SOCIETY FOR MOLECULAR BIOLOGY & EVOLUTION
DUBLIN 2012

Smbe
Society for Molecular Biology & Evolution

SMBE 2012 • Society for Molecular Biology & Evolution

June 23rd - 26th 2012 • The Convention Centre Dublin

CONFERENCE ABSTRACTS

F-1

Where classic theory meets genomic technology: the distribution of fitness effects in yeast

Claudia Bank[1], Jeffrey D. Jensen[1], Daniel N. A. Bolon[2]
[1]School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland,[2]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA, USA

One of the most controversial questions in evolutionary biology is the role of adaptation in molecular evolution. After decades of debate between selectionists and neutralists, new high-throughput methods are beginning to illuminate the full distribution of fitness effects (DFE) of new mutations. Using an innovative approach called EMPIRIC ("extremely methodical and parallel investigation of randomized individual codons"), it is possible to gain data on the fitness of all possible single point mutations in *Saccharomyces cerevisiae*, under a variety of environmental conditions. We use a uniquely large dataset from eight regions of four different proteins, yielding a total of 4800 amino acid substitutions and more than 14000 codon variants - for each of which we obtain a selection coefficient. In addition, this data is available not only for equilibrium growth conditions, but also for high and low temperature, and high and low salinity – allowing for the direct evaluation of hypotheses governing adaptation to novel environments. Hence, this dataset provides us with the opportunity to address a number of fundamental questions: What is the shape of the DFE? Does it differ between proteins, and between conserved and divergent regions of a protein? What is the relative proportion of deleterious, neutral and beneficial mutations? How do novel selective pressures alter the DFE? Do beneficial mutations tend to be locally or globally advantageous? Directly associated with these questions are several classical theories of population genetics, in particular the neutral and nearly-neutral theories of molecular evolution, and a proliferating literature for statistically inferring the DFE. Thus far, results show a remarkable correspondence to the world-view of Kimura/Ohta – an observation which we place in the context of modern genomic analyses.

F-2

Decoding the genomic and transcriptional changes underlying the transition to multicellularity in experimentally-evolved *Saccharomyces cerevisiae*.

Johnathon D. Fankhauser, William C. Ratcliff, R. Ford Denison, Michael Travisano
University of Minnesota, St Paul, MN, USA

The transition from unicellular to multicellular life is one of the most important evolutionary innovations in the history of life on earth. Starting with unicellular *Saccharomyces cerevisiae,* we experimentally evolved simple multicellular strains with specialized cellular behavior and gene regulation. Using RNA-seq and genome sequencing we are able to dissect the genetic mechanism of this evolutionary transition. We have identified at least three alternative evolutionary directions in ten independently evolving lines ranging from large-scale transcriptional changes and hypermutation to convergent genomic variations conferring a selective advantage. The multiple genetic routes to multicellularity suggest that this transition may not be as constrained as previously expected. We also explore constraints on genomic stability, dispensability, and mutation rates in different genetic functional groups. We investigate the molecular evolution of regulatory sequences, identifying the complex phenotypic consequences of differential gene expression and the influence of transcriptional regulation on adaptability. Investigating regulatory dynamics were essential to understanding the rapid adaption to multicellular life. An exciting observation from whole transcriptome analysis is we find that cellular adhesion in our multicellular yeast is not due to flocculation. Indeed, loci associated with flocculation are down-regulated in multiple lines, potentially reducing exploitation by adhesive unicellular competitors. The importance of reduced flocculation is illustrated by evidence of convergent evolution in multiple gene types from cell surface glycoproteins to transcription factors.

F-3

Reconstructing demographic histories from long tracts of DNA sequence identity

Kelley Harris[1], Rasmus Nielsen[1,2]
[1]UC Berkeley, Berkeley, CA, USA, [2]University of Copenhagen, Copenhagen, Denmark

There has been recent excitement and debate about the details of human demographic history, involving gene flow that has occurred between populations as well as the extent and timing of bottlenecks and periods of population growth. Much of the debate concerns the timing of past admixture events; for example, whether Neanderthals exchanged genetic material with the ancestors of non-Africans before before or after they left Africa. Here, we present a method for using sequence data to jointly estimate the timing and magnitude of past genetic exchanges, along with population divergence times and changes in effective population size. To achieve this, we look at the length distribution of regions that are shared identical by state (IBS) and maximize an analytic composite likelihood that we derive from the sequentially Markov coalescent (SMC). Recent gene flow between populations leaves behind long tracts of identity by descent (IBD), and these tracts give our method its power by influencing the distribution of shared IBS tracts. However, since IBS tracts are directly observable, we do not need to infer the precise locations of IBD tracts. In this way, we can accurately estimate admixture times for relatively ancient events where admixture mapping is not possible, and in simulated data we show excellent power to characterize admixture pulses that occurred 100 to several hundred generations ago. When we study the IBS tracts shared between and within the populations sequenced by the 1000 Genomes consortium, we find evidence that there was no significant gene flow between Europeans and Asians within the past few hundred generations. It also looks unlikely that the Yorubans of Nigeria interbred with Europeans or Asians in a population-specific way, though there may have been admixture between Africans and an ancestral non-African population.

F-4

How do hosts shape their microbial communities?

Matthew Horton, Natacha Bodenhausen, Joy Bergelson
University of Chicago, Chicago, Il, USA

Earlier results from greenhouse experiments suggest that individual accessions of *Arabidopsis thaliana*, the plant genetic model, harbor distinct microbial communities. To date, however, it is unclear whether the composition of a host's microbial community is a heritable trait in a natural setting, or which (if any) host factors are responsible. We have conducted a field experiment in which we allowed 4 replicates of each of ~200 accessions of *A. thaliana* to recruit their bacterial and fungal communities. Characterizing each individual accession's microbial community with next generation sequencing allows us unparalleled insights into the relationship between a host in its natural setting, and its associated microbial community. We find a strong and statistically significant association between host-genotype and the ordinations produced from three complementary community ecology techniques (non-metric multidimensional scaling, correspondence analysis, and canonical correspondence analysis). We therefore scanned the genome of *A. thaliana* using genome-wide polymorphism data to identify the loci associated with the abundances of individual microbial species. I will discuss the most interesting candidate loci, as well as the gene ontology categories that show enrichment for these associated SNPs. Our results are consistent with the idea that host associated microbial communities are heritable traits, and provide clear genetic candidates for follow-up study.

F-5

Repeatability in evolution varies with scale, organism, and the nature of selection.

Elizabeth B. Perry, Kristin Alligood, Brendan J.M. Bohannan
University of Oregon - Institute of Ecology and Evolution, Eugene, Oregon, USA

Understanding the patterns and causes of genome evolution is a major challenge in evolutionary biology. We allowed seven replicate communities of bacteria (*Escherichia coli B*) and bacteriophage (T3) to evolve in continuous culture (chemostats) for approximately 200 bacterial generations. We used Illumina re-sequencing technology to sequence whole-genomes of 14 phage-resistant bacteria and 38 host-range mutant bacteriophage that evolved in the replicate environments. By comparing the genomes of derived phenotypes to the ancestral genomes, we were able to identify the genomic changes associated with new phenotypes. By comparing genomic changes across replicate experiments, we were able to assess the repeatability of evolution at different biological scales - from biological pathways to genes, codons, and nucleotides.

In each chemostat, bacteria evolved that are resistant to infection by the ancestral phage. All of the first-order resistant bacteria displayed mutations in a gene that codes for glucosyltransferase I (*waaG*). This enzyme is involved in synthesis of the *E.coli* lipopolysaccharide (LPS), which is the surface structure to which T3 binds in adsorption. Each mutation occurred at a unique position within the *waaG* gene, and a variety of mutation types were observed including missense and nonsense mutations, insertions and deletions. Host-range mutant phage also evolved in each chemostat that infect first-order resistant bacteria as well as the ancestral bacterial type. Each of the host-range mutant phage has one of two substitutions. Both of the substitutions are non-synonymous and fall within the same codon of the phage tail fiber gene 17, which is the protein that binds to the bacterial surface during adsorption. Second-order resistant bacteria evolved in each chemostat that are resistant to the ancestral phage as well as host-range mutant phage. The mutations in these bacteria are dominated by large deletion events facilitated by insertion sequence (IS) transposable elements. The second-order resistant strains evolve resistance to phage by altering at least two different pathways.

We observed a striking degree of repeatability in this evolutionary system. The scale at which evolution is repeatable varies with organism and the nature of selection.

F-6

Which way did they go? Detecting directional migration from genetic data

Benjamin Peter, Montgomery Slatkin
University of California, Berkeley, Berkeley, USA

Range expansions and colonizations are ubiquitous in many species and are studied from many different perspectives in e.g. anthropology, biogeography and invasion biology. It has been well established that these colonization events lead to a loss of genetic diversity and that in many cases it is possible to infer the history of a species' range from present-day genetic data. Previous approaches were mainly based on within-population measures of diversity such as heterozygosity, which then have been compared between populations. However, it is also well established that these statistics are susceptible to confounding demographic factors such as unequal subpopulation sizes or population size changes.

In this study, we propose a novel method using data from multiple populations to infer a population's history. Our approach is based on a statistic that detects asymmetries in the 2D-allele frequency spectrum that occur when one population consists mostly of offspring of another population, as we expect in an expanding population. We show that our statistic is able to detect the direction of an expansion using data from multiple populations. Using simulations, we further show that our statistic is generally more powerful than previous approaches and that it is robust to a wide array of confounding demographic factors. We further illustrate the use of our statistic on several data sets for humans, Drosophila and Neurospora and show that we are both able to detect global patterns of colonization and fine-scale population structure.

F-7

Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency

Wenfeng Qian, Jianzhi Zhang
University of Michigan, Ann Arbor, MI, USA

Cellular efficiency in protein translation is an important fitness determinant in rapidly growing organisms. It is widely believed that synonymous codons are translated with unequal speeds and that translational efficiency is maximized by the exclusive use of rapidly translated codons. Using next-generation-sequencing-based ribosomal profiling data, we estimated for the first time the *in vivo* translational speeds of all 61 sense codons from the budding yeast *Saccharomyces cerevisiae*. Surprisingly, preferentially used synonymous codons are not translated faster than unpreferred ones, and no correlation exists between the translational speed of a codon and the concentration of its cognate tRNA. We hypothesize that the phenomenon of similar translational speeds of different synonymous codons is a result of proportional use of synonymous codons according to their cognate tRNA concentrations, the optimal strategy in enhancing translational efficiency under tRNA shortage, which is a cellular condition that has circumstantial empirical evidence but has not been seriously considered in the past. Our hypothesis predicts that, for each amino acid, the fractional use of a codon among its synonymous codons equals the fractional concentration of its cognate tRNA among all isoaccepting tRNAs, and this is indeed the case in all eukaryotic model organisms examined (*S. cerevisiae*, *S. pombe*, *A. thaliana*, *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. sapiens*). We further tested our hypothesis by a manipulative experiment in which multiple synonymous versions of a heterologous red florescent protein gene were highly expressed to induce different levels of codon-tRNA imbalance in yeast. We measured the expression level of a yellow florescent protein gene serving as a reporter that indicates the overall cellular translational efficiency. This inducer-reporter experimental system excluded multiple confounding factors such as the potentially different translational accuracies of synonymous codons. Our results unambiguously support the hypothesis that codon-tRNA balance, rather than exclusive use of preferred codons, optimizes cellular translational efficiency. Our hypothesis also applies to amino acid usage, suggesting that it again is shaped by selection for translational efficiency. Together, our study reveals a previously unsuspected mechanism by which unequal codon usage increases translational efficiency, demonstrates widespread natural selection for translational efficiency, and offers new strategies to improve synthetic biology.

F-8

Successive masculinization of the avian Z chromosome

Alison Wright[1,2], Hooman Moghadam[1], Judith Mank[2]
[1]University of Oxford, Oxford, UK, [2]University College London, London, UK

Due to their unequal pattern of inheritance in males and females, the sex chromosomes are subject to unbalanced sex-specific selection and thus exhibit a non-random distribution of sex-biased genes compared to the remainder of the genome. Theoretically, the degree of sex-bias on the sex chromosomes should increase over time with cumulative exposure to sex-specific selection, although this has not been definitively documented. The chicken Z chromosome provides a useful opportunity to study the role of male-specific selection in shaping the evolution of the genome as it is present twice as often in males than females, and therefore predicted to be enriched for dominant male-specific effects. The avian Z chromosome is separated into multiple strata, formed by successive inversions events over the course of roughly 130 million years of avian evolution. Here, we expand and refine our understanding of chicken Z chromosome geology using several newly-discovered Z-W orthologs, and find support for at least three and possibly four strata (formed approximately 133, 73, 63 and 46 million years ago). By comparing the magnitude of sex-biased expression among the strata, we uncover a pattern consistent with masculinizing selection, where the magnitude of male-biased expression is predicted by the age of the stratum (R^2=0.99). Additionally, the rate of this masculinization appears to decline over time, a finding that has important implications regarding the genetic basis of sexual dimorphism.

O-1

A Whole-genome Polymorphism Scan in Two *Aquilegia* (Columbine) Species Identifies Putative Speciation Genes.

Danièle Filiault[1], Elizabeth Cooper[2], Nathan Derieg[3], Scott Hodges[3], Magnus Nordborg[1]
[1]Gregor Mendel Institute for Plant Molecular Biology, Vienna, Austria, [2]University of Southern California, Los Angeles, CA, USA, [3]University of California, Santa Barbara, Santa Barbara, CA, USA

The genus *Aquilegia* has undergone a recent adaptive radiation in North America, making this group of plants an excellent choice for studying speciation. Two of these species, *A. pubescens* and *A. formosa*, have been the focus of many studies aimed at understanding the genetics of reproductive isolation. The flower morphologies of these two species reflect pollinator differences; hummingbird-pollinated *A. formosa* has red pendant flowers with short nectar spurs, while hawkmoth-pollinated *A. pubescens* has white upright flowers with long spurs. Despite marked differences in pollination syndrome and habitat, these species are interfertile, both in the laboratory and in the field. In fact, a study of genetic variation in nine nuclear genes concluded that little genetic differentiation exists between these species. This result suggested that a whole-genome scan for fixed polymorphisms could be a useful tool for identifying speciation genes. In this work, we performed a scan looking for single nucleotide polymorphisms (SNPs) showing high differentiation between *A. pubescens* and *A. formosa*. We sequenced six population pools of each of the two species with next generation sequencing and mapped reads to the *A. coerulea* reference sequence. Although more than a half-million unique polymorphisms were identified within each of the two species, fewer than 1000 of these SNPs were fixed or nearly-fixed between the species. Many of these SNPs occurred in distinct “peaks” of genetic differentiation that could represent areas of high linkage disequilbrium associated with loci under selection. One of these regions corresponds to the location of a major-effect QTL for flower color mapped in an *A. pubescens* x *A. formosa* F2 population. Interestingly, one nearly-fixed SNP in this region is predicted to encode a non-synonymous change in a MYB transcription factor that has been shown to be involved in the synthesis of anthocyanin pigments in other plant species. Since flower color often plays a key role in pollinator attraction, changes in this putative speciation gene could have lead to increased reproductive isolation between *A. pubescens* and *A. formosa*. Highly differentiated SNPs coding for non-synonymous changes in a number of other genes were also identified in the genome scan. An understanding of whether these genes contribute to speciation will be facilitated both by additional functional genetics work and through a forthcoming increase in the number and quality of *Aquilegia* reference sequences.

O-2

Drug resistance in malaria - Revealing differences in selective pressure on *dhps* and *dhfr* mutations in western Kenya based on a new theoretical model

Andrea M. McCollum[1], Kristan A. Schneider[2], Ananias A. Escalante[3]
[1]Malaria Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA, [2]Department of Mathematics, University of Vienna, Vienna, Austria, [3]Center for Evolutionary Medicine & Informatics, The Biodesign Institute, Arizona State University, Tempe, AZ, USA

Kristan A. Schneider

Assisant Professor, Department of Mathematics, University of Vienna

Organizer SMBE Symposium: Population genetics of demography and adaptation

Abstract:

Anti-malarial drug resistance is the most severe threat to successful malaria control and eradication efforts. Indeed the recent emergence of artemisinin-tolerant parasite strains raises the concern that the repertoire of save and effective antimalarials is soon exhausted. Understanding the origin and spread of mutations associated with drug resistance, especially in the context of combination therapy, will help guide strategies to halt and prevent the emergence of resistance. Unfortunately, information on the evolutionary dynamics leading to multidrug-resistant parasites is scattered and limited to areas with low or seasonal malaria transmission.
We describe the dynamics of strong selection for mutations conferring resistance against sulphadoxine-pyrimethamine (SP), a combination therapy, in a hyper-endemic area in western Kenya between 1992 and 1999, just before SP became first-line therapy (1999). Importantly, the study is based on longitudinal data, which allows for a comprehensive analysis that contrasts with previous cross-sectional studies carried out in other endemic regions. Based on a novel genetic model tailored to malaria, we are able to estimate selection pressures on various *dhfr* and *dhps* mutants and link these estimates to the pattern of STR variation. We discuss how transmission intensities can affect the onset and dispersion of resistance-associated mutations and the pattern of neutral genetic variation flanking resistance-associated loci.

References:

Andrea M. McCollum, Kristan A. Schneider, et al. **Differences in selective pressure on *dhps* and *dhfr* drug resistant mutations in western Kenya.** Malaria Journal, In Press.

Kristan A. Schneider and Yuseob Kim (2011). **Approximations for the Hitchhiking Effect caused by the Evolution of Antimalarial-Drug Resistance**. J. Math. Biol. 62:789-832.

Kristan A. Schneider and Yuseob Kim (2010). **An Analytical Model for Genetic Hitchhiking in the Evolution of Antimalarial-Drug Resistance**. Theor. Popul. Biol. 78(2), 93-108.

O-3

Evolution of *de novo* originated overlapping genes: a molecular tug of war

Niv Sabath[1,2], Andreas Wagner[1,2], David Karlin[3]
[1]University of Zurich, Zurich, Switzerland,[2]The Swiss Institute of Bioinformatics, Basel, Switzerland,[3]Oxford University, Oxford, UK

Most studies to understand the origin and evolution of new genes have focused on those originated through modifications of existing genes (e.g. gene duplication or fusion). In contrast, little is known about genes originated *de novo*. Viral genomes often contain a special case of such genes: those originated by overprinting an existing reading frame, forming overlapping genes. We present an evolutionary analysis of 14 viral genes originated *de novo* (*De novo* genes) by overprinting, whose existence is experimentally proven. We estimated their relative ages by using a domain conserved in all, the polymerase domain. Further, we examined their codon usage, evolutionary rate, and selection pressure, using a model especially adapted to overlapping genes. We found that the young *De novo* genes have a different codon usage relative to the rest of their genome, evolve fast, and are under positive or weak purifying selection. Old *De novo* genes have a codon usage that is similar to that of their genome, evolve slowly, and are under stronger purifying selection. Some of the old *De novo* genes were found to evolve under stronger selection pressure than the overlapping ancestral gene, suggesting an evolutionary tug of war between the two genes over the dominance of the sequence.

O-4

Unraveling the life dynamics of Sirevirus LTR Retrotransposons: major players in the organization and evolution of the maize and other angiosperm genomes

Alexandros Bousios[1], Yiannis Kourmpetis[2], Pavlos Pavlidis[3], Evangelia Minga[1], Nikos Darzentas[1]
[1]Bioinformatics Analysis Team, Institute of Agrobiotechnology, CERTH, Thessaloniki, Greece,[2]Laboratory of Bioinformatics, Wageningen University, Wageningen, The Netherlands,[3]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Sireviruses is an ancient, and with a unique genome structure, LTR retrotransposon genus of the Copia superfamily, and the only one that has exclusively proliferated within plant genomes. However, as a result of receiving little research interest until now, the extent of their colonization and impact on their host genomes remained unclear. Here, aided by the recent development of a purpose-built algorithm, we report that Sireviruses have infiltrated most phylogenetic branches of the plant kingdom, extensively colonizing genomes such as that of soybean, sorghum and lotus. In maize, Sireviruses reached massive numbers to form a plethora of autonomous and non-autonomous families with distinct genome characteristics, some outlined herein for the first time. Sireviruses occupy 21% of the maize genome and comprise 90% of the Copia population, experiencing intense amplification during the past 600,000 years, and mediating the formation of gene islands by targeting their own genomes in chromosome-distal gene-rich areas. Intriguingly, this spatial and temporal integration pattern is not universal, as evident by their pericentromeric preference in soybean and by their much older amplification burst in cacao. Maize Sireviruses are constantly recycled by host mechanisms, exhibiting a significantly higher solo LTR formation rate than previously reported for known maize LTR retrotransposons. Their LTRs are heavily methylated, whilst they also form recombination hotspots by producing vast numbers of indels with specific lengths of 19-22bp. Finally, there is evidence for a palindromic consensus target sequence. To support further studies into these infiltration patterns, their evolutionary depth and impact on their hosts, and also facilitate genome annotation projects, we developed a highly curated database that catalogues Sireviruses in eleven fully-sequenced plants, currently housing approximately 16,200 elements. Overall, this multi-faceted work brings for the first time Sireviruses, together with a unique set of tools and data for the scientific community, under the spotlight (http://bat.infspire.org/sireviruses/).

O-5

Is dioecy an evolutionary dead end for plant species? A case study in the genus Silene

Jos Käfer[1], Martina Talianova[2], Alex Widmer[3], Gabriel Marais[1]
[1]Laboratoire de Biométrie et Biologie Évolutive, UMR 5558 CNRS/Université Lyon 1, Lyon, France, [2]Department of Plant Developmental Genetics, Institute of Biophysics, Brno, Czech Republic, [3]Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

Dioecy (i.e. having separate sexes) is a rather rare breeding system in flowering plants, probably due to a high probability of extinction in dioecious species. This may result from less efficient dispersal and the costs of sexual selection, which are expected to harm dioecious species survival on the long term. These handicaps should decrease the effective population size (Ne) of dioecious species, which, in turn, should reduce both polymorphism levels and the efficacy of selection.
Here we tested these predictions using sequences of various sources from the Silene genus (Caryophyllaceae), where dioecy has evolved at least twice. For the dioecious species in the section Melandrium, where dioecy is the oldest, the chloroplast sequences show a reduced efficacy of purifying selection, while there are conflicting forces on the nuclear genomes: a global reduced purifying selection and an enhanced positive selection for maleness on some genes. For the more recently dioecious section Otites, we found no significant effect of dioecy on genome evolution. Our results are thus in agreement with the hypothesis that dioecy is an evolutionary dead end in flowering plants. However, they also show that contrasting forces act on the genomes, and suggest that some time is required before the genome of dioecious plants bears the footprints of reduced Ne.

O-6

Devils and disease: Tracing spread and impacts of Devil Facial Tumour Disease on Tasmanian devil populations

Anna Brüniche-Olsen[1], Elizabeth Murchison[3], Jeremy Austin[2], Chris Burridge[1], Menna Jones[1]
[1]University of Tasmania, Hobart, TAS, Australia,[2]University of Adelaide, Adelaide, SA, Australia,[3]The Wellcome Trust Sanger Institute, Hinxton, CB, UK

Emerging infectious diseases of wildlife are recognized as one of the greatest threats to global biodiversity. The clearest case yet of the devastating consequences of low genetic diversity following the emergence of infectious disease is the current extinction threat to the world's largest marsupial carnivore, the Tasmanian devil, caused by the novel and unusual contagious cancer, Devil Facial Tumour Disease (DFTD). The Tasmanian devil was widespread in Australia 400 years ago, but is now only found in Tasmania. The species has unusually low genetic diversity, at both nuclear microsatellite and immune-gene loci, which may have facilitated the spread of DFTD.

Since DFTD was first detected in 1996 at Mt. William, it has spread to the majority of the species' geographic range, causing more than 80% overall population decline reducing the total population to 10,000–25,000 individuals. DFTD has led to increased inbreeding and geographic fragmentation of genetic variation, and life history changes, with a transition from iteroparity towards precocious single breeding. The rapid spread of DFTD is expected to cause further loss of genetic diversity, leading to a reduction in adaptive potential and possibly extinction.

Here we present the most detailed accession of DFTD impacts on mitochondrial genetic diversity. Using comprehensive geographic and temporal sampling we compare changes in mitogenome diversity at DFTD affected and unaffected populations through space and time. We sequenced mitogenomes from 500 individuals, representing the species current geographic distribution, sampled from DFTD emergence until present. Population structure, gene flow and effective population sizes were estimated pre- and post DFTD arrival for affected populations with comparisons against control populations, which remained DFTD free during the same time interval.

The results from this study provide a detailed understanding of the effect of DFTD-related population declines on genetic diversity of the Tasmanian devil. The results will inform strategies to maintain, restore, or enhancement genetic diversity of both wild and captive insurance populations, which in turn will assist with the maintenance of devil populations in the wild.

O-7

Paleogenomics: hunting our genetic past.

Eske Willerslev
Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen University, Copenhagen, Denmark

The advent of Next-Generation Sequencing has propelled ancient DNA research in the genomic era, opening new avenues for identifying and timing the origin of human populations. By characterizing the complete genomes of one paleo-Eskimo of the Saqqaq culture, and of one ancient Australian Aborigine, our laboratory has participated in the emergence of paleogenomics as a new field of research, aiming at sequencing and analyzing the complete genomes of individuals from the past.
These approaches have enabled us to revisit current models of the colonisation of the Arctic Greenland as well as to decipher with more details the chronology of the 'Out-of-Africa' event. Ancient genomes have revealed unexpected features, that were virtually untraceable from the sole analysis of modern genomes. I will review these achievements and will present new lines of research currently developped in our laboratory.

O-8

Next generation sequencing of archaeological strains of *Mycobacterium tuberculosis*

Terry Brown
University of Manchester, Manchester, UK

Tuberculosis (TB) is a re-emerging disease that affects one-third of the world's population and was responsible for 1.8 million deaths in 2008. Human TB is caused by *Mycobacterium tuberculosis*, a member of a related complex of species that give rise to the disease in different animals. Increasingly detailed phylogenetics is revealing that *M. tuberculosis* has extensive genetic variation and that this variation has, at least in part, a geographical basis. These studies suggest that *M. tuberculosis* has been in existence for 2.5–3.0 million years, but tell us little about the evolution of the bacterium. Palaeogenetics therefore has the potential to provide information directly relevant to our understanding of the modern disease. Although TB is mainly a pulmonary disease, in some patients the bacteria spread to bones and cause lesions that can be recognized in archaeological skeletons. Ancient DNA (aDNA) from *M. tuberculosis* has been detected in skeletons displaying TB lesions, but analysis of archaeological TB has been hampered by the problems that have afflicted aDNA research, compounded by the presence in bones of contaminating environmental mycobacteria, which make it difficult to design PCR systems that are specific for *M. tuberculosis*. We have shown that these problems can be circumvented by use of next generation sequencing. We designed a target enrichment system directed at 252 SNPs and 33 other variable loci in the *M. tuberculosis* genome, and then used the SOLiD platform to obtain sequences from the enriched aDNA extracts. We used this approach with 100 skeletons displaying TB lesions, dating from 100 to 2000 years old. In all cases we obtained sequence reads that matched *M. tuberculosis* DNA, though the number of SNPs that were covered was variable, ranging from zero to 252. We have shown that a variety of strains of TB were present in England during the last 2000 years, and that multiple strains were present at the same time. Some of these historic strains are closely related to modern varieties. We identified one individual from the 3rd century AD who was coinfected with two strains. Our results demonstrate that sufficient information can be obtained from archaeological remains to attempt detailed comparisons between historic strains of *M. tuberculosis*, and to map changes in the population genetics of those strains over time.

O-9

Coprolites as a source of information on the genome and diet of the cave hyena

Jean-Marc Elalouf
CEA Saclay, Saclay, France

We performed high-throughput sequencing of DNA from fossilized faeces to evaluate this material as a source of information on the genome and diet of Pleistocene carnivores. We analysed coprolites derived from the extinct cave hyena (*Crocuta crocuta spelaea*), and sequenced 90 million DNA fragments from two specimens using the Illumina strategy. The DNA reads enabled reconstruction of complete cave hyena mitochondrial genomes with up to 158-fold coverage. Together with complete mitochondrial genomes for the spotted (*Crocuta crocuta*) and stripped (*Hyaena hyaena*) hyena that we sequenced from extant specimens, the cave hyena sequences allow establishing a robust phylogeny that supports a close relationship between the cave and spotted hyenas. Through identification of nuclear DNA, we demonstrate that the high-throughput strategy yields sequence data for multicopy as well as for single-copy cave hyena nuclear genes, and that about 50% of the coprolite DNA can be ascribed to this species. Analysing the data for additional mammalian species that may highlight the cave hyena diet, we retrieved abundant sequences for the red deer (*Cervus elaphus*), and characterized its mitochondrial genome with up to 3.8-fold coverage. We conclude for the presence of abundant ancient DNA in the coprolites surveyed. Shotgun sequencing of this material yielded a wealth of DNA sequences for a Pleistocene carnivore, and allowed unbiased identification of its diet.

O-10

Ancient DNA and the history of domestic cattle

Dan Bradley, Matthew Teasdale, Frauke Stock
Trinity College, Dublin, Ireland

Cattle and the other major domestic animals emerge in the remarkably innovative early agricultural communities of the Near East over 10,000 years ago and proceed to dominate food, economy and culture. Through their domestication humans profoundly altered their relationship with nature, controlling the breeding of their major food sources for material, social or symbolic profit. Understanding this complex process is a compelling research aim. mtDNA sequence diversity has been valuable in sketching out aspects of this process. Recent developments include the ramping up of sequence output to the level of whole mtDNA genomes, 50k SNP chip data and autosomal gene resquencing. These will be described. Time-stamped genetic variants are particularly valuable for calibrating the molecular clock, eg with mtDNA chromosome evolution, as the temporal distinctions required are often fine grained. Inference from variation points toward a domestication process with focus but not without some biogeographical complexity.

O-11

Populations genetics of the Neolithic transition

Joachim Burger[1], Mark Thomas[2,3]
[1]Johannes Gutenberg University, Institute of Anthropology, D-55128 Mainz, Germany, [2]Research Department of Genetics, Evolution and Environment, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK, [3]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvagen 18D, SE-752 36 Uppsala, Sweden

About 11,000 years ago, a change in human lifestyle took place in the territories of present-day western Iran, the Levant region and south-east Anatolia, which is characterised particularly by four factors: the people founded permanent settlements with buildings for various functions; plants such as Einkorn and beans were cultivated; goats, sheep, pigs and cattle were domesticated; a new kind of culture evolved, that became conspicuous with the appearance of a new material culture including ground stone tools and later, pottery products. The transition from the partly nomadic hunter-gatherer subsistence strategy to a settled lifestyle based on food production is also known as the “Neolithic Revolution”. About 8,500 years ago, the Neolithic culture spread to the southeast of Europe and later expanded episodically across central and northern Europe. The extent to which this movement of a farming culture was accompanied by a movement of people, as opposed to just a spread of ideas and skills, has been a subject of considerable debate and dispute over the last 100 years. Population genetic computer simulations of genetic data from ancient human remains, based on coalescent theory have shown that the early Neolithic farmers could not have been descended just from the later hunter-gatherers of central Europe (Bramanti et al. 2009). As the hunter-gatherers had already been settled in Central Europe since the retreat of the glaciers 20 kya, Neolithic famers must have migrated into this area.

There is good evidence of cultural contact between hunter-gatherers and early farmers in central Europe. Whether the exchange of hunting tools also led also to the exchange of men is still not clear, as Y-chromosomal DNA has not yet been studied systematically in ancient human remains. Moreover, ancient DNA evidence is now emerging that other regions don´t follow the patterns of population discontinuity observed in Central Europe. While the overall results support a model of demic diffusion of farmers from southeastern Europe, or even further East, in to Central Europe, it is very likely that modern populations in most parts of Europe were formed by varying degrees of admixture between incoming farmers and indigenous hunter-gatherers. Analyses of the appropriate neutral and phenotypically informative markers using next generation sequencing technologies will provide more information on this in the near future.

O-12

Archaeogenomic evidence of local adaptation in crops

Robin Allaby, Sarah Palmer, Oliver Smith
University of Warwick, Coventry, UK

Evidence from phylogeographic patterns suggests that many crop lineages existed *in situ* at geographic locations for millennia, giving the opportunity for specific local adaptation. The genetic variation associated with useful adaptations may be applicable to modern crop development. Barley grown by the ancient Nubians is interesting because of a phenotypic abnormality that gives the crop a 2-row appearance, despite archaeogenetic evidence that the crop was originally 6-row. Moreover, apparently the same crop lineage was passed through five successive cultures, even through violent transitions. This evolutionary sequence is likely to be explained by drought adaptation, which was undoubtedly a source of stress in the area in which the archaeological site, Qasr Ibrim, is situated, upstream of the first cataract of the River Nile. Using next generation sequencing with archaeobotanical material we have established that large-scale genomic level change in plants of this region has occurred in a punctuated fashion illustrating the rapidity with which domesticated plants have evolved within the Holocene. We have also studied the archaeoepigenome and found fundamental differences in microRNA levels between the Nubian archaeological barley, and modern barley, which centre around the 'Green Revolution' gene interactions. These suggest that the archaeological barley was more primed than modern for germination, perhaps linked to the very short growing season available, but also had down regulation of mitochondria, which may provide the first clues as to how the abnormal phenotype was generated. We are also employing a DNA capture approach to study the portion of the barley genome most likely to be involved with drought adaptation. The archaeogenomic evidence so far has turned up a number of surprising and independent connections to mutations associated with that part of the genome associated with the 'Green Revolution' that is leading us to speculate about Green Revolutions across the ancient world.

O-13

Evolutionary approaches to vaccine design: Influenza A and *Chlamydia trachomatis.*

Robin Bush
University of California, Irvine, Irvine, CA, USA

Genomic sequencing has allowed us to track the spread of infectious disease on an ever more detailed level. However, we are still far from understanding the relationship between pathogen evolution on a genomic scale and the immunological response of the host. I use the influenza A virus and the bacterium *Chlamydia trachomatis* to illustrate methods for the identification of epitopes important in vaccine development using a synthesis of structural, immunological, genetic and epidemiological data.

O-14

Discovery and population genomics of *Drosophila* viruses: No evidence for an arms-race

Darren Obbard, Claire Webster
University of Edinburgh, Edinburgh, UK

Drosophila melanogaster is an important model for innate immunity, and is arguably our primary model for antiviral resistance in arthropods. Several groups have used population-genetic and phylogenetic approaches to show that some antiviral immune genes in *Drosophila* (notably the antiviral RNAi pathway) display highly elevated rates of adaptive evolution. However, although this is consistent with a host-virus arms race, the evolutionary genetics of Drosophila viruses are almost unstudied - only a handful of viruses which naturally infect *Drosophila melanogaster* are known, and only Drosophila Sigma Virus (a Rhabdovirus) has been regularly isolated from wild populations. Other viral pathogens (such as Drosophila C Virus, Drosophila A Virus and Drosophila Nora virus) occur at unknown prevalence and are little-studied outside of the lab.

In an attempt to understand the evolutionary genetics of Drosophila viruses, we have sequenced both rRNA-depleted RNAseq, and small-RNA, libraries from large pooled samples of wild-caught *D. melanogaster*. This has allowed us to identify several new viruses, including a wide diversity of RNA viruses (viruses with sequence-similarity to Sacbrood Virus, Slow Bee Paralysis Virus, Chronic Bee paralysis virus, Acyrthosiphon Pisum Virus, Brevicoryne brassicae picorna-like Virus, Flaviviruses, and Cypoviruses) and a DNA virus (Nudivirus).

Following a wider geographic survey of *D. melanogaster* using RT-PCR, we find that the previously known viruses of *D. melanogaster* (including DAV, Sigma and Nora) are widespread at low to intermediate prevalence, but that the new viruses vary widely in prevalence and distribution. None of the viruses shows high rates of adaptive evolution, and in general (despite substantial synonymous divergence) protein sequences are very highly conserved. However, while this may indicate that these viruses are not engaged in ‘arms race'-like coevolution, we suspect that the short timescale of viral co-ancestry (tens to hundreds, rather than thousands, of years) makes this process extremely difficult to detect. This is in sharp contrast to viral evolution in response to vertebrate adaptive immunity, which adapts plastically on the same timescale as viral evolution.

O-15

Molecular evolution of viral epitopes: frequent associations and selection

Helen Piontkivska, Sinu Paul, Reeba Paul, Patrice Conway, Michael Rose, Joel Serre
Kent State University, Kent, USA

During viral infection, an important role in controlling the infection is played by recognition of viral peptides by class I major histocompatibility complex (MHC) molecules and subsequent interactions with the cytotoxic T lymphocytes (CTLs). Viral epitopes are a critical part of these interactions, and persistent selective pressure on epitopes can enable viral "escape" through amino acid changes that disrupt recognition by the immune system. Notably, some CTL epitopes harbor very few amino acid substitutions despite ongoing pressure from the immune system. We have recently described a set of so-called "associated epitopes" (Paul and Piontkivska 2009, 2010) that consists of CTL epitopes that frequently co-occur together among different subtypes of HIV-1, including circulating recombinant forms, and exhibit signs of strong purifying selection acting at both the amino acid as well as nucleotide levels. While the unusually low level of sequence variability at these epitope regions can be attributed to the strong structural/functional constraints, it is important to understand whether and how the selective pressure acts across different associated epitopes. Patterns of nucleotide substitutions in epitope regions of human immunodeficiency virus (HVI-1) are contrasted to better understand the molecular evolutionary forces driving sequence changes at individual epitopes.

O-16

The rate of adaptation of pandemic H1N1 influenza virus.

Jessica Hedge, Samantha Lycett, Andrew Rambaut
University of Edinburgh, Edinburgh, UK

A novel H1N1 influenza A virus (H1N1pdm) emerged in April 2009, giving rise to two waves of the pandemic in 2009. For many countries in the Northern hemisphere, these outbreaks appeared in the summer months, when influenza transmission is typically very rare. Subsequent H1N1pdm epidemics have occurred within typical flu seasons, in which the pre-pandemic H1N1 strain has only been detected at very low levels while H1N1pdm has been the cause of many of the cases identified. In contrast, the existing H3N2 subtype remained to circulate throughout the pandemic in populations worldwide and has continued to do so. We use whole genome sequences of H1N1pdm and population genetics methods to estimate the rate of adaptation of this strain from emergence to it's transition to the endemic strain that currently circulates in temperate regions of the world. We find variation in the rate of adaptation between different segments of the genome and investigate the emergence and fate of new mutations that arose in the pandemic stages of it's evolutionary history. We combine these analyses with Bayesian phylogenetic analysis to estimate the population dynamics of H1N1pdm and explore the coevolution between the host population immunity and this novel strain.

O-17

The genomic accordion model of large DNA virus evolution

Nels Elde[1,2], Stephanie Child[2], Michael Eickbush[2], Jacob Kitzman[3], Jay Shendure[3], Adam Geballe[2,3], Harmit Malik[2,4]
[1]University of Utah, Salt Lake City, UT, USA, [2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [3]University of Washington, Seattle, WA, USA, [4]Howard Hughes Medical Institute, Seattle, WA, USA

Host-pathogen relationships often put populations on the brink of existence. These ongoing struggles for survival place intense selective pressure on both pathogens and hosts, and spur some of the most radical adaptations observed in nature. A current gap in our knowledge of virus evolution is an explanation for the ability of medically relevant, large DNA viruses like poxviruses to adapt and infect large host ranges. How do large DNA viruses evade diverse host defenses despite an inability to generate mutations at rates necessary for other classes of viruses to avoid host immunity and propagate? We devised an experimental system using serial passaging of vaccinia, the model poxvirus, to investigate mechanisms of adaptive evolution. We placed selective pressure on the viral K3L protein, a weak but specific inhibitor of human Protein kinase R (PKR), a key host immunity factor. Remarkably, vaccinia evolved significantly higher fitness after fewer than ten passages in human cells. Deep sequencing of replicate virus populations suggested the basis of adaptation was extensive and recurrent copy number variation (CNV) precisely at the K3L locus. Some viral genomes in adapted populations harbored greater than 10-fold expansions at K3L, reflecting 7-9% increases in genome size. Further analysis, including siRNA knockdown of K3L, showed that CNV was both necessary and sufficient for higher viral fitness in human cells. Moreover, K3L expansion preceded, and seemed to facilitate, the subsequent appearance of an adaptive amino acid substitution in K3L, which also defeats human PKR. A genomic accordion model emerges for adaptive evolution of large DNA viruses where CNV transiently provides increased expression of specific viral factors while concurrently providing expanded substrates for sampling mutations. Further analysis revealed rare gene duplications at many genomic locations in vaccinia populations suggesting that CNV is a common mechanism of adaptation in this class of viruses. As hosts and viruses compete in seemingly endless evolutionary arms races, this newly described mechanism of vaccinia evolution reveals a potent weapon in the arsenal of large viruses.

O-18

CICHLIDomics: uncovering the genetic and developmental basis of diversification in East African cichlid fishes

Walter Salzburger
University of Basel, Basel, Switzerland

The question of how variation in the DNA translates into organismal diversity has puzzled biologists for decades. Despite recent advances in evolutionary and developmental biology, the mechanisms that underlie adaptation, diversification and evolutionary innovation remain largely unknown. The exceptionally diverse species flocks of cichlid fishes in the East African Great Lakes are textbook examples of adaptive radiation and explosive speciation and emerge as powerful model systems to study the molecular basis of animal diversification. East Africa's hundreds of endemic cichlid species are akin a natural mutagenesis screen and differ greatly in ecologically relevant and, hence, naturally selected characters such as mouth morphology, but also in sexually selected traits such as coloration. One of the most fascinating aspects of cichlid evolution is the frequent occurrence of evolutionary parallelisms, which has led to the question whether selection alone is sufficient to produce these parallel morphologies, or whether a developmental or genetic bias has influenced the direction of diversification. The availability of five cichlid genomes now greatly enhances the quest for the genetic and developmental basis of diversification and convergent evolution in cichlid fishes.

O-19

Evolution of the endometrial transcriptome during the origin of pregnancy in mammals

Vincent Lynch, Gunter Wagner
Yale University, New Have, CT, USA

How morphological innovations originate is an enduring question in biology. While it's clear that gene regulatory evolution is ultimately responsible for phenotypic differences between species, the molecular mechanisms that underlie the evolution of innovations and major morphological transitions are obscure. Here we use mRNA-Seq to compare gene expression in the pregnant uterine endometrium across the major mammalian lineages, including eight placental mammals, two marsupials, and the egg-laying platypus as well as a chicken. We show that thousands of genes were recruited into uterine expression during the evolution of pregnancy, including cell surface proteins important for implantation of the mammalian embryo (LAMA1/LAMB1/LAMC1), essential signaling pathways that mediate maternal-fetal communication (cAMP, IL11, LIF), and gene regulatory networks that regulate hormone responses to pregnancy (HAND2, AHR). Using experimental methods we show that the transcription factors HAND2 and AHR were recruited into uterine expression through the domestication of placental-mammalian specific transposons into repressor and insulator elements, respectively. Finally we use ChIP-Seq data to demonstrate that >13,300 H3K4me3 marked regulatory elements in human uterine cells, ~59% of all H3K4me3 peaks in endometrial cells, are derived from transposable elements indicating that the regulatory landscape of the mammalian genome has been dramatically reorganized by transposons.

O-20

The Homology of Feathers and Scales: Using New High-throughput Methods to Address a Classic Question

Jacob Musser[1,2], Gunter Wagner[1,3], Richard Prum[1,2]
[1]Dept. of Ecology & Evolutionary Biology, Yale University, New Haven, CT, USA, [2]Peabody Museum of Natural History, Yale University, New Haven, CT, USA, [3]Yale Systems Biology Institute, New Haven, CT, USA

Feathers are an important anatomical innovation that evolved in the ancestors of birds and facilitated the evolution of flight, greater thermoregulation, and other facets of modern avian life. However, the molecular basis for the evolution of feathers is poorly understood, and the homology of feathers to other skin derivatives, especially scales, remains contentious. Here, we take a new approach to answering these questions by comparing transcriptomes from different stages of developing feathers, different avian and reptilian scales, and claws. We performed mRNA-seq on embryonic epidermal samples of these different developing skin appendages collected from two distantly related birds, Chicken (*Gallus gallus*) and Emu (*Dromaius novaehollandiae*), and developing scale epidermis from American Alligator (*Alligator missippiensis*), a member of the extant clade most closely related to birds. Comparison of these transcriptomes allows us to address several important questions, including the homology of feathers and scales at both a very early stage, when they share developmental similarity, and at a more advanced stage when feathers are developmentally unique. Further, these comparisons allow for the identification of feather regulatory molecules, including transcription factors and signaling pathways, that underlie feather development and evolution. Finally, to complement our comparative transcriptomics approach, we used immunohistochemistry to compare patterns of expression and subcellular localization of the earliest known feather regulatory molecule, the transcription cofactor β-catenin. Nuclear β-catenin has previously been shown to be necessary for feather development. Our preliminary evidence suggests nuclear β-catenin is also present in early developing avian scales and alligator scales, suggesting these skin appendages use similar molecular pathways at the beginning of their development.

O-21

Emergence of new *cis*-regulatory modules in fish and their role in the evolution of innovation after genome duplication

Ingo Braasch[1], Jason Sydes[1,2], Angel Amores[1], John H. Postlethwait[1]
[1]Institute of Neuroscience, University of Oregon, Eugene, Oregon, USA, [2]Computer and Information Science Department, University of Oregon, Eugene, Oregon, USA

More than 40 years since the publication of Ohno's famous book 'Evolution by gene duplication', the role of *cis*-regulatory innovation for the functional evolution of gene duplicates and morphological evolution in vertebrates remains elusive. We are investigating the emergence of *cis*-regulatory novelties and their contribution to phenotypic innovation after a teleost fish-specific genome duplication (TGD). Teleosts are the most species-rich group of vertebrates and offer unique advantages for the study of the genomic basis of gene regulation, morphological evolution and biodiversity at the functional level.

Among basal rayfin fish that diverged just prior to the TGD, the spotted gar *Lepisosteus oculatus* is the species that can most readily be fertilized and grown in the lab, providing embryos and adults amenable to gene expression studies. Thus, the gar is an ideal genomic and Evo-Devo outgroup model for the investigation of the mechanisms of gene function evolution after the TGD.

We developed a bioinformatic pipeline to survey vertebrate genomes for conserved non-coding elements (CNEs) that are conserved specifically in teleost genomes but are absent in tetrapods and other vertebrates. Adding next generation genome sequence data from the spotted gar as well as from the goldeye *Hiodon alosoides*, a basal teleost species that diverged from other teleost lineages shortly after the TGD, enables us to pinpoint the emergence of CNEs with respect to the TGD.

We show here that the majority of CNEs present in teleost genomes arose after teleosts diverged from gar. Furthermore, the gar genome is a 'missing link' that is particularly powerful to detect otherwise unrecognized homologies between teleost and tetrapod CNEs. Finally, the gar genome serves as an unduplicated outgroup to study teleost CNEs in paralogous genomic regions from the earlier vertebrate genome duplications that were lost specifically in the lobefin or tetrapod lineage (so-called ohnologs gone missing).

We are functionally testing CNEs with reporter constructs in zebrafish embryos. Comparing the expression of the associated genes in teleosts (e.g. zebrafish, stickleback, medaka) and the non-teleost gar shows whether the teleost-specific elements drive expression in teleost-specific domains. We hypothesize that such elements are involved in neofunctionalization after the TGD and contributed to the evolution of morphological key innovations of the teleost lineage such as their truly symmetric tail fin.

Our project is significant for understanding the role of *cis*-acting elements for sub- and neofunctionalization, evolutionary innovations, and the principles that govern gene function evolution after genome duplication in general.

O-22

The evolution of gene expression underlying sexual development in fungi

Nina Lehr[1], Usha Sikhakholli[2], Zheng Wang[1], Francesc Lopez-Giraldez[1], Ning Li[1], Frances Trail[2], Jeffrey Townsend[1]
[1]Yale University, New Haven, CT, USA, [2]Michigan State University, East Lansing, MI, USA

Studies linking the evolution of gene expression and the evolution of development are challenging in complex organisms. Of animal, plant, and fungal multicellular development, the genomic basis of fungal development is arguably the most different and the least well understood, but also has tremendous potential for illumination of evolutionary principles. Fungi can provide ideal systems for these studies as they are often easily manipulated, develop fruiting structures with a few well-characterized tissue types, feature many available and complete genome sequences. Moreover, in many cases, multiple species can thrive in a single simple heterotrophic environment. We undertook to reveal elements of the underlying program of fruiting body development by transcriptomic sequencing of genome-wide gene expression in a set of five closely-related fungi with differing morphological development. These developmental processes are fundamental to sexual reproduction, recombination, and to the adaptive dynamics of pathogens and hosts. Because we maintained a strictly common medium across experiments, our transcriptomic data revealed solely evolved differences in the transcriptional basis of morphological changes. We assayed the transcriptome of two plant pathogenic *Fusarium* species and three species of *Neurospora* during fruit body development. We estimated the ancestral evolutionary transitions that resulted in the shifts in morphology of these two genera, knocking out genes with evolved function and assaying phenotype. Our transcriptional studies and candidate gene knockouts reveal regulatory circuits that control development and clearly identify many relevant genes without prior annotation. These results provide a model for how transcriptional shifts drive the evolution of morphological variation.

O-23

Modeling the contributions of GC-biased gene conversion and selection to fast-evolving regions of primate genomes

John A. Capra[1], Melissa J. Hubisz[2], Dennis Kostka[1], Katherine S. Pollard[1], Adam Siepel[1]
[1]Gladstone Institutes, University of California, San Francisco, San Francisco, CA, USA,[2]Cornell University, Ithaca, NY, USA

Patterns of DNA sequence variation within and between species provide evidence about the evolutionary processes that drive speciation and the evolution of lineage-specific traits. However, it can be difficult to disentangle the action of adaptive and non-adaptive evolution, because non-adaptive processes can produce patterns of variation that resemble those produced by selection. For example, GC-biased gene conversion (gBGC) is a recombination associated bias that favors the fixation of G and C alleles near the double strand breaks that induce recombination. The action of gBGC on a locus can potentially produce an excess of substitutions, and can yield false positives in tests for selection. Additionally, in contrast to selection, gBGC is neutral to the fitness of alleles, and simulation studies have shown that gBGC can lead the the fixation of mildly deleterious alleles. Based on evidence in many species, we showed that gBGC's impact on sequence evolution may be widespread in eukaryotes. Thus, knowledge of regions influenced by gBGC is crucial to interpreting the causes and effects of variation between genomes.

To study the prevalence of gBGC in humans and investigate the interaction of adaptive and non-adaptive processes, we generated the first model-based, unbiased, genome-wide predictions of regions likely to have recently experienced gBGC (gBGC tracts). We designed a phylogenetic hidden Markov model with both selection and gBGC states. We fit this model to multiple sequence alignments of human, chimpanzee, orangutan, and rhesus macaque and used the results to segment the human and chimp genomes into tracts with or without evidence of gBGC. The predicted gBGC tracts are fast-evolving, have high recombination rate, and show strong evidence of historical and on-going GC-biased evolution. We used the tracts to investigate the the functional effects of gBGC and its interaction with selection. We found that the tracts significantly overlap human accelerated regions (HARs) and positively selected genes (PSGs). These findings agree with previous results and suggest that gBGC can confound classic tests for adaptive evolution. By analyzing patterns of recent polymorphism in the tracts, we found that regions recently affected by gBGC harbor significantly more disease-causing mutations than similar regions without recent gBGC. Overall, our models and analysis enable simple tests for the possible influence of non-adaptive evolution via gBGC on a locus and provide a resource for investigating the influence, extent, and interplay of adaptive versus non-adaptive evolution in human and primate genomes.

O-24

Evolutionary consequences of variation in the processing rate of the *Arabidopsis thaliana* microRNA824

Jinyong Hu[1], Filippos Klironomos[2], Johannes Berg[2], Juliette de Meaux[3]
[1]MPIZ, Cologne, Germany, [2]Institute Theoretical Physics, Cologne, Germany, [3]IEB, Münster, Germany

Non-coding RNAs are post-transcriptional regulators of gene expression enjoying unique evolutionary attributes because of their structural properties. Yet, evidence for their involvement in adaptive evolution is lacking. Here, we analyze the effect of variation at *ath-miR824* on the expression level of its target, AGL16, a MADS-box transcription factor controlling stomata patterning in *Arabidopsis thaliana*. We find that a structural polymorphism in the miRNA precursor molecule affects the synthesis of mature miRNA. Using mathematical modeling of the regulatory network, in which the miRNA is embedded, we show how changing the miRNA processing rate alters the level of target protein. With a combination of mutants, near-isogenic and transgenic lines, we demonstrate that *miR824* polymorphism affects stomata patterning, and we investigate the fluctuations of selection pressure on the *miR824* alleles in a field trial over three consecutive generations. Our work showcases the unique adaptive properties of non-coding RNAs for the adjustment of gene expression at the post-transcriptional level and illustrates how these properties come into play in a natural population.

O-25

The role of selection, biased gene conversion and mutation bias in the evolution of genomic GC content in bacteria

Adam Eyre-Walker[1], Falk Hildebrand[1,2]
[1]*University of Sussex, Brighton, UK,* [2]*Vrije Universitet Brussel, Brussels, Belgium*

Genomic GC contents vary dramatically between species from less than 20% to more than 70%. The reasons for this variation have remained unclear since it was first described more than 50 years ago. Several recent analyses have suggested that mutation bias alone cannot explain the variation and that either selection or biased gene conversion (BGC) must elevate genomic GC content in GC rich bacteria. In my talk I will present evidence that a simple selection or BGC model does not fit patterns of synonymous polymorphism unless we assume that the pattern of mutation also varies between bacterial species, such that GC rich species have a relatively GC biased mutation pattern. Thus variation in genomic GC content appears to be a consequence of both mutation bias and selection. The model of best fit suggests the pattern of mutation in the most GC rich species are likely to be GC biased, thus contradicting the recent hypothesis that mutation patterns are universally AT biased.

O-26

Invariants of genome evolution: selection, stochasticity or both?

Eugene Koonin
National Center for Biotechnology Information, NIH, Maryland, USA

Research in quantitative evolutionary genomics and systems biology led to the discovery of several universal regularities connecting genomic and molecular phenomic variables. These universals include the log-normal distribution of the evolutionary rates of orthologous genes; the U-shaped distribution of the size of orthologous gene sets; the power law-like distributions of paralogous family size and node degree in various biological networks; the negative correlation between a gene's sequence evolution rate and expression level; and differential scaling of functional classes of genes with genome size. These universals of genome evolution can be accounted for by simple mathematical models similar to those used in statistical physics, such as the birth-death-innovation model. Some of these models do not explicitly incorporate selection, therefore the observed universal regularities do not appear to be shaped by selection but rather are emergent properties of gene ensembles. In those cases where selection is involved, it appears to be directed primarily at prevention of malfunction rather than adaptive gain of function. Although a complete physical theory of evolutionary biology is inconceivable, the universals of genome evolution might qualify as 'laws of evolutionary genomics' in the same sense 'law' is understood in modern physics.

References

Koonin EV. Are there laws of genome evolution? PLoS Comput Biol. 2011Aug;7(8):e1002173

Koonin EV (2011) The Logic of Chance: The nature and origin of biological evolution. Upper

Saddle River (NJ): FT Press

O-27

Variation in exonic splice related constraints across taxa: why the impact of selection can be more evident when the effective population size is low

Laurence Hurst, Joanna Parmley, Toby Warnecke, Andreas Schueler
University of Bath, Bath, Somerset, UK

Classical theory holds that as the effective population size decreases so the relative importance of selection over drift also diminishes. This, for example, is evoked to explain why, comparing between species, intron sizes tend to increase as the effective population size goes down, mammals for example, having very large introns on the average. In this talk I show that such genome degradation has, perhaps paradoxically, resulted in more, not less evidence for some forms of selection in exons in species with low Ne. I hypothesize that a) large introns are harder to splice using information held solely in introns (as yeast does) and b) to compensate, species with large introns tend to move more information to help identify splice borders into the ends of exons in the form of exonic splice enhancer (ESEs) motifs. I show that, as predicted, within a species (humans) it is indeed the case that exons flanked by larger introns have a higher density of ESEs. These enhancers have highly skewed nucleotide content and thus in turn affect in predictable ways codon usage bias and amino acid choice. Rates of evolution at synonymous sites, non-synonymous sites and in non-coding (but spliced) RNAs are all lower near exon ends and are lower in ESE motifs than in non-ESE sequence. Using the extent of the skew in amino acid content as a proxy for the impact of ESEs on sequence evolution, I show that between species there is a correlation between the proportion of gene sequence that is intronic and the extent of the skewed amino acid usage, humans having both the largest introns and the most skewed amino acid usage at exon ends, while yeasts have no skew and have little intronic DNA. This can equally well be plotted as more amino acid skews when Ne.u is low. Selection within exons for maintaining accurate splicing is thus most acute when Ne is low as this is when intronic information is inadequate to accurately specify exon-intron junctions. In considering the effects of low Ne on selection it is thus important to consider the knock-on consequences of genomic deterioration. Assertions that a logical corollary of low Ne is an absence of selection on synonymous sites are spurious.

O-28

Sequencing of a Darwin's finch genome: evidence against the selective thermal stability hypothesis, and evidence for a neutral explanation for the origin of isochores

Chris Rands[1], Matthew Fujita[2,1], Lesheng Kong[1], Aaron Darling[3], Céline Clabaut[2], Richard Emes[4], Andreas Heger[1], Tim Harkins[5], Clotilde Teiling[5], Matthew Webster[6], Stephen Meader[1], Peter Grant[7], Rosemary Grant[7], Jonathan Eisen[3], Arkhat Abzhanov[2], Chris Ponting[1]
[1]University of Oxford, Oxford, UK, [2]Harvard University, Cambridge, Massachusetts, USA, [3]University of California, Davis, Davis, California, USA, [4]University of Nottingham, Nottingham, UK, [5]Roche Applied Science, Indianapolis, Indiana, USA, [6]Uppsala University, Uppsala, Sweden, [7]Princeton University, Princeton, New Jersey, USA

The Galápagos' Darwin's finches, with their widely different beak sizes and morphologies, are a text-book example of how species radiate and adapt to novel environments. We sought to investigate the impact of selection and mutation on the genome of a large ground finch (*Geospiza magnirostris*), the first whole genome sequence from a vertebrate archipelago species. From 454 sequence reads assembled over 12,957 scaffolds we predicted over 13,000 protein coding genes in the Darwin's finch genome, and defined a conservative set of simple 1:1 orthologs that have been retained without duplication or loss across seven diverse amniotes. We conservatively filtered protein coding sequence alignments from this ortholog set to remove poorly aligning sequence. Our analysis suggests that less stringent filtering schemes, such as those used in many previous studies, will lead to spurious inference of elevated substitution rates. Comparing the lineage-specific non-synonymous (dN) and synonymous (dS) substitution rates in the filtered alignments, we find a similar dN/dS ratio along the Darwin's finch and zebra finch branches. This implies that similar patterns of natural selection operated across Darwin's and zebra finch protein coding regions, and therefore that the long-term effective population size of the Darwin's finch was as high as that of zebra finch. Unexpectedly, GC-content in the Darwin's finch genome was found to be considerably more homogeneous than in the genomes of zebra finch or other sequenced birds or mammals, making Darwin's finch the only sequenced endotherm to lack prominent isochores. This provides evidence against the thermal stability hypothesis, which proposes that GC-rich isochores in genomes of endothermic vertebrates have arisen by selection to engender thermal stabilisation of active, open chromatin, regions. Examining patterns of sequence constraint, we estimate that ~15% genomic sequence is constrained with respect to insertions and deletions between Darwin's and zebra finch. This suggests that there is a large functional non-coding component of the genome, but nonetheless implies that the vast majority of genomic sequence is neutrally evolving. This is inconsistent with a selective hypothesis for isochore origin, which requires selection to be operating across a large proportion of the genome. Our results shed new light onto the origin of isochores, and underline the importance of whole genome sequences for investigating fundamental biology that cannot currently be predicted from study at the whole organism level.

O-29

Tracking the molecular mechanisms responsible for GC-biased gene conversion in yeast

Yann Lesecque, Laurent Duret
Laboratoire de Biométrie et Biologie Evolutive, Lyon, France

Recently, a new evolutionary force acting on genomes has been identified in addition to mutation, selection and drift. This process, called GC-Biased Gene Conversion (gBGC), is associated with recombination and is known to be acting in yeast and mammalian genomes. In those genomes, gBGC favours the spreading of GC-rich alleles within populations, regardless of their fitness effects. This force is even able to counteract selection by allowing the fixation of deleterious alleles. However, despite its major impact on genome evolution, the molecular causes and mechanisms responsible for gBGC are still unknown.

In theory, two steps of the recombination process might be responsible for a bias in the process of gene conversion: (i) formation of double strand breaks (DSBs); (ii) repair of mismatches occurring in heteroduplex DNA. Interestingly, in mammals, the repair of G:T mismatches by the Base Excision Repair (BER) is intrinsically biased towards G:C, which led to the proposal that this process might be responsible for gBGC. To investigate these different hypotheses, we analysed the high-resolution yeast recombination data of Mancera et al. (2008)[*]. In agreement with previous findings, we observed a significant over-transmission of GC-alleles at GC/AT heterozygote sites. We show that this bias is mainly associated with simple conversion tracts, i.e. tracts involving only one donor genotype. This is not consistent with the hypothesis of gBGC being caused by BER, because this repair system cannot act at the whole tract scale. Interestingly, we observed that gBGC is specifically associated with crossover, and not with non-crossover recombination events. Moreover, this bias seems to be driven by an over-transmission of haplotypes flanked by GC-rich heterozygous sites. These two last observations are not consistent with the hypothesis of gBGC being caused by a bias in the formation of double strand breaks (DSBs). Finally, we propose that gBGC is driven by a bias of the Mismatch Repair system (MMR), after the introduction of nicks in the homologue, specifically during the formation of crossovers.

* Mancera, E., R. Bourgon, A. Brozzi, W. Huber, and L.M. Steinmetz (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, **454**: p. 479-85.

O-30

The evolution of alternative splicing patterns and splicing regulatory mechanisms in vertebrates.

Adem Bilican[1,2], Anamaria Necsulea[1,2], Henrik Kaessmann[1,2]
[1]University of Lausanne, Lausanne, Switzerland, [2]Swiss Institue of Bioinformatics, Lausanne, Switzerland

Through alternative splicing, a single gene can produce many different isoforms, thus multiplying its potential functional roles. Alternative splicing is known to affect the great majority of protein-coding genes in many species, but it is unclear what proportions of alternative isoforms are functional. Moreover, the evolutionary forces that act on alternative splicing patterns are unknown, and it is unclear to what extent evolutionary changes in alternative splicing patterns may lead to lineage-specific phenotypic innovations. Here, we address these questions by studying the evolution of alternative splicing patterns in conjunction with the evolution of splicing regulatory mechanisms.
To study the evolution of transcriptomes, we generated an extensive RNA-Seq dataset, comprising 11 species (great apes, macaque, mouse, opossum, platypus, chicken and xenopus) and 8 tissues (cerebellum, cortex, heart, kidney, liver, ovary, testis and placenta). Using these data, we defined for each species a complete catalogue of cis-acting regulatory elements of splicing, i.e. splicing enhancers or silencers that are located in intronic or exonic sequences flanking the splice site.
We defined these regulatory elements using a previously proposed in silico method, which relies on a principle of compensation between splice site strength and presence of additional splicing regulators (e.g., constitutive exons with weak splice sites are expected to have more enhancers for compensation). We found that the presence of in silico detected splicing enhancers is correlated with high levels of exon inclusion frequency, thus confirming the validity of our approach.
Our analyses revealed that the sets of splicing regulators are highly conserved between species (for example, 81% of human exonic splicing enhancer motifs are also predicted to be enhancers in the macaque). Furthermore, the individual occurrences of splicing regulatory motifs also display high levels of sequence conservation, which confirms their functional significance. However, we also detected many species-specific regulatory motifs, which may explain the rapid evolution of alternative splicing patterns.
Finally, we searched for substitutions that disrupt or create splicing regulator motifs in splice site flanking regions, and we found that the presence of such substitutions correlates with important changes in the pattern of alternative splicing.

O-31

Non-random mate choice in humans: insights from a genome scan

Romain Laurent, Bruno Toupance, Raphaëlle Chaix
Eco-Anthropologie et Ethnobiologie, UMR 7206 CNRS, MNHN, Univ Paris Diderot, Sorbonne Paris Cité, Paris, France

Little is known about the genetic factors influencing mate choice in humans. Still, there is evidence for non-random mate choice with respect to physical traits. In addition, some studies suggest that the Major Histocompatibility Complex may affect pair formation. Nowadays, the availability of high density genomic data sets gives the opportunity to scan the genome for signatures of non-random mate choice without prior assumptions on which genes may be involved, while taking into account socio-demographic factors. Here, we performed a genome scan to detect extreme patterns of similarity or dissimilarity among spouses throughout the genome in three populations of African, European American, and Mexican origins from the HapMap 3 database. Our analyses identified genes and biological functions that may affect pair formation in humans, including genes involved in skin appearance, morphogenesis, immunity and behaviour. We found little overlap between the three populations, suggesting that the biological functions potentially influencing mate choice are population specific, in other words are culturally driven. Moreover, whenever the same functional category of genes showed a significant signal in two populations, different genes were actually involved, which suggests the possibility of evolutionary convergences.

O-32

Recent ecological speciation and local adaptation along an altitudinal gradient on Mount Etna.

Owen Osborne[1], Graham Muir[1,3], Jonas Sarasa-Marcuello[1], Simon Hiscock[2], Dmitry Filatov[1]
[1]University of Oxford, Oxford, UK, [2]University of Bristol, Bristol, UK, [3]Heidelberg University, Heidelberg, Germany

To what extent does natural selection contribute to the origin of new species? It is now accepted that speciation can occur via divergent selection imposed by contrasting environments. Identifying cases of such ecological speciation and, crucially, identifying their underlying genetic mechanisms are now key goals of evolutionary research. On the slopes of Mount Etna, Sicily, two species of ragwort (*Senecio*) occur. *Senecio aethnensis* grows at high altitude, while *S. chrysanthemifolius* inhabits lower elevations, and the species display substantial phenotypic divergence. At intermediate altitudes they form a hybrid zone with altitude-associated gradients in both genetic and phenotypic characters. This cline could be the result of ecological speciation. Alternatively, it could simply have arisen by recent admixture of species which diverged in allopatry. To differentiate between these possibilities, we performed a variety of population genetic analyses on DNA sequence and microsatellite polymorphism datasets. MCMC-based analyses in an isolation with migration framework suggested that the species actually diverged extremely recently (~45,000 ya) and that gene flow has been on-going since the split, supporting ecological speciation as the mechanism of their divergence. Reflecting their recent speciation, the species are extremely close genetically and we detected no fixed genetic differences. This raises the question of how to account for their substantial phenotypic divergence. We found that genetic differentiation was significantly higher in differentially expressed genes compared to those which were expression neutral between the species, suggesting that diversifying selection is acting at these loci to suppress the homogenising effect of gene flow. Taken together, our results suggest that the species are the result of ecological speciation and this has at least partially been driven and maintained by diversifying selection on differentially expressed genes.

O-33

Genome-scale simultaneous inference of species and gene trees

Bastien Boussau[1,2], Gergely Szollosi[1], Laurent Duret[1], Manolo Gouy[1], Eric Tannier[3], Vincent Daubin[1]
[1]LBBE, CNRS, Université Lyon 1, Lyon, France,[2]UC Berkeley, Berkeley, CA, USA,[3]LBBE, INRIA, Université de Lyon, Lyon, France

Comparisons of gene trees and species trees are key to the understanding of major processes of genome evolution such as gene duplication and gene loss. Because current methods fail to model the two-way dependency between gene and species histories, these mechanisms can be misrepresented in genome studies. We present a new probabilistic method to jointly infer species and gene trees for dozens of genomes and thousands of gene families. In our model, each branch of the species tree is associated to a particular birth-death model with specific duplication and loss parameters, thus accommodating heterogeneity in the processes of genomic evolution. Trees and parameters are estimated in the maximum likelihood framework, using hundreds of processors in parallel. We use realistic simulations of gene family evolution to show that our approach accurately recovers the species tree and the numbers of gene duplications and losses. In addition we find that our approach outperforms state-of-the-art algorithms to reconstruct gene trees. We confirm the quality of our inferences on thousands of gene families in 36 mammalian species. Importantly, several measures of ancestral genome reconstruction quality show that our approach outperforms phylogenetic algorithms routinely used to create trees for databases of homologous gene families. As a consequence, we believe our method could be advantageously used in databases of homologous gene families, for comparative studies of genome dynamics, but also for detecting selection on gene sequences and gene numbers.

O-34

Age and structural characteristics as determinants of protein evolutionary rate

Macarena Toll-Riera[1], David Bostick[2], M.Mar Albà[1,3], Joshua B. Plotkin[2]
[1]Evolutionary Genomics Group, Fundació Institut Municipal d'Investigació Mèdica (FIMIM)- Universitat Pompeu Fabra (UPF), Barcelona, Spain, [2]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA, [3]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

Which are the determinants of protein evolution is a long-standing question which is still under debate. To date several determinants have been proposed such as mRNA expression level, protein dispensability, number of protein-protein interactions, protein age, structural characteristics and folding robustness. However, many of these factors are not independent. For example younger proteins have been reported to evolve faster (Albà & Castresana, 2005) but they also have less compact structures and thus a higher number of exposed residues (Choi & Kim, 2006; Toll-Riera, Radó-Trilla, Martys, & Albà, 2011). Can we separate the influence of protein structure to that of protein age? In order to do that we have compiled one-to-one human and mouse orthologous proteins from Ensembl with mapped PDB structures: 1,899 proteins and 2,145 structures. We have assigned an age (Eukarya, Metazoan, Vertebrates) to each PDB entry and calculated several structural properties: solvent accessibility, designabilty (roughly the number of sequences that can fold into a given structure), stability, and secondary structure. We have found that the reported effects of these properties on evolutionary rates are maintained within age groups but not between age groups. For example, the previously found positive correlation between protein designability and evolutionary rate only holds within age groups. This is because younger structures, despite being less designable than older ones, are still evolving faster. Similarly, we have found that in younger proteins a higher fraction of the substitutions are potentially destabilizing but, nevertheless, these proteins are accumulating many more amino acid substitutions. The effect of age can also be observed when we group the residues according to secondary structure or solvent accessibility: differences between proteins of different age remain highly significant within each structural group. Our results show that both historical factors (protein age) and intrinsic properties of present-day proteins (protein structure) contribute to the differences in protein evolutionary rates.

Albà, M. M., & Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular biology and evolution*, *22*(3), 598-606.

Choi, I.-G., & Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(38), 14056-61.

Toll-Riera, M., Radó-Trilla, N., Martys, F., & Albà, M. M. (2011). Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. *Molecular biology and evolution*.

O-35

Population genetic properties of time serial data with examples from ancient population-genomic data

Mattias Jakobsson
Uppsala University, Uppsala, Sweden

Extracting genetic information from ancient material has for long been hampered by numerous difficulties since its first steps some two decades ago, but in the last few years, many of these problems have been solved and the use of ancient DNA (aDNA) is now beginning to show its full potential. We will likely see a wealth of genomic data from ancient populations, but the statistical properties of time-structured genetic samples are considerably less explored than population genetic patterns arising from spatial structure. Using simulations, we explore and highlight features of temporal structure and spatial structure, such as an 'isolation-by-time' effect that is similar to isolation-by-distance. Using model- and simulation-based approaches, we can now make novel inferences about demographic and evolutionary questions from time serial data. We will discuss examples from the long standing debate about the introduction of farming in Europe and question about archaic ancestry in East Asia using paleogenomic data.

O-36

A multi-locus perspective of the Late Pleistocene demography of bison and horses in the Klondike, Yukon Territory, Canada

Beth Shapiro[1,3], Mathias Stiller[1,3], Matthew Wooller[5], Robert Wayne[4], Duane Froese[2]
[1]Penn State University, University Park, PA, USA, [2]University of Alberta, Edmonton, Alberta, Canada, [3]University of California Santa Cruz, Santa Cruz, CA, USA, [4]University of California Los Angeles, Los Angeles, CA, USA, [5]University of Alaska Fairbanks, Fairbanks, Alaska, USA

Previous work to infer the evolutionary dynamics of bison and horses in Eastern Beringina has focused exclusively on mitochondrial DNA. Here, we present a preliminary analysis of a large, multi-locus data set isolated from ca. 100 horses and bison recovered from permafrost deposits near Dawson City, Yukon Territory, Canada. Bones from which DNA was extracted and sequenced range in age from the early Holocene to around 80,000 years old. The results reveal a much more complex demographic history over this time period for these two species than was possible to estimate from mitochondrial data alone. We correlate changes observed in population size and structure with changes in climate, including in the distribution of habitat, throughout Marine Isotope Stages 4-1.

O-37

Inferences on dog domestication - genetic analysis of the most ancient dogs utilizing DNA capture arrays

Olaf Thalmann[1,2], Daniel Greenfield[2], Matthias Meyer[3], Susanna Sawyer[3], Pin Cui[3], Mietje Germonpré[4], Mikhail V. Sablin[5], Francesc López-Giraldez[9], Daniel LePont[1], Brian Worthington[10], Jeff P. Blick[6], Jeniffer A. Leonard[7], Richard E. Green[8], Robert K. Wayne[2]
[1]University of Turku, Turku, Finland, [2]University of California, Los Angeles, USA, [3]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [4]Royal Belgian Institute of Natural Sciences, Brussels, Belgium, [5]Zoological Institute RAS, Saint-Petersburg, Russia, [6]Georgia College & State University, Milledgeville, USA, [7]Estación Biológica de Doñana- CSIC, Seville, Spain, [8]University of California, Santa Cruz, USA, [9]Yale University, New Haven, USA, [10]Southeastern Archaeological Research, Inc., Newberry, USA

The geographical and temporal origin of the dog is controversial. Genetic data suggest a domestication event in Asia or the Middle East about 15,000 - 30,000 years ago, whereas the oldest dog-like fossils are found in Europe dating to over 30 thousand years ago. We genetically analyzed the remains of 14 prehistoric wolves and dogs including some of the oldest dog remains described from the New and Old World. Utilizing array based DNA capture techniques coupled with Illumina double indexed sequencing, we targeted a total of ~750,000 nucleotides in each of the ancient canids and additional 20 contemporary wolves from North America and Eurasia. The sequence information comprised the complete mitochondrial genome, 3,000 SNPs previously identified as highly informative for differentiating dogs from wolves, exonic sequences from 62 potential domestication genes and ~150,000 nucleotides of non-coding regions spread throughout the genome.
Initial analyses reveal that we have successfully captured and sequenced the complete mitochondrial genome with high coverage as wells as a substantial number of autosomal fragments from ten prehistoric canids and all contemporary wolves. Phylogenetic analysis combining the complete mitochondrial genomes of the prehistoric canids with those of a large collection of modern dogs and wolves result in a statistically well supported tree. While some haplotypes cluster within modern dogs or wolves, others show a basal placement in the phylogeny. The latter finding might support a previous notion that an aberrant lineage of dog-like canids might have existed throughout the northern hemisphere during the late Pleistocene and became globally extinct during the last 20,000 years. We will test this hypothesis by investigating the autosomal loci and employ sophisticated phylogenetic analyses, demographic modeling and selection scans to better understand the influence of early human society and artificial selection on the canine genome.

O-38

Ancient DNA from museum skins reveals the progression retroviral endogenization

María Ávila-Arcos[2], Simon Ho[3], Yasuko Ishida[5], Nikolas Nikolaidis[4], Kyriakos Tsangaras[1], Karin Hönig[1], Rebeca Medina[4], Morten Rasmussen[2], Sarah Fordyce[2], Sébastien Calvignac-Spencer[6], Eske Willserslev[2], M. Thomas Gilbert[2], Kristofer Helgen[7], Alfred Roca[5], Alex Greenwood[1]
[1]Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany, [2]Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark, [3]University of Sydney, New South Wales, Australia, [4]California State University, Fullerton, California, USA, [5]University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, [6]Robert-Koch-Institut, Berlin, Germany, [7]Smithsonian Institution, Wasshington D.C., USA

Retrovirus-like elements constitute close to 8% of the human genome and may play a role in health and disease. However, most endogenous retroviruses are millions of years old, making it difficult to study the process by which they invaded and shaped the genomes of mammals. The koala retrovirus (KoRV) is the only known infectious retrovirus currently in the midst of invading the germ line of its host. The process is thought to have begun within the last 200 years. KoRV is linked to leukemias and immune suppression. Understanding the evolutionary events affecting the genetic diversity of KoRV, and how the virus changed as it invaded the host germ line, provides an opportunity for research into how endogenous retroviruses evolve and affect the genomes of all vertebrates, including humans. We analysed museum koala samples to examine the process of endogenization in real time. Using ancient DNA methods and GS FLX high-throughput sequencing, we determined the sequence of multiple full-length retroviral envelope genes, including important viral functional motifs. Our results suggest that KoRV endogenization is not a recent event. The data are consistent with a model whereby the endogenization of retroviruses proceeds over long periods of time during which proviral sequences remain stable, host species may experience a prolonged stage of reduced fitness, and the virus remains infectious.

O-39

Admixture inference using a sequentially Markov coalescent

Joshua Schraiber, Eric Durand, Montgomery Slatkin
UC Berkeley, Berkeley, CA, USA

Recent advances in ancient DNA technology have made it possible to obtain genome-scale sequence from individuals of extinct lineages. Many analyses of ancient samples have revealed evidence of admixture between the ancient lineages and modern lineages. In an effort to characterize these admixture events, we developed a hidden markov model that explicitly models the history of a single chromosome sampled from a modern population and a single chromosome sampled from an extinct lineage. By assuming that recombination acts as a Markov process along the genome, we are able to infer not only the parameters of the admixture, including the timing and admixture proportion, but also identify admixture tracts with high confidence. Applications to both synthetic and real data will be presented.

O-40

A screen for selection in humans based on archaic genome sequences.

Daniel Falush, Michael Lachmann
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Studies of variation amongst modern humans have more power to detect recent selection that occurred during our expansion out of Africa than on variants that have already reached fixation. We present novel statistical genetic methodology to infer selection based on identifying genomic tracts where sequences from closely related outgroups fall outside variation within the population. The method is applied to recently sequenced archaic hominid sequences (Neanderthal and Denisovans) and to data from the 1000 Genomes Project. Our screen differs from previous scans for positive selection in humans in both the genomic regions found and the functions of genes that are most overrepresented in those regions. We also provide new insight into the role of background selection in shaping variation over the last 300,000 years.

O-41

The antigenic evolution of influenza viruses

Derek Smith
University of Cambridge, Cambridge, UK

Thirty plus years of global influenza virus surveillance, in multiple species, provides a remarkable dataset for the study of influenza virus evolution. Because the purpose of much of this surveillance is vaccine strain selection, these data have been analyzed antigenically as well as genetically. These studies of human influenza viruses will be contrasted with studies in other species, to show the importance of the coevolution of the virus and population-level immunity to the virus, and some of the governing mechanisms for the evolution of virus.

O-42

A baculovirus gene inducing hyperactive behavior in caterpillars

Stineke van Houte, Vera Ros, Monique van Oers
Laboratory of Virology, Wageningen University, Wageningen, The Netherlands

Many parasites manipulate host behavior to increase the probability of transmission. To date, however, knowledge on parasitic genes underlying such behavioral manipulations is scarce. Here we show that the baculovirus *Autographa californica* nuclear polyhedrovirus (AcMNPV) induces hyperactive behavior in *Spodoptera exigua* caterpillars at three days after infection. Furthermore, we identify the viral protein tyrosine phosphatase (*ptp*) gene as a key player in the induction of hyperactivity in larvae, and show that the phosphatase activity of the encoded enzyme is crucial for this behavioral change. Phylogenetic inference points at a lepidopteran origin of the viral *ptp* gene. Our study suggests that *ptp*-induced behavioral manipulation is an evolutionary conserved strategy of baculoviruses to enhance virus transmission. In addition, we show that larvae move downward after AcMNPV infection, which appears to be a *ptp*-independent process. Overall, these data provide a firm base for a deeper understanding of the mechanisms behind baculovirus-induced insect behavior.

O-43

Antigenic flux of the influenza virus population

Trevor Bedford[1], Marc Suchard[2], Philippe Lemey[3], Colin Russell[4], Derek Smith[4], Andrew Rambaut[1]
[1]University of Edinburgh, Edinburgh, UK, [2]University of California, Los Angeles, USA, [3]Katholieke Universiteit Leuven, Leuven, Belgium, [4]Cambridge University, Cambridge, UK

The influenza A virus infects approximately 500 million individuals each year. Owing to its RNA makeup, influenza mutates extremely rapidly allowing the virus population to escape the pull of the human immune system. A single individual may be infected year after year by antigenically novel strains. As result of this rate of mutation, the timescale of influenza evolution is a human timescale. We get the chance to observe the process of evolution in action. Here, we investigate antigenic and genetic evolution in influenza A (H3N2) from 1968 to 2012. Genetic evolution can be modeled through standard phylogenetic techniques, but antigenic evolution requires novel methods of analysis. Antigenic data exists as a series of pairwise measurements of cross-reactivity between one virus strain and ferret anti-sera harvested after infection by a second virus strain. Here, we use Bayesian multi-dimensional scaling (BMDS) to transform this series of pairwise measurements into a set of locations, one for each virus strain. The distance between strains in this antigenic space is inversely proportional to their degree of cross-reactivity. This transformation allows us to easily model antigenic evolution as diffusion across an antigenic landscape. We fit phylogenetic models that jointly describe sequence change and antigenic change to see what evolutionary histories result in persistent viral lineages and what histories lead to lineage extinction. Through this analysis, we retrospectively assess vaccine strain selection in the context of antigenic match between vaccine strain and the global influenza population, and in doing so, provide metrics for future vaccine choice.

O-44

Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints.

Nuno R Faria[1], Marc A Suchard[2,3], Daniel G Streicker[4], Philippe Lemey[1]
[1]Rega Institute for Medical Research, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium, [2]Department of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, USA, [3]Department of Biostatistics, UCLA School of Public Health, University of California, Los Angeles, USA, [4]Odum School of Ecology, University of Georgia, Athens, USA

Understanding the factors that determine the fate of zoonotic transmission is essential for the development of predictive models and for assessing the efficacy of prevention strategies. However, the determinants of cross-species transmission and its outcome in new host species are still unknown for the majority of important human pathogens. A recent analysis of rabies virus gene sequences in North American bat species demonstrated that host divergence posed a more important barrier than ecological variables for the emergence of a rapidly evolving virus in a new host. This analysis involved population genetic inference on subsets of the data, followed by rescaling and statistical analysis of the resulting mean estimates.

Here, we describe a flexible Bayesian statistical framework to reconstruct virus transmission between different host species while simultaneously testing and estimating the contribution of several potential predictors of cross-species transmission. For this purpose, we parameterize a stochastic discrete diffusion process as a generalized linear model and perform Bayesian model averaging over several potential predictors. Importantly, we extend this model to allow different host transition processes on external and internal branches in order to discriminate between recent cross-species transmission events, many of which are likely to result in dead-end infections, and host shifts, which reflect successful onwards transmission in the new host species.

Our integrated approach corroborates genetic distance between hosts as a key determinant of both host shifts and cross-species transmission of rabies virus in bats. In addition, we confirm that geographic range overlap is an additional barrier for cross-species transmission but not for historical host shifts. Although our framework focused on the multi-host reservoir dynamics of bat rabies virus, it is applicable to other pathogens and to other discrete state transition processes. We anticipate that this framework will be useful to understand the key drivers of cross species dynamics for a broad range of zoonotic pathogens.

O-45

Coevolution between C2H2 zinc fingers and retroelements in vertebrates

James Thomas, Sean Schneider
University of Washington, Seattle, WA, USA

Vertebrate genomes encode large and highly variable numbers of tandem C2H2 zinc finger (tandem ZF) transcription factor proteins. In mammals, most tandem ZF genes also encode a KRAB domain (KZNF proteins). Very little is known about what forces have driven the number and diversity of tandem ZF genes. Recent papers suggest that KZNF proteins can not only bind and repress transcription of exogenous retroviruses and their endogenous counterpart LTR retroelements, but also block their integration into the host genome. We report a striking correlation across vertebrate genomes between the number of LTR retroelements and the number of host tandem ZF genes. This correlation is specific to LTR retroelements and ZF genes and was not explained by covariation in other genomic features. We further show that recently active LTR retroelements are correlated with recent tandem ZF gene duplicates across vertebrates. On branches of the primate phylogeny, we find that the appearance of new families of endogenous retroviruses is strongly predictive of the appearance of new duplicate KZNF genes. We hypothesize that retroviral and LTR retroelement burden drives evolution of host tandem ZF genes. This hypothesis is consistent with previously described molecular evolutionary patterns in duplicate ZF genes throughout vertebrates. Our results support a host-pathogen model for tandem ZF gene evolution, in which new LTR retroelement challenges drive duplication and divergence of host tandem ZF genes.

O-46

Duplicate and conquer: phylogenomics of MADS-box genes meets evo-devo of plants

Guenter Theissen
Friedrich Schiller University, Jena, Germany

MADS-box genes encoding MADS-domain transcription factors are involved in controlling almost all aspects of plant development. Several lines of evidence indicate that an increase of the MADS-box gene family during the evolution of land plants increased the capacity of plants to control different developmental processes and thus contributed to an increase in biodiversity. In my talk I will summarize several case studies revealing different duplication events in the MADS-box gene family that may have contributed to a diversification of flowering plant morphologies during evolution, ranging from single gene to whole genome duplications. My examples comprise ancient and recent events and include the occurrence of four paralogous *DEF*-like class B floral homeotic genes in Orchidaceae that facilitated diversification of the orchid flower; the origin of *GORDITA*-like genes within B_{sister} genes in Brassicaceae that may have affected the evolution of fruit size; and a promoter rearrangement and duplication of an *StMADS11*-like gene that generated pod corn (*Tunicate* maize).

O-47

The ancient syntenic roots of the human genome

Manuel Irimia[1,2], Juan J. Tena[3], Maria S. Alexis[1], Ana Fernandez-Miñan[3], Ignacio Maeso[4], Ozren Bogdanović[3], Elisa de la Calle-Mustienes[3], Scott W. Roy[1], Jose L. Gómez-Skarmeta[3], Hunter B. Fraser[1]
[1]Stanford University, Stanford, USA, [2]University of Toronto, Toronto, Canada, [3]Centro Andaluz de Biología del Desarrollo (CABD), Sevilla, Spain, [4]University of Oxford, Oxford, UK

Gene order, or microsynteny, is generally thought to be neutral, and thus is not expected to be conserved across many metazoan phyla. Only a handful of exceptions, typically of tandemly duplicated genes such as Hox genes, have been discovered. Here, we performed a systematic survey for microsynteny conservation in 17 genomes and identified nearly 600 pairs of unrelated genes that have remained together across over 600 million years of evolution. Using multiple genome-wide resources, including several genomic features, epigenetic marks, sequence conservation and microarray expression data, we provide extensive evidence that many of these ancient microsyntenic arrangements have been conserved in order to preserve either (i) the coordinated transcription of neighboring genes, or (ii) Genomic Regulatory Blocks (GRBs), in which transcriptional enhancers controlling key developmental genes are contained within nearby "bystander" genes. In addition, we generated ChIP-Seq data for key histone modifications in zebrafish embryos to further investigate putative GRBs in embryonic development. Finally, using chromosome conformation capture (3C) assays and stable transgenic experiments, we demonstrate that enhancers within bystander genes drive the expression of genes such as Otx and Islet, critical regulators of central nervous system development across bilaterians. These results show that ancient genomic functional associations may be far more common in modern metazoans than previously thought – involving up to 12% of the ancestral bilaterian genome – and that cis-regulatory constraints have played a major role in conserving the architecture of metazoan genomes.

O-48

Combining genomic sequencing and gene expression studies to identify the genetic basis of colour pattern diversity in *Heliconius* butterflies.

Nicola Nadeau[1], Carolina Pardo-Diaz[1], Grace Wu[1], Richard Wallbank[1], Kanchon Dasmahapatra[2], Simon Martin[1], Simon Baxter[1], Chris Jiggins[1]
[1]Department of Zoology, University of Cambridge, Cambridge, UK, [2]University College London, London, UK

The *Heliconius* butterflies are a system that has fascinated biologists since the early days of evolutionary biology because of their close mimicry between species and striking divergence within species in colour pattern. We have a relatively good understanding of the ecological factors driving the evolution of colour pattern and also its relatively simple Mendelian genetic control. However, identifying the particular genes controlling these traits has proven difficult. Whole genome and targeted re-sequencing of multiple colour pattern races and species of *Heliconius* has revealed narrow "islands of divergence" containing loci that control colour pattern. We also find particular alleles that are shared between populations and even species that share the same colour patterns, either ancestrally or through adaptive introgression. Combined with gene expression analyses this has allowed us to identify particular genes and even suggest candidate SNPs that control colour pattern.

O-49

Vertebrate ultraconserved noncoding elements are ultra-cooperative

Slavica Dimitrieva[1,2], Philipp Bucher[1,2]
[1]EPFL, Lausanne, Switzerland,[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Vertebrate ultraconserved non-coding elements (UCNEs) constitute one of the biggest mysteries in current biology. They are almost 100% identical between species as distant as human and shark, but no molecular mechanism is known that would require such a high degree of conservation. Experiments suggest that most UCNEs act as transcriptional regulators of master regulatory genes.

Our work follows up on the striking observation that UCNEs are highly clustered around a few key developmental genes. We were asking whether this striking co-localization reflects functional cooperativity between UCNEs. An alternative hypothesis would be that each UCNE acts independently on its target gene, in which case the clustering would merely reflect the importance of the target gene.

To address this question, we analyzed the fate of these elements after the whole genome duplication event that occurred in the fish lineage. We reasoned that UCNEs that are jointly retained in one duplicated gene but lost in the other could function in a cooperative manner. We were able to collect 385 cases where the potential target gene containing UNCEs was retained in two or more copies in a completely sequenced fish genome. We found that in most cases (>80%) all UCNEs are either jointly retained or jointly lost in the respective daughter genes, suggesting that these elements truly cooperate with one another. This finding is significant for two main reasons:

a) It explains the extraordinarily high degree of conservation. According to the scenario we propose, the base sequence of a UCNE cannot change during evolution because it has to properly interact with so many partner elements. There is a precedent of this phenomenon in the protein world. The ultraconservation of histones is generally explained by the fact that these proteins have to properly interact with hundreds of gene regulatory proteins.

b) It has important implications regarding the choice of experimental strategies to elucidate UCNE function. Reporter gene expression constructs driven by a single UCNE may not reproduce its behavior in the native genomic context. More promising are 3C experiments that can reveal physical interactions between distant DNA elements.

Our work furthermore constitutes “proof of concept” that comparative genomics approaches like “genomic context analysis” or “phylogenetic profiling”, that were successfully applied to infer protein-protein interaction networks, can also provide insights about interactions between non-coding sequences. However, to our knowledge, the same principles have never been used before to infer cooperative interactions between *cis*-regulatory elements.

O-50

Parallel evolution by parallel genetic mechanisms, or is evo-devo different?

Axel Meyer, Kathryn Elmer
Univ. of Konstanz, Konstanz, Germany

Understanding the interplay of natural selection and genomic variation in generating new phenotypes is a major goal of modern evolutionary biology. But, the connection between genotypic and phenotypic variation is almost completely lacking, at least for non-model systems.

Parallel evolution - the repeated and independent evolution of similar phenotypes - greatly improves our power to understand the connection of genome variation, natural selection and evo-devo mechanims by offering replicate experiments in adaptation and evolution. Foremost in the mind of many ecologists and evolutionary biologists is the hypothesis that parallel phenotypic evolution in closely related species is accomplished by recruiting homologous genomic and developmental pathways – evo-devo is conservative and does not change.

Presently, genomic resources and experiments, such as whole genome and transcriptome approaches with 'next-generation' methodologies, are getting increasingly within reach of more and more biologists studying ecological speciation and now permit one to test these ideas. I will review how studying parallel phenotypic evolution in conjunction with genomic work can inform about how selection in conjunction with genomic changes could result in repeated evolution. The knowledge existing so far – however does not reveal a consistent pattern about the genetic bases of parallel phenotype evolution.

O-51

Near Neutrality: the Mutational-Hazard Theory of Genome Evolution, and the Drift-Barrier Hypothesis

Micheal Lynch
Indiana University, Indiana, USA

Understanding the mechanisms of evolution and the degree to which phylogenetic generalities exist requires information on the rate at which mutations arise and their effects at the molecular and phenotypic levels. Although procuring such data has been technically challenging, high-throughput genomic sequencing is rapidly expanding our knowledge in this area. Most notably, information on spontaneous mutations, now available in a wide variety of organisms, implies an inverse scaling of the mutation rate (per nucleotide site) with the effective size of a lineage. The argument will be made that this pattern naturally arises as natural selection pushes the mutation rate down to a lower limit set by the power of random genetic drift rather than by intrinsic molecular limitations on repair mechanisms. Additional support for this idea derives from the relative levels of efficiency of DNA polymerases and mismatch-repair enzymes in eukaryotes relative to prokaryotes.

This drift-barrier hypothesis has general implications for all aspects of evolution, including the performance of enzymes and the stability of proteins. The fundamental assumption is that as molecular adaptations become more and more refined, the room for subsequent improvement becomes diminishingly small. If this hypothesis is correct, the population-genetic environment imposes a fundamental constraint on the level of perfection that can be achieved by any molecular adaptation. It also implies that effective neutrality is the expected outcome of natural selection, an idea first suggested by Hartl et al. in 1985.

Although generally viewed as an independent process, mutation also operates as a weak selective force, thereby playing a central role in "nearly neutral" hypotheses in evolution. Most notably, genes and proteins with more complex structures are subject to higher rates of mutational degeneration simply because they are larger mutational targets. However, because the mutation rate is very low at the nucleotide level, the efficiency of such mutation-associated selection becomes of diminishing significance in populations with small effective sizes. Thus, mutationally hazardous genomic features, which may or may not be adaptive, are expected to passively arise in lineages with small effective sizes. This general principle, the mutational-hazard theory, will be illustrated with examples including: 1) the differential expansion of intron numbers in various phylogenetic lineages; and 2) the diversification of protein-architectural features.

O-52

NATURAL SELECTION IN GENOME EVOLUTION

Giorgio Bernardi
Biology Dept., Rome 3 University, Rome, Italy, Italy

"Most of the familiar features of living organisms show clear signs of adaptation of structure to function. There is overwhelming evidence that this is the outcome of evolution by natural selection" (B. Charlesworth). Whether the same applies to the genomes of vertebrates (and other eukaryotes) with their overwhelming amount of non-protein-coding DNA is, however, an open question. Many years ago, we developed a compositional strategy which revealed that those genomes are mosaics of isochores. These are long DNA stretches fairly homogeneous in base composition that belong to a small number of families characterized by different GC levels and different short-sequence patterns (*i.e.,* different DNA structures and different DNA-protein interactions). This genome organization led us to two discoveries: (i) the genomic code is a collective definition for the correlations that hold between coding and non-coding sequences; between coding sequences and the structural properties of the encoded proteins; and between the frequencies of short sequences of isochore families and nucleosome positioning/transcription factor binding. (ii) The genome phenotypes correspond, at low resolution, to the patterns of isochore families in the genome; at high resolution, to isochore maps on chromosomes. While the latter may be used to study genomic variation and genomic diseases, the former showed that genome evolution may proceed according to a conservative mode or to a transitional (shifting) mode. The conservation and the changes of isochore patterns depend upon whether the environment is constant or shifting. According to the "neo-selectionist theory", natural selection is responsible for both modes and plays a dominant role in genome evolution.

O-53

Are there severe limits to the power of selection?

Brian Charlesworth
University of Edinburgh, Edinburgh, Scotlan, UK

Fisher argued that population sizes are so large and mutation rates are so low that natural selection is in control of most evolutionary changes of any biological significance. At the level of individual amino-acid or nucleotide variants, this conclusion is almost certainly correct when the product of effective population size (Ne) and selection coefficient (s) is of the order of one or more. But work on the intensity of purifying selection against amino-acid mutations, and mutations in functionally important noncoding sequences, suggests that there is some fraction of such mutations that violate this condition, especially if Ne is low, either for demographic reasons or because of a low rate of genetic recombination. The degree of adaptation at the molecular sequence level is then jointly determined by mutation, selection, genetic drift, and other factors such as biased gene conversion and the hitchhiking effects of selection at linked sites. In extreme cases, this can lead to complete loss of functionality, as in degenerating Y chromosomes.
However, most phenotypes of interest are controlled by many genetic loci, and many of the DNA sequences whose molecular evolution and variation have been studied probably contribute to variation in such phenotypes. Relatively low levels of effective selective constraints on individual sequences affecting a complex phenotypic trait do not imply a low level of effectiveness of selection at the level of the phenotype, as this depends on the product of Ne and the intensity of selection on the phenotype as a whole, not on the individual genetic variants. The existence of highly adapted phenotypes in species with low Ne values, such as humans , is therefore not unexpected.

O-54

The known unknowns of complex environments: how reliably can we place metagenomic samples onto reference phylogenetic trees?

Nick Goldman[1], Eliza Loza[2]
[1]EMBL-European Bioinformatics Institute, Hinxton, Cambs, UK, [2]Department of Computational & Systems Biology, Rothamsted Research, Harpenden, UK

Metagenomics is the study of microbial DNA directly extracted from a habitat (e.g. agricultural soil, ocean water, or the human gut). When using second-generation sequencing technologies, the observed data are thousands of short DNA sequences that originate from the genomes of the microorganisms that populate the sampled habitat. A typical objective in a metagenomic study is to associate these metagenomic fragments with the operational taxonomic units (e.g. species, strains, or populations) or functions of their origin. Some existing methods build upon the strength and reliability of likelihood-based phylogenetic approaches, such as the placement of metagenomic sequences onto a reference phylogeny.

The input in phylogenetic placement consists of a reference tree and the metagenomic data. The parameter of primary inferential interest is the set of assignments of the metagenomic fragments onto the reference tree. Once the assignments are observed, their distribution along the edges of the reference tree is used as an indication of biodiversity and, in some cases, of relative abundance of the microbes in the habitat. A variety of approaches has been used to arrive at the reference trees used in applications of phylogenetic placement, including a reference tree of all the organisms whose genome has been fully sequenced, and trees constructed from alignments of concatenated marker genes. However, no study that we are aware of has assessed the effects of the reference tree on the inferential process. In studies of exceptionally rich and biodiverse habitats assignments of metagenomic fragments onto a reference tree will represent habitat biodiversity as poorly/adequately as the reference tree itself represents life biodiversity.

In this talk I will discuss the phylogenetic placement of metagenomic data onto reference trees and will argue that principles of experimental design need to be considered for optimal choices of reference trees. I will make reference to the long history of experimental design at Rothamsted Research, since its conception by Ronald A. Fisher in the 1930's, and to the metagenomic experiments we are currently conducting.

O-55

Uncovering the Tree of Life in the genomic era: separating the wheat from the chaff

Jeremy Brown
Louisiana State University, Baton Rouge, LA, USA

Massive genome-scale datasets generated by high-throughput DNA sequencing offer an unprecedented wealth of information for inferring phylogenies. However, proper interpretation of this information is dependent on the stochastic model of genome evolution used for phylogenetic inference. While much biological realism is incorporated into models, some assumptions must always be made to ensure tractability. The extent to which these assumptions match the actual evolutionary process varies across the genome, as does the quality of both the sequence data and alignment. Therefore, the reliability of the inferences returned by a model also varies. Inferences drawn from genomic regions where the data do not match the model may be systematically biased. Despite such variation in the fit of a model across the genome, little effort has been devoted to developing quantitative methods for directly assessing when inference has been compromised. Such methods could be used to filter genomic data and discard misleading signal, to identify unforeseen (and potentially important) biological processes, to compare the reliability of conflicting studies and datasets, and to design better models of sequence evolution. Here, I propose a robust and flexible approach to directly evaluating the reliability of phylogenetic inferences. The proposed approach employs Bayesian posterior predictive simulation and applies test statistics that can be tailored to assess the reliability of any inference of interest. These statistics compare inferences drawn from empirical data to those drawn from data simulated under the assumed model. For example, genomic data may be filtered by the overall reliability of the inferred tree topology, by the reliability of inferred support for particular branches, by the reliability of inferred tree (or branch) lengths, and even by the reliability of inferred model parameter values. I illustrate the flexibility and utility of the new tests through analyses of simulated data and application to exemplar empirical data sets, including ongoing work on HIV strains in small transmission clusters. Analyses of simulated data show that biased phylogenetic signal resulting from oversimplified models can robustly be identified. The detection of bias is also inference-specific (e.g., if only branch lengths are biased, tests do not identify the topology as biased, and vice versa). Analyses of empirical data show that the method is able to reveal variation in the reliability of phylogenetic conclusions across data sets. I also outline simple extensions of this approach to population genetic and comparative models.

O-56

Quantifying phylogenetic signal

Mike Steel
University of Canterbury, Christchurch, New Zealand

Molecular systematics relies on the notion that one can reliably extract phylogenetic signal from genetic data in the presence of random 'noise' and other confounding influences (such as alignment errors, model misspecification etc). Even for 'ideal' data, evolving on a tree under a well-behaved model, there are fundamental information-theoretic limits to resolving evolutionary relationships, especially those that involve short branches deep within a tree. I will outline these results, and show how they provide easily computable mathematical bounds on the sequence length requirements for resolving deep phylogenetic divergences.

O-57

Phylogenetic signal and noise: predicting the power of a dataset to resolve molecular phylogeny

Jeffrey Townsend[1], Zhuo Su[1], Yonas Tekle[2]
[1]Yale University, New Haven, CT, USA, [2]Spelman College, Atlanta, GA, USA

A principal objective for maximizing the impact of molecular evolutionary and phylogenetic studies is to predict the power of a dataset to resolve nodes in a phylogenetic tree. However, marshalling data to proactively assess the potential for phylogenetic noise compared to signal in a candidate dataset presents a formidable challenge. We will present a theory that applies estimates of the state space and the rates of evolution of characters in a dataset to predict phylogenetic signal and phylogenetic noise and therefore to predict the power to resolve internodes. The theory is implemented as a Monte Carlo approach to estimating power to resolve, but we also have derived an equivalent, faster deterministic calculation. These approaches are applied to describe the distribution of potential signal, polytomy, or noise for several example datasets, including at least one recent (cytochrome c oxidase I and 28S ribosomal rRNA sequences from Diplazontinae parasitoid wasps) and one deep (eight nuclear genes and a phylogenomic sequence for diverse microbial eukaryotes including Stramenopiles, Alveolata, and Rhizaria). The predicted power of resolution for the loci analyzed is consistent with the historic use of the genes in phylogenetics. Because high-throughput data sets yield such high power in terms of sample size, many problems of inference need to be evaluated in terms of parsing signal from noise, rather than applying traditional rubrics for nodal support. Understanding the impact of collection of additional sequence data to resolve recalcitrant internodes at diverse historical times will facilitate increasingly accurate and cost-effective, accurate research on molecular evolution and phylogenetics.

O-58

Fundamentals of Molecular Evolution Revisited

Consortium of Wen-Hsiung's Research Phylogeny, Soojin Yi
Georgia Institute of Technology, Atlanta/GA, USA

In honor of Wen-Hsiung Li's long career in molecular evolution, we present the remarkable contributions of Dr. Wen-Hsiung Li to studies of molecular and genome evolution. This presentation will be an opportunity for molecular evolutionists to review some of his many achievements during his career and discuss future studies to follow up new questions stemming from his work.

O-59

Molecular and Genomic Evolution: Personal Retrospects and Prospects

Wen-Hsiung Li
University of Chicago, Chicago, Illinois, USA

I was fortunate to be "there" at the very begging of the "DNA era" of molecular evolution and at the very begging of genomic evolution. I shall briefly describe a few examples of how I used newly available DNA sequence data or genomic data to solve some outstanding problems. I shall then discuss two major current projects in my lab. One is "Avian Evolutionary Genetics and Genomics". Birds show tremendous phenotypic diversity, providing excellent materials for evolutionary study. Currently we focus on the genetic and molecular basis of feather evolution because feather materials are not difficult to collect and feather morphological variation has always been an interesting topic in evolution. The other major project is "Genetic Basis of the Evolution of C4 Photosynthesis in Grasses". C4 photosynthesis is more efficient than C3 photosynthesis. The evolution from C3 to C4 photosynthesis requires not only changes in the enzymes involved in photosynthesis, but also changes in cell structure, known as the kranz anatomy. While the evolution of C4 enzymes has been well studied, the genetic and cellular changes responsible for the emergence of the kranz anatomy are virtually unknown. The emergence of C4 photosynthesis has occurred many times in grasses and that is one reason for our focus on grasses. We are especially interested in identifying the regulators responsible for the development of the kranz anatomy. Our eventual goal is to transform rice, a C3 crop, into a C4 crop to increase rice production. I shall explain the issues and approaches of each project and show some results we have obtained so far.

O-60

Dosage sensitive genes in evolution and disease.

Aoife McLysaght
Trinity College, Dublin, Ireland

Evolutionary change of gene copy number through gene duplication is a relatively pervasive phenomenon in eukaryotic genomes. However, for a subset of genes such changes are deleterious because they result in imbalances in the cell. Such dosage-sensitive genes have been increasingly implicated in disease, particularly through the association of copy number variants (CNVs) with pathogenicity.

We have previously shown that many genes in the human genome which were retained after whole genome duplication (WGD) are refractory to gene duplication both over evolutionary timescales and within populations. These are expected characteristics of dosage-balanced genes. Many of these genes are implicated in human disease. I now show how this evolutionary information can be used to narrow down the list of candidate genes from CNV studies in cases and controls.

O-61

Hypothesis of steady-state genetic robustness implies a dynamic balance between gene essentiality turnovers through gene/genome duplications

Xun Gu[1,2], Zhixi Su[2], Wei Huang[1], katie Gu[1]
[1]*Iowa State University, Ames, Iowa, USA,* [2]*Fudan University, Shanghai, China*

Complex networks of various biological systems show extraordinary robustness against deleterious mutations. Simply speaking, two major mechanisms were proposed for a gene to be nonessential: the genetic buffering from redundant gene networks and the functional compensation of duplicate genes. This seemingly intuitive statement has been recently challenged by two observations: the mouse knockout data that the proportion of essential genes (P_E) is similar between (ancient) duplicate and single-copy genes, and RNAi knockout of *Drosophila* that P_E is similar between young duplicate and single-copy genes. To resolve this problem, we develop a probabilistic model of duplicate essentiality by considering the ancestral essentiality prior to the gene duplication. We then propose the hypothesis of steady-state genetic robustness for similar P_E between duplicates and single-copy genes: The genomic flux from essential single-copy genes to compensated duplicates is roughly equal to that from nonessential single-copy genes to essential duplicates. Moreover, we estimated that the rate of a duplicate gene to be essential is roughly $3\text{-}4\times10^{-9}$ per year per gene.

O-62

Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification

Robert Meredith[1], Jan Janecka[2], John Gatesy[1], Oliver Ryder[3], Colleen Fisher[2], Emma Teeling[4], Alisha Goodbla[4], Eduardo Eizirik[5], Taiz Simao[5], Tanja Stadler[6], Daniel Rabosky[7], Rodney Honeycutt[8], John Flynn[9], Colleen Ingram[9], Cynthia Steiner[3], Tiffani Williams[2], Terence Robinson[10], Angela Burk-Herrick[1], Michael Westerman[11], Nadia Ayoub[1], Mark Springer[1], William Murphy[2]
[1]University of California, Riverside, Riverside, CA, USA,[2]Texas A&M University, College Station, TX, USA,[3]San Diego Zoo's Institute for Conservation Research, Escondido, CA, USA,[4]University College Dublin, Belfield, Dublin, Ireland,[5]PUCRS, Porto Alegre, RS, Brazil,[6]Eidgenössiche Technische Hochschule Zurich, Zurich, Switzerland,[7]University of California, Berkeley, Berkeley, CA, USA,[8]Pepperdine University, Malibu, CA, USA,[9]American Museum of Natural History, New York, NY, USA,[10]University of Stellenbosch, Matieland, South Africa,[11]LaTrobe University, Bundoora, Victoria, Australia

Previous analyses of relationships, divergence times, and diversification patterns among extant mammalian families have relied on supertree methods and local molecular clocks. We constructed a molecular supermatrix for mammalian families and analyzed these data with likelihood-based methods and relaxed molecular clocks. Phylogenetic analyses resulted in a robust phylogeny with better resolution than phylogenies from supertree methods. Relaxed clock analyses support the Long Fuse Model of diversification and highlight the importance of including multiple fossil calibrations that are spread across the tree. Molecular timetrees and diversification analyses suggest important roles for the Cretaceous Terrestrial Revolution and Cretaceous/Paleogene mass extinction in opening up ecospace that promoted interordinal and intraordinal diversification, respectively. By contrast, diversification analyses provide no support for the 'Eocene delayed rise of present-day mammals' hypothesis. Prospects for fully resolving the mammal tree of life will be presented.

O-63

The use of Heterogeneous Evolutionary models to infer the Placental Mammal Phylogeny.

Claire C. Morgan[1,2], Peter G. Foster[3], Andrew E. Webb[1,2], Davide Pisani[4], James O. McInerney[4], Mary J. O'Connell[1,2]
[1]Dublin City University, Dublin, Ireland, [2]Centre for Scientific Computing & Complex Systems Modeling (SCI-SYM), Dublin, Ireland, [3]Natural History Museum, London, UK, [4]Molecular Evolution and Bioinformatics Unit, Co. Kildare, Ireland

Resolving the order in which the placental mammal Superorders arose has proven extremely difficult. There are currently four major competing hypotheses for the topology of placental mammals each of which have been generated using molecular data and various phylogenetic approaches. Variation in body size, longevity, metabolic rates and age at which reproductive maturity is reached can greatly impact the accumulation of mutations in a genome and results in mixed phylogenetic signal. These variations in life traits are seen across the placental mammals, however all methods to analyse their relationships to date have applied homogeneous models. In this presentation we will describe the analysis of a previously published dataset and a novel dataset (composed of single gene orthologous families). We address the issue of the suitability of the datasets and phylogenetic models to answer the question of the placental root position. We have used two heterogeneous modelling approaches that allow for multiple rate matrices and composition vectors (p4), and variation in the rate matrix (CAT model PhyloBayes). We show strong statistical support for the "Atlantogenata Hypothesis" and the rejection of all other hypotheses. The "Atlantogenata Hypothesis" places the root of the placental mammals on the lineage leading from the common ancestor of Afrotheria (e.g. elephants) and Xenarthra (e.g. armadillos). Heterogeneous modelling, accounting for variations in both the rate and composition of placental mammal data, together with a large, novel, and high-quality mammal dataset have allowed us to address the longstanding question of the position of the placental mammal root.

O-64

A combined analysis of Eutherian relationships: microRNAs and Phylogenomics.

James Tarver[1,2], Philip Donoghue[1], Kevin Peterson[2]
[1]University of Bristol, Bristol, UK, [2]Dartmouth College, Hanover, USA

microRNAs are a rare genomic character that has been used to resolve the position of some of the most troublesome nodes within the metazoan tree of life - such as the monophyly of cyclostomes and the origin of turtles and tardigrades - nodes which have long troubled both palaeontologists and neontologists alike.

Here we discuss the use of microRNAs in phylogenetic analyses by considering four unique properties. Firstly, substitutions to the mature sequence are rare. Secondly, microRNAs are continuously acquired through evolutionary time. Thirdly, there is minimal secondary loss, and finally, the likelihood of convergence evolution in generating two identical miRNAs is exceedingly small.

Given these characteristics, we use microRNAs to resolve the relationships amongst Eutherian mammals. The relationships between the four major clades of Eutherians (Xenarthrens, Afrotherians, Laurasiatheria and Euarchontoglires) are still contentious, with various datasets suggesting, that either, the Afrotherians, Xenarthrens or Atlantogeneta (Afrotherians + Xenarthrens), are the earliest diverging lineage. However, the miRNA dataset suggested a fourth hypothesis, that the murid rodents are basal and that Rodentia itself is paraphyletic.

This is a highly controversial result with wide ranging implications for many aspects of mammalian biology. Thus a reanalysis of the largest mammalian phylogenomic dataset to date (Hallström and Janke, 2010 MBE) was conducted. Analyses using Phylobayes under a CAT model showed that there is little support for any one topology and that slight perturbations to the data set can result in a variety of alternative hypotheses. In addition the branch leading to Eutherian mammals was considerably longer than the internal branches leading to the four major clades, suggesting that current phylogenomic datasets are unable to accurately resolve such rapid speciation events.

O-65

The Timescale of Mammalian Phylogeny

Mario dos Reis[1], Jun Inoue[2,1], Masami Hasegawa[3], Robert Asher[4], Philip Donoghue[5], Ziheng Yang[1]
[1]University College London, London, UK, [2]University of Tokyo, Tokyo, Japan, [3]Fudan University, Shangai, China, [4]University of Cambridge, Cambridge, UK, [5]University of Bristol, Bristol, UK

The fossil record shows the sudden appereance of Placental mammals 65 My ago, soon after the Cretaceous-Paleogene mass extinction event when 76% of species, including non-avian dinosaurs, died out. Some molecular studies have placed the origin and diversification of placental mammals deep in the Cretaceous, before the mass extinction event. On the other hand, paleontological analysis of the fossil record indicate a diversification peak of placental orders in a short 16 My post-Cretaceous window. The discrepancies between estimates of mammalian divergence between molecular and paleontological studies have been considered unacceptably large. We revised the timeline of mammalian diversification by Bayesian analysis of a large alignment (21 million sites) from 36 mammalian genomes, together with mitochondrial data from 274 mammal species. We used 26 soft, non-limiting fossil constraints to calibrate the molecular clock. Our analysis indicates that intraordinal diversification of placental mammals occurred in a 20 My post-Cretaceous window, in accordance with paleontological estimates. We find that genomic data reduces uncertainty in divergence time estimates towards the theoretical limit of precision and allowed us to confidently reject a pre-extinction diversification of placentals. Other famous controversies of mismatch between paleontological and molecular time estimates, such as the origin of animal phyla or of flowering plants, are likely to be resolved with a genomic-scale approach.

O-66

Estimating species divergence times with molecular sequence data but without fossils.

Hui-Jie Lee[1], Hirohisa Kishino[2], Jeffrey L. Thorne[1]
[1]North Carolina State University, Raleigh, NC, USA, [2]University of Tokyo, Tokyo, Japan

The amount of sequence divergence depends on the product of the rate of molecular evolution and the time since common ancestry. When only sequence data are available, these rates and times are confounded. Conventionally, rates and times are disentangled by supplementing interspecific sequence information with additional time information (e.g., information from fossils or from "ancient" DNA). We are exploring the alternative approach of separating evolutionary rates and divergence times by supplementing interspecific sequence data with information about rates. The basic idea is that the rate of selectively neutral molecular evolution equals the mutation rate and, with the advent of high-throughput sequencing, it is becoming increasingly practical to characterize mutation rates and patterns from parent-offspring data and/or from mutation accumulation lines.

We have modified the BEAST software package so that divergence times can be estimated from interspecific sequence data together with information on mutation that is collected by high-throughput sequencing. The approach can be applied when evolutionary rates are treated as being constant through time and also when the evolutionary rates themselves evolve. We believe that this research direction will have particular value when studying parts of the tree of life that have poor fossil records and also when studying how mutation rates evolve.

Our preliminary work has been to assess when this technique for inferring evolutionary times and rates does and does not succeed. To broaden the applicability of our approach, we are working to replace the assumption of selective neutrality with one of mutation-selection balance. We are also striving to make our treatments of the mutation process more realistic and to consider the potential impact of sequencing errors.

O-67

Phylogenetic network construction as default method for reticulated evolution

Naruya Saitou
National Institute of Genetics, Mishima, Japan

Good old days of constructing phylogenetic trees from relatively short sequences were over. Reticulated or "non-tree" structures are omnipresent in genome sequences, and construction of phylogenetic network is now default for describing these complex realities. Recombinations, gene conversions, and gene fusions are biological mechanisms to produce non-tree structures to gene phylogenies, while gene flow is well known factor for creating reticulations to population phylogenies. I will mostly focus on detection of recombination via phylogenetic network constructions. Kitano, Blancher, and Saitou (Jan. 27, 2012; Mol. Biol. Evol., published online) found that A allele of ABO blood group gene in modern human was resurrected by recombination of B and O alleles by applying Kitano et al.'s (2009; Mol. Phyl. Evol., vol. 51, pp. 465-471) method for detecting recombinants from phylogenetic network. I will demonstrate generality of this method by showing examples from viral sequences and HLA haplotypes.

O-68

Inferring the network of life via agreement forest based models

Christopher Whidden, Norbert Zeh, Robert Beiko
Dalhousie University, Halifax, Nova Scotia, Canada

Evolutionary trees are increasingly regarded as an inadequate representation of the history of life due to lateral gene transfer, recombination, hybridization and other inherently non-treelike evolutionary processes. Individual genes can potentially have tree-like histories but a forest of thousands (or even millions) of gene trees is not a practical representation for understanding evolution. To display meaningful representations of how life evolved, we propose using models based on agreement forests to construct networks that minimize the number of inferred reticulation events.

First, we present our unifying framework for efficiently computing maximum agreement forests (MAFs) and maximum acyclic agreement forests (MAAFs) of two evolutionary trees. The sizes of these forests are equivalent to the subtree-prune-and-regraft (SPR) distance and hybridization number, respectively. Moreover, an agreement forest itself represents a network of the two trees. In particular, an MAAF represents a most parsimonious scenario of reticulation events that does not violate time constraints. Although NP-hard to compute, we show that these distances can often be computed efficiently in practice. Using our framework we compute SPR distances as large as 46 in less than one second; other approaches require 5 hours or more to compute SPR distances larger than 20.

Second, we apply our efficient algorithms for computing maximum agreement forests to the construction of SPR supertrees - supertrees with the minimum SPR distance to a set of gene trees. Although supertrees may be inadequate to represent the history of life, their construction provides a useful demonstration of our inference methods and a first step towards constructing networks. Our results using a simulated history of gene births, losses, and duplications suggest that SPR supertrees are more similar to the species history than Robinson-Foulds supertrees or MRP supertrees under plausible rates of lateral gene transfer.

Finally, building on our framework for inferring SPR supertrees, we propose inferring networks of life that minimize the number of reticulation events required to explain the individual gene histories. To realize the goal of constructing meaningful but understandable networks of life, our proposed SPR network framework allows restricting the number of edges in the resulting network. We close with a discussion on possible visual representations of such networks that provide more information about the underlying gene trees using observations based on agreement forests. Such representations could be used to identify frequent paths ("highways") of gene sharing.

O-69

Assessing and Visualising Genetic Diversity at the Population Level

Jessica W. Leigh, David Bryant
University Of Otago, Dunedin, New Zealand

Haplotype networks are a powerful tool for visualising gene genealogies at the population level and for inferring demographic and phylogeographic scenarios. These are graphs in which sequences sampled from a population are represented by vertices connected to one another and to vertices representing unsampled sequences by edges that each represent a single mutation event. Generally these networks also contain information on frequency with which different haplotypes were sampled and often other information such as sampling location or associated phenotypic traits. One of the most popular methods used for inferring haplotype networks is statistical parsimony, generally using the TCS algorithm (1). The intent of the TCS algorithm is to estimate a network in which the most parsimonious genealogical trees are embedded, though this is not always the result. In addition, TCS has been shown to perform less well than other network algorithms (2) as well as traditional phylogenetic inference methods (3). The sometimes poor performance of TCS is partly due to a poorly defined optimality criterion and the arbitrary nature of penalties used when producing connections. We will present a new algorithm for inferring haplotype networks with more clearly defined decision-making criteria, implemented in an open-source software package. This software will also implement TCS, as well as other available haplotype inference methods, providing a flexible tool for population genetic and phylogenomic analysis.

1. Clement M, Snell Q, Walker P, Posada D, Crandall K. IEEE Trans Parallel Distrib Syst. 2002.

2. Cassens I, Mardulyn P, Milinkovitch MC. Syst Biol 2005 54(3):363-72.

3. Salzburger W, Ewing GB, Von Haeseler A. Mol Ecol 2011 20(9):1952-63.

O-70

Phylogenomic networks provide insights into the chimerical origin of eukaryotes

David Alvarez-Ponce[1], Eric Bapteste[2], Philippe Lopez[2], James McInerney[1]
[1]National University of Ireland Maynooth, Maynooth, Ireland, [2]Université Pierre et Marie Curie, Paris, France

The relationships between the three domains of cellular life (Eubacteria, Archaebacteria and Eukaryotes) have been the subject of debate ever since their definition. In particular, the events that led to the emergence of Eukaryotes, and their relatedness to the other two domains, remain controversial. Some models have placed Eukaryotes as the sister group of Archaebacteria, or within Archaebacteria, whereas others state that Eukaryotes arose from a fusion event involving at least an Archaebacterium and a Eubacterium, and others propose that Archaebacteria and Eubacteria arose from a Eukaryote-like ancestor. When it comes to establishing the relationships between lineages as distant as the three domains of cellular life, using phylogenetic trees can entail a series of shortcomings, including difficulties with obtaining accurate multiple sequence alignments, and choosing suitable models of evolution. Phylogenomic networks (whose nodes and edges represent genes and homology statements, respectively) are an interesting alternative that can overcome these shortcomings. Here, we have constructed a phylogenomic network including the proteomes of 9 eukaryotes, 52 archaebacteria, 52 eubacteria, 2390 viruses and 469 plasmids. In this talk, it will be discussed how the structure of this phylogenomic network can provide insight into the origin and evolution of eukaryotes.

O-71

A network of everything? Open issues in lateral gene transfer and network construction

Robert Beiko
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

Recent work in microbial phylogenomics has demonstrated that a canonical Tree of Life cannot be convincingly reconstructed using aggregated sequence data, and is not a valid representation unless certain genes are accorded a privileged status as the true indicators of vertical descent. Even in such cases, the recovered tree can be very sensitive to small perturbations in the data, such as the inclusion or exclusion of certain critical loci, the use of different parameters for sequence alignment or phylogenetic inference, or recoding of the input sequences.

Switching to a network paradigm is justified, but network methods differ in their scalability and their ability to handle phylogenetic data in an appropriate matter. For example, maximum likelihood and Bayesian analysis typically infer unrooted trees, but many network approaches require rooted trees, and identifying the correct root position in a gene tree, especially when set of genes have undergone a great deal of lateral gene transfer in their history, is not trivial.

Two major challenges present themselves. The first lies in the definition of suitable phylogenomic data sets to serve as raw material for network reconstruction, and the second lies in the choice of appropriate network construction algorithms and visualizations that can adequately represent important reticulations without descending into a "hairball" of inscrutable connections.

In this talk I will present these two major challenges, and provide some examples from microbial data that demonstrate how widespread these problems are. In the extreme, some taxonomically disparate organisms such as acidophiles show evidence of massive, likely habitat-directed gene transfer. Acidithiobacillus alone counts over 200 other genera as partners in phylogenetic trees; how can we show all these relationships, or decide on a particularly compelling subset?

O-72

Maximum-likelihood methods for the estimation of population-genetic parameters from high-throughput sequencing data

Micheal Lynch
Indiana University, Indiana, USA

High-throughput sequencing strategies are now leading to the acquisition of high-coverage genomic profiles of hundreds to thousands of individuals within species, generating unprecedented levels of information on patterns of nucleotide heterozygosity and linkage disequilibrium and on the frequencies of nucleotides segregating at individual sites. While offering unprecedented power for the acquisition of population-genetic parameters, these new methods also introduce a number of challenges, most notably a need to account for the sampling of alternative parental alleles at diploid nucleotide sites and the introduction of spurious variation by read errors.

To minimize the effects of both problems and to avoid *ad hoc* decisions on data utilization and error rates, we are developing maximum-likelihood methods for generating unbiased and nearly minimum-variance estimates of a number of key parameters, including average nucleotide heterozygosity and its variance among sites, the pattern of decomposition of linkage disequilibrium with physical distance, the rate and molecular spectrum of spontaneously arising mutations, and the allele-frequency spectrum for segregating polymorphisms. These methods define the limits to our ability to estimate population-genetic parameters, while also providing a formal basis for identifying optimally efficient experimental designs, e.g., the tradeoff between numbers of individuals sampled and depth of sequence coverage per individual.

A general description of the methodology will be presented, along with numerous applications to fully sequenced genomes. These methods provide a powerful platform for estimating population-genetic parameters even when the sample consists of a single individual.

O-73

Inferring adaptive evolution in the murid genome

Peter Keightley[1], Daniel Halligan[1], Rob Ness[1], Athanasios Kousathanas[1], Bettina Harr[2], David Adams[3], Thomas Keane[3]
[1]University of Edinburgh, Edinburgh, UK, [2]Max-Planck-Institute for Evolutionary Biology, Ploen, Germany, [3]Wellcome Trust Sanger Institute, Hinxton, UK

We have attempted to quantify the number and fitness effects of adaptive nucleotide substitutions originating from amino acid and regulatory mutation by comparing the genome-wide frequency distributions of segregating nucleotide sites in wild house mice and related murid rodents. To do this, we have analysed the genome sequences of 10 wild Mus musculus castaneus individuals from the species' ancestral range in NW India along with a sequence of M. famulus. We have developed an extension of the McDonald-Kreitman test to estimate the frequency and effects of adaptive mutations. There appears to be substantial adaptive regulatory and protein evolution, and moreover, we estimate that X-linked proteins experience ~1.7 times more adaptive change than autosomal proteins. By comparing numbers of adaptive substitutions in coding and conserved noncoding sequences, we infer that there are at least 9 times as many adaptive substitutions in noncoding elements, including regulatory sequences upstream and downstream of genes, as in protein-coding genes. On average, the fitness effects of regulatory mutations appear to be smaller, but their overall contribution to adaptive fitness change is estimated to be ~7 times greater than amino acid-changing mutations.

O-74

Admixed human genomes reveal complex demographic patterns from early modern humans to the contemporary era

Simon Gravel[1], Jeffrey M Kidd[2], Jake K Byrnes[1], Andres Moreno Estrada[1], Fouad Zakharia[1], Shaila Musharoff[1], Francisco M De La Vega[1], Carlos D Bustamante[1]
[1]Stanford University, Stanford, CA, USA,[2]University of Michigan, Ann Arbor, MI, USA

A substantial proportion of humans are "admixed", in the sense that their recent ancestors belong to statistically distinct groups. This needs to be accounted for if unbiased inference and associations are to be performed. We present a diversity of methods for the analysis of whole-genome sequence data from admixed individuals, and apply them to 50 genomes sequenced by Complete Genomics, including 4 Mexican-Americans, 4 African-Americans and 2 individuals from Puerto Rico, together with SNP genotype data from hundreds of additional samples.

Many methods have been presented recently to infer the population of origin of specific loci along the genomes of admixed individuals, leading to inferred mosaics of ancestry. We first propose a simple Markov model that relates the time-dependent migration history to the inferred patterns of local ancestry. We use this framework to infer the timing of admixture and to differentiate between punctual and continuous models of migration: using demographic models that are consistent with both historical records and genetic data, we find evidence for continuous migration patterns in both Mexican and African-American populations.

We also propose models to study the longer-term evolution of the ancestral populations, by considering the allele frequency distribution, pairwise TMRCA's, and a simple extension of the recently introduced Pairwise Sequentially Markovian Coalescent approach for demographic inference. The inferred source population demographic histories are in broad agreement with previous results for European and West-African populations, and the inferred demography for the Native source population closely follows the European one until about 20,000 years ago. Taken together, whole genome sequencing and local ancestry assignment therefore permit inferences about long-term histories of unsampled ancestral populations and highlights recent historical demographic processes that altered patterns of variation observed in admixed populations.

O-75

Characterization of Great Ape genetic diversity by whole genome sequencing.

Peter Sudmant[1,3], Javier Prado[2,3], Tomas Marques-Bonet[2,3], Evan Eichler[1,3]
[1]University of Washington Genome Sciences, Seattle, Wa, USA,[2]Institut de Bioligia Evolutiva, Barcelona, Spain,[3]Great Ape Sequencing Consortium, Seattle, Wa, USA

Although human genome diversity has been the subject of extensive research, the diversity of non-human Great Apes has garnered much less attention. We set out to understand the full spectrum of genetic diversity from single nucleotide to large structural variation using high-coverage next-generation sequencing of a diverse sample of 109 Great Apes. A total of 8.2 tera-bases were sequenced in total with a mean of 25x sequence coverage per individual. The set includes wild-born/unrelated specimens from all major subspecies including 28 chimpanzees (10 *P.t.elioti*, 6 *P.t. verus*, 8 *P.t.schwenifurthii*, 4 *P.t.troglodytes*), 16 bonobos (*P.paniscus*), 39 gorillas (35 *G.gorilla* and 4 *G.graueri*) and 16 orangutans (8 *P.abelii* and 8 *P.pygmaeus*). We characterize the SNP diversity of each of these species and define lineage specific duplications and deletions. We have accurately (>98.5%) cataloged all fixed differences between species and subspecies and generated the first detailed map of SNPs among chimps (34M/million SNPs), gorillas (25M), orangutans (36M) and bonobos (11M). Comparisons within species revealed unexpected differences in genetic diversity. Chimpanzee subspecies show the greatest population structure with heterozygosities ranging from approximately equal to twice that of humans. *P.t.verus* show the greatest number of fixed differences compared to sister subspecies likely due to their geographic isolation in Western Africa. We find that *G.gorilla* individuals are almost twice as diverse at the SNP level than *G.b.grauri,* suggesting an ancient bifurcation and bottleneck similar to that of the two orangutan species. Overall, we predict 923 genes have been lost within different ape lineages due to fixed disruptive mutations. While most forms of genetic diversity have behaved in a clock-like manner, both within and between species, copy-number variation especially that associated with segmental duplications appears more episodic. We can now date all gene duplications and determine whether they are fixed or polymorphic between subspecies. For example, we identify 3342 regions corresponding to 1318 genes in orangutans which have undergone deletion or duplication exclusively in the Bornean or Sumatran lineage. Similarly, we identified 53 human specific gene duplications that emerged after divergence with chimpanzees including several genes which are fixed in all humans (n=801) and have specific roles in neuronal development. These data provide a rich resource to understand the biology of great apes, reconstruct their population history and improve conservation efforts for these remarkable species.

O-76

Challenges and prospects for molecular phylogenetics on disparate branches of the eukaryotic Tree of Life

Eduardo Eizirik, Laura Utz, Taiz Simão, Tatiane Trigo, Alexsandra Schneider
PUCRS, Porto Alegre, RS, Brazil

Molecular phylogenies illuminate many aspects of the history of life on Earth, from the origin of major groups and the dynamics of rapid radiations, to the temporal and spatial contexts of adaptive changes across all levels of the biological hierarchy. Although recent decades have seen tremendous progress in reconstructing the Tree of Life, many challenges remain, and they differ substantially among the various groups of organisms that comprise the Earth's biodiversity. Some groups have been the focus of intense phylogenetic study, which has resulted in robust knowledge of their evolutionary relationships and timing of divergence. Still, are there phylogenetic challenges remaining in such groups? The answer is 'yes'. Using the mammalian order Carnivora as a case study, we will show the current stage of resolution of its phylogenetic structure and divergence dating, highlighting the robustness of nodes above the family level, and the ongoing challenges to reconstruct portions of the tree generated by rapid and recent radiations. We will compare the use of nuclear supermatrices and 'species tree' methods to address different portions of the carnivoran phylogeny, and also compare the performance of nuclear markers and mitochondrial DNA segments. We discuss the effects of taxon and character sampling, and also illustrate the impacts of important challenges to the accurate reconstruction of recent radiations, including incomplete lineage sorting and inter-species hybridization. If challenges are still present in such a well-known group, what is to be expected in other portions of the Tree of Life? We illustrate this situation with another case study, focusing on ciliates (phylum Ciliophora), and more specifically on a subclass (Peritrichia) which has so far received little attention in terms of molecular phylogenetics. Although some molecular phylogenies have been produced, they rely on a single genomic region (18S rDNA), and taxon sampling is far from complete. Still, initial studies have revealed widespread occurrence of non-monophyly at the level of genera and families, prompting the need for taxonomic revision in this group. As an illustration of the magnitude of the challenges ahead, our initial molecular analyses of this group have indicated that two novel orders should be recognized, and that morphological characters so far considered to be diagnostic for one genus are in fact plesiomorphies retained in both of these ordinal-level clades. By comparing these two disparate clades (Carnivora and Peritrichia), we will draw conclusions regarding the overall goal of reconstructing the eukaryotic Tree of Life.

O-77

Diagnosing phylogenetic conflict among genes and proteins: evidence for the origin of plastids and the early diversification of plants.

Blaise Li[1], Peter Foster[3], Martin Embley[2], Cymon Cox[1]
[1]Centro de Ciencias do Mar (CCMAR), Faro, Portugal, [2]Newcastle University, Newcastle, UK, [3]Natural History Museum, London, UK

Plants (Archeaplastida) evolved from a heterotrophic eukaryotic ancestor approximately 1.5 Gya following the endosymbiotic capture of a cyanobacterium and its subsequent cellular integration as a photosynthetic organelle, or plastid. Such ancient events are especially challenging for phylogenetic reconstruction and, in the case of plant evolution, has resulted in a number of contradictory hypotheses concerning the origin and early history of the plant lineage. Here we present analyses of cyanobacterial and plastid data aimed at identifying and reconciling sources of phylogenetic conflict concerning the identity of the closest extant cyanobacterium to the plastids and the relationships among the glaucocystophytes, and the green and red lineages of plants. We obtained strongly conflicting phylogenies using 75 plastid (or nuclear plastid-targeted) genes and their direct translations to proteins. The conflict between genes and proteins is largely robust to the use of models accounting for composition heterogeneities across the data or across the tree, site-stripping techniques based on site variability, and data-recoding strategies. We identify a composition bias and concomitant codon-usage bias resulting from synonymous substitutions at third codon positions, and crucially also at the first positions of Leucine and Arginine codons, as the source of the conflict. Although the removal of third codon positions is a common strategy in analyses of protein-coding genes, our analyses demonstrate that failing to account for composition biases caused by synonymous substitutions in first positions of Leucine and Arginine codons - which in our data occur at more than a hundred times the median non-synonymous substitution rates - can also lead to a systematically biased tree. Consequently, we argue that our gene data analyses using conventional modes of analysis and character coding are likely misleading, and recommend that strategies for recoding substitutions among synonymous codons be explored when analysing protein-coding genes. Our analyses identify the closest extant taxon to the plastid lineage as a clade formed by all cyanobacteria present in our data, except those belonging to the GBACT group, and also suggest that the glaucocystophytes are most closely related to the red algal lineage.

O-78

Horizontal Gene Transfer (HGT), Homeoalleles, NISEs, and the Formation and Maintenance of Higher Taxonomic Units in Bacteria and Archaea

J. Peter Gogarten[1], Cheryl P. Andam[1,2], David Williams[1], Erica Lasek-Nesselquist[1], Seila Omer[1]
[1]*University of Connecticut, Storrs, Connecticut, USA,* [2]*Cornell University, Ithaca, New York, USA*

How are higher taxonomic units in microbial evolution created and maintained? Does the central tendency in the forest of gene trees correspond to a vertical signal created by shared ancestry? Does biased gene transfer contribute to this signal? Or could it be that the observed signal is created mainly through biased gene transfer? If gene transfer is biased towards similar organisms (those that we usually consider as related), then gene transfer will tend to reinforce the signal created through shared ancestry. Only if gene transfer bias is due to a shared ecological niche and creates a highway of gene sharing between unrelated organisms, does HGT and shared ancestry create conflicting signals, as was suggested for the extremely thermophilic bacteria [1,2].

Until recently the hypothesis that biased HGT contributed to the formation and maintenance of the central tendency observed in comparative genome analyses was difficult to test, because the predictions from this hypothesis were indistinguishable from the claim that the central tendency found in genome phylogenies is due to shared ancestry only. For a group held together by frequent within group gene transfer followed by homologous recombination in the recipient, the within group gene trees should conflict [3]; however, if gene transfer is biased towards close relatives, the phylogenetic conflict created through transfer frequently is not significantly supported, and the compositional bias between donor and recipient is too similar to allow for the detection of transfer events.

This state of affairs changed with the discovery of homeoalleles in aaRS phylogenies [4,5]. Homeoalleles were defined as divergent, largely isofunctional homologs that frequently replace one another. Homeoalleles are similar to alleles in a population; however, all members of a population or species usually possess the same type of homeoallele, and the replacements occur through transfer from a related group of organisms. Only the replacement of one type of homeoallele with the other reveals the transfer – within each type phylogenetic conflicts to the reference phylogeny are not strongly supported.

The presentation will discuss inferences from the analysis of replacement between homeoalleles and between non-homologous isofunctional enzymes (NISEs)[6].

1. Boussau *et al BMC Evol Biol* 2008, **8**:272.

2. Zhaxybayeva *et al*: *PNAS* 2009, **106**:5865.

3. Dykhuizen & Green *JBact* 1991, **173**:7257.

4. Andam *et al: PNAS* 2010, **107**(23):10679.

5. Andam & Gogarten *Nature RevMicr* 2011, **9**(7):543-555.

6. Omelchenko *et al: Biology direct* 2010, **5**:31.

This research was supported through NSF DEB 0830024

O-79

Nested Phylogenetic Reconstruction: scalable resolution in large phylogenies

Jaime Huerta-Cepas, Marina Marcet-Houben, Toni Gabaldón
Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain

Abstract A pressing challenge in phylogenetics is the need to cope with the massive production of complete genomic sequences, especially after recent technological developments. Problems that are particularly affected by the increasing flow of genomic data and that require continuous update are: i) the establishment of evolutionary relationships between species (the so-called Tree Of Life (TOL)) and ii) the study of the evolution of large, widespread super-families that evolved through complex patterns of duplications and losses. Regardless of the complexity and the number of tasks involved in the classic phylogenetic pipeline, current methods undertake phylogenetic reconstruction as a single step procedure, considering the whole set of species or homologous sequences as an indivisible dataset.

We propose here the Nested Phylogenetic Reconstruction (NPR) approach, which can be used to reconstruct large phylogenies by means of an iterative strategy that provides scalable and sustained resolution in all tree nodes. Using the NPR strategy, all internal nodes in a tree can be automatically re-evaluated in order to optimize and fine tune the parameters of the phylogenetic pipeline applied to different levels of the same tree. To illustrate its use, we apply NPR to the reconstruction of a highly resolved eukaryotic TOL using a total of 216 fully sequenced species. Positive effects of the increased gene sampling at each iteration are clear, both in terms of balanced functional representation and of accuracy, as judged from the overall agreement with taxonomic classifications and established relationships. The final topology is highly resolved, with all but 6 nodes in the tree receiving the highest statistical support. Agreement with the established taxonomic divisions is remarkably high, considering our completely automated and unsupervised approach. Finally, we present also a command line application to apply the NPR approach to large gene families.

O-80

Genes, genomes and efficacy in discerning dimensionalities of the tree of life

Khidir Hilu, Sunny Crawley
Virginia Tech, Blacksburg, Virginia, USA

A surge in the number of organelle genome sequences, spurred by the decline in sequencing cost and technological advances, has provided an abundance of molecular characters for phylogenetic reconstruction. Such types of datasets are thought to provide close approximation of species phylogenies. However, phylogenetic reconstruction based on whole genome matrices is generally strongly skewed toward relatively few representative taxa. In contrast, datasets based on a few genomic regions, although promoting a denser taxon sampling, are construed as potentially too character-poor to accurately resolve phylogenetic histories. These approaches are seemingly mutually exclusive due to the extraordinarily large amount of missing data that may result from combining them. Using the angiosperm order Caryophyllalles as a taxonomic platform we contrast the relative efficacy in phylogenetic reconstruction of whole genome/narrow taxon representation vs. few genomic regions/denser taxon sampling, and explore the potential of combining these two types of datasets, allowing for the inherently large proportion of missing data. Our preliminary results demonstrate that a prudent selection of limited number of genomic regions can provide a phylogenetic tree that approaches the genome-based tree with the added benefit of considerable divergence details at recent histories. Phylogenetic signal from the combined dataset recovered the backbone and provided valuable information on the terminal leaves despite extensive punctuation in missing characters/genomic regions. Such a supermatrix may provide a cost- and time-effective platform for discerning the dimensionality of the tree of life.

O-81

Evolutionary contrasts of somatic and imprinted X chromosome inactivation

Andrew Clark, Xu Wang
Cornell University, Ithaca, NY, USA

Dosage compensation in somatic tissues of mammals is achieved by the well studied mechanism of X-chromosome inactivation. There is however wide variation in the regulation of X-linked gene expression in placental tissue of embryonic origin (trophoblast tissue). In the mouse trophobast, the paternal X remains inactivated, a phenomenon known as imprinted X inactivation. In the horse trophoblast tissue, we show that there is random X-inactivation, just like in somatic tissue. Both systems of X inactivation have genes that escape the inactivated state, and it is of great interest to understand the evolutionary constraints on these genes (e.g. haploinsufficiency), or whether the escaper status is in any way adaptive. Analysis of extensive sets of RNA-seq runs from reciprocal crosses of mouse, horse/donkey, and opossum identify not only the inactivation status, but also the complete collection of escaper genes. It is clear that the imprinted X inactivation escaper genes are a distinct set from the random X inactivation escapers, suggesting that the mechanisms of inactivation and escape are independent. Site frequency spectra of genes that do and do not escape X inactivation might be expected to differ, and we examine this issue as well.

O-82

Rapid *De Novo* Evolution of X Chromosome Dosage Compensation in *Silene latifolia*, a Plant with Young Sex Chromosomes

Aline Muyle[1], Niklaus Zemp[2], Clothilde Deschamps[3], Sylvain Mousset[1], Alex Widmer[2], Gabriel Marais[1]
[1]CNRS / Université Lyon 1, Villeurbanne, France, [2]ETH Zurich, Zürich, Switzerland, [3]Pôle Rhône-Alpes de Bioiformatique, Villeurbanne, France

Silene latifolia **is a dioecious plant with heteromorphic sex chromosomes that have originated only ~10 MYA and is a promising model organism to study sex chromosome evolution in plants. Previous work suggests that *S. latifolia* XY chromosomes have gradually stopped recombining and the Y chromosome is undergoing degeneration as in animal sex chromosomes. However, this work has been limited by the paucity of sex-linked genes available. Here, we used 35 Gb of RNA-seq data from multiple males (XY) and females (XX) of a *S. latifolia* inbred line to detect sex-linked SNPs and identified more than 1700 sex-linked contigs (with X-linked and Y-linked alleles). Analyses using known sex-linked and autosomal genes, together with simulations indicate that these newly identified sex-linked contigs are reliable. Using read numbers, we then estimated expression levels of X-linked and Y-linked alleles in males and found an overall trend of reduced expression of Y-linked alleles, consistent with a widespread ongoing degeneration of the *S. latifolia* Y chromosome. By comparing expression intensities of X-linked alleles in males and females, we found that X-linked allele expression increases as Y-linked allele expression decreases in males, which makes expression of sex-linked contigs similar in both sexes. This phenomenon is known as dosage compensation and has so far only been observed in evolutionary old animal sex chromosome systems. Our results suggest that dosage compensation has evolved in plants and that it can quickly evolve *de novo* after the origin of sex chromosomes.**

O-83

Recombination in the Human Pseudo-Autosomal Region (PAR1)

Anjali Hinch[1], Arti Tandon[2], David Reich[2], Simon Myers[1]
[1]Oxford University, Oxford, UK,[2]Harvard Medical School, Boston, MA, USA

The pseudo-autosomal region (PAR1) is a 2.7 Mb region that has seen rapid evolution in mammalian genomes[1]. It is a short region of homology between the X and Y chromosomes. A cross-over in PAR1 is essential for the proper disjunction of X and Y chromosomes in male meiosis[2] and its deletion results in male sterility[3]. This leads to PAR1 having a male cross-over rate 18-fold higher than the genome average. It is not known how this high rate of recombination is achieved biologically. Recent research in mice[4] suggests the existence of PAR-specific recombination machinery responsible for programmed double-strand breaks in male meiosis.

Despite its importance, no fine-scale pedigree-based maps are available for the PAR[5], while the map based on breakdown of linkage disequilibrium (LD)[6] is unreliable due to little LD present in the PAR. We have leveraged genetic and next-generation sequencing data to produce the first detailed genetic maps available for the PAR. Using sex-specific maps we built using 137 African-American pedigrees, we find that PAR1, similar to the autosomes, shows considerable fine-scale variation in rates. Males have intense recombination throughout PAR1, while in females recombination is concentrated near the Pseudo-Autosomal boundary. Further, we have used sequence data from the 1000 genomes project[8] to map ancestry tracts and identify crossovers in 22,000 unrelated African Americans. In contrast with the autosomes where nearly all hotspot locations are determined by the gene PRDM9 [7,9,10,11], we find that cross-over locations in PAR1 are not associated with PRDM9. Our results support the existence of an independent recombination machinery for the PAR. This work will enable us to understand the role of recombination in the sequence evolution of the PAR in the human lineage.

[1] Graves J. *et al.* Human Molecular Genetics. 7(13), 1991–1996 (1998).

[2] Rouyer F. *et al.* Nature 319, 291–295 (1986).

[3] Mohandas T.K. *et al*. American Journal of Human Genetics. 51(3), 526-533 (1992).

[4] Kauppi L. *et al*. Science 331, 916-920 (2011).

[5] Flaquer A. et al. European Journal of Human Genetics. 16, 771-779 (2008).

[6] Myers, S. *et al*. Science 310, 321–324 (2005).

[7] Hinch A.G. *et al*. Nature 476, 170–175 (2011).

[8] The 1000 Genomes Project Consortium. Nature 467, 1061-1073 (2010)

[9] Myers, S. *et al*. Science 327, 876–879 (2010).

[10] Baudat, F. *et al*. Science 327, 836–840 (2010).

[11] Parvanov, E.D. *et al*. Science 327, 835 (2010).

O-84

The functional evolution of mammalian Y chromosomes

Diego Cortez[1], Laure Froidevaux[1], Angélica Liechti[1], Frank Grützner[2], Henrik Kaessmann[1]
[1]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland, [2]The Robinson Institute, School of Molecular and Biomedical Science, University of Adelaide, Adelaide, Australia

The Y chromosome is specific to males and plays a dominant role in (male) sex determination and fertility in mammals. However, due to the overall highly repetitive nature of Y chromosomes, resulting from differentiation processes that occurred since sex chromosome origination, complete genomic sequences of only a few mammalian Y chromosomes have been determined so far. Thus, the evolution of mammalian Y chromosomes remains little understood. Using extensive RNA sequencing data for somatic and germline tissues from males and females, we have reconstructed Y chromosomal transcriptomes of eight mammals (human, gorilla, orangutan, macaque, marmoset, mouse, opossum, platypus) that represent all major mammalian lineages and sex determination systems. Y chromosome transcripts were validated using a large-scale PCR and resequencing approach as well as fluorescent in situ hybridization. Our analyses of these data provide fundamental novel insights into the functional evolution of mammalian Y chromosomes and the associated selective forces.

O-85

Phylogenomics, and microRNAs congruently resolve the ecdysozoan phylogeny

Davide Pisani[1], Omar Rota-Stabelli[2], Lahcen Campbell[1]
[1]The National University of Ireland Maynooth, Maynooth, Ireland, [2]IASMA Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy

Ecdysozoa, the group including Nematoda (roundorms) and the Arthropoda (e.g. spiders, centipedes, insects and crustaceans), represents include the largest majority of Earth biodiversity and biomass. The monophyly of this assemblage of phyla as long been debated and has only recently been confirmed by the analyses of complete animal genomes. However, the relationships within this group remain uncertain. We assembled large-scale phylogenomic data sets, and identified the nearly complete microRNA repertoire (also sequencing small RNA libraries) for a representative sample of ecdysozoan species. These independent genomic-scale data sets were analysed and the relationships among the Ecdysoza resolved by means of congruence. We were able to show that of the two generally accepted ecdysozoan clades (the Panarthropoda and the Cycloneuralians) only the first, which includes the Arthropoda, the Tardigrada (water bears) and the Onychophora (velvet worms), is monophyletic. The second, Cycloneuralia, most likely represents a paraphyletic assemblage of phyla. In addition, we were able to show that within Panarthropoda, the velvet worms represent the sister group of the Arthropoda with the water bears representing the sister group of the Arthropoda plus Onychophora. Our results substantially clarify the relationships among the Ecdysozoa and provide an invaluable framework to understand evolutionary patterns (both genomic and morphological) within this group.

O-86

Pathways to the Amphibian Tree of Life

Mark Wilkinson[1], Karen Siu Ting[1]
[1]The Natural History Museum, London, UK,[2]National University of Ireland, Maynooth, Ireland

We review the growth of understanding of amphibian phylogeny and ask what this can tell us about how best to construct a comprehensive phylogeny of the group. The living amphibians (Class Amphibia) comprise three main groups , frogs (Order Anura), salamanders (Order Caudata) and caecilians (Order Gymnophiona) each of which are well-supported monophyla as is the Batrachia - the group ,comprising frogs and salamanders. Amphibians are subject to current concerns over population declines, emerging diseases and extinction and phylogenetics has been used in determining conservation priorities. The group is also experiencing sustained rapid expansion through the discovery of new taxa. Previous phylogenetic wok (and much progress) has come from studies that are limited in taxonomic scope and which focus on particular sub-groups (e.g. Orders or other well supported monophyla). Combining this prior phylogenetic knowledge provides one possible pathway to the Amphibian Tree of Life. An alternative pathway is through de novo phylogenetic analyses of data sets of less limited taxonomic scope. We illustrate some of the potential issues raised by these different approaches.

O-87

1000 Insect Transcriptomes - the 1KITE project takes the next step in insect phylogenomics

Karen Meusemann[1], Ralph S. Peters[2], Karl M. Kjer[3], Xin Zhou[4], Bernhard Misof[1]
[1]Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für molekulare Biodiversitätsforschung (zmb), Bonn, Germany, [2]Zoologisches Forschungsmuseum Alexander Koenig, Abteilung Arthropoda, Bonn, Germany, [3]Rutgers University, Department of Ecology, Evolution and Natural Resources, New Brunswick, NJ, USA, [4]Beijing Genomics Institute (BGI), Shenzhen, China

Insects are the most species-rich group of metazoan organisms. To infer insect phylogeny, molecular sequence data have become an indispensable tool. However, phylogenomic molecular sequence data of considerable extent are only available for a small number of mostly holometabolous insect groups (e.g., Lepidoptera, Hymenoptera, Diptera) and for few model organisms (e.g., human body louse, fruit fly).

1KITE (1K Insect Transcriptome Evolution), an international research project started in 2011, aims to study the transcriptomes of 1,000 insect species encompassing all recognized insect orders. The molecular sequence data (Expressed sequence Tags, ESTs) obtained using next generation sequencing (NSG) techniques will allow inferring for the first time a robust phylogenetic backbone "insect tree of life". This backbone tree will help to unraveling the evolution of insects which is essential for understanding how life in terrestrial and limnic environments evolved. Preliminary analyses of already available sequences show that the obtained data are of yet unparalleled size and quality. We aim to generate a 1,000-gene data set with a low proportion of missing data.
Scientists involved include not only experts in molecular biology, but also experts in insect morphology, taxonomy, paleontology, embryology, bioinformatics, and scientific computing. Overall, scientists from eight nations (Australia, Austria, China, Germany, Japan, Mexico, New Zealand, and the US) are collaborating in 1KITE. All transcriptome sequence data will be collected by the Beijing Genomics Institute (BGI), Shenzhen, China. 1KITE is divided into several subprojects, each of which focusing on specific phylogenetic groups, for example apterygote hexapods, Odonata, Dictyoptera, Neuropterida, Coleoptera, and Antliophora. Additionally, 1KITE includes the development of new software for data quality assessment, phylogenetic reconstruction, and molecular dating that will allow for advanced and accelerated analyses of such large amounts of sequence data. Sequencing is scheduled to be completed by the end of 2012. The results are expected to have an extraordinary and long-standing high impact on entomological and phylogenetic research.
Further information: www.1kite.org.

O-88

Conservation and Adaptation in Blindness Genes: A Phylomedicine Approach to Understanding the Evolution of Sensory Perception in Mammals

Emma Teeling[1], Michael Bekeart[2], Alisha Goodbla[1]
[1]*University College Dublin, Dublin, Ireland,* [2]*University of Stirling, Scotland, UK*

In 2009, the World Health Organisation estimated that 314 million people worldwide are visually impaired and 45 million of them are blind. Only 10% of these people know their underlying genetic disorder. Although 50-90% of the causal genes have been identified, the majority of disease causing single-nucleotide polymorphisms (SNPs) and crucial genomic regions that lead to inherited blindness are still not known . Using a phylomedicine approach we investigated the molecular evolution of vision across mammals focusing on the sensory specialists, the bats. We designed universal primer pairs for over 33 genes involved in non-syndromic blindness in humans; amplified and sequenced these genes in bats and other mammals (69 mammals,~ 28,000bp); and, examined if these genes or regions of these genes are under selection using phylogenetic methods and multiple selection tests. We located known disease SNPS sites within our alignments and located regions of high evolutionary conservation and divergence. We found evidence for positive selection occurring across mammals in 12 visual genes with different genes showing evidence for clade specific selection. Within bats, genes involved in night vision pathways showed significantly more evidence of positive selection than other visual functions. We identified regions of evolutionary conservation (Regions of Extreme Purifying Selection, REPS) across all genes that are essential for correct mammalian visual function. As predicted the disease associated H. sapiens mutations (DAMS) mostly occur in the REPSs (594 REPSs identified covered 1,979 nt of the 28,798 nt of the 34 genes fragment alignments, 105 SNPs are in REPS, 62 are outside, $P<0.001$). Hence, identifying these REPS can predict the cause of genetic pathology in H. sapiens and therefore, provide new potential causative sites for blindness in man. We located REPs where no DAMS have yet been reported and predict that these regions are important for visual function, therefore, are potential targets for disease diagnostics. We also identified nucleotide positions in these genes that are important for divergent function of visual capabilities in mammals.

O-89

Phylomedicine: Evolutionary Diagnosis of Variants in Personal Exomes

Sudhir Kumar
Center for Evolutionary Medicine & Informatics, Arizona State University,, Tempe, AZ 85287, USA

Every personal exome contains thousands of nonsynonymous variants (nSNVs) of unknown effect, which compels the use of computational tools to gauge the potential of these alleles in altering protein function. Because of increasing sophistication and prediction accuracies of available tools, their application is now common in prioritizing nSNVs for laboratory investigations and assessing them in clinical genomic profiles. I will discuss the power and pitfalls of current computational tools for assessing the neutrality of nSNVs. I will present a novel evolutionary diagnosis (EvoD) method that remedies some of the major pitfalls of current tools by employing a purely phylogenetic approach. Results from the application of EvoD to about quarter million nSNVs from the 1000 genomes data will be presented, which demonstrates the usefulness of historical molecular evolutionary patterns in predicting the neutrality of novel protein variants.

O-90

Rooted Phylogenetic Networks From Real Trees

Daniel Huson
University of Tuebingen, Tuebingen, Germany

Phylogenetic networks promise to facilitate the explicit representation of reticulate evolutionary events, such as hybridization,

horizontal gene transfer, recombination, reassortment or incomplete lineage sorting. While a number of methods

for computing unrooted phylogenetic networks are available and commonly used, robust and widely-used methods for calculating

rooted phylogenetic networks are only beginning to emerge. In this talk we will look at some recent advances in this area and

will describe a new method for computing a rooted phylogenetic network from two ``real'' phylogenetic trees, that is, from

phylogenetic trees as they occur in most studies, namely with multifurcations, missing taxa and unrooted. This is based on joint work

with Simone Linz.

O-91

The probability of gene trees in the presence of hybridization

Yun Yu[1], James Degnan[2], Luay Nakhleh[1]
[1]Rice University, Houston, Texas, USA, [2]University of Canterbury, Christchurch, New Zealand

Gene tree probabilities allow for inferring evolutionary histories from multi-locus data in a probabilistic manner. The coalescent has been used extensively to derive these probabilities in a population, and the multi-species coalescent provides a generalization that enables computing gene tree probabilities within the branches of a phylogenetic tree. When reticulate evolutionary events, such as hybridization, occur, gene tree probability computation must be extended to incorporate the multiple evolutionary trajectories that a gene could have followed. Here, I will describe our recent work on developing new algorithms for computing gene tree probabilities within the branches of a phylogenetic network. These algorithms can be used as the engine of a tool for inferring reticulation evolutionary histories, while simultaneously accounting for hybridization and incomplete lineage sorting as two causes of gene tree incongruence.

O-92

Fast computation of minimum hybridization networks

Celine Scornavacca[1], Benjamin Albrecht[4], Alberto Cenci[3], Daniel H. Huson[2]
[1]CNRS, ISEM, Montpellier, France, [2]ZBIT, Tübingen University, Tübingen, Germany, [3]IRD, Montpellier, France, [4]LMU, München, Germany

Hybridization events in evolution may lead to incongruent gene trees. One approach to determining possible interspecific hybridization events is to compute a hybridization network that attempts to reconcile incongruent gene trees using a minimum number of hybridization events. We describe how to compute a representative set of minimum hybridization networks for two given bifurcating input trees, using a parallel algorithm and provide a user-friendly implementation. A simulation study suggests that our program performs significantly better than existing software on biologically relevant data. Finally, we demonstrate the application of such methods in the context of the evolution of the Aegilops/Triticum genera. The algorithm is implemented in the program Dendroscope 3, which is freely available from www.dendroscope.org and runs on all three major operating systems.

O-93

Phylogenomic Networks

Tal Dagan
Heinrich-Heine University Düsseldorf, Duesseldorf, Germany

Phylogenomics is aimed at studying functional and evolutionary aspects of genome biology using phylogenetic analysis of whole genomes. Current approaches to genome phylogenies are commonly founded in terms of phylogenetic trees. However, several evolutionary processes are non tree-like in nature, including recombination and lateral gene transfer (LGT). Phylogenomic networks are a special type of phylogenetic networks reconstructed from fully sequenced genomes. The network model, comprising genomes connected by pairwise evolutionary relations, enables the reconstruction of both vertical and LGT events. Modeling genome evolution in the form of a network enables the use of an extensive toolbox developed for network research. The structural properties of phylogenomic networks open up fundamentally new insights into genome evolution.

O-94

Markov models for phylogenetic networks

Barbara Holland, Jeremy Sumner, Peter Jarvis, Jonathan Mitchell
University of Tasmania, Hobart, Tasmania, Australia

We show how to express the two-state continuous-time general Markov model on trees in such a way that allows its extension to more general network models. In this framework we can model convergence of subsets taxa as well as divergence (speciation). This could be highly applicable to situations where species boundaries are incomplete. However, these exciting possibilities must be tempered by analysis of the identifiability (or otherwise) of the models that arise by taking more general networks.

O-95

Demographic inference from the site frequency spectrum with ascertainment bias

Laurent Excoffier[1]
[1]*University of Berne, Berne, Switzerland,* [2]*Swiss Institute of Bioinformatics, Lausanne, Switzerland*

The site frequency spectrum (SFS) summarizes the distribution of allele frequencies within and between populations, and is therefore informative about past demographic and selective events. Several composite-likelihood methods have been proposed to reconstruct demographic history from the SFS, but their accuracy has not been fully evaluated. We introduce here a very flexible simulation-based method to estimate demographic parameters from SFS derived from genomic data, and we examine its accuracy relative to a diffusion-based method (dadi) under several scenarios involving one to three populations. We also check the ability of our simulation-based approach to estimate parameters when more than 3 populations are involved. We finally introduce the possibility to infer parameters from SNP chip data under a known ascertainment scheme.

O-96

Mining human population-genomic data for demographic inference

Mattias Jakobsson
Uppsala University, Uppsala, Sweden

Population genomic data is accumulating for a multitude of species from various new technologies. For model species, including humans, rich population genomic data have become available, for example from SNP-chip panels that target increasingly larger sets of SNPs. These data contain vast amounts of information that can be mined in concert with new generation sequencing data. We explore possibilities, methods and tools for making inference from new generation sequencing data together with population-genomic data from SNP-chips. For example, using simulations, we evaluate the performance of the Approximate Bayesian Computation (ABC) approach in the case where the data consists of large-scale population-genomic data. We show that the ABC approach can accommodate realistic genome-wide population genetic data, which may be difficult to analyze with full likelihood approaches. Furthermore, we investigate the impact of serial founder models, archaic admixture and ascertainment bias on patterns of allele sharing, inference of time to most recent common ancestor, and inference tools such as principal component analysis.

O-97

Robust identification of local adaptation to environmental gradients with an application to *Arabidopsis thaliana*.

Torsten Günther[1], Christian Lampei[1], Oliver Simon[1], Karl J. Schmid[1], Graham M. Coop[2]
[1]*University of Hohenheim, Stuttgart, Germany,* [2]*University of California, Davis, USA*

Identifying the genetic changes that are responsible for local adaptation is a major challenge in population genetics. We can hope to identify such loci from strong correlations between allele frequencies and relevant ecological or environmental variables. However, these analyses are complicated by an imperfect knowledge of population allele frequencies and neutral correlations of allele frequencies among populations by population history and gene flow. The former confounding effect is particularly strong in next-generation sequencing (NGS) data, where sample coverage varies dramatically across the genome, a problem further amplified in cost effective pooled sequencing studies. We extend a Bayesian method (Bayenv) that takes both various levels of sampling noise and relationship between populations into account while looking for correlations between allele frequency and environmental factors. Under our null hypothesis, the covariance of allele frequencies among populations is accounted for using a hierarchical Bayesian model where the transformed population frequencies are multivariate normal. Then for each individual SNP, the null model is tested against an alternative model where the environmental variable has a linear effect on allele frequencies, the support for this alternative model is provided in form of a Bayes Factor. The Bayes Factor as well as all other parameters are estimated using a Markov chain Monte Carlo (MCMC). As the linear model may not always be biologically realistic, we also generate a set of transformed allele frequencies, where the effect of unequal sampling variance and covariance among populations has been removed. This allows users a general framework to utilize non-parameteric statistics to investigate various environmental correlations. We illustrate the utility of the method by identify putative targets of selection in pooled NGS data from *Arabidopsis thaliana* populations collected along a strong environmental gradient.

O-98

Dissecting the Divergence of Dogs from Their Wild Ancestors: Complete Genome Sequences of 5 Close Relatives to Modern Dog Breeds

Adam Freedman[1], Rena Schweizer[1], Pedro Silva[2], Marco Galaverni[3], Zhenxin Fan[4], Eunjung Han[1], Diego Ortega-Del Vecchyo[1], Ilan Gronau[5], Belen Lorente-Galdos[6], Can Alkan[7], Kevin Squire[1], Robert Chin[1], Adam Boyko[5], Zugen Chen[1], Heidi Parker[8], Christina Chung[9], Clarence Lee[9], Adam Siepel[5], Tomas Marques-Bonet[6], Carlos Bustamante[10], Elaine Ostrander[8], Timothy Harkins[9], Stanley Nelson[1], Robert Wayne[1], John Novembre[1]
[1]University of California, Los Angeles, Los Angeles, California, USA, [2]Universidade do Porto, Vairão, Portugal, [3]University of Bologna, Bologna, Italy, [4]Sichuan University, Chengdu, China, [5]Cornell University, Ithaca, New York, USA, [6]Institut de Biologia Evolutiva, Barcelona, Spain, [7]University of Washington, Seattle, Washington, USA, [8]National Institutes of Health/NHGRI, Bethesda, Maryland, USA, [9]Life Technologies, Foster City, California, USA, [10]Stanford School of Medicine, Stanford, California, USA

The domestication of dogs from wild canid ancestors has resulted in one of the most striking and rapid episodes of phenotypic divergence in evolutionary history. To provide more insight on the genetics of early dog domestication and a resource for future canid genomics studies, we have sequenced 5 novel canid genomes to greater than 20x coverage: 2 gray wolves, 1 golden jackal, and 2 divergent dog breeds, the Basenji and Dingo. Given the recent divergence involved, we have aligned all sequences with the published genome of the Boxer (7.5x Sanger) and identified variants. The sequences of the 2 gray wolves provide the first genomes of the closest extant relative of the domestic dog and thus provide a unique perspective on domestic dog-derived changes. Using the sequence of the golden jackal (a close outgroup to dogs and wolves) we assign observed wolf-dog variants to the dog or wolf lineage, and identify regions that have had an excess of variants private to the dog lineage and thus may play a role in the genetic basis of domestication traits. We present four major findings. First, using recently developed coalescent-based approaches, we estimate the timing of divergence between dogs and wolves, and reveal periods of both congruence and divergence in population size change between wild and domestic canids. Second, we present results indicating that post-divergence gene flow took place between dogs and wolves early on in the domestication process with particular wolf populations. Third, we show the extent to which selection has reduced diversity around nonsynonymous substitutions on the dog lineage. And, fourth, using a combination of selection scan approaches, we identify the loci underlying signals of selection on the dog lineage, and discuss how they inform our understanding of the genetic architecture underlying the rapid phenotypic divergence of dogs from wolves.

O-99

Parallel evolution of Mycobacterium tuberculosis and modern humans

Sebastien Gagneux
Swiss Tropical and Public Health Institute, Basel, Switzerland

Mycobacterium tuberculosis (Mtb) is an obligate pathogen of humans and the most important cause of human tuberculosis (TB). Left untreated, TB kills up to 50% of patients, and despite available treatment, close to 2 million people die of TB each year. TB has been regarded as a typical "crowd disease", thought to have originally emerged following the initiation of animal domestication 10,000 years ago. Recent comparative genomic data however indicate that Mtb originated as a human pathogen in Africa, suggesting that human TB might predate the Neolithic Revolution.

We resequenced the genomes of 220 clinical isolates representative of Mtb's global phylogeographic diversity. Comparison to a corresponding set of 423 human mitochondrial genomes (mtDNAs) revealed a striking congruence between the Mtb and mtDNA phylogenies. We used the point of phylogenetic separation between the exclusively African mitochondrial lineages and the mtDNAs associated with the Out-of-Africa migration of modern humans dated at 70,000 years ago to calibrate our Mtb phylogeny. This allowed us to define divergence times in the Mtb tree which were strikingly similar to previous estimates using human DNA. In particular, our Mtb data support the "multiple dispersal model" of modern humans out of Africa, with a first migration along the Indian Ocean around 65,000 years ago, followed by second wave into Eurasia around 40,000 years ago. Moreover, calculation of the effective population size revealed an almost perfect correlation between the values predicted for Mtb and human mtDNAs, with both measures exhibiting a strong increase around 8-10,000 years ago.

These data suggest that in contrast to other typical "crowd diseases", Mtb has co-existed with modern humans for tens of thousands of years, most of the time in a quasi-commensal relationship and with no obvious impact on human population size. We hypothesize that the high TB mortality observed in 18th-19th century Europe, which continues in many parts of the developing word to date, represents a comparably recent phenomenon linked to the massive and unprecedented human population expansions of the last two centuries.

O-100

Detection of recombination events in bacterial genomes from large population samples

Jukka Corander
University of Helsinki, Helsinki, Finland

Analysis of important human pathogen populations is currently under transition towards whole-genome sequencing of growing numbers of samples collected on a global scale. Since recombination in bacteria is often an important factor shaping their evolution by enabling resistance elements and virulence traits to rapidly transfer from one evolutionary lineage to another, it is highly beneficial to have access to tools that can detect recombination events from whole-genome sequence data for bacterial population samples on a large-scale. We discuss a recently introduced Bayesian approach that can efficiently handle hundreds of whole genome sequenced population samples and identify separate origins of the recombinant sequence. The necessary steps towards making the analysis of thousands, or even tens of thousands of genomes feasible will also be considered.

O-101

Genomic sequencing of *Plasmodium falciparum* malaria parasites from Senegal reveals the demographic history of the population

Hsiao-Han Chang[1], Daniel Park[1,2], Kevin Galinsky[2], Stephen Schaffner[2], Daouda Ndiaye[3], Omar Ndir[3], Soulyemane Mboup[3], Roger Wiegand[2], Sarah Volkman[2,4], Pardis Sabeti[1,2], Dyann Wirth[2,4], Daniel Neafsey[2], Daniel Hartl[1]
[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA, [2]Broad Institute, Cambridge, MA, USA, [3]Cheikh Anta Diop University, Dakar, Senegal, [4]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA, USA

Malaria is a deadly disease that causes nearly one million deaths each year. It is important to understand the genetic basis of *P. falciparum* adaptations to antimalarial treatments and the human immune system while taking into account its demographic history in order to develop methods to control and eradicate malaria. To study the demographic history and identify genes under selection more efficiently, we sequenced the complete genomes of 25 culture-adapted *P. falciparum* isolates from three sites in Senegal. We show that there is no significant population structure among these Senegal sampling sites. By fitting demographic models to the synonymous allele-frequency spectrum, we also estimated a major 60-fold population expansion of this parasite population approximately 20,000–40,000 years ago. Using inferred demographic history as a null model for coalescent simulation, we identified candidate genes under selection, including genes identified before, such as *pfcrt* and *PfAMA1*, as well as new candidate genes. Interestingly, we also found selection against G/C-to-A/T changes that offsets the large mutational bias toward A/T, and similar synonymous and nonsynonymous allele-frequency spectra and high nonsynonymous polymorphism that suggest pervasive diversifying selection in the genome.

O-102

Population genomic evidence for widespread selection on symbiosis-related genes in the facultatively mutualistic rhizobia Sinorhizobium meliloti and S. medicae

Brendan Epstein, Michael Sadowsky, Peter Tiffin
University of Minnesota, Saint Paul, MN, USA

The symbiosis between rhizobia bacteria and legume plants has served as a model for investigating the genetics of nitrogen fixation and the genetics and evolution of facultative mutualism. We used deep sequence coverage (> 100X) to characterize genomic diversity at the nucleotide level among 12 Sinorhizobium medicae and 32 S. meliloti strains. Although these species are closely related and even share host plants, based on the ratio of shared polymorphisms to fixed differences we found that horizontal gene transfer between these species was confined to the symplasmids and limited exclusively to genomic regions that harbor genes with direct functions in mutualism. Nucleotide diversity in S. meliloti is highly structured along the chromosome - with an entire half the chromosome (1.8 Mb) showing severely reduced diversity, consistent with extensive hitchhiking along with a selective sweep. Using frequency-spectrum based statistics we identified 101 genes that bear a signature of having evolved in response to recent positive selection. Thirty-six of the 41 named genes that were identified as targets of selection have characterized biochemical or phenotypic functions are functionally involved in the mutualism with plant hosts. The legume-rhizobia symbiosis first evolved ~ 60 million years ago and is often thought as being evolutionary stable; the large number of symbiosis related genes identified as targets of selection suggests that while this mutualism may be evolutionarily stable but is far from evolutionarily static.

O-103

Towards population genomics with Approximate Bayesian Computation

Daniel Wegmann
École polytechnique fédérale de Lausanne, Lausanne, Switzerland

Approximate Bayesian Computation (ABC) is an approach to numerically approximate the posterior distribution through simulations and is well suited for parameter inference of complex models for which analytical likelihoods are not available. While not limited to population genetic questions, it has been used to infer a wide range of demographic scenarios from genetic data, including bottlenecks, population splits and migration. However, ABC has rarely been applied to truly genome-wide data sets, despite the wealth of sophisticated software packages to simulate such data. This is mainly because the simulation of such large data sets is computationally prohibitive, particularly if complex genomic features such variable recombination and mutation rates or selection are considered. I will present and discuss several ideas on how to render ABC fit for genome-wide analysis such as ways to recycle simulations efficiently or to hybridize ABC with full-likelihood approaches. Such advances will bring us closer to a joint inference of both, the demographic and selective history of a species or population.

O-104

Dawg 2.0: New methods for sequence simulation.

Reed Cartwright[1,2]
[1]*Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, Arizona, USA,* [2]*Genomics, Evolution, and Bioinformatics, School of Life Sciences, Arizona State University, Tempe, Arizona, USA*

Simulations of sequence evolution are an important component of research in molecular evolution, comparative genomics, and bioinformatics. Existing simulation technologies cannot reliably replicate critical characteristics of real sequences. Dawg 2.0 will provide a framework to solve these problems. Currently the gold standard of simulating sequence evolution is using the Gillespie algorithm to process both point mutations and indels as a sequence evolves along a branch. Dawg 2.0 turns the Gillespie algorithm on its side and processes events as they happend along a sequence. Thus the new algorithm avoids complex bookkeeping and search trees, effeciently processing both indels and rate variation. Speed-ups of 1200% are seen for some tasks. Significant new features of Dawg 2.0 include support for nucloetide, protein, and codon evolutionary models, and model heterogenity along a tree and sequences.

O-105

From trajectories to averages: an improved description of the heterogeneity of substitution rates along lineages

Stephane Guindon[1,2]
[1]*University of Auckland, Auckland, New Zealand,* [2]*LIRMM, Montpellier, France*

The accuracy and precision of species divergence date estimation from molecular data strongly depend on the models describing the variations of substitution rates along a phylogeny. These models generally assume that rates randomly fluctuate along branches from one node to the next. However, for mathematical convenience, the stochasticity of such process is ignored when translating such rate trajectories into branch lengths. I show that this simplification has important consequences on the precision on the estimation of substitution rates, node ages and the autocorrelation of rates along a phylogeny. I also describe a new approach that explicitly considers the average substitution rates along branches as random quantities, resulting in a more realistic description of the time-dependent variations of evolutionary rates. The new method provides more precise estimates of the rate autocorrelation parameter as well as divergence date estimates. Altogether, this approach is a step forward to designing biologically relevant models of rate evolution that are well-suited to data sets with dense taxon sampling which are likely to present rate autocorrelation. The computer programme PhyTime, part of the PhyML package and implementing the new approach, is available from http://code.google.com/p/phyml

O-106

Estimating empirical codon mutation models from re-sequencing data

Carolin Kosiol, Nicola De Maio
Vetmeduni Vienna, Wien, Austria

Empirical models summarize the pressures and mutational patterns of sequence evolution from large data sets. In this talk, I will present how we take advantage of the new re-sequencing data to estimate empirical codon models (ECMs). In particular, I will present results for empirical Hidden Markov Models (ecHMMs) that allow for lineage and site specific variation on the 12 Drosophila Genome Phylogeny. Generally, we find that more complex models fit the data better. Thereby, models allowing for variation in rates, selective pressure and codon usage among sites as well as non-reversibile ECMs lead to the largest improvements. For ECMs estimated from different clades, we observe significantly different parameter estimates, showing that it is feasible and useful to estimate the empirical models for each genomic data set anew. Furthermore, it is noteworthy that we estimate 1%-3% multiple nucleotide substitutions (MNS), a proportion that is agreement with estimates derived with different methods. However, when applying ECMs to the problem of identifying positive selection with likelihood ratio tests the inclusion of MNS leads to a noticeable decrease of power of the tests.

We will also introduce a new model that simultaneously uses inter- and intraspecific data to estimate transition probabilities between polymorphic and fixed states. The calculation of the transition probabilities is based on an approximation of the multi-allele model by Kai Zeng (2010) and disentangles mutational and selective forces. Since the new model also allows for polymorphisms at internal nodes of the tree, we can account for incomplete lineage sorting. Preliminary results for the estimation of empirical polymorphism models from Drosophila melanogaster data of the Drosophila Population Genomic Project (DPGP) and the African Survey (DPGD2) will be discussed.

O-107

Efficiently and rapidly modeling the non-homogeneity of evolutionary processes

Mathieu Groussin, Laurent Gueguen, Bastien Boussau, Manolo Gouy
Lyon 1 University - Laboratoire de Biométrie et Biologie Évolutive, Lyon, France

The advent of Next-Generation Sequencing (NGS) techniques has produced a colossal amount of genetic data that can be used to investigate the evolution of species through phylogenomic approaches. Meanwhile, more and more sophisticated evolutionary models have been proposed to better capture the phylogenetic signal contained in biological sequences. Classically, these models assume the evolutionary process to be homogeneous, leading to phylogenetic reconstruction artifacts and inaccurate estimation of ancestral frequencies. Therefore, some new models focus on the non-homogeneity of the evolutionary process through time and/or sites. The rationale of the non-homogeneous approach lies in the fact that molecular compositions are not shared among lineages and that sites have very different evolutionary patterns due to local characteristics of the protein. Non-homogeneous models allow the equilibrium frequencies to vary over time, or among sites, or both. However, modeling the non-homogeneity of equilibrium frequencies usually requires the optimization of many free parameters, possibly reducing the efficiency of parameter space exploration and hugely increasing time computation. These reasons hampered the development of efficient branch and site-wise non-homogeneous models for proteins in the maximum likelihood framework. We propose here a new way to model the branch-wise non-homogeneity of the evolutionary process that considerably reduces the number of free parameters, with the use of a multivariate analysis that allows to explore a sub-space of the total variance present in the data. We use this branch-wise non-homogeneous model to modulate site-specific equilibrium frequencies, as in Blanquart and Lartillot, 2008. As a result, we provide the first maximum-likelihood implementation of a branch- and site-wise non-homogeneous model for protein sequences, and we present a few computing tricks that provide speed improvements without reducing accuracy. It then becomes feasible to use such branch- and site-wise non-homogeneous models in a very reasonable time.

O-108

The Protein Diversity Enhanced by Young Chimera in Rice Genome Through DNA-based and RNA-based Recombination

Chuanzhu Fan[1], Manyuan Long[2], Rod Wing[3]
[1]*Wayne State University, Detroit, MI, USA,* [2]*University of Chicago, Chicago, IL, USA,* [3]*University of Arizona, Tucson, AZ, USA*

By genomic pairwise comparison of 20Mb DNA sequence of chromosome 3 short arm (chr3s) in four rice species, *Oryza sativa* ssp. *japonica*, *O. glaberrima*, *O. punctata*, and *O. officinalis*, we searched for the cultivated rice *O. sativa* ssp. *japonica* lineage specific genes. Combining with the synonymous substitution rate tests and other evidence, we were able to identify potential recent duplicated genes evolved within 1 million years (Myr). We found as many as 25 *O. sativa* ssp. *japonica* specific functional young gene candidates that were likely originated after divergence between *O. sativa* and *O. glaberrima*, which accounts for around 1% of all annotated genes in chr3s. Among 25 young gene candidates, two recently duplicated segments contained 16 genes. Although the majority of these 25 young gene candidates were created by single or segmental DNA-based gene duplication and recombination, we also found two genes were originated through exon shuffling and RNA-based retroposition, respectively. Sequence divergence test between young genes and their putative progenitors indicated that all young genes were evolving under natural selection. We found 11 of 25 identified genes hold a chimeric gene structure derived from multiple parental genes/flanking targeting sequences and 22 out of the 25 genes seem to be functional, as suggested by dN/dS analysis and the presence of cDNA, EST, and MPSS data. The high rate of young gene origination and of chimeric gene formation in rice may demonstrate the rice broad diversification, domestication, its environmental adaptation, and the role of young genes in the rice speciation.

O-109

Spot the difference! The genetic basis of an evolutionary novelty in cichlid fishes

M. Emilia Santos, Walter Salzburger
University of Basel, Basel, Switzerland

The origin and modification of novel traits is a key topic in evolutionary biology. Despite this the genetic and developmental mechanisms driving these processes remain largely unknown. The spectacularly diverse adaptive radiations of cichlid fishes in the East Africa Great Lakes provide an ideal system to study the molecular basis of evolutionary novelties in the context of adaptation and explosive speciation. One characteristic innovation of the most species-rich lineage of cichlids, the haplochromines, are brightly pigmented spots on male anal fins, known as “egg-spots”. Egg-spots are a diverse trait (number, colour and shape) that plays a key role in the territorial and breeding behaviour of about 1500 species of cichlids. Here we report the identification of several egg-spot candidate genes by quantitative next generation sequencing of RNA from egg-spot tissue in the haplochromine cichlid *Astatotilapia burtoni*. We confirmed these results in other haplochromine species through quantitative gene expression analysis (qPCR), and narrowed down our study to one gene – an androgen receptor (AR) cofactor. A comparative genomic analysis between haplochromines and egg-spot-less non-haplochromine species reveals that the coding region of this gene cannot explain the origin and diversity of this trait. However, the upstream regulatory region of the AR cofactor differs dramatically between these groups: haplochromines bear a unique transposable element insertion in the proximity of the transcription initiation site of the AR cofactor. We thus propose that this transposable element insertion changed the expression pattern of the AR cofactor, thereby initiating the morphogenesis of an evolutionary key innovation of one of the presumably single-most species-rich lineages of vertebrates.

O-110

MADS microRNAs: A phylogenomic survey of spliced microRNAs regulating plant developmental control genes

Lydia Gramzow, Dajana Lobbes, Peer Aramillo Irizar, Sophia Walter, Günter Theißen
Friedrich Schiller University, Jena, Germany

MicroRNAs (miRNAs) are small, non-coding RNAs which generally negatively regulate gene expression. In plants, many miRNAs are involved in controlling developmental processes, ranging from root and vascular development to leaf and flower development and including phase change from vegetative to reproductive development. A lot of these miRNAs do so by regulating the expression of developmental control genes. One family of developmental control genes with crucial functions in plant development are the MADS-box genes. They include the clade of *AGL17*-like genes involved in root and leaf development.
AGL17-like MADS-box genes have been shown to be regulated by miRNAs in *Arabidopsis thaliana* and *Oryza sativa*, two distantly related flowering plants. The corresponding miRNA genes in *O. sativa* contain an intron in the section of the gene which later forms the loop in the characteristic hairpin structure of the precursor miRNA. Due to the intron these miRNAs were unpredictable using bioinformatics methods until the recent publication of the algorithm SplamiR. Regulation of *AGL17*-like genes by miRNAs is not known from any other flowering plant species. However, this regulation might have been simply overlooked so far due to the difficulties in prediction.
Here, we investigate the presence of spliced miRNAs regulating *AGL17*-like MADS-box genes in other plant species for which whole genome information is available. To do so, we have first identified *AGL17*-like genes in a number of plant species and constructed a comprehensive phylogeny of these genes. Then, we used SplamiR to predict miRNAs which may regulate these *AGL17*-like genes. We have detected candidate miRNAs for a number of species, among them one miRNA in *Glycine max* which is encoded in natural antisense direction to its target gene. Our results demonstrate that the regulation of *AGL17*-like genes by miRNAs may be common and necessary for proper root and/or leaf development in flowering plants. Furthermore, they illustrate that spliced miRNAs may be quite abundant and many more of these regulators may remain to be discovered.

O-111

The role of prolactin in male pregnancy

Camilla Whittington, Marie-Emilie Gauthier, Anthony Wilson
University of Zurich, Zurich, Switzerland

Prolactin (PRL) is an important pituitary hormone with a diverse range of activities, acting in varied tissue types in all vertebrates via its receptor. PRL is often associated with reproductive function, and has been shown to have a role in oestrous, pregnancy, lactation and induction of parental behaviour in a number of species. The hormone has been most widely studied in mammals, but studies in several species of fish have shown that PRL also induces paternal behaviour, including egg fanning, nest building, and nutrient provisioning to young.
Male syngnathids (seahorses and pipefish) have specialised brooding structures (pouches) that provide protection, aeration, osmoregulation and possibly nutrient provisioning to developing embryos. Previous work has shown that male reproductive function is compromised in hypophysectomised seahorses, but can be rescued by treatment with exogenous PRL. In order to identify the function of this hormone and its mode of action during syngnathid pregnancy, we have identified the genes encoding PRL and its receptor (PRLR) in this group. Here we investigate gene expression patterns of PRL and PRLR in brooding males of two syngnathid species differing in pouch complexity, targeting key stages of pregnancy in order to determine the mode of action of PRL during syngnathid reproduction. This is part of a broader project looking at differential gene expression during syngnathid pregnancy, and paves the way for future investigations into how this novel trait has evolved.

O-112

Evolution and Architecture of the Inner Membrane Complex of the Malaria Parasite

Maya Kono[1], Susann Herrmann[2], Noeleen Loughran[3], Ana Cabrera[2], Klemens Engelberg[1], Christine Lehmann[1], Dipto Sinha[1], Boris Prinz[1], Ulrike Ruch[1], Volker Heussler[1,4], Tobias Spielman[1], John Parkinson[3], Tim Gilberger[1,2]
[1]Bernhard-Nocht-Institute for Tropical Medicine, Hamburg, Germany, [2]McMaster Universtiy, Hamilton, Ontario, Canada, [3]Hospital for Sick Children and University of Toronto, Toronto, Ontario, Germany, [4]University of Bern, Bern, Switzerland

The Apicomplexa (superphylum Alveolata) are a large group of unicellular eukaryotes, the majority of which are obligate intracellular parasites. These parasites represent a significant global healthcare burden with the increasing prevalence of infections involving parasites such as *Plasmodium*, *Toxoplasma* and *Cryptospordium* (causative agents of malaria, toxoplasmosis and cryptosporidium respectively). Genome sequencing efforts are beginning to reveal an arsenal of specialized proteins associated with distinct life cycle strategies mediating host cell invasion and persistence. For example, *Toxoplasma* possesses a large superfamily of surface proteins, termed SAG1 related sequences (SRS), associated with host cell recognition and immune modulation. Similarly the *Plasmodium*-specific var protein family is involved in evasion of the host's immune response allowing infection to persist. While such taxon-specific innovations required for pathogenesis are apparent across this phylum, it may not be the case that all machinery required for pathogenesis is taxon-specific. Previous studies of the inner membrane complex (IMC), an important cellular machine required by the malarial parasite, *Plasmodium falciparum*, for successful invasion, suggest that some elements are more widely conserved. Here we were interested in examining the extent to which the components of this complex are evolutionary conserved across the Apicomplexa or if taxon-specific adaptations are present. We investigate the evolution of 17 known members of the complex in an attempt to explore its origins. A series of systematic sequence searches revealed that the structurally diverse IMC proteins display a mosaic pattern of evolutionary trajectories. A global homology network suggests that the majority of these proteins represent a core conserved complex, with few lineage-specific innovations that may be required for unique invasion strategies. Interestingly, these profiles do not correlate with cell localisation data - both non-alveolin and alveolin-like proteins arose at different evolutionary time points. Together these findings support an emerging picture that the evolution of the IMC required both the recruitment and subsequent diversification of ancient eukaryotic proteins, as well as the innovation of novel apicomplexan-specific proteins.

O-113

Positive Selection on Genes involved in Zinc Homeostasis in Modern Humans

Johannes Engelken[1,2], Elena Carnero-Montoro[1], Marc Pybus[1], Manu Uzcudun[1], Giovanni M. Dall'Olio[1], Pierre Luisi[1], Jaume Bertranpetit[1], Mark Stoneking[2], Elena Bosch[1]
[1]Institute of Evolutionary Biology (CSIC-UPF), Universitat Pompeu Fabra, 08003 Barcelona, Spain,[2]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

Micronutrients such as the essential trace element zinc play an important role in human physiology and their physiological and cellular concentrations are tightly controlled. Limited availability of micronutrients has many adverse effects on human health. Interestingly, a number of micronutrients show geographic patterns of abundancy. Some micronutrient deficiencies are known from modern agriculture as well as from international health records. Consequently, we hypothesize that certain human genes involved in micronutrient metabolism may have been under natural selection as an adaptation to such micronutrient availability issues. Here, we combine population genetic analyses with functional approaches in order to test whether recent positive selection has acted on micronutrient-related genes, and we characterize functional genetic variants related to micronutrient homeostasis in humans.

Using population worldwide data from the 1000genomes project as well as from the HGDP panel, we interrogated a set of 50 micronutrient-related candidate genes (such as those involved in zinc metabolism) for signatures of recent positive selection. A number of zinc-related genes are known to be transcriptionally regulated by dietary components and others possess an expression QTL. Therefore, to test for functional diversity, mRNA expression analyses of the candidate genes were carried out in a collection of human kidney (n=120) and liver (n=150) tissue samples in which we had previously quantified the zinc concentrations. SNP genotyping of the samples could confirm and newly identify eQTLs as well as genes with mRNA expression levels that are correlated with tissue zinc concentration. In order to link the known and newly identified functional variants back to our evolutionary hypothesis, we will explore their possible overlap with the genomic footprints of natural selection.

O-114

The evolution of silencing suppression and other ways viruses have to escape from RNA silencing

Santiago F. Elena[1,2]
[1]*IBMCP (CSIC-UPV), Valencia, Spain,* [2]*Santa Fe Institute, Santa Fe, New Mexico, USA*

It is well established now that viruses are both initiators and targets of RNA silencing, a mechanism used by plants as an antiviral adaptive response. However, viruses have evolved counterdefenses that allow them to escape from silencing surveillance and successfully infect the plant. Such diverse mechanisms range from the high genetic variability and adaptability bestow by the quasispecies nature of viral populations to the acquisition of suppressor activities, usually encoded by multifunctional proteins. In the laboratory, we took an experimental approach to understand the implications of these two mechanisms on the evolution of two potyviruses (TEV and TuMV).

In a first set of experiments, we have created a collection of 27 amino-acid replacement mutants in TEV's silencing suppressor, HC-Pro, and evaluated their efficiency as suppressors in transient-coexpression assays with GFP in *Nicotiana benthamiana* plants. In parallel, for each mutant protein we have evaluated the ability of the virus to move systemically, their accumulation and the severity of the symptoms induced. Nine mutants showed no activity at all, three showed significant reduction in suppressor activity (hyposuppressors), nine mutations were neutral, and five were stronger suppressors than wildtype (hypersuppressors). All hyposuppressors accumulated less and induced milder symptoms than wildtype TEV. Hypersuppressor were most variable in their accumulation and symptomatology. Next, we ran serial transfer evolution experiments with three mutants from each category and found that different lineages converged to the phenotypic value of the wildtype virus (stabilizing selection).

It was proposed ta few years ago hat expression of amiRNAs targeting TuMV HC-Pro might be an efficient way of generating resistant plants. However, *in vitro* experiments with some animal viruses had shown rapid generation of escape mutants. These observations prompt caution against the use of amiRNA-derived resistances. We have done a large-scale evolution experiment in which TuMV populations have been serially transferred either on fully susceptible *Arabidopsis thaliana* Col-0 or on partially resistant 10-4, after each transfer, we evaluated their ability to replicate on the amiRNA-HC-Pro transgenic plants. These experiments simulate the generation of neutral or quasi-neutral genetic variability in viral populations replicating on a wildtype or partially susceptible host that periodically are challenged to replicate on the resistant plants. We found that lineages evolved in Col-0 plants accumulated escape mutants in a steady manner, taking an average of 15 passages to break resistance. By contrast, evolution in partially resistant plants speeded up the process, taking only 2 passages to break resistance.

O-115

Aberrant piRNA Production and Global TE Derepression in *Drosophila* Interspecific Hybrids Suggests that Rapidly Evolving piRNA Proteins Contribute to Genome Defense

Erin Kelleher, Daniel Barbash
Cornell University, Ithaca, NY, USA

Exceptional rates of germline transposition can be highly deleterious to eukaryotic hosts, causing mutations that disrupt gene function and lead to chromosomal damage and gonadal atrophy. The activity and distribution of different transposable element (TE) families often evolve rapidly, inducing strong selection on host genomes to maintain control of germline transposition. Recent research on the Piwi-interacting RNA (piRNA) pathway has made the exciting discovery that these proteins act in concert with TE-derived small RNAs (piRNAs) to silence active TEs in male and female germlines. piRNAs and their associated proteins are maternally deposited, thus priming the silencing process in the offspring's primordial gonads.

It has therefore been suggested that the piRNA pathway adapts to the dynamic pool of TEs by modulating the composition of maternally deposited piRNAs. In *Drosophila*, however, many piRNA effector proteins show signatures of adaptive evolution, implying coding sequence changes may also contribute to the evolution of host genome defense. To test this hypothesis we compared ovarian piRNA and mRNA pools between parental strains and their F1 offspring for crosses at two different degrees of divergence: between two strains of *D. melanogaster*, and between *D. melanogaster* and its sibling species *D. simulans.* In the intraspecific cross, we observed that differences in ovarian piRNA composition between strains is predictive of TE derepression, consistent with the role of the piRNA pool in determining the germline activity of TEs. In contrast, in the interspecific cross, dramatic differences in ovarian piRNA pools are rarely predictive of piRNA abundance or TE expression levels in hybrid offspring. Rather, interspecific hybrids phenocopy piRNA pathway mutants in terms of aberrant piRNA protein localization, reduced transcription of precursor piRNAs, disrupted piRNA processing, and global TE derepression. We propose that piRNA pathway dysfunction in interspecific hybrids reflects incompatibilities arising from the combined effects of interspecific divergence in both the small RNA and protein components of the piRNA pathway.

O-116

Evolution of Piwi-interacting RNAs in Humans

Sergio Lukic, Kevin Chen
Rutgers University, Piscataway, NJ, USA

Piwi-interacting RNAs (piRNAs) are a recently discovered class of 24- to 30-nt noncoding RNAs whose best-understood function is to repress transposable elements (TEs) in animal germ lines. In humans, TE-derived sequences comprise

~45% of the genome and there are several active TE families, including LINE-1 and Alu elements, which are a

significant source of de novo mutations and intrapopulation variability. In the “ping-pong model,” piRNAs are thought to alternatively cleave sense and antisense TE transcripts in a positive feedback loop. Because piRNAs are poorly conserved between closely related species, including human and chimpanzee, we took a population genomics approach to study piRNA function and evolution. We found strong statistical evidence that piRNA sequences are under selective constraint in African populations. We then mapped the piRNA sequences to human TE sequences and found strong correlations between the age of each LINE-1 and Alu subfamily and the number of piRNAs mapping to the subfamily. This result supports the idea that piRNAs function as repressors of TEs in humans. Finally, we observed a significant depletion of piRNA matches in the reverse transcriptase region of the consensus human LINE-1 element but not of the consensus mouse LINE-1 element. This result suggests that reverse transcriptase might have an endogenous role specific to humans. In addition to our SNP-based analyses, we have also found evidence for positive selection on human piRNAs at the level of copy number variation. Finally, we have compared the evolution of the piRNA system to the evolution of the CRISPR system in prokaryotes which also acts as an RNA-based adaptive immune system against foreign nucleic acids, specifically phage and plasmids.

O-117

Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement

Wenwen Fang[1], Xing Wang[1], John Bracht[1], Mariusz Nowacki[1,2], Laura Landweber[1]
[1]Princeton University, Princeton NJ, USA, [2]University of Bern, Bern, Switzerland

Genome duality in ciliates offers a unique system to showcase their epigenome as a model of inheritance. In *Oxytricha,* the somatic genome is responsible for vegetative growth, while the germline contributes DNA to the next generation. Somatic nuclear development eliminates all transposons and other "junk DNA", which constitute ~95% of the germline. Here we demonstrate that Piwi-interacting small RNAs (piRNAs) from the somatic nucleus specify genomic regions for retention in this process. *Oxytricha* piRNAs map primarily to somatic genome regions, representing the minority fraction (5%) of the germline that is retained. This marking is orthogonally different from the ciliates *Paramecium* and *Tetrahymena*, where Piwi-interacting RNAs map primarily to deleted regions instead, suggesting that a "sign change" occured during the evolutionary divergence of these lineages. Furthermore, injection of synthetic piRNAs corresponding to normally-deleted regions in *Oxytricha* leads to their retention in the next generation. Remarkably, the DNA retention pattern also transfers to the next sexual generation (F2 generation). Therefore, transiently available small RNAs can effect heritable DNA sequence change in the somatic genome across multiple generations. Our findings highlight small RNAs as powerful transgenerational carriers of epigenetic information for genome programming and the evolutionary plasticity of RNA-mediated mechanisms of genome surveillance.

O-118

Statistical Methods for Detecting Selection Associated with High Altitude in Tibetans and Ethiopians

Rasmus Nielsen
UC Berkeley, Berkeley, USA

The hypoxic conditions of high altitude regions are among the strongest environmental challenges that humans have faced in their recent evolutionary history. Previous studies have shown that populations in the Andes and in Tibet have a number of genetic adaptations that allow them to function better in this environment. Using targeted new-generation sequencing data we further analyze several of the candidate genes that have been associated with altitude adaptation in Tibet to determine if selection has been acting on standing variation or de novo mutations and to determine the strength of selection. We also further analyze the demographic history of the Tibetan populations and compare them to a group of Southern Han individuals. Finally, we analyze SNP genotyping data from Ethiopian populations that may also be adapted to the high altitude environment of the Semien Plateau of Ethiopia. These populations are admixed and have a certain fraction of non-African ancestry. After correcting for admixture, we find that the gene showing the strongest allele frequency difference between high altitude and low altitude populations is a gene previously known to play a role in the response to hypoxia.

O-119

Hitch-hiking in spatially extended populations

Nick Barton[1], Jerome Kelleher[2], Amandine Véber[4], Alison Etheridge[3]
[1]IST Austria, Klosterneuburg, Austria, [2]IEB, University of Edinburgh, Edinburgh, UK, [3]Oxford University, Oxford, UK, [4]Ecole Polytechnique, Paris, France

The advance of a favourable allele through a spatially continuous population is driven by the reproduction of a small number of individuals at the leading edge. Thus, a selective sweep causes much more coalescence than in a panmictic population. We give a simple analytical approximation for the rate of coalescence due to a sweep through a one-dimensional population. In two dimensions, the effect of a sweep is qualitatively different, with coalescence being dominated by rare fluctuations in which a few genes jump well beyond the bulk of the wave. This further strengthens the average effect, and makes it highly stochastic.

O-120

Unbiased and Exhaustive Experimental Measurement of Point-mutant Fitness Effects

Dan Bolon
University of Massachusetts Medical School, Worcester, MA, USA

Two of the most fundamental problems in evolutionary biology involve characterizing the relative importance of adaptive (positive selection) and non-adaptive (genetic drift) processes in the evolution of natural populations, and understanding the molecular basis of adaptive phenotypic variation. Specifically, we would like to know, for a given selective pressure, what mutational changes cause fitness effects that are deleterious, neutral, or beneficial to the organism (and what proportion of changes fall in to these respective classes). To address this question, we developed a novel technology that we term EMPIRIC (Exceedingly Meticulous and Parallel Investigation of Randomized Individual Codons) allowing for the systematic generation and fitness characterization of all possible point mutations at all possible sites along a genomic region. Our experimental results with two genes (yeast ubiquitin and Hsp90) indicate that this technology is highly efficient and has the opportunity to indeed revolutionize experimental evolutionary studies. Further, our results are suggestive that strongly conserved regions may in fact be well predicted by the nearly neutral model of molecular evolution, a long contentious issue (Hietpas et al., PNAS 2011), and that when analyzed at appropriate structural and mutational resolution that the density of molecular interactions strongly correlates with fitness sensitivity to mutation (Zuckerkandl, J. Mol. Evol. 1976).

O-121

McDonald-Kreitman Test under Frequent Adaptation: Problems and Solutions

Philipp Messer, Dmitri Petrov
Stanford University, Stanford, CA 94305, USA

Application of the McDonald-Kreitman (MK) test to population polymorphism data has revealed a surprisingly high rate of adaptive evolution in many species. For example, in *Drosophila melanogaster*, MK-type tests suggest that α - the proportion of nonsynonymous substitutions that were adaptive - is on the order of 0.5 or larger. MK-type tests are based on the assumption that polymorphisms at different sites in the genome evolve independently of each other, i.e., that there is no Hill-Robertson interference (HRI) between sites. However, under frequent adaptation, neutral and slightly deleterious mutations could often hitchhike together with adaptive mutations, raising the question of whether MK-type tests remain unbiased. We performed an extensive simulation study to investigate the effects of HRI on the genome-wide patterns of polymorphism and divergence. Our forward simulation models the evolution of a chromosome with realistic gene structure, incorporating mutation, recombination and selection and allowing for an arbitrary distribution of fitness effects of new mutations. Using evolutionary parameters resembling those inferred for human evolution, we find that estimates of α from MK-type tests are substantially confounded by HRI when selective sweeps are frequent. Specifically, in the presence of slightly deleterious mutations, estimates of α are often below zero although the actual α can be substantial (e.g. α = 0.5 in our simulations). Moreover, we find that already for intermediate α, the resulting HRI can strongly distort the site frequency spectrum of neutral polymorphisms at synonymous sites from that expected under random genetic drift alone, casting doubt on the accuracy of the commonly used approach to infer demography from such data. Interestingly, methods such as DFE-alpha which first infer demography from synonymous sites and then use the inferred demography to correct the estimation of α at nonsynonymous sites, obtain almost the correct α but entirely incorrect estimates of demography. When applied to our simulation data, these methods typically infer an extreme past population expansion although there was no such expansion in the simulations. We finally provide a simple extension of the MK-test where different intervals of derived allele frequencies are treated separately. We show that this new approach allows for an accurate asymptotic estimation of α even in the presence of weakly deleterious mutations and strong HRI.

O-122

Genome-wide patterns of natural variation reveal ongoing genomic conflict in *Drosophila mauritiana*

Viola Nolte, Ram Vinay Pandey, Robert Kofler, Christian Schlötterer
Vetmeduni, Vienna, Austria

Drosophila mauritiana, a close relative of *D. melanogaster*, is endemic to a few islands in the Indian Ocean. Despite the importance of *D. mauritiana* as a model for understanding the genetic basis of speciation processes, its genome sequence is not yet available, and natural variation has been characterized only at a few loci. We generated a draft genome sequence for *D. mauritiana* and characterized genome-wide polymorphism by sequencing pooled individuals (Pool-Seq). We address the long-debated phylogenetic relationship within the *D. simulans* clade by showing that *D. mauritiana* and *D. simulans* are more closely related to one another, while *D. sechellia* is the more divergent species. In consequence of the large amounts of shared polymorphism within the *D. simulans* group, we find no evidence for faster X chromosome evolution, but a very pronounced effect in comparisons involving either of the more distantly related *D. melanogaster* or *D. yakuba*. We demonstrate how the well-documented change in recombination landscape in *D. mauritiana* affects the distribution of polymorphisms across the genome: regions close to the centromere which exhibit low variation and reduced selection efficacy in *D. melanogaster* show normal polymorphism levels and no enrichment of non-synonymous changes in *D. mauritiana*. Finally, we report the genomic signatures of ongoing genomic conflict in *D. mauritiana*. Two large regions (> 500 kb) near the genes *Dox* and *OdsH* show the signature of almost complete selective sweeps. In addition to *Dox* and *OdsH*, we find that another class of genes thought to be involved in genomic conflict, nucleoporin genes, are among the genes showing the strongest evidence of recurrent adaptive evolution.

O-123

Genome-wide quantification of bacterial mutational-biases

Ruth Hershberg
Technion - Israel Institute of Technology, Haifa, Israel

Mutation is the ultimate source of all genetic variation, and as such propels all variation dependent evolutionary processes. Different types of mutations occur at different frequencies, inserting biases into the patterns of genetic variation generated by mutation. Such mutational biases can greatly affect evolutionary outcomes. To date, however, mutational biases remain largely uncharacterized even within individual organisms, and our understanding of how these biases vary between organisms is even more limited. I will present an approach to study mutational biases at a genome-wide level, by focusing on bacterial lineages for which natural selection is extremely relaxed. When selection is relaxed, it affects only mutations conferring extremely high fitness effects. Thus, patterns of variation observed under relaxed selection are expected to better reflect mutational patterns. I will describe how we used this approach to study two questions regarding mutational biases: (1) Does variation in mutational biases drive variation in bacterial nucleotide content? The extreme variation observed in nucleotide content in bacteria (<25% to >75% GC), has long been assumed to be the result of extreme variation in mutational biases. It was widely thought that in GC-rich bacteria, mutations from AT to GC occur more frequently, while in AT-rich bacteria mutation from GC to AT are the most frequent. I will describe how we showed that this is in fact not the case. Rather, mutation is AT-biased across all studied bacterial lineages, irrespective of their current nucleotide content. These findings demonstrate that selection or a selection like process must be involved in determining bacterial nucleotide content. (2) How do mutation and natural selection each affect transition / transversion biases in bacteria? It has long been noted that transition substitutions (changes from Purine to Purine, or from *Pyrimidine to Pyrimidine)* are much more frequent than transversion substitutions (Changes from Purine to *Pyrimidine or vice versa)*. Both mutation and natural selection are expected to contribute to this bias. I will describe how we disentangled the effects of mutation and selection on creating this widely observed bias, with surprising results.

O-124

Metagenomics analysis of the human gut

Peer Bork
EMBL, Heidelberg, Germany

As the number of individual human genome sequences increase, our knowledge about the human microbiome, that is all the microbes that live in and around us, is likewise expanding. The main access to this invisible world is currently through 16S/OTU profiling or metagenomics, environmental shotgun sequencing of various human body sites whereby the gut is by far the most prominent one. In metagenomics, data across samples are compared by using the presence and abundance of species or functions, mostly at the genus level, usually with the goal of finding biomarkers for various disease states although protocols are not standardised yet and the knowledge about variation in the human population is still very limited. For example, we recently identified three microbial gut community types across several western countries form three continents, which we dubbed enterotypes (Arumugam et al, Nature, 2011) indicating a stratification of the human population with consequences for biomarker discovery as well as for responses to diet and drug intake. We also analyzed (meta)genomic variation in gut microbial communities to test individuality beyond the species level 16S which is impossible using 16S profiling. In a sizable cohort of 252 stool samples, these variation patterns were individual and appear, at least in healthy people, stable over a considerable time span. For diagnostic purposes, it is unclear how much biology they reveal, and we have also started to compare them with samples from biopsies where we found, as expected, quite a few differences.

O-125

Tracking protein length changes reveals pseudogenization in action.

Wilson Sung, Brian Golding
McMaster University, Hamilton ON, Canada

A method for identifying and mapping fusion and fission events onto a phylogeny was developed and applied to Bacillaceae genomes. In contrast to previous studies across longer evolutionary time scales, we found that gene fission is more common than fusion in bacterial genomes. Fusion and fission events are generally rare across the Bacillaceae and are genome-specific, but we unexpectedly uncovered a large number of fissions specific to the genetically monomorphic *Bacillus anthracis* lineage. Our results suggest that the *B. anthracis* lineage may be under an accelerated rate of gene fragmentation, which is a common evolutionary trend found in lineages that have recently become host-restricted. We hypothesize that other bacteria evolving under similar conditions would also exhibit detectable fission patterns. *Salmonella enterica* and *Yersinia pestis* genomes are used to confirm this hypothesis.

O-126

Bacterial interactome dynamics: evolutionary and co-evolutionary trends in gene family acquisitions

Ofir Cohen, Tal Pupko
Tel Aviv University, Tel Aviv, Israel

Comparative genomics studies of prokaryote genomes revealed a huge variance in gene content, and uncovered the ubiquity and importance of macro evolutionary events in which genes are gained and lost. These studies further discovered the significance of gene family acquisitions, a subtype of horizontal gene transfer, baring exceptional contribution to biological innovation and niche adaptation. While virtually all prokaryote genomes were shaped by gene transfer during their evolution, other and we have previously shown that gain and loss dynamics substantially differ among gene families. Our motivation is to better understand these evolutionary dynamics in the context of the protein interactome.

We have recently published a study, in which we examined the biological factors that determine transferability. The prevailing hypothesis (known as the 'complexity hypothesis') was that the difference in transferability among genes might be explained by both their functional category (biological process) and their number of interacting partners (connectivity). We disentangled these two factors by using advanced probabilistic models to quantify the transferability of gene families and rigorous statistical analysis to examine the associations. By measuring the impact of each factor while accounting for the confounding effects of the other, we revealed that high connectivity constitutes a strong barrier to transferability, whereas functional category is mainly a by-product due to the variability in connectivity among functional categories.

In our current research we study co-evolutionary interactions and cooperation among gene families. In this approach, co-evolutionary force among gene families is detected by patterns of correlated gains and losses along the evolutionary tree. Applying our model-based approach to study the evolution of hundreds of microbial species enabled us to reliably reconstruct the co-evolutionary network among all gene families. Our findings suggest a pervasive and important role for co-evolution in shaping microbial genomes, manifested in a dense co-evolutionary network connecting most gene families. This co-evolutionary network re-affirms known associations among genes in close genomic proximity, physically interacting genes, and members of the same metabolic pathways, while uncovering many additional co-evolutionary associations. The biological significance is further demonstrated by strong functional associations found among co-evolving genes. Furthermore, using these co-evolutionary interactions we can reliably classify many poorly annotated gene families. This function classification approach provides a significant augmentation to the existing annotation system and suggests multiple-functions for many gene families that were previously classified with a single function.

O-127

Fitting structured models of substitutions and indels(invited talk, symposium: "Estimating and simulating models of molecular evolution")

Ian Holmes
University of California, Berkeley, CA, USA

Many statistical models used in molecular evolution and genome annotation can be represented using phylogenetically parameterized grammars. These include Yang's "space-time" autocorrelated discretized gamma distribution over rates; Siepel and Haussler's "phylo-HMMs"; Thorne, Goldman and Jones' HMM for protein secondary structure; the "X-Decoder" model for annotating RNA structures in viral genomes; and all point substitution models. I will describe how these models can be prototyped, fit to data (yielding meaningful parameter estimates), and simulated from, using the XRATE package. This package has a flexible Scheme-based scripting language that allows the construction of elaborate models with extensive repetition, as well as data-dependent models. I will then discuss extension of the literature on substitution models to the world of indel models. The indel model has received scant attention and usage compared to the substitution model; some reasons for this are inherent (indel rates cannot be measured with pinpoint accuracy like substitution rates, but must be averaged over a longer region of sequence) but some, I will argue, are technological (nearly all existing multiple alignment programs introduce systematic biases for the purposes of estimating indel rates; the main exceptions are the "Statistical Alignment" tools for co-sampling alignment and phylogeny, such as HandAlign and BaliPhy, however these can be very time-consuming to run). New programs (specifically the ProtPal software from our lab) permit the rapid AND accurate estimation of indel rates in proteins, suggesting the possibility of including indel rates in both genome-wide and gene-specific analyses of evolutionary pressure in proteins.

O-128

Harnessing Non-Local Evolutionary Events for Tree Inference

Liangliang Wang, Alexandre Bouchard-Côté
University of British Columbia, Vancouver, Canada

Current tree inference methods ground their analysis on a relatively restricted range of evolutionary phenomena, limited in most cases to substitution events and, less frequently, to context independent long indels. However, sequence change is caused by many other important molecular mechanisms, including slipped strand mispairing (SSM), a well known explanation for the evolution of repeated sequence. However, SSMs have not yet been exploited as phylogenetically informative events, since it is generally non-trivial to extend tractable sequence evolution models beyond point mutations.

In this work, we present a tree sampling algorithm allowing non-local sequence evolution models. By non-local, we mean that sequence changes can depend on contexts of unbounded random size. More precisely, the model is based on an extension of the Doob-Gillespie construction for continuous time Markov process, where the exponential rates are allowed to depend on an unbounded context. To compute the likelihood, our algorithm uses a recursive particle approximation instead of depending on complicated analytic results or biased numerical integration. Using a Bayesian approach, we aim to estimate the posterior distribution of phylogenetic trees. This is accomplished using a particle Markov chain Monte Carlo (PMCMC) method, which transforms the approximate likelihood calculation into a valid, consistent MCMC algorithm for the SSM-aware model.

We apply our method to study how the complex interplay between SSMs and point mutations affects the inferred trees. Our experiments involve simulated sequence data with realistic patterns of SSMs. We compare the tree reconstructions obtained with our model incorporating SSMs to other models that ignore SSMs. We also apply our method to plant intron datasets, where SSMs play an important role in sequence change.

As an extra advantage, our proposed model and algorithms can be adapted to do joint estimation of multiple sequence alignments (MSAs), evolutionary parameters, and phylogeny. In contrast to methods based on an MSA point estimate, joint estimation avoids guide tree bias and over-confidence problems. Moreover our method can do so and at the same time take into account non-local evolutionary events such as SSMs, which are known to affect MSAs and are already used to do manual alignment.

The proposed model and algorithms are not limited to SSM modelling. It can also be applied to context-sensitive substitutions and to incorporate structural constraints in RNA evolution.

O-129

Advanced probabilistic models to study gain and loss dynamics of gene families among microbial species

Tal Pupko, Ofir Cohen
Tel-Aviv University, Tel-Aviv, Israel

In microbial species, genes are frequently gained and lost. This leads to high genome plasticity, which plays a major part in shaping microbial species genomes. The most prominent means by which genes are gained in prokaryotes is Horizontal Gene Transfer (HGT). The study of the evolutionary dynamics of gains and losses of gene families is facilitated by a compact representation called phyletic pattern depicting gene families' content of all genomes in the data set. This binary presence\absence matrix is analyzed by modeling gain and loss dynamics as continues-time Markov chains. While pioneering models assumed that gain and loss rates are equal, we have shown that allowing gain and loss rates to differ had significantly improved the likelihood of the models with loss rate typically substantially higher than gain rate. The observation that some gene families are ubiquitously present (‘core' of the genome) while other gene families sporadically appear in genomes, others and we had developed models that allow for variability of rates among gene families. Furthermore, since gene-specific higher tendency for gain does not necessarily entails higher loss propensity, we developed an advanced model that allows for both the gain and the loss rates to vary among genes, a further step toward more realistic modeling. Importantly, we integrated these models within a stochastic mapping approach, in which probabilities and expectations of gain and loss events are estimated for each gene family and for each branch of an underlying phylogenetic tree.

In a recent study we compared the performance of our model-based stochastic mapping approach to the classical and prevalent approach of maximum parsimony in detecting gene family gains and losses on specific branches. Our simulation-based findings suggest that over vast range of evolutionary scenarios the model-based approach outperforms maximum parsimony. By applying our methodology to an extensive across-species genome-wide dataset we quantified transferability levels and propensity for loss of gene families. Additionally, we studied the factors that determine the observed variability in evolutionary dynamics among genes. We also employ lineage-specific quantification of events to detect lineages with either exceptional propensity for gains or for losses. Finally, we propose a likelihood based method to detect co-evolving gene families (gene families that tend to be lost and gained simultaneously). This method can have broad application for genomic annotations.

O-130

RevBayes: One program for your whole phylogenetic analysis

Sebastian Höhna[1], Fredrik Ronquist[2], John Huelsenbeck[3]
[1]Stockholm University, Stockholm, Sweden,[2]Swedish Museum of Natural History, Stockholm, Sweden,[3]University of California, Berkeley, Berkeley, California, USA

The amount of models has increased rapidly in recent years. The models used in molecular evolutionary analyses range from the gene level (e.g. substitution models) to the species level (e.g. diversification models and phylogeographic models). This plethora of models has retained a rich landscape of computer programs where most models are only implemented in one single program. Hence, most analyses are performed gradually using different computer programs. However, there are several reasons why a single software unifying as many models as possible is beneficial:

First, an integrative analysis can combine several models and therefore take the uncertainty in the different steps of the analysis into account. Second, model selection and model choice is in most situations only feasible if competing models are implemented in the same computer program. Furthermore, mixture models are easier to implement if more models are available.

Our new computer program, RevBayes (expected realease: June 2012), is an R-like environment for Bayesian phylogenetic inference which aims to integrate the whole analysis. Most standard models are implemented and it already offers more models than its predecessor MrBayes. Even if a model is missing, adding the new model is as easy as writing a function in R. RevBayes provides a very flexible environment to specify a model and run a phylogenetic analysis. Once a model is specified, RevBayes offers not only the inference machinery but also simulation tools for observations under the model.

In this talk I will give a brief overview of RevBayes and demonstrate its abilities on an integrative, fully hierarchical Bayesian analysis where, amongst others, the tree topology, the divergence times, clock rates and diversification rates are estimated simultaneously. Several models and mixtures thereof are considered and evaluated using marginal likelihoods. Our preliminary results show that the tree topology, divergence times and diversifications rates can be biased if estimated independently.

O-131

Phylogeny-aware Analysis of Metagenomic Samples

Alexandros Stamatakis, Simon Berger, Nikos Alachiotis
Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

In this talk I will discuss new methods for phylogeny-aware analysis of metagenomic short read samples and try to outline the potential methodological advantages of this approach.

More specifically, I will address algorithms for evolutionary placement of short reads into given reference phylogenies and phylogeny-aware methods for aligning short reads to given reference phylogenies and reference alignments. I will also address how these methods can be parallelized. Furthermore I will briefly discuss the common output format that is implemented by several evolutionary placement and post-analysis tools.

Finally, I will address some of the future challenges in the field with respect to designing post-analysis methods and making the transition to production-level clinical deployment of these new approaches.

O-132

The BH Mixture Model

Liwen Zou[2], Edward Susko[1], Chris Field[1], Andrew J. Roger[1]
[1]Dalhousie University, Halifax, NS, Canada,[2]North Carolina State University, Raleigh, NC, USA

The default general Markov model described by Barry and Hartigan (1987), known as the BH model, assumes that while evolutionary processes may vary over time, they are identical among sites in DNA sequences. However, it has long been recognized that the evolutionary dynamics of sites differ depending on a variety of factors. For instance, in Fitch and Margoliash (1967), invariant sites in the amino acid sequences of Cytochrome C were identified. Rates-across-sites variation is usually modeled by assigning rates to sites according to a discretized approximation to the Γ distribution (Yang 1994b) or a mixture of a discretized Γ distribution and an invariable site class, "I", (Gu et al. 1995) and has been shown to be important for accurate phylogenetic estimation (c.f. Yang 1994b, Thomas et al. 2006, Bromham 2009).

The YR nonstationary model in Yang and Roberts (1995) and the GG nonstationary model in Galtier and Gouy (1995) as implemented by Boussau and Gouy (2006) incorporated rate-across-sites variation using a discretized Γ distribution. In Jayaswal et al. (2007), a BH + I model was proposed. This model treated the invariable and variable sites differently through a mixture model. In a follow up study, Jayaswal et al. (2011) suggested two stationary models in order to simplify the BH model. These two simplified models both allow invariable and variable sites.

While adjusting for rates-across-sites variation is now common practice in phylogenetic analyses, functional constraints on sites in a gene sequence can also change over time, causing shifts in site-specific evolutionary rates. This process is often referred to as heterotachy (Lopez et al. 2002) and has been modeled in various ways (c.f. Tuffley and Steel 1998, Galtier 2001, Huelsenbeck 2002, Susko et al. 2003). All of these models effectively allow rates and hence substitution matrices to vary across both sites and lineages (Wu and Susko 2010). Here we introduce a BH mixture model that is more general than the BH+I model in that it not only allows completely different models along edges of a topology, but it also allows for different site classes whose evolutionary dynamics (e.g., rates, frequencies over sites and edges) can take any form. We also provide a method that can be used in any nonstationary model but that has been missed by the nonstationary models for estimating the edge length that can be interpreted as the expected number of substitutions in any nonstationary models.

O-133

A genomewide map of Neandertal ancestry in modern humans

Sriram Sankararaman[1,2], Nick Patterson[2], Swapan Mallick[1,2], Svante Paabo[3], David Reich[1,2]
[1]Harvard Medical School, Boston, USA, [2]Broad Institute of Harvard and MIT, Cambridge, USA, [3]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Analysis of the genomes of archaic hominins, such as Neandertals and Denisovans, has revealed that these groups have contributed to the genetic variation of modern human populations. Yet, we know little about how these ancient mixtures have shaped the genetic structure of human populations and even less about their impact on human evolution. To answer these questions systematically, we need a map of archaic ancestry *i.e.*, a map that labels whether each region of an individual genome is descended from these archaics.

Building such a map is technically challenging because of the antiquity of these gene flow events. We have identified signatures based on patterns of variation at single SNPs as well as haplotypes that are informative of ancient gene flow. We propose a principled method based on the statistical framework of Conditional Random Fields (CRFs) that integrates these patterns leading to highly accurate predictions.

We applied our method to polymorphism data in European and East Asian individuals from the 1000 genomes project, in conjunction with the draft sequence of the Neandertal genome, to obtain the first genomewide map of Neandertal ancestry. Analysis of this map reveals several findings:

1. We identify around 35,000 Neandertal-derived alleles in Europeans and 21,000 in East Asians.

2. The map allows us to identify Neandertal alleles that have been the target of selection since introgression. We identified over 100 regions in which the frequency of Neandertal ancestry is extremely unlikely under a model of neutral evolution. The highest frequency region on chromosome 4 has a frequency of Neandertal ancestry of about 85% in Europe and overlaps CLOCK, a key gene in Circadian function in mammals. The high frequency, Neandertal-derived variant is specific to Europeans; it is not very common in East Asians. This gene has been found in other selection scans in Eurasian populations, but has never before been linked to Neandertal gene flow.

3. Several of the Neandertal-derived alleles identified in 1) above are found in the >6,000 SNPs associated with common diseases listed in the NHGRI catalog. These Neandertal derived variants are found to be risk variants associated with obesity and protective variants against breast cancer.

4. We also investigate the possibility of using this map to reconstruct the genome of the introgressing Neandertal. Using the ancestries in Europe and East Asia, we can reconstruct about 600 Mb which we expect to increase with larger samples and additional populations.

O-134

Gene duplicability in the primate protein-protein interaction network

Aoife Doherty, David Alvarez-Ponce, James McInerney
National University of Ireland, Maynooth, Maynooth, Ireland

Gene duplication followed by divergence is an important source of genetic novelty. Although duplications occur at a high rate, only a small fraction of duplicated genes are retained. Gene duplicability is the tendency to retain multiple copies of a gene after gene duplication. A number of factors are known to correlate with gene duplicability, such as gene function or protein complexity. Genes and their encoding proteins generally operate as networks of interacting molecules. The position of a gene's encoded product in the protein-protein interaction network has recently emerged as a determining factor of gene duplicability. However, the direction of the relationship between centrality and duplicability is not universal: in *Escherichia coli*, yeast, fly and worm, duplicated genes tend to occupy more peripheral positions in the network, whereas in humans it has been observed that duplicated genes tend to occupy the most central positions. Here, we have inferred the duplication events that took place in all the branches of the primate phylogeny. In agreement with previous observations, we found that duplications generally affected the genes that are more central in the protein-protein interaction network. However, the opposite trend, i.e. duplication being more common in genes whose protein product is peripheral in the network, is observed in the external branch leading to humans, the branch subtending the human/chimpanzee clade and the branch subtending the human/chimpanzee/gorilla clade. This indicates that the relationship between centrality and duplicability underwent modification during primate evolution. Furthermore, we found that genes encoding interacting proteins exhibit phylogenetic tree topologies that are significantly more similar than expected for random pairs of trees, and that genes that duplicated in a given branch of the phylogeny tend to interact with those that duplicated in the same branch. These results indicate that duplication of a gene increases the likelihood of duplication of its interacting partners. Taken together, our observations indicate that the structure of the primate protein-protein interaction network affects gene duplicability in previously unrecognized ways.

O-135

Evolutionary transitions in fungus-farming ants: A whole-genome sequencing approach

Sanne Nygaard[1], Cai Li[2], Morten Schiøtt[1], Guojie Zhang[1,2], Jun Wang[2], Jacobus Boomsma[1]
[1]University of Copenhagen, Copenhagen, Denmark, [2]BGI-Shenzhen, Shenzhen, China

Ant societies, with their complex social organization and diverse life-histories, are obvious model systems for the evolutionary study of social adaption and microbial symbiosis. One of their most spectacular symbiotic innovations is fungus farming, a mutualism based on ants delivering plant material to subterranean gardens where they grow specialized fungi for food.

Since this symbiosis first evolved in the Amazon basin some 50 mya, the attine ant clade has undergone a series of evolutionary transitions, of which the use of specialized rather than generalist fungal strains, active herbivory rather than using dead plant parts, polymorphic rather than monomorphic worker castes, and multiple rather than single mating of queens are the most important.

In a recent study, we showed that the (100x) genome sequence of the leafcutter ant *Acromyrmex echinatior*, a representative of the most highly derived fungus farming ants, contains characteristic changes in detoxification pathways, loss of function in arginine metabolism pathways, and expansion of specific gene-families of peptidases compared to other ant genomes.

We have now sequenced and partly analyzed the genomes of five additional fungus farming ant species, representing all branches in the phylogeny of the higher attine ants and most of the major evolutionary transitions in the entire fungus farming clade. These data allow us to assess genomic changes in an evolutionary context and to link genomic and life-history changes in a comparative manner.

O-136

VLR-encoding loci in lamprey: Illuminating the origin and early evolution of adaptive immunity

Sabyasachi Das, Masayuki Hirano, Qifeng Han, Jianxu Li, Brantley R. Herrin, Peng Guo, Max D. Cooper
Emory Vaccine Center and Department of Pathology and Laboratory Medicine, Emory University, Atlanta, USA

Approximately 500 million years ago two types of recombinatorial adaptive immune systems (AIS) arose in vertebrates. The jawed vertebrates diversify their repertoire of immunoglobulin (Ig)-domain-based T and B cell antigen receptors mainly through the rearrangement of V-(D)-J gene segments and somatic hypermutation, but none of the Ig-based antigen receptor gene of jawed vertebrates has been found in jawless vertebrates. Instead, the AIS of jawless vertebrates is based on variable lymphocyte receptors (VLR) that are generated through recombinatorial usage of a large panel of highly diverse leucine rich-repeat (LRR) sequences. The germline VLR gene has an incomplete structure that is incapable of encoding functional proteins; however, it is, flanked by hundreds of LRR-encoding modules exhibiting remarkable sequence diversity. During the development of lymphocyte-like cells, functional VLR genes are produced by incorporation of the flanking LRR modules into the incomplete VLR gene by a gene conversion-like process. Three VLR genes (VLRA, VLRB and VLRC) have been found in lampreys and are expressed in monogenic and monoallelic fashion by discrete populations of lymphocytes. Unlike VLRB, the genomic organization and promoter architecture of VLRA are similar to that of VLRC. Both in-silico and in-vivo data suggest that VLRB-expressing cells are B- cell like and the VLRA- and VLRC- expressing cells are evolutionary linked with αβ and/or γδ lineages of T lymphocytes. The basic AIS design featuring two interactive T and B lymphocyte arms apparently evolved in an ancestor of jawed and jawless vertebrates within the context of preexisting innate immunity and has been maintained since as a consequence of powerful and enduring selection for pathogen defense purposes.

O-137

Reconstructing cell-lineages trees using somatic microsatellite instability

Yosef Maruvka[1,2], Yitzhak Reizel[3], Noa Chapal-Ilani[3], Ehud Shapiro[3]
[1]Department of Biostatistics and Computational Biology Dana-Farber Cancer Institute, Boston MA, USA, [2]Department of Biostatistics, Harvard School of Public Health, Boston MA, USA, [3]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science,, Rehovot, Israel

The cells inside multicellular organisms develop and mutate like other populations, and their development from the zygote can be described by a binary-rooted tree. The genomic signature differs between the cells due to the accumulation of somatic mutations since the zygote. These differences can be used to reconstruct the cell-lineage tree of different cells. While most types of genetic and epigenetics mutations are too slow for such a fine resolution, there are microsatellites that do mutate fast enough for this purpose.

The presentation will be divided into two parts. The first is methodological: an analysis of the somatic MS mutational process, and the best algorithm choice for reconstructing the cell-lineage tree. The second is applicatory: two concrete examples where the method was applied to outstanding questions in developmental biology.

We analyzed the MS mutational process using 9 mice and 5 humans, each with about 80 cells, and about 90 MSs per cell. We provide the first direct proof for the growth in the mutation rate as the MS loci elongate. The mutation rate differs between different unit sizes, similar to in the germ-line. However, we did not identify bias towards elongation or subtraction as was reported for the germ-line.

We tested for the best algorithm for cell-lineage reconstruction by taking cells from different organisms that should clearly separate into two sub-trees. However, because the different mice we used are related, and due to the randomness of the mutational process, some mixture appears. By testing all the combinations we can create from the different mice (500 combinations), the best algorithm was identified.

We applied this approach to analyze the mammalian female germ-line. For example, we show that on the reconstructed mouse cell lineage trees, oocytes form a cluster separate from hematopoietic and mesenchymal stem cells, both in young and old mice, indicating that these populations belong to distinct lineages.

We also analyzed colon stem-cell and crypt dynamics. Among other findings, we found that the colon epithelium is clustered separately from hematopoietic and other cell types, indicating that the colon is constituted by a few progenitors, and rule out significant renewal of colonic epithelium from hematopoietic cells during adulthood.

“Cell lineage analysis of the mammalian female germline” PLoS Genet, in press.

"Colon Stem Cell and Crypt Dynamics Exposed by Cell Lineage Reconstruction" (2011) PLoS Genet 7(7).

O-138

RNA silencing, DNA methylation and its effects on plant genome size, structure and function

Brandon Gaut
University of California, Irvine, Irvine, CA, USA

DNA methylation is common in all eukaryotes but particularly prevalent in plants. Plants methylate the DNA of both transposable elements (TEs) and genes. For the former, DNA methylation generally suppresses TE activity, acting as a critical component to prevent TE proliferation and concomitant genome size expanxion. However, the methylation of TEs near genes affects gene expression, leading to a trade-off between the largely beneficial effects of TE suppression and potentially deleterious effects on gene expression. The function of DNA methylation on coding genes is not yet fully elucidated, but there is growing evidence both that methylation targets genes of essential function and that such targeting is evolutionary conserved. Altogether, the methylation of TEs and genes plays a prominent role in shaping genome size, structure and function; these effects on genomes from the genera *Arabidopsis*, *Oryza* and *Zea* will be discussed.

O-139

Long-term and Short-term Evolutionary Impacts of Transposable Elements on Drosophila

Grace Yuh Chwen Lee, Charles H. Langley
Center for Population Biology, University of California, Davis, Davis, CA, USA

Transposable elements (TEs) are ubiquitous genomic parasites and their interactions with hosts have long been likened to the coevolution between host and other nongenomic, horizontally transferred pathogens. TE families, however, are vertically inherited as integral segments of the nuclear genome. This transmission strategy has been theoretically suggested to weaken the selective benefits of host alleles that can repress TE transposition. On the other hand, the elevated rates of TE transposition and high incidences of deleterious mutations observed during the rare cases of horizontal transfers of TEs families between species could create at least a transient process analogous to the influence of horizontally transmitted pathogens. Here, we formally address such analogy using empirical and theoretical analysis to specify the mechanism of how host-TE interactions may drive the evolution of host genes. We found that TE-interacting host genes, particularly genes involved in the *piRNA* generation pathway, actually have more pervasive evidence of adaptive evolution than immunity genes that interact with nongenomic pathogens in Drosophila. Yet, both our theoretical modeling and empirical observations by comparing *D. melanogaster* populations before and after the invasion of *P elements* demonstrated that horizontally transferred TEs only have a limited influence on host TE-interacting genes. Our result suggested that the more prevalent and constant interaction with multiple vertically transmitted TE families should instead be the main force driving the fast evolution of TE-interacting genes, though with a fundamentally different mechanisms from that of the host-pathogen coevolution.

O-140

The epigenetics and evolution of TE control by piRNA: The role of dose.

Justin Blumenstiel, Dean Castillo, Chris Harrison, Christine Yoder, Kim Box, Jianwen Fang
University of Kansas, Lawrence, KS, USA

Transposable elements (TEs) are generally harmful genetic parasites that can cause mutation, shape genomes, and contribute to the architecture of gene expression networks. Historically, natural selection has been considered to play the key role in limiting TE proliferation in populations. However, recent studies have demonstrated that an adaptive system of genome defense by piRNA also limits TE proliferation. Using hybrid dysgenesis in *Drosophila virilis* as a model, we are examining how asymmetric inheritance of maternally provisioned piRNAs determines patterns of TE induced hybrid sterility. Our studies indicate that TE dosage likely plays a key role both in the induction of TE mediated hybrid sterility and in maternally provisioned protection against it. Furthermore, TE instability can be a general genomic property the can be propagated across generations, but repressed epigenetically. In light of the key role that dosage plays, I will present studies on the molecular evolution of the piRNA machinery that suggest a complex co-evolutionary dynamic between TEs and the machinery of genome defense. In particular, the dominant evolutionary response to increasing TE burden across the Drosophila genus seems to be improved translational efficiency in the piRNA machinery, not an increased rate of evolution.

O-141

Genomic satellite DNA repeats and small RNAs: bioinformatic, molecular and cytological evidence for the expression of *Responder* rasiRNAs in the *Drosophila melanogaster* testis.

Amanda Larracuente, Daven Presgraves
University of Rochester, Rochester, NY, USA

Responder (*Rsp*) is a satellite DNA repeat found in the pericentric heterochromatin of chromosome *2* in *Drosophila melanogaster*. *Rsp* is well-known for being the target of *Segregation Distorter* (*SD*)– a meiotic drive system found on chromosome *2* of *D. melanogaster*. *SD/SD*$^{+}$ heterozygous males transmit the *SD* chromosome to >95% of their progeny when the *SD*$^{+}$ chromosome bears a sensitive *Rsp* allele. *Rsp* copy number in the pericentric heterochromatin of chromosome *2* is positively correlated with the sensitivity to segregation distortion. The molecular relationship between *Rsp* repeat number and segregation distortion is not understood. Recently, RNAi has been suggested to have a role in Drosophila meiotic drive systems, including *SD*. Small RNAs corresponding to *Rsp* repeats are found in both female and male flies, consistent with the involvement of RNAi in *SD*. We conducted a bioinformatics study of genomic *Rsp* satellite repeats, and explore the possibility that *Rsp* short RNAs are expressed in the testis using a bioinformatic, molecular and cytological approach. We found several *Rsp*-like repeat families on all major chromosome arms in *D. melanogaster*. Although components of the *SD* system are assumed to be specific to *D. melanogaster*, we also find *Rsp*-like repeats in other Drosophila species. We present bioinformatic, molecular and cytological evidence for the expression of repeat associated small interfering RNAs (rasiRNAs) corresponding to *Rsp* in the adult testis. To determine if *Rsp* rasiRNAs correspond to the *Rsp* repeats on chromosome *2* targeted by *SD*, we mapped rasiRNAs to their respective genomic locations. We studied the spatial distribution of *Rsp* rasiRNAs in the testis using fluorescence *in situ* hybridization and find evidence for *Rsp* rasiRNA expression in both pre- and postmeiotic cell populations in a distinct pattern. The implication of these rasiRNAs in the *SD* system will be discussed.

O-142

Molecular Evolution of piRNA pathway genes in *Drosophila simulans* and *D. mauritiana*

Jeffrey Vedanayagam, Daniel Garrigan
University of Rochester, NY, USA

Piwi-interacting RNAs (piRNAs) defend against transposable elements (TEs) in the *Drosophila* germline. The piRNA pathway is a complex interplay of several genes whose function is essential for both piRNA biogenesis and TE silencing. Here, I analyze coding sequence polymorphism data from genes involved in piRNA pathway in the closely related species *D. simulans* and *D. mauritiana*. Application of the McDonald-Kreitman test, using *D. melanogaster* as the outgroup, finds that four of the genes show evidence of a significantly increased rate of amino acid substitution. *D. simulans* has an increased rate in Armitage (Armi) and Maelstrom (Mael), while *D. mauritiana* has an increased rate of amino acid substitution in the Aubergine (Aub) and Cutoff (Cuff) proteins. Armi and Mael are putative RNA helicase and nuclease respectively, while Cuff is a putative transcription termination factor. All three proteins are known to be crucial for piRNA biogenesis. Aub is a RNA-binding protein which is essential for ping-pong amplification of piRNAs. Analyses of the allelic frequency spectrum suggest that both Argonaute3 (Ago3) in *D. mauritiana* and Aub in *D. simulans* have an excess of low-frequency derived mutations and therefore may have experienced recent bouts of positive selection. Both Ago3 and Aub are Piwi family proteins that play a vital role in the ping-pong amplification of piRNAs. To identify the fixed amino acid substitutions contributing to species differences, I analyze the non-synonymous changes in these proteins that are fixed in one species and either absent (or present in low-frequency) in the other species. Also, I analyze whether the piRNA pathway is enriched for genes experiencing positive selection relative to a randomly chosen molecular pathway. These results provide candidate loci for introgression studies of the effect of rapidly diverging piRNA machinery on hybrid fitness in interspecific crosses.

O-143

Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory

Daniel Zivkovic, Pablo Duchen, Wolfgang Stephan
Biocenter, LMU Munich, Munich, Germany

The allele frequency spectrum has attracted considerable interest for the simultaneous inference of the demographic and adaptive history of populations. In a recent study, Evans et al. (2007) developed a forward diffusion equation describing the allele frequency spectrum in non-equilibrium populations under selection and mutation. We present here an explicit solution of their moment equations and of the allele frequency spectrum for the case of variable population size and neutrality. We also discuss the applicability of the theoretical results to the analysis of nucleotide polymorphism data using re-sequencing data from Drosophila.

References:

Evans, SN et al. 2007. Non-equilibrium theory of the allele frequency spectrum. TBP 71: 109-119
Živković, D, and W Stephan. 2011. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. TPB 79: 184-191

O-144

Approximate genealogical models for large-scale population genomic inference

Yun Song, Joshua Paul, Matthias Steinrücken, Kelley Harris, Sara Sheehan
University of California, Berkeley, California, USA

Key population-genetic applications involve inference under the coalescent and can be formulated in terms of the so-called conditional sampling distribution (CSD). Briefly, the CSD describes the probability of sampling an individual with a particular genomic sequence, having already observed a collection of individuals. We employ a sequentially Markov framework in conjunction with a previously introduced mathematical approximation to obtain a highly accurate approximation to the true CSD. Our approximate CSD, which admits a natural genealogical interpretation, is formulated as a hidden Markov model, yielding an algorithm with time complexity linear in both the number of loci and the number of haplotypes, thereby facilitating application on a genome-wide scale. In this talk, I will give an overview of the approach, and illustrate how past population size changes and subdivided population structure can be incorporated into our framework. The approach presented here can be readily adopted in a wide range of population-genomic applications, including genotype phasing and imputation, demographic and ancestral inference, and admixture mapping.

O-145

A new Markovian approximation to the coalescent with recombination.

Gerton Lunter
WTCHG, University of Oxford, Oxford, UK

Inferring population genetic parameters from whole-genome sequences requires dealing with the interplay between the coalescent process and recombination, as modeled by the Ancestral Recombination Graph (ARG). For long sequences, the ARG and its relatives, including Hudson's 1983 algorithm and Wiuf and Hein's sequential algorithm, are slow even for simulations. The reason is that each of these approaches operate on a large graph, encoding both coalescence events and recombination events across a range of loci. In 2005, McVean and Cardin introduced the Sequentially Markov Coalescent (SMC), an approximation that operates on a single local tree. It is sufficiently efficient to be used for inference on whole genomes, and Li and Durbin used this model in 2011 to infer from a single diploid genome, how human effective population size has changed through time.

The SMC model is a "fixed" approximation: it does not have a parameter to improve the approximation where needed. Here I introduce a new sequential model that is more efficient than both Hudson's and Wiuf and Hein's algorithm, and admits a range of approximations, including the SMC model as a special case.

The model differs from previous models in that the current state is formulated in terms of genealogical relationships at a single locus, rather than a graph describing both coalescence and recombination events. States take the form of a forest of trees, one of which contains the marginal genealogy of the sampled sequences as a sub-tree. The model is Markovian, and in contrast to Hudson's and Wiuf and Hein's algorithm, states of the model only contain "local" information relating to the current position in the sequence.

The model is exact in the sense that it produces distributions of marginal genealogies that are identical to the ARG. It is arguably conceptually simpler than Wiuf and Hein's algorithm, and encodes less information in its states than either Wiuf and Hein's or Hudson's algorithm so is likely to be faster, but is still computationally costly. A range of approximations are obtained by pruning selected trees and non-contemporary tips, and the algorithms so obtained are much more efficient than the full algorithm, as well as including very good approximations. The SMC model is obtained by aggressively y pruning all branches and trees that are not part of the marginal genealogy. Applications of the model include accurate simulation of long sequence, and inference.

O-146

The Ties That Bind: Gene genealogies within a fixed pedigree

Sohini Ramachandran[1], Leandra King[2], Peter Wilton[2], John Wakeley[2]
[1]Brown University, Providence, RI, USA, [2]Harvard University, Cambridge, MA, USA

The evolutionary forces of reproduction, migration, mutation, selection, and recombination, as well as events such as population bottlenecks, produce the genome-wide patterns of genetic variation we observe within any given population or species. It is well known that the structure of genetic variation is mediated by gene genealogies, which are the ancestral genetic relationships among samples of genetic data. Coalescent theory, the backward-time approach to population genetics, provides a conceptual framework for describing the gene genealogy of a sample and making predictions about patterns of genetic variation.

We address a conceptual flaw in coalescent theory as it is applied to diploid bi-parental organisms. In diploid bi-parental organisms, the structure of gene-genealogies is mediated by the population pedigree. Population pedigrees — the set of all family relationships among members of the population — are modeled as random quantities, but really they should be taken as given, having been fixed by past events. Gene genealogical models should describe the outcome of the percolation of genetic lineages through the population pedigree according to Mendelian inheritance. Yet, the effects that population pedigrees have on gene genealogies are largely unknown.

We study the differences between the fixed-pedigree coalescent and the standard coalescent by analysis and using simulated pedigrees, some of which are based on human family data. Differences are apparent in recent past generations, until the numerous ancestors of the sample likely overlap completely, at which point the standard coalescent provides a surprisingly accurate description of gene genealogies on a fixed pedigree. Because all loci in the genome share the same population pedigree, it is especially important to condition on the population pedigree in multi-locus population genetics. We also investigate the effects of pedigree structure on genetic variation during a population bottleneck, and consider loci that undergo recombination. As human ancestors went through multiple bottlenecks during the diaspora out of Africa, and through exponential growth in the last 10,000 years, pedigree structure will be especially important in studying human evolutionary history.

O-147

Emergence and global spread of seventh pandemic cholera

Julian Parkhill
The Sanger Institute, Cambridge, UK

Vibrio cholerae is the cause of cholera, and has been responsible for seven historical pandemics. Although *V. choleae* is genetically and antigenically diverse, only isolates of serogroup O1 (consisting of two biotypes known as 'classical' and 'El Tor') and its derivative O139 can cause epidemic cholera. It is believed that the first six cholera pandemics were caused by the classical biotype, while the seventh has been caused by El Tor, which has spread globally and replaced the classical biotype. Previous attempts to understand the transmission of cholera, and answer the question of whether novel outbreaks are caused by local arisal of extant strains or new introductions, have been limited by the use of sub-genomic typing techniques based on mobile elements, obscuring the true relationships among strains. To understand the underlying phylogeny of the lineage responsible for the current pandemic, we generated whole genome sequences for 136 isolates from a global collection spanning 70 years. The results show that seventh pandemic cholera is monophyletic, that there have been multiple wave of transmission from a common source in South Asia, and that long-distance transmission events are ongoing.

O-148

Genome-wide association mapping in *Campylobacter*: genetic signals of host adaptation.

Samuel Sheppard[1,2], Xavier Didelot[2], Daniel Falush[3]
[1]Swansea University, Swansea, UK,[2]Oxford University, Oxford, UK,[3]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

The ability to adapt to multiple host species is a fundamental property associated with the emergence of zoonotic pathogens such as *Campylobacter.* Advances in whole genome sequencing are providing new opportunities for improved understanding of bacterial epidemiology and evolution. However, while completion of the first 1000 *Campylobacter* genomes is now imminent, developing methods for analyzing this data remains a significant challenge. Consistent with techniques designed to meet similar challenges in human genome sequencing projects, we have developed an association mapping approach to identify genetic elements associated with particular phenotypes. 40 *C. jejuni* from the multihost ST-45 clonal complex were isolated from chicken and cattle. The genomes of these isolates were sequenced and divided into overlapping 30bp words to allow simultaneous analysis of homologous and non-homologous sequence variation. Words that were significantly overrepresented in one or other host, compared to expectation based upon the population structure, were considered as candidates for adaptation and mapped onto an annotated reference genome to investigate their functionality. Ten genomic regions were shown to segregate strongly by host and 75% of the words mapped to a single large host-associated region containing 16 genes. Greatly reduced sequence variation in this region in isolates from cattle provided evidence for selection. Seeking the host-associated words of the ST-45 complex in another 173 *Campylobacter* genomes, from other clonal complexes and sources (chicken, cattle, pigs, wild birds, clinical samples), revealed that many of them were also host-associated in these genomes. Furthermore, there was evidence that genetic elements associated with cattle and chicken were more common in isolates from other mammals and birds respectively. Using this association mapping technique we identified some of the genetic changes associated with the adaptation of this important human pathogen from its ancestral host (birds) to mammals.

O-149

Tracing the emergence, transmission and adaptation of pandemic MRSA

Paul R. McAdam[1], Kate E. Templeton[2], Giles F. Edwards[3], Matthew T. G. Holden[4], Edward J. Feil[5], David Aanenson[6], Mark C. Enright[7], Anne Holmes[2], E. Kirsty Girvan[3], Paul A. Godfrey[8], Michael Feldgarden[8], Angela M. Kearns[9], Andrew Rambaut[10], D. Ashley Robinson[11], J. Ross Fitzgerald[1]
[1]Roslin Institute, University of Edinburgh, Edinburgh, UK, [2]Microbiology, Royal Infirmary of Edinburgh, Edinburgh, UK, [3]Scottish MRSA Reference Laboratory, Glasgow, UK, [4]Wellcome Trust Sanger Institute, Hinxton, UK, [5]University of Bath, Bath, UK, [6]Imperial College, London, UK, [7]AmpliPhi Biosciences, Bedfordshire, UK, [8]Broad Institute, Massachusetts, USA, [9]Microbiology Services Division, Health Protection Agency, London, UK, [10]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK, [11]Department of Microbiology, University of Mississippi Medical Center, Mississippi, USA

Hospital-associated infections caused by methicillin-resistant Staphylococcus aureus (MRSA) are a global health burden dominated by a small number of specialized clones. For example, the pandemic EMRSA-16 clone was widespread in UK hospitals for 20 years but its evolutionary origin and the molecular basis for its hospital association are unknown. We carried out a Bayesian phylogenetic reconstruction based on the genome sequences of 89 S. aureus isolates including 60 EMRSA-16 and 29 additional clonal complex 30 (CC30) isolates from patients in 3 continents over a 53 year period. The 3 major pandemics to originate from the CC30 lineage, including phage type 80/81, South-West Pacific, and EMRSA-16, shared a most recent common ancestor that existed about 150 years ago while the hospital-associated EMRSA-16 clone is estimated to have emerged about 35 years ago, almost 2 decades prior to its first description. In order to identify genetic events which contributed to the hospital association of EMRSA-16, we carried out a genome-wide comparison of EMRSA-16 isolates and basal CC30 community-associated strains. Our analyses reveal several non-synonymous mutations in loci affecting antibiotic resistance and virulence. as well as differences in mobile genetic element content which contribute to community or hospital specialisation. Importantly, phylogeographic analysis indicates that EMRSA-16 spread within the UK by transmission from hospitals in the major population centers of London and Glasgow to regional health-care centers, implicating patient referrals as an important cause of intra-national transmission. Taken together, the high-resolution phylogenetic approach employed resulted in a new understanding of the emergence and transmission of a major MRSA clone and provided molecular correlates of its hospital specialization. Similar approaches for hospital-associated clones of other bacterial pathogens may inform appropriate measures for controlling intra- and inter-hospital spread of infection.

O-150

The population structure and genome architecture of Staphylococcus aureus

Richard Everitt, Bernadette Young, Ruth Miller, Antonina Votintseva, Rory Bowden, Kyle Knox, Derrick Crook, Daniel Wilson
University of Oxford, Oxford, UK

Whole genome sequencing provides an unprecedented opportunity to characterize patterns of genetic diversity and genome evolution in natural populations of bacteria. Staphylococcus aureus is a major hospital-associated human pathogen that has developed resistance to a diverse range of antimicrobials. Understanding the genetic basis of virulence and pinpointing dominant modes of transmission are translational priorities that require an intimate understanding of population structure. Despite the notoriety of this pathogen, natural populations of S. aureus are predominantly commensal, colonizing the nose of 28% of healthy adults, causing disease infrequently relative to the prevalence of carriage. In this study we used Illumina sequencing to characterize the genomes of 100 S. aureus isolated from asymptomatic nasal carriers in Oxfordshire, United Kingdom. I will discuss the diversity and structure of this population, including the frequency of genes known to encode virulence and drug resistance. I will also investigate the recent evolution of the S. aureus core and accessory genome, including the distribution of mobile genetic elements, and examine the evolutionary forces that have helped shape the contemporary population.

O-151

Evolutionary dynamics of bacteria in a human host environment

Rasmus Lykke Marvig, Lars Jelsbak, Søren Molin
Technical University of Denmark, Lyngby, Denmark

Genome-wide characterization of the molecular change in evolving bacterial populations is important in our understanding of microbial evolution. Here we detect mutations that accumulate in the same pathogenic strain during infection of multiple patients. We conducted a retrospective study of the dissemination of the DK2 lineage of *Pseudomonas aeruginosa* among patients with cystic fibrosis, sequencing the genomes of 44 isolates collected from 16 individuals over 35 years. Phylogenetic analysis of the single nucleotide polymorphisms (SNPs) revealed a network of transmission between patients. Isolates from the same patient tended to form monophyletic groups, exemplified by one patient colonized with a single homogeneous sub-lineage over 18 years of infection. However, we also observed a case with a patient colonized with three different sub-lineages each being dominant over several years. Despite an overall molecular signature of genetic drift, we identified both known and novel genes to be involved in host adaptation by tracking recurrent patterns of mutations in the same bacterial genes. A low sign of homoplasy underscores vertical genomic inheritance to be dominant during the long-term infection of the DK2 lineage, but a single event of horizontal gene transfer from a co-existing but distant strain of *P. aeruginosa* was also observed. Several sub-lineages evolved as hypermutators due to mutations in the DNA mismatch repair system. The hypermutators had increased mutations rates with distinct mutational signatures that might be of importance for the genetic adaptation of the pathogen to its host. In conclusion genome sequencing of longitudinal *P. aeruginosa* isolates from chronic infections can serve as a system to study evolutionary mechanisms *in vivo*.

O-152

Evolvability and the space of possible metabolisms

Andreas Wagner
University of Zurich, Zurich, Switzerland

Spaces of the possible', such as ‘morphospace' and ‘protein space', have long been of interest to evolutionary biologists. A thorough analysis of such spaces is usually prevented by ignorance about how information in genotypes gives rise to phenotypes. In recent years, new computational methods have alleviated this problem in some classes of systems, such as regulatory circuits and metabolic reaction networks, and made it possible to study the space of the possible in these systems. My focus in this talk is on metabolic systems, complex systems of chemical reactions that sustain all of life. I will discuss recent insights into the organization of the space of possible metabolism, and on aspects of this organization that may affect an organism's ability to bring forth novel phenotypes.

O-153

Bi-stability in proteins as a potential evolutionary response to adaptive conflicts

Tobias Sikosek[1], Hue Sun Chan[2], Erich Bornberg-Bauer[1]
[1]Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany, [2]Departments of Biochemistry, Molecular Genetics, and Physics, University of Toronto, Toronto, ON, Canada

Many organisms live under complex and changing environmental conditions, while having a limited number of proteins to deal with these conditions. Multi-functionality, as exhibited by many functionally promiscuous enzymes, has been hypothesised as an advantageous compromise whenever the same protein is under selection to conserve an existing function while adapting towards a new function (adaptive conflict). A stage of multi-functionality may or may not be followed by gene duplication and divergence. We use computational biophysical models to analyse multi-functionality of proteins that can fold into more than one stable structure (using structure formation as a proxy for functionality). Our model predicts that proteins evolving under selection for two alternative structures can follow gradients of stability shift from the formation of only one stable structure towards an equilibrium state between two stable structures (bi-stability). Population dynamics simulations show that weak conflicting selection pressures may be sufficient to direct protein evolution towards bi-stability. Our results also suggest that models of protein evolution may underestimate evolvability if they do not account for bi-stability. However, while bi-stable proteins provide many more mutational connections to other protein structure phenotypes in genotype space, they are also less stable. This shows the inherent conflict between conservation of structure (by maximising stability), and adaptation towards new structures (which requires some destabilisation). Bi-stable proteins may provide the necessary compromise. Furthermore, bi-stable proteins may provide an additional advantage after gene duplication, because they provide excellent starting points for subfunctionalisation (functional divergence driven by adaptation and/or genotype space entropy), as consistent with the recently proposed Escape from Adaptive Conflict model. The potential for increased evolvability due to bi-stable proteins is thus two-fold by allowing adaptation before and after gene duplication.

Publications:

- Bornberg-Bauer E, Chan HS (1999) Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc Natl Acad Sci U S A 96: 10689–10694.
- Bornberg-Bauer E, Huylmans A-K, Sikosek T (2010) How do new proteins arise? *Curr Opin Struct Biol* 20:390-396.
- Sikosek T, Chan HS, Bornberg-Bauer E (2012) Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness.
- Sikosek T, Bornberg-Bauer E, Chan HS (2012) Evolutionary dynamics on bi-stability landscapes have the potential to resolve adaptive conflicts.

O-154

Enzyme flux control predicts polymorphism, divergence between species, and fitness in the field

Carrie F. Olson-Manning, Kasavajhala Prasad, Bao-Hua Song, S. Thomas Mitchell-Olds
Duke University, Durham, North Carolina, USA

In biochemical pathways, the widespread occurrence of genetic dominance suggests that pathway flux is robust to many changes in catalytic activity. However, this is contrary to a growing body of evidence that one or a few enzymes in a metabolic pathway are the primary targets of natural selection. Studies in diverse systems (plants and animals) have found a correlation between enzyme pathway position and rates of sequence evolution. These analyses have found that enzymes at the beginning or at branch-points of pathways may show signatures of strong purifying selection or positive selection, in comparison to other enzymes in the same pathway. The pattern depends on whether the pathway is at an optimum or is under directional selection for a new optimum, respectively, but both are consistent with a pathway with uneven flux control. This correlation suggests that key enzymes in metabolic pathways show strong flux control and low robustness to mutations, and therefore have a higher capacity for evolutionary change. However, the factors responsible for this correlation between pathway position and evolutionary rate have not been tested.

Here we empirically perturb the enzyme activity of each step in an important plant defense pathway in *Arabidopsis thaliana* and measure the resulting change in the metabolic phenotype. McDonald-Kreitman analyses show that the enzyme with the highest flux control displays a significant signature of non-neutral protein evolution. Elsewhere in the pathway, biosynthetic flux is robust to mutations that severely disrupt enzyme function, and sequence signatures are consistent with neutral evolution. We also study the evolution of this plant defense pathway in a closely related species, *Boechera stricta*, and show that natural selection has driven the evolution of a novel catalytic function at the same enzyme that shows majority control over flux in *A. thaliana*. This suggests that enzyme control over flux and defensive phenotype is consistent over millions of years, and that the same enzymes may be exploited multiple times by natural selection. Finally, with ancestral state reconstruction and *in vitro* assays, we explore the causal mutations responsible for this novel activity, to test how many mutations are required to evolve the novel function.

Our results suggest an interesting emergent property of biochemical pathways: variation in sequence signatures of selection is predicted by functional control over pathway flux, and the enzymes with greatest control over phenotype are most likely to be exploited by natural selection.

O-155

A comparative evolutionary model for the study of meiotic mechanisms

Joshua Bayes[1,2], Clara Wang[1], Abby Dernburg[1,2]
[1]Howard Hughes Medical Institute, Berkeley, CA, USA, [2]Univ. of California, Berkeley, Berkeley, CA, USA

Meiosis is a specialized cell division process that reduces a diploid cell to haploid gametes. This evolutionary innovation underlies sexual reproduction and genetic diversity. Conserved features of the meiotic program include homologous chromosome recognition (pairing), establishment of the synaptonemal complex (SC) between paired homologs (synapsis), and formation of physical linkages, or chiasmata, between paired chromosomes through crossover recombination. Despite its important function and wide conservation, many meiotic mechanisms and genes diverge rapidly.

To address why these differences exist and how evolution has tailored the mechanistic details of meiosis among different species, we have initiated efforts to develop the satellite species *Pristionchus pacificus* as a comparative model of meiosis. Cytologically, meiosis in *P. pacificus* closely resembles the progression observed in *C. elegans*, but these nematodes are separated by a large evolutionary distance. *In silico* searches of the *P. pacificus* genome using known meiotic protein sequences from *C. elegans* and other genera have revealed interesting avenues for investigation. The *P. pacificus* genome lacks apparent orthologs of many genes that mediate homolog pairing and synapsis in *C. elegans,* including the pairing center zinc finger proteins characterized by our laboratory. More surprisingly, meiotic recombination mechanisms in *P. pacificus* may also contrast with the pathways elucidated in *C. elegans*. *P. pacificus* appears to express two members of the RecA recombinase superfamily: RAD-51, which is required for somatic DNA repair and meiotic recombination, and the meiosis-specific DMC-1 protein, which is absent in *C. elegans.* Moreover, *P. pacificus* contains genes encoding members of the Hop2 and Mnd1 protein families, which appear to have been lost in *C. elegans*.

To identify novel or highly divergent genes involved in pairing, synapsis, and recombination in *P. pacificus*, a “High incidence of males” screen was carried out. Mutants identified in this screen are currently being characterized using cytological tools I have developed in *P. pacificus*. Additionally, the types of selection acting on meiotic components are being inferred by comparing orthologous meiotic gene sequences between closely related *Pristionchus* species. Ongoing work will provide insight into the evolution of meiotic mechanisms over long- and short-term evolutionary time scales.

O-156

Sex can restrict evolvability

Ricardo Azevedo[1], Tiago Paixão[1,2]
[1]*University of Houston, Houston, Texas, USA,* [2]*Institute of Science and Technology, Vienna, Austria*

Evolution is the movement of populations through a space of genotypes. This space can be modeled as a network connecting genotypes that can be reached through mutation. In this view, the mutational robustness of a genotype is the proportion of its mutational neighbors that are viable. Robustness can facilitate the exploration of genotype networks, or evolvability. Sexual reproduction is also widely believed to promote evolvability, for two reasons. First, because it allows long jumps through genotype space. Second, because it selects for mutational robustness. Here, we show that, depending on the structure of the genotype network, sexual reproduction may not select for the highest mutational robustness, and can actually reduce evolvability.

O-157

Orthology: a simple concept marred in complexity but still holding

Eugene Koonin
National Center for Biotechnolgoy Information, NIH, Bethesda, MD, USA

Accurate inference of orthologous genes is the cornerstone of comparative genomics.

Orthology is a very simple concept in principle: a set of orthologs is simply the set of descendants of the same ancestral gene. However, in the real world, orthologous relationships between genes are dramatically complicated by several major evolutionary processes including lineage-specific gene duplication (paralogy); differential loss of paralogs in different lineages; horizontal gene transfer (xenology); and gene fusion and fission. In prokaryotes, horizontal gene transfer is the main culprit leading to the wide spread of obvious or hidden xenology that is detectable in the history of nearly all genes. In eukaryotes, the principal factor complicating orthologous relationships is lineage-specific accretion of protein domains which requires a new, dynamic approach to the definition and identification of orthology. All these complications notwithstanding, gene (or domain) orthology remains tractable and essential for evolutionary genomic studies. It is usually tacitly assumed that orthologs perform 'the same' biological function in different organisms although the definition of orthology is purely evolutionary and does not directly imply conservation of function. The conjecture on the functional conservation among orthologs is central both to the understanding of the fundamental connections between genome and phenome evolution, and more practically, for genome annotation. Recently, this conjecture has been challenged by the demonstration that at the same level of divergence paralogs appear to show more functional similarity than orthologs as inferred primarily from Gene Ontology annotation. A comprehensive, genome wide analysis of expression profiles of orthologous and paralogous genes in organisms from diverse taxa will be presented shedding light on this critical problem.

O-158

Changes in gene expression following segmental duplication in mammals

Katerina Guschanski, Julien Meunier, Henrik Kaessmann
University of Lausanne, Lausanne, Switzerland

Gene duplications are powerful drivers of evolution that allow novel functions to emerge from a state of initial genetic redundancy. They were shown to be abundant in many organisms and to have contributed to their phenotypic diversity. Here, we focus on genes that emerged through DNA-based (segmental) duplication within mammals and study the evolution of their expression patterns in representatives of all major mammalian lineages; placental mammals, marsupials and the egg-laying monotremes. Using gene family trees retrieved from the Ensembl database, we identified thousands reliable duplication events belonging to almost three thousand different gene families. Relying on a unique transcriptome dataset generated for six organs from eight mammalian species and a bird (the evolutionary outgroup) using RNA sequencing, we computed expression profiles for all protein-coding genes in these families. This allowed us to not only compare expression profiles between the paralog sets but also to study expression profiles of non-duplicated orthologs within the same gene family. We observe that mammalian lineage-specific duplicated genes have on average lower expression levels and are more tissue-specific than their non-duplicated orthologous counterparts. After duplication, sets of paralogs tend to diverge in their expression levels, with one set of paralogs being expressed at a higher level and in fewer tissues (i.e., with a reduced expression breadth) than the other. This difference increases with the age of the duplication and is highest in old duplications that happened before the emergence of mammals, moderate in mammals, therians and eutherians, and lowest in younger duplications specific to the primate lineage. Genes within paralog sets tend to conserve tissue-specific expression profiles: correlations of tissue-specific expression profiles are higher within paralog sets than between them. Again, this difference is age-dependant. However, changes in tissue-specific expression and in expression levels occur independently. In summary, using newly generated transcriptome data, we provide initial insights into the expression evolution of segmentally duplicated genes in mammals. Our first results suggest that divergence between paralogs becomes more pronounced with evolutionary time and can occur through expression level changes, modifications of tissue-specific expression profiles, and changes in expression breadth. These changes can occur together or individually, thus providing a variety of modification mechanisms that may facilitate the fixation of duplications.

O-159

Exploring The Evolution Of Novel Enzyme Functions Within Structurally Defined Protein Superfamilies

Nicholas Furnham[1], Ian Sillitoe[2], Gemma Holliday[1], Alison Cuff[2], Roman Laskowski[1], Christine Orengo[2], Janet Thornton[1]
[1]EMBL-EBI, Cambridge, UK,[2]University College London, London, UK

Understanding of how enzymes have evolved to undertake the wide variety of reactions they perform is essential to many studies in biology and medicinal chemistry. To unravel this problem requires the combination of protein three-dimensional structural, sequence, phylogenetic and chemistry information. We have combined this variety of data in an automatic pipeline for investigating enzyme functional evolution within structurally defined protein superfamilies. The results of this pipeline, called FunTree, can be accessed at http://www.ebi.ac.uk/thornton-srv/databases/FunTree/.

The development of this resource has permitted us to analyze 276 enzymatic superfamilies cataloged by the CATH database. In addition to showing relationships between structures and sequences though phylogeny, we are able to show relationships of the small molecule metabolites the enzymes are acting on. To demonstrate the power of this approach in identifying some of the unique features of enzymes we use a number of distinct superfamilies including the phosphatidylinositol-phosphodiesterase (PPD), the phylogenetic tree of which reveals a number of distinct clades corresponding to distinct changes in overall reactions and reaction mechanisms.

By bringing together so much data, we can provide a comprehensive overview of the most common and rare types of changes in function as well as providing a framework for detecting changes in function between orthologs and paralogs. Our analysis demonstrates on a larger scale than previously studied, that modifications in overall chemistry still occur, with all possible changes at the primary level of the Enzyme Commission (E.C.) classification observed to a greater or lesser extent. The phylogenetic trees map out the evolutionary route taken within a superfamily, as well as all the possible changes within a superfamily. This has been used to generate a matrix of observed exchanges from one enzyme function to another, revealing the scale and nature of enzyme evolution and that some types of exchanges between and within E.C. classes are more prevalent than others. Surprisingly a large proportion (71%) of all known enzyme functions are performed by this relatively small set of superfamilies. This reinforces the hypothesis that relatively few ancient enzymatic domain superfamilies were progenitors for most of the chemistry required for life. Moreover, by using new tools that can automatically compare enzyme reactions solely based on changes in valence and bond orders as well as metabolite sub-structure similarities we can provide a robust means of identifying conservation in reactions, thus providing a means to improve function prediction and contribute to the design of novel enzyme functions.

O-160

Experimental annotations in 13 model organisms corroborate ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs

Adrian Altenhoff[1,2], Romain Studer[2,3], Marc Robinson-Rechavi[2,3], Christophe Dessimoz[1,4]
[1]ETH Zurich, Zurich, Switzerland, [2]Swiss Institute of Bioinformatics, Zurich, Switzerland, [3]University of Lausanne, Lausanne, Switzerland, [4]EMBL-EBI, Hinxton, UK

The function of most proteins is not determined experimentally, but is extrapolated from homologs. According to the "ortholog conjecture", or standard model of phylogenomics, protein function changes rapidly after duplication, leading to paralogs with different functions, while orthologs retain the ancestral function. But despite its prevalence, this model mostly rests on first principles, as for a long time we have not had sufficient data to test it empirically. Recently, some studies began investigating this question and have cast doubt on the validity of this model. In this talk, we report that a comparison of experimentally supported functional annotations among homologs from 13 genomes mostly supports this model. We show that to analyse GO annotation effectively, several confounding factors need to be controlled: authorship bias, variation of GO term frequency among species, variation of background similarity among species pairs, and propagated annotation bias. After controlling for these subtle biases, we observe that orthologs have generally more similar functional annotations than paralogs. This is especially strong for sub-cellular localisation. We observe only a weak decrease in functional similarity with increasing sequence divergence. These findings hold over a large diversity of species; notably orthologs from model organisms such as E. coli, yeast or mouse have conserved function with human proteins.

O-161

A comparative analysis of orthologs and paralogs in Animals: from sequence to structure.

Romain Studer, Christine Orengo
University College London, London, UK

The selective regimes in a gene family can be divided into three different types: purifying selection; neutral evolution and positive selection, where amino acid changes are selected for their new role in the protein and are thus preserved in the genome. Positive selection can be transient and it affects only a small subset of amino acids sites. Substitution codon models and amino acid replacement rate models can describe this mechanism quantitatively.

In previous studies, work undertaken in the Group of Marc Robinson-Rechavi (Studer et al., Genome Res 2008, Studer et al., Mol Biol Evol 2010), we identified amino acid sites that are suggested to promote functional divergence between orthologs and between paralogs. The dataset was based on hundreds and sometimes thousand of genes families in Animals (mainly Vertebrates and Insects). These sites have been detected using branch-site codon models for nucleotides (from PAML, developed by Ziheng Yang's group) and covarion-like and constant-but-different substitutions models for amino acids. While our initial question was to assess the role of duplication in shaping genomes, we found no significant differences between the evolutionary rate of orthologs and paralogs. This suggests that if any differences exist, it happens elsewhere than in the primary sequence.

In the present study, we investigated the tertiary level of proteins, as the biochemical function is directly linked to the 3D structure. Changes in the structure, caused by mutations, will probably change the characteristics of the biochemical function, such as the enzymatic activity or interaction partners. Working on these datasets, we analysed these sites inside the structure of proteins and estimated their potential effect on the evolution of protein function. Inside the protein, we identified various patterns of positively selected sites, ranging from isolated sites to big clusters of sites, suggesting a correlated pattern of mutation under functional diversification. These results are put in the context of orthology and paralogy.

O-162

Transposable element mediated gene regulatory network innovation: the evolution of pacental mammals

Gunter Wagner, Deena Emera, Vincent Lynch
Yale University, New Haven, USA

Evolutionary innovation is, at the molecular level, caused by innovations at level of gene regulatory networks. Evolution of gene regulation takes place on two principal levels, the level of cis-regulatory elements and at the level of transcription factors and other trans-acting factors such as protein modifying enzymes and receptors. Here we focus on the role of transposable elements in creating novel cis-regulatory elements. Comparative transcriptomic data shows that major transitions in the evolution of mammalian reproduction are associated with the recruitment of hundreds of gene into the cells of the endometrium, the cell layer covering the inside of the uterus. Detailed studies of some examples of recruited gene show that the derived CREs often if not always derive from transposable elements. We find two modes of CRE evolution. One which affects the chromatin state and thus changes the gene regulatory “landscape” on the chromosome and one where a transposable element gets modified until it turns into either an enhancer or a promoter. In the latter case we observe a mode of molecular evolution that may be more broadly applicable: *epistatic capture*. We call epistatic capture the process by which a transcription factor binding site, that was present in the ancestral transposable element and which is variable up to a certain point in phylogeny, gets fixed in a derived clade by epistatic interaction with derived transcription factor binding sites. We observe this mode of evolution in the PRL promoter in the primate lineage, where an ancestral ETS1 site is variable in primates outside the apes, but interacts strongly with derived binding sites within the apes, supposedly fixing its presence after the acquisition of the derived binding sites.

O-163

Searching the genetic basis underlying the evolution of the human brain.

Lucía Franchini
INGEBI-CONICET, Buenos Aires, Argentina

It has been hypothesized that the evolution of the unique human cognitive capacities is due to the acquisition of new temporal and spatial expression patterns of preexisting genes rather than changes in the protein-coding sequences. Using a combination of bioinformatics and functional studies including the generation of transgenic zebrafish and mice we are investigating differences in gene regulation which may have contributed to the evolution of the human brain.

I will present here our results involving the functional characterization of the largest cluster of the most rapidly evolving human elements yet identified [termed human accelerated elements (HAEs)] located within 648 kb of the *neuronal PAS domain-containing protein 3 (NPAS3)* gene. We tested the ability of the 14 *NPAS3*-HAEs to function as developmental enhancers using a transposon-based transgenic assay in zebrafish. Our results indicated that 9 HAEs behave as developmental enhancers. The finding that the *NPAS3* gene shows a set of highly conserved regulatory regions that evolved faster in the human lineage suggests it might have acquired a new expression pattern and probably a novel function. In order to test this hypothesis, we performed a comparative expression analysis over selected *NPAS3* elements in transgenic mice. We found that one of the selected HAEs acts as a neurodevelopmental enhancer driving the expression of the reporter gene lacZ between E9.5 and E14.5 in a subdomain of the *NPAS3* expression pattern in mice. We also observed that the human ortholog of this HAE shows a new expression territory in the developing cortex. The *NPAS3* gene is a transcription factor of the bHLH-PAS family that is broadly expressed in the developing mouse nervous system playing an important role in normal brain development and neurosignaling pathways. In addition, its dysfunction has been associated with schizophrenia in humans. Our results indicated that the regulation of *NPAS3* has been shaped during human evolution. In addition, our data suggest that changes in the expression pattern acquired by at least one of the *NPAS3* regulatory regions could have contributed to changes in the spatio-temporal expression of *NPAS3* in the human lineage and, therefore, to play unique roles during human brain development and mental illness.

O-164

Selective Sweep of a cis-Regulatory Sequence in a Non-African Population of Drosophila melanogaster

Sarah Saminadin-Peter, Claus Kemkemer, Pavlos Pavlidis, John Parsch
University of Munich, Munich, Germany

Although it is thought that changes in gene expression play an important role in adaptation, the identification of gene-regulatory sequences that have been targets of positive selection has proven difficult. Here we identify a cis-regulatory element of the Drosophila melanogaster CG9509 gene that is associated with a selective sweep in a derived, non-African population of the species. Expression analyses indicate that CG9509 consistently shows greater expression in non-African than in African strains of D. melanogaster. We find that a 1.8-kb region located just upstream of the CG9509 coding region is devoid of DNA sequence polymorphism in a European population sample and that this is best explained by the recent action of positive selection (within the past 4,000–10,000 years). Using a reporter gene construct and phiC31-mediated site-specific integration, we show that the European version of the CG9509 upstream region drives 2-3 times greater expression than the African version in an otherwise identical genetic background. This expression difference corresponds well to that of the native gene and indicates that sequence variation within the CG9509 upstream region can completely account for its high expression in the European population. Selection appears to have favored a quantitative increase in gene expression in the Malpighian tubule, the tissue where CG9509 is predominantly expressed.

O-165

Predicting and Testing Human-specific Developmental Enhancers

John Capra, Genevieve Erwin, Katherine Pollard
Gladstone Institutes, University of California, San Francisco, San Francisco, CA, USA

The dramatic diversity of form and function found between closely-related species is likely driven by changes to non-coding DNA that modify the complex patterns of gene expression observed throughout development. Indeed, in the human genome, the vast majority of regions that exhibit human-specific accelerated evolution (HARs), are found in non-coding contexts and are enriched nearby transcription factors and developmental genes. Detailed study of HAR2, one of the most accelerated HARs, demonstrated that it is a developmental enhancer that produces unique expression patterns in human compared to other primates. These observations suggest that many of the most accelerated regions of the human genome may function as developmental enhancers.

To investigate this hypothesis, we combined computational analysis of functional genomics data with enhancer assays in transgenic mice. First, we developed a machine learning algorithm that predicts tissue-specific developmental enhancers. The algorithm integrates DNA sequence data, cell type specific histone modifications, data on chromatin state, transcription factor (TF) binding sites, and gene expression. We predicted enhancers across the human genome and intersected these predictions with the HARs to create a list of candidate human-specific enhancers. We prioritized regions on this list for further study using a method we developed for predicting functional divergence in non-coding sequence from TF binding site divergence between human and chimp. We then tested more than 20 of the predicted enhancers in transgenic mice with the human or chimp sequences. We found several human-specific neural and cardiac enhancers. These examples highlight how enhancer activity has evolved since divergence from our last common ancestor with chimp. They are exciting candidates for understanding the genetics of human-specific traits. This study provides a novel approach to modeling the functional effects of non-coding sequence changes between species.

O-166

Comparative Genomics Reveals Birth and Death of Fragile Regions in Mammalian Evolution

Pavel Pevzner
University of California at San Diego, La Jolla, CA, USA

An important question in genome evolution is whether there exist fragile regions (rearrangement hotspots) where chromosomal rearrangements are happening over and over again. We demonstrate that fragile regions are subject to a ``birth and death'' process, implying that fragility has limited evolutionary lifespan. This finding implies that fragile regions migrate to different locations in different mammals, explaining why there exist relatively few chromosomal breakpoints shared between distant branches of the evolutionary tree. The birth and death of fragile regions phenomenon reinforces the hypothesis that rearrangements are promoted by matching segmental duplications and suggests putative locations of the currently active fragile regions in the human genome.

This is a joint work with Max Alekseyev at University of South Carolina.

O-167

A GENOME-WIDE ANALYSIS OF Common Fragile Sites: WHAT FEATURES DETERMINE CHROMOSOMAL instability IN A MAMMALIAN GENOME?

Arkarachai Fungtammasan, Erin Walsh, Francesca Chiaromonte, Kristin Eckert, Kateryna Makova
Penn State University, University Park, USA

Chromosomal Common Fragile Sites (CFSs) are unstable genomic regions that break under replication stress and are involved in structural variation. They frequently become sites of chromosomal rearrangements in cancer and of viral integration. However, CFSs are under-characterized at the molecular level and thus difficult to predict computationally. Newly available genome-wide profiling studies provide us with an unprecedented opportunity to associate CFSs with features of their local genomic contexts. Here we contrasted the genomic landscape of cytogenetically defined human aphidicolin-induced CFSs (aCFS) to that of non-fragile sites, using multiple logistic regression. We also analyzed aCFS breakage frequencies as a function of their genomic landscape, using standard multiple regression. We show that local genomic features are effective predictors both of regions harboring aCFSs (explaining approximately 77% of the deviance in logistic regression models) and of aCFS breakage frequencies (explaining approximately 45% of the variance in standard regression models). In our optimal models (having highest explanatory power), aCFSs are predominantly located in G-negative chromosomal bands and away from centromeres, are enriched in *Alu* repeats and have high DNA flexibility. In alternative models, CpG island density, transcription start site density, recombination rate, H3K4me1 coverage, distance from telomere and mononucleotide microsatellite coverage are significant predictors. Also, aCFSs have high fragility when co-located with evolutionarily conserved chromosomal breakpoints. Thus, our modeling suggests that some features of chromosomal regions that are conserved in their evolutionary fragility across species separated by hundreds of million of years are also associated with fragility of these regions in the human genome under conditions of replication stress. This implies similarity in the mechanisms of chromosomal fragility at micro- and macro-evolutionary levels. Our models are predictive of the fragility of aCFSs mapped at a higher resolution. Importantly, the genomic features we identified here as significant predictors of fragility allow us to draw valuable inferences on the molecular mechanisms underlying aCFSs.

O-168

Positionally-biased gene loss after whole genome duplication: evidence from human, yeast and plant

Takashi Makino[1,2], Aoife McLysaght[1]
[1]*Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland,* [2]*Graduate School of Life Sciences, Tohoku University, Sendai 980-8578, Japan*

Whole genome duplication (WGD) has made a significant contribution to many eukaryotic genomes including yeast, plants and vertebrates. Following WGD, some ohnologs (WGD paralogs) remain in the genome arranged in blocks of conserved gene order and content (paralogons). However the most common outcome is loss of one of the ohnolog pair. It is unclear what factors, if any, govern gene loss from paralogons. Recent studies have reported physical clustering (genetic linkage) of functionally linked (interacting) genes in the human genome and propose a biological significance for the clustering of interacting genes such as co-expression or preservation of epistatic interactions. Here we conduct a novel test of the biological significance of gene clustering in the genome by examining patterns of gene loss after WGD. Immediately following WGD all interacting gene clusters are duplicated and the protein products from the linked and the unlinked interacting partner are equivalent. If the physical clustering of the genes is significant then we expect biased preservation of genetic linkage. Therefore we test a hypothesis that functionally linked genes in the same paralogon are preferentially retained in cis after WGD. We compare the number of protein-protein interactions (PPIs) between linked singletons within a paralogon (defined as cis-PPIs) with that of PPIs between singletons across paralogon pairs (defined as trans-PPIs). We find that paralogons in which the number of cis-PPIs is greater than that of trans-PPIs are significantly enriched in human, yeast and plant. In particular, singletons participating in cis-PPIs tend to be classified into "response to stimulus". We uncover strong evidence of biased gene loss after WGD which further supports the hypothesis of biologically significant clustering of genes in eukaryotic genomes. These observations give us new insight for understanding the evolution of genome structure and of protein interaction networks.

O-169

Radical Genome Architectures in Oxytricha: Diminution, Unscrambling, and Comparative Genomics in a Single Cell

Xiao Chen[1], John Bracht[1], Aaron Goldman[1], Egor Dolzhenko[1,2], Estienne Swart[1], Laura Landweber[1]
[1]Princeton University, Princeton NJ, USA, [2]University of South Florida, Tampa, FL, USA

Oxytricha and its relatives possess an extreme genome architecture that relies on two nuclear genomes, a diploid germline micronucleus and a highly polyploid somatic nucleus, as well as an RNA cached copy of the maternal macronucleus. Developmental genome rearrangement, amplification, and massive DNA elimination form the somatic, macronuclear genome from the germline, micronucleus through a process that appears guided by an RNA copy of the maternal macronuclear genome (Nowacki et al. 2008 Nature 451:153-8). Programmed rearrangement eliminates ~95% of the ~1Gb germline DNA, including all transposons and satellite DNA, and must piece together hundreds of thousands of often-scrambled gene segments to form predominantly single-gene, telomere-bearing molecules (nanochromosomes) that average just 3.2 kb in the macronucleus. Thousands of micronuclear genes are scrambled with respect to their macronuclear counterparts, with the gene segments present in a permuted or inverted order in the micronucleus relative to their order in the macronucleus. Hence, Oxytricha genome rearrangement requires an astounding number of inversion or translocation recombination events in a single cell that houses both the precursor and the product genomes. Furthermore, we are finding segments for different somatic genes frequently intertwined on the same genetic locus in a micronuclear chromosome. This talk will present our current knowledge of Oxytricha's scrambled germline genome and its cascade of programmed rearrangements that produce the much reduced but highly polyploid somatic macronucleus.

O-170

The origin and maintenance of a butterfly wing pattern supergene

Annabel Whibley[1], Lise Frezal[1], Robert Jones[1], Nicola Nadeau[2], Chris Jiggins[2], Mathieu Joron[1]
[1]CNRS/Muséum National d'Histoire Naturelle, Paris, France,[2]Department of Zoology, University of Cambridge, Cambridge, UK

Supergenes are tight clusters of loci that facilitate the co-segregation of adaptive variation, providing integrated control of complex adaptive phenotypes. A supergene architecture underlies the spectacular mimetic wing pattern variation in the neo-tropical butterfly *Heliconius numata*, in which several discrete forms, each resembling a different model, are maintained in sympatry. We have recently demonstrated that a series of chromosomal inversions exist at the supergene locus (P) in *H. numata*. These rearrangements tighten the genetic linkage between colour pattern loci that are known to recombine in closely-related species, resulting in complete suppression of recombination across a 400kb region in *H. numata*. Different haplotypes, each containing a different organization of genes within the supergene, show complete association with specific wing pattern morphs. Thus, the chromosome inversions provide a mechanism by which adaptive combinations of alleles can become locked together to function effectively as a single unit.

We have extended our initial genomic organization findings by applying multiple Next Generation Sequencing-based strategies to fully characterize the rearrangements and to investigate their impact on gene function and the recombination dynamics of the region. This analysis promises to shed light on the functional components of the supergene and the population processes maintaining wing pattern polymorphism in *H. numata*. Moreover, within the silvaniform clade of *Heliconius*, there are remarkable instances of shared wing patterns which may reflect the parallel evolution of adaptive mimetic patterns. By extending our genetic analyses to closely related species, we instead provide evidence for the wholesale introgression of an inverted chromosome segment from a second species, consistent with an hypothesis of promiscuous exchange of colour pattern alleles via hybridization.

O-171

Philosophy of paradigms and phylogeny

Maureen O'Malley
University of Sydney, NSW, Australia

The notion of paradigm came into prominence in Thomas Kuhn's work on scientific revolutions. Although Kuhn has been much criticized by philosophers, historians and scientists, some of his terminology has withstood these critical tests and become popularly used. Paradigm shifting is a term that has often been applied to molecular phylogeny, as new methods, data and concepts have altered and even transformed the field since the 1960s. In the introduction to this symposium on paradigms and phylogeny, I will discuss some putative paradigm shifts and raise questions about how useful it is to think of them in this Kuhnian light. I will propose alternative ways of understanding the dynamics of molecular phylogeny as it has grown and changed, with a focus on the concept of *integration*. One situation in which the term paradigm shift is commonly used is when integration fails. This occurs when one framework of inquiry cannot accommodate or contain (or perhaps control) a programme of research. Rather than thinking of paradigm-shifting revolutions, I will suggest that we might want to consider integration and its implications for phylogeny, especially in regard to prokaryote vis-à-vis eukaryote phylogeny.

O-172

The evolution of eukaryotes from archaea

Tom Williams[1], Peter Foster[2], Cymon Cox[3], T. Martin Embley[1]
[1]Newcastle University, Newcastle upon Tyne, UK,[2]Natural History Museum, London, UK,[3]University of Algarve, Faro, Portugal

The relationship of eukaryotes to other life forms remains a fundamental and unresolved question in evolutionary biology. Under the classic “three domains” paradigm, eukaryotes and archaea share a common ancestor after their divergence from bacteria, but before the diversification of either group. In contrast, some recent phylogenetic analyses have placed the eukaryotes within an already diversified archaea, as a sister group to one of several extant archaeal lineages. These competing hypotheses make very different predictions about the nature of the common ancestor of archaea and eukaryotes, and therefore about early eukaryotic evolution. A major difficulty for all current analyses is the poor sampling of archaeal diversity by genome sequencing, although this situation is now improving. We are exploring eukaryotic origins in the light of these new genomes by investigating the effect of recently sequenced, deep-branching archaeal genomes on two- and three-domain trees constructed from the “genealogy-defining core” of ribosomal RNA and protein-coding genes. We combine formal tests of congruence with phylogenetic networks to quantify the level of uncertainty and conflict in phylogenies of these core genes. Although core genes are not immune to horizontal transfer, we show that much of the observed incongruence is weakly supported or associated with poorly fitting evolutionary models. Our analyses suggest that eukaryotes are the sister group to the “TACK” (Thaum-, Aig-, Cren- and Korarcheota) superphylum within the archaea. We then consider the implications of this phylogenetic position for the gene content and biology of the common ancestor of archaea and eukaryotes.

O-173

Is the Tree of Life still a workable paradigm for evolution and phylogeny ?

Eric Bapteste
CNRS, Paris, France

Phylogenetics is rooted in a well-known style of scientific reasoning, called tree-thinking. In theory, tree-thinking has a very high explanatory power; in practice evolutionary studies inspired only by tree-thinking dramatically underestimate the diversity of evolutionary objects and processes identified by approaches of genomics and metagenomics. Thus, major evolutionary phenomena are not given as much room as they could in phylogenetic-based descriptions of the biological history of life. To effectively account for these major objects and processes, I will suggest that two additional styles of reasoning are required: the integrative-thinking, and the exploratory-thinking. I will introduce these styles and question whether their recognition in addition to tree-thinking is likely to provoke a change of paradigm in phylogenetics.

O-174

The tree of Life *circa* 2012: Is there reasonable middle ground between tree-hugging and clearcutting?

Andrew Roger
Dept. of Biochemistry and Molecular Biology, Dalhousie University, Halifax, N.S., Canada

The concept of a unique tree of Life (ToL) that sufficiently summarizes the relationships between organisms on Earth has been hotly debated over the last two decades. Challenges have come from the growing recognition of the importance of two kinds of evolutionary processes that cannot be represented by a strictly bifurcating ToL: horizontal gene transfer and endosymbiosis. It is now clear that horizontal (lateral) gene transfer and recombination between both closely- and distantly-related microbial genomes has been a widespread and important evolutionary mechanism for much of Life's history. Similarly, primary and secondary endosymbiosis followed by cellular and genomic integration of the symbionts has played a major role in generating the diversity of extant eukaryotes. Although these processes are well-accepted, the meaning and usefulness of a ToL concept in light of them remains extremely controversial.

Several common positions are taken by participants in this debate. At one extreme lies a kind of strict 'tree quasi-realism' that holds that events that lead to reticulation are a kind of white noise that can be ignored in the search for a major bifurcating ToL pattern. At the other end of the spectrum, 'tree anti-realists' claim that the ToL is false as it applies to the whole of Life's diversity and is, therefore, a positively misleading concept. I will argue that because reticulation events occur at relatively high frequency and affect virtually all major organismal lineages often resulting in profound evolutionary transitions, 'tree quasi-realism' is indefensible. However, several flavours of 'tree anti-realism' are similarly problematic because of their reliance on misleading metaphors, inappropriate generalizations from case examples of HGT and their ill-conceived attempts to vanquish 'tree-thinking' from microbial systematics. I will argue that it is unknown whether a 'vertical' signal of ancient microbial history can be recovered from genomic data from extant organisms and that discerning this signal is of critical importance to understanding evolutionary history.

O-175

Goods-thinking versus Tree-thinking

James McInerney
National University of Ireland Maynooth, Maynooth, Ireland

In an interview with Richard Dawkins, the great British evolutionary biologist John Maynard-Smith offered the opinion that a bacterial cell such as Escherichia coli could effectively behave like a football team, making changes to the personnel on the team (the genes in the genome) as it liked and if these changes were successful, then they were retained. The machanism for swapping out and in new genetic personnel is of course horizontal (or lateral) gene transfer. Maynard-Smith clearly had a goods-thinking approach to understanding genome evolution and this viewpoint resembles our proposal in 2011 that we should view evolving entities (genomes, plasmids, phage, viruses, etc.) through a prism of gene exchange in much the same way that economists view goods. In this talk I shall expand on our vision of goods-thinking to include one of the paradigms of goods-thinking, which is the notion of the "tragedy of the commons". Briefly, in the even that a good is publicly available, there is a danger that this good can fall victim to being overused and that "free-riders" would arise that make use of a good without contributing enough to its production.

O-176

It's OK to say 'prokaryote': paradigm wars in microbial systematics

W. Ford Doolittle
Dalhousie University, Halifax, Nova Scotia, Canada

@font-face { font-family: "Arial"; }@font-face { font-family: "Cambria"; }p.MsoNormal, li.MsoNormal, div.MsoNormal { margin: 0cm 0cm 0.0001pt; font-size: 12pt; font-family: "Times New Roman"; }div.Section1 { page: Section1; }

Norman Pace has in several publications asserted that the use of ‘prokaryote’ imparts on our students and ourselves a “wrong idea”, because there are on this planet *three* kinds of living thing (Bacteria, Archaea and Eukarya), not two (prokaryotes and eukaryotes). But “kinds of living thing” is surely not an unproblematic notion, and mutually cancelling distinctions of grade and clade figure prominently in Pace’s attack on ‘prokaryote’. Underneath, I suggest, there is an unarticulated belief, common among microbiologists, that taxa (from species to domains) are some sort of “natural kind”. This might make sense if the Tree of Life depicted the unique pattern of relationships between them, but in fact it only shows the relationships of some (a minority) of their genes. To make more of it is an act of reification. After a detailed consideration of the arguments pro and con, I will conclude, unsurprisingly that ‘prokaryote’ is OK, but ‘Prokaryota’ is not.

O-177

Evolution of evolvability of the var gene family in malaria

Adam F. Sander[1], Thomas Lavstsen[1], Thomas S. Rask[2], Michael Lisby[4], Thomas P. Gilbert[3], Ali Salanti[1], Sarah L. Fordyce[3], Jakob S. Jespersen[1], Eske Willerslev[3], Richard Carter[5], Thor G. Theander[1], Anders Gorm Pedersen[2], David E. Arnot[1,5]
[1]Centre for Medical Parasitology, University of Copenhagen, Copenhagen, Denmark, [2]Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark, [3]Centre for GeoGenetics, University of Copenhagen, Copenhagen, Denmark, [4]Department of Biology, University of Copenhagen, Copenhagen, Denmark, [5]Institute of Infection & Immunology Research, University of Edinburgh, Edinburgh, UK

Several human pathogens - bacteria, vira and parasites – undergo antigenic variation to counter host immune defense mechanisms. In *Plasmodium falciparum*, the most lethal of the human malaria parasites, the approximately 60 genes in the var gene family encode a set of antigenically diverse adhesion proteins called Plasmodium falciparum-erythrocyte membrane protein 1 (PfEMP1). Switching of var gene expression results in alternate expression of PfEMP1 proteins on the surface of infected erythrocytes, thereby allowing the parasite to evade the immune system.

Here we show that diversity in the var gene family is created by ectopic recombination between var gene paralogs during the parasite's sexual stages. The recombination events occur in close proximity to quasi-palindromic DNA motifs that are predicted to be capable of folding into hairpin structures. We show that these predicted secondary structure elements are concentrated at the boundaries of structural domains of the encoded PfEMP1 protein. Using an assay commonly used to study homologous recombination in *Saccharomyces cerevisiae*, the ability of a predicted secondary structure element to cause an increase in recombination frequency was furthermore verified experimentally.

These observations indicate that a DNA structure-dependent recombination mechanism generates antigenically diverse and functional *P. falciparum* adhesion antigens during sexual reproduction. It therefore appears that the DNA sequence of the var gene family has evolved to control its own mutation rate. The preferred location of the quasi-palindromic sequences at the boundaries of PfEMP1 structural domains further suggests that the recombination mechanism works in a manner that ensures the structural integrity of the resulting encoded protein. We suggest that these evolved characteristics of the var gene family result in a faster and less wasteful evolutionary search of sequence space compared to random mutation.

This is the first example of a large gene family evolving by DNA secondary structure-induced recombination, and may represent a common mechanism for facilitating evolvability of virulence genes of other pathogenic organisms.

O-178

"Proto-genes" and *de novo* gene birth

Anne-Ruxandra Carvunis[1,2], Thomas Rolland[1], Ilan Wapinski[3], Muhammed Yildirim[1], Nicolas Simonis[1], Benoit Charloteaux[1], Cesar Hidalgo[4], Michael Calderwood[1], Balaji Santhanam[1], Gloria Brar[5], Jonathan Weissman[5], Aviv Regev[6], Nicolas Thierry-Mieg[2], Michael Cusick[1], Marc Vidal[1]
[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, MA, USA, [2]UJF-Grenoble 1 / CNRS / TIMC-IMAG UMR 5525, Computational and Mathematical Biology Group, Grenoble, France, [3]Department of Systems Biology, Harvard Medical School, Cambridge, MA, USA, [4]The MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA, [5]Howard Hughes Medical Institute, Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California, USA, [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA

Novel protein-coding genes can arise either through re-organization of pre-existing genes or *de novo* in sequences previously devoid of genes. Processes involving re-organization of pre-existing genes, notably following gene duplication, have been extensively described. In contrast, *de novo* gene birth remains poorly understood, mainly because translation of sequences devoid of genes, or "non-genic" sequences, is expected to produce insignificant polypeptides rather than proteins with specific biological functions. We have tested an evolutionary model for *de novo* gene birth according to which functional genes evolve through transitory "proto-genes" generated by translational activity of non-genic sequences. In support of this model, open reading frames (ORFs) in *Saccharomyces cerevisiae* can be placed in an evolutionary continuum ranging from non-genic sequences to genes. Moreover, we detect translation of hundreds of short species-specific ORFs located in non-genic sequences. These translation events appear to provide adaptive potential, as suggested by their differential regulation upon stress and by signatures of retention by natural selection. Overall, we identify ~1,900 candidate proto-genes among *S. cerevisiae* ORFs. We estimate that *de novo* gene birth from such reservoir might be more prevalent than sporadic gene duplication. The proto-gene model is consistent with the notion that exposing hidden genetic variations facilitates species evolution.

O-179

The microbial lag phase is an evolvable trait that underlies specialist and generalist growth strategies.

Aaron New[1,2], Sander Govers[1,2], Bram Cerulus[1,2], Joao Xavier[3], Kevin Verstrepen[1,2]
[1]VIB Laboratory of Systems Biology, Heverlee, Belgium, [2]CMPG Laboratory of Genetics and Genomics, Heverlee, Belgium, [3]Program in Computational Biology, Memorial Sloan Kettering Cancer Center, New York City, USA

Organisms are faced with a tradeoff between specialized growth in specific niches, or more "generalist" strategies that optimize growth across variable conditions. Many microbes, for example, are glucose specialists because preference for this carbon source allows high growth rates. The fitness tradeoff for this specialist strategy can come in environments where a population of cells must transition from growth on preferred to non-preferred nutrients (commonly called diauxic shift). Specifically, as the preferred nutrient is depleted, delayed activation of the genes necessary to grow on non-preferred nutrients results in a "lag" phase: a period of adaptation during which cells temporarily slow growth or cease division altogether. Despite its biological and industrial relevance, little is known about the factors that determine the duration of the lag. Here, we show that the length of the lag phase is a variable and evolvable trait that balances an evolutionary tradeoff between generalist and specialist survival strategies. In wild isolates of the yeast Saccharomyces cerevisiae, we find remarkable variation in lag duration at the single-cell level that corresponds to glucose specialization at the population level. Moreover, within a genetically identical population, the length of the lag phase of individual cells can differ greatly, resulting in a bottleneck where the cells with the shortest lag phases contribute the most descendants to the final population. Using experimental evolution, we exploit this principle to select for mutants that have shorter lags than their wild-type ancestor. The independently arising mutants exhibit a spectrum of tradeoffs between shorter lag phases and glucose specialization. Specifically, two classes of generalist strategies emerge. The most common mutants are of a “jack of all trades” phenotype, many of which have lost preference for glucose. These mutants reduce variation in growth rates across different environments at the cost of maximal growth rate in glucose. A second class of “bet-hedgers” averages out the tradeoffs of specialization and generalization by a chromatin-mediated stochastic switching mechanism. Taken together, our quantitative measurements of single-cell and population level growth unmask the evolutionary forces shaping the speed at which organisms adapt to new environments. The results reveal that specialist and generalist strategies are subject to strong selective pressure and are evolutionarily plastic phenotypes.

O-180

Quantitative analysis of genetic and environmental factors determining variation in cell growth

Naomi Ziv, Mark Siegal, David Gresham
New York University, New York, USA

Growth is a fundamental property of cells. Complex networks of genetic and environmental factors regulate both the initiation of growth and the rate of proliferation. One of the major growth regulators in the budding yeast (*Saccharomyces cerevisiae*) is carbon availability. How a cell regulates its rate of growth in response to carbon availability and the extent to which this regulation varies is currently unknown. Furthermore, while microbial population growth has been studied for decades, it is still unknown how much inter-individual variation exists in growth rate and how this variation is effected by genetic or environmental variation. Phenotypic variation between cells of the same genotype (in the same environment) may be the result of stochastic events or a consequence of deterministic processes involving uncharacterized molecular regulatory mechanisms and is of unknown evolutionary significance. We have developed a novel high-throughput individual-based phenotyping assay, which enables estimation of cell-to-cell variability and as thousands of cells are analyzed, simultaneous accurate population measurements. I have used this assay to investigate the genetic basis of variation in growth between and within natural isolates of *Saccharomyces cerevisiae* with different ecological histories. This analysis has shown that cells tune their growth rate according to the nutrient concentration in the environment. Additionally, genetic determinants affect this response, leading to fitness advantages. The magnitude of these advantages again depends on the amount of carbon found in the environment. Remarkably, individual cells of the same genotype can have different growth rates and the amount of variation differs between genotypes. The evolutionary implications of this variation are unknown. Quantitative trait locus (QTL) mapping has revealed several loci affecting growth rate in different environments. Some loci are shared between phenotypes (glucose transporters) while others are unique. There are also genetic loci affecting the amount of within-genotype variation. Resolving these loci to gene and even nucleotide level will elucidate how the amount of phenotypic variation is controlled in natural populations and how this control has been modified during evolution. Regulation of growth in response to the nutrient environment is an ideal system in which to study microbial evolution. Nutrient limitation is an ecologically relevant selective pressure while growth rate is a major element of microbial fitness.

O-181

Molecular basis of increased evolvability as a result of multicellularity evolution

Maria Rebolleda-Gomez[1], Johnathon Fankhauser[3], William C. Ratcliff[1], Michael Travisano[1,2]
[1]Ecology, Evolution and Behavior deppartment, University of Minnesota, Minneapolis, USA, [2]Biotechology Institute, University of Minnesota, Minneapolis, USA, [3]Plant Sciences, University of Minnesota, Minneapolis, USA

Observations from the fossil record suggest that evolution of innovations may increase the ability of a clade to generate adaptive diversity (i.e. evolvability). Multicellularity is one of the major evolutionary innovations in the history of life; it allowed for the rapid diversification of metazoans. In contrast, other multicellular lineages like the volvocine algae have relative low diversity. The first appearance of multicellularity and its effects on evolutionary dynamics are difficult to evaluate from the fossil record. Thus, to evaluate the underlying molecular basis of innovations and their impact in evolvability, we compared the genomes of replicate populations of multicellular yeast obtained via experimental evolution under uniform selective conditions. We identified an important role of regulatory elements as well as the co-option of unicellular molecular “tool-kits” on the evolution of multicellular phenotypes. Despite the generality of this mechanism, different mutations, and even different genes, played a role on the evolution of multicellularity in different populations. Consistent with these results, we observed the evolution of large amounts of phenotypic divergence across populations. These results support the idea that unicellular components largely determine the evolutionary potential for multicellularity as well as the resulting increase in evolvability. Divergence between replicate populations could be due to neutral processes like historical contingency. Nonetheless, within populations we found evidence for adaptive diversification. To further understand the molecular basis of intra-population divergence, we compared the genomes of two genotypes coexisting in one of the replicate populations at the time when the diversification event first occurred and after 60 days of selection. We found changes involved with cluster traits regulation. Differences between these isolates are explained, partly, by a trade-off between growth and settling rates leading to ecological specialization. Thus, studying the evolutionary dynamics of such a simple model provides insights on the evolution of complex traits. In this model, such novelties expand the phenotypic space, allowing for increased evolvability and ecological diversification by co-option of molecular toolkits present in the unicellular ancestor.

O-182

Experimental evolution of a viral protease

Thomas Shafee, Pietro Gatti-Lafranconi, Florian Hollfelder
University of Cambridge, Cambridge, UK

Directed evolution in the laboratory provides a useful complement to the study of natural variation. Single enzymes can be artificially evolved in isolation by repeated rounds of mutagenesis of their gene, expression and selection/screening. Focusing on a single enzyme allows the processes of evolution to be followed in molecular detail using an integrated set of kinetic, biophysical and structural characterisations.

Tobacco etch virus (TEV) protease is a highly sequence specific cysteine protease. An activity assay using a FRET-pair of GFP variants (linked by a cleavage recognition sequence) as a substrate gives enormous substrate variety and detail on enzyme promiscuity. Here, we use TEV protease's laboratory evolution as a model to understand the details of how cryptic variation is accumulated when selecting for maintenance of function and how this affects protein evolution.

Libraries of 10^6 TEV protease variants are created by mutagenic PCR (at either $2x10^{-3}$or $5x10^{-3}$ substitutions per bp) and co-expressed with the FRET-pair substrate in *E.coli*. The cells are screened by FACS and populations of either 200 or 20000 variants retaining wt levels of activity are selected and used as the template for the next round of mutagenesis. This cycle is iterated to allow the pool to evolve away from the original sequence.

The distribution of phenotypes in the libraries created by mutagenesis was found to be bimodal with the majority of mutations display either near-neutral or completely deleterious effects. This bimodality was maintained over the 7 rounds of evolution performed so far.

The changes in distribution of activity levels in the library show that, though the proportion of mutations that are neutral decreases as the sequence drifts from the wt, the larger populations can sustainably tolerate mutagenesis. The ratio of neutral mutations plateaus at a level proportional to the mutation rate (40% and 10% for the low and high mutations rates respectively). The properties of these evolving lineages will be studied from a biochemical angle to address what types of mutations accumulate and how these affect the properties of the enzyme.

These experiments are being complemented by parallel studies on substrate specificity changes and for adaptation to a changed active site nucleophile. Information from targeted and neutral enzyme evolution is integrated to better understand the interplay of stability, promiscuity and evolvability and how these can be affected by the accumulation of neutral mutations.

O-183

Evolution of orthologous and paralogous genes: expression, function, and fitness effect

Jianzhi Zhang
University of Michigan, Ann Arbor, MI, USA

Orthology and paralogy are basic concepts in molecular evolution. It is generally assumed that orthologous genes have lower evolutionary rates than paralogous genes in expression, function, and fitness effect. Although this assumption has important implications for many areas of biology, its validity has just begun to be examined empirically. I will summarize the current standing of this assumption and report our new findings.

O-184

The evolutionary fate of short linear motifs in paralogous proteins

Alex N. Nguyen Ba, Alan M. Moses
University of Toronto, Toronto, Ontario, Canada

Gene duplication is an important evolutionary mechanism that can generate new protein function by neo-functionalization or sub-functionalization. Neo- and sub-functionalization are usually thought to occur through amino acid substitutions that change enzymatic activity, reorganization of protein domains or changes in transcriptional regulation. However, protein activity is also controlled by post-translational regulation (protein interactions and modifications), often mediated through ‘short linear motifs' in the primary amino acid sequence. We hypothesize that changes in these motifs might be another important source of functional diversity, because protein regulation is intimately related to protein function. We therefore sought to characterize the patterns of evolution of short linear motifs following gene duplication, specifically using retained duplicates created during the whole genome duplication in budding yeast. In contrast to strong constraint observed on short linear motifs in single-copy proteins, when paralogous gene copies have been retained, we find that a large proportion of the short linear motifs rapidly diverge after duplication, consistent with neo- or sub-functionalization. Interestingly, we find that these changes are often correlated with changes in protein domains. These results provide evidence that rapid post-translational regulatory changes are also important in the evolution of novel protein function after gene duplication.

O-185

Functional, transcriptional and sequence evolution asymmetry between paralogs and between orthologs in 12 Drosophila genomes

Lev Yampolsky[1], Michael Bouzinier[2]
[1]*East Tennessee State University, Johnson City, TN, USA,* [2]*InterSystems Corp., Boston, MA, USA*

Recent research demonstrates that despite widespread sub- and neofunctionalization of duplicated genes the difference in functional divergence between paralogs and between orthologs is far from simple. Part of the complexity of the problem is rooted in the fact that duplications are more likely to be retained in more conserved genes, thus creating an apparent conservatism of paralogs divergence. Further, patterns of sequence and functional divergence may be fundamentally different between old and recent duplications and between duplicated genes with different expression patterns. Analysis of sequence divergence on a rich phylogenetic context allows the comparison between pairs of paralogs and pairs of sister orthologous clades, which ameliorates these problems. We created a database of over 8 mln of amino acid substitutions occurring on the phylogenetic tree of 12 Drosophila genomes and analyzed their rates and radicalities in singletons and duplicated genes in conjunction with ontology and expression data. The analysis suggests that 1) the rate and radicality of paralogs divergence is greater than those of orthologs in slowly evolving proteins, but not in fast evolving proteins, regardless of the age of the duplication; 2) the asymmetry of paralogs divergence is greater than the asymmetry of comparable orthologs divergence in proteins with Ka/Ks < 0.5 and in proteins with Ka/Ks>1, but not in the ones with Ka/Ks close to neutrality; 3) asymmetry of paralogs divergence is greater in paralogs pairs with independently confirmed functional diversity; 4) asymmetry of paralogs sequence divergence is the highest in genes with intermediate expression level, in genes with the most uniform expression across tissues and in pairs of paralogs with the highest divergence of expression (measured as the coefficient of correlation across tissues) and 5) paralogs with the highest relative rate of sequence evolution have relatively lower expression rate in the young, but not in the old duplications. We will discuss the implications of these results for the understanding of functional divergence in retained duplicated genes and for identification of duplications retained via different mechanisms.

O-186

Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication

Toni Gabaldón[1], Jaime Huerta-Cepas[1], Joaquin Dopazo[2], Martijn Huynen[3]
[1]Centre for Genomic Regulation (CRG), Barcelona, Spain,[2]Centro de Investigación Príncipe Felipe, Valencia, Spain, [3]Centre for Molecular Life Sciences, Nijmegen, The Netherlands

Gene duplication is one of the main mechanisms by which genomes can acquire novel functions. It has been proposed that the retention of gene duplicates can be associated to processes of tissue expression divergence. These models predict that acquisition of divergent expression patterns should be acquired shortly after the duplication, and that larger divergence in tissue expression would be expected for paralogs, as compared to orthologs of a similar age. Many studies have shown that gene duplicates tend to have divergent expression patterns and that gene family expansions are associated with high levels of tissue specificity. However, the timeframe in which these processes occur have rarely been investigated in detail, particularly in vertebrates, and most analyses do not include direct comparisons of orthologs as a baseline for the expected levels of tissue specificity in absence of duplications. To assess the specific contribution of duplications to expression divergence, we combine here phylogenetic analyses and expression data from human and mouse. In particular, we study differences in spatial expression among human-mouse paralogs, specifically duplicated after the radiation of mammals, and compare them to pairs of orthologs in the same species. Our results show that gene duplication leads to increased levels of tissue specificity and that this tends to occur promptly after the duplication event.

O-187

On the use of Gene Ontology annotations to assess functional similarity among orthologs and paralogs

Paul Thomas[1], Valerie Wood[2], Christopher Mungall[3], Suzanna Lewis[3], Judith Blake[4]
[1]University of Southern California, Los Angeles, CA, USA, [2]University of Cambridge, Cambridge, UK, [3]Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [4]The Jackson Laboratory, Bar Harbor, ME, USA

Gene Ontology (GO) annotations have been used to estimate the functional similarity between homologous genes, notably to assess whether orthologs tend to be more similar in function than paralogs. We discuss recent work showing the effects of the "open world assumption" on some GO-based metrics of functional similarity. In particular, these metrics are influenced by ascertainment bias in experimentally-supported GO annotations. We also discuss GO Consortium efforts to model functional evolution through gene families, and how these models may bear on the relationship between orthology, paralogy and gene function.

O-188

Emergence of Mammals by Emergency: Exaptation

Norihiro Okada
Tokyo Inst.Tech,, Yokohama, Japan

P (Permian)-T (Triassic) mass extinction is the most extensive extinction we have ever experienced on the earth. This occurred 250 Ma (million years ago). Since reptiles and mammals diverged about 310 Ma, ancestors of these two lineages struggled to survive after this event under. The most extensive environmental change occurred at the boundary of P-T mass extinction is the decrease of oxygen concentration. It is believed that it was up to 30% during the Permian era but after this boundary it decreased to 10%. Considering the concentration of oxygen at present is 21%, this environment should have been severe to all the survivals. Ancestors of reptiles and mammals should have adapted to this superanoxia.

For these 3 years, our group has asked what happened on the DNA level for mammals to adapt to this extreme environment. We discovered that more than 100 loci of a new SINE family called AmnSINE1 were exapted (got function) possibly for this adaptation, because these SINE loci constitute a part of CNE that are conserved among mammals. We have now three concrete examples. One locus function as an enhancer for fgf8, which is involved in patterning of diencephalon. This enhancer has unique modular organization. The second functions as an enhancer for Satb2, which is involved in axon formation of corpus callosum. Corpus callosum is specific to mammals, being involved in interhemispheric communication between the left and right cerebral hemispheres. It is the largest white matter structure in the brain, consisting of 200–250 million contalateral axonal projections, a lower half of formation of which was shown to be enhanced by our SINE. I speculate that it might be required for early mammals to stimulate communication between both hemispheres of our brain. The last example functions as an enhancer for Wint5a, which is involved in closure of the secondary palate of our jaw. Secondary palate is responsible for separation between the nasal and oral cavities, and is assumed to have evolved just after P-T mass extinction for our ancestor to obtain more efficient respiration. I am confident with the notion that this example is one of the most typical ones of exaptation which links to geological events we have ever experienced.

O-189

Molecular And Functional Evolution Of Two Retroposon-Derived Enhancers Controlling Neuron-Specific Expression Of The Same Brain Gene

Marcelo Rubinstein[1,2], Lucía Franchini[1], Flavio de Souza[1,2], Daniel Lam[3], Malcolm Low[3]
[1]INGEBI-CONICET, Buenos Aires, Argentina,[2]University of Buenos Aires, Buenos Aires, Argentina,[3]University of Michigan, USA, Michigan, USA

The proopiomelanocortin gene (POMC) is expressed in a group of neurons present in the arcuate nucleus of the hypothalamus. Neuron-specific POMC expression in mammals is conveyed by two conserved distal enhancers, named nPE1 and nPE2. Previous transgenic mouse studies showed that nPE1 and nPE2 independently drive reporter gene expression to POMC neurons. Here, we investigate the evolutionary mechanisms that shaped not one but two neuron-specific POMC enhancers, and tested whether nPE1 and nPE2 drive identical or complementary spatio-temporal expression patterns. Sequence comparison among representative genomes of most vertebrate classes and mammalian orders showed that nPE1 is a placental novelty. Using in silico paleogenomics we discovered that nPE1 originated from the exaptation of a Mammalian-apparent LTR Retrotransposon (MaLR) sometime between the metatherian/eutherian split (147 MYA) and the placental mammal radiation (~90 MYA). Thus, the evolutionary origin of nPE1 differs, in kind and time, from that previously demonstrated for nPE2 which was exapted from a CORE-SINE retroposon before the origin of prototherians, 166 MYA. Analysis of compound transgenic mice expressing the fluorescent markers tomato and EGFP under the transcriptional control of nPE1 or nPE2, respectively, demonstrated coexpression of both reporter genes along the entire arcuate nucleus from the onset of Pomc expression in the presumptive mouse diencephalon. Mice deficient in either nPE1, nPE2 or both enhancers simultaneously show unique differential phenotypes that are beginning to reveal the importance of functionally overlapping enhancers. Thus, the independent exaptation of two unrelated retroposons into functional analogs acting as neuron-specific enhancers of POMC constitute an authentic first example of convergent molecular evolution of cell-specific enhancers to govern gene expression throughout the entire lifespan.

O-190

Brain region specific allelic imbalance for a SNP in the prairie vole oxytocin receptor gene

Lanikea King[1,2], Kiyoshi Inoue[1,2], Larry Young[1,2]
[1]Center for Translation Neuroscience, Atlanta, GA, USA, [2]Behavioral Neuroscience and Psychiatric Disorders, Atlanta, GA, USA

Diversity in brain oxytocin receptor (*Oxtr*) expression is associated with species differences and individual variation in social behavior. Monogamous prairie voles have high densities of OXTR in the nucleus accumbens (NA) compared to non-monogamous species, and OXTR signaling in NA is necessary for partner preference formation. There is remarkable individual variation in OXTR density in the NA that is not present in non-striatal brain regions. Females with higher OXTR density in the NA display higher levels of alloparental nurturing behavior, and increasing OXTR density in the NA using gene transfer facilitates partner preference formation. To test the hypothesis that genetic polymorphisms contribute to this variation in *Oxtr* expression, we used pyrosequencing to survey for allelic expression imbalance (AEI) of *Oxtr* mRNA in the NA of individual prairie voles. AEI compares mRNA levels derived from each allele, identified by a single nucleotide polymorphism (SNP) within an individual. AEI is indicative of linked regulatory elements contributing to variation in gene expression rather than epigenetic factors. We resequenced the *Oxtr* from 19 prairie voles to identify common SNPs within *Oxtr* exons. We then identified five animals heterozygous at SNP 2 (C/T), which lies within the 3' untranslated region of the *Oxtr*. Genomic DNA (gDNA) and cDNA derived from NA and three other brain regions were amplified by PCR and pyrosequenced. The relative abundance of the T and C alleles was calculated. In the NA, The cDNA T/C ratios for SNP 2 ranged from 1.16 – 9.00, indicating consistent AEI, with T always greater than C. T/C ratios were greater for cDNA than gDNA from the same individuals, confirming that *Oxtr* exhibits AEI in prairie vole NA. In three other brain regions, there was no AEI observed. We next genotyped 10 voles with known extremes of NA OXTR density: 5 High and 5 Low. Each of the 5 High OXTR voles were heterozygous C/T while all 5 Low voles were homozygous C/C at SNP 2. The C/T genotype was associated with high OXTR density. These data suggest the T allele of SNP 2 is either causative or linked to a regulatory variant that potentiates *Oxtr* expression specifically within a single brain region. Thus SNP2 is a useful marker to predict OXTR density from birth. The observed regulatory variation may potentially, through this modulation of OXTR, contribute to the organization of prairie vole social behavior.

O-191

Evolution of Regulatory Domains of the Transcrption Factor HoxA-11 in Mammalian Evolution.

Mauris Nnamani[1], Jens Meiler[2], Laura Mizoue[2], Gunter Wagner[1]
[1]Yale Systems Biology Institute, Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA, [2]Center for Structural Biology, Vanderbilt University., Nashville, TN, USA

HoxA-11 is a member of a large class of transcription factors called homeobox genes. It is involved in the regulation of placenta development and required for proper fetal implantation. We have recently shown that this gene experienced strong directional selection in the stem lineage of mammals. We also reported that HoxA-11, although a repressor on the decidual prolactin promoter, is converted to an activator when co-expressed with FOXO1A. This switch in regulatory activity was only noticed in HoxA-11 from placental mammals and all the amino acid substitutions coincidental with this switch are outside the homeodomain. It is clear how gene regulation evolves through changes in the *cis*-regulatory elements, but it is not clear how amino-acid substitutions in transcription factors influence gene regulation. Here we report on functional and structural studies to uncover the mechanisms of the derived cooperative activation function of HoxA-11 and Foxo1a.

Our goal in this study is to further understand the functional architecture of the HoxA-11 protein in order to understand the functional significance of the derived amino acid residues. Computational structure predictions suggest the existence of a structured domain in the N-terminal part of the HoxA-11 protein, which we call MD, which contains 10 of the derived amino acids. Here we show that the HoxA-11 and FOXO1A cooperative activation of prolactin is mediated through an activation domain and regulatory motif of HoxA-11. By functionally analyzing resurrected HoxA-11 ancestral protein, we were able to identify key evolutionary amino-acid substitutions that resulted in a lose/gain of phosphorylatable sites within the regulatory motif that lead to the derived co-operative gene regulation with Foxo1A. This regulatory region appears to be crucial in the regulation of HoxA-11 transcriptional activity and provides a second layer of evolutionary control for gene regulation.

O-192

The Mechanisms and Expression of Copy Number Variation in Drosophila

Manyuan Long
The University of Chicago, Chicago, Illinois, USA

Copy number variation (CNV) has been shown to be prevalent in Drosophila populations. A few lines of evidence revealed that both the level and structure of CNV were shaped by natural selection, due to the functional consequence of CNV. These include the uneven distribution of CNV across genomes, population genomic analyses of CNV and the functionality analyses of new duplicates using both microarray-based and deep sequencing approaches (Emerson et al, 2008, Science; Dopman et al, 2007, PNAS; Landback et al, unpublished data). A neglected line of analysis was the understanding of the impact on the level and property of CNV from the mutational process as manifested in genome replication. We provided evidence by combining the replication analysis and CNVs in the natural populations of D. melanogaster and D. simulans, revealing that the mutational process also played a role in shaping the CNV, in conjunction with the eminent role of natural selection (Cardoso-Moreira et al, 2010, Trend Genet; Cardoso-Moreira et al, 2011, PLoS Genet). The impact of expression on the non-fixated duplicates and determination of its evolutionary fates will be also discussed (Vankuren et al, non-published data).

O-193

Gene expression, copy number variants and selective sweeps in wild mice

Jarosław Bryk, Diethard Tautz
Max Planck Institute for Evolutionary Biology, Plön, Germany

Copy number variatnts (CNV) were until recently an under appreciated contribution to genome diversity and function. In humans and mice, CNV affect up to 12% of genome and can influence transcription levels of genes within and outside CNV loci. CNVs are also highly mutable, with mutation rates even in inbred mice being as high as for microsatellites. High mutagenicity and functional impact on the genome make CNVs a potentially important source of adaptive variation.

Here we present an investigation of CNVs influence on gene expression divergence and a search for CNV under selective sweeps in two recently diverged, natural populations of *Mus musculus domesticus*. We analysed three tissues (brain, liver, testis) from twelve animals (six from each population). We identified both differentially expressed and invariant genes between the two populations using microarray platforms from Agilent and Affymetrix and subsequently checked whether these genes are copy–number variable using Agilents' custom, high-resolution comparative genome hybridization microarrays on 10 unrelated samples from the same populations. While we identified about 1000 unique abberrations altogether, gene expression differences were not predictive of copy number differences. On the other hand, almost all of about 20 copy number variants that differentiated the two populations influenced expression levels. Finally, we screened the CNVs that differentiated the two populations for selective sweeps using microsatellite genotyping and identified several loci with evidence for reduced heterozygosity in one population versus the other and potential phenotypic significance.

O-194

A bias toward inflation of the *Drosophila melanogaster* genome is countered by strong selection

Daniel R. Schrider[1,2], David Houle[3], Michael Lynch[1], Matthew W. Hahn[1,2]
[1]Department of Biology, Indiana University, Bloomington, Indiana, USA, [2]School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA, [3]Department of Biological Science, Florida State University, Tallahassee, Florida, USA

Because spontaneous mutation is the source of all genetic diversity, measuring mutation rates can reveal how natural selection drives patterns of variation within and across species. Here we present results from a mutation accumulation experiment conducted in *Drosophila melanogaster* that is able to uncover mutations of all sizes-not just point mutations-and captures more mutations than all previous genome-wide mutation accumulation studies in eukaryotes combined. We find substantial variation (> twofold) in mutation rates measured from different genetic backgrounds, suggesting that rates may vary highly among individuals. By comparing mutation data to polymorphism segregating in the North American population of *Drosophila*, we are able to present the first direct inference of the proportion of substitutions that are deleterious enough to be eliminated by purifying selection. Given the potential phenotypic and adaptive importance of genomic copy number variants (CNVs), we also measured the rates of large genomic duplications and deletions, and present the first accurate genome-wide estimates of these rates in any species. We show that the previously observed mutational bias towards DNA loss in *Drosophila* is dramatically reversed when these large mutations are included, with twice as many base pairs added to the genome by duplication as removed by deletion each generation. This implies that mutational forces alone would cause the genome to grow rapidly, doubling in size every ~250,000 years. However, we estimate that ~99% of duplications and deletions are deleterious, making them 10 times more likely to be removed by natural selection than nonsynonymous mutations; this strong purifying selection slows the growth of the *Drosophila* genome by over two orders of magnitude, increasing the doubling time to ~30 million years. These results have a profound impact on our understanding of not only the rates of new mutations of all sizes, but the selective forces they encounter once they arise.

O-195

Structural variation within and between natural stickleback populations

Philine Feulner[1], Frédéric Chain[2], Mahesh Panchal[2], Marvin Mundry[1], Christophe Eizaguirre[3], Martin Kalbe[2], Tobias Lenz[2], Andrew Moore[1], Irene Samonte-Padilla[2], Erich Bornberg-Bauer[1], Thorsten Reusch[3], Manfred Milinski[2]
[1]Westfaelische Wilhelms University, Muenster, Germany, [2]Max Planck Institute for Evolutionary Biology, Ploen, Germany, [3]Institute for Marine Sciences, Kiel, Germany

Recent advances in technology have enabled the characterization of various types of genome architectural variation and their prevalence in populations. It has been revealed that many such structural differences have significant functional impacts. The tree-spined stickleback has been established as a super model for evolutionary ecological genomics. However, genome-wide structural variation data from natural populations remain scarce. Here we present whole genome data from 66 three-spined stickleback individuals representing three ecotypes from a broad geographic range, focussing on lake-stream population pairs. For each population, the sequencing of six individuals per population allows an in-depth evaluation of genome-wide polymorphism. Population diversity is contrasted between parapatric and geographically distant (allopatric) population pairs undergoing parallel ecological adaptation. Large structural variations such as deletions, insertions, copy number variations, inversions, and translocation are accessed exploiting different signatures apparent in short-read sequencing data. Utilizing paired-end mapping, depth of coverage, and split read analysis, we are constructing a structural variation discovery set for the three-spined stickleback. We have identified population-specific structural variations, many of which interfere with known protein coding regions. Evaluating structural variation of the stickleback genome within and between natural populations will contribute towards a better understanding of the evolution of genomes. In addition, our nested experimental setup allows us to gain insights about the importance of structural variations for an adaptive response in the wild.

O-196

Structural haplotypes and recent evolution of the human 17q21.31 locus

Linda Boettger[1,2], Robert Handsaker[1,2], Michael Zody[1,2], Steven McCarroll[1,2]
[1]Department of Genetics, Harvard Medical School, Boston, MA, USA, [2]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA

Structurally complex regions are thought to be dynamic and rapidly evolving, but to date the underlying genome structures segregating at such loci have not been known. Human 17q21.31 is one such structurally complex region in the human genome: it contains a megabase-long inversion polymorphism and many complex, uncharacterized copy-number variations (CNVs). The inverted form of this locus, H2, is rare in most of the world but reaches a 20% frequency in Europe, where this haplotype is reported to be positively selected. Markers in the 17q21.31 region associate with female fertility and the female meiotic recombination rate, but the functional variants behind these associations are not known.

As with other structurally complex regions of the human genome, the underlying molecular genome structures that segregate at 17q21.31 have not been known. We developed a population-genetic approach to this problem based on (i) accurate typing of structural features in large populations, and (ii) phasing of these structural features using populations and families, to identify the structural haplotypes that segregate at this locus.

We describe nine segregating structural forms of 17q21.31, each distinguished by the gain, loss, or rearrangement of a large genomic segment (>100 kb) relative to each of the other structural forms. These nine structural haplotypes form a phylogeny that offers a structural history of the rapidly evolving 17q21.31 locus. We identify an older H2 structure (not previously described) that is present at low frequency in Europeans and in Central African hunter-gatherer populations. Five structural forms of 17q21.31 (including forms of both the H1 and H2 inversion types) are present primarily in populations with West Eurasian ancestry.

Surprisingly, both the H1 and the H2 forms of the inversion contain partial duplications of the same gene, KIAA1267. These duplications have arisen independently on the H1 and H2 backgrounds; both have subsequently risen to high frequency (26% and 19%) in West Eurasian populations, though they are nearly absent in other parts of the world and appear to coalesce to distinct, recent ancestors. The duplications both produce novel, truncated transcripts of the KIAA1267 gene, whose Drosophila ortholog is involved in maintenance of female germ cells and affects fertility. Although these two overlapping duplications arose separately, they converge upon a common molecular expression phenotype.

O-197

Evolution of antibiotic resistance: The network of trade-offs

Csaba Pal
Biological Research Center, Szeged, Hungary

Evolution of resistance towards a single antibiotic can simultaneously increase (cross-resistance) or decrease (collateral sensitivity) fitness to multiple other antimicrobial agents. However, it remains unclear how frequent these evolutionary interactions are, and how far they are understandable based on accumulated knowledge on individual compoundsdrugs. Using combination of laboratory experimental evolution and high-throughput phenotypic screens in *Escherichia coli*, we systematically explored the map of these evolutionary trade-offs across 24 clinically widely employed antibiotics. The network is very dense and strikingly, not only cross-resistance but also collateral sensitivity occur at high rates. By integrating chemical, physiological and chemogenomic data, evolution of cross-resistance is predictable. Whole-genome sequencing of the evolved strains revealed a diverse set of parallel mutations with pleiotropic effects. Most notably, changes in membrane proton-motive force underlie the observed negative trade-off between aminoglycoside uptake and multi-drug efflux pump activity, explaining why strains adapted to aminoglycosides are hypersensitive to several other agents. Furthermore, we show that these trade-offs also have implications on bacterial persistence and evolvability under severe antibiotic stress.

O-198

Communities in Dense Networks

Sune Lehmann[1,4], James P Bagrow[2,4], Yong-Yeol Ahn[3,4]
[1]Technical University of Denmark, Lyngby, Denmark,[2]Northwestern University, Evanston, IL, USA,[3]Indiana University Bloomington, Bloomington, IN, USA,[4]Northeastern University, Boston, MA, USA

We know that communities in networks often overlap such that nodes simultaneously belong to several groups. Additionally, many networks are known to possess hierarchical organization, where communities are recursively grouped into a hierarchical structure. However, when each and every node belongs to more than one group, a single global hierarchy of nodes cannot capture the relationships between overlapping groups. Here we define communities as groups of links rather than nodes and show that this approach reconciles the ideas underlying overlapping communities and hierarchical organization. Further, we discuss the issue of sampling in graphs with pervasive overlap and implications for non-overlapping communities and network analysis. Finally, the talk will address the general problem of proper validation of detected communities and suggest a solution, based on real networks with available metadata (which will be applied to assess the quality of the link-communities).

O-199

Gene loss promotes functional complexity in bacteria

Yogeshwar Kelkar, Howard Ochman
Yale University, New Haven, CT, USA

The genomes of many host-associated bacteria experience extensive gene loss resulting from two sources: (1) the inactivation of genes that are no longer needed in the host, and (2) the accumulation of deleterious mutations in beneficial, but not essential, genes due to genetic drift - a consequence of host-restriction and population bottlenecks during host-to-host transfers. Because drift will erode useful genes, we asked whether the genes maintained in reduced genomes assume lost functions by increasing their functional complexity. Using protein-protein interactions (PPI) as a proxy for gene function and complexity, we compared the experimentally determined PPI networks of bacteria spanning a broad array of genome sizes. We determined that surviving proteins in small genomes interact with proteins from a wider range of biological processes than do their orthologs in the larger genomes, implying that genome reduction and gene loss lead to increased functional complexity in bacteria. In addition, the conclusion that pathogens with reduced genomes may have independently evolved novel variations in biological pathways has consequences in the development of highly targeted antimicrobial treatments.

O-200

Superessential reactions in metabolic networks and the evolution of drug resistance

Aditya Barve[1,3], João F. Matias Rodrigues[2,3], Andreas Wagner[1,3]
[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland,[2]Institute of Molecular Life Sciences , University of Zurich, Zurich, Switzerland,[3]The Swiss Institute of Bioinformatics, Lausanne, Switzerland,[4]The Santa Fe Institute, Santa Fe, USA

The metabolic genotype of an organism can change through loss and acquisition of enzymecoding genes, while preserving its ability to survive and synthesize biomass in specific environments. This evolutionary plasticity allows pathogens to evolve resistance to antimetabolic drugs, by acquiring new metabolic pathways that by-pass an enzyme blocked by a drug. We here study quantitatively the extent to which individual metabolic reactions and enzymes can be by-passed. To this end, we use a recently developed computational approach to create large metabolic network ensembles that can synthesize all biomass components in a given environment, but that contain an otherwise random set of known biochemical reactions. Using this approach, we identify a small core of 125 reactions that are absolutely superessential, that is, required in all metabolic networks. Many of these reactions have been experimentally confirmed as essential in different organisms. We also report a superessentiality index for thousands of reactions. This index indicates how easily a reaction can be by-passed. It correlates with the number of sequenced genomes that encode an enzyme for the reaction. Superessentiality can help choose an enzyme as a potential drug target, especially since the index is not highly sensitive to the chemical environment a pathogen requires. Our work also shows how analyses of large network ensembles can help understand the evolution of complex and robust metabolic networks.

O-201

Genomic sources of regulatory variation: mutation, polymorphism, and divergence

Patricia Wittkopp[1], Joseph Coolon[1], Jonathan Gruber[1], Kraig Stevenson[1], David Yuan[1], Brian Metzger[1], Joel McManus[2], Brenton Graveley[2], Kara Vogel[1], Gizem Kalay[1]
[1]*University of Michigan, Ann Arbor, MI, USA,* [2]*University of Connecticut, Farmington, CT, USA*

During the last decade, various methods (most recently, RNA-seq) have been developed for quantifying levels of gene expression (mRNA abundance) on a genomic scale in diverse organisms. Application of these techniques has shown that variation in gene expression is common both within and between species. These expression differences can arise from *cis*- and/or *trans*-regulatory changes, and methods are now available for distinguishing between them. Using such approaches to compare the genetic basis of polymorphic and divergent expression in flies (*Drosophila*) and yeast (*Saccharomyces*) has shown that *cis*-regulatory changes account for a greater proportion of expression differences between than within species. In this talk, I will use RNA-seq analysis of closely related *Drosophila* species to illustrate this and other patterns of regulatory evolution and then show how quantifying properties of new regulatory mutations (in *S. cerevisiae*) is providing insight into the evolutionary processes responsible for these patterns.

O-202

Faster-X evolution of gene expression in *Drosophila*

Richard Meisel[1], John Malone[2], Andrew Clark[1]
[1]*Cornell University, Ithaca, NY, USA,* [2]*Florida State University, Tallahassee, FL, USA*

Our understanding of the evolution of gene expression is naïve relative to what we know about DNA sequence evolution. Contrasting patterns between X-linked and autosomal genes has proven informative of the evolutionary dynamics of DNA sequences because of differences in the demographic and selective environments on the X and the autosomes. X chromosomes and autosomes also differ in their chromatin structure, and gene expression levels are affected by these chromatin modifications. To test whether the unique environment of the X chromosome affects the evolution of gene expression, we interrogated RNA-seq and microarray data collected from males and females of six *Drosophila* species. The expression levels of X-linked genes diverge faster than autosomal gene expression, similar to the "faster-X" effect often observed in DNA sequence evolution. In addition, we detect faster-X evolution of gene expression in non-reproductive tissues, suggesting that this pattern is not driven by the rapid turn-over of X-linked reproductive genes. Faster-X evolution of gene expression was recently described in mammals, but it was limited to the evolutionary lineages shortly following the creation of the therian X chromosome. In contrast, we detect faster-X evolution along all evolutionary lineages, including those at the tips of the phylogeny. Additionally, we fail to detect a faster-X effect along a lineage containing a recently created X chromosome, suggesting that the faster-X evolution of gene expression in *Drosophila* is not the result of acceleration immediately following X-linkage. The *Drosophila* dosage compensation complex (DCC) binds the X chromosome in males, inducing chromatin modifications that lead to the hyper-expression of X-linked genes. We find that genes furthest from DCC binding sites have expression levels that diverge faster than genes bound by the DCC, and this is independent of absolute expression level. Furthermore, the expression levels of genes bound by the DCC evolve slower than autosomal genes. Remarkably, we observe these patterns when expression is measured in either males or females. It was recently demonstrated that male-specific chromatin modifications associated with dosage compensation can have "bleed-over" effects into females, and our results suggest that these modifications can attenuate the evolution of expression levels in both sexes. We hypothesize that the faster-X evolution of gene expression in Drosophila is the result of looser regulation of genes that are not directly controlled by the DCC, implying that the chromatin environment can affect macro-evolutionary genomic patterns.

O-203

Transcriptional trade-offs across tissues in human evolution

Gregory Wray, Lisa Pfefferle, Olivier Fedrigo, Courtney Babbitt
Duke University, Durham, NC, USA

There are a number of prominent hypotheses in the anthropological literature concerning the importance of diet in human evolution. Comparisons with extant great apes as well as the fossil and archaeological record suggest that among the most important changes in diet there was an increase in animal products (meat and fat) and starchy plant products during human evolution. Diet is thought to have had an important influence on the human phenotype, and dietary differences have been hypothesized to contribute to the dramatic morphological changes seen in modern humans. These phenotypic differences have been hypothesized by a number of authors to be due to metabolic “trade-offs” between tissues.

In this study, we integrate the results of genomic studies within this well developed anthropological context. We used directional paired-end RNA-Seq data to assess changes in global transcript abundance in five metabolically critical tissues (cortex, cerebellum, liver, muscle, and white adipose tissue) from 4 individuals each of humans, chimpanzees, and rhesus macaques. We see an enrichment of tissue-specific GO and PANTHER categories related to metabolism (such as cholesterol metabolism in the adipose tissue) in the significantly differentially expressed genes. Using a novel multi-dimensional analysis we are quantifying the nature of these tradeoffs between all five tissues. Additionally, we have detailed changes in gene expression and evidence for adaptation in regulatory regions for two specific metabolic pathways involved in brain function: the GLUT family of glucose transporters and the phosphocreatine circuit. In both cases, we identified differential expression of multiple members of these metabolic pathways critical to increase energy availability. Identifying the specific metabolic shifts that have occurred in primate and human evolution may be a powerful tool for understanding the specific metabolic changes that have allowed for the modern human phenotype.

O-204

Polygenic *cis*-regulatory adaptation in the evolution of yeast pathogenicity

Hunter Fraser[1], Sasha Levy[1,4], Arun Chavan[2], Hiral Shah[2], Christian Perez[3], Yiqi Zhou[1], Mark Siegal[4], Himanshu Sinha[2]
[1]Stanford University, Stanford, USA,[2]Tata Institute of Fundamental Research, Mumbai, India,[3]UCSF, San Francisco, USA,[4]New York University, New York, USA

The acquisition of new genes, via horizontal transfer or gene duplication/diversification, has been the dominant mechanism thus far implicated in the evolution of microbial pathogenicity. In contrast, the role of many other modes of evolution—such as changes in gene expression regulation—remains unknown. A transition to a pathogenic lifestyle has recently taken place in some lineages of the budding yeast *Saccharomyces cerevisiae*. Here we identify a module of physically interacting proteins involved in endocytosis that has experienced selective sweeps for multiple *cis*-regulatory mutations that down-regulate gene expression levels in a pathogenic yeast. Genetic variants at these loci are associated with pathogenicity across 88 diverse yeast strains, suggesting the adaptations may have increased virulence. To test this, we created a panel of single-allele knockout strains whose hemizygous state mimics the genes' adaptive down-regulations, and measured their virulence in a mammalian host. Despite having no growth advantage in standard laboratory conditions, nearly all of the strains were more virulent than their wild-type progenitor, suggesting that these adaptations likely played a role in the evolution of pathogenicity. We also detected pleiotropic effects of these adaptations on a wide range of morphological traits, which appear to have been mitigated by compensatory mutations at other loci. These results suggest that *cis*-regulatory adaptation can occur at the level of physically interacting modules, and that one such polygenic adaptation led to increased virulence during the evolution of a pathogenic yeast.

O-205

Characterizing regulatory mechanisms underlying gene expression variation and adaptation

Athma Pai[1], Roger Pique-Regi[1], Jacob Degner[1], Carolyn Cain[1], Noah Lewellen[2], Katelyn Michelini[2], Jonathan Pritchard[1,2], Yoav Gilad[1]
[1]University of Chicago, Chicago, IL, USA, [2]Howard Hughes Medical Institute, Chicago, IL, USA

Changes in gene regulation are thought to play an important role in adaptation and speciation, especially in primates. However, the extent to which changes in different regulatory mechanisms underlie variation in gene expression levels within and between species is not yet known. To address this, we used a combination of genomic approaches to investigate the extent to which variation in transcriptional mechanisms influence both fine scale gene regulatory variation within humans and variation in gene expression levels across primates. Specifically, we assayed transcript expression levels, chromatin sensitivity, transcription factor binding footprints, DNA methylation, and markers of chromatin state across panels of human, chimpanzee, and rhesus macaque lymphoblastoid cell lines. Within humans, we found that greater than 10% of heritable gene expression levels might be due to variation in DNA methylation levels. Variation in chromatin sensitivity associated with transcription factor binding footprints could be responsible for as much as 55% of inter-individual variation in expression levels. Across primates, we found that greater than 60% of DNA methylation and chromatin sensitivity profiles are conserved in putatively functional regions. Differences in epigenetic markers and transcription factor binding footprints across species independently explain as much as 20% and 30% of inter-species differences in gene expression levels, respectively. Using these data, we were also able to characterize the extent to which changes in either *cis*- or *trans*- elements underlie inter-species differences in transcription factor binding locations. In these studies we have taken the first steps towards understanding the genetic and mechanistic basis of variation in transcriptional regulation. Our data also allowed us to develop a basic understanding of the relative importance of changes in different transcriptional mechanisms to regulatory evolution in primates.

O-206

Birth-and-Death Evolution of Multigene Families: An Overview and Update

Alejandro Rooney
U.S. Department of Agriculture, Peoria, Illinois, USA

Since it was first put forth in the mid-1990's, the birth-and-death model of evolution has been shown to provide a reasonable explanation for the observed relationships among the members of a variety of medium to large-sized multigene families. Nevertheless, work still remains to be done in certain areas. For instance, most examples of multigene families that evolve under a birth-and-death model are from eukaryotes, but the extent to which this model characterizes prokaryotic gene families is unknown. The purpose of this talk is to provide an overview of the birth-and-death model and possible avenues for future research.

O-207

Reconstruction of ancestral maltase enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication

Karin Voordeckers[1,2], Chris A. Brown[1,3], Kevin Vanneste[4,5], Elisa van der Zande[1,2], Arnout Voet[6], Steven Maere[4,5], Kevin J. Verstrepen[1,2]
[1]VIB Laboratory for Systems Biology, Leuven, Belgium, [2]CMPG Laboratory for Genetics and Genomics, K.U.Leuven, Leuven, Belgium, [3]Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, MA, USA, [4]VIB Department of Plant Systems Biology, Gent, Belgium, [5]Department of Plant Biotechnology and Genetics, Ghent University, Gent, Belgium, [6]Laboratory for Molecular en Structural Biology, K.U.Leuven, Leuven, Belgium

Gene duplications are believed to facilitate evolutionary innovation. However, the mechanisms shaping the fate of duplicated genes remain heavily debated since the molecular processes and evolutionary forces involved are difficult to reconstruct. We studied a large family of fungal glucosidase genes. We reconstructed all key ancestral enzymes and show that the very first pre-duplication enzyme was primarily active on maltose-like substrates, but also had a trace activity for isomaltose-like sugars. Structural analysis and activity measurements on resurrected and present-day enzymes show that both activities cannot be optimized in a single enzyme. However, gene duplications repeatedly spawned daughter genes in which adaptive mutations optimized either isomaltase activity or maltase activity. Interestingly, similar shifts in enzyme activity were reached multiple times via different evolutionary routes. In summary, our results provide definitive experimental evidence for the 'Escape from Adaptive Conflict' model of duplicate gene evolution.

O-208

Functional evidence that a newly evolved *Drosophila* sperm-specific gene boots sperm competition

Shu-Dan Yeh[1], Tiffanie Do[1], Carolus Chan[1], Adriana Cordova[1], Francisco Carranza[1], Eugene Yamamoto[1], Mashya Abbassi[1], Kania Gandasetiawan[1], Pablo Librado[2], Elisabetta Damia[3], Patrizio Dimitri[3], Julio Rozas[2], Daniel Hartl[4], John Roote[5], Jose Ranz[1]
[1]University of California, Irvine, CA, USA, [2]Universitat de Barcelona, Barcelona, Spain, [3]Sapienza Università di Roma, Rome, Italy, [4]Harvard University, Cambridge, MA, USA, [5]University of Cambridge, Cambridge, UK

In many species, both morphological and molecular traits related with sex and reproduction evolve faster in males than in females. Ultimately, rapid male evolution relies on the acquisition of genetic variation associated with differential reproductive success. Many newly evolved genes are associated with novel functions that might enhance male fitness. Nevertheless, functional evidence of the adaptive role of recently originated genes in males is still lacking. The *Sdic* multigene family, which encodes a sperm dynein intermediate chain presumably involved in sperm motility, originated from complex genetic rearrangements in the lineage that leads to *D. melanogaster* within the last 5.4 myr since its split from *D. simulans*. We have deleted all the members of this multigene family resident on the *X* chromosome of *D. melanogaster* by chromosome engineering and found that, although the deletion does not result in a reduction of the progeny number, it impairs the competence of the sperm in the presence of sperm from wildtype males. Therefore, the *Sdic* multigene family epitomizes the birth and integration of a new gene function into the genetic network underlying male-related fitness over a very short evolutionary timespan. Taken together, our results not only uphold the notion that Darwinian selection actually operated in this genomic region but also provide functional evidence for the relevance of *Sdic* in intrasexual selection, an evolutionary mechanism that fuels the steady and rapid evolution of genomes.

O-209

Inference of co-evolving Drosophila gene family pairs by likelihood-based methods

Pablo Librado[1,3], Filipe G. Vieira[2], Julio Rozas[1,3]
[1]Departament de Genètica, Universitat de Barcelona, Barcelona, Spain, [2]Department of Integrative Biology, University of California Berkeley, Berkeley, USA, [3]Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Barcelona, Spain

Most gene families are highly dynamic, and usually evolve under a birth-and-death process, where genes mainly arise by duplication (birth) and are eliminated via deletion or pseudogenization (death) [1, 2]. Currently, several stochastic birth-and-death models and programs have been developed to analyze gene family dynamics. Nevertheless, these programs have some drawbacks, including unrealistic assumptions about gene family evolution, or the inability to specify branch models.

We have developed BadiRate [3], a likelihood-based method that implements five different stochastic population models to estimate family turnover rates (birth, gain, death, innovation and lambda), as well as the most likely ancestral family sizes. Moreover, the method allows detecting expanding/contracting gene families and phylogenetic lineages with extreme turnover rates. These families and lineages are good candidates to further analyse the evolutionary processes shaping gene copy number variation. Here, we have used BadiRate and the mirror-tree approach [4] to infer functional interactions among Drosophila gene family pairs, assuming that co-evolving gene family pairs exhibit correlated turnover rates. In particular, we first used BadiRate to estimate the turnover rates for each gene family and phylogenetic lineage. Then, for all the gene family pairs, we calculated the correlation of the turnover rates. Lastly, after correcting by multiple testing and for the evolutionary relationship of the Drosophila species (which might produce some false positive correlations), we obtained the gene family pair candidates for functional interaction.

1. Nei M, Rooney AP: Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 2005, 39:121-152.
2. Vieira FG, Rozas J: Comparative Genomics of the Odorant-Binding and Chemosensory Protein Gene Families across the Arthropoda: Origin and evolutionary history of the chemosensory system. Genome Biol Evol 2011, 3.
3. Librado P, Vieira FG, Rozas J: BadiRate: Estimating Family Turnover Rates by Likelihood-Based Methods. Bioinformatics 2012, 28(2):279-281.
4. Pazos F, Valencia A: Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Engineering 2001, 14(9):609-614.

O-210

Polycistronic microRNAs arise by tandem duplication and *de novo* emergence

Antonio Marco, Sam Griffiths-Jones
University of Manchester, Manchester, UK

MicroRNAs, small endogenous RNA molecules, are emerging as crucial regulators of gene expression. Approximately half of the microRNAs in animals are clustered in the genome. Clustered microRNAs are often transcribed as a single RNA under the control of a single promoter. These polycistronic microRNA transcripts are believed to encode mostly functionally related microRNAs. This is analogous to prokaryotic operons, in which proteins involved in the same biochemical pathway are often controlled by the same promoter. On the other hand, comparative genomics shows that the relatively rare operons seen in eukaryote genomes are a by-product of genome reduction. However, microRNAs are commonly clustered in even large genomes. So, which evolutionary forces have driven the formation of microRNA clusters? To tackle this question we reconstructed the evolutionary history of all microRNA clusters in *Drosophila melanogaster*. We make the following observations: (1) clusters arise by either tandem duplication of existing microRNAs or *de novo* emergence of microRNAs near existing ones; (2) not a single case of cluster formation by fusion of non-clustered microRNAs can be seen; (3) likewise, we do not see any events that have split a cluster into smaller clusters or single microRNAs; (4) although the proportion of microRNAs located in clusters is relatively constant within animal genomes, only a few clusters are conserved across evolutionarily distant species; (5) non-homologous clustered microRNAs are mostly functionally unrelated as they have different targeting properties. Our observations can be summarized in the following simple model. Clusters of microRNAs appear by non-adaptive tandem duplication or *de novo* microRNA emergence. If the new microRNA in the cluster acquires a function, purifying selection keeps this cluster together, since these microRNAs are under the control of the same promoter. Otherwise, clustered microRNAs follow a birth-and-death dynamics driven by random drift. Further explorations of the human and *Caenorhabditis* genomes are also consistent with our model. This work represents the first evolutionary study of the dynamics of polycistronic microRNAs in animal genomes, and has wide-ranging implications for our understanding of evolutionary and functional genomics.

O-211

Characterizing the effect of multiple sequence alignment on downstream analyses

Simon Whelan, Ben Blackburne
University of Manchester, Manchester, UK

Many molecular evolutionary analyses start with a multiple sequence alignment, which is usually accepted as known despite wide recognition that errors may impact downstream phylogenetic analysis. Many statistical methods in molecular evolution have been developed to (e.g.) estimate phylogenetic trees or infer adaptation, but these methods are dependent on the accuracy of sequence alignment. Several studies have demonstrated that the results obtained are dependent on the sequence alignment chosen. In cases where systematic error occurs, these differences can be attributed to non-homologous characters being placed together.

To characterize properties of multiple sequence alignment and its effect on downstream evolutionary analysis we examine 200 sets of sequences extracted from The Adaptive Evolution Database, with strict sampling criteria to ensure high quality sequences and reasonable evolutionary divergence. For each set of sequences we apply a range of out-of-the-box popular multiple sequence alignment tools, splitting broadly into 'algorithm-based' aligners (e.g. ClustalW; Muscle; ProbCons; MAFFT; T-Coffee) and phylogenetically-aware aligners (Prank and BAliPhy). We also include samples from the posterior distribution of the statistical aligner BAli-Phy to quantify the degree of uncertainty associated with alignment. We apply recently developed metrics to investigate the similarities and differences between these alignments, finding under multidimensional scaling that algorithm-based and phylogenetically-aware aligners tend to form discrete clusters.

To investigate the effect of alignment on downstream methods we examine two common evolutionary analyses: the inference of a maximum-likelihood tree and a test for adaptive evolution (M7 vs M8 in PAML). For tree estimation algorithm-based and phylogenetically-aware aligners tend to yield noticeably different results, and the distances between alignments seems reasonably correlated with the geodesic distance between trees. To examine the effect of alignment on the inference of adaptive evolution we develop a conservative marginal likelihood approach for integrating across the uncertainty in alignment, based on samples from the statistical aligner BAliPhy. We find that positive results from phylogenetically-aware aligners tend to agree with the results from our marginal likelihood approach, with moderate numbers of additional inferences. Positive results from algorithm-based aligners tend to include those from the marginal likelihood approach, but also include large numbers of additional inferences.

O-212

Alignment error and its impact on the sitewise detection of positive selection

Gregory Jordan, Nick Goldman
EMBL European Bioinformatics Institute, Hinxton, Cambridgeshire, UK

Errors in alignment are expected to cause errors in detecting positive selection, but the magnitude and nature of this impact is poorly understood. Focusing on the detection of sitewise positive selection, we used a simulation approach to measure the effect of alignment error and to identify trends with respect to sequence divergence, tree size, aligner, and the rate of biological insertions and deletions.

We found that some aligners were prone to produce false positive results even at a relatively low sequence divergence and insertion and deletion rate, while other aligners yielded few false positives under all but the most extreme conditions. On the other hand, an increase of false negatives with sequence divergence was universal. We also tested the ability of a number of alignment filters to improve power by removing alignment errors; most filters showed little, if any, improvement over the best unfiltered alignments.

These results provide a baseline quantification of expected error rates due to misalignment across a wide range of plausible parameter ranges. Furthermore, they support the view that non-biological factors, such as sequencing error and assembly error, may be more likely to cause false positives than biological indels in many large-scale studies. Future work will focus on the design of methods for removing non-biological sources of alignment error that occur during the sequencing, assembly and annotation of genes across many species.

O-213

Detecting and correcting alignment overcompression

Daniel Money, Mark Holder
University of Kansas, Lawrence, KS, USA

A reliable sequence alignment is a pre-requisite for most phylogenetic studies, and it has been shown that a sub-optimal alignment can have a significant effect on tree inference and other downstream analyses. Although previous work has shown that alignment errors can have significant consequences, there has still been relatively little research into detecting and fixing such errors - especially when compared to the significant amount of research on inferring good trees. We propose and investigate a method for detecting and correcting a specific class of alignment errors: overcompression. Each column in an alignment represents residues that should be inherited from a common ancestor. Under a strict interpretation of homology, each site in an alignment can be explained at most one insertion, but possibly multiple deletions. In a compressed alignment, residues inherited from an insertion event appear in the same column as residues that did not arise from that event. Our method compares site-likelihoods under a model that allows multiple insertions to site-likelihoods under a form of stochastic Dollo model which prohibits multiple insertions in the same column. Site-likelihood comparisons that favour multiple insertions can be used to guide the splitting of a single column into two or more columns in which all of the residues do appear to be homologous. Study of recent additions to TreeBase suggest that a large proportion of alignments used in studies in the past year suffer from overcompression. We study the effect on downstream analyses of correcting these apparent errors.

O-214

Facultative exons are enriched for adaptively evolving codons, but much of the pattern is driven by alignment errors

Ryan McGee, Matthew Dean
Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA

Mutations that are beneficial in one context but deleterious in another have a more difficult time reaching fixation than those that are universally beneficial. In this context, we predicted that adaptive evolution more frequently affected codons that occur in facultative exons, with the idea that facultative exons might participate in fewer pleiotropic processes than constitutive exons, which are included in every transcript of a gene. To test this hypothesis, we perform genome-wide estimates of adaptive evolution from six complete mammalian genomes - human, dog, human, cow, mouse, rat, and rabbit. We employ six different alignment strategies meant to control alignment errors at different levels. We find that facultative exons are significantly enriched for codons experiencing recurrent positive selection, however most of the pattern is driven by alignment errors. In other words, facultative exons are more subject to alignment errors, which leads to more false positive identification of recurrent and/or rapid evolutionary change. In addition to revealing an excess of adaptive evolution acting on facultative exons, this study underscores the importance of controlling alignment errors that will falsely identify codons as adaptively evolving.

O-215

DACTAL: Divide-and-conquer trees (almost) without alignments.

Tandy Warnow[1], Serita Nelesen[3], Kevin Liu[2], Li-San Wang[4], C. Randal Linder[1]
[1]University of Texas at Austin, Austin, TX 78712, USA, [2]Rice University, Houston, TX, USA, [3]Calvin College, Grand Rapids, MI, USA, [4]University of Pennsylvania, Philadelphia, PA, USA

Large-scale phylogeny alignment, especially of datasets with many thousand sequences, can be highly inaccurate due to alignment error. The SATe method of Liu et al (Science 2009) can co-estimate trees and alignments and can run on very large datasets with 10,000 sequences, but larger datasets can be quite computationally challenging. Here, we introduce a new method, called DACTAL. DACTALis an iterative method, where each iteration utilizes a divide-and-conquer method, constructs alignments and trees only on subsets of 250 sequences at a time, and then mergers the trees together into a tree on the full dataset. We show that DACTAL produces extremely accurate trees, matching the accuracy of SATe on large biological and simulated datasets, but achieving the same accuracy in only 10% of the time. DACTAL outperforms all two-phase methods (that first align and then estimate a tree on the alignment) in terms of accuracy. DACTAL can also analyze datasets with 28,000 taxa that are too big for SATe (each iteration of SATe uses RAxML, which is computationally limiting). DACTAL, however, achieves this accuracy without ever estimating an alignment on the full dataset.

This result shows that large-scale tree estimation is feasible for very large datasets, even without the need for a full multiple sequence alignment of the entire dataset.

O-216

Maternal siRNAs as regulators of parental genome imbalance and gene expression during seed development in angiosperms

Jie Lu[1], Changqing Zhang[1], David Baulcombe[2], Z. Jeffrey Chen[1]
[1]*The University of Texas at Austin, Austin, Texas, USA,* [2]*University of Cambridge, Cambridge, UK*

In angiosperms the endosperm is formed after pollination of the egg by a male gamete (pollen) that contains two sperm nuclei. One sperm fertilizes the egg to form a zygote with a 1:1 maternal to paternal genome ratio (1m: 1p), whereas the other fertilizes two central cell nuclei to form an endosperm cell with a 2:1 maternal to paternal genome ratio (2m:1p). In *Arabidopsis thaliana*, increasing the paternal genome ratio (2m:2p) in endosperm by pollinating a diploid "mother" with a tetraploid "father" (2X4) delays endosperm cellularization (EC) and produces larger seeds. In contrast, increasing the maternal genome ratio (4m: 1p) in endosperm by pollinating a tetraploid mother with a diploid father (4X2) leads to precocious EC and smaller seeds. The mechanisms for responding to the parental genome dosage imbalance and for gene expression changes in endosperm are unknown. In plants, RNA polymerase IV (PolIV or p4) encoded by *NRPD1a* is required for biogenesis of a major class of 24-nt siRNAs (p4-siRNAs), which are predominately expressed in developing endosperm. Here we show that p4-siRNA accumulation depends on the maternal genome dosage, and maternal p4-siRNAs target transposable elements (TEs) and TE-associated genes (*TAGs*) in seeds. The p4-siRNAs correlate negatively with expression levels of AGAMOUS-LIKE (*AGL*) genes in endosperm of interploidy crosses. Moreover, disruption of maternal *NRPD1a* expression is associated with p4-siRNA reduction and *AGL* upregulation in endosperm of reciprocal crosses. This is the first genetic evidence for maternal siRNAs in response to parental genome imbalance and in control of transposons and gene expression during endosperm development.

O-217

Using semi-natural enclosure experiments to study the role of imprinting and epigenetic modifications in the evolution of house mouse populations

Diethard Tautz
Max-Planck Institute for Evolutionary Biology, Plön, Germany

We have performed a series of semi-natural enclosure experiments with house mice (*M. m. domesticus*) derived from wild caught populations to assess the role of imprinting in generating population divergence and the role of epigenetic modifications in response to environmental changes. Groups of mice could freely reproduce for several generations in the enclosures and mate choice preferences as well as epigenetic modifications were traced. In a first set of experiments, we found evidence for a role of paternal imprinting in differential mate choice associated with populations that have only recently diverged. This implies that fast evolving imprinted cues play a role in population divergence. In a second experiment we were interested in epigenetic chromatin modifications associated with environmental changes. Mice were subjected to different environmental conditions (high energy food and different daylight regime) for several generations and then returned to standard conditions. Studying histone H3K4 methylation (a signal for active chromatin) via CHIP-Seq analysis, we found evidence that changes in histone methylation status of promotors occur in the comparison between high energy and normal diet, with many of the affected genes being involved in metabolic processes. A subset of these changes can still be traced in the offspring of animals after returning them to standard conditions. We conclude that imprinting and epigenetic modifications may be of relevance when studying short term adaptations.

PIs on the projects: Angelika Boersch-Haubold (epigenetics), Inka Montero and Meike Teschke (mate choice and imprinting)

O-218

Brain-specific epigenetic adaptations in humans

Jessica L. Crisci[1], Hennady P. Shulla[1], Denis Reshetov[2], Iris Cheung[1], Rahul Bharadwaj[1], Hsin Jung Chou[1], Isaac Houston[1], Cyril J. Peter[1], Wei-Dong Yao[3], Richard H. Meyers[4], Jiang-fan Chen[4], Todd Preuss[5], Evgeney Rogaev[1,2], Jeffrey D. Jensen[1,6], Zhiping Weng[1], Schahram Akbarian[1]
[1]University of Massachusetts Medical School, Worcester, MA, USA, [2]Vavilov Institute of General Genetics, Moscow, Russia, [3]New England Primate Center, Southboro, MA, USA, [4]Boston University, Boston, MA, USA, [5]Yerkes National Primate Research Center/Emory University, Atlanta, GA, USA, [6]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

The majority of functionally validated adaptive mutations in humans are located within coding regions of the genome. Indeed, this is often the first (or only) place researchers look in order to explain species- or population- specific adaptation.. However, there is evidence of selective sweeps throughout the genome, not all of which correspond to coding changes. This suggests that adaptive phenotypes in humans may not all correspond to changes in protein-coding sequence, and indeed there is emerging interest in the possibility of adaptation in the epigenome. In order to begin to evaluate this hypothesis, we focus here on the observed paradox of deep conservation of neuronal genes across primates, and the rapid phenotypic evolution of the human brain. We compare peaks of trimethylated histone H3-lysine4 (H3K4met3) from the prefrontal neurons of human, chimpanzee, and macaque. These peaks are often found near transcriptional start sites and 5' ends of genes. We highlight 33 peaks which show 2-fold greater expression in humans than in chimpanzee and macaque, and similar H3K4met3 landscapes between all 3 species in non-neuronal tissues - suggesting a human-specific expression pattern in neurons. Of these 33 peak regions, 8 show an accelerated nucleotide substitution rate along the human branch. We discuss these results in the light of human specific cognitive abilities, and the relative importance of genetic vs. epigenetic adaptation.

O-219

Coordinated histone modifications and small RNAs in gene and genome copy number variations

Misook Ha
Samsung Advanced Institute of Technology, Yonging-si, Gyeonggi-do, Republic of Korea

Genome segmental duplication is a fundamental evolutionary mechanism for most organisms including many plants and some animals and in human diseases. Duplicate genes and copy number variations from segmental genome duplication are major sources of expression and functional diversity of genes. To investigate evolutionary roles of retained duplicate genes, we examined how duplicate genes respond to developmental and environmental changes within species and how ancient duplicate genes contribute to gene expression diversity in resynthesized allopolyploids. We found that expression divergence between gene duplicates was significantly higher in response to environmental stress than to developmental process. Furthermore, duplicate genes related to external stresses show higher expression divergence between two closely related species and in resynthesized and natural allotetraploids than single-copy genes. A slow rate of expression divergence of duplicate genes during development may offer dosage-dependent selective advantage, whereas a high rate of expression divergence between gene duplicates in response to external changes may enhance adaptation. To investigate epigenetic mechanisms of expression diversity in hybrids containing multiple divergent genomes, we analyze high-throughput sequencing data of small RNAs and chromatin modifications combining gene expression measurements in hybrids and the parental species. Our results suggest that stable inheritance of parental siRNAs in hybrids helps maintain genome stability in response to genome duplication, whereas expression diversity of miRNAs leads to interspecies variation in gene expression, growth, and development. Integrating ChIP-seq and ChIP-chip of histone modifications, we find distinct patterns of histone modifications at the genic region, and observe that the chromatin modification patterns are marker of gene expression profiles and functions. Moreover, dynamic remodeling of histone acetylations mediated by histone deacetylases is a key mechanism of expression variation within and between species. Our data suggest that coordinated modifications of histone acetylations and methylations and miRNAs and siRNAs provide epigenetic and evolutionary mechanisms for gene expression diversity and morphological variations in genomic structural changes.

O-220

Deciphering the porcine imprintome

Ole Madsen[1], Djawad Radjabzadeh[1], Hendrik-Jan Megens[1], Mirte Bosse[1], Laurent Frantz[1], Yogesh Paudel[1], Richard Crooijmans[1], Laurie Rund[2], Lawrence Schook[2], Martien Groenen[1]
[1]Wageningen University, Wageningen, The Netherlands, [2]University of Illinois, Illinois, USA

Genomic imprinting is an epigenetic phenomenon in which the level of allelic expression is dependent upon parental origin. In mammals, >100 genes have been shown to be imprinted, mainly through studies with humans and mice, and recently as many as 1,300 loci were identified as being expressed in a parentally biased manner in the mouse brain. Studies in other mammals for imprinted genes are relatively sparse and have generally been limited to either an *ad hoc* single gene analysis or analyses on "unnatural" systems such as uniparental embryos. Imprinted genes play important roles in many stages of development, and dysregulation can result in diseases such as cancer and neurological disorders. Comparative studies, mainly between human, mouse and amongst marsupials, indicate a relatively large fraction of imprinted genes as clade- or species-specific. This suggests that the evolution of the imprintome is a dynamic and stepwise ongoing adaptive process that is still poorly understood. Since the pig is widely used as a model organism for human diseases and biomedical research and phylogenetically distant from the well studied rodents and humans, a comprehensive investigation of imprinted genes in pigs contributes to a more detailed understanding of the evolution of imprinted genes in relation to their key role in development and reproduction.

RNA-seq is a powerful tool to analyze allele specific expression and therefore provides an excellent opportunity for detecting imprinted genes, by comparing the genetic variation present in a genome with the variation in allelic expression. By combining this information with the variation present in the genome of the parents, it will be possible to detect imprinted genes.

This study aims to detect the entire range of imprinted genes (the imprintome) in pigs. This is being done using whole genome sequencing of two family trios (father, mother and offspring) plus RNA-seq of tissues from the two offspring (starting out with placenta, brain and muscle). Results from these analyses will be presented and related to the imprintome of other eutherian mammals.

O-221

The Evolution of Species Interaction Networks

Miguel A. Fortuna
Estacion Biologica de Doñana (EBD-CSIC), Seville, Spain

Species do not live in isolation but are part of a broader ecological community that researchers often depict as a network. Although the structure of these ecological networks has recently been well described and we are starting to understand its dynamical implications, we know almost nothing about the role evolution played in shaping these species interaction networks. I will start by giving an overview of the signal left by evolution on the pattern of who interacts with whom in food webs, plant-animal mutualistic networks, and host-parasite communities. In order to complement this static approach, we will go a step further exploring the dynamics of the coevolutionary process in species rich communities. These ongoing dynamics will be tackled using artificial life approaches that will help us understand the evolutionary emergence of the nested structure in host-parasite antagonistic networks we observe in nature.

O-222

Network analysis of ocean plankton communities: from viruses to fish larvae

Gipsi Lima-Mendez[1,2]
[1]*Vrije Universiteit Brussel (VUB), Brussels, Belgium,* [2]*VIB, Brussels, Belgium*

Metagenomics and rRNA gene sequencing are powerful tools for the analysis of microbial communities. Due to the molecular nature of these techniques ecological questions can be addressed at the genetic level. Because of its

complexity, meta-omics data have required the development of novel computational tools to analyze and determine the

functional and phylogenetic composition of the sampled community. However, current tools are not sufficient to go from a metagenomic 'parts list' (i.e. a bag of genes) to an initial understanding of the ecosystem structure and functioning. Using networks, here I represent pair-wise relationships between taxonomic units that tend to co-occur across different samples. These networks are analyzed to identify community structure across different environmental gradients, predict the most relevant members of the community and the type of interaction between linked pairs of organisms. Here I will present the results of applying such methods to data gathered during the Tara oceans expedition [1].

1. Karsenti, E., et al., *A holistic approach to marine eco-systems biology.* PLoS Biol, 2011. **9**(10): p. e1001177.

O-223

Looking for hints of new Domains of Life with similarity networks

Philippe Lopez[1], Sébastien Halary[3], Eric Bapteste[2]
[1]Université Pierre et Marie Curie, Paris, France,[2]CNRS, Paris, France,[3]Institut de Recherche en Biologie Végétale, Montréal, Canada

Metagenomics data have revealed an unprecedented amount of environmental genetic diversity, mostly coming from uncultured organisms. The extent of this diversity is so great that a large number of these environmental sequences bear no similarity to previously known sequences, and are thus called metaorfans. An explanation for the existence of metaorfans could be that they come from highly divergent but still unknown organisms. Levels of divergence higher than those observed between the three Domains of Life could then hint for new Domains out in the environment. Phylogenetic approaches, requiring that all sequences under study be homologous one to another, are by definition unable to accommodate metaorfans. Similarity (or homology) networks, however, are much less constrained and can detect extremely divergent variants of known sequences. Here, we devised an original exploratory network-based strategy to structure the environmental sequence space, looking for very divergent sequences possibly hinting for novel life forms outside the three Domains of Life through a simultaneous study of genetic diversity from cultured organisms and environmental samples. We selected 10,822 sequences from 86 extremely conserved gene families, and compared them to more than 5 million predicted ORFs coming from microbial metagenomes. The resulting similarity networks show that very distant homologs of known conserved proteins are indeed present in large numbers in the environment, opening the possibility of new Domains of Life.

O-224

Reticulate evolutionary history of eudicot model species

Martin Lercher, Christian Esser, Daniel Hartleb
Heinrich-Heine-University, Duesseldorf, Germany

Plant genomes show extensive evidence of multiple past polyploidizations ('whole-genome duplications'). Polyploidization via fusion of unreduced gametes – either from the same or from separate species – was likely involved in at least 15% of speciations in flowering plants. At the same time, genome size remained roughly constant due to the subsequent loss of most duplicates. Both hybridisation and reciprocal deletion of DNA in different lineages after duplication may have lead to distinct evolutionary histories of chromosomal regions. However, evolutionary relationships among plant model species are still portrayed using bifurcating phylogenetic trees. These trees are based dominantly on plastid gene sequence analysis and hence reflect maternal inheritance only.

To test for deviations from tree-like evolution, we analysed genomic datasets of 9 plants and algae. We find that alignments of individual nuclear-encoded eudicot gene families tend to support one of two drastically different tree topologies with almost equal probability. Only one of these trees corresponds to the maternal tree. Genes supporting the different topologies tend to form genomic clusters, and chromosomal regions tend to show synteny signals consistent with the regionally dominant tree. Thus, sequence and gene order evolution provide independent, consistent evidence for drastically different evolutionary histories of neighbouring genomic regions within eudicot plants. This observation enforces a general re-thinking of plant history in terms of reticulate evolution, either caused by orchestrated deletions of genomic regions or by wide hybridisations.

O-225

Difference in gene duplicability may explain the difference in overall structure of protein-protein interaction networks among eukaryotes

Takeshi Hase[1,2], Yoshihito Niimura[1]
[1]Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, [2]The Systems Biology Institute, Tokyo, Japan

A protein-protein interaction network (PIN) was suggested to be a disassortative network, in which interactions between high- and low-degree nodes are favored while hub-hub interactions are suppressed. It was postulated that a disassortative structure minimizes unfavorable cross-talks between different hub-centric functional modules and was positively selected in evolution. However, by re-examining yeast PIN data, several researchers reported that the disassortative structure observed in a PIN might be an experimental artifact. Therefore, the existence of a disassortative structure and its possible evolutionary mechanism remains unclear. In this study, we investigated PINs from the yeast, worm, fly, human, and malaria parasite including four different yeast PIN datasets. The analyses showed that the yeast, worm, fly, and human PINs are disassortative while the malaria parasite PIN is not. By conducting simulation studies on the basis of a duplication-divergence model, we demonstrated that a preferential duplication of low- and high-degree nodes can generate disassortative and non-disassortative networks, respectively. From this observation, we hypothesized that the difference in degree dependence on gene duplications accounts for the difference in assortativity of PINs among species. Comparison of 55 proteomes in eukaryotes revealed that genes with lower degrees showed higher gene duplicabilities in the yeast, worm, and fly, while high-degree genes tend to have high duplicabilities in the malaria parasite, supporting the above hypothesis. These results suggest that disassortative structures observed in PINs are merely a byproduct of preferential duplications of low-degree genes, which might be caused by an organism's living environment.

O-226

The evolution of mammalian tissue transcriptomes

Henrik Kaessmann[1,2]
[1]*University of Lausanne, Lausanne, Switzerland,* [2]*Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Shared mammalian traits include lactation, hair and relatively large brains with unique structures. Individual lineages have, in turn, evolved distinct anatomical, physiological and behavioral characteristics that relate to differences in reproduction, life span, cognitive abilities and disease susceptibility. The molecular changes underlying these phenotypic shifts and the associated selective pressures have begun to be investigated based on a growing number of fully sequenced mammalian genomes. Genome analyses may uncover protein-coding changes that potentially underlie phenotypic alterations. However, regulatory mutations affecting gene expression probably explain many or even most phenotypic differences between species. The advent of high-throughput RNA sequencing (RNA-seq) approaches now allows for accurate and sensitive assessments of transcript sequences and expression levels at a genome-wide scale. We have generated comprehensive sets of RNA-seq data for a large collection of germline and somatic tissues from representatives of all major mammalian lineages (placental mammals, marsupials, and the egg–laying monotremes) and evolutionary outgroups (e.g., birds). On the basis of these data, we are investigating various aspects of transcriptome evolution, including the evolution of protein-coding gene expression levels, long noncoding RNAs, microRNAs, alternative splicing, and dosage compensation. I will present highlights of these analyses.

O-227

Evolution of transcription factor binding and gene expression in the genus *Drosophila*

Mathilde Paris, Tommy Kaplan, Susan Lott, Jacqueline Villalta, Michael Eisen
University of California Berkeley, Berkeley, France

To better characterize how variation in regulatory sequences drives divergence in gene expression, we undertook a systematic study of transcription factor binding and gene expression in four species that sample much of the diversity in the 60 million-year old genus Drosophila: *D. melanogaster*, *D. yakuba*, *D. pseudoobscura* and *D. virilis*. We used ChIP-Seq to examine the genome-wide binding of four transcription factors (BCD, HB, GT and KR) regulating segmentation along the anterior-posterior axis. As expected, genome-wide transcription factor binding divergence correlates with phylogenetic distance; regions more highly bound in *D. melanogaster* were more likely to be bound in other species, and to have similar overall levels of binding; and increases/decreases in binding were associated with gain/loss of transcription factor binding sites.

To examine the consequences of these changes, we used single embryo mRNA-Seq to measure gene expression in sex individual blastoderm embryos of each species. Surprisingly, we found relatively few changes in gene expression, suggesting that differences in sequence and binding have limited effect on gene expression or act in a compensatory manner to maintain the overall expression levels of regulated genes. Using *in situ* hybridization, we also compared the expression pattern of genes showing variation in nearby regulatory TF binding among species. Finally, we used an evolutionary model of quantitative traits to link the evolution of gene expression with the evolution of regulatory TF binding. This analysis unravels the evolutionary links between various levels of transcriptional regulation, from DNA sequence to gene expression through protein binding.

O-228

Different selective forces on cis and trans acting factors shape gene expression evolution in yeasts

Bernhard Schaefke[1,2], J.J. Emerson[3], Tzi-Yuan Wang[4,2], Wen-Hsiung Li[2,5]
[1]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan, [2]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan, [3]Center for Theoretical Evolutionary Genomics, University of California, Berkeley, Berkeley, California, USA, [4]Genomics Research Center, Academia Sinica, Taipei, Taiwan, [5]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA

Gene expression differences between organisms can be caused by differences in factors acting either in cis (e.g., promoters) or in trans (e.g., transcription factors) or a combination of both (e.g., interactions between transcription factors and their binding sites). For understanding processes like speciation and the evolution of new phenotypes it is crucial to understand the selective forces acting on these regulatory elements and how they shape gene expression variation within species as well as divergence between species. Previous studies suggest that cis regulatory differences play a dominant role in interspecific divergence, but the relative contributions of positive Darwinian selection and selective constraint as well as the evolutionary dynamics of cis-trans interactions remain active areas of research.

We investigated gene expression differences between two Saccharomyces cerevisiae strains, RM and BY, and estimated the contribution of cis and trans acting factors, using a hybrid experimental approach to obtain large amounts of transcriptome data. A comparison of these data with gene expression divergence between S. cerevisiae and S. paradoxus showed a marked difference between cis and trans regulatory elements. While trans regulators of essential and conserved genes are subjected to stronger selective constraint, differences in cis elements largely conform to a neutral model of gene regulatory evolution.

Furthermore, we conducted an analysis of transcription factor and promoter sequence evolution in both S. cerevisiae strains and S. paradoxus and evaluated instances in which positive selection might have contributed to gene expression differences as well as cases of cis-trans interactions and compensatory evolution.

Additionally, we examined the relationship between regulatory differences in cis and trans and the inheritance patterns of total gene expression levels in hybrids. Our results indicate that misexpressed genes, with either higher or lower total expression levels in the hybrid than in both parental strains, are more likely to exhibit antagonistic or compensatory cis-trans interactions. In contrast to previous studies there was no evidence that cis regulatory changes contribute more to expression differences in genes that show additive inheritance in hybrids as compared to those with non-additive inheritance patterns.

In summary, this study provides new insights into how selective forces on cis and trans elements shape the evolution of gene expression in yeasts.

O-229

An Understanding of Alternative Splicing is Essential in Understanding Gene Evolution

Toby Hunt, Jonathan M Mudge, Adam Frankish, Jennifer Harrow
Wellcome Trust Sanger Institute, Cambridge, UK

Comparative analysis of alternative splicing (AS) can help us understand a mode of gene evolution that remains largely unexplored and therefore underappreciated. As part of the GENCODE consortium we are creating the reference human gene annotation for the ENCODE project, using highly descriptive manual annotation from the HAVANA group supplemented by computationally generated Ensembl gene models. This annotation has been adopted by studies such as the 1000 genomes project and the cancer genome project.

Our annotation captures the extent to which AS expands the human transcriptome; we observe 6.3 AS variants for every protein-coding locus. However, the human genebuild will not be complete until we have separated those AS transcripts that represent 'noise' from those that make a genuine contribution towards phenotypic complexity. In the absence of experimental data, the evolutionary conservation of an AS event provides a strong proxy for the confirmation of its functionality. Since we are annotating the mouse and zebrafish genomes in parallel to human, we have the opportunity to undertake an in-depth comparative analysis of the conservation and evolution of AS, and to estimate its true contribution to functionality.

Our data confirm that protein-coding genes contain both ancient and more recent AS events; we observe events common to vertebrates, others restricted to mammals and many that are lineage-specific. Clearly, if we consider a protein-coding locus to be represented by a single transcript encoding a single protein produce, we will typically be under-representing the complex evolutionary changes present. For example, the main protein product of the *RBM39* gene is under strong purifying selection, allowing for the claim that this locus is inert in evolutionary terms. However, when AS is considered we observe large variation between species and clear evidence of significant evolutionary change. Here, we will describe the variety of methods by which genes can accumulate new AS events, and our continuing efforts to infer the functional potential of these transcripts.

Our work demonstrates that a complete knowledge of the transcriptome of a species is required to truly understand its genome and that understanding the true variation between the transcriptomes of different species is necessary when interpreting the evolvability of a gene between species.

O-230

Testing the link between human cognitive evolution and cognitive disorders

Xiling Liu[1], Philipp Khaitovich[1,2]
[1]Partner Institute for Computational Biology, Chinese Academy of Sciences, 320 Yue Yang Road, 200031 Shanghai, China,[2]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

Human cognitive abilities are outstanding among primates. Healthy development of the human brain is indispensable for successful acquisition of these abilities. Here, we analyzed changes in gene expression and chromatin modifications associated with healthy human brain and compared them to the changes found in the brains of non-human primates, as well as patients suffering from two common cognitive disorders: autism and schizophrenia. We found that genes showing human-specific expression in the developing brain are significantly over-represented among genes showing pathological changes in these two debilitating cognitive disorders. Specially, the average expression and chromatin modification levels of these genes are consistently higher in humans, compare to other primates or to patients suffering from autism or schizophrenia. Functionally, many of these genes are associated with synapses and synaptic transmission. The regulation mechanisms controlling these changes were also explored. Taken together, this study provides new insights into molecular mechanisms responsible for human cognitive function and dysfunction.

O-231

Estimating gene gain and loss rates in the presence of error in genome assembly and annotation

Matthew Hahn
Indiana University, Bloomington, IN, USA

Next-generation sequencing methods produce large amounts of data, but the resulting genome assemblies are often woefully incomplete. These incomplete and error-filled assemblies result in many annotation errors, especially in the number of genes present in a genome. In this talk I describe the magnitude of the problem, both in terms of total gene number and the number of copies of genes in specific families. Then I describe a method for taking these errors into account, allowing one to accurately infer rates of gene gain and loss among genomes with heterogeneous quality levels. I demonstrate the utility of this method (made available in the newest version of the software package CAFE) on many newly sequenced mammalian genomes.

O-232

Uncovering the molecular basis of the centromere paradox

Benjamin Ross[1,2], Harmit Malik[1,2]
[1]*University of Washington, Seattle, USA,* [2]*Fred Hutchinson Cancer Research Center, Seattle, USA*

Centromeres are chromosomal loci required for chromosome segregation during mitosis and meiosis in all eukaryotes. Despite this functional conservation, centromeric DNA varies dramatically in both sequence and abundance across taxa and even between closely related species. Furthermore, proteins that interact with centromeres evolve rapidly in plants and animals. This paradox of rapid evolution but conserved function is exemplified by *CenH3*, which encodes the centromere-specific histone that replaces canonical histone H3 in centromeric nucleosomes, and is essential for chromosome segregation. In stark contrast to canonical H3, which evolves under purifying selection, *CenH3* evolves rapidly under positive selection in many taxa, including *Drosophila*, primates, and plants. How can the centromere paradox be reconciled? The decade-old centromere-drive hypothesis posits that competition between centromeric DNA sequences for transmission during female meiosis (centromere drive) can result in reduced male fitness. This could place strong selective pressure on genes that encode centromere-binding proteins (like *CenH3s*) to act as suppressors of drive. However, direct experimental evidence for this hypothesis is lacking. *CenH3 (*called *Cid* in *Drosophila)* has been shown to evolve under positive selection in *D. melanogaster* and *D. simulans*, with 18 amino acid differences in predicted DNA-binding surfaces fixed between species. To uncover the selective pressure driving the evolution of centromere proteins*,* we reversed the most recent evolution of the *Cid* gene in the *D. melanogaster* genome. Using parsimony and gene synthesis, we "resurrected" ancestral *Cid* (*Cid**)*,* which differs by 11 amino acids from $Cid^{melanogaster}$, and introduced either the ancestral allele or the native $Cid^{melanogaster}$ into the same genetic location of the *D. melanogaster* genome. Transgenic $Cid^{melanogaster}$ completely complemented a genetic knockout of *Cid*. In contrast, we found that *Cid** function was compromised. In particular, we noticed a striking female-biased sex-ratio in the degree of complementation. The dearth of males among *Cid**-rescued flies could be due to drive against the Y chromosome, or due to Y-specific mitotic defects. Thus, reversing 2.5 million years of positive selection of *Cid* unleashed deleterious effects on males, suggesting that *Cid* adapted to mitigate the effects of centromere drive. Our results indicate that intragenomic conflict has recurrently altered the centromeric landscape and the function of the centromeric histone in *Drosophila.* Our approach provides experimental support for the centromere drive model, and a powerful means for the functional dissection of the consequences of positive selection *in vivo*.

O-233

Ocean Drive: A major factor in opsin diversification in vertebrates was the availability of new ocean ecosystems

Roberto Feuda, Sinéad Hamilton, Davide Pisani, James McInerney
NUI Maynooth, Maynooth, Ireland

Animals have large numbers of gene families that arose as a result of massive amounts of gene duplication and diversification. These genes can differ greatly in function and genomic location while still sharing the same evolutionary history. It is likely that gene duplication and diversification are linked to certain environmental factors such as global temperature changes or varying levels of available oxygen in the ocean or atmosphere. We wanted to see if it was possible to detect genomic changes, such as increased duplication rates, occurring around the same time as known global environmental events, such as the oxygenation of the ocean, to give some explaination as to how and why these changes occurred. To do this we used a very important, large protein family, the opsins. These proteins are used for light detection and vision, allowing organisms to interact with their environment to a greater extent than before.Opsins are light sensitive GPCRs (G-protein coupled receptors) that can react to light, propogating a signal that is sent to the brain for image processing or entrainment of circadian rhythms. Using a baysian molecular dating approach and a series of calibrations taken from the fossil record, we analysed the time at which each of the opsin duplications occurred to see if they coincide with any of the major global events at that time. We found that the vertebrate C-opsins and the arthropod R-opsins, used for image forming, all duplicated around the same time that major environmental changes were occurring in the oceans. During the transition between the Proterozoic and the Phanerozoic eras, ocean geochemistry was changing. The waters were becoming more oxygenated and metazoan life was beginning to flourish. This would allow animals to move into deeper oceans as they became more oxygenated. This trend can be seen in the duplication pattern of the visual opsins in both the arthropods and the vertebrates simultaneously. In both the vertebrate visual C-opsins and the arthropod visual R-opsins, the clades that emerge as a result of the duplications each have a specific maximum wavelength absorbancy that correlates with the penetration depth of each wavelength of light (colour) in coastal water. Therefore, as animals moved into deeper waters after the oxygenation event, they became capable of maximally detecting each wavelength of light that was most abundant at that depth.

O-234

Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution

Anna R. Kersting[1], Andrew D Moore[1], Sonja Grath[1], Alexander A. Myburg[2], Erich Bornberg-Bauer[1]
[1]Institute of eVolution and Biodiversity, University of Münster, Münster, Germany, [2]Forest Molecular Genetics Programme, Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

Plant genomes are typically very large; several have undergone paleopolyploid events and thus have numerous gene duplicates. These duplication events have been hypothesized to allow plants, which are sessile organisms, to rapidly adapt to environmental challenges. In other kingdoms such as bacteria, animals and fungi another mechanism for fast adaptation has recently been found to be predominant: the modular rearrangement of protein-coding genes by domain shuffling. To our knowledge this study is the first plant specific comparable investigation of domain arrangement on whole genome level. By reconstructing the ancestral domain content within plant phylogeny, we analyze the dynamics of domain emergence, loss, expansion, and rearrangements and explore their adaptive benefits across 27 plant and 3 algal genomes.
By using a phylogenomic approach, we show that 88% of domain arrangements were created by single-step events such as fusion, fission and terminal loss of domains. We observe that domain loss is a frequent event along every lineage. Moreover, we could detect more than 500 domains newly arising within the Viridiplantae at different times, which are unique to green plants. These novel domains are spreading rapidly, and are over-proportionally involved in stress response and developmental innovations. While a relatively large and well-conserved core set of single domain proteins exists, we could also observe more than 7000 species-specific arrangements.. Duplicated genes are frequently involved in rearrangements. Fission events are more likely to impact metabolic proteins, while fusion events often create new signaling proteins essential for environmental sensing.
We used the newly sequenced *Eucalyptus grandis* genome to test our predictions for a plant which has a high agro-economic relevance to the biofuel and pulp and paper industries. A prominent result we observed was an expansion of several domains related to terpene synthesis compared to the other rosids.
In summary, the high variability of domains and domain arrangements in plant genomes offer a high flexibility to adapt to biotic and abiotic factors.

These studies have been carried out in collaboration with the following colleagues: A. D. Moore, S. Grath, E. Bornberg-Bauer and A.A. Myburg
Kersting et al (2012) : Dynamics and Adaptive Benefits of Protein Domain Emergence and Arrangements during Plant Genome Evolution. GBE 10.1093/gbe/evs004

O-235

The Fates of Duplicate Genes at the Subcellular Level

Lydia Bright, Casey McGrath, Tom Doak, Michael Lynch
Indiana University, Bloomington, IN, USA

The *Paramecium aurelia* complex of closely related species has undergone at least three whole genome duplications (WGDs) in its history. Tracing the fates of paralogous products of these duplications at the subcellular level, between closely related species, promises to provide insight into both stochastic and adaptive processes that drive the evolution of duplicate genes on a fine scale. Members of the *aurelia* complex are indistinguishable from one another morphologically and are genetically tractable. Our group is currently sequencing and assembling the genomes of all members of the *P. aurelia* complex as well as three outgroup *Paramecium* species.

Individual steps in membrane trafficking are driven by protein determinants, acting in concert with one another and effectors. These determinants are often found in large gene families, whose size is related to the trafficking complexity of a given organism. The duplication and divergence of genes within these families has paralleled and driven the elaboration of membrane trafficking pathways.

We are pursuing a combined evolutionary and comparative cell biological investigation of two types of trafficking determinants. We have traced the paralog preservation patterns of members of the adaptor protein and Rab GTPase families across three closely related species of *Paramecium,* and found that many paralogs are retained differentially among the different species.

The *Paramecium* genome contains an extremely large Rab GTPase family (~220 genes), with a high rate of paralog retention among its members. We first focused on the large Rab11 subfamily, a clade of proteins involved in endocytic recycling and receptor turnover in animal cells. Duplicates in this subfamily show differential patterns of paralog retention. In particular, the Rab11h clade of ortho/paralogs has a pattern that hints at innovation in one of its members. Three conserved orthologs—one in each *aurelia* species examined,—are retained, and have highly conserved coding sequences (~87% amino acid identity). On the other hand, one paralog in *P. biaurelia* has diverged significantly (~63% identity), and has changes in key amino acid residues putatively involved in effector binding, and membrane attachment.

By creating fluorescently tagged versions of these proteins, we will be able to trace how their localization patterns differ between species, and between paralogs retained in the same species. In the future, we will further explore their functional roles through RNAi knockdown, knockouts, and rescue of knockouts with site-directed mutants.

O-236

Filtering multiple sequence alignments to avoid artifacts in downstream analyses

Eyal Privman[1,2], Osnat Penn[3], Haim Ashkenazy[3], Tal Pupko[3]
[1]University of Lausanne, Lausanne, Switzerland,[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland,[3]Tel Aviv University, Tel Aviv, Israel

Recent studies have emphasized the sensitivity of phylogenetic and other evolutionary analyses to errors in the multiple sequence alignment. We developed the GUIDANCE algorithm for scoring the reliability of each position in the alignment by measuring robustness to perturbation of the guide tree used by all progressive alignment algorithms. As such, GUIDANCE is widely applicable. We present benchmark and simulation studies to show that GUIDANCE outperforms other methods in correctly identifying the large majority of alignment errors. Our analysis demonstrates that filtering or masking of low scoring positions improves the accuracy of downstream evolutionary analyses, notably so in positive selection inference. In this study, we used evolutionary simulations in a novel, more realistic simulation scheme that is necessary to correctly capture the challenge of alignment and the impact of alignment errors on the accuracy of inference. Conversely, we find less sensitivity to alignment errors in phylogeny reconstruction. These studies may be summarized as a list of recommendations for the different alignment filtering or masking strategies that are most appropriate for different types of evolutionary analysis.

O-237

Robust handling of alignment uncertainty when inferring positive selection

Benjamin Redelings
NEScent, Durham, NC, USA

Estimates of positive selection are sensitive to alignment errors because incorrectly aligned residues may imply the presence of imaginary non-synonymous substitutions. We propose to incorporate alignment inference into the inference of positive selection using a Bayesian co-estimation framework. This framework can account for alignment uncertainty in the inference of positive selection without the necessity for a two-stage procedure in which ambiguous sites are first removed from the alignment. Thus, in theory, the model of positive selection can influence the inferred alignment, as well as vice versa. We implement the Branch-Site model for detection of positive selection within the software package BAli-Phy, and assess changes in sensitivity and specificity for identifying positively selected sites in simulated sequence data.

O-238

Uncertainty on multiple sequence alignment: an approach using the bootstrap

Jia-Ming Chang, Salvador Capella, Jean-Francois Taly, Toni Gabaldon, Cedric Notredame
Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain

Several highly publicised works have recently reported on the influence of multiple sequence alignment on downstream analysis, and on the uncertainty inherent to their estimation. The lack of robustness of MSA methods is illustrated in the recent Heads-or-Tails methodology (HoT), where it has been shown that substantial variation can occur when estimating an MSA on a set of sequences and subsequently on the same sequences reversed from left to right [1]. We show here that the HoT observation merely results from arbitrary tie breaking in the dynamic programming algorithm (low/high roads). This make the HoT protocol a one time sampling within a space defined by the entire set of ambiguous trace-back cells in the pairwise dynamic programming. A more generic sampling can be done that would involve breaking all ties arbitrarily, as done in the GUIDANCE bootstrap strategy [2]. This approach, however, is invasive and requires a substantial code modification in most MSA packages. We show here that shuffling the order in which the sequences are provided to the aligner can approximate a similar result. In most dynamic programming implementation, a swap of the order in which two sequences are aligned effectively amounts to inverting the tie break priority of each tie (low roads become high roads and inversely).

We applied this strategy on two popular aligners: ClustalW2.1 and MAFFTv6.815b, using the BAliBASE3.0 reference dataset. Our approach involves shuffling the sequences and realigning them 100 times. The resulting MSAs are then combined using M-Coffee, a flavor of the T-Coffee package able to combine several alternative models into a consensus model. The M-Coffee MSA model comes along with an estimation of the consistency between each position and the combined alignments, a bit like a consensus tree when combining the replicates. Our results indicate a variation among alignments as a consequence of the order shuffling. For instance, MAFFT replicates are on average 92.1% consistent, while ClustalW alignments are 90.2% consistent. Interestingly, our analysis also suggests a correlation between alignment accuracy and overall consistency (MAFFTr=0.52,ClustalWr=0.49). This result also holds locally. For instance pairs of residues having a score of 5 or higher are 96.9% likely to be correctly aligned. Likewise, pairs of residues having a score of 4 or lower are 27.6% likely to be accurately aligned.

1. Landan G, Graur D., Molecular Biology and Evolution 2007.
2. Penn O, Privman E, Landan G, Graur D, Pupko T., Molecular Biology and Evolution 2010.

O-239

Simultaneous protein multiple sequence alignment and tree construction using hidden Markov models.

Kimmen Sjolander
University of California Berkeley, Berkeley, CA, USA

Protein multiple sequence alignments and phylogenetic trees are used in a plethora of bioinformatics analyses. The standard protocol for constructing a phylogenetic tree involves a two-step process: first, construct (and mask) an MSA and second, estimate a phylogeny based on the MSA. The intent of masking is to remove columns in the MSA reflecting significant variation across the family, under the assumption that they contribute more noise than signal. However, in a protein superfamily (containing groups of paralogous genes, with potentially high levels of sequence divergence), positions that are variable across the family may be of critical importance, contributing to the functional specificity of individual clades or subfamilies: sequences that agree at these positions ought to be clustered together in the tree.
SATCHMO (Simultaneous Alignment and Tree Construction using hidden Markov models) (1) uses profile-profile scoring and alignment to progressively and simultaneously align a set of sequences and derive a phylogenetic tree. The fundamental innovation in SATCHMO is the use of subtree-specific masking: at subtrees near the leaves (with closely related sequences that align along their entire lengths), very few or no columns may be masked. As sequences spanning greater degrees of structural divergence are clustered into subtrees, additional positions are masked. Towards the root, masking may be quite extreme, particularly in cases of extreme structural divergence across the family. This subtree-specific masking protocol maintains maximum information in the HMM constructed for each subtree, retaining regions that define each subgroup, even if they are variable across the family as a whole. SATCHMO was recognized by the Faculty of 1000, who selected it as a "Must Read" for Technological Advance (rating 6.0).
SATCHMO-JS uses a divide-and-conquer protocol to reduce the computational complexity of all-vs-all profile-profile scoring and to improve alignment accuracy. Our results on a benchmark dataset of structurally aligned proteins show a statistically significant improvement ($P < 0.05$) in alignment accuracy relative to MAFFT, MUSCLE and ClustalW. The SATCHMO-JS paper (2) was selected by the editors of Nucleic Acids Research as a Featured Article.

1. Edgar, R.C. and Sjolander, K. (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics (Oxford, England), 19, 1404-1411, PMID: 12874053.
2. Hagopian, R., Davidson, J.R., Datta, R.S., Samad, B., Jarvis, G.R. and Sjolander, K. (2010) SATCHMO-JS: a webserver for simultaneous protein multiple sequence alignment and phylogenetic tree construction. Nucleic acids research, 38, W29-34.

O-240

Clustal Omega for fast and accurate alignment of many protein multiple sequences

Des Higgins
University College Dublin, Dublin, Ireland

Most Multiple Sequence Alignments (MSA) are made using a range of related heuristics that involve clustering the sequences and building an alignment that follows the clusters. These methods have served us well for the past 20 years but are now starting to creak. I will describe and demonstrate a new program called Clustal Omega which can make alignments of any number of sequences. It gives good quality alignments in reasonable times and has extensive features for adding new sequences to or for exploiting information in existing alignments. Currently, the program is for proteins only and is mainly designed to be run from the command line on Unix operating systems.

Clustal Omega uses the mBED algorithm (Blackshields, et al, 2010) for clustering sequences initially. This method can cluster very large numbers of sequences in O(NlogN) time and memory and is the main reason why Clustal Omega can make large alignments. Alignments of 100,000 or more proteins are routine and can be made on a single core of a desktop computer in 3-8 hours. The program uses multithreading to reduce these times if more processor cores are available. Crucially, mBED introduces almost no loss in accuracy compared to running default clustering from the older Clustal W program.

Once the sequences are clustered, they are aligned in larger and larger groups of sequences using the HHalign algorithm from Söding (2005). This is a high accuracy HMM alignment package which is normally used for searching collections of HMMs. With Clustal Omega, the sequences are converted into HMMs and aligned using HHalign. This allows alignments of high accuracy to be generated. Using standard benchmark methods, the program is now one of the most accurate protein alignment packages available.

Finally, Clustal Omega has extensive facilities for aligning new sequences using existing alignments to help guide the alignment. This allows sequences to be added to an older alignment or allows users to maximise the accuracy of a new alignment, given access to an existing high-quality alignment.

Blackshields, G., Sievers, F., Shi, W., Wilm, A., and Higgins, D. (2010)
Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms for Molecular Biology *5*, 21.

Söding, J. (2005) Protein homology detection by HMM–HMM comparison.
Bioinformatics *21*, 951 -960.

O-241

Efficient representation and propagation of posterior alignment uncertainty

Joseph Herman, Ádám Novák, Rune Lyngsø, Jotun Hein
University of Oxford, Oxford, UK

A number of methods have recently emerged for probabilistic sampling of multiple sequence alignments, enabling an ensemble of alignments to be generated according to their probability. Although such methods provide much more information than is contained in a single alignment, there has been no clear consensus as to how to interpret and make use of this information. A typical approach has been to heuristically generate a single alignment as a representative of the ensemble, but it is often difficult to place a probabilistic interpretation on such a summary. Moreover, due to the size of the alignment space, each observed alignment is typically sampled with a very low frequency, hence summary statistics such as the estimated maximum a posteriori alignment are also poor summaries of the distribution.

We recently developed an equivalence-class framework for coding the space of sampled alignments in order to permit a directed acyclic graph representation of the posterior in the space of alignment columns. This representation enables the use of dynamic programming algorithms for finding single alignments that minimise the posterior risk under a wide range of loss functions. For example, this allows us to generate alignments that minimise the posterior probability of false positive or false negative homology statements, given prior information regarding expected homology.

The DAG representation also allows for the formulation of efficient algorithms for other types of downstream analysis, enabling the posterior uncertainty associated with the alignment step to be incorporated into the downstream inference.

As a specific example, we discuss the annotation of multiple sequence alignments via hidden Markov models. This usually involves conditioning on a single alignment, but in many cases the choice of alignment can heavily influence the downstream analysis. As such, one would ideally perform joint inference on the alignment, phylogenetic tree, and HMM parameters. However, this requires formulation of complex hierarchical models, which may be highly computationally intensive to run.

The use of the posterior alignment DAG represents an intermediate solution between these two extremes, whereby alignments are sampled according to a probabilistic evolutionary model, and all the samples can then be used for downstream analysis. Any downstream algorithm that can be formulated in terms of a linear recursion on an alignment (e.g. the forward-backward algorithm for HMMs) can be easily adapted to run on the alignment DAG, such that averaging over alignments according to their probability can be carried out without incurring significant additional computational cost.

O-242

Evolutionary and functional investigations of genes subject to regulation by genomic imprinting in plant seeds

Peter McKeown[1], Reetu Tuteja[1], Martin Braud[1], Antoine Fort[1], Mary O'Connell[2], Mark Donoghue[3], Charles Spillane[1]
[1]NUI Galway, Galway, Ireland,[2]Dublin City University, Dublin, Ireland,[3]Cold Spring Harbor Laboratory, New York, USA

The laws of Mendelian inheritance state that alleles of genes are equally expressed regardless of whether they are inherited from the maternal or paternal parent. However, this rule is violated for certain genes in flowering plants and eutherian mammals which are marked by differential 'imprints' established in the gametes which lead to parent-of-origin specific uniparental expression effects in the F1 offspring. Imprinted plant genes such as *MEDEA* are fast-evolving under Positive Darwinian Selection. Mammalian imprinted genes display variable patterns of evolution and selective pressures across different lineages. Recently, additional candidate imprinted genes have been identified in the model plant, *Arabidopsis thaliana* L, and in grain monocots (rice and maize). However, only low levels of conservation of imprinting status has been detected between monocot and dicot species. To gain understanding of the evolutionary forces associated with imprinting, we assess current progress in identifying selective pressures acting upon plant imprinted genes and the stability of imprinting over evolutionary timescales.

O-243

'Archaeoepigenetics: Evidence of epigenetic response to drought stress and viral infection in archaeological plant material'

Oliver Smith, Sarah Palmer, Robin Allaby
The University of Warwick, Coventry, UK

MicroRNA (miRNA)-based responses to environmental stresses have been well documented in plants. Here we show that second-generation sequencing of an archaeological (~1,000 BP), extinct variety of H. vulgare (barley) from Qasr Ibrim, Egypt, exhibits some significant changes in expression profiles of miRNA and short interfering RNA (siRNA) when compared to a modern, unstressed plant from the same region. In particular, a highly conserved TCP domain-targeting microRNA, miR319, is highly abundant in the archaeological sample. MiR319 is thought to negatively regulate the mRNA transcript of intermedium-c, a recently characterised gene in barley which is closely linked to a previously described genotype / phenotype discrepancy in these samples.

The archaeology and paleoclimate of the site suggests significant dehydration stress was a recurring issue in antiquity, and consequently the cultivated barley grown there underwent a miRNA- or epigenetically-mediated response to such conditions. This response perhaps resulted in the unusual genotype for the observed phenotype, forming the original hypothesis for this project. The phenomenon, which recurred over an extended period of occupation (2,600 years), has not been observed in any other variety since and implies localized heritability of this trait.

Investigation of allelic variants of the intermedium-c locus is ongoing, alongside comparisons of methylation patterns throughout the sample and control genomes to identify epigenetic variation. Further results indicate the presence of a smaller, ssRNA viral genome (Barely stripe mosaic virus) in the archaeological sample that can be reconstructed with high (98%) coverage. Certain motifs of siRNA spikes along that genome are indicative of simultaneous RNAi activity in the archaeological material. These results also serve to validate the use of second-generation sequencing technologies with ancient RNA, a previously unexplored area of study.

O-244

Human postmeiotic sex chromosome inactivation and its impact on sex chromosome evolution

Ho-Su Sin[1,2], Eitetsu Koh[3], Satoshi Namekawa[1,2]
[1]Division of Reproductive Sciences, Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA, [2]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA, [3]Department of Integrative Cancer Therapy and Urology, Andrology Unit, Kanazawa University Graduate School of Medical Science, Kanazawa, Japan

Sex chromosome inactivation is essential epigenetic programming in male germ cells. However, it remains largely unclear how epigenetic silencing of sex chromosomes impacts the evolution of the mammalian genome. Here we demonstrate that male sex chromosome inactivation is highly conserved between humans and mice and has an impact on the genetic evolution of human sex chromosomes. We show that, in humans, sex chromosome inactivation established during meiosis is maintained into spermatids with the silent compartment postmeiotic sex chromatin (PMSC). Human PMSC is illuminated with epigenetic modifications such as trimethylated lysine 9 of histone H3, HP1b, and HP1g, which implicate a conserved mechanism underlying the maintenance of sex chromosome inactivation in mammals. Furthermore, our analyses suggest that male sex chromosome inactivation has impacted multiple aspects of the evolutionary history of mammalian sex chromosomes: amplification of copy number, retrotranspositions, acquisition of *de novo* genes, and acquisition of different expression profiles. Most strikingly, profiles of escape genes from postmeiotic silencing diverge significantly between humans and mice. Escape genes exhibit higher rates of amino acid changes compared to non-escape genes, suggesting they are beneficial for reproductive fitness and may allow mammals to cope with conserved postmeiotic silencing during the evolutionary past. Taken together, we propose that the epigenetic silencing mechanism impacts the genetic evolution of sex chromosomes and contributed to speciation and reproductive diversity in mammals.

O-245

Contribution of epigenetic modification to subfunctionalization after whole genome duplication

Lidija Berke, Gabino F. Sanchez-Perez, Berend Snel
Utrecht University, Utrecht, The Netherlands

Gene duplications are an important source for evolutionary innovation. Differences in gene expression patterns reflect the functional divergence of paralogs such as neofunctionalization and subfunctionalization. Functional divergence has been linked to sequence divergence and is influenced by a plethora of factors such as protein function. We investigate whether epigenetic modification could be a factor in evolution of expression patterns, and focus on trimethylation of histone H3 at lysine 27 (H3K27me3). Its function as a repressive mark for expression is conserved from animals to plants, however in plants the mark is almost exclusively localized at specific genes whereas in animals it covers continuous regions of DNA containing larger numbers of genes. This gene-specific localization together with well-annotated recent whole-genome duplication makes Arabidopsis thaliana uniquely suitable for studying the epigenetic fate of individual genes.
We merged results from several recent whole-genome analyses reporting genes which are regulated by H3K27me3, and added this information to paralogous pairs from the latest A. thaliana whole-genome duplication, a set of paralogs of the same age. Our results show that paralogous pairs with H3K27me3 exhibit the highest similarity in expression patterns. The most divergent patterns belong to paralogous pairs where only one member of the pair has H3K27me3. For transcription factors, the differences in expression correlation between pairs with and without H3K27me3 are even more striking. The effect of the epigenetic modification on differences in expression divergence did not depend on coding sequence divergence or differences in expression levels. However, conservation of upstream regulatory regions is higher in paralogous pairs that share H3K237me3. We thus show that, in addition to other factors, H3K27me3 is strongly linked to expression divergence between paralogs.

O-246

Genome-wide analysis of *cis*-regulatory divergence between species in the *Arabidopsis* genus

Fei He[1], Juliette de Meaux[1], Justin Borevitz[2]
[1]*University of Münster, Münster, Germany,* [2]*University of Chicago, Chicago, USA*

Cis-regulatory DNA has been suspected to play a pre-eminent role in adaptive evolution. Understanding the role of *cis*-regulatory mutations in gene expression divergence requires in the first place an accurate analysis of the functional differences associated with these regions. We analyzed allele-specific expression (ASE) in leaf and floral tissues of F1 interspecific hybrids generated between the two closely related species *Arabidopsis thaliana* and *A. lyrata* with a whole-genome SNP (Single Nucleotide Polymorphism) tiling array. We observed 2205 genes showing allele-specific expression (ASE) pattern in at least one tissue. Nearly 90% of genes displaying ASE preferentially expressed the allele of *A. lyrata*. Genome-wide comparison of sequence divergence revealed that genes displaying ASE had a higher ratio of non-synonymous to synonymous substitutions in coding regions and local modifications in CG di-nucleotides contents in upstream regions. We further observe that the epigenetic landscape of histone methylation in *A. thaliana* genome associate with ASE. The asymmetric pattern of allele-specific expression suggests that epigenetic landscapes could differ greatly between species.

O-247

Monoallelic gene expression contributes significantly to expression divergence and diversity in primates

Lisa Stubbs[2], Katja Nowick[1]
[1]*University Leipzig, Bioinformatics Group, IZBI, Paul-Flechsig-Institute for Brain Research, Leipzig, Germany,*
[2]*University of Illinois, Institute for Genomic Biology, Urbana, IL, USA*

How do genotypic differences translate into phenotypic differences within and between species? Are there classes of genes that contribute more to transcriptomic and phenotypic variation than others? In diploid organisms, a surprisingly large number of genes are expressed from only one chromosome (monoallelically expressed genes (MEG)). For humans, the number of genes with random monoallelic expression has been estimated to be at least 1000 based on data from B-lymphoblastoid cells. Interestingly, MEGs are enriched near conserved non-coding sequences with human-specific accelerated evolution, raising the question if such genes are differentially expressed between humans and other primates. To address this question, we generated genome-wide expression profiles from multiple human and chimpanzee B-lymphoblastoid cell lines.

We found that MEGs have significantly higher expression divergence than biallelically expressed genes between humans and chimpanzees. Moreover, MEGs are also significantly enriched among the genes that show signs of adaptive expression evolution (high divergence/diversity ratios). Strikingly, MEGs are also significantly enriched among the genes with higher expression diversity among humans, wherein the set of MEGs displaying high expression diversity does not overlap with the set of MEGs showing signs of adaptive evolution. We confirmed the finding of higher expression divergence and diversity of MEGs in multiple tissues, including brain.

Theoretically, mutations in MEGs have a larger phenotypic impact, because the other allele is not available as a backup. We therefore asked if MEGs often contain non-synonymous sequence differences between the two alleles of an individual. Taking advantage of data from the human 1000 Genomes Project, we found that MEGs are significantly overrepresented among the genes that encode for two different proteins in the majority of the individuals. Randomly expressing only one of the alleles can thus significantly increase transcriptome diversity between cells.

Among the MEGs with high sequence and expression variation we revealed significant enrichment for genes involved in neurological systems processes. For instance, MEGs important for the auditory system often encoded two different proteins in human individuals and display high expression diversity and divergence. MEGs implicated in Alzheimer's disease were often variable in expression among humans or had high divergence/diversity ratios between humans and chimpanzees.

Taken together, MEGs are good candidates to be responsible for phenotypic differences within and between species. The mode of expressing genes randomly from only one or the other allele adds another layer of complexity to the challenge of understanding how genotypic differences translate into phenotypic differences.

P-1000

Bioinformatics Resources for Enabling Discoveries from Genomes to Phenotypes (*MEGA*, *TimeTree*, and *FlyExpress*)

Sudhir Kumar
Center for Evolutionary Medicine & Informatics, Arizona State University, Tempe, Arizona, USA

Primary databases containing DNA sequences from diverse species, patterns of expression for thousands of genes, and scientific literature capturing primary inferences are growing at an unprecedented rate. Consequently, there is an acute need for tools that facilitate evolutionary data analysis and enable the mining of image and divergence time data. We have responded to this need by continuously developing **MEGA** (software for comparative analysis of molecular sequences; www.megasoftware.net), establishing a literature-mining knowledge-base for information on the species timetrees (**TimeTree**, www.timetree.org), and a web-tool for finding genes with overlapping patterns by image-matching (**FlyExpress**; www.flyexpress.net). I will highlight salient features of these resources, including the biologist-centric designs of our MEGA software and advances in TimeTree and FlyExpress resources that break domain nomenclature barriers and go beyond classical data-mining to providing answers to specific biological questions.

P-1001

Molecular biology and evolution of a candidate immune receptor in Drosophila

Erin Keebaugh, Todd Schlenke
Emory University, Atlanta, Georgia, USA

To gain a better understanding of non-self recognition between eukaryotes we study the interaction between fruitflies and one of their natural metazoan pathogens, parasitic wasps. Parasitic wasps lay eggs in Drosophila larvae that hatch, consume larval tissues, and eclose from the fly pupal case, killing the fly in the process. Drosophila larvae can mount a robust cellular immune response against the wasp eggs termed melanotic encapsulation, where fly hemocytes (blood cells) form a capsule around and kill the entrapped wasp egg. As a first step in the encapsulation response, the host must be able to recognize the parasite as foreign. We are interested in identifying the immune receptors Drosophila use to identify parasitic wasps as non-self, and uncovering the evolutionary history of such genes as a first step towards understanding the selective forces parasites impose on their hosts in nature. Microarray analysis of Drosophila larvae post-wasp attack identified several promising candidate immune receptors including a C-type lectin, lectin-24A. We found that lectin-24A is a new gene present in only four Drosophila species and shows signs of recent and recurrent selective pressures in D. melanogaster and D. simulans. Expression analysis of this candidate immune receptor shows enriched expression in immune tissue and a wasp-specific regulatory response. We are currently investigating the effects of mutant levels of lectin-24A on hemocyte viability, structure, and ability to form melanotic capsules, and have designed experiments to test Lectin-24A's binding specificity.

P-1002

Recreating the Past: Reconstruction of a One-Billion-Year-Old-Enzyme

Joanne Hobbs[1], Charis Shepherd[2], David Saul[3], Nicholas Demetras[1], Svend Haaning[1], Colin Monk[1], Roy Daniel[1], Vickery Arcus[1]
[1]University of Waikato, Hamilton, New Zealand, [2]Otago Medical School, Dunedin, New Zealand, [3]ZyGEM Corporation Ltd., Hamilton, New Zealand

Ancestral sequence reconstruction (ASR) is a molecular technique that allows proteins from extinct organisms to be reconstructed, and subsequently characterised, in the laboratory. It uses 'genetic souvenirs' present in the sequences of extant proteins, and the phylogenetic relationships between them, to trace their evolutionary history and infer the sequences of their ancestors. ASR has revealed otherwise unobtainable details about the evolution of several binding proteins but few ancestral enzymes have been reconstructed. The reconstruction of ancestral enzymes has the added advantage of catalytic activity, which can act as an internal control for accurate inference.

Here we report the reconstruction of the structurally-complex core metabolic enzyme LeuB from the last common ancestor (LCA) of extant *Bacillus* species using both maximum likelihood (ML) and Bayesian inference. ML LeuB from the LCA of *Bacillus* is approximately one billion years ago and shares only 76% amino acid sequence identity with its closest extant homologue, yet it is fully functional, thermophilic and exhibits a high turnover rate (k_{cat}), catalytic efficiency (K_M/k_{cat}) and free energy for unfolding. The Bayesian version of this enzyme is also thermophilic but exhibits anomalous catalytic kinetics. In addition, we have determined the three-dimensional structure of the ML enzyme and found that it is more closely aligned with LeuB from deeply-branching bacteria, such as *Thermotoga maritima,* than extant *Bacillus* species, even though sequences from these ancient bacteria were not used in the reconstruction process.

LeuB from the LCA of *Bacillus* is thermostable and has a high optimum temperature for activity, which suggests that its host was thermophilic. To investigate the evolution of thermophily within the *Bacillus* genus, we also reconstructed three ancestral descendents of ML LeuB. These enzymes reveal a fluctuating trend in thermal evolution, with a temporal adaptation towards mesophily followed by a more recent return to thermophily. This contrasts with previous studies which have reported a gradual loss of thermophily over evolutionary time. Structural analysis of our reconstructed enzymes suggests that the determinants of thermophily in LeuB from the LCA of *Bacillus* and the more recent thermophilic enzyme are distinct, and that thermophily has arisen at least twice within the *Bacillus* genus via independent evolutionary paths.

P-1003

Investment in fast growth modulates the evolutionary rates of essential proteins

Sara Vieira-Silva[1,2], Marie Touchon[1], Sophie S. Abby[1], Eduardo P.C. Rocha[1]
[1]Institut Pasteur, Paris, France, [2]VIB/VUB, Brussels, Belgium

Different proteins evolve at very different rates and these rates can also change through time and across lineages. Most notably, protein evolutionary rates are inversely proportional to protein expression level. The relative frequency of highly expressed proteins in the proteome, and thus their cost on fitness, increases steeply with cellular growth rate. However, the potential to reach high growth rates varies among microbial lineages. Indeed, the maximal growth rate is a key life-history trait determined by trade-offs between rapid growth and other fitness components. Because of the negative correlation between protein expression levels and evolutionary rate and the positive correlation between expression levels of highly expressed proteins and growth rates, we hypothesize that investment in fast growth affects the evolutionary rate of proteins, especially highly expressed ones. We analyzed 61 families of orthologs in 74 different proteobacterial species and found that indeed differences in evolutionary rates between lowly and highly expressed proteins depend on maximal growth rates. These results were confirmed when analyzing complexes with key roles in bacterial growth and strikingly different expression levels: the ribosome and the replisome. These patterns suggest that growth-related sequence conservation is associated with protein synthesis. We extended this observation to other bacterial clades. We propose that long-branch attractions associated with growth-related evolutionary rate variation may explain the observation that clades with persistent history of slow growth are attracted to the root when the tree of prokaryotes is inferred using highly, but not lowly, expressed proteins. These results indicate that reconstruction of deep phylogenies can be affected by maximal growth rates, and highlight the importance of life-history traits and their physiological consequences for protein evolution.

P-1004

Evolution of physical interactions between the transcription factor protein HoxA11 and the long non-coding RNA molecule Steroid Receptor Activator 1

Kathryn Brayer, Gunter Wagner
Yale University, New Haven, CT, USA

Long non-coding RNAs (lncRNAs) have gained recent attention for their role in gene and genome regulation; and, there is growing recognition that lncRNAs play a critical role in mediating transcriptional changes. Despite this, very little is known about the functional evolution of these molecules as it pertains to the evolution of novelty. Steroid Receptor Activator 1 (SRA1) is a lncRNA known to interact with, and activate, various nuclear hormone receptors including progesterone receptor and estrogen receptor. Recently, using RNA-IP we have determined that SRA1 also interacts with human HoxA11 in human endometrial stromal cells and that this interaction is cell-state dependent whereby human HoxA11 captured significantly more SRA1 in differentiated cells versus undifferentiated cells. Although the mechanisms that regulate endometrial cell differentiation are poorly understood, differentiation is a critical step in the successful establishment of pregnancy, and previous work in our lab has demonstrated that the functional evolution of HoxA11 coincided with the evolution of pregnancy. Here, we examine the evolution of the HoxA11-SRA1 interaction. Preliminary data suggests that the interaction is derived in placental mammals since chicken HoxA11 failed to capture human SRA1; and opossum HoxA11, although able to capture small amounts of human SRA1, it failed to do so in a cell-state dependent manner. Experiments to determine whether or not the derived human HoxA11-SRA1 interaction is an example of coevolution are underway, and will also be discussed.

P-1005

Phylogeography: a new statistical approach for measuring dispersal ranges and competitive exclusion.

Louis Ranjard, Stephane Guindon, David Welch
University of Auckland, Auckland, New Zealand

Extensive sampling of genetic material is now common place in ecological studies. It is then straightforward to infer the evolutionary relationships between individuals, populations or species. Moreover, by combining phylogenetic and geographic information, one can decipher the history of colonisation of a given region and answer important questions about the many factors influencing this complex process.

We propose a new model of colonisation in which each node of a phylogeny corresponds to a past dispersal event. In that model, the rate of migration from one geographic location to another depends on the physical distances between locations. Moreover, the probability for a given individual to settle in a new location depends on the density of individuals currently occupying this spot. Our approach therefore provides a sound statistical tool to quantify the effect of competitive exclusion.

We simulated migration histories in a two-dimensional finite space and use a Bayesian approach to estimate the model parameters. Simulation results as well as potential model extensions of the model will be presented.

P-1006

Is there really a universal deletion bias?

Steven Laurie[1], Macarena Toll-Riera[1], Núria Radó-Trilla[1], Mar Albà[1,2]
[1]Fundació Institut Municipal d'Investigació Mèdica (FIMIM)-Universitat Pompeu Fabra (UPF), Barcelona, Catalunya, Spain, [2]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Catalunya, Spain

It has been proposed that there is a universal bias towards deletions rather than insertions when considering short indels at the genomic level, though some studies of coding sequences have not observed such a bias. We have investigated this bias through analysis of ~20,000 orthologous non-coding ancestral-repeat regions, and ~6,000 orthologous proteins in human, macaque, mouse, and rat, using cow as our outgroup. We have used a relatively novel multiple-alignment algorithm, Prank+F, for the first time in this type of large-scale comparative analysis. Considering ancestral repeats, only in the rodent ancestral branch do we observe a deletion bias that results in a significant loss of sequence (approximately 2.5% of mammalian syntenic sequence), while in the primate ancestral branch we observe the inverse relationship (i.e. insertions outnumber deletions), though it results in much less sequence loss. In the orthologous protein sequences we do observe a significant deletional bias in every branch apart from human and mouse, but this does not lead to a significant reduction in protein length in any branch. Curiously, when we compare the two datasets, we observe that negative selection appears to act more strongly against insertions than deletions in coding sequence, suggesting that the former are more deleterious. The strength of negative selection is found to be higher in the rodent lineages than in the primate lineages, which is consistent with the role of effective population size upon selection. We propose to investigate further the role of short-indels in the evolution of novel genes following duplication, to determine their relevance in this scenario.

P-1007

Assessing the impacts of mating system and natural selection on genome-wide patterns of nucleotide diversity in wild tomatoes

Margot Paris, Ana Marcela Florez-Rueda, Thomas Städler
ETH Zurich, Institute of Integrative Biology, Plant Ecological Genetics, Zürich, Switzerland

Mating system variation has several important effects on the genetic properties of populations by affecting levels of nucleotide polymorphism and linkage disequilibrium among loci. Wild tomatoes (*Solanum* section *Lycopersicon*, Solanaceae) comprise a small monophyletic clade within the large genus *Solanum* which is characterized by an ancestral self-incompatibility (SI) system, i.e. the ability of a plant to recognize and reject its own pollen, enforcing obligate outcrossing. The evolutionary transition to self-compatibility (SC) has been common in wild tomato species, leading to the presence of SI, SC, and realized mixed mating (i.e. variable rates of self-fertilization) in this clade. Due to their recent divergence, the differences in their mating system, and the abundant genomic resources available, wild tomatoes appear ideally suited to study the impact of mating systems on genetic diversity and linkage disequilibrium at a genome-wide scale. Using individual tagging with the Illumina RNAseq procedure, we sequenced the transcriptomes of wild tomato accessions covering 8 species (up to 5 individuals per species), two of them with a highly selfing mating system. For each accession, millions of reads were mapped to the cultivated tomato (*S. lycopersicum*) reference genome, and SNPs were detected for thousands of genes. Using analyses at the sequence level we aim to characterize patterns of nucleotide polymorphism and linkage disequilibrium. The comparison between SI and SC species should allow testing theoretical expectations on the effect of mating system changes on genome-wide patterns of nucleotide variation. Given the higher level of linkage disequilibrium in selfing species, natural selection is expected to affect larger chromosomal regions in the vicinity of the targets of selection (whether positive or purifying), thus reducing nucleotide variation in these linked regions. Taking into account mating system differences is therefore important for the study of speciation processes and the genome-wide distribution of divergence between species, as well as for the study of genomic signatures of natural selection.

P-1008

Characterizing the molecular basis of hybrid seed failure between two wild tomato species

Ana Marcela Florez-Rueda, Margot Paris, Thomas Städler
ETHZ, Zürich, Switzerland

The establishment of reproductive isolation between diverging lineages is a crucial component of the speciation process and thus of major interest in evolutionary biology. These barriers can be classified as either prezygotic or postzygotic, and it has been argued that the latter are more important because hybrid inviability or sterility are unlikely to be reversible, and therefore are more likely permanent barriers to gene flow between nascent species. Hybrid seed failure (i.e. embryo inviability) is a common interspecific barrier in angiosperms, and there is substantial evidence that imbalances in endosperm–embryo interactions early in seed development mediate such barriers. In *Arabidopsis* it has been shown that misregulation of normally imprinted genes (i.e. the parent-of-origin dependent expression of alleles) accompanies hybrid embryo death and thus hybrid seed failure. Wild tomatoes offer a good opportunity to characterize these mechanisms outside the model system *Arabidopsis.* Many decades ago, Rick and Lamm (1955)documented strong reproductive barriers separating the sister species *Solanum peruvianum* and *Solanum chilense*. The nature of their reproductive isolation is evidently postzygotic, as following controlled pollinations there is production of fruits that contain aborted seeds. By means of histological procedures and seed counts in fully developed fruits, we have begun to examine the time course and end fate of seed development. Within-species crosses were performed to serve as a baseline of the normal number of seeds per fruit, normal histology, and the normal transcriptomes of embryo and endosperm tissues. In hybrid seeds the embryos reach only the globular stage and at that point development appears to be arrested. There are clear histological differences in reciprocal crosses, consistent with a misregulation of the still-unknown normal imprinting pattern. In order to molecularly characterize such barriers and potentially identify causal genes, separation of embryo and endosperm tissue is performed with a laser dissection procedure following reciprocal crosses between *S. chilense* and *S. peruvianum*. After RNA extraction and whole-transcriptome sequencing using the Illumina platform, we aim to identify and quantify the expression of tissue-specific genes. Comparison of the gene identity and transcript levels between intraspecific crosses and those of interspecific crosses will provide the basis for the identification of candidate genes for postzygotic reproductive isolation.

P-1009

Genes important to C_4 evolution identified by comparative genomics

Janina Maß[1], Alisandra Denton[2], Martin J. Lercher[1]
[1]Institute for Computer Science, Heinrich-Heine-University Duesseldorf, Düsseldorf, Germany,[2]Institute for Plant Biochemistry, Heinrich-Heine-University Duesseldorf, Düsseldorf, Germany

C_4 photosynthesis is an add-on to the evolutionary older and wide-spread C_3 pathway. In contrast to C_3 plants, C_4 plants show increased photosynthetic efficiency by reducing the rate of photorespiration. Intriguingly, the C_4 pathway evolved more than 60 times independently among the angiosperms, which leads to the assumption that the transition from C_3 to C_4 photosynthesis might be relatively easy. The evolution of the C_4 pathway has left traces in the genomes and transcriptomes of affected species. Gene duplication, for instance, is commonly accepted as a preconditioning factor necessary in the C_3 stage before C_4 emergence.

Within the grasses, the PACMAD clade contains many lineages with C4 photosynthesis, whereas the BEP clade does not show a single occurrence. Concentrating on a subset of C_3 and C_4 species in these clades, we aimed to detect candidate genes involved in C_4 evolution. We first clustered orthologous genes from both clades. To identify relevant gene duplications or gene losses, we then analyzed clade-specific differences in presence and number of genes in orthologous groups. Furthermore, we conducted phylogenetic analyses of gene families to search for signs of distinct selection pressure in branches with C_4 species.

P-1010

The lifestyles of vertically and horizontally transmitted endosymbiotic bacteria affect the balance between selection and drift acting on their genomes.

Meg Woolfit, Elizabeth McGraw
Monash University, Melbourne, Victoria, Australia

Taxonomically diverse endosymbiotic bacteria infect a wide range of eukaryotes, and have a profound impact on host biology. They may be parasitic or beneficial to their hosts, and exhibit highly variable evolutionary and ecological characteristics. Some of these taxa have unusually rapid rates of evolution and small genome sizes when compared to non-endosymbiotic bacteria. These are the molecular signatures of an increase in genetic drift caused by small effective population sizes. However, other evolutionary processes may also affect the rate and pattern of evolution across the genomes of some or all endosymbionts. The interacting genomic effects of relaxation of purifying selection due to host protection, intensification of purifying selection driven by inter-host competition and positive selection due to conflict with the host have not yet been quantified for most endosymbionts. Here we use publicly available genome sequences of free-living, vertically-transmitted and horizontally-transmitted endosymbiotic bacteria to examine how these factors can affect the balance between selection and drift across bacterial genomes. After taking into account the effects of recombination, and testing the power of our analyses using simulations, we compare the distributions of per-gene dn/ds values across taxa, and link patterns of positive and purifying selection to aspects of bacterial lifestyles.

P-1011

HuntMi: a novel machine learning method for miRNA classification

Michal Szczesniak[1], Adam Gudys[2], Marek Sikora[2], Izabela Makalowska[1]
[1]Adam Mickiewicz University, Poznan, Poland, [2]Silesian University of Technology, Gliwice, Poland

miRNAs regulate the expression of thousands of genes in plants and animals and are key players in developmental, stress-related and signalling processes. A growing number of miRNAs have been associated with diseases in human, e.g. leukemia, pancreatic cancer, or Alzheimer's disease. Knowledge on miRNAs, their functions and evolution is critical for understanding the gene expression regulation in different species. Moreover, the evolution of miRNA genes themselves is still poorly understood. As a result, identification of miRNAs, both in plants and animals, became a critical issue in molecular biology, medical research and agriculture.

Machine learning techniques have been widely used in miRNA search field, as they help avoid some of the problems associated with homology-based approaches, like incapability of detecting species-specific miRNAs. However, machine learning tools usually suffer from some drawbacks as well. These include not addressing the class imbalancement problem, which may lead to overlearning the majority class and/or incorrect assessment of classification performance. Moreover, the tools tend to be usable in a very narrow range of species, usually the model ones.

Our goal was to improve miRNA classification procedure in terms of sensitivity, specificity as well as usability and computational time. To this point we implemented a new method of dealing with the class imbalancement problem, referred to as ROC-select. We also introduced novel miRNA features into the data representation. Several classification algorithms in combination with ROC-select were tested and random forest was picked as the one balancing best sensitivity and specificity. As a result, our method achieved significantly better performance when compared to other miRNA classification tools and can be applied in miRNA search experiments in a wide range of species. Reliable assessment of classification performance was guaranteed by using large, quality-filtered, and taxon-specific datasets in experimental procedure.

The miRNA classification method is freely available as a framework called HuntMi. HuntMi comes with trained models for animals, plants, viruses and separately for H. sapiens and A. thaliana. As a result, the tool can be used in miRNA classification experiments in a wide range of species. The user can use the models in experiments or train new ones before classification, using custom datasets.

P-1012

Population dynamics of bacteriophage T7 on heterogeneous *Escherichia coli* substrates

Wolfram Möbius[1], Andrew W. Murray[2], David R. Nelson[1]
[1]Department of Physics and FAS Center for Systems Biology, Harvard University, Cambridge, MA, USA, [2]FAS Center for Systems Biology and Department of Molecular and Cellular Biology, Harvard University, Cambrdige, MA, USA

How species invade new territories and how these range expansions influence the population's genotypes are important questions in the field of population genetics. The majority of work addressing these questions focuses on homogeneous environments. Much less is known about the population dynamics and population genetics when the environmental conditions are heterogeneous in space.

To better understand range expansions in two-dimensional heterogeneous environments, we employ a system of *Escherichia coli* and bacteriophage T7. Thereby, the bacteria constitute the environment in which a population of bacteriophages expands, forming a plaque whose growth is followed over time. We investigate which *Escherichia coli* strains are suited to generate a heterogeneous environment for phage T7 and study the expansion of the phage for a wide range of patterns of environments.

P-1013

Improving metabolic flux predictions using gene expression data

Abdelmoneim Desouki, Gabriel Gelius-Dietrich, Martin Lercher
Heinrich-Heine-University, Duesseldorf, Germany

Based on reasonable biological assumptions, flux-balance analysis (FBA) estimates steady-state flux distributions in metabolic networks without knowledge about kinetic parameters. It does this by maximizing a trait relevant for fitness (e.g., biomass yield) under biochemical and environmental constraints. However, solutions are not unique: several distinct metabolic flux distributions may result in the same biomass yield, and it is unclear which of them corresponds to the ‘real' biological fluxes.
We propose an FBA variant that uses additional gene expression data to select among alternative flux distributions. In contrast to a popular previous variant of the FBA scheme that considers ON/OFF expression status (expression-based FBA, or eFBA, Shlomi et al. 2008), we include quantitative expression data. To obtain a scale linking flux values to gene expression levels (as measured, e.g., by microarray experiments), we start with a flux variability analysis (FVA, Mahadevan et al 2003) for each reaction under a range of assayed conditions. We then find a best linear fit between the resulting flux ranges (which encompass fluxes that all support maximal biomass yield) and measured expression levels across conditions. In a given condition, this relationship is then used to assess the agreement of flux distributions predicted from quantitative gene expression data with all flux distributions consistent with maximal biomass yield (FBA solutions). The FBA solution with minimal distance to expression-predicted fluxes is considered to be close to the biologically realised flux distribution.

References

Mahadevan, R. & Schilling, C.H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab. Eng. 5, 264-276 (2003).

Shlomi, T., Cabili, M.N., Herrgard, M.J., Palsson, B.O., and Ruppin, E. 2008. Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26: 1003-1010.

P-1014

Genomic structures and gene evolution on mammalian X chromosomes

Yukako Katsura, Yoko Satta
The Graduate University for Advanced Studies(SOKENDAI), Hayama, Kanagawa, Japan

Of genomic structures, intrachromosomal segmental duplications (ISDs) are relatively large tandem and/or inverted repeats in neighboring regions on a chromosome. Counting of the number of ISDs on each human chromosome found more ISDs on X chromosome than any of the autosomes and Y chromosome. In this study, we show that the concentration of ISDs are not a general characteristic for X chromosomes, or not due to sequence-specificity of the human X chromosome, but are correlated with gene evolution and expression control.

ISDs on the X chromosome in four mammalian species with different origins or evolutionary histories (human, mouse, opossum, and platypus) were identified and characterized. The number or the size of ISDs was different among these species; ISDs on human and mouse X chromosomes were much larger in size, in the number, and more structurally complex than those in opossums. Moreover, gene density and the number of different gene families in ISDs-containing regions were larger in the human and mouse, whereas in the opossum, ISD-regions were gene-poor. Interestingly multiple X chromosomes of platypus did not have any ISDs, except for one on the X1 chromosome. Furthermore, autosomal regions syntenic to the human X chromosome in opossum and platypus were analyzed, but no concentration of ISDs was founded in these two species. These observations indicated that ISDs accumulated on the X chromosome in the eutherian ancestor.

In the human, more than 70% of the genes within X chromosomal ISDs were cancer-testis antigen (*CTA*) genes, and they are highly expressed in testis and cancer cells. The *CTAs* showed primate or eutherian-lineage specific, suggesting the recent origin and rapid evolution within ISDs. The amplification and complexity of ISDs can be evolutionally maintained by the emergence and functional constraint of *CTA*s, respectively. In fact, the evolution of one family of human *CTA*s, melanoma antigen A genes (*MEGA-A*), affected species-specific rearrangements of X chromosomal ISDs; the ISD-region possessing *MAGE-A* genes was under negative selection to maintain the immunological function of *MAGE-A*. In addition, the accumulation of X chromosomal ISDs might be involved in the mechanism for expression control, because the ISD-regions on X chromosomes were hypomethylated and genes within the ISDs were expressed specifically in germ cells.

P-1015

The Effects of Purifying Selection on Genealogies

Lauren Nicolaisen, Michael Desai
Department of Physics and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

Purifying selection has the potential to significantly distort patterns of molecular evolution relative to neutral expectations. These distortions form the basis for many of the statistics and inference methods used to detect the presence of purifying selection in the history of a population. Although several numerical methods and simulation studies have been successful at describing these distortions across a broad range of parameters, it is difficult to use these methods to make simple analytical conclusions about the effects of purifying selection on genealogies.

Building off earlier structured coalescent work, and introducing additional approximations valid in the strong selection regime, we recently derived simple analytical expressions for a time-dependent effective population size and time-dependent effective mutation rate that describe a population undergoing purifying selection. We calculate explicit formulas for various statistics of interest, as well as describe the dependence of these distortions on the parameters involved.

We extend our analysis to data sampled at different time points, and explicitly calculate a time-dependent coalescent rate as a function of sampling times. We show that the distortions due to purifying selection depend upon the time between sampling, thus providing potential power to distinguish between population expansion and purifying selection.

P-1016

Speciation within genomic networks: a case study based on lower Congo cichlidsp { margin-bottom: 0.21cm; }a:link {

Julia Schwarzer[1,2], Bernhard Misof[1], Ulrich K. Schliewen[2]
[1]*Zoologisches Forschungsmuseum A. Koenig, Zentrum für molekulare Biodiversitätsforschung, Bonn, Germany,*
[2]*Zoologische Staatssammlung München, München, Germany*

p { margin-bottom: 0.21cm; }

Hybridization is a common phenomenon both in plants and animals. Its impact on species diversity is controversially discussed as the interplay of homogenising gene-flow and speciation appears counterintuitive. Gene-flow between lineages with distinct ancestry, however, can increase genetic and phenotypic variation and has been hypothesized to facilitate adaptive radiation by providing increased standing genetic variation. Supporting empirical data remains however scarce. The riverine cichlid genus *Steatocranus* (Teleostei: Cichlidae), closely related to members of the East African cichlid radiation, radiated produced a small species flock in the lower Congo rapids. Phylogenetic analyses suggest that hybridization potentially contributed to the comparatively high species diversity in this genus. Testing for signals of of ancient gene flow based on a 2000 loci-AFLP data set provided strong evidence for a reticulate phylogenetic history of the genus *Steatocranus*. With the present study we give – to our knowledge, the first example of a complex reticulate network in vertebrates.

P-1017

Functional evolution of an anthocyanin pathway enzyme during a flower color transition

Stacey Smith[1,2], Shunqi Wang[2,3], Mark Rausher[2]
[1]University of Nebraska, Lincoln, NE, USA, [2]Duke University, Durham, NC, USA, [3]Nanchang Univeresity, Nanchang, Jiangxi, China

Dissecting the genetic basis for the evolution of species differences requires a combination of phylogenetic and molecular genetic perspectives. By mapping the genetic changes and their phenotypic effects onto the phylogeny, it is possible to identify changes which may have directly contributed to transitions to a new character state. Here we use phylogenetic and functional methods to trace the evolution of substrate specificity in dihydroflavonol-4-reductase (*Dfr*), an anthocyanin pathway gene known to be involved in the transition from blue to red flowers in *Iochroma*, a genus in the potato family, Solanaceae. Ancestral state reconstruction indicates that three substitutions occurred along the lineage where flower color changed while several additional substitutions followed the transition. Comparisons of enzyme function between ancestral proteins in blue- and red-flowered lineages and enzymes from present-day taxa demonstrate that evolution of specificity for red pigment precursors was caused by the first three substitutions, which were fixed by positive selection and which differ from previously documented mutations affecting specificity in this enzyme. The substitutions subsequent to the initial flower color transition were also adaptive and resulted in an additional increase in specificity. This pattern suggests that the present-day species differences in specificity for floral pigment precursors arose by a combination of selection on flower color and selection for improved pathway efficiency.

P-1018

The horizontal gene transfer events in the genus Aspergillus

Yi-Pei Ho, Chih-Hang Chen, Yeong-Shin Lin
National Chiao Tung University, Hsinchu, Taiwan

Horizontal Gene Transfer (HGT), also called Lateral Gene Transfer (LGT), plays an important role in evolution of prokaryotes and part of eukaryotes. HGT is the nonsexual process of genetic material transfer across species. In the phylogenetic view, the tree will show a network-like diagram distinct from the vertical transfer. It is well known that HGT events occur frequently among prokaryotes that can often increase fitness to colonize new environment. The common transfer events that people often mentioned is developing drug resistance. However, as genome sequencing era coming, HGTs found in eukaryotes increasingly. HGT events may explain some new traits of organisms and even new disease to human, animals and plants. There are several approaches, which based on genome-wide features, sequence similarity and phylogeny incongruence, can screen out potential HGT cases. Since, there are different possibilities that can contribute to the gene anomalies, HGTs should be confirmed by different approaches, especially phylogeny analysis. In this study, We find one HGT event, transferred from fungi to Stigmatella aurantiaca. From the analysis of phylogeny and genome characteristics, the orthologous gene of Aspergillus clavatus is the most closed gene sequence to the HGT gene, STAUR_2131. To find out where the sequence come from and any other gene take part in HGT event, we perform a large-scale scan for further study.

P-1019

The Role, Function, and Conservation of Upstream Open Reading Frames in Human

Yi-Chiao Fang[1], Arthur Chun-Chieh Shih[2], Yeong-Shin Lin[1]
[1]National Chiao Tung University, Hsinchu City, Taiwan,[2]Academia Sinica, Taipei, Taiwan

Upstream open reading frame (uORF) is an open reading frame whose start codon (Upstream AUG, uAUG) locate in the 5'UTR of an mRNA. Upstream ORF may affect the translation of main coding sequence by interfering the start of its translation. Although about half of known human transcripts contain at least one uORF, we don't know what roles uORFs play and how uORFs function in human. By analyzing the significant terms of Gene Ontology, we found that uORFs play important roles in metal ion binding and regulation of cellular processes, especially in the regulation of transcription. A uORF that interfere the translation of main coding sequence must be recognized by the ribosome on their uAUG. The features of uAUGs decide which uAUG may be recognized frequently. Beside the uAUG, there are several known features of uORFs that affects the starting of translation of uORF itself and the re-initiation of translation of following ORFs. According to these features, we can estimate how the main coding sequence of an mRNA is affect by the uORFs starting from 5'UTR. We found that dozens of uORFs that frequently affect the translation of main coding sequences are starting from a conserved uAUG. Our results suggest that uORFs may be conserved for their regulation functions and also influence the protein expression of many genes, play an important role in human.

P-1020

Revisit the origin of mitochondria using universal mitochondrial proteins.

Ding He
Uppsala University, Uppsala, Sweden

Despite the strong genomic evidence suggest the obligate parasitic Rickettsiales as a sister group to mitochondria, the origin of mitochondria has been a major dispute in biology. Phylogenetic studies often attempted to address this question with mitochondrial-encoded proteins resulting in unbalanced taxa sampling of eukaryotes and accelerated evolutionary rate that may ultimately lead to phylogenetic artefact due to long-branch attraction (LBA) or amino-acid compositional bias. In the present study I selected 20 (3 mitochondrial-encoded + 17 nuclear-encoded) universal mitochondrial proteins (UMPs) with a balanced taxonomic sampling of proteobacteria (with focus on alpha-proteobacteria) and eukaryotes. All 20 UMPs were thoroughly vetted for monophyletic eukaryote (with > 70% maximum likelihood bootstrapping support) and non-canonical grouping to rule out potential lateral gene transfer events and paralogy. The concatenated 20-UMP supermatrix produced phylogeny with maximum-likelihood bootstrapping using RAxML (mlBP) puts monophyletic eukaryote sister to monophyletic alpha-proteobacteria, whereas phylogeny with Bayesian-inferential posterior probability (biPP) using PhyloBayes with heterogeneous amino-acid substitution CAT model puts monophyletic eukaryote sister to Rickettsiales, both with good supports (mlBP 89% and 0.99 biPP). Discrepancy of the preliminary results suggests that more detail analyses are needed before any concrete conclusion could be drawn.

P-1021

Phylogenetic analysis of cave bear specimens from Niedźwiedzia Cave, Sudetes, Poland

Pawel Mackiewicz[1], Mateusz Baca[2], Anna Stankovic[3], Krzysztof Stefaniak[4], Adrian Marciszak[4], Michael Hofreiter[5], Adam Nadachowski[6], Piotr Weglenski[3]
[1]Faculty of Biotechnology, University of Wroclaw, Wroclaw, Poland,[2]Center for Precolumbian Studies, University of Warsaw, Warsaw, Poland,[3]Institute of Biochemistry and Biophysics, Polish Academy of Science, Warsaw, Poland, [4]Zoological Institute, University of Wroclaw, Wroclaw, Poland,[5]Department of Biology (Area 2), The University of York, Heslington York, UK,[6]Institute of Systematics and Evolution of Animals, Polish Academy of Sciences, Cracow, Poland

The vast majority of fossil remains in Late Pleistocene deposits from Niedźwiedzia Cave in Kletno, Sudetes, Poland, belong to the cave bear. Phylogenetic analyses based on a fragment of the mitochondrial D-loop region extracted from two cave bear samples showed unambiguously their close affiliation with the *Ursus ingressus* haplogroup. This taxonomic affiliation of the cave bear remains from Niedźwiedzia Cave was further confirmed by biometrical analyses of molar teeth and skulls. Our results represent the first record of *U. ingressus* north of the Carpathian Arch, while radiocarbon dating (> 49,000 yr BP) of the samples indicates that they represent some of the oldest specimens of this cave bear taxon known so far. Multi-method phylogenetic approach allowed analysis of the relationships between our samples and numerous publicly available cave bear sequences, considering the significance of particular clades, and discussing some aspects of cave bear phylogeography. The sequences of *U. ingressus* from Poland are most closely related to the specimens from the Ural Mountains and next to Slovenian samples, which may indicate migrations between Central and Eastern European populations. The internal placement of Ural samples among European specimens in phylogenetic trees and the older age of Polish samples than those from Urals suggest that the eastward expansion of *U. ingressus* may have started from Central Europe.

P-1022

Origins and evolution of the Etruscans' DNA

Silvia Ghirotto[1], Francesca Tassi[1], Erica Fumagalli[2,1], Vincenza Colonna[3,1], Anna Sandionigi[4], Martina Lari[4], Stefania Vai[4], Emmanuele Petiti[4], Giorgio Corti[5], Ermanno Rizzi[5], Gianluca De Bellis[5], David Caramelli[4], Guido Barbujani[1]
[1]Department of Biology and Evolution, University of Ferrara, Ferrara, Italy, [2]Department of Biotechnologies and BiosciencesUniversity of Milano-Bicocca, Milano, Italy, [3]Institute of Genetics e Biophysics "Adriano Buzzati-Traverso", National Research Council, Napoli, Italy, [4]Department of Evolutionary Biology, University of Firenze, Firenze, Italy, [5]Institute for Biomedical Technologies (ITB), National Research Council (CNR), Milano, Italy

The Etruscan culture is documented in Etruria, Central Italy, from the 7th to the 1st century BC. For more than 2,000 years there has been disagreement on the Etruscans' biological origins, whether local or in Anatolia. Genetic affinities with both Tuscan and Anatolian populations have been reported, but so far all attempts have failed to fit the Etruscans' and modern populations in the same genealogy. We extracted and typed mitochondrial DNA of 14 individuals buried in two Etruscan necropoleis, analyzing them along with other Etruscan and Medieval samples, and 4,910 contemporary individuals. Comparing ancient and modern diversity with the results of millions of computer simulations, we show that the Etruscans can be considered ancestral, with a high degree of confidence, to the modern inhabitants of two communities, Casentino and Volterra, but not to most contemporary populations dwelling in the former Etruscan homeland. We also estimate that the genetic links between Tuscany and Anatolia date back to at least 5,000 years ago, strongly suggesting that the Etruscan culture developed locally, without a significant contribution of recent Anatolian immigrants.

P-1023

2012 sees the 50th birthday of the molecular clock, but has it reached maturity?

Rachel C M Warnock, Philip C J Donoghue
University of Bristol, Bristol, UK

Given the vagaries of the fossil record, the molecular clock represents the only viable means of establishing an evolutionary timescale, and the past fifty years have witnessed significant methodological advances to improve its efficacy. Knowledge and assumptions about the underlying evolutionary processes can be modeled with increasing realism. In the absence of the known evolutionary timescale, which – if any – of the available methods are able to estimate divergence times with accuracy or precision? Robust testing of the molecular clock can only be achieved using simulated data, where the relationship between fossil evidence, rate variation and genetic divergence is known. Methods for simulating molecular and palaeontological data have never previously been combined. We apply the first test of the molecular clock using both simulated molecular and fossil data. Our results demonstrate that when the distribution of fossils is non-random, current approaches to calibration fail to generate reliable estimates of substitution rates and node ages. Given that the real distribution of fossils is highly non-random, this represents a ubiquitous problem for molecular clock studies. We show that probabilistic models based on fossil occurrence data can be used to obtain more accurate estimates of divergence times. Approaches to better summarize the relationship between palaeontological and molecular evidence demand critical attention.

P-1024

Human population genomics in time and space: paleogenomics of populations in Bulgaria

Meredith L. Carpenter[1], Hannes Schroeder[2], Nikola Theodossiev[3], M. Thomas P. Gilbert[2], Carlos D. Bustamante[1]
[1]Department of Genetics, Stanford University, Stanford, CA, USA, [2]Centre for Geogenetics, University of Copenhagen, Copenhagen, Denmark, [3]Department of Archaeology, Sofia University, Sofia, Bulgaria

With a few exceptions, most ancient human DNA studies to date have restricted their analysis to the mitochondrial DNA (mtDNA) and Y chromosome. These approaches have led to some interesting theories regarding the spread of human populations; however, they are inherently limited by their use of these two non-recombining markers, which are subject to forces such as genetic drift and natural selection and also represent only the histories of the female and male lines, respectively, from which they descend. Recently, the whole genomes of several ancient individuals have been sequenced. These genomes yielded much more information about the individuals' ancestry than their mtDNA alone; nevertheless, a single ancient individual is not sufficient for population genetic analyses. Thus, the goal of our study is to sequence the genomes of multiple ancient individuals from the same population. This type of study has the potential to dramatically improve our ability to address demographic questions about ancient human populations. We have begun the low-coverage sequencing of genomic DNA from the teeth of 16 individuals from different time periods (1500 BC-400 BC) in Bulgaria, and we plan to ultimately extend the study to at least 50 ancient Bulgarian individuals from the Neolithic to the Iron Age (6300 BC-400 BC). The results of these initial experiments will be presented, including the identification of mtDNA haplogroups and initial population genetic analyses. Ultimately, we plan to analyze whole-genome sequence variation in these individuals and to compare it to variation present in modern populations. This will be the first systematic population-level study of ancient human genomes and therefore will allow us address demographic questions that have previously been restricted to the realm of theoretical modeling using extant populations.

P-1025

Horse domestication: a computer simulation approach

Michela Leonardi[1], Christine Weber[1], Norbert Benecke[2], Mark G. Thomas[3,4], Joachim Burger[1]
[1]AG Palaeogenetik, Institute of Anthropology, SBII, Johannes Gutenberg University, Colonel Kleinmann-Weg 2, 55128, Mainz, Germany,[2]German Archaeological Institute, Im Dol 4-6, 14165, Berlin, Germany,[3]Research Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, WC1E 6BT, London, UK,[4]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvagen 18D, SE-752 36, Uppsala, Sweden

The domestication of horse played a key role in human history. It seems to have happened far both in time and space from the domestication of other ungulates such as cattle, pig, sheep and goat. Archaeological studies, nevertheless, failed in determining exactly the region and modality for horse domestication: several centers have been proposed (at least one in Europe and one in Central Asia) and the relationship between wild and early domestic populations are not clear. From a genetic point of view a phylogenetic approach on modern mitochondrial diversity could not find any structure related with geography or breeds.

In the last decade ancient DNA became an important tool to reconstruct past demography. We obtained more than 100 HVR I sequences from pre domestic and domestic specimens found in Europe and Central Asia. After collecting all the previously published ancient and modern comparable sequences from the sub mentioned regions, computer simulations with a Bayesian approach were performed in order to test demographic models related with single or multiple domestications with or without gene flow. A single domestication appears to be unrealistic on the basis of mitochondrial data, while possible model of multiple domestications will be discussed.

P-1026

Late Pleistocene population dynamics in the European collared lemmings (Dicrostonyx torquatus)

Selina Brace[1], Eleftheria Palkopoulou[2], Rebecca Miller[3], Mietje Germonpré[4], Love Dalen[2], John Stewart[5], Ian Barnes[1]
[1]Royal Holloway, University of London, London, UK, [2]Swedish Museum of Natural History, Stockholm, Sweden, [3]University of Liege, Liege, Belgium, [4]Royal Belgian Institute of Natural Sciences, Brussels, Belgium, [5]Bournemouth University, Bournemouth, UK

The dramatic climatic cycles during the Late Pleistocene are suggested to have had a major impact on the evolution and demography of arctic species. By comparing their past and modern geographic distributions, it is obvious that species either respond to such climatic changes by moving or become extinct. Thus, it is of great importance to understand the patterns of species response in order to make future predictions.
In this study we use ancient DNA from subfossil remains of collared lemmings (D. torquatus) up to ~50,000 years old, excavated from two caves in Belgium to investigate their population history. According to refugial theory, arctic species are presumed to exhibit population continuity in Central Europe during the Late Pleistocene. Our phylogenetic and population genetic analyses show that the history of the collared lemming Belgian population is rather complex, with population bottlenecks and several population replacements through time. Pinpointing the exact timing of these events will enable us to understand the processes driving the dynamics of the species.

P-1027

Improving the performance of true-Single Molecule Sequencing for ancient DNA

Aurélien Ginolhac[1], Julia Vilstrup[1], Jesper Stenderup[1], Maanasa Raghavan[1], Morten Rasmussen[1], Mathias Stiller[2], Beth Shapiro[2], Grant Zazula[3], Duane Froese[4], Kathleen E. Steinmann[5], John F. Thompson[5,6], Khaled A.S. AL-Rasheid[7], Tom Gilbert[1], Eske Willerslev[1], Ludovic Orlando[1]
[1]Centre for GeoGenetics; Natural History Museum of Denmark, Copenhagen, Denmark, [2]Department of Biology; The Pennsylvania State University, University Park, PA, USA, [3]Department of Tourism and Culture; Yukon Palaeontology Program, Whitehorse, Yukon Territory, Canada, [4]Department of Earth and Atmospheric Sciences; University of Alberta, Alberta, Canada, [5]Helicos BioSciences, Cambridge, MA, USA, [6]NABsys Inc, Providence, RI, USA, [7]Zoology Department; College of Science King Saud University, Riyadh, Saudi Arabia

Helicos true Single-Molecule DNA Sequencing (tSMS) has recently shown great potential for sequencing DNA molecules from Pleistocene fossils. We have previously demonstrated that tSMS outcompeted Illumina GAIIx platforms with regards to percentage of endogenous sequences recovered, when sequencing DNA extracted from a Pleistocene horse bone preserved in permafrost. Here we test the performance of two novel tSMS template preparation methods on ancient DNA. We show that DNA phosphatase treatments and mild denaturation temperatures increase the amount of sequence information recovered from ancient bones. Together with molecular preservation niches identified in large bone crystals, these procedures could be used to increase sequence coverage of ancient genomes.

P-1028

Origins and genetic legacy of Northern European Neolithic farmers and hunter-gatherers

Pontus Skoglund[1], Helena Malmström[1], Maanasa Raghavan[2], Jan Storå[3], Per Hall[4], Eske Willerslev[2], M. Thomas P. Gilbert[2], Anders Götherström[1], Mattias Jakobsson[1]
[1]Uppsala University, Uppsala, Sweden,[2]University of Copenhagen, Copenhagen, Denmark,[3]Stockholm University, Stockholm, Sweden,[4]Karolinska Institutet, Stockholm, Sweden

Farming culture originated in the Near East some 11,000 years ago, and had reached most of the European continent 5,000 years later. However, the impact of the agricultural revolution on demography and patterns of genomic variation in Europe remains unknown. We obtained 249 million base pairs genomic DNA from ~5,000 year-old remains of three hunter-gatherers and one farmer excavated in Scandinavia. Using new approaches for studying population structure using low-coverage genomic data, we find that the farmer is genetically most similar to extant southern Europeans, contrasting sharply to the hunter-gatherers whose unique genetic signature is most similar to extant northern Europeans. We use nucleotide misincorporation patterns in the data to show that this result is not affected by modern human contamination, and apply admixture models of population history to investigate the potential impact of the Neolithization on European genetic variation. Our results suggest that migration from southern Europe catalyzed the spread of agriculture, and that admixture in the wake of this expansion eventually shaped the genomic landscape of modern-day Europe.

P-1029

Plant population paleogenomics using long-term balanced polymorphisms: detection of old genetic bottlenecks using allelic genealogies at the self-incompatibility locus in Brassicaceae

Xavier Vekemans, Céline Poux, Pauline Goubet, Sophie Gallina, Vincent Castric
UMR CNRS GEPV, University Lille 1, France

Polymorphisms maintained by long-term balancing selection offer unique opportunities to detect drastic demographic changes in the distant past. In contrast, such signatures would be erased at most genomic loci, through the action of the neutral coalescent process. Genomic regions controlling homomorphic self-incompatibility in plants constitute unique opportunities to test a paleogenomic approach. Indeed the self-incompatibility locus is subject to very strong balancing selection leading to the co-occurrence within species of allelic lineages that diverged more than 20 millions of years ago in several plant families. In Brassicaceae, phylogenetic patterns of allelic genealogies at the self-incompatibility locus show striking variation among tribes. In the tribe Calamineae (Arabidopsis), many highly divergent lineages co-occur within allelic genealogies, which suggests that members of this tribe did not experience strong genetic bottlenecks in the distant past. In contrast, in the tribes Brassiceae (Brassica) and Biscutelleae (Biscutella), alleles cluster in two separate lineages, with large divergence between lineages, and low divergence among alleles within clusters. We suggest that these observed patterns reflect strong genetic bottlenecks that occurred independently in the distant past in these two tribes. We tried two approaches to estimate the time of these demographic events. One is based on comparing patterns of allelic genealogies between increasingly distant taxonomic groups, the other is based on molecular dating methods to estimate the time of burst of allelic diversification within each surviving allelic lineage. In Brassica, these time estimates were compared to the estimate of whole genome duplication events that have been recently determined.

P-1030

Using the doubly-conditioned site frequency spectrum to distinguish between alternative demographic models

Melinda Yang, Anna Sapfo-Malaspinas, Eric Durand, Montgomery Slatkin
UC Berkeley, Berkeley, CA, USA

Recent analyses of the genomes of the ancient hominins Neanderthals and Denisovans show that there exists a slight but significantly greater genetic similarity between non-African humans and Neanderthals that is not observed between African humans and Neanderthals. This suggests that the ancestors of non-African humans interbred with Neanderthals. However, an alternative explanation is that ancient structure in African human populations may account for the observations. We use the doubly conditioned frequency spectrum to determine whether we can distinguish between these two models of human demographic history—a model of recent admixture between Neanderthals and non-African humans and a model of ancient structure in Africa. Using simulations, we determine how the different demographic models affect the shape of the doubly conditioned frequency spectrum. Then, using data from the Complete Genomics Diversity Panel, we find both the doubly conditioned frequency spectra for each population and the individual pair-wise admixture rates. We show that the recent admixture model is the most parsimonious and plausible explanation and assess the admixture rate across multiple individuals from multiple populations.

P-1031

Targeted sequencing of ancient DNA on a population level scale

Mathias Stiller[1,6], Duane Froese[2], Grant Zazula[3], Matthew Wooller[4], Robert Wayne[5], Beth Shapiro[1,6]
[1]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, California, USA, [2]Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta, Canada, [3]Department of Tourism and Culture, Yukon Palaeontology Program, Whitehorse, Yukon, Canada, [4]Water and Environmental Research Center, University of Alaska Fairbanks, Fairbanks, Alaska, USA, [5]Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, California, USA, [6]Department of Biology, The Pennsylvania State University, University Park, Pennsylvania, USA

Advances in target enrichment protocols and high-throughput DNA sequencing technologies in recent years have allowed the targeted resequencing of specific genomic areas of interest from a variety of organisms. By adopting and refining those techniques, we here present preliminary results on the targeted sequencing of mitochondrial as well as nuclear loci from ancient DNA (aDNA). Several hundred-thousand bases of aDNA sequence data have been successfully retrieved from a large population sample of horses and bison from the Yukon Territory, Canada, dating back as far as ~80,000 years before present. A multi-locus aDNA population data set like this will enable us to precisely reconstruct the genetic make-up of entire populations and thus significantly enhance our understanding of how species and populations change over evolutionary timescales.

P-1032

Did Quaternary glacial cycles drive walrus diversity? A paleogenomic approach.

Tara Fulton[1,2], Grant Zazula[3], Michael Cousar[1], Beth Shapiro[1,2]
[1]The Pennsylvania State University, University Park, PA, USA, [2]University of California Santa Cruz, Santa Cruz, CA, USA, [3]Government of Yukon, Whitehorse, YT, Canada

Arctic sea ice is a key component to walrus survival. Moving pack ice provides a surface for molting, pupping, and resting between foraging trips on the sea floor. However, the amount of Arctic sea ice has varied dramatically during the Quaternary (~2.6 Ma to present), as compared to the levels observed at present. We hypothesize that sea ice distribution affects the spatial distribution and population connectivity of walruses (*Odobenus rosmarus*). During warm interglacial periods when Arctic sea ice was much reduced, walrus populations probably shifted northward and were more likely to come in contact throughout the Arctic. Conversely, in glacial periods, walrus populations may have become more isolated when the Arctic Ocean was covered in a continuous mass of sea-ice and populations became displaced into southern refugia. We are using targeted resequencing to obtain genome-wide data for >100 modern and historic and >60 radiocarbon-dated ancient individuals from across the Holarctic range. This includes nine radiocarbon-infinite (>55 thousand years ago; ka) samples from the northern coast of Yukon, Canada. As this region is more northern than most of the modern walrus range, this population probably lived during the previous interglacial period (~120 ka). With these data, we address (1) how walruses from the last interglacial period are related to modern walruses, (2) how genetic diversity has changed through the past ~120,000 years, and (3) when the modern subspecies (Atlantic, Pacific, Laptev Sea) diverged from one another. Through analyses of these data within the framework of known paleoclimatic and paleoceanographic variables, we hope to obtain key insights into how walruses have responded to past glacial cycles and a basis for predicting how the species will respond to ongoing climate change.

P-1033

Neolithic expansion into Europe is visible in complete mitochondrial genomes

Qiaomei Fu[1], Pavao Rudan[0,2], Svante Pääbo[0,1], Johannes Krause[0,3]
[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Croatian Academy of Sciences and Arts, Zagreb, Croatia, [3]Institute for Archaeological Sciences, University of Tübingen, Tübingen, Germany

The Neolithic transition from hunting and gathering to farming and cattle breeding marks one of the most drastic cultural changes in European prehistory. Short stretches of ancient mitochondrial DNA (mtDNA) from skeletons of Mesolithic hunter-gatherers as well as early Neolithic farmers support the demic diffusion model where a migration of early farmers from the Near East and a replacement of Mesolithic hunter-gatherers are largely responsible for cultural changes in subsistence strategies in Europe. In order to test if a signal of population expansion is still present in modern European mitochondrial DNA, we identified mtDNA haplogroups in the ancient DNA data that are typical for early farmers and hunter-gatherers, such as H and U, respectively. We then analyzed these haplogroups in 1,151 complete mtDNAs from present-day Europeans. Using Bayesian skyline coalescence on subsets of the complete modern mtDNAs we found evidence for a population expansion between 15,000 and 10,000 years ago in mtDNAs typical for hunter and gatherers, with a decline between 10,000 and 5,000 years and starting at 9,000 years ago a population increase for mtDNA typical of early farmers. This suggests that the spread of agriculture involved the expansion of farming populations into Europe but also the assimilation of resident hunter-gatherers. Our data show that contemporary mtDNA datasets can be used to study ancient population history when limited ancient genetic data are available.

P-1034

An improved library preparation method for ancient DNA

Marie-Theres Gansauge, Svante Pääbo, Matthias Meyer
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Despite huge advances in sequencing technology during the last decade, sequencing of ancient DNA still imposes great challenges. DNA repair mechanisms of organisms cease function *post mortem*, leading to a rapid decay of genetic substance. Ancient DNA is therefore usually present only in trace amounts and characterized by double-stranded and single-stranded breaks and alteration of DNA bases, which make it harder to generate DNA sequences from a fossil.

Most current high-throughput sequencing technologies require the conversion of template molecules into sequencing libraries by attaching two different artificial adaptor sequences to both molecule ends. Despite being originally developed for high-quality modern DNA, this process has been applied, albeit with modifications, to ancient samples.

We describe a new library preparation method specifically designed and optimized for highly fragmented DNA, which greatly increases the amount of sequence data that can be generated from small amounts of ancient DNA extract. We already demonstrated the power of this method by generating a high-coverage genome sequence of a Denisovan, an extinct human group related to Neandertals. We hope that this method will make additional samples accessible to genetic studies, which previously did not yield sufficient material for sequencing. In addition, the method offers the possibility to answer still open questions regarding fragmentation patterns in ancient DNA.

P-1035

The complete mitochondrial genome of a third individual from Denisova Cave

Susanna Sawyer[1], Bence Viola[1], Marie-Theres Gansauge[1], Michael Shunkov[2], Anatoly Derevianko[2], Svante Pääbo[1]
[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Paleolithic Department, Institute of Archaeology and Ethnography, Russian Academy of Sciences, Siberian Branch, Novosibirsk, Russia

A draft genome sequence was determined in 2010 from a small finger bone found in Denisova Cave in southern Siberia and was recently completed to 30-fold coverage. Its analysis reveals that it derived from an individual that belonged to a population related to, but distinct from, Neandertals. A large molar has also been described from Denisova Cave and shown to carry an mtDNA genome closely related to that of the finger bone.
A second molar was found in Denisova Cave in 2010. We have captured and sequenced the complete mitochondrial genome of this tooth. While the mtDNAs of the finger bone and the first molar differ at only two nucleotide positions, they carry 86 and 84 differences, respectively, to the second molar. Thus, the maximum amount of mtDNA differences observed among these three Denisovans found within one cave is almost twice as large as the maximum differences seen among six Neandertals for which complete mtDNAs are available. Interestingly, the mtDNA of the second molar has a shorter branch than the other two Denisovan mtDNAs, suggesting that it may be older than the others.

P-1036

The crucial first step: Computational analysis of ancient DNA reads

Stinus Lindgreen[3,1]
[1]University of Copenhagen, Copenhagen, Denmark, [2]Centre for GeoGenetics, Copenhagen, Denmark, [3]University of Canterbury, Christchurch, New Zealand

When working with ancient DNA reads obtained using next-generation sequencing tehcnologies, there are numerous challenges that need to be overcome before the downstream analyses of the data can be carried out. This is due to many factors including the limited amount of material, the fragmentation of the DNA, and the damage through deamination of the nucleotides. If not carefully dealt with, everything from mapping of the reads to genoptyping of the individual can be compromised.

As part of a number of ancient DNA studies, our group has developed different tools to help deal with these issues. This includes finding and removing fragments of the adapter sequences from the sequencing data before mapping, mapping tools that are tailored for short, damaged DNA reads, and the development of probabilistic tools for genotyping that can take damage into account.

For any group working with ancient DNA, it is of great importance to be aware of these issues and think of ways to deal with them. I will present our work in this area and present our solutions used in various projects including the sequencing of the first ancient human genome, and the sequencing of an Aborigine genome.

This talk will cover the algorithmic details of our genotyper, SNPest, which is based on a probabilistic graphical model that simulates the process from sampling to genotype, making it possible to model technology specific errors and other sources of variation. Also our new version of bwa that utilizes position specific scoring matrices to model the nucleotide distribution in the reads when mapping to a reference genome. This makes it possible to directly model e.g. position specific error rates, and this procedure can also be used to efficiently locate adapter fragments that have been sequenced but should be removed from the reads.

P-1037

Efficient Cross-Species Capture Hybridization and Next Generation Sequencing of Mitochondrial Genomes from Non-Invasively Sampled Colugo Museum Specimens.

Victor Mason[1,2], Gang Li[1], Kristofer Helgen[3], William Murphy[1,2]
[1]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA,[2]Interdisciplinary Program of Genetics, Texas A&M University, College Station, TX, USA,[3]Smithsonian Institution, National Museum of Natural History, Washington, D.C., USA

The advent of next-generation sequencing technologies (NGSTs) has transformed the way in which scientists approach a myriad of biological questions. Difficulties with applying NGSTs to phylogenetic problems does not lie with the sequencing technology itself, but with the preparative procedures for isolation of large, orthologous DNA regions across multiple divergent species. This problem is exacerbated for museum specimens, where contamination levels are often high and DNA quality varies greatly between samples, which makes traditional PCR based enrichments/amplifications inconsistent. However, capture-hybridization procedures allow for selection and enrichment of large orthologous regions even when samples have poor DNA quality and high levels of contamination. We report the development of a comprehensive capture-hybridization approach for sequencing mitochondrial genomes and nuclear genes from large collections of museum specimens that combines DNA-capture techniques with NGSTs. We successfully extracted, captured, and sequenced the mitochondrial DNA genome and nuclear gene fragments from 13 Sunda colugo museum samples (50-150 years old) from various locales throughout the South-east Asian archipelago with upwards of 13% pairwise sequence divergence. Colugos (flying lemurs) are a group of arboreal mammals that have poor dispersal capabilities outside of forested areas, have the largest patagium of any known mammal (enabling them to glide for 136m), are widely distributed across the Southeast Asian mainland and archipelago, and are the closest living relative to primates. They make up their own order, Dermoptera, and under current taxonomy two species are recognized, Galeopterus variegatus, the Sunda colugo, and Cynocephalus volans, the Philippine colugo. Evidence for multiple, deeply divergent colugo populations/species endemic to distinct geographic regions is supported by raw sequence divergence, biogeographic correlations, and maximum likelihood phylogenetic reconstructions showing distinct clades with high support values both across island systems and even within islands themselves (Borneo, Java). Further museum sampling brings our analysis to ~65 individuals, allow us to elucidate the complex phylogeographic history of this species in the context of Plio-Pliestocene SE Asia biogeography. Ongoing analysis and annotation of the first colugo genome provides opportunities for the development of nuclear probes for large-scale targeted resequencing. Our approach combining cross-species hybrid capture and next generation sequencing-based technologies with DNA isolated from museum samples that are well over 100 years old, opens up new avenues of research for many other historical and ancient specimens of mammals, and has potentially broad conservation implications for colugos as well as many other taxa where sampling in the wild is difficult.

P-1038

Molecular genetic analysis of ancient pig remains excavated from the Pong Takhop archaeological site in Saraburi Province, Thailand

Wunrada Surat[1], Mattana Wannajuk[1], Pradit Sangthong[1], Surapol Natapintu[2], Anchanee Kubera[1], Mingkwan Mingmuang[1]
[1]Faculty of Science, Kasetsart University, Bangkok, Thailand,[2]Faculty of Archaeology, Silpakorn University, Bangkok, Thailand

Ancient DNA isolated from thirteen pig remains (approximately 3,000 years BP) excavated from the Pong Takhop archaeological site in Wang Muang District, Saraburi Province, Thailand. Seven of thirteen ancient pig remains were successful to amplify and determine partial sequences (131 bp) of *cytB* gene. Analysis of the seven nucleotide sequences revealed that these ancient remains were *Sus scrofa*. To gain the information about the origin and dispersal history of Thai pigs, partial *cytB* sequences of twenty seven Asian and European pigs from GenBank database were incorporated in this study. Nucleotide variation, haplotype diversity and phylogenetic relationship among the ancient samples and the modern samples were investigated. A total of eleven polymorphic sites were detected and five nucleotide variations were identified in the ancient samples. Eight haplotypes were classified and the ancient samples belonged to five haplotypes (H2-H6). Two ancient samples (PTK2_1-3_1 and PTK1_6_1) were clustered into H2 with Chinese domestic pigs while an ancient sample (PTK2_1-3_2) was clustered into H5 with Asian wild boars. The other four ancient samples were classified into the unique haplotypes. In the maximum-likelihood phylogenetic tree, all of the ancient samples and Asian pigs were located in Asian clade while all of European pigs were located in European clade indicating the division of European and Asian pig origins. In Asian clade, two Asian subclades; Ancient group (3 ancient samples) and Asian group (4 ancient samples, Asian wild boars and Asian domestic pigs) were divided and Asian domestic pigs were discriminated from Asian wild boars with 64% bootstrap value. Our findings suggest that some ancient pigs located in Asian subclade might originate from Asian domestic pigs (especially Chinese domestic pigs) and/or Asian wild boars. In particular, some ancient pigs located in Ancient subclade might originate from local wild boars indicating that there were at least two maternal lineages of these ancient pigs.

P-1039

Characterizing recent evolutionary changes on the human lineage using the high-coverage Denisovan genome

Fernando Racimo, Martin Kircher, Janet Kelso, Svante Pääbo
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

We have increased the sequence coverage of the previously published genome of an archaic human from Denisova cave in southern Siberia to approximately 30-fold coverage. This high-coverage genome provides the opportunity to identify the complete set of sequence changes that have risen to high frequency or fixation in modern humans since the split from the common ancestor with Denisovans. These recent changes may set anatomically modern humans apart from other extinct hominin forms.

Here we present an updated catalog of all single nucleotide changes (SNCs) and insertion/deletion events (InDels) now identified using the high-coverage genome data. Using variation data for modern humans we are able to identify sites where the Denisova genome sequence is ancestral while the derived allele is either fixed or at high frequency (>90%) in modern human populations. We can confidently identify 258 fixed and 393 high-frequency amino acid substitutions in protein-coding genes. A further set of nucleotide changes and InDels in 3' and 5' UTRs, splice sites, miRNAs and motif patterns in regulatory regions were also identified. An analysis of the chemical impact of specific amino acid changes, as well as an assessment of evolutionary conservation allows us to prioritize those substitutions that might have the largest phenotypic effect. This catalog is a resource for researchers wishing to determine those genome sequence changes that have occurred recently on the human lineage. Analyses are underway to prioritize particularly interesting candidates for functional studies.

P-1040

Maximizing the information of DNA extracts obtained from skeletalremains

Christian Sell, Susanne Kreutzer, Melanie Strobel
Johannes Gutenberg University Mainz, Mainz, Rheinland-Pfalz, Germany

A great advantage of next-generation technologies for the field of ancient DNA is the creation of DNA libraries. By transferring extracted DNA into a library, molecules of all sizes and quantities are made available for subsequent applications.

Our goal is to obtain as much information as possible from DNA included in an extract through library preparation and in-solution capture enrichment. To increase the amount of data from endogenous DNA towards contaminations through microorganisms and to ensure that only specific genomic regions of interest are sequenced, we aim to apply two different in-solution capture arrays with one single library: One array for human nuclear loci and one for the whole mitochondrial DNA.

Our protocol is based on the library protocol by Meyer and Kircher 2010 for ancient DNA[1]. However, we included a self-developed modification in one of the enzymatic steps. This enables us to reduce the recommended purification steps in the original protocol. The DNA extract is first purified after the adapter ligation step, at the time when molecules show required base pair length. Sample specific indices are attached during a PCR as part of the full adapter sequence on one end of each fragment. This approach enables simultaneous sequencing on NGS platforms.

For the following in-solution capture self-designed single cRNA oligo sequences of defined length of 120 bases are used as baits. We designed a bait library targeted the complete human mitochondrial genome. Ancient DNA libraries are enriched separately prior to pooling for sequencing. In our approach we enrich samples with low quantity of DNA separately in a PCR plate without pooling. Additionally, specific blocking oligos for each index were designed to block the entire adapter sequence during hybridization step and thus increase the enrichment efficiency.

So far capture enrichments were performed for over 50 archaeological samples in different states of preservation and from different time periods. Furthermore, there is another capture assay in development. This assay for human nuclear loci combines nearly 400 neutral regions and more than 160 SNPs under selection.

We will present summarized results of the first data analyses, discuss successful experiments and point out possible pitfalls during the workflow.

[1] Meyer M, Kircher M (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harbor Protocols doi:10.1101/pdb.prot5448.

P-1041

Linking Codon Usage and tRNA Prevalence In The Smallest Photosynthetic Eukaryotes And Their Giant Viruses

Stephanie MICHELY[3], Eve TOULZA[1,2], Lucie SUBIRANA[1,2], Uwe JOHN[4], Valérie COGNAT[5], Laurence MARECHAL-DROUARD[5], Herve MOREAU[1,2], Gwenael PIGANEAU[1,2]
[1]CNRS UMR 7232, Banyuls sur mer, France, [2]UPMC, Banyuls sur mer, France, [3]INRA, UMR1319 Micalis, Jouy-en-Josas, France, [4]Alfred Wegener Institute for Polar and Marine Research, Bremerhafen, Germany, [5]UPR 2357 CNRS, Strasbourg, France

Codon usage bias is a prevailing genomic feature in a broad range of organisms as a result of a mutation-selection-drift balance. Selection on codon usage bias is the consequence of selection for translational optimization, triggering coevolution between codon usage and the most abundant tRNAs in highly expressed genes. Prasinoviruses are large (>200 kb) double stranded DNA viruses coding for several hundreds of protein coding genes and six viruses infecting small eukaryotic planktonic green algae in the Bathycoccus, Micromonas and Ostreococcus lineages have been sequenced recently. Their genomes contain most of the DNA replication machinery and several genes involved in transcription and translation, like tRNAs. Here we analyze expression data during a viral infection to investigate the role of selection on codon usage bias in one prasinovirus, OtV5, and its host *Ostreococcus tauri*. Codon usage bias in the host and in the viral genes increases with expression levels. The annotation of 47 tRNAs in the 13 Mbp genome of *O. tauri* provides evidence of a rare tRNA sharing strategy in this green algae and an over-representation of tRNAs corresponding to preferred codons. We also suggest that the viral tRNA pool complements the host tRNA pool to optimize the number of host tRNAs per viral amino acid. We further investigate the relationship between host and prasinovirus codon usage bias in 3 host-virus specific pairs. We show that selection for codon usage in these slow growing streamlined eukaryotes and their giant viruses can hardly be detected from whole genome host-virus codon usage correlation analysis.

P-1042

Cell Tropism Influences RNA Virus Nucleotide Substitution Rates

Allison Hicks, Siobain Duffy
Rutgers University, New Brunswick, New Jersey, USA

The high rates of evolution of RNA viruses are generally attributed to their replication with error-prone RNA-dependent RNA polymerases (RdRps). However, their long-term nucleotide substitution rates span three orders of magnitude and do not correlate well with per-base mutation rate variation. This variation could be explained by differences in virus lifestyle: cell and host tropism, transmission mode, and whether the infection is acute or persistent. Here we present a modern survey of the nucleotide substitution rates of RNA viruses of mammals, comparing novel and published Bayesian rate estimates. We examined evolutionary rates of structural genes from more than 70 viruses. Non-structural gene rates were found for half of these viruses and compared independently. All of our data sets were under substantial purifying selection ($dN/dS << 1$), eliminating selection pressure as a major source of substitution rate variation. Instead we found a significant association between certain cell tropisms and higher rates of evolution (intestinal and respiratory epithelial cells) or lower rates of evolution (neural and endothelial cells). Cell tropism appears to explain more rate variation among RNA viruses than genome length or measured mutation rate.

P-1043

A Replicate Experiment in Nature: Human Cytomegalovirus Evolution in Congenitally Infected Twins

Nicholas Renzette[1], Jeffrey Jensen[2], Timothy Kowalik[1]
[1]University Of Massachusetts Medical School, Worcester, MA, USA,[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Human Cytomegalovirus (HCMV) is a large, dsDNA virus that is the leading source of birth defects caused by an infectious agent. HCMV, unlike most viruses, is able to cross the placenta during pregnancy and establish a lifelong infection during fetal development. The viral population dynamics associated with this unique route of infection are poorly understood. Here, we present a study of HCMV genomic populations sampled from the urine of congenitally infected monozygotic twins at 1 month, 2 months, and 11 months of life. In agreement with our previous study, we find that the populations exhibit levels of nucleotide and amino acid diversity that are comparable to those of RNA virus populations. Data from the 1 and 2 months samples revealed the populations to be stable across time and hosts. The last timepoints, however, showed significant increases in intrapopulation diversity and interpopulation differentiation, as measured by nucleotide diversity and F_{ST}, respectively. The large amount of polymorphism data allowed us to develop a high resolution model of the population dynamics associated with infection of fetal and neonatal hosts. A maximum likelihood estimate of initial infection was 16.5 weeks gestational age (95% CI: 15.9 – 17.1 weeks), which is in good agreement with previous estimates based on patterns of maternal seroconversion data. We found evidence of three sequential population bottlenecks and expansions in both twins, though the magnitude of these events differed between hosts. The timing of these events is consistent with population bottlenecks and expansions that would occur during viral colonization of the placenta, fetal blood and fetal renal cells. Lastly, there is strong evidence of migration of alleles from a “ghost” population that occurred between 2 and 11 months of life within the viral population history from a single twin. This migration is interpreted to be a re-infection event (a phenomenon which has previously been difficult to identify and quantitate in HCMV infections). We currently are attempting to identify migrant tracts from the ghost population and more accurately calculate the timing of the reinfection. In total, this study provides a highly detailed model of the population dynamics of HCMV during congenital infections – and is an important case study for the applicability of commonly used methods of population genetic inference to illuminate medically important parameters of viral infection.

P-1044

Global Diversity of the smallest planktonic green alga (Mamiellophyceae) and their giant prasinoviruses from marine metagenomes

Eve Toulza, Nigel Grimsley, Gwenael Piganeau
Observatoire Océanologique, UPMC Univ Paris 06, UMR 7232, BIOM, Banyuls-sur-Mer, France

Photosynthetic picoeukaryotes like Mamiellophyceae (Prasinophytina) have small cell sizes (<3μm) and small genomes (13-20 Mbp) (1). They play a key role in oceanic ecosystems especially in coastal areas, where they can account for most of the autotrophic biomass. Their population dynamics is largely influenced by large double stranded DNA viruses (200 kbp) (2), found at high density in surface waters. Metagenomic approaches have provided the DNA content of filtered seawater samples in many oceanic areas and provide access to the gene pool of the community. Here we screened the metagenomes from the GOS expedition (3) and data hot from the Tara Oceans campaign (4) to extract Mamiellophyceae and Prasinovirus sequences. This was done by comparing the identity and coverage of the sequences with a database of available picoalgal and viral complete genomes. This method corresponds to a very low false positive rate (5) and is therefore well adapted to evolutionary genomic analyses. We used these sequences (i) to test the neutral hypothesis of biogeography for the distribution of Mamiellophycean-like and prasinovirus-like sequences and (ii) to test the role of the environmental variables on their base composition.

(1) Piganeau G, Grimsley N, Moreau H. Genome diversity in the smallest marine photosynthetic eukaryotes. Res Microbiol. 2011 Jul-Aug;162(6):570-7.

(2) Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, Grimsley N. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J Virol. 2010 Dec;84(24):12555-63.

(3) Rusch DB, Halpern AL Sutton G et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biology 5(3): e77.

(4) Karsenti E, Acinas SG, Bork P et al. A holistic approach to marine eco-systems biology. PLoS Biol 9(10): e1001177.

(5) Vaulot D, Lepère C, Toulza E et al. Metagenomes of the picoalga Bathycoccus from the Chile coastal upwelling. Submitted.

P-1045

The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella.

Touchon marie, Rocha Eduardo
Département Génomes et Génétique, Institut Pasteur, Microbial Evolutionary Genomics, Paris, France

Prokaryotes thrive in spite of the vast number and diversity of their viruses. This partly results from the evolution of mechanisms to inactivate or silence the action of exogenous DNA. Among these, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are unique in providing adaptive immunity against elements with high local resemblance to genomes of previously infecting agents. Here, we analyze the CRISPR loci of 51 complete genomes of Escherichia and Salmonella. CRISPR are in two pairs of loci in Escherichia, one single pair in Salmonella, each pair showing a similar turnover rate, repeat sequence and putative linkage to a common set of cas genes. Yet, phylogeny shows that CRISPR and associated cas genes have different evolutionary histories, the latter being frequently exchanged or lost. In our set, one CRISPR pair seems specialized in plasmids often matching genes coding for the replication, conjugation and antirestriction machinery. Strikingly, this pair also matches the cognate cas genes in which case these genes are absent. The unexpectedly high conservation of this anti-CRISPR suggests selection to counteract the invasion of mobile elements containing functional CRISPR/cas systems. There are few spacers in most CRISPR, which rarely match genomes of known phages. Furthermore, we found that strains divergent less than 250 thousand years ago show virtually identical CRISPR. The lack of congruence between cas, CRISPR and the species phylogeny and the slow pace of CRISPR change make CRISPR poor epidemiological markers in enterobacteria. All these observations are at odds with the expectedly abundant and dynamic repertoire of spacers in an immune system aiming at protecting bacteria from phages. Since we observe purifying selection for the maintenance of CRISPR these results suggest that alternative evolutionary roles for CRISPR remain to be uncovered.

P-1046

Genetic specificity in an insect virus interaction

Lena Wilfert[3,1], Mike Magwire[2], Frank Jiggins[3]
[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK,[2]Department of Genetics, North Carolina State University, Raleigh, USA,[3]Department of Genetics, University of Cambridge, Cambridge, UK

Coevolutionary arms races are predicted to lead to rapid evolutionary changes in both competing partners, with the host evolving resistance mechanisms giving rise to counter-adaptations in the parasite and vice versa. This process may lead to specificity in genetic interactions, where the outcome of an infection depends on the genotype of both the host and parasite. We have studied the genetic basis of resistance to parasite replication in an insect virus. The sigma virus is a natural parasite of *Drosophila melanogaster*, which is vertically transmitted and tightly coevolved with its host. We mapped a major-effect polymorphism previously known as the *ref(3)d* locus to a region containing the two paralogous genes CHKov1 and CHKov2. In a panel of inbred fly lines, we found that a transposable element insertion in the protein coding sequence of CHKov1 is associated with increased resistance to infection. This resistant allele has rapidly increased in frequency under directional selection and is now the commonest form of the gene in natural populations. Using genetic mapping and site-specific recombination, we identified a third genotype with considerably greater resistance that is currently rare in the wild. By screening a panel of wild viral isolates, we found that there is indeed evidence for a viral counter-adaptation to this resistance genotype. Candidate mutations for this phenotype include a non-synonymous mutation in the *m*-gene, which produces the viral matrix. This currently rare viral variant illustrates how parasites can rapidly overcome host resistance mutations.

P-1047

Identifying functional variation in viral phylogenies

Kevin Rattigan[1], Xiaowei Jiang[1,2], David Robertson[2], Mario Fares[1]
[1]Department of Genetics, Trinity College Dublin, Dublin, Ireland, [2]Faculty of Life Sciences, University of Manchester, Manchester, UK

Understanding the genetic variation underlying viral persistence, immune escape and drug resistance is increasingly important in biomedicine. For example, detecting patterns of variation in pathogen genomes that are due to drug-induced selection can help us understand the emergence of resistance to current drugs. Here, we present a study, based on functional divergence, of HIV-1 variation. The classification system, however, that places HIV-1 into distinct subtypes, is hindered by many complications such as the emergence of recombinant subtypes. Moreover different clusters of viruses in phylogenies can potentially have different properties (correlated with subtypes or not) and, thus, it is important to use a classification system that is based upon meaningful biological information. This includes properties such as the dynamics of infection, replication and the clinical outcome of infection. The aim of this study is twofold. Firstly we use functional divergence (applied to individual functional domains of the main viral proteins Env, Gag and Pol) to identify unique clusters of viruses at different levels in the HIV-1 phylogeny. Secondly, we use heatmaps to study enrichment patterns of functional divergence and subsequently cluster together potential functional clusters based on similarities in these patterns. This analysis indicates that the different clusters often have more similar enrichment patterns of functional divergence. Interestingly the sites found to be under functional divergence included N-linked glycosylation sites, drug resistance mutations, and one RNA interacting domain. Our approach, thus, identifies genetic variation that in some cases follows the traditional subtype classification, other distinct clusters in the phylogeny and convergent changes.

P-1048

A Bayesian Phylogenetic Mixture Model to Detect the Differential Selection Patterns of HIV Sequence According to Host Genetic Background

Sahar Parto, Nicolas Lartillot
University of Montreal, Montreal, Canada

The extensive rate of HIV mutation and adaptation makes the design of vaccine difficult, as it enables the virus to escape from the immune system (escape mutation). Frequent recombination and the natural selection driven by immune system even intensify this diversity. So the first step for designing efficient vaccines is to identify consistent patterns in viral adaptation, as a function of the specific genetic background of the host. It has been shown that polymorphisms in HIV-I are associated with particular host HLA (Human Leukocyte Antigen) alleles. For example, HLA-B57 and B35 are associated with long-term and short-term HIV control, respectively, and are likely to exert strong selection pressure on the virus. These associations confirm the effect of HLA-restricted CTL (Cytotoxic T-Lymphocyte) response on HIV evolution.
The interplay between mutation and selection at the molecular level has been modeled extensively in molecular evolutionary studies. Estimation of evolutionary patterns from homologous sequences is crucial for understanding the evolutionary processes like mutation rate and selection pressures. To do so, codon-based evolutionary modeling is a more realistic description of the substitution process in protein coding genes.
In this study, a differential mutation-selection model is developed which parameterizes mutational and selective effects bearing on the overall substitution process. This Bayesian mixture model is presented to detect the differentially selected amino acids of HIV coding sequence according to different host immune characteristics. This site-specific mixture model accounts for the heterogeneity of selection coefficients across the sequence and allows the selection for different amino acids to be classified as no selection, positive and negative selection.
This model is implemented in a MCMC framework for gag sequences from patients with identified genetic immune profile and phylogenetic dependencies. The significant (according to q-values) differential selection profiles are estimated for sequences in B-57+ and B-35+ individuals specifying which amino acids are selected for or against at each position of HIV sequence. We also estimated the HIV mutation rate and codon usage bias on HIV and compared it with that of human in which the virus replicates.
Keywords: selection, escape mutation, virus adaptation, HLA, Bayesian, phylogenetic, MCMC

P-1049

Interspecies transmission of viral haemorrhagic septicaemia virus between marine and freshwater fishes.

Anna Amanda Schönherz[1], Niels Lorenzen[2], Katja Einer-Jensen[2]
[1]Aarhus University, Faculty of Science and Technology, Department of Molecular Biology and Genetics, Tjele, 8830, Denmark,[2]Technical University Denmark, National Veterinary Institute, Department of Poultry, Fish and Fur Animals, Aarhus, 8200, Denmark

Viral haemorrhagic septicaemia virus (VHSV) is an RNA virus belonging to the family *Rhabdoviridae* that causes disease in teleost fish. Compared to other rhabdoviruses that infect fish, VHSV has an exceptional wide host range of more than 70 species spread across the Northern hemisphere. It occurs in marine and aquatic environments and causes disease in wild as well as cultured fish. *In vivo* studies indicate that the virus has adapted to different host species. One of the most definite adaptation processes occurred in host shift between European marine fish species and farmed rainbow trout. VHSV strains originating from European marine fish species are non-pathogenic to rainbow trout and vice versa. Markers for this phenotypic difference remain to be detected at the genetic level, as the four major genotypes correlate with geographic region rather than with host species.

In this study we investigate the possibility of interspecies transmission of VHSV between a marine fish species (turbot, *Scophthalmus maximus*) and rainbow trout (*Oncorhynchus mykiss*) using a cohabitation design. Furthermore, we explore differences in replication kinetics between marine and freshwater VHSV strains with or without adaptation to the host by analyzing the kinetics of viral shedding. We also determined whether a single infection cycle of a freshwater VHSV strain in a marine host can affect pathogenicity of the freshwater VHSV strain to its adapted host (rainbow trout).

The two fish species, turbot (marine) and rainbow trout (freshwater), were exposed to a marine strain (DK-4p168) adapted to turbot, and a freshwater strain (DK-3592B) highly adapted to rainbow trout, using a cohabitation design where turbot were ip challenged with the two strains, respectively, and grouped with naive rainbow trout. Mortality was recorded daily and water samples were taken between day 1 and 18 post infection to quantify viral shedding and thereby identify replication kinetics.

We identified that both VHSV strains replicate in turbot resulting in viral shedding. Viral replication of both strains show similar kinetic patterns but only the freshwater strain resulted in infection of cohabitating rainbow trout.

P-1050

Phylodynamics of influenza viruses before, during and after the 2009 swine flu pandemic

Gytis Dudas
University of Edinburgh, Edinburgh, UK

Influenza pandemics occur when influenza A viruses of reassortant and/or zoonotic origin enter naïve human populations. The three widely recognised influenza pandemics of the 20th century (occurring in 1918, 1957 and 1968, caused by influenza A virus subtypes H1N1, H2N2 and H3N2, respectively) were associated with increased morbidity and mortality compared to seasonal influenza epidemics. Of the three pandemics, two are known to have resulted in the replacement of previously circulating seasonal influenza viruses with the new pandemic viruses. This has been suggested to be the result of cross-reactive immune responses which can also limit influenza virus diversity at any given time. A minor pandemic in 1977 - caused by the reemergent H1N1 viruses which were thought to be extinct since 1957 - resulted in the co-circulation of post-1968 H3N2 and 1977 H1N1 influenza A viruses. Influenza A viruses also circulate with influenza B viruses which almost exclusively infect humans. Influenza B viruses exist as two clades known as Victoria and Yamagata lineages, which occasionally reassort.
The recent "swine flu" pandemic, originating in Mexico at the beginning of 2009, spread globally and became the first pandemic of the 21st century. To investigate whether pre-pandemic seasonal influenza viruses (influenza A H1N1, H3N2 subtypes and influenza B Victoria and Yamagata lineages) were affected by the 2009 pandemic, haemagglutinin sequences of influenza viruses were used to reconstruct the demographic and phylogeographic history of influenza A and B viruses over the past 5 years (2006-2011). The findings suggest that cross-reactive immune responses induced by either antigenically drifted seasonal or pandemic influenza A viruses can significantly affect other circulating human influenza A viruses. These interactions appear to have resulted in local extinctions of seasonal influenza A viruses in the Americas and Asia in 2007 and 2009, respectively. The effects of the 2009 pandemic on influenza B viruses, on the other hand, appear to have been minimal, suggesting the involvement of the adaptive, and not the innate, branch of the immune system in the extinction of seasonal influenza A viruses.

P-1051

Human evolution and HIV: A deep-time perspective on pathogenesis

M. Cyrus Maher, Ryan Hernandez
University of California, San Francisco, San Francisco, USA

Appreciating how hosts and pathogens interact is crucial for building a more thorough understanding of infectious disease. Despite the fact that HIV is perhaps the most thoroughly studied virus in history, many details about how HIV interacts with host cell-machinery remain uncharacterized. Further, the degree to which human evolution has been shaped by the ancestors of such interactions is similarly unclear. Conveniently, it may be fruitful to address these two questions at the same time. While modern HIV is a relatively new human pathogen, it is nonetheless part of a large family of lentiviruses that have interacted with primates for millions of years. To examine the extent to which lentiviruses have been a driver of primate evolution, we developed a regression-based technique for comparing the signatures of natural selection found in interacting host proteins to those in the rest of the genome (taking into account a wide range of attributes, such as gene length, GC-content, and tissue expression). Using phylogenetic techniques and an alignment of 7 primate exomes, we found that HIV-interacting proteins simultaneously show stronger signals of both positive and negative selection compared to the rest of the genome. This effect remains even after removing all genes known to interact with any other human pathogen. We then combined divergence data with polymorphism data from the 1000 Genomes Project to test for more recent signatures of selection on the human lineage. Together, these analyses provide insight into the evolutionary history of a pair of host-pathogen families, and in doing so suggest proteins and even specific amino acids that may be particularly important to the acquisition and progression of HIV in humans.

P-1052

Roadmap to a Bat's Major Histocompatibility Complex (MHC) Class I Region: Organization & Characterization

Justin Ng[1,2], Katherine Belov[2], Lin-Fa Wang[1], Michelle Baker[1]
[1]CSIRO (Livestock Industries) - Australian Animal Health Laboratory, Victoria, Australia, [2]Faculty of Veterinary Science, University of Sydney, New South Wales, Australia

In recent years, bats have been widely acknowledged as reservoir hosts to numerous high profile emerging and re-emerging viruses, such as Ebola virus, Rabies virus, Henipaviruses and SARS-like Coronaviruses. How bats coexist with these viruses without any clinical symptoms of disease remains a mystery. The long coevolutionary history of bats with viruses may have resulted in the evolution of immune adaptations that allow bats to control viral replication more effectively than other mammals. How bats control viral replication is poorly understood but could provide significant insights into the evolution of antiviral mechanisms in mammals. Towards understanding the control of viral replication in bats, we have focused on the adaptive aspect of the immune system, the MHC class I (MHC-I) system in particular. MHC-I molecules play an important role in the immune response to viral infection, presenting endogenously derived peptides to cytotoxic ($CD8^+$) T cells. $CD8^+$ T cells then release cytotoxins to targeted infected cells inducing apoptosis, hence controlling viral replication. To date, no studies have documented the diversity and repertoire of the MHC-I region in any bat species. Here, for the first time, we present a detailed characterization and organization of a megabat's (*Pteropus alecto*) MHC-I region. Preliminary findings suggest that bats may have a larger and more diverse repertoire of classical MHC-I genes, which are essential in viral clearance. Further characterisation of the bat MHC-I region will assist in gaining a deeper understanding of how bats coexist asymptomatically with viruses.

P-1053

Evolution of dUTPase from Chlorellaviruses and host *Chlorella*

Hideaki Moriyama
University of Nebraska-Lincoln, Nebraska, USA

Comprehensive understanding of host and virus relationships will allow us to find clinical targets for treatments and preventions of infectious disease. The most important information we can acquire is mechanical differences, such as in authentication and metabolic pathways, between host and virus. Similarity in the biological systems of hosts and pathogens makes the maneuvering window extremely small in effective drug development. Eve so, use of early stage proteins after the infection can be use to enlarge the window. We investigated roles of early protein with taking deoxyuridine triphosphatase (dUTPase) and *Chlorella* and Chlorelaviruses as the model system.

Chlorelaviruses posses huge dsDNA genome, and encode more than 200 expressing genes. Physiological demand of redundant genes involves better fitness to the environments. One way to asses the role of such genes is comparative studies of the gene products. We chose to compare dUTP pyrophosphatase (dUTPase) between Chlorelaviruses and their host Chlorella.

dUTPase hydrolyze dUTP, and produce dUMP and diphosphate. Most of the organisms have dUTPase as a housekeeping enzyme, where the enzyme plays a key role in keeping a balanced pool of dUTP and dTTP. It is because higher dUTP concentration increases a risk of uracil disincorporation during DNA synthesis. Viruses replicate in dividing cells can depend on the host in dUTPase activity. However, viruses target none-dividing cells sometimes encode dUTPase to avoid the risk of uracil misincorporation.

In this report, we describe the specific activity of dUTPase from host *Chlorella* showmen 7 fold higher than that of Chlorellaviruses. The relationship of enzymatic performance was reversed from human and human viruses, including HSV-1. With newly identified sequences of Chlorellaviruses were analyzed, phylogenic relations were reconstructed. It has been suggested that *Chlorella* virus acquire the enzyme function independently.

P-1054

Examining Putative SNPs Across Parasite Populations:The First Step in an Examination of Migration Across Populations

Christina Jenkins[1,2], Matt Settles[1]
[1]*University of Idaho, Moscow, ID, USA,* [2]*Washington State University, Pullman, WA, USA*

Much theoretical work has been dedicated to understanding how migration among isolated populations effects coevolutionary local adaptation. However, little empirical work corroborates the patterns predicted by theory, owing largely to the lack of genetic tools available to study coevolution in natural populations. Here we present findings of SNP ascertainment that will provide a critical genomic resource for linking theory and data. We use the New Zealand Mudsnail *Potamopyrgus antipodarum* and its closely coevolved trematode parasite *Microphallus*. This is not only a well established natural system for studying coevolution, but it is an excellent system in which to study the effects of migration of the parasite on local adaptation. We sequenced the transcriptomes of 4 different parasite populations and assembled the sequences de novo. We then looked for SNPs segregating within and among populations. These data will next be combined with GIS data to examine how the genetic material is moving among populations, establishing migration patterns and magnitude. This is the first step in using genomic data to evaluate theoretical coevolutionary models of natural systems.

P-1055

Detecting HIV-1 co-receptor usage from sequence variation and V3 structural information

Felix Feyertag, John Archer, David L Robertson
University of Manchester, Manchester, UK

HIV type 1 gains entry to $CD4^+$ cells by binding to a CD4 receptor and a chemokine co-receptor. The CCR5 co-receptor is most frequently utilized in early stages of infection, although in the majority of patients the virus evolves to utilize the CXCR4 co-receptor in later stages of infection. This tropism switch is associated with increased disease progression and the onset of AIDS. Depending on the co-receptor utilized, CCR5 or CXCR4, HIV-1 can be classed as either R5-tropic or X4-tropic, respectively, or dual-tropic if both receptors can be used. Determining tropism of HIV-1 is of clinical relevance during the application of CCR5-antagonist drugs, which inhibit R5-tropic HIV-1 from gaining cell entry. These drugs have no effect on CXCR4-using viruses and so this type of virus is routinely screened for before treatment with a CCR5-antagonist. Currently genotypic and phenotypic methods can be used to screen for tropism, although genotypic methods are preferable, due to lower costs and faster turnaround time. Here we present TroGene, a novel genotypic tropism prediction algorithm. TroGene classifies viral sequences as R5-tropic or CXCR4-using, based on transitions between 35 residues in close structural proximity in the V3 loop of the envelope gene. A reference structure is used from the protein databank to incorporate structural information about the V3 loop. In addition to the TroGene algorithm, we implement the charge rule, PSSM and SVM algorithms. The SVM is implemented in concordance with the method used by Geno2Pheno, while PSSM is implemented as in WebPSSM; the two methods that are commonly applied to tropism prediction. Implementing these algorithms allows for the direct comparison of the different methods when trained on identical training sets. We demonstrate that a combination of predictors provides increased reliability in predicting clinical outcome.

P-1056

Positive selection for gains of N-linked glycosylation sites in hemagglutinin during evolution of H3N2 human influenza A virus

Yoshiyuki Suzuki
Nagoya City University, Nagoya-shi, Aichi-ken, Japan

The number of N-linked glycosylation sites in the globular head of hemagglutinin (HA) has increased during evolution of H3N2 human influenza A virus. Here natural selection operating on the gains of N-linked glycosylation sites was examined by using the single-site analysis and the single-substitution analysis. In the single-site analysis, positive selection was not inferred at the amino acid sites where the substitutions generating N-linked glycosylation sites were observed, but was detected at antigenic sites. In contrast, in the single-substitution analysis, positive selection was detected for the amino acid substitutions generating N-linked glycosylation sites. The single-site analysis and the single-substitution analysis appeared to be suitable for detecting recurrent and episodic natural selection, respectively. The gains of N-linked glycosylation sites were likely to be positively selected for the function of shielding antigenic sites from immune responses. At the antigenic sites, positive selection appeared to have operated not only on the radical substitution but also on the conservative substitution in terms of the charge of amino acids, suggesting that the antigenic drift is not a by-product of the evolution of receptor binding avidity in HA of human H3N2 virus.

P-1057

Evidence for N-Glycan Shielding of Antigenic Sites during Evolution of Human Influenza A Virus Hemagglutinin

Yuki Kobayashi[1,2], Yoshiyuki Suzuki[2]
[1]Department of Zoology, Oxford, UK, [2]Nagoya City University, Nagoya, Aichi, Japan

After the emergence of influenza A viruses in the human population, the number of N-glycosylation sites (NGS) in the globular head region of hemagglutinin (HA) has increased continuously for several decades. It has been speculated that the addition of NGS to the globular head region of HA has conferred selective advantages to the virus by preventing the binding of antibodies (Ab) to antigenic sites (AS). Here, the effect of N-glycosylation on the binding of Ab to AS in human influenza A virus subtype H3N2 (A/H3N2) was examined by inferring natural selection at AS and other sites (NAS) that are located close to and distantly from the NGS in the three-dimensional structure of HA through a comparison of the rates of synonymous (d_S) and nonsynonymous (d_N) substitutions. When positions 63, 122, 126, 133, 144, and 246 in the globular head region of HA were non-NGS, the d_N/d_S was >1 and positive selection was detected at the AS located near these positions. However, the d_N/d_S value decreased and the evidence of positive selection disappeared when these positions became NGS. In contrast, d_N/d_S at the AS distantly located from the positions mentioned above and at the NAS of any location were generally <1 and did not decrease when these positions changed from non-NGS to NGS. These results suggest that the attachment of N-glycans to the NGS in the globular head region of HA prevented the binding of Ab to AS in the evolutionary history of human A/H3N2 virus.

P-1058

Identification and analysis of *trans*-splicing in human embryonic stem cells by transcriptome sequencing

Chan-Shuo Wu, Chun-Ying Yu, Ching-Yu Chuang, Hung-Chih Kuo, Trees-Juen Chuang
Academia Sinica, Taipei, Taiwan

Trans-splicing is a small class of the post-transcriptional events in higher eukaryotes, which generates transcripts that are orderly inconsistent with their corresponding DNA templates (in a non-co-linear fashion). Detection of *trans*-splicing is usually severely hampered by false positives arising from experimental artifacts and genetic rearrangements. Until recently only exceedingly rare *trans*-splicing events have been experimentally characterized. Here we develop a pipeline (designated as "TSscan"), which integrates different types of high-throughput transcriptome sequencing of different human embryonic stem cell (hESC) lines to effectively minimize potential false positives and identify novel *trans*-splicing events. Our results show that experimental artifacts may highly vary between different mRNA products and such an integrative analysis of varied NGS data can effectively rule out a tremendous amount of potential experimental artifacts. On the basis of the TSscan identification, we successfully experimentally confirm four previously uncharacterized intragenic *trans*-spliced isoforms of *CSNK1G3*, *ARHGAP5*, *FAT1* and *NCRMS* in hESCs. We find that these *trans*-spliced isoforms are all highly expressed in pluripotent stem cells (including hESCs and induced pluripotent stem cells (iPSCs)) and differentially expressed during the transition of pluripotent to differentiated statuses. Furthermore, we find that the *trans*-spliced isoform of *NCRMS* (ts*NCRMS*) tends to be specifically transcribed in hESCs/iPSCs and disruption of ts*NCRMS* expression can conspicuously affect the pluripotency maintenance of hESCs, suggesting that *trans*-splicing may be involved in pluripotency and early lineage differentiation-related regulation. These results reveal the functional significance of *trans*-splicing. TSscan thus helps expand the discovery of *trans*-splicing, opening up this important but rarely-investigated class of post-transcriptional events for comprehensive characterization.

P-1059

Genomic Sequencing of *Patella vulgata*: a Further Window into Mollusc and Lophotrochozoan Evolution

Nathan Kenny, Jerome Hui, Sebastian Shimeld
Department of Zoology, University of Oxford, Oxford, UK

The common limpet *Patella vulgata* is widely distributed around the European coast, and has acted as an excellent model for developmental biology research for more than half a century. As genomes of a small number of lophotrochozoans have now been made publicly available, we have sequenced the *P. vulgata* genome using the Illumina HiSeq2000 platform in order to further assess the process of genomic evolution in this neglected superphylum.

Two libraries of 180 bp and 500 bp were constructed from DNA extracted from the sperm of a single male and sequenced, providing 30x genomic coverage. Developmental genes (homeobox genes and those involved in the establishment of left/right asymmetry), genomic synteny, and other conserved regulatory elements were then compared to other metazoans. The *P. vulgata* genomic resources presented here will provide a better understanding of how genomes in the Lophotrochozoa evolve, and will be useful for a wide range of further developmental and genomic research.

P-1060

High resolution association mapping in an outbred *Drosophila melanogaster* population using Pool-Sequencing

Héloïse Bastide, Viola Nolte, Martina Visnovska, Ray Tobler, Andrea Betancourt, Christian Schlötterer
Vetmeduni, Vienna, Austria

Next Generation Sequencing (NGS) techniques provide powerful tools for the identification of the genetic basis responsible for variation in quantitative phenotypes (QTLs). Here, we test the performance of NGS in association mapping studies in an outbred population of *Drosophila melanogaster*. Since, *D. melanogaster* has extremely low levels of linkage disequilibrium, we reasoned that GWAS should be extremely powerful at an unprecedented level of resolution. To test our approach, we focused on an extremely well-studied trait, abdominal pigmentation variation in *D. melanogaster* females. About 5,000 F1 females obtained from naturally inseminated flies were scored for pigmentation and two replicates each of the 100 most extreme phenotypes were sequenced. Our results confirm the efficiency of our approach. In addition to several genes with a proven role in pigmentation (e.g. bab), we identified additional candidates, which are currently being functionally tested. Our analyses suggest that levels of linkage disequilibrium were low enough to identify the genes, if not the causative SNPs. Hence, we propose that our new approach (NGS speed mapping) provides an excellent tool for GWAS studies, in particular for species with low levels of linkage disequilibrium.

P-1061

Does convergent morphology implies convergent development? Towards a comparative transcriptomics of tooth development.

Marie Sémon, Vincent Laudet, Sophie Pantalacci
IGFL, ENS de Lyon, Lyon, France

Morphologies are the product of developmental mechanisms, underpinned by developmental genes and pathways. Conserved morphologies are thought to be underlaid by similar developmental mechanisms, although the degree of similarity can vary. Indeed, the developmental pathways underpinning homologous characters are not static, even when the phenotype does not change (so-called developmental drift). New morphologies are built by the tinkering of existing developmental pathways (rewiring of gene networks, co-option of genes or gene networks at work in other parts of the organism, novel genes...). In this context our question is to decipher wether the developmental mechanisms involved in a morphological convergence are similar or different.

Our model is the study of the convergence of the upper molar morphology in two rodent species, the mouse and the spiny mouse : their upper molar crown displays two additional cusps (hills). This convergence is supported by paleontological data, and the development of rodent tooth benefits from a strong developmental background. Development can be seen as a temporal sequence of pattern transformations, in which gene expression plays a crucial role. We aim therefore at comparing the development of these teeth by a time-series comparative transcriptomic analysis involving these two species and two control-species retaining the ancestral cusp number, gerbil and hamster. The lower molars of the four species also retained the ancestral cusp number, and will thus serve as internal controls.

With our model and approach, we aim ultimately at answering 3 main questions. 1) Does a convergent morphology imply a convergent development at the morphological level? 2) Can we find genes (or genes of the same pathways) showing evidence of convergence in their expression profiles? 3) At the whole-transcriptome level, is there is evidence of some level of convergence, notably at specific stages of the development? This poster will present the project and preliminary data on this dataset (developmental data for mouse and first comparative data for mouse/hamster comparisons).

P-1062

A tale of four phyla: Contrasts between adaptive coding and noncoding changes during metazoan evolution.

David Garfield[1], Courtney Babbitt[2], Olivier Fedrigo[2], Eileen Furlong[1], Gregory Wray[2]
[1]European Molecular Biology Laboratory, Heidelberg, Baden-Württemberg, Germany,[2]Duke University, Durham, NC, USA

Whole-genome scans for selection are useful for identifying novel loci that may contribute to adaptive processes. They can also reveal evolutionary patterns that are not visible from the perspective of individual genes, such as evidence of recurrent selection on functional categories of genes. The vast majority of these scans have focused on coding DNA. However, changes in non-coding, regulatory DNA also play important, and potentially distinct, roles in adaptive evolution. Here we use ML-based evolutionary models and a formal statistical meta-analysis to compare signatures of adaptive evolution in coding and regulatory (5' flanking) DNA using representatives from four clades: sea urchins, hominids, *Drosophila*, and nematodes. Across all groups, we find a consistent enrichment of adaptive changes in coding genes associated with immunity, cell adhesion, and gamete recognition, suggesting functional "rules" for how selection acts on coding DNA. As previously reported, we identify a strong signature of adaptive evolution in regulatory DNA associated with neurogenic genes along the human lineage not seen in surveys of coding DNA, suggesting an important role for non-coding changes in the evolution of human-specific traits. Surprisingly, we observe similar enrichments for regulatory regions of neurogenic genes across taxa, as well as evidence for selection on regulatory DNA for metabolic genes. These data support the perhaps obvious observation that both coding *and* non-coding changes contribute to adaptation. However they do so in distinct and repeatable ways independent of organisms' particular life-histories. Additionally, scores for adaptive evolution for homologous genes in sister species are often correlated, while scores for homologous regulatory regions are not, suggesting non-coding adaptations are more species-specific than are coding adaptations. Across all taxa, we observe a negative correlation between the extent of adaptive evolution in a coding gene and the number of tissues or developmental stages in which it is expressed. No such correlation is observed for non-coding regions, supporting the hypothesis that pleiotropy and the modular organization of regulatory elements influence how selection operates on mutations in coding and regulatory regions. To better understand regulatory evolution, we analyzed ChIP-Chip experiments in *Drosophila*. We observe that the degree of cooperative binding of transcription factors at regulatory elements influences their evolutionary rate, with cooperative binding allowing for more rapid sequence turnover. Together, these data highlight how the distinct functional properties of coding and non-coding DNA influence their evolution and provide sets of phenotypic traits that might be modulated by these different genomic regions.

P-1063

Developmental constraints on transcriptome evolution vary for different molecular features

Barbara Piasecka[1,2], Pawel Lichocki[3], Marc Robinson-Rechavi[1,2]
[1]University of Lausanne, Lausanne, Switzerland,[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland,[3]Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland

Two main hypothesis of the evolution of embryonic development have been put forward so far. First, an early conservation model predicts that the highest conservation occurs at the beginning of embryogenesis. It dates back to Karl von Baer who postulated that embryos of different species progressively diverge from one another during ontogeny. Second, an hourglass model predicts that the highest conservation can be found during mid-embryogenesis. It has been proposed when the morphological variation in the early stages of development was observed. Nowadays, the hourglass model is commonly accepted, although a formal characterization has been elusive. Recent studies have reported several molecular characteristics supporting the hourglass model. To this aim they compared descriptive statistics of expression values of all genes between developmental time-points. Such a methodology introduces dependencies between the sets of expressed genes which are compared, and consequently can produce results biased by genes expressed in many time-points. To overcome this problem, we used an alternative "modularization" approach to study the evolution of zebrafish development. We first decomposed the genes into different modules, which contained genes that were expressed only in one of the six different developmental stages (i.e., cleavage, gastrula, pharyngula, larva, juvenile and adult). Next, for every module we obtained five characteristics: 1) gene sequence conservation, 2) genes' age, 3) gene expression conservation, 4) one-to-one orthologs, and 5) non-coding sequence conservation. The first three characteristics suggest that all developmental stages are conserved at the same level. The number of one-to-one orthologs (i.e., genes without secondary duplication) supports the early conservation model, whereas the regulatory region conservation supports the hourglass model. Thus different levels of molecular evolution seem to follow different patterns of developmental constraints.

P-1064

Identification of Pigment QTL in African Cichlids from Lake Malawi

Claire O'Quin, Alexi Drilea, Kelly O'Quin, Matthew Conte, Thomas Kocher
University of Maryland, College Park, MD, USA

Pigmentation patterns are one of the most recognizable phenotypic traits of organisms. Research on the genetic basis of pigment patterning in vertebrates has mostly involved the analysis of mutants in laboratory models. The genetic basis of pigment pattern formation in natural populations is still not fully understood. The Lake Malawi cichlids, *Metriaclima zebra* and *M. mbenjii*, differ in several aspects of their pigmentation, including differences in the color of fins, barring on the body, and color of the head. Analysis of an F_2 hybrid cross between these species suggests a small number of genes (1-4) control each pigmentation trait. Several of the traits are highly correlated in the F_2, suggesting shared developmental pathways may control multiple traits. To identify the quantitative trait loci associated with these different pigmentation traits, we first used RAD tag sequencing to identify ~700 useable single nucleotide polymorphisms (SNPs). We genotyped these SNPs in 150 F2 males and created a linkage map using JoinMap. We then performed a QTL analysis using rQTL, and identified several genomic regions and candidate genes associated with our pigmentation traits.

P-1065

Sequencing and analysis of the European amphioxus (*Branchiostoma lanceolatum*) transcriptome

Silvan Oulion, Stéphanie Bertrand, Mohamed Belgacem, Yann Le Pétillon, Hector Escriva
CNRS, Banyuls-sur-Mer, France

The basally divergent phylogenetic position of amphioxus (Cephalochordata), as well as its conserved morphology, development and genetics, make it the best proxy for the chordate ancestor. Particularly, studies using the amphioxus model help our understanding of vertebrate evolution and development. Thus, interest for the amphioxus model led to the characterization of the transcriptome and complete genome sequence of the American species: *Branchiostoma floridae*. However, recent technical improvements, allowing induction of spawning in the laboratory on a daily basis with the Mediterranean species *Branchiostoma lanceolatum*, have pushed European evo-devo researchers to adopt this model although no genomic neither transcriptomic data are yet available. To fill this lack we used pyrosequencing method in order to characterize the B. lanceolatum transcriptome that we further compared with the one of B. floridae.

Starting from total RNA from nine different developmental stages of *B. lanceolatum*, (i.e. from 8 cell embryos to adult), a normalized cDNA library was constructed and sequencing on Roche GS FLX (Titanium mode) was performed. Around 1,4 million of reads were produced, and assembled into 70,530 contigs (with an average length of 490 bp). Overall 37% of the assembled sequences were annotated by BlastX, and their Gene Ontology terms were determined. Our results were then compared to genomic and transcriptomic data of B. floridae to underline similarities and specificities of both species.

High throughput sequencing approach used in this study allowed us to obtain a high-quality amphioxus (*B. lanceolatum*) reference transcriptome. Indeed, 83% of the predicted genes in the B. floridae complete genome sequence are found in the *B. lanceolatum* transcriptome while only 41% were found in the *B. floridae* transcriptome obtained with traditional Sanger based sequencing. Therefore, this set of ESTs could be used now as a reference transcriptome for the amphioxus.

P-1066

A social chromosome in the red fire ant

Yannick Wurm[1,2], John Wang[4], Oksana Riba-Grognuz[2,3], Mingkwan Nipitwattanaphon[2], DeWayne Shoemaker[5], Laurent Keller[2]
[1]Queen Mary, University of London, London, UK, [2]University of Lausanne, Lausanne, Switzerland, [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [4]Academia Sinica, Taipei, Taiwan, [5]USDA-ARS, Gainesville, Florida, USA

Explaining how interactions between genes and the environment influence social behavior is a fundamental question, yet there is limited relevant information for species exhibiting natural variation in social organization. The fire ant *Solenopsis invicta* is characterized by a remarkable social polymorphism: The presence of one or multiple reproductive queens within a colony as well as other phenotypic and behavioral differences are completely associated with allelic variation at a single Mendelian factor marked by the gene *Gp-9*. Furthermore, the *b* allele at *Gp-9* is a rare example of a "green beard gene" because *b* workers favor the reproduction of *b* queens by executing queens that do not carry *b*. This selfish allele has not reached fixation because of balancing selection: The phenotypes associated with *b* are counter-selected in certain environments, and *bb* homozygotes are lethal (Keller & Ross, 1998).

The fire ant genome project focused on a single male who had the *Gp-9 BigB* genotype. We subsequently sequenced and assembled *de novo* the genome of a male with the alternate genotype, *Gp-9 b*. In addition, we used restriction site associated DNA sequencing (RAD, Baird *et al* 2008) to obtain genotypes at over 6,000 loci across the genome for 80 males that are *Gp-9 BigB* and 80 males that are *Gp-9 b.*

We find that *Gp-9* is in complete linkage disequilibrium with 6% of the genome. The two haplotypes do not recombine between social forms and show extensive differences in structure and gene content, similar to sexual chromosomes. We thus will provide the first evidence for a genomic region having the properties of a social chromosome.

P-1067

Elucidating the genetic landscape of cellular morphogenesis in an obligate fungal symbiont

Henrik H. De Fine Licht, Anders Tunlid
Lund University, Lund, Sweden

Fungi are masters of evolutionary adaptation, which have led to their enormous diversification and utilization as chemical factories of many potent secondary metabolites. Indeterminate growth and the capacity to produce distinct cell types are important sources of variation in fungal development. However, it remains largely unknown how natural selection on the genome drives adaptation of cellular morphogenesis. In many obligate mutualistic interactions the strong, but often narrow, selection pressure for a specific form or function provides an ideal framework for studying the genomic basis for fungal development. Our research identifies the key genetic adaptations of symbiotic fungi cultivated by fungus-growing ants to better understand the underlying evolutionary forces shaping fungal morphogenesis. A sub-group of these fungi form a distinct morphological adaptation to the symbiosis, i.e. inflated hyphal tips, called gongylidia. These structures are preferentially eaten by the cultivating ants and serve both as nutrition and as a source of fungal enzymes. These enzymes are important for efficient degradation of the plant substrate the ants use as substrate for cultivating the fungus garden. Using transcriptome sequencing and expressional profiling (RNA-seq) we identify and compare signatures of structural and expressional genetic adaptation in the evolution of these unique fungal structures.

P-1068

Using comparative genomics to disentangle the evolution of the Animal Kingdom

Jordi Paps Montserrat, Peter W.H. Holland
Department of Zoology, University of Oxford, Oxford, UK

Our knowledge of the evolution of the Metazoa is currently changing dramatically thanks to the spectacular advances in molecular biology. Molecular phylogenetics has provided a new Tree of Life, while research into the relationship between evolution and development (evodevo) explores how genes and development patterns have diversified through animal evolution. Moreover, the recent transformation of evolutionary biology induced by next generation sequencing has produced a plethora of genomic information for the major branches of the tree of the animals. However, the genome sampling is currently biased towards model organisms and the study of the interaction between genome biology and evolution, phylogenetics and development is still in its infancy.

To fill in those gaps, we have carried out a comparative genomic analysis of the major branches of Animal Kingdom to better understand the evolution of metazoan genomes and the origin of animal diversity. We have incorporated newly sequenced genomes that greatly increase the sampling of the metazoan superclades, mainly the usually neglected Lophotrochozoa. Specifically, we have identified orthologs for multiple genes families in different animal taxa using cutting edge methods. Then we have investigated the patterns of gene gains and losses in a well known phylogenetic framework, allowing us to recognize gene families whose evolutionary dynamics may be tied to the origin of the main metazoan clades. We also have tested the suitability of genomic qualitative features to address the evolutionary relationships between taxa. Finally, we have assessed whether there is a relationship between the rise and fall of this gene families, their biological functions and the morphological evolution of the animal groups. Our results enhance our understanding of the interplay between genomics, evodevo and the diversity of the Animal Kingdom.

P-1069

Isolation barriers in *Spodoptera frugiperda* - a genomic approach

Sabine Haenniger[1], Gerhard Schoefl[2], Steffi Gebauer-Jung[1], Heiko Vogel[1], David G. Heckel[1], Astrid T. Groot[1,3]
[1]MPICE, Dept. of Entomology, Jena, Germany, [2]Hans Knoell Institute, Dept. of Cell and Molecular Biology, Jena, Germany, [3]University of Amsterdam, IBED, Amsterdam, The Netherlands

In *Spodoptera frugiperda* (Lepidoptera: Noctuidae) two sympatric and morphologically identical, but genetically and behaviourally different "host-strains" have been identified: the corn-strain, occurring on larger grasses (e.g. corn, sorghum) and the rice-strain inhabiting smaller grasses (e.g. rice, bermudagrass). However, despite the distribution on different host plants in the field, various behavioural assays have failed to show a difference between the strains in female oviposition choice, larval preference or performance (Haenniger, unpubl. res; Meagher, pers. comm). Also, recent field studies show many exceptions from the reported distribution (Haenniger, Marr, and Juarez, unpubl. res), so a strict host association of the two strains can be questioned. However, the strains *do* differ in their timing of reproductive activity with the corn strain mating early and the rice-strain mating late at night. Our research aims to find the genetic basis of this timing difference. We are also taking an unbiased genomic approach to identify additional isolation barriers between the two strains.
In assessing the genetic basis of the timing difference, we identified three QTL involved in the phenotype. Utilizing RAD sequencing, we constructed a sequence-based linkage map that allowed us to homologize the QTL to the *Bombyx mori* genome and identify candidate genes from the circadian rhythm in these regions. One candidate gene, *vrille*, is located on the major QTL for the timing difference. Its structure in both strains, as well as expression levels over time, is now beeing investigated for strain-specific differences.
To determine which regions in the genome are differentiated between the two strains, we are performing a genome-wide scan, using RAD sequencing as well. We are analysing genome-wide SNP differences between corn-strain and rice-strain populations from 8 different geographic regions in the Americas. By analyzing the two strains from independent geographic regions, we are disentangling geographic from strain-specific SNP variation. Variations will then be mapped to the linkage map. Combining the two steps, we will assess whether strain-specific SNPs are clustered in specific regions in the genome or spread throughout, and whether these regions hit additional candidate genes. This analysis will thus identify genomic regions under differential selection, and shed light on the importance of habitat choice and timing for the strain divergence. On a larger scale, this project will provide novel insights into mechanisms of sympatric speciation and reproductive isolation on a molecular level.

P-1070

The unusual *hox* clusters of the basal teleost, *Pantodon bucholzi*, reveal tetralogy as a cryptic subtype of homology.

Kyle Martin[1], Peter Holland[1]
[1]*University of Oxford, Oxford, UK,* [2]*University of Oxford, Oxford, UK*

Whole genome duplications (WGD) are massive-scale genetic changes and were instrumental in shaping the evolutionary history of vertebrate genomes. Two WGDs occurred early in vertebrate evolution (2R) and a third WGD (3R) took place in the common ancestor of the ~26,000 species of teleost fish; these were responsible for the duplication of all genes including the *hox* gene clusters. After the 3R event, the now tetraploid genome reverted to a diploid state in a process known as diploidization, with extensive duplicate gene divergence and loss, often in species-specific patterns. Since the 3R event occurred at the base of teleost fish, over 300 million years ago, subsequent genomic changes obscured the initial stages of diploidization and patterns of divergence. To investigate these ancient events, we exploited the fact that the osteoglossomorph (bony-tongue) fish diverged from the majority of teleosts, soon after the 3R event. We show that following whole genome duplication, the resolution of the *hox* gene clusters of *Pantodon bucholzi* was radically different to those of other studied teleosts, including all current genomic models. Using next-generation sequencing, we reconstructed the entire hox gene complement of *Pantodon* and found it to contain 45 full length *hox* genes, 3 *eve/evx*-like genes and 7 miRNAs which assemble into only 5 clusters, the fewest found in any teleost to date. The discovery of "missing ohnologs" in the *hoxb2*, *hoxb4*, *hoxb9* and *hoxb13* families also makes the duplicated hoxb clusters in *Pantodon* the most complete pair of 3R-duplicated hox clusters described to date. Interestingly, our phylogenetic analyses indicate that osteoglossomorph fish diverged from other teleosts before the diploidization of all *hox* gene clusters was complete. We propose, therefore, that genes of osteoglossomorphs and other teleosts can show 'tetralogy', a novel type of homology defined as the relationship between pairs of duplicated genes in two species whose most recent common ancestor had not diploidized the locus in question, making each individual duplicate in one species equally related to both duplicates in the other, and obfuscating any assignment of 1:1 orthology. This proposal has potentially broad implications towards shaping our understanding of gene family evolution in species which diverged rapidly following a WGD event.

P-1071

Alternation of ploidal phases and genome-wide molecular evolution

Peter Szovenyi[1,4], Jonathan Shaw[2], Gregory Wray[2], Andreas Wagner[1,6]
[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland,[2]Biology Department, Duke University, Durham, NC, USA,[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland,[4]Institute for Systematic Botany, University of Zurich, Zurich, Switzerland,[5]MTA-ELTE-MTM Ecology Research Group, Budapest, Hungary,[6]The Santa Fe Institute, Santa Fe, NM, USA

All sexually reproducing eukaryotes experience an alternation of two ploidal phases separated by the processes of meiosis and syngamy. In organisms exhibiting biphasic life cycles alternating generations build up a multi cellular body and thus natural selection has the opportunity to shape the evolution of both phases. Genes showing generation-specific deployment are expected to experience different evolutionary forces owing to the functional, morphological and ploidal divergence of the alternating phases. Evolutionary theory predicts that ploidal difference has a profound effect on the evolution of organismic life cycles by influencing the number of mutations that arise and the efficacy of natural selection acting on them. Therefore, ploidal difference between phases is thought to represent one of the most important forces governing the evolution of diverse life cycles. Nevertheless, it is still unclear whether ploidal difference per se is sufficient to exert a general effect on the molecular evolution of the genome that is fundamental to the understanding of the evolution of eukaryotic life cycles.

Photosynthetic eukaryotic organisms represent an outstanding model group to explore this question exhibiting an amazing diversity of biphasic life cycles. Here we use genome-wide gene expression and sequence polymorphism data for green plants and algae having sporophyte (diploid phase) and gametophyte (haploid phase) generations with highly variable dominance, morphological complexity, function and interdependency to investigate this evolutionary issue in details. Applying various measures of selection we show that ploidal level of the generations appears to have a unified effect on the evolution of genes in photosynthetic eukaryotic organisms when both generations are relatively well-developed. By contrast, when one generation is highly reduced and experiences highly specialized function this effect may disappear and thus function may override the effect of polidal difference. Altogether, our analyses suggest that ploidal differences between generations may be an important driving force of life cycle evolution in organisms where phases are not highly specialized to one particular function. By contrast, extreme functional specialization may override the effect of ploidy and thus this has to be incorporated into theoretical models of life cycle evolution.

P-1072

Ancestral peptidergic signaling systems of bilaterian animals

Olivier Mirabeau, Jean-Stéphane Joly
CNRS, Gif-sur-Yvette, France

Neuropeptides are small secreted proteins that signal through G protein-coupled receptors (GPCR) to modulate the activity of behavior and physiology-controlling neurons in the brain of animals. We define a peptidergic system (PS) to be the association of a group of closely related peptides with their receptors. We have screened the sequence databases to better understand the origin of vertebrate PS and clarify the evolutionary relationship between all the very diverse PS that have been described in both protostomes and deuterostomes.
We used human annotated GPCR sequences as baits, using BLAST, to retrieve potential peptide GPCRs from the genome of 14 other bilaterian species. Clustering of receptors sequences allowed us to recover three main groups that corresponded to rhodopsin beta, rhodopsin gamma, and secretin GPCR subfamilies. Sequences from each list were aligned to build phylogenetic trees that were analyzed to infer the presence of ancestral vertebrate or arthropod-type GPCRs. We then used a Hidden Markov Model trained on known vertebrate peptide precursors to derive a shortlist of potential peptide precursor candidates that were screened for the presence of short conserved motifs of known vertebrate and arthropod-type peptides. Alignments of homologous precursors were built in a multi-step trial-and-error process, gradually integrating or discarding sequences. Peptide precursor homology was assessed through the analysis of an alignment-free Neighbour-Joining tree that was constructed from the list all predicted peptides using a non-standard distance adapted to the study of short peptides.

With this analysis we confirm the presence of at least 9 bilaterian systems previously hypothesized. We propose 4 novel homology associations between protostome and deuterostome systems and recover at least 11 systems that have not been hypothesized to have been present in the most recent common ancestor of bilaterians (the urbilaterian). These include systems that have only been characterized in either a protostome (e.g. Leucokinin) or a vertebrate (e.g. TRF). They also include 5 uncharacterized bilaterian systems that have apparently been lost in ecdysozoans and vertebrates.

Our results lend further support to the theory that the urbilaterian was an animal with a sophisticated physiology and nervous system. We believe that these newly established homologies will provide the evo-devo community with new markers to study ancestral cell types, yield insights into the fundamental functions of vertebrate peptidergic systems and offer new training data for computational biologists interested in peptide-receptor coevolution studies.

P-1073

Molecular evolution of Hox genes after whole genome duplication in a basal ray-finned fish, and surprising patterns of Hox expression in various body plan features at later stages of development.

Karen Crow
San Francisco State University, San Francisco, CA, USA

Vertebrates have experienced several rounds of whole genome duplication in the stem lineages of deep nodes within the group, and a subsequent event in the stem lineage of the teleosts-a diverse group of ray-finned fishes. Based on the first full Hox gene sequences for any member of the Acipenseriformes, the American paddlefish, we confirm that an independent whole genome duplication occurred in the paddlefish lineage approximately 42 million years ago. We obtained sequences spanning the entire HoxA cluster and six genes on the HoxD gene cluster. These clusters are located on different chromosomes, and maintain conserved synteny relative to bichir, zebrafish, stickleback and pufferfish, as well as human, mouse, and chick. We also provide a gene genealogy for the duplicated *fzd8* gene in paddlefish. Taken together, we clarify that the American paddlefish has a duplicated genome and highlight implications on comparative analyses in the study of the “fin-limb transition”, as well as gene and genome duplication in bony fishes. There were interesting trends in the molecular evolution of the alpha and beta paralog clusters for both the HoxA and HoxD genes. The posterior Hox genes, or 5’ genes that share homology with the Drosophila *Abd-B* gene, are expressed during pectoral fin and limb development. And we found interesting patterns of expression for these same genes in various body plan features at later stages of development, that have not been previously described.

P-1074

The genetic basis of the evolution of eye and head morphology in *Drosophila*

Maarten Hilbrant[1], Saad Arif[1], Corinna Hopfen[2], Linta Kuncheria[1], Nico Posnien[1], M. Daniela Nunes[1], Christian Schlötterer[2], Alistair P. McGregor[1]
[1]*Oxford Brookes University, Oxford, UK,* [2]*Veterinärmedizinische Universität Wien, Vienna, Austria*

A huge diversity of compound eyes has evolved among insects. These changes have often involved differences in ommatidia number and ommatidia size: two important parameters for visual acuity and light sensitivity respectively. However, the eyes of insects develop in concert with other head capsule tissues. Therefore, there may be a trade-off between eye size and other head traits, such as face width, that could affect the evolution of overall head morphology. Indeed, we have found that differences in eye morphology in *D. mauritiana* and *D. simulans* are associated with reciprocal changes in face width: *D. mauritiana* generally has larger eyes and a narrow face while *D. simulans* exhibits smaller eyes and a wider face. These differences in overall eye size are caused by changes in both facet size and ommatidia number. We subsequently mapped eye size and face width differences between *D. simulans* and *D. mauritiana* and discovered that different loci are responsible for inter-specific variation in these two traits. Furthermore, introgression of the region responsible for larger eyes in *D. mauritiana* into *D. simulans* increases eye size without affecting face width. This suggests that the size of the eyes is genetically de-coupled from the width of the face, which implies that eye size could potentially evolve without changing the size of the remaining head cuticle. We are currently performing high-resolution mapping of the evolved loci to determine the precise genetic basis for differences in these traits and how they shape coordinated differences in eye size and face width during development.

P-1075

Transcriptional profiling of the choanoflagellate *Salpingoeca rosetta* during simple multicellular development

Tera Levin, Nicole King
University of California, Berkeley, Berkeley, California, USA

The origin of animals was marked by the evolution of multicellularity in a previously unicellular or colonial progenitor. The study of choanoflagellates, the closest living relatives of animals, may reveal cellular and molecular innovations that contributed to this evolutionary transition. The choanoflagellate *Salpingoeca rosetta* can reversibly transition between unicellular and multicellular states, although it is unknown whether (or to what degree) the molecular mechanisms regulating *S. rosetta* multicellularity are homologous with those required for animal development. The recently sequenced genome of *S. rosetta* allows for investigation into the genes involved in the regulation, formation, and maintenance of this simple multicellular state. With improved techniques for inducing synchronized colony development in *S. rosetta* cultures, we are investigating genome-wide transcriptional changes across a time course of *S. rosetta* colony formation. I will present the results of an RNA-seq time course of multicellular development in a choanoflagellate, focusing on findings with potential regulatory or evolutionary significance.

P-1076

Mechanism Underlying Convergence of Albinism in Cave Adapted Animals

Helena Bilandžija[1], Helena Cetkovic[1], William R. Jeffery[2]
[1]Rudjer Boskovic Institute, Zagreb, Croatia,[2]University of Maryland, College Park, USA

Caves are unique environments that drive the same or similar morphological, physiological and behavioral adaptations in every animal group that has successfully invaded them. Albinism, the regression and loss of melanin pigmentation, is an omnipresent feature that has evolved by convergence in all phyla with cave dwelling representatives. However, the molecular basis of albinism is currently known in only one cave adapted animal: Astyanax mexicanus cavefish. In this species, different loss-of-function mutations in the oca2 gene cause albinism in at least three independently evolved cavefish lineages. OCA2 functions during the first step of melanin biosynthesis, the conversion of L-tyrosine to L-DOPA. Here we ask what is the molecular defect resulting in the evolution of albinism in other cave animals? Since the melanin synthesis pathway is generally conserved among animals, we have used a melanogenic substrate assay to survey for defects in this pathway in albino cave animals belonging to many different phyla. The assay involves supplying exogenous substrates, such as L-tyrosine or L-DOPA, to lightly fixed specimens, and subsequently detecting the presence of melanin as deposits of black pigment. The addition of L-DOPA, but not L-tyrosine, produced black pigment in diverse albino cave animals, including a sponge, a planarian, annelids, mollusks, arthropods, and several vertebrates other than Asytanax, indicating that the initial step of the pathway is defective in all these animals. In some of the cases, L-DOPA treatment restored melanin pigmentation in patterns resembling those of closely related surface-dwelling relatives. Therefore, albinism has evolved by a defect at the first step of melanin biosynthesis in all cave-adapted animals tested thus far, and L-DOPA can restore pigmentation, indicating that all downstream components of the pathway are present and potentially functional in these animals. Our results show that albinism has evolved by a defect in the first step of melanin biosynthesis in cave animals belonging to many different phyla, suggesting that there is an adaptive advantage to a block at the beginning of the pathway.

P-1077

The Adaptive Evolution Database: An Update Russell Hermansen and David A. LiberlesDepartment of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA

Russell Hermansen, David Liberles
University of Wyoming, Laramie, WY, USA

The Adaptive Evolution Database (TAED) is a gene family database that is phylogenetically indexed through gene tree/species tree reconciliation between underlying gene trees and a reference species tree. The database includes multiple sequence alignments, maximum likelihood phylogenetic trees, dN/dS values from the branches model of PAML, and gene tree/species tree reconciliation to place events in the context of the species tree and identify gene duplication events. An update on the latest version will be provided in addition to ongoing methodological research to improve inference in the database, including the detection of positive selection and a characterization of duplication events.

P-1078

Late replicating domains have higher divergence and diversity in Drosophila melanogaster

Claudia Weber[1,2], Catherine Pink[1], Laurence Hurst[1]
[1]*University of Bath, Bath, UK,* [2]*Uppsala University, Uppsala, Sweden*

Several reports from mammals indicate that an increase in the mutation rate in late-replicating regions may, in part, be responsible for the observed genomic heterogeneity in neutral substitution rates and levels of diversity, although the mechanisms for this remain poorly understood. Recent evidence also suggests that late replication is associated with high mutability in yeast. This raises the question whether a similar effect is operating across all eukaryotes. Limited evidence from one chromosome arm in D. melanogaster suggests the opposite pattern, with regions overlapping early-firing origins showing increased levels of diversity and divergence. Given the availability of genome-wide replication timing profiles for D. melanogaster, we now return to this issue. Consistent with what is seen in other taxa, we find that divergence at synonymous sites in exon cores, as well as divergence at putatively unconstrained intronic sites is

elevated in late-replicating regions. Analysis of genes with low codon usage bias suggests a ~30% difference in

mutation rate between the earliest and the latest replicating sequence. Intronic sequence suggests a more modest difference. We additionally show that an increase in diversity in late-replicating sequences is not owing to replication timing covarying with the local recombination rate. If anything, the effects of recombination mask the impact of replication timing. We conclude that, contrary to prior reports and consistent with what is seen in mammals and yeast, there is indeed a relationship between rates of nucleotide divergence and diversity and replication timing that is consistent with an increase in the mutation rate during late S-phase in D. melanogaster. It is therefore plausible that such an effect might be common amongst eukaryotes.

P-1079

Genome-wide mutational dynamics in *Saccharomyces cerevisiae* lead to complex patterns of selection and coadaptation

Mario Fares[2,1], Orla Keane[4], Lorenzo Carretero-Paulet[1], Gary Jones[3]
[1]Consejo Superior de Investigaciones Cientificas, Valencia, Spain, [2]University of Dublin, Trinity College, Dublin, Ireland, [3]National UNiversity of Ireland Maynooth, Maynooth, Ireland, [4]Animal & Bioscience Department, Teagasc, Dunsany, Ireland

Discerning the spectrum of fitness effects of mutations is an important aim in evolutionary genomics, as it allows understanding of how novel functions and adaptations emerge. The fixation dynamics of mutations underlying evolutionary adaptations remain controversial. In particular, researchers have classified mutations into either being neutral or adaptive. To understand the interplay between mutations, fitness consequences and compensatory evolution, we interrogated the dynamics of genome evolution in five unstressed lines of *Saccharomyces cerevisiae* evolving neutrally for 2200 generations. Sequencing and analysis of 27 *S. cerevisiae* genomes at different time points of their evolutionary history unravelled a striking and complex distribution of mutations. We found that non-synonymous mutations accumulated linearly with time, in agreement with the neutral theory but their effects presented evidence of selection. Mutations in highly connected genes in parallel experimental lines suggest complex adaptive leaps in a highly rugged fitness landscape. Despite the small number of fixed mutations their effects on fitness were significant suggesting that the number of essential genes is greater than previously predicted. The mutated genes displayed a significant number of protein-protein interactions and genetic interactions, supporting the role of genome-wide compensatory epistasis and protein co-adaptation in all five experimental evolution lines. Our results shed light on the interplay dynamics of drift and selection, unearth complex patterns of genetic and protein interactions, contribute to the understanding of functional complementation between duplicated genes and have important implications in the dissection of the evolutionary dynamics of endosymbiotic genomes.

P-1080

Population genetic inference from single genomes

Priyanka Sinha[1], Aslihan Dincer[2], Daniel Virgil[2], Guang Xu[2], Yu-Ping Poh[2], Jeffrey Jensen[1]
[1]ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, Lausanne, Vaud, Switzerland,[2]Univ of Massachusetts Medical School, Worcester, Massachusetts, USA

Describing the adaptive and demographic history of human populations has remained a central focal point of modern population genetics. Two major areas of interest have been the use of polymorphism data to detect so-called 'footprints' of selective sweeps - patterns produced as a beneficial mutation arises and rapidly fixes in the population, as well as to quantify the population history of modern humans. Based on numerous simulation studies and power analyses, the necessary sample size for achieving appreciable power has been shown to vary from a few individuals to a few dozen, depending on the test statistic. And yet, the sequencing of multiple copies of a single region, or of multiple genomes as is now often the case, incurs considerable cost. Two recent studies have proposed methods to perform population genetic inference based on single genomes: one proposing a modified HKA-approach to detect recent selective sweeps comparing between species, and the other a pairwise sequentially Markovian coalescent model to estimate historical population size change. Given the potential importance of these findings, we here investigate the properties of the proposed statistics taking a simulation approach, in order to quantify type-I and type-II error under a variety of population genetic models. Indeed, results highlight a number of areas of critical concern, which may indeed lead to highly biased results suggesting the need for great caution when using single-genome based approaches.

P-1081

RECURRENT MUTATIONS IN THE HUMAN MITOCHONDRIAL GENOMIC PHYLOGENY - EVIDENCE FOR CONVERGENT EVOLUTION?

Liron Levin, Ilia Zhidkov, Yotam Gurman, Dan Mishmar
Ben Gurion University of the Negev, Beer Sheva, Israel

Mutations frequently reoccur in the human mitochondrial DNA (mtDNA) phylogeny mainly due to the high mtDNA mutation rate. However, it is unclear whether recurrent mtDNA mutations that became fixed in human phylogenetic sub-groups (RFMs) are potentially functional, thus being subjected to selective constraints. We hypothesized that a subset of the human mtDNA RFMs carries functional attributes. To test our hypothesis we constructed a phylogenetic tree from 5067 publicly available non-redundant whole human mtDNA sequences. Secondly, comprehensive parsimony and maximum likelihood analysis revealed 24,115 mutational events that occurred throughout the human mtDNA phylogeny, of which 121 RFMs were either non-synonymous (N=81) or RNA gene mutations (N=40). To evaluate the functionality of non-synonymous changes we assessed evolutionary conservation among 296 mammalian species (using ConSurf), and used a previously described pathogenicity score (SIFT). Functional potential of RFMs in RNA genes was assessed by evolutionary conservation and calculation of the potential effect on structural stability (delta-G). Since disease-causing mutations are clearly functional we assessed their functional potential in protein coding genes (N=35) and RNA genes (N=25). Functionality values were significantly different between RFMs and disease-causing mutations ($p<0.001$). However, 20\121 RFMs (11 in non-synonymous, 9 in RNA genes) had indistinguishable functionality values from disease-causing mutations. Since these RFMs possess strong functional potential they could reflect convergence evolution.

P-1082

Insertion-biased gene conversion for short indels

Evgeny Leushkin[1,3], Alexey Kondrashov[1,2], Georgii Bazykin[1,3]
[1]M.V. Lomonosov Moscow State University, Moscow, Russia, [2]University of Michigan, Ann Arbor, USA, [3]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia

Recombination between homologous loci is accompanied by heteroduplex formation. Repairing mismatches in heteroduplexes often leads to single nucleotide substitutions known as gene conversion events. Gene conversion was shown to be GC-biased in different organisms; i.e., an AT->GC substitution is more probable in this process than GC->AT substitution. We observed that the insertion/deletion ratio for short indels is positively correlated with the recombination rate in *Drosophila melanogaster, Homo sapiens* and *Saccharomyces cerevisiae.* Whole-genome data on indel polymorphism and divergence in *D. melanogaster* rule out mutation biases and selection as the cause of this trend, pointing to insertion-biased gene conversion as the most likely explanation.

P-1083

Positive and negative selection in splice sites

Stepan Denisov[1], Georgii Bazykin[1,2], Alexander Favorov[3,4], Andrey Mironov[1,2], Mikhail Gelfand[1,2]
[1]A.A. Kharkevich Insitute for Information Transmission Problems, Moscow, Russia, [2]M.V. Lomonosov Moscow State University, Moscow, Russia, [3]The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore/Maryland, USA, [4]N.I. Vavilov Institute of General Genetics, Moscow, Russia

Splice sites (SSs) are crucial for proper mRNA splicing in eukaryotic cells, and therefore can be expected to be shaped by strong selection. Nevertheless, in mammals and in other intron-rich organisms, beyond the two critical nucleotides, SSs have relatively low fidelity to the consensus sequence within a species, and are not very well conserved between species. Some data indicate that selection does not necessarily favor the highest closeness to the consensus. It is well known that SSs of alternative exons are weaker than of constitutive ones. Here, we study the patterns of selection in SSs in the evolution of human, dog and mouse and of different *Drosophila* species. Overall, the alternative SSs are more conserved than the constitutive SSs. While the constitutive SSs invariably favor the consensus nucleotide, the selection in alternative SSs favors the existing nucleotide, no matter whether consensus or not. The mode of selection in alternative SSs is similar between weak and strong sites, suggesting inherent differences in the fitness landscapes of different SSs, rather than uniform selection for low site weight in cassette-exon SSs. Stronger and site-specific selection in alternative SSs is consistent with the existing data on tight regulation of alternative splicing.

P-1084

Looking for evidence of co-evolution between nuclear and mitochondrial genomes in a situation of interspecific hybridization

José Melo-Ferreira[1], Pierre Boursot[2], Nicolas Galtier[2], Paulo C. Alves[1,3]
[1]*CIBIO, Centro de Investigacao em Biodiversidade e Recursos Geneticos, Universidade do Porto, Vairao, Portugal,* [2]*Universite Montpellier 2, CNRS UMR 5554, Institut des Sciences de l'Evolution, Montpellier, France,* [3]*Dpto. de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal*

Although the mitochondrial genome is the most studied in biodiversity surveys, the relationship between its sequence variation and functional evolution, as well as its co-evolution with the nuclear genome, remain little documented and tested. We here take advantage of a peculiar biological situation to tackle this question. In northern but not in southern Iberian Peninsula, the Iberian hare, *Lepus granatensis*, harbors high frequencies of mtDNA from *Lepus timidus*. Since the latter is an arctic/boreal species and there is evidence in other species for a role of mtDNA in adaptation to cold, some functional differences may exist between the mtDNAs of the two species. We therefore ask whether the strong north-south differentiation for mtDNA in *L. granatensis* is paralleled by differentiation at some nuclear genes, as a result of either co-introgression with mtDNA or evolution after mtDNA introgression. If genes with such patterns were preferentially involved in interactions with mitochondrial genes, we would have a strong case for coevolution between the two genomes, for the functional role of mtDNA variation, and perhaps the adaptive nature of interspecific mtDNA introgression. We collected transcriptome sequencing data from 5 *L. granatensis* sampled in northern and 5 in southern populations, and also from *L. timidus*. Mapping of the *Illumina* sequence reads onto a de novo assembly of hares' transcripts was performed using *bwa*, and SNP and genotype calling was done using *samtools*. Preliminary analyses suggest that F_{ST} values between northern and southern samples based on over 3000 nuclear genes are generally very low, with an average close to zero. However, a small number of genes present high levels of geographic structure, and some of these have variation consistent with introgression from *L. timidus*. These results suggest that introgression of nuclear genes co-occurred with mtDNA but affected a small number of genes. Interestingly, a gene ontology analysis revealed a significant excess of functions related with mitochondria in the 1% of nuclear genes with the highest F_{ST} (irrespective of evidence for introgression) which may indicate adaptive co-evolution between the mitochondrial and nuclear genomes, accompanying introgression of mtDNA.

P-1085

Strong signatures of natural selection driving the evolution of *MC1R* in the Spanish population

Saioa López[1], Arrate Sevilla[1], Neskuts Izagirre[1], Conrado Martínez-Cárdenas[5], Gloria Ribas[2], Concepción de la Rúa[1], Isabel Smith[1], Montserrat Hervella[1], Alicia Galdeano[1], Mari Luz Cañavate[1], Oscar García[4], Jesús Gardeazabal[3], Santos Alonso[1]
[1]Universidad del País Vasco/Euskal Herriko Unibertsitatea, Vizcaya, Spain,[2]Fundación Hospital Clínico Universitario-INCLIVA, Valencia, Spain,[3]Hospital de Cruces, Vizcaya, Spain,[4]Área de Laboratorio Ertzaintza. Sección de genética forense, Vizcaya, Spain,[5]Fundación Hospital Provincial de Castellón, Castellón, Spain

MC1R is a highly polymorphic gene that regulates the production of melanin. Many of its variants have been correlated to a fair skin and red hair phenotype, thus providing a low tanning ability and a higher skin cancer risk. It is in European populations where it shows the highest variability, and two alternative hypotheses have been proposed to explain this high level of polymorphism: local selection for specific alleles or relaxation of functional constraints. The objective of this study was to determine which selective forces have shaped *MC1R* evolution in the Spanish population.

We analyzed 2146 human alleles from different regions of Spain through direct DNA sequencing of the coding region of *MC1R*. We detected 31 polymorphisms, 6 of which have not been previously reported, and we also found 2 new indels. The most frequent variant was V60L, consistent with other studies made in European populations. Neutrality tests were calculated with Zeng's DH software. p values for Tajima's D and E tests were 0.006 and 0.005, respectively. We performed simulations with Hudson's ms, and obtained a distribution of values of Tajima's D and E test with a modification of Zeng's DH software. The p values obtained confirmed the significance of the tests: <0.006 for Tajima's D and <0.003 for E test. The compound statistics DH, EW, HEW and DHEW, which and are relatively insensitive to background selection and demography and thus show higher specificity to positive selection, were also calculated, obtaining p values of 0.047, 0.005, 0.049 and 0.028 respectively. In order to explore for association to melanoma susceptibility we analyzed a second sample of 278 alleles from Spanish melanoma patients. We found 12 variants, 2 of which were monomorphic in the control population. The most frequent polymorphism was also V60L, but in a significantly higher frequency than in the healthy samples (0.201 vs. 0.152; p allelic exact test =0.035).

Our data showed significant departure from neutrality, so they robustly confirm the action of selection on *MC1R* in the Spanish population. This work also supports the association of the V60L variant with risk of melanoma, in accordance with some previous studies. The lower frequency of this mutation in healthy individuals might suggest a selective process favoring the non-risky variant.

P-1086

De novo analysis of the marine bivalve *Macoma balthica* transcriptome with the 454 technology

Vanessa Becquet[1], Audrey Rohfritch[2], Eric Pante[1], Pascale Garcia[1]
[1]*Laboratoire LIENSs UMR 7266 CNRS-Université de La Rochelle, La Rochelle, France,* [2]*CBPG, Montferrier sur lez, France*

The Baltic clam, *Macoma balthica* (L.) is a key species of intertidal mudflats commonly present in marine and estuarine soft-bottom habitats of the northern hemisphere. In Europe, it is currently widely distributed from the eastern Pechora Sea in Russia (northern range limit) to the Gironde estuary in France (southern range limit). Nevertheless, during the past five decades, the natural range of *Macoma balthica* has undergone a significant shift towards the northeast of the European coasts (Hummel et al., 2000). Abundant along the Atlantic coasts of the Iberian Peninsula more than 40 years ago (Otero and Milan, 1970), it has now completely disappeared from this area. This shift seems to be correlated with the increased sea surface temperature observed in the Bay of Biscay (France). We conducted a *de novo* sequencing of the transcriptome of *Macoma balthica* in search for the molecular signature of selection and local adaptation in this species. We analysed three natural populations sampled along the european coast in constrating environments: the Marennes-Oleron Bay in France (at the southern range limit of the species), the Gdansk Bay in Poland (a site that is highly impacted by anthropogenic activity) and the Somme Bay in France (considered as a reference population, as individuals were sampled in a natural reserve). A single 454 pyrosequencing run was conducted on these three isolated transcriptomes. 871,962 reads corresponding to 277 M bases were assembled with MIRA. Functionnal annotation resulted in the isolation of over 30,000 contigs. Among them, 1,000 genes implicated in stress response were characterized and differential gene expression was detected among sites. This preliminary study shows that genes implicated in response to heat (like Heat Schock Protein) seemed to be overexpressed at the southern range limit of the species whereas genes implicated in DNA repair and stress response were overexpressed in the highly impacted site. These results will be used in a qPCR approach to estimate the cumulative effects of pollution and temperature on cellular stress.

P-1087

Role of positive selection versus relaxed negative selection in genes showing human-specific evolutionary rate acceleration

Magdalena Gayà-Vidal[1], M. Mar Albà[1,2]
[1]*Evolutionary Genomics Group IMIM-UPF Research Programme on Biomedical Informatics. Barcelona Biomedical Research Park (PRBB), Barcelona, Catalonia, Spain,* [2]*Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Catalonia, Spain*

The rate of evolution of some protein-coding genes may present significant deviations (acceleration/deceleration) compared to the species clock due to functional and/or structural alterations. Lineage-specific gene evolutionary rate acceleration may be the result of positive selection or relaxed negative selection. Comparison of divergence and polymorphism data can help to distinguish between these two possibilities. The McDonald and Kreitman test (MK test) compares the numbers of non-synonymous and synonymous substitutions ratio (Dn and Ds, respectively) with the number of non-synonymous and synonymous polymorphisms (Pn and Ps, respectively). Dn/Ds>Pn/Ps will indicate fixation of advantageous mutations in the branches separating the two species. In this work, we first identify genes showing human branch specific acceleration - out of a set of 10,071 orthologs from human, macaque, mouse and cow - and then use available human polymorphism data to perform the MK test. A set of 620 out of the 10,071 genes presented acceleration in the human branch. In these genes, dN/dS was significantly larger in the human branch than in the ancestor primate branch using the Fisher test, an effect that was not observed when we compared the macaque and the ancestor primate branches. When dN/dS was significantly larger in the human branch than in both the ancestor primate and the macaque branches, and with no differences between macaque and the ancestor primate branches, the set of interest included 239 accelerated genes. Positive selection (PS) seems to have played an important role in the accelerated genes set compared to the control set according to preliminary results. The set of interest shows about twice higher fixation index (FI) than the control set. The direction of selection (DoS) indicates a general presence of PS (positive DoS value) in the set of interest, in contrast to the control set, which shows a negative value indicating negative selection. Likewise, the comparison of the two FI (or DoS) distributions (control set vs. set of interest) reveals that between 25 and 35% of the accelerated genes set are influenced by positive selection.

P-1088

Introgression of haplotypes with elevated mutation rate drives radiation of obligate asexuality in *Daphnia pulex*

Abraham Tucker, Matthew Ackerman, Brian Eads, Michael Lynch
Indiana University, Bloomington, IN, USA

Most lineages of the microcrustacean Daphnia pulex are cyclically parthenogenetic, alternating between sexual and asexual generations. However, obligate asexual lineages have arisen polyphyletically across North America, dominating many small ponds. Obligate asexual populations do not recombine and are thought to suffer from elevated deleterious mutation accumulation, since targets of positive and negative selection remain permanently linked (selective interference), reducing the effective population sizes of asexual lineages. High-throughput sequencing of 11 sexual and 11 asexual North American D. pulex isolates reveals 33,528 SNP markers shared by the obligate asexual phenotype, clustered on 13 large scaffolds (~ 15 Mb), most of which map to chromosomes 8 and 9. These results indicate that contemporary asexual lineages have not evolved independently, but have resulted from the spread of a common set of haplotypes that have introgressed into sexual populations, converting them to obligate asexuality. Asexual-linked haplotypes from 11 geographically distinct asexual populations have diverged from each other by only 9.1 x 10-5 changes per silent site, suggesting that the radiation of asexuality across the continent is quite recent. However, phylogenetic analysis indicates an ancient origin of the asexual-linked haplotype, which makes up roughly 20% of the genome of asexuals. The high divergence of the asexual-linked haplotype from all sexual haplotypes (> 0.04 per silent site) explains the prominent elevation of heterozygosity (~ 2-fold higher) in marker regions of asexuals, where the introgressed haplotype is paired with a random sexual background haplotype. While most asexual-linked SNP markers are concentrated on two linkage groups, more than 10 distinct regions of the genome appear linked to the asexual phenotype, suggesting that multiple loci are involved in the transition to obligate asexuality. Patterns of variation within protein-coding genes also indicate that, relative to the same genomic regions of sexual lineages, the introgressed haplotypes have a history of elevated mutation rate as well as relaxed selection against replacement substitutions, suggesting that obligate asexuals are less efficient at purging deleterious mutations, including mutator alleles. Molecular evolutionary analyses point to a number of specific genes likely to be involved in asexuality. While rates of amino acid-replacing substitutions are also elevated on the external branches of the asexual-linked haplotypes, the very young age of current asexual lineages limits our ability to measure the recent effects of deleterious mutation accumulation in contemporary obligate asexual genomes.

P-1089

Color vision variation of trichromacy by L/M hybrid genes in wild populations of New World howler monkeys

Yuka Matsushita[1], Hiroki Oota[1,2], Barbara Welker[3], Mary Pavelka[4], Shoji Kawamura[1]
[1]The Uniersity of Tokyo, Kashiwa, Japan, [2]Kitasato University, Sagamihara, Japan, [3]State University of New York at Geneseo, Geneseo, USA, [4]University of Calgary, Calgary, Canada

Most New World monkeys have highly polymorphic color vision achieved by allelic variation of the single-locus L/M opsin gene on the X chromosome. Among them, howler monkeys (*Alouatta*) are an exception having presumably a uniform trichromacy achieved by the L/M opsin gene duplication as in Old World primates (Old World monkeys, apes and humans). A previous *in vivo* (electroretinogram; ERG) measurement of howlers' visual sensitivity indicated that the sensitivity was explainable if the L and M opsins of howlers had similar spectral sensitivities with those of Old World primates, respectively, and both were expressed. Therefore, howler monkeys have been supposed to have same uniform trichromatic color vision as Old World primates and are important to study the evolutionary condition for trichromacy in primates. However, in fact nucleotide variation of the L and M opsin genes in wild howler populations has not been studied, and spectral sensitivities of howlers' L and M opsins have not been directly measured. To verify howlers' uniform trichromacy non-invasively, we examined fecal DNA samples of howlers (*A. palliata* and *A. pigra*) collected in Costa Rica, Nicaragua and Belize. We confirmed that the absorption spectra of L and M opsins are equivalent to those of Old World primates by reconstitution of these photopigments *in vitro*. Surprisingly, we found various types of L/M hybrid gene in the population samples. These hybrid L/M opsins showed the shift of absorption spectra from normal L and M opsins *in vitro*. This indicates that color vision of howlers should be polymorphic among at least trichromatic phenotypes, prompting us to reconsider an evolutionary significance of uniform trichromacy in primates.

P-1090

Mutations improving chromatic resolution of L/M opsin alleles in atelid New World monkeys

Shoji Kawamura[1], Yoshifumi Matsumoto[1], Yuka Matsushita[1], Norihiro Ozawa[1], Makiko Nakata[1], Satoshi Kasagi[1], Chihiro Hiramatsu[1,2]
[1]*University of Tokyo, Kashiwa, Chiba, Japan,* [2]*Kyoto University, Kyoto, Japan*

New World monkeys exhibit prominent variation in color vision. This polymorphism is due to allelic variation of the L/M opsin gene on the X chromosome. It has been shown that spectral variations of L/M opsins in primates result from substitutions at three amino acid sites: 180, 277 and 285 (the three-site rule). We previously reported, however, that two alleles of the L/M opsin gene of black-handed spider monkeys (*Ateles geoffroyi*) were the exceptions to the rule. The peak absorption spectra (λ_{max}) of the two allelic photopigments were predicted to be 560 and 552 nm based on their three site compositions SYT and SFT, respectively, while those of their reconstituted photopigments were measured to be 553 and 538 nm. Thus, the spectral difference between the two alleles expanded from 8 nm to 15 nm by some other mutations. In the present study, we identified these two alleles in two other atelid species, long-haired spider monkeys (*Ateles belzebuth*) and common woolly monkeys (*Lagothrix lagotricha*), indicating a common origin of the alleles in the two genera. Via mutagenesis, we identified mutations, in particular Y213D, occurred at amino acid sites conserved among primate L/M opsin genes, as playing major roles in expanding the spectral separation between the two alleles. These modifications appear to be adaptive, with heterozygous females having a better red-green chromatic resolution in detecting reddish or yellowish objects against background foliage.

P-1091

Relaxation of selective constraint in the human genome

Mehmet Somel, Emilia Huerta-Sanchez, Kirk Lohmueller, Melissa Wilson Sayres, Anna Ferrer-Admetlla, Matteo Fumagalli, Rori Rohlfs, Rasmus Nielsen
UC Berkeley, Berkeley, USA

It has been argued that ecological change and environmental niche construction could have relaxed selective constraint on specific physiological processes in human evolution, allowing accumulation of mutations at previously conserved sites. However, such shifts in selective pressure have not yet been studied at a genome-wide level. Here we developed a statistical approach to estimate change in selective constraint in the human lineage at each site, where we compare the selection coefficient estimated from the mammalian phylogeny, to that estimated from human population genetic data. Analyzing 54 individual genomes from the Complete Genomics deep sequencing dataset and 13 non-human eutherian genomes, we thus identified sites under negative selection among mammals, while exhibiting non-synonymous polymorphism in humans. We then sought for functional pathways showing high levels of such non-synonymous polymorphism at conserved sites, compared to other pathways. Consistent with previous work, we found that olfactory receptors display the highest levels of such polymorphism in humans. More interestingly, we identified a number of functional gene sets showing strong overall conservation in their protein sequence among non-human mammals, while exhibiting high levels of non-synonymous polymorphism in humans. Our results reveal unexpected functional consequences of relaxation of constraint in human evolution.

P-1092

Molecular basis of convergent adaptation to hydrogen sulfite extreme habitats in the Poecilia mexicana complex

Markus Pfenninger[1], Hannes Lerp[2], Elisabeth Funke[1], Patrick Slattery[2], Kaan Erkoc[2], Martin Plath[2]
[1]Biodiversity and Climate Research Centre by Senckenberg Naturforschende Geselleschaft and Goethe University, Frankfurt, Germany,[2]Evolutionary Ecology Group, Goethe University, Frankfurt, Germany

Adaptations to extreme environments have long since been in the focus of evolutionary biology. The ability of higher eukaryotes to live in hydrogen sulfite saturated environments is certainly one of the most spectacular examples of adaptations to extremely hostile environments.
In a comparative study, we present evidence for independent, convergent adaptation to such extreme habitats in several populations of the *Poecilia mexicana* (Atlantic molly) complex in Mexico. In combined approach of NGS, traditional Sanger sequencing and protein modeling of several mitochondrial genomes and nuclear candidate loci, we elucidate the molecular basis of these adaptations in the catalytic core of the COX-OXPHOS complex. We show that the fixation of the respective mutations was driven by positive selection. Different degrees of reproductive isolation among the respective population/species pairs associated with increasing secondary adaptations make this an excellent system to study ecological speciation over a steep environmental gradient.

P-1093

Inversion and nucleotide polymorphisms in Drosophila subobscura related with thermal adaptation

Olga Dolgova[1], Gemma Calabria[2], Marta Pascual[2], Joan Balanya[2], Mauro Santos[1]
[1]Universitat Autònoma de Barcelona, Barcelona, Spain, [2]Universitat de Barcelona, Barcelona, Spain

Drosophila subobscura is a native Paleartic species, which colonized the South and North Americas three decades ago. The compelling evidence that the latitudinal clines of inversion frequencies found along the distribution area of this species are adaptive (Prevosti et al. 1985, 1988) gave us a great opportunity to study thermal adaptation since temperature is the main selective factor causing the clines.

Recent work in our group has captured two components of the multivariate space linked to chromosomal inversion polymorphisms correlated with thermal adaptation in natural populations: thermal preference and heat tolerance (Rego et al. 2010, Dolgova et al. 2010).

Our aim was to understand the genetic basis of thermal adaptation in D. subobscura by means of the exploration of Palearctic populations along a latitudinal gradient.
To achieve our objectives, we focused on the chromosome O of D. subobscura, for which a balancer stock is available. Outbreed flies were collected from seven natural populations that are known to be representative for the Palearctic cline in chromosome arrangements in order to characterize their chromosomal inversion frequencies and nucleotide variation.

A physical mapping of 30 candidate genes that differentially respond to thermal adaptation in the laboratory was conducted for O chromosome. Nine of them are corresponded to the Segment I and 21 for Segment II.

We sequenced five candidate genes which were mapped inside and outside of the three most frequent gene arrangements in the O chromosome to elucidate the nucleotide variation along the cline and distiguish genètic content of diferent inversions. These gene arrangemens demonstrate the latitudinal clines in their frequency distributions along the Europe and appeared to be associated with behavioural and physiological thermoregulation in D. subobscura.

P-1094

Convergence of life-history traits and mutations roads in experimentally evolving populations of *Saccharomyces cerevisiae*

Aymé SPOR[1], Daniel KVITEK[2], Tibault NIDELET[1], Christine DILLMANN[1], Aurélie BOURGAIS[1], Dominique De VIENNE[1], Gavin SHERLOCK[2], Delphine SICARD[1]
[1]Université Paris Sud, Orsay, Essonne, France, [2]Stanford University, Stanford, California, USA

Phenotypic convergence is thought to occur because of natural selection in a given environment and/or because of physical and biological constraints limiting the number of possibilities. If natural selection was the only force responsible for convergent evolution, one would expect to detect convergent evolution in environments that are somewhat similar in many aspects (resource quality/quantity, seasonality, competition), *i.e.* similar ecological niches. By contrast if convergent evolution resulted from organisms structural constraints, convergent evolution would not need to be driven by the ecological niche. Using *Saccharomyces cerevisiae* as a model system; we investigate how much convergent evolution is specific to a given environment and how convergent evolution for a phenotypic trait may be constrained by the evolution of other correlated traits. By carrying 72 independent experimental evolutions with several yeast strains in four selection regimes, we found evidence for multiple life-history traits convergence. We sequenced the genome of evolved and ancestral strains. We found that in several strains and selection regimes, the same gene had been mutated, and additionally, some pathways were targeted by mutations multiple times independently. This reveals that some genotypic convergence underlies multiple traits convergence suggesting that there exists constraints on the adaptive landscape that require mutation of certain genes/pathways to improve fitness in these conditions, regardless of genetic backgrounds and environments. To our knowledge, this laboratory evolution experiment is the first to analyze convergence of multiple traits for several phenotypes/genotypes in several controlled environments, and to couple these observations to their underlying mutations.

P-1095

Molecular evolution and variation in regions of the *Drosophila melanogaster* genome that lack crossing over.

Jose L. Campos, Penelope R. Haddrill, Brian Charlesworth
Institute of Evolutionary Biology, Edinburgh, UK

The frequency of genetic recombination is expected to affect the efficiency of natural selection. This can be studied by comparing patterns of DNA sequence variability and evolution in regions with no crossing over versus regions where crossing over occurs. The recent annotation of more than 200 genes in *D. melanogaster* in the heterochromatin (where crossing over is absent) allows a better assessment of the importance of recombination than was possible before. We have previously shown that levels of protein sequence divergence and codon usage bias are consistent with selection being less efficient in the heterochromatin, although differences were observed among different non-crossover regions. We have extended our analyses of these genes by examining their polymorphism levels in an African population from the Drosophila Polymorphism Genome Project (DPGP). Average values for synonymous diversity were tenfold lower than the values for normally recombining regions. There was a more than a 2-fold increase in the ratio of replacement to silent polymorphism compared to normally recombining regions, and a more distorted site frequency spectrum (SFS). However, the patterns were different across the different non-crossover regions, suggesting that the efficiency of natural selection does indeed vary among them. Contrasting patterns of polymorphism and divergence also indicate lower levels of adaptive protein sequence evolution in the non-crossover regions. The results suggest that both positive selection and purifying selection are compromised when recombination is rare or absent, due to greater effects of Hill-Robertson interference compared with the rest of the genome.

P-1096

Sergey A. Naumenko and Alexey S. Kondrashov. Rate and breadth of protein evolution are only weakly correlated

Sergey Naumenko[1,2], Alexey Kondrashov[2]
[1]IITP RAS, Moscow, Russia,[2]Moscow State University, Moscow, Russia,[3]3Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, USA

Evolution at a protein site can be characterized from two different perspectives, by its rate and by the breadth of the set of acceptable amino acids. We used two datasets of protein-coding genes from 13 genomes of placental mammals and 11 genomes of species from genus Drosophila to study the correlation between these two characteristics. We reconstructed ancestral states using multiple alignments and maximum likelihood method implemented in PAML package and then calculated the number of amino acid replacements and the diversity of acceptable amino acids for each site. Using this data we show that there is a weak positive correlation between rates and breadths of evolution, both across individual amino acid sites and across proteins.

P-1097

Parallel reduction in *opsin* and *hedgehog* gene expression in independently derived cave populations of the amphipod *Gammarus minus*

Ariel Aspiras, Rashmi Prasad, Daniel Fong, David Angelini, David Carlini
American University, Washington, DC, USA

The subterranean realm is an excellent system for studying the adaptive and/or neutral evolution of regressive morphological traits at the molecular level. *Gammarus minus*, a freshwater amphipod living in the cave and surface streams in the eastern United States, is a premier candidate for studying the evolution of troglomorphic traits such as pigmentation loss, elongated limbs, and reduced or absent eyes. At present, relatively little is known about the evolution and development of troglomorphic traits in invertebrates. In *G. minus*, multiple pairs of genetically related, physically proximate cave and surface populations exist which exhibit an astounding degree of intraspecific morphological divergence. The morphology, ecology, and genetic structure of these sister populations is well characterized, yet the genetic basis of their morphological divergence remains unknown. In this study, we characterized nucleotide sequence variation in two paralogous *opsin* genes in individuals from multiple independently derived cave populations of the amphipod *Gammarus minus*. We also quantified gene expression of both *opsin* genes, as well as that of the eye development genes *hedgehog, pax6, sine oculis*, and *dachshund*. Low levels of nucleotide sequence variation was detected in both *opsin* genes, regardless of habitat, and were not suggestive of a relaxation of functional constraint in the cave populations with regressed or absent eyes. However, both *opsin* and *hedgehog* expression were significantly reduced in cave populations relative to their sister surface populations. No differences in the expression of the *pax6, sine oculis* or *dachshund* genes were detected. Since *hedgehog*-related genes are also involved in eye reduction in the Mexican cavefish *Astyanax mexicanus,* these genes may be common targets of evolution during cave adaptation, providing support for the hypothesis of genomic "hotspots" of evolution.

P-1098

Adaptive evolution of the Matrix Extracellular Phosphoglycoprotein in mammals

João Paulo Machado[1,2], Warren Johnson[3], Stephen O'Brien[3], Vítor Vasconcelos[1,4], Agostinho Antunes[1,3]
[1]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal, [2]Instituto de Ciências Biomédicas Abel Salazar (ICBAS), Universidade do Porto, Porto, Portugal, [3]Laboratory of Genomic Diversity, National Cancer Institute, Frederick, USA, [4]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal

The Matrix Extracellular Phosphoglycoprotein (MEPE) belongs to a family of small integrin-binding ligand N-linked glycoproteins (SIBLINGs) that play a key role in skeleton development, particularly in mineralization, phosphate regulation and osteogenesis. MEPE associated disorders cause various physiological effects, such as loss of bone mass, tumors and disruption of renal function (hypophosphatemia). The study of this developmental gene from an evolutionary perspective could provide valuable insights on the adaptive diversification of morphological phenotypes in vertebrates. Here we studied the adaptive evolution of the MEPE revealing that the majority of the coding region had a fast evolutionary rate, particularly within the largest exon (1467 bp). Rodentia and Scandentia had distinct substitution rates with an increased accumulation of both synonymous and non-synonymous mutations compared with other mammalian lineages. We identified 20 sites with significant positive selection signatures (codon and protein level) outside the main regulatory motifs (dentonin and ASARM) suggestive of an adaptive role. Conversely, we find three sites under selection in the signal peptide and one in the ASARM motif that were supported by at least one selection model. The MEPE protein shows a high number of selection signatures, revealing the crucial role of positive selection in the evolution of this SIBLING member. The selection signatures were found mainly outside the functional motifs, reinforcing the idea that other regions outside the dentonin and the ASARM might be crucial for the function of the protein and future studies should be undertaken to understand its importance.

P-1099

Drift may dominate the evolution of fidelity in unicellular eukaryotes with large genomes

Matthew Ackerman, Samuel Miller, Micheal Lynch
Indiana University, Bloomington, USA

Microbes with DNA-based genomes ≤4×10^7 bp experience roughly 1 mutation per 300 genome replications, regardless of the size of their genome. In contrast to this remarkably consistent rate, eukaryotes with large genomes experience between 0.84 (*A. thaliana*) and 75 (*H. sapiens*) mutations per generation. The consistency of microbial mutation-rates may be explained by a trade-off between the cost of mutations and the intrinsic cost of high-fidelity replication, by the rate that maximizes the speed of adaptation, or by the process of drift; it is not clear why eukaryotes with large genomes depart from microbial mutation patterns. However, only metazoans and metaphytes have been examined, confounding a number of traits (e.g. small population size, long generation times, and multicellularity) with large genome size, making it difficult to evaluate theories of mutation rate evolution.

In order to address the phylogenetically poor sampling of eukaryotes, we performed a mutation accumulation (MA) experiment with *Chlamydomonas reinhardtii,* a unicellular algae separated from land plants by ~1 billion years and with a GC rich genome of 1.2×10^8 bp. Typical eukaryotic mutation rates vary between ~5×10^{-10} and 5×10^{-8} mutations per base pair per generation; rates of transitions are nearly universally higher than rates of transversions, and mutations occur at GC sites more frequently than at AT sites—leading to an AT bias. However, the rate and spectrum of mutation in *C. reinhardtii* is not typical. Based on whole genome re-sequencing of four MA lines maintained for 1600 generations, the inferred rate of mutation is surprisingly low, 5.13×10^{-11}, consistent with the low mutation rates recently described in *Paramecium* (Sung et al. submitted for publication). While *Chlamydomonas reinhardtii* has a surprisingly low mutation rate when compared to other organisms with 1.2×10^8 bp genomes, it experiences 2 mutations per 300 genome replications, only twice the rate seen in other DNA-based microbes. In addition, the spectrum of mutation is unique: AT:GC transitions occurred at rates similar to GC:AT transitions, but AT:CG transversions occurred at a higher rate than either transition—generating a GC bias. As a result, AT sites become GC sites at 2.5 times the rate that GC sites become AT sites. *C. reinhardtii* genome's GC content of 66.8% is close to the mutation equilibrium of ~70% GC. These findings are consistent with a drift-barier theory of mutation rate evolution, but the implications of these observations for other theories of mutation rate evolution are also reviewed.

P-1100

Adaptation in the genic control regions of murids

Daniel Halligan[1], Athanasios Kousathanas[1], Bettina Harr[2], Thomas Keane[3], David Adams[3], Peter Keightley[1]
[1]University of Edinburgh, Edinburgh, UK, [2]Max-Planck-Institute for Evolutionary Biology, Ploen, Germany, [3]Wellcome Trust Saner Institute, Hinxton, UK

Few attempts have been made to quantify selection and adaptation in mammalian noncoding DNA. Furthermore, the extent of adaptation attributable to control regions both upstream and downstream of protein coding genes is currently unknown. Using whole genome polymorphism data from ten wild caught mice (Mus musculus castaneus) and divergence to a close relative (rat), we investigate the extent of purifying selection and the rate of adaptive evolution in the sequences immediately flanking the translation start/stop codons of protein-coding genes. Inferred adaptive substitutions in these regions can then be partitioned into those occurring within and outside of UTR regions. We investigate the relative rates of adaptation observed in these regions between the X chromosome and autosomes, and in different functional classes of genes.

P-1101

Profiling diverse levels of natural selection across regions of coding gene sequences

Ning Li[1], Zhang Zhang[2], Jeffrey Townsend[1]
[1]*Yale University, New Haven, CT, USA,* [2]*Beijing Institute of Genomics, Beijing, China*

One of the major barriers to identifying the relative roles of natural selection and genetic drift in the evolution of gene sequence has been the inability of established methodologies to overcome issues of statistical power in order to accurately and precisely characterize the level of selection on regions and sites within gene sequences. One of the most effective approaches to detect a history of natural selection is based on contrasting patterns of polymorphism and divergence in DNA sequences. By comparing putatively neutral polymorphism to levels of observed divergence in sequence, the gene-wide McDonald-Kreitman test and homogeneous Poisson Random Field models have identified numerous gene-wide targets of strong, diversifying selection that play novel and important roles in ecological and evolutionary processes. However, genome-wide scans of polymorphism and divergence in synonymous and replacement sites within genes have yielded surprisingly few genes whose tallies reject a null hypothesis, even in organisms with very high effective population sizes such as the wine yeast Saccharomyces cerevisiae. The surprising dearth may arise because targets of selection are often likely to be restricted to a single functional domain within a protein. Unfortunately, most data sets lack sample size under previous methodologies to effectively detect selection at these levels, leading to a tyranny of the null hypothesis. Additionally, local variation in mutation rate of synonymous sites presents a challenge for fine-scale detection of a history of natural selection. To address these challenges, we have developed an approach that performs model-averaged clustering of intragenic polymorphism and divergence. Applying Poisson Random Field theory to the model-averaged polymorphism and divergence over all potential clusters, we graphically profile the model-averaged level of selection and its 95% confidence intervals for each site. Our analysis of individual genes such as Cdc6 demonstrates the action of natural selection on a much finer scale than previous approaches. Our genomic analysis of genes across the S. cerevisiae genome demonstrates that although much of the coding sequence is under purifying rather than adaptive natural selection, nearly 50% of coding sequence overall is under an adaptive regime of selection. Moreover, adaptively evolving sites are distributed across about 80% of the genes in the whole genome. The presence of extensive regions of purifying selection explains that few genes are significant under traditional tests of polymorphism and divergence, even though many critical sites or regions within individual genes are playing important adaptive roles in an evolutionary history.

P-1102

A mathematical approach to study the effects of the distribution of selection coefficients associate with allele replacements at different sites on the overall rate of evolution

Alice Ledda[1], Fyodor Kondrashov[2,3]
[1]INSERM, Paris, France,[2]CRG, Barcelona, Spain,[3]ICREA, Barcelona, Spain

The ratio of nonsynonymous to synomymous evolution (dn/ds) is often used as a measure of the strength of selection. However, there are many different selective regimes that can lead to the same observed dn/ds. For example, a dn/ds ratio of 0.1 can be just as easily explained by 10% neutral nonsynonymous sites with the remaining 90% of sites being under strong negative selection, or with a uniform weak negative selection across these sites that result in the rate of evolution, R, being equal to 10% of the neutral rate. When a dn/ds ratio is measured for a specific protein coding sequence, are any generalizations that are possible about the underlying distribution of average selection pressures in that sequence? Here we use a mathematical approach to study the effects of the distribution of selection coefficients associate with allele replacements at different sites in the sequence on the overall rate of evolution of the sequence. We solve for the expected relative rate of evolution of two different cases. The first case is when the fitness of each possible allele replacement is drawn from a gaussian distribution of selection coefficients with an average of X. The second, where the fitness of each possible allele replacement is equal to X. For the case of two alleles per locus we show analytically that R(E(X)) <= R(s=X). For cases with more than two alleles per locus we show the same expected outcome computationally. We show that the difference between R(E(X)) and R(s=X) increases with the number of alleles per locus. We also show that the rate of evolution is influenced by the variance of the mutation rates distribution. Thus, for a specific dn/ds in a sequence it is possible to estimate the upper bound of the average selection coefficient associated with allele replacements in that sequence.

P-1103

Analysis of sites of polymorphic deletions in noncoding regions of D.melanogaster genome reveals pervasive epistatic selection

Albert Khajrullin[1], Evgeny Leushkin[1,3], Georgii Bazykin[1,3], Alexey Kondrashov[1,2]
[1]Moscow State University, Moscow, Russia,[2]University of Michigan, Michigan, USA,[3]Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia

Numerous observations suggest that selection is often epistatic, so that evolution at one particular locus is highly dependent on the rest of the genome. Here, we use whole-genome data on deletion polymorphisms, and on single-nucleotide polymorphisms (SNPs) contained within DNA segments of polymorphic deletions, to estimate the extent of epistatic selection. Allele frequency spectra indicate that selection against deletion increases with its length, which is not surprising, as a longer deletion is more likely to overlap a functional region. However, an average deletion is much less deleterious than expected from the SNP data, and this contrast is most pronounced for longer deletions. Thus, neighboring SNPs interact epistatically, and this epistasis is antagonistic. Evolution of nucleotides within the segments of polymorphic deletions and in their closest vicinity approaches selective neutrality as the deletion allele frequency increases, suggesting that, in addition to weaker negative selection, nucleotides contained within more frequent deletions also experience weaker epistatic interactions.

P-1104

Polymorphism, recombination and selection efficiency along the complete genome of Drosophila Melanogaster.

Maite Garazi Barron Aduriz[1,2], David Castellano Esteve[1,2], Miquel Ramia Jesus[1,2], Antonio Barbadilla Prados[1,2]
[1]Universitat Autonoma de Barcelona, UAB, Barcelona, Spain,[2]Institut de Biotecnologia i Biomedicina, IBB, Barcelona, Spain

Whether diversity in the genomes are shaped mainly by neutral processes or selection is a recurrent theme in population genetics. Under neutrality, no relationship between levels of polymorphism and recombination is expected. However, the correlation between polymorphism and recombination appears to be one of the most consistent patterns of population genetics, with similar relationships found in every species examined. The neutral model predicts that if recombination itself is mutagenic then high recombination regions will exhibit higher levels of both polymorphism and divergence. The alternative hypothesis is that some form of linked selection is acting across the genome such that loci within higher recombination are more likely to escape from the effects of nearby selection, whether advantageous or deleterious.
We used the 158 DGRP Illumina sequence data and genome sequences from D. simulans and D. yakuba to perform genome wide analyses of polymorphism and divergence and assessed the association of these parameters with the recombination landscape on a much larger scale than had been previously done. We used two recombination estimates, recombination map estimates (in cM/Mb units) and the population recombination rate (4Ner) based on linkage disequilibrium in 50kbp non-overlapping windows.
Regions whose recombination is < 2cM/Mb there is a positive correlation between polymorphism and recombination rate, while divergence levels remain unaffected. Above this threshold, recombination and polymorphism behaves independently. would expect to have a stronger correlation between polymorphism and recombination when selection efficiency increases. However, we find a greater adaptive propensity in genes whose recombination context is > 2cM/Mb. We explore this apparent contradiction establishing a specific threshold for each chromosome arm and analyzing different selection regimes above and below these thresholds. While in autosomes less than 1/3 of the chromosome is above this threshold, it spans more than 2/3 on the X chromosome.
We hypothesize that in genomics regions where polymorphism and recombination are uncorrelated, sites behaves as effectively independent between them. Thus, in those regions, the expectations of Neutral Theory are fulfilled even though the efficiency of selection is higher. Below this threshold, sites are linked and hitchhiking and/or background selection can reduce levels of polymorphism by an amount proportional to the strength of selection and the recombination rate. The independence between sites is achieve depending on the interactions of levels of polymorphism and recombination rate. These results throw new insight in the relationship between demographic changes and natural selection as well as in sex chromosome evolution.

P-1105

Comparative genomics of *Saccharomyces cerevisiae* isolates sheds light on the role of neutrality and selection in the emergence of adaptations

Christina Toft[1], Clara Ibañez[2], Amparo Querol[2], Eladio Barrio[1]
[1]*University of Valencia, Valencia, Spain,* [2]*Consejo Superior de Investigaciones Científicas, Valencia, Spain*

The mechanisms of natural selection in mediating the origination of novel ecological adaptations remain elusive. The reason underlying our inability to understand how selection operates is that analyses and experiments have been mostly driven by data, in which rapid periods of episodic selection are always obscured by large periods of evolutionary stasis. The study of some evolutionary phenomena, such as gene and genome duplication, has revolutionized our understanding of the mechanisms of natural selection. The Baker's yeast *Saccharomyces cerevisiae* has undergone whole-genome duplication 100 MYA, and a number of studies have explored how selection has operated on the enormous amount of genetic material generated. Our knowledge on the precise mutational dynamics that originate adaptations remains, however, highly fragmented. In this study we explored the selective constraints that have been operating on 26 *S. cerevisiae* genomes isolated from different environments, including wine yeast, brewing yeast, natural isolates, biotechnological environments, pathogenic yeast and yeast from the laboratory. Genome-wide exploration of the mutational dynamics reveals complex patterns of selection and coadaptation. We explored how these patterns may be related to the environmental conditions of yeast and the relationship between gene duplication and the emergence of ecological novelties.

P-1106

Divergent mitochondrial haplotypes convey clearadjustments in metabolic phenotypes.

Pierre Blier[1], Nicolas Pichaud[1,2], Robert M Tanguay[1,3], J William O Ballard[1,2]
[1]Université du Québec, Rimouski, Qc, Canada,[2]University of New South Wales, Sydney, Australia,[3]Université laval, Québec, Qc, Canada

Linking the mitochondrial genotype and the organismal phenotype is of paramount importance in evolution of mitochondria. In this study, we determined the differences in catalytic properties of mitochondria dictated by divergences in the siI and siIII haplogroups of Drosophila simulans using introgressions of siI mtDNA type into the siIII nuclear background. Our results showed that the catalytic properties of the electron transport system are not impaired by introgressions, suggesting that the observed divergences in mitochondrial functions are mainly confered by mtDNA differences and not of nuclear DNA or mito-nuclear interactions. This is the first study in our knowledge that make a clear demonstration that one haplogroup observed in a natural population can confer phenotypic divergences in terms of functional properties of the mitochondrial metabolism.

P-1107

Drift constrained by selection: the dynamics of small RNA copy number and expression patterns over evolutionary timescales.

Anthony Poole[1], Marc Hoeppner[2]
[1]*University of Canterbury, Christchurch, New Zealand,* [2]*Uppsala University, Uppsala, Sweden*

Small RNAs are often found in multiple copies in eukaryote genomes. Such copies may be the result of functional diversification, or they may be functionally redundant and short-lived on evolutionary timescales. We analysed the evolutionary stability of multicopy, essential mammalian RNAs and compared their expression profiles to test whether functional redundancy or diversification best explain their evolution. Under the former model, the expectation is that copies are equally expressed across tissues and subject to high turn-over. The latter model, in contrast, predicts that sub- or neofunctionalization following duplication may lead to a range of complementary expression profiles across tissues. Detailed study of spliceosomal snRNA U1 and snoRNA U3, both of which are found in dozens of copies across mammalian genomes, indicates that few loci are stable over the course of mammalian evolution. Interestingly, the most deeply conserved loci are located within the introns of a testis-expressed gene, suggesting a possible driving force for copy number proliferation through retrotransposition early in development. Analysis of human expression data indicates the majority of copies show little or no expression, with the remainder showing indistinguishable and ubiquitous expression profiles, suggesting that these high copy number RNAs are redundant.

P-1108

Opsin Phylogenetics Iluminates the Evolution of Colour Vision in Mammals

Bruno Simoes[1], Huabin Zhao[2], Shuyi Zhang[2], Emma Teeling[1]
[1]UCD School of Biology and Environmental Science, University College Dublin, Dublin, Ireland,[2]School of Life Science, East China Normal University, China, Shanghai, China,[3]School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

Mammals have successfully colonized a vast range of ecological niches and have highly developed sensory capabilities, ranging from a wide olfactory repertoire to echolocation. Vision provides information fundamental for the survival of the almost 5000 mammal species that currently exist. Studies of mammalian colour vision have linked changes in opsin genes to differences in evolutionary history, ecology, as well as other sensory capabilities. With the aim of establishing the evolution and functionality of mammalian photopigments, we sequenced and gathered a molecular dataset that included 207 short-wavelength opsin genes (SWS1) and 85 medium-to-long wavelength opsin genes (MWS/LWS) from across the major mammalian groups spanning 80 million years of mammal evolution. We focused particularly on bats due to their great level of sensory specialization and wide diversity of ecological niches inhabited. These molecular data show the SWS1 opsin has undergone multiple independent losses in many mammalian orders, but MWS/LWS is conserved and functional throughout mammalian evolution. The ancestor of all mammals had a 555nm sensitive MWS/LWS whereas the SWS1 was UV sensitive in the ancestor of all mammals, Laurasiatheria, Afrotheria and Euarchontoglires. The SWS1 is UV sensitive in most bats, however in some Yinpterochiropteran and Yangochiropteran lineages it has undergone mutations, leading to a loss-of-function, which are associated with evolutionary events in their natural history and roosting. This presentation concerns the evolution of opsins among mammals and the genetic consequences of evolutionary history, echolocation, nocturnality, and ecological niche specialization on colour vision.

P-1109

Pervasive epistatic interactions between nearby sites in coding and non-coding sequences of *D. melanogaster*

Vladimir Seplyarskiy[1,2], Alex Kondrashov[1,3], Egor Bazykin[1,2]
[1]Moscow State University, Moscow, Russia,[2]IPPI, RAS, Moscow, Russia,[3]University of Michigan, Michigan, USA

Most models of sequence evolution assume independence of mutations at different sites; however, there are multiple possible reasons for deviations from independence. In particular, epistatic interactions between sites may lead to non-independence of SNPs and of nucleotide replacements at distinct sites. Here, we show that epistatic interactions between nearby (at distances of up to 5 bp) nucleotide sites shape the patterns of polymorphism and divergence in coding (CDS) and noncoding (NDS) DNA sequences of *D. melanogaster*. Both in within-species polymorphism and in interspecies divergence, the clustering of SNPs and substitutions at nearby sites is positively correlated with the degree of conservatism of the sequence segment. Moreover, clustering of substitutions in the same of the two evolving lineages, suggestive of positive selection or epistatic interactions (Bazykin et al. Nature 2004 Jun 3; 429:558-562), is twice as strong in the conserved regions of NDS, and 3 times as strong in the conserved regions of CDS, compared to the corresponding non-conserved regions. These results show that epistatic interactions between nearby CDS and NDS sites are prevalent genome-wise, and may be an underappreciated contributor to sequence evolution.

P-1110

Role of *MT3* gene in metalophyte plant, *S. vulgaris*

Eva NEVRTALOVA, Roman HOBZA, Jiri BALOUN, Vojtech HUDZIECZEK, Radim Cegan
Institute of Biophysics, Brno, Czech Republic

Many *Silene* species present a valuable model of metallophyte plants. *S. vulgaris* is a traditional plant of choice to study aspects of plant tolerance to various heavy metals. Recent results suggest that copper tolerance in *S. vulgaris* is governed by a single dominant gene with minor effect of a number of modifiers. Other studies propose association of copper tolerance with expression of a specific metallothionein gene. Until recently, no large scale genomic resources were available for this species.

We have employed massive transcriptome sequencing by Illumina technology to establish transcriptome libraries of two *S. vulgaris* populations: LH metalliferous population (plants collected in Cu-contamined site in Lubietova Halda, Slovakia) and SS non-metalliferous population (plants collected in Stranska Skala, Brno, Czech Republic). Detailed comparative analysis of candidate genes between populations and tissues (+/- copper) revealed candidates for copper tolerance in metalliferous *S. vulgaris* population.

Further, we have identified and characterized genomic locus containing *MT3* gene in *S. vulgaris*. Our data show that *MT3* gene is locally duplicated in all studied *S. vulgaris* accessions. Both paralogues present functional copy of a *MT3* gene based on complementation analysis with yeast mutants. *In silico* expression analysis showed that *MT3* gene increases transcription level under copper stress condition. By Q-PCR, we observed different *MT3* expression regulation after Cu treatment in studied populations. We show that down-regulation of *MT3* expression in higher copper concentration might be a key characteristic of *S. vulgaris* copper tolerant plants. Our data evoke that duplication of *MT3* gene is a species specific trait which makes *S. vulgaris* predisposed to tolerate elevated copper concentration in soil. However, tolerance to high copper content differs among populations and it correlates with expression level control of *MT3* gene.

Acknowledgement: The work was supported by the Czech Science Foundation grant 522/09/0083 and Grant Agency of the Czech Republic P501/12/G090

P-1111

Selection against loss-of-function alleles in Drosophila melanogaster populations

Aleksandra I. Akhmadullina, Georgii A. Bazykin, Aleksey S. Kondrashov
Moscow State University, Moscow, Russia

An average individual in human and Drosophila populations carries ~100 loss-of-function alleles of protein-coding genes. Only a small fraction of these alleles have severe impact on fitness even when homozygous, and most of them are under only a weak negative selection. In this work we studied complete genomes of African and North American populations of D. melanogaster to examine the frequency of accumulating nonsense mutations. We found that 10% of gene D. melanogaster contain stop codons, which reach high frequencies in the population. Indeed, after a locus accepts a loss-of-function mutations, all other mutations in it become selectively neutral and start accumulating freely. Nonsense alleles carry 0.0014 allele specific missense mutations, and 0.00098 synonymous mutations, indicating free accumulation of missense mutations. Here, we estimate the arithmetic mean persistence time, and a harmonic mean selection coefficient, of a nonsense allele in African and North American populations of D. melanogaster, using negative epistasis between a loss-of-function and any other non-neutral mutation within a locus.

P-1112

Adaptive evolution and positive selection in rhodopsin: the dN/dS debate

D. David Yu[1], Belinda S.-W. Chang[1,2]
[1]Dept. of Cell & Systems Biology, University of Toronto, Toronto, Ontario, Canada, [2]Dept. of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada

The relative contributions of positive selection and genetic drift to evolutionary divergence is one of the central unresolved debates in molecular evolution. Positive selection at specific amino acid sites can be detected by comparing the ratio of nonsynonymous and synonymous substitutions (dN/dS) using likelihood codon models. Despite the widespread use of dN/dS-based methods, experimental validations have been lacking, with the majority of claims of positive selection based on statistical analyses alone.
Rhodopsin is the dim-light sensor in vertebrates whose wavelength of maximum absorbance (λmax) has been shown evolve adaptively to an organism's spectral environment. Two recent studies on rhodopsin have cast doubt on the validity of dN/dS-based methods by finding no correspondence between dN/dS-based site predictions and amino acids known to adaptively shift λmax.
We have performed dN/dS analyses on two large teleost rhodopsin datasets (>700 taxa combined). This allowed us to identify a number of sites that are robustly-predicted to be undergoing positive selection. Positively-selected sites were found to be non-randomly distributed in rhodopsin's tertiary structure, with a large subset located on the membrane-facing portions of transmembrane helices 4, 5, and 6. Site-directed mutagenesis was used to generate natural variants at positively-selected sites. Variants at positively-selected sites alter a number of different aspects of rhodopsin function (including λmax). The implications of our findings towards the current debate over dN/dS-based methods will be discussed.

P-1113

Molecular evolution of Rhodopsin in Neotropical cichlids: Differences in selective constraint between African and Neotropical dim-light visual pigments.

Shannon Refvik[1], Hernán López-Fernández[1,2], Belinda Chang[1]
[1]*University of Toronto, Toronto, Ontario, Canada,* [2]*Royal Ontario Museum, Toronto, Ontario, Canada*

The cichlid fishes of the African rift lakes are model systems of adaptive radiation in vertebrates, and properties of their visual systems have been implicated in their diversification and speciation. Neotropical cichlids are their most closely related relatives, and are also characterized by diversity in life history, feeding mode, and habitat. However, relatively little is known of their visual systems. We sequenced the Rhodopsin gene for 31 species of Neotropical cichlids, focusing on the speciose and morphologically diverse Geophagini clade. With this data set, we provide the first comparative study of a visual protein between the well-studied African clade and their closest sister group. Using random site, branch site, and clade based likelihood codon models of evolution, we investigated differences in selective constraint between African and Neotropical cichlid Rhodopsins. Consistent with previous studies, we found evidence for positive selection in African rift lake cichlid Rhodopsin. Using clade models on a combined alignment as well as random sites models on separate African and Neotropical datasets, we also found strong evidence for positive selection in Neotropical cichlid Rhodopsins. This selection is associated with different amino acid sites than in the African group. Based on prior studies of Rhodopsin structure and function, we hypothesize that substitutions at these sites may be influencing non-spectral properties of Rhodopsin function.

P-1114

Systematic error in seed plant phylogenomics

Bojian Zhong[1,2], Oliver Deusch[1], Vadim V Goremykin[3], David Penny[1], Peter Lockhart[1]
[1]Institute of Molecular BioSciences, Palmerston North, New Zealand,[2]Allan Wilson Centre for Molecular Ecology and Evolution, Palmerston North, New Zealand,[3]Istituto Agrario San Michele all'Adige Research Center, San Michele all'Adige, Italy

Resolving the closest relatives of Gnetales has been an enigmatic problem in seed plant phylogeny. The problem is known to be difficult because of the extent of divergence between this diverse group of gymnosperms and their closest phylogenetic relatives. Here we investigate the evolutionary properties of conifer chloroplast DNA sequences. To improve taxon sampling of Cupressophyta (non-Pinaceae conifers) we report sequences from three new chloroplast (cp) genomes of Southern Hemisphere conifers. We have applied a site pattern sorting criterion to study compositional heterogeneity, heterotachy and the fit of conifer chloroplast genome sequences to a GTR + G substitution model. We show that non-time reversible properties of aligned sequence positions in the chloroplast genomes of Gnetales mislead phylogenetic reconstruction of these seed plants. When 2250 of the most varied sites in our concatenated alignment are excluded, phylogenetic analyses favour a close evolutionary relationship between the Gnetales and Pinaceae – the Gnepine hypothesis. Our analytical protocol provides a useful approach for evaluating the robustness of phylogenomic inferences. Our findings highlight the importance of goodness of fit between substitution model and data for understanding seed plant phylogeny.

P-1115

The Puerto Rican parrot genome project: an international effort supported by the local community

Taras K. Oleksyk[1], Jean-François Pombert[2], Daniel Siu[3], Anyi Mazo-Vargas[1], Brian Ramos[1], Wilfried Guiblet[1], Yashira Afanador[1], Christina T. Ruiz-Rodriguez[4], Michael L. Nickerson[4], Michael Dean[4], David Logue[1], Luis Figueroa[5], Ricardo Valentin[6], Juan L. Rodriguez-Flores[7], Steven E. Massey[8], Juan Carlos Martinez-Cruzado[1]
[1]UPRM, Mayaguez, Puerto Rico,[2]UBC, Vancouver, Canada,[3]Axeq Technologies, Rockville, MD, USA,[4]NCI-Frederick, Frederick, MD, USA,[5]Compania De Parques Nacionales, San Juan, Puerto Rico,[6]DNER, San Juan, Puerto Rico, [7]Cornell University, New York, NY, USA,[8]UPRRP, Rio Riedras, Puerto Rico

Amazona vittata, the only surviving native parrot in the U.S. and a beloved mascot of Puerto Rico, is a critically endangered species. By the late 20th century, human activity had reduced the once global population of Puerto Rican Parrots to 16 individuals in the El Yunque National Forest. While the *A. vittata* population is slowly recovering due to a successful captive breeding program, it is unclear how much genetic variation is left in the species and what implications it would have for current recovery efforts. The local community is highly concerned about the danger of losing its endemic species and strongly supports the ongoing efforts towards its conservation. We initiated the Puerto Rican parrot genome project as the first genome project fully supported by donations from the Puerto Rican community. This project is also the first example of the direct public involvement into the sequencing of the genome an endangered animal. Genomic DNA from one *A. vittata* female (Rio Abajo breeding facility, Puerto Rico) was sequenced using the Illumina HiSeq platform to an average coverage depth of 26.9X (based on a predicted size of 1.58 Gb) and assembled *de novo* using Ray. Early assemblies with N50s of 6,983 and 19,470 bp for the contigs and scaffolds, respectively, along with their validation show promising results. Additional public fundraising efforts are on the way to increase the depth of coverage, improve the assemblies, and annotate the *A. vittata* genome. The initial data has been deposited in public databases and is also freely accessible via our genome browser (http://genomes.uprm.edu). The early annotation efforts displayed online include gene models, ESTs, and SNPs. The browser also contains a complete physical map of the species' mitochondrial genome. The knowledge acquired from the analysis of this genome, and its comparison with genomes of other Amazon parrots, will contribute to the conservation efforts leading to the species' recovery.

P-1116

NIN, a gene necessary for maintaining neurogenic divisions of radial glial cells, evolved adaptively in anthropoid primates

Stephen Montgomery[1,2], Nicholas Mundy[1]
[1]*University of Cambridge, Cambridge, UK,* [2]*University College London, London, UK*

Until recently a long held dogma in comparative neurobiology has been that the number of neurons under any given area of cortical surface was constant. As such, the attention of those seeking to understand the genetic basis of brain evolution has focused on genes with possible functions in the lateral expansion of the cortex. However, new data from primates suggests cortical cytoarchitecture is not constant across species, raising the possibility that evolutionary changes in the radial development of the cortex may have played an important role during primate evolution. Here we present the first analysis of a gene with a function of possible relevance to this dimension of brain evolution. We utilized genomic data to inform targeted re-sequencing of partial coding sequence of *NIN*, a gene encoding a protein necessary for maintaining asymmetric and neurogenic divisions of radial glial cells, across a diverse range of 22 anthropoid primates. We show that *NIN* evolved under positive selection during anthropoid primate evolution. Using phylogenetic comparative analysis we explore how this selection may relate to neurological phenotypes including brain mass, neuron number and the variation in neuron number not explained by variation in cortical surface area. The results suggest a role for *NIN* in the evolution of radial cortical development.

P-1117

Phylogenomics on African lake cichlids: the evolution of transcriptomes during explosive speciation

Zuzana Musilova, Walter Salzburger
University of Basel, Basel, Switzerland

The East African Great Lakes (Malawi, Tanganyika and Victoria) host several hundreds of recently evolved cichlid species, which form the most impressive examples of adaptive radiation and rapid organismal diversification. These species flocks exhibit an exceptionally high proportion of endemism and not a single cichlid species is shared between the three Great Lakes. The triggers of the cichlid radiations, and the genetic or genomic basis of their evolutionary success, are largely unknown, however. Here, we apply next-generation sequencing techniques on genomes and transcriptomes from a phylogenetic representative set of East African cichlid species in order to pursue genome and transcriptome evolution during explosive speciation. Our data, combined with the five available whole-genome sequences, provide new insights into the molecular evolution of a rapidly evolving clade and shed light on some of the molecular factors that lead to accelerated rates of speciation. We thus present the first broad phylogenomic study focused mainly on the evolutionary history of an adaptive radiation.

P-1118

Ensembl Quality Database

Michal Kabza, Izabela Makalowska
Adam Mickiewicz University, Poznan, Poland

Ensembl project [1], since its first release in 2000, has become an important source of data for multiple experimental and computational genome-scale analyses, providing the rich set of annotations, such as genes, transcripts, proteins, homology relations and data regarding syntenic regions. In this way, Ensembl is extremely useful for many types of analyses. It does, however, come with a unique set of challenges. Many genome annotations, particularly these of low-coverage genomes, contain multiple artifacts such as very short exons and introns or protein coding genes with CDS lacking proper start and stop codons. Those artifacts are the result of factors like Ensembl pipeline's bias towards specificity or paucity of organism-specific biological sequences. In consequence, it is usually necessary to filter the data before conducting analysis to obtain biologically meaningful results. This need has been recently recognized in regards to estimating isoforms expression from RNA-Seq data [2]. Simple filtering can be achieved by using BioMart, but more complicated queries still require a lot of work. This is the reason why we developed the Ensembl Quality Database. For now the database is primarily focused on genomic features (i.e. genes, transcripts and exons) and contains information regarding annotation quality as well as some additional data like GC content. However, we are planning to broaden the extent of data covered in the database by including information stored in other databases (UCSC Known Genes or NCBI Gene) and results of additional local analyses. Important hallmark of Ensembl Quality Database is it's query language and data distribution system. Contrary to most databases, in case of which data is stored on a server and accessed via web interface, we decided to implement query tool as a domain-specific language, written in Python. This allowed us to create a tool that is much more extendable and enables users to perform the most complicated queries easily. Database data for each organism is stored in separate file that needs to be downloaded from our server, similarly to the data policy adopted in many Biodonductor [3] packages.

References:
1. Hubbard T. et al.: The Ensembl genome database project. Nucleic Acids Res. 2002 Jan 1;30(1):38-41.
2. Garber M. et al.: Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011 Jun;8(6):469-77.
3. Gentleman R.C. et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

P-1119

Genetic diversity of the dinoflagellates (Alveolata: Dinophyceae) from the ancient lake Baikal

Natalia Annenkova[1,2]
[1]*Limnological Institute, Siberian Division of Russian Academy of Science, Irkutsk, Russia,* [2]*Lund University, Lund, Sweden*

Single-celled eukaryotes (protists) play a key role in the biosphere because they are numerous and carry out almost all ecological functions. However, many questions concerning their evolution, biogeography and even real number of species (from 90,000 to 300,000) stay unresolved due to undersampling (protists have not been searched for thoroughly in many habitats, especially, in freshwaters) and difficulties in distinguishing species of similar morphology without DNA analysis. In particular, there are only handful studies of protists from ancient lakes.

The lake Baikal is the oldest (~25 million years) and the deepest lake in the world. There are more than 2500 animal species in it including both ancient polyphyletic groups and young monophyletic taxons, within which species still can not be divided for sure because they originated from rapid radiation. Baikalian protists are almost undescribed. Dinoflagellates are the protists closely related to endoparasites from Apicomplexa. They occupy every major ecological niche in water environments as primary producers, predators, free-living, symbionts, parasites and have huge genome (3000-215000 Mb). By 2008 more than 1700 species of free-living marine dinoflagellates and only 220 of freshwater species were described. There are no detailed studies of the dinoflagellates in Baikal, though they are members of its plankton community. Current study represents the first molecular genetic description of the dinoflagellates from the lake Baikal and also the first study of the diversity of dinoflagellates in freshwater sponges.

Dinoflagellate DNA fragments from both the environmental samples (water and endemic Baikalian sponges) and from the real organisms living in the spring plankton under ice were obtained. In the latter case single-cell PCR was used. Fragments of rRNA genes, as well as ITS-2 were sequenced with following Bayesian and Maximum-Likelihood analyses. It has been demonstrated that Baikalian dinoflagellates belong to phylogenetically distinct groups related to the genera from Gymnodiniaceae, Pfiesteriaceae, Suessiaceae families. Special interest is addressed to two groups of DNA sequences related to Suessiaceae because they may belong to putative symbionts of freshwater sponges. The most numerous dinoflagellate species that regards as Baikalian endemic, was found to be a true member of the Gymnodiniales order and appears to be relict but not endemic species. It has close relatives in the glacial melt waters of the Arctic Ocean, that is unusual because many freshwater dinoflagellates form groups phylogenetically distinct from marine ones. In general, Baikalian dinoflagellates have multiple origins and different timing of appearance in the lake.

P-1120

Moving beyond ribosomal genes: Metagenomics for microbial eukaryotes

Holly Bik, Aaron Darling, Guillaume Jospin, Jonathan Eisen
University of California, Davis, Davis, CA, USA

Microbial eukaryotes (e.g. nematodes, fungi, protists, and other 'minor' metazoan phyla) are diverse and abundant in every habitat on earth, yet for most groups we possess scant knowledge of species distributions and global diversity. Underdeveloped bioinformatic pipelines and the lack of reference genomes have severely hindered the utility of true metagenomic approaches for these taxa; instead, high-throughput studies of environmental diversity (454, Illumina) to date have relied heavily on ribosomal marker genes. Building a robust picture of ecosystem function necessitates the inclusion of disparate data types such as whole-genome and transcriptome sequences, in addition to surmounting the significant informatic challenges for interpreting rRNA datasets (stemming from the existence of intragenomic variation across rRNA genes in eukaryotic genomes). Here we present recent efforts towards tackling the perpetual computational bottlenecks associated with increasingly large sequence datasets, specifically focusing on advances in the analysis and biological interpretation of environmental, eukaryotic datasets. Thorough expansion of the AMPHORA2 pipeline (https://github.com/gjospin/Amphora-2) we have greatly improved our capacity for accurate taxonomic assignments in eukaryotes using both ribosomal OTUs and conserved, protein-coding marker genes mined from metagenome data.

P-1121

Evolution of microsatellite DNA in human and chimpanzee genomes

Naoko Takezaki
Kagawa University, Kitagun, Kagawa, Japan

In this study evolutionary change of microsatellite loci (tandem repeats of short nucleotides) was investigated in genomic sequences of human and chimpanzee with orangutan as outgroup.

Proportions of conserved loci between species (P_C) for repeat unit size 1-5 bp and those of invariable loci (P_I) for loci with repeat unit 1-3 bp showed exponential decay with time; P_I for loci with repeat unit 4-5 bp deviated from exponential decay probably because P_I quickly reached the minimum for these loci. The rates of turnover (gain and loss) of loci were in the order of $10^{-2} – 10^{-4}$ per genome per year. They were slightly lower for human and chimpanzee lineages than for their common ancestor as well as for the human lineage than for the chimpanzee lineage. The rates of gain of loci are higher than those of loss in the human and chimpanzee lineages. In contrast, the rates of loss of loci are higher than the rates of gain in the common ancestor.

The rates of gain and loss of repeats are in the order of $10^{-8} – 10^{-9}$ per locus per year. Similarly to the rates of gain and loss of loci, the rates of gain are higher than those of loss in the human and chimpanzee and the rates of loss are higher than those of gain in the common ancestor. However, the rates for gain and loss of repeats are higher in the human and chimpanzee lineages than the common ancestor and the rates of gain tend to be higher in human than in chimpanzee and the rates of loss are higher in chimpanzee than in human.

P-1122

From Gene Trees to the Species Tree in the Adaptive Radiation of Cichlid Fishes from Lake Tanganyika, East Africa

Britta S. Meyer, Walter Salzburger
University of Basel, Basel, Switzerland

The Great Lakes in the East African Rift Valley harbour an astonishing biodiversity of about 2,000 endemic species of cichlid fishes. Cichlids are one of the most important model systems in evolutionary biology, which is mainly due to their species-richness, the unparalleled ecological and morphological diversity, and the rapidity of the formation of new species. For exactly the same reasons, it has been notoriously difficult to resolve phylogenetic relationships within East African cichlid species flocks, although robust phylogenies are an important basis for evolutionary studies regarding e.g. the triggers of explosive speciation and adaptive radiation or the function of evolutionary key innovations. Here, we present a new phylogenetic hypothesis for the cichlid species flock of Lake Tanganyika based on a novel multi-marker dataset. We first designed a new set of 46 nuclear markers and developed a strategy for massive parallel sequencing on the basis of the 454 technology. Our multi-marker matrix contains close to 20 kb each from 500 individual specimens, representing the phylogenetic spectrum of East African cichlid fishes. This matrix yields the so far most robust phylogenetic hypothesis for Lake Tanganyikan cichlids and is highly suitable for gene tree/species tree comparisons. At the same time, our sequencing strategy allows us to call SNPs and identify nuclear haplotypes in the 46 loci. Our analyses at the interface between phylogenetics and population-genomics emerges as the ideal strategy to resolve the evolutionary history of a rapid radiation.

P-1123

Decreasing speciation or incomplete sampling?

Disa Hansson, Sebastian Höhna
Dept. of Mathematics, Stockholm University, Stockholm, Sweden

Studies of species diversification often use the birth and death process. The diversification rates can be estimated even if only a reconstructed phylogeny without extinct taxa is available, e.g. a phylogeny estimated from molecular data. Previously, most models assumed constant diversification rates. However, under a constant rates model, and a speciation rate higher than the extinction rate, the number of extant taxa would increase infinitely. Furthermore, constant rate models do not agree with rapid radiation, which have been observed many times.

When performing analysis of diversification patterns on reconstructed phylogenetic trees, models without extinction rates tend to be chosen in spite of what we know from the fossil record. We argue that this result could be tied to the assumption that all extant species are included in the tree, i.e. incomplete sampling has not been taken into account. In this project we study which effect incomplete sampling on the inference of diversification rates under time varying rate models has and if incomplete phylogenies are needed for unbiased estimates.

We use birth-death process to derive the likelihood equation under time varying rates and incomplete sampling, given the divergence times. We analyzed 32 bird phylogenies and conclude that even when incomplete taxon sampling is considered, the pure birth process with a decreasing speciation rates is selected for most datasets. However, models including extinction rates are very close to the best model and the extinction is not estimated to be zero anymore. If the number of missing taxa is actually underestimated, it could explain the very small extinction rates and the preference of the pure birth model. Realistic estimates do not need more taxa but better estimates of the total number of species in a clade. Furthermore, we show by simulations that it is possible to distinguish incomplete sampling from decreasing speciation, though the diversifications rates are effected by incomplete sampling.

P-1124

Loss of Different Inverted Repeat Copies from the Chloroplast Genomes of Pinaceae and Cupressophytes and Influence of Heterotachy on the Evaluation of Gymnosperm Phylogeny

Chung-Shien Wu[1], Ya-Nan Wang[2], Chi-Yao Hsu[1], Ching-Ping Lin[1], Shu-Miaw Chaw[1]
[1]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan, [2]School of Forestry and Resource Conservation, National Taiwan University, Taipei, Taiwan

The relationships among the extant five gymnosperm groups-gnetophytes, Pinaceae, non-Pinaceae conifers (cupressophytes), Ginkgo, and cycads-remain equivocal. To clarify this issue, we sequenced the chloroplast genomes (cpDNAs) from two cupressophytes, *Cephalotaxus wilsoniana* and *Taiwania cryptomerioides*, and 53 common chloroplast protein-coding
genes from another three cupressophytes, *Agathis dammara, Nageia nagi*, and *Sciadopitys verticillata*, and a non-Cycadaceae cycad, *Bowenia serrulata*. Comparative analyses of 11 conifer cpDNAs revealed that Pinaceae and cupressophytes each lost a different copy of inverted repeats (IRs), which contrasts with the view that the same IR has been lost in all conifers. Based on our structural finding, the character of an IR loss no longer conflicts with the "gnepines" hypothesis (gnetophytes sister to Pinaceae). Chloroplast phylogenomic analyses of amino acid sequences recovered incongruent topologies using different tree-building methods; however, we demonstrated that high heterotachous genes (genes that have highly different rates in different lineages) contributed to the long-branch attraction (LBA) artifact, resulting in incongruence of phylogenomic estimates. Additionally, amino acid compositions appear more heterogeneous in high than low heterotachous genes among the five gymnosperm groups. Removal of high heterotachous genes alleviated the LBA
artifact and yielded congruent and robust tree topologies in which gnetophytes and Pinaceae formed a sister clade to cupressophytes (the gnepines hypothesis) and Ginkgo clustered with cycads. Adding more cupressophyte taxa could not improve the accuracy of chloroplast phylogenomics for the five gymnosperm groups. In contrast, removal of high heterotachous genes from data sets is simple and can increase confidence in evaluating the phylogeny of gymnosperms. Key words: phylogenomics, chloroplast genome, gymnosperms, heterotachy, long-branch attraction.

P-1125

Accelerated exon evolution in duplicated regions in hominids

Belen Lorente-Galdos[1,2], Jonathan Bleyhl[3], Laura Vives[3], Gregory Cooper[1], Arcadi Navarro[1,4], Evan Eichler[3,5], Tomas Marques-Bonet[1]
[1]IBE, Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, PRBB, Barcelona, Spain,[2]National Institute for Bioinformatics (INB), Barcelona, Spain,[3]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA,[4]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain,[5]Howard Hughes Medical Institute, Seattle, Washington 98195, USA

Identification of signatures of positive selection has long been a major issue for understanding the unique features of any given species. However, only a fraction of human genes have been interrogated. Genes within segmental duplications are usually omitted due to the limitations of draft nonhuman genome assemblies and the methodological reliance on accurate gene trees. In this work, we show the feasibility of a new method that does not need accurate gene trees or individual high-quality assemblies. We consider all high-quality nucleotide differences between shotgun sequencing reads from single human and macaque individuals relative to the human assembly. Comparing the observed rates of nucleotide differences between coding exons and their flanking intronic sequences with a likelihood ratio test, we identified 74 exons with evidence for rapid coding sequence evolution during human and Old World monkey evolution and validated five of them by means of experimental analysis. Strikingly, compared to only 6% of duplicated exons initially analyzed, 55% (41/74) of rapidly evolving exons were either partially or totally duplicated. Our results suggest abundant accelerated coding sequence evolution within these duplicated and highly dynamic regions of the genome and provide a more comprehensive view of the role of selection on human genome evolution.

P-1126

The Neolithic trace in mitochondrial haplogroup U8

Joana Barbosa Pereira[1,2], Marta Daniela Costa[1,2], Pedro Soares[2], Luísa Pereira[2,3], Martin Brian Richards[1,4]
[1]Institute of Integrative and Comparative Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK, [2]Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal, [3]Faculdade de Medicina da Universidade do Porto, Porto, Portugal, [4]School of Applied Sciences, University of Huddersfield, Huddersfield, UK

The mitochondrial DNA (mtDNA) still remains an important marker in the study of human history, especially if considering the increasing amount of data available. Among the several questions regarding human history that are under debate, the model of expansion of agriculture into Europe from its source in the Near East is still unclear. Recent studies have indicated that clusters belonging to haplogroup K, a major clade from U8, might be related with the Neolithic expansions. Therefore, it is crucial to identify the founder lineages of the Neolithic in Europe so that we may understand the real genetic input of the first Near Eastern farmers in the current European population and comprehend how agriculture spread so quickly throughout all Europe.
In order to achieve this goal, a total of 55 U8 samples from the Near East, Europe and North Africa were selected for complete characterisation of mtDNA. A maximum-parsimonious phylogenetic tree was constructed using all published sequences available so far. Coalescence ages of specific clades were estimated using ρ statistic, maximum likelihood and Bayesian methods considering a mutation rate for the complete molecule corrected for purifying selection.
Our results show that U8 dates to ~37-54 thousand years ago (ka) suggesting that this haplogroup might have been carried by the first modern humans to arrive in Europe, ~50 ka. Haplogroup K most likely originated in the Near East ~23-32 ka where it might have remained during the Last Glacial Maximum, between 26-19 years ago. The majority of K subclades date to the Late Glacial and are related with the repopulation of Europe from the southern refugia areas. Only a few lineages appear to reflect post glacial, Neolithic or post-Neolithic expansions, mostly occurring within Europe. The major part of the lineages dating to the Neolithic period seems to have an European origin with exception of haplogroup K1a4 and K1a3. Clade K1a4 appears to be originated from the Near East where it also reaches its highest peak of diversity. Despite the main clades of K1a4 arose in the Near East during the Late Glacial, its subclade K1a4a1 dates to ~9-11 ka and is most likely related with the Neolithic dispersal to Europe. Similarly, K1a3 probably originated in the Near East during the Late Glacial and its subclade K1a1a dispersed into Europe ~11-13 ka alongside with the expansion of agriculture.

P-1127

Late Glacial Expansions in Europe revealed through the fine-resolution characterisation of mtDNA haplogroup U8

Marta Daniela Costa[1,2], Joana Barbosa Pereira[1,2], Pedro Soares[2], Luísa Pereira[2,3], Martin Brian Richards[1,4]
[1]Institute of Integrative and Comparative Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK, [2]IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal, [3]Faculdade de Medicina, Universidade do Porto, Porto, Portugal, [4]School of Applied Sciences, University of Huddersfield, Huddersfield, UK

The maternally inherited and fast evolving mitochondrial DNA (mtDNA) molecule is a highly informative tool with which to reconstruct human prehistory. This has become even more true in recent years, as mtDNA based studies are becoming more robust and powerful due to the availability of complete mtDNA genomes. These allow better mutation rate estimates and fine-resolution characterisation of the phylogeography of mtDNA haplogroups, or named clades. MtDNA haplogroup K, the major subclade of U8, occurs at low frequencies through West Eurasian populations, and is much more common in Ashkenazi Jews. However, the lack of variation on the first hypervariable segment (HVS-I) has precluded any meaningful phylogeographic analysis to date. We therefore completely sequenced 50 haplogroup K and 5 non-K U8 mtDNA samples from across Europe and the Near East, and combined them with 343 genomes previously deposited in GenBank, in order to reconstruct a detailed phylogenetic tree. By combining several inference methods, including maximum parsimony, maximum likelihood and Bayesian inference it was possible to trace the timescale and geography of the main expansions and dispersals associated with this lineage. We confirmed that haplogroup K, dating to ~32 thousand years (ka) ago, descended from the U8 clade, which coalesces ~48 ka ago. The latter is close to the timing of the first arrival of modern humans in Europe and U8 could be one of the few surviving mtDNA lineages brought by the first settlers from the Near East. U8 split into the widespread U8b, at ~43 ka, and U8a, which seems to have expanded only in Europe ~24 ka ago. Considering the pattern of diversity and the geographic distribution, haplogroup K is most likely to have arisen in the Near East, ~32 ka ago. However, some subclades were evidently carried to Europe during the Last Glacial Maximum (LGM). We observed significant expansions of haplogroup K lineages in the Late Glacial period (14-19 ka), reflecting expansions out of refuge areas in southwest and possibly also southeast Europe.

P-1128

Probabilistic parsimonious reconciliation between genes and species trees

Leonardo de Oliveira Martins, David Posada
University of Vigo, Vigo, Galicia, Spain

With the accumulation of genomic data we can expect phylogenetic inferences between distinct genes to disagree. Such disagreement is not only an artifact due to the inference procedure or limited amount of data, but more likely represents biological phenomena like lateral transfers, ancestral polymorphisms or duplications and losses. When doing phylogenomic analyses we should, therefore, take these events into account when estimating the species tree. This can be done by e.g. finding the species trees that minimize the overall number of such events, that is, the most parsimonious reconciliations between genes and species. On the other extreme fully probabilistic models have been developed, that describe the distribution P(G/S) of genes trees G for a given species tree S and can consider all possible reconciliations.

We decided to take an intermediate approach, where we have a probability distribution P(G/S) of gene trees under a species tree, but this distribution only considers the minimum distance between the trees G and S. This distance is the cost -- number of events like duplications and losses -- of the most parsimonious reconciliation scenario between the unrooted gene tree and the species tree.

Using this distribution we have built a Bayesian hierarchical model to estimate the posterior distribution of species trees given a collection of gene families, assuming that the disagreements between genes and species trees are due to duplications and losses or due to incomplete lineage sorting. When calculating P(G/S) the root position for G that minimizes the reconciliation cost must be found since we work with unrooted gene trees, and thus our procedure can also output this collection of likely rootings for each gene.

P-1129

MiRCandRef a local assembly algorithm for description of microRNAs in non-genome organisms

Bastian Fromm[1], Thomas D. Otto[2], Christoph Hahn[1], Philip D. Harris[1], Lutz Bachmann[1]
[1]Natural History Museum, University of Oslo, Oslo, Norway, [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

MiRNAs are single-stranded, 22 nucleotide long, noncoding transcripts derived from different genome-encoded hairpin precursors that regulate gene expression negatively. They represent the most recently discovered gene regulators and are involved in a broad variety of biological processes. Due to the considerable interest in the role of miRNA in animal development, many bioinformatic genome- and non-genome based methods are available to characterise the miRNA complement of tissues or organisms of choice. These methods rely upon data from high-throughput miRNA samples but they differ considerably in the number of mirnas they predict, because of either the incompleteness, or the the lack of suitable reference genomes
To overcome the problems caused by the quality/existence of the reference genome, we are currently developing an algorithm to assemble the genomic region around all unique miRNA reads, using local assembly of genomic reads (MiRCandRef). The algorithms is programmed in PERL and uses third party tools such SMALT, BLAST, VELVET and IMAGE. With this method we create a miRNA-candidate reference that is used for remapping all available miRNA reads using state-of the art miRNA prediction software.
Preliminary analyses with the platyhelminth *Gyrodactylus salaris* show several advantages: using MiRCandRef a complete reference genome is not necessary for identification of all miRNA precursors from an organism. This has great potential to save resources and time for other projects and makes miRNA analyses possible without the need of huge-memory computing. Furthermore the number of identified miRNAs is substantially higher than that derived from global alignments of the same data-set.

P-1130

Developing Maximum Likelihood and Bayesian Supertrees

Wasiu Akanni[1], Peter Foster[2], Mark Wilkinson[2], Davide Pisani[1]
[1]NUI, Maynooth, Ireland,[2]NHM, London, UK

Little work has been done on the development of supertree methods in the Likelihood and Bayesian frameworks. Recently, it has been proposed that Maximum Likelihood (ML) supertrees could be developed by using an exponential distribution to model the probability that the input trees could be erroneous. When the tree-to-tree distances used in the ML computation are calculated using the Symmetric Difference, the ML supertree has been shown to be equivalent to a Majority Rule Consensus Supertree, and hence, exactly as the latter, the ML supertree must have the desirable property of being a median tree – with reference to the input set. In addition, the ability to estimate the likelihood of supertrees, will allow the implementation of Bayesian MCMC approaches, which have the advantage of allowing for a natural estimation of the support for the nodes on the recovered supertree. We have developed the first software for the estimation of Maximum Likelihood and Bayesian s
upertrees. The program is being written in Python and will also exploit the capabilities of already available software, i.e. P4.

Here, we present results of reanalyses of the datasets of Holton and Pisani (2010) and Pisani(2007) and present the first Bayesian Genomic Supertrees generated using our new software. In addition, we shall compare these supertrees with those derived using other common supertree methods (e.g. Matrix Representation with Parsimony and Average Consensus). We show that results obtained using our method compared well with those obtained using MRP, and significantly outperforms all other methods.

P-1131

Investigating deep nodes in Polyporales

Elisabet Sjökvist[1], Bernard Pfeil[1], Ellen Larsson[1], Karl-Henrik Larsson[2]
[1]University of Gothenburg, Gothenburg, Sweden, [2]University of Oslo, Oslo, Norway

Polyporales is a fungal order comprising roughly 1800 species of wood-decaying basidiomycetes. These are primarily polyporoid and corticioid, but a few species have lamellae. Sporocarp morphology appears to be misleading for resolving the relationships within the order, most genera are polyphyletic and gene tree phylogenies are poorly resolved.
This study aims at exporing methodologies to solve gene-tree incongruences on the deeper nodes.
Preliminary results will be presented from a dataset using
both protein coding genes, nrDNA and mitochondrial DNA from a broad sample of taxa, representing the major lineages in Polyporales.

P-1132

The poor man's 1000 genomes project: Recent human population expansion confounds detection of disease alleles in 7,099 complete mitochondrial genomes

Hie Lim Kim, Stephan C. Schuster
Pennsylvania State University, University Park, PA, USA

We analyzed 7,099 complete human mitochondrial (mt) genomes, which were retrieved from the GenBank, to find out how many, and how often disease-related mutations existed in the mt genomes of the human population. In the 7,099 genomes, 6,110 Single Nucleotide Variants and their frequency of occurrence were identified with the exception of the heteroplasmy and indel sites. Among them, 4,092 variants were observed in at least two genomes. Most of the variants were rare variants, indicating recent expansions of the human population. The lack of sharing of the variants between populations was likely a result of population bottlenecks and migrations since the out of Africa of modern humans. It is important to estimate population history to properly examine the distribution of a pathogenic mutation in the human population. The extent of nucleotide diversity in the 7,099 genomes was used to determine the best-fit demographic model of the human population for mt genomes. Based on this model, we simulated the evolution of mt genomes within the human population to ascertain the behavior of deleterious mutations. The exponential population expansions of the non-African populations generated a lot of numbers of new mutations. Most of them quickly disappeared by drift before their frequency increased even under neutrality, because of the rapid population growth. A deleterious mutation was under the selection constrains limiting its increase in frequency, was more rarely found in a population than a neutral mutation. From our simulations, we derived the threshold frequency of a deleterious mutation for each of the African, European, and Asian populations and applied it to identify mt pathogenic mutations. Among the 6,110 variants we identified in the 7,099 genomes, each of the African, European, and Asian populations contained 2,017, 3,596, and 4,003 variants respectively, and 67%, 82%, and 74% of the corresponding variants of each of those respective populations, showed a lower frequency than the threshold. In the large number of candidate variants, only 12 known pathogenic mutations were detected in the 7,099 genomes. Our results showed a difficulty in detecting a disease-related mutation within the abundance of rare variants in the human population even with the large number of genomes in our sample.

P-1133

Deep relationships of Rhizaria revealed by phylogenomics

Roberto Sierra, Jan Pawlowski
University of Geneva, Geneva, Switzerland

Rhizaria is one of the six supergroups of eukaryotes, which comprise the majority of amoeboid and skeleton-building protists. The overall lack of molecular data for the group results in unresolved deep phylogeny of rhizarians. Molecular data are particularly scarce for the clade of Retaria, which include two most important groups of microfossils: Foraminifera and Radiolaria. To fill this gap, we have produced and sequenced EST libraries for 14 rhizarian species including Foraminifera and several taxa of traditional Haeckel's Radiolaria: Acantharea, Collodaria, Spumellaria, and Phaeodarea. A matrix was constructed for phylogenetic analysis based on 109 genes and a total of 56 species, of which 22 are rhizarians. The phylogenomic data set presented here constitutes the largest and most complete available for Rhizaria to date. Our analyses confirm the hypothesis of polyphyly of Haeckel's Radiolaria providing the first multigene evidence for branching of Phaeodarea within Cercozoa. We also confirm the monophyly of Retaria, a clade grouping Foraminifera with other lineages of Radiolaria.

P-1134

Reticulated origin of domesticated tetraploid wheat

Peter Civan
Centro de Ciencias do Mar, Universidade do Algarve, Faro, Portugal

The past 15 years have witnessed a notable scientific interest in the topic of crop domestication and the emergence of agriculture in the Near East. Multi-disciplinary approaches brought a significant amount of new data and a multitude of hypotheses and interpretations. However, some seemingly conflicting evidence, especially in the case of emmer wheat, caused certain controversy and a broad scientific consensus on the circumstances of the wheat domestication has not been reached, yet.

The past phylogenetic research has translated the issue of wheat domestication into somewhat simplistic mono-/polyphyletic dilemma, where the monophyletic origin of a crop signalizes rapid and geographically localized domestication, while the polyphyletic evidence suggests independent, geographically separated domestication events. Interestingly, the genome-wide and haplotypic data analyzed in several studies did not yield consistent results and the proposed scenarios are usually in conflict with the archaeological evidence of lengthy domestication.

Here I suggest that the main cause of the above mentioned inconsistencies might lie in the inadequacy of the divergent, tree-like evolutional model. The inconsistent phylogenetic results and implicit archaeological evidence indicate a reticulate (rather than divergent) origin of domesticated emmer. Reticulated genealogy cannot be properly represented on a phylogenetic tree; hence different sets of samples and genetic loci are prone to conclude different domestication scenarios. On a genome-wide super-tree, the conflicting phylogenetic signals are suppressed and the origin of domesticated crop may appear monophyletic, leading to misinterpretations of the circumstances of the Neolithic transition.

The network analysis of multi-locus sequence data available for tetraploid wheat clearly supports the reticulated origin of domesticated emmer and durum wheat. The concept of reticulated genealogy of domesticated wheat sheds new light onto the emergence of Near-Eastern agriculture and is in agreement with current archaeological evidence of protracted and dispersed emmer domestication.

P-1135

Using a High Powered Computing Network to Facilitate Genomics Focused Research at a Medium Sized Research Institute

Daniel White
Landcare Research New Zealand, Auckland, New Zealand

Today, with the establishment of next generation sequencing technologies as a regular part of biological research, and considering these technologies are constantly being developed and upgraded, vast amounts of DNA sequence data are being generated at an ever increasing rate. As such, the bottleneck in biological research that utilises DNA sequence data has transitioned from molecular procedures in the lab, to the processing of this data, and further downstream analyses. Research undertaken at Landcare Research New Zealand, a crown research institute that focuses on the terrestrial environment, is no exception. Landcare Research, which employs around 400 staff, consists of 8 teams of integrated scientific disciplines, all working towards Landcare's core objective of conserving New Zealand's terrestrial biodiversity. Of these teams, 5 are dedicated to biological or ecological research and, of these, 4 are utilising next generation sequencing technologies to explore relevant questions. To facilitate the processing of large datasets from whole genomes, which grossly over burden desktop machines, Landcare Research is in a unique position where it can harness the resources of a national high power computing network established, in part, for the research community of New Zealand. The New Zealand eScience Infrastructure (NeSI), managed by the Centre for eResearch from the University of Auckland, consists of a networked grid of high powered computer clusters. In particular, the most recent addition to the grid is an Intel© cluster consisting of 80 compute nodes, each with 96GB of RAM and 12 cores. The challenge for Landcare Research is to develop workflows that utilise these computing resources for biologists with little or no informatics expertise, and who may need to use multiple programs as part of their analysis. Here, I demonstrate such a workflow and include a case study as an example that reduces computation time from around one week to just several hours. This work shows how researchers at institutes of the size of Landcare Research,where cost effectiveness is an important consideration, can realise the full potential of sequencing technologies currently available, by establishing a link between biological research and remote, high powered informatics tools.

P-1136

Mining Genomes for Microinversions: Rare Genomic Characters for Phylogenomics

Gary Stuart, Arun Seetharam
Indiana State University, Terre Haute, IN, USA

Rare genomic characters in general and microinversions in particular are of interest because they are best candidates for "near perfect" phylogenetic characters due to multiple alternative states and very slow conversion rates. There is currently a very broad size range definition for microinversions (5bp to over 50kb), and it appears that the smallest inversions are generally the most frequent class. We have begun to search systematically for the smallest microinversions: those under 50 bp in length. A prototype microinversion search program was developed and tested on two yeast genomes (S. cereviseae and S. paradoxus) and three fruit fly genomes in pairwise combination (*D. melanogaster*, *D. simulans*, and *D. sechillia*). A divergence time for the two yeast genomes is difficult to estimate but might be roughly10-30 MYA. In contrast, divergence time estimates for the three fly genomes are considered to be more accurate: *D. simulans* and *D. sechillia* diverged roughly 5 MYA, while *D. melanogaster* diverged from these roughly 10 MYA. Hence the fly genomes are closely related and would have sufficient sequence similarity to presumably make pairwise microinversion detection a reasonable task. This should be the case for the yeast genomes as well. In fact, a large-scale scan of the two yeast genomes revealed multiple examples of likely orthologs that appear to be good candidates for microinversions. In addition, early attempts at small scale scans of the three larger fly genomes also provided several examples of microinversions shared between these species. Many of these can be expected to be useful as RGC's and should help to provide a novel reconsideration of the currently accepted phylogenetic hypothesis for these species. In the near future, we intend to scale up to full genome comparisons and begin to provide more accurate estimates of the frequency of observable microinversions of this type within genomes. It should also be of interest, as we uncover larger numbers of these microinversions, to determine where in the genome they tend to be found and in what sequence context they are likely to occur.

P-1137

Analysis of ancient transposon families using molecular dating and a population genetics paradigm

Elizabeth Hellen, John Brookfield
University of Nottingham, Nottingham, Nottinghamshire, UK

Population genetics techniques, such as molecular dating using Monte Carlo Markov Chain methods, can be used to analyse ancient transposon families, in which each element copy is treated as an individual genetic lineage in a population containing all copies of the transposon. Our previous work concentrated on using Primate lineages, with the divergence date between Primate species being used as a constraint to time the overall tree. This yielded promising results in the analysis of the origin dates and evolutionary rates of the Golem transposon families. However, this previous analysis had a number of assumptions associated with it, such as the assumption that modern primate evolutionary rates are to the same as those in mammals more than 100 MY ago and that large scale deletions of transposons have not occurred.

To reduce the impact these assumptions have on the predicted origin dates of the transposon families, we have independently analysed the transposon families from Primate, Carnivora and Artiodactyla lineages. A large-scale molecular dating analysis of class II transposon families found in the modern human genome, has been carried out using BEAST. The results allow us to describe the origin dates of mammalian transposons and to analyse how and when new transposon families developed. An analysis of the interior branches and the structure of the trees produced allows prediction of the dates when inactivation of the transposons occurred. Evolutionary rates are also analysed to determine whether different rates can be seen across the different mammalian lineages, whether the transposons evolve at similar rates in different positions in the genome and whether there are systematic differences in evolutionary rates in the different transposon families.

P-1138

How protein targeting to primary plastids via the endomembrane system could have evolved?

Przemyslaw Gagat[1], Pawel Mackiewicz[1], Andrzej Bodyl[2]
[1]Department of Genomics, Faculty of Biotechnology, University of Wroclaw, Wroclaw, Poland,[2]Department of Biodiversity and Evolutionary Taxonomy, University of Wroclaw, Wroclaw, Poland

Before 1.5 billion years ago a phagotrophic eukaryotic ancestor of glaucophytes, red algae, and green plants engulfed a cyanobacterium that was transformed into a primary plastid with two envelope membranes. Gene transfer from the cyanobacterial endosymbiont genome to the host nucleus fostered the integration of both endosymbiotic partners, but it is still unclear how protein products of the transferred and other host nucleus-encoded genes were initially transported into the cyanobacterial endosymbiont/primary plastid. At present, almost all nucleus-encoded proteins are imported into primary plastids post-translationally using an N-terminal transit peptide and Toc and Tic translocons. However, there are several proteins with an N-terminal signal peptide that are directed to higher plant plastids in vesicles derived from the endomembrane system, such as the *Arabidopsis thaliana* α-carbonic anhydrase (CAH1), the *Oryza sativa* nucleotide pyrophosphatase/phosphodiesterase (NPP1), and two *O. sativa* α-amylases (αAmy3, αAmy7). The existence of such proteins inspired the formulation of an "early endomembrane trafficking" hypothesis postulating that nuclear-encoded, plastid-targeted proteins carried initially signal peptides and were transported to the ancestral primary plastid via the endomembrane system. To test if these proteins are indeed relics of the hypothesised ancestral import pathway we performed their phylogenetic analyses. Our results indicate that CAH1, NPP1, αAmy3, and αAmy7 are of the eukaryotic (not cyanobacterial) origin and that their non-plastid homologs are also equipped with signal peptides responsible for their co-translational import into the endoplasmic reticulum and subsequent compartments of the endomembrane system including the extracellular space. Their vesicular trafficking to the plastid must have evolved secondarily after the primary endosymbiosis, and probably only in the land plant lineage, to satisfy their needs for glycosylation and/or their transport to more than one cellular compartment. Therefore CAH1, NPP1, αAmy3, and αAmy7 cannot be considered a relic of the hypothetical endomembrane system-mediated protein targeting to the ancestral primary plastid. The protein import into primary plastids was dominated right from the beginning by the Toc- and Tic-based route, and only few host proteins, previously targeted to distinct endomembrane system compartments, exploited their signal peptides to reach the plastid via the vesicular trafficking.

P-1139

Distribution of Plastid Gene Signals among Chromalveolates without Plastids

Christian Woehle, William Martin, Sven B. Gould
Molecular Evolution, Heinrich-Heine-University, Duesseldorf, Germany

The infrakingdom chromalveolata unites a potpourri of protists whose evolutionary origin is still actively discussed. Many chromalveolate lineages harbor plastids or relicts thereof, but in some there is no morphological trace of the plastid left. In a few of species however, footprints of a potential plastid organelle have been identified in forms of nuclear-encoded genes tracing back to a once photosynthetic lifestyle. Nevertheless, convincing evidence for the presence of such evolutionary markers does not exist for all of them, and with current data suggesting cryptophytes and haptophytes to share a plastid of secondary red origin that evolutionary differs from that of heterokonts and alveolates, the chromalveolate debate continues. The oomycetes, which consist mainly of plant pathogens, represent one of the more problematic phyla among chromalveolates. These protozoa group to the base of heterokonts, but they do not house a plastid organelle, and it is discussed whether they ever did. To shed light on the plastid-associated ancestry of oomycetes we examined the phylogenetic position of this phylum in comparison to other chromalveolates and more generally archaeplastida. We clustered orthologous genes based on 536692 individual proteins of 35 different organisms and constructed more than 10.000 phylogenetic trees. We focused in particular on those results, in which oomycetes grouped within the archaeplastida. This was the case for, on average, 180 oomycete genes among trees containing archaeplastida, chromalveolates and at least one outgroup species. These results suggest that the amount of potential plastid gene signal in oomycetes is only slightly smaller than in other heterokonts, where on average 250 genes sit within the archaeplastidal trees. Here we present our latest results in comparing the plastid gene signal in oomycetes with that of other chromalveolates, who might also have once possessed a plastid, or maybe not.

P-1140

Identification and characterization of angiosperm single-copy gene families

Riet De Smet[1,2], Keith Adams[3], Klaas Vandepoele[1,2], Yves Van de Peer[1,2]
[1]VIB, Ghent, Belgium,[2]Ghent University, Ghent, Belgium,[3]University of British Columbia, Vancouver, Canada

Gene retention and gene loss patterns following duplication have been extensively studied to reveal the consequences of altered gene dosage (here represented by changes in gene copy number) on organism's fitness and evolutionary adaptation. In particular, contrasting gene retention patterns following whole-genome duplication (WGD) and tandem duplications in human, yeast, fish and plants have revealed that duplication of some genes might be better tolerated than that of others. Duplicates of regulatory genes and signal transducers, for instance, seem to be only retained in case of WGD events. The ‘gene balance hypothesis' provides an explanation for this and states that these genes are part of protein complexes or regulatory networks for which it is detrimental that they are present in stoichiometric balances. Less attention has been paid to genes for which duplication is never tolerated (i.e. genes that are always single-copy), the so-called duplication-resistant genes.

Here we took advantage of the large number of available angiosperm genomes in the PLAZA 2.5 database to identify duplication-resistant genes with a high accuracy. These genes can be identified as gene families with a one-to-one orthology relationship in all available genomes (single-copy gene families). In particular, using the OrthoMCL method, we identified such gene families in 17 angiosperm genomes, covering a wide range of evolutionary distance. By using such a large set of genomes we can filter out genes that randomly returned to single-copy status and only focus on those that are most likely single-copy due to evolutionary constraints. In addition, we mainly incorporated genomes of high annotation quality as to minimize the errors in single-copy gene family prediction due to annotation errors. To accommodate for mispredictions of orthology/paralogy relationships by OrthoMCL we included a phylogeny-based quality-check of the obtained gene families. The identified single-copy gene families show a bias towards certain functional processes and will be further investigated in terms of their sequence and functional conservation.

P-1141

Recurrent patterns in the evolution of self-compatibility in Arabidopsis (Brassicaceae)

Takashi Tsuchimatsu, Pascal Kaiser, Julien Bachelier, Chow Lih Yew, Kevin Ng Kit Siong, Kentaro Shimizu
University of Zurich, Zurich, Switzerland

Ever since Darwin's pioneering research, the evolutionary transition from outcrossing to self-fertilization (selfing) through the loss of self-incompatibility (SI) has been considered one of the most prevalent events in flowering plants, and its genetic basis has been a major focus in evolutionary biology. SI systems generally consist of male and female specificity genes at the *S*-locus, and SI modifier genes that are not linked to the *S*-locus. Over the past decades, much attention has been paid particularly to clarify the mutations of which genes were responsible for the loss of SI. In the Brassicaceae, the female and male components for recognition are determined by the *S*-locus receptor kinase (SRK) and by the *S*-locus cysteine-rich protein (SCR; also known as *S*-locus protein 11, SP11), respectively. Here, we investigated the pattern of polymorphism and functionality of the *S*-locus components in allotetraploid *Arabidopsis kamchatica*, which is self-compatible in contrast to its parental species, *A. lyrata* and *A. halleri*. We identified five highly diverged *SRK* haplogroups and surveyed the geographic distribution of *SRK* at the two homeologous *S*-loci across the species range, which is the first time this has been conducted in a wild allotetraploid species. We found intact full-length *SRK* sequences in many accessions. Through interspecific crosses with the self-incompatible congener *A. halleri*, we found that the female components of the SI system, including *SRK* and the modifier genes, are still functional in these accessions, suggesting that the degradation of male components was responsible for the loss of SI in *A. kamchatica.* We found that the distribution of three *A. halleri*-derived *S*-haplogroups was significantly correlated with the population structure. In contrast, for the *A. lyrata*-derived *S*-locus, the markedly high frequency of *AkSRK-D* is consistent with the scenario that its prevalence over population clusters might have been driven by positive selection before or after the origins of *A. kamchatica*. In *Arabidopsis thaliana*, we have reported similar patterns: the 213-bp inversion mutation in the male specificity gene *SCR* was responsible for the evolution of self-compatibility and was found in >95% of European accessions. Our compilation and statistical analysis of recent extensive studies in multiple Brassicaceae species demonstrate that in contrast to cultivated species, the loss of SI tends in wild species to be derived from mutations in the male components. This is consistent with theoretical predictions that mutations disabling male components are more strongly selected under natural selection than mutations disabling female components.

P-1142

Evolution of microRNA in primates

Jennifer McCreight, Willie Swanson
University of Washington, Seattle, WA, USA

Comparative genomics is an indispensable tool for studying primate evolution. Differences in protein-coding sequences and gene expression have uncovered the molecular basis for numerous phenotypic differences. Another regulator of gene expression that has been mostly neglected is microRNA (miRNA), a short, noncoding, single-stranded RNA involved in post-transcriptional regulation in eukaryotes. Immature pre-miRNA forms a hairpin structure that is cleaved into mature miRNA of about 22 nucleotides in length and functions as part of the RNA-induced silencing complex. The mature miRNA has a seven nucleotide "seed region" that complementary base-pairs with the 3' untranslated region (UTR) of messenger RNA (mRNA). This process guides the RNA-induced silencing complex to specific transcripts, resulting in the degradation and therefore down-regulation of the targets. A single type of miRNA can have anywhere from one to thousands of mRNA targets, which establishes the potential for small changes in miRNA to have profound phenotypic effects.

Most miRNA is highly conserved across species, which makes nonconserved miRNA intriguing from an evolutionary perspective. Previous research has focused on differences in miRNA expression, but this technique would miss any phenotypic differences caused by changes in miRNA target specificity. Target specificity could change due to sequence differences in the seed region, or sequence differences that change the secondary structure of pre-miRNA.

To investigate such changes, I used high quality pre-miRNA sequences from seven primate species: Tamarin (*Saguinus midas*), Colobus (*Colobus angolensis*), Vervet (*Chlorocebus aethiops*), Rhesus Macaque (*Macaca mulatta*), Orangutan (*Pongo pygmaeus*), Chimpanzee (*Pan troglodytes*), and Human (*Homo sapiens*). This data was recently obtained through targeted capture methods designed for human exomes and miRNA and then sequenced using next generation sequencing. To discover if any specific miRNA are evolving rapidly, I compared the number of nucleotide differences between species per miRNA to the mutation rate of known neutrally evolving portions of the genome. I predicted miRNA secondary structure using the Vienna RNA Package, with differences in secondary structure measured by the subprogram RNAdistance. I found a number of divergent miRNA with striking structural differences, and the targets of divergent miRNA were enriched for genes involved in neural development. I was then able to compare these findings to miRNA from a dataset of 5,500 humans in order to discover any miRNA that were significantly divergent or conserved in the human lineage. These yet-to-be published results shed new light on the effect of miRNA on primate evolution and human diversity.

P-1143

Evolutionary mitogenomics of the phylum Mollusca

Daniela Almeida[1,2], Vítor Vasconcelos[1,2], Agostinho Antunes[1,2]
[1]CIMAR/CIIMAR - Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Rua dos Bragas, 177, 4050-123, Porto, Portugal,[2]FCUP - Faculty of Sciences of University of Porto, Department of Biology, University of Porto, Rua Campo Alegre s/n, 4169-007, Porto, Portugal

Mollusca is the second largest phylum of animals, with about 200,000 living species and a remarkable fossil record with 543 million years. Mollusks exhibit a wide range of behavioral, ecological adaptations and habitats, comprising some of the most diverse creatures in the sea and also being major components of terrestrial and freshwater ecosystems. Aside from their high importance for humans as food, art, pests and disease vectors, mollusks play an important role as model organisms in science, particularly in neurobiology and evolutionary biology. However, few is actually known about the molecular evolution of this phylum, with only two nuclear genomes (ngenomes) fully sequenced at the present, and 121 mitochondrial genomes (mtgenomes) available but that represent only approximately 0.06% of the mollusks species diversity. Gene synteny within mollusks mtgenomes display extraordinary differences, as a result of gene loss or gene duplication, changes in the position and strand specificity of transfer ribonucleic acid (tRNA) genes, protein-encoding genes and ribosomal ribonucleic acid (rRNA) genes. Hence, representatives of several classes of mollusks share remarkably few mitochondrial gene boundaries, with gene orders varying extensively even across closely related lineages. This contrasts with the highly conserved gene content and gene order in the vertebrates and arthropods. Further, some mollusks (orders: Mytiloida, Unionoida and Veneroida) possess an unusual system termed doubly uniparental inheritance (DUI) of mtDNA. Accordingly, females transmit their mtDNA to both sons and daughters (as in conventional strict maternal inheritance) but males pass on their mtDNA only to sons. Therefore, we analyzed the available molluscan mtgenomes to obtain insight into the controversial phylogeny of this phylum. We analyzed all the 13 protein-encoding genes - cytochrome c oxidase subunit 1-3 (cox1, cox2, cox3), cytochrome b (cytb), NADH dehydrogenase subunit 1-6 (nad1, nad2, nad3, nad4, nad5, nad6), NADH dehydrogenase subunit 4L (nad4L), ATPase F0 subunit 6 (atp6) and ATPase F0 subunit 8 (atp8) - for the available molluscan classes (Aplacophora, Bivalvia, Cephalopoda, Gastropoda, Polyplacophora, and Scaphopoda) to characterize the effect of mutation and gene duplication in the evolution of these species. Moreover, we performed detailed mitogenomic analyses to improve the conflicting phylogeny of these organisms.

P-1144

Characterisation of the skin immunones in the most archaic frogs in the world, *Leiopelma archeyi* and *L. hochstetteri*

Mette Lillie[1], Phillip Bishop[2], Dianne Gleeson[3], Katherine Belov[1]
[1]Faculty of Veterinary Science, University of Sydney, Sydney, Australia, [2]Department of Zoology, University of Otago, Dunedin, New Zealand, [3]Ecological Genetics Laboratory, Landcare Research, Auckland, New Zealand

The endemic frogs of New Zealand (Family: Leiopelmatidae; Genus: *Leiopelma*), along with the North American genus *Ascaphus*, are the most archaic lineage of modern frogs in the world. The four *Leiopelma spp*. are all classified as endangered by the IUCN with small, fragmented or declining populations. The pathogenic fungus, *Batrachochytrium dendrobatidis* (*Bd*), which causes the skin infection chytridiomycosis, has been a significant contributing factor in the declines and extinctions of amphibian species worldwide. This disease has also been detected in populations of *L. archeyi*. Although Bd has been implicated in population declines of this species, recent studies have suggested that *L. archeyi* are not susceptible to fatal infections of chytridiomycosis. Disease is also absent from populations of *L. hochstetteri* that are sympatric with infected populations of *L. archeyi*. Given the proposal that these species may have e resistance to this globally significant disease, we have sequenced the transcriptomes of the skin of *L. archeyi* and *L. hochstetteri*.

The transcriptomes of the ventral and dorsal skin from both frog species were separately sequenced on a GS Junior 454 Sequencing System (Roche). Two gene families of key interest include the major histocompatibility complex (MHC) and antimicrobial peptide genes. These immune gene families have been implicated in *Bd* resistance in other amphibian species. Blast databases were constructed from the assembled transcriptome contigs and queried with amphibian MHC gene sequences and antimicrobial peptide gene sequences available on Genbank (NCBI). MHC class I and II genes and a large number of antimicrobial peptides were identified in *L. hochstetteri*. MHC class II genes and antimicrobial peptides were identified in *L. archeyi*. We will present the characterisation and evolutionary analyses of these genes with a specific focus on their roles in the amphibian innate and adaptive immune response. This work will result in the development of adaptive immunogenetic population markers to better inform conservation management of these species with information on the diversity and differentiation at functionally significant loci.

P-1145

High-coverage population genomics of diverse African hunter-gatherers

Joseph Lachance[1], Benjamin Vernot[2], Clara Elbers[1], Bart Ferwerda[1], Alain Froment[3], Jean-Marie Bodo[4], Godfrey Lema[5], Thomas Nyambo[5], Timothy Rebbeck[1], Kun Zhang[6], Joshua Akey[2], Sarah Tishkoff[1]
[1]University of Pennsylvania, Philadelphia, PA, USA, [2]University of Washington, Seattle, WA, USA, [3]IRD-MNHN, Musée de l'Homme, Paris, France, [4]Ministère de la Recherche Scientifique et de l'Innovation, Yaoundé, Cameroon, [5]Muhimbili University College of Health Sciences, Dar es Salaam, Tanzania, [6]University of California at San Diego, San Diego, CA, USA

In addition to their distinctive subsistence patterns, African hunter-gatherers belong to some of the most genetically diverse populations on Earth. To infer demographic history and detect signatures of natural selection, we sequenced the whole genomes of five individuals in each of three geographically and linguistically diverse African hunter-gatherer populations at >60x coverage. In these 15 genomes we identify 13.4 million variants, many of which are novel, substantially increasing the set of known human variation. These variants result in allele frequency distributions that are free of SNP ascertainment bias. This genetic data is used to infer population divergence times and demographic history (including population bottlenecks and inbreeding). We find that natural selection continues to shape the genomes of hunter-gatherers, and that deleterious genetic variation is found at similar levels for hunter-gatherers and African populations with agricultural or pastoral subsistence patterns. In addition, the genomes of each hunter-gatherer population contain unique signatures of local adaptation. These highly-divergent genomic regions include genes involved in immunity, metabolism, olfactory and taste perception, reproduction, and wound healing.

P-1146

An overlooked function for Vaults - Lessons from phylogeny

Asfa Alli Shaik[1], Chengcheng Liu[2], Christopher Hogue[1,2]
[1]Department of Biological Sciences, National University of Singapore, Singapore, Singapore, [2]Center for Singapore-MIT Alliance, National University of Singapore, Singapore, Singapore, [3]Mechanobiology Institute of Singapore. National University of Singapore, Singapore, Singapore

Despite its discovery in 1986, the function of the vault complex, a 13 million Dalton macromolecule and the largest ribonucleoprotein complex, today still remains a mystery. A growing amount of data from diverse species and systems are available, yet the definition of precise vault function is still unknown. Hypotheses that vaults act as drug shuttle or nuclear pore plug or cargo carriers have been proposed, owing to its size and hollow interior, however, the arguments are still under debate. Our analyses of the major vault protein (*MVP*), that makes up 70% of this massive complex has revealed a unique and evolutionarily conserved nutritionally-biased amino acid composition. Phylogenomic study of vault complex suggests that evolution of vaults matches eukaryotic heterotrophs with ancestral loss of essential amino acid biosynthetic pathways, while loss of *MVP* in insects and nematodes seems to be widely complemented by gain of amino acid biosynthesis through gut intracellular bacterial endosymbionts. Vaults being present in high copy number and lacking other observable functions, we believe, may have been selected as a transportable cache of nutrient amino acids and ribonucleotides and a synthesis-turnover based nutrient absorption function may explain its abundance in oocytes, accumulation at the axons on neuronal excitation, nutrient phenotype in *Dictyostelium* and its role against intracellular bacterial pathogens.

P-1147

Investigating determinants of large-scale recombination rate variation using a recombination map of the human-chimpanzee ancestor.

Kasper Munch[1], Thomas Mailund[1], Julien Y Dutheil[2], Asger Hobolth[1], Mikkel H Schierup[1]
[1]*Bioinformatics Research Center, Aarhus University, Aarhus, Denmark,* [2]*Intitute of Evolutionary Sciences, University of Montpellier 2, Montpellier, France*

Most species and subspecies of apes have now been sequenced and the genomes of these closely related species are easily aligned. Along such an alignment, divergence times between species differ due to segregating polymorphism in the ancestral species. For some species the population size of the ancestral species is sufficiently large and the time span between speciation events is sufficient small that ancestral polymorphism may lead to gene trees with a topology different from the species tree. This phenomenon is termed incomplete lineage sorting (ILS) and implies that segments of the genome will display a closer relation to species other than the sister species. ILS is well established between human, chimpanzee and gorilla.

We apply a coalescent hidden Markov model where the hidden states along the alignment represent gene trees with separate topologies and separate coalescent times. Using simulations we establish that the recombination events constituting the transitions between different trees along the alignment almost exclusively occur in the human-chimpanzee ancestor allowing us to construct a recombination map of this ancestral species. We describe how this recombination map differs from that of extant humans and how these differences correlate with genomic features hypothesized to affect the large-scale variation in recombination rate across the genome.

P-1148

Search for antimicrobial peptide against fungal-infection in *Drosophila virilis*.

Yosuke Seto, Koichiro Tamura
Tokyo Metropolitan University, Hachioji, Tokyo, Japan

Antimicrobial peptides are essential components in *Drosophila* innate immune systems against bacteria and fungi. Expressions of a repertoire of antimicrobial peptides are important for an individual fly to survive under infection of microorganisms. For example, the mutant fly in which the expressions of all antimicrobial peptide genes were inhibited was killed by infection of microorganisms within a few days, whereas the wild type fly survived after the same infection. The survival rate of the fly in which the expressions of one or two antimicrobial peptide genes were rescued was raised to almost the wild type level. These results show that antimicrobial peptide plays an important role to defense against invading of microorganisms. To date, seven antimicrobial peptides have been identified in *D. melanogaster*, in which Metchnikowin and Drosomycin have been identified to have an antifungal activity. From the 12 Drosophila genome project, it was turned out that the metchnikowin gene is present in all 12 Drosophila species as a single-copy gene, whereas the drosomycin gene is found in only closely related species to *D. melanogaster* and composed of 7 copies. Although *D. virilis*, which is distantly related to *D. melanogaster*, does not have drosomycin genes, this species shows higher resistance to fungal infection than *D. melanogaster* shows. This observation motivated us to search for novel innate immune genes in *D. virilis* using pyrosequencing technology. As the results, we identify some candidate genes for new antimicrobial peptides.

P-1149

The sex-specific impact of meiotic recombination on nucleotide composition.

Alexandra Popa[1], Paul Samollow[2], Christian Gautier[1], Dominique Mouchiroud[1]
[1]Laboratoire de Biometrie et Biologie Evolutive, Villeurbanne, France, [2]Department of Veterinary Integrative Biosciences, Faculty of Genetics, Texas, USA

Meiotic recombination is an important evolutionary force shaping the nucleotide landscape of genomes. For most vertebrates, the frequency of recombination varies slightly or considerably between the sexes (heterochiasmy). In humans, male, rather than female, recombination rate has been found to be more highly correlated with the GC content across the genome. In the present study, we review the results in human and extend the examination of the evolutionary impact of heterochiasmy beyond primates to include four additional eutherian mammals (mouse, dog, pig, and sheep), a metatherian mammal (opossum), and a bird (chicken). Specifically, we compared sex-specific recombination rates with nucleotide substitution patterns evaluated in transposable elements. Our results, based on a comparative approach, reveal a great diversity in the relationship between heterochiasmy and nucleotide composition. We find that the stronger male impact on this relationship is a conserved feature of human, mouse, dog, and sheep. In contrast, variation in genomic GC content in pig and opossum is more strongly correlated with female, rather than male, recombination rate. Moreover, we show that the sex-differential impact of recombination is mainly driven by the chromosomal localization of recombination events. Based on our results from these seven vertebrate species, we propose a new explanation for the evolutionary impact of heterochiasmy on nucleotide composition. Independent of sex, the higher the recombination rate in a genomic region, and the longer this recombination activity is conserved in time, the stronger the bias in nucleotide substitution pattern, through such mechanisms as biased gene conversion. Over time this bias will increase the local GC content of the region.

P-1150

Theileria genome architecture: unique structure and divergence between species

Richard Perry, Paul Sharp
University of Edinburgh, Edinburgh, UK

Theileria species are tick-borne apicomplexan parasites of cattle, distantly related to *Plasmodium*. *Theileria parva and T. annulata* are closely related species causing East Coast fever in Africa, and theileriosis in Asia, respectively. The ~8 Mbp genomes of both species comprise four, highly syntenic, chromosomes [1,2]. Previous comparative analyses have attempted to identify (i) selective constraints on genes, and (ii) patterns of selected codon usage bias [3]. However, the results of these analyses appear to have been confounded by a very unusual genome architecture, with discrete regions of significantly divergent base composition. In both species, the background G+C content is 35% (all G+C values here refer to synonymously variable third codon positions within genes). The *T. parva* genome contains 15 regions, comprising 25% of the genome, where the average G+C is elevated to 42%. Remarkably, the orthologous regions in the *T. annulata* genome have a decreased average G+C value of 23%. Levels of divergence within these anomalous regions are twice those of the background regions. In both species, the unusual regions have a substantially increased frequency of tandem repeats. Recombination rate data available for *T. parva* reveal that the anomalous regions coincide with recombination hot-spots. In *T.annulata*, the anomalous regions are enriched for two specific DNA sequence motifs. The evolutionary origins of this remarkable pattern of genome architecture and divergence will be discussed.

[1] Pain et al., 2005, *Science* **309**:131-133

[2] Gardner et al., 2005, *Science* **309**:134-137

[3] Weir et al., 2009, *Infect. Genet. Evol.* **9**:453-61

P-1151

Genome-wide patterns of codon usage bias in sequenced flatworms

Heidi Aisala, Tiina M. Mattila, Jaakko Lumme
Department of Biology, University of Oulu, Oulu, Finland

The biased codon usage, observed in all kingdoms of life, suggests selective non-neutrality of synonymous changes. The fitness variation may be caused by translation efficiency, and therefore, highly expressed genes should contain the most optimal codons. Further, the efficiency of selection for optimal codon usage is influenced by mutational pressure, population structure and demographic factors. We used the predicted protein coding genes of species representing Trematoda, Cestoda, Monogenea and Turbellaria to contrast the patterns of codon usage. The flatworms have divergent reproductive systems (obligatory sexual vs. more or less parthenogenetic) and life histories (endoparasitic vs. ectoparasitic vs. free living). These variables are predicted to cause extreme differences in the long-term effective population size of the organisms, as well as in the recombination rate. The codon frequencies are also compared with the estimated numbers of tRNAs with different anticodons.

P-1152

Selective constraint beyond apparent sequence conservation

Olga Vakhrusheva[1,2], Georgii Bazykin[1,2], Alexey Kondrashov[1,3]
[1]Department of Bioengineering and Bioinformatics, M. V. Lomonosov Moscow State University, Moscow, Russia, [2]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia, [3]Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA

At moderate evolutionary distances, functional significance of non-coding sequences can be readily assessed through their above-random conservation between genomes. However, there is accumulating evidence that regulatory sequences are capable of rapid turnover, and, generally, conservation of sequence is not a necessary prerequisite for conservation of function.
For example, sequences at orthologous loci conserved among teleosts and mammals, but lacking significant sequence similarity between mammals and teleosts, can drive similar patterns of gene expression in zebrafish transgenic assays[1]. These findings may point to insufficiency of sequence similarity-based approaches to analysis of functional conservation. We used a bioinformatic approach to the question of whether functional conservation is possible without sequence conservation, and whether it can be inferred from genome wide comparative studies.
We considered two pairs of species chosen so that the evolutionary distance between species in a pair is significantly less than between species from two different pairs, though large enough to ensure that conservation in noncoding regions is mainly due to their functional role. We studied three such quartets, of dipterans, vertebrates, and chordates. We hypothesized that functional conservation between pairs of species should lead to above-average occurrence of conserved (within each pair) sequences at loci which are orthologous between pairs, even when no conservation between pairs is observed. Orthologous introns were chosen as sequence units within which conservation was analyzed, as their orthology can be inferred even at large phylogenetic distances through flanking exons.
We selected introns orthologous in all the four species (defined as the introns in orthologous positions of coding sequences in orthologous proteins). Then excluded from the analysis introns with significant local similarities between species from different pairs and, thus, considered only those introns that are unalignable between pairs. Nevertheless, there was a 2-4-fold excess of 4-sets of orthologous introns in which each species pair carried a conserved non-coding element, compared to the random expectation in dipteran, chordate and vertebrate species quartets. We also have shown a strong correlation between intron sequence similarity within one species pair and its retention in the other pair (again taking into consideration only introns unalignable between different pairs) for the dipteran quartet. These results indicate that selective constraint, presumably caused by retention of the ancestral function, often persists even in unalignable DNA segments.

1. Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L. & McCallion, A.S. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312, 276-279 (2006).

P-1153

Genomic insights into the evolution of the economically important monogenean flatworm Gyrodactylus salaris

Christoph Hahn, Philip D. Harris, Bastian Fromm, Tor A. Bakke, Lutz Bachmann
Natural History Museum, University of Oslo, Oslo, Norway

We present the first draft genome for the monogenean flatworm Gyrodactylus salaris, an economically important, notifiable ectoparasite of Atlantic salmon (Salmo salar) in Norway and Russia and a substantial threat to British and impaired Central European salmon populations. The high quality draft (>70 x coverage), the first for a monogenean species, was assembled based on data generated from both Roche 454 and Solexa next generation sequencing platforms. A bioinformatics pipeline was designed to detect and remove ubiquitous host-, as well as bacterial, contamination from the dataset. We identify single nucleotide polymorphisms (SNPs) in the diploid genome of G. salaris and compare a set of monogenean core genes to genes of the annotated genome of the digenean blood fluke Schistosoma mansoni. Furthermore, we generated low coverage (<10 x coverage) genomic data for several geographic strains of five closely related non-pathogenic gyrodactylids infecting salmonids, including G. thymalli (2 strains), G. truttae (2 strains), G. teuchis (3 strains) and G. derjavinoides, using a Solexa multiplex approach. Subsequent comparative genomic and mitogenomic analyses will focus on aspects of the genome likely to be associated with differences in pathogenicity and host specificity in Gyrodactylus species as well as the evolution of the pronounced progenetic lifestyle seen in the group. Our data and analyses will provide a foundation for studies of genome evolution in monogeneans, an important group of predominantly fish-parasitizing flatworms, which as the sister group to tapeworms contains a major portion of overall flatworm diversity.

P-1154

Signatures of genome-wide convergent molecular evolution: initial results.

Joe Parker[1], Georgia Tsagkogeorga[1], James A. Cotton[3], Elia Stupka[2], Stephen J. Rossiter[1]
[1]School of Biological and Chemical Sciences, Queen Mary, University of London, London, UK, [2]Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Milano, Italy, [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK

Recent studies have demonstrated that convergent sequence evolution can be detected in vertebrates using statistically robust phylogenetic methods that model parallel substitutions in genetic data. We are scaling this approach to a genomic level using 5 novel bat genomes and orthologous genes in over 30 other published genomes. Focusing on genes that have undergone convergence during the independent evolution of echolocation, we describe the computational algorithms and informatics framework we have developed, and our initial results. These early findings give promising indications that signatures of convergent molecular evolution are more prevalent in vertebrate genomes than previously recognised.

P-1155

Genome-wide association mapping of fertility traits in a natural house mouse hybrid zone

Bettina Harr[1], Leslie Turner[2]
[1]*Max-Planck-Institut fuer Evolutionsbiologie, Ploen, Germany,* [2]*University of Wisconsin, Madison, USA*

The house mouse hybrid zone is a natural laboratory in which the genetics of speciation can be studied. We have recently documented that hybrid male fertility is commonly reduced in hybrids that are collected directly in the wild. Here we use whole genome SNP chips (the mouse diversity array) to map genetic determinants of reduced hybrid fitness using techniques such as admixture mapping and classical genome-wide association tests. We identify several regions on the X chromosome and several regions on autosomes that are significantly associated with testis weight. Interaction analysis reveals a strong epistatic interaction between a region on chromosome 17, originating from subspecies *M. m. musculus* and a region on the X-chromosome, origination from *M. m. domesticus*, consistent with expectations of the Dobzhansky-Muller model of hybrid incompatibilities. We discuss candidate genes and potential mechanisms that lead to reproductive isolation and also the unique challenges and advantages of using the house mouse hybrid zone as a mapping population to answer important questions in speciation.

P-1156

Divergent resolution of duplicate genes among three *Paramecium* species following whole-genome duplication

Casey L. McGrath, Thomas G. Doak, Michael Lynch
Indiana University, Bloomington, IN, USA

The study of gene and genome duplications has long been of interest due to the potential evolutionary consequences of these events. Duplicate genes provide substrates for the evolution of novel genes and gene functions, and differential gene loss among subpopulations may lead to reproductive isolation and rapid speciation. The most recent genome duplication of *Paramecium tetraurelia* is thought to shortly predate the rise of the *P. aurelia* species complex, a group of 15 morphologically identical species. Additionally, 51% of the pre-duplication genes are still present in two copies in *P. tetraurelia*. Comparative genomic analysis of *P. aurelia* species, therefore, provides a unique opportunity to assess how a single genome duplication is resolved in multiple lineages.

We have completed macronuclear genome sequences for *P. biaurelia* and *P. sexaurelia*. Along with the published genome of *P. tetraurelia*, this allows us to investigate duplicate gene evolution in three lineages that share a whole-genome duplication. *P. biaurelia* and *P. tetraurelia* exhibit a surprisingly low rate of differential gene loss (~5% of all paralog pairs) given that the two taxa are considerably diverged (average dS = 0.70). *P. sexaurelia*, however, is even more distantly related (average dS from *P. tetraurelia* = 1.74), and exhibits substantial differential gene loss with *P. tetraurelia*. Indeed, gene retention/loss data indicate that *P. sexaurelia* may have diverged from the lineage leading to *P. tetraurelia* immediately following (or perhaps concurrent with) the whole-genome duplication. The pattern of gene loss among the three species, therefore, suggests that gene loss is rapid immediately following a whole-genome duplication but slows down dramatically afterwards.

By comparing the divergences of orthologs and paralogs among and within species, we see a significant effect of gene conversion acting to homogenize paralogs within lineages, and the strength of this effect appears to vary among species. Indeed, this effect is so pronounced that previous calculations based on *P. tetraurelia* paralog divergences alone appear to have underestimated the age of the whole-genome duplication by ~35%.

Finally, functional analysis of duplicate gene retention and loss supports the dosage balance hypothesis - the theory that following a whole-genome duplication, duplicate genes will be maintained due to stoichiometric constraints among interacting proteins. The sequencing of additional *P. aurelia* genomes in the future will add additional power to our analysis of the evolution of gene duplicates over time following a whole-genome duplication event.

P-1157

Capuchin monkey transcriptome provides insight into primate brain evolution

Amy M. Boddy[1], Mary Ann Raghanti[2], Chet Sherwood[3], Kimberley A. Phillips[4], Stephen H. Montgomery[5,6], Nicholas I. Mundy[6], Lawrence I. Grossman[1], Derek E. Wildman[1,7]
[1]Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, MI, USA, [2]Department of Anthropology, Kent State University, Kent, OH, USA, [3]Department of Anthropology, The George Washington University, Washington, DC, USA, [4]Department of Psychology, Trinity University, San Antonio, TX, USA, [5]Department of Genetics, Ecology & Evolution, University College London, London, UK, [6]Department of Zoology, University of Cambridge, Cambridge, UK, [7]Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, USA

Encephalization, defined as a higher than expected brain mass relative to total body mass, is a distinguishing trait in anthropoid primates compared to most other mammals. Neocortical expansion evolved early in the primate clade and has resulted in varying degrees of encephalization among anthropoids, with *Homo* as the most encephalized genus, and capuchin monkeys (Platyrrhini: *Cebus*), as the second most encephalized mammal genus. Interestingly, capuchin monkeys, like humans, exhibit many complex cognitive skills including tool use and complex social behavior. Evidence for an encephalized neocortex, along with the presence of those abilities, suggests capuchin monkeys represent an excellent taxon to investigate primate brain evolution; however, there has yet to be any large genomic studies published in *Cebus*. To address this question at the molecular level we present sequences derived from a neocortical transcriptome of a male tufted capuchin monkey, *Cebus apella*. We have collected and annotated over 60,000,000 high quality paired-end reads. Due to the absence of reference genomes for any *Cebus* species, we constructed a *de novo* assembly of the capuchin transcriptome. This assembly resulted in 78,737 contigs, with an N50 length of 882 bp and an L50 of 13,213. The mean contig size is 604 bp and the longest transcript assembled is 12,144 bp. To annotate the assembled transcripts, we used BLAST to compare capuchin sequences with those from the closely related marmoset genome, *Callithrix jacchus*. Hits with an e-value of less than 0.10, and a match length greater than 100 bp were accepted as putative homologous sequences. Using this strategy, the capuchin data were aligned to 34,103 *C. jacchus* Ensembl transcripts, which map to 14,914 marmoset genes. These data allow us to conduct large-scale analyses of protein-coding evolution using publically available transcript data from several primate species to identify genes that have undergone adaptive evolution on the capuchin lineage. We discuss the new brain transcriptome data in the context of adaptive protein evolution in the capuchin monkey with a focus on those genes that also underwent positive selection on other highly encephalized lineages, such as human.

P-1158

Selection on Wobble Sites in tRNA Anticodons in Vertebrate Mitochondrial Genomes with Codon Usage Reversal

Miguel M Fonseca[1,2], Sara Rocha[1,2], David Posada[2]
[1]CIBIO-UP, Research Center in Biodiversity and Genetic Resources, Porto, Portugal,[2]Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain

Vertebrate mitochondrial genomes usually have one transfer RNA (tRNA) for each synonymous codon family. This limited anticodon repertoire implies that each tRNA anticodon needs to wobble (a non-Watson-Crick base pairing between two nucleotides in RNA molecules) to recognize two or four synonymous codons. Different hypotheses have been proposed to explain the factors that determine the nucleotide composition of wobble sites of the mitochondrial tRNA anticodons. Until now, the two major hypotheses - the "codon-anticodon adaptation hypothesis" and the "wobble versatility hypothesis" - have not been formally tested in vertebrate mitochondrial data because both make the same predictions regarding the composition of anticodon wobble sites. The same is true for the more recent "wobble cost hypothesis".
In this study we have analyzed the frequencies of synonymous codons and tRNA anticodon wobble sites in 1553 complete vertebrate mitochondrial genomes, focusing specially on three fish species, whose mitogenomes present codon usage bias reversal (L-strand rich in T/G). These genomes constitute an excellent opportunity to study the evolution of the wobble nucleotide composition of tRNA anticodons because due to the reversal the predictions for the anticodon wobble sites differ between the existing hypotheses. We observed that none of the wobble sites of tRNA anticodons in these unusual mitochondrial genomes coevolved to match the new overall codon usage bias, suggesting that nucleotides at the wobble sites of tRNA anticodons in vertebrate mitochondrial genomes are determined by wobble versatility. Our results suggest that, at wobble sites of tRNA anticodons in vertebrate mitogenomes, selection favors the most versatile nucleotide in terms of wobble base-pairing stability and that wobble site composition is not influenced by the codon usage.

P-1159

Investigating the Impacts Recombination has on Origin Estimates

Crystal Hepp, Michael Rosenberg
Arizona State University, Tempe, AZ, USA

Despite the impacts that phylogenetic analyses have made on the estimation of geographical and temporal origins, relevant molecular evolutionary processes are frequently overlooked. Recombination is among these processes, due an integral assumption made within the framework of phylogenetic reconstruction; the segment of genetic information under scrutiny represents a single evolutionary history. Given that recombination can merge genomic segments from multiple entities (i.e. organisms derived from distinct geographical locations or hosts), a segment of genetic information that includes one or many recombination breakpoints is also representative of multiple evolutionary histories.

Haiti has long been suspected of playing an integral role in the successful spread of the worldwide HIV subtype B epidemic, resulting in a stigma carrying sociological and economic consequences. Molecular phylogenetic analyses of HIV-1 have shown that sequences from a limited number of individuals infected in Haiti form a paraphyletic and basal clade relative to other subtype B sequences (1, 2). A recent study clearly demonstrated a well-supported Haiti-first model, indicating that the worldwide subtype B epidemic is a result of a single migration event out of Haiti (3). Although multiple studies are in agreement regarding the ancestral placement of Haiti-originating sequences, they also collectively share a single major caveat: each disregards recombination.

We have reanalyzed a dataset composed of globally distributed HIV-1 subtype B sequences. When recombination breakpoint detection methods are used in conjunction with phylogenetic inference, the resulting topologies constructed from the majority of positions within the env gene support independently initiated introductions of the subtype B epidemic into Haiti and the rest of the world. One perplexing question, stemming from our analysis, is how the fusion of recombinant segments results in one clade systematically clustering both basal and paraphyletically to another when they are actually sister clades. I will address the fundamental effects that recombination has on clustering behavior and temporal estimates of artificially and naturally recombined HIV taxa.

1. W. H. Li, M. Tanimura, P. M. Sharp, Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5, 313 (Jul, 1988).
2. T. Gojobori et al., Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci U S A 87, 4108 (Jun, 1990).
3. M. T. Gilbert et al., The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A 104, 18566 (Nov 20, 2007).

P-1160

From Genome to Population Dynamics; the Glanville fritillary butterfly (*Melitaea cinxia*) as a model system

Rainer Lehtonen[1], Panu Somervuo[0,1], Lars Paulin[0,1], Leena Salmela[0,1], Virpi Ahola[0,1], Patrik Koskinen[0,1], Minna Taipale[0,2], Anne Duplouy[0,1], Pasi Rastas[0,1], Jouni Kvist[0,1], Swee Wong[0,1], Niko Välimäki[0,2], Jussi Nokso-koivisto[0,1], Veli Mäkinen[0,1], Mikko Frilander[0,1], Jussi Taipale[0,2], Esko Ukkonen[0,1], Liisa Holm[0,1], Petri Auvinen[0,1], Ilkka Hanski[0,1]
[1]University of helsinki, Helsinki, Finland, [2]Karolinska Instututet, Stocholm, Sweden

Our main research interests are the effect of spatial population structure on natural selection and demographic and evolutionary dynamics using the Glanville fritillary butterfly (*Melitaea cinxia*) as a model system. Long-term research conducted on the Glanville fritillary metapopulation in the Åland islands has yielded examples of landscape structure maintaining nonsynonymous polymorphisms in candidate genes, genetic effects on population dynamics as well as putative examples of coupling between demographic and evolutionary dynamics. The Glanville fritillary genome which have been sequenced using the newest Illumina, SOLiD and Roche 454 platforms, will be completed during this spring. In parallel with the assembly of the genome, we are reconstructing a genetic map by next-generation RADtag sequencing and targeted genotyping the parents and the offspring from carefully selected crosses between distant populations. We have conducted pioneering transcriptome-wide analyses of pooled population samples which were used to identify loci under selection using cross-population SNP allele comparison and Gene Ontology enrichment analysis. We have been able to find significant genetic associations to phenotypic and life-history traits, such as larval growth, flight metabolism and total reproductive output, using the carefully selected SNP panels. Additionally, we have already identified interesting phenotypic and genotypic differences between old vs. newly established populations, fragmented vs. continuous landscape and isolated single population vs. large metapopulation. The study metapopulations originate from two highly fragmented landscapes in Finland and Sweden, two extensive continuous landscapes in Estonia and Sweden, and one extremely isolated small population in Russia.

We are now extending the variation analysis to the whole transcriptome by ultra-deep Illumina RNA sequencing in 5 wild butterfly populations (~40 samples/population) living in different kinds of landscapes. We are aiming to gain knowledge about local adaptation, micro-evolutionary effects and gene regulations using SNP allele, splicing and gene expression variation data.

This project will yield rich molecular tools and material to study adaptation and evolution in contrasting environments at the genomic level as well as candidate genes for organismal studies.

P-1161

Exceptional sex: the evolution of meiosis when it breaks the rules

John Logsdon
University of Iowa, Iowa City, IA, USA

Meiosis is a highly conserved cellular process, and homologs of many meiotic genes are shared across diverse eukaryotes. In order to better understand the evolution and function of meiotic sex, we are investigating a number of eukaryotic lineages that exhibit unusual meiosis. These include male achaismatic *Drosophila* as well as putatively asexual snails, fungi and rotifers. Results from our ongoing analyses of the presence, absence and rates of evolution of meiotic genes in diverse organisms is revealing clues about the evolutionary history of this key eukaryotic innovation.

P-1162

The Amborella Genome: An Evolutionary Reference Sequence for Comparative Plant Genomics

Jim Leebens-Mack[1], Josh Der[2], Saravanaraj Ayyampalayam[1], James Burnett[5], Srikar Chamala[3], Andre Chanderbali[3], James Estill[1], Yuannian Jiao[2], Kun Liu[5], Tianying Lan[4], Eric Lyons[6], Lynn Tomsho[2], Hiabao Tang[8], Eric Wafula[2], Brandon Walts[3], Victor Albert[4], Brad Barbazuk[3], Hong Ma[2], David Sankoff[7], Stephan Schuster[2], Doug Soltis[3], Pam Soltis[3], Sue Wessler[5], Claude dePamphilis[2]
[1]University of Georgia, Athens, GA, USA,[2]Pennsylvania State University, University Park, PA, USA,[3]University of Florida, Gainesville, FL, USA,[4]University of Buffalo, Buffalo, NY, USA,[5]University of California, Riverside, Riverside, CA, USA,[6]University of Arizona, Tucson, AZ, USA,[7]University of Ottawa, Ottawa, ON, Canada,[8]J. Craig Venter Institute, Rockville, MD, USA

The Amborella Genome: An Evolutionary Reference Sequence for Comparative Plant Genomics

Comparative analyses have been shedding light on the drivers of gene family and genome evolution within some groups of flowering plants (i.e. core eudicots and cereal grains (Poaceae)), but the recent release of a draft genome sequence for *Amborella trichopoda*, the sole sister species to all other extant angiosperms is enabling more comprehensive analyses of angiosperm gene family and genome evolution (http://www.amborella.org). Cross-species comparisons of syntenic chromosomal segments indicate that all genome duplications that have been previously identified in eudocot and monocot genomes sequence analyses occurred after the divergence of lineages leading to *Amborella* and all other extant angiosperms. Inclusion of *Amborella* gene sequences in gene family analyses is elucidating the timing of gene duplication events relative to the origin and diversification of angiosperms. These results illustrate the utility of *Amborellla* gene and genome sequence data for understanding the origin and diversification of flowering plants.

P-1163

Phenotypic Evolution Driven by New Gene Origination

Manyuan Long
The University of Chicago, Chicago, Illinois, USA

The genetic basis for phenotypic evolution has been a major interest in evolutionary biology. The rapid growth in molecular and genetic analyses has provided powerful tools to dissect the phenotypic effects of genetic changes in model organisms, whereas the population genomics with deep sequencing in the model organisms have created unpresedented opportunity to examine the evolutionary forces responsible for phenotypic evolution. The new genes which originated in recent evolution in Drosophila and human have been investigated for their general roles in the evolution of important phenotypic or morphological traits, including sex dimorphisms, developmental process, central nervous systems and behaviors (e.g. Emerson et al, 2004, Science; Vibranovski et al, 2009, PLoS Genet; Chen et al, 2010, Science; Zhang et al, 2010, PLoS Biol; Zhang et al, 2011, PLoS Biol; Chen et al, 2012, Cell Rep). Integrative approach have been developed by combining genetic silencing techniques, tissue expression detecion, trait genetic analysis, population genetics and molecular evolutionary analyses and applied to the evolution of these phenotypic and morphological traits in Drosophila and humans. Unexpectedly critical roles of new genes were detected in the genetic control of these traits, revealing unusually rapid evolution in the traits and underlying genetic pathways under strong natural selection. The findings of species-specific and lineage-specific components in the genetic control of development, brain and behavior redefined a new paradigm to understand the phenotypic evolution.

P-1164

Tracing the green algal origin of embryophytes using a comparative mitochondrial genomics approach

Monique Turmel, Christian Otis, Claude Lemieux
Universite Laval, Quebec, Quebec, Canada

Six monophyletic groups of charophycean green algae are currently recognized within the Streptophyta: Mesostigmatales, Chlorokybales, Klebsormidiales, Zygnematales, Coleochaetales and Charales. Which one gave rise to the land plants? Based on a study of four genes from three cellular compartments, it was initially thought that the morphologically complex green algae belonging to the Charales are the closest relatives of land plants. However, a whole-chloroplast genome study and recent phylogenomic analyses of nuclear-encoded genes independently identified the Zygnematales or a clade consisting of the Zygnematales and Coleochaetales as sister to the land plants. In the present investigation, comparative mitochondrial genome analyses were undertaken in order to explore the evolutionary patterns of mitochondrial DNA (mtDNA) in the Streptophyta and determine whether a streptophyte mitochondrial phylogeny congruent with those derived from the chloroplast and nuclear phylogenomic data can be reconstructed.

As the Klebsormidiales and Zygnematales had not been previously sampled for their mitochondrial genome, we sequenced the mtDNAs of the klebsormidialeans *Entransia fimbriata* and *Microspora stagnorum* and of the zygnemateleans *Closterium baillyanum* and *Roya obtusa* and compared them to those previously reported for representatives of the four other streptophyte algal lineages as well as to mtDNAs of selected embryophytes and chlorophytes. The new sequence data confirm the notion that charophycean mtDNAs are less uniform in size than previously thought. Despite their five-fold size variation (42-202 kb), the eight analyzed charophycean mtDNAs exhibit a well conserved gene repertoire, which is similar to that found in liverworts and mosses. At the gene order level, the *Chara* and zygnematalean genomes are the charophycean mtDNAs most closely resembling those of bryophytes; however, neither algal genome displays striking similarity in intron content with land plant mtDNAs. Mitochondrial phylogenies were inferred from a data set of 40 concatenated proteins from eight charophyceans, 12 land plants and four chlorophytes using two different methods (maximum likelihood [ML] and Bayesian inference) and various models. The ML phylogenies inferred using LG+G4+F and GTR+G4 identified with robust support the Charales as sister to land plants. When the CAT-Poisson+G4 and CAT-GTR+G4 models were used, however, the level of support for this hypothesis dropped substantially and consistent with the abovementioned chloroplast and nuclear phylogenomic studies, an important fraction of the trees in bootstrap replicates recovered the Charales before the divergence of the Zygnematales and Coleochaetales. Importantly, this basal placement of the Charales received support from mitochondrial gene distribution and gene order data.

P-1165

Successive gain of insulator proteins in arthropod evolution

Peter Heger, Thomas Wiehe
Universitaet zu Koeln, Koeln, Germany

Chromatin insulation is mediated by insulator proteins and plays a fundamental role in organising gene expression. While a single insulator protein, CTCF (CCCTC-binding factor), is known in vertebrates, Drosophila melanogaster utilises six additional factors: Su(Hw), Mod(Mdg4), Gaga factor, ZW5, CP190, and BEAF-32. To explain the difference, we studied the phylogenetic distribution of these proteins. We found that all known chromatin insulators except CTCF are arthropod-specic and have been acquired successively during arthropod evolution. The full set of proteins is present exclusively in the genus Drosophila, whereas older insect orders and arachnids are equipped with only two known insulators, Su(Hw) and CTCF. In addition, our results suggest the loss of Su(Hw) and Gaga factor in several arthropod lineages. Thus, the D. melanogaster insulator mechanisms are not generally relevant in arthropods and so far unknown insulators may have evolved in other lineages.

P-1166

Searching - and finding - bilaterian specific proteins

Andrea Kraemer-Eis, Peter Heger, Thomas Wiehe
Universitaet zu Koeln, Koeln, Germany

Over 99% of all metazoa belong to the super-phylum of bilateria. In the early Cambrian essentially all fossilisable animal body plans and all today's phyla appear during a short period of a few Mill years. Since all evolutionary changes must have a molecular cause it seems reasonable to expect this to be true for the massive radiation of bilateria as well. Comparing the gene and protein repertoire of bilateria and searching for lineage specific proteins will help to identify potential molecular causes of the Cambrian explosion and to understand why bilaterians are so prevalent on earth today. We have taken a comparative genomics approach to identify proteins which are shared among bilateria, but absent in non-bilateria. Using extensive data mining approaches including reciprocal Blast, orthologue clustering, Gene Ontology annotation and literature searches we have obtained and analyzed several candidate clusters, composed mainly of transcription factors, binding proteins and proteins involved in cell differentiation and signalling. Interestingly, almost all are involved in developmental processes.

P-1167

Clustering Gene Trees Into Topological Classes

Kevin Gori, Christophe Dessimoz
European Bioinformatics Institute, Cambridge, UK

To uncover evolutionary relationships between species, we can no longer assume that all genetic loci in a genomic dataset support the same underlying tree topology: effects such as incomplete lineage sorting, introgression and horizontal transfer can cause incongruence to occur between gene trees. Here, we explore ways to identify multiple topologies present in the data by clustering trees reconstructed from individual loci into classes with common underlying topologies.

Using simulation and empirical data, we investigated which is the best combination of distance metric (Robinson-Foulds, Weighted Robinson-Foulds and Felsenstein's Branch Length Distance) and clustering method (single-linkage, complete-linkage, UPGMA and Ward's method) to partition the genes into topological classes. On real data, where the true partition is unknown, we computed trees for each cluster from concatenations of its constituent loci. We then used the sum of log-likelihoods of the resulting cluster trees as a scoring function to assess the relative performance of each method.

Finally, we explored ways of estimating the optimal number of topologies, using information criteria and other model selection techniques.

To build on these results we propose an iterative optimisation procedure, using the sum of log-likelihoods as an objective function, to further improve the partitioning.

P-1168

The use of mitochondrial data to infer the mammalian phylogeny: a cautionary tale.

Claire C. Morgan[1,2], Christopher J. Creevey[3], Paul Kilroy-Glynn[4], Mary J. O' Connell[1,2]
[1]Dublin City University, Dublin, Ireland,[2]Centre for Scientific Computing & Complex Systems Modeling (SCI-SYM), Dublin, Ireland,[3]Teagasc, Co. Meath, Ireland,[4]Comparative Genomics Unit NIH, Washington DC, USA

The number of fully sequenced mammal genomes restricts extensive taxon sampling within Superorders. To increase the number of mammals sampled, mitochondrial data has been employed. Mitochondrial genes have been used to resolve the phylogenetic relationships at the root of the mammalian phylogeny and also the more shallow relationships such as those within the Cetacea and the Rodentia. However, it has been demonstrated that in a given time frame mtDNA undergoes more mutations compared to nuclear DNA - leading to saturation for base changes. This results in corruption of phylogenetic signal. Here we have assessed the suitability of mitochondrial genes as phylogenetic markers in the mammal tree of life. We find that mitochondrial data is not suitable for the resolution of the basal positions in mammalian phylogeny. However, we find a trend towards improved signal upon removal of saturated sites and highly divergent taxa. We find that phylogenetic signal improves as gene coverage increases across the taxa sampled and on dealing individually with Suborders. Our study highlights the importance of having suitable data to address the phylogenetic question at hand.

P-1169

Phylogenomic analysis of myzostomid transcriptome data supports an annelid origin of myzostomids

Stefanie Hartmann[1], Christoph Bleidorn[2]
[1]University of Potsdam, Potsdam, Germany, [2]University of Leipzig, Leipzig, Germany

In trying to understand the evolutionary relationships of organisms, the current flood of sequence data offers great opportunities, but also reveals new challenges with regard to data quality, the selection of data for subsequent analysis, and the automation of steps that were once done manually for single-gene analyses. Even though genome or transcriptome data is available for representatives of most bilaterian phyla, some enigmatic taxa still have an uncertain position in the animal tree of life. This is especially true for myzostomids, a group of symbiotic (or parasitic) protostomes that are either placed with annelids or flatworms.

Based on similarity criteria, Illumina-based transcriptome sequences of one myzostomid were compared to protein sequences of one additional myzostomid and 29 reference metazoa and clustered into gene families. These families were then used to investigate the phylogenetic position of Myzostomida using different approaches: Alignments of 989 sequence families were concatenated, and the resulting superalignment was analyzed under a Maximum Likelihood criterion. We also used all 1,878 gene trees with at least one myzostomid sequence for a supertree approach: the individual gene trees were computed and then reconciled into a species tree using gene tree parsimony.

Superalignments require strictly orthologous genes, and both the gene selection and the widely varying amount of data available for different taxa in our dataset may cause anomalous placements and low bootstrap support. In contrast, gene tree parsimony is designed to accommodate multilocus gene families and therefore allows a much more comprehensive data set to be analyzed. Results of out supertree approach showed a well-resolved phylogeny, in which myzostomids were part of the annelid radiation, and major bilaterian taxa were found to be monophyletic.

P-1170

The evolution of cold tolerance in *Drosophila*: from phylogenetics to gene expression

Ramiro Morales-Hojas, Micael Reis, Cristina P. Vieira, Jorge Vieira
IBMC - University of Porto, Porto, Portugal

We present a study of the evolution of climate adaptation in *Drosophila* within a phylogenetic context. The goal is to investigate the basis of adaptation to different climatic conditions and to understand why some species are capable to expand their ranges to other latitudes. For this purpose we have first estimated the phylogeny of 122 species of the subgenus *Sophophora* including representatives of all the major lineages. We used GenBank sequence data from six genes (*COI*, *COII*, *Adh*, *hunchback*, *Gpdh* and *Amyrel*) and two phylogenetic reconstruction methods (ML and BI) with partitions. Both methods result in a highly congruent phylogeny. Ancestral reconstruction of climatic distribution done using Bayesian and Likelihood methods (SIMMAP and Mesquite, respectively) indicate that the ancestral species of the *Sophophora* had a tropical distribution. Furthermore, adaptation to temperate climatic conditions has occurred several times independently. Some lineages have shifted their distribution to temperate latitudes and are not found in tropical regions. However, other species have expanded their ranges to temperate areas while retaining the tropical distribution as well (cosmopolitan species). In order to investigate the behavioural and molecular basis of this difference in climatic range tolerance, we have performed chill-coma recovery experiments and gene expression assays of the candidate gene *frost* in 10 species with different climatic distribution. For cosmopolitan species, tropical and temperate strains were tested. Results show that tropical and cosmopolitan species have a positive correlation between cold exposure time and chill-coma recovery time, with cosmopolitan species showing a lower slope. This may explain why some species have been able to colonise temperate regions and others not. In contrast, temperate species show no correlation between cold exposure time and recovery time. Fold change in expression of *frost* after cold exposure is also discussed.

P-1171

Monophyly and placement of the *Drosophila inca* species subgroup suggests an early South American diversification of the *Drosophila repleta* lineage.

Andrea Acurio[1], Deodoro Oliveira[1], Violeta Rafael[2], Alfredo Ruiz[1]
[1]*Universitat Autònoma de Barcelona, Barcelona, Spain,* [2]*Laboratorio de Genética Evolutiva, Escuela de Ciencias Biológicas, Pontificia Universidad Católica del Ecuador, Quito, Ecuador*

We generated a phylogeny of the *Drosophila repleta* group using molecular characters and included for the first time three species belonging to the *inca* subgroup. The *inca* subgroup is the less well-known of the six species subgroups of the *repleta* group. It was defined in 1989 to include three cactophilic species native to Ecuador, *D. inca*, *D. huancavilcae* and *D. yangana.* Although the affiliation of these three species was apparent based on shared morphological traits, no molecular phylogenetic approach had been made to confirm the monophyly of the *inca* subgroup. Further, the position of the *inca* subgroup inside the *repleta* group was unclear. Our study recovered a monophyletic, basal *inca subgroup*, a result that sheds light on the origin and early diversification of the *repleta* group. Collections of wild flies were performed on desertic habitats of North Coast, Central and South of Ecuador. Sequences of two mitochondrial (COI, COII) and two nuclear genes (Marf, SinA) were generated. Sequences were concatenated to obtain a 2348-bp long combined matrix for 44 *Drosophila* species. The dataset includes selected representatives from the other five *repleta* species subgroups (*mulleri*, *fasciola*, *hydei*, *mercatorum, repleta*) taken from Oliveira *et al*. (in review). Using BEAST, we set different substitution models and molecular clocks for the nuclear and the mitochondrial partitions. The nucleotide substitution model to mitochondrial partitions was GTR, with empirical base frequencies plus Gamma model of site heterogeneity (four categories) with 2 partitions into codon positions (1+2), 3. The nuclear partition has the same settings but without codon partition. We used a strict clock with universal rate to mitochondrial partition, and estimated rate to nuclear partition. The Yule process speciation was used as prior. A MCMC run with 10 millions generations was performed sampled every 1000 generations. Measures of effective sample sizes were determined with TRACER. The same dataset was used to build a Maximum Likelihood tree in order to test different topologies. Both ML and partitioned Bayesian analyses produce single tress with the same well-supported topology. The three *inca* species: *D. inca*, *D. huancavilcae* and *D. yangana* form a monophyletic clade distant from the other five *repleta* species subgroups. Our results agree with the assertion of previous taxonomic and cytological work argued that the *inca* species subgroup was one of the most basal subgroups in the radiation of *repleta* group suggesting an early South American diversification of the *Drosophila repleta* lineage.

P-1172

Whole-genome sequences of 6 pigs species shed light on natural and Human-mediated migration routes across insular South-East Asia.

Laurent Frantz[1], Joshua Schraiber[2], Ole Madsen[1], Hendrik-Jan Megens[1], Mirte Bosse[1], Yogesh Paudel[1], Richard Crooijmans[1], Greger Larson[3], Montgomery Slatkin[2], Martien Groenen[1]
[1]Wageningen University, Wageningen, The Netherlands,[2]University of California, Berkeley, Berkeley, USA,[3]Durham University, Durham, UK

Next-generation sequencing enables sequencing of complete genomes at an affordable price. However, genome sequence analysis presents several challenges in particular with the aim to improve our understanding of present day biodiversity. In this study, we use 10 whole-genome sequences of 6 species from the genus Sus (Cetartiodactyla; pigs and related species) from insular South-East Asia (ISEA) to explore this question. The Plio-Pleistocene glaciation cycles that repeatedly isolated and connected islands in ISEA created the perfect environment for allopatric speciation and shaped the repartition of pigs in the region. Moreover, Human migrations were often accompanied by pigs which influenced their dispersal. This tight link between pigs and humans has provided information about our early history. However, the phylogeny and phylogeography of these species is debatable and little is known about the influence of anthropogenic activities on their dispersal. Despite several challenges arising from phylogenomic methods, incomplete lineage sorting and possible admixture, we show that previous phylogenetic inference have been misled by the use of small genetic markers. Our results demonstrate the importance of combining both supermatix and supertree methods to resolve difficult phylogenies. The wealth of information contained in those sequences allowed us to show the impact of past climatic fluctuation on the demography of these species. Using a hidden-markov-model (HMM) we were able to show the drastic effect of the sharp mid-Pleistocene climatic transition on the population size of our species. We applied a novel method to detect admixture across our samples in order to identify new migration routes within ISEA that could not be detected solely with a phylogenetic tree. Our results show conclusive evidence of facilitated migration across the shallow sea of the Sunda-Shelf (Borneo-Malaysia-Sumatra) due to its repeated exposure during glacial periods. Moreover, our analysis shows a possible early human-mediated displacement, across all ISEA and mainland Asia, of the species inhabiting Sulawesi. We interpret this result as a possible new center of domestication. Finally, we were able to show that during the last centuries, Europeans have brought and released pigs in this area. The result of our study shed light on the complex natural history of the genus and the influence of anthropogenic activities throughout our history. Together, our findings highlight the importance of phylogenomic methods to comprehensively infer the natural history of present day biodiversity.

P-1173

Performance Degradation in Molecular Divergence Time Estimation with t he Increase in Calibrations with Large Uncertainty

Kazutaka Takeshita, Hidemi Watanabe
Hokkaido University, Sapporo, Japan

The crucial role of calibration in molecular divergence time estimation has been well recognized and has attracted a lot of interest. Most of the molecular divergence time estimations have been based on fossil-based calibrations with a wide range of uncertainty. It is an often employed for molecular divergence time estimation to use as many calibrations as possible in anticipation of improving the accuracy with more calibrations. However, it has not been proved if the accuracy is improved with the increase in the number of calibrations regardless of calibration uncertainty. In this study, we assessed the impacts of changes in calibration in terms of both number and uncertainty on molecular divergence time estimation with sequence data produced along a given phylogeny. As a set of the calibration uncertainties to be tested, we used two types of calibration, small- and large-uncertainty models. The former was specified with a uniform distribution with minimum and maximum bounds of a small range and the latter with a uniform distribution with a minimum bound and an arbitrary large value as a maximum bound similarly to a model for fossil-based calibrations. This simulation study clearly demonstrated that the increase in the number of calibrations did not improve or even worsened the accuracy of the divergence time estimations when the calibrations were based on the large-uncertainty model, while those based on the small-uncertainty model significantly improved the accuracy, and that the estimations with the small-uncertainty model were more accurate than with the large-uncertainty model when the same number of calibrations was used. In addition, as expected, the smaller the uncertainty range was, the more accurate the estimate was. Therefore, we concluded that it is strongly encouraged to employ calibrations with small uncertainties, instead of many calibrations with large uncertainties, even if only a small number of such calibrations are available.

P-1174

Impact of model selection across the Amphibian Tree of Life

Karen Siu Ting[1], Mark Wilkinson[2]
[1]National University of Ireland, Maynooth, Ireland, [2]The Natural History Museum, London, UK

Living amphibians (Class Amphibia) comprise three main groups, frogs (Order Anura), salamanders (Order Caudata) and caecilians (Order Gymnophiona) each of which are well-supported monophyla as is the Batrachia - the group comprising frogs and salamanders. Broad phylogenetic studies of amphibians have included progressively larger numbers of taxa. Large-scale phylogenetic analyses confront some well-known issues of problem size due simply to the number of taxa. Additionally, large-scale phylogenetic analyses may present additional challenges related to the breadth (diversity) of the taxon sampling, the diversity of markers, and the distribution and abundance of missing data in multigene data sets where taxon sampling is incomplete for some or all of the genes. Using amphibians as an example, we present a study on the impact of taxon sampling upon model selection, model accuracy, and on inferred phylogenetic relationships.

P-1175

The identification of putative dihydrofolate reductase (DHFR) duplicates and their phylogenetic distribution

Robert Carton, Gráinne McEntee, Anne Parle-McDermott, Mary O'Connell
Dublin City University, Dublin, Ireland

Dihydrofolate reductase (DHFR) is an enzyme essential for the conversion of dihydrofolate to tetrahydrofolate. As such, this enzyme plays a key role in DNA synthesis. Until recently it was thought that DHFR was encoded by a single gene. A number of processed pseudogenes have since been identified in human and one of these (DHFRL1) has been shown to serve the same function as DHFR (McEntee et al. 2011). This discovery raises an important question: are there multiple functional copies of this essential enzyme in organisms such as our model species mouse and rat? We have identified additional DHFR like sequences (DHFRLSs) across the *vertebrata* and have investigated their phylogenetic distribution. From our analysis of the phylogenetic and syntenic distribution of these newly identified DHFRLSs, we can show that the known pseudogenes in human have arisen from separate retrotransposition events. The human DHFRLS, known as DHFRL1, was derived from a duplication that occurred within the primates, most likely in the *Catarrhini* ancestor, and is a functional gene. Other DHFRLSs we have identified have emerged in many species across the *Eutheria* in multiple independent duplication events. We predict that the DHFR gene has duplicated on at least four separate occasions, producing many processed pseudogenes – warranting further functional assessment.

Reference

McEntee G, Minguzzi S, O'Brien K, Ben Larbi N, Loscher C, Ó'Fágáin C & Parle-McDermott A (2011). The former annotated pseudogene dihydrofolate reductase-like 1(DHFRL1) is expressed and functional. PNAS, 108(37): p.15157–15162.

P-1176

Testing the status of infra generic taxonomic ranks in Triatomini (Hemiptera: Reduviidae)

Silvia Justi, Claudia Russo
Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Triatominae (Hemiptera: Reduviidae) diversity comprises hematofagous assassin bugs that are potential vectors of the Chagas Disease. The *Triatoma* genus alone includes more than half of the subfamily diversity and it has been divided into species complexes and subcomplexes. These infrageneric taxonomic ranks, however, reflect geographical distribution or questionable shared morphological traits. No rigorous phylogenetic investigation has been performed to ascertain the formal taxonomic status of these ranks. Here, we investigate the validity of the species complex Infestans and its six species subcomplexes (Brasiliensis, Infestans, Maculata, Matogrossensis, Rubrovaria and Sordida). Since many cryptic species are currently being investigated in this important genus, we have decided to include specimens from different localities along South, Central and North America to test the specific status of the 35 putative species of Triatomini. The data set includes four mitochondrial (COI, COII, CytB and 16S) and one nuclear (18S) fragments, in a total of 2970 bp. A maximum likelihood tree was the basis for our phylogenetic hypotheses with the GTR model. Bootstrap test was used to evaluate the repeatability of our clusters. Our results show that Infestans is not a natural complex, as two species assigned to the Brasiliensis subcomplex clustered elsewhere. First, *T. tibiamaculata* appears as sister taxa to *Panstrongylus megistus*, and second *T. vitticeps* falls outside the clade containing the remaining Infestans species. On the other hand, Infestans and Rubrovaria seem to be natural subcomplexes. Part of the Matogrossensis diversity clusters within Sordida subcomplex whilst the remaining Matogrossensis is tightly grouped with a Maculata lineage. Additionally, it was concluded that specimens formally recognized as *T. sordida* are in fact a paraphyletic cryptic species assemblage. Finally, diversity within *T. guasayana*, *T. costalimai*, *T. vandae* and *T. matogrossensis* are not clustered monophyletically. These results alone confirm what is long known by Triatominae specialists: an extensive and thorough systematic revision of the tribe taking into account morphological and molecular data is long overdue.

P-1177

Coral reef fishes: the origins of biodiversity hotspots and biogeographic patterns

Peter Cowman[1,2], David Bellwood[1,2]
[1]School of Marine and Tropical Biology, James Cook University, Townsville, Queensland, Australia, [2]Australian Research Council Centre of Excellence for Coral Reef Studies, Townsville, Queensland, Australia

The world's largest marine biodiversity hotspot is centred in the Indo-Australian Archipelago (IAA). To date, much of the study on this hotspot has concentrated on processes that maintain species richness within the hotspot. Less emphasis has been placed on the origin of the hotspot and evolution of taxa that form the hotspot. To address this issue, we used phylogenetic methods and age estimation techniques to explore the origins of four diverse reef fish families (Labridae, Pomacentridae, Apogonidae, Chaetodontidae) that form much of the biodiversity found both in the hotspot and on tropical coral reefs around the globe. We examine patterns of origination, diversification and ancestral biogeography to explore the links between the evolutionary history of coral reefs and their associated fish fauna. Examination of lineage through time plots reveal a possible late Eocene/early Oligocene cryptic extinction event coinciding with the collapse of an ancestral Tethyan/Arabian palaeodiversity hotspot. Rates of diversification analysis reveal elevated cladogenesis in all four families in the Oligocene/Miocene. In this period, lineages with a high percentage of coral reef associated taxa display significantly higher diversities and net diversification rates than expected. Patterns of diversification among taxa also suggest that coral reefs may have acted as a refuge from high extinction, as reef taxa were able to sustain significantly higher diversities at higher simulated extinction rates than non-reef counterparts. Biogeographic reconstruction of the Labridae, Pomacentridae and Chaetodontidae reveal marked temporal congruence in origination and dispersal between the East Pacific, Atlantic, Indian Ocean, the IAA hotspot and Central Pacific regions. The East Pacific and Atlantic have a history of isolation, developing from broader connectivity with the Indo-Pacific from the early Eocene. The IAA has a history of connectivity with adjacent regions. It has sequentially and then simultaneously acted as a centre of accumulation (Palaeocene/Eocene onwards), survival (Eocene/Oligocene), origin (Miocene onwards), and export (Pliocene/Recent) for reef fishes. While association with coral reefs may provide a mechanism for the cladogenesis of several reef fish lineages in the Miocene, it appears to be the unique attribute of expanding reef habitat in the mosaic of island archipelagos in the IAA hotspot that has allowed the survival, proliferation and expansion of coral reef fish lineages to form the hotspot that wee see today.

P-1178

Ignoring heterozygosity biases phylogenomic estimates of divergence times

Heidi E.L. Lischer[1,2], Laurent Excoffier[1,2], Gerald Heckel[1,2]
[1]CMPG, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland, [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Phylogenetic reconstruction of closely related species may be difficult, as they contain potentially unsorted lineages and a relatively high proportion of heterozygous sites that are usually not handled in phylogenetic methods. Genomic analyses promise to provide a sufficiently large number of fixed differences to resolve branching orders, but the analysis of the sequences of diploid individuals is still challenging. Here we performed a phylogenomic study of the influence of heterozygosity on the estimation of intra-specific divergence times between allopatric populations of the common vole (Microtus arvalis). We used the Roche 454 Titanium technology to generate genome-wide sequence information from AFLP fragments analysed in 15 voles belonging to old evolutionary lineages distributed across Europe. To evaluate the impact of heterozygosity on phylogenetic parameter estimates, we used a procedure involving repeated random sampling of haplotypes generated from sequences with multiple heterozygous sites, followed by maximum-likelihood and Bayesian tree estimations of the partitioned data set. We compared these results with phylogenetic trees inferred from data encoding heterozygous sites with ambiguity codes and from data totally excluding these positions. All results show a clustering of individuals in four deep evolutionary lineages, but the incorporation of information from alternative alleles at heterozygous sites has a significant impact on absolute and relative branch lengths. Thus the exclusion of heterozygous sites from evolutionary analyses may lead to biased and misleading estimations of divergence times particularly for closely related taxa. With the increasing accessibility and availability of genome-wide data there is a growing need for a better handling and integration of diploid sequence information into phylogenetic methods.

P-1179

Phylogenetic reconstruction of malaria vectors in Thailand using multilocus DNA sequences

Uraiwan Arunyawat, Prin Phunngam, Theeraphap Chareonviriyaphap
Kasetsart University, Bangkok, Thailand

Malaria spreads through mosquitoes, belonging to *Anopheles* genus. Understanding the evolutionary and taxonomic status of closely related malaria vector species is the initial step in malaria vector control program. In this study, we perform four different approaches (neighbor joining, minimum evolution, maximum likelihood and maximum parsimony) to reconstruct phylogenetic trees of main malaria vectors presented in Thailand based on sequence information of five DNA fragments from both nuclear and mitochondrial regions. Our results reveal clear evidence that *Anopheles* species separate into three distinct clades; Dirus group, Minimus group and Maculatus group. Interestingly, phylogenetic trees based on different reconstructed algorithms and different gene regions provide congruent phylogenetic status of studied mosquito species. The phylogenetic relationship of these malaria vector species follow the pattern based on morphological identification. Moreover, estimation of divergence time among studied species inferred that *Anopheles* species probably have occurred during Eocene to early Plesitocene.

P-1180

A Family Affair: Resolving the taxonomic position of the Old World Leaf-Nosed Bats, the hipposiderids

Nicole Foley, Sebastien Puechmaille, Emma Teeling
University College Dublin, Dublin, Ireland

The hipposiderids comprise 9 genera and 69 species. In stark contrast, their nearest relatives the rhinolophids are monogeneric consisting of a single genus, Rhinolophus which comprises 62 species. These two groups of bats are most notable for their unique form of echolocation, which unlike others bats involves nasal emission of echolocation calls. Within the phylogeny of the Chiroptera the relationship between the hipposiderids and the rhinolophids remains a source of taxonomic conflict. Some regard these bats as two separate families while others argue that the hipposiderids represent a sub-family of the Rhinolophidae. While extreme morphological similarities has prevented morphological approaches from resolving this taxonomic conflict, molecular studies to date have failed to account for the entire generic diversity of these groups. Determining the relationship between these two closely related groups of bats has implications for conservation and will be pivotal in furthering our understanding of the evolution of echolocation and the evolution of SARS. Data from nuclear introns (9) and exons (13) were used to generate the first resolved phylogeny that comprises species representing all hipposiderid and rhinolophid genera. Phylogenetic analysis revealed strong support for the monophyly of the hipposiderids and rhinolophids respectively. Contrary to previous findings, the monophyly of the hipposiderids most speciose genus, Hipposideros, was not supported. Molecular Clock Dating combined with appropriate fossil constraints was used to date the divergence times between all major nodes on the tree. Character State Mapping was used to reveal the biogeographical origin of the hipposiderids and rhinolophids respectively. In combination, these analyses hint that diversity within the taxonomic ranks of the hipposiderids is not fully described by current taxonomic classifications.

P-1181

Molecular phylogenetics of Passerine birds: Sibley and Ahlquist's Tapestry reloaded

Alexandre Pedro Selvatti, Claudia Russo
Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Passerines represent nearly half of the avian extant diversity. Hence, it is not surprising that it still holds major phylogenetic ambiguities even after decades of molecular phylogenetics. Here, we present the largest molecular sequence dataset to date, as 1.119 genera (12,575 bp) were analysed in a likelihood framework. The order Passeriformes is divided into the two well-established suborders, Oscines and Suboscines. One exception is the *incertae sedis Acanthisitta* genus for which phylogenetic position varies with analysis. In our topology, suborders were recovered as clades, but *Acanthisitta* was sister to the remaining passerines. Among Suboscines, Old World (OW, Eurylaimides) and New World (NW, Tyrannides) groups were monophyletic. The enigmatic NW *Sapayoa*, however, was included in the OW clade. Tyrannides were divided in two monophyletic lineages, the Furnariida (including Thamnophilida) and the Tyrannida. On the other hand, Oscines are traditionally divided in two parvorders (and six superfamilies): Corvida (Corvoidea, Meliphagoidea, Menuroidea) and Passerida (Muscicapoidea, Passeroidea, Sylvioidea). Corvida diversity appears as a paraphyletic assemblage with a monophyletic Passerida clustered within. Corvidan lineages are mostly endemic to the Australo-Papuan region and secondary radiations in Africa, Eurasia, and NW are noticeable. In this parvorder, Meliphagoidea was strictly recovered, but a "core Corvoidea" could also be distinguished. Conversely, the Menuroidea were paraphyletic, with Menurae as sister to Oscines and the remaining Menuroidea clustered in the following lineage to diverge. Also, as previously reported, the problematic *Chaetops*, *Picathartes* and *Eupetes* seem closer to Passerida than to the other Corvida. The geographical pattern for Passeridan origin is not clear-cut, as these genera are endemic to small but remarkably disjunct OW regions (Africa, Indonesia). Among Passerida, the three superfamilies were each recovered as monophyletic groups with Sylvioidea as sister to the Passeroidea and Muscicapoidea clade. The latter clade was probably originated in Eurasia, but it exhibits a NW radiation that is unique among the oscines. In this study, we show that Sibley and Ahlquist's subdivisions to be largely consistent with our molecular phylogenetic results, particularly within Passerida. Since that groundbreaking study, the amount of information increased and diversified public data banks, setting the stage for a consistent Passeriformes tree of life.

P-1182

Bioinformatic Challenges in the 1KITE Project

Alexander Donath
Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany

1KITE (1K Insect Transcriptome Evolution) is an international research initiative that aims to study the transcriptomes of 1,000 insect species encompassing all recognized insect orders. The expected sequence data will allow inferring for the first time a robust phylogenetic backbone tree of insects, one of the most species-rich groups of metazoan organisms. Preliminary analyses of the already available sequences show that the obtained data are of yet unparalleled size and quality.
Thus, 1KITE not only shows the need for but explicitly includes the development of new software for data quality assessment, phylogenetic reconstruction, and molecular dating that will allow for advanced and accelerated analyses of such large amounts of sequence data.

P-1183

Gene transpositions in the primate genomes

Mira Han
NESCent, Durham NC, USA

segmental duplications and genome rearrangements have been well studied in the primates, but these studies have not examined the evolution of the genes involved in the rearrangements. Using a map of conserved anchors across the primate genomes, I've reconstructed the ancestral distribution of gene families on the primate phylogeny. The ancestral reconstruction allows estimation of the rate of transposition in the unit of genes and identification of transposed genes along the phylogeny. After identifying the genes, I studied the functions, and sequence evolution of the transposed genes.

P-1184

On the tree of life size distribution: A generalization of Yule's model of genera size (how many species are in a genus) to include species extinction.

Yosef Maruvka[1,2], Nadav Shnerb[3], David Kessler[3], Robert Ricklefs[4]
[1]Department of Biostatistics and Computational Biology Dana-Farber Cancer Institute, Boston, MA, USA, [2]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA, [3]Department of Physics, Bar-Ilan University, Ramat-Gan , Israel, Israel, [4]Department of Biology, University of Missouri St. Louis, St. Louis, MO, USA

The highly skewed distribution of species among genera poses a continuing challenge to macroevolutionists, but also an opportunity to understand the dynamics of diversification. Minimal models considered so far have been based either on Yule's (1925) work, which neglects extinction, or on a simple birth-death (speciation-extinction) process. Here we present a generic, neutral, speciation-extinction (of species)-origination (of genera) (SEO) model for macroevolutionary dynamics of taxonomic diversification. This new model fits well the observed species-per-genus distribution for class-to-kingdom sized taxonomic groups. The model's predictions for the appearance times (the time of the first existing species in this group) of the taxonomic groups also approximate match estimates based on molecular inference and fossil records, along some other predictions. Finally, fitted extinction rates for large clades are close to speciation rates, consistent with high rates of species turnover observed in the fossil record.

[1] **Yosef E. Maruvka**, Nadav M. Shnerb, David A. Kessler. (2010) J Theor Biol 262(2) 245-56).

[2] **Yosef E. Maruvka**, Nadav M Shnerb, David A. Kessler. (2011) PLoS ONE 6(11)

[3] **Yosef E. Maruvka**, Nadav M. Shnerb, David A. Kessler, Robert E. Ricklefs. Submitted.

P-1185

Comparative Genomics and Transcriptomics of Parasitic Nematodes

James Cotton, Adam Reid, Isheng Tsai, Nancy Holroyd, Matthew Berriman
Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Parasitism of vertebrates by nematodes has evolved at least 4 times independently, and plant parasitism at least 3 times. Within each of these parasitic groups, the mode of parasitism, life history and host species usage can all vary widely. These highly labile characters include those most relevant to the different pathologies caused by different species, and so include some of the most medically relevant traits. No parasitic nematodes are as easily manipulated as standard model organisms: in vitro culture is impossible for most species, and forward and reverse genetics approaches to studying gene function are not yet available. While a few nematode parasite species are studied extensively and are fairly well understood, almost nothing is understood at the molecular level beyond a few of the human-infective species.

Developments in sequencing technology mean that whole-genome data is actually cheaper and easier to obtain than much parasitological or molecular biology investigation. By sequencing multiple genomes of nematode parasites of medical and veterinary importance, as well as some key comparator species, we hope to exploit comparative genomic approaches, as well as transcriptomic data where available, to generate novel hypotheses about gene function in parasitic nematodes. Here, I present some results from comparisons of some plant-parasitic nematodes for which we have recently completed genome sequencing projects, and discuss how our current sequencing strategy will allow both broad and deep comparisons between different parasitic nematodes. The broad comparison will enable a genuinely comparative approach to parasite gene function, by finding genomic correlates between features of parasite biology and genome content. Deeper comparisons within particular clades of medically important organisms will enable us to exploit information in the pattern of molecular evolution between species, for example identifying highly-conserved non-genic regions that may be regulatory regions involved in changes in life history.

P-1186

Simultaneously reconstructing species and gene histories by modelling gene duplication, transfer and loss.

Gergely J Szollosi[1], Bastien Boussau[1,2], Eric Tannier[1], Vincent Daubin[1]
[1]UMR CNRS 5558 LBBE, Lyon, France, [2]UC Berkeley, Berkeley, USA

The success of lateral gene transfer (LGT) as an evolutionary process predicts that each gene possess its own, unique history. This situation makes the reconstruction of any scenario of transfer difficult, because no species tree can be reliably inferred from molecular data without explicitly accounting for LGT, and the interpretation of LGT routes can only concern branches of a known species tree.

The solution to this conundrum is the simultaneous reconstruction of all gene trees and their underlying species tree. We describe ODT, a probabilistic model of genome evolution that accounts for differences between gene phylogenies and the species tree as series of origination, duplication, transfer and loss events. We show, both with simulations and with biological data that we can robustly infer a maximum likelihood species tree that not only describes the pattern of speciations, but also their order in time [1]. Focusing on Cyanobacteria, distinguished among prokaryotes by a relative abundance of microfossil and biomarker records, we use 8332 homologous gene families from 36 genomes to demonstrate that our estimates are completely consistent with the fossil record and derive a set of statistically supported relative constraints that can be used to refine molecular dating.

Although ODT has proven robust to errors in gene trees, understanding the evolution of genomes requires high quality gene trees. We thus explore the possibility of improving gene tree reconstruction while at the same time inferring the chronologically ordered species phylogeny. We demonstrate that it is possible to combine the ODT model [1] (giving the likelihood of the gene tree - species tree reconciliation) with conditional clade probabilities [2] (giving the likelihood of gene tree topologies based on alignments) to efficiently infer improved gene trees given a species tree. Implemented in a parallel computing framework, our approach makes joint inference of gene trees and species trees feasible for several dozen complete prokaryotic genomes.

Simultaneous reconstruction does not only provide a single species tree, but also an ensemble of improved gene trees annotated with gene transfer, duplication and loss events that together render the historically unique Tree of Life as the emergent result of stochastic genome evolution processes.

[1] Szollosi, Boussau, Abby, Tannier, Daubin: Reconstructing the chronology of speciations by modelling genome evolution (submitted)

[2] Hohna, Drummond: Guided Tree Topology Proposals for Bayesian Phylogenetic Inference. (2012) *Syst. Biol.*

[3] Boussau, Daubin: Genomes as documents of evolutionary history. (2009) *TREE*

P-1187

Computing Phenotypes from Nucleotide Sequences: Disentangling the Role of Duplication of Network Components in the Navigability of Fitness Landscapes

Jayson Gutierrez Betancur[1,2], Steven Maere[1,2]
[1]*VIB Department of Plant Systems Biology, Technologiepark 927, B-9052, Gent, Belgium,* [2]*Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, B-9052, Gent, Belgium*

Genotype-phenotype mappings (GPMs) result from a highly multilayered and hierarchically organized process, wherein complex gene regulatory networks (GRNs) play a crucial role. Several mathematical representations of GRNs have been developed in the past, many of which have been adopted as standard computational tools for studying GPMs. Due to the generic nature of these models, however, critical levels of organization are overlooked or otherwise assumed to be fairly well captured in phenomenological parameters and/or aggregated model components. To deepen our understanding on the organization of GPMs in GRNs, we have developed a sequence-based dynamic modeling framework built upon first principles of biochemical reaction processes and protein-DNA interactions. The components of this novel modeling framework emphasize sequence-encoded molecular features that play critical roles in the dynamic behavior of GRNs in plant and animal genomes. This feature permits the reading of canonical network control parameters directly from nucleotide sequences, followed by the computation of the corresponding dynamic behavior. As a proof of concept, we focus on oscillatory GRN motifs composed of two cross-regulated transcription factor coding genes, which are arranged in minimal linear genomes. The modeling framework is implemented to disentagle the role of duplication of network components in the navigability of fitness landscapes. Our simulation results indicate that the navigability of the fitness landscape of small GRNs can be dramatically altered when its dimensionality is modified via duplication of network components.

P-1188

Phylogenomic analysis of the diverse evolutionary origins of eukaryotic genes.

Nicolas Rochette, Manolo Gouy
Laboratoire de Biométrie et Biologie Évolutive; CNRS; Université Lyon 1; Université de Lyon, Lyon, France

Although it is widely regarded as a critical step in the evolution of life, early eukaryotic evolution remains essentially unknown. The last eukaryotic common ancestor (LECA) was fully compartmentalized and bore a mitochondrion derived from alphaproteobacteria. Its genome is currently viewed as a mosaic of archaeal-like, bacterial-like and eukaryote-specific genes, suggesting one or several large scale reticulation event(s) in the tree of life.

We used an original approach to obtain a genome-wide view of eukaryotic genes origins and test hypotheses regarding the origin of eukaryotes. In the HoGenom database of clusters of homologous sequences, 554 eukaryotic genes families were identified as traceable to LECA and having archaeal or bacterial homologues. The evolutionary history of each cluster of interest was reconstructed using Bayesian and bootstrapped maximum likelihood analyses. Resulting trees were systematically parsed to investigate the relationships between eukaryotic archaeal and bacterial sequences.

Unexpectedly, very few genes supported the 3-domain topology of the tree of life. Instead, eukaryotic sequences were branching within bacteria (283) or archaea (123). The exact position of eukaryotic genes branching within archaea was unstable and not supported. Eukaryotes branched either within one of the extant archaeal phyla or as a deep-branching archaeal clade.

Regarding « bacterial » eukaryotic genes, a connection to alphaproteobacteria was prominent, with 59 eukaryotic gene families closely related to this class. Yet most genes showed no particular relationship to it and pointed to various phyla (none dominating the others) or could not be related to any particular phylum. This diversity can be explained either by the mitochondrion ancestor having an extraordinarily mosaic genome, or by multiple independent horizontal gene transfers from these phyla to the ancestors of eukaryotes.

Our results suggest a late branching of eukaryotes, close to the last archaeal ancestor, possibly as the present-day descendants of a fifth archaeal phylum. They exclude the intimate involvement, in the evolution of eukaryotes, of any other bacterium than the mitochondrion ancestor.

P-1189

What numerical taxonomy has to say regarding phylogenetics

Alex Griffing, Benjamin Lynch, Eric Stone
North Carolina State University, Raleigh, NC, USA

Numerical taxonomy describes a decades-old suite of approaches designed to group and order taxonomic units on the basis of their quantitative similarities. Phylogenetics accomplishes a similar goal through the reconstruction of an evolutionary tree, but numerical taxonomy is distinct in its explicit avoidance of phylogenetic assumptions. The methods of numerical taxonomy and phylogenetics appear to have little overlap, and there remains some ambiguity about how the two fields relate. To clarify the relationship, we focus on a dimension reduction approach to numerical taxonomy as applied to phylogenetic distance data. Our results pertain to the popular ordination technique known as principal coordinate analysis (PCoA). Intimately related to principal component analysis, PCoA uses the machinery of metric multidimensional scaling to find a low dimensional projection of organismal data for which pairwise distance relationships are in a sense maximally preserved. Classification can be achieved by clustering organisms in this lower dimensional space, and applications spanning four decades provide empirical evidence that the resulting clusters are often phylogenetically coherent. We appeal to spectral graph theory to provide support for these observations, showing how and why evolutionary trees are preserved through their projection into low-dimensional Euclidean space.

P-1190

A unified framework for inferring population genetic structure and gene-environment associations

Eric Frichot, Sean Schoville, Guillaume Bouchard, Olivier Francois
University Joseph Fourier, Grenoble, France

Local adaptation through natural selection plays a central role in shaping the genetic variation of populations. A way to investigate signatures of local adaptation is to identify polymorphisms that exhibit high correlation with environmental variables. However the geographical basis of both environmental and genetic variation can confound interpretation of these associations.

We propose to extend statistical learning approaches to detecting signatures of local adaptation in genomes, and perform inference of population genetic structure and gene-environment associations using probabilistic matrix factorization algorithms. We propose an integrated framework based on spatial statistics, population genetics and ecological modeling for scans for signatures of local adaptation from genomic data. We present a novel class of algorithms to detect correlations between environmental and genetic variation that take account background levels of population structure and spatial autocorrelation in allele frequencies generated by isolation-by-distance mechanisms.

Our framework uses probabilistic matrix factorization, a hierarchical Bayesian mixed model in which environmental variables are fixed effects and population structure is introduced as random effects. We implement fast algorithms that simultaneously estimate scores and loadings, the effects of environmental variables and a local autocorrelation scale parameter. We show that our algorithm can used to 1) correct for spatial autocorrelation when running principal component analysis, and 2) control random effects due to population history when estimating gene-environment correlations. We give examples of applications to simulated data and to human genetic data and climatic variables.

P-1191

Population structure and demographic pre-history of the Solomon Islands

Ana Duggan, Mark Stoneking
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

The Solomon Islands are unique in both their geographical situation and population composition. This collection of islands stretch through Near Oceania and into Remote Oceania, the area that spans the boundary of these two regions is both rich in diversity and understudied from a population genetics research perspective. The region is comprised of populations with varied histories, Papuan populations are thought to be descendant from the first peopling of Sahul ~50kya, the Austronesians settled in many coastal areas following their expansion through Island South East Asia ~3kya, and Polynesian Outliers are the most recent settlers, descendants of populations who back-migrated after colonizing Polynesia. The region thus encompasses great variety in population histories, cultures and languages and provides an excellent opportunity to study the effects of multiple migrations and population admixture. Extensive archaeological and linguistic research in the area offers additional sources for comparison and support of genetic findings. Our study has used next-generation sequencing technology to obtain the complete mitochondrial genome of more than 700 individuals from 18 groups spanning Austronesian, Papuan and Polynesian Outlier populations, making it both the largest and most fine grained study of the region to date. Results show that despite retaining language and culture, Papuan populations have mixed substantially with and have no greater Near Oceanic ancestry than Austronesian populations. In striking contrast to all other groups, the Austronesian population from Santa Cruz Islands shows significant Near Oceanic ancestry that confirms archaeological inferences of their unique settlement and early relationship with the Bismarck Archipelago more than 2000km to the Northwest. The mtDNA analyses of Polynesian Outlier groups indicate both recent origins and bottlenecks in these populations in confirmation of their oral traditions. Additionally, we have identified a position in the control region with an unusual mutation pattern; diagnostic for the B4a1a1a haplogroup, the transition at position 16247 appears to have arisen recently and then reverted to the ancestral form several times.

P-1192

Strong Purifying Selection at Synonymous Sites in *D. melanogaster*

David Lawrie[1], Philipp Messer[1], Ruth Hershberg[2], Dmitri Petrov[1]
[1]Stanford University, Stanford, CA, USA,[2]Technion, Israel Institute of Technology, Haifa, Israel

Synonymous sites are generally assumed to be subject to weak selective constraint. For this reason, they are often neglected as a possible source of important functional variation. We use site frequency spectra from deep population sequencing data to show that contrary to this expectation a substantial proportion of synonymous sites in *D. melanogaster* evolve under very strong selective constraint while few, if any, synonymous sites appear to be under weak constraint. Linking polymorphism with divergence data, we further find that the proportion of synonymous sites exposed to strong purifying selection is higher for those positions that show slower evolution on the Drosophila phylogeny. However, fewer synonymous sites are conserved across the entire tree than expected given the estimated percentage of strongly constrained sites. Together, these results suggest that the strong constraint at synonymous sites is episodic, such that any particular position may spend part of its evolutionary history evolving under tight constraint and part evolving neutrally or almost neutrally. This model of episodic strong selection appears to explain the rates of evolution of synonymous sites better than the alternative model of consistent weak selection.

P-1193

Detecting selective sweeps from pooled next generation sequencing samples

Simon Boitard[1], Christian Schlötterer[2], Viola Nolte[2], Ramvinay Pandey[2], Andreas Futschik[1]
[1]INRA, Laboratoire de G\énétique Cellulaire, Castanet-Tolosan, France,[2]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria,[3]Institute of Statistics and Decision Support Systems, University of Vienna, Vienna, Austria

Due to its cost effectiveness, next generation sequencing of pools of individuals (Pool-Seq) is becoming a popular strategy for characterizing variation in population samples. Since Pool-Seq provides genome-wide SNP frequency data, it is possible to use them for demographic inference and/or the identification of selective sweeps. Here, we introduce a statistical method that is designed to detect selective sweeps from pooled data by accounting for statistical challenges associated with Pool-Seq, namely sequencing errors and random sampling among chromosomes. This allows for an efficient use of the information : all base calls are included in the analysis, but the higher credibility of regions with higher coverage and base calls with better quality scores is accounted for. Computer simulations show that our method efficiently detects sweeps even at very low coverage (0.5X per chromosome). Indeed, the power of detecting sweeps is similar to what we could expect from sequences of individual chromosomes. Since the inference of selective sweeps is based on the allele frequency spectrum, we also provide a method to accurately estimate the allele frequency spectrum provided that the quality scores for the sequence reads are reliable. Applying our approach to Pool-Seq data from Drosophila melanogaster we identify several selective sweep signatures on chromosome X that include some previously well characterized sweeps like the wapl region.

P-1194

Demographic Inference Using Skyline Plots on Approximate Bayesian Computation

Miguel Navascués[1], Concetta Burgarella[2]
[1]UMR CBGP - INRA, Montpellier, France,[2]UMR AGAP - INRA, Montpellier, France

Bayesian Skyline Plots (BSPs) are representations of the posterior probability density of the effective population size in function of time; i.e. a graphical representation of the fluctuation of the effective population size with time based on the estimates obtained from Bayesian inference. The interest of BSPs is that they allow to infer gradual changes of the effective population size without the need of a specific mathematical function determining the shape of the demographic change. For instance, a population expansion can be well characterized by a BSP either if the change had ocurred instantaneously, exponentially or logistically.. Currently, the only implementation of this analysis has been done within the MCMC-based estimation of likelihood approach and is restricted to non-recombining DNA sequence data. We have explored how to implement BSP within the approximate Bayesian computation (ABC) framework, with promising preliminary results. An implementation in ABC allows to obtain BSP from any multilocus molecular data, e.g. recombining DNA sequence data, microsatellites, AFLPs or SNPs.

P-1195

Analysis of inter and intra-species NGS data with a phylogenetic approach.

Nicola De Maio, Carolin Kosiol
Institute of Population Genetics VetMedUni, Vienna, Austria

NGS genome-wide data from both within and between species represents both an important opportunity and a challenge for comparative genome analysis. While divergence data is usually dealt with phylogenetic models, polymorphism data tends instead to be analyzed with population genetics statistical tools. Phylogenetic models usually ignore polymorphisms and assume substitutions to be instantaneous. In a similar way population genetics methods do not always generalize to multiple populations or to species divergence times.

We propose new phylogenetic methods that can handle data from both within and between species, even from pooled NGS data.
Our models relax the assumption of instantaneous substitutions adding new states to the phylogenetic Markov models. These states represent polymorphic sites at different allele frequencies. Substitutions are therefore replaced by mutation events and changes in allele frequency. We call them Polymorphism Models, or PoMos.

With PoMos we can infer phylogenies and estimate mutation rates and selection coefficients on datasets from any combination of populations and reference genomes. We do not need to discard polymorphism information and we intrinsically account for incomplete lineage sorting.
Maximum likelihood estimation of our model parameters is efficiently performed with the EM algorithm built in XRATE (Klosterman et al. 2006).

Drosophila melanogaster is a model organism with a very well annotated genome. Furthermore many individuals from several natural populations have been, or will be soon, sequenced, both with individual or pooled NGS.
We applied nucleotide PoMos to a whole genome NGS sequencing dataset composed of 3 populations (Rwanda, Congo and North Carolina) of *D. melanogaster* from the DPGP projects aligned to the *D. simulans* reference genome (outgroup). We extract 4-fold degenerate sites and from these we estimate mutation rates and selection coefficients, confirming previous results, but also differentiating among nucleotides and investigating strand-specific and non-equilibrium patterns. We further infer phylogenetic trees relating the populations and the outgroup.
We do not assume any equilibrium or reversibility: ancestral base frequencies are estimated as free parameters together with all mutation rates and selection coefficients.

We plan to extend these models with an HMM structure describing variation in evolutionary forces along the genome. With such a feature we could test for the presence of positive or purifying selection and other site and branch-specific patterns along the genome.

P-1196

Estimating population mutation parameters from a single shotgun-sequenced diploid Bactrian camel (*Camelus bactrianus*) genome

Pamela Burger, Nicola Palmieri
Institute of Populationsgenetics, Vetmeduni Vienna, Vienna, Austria

The Bactrian camel and the dromedary are among the last species that have been domesticated around 3,000 to 6,000 years ago. During domestication, strong artificial (anthropogenic) selection has shaped our livestock creating a huge amount of phenotypes and breeds. Hence, domestic animals represent a unique resource to understand the genetic basis of phenotypic variation and adaptation. Similar to its late domestication history, the Bactrian camel is also among the last livestock animals to have its genome sequenced and deciphered. As no genomic data have been available until now, we adopted the development of whole-genome analyses and generated a *de novo* assembly by shotgun sequencing of a single male Bactrian camel. We obtained 1.6 Gb genomic sequences, which correspond to more than half of the Bactrian camel's genome. The aim of this study was to identify heterozygous single nucleotide polymorphisms (SNPs) and to estimate population parameters and nucleotide diversity based on a single individual camel. With an average 6.6-fold coverage we detected over heterozygous 153,200 SNPs and recorded a genome-wide nucleotide diversity comparable with that of other domesticated ungulates. Our results provide a template for future association studies targeting economically relevant traits and to identify the changes underlying the process of camel domestication and environmental adaptation.

P-1197

Parallel differentiation in Australian and North American populations of *Drosophila simulans*

Alisa Sedghifar, David Begun
University of California, Davis, CA, USA

Patterns of latitudinal differentiation in species experiencing high gene flow provide evidence for spatially varying selection. Such variation has been extensively studied in *Drosophila melanogaster*, but much less is known about *Drosophila simulans*, which has a similar distribution. We have used next-gen sequencing to compare genome-wide patterns of differentiation in *D. simulans* populations from Australia and North America and have identified regions showing high levels of differentiation on both continents as likely targets of spatially varying selection. We describe patterns of continent level convergent adaptive evolution at the nucleotide, gene and pathway levels and relate these patterns to the selection response to variable environments.

P-1198

Genome features of "Dark-fly", a *Drosophila* line reared long-term in a dark environment

Minako Izutsu[1,2], Yuzo Sugiyama[1], Osamu Nishimura[1], Kiyokazu Agata[1,2], Naoyuki Fuse[1]
[1]Laboratory for Biodiversity, Global COE Program, Kyoto, Japan,[2]Laboratory for Molecular Developmental Biology,Kyoto University, Kyoto, Japan

A laboratory at Kyoto University has maintained a *Drosophila melanogaster* line in a constant dark condition for 57 years (1400 generations). We designate this fly line "Dark-fly" and utilize it to investigate molecular mechanisms underlying environmental adaptation. We initially examined the heritability of Dark-fly in dark and light conditions. Under mating competition with the wild-type strains, Dark-fly exhibited about 2% higher heritability in a dark condition, indicating that Dark-fly possesses some traits advantageous in the dark. To address the molecular nature of the environmental adaptation of Dark-fly, we performed whole genome sequencing for Dark-fly and identified 220,000 SNPs and 4,700 InDels in the Dark-fly genome compared to the genome of Oregon-R-S, a control line. 1.8% of SNPs were classified as non-synonymous SNPs (nsSNPs). Among them, we detected 28 nonsense mutations in the Dark-fly genome. These included genes encoding an olfactory receptor and a light receptor. We also searched runs of homozygosity (ROH) regions as putative regions selected during the population history, and found 21 ROH regions in the Dark-fly genome. We identified 241 genes carrying nsSNPs or InDels in the ROH regions. These included a cluster of alpha-esterase genes that are involved in detoxification processes. These genes in the ROH regions are potential candidates related to the Dark-fly's adaptive traits. To further characterize the Dark-fly genome, we have maintained large-mixed populations (1000 flies) of Dark-fly and Oregon-R-S in a normal light-dark cycling (LD) condition and in a constant dark (DD) condition. We expected that if one of Dark-fly's SNPs (or InDels) is advantageous for living in the dark, that SNP would dominate over the genome pool of the population reared in the DD but not that reared in the LD condition. After the 22nd generation, we carried out whole genome sequencing for the mixed populations and obtained data covering the genome with mean depth of 160. Analyses of SNPs' frequencies and SNPs' linkages identified some genome regions selected in the dark. We will examine the dynamics of each SNP in the population genome through consecutive generations. These experiments will identify genes involved in the adaptive traits of Dark-fly and will provide new insights into experimental evolution.

P-1199

How ecological factors influence population phylogenomics?

Philippe Gayral[1], Vincent Cahais[2], Georgia Tsagkogeorga[3], José Melo-Ferreira[4], Marion Ballenghien[2], Lucy Weinert[5], Ylenia Chiari[4], Khakid Belkhir[2], Vincent Ranwez[2], Nicolas Galtier[2]
[1]IRBI - Université François Rabelais, Tours, France, [2]ISEM - Université Montpellier 2, Montpellier, France, [3]Queen Mary University, London, UK, [4]CIBIO Universidade do Porto, Vairão, Portugal, [5]Centre for Outbreak Analysis and Modelling, Imperial College, London, UK

Molecular evolutionary processes are governed by biological and ecological factors; such link is however poorly understood. The aim of the PopPhyl (Population Phylogenomics) project is to characterize and quantify the influence of ecological and genetical factors on comparative genomic patterns inferred from transcriptomics data of a large (~100 species) animal sampling. For each taxa, the transcriptome of 10 individuals of the focal species and 2 individuals of outgroups species is sequenced with illumina technology. Half a 454 run is also performed in one individual of the focal species. We propose a strategy for de novo transcriptome assembly from non-model organisms. Optimal assembly was achieved using a combination of assembler on illumina and 454 reads. Individual illumina reads were then mapped on de novo transcriptome. Sequencing errors, allelic expression imbalance and stochastic error (wrong sampling influenced by low read coverage) where taken into account to call SNPs. This strategy was validated on three pilot taxa, ciona, oysters and hares, showing contrasted species ecology and life history traits (marine and abundant vs. terrestrial and less abundant). Coding sequences are used to estimate and compare population genetic parameters (effective population size, mutation rate, recombination rate, strength of purifying selection -pn/ps and adaptative selection - Dn/Ds) and assess the influence of species life-history traits on molecular evolution genome-wide.

P-1200

Patterns of shared polymorphism in humans and apes

João C Teixeira, Juan R Meneu, Genís Parra, Aida M Andrés
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

The evolution of species and populations occurs through the influence and interplay of different evolutionary forces. In particular, mutations arising in a DNA sequence are responsible for the emergence of genetic polymorphisms upon which natural selection can act, driving the adaptation of organisms to their environment. This continuous process had a key role in the differentiation of populations and species throughout evolutionary history. Nonetheless, some polymorphisms are shared between closely related species, and their presence might be due to a number of factors, including recurrent mutation or balancing selection (the conservation of advantageous genetic and phenotypic diversity over long evolutionary times). In particular, there are known instances of unexpected trans-species polymorphisms (shared ancestral polymorphisms maintained since the common ancestor of different species) in humans and apes, such as those present in the MHC locus, the best-known example of balancing selection in vertebrates.

Nevertheless, the scarcity of available genome-wide, high-quality polymorphism data in apes has so far hampered the emergence of comprehensive and unbiased studies on the patterns of shared polymorphisms between these species.

Our study consists of a genome-wide approach aiming to determine the level of shared polymorphisms in the *Homo-Pan* clade, and better understand their origin. For that purpose we used exome-wide high-quality single nucleotide polymorphism (SNP) data for a sample of African humans, chimpanzees and bonobos. The exome data was obtained through whole-exome capture and high-coverage next-generation sequencing. We applied several quality-controls to our SNP dataset to retain only real polymorphisms, taking particular care to avoid the effects of technical artifacts such as duplication events and systematic errors. This provided a high-quality, comprehensive and unbiased shared coding polymorphism dataset for the three species. A combination of statistical and population genetics analyses allowed us to investigate the polymorphisms' genomic context, uncover regions that contained an excess of shared polymorphisms compared to neutrality, and infer biological interpretations for their emergence.

P-1201

Quantifying genome-wide fine-scale recombination rates in *Saccharomyces cerevisiae* from advanced intercross lines

Christopher Illingworth[1], Leopold Parts[2], Gianni Liti[3], Ville Mustonen[1]
[1]*Wellcome Trust Sanger Institute, Cambridge, UK,* [2]*University of Toronto, Ontario, Canada,* [3]*University of Nice Sophia-Antipolis, Nice, France*

Accurate estimates of bare rates of evolutionary processes such as mutation and recombination are important building blocks in our attempt to understand how organisms evolve. These rates are known to vary across genomes; in the case of recombination values can be orders of magnitude different between nearby loci due to the presence of recombination hotspots.

Here we describe a genome-wide, fine-scale inference of recombination rate in the yeast *S.cerevisiae*. A large pool of recombinants was generated through a twelve-round crossing between two diverged parental strains, carried out in two biological replicates (Parts et al. Genome Res. 2011). We sequenced full genomes of 96 isolates from each replicate pool and used haplotype data from segregating sites to infer recombination rates. We derive estimates of bare per-generation recombination rates at over thirty thousand locations across the yeast genome, and here report their statistics.

P-1202

Population genomic inference for recently bottlenecked selfing species.

Yaniv Brandvain[1], Stephen Wright[2], Graham Coop[1]
[1]*University of California - Davis, Davis, CA, USA,* [2]*University of Toronto, Toronto, ON, Canada*

Freed from the constraints of mate limitation entire populations of self-compatible plants can be established by a single founder. Such bottlenecks have frequently accompanied the transition from self-incompatibility, and subsequent rapid range expansion of selfing species. The genomic effects of these bottlenecks is profound, with diversity within these species composed of long runs of the alternative haplotypes present in the founder(s), and rare haplotypes introduced by subsequent introgression.

We present a general method to identify haplotypic structure within selfing species by integrating genome-wide polymorphism data from recently derived selfing species and their outcrossing progenitors. We then utilize patterns of polymorphism within and between haplotypes to estimate an accurate high-resolution view of the history of selfing species recently derived from an extreme bottleneck.

We illustrate these methods utilizing transcriptome resequencing data from *Capsella rubella* and its outcrossing progenitor, *C. grandiflora*. We show that for most of the genome, *C. rubella* individuals can be assigned to one of two haplotypes, and that at silent sites diversity between haplotypes is comparable to diversity at silent sites between species, suggesting that most of the ancestry across the *C. rubella* genome can be traced to two founding chromosomes. Using diversity within *C. rubella* haplotypes, we accurately date the recent emergence of the species and demonstrate the extremely rapid rate of population expansion of *C. rubella* as evidenced by the extreme elevation of the frequency of rare variants within haplotypic comparisons. Finally, we examine the extent of introgression from *C. grandiflora* by evaluating the support for rare additional haplotypes. Together, the methods presented herein, facilitate a population genomic perspective on the demographic and genomic changes associated with the transition to a selfing life.

P-1203

Extreme SNP density on the genome of Macoma balthica, a model system for the study of marine hybrid zone dynamics

Eric Pante[1], Audrey Rohfritsch[1], Vanessa Becquet[1], Nicolas Bierne[2], Pascale Garcia[1]
[1]*Laboratoire LIENSs, UMR7266 CNRS, Université de La Rochelle, La Rochelle, France,* [2]*CNRS, Institut des Sciences de l'Evolution, UMR5554, Station Méditérranéenne de l'Environnement Littoral, Sète, France*

Macoma balthica, an infaunal bivalve from marine and estuarine soft-bottom habitats of the northern hemisphere, is emerging as a strong model system to study the dynamics of marine hybrid zones. To investigate the shape, dynamics and functioning of these zones in Europe, we used a single Roche 454 FLX Titanium run to detect single nucleotide polymorphisms (SNPs) from the transcriptome of *M. balthica*. Ten individuals were pooled for each of three populations, spanning 24° latitude and two hybrid zones. This ru n generated 871,962 reads corresponding to 277 M bases from which adaptors, low-quality reads, and repetitive regions were removed. Based on a MIRA assembly and a SMALT mapping, 50 717 SNPs were detected along 558 821 nt, resulting in a density of 1 SNP every 11 bp (quality filter: phred alignment quality ≥ 20, phred base call quality ≥ 20, depth of coverage ≥ 10 reads, minor allele count ≥ 2). SNP prevalence was variable among and within assembled contigs, few being highly conserved and most being hyper variable (most variable contig: 793 SNPs along 1168 nt). Allelic diversity was high as well, with 42, 32, 26% of SNPs being di-, tri-, and quadri-allelic, respectively. In the selection of candidate SNPs for population-scale genotyping, this extreme SNP density is a severe bottleneck, as fewer than 1% of the detected SNPs are separated by ≥ 50 nucleotides. The record SNP prevalence reported here is in par with, but higher than what was previously reported for the Pacific oyster *Crassostrea gigas*, and adds to the existing view that bivalves are champions of genetic diversity among animals.

P-1204

Assessing the Relative Effects of Demography versus Selection on the Human X Chromosome Using Next-Generation Sequencing

Shaila Musharoff[1], Jeffrey M. Kidd[2,1], Brenna M. Henn[1], M.C. Yee[1], Howard M. Cann[4], Ghia Euskirchen[1], Michael Snyder[1], Carlos D. Bustamante[1], Sohini Ramachandran[3]
[1]Stanford University, Stanford, CA, USA,[2]University of Michigan, Ann Arbor, MI, USA,[3]Brown University, Providence, RI, USA,[4]Fondation Jean Dausset – Centre d'Étude du Polymorphisme Humain, Paris, France

The human X chromosome presents a unique opportunity to study the effect of population genetic forces. Because the X chromosome is carried in two copies in females and one copy in males, it has a smaller effective population size than the autosomes and is more affected by drift. As a result, the X chromosome is sensitive to sex-biased demographic processes (e.g. those involving unequal numbers of breeding males and females). Previous estimates of sex-bias in human history have shown evidence for an early persistent female bias, possibly corresponding to greater variance in male reproductive success, and a more recent male bias, possibly corresponding to excess male migration out of Africa. In addition, since males are effectively haploid on the X chromosome, there may be increased selection relative to the autosomes. Here we analyze the complete genomes of 53 individuals from 7 divergent human populations included in the HGDP-CEPH Diversity Panel (KhoeSan, Mbuti-Pygmy, Mozabite, Pathan, Cambodian, Yakut, and Maya) sequenced to 6x-15x coverage. We estimate the level of sex-bias in populations, estimate the number of deleterious mutations on the X chromosome, and perform simulations to characterize the interplay between sex-biased demography and selection.

We find that two estimators of Q, the effective population size of the X chromosome divided by that of the autosomes, are consistent with previous findings: a pi-based estimator implies a female sex-bias and an Fst-based estimator implies a male sex-bias. We next extend the demographic inference program dadi to account for dominance on the X chromosome. By analyzing the X chromosome and autosomes separately, we estimate demographic parameters of the sequenced HGDP populations and take the ratio of effective population sizes on X and autosome to be an estimator of sex-bias. The studied populations represent a series of nested bottlenecks and some have been isolated. The estimate of the number of deleterious alleles on the X chromosome in each population increases with distance from Africa, which is consistent with a serial founder effect. Finally, we perform simulations to characterize the relative effects of demographic history and selection on the amount of variation on the X chromosome using the forward simulator SFScode and assess the sensitivity of the estimation of Q. These results place the X chromosome in the context of genomic study and are of fundamental importance to the reconstruction of human history as well as to analysis that depends on estimates of human demographic parameters.

P-1205

Screening for transposable element-induced adaptations in *Drosophila melanogaster* using next-gen sequencing data

Anna-Sophie Fiston-Lavier[1], Josefa González[2], Dmitri Petrov[1]
[1]*Dept. of Biology, Stanford University, California, USA,* [2]*Institute of Evolutionary Biology, (CSIC-UPF), Barcelona, Spain*

Transposable elements (TEs) are known to be potent sources of mutation. Screening for TE-induced adaptations in *Drosophila melanogaster* that have taken place during or after the spread of this species out-of-Africa, we previously conducted a genome-wide analysis using a PCR-based approach and detected 18 putatively adaptive TE insertions, increased in population frequency after the spread out-of-Africa. Unfortunately, this analysis was limited by the genomic data available at this time (i.e, TE annotation and sequencing data).

We designed a package called "T-lex" composed of two pipelines for TE discovery and annotation using next-gen sequencing data. The "T-lex *de novo"* pipeline uses paired-end data to screen for TE one-end anchored pairs –only one of the reads mapping to the non-TE reference sequence, while the unmapped read corresponding to a TE sequence. The "T-lex pipeline" ascertains the presence/absence of annotated TE insertions in the sequenced genome(s) using two distinct approaches: While the “presence” approach looks for reads overlapping the TE junctions, the “absence” approach looks for reads corresponding to the genome sequence without the TE insertion. T-lex detected known TE insertions with 100% sensitivity and 97% specificity. Using the reads from the "absence" detection as proxy of the ancestral genome sequence prior the TE insertion, T-lex identifies the target site duplications (TSD), traces of the transposition mechanism and thus can be used for TE re-annotation. Combining the results from population data sequenced individually or as a pool, T-lex finally estimates the frequencies for each TE insertion.

Using a pooled sequence library of 92 North America strains (DGRP) and sequencing data from 20 African population (DPGP2), we analyzed 1,826 selected known TE insertions. We ended up re-annotating 108 TE insertions, mostly non-LTR elements, due to a longer poly-A tail. The T-lex results also support the conserved insertion sites for DNA and LTR elements, and show new TSD motifs. We identified 47 new putatively adaptive TE insertions being likely involved in regulatory changes. Positive selection traces combined with clinal patterns of variation along the America eastern coast provide strong evidences of their recent adaptive roles in the out-of-Africa *D. melanogaster* population. Using the T-lex *de novo* pipeline and 12 individual DGRP lines*,* we discovered more than 2000 new TE insertions in North America, mostly detected as rare in the American population. Analyzing 121 polymorphic new TE insertions, we expect to detect additional TE-induced adaptation and then end up with the largest set known to date.

P-1206

Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data

Jacob Crawford, Brian Lazzaro
Cornell University, Ithaca,NY, USA

Next-generation sequencing technologies have made it possible to address principle population genetic questions in almost any system, but high error rates associated with this data can introduce significant biases into downstream analyses, so careful consideration of experimental design and interpretation in essential in studies based on short-read sequencing. Exploration of population genetic analyses based on next-generation sequencing, has revealed some of the potential biases, but previous work has emphasized human population genetics, and further examination of parameters relevant to other systems is necessary, including when sample sizes are small and genetic variation is high. To assess experimental power to address several principal objectives of population genetic studies under these conditions, we simulated population samples under selective sweep, population growth, and population subdivision models and tested the power to recover the correct model from sequence polymorphism data inferred from 4x, 8x, and 15x short-read data. We found that estimates of population genetic differentiation and population growth parameters were systematically biased when inference was based on 4x sequencing, but biases were markedly reduced at even 8x read depth. We also found that the power to identify footprints of positive selection depends on an interaction between read depth and the strength of selection, with strong selection being recovered consistently at all read depths, but weak selection requiring deeper read depths for reliable detection. Although we have only explored a small subset of the many possible experimental designs, population genetic models and SNP calling approaches, our results reveal some general patterns and provide some assessment of what biases could be expected under similar experimental structures.

P-1207

Population mitogenomics of Eurasian mammoths using barcoded microarray capture and high-throughput sequencing

Michael Knapp[1], Matthias Meyer[2], Sebastian Lippold[2], Martin Kircher[2], Beth Shapiro[3], Michael Hofreiter[4]
[1]University of Otago, Dunedin, New Zealand, [2]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [3]Pennsylvania State University, University Park, USA, [4]University of York, York, UK

Mammoths are no doubt the most iconic of all Ice Age species. They are met with great interest not only by the lay audience but also by scientists as testified by the fact that woolly mammoths were among the first Pleistocene species from which DNA was sequenced and the first for which the complete mitochondrial genome, a complete nuclear gene and large parts of the nuclear genomes were sequenced. There have also been a number of population genetics studies addressing the population history of woolly mammoths from parts of their former geographical range. However, all these studies investigated almost exclusively mammoth specimens from the permafrost regions of northern Siberia and Alaska. This was mainly due to the challenges associated with obtaining DNA sequence data from poorly preserved mammoth remains from temperate regions. To overcome these difficulties we have used barcoded microarray DNA hybridization capture and high-throughput sequencing to sequence more than 50 complete mammoth mitochondrial genomes from most of the woolly mammoth's former range which extended from Spain to Alaska. Using these data we have reconstructed mammoth phylogeography and population dynamics across Eurasia and throughout the late Pleistocene. The results help analysing the mammoth's response to environmental change and contribute to evaluating different hypothesis explaining the extinction of the woolly mammoth.

P-1208

Parallel tagged next-generation sequencing for population genetics and phylogeography: a case study using the New Zealand frog *Leiopelma hochstetteri*.

Monika Zavodna, Catherine Grueber, Neil Gemmell
Centre for Reproduction and Genomics, Department of Anatomy, University of Otago, Dunedin, New Zealand

Rapid improvements in next-generation sequencing (NGS) platforms have greatly increased the quantities of genetic data that can be affordably processed and obtained in relatively short time. These technologies have already found application in evolutionary biology, ecology and conservation biology using genomic approaches and parallel tagging (multiplexing) of individual samples. However, only a few studies have employed parallel tagged amplicon NGS in population genetics and phylogeography. One reason for this might be the cost and time related to tagging the large number of individual samples typically required for such investigations. Here we examine how accurate and appropriate parallel tagged NGS on pooled population samples is for estimating species diversity and reconstructing phylogeographic patterns. To do so, we have used the endemic New Zealand frog *Leiopelma hochstetteri,* a species in which strong genetic structure has previously been documented. We have employed a pooling approach, in which the individual mtDNA gene amplicons have been combined by population and tagged with population specific barcodes for 454 sequencing (NGS). We have then analysed population NGS data and compared with those previously documented on the individual level. We found that population SNP frequencies from NGS data were concordant with those obtained from individual data, indicating that NGS on pooled data is accurate. However, to achieve high accuracy a number of parameters have to be considered and optimized before this population (rather than individual) parallel tagged NGS for population genetic and phylogeographic analyses can gain wide application

P-1209

Identification of elite variety tag SNPs (ETASs) - a new approach unraveling loci underlying crop improvement

Jun Lv, Wen Wang
Kumming Institute of Zoology, Kunming, Yunnan, China

Elite crop varieties usually fix elite alleles that occur with low frequency within non-elite gene pool. Hence a powerful way of dissecting these alleles for desirable agronomical traits is to compare the genomes of elite varieties with the non-elite population, and then identify the elite variety tag alleles. In this study, we sequenced deeply six elite rice varieties and used two large control panels to identify the elite varietytag SNPs (ETASs). We identified many ETASs resulting in either amino acid changes or even gene structure disruption, providing a valuable checklist for quick identification of targeted genes during these elite rice improvement. As an example, we comprehensively characterized one such ETAS, which occurred in the 9-cis-epoxycarotenoid dioxygenase gene (*Nced*) of the elite upland rice variety, IRAT104. This site shows a drastic frequency difference between upland and irrigated rice, and the striking selective sweep around it strongly suggests its association with upland rice suitability. Embedded in the only large-effect QTL identified for upland rice yield under water stress, this ETAS is associated with significant constitutive reduction of ABA level in upland rice under the regular growth condition, suggesting it probably enhances yield by relieving ABA's growth inhibition both for shoot and especially lateral roots. Our work demonstrates a novel and effective strategy to mine rare agronomically important alleles.

P-1210

Genomics of adaptation in *Arabidopsis lyrata*

Tiina Mattila, Tuomas Toivainen, Outi Savolainen
University of Oulu, Oulu, Finland

Populations of a species inhabiting diverse environments are likely to face differential selective pressures leading to local adaptation. Genetic changes accounting for these adaptations have received lots of interest in the resent years. *Arabidopsis lyrata*, a small perennial herb, occurs in fragmented populations in a wide spectrum of habitats in the Palearctic region. The populations of the species are phenotypically diverged in multiple phenotypes, such as flowering time and morphology. In addition, the different populations are genetically diverged and vary in the level of genetic variation. Thus, the species serves a good model system to study adaptive genomics in plants. We use large scale population re-sequencing of six phenotypically diverged and locally adapted *A. lyrata* populations to shed light on the genomic patterns of recent and ongoing natural selection in this species.

P-1211

Population genomics of Ostreococcus tauri

Blanc-Mathieu Romain, Derelle Evelyne, Grimsley Nigel, Moreau Hervé, Piganeau Gwenael
Observatoire Océanologique, UPMC Univ Paris 06, UMR 7232, BIOM, Banyuls-sur-Mer, France

Ostreococcus is a genus comprising a minimum of four distinct species of planktonic photosynthetic eukaryotes of extremely small size (<2 μm)[1].

Here, we investigate the power of *Illumina* Second Generation Sequencing (SGS) platform and associated tools to assemble, *de novo*, an *Ostreococcus* genome and to improve the published Sanger reference draft genome sequence [2].

The reference *Ostreococcus tauri* strain initially used to produce the Sanger reference has been re-sequenced using *Illumina* SGS (76 bp paired-end reads, 180X). The resulting dataset was assembled using 2 different *de novo* assemblers: Velvet [3] and Abyss [4]. Comparison of our *de novo* assemblies with the Sanger reference shows that up to 87% of the Sanger reference was covered with an assembly error rate estimate of 3%. These data enabled us to provide sequences for 60% of the 1678 gaps in the Sanger reference, and we used independent PCR to estimate the corresponding error rate.

We used this reference genome to map pair-end reads produced by *Illumina* sequencing of 12 *O. tauri* strains sampled in the Golf of Lion (Mediterranean sea) to infer the polymorphism spectrum in this species. I will present the first estimate of within genome polymorphism in this genus and its use to test mechanisms of GC content evolution and the strength of selection on protein coding genes.

1 Piganeau, G. *et al.* Genome diversity in the smallest marine photosynthetic eukaryotes. *Res Microbiol* (2011) **162**, 570-577

2 Derelle, E. *et al.* Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *PNAS* (2006) **103**, 11647-11652

3 Zerbino, D. R. *et al.* Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* (2008) **18**, 821-829

4 Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* (2009) **19**, 1117-1123

5 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) **25**, 1754-1760

P-1212

Regions of Homozygosity in the Porcine genome: interactions between demography and the recombination landscape

Mirte Bosse, Ole Madsen, Hendrik-Jan Megens, Richard Crooijmans, Laurent Frantz, Yogesh Paudel, Martien Groenen
Wageningen University, Wageningen, The Netherlands

Inbreeding has long been recognized as a primary cause of fitness reduction in both wild and domesticated populations. Consanguineous matings cause inheritance of similar haplotypes that are identical by descent (IBD) and result in homozygous stretches in the genome of the offspring. The size and position of regions of homozygosity (ROHs) are expected to correlate with genomic features such as GC content and recombination rate, and should be non-randomly distributed across the genome. This has important implications for the chance of particular regions to become depleted of variation. Demographics alone, therefore, may be a poor predictor of effects of inbreeding and inbreeding depression. Recent advances in sequencing technology enable a thorough investigation of genome-wide SNP distributions, and extends the use of SNP chips for ROH identification. Moreover, re-sequencing strategies should enable an unbiased characterization of variation, whereas SNP chips usually suffer from ascertainment bias. Next-generation sequencing can now be applied to characterize genome-wide patterns of homozygosity in domesticated species such as the pig, that have complex population structures and selection histories. The porcine genome is known to have a relatively heterogeneous distribution of recombination rate, making *Sus scrofa* an excellent model to study the influence of both recombination landscape and demography on genomic variation. This study utilizes re-sequencing data for the analysis of genomic ROH patterns, using a novel comparative sliding window approach. We identified an abundance of ROHs in all genomes of multiple pigs from commercial breeds and wild populations from Eurasia. Size and number of ROHs are in agreement with known demography of the populations, with population bottlenecks highly increasing ROH occurrence. The nucleotide diversity outside ROHs is high in populations derived from a large ancient population, regardless the current population size. In addition, we show an unequal genomic ROH distribution, with strong correlations of ROH size and abundance with recombination rate and GC content. Genomic ROH distribution and additional nucleotide diversity is therefore an interaction between demography and the recombination landscape. The effects of the genomic distribution of variation is often neglected in historical estimates of inbreeding. We propose a novel measure of genome variation based on nucleotide diversity outside ROHs, the number of ROHs in the genome and the average ROH size. This Genomic Variation Score (GVS) shows a high correlation with recombination rate, highlighting the importance of understanding the interaction of demography and recombination to assess effects of inbreeding.

P-1213

mtDNA genomes of Khoisan populations deciphered under a linguistic and social anthropological perspective

Chiara Barbieri[1], Blesswell Kure[4], Tom Güldemann[2], Christfried Naumann[1], Linda Gerlach[1], Falko Berthold[1], Hirosi Nakagawa[3], Sunungukuo W. Mpoloka[5], Mark Stoneking[1], Brigitte Pakendorf[6]
[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Humboldt University, Berlin, Germany, [3]Tokyo University of Foreign Studies, Tokyo, Japan, [4]Aarhus University, Aarhus, Denmark, [5]University of Botswana, Gaborone, Botswana, [6]CNRS / Université Lumière Lyon 2, Lyon, France

The ethnic groups often labeled 'Khoisan' represent a heterogeneous assortment defined by a linguistic feature: the term 'Khoisan' is in fact used as a cover term for all non-Bantu languages characterized by click consonants and spoken in Southern and East Africa. Khoisan populations differ in their ways of subsistence (being hunter-gatherers or pastoralists), phenotypical traits (the characteristic 'Khoisan' phenotype with on average lighter skin pigmentation vs. the on average darker pigmentation commonly found in sub-Saharan Africa) and in particular in the language they speak. Many specialists of Khoisan languages assume that the similarities amongst them (such as the presence of click consonants) may be the result of areal diffusion through contact rather than shared inheritance, suggesting a subdivision in three potentially unrelated language families: Tuu, Ju-ǂHoan and Khoe-Kwadi.
Genetic investigations of Khoisan populations up to now consistently showed the presence of the most ancestral human lineages for both the Y chromosome and mtDNA. A critical weakness of these studies comes from the very limited number of populations included as well as from the description of the Khoisan populations themselves (Mitchell 2010): in almost all cases, geographic and linguistic characterizations are too generic and approximate.
In this study, we employ in-solution capture and Illumina GAIIx Solexa technology to generate full mtDNA genome sequences for more than 500 individuals from 15 Khoisan populations which encompass all three of the proposed language families, as well as from neighboring Bantu-speaking populations. The uniparental perspective of mtDNA has the advantage of revealing sex-biased demographic patterns, which may be of particular importance for elucidating the social context of prehistoric contact. With these data we will investigate how people belonging to the three linguistic groups differentiated and dispersed. We will explore hypotheses of contact, late migrations and language shift in the light of the different phenotypes and ways of subsistence as well as the persistence of the extremely ancient lineages associated with the Khoisan genotype. Lastly, we will evaluate the degree of exchange with Bantu speaking neighbors.
This work has been carried out within the EUROCORES Programme EuroBABEL of the European Science Foundation.

P-1214

Inferring the history of speciation from next-generation sequence data in two nightingale species

Libor Mořkovský[1], Jakub Rídl[2], Jan Pačes[2], Lukáš Choleva[3], Marcin Antczak[4], Jiří Reif[5], Karel Janko[3], Petr Ráb[3], Radka Reifová[1]
[1]Department of Zoology, Faculty of Science, Charles University in Prague, Praha, Czech Republic, [2]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Praha, Czech Republic, [3]Institute of Animal Physiology and Genetics, Academy of Sciences of the Czech Republic, Liběchov, Czech Republic, [4]Department of Behavioural Ecology, Adam Mickiewicz University, Poznan, Poland, [5]Institute for Environmental Studies, Faculty of Science, Charles University in Prague, Praha, Czech Republic

Two closely related bird species, the Thrush Nightingale (*Luscinia luscinia*) and the Common Nightingale (*Luscinia megarhynchos*), are suitable model species for studying the genetic basis of reproductive isolation. They diverged in geographical isolation approximately 1.8 Mya and subsequently came into secondary contact in Central and Eastern Europe where they form a narrow hybrid zone. Our previous study involving eight autosomal and four Z-linked loci sampled from allopatric populations revealed substantial interspecific gene flow, which was limited mainly to autosomal loci. In this study, we used Roche 454 GS FLX+ System to sequence normalized cDNA from liver in 8 individuals of each species. *De novo* assembly of the obtained reads yielded between 5.000 and 15.000 contigs longer then 500 bp for each individual. The sequence data will be used to study the genome-wide patterns of genetic divergence and interspecific gene flow. We are particularly interested in the differences in these patterns between autosomes and the Z chromosomes. Preliminary results will be presented.

P-1215

Signatures of divergence and selection in *Cardiocondyla obscurior*

Lukas Schrader[1], Antonia Klein[1], Jürgen Heinze[1], Jürgen Gadau[2], Jan Oettler[1]
[1]*University Regensburg, Regensburg, Germany,* [2]*Arizona State University, Tempe, USA*

The tramp ant species *Cardiocondyla obscurior* is distributed across the tropics and subtropics. For introduced species, colonization events are associated with genetic bottlenecks drastically reducing allele diversity in the population, a phenomenon referred to as founder effect.

Studies comparing two introduced populations of *C. obscurior* from Brazil and Japan revealed significant phenotypic differences (e.g. in CHC profiles), indicating an early stage of divergence. Furthermore, crossing experiments between the two populations showed fitness defects in outbred queens. This susceptibility to outbreeding depression might be related to the evolutionary history of *C. obscurior*, in which inbreeding in the maternal nest is favoured.

The founder effect of low genetic diversity together with the susceptibility for outbreeding depression in *C. obscurior* provides a framework for studying the evolution of genetic incompatibility and of the very early steps of speciation processes.

By comparing the sequenced genomes from the two different populations we aim to identify genetic differences and signatures of selection potentially accounting for the emerging incompatibility.

P-1216

Likelihood-based inference of population relationships in a non-model insect using whole genomes

Konrad Lohse[1], Jack Hearn[1], Graham Stone[1], Nicholas Barton[2]
[1]*University of Edinburgh, Edinburgh, UK,* [2]*IST, Vienna, Austria*

We investigate the historical relationships between major refugial populations in Southern Europe in the oak gall wasp *Biorhiza pallida*. A recently developed likelihood method makes it possible to quantify the support for a range of non-equilibrium scenarios including divergence, migration and admixture in this species using de novo assemblies of a small number of individual haploid genomes. We consider the effects of linkage and assembly biases on such analyses.

P-1217

Pathway analysis of the action of natural selection on coding and non-coding genomic elements

Gabriel Santpere[1], Elena Carnero[1], Natalia Petit[1], Jordi Rambla[1,2], François Serra[5], Christina Hvilsom[4], Hernan Dopazo[5], Arcadi Navarro[1,3], Elena Bosch[1]
[1]Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Catalonia, Spain, [2]National Institute for Bioinformatics, Universitat Pompeu Fabra, Barcelona, Spain, [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain, [4]Research and Conservation Copenhagen Zoo, Copenhagen, Denmark, [5]Evolutionary Genomics Lab, Bioinformatics & Genomics Department, Centro de Investigación Principe Felipe, Valencia, Spain

The signature of Natural Selection may be detected not only upon coding regions of individual genes, but also in related functional elements (e.g. regulatory regions or introns) or multiple genes (e.g. functional modules or metabolic pathways). Indeed, genome-wide comparative genomics, using alignments between reference assemblies has allowed the identification of regulatory elements¹ and functional modules² bearing the signatures of adaptation and/or purifying selection in the lineages of various species, particularly humans and great apes.

In the present study, we took a two-pronged approach:

First, we focused on the coding regions of genes in four functional modules reported to be enriched with genes presenting signatures of different types of selection in the human and chimpanzee lineages. These modules are complement and coagulation cascade (with genes under positive selection), regulation of actin cytoskeleton (under purifying selection) and two KEGG pathways of common human neurodegenerative diseases, Alzheimer's and Parkinson, with no clear trends of selection.

Second, we studied regulatory regions around these genes, including introns, UTRs, miRNA, lncRNA, promoters (5kb) and trailers (5kb).

We have performed genomic capture and sequencing for the above target regions (totalling ~450 genes) in 20 chimpanzee individuals (*Pan troglodytes troglodytes*) so that, after SNP calling, we are able to study the signature of selection using both polymorphism and divergence data. By means of a combination of tests (frequency spectra and variations of the McDonald-Kreitman test) we are able to study the ralationship between ancestral and recent selective patterns and to dissect whether selection acted preferentially in conding or non-coding regions of the pathways under study.

1) Serra F, Arbiza L, Dopazo J, Dopazo H. Natural selection on functional modules, a genome-wide analysis. PLoS Comput Biol. 2011 Mar;7(3):e100109

2) Haygood R, Babbitt CC, Fedrigo O, Wray GA. Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci U S A. 2010 Apr 27;107(17):7853-7.

P-1218

Population genomic analysis and *de novo* genome assembly of the mangrove rivulus, *Kryptolebias marmoratus*, a self-fertilizing hermaphroditic fish

Joanna L. Kelley[1], Muh-Ching Yee[1], Clarence Lee[2], Elizabeth Levandowsky[2], Minita Shah[2], Craig Cummings[2], Ryan Earley[3], Carlos Bustamante[1], Timothy Harkins[2]
[1]Stanford University, Stanford, CA, USA,[2]Life Technologies Corporation, Carlsbad, CA, USA,[3]University of Alabama, Tuscaloosa, AL, USA

Elucidating the genetic basis of adaptation is a difficult task, especially when the targets of selection are not known. Emerging sequencing technologies and assembly algorithms facilitate the genomic dissection of adaptation and differentiation in a vast array of organisms. The mangrove rivulus, *Kryptolebias marmoratus,* is one of only two known internally self-fertilizing hermaphroditic vertebrates. Individuals are either hermaphroditic or male**.** Many wild caught individuals are homozygous, however, heterozygous individuals also exist, indicating that out-crossing is a relatively common occurrence with estimated frequencies of males ranging from nearly 0% to 20%. The presence of homozygous strains from different environments with distinct, and apparently heritable, phenotypes make this a tractable system to study differentiation within and among populations. The unique nature of these killifish makes it an emerging model vertebrate system, providing many opportunities to answer population genomic questions. To date, microsatellite markers distinguish wild clonal strains. Here, we leverage multiple sequencing technologies and assembly algorithms to assemble the approximately 1 gigabase genome of *Kryptolebias marmoratus.* We combine long- and short-insert libraries sequenced on multiple platforms: SOLiD sequencing of long mate pair libraries, Illumina HiSeq 2000 sequencing of short-insert libraries and Ion Torrent PGM 400 base pair long reads and long mate pair sequencing. We generate a high quality draft assembly with a total of 50X coverage of the genome and an initial combined assembly with N50 greater than 90kb and several scaffolds of over 300kb. Gene annotation is accomplished with RNA-sequencing data from multiple tissues in combination with the *de novo* genome assembly. We also present diversity estimates between lineages. We lay the groundwork using emerging genomic technologies for developing *K. marmoratus* as the new model organism for evolutionary genomics research.

P-1219

Assembly, annotation and SNP detection in Douglas-fir using cDNA libraries of control and drought stressed trees

Thomas Müller[1], Ingo Ensminger[2,3], Karl Schmid[1]
[1]University of Hohenheim, Stuttgart, Germany, [2]University of Toronto Mississauga, Mississauga, Canada, [3]Forest Research Institute of Baden-Wuerttemberg, Freiburg i. Brsg., Germany

Douglas-firs (*Pseudotsuga menziesii*) are located over a large natural range in North America where they occur in two varieties, the coastal (*Pseudotsuga menziesii* var. *menziesii*) and the interior Douglas-fir (*Pseudotsuga menziesii* var. *glauca*). Within their natural range, both varieties of Douglas-firs have evolved into a number of genetically diverse populations that are adapted to different ecozones. Douglas-firs can therefore serve to study the adaptation of conifer trees to contrasting climates using the ecological and genetical diversity.

As an increase in summer temperatures and a decrease of precipitation is expected in the coming years by IPCC in Central Europe, it is essential for forest managers to select suitable tree species that are adapted to the changing environmental conditions. The identification and analysis of differentially adapted coastal and interior Douglas-fir provenances has therefore also a practical benefit.

This study was conducted to describe an unigene set of Douglas-fir sequences for further studies and to identify genetic variations in the transcriptome that can be used as genetic markers in future analysis. Twelve pooled and normalized cDNA libraries were sequenced using Genome Analyzer FLX and 454 titanium chemistry. The libraries were constructed using needle and wood tissue from coastal and interior Douglas-fir seedlings. The seedlings were grown under different conditions (no, mild, and severe water stress). We expected to cover a high proportion of the protein-coding sequences of the Douglas-fir genome because the libraries represented a high level of genetic, morphological, and physiological diversity.

After the assembly of the twelve libraries using Newbler, we constructed an unigene set consisting of 170,859 sequences. We were able to annotate 39,624 sequences of the unigene set. Furthermore, we identified 187,653 SNPs using three different approaches (Newbler, ssahaSNP, and bwa/SAMTools), but only 27,688 SNPs were detected by all three tools.

Most SNPs detected by all three tools were found to be in both coastal and interior Douglas-firs. Considering the remaining variations, it seems that more variations can be found within coastal than within interior Douglas-firs. But those results need to be verified in future studies, as a higher number of coastal reads were used in the analysis probably introducing a bias (approx. 1.8 million coastal vs. approx. 1 million interior reads).

The data presented in this study are a necessary and useful resource for future projects like resequencing projects, association studies and studies of expression patterns.

P-1220

Next-generation sequencing and SNP genotyping of poplars reveal chloroplast capture and admixture

Stacey Lee Thompson
Umeå University, Umeå, Sweden

The use of next-generation sequencing technologies allows an opportunity to examine intra- and interspecific evolution of the chloroplast genome and to develop SNP-based assays for broader genotyping studies of population structure and admixture. Using the Illumina GAII platform, we sequenced 20 trees collected from throughout the natural ranges for each of four poplar species: Populus balsamifera, P. deltoides, P. fremontii, and P. trichocarpa (80 trees total). Sequences from European aspen (P. tremula) were also included as an outgroup. Whole chloroplast genomes were assembled for each species and annotated based on the P. trichocarpa assembly. Genomes were highly collinear with only small indels detected among species. Using a conservative SNP calling strategy, 871 SNP were identified across all 80 chloroplast genomes. Nucleotide polymorphism was generally low, averaging 6.8-9.9 x10^-4 across all polymorphic genes, compared with 4.1-8.2 x10^-4 across the entire chloroplast. Sliding-window analyses of Tajima's D show some small deviations from neutrality, with the overall low values for P. trichocarpa, likely resulting from introgression and chloroplast capture from P. balsamifera. Introgression was confirmed with additional analyses of 22 cpDNA SNP from 1067 samples of P. trichocarpa and P. balsamifera from across their ranges, revealing a deep phylogeographic split within the latter species (FCT=0.139). Similar patterns of admixture could also be observed across 94 nuclear SNP (Rho = 0.121). Additional work includes sequencing from P. tremula and P. tremuloides, as well as a deeper examination of evolutionary signatures of selection within different functional classes of genes.

P-1221

X-chromosome haplotypes as markers for population history

Pierpaolo Maisano Delser[1], Pille Hallast[1], Rita Neumann[1], Stéphane Ballereau[1,2], Mark Jobling[1]
[1]University of Leicester, Leicester, UK, [2]CNRS, Lyon, France

Inferring human population history is one of the aims of population genetics, and has been pursued by analysing different genetic markers. Much attention has been focused on haploid markers (such as mtDNA and Y chromosome) and on autosomes, but this study addresses the X chromosome in order to describe genetic diversity and to infer human population history in Western Europe. This chromosome has been chosen for its peculiar features: its haploidy in males allows phase reconstruction problems to be avoided, and it describes a greater effective population size compared to the classical haploid markers, with a bias towards female histories.

On the X chromosome several specific regions which we call PHAXs (Phylogeographically informative Haplotypes on Autosomes and X chromosome) have been selected. In these regions recombination is historically absent in HapMap Phase II samples, and mutations are likely to be the only source of genetic variation; Single Nucleotide Polymorphisms (SNPs) provide a robust phylogeny, while short-tandem repeats (STRs) provide greater discrimination and the potential for dating. Within a subset of PHAXs, SNPs and STRs are being jointly typed using a re-sequencing approach, an approach that will also discover new SNPs with low minor allele frequencies. Among several next-generation sequencing platforms, the Ion Torrent machine has been chosen, since it provides an excellent combination of features (read-length, accuracy, throughput and cost) for re-sequencing specific regions like PHAXs.

These new genetic markers will be tested on different geographical and temporal scales, with a focus on Western European populations. Two major past events will be analysed: the genetic impact of the Neolithic transition, and the Viking migrations to the British Isles. The Neolithic expansion affected almost all of Europe and has been dated 10,000-6,000 years ago, so represents a good example of an old event on a large geographical scale. On the other hand the Viking migration represents a more recent historical event (eighth century) affecting a more confined part of Europe, including the Scandinavian Peninsula and the British Isles.

This work is a good basis for exploring these novel markers and novel sequencing methods in order to understand their advantages and limits in population genetics studies, and to illuminate some aspects of human population histories.

P-1222

The landscape of recombination rate variation in Western Lowland Gorillas

Laurie S. Stevison[1], Jeffrey M. Kidd[4], Joanna L. Kelley[2], August E. Woerner[3], Laurel Johnstone[3], Carlos D. Bustamante[2], Michael Hammer[3], Jeff Wall[1], the Great Apes Genome Diversity Consortium[5]
[1]University of California, San Francisco, San Francisco, CA, USA, [2]Stanford University, Stanford, CA, USA, [3]University of Arizona, Tucson, AZ, USA, [4]University of Michigan Medical School, Ann Arbor, MI, USA

Understanding the landscape of recombination rates across the genome is fundamental to our understanding of other genome features such as GC content, nucleotide diversity, and recent structural changes. Empirical approaches used for some model species to estimate recombination rates are less feasible in non-model systems with long generation times and/or endangered species status. Therefore, it has become more common to infer population recombination rates from NGS, but small sample sizes and poorly assembled genomes present challenges to this approach. We present here the results of population recombination rates inferred from patterns of linkage disequilibrium using 6-10x coverage of full genome sequence data from 25 wild-caught Western Lowland Gorillas (*Gorilla gorilla gorilla*). We discuss how to overcome the difficulties associated with using a non-model system and a fragmented genome assembly to calculate accurate estimates of genetic distance. While recent studies have shown differences in hotspot usage between human populations, we compare hotspot usage in gorillas to these and other data available in primates. We also compare the recombination landscape of gorillas to that of humans to look at differences in evolution of broad- vs. fine-scale recombination rates between species, especially near known structural differences between genomes and known human hotspots. By examining the sequence composition of gorilla hotspots, we search for novel motifs associated with recombination enrichment in this species. Our results support the feasibility of obtaining accurate recombination rate estimates in non-model systems where experimental approaches to calculate recombination rates are increasingly difficult.

P-1223

Genomewide variation in natural populations of Neurospora tetrasperma

Padraic Corcoran[1], Fen Chen[2], Martin Lascoux[1], Peixiang Ni[2], Hanna Johanesson[1]
[1]Uppsala University, Uppsala, Sweden,[2]BGI, Hong Kong, Hong Kong

Fungal population genomics as a field of inquiry has seen a rapid growth in the recent years, with ability to sequence the genomes of multiple stains of fungi sampled from many populations. These studies have aimed at utilising a population genomic approach to understanding the evolutionary forces that have had the greatest effect in shaping the genomes of species that often display complex life cycles and novel features in comparison to plants and animals. Here we extend the use of population genomics to help understand the evolutionary history of *Neurospora tetrasperma* and its close relatives. *N. tetrasperma* has been been the focus of much research in recent years that has focused on the predominantly non-recombining mating type chromosomes, which shares features analogous to the evolution of sex chromosomes in animals and plants. Here we will present the early results of analysis on the whole genome resequencing of 58 haploid homokaryotic strains of N. tetrasperma, sampled from two populations in England and New Zealand. We wish to use the populatiom genomic dataset generated in this study to quantify the lifecycle of this species and the to investigate the rate of recombination across the genome, and the impact of such rate variation. We shall present analysis of genome widepatterns of polymorphism and divergence, with the aim of disecting which regions of the genome have been recently acted on by natural selection. In particular, we shall focus on the what evolutionary force haves shaped patterns of genetic variation on the mating type chromosomes of this species. Results and insights gained in this study can help create a greater understanding genomic impact of a recent mating system shifts in fungi. This study also represents an important step in the further development of Neurospora as a model genus for comparative population genomics.

P-1224

Whole-genome sequence versus SNP-array insights into the genetics of phenotypic variation and local adaptation in the model legume *Medicago truncatula*

Peter Tiffin[1], John Stanton-Geddes[1], Jeremy Yoder[1], Tim Paape[1], Brendan Epstein[1], Joann Mudge[2], Andrew Farmer[2], Arvind Bharti[2], Peng Zhou[1], Roman Briskine[1], Nevin Young[1]
[1]University of Minnesota, Saint Paul, MN, USA, [2]National Center for Genomic Resources, Santa Fe, NM, USA

The vast majority of genome-wide analyses of selection and association mapping have been conducted using SNP arrays. The data from arrays are limiting due to ascertainment bias associated with SNP selection and that assayed SNPs represent only a small portion of nucleotide variation. Here we compare results from both genome-wide association mapping (GWAS) and selection scans that are obtained using whole-genome sequence data (from 250 lines of the model legume *Medicago truncatula*, mean sequence coverage ~ 6X) to results obtained from insilico SNP-arrays. The resequence data also allow us to compare the putative contribution that common versus uncommon SNPs and coding versus non-coding variants make to make to phenotypic variation. The phenotypic data for these comparisons come from a greenhouse experiment in which we measured plant growth rates and the relative benefits plants obtained when grown with different strains of rhizobial mutualists. Finally, we compare insights into the genomic basis of phenotypic variation that are obtained from GWAS (in which we know the phenotype but collect data in an unnatural environment) to those obtained from local adaptation scans (for which we know the natural environment, but do not necessarily know the phenotype).

P-1225

One lane, one population: Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*.

Yuan Zhu[1], Alan Berglans[1], Josefa González[2], Dmitri Petrov[1]
[1]Stanford University, Stanford, USA,[2]Passeig Maritim de la Barceloneta, Barcelona, Spain

Sequencing of pooled, non-barcoded individuals holds great promise for the inexpensive and efficient assessment of genome-wide population allele frequencies. This approach has yet to be tested with whole, complex eukaryotic genomes. In order to carry out such a test we re-sequenced a series of libraries made from pools of largely isogenic, individually sequenced *Drosophila melanogaster* strains. We compared allele frequency estimates derived from these "pooled" libraries to corresponding estimates derived from individually sequenced strains. We demonstrate that pooled sequencing provides a faithful estimate of population allele frequency with the error well approximated by binomial sampling. We identify variation in the amount of DNA derived from individual strains as the key source of noise and show that binomial approximations works well once a sufficient numbers of strains are used in the pooling. We believe that pooling is a very powerful and cost-effective technique for detecting unusual genomic patterns in populations on genome-wide scales, and is applicable to any dataset where sequencing individuals is impossible, difficult, time consuming, or just expensive.

P-1226

Mapping the human genome's missing pieces using population admixture

Giulio Genovese[1,2], Steven McCarroll[1,3]
[1]Stanley Center, Cambridge, MA, USA, [2]Broad Institute, Cambridge, MA, USA, [3]Harvard Medical School, Boston, MA, USA

Many megabase pairs of sequence missing from the reference genome and laying within repetitive heterochromatic regions and consisting mainly of multiple inter- and intra-chromosomal duplications, including many protein-coding genes, have no known location in the human genome. We describe an approach for localizing the human genome's missing pieces by utilizing the linkage disequilibrium structure of genome sequence variation that have been created by recent admixture of historically separated human populations. By realigning unmapped sequence reads from the 1000 Genome project to sequences unplaced and missing from GRCh37/hg19 and genotyping variants that could have not been ascertained otherwise, we utilized the unique linkage disequilibrium structure in African Americans chromosomes to map several megabase pairs of the human genome's unplaced euchromatic sequence. We identified several segmental duplications of euchromatic sequence paralogous to regions of the genome that appear as unique in the reference genome but are not so in the human genome. We identified and mapped a recent ~240kbp inter-chromosomal duplication containing a partial copy of an evolutionary conserved gene, and never described before in the literature, that has no place in the current reference genome. We find that most of these paralogous sequences are hidden in the genome's heterochromatic regions, particularly its centromeres, and are affected by duplications and deletions more often than the rest of the euchromatic genome. We speculate that an approach based on mapping clones through population admixture, while being complementary to the more conventional clone tiling path approach based on overlapping sequence at the end of clones, might have an important role in finishing the euchromatic part of the genome, especially for those sequences embedded in the heterochromatic regions of centromeres and the short arms of acrocentric chromosomes. This, in turn, will allow to investigate the part of the human genome consisting of recent segmental duplication with less than 2% sequence divergence, which are currently under-represented in human genome assemblies.

P-1227

Exome Capture from Saliva Produces High Quality Genomic and Metagenomic Data on South African Genetic Diversity

Brenna Henn[1], Jeff Kidd[1,2], Thomas Sharpton[3], Meredith Carpenter[1], Paul Norman[1], Christopher Gignoux[3], Marcus Feldman[1], Jeff Wall[3], Eileen Hoal van Helden[4], Carlos Bustamante[1]
[1]Stanford University, Stanford, CA, USA, [2]University of Michigan, Ann Arbor, USA, [3]University of California, San Francisco, San Francisco, USA, [4]Stellenbosch University, South Africa, South Africa

Targeted capture of genomic regions reduces sequencing cost while generating higher coverage by allowing biomedical researchers to focus on specific loci of interest, such as exons. Targeted capture also has the potential to facilitate the generation of genomic data from DNA collected via saliva or buccal cells. DNA samples derived from these cell types tend to have a lower human DNA yield, may be degraded from age and have contamination from bacterial or other ambient oral flora. However, thousands of samples have been previously collected from these cell types and saliva collection has the advantage that it is a non-invasive form, appropriate for a wide variety of research. We demonstrate successful amplification and sequencing of 8 human exomes (at ~25x coverage for 2 South African KhoeSan families) with samples initially derived from saliva. The genotype concordance rate with Illumina SNP microarrays met 99% after filtering. This approach also captures a variety of non-human DNA reads, set aside after mapping to the human genome, and holds promise for future metagenomic studies as a 'free' addition to human exome sequencing. For example, we detect the presence of tuberculosis genomes at a low level in our metagenomic data. We develop additional methods to accurately map and call the highly polymorphic HLA and KIR loci. With an expanded exome dataset, we assess positive selection and loss-of-function mutations in the most diverse population sampled to date, the South African KhoeSan hunter-gatherers.

P-1228

European Americans who have rare variation in common.

Timothy O'Connor[1], Paul Auer[2], Benjamin Logsdon[2], Eimear Kenny[3], - NHLBI GO Exome Sequencing Project[1,4], Josyf Mychaleckyj[5], Carlos Bustamante[3], Christopher Carlson[1,2], Joshua Smith[1], Mark Rieder[1], Michael Bamshad[1], Joshua Akey[1]
[1]University of Washington, Seattle, WA, USA, [2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [3]Stanford University, Palo Alto, CA, USA, [4]Broad Institute, Cambridge, MA, USA, [5]University of Virginia, Charlottesville, VA, USA

Due to the recent dramatic expansion of human populations, most variation is exceptionally rare, which is only now becoming accessible to systematic population genetics analyses with the advent of next-generation sequencing technology. Here, we evaluate signatures of population structure in rare and common variants in 2,440 exomes sequenced to a median depth of 111x as part of the NHLBI Exome Sequencing Project. We found striking differences in patterns of structure for common (>10%) and rare (<0.5%) variants in a principal components analysis (PCA). In particular, we identified a group of 59 European Americans who tightly cluster in the PCA with rare variation, which is not observed in the PCA of common variation. To gain insights into the ancestry of these individuals, we obtained high-density SNP genotype data from a 1M Illumina bead chip, and intersected this data with additional large-scale publicly available SNP data sets. We pursued four complimentary analysis methods of this data including a global principal components analysis, Procrustes projection, admixture analysis using FRAPPE, and phylogenetic inference. All analyses provide strong evidence that these 59 individuals are of Ashkenazi Jewish ancestry. In addition, we used the exome sequence data from these 59 individuals to estimate demographic parameters and contrasted this with the demographic history inferred from other European Americans included in the Exome Sequencing Project. We find that the site frequency spectrum is shifted towards intermediate frequency alleles in the individuals inferred to have Ashkenazi ancestry, consistent with founder effects. Finally, we performed a genome-wide scan for adaptive evolution in these 59 individuals and identified several loci that have a signature of local adaptation in individuals of Ashkenazi Jewish ancestry. In summary, our comprehensive analyses demonstrate that rare variation will be a powerful tool to infer recently formed population structure.

P-1229

Estimating genome-wide fine-scale recombination rates in Drosophila melanogaster

Andrew H. Chan, Paul A. Jenkins, Yun S. Song
University of California, Berkeley, CA, USA

We perform a genome-wide analysis to estimate the recombination rate map from genetic variation data for Drosophila melanogaster. Our method is based on that of McVean et al. (2004) with several key improvements that allow accurate inference on D. melanogaster, including a more realistic mutation model, exact computation of pairwise likelihoods, incorporation of ancestral allele estimates, and the use of asymptotic sampling formulas to accommodate the higher background recombination rate observed in D. melanogaster. Through a simulation study, our method demonstrates greater robustness to the effects of selection. We find evidence for fine-scale recombination rate variation throughout the D. melanogaster genome, particularly in the X chromosome, but not to the extent found in human (McVean et al., 2004).

P-1230

A second generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage- and chromosome-specific divergence.

Tina Hu[1], Michael Eisen[2], Kevin Thornton[3], Peter Andolfatto[1]
[1]Princeton University, Princeton, NJ, USA,[2]Howard Hughes Medical Institute and the Lawrence Berkeley Laboratory, University of California Berkeley, Berkeley, CA, USA,[3]University of California Irvine, Irvine, CA, USA

We create a new assembly of the *D. simulans* genome using 142 million paired short-read sequences and previously published data for strain *w501*. Our assembly represents a higher quality genomic sequence with greater coverage, fewer misassemblies and, by several indexes, fewer sequence errors. Evolutionary analysis of this genome reference sequence reveals interesting patterns of lineage-specific divergence that are different than those previously reported. Specifically, we find that *D. melanogaster* evolves faster than *D. simulans* at all annotated classes of sites, including putatively neutrally evolving sites found in minimal introns. While this may be partly explained by a higher mutation rate in *D. melanogaster*, we also find significant heterogeneity in rates of evolution across classes of sites, consistent with historical differences in the effective population size for the two species. Also contrary to previous findings, we find that the X chromosome is evolving significantly faster than autosomes for most noncoding DNA sites and significantly slower for synonymous sites, but only marginally faster for nonsynonymous sites. The absence of a X/A difference for putatively neutral sites and the robustness of the pattern to gene ontology and sex-biased expression suggests that partly recessive beneficial mutations may comprise a substantial fraction of non-coding DNA divergence observed between species. Our assembly and analyses of the *D. simulans* genome reveal a different picture of evolutionary patterns in this lineages than suggested by previous analyses, and this has more general implications for the interpretation of evolutionary analyses of genomes of different quality.

P-1231

A new composite method to detect genomic regions under positive selection in 1000genomes data

Marc Pybus[1], Manu Uzcudun[1], Giovanni Dall'Olio[1], Pierre Luisi[1], Jaume Bertranpetit[1], Johannes Engelken[1,2]
[1]Institute of Evolutionary Biology (CSIC-UPF), Dr. Aiguader 88, 08003 Barcelona, Spain,[2]Department of Evolutionary Genetics, 04103 Leipzig, Germany

A major challenge in population genetics is the inference of naturally selected genetic variants. Recent progress in the development of new statistics as well as more accurately inferred demographic histories and unprecedented amounts of data hold great promise for the future. Specifically, the human 1000genomes project aims to make available the majority of genetic variants with an allele frequency of >1% from numerous worldwide populations (www.1000genomes.org).
We have implemented a large number of tests for natural selection in a bioinformatic pipeline, including XP-CLR, ω, CLR, Tajima´s D, Fay & Wu´s H, Fu & Li´sD, iHS, δiHH, XPEHH, Fst, δDAF etc. These statistics are mainly based on population differentiation, long range haplotype and allele frequency spectrum models. We used extensive coalescent simulations of neutral and selected genomic regions in order to evaluate each statistic for (i) their sensitivity to detect selection and for (ii) their power to localize the center of selection as well as for (iii) their robustness in diverse demographic scenarios.
Further, by combining the statistics in a single composite score through machine learning, we both improve the power and facilitate the interpretation of the diverse tests for selection. Specifically, we have obtained and tested classifiers that are optimized for detecting regions with different selective scenarios, including complete and partial sweeps as well as differentiating ancient and recent selection events. We have applied our methods successfully to experimental data from the 1000genomes project and have found interesting new patterns of selection. Results are visualized on a local version of the UCSC genome browser. We will discuss the latest results that we obtain from this ongoing project, showing that different layers of selection can be observed in the human genome.

P-1232

Measuring mixtures of microbes: Bayesian phylogenetics for metagenomic data

John O'Brien[1], Xavier Didelot[1], Daniel Falush[2]
[1]University of Oxford, Oxford, UK,[2]Max Planck Institute, Leipzig, Germany

In microbial ecologies, evolutionary change often occurs at the same temporal and spatial scales as ecological mixing. Practically, this means that for a variety of microbial environments, metagenomic samples may be composed of mixtures of closely-related species that cannot be analyzed using either traditional phylogenetic methods or more recent reference-based-mapping approaches. In this context, I show how by simultaneously using a mixture model together with a latent coalescent process it is possible to estimate the underlying haplotypes present across all samples as well as estimate the frequency of each haplotype in each sample. In the appropriate sampling regime, this analysis can provide strong insight into both evolutionary and ecological relationships at work and I provide several descriptive examples, from human gut microbiomes, Antarctic lakes, and mixed malarial infections. I conclude by detaling some interesting connections to other problems in statistical genetics, including the island coalescent process and phasing.

P-1233

Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture

John Pool[1], Kristian Stevens[2], Charis Cardeno[2], Ryuichi Sugino[1], Russell Corbett-Detig[3], Marc Crepeau[2], Pablo Duchen[4], James Emerson[5], Perot Saelao[2], David Begun[2], Charles Langley[0]
[1]University of Wisconsin - Madison, Madison, WI, USA, [2]University of California - Davis, Davis, CA, USA, [3]Harvard University, Cambridge, MA, USA, [4]Ludwig Maximilians Universitat, Munich, Germany, [5]University of California - Berkeley, Berkeley, CA, USA

Drosophila melanogaster has played a pivotal role in the development of modern population genetics. However, many basic questions regarding the demographic and adaptive history of this species remain unresolved. We report the genome sequencing of 139 wild-derived strains of *D. melanogaster*, representing 22 population samples from the sub-Saharan ancestral range of this species, along with one European sample. The bulk of these genomes were sequenced at >25X depth from haploid embryos.

Results indicated a pervasive influence of non-African admixture in many African populations, motivating the development and application of a novel admixture detection algorithm. Most inferred admixture tracts were on the scale of megabases/centiMorgans, suggesting very recent introgression. Admixture proportions were found to vary greatly among population samples (with higher levels occurring in larger towns) and even among genomes from the same population (with some collections containing a mixture of non-admixed and highly admixed individuals). Across all African genomes, admixture levels differed between chromosomes, with local regions of sharply increased or decreased admixture also apparent.

After filtering putatively admixed regions, the greatest genetic diversity was observed in southern Africa (*e.g.* Zambia), while diversity in other populations was consistent with a recent geographic expansion from this region. The European population showed fundamentally different levels of diversity reduction on each major chromosome, retaining 41% of an African sample's nucleotide diversity on the X chromosome, 62% on chromosome 2, and 80% on chromosome 3. Some African populations displayed diversity reductions specific to one or two chromosome arms, potentially reflecting a strong influence of inversion-related selection on genome-scale diversity.

Finally, genomic scans were conducted to illuminate the genetic basis of directional selection (1) within an African population, (2) differentiating African populations, and (3) differentiating European and African populations. Unique sets of gene functions were observed to be enriched for each set of genomic outliers.

P-1234

Reconstructing past Native American genetic diversity in Puerto Rico from contemporary populations

Marina Muzzio[1,2], Fouad Zakharia[1], Karla Sandoval[1], Jake K. Byrnes[3], Andres Moreno-Estrada[1], Simon Gravel[1], Eimear Kenny[1], Juan L. Rodriguez-Flores[5], Chris R. Gignoux[6], Wilfried Guiblet[4], Julie Dutil[7], The 1000 Genomes Consortium[0], Andres Ruiz-Linares[8], David Reich[9,10], Taras K. Oleksyk[4], Juan Carlos Martinez-Cruzado[4], Esteban Gonzalez Burchard[6], Carlos D. Bustamante[1]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA, [2]Facultad de Ciencias Naturales, Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina, [3]Ancestry. com®, San Francisco, California, USA, [4]Department of Biology, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico, [5]Department of Genetic Medicine, Weill Cornell Medical College, New York, New York, USA, [6]Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA, [7]Ponce School of Medicine, Ponce, Puerto Rico, [8]Department of Genetics, Evolution and Environment. University College London, London, UK, [9]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA, [10]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

The Caribbean region has a rich cultural and biological diversity, including several countries with different languages, and important historical events like the arrival of the Europeans in the late fifteenth century affected it deeply. Although it has been said that two main Native American groups peopled the Caribbean at the time of Columbus's voyages—the Arawakan-speaking Tainos and the Caribs—this model has been questioned because it comes from the descriptions written by the conquerors. The archaeological record shows a richer picture of trade among the islands, cultural change and diversity than what colonial documents depict, from the early settlements around 8000 B.P. to the chiefdoms and towns at the time of contact. How this area was peopled and how its inhabitants interacted with the surrounding continent are questions that remain to be answered due to the fragmentary nature of the historical and archaeological records.

We aim to reconstruct the Native American genetic diversity from the time of the Spanish arrival at the island of Puerto Rico from its contemporary population. We seek to find out how the original peopling of Puerto Rico occurred, along with which contemporary Native American populations are the most closely related to the Native tracks found. We used PCAdmix to trace Native American segments in admixed individuals, thus enabling us to reconstruct the original native lineages previous to the European and African contact.

Specifically, we generated local ancestry calls for the 70 parents of the 35 complete Puerto Rican trios from the whole-genome and Illumina Omni 2.5M chip Genotype data of the 1000 Genomes Project, both to examine genome-wide admixture patterns and to infer demographic historical events from ancestry tract length distributions and an ancestry-specific PCA approach, adding 55 Native American groups as potential source populations (N=475 genotyped through Illumina's 650K array) and 15 selected Mexican trios (genotyped on Affymetrix's 6.0 array, including about 906,000 SNPs) to provide population context. ADMIXTURE analysis has shown that in Puerto Rico there is no single source of contribution for the Native component. Rather, this component seems to include a mixture of major Mexican and Andean components with little contributions from the Amazonian isolates. On the other hand, the ancestry-specific PCA plotted the Puerto Rican Native segments tightly clustered with the Native segments of groups from the same language family as the Tainos (Equatorial-Tucanoan), showing a clear association between linguistics and genetics instead of a geographical one.

P-1235

MicroRNA-driven differentiation of the placental transcriptome among human populations

Song Guo[1], David Hughes[2,3], Mark Stoneking[2], Philipp Khaitovich[1,2]
[1]Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai, China,[2]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany,[3]1Department of Anthropology, University of Florida, Florida, USA

Transcriptome differences among human populations have been largely studied in cell lines. Here, we present a transcriptome-wide study conducted in placenta samples from four major human populations (Europe, Africa, South Asia, and East Asia), using high-throughout sequencing (RNA-seq).

By sequencing both long- and short- RNA, we obtained a comprehensive picture of the transcriptome differences among populations, including differences in mRNA expression and expression of known and novel non-coding RNAs. We further identify regulatory links between differences in the expression of microRNAs (miRNAs, small non-coding RNAs that mediate post-transcriptional gene silencing) and their target genes among populations.

Our study provides the first comprehensive insight into transcriptome differences among major human populations, and regulatory changes controlling these differences, in an actual tissue.

P-1236

Mining microsatellites in seabuckthorn (Hippophae rhamnoides L.) transcriptome

Ankit Jain, Rajesh Ghangal, Saurabh Chaudhary, Ram Singh Purty, Prakash Chand Sharma
University School of Biotechnology, Guru Gobind Singh Indraprastha University, Dwarka Sector-16C, Delhi, India

Development of molecular markers has been an important area of plant research considering their wide applications in molecular plant breeding, germplasm charaterization and phylogenetic studies. Among different classes of molecular markers, microsatellite based markers are the most useful due to their codominant inheritance, reproducibility, polymorphic nature and abundance. Further, gene based microsatellite markers are becoming more popular as compare to traditional anonymous random microsatellite markers, due to their rapid and inexpensive method of isolation and cross-species transferability. The next generation sequencing technology is a robust and cost-effective tool of transcriptome profiling, that has further allowed large scale development of gene based microsatellite markers at large scale at much reduced cost. Seabuckthorn (Hippophae rhamnoides L.) is a medicinally and ecologically important plant adapted to diverse and extreme environments.
Despite great advances in plant genomics, only limited marker resources are available in seabuckthorn (Hippopheae L.). Enrichment of these marker resources may have a major impact on genetic analysis, gene mapping and marker assisted selection of seabuckthorn and other related genomes.
We used Illumina GA platform sequencing data amounting to 88297 putative unigenes after assembly for mining microsatellite repeats. Of these, 7.69% unigenes contained microsatellite repeats with a frequency of one SSR every 6.704 Kb. Dinucleotide repeats were most abundant with a frequency of 63.46%, followed by trinucleotide repeats (31.35%) while a very small share was contributed by tetra- ,penta- and hexa-nucleotide repeats (5.19%) in seabuckthorn transcriptome. Exons were found densely populated with microsatellite repeats as compared to the individual untranslated regions. On an average, AG and GAA were amongst the frequent repeat types throughout the transcriptome. Out of 6787 putative unigenes containing microsatellites, 3313 unigenes (48.81%) could be assigned gene ontology (GO) terms in order to assess association of SSR positive unigenes with biological processes, cellular components and molecular function of known genes. In order to develop unigene based microsatellite markers, microsatellite carrying unigene sequences were selected on the basis of their number of repeat units, repeat length and purity by estimating their repeat variability (Var Score) using SERV. These putative markers are currently being evaluated to assess genetic diversity in natural populations of seabuckthorn collected from diverse areas across Indian Himalayas.

P-1237

Is RAD-seq suitable for short-scale phylogenetic inference ? An in silico assesment

Marie Cariou, Laurent Duret, Sylvain Charlat
Laboratoire de Biometrie et Biologie evolutive UMR5558, Lyon, France

Resolution of phylogenetic relationships between closely related species can be hindered by several problems. First, most nuclear markers lack informative variation at short evolutionary timescale. Second, because of incomplete lineage sorting and introgression, phylogenies infered from particular loci can differ from the average genome phylogeny. Finally, PCR-primer information can be lacking in poorly studied taxa. In this context Restriction site Associated DNA sequencing (RADseq) seems promissing (Davey *et al.* 2011). This technique can generate sequence data from DNA fragments flanking restriction sites, thus randomly distributed throughout the genome, from a large number of samples and without preliminary knowledge on the taxa under study.

RADseq was first developed for population genetics and quantitative trait mapping. The suitability of this method for phylogenetic inference, relying on the presence of conserved restriction sites in different species, thus remains to be evaluated. This need motivated the present study, where we simulated a RADseq experiment using the 12 *Drosophila* genomes. More than 100 restriction sites were conserved between the most distant species which diverged 40 million years ago. Inter-individual clustering of RAD sequences retrieved the majority of known orthologs. Using these data, we were able to recover the expected phylogenetic relationships between the 12 *Drosophila* species, with strong statistical support. This study therefore validates the suitability of the RADseq technique for phylogenetic inference between closely related species.

P-1238

Inferences of wolf population genetics using next-generation sequencing data

Violeta Munoz-Fuentes, Carles Vila
Estacion Biologica de Donana-CSIC, Sevilla, Spain

Grey wolves constitute a good system in which to evaluate population genetic inferences made from next generation sequencing data as many grey wolf populations have been studied extensively using traditional population genetic markers. Interestingly, different populations of wolves may live in very different habitats (arctic tundra, boreal forests, plains, mountains, deserts, temperate rain forests), some are prey specialists while others are generalists, and they may have very different demographic histories (from stable or even increasing populations to human-induced population fragmentation, density alterations and/or disruption of their social structures). Traditional population genetics have relied on a few number of loci (typically 10-20) that were assumed to be selectively neutral (e.g. control region mitochondrial DNA, microsatellites) on an extensive number of individuals per population (mostly several tens to a few hundreds). Here we make population genetic inferences based on 2.5 Mbp of sequence data per individual, encompassing both coding and non-coding regions, in circa two individuals from geographically, demographically and ecologically very distinct populations to make inferences about the demographic and evolutionary history of wolves and compare with those made from traditional population genetics.

P-1239

Diversity of immune-related genes in the yellow fever mosquito Aedes aegypti as revealed by resequencing of a wild pool

William J Palmer[1], Jewelna Osei-Poku[1], Yung Shwen Ho[2], Arnab Pain[2], Francis Jiggins[1]
[1]*University of Cambridge, Cambridge, Cambs, UK,* [2]*King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*

Aedes aegypti is a major disease vector, transmitting many human and animal pathogens including dengue, yellow fever, and dirofilaria. In the wild, refractoriness to such pathogens is highly variable between both individuals and populations. Such resistance is often controlled by variation in genes related to the function of the innate-immune system and is thought to be driven by ongoing co-evolution between host and parasite. To what extent such immune genes differ, and the evolutionary processes acting on them, have not been well quantified in natural populations. We use whole genome resequencing of a pool of wild-caught individuals (Pool-Seq) to identify polymorphism and divergence in over one hundred immunity-related genes and locate the signatures of selection acting on them.

P-1240

Detection of traces of selection with numerous SNP in small experimental chicken populations undergoing directional selection.

Frédéric Hospital[1], Bertrand Bed'Hom[1], Mathieu Gautier[2], Licia Silveri[1], Nicolas Bruneau[1], Jean-Luc Coville[1], David Gourichon[3], Marie-Hélène Pinard-van der Laan[1]
[1]INRA, Jouy en Josas, France,[2]CBGP, Montferrier-sur-Lez, France,[3]INRA, NOUZILLY, France

Three lines of White Leghorn Chickens have been selected for 12 generations for one of three different immune response traits, high antibody response (ND3), cell mediated activity (PHA) and phagocytic activity (CC). Line ND3-L was selected on ND3, line PHA-L was selected for PHA, and line CC-L for CC. A fourth line was a contemporary random bred Control maintained throughout the selection experiment (Minozzi et al. 2008). Each generation, 200 chicks per line were hatched (800 chicks in total) in a single batch. Selection for each trait was done by within-family mass selection based on individual phenotype. Heritabilities estimated for the three selection criteria ND3, PHA and CC were 0.35, 0.13 and 0.15, respectively, and correlations between the traits were not significant.

Individuals from the three selected lines and the control line at generation G9, as well as individuals from the founding population (G0) were sampled (about 20 individuals/line) and genotyped with a 60K SNP chip. We present the use of this dataset to detect traces of selection in the three selected Chicken lines.

An original Bayesian method was designed to detect the possible effects of selection by comparing the SNP allele frequencies between generations G0 and G9 for each line. The method was able to pinpoint a dwarfing gene that is known to have undergone strong selection in the experiment – hence serving as a validation control. In addition it highlights numerous SNPs that seem to behave non-neutrally, hence providing candidate regions for future search for selected genes.

While classical approaches generally focus on 'historical' (long term) traces of selection, this experiment demonstrates that it is possible to detect short-term selection in experimental population using SNPs.

Minozzi, G, et al., BMC GENETICS, 9:5, 2008.

P-1241

Population genetic structure of genome-wide SNP variation in wild *Arabidopsis thaliana* from the Western Mediterranean region

Adrian C. Brennan[1], Belen Méndez-Vigo[2], Abdel Haddioui[3], José M. Martínez-Zapater[4], F. Xavier Picó[1], Carlos Alonso-Blanco[2]
[1]Estación Biológica de Doñana (EBD), Consejo Superior de Investigaciones Científicas (CSIC), Seville, Spain,[2]Centro Nacional de Biotecnología (CNB), Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain,[3]Faculté des Sciences et Techniques, Université Sultan Moulay Slimane, Beni Mellal, Morocco,[4]Instituto de Ciencias de la Vid y del Vino (ICVV), Consejo Superior de Investigaciones Científicas (CSIC), Logroño, Spain

Thanks to its small genome and facility for experimental studies, the small, shortlived, selfing weed, *Arabidopsis thaliana*, is the dominant plant model system for molecular genetics. The advent of high throughput genotyping and decreasing costs of next generation sequencing have now helped establish *A. thaliana* as an important model system for population genetics and molecular ecology.

Here we present a population genetic study focussed on new *A. thaliana* samples from Morocco at the southern edge of the native range that have been high-throughput genotyped with genome-wide distributed SNP markers. Until now, less than a handful of individuals from the whole of the North African range have been available for study, limiting our understanding of the population history of the species in this region. Our study of the patterns of genetic diversity and structure of Moroccan A. thaliana is an important step in closing this gap in our knowledge about the worldwide population genetic structure of *A. thaliana*.

In addition, we compare the population genetic diversity and structure of Moroccan samples with geographically close Iberian samples and with samples from the rest of the worldwide northern hemisphere range. Thus, we address also questions about the relationships and likely origins of Moroccan *A. thaliana* relative to other regions and the influence of latitude on population genetic structure across the entire latitudinal range of this species.

P-1242

Sex specific gene expression and haplodiploidy purging in an ant

Kalevi Trontti, Heikki Helanterä, Christopher Wheat
Department of Biosciences, University of Helsinki, Finland

In populations with small effective sizes (N_e), the loss of allelic variation increases the risk of inbreeding depression by expression of harmful or even lethal recessive alleles in diploid individuals. Effects of inbreeding are under investigation in many hymenopteran species (ants, bees, and wasps) due to habitat fragmentation and for economic reasons (e.g. the honey bee). In hymenoperans only females are diploid whereas males are produced from unfertilized haploid eggs. Thus harmful recessive alleles are routinely exposed to natural selection and likely purged from populations when expressed in hemizygous males. Due to this 'haplodiploidy purging' it has been hypothesized that hymenopterans are less prone to inbreeding depression than other species.

In this study we used RNA-seq to identify genes that exhibit preferential male or female expression in the ant *Formica exsecta*. We then compared the nucleotide diversity in these genes among Scandinavian populations of the species. Given the haplodiploidy purging hypothesis, we predict that genes that are expressed in the male sex have less non-synonymous nucleotide changes that affect the protein structure than genes that are only or mainly expressed in females.

P-1243

Rapid development of novel protein-coding gene markers for population genetics and phylogenetics in crucifers (Brassicaceae) using a next generation sequencing approach

Stefan Zoller[1], Martin C. Fischer[2], Rie Inatsugi[4], Felix Gugerli[3], Rolf Holderegger[3], Kentaro Shimizu[4], Alex Widmer[2]
[1]Genetic Diversity Centre, ETH Zurich, Zurich, Switzerland, [2]Plant Ecological Genetics, ETH Zurich, Zurich, Switzerland, [3]Biodiversity and Conservation Biology, WSL Federal Research Institute, Birmensdorf, Switzerland, [4]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

Single-copy protein coding genes are becoming more and more important to elucidate genetic variation and phylogenetic relationships at all taxonomic levels. Finding and testing new molecular markers, however, is still a very time consuming procedure. We present and validate a method for rapidly generating a high number of such markers using next generation sequencing data. We used two Roche/454 sequencing runs to collect approximately 1 million raw sequence reads for Cardamine hirsuta and Arabis alpina. The raw reads were quality-filtered and mapped onto the genome of Arabidopsis thaliana. Using samtools and a number of perl scripts we extracted all exact matches of 20 base pairs and longer that were present and identical in all three species. We then applied primer3 to design primer-pairs on these identical regions. Pairs had to be at least 300 but not more than 700 base pairs apart, as measured on the A. thaliana reference genome and satisfy the same PCR settings (e.g., annealing temperature optimum). The primer-pairs were tested in-silico on the A. thaliana genome using Blast and a perl script, in order to validate that a single PCR fragment is potentially amplified. This resulted in 2578 primer-pairs with one potential PCR product, covering parts of 1375 unique genes in A. thaliana. We then tested 48 primer-pairs on genomic DNA extracts of 17 Brassicaceae species. 24 of these primer-pairs targeted exons only, and 24 included at least one intron or part of an UTR. 82% of the PCR reactions yielded a well-defined PCR fragment, a further 6% showed at least a weak amplification product. Sanger sequencing of the well-defined fragments resulted in high quality sequences in all cases. Usefulness for phylogenetic analysis (e.g., resolution and support of sub-clades) was tested using maximum likelihood and Bayesian methods and compared to preexisting markers.

P-1244

Population genomics of highly polymorphic species *Ciona savignyi* using high-throughput sequencing

Mariya Baranova[1], Yegor Bazykin[1], Alexei Kondrashov[1,2]
[1]*MSU, Moscow, Russia,* [2]*University of Michigan, Ann Arbor, USA*

Ciona savignyi is a marine ascidian which has the highest confirmed nucleotide diversity among multicelular organisms, probably due to exceptionally large effective population size. We used next generation sequencing to obtain the whole-genome sequences of 6 individuals of *C. savignyi* from 3 geographically remote locations, for a total of 12 haploid genotypes (haplomes). On average, two haplomes of a single individual differ from each other at 7% of nucleotide sites. The difference between genotypes of geographically remote populations is only slightly higher, suggesting limited global population subdivision or lack thereof. The ratio of the numbers of nonsynonymous and synonymous coding polymorphisms was 0.051, indicative of prevalent strong negative selection. The allele frequency spectra of nonsynonymous, conserved noncoding and splicing sites also reveal abundant negative selection. Furthermore, we detected multiple regions of polymorphism locally reduced by selective sweeps. In particular, polymorphism is reduced around the sites of nonsynonymous replacements between *C. savignyi* and *C. intestinalis,* indicative of recurrent positive selection.

P-1245

Using high density 3D image registration to identify genetic loci associated with facial morphological variations

Shouneng Peng[1], Jingze Tan[2], Hang Zhou[1], Sile Hu[1], Jing Guo[1], Li Jin[1,2], Kun Tang[1]
[1]CAS-MPG Partner Institute for Computational Biology,SIBS,CAS, Shanghai, China,[2]School of Life Sciences, Fudan University, Shanghai, China

Human facial morphological variations are often studied by examining dozens of common landmarks, which ignores the fine details of the complex traits and has low power towards poorly surrogated features. Recently we developed a 3D face registration method that aligns the high density shape meshes across hundreds of 3D facial images. This method enables us to study the fine variations of human facial morphology in fully quantitative way. We selected 10 candidate SNPs that potentially modulate facial morphology. These Marks are checked for associations with high density 3D facial surface models collected from a Han Chinese group in Jiangsu province, China. Several markers are found to be associated with changes in different facial features. One SNP in particular, rs642961 in the IRF6 gene was previously reported of significant associations with non-syndrome lip cleft. We found that the genotypes of this gene strongly predict the shape around the lip and neighboring areas (P value <0.000005). This is the first study that provides fine details of genetic modulation of common variants on complex facial surfaces. Such high-resolution signatures of genetic association may have important applications in forensics and disease diagnosis.

P-1246

Mammalian X Chromosome Inactivation evolved as a dosage compensation mechanism for dosage-sensitive genes on the X chromosome

Eugénie Pessia[1], Takashi Makino[2,3], Marc Bailly-Bechet[1], Aoife McLysaght[3], Gabriel A.B. Marais[1,4]
[1]Université Lyon 1, Centre National de la Recherche Scientifique, UMR5558, Laboratoire de Biométrie et Biologie évolutive, Villeurbanne, F-69622 cedex, France, [2]Department of Ecology and Evolutionary Biology, Graduate School of Life Sciences, Tohoku University, 6-3, Aramaki Aza Aoba, Aoba-ku, Sendai 980-8578, Japan, [3]Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland, [4]Instituto Gulbenkian de Ciência, Rua da Quinta Grande, 6, P-2780-156 Oeiras, Portugal

How and why female somatic X-chromosome inactivation (XCI) evolved in mammals remains poorly understood. Ohno proposed a two-step process where XCI is a dosage compensation mechanism that evolved to equalize expression levels of X-linked genes in females (2X) and males (1X), with a prior two-fold increase in expression of X-linked genes in both sexes.
Whereas the parity of X chromosome expression between the sexes has been clearly demonstrated, tests for the doubling of expression levels globally along the X chromosome have returned contradictory results. However, changes in gene dosage during sex chromosome evolution are not expected to impact on all genes equally, and should have greater consequences for dosage-sensitive genes.
We show that, for genes encoding components of large protein complexes (7 members) –a class of genes that is expected to be dosage-sensitive– expression of X-linked genes is similar to that of autosomal genes within the complex. These data support Ohno's hypothesis that XCI acts as a dosage compensation mechanism, and allow us to refine Ohno's model of XCI evolution.
We also explore the contribution of dosage-sensitive genes to X aneuploidy phenotypes in humans such as Turner (X0) and Klinefelter (XXY) syndromes. X aneuploidy in humans is common and is known to have mild effects because most of the supernumerary X genes are inactivated and not affected by aneuploidy. Only genes escaping XCI experience dosage changes in X-aneuploidy patients. We combined data on dosage sensitivity and XCI to compute a list of candidate genes for X-aneuploidy syndromes.

P-1247

Mechanisms and evolutionary patterns of mammalian and avian dosage compensation

Philippe Julien[1,2], David Brawand[1,2], Magali Soumillon[1,2], Anamaria Necsulea[1,2], Angelica Liechti[1], Frédéric Schütz[1,2], Tasman Daish[3], Frank Grützner[3], Henrik Kaessmann[3]
[1]University of Lausanne, Lausanne, Switzerland,[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland,[3]University of Adelaide, Adelaide, Australia

As a result of sex chromosome differentiation from ancestral autosomes, male mammalian cells only contain one X chromosome. It has long been hypothesized that X-linked gene expression levels have become doubled in males to restore dosage balance between the X and autosomes, and that the resulting X overexpression in females then drove the evolution of X inactivation (XCI). However, this model has never been directly tested and patterns and mechanisms of dosage compensation across different mammals and birds generally remain little understood. We have traced the evolution of dosage compensation using extensive transcriptome data from males and females representing all major mammalian lineages and birds. Our analyses suggest that the X has become globally upregulated in marsupials but probably not in placental mammals, where instead at least a subset of autosomal genes interacting with X-linked were downregulated. Thus, different driving forces may underlie the evolution of XCI and the highly efficient equilibration of X expression levels between the sexes observed for both of these lineages. In the egg-laying monotremes and birds, which have homologous sex chromosome systems, partial upregulation of the X (Z in birds) evolved but is largely restricted to the heterogametic sex, which provides an explanation for the partially sex-biased X (Z) expression and lack of global inactivation mechanisms in these lineages. Our findings suggest that dosage reductions imposed by sex chromosome differentiation events in amniotes were resolved in strikingly different ways.

P-1248

The impact of selection for female fitness on autosomal and W-linked genes

Marie Pointer[1,2], Hooman Moghadam[1], Alison Wright[1], Judith Mank[2]
[1]University of Oxford, Oxford, UK, [2]University College London, London, UK

Many animals (e.g. birds, lepidoptera, some reptiles) have female heterogametic (ZW) sex chromosomes. The W chromosome is analogous to the Y chromosome in that it is sex-limited, but differs in that it is found solely in females rather than males. It is therefore predicted that the W chromosome is subject to strong female-specific selection, and may play an important role in female fitness. The overall importance of directional selection in shaping the W chromosome is unknown given the powerful degradative forces that act on non-recombining sections of the genome, such as the Y and W chromosomes. We tested the role of sex-specific selection on the W chromosome and the genome as a whole in four poultry breeds that have been selected for either female or male fitness traits for over 100 generations in conjunction with the wild ancestor of chicken, the Red Jungle Fowl. RNA-Seq data from gonad tissue enabled us to identify 16 new W-linked genes, which combined with previously identified W-genes, gave us expression data on 26 loci. Although little sequence variation on the W, there was marked variability in W chromosome gene expression, with convergent directional up-regulation in replicate female-selected breeds on 73% of W genes compared to the ancestral Red Jungle Fowl. These changes are likely due to *trans* effects from the rest of the genome. Selection is therefore likely to have acted upon autosomal gene sequence to shape W chromosome gene expression, suggesting the importance of both autosomal and W-linked genes in female fitness. Taken together, these data provide the first evidence for the chromosome-wide importance of the sex-limited chromosome in a female heterogametic species, and the importance of sex-specific selection in shaping gene expression patterns.

P-1249

The evolution of sex determination and sexual differentiation in species-flock of cichlid fishes of Lake Tanganyika, East Africa

Astrid Boehne, Corina Heule, Walter Salzburger
University of Basel, Basel, Switzerland

The East African Great Lakes Tanganyika, Malawi and Victoria are home to one of the most fascinating examples of adaptive radiations: more than 2000 endemic cichlid species evolved in a few million years only making these cichlid species-flocks prime model systems in evolutionary biology. Yet, the abiotic and biotic factors that triggered explosive speciation in cichlids remain elusive. It has been proposed that newly evolving or changing sex-determining systems could be coupled to speciation.

We investigate sex determination in the genetically, morphologically and ecologically most diverse cichlid radiation, the one in Lake Tanganyika. One striking difference between species of this radiation can be seen in their mode of reproduction: one can divide cichlids of Lake Tanganyika into substrate and mouthbrooders. Interestingly, very often mouth-breeding species show strong sexual dimorphism whereas substrate breeders do not or very little and we think that how sex is determined in these species might be very different. We chose representative species for both breeding systems with available genetic resources such as genome and transcriptome sequences to study the sequence evolution and expression of genes known to play a role in sexual differentiation pathways in other vertebrate groups with a special focus on additional fish specific gene copies derived from the fish specific whole genome duplication. This approach is then applied to more species.

In a second project we study changes in genomic structure using comparative genomics to analyze syntenic relationships between different cichlid species: currently, we have access to the genome sequences of one substrate breeding species, *Neolamprologus brichardi*, and a mouthbrooder, *Astatotilapia burtoni*, both from Lake Tanganyika. Furthermore a sequenced genome is published for *Oreochromis niloticus*, a tilapiine cichlid with an XX-XY male heterogametic sex-determining system, which serves as outgroup in our analysis. This genome approach will characterize synteny breakpoints and major changes, regions that could be prone to harbor newly evolving sex-determining genes.

P-1250

Genome-wide evidence for faster-X adaptive protein evolution in mice

Athanasios Kousathanas[1], Dan Halligan[1], Rob Ness[1], Thomas Keane[2], Dave Adams[2], Bettina Harr[3], Peter Keightley[1]
[1]University of Edinburgh, Edinburgh, UK,[2]Wellcome Trust Sanger Institute, Cambridge, UK,[3]Max Planck Institute for Evolutionary Biology, Ploen, Germany

The 'faster-X' hypothesis posits that if novel mutations are at least partially recessive and the distribution of selection coefficients does not systematically differ between X-linked and autosomal loci, then X-linked loci are expected to experience more selective sweeps per generation and have higher rates of adaptive evolution than autosomal loci (Charlesworth, Coyne, and Barton; 1987). Previous studies that compared the nonsynonymous to synonymous divergence ratio (d_N/d_S) at X-linked genes versus autosomal genes have suggested a faster-X effect in mammals. However, a higher d_N/d_S ratio for the X-linked genes can be the result of a greater rate of fixation of slightly deleterious mutations and not a higher rate of adaptive evolution. We have obtained whole genome polymorphism data from ten wild caught mice (*Mus musculus castaneus*) by next-generation sequencing, and have contrasted polymorphism and divergence for ~18,500 autosomal (A) and ~750 X-linked genes which are orthologous between mouse and rat. We find that the average synonymous diversity X/A ratio is lower than the expected ¾ ratio, even after controlling for different mutation rates between the chromosomes. The greater than expected reduction in synonymous diversity of the X-linked genes might be caused by a higher rate of selective sweeps on the X-chromosome reducing linked neutral diversity. We find that the average selective constraint and the distribution of fitness effects of new deleterious mutations do not significantly differ between X-linked and autosomal genes. However, we find that the estimated fraction of nucleotide substitutions driven to fixation by positive selection (α) is significantly higher for X-linked genes than autosomal genes. We use gene expression and gene ontology data to show that the faster-X effect holds for several functional categories of genes, including genes that have high expression levels in the testis. Our results suggest that X-linked genes in mice experience an accelerated rate of adaptive evolution which could explain why the X chromosome harbors a disproportionate number of genes that contribute to reproductive isolation.

P-1251

Sex chromosome differentiation and evolution across multiple natural stickleback populations

Frédéric Chain[1], Philine Feulner[2], Mahesh Panchal[1], Christophe Eizaguirre[3,1], Martin Kalbe[1], Andrew Moore[2], Erich Bornberg-Bauer[2], Thorsten Reusch[3], Manfred Milinski[1]
[1]Max Planck Institute for Evolutionary Biology, Ploen, Germany, [2]Westfaelische Wilhelms University, Muenster, Germany, [3]IFM-GEOMAR, Institute for Marine Sciences, Kiel, Germany

The recent advancements of technologies and genomic resources have facilitated the investigation of sex chromosome evolution across different evolutionary time scales. Here we conduct a comparative analysis of sex chromosome variation among 66 individuals from 11 different natural populations of three-spined sticklebacks (*Gasterosteus aculeatus*). Stickleback populations have recently and repeatedly colonized and adapted to several different aquatic environments, making these fish a great system to study how natural selection under various adaptive scenarios can affect the evolutionary trajectories of sex chromosomes. Furthermore, sticklebacks have relatively young XY chromosomes, allowing an evaluation of the early stages of chromosomal degeneration and divergence. Using next-generation whole genome sequencing, we measure the degree to which young vertebrate sex chromosomes differ within and between individuals from various ecological backgrounds. The staggered divergence times between our sampled populations allows us to estimate the rate at which sex chromosome differentiation can progress, the amount of variation that can accumulate along different lineages, and the extent to which recurrent patterns occur in parallel adaptive systems. Whole genome sequencing allows us to identify chromosomal regions evolving under selection or undergoing degradation and differentiation based on substitution rates and the inference of functional constraints on sex-linked genes. In addition, our paired-end reads enable us to enumerate many types of structural variation, including inversions, duplications, deletions and transpositions among and between the sex chromosomes and autosomes. With these data, we have identified sex-specific genetic traits and characterized various aspects of sex chromosome polymorphism in recently diverged populations. Our study offers some insight into the evolution of young vertebrate sex chromosomes in an ecological context, as well as into the mechanisms and extent of differentiation across populations.

P-1252

Frequency and genomic distribution of genes with sex-biased expression in the dioecious plant *Silene latifolia*

Niklaus Zemp[1], Gabriel Marais[2], Alex Widmer[1]
[1]*ETH Zurich, Zurich, Switzerland,* [2]*University of Lyon, Lyon, France*

Sex-biased genes are predominantly expressed in one sex over the other and are expected to underlie sexual dimorphism and ecological differences between the sexes. Animal model systems with evolutionary old sex chromosomes display an accumulation of sex-biased genes on sex chromosomes, particularly female-biased genes in organisms with heterogametic (XY) males. In plants, the frequency and genomic distribution of sex biased genes has not been studied in detail. *Silene latifolia* is a dioecious plant with heteromorphic sex chromosomes that are evolutionary young (~10 MYA). Female and male plants display strong floral sexual dimorphism. A major problem for the study of sex-biased gene expression in this system is that no reference genome is available. To overcome this problem we used a next-generation transcriptome sequencing approach based on multiple male and female plants. We found that a substantial part of the *S. latifolia* transcriptome is sex-biased and the proportion of sex-biased genes is significantly higher on sex chromosomes than across the entire transcriptome. Moreover, the direction of the bias differs between the sex chromosomes and the entire transcriptome. Our results thus reveal that sex-biased gene expression has evolved quickly after the evolution of plant sex chromosomes and underlines the importance of sex chromosomes for the evolution of gene expression differences between the sexes.

P-1253

The *Drosophila miranda* neo-Y transcriptome: *De novo* assembly and expression analysis of neo-sex linked genes

Vera Kaiser, Doris Bachtrog
UC Berkeley, California, USA

The *Drosophila miranda* neo-sex chromosomes were formed about 1.5 MYA, when an autosomal chromosome arm (Muller element C) became fused to the existing Y chromosome. *D. miranda* is used as a model system for studying the early stages of Y chromosome degeneration and X-Y differentiation. Based on the genome assembly of the species, about 40% of all neo-Y genes have become putatively non-functional, containing frame-shift mutation and/or premature stop codons. Diagnostic SNPs observed in RNA-Seq experiments can provide estimates of transcript levels from the neo-X and neo-Y respectively. However, these SNPs alone provide little insight into the changes in the actual mRNA sequences transcribed from the neo-sex chromosomes, i.e, transcript lengths and completeness, and divergence with respect to splicing events on the neo-X and neo-Y copies. Here, we use RNA-Seq data of male and female *D. miranda* flies to reconstruct a *de novo* transcriptome assembly of the neo-Y chromosome. One major challenge is associated with the short length of Illumina RNA-Seq reads and low neo-sex chromosome divergence (about 1.1% in the coding regions). We use Trinity for a *de novo* assembly of the male transcriptome (containing hybrid assemblies of neo-X and neo-Y transcripts), and re-map male and female reads against the male assembly, extracting reads containing either diagnostic SNPs or reads spanning regions that are not differentiated between the neo-X chromosomes. Comparison with genomic reads provides us with a transcriptome of the neo-Y, separate of the neo-X, which will be used for further analyses of dosage compensation, expression changes, and degeneration of the neo-Y chromosome.

P-1254

Contrasted Patterns of Molecular Evolution in Dominant and Recessive Self-Incompatibility Haplotypes in *Arabidopsis*

Pauline Goubet[1], Hélène Bergès[2], Arnaud Bellec[1], Elisa Prat[1], Nicolas Helmstetter[1], Sophie Mangenot[1], Sophie Gallina[1], Anne-Catherine Holl[1], Isabelle Fobis-Loisy[1], Xavier Vekemans[1], Vincent Castric[1]
[1]Laboratoire de génétique et évolution des populations végétales (CNRS-Université Lille1), Lille, France, [2]Centre National des Ressources Génomiques Végétales (INRA), Toulouse, France, [3]Genoscope (CEA), Evry, France, [4]Laboratoire reproduction et développement des plantes (CNRS-INRA-ENS), Lyon, France

Genomic regions involved in the determination of mating-types such as sex chromosomes or plants self-incompatibility loci typically show very unusual patterns of molecular evolution. In particular, they are typically characterized by high levels of sequence divergence and a high density of repeated DNA, and as such they represent challenges for empirical investigation by sequencing. Self-incompatibility has been considered by geneticists a model system for reproductive biology and balancing selection, but our understanding of the genetic basis and evolution of this molecular lock-and-key system has remained limited by the extreme level of sequence divergence among haplotypes, resulting in a lack of appropriate genomic sequences. In this talk, we report and analyze the full sequence of eleven distinct haplotypes of the self-incompatibility locus (S-locus) in two closely related *Arabidopsis* species with a functional S-locus, obtained from individual BAC libraries and deep sequencing. We also expore how full genome sequencing can help us evaluate the level of degeneracy of the S-locus region in a species where self-incompatibility has been lost recently. We use this extensive dataset to highlight sharply contrasted patterns of molecular evolution of each of the two genes controlling self-incompatibility themselves as well as of the genomic region surrounding them. We find strong collinearity of the flanking regions among haplotypes on each side of the S-locus together with high levels of sequence similarity. In contrast, the S-locus region itself shows spectacularly deep gene genealogies, high variability in size and gene organization, as well as complete absence of sequence similarity in intergenic sequences and striking accumulation of transposable elements. Of particular interest, we demonstrate that dominant and recessive S-haplotypes experience sharply contrasted patterns of molecular evolution. Indeed, dominant haplotypes exhibit larger size and a much higher density of transposable elements, being matched only by that in the centromere. Overall, these properties highlight that the S-locus presents many striking similarities with other regions involved in the determination of mating-types, such as sex chromosomes in animals or in plants, or the mating-type locus in fungi and green algae.

P-1255

Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing

Michaël Bekaert[1], Christos Palaiokostas[1], Mairi Cowan[1], Andrew Davie[1], John B. Taggart[1], Karim Gharbi[2], Brendan J. McAndrew[1], David J. Penman[1], Hervé Migaud[1]
[1]Institute of Aquaculture, University of Stirling, Stirling FK9 4LA, Scotland, UK, [2]The Gene Pool, Ashworth Laboratories, The King's Buildings, University of Edinburgh, Edinburgh EH9 3JT, Scotland, UK

The Atlantic halibut (*Hippoglossus hippoglossus*) is a high value North Atlantic aquaculture species. Males mature before harvest size and thereafter show reduced growth, rendering them uneconomic for culture. Monosex female culture is therefore desirable. In this study, halibut juveniles were masculinised with MDHT (17α-methyldihydroxytestosterone) and grown on to maturity. From the first group to mature, progeny groups from four males were reared and sexed. Two of these progeny groups (n = 26 and 70) consisted of only females, while the other two (n = 30 and 71) contained balanced sex ratios (50% and 47.5% females respectively). DNA from parents and offspring from the two mixed-sex families were used as template for RAD (Restriction Associated DNA) sequencing. The 348 million raw reads produced 67,466 unique RAD tags (loci). A linkage map was constructed based on >3000 informative SNP markers. The map has 24 linkage groups, corresponding to the number of chromosome pairs in this species, and a single sex determination locus was mapped to LG7 in both families. A further set of 11 markers that were present only or predominantly in DNA from male fish was isolated from these two families. Synteny analysis showed that DNA sequences containing Atlantic halibut sex-associated SNPs were consistently clustered in several other genomes (medaka, fugu, tetraodon and stickleback). These data support the hypothesis that the Atlantic halibut has an XX/XY sex determination system. Assays are being developed for sex-associated DNA markers developed from the RAD sequencing analysis, for use in further development of monosex female halibut culture.
We are grateful for support from the Marine Alliance for Science and Technology for Scotland, the Scottish Aquaculture Research Forum and a SPARK award from the Biosciences Knowledge Transfer Network.

P-1256

Preservation of the Y transcriptome in a 10 million-year-old plant sex chromosome system

Roberta Bergero, Deborah Charlesworth
Inst. Evolutionary Biology, University of Edinburgh, Edinburgh, UK

The non-recombining chromosomes of the ancient sex chromosome systems of mammals and birds are highly degenerated, with low gene content compared to their recombining homologues, and accumulation of transposable elements. In humans, the Y chromosome has lost more than 95% of the ancestral genes, whereas the X has retained functional copies. Although theoretical studies have proposed several evolutionary causes for the genetic degeneration associated to the lack of recombination, little is known about the rate of gene erosion in non-recombining chromosomes. The XY sex chromosome pair of the dioecious plant *Silene latifolia* is at an early stage of evolution, making it interesting for studying the earliest consequences of suppressed recombination. As this plant has a large genome, and no whole-genome sequences are available, identification of sex-linked genes has been a slow process, with fewer than 20 genes identified by genetic studies of single genes. By analysing segregation patterns of SNPs from high-throughput RNA-seq transcriptome sequencing in a single full-sib family, we identified hundreds of fully sex-linked genes. Most of the genes have X and Y copies, implying that many of them are still maintained on this plant's Y chromosome. Expression levels of many Y-linked alleles are lower than those of their X-linked alleles, but the average difference is small. Also, the proportion of sex-linked genes that appear to have lost the Y-linked counterparts is estimated to be low (<20%), again suggesting that genetic degeneration is slight. Nevertheless, by using orthologous sequences from an outgroup species that does not have sex chromosomes (*S. vulgaris*), we infer that Y-linked coding sequences have incorporated a great excess of non-synonymous changes relative to their X-linked alleles, suggesting an impairment of Y-linked genes' functions.

P-1257

X/Autosome diversity in Bonobos as compared to Humans

Ines Hellmann[1], Kay Pruefer[1,2], Bonobo Genome Consortium[1]
[1]MFPL, Vienna, Austria,[2]MPI-EVAN, Leipzig, Germany

Many aspects of behavior and social structure (patri/matrilocality and polygynie/polyandry) will be reflected in the ratio of X/ autosome diversity. However, before we can make inferences about social structure from X/Autosome comparison, we need to correct for a variety of confounding factors, including male mutation bias, recent population size changes and selection.

After correcting for those factors, we find that bonobos and humans have comparable X/Autosome ratio of 0.87 and 0.85, respectively, which would translate into a roughly 2 times bigger reproductive variance for males than for females.

P-1258

Sex determination in the brown alga Ectocarpus

Sophia Ahmed[1], Rémy Luthringer[1], Alexandre Cormier[1], Akira Peters[2], Marine Robouchon[1], John Bothwell[4], Denis Roze[1], Gabriel Marais[3], Cock Mark[1], Coelho Susana[1]
[1]Station Biologique de Roscoff, CNRS and UPMC, Roscoff, France,[2]Bezhin Rosko, Roscoff, France,[3]UMR CNRS 5558 - LBBE, Lyon, France,[4]Queen's University, Belfast, UK

In genetically controlled sexual systems, gender is determined either by defined chromosomal regions or by complete sex chromosomes. Both types of structure have emerged independently and repeatedly during evolution. The structure, function and evolution of a number of such sex determining regions (SDRs) have been studied in animals, plants and fungi, but little is known about how sex is determined in other eukaryotic lineages. The brown algae represent an interesting group to study sex not only for their phylogenetic position but also because they exhibit a broad range of different sexual characteristics (isogamy/anisogamy, sex determination in either the diploid or the haploid phase, varying levels of sexual dimorphism, etc.). The filamentous brown alga Ectocarpus is particularly interesting because it has a very primitive sexual system, with minimal differences between male and female individuals. Moreover, sex in Ectocarpus is determined during the haploid, gametophyte phase, a feature that has important repercussions for SDR evolution. By exploiting resources generated as part of the Ectocarpus genome project (Cock et al., Nature 2010), we have identified a sex determining region of 1 Mbp on the male genome, where recombination is totally suppressed. The recent availability of a female Ectocarpus genome sequence is allowing us to explore the differences between male and female haplotypes of the sex locus to retrace their evolutionary history. In parallel, deep transcriptome analysis (RNA-seq) is unraveling the transcriptional network involved in Ectocarpus sex determination and differentiation. We will discuss how the elucidation of the mechanisms of sex determination in another major eukaryotic lineage will help to test existing theories of the evolutionary dynamics of sex determining regions.

P-1259

Transition between environmental and genetic sex determination in *Daphnia*: Evolution of a novel, partially sex-determining chromosome region

Celine Reisser[1], Yan Galimov[2], Cathy Haag-Liautard[1], Dominique Fasel[1], Anne Roulin[3], Jarkko Routtu[3], Christoph Haag[1]
[1]University of Fribourg, Fribourg, Switzerland, [2]Institute of Developmental Biology, Russian Academy of Sciences, Moscow, Russia, [3]University of Basel, Basel, Switzerland

In *Daphnia magna* (Cladocera, Crustacea), clonal reproduction alternates with sexual reproduction. Individuals of both sexes belonging to the same clone are genetically identical and their sex is determined by the environment. However, some clones have been shown to never produce males. These Non Male Producing (NMP) clones can only persist through phases of sexual reproduction if they co-occur with Male Producing (MP) clones. The NMP phenotype was found in divergent mitochondrial lineages, possibly indicating that the trait has evolved several times independently and that it could be more common than previously thought. Breeding experiments suggest that NMP is determined by a single autosomal region, with NMP showing dominance over MP. By using two crosses involving NMP females from different mitochondrial lineages, we mapped this region to linkage group 4 of *D. magna* in both crosses. Ongoing research focuses on fine mapping of this region to try to identify candidate loci that may contain the mutation causing the NMP phenotype. More specifically, we are employing a combination of association and classical genetic mapping using information from the scaffold sequences of the *D. magna* genome partly obtained through NGS technologies. We are currently characterizing the size, potential inversion break-points, and nucleotide diversity of the putative non-recombining region for each of the different mitochondrial lineages of NMP. This will allow us to test if recombination suppression occurs progressively in evolutionarily independent cases and whether or not selection drives loss of recombination over increasingly large regions in *D. magna*. Hence, this breeding system polymorphism may be a model for an evolutionary transition from a purely environmental to partially genetic sex determination.

P-1260

***Teximy* - a newly identified gene on the Y chromosome of the platyfish is a member of an SGNH hydrolase multigene family mainly expressed in gonads**

Marta Tomaszkiewicz[1], Manfred Schartl[2], Delphine Galiana-Arnoux[1], Jean-Nicolas Volff[1]
[1]The Institute of Functional Genomics of Lyon (IGFL), ENS de Lyon, Lyon, France, [2]University of Würzburg, Physiologische Chemie I, Biozentrum, Würzburg, Germany

Homomorphic sex chromosomes are present in several fish species including the platyfish *Xiphophorus maculatus*, making them good models to study evolutionary young gonosomes. Bacterial artificial chromosome (BAC) contigs covering several megabases from the sex-determining region of the X and Y sex chromosomes of the platyfish have been constructed and sequenced. Bioinformatic and molecular analyses of these contigs are being performed in order to identify gene candidates for the male sex-determining gene in this species.

A new gene named *teximy* (**t**estis-**e**xpressed in ***X****iphophorus* ***m****aculatus* **Y**) has been recently identified in the sex-determining region of the Y chromosome. Three very similar gene copies are organized in tandem on the Y. In silico analyses demonstrated that *teximy* is absent from the X chromosome, but four additional copies are present on autosomes. One of the autosomal copies shows 95% identity at the nucleotide level with the Y-linked copies, whereas the others are more divergent. Homologous sequences have been found in amphioxus, sea urchin and Atlantic cod, but nothing is known about their function. RT-PCR analyses of one of the Y-linked *teximy* copy have shown its exclusive expression in testes, whereas the same analyses for all the autosomal copies have demonstrated preferential expression in male and female gonads. Protein prediction from the platyfish sequences shows an SGNH hydrolase domain with its characteristic critical active site residues within four highly conserved blocks. Phylogenetic analyses of the SGNH hydrolase domain of Teximy and its homologs are performed to infer its evolutionary history and distribution within fish and other organisms.

Different studies have shown that new sex-determining genes can be formed through the duplication of autosomal genes. Such a scenario has been shown for the male sex-determining gene in the medaka (*dmrt1bY* is a duplicate gene of the autosomal gene *dmrt1*). We therefore propose the testis-expressed Y-linked *teximy* copy as a candidate for the male sex-determining gene of the platyfish.

P-1261

Masculinization of the X chromosome in the pea aphid, a system with an unusual sex-chromosome inheritance

Julie Jaquiéry[1], Claude Rispe[1], Fabrice Legeai[1,2], Solenn Stoeckel[1], Lucie Mieuzet[1], Corinne Da Silva[3], Julie Poulain[3], Nathalie Prunier-Leterme[1], Béatrice Ségurens[3], Denis Tagu[1], Jean-Christophe Simon[1]
[1]INRA, Le Rheu, France, [2]INRIA, Rennes, France, [3]Genoscope, Evry, France

Males and females differ in their optimal values at most phenotypic traits, so genetic conflicts among sexes are common. Sex-chromosomes have a sex-biased transmission, a pattern which creates favourable conditions for sexually antagonistic alleles (i.e. alleles beneficial for one sex but deleterious for the other). In particular, X- (e.g. mammals, *Drosophila*) and Z-chromosomes (e.g. birds) are predicted to be either enriched with male-beneficial or female-beneficial alleles depending on the dominant or recessive character of mutations. Sex-biased gene expression could solve conflicts raised by the spread of sexually antagonistic alleles and accordingly, there is ample evidence for a non-random chromosomal distribution of sex-biased genes in XY and ZW systems. However, several additional factors can affect this distribution, making difficult to validate such hypothesis. In this context, organisms with alternative modes of inheritance of sex-chromosomes could offer ideal conditions for studying sex conflicts. Aphids display an XX/X0 system and combine an unusual inheritance of the X chromosome with the alternation of sexual and asexual reproduction. Hence, predictions made for the genomic location of sexually antagonistic alleles in X0, XY or ZW systems do not hold for aphids. Here, we investigate theoretically the accumulation of sexually antagonistic mutations on the aphid X-chromosome. We then test the model predictions by determining the chromosomal location of genes with sex-biased expression identified from genomic data collected on males and females. In contrast to predictions for standard systems, our simulations show that the aphid X accumulates sexually antagonistic mutations beneficial for males, irrespective of the dominant or recessive character of alleles. As expected if the evolution of sex-biased expression solves the conflict raised by the spread of sexually antagonistic alleles, genes overexpressed in males were overrepresented on the X, which contains over 25% of such genes (n=286) against only 9% of the genes with unbiased expression (n=997). We report here theoretical and empirical evidence for a substantial masculinization of the X-chromosome in aphids. The lack of global hyperexpression of the haploid X in males to compensate for gene dose together with the putative absence of meiotic sex-chromosome inactivation (two factors involved in the demasculinization of the X in *Drosophila*) should further favor the masculinization of the aphid X. Our results contrast with previous observations of demasculinization of the X in non-mammalian species and highlight the relevance of organisms with alternative inheritance patterns of sex-chromosomes to unravel the mechanisms of genome evolution.

P-1262

Gene conversion and evolution of sex chromosomes: an NGS approach.

Chiara Batini, Daniel Zadik, Pille Hallast, Gurdeep Lall, Mark Jobling
University of Leicester, Leicester, UK

p { margin-bottom: 0.21cm; }

Human sex chromosomes share a long-term evolutionary history, originating ~300MYA as a pair of autosomes. Regions within them subsequently went through a series of duplications and inversions, leaving only two surviving pseudoautosomal regions in which crossing over normally occurs, and establishing other homologous but diverging (gametologous) regions in the current chromosomal organization. These are categorized into five strata of increasing evolutionary similarity, depending on the time elapsed since the crossing over was suppressed between the two. With the exception of the pseudoautosomal regions, the traditional view has been that these two chromosomes do not interact, and that the Y chromosome is on a path to degeneration. Recently, however, this view has been changed by small-scale observations of non-reciprocal transfer of variants (gene conversion) between the sex chromosomes. The larger-scale influence and implications of this process have not yet been investigated.

The aim of this work is to identify gene conversion events in human and primate sex chromosomes, in order to disentangle recent and more ancient instances in the evolutionary history of our species. Gametologous regions have been defined on the human X-chromosome reference sequence, and have been located on the Y chromosome through alignment, consisting respectively of 5.8 and 4.1 Mb.

A Next Generation sequencing methodology has been applied, using a combination of customized target enrichment (Agilent SureSelect) and the Illumina sequencing platform. The combination of a relatively small number of samples in the same lane allowed us to reach a deep coverage of these regions (between 50 and 100X on average), thus reducing the possibility of ambiguities in the final consensus sequence. An ad-hoc pipeline has been built for the analysis of the reads in the context of duplicated regions and is presented here through a critical description of pros and cons of different methods.

P-1263

Dynamics of intrachromosomal recombination in human Y-chromosome palindrome P6

Pille Hallast, Georgina Bowden, Patricia Balaresque, Stéphane Ballereau, Mark A Jobling
University of Leicester, Department of Genetics, Leicester, UK

The human sex chromosomes arose from a pair of autosomes some 300 MYA. Over time the Y chromosome has lost most of its original content and, except for the two pseudoautosomal regions, the ability to recombine normally with the X chromosome, being therefore characterized by gradual degeneration and isolation. The rest of the Y chromosome has classically been considered recombinationally inert and consequently evolving towards loss of functional genes. However, this view has been changed by the finding of intrachromosomal recombination within the Y chromosome. The majority of functional Y-linked testis-specific genes have duplicate copies and are located within large palindromes composed of two highly similar (>99.9%) arms separated by spacer sequence and spanning about 30% of the male-specific region (MSY) of human Y-chromosome euchromatin. The human Y contains 8 palindromes ranging from 30 kb to 2.9 Mb, with arm lengths from 9 kb to 1.45 Mb. The extremely high DNA sequence identity between arms is maintained by gene conversion (non- reciprocal transfer). Interspecific comparisons of palindrome arms and spacers have demonstrated a significant reduction of divergence between arms compared to spacers, suggesting that conversion may be biased towards the retention of function of genes within the arms. As well as gene conversion, reciprocal exchange could also occur, leading to intrapalindrome inversions.

We chose palindrome P6 to study in detail possible biases and outcomes of intrapalindrome recombination during human evolution. P6 spans 266 kb, with arm lengths of 110 kb and is one of the two Y-chromosome palindromes not containing any genes. We have genotyped 10 paralogous sequence variants (PSVs) from 378 individuals representing 63 distinct branches of the Y-chromosome genealogy. We demonstrate through a phylogenetic analysis of conversion events three cardinal features of the palindrome conversion process: (i) the conversion process has been rapid throughout the evolution of modern human Y-chromosomal lineages, and shows significant bias to the fixation of GC base pairs; (ii) conversion tracts can encompass many kilobases; and (iii) despite the high frequency of recombination events within palindrome arms, these resolve overwhelmingly via non-reciprocal exchange (conversions) rather than reciprocal exchange (inversions).

P-1264

Natural selection on X-linked vs. autosomal genes in the *D. pseudoobscura* subgroup

Sophie Marion de Procé[1], Andrea Betancourt[2], Nicola Palmieri[2], Christian Schloetterer[2], Brian Charlesworth[1]
[1]*University of Edinburgh, Edinburgh, UK,* [2]*Veterinärmedizinische Universität, Vienna, Austria*

The X chromosome has peculiar evolutionary properties, such as a lower effective population size and a higher exposure to natural selection in males. The hypothesis of faster-X evolution states that X-linked genes evolve faster than autosomal genes due to more adaptive evolution. However, studies investigating X-autosome differences in rates of evolution have not found conclusive evidence for or against this hypothesis. Since the XR chromosome in the *D. pseudoobscura* subgroup is orthologous to the 3L autosome in the *D. melanogaster* subgroup, we can compare the same genes in both autosomal and X-linked contexts. Using a polymorphism and divergence dataset for 123 fast-evolving genes, we estimated the proportion of mutations due to the fixation of beneficial mutations to determine whether X-linked genes experienced more adaptive evolution.

We also obtained divergence for a genome-wide analysis in which we added *D. affinis* and *D. lowei* genes to alignments of the 12 *Drosophila* genomes. This data was used for PAML analysis, which allows the use of likelihood ratio tests to test whether differences in K_a between the *D. pseudoobscura* subgroup and the *D. melanogaster* subgroup for XR/3L genes are larger than those for autosomal genes.

P-1265

The impact of recombination on the X to autosome diversity ratio in *Drosophila melanogaster* and *D. pseudoobscura*

Penelope Haddrill, Jose Campos, Brian Charlesworth
University of Edinburgh, Edinburgh, UK

Under the simplest population genetics model, the amount of neutral nucleotide diversity on the X chromosome is expected to be three-quarters of that found on the autosomes, as a result of differences in effective population size of the two types of chromosome. Recent extensions of these models that include the effects of background selection can increase this ratio closer to unity, and indicate that it is likely to be influenced by the rate of recombination. We examine this ratio in two species of fruit fly that are known to differ in their recombination rates, *Drosophila melanogaster* and *D. pseudoobscura*. In *D. pseudoobscura*, which has the higher recombination rates, the ratio of X-linked to autosomal diversity is not different from the expectation of three-quarters, whereas in *D. melanogaster* it is significantly higher. However, when we examine the impact of variation in recombination rates on a finer scale in *D. melanogaster*, we find that much of this departure from expectations can be explained by differences in the effective recombination rate experienced by the two types of chromosome, as a result of the lack of recombination in male *Drosophila*.

P-1266

Identification of regulatory networks involved sex-specific mRNA expression

Heidi Viitaniemi[1], Anti Vasemägi[1], Juha Merilä[2], Craig Primmer[1], Erica Leder[1]
[1]University of Turku, Turku, Finland,[2]University of Helsinki, Helsinki, Finland

The evolution of sexual dimorphism, including sex-specific behaviors and sexual ornamentation, is of great interest in the scientific community. However, despite the large amount of knowledge on how sexually dimorphic traits affect ecology and behavior of various organisms, the molecular mechanisms responsible for these traits remain mostly undetermined. Given that the genome of males and females are almost identical with the exception of the few genes on the Y- (or W-) chromosome or the sex-determining alleles (in the case of organisms without sex chromosomes), it is likely that many of the downstream processes resulting in sex-specific gene expression patterns are produced by changes in gene regulation. Yet, the regulatory networks involved in sex-specific gene expression patterns are generally undetermined in non-mammalian vertebrates.
A bias in mRNA expression towards females has been observed in threespine stickleback, *Gasterosteus aculeatus*, with almost a quarter of the genes having sex-biased expression concentrating on chromosome group XIX, the sex chromosomes. Compared to females, males have gaps and duplications in the chromosome sequence indicating Y-chromosome degeneration which could affect gene regulation. This provides an excellent opportunity to investigate the details of how sex-specific regulatory differences arise. To investigate regulatory networks involved in sex-specific mRNA expression in threespine stickleback, transcription patterns and SNP genotypes from approximately 600 fish from 60 families were examined. We identified chromosome regions that explain a significant proportion of variation in mRNA expression for various sex-biased genes. These regions are likely to contain important regulatory regions and allow for insights into the regulatory networks governing sexual dimorphism in gene expression.

P-1267

Gene conversion and evolution of sex chromosomes: an NGS approach.

Chiara Batini, Daniel Zadik, Pille Hallast, Gurdeep Lall, Mark Jobling
uni of leicester, leicester, UK

Human sex chromosomes share a long-term evolutionary history, originating ~300MYA as a pair of autosomes. Regions within them subsequently went through a series of duplications and inversions, leaving only two surviving pseudoautosomal regions in which crossing over normally occurs, and establishing other homologous but diverging (gametologous) regions in the current chromosomal organization. These are categorized into five strata of increasing evolutionary similarity, depending on the time elapsed since the crossing over was suppressed between the two.

With the exception of the pseudoautosomal regions, the traditional view has been that these two chromosomes do not interact, and that the Y chromosome is on a path to degeneration. Recently, however, this view has been changed by small-scale observations of non-reciprocal transfer of variants (gene conversion) between the sex chromosomes. The larger-scale influence and implications of this process have not yet been investigated.

The aim of this work is to identify gene conversion events in human and primate sex chromosomes, in order to disentangle recent and more ancient instances in the evolutionary history of our species. Gametologous regions have been defined on the human X chromosome reference sequence, and have been located on the Y chromosome through alignment, consisting respectively of 5.8 and 4.1 Mb.

A Next Generation sequencing methodology has been applied, using a combination of customized target enrichment (Agilent SureSelect) and the Illumina sequencing platform. The combination of a relatively small number of samples in the same lane allowed us to reach a deep coverage of these regions (between 50 and 100X on average), thus reducing the possibility of ambiguities in the final consensus sequence. An ad-hoc pipeline has been built for the analysis of the reads in the context of duplicated regions and is presented here through a critical description of pros and cons of different methods.

P-1268

Learning from genetic fossils on the human Y chromosome

Melissa Wilson Sayres[1,2], Kateryna Makova[2]
[1]University of California, Berkeley, Berkeley, California, USA, [2]The Pennsylvania State University, State College, Pennsylvania, USA

The Y chromosome has long been dismissed as a graveyard of genes, but there is still much to be learned from the genetic relics of genes that were once functional on the human Y. We identify human X-linked genes whose gametologs have been pseudogenized or completely lost from the Y chromosome and infer which evolutionary forces may be acting to retain genes on the Y chromosome. Although this gene loss appears to be largely correlated with the suppression of recombination, we observed that X-linked genes that are expressed in male-specific tissues may be more likely to have retained functional Y homologs. Curiously, differences in the function and molecular evolution of X-linked genes do not distinguish whether their Y-linked homologs are likely to have been retained or lost, suggesting that a large proportion of gene loss on the human Y chromosome may be stochastic. Importantly, we both support and expand upon previous suggestions that X chromosome inactivation is primarily driven by gene loss on the Y. According to our linear discriminatory analysis, the inactivation status of an X-linked gene is sufficient to discriminate between X-linked genes with functional vs. pseudogenized Y homologs with a 90% success rate.

P-1269

Sex-specific embryonic gene expression at different stages of sex chromosome evolution

Susan E. Lott, Jacqueline E. Villalta, Doris Bachtrog, Michael B. Eisen
University of California, Berkeley, CA, USA

Sex chromosome dosage differences between males and females are a major form of natural genetic variation in many species. In *Drosophila*, females have two X chromosomes, while males have one X and one Y. Several fusions of X chromosomes with autosomes have occurred along the branches leading to *D. pseudoobscura* and *D. miranda*. The resulting neo-X chromosomes are gradually acquiring the properties of classical sex chromosomes, and becoming targets for the complex molecular mechanisms that have evolved to compensate for the differences in X chromosome dose between sexes. We have recently shown that *D. melanogaster* possess at least two dosage compensation mechanism: the well-characterized MSL-mediated dosage compensation active in most somatic tissues, and a second, which acts during early embryogenesis. To better understand the evolutionary constraints on sex chromosome expression and evolution, we have used single embryo mRNA-seq to characterize gene expression in female and male embryos of *D. pseudoobscura* and *D. miranda*, from ~0.5-8 hours of development. Examining expression from these X chromosomes throughout embryonic development, we observe a relationship between the age of the X chromosome and the number of genes that are compensated. We also characterize what kinds of genes and processes are more likely to be compensated or have their expression level more constrained, at various stages in embryonic development, which we can then use to test whether expression constraint is predictive of fitness consequences.

P-1270

Dynamics of Y chromosome evolution in *Drosophila affinis*

Robert Unckless, Andrew Clark
Cornell University, Ithaca, NY, USA

A neo-Y chromosome degenerates through a combination of transcriptional silencing and mutation until there are very few functional genes remaining. In the *Drosophila obscura* group, a neo Y-chromosome was formed about 10 million years ago when the fusion of Muller elements A and D formed a neo-X chromosome. The makeup and degeneration of the neo-Y chromosome has been studied in several members of the obscura group. *Drosophila affinis* presents an opportunity to gain further insight into the evolution of the Y chromosome because of three characteristics. First, males without a Y chromosome are fertile, suggesting that no essential fertility factors reside on the *D. affinis* Y chromosome. Second the Y chromosome exhibits considerable morphological diversity in and between wild populations. Finally, *D. affinis* is host to an X-linked sex-ratio meiotic drive system, presumably placing considerable selective pressure on the Y chromosome for resistance. To begin to develop the *D. affinis* system, we used next generation sequencing of both the genome and transcriptome of males and females to determine the genomic location of canonical Y-linked fertility factors and to determine the genomic makeup of the Y chromosome in *Drosophila affinis.* Preliminary results suggest that male fertility factors are scattered throughout the autosomes.

P-1271

Duplicate gene evolution on the Y-chromosome: insights from ampliconic genes of Sulawesi macaques.

Ana-Hermina Ghenu, Ben Evans
McMaster University, Hamilton, Ontario, Canada

Recombination facilitates natural selection by increasing the genetic variance in fitness of a population. For this reason, haploid portions of the genome, such as the Y-chromosome in primates, potentially experience a decreased efficiency of natural selection relative to diploid genomic regions. This could drive Y-chromosome "degeneration", for example by contributing to the loss of genes on this chromosome. However, the "ampliconic" class of genes on the Y-chromosome may evade degeneration by having duplicate copies that undergo a form of ectopic recombination called gene conversion.

With an aim of better understanding the evolutionary interplay between gene duplication, gene conversion, and natural selection on a haploid chromosome, we examined sequence evolution and copy number polymorphism on the Y-chromosome in a group of closely related macaque species. We quantified copy number variation within and between species of Sulawesi macaques for seven genes homologous to human ampliconic genes. We recovered evidence for gene conversion and observed a high degree of conservation in copy number between these closely related species, even though they have small effective population sizes and diverged from one another ~20Ne generations ago.

The conservation of gene duplicates in species with small effective population sizes suggests either that the rates of gene duplication and loss are slow on the Y-chromosome, that ampliconic gene copy numbers are under purifying selection in macaques, or some combination of these alternatives.

P-1272

Recent specialization of the human and mouse X chromosomes for the male germline

Jacob Mueller[1], Helen Skaletsky[1], Laura Brown[1], Sara Zaghul[1], Susie Rock[2], Wesley Warren[2], Richard Wilson[2], David Page[1]
[1]Whitehead Institute, Cambridge, MA, USA, [2]Washington University Genome Center, St. Louis, MO, USA

Roughly 45 years ago, Susumu Ohno posited that the gene content of X chromosomes would be conserved across placental mammals, a statement often referred to as Ohno's law. However, a systematic and rigorous test of Ohno's law has yet to be performed. Such a test requires the comparison of two high-quality sequence assemblies in order to uncover notable gene exceptions to Ohno's law. We thus generated targeted improvements of haplotype-specific clone-based sequence across regions of the human X chromosome assembly in order to compare it to the haplotype-specific clone-based mouse X chromosome assembly. We targeted our improvements to large, nearly-identical, segmental duplications, termed amplicons, which permitted us to resolve gaps in the current reference sequence, reconstruct previously misassembled regions, and reveal new palindromic amplicons. We compared this revised sequence to the mouse X chromosome sequence and found that single-copy genes tend to follow Ohno's law of conservation; ~95% are orthologous between humans and mice. In striking contrast, genes within amplicons are the primary violators of Ohno's law; only 31% and 22% are conserved in humans and mice, respectively. We find that X-amplicon genes represent the majority of newly acquired genes in both lineages. Furthermore, we find that newly acquired genes are expressed almost exclusively in testicular germ cells, suggesting specific roles in the male germline. Thus, despite the X chromosome's reputation for conservation, large portions of its sequence were recently sculpted into X-amplicons harboring novel testicular germ cell-specific genes.

P-1273

Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*

Qi Zhou[1,2], Wen Wang[2], Doris Bachtrog[1]
[1]*University of California, Berkeley, Berkeley, USA,* [2]*Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China*

Drosophila albomicans is a unique model organism for studying both sex chromosome and B chromosome evolution. A pair of its autosomes comprising roughly 40% of the whole genome has fused to the ancient X and Y chromosomes only about 0.12 million years ago, thereby creating the youngest and most gene-rich 'neo-sex' system reported to date. This species also possesses recently derived B chromosomes that show non-Mendelian inheritance and significantly influence fertility. We sequenced male flies with B chromosomes at 124.5-fold genome coverage, and inbred female flies derived from the same strain but without B's at 28.5-fold using next-generation sequencing. We assembled a female genome and placed 53% of the sequence and 85% of the annotated proteins into specific chromosomes, by comparison with the 12 *Drosophila* genomes. Despite its very recent origin, the non-recombining neo-Y chromosome shows various signs of degeneration, including a significant enrichment of non-functional genes compared to the neo-X, and an excess of tandem duplications relative to other chromosomes. We also characterized a B-chromosome linked scaffold that contains an actively transcribed unit and shows sequence similarity to the subcentromeric regions of both the ancient X and the neo-X chromosome. Our results provide novel insights into the very early stages of sex chromosome evolution and B chromosome origination, and suggest an unprecedented connection between the births of these two systems in *D. albomicans*.

P-1274

PRELIMINARY SEQUENCING AND COMPARATIVE ANALYSIS OF CARNIVORE Y CHROMOSOMES

GANG LI[1], Alison Pearks-Wilkerson[1], Tess Crider[1], Terje Raudsepp[1], Victor Mason[1], Brian Davis[1], Bhanu Chowdhary[1], Malcolm Ferguson-Smith[2], Patricia O'Brien[2], Tina Graves[3], Derek Albracht[3], Wesley Warren[3], William Murphy[1]
[1]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA,[2]Cambridge Resource Centre for Comparative Genomics, Cambridge, UK,[3]Washington University Genome Sequencing Center, St. Louis, MO, USA

Y chromosomes have arisen independently in divergent evolutionary lineages across the eukaryotic tree of life. Despite numerous origins, these chromosomes appear to share several fundamental properties, such as 1) they contain genes that may play a role in sex determination, 2) their DNA content is highly repetitive, and 3) they accumulate male benefit genes that enhance gametogenesis. While many genes are known to be Y linked, the actual number of Y chromosomes that have been sequenced is extremely small in relation to the bewildering rate at which genomes are presently being sequenced. The most completely sequenced and annotated Y chromosomes are from two recently diverged eutherian mammals, human and chimpanzee, and offer a powerful glimpse into the immense structural variation and complexity that can emerge on the Y within a very short period of evolutionary time. However, this lack of phylogenetic scope has hampered identification of broader evolutionary patterns, the putative structure and gene content of the ancestral Y chromosome, and identification of conserved non-coding elements that may regulate Y-linked genes. Here we describe the first sequence-based analysis of two carnivoran Y chromosomes: the domestic cat, *Felis silvestris catus*, and domestic dog, *Canis lupus familiaris*. These two carnivores diverged from each other ~55 Mya, and from primates ~92 Mya, hence providing a phylogenetically distinct vantage point with which to interpret the evolutionary patterns observed in the recent human and chimpanzee y chromosome comparisons. Further, by including the mouse X-degenerate sequence, our 5-way multispecies comparison allows for the first accounting of levels of selection constraining the evolution of mammalian Y chromosomes, as well as the divergent and convergent evolutionary forces that allow for the retention of sequences in a non-recombining genomic environment. These studies extend observations that Y chromosomes have undergone radical restructuring across and within mammalian orders, and harbor many lineage-specific genes that are important for fertility as well as other non-reproductive functions.

P-1275

Sequencing and analysis of the whole genome of the guppy, a model for natural variation and sex chromosome evolution

Axel Künstner, Margarete Hoffmann, Christine Dreyer
Max Planck Institute for Developmental Biology, Tübingen, Germany

The teleost guppy (Poecilia reticulata) has been widely studied because of its natural variation and adaptation in morphology, behaviour, and life history traits observed within and between different populations. Due to the strong genetic linkage of some male-advantageous pattern formation genes to the putative sex-determining locus on Y, the guppy may be regarded as a model for sex chromosome differentiation and evolution. To augment the molecular tools for studying the genetic basis of such variation, we initiated a next generation (Illumina) whole genome-sequencing project of the guppy. First, we sequenced the genome of a female from the Guanapo population (West Trinidad). Using SOAPDENOVO software, we assembled 690MBb of sequence into 4,684 scaffolds with an L50 of 1.4Mb. We also link the assembled genomic scaffolds to a dense genetic map that comprises >5000 SNP markers on the 23 linkage groups. Additionally, it is planned to do whole genome alignments to closely related fish genomes to study species-specific adaptations and features of the guppy genome.

To gain information on the highly differentiated and repetitive Y chromosome, we have also sequenced the genome of a closely related inbred male Guanapo guppy, and will also sequence genomes of male guppies of different clades. Together with RNA-seq data from male and female guppies, this project provides a genomic resource to get deeper knowledge of the genetic differences within the guppy species.

P-1276

The Relative Contributions of Selection for Speed and Accuracy of Translation to Codon Bias in Bacteria

WENQI RAN[1], PAUL HIGGS[2]
[1]NCBI,NIH, BETHESDA,MD, USA, [2]McMaster Univ, Hamilton, Canada

It has been known for several decades that translational selection influences the frequencies of synonymous codons in the genes of many organisms, although the causes for this are still under debate. Several commonly used measures of codon bias are not easy to compare among organisms. Here we present a simple statistical method for assessing the relative strength of selection on codon usage in different organisms and comparing the relative contributions of selection on speed and accuracy to codon bias. Among bacteria, we have previously shown that species that are capable of rapid growth usually have stronger selection on codon usage than slow growing species, and also possess higher numbers of rRNA and tRNA genes. This suggests that fast-growers are adapted for fast protein synthesis because translational speed has a direct influence on the rate of cell growth and division. There is also considerable evidence that codon usage is influenced by accuracy of translation, and some authors have argued that accuracy is more important than speed. Here we use a comparison of codon usage in high- and low expression genes as a measure of selection for speed and a comparison of codon usage at conserved and variable sites within high expression genes as a measure of selection for accuracy. Across a large sample of bacterial genomes, both effects are clearly visible, although the speed effect appears to be much stronger than that of accuracy and is found to be significant in a larger proportion of genomes. It is also difficult to explain the correlation of codon bias with growth rates and numbers of copies of tRNA and rRNA genes on the basis of selection for accuracy. Hence we conclude that that selection for translational speed is a dominant effect in driving codon usage bias in fast-growing bacteria.

P-1277

Parallel evolution of *rpoB* mutants of *Escherichia coli* during thermal stress adaptation

Alejandra Rodriguez Verdugo[1], Brandon Gaut[1], Pamela McDonald[1], Rebecca Gaut[1], Albert Bennett[1], Anthony Long[1], Olivier Tenaillon[1,2]
[1]*University of California Irvine, Irvine, CA, USA,* [2]*INSERM U722 and Université Paris 7, Paris, France*

Microbial evolution experiments allow us to understand better the dynamics and genetic basis of adaptation, including the contribution of individual mutations to fitness improvement. To assess the genetics of adaptation, we evolved 114 replicate populations of *Escherichia coli* B for 2000 generations at a high and stressful temperature (42℃). Analyses of the genomes revealed parallel mutations within the *rpoB* gene encoding the beta subunit of RNA polymerase (RNAP). Some of these amino acid substitutions, present in 12 of 114 evolved populations, are in the active site of the RNAP and confer resistance to the antibiotic rifampicin. To investigate the emergence of resistance, we estimated the frequency of rifampicin-resistant clones through time. We find that the resistance to rifampicin appeared before 500 generations in most of the lines. To our surprise, the resistance appeared before 100 generations in one line, suggesting a strong selective advantage of the *rpoB* mutation under our experimental conditions. To measure the selective advantage of these mutations directly, we moved the single amino-acid changing mutations into the ancestral background. With this construct, we could compare the effect of the *rpoB* mutations against the common ancestor, confirming that the *rpoB* mutations confer a fitness advantage ranging from 18.2 to 30.2% at 42℃ in both the ancestral background (REL1206) and the REL606 background. At 37℃, however, these *rpoB* mutations come with a fitness cost of 5.7%. The pleiotropic effects observed suggest that the high advantage confer by the *rpoB* mutations is specific to our experimental conditions.

P-1278

The genomics of *Acinetobacter baumannii*: insights into genome plasticity, antimicrobial resistance and pathogenicity

Luisa Antunes[1], Francesco Imperi[2], Jochem Blom[3], Laura Villa[4], Alessandra Carattoli[4], Paolo Visca[1]
[1]Department of Biology, University Roma Tre, Rome, Italy, [2]Department of Biology and Biotechnology Charles Darwin, Sapienza University of Rome, Rome, Italy, [3]Computational Genomics, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany, [4]Department of Infectious, Parasitic and Immune-Mediated Diseases, Istituto Superiore di Sanità, Rome, Italy

Multidrug-resistant *Acinetobacter baumannii* is rapidly becoming the prevalent agent threatening intensive care units, where it is a major cause of healthcare-associated infections. In the last years, the genome sequences of a number of *A. baumannii* strains, including representatives of main epidemic international clonal complexes, were determined, and several others are in progress. In this work 12 available *A. baumannii* chromosomal genome sequences were compared for orthologous protein coding sequences (CDSs), using the EDGAR comparative genomics software. Whole-genome-based analysis of their phylogenetic relationships confirmed the existence of two main clusters, corresponding to the large clonal complexes 1 and 2, as well as minor lineages composed of individual strains. The calculation of the *A. baumannii* pan genome revealed an open pan genome structure, with an impressively large and expanding dispensable genome, and a rather undersized core genome. Analysis of the content of the core genome evidenced a series of putative virulence determinants, many of which encode proteins involved in drug-resistance, iron uptake, motility and adhesion. Comparative analysis of the dispensable genome predicted that epidemic *A. baumannii* strains have not acquired additional virulence determinants with respect to non-epidemic (sporadic) isolates. This work confirms the multifactorial nature of *A. baumannii* virulence, and provides evidence that the clinical success of *A. baumannii* mainly relies on its ability to persist in the hospital setting and to survive antibiotic treatment. Its apparent low-grade pathogenicity could represent a strategy to persist in health-care facilities. However, the high plasticity of the *A. baumannii* genome might be a favorable prerequisite to the acquisition of novel virulence traits, ultimately transforming this opportunistic pathogen into an even greater threat to human health.

P-1279

Codon usage in large modular proteins in bacteria

John Cullum[1], Heba Al-Hashmi[1], Mohamed Lisfi[1], Jurica Zucko[3], Daslav Hranueli[3], Paul Long[2]
[1]University of Kaiserslautern, Kaiserslautern, Germany,[2]King's College, London, UK,[3]University of Zagreb, Zagreb, Croatia

Modular polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS) consist of large polypeptides, which can have more than 10000 amino acids. The polypeptides usually contain several modules each divided into domains. The fundamental structure of each module is similar, but they differ in detail according to the nature of the extender unit that they incorporate. A large number of PKS and NRPS clusters are known, which provides a lot of data for testing evolutionary hypotheses about codon usage. Both types of clusters are present in actinomycetes, which have a high G+C-content and a highly biased codon usage. Many NRPS clusters are also present in *Bacillus* species and PKS clusters are common in myxobacteria. Comparison of the clusters in these distantly related bacteria with the similar clusters in actinomycetes shows whether the observed codon usage phenomena are determined by the evolutionary pressures on the proteins or caused by the host organisms.
Use of common codons for an amino acid is associated with high efficiencies of translation, which can be estimated using the codon adaptation index (CAI). The CAI was calculated in sliding windows for the modular proteins and showed characteristic short regions of slow translation at conserved positions in modules, which may be associated with allowing time for protein folding. Selective pressures acting on the modules and domains were estimated using the ratio of non-synonymous/synonymous codons in sliding windows. This confirmed that most of the sequences had low ratios suggesting strong purifying selection.

P-1280

Phylogenetic and Environmental Influences on the Saliva Microbiome of Pan and Homo

Jing Li[1], Ivan Nasidze[1], Dominique Quinque[1], Mingkun Li[1], Michel Halbwax[1], Anne Fischer[1,2], Mark Stoneking[1]
[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Microbiology and Biotechnology Department, International Center for Insect Physiology and Ecology, Nairobi, Kenya

A major effort is underway to categorize the human microbiome and understand the factors that can influence the distribution of microbial taxa within and among individuals. Interspecies comparisons can help sort out the relative influence of phylogenetic (i.e., vertical transmission) vs. environmental (i.e., horizontal transmission) factors on the microbiome. Previous studies have documented substantial diversity in the human saliva microbiome; to investigate the relative importance of phylogenetic vs. environmental factors on saliva microbiome diversity, we here analyze the saliva microbiomes of chimpanzees (*Pan troglodytes)* and bonobos *(Pan paniscus)* from two sanctuaries in Africa, and from human workers from each sanctuary. We find that the saliva microbiomes of the two *Pan* species are more similar to one another, and the saliva microbiomes of the two human groups are more similar to one another, than are the saliva microbiomes of human workers and apes from the same sanctuary. These results suggest that phylogeny (i.e., vertical inheritance) plays a substantial role in determining the saliva microbiome of *Pan* and *Homo*, as found previously for the fecal microbiome of wild apes. We also studied the saliva microbiome from chimpanzees, bonobos, gorillas, and orangutans from the Leipzig Zoo, and found an extraordinary diversity in the zoo ape saliva microbiomes that is not found in the saliva microbiomes of the sanctuary animals. These results suggest that microbiome analyses based on captive animals should be viewed with caution, as they may not reflect the microbiome of animals in the wild.

P-1281

Rewiring of the bacterial core gene regulatory network by homologous recombination may facilitate niche adaptation

Yaara Oren, Tal Pupko
The Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel-Aviv University,, Tel-Aviv, Israel

Comparative genomics studies often search for genes that are unique to a certain bacterial strain, which may account for its specific niche adaptions. However, there is a growing number of bacterial phenotypic adaptations that comparative genomics approaches fail to explain. We hypothesize that this stems from the fact that genes shared among strains are largely overlooked in such approaches and that differences in regulation may account for such differences. To test this hypothesis, we developed a "comparative regulomics" approach, in which regulatory regions of genes that are shared among all strains (i.e. core genes) are analyzed. We specifically searched for evolutionary changes in such regions that might have phenotypic consequences. We show that the core genome of E. coli (both genes and their regulatory regions) is highly dynamic and is characterized by extensive horizontal gene transfer (HGT). We found specific cases of "promoter switching" in which evolutionary distant strains share common regulatory motifs as a result of this HGT. Further, there is enrichment for metabolic genes in the group that exhibits such "promoter switching". In addition, we could show that many of these switching events lead to changes in transcription factor binding sites. Taken together, our results suggest that homologous recombination is a major force shaping bacterial core gene regulatory network and may thus enable niche adaption.

P-1282

Adaptation of intracellular parasitic bacteria to acidic vacuoles

Pedro M. B. M. Coelho[1], Mafalda Silva[2], Miguel C. Seabra[2], José B. Pereira-Leal[1]
[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal, [2]Centro de Estudos de Doenças Crónicas, Faculdade de Ciências Médicas da Universidade Nova de Lisboa, Lisboa, Portugal

How do parasitic bacteria adapt to their hosts? Some parasites can opt for an (facultative) intracellular life, others lost that choice (obligate intracellular parasites) and experienced extensive gene loss. There is considerable evidence for general loss of biosynthetic pathways and transcriptional regulation, but comparisons between different intracellular parasites suggest additional niche-specific adaptations that have never been elucidated.

Bacterial parasites frequently invade the host cell through the endocytic/phagocytic pathway, where they employ different strategies to subvert it and reach their chosen replication niche. While the nature of this replication niche is frequently unclear, we can classify them in four main classes: cytosol, endocytic compartment, vacuole without endocytic character, and phagolysosomes. Therefore, an open question is whether there are common adaptations to common niches. We will ascertain whether an acidic intracellular niche determines specific parasite adaptations.

To address this question, we compiled a list of sequenced parasites, their taxonomic classification, facultative vs. obligate nature, specific intracellular niche(s) occupied, targeted cell type, and host range. Then, we were able to compare the genome of organisms that occupy an acidic environment, whose niche is marked with the host protein lysosomal-associated membrane protein 1 (LAMP1+), with the genome of organisms that occupy a non-acidic environment (LAMP1-).

Our strategy involves identifying the protein families that distinguish the genome of the intracellular parasites in different niches. Protein families, defined based on structural domain architectures, represent groups of proteins that share a common evolutionary history and for which there is conservation of structure and biochemical function. We considered that a protein family/domain is part of the acidic genomic signature if their frequency distributions between LAMP1+ and LAMP1- organisms differs significantly and the median is higher for LAMP1+ organisms, and it is not a by-product of phylogenetic bias.

Using this strategy, we assigned a genomic signature to intracellular parasites that inhabit an acidic environment that is composed of 10 proteins families and 3 domains which include genes involved in free radical defense, specialized nutrient acquisition and replication regulation, among others. This signature is distinct from one we determined for free-living bacteria that inhabit extreme acidic environments, reflecting different biophysical and nutrient availability constraints.

We further observed that closely related organisms do not necessarily occupy similar niches, nor do obligate intracellular parasites occupy the same niche as their related facultative intracellular parasites. This suggests that changes of replication niche may be possible even in organisms that have suffered extreme adaptations.

P-1283

Transposable element evolution in Wolbachia bacterial endosymbionts

Nicolas Cerveau, Didier Bouchon, Richard Cordaux
CNRS/Université de Poitiers, Poitiers, France

Transposable elements (TE) are one of the major driving forces of genome evolution, raising the question of the long-term dynamics underlying their evolutionary success. Long-term TE evolution can readily be reconstructed in mammals thanks to many degraded copies constituting genomic fossil records of past TE proliferations. By contrast, bacterial genomes usually experience high sequence turnover and short TE retention times, thereby obscuring ancient TE evolutionary patterns. The genomes of Wolbachia, one of the most abundant bacterial endosymbionts on Earth, are littered with TEs. The question arises as to why there are so many TEs elements in the genomes of this ancient endosymbiont. To address this question, we investigated the dynamics of insertion sequence (IS) TEs using an evolutionary perspective. Our results indicate that several processes may explain TE abundance in Wolbachia, including recent activity, along with recurrent invasions through horizontal transfers and gene conversion. Remarkably, we found that 70% of Wolbachia IS are nonfunctional. They constitute an unusual bacterial IS genomic fossil record providing direct empirical evidence for a long-term IS evolutionary dynamics following successive periods of intense transpositional activity. The identification of an important IS genomic fossil record in Wolbachia demonstrates that IS elements are not always of recent origin, contrary to the conventional view of TE evolution in prokaryote genomes.

P-1284

The evolutionary consequences of DNA replication onto the mutational profile of yeast genomes

Gilles Fischer, Nicolas Agier, Fabrice Touzain, Marco Cosentino Lagomarsino, Orso Maria Romano
Université Pierre et Marie Curie, Paris, France

We will present some evidences that the mutational profile of yeast genomes is shaped by DNA replication. We showed that substitution rate increases with replication timing by more than 30% between the earliest and the latest replicating regions in the genome of *S. cerevisiae*. We also found a mutational asymmetry associated with the polarity of replication resulting in higher rates of substitutions towards C and A than towards G and T in leading strands. Such mutational asymmetries applied over long evolutionary periods have generated a compositional skews between the leading and the lagging strands in the present day genome of *S. cerevisiae*. In addition, we discovered a very unusual nucleotide composition in the genome of another yeast species, *Lachancea kluyveri,* where a whole chromosomal arm harbors a GC content very different from the rest of the genome. We established the complete temporal program of replication in this genome. We identified all replication origins and show that the GC-rich chromosomal arm present unique replicative properties that are linked to its unusual base composition.

P-1285

High rates of retrogene formation in the giant nuclear genomes of dinoflagellates.

Renny Lee, Claudio Slamovits
Dalhousie University, Halifax, Nova Scotia, Canada

Dinoflagellates comprise a large and diverse group of protists, consisting of over 2500 described species in 125 genera. They reside in all aquatic environments and include photosynthetic and heterotrophic species. Important ecological attributes include bioluminescence, algal-coral symbioses, major contribution to marine phytoplankton, and toxic algal blooms. Another remarkable hallmark of dinoflagellates is the possession of enormous nuclear genomes, with size estimates at over 200 Giga bases. Other oddities in their nuclear biology are permanently condensed chromosomes and the apparent absence of nucleosomes. At the level of gene organisation they exhibit some unusual characteristics such as long tandemly-arranged protein-coding gene repeats for some genes, and the requirement for every mRNA transcript to be trans-spliced to a common 22 bp leader (Dino-SL).

A previous analysis (Slamovits and Keeling, 2008) on a limited set of dinoflagellate ESTs revealed the presence of additional truncated and degenerated sequence (termed 'SL relics') with similarity to Dino-SL, immediately downstream and adjacent to the Dino-SL. Being genome-encoded, the SL relics are hypothesized to originate when the trans-spliced mRNA converts into cDNA and reintegrates back into the genome (cDNA recycling). Taking advantage of the increasing availability of sequence data, we have reassessed the prevalence of cDNA recycling. Bioinformatically, motif-searching for Dino-SL and SL relics has led us to conclude that recycled transcripts comprise a significant proportion of dinoflagellate transcriptomes. Indeed, this high level of retrogene formation seems unprecedented. In order to gain deeper insight into the mechanisms and consequences of cDNA recycling we also examine other correlations, such as those between recycled genes and functional category, expression levels and copy numbers. The work discussed in this presentation provides important evidences for the involvement of cDNA recycling in shaping the remarkable genomes of these extraordinary organisms.

P-1286

The evolution of CRISPR-Cas in *Mycobacterium tuberculosis*

Daniela Brites, Mireia Coscolla, Sebastien Gagneux
Swiss Tropical and Public Health Institut, Basel, Switzerland

The genomic regions encoding CRISPR-Cas (clustered regularly interspaced short palindromic repeats - CRISPR associated proteins) are involved in generating adaptive immunity against invading genetic elements in many bacterial species. In Mycobacterium tuberculosis, the causative agent of human tuberculosis, the CRISPR-Cas region has a relatively high level of polymorphism in the number of repeats and spacer sequences. Yet, a relation between the CRISPR-Cas region in M. tuberculosis and immunity to foreign genetic elements has not been demonstrated. A search for an homologue region in other related mycobacteria revealed that the CRISPR-Cas region is only present in M. tuberculosis and in M. canetti, the putative ancestral species from which M. tuberculosis evolved. This suggests that the CRISPR-Cas region might have functional importance and that the genetic variation found might be of relevance for M. tuberculosis populations. We have measured the genetic variation of the CRISPR region in M. tuberculosis strains collected worldwide, both by assessing the variation in the numbers of repeats and spacers as well as indels and SNPs in the Cas associated proteins. These results allow to us to infer the mode of evolution of this genomic region and bring some insight into its function.

P-1287

Phylogeny of Prokaryotic Cyanobacteria: a tale of two species

Cristiana Moreira[1,2], Vitor Vasconcelos[1,2], Agostinho Antunes[1,2]
[1]CIMAR/CIIMAR/ Laboratory of Ecotoxicology, Genomics and Evolution, Porto, Portugal,[2]Faculty of Sciences University of Porto, Porto, Portugal

Cyanobacteria are photosynthetic prokaryotic microorganisms that exist in our planet for over 3 billion years. Interest in studying this group of microorganisms is due to the production of toxic secondary metabolites that can kill animals including humans. The cyanobacteria species that produce them are ubiquitous and can be found in aquatic, terrestrial and aerial environments. Two of the most studied cyanobacterial species associated with toxicological incidents are the colonial *Microcystis aeruginosa* and the filamentous *Cylindrospermopsis raciborskii*. Both have distinct life histories with the first having a cosmopolitan distribution and the second occurring mainly in tropical ecosystems being currently also an invader in both subtropical and temperate environments. Topics addressing their genetic diversity, population structure and phylogeography in a worldwide perspective are reduced and knowledge about its origin and evolution are unknown. In this study we assessed the phylogenetic relationships within these two species by comparing isolates obtained from all of the five continents using a concatenated system comprehending four distinct genetic markers. Our results will provide insight to understand how these species have evolved through time, its population structure and how phylogeographic patterns may explain their current distribution.

P-1288

Evolutionary pathways of *Pseudomonas aeruginosa* during long-term infection of CF airways

Trine Markussen[1], Rasmus Lykke Marvig[1], Niels Høiby[2,3], Helle Krogh Johansen[2], Søren Molin[1], Lars Jelsbak[1]
[1]Department of Systems Biology, Technical University of Denmark, Lyngby, 2800, Denmark, [2]Department of Clinical Microbiology, University Hospital, Rigshospitalet, Copenhagen, 2100, Denmark, [3]Institute for International Health, Immunology and Microbiology, University of Copenhagen, Copenhagen, 2200, Denmark

The dominant cause of premature death in patients suffering from cystic fibrosis (CF) is chronic lung disease caused by infections with *Pseudomonas aeruginosa*. Chronic lung infections often last for decades with single clones. During chronic lung infection, the bacterial population undergoes significant genetic changes as a result of multiple selective pressures. At the Copenhagen CF Clinic two dominating genotypes, DK1 and DK2, have been found to colonize and persist in many patients suffering from CF. Investigation of the DK2 lineage showed that adaptation leading to persistence is not associated with high-level population diversity. Instead it seems that DK2 has reached a major adaptive peak in the fitness landscape that made it possible for the lineage to colonize all major niches of the CF airways and in several different hosts.

In the present study adaptation and evolution of the dominating opportunistic *P. aeruginosa* DK1 lineage was phenotypically and genotypically investigated using isolates from six CF patients covering a time span of 38 years corresponding to ~200,000 generations. Isolates were phenotypically characterized according to colony morphology, motility, quorum sensing, virulence, doubling time, and catabolic functions. Furthermore, genome sequencing was performed on isolates in order to characterize the evolutionary trajectory associated with the DK1 lineage.

Phenotypic examination revealed prevalence of mutators, reduced virulence and decline in catabolic function during infection. These data corroborate the genome sequencing results showing deep branching of the phylogenetic tree. Surprisingly, there was maintenance of antibiotic susceptibility over the entire infection period.

This study shows that adaptation of the DK1 lineage to life in the CF lung is achieved by high degree of population diversity that made it possible for different subpopulations to co-exist as sub-niche specialists.

P-1289

Empirical fitness landscapes and the predictability of evolution

Lilia Perfeito[1,2], Stephan Schiffels[1,3], Michael Lässig[1]
[1]Institute of Theoretical Physics, University of Cologne Physics, Cologne, Germany,[2]Instituto Gulbenkian de Ciência, Oeiras, Portugal,[3]Welcome Trust Sanger Institute, Hinxton, Cambridge, UK

Natural selection acts on genes through the phenotypes they encode. Population genetics has made significant progress in understanding the evolution of genomes, but we can now explicitly link sequence changes to phenotypes and fitness. We have studied the fitness effects of two generic phenotypes: protein production and protein activity. For the lac utilization pathway in Escherichia coli, we can manipulate these phenotypes and build an empirical fitness landscape. A simple model of the lac pathway, based on elementary biophysical processes, predicts the growth rate of all observed genotypes and environments and explains a striking non-linearity in the landscape. We use this empirical fitness landscape to design selective conditions where we expect one or two of these phenotypes to evolve. We follow the evolution of different genotypes, located at different points on the fitness landscape as they adapt to their environment. In order to observe how populations evolve, we measure changes in the phenotypes related to the lac operon in combination with whole-genome sequencing. We obtain a quantitative picture of the adaptive process jointly at the sequence and at the phenotype level. This approach quantifies our understanding of the fitness effects of mutations, epistasis and genotype-environment interactions. It will also allow us to quantitatively predict how populations will evolve.

P-1290

Comparative genome analysis of Microsporidia reveals gain and expansion of gene families

Sirintra Nakjang[1], Eva Heinz[1], Tom A. Williams[1], Christophe J. Noël[1], Daniel C. Swan[1], Alina V. Goldberg[1], Simon R. Harris[1], Thomas Weinmaier[2], Stephanie Markert[3], Tal Dagan[4], Thomas Schweder[3], Thomas Rattei[2], Neil Hall[5], Robert P. Hirt[1], T. Martin Embley[1]
[1]Newcastle University, Newcastle Upon Tyne, UK, [2]University of Vienna, Vienna, Austria, [3]Ernst-Moritz-Arndt-Universitaet Greifswald, Greifswald, Germany, [4]Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [5]University of Liverpool, Liverpool, UK

Microsporidia are a group of highly successful obligate intracellular parasites, related to fungi, which affect a wide range of hosts and present a variety of disease phenotypes. Their genomes sizes vary, but tend to be compact and lack many key metabolic genes; *Encephalitozoon cuniculi* (2.5Mb) has the smallest known eukaryotic genome. The reduction of biological processes and metabolic pathways are typical characteristics of intracellular parasites that rely on their hosts for nutrients. Here, we present a comparative analysis of the Microsporidia genomes that have been sequenced to date, including the recently sequenced genome (in our lab) of *Trachipleistophora hominis* (8.5Mb). In contrast to *E. cuniculi*, *T. hominis* has a greater number of transporters, peptidases and other metabolic enzymes. Also present in *T. hominis*, but absent in many other Microsporidia are transposable element encoded proteins and the machinery for RNA interference. In addition to gene loss during the reductive genome evolution of Microsporidia, our comparative analysis reveals the presence of lineage-specific innovation and selective expansion of their protein repertoires. These expansions may be due in part to adaptation to the evolutionary pressures encountered by Microsporidia in their different hosts. Our analyses show that the microsporidian common ancestor underwent extensive proteome reduction as well as the acquisition of new genes. Subsequently, some microsporidial lineages then underwent further genome compaction.

P-1291

A site-dependent evolutionary model for CRISPR spacer arrays

Anne Kupczok, Jonathan P. Bollback
IST Austria, Klosterneuburg, Austria

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is a locus found in Eubacteria and Archaea. It consists of an array of repeats and spacers that acts as an adaptive heritable immune system. The spacers between the repeats represent viral/plasmid targeting sequences and the system functions in an analogous way to the eukaryotic siRNA system. The length and content of the spacer array varies considerably among individuals within species (suggesting a rapid arms race) and it has been suggested that there is a selective cost, in the absence of parasites, associated with maintaining these arrays.
Therefore, the rate at which spacers are gained and lost from these arrays provides insight into the evolutionary dynamics of host-parasite interactions. To this end we develop a probabilistic model for the change of CRISPR spacer arrays over time. To describe biological observations, the model
differs in two ways from standard phylogenetic models. First, insertion occurs only at the beginning of the array and second, one deletion event can affect multiple consecutive spacers. Parameter estimation under the model is done by maximum likelihood accounting for unobserved insertions and deletions. Thereby all pairs of arrays having spacer overlap are evaluated. The set of valid ancestors for each pair includes valid orderings and additional spacers deleted in both lineages.
Simulating under this model shows that the ratio of insertions over deletions can be recovered well in the estimation. Furthermore, the model differs considerably from site-independent models. The site-dependent and site-independent model are compared for bacterial CRISPR data sets.

P-1292

Modelling the Growth and Transmission of Infectious Disease: Linking epidemiology and population genetics

Bethany Dearlove, Daniel Wilson
University of Oxford, Oxford, UK

Understanding the transmission of infectious disease is important for monitoring outbreaks, informing public health policy, and improving intervention strategies. Traditionally the fields of population genetics and epidemiology have been studied separately; however it is clear that using genetic information alongside epidemiological models has great potential for understanding the dynamics of infectious disease. Directly estimating epidemiological parameters such as transmission rates can be difficult, as it relies on comprehensive monitoring during an outbreak where relevant processes may be hidden or undetectable. However, genetic information provides an alternative window into the past. Here we describe a combined coalescent-based meta-population model for estimating the parameters of the standard susceptible-infected-susceptible (SIS) epidemiological model from genetic data. We apply this model to conduct a meta-analysis of Hepatitis C virus (HCV), with the aim of explaining differences in patterns of genetic diversity between populations in terms of the underlying epidemiological dynamics. We show that there is correlation between diversity and the growth rate of local epidemics. However, differences between datasets in the growth rate of HCV do not appear to be explained by host population size or prevalence of disease. We conclude that global variation in Hepatitis C diversity is driven by other, unobserved factors whose effects are mediated via the intrinsic growth rate of local epidemics.

P-1293

Evolution of the Cas/CRISPR immune locus in *Staphylococcus*

Jonathan P. Bollback[1], Anne Kupczok[1], Flavia G. G. Leite[1,2]
[1]IST Austria, Klosterneuburg, Austria,[2]Medical University of Vienna, Wien, Austria

Eubacteria have evolved numerous defense mechanisms against viruses, plasmids, and other genomic parasites. The majority of these mechanisms are innate, but recently, an adaptive heritable immune system has been discovered in Eubacteria and Archaea called Cas/CRISPR. The Cas/CRISPR system operates in an analogous manner to siRNA anti-viral systems: upon initial infection, short regions of the invading genome are inserted into the host genome's CRISPR array, transcribed, and then used as targets for Cas endonuclease destruction of the invading parasite. Immunity gained in this fashion is then transmitted vertically to daughter cells, providing acquired heritable immunity.

Here, we describe the evolutionary dynamics of this locus in *Staphylococcus intermedius*, *S. pseudintermedius*, and *S. delphini*. Specifically, we address the following questions. What are the parasitic origins of the spacer elements in the CRISPR array, and how rapidly do they experience turnover? Using a probabilistic model, we infer the rates at which spacer elements are gained and lost: high rates of turnover would suggest that bacterial parasites, such as viruses, rapidly evolve escape mutants resistant to the Cas/CRISPR system. Finally, we have detected at least one event in which the locus has mobilized within staphylococcal genomes, a handful of events in which the locus has been inactivated and then regained, and three events of horizontal gene transfer between these staphylococcal species.

P-1294

The Promotion of Microbial Sociality by Horizontal Gene Transfer and Gene Dosage

Sorcha Mc Ginty[1,2], Evandro Ferrada[3], Andreas Wagner[1,2], Daniel Rankin[1,2]
[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zürich, Switzerland, [2]The Swiss Institute of Bioinformatics, Luasanne, Switzerland, [3]The Santa Fe Institute, New Mexico, USA

It is widely recognized that bacteria display a diverse range of social traits. Such traits are exchanged between cells through both vertical and horizontal transmission. Plasmids are ubiquitous among bacteria and act as vectors for horizontal gene transfer. Transmission via plasmids provides many advantages for potentially social traits. Firstly, infection by a plasmid forces a cell to carry particular genes, even if they are costly, as is frequently the case with social traits. Secondly, the spread of plasmids can increase local relatedness by increasing the numbers of carriers of social genes within an interacting population. A third advantage to plasmid-based spread comes from the fact that multiple copies of plasmids are frequently carried within a single cell, thus allowing increased expression of plasmid-based genes.

Here we use secreted proteins as a proxy for sociality and, using a combination of theoretical modeling and analysis of a dataset of >1.2 million proteins from 351 bacterial strains, we investigate the factors which may promote the spread of genes coding for secreted proteins. Our model suggests that plasmid-based gene duplications are more likely to invade than those based purely on vertical transmission. We find that secreted proteins tend to be more recently acquired, suggesting possible horizontal spread, and are more likely to be carried on plasmids than on the bacterial chromosome. Smaller plasmids, which can have higher copy numbers within a cell, also tend to carry proportionately more secreted proteins. Additionally secreted proteins tend to be more highly expressed and less costly. Overall our results suggest that horizontal transmission and high product expression through gene dosage and gene expression levels may promote the spread of secreted proteins and thus social behaviour among bacteria.

P-1295

Analysis of genetic composition of recombined regions in bacteria through whole genome comparisons

Sebastien Leclercq, Daniel Falush
Max Plank Institute for Evolutionary Anthropology, Leipzig, Germany

Contrary to eukaryotes, bacteria essentially use homologous recombination (HR) as a DNA break repair system. Nevertheless, it sometimes produces genetic admixture when foreign DNA is used as a template and is therefore a major process for bacterial genetic innovation. The HR events depend on the genetic composition of the recombined segment, such as the presence of Chi sequences in Escherichia coli. Although a number of molecular mechanisms of recombination have been studied in depth in vitro, the genetic factors that are actually responsible for promoting bacterial HR in nature are still not well understood.

With the advent of the post-genomic era, detection of recombined regions in bacteria at the genomic scale now becomes feasible, by comparing closely related sequenced strains. For example, ClonalOrigin is a method designed to detect recombined segments in a set of related bacterial genomes from their whole alignment. Here we take advantage of the potentially recombined regions provided by ClonalOrigin to study the genetic factors that influence bacterial HR at the genomic scale in different bacterial species.

P-1296

Genome-wide survey of homologous recombination and diversifying selection in an extremely sexual bacterial species

Koji Yahara[1], Mikihiko Kawai[2], Yoshikazu Furuta[1], Noriko Takahashi[1], Naofumi Handa[1], Takeshi Tsuru[1], Kenshiro Oshima[1], Masaru Yoshida[3], Takeshi Azuma[3], Masahira Hattori[1], Ikuo Uchiyama[2], Ichizo Kobayashi[1]
[1]University of Tokyo, Tokyo, Japan, [2]National Institute for Basic Biology, Okazaki, Japan, [3]Kobe University, Kobe, Japan

The nature of a species has been a fundamental and controversial question in biology. The era of genome/metagenome sequencing has intensified the debate for prokaryotes because extensive horizontal gene transfer across species boundaries makes the very existence of separate species questionable. Currently, more than a dozen attempts have been made to establish a conceptual framework to identify species in prokaryotes. However, none of these models has been examined based on whole genome sequence data. In the present work, we conducted genome-wide survey of homologous recombination and diversifying selection in the extremely sexual bacterial species *Helicobacter pylori*.

Through extensive analysis of both the multiple genome alignment and entire dataset of one-to-one orthologous genes of global strains of *H. pylori*, mosaic structures from recombination events within the species were detected throughout the genomes. Gene trees with identical topology to the core tree were rare and those with deviated topology were conspicuous and scattered. Almost all (99.8%) of these genes including the "core" set of genes and horizontally transferred genes showed a signal of at least one recombination event.

Meanwhile, genome-wide search for diversifying selection identified genes (about 3% of all the orthologous genes) with adaptive amino acid changes. It was as expected that they included genes for host interaction such as lipopolysaccharide synthesis, outer membrane proteins, and toxins. More remarkably, it also identified genes involved in maintenance of genome, epigenome and proteome. For example, genes for recombination/repair and those responsible for translation fidelity were included, which suggests a new mechanism of diversification at genome/ epigenome/ proteome level.

These results provide a genome-wide view of homologous recombination and diversifying selection in a bacterial species. Our results may lead to re-examination of the concepts of the species.

P-1297

Comparative genomics of louse primary endosymbiotic bacteria

Bret Boyd, Julie Allen, David Reed, Valerie de Crecy-Lagard, Lauren McIntyre
University of Florida, Gainesville, Florida, USA

The sucking lice of mammals (Insecta: Anoplura) possess obligate associations with intracellular endosymbiotic bacteria to supplement their diet with vitamins. The life-history strategies of louse species are canalized as blood feeding ectoparasites; therefore, we expect their endosymbionts to provide the same or very similar functions for their hosts. The louse endosymbiont system is a unique system in which distantly related bacteria have recently been acquired as new endosymbionts in closely related louse species, with at least 10 separate origins. This is unlike other insect-bacterial endosymbiont systems were associations have persisted for 50-200 million years and endosymbiont replacements are rare. Recently acquired bacterial endosymbionts undergo a massive genome reduction with a corresponding increase in substitution rate. They maintain the genes necessary for mutualism and those necessary for life inside the host. Here we have a system where multiple independent lineages of bacteria evolved to fill the same symbiotic role with their hosts. By examining the genomes of distantly related louse endosymbionts we can determine whether they are using the same genes in the same way, or if they arrived via different evolutionary paths. Further we can also get a more in-depth look at the process of genome reduction in these endosymbionts. We have sequenced the genomes of distantly related louse endosymbionts and examined genomes for commonalities in gene content. Our results provided insights into functional roles of louse endosymbionts, as well as insight into gene retention by bacteria entering into obligate symbiosis.

P-1298

A whole genome phylogeny of the staphylococci

Apurva Narechania[1], Paul Planet[2,1], Sam Boundy[3], Rob DeSalle[1], Gordon Archer[3], Barry Kreiswirth[4]
[1]American Museum of Natural History, New York, NY, USA, [2]Columbia University, New York, NY, USA, [3]Virginia Commonwealth University, Richmond, VA, USA, [4]Public Health Research Institute, Newark, NJ, USA

Using next generation sequencing techniques, we sequenced 87 new *Staphylococcus* strains from the species *S. aureus* (n=51), *S. epidermidis* (n=27), *S. capitis* (n=1), *S. lugdunensis* (n=4), and *S. hominis* (n=1), and *S. warneri* (n=1). These genomes nearly double the number of draft genomes currently available in public databases, allowing for the largest whole genome phylogenetic analysis of a single prokaryotic genus to date. We combined data from these 87 new *Staphylococcus* genomes with publicly available sequences from 102 strains. Distance-based analysis of raw sequence reads from this collection produced a tree that was in complete agreement with recognized species, clonal complexes, and sequence types. Further maximum likelihood and parsimony phylogenetic analysis using concatenated gene datasets revealed robust phylogenetic signal that agreed with known relationships. Mapping of SCCmec types onto the phylogeny strongly suggested independent acquisitions of same methicillin resistance element in disparate lineages, implying that SCCmec types are not strongly associated with particular clones. Sequence diversity in this dataset was greatly increased by the addition of the new genome sequences. For instance, within the *S. aureus* clade sequence diversity increased between 2 and 10 fold per gene. We hope that further analysis of these data will yield useful insights into the evolution of this important genus.

P-1299

Alignment-free Comparison of Microbial Genomes

Bernhard Haubold[1], Mirjana Domazet-Loso[2], Peter Pfaffelhuber[1]
[1]Max-Planck-Institute for Evolutionary Biology, Ploen, Germany,[2]University of Zagreb, Zagreb, Croatia,[3]University of Freiburg, Freiburg, Germany

The first step in almost all genome comparisons is to align the sequences in question. However, despite much progress in the development of algorithms and software tools, genome alignment remains computationally challenging. We have therefore developed an alignment-free strategy for genome comparison over the past few years. It is based on the distribution of the lengths of exact matches between two or more genomes. Thanks to modern string indexing algorithms, this distribution can be looked up in time that grows only linearly with the number of nucleotides analyzed. Moreover, it is independent of synteny and can therefore be looked up from sets of unmapped contigs. This algorithmically optimal strategy can be implemented in very fast computer programs. I will present three such programs, which we have developed to address the following three problems, given a set of unaligned microbial genomes: (i) estimate the pairwise genetic distances for subsequent phylogenetic analysis; (ii) identify candidate regions for horizontal gene transfer; (iii) calculate local genetic diversity in sliding windows. Finally, I discuss the prospect of using our approach to look for recombination in unaligned bacterial genomes.

P-1300

Tit for tat - genome analysis of the bacterial endosymbionts of reed beetles (Donaciinae)

Gregor Kölsch
Zoological Instutute, University of Hamburg, Hamburg, Germany

In most mutualistic symbioses of insects and intracellular bacteria, the endosymbionts provide additional nutrients to a host that feeds on an unbalanced diet. In contrast to this, we investigated an insect-bacteria relationship with a primarily non-nutritional basis. The reed beetles (Coleoptera, Chrysomelidae, Donaciinae) harbour bacteria that produce a secretion, which is used by the larvae for building a cocoon for pupation in the sediment under water. Beetles and symbionts went through a long period of strict co-speciation that lasted approximately 100 million years (age of the subfamily Donaciinae), which marks an intermediate age for associations of bacteria and insects. The bacteria belong to the Enterobacteriaceae (gamma-proteobacteria), more specifically to a clade that is made up of various symbionts of invertebrates (among them Buchnera). Both the intracellular endosymbiotic life style in general and the specific type of interaction (provision of the secretion) motivated our genome sequencing project to elucidate the evolution of the bacterial genome. We hypothesized that the bacterial genome experienced considerable reduction, and that the specific gene content is related to this peculiar interaction. It is of interest if the bacterial genome bears the signature of additional, nutritional interaction between host and symbionts. The genome will form the basis of investigations into the stage specific transcription pattern. At the time of writing, the project is at an advanced stage of assembly with interesting results emanating from a preliminary annotation of the large contigs. In addition to the data on genome evolution, the potential of this new study system for transcription analysis at different life stages of host and symbionts, as well as for comparative genomics will be discussed.

P-1301

Evolution of the cichlid fishes gut microbiota

Laura Baldo, Walter Salzburger
University of Basel, Basel, Switzerland

One of the most fascinating and still poorly understood symbioses has evolved between both vertebrate and invertebrate hosts and complex microbial communities that colonize the internal body surfaces of the digestive tract, where they form the gut microbiota. Recent research indicates both diet and phylogeny as crucial predictors of the gut microbiota features, while the relative contribution of these two factors in recapitulating microbial communities relationships is currently under debate. Therefore the study of phylogenetically closely related species that show a large differentiation of diets can represent an especially interesting case to explore microbiota dynamics in response to concurrent phylogenetic constraints and selective pressures for rapid adaptation to different feeding habits. East African Cichlid fishes are a large group of genetically very closely related species that underwent a rapid dietary niche radiation, with examples of diet convergence/divergence that can be found all along the cichlids tree. Such diet shifts represented a main drive in the process of ecological speciation of this group. Here I provide the first snapshot of the cichlid gut microbiota and explore its dynamics in the context of the cichlid phylogeny and diets.

P-1302

A phylogenomic analysis of replication in the Archaea

Kasie Raymann, Celine Brochier-Armanet, Patrick Forterre, Simonetta Gribaldo
Institut Pasteur, Paris, France

Archaea have unique mechanisms for the transmission of genetic information (translation, transcription, replication), which show similarity to their eukaryotic counterparts.

In previous analyses, we have reconstructed the evolutionary history of translation and transcription and used the sequences of their components to build a robust phylogeny of the archaeal domain.

There are now over one hundred complete archaeal genomes available. Albeit overall resolved, the archaeal phylogeny still contains a few nodes that remain to be clarified, such as the assignement of newly sequences taxa, the position of recently proposed new lineages, and the relationship among archaeal phyla.

Here, we will present results of an exhaustive phylogenomic investigation of all known components of archaeal replication.

We will seek to identify true orthologues versus copies of extrachromosomal origin and use the first to reconstruct the archaeal phylogeny.

P-1303

Microevolution of *Staphylococcus aureus* during progression from carriage to disease

Bernadette Young, Rowena Fung, Elizabeth Batty, Rory Bowden, Derrick Crook, Daniel Wilson
University of Oxford, Oxford, UK

Background: Staphylococcus aureus bacteremia is a common blood stream infection with high mortality. S. aureus nasal carriage is an established risk factor for S. aureus infections, however the mechanisms by which nasally carried S. aureus develops into invasive disease are unknown. Mouse model work suggests small genomic differences can radically alter virulence of S. aureus. Regulatory protein dysfunction has recently been demonstrated to increase mortality in S. aureus bacteremia.

Methods: We use next-generation sequencing to investigate the evolution of a single strain within an individual who carried the organism over 12 months before developing a fatal S. aureus bacteremia in the absence of any surgical risks or invasive devices. Seventy-two isolates were cultured from 7 time points : 6 nasal swab cultures taken over 12 months and 2 blood cultures from the day of admission to hospital with septicemia. Extracted DNA from each isolate was sequenced using the Illumina HiSeq platform. Reads were mapped to the MRSA252 reference genome using STAMPY, with base and variant calls made using SAMTOOLS and bespoke Python scripts. De novo assembly of reads and detection of variants was performed using Cortex.

Results: This analysis found that the isolates were highly similar, with only 29 single nucleotide variants found between the isolates, with no large insertions or deletions acquired during carriage. More than half the observed variants (16/29) occurred in the 23 isolates cultured from the last nasal swab and from blood culture, compared with only 14 variants found between 49 isolates from earlier carriage ($p= 0.03$). Ten variants, including all three of the observed premature stop codons, segregate disease causing isolates from the carriage strains. One premature stop codon occurs in an AraC regulatory protein, the best current candidate for a phenotype-altering mutation.

Conclusion: This case is a unique opportunity to use whole-genome sequencing to investigate micro-evolution of a single strain of S. aureus during both colonisation and disease. We have established that case-studies can yield candidate virulence loci.

P-1304

Tracing the evolutionary history of the *Escherichia coli* O104:H4 outbreak strains

Lionel Guy[1], Cecilia Jernberg[2], Jenny Arvén Norling[1], Britta Björkholm[2], Sofie Ivarsson[2], Ingela Hendenström[2], Lars Engstrand[2,3], Siv Andersson[1]
[1]Department of Molecular Evolution, Uppsala University, Uppsala, Sweden, [2]Department of Preparedness, Swedish Institute for Communicable Disease Control, Solna, Sweden, [3]Department of Microbiology, Tumour- and Cell Biology, Karolinska Institute, Stockholm, Sweden

The recent emergence of a highly virulent *Escherichia coli* O104:H4 strain in Germany causing the largest outbreak of bloody diarrhoea and haemolytic uraemic syndrome (HUS) observed so far in Europe needs to be placed in an evolutionary context. Here, we present the genome of an O104:H4 strain isolated a year prior to the outbreak (E112/10) along with the genomes of seven *E. coli* strains isolated in Sweden during the outbreak, including a strain from the only non-German casuality. A phylogenomic analysis based on nucleotide substitutions in single copy genes showed that E112/90 was 10 to 100-fold less divergent from the outbreak strains than any previously sequenced genome in the O104:H4 group. A majority of substitutions acquired in either of the two strains were at non-synonymous sites in genes coding for surface proteins, transporters and carbohydrate metabolism, indicative of adaptive changes. We also sequenced the genomes of two *E. coli* strains with an enteroaggregative phenotype isolated in 1990 an 1993. Despite a clustering with commensal K12 strains, these strains contained plasmid-borne aggregation genes and chromosomal genes for adhesins and antibiotic resistances that were nearly identical in sequence with those of the outbreak strains. The emergence of the same *E. coli* pathotype from different genomic backgrounds provides insight into the flux of genes for host-adaptive processes. The analysis highlights the benefits of using whole-genome sequencing as an epidemiologic tool to capture the whole array of virulence features, and underlines the limitations of standard typing methods with less explanatory power.

P-1305

The effects of non-protein-coding RNA structure and function on intron evolution

Katarzyna Hooks, Daniela Delneri, Sam Griffiths-Jones
University of Manchester, Faculty of Life Sciences, Manchester, UK

It is well established that introns were present in the last common ancestor of Fungi. However, intronic sequences are continually being lost in yeast lineages, such that only 5% of *Saccharomyces cerevisiae* genes contain introns. We hypothesize that selectively retained introns contain RNA structures that act at the pre-mRNA level or that are processed to produce functional non-protein-coding RNA transcripts. So far, eight ncRNAs and two cis-acting RNA structures have reported to be encoded by introns in *S. cerevisiae*. We have used RNA gene prediction software to identify nineteen highly significant structures within introns for experimental validation. Preliminary experiments using reverse-transcription PCR strongly validate the computational screen; we found that at least eleven out of fourteen of the novel intronic sequences predicted have been shown to be either processed from the introns in *S. cerevisiae*, or splicing of their host introns appears to be differentially regulated. Using state-of-the-art RNA homology tools, including the INFERNAL package, we find that the majority of the predicted structures are well conserved in *S. cerevisiae* clade, but conservation extends to the *Candida* clade in at least two examples. Even though predictions were made in only 18% of all introns, 60% of long introns of non-ribosomal protein genes contained at least one high-scoring structure prediction. These data lead us to speculate that there might be subclasses of introns performing regulatory functions through RNA secondary structure. Further experiments, including deletion of introns containing RNA structures, are under way to describe the function of novel RNAs predicted in this study.

P-1306

Genome-wide analyses of recombination suggest that *Giardia intestinalis* assemblages represent different species

Feifei Xu, Jon Jerlström-Hultqvist, Jan O. Andersson
Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Giardia intestinalis is a major cause of waterborne enteric disease in humans. The species is divided into eight assemblages suggested to represent separate *Giardia* species based on host specificities and the genetic divergence of marker genes. We have investigated whether genome-wide recombination occurs between assemblages using the three available *G. intestinalis* genomes. First the relative non-synonymous substitution rates of the homologs were compared for 4,009 positional homologs. The vast majority of these comparisons indicate genetic isolation without inter-assemblage recombinations. Only a region of 6 kbp suggests genetic exchange between assemblages A and E, followed by gene conversion events. Second, recombination detecting software fails to identify within-gene recombination between the different assemblages for most of the homologs. Our results indicate very low frequency of recombination between the syntenic core genes, suggesting that *G. intestinalis* assemblages are genetically isolated lineages and thus should be viewed as separated *Giardia* species.

P-1307

Genome Sequence of the Cyanobacterium *Scytonema hofmanni* PCC 7110: The Largest Prokaryotic Proteome Sequenced to Date.

Robin Koch[1], Karina Stucken[1], Rosmarie Rippka[2], Nicole Tandeau de Marsac[2,3], Muriel Gugger[4], Tal Dagan[1], William F Martin[1]
[1]Institute of Molecular Evolution, Heinrich-Heine University, Duesseldorf, Germany, [2]Unité des Cyanobactéries, Institut Pasteur, CNRS URA 2172, Paris, France, [3]Aix-Marseille Univ, LCB, CNRS UPR 9043, Marseille, France, [4]Institut Pasteur, Laboratoire Collection des Cyanobactéries, Paris, France

Cyanobacteria are ancient and highly successful oxygenic photosynthetic prokaryotes, residing in diverse terrestrial and aquatic habitats. Cyanobacteria are divided into five subsections based on their morphology, cell arrangements, and developmental cycles. They comprise unicellular, single or colony forming members, and filamentous types, with or without cellular differentiation. Subsections IV and V contain the filamentous forms able to differentiate cells into heterocysts (nitrogen fixation). Many also form akinetes (resting cells) and hormogonia (vegetative reproduction). Members of Subsection V differ from those of Subsection IV by forming true branches on the primary trichomes. The cyanobacterial extreme versatility and physiological adaptation to different environments are reflected in their large genotypic diversity and genome sizes (marine picocyanobacteria: 1.5 - 3 Mb; multicellular cyanobacteria: 3.1 - 8.4 Mb). The largest known genome (8.4 Mb) was that of *Nostoc punctiforme* PCC 73102 (Subsection IV), originating from a symbiotic association with *Cycas sp.*, and capable of differentiating all three types of specialized cells known to occur in heterocystous cyanobacteria.

Here, we present the genome sequence of *Scytonema hofmanni* PCC 7110 (Subsection IV), a free-living and heavily ensheathed species, in which akinetes are not formed. With a genome size of 11,924,780 base pairs and 12,356 predicted protein coding genes, this is the second largest bacterial genome sequenced to date and the largest proteome described so far. A search for homologs to *S. hofmanni* ORFs in fully sequenced cyanobacterial genomes revealed that 7,344 (59%) have a cyanobacterial one. A phylogenetic reconstruction of the latter proteins revealed that 3,056 (42%) have their nearest neighbor among members of Subsection V, and 2,329 (32%) have their nearest neighbor among cyanobacteria of Subsection IV. A total of 5,038 (41%) ORFs have homologs in other completely sequenced genomes outside the cyanobacterial phylum, or in viral genomes and environmental DNA databases. The remaining 4,478 (36%) ORFs are singletons with no identified homologs. Singleton proteins were found to be shorter, and to display higher codon usage biases by comparison to proteins with homologs in other genomes.

Our phylogenetic analyses suggest that the "gargantuan" proteome of *S. hofmanni* represents a hybrid evolved from cyanobacteria of Sections IV and V. The reason behind the high frequency of singletons within that genome, and the putative functions of the singletons remains unclear. A comprehensive proteomic analysis of *Scytonema hofmanni* is currently under way, and is expected to shed light on the biochemistry and evolution of this extraordinary cyanobacterium.

P-1308

Estimate of the spontaneous mutation rate by whole genome sequencing of experimental lines in the green alga *Chlamydomonas reinhardtii*

Rob Ness, Andrew Morgan, Nick Colegrave, Peter Keightley
University of Edinburgh, Edinburgh, UK

Spontaneous mutation provides the raw genetic variation on which natural selection is dependent to function effectively. Despite the central importance of mutation to genetic, medical and evolutionary research there are relatively few detailed studies of the nature and impact on fitness of mutation across the genome. The rarity and stochastic nature of mutation make it a particularly difficult phenomenon to study. As a result direct estimates of the mutation rate are available from only a handful of model organisms. However, advances in DNA sequencing technology and the genomic sequences of an increasing number of organisms make it feasible to gather whole genome information from natural collections and experimental populations. Here, we present an estimate of the spontaneous mutation rate in the green alga *Chlamydomonas reinhardtii*. The mutation rate was estimated via whole genome sequencing of two experimental lines. These lines were maintained for ~250 generations in a mutation accumulation experiment, where the effectiveness of selection on mutations is much reduced. Two lines were then sequenced to ~10x coverage using Illumina HiSeq technology and assembled using the publicly available *C. reinhardtii* genome. From these data we were able to identify specific nucleotide and insertion-deletion mutations throughout the genome. Our findings are discussed in the context of genome evolution and the consequences mutation has for evolution in *C. reinhardtii*. Our data provide an accurate measurement of the spontaneous mutation rate for this strain, which can be applied to better understand the distribution of mutation effects on fitness, the evolution of sexual reproduction, the ability of populations to respond to natural selection and how the genome is shaped by the mutational process.

P-1309

'Genus wide sequence informed design of an evolutionary informative bacterial MLST scheme'

Miquette Hall[1], Sandra Reuter[2], Thomas Connor[2], Nick Thomson[2], Alan McNally[1]
[1]Nottingham Trent University, Nottingham, UK, [2]Wellcome Trust Sanger Institute, Cambridge, UK

Multilocus sequence typing (MLST) is a rapid, reliable technique used for determining the genetic identity of an isolate. MLST is based on the analysis of ~500bp regions from 7 unlinked housekeeping genes and can also be used to determine evolutionary relationships of genera. Previously, it has been impossible to design a universal MLST scheme for the *Yersinia* genus, due to high levels of genetic diversity. This project aims to design a method that will enable MLST schemes to be developed for genetically diverse genera. Next generation sequencing technology was used to sequence 200 *Yersinia* strains representative of the whole *Yersinia* genus; from these sequences the *Yersinia* phylogeny was established, using 85 common housekeeping genes. The MLST scheme was designed from these 85 housekeeping genes. The 7 genes chosen were unlinked genes representative of the whole genome, suitable for both primer design and also maintaining the established Yersinia phylogeny. This theoretical work has been used to optimise an experimentally viable MLST scheme, and is able to depict the evolutionary relationships between *Yersinia* species.

P-1310

Frequent and biochemically conservative amino-acid changes have the most moderate impact on codon bias.

Richard Cousins[1,2], Dawn Field[2], Laurence Hurst[1], Edward Feil[1]
[1]University of Bath, Bath, Somerset, UK,[2]Centre for Ecology and Hydrology, Oxford, Oxfordshire, UK

Codon bias refers to the unequal use of synonymous codons and is thought to result from a combination of mutational biases and selective pressures relating to the speed and accuracy of translation. Although mutational effects should clearly impact equally across all codons, codon preferences are commonly viewed as independent, such that the increased frequency of a synonymous codon corresponding to one amino-acid is not thought to impact on the frequencies of codons corresponding to other amino-acids. Here we challenge this view by exploiting 653 bacterial genome sequences to show that codon preferences are more similar between pairs of codons corresponding to more biochemically conservative, and more frequently observed amino-acid changes. Whilst this effect may have the advantage of minimizing the loss of codon bias resulting from the most frequent non-synonymous changes, we argue that selection is not required to explain our results. Instead we demonstrate through a simple simulation that a high rate of mutation between conservative amino-acids will itself lead to similar biases in the underlying codons. The simulation suggests that the effect is dramatically amplified by a positive feedback between codon preference and tRNA abundance, but the relevance of this remains unclear as a similar association is observed for both highly and lowly biased genes.

P-1311

Evolution of prophages in *Escherichia* and *Salmonella*

Louis-Marie Bobay[1], Marie Touchon[2,3], Eduardo Rocha[2,3]
[1]*Univ. Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, France,* [2]*CNRS, Paris, France,* [3]*Institut Pasteur, Paris, France*

Bacterial gene repertoires are extremely fluid largely by the action of mobile genetic elements. Indeed, most genomes present prophages, and this has been shown to be relevant for adaptation, virulence and defense against other phages. We are studying how prophages integrate genomes and how they subsequently evolve while integrated. We identified 500 prophages in 69 *Salmonella* and *Escherichia* genomes and analyzed their composition, signs of pseudogenization and evolution when conserved among closely related genomes. This investigation showed that integration of prophages respects genome organization, e.g. gene strand bias. The putative orthologous relationships between prophages provide a first glimpse on prophage evolution. The features of several prophages suggest that a large part of them are non-functional remnants and that their decay is discontinuous with initial long deletions and later slow pseudogenization. This might reveal different selective pressures for deletion along prophages and even selection for the maintenance of small non-functional phages for the protection of the bacteria from large functional phages. Moreover, we propose a phylogenetic classification of the 500 detected prophages and the 316 phages of *Enterobacteriaceae* available in the databanks. Taken together these results improve the understanding of the dynamics of prophages evolution in *Escherichia* and *Salmonella* genomes.

P-1312

Stop codons in bacteria are not selectively equivalent.

Inna Povolotskaya[1], Fyodor Kondrashov[1,2], Peter Vlasov[1]
[1]*Bioinformatics and Genomics Programme, Centre for Genomic Regulation, Barcelona, Spain,* [2]*Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain*

Many global patterns in molecular evolution are defined by the genetical code, including rates of nonsynonymous and synonymous evolution, synonymous codon usage and the optimality of the genetic code. The evolution and usage of stop codons, however, have not been rigorously studied with the exception of coding of non-canonical amino acids. Here, we study the rate of evolution and genomic frequency of TAA, TGA and TAG canonical stop codons in bacterial genomes. We find that stop codons evolve slower than synonymous sites, suggesting the action of weak negative selection. However, the frequency of stop codon usage relative to genomic nucleotide content indicates that this selection regime is not straightforward. The usage of TAA and TGA stop codons is GC-content dependent, with TAA decreasing and TGA increasing with GC-content, while TAG frequency is independent of nucleotide content. We thus modeled stop codon usage and nucleotide content with mutation rates and two selection on nucleotide content and TAG frequency as parameters. We found that the relationship between stop codon frequencies and nucleotide content cannot be explained by mutational biases or selection on nucleotide content. However, with weak nucleotide content-dependent selection on TAG, $-0.5 < N_e s < 1.5$, the model fits all of the data and recapitulates the lack of a relationship of TAG and nucleotide content. For biologically plausible rates of mutations we show that, in bacteria, TAG stop codon is universally associated with lower fitness, with TAA being the optimal stop codon for G-content < 16% while for G-content > 16% TGA has a higher fitness than TAG.

P-1313

Genome histories reveal adaptive cladogenesis in A. tumefaciens

Florent Lassalle[1,2], Xavier Nesme[2], Vincent Daubin[1]
[1]Laboratoire de Biométrie et Biologie Evolutive, Lyon, France, [2]Laboratoire d'Ecologie Microbienne, Lyon, France

The role of adaptation vs. neutral processes in bacterial cladogenesis is debated. We used the *Agrobacterium tumefaciens* species complex, a diverse group of plant-associated bacteria as a model to search for genomic signatures of adaptation in relation to diversification. The evolutionary histories of genomes of the entire taxon was reconstructed using an original dataset of 31 Rhizobiaceae, including 17 *A. tumefaciens* representing the diversity of the group. We designed a new phylogenetic approach for ancestral genome reconstruction, accounting for events of gene gain and loss as well as horizontal gene transfer. This approach identifies groups of co-transferred genes, providing better confidence and accuracy for the identification of donors and acceptors of transfers. A comparison of the observed patterns of genomic innovations with the expectations of a neutral model revealed putative adaptive events. We further used manually curated functional annotations to understand these genomic marks of adaptation and their role in the diversification of key clades within *A. tumefaciens.* Several blocks of co-transferred genes encode complete metabolic pathways, such as chemotaxis regulation, production of extracellular secondary metabolites (lipo-polysaccharide, siderophore) or catabolism of plant-derived compounds (phenolics, amino-acids, complex sugars). The molecular function and ecological role of several candidate genes has been experimentally characterized, revealing previously unknown metabolic activities. These results suggest that diversification in the *A. tumefaciens* species complex is generally marked by ecological adaptations related to interaction with a host plant and competition between rhizospheric bacteria.

P-1314

High-throughput sequencing to identify lineage-specific transcriptomes of *M. tuberculosis* and the contribution of background genetic diversity

Graham Rose[1], Iñaki Comas[1], Douglas Young[1], Sebastien Gagneux[1]
[1]MRC National Institute for Medical Research, London, UK,[2]Swiss Tropical and Public Health Institute, Basel, Switzerland

A global picture of *M. tuberculosis* strain variation has emerged, with clinical strains grouping into six major phylogenetic lineages that display a strong geographic distribution. Driven by advances in high-throughput sequencing, numerous whole genome sequences from each lineage now exist, enabling the determination of each lineage-specific genetic background. Focusing on Lineages 1 and 2, we asked how the lineage-specific SNPs might be contributing to functional diversity at the level of transcription. RNA-sequencing was used to uncover the total transcriptomic landscape of several clinical strains from these two lineages, grown *in vitro* in 7H9 media. Using this transcriptomic data we found many conserved RNA expression profiles that exist within but not across lineages, or lineage-specific expression, ranging from conserved expression of mRNAs, to non-coding RNAs and antisense expression. From this conservation we generated a panel of genes that belong to each lineage-specific transcriptome. Interestingly, dividing these genes into functional categories revealed a significant overrepresentation of those belonging to the regulation class, including several global transcriptional regulators. We have also found direct associations between lineage-specific transcription and lineage-specific genetic background. We are exploring these findings using a mixture of *in silico* and experimental methods. Overall, these findings provide insight into the functional differences between the lineages and suggest the potential global impact of this diversity.

P-1315

Studying Recoding in Bacterial genes

Virag Sharma[1], David Murphy[1], Christophe Penno[1], John F Atkins[1,2], Pavel V Baranov[1]
[1]University College Cork, Cork, Ireland,[2]University of Utah, Salt Lake City, USA

A number of annotated bacterial protein coding genes consist of more than a single continuous open reading frame (ORF). ORF disruptions in annotated genes arise due to two major causes: i. Sequencing errors and recent mutations. ii. Utilization of non-standard decoding (Recoding). Both, transcriptional and translational mechanisms of Recoding are known. During Programmed Transcriptional Realignment (PTR), indels are introduced into RNA molecule due to RNA polymerase slippage that usually takes place on simple mononucleotide repeats. During Programmed Ribosomal Frameshifting (PRF), a proportion of ribosomes shift reading frame at a specific location in response to Recoding signals embedded in mRNA. The analysis of annotated bacterial genomes allowed us to identify ~6,000 genes utilizing recoding. These genes can be clustered into 64 groups based on their sequence similarity [1].

Experimental investigation of nucleotide context surrounding PTR patterns revealed that short RNA secondary structures upstream of transcription slippage patterns can modulate the efficiency of PTR. Therefore, we performed a systematic bioinformatics analysis to identify the presence of evolutionary conserved RNA secondary structures in the proximity of potential slippery patterns. We screened bacterial genomes for the presence of genes with polyA runs (where the minimum length of polyA run is 8 residues). PolyA containing genes were subsequently clustered based on sequence similarity and screened for the presence of RNA secondary structure upstream polyA runs. We found the presence of evolutionary conserved RNA secondary structures in 20 gene families/clusters.

To enable the visualization of recoding patterns that are prone to PRF, which usually are combinations of specific codons, we have developed a software tool called CodonLogo. CodonLogo measures Shannon Entropy in a multiple codon alignment, in a manner similar to Sequence Logos [2]. CodonLogo is freely available at http://lapti.ucc.ie/CodonLogo/

References:

1. Sharma V, Firth AE, Antonov I, Fayet O, Atkins JF, Borodovsky M, Baranov PV. A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. Mol Biol Evol. 2011 Nov;28(11):3195-211.

2. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990 Oct 25;18(20):6097-100.

P-1316

Genome evolution of cells with two nuclei: double peaks reveal rare sex in the protist group diplomonads

Jan O. Andersson
Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

Diplomonads are unusual in having two nuclei. Each nucleus of these anaerobic flagellated single-celled eukaryotes appears to contain at least two copies of the genome and is transcriptionally active. It has long been assumed that diplomonads are asexual. However, population genetic studies have indicated recombination between natural isolates, and meiosis-specific genes have been identified in the human intestinal parasite *Giardia intestinalis*. Epidemiological studies suggest exchange of alleles between lineages, and nuclear fusions have been observed within *G. intestinalis* cysts. Together, these data have led to the suggestion that *Giardia* undergo recombination of some sort. It is currently unknown how the genome copies in the two nuclei are maintained and affected by recombination.

Different members of the group indeed show large variation in the degree of genome heterogeneity between the nuclei. A very low amount of allelic variation (<0.01%) was reported from the WB isolate of *G. intestinalis*. In contrast, the GS isolate showed an average allelic variation of 0.5% with distinct patterns along the chromosomes; large regions that lack variation are interspersed with regions with up to 5% differences. ‘Double peaks' - mixed signals in distinct sequence positions in sequences from PCR products from stool samples - are commonly reported in epidemiological studies of G. intestinalis suggesting that allelic variation is frequent within the species. The morphologically indistinguishable Atlantic Salmon parasite *Spironucleus salmonicida* and the fish commensal *S. barkhanus* show similar differences: *S. salmonicida* lack allelic variation whereas *S. barkhanus* show large frequencies of variations in some, but not all, genes. Curiously, more than four alleles were detected for several single-copy genes suggesting a ploidy of more than four in *S. barkhanus*. Indeed, flow cytometry analyses indicted a 50% higher DNA content for *S. barkhanus* than *S. salmonicida*.

Here I propose, in contrast to common assumptions, that the observed allelic differences in diverse diplomonads indicate recent sexual events. The cell and nuclei fuse, possibly at the cyst stage in the lifecycle, followed by random chromosome losses. During this reduction process the chromosomes may exchange genetic material. In the absence of additional cell fusions the descendant will reach tetraploidy, and the allelic sequence heterogeneity will be purged from descendants growing asexually via within-cell recombinatory events, such as gene conversions. Directed studies of the allelic sequence heterogeneity in diverse diplomonad lineages are likely to reveal details about the enigmatic diplomonad sexual life cycle.

P-1317

Uncovering the Molecular Evolution of Plasmid 'Species'

Edel Hyland, Andrew Murray
Harvard Univeristy, Cambridge, MA, USA

We are interested in the molecular details of evolution; how specific DNA base pair changes give rise to selectable phenotypes, and ultimately new species. We study the evolution of plasmid segregation in bacteria as a minimal network of proteins that produce a clear and selectable biological function. We focus on the Type II plasmid partitioning system, which is composed of two trans-acting elements, ParM and ParR and a cis-acting sequence, ParC, encoded by a single operon. These elements comprise a primitive mitotic spindle in prokaryotes. It is known that two plasmids encoding identical partitioning operons in the same cell are in competition with each other, and are said to be incompatible. Sequence diversity within the partitioning operon that eliminates such competition gives rise to compatible plasmids, or new plasmid 'species'. Our goal is to understand the molecular basis of this partition-based plasmid incompatibility, and to ask how many interfaces in the interactions between the ParM, ParR and ParC components are evolutionarily malleable to facilitate the differentiation into new plasmid species.

Taking a synthetic biology approach, we engineered a library of twelve synthetic, minimized plasmids each expressing a unique Type II partitioning sequence obtained from endogenous plasmids of various bacterial species. Using this library we determined the extent of incompatibility, or competition, between pairs of these synthetic plasmids, allowing us to identify sequences within this operon that govern partition-based plasmid incompatibility. Guided by these results we demonstrate that for certain incompatible plasmid pairs, competition is defined by sequence similarities at the interface between ParR and ParM. Currently, we hope to evolve novel plasmid species by targeted mutagenesis of this interface, revealing the least amount of sequences changes required for speciation in this very simple model.

P-1318

Genomic diversity of Staphylococcus aureus in singly-colonized nasal carriers and the implications for detecting recent transmission

Elizabeth Batty[1], Tanya Golubchik[1], Ruth Miller[2], Helen Farr[2], Hanna Larner[3], Rowena Fung[2], Heather Godwin[2], Daniel Wilson[2,3], Derrick Crook[2], Rory Bowden[1,2]
[1]Department of Statistics, University of Oxford, Oxford, UK,[2]NIHR Oxford Biomedical Research Centre, Oxford, UK, [3]Wellcome Trust Centre for Human Genetics, Oxford, UK

Staphylococcus aureus is a commensal of humans and several other animal species which can also cause a range of mild to severe diseases. In obligate parasites such as S. aureus, genetic variation arising within individual hosts provides the only raw material for evolution. Mutations accumulated during colonization and disease may play a role in disease progression and can facilitate reconstruction of transmission chains. Here we used bacterial whole-genome sequencing to investigate intra-host genetic variation in 13 asymptomatic individuals, each colonized by a single clone of S. aureus. We discovered 162 single nucleotide polymorphisms and 22 short insertions/deletions among ten hosts, revealing microvariation as a common feature of nasal carriage. Structural variation in the S. aureus genome associated with insertion or excision of mobile elements was also apparent. The fitness landscape during colonization was dominated by purifying selection and reduced diversity was associated with MRSA carriage and the recent use of antibiotics. Total diversity and variation in allele frequencies across individuals were consistent with longitudinal fluctuation in intra-host population size. By identifying the scale of within-host variation, even in clonally colonized carriers, this work demonstrates the ability of next-generation sequencing to identify microvariation within a population and provides a molecular genomic basis for future studies of transmission in S. aureus and other organisms.

P-1319

Genome-wide insights into the evolution of regulatory elements in diverse yeast strains

Caitlin Connelly, Joshua Akey
University of Washington, Seattle, WA, USA

Noncoding genetic variation makes a significant contribution to phenotypic diversity and disease susceptibility by modulating gene expression, and examples of noncoding variants causing phenotypic differences within and between species are rapidly accumulating in diverse lineages. However, the precise molecular mechanisms that noncoding variants act through, the evolutionary forces that govern the trajectory of noncoding variation, and the relative effects of different noncoding variants on gene expression are not well characterized. We addressed these questions by comprehensively characterizing regulatory regions in diverse strains of *S. cerevisiae* using FAIRE-Seq (Formaldehyde-Assisted Isolation of Regulatory Elements), investigating the evolutionary forces acting on noncoding variants within active regulatory regions, and characterizing the effects of changes to regulatory elements on gene expression differences between strains. Population genetics analyses have previously revealed that on a genome-wide scale, regulatory motifs are under strong purifying selection, and there is considerable heterogeneity in the magnitude of selection across different motifs. Here, we expand on these previous studies by characterizing the evolutionary forces acting at active regulatory elements. Finally, we use RNA-Seq to measure genome-wide gene expression levels, and quantify the extent to which differences in regulatory elements between strains influence gene transcriptional variation. Our results provide new insights into the evolution of functional noncoding DNA, the contribution of transcriptional variation to phenotypic diversity, and the characteristics of regulatory alleles in global yeast strains.

P-1320

The Chimera Hypothesis for the Thermotogae Phylum

Kristen Swithers, J. Peter Gogarten
University of Connecticut, Storrs, Ct, USA

The Thermotogae phylum is a deeply branching bacterial lineage found in anaerobic marine and fresh water geothermal environments. Organisms in this phylum thrive in an extreme temperature range; mesophilic to hyperthermophilic, and are characterized by a unique toga-like outer envelope. The ancestral genome state for the phylum has been suggested to have a higher optimal growth temperature than the extant lineages. Horizontal gene transfer (HGT) played a major role in shaping the genomes of the extant lineages of the phylum: 48% of the genes in a given genome were transferred from the Firmicutes, and 11% of the genes were transferred from the Archaea. This extensive amount of HGT is only seen in one other phylum in the prokaryote domain; the Aquificae. The large contributions from the Archaea and Clostridia leads to the hypothesis that the extant genomes of this group resulted from an ancient chimerization event between a Firmicute and an Archaeon giving rise to the ancestor of the Thermotogae. If this chimera hypothesis is correct then all of the gene transfers from Firmicute and Archaea should have been gained in one event in the ancestor of all the Thermotogales. In a preliminary parsimony analysis of gene absence/presence data Firmicute and archaeal genes appear to have been acquired in multiple events on a reference tree proving evidence against the chimera hypothesis. The lineage leading to the more thermophilic *Thermotoga* genus appears to have acquired more archaeal genes, and the other lineages in the phylum, which are characterized by a lower temperature optimum, acquired more clostridial genes. These Clostridia gene transfers may have allowed for the mesophilic life style adaptation of the Thermotogales lineages with a lower temperature optimum. Under the assumption of the chimera hypothesis, all Clostrdia and Archaea genes are present in the ancestor of the Thermotogae and extant gene absence is strictly due to loss. Here we provide a rigorous statistical test for the chimera hypothesis using Bayesian and maximum likelihood ancestral state reconstructions, and we provide evidence of how these archaeal and bacterial HGTs have affected the various genomes of the phylum over time.

This research was supported through NSF DEB 0830024 and NASA Exobiology Program NNX08AQ10G.

P-1321

p { margin-bottom: 0.21cm; } Rapid desktop sequencing transforms *Staphylococcus aureus* and *Clostridium difficile* outbreak detection and surveillance

David Eyre[1,3], Tanya Golubchik[2], Claire Gordon[1,3], Rory Bowden[2,3], Paolo Piazza[4], Elizabeth Batty[2], Camilla Ip[2], Daniel Wilson[1,3], Xavier Didelot[1,3], Lily O'Connor[1,5], Rochelle Lay[5], David Buck[4], Angela Kearns[6], Angela Shaw[7], John Paul[8], Mark Wilcox[9], Peter Donnelly[4], Tim Peto[1,3], Sarah Walker[1,10], Derrick Crook[1,3]
[1]1. Nuffield Department of Clinical Medicine, Experimental Medicine Division, Oxford, UK,[2]2. Department of Statistics, University of Oxford, Oxford, UK,[3]3. NIHR Oxford Biomedical Research Centre, Oxford, UK,[4]4. Wellcome Trust Centre for Human Genetics, Oxford, UK,[5]5. Oxford University Hospitals NHS Trust, Oxford, UK,[6]6. Health Protection Agency Centre for Infections, Staphylococcus Reference Unit, London, UK,[7]7. Ashford and St Peter's NHS Foundation Trust, Department of Microbiology, Surrey, UK,[8]8. Health Protection Agency, Brighton, UK,[9]9. Leeds Teaching Hospitals & Univ of Leeds, Microbiology, Leeds, UK,[10]10. Medical Research Council, Clinical Trials Unit, London, UK

p { margin-bottom: 0.21cm; }

Whole-genome sequence data can be combined with epidemiological information to provide precise and rapid detection of healthcare-associated outbreaks. To investigate the potential of this approach, we used the Illumina MiSeq platform to sequence bacterial isolates from clusters of patients with methicillin-resistant *Staphylococcus aureus* (MRSA) and *Clostridium difficile*, two clinically important pathogens. In addition, we simulated a real-time surveillance situation by sequencing all *C. difficile* cases over six weeks in a single hospital and comparing these to existing sequences from the same hospital group in the preceding three years. Two clusters each for MRSA (26 isolates) and *Clostridium difficile* (15 isolates) were sequenced and analysed within 5 days of initial culture. Both clusters of MRSA were confirmed as outbreaks, with most sequences in each cluster identical and all within 3 single nucleotide variants (SNVs). Epidemiologically unrelated isolates that were identical by *spa* typing were shown to be genetically distinct (>=21SNVs). Similarily, in both *C. difficile* clusters, closely epidemiologically linked cases (in one case sharing the same strain type) were also genetically distinct (>=144SNVs). The surveillance simulation demonstrated that early outbreak detection is possible for *C. difficile,* and identified a previously undetected probable community transmission. Our findings are widely generalisable to other human pathogens.

P-1322

Analyzing recombination and phylogeography in bacteria at once: the case of a recently emerged clone of *Staphylococcus aureus*.

Santiago Castillo[1], Pekka Marttinen[4], Jukka Corander[2], Matt Holden[3], Mona Aldeljawi[1], Edward Feil[1]
[1]*University of Bath, Bath, UK,* [2]*University of Helsinki, Helsinki, Finland,* [3]*The Wellcome Trust Sanger Institute, Cambridge, UK,* [4]*Aalto University, Helsinki, Finland*

The amount of molecular data generated through genome sequencing has significantly impacted on our understanding of the microevolution of many bacterial species. Firstly the molecular processes introducing genetic variation (mutation and recombination) and the evolutionary fate of such variation over time (drift or selection) have been studied at an unprecedented level due to these powerful data sets. Just as well, these data have been very useful for illuminating the transmission and spread of some pathogenic bacteria. For instance, a recent study employing a next-generation platform described the hospital transmission and the intercontinental spread of the recently emerged methicillin-resistant *S. aureus* ST239 clone. Nonetheless, the marrying of these two perspectives is a major task that remains largely unaddressed. Here we address this by supplementing the data of that recent study with a further 127 isolates, most of which were collected from four hospitals in three major Turkish cities (Istanbul, Ankara and Izmir); this was done to analyze the degree of structuring on a national level, a level not addressed in the initial study. First, we investigate the impact of recombination in this recently emerged clone and find that more than 30% of the genes have been affected. These genes are significantly more diverse, however most of them belong to the accessory genome. Secondly, using the SNPs not affected by recombination, maximum likelihood and Bayesian MCMC phylogenetic analyses were carried out to study the transmission and spread of this clone. The Turkish and Eastern European isolates correspond to clear monophyletic groups within the global dataset and, within the Turkish sample, isolates from Izmir are clearly distinct from those from Ankara and Istanbul, which is consistent with the relative isolation of this city. In general, we find that most of the isolates clustered according to geographic source (the main clades being Europe, East Asia, South America and Turkey) suggesting a regional rather than global dispersion. Analyzing the age and the rate of recombination in the core genome for each one of clades, we note that the rate of recombination, in comparison with mutation, increases with time, although the level of recombination is lower than expected in the Turkish clade. Finally, a demographic analysis suggests that the effective population size of the ST239 clone has been decreasing, which agrees with the idea of increased drift acting on this clone.

P-1323

Tracking the emergence of new virulent lineages of Staphylococcus aureus using whole-genome sequencing.

Ruth Miller[1], James Price[2], Antonina Votintseva[1], Xavier Didelot[1], Daniel Wilson[1], David Wyllie[3,1], Liz Batty[1], Tanya Golubchik[1], John Paul[3], Martin Llewelyn[2], Derrick Crook[1], Rory Bowden[1]
[1]University of Oxford, Oxford, UK,[2]Royal Sussex County Hopital, Brighton, UK,[3]Health Protection Agency, London, UK

Staphylococcus aureus is a commensal of humans and other animals that can also cause serious disease and for which treatment can be difficult, particularly for drug-resistant forms such as methicillin resistant S. aureus (MRSA). The epidemiology of healthcare-associated S. aureus is characterized by waves of infection caused by the emergence and spread of clones at national and even global scales.
Large-scale whole-genome data, generated by next-generation techniques, has great potential to help in understanding the epidemiology and spread of pathogens, by identifying changes in gene content and measuring genomic relationships at high resolution. We used Illumina sequencing of 162 MRSA isolates to investigate the local emergence of a new variant of ST36 / EMRSA-16, one of two dominant UK MRSA lineages, that was initially identified because of an exceptional ‘spike' in local MRSA bacteraemia rates.
The new clone was indistinguishable from other circulating ST36 isolates by conventional sequence-based typing and quickly became the locally dominant MRSA. Genomic comparisons with other CC30 isolates confirm that this was a recently emerged lineage (lineage MRCA c.2004; MRCA with other lineages c.2001). The inferred genealogy of the new lineage suggests that genetic innovations affecting virulence or a small number of prolific sources of infection may be more likely explanations for its success than failures in infection control practices. A short list of genetic changes identified in the new virulent lineage provides candidate risk factors for transmissibility and virulence, however assessing their statistical and functional significance is not straightforward.
This work demonstrates that it will be possible to detect new virulent lineages of pathogens such as MRSA with potentially serious outcomes at an early stage in their emergence. Identifying plausible genetic determinants of emergence will likely require combining data from multiple such examples.

P-1324

Detection of *Mycobacterium tuberculosis* outbreaks using whole-genome sequencing

Timothy M. Walker[1], Camilla L.C. Ip[2], Jason T. Evans[3], Philip Monk[4], Peter M. Hawkey[3], Julian Parkhill[5], Ruth Harrell[3], A. Sarah Walker[1], Derrick W. Crook[1], Tim E. A. Peto[1], Rory Bowden[2], E. Grace Smith[3]
[1]Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Headley Way, Headington, Oxford, OX3 9DU, UK, [2]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK, [3]Health Protection Agency, West Midlands Public Health Laboratory, Heart of England NHS Foundation Trust, Bordersley Green East, Birmingham B9 5SS, UK, [4]Health Protection Agency, East Midlands South, Atelius House, 2 Smith Way, Grove Park, Enderby, Leicester, LE19 1SX, UK, [5]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1HH, UK

Outbreak investigation of *Mycobacterium tuberculosis* (MTB) in the UK is currently based on contact tracing among patients who share a 24-locus Mycobacterial Interspersed Repetitive Unit - Variable Number Tandem Repeat (MIRU-VNTR) genotype. Whole-genome sequencing (WGS) of pathogens is now sufficiently fast and cheap that it is a practical alternative to the variety of routine tests used in infection control. The data can be used to study evolution of MTB in the community and improve patient care and infection control.

We sequenced 109 MTB isolates from 85 patients from the Midlands (UK) on the Illumina platform. Variants were inferred using an in-house bioinformatics pipeline based on read mapping to the H37Rv reference genome and used these for phylogenetic analyses and molecular clock estimates. Epidemiological and clinical information was available for each isolate.

Over 85% of the sample sites shared with the H37Rv reference strain were called, corresponding to >92% of the non-repeated regions of the reference. We estimated a molecular clock rate of 0.6 nucleotide variants per year from serially sampled isolates from 7 patients having a mean sampling interval of 33 months. We applied this to 16 household clusters, the epidemiological time-frames of which were consistent with the clock. Our genomes were representative of global diversity. Among those with shared MIRU-VNTR profiles, WGS demonstrated sufficient resolution to determine which isolates were closely related and infer transmission events between them.

Our study demonstrates that the additional information in the WGS of MTB isolates is of sufficient quality to exclude false matches and identify samples not linked through traditional cluster investigation protocols.

P-1325

The mitochondrion-related organelles of a novel breviate protist: Enigmatic organelles in a microaerophilic lineage of eukaryotes.

Courtney Stairs[1], Matthew Brown[1], Mark van der Giezen[2], Andrew Roger[1]
[1]Dalhousie University, Halifax, NS, Canada, [2]University of Exeter, Exeter, UK

Mitochondria are powerhouses of eukaryotic cells responsible for the vast majority of their ATP production. In eukaryotes that have evolved in low oxygen environments, many metabolic pathways of mitochondria have been retailored by both loss of aerobic capacities and acquisition of novel enzymes to survive. These mitochondrion-related organelles (MROs such as mitosomes and hydrogenosomes) appear to have evolved independently multiple times in different lineages of eukaryotes. The metabolic diversity and punctate distribution of MROs across the tree of eukaryotes has made determining their origins difficult and extremely controversial. Here we report an investigation of the MROs of a novel breviate-like organism (PCB) closely related to Breviata anathema, a putative early-branching eukaryote of controversial affinities. Preliminary electron microscopy of PCB revealed an oblong double membrane bound organelle that lacks mitochondrial cristae and spans the length of the cell (5-10 µm). From 38000 clusters in 86 million reads of HiSeq (RNAseq) transcriptomic data, we made an in silico prediction of the metabolism of the MRO in PCB using MitoProt and TargetP to assess the subcellular localization of different proteins based on N-terminal targeting peptides. Using traditional BLAST and Hidden Markov Model techniques, we identified genes encoding typical mitochondrial processes including amino acid metabolism, protein import machinery, a partial electron transport chain and a partial Kreb's cycle but not ATP synthase. We also identified genes encoding the core anaerobic energy generation system found in the hydrogenosomes of Trichomonas vaginalis and the unique iron-sulfur cluster biosynthesis strategy recently reported in the cytosol of Blastocystis sp.. Preliminary phylogenetic analysis revealed that many of the aforementioned MRO genes (of non-mitochondrial origin) have been acquired from both prokaryotic and eukaryotic sources via lateral gene transfer. The PCB organelle shows a unique suite of both ancestral and recently-acquired metabolic properties as well as metabolic redundancy never before observed in MROs underscoring the importance of genomic studies of newly-discovered 'deep' protistan lineages in clarifying the early course of eukaryote evolution.

P-1326

Phylogenetic analysis and alignment of NGS sequence data

Tandy Warnow, Nam-phuong Nguyen, Siavash Mirarab
University of Texas at Austin, Austin, TX 78712, USA

NGS technologies produce short, fragmentary data, which presents enormous challenges for alignment and phylogenetic analysis. Here we present a new method for alignment and phylogeny estimation for use with NGS data, and in particular for the use with metagenomic data. Phylogenetic placement is the problem in which the objective is to insert short molecular sequences (called "query sequences") into an existing phylogenetic tree and alignment on full-length sequences for the same gene. We present SEPP, a general "boosting" technique to improve the accuracy and/or speed of phylogenetic placement techniques. The key algorithmic aspect of SEPP is a dataset decomposition technique in SATe (Liu et al., Science 2009 and Systematic Biology (in press), a method that utilizes an iterative divide-and-conquer technique to co-estimate alignments and trees on large molecular sequence datasets. We show that SEPP improves current phylogenetic placement methods, placing short sequences more accurately when the set of input sequences has a large evolutionary diameter and produces placements of comparable accuracy in a fraction of the time for easier cases. We then show that we can use SEPP to perform taxon identification of metagenomic data, using statistical support estimations from each of its steps, to produce dramatically improved accuracy over current taxon identification methods.

P-1327

Shaping microbial genomes: the source of each gene and the impact of habitat and evolutionary distance on lateral gene transfer

Conor Meehan, Robert Beiko
Dalhousie University, Halifax, NS, Canada

Microbial genomes evolve through vertical inheritance of gene content and also lateral gene transfer (LGT) from other species. Estimation of LGT contribution to genome content is thought to be up to 30% for prokaryotic genomes. However the origin and boundaries of such gene acquisitions is still not fully understood. Certain habitats, such as the human gut, are thought to permit greater levels of gene exchange and that such exchange is likely to be between close relatives.

Here we have attempted to study species from two environments, the human gut and cow rumen, to determine the influence of habitat and taxonomy on genome evolution. As many species have not been definitively assigned to a habitat, homology matching of microbiome samples to 3,050 sequenced genomes was utilised to assign such species to either being present in the human gut, the cow rumen, both or neither. This approach was successful in correctly identifying many known commensal human gut bacteria, and indicated a low overlap of species between environments.

Four species, two bacteria and two archaea, were subsequently studied to determine the likely source of every gene within their genome. Each species pair is closely related taxonomically (within the same genus) but reside in different habitats (human gut and cow rumen). Comparison of protein coding genes to all sequenced genomes revealed potential LGT events likely contributed 15-42% of the genes within each genome, with the rate of LGT being twice as high in bacterial species than archaeal. Although the majority of genes were inherited vertically, several cases of orthologous replacement from taxonomically distant partners were found through phylogenetic investigation. Gene sharing networks showed that frequent interacting partners were likely to be residing in the same habitat and many were from taxonomically dispersed clades.

The higher rate of LGT found here in the bacteria over the archaea, regardless of habitat, indicates that fundamental differences between these domains may influence the level of lateral gene acquisition more so than environment. However any LGT that does occur is likely habitat-driven with interacting partners being dispersed throughout the domains of life. These results demonstrate the complex patterns of vertical inheritance, lateral transfer and orthologous replacements that create the mosaic that is each microbial genome.

P-1328

Levels and patterns of genetic diversity of AI-2 quorum sensing in Escherichia coli

Patricia H. Brito[1], Eduardo P. C. Rocha[2], Isabel Gordo[1], Karina B. Xavier[1,3]
[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal, [2]Institut Pasteur, Paris, France, [3]Instituto de Tecnologia Quimíca e Biológica, Oeiras, Portugal

Although single cell organisms, bacteria can express coordinated multicellular behaviors, such as virulence, biofilm formation, and quorum sensing (QS). Autoinducer-2 (AI-2) represents a nonspecies-specific signal produced to mediate both intra- and interspecies communication in bacterial QS. We analyzed the genetic diversity of AI-2 QS genes that includes the luxS gene (the signal synthase) as well as all the genes of the *lsr* regulon that are responsible for the reception, internalization and processing of the signal across all available genomes of *Escherichia coli* and *Shigella*. Many strains do not hold a complete functional *lsr* operon, and the presence of genes does not correlate with phylogenetic history or pathogenicity. The functional operon is present at a frequency of 0.63, strongly suggesting that these genes are neither strain-specific nor volatile. We hypothesize that selection actively maintains a balanced polymorphism for the presence/absence of a functional lsr regulon. This pattern of loss-of-function in the lsr regulon contrasts with the presence of the signal synthase (luxS) in all strains. Analysis of selection at the nucleotide level in organisms that have the complete system shows that two genes are under balancing selection: luxS and lsrA. Interestingly these genes code for the production of the signal and for the ATPase enzyme that produces the energy necessary for the internalization of AI-2 into the cell. This pattern of polymorphism suggests evolution on fluctuating environments or alternatively, the consequence of frequency-dependent selection acting on the efficiency of signal production and internalization. The later would support the hypothesis of AI-2 QS as a cooperative behavior vulnerable to the evolution of cheating, as has been show for other bacterial species.

P-1329

Gene numbers are underestimated in bacterial genomes

Klaus Neuhaus[1], Katharina Mir[3], Svenja Simon[2], Richard Landstorfer[1], Steffen Schober[3], Martin Bossert[3], Daniel Keim[2], Siegfried Scherer[1]
[1]Chair for Microbial Ecology, Department for Biosciences, Technische Universität München, Freising, Germany, [2]Department for Informatics, Universität Konstanz, Konstanz, Germany, [3]Institute of Communications Engineering, Universität Ulm, Ulm, Germany

The existence of non-trivially overlapping genes is widely accepted in viruses and phages but not in bacteria. Current genome annotation programs discard overlapping genes due to constraints of such gene pairs. Since bacteriophages are in a long-term equilibrium with their host genomes we hypothesize the existence of additional (overlapping) genes in bacterial genomes.

Bacterial genomes are densely covered by genes, thus, codon usage is the most influential parameter. Based on more than 50 bacterial genomes of different GC-content (21.4% - 74.9%) we build randomized model genomes of the same codon usage as their natural counterparts. Global parameters (e.g., predicted number of ORFs, length distributions of the "annotated" ORFs, etc.) show close correspondence in each case. However, the length distributions of overlapping ORFs in alternative reading frames differed between natural genomes and the randomized counterparts. Significantly more of the longer ORFs in alternative reading frames exist in natural genomes compared to the random model. We found up to several hundred unexpectedly long ORFs in alternative reading frames per genome, depending on the organism. We hypothesize that selection "keeps" the longer overlapping ORFs intact. Otherwise, unused ORFs in the alternative reading frames would be shortened to the expected stochastic distribution lengths by random mutations.

A second possibility to detect overprinted genes is to find suspected protein-coding ORFs using a blastp-search. Using a cut-off E value of 10-10 or less, we received numerous hits, depending on species and GC-content. Organisms with 30% GC or less obtained up to about 20 hits, organisms with more than 30% GC had up to about 200 hits and more per Mbp of genome sequence.

Analysis of published transcriptomes from *Nostoc* PCC7120 (GC 41.3%, 6.4 Mbp) showed at least 150 transcripts of overlapping ORFs longer than 200 bp in length. Such transcripts are generally assumed to be non-coding regulatory RNA. However, based on the above we hypothesize that some of these transcripts are protein coding.

In conclusion, we believe that the coding capacity of bacterial genomes has been underestimated.

P-1330

pH adaptation of archaeal ammonia oxidizers

Cecile Gubry-Rangin[1], Christopher Quince[3], Robert Griffiths[4], Michael Schloter[2], James Prosser[1], Graeme Nicol[1]
[1]University of Aberdeen, Aberdeen, UK, [2]Helmholtz Zentrum München, Munich, Germany, [3]University of Glasgow, Glasgow, UK, [4]Centre for Ecology and Hydrology, Wallingford, UK

Prokaryotes, the Bacteria and the Archaea, comprise two of the three domains of life and are the most abundant and diverse organisms on earth. Bacteria have been widely used to study adaptation, but with emphasis on pathogenic bacteria or readily cultivated organisms, such as *Escherichia coli* and *Pseudomonas fluorescens*. Few studies have investigated the archaeal domain, even though truly universal models of adaptation must apply to all three biological domains. In this study we focused on the thaumarchaea, an abundant archaeal evolutionary lineage, which contribute significantly to nitrification, a critical step in terrestrial nitrogen cycling. Whereas the limited number of cultivated thaumarchaea has seriously limited investigation of these organisms, high-throughput sequencing of amplified functional genes makes it possible to understand their diversity, distribution and adaptation to environment.

Soil pH is a major determinant of microbial ecosystem processes and potentially a major driver of evolution, adaptation and diversity. To determine whether pH drives evolutionary adaptation and community structure of soil archaeal ammonia oxidizers, sequences of *amoA*, a key functional gene of ammonia oxidation, were examined in soils at global, regional and local scales. Globally distributed database sequences clustered into 18 well-supported phylogenetic lineages that dominated specific soil pH ranges classified as acidic (pH<5), acido-neutral (5≤pH<7) or alkalinophilic (pH≥7). To determine whether patterns were reproduced at regional and local scales, *amoA* gene fragments were amplified from DNA extracted from 47 UK soils (pH 3.5 - 8.7) including a pH-gradient formed by seven soils at a single site (pH 4.5 - 7.5). High-throughput sequencing and analysis of *amoA* gene fragments identified an additional, previously undiscovered phylogenetic lineage and revealed similar pH-associated distribution patterns at global, regional and local scales, especially for the five most abundant clusters. Furthermore, archaeal *amoA* abundance and diversity increased with soil pH, which was the only physicochemical characteristic measured that significantly influenced community structure. The results suggest evolution based on specific adaptations to soil pH and niche specialization resulting in a global distribution of archaeal lineages that have important consequences for soil ecosystem function and nitrogen cycling.

The analysis of the ecological coherence at different taxonomic levels is currently being analysed to determine whether different environmental factors influence thaumarchaeal adaptation at different taxonomic levels. This study provides insights into the relative importance of different factors influencing microbial adaptation in the environment for an ecologically important and ubiquitous archaeal kingdom.

P-1331

On the origin of leprosy: a genomics perspective

Luz-Andrea Pfister, Anne C. Stone
Arizona State University, Tempe, Arizona, USA

The origin of leprosy, one of the most ancient scourges of humanity, remains unknown. *Mycobacterium leprae*, the causative agent of leprosy, exhibits an ancient parasitic lifestyle. Based upon the timing of *M. leprae*'s massive genome decay, its obligate intracellular lifestyle originated millions of years ago, long before the origins of modern humans suggesting its presence in primates for a long time or a more recent jump to humans which is also supported by very limited genetic variation among strains. To address questions about the evolutionary history of leprosy, we sequenced the genome of a strain of *M. leprae* isolated from a West African mangabey (Cercocebus atys). In addition, to begin to assess whether other non-human primates are also affected by leprosy, we have surveyed wild chimpanzees and ringtail lemurs using qPCR of DNA extracted from cheek swabs and wadges. Our preliminary results show that the mangabey strain is closely related to human strains. Further, the human West African and mangabey strains are ancestral to European and Asian strains, supporting an Africa origin for the disease. To determine the timeframe of disease spread, we calculated *M. leprae's* substitution rate based on the estimated divergence time between *M. leprae* and *M. tuberculosis*. The latter was obtained using a set of 101 homologous proteins across 27 bacterial taxa and a Bayesian relaxed-clock framework calibrated with reliable biogeochemical events. Our findings support the spread of leprosy from Africa to Europe and Asia around 10,000 years ago.

P-1332

Genome-wide mutation spectrum and rate of *Bacillus subtilis*

Way Sung, Emily Williams, Michael Lynch
Indiana Univeristy, Bloomington, USA

Mutations are the basis for all evolutionary processes, and understanding mutation enhances our knowledge of inheritance, divergence, and genetic disorders. Prior estimates of mutation rates using fluctuation assays and reporter constructs have yielded spontaneous mutation rates of approximately 0.0033 mutations per genome per generation across microbes. Here, as a complement to existing research on microbial mutation rates, we provide the first genome-wide rate and molecular spectrum of mutations in the gram-positive bacteria *Bacillus subtilis* strain 168. After 5000 generations of mutation accumulation (MA) with extreme bottlenecking we identify 353 base substitutions and 72 small insertion deletions (indels) across 50 mutation accumulation lines, yielding in a spontaneous mutation rate of 0.0018 per genome per generation; significantly lower than previously estimated. The observed ratio of non-synonymous and synonymous mutations do not significantly differ from neutral expectation (χ2 test, $P = 0.97$, 2 df), suggesting that purifying selection was not a significant force during the MA process. We observe a greater number of base substitutions in coding regions than expected by random chance, with a significant increase in G:A and T:C transitions on the non-coding strand (χ^2 test, $P < 0.05$, 2 df), which may arise from differences in either transcription-coupled mutation or repair. In addition, we also observe an elevated number of coding indels, which can be attributed to a large class of bacteriophage-like genes enriched with highly mutable repetitive AT residues. In this study, we provide the first genome-wide direct evidence that indels are biased towards deletions (χ^2 test, $P < 0.01$, 2 df). Although this study is limited to a single species, an observed deletion bias is consistent with the overall correlation for microbial genome sizes to decline with decreasing effective population size, as reduced selective pressure on neutral genomic features will drive the genome in the direction of mutation bias. Based on the observed MA base-substitution spectrum, if the nucleotide content of the *B. subtilis* genome is reached from mutation pressure alone, we would expect the genome to be composed of 56% A/T. Remarkably, this reflects the actual A/T base composition of *B. subtilis*, suggesting that the genome is very close to mutation equilibrium, and that the primary force driving G/C content in this organism is mutation.

P-1333

Genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* reveals compensatory mutations in the RNA polymerase

Iñaki Comas[1,2], Sonia Borrell[3,4], Andreas Roetzer[8], Graham Rose[2], Bijaya Malla[3], Midori Kato-Maeda[5], James Galagan[6,7], Stefan Niemann[8], Sebastien Gagneux[3,4]
[1]Center for Public Health Research (CSISP), Valencia, Spain, [2]National Institute for Medical Research, London, UK, [3]Swiss Tropical and Public Health Institute, Basel, Switzerland, [4]University of Basel, Basel, Switzerland, [5]San Francisco General Hospital- Univeristy of California San Francisco, San francisco, USA, [6]The Broad Insitute of MIT and Harvard University, Cambridge, USA, [7]Boston University, Boston, USA, [8]Research Center Bosrtel, Borstel, Germany

Like many other pathogens, *Mycobacterium tuberculosis*, the causative agent of tuberculosis, accumulates drug resistance mutations in response to antibiotic pressure. However, drug resistant strains are thought to carry a fitness cost in the absence of antibiotic, which can be restored by additional mutations. We sequenced rifampicin resistant strains of *Mycobacterium tuberculosis* and their susceptible counterparts. We were able to identify potential compensatory mutations in the *rpoC* and *rpoA* subunits of the RNA polymerases. We corroborated the importance of these mutations in two ways. *In vitr*o we measured the fitness gain of drug resistance strains carrying the mutation when compared to their susceptible counterparts. *In vivo* we demonstrated their importance in clinical settings by sequencing the corresponding genomic regions in drug resistant strains from countries with a high- and low-burden of drug resistant tuberculosis. We found that these compensatory mutations where significantly overrepresented in those countries showing the highest drug resistance burden, suggesting these mutations contribute to increased transmission of drug resistant strains. These findings have implications for our understanding of the transmissibility of drug resistant pathogens and also on how drug resistance evolves. From an evolutionary point of view, it reveals the potential of whole genome sequencing approaches to disentangle micro-evolutionary events in microbial populations to a detail and scale not possible before.

P-1334

Adaptive Regulatory Substitutions Affect Several Fitness Components of the Bacteriophage φX174

Celeste Brown, Amber Stancik
University of Idaho, Moscow, Idaho, USA

Gene regulation plays a central role in the evolutionary adaptation of organisms to their environments. In multiple evolution experiments with the bacteriophage φX174, adaptive substitutions in *cis*-acting regulatory sequences sweep through the phage population due to strong selection at high temperatures that are non-permissive for laboratory-adapted phage. For one *cis*-regulatory region, we have shown that each of four adaptive substitutions decrease transcript levels, and individually increase fitness for phages growing at a high, but permissive temperature. To determine the connection between reduced transcript levels and fitness, we have compared various fitness components between the ancestral strain and each of two (mutant) strains that differ from the ancestor by one substitution in this regulatory region. We have previously shown that the times to lysis for the mutant phages are earlier than the ancestor at the high temperature. We have since tested the adsorption rate, genome ejection rate, eclipse time (when viable capsid are first found inside the cell), capsid assembly rate and burst size for the ancestor and the mutant phages at permissive (37℃) and high (42℃) temperatures. We expected that there would be no significant difference in adsorption or ejection rate, because none of the adaptive substitutions are in the proteins that are involved in these processes, and this was true at both temperatures. The eclipse time of the ancestor was significantly earlier than the mutants at both temperatures, however, the rate of capsid assembly was faster for the mutants at 42℃. This resulted in a greater burst size for the mutants than the ancestor at high temperature. Our results show that decreasing gene transcription influences several fitness components in the bacteriophage φX174, resulting in greater fitness at high temperature. We speculate that the adaptive nature of these substitutions is due to the physiology of the host at high temperature or the need to maintain particular ratios of phage proteins during capsid assembly. Our continuing investigations will test these hypotheses.

P-1335

Rapid and Dynamic Evolution of Biosynthetic Gene Clusters in Microbial Genomes

Marnix H Medema[1,2], Peter Cimermancic[3], Michael A Fischbach[3], Rainer Breitling[2,4], Eriko Takano[1]
[1]Department of Microbial Physiology, University of Groningen, Groningen, The Netherlands, [2]Groningen Bioinformatics Centre, University of Groningen, Groningen, The Netherlands, [3]Department of Bioengineering and Therapeutic Sciences, UCSF, San Francisco, USA, [4]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

Huge, functionally tightly linked multi-gene clusters are fundamental units of microbial evolution. These clusters, of up to 200 kb in length and sometimes containing more than 40 genes, frequently encode the biosynthetic pathways leading to important bioactive secondary metabolites, such as signaling factors and antibiotics. Such secondary metabolites are often involved in intra-species and inter-species interactions that strongly influence organism survival and reproductive success. The reshuffling, recombination and horizontal transfer of secondary metabolite biosynthesis gene clusters underpins the accelerated evolution of the novel chemistry necessary to succeed in this continuous arms race [1,2].

Here we present the results of the first global analysis of the resulting evolutionary patterns, exploiting the completed genome sequences of 1154 eubacterial and archaeal species, which represent a large fraction of microbial diversity. Using two complementary algorithms for the genome-wide detection of biosynthetic gene clusters [3,4], we systematically identified almost 10,000 biosynthetic gene clusters across all species, which allowed a systematic analysis of their diversity and evolution. We found that biosynthetic gene clusters indeed evolve rapidly, with rearrangements and insertions/deletions occurring 5-10 times more frequently than in chromosomal regions involved in primary metabolism. Large (>20 kb) gene clusters encoding the biosynthesis of complex molecules appear to evolve by the successive merger of smaller subclusters coding for their constituent chemical moieties. Taxonomically, we observe that even species within the same genus sometimes differ dramatically in the number of biosynthetic gene clusters they encode, up to more than 10-fold within a single genus. This suggests that the gain and loss of biosynthetic gene clusters can play an important role in ecological differentiation and niche adaptation even on a relatively short evolutionary timescale.

The detailed knowledge gained about biosynthetic gene cluster evolution provides unprecedented insight into nature's strategies for chemical tinkering and diversification, which can now be practically applied in synthetic biology approaches for drug discovery and production [5,6].

1. Medema et al. (2010) Genome Biol. Evol. 2: 212-224.
2. Fischbach (2008) PNAS 105: 4601-4608.
3. Medema et al. (2011) Nucl. Acids. Res. 39: W339-W346.
4. Cimermancic, Medema et al. (2012), in preparation.
5. Medema et al. (2011) Nature Rev. Microbiol. 9: 131-137.
6. Medema et al. (2012) Nature Rev. Microbiol. 10: 191-202.

P-1336

Inter-domain horizontal gene transfer, a major role in the adaptation of archaea to mesophilic lifestyles?

Purificacion Lopez-Garcia[1], Philippe Deschamps[1], Yvan Zivanovic[0,2], David Moreira[1]
[1]Unite Ecologie, Systematique et Evolution, CNRS & Universite Paris-Sud, Orsay, France, [2]Institut de Genetique et Microbiologie, CNRS & Universite Paris-Sud, Orsay, France

Twenty years ago, the discovery of widespread marine planktonic archaea via their 16S rRNA genes in environmental samples abolished the traditional view that all archaea were extremophiles thriving in, often anoxic, hot, acid/alkaline or hypersaline environments. Previously, mesophilic archaea (Halobacteriales and some methanogens) were exclusively known within the Euryarchaeota. The newly discovered marine archaeal sequences fell into two different major groups: Group I Crenarchaeota, which was recently proposed to form an independent phylum, the Thaumarchaeota, and the Groups II/III Euryarchaeota. Thaumarchaeota have attracted much attention because of their ubiquitous presence in oceans and soils and the discovery that many of its members are ammonia oxidizers, thus contributing essentially to the N cycle. Despite their importance, only very few Thaumarchaeota have been isolated in pure culture and the information about their genome content is still limited. Marine Groups II and III are much more enigmatic and no member of these groups is available in culture. Metagenomic analyses are therefore appropriate tools to get access to the genes and genomes of these organisms. Phylogenetic analyses of fosmid ends and individual genes in 16S rRNA-gene-containing fosmids from deep-sea metagenomic libraries suggested a high level of horizontal gene transfer, mostly from bacteria. This appears to be further confirmed by the recent assembly of a consensus genome for Group II archaea from metagenomic data. Using more extensive sequencing of deep-sea metagenomic libraries, we show that inter-domain horizontal gene transfer is an ongoing process in marine archaea. Many of the transferred genes correspond to energy metabolism and metabolite transport across membranes, suggesting that horizontal gene transfer plays an important role in the environmental adaptation of these marine archaea. A preferential bacteria-to-archaea gene transfer has been also observed in halophilic archaea, which allows hypothesizing that the acquisition of foreign genes may have facilitated the independent adaptation of different archaeal phyla to mesophilic lifestyles.

P-1337

Mealybugs nested endosymbiosis: going deeply into functional and evolutionary aspects of the 'matryoshka' system in *Planococcus citri*.

Sergio López-Madrigal[1,2], Amparo Latorre[1,2], Manuel Porcar[3], Andrés Moya[1,2], Rosario Gil[1,2]
[1]*Institut Cavanilles de Biodiversitat i Biologia Evolutiva, UVEG, Valencia, Spain,* [2]*Departament de Genètica, UVEG, Valencia, Spain,* [3]*Fundació General de la Universitat de València, Valencia, Spain*

Many insect species are engaged in obligatory mutualistic symbiosis with intracellular bacteria that complement their unbalanced diets[1]. In some cases, two bacteria coexist in the same host, establishing a consortium that is needed for host fitness. In this context, mealybugs of the subfamily Pseudococcinae present a complex nested endosymbiotic system where a b-proteobacterium, *Tremblaya princeps,* harbours a g-proteobacterium, *Moranella endobia*[2]. Recent genome sequencing of two strains (PCIT and PCVAL) of *T. princeps* and *M. endobia* from *Planococcus citri* revealed an unprecedented functional complementation between them[3,4]. With the smallest bacterial genome described so far (139kb), *T princeps* has lost the ability for DNA replication and transcription, as well as part of its translational machinery and most metabolic functions (although it is still able to synthesize several essential amino acids). In contrast, *M. endobia* still retains most of the essential functions lost by *T. princeps*, suggesting that both bacteria could not be considered as independent organisms but as part of an unprecedented composed entity. Genome-driven reconstruction of the consortium functions, as well as comparative genomics and evolutionary analyses on these consortium genomes are revealing some clues of the meaning and evolutionary path of this peculiar system.

1.Baumann *et al*., 2005. Annu Rev Microbiol. 59:155-89

2. von Dohlen *et al*; 2001. Nature. 412(6845):433-6

3. McCutcheon JP and von Dohlen CD, 2011. Curr Biol. 21(16):1366-72

4.López-Madrigal S *et al*; 2011. J Bacteriol. 193(19):5587-8

P-1338

The origins of spontaneous mutations and the role of mismatch repair in shaping the genome of the model prokaryote, *Escherichia coli*

Patricia Foster, Ellen Popodi, Heewook Lee
Indiana University, Bloomington, IN, USA

The generation of genetic variation underlies evolution, and so accurate measures of mutation rates and mutational spectra are essential for understanding evolutionary processes. Our current knowledge of genomic mutation is largely based on two types of studies: comparisons of putatively neutral sequence changes that have accumulated between divergent species; and, extrapolation of experimental results from reporter constructs. Both methods can be subject to substantial uncertainties. With the advent of high-throughput sequencing and the application of mutation accumulation (MA) protocols, direct determinations of mutation rates and spectra have become possible. Here are reported the results of a 3,000 generation MA experiment on 35 lines of the model prokaryote *Escherichia coli* K12 strain MG1655. Whole-genome sequencing at 100x coverage revealed a base-pair substitution rate per nucleotide per generation of 1.7E-10, less than the 5.4E-10 estimated by Drake (1991, PNAS 88:7160), but in line with the 0.89E-10 recently estimated by Wielgoss et al (2011,G3 1:183). Sequenced MA lines of a mismatch-repair defective strain revealed the expected 150x increase in base-pair substitution rate. Interestingly, the spectrum of mutations in the absence of mismatch-repair was dramatically different from that in the wild-type strain. Loss of mismatch repair shifted the mutational bias from that of creating AT base pairs to that of creating GC base pairs. In addition, there were strong sequence-context specificities for mutations in the mismatch-repair defective background. The implications of these results for the mechanisms by which spontaneous mutations arise and for the impact of mismatch repair on genome evolution will be discussed.

P-1339

Sequence conservation and gene-conversion; insights into highly conserved elements in mating-type loci in *Saccharomyces cerevisiae*

Yutaka Watanabe, Denise Brooks, Alexander S. Mikheyev
Okinawa Institute Of Science And Technology, Onna-son, Japan

Comparative genomics, made possible by recent advances in sequencing, revealed the existence of numerous sequences, often non-protein coding, conserved at the nucleotide level across distantly related taxa. Although such ultra-conserved sequences are predicted to experience strong stabilizing selection, their functions remain largely unknown. Curiously, in some cases, experimental deletion of ultra-conserved regions has had no obvious phenotypic consequences. To understand this phenomenon, we examined the *HMRa2* locus of the budding yeast *Saccharomyces cerevisiae*. This gene shows 100% conservation at nucleotide level for 371bp among 5 different species in genus *Saccharomyces*. However, previous studies showed that deletion of *MATa2* has no phenotypic consequences under laboratory conditions. *HMRa2* acts as a donor sequence of *MAT* during mating-type switching, which occurs through a gene conversion event. Here we constructed a series of *HMRa2* mutants, ranging from complete deletion of the locus to a range of more subtle changes, and analyzed how those sequence modifications affect at mating-type switching. The fitness and mating-type switching capability of mutant strains were not different from those of the control strain. We do not find any evidences that mating-type switching (gene-conversion) events repair the introduced mutations. Thus, other mechanism(s) may have responsible for the highly conserved region.

P-1340

The genome of a gut parasite of *Daphnia* provides evidence for an ancient origin of Microsporidia.

Karen Haag[1], Tobias Schaer[1], Dominik Refardt[2], Dieter Ebert[1]
[1]University of Basel, Basel, Switzerland,[2]ETH, Zurich, Switzerland

Microsporidia are obligate intracellular parasites of animals well known for their reduced genomes, lack of mitochondria and for being transmitted by means of the ejection of a coiled structure known as the polar tube. All the most recent phylogenetic studies support the placement of these organisms on an early diverging branch within fungi. We employed the Illumina technology for a *de novo* sequencing of the genome of an undescribed intracellular gut parasite of *Daphnia magna*, showing several ultrastructural features of Microsporidia. The spores are surrounded by a two-layered cell wall, and a coiled structure resembling a polar tube is observed. However, our sequencing data indicate the presence of mitochondria. A contig of 13,145bp contains 6 typical mitochondrial genes: atp6, atp8, cox1, cox2, cox3 and cytB. Interestingly, ribosomal gene organization in the genome of the *Daphnia* gut parasite has a hybrid appearance. Microsporidia have typically reduced 23S and 16S ribosomal genes, separated by a very short intergenic spacer (100-200bp), and lacking the 5.8S gene, which in fungi is located between 28S and 18S. The 5.8S subunit in microsporidia is reduced to 5S and is normally encoded by multiple gene copies scattered throughout the genome. The *Daphnia* gut parasite genome shows six 5S genes in five different contigs, and the predicted 28S and 18S genes are separated by 1,174bp of non-coding sequence. Moreover, the 18S ribosomal gene shows significant hits to similar sequences from an ancestral branch of fungi (Chytridiomycota) as well as to protozoa (Eimeriidae). Interestingly, the 28S gene has no similarity to any microsporidian or fungal sequence, but shows highly significant blast hits to cnidarian and cercozoan 28S. We are now mining the genome for other microsporidian-like features (*e.g.* presence of genes encoding polar tube and spore wall proteins). At the moment, our data strongly suggest that the *Daphnia* gut parasite represents an ancestral lineage that precedes the origin of fungi.

P-1341

RNA-level unscrambling of systematically fragmented genes

Gertraud Burger[1], Georgette Kiethega[1], Yifei Yan[1], Marcel Turcotte[0]
[1]Universite de Montreal, QC, Canada, [2]University of Ottawa, ON, Canada

We previously reported a unique genome with systematically fragmented genes and gene pieces dispersed across numerous circular chromosomes, occurring in mitochondria of diplonemids. Genes are split into up to twelve short fragments (modules), which are separately transcribed and joined in a way that differs from known trans-splicing. Further, cox1 mRNA includes six non-encoded uridines indicating RNA editing. In the absence of recognizable cis-factors, we postulated that trans-splicing and RNA editing are directed by trans-acting molecules.
Our current work aims at understanding the post-transcriptional processes by investigating transcription, RNA processing, trans-splicing, and RNA editing in cox1 and at a newly discovered site in cob. We will report that module precursor transcripts are up to several thousand nt long and processed accurately at their 5' and 3' termini to yield the short coding-only regions. Processing at 5' and 3' ends occurs independently, and a processed terminus engages in trans-splicing even if the module's other terminus is not yet processed. Moreover, only cognate module transcripts join, though without directionality. In contrast, module transcripts requiring RNA editing only trans-splice when editing is completed. Finally we will show experimental and computational evidence for the existence of trans-factors with the potential for guiding both trans-splicing and RNA editing.

P-1342

Cloning, characterization and expression of an insecticidal crystal protein gene from *Bacillus thuringiensis* isolates of Andaman and Nicobar Islands

H.M.MAHADEVA SWAMY[1], RAMASAMY ASOAKN[1], GEETHA G. THIMMEGOWDA[1], RIAZ MAHMOOD[2], NAGESHA S N[1], DILIP KUMAR ARORA[3]
[1]INDIAN INSTITUTE OF HORTICULTURAL RESEARCH (IIHR), BANGALORE, KARNATAKA, India, [2]KUVEMPU UNIVERSITY, SHIMOGA, KARNATAKA, India, [3]NATIONAL BUREAU OF AGRICULTURALLY IMPORTANT MICROORGANISMS, MAU NATH BHANJAN, UTTAR PRADESH, India

Biocontrol of pests via *Bacillus thuringiensis* (Bt) d-endotoxins represents the most successful use of a biological control agent to date. The most notable characteristic of *Bacillus thuringiensis* is its ability to produce insecticidal proteins. More than 300 different proteins have been described with specific activity against insect species. The six isolates of *Bacillus thuringiensis* from Andaman and Nicobar Islands which were previously characterized by PCR analysis for the presence of Coleopteran active *cry* genes were used for *Cry1I* full length gene amplification. A 2.16-kb DNA fragment of *Cry1I* gene was PCR amplified, cloned in expression vector pQE 80 L, and then used for transformation of *E. coli M15* cells. The optimum expression was obtained with 1 mM IPTG at 37^0C for 3 h. The sequence of the cloned crystal protein gene showed almost complete homology with a *Cry1I* toxin gene from *Bacillus thuringiensis* var. *kurstaki*, with scattered mutations in the toxic region. The deduced sequence of the protein has homologies of 91.0% with *Cry1I* and *Cry1Ia*, and 98.0% with *Cry1Ib*.Cloning of this gene may help to overcome the increasing resistance of pests to currently used insecticides. Based on the results obtained, the PCR method may be a valuable and reliable tool for specific detection and identification of *cry1I* genes. The toxicity of Bt recombinant protein was determined against first instar larvae of *Myllocerus undecimpustulatus undatus* Marshall (Coleoptera: Curculionidae) and Adults; *Helicoverpa armigera* Hübner (Noctuidae: Lepidoptera) at 310*µ*g/mL and 15.5*µ*g/mL respectively. The novel *cry*1I gene will be an important resource in constructing genetically engineered bacteria and transgenic plants for biocontrol of insect pests and Bt based biopesticidal formulations, aiming to reduce the use of chemical insecticides.

P-1343

Comparative Analysis of Gene Content Evolution in Phytoplasmas and Mycoplasmas

Ling-Ling Chen[1], Wan-Chia Chung[1], Chan-Pin Lin[0,2], Chih-Horng Kuo[1]
[1]Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan, [2]Department of Plant Pathology and Microbiology, National Taiwan University, Taipei, Taiwan

Phytoplasmas and mycoplasmas are two groups of important pathogens in the bacterial class Mollicutes. Because of their economical and clinical importance, these obligate pathogens have attracted much research attention. However, difficulties involved in the empirical study of these bacteria, particularly the fact that phytoplasmas have not yet been successfully cultivated outside of their hosts despite decades of attempts, have greatly hampered research progress. With the rapid advancements in genome sequencing, comparative genome analysis provides a new approach to facilitate our understanding of these bacteria. In this study, our main focus is to investigate the evolution of gene content in phytoplasmas, mycoplasmas, and their common ancestor. By using a phylogenetic framework for comparative analysis of 12 complete genome sequences, we characterized the putative gains and losses of genes in these obligate parasites. Our results demonstrated that the degradation of metabolic capacities in these bacteria has occurred predominantly in the common ancestor of Mollicutes, prior to the evolutionary split of phytoplasmas and mycoplasmas. Furthermore, we identified a list of genes that are acquired by the common ancestor of phytoplasmas and are conserved across all strains with complete genome sequences available. These genes include several putative effectors for the interactions with hosts and may be good candidates for future functional characterization.

P-1344

Aspergillus genes found in *Zea Mays* shotgun reads

Chun-Lin Wang[1,2], Yi-Pei Ho[1], Chien-Chi Chen[2], Shih-Hau Chiu[2], Li-Min Sung[2], Tsu-pei Chiu[2], Yeong-Shin Lin[1]
[1]*Institute of Bioinformatics and Systems Biology, National Chiao-Tung University, 75 Po-Ai Street, Hsinchu, Taiwan,*
[2]*Bioresource Collection and Research Center, Food Industry Research and Development Institute, 331 Shih-Pin Road, Hsinchu, Taiwan*

Aspergillus species play an influential role in medical, agricultural and commercial areas. We retrieved the predicted *Aspergillus* open reading frame (AORF) sequences of three species (*A. fumigatus*, *A. nidulans* and *A. niger*) from *Aspergillus Genome Database* and performed TBLASTN analysis against NCBI Whole Genome Database. Surprisingly, we found that some of the AORF sequences were highly conserved in the *Zea mays* genome and EST shotgun reads.

There are at least two possibilities to account for this phenomenon: (1) Horizontal gene transfer events might happen between *Aspergillus* and *Zea mays*. (2) *Aspergillus*-contaminated samples of *Zea mays* were used for conducting high throughput sequencing.

Our further analyses indicated that although these AORFs hit *Zea mays* EST reads, we could not map them to *Zea mays* genome draft. At present, we postulate that the *Aspergillus* genetic material found might be unexpected contamination in *Zea mays* since *Aspergillus* sp. is usually found to live with crops. Furthermore, we performed *Gene Ontology* analysis and found that many AORFs were related to primary metabolic processes (e.g., carbohydrate, lipid, protein, or amino acid) while some others may have cation transmembrane activity.

This unintended finding is important since it provides direct evidence for the expression of some *Aspergillus* genes in crops. Further studies could be useful for *Aspergillus* related researches.

P-1345

Divergence of bacterial gut symbionts between two floricolous flies, *Colocasiomyia alocasiae* and *C. xenalocasiae*

Jia-Syuan Chen[1], Chau-Ti Ting[1,2], Shu Fang[3], Shun-Chern Tsaur[4]
[1]Institute of Zoology, National Taiwan University, Taipei, Taiwan, [2]Department of Life Science & Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan, [3]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan, [4]Department of Life Sciences & Institute of Genome Sciences, National Yang-Ming University, Taipei, Taiwan

Drosophilids could utilize a variety of food resources, and therefore is a powerful model system in understanding the molecular mechanisms and ecological roles of host-symbionts interactions. Previous studies have been focused on the different laboratory strains to investigate the functional gut microbial fauna from various food sources and adapt to the variant environments. However, forces shaping the host-microbe relationship and driving the divergence of gut bacterial symbionts in the natural population are largely unknown. To address this question, we compared the bacterial communities of two diet specific fly species, *Colocasiomyia alocasiae* and *C. xenalocasiae*, both feeding on the inflorescences of *Alocasia odora* and *Colocasia formosana*, and the surface of the feeding area of the two host plants by bacterial specific 16S rDNA sequencing. We found that the bacteria taxa were restricted in only few dominant bacterial families, including *Bartonellaceae, Enterobacteriaceae, Enterococcaceae*, ,and *Lactobacillaceae*. Diet was the primary factor influencing the characterizing of gut bacterial community. In addition, the similar bacterial taxa were observed from the intestines and the surface of fly host plants with different abundance. The results reveal that although the diet source is the main element affecting the gut bacterial community, the internal conditions of fly also appear to exert some level of control over the bacteria in shaping the bacterial communities in the digestive tract.

P-1346

Computational Prediction and Experimental Validation of Hydrogenosomal Proteins Coded in the Nuclear Genome of *Trichomonas vaginalis*

David Burstein[1], Sven Gould[2], Verena Zimorski[2], Thorsten Kloesges[2], Fuat Kiosse[2], Peter Major[2], William Martin[2], Tal Pupko[1], Tal Dagan[2]
[1]Tel Aviv University, Tel Aviv, Israel,[2]Heinrich-Heine University, Düsseldorf, Germany

Hydrogenosomes are mitochondrion-like organelles that produce ATP under anaerobic conditions. They share a common ancestor with mitochondria, but are scattered over the eukaryotic domain, suggesting that the organelle specialization to anaerobic lifestyle occurred several times during evolution. *Trichomonas vaginalis,* a unicellular flagellate that bears hydrogenosomes, is the causative agent of the most prevalent non-viral sexually transmitted disease in human. The hydrogenosomes of Trichomonad are devoid of a genome, so the *T. vaginalis* nuclear genome is a mosaic composed of protein coding genes of eukaryotic origin and of hydrogenosome origin. Our goal was to identify novel *T. vaginalis* proteins targeted to the hydrogenosome on a genomic scale. We formulate the task as a classification problem: a set of features potentially distinguishing hydrogenosome-targeted proteins from other proteins was established. The values of these features for proteins with known localization were used to train a variety of classification algorithms. Feature assessment showed that evolutionary-based distance to proteobacterial proteins was among the most informative features. The classifiers were used to predict for each *T. vaginalis* open reading frame its likelihood to encode a protein targeted to the hydrogenosome. Ten high scoring predictions were experimentally validated by *in vivo* localization studies, yielding the identification of six novel hydrogenosomal proteins. Analysis of such proteins provides insights into the evolutionary origin of hydrogenosomes.

P-1347

Signatures of nitrogen limitation in the Bison Pool metagenomeSignatures of nitrogen limitation in the Bison Pool metagenome

Claudia Acquisti, Tabea Hoemann
IEB, WWU Muenster, Muenster, Germany

Recent advances have shown a direct impact of resource constraints from the environment on the evolution of genes and proteins of species, suggesting that the material costs of evolutionary change play a pivotal role in constraining the evolution of species in response to nutrient limitations in natural ecosystem [1]. For example, the nitrogen (N) content of molecular sequences has been established as a marker to trace connections between the genome and the eco-phisiology of the organisms, using few bacteria and plant model organisms as representative of primary producers and contrasted to few animal and fungi model organisms, considered as consumers [2-3]. However, these results have primarily relied on few well established genetic model organisms, leaving the question of the relevance of adaptation to nutrient availability in natural environments only partially addressed. Recent advances in metagenomics allow to extend our understanding of the impact of the evolutionary history of nutrient limitation on molecular evolution in a biogeochemical framework, providing a major arena to directly quantify the allocation of nutrients from the abiotic habitat to genes and proteins in environmental samples.

Bison Pool, a flowing alkaline hot spring in the Lower Geyser Basin of Yellowstone National Park, represents an ideal ecosystem to investigate the role of nitrogen availability in shaping evolutionary change in natural communities of microorganisms, owing to the availability of metagenomic and geochemical data along a temperature, and a N availability gradient. The combination of very low N availability in the source waters (both in mineral and organic forms) and the high temperature (above 95 degrees) hindering N fixation, makes the hotter part of the pool a severely N limited environment to the bacterial communities. As the temperature decreases along the flow of the water N fixation becomes possible (at temperatures below 68 C), reducing the severity of N limitation in the colder spots. Analyzing metagenomic samples along a temperature gradient, we have found that the allocation of N in ribosomal proteins follows the environmental availability of N. We have corrected our estimates of effective N allocation in ribosomal proteins for phylogenetic and temperature effects, and still found a significant difference between the sites with different rates of N fixation.

1. Elser JJ, Acquisti C, Kumar S. 2011. *TREE* doi:10.1016/j.tree.2010.10.006
2. Acquisti C, Elser JJ, Kumar S 2009 *Mol. Biol. Evol.* 26, 953-956
3. Acquisti C, Kumar S, Elser JJ 2009 *Proc. R. Soc. London B* 276, 2605-2610

P-1348

The PhyloFacts Phylogenomic Encyclopedia of Gene Families Across the Tree of Life

Jonathan Dobbie, Cyrus Afrasiabi, Ajithkumar Warrier, Bushra Samad, Kimmen Sjölander
University of California, Berkeley, Berkeley, CA, USA

PhyloFacts is a database of gene families across the Tree of Life, available at http://phylogenomics.berkeley.edu/phylofacts. PhyloFacts is designed to serve the needs of evolutionary biologists needing high-accuracy orthologs and phylogenetic trees to reconstruct species phylogenies, and of molecular biologists and genome centers interested in predicting gene function in an evolutionary context. Most PhyloFacts families span multiple paralogous groups (i.e., including gene duplication) and represent many distinct functions. Phylogenetic trees are analyzed using the PHOG algorithm to identify subtrees corresponding to groups of orthologs, allowing both evolutionary and functional inferences.

As of April 2012, PhyloFacts contains >8M proteins from the UniProt resource clustered into >92K gene families spanning >99K unique taxa (including strains). Each PhyloFacts family includes a phylogenetic tree, a multiple sequence alignment, hidden Markov model, predicted Pfam domains, homologous PDB structures, Gene Ontology annotations, BioCyc pathway data, and various other sources of functional annotations. Roughly 2/3 of PhyloFacts families represent multi-domain architectures (in which sequences are required to align well along their entire lengths), with the remainder representing trees constructed for individual Pfam domains. These two distinct types of phylogenetic clustering allow biologists to explore the evolution of gene function following exon shuffling and other sources of domain architecture rearrangements.

As mentioned above, PhyloFacts includes >7M proteins spanning >99K unique taxa giving us close to complete coverage of many genomes. For instance, within Eukarya, >91% of Homo sapiens, >87% of Mus musculus, >91% of Drosophila melanogaster, >90% of Saccharomyces cerevisiae and >86% of Arabidopsis thaliana genes are represented in one or more PhyloFacts trees. Within Archaea, >90% of Halobacterium salinarum and >87% of Sulfolobus solfataricus are represented by at least one PhyloFacts family. Within Bacteria, 100% of Escherichia coli K12, >94% of Bacillus subtilis, >91% of Thermotoga maritime, >90% of Geobacter sulfurreducens, >87% of Sulfolobus solfataricus and >79% of Deinococcus radiodurans are represented. Many more genomes are represented at different levels of coverage; details are at http://phylogenomics.berkeley.edu/phylofacts/coverage/.

We are developing a novel system to automate the functional annotation of genomes and to provide simultaneous functional and taxonomic annotation of metagenome datasets. This system, called FAT-CAT, for Fast Approximate Tree Classification, uses HMMs at internal nodes of PhyloFacts trees. Internal nodes are annotated using functional and taxonomic data for sequences descending from that node, and enable us to annotate new sequences based on their top-scoring subtree HMM.

P-1349

GC bias in recombining loci of bacteria and archaea: Evidence that biased gene conversion is universal

Christopher Graves[1,2], Dustin Brisson[2]
[1]Brown University, Providence, RI, USA, [2]University of Pennsylvania, Philadelphia, PA, USA

Local GC content is correlated with increased rates of homologous recombination in eukaryotic genomes. It is hypothesized that this relationship is caused by a systematic bias toward GC substitutions in the process of homologous recombination. While supported across eukaryotes, the hypothesis that biased gene-conversion affects the nucleotide composition of bacteria and archaea has not been thoroughly examined. Although these organisms do not undergo meiotic recombination, there is substantial evidence of recombination among multi-copy gene families. We compared GC content across the genomes of over 2000 species of bacteria and archaea and observed a strong bias in local GC content in the 16S rRNA genes. Similar biases in unrelated gene families shown to frequently undergo homologous recombination suggest that the GC bias is a general feature of recombining loci rather than a result of stabilizing selection on the rRNA. In addition, analysis of template-independent sequence changes occurring in a recombination prone gene family of the Lyme disease bacterium, *Borrelia burgdorferi*, suggest that substitution biases during gene conversion may explain the strong GC bias at these loci. We conclude that biased gene conversion may be an important process shaping nucleotide composition within bacterial genomes and highlight possible implications of our results for studies of molecular evolution and pathogenesis in bacteria.

P-1350

Thermodynamic constraints on protein evolution: sequences, structures and models

Johan Grahnen
University of Wyoming, Laramie, WY, USA

Insight into the thermodynamics of protein folding and function is crucial to understanding the underlying mechanistic processes that drive the evolution of individual proteins. On a larger scale, proteins function within the context of the full cellular complement of molecules, which ultimately give rise to organismal fitness effects and population-level changes in allele frequencies. Using a novel biophysical model of protein evolution that takes population genetics into account, we show that thermodynamics of folding and binding can drive many of the observed characteristics of protein sequences: sequence composition, the evolutionary rate and its distribution across sites, binding specificity and pleiotropic effects, etc. Recently, there have been a number of attempts to explicitly model site-interdependence due to thermodynamics in a phylogenetic context, but success has been somewhat elusive. Our model suggests that the relatively low utility compared to site-independent approaches stems from a violation of the fundamental assumption of high stability of the native state. We also apply this biophysical perspective to studying the mechanistic causes of co-variation between residues, with particular attention paid to the claim that rate heterotachy is a strong predictor of functional change. Our simulation studies are mostly consistent with this view, but also emphasize that intra-molecular co-evolution is a constantly ongoing process even in the absence of diversifying selection. Finally, we present some preliminary estimates of the ease of neofunctionalization and the probability of convergent evolution in a model system of protein-ligand interaction.

P-1351

Not Genetic Hitchhiking but Positive Selection Plays a Major Role in Influenza A Hemagglutinin Evolution

Li-Ching Hsieh[1,2], Yi-Chiao Fang[3], Wen-Hsiung Li[4,5], Arthur Chun-Chieh Shih[3]
[1]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung, Taiwan, [2]Biotechnology Center, National Chung Hsing University, Taichung, Taiwan, [3]Institute of Information Science, Academia Sinica, Taipei, Taiwan, [4]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan, [5]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA

For understanding influenza A/H3N2 virus evolution and facilitating vaccine strain prediction, several studies have been conducted to determine which sites on the HA1 domain of the hemagglutinin (HA) are under positive selection pressure. A number of positive selection sites have been identified by using the criteria that a positively selected codon may exhibit a significant excess of nonsynonymous over synonymous nucleotide substitutions throughout individual branches of a phylogenetic tree. However, many substitutions fixed in the population did not occur on these sites and most of the substitutions lead to simultaneous (parallel) multiple amino acid fixations.

Our previous study proposes that the simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. However, theoretically, evolution may occur without positive selection, called hitchhiking. Therefore, whether or not the majority of these parallel substitutions were due to hitchhiking is still controversial. In this study, we propose a new method to examine whether a substitution is due to hitchhiking. We design a frequency-based method without utilizing a phylogenetic tree to estimate the selection pressure acting at each single amino acid site of the H3HA1. Then, using the latest data, we identify 39 substitutions at 35 sites from 1993 to 2009 and show that most substitutions occurred very rapidly. Among the 35 sites, 32 sites may have undergone positive selection while only three sites were under slight negative selection. Therefore, for the 35 sites of simultaneous substitutions on the H3HA1, hitchhiking likely plays a minor role; instead, most of the 35 sites probably have undergone positive selection.

P-1352

Biased gene conversion in codon substitution models

Laurent Guéguen, Laurent Duret
LBBE - UMR CNRS 5558, Lyon, France

Biased gene conversion is a major factor in the process of molecular evolution in many species. This mechanism is directly associated to recombination, which is known to be highly heterogeneous, both in space (recombination events tend to cluster in hotspots within the genome) and in time (the location of recombination hotspots varies rapidly). As a consequence, the base composition (G+C content) is heterogeneous inside these genomes, and not at the equilibrium of the process.

Although usual models of codon substitution take into account the base composition of the sequences, they do not explicitly consider the dynamics due to biased gene conversion, neither the mechanism itself nor the non-stationarity of G+C content in the sequences.

We show on simulated data how much the estimation of selection is biased when biased gene conversion is not considered in codon substitution models, considering both the specificity of the mechanism and the non-stationarity of the process.

To adjust this estimation, we first show on human genes with several G+C contents how taking into account the non-stationarity of the process improves the likelihood of the models. Then we present a new probabilistic modelling of codon substitution that explicitly takes into account biased gene conversion, and proceed with this model on the estimation of selection in human genes together with biased gene conversion.

P-1353

Determining the power and robustness of the branch-site test of positive selection under conditions of strong divergence

Walid Gharib[1,2], Marc Robinson-Rechavi[1,2]
[1]*University of Lausanne, Lausanne, Switzerland,* [2]*SIB Swiss institute of bioinformatics, Lausanne, Switzerland*

Positive selection is widely estimated from protein coding sequences by the non synonymous / synonymous ratio w. Increasingly elaborate codon models are used in a likelihood framework for this estimation. While there is widespread concern about the robustness of the estimation of the w ratio, there have been few efforts to evaluate this robustness for the more elaborate, and biologically realistic, models. Here we focus on the branch-site codon model.

We investigate the robustness of the branch-site model to detect selection on a large set of simulated data. We performed simulation studies based on parameters from real vertebrate data (i.e., distributions of gene lengths, tree topologies, etc.).

First, we investigate the impact of sequence divergence on the estimation of the synonymous substitution rate (dS). We found evidence of under-estimation of dS for values ~0.8, and at this point the branch-site test has a maximum of false positives (but still under the FDR fixed at 10%). When dS increases further, under-estimation of dS is worse, but false positives decrease. Interestingly, the detection of true positives follows a similar distribution, with a maximum for intermediary values of dS. Thus high dS is more of a concern for a loss of power than for false positives of the test. This confirms that the branch-site is a very conservative test.

Second, we investigate the impact of GC content. Mammalian genomes are highly variant in G+C content (~30-60% in a typical mammal). Moreover, studies have shown that GC-BGC (GC-biased gene conversion) both contributes to this variation in GC, and can be confounded with natural selection. We show that high GC content (~65%) genes generate considerably less false positives then low GC (~30%). On the other hand, a shift of average GC content on a specific branch (~40 to 30 %), simulating the effect of GC-BGC, does not generate many false positives. Similarly, major shifts in GC along the gene sequence do not generate many false positives.

Third, strong positive selection on one branch does not create false positive detection of positive selection on other branches of the same tree.

In conclusion, we have delineated a parameter space inside which the branch-site appears reliable. This space extends further than expected for sequence divergence or GC shifts, but the test appears very sensitive to extreme base composition of sequences.

P-1354

SlimCodeML: An Optimized Version of CodeML for the Branch-Site Model

Hannes Schabauer[1,2], Mario Valle[3], Christoph Pacher[4], Heinz Stockinger[2], Alexandros Stamatakis[5], Marc Robinson-Rechavi[1,2], Ziheng Yang[6], Nicolas Salamin[1,2]
[1]University of Lausanne, Lausanne, Switzerland, [2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3]Swiss National Supercomputing Center, Lugano, Switzerland, [4]AIT Austrian Institute of Technology, Vienna, Austria, [5]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany, [6]University College London, London, UK

The recent advances in high-throughput sequencing techniques are providing researchers with a wealth of genome scale data that allows us to study evolutionary questions that were not feasible a decade ago. Notably, the increase in molecular data available for non-model organisms is creating a surge toward comparative genomics that is strengthened by recent new theoretical developments. However, the full potential of these new data will only be achieved by the developments of further computational and mathematical techniques. The increasing number of available genomes entails a dramatic growth in computational complexity of phylogenetic trees. One common approach to assess a phylogenetic tree is based on the maximization of its likelihood. Our aim is to reduce the computational time when optimizing the phylogenetic likelihood function used to estimate the parameters of the substitution matrix. This is done through a series of novel optimizations which we implemented in existing software. Our new approach makes it feasible to use these codon models on large-scale genomic data.

CodeML (PAML) implements different codon substitution models by means of continuous-time Markov models. We investigate specifically the branch-site model (BSM). The likelihood is computed using Felsenstein's pruning algorithm, which conducts a post-order tree traversal propagating from the leaves toward the root. Along each branch, a partial conditional probability vector for the branch's parent node is computed by applying the transition probability matrix to the child's conditional probability vector. The computationally most demanding parts to estimate the likelihood of the BSM are the matrix exponential to obtain the transition probability matrix and the computation of the conditional probability vector at each branch.

We present SlimCodeML, an optimized version of CodeML for the BSM. SlimCodeML features an improved likelihood computation, including a novel approach to compute the matrix exponential exp(Qt), implemented utilizing the BLAS and LAPACK libraries. We evidence the superiority of our approach by discussing accuracy and runtimes on four different datasets; SlimCodeML has very encouraging performance with speedups up to 9.38. SlimCodeML represents the first step toward FastCodeML which will be a new parallel and distributed version of CodeML adapted to the BSM test. While we focus here on the BSM, the optimized likelihood computation can also be applied to other maximum likelihood-based evolutionary models.

P-1355

Simulating genome evolution with ALF

Daniel Dalquen[1,2], Maria Anisimova[1,2], Gaston Gonnet[1,2], Christophe Dessimoz[1,2]
[1]ETH Zurich, Zürich, Switzerland, [2]Swiss Institute of Bioinformatics, Zürich, Switzerland

Validation and benchmarking are challenging tasks in computational evolutionary biology because, except for some cases where data could be extracted from fossils, the evolutionary history of biological entities studied is usually not known. Using computer programs for simulating sequence evolution *in silico* is therefore an accepted and widely used way to characterize newly developed models and methods under controlled conditions. However, current simulation packages tend to focus on gene-level aspects of genome evolution such as character substitutions and indels, population-level events such as recombination and gene conversion, or on genome-level aspects such as genome rearrangement and speciation events.

In this talk, we introduce a new simulation program called *ALF* (for artificial life framework), which we developed with the long-term goal of simulating the entire range of evolutionary forces that act on genomes (Dalquen et al. Mol Biol Evol. 2011). In the first release, we primarily focussed on species-level evolution where an ancestral genome, represented by an ordered set of sequences, is evolved along a tree into a number of descendant synthetic genomes. At the gene-level, ALF can simulate evolution at the nucleotide, codon or amino acid level with indels and among-site heterogeneity, supporting most established models of character substitution. Different types of sequences can be mimicked by defining several sequence classes with separate models of substitution, insertion-deletion and among-site rate variation. At the genomic level, ALF simulates GC content amelioration, gene duplication and loss, genome rearrangements and lateral gene transfer. A user-friendly web interface facilitates the setup of new simulations.

We illustrate the utility of ALF with an example study demonstrating that lateral gene transfer can dramatically decrease the accuracy of two well established methods for orthology inference. We augment this study with new, unpublished results that give a broader view on the effects of evolutionary events on orthology prediction. Finally, we discuss recent improvements of ALF towards the simulation of population-level events like recombination.

P-1356

Bayesian phylogenomics under mutation-selection codon substitution models

Nicolas Rodrigue[1,2], Daniel Stubbs[3], Jacques Richer[3], Nicolas Lartillot[4]
[1]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada, [2]University of Ottawa, Ottawa, Ontario, Canada, [3]Calcul Québec, Montréal, Québec, Canada, [4]Université de Montréal, Montréal, Québec, Canada

Bayesian codon substitution models are of great interest for the study of protein coding DNA sequence evolution. However, the computational cost of these models remains a significant hurdle to their wider use on modern data sets. Here, we combine several technical advances to implement rich, computationally intensive, codon substitution models, and focus on those based on the mutation-selection principle. The models of interest include the homogeneous model of Yang & Nielsen (Mol. Biol. Evol., 2008, 25:568-579), the fully site-heterogeneous model of Halpern & Bruno (Mol. Biol. Evol., 1998, 15:1499-1505), and the infinite mixture model based on the Dirichlet process (DP) of Rodrigue et al. (PNAS, 2010, 107:4629-4634). Though many applications are possible, the models are used here to evaluate distributions of selection coefficients, while accounting for uncertainty regarding substitution parameters and tree topology. The computational advances enabling such applications include: i) a Gibbs subtree pruning and regrafting update operator on the tree topology (conditional on the parameters of the substitution model) that efficiently exploits conditional likelihood caching; ii) data-augmentation-based Gibbs and Metropolis-Hastings operators on the parameters of the substitution model (conditional on the tree topology); and iii) full multi-core across-site parallelization of likelihood calculations for tree topology moves using a message passing interface (MPI) system and full (under site-specific models) or partial (under DP models) parallelization for moves on parameters of the substitution process. We apply our implementation to a data set comprised of 244 mammalian mitochondrial genomes (taken from Tamuri, dos Reis & Goldstein, 2012, Genetics, in press) to demonstrate its usefulness and scalability.

P-1357

The Likelihood Landscape for Phylogenetic Models

Barbara Wilhelm, Peter F. Arndt
Max Planck Institute for Molecular Genetics, Berlin, Germany

To unravel the evolution of genomes has always been a topic of high scientific interest. Probabilistic models have proven to be an efficient tool to break down the complexity of this process when analyzing phylogenetic data. A matrix of rate parameters for nucleotide substitutions, the topology of the phylogeny, branch lengths, and other parameters summarize the given information. It is an important task to estimate all these model parameters on the basis of given sequence data. In this field maximum likelihood estimation is a popular statistical method. For many years now it has also been applied in phylogenetic analysis. But it is important to note that the maximum of the likelihood function does not necessarily exist. There are cases where the likelihood function does not nicely peak at a particular value but where the maximum is degenerate, i.e. where the maximum is a ridge. In the latter case the parameter estimation will fail. We investigate this issue and the influence of the tree topology on the ability to estimate all parameters for phylogenetic models.

P-1358

The evolution of GC content in avian genomes

Carina Mugal[1], Peter Arndt[2], Hans Ellegren[1]
[1]Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden,[2]Max Planck Institute for Molecular Genetics, Berlin, Germany

The genomes of many vertebrates, including mammals and birds, show a characteristic heterogeneous distribution of the local GC content, the so-called isochore structure of the genome. By now, the origin of isochores has been explained via the mechanism of GC-biased gene conversion (gBGC), i.e. short-scale, unidirectional exchanges between homologous chromosomes in the neighborhood of recombination-initiating double-strand breaks, where AT/GC heterozygotes produce more GC- than AT-gametes. Whereas the isochore structure is declining in many mammalian genomes, the heterogeneity in GC content is being reinforced in the avian genome. Despite this discrepancy, examinations of individual mammalian and avian substitution frequencies, are both consistent with the gBGC model of isochore evolution. However, a negative correlation between the local substitution rate and the local recombination rate present in the avian genome appears to be inconsistent with the gBGC model of evolution. Hence, it seems important to consider along with gBGC other consequences of recombination on the origin of mutations and their probability of fixation, as well as to take relationships of recombination rate to other genomic features into account.

In order to investigate the negative correlation between the local substitution rate and the local recombination rate, we developed a minimal analytical model to describe the substitution pattern found in the avian genome and compared simulated data to observed data. This analysis sheds light into which other genomic features and aspects of recombination impact on the local substitution pattern and the evolution of GC content in the avian genome. The results indicate that the local GC content itself, either directly or indirectly via interrelations to other genomic features, has an impact on the local substitution pattern by affecting the rate of mutation. Further, we suggest that this phenomenon is specific to avian genomes due to their unusually slow rate of chromosomal evolution, where many chromosomes have remained more or less intact during avian evolution. Because of this, interrelations between the local GC content and other genomic features are more pronounced in the avian genome.

P-1359

Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species

Suo Qiu[1,2], Kai Zeng[1], Tanja Slotte[4], Stephen Wright[3], Deborah Charlesworth[1]
[1]University of Edinburgh, Scotland, UK, [2]Sun Yat-Sen University, Guangzhou, China, [3]University of Toronto, Ontario, Canada, [4]Uppsala University, Uppsala, Sweden

A reduced efficacy of natural selection is predicted in self-fertilizing species, because of their low effective population size (N_e). We tested this by examining silent site polymorphisms (at synonymous and intron sites), using two comparisons between closely related selfing and outcrossing plants for which extensive DNA sequence polymorphism datasets have been obtained: *Arabidopsis thaliana* (selfer) with *A. lyrata* (outcrosser), and *Capsella rubella* (selfer) versus *C. grandiflora* (outcrosser). Our analysis employed extensions of two recently developed models that allowed us to investigate selective differences between synonymous codons, mutational biases, and biased gene conversion, taking into account possible recent changes in population size. We found evidence for selection on synonymous codons, and evidence that this is significantly weaker in the selfers compared to the outcrossers, and that this is not due purely to mutational biases or biased gene conversion.

P-1360

Toward Computationally Feasible Evolutionary Inference with Biologically Plausible Models

Cory L. Strope, Jeffrey L. Thorne
Bioinformatics Research Center, North Carolina State University, Raleigh, USA

Conventional methods for making inferences about molecular evolution assume that sequence positions evolve independently of one another. This assumption can be computationally convenient because it allows the likelihood of a set of aligned sequences on a phylogenetic tree to be expressed as a product of likelihoods for individual sites or codons. However, both natural selection and context-dependent mutation can violate this independence assumption.

We are interested in making evolutionary inferences about natural selection. We do this by connecting genotypes (i.e., DNA sequences) to some aspect of the phenotypes that they encode (e.g., protein tertiary structure) and we then try to translate phenotypic values to relative fitnesses. Because sequence positions do not make independent contributions to phenotype (or fitness), we accommodate the dependent evolution among sites by augmenting the sequences observed at the tips of an evolutionary tree with a history that could have generated the observed sequences. Each history or "stochastic mapping" consists of specific sequence changes that occurred at specific points in time on a phylogeny. Because many possible histories are consistent with a set of observed sequences, Markov chain Monte Carlo (MCMC) techniques can be employed to approximate the posterior distribution of sequence histories by taking a random walk among possible histories. At each step in the random walk, a change to the current sequence history is randomly proposed and then the proposed change is accepted or rejected with the appropriate probability.

We have been exploring strategies for constructing the random walk. Rather than proposing a small change to a sequence history such as would occur if a proposed sequence history differed from the current sequence history only in the history of one site, we are interested in whether we can propose more dramatic changes in sequence history that still have a relatively large chance of being accepted. Success with this proposal strategy would improve the computational tractability of statistical inference with biologically appealing treatments of natural selection.

To date, our experiments have resulted in mixed success and, as should be anticipated, our strategy is more successful when evolutionary dependence is weak. In this talk, we summarize our experiments with simple models of evolutionary dependence among sites. Future directions of this line of research include efforts to improve ancestral sequence inference and better characterization of the fitness landscape.

P-1361

Goodness of fit tests and outlier identification in phylogenetics

Steffen Klaere
Department of Statistics, University of Auckland, Auckland, New Zealand

A goodness of fit test evaluates whether or not a datum set supports a model inferred by, e.g., maximum likelihood. Thus, it provides the means to reject a null hypothesis. In the case of phylogenetics this null hypothesis combines tree-like evolution and Markovian models of mutation along branches.

Many standard tests have been applied to the field, with varying success. The nature of the data in phylogenetics (sparse sampling of state space) leads to bad performances of most of these tests. They reject the null hypothesis too often, and if they don't their power does not increase the trust in the null hypothesis. Attempts at increasing the power by using marginalised tests have proven promising.

Another more recent approach is to reduce the noise in the data by outlier detection and a subsequent modification of these outliers. While some methods suggest the removal of outliers and inferring a phylogeny from the reduced data (e.g., Goremykin et al. 2010), other methods suggest to accept the outliers as outcome of a second process acting on the data (e.g., Nguyen et al. 2011).

In this presentation I will compare the performance of these methods on different datum sets, discuss their merits and shortcomings, and suggest modifications and alternatives.

P-1362

Computing the joint distribution of tree shapes and tree distances for multiple tree inference with tree metric based models

Yujin Chung[1], Cécile Ané[1,2]
[1]Department of Statistics, University of Wisconsin - Madison, Madison, WI, USA, [2]Department of Botany, University of Wisconsin - Madison, Madison, WI, USA

Recombination events and other biological processes can cause the topologies of phylogenetic trees to be different for different genes. The dissimilarity among gene trees can be incorporated in a model to improve the accuracy of tree inference, when we seek to simultaneously detect recombination breakpoints along an alignment and infer phylogenetic trees of segments defined by the recombination breakpoints. Modeling the Robinson-Foulds (RF) distance between tree topologies of neighboring segments allows the detection of recombination breakpoints between short segments with similar tree topologies. When taking into account the RF distance between trees of neighboring segments, a major difficulty is the calculation of the "partition function", which works as a normalizing constant for the tree distribution. When the partition function is overlooked or miscalculated, an incorrect maximum likelihood estimate or an incorrect Bayesian posterior distribution may be obtained. Calculating the partition function in the naive way is computationally prohibitive.
We derive here an algorithm to calculate the partition function exactly, based on the calculation of the joint distribution of the tree shapes and of the RF distance between two random trees. We also propose approximations to the partition function, which are computationally fast and very accurate. Finally, we tie this work back to the problem of recombination breakpoint detection and tree reconstruction and its benefits to the development of correct MCMC Bayesian estimation.

P-1363

PopGenome: An R-library for large-scale population genetic analysis

Bastian Pfeifer, Ulrich Wittelsbürger, Martin Lercher
Institute for Computer Science Heinrich-Heine-University, Düsseldorf, Germany

Population genetics software comes in a wide range of implementations, written in various programming languages for different operating systems, and accepting diverse input formats. This diversity often hinders efficient research by the software users. At the same time, it does not aid theoreticians in the quick and effortless implementation of new population genetics tests. We have started to develop a new population genetics analysis and development package, named PopGenome, based on the powerful, open-source, statistical computing environment R. R is available for all major operating systems and has built-in high level scientific graphics capabilities. Its architecture allows easy management of large and complex datasets.

The first release of PopGenome includes, e.g., a wide range of polymorphism and neutrality statistics and F_{ST} estimates. It is linked to Hudson's MS program for significance tests using coalescent simulations. An integrated sliding window method can be used to scan genomic data.

We envision the open source PopGenome project to form a basic framework for the implementation of new methods by population geneticists worldwide, much as the BioConductor R project provides a framework for, e.g., new microarray analysis methods. We are currently implementing Bayesian methods as well as the data handling and analysis capabilities needed for genome-wide resequencing projects.

PopGenome is freely available under the GNU General Public License from our website *(http://www.cs.uni-duesseldorf.de/AG/BI/Software/PopGenome)*, and its components can be freely extended and reused.

P-1364

The 'State' of RNA Model Selection

James Allen, Simon Whelan
University of Manchester, Manchester, UK

Alignments of RNA, particularly ribosomal RNA, are frequently used to infer evolutionary trees, where the process of evolution is described by a substitution model, and a method of statistical inference is used to estimate the tree and the parameters associated with the model. In recent years the importance of non-coding RNA has become more evident, and has led to a renewed interest in understanding the selective forces acting during RNA evolution, which are derived from the base-pairing in stems that define the structure (and function) of many RNA molecules. Consequently, dinucleotide substitution models of varying levels of complexity have been developed to describe patterns of change in RNA stems. No statistically rigorous methods are currently available for comparing nucleotide and RNA models, due to the differing structure of the models.

Several studies have demonstrated that dinucleotide (4x4 = 16-state) stem substitution models provide a better statistical fit to RNA alignments than standard nucleotide (4-state) models. These studies typically, however, only test one or two RNA alignments and it is unclear whether their conclusions can be generalised. There are also difficulties choosing between RNA models with different numbers of states. A dinucleotide (16-state) model may be over-parameterised and lead to unreliable results. Simpler 7-state models also exist, where each of the 6 stable base pairs have a distinct state, and the other 10 nucleotide pairs are aggregated into a single mismatch state. Current methods for comparing these models rely on complex resampling methods, which are computationally slow and cannot be applied to large data sets.

In this study we develop fast and statistically rigorous methods for comparing all nucleotide and RNA models, and apply them to large numbers of mammalian RNA genes. First, we collect a large set of alignments of mammalian non-coding RNA, using manually-curated RNA sequences to retrieve appropriate sections from genomic alignments, which are then filtered to provide good quality datasets. We then develop a method, based on the results of Seo and Kishino (2008; 2009), to compare 7- and 16-state RNA models to each other and to 4-state nucleotide models, within a standard model selection framework. We use this approach to investigate model fit across our mammalian alignments. Finally, we assess the effect of model choice on downstream phylogenetic inference.

P-1365

p { margin-bottom: 0.21cm; } Methods to detect regional genomic variation in evolutionary history

Daniel Hartleb, Christian Eßer, Martin J. Lercher
Institute for Computer Science Heinrich-Heine-University Duesseldorf, Duesseldorf, Germany

Allo- and autopolyploidisations (also known as hybridizations and whole genome duplications) have strongly affected the evolution of plants. They seem necessary for the rapid invasion and contemporary dominance of angiosperms. After these polyploidisations, gene losses and chromosomal rearrangements have shaped the plant genome further. Due to hybridizations or reciprocal loss of duplicates, we expect different chromosomal regions to display distinct evolutionary histories. However, little is known about these effects. Here we show methods to detect such regional genomic variation.
Comparisons of gene trees are often complicated by the existence of gene duplicates. To solve this problem, we collapse each gene family tree by merging identical subtrees, thereby removing obvious duplications. Subsequently the collapsed tree has to fulfill two criteria: a) the ingroup (for which we suspect variation in evolutionary histories) must consist of exactly one gene of every species, b) no ingroup gene is allowed to be in the outgroup (which we assume to be stable, i.e., be located before the polyploidization). As individual gene trees may suffer from artefacts of tree reconstruction, we combine evidence from several neighbouring trees. To confirm that a locally dominant gene tree topology reflects the true evolutionary history, we correlate the topology to local synteny patterns. For this, we compare all genomic regions (e.g., 100-gene windows) in a reference genome with each other ingroup genome and identify for which species we find the highest synteny score. We find that local tree topology is correlated with expected synteny patterns, such that species closer on the locally dominant gene tree show stronger synteny patterns.

P-1366

How stable are protein domains in the course of evolution?

Tina Koestler[1,2], Bui Quang Minh[1,2], Arndt von Haeseler[1,2], Ingo Ebersberger[1,2]
[1]University of Vienna, Vienna, Austria, [2]Center for Integrative Bioinformatics Vienna, Vienna, Austria

Protein domains can be considered as evolutionary modules taking over individual functions within a protein. The characteristic features of these sequences, as summarized in profile Hidden Markov Models (pHMMs), are commonly used to search for new domain instances in yet uncharacterized proteins. However, it is unclear how many mutations can happen until a sequence loses its domain specific characteristics. To mimic the evolutionary process acting on a domain in a biologically realistic way, we have developed REvolver. REvolver simulates protein sequences under domain constraints (e.g. site-specific amino acid frequencies or the placement of insertions and deletions) that are extracted from a pHMM. With the help of this tool, we now assess the extent of evolutionary change a domain can bear before it is no longer detected by a standard pHMM search. Our analysis of 11,912 Pfam domains reveals three main types. Stable domains can be detected even after many mutations, thus over large evolutionary distances. Evolutionary unstable domains become undetectable already after few sequence changes. An interesting third type, are domains that repeatedly disappear and reappear in the course of simulated evolution. This temporary disappearance can be interpreted as a lack of sensitivity in the domain search. We now use this information to parameterize an evolutionary model of domain loss and gain. This serves as a null model to distinguish between cases where the domain is truly absent in a phylogenetic tree and cases where it is overlooked in the domain search.

P-1367

One step mutation matrices: A fresh view on modeling sequence evolution

Arndt von Haeseler[1], Minh Anh Thi Nguyen[1], Tanja Gesell[1], Steffen Klaere[2]
[1]CIBIV-MFPL, Vienna, Austria, [2]Dept Mathematics, Auckland, New Zealand

We will describe our recent development to model the process of DNA evolution. Contrary to classical approaches, that assume a rate matrix substitution scheme, we introduce the one step mutation (OSM) matrix, that describe the change of an alignment column (pattern) if an mutation is put on an arbitrary branch of the underlying tree. The applications of OSM-matrices are manifold.

Firstly, we use OSM-matrices to evaluate the goodness of fit of an evolutionary model together with the underlying tree. Contrary to standard statistical approaches that typically reject an evolutionary model, we suggest introduce a minimum number of “extra substitutions” on the inferred tree to provide a biologically motivated explanation why the alignment may deviate from expectation. We illustrate the method on several examples and give a survey about the goodness of fit of the selected models to the alignments in the PANDIT database.

Secondly, OSM-matrices are used to systematically add well-defined model violations to simulated data, generated along a tree and an evolutionary substitution model. The extra substitutions are placed on the tree and they disturb the otherwise perfect alignments. We then discuss the robustness of widely used tree reconstruction methods (maximum likelihood, maximum parsimony, neighbor joining). We show that under some conditions both maximum likelihood and maximum parsimony fail to reconstruct the underlying tree, whereas BIONJ still recovers the tree.

References

Klaere et al. (2008) Phil. Trans. Roy. Soc B 363:4041-4047

Nguyen et al. (2011) Mol Biol Evol 28 (1):143-152

Nguyen et al. (2012) Mol Biol Evol 29 (2): 663-673

P-1368

Model selection in EnsemblCompara GeneTrees

Mateus Patricio[1], Matthieu Muffato[2], Javier Herrero Sanchez[2], David Posada[1]
[1]University of Vigo, Vigo, Spain,[2]EMBL-EBI European Bioinformatics Institute, Cambridge, UK

EnsemblCompara GeneTrees is a phylogenetic reconstruction resource developed to better understand the evolutionary history of gene families distinguishing both duplication and speciation events. This resource uses a pipeline where the method *TreeBest* from TreeFam plays a central role in the generation of maximum likelihood (ML) family gene trees, using predefined models of nucleotide substitution or amino acid replacement –HKY for DNA and WAG for proteins– regardless to their fit to the data. Because computer simulations have shown that best-fit models can provide more accurate estimates of topology and especially branch lengths, we have introduced model selection procedures in EnsemblCompara GeneTrees. With this we hope not also to obtain better trees but also to obtain a new annotation for each alignment, the best-fit model and estimated model parameters.

After implementing a pipeline to estimate ML parameters, best-fit models and model-averaged phylogenies with jModeltest (DNA data) and ProtTest (protein data), we analyzed all the gene family alignments available from EnsemblCompara GeneTrees. A clear model preference trend was observed; for DNA and protein data the GTR and JTT models were selected ~30% and 65% of the time, respectively. We then used the KH-SH tests to measure the statistical differences between the current gene trees estimated with pre-specified models and the new gene trees estimated with the best-fit models. Remarkably around 20% of the new trees were statistically different from their counterparts. In addition, using best-fit models results in a decreased number of inferred duplication/losses. Finally, we also observed some correlations in the data, where topological differences among trees increased with alignment length but decreased with increasing number of taxa.

P-1369

Evolutionary impact of transcription factor binding

Clemens Lakner, Eric A. Stone, Jeffrey L. Thorne
North Carolina State University, Raleigh, NC, USA

Statistical inference methods that accurately account for the balance between evolutionary processes allow us to detect the footprint of natural selection and estimate its effects. Rates of evolution between genotypes depend on the mutation rate and the probability with which new genotypes will replace all existing variants in a population. The latter probability is contingent on the fitness difference between the sequences in question: mutations that increase an organism's relative fitness are more likely to replace less fit variants.

Current models of molecular evolution that explictly employ this connection between genotype and fitness rely on computational biology to predict phenotype from genotype (e.g. effects of mutations on protein structure). Fitness differences are subsequently taken to be a simple function of the predicted phenotypic effects of mutations.

Eukaryotic transcription factors bind to relatively short DNA motifs. Recently, in-vitro binding affinities to all possible DNA oligomers of length k have become available for a number of transcription factors and species. This constitutes a complete representation of a genotype-to-phenotype map. Using these binding affinities, our motivation was the possibility of detecting and characterizing weak selection against the appearance of transcription factor binding motifs in putatively non-functional regions of genomes.

We were particularly interested in how natural selection affects binding motifs for transcription factors that are involved in chromatin remodeling. In this talk we will summarize our approach and the results that we have obtained.

P-1370

Evaluating epistatic models of protein evolution based on comparative genomics data

Daniel Jordan[1,2], Ivan Adzhubey[1], Shamil Sunyaev[1]
[1]*Div of Genet, Dept of Med, Brigham & Women's Hosp and Harvard Med School, Boston, Massachusetts, USA,*
[2]*Program in Biophysics, Harvard University, Cambridge, Massachusetts, USA*

Mutations that cause disease with high penetrance in humans are known to occasionally reach fixation in other species. This observation has interesting implications for both medical genetics and models of protein evolution. Fixation of human disease mutations has previously been interpreted as evidence of compensatory amino acid changes, such as those involved in Dobzhansky-Muller incompatibility. We analyzed a large dataset of well-annotated mutations involved in Mendelian diseases using new comparative genomics data on 46 vertebrate species. We observed that over 8% of mutations correspond to wild type amino acids in vertebrate species. We tested whether these cases can be explained by compensatory effects of other amino acid changes in these species. For a large fraction of observations, the presence of a human disease variant in vertebrate orthologs is accompanied by synchronous changes at many other amino acid sites. This observation is limited to disease mutations and is in sharp contrast to human neutral sequence variants. We show that a simple statistic that counts co-evolving sites can differentiate between neutral and pathogenic mutations that are observed as wild-type amino acids in vertebrate orthologs. We propose two models to explain the success of this statistic. In one model, the pathogenic variant becomes neutral due to a small number of true compensatory changes within the larger set of co-evolving sites; in the other, the pathogenic variant becomes neutral due to a global change in the configuration of the protein. We present simulations of each model and discuss their implications for evolution and population genetics, quantifying the inaccuracy of models that assume that multiple sites evolve independently in a constant evolutionary landscape. We conclude that an epistatic model is essential for understanding protein evolution. We also conclude that using a co-evolution statistic like that reported here can greatly increase accuracy of computational predictions of the pathogenic effect of human missense mutations based on comparative genomics data.

P-1371

Combining protein structure and sequence variation to identify functionally important sites and sites under positive or negative selection

Austin Meyer, Claus Wilke
The University of Texas at Austin, Austin, TX, USA

Conventional approaches to identify sites under positive selection aim to identify sites with ω=dN/dS significantly larger than one. Maximum-likelihood models are fit to sequence alignments, and ω values are estimated for each site using either fixed-effects or random-effects methods. One limitation of these approaches is that they don't provide a baseline of what kind of ω values should be expected for a given site. For example, a site with ω=0.9 would not be identified as being under positive selection, yet 0.9 might be unusually high (and a sign of positive selection) if the baseline expectation for this site in this protein was 0.1. Likewise, sites with particularly low ω (indicative of negative selection and likely functional importance) cannot be identified at all without a baseline expecation.

Here, we develop maximum-likelihood models that can provide a baseline expecation for ω and can identify sites that significantly deviate from this baseline. Our method is based on the observation that the evolutionary conservation of a site is correlated with the site's relative solvent accessibility (RSA, a measure of solvent exposure of the focal amino acid in the folded, 3-dimensional protein structure). In our models, ω is described by linear functions of RSA. We use a model-fit criterion to identify the optimal number of linear functions required to describe all sites in a protein, and we identify for each site to which linear function it belongs.

We apply our method to several viral proteins, including influenza hemagglutinin and neuraminidase, and HIV reverse transcriptase. We find that models in which ω is RSA dependent always provide a better model fit than conventional, RSA-independent models. Further, we find that the number of different linear functions needed to describe typical viral proteins is small, on the order of 6-10. In general, most sites in a protein fall onto 2-3 linear functions, which can be considered baseline. Sites that don't fall onto the baseline are either positively or negatively selected, depending on whether their ω falls above or below baseline. For the viral proteins we studied, these off-baseline sites are enriched in sites of known function, including hemagglutinin cleavage sites and sites known to be important for the evolution of tamiflu resistance. Our method is easily implemented and broadly applicable to a wide range of scenarios, as long as a crystal structure is available for the protein of interest.

P-1372

Evolution of clonal populations approaching a fitness peak.

Isabel Gordo[1], Paulo Campos[2]
[1]*Instituto Gulbenkian de Ciência, Oeiras, Lisbon, Portugal,* [2]*Universidade Federal Rural de Pernambuco, Recife, Pernambuco, Brazil*

When a population faces a new environment it is expected to evolve through the accumulation of adaptive substitutions. The dynamics of such adaptation depends on the fitness landscape and possibly on the genetic background on which new advantageous mutations arise. Here we study theoretically the dynamics of adaptive evolution at the phenotypic and genotypic levels. We focus on a Fisherian landscape characterized by a single fitness peak. Motivated by major evolutionary patterns of asexual populations observed in laboratory conditions, we find that Fisher's model of adaptation, extended to allow for small random environmental variations, is able to explain these patterns. Consistent with the observations of populations evolving under controlled conditions, the model predicts that: mean population fitness increases rapidly when populations face novel environments and then achieves a dynamic plateau; the rate of molecular evolution is remarkably constant over long periods of evolution; mutators are expected to invade and patterns of epistasis vary along the adaptive walk. Negative epistasis is expected in the initial steps of adaptation but not at later steps, a prediction that remains to be tested. Furthermore, populations are expected to exhibit high levels of phenotypic diversity at all instances of their evolution. This implies that populations are possibly able to adapt rapidly to novel abiotic environments.

P-1373

What is evolutionary time in a non-stationary substitution process?

Von Bing Yap[1], Gavin Huttley[2]
[1]National University of Singapore, Singapore, Singapore, [2]Australian National University, Canberra, Australia

The idea of evolutionary time goes back to the PAM matrices for comparison of amino acid sequences (Dayhoff et. al, 1960's). For a group of extant taxa, the chronological phylogeny not surprisingly is clocklike: the distance from the root to any taxon is constant. However, it is customary to measure branch lengths in evolutionary time, so that lineages with different evolution rates can be picked out at a glance. In particular, an evolutionary phylogeny with the clocklike property suggests a molecular clock. For a stationary substitution process, the relationship between chronological time and evolutionary time is linear, where the proportionality constant is easily computed from the rate matrix. For a non-stationary process, the corresponding relationship is rather complicated, and we need to go back to the basics: the expected number of substitutions per site in a given chronological time interval. In this presentation, we introduce a method to approximately calculate this quantity, and demonstrate its relevance to a definition of evolutionary time for a large class of substitution processes, which generalises the existing definition for stationary processes. Its utility in presenting and interpreting non-stationary evolutionary phylogenies will be described with real examples.

P-1374

TAKING A MORE DETAILED LOOK AT MODEL ADEQUACY: RESIDUAL DIAGNOSTICS FOR PHYLOGENETICS

Sandra Meid[1], Bernhard Misof[1], Christoph Mayer[1], Barbara Holland[0,2]
[1]ZFMK, zmb, Bonn, NRW, Germany,[2]UTAS, Hobart, TAS, Australia

Maximum Likelihood methods are known to be consistent. However, the method accuracy depends strongly on the quality of the multiple sequence alignment and the ability of the selected evolution model to reflect the underlying historical processes. Several methods have been established to check which model covers the data and others to test how well a model fits. In case a substitution model is not adequate for the dataset, there is no way to achieve consistent and reliable results.
In case that a model fitting to a real data set is correct, split spectra should be similar to the ones obtained for data simulated based on the identical tree topology and identical model and model parameters (parametric bootstrapping). We will show, that for a lot of datasets this is not the case. The detected over- or underrepresented splits of the simulated data sets give some hints in which way chosen models may misfit the underlying evolutionary models.

P-1375

IQ-TREE: An adaptive and efficient maximum-likelihood tree inference package

Bui Quang Minh, Lam Tung Nguyen, Heiko Schmidt, Arndt von Haeseler
Center for Integrative Bioinformatics, Vienna, Austria

IQPNNI (Important Quartet Puzzling with Nearest Neighbor Interchange) is a heuristic method to efficiently reconstruct large phylogenetic trees using the likehood principle. It is based on the fact that the fast NNI method, introduced in PhyML, may get stuck in local optimal trees. To avoid such local optima, one can apply more general tree-rearrangement methods like subtree pruning and regrafting, which is employed in RAxML and the new PhyML. IQPNNI uses a different strategy; Here the current best tree is perturbed by a quartet-puzzling type algorithm and further optimized by the fast NNI. The new tree is accepted if it shows higher likelihood. We repeat this process for a number of iterations or until the stopping criterion is met.

Here we present IQ-TREE, a complete reimplementation of the IQPNNI method with substantial improvements with respect to computing time and model flexibility. We achieve better performance by a so-called adaptive NNI, which predicts if subsequent NNIs will improve the current tree or not. First, we observe that if the tree likelihood is significantly higher than the likelihood of the same tree where one branch length collapses to zero, the NNI on such branch will likely not improve the tree. This is similar to the likelihood ratio test but we estimate the cutoff in the log-likehood difference based on the data instead of using the chi-square distribution. Hence, we will not evaluate the NNI for such branches. Second, we predict the maximally achievable tree likelihood during each IQPNNI iteration based on the observed likelihood improvements per NNI and the number of NNIs per iteration. If such predicted likelihood upper-bound is still lower than the highest likelihood obtained, we interrupt the current iteration. That way, the adaptive NNI helps to avoid non-promissing NNIs.

The performance analysis shows that the adaptive NNI reduces the average running time by a factor of 5 while maintaining the accuracy of the IQPNNI algorithm. Moreover, the computation of the likelihood function is optimized by the SSE instructions that results in a further speedup of 2. Overall, IQ-TREE is 10 times faster than IQPNNI. IQ-TREE also compares favourably with RAxML on protein alignments. In addition to most features in IQPNNI, IQ-TREE now provides a partition model with suitable data structures for efficiently analyzing phylogenomic data. Finally, IQ-TREE includes a test of model homogeneity assumption along the tree.

P-1376

Toward identifying Bateson-Dobzhansky-Muller incompatibilities in yeast: A simulation study

Chuan Li[1], Zhi Wang[2,1], Jianzhi Zhang[1]
[1]*University of Michigan, Ann Arbor, MI, USA,* [2]*Sage Bionetworks, Seattle, WA, USA*

The Bateson-Dobzhansky-Muller (BDM) model of reproductive isolation by genic incompatibility is a widely accepted model for speciation. Despite repeated efforts, BDM incompatibilities between nuclear genes have never been identified between the budding yeast *Saccharomyces cerevisiae*, a genetic model organism, and its sister species *S. paradoxus*. Such negative results have led to the hypothesis that simple nuclear BDM incompatibilities do not exist in yeast. Here we explore an alternative explanation that simple BDM incompatibilities exist but were undetectable due to limited statistical power. We found that previously employed statistical methods were not ideal and a redesigned method improves the statistical power. Furthermore, we determined the relationship between the required sample size and the probability of identifying a BDM incompatibility of a given effect size. Our results show that the previous failure to detect BDM incompatibilities is likely caused by the limited sample sizes and/or inappropriate statistical methods. This finding should reopen the experimental search for nuclear BDM incompatibilities in yeast. Given the exceptionally rich biological information about yeast, the identification of BDM incompatibilities would greatly deepen our understanding of the genetic basis of reproductive isolation and speciation. Because the BDM incompatibility is a form of epistasis, the improved methodology can also help identify epistasis in general.

P-1377

A novel method for measuring codon usage bias and estimating its statistical significance

Zhang Zhang
CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

Codon usage bias is generally believed to be a combined outcome of mutation pressure, natural selection, and genetic drift. Thus, accurate estimation of codon usage bias is of fundamental importance in improving our knowledge on gene function and evolution. However, extant measures have not fully considered background nucleotide composition when estimating codon usage bias and have not statistically evaluated the significance of codon usage bias. Here we devise a novel measure-Codon Deviation Coefficient (CDC)-that measures codon usage bias and estimates its statistical significance without requiring any prior knowledge. Unlike extant measures, CDC measures codon usage bias by accounting for specified background nucleotide composition in any given sequence and employs the bootstrapping to assess the statistical significance of codon usage bias. We test the performance of CDC by comparison with extant measures on simulated sequences and empirical data. Our results show that CDC is superior to extant measures by achieving a more informative estimation of codon usage bias and its statistical significance. Therefore, CDC is a highly informative measure of codon usage bias, useful for determining comparative magnitudes and patterns of biased codon usage for genes or genomes with diverse sequence compositions.

P-1378

Agrisims: Simulating crop domestication using individual based modeling

James Kitchen[1], Dorian Fuller[2], Terry Brown[3], Robin Allaby[1]
[1]School of Life Sciences, University of Warwick, Coventry, UK,[2]Institute of Archaeology, University College London, London, UK,[3]Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, UK

Since cultivation practices began, our crops have acquired a number of traits making them better suited to farming processes: these traits include the tough rachis (non-shattering) phenotype and larger seed sizes. This process is known as domestication. We aim to better understand the process of domestication and therefore seek to determine which factors have given rise to patterns of domestication that are consistent with those in the archaeological record. To this end we have developed a spatial individual based simulation model, Agrisims. Simulated individuals contain gene networks that interact with the surrounding environment: through mutation and gene flow these networks allow individuals to become better adapted to their environment. We first tested the Agrisims model to check that it behaved in expected ways such as achieving Hardy-Weinberg equilibrium, and gained some insight into spatial population dynamics. Agrisims was then used to study the spread of the tough rachis mutant in a regularly harvested crop field surrounded by wild plants: here we have modeled the effects of varying the proportion of dispersed seeds that are predated, the proportion of harvested seeds that are sown for the next year and the proportion of seeds that are retained on non-mutant crops. We have also varied the amounts of gene flow between the surrounding wild plants and the field crops. Higher proportions of dispersed seeds predated and harvested seeds sown increases selection for the tough rachis mutant, whereas more seeds retained by non-mutants (earlier harvesting) decreases selection. Moderate to high levels of gene flow from surrounding wild plants allows the mutant to reach high levels with the field (>90%), yet plateau and never fixate. We have used approximate Bayesian computation to assess which combinations of these parameters may have given rise to summary data from the archaeological record.

P-1379

A divide and concatenate strategy for the phylogenetic reconstruction of large orthologous datasets

Jia-Ming Chang[1,2], Matthieu Muffato[2], Paolo Di Tommaso[1], Jean-Francois Taly[1], Javier Herrero[2], Cedric Notredame[0]
[1]Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain,[2]European Bioinformatics Institute, Hinxton, UK

Thanks to next-generation sequencing techniques, more and more sequences have become available. Reconstructing the phylogenetic tree of related sequences is a regular analysis of NGS data. Moreover, this overwhelming amount of data is challenging even for the fastest methods and the most accurate ones, like Maximum Likelihood, can not deal with some important multi-genetic families like olfactory receptors.

Here we show how a simple Divide and Concatenate (D&C) strategy can be applied to this issue by breaking it down in smaller independent problems. Our approach relies on the ability to identify clusters of orthologous (CO) genes within a large dataset. For each CO, a phylogenetic tree is build using a typical approach (M-Coffee + TreeBest). The upper level of the tree (super-tree) is resolved in a second stage. One protein per species is chosen from each subtree. All proteins from the same species are aligned together. The alignment used for building the super-tree results from concatenating all these alignments, where within-species paralogues appear in the same columns and orthologues appear in the same row. The advantage is that we reduce the number of sequences to classify without losing information as all sequences are represented in the final alignment. This approach can easily deal with lineage specific duplications, but not with lateral transfers. It is therefore better suited for the analysis of eukaryotic large families like the kinases or the olfactory receptors.

We applied this approach for classifying 858 olfactory receptors from 13 Drosophila species. We could reconstruct the tree of all these sequences (403 amino acids long on average) in roughly 67 minutes in a normal workstation. The log likelihood of the final tree by applying traditional neighbor joining in D&C strategy is -517,952.30, superior to -575,824.05 with the single neighbor joining method. The web service is available in http://tcoffee.crg.cat/apps/bigtree/.

P-1380

PartitionFinder and PartitionFinderProtein - combined selection of partitioning schemes and substitution models for phylogenetic analyses.

Robert Lanfear[1], Brett Calcott[1], Simon Ho[2], Stephane Guindon[3]
[1]Australian National University, Canberra, ACT, Australia,[2]Sydney University, Sydney, NSW, Australia,[3]University of Auckland, Auckland, New Zealand

This poster describes the use and utility of two programs for objectively choosing partitioning schemes and substitution models for phylogenetic analyses. In phylogenetic analyses of molecular sequence data, partitioning involves estimating independent models of molecular evolution for different sets of sites in a sequence alignment. Choosing an appropriate partitioning scheme is an important step in many analyses because it can affect the accuracy of phylogenetic reconstruction. Despite this, partitioning schemes are often chosen without explicit statistical justification. Here, we describe two new programs for the combined selection of best-fit partitioning schemes and substitution models for nucleotide and amino acid alignments. These programs combine new methods and algorithms with multi-processor analyses to allow millions of partitioning schemes to be automatically compared in realistic timeframes. They allow users to quickly, easily, and objectively select best-fit partitioning schemes and substitution models even for large multilocus datasets. We demonstrate both programs on a large range of datasets, and show that they significantly outperform previous approaches, including the ad-hoc selection of partitioning schemes (e.g. partitioning by gene or codon position), and other recently proposed objective methods. The programs are user friendly and open source, and will run on Windows and Mac computers. Both programs have a flexible range of options, allowing users to select partitioning schemes and substitution models using a range of information theoretic metrics (e.g. the BIC, AIC, and AICc). The programs, source code, and a detailed manual are freely available from www.robertlanfear.com/partitionfinder.

P-2000

Can we understand bacterial evolution from MLST data?

Ferran Palero[1,2], Fernando Gonzalez-Candelas[1,2]
[1]University of Valencia, Valencia, Spain,[2]CSISP, Valencia, Spain

Molecular techniques have only recently started to unveil genetic diversity levels in bacterial species. Since its proposal by Maiden et al. (1998), Multilocus Sequence Typing (MLST) is the most generalised way for assessing molecular diversity of bacterial populations. Under the MLST scheme, isolates are characterized using the sequences of internal fragments (approx. 450-500 bp) of six or seven house-keeping genes. Nucleotide differences between alleles are usually ignored and each isolate of a species is characterised by an array of integers that correspond to the alleles at the loci used in typing. A caveat that pervades most studies on bacterial population genetics is to which extent the MLST approach is providing an accurate representation of bacterial genetic diversity. The present work aims at defining how representative are the genetic diversity levels provided by the MLST house-keeping loci as compared with the global diversity levels provided by complete genome sequences.

Several public databases containing complete sequences of bacterial genomes and MLST isolate information were accessed using a computer pipeline written in Perl and Mathematica. Only bacterial species for which MLST schemes have been developed previously and for which more than five genome sequences are available were included in this study. The "core genome" for each species, containing genes present in all strains within a species, was analysed. Several summary statistics of nucleotide polymorphism levels were estimated per gene, such as: the total number of segregating sites (S), the population mutational parameter (θ), and the nucleotide diversity (π). Moreover, statistics for the neutrality-tests of Tajima (D) and Fu and Li (D* and F*) were calculated for each gene along the genome.

Our results show that genetic diversity estimates vary considerably along the genome of bacterial species. Interestingly, different levels of skewness for the distribution of diversity estimates and neutrality-test summary statistics were observed among species. In some cases, house-keeping genes showed higher genetic diversity levels than the average and presented neutrality-tests statistics far from zero, but the relative distance of the MLST loci to the average value over the genome varied depending on the species. An overestimate of diversity levels could result from a bias towards polymorphic markers for typing. MLST genes with extremely high and positive Tajima's D values were found in B. cereus, H. pylori and C. jejuni.

P-2001

Evolution of Argonaute genes in Drosophila obscura

Samuel Lewis, Darren Obbard
The University of Edinburgh, Edinburgh, UK

RNAi is a major invertebrate immune pathway, in which small RNAs are derived from a target RNA and guide an Argonaute-family protein to cleave and subsequently degrade the target. This mechanism defends against both viruses and transposable elements, the latter thought to be especially costly in the germline. These varied and sometimes conflicting selective pressures have driven rapid adaptive evolution in Argonaute family genes, and may have led to the expansion of the Argonaute family in *Aedes aegypti* and *Culex pipiens*, as well as the subfunctionalization of Piwi genes recently documented in the pea aphid *Acyrthosiphon pisum*. This rapid evolution is evidenced by Argonautes being among the top 3% of fastest evolving *Drosophila* genes, potentially driven by an arms race between viral suppressors of RNAi (VSRs) and the RNAi mechanism.

In *Drosophila* the Argonaute 2 (Ago2) gene functions in siRNA-mediated antiviral defence and siRNA-mediated anti-TE defence (including when the genome is colonized by a new TE). *Drosophila* Ago2 is represented by a single 1:1 ortholog in most insect genomes, but in some lineages it has undergone duplication events, for example two duplicates are found in *C. pipiens* and *D. willistoni*. Ago2 has also been duplicated numerous times in the *obscura* group of *Drosophila*, producing at least four copies in *D. pseudoobscura* and *D. persimillis*. Combined with rapid adaptive evolution and recent selective sweeps in the Argonaute family, and the possibility of both antiviral and anti-TE roles, these duplication events make diversification and specialization in *obscura* Ago2 worthy of investigation.

Here we present a large-scale gene tree for Piwi and Argonaute-family genes in insects. We make use of publically available genome sequences, supplemented by targeted PCR and novel transcriptomic data, to elucidate the evolutionary history of Argonaute duplications in *Drosophila*, focusing particularly on the multiple duplications in the *Drosophila pseudoobscura* subgroup. We find that at least one duplicate in *D. pseudoobscura* predates the most recent common ancestor of *D. pseudoobscura* and *D. affinis*, and that the rate of protein evolution remains high in all paralogs.

P-2002

Diversity of endogenous retroviruses in the saltwater crocodile (Crocodylus porosus) in the Northern Territory of Australia

Amanda Y. Chong[1], Sarah Atkinson[1], Sally Isberg[1,2], Lorna Melville[3], Jaime Gongora[1]
[1]University of Sydney, Faculty of Veterinary Science, Sydney, NSW, Australia,[2]Porosus Pty Ltd, Palmerston, NT, Australia,[3]OIC Berrimah Veterinary Laboratories Department of Resources, Darwin, NT, Australia

The saltwater crocodile (Crocodylus porosus) is one of two species of crocodile found in Australia and the only one that is commercially farmed. It has been proposed that retroviruses may play a role in the high incidence of runtism seen in farmed saltwater crocodile hatchlings [1]. Endogenous retroviruses (ERVs) are remnants of exogenous retroviruses that have integrated into the DNA of a germ-line cell and are therefore inheritable [2]. Here we investigate the distribution of ERVs of the saltwater crocodile from the Northern Territory of Australia compared with ERVs from other species of Crocodylia (Crocodilians; alligators, caimans, gharials and crocodiles).

This study has identified two major clades of ERVs in the saltwater crocodile. This is consistent with previous investigations which have revealed two major lineages of ERVs, one of which appears to specific to the family Crocodylidae (CERV1), and another which is common among Crocodilian species (CERV2) [3]. We have isolated a number of potentially functional fragments that appear to belong to the CERV1 clade. These are of particular interest due to their relevance to the assessment of expression of ERVs and their role in diseases. Overall, sequences isolated in this study show a large amount of genetic variation, but very little phylogenetic resolution. Multiple ERV integration events have occurred throughout the evolutionary history of crocodilians. Of the two major lineages, the CERV2 lineage is likely to be an older integration event predating the speciation of modern day Crocodilians. On the other hand, CERV1 appears to be a more recent integration. In addition to this, we have identified novel sequences that appear to be related to other retroviral genera.

1. Isberg, S., C. Shilton, and P. Thomson, Improving Australia's crocodile industry productivity: Understanding runtism and survival, 2009, Rural Industries Research and Development Corporation Union Offset Printing Canberra.
2. Bishop, J.M., Retroviruses. Annual Review of Biochemistry, 1978. 47: p. 35-88.
3. Jaratlerdsiri, W., et al., Distribution of endogenous retroviruses in crocodilians. Journal of Virology, 2009. 83(19): p. 10305-8.

P-2003

Widespread interspecific divergence in *cis*-regulation of transposable elements in the *Arabidopsis* genus

Fei He[1], Justin Borevitz[2], Juliette de Meaux[1]
[1]IEB, Münster, Germany, [2]University of Chicago, Chicago, USA

Transposable elements (TEs) are so abundant and variable that they count among the most important mutational sources in genomes. Nonetheless, little is known about the genetics of their variation in activity or silencing across closely related species. Here, we demonstrate that regulation of transposable element genes can differ dramatically between the two closely related *Arabidopsis* species *A. thaliana* and *A. lyrata*. In leaf and floral tissues of F1 interspecific hybrids, about 47% of TEs show allele-specific expression, with the *A. lyrata* copy being generally expressed at higher level. We confirm that TEs are generally expressed in *A. lyrata* but not in *A. thaliana*. Allele-specific differences in TE expression are associated with divergence in epigenetic modifications like DNA and histone methylation between species as well as with sequence divergence. Our data demonstrates that *A. thaliana* silences TEs much better than *A. lyrata*. For LTR retrotransposons, these differences are more pronounced for younger insertions. Interspecific differences in TE silencing may have a great impact on genome size changes.

P-2004

Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila melanogaster

Robert Kofler, Andrea Betancourt, Christian Schlötterer
Institute for Population Genetics, Vienna, Austria

Transposable elements (TEs) are mobile genetic elements that parasitize genomes by semi-autonomously increasing their own copy number within the host genome. While TEs are important for genome evolution, appropriate methods for performing unbiased genome-wide surveys of TE variation in natural populations have been lacking. Here, we describe a novel and cost-effective approach for estimating population frequencies of TE insertions using paired-end Illumina reads from a pooled population sample. Importantly, the method treats insertions present in and absent from the reference genome identically, allowing unbiased TE population frequency estimates. We apply this method to data from a natural *Drosophila melanogaster* population from Portugal. Consistent with previous reports, we show that low recombining genomic regions harbor more TE insertions and maintain insertions at higher frequencies than do high recombining regions. We conservatively estimate that there are almost twice as many “novel” TE insertion sites as sites known from the reference sequence in our population sample (6,824 novel versus 3,639 reference sites, with on average a 31-fold coverage per insertion site). Different families of transposable elements show large differences in their insertion densities and population frequencies. Our analyses suggest that the history of TE activity significantly contributes to this pattern, with recently active families segregating at lower frequencies than those active in the more distant past. Finally, using our high-resolution TE abundance measurements, we identified 13 candidate positively selected TE insertions based on their high population frequencies and on low Tajima's *D* values in their neighborhoods.

P-2005

Comprehensive population genetic modelling unravels the invasion history of four *Pristionchus pacificus* lineages on La Réunion Island

Angela McGaughran, Katy Morgan, Ralf Sommer
Max Planck Institute for Developmental Biology, Department for Evoutionary Biology, Tuebingen, Germany

Nematodes are a ubiquitous animal group, highly successful in terms of both biodiversity and abundance. In the globally widespread hermaphrodite *Pristionchus pacificus*, this success is likely enhanced as self-fertilisation aids establishment and high genetic diversity increases adaptive potential. Here, we examine the demographic and evolutionary features influencing establishment success following invasion in four lineages of *P. pacificus* present on young volcanic La Réunion Island.
Our approach exploits a model system that includes fully developed genetic tools, a sequenced genome, population genetics and ecological knowledge, to examine species invasion using approximate Bayesian computation (ABC) and complementary demographic and divergence analyses with both mitochondrial and microsatellite data.
We find that establishment of the four island lineages occurred via at least four independent invasion events. We identify the most likely order of invasion and find that divergence in the 'metapopulation' occurred early (TMRCA = ~500,000 generations) in the evolutionary history of *P. pacificus*. All four island lineages are shown to have undergone demographic and spatial expansion following foundation, with the dated expansion signals corresponding well with the islands volcanic history.
Our comprehensive analyses highlight *P. pacificus* as a successful island colonist and demonstrate how differences in the timing and strength of evolutionary events can result in unique patterns among lineages. The special nature of our system will enable future investigation of the factors that influence invasion success and adaptation in different environments, providing further insight into evolution in an island setting.

P-2006

Recent and contrasting evolutionary change in human and chimpanzee bone phenotypes: primate population genetics of type I collagen (*COL1A1*)

Daryn Stover[2,1], Anne Stone[1], Brian Verrelli[1]
[1]*Arizona State University, Tempe, AZ, USA,* [2]*Simon Fraser University, Burnaby, BC, Canada*

The most abundant structural protein in vertebrates, type I collagen, is encoded in part by the *COL1A1* gene, which harbors >600 mutations linked to human skeletal and connective-tissue diseases like osteoporosis. Our previous comparative species analyses showed that not only has the COL1A1 protein been impacted by varying selective constraint across the amino acid sequence over ~450 My of vertebrate evolution, but that noncoding regions show surprisingly strong evidence of functional constraint across species. In addition, our population genetic analyses of a global human sample reveals that amino acid variation is slightly deleterious today, but also an unusual haplotype structure for noncoding regions that is associated with observed geographic differences in bone strength. To determine whether these varying selective pressures are unique to humans, here we conduct an analysis of the ~17-kb *COL1A1* locus in 40 chromosome sequences from the chimpanzee *Pan troglodytes verus*, our closest-living relative. While hundreds of bone disease genotyping studies have been performed in our species, this is the first to characterize population variation and inferences of selection in a non-human primate. Although we find no amino acid variation, there is an unusual haplotype structure in chimpanzees associated with noncoding regions that is characterized by an excess of intermediate frequency polymorphism segregating between two haplogroups. Most surprisingly, although COL1A1 protein length variants are extremely rare and highly deleterious in humans, we have also discovered a partial exon duplication at 18% frequency in chimpanzees. An examination of long-range LD of flanking regions spanning ~180-kb of the *COL1A1* chromosomal region in chimpanzees and bonobos suggests an ancient age (>2 My) for the main *COL1A1* haplogroups, predating the chimpanzee-bonobo divergence. Additionally, we find evidence of long-range haplotype structure in association with the exon duplication, which shows evidence of having risen to high frequency recently and rapidly. The potential for adaptive exon duplication in chimpanzees offers a unique model with which to study bone-related protein diversity, given its rare occurrence in humans. In conclusion, we find that humans, chimpanzees, and bonobos have contrasting evolutionary signatures for this ubiquitously expressed gene, which suggests that functional differences have accumulated recently within primates for a phenotype that shows a long history of high constraint in vertebrates overall.

P-2007

Matching minor allele frequency rectifies confounding effects of purifying selection on population differentiation measures

Takahiro Maruki[1], Jesse Taylor[2]
[1]*Indiana University, Bloomington, USA,* [2]*Arizona State University, Tempe, USA*

Quantifying population differentiation is important for understanding the mechanisms of evolution. Researchers frequently use Wright's F_{ST} and its analogues for this purpose. Their main applications include identification of putative targets of local adaptation and estimation of the amount of gene flow between populations. Most theoretical studies of these measures assume neutral evolution and very few of them have investigated the effect of purifying selection on population differentiation measures. However, many empirical studies report evidence of widespread purifying selection, especially at functionally important positions (loci). Furthermore, some recent studies have reported that purifying selection significantly influences population differentiation measures at SNP sites. In this study, the effect of purifying selection on population differentiation measures at bi-allelic loci is investigated. We show that purifying selection intensity significantly affects population differentiation measures mainly through modulating minor allele frequencies across populations. This effect is found to be profound when the migration rate between populations is small. We also show that the confounding effects of purifying selection on measures of population differentiation can in part be controlled by conditioning on the minor allele frequency at each site.

P-2008

Estimated demographic parameters of modern human migration to the New World by whole mtDNA sequences of non-admixture Mesoamericans

Jun Gojobori[1], Shintaroh Ueda[2]
[1]*Graduate School for Advanced Studies, Hayama, Kanagawa, Japan,* [2]*The University of Tokyo, Bunkyo-ku, Tokyo, Japan*

It is widely accepted that the Anatomical Modern Humans (*Homo Sapiens*) colonized New World continent from Asia through Beringia at the ice age. It is under debate on when or how large this migration event was. We sequenced the whole mtDNA genome from individuals of two populations of Mesoamerican, Zapotec and Mazahua. They are said not to be admixtures with Europeans. Haplogroups A, B, C and D are found in these populations. We employed Bayesian Skyline Plot to estimate the demographic history of these populations. The result shows that the population expansion started about 20,000 years ago, at the end of Last Glacial Maximum. Our result also shows that the initial female effective population size was 1,500 ~ 1,800 and it expanded to 18,000 ~ 21,000. We further found that the estimated coalescent times of the four haplogroups are almost the same. These findings support the “single migration hypothesis” for the peopling to the New World.

P-2009

Jaatha, a fast method to infer demographic parameters - not only for wild tomatoes

Lisha Mathew[1], Laura E. Rose[2], Dirk Metzler[1]
[1]Ludwig-Maximilians-Universität, Munich, Germany,[2]Heinrich-Heine-Universität, Duesseldorf, Germany

Demography can leave similar signatures in the genome as natural selection. Therefore before we can study the effect of natural selection on genome evolution we have to account for the species' demography. Here we focus on elucidating the demography of a pair of closely related species.

We present Jaatha (**J**oint site frequency **a**ssociated **a**pproximation of **th**e **a**ncestry, Malayalam word for 'past') which is a fast composite-likelihood method to estimate parameters of a divergence model based on single nucleotide polymorphism data. Examples for model parameters are divergence time, population sizes, size changes, and migration rates. In contrast to most previously available software packages, Jaatha allows for recent divergence and intralocus recombination, and can be applied to any two-species divergence model. Furthermore, Jaatha's results can be obtained within a few of hours while some methods take several weeks to converge. The accuracy of Jaatha is comparable to that of other methods (IM and ∂a∂i), except for the estimation of divergence time where Jaatha is superior.

Jaatha enables us now to estimate demographic parameters of the two recently diverged South American wild tomatoes *Solanum chilense* and *S. peruvianum* which show high levels of intralocus recombination. Our results indicate that *S. peruvianum* underwent a size expansion and that a significant amount of gene flow continued after the split of the two species. The estimated divergence time is less than $4N_e$ generations which corresponds to ~700.000 years, where N_e is the effective population size of *S. chilense*. Conclusions did not change qualitatively when we permitted multiple and back mutations (finite sites models).

Currently we are building a freely available R package so that the software can be easily adapted to other systems. Our next step is to extend Jaatha such that the estimated demography can be used as a basis to detect selection on genes of interest.

P-2010

Detecting and characterizing selection at loci subject to recurrent mutation.

Ryan Haasl, Bret Payseur
University of Wisconsin - Madison, Madison, WI, USA

Recurrent mutation at a selected locus due to high mutation rate violates a key assumption of the standard selective sweep model. Therefore, new predictive models are needed to detect and characterize selection at loci that experience recurrent mutation and are commonly found throughout genomes, such as microsatellites and copy number variants (CNVs). We focus on functional microsatellites and develop models of the diploid multiallelic fitness surface, which we use to simulate microsatellite variation and flanking sequence variation. The joint pattern of variation is dependent on the shape of the fitness surface, variability of the microsatellite when selection begins, and microsatellite mutation rate. Still, under a large variety of realistic parameter values, selected microsatellites fail to generate stereotypical signatures of a selective sweep. We attribute this result to recurrent mutation, which, like recombination, causes favored allele(s) to occupy numerous haplotypic backgrounds and limits loss of linked variation due to selection. Moreover, this result suggests typical scans for selection will often fail to identify selected microsatellites and motivates development of an inferential pipeline applicable to loci subject to recurrent mutation. To this end, we describe a rapid simulation method based on multiallelic mutation-selection balance. We adopt the framework of approximate Bayesian computation (ABC), and, using a rejection algorithm, compare empirical and simulated data to identify relevant volumes of parameter space. Then, we use ABC coupled with Markov Chain Monte Carlo to efficiently explore the identified range of parameter space, choose between selection and neutrality, and obtain approximate posterior distributions of mutation and selection parameters. We apply our method to new polymorphism data for 200 microsatellite loci sampled from eight of the human populations included in the 1000 Genomes Project. Surveyed loci include coding dinucleotides and intergenic loci as well as microsatellites that affect gene expression in vitro, are linked to putative selective sweeps identified by SNP surveys, or cause trinucleotide diseases. We generate approximate fitness surfaces for all sampled microsatellites and - based on the subset of microsatellites where selection is inferred - characterize the strength and dynamics of microsatellite selection for the first time. Our models and method address a complex but under-appreciated class of adaptive variation and may be easily extended to similar classes of variation including CNVs. In addition, our results emphasize that interesting examples of adaptive evolution are undetectable by typical scans for selection regardless of the marker density used.

P-2011

The joint effects of background selection and genetic recombination on local gene genealogies

Kai Zeng[1], Brian Charlesworth[2]
[1]*University of Sheffield, Sheffield, UK,* [2]*University of Edinburgh, Edinburgh, UK*

Background selection, the effects of the continual removal of deleterious mutations by natural selection on variability at linked sites, is potentially a major determinant of DNA sequence variability. However, the joint effects of background selection and genetic recombination on the shape of the neutral gene genealogy have proved hard to study analytically. The only existing formula concerns the mean coalescent time for a pair of alleles, making it difficult to assess the importance of background selection from genome-wide data on sequence polymorphism. Here we develop a structured coalescent model of background selection with recombination, and implement it in a computer program that efficiently generates neutral gene genealogies for an arbitrary sample size. We check the validity of the structured coalescent model against forward-in-time simulations, and show that it accurately captures the effects of background selection. The model produces more accurate predictions of the mean coalescent time than the existing formula, and supports the conclusion that the effect of background selection is greater in the interior of a deleterious region than at its boundaries. The level of linkage disequilibrium between sites is elevated by background selection, to an extent that is well summarized by a change in effective population size. The structured coalescent model is readily extendable to more realistic situations, and should prove useful for analyzing genome-wide polymorphism data.

P-2012

Genetic structure of spatially overlapping contact zones of two Iberian amphibians: a comparative approach

Helena Gonçalves[1], Jesús Díaz-Rodríguez[1,2], Alexandre Silva-Ferreira[1,3], Tiago S. Neves[1], Miguel Tejedo[2], Iñigo Martínez-Solano[4], Fernando Sequeira[1]
[1]CIBIO/UP, Centro de Investigação em Biodiversidade e Recursos Genéticos da Universidade do Porto, Vairão, Portugal,[2]Departamento de Ecología Evolutiva, Estación Biológica de Doñana-CSIC, Sevilla, Spain,[3]Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Porto, Portugal,[4]Instituto de Investigación en Recursos Cinegéticos (CSIC-UCLM-JCCM), Ciudad Real, Spain

Spatially coincident hybrid zones for different organisms (*sensu* suture zones) provide natural replicates to analyze the dynamics of lineage divergence and admixture patterns. Using two ecologically distinct amphibian species (parsley frog, *Pelodytes* spp.; and Bosca's newt, *Lissotriton boscai*), both presenting coincident secondary contact zones between two highly differentiated lineages (≈ 3-6 Myr.) in the southwest of the Iberian Peninsula , we address: *i*) how species-specific genetic cohesiveness could be differently affected by similar historical and current environmental factors; and *ii*) whether the observed dynamics and patterns of genetic structure are in agreement with previously described patterns in other Iberian hybrid organisms exhibiting lower levels of lineage divergence. Individuals collected along transects defined independently for each species' and across the contact zones were typed for a set of SNPs and STR loci. Individual multilocus genotypes were assigned to each of the forms using clustering methods. Our preliminary results revealed highly contrasting modes of introgression across markers in the hybrid zones of both species, in agreement with the general patterns described for other hybrid zones of Iberian organisms. The complex patterns of admixture detected in these hybrid zones will be discussed in light of the specific evolutionary dynamics associated with the long-term persistence of populations in refugial southern Peninsulas.

P-2013

Private haplotypes as detectors for population-specific selection

Agnès Sjöstrand[1,2], Per Sjödin[2,3], Mattias Jakobsson[2]
[1]MNHN, Paris, France,[2]EBC, Uppsala University, Uppsala, Sweden,[3]SLU, Uppsala, Sweden

Detecting genes targeted by selection in the Human genome has been a challenge for population geneticists and several tools have been developed for this purpose. The following study describes a new statistical tool - MFPH - to scan genomes for signal of recent population-specific selection. MFPH, which is the maximum frequency of private haplotypes, is easy to compute on phased SNP and sequence data and is equally robust as SNP-based analyses. Comparing with other statistics (F_{st}, Tajima's D, Fay and Wu's H, iHS), both on simulated and empirical data, we show that MFPH is a powerful statistic to detect strong and recent population-specific selection events when migration among populations is restricted.

P-2014

Inferring human population history from multiple genome sequences

Stephan Schiffels, Richard Durbin
Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

The Pairwise Sequentially Markovian Coalescent (PSMC) model was introduced by Li and Durbin (2011) to fit a history of ancestral effective population sizes from a single diploid genome sequence. This was achieved by modelling the pattern of mutations between the two chromosomes of the individual by accounting for the segmentation of the genome into regions with a shared common ancestor, separated by ancestral recombinations. While the method pioneered the use of individual genome sequences for demographic inference and yields accurate insights into human evolutionary history older than about 20ky ago, the more recent past can only be covered with very poor resolution. Here we present an extension of PSMC towards inference from multiple chromosomes (MSMC). Instead of modelling the full genealogical tree connecting several individuals, we reduce the complexity of the problem by considering only the first coalescence between any pair of chromosomes. Because the expected coalescence time of any two pairs of multiple sequences is shorter than with just two chromosomes, the inferred population size history is more focussed on the recent past. We use numerical and analytical methods to validate our method and develop an inference protocol on data from multiple phased haplotypes. We discuss the application of this method to published human genome sequences from father-mother-child trios within the 1000 genomes project and other sources.

P-2015

Signs of adaptation in interacting genes

Joséphine Daub[1,3], Marc Robinson-Rechavi[2,3], Laurent Excoffier[1,3]
[1]University of Berne, Berne, Switzerland, [2]University of Lausanne, Lausanne, Switzerland, [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland

In order to find genes that are involved in adaptive events, most approaches try to detect outlier loci, and these attempts often result in the discovery of a given number of ´significant´ genes. However, many small effect mutations can have a large effect on a given pathway involving several genes. We therefore propose to uncover signals of natural selection in pathways or gene sets instead of looking at single independent genes. We used a simple but fast gene set enrichment test (SUMSTAT) to identify gene sets enriched for adaptive signals among human populations. First results reveal several candidate pathways for having been the target of positive selection in recent human evolution. The genes that give the enrichment signal in these pathways are found on different chromosomes, suggesting functional epistatic interactions, which is confirmed by the detection of long distance linkage disequilibrium between SNPs of different genes.

P-2016

A new haplotype test for the detection of selection signatures using data from multiple populations.

Maria Ines Fariello, Magali SanCristobal, Simon Boitard, Bertrand Servin
INRA, Toulouse, France

The genomic diversity of populations is affected by several evolutionary forces (mutation, drift, migration, selection ...). Disentangling 'neutral' forces (acting on the whole genome) from selective ones (targeting few and precise genomic regions) is of great interest in human, plant or animal genetics, to detect genes controlling selected traits. When selected loci are detected based on the divergence between several populations, statistical methods must account for two types of correlations : the correlation between populations, which arises from their phylogeny, and the correlation between SNPs.
Here we present a new haplotype test for the detection of selection signatures using data from multiple populations. This test extends the one of Bonhomme et al (2010), which already considers correlations between populations but focuses on single SNPs. We account for the dependence between SNPs by using the haplotype clustering model of Scheet and Stephens (2003). Using simulations, we show that the detection power of the new test exceeds that of the F-LK test in various situations. One important difference is also that our new test identifies entire regions, while the F-LK test only detects particular SNPs. Finally, we show an application of both tests to a set of sheep populations, and explain how the selected population(s) can be pinpointed.

P-2017

EVIDENCE FOR WIDESPREAD POSITIVE AND PURIFYING SELECTION ACROSS THE EUROPEAN RABBIT (ORYCTOLAGUS CUNICULUS) GENOME

Miguel Carneiro[1,7], Frank Albert[5,6], José Melo-Ferreira[1], Nicolas Galtier[4], Philippe Gayral[4], Jose Blanco-Aguiar[2], Rafael Villafuerte6[2], Michael Nachman[3], Nuno Ferrand[1,7]
[1]CIBIO-UP, Vairão, Portugal, [2]IREC, Ciudad Real, Spain, [3]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, USA, [4]Institut des Sciences de l'Evolution Université Montpellier 2, Montpellier, France, [5]Princeton University, Princeton, USA, [6]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [7]Departamento de Biologia, Faculdade de Ciências Universidade do Porto, Porto, Portugal

The nearly neutral theory of molecular evolution predicts that the efficacy of both positive and purifying selection is a function of the long term effective population size (N_e) of a species. Under this theory, the efficacy of natural selection should increase with N_e. Here, we tested this simple prediction by surveying ~1.5-1.8 Mb of protein coding sequence in the two subspecies of the European rabbit (*O. c. algirus* and *O. c. cuniculus*), a mammal species characterized by high levels of nucleotide diversity and Ne estimates for each subspecies on the order of $1X10^6$. When the segregation of slightly deleterious mutations and demographic effects were taken into account, we inferred that >60% of amino acid substitutions on the autosomes were driven to fixation by positive selection. Moreover, we inferred that a small fraction of new amino acid mutations (< 4%) are effectively neutral (defined as $0 < N_e s < 1$), and that this fraction was negatively correlated with a gene's expression level. Consistent with models of recurrent adaptive evolution, we detected a negative correlation between levels of synonymous sites polymorphism and the rate of protein evolution, although the correlation was weak and non-significant. No systematic X-chromosome-autosomes difference was found in the efficacy of selection. For example, the proportion of adaptive substitutions was significantly higher on the X-chromosome compared to the autosomes in *O. c. algirus*, but not in *O. c. cuniculus*. Our findings support widespread positive and purifying selection in rabbits, and add to a growing list of examples suggesting that differences in Ne among taxa play a substantial role in determining rates and patterns of protein evolution.

P-2018

Inference of demographic history and natural selection in African Pygmy populations from whole-genome sequencing data

Martin Sikora[1], Etienne Patin[2], Helio Costa[1], Katherine Siddle[2], Brenna M Henn[1], Jeffrey M Kidd[1,3], Ryosuke Kita[1], Carlos D Bustamante[1], Lluis Quintana-Murci[2]
[1]Department of Genetics, School of Medicine, Stanford Uni, Stanford, CA, USA, [2]Unit of Human Evolutionary Genetics, Institut Pasteur, CNRS URA3012, Paris, France, [3]Departments of Human Genetics and Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

The Pygmy populations of Central Africa are some of the last remaining hunter-gatherers among present-day human populations, and can be broadly classified into two geographically separated groups, the Western and Eastern Pygmies. Compared to their neighboring populations of predominantly Bantu origin, Pygmy populations show distinct cultural and physical characteristics, most notably short stature, often referred to as the “Pygmy phenotype”. Given their distinct physical characteristics, the questions of the demographic history and origin of the Pygmy phenotype have attracted much attention. Previous studies have shown an ancient divergence (~60,000 years ago) of the ancestors of modern-day Pygmies from non-Pygmies, and a more recent split of the Eastern and Western Pygmy groups. However, these studies were generally based on a relatively small set of markers, precluding accurate estimations of demographic parameters. Furthermore, despite the considerable interest, to date there is still little known about the genetic basis of the small stature phenotype of Pygmy populations.

In order to address these questions, we sequenced the genomes of 47 individuals from three populations: 20 Baka, a Pygmy hunter-gatherer population from the Western subgroup of the African Pygmies; 20 Nzebi, a neighboring non-Pygmy agriculturist population from the Bantu ethnolinguistic group; as well as 7 Mbuti, Eastern Pygmy population, from the Human Genome Diversity Project (HGDP). We performed whole-genome sequencing using Illumina Hi-Seq 2000 to a median sequencing depth of 5.5x per individual. After stringent quality control filters, we call over 17 Million SNPs across the three populations, 32% of them novel (relative to dbSNP 132). Genotype accuracy after imputation was assessed using genotype data from the Illumina OMNI1 SNP array, and error rates were found to be comparable to other low-coverage studies (< 3% for most individuals). Preliminary results show relatively low genetic differentiation between the Baka and the Nzebi (mean F_{ST} = 0.026), whereas the Mbuti show higher differentiation to both Baka and Nzebi (mean F_{ST} = 0.060 and 0.070, respectively). Furthermore, we find that alleles previously found to be associated with height in other populations are not enriched for the “small” alleles in the Pygmy populations. We find a number of highly differentiated genomic regions as candidate loci for height differentiation, which will be verified using simulations under the best-fit demographic model, inferred from multi-dimensional allele frequency spectra using DaDi. Our dataset will allow a detailed investigation of the demographic history and the genomics of adaptation in these populations.

P-2019

On the prospect of identifying adaptive loci in recently colonized populations

Yu-Ping Poh[1,2], Vera Domingues[2], Hopi Hoekstra[2], Jeffery Jensen[1,3]
[1]UMass Medical School, Worcester,MA, USA,[2]Harvard University, Cambridge,MA, USA,[3]EPFL, Lausanne, Switzerland

Our ability to detect the footprint of a selective sweep in genomic data has been one of the most important recent advances in population genetics. However, distinguishing the effects of selective from demographic factors (e.g., a population bottleneck) is a particular challenge—both factors can cause a local reduction in nucleotide variation, skews in the site frequency spectrum, and elevated levels of linkage disequilibrium. This can be particularly difficult when a population colonizes a novel habitat – experiencing potentially severe demographic and selective forces simultaneously. Here we share the results of a detailed simulation study, which models a range of selective strengths and bottleneck severities. First, in small populations, we found no power to detect even strong and recent selective sweeps in recently founded populations. Second, although the true positive rates increased for moderate-sized populations, the false positive rate also increased dramatically. Third, the false positive rates were particularly high when the selective pressure and bottleneck take place simultaneously (i.e., colonization event) in the recent past. We also evaluated the relative advantage of linkage disequilibrium vs. site frequency spectrum based approaches – and found that while the former outperforms the latter under such models – there is still a very large parameter space which presents great difficulty. Thus, we propose that the commonly used tests of selection are at best ineffective, and at worst misleading, when applied to populations that have experienced a recent size reduction – a result which has major implications for studies ranging from the process of domestication to the colonization of novel habitats.

P-2020

The story of wheat domestication and improvement as told by the nutrient content gene *NAM-B1*

Jenny Hagenblad
Linkoping University, Linkoping, Sweden

The transition of domestic crop species from wild progenitor to high-yielding modern varieties is characterized by two major events both involving the introduction of new and strong selection pressures acting on the species. The first, the domestication, occurred some 10000 years ago while the second, modern plant improvement, has taken place during the past 150 years. A few genes have been tied to selection during domestication and plant improvement, most notably the *tb1* gene in maize.

One suggested domestication gene is *NAM-B1* in wheat. It has antagonistic effects with the wildtype allele giving faster maturation time and higher nutrient content in the seed. Two null alleles instead lead to slower maturation time causing larger yield, but with a lower nutrient content. Although originally suggested to be a domestication gene we could show, using historical collections of 19th century seeds that the wildtype allele was still present in widely cultivated wheats during the latter half of the 19th century. Screening accessions from different species of extant wheats from across the world for genetic variation in and around the *NAM-B1* region further paints a picture of the evolutionary history of this gene and the interacting roles of natural and artificial selection.

P-2021

1001 queens: spatial and temporal variation of kin structure in super-colonial ants

Eva Schultner[1], Jari Saramäki[2], Heikki Helanterä[1]
[1]University of Helsinki, Helsinki, Finland,[2]Aalto University, Aalto, Finland

Ant super-colonies are the largest cooperative units known in nature, with large networks of interconnected nests spanning vast areas. In these systems, individuals move freely between nests and each nest contains hundreds of reproductive queens that all contribute to the production of offspring. Levels of relatedness between nestmates are thus extremely low, and workers may incur high fitness costs by rearing unrelated brood. One way for helping individuals to avoid these costs and preferentially direct their efforts towards relatives may be by remaining in their natal nest until the rearing of sexual brood is completed, and only then moving into other nests. If this is the case, kin structure in nests within ant super-colonies should change over time.

We are analyzing the spatial and temporal variation of kin structure in two *Formica aquilonia* super-colonies using polymorphic microsatellite loci. Our study is the first to investigate how kin structure in a population of super-colonial ants may vary over time, and will shed light on the mechanisms responsible for maintaining cooperation when levels of relatedness are extremely low.

P-2022

The effects of sweeps on surrounding loci in large populations

Daniel Weissman, Nicholas Barton
IST Austria, Klosterneuburg, Austria

Selective sweeps reduce the effectiveness of selection and the amount of neutral genetic diversity at linked loci. A simple model of an adapting population shows that interference among sweeps prevents the rate of adaptive substitutions from much exceeding one per chromosome per sexual generation. Even low densities of strongly-selected sweeps are sufficient to interfere with more weakly-selected loci and to greatly reduce neutral diversity. The stochasticity due to linked sweeps differs fundamentally from that due to drift, and cannot be described in terms of a reduced "effective population size." Statistics based on the distribution of locations of ancestral crossovers around a substitution may provide a clear signal of positive selection, robust to the details of the sweep dynamics.

P-2023

An evolutionary history of the selectin gene cluster in humans

Rachele Cagliani[1], Matteo Fumagalli[1], Marco Fracassetti[1], Diego Forni[1], Uberto Pozzoli[1], G. Pietro Comi[2], Federico Marini[1], Nereo Bresolin[1,2], Mario Clerici[3], Manuela Sironi[1]
[1]Scientific Institute IRCCS E. Medea, Bosisio Parini (LC), Italy, [2]Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, Milano, Italy, [3]Chair of Immunology, Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, Milano, Italy, [4]Fondazione Don C. Gnocchi, IRCCS, Milano, Italy

Selectins are C-type lectins playing a central role in immune/inflammatory processes. In humans three selectin genes are located in a ~150 kb region on chromosome 1. We performed a sliding window analysis of FST along the selectin cluster. Peaks of significantly high population genetic differentiation were restricted to two regions in SELP. Sanger resequencing data indicated that the region covering SELP exons 11-13 displays high nucleotide diversity in Africans and Europeans, and a high level of within-species diversity compared to inter-specific divergence. Analysis of inferred haplotypes revealed a complex phylogeny with two deeply separated clades that coalesce at ~3.5 million years (MY) plus a minor clade with a TMRCA of ~2.2 MY. Overall, these data are consistent with a model of multiallelic balancing selection, and SNP analysis indicated that the Val640Leu variant represents a likely selection target. In populations of Asian ancestry a distinct haplotype, possibly carrying regulatory variants, has been driven to high frequency by positive selection. Resequencing of SELP in chimpanzees revealed a haplotype phylogeny with extremely deep basal branches, suggesting either long-standing balancing selection or ancestral population structure. No allele was shared among humans and Pan troglodytes. Thus, SELP has experienced a complex selective history, possibly as a result of local adaptation. Variants in the gene have been associated with autoimmune and cardiovascular diseases. Data herein suggest that association studies would benefit from both taking the complex SELP haplotype structure into account and from the analysis of possible regulatory variants in the gene.

P-2024

Integration of selection signatures and GWAS data identifies novel susceptibility loci for Crohn's disease

Rachele Cagliani[1], Diego Forni[1], Uberto Pozzoli[1], Andrea Cassinotti[4], Matteo Fumagalli[1], Matteo Giani[1], Stefania Riva[1], Rosanna Asselta[6], Giacomo P Comi[5], Nereo Bresolin[5], Mario Clerici[3,2], Manuela Sironi[1]
[1]Scientific Institute IRCCS E. Medea, Bosisio Parini, Italy, [2]Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy, [3]Department of Biomedical Sciences and Technologies LITA Segrate, University of Milan, Milan, Italy, [4]L. Sacco University Hospital, Milan, Italy, [5]Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, Milan, Italy, [6]Department of Biology and Genetics for Medical Science, University of Milan, Milan, Italy

Crohn's disease (CD) is a chronic inflammatory disease of the intestines which is though to result from a combination of genetic and environmental risk factors. GWAS have identified several susceptibility loci for CD, but these explain less than 25% of disease heritability. A portion of missing risk loci is likely to be accounted for by common variants with modest effect which have been treated as false-negatives in GWAS due to the heavy multiple test correction. Thus, leveraging GWAS results with other types of data might allow the identification of additional risk variants. Previous studies have indicated that risk alleles for autoimmune diseases have been targets of pathogen-driven selective pressure. We estimated pathogen-driven selection for single SNPs in the HGDP-CEPH panel using the diversity of virus, protozoan, bacterium or helminth species transmitted in distinct geographic locations. We next retrieved all GWAS SNPs from the NHGRI Catalog of Published Genome-Wide Association Studies, and collapsed SNPs in linkage disequilibrium into single loci. The percentage of CD SNPs that was significantly associated with virus, bacterium, protozoan and helminth diversity were 12.8, 4.2, 19.1 and 6.3, respectively, suggesting a strong enrichment for SNPs targeted by protozoa-driven selection. The empirical probability of obtaining 19.1% of variants significantly associated with protozoan diversity amounted to 0.003. Analysis of association p values from a CD GWAS meta-analysis indicated that protozoan-selected SNPs display significantly stronger association compared to non-selected variants. We reasoned that true-positive associations might be enriched among SNPs that display signatures of protozoa-driven selection and do not reach the meta-analysis threshold for significance. We extracted all SNPs showing a meta-analysis p vlaue higher than 5×10^{-5}; those associated with protozoa diversity were ranked according to their meta-analysis p value, and the top 5 genic SNPs were selected for replication in an Italian cohort of 750 CD cases and 650 controls. Three SNPs (in *ARHGEF2*, *NSF*, and *HEBP1*) were significantly associated with CD risk. Under the null hypothesis with 5 tested SNP, no variant would be expected to yield a significant association with $p\leq0.05$, suggesting that these variants represent true positive associations. We finally applied unsupervised Ingenuity Pathway Analysis (IPA) using the three genes as input; IPA identified one network with significant score ($p=10^{-6}$) which contains an additional known CD gene (*VAMP3*), and is centred around mir-31, a microRNA known to be dysregulated in CD.
Thus, by integrating selection signatures and GWAS data we identified novel susceptibility loci for CD.

P-2025

North American *Drosophila melanogaster* produced by admixture between African and European populations: An Approximate Bayesian Computation approach

Pablo Duchen, Stefan Laurent, Daniel Zivkovic, Wolfgang Stephan
University of Munich, Munich, Germany

Drosophila melanogaster spread from sub-Saharan Africa to the rest of the world colonizing and adapting to new environments. In order to find signatures of adaptation it is necessary to disentangle the confounding effects of demography and selection. Here, we modeled the joint demography of Africa (Zimbabwe), Europe (Netherlands) and North America (North Carolina) using an Approximate Bayesian Computation (ABC) approach. By testing different models we found that an admixture between Africa and Europe most likely produced the North American population as opposed to migration or no migration. We also revisited the demography of the ancestral population (Africa) and found that a bottleneck and not a population expansion fits better the history of Zimbabwe. We computed the observed site-frequency spectrum of the ancestral population and compared it to analytical predictions under a bottleneck model using the parameter estimates of our ABC approach. The observed site-frequency spectrum showed an excess in the high-frequency classes, which is not expected under a bottleneck model. We found that this excess is the result of ancestral misidentification and we performed an analytical correction of the observed site-frequency spectrum.

P-2026

Genetic structure in North African human populations and the gene flow to Southern Europe

Laura R Botigué[1], Brenna M Henn[2], Simon Gravel[2], Jaume Bertranpetit[1], Carlos D Bustamante[2], David Comas[1]
[1]Institut de Biologia Evolutiva (IBE, CSIC-UPF), Barcelona, Spain,[2]Stanford University, Stanford CA, USA

Despite being in the African continent and at the shores of the Mediterranean, North African populations might have experienced a different population history compared to their neighbours. However, the extent of their genetic divergence and gene flow from neighbouring populations is poorly understood. In order to establish the genetic structure of North Africans and the gene flow with the Near East, Europe and sub-Saharan Africa, a genomewide SNP genotyping array data (730,000 sites) from several North African and Spanish populations were analysed and compared to a set of African, European and Middle Eastern samples. We identify a complex pattern of autochthonous, European, Near Eastern, and sub-Saharan components in extant North African populations; where the autochthonous component diverged from the European and Near Eastern component more than 12,000 years ago, pointing to a pre-Neolithic "back-to-Africa" gene flow. To estimate the time of migration from sub-Saharan populations into North Africa, we implement a maximum likelihood dating method based on the frequency and length distribution of migrant tracts, which has suggested a migration of western African origin into Morocco ~1,200 years ago and a migration of individuals with Nilotic ancestry into Egypt ~ 750 years ago.
We characterize broad patterns of recent gene flow between Europe and Africa, with a gradient of recent African ancestry that is highest in southwestern Europe and decreases in northern latitudes. The elevated shared African ancestry in SW Europe (up to 20% of the individuals' genomes) can be traced to populations in the North African Maghreb. Our results, based on both allele-frequencies and shared haplotypes, demonstrate that recent migrations from North Africa substantially contribute to the higher genetic diversity in southwestern Europe.

P-2027

The influence of natural selection in the antigen presentation pathway

Felix M Schulze, Genís Parra, Cesare de Fillipo, Aida M Andrés
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Natural Selection has continuously shaped the functionality of the immune system. The evolutionary development of the adaptive immunity with its seemingly infinite pathogen recognition motifs is an outstanding achievement of life in combating the continuous stream of infectious agents. To deploy the pathogen recognition to its maximal effectiveness, a versatile antigen presentation pathway is crucial. Meaning, a system which is able to process proteins originating from sources as diverse as viruses, bacteria or fungi, and effectively present them to the adaptive immune system.

We focus in our work on the mayor histocompatibility complex I (MHC I) peptide loading complex (MHCPLC) pathway. In brief, this machinery is responsible for degrading cytosolic proteins into peptides which become loaded in a stable MHC I peptide complex and subsequently transported to the cell surface. A recognition of the peptide presented by MHC I as foreign by the adaptive immune system elicits an immune response necessary to overcome an infection.

Our interest in this pathway is based on several interesting facts of its evolution. First, the MHC I is the most variable locus in the genome due to the action of balancing selection, likely due to selective pressures to improve antigen recognition and thus immunity. Second, it is now known that other key players in the MHCPLC, like ERAP1 and ERAP2, show also signals of balancing selection in humans, suggesting that other steps in this pathway may benefit from the presence of advantageous genetic diversity.

Our research focuses on elucidating the selective forces behind the diversity of all proteins involved in the MHCPLC. To address this question, we have re-sequenced, using target capture and next-generation sequencing, the 13 key players in the MHCPLC in 202 HapMap samples representing six human populations. We generated a high quality dataset of polymorphisms for our genes, characterized their patterns of polymorphism through a variety of population genetics approaches, and inferred their signatures of selection by comparison with neutral coalescent simulations and neutral genomic regions. This provided a global picture of the evolutionary history of the different steps of the MHCPLC, further clarifying the selective pressures affecting this important immunological pathway.

By integrating information about the signatures of selection of the complete pathway, this work helps elucidate the crucial mechanism of antigen presentation, and puts our present knowledge about the evolution of the MHC-I antigen presentation pathway in its functional and mechanistic context.

P-2028

Genetic variability of Calvin cycle genes in accessions of *Arabidopsis thaliana*

Madlen Stange, Sandra Schwarte, Fanny Wegner, Ralph Tiedemann
University of Potsdam, Potsdam, Germany

The Calvin cycle is the essential pathway for carbon fixation in green plants and is part of the primary metabolism. We analyzed 18 genes regarding genetic variation among 26 *A. thaliana* accessions. These genes code for enzymes which catalyze the reactions of the Calvin cycle, additionally 4 genes encoding regulatory enzymes. The hitherto unknown nucleotide diversity was similar to those of secondary metabolism genes, e.g. from the phenylproponoid pathway or disease resistance genes. This was surprising due to the fact that it was commonly expected that the primary metabolism is rather conserved. Several polymorphisms were detected in functional relevant regions for instance in promoter or coding regions. Some polymorphisms occurring in coding regions were predicted to possibly have implications on protein structure and presumably on protein function. Applying one-tailed *Z*-tests and Tajima's *D* confirmed most genes to be under purifying selection. A Tajima's *D* test supports the findings of the *Z*-test. Although all genes are essential for a proper function of the Calvin cycle, the tolerance towards nonsynonymous SNPs in coding regions differed among the genes analyzed here. Evolutionary implications of this observation are discussed.

P-2029

Patterns of strong selective sweeps in Northern Swedish *Arabidopsis thaliana*

Christian Huber[1], Quan Long[2], Magnus Nordborg[2], Ines Hellmann[1]
[1]*Max F. Perutz Laboratories, Vienna, Austria,* [2]*Gregor Mendel Institute of Molecular Plant Biology GmbH, Vienna, Austria*

A. thaliana is one of the best characterized population genetics model systems for which there is abundant sequence data available. Despite this, there is very little evidence for hard selective sweeps. Thus far, there has been only one global selective sweep identified. Here, we explore the possibility that selective sweeps occur by and large locally, and are therefore not detected by scans based on samples with a wide geographic spread. Using *A. thaliana* plants from N. Sweden, we find several strong signals of selective sweeps, identified through a shift of the site frequency spectrum towards low and high frequency derived alleles. Sweep signals in the N-Sweden are an order of magnitude stronger than in a Southern Swedish population, where we detect hardly any. This suggests that the unusual population history of the Northern Swedish population magnifies sweep signals. This is exemplified by a previously detected worldwide sweep on chromosome 1. The footprint of this sweep in N-Sweden spans 3 Mbp, while it only stretches over 0.5 Mbp in the S-Sweden. However, except for this unique region, all other putative sweep regions show a significantly higher degree of genetic differentiation between N- and S-Sweden than other regions, which is expected under a model of local selective sweeps. Further evidence for the mainly local nature of adaptation in *A. thaliana* comes from the lack of correlation between several sweep detection statistics calculated for N- and S-Swedish populations. In conclusion, we suggest that selective sweeps are predominantly local in extent, and are magnified in N-Sweden due to the specific population history and demography of that population.

P-2030

Salt tolerance in *Arabidopsis thaliana*: polymorphism and phenotypic response to salt stress

Eva Puerma, Montserrat Aguadé
Universitat de Barcelona, Barcelona, Spain

Soil salinity is one of the most significant abiotic stressses for plants. The genetic basis of this character has been studied in several model species including *Arabidopsis thaliana*, which is also a model species for population genetic studies. Its complex demographic history complicates efforts to uncover the recent action of natural selection from levels and patterns of nucleotide variation. However, the availability of extensive genome-wide information in this species facilitates obtaining empirical distributions of different summary statistics with which to contrast those obtained from variation at particular genes or sets of genes. We have surveyed DNA sequence variation at nine genes involved in salt tolerance (~32 kb) in 20 ecotypes of *A. thaliana* with a worldwide geographical distribution. Variation at these nine genes is characterized by a frequency spectrum skewed towards polymorphisms with low frequency variants, as revealed by average negative values of Tajima's *D* and Fay and Wu's *H* test statistics. From comparison with empirical distributions, we could infer that this pattern mainly reflects the demographic history of the species. We have also measured the phenotypic response of the studied ecotypes to different salt conditions in order to explore the relationship between polymorphism at the nine genes and the phenotypic response to salt stress.

P-2031

Identification of Soft Sweeps in the *Drosophila melanogaster* Genome

Nandita Garud, Philipp Messer, Dmitri Petrov
Stanford University, Stanford, California, USA

We aim to determine whether recent adaptation in Drosophila results in primarily hard or soft sweeps. A soft sweep differs from the standard hard sweep model of adaptation in that multiple haplotypes harboring the same adaptive mutation rise in frequency simultaneously. Soft sweeps can arise when multiple adaptive mutations at the same position occur practically simultaneously in large populations. Alternatively, soft sweeps can be generated by adaptive mutations that are initially neutral or slightly deleterious and have enough time to recombine onto multiple haplotypes prior to a change in the environment resulting in the mutations becoming adaptive. In contrast to a classical hard sweep, heterozygosity is not completely lost during a soft sweep as multiple adaptive haplotypes increase in frequency and remain in population upon completion of the sweep.

We developed a haplotype-based test statistic to first detect both hard and soft sweeps and then to distinguish them from each other. First, we identify multiple genomic regions in the *Drosophila* Genetic Reference Panel that have extreme values of haplotype homozygosity as compared to neutral demographic scenarios. Then, we apply our test statistic to each of these unusual regions to distinguish whether the selective event in question generated a hard or soft sweep. Surprisingly, we rejected the null hypothesis that these regions are undergoing hard sweeps in almost every case. The alternative hypothesis that these regions are undergoing a soft sweep appears more likely overall. We conclude that recent adaptation in North American populations of *D. melanogaster* has led primarily to soft sweeps either because it utilized standing variation or because short-term effective population sizes in *D. melanogaster* are on the order of billions rather than millions, as we suggested previously.

P-2032

Genome-wide effects of sex-specific hybrid incompatibility on neutral introgression

Marcy Uyenoyama, Seiji Kumagai, Liuyang Wang
Duke University, Durham, North Carolina, USA

A widely-used paradigm holds that while selection affects variation in a region-specific manner, all regions of the genome experience the same demographic history. This view justifies identification of outlier segments as candidates for regions that are associated, through tight linkage or epistasis, with targets of selection. On the contrary, we show that the strong, sex-specific selection induced in many species upon hybridization generates region-specific barriers to neutral introgression. For example, an autosomal incompatibility factor that induces more severe effects in male than in female hybrids can induce a higher barrier against introgression of Y chromosomes than unlinked autosomal regions, and even lower barriers to X-linked regions. Further, Y-linked factors inhibit introgression by autosomes but pose no barrier to mitochondria. These examples illustrate that sex-specific incompatibility factors can induce region-specific backward migration rates throughout the genome, even to the extent of generating higher F_{ST} in autosomal than X-linked regions, a pattern that might be interpreted as a signature of selection or sex-biased dispersal. We present analytical expressions for backward migration rates induced by sex-specific incompatibility. We then use these results to develop a coalescent-based algorithm for sampling from the exact distributions of topologies and internode lengths of the gene genealogy of a sample of neutral sites. We present preliminary results from application of our methods to genomic sequences sampled from Drosophila pseudoobscura and D. persimilis, a model system for introgression between recently diverged species for which locations of hybrid incompatibility factors have been determined by direct experiment.

P-2033

Estimating a date of mixture of ancestral South Asian populations

Priya Moorjani[1,2], Nick Patterson[2], Periasamy Govindaraj[3], Danish Saleheen[4], John Danesh[4], Lalji Singh*[3,5], Kumarasamy Thangaraj*[3], David Reich*[1,2]
[1]Harvard University, Boston, Massachusetts, USA, [2]Broad Institute, Cambridge, Massachusetts, USA, [3]Centre for Cellular and Molecular Biology, Hyderabad, Andhra Pradesh, India, [4]Dept of Public Health and Care, University of Cambridge, Cambridge, UK, [5]Genome Foundation, Hyderabad, Andhra Pradesh, India

Linguistic and genetic studies have demonstrated that almost all groups in South Asia today descend from a mixture of two highly divergent populations: Ancestral North Indians (ANI) related to Central Asians, Middle Easterners and Europeans, and Ancestral South Indians (ASI) not related to any populations outside the Indian subcontinent. ANI and ASI have been estimated to have diverged from a common ancestor as much as 60,000 years ago, but the date of the ANI-ASI mixture is unknown. Here we analyze data from about 60 South Asian groups to estimate that major ANI-ASI mixture occurred 1,200-4,000 years ago. Some mixture may also be older—beyond the time we can query using admixture linkage disequilibrium—since it is universal throughout the subcontinent: present in every group speaking Indo-European or Dravidian languages, in all caste levels, and in primitive tribes. After the ANI-ASI mixture that occurred within the last four thousand years, a cultural shift led to widespread endogamy, decreasing the rate of additional mixture.

*- These authors co-mentored the project.

P-2034

Extreme (X)-QTL Mapping Seed and Seedling Traits in Arabidopsis thaliana

Wei Yuan, Michael Purugganan
New York University, New York, USA

An important goal of molecular evolutionary biology is to understand the genetic basis of life history trait variation, an important target for natural selection and adaptive evolution. Extreme (X)- QTL mapping is a new approach of identifying genes for quantitative traits that employs high-throughput bulk-segregant-analysis and large-scale selection-based phenotyping. We will implement X-QTL mapping in *A. thaliana*, and use it to explore variation in germination salt tolerance and temperature-regulated germination in this model plant species. We have developed X-QTL mapping populations using a cross between the natural accessions Col-0 and Bs-2, and will screen 10^5 individuals in a mapping experiment. For genotyping, we have also developed an isothermal SNP mapping array that can estimate SNP frequency in polled populations at 30,000 SNPs throughout the genome, with an average SNP marker density of 1 SNP per 2 kb. This new method is expected to provide a better understanding of the genetic basis for life history traits and their evolution in natural populations.

P-2035

Long-term balancing selection may be more common than previously believed: several examples from apes

Ellen M. Leffler[1], Ziyue Gao[1], Susanne Pfeifer[2], Laure Ségurel[1], Adam Auton[3], Ryan Hernandez[4], Joanna L. Kelley[7], Jeffrey Kidd[7], S. Cord Melton[1], Laurie Stevison[4], Aarti Venkat[1], Oliver Venn[3], Ronald Bontrop[6], Carlos Bustamante[7], Michael Hammer[8], Jeffrey Wall[4], Peter Donnelly[2,3], Gilean McVean[2,3], Molly Przeworski[1,5]
[1]University of Chicago, Chicago, IL, USA, [2]University of Oxford, Oxford, UK, [3]Wellcome Trust Centre for Human Genetics, Oxford, UK, [4]University of California, San Francisco, San Francisco, CA, USA, [5]Howard Hughes Medical Institute, Chicago, IL, USA, [6]Biomedical Primate Research Centre, Rijswijk, The Netherlands, [7]Stanford University, Stanford, CA, USA, [8]University of Arizona, Tucson, AZ, USA

Balancing selection, in which two or more alleles are maintained in a population by selection, is predicted to lead to high levels of diversity and to haplotypes with deep coalescence times. Moreover, if balancing selection pressures are older than species split times, species may share alleles identically by descent (i.e., there may be a “trans-species polymorphism”). Such long-term balanced polymorphism is thought to be extremely rare, with the most notable examples being the MHC in primates and the self-incompatibility loci in plants. However, modeling suggests that the footprint of balancing selection may be difficult to detect, leaving open the possibility that this mode of selection is more common than appreciated. Using genome-wide data from 10 western chimpanzees (*Pan troglodytes verus*) sequenced as part of the PanMap project and 1000 Genomes low coverage YRI data, we undertook a genome-wide search for orthologous sites that are polymorphic for the same alleles in both chimpanzee and human. We found that SNPs are shared in excess of what is expected by chance after accounting for variation in the mutation rate. While it is difficult to distinguish balanced polymorphism from recurrent mutation based on a single SNP, the short ancestral segments on which a balanced polymorphism resides may contain additional ancestral polymorphisms shared between species. We therefore focused on cases of two or more shared polymorphic sites in close proximity and with the same linkage disequilibrium patterns in both species, a scenario that is exceedingly unlikely to occur by recurrent mutation. Besides the MHC, we identified over 70 such loci. To rule out the possibility that they represent the rare regions of the genome where, by chance, coalescent events are deep in both humans and chimpanzees, we looked at patterns of variation in seven samples of *Gorilla gorilla*. In several cases, pairs of polymorphisms were shared among all three species, providing clear-cut evidence for trans-species polymorphisms. This scan points to new targets of selection in apes. Given our conservative criteria, our results further suggest long-term balancing selection may be much more common than previously believed.

P-2036

A computational approach to detect population structure and signatures of local adaptation from high-throughput re-sequencing data

Matteo Fumagalli[1], Thorfinn Korneliussen[2], Tyler Linderoth[1], Anders Albrechtsen[2], Rasmus Nielsen[1,2]
[1]*University of California, Berkeley, CA, USA,* [2]*University of Copenhagen, Copenhagen, Denmark*

New, high-throughput, DNA re-sequencing technologies provide a cost-effective means of obtaining large-scale genetic data, and are now a primary resource for population genetics studies. Some drawbacks for being able to fully benefit from this technology, however, is the computational challenge associated with large datasets and high sequencing error rate. Moreover, many re-sequencing studies employ medium to low sequencing coverage. Under such circumstances, SNP and genotype calling can be unreliable and lead to a biased allele frequency distribution. Accurate estimation of the site frequency spectrum is vital to population genetic inference of demography, natural selection, and population subdivision, as well as the statistics used to summarize these processes.

Many previously used methods for estimating allele frequencies that are primarily based on direct counting of sequencing reads, lead to inaccurate estimates of local nucleotide diversity indices, and F_{ST}, a measure of genetic population differentiation. Recently, more sophisticated algorithms for SNP and genotype calling have been proposed. Most promising are those that use a Bayesian framework to incorporate base quality scores and statistical uncertainty in order to obtain posterior probabilities associated with each genotype.

Accordingly, I propose several new strategies for computing unbiased estimates of F_{ST} from estimation of allele frequencies per site. I will demonstrate the high performance of these novel methods for detecting population subdivision, especially when facing high error rates and low sequencing coverage. I will also discuss how population structure can be investigated with principle component analysis within this probabilistic framework. In addition, I will perform sliding-window analyses on previously published low-coverage re-sequencing data from multiple populations to demonstrate an efficient and reliable strategy for genome-wide scans of natural selection. This analysis will highlight genomic regions displaying exceptionally high genetic differentiation along with reduced levels of nucleotide diversity among pairs of populations. Confidence intervals will be determined by an empirical moving block bootstrap approach.

Methods herein presented are powerful and reliable tools for investigating population genetic variation on a large scale directly from high-throughput re-sequencing data.

P-2037

Long IBD in Europeans and recent population history

Peter Ralph, Graham Coop
UC Davis, Davis, CA, USA

Numbers of common ancestors shared at various points in time across populations can tell us about recent demography, migration, and population movements. These rates of shared ancestry over tens of generations can be inferred from genomic data, thereby dramatically increasing our ability to infer population history much more recent than was previously possible with population genetic techniques. We have analyzed patterns of IBD in a dataset of thousands of Europeans from across the continent, which provide a window into recent European geographic structure and migration.

P-2038

Fine-scale mapping complex signatures of positive selection in the human genome

Leslie Emery, Joshua Akey
University of Washington Department of Genome Sciences, Seattle, WA, USA

Despite the recent success of genome-wide scans for signatures of positive selection, two major obstacles exist in comprehensively identifying human adaptive alleles. First, current statistical methods are designed to detect classical selective sweeps, which appear to be relatively rare in human evolutionary history. Second, most genome-wide scans for selection rely on one or two neutrality test statistics that only utilize a subset of the available genomic information. To address these issues, we develop and optimize several statistics to detect and fine-scale map recent positive selection in human populations. Using demographically calibrated coalescent simulations over a broad range of selective models, we extensively validate the power and operating characteristics of these statistics in detecting positive selection in more complicated models of adaptation, such as selection from standing variation. Finally, we develop a composite likelihood ratio method for combining information from several neutrality test statistics based on linkage disequilibrium, the site frequency spectrum, population structure, and haplotype structure. We apply our method to whole-genome sequences from the 1,000 Genomes Project data set and compare our results to those of previous scans for positive selection.

P-2039

Heterozygote advantage as a natural consequence of adaptation in diploids

Diamantis Sellis[1], Benjamin Callahan[2], Dmitri Petrov[1], Philipp Messer[1]
[1]Department of Biology, Stanford University, Stanford, CA 94305, USA, [2]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA

Molecular adaptation is typically assumed to proceed by sequential fixation of beneficial mutations. In diploids, this picture presupposes that for most adaptive mutations, the homozygotes have a higher fitness than the heterozygotes. We show that contrary to this expectation, a substantial proportion of adaptive mutations should display heterozygote advantage. This feature of adaptation in diploids emerges naturally from the primary importance of the fitness of heterozygotes for the invasion of new adaptive mutations. We formalize this result in the framework of Fisher's influential geometric model of adaptation. We find that in diploids, adaptation should often proceed through a succession of short-lived balanced states that maintain substantially higher levels of phenotypic and fitness variation in the population compared with classic adaptive walks. In fast-changing environments, this variation produces a diversity advantage that allows diploids to remain better adapted compared with haploids despite the disadvantage associated with the presence of unfit homozygotes. The short-lived balanced states arising during adaptive walks should be mostly invisible to current scans for long-term balancing selection. Instead, they should leave signatures of incomplete selective sweeps, which do appear to be common in many species. We attempt to directly measure the fitness advantage of heterozygotes in *Saccharomyces cerevisiae* by competitive growth experiments. We are currently evolving diploid yeast strains in a low glucose environment and screening the adapted clones for overdominant mutations. Our theoretical results raise the possibility that balancing selection, as a natural consequence of frequent adaptation, might play a more prominent role among the forces maintaining genetic variation than is commonly recognized.

P-2040

Signals of positive selection in North African human populations

Laura R. Botigué[1], Brenna M. Henn[2], David Comas[1], Carlos D. Bustamante[2]
[1]IBE (UPF-CSIC), Barcelona, Spain,[2]Department of Genetics, Stanford University, Stanford, California, USA

Patterns of human structure in North Africa are highly complex and cannot be explained by models involving two or even three admixture events, as has been shown by recent studies. Genetic ancestries from Arabic, European, eastern and western sub-Saharan sources, as well as an autochthonous Maghrebi ancestry have been identified in North African populations. North Africa is also characterized by high insolation and predominance of desertic or semidesertic climate, for which it is probable that some adaptive genetic signal is found. Also, a long and old herding tradition has been described in the region, even if intermediate frequencies of lactase persistence have been previously described. In the present work we aim to investigate the mechanisms by which people living in North Africa have adapted to these situations and also, how gene flow coming from neighboring regions has affected the adaptive architecture of North Africans.

P-2041

Gene flow between human populations during the exodus from Africa, and the timeline of recent human evolution

Aylwyn Scally, Richard Durbin
Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK

We present a novel test for historical gene flow between populations using unphased genotypes in present-day individuals, based on the sharing of derived alleles and making a minimal set of assumptions about their demographic history. We apply this test to data for three human individuals of African, European and Asian ancestry. We find that the joint distribution of European and Asian genotypes is compatible with these populations having separated cleanly at some time in the past without subsequent genetic exchange. However the same is not true of the European-African and Asian-African distributions, which instead suggest an extended period of continued exchange between African and non-African populations after their initial separation.

We discuss this in comparison with recent models and estimates of separation time between these populations. We also consider the impact of recent direct experimental studies of the human mutation rate, which suggest rates of around 0.5×10^{-9} bp^{-1} y^{-1}, substantially lower than prior estimates of 1×10^{-9} bp^{-1} y^{-1} obtained from calibration against the primate fossil record. We show that in several places the lower rate, implying older dates, yields better agreement between genetic and non-genetic (paleoanthropological and archaeological) evidence for events surrounding the exodus of modern humans from Africa and their dispersion worldwide.

P-2042

Adaptation to urban environments in *Arabidopsis thaliana*

Angela Hancock[1,2], Felice Sperone[2], Joy Bergelson[2]
[1]*University of Vienna, Vienna, Austria,* [2]*University of Chicago, Chicago, IL, USA*

Plants growing in urban environments face unique challenges, and their abilities to thrive in these conditions may be mediated by natural selection. We scanned the *Arabidopsis thaliana* genome to identify genetic variants with higher frequencies in urban compared to non-urban environments. We found that truncating amino acid changes were significantly over-represented among the variants most strongly associated with urban environments. Furthermore, when we compared the selection scan results to those from genome-wide association studies to ask which phenotype-associated variants were also associated with urban environments, we found significant enrichments of several phenotypes related to flowering and development. These results are consistent with the idea that urban environments pose novel challenges and lend insight into the physiological impacts of these environments.

P-2043

Inference in the extinction/recolonisation model

Nick Barton[2], Alison Etheridge[3], Jerome Kelleher[1], Amandine Veber[4]
[1]University of Edinburgh, Edinburgh, UK, [2]IST Austria, Vienna, Austria, [3]University of Oxford, Oxford, UK, [4]Ecole Polytechnique, Paris, France

The extinction/recolonisation model solves many problems associated with the classical models of isolation by distance. In this model, a population evolves through recurrent events falling at random locations in a spatial continuum. In each event, some fraction of the local population dies and is replaced by the offspring of a small number of parents. We discuss the inference of the parameters of this model from sequence data, and, in particular, the inference of the parameters driving local structure from patterns of recombination.

P-2044

Long-term presence versus recent admixture: Bayesian and approximate-Bayesian analyses of genetic diversity of human populations in Central Asia

Friso Palstra, Evelyne Heyer, Frederic Austerlitz
Eco-anthropologie et Ethnobiologie UMR 7206 CNRS, Equipe Genetique des Populations Humaines, Museum National d'Histoire Naturelle, Paris, France

A long-standing goal in population genetics is to unravel the relative importance of evolutionary forces that shape genetic diversity. Here we focus on human populations in Central Asia, a region that has long been known to contain the highest genetic diversity on the Eurasian continent. However, whether this variation principally reflects long-term presence, or rather the result of admixture associated with repeated migrations into this region in more recent historical times, remains unclear. Here we investigate the underlying demographic history of Central Asian populations in explicit relation to Western Europe, Eastern Asia and the Middle East. For this purpose we employ both full Bayesian and approximate-Bayesian analyses of nuclear genetic diversity in 20 unlinked non-coding resequenced DNA regions, known to be at least 200 kb apart from any known gene, mRNA or spliced EST (total length of 24 kb), and 22 unlinked microsatellite loci.
Using an approximate Bayesian framework, we find that present patterns of genetic diversity in Central Asia may be best explained by a demographic history which combines long-term presence of some ethnic groups (Indo-Iranians) with a more recent admixed origin of other groups (Turco-Mongols). Interestingly, the results also provide indications that this region might have genetically influenced Western European populations, rather than vice versa. A further evaluation in MCMC-based Bayesian analyses of isolation-with-migration models confirms the different times of establishment of ethnic groups, and suggests gene flow into Central Asia from the east. The results from the approximate Bayesian and full Bayesian analyses are thus largely congruent. In conclusion, these analyses illustrate the power of Bayesian inference on genetic data and suggest that the high genetic diversity in Central Asia reflects both long-term presence and admixture in more recent historical times.

P-2045

Demography and natural selection in chimpanzee populations inferred from full exome sequencing

Christina Hvilsom[2,3], Jinjie Duan[1], Thomas Bataillon[1], Yingrui Li[4], Thomas Mailund[1], Kasper Munch[1], Yu Qian[1], Asger Hobolth[1], Jun Wang[4], Hans R. Siegismund[2], Mikkel H. Schierup[1]
[1]Aarhus University, Aarhus, Denmark, [2]University of Copenhagen, Copenhagen, Denmark, [3]Copenhagen Zoo, Copenhagen, Denmark, [4]BGI, Shenzhen, China

We captured and sequenced at ≈30X coverage the exomes of six western chimpanzees (*Pan troglodytes verus*), 11 eastern chimpanzees (*Pan troglodytes schweinfurthii*) and 12 central chimpanzees (*Pan troglodytes troglodytes*) and called ≈120,000 coding SNPs. We contrast the unfolded site frequency distributions (SFS) of synonymous and non-synonymous SNPs as well as the X-linked versus autosomal SFS and diversity. We apply McDonald-Kreitman tests using the human genome as outgroup for evidence of adaptive evolution for different gene ontology classes and for the comparison of the X chromosome and the autosomes. We find that the very different demographic history of the populations is reflected in different SFSs and different estimated amounts of purifying selection. Differentiation measured by F_{ST} is higher for synonymous SNPs than for non-synonymous SNPs. In all populations the X chromosome harbors only about half the diversity of the autosomes but display similar amounts of differentiation, suggesting that the reduction in X-linked diversity is not due to sex-specific migration. We scan the genome for regions of loss of diversity for both population specific and species wide evidence of selective sweeps. The most striking example of a population specific sweep is a 5 Mbp region on chromosome 3 with several immunity related genes which has severely lost diversity in central chimpanzees and also display the largest values of differentiation as measured by F_{ST}.

P-2046

Patterns of genetic diversity across the European rabbit (Oryctolagus cuniculus) hemoglobin beta (HBB) cluster suggest a recent and incomplete selective sweep at the HBB gene

Rita Campos[1,2], Nuno Ferrand[1,2]
[1]CIBIO - Centro de Investigacao em Biodiversidade e Recursos Geneticos, Vairao, Porto, Portugal, [2]Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Porto, Portugal

The European rabbit, Oryctolagus cuniculus, is composed of two subspecies that recently established a hybrid zone, thus being an excellent model to study the dynamics of speciation. A preliminary analysis of population differentiation given by 25 allozyme loci showed that beta-globin gene (HBB) exhibit lower-than-expected Fst values compared to the background polymorphism, suggesting that some form of balancing selection is driving the evolution of this locus. To test this hypothesis, we conducted a fine scale analysis of the HBB gene cluster, including the complete HBB gene, a microsatellite locus located upstream HBB and intron II of pseudogene delta-globin (HBD). HBB and microsatellite variation allowed identifying the existence of two well differentiated allelic lineages, HBB1 and HBB2. However, this dichotomy does not extend much further along the HBB cluster, probably due to a recombination hotspot in HBD. Coalescent analysis results indicates that the two lineages differentiated during the Pleistocene, and are consistent with a scenario of population decline of HBB1 lineage and of recent population growth of HBB2. To reconcile the observed patterns with the evolutionary history of the European rabbit, we hypothesize that both lineages diverged during the period of isolation of rabbit subspecies and that the recent contact of both population groups promoted the occurrence of an incomplete selective sweep of lineage HBB2 leading to a young balanced frequency of both HBB lineages across populations. Given the functional relevance of HBB gene (e.g. in oxygen transport) and the geographic distribution of the stable allelic frequencies, the balanced polymorphism observed at this locus may be maintained across rabbit populations by heterozygote advantage. Although this hypothesis and the exact nature of the selective sweeps need further investigation, the results obtained here demonstrate how selection can accelerate the rate of introgression in the initial stages of speciation.

P-2047

How the colonization of subantarctic islands shaped mouse populations

Anna Lorenc[1], Ines Hellmann[2], Emilie Hardouin[1], Diethard Tautz[1]
[1]MPI for Evolutionary Biology, Plön, Germany,[2]MaBS at MFPL, Vienna, Austria

Mice, traveling along with humans, colonized almost the whole world. They also reached several uninhabited islands in the subantarctic area (between the 40th and the 60th south parallel), probably as independent settling events from European ships with little gene-flow afterwards. In addition to the strong founder effects, those island populations had to adapt independently to very different environments in terms of climate and food.
Here, we reconstruct the relationship among 69 animals from 6 island populations and events that shaped those populations' histories, using genome-wide SNP data. In particular, we are trying to disentangle the footprints of selection from demographic events. To this end we will also identify genomic regions independently subjected to repeated selection in the separate populations. The data will also help to reconstruct history of colonization and identify the source populations.

P-2048

Polymorphism Pattern of a MITE Locus Downstream of the Domestication Gene *Teosinte-Branched1* in Wild and Cultivated Pearl Millet

Yann Dussert[1], Marie-Stanislas Remigereau[2], Fontaine Michaël[1,3], Alodie Snirc[1], Ghayas Lakis[1], Solenn Stoeckel[4], Thierry Langin[5], Aboubakry Sarr[1], Thierry Robert[1]
[1]Ecologie, Systématique et Evolution, UMR 8079, Univ. Paris Sud, Orsay, France, [2]University of Southern California Molecular & Computational Biology, Los Angeles, CA, USA, [3]Eco-Anthropologie et Ethnobiologie, UMR 7206 CNRS, MNHN, Univ. Paris Diderot, Paris, France, [4]INRA, UMR 1099, Biology of Organisms and Populations applied to Plant Protection, Le Rheu, France, [5]INRA, UMR 1095, Génétique, Diversité et Ecophysiologie des Céréales, Clermont-Ferrand, France

Unraveling the mechanisms involved in adaptation processes to understand plant morphological evolution is a challenging goal. For crop species, the genetic dissection of domestication syndrome traits and the identification of molecular causal polymorphisms are central to this issue. Pearl millet, a domesticated grass found in semi-arid areas of Africa and India, is an interesting model for addressing these questions: the cultivated form shares common derived phenotypes with other cereals (*e.g.* maize, sorghum or foxtail millet) such as a decreased ability to develop axillary branches and tillers in comparison to the wild phenotype. The orthologue of the maize gene *Teosinte-Branched1* in pearl millet (*PgTb1*) has likely been involved in tillering evolution during domestication. An insertion of a miniature inverted-repeat transposable element (MITE) of the *Tuareg* family has also been found in some accessions in the 3'-untranslated region of *PgTb1*. For a set of 35 wild and cultivated populations, we compared the polymorphism pattern at this MITE element (for 1357 individuals) and at presumably neutral microsatellite loci (for 1068 individuals). Population genetic structure of wild and domesticated pearl millet was assessed on the basis of microsatellite data using a Bayesian approach. The *Tuareg* insertion was nearly absent in the wild populations throughout their distribution area, whereas a strong longitudinal frequency cline was observed in the cultivated populations. The geographic pattern revealed by neutral microsatellite loci clearly demonstrated that isolation by distance does not account for the existence of the cline in domesticated populations. However, analyses of the nucleotide polymorphism in the region downstream of *PgTb1* did not show evidence that the cline at the MITE locus has been shaped by selection, suggesting the implication of a neutral process. Alternative hypotheses are discussed.

P-2049

Assessing the molecular evolution of the invasive bivalve *Corbicula fluminea* in Portugal

Cidália Gomes[1,2], Vítor Vasconcelos[1,3], Lúcia Guilhermino[1,2], Agostinho Antunes[1,3]
[1]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal, [2]ICBAS - Instituto de Ciências Biomédicas de Abel Salazar, Departamento de Estudos de Populações, Laboratório de Ecotoxicologia, Porto, Portugal,[3]FCUP - Faculdade de Ciências da Universidade do Porto, Departamento de Biologia, Porto, Portugal

The *Corbicula fluminea* is an Asian non-indigenous invasive species that has rapidly spread worldwide in freshwater and estuarine environments. It is considered one of the most important invasive bivalves. This invasive species was detected in two Portuguese estuaries namely, the Minho and Lima rivers (NW of Iberian Peninsula) in 1989 and 2002, respectively. Even though the population density of *C. fluminea* in the Lima river is confined to a reduce area in comparison to the Minho river, previous research, indicated that these two populations are morphologically distinct but have the same mitochondrial (mt) DNA *COI* haplotype FW5. The observed ecophenotypic shell variation has been attributed to an adaptation to distinct ecological conditions, different origins and/or genetic alterations during distinct migration, or differential selection processes in the two rivers. Knowledge on the genetic variability of nuclear DNA (nDNA) markers between populations of *C. fluminea* in the Minho and Lima rivers is absent, which raises the question of how concordant is the information provided by the mitochondrial genome and the nuclear genome in these populations. Therefore, to better understand the genetic processes that may contribute to the invasive behaviour of this species, the genetic variability between Minho and Lima rivers populations of *C. fluminea* was assessed using a multi-genic approach employing mtDNA and microsatellite markers. Detailed population genetics analyses provided insight to understand the invasive response of this species in Portugal. Future studies would be important to determine how the observed differentiation patterns may contribute to the invasive potential of the *C. fluminea*.

P-2050

Genetic variation in *Drosophila subobscura*: a multilocus analysis along chromosome J

Roser Pratdesaba, Carmen Segarra, Montserrat Aguadé
Universitat de Barcelona, Barcelona, Spain

Drosophila subobscura is a member of the obscura group that is distributed in a wide area of the paleartic region and presents a rich chromosomal polymorphism. Inversion polymorphism affects nucleotide variation at loci associated with inversions. These effects depend on the age of the inversion and the rate of genetic exchange between inverted and standard chromosomes. New adaptive inversions as a result of their rapid increase in frequency, show a depletion of variation. The gradual recovery of variation by mutation leads to genetic differentiation between inverted and standard chromosomes, given the strong reduction of recombination along the inversion loop. Nevertheless, genetic exchange by gene conversion or double crossover might contribute to erode this genetic differentiation. In the present study, we surveyed nucleotide variation in non-coding regions distributed along the J chromosome in a population of *Drosophila subobscura* from the outskirts of Barcelona. Given the presence of a small polymorphic inversion in this chromosome, we analyzed separately levels and patterns of variation in regions inside and outside this inversion.

P-2051

Signs for balancing selection in S. paradoxus

Na Gao, Martin J. Lercher
Institute for Computer Science Heinrich-Heine-University Duesseldorf, Duesseldorf, Germany

To detect signs of population-specific or balancing selection, one can use variants of FST, which compare diversity within and between populations. To assess statistical signficance of non-neutrality, Bayesian methods can be employed that estimate FST as a combination of population-specific and locus-specific factors. One popular such approach is implemented in BayeScan (CITATION).

Applying BayeScan to genomic data from 25 Saccharomyces paradoxus strains (CITATION), we found a surprisingly large fraction of genes apparently under balancing selection. However, we also observed a strong correlation between gene length and the FST coefficient estimated by BayeScan for haplotype blocks defined a priori as complete genes. This correlation is an artifact, caused by the fact that longer genes on average contain more polymorphic sites and thus more haplotypes.

To circumvent this problem, we calculated FST for individual SNPs and for haplotype blocks defined by sliding windows. We thus identified a small number of genes with highly significant signs of balancing selection. The strongest signal was found for spar435-g2.1, which is orthologous to the S. cerevisiae genes YPL277C and YOR389W. All three are proteins of unknown function. Based on protein-protein interaction data, we conclude that YPL277C and YOR389W interact with type I myosins, as does spar435-g2.1.

However, spar435-g2.1 shows very high sequence similarity to spar_93-g6.1, a likely gene duplicate. Thus, polymorphism patterns may be influenced by assembly errors that wrongly assign nucleotide variants of one copy to its duplicate. We test this possibility by careful manual inspection of sequencing read assembly and alignment.

P-2052

The role of epistasis on the response to selection

Tiago Paixao, Nick Barton
IST Austria, Klosterneuburg, Austria

Quantitative genetics largely abstracts the genetic basis of the phenotype either by assuming that the distribution of breeding values is Gaussian or that genes act independently. However, gene interactions seem to be pervasive in real organisms. This “physiological” or “functional” epistasis makes the effect of an allelic substitution dependent on the background. Because over the course of a response to selection the mean background will change, as different alleles increase or decrease in frequency, the effect of an allelic substitution will also change. In particular, the nature (positive or negative) of the effect of an allelic substitution can change over the course of a response. This allows for the possibility of a “prolonging” of the response in the sense that exhaustion of genetic variation will occur at a slower pace.

Here we address the relation between epistasis at the trait level and the dynamics of the response to selection. We consider adaptation of sexual populations from standing genetic variation and under different population genetic regimes. We use individual based simulations to show how departures from additive models of interaction affect the dynamics of additive genetic variance, and pay special attention to cases where the nature of the effect of an allelic substitution changes over the course of the response.

P-2053

The WFDC locus: a preferred target of natural selection in humans

Zelia Ferreira[1,2], Susana Seixas[2], Aida Andres[1], Warren Kretzschmar[1], Jim Mullikin[4], Pedro Cruz[4], Praveen Cherukuri[4], Eric Green[1,4], NISC Comparative Sequencing Program[1,4], Belen Hurle[1]
[1]National Human Genome Research Institute (NHGRI), Bethesda, Maryland, USA, [2]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal, [3]Department of Zoology and Anthropology, Faculty of Sciences, University of Porto, Porto, Portugal, [4]NIH Intramural Sequencing Center (NISC), Bethesda, Maryland, USA

The whey acidic protein (WAP) four-disulfide core domain (*WFDC*) gene locus on human chr20q13 spans 19 genes with WAP and/or Kunitz domains that confer serine protease inhibitor activity. *WFDC* genes have evolved to exhibit core functions involving antimicrobial, immune and tissue homoeostasis activities that in most cases remain poorly understood. WFDC-related genes also include nearby genes encoding seminal proteins Semenogelin 1 and 2 (SEMG1 and SEMG2). At ejaculation, SEMG proteins form a transient sperm-entrapping gel, later hydrolyzed by the Prostate Specific Antigen (PSA). *WFDC* and *SEMG* genes stand out for their concomitant key roles in reproduction and innate immunity, and because of reports of striking adaptive pressures acting on them during vertebrate evolution, as measured by their high K_A/K_S values. Our goals are to understand the selection pressures acting on *WFDC* genes in human populations based on large-scale polymorphism data, and to functionally validate candidate variants driving the selective signals.

To catalog the genomic variation across the human *WFDC* locus, 17 genes and 54 non-coding tags were sequenced in 73 European (CEU), African (YRI) and Asian (CHB+JPT) HapMap individuals. A set of 47 neutrally evolving loci (pseudogenes) was also surveyed to assess baseline genomic diversity. Overall, we generated ~15 Mb of high-quality sequence and identified 765 Single Nucleotide Polymorphisms (SNPs), including 74 coding mutations, of which 54 were non-synonymous substitutions. Using classic neutrality tests (Tajima's D and Fay and Wu's H), we confirmed a signature of short-term balancing selection on *WFDC8* in the CEU population; and we pinpointed a signal of positive selection signature spanning genes *PI3*, *SLPI*, *SEMG1* and *SEMG2.* Associated to the latter signal and exclusive to Asians, we identified an unusually homogeneous haplotype with a high frequency of 88%. Based on the strong antimicrobial activity of SEMG1 peptides, we propose Thr56Ser in SEMG1 as the candidate variant responsible for the selective footprint. Specifically, we hypothesize that Ser56 modifies the likelihood of PSA hydrolyzing nearby PSA-cleavage site Tyr63, which in turn could alter the peptide profile and antimicrobial activities of semen in Asian populations. Ancestral Thr56 was traced to the last common ancestor of SEMG1 in Old World Monkeys and hominoids ~25 million years ago (MYA) - with derived allele Ser56 arising less than 0.25 MYA in Asians.

This study provides further evidence that the *WFDC*s and *SEMG*s are under strong adaptive pressures, evolving in ways that are perceptible within the short timescale of modern humans.

P-2054

MITOCHONDRIAL EVOLUTION ON THE SUB-ANTARCTIC KERGUELEN ARCHIPELAGO: POSITIVE SELECTION OR NEUTRAL EVOLUTION?

Emilie Hardouin, Diethard Tautz
Max Planck Institute for Evolutionary Biology, Plön, Germany

Mitochondrial DNA (mtDNA) is a commonly used marker in population genetics studies as it is relatively easy to amplify and highly variable in natural populations. Until recently, it has been accepted that it has clonal inheritance, a constant mutation rate and evolves neutrally. However, several recent studies proposed that the mtDNA diversity pattern is largely influenced by natural selection. For example, it has been suggested, for the human population, that advantages mutations occur and even have been fixed in individuals living at higher latitude i.e. colder environment. These finding questions the relevance of the use of mtDNA as neutral markers for population history and phylogeography studies.

House mouse living on the Kerguelen Archipelago (48°25'-50°S; 68°27'-70°35'E) provides the opportunity to investigate mechanisms of early adaptation to a new environment. Indeed coming from Western Europe no more than 300 years ago, they had to adapt to extreme conditions such as cold climate, new food resources and a feral lifestyle. Interestingly, previous data on the mitochondrial D-loop sequences suggested an unusual high mutation rate or the putative presence of selective sweeps in the mitochondrial genome of sub-Antarctic mice (Hardouin et al. 2010). In order to understand better the evolutionary processes involved in the mitochondria evolution, 16 mitochondrial genomes from Kerguelen mice were sequenced and their mutational patterns were systematically investigated. We conclude that the patterns we see are well in line with neutral expectations.

Hardouin EA, Chapuis JL, Stevens MI, van Vurren JB, Quillfeldt P, Scavetta RJ, Teschke M, Tautz D: House mouse colonization patterns on the sub-Antarctic Kerguelen Archipelago suggest singular primary invasion and resilience against re-invasion. BMC Evolutionary Biology 2010, 10:325 doi:10.1186/1471-2148-10-325.

P-2055

Characterizing adaptively relevant alleles in wild mice (*Mus musculus* L.)

Natascha Hasenkamp, Diethard Tautz
Max Planck Institute for Evolutionary Biology, Plön, Germany

One of the main themes in evolutionary biology is to understand the molecular basis of adaptive changes in populations. Positive selection drives adaptive evolution by increasing the frequency of beneficial alleles in gene pools. This process always depends on environmental conditions, and can only happen if alleles of a gene are functionally distinct. However, to date little is known about the nature and magnitude of the allele-dependent functional variation that can be targeted by positive selection and about the geographic distribution of such events in the wild. We address these questions by focusing on the allelic variation of the gene *dmrt1* (*doublesex and mab-3 related transcription factor 1*) in mainly two wild-derived populations of the house mouse. This transciption factor is a highly conserved component of the sex-determination cascade in vertebrates and is required for testis differentiation in mice. Notably, despite the high degree of conservation, a positive selection event in a single mouse population was detected by a previous microsatellite screen and SNP chip data. Consequently, we conducted a microsatellite analysis in a window of 200 kb around the gene for an extended set of 9 geographically distinct, wild mouse populations. The results indicate that several populations experienced a selective event in this genomic region. Subsequent sequencing analysis of the coding region of *dmrt1* was conducted in the two main study populations that were analyzed in the original microsatellite screen. This interestingly revealed the presence of a nonsynonymous SNP in the second exon of the gene. The SNP is fixed in the sweep population, but polymorphic in the other one, and causes an exchange of the amino acids serine and asparagine. The role of this particular SNP as the driver of the sweep is additionally supported by the fact that so far no evidence for differential expression, copy number variation or alternative splicing of *dmrt1* was found among the two mouse populations. Therefore, further analyses are focusing on potential SNP status-dependent changes in target gene-specificity and target gene-activation.

P-2056

Robust estimates of heterozygosity from low coverage sequencing data

Katarzyna Bryc[1], Nick Patterson[2], David Reich[1,2]
[1]Harvard Medical School, Boston, MA, USA, [2]Broad Institute, Cambridge, MA, USA

Heterozygosity, or the fraction of nucleotides within an individual that differ between the chromosomes they inherit from their parents, is a crucial number for understanding human variation. Estimating this simple statistic from any type of sequence data is confounded by sequencing errors, mapping errors, and imperfect power for detecting polymorphisms. Obtaining an unbiased estimate is especially difficult for ancient genomes (such as Neandertal or Denisova) where the sequences have a higher error rate, for low-coverage sequence data where there is low power to detect heterozygous sites, and for hybrid capture where there may be additional biases due to the oligonucleotides used for fishing out sequences of interest.

We present a method that estimates the heterozygosity for an individual of interest by leveraging the genome-wide joint information across sequence reads from a panel of individuals. We use an Expectation-Maximization (EM) algorithm to estimate the most likely distribution of counts across the unknown underlying genotypic states, from which we obtain an estimate of the proportion of loci that are heterozygous in the target individual. An advantage of this method is that it returns an unbiased and accurate estimate of heterozygosity even when the individual has low sequence coverage. Our method learns the distribution of alleles directly from the sequence read data, and does not require modeling demographic relationships among the individuals nor genotype calls from the sequence reads. We validate our EM method on 1 Gb of simulated sequence data of 5X, 10X, and 20X coverage, and find that our method performs well at estimating the true heterozygosity even when the sequence error rate is extreme and mean coverage is low.

We apply our method to estimate heterozygosity for the genomes of 11 modern humans from geographically dispersed populations, sequenced to a mean depth of 18.8X-28.4X. We obtain estimates that are consistent with previous results, and match the expected loss of heterozygosity following subsequent OOA bottlenecks. Stratifying the genomes by regional coverage, we find estimates of heterozygosity are consistently elevated for regions of the genome with higher or lower coverage relative to the individual's mean coverage. We examine systematic differences in these regions and posit that strong biases by GC content, lack of alignment of reads due to high divergence from the reference, and segmental duplications may all strongly influence estimates of heterozygosity.

P-2057

Population structure and evidence of selection in the Khoe-San and Coloured populations from southern Africa

Carina Schlebusch[1], Pontus Skoglund[1], Per Sjödin[1], Lucie Gattepaille[1], Sen Li[1], Flora Jay[2], Dena Hernandez[3], Andrew Singleton[3], Michael Blum[2], Himla Soodyall[4,5], Mattias Jakobsson[1]
[1]Uppsala University, Uppsala, Sweden,[2]Université Joseph Fourier, Grenoble, France,[3]National Institute on Aging (NIH), Bethesda, USA,[4]University of the Witwatersrand, Johannesburg, South Africa,[5]National Health Laboratory Service, Johannesburg, South Africa

The San and Khoe people currently represent remnant groups of a much larger and widely distributed population of hunter-gatherers and pastoralists who had exclusive occupation of southern Africa before the arrival of Bantu-speaking groups in the past 1,200 years and sea-borne immigrants within the last 350 years. Mitochondrial DNA, Y-chromosome and autosomal studies conducted on a few San groups revealed that they harbour some of the most divergent lineages found in living peoples throughout the world.

We used autosomal data to characterize patterns of genetic variation among southern African individuals in order to understand human evolutionary history, in particular the demographic history of Africa. To this end, we successfully genotyped ~ 2.3 million genome wide SNP markers in 220 individuals, comprising seven Khoe-San, two Coloured and two Bantu-speaking groups from southern Africa. After quality filtering, the data were combined with publicly available SNP data from other African populations to investigate stratification and demography of African populations. We also applied a newly developed method of estimating population topology and divergence times. Genotypes and inferred haplotypes were used to assess genetic diversity, patterns of haplotype variation and linkage disequilibrium in different populations.

We found that six of the seven Khoe-San populations form a common population lineage basal to all other modern human populations. The studied Khoe-San populations are genetically distinct, with diverse histories of gene flow with surrounding populations. A clear geographic structuring among Khoe-San groups was observed, the northern and southern Khoe-San groups were most distinct from each other with the central Khoe-San group being intermediate. The Khwe group contained variation that distinguished it from other Khoe-San groups. Population divergence within the Khoe-San group is approximately 1/3 as ancient as the divergence of the Khoe-San as a whole to other human populations (on the same order as the time of divergence between West Africans and Eurasians). Genetic diversity in some, but not all, Khoe-San populations is among the highest worldwide, but it is influenced by recent admixture. We furthermore find evidence of a Nilo-Saharan ancestral component in certain Khoe-San groups, possibly related to the introduction of pastoralism to southern Africa.

We searched for signatures of selection in the different population groups by scanning for differentiated genome-regions between populations and scanning for extended runs of haplotype homozygosity within populations. By means of the selection scans, we found evidence for diverse adaptations in groups with different demographic histories and modes of subsistence.

P-2058

Impacts of life-style on human evolutionary history: A genome-wide comparison of herder and farmer populations in Central Asia

Michael C. Fontaine[1,2], Laure Segurel[2,3], Christine Lonjou[4], Tatiana Hegay[5], Almaz Aldashev[6], Evelyne Heyer[2], Frederic Austerlitz[1,2]
[1]Ecology, Systematics & Evolution. UMR8079 Univ. Paris Sud - CNRS - AgroParisTech, Orsay, France, [2]Eco-Anthropologie et Ethnobiologie, UMR 7206 CNRS, MNHN, Univ Paris Diderot, Sorbonne Paris Cité, Paris, France, [3]Department of Human Genetics, University of Chicago, Chicago, USA, [4]C2BiG (Centre de Bioinformatique/Biostatistique Génomique d'Ile de France), Plateforme Post-génomique P3S, Hôpital Pitié Salpêtrière, Paris, France, [5]Uzbek Academy of Sciences, Institute of Immunology, Tashkent, Uzbekistan, [6]Institute of Molecular Biology and Medicine, National Center of Cardiology and Internal Medicine, Bishkek, Kyrgyzstan

Human populations use a variety of subsistence strategies to exploit an exceptionally broad range of habitats and dietary components. These aspects of human environments have changed dramatically during human evolution, giving rise to new selective pressures. Here we focused on two populations in Central Asia with long-term contrasted lifestyles: Kyrgyz's that are traditionally nomadic herders, with a traditional diet based on meat and milk products, and Tajiks that are traditionally agriculturalists, with a traditional diet based mostly on cereals. We genotyped 93 individuals for more than 600,000 SNP markers (Human-660W-Quad-V1.0 from Illumina) spread across the genome. We first analysed the population structure of these two populations in the world-wide context by combining our results with other available genome-wide data. Principal component and Bayesian clustering analyses revealed that Tajiks and Kirgiz's are both admixed populations which differed however from each other with respect to their ancestry proportions: Tajiks display a much larger proportion of common ancestry with European populations while Kirgiz's share a larger common ancestry with Asiatic populations. We then examined the region of the genome displaying unusual population differentiation between these two populations to detect natural selection and checked whether they were specific to Central Asia or not. We complemented these analyses with haplotype-based analyses of selection.

P-2059

On the move: sex-biased migration during the Neolithic transition

Rita Rasteiro[1,2], Pierre-Antoine Bouttier[1,3], Vítor C. Sousa[1,4], Lounès Chikhi[1,5]
[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal, [2]Department of Genetics, University of Leicester, Leicester, UK, [3]Laboratoire Jean Kutzman, Université de Grenoble, Grenoble, France, [4]Department of Genetics, Rutgers University, New Jersey, United States Minor Outlying Islands, [5]CNRS, Université Paul Sabatier, ENFA, UMR 5174 EDB, Toulouse, France

The Neolithic transition is probably the most important cultural, economic and demographic revolution in human prehistory. It profoundly modified the distribution of human genes, languages and cultures worldwide. Archaeological and anthropological data suggest that changes in post-marital residence rules between males and females took place as a consequence of sedentism and new rules of land control by men. However, very few studies address this change from a genetic viewpoint.

We developed a new individual-based simulation approach to explore the genetic consequences of 45 different scenarios, where we varied the patterns of post-marital residence and admixture between hunter-gatherers and farmers. We recorded mtDNA and Y-chromosome data and analysed their diversity patterns within and between populations, through time and space. We also collected published mtDNA and Y-chromosome data from European and Near-Eastern populations in order to identify the scenarios that would best explain them. ´

Our results show that the Neolithic transition must have left its mark in the genome of Europeans and we confirm that farming was accompanied by reduced male migration and a movement of females to their husband's birthplace.

P-2060

Demographic inference and scans for selection in 28 sequenced gorillas

Jeffrey Kidd[1], Joanna Kelley[2], Laurie Stevison[3], August Woerner[4], Laurel Johnstone[4], Michael Hammer[4], Jeff Wall[3], Carlos Bustamante[2], Great Apes Genome Diversity Consortium[5]
[1]University of Michigan, Ann Arbor, MI, USA,[2]Stanford University, Stanford, CA, USA,[3]University of California, San Francisco, San Francisco, CA, USA,[4]University of Arizona, Tucscon, AZ, USA

The genetic diversity and recent population history of the Gorilla remains largely unexplored using genome sequence data. We have conducted an population genomic analysis of this species based on 6-20x full genome sequence data from 28 wild-caught Gorillas including 25 Western (Gorilla gorilla gorilla) and 3 Eastern (Gorilla beringei graueri). Following standard mapping and variant calling procedures we identified over 23 million single nucleotide polymorphisms. We use the resulting site frequency spectra to make inferences about the recent history of these populations and to conduct scans for signatures of recent selection in the gorilla genome. It is increasingly recognized that background selection on linked deleterious variants and non-equilibrium demographic processes such as bottlenecks and rapid population growth complicate the interpretation of identified signatures of selective sweeps. Comparison of selected loci between and among non-human primate species that have experienced different demographic histories, exist in diverse environments, and live dissimilar life styles offers guidance for the interpretation of signatures of recent selection.

P-2061

How often does natural selection targets multiple, interacting genes? The prevalence of epistasis in recent human evolution

Natalia Petit[1], Arcadi Navarro[1,4]
[1]Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Barcelona, Spain,[2]Departament de Ciències Experimentals i de la Salut (DCEXS).Universitat Pompeu Fabra, Barcelona, Barcelona, Spain,[3]National Institute for Bioinformatics (INB), Population Genomics Node, Barcelona, Barcelona, Spain,[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Barcelona, Spain

Epistasis in its broadest sense could be defined as the dependence of the outcome of a mutation on its genetic background. Mutations in functionally related genes could be selected jointly if they jointly affect the total fitness of individuals carrying them. Whenever this happens, the signature that selection may leave in patterns of genome variability might differ dramatically from the signature of selection when it acted upon a single gene. An excellent case-study is provided by ancestral and current human populations, which experienced significant changes in their selective pressures after leaving Africa to colonize the whole Planet. In this study, we studied derived human populations looking for evidence of recent positive selection acting upon pairs of protein-coding genes that are known interact within pathways ("interacting genes"). Evidence of recent positive selection events was investigated using Tajima's D and other frequency-spectrum statistics. We found an excess of interacting pairs of genes with evidence of recent positive selection in derived populations when compared with African populations. These observations cannot be accounted for neither by (1) the demographic history of populations nor by (2) an excess of outlier genes in derived relative to African populations. Further analyses showed that the pairs of outlier interacting genes tend to be more central in the networks than the rest of outliers genes, and that this trend is stronger in derived than in African populations. Taken together, our results suggest that changes in selective pressures in human derived populations affect groups of interacting genes.

P-2062

Estimating selection coefficients in structured populations

Iain Mathieson, Gil McVean
University of Oxford, Oxford, UK

Many traits in both humans and other natural populations are thought to be under spatially varying selection arising from, for example, exposure to pathogens or other environmental factors. In some cases the selective pressure is well understood, for example selection on skin colour or for malaria resistance. In other cases, such as the CCR5Δ32 HIV resistance allele, or the DRB15*01 MS risk haplotype, the selective pressure is less certain. In general we might think that the recent expansion of modern humans into new environments will have created exposure to new and spatially varied selective pressures. However, characterizing the spatial distribution of selective pressure from patterns of genetic variation alone is potentially confounded by the need to jointly infer demographic processes. To understand the limits of our ability to learn about structured selection pressures we consider evolution in a simple lattice model of population structure and measure the amount of information available about underlying parameters from different possible sources of information.

We first demonstrate that for a single deme, if the entire allele frequency trajectory is known then, conditional on eventual fixation of the allele, the Wright-Fisher maximum likelihood estimator (MLE), which is just one divided by half the total heterozygosity, converges rapidly to the true selection coefficient s. Specifically we show that the inverse of the MLE converges to s^{-1} with error $O(N^{-1}s^{-2})$ where N is the effective population size. In the structured case we allow s to vary across demes, and the equivalent estimator for s has an elegant form, which is just the single-deme estimator plus a correction which is proportional to the migration rate. If we don't observe the whole trajectory, but only an interval or a discrete set of points, we can use a similar estimator, which still converges rapidly as N increases.

Of course in practice, and certainly for humans, we do not have this much information to base inference on. With only a single time-point, we must estimate the history of an allele by building genealogies. Further complicating estimation, in a structured population this process is confounded by migration. We will quantify the amount of information about s available when we have different types of data, and discuss the limits on our ability to estimate it accurately.

P-2063

Approximate Bayesian Computation Sheds Light on the Complex Demographic History of Orangutans (genus: *Pongo*)

Alexander Nater, Maja Greminger, Natasha Arora, Carel van Schaik, Michael Krützen
Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland

Investigating how different evolutionary forces have shaped patterns of DNA variation within and among extant species requires detailed knowledge of demographic history, as these patterns are the result of demographic, selective and random processes. Reconstructing demographic history from genetic data poses a significant challenge, as many species have experienced complex demographic scenarios throughout their evolutionary history. Investigating such scenarios requires parameter-rich models, which are computationally extremely intensive when applied to large data sets. Orangutans, whose distribution is currently restricted to the Southeast Asian islands of Borneo (*Pongo pygmaeus*) and Sumatra (*Pongo abelii*), have likely experienced a complex demographic history, influenced by fluctuating sea levels, climate changes and volcanic activity. Previous studies have tested simplified demographic models, often using a small number of samples with unknown geographic origin. Thus, it remains unknown to what extent unrepresentative population sampling, population substructure and oversimplified models have led to misleading conclusions. Approximate Bayesian Computation (ABC) allows testing complex demographic models using different types of genetic markers in a combined analysis. In our ABC approach, we employed the most extensive genetic data set of wild orangutans to date, including autosomal and Y-chromosomal microsatellite genotypes, as well as autosomal, X-chromosomal and mitochondrial sequence data. We tested the fit of 8 different demographic models, which we designed based on current knowledge of orangutan genetics and behavioral ecology. We found that a complex demographic model, incorporating population substructure within Bornean and Sumatran orangutans, as well as a recent bottleneck on Borneo, best explains the currently observed patterns of genetic variation. Based on this demographic model, we estimated that the two orangutan species diverged ~0.9 Ma, with subsequent heavily male-biased migration between the two islands until ~110 kya. Sumatran orangutans showed a deep split of populations north and south of Lake Toba, probably caused by multiple eruptions of the Toba volcano. On Borneo, orangutans experienced a severe bottleneck ~60 kya, followed by a population expansion and substructuring ~20 kya, which we link to a refugium during the last glacial period. Thus, we show that orangutans, like other non-human great apes, went through drastic changes in population size and connectedness caused by recurrent contraction and expansion of rainforest habitat during Pleistocene glaciations. Our results contrast with a previous study, which tested simplified demographic models on genomic data. This discrepancy demonstrates that caution has to be exerted when using oversimplified demographic models and potentially inappropriate sampling schemes to reconstruct demographic history.

P-2064

Selective processes affecting candidate genes in contrasting adaptive chromosomal inversions

Marta Pascual, Pedro Simoes, Gemma Calabria, Cinta Pegueroles, Francesc Mestres, Joan Balanya
Universitat de Barcelona, Barcelona, Spain

There is compelling evidence supporting an adaptive explanation for the evolution of inversion polymorphisms in *Drosophila subobscura.* They present latitudinal frequency clines and vary in frequency in response to global warming. Sequencing candidate genes for thermal adaptation mapping inside and outside the inversions can help identify regions that might be under selection. We have analysed four genes located in the sex chromosome and four genes in the O chromosome (autosome) in several European populations distributed latitudinally for the most frequent arrangements. Strong differentiation is found between inversions with fixed nucleotide differences for some genes located inside inversions. However some genes located outside inversions also presented strong differentiation showing their linkage with inversions. Despite the recombination reduction mediated by inversions, recombinant sequences (either by gene conversion or crossover) were detected in almost all genes and populations for most arrangements. Generally individuals bearing the same arrangement presented no genetic differentiation between populations, with some exceptions. We will discuss these results considering the different hypotheses explaining the mechanisms of adaptive chromosomal polymorphism.

P-2065

A new linkage disequilibrium based method for estimating effective population size

Lucie Gattepaille, Mattias Jakobsson
Uppsala University, Uppsala, Sweden

The effective population size (Ne) is an essential concept in population genetics. Inferring this mathematical and biological parameter is both challenging and of critical relevance, as it can give insights into the strength of adaptive evolution acting upon species as well as the demographic history of populations. Some methods for estimating Ne in natural populations have been proposed, based for example on nucleotide diversity, variance in allele frequencies over generations, linkage disequilibrium (LD) or recombination patterns. Here, we introduce a new method to infer Ne. The method relies on observed patterns of LD in populations. We simulate data from standard neutral coalescent models for different constant population sizes and recombination rates. For a given recombination rate, there is one population size that best fit the observed LD pattern, so uncertainty about recombination rate creates uncertainty on the estimate of Ne. The negative correlation existing between recombination rate and Ne is fitted by a linear model on their logarithmic values. By considering the ratio of Ne estimates to the Ne estimate of a reference population, and assuming that the recombination rate is the same across populations, we remove the dependence on recombination rate. Using data from 70 human populations worldwide, we apply our method to compute the Ne of each population relative to the Yoruban effective population size. We found that the Ne values averaged over continental regions are coherent with previous results, with a mean effective size, relative to Yoruban population size, of 1.1 for African populations, 0.45 for Middle-Eastern populations, 0.44 for Central/South Asian populations, 0.37 for European populations, 0.25 for East Asian populations, 0.12 for Oceanian populations and 0.07 for Native American populations. This method can be particularly useful in data sets containing ascertainment bias, or uncertainty about recombination rates. Finally, our investigation improve our understanding on how LD patterns can be influenced by the strength of genetic drift.

P-2066

Estimating population sizes using the coalescent with recombination

Sara Sheehan, Kelley Harris, Yun S. Song
UC Berkeley, Berkeley, CA, USA

Throughout history, the population size of modern humans has varied considerably due to events such as the out of Africa bottleneck, and has been further shaped by recent super-exponential growth [1]. More accurate estimates of population size changes and when they occurred could provide a clearer picture of human colonization history and shed light on the concept of effective population size. Li and Durbin [2] recently estimated past population sizes by inferring coalescence times between a pair of haplotypes, but this approach is hampered by the fact that only few coalescence events occur in the very recent and very ancient past, thus impeding inference for those epochs. Genealogies of multiple (greater than 2) haplotypes contain larger numbers of very recent and very ancient coalescence events, and here we present a general hidden Markov model that can infer population sizes from this store of data.

Our work generalizes the sequentially Markov conditional sampling distribution (CSD) that was recently proposed by Paul et al. [3] The CSD describes the probability of observing a newly sampled haplotype given a set of previously sampled haplotypes, and it allows us to compute the joint likelihood of multiple haplotypes as a product of approximate conditionals (PAC) [4]. Because the CSD proposed by Paul et al. was derived from the diffusion process dual to the coalescent and the construction admits a natural genealogical interpretation, it can be rigorously modified to incorporate past population size changes, and these size changes can be inferred within an expectation-maximization framework. Empirical results demonstrate that we can accurately reconstruct the true population size function, even for very recent and very ancient times.

[1] Coventry et al.
[2] Li and Durbin. Nature
[3] Paul, Steinrucken, and Song. Genetics.
[4] Li and Stephens.

P-2067

Long-term evolution of quantitative traits in finite populations.

Harold P. de Vladar, Nick Barton
IST Austria, Klosterneuburg, Austria

When polygenic quantitative traits are subject to selection, their response in a population may be complicated for several reasons, amongst which we highlight three factors: (1) the current allele frequencies of each gene influencing the trait are unknown, (2) the distribution of mutational effects is usually unknown, and (3) stochastic factors like genetic drift perturb the response. Consequently, approximations such as the infinitesimal model, fail. Thus, we seek approximations that only rely on readily measurable variables, such as the trait mean and its variance. Previous approximations that employed this rationale failed because they required complete knowledge of all the moments of the distribution of the trait. However, previously we devised a methodology that employs only a small set of these measurable quantities. In our method, instead of attempting to track a single trajectory, we average over genetic drift (i.e. an ensemble of populations), and track in time the expectation of the measureable quantities. We achieved this for general cases of directional selection and of stabilizing selection but assuming equal effects. We have now extended our method to include arbitrary additive effects for a trait under stabilizing selection, as well as pleiotropic factors induced by multivariate selection. Our method can predict the long-term response to selection, including that of the genetic variance, which may change through time. The approximations, which assume Hardy-Weinberg and linkage equilibria, are surprisingly good, at least compared with simulations, and fall within the expected variability due to drift. We address classic questions such as: how does the rate of adaptation depend on the distribution of genetic effects? When, and why does genetic variance remain constant during the adaptive process? In addition, we also point other relationships amongst the quantitative quantities that were not previously sought, which exploit properties of the ensemble, but which are measurable from the populations. We argue that these might be the basis for new ways to understand how selection acts on complex traits.

P-2068

Lizards and LINEs: Host Demography and Selection Affect the Fate of non-LTR Retrotransposons in the *Anolis* Genome

Marc Tollis[1,2], Stephane Boissinot[1,2]
[1]*Queens College, City University of New York, Flushing, NY, USA,* [2]*The Graduate Center, City University of New York, New York, NY, USA*

Anolis carolinensis, or the green anole lizard, is the first lepidosaurian reptile to have its entire genome sequenced. As a species that is widespread and abundant in the southeastern United States, the *Anolis* genome offers an opportunity to study the population-level forces that shape genome evolution. Non-LTR retrotransposons, or LINEs, have significantly impacted the genomes of their hosts and account for much of the variability in genome size and structure among vertebrates. Possibly over 50% of the human genome is comprised of hundreds of thousands of copies of a single LINE clade known as L1. The lizard genome is more similar to teleost fish, containing a wider diversity of LINEs with relatively fewer copies. The differences in abundance and diversity of LINEs suggest that the vertebrates interact quite differently with their intragenomic parasites. In order to determine the strength of selection against harmful alleles, it is important to first determine the evolutionary history of populations since demography can affect the efficiency of selection against deleterious alleles. To this end, we performed a phylogeographic analysis of the green anole using one mitochondrial and 10 nuclear loci. Green anoles are a surprisingly polytypic species, as we describe four evolutionarily distinct lineages, the divergences of which date to the early-to-mid Pleistocene. All lineages are characterized by population size expansions, and one in particular has more recently expanded westward across the Gulf Coastal Plain. We demonstrate that LINEs are polymorphic in green anole populations, and that full-length elements (FL) are found at lower population frequencies that truncated (TR) ones. This suggests that the strength of selection against LINEs in *Anolis* is related to element length and their ability to mediate ectopic recombination. We further found significant differences in the frequency of LINE insertions among populations, supporting the hypothesis that the fate of transposable elements in a host genome depends on a selection-drift balance.

P-2069

Assessing the evolutionary history of Theobroma cacao from resequenced genomes

Omar Cornejo[1], Keithanne Mockaitis[2], Muh-Ching Yee[1], Stefan Royaert[3], Donald Livingstone III[3], Ram Podicheti[2], David Kuhn[3], Carlos Bustamante[1], Juan Carlos Motamayor[3]
[1]Stanford University, Stanford, CA, USA,[2]Indiana University Center for Genomics and Bioinformatics, Bloomington, IN, USA,[3]USDA ARS SHRS, Miami, FL, USA

Increased access to DNA sequence and genotype data has created an opportunity to answer questions regarding the origin of domestication and the demographic history of plant and animal species, for example in rice and dogs. In the case of cacao, the chocolate tree, recent work based on microsatellite markers has provided insights into the geographic and genetic differentiation of cacao. Here, we report preliminary results from 17 fully re-sequenced genomes of cacao, *Theobroma cacao* L. Our analyses of more than 6.3 million bi-alellic single nucleotide polymorphisms (SNPs) across samples allowed us to examine a variety of demographic models and revealed that cacao populations underwent an ancient reduction in population size. We highlight several of the challenges associated with the inference of demographic history in highly heterozygous systems. We discuss the relevance of accounting for the demographic history of cacao while attempting to discover the genetic basis for adaptation during its domestication. Specifically, we discuss how taking into account the demographic history of cacao will allow us to generate adequate null models to identify regions under putative selection and understand the basis of adaptation during cacao domestication.

P-2070

Elucidating the history of a rare spruce species endemic to Taiwan using Approximate Bayesian Computation

Sofia Bodare, Michael Stocks, Martin Lascoux
Uppsala University, Uppsala, Sweden

Picea morrisonicola **is the most southernly distributed spruce species and is endemic to Taiwan. The species is listed as vulnerable by the IUCN and faces an uncertain future that comes partly from logging that occurs in the region, but also from a changing climate.** ***P. morrisonicola*** **occurs mainly in the higher altitudes of the island, so any increase in temperature decreases the distribution of the species further. To assess the evolutionary history and potential fate of this species we re-sequenced nuclear loci in individuals from a population from Taiwan. Compared to previously studied spruce species from the northern hemisphere, nucleotide diversity in** ***P. morrisonicola*** **is considerably lower. The species exhibits levels of genetic diversity that are similar to other spruce species with more restricted distributions such as** ***P. breweriana*** **and** ***P. schrenkiana*****. Additionally, using an Approximate Bayesian Computation approach, we show that the species has suffered a decrease in the effective population size, giving rise to the levels of genetic variation we observe today.**

P-2071

Evolutionary and functional evidence for positive selection at the human *CD5* immune receptor gene

Elena Carnero-Montoro[1], Lizette Bonet[2], Johannes Engelken[1,3], Mario Martínez-Florensa[2], Francisco Lozano[2,4], Elena Bosch[1]
[1]Institut de Biologia Evolutiva (UPF-CSIC), Barcelona, Spain, [2]Institut d'Investigacions Biomèdiques August Pi i Sunyer, Centre Esther Koplowitz, Barcelona, Spain, [3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [4]Servei d'Immunologia, Hospital Clínic, Barcelona, Spain, [5]Departament de Biologia Cel·lular, Immunologia i Neurociències, Facultat de Medicina, Universitat de Barcelona, Barcelona, Spain

CD5 is a surface correceptor of T and B1-cells. Although its function is still poorly understood, it is known to participate in the adaptive immune system by modulating lymphocyte activation and/or differentiation and in the innate immune system by specifically recognizing conserved pathogen components such as fungal cell wall molecules. Previous SNP data analysis provided evidence for a recent selective sweep in East Asia and suggested a Val to Ala substitution at position 471 (A471V, rs2229177) of the CD5 cytoplasmic region as the most plausible target of selection. The present study further investigates the role of natural selection by fully describing the genetic variability at this locus and by functionally assaying the immunological differences among the two alleles at the A471V site. We have obtained *CD5* sequences of 60 individuals from 3 different populations (Africans, Europeans and East Asians). The sequence analysis revealed several lines of evidence for positive selection at the A471V site, namely i) high population differentiation for this site; ii) departure from neutrality as indicated by several tests (Tajima´s *D,* Fu and Li´s *F** and *D**, and Fu´s *F*s) and iii) the predominance of a major haplotype in East Asia linked to the V allele. Interestingly, MAPK kinase activity, cytokines release and lymphocyte proliferation measured on transfected mammals cells and/or on human peripheral blood mononuclear cells post-stimulus via CD5, T cell receptor (TCR) and B cell receptor (BCR) demonstrate functional differences between the two *CD5* variants at the A471V site and provide evidence for an adaptive role of the A471V substitution to environmental challenges in East Asian populations.

P-2072

Dispersal processes underlying the recent pandemic caused by the plant pathogenic fungus Mycosphaerella fijiensis

Jean Carlier[1], Stéphanie Robert[1], Adrien Rieux[1], Fabien Halkett[2], Marie-Françoise Zapater[1], Luc De Lapeyre de bellaire[1], François Bonnot[1], Catherine Abadie[1], Virginie Ravigné[1]
[1]CIRAD, UMR BGPI, Montpellier, France,[2]INRA, UMR 1136, Nancy, France,[3]CIRAD, UPR 26, Montpellier, France

How plant pathogenic fungi spread is the first question to consider for understanding the emergence of diseases caused by such organisms. Mycosphaerella fijiensis causing the black leaf streak disease of banana is an example of a recent pandemic in agriculture and a good model to address this question in the case of an aerial plant pathogen. The pandemic started around 1960 from the South-East Asia. Samples from various populations around the world at different geographical scales were analyzed using nuclear sequence-based and microsatellite markers. Demographic events (founder effects or admixture) were detected at global and continental scales following introductions of the disease. These introductions were more likely due to movement of infected plant materials. At lower scale, the structure of the M. fijiensis populations in two recently (~1979-1980) colonised areas in Costa Rica and Cameroon was analysed. Genetic differentiation and isolation by distance (IBD) were detected in both countries along a ~250-300km-long transect, suggesting continuous range expansion through gradual dispersal of spores over a few hundred kilometres. Furthermore, a discontinuity in gene frequencies was observed along the Cameroon transect. A landscape genetic study was recently conducted around this discontinuity. No landscape features matched the genetic discontinuity supporting it could result from a demographic event during the spread of M.fijiensis in the country rather than a physical barrier impeding contemporary gene flow. The genetic structure observed in M. fijiensis populations at different geographical scales has allowed a better understanding of dispersal processes in such an organism

P-2073

Climatic conditions during the Last Glacial Maximum affect the current genetic diversity of the European wild boar (*Sus scrofa*)

Sibelle Vilaca[1], Laura Iacolina[2], Daniela Biosa[2], Frank Zachos[2], Marco Apollonio[2], Massimo Scandura[2], Giorgio Bertorelle[1]
[1]Università di Ferrara, Ferrara, Italy, [2]Università di Sassari, Sassari, Italy

In many terrestrial species, the geographic distribution of DNA lineages was heavily affected by the climatic fluctuations that occurred during the Quaternary, although the impact of human populations in more recent times, especially on harvested species, might have confounded the pattern. The wild boar (*Sus scrofa*) is one the most widely distributed terrestrial animals, naturally occurring from Western Europe to Japan. Previous studies suggested that Iberia, Balkans and Italy played a major role as European refugia during the Last Glacial Maximum (LGM), and these three areas were the responsible for the recolonization of the continent. We tested this hypothesis using 770 mtDNA sequences (HVR) from 75 sites covering the entire European continent. Northern populations show lower genetic diversity when compared to southern populations. Surprisingly, Iberian and some Balkans populations share haplotypes that were only found in these two areas, while Italian populations are more similar to central European region (France and Germany). To predict the distribution of the wild boar during the LGM and its relationship with the current distribution of genetic diversity, a maximum entropy method based on 11 climatic variables was used. The model predicts that Iberia, Balkans and Italy had great habitat suitability during the LGM. Italy and Iberia shows greater suitability when compared to other regions of Europe, which can indicate that these two areas may have retained large population size even during cold periods. The suitability map and the current distribution of the genetic diversity show parallel geographic patterns, also within the non-homogeneous Iberian refugium. This result suggests that the climatic conditions in Europe during the LGM affected the current geographic pattern of genetic variation. The genetic similarity between some Western (Iberia) and some Eastern (Balkans) populations suggests that these areas were not isolated during the LGM, but this hypothesis requires additional testing.

P-2074

The effect of recurrent partial sweeps on patterns of neutral diversity

Graham Coop, Peter Ralph
Dept. of Evolution and Ecology, UC Davis, Davis, CA, USA

Two major sources of stochasticity in the dynamics of neutral alleles result from the sampling due to finite population size (genetic drift) and the random genetic background of selected alleles on which neutral alleles are found (linked selection). There is now good evidence that linked selection plays an important role in shaping polymorphism levels in a number of species. One of the best investigated models of linked selection is the recurrent full sweep model, in which newly arisen selected alleles fix rapidly. However, the bulk of selected alleles that sweep into the population may not be destined for rapid fixation in the species. Here we develop a coalescent model that generalizes the recurrent full sweep model to the case where selected alleles do not sweep to fixation. We show that in a large population, only the initial rapid increase of a selected allele affects the genealogy at partially linked sites, such that the subsequent fate of the selected allele often does not matter. We investigate the impact of recurrent partial sweeps on levels of neutral diversity, and show that for a given reduction in diversity, the impact of recurrent partial sweeps on the frequency spectrum at neutral sites is determined primarily by the frequency reached by these partial sweeps. Recurrent sweeps of selected alleles to low frequencies can have a profound effect on levels of diversity but can leave the frequency spectrum relatively unperturbed. Indeed we show that in the limit of a high rate of sweeps to low frequency, the resulting coalescent model is identical to the standard neutral model. This generalized model goes some way towards providing a more flexible framework to describe genomic patterns of diversity.

P-2075

Genome-Wide Data Reveals the History and Structure of Mexican Populations

Christopher Gignoux[1], Andres Moreno[2], Fouad Zakharia[2], Juan Carlos Fernández López[3], Simon Gravel[2], Patricia Ortiz-Tello[2], Eimear Kenny[2], Karla Sandoval[2], Martin Sikora[2], Alejandra Contreras[3], Victoria Robles[3], Sandra Romero Hidalgo[3], Rodrigo García Herrera[3], Alessandra Carnevale[3], Hector Rangel-Villalobos[4], Victor Acuña-Alonzo[5], Samuel Cañizales-Quinteros[6], Irma Silva-Zolezzi[3], Esteban González Burchard[1], Carlos Bustamante[2]
[1]UCSF, San Francisco, CA, USA,[2]Stanford University, Stanford, CA, USA,[3]National Institute of Genomic Medicine, Mexico DF, Mexico,[4]Universidad de Guadalajara, Guadelajara, Ocotlan, Mexico,[5]Escuela Nacional de Antropología e Historia, Mexico City, Mexico,[6]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico

The rich history of the populations of the Americas has been under-studied in both population and medical genetics. Here we investigate both early demographic history of the Americas as well as the mixed ancestry of the past 500 years from a diverse set of individuals sampled across Mexico. We generated Affymetrix 6.0 data for 492 Native American individuals from 18 different groups representing diverse geographic regions and linguistic affiliations, as well as 362 Mexican Mestizo, or admixed, individuals from 8 different states.

In first looking at Native American individuals, we detected structure with strong geographic patterning consistent with isolation by distance. To investigate the level of isolation across the country we implemented simulations to estimate current N_e values and the strength of the bottleneck into the Americas via approximate Bayesian computation. Current estimates of Ne tend to be small but vary greatly between populations: geographically isolated populations such as the Seri and Lacandon have MLE(N_e) 1,000 while larger groups such as the Nahuans have a MLE(N_e) above 3,000.

The Mestizo individuals received varying proportions of European and Native American ancestry, along with a smaller but detectable African component, consistent with prior studies. When we compare these individuals with hapmap reference panels, shared haplotype analysis demonstrates that the Native component of ancestry is not well captured by East Asians. To look at potential substructure within each ancestral component we developed an ancestry-specific extension of PCA to look at the ancestral components of admixed individuals within their ancestrally relevant PC context. When comparing the Native component of Mestizos with our Native American data, we observe a strong correlation in the Mestizo samples between ancestry-specific PCA values and geography, allowing us to make inference of geographic origin for Mestizo individuals based solely on their Native American ancestry tracts. European ancestry-specific PCA on the other hand shows a strong signal of Iberian ancestry for all individuals. We extend this method to larger studies of Mexicans both within Mexico as well as individuals currently residing in the USA.

To our knowledge this is the first study to combine fine-scale structure investigations of both Native Americans and Mestizos in Mexico using genome-wide data. The structure of Native American ancestry across the country in particular demonstrates a need for more sequencing and optimized array design in genetic studies.

P-2076

The genomic basis of local adaptations in orangutans (*Pongo* spp.)

Maja P. Greminger[1], Kai N. Stölting[2], Alexander Nater[1], Maria Anisimova[3,4], Giada Ferrari[1], Heidi E. L. Lischer[8], Remy Bruggmann[5,6], Benoit Goossens[7], Natasha Arora[1], Carel P. van Schaik[1], Michael Krützen[1]
[1]Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland, [2]Unit of Ecology and Evolution, Department of Biology, University of Fribourg, Fribourg, Switzerland, [3]Computational Biochemistry Research Group, Department of Computer Science, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, [4]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, [5]Functional Genomics Center, University of Zurich, Zurich, Switzerland, [6]Department of Biology, University of Bern, Bern, Switzerland, [7]School of Biosciences, Cardiff University, Cardiff, UK, [8]Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

One of the major tasks of evolutionary biology is to understand how adaptive variation in phenotypic traits evolves by natural selection and how it leads to speciation. However, we still have very limited understanding of the genetic basis of local adaptations in natural populations and the process of adaptive evolution and speciation at the proximate level. Among great apes, orangutans (genus: *Pongo*) represent a unique model to study the genetic basis of local adaptation between and within species, as they show remarkable systematic geographic variation in traits such as brain size, fat storage ability, male developmental arrest and social organization. To identify the genomic regions affected by natural selection in orangutans, we generated a large-scale SNP data set, focusing on two populations of the orangutan taxa with most pronounced differences in functionally linked traits, i.e. *P. abelii* (Sumatra) and *P. pygmaeus morio* (northeastern Borneo). We developed a new protocol to establish reduced representation libraries (RRLs) and sequenced >330,000 RRL stacks (~1.5% of the genome) for 18 individuals from each of the two populations as well as for at least one individual of all other major orangutan populations. We obtained >150,000 high quality SNPs of which 1.77% appear fixed between the Bornean and Sumatran populations. We detected signals of selective sweeps using sliding window outlier analyses. To disentangle patterns of selection from those produced by neutral events, we simulated confidence intervals under neutrality based on the demographic history of orangutans, as inferred by Approximate Bayesian Computation. We identified several hundred candidate genes and regulatory elements, a large fraction of which appears to be involved in immune defense, metabolic processes and nervous system development. To identify potential functional changes we took advantage of available whole genome data of ten orangutans. In addition, we used the whole genome data to identify genes under positive selection by codon-based modeling. The results of this study contribute to continued comparative analyses of primates and humans which in turn will ultimately provide further insights into human evolution and expand our understanding of the nature of intra-specific variation and evolution by natural selection in general.

P-2077

Chromosome-scale selective sweeps in Caenorhabditis elegans

Joshua Shapiro[1], Erik Andersen[1], Justin Gerke[1], Marie-Anne Félix Fèlix[2], Leonid Kruglyak[1,3]
[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA, [2]Institute of Biology of the Ecole Normale Superieure, Paris, France, [3]Howard Hughes Medical Institute, Princeton University, Princeton, New Jersey, USA. 4 Department of Molecular Biology, Princeton University, Princeton, NJ, USA

Despite *Caenorhabditis elegans* central role in molecular, cell, and developmental biology, there has until recently been only limited study of genomic and phenotypic variation in the species. In particular, we have lacked an unbiased genome-wide set of polymorphic loci for thoroughly exploring the evolutionary history of this partially selfing species. To provide an improved set of single nucleotide polymorphisms (SNPs) for analyses of *C. elegans* variation, we used high-throughput selective sequencing to identify a set of 41,188 SNPs from a worldwide collection of 200 wild strains. Our results define the currently available variation in the species, highlighting a set of highly diverged strains with large numbers of rare variants. Overall, polymorphism rates are strikingly non-uniform across each autosome, correlating strongly with variation in recombination rate. While much of this pattern can be explained by background selection, there is also clear evidence for the effects of positive selection. *C. elegans* genomic variation is dominated by a set of commonly shared haplotypes, each spanning many megabases and present throughout the world. This pattern was most likely generated by multiple strong, recent selective sweeps, with the effects amplified by the low outcrossing rate in the species. At least one of these sweeps probably occurred within the past few hundred years, and may reflect the effects of human activity.

P-2078

Demographic inference using the transition matrix of allelic types from diploid genomic sequences

Yong Wang, Kirk Lohmueller, Rasmus Nielsen
Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA

We developed a novel population genetic method for inferring demographic parameters from pairs of diploid genomes. Our method uses information from both linkage disequilibrium (LD) patterns and differences in allele frequency between populations. We summarized the genomic information by numbers of different transitions between allelic types. These allelic types contain information about the extent of population differentiation between the two genomes. The transition pattern from one allelic type to another allelic type moving along a chromosome is a measure of LD across the genome. Transitions between very different allelic types indicate that there is little LD in the genomic region. We tabulated the number of different transitions from a pair of genomes and applied standard coalescent simulation to estimate demographic parameters. Our method explicitly models variation in recombination rates including recombination hotspots, and further takes sequencing quality and missing data into account. We evaluated the method by simulations and applied the method to an Aboriginal Australian genomic sequence obtained from a 100-year-old lock of hair. We show that Aboriginal Australians are descendants of an early human dispersal into eastern Asia, around 62,000 to 75,000 years ago. We also find evidence of gene flow between populations of the two dispersal waves. Our findings support the hypothesis that present-day Aboriginal Australians descend from the earliest humans to occupy Australia, likely representing one of the oldest
continuous populations outside Africa.

P-2079

Charting the footprint of natural selection through the genome of *Drosophila melanogaster*

Antonio Barbadilla, David Castellano, Maite Barrón, Miquel Ràmia, Sònia Casillas
Institut de Biotecnologia i de Biomedicina - IBB / Department of Genetics and Microbiology, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain, Cerdanyola, Catalonia, Spain

The Drosophila Genetic Reference Panel (DGRP) has sequenced 168 inbred lines of *Drosophila melanogaster* from a single natural population. This dataset has allowed the unprecedented opportunity to perform the most comprehensive nucleotide variation study done so far. By incorporating information on the site frequency distribution to the framework of the McDonald Kreitman test, we have been able to estimate for any given region of the genome, ranging from a single gene to the whole genome, five different regimes of selection, namely, the fraction of new mutations that are strongly deleterious, weakly deleterious, neutral, and recently neutral, as well as unbiased estimates of the proportion of adaptive divergence. The description was done for the five site classes defined by a coding gene: synonymous, non-synonymous, UTRs, intron and intergenic sites. This comprehensive map of natural selection along the genome of *D. melanogaster* is a treasure trover to unveil the way natural selection operates both at the genome and the gene levels. The footprint of natural selection is pervasive all along the genome, although the importance of different selection regimes depends on the functional site classes and the genomics regions under consideration. Natural selection is more intense and effective on chromosome X than in the rest of chromosomes. Within autosomes, selection is much less effective in around one third of the length of the chromosome arm spanning the centromere. Another important observation is that recombination rate (variable along the genome) plays a key role in the ability of natural selection of improving the adaptation of different genomics regions. The probability that a genomic region responds effectively to natural selection depends on its recombination context. In a region with little or no recombination, selection cannot prevent the functional degradation of the region.

P-2080

Population Genetics of the Sex-Locus in a closed Population of the Honey Bee (Apis mellifera) Marcel Medrano, Dr Michael Lattorff, Pr Dr Robin Moritz

Marcel Medrano, Michael Lattorff, Robin Moritz
Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

In the honeybee (Apis mellifera) haplodiploidy is the mode of sex determination. Thereby heterozygosity at the sex-locus – the complementary sex determiner (csd) – induces the production of females whereas hemi- or homozygosity initiates the production of male individuals. Since possible diploid males do not survive to the stage of maturity, negative selection towards frequent alleles but positive selection towards rare alleles (balancing selection) for this particular locus is predicted. To this day, the molecular character providing allelic identity still remains unknown. In this study on an isolated honeybee population from the North Sea island Schiermonnikoog (NL) a set of six tightly linked microsatellites is used to constitute haplotypes, which qualitatively represent the enclosed csd-alleles. The low non-sampling error suggests that even rare csd-alleles are detected (AF0,95 = 0,019). The reliability of this method is also expressed by a low non-detection error (d = 0.00024) ensuring that one haplotype unambiguously represents one csd-allel. Employing this robust haplotype-tool on a great sample size from all but one colony of the population, 15 functionally different alleles have been identified which thus fits in the range of earlier estimates in the literature. Allele frequencies and distribution mainly confirm the prediction of balancing selection at the csd-locus. At the same time they clearly show characteristics of small populations dwelling on marine islands.

P-2081

Aging and the loss of immune repertoire genetic diversity

Philip Johnson[1], Andrew Yates[2], Jorg Goronzy[3], Rustom Antia[1]
[1]Emory University, Atlanta, GA, USA, [2]Albert Einstein College of Medicine, New York, NY, USA, [3]Stanford School of Medicine, Palo Alto, CA, USA

A diverse array of CD4 and CD8 T cells is required for defense against pathogens. The naive cell diversity reaches its peak by early human adulthood, and diversity remains approximately constant until older age. However, around age 70, this diversity plummets abruptly. A similar qualitative pattern holds for the memory cell population as well. We use mathematical models to explore different hypotheses for how such a loss of diversity might occur. The prevailing hypotheses suggest that the loss of diversity is due to a decline in emmigration of cells from the thymus or a contraction in total number of cells. Our model rejects these mechanisms since they yield only a gradual decline in the repertoire instead of the observed sudden decrease. We propose that this abrupt decline in the repertoire may be caused by the accumulation of mutations that provide a short-term fitness advantage to a small number of T cell clones (e.g., an increased division rate or decreased death rate) despite incurring the long-term cost of decreased ability to fight a variety of infections.

P-2082

Exploring genetic interaction as a predictor for non-random association among loci

Ching-Hua Shih[1], Shuwei Li[1], Hyun Jung Park[2], Luay Nakhleh[1,2], Michael Kohn[1]
[1]Department of Ecology and Evolutionary Biology, Rice University, Houston, TX, USA, [2]Department of Computer Science, Rice University, Houston, TX, USA

Non-random association of alleles among unlinked loci can be generated by either selection, intrapopulation heterogeneity due to recent admixture of genetically differentiated subpopulations, or demographics of populations. False discovered non-random association is expected when enough loci are analyzed. Estimation of association among loci is ordinarily done by calculation of non-random association, here called linkage disequilibrium (LD), from multilocus genotype (or haplotype) datasets. A large fraction of human gene pairs in LD currently is attributed to statistical and demographical factors. However, the contribution of biological factors to LD, such as genetic interactions, remains to be detected. Here we tested whether genetic interactions can partially explain LD in the human genomes in different populations.

We utilized 'degree distribution' and 'geodesic distance' characteristics from a human KEGG network. It has been shown that biological attributes of genes, such as essentiality and susceptibility to disease, may affect these characteristics. We calculated pairwise LDs for the SNPs in protein-coding regions from the genotype data from HapMap into eleven population LD networks to obtain the corresponding properties. To detect correlations between functional properties and non-random association of alleles among genes, we compared the corresponding properties of genes in KEGG network and in LD networks.

Both characteristics of KEGG network predicted their counterparts in LD networks. The power of KEGG network to predict LD among interacting genes was weak; one order of magnitude lower in terms of correlation coefficients, when compared to the power of physical linkage to predict LD among genes. Interchromosal LDs also indicate non-random associations between genes. While results were generally robust to demography, we detected a significant number of population-specific genes and interactions in LD networks.

Despite the weak correlation between network characteristics, a small portion of LD in the human genome appears to be due to genetic interactions. To better quantify the effect, analyses need to be expanded to different types of interaction networks. An unspecified fraction of population-specific gains and losses of LD among interacting genes hinted at genetic interactions of relevance to complex trait differences and medical conditions that distinguish human populations. The excess of interacting gene pairs in the genetic interaction network co-locating on the same haplotype indicated that LD among them should not automatically be attributed to physical linkage alone. While inconclusive, we hope that our results will spark explorations of additional network properties and of LD measures most powerful to detect LD due to genetic interactions effectively.

P-2083

Human evolutionary processes and genetic variation as revealed by mtDNA simulations

Aida Miró-Herrans, Connie Mulligan
University of Florida, Gainesville, FL, USA

Human evolutionary processes can be understood by describing the demographic parameters in these evolutionary processes, such as migration rates and bottleneck sizes. Demographic parameters are inferred from patterns of genetic variation, generally from statistics that summarize the genetic variation. In this study, we simulated mitochondrial DNA for 42 scenarios describing colonization of modern humans out of Africa to investigate the effects of parameter combinations on genetic variation. The 42 parameter combinations include three values for colonization size (1%, 10% and 30%), seven values for rate of gene flow (10^{-6}-10^{-1} and 0.5) and two values for time of colonization (50,000 and 100,000 years ago). We estimated genetic summary statistics on the simulated datasets to reflect the genetic variation of each parameter combination. We analyzed the summary statistics to calculate the percent of explained variation by each parameter and to identify which parameter combinations generated distinguishable differences in genetic variation. The results demonstrate that the summary statistics differentially explain variation depending on the parameter. However, colonization size, gene flow, and their interaction were significant across all summary statistics and yielded the highest percents of explained variation. The comparisons of the 42 parameter combinations show that scenarios that have a high colonization size (30%) generate similar patterns of genetic variation, while scenarios that have moderate to low colonization size (1% and 10%) and high levels of gene flow (greater than 10^{-3}) generate patterns of genetic variation that are distinguishable by gene flow category, but not by time category. These results suggest that some evolutionary scenarios, and particular parameters and questions of interest, may not be able to be distinguished based on mitochondrial DNA data, e.g. our results suggest there are few conditions under which it is possible to distinguish between human colonization out of Africa 50,000 vs 100,000 years ago, despite this being a pressing question in human evolution.

P-2084

Sex ratio between Y/A chromosomes, Y-haplotype diversity and frequency spectrum patterns confirms a marked ancient isolation in pigs (*Sus scrofa*)

Oscar Ramirez[1], Ana Ojeda[2], Marcel Amills[2,3], Greger Larson[4], Miguel Pérez-Enciso[2,3], Sebastian Ramos-Onsins[3]
[1]Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Spain, [2]Departament de Ciència Animal i dels Aliments Facultat de Veterinaria. Universitat Autònoma de Barcelona (UAB), Cerdanyola, Spain, [3]Department of Animal Genetics Centre for Research in Agricultural Genomics (CRAG) Consortium CSIC-IRTA-UAB-UB Edifici CRAG, Campus UAB, Cerdanyola, Spain, [4]Department of Archaeology, University of Durham, Durham, UK, [5]Institut Català de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Pigs (*Sus scrofa*) originated in Southeast of Asia and expanded to the rest of Eurasia and North Africa ca 0.1-0.9 MYA. In this work we studied the levels and patterns of nucleotide variability at seven Y-linked locus and also at 34 Y-linked SNPs in a world-wide sample, comprising 10 different groups of local, commercial pigs and wild boars mainly from Asia, Near East, Europe and Africa.

The nucleotide variability at Y-linked loci (0.0016/bp in Asian pigs) is of the same order than that reported at for autosomal loci (~ 0.0021), suggesting that the high Y-linked variability is consequence of a marked isolation process between the two-haplotype groups. We also observed two predominant haplotypes, which are concordant with previous mtDNA observations. Furthermore, the genotyped data showed that the closest outgroup *Sus celebensis* is an intermediate haplotype between the two predominant haplotypes, suggesting an ancient isolation in the Asian population. We discuss these results in relation with mitochondrial, autosomal and X-linked loci data in order to explain the history of this species.

P-2085

Depth-related changes in coral zooxanthellae

Nikolaos Schizas[1], Matthew Lucas[1], Matthew Smith[1,2], Ernesto Weil[1]
[1]University of Puerto Rico, Mayaguez, USA, [2]University of Wisconsin - Milwaukee, Milwaukee, USA

The extraordinary physiological and depth range of some corals distributed from shallow waters to mesophotic depths is likely attributed to adaptations involving both the endosymbiotic dinoflagellates and the coral host. Mesophotic coral ecosystems (MCEs, reef habitats found in > 50 m) of southwestern Puerto Rico and their adjacent shallow water counterparts provide a unique system to examine the patterns of genetic connectivity for scleractinian corals and their algal symbionts (*Symbiodinium* spp.). *Agaricia lamarcki* (Cnidaria: Scleractinia) harbors algal endosymbionts and inhabits both shallow and mesophotic habitats. In this study, *Symbiodinium* lineages from *A. lamarcki* were estimated between shallow (< 25 m) and mesophotic populations (50-70 m) in Mona Island, southwestern Puerto Rico, and St. Thomas, USVI. DNA sequences for *Symbiodinium* communities were obtained with the commonly used internal transcribed spacer of rDNA (ITS2). To detect the presence of multiple *Symbiodinium* lineages within a single coral host we used molecular cloning. Thus far, 280 bp of the ITS2 region shows that *A. lamarcki* populations > 50 m harbor “deep specialists” symbionts. Continuing work will include tests for population genetic structure among different depth habitats within a region and between regions. These data are important since coral-algal associations are hypothesized to provide an evolutionary plasticity for corals responding to further environmental degradation and climate change.

P-2086

The joint allele frequency spectrum of multiple populations: A coalescent theory approach

Hua Chen
Department of Epidemiology and Biostatistics, Harvard School of Public Health, Boston,MA, USA

The allele frequency spectrum is a series of statistics that describe genetic polymorphism, and is commonly used for inferring population genetic parameters and detecting natural selection. Population genetic theory on the allele frequency spectrum for a single population has been well studied using both coalescent theory and diffusion equations. Recently, the theory was extended to the joint allele frequency spectrum (JAFS) for three populations using diffusion equations and was shown to be very useful in inferring human demographic history. In this paper, I show that the JAFS can be analytically derived with coalescent theory for a basic model of two isolated populations and then extended to multiple populations and various complex scenarios, such as those involving population growth and bottleneck, migration, and positive selection. Simulation study is used to demonstrate the accuracy and applicability of the theoretical model. The coalescent theory-based approach for the JAFS can characterize the demographic history with comprehensive statistical models as the diffusion approach does, and in addition gains several novel advantages: the computational complexity of calculating the JAFS with coalescent theory is reduced, and thus it is feasible to analytically obtain the JAFS for multiple populations; the hitchhiking effect can be efficiently modeled in coalescent theory, enabling the development of methodologies for detecting selection via multi-population polymorphism data. As an alternative to the diffusion approximation approach, the coalescent theory for the JAFS also provides a foundation for population genetic inference with the advent of large-scale genomic polymorphism data.

P-2087

Latitudinal clinal variation in a blue-absorbing visual pigment of butterflies

Furong Yuan, Adriana Briscoe
University of California, Irvine, Irvine, USA

Clinal variation provides strong evidence of natural selection and adaptation to the environment. Such variation has been reported in natural populations of numerous species for morphological and genetic traits. Perhaps the most compelling evidence for clinal selection comes from the many studies on latitudinal clinal variation in *Adh* in *Drosophila melanogaster*. Latitudinal clines have also been studied in other insects such as butterflies for genetic traits including allozyme loci and a heat-shock gene HSP70. Here we report a novel latitudinal cline in butterflies in the blue-absorbing visual pigment (B opsin) gene of *Limenitis arthemis* butterflies. In the study, we sampled genomic DNAs from nine wild populations of *L. arthemis* along the east coast of North America. We sequenced the exons of the B opsin gene of these individuals and analyzed the data on a site-by-site base. 22 polymorphic sites were scored along the1143-bp coding region in 228 individuals. Next, we plotted the frequencies of each polymorphic site against latitude and carried out logistic regression analysis. Our regression analysis revealed a highly significant polymorphic site, A901G, encoding threonine / alanine (Thr/Ala) at amino acid position 301. The frequency of Ala-bearing alleles in each population declined with decreasing latitude, such that the southern-most populations were mostly fixed for Thr. Such geographical heterogeneity implies the action of positive selection on the blue opsin gene, specifically on the A901G site. By contrast, the remaining 21 polymorphic sites showed no significant latitudinal variation in frequencies, implying that they have evolved neutrally. To investigate if the Thr/Ala site has any impact on spectral tuning, we functionally expressed and characterized a natural variant of the *L. arthemis astyanax* B opsin and its single mutant Thr301Ala in HEK293 cells. The wavelength of peak absorbance, lambda-max, of the resulting mutant photopigment was significantly blue-shifted by 3 nm. Taken together, our study provides both genetic and functional evidence for selective pressure acting on the polymorphic site A901G in the opsin gene of the butterfly *Limenitis arthemis*.

P-2088

Coalescent analysis of archaic hominin data

Swapan Mallick, David Reich
Harvard Medical School, Boston, MA, USA

The analysis of genetic data from closely related organisms can provide insights into the nature, timing and dynamics of their separation. To analyse the mechanics of living and extant hominins, we constructed alignments of denisova data, a haploid gujarati, african sections of the human reference, a san individual and a ceu individual using chimpanzee as an outgroup. We restrict our analysis to regions of very low recombination, and assume that regions which have low recombination in humans are also low in denisova. This allows us to perform a haploid coalescent analysis with which we can estimate effective population sizes and the population split times of archaic hominins from modern day humans.

P-2089

Using deterministic functions to derive approximate coalescent distributions

Ethan Jewett, Noah Rosenberg
Stanford University, Stanford, CA, USA

Maruvka et al. (2011) observed that, under the coalescent model, the number $n(t)$ of coalescent lineages as a function of time is nearly deterministic and is well approximated by its expected value $E[n(t)]$. In turn, $E[n(t)]$ is well approximated by simple deterministic functions that are fast and numerically stable to compute. Here, we show how the approximation $n(t) \approx E[n(t)]$ can be used to derive accurate, computationally fast, and numerically stable approximations to a variety of coalescent probability distributions and expectations. These approximations are generally very good, even when the number $n(0)$ of lineages sampled at time $t = 0$ is small. Moreover, the computational complexity of these approximations remains constant as $n(0)$ increases. Such approximations provide alternatives to exact formulas that are computationally intractable or numerically unstable when the number of sampled lineages is moderate or large.

YE Maruvka, NM Shnerb, Y Bar-Yam, and J Wakeley (2011) Recovering population parameters from a single gene genealogy: An unbiased estimator of the growth rate. *Mol. Biol. Evol.* 28:1617-1631.

P-2090

Parallel evolution in Drosophila species along a latitudinal cline

Heather Machado[1], Alan Bergland[1], Paul Schmidt[2], Dmitri Petrov[1]
[1]Stanford University, Stanford, CA, USA,[2]University of Pennsylvania, Philadelphia, PA, USA

Although there is much stochasticity in the processes underlying evolution, there is also predictability due to the physiological and genetic similarities of species with shared evolutionary histories, and due to shared environments exerting similar selective pressures. Molecular evolution experiments have sought to quantify this predictibility, and comparative genomic studies have identified instances of genetic convergence. However, this has been less studied in natural populations and, importantly, is understudied in the context of local adaptation of populations. Here, we ask "how predictable is the genetic basis of local adaptation?" by studying genome wide patterns of parallel local adaptation along a latitudinal cline in two sister species, Drosophila melanogaster and Drosophila simulans. These species represent an ideal system for addressing parallel adaptation, as they possess consistent morphological differences associated with the latitudinal cline (e.g., body size, development time, stress tolerance), are still physiologically similar, and can be collected in large numbers, facilitating accurate estimation of allele frequencies. We perform deep population genomic sequencing of four populations of each species along a 17˚ latitudinal cline and identify 1000's of SNPs that vary clinally. We use these data to quantify the strength and extent of selection producing local adaptation and the extent of parallelism of adaptation at the functional and genomic levels.

P-2091

A Simple Method for Finding Explicit Solutions to the Dynamics of Diffusion Processes with General Diploid Selection

Yun Song[1,2], Matthias Steinruecken[1]
[1]UC Berkeley, Department of Statistics, Berkeley, CA, USA, [2]UC Berkeley, Computer Science Division, Berkeley, CA, USA

The Wright-Fisher diffusion describes the evolution of population-wide allele frequencies over time and has been successfully applied in various population genetic analyses in the past. Examples include finding the stationary distribution of allele frequencies, and approximating fixation times and probabilities. The diffusion also allows one to describe the development of frequency spectra in time, which has recently been successfully applied to the inference of human demography. Although the diffusion has important practical applications in population genetics, finding explicit formulas for the dynamics under a general diploid selection model has remained a difficult open problem.
In this work, we develop a new computational method to tackle this classic problem. Using spectral methods, we find explicit expressions for the dynamics of the Wright-Fisher diffusion with recurrent mutation and arbitrary diploid selection. Simplicity is one of the appealing features of our approach. Although our derivation involves somewhat advanced mathematical concepts, the resulting algorithm is quite simple and efficient. Furthermore, unlike previous approaches, which were only accurate when the population-scaled selection coefficient is small, our method is valid for a broader range, including biologically relevant values.
We show how our solution can be applied to obtain the rate of convergence to the stationary distribution under mutation-selection balance. We also discuss the application of our solution in a hidden Markov model framework to infer the strength of selection from population samples taken at different points in time, for example from ancient human DNA or viral populations.

P-2092

Formation of reproductive barriers in a hybrid zone of American and Caribbean *Drosophila melanogaster*

Joyce Kao, Sergey Nuzhdin
University of Southern California, Los Angeles, USA

Studying reproductive barriers is a valuable asset to understanding the dynamics of hybrid zones where two allopatric populations interbreed again after being separated for a long period of time adapting to their separate environments. The southeast United States (US) and Caribbean islands is a recent hybrid zone between two distinct populations of *Drosophila melanogaster* that have been diverged for 10,000 years: the European flies which colonized the US with the European settlers and the African flies which were introduced into the Caribbean islands via the trans-Atlantic slave trade. Previous studies have established the existence of clines in several pre-mating or pre-zygotic reproductive traits from this area of the world. We are interested in determining whether there is a similar geographical pattern of post-mating behavior (i.e. post-zygotic reproductive traits) among these flies. After mating, females exhibit a variety of behavioral changes, but we are most interested in the reduced receptivity to re-mating and increase in egg laying. We will be measuring the effects of reduced receptivity to re-mating by introducing males to mated females three days and seven days after initial mating and recording when females choose to re-mate. Fecundity and fertility measurements involve counting the number of egg laying for 10 days and counting the number of flies that eclose from those eggs, respectively. Furthermore, we are interested in examining these established and potential reproductive barriers on a genomic level. We are re-sequencing genomes of the 23 isofemale lines used in the post-mating behavioral assays to look for signatures of selection in genes that possibly play a role in the formation of reproductive barriers in this hybrid zone.

P-2093

Genetic evidence of two major prehistoric migrations of modern humans into the Tibetan plateau

Xuebin Qi[1], Chaoying Cui[2], Yi Peng[1,5], Xiaoming Zhang[1,5], Zhaohui Yang[1,5], Hui Zhang[1], Kun Xiang[1,5], xiangyu Cao[1,5], Yi Wang[1,5], Ouzhuluobu .[2], Basang .[3], Ciwangsangbu .[3], Bianba .[2], Gonggalanzi .[2], Tianyi Wu[4], Hua Chen[6], Hong Shi[1], Bing Su[1]

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China,[2]School of Medicine, Tibetan University, Lhasa, China,[3]People's Hospital of Dangxiong County, Dangxiong, China,[4]National Key Laboratory of High Altitude Medicine, High Altitude Medical Research Institute, Xining, China,[5]Graduate School of Chinese Academy Sciences, Beijing, China,[6]Department of Epidemiology, Harvard School of Public Health, Boston, USA

The prehistoric peopling of modern humans at the Tibetan plateau remains one of the unsolved mysteries in human inhabitation history. With the use of genetic markers representing both the paternal (Y chromosome) and the maternal (mitochondrial DNA) lineages, we analyzed the genetic diversity of 44 Tibetan populations (a total of 6,109 individuals) across the northern Himalayas. In these Tibetans, we identified ancient Tibetan-specific genetic signatures which were dated to about 30 thousand years ago (kya), suggesting the initial peopling of modern humans in Tibet occurred during the Paleolithic time. There are also shared Y chromosome and mitochondrial DNA haplotypes between Tibetans and northern Han Chinese, and these haplotypes are relatively young (8-10 kya), indicating a second wave of migration into the Tibetan plateau resulting from the Neolithic expansion of agriculture starting from northwestern China. Hence, the genetic data supports two major prehistoric migrations of modern humans into the Himalayas.

P-2094

Adaptive evolution of loci covarying with the human African Pygmy Phenotype

Isabel Mendizabal[1], Urko M. Marigorta[1], Oscar Lao[0,2], David Comas[1]
[1]Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain, [2]Department of Forensic Molecular Biology, Erasmus University Medical Center, Rotterdam, The Netherlands

African Pygmies are hunter-gatherer populations from the equatorial rainforest that present the lowest height averages among humans. The biological basis and the putative adaptive role of the short stature of Pygmy populations has been one of the most intriguing topics for human biologists in the last century, which still remains elusive. Worldwide convergent evolution of the Pygmy size suggests the presence of strong selective pressures on the phenotype. We developed a novel approach to survey the genetic architecture of phenotypes and applied it to study the genomic covariation between allele frequencies and height measurements among Pygmy and non-Pygmy populations. Among the regions that were most associated to the phenotype, we identified a significant excess of genes with pivotal roles in bone homeostasis such as PPPT3B and the height associated SUPT3H-RUNX2. We hypothesize that skeletal remodeling could be a key biological process underlying the Pygmy phenotype. In addition, we showed that these regions have most likely evolved under positive selection. These results constitute the first genetic hint of adaptive evolution in the African Pygmy phenotype, which is consistent with the independent emergence of the Pygmy height in other continents with similar environments.

P-2095

Rapid fixation of duplications and deletions in experimental *Caenorhabditis elegans* populations during adaptive recovery.

Vaishali Katju, James Farslow, Ulfar Bergthorsson
University of New Mexico, Albuquerque, NM, USA

@font-face { font-family: "M S 明朝"; }p.MsoNormal, li.MsoNormal, div.MsoNormal { margin: 0in 0in 0.0001pt; font-size: 12pt; font-family: "Times New Roman"; }div.Section1 { page: Section1; }

Selection on gene dosage has been hypothesized to play an important role in the maintenance of duplicate genes in populations and in the evolution of new genes. Here we investigate if adaptive gene copy-number changes appear in populations of *Caenorhabditis elegans* undergoing compensatory evolution following fitness decline during mutation accumulation. Independent replicate lines of the obligately outcrossing *fog-2* mutant strain were subjected to 50 generations of mutation accumulation (MA) via repeated bottlenecking at $N_e = 2$ (single-sib pair) and RNAi-induced knockdown of the key mismatch repair gene *msh-2*. Subsequent to the MA regime, each line was independently expanded and maintained at large population size ($N_e > 5{,}000$) for more than 200 generations to enable fitness recovery via compensatory mutations. Olignonucleotide array Comparative Genomic Hybridization (oaCGH) analyses on these recovery population found multiple instances of large duplications in high frequency. The oaCGH results were confirmed by qPCR, and when possible, PCR and sequencing across duplication and deletion breakpoints. Single worm PCR and qPCR of DNA from population samples at different time points during the fitness recovery experiments showed that the frequency of copy-number variants increased gradually with some reaching fixation at the population-level. Two regions were duplicated in more than one population, but with different breakpoints. In three independent populations, the breakpoints of duplications were located within the same repeat on chromosome V, but the breakpoint is at different locations within the repeat in all instances. The occurrence of several parallel copy-number changes in high frequency in independent populations suggests a strong role of natural selection in their spread towards fixation.

P-2096

Coalescent-based Species Tree Inference from Gene Tree Topologies Under Incomplete Lineage Sorting by Maximum Likelihood

Yufeng Wu
University of Connecticut, Storrs, CT, USA

It is commonly observed that phylogenetic trees (called gene trees) inferred from DNA sequences of individual genes are not always identical. One of the main causes is incomplete lineage sorting. Incomplete lineage sorting is caused by the inherent stochasticity of population genealogical processes and is believed to be widespread in the evolution of species and populations. The incongruence between gene trees and the species tree caused by incomplete lineage sorting greatly complicates species tree inference. During the past several years, several methods have been developed to infer the species tree from discordant gene trees.

In this presentation, I present a new coalescent-based algorithm for species tree inference with maximum likelihood from a set of gene tree topologies.
I first describe an improved method for computing the probability of a gene tree topology given a species tree with incomplete lineage sorting, which is much faster than an existing algorithm. This new peeling-style method for computing the probability of a gene tree topology given a species tree. The gene tree probability algorithm is not an approximation: it computes exactly the same gene tree probability as Degnan and Salter (2005). Based on this method, I develop a practical algorithm that takes a set of gene tree topologies and infers species trees with maximum likelihood. The likelihood being computed in STELLS is the likelihood of the gene tree topologies assuming that the gene tree topologies are correctly inferred. This algorithm searches for the best species tree by starting from initial species trees and performing heuristic search to obtain better trees with higher likelihood. This algorithm infers a species tree with branch lengths. Branch lengths on the estimated species tree are in coalescent units. This algorithm, called {STELLS}, has been implemented in a program that is downloadable from the author's web page. The simulation results show that the STELLS algorithm is more accurate than an existing maximum likelihood method for many datasets, especially when there is noise in gene trees. I also show that the STELLS algorithm is efficient and can be applied to real biological datasets.

P-2097

Recurrent positive selection at the same amino acids in primates and human populations

Renee George[1], Graham McVicker[2], Josh Akey[1], Willie Swanson[1]
[1]University of Washington, Seattle, WA, USA,[2]University of Chicago, Chicago, IL, USA

Positive selection that acts over long evolutionary time periods can be detected by an elevated number of protein coding sequence differences between species. An important question in evolutionary biology is whether positive selection repeatedly targets the same amino acids at different evolutionary time periods along the same lineage or along multiple independent lineages. To investigate this possibility, we tested whether genes that are rapidly evolving in primates are also under positive selection in modern human populations. By comparing the protein coding sequences of 6 non-human primate species, we identified 473 genes containing 1249 codons with strong evidence for prior episodes of positive selection. We then looked for signs of recent adaptive evolution in the same codons using a large exome data set of 2440 individuals. Of the putatively positively selected sites (PSSs), 0.83% are fixed for non-synonymous mutations that occurred on the human lineage. This is far greater than the number of fixed non-synonymous mutations at other codons (0.23%) and even exceeds the number of fixed synonymous mutations at four-fold degenerate sites (0.57%). The PSSs have an elevated non-synonymous diversity (0.168%) compared to non-positively selected sites (0.032%) and synonymous diversity at neutral sites (0. 078%). Variants at PSSs are skewed towards higher frequencies than those at non-positively selected sites, and the differences between PSSs and non-positively selected codons cannot be explained by regional variation in the mutation rate, hypermutable CpG dinucleotides or population substructure. These results are evidence that many amino acid sites that have previously undergone positive selection in primates have experienced similar selection pressure in the human lineage and continue to evolve adaptively.

P-2098

Phylogeography in habitat generalists and specialists in the granite inselbergs of Namibia.

Sara Tromp, Anne Goldizen, Jenny Seddon
The University of Queensland, QLD, Australia

In the Northern Hemisphere, severe Pleistocene climate cycles resulted in strong concordance of refugia and hence in similar genetic expansion patterns. In contrast, in the southern hemisphere and tropical regions, phylogeographic patterns suggest that large-scale migration was unusual and that the majority of biota persisted in multiple localized refugia. Here we examined the little-studied arid region of north-western Namibia's granite inselbergs. The inselberg are large islands of granite rising from a desert plain and strong phylogeographic patterns may be expected. Strikingly different mtDNA phylogeographic patterns were observed for two rodent species: a habitat specialist, *Petromyscus shortridgei*, showed strong phylogenetic and geographic structure, whilst a habitat generalist, *M. namaquensis*, lacked any clear geographic structure in the phylogeny. Three habitat specialists, but of different vagility, were sampled – the black mongoose, *Galerella nigrata*, the mouse *Petromyscus shortridgei*, and the lizard, A*gama planiceps*. The pattern of divergence was strongest in the least vagile species yet the inferred barriers to dispersal varied among the species, indicating there were no specific habitat refugia in the landscape. Further, the importance of species' ecological plasticity, ability to cross geographic barriers and responses to stochastic effects were highlighted.

P-2099

Genetics of malaria susceptibility across three African populations: Kenya, Malawi and Gambia.

Chris Spencer
Wellcome Trust Centre for Human Genetics, University of Oxford, UK

Through the MalariaGEN initiative (www.malariagen.net) large cohorts of individuals diagnosed with severe malaria have been collected, along with matching population controls. We analysed data from three populations, totalling 10,000 samples, assayed on Illumina genotyping arrays. We describe at high resolution the genetic diversity within these samples; it shows strong population structure both within and between collections, which correlates with ethnicity; and highlights regions with extreme allele frequency differences, consistent with natural selection. Using imputation we investigate the signals of association at previously reported susceptibility loci and search for new signals of association. Both existing approaches to combining data from different populations, and new tools for assessing evidence for susceptibility variants are employed. The results are very promising for the application of genome-wide association studies to infectious disease in diverse populations. These obsevations provide insights into the role of pathogens in shaping human genetic diversity.

P-2100

Understanding genetic adaptations to varying dietary selenium intakes in human populations

Louise C White, Genís Parra, Sergi Castellano
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Diet is an important environmental factor that all humans are exposed to throughout life. Because diet varies with culture and the environment, and has a crucial role in subsistence, it likely represents a powerful selective force. A particularly interesting dietary component is the essential trace nutrient selenium. Selenium metabolism in humans is tightly regulated by levels of dietary selenium, which is in turn largely dependent on soil and water selenium concentrations. Daily selenium intake ranges from 3μg in geologically selenium-poor regions of the world to 6700μg in selenium-rich regions. As humans spread out of Africa, they experienced a vast range of climates, diets and ecosystems, including soils with very different levels of selenium. We therefore hypothesise that geographical or temporal differences in selenium availability to human populations have led to regulatory adaptations in selenium dependent genes.

In order to test this hypothesis, we have captured and sequenced to high coverage (20x) exons and regulatory regions from 52 genes that either incorporate selenium or are involved in its metabolism. This list includes the 25 proteins known to incorporate selenium in the form of the amino acid selenocysteine, the 21st amino acid in the genetic code, and a number of selenium-binding and selenium-transport proteins. These genes were sequenced in the complete HGDP-CEPH panel, which provides a worldwide sample of 52 populations with a wide diversity of diets. With this dataset, we evaluate the extent to which recent natural selection has shaped the evolution of selenium homeostasis amongst human populations. Our analytical approach uses tests designed to detect targets of differential selection among populations (e.g. XP-EHH and XP-CLR) in addition to tests for shared selective sweeps at our loci of interest.

P-2101

Establishment of new mutations in changing environments

Stephan Peischl[1,2], Mark Kirkpatrick[2]
[1]*University of Bern, Bern, Switzerland,* [2]*University of Texas at Austin, Austin, TX, USA*

A factor that may limit the ability of many populations to adapt to changing conditions is the rate at which beneficial mutations are able to become established. We study the probability that mutations become established in changing environments by extending the classic theoretical framework for branching processes. We derive simple analytical results for a variety of interesting cases. When environments change in time, under quite general conditions, the establishment probability is approximately twice the “effective selection coefficient”, whose value is an average that gives most weight to a mutant’s fitness in the generations immediately after it appears. This generalizes Haldane's classical result for the probability of fixation to arbitrary patterns of temporal change in selection intensity. When fitness varies along a gradient in a continuous habitat, increased dispersal generally decreases the chance a mutation establishes because mutations move out of areas where they are most adapted. When there is a patch of favorable habitat that moves in time, there is a maximum speed of movement above which mutations cannot become established regardless of when and where they first appear. This critical rate of change, which is proportional to the mutation’s maximum selective advantage, represents an absolute constraint on the potential of locally-adapted mutations to contribute to evolutionary rescue.

P-2102

Molecular Evolution of DNA Sequences on Neo-X and Neo-Y Chromosomes in *Drosophila albomicans*.

Kazuhiro Satomura, Koichiro Tamura
Tokyo Metropolitan University, Tokyo, Japan

Meiotic recombination is a universal phenomenon in the replication of eukaryote genomic DNA and believed to have influence on molecular evolution of genes and genomes by causing mutations and relaxing the effect of linkage on natural selection. It is known that the repair error on the double-strand breaks generated in the process of meiotic recombination can lead to nucleotide substitutions. The fate of a gene is expected to be affected by its linked genes under strong selection without meiotic recombination as referred to as 'hitchhiking effect.' However, it is usually very difficult to distinguish the effects on these two different driving forces of molecular evolution. Therefore, the relative importance of meiotic recombination on mutation and natural selection remain to be examined for actual evolution of genes and genomes. In order to distinguish them clearly, comparing the molecular evolution between physically and functionally the same genes but located in different conditions of meiotic recombination is required.

A fruit fly species, *Drosophila albomicans*, has neo-X and neo-Y chromosomes originated by a fusion of X and an autosome and another fusion of Y and the autosome, respectively, resulting in a situation that so many genes derived from the autosome are shared by the neo-X and neo-Y chromosomes. As, generally in *Drosophila*, meiotic recombination does not occur in the male germline, the genes on the neo-Y chromosome must have been in a recombination-free environment, while their homologues on the neo-X chromosome have continuously experienced meiotic recombination in the female germline. Therefore, comparing DNA sequence variation of the genes on the neo-Y chromosome with that of the homologues on the neo-X chromosome, we can directly examine the effects of meiotic recombination. In this study, we performed comparative analyses of genetic variations for over forty genes, half of which are located on the neo-X and neo-Y chromosomes in *D. albomicans* in order to clarify the effects of meiotic recombination on the molecular evolution of genes and genomes.

The results showed that the difference in the mutation rate made little difference in the genetic variation of the genes on the neo-Y chromosome, whereas the genetic variation was largely affected by natural selection on the linked genes on the neo-Y chromosome. These results suggest that meiotic recombination takes effect on the molecular evolution by relaxing the effect of linkage to enhance the efficacy of natural selection on individual genes.

P-2103

Dynamic transmission, host quality and population structure in a multi-host parasite of bumble bees

Mario X. Ruiz-González[1,2], John Bryden[3], Yannick Moret[4], Christine Reber-Funk[5], Paul Schmid-Hempel[5], Mark J. F. Brown[1,3]
[1]Department of Zoology, School of Natural Sciences, Trinity College Dublin, Dublin, Ireland,[2]Instituto de Biología Molecular y Celular de Plantas, C.S.I.C. – U.P.V., Valencia, Spain,[3]School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey, UK,[4]Université de Bourgogne, UMR CNRS, Dijon, France,[5]ETH Zurich, Institute of Integrative Biology (IBZ), Zürich, Switzerland

The evolutionary ecology of multi-host parasites is predicted to depend upon patterns of host quality and the dynamics of transmission networks. Depending upon the differences in host quality and transmission asymmetries, as well as the balance between intra- and inter-specific transmission, the evolution of specialist or generalist strategies is predicted. Using a trypanosome parasite of bumble bees we ask how host quality and transmission networks relate to parasite population structure across host species, and thus the potential for the evolution of specialist strains adapted to different host species. Host species differed in quality, with parasite growth varying across host species. Highly asymmetric transmission networks, together with differences in host quality, likely explain local population structure of the parasite across host species. However, parasite genetic population structure across years was highly dynamic, with parasite populations varying significantly from one year to the next within individual species at a given site. This suggests that, whilst host quality and transmission may provide the opportunity for short-term host specialisation by the parasite, repeated bottlenecking of the parasite, in combination with its own reproductive biology, overrides these smaller scale effects, resulting in the evolution of a generalist parasite.

P-2104

The demographic and adaptive history of domesticated and wild yeast

Angela Bean[1], Nicholas Renzette[1], Oliver Rando[1], Jeffrey Jensen[1]
[1]University of Massachusetts Medical School, Worcester, MA, USA,[2]EPFL, Lausanne, Switzerland

The identification of adaptive changes in a genome is crucial for understanding the process of selection and the evolutionary forces that shape genetic variability. With the availability of whole genome sequence data for over seventy isolates; *S. cerevisiae* has become an invaluable tool to study these selective pressures. The species as a whole occupies a variety of ecological, geographical, clinical, and industrial niches providing a wide range of opportunities for adaptive changes to occur allowing for the opportunity to study how different pressures shape genomic variability. Recent studies have shown significant differences in levels of variation between lab and wild strains – a prediction of classic hitchhiking models, and have identified putative adaptive regions in the former. However, it is also of note that in the process of moving wild strains into the lab, they likely underwent a diversity reducing domestication event - an important consideration given the well-known capacity of population bottlenecks to replicate genomic patterns of positive selection (i.e., false positives). Using whole genome polymorphism data from both wild and lab strains, this study systematically estimates the demographic history of each population. With these parameters in hand, we perform a corrected scan for recent adaptive events important in the domestication of yeast, and identify a number of genomic regions both putatively important for lab strains, as well as of shared significance between lab and wild. These results allow for the evaluation of a number of recent publications making claims regarding the pervasiveness of positive selection in the yeast genome.

P-2105

A Hidden Markov Model for inferring natural selection and other demographic parameters from allele frequency data

Anand Bhaskar[1], Yun Song[1,2], Matthias Steinruecken[2]
[1]Computer Science Division, University of California, Berkeley, Berkeley, California, USA, [2]Department of Statistics, University of California, Berkeley, Berkeley, California, USA

We consider the problem of inferring signatures of natural selection, divergence times and other demographic parameters given diallelic samples taken from a single locus of a population at different points in time. Such samples are available through various means, for example, from samples of ancient DNA, or from lab strains observed over several generations of reproduction. The trajectory of allele frequency over time can be modeled by a Wright-Fisher diffusion process incorporating recurrent mutation and general diploid selection. Recently, techniques from linear algebra were used to derive a novel analytic expression for the transition density function of this diffusion process, and we employ this representation in a Hidden Markov Model (HMM) framework to track the dynamics of the allele frequency over time in order to efficiently evaluate the likelihood of observed data. Previous work on using HMMs to infer the strength of selection from temporal allele frequency data has relied on numerical methods to evaluate the transition density function and has only dealt with the case of a panmictic population. On the other hand, our method allows analytic computation in the transitions of the HMM, and can also be applied to structured populations with point migration events. As a result, we can use our method to efficiently infer branch lengths and other demographic parameters for fitting demographic models of structured populations. We demonstrate our method on both simulated and real biological data.

P-2106

Genomics of Oil Exposure

Marjorie Oleksiak, Douglas Crawford
University of Miami, Miami, FL, USA

@font-face { font-family: "Cambria"; }p.MsoNormal, li.MsoNormal, div.MsoNormal { margin: 0in 0in 0.0001pt; font-size: 12pt; font-family: "Times New Roman"; }div.Section1 { page: Section1; }

The Deepwater Horizon oil spill released millions of gallons of oil and dispersants into the Gulf of Mexico, with potential short and long-term impacts on multiple species and populations. To explore the physiological and evolutionary impacts of the *Deepwater Horizon* oil release, we combined microarray studies of gene expression with high throughput genotyping technologies (genotyping by sequencing or GBS) using the teleost *Fundulus grandis*, a natural inhabitant of estuaries along the Gulf of Mexico. Data on *F. grandis* populations prior to impact in combination with unimpacted populations provide a powerful experimental design to ascertain the effects of the Deepwater Horizon oil release. Thus, using genomic technologies, we can describe the genetics of mans' impact on his environment and provide the baseline data to document the effect of oil pollution and effectiveness of remediation.

P-2107

Demographic processes or mitochondrial selective sweep in shaping patterns of genetic diversity in a Neotropical treefrog endemic from the Brazilian Atlantic forest

Tuliana Oliveira Brunes[1,2], João Paulo Soares de Cortes[2], João Alexandrino[3], Célio Fernando Baptista Haddad[2], Fernando Sequeira[1]
[1]CIBIO-UP/PT, Vairão, Porto, Portugal, [2]UNESP, Rio Claro, São Paulo, Brazil, [3]UNIFESP, Diadema, São Paulo, Brazil

Low mitochondrial DNA variation has been traditionally explained by demographic expansions and rarely by selective processes. Recent molecular analysis of an endemic treefrog from the Brazilian Atlantic Forest, Phyllomedusa distincta, revealed two highly differentiated mtDNA evolutionary groups with contrasting levels of genetic variability. While the northern group presents high levels of genetic variation, the southern group, that covers half of the P. distincta range, is represented by only a single haplotype. Here, we extended the mtDNA (ND2, 993bp) sequencing analysis to additional samples and included a nuclear marker (β-fibrinogen intron 7; β-fibint7) to examine in detail the phylogeographic pattern of P. distincta. Our results corroborate previous mtDNA analysis in revealing two highly differentiated groups and the striking absence of ND2 variability in the southern group. Nuclear marker analysis revealed the presence of exclusive haplotypes in the two mtDNA-defined evolutionary groups, but in contrast, β-fibint7 showed similar levels of genetic variability in both groups. We hypothesized that this unexpectedly absence of mtDNA variation in the southern group could be the result of a sudden bottleneck followed by a demographic expansion or a selective sweep. We used the Hudson-Kreitman-Aguadé (HKA) and neutrality tests to discriminate between the two hypotheses. No departure from neutrality was found either for the nuclear marker in both groups or for ND2 in the northern group. However, HKA test was significant for the southern group. Although further analyses are needed, the discordant patterns found in the extent and nature of mtDNA genetic diversity favored the hypothesis of a selective sweep on mtDNA. To investigate if some bioclimatic difference related with temperature and ecofisiological adaptation could be associated with a putative mitochondrial selective sweep, we used the ENMtools implemented Hellinger´s I distance and Schoener´s D metrics to verify the niche similarity between the two genetic groups. Results suggested some niche dissimilarity between the Northern and Southern groups, which may be further explored in trying to unravel the processes underlying mtDNA selection in the southern populations.

P-2108

Kernel Approximate Bayesian Computation

Shigeki Nakagome, Kenji Fukumizu, Shuhei Mano
The Institute oif Statistical Mathematics, Tokyo, Japan

Bayesian inferences are very popular to estimate evolutionary parameters. Population genetics models are generally complicated, and it is difficult to obtain explicit likelihood function forms. Approximate Bayesian Computation (ABC) is an alternative method for Bayesian inferences without likelihoods. ABC is a rejection method with tolerance of dissimilarity between summary statistics of observed data and those of simulated data. ABC gives an exact sampler from the posterior density in the limit of zero tolerance. Acceptance rates, however, decrease with increasing number of summary statistics. Further, choices of summary statistics and metric of dissimilarity are ambiguous. In practice, therefore, it is difficult to keep the consistency of estimators in ABC. In this work, we apply the kernel Bayes' rule proposed by Fukumizu et al. (2011) to the ABC framework and develop a new method, kernel ABC. Advantages of kernel ABC are that (i) no tolerance is needed, (ii) the consistency of estimators is upheld, and (iii) a large number of summary statistics is tractable. We demonstrate the efficiency of kernel ABC in the inference of population demography; costs of computing times are reduced to achieve the same accuracy of parameter estimation with conventional ABC methods.

P-2109

Patterns of genetic diversity and population structure in domestic rabbits: early steps of domestication and the subsequent process of breed formation

Joel Alves[1,2], Miguel Carneiro[1,2], Sandra Afonso[1], Susana Lopes[1], Hervé Garreau[3], Samuel Boucher[4], Daniel Allain[3], Guillaume Queney[5], Pedro Esteves[1,6], Gerard Bolet[3], Nuno Ferrand[1,2]
[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Vairão, Portugal,[2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal,[3]INRA, Institut National de la Recherche Agronomique, Castanet-Tolosan Cedex, France,[4]Labovet Conseil, Les Herbiers, France,[5]ANTAGENE, Wildlife Genetics Laboratory, Lyon, France,[6]CITS, Centro de Investigação em Tecnologias da Saúde, IPSN, CESPU, Gandra, Portugal

For thousands of years man shaped the genetic and phenotypic composition of several organisms by transforming wild species into domesticated forms. From this close association, domestic species emerged as important models in biomedical and fundamental research, in addition to their intrinsic economical value. The domestic rabbit is no exception but few studies have investigated the impact of domestication across its genome. Here, in order to study patterns of genetic structure in domestic rabbits and to quantify the amount of genetic diversity lost with domestication, we genotyped 45 microsatellites for 471 individuals belonging to 16 breeds and several wild populations. We found that both the initial domestication event and the subsequent process of breed formation culminated in losses of ~20% of the genetic diversity present in wild rabbits and domestic rabbits as a whole, respectively, suggesting that these bottlenecks were of equivalent magnitude. Despite the short time elapsed since breed diversification, we uncovered a well-defined breed structure in domestic rabbits. However, in contrast with other domesticated species we failed to detect deeper levels of structure, probably as a consequence of a recent and single domestication event. Finally, we found evidence for intrabreed stratification that is directly associated with demographic and selective causes such as formation of strains, colour morphs within the same breed, or country/breeder of origin. These additional layers of population structure within breeds should be taken into account in future mapping studies.

P-2110

Inferring natural selection by stratifying a single genome sequence

Naoki Osada
National Institute of Genetics, Mishima, Shizuoka, Japan

Identifying genomic regions under natural selection usually requires data from divergent species or polymorphisms within populations. Recently, a genome-scale sequential Markov coalescent method was efficiently implemented by Li and Durbin (2011). The algorithm stratifies individual diploid genome sequences into discrete time intervals of pairwise coalescent time between two chromosome fragments. I investigated whether the method could be useful for detecting natural selection acting on genomes. Although we have to be careful for some artifacts arising from the methods, the method found that the regions with unusually deep coalescence in human and macaque genomes are enriched with genes of particular functions such as immune response and metabolism and those genes contained many non-synonymous polymorphisms than other genes, suggesting that balancing-selection-like mechanism is acting on those regions. The pattern was robust against masking CpG sites, protein-coding sequences, and recently duplicated sequences. The method could be applicable to many other genomes of non-model organisms, of which polymorphism data is hard to obtain practically.

P-2111

Combinatorial properties of coalescent trees and relatives

Filippo Disanto, Thomas Wiehe
Universitaet zu Koeln, Koeln, Germany

Models in evolutionary biology are intimately linked to the tree paradigm. Given a direction by time, ancestry relationship between species, individuals, alleles or cells can be depicted as a rooted tree. Of particular interest are binary rooted unordered trees, such as coalescent trees. The coalescent with *n* leaves is a standard null model for the evolutionary history of a sample of *n* genes. In contrast to phylogenetic trees -which carry leaf labels- and to shape trees -which are unlabelled- only the internal nodes of coalescent trees are labelled. This leads to deep combinatorial differences between the different types of binary trees, regarding their enumeration, their equivalence to certain types of permutations, their metric and algebraic properties. We study coalescent trees introducing the technique of generating functions. With their help we derive results concerning several topological properties of coalescent trees, such as the distribution of the number of certain subtrees, numbers of so-called caterpillars, pitchforks and cherries. These distributions are valid for trees generated under the neutral coalescent. More importantly, they are independent of demographic changes, but they are affected by adaptive changes in the evolutionary history of a sample. Topological properties of coalescent trees harbour essential pieces of information which can be exploited to test the neutral evolution hypothesis.

P-2112

Correcting principal components of spatial population genetic variation under isolation by distance models

Eric Frichot[1], Sean Schoville[1], Guillaume Bouchard[0,2], Olivier François[1]
[1]Université Joseph Fourier, CNRS, TIMC-IMAG UMR 5525, Grenoble, France, [2]Xerox Research Center Europe, Grenoble, France

In many species, spatial population genetic variation displays patterns of isolation by distance. Characterized by locally correlated allele frequencies, these patterns are known to create periodic shapes in geographic maps of principal components which confound signatures of specific migration events and influence interpretations of principal component analyses (PCA). In this presentation, we introduce a new model combining probabilistic PCA and Bayesian kriging to infer population genetic structure from spatial genetic data while correcting for errors introduced by isolation by distance. The proposed algorithm is based on matrix factorization and low rank approximations, and scales with the dimension of the data set. To illustrate our method, we generated patterns of isolation by distance and broad-scale geographic clines using simulations of spatial Markov models. We show that our method improves the interpretation of PC maps, and is able to remove the horseshoe patterns usually observed in those maps for spatially correlated data. We present an application to human SNP data from the Human Genome Diversity Panel.

P-2113

Climate adaptation of flowering time in the Mediterranean model species *Medicago truncatula*

Concetta Burgarella[1], Nathalie Chantret[1], J. Marie Prosperi[1], Stephane De Mita[2], Peter Tiffin[3], Joelle Ronfort[1]
[1]INRA, Montpellier, France, [2]INRA, Nancy, France, [3]Univ. of Minnesota, St. Paul, USA

Flowering time is a phenotypic character potentially implied in the adaptive responses to different selective pressures like climate, pollinators, herbivores and pathogens. The Mediterranean basin, which is the area of distribution of the legume model *Medicago truncatula*, constitutes an ideal framework to address the effect of heterogeneous climatic conditions on adaptive characters. In this study, we examine the natural diversity of flowering genes in the annual autogamous *M. truncatula* and put it in relation with climatic variables and phenotypic characteristics. For this, we use next generation sequencing data for a sample of 192 accessions covering the whole distribution area of the species. First, we used a multivariate method (the Discriminant Analysis of Principal Components) and nucleotide polymorphism at 34,550 intergenic polymorphic sites distributed over the entire genome to check for the presence of neutral genetic structure within the sample. We found two major genetic groups, corresponding to western populations (from the Iberian Peninsula and Morocco) and eastern populations (from France and Algeria to the Middle East) respectively. Second, in each group, we applied different regression methods to identify sites potentially involved in the adaptation to climatic conditions. Preliminary results on a small subset of loci (330 SNP) point to 8-30% significant associations, depending on the method. These loci are potential candidates to be involved in local adaptation.

P-2114

Bayesian inference of the demographic history of Niger-Congo speaking populations

Isabel Alves[1,2], Lounès Chikhi[2,3], Laurent Excoffier[1,4]
[1]CMPG, Institute of Ecology and Evolution, Berne, Switzerland,[2]Population and Conservation Genetics Group, Instituto Gulbenkian de Ciência, Oeiras, Portugal,[3]CNRS, Université Paul Sabatier, ENFA, Toulouse, France,[4]Swiss Institute of Bioinformatics, Lausanne, Switzerland

The Niger-Congo phylum encompasses more than 1500 languages spread over sub-Saharan Africa. This current wide range is mostly due to the spread of Bantu-speaking people across sub-equatorial regions in the last 4000-5000 years. Although several genetic studies have focused on the evolutionary history of Bantu-speaking groups, much less effort has been put into the relationship between Bantu and non-Bantu Niger-Congo groups. Additionally, archaeological and linguistic evidence suggest that the spread of these populations occurred in distinct directions from the core region located in what is now the border between Nigeria and Cameroon towards West and South Africa, respectively. We have performed coalescent simulations within an approximate Bayesian computation (ABC) framework in order to statistically evaluate the relative probability of alternative models of the spread of Niger-Congo speakers and to infer demographic parameters underlying these important migration events. We have analysed 61 high-quality microsatellite markers, genotyped in 130 individuals from three Bantu and three non Bantu-speaking populations, representing a "Southern wave" or the Bantu expansion, and a "Western wave", respectively. Preliminary results suggest that models inspired by a spatial spread of the populations are better supported than classical isolation with migration (IM) models. We also find that Niger-Congo populations currently maintain high levels of gene flow with their neighbours, and that they expanded from a single source between 200 and 600 generations, even though available genetic data do not provide enough information to accurately infer these demographic parameters.

P-2115

The porcine colonization of the Americas: A 60k SNP story

William Burgos-Paz[1], Carla Souza[1,2], Hendrik-Jan Megens[3], Samuel Paiva[2], Martien Groenen[3], Miguel Pérez-Enciso[1,4], AmMap Consortium[0]
[1]Center for Research in Agricultural Genomics (CRAG) – Universitat Autònoma de Barcelona (UAB), Bellaterra, Spain, [2]Embrapa Recursos Genèticos e Biotecnologia – CENARGEN, Brasilia DF, Brazil, [3]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands, [4]Institut de Recerca i Estudis Avançats de Catalunya, ICREA, Barcelona, Spain

The pig, *Sus scrofa*, is a foreign species to the American continent. Although pigs originally introduced in the Americas should be related to those from the Iberian Peninsula, the phylogeny of current creole pigs is likely to be very complex. Because of the extreme climates that America harbors, these populations also provide a unique example of fast adaptation. Here, we provide a genome wide study of these issues by genotyping 60k SNP in 206 village pigs sampled across 14 countries and 183 pigs from outgroup breeds that are potential founders of the American populations, including Chinese breeds. Results show that American village pigs are primarily of European ancestry, although the observed genetic landscape is that of a complex conglomerate. There was no correlation between genetic and geographical distances, neither continent wide nor when analyzing specific areas. Most populations showed a clear admixed structure where the Iberian pig was not necessarily the main component: international breeds, but also Chinese pigs, have contributed to extant genetic composition of village pigs. In fact, our data support that the influence of Asia in village American pigs may be much more widespread than previously acknowledged, especially in the Caribbean and in Brazil. In response to altitude, many genes related to the cardiovascular system have increased differentiation between altiplano and genetically related pigs living near sea level. When comparing village vs. European pigs, genes involved in limb development seem to have played a major role, whereas a variety of signals is observed in miniature pigs.

AmMap consortium members: Y. Ramayo-Caldas (Spain/Cuba), M. Melo (Peru), C. Lemús-Flores (Mexico), H.W. Soto (C Rica), R. Martínez and L.A. Álvarez (Colombia), V. Iñiguez (Bolivia), M.A. Revidatti (Argentina), O.R. Martínez-López (Paraguay), A. Esteve-Codina (Spain), R.P.M.A. Crooijmans (Holland), L.B. Schook (USA)

P-2116

The worldwide spread of transposable element insertions associated with pesticide exposure and impact on linked polymorphism in Drosophila simulans.

Pierre Gerard, Francois Wurmser, Romain Fougeyrollas, David Ogereau, Catherine Montchamp-Moreau
Lab Evolution Genome Speciation, CNRS-Universite Paris-Sud, Gif-sur-Yvette, France

The evolution of insecticide resistance is a good example of rapid adaptation to human-mediated environmental changes. The upregulation of the Cytochrome P450 gene Cyp6g1 associated with a transposon insertion in Drosophila melanogaster, and the subsequent correlation with DDT resistance, has become a classic case of adaptation involving cis-regulatory changes and gene expression variation. Using high-throughput gene expression analysis in a comparative study between a population from the African ancestral range of Drosophila simulans and a derived population from Europe, we revealed a strong increase in expression of detoxification genes in Europe, such as Gluthatione transferases and Cytochromes P450. We examined the worldwide polymorphism in transposable element insertions at the Cyp6g1 locus and show that different elements have been recruited and spread in regions where pesticide exposure should have been high. We further analyzed the polymorphism in the Cyp6g1 region and at other non-coding loci scattered along the second chromosome to detect potential selective sweeps associated with the spread of the insertions in anthropized environments.

P-2117

Transformations of the site frequency spectrum and application to human populations

Alexander Klassmann, Thomas Wiehe
Universitaet zu Koeln, Koeln, Germany

Some commonly used population genetic tests on neutrality such as Tajima's D
rely on weighted linear combinations of site frequency estimators
with equal expectation value Theta (Achaz 2009). Ferretti (2010) explored generalisations of such
tests. We follow a similar line and show that deriving the tests as a function of estimators with equal variance
leads to weighting schemes which depend on Theta. In whole-genome scans, Theta depends in turn on the chosen
sliding window size, which is usually sought to minimize recombination within a window. In simulations and data from
the 1000 genome project we investigate the interplay of the two effects.

P-2118

Population genomics of threespine stickleback after 30-year evolution experiment

Nadezhda Terekhanova[1], Nicolai Mugue[2], Georgii Bazykin[1,4], Alexey Kondrashov[1,3]
[1]Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow, Russia, [2]Institute of Developmental Biology RAS / VNIRO, Moscow, Russia, [3]Life Sciences Institute, Michigan, USA, [4]A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia

Threespine stickleback G. aculeatus is a model species in the studies of adaptation. It exists in two phenotypic forms, freshwater and marine, that differ in their morphological, physiological and behavioral traits, and these traits have evolved repeatedly in freshwater populations from marine ancestors all over the world. We use next-generation sequencing to study the genetic variation of the natural marine and freshwater stickleback populations, and of experimental freshwater populations formed 30 years ago from crosses of marine and freshwater ancestors. We confirm the previously identified QTLs, and discover novel QTLs responsible for stickleback freshwater adaptation. Furthermore, our results show that in the experimental freshwater populations, strong selection has favored the freshwater variants in these QTLs. Over the 30 years of the experiment, selection has increased the mean frequency of the freshwater alleles in most of these loci from 50% to 70-80%. Selection was especially prominent in the experiment where the founder population consisted of 20 individuals; by contrast, drift prevailed in the experimental population founded by a cross of only two individuals.

P-2119

A genetic study of skin pigmentation variation in India

Mircea Iliescu[1], Chandana Basu Mallick[2,3], Niraj Rai[4], Anshuman Mishra[4], Gyaneshwer Chaubey[2], Rakesh Tamang[4], Märt Möls[3], Rie Goto[1], Georgi Hudjashov[2,3], Srilakshmi Raj[1], Ramasamy Pitchappan[5], CG Nicholas Mascie-Taylor[1], Lalji Singh[4,6], Marta Mirazon-Lahr[7], Mait Metspalu[2,3], Kumarasamy Thangaraj[4], Toomas Kivisild[1,3]
[1]Division of Biological Anthropology, University of Cambridge, Cambridge, UK, [2]Evolutionary Biology Group, Estonian Biocentre, Tartu, Estonia, [3]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, [4]Centre for Cellular and Molecular Biology, Hyderabad, India, [5]Chettinad Academy of Research and Education, Chettinad Health City, Chennai, India, [6]Banaras Hindu University, Varanasi, India, [7]Leverhulme Centre for Human Evolutionary Studies, Division of Biological Anthropology, University of Cambridge, Cambridge, UK

Human skin colour is a polygenic trait that is primarily determined by the amount and type of melanin produced in the skin. The pigmentation variation between human populations across the world is highly correlated with geographic latitude and the amount of UV radiation. Association studies together with research involving different model organisms and coat colour variation have largely contributed to the identification of more than 378 pigmentation candidate genes. These include TYR OCA2, that are known to cause albinism, MC1R responsible for the red hair phenotype, and genes such as MATP, SLC24A5 and ASIP that are involved in normal pigmentation variation. In particular, SLC24A5 has been shown to explain one third of the pigmentation difference between Europeans and Africans. However, the same gene cannot explain the lighter East Asian phenotype; therefore, light pigmentation could be the result of convergent evolution. A study on UK residents of Pakistani, Indian and Bangladeshi descent found significant association of SLC24A5, SLC45A2 and TYR genes with skin colour. While these genes may explain a significant proportion of interethnic differences in skin colour, it is not clear how much variation such genes explain within Indian populations who are known for their high level of diversity of pigmentation. We have tested 15 candidate SNPs for association with melanin index in a large sample of 1300 individuals, from three related castes native to South India. Using logistic regression model we found that SLC24A5 functional SNP, rs1426654, is strongly associated with pigmentation in our sample and explains alone more than half of the skin colour difference between the light and the dark group of individuals. Conversely, the other tested SNPs fail to show any significance; this strongly argues in favour of one gene having a major effect on skin pigmentation within ethnic groups of South India, with other genes having small additional effects on this trait. We genotyped the SLC24A5 variant in over 40 populations across India and found that latitudinal differences alone cannot explain its frequency patterns in the subcontinent. Key questions arising from this research are when and where did the light skin variant enter South Asia and the manner and reason for it spreading across the Indian sub-continent. Hence, a comprehensive view of skin colour evolution requires that in depth sequence information be corroborated with population (genetic) history and with ancient DNA data of past populations of Eurasia.

P-2120

Stepwise colonization of the Andes by Ruddy Ducks dispersing from North America accompanied by the evolution of novel beta-globin variants

Violeta Munoz-Fuentes[1,4], Maria Cortázar Chinarro[4], Maria Lozano[3], Kevin McCracken[2]
[1]Estacion Biologica de Donana-CSIC, Sevilla, Spain, [2]University of Alaska Fairbanks, Alaska, USA, [3]Universidad de Los Andes, Bogota, Colombia, [4]Uppsala University, Uppsala, Sweden

Andean uplift played a key role in Neotropical bird diversification. Yet patterns of dispersal to and from montane habitats as well as genetic adaptation to high-altitude environments, remain little understood. Here we use multilocus population genetics to study historical gene flow in the Ruddy Duck (*Oxyura jamaicensis*), distributed from southern Canada to Tierra del Fuego and inhabiting wetlands from sea level to 4,500 meters in the Andes. We sequenced the mitochondrial DNA control region (mtDNA), four autosomal introns and three hemoglobin genes and used isolation-with-migration (IM) models to study gene flow between North America and South America, and between the northern and southern Andes. Our analyses indicated that Ruddy Ducks colonized the Andes in a stepwise fashion, first from North America into high-altitude regions in the northern Andes, then by secondary dispersal to low-altitude regions in the southern Andes. While no nonsynonymous substitutions were found in either alpha-globin, three amino acid substitutions were observed in the beta-globin. Based on phylogenetic reconstruction and power analysis, the most parsimonious explanation is that the first β^A-globin substitution (Ser-b69), found in all Andean individuals, was acquired when Ruddy Ducks expanded their range from low-altitude habitats in North America to high-altitude sites in the northern Andes, while the two additional amino acid substitutions (Ser-b13, Ile-b14), found in birds from the lowlands of the southern Andes, occurred more recently, when Ruddy Ducks dispersed from high-altitude sites in the northern Andes to low-altitude sites in the southern Andes. Stepwise colonization of high- and low-altitude habitats coupled with amino acid replacements in the β^A-globin suggests that Ruddy Ducks first adapted to the Andean highlands and then again to the lowlands. In addition, our results showed that Ruddy Ducks colonized the Andes via a less common route as compared to most other waterbird species that likely colonized the Andes northwards from the southern cone of South America.

P-2121

Genetic Differentiation and Local Adaptation of an Ecologically Dominant Prairie Grass *Andropogon gerardii* (Big Bluestem) Occurring Along a Natural Precipitation Gradient in Midwest US Grasslands

Loretta Johnson[1], Miranda Gray[1], Paul St Amand[2], Eduard Akhunov[1], Karen Garrett[1], Mary Knapp[1], Ted Morgan[1], Bai Guihua[2], Sara Baer[4], Brian Maricle[3], Hannah Tetreault[1]
[1]Kansas State University, Manhattan KS, USA, [2]United States Department of Agriculture Agricultural Research Service, Hard Winter Wheat Genetics Research Unit,, Manhattan KS, USA, [3]Fort Hays State University, Hays KS, USA, [4]Southern Illinois University, Carbondale Illinois, USA

Big bluestem is a widely-distributed dominant grass in Midwest US grasslands. Its productivity is dependent upon precipitation, and with wide distribution across a sharp precipitation gradient (400-1200mm yr^{-1}), we expect ecotypic variation in drought tolerance and potentially, local adaptation. A better understanding of ecotypic variation will help predict how a dominant prairie grass may respond to climate change and will inform prairie restoration. We investigate the linkage of phenotypic variation and genetic diversity using reciprocal common gardens across the precipitation gradient. Sites were planted in Carbondale, Illinois, Manhattan and Hays KS and a site in Colby, KS (to test limits of tolerance into drier areas). At these locations, plants of three ecotypes (each comprised of seeds collected from four pristine populations in Hays, Manhattan, and Illinois) were reciprocally planted in replicated seeded communities ($16m^2$ plots). We measured phenotypic variation in drought tolerance across ecotypes and sites. Because genetic diversity may be critical for predicting a species' ability to adjust/adapt to climate change, we assess genetic diversity and population differentiation (using AFLP markers) of Andropogon gerardii in the 12 source populations used in the reciprocal gardens. Our data demonstrate a strong phenotypic cline in drought tolerance of the three ecotypes. Establishment and cover in the seeded plots showed a significant ecotype ($p<0.0001$), site ($p<0.0001$) and interaction effect ($p<0.0001$). The Hays ecotype had disproportionate cover in western regions relative to the Illinois and Manhattan ecotypes (GXE), indicating local adaptation to drought. Thus, the Hays ecotype had 2x and 3x the cover compared to the other ecotypes in Hays and Colby, respectively. We analyzed 483 plants (>20 plants per population) and identified 387 markers using 2 primers. Several lines of evidence (neighbor joining trees of genetic similarity, STRUCTURE and PCA) provide support for genetic differentiation of the ecotypes. Based on PCA, 47% of the variation in the axis one eigenvalue was explained by mean annual rainfall. Based on outlier Fst analyses using Bayescan, we identified 11 ecotype-specific loci under diversifying selection. In spite of the genetic differentiation among ecotypes, molecular analyses of variance indicates that most of the genetic variation lies within populations. This high, within-population, genetic diversity may allow bluestem populations to withstand diverse climate change and has implications for prairie restoration. Next-generation sequencing studies of functional transcriptomic variation among ecotypes in response to drought are ongoing.

P-2122

Purifying and positive selection in the *Medicago truncatula* genome

Timothy Paape[1], Peng Zhao[1], Thomas Batiallon[2], Nevin Young[1], Peter Tiffin[1]
[1]University of Minnesota, Saint Paul, MN, USA, [2]University of Denmark-Aarhus, Aarhus, Denmark

Using whole genome re-sequencing data from 56 accessions of the model legume *Medicago truncatula* and the outgroup *M. sativa* (alfalfa), we conducted genomic scans to i) identify targets of positive selection as well as gene classes that have evolved under strong purifying selection in 20,000 annotated coding sequences, ii) use MKtests and estimates of deleterious fitness effects to estimate proportions of adaptive and slightly deleterious substitutions correcting for demographic changes, and iii) a combined neutrality statistic using Tajima's D and Fay Wu's H tests to identify genes that have undergone selective sweeps. We examine results from these analyses with expression patterns (total expression and tissue-specific expression) to evaluate relationships between expression and selective constraints. Finally, to empirically evaluate the effect that weak population structure and sampling stochasticity may have on results of these analyses we compare the results from the entire sample to results from subsets that partition weak population structure that is present in our range-wide sample.

P-2123

Robust Codon Usage in Protein Ligand Binding Sites

Tugce Bilgin[1,3], Isil Aksan Kurnaz[4], Turkan Haliloglu[5], Andreas Wagner[1,2]
[1]Institute of Evolutionary Biology and Environmental Sciences, University of Zurich, Zürich, Switzerland,[2]The Santa Fe Institute, New Mexico, USA,[3]The Swiss Institute of Bioinformatics, Zürich, Switzerland,[4]Department of Genetics and Bioengineering, Yeditepe University, Istanbul, Turkey,[5]Polymer Research Center, Bogazici University, Istanbul, Turkey

Codon usage bias is associated with various biological factors, such as gene expression level, gene length, translational accuracy, protein amino acid composition, protein structure, tRNA abundance, mutation frequency and patterns, and GC compositions. Here, we present a confounding factor in explaining codon usage bias from an evolutionary point of view. Codons vary in their tolerance to error, that is, the degree of change in the encoded amino acid after a point mutation. We refer to this property as codon robustness and suggest that depending on their robustness, codons are used differentially. More specifically, we focused in ligand binding sites of proteins, where evolutionary forces act more strongly.

We first developed a measure of codon robustness, that evaluates for each codon the phenotypic change after three successive point mutations based on an amino acid substitution matrix. Next, we analyzed 275 non homologous exons from various organisms, whose encoded proteins are known to bind to only one of the 24 ligands we selected. We chose the exons based on the data availability, including protein crystall structure. We determined ligand binding sites based on the ligand proximity to residue. A comparison of the codon robustness in binding sites and non binding sites demonstrates that codons encoding ligand binding sites of proteins tend to be more robust compared to the codons encoding non binding sites. We also compared accessive surface area, amino acid conservation and hydrophobic index for binding and non binding sites of proteins. The codon robustness seems to be more distinguishing ($P<10^{-39}$) than the accessive surface area ($P<10^{-26}$) and the hydrophobic index ($P=0.99$), but less distinguishing than the conservation ($P<10^{-257}$) for binding and non binding residues. Overall, our results suggest that codon robustness may not only bias codon usage, but it may also help to better distinguish between protein ligand binding and non binding sites. The latter observation makes codon robustness a promising target for protein binding site predictions especially for prediction algorithms based solely on DNA sequence.

P-2124

Functional evolution of nociceptive receptors TRPA1 and TRPV1 in vertebrates

Shigeru Saito[1], Kazumasa Nakatsuka[2], Masashi Ohkita[2], Kenji Takahashi[2], Naomi Fukuta[1], Toshio Ohta[2], Makoto Tominaga[1,3]
[1]Okazaki Institute for Integrative Bioscience (National Institute for Physiological Sciences), Okazaki, Aichi, Japan, [2]Tottori University, Tottori, Tottori, Japan, [3]The Graduate University for Advanced studies, Okazaki, Aichi, Japan

TRP channels perceive a wide variety of sensory stimuli, serving as multimodal receptors in vertebrates. Among these, TRPA1 and TRPV1 perceive noxious temperatures and chemical stimuli and are involved in pain sensations in mammals. These two channels thus provide a model for understanding how different genes with similar biological roles may influence the function of one another during the course of evolution. In mammals, both TRPA1 and TRPV1 are activated by noxious chemicals. In contrast, regarding temperature sensitivities, TRPA1 is activated by cold, while TRPV1 is activated by heat. To elucidate the functional evolution of these two channels in vertebrates, we cloned them from western clawed (WC) frog, which diverged earliest among terrestrial vertebrates and characterized channel properties. We also cloned TRPA1 from green anole lizard.

Both TRPA1 and TRPV1 of WC frog were activated by heat stimulation with activation temperature threshold for 40℃ for TRPA1 and 38℃ for TRPV1. We also found that bo th channels were functionally co-expressed in the same native sensory neurons similar to mammals. Given that temperature above 38℃ elicited nocifensive behaviors i n WC frog, both TRPA1 and TRPV1 cooperatively serve as noxious heat receptors. In addition, green anole TRPA1 was activated by heat stimulation. TRPA1s of WC frog and green anole were activated by noxious chemicals that activate mammalian TRPA1, indicating that noxious chemical sensitivity has been conserved throughout vertebrate evolution. Taking into consideration that *Drosophila* TRPA1 is also heat- and chemical-activated channel, TRPA1 acquired heat and chemical sensitivities in an early stage of animal evolution and these traits were retained in ancestral vertebrates. Phylogenetic analysis showed that TRPV1 gene emerged in ancestral vertebrates, suggesting that TRPV1 became co-expressed with the pre-existing TRPA1 in the same sensory neurons at latest in the common ancestor of tetrapods and have been conserved its heat sensitivity during vertebrate evolution.

Emergence of TRPV1 as a novel noxious heat receptor may have significantly influenced the functional evolution of TRPA1, regarding temperature sensitivity, resulting in different evolutionary consequences in the respective vertebrate lineages, such as loss of temperature sensitivity in zebrafish, change in temperature sensitivity in mammals, co-option in a novel thermosensory organ (pit) in snakes, and functional compensation in WC frog. Here we discuss the significance of the evolution of two nociceptive receptors during the course of vertebrate evolution.

P-2125

Genomics and physiology to investigate the lipid metabolism of 9 oleaginous yeasts

Stéphanie Michely[1], Claude Gaillardin[2], Jean-Marc Nicaud[3], Cécile Neuvéglise[1]
[1]INRA, UMR 1319 Micalis, Jouy-en-Josas, France,[2]AgroParisTech, Micalis, Jouy-en-Josas, France,[3]CNRS, Micalis, Jouy-en-Josas, France

Yarrowia lipolytica belongs to a group of yeasts that has diverged very early from most other hemiascomycetous yeasts. This yeast is able to use various hydrophobic substrates as unique carbon source and to synthesize new free fatty acid from non hydrophobic compounds. These characteristics make *Y. lipolytica* a known oleaginous model for the lipid metabolism survey of yeasts.

Its genome has been entirely sequenced within the framework of the Génolevures consortium. This sequencing highlighted several peculiarities of its genome organisation, clearly setting it apart from most other hemiascomycetous yeasts so far analysed: unusually large genome size, atypical organisation of tRNA genes, relatively high frequency of spliceosomal introns and diversified content of transposable elements.

Several yeast species recently described were proposed to be close relatives of *Y. lipolytica*. All of them are from different geographical and biological origins. Here we study, for 9 of them, physiological characteristics such as their growth capacities on different media (lipids, alkanes and sugars), their fatty acid synthesis and storage capacities as well as their own lipid composition. As an example we discovered that these 9 yeasts were able to grow on fatty acids whereas some of them had lost the capacity to grow on alkanes. These findings highlight differences at the lipid metabolism level.

The recent sequencing of 5 genomes within the *Yarrowia* clade enables to search homologues of *Y. lipolytica* genes known to be implicated in the lipid metabolism and allows to study their genetic environment, to compare these data with the physiological characteristics and finally to deduce the relative rule of the known mechanisms of evolution (gene duplications, mutations, chromosomal rearrangements). The contractions and expansions of key protein families (lipases, cytochrome P450, acyl-CoA oxydases, etc.) and their synteny are particularly investigated.

The combination of these developed approaches will improve the comprehension of both evolution and adaptation of the lipid metabolism of these different yeasts within the *Yarrowia* clade.

P-2126

Impact of the *M. tuberculosis* Genetic Background on the Acquisition of Drug Resistance-conferring Mutations

Sonia Borrell[1], Graham Rose[2], Julia Feldmann[1], Sebastien Gagneux[1]
[1]Swiss Tropical and Public Health Institute. University of Basel., Basel, Switzerland, [2]MRC National Institute for Medical Research., London, UK

The extent of strain genetic diversity in *M.tuberculosis* (*Mtb*) is more pronounced than was traditionally believed, being classified into different geographically distributed lineages. One of these lineages; the Beijing lineage (L2) has repeatedly been associated with drug resistant tuberculosis (DR-TB) and has been reported as being more likely to harbour deleterious mutations in *katG*, the only known catalase-peroxidase gene in *Mtb*. Mutations (DRm) in this region have been related to confer high level of isoniazid (INH) resistance. Another phylogenetical related lineage; the Indo-Oceanic lineage (L1) has been related to harbour more frequently a less costly low-level INH-resistance-conferring mutation in -15*InhApro*. Altogether suggests that lineages genetic differences, either due to epistasis or to a better adaptation to the physiological effects of INH-resistant, might influence the propensity of *Mtb* towards particular INH-DRms.

To evaluate the influence of the genetic background on the acquisition of specific INH-DRms, we performed a Lüria-Delbruck fluctuation assay on regular medium with and without catalase at two different INH concentrations, using two pan-susceptible clinical strains belonging to Lineage 1 and 2 of *Mtb*, respectively. We found no lineage-specific differences in mutation rate with respect to the different types of INH-DRms. Interestingly, no *inhA* promoter mutants were isolated, illustrating the differences between the *in vitro* compared to the *in clinico* environments. As both lineages showed different requirements for catalase during *in vitro* growth, the complete RNAseq transcriptional profiles of the two strain backgrounds were compared. We found significant lineage-specific differences in expression of *ahpC* and *katG*, suggesting that *Mtb* lineages might differ in their baseline capacity to respond to oxidative stress. We are exploring this possibility using various experimental models.

Overall, our findings support a role for *Mtb* lineage diversity in the emergence of drug resistance in Tuberculosis.

P-2127

Functional retrogenes - identification and analysis.

Joanna Ciomborowska, Michal Kabza, Kamil Sambor, Izabela Makalowska
Adam Mickiewicz University, Faculty of Biology, Laboratory of Bioinformatics, Poznan, Poland

Retrogenes are created through retroposition in which mRNA is reversely transcribed into cDNA and in this form is inserted into a new place in the genome. Majority of created in that way retrocopies end up as pseudogenes. However, many studies show that they often become a source of new genes, protein domains or regulatory elements. At the moment it is estimated that human genome contains about 19,000 retrocopies from which about 2% are considered as putative functional retrogenes. At the same time, a genome-wide study showed that 20% of mammalian protein encoding genes lack introns in their ORFs. Therefore, it is conceivable that many genes lacking introns arose by retroposition and the fraction of functional retrogenes is higher than estimated.

Identification of functional retrogenes is a challenging task and usually related to two main steps: (i) looking for pairs of putative retrogene and its parental counterpart in the genome and (ii) searching for evidences of retrogenes expression to confirm functionality. We applied new and effective methods combining various elements to look for functional retrogenes in human and other animal genomes. In our approach the identification was performed at two levels: genomic and orthology groups level. At the first level, our method was utilizing Splign and BLAST to map all multi exon genes (potential parental genes) to the genome. Then, results were strictly filtered to get the most likely pairs of parental gene and its retrocopies (single exon genes). Next we mapped to all retrogenes RNA-Seq reads from several libraries in order to obtain information about their expression. For some of identified human retrogenes we performed experimental studies and confirmed their functionality.

Second way of identification was based on orthology groups. In this step we analyzed orthologous pairs from OMA database under the assumption that retroposition is associated with intron loss. We compared the structure and genomic localization of human single exon genes and assigned multi exon orthologs from several animal species. That gave us an opportunity to follow the evolutionary history of retroposition as well as identify retrogenes that lost their parental genes.

Our analyses revealed several hundreds of novel potentially functional retrogenes in human genome. We also showed that numerous retrogenes replaced their parental genes.

P-2128

Amino acid polymorphism at *Drosophila methuselah*-like genes is associated with lifespan differences.

Micael Reis, Ana Araújo, Bruno Aguiar, Helder Rocha, Cristina Vieira, Jorge Vieira
IBMC, Porto, Portugal

The study of traditional model species only makes sense if the findings can be generalized to distantly related ones. However when a candidate gene belongs to a gene family it is conceivable that only paralogous copies of the candidate gene can be found in more divergent species. Thus it is unclear whether other distant members of a gene family should be considered candidate genes as well. This is the case of the *Drosophila melanogaster methuselah* (*mth*) gene which is a member of a family comprising 16 elements. This gene encodes a putative G protein-coupled receptor (GPCR) required in the presynaptic motor neuron to acutely upregulate neurotransmitter exocytosis and naturally occurring aminoacid variation has been previously associated with lifespan differences. In this work, we show that *mth* has an estimated age of 10 million years since it is present only in the *melanogaster* subgroup of species. So we were interested in determining if natural occurring variation at *mth* paralogous genes (*mth*-like) could explain lifespan differences in a species (*D. americana*) that has been diverging from *D. melanogaster* for 40 million years. Here it is shown that *mth*-like genes encoding proteins that share less than 50% amino acid identities with the *D. melanogaster* Mth protein show aminoacid polymorphism which explains a considerable amount of the lifespan differences observed in a F2 association experiment. For *D. americana mth*-like gene *GJ23561*, we found a likely truncated allele that decreases lifespan by 22.3%, and an amino acid polymorphism, present at about 55% frequency in natural populations, that in heterozygosity increases lifespan by 12.0%. For *GJ12490* gene, individuals having at least one allele that codes for a protein without a highly conserved putative N-glycosilation site (present at about 50-67% frequency in natural populations) live 21.7% longer (although they take longer to develop and are smaller) than individuals with two functional alleles. We finally suggest that lifespan is a by-product of selection for other phenotypic traits such as body size and developmental time.

P-2129

Why have sex? Mutational meltdown in asexual organisms

Philipp H. Schiffer, Einhard Schierenberg
Institute for Zoology, University of Cologne, Cologne, Rheinland, Germany

Despite the two-fold cost of sex an overwhelming number of metazoan species reproduce sexually. Among theories for mechanisms depriving parthenogenetic species of their two-fold advantage mutational meltdown is the most prominent one. In the absence of outcrossing slightly deleterious mutations will accumulate in genomes on an evolutionary very short time-scale culminating in non-viable genotypes species-wide. Thus, asexual taxa should be in existence only transiently. To counteract such an effect and evade short-term extinction, it has been hypothesised, that parthenogenetic species have lowered mutation rates. However, to be a general mechanism, this implies the independent evolution of decreased mutation rates in a large number of diverse taxa after spontaneous transition to asexuality. At the same time reduced mutation rates would hamper evolutionary change and thus decrease the chance of adaptation (and consequently further speciation) in nascent parthenogenetic species. To date mutation rates in asexual taxa could only be inferred phylogenetically from analysis of a small number of genes coupled to molecular-clock estimates. Now however, with the advent of 2nd generation sequencing, it is possible to directly measure mutation accumulation on a whole-genome scale under controlled laboratory conditions. Here I present data from an on-going mutation rate analysis in nematodes of the genus *Panagrolaimus*, bi-partitioned in: a) A mutation accumulation (MA-line) experiment conducted in both the parthenogenetic species, *P.* strain PS1159, and an androdioecious species, *P.* strain JU765. b) Genome sequences from several closely related parthenogenetic and dioecious *Panagrolaimus* strains of different geographic origin that will elucidate the temporal origin of parthenogenesis in this genus and the adaptive potential of asexual taxa. Correlating results from both analyses makes it possible to gain important insights on the longstanding question why having sex is better than to not having sex.

P-2130

A Screen for QTLs that Influence the Degree of Morphological Variation in Natural Yeast Isolates

Kerry Geiler-Samerotte, Naomi Ziv, Mark Siegal
New York University, New York, New York, USA

Phenotypic variation caused by heritable genetic differences is an essential prerequisite for evolution by natural selection. However, some genes can buffer phenotypic variation by masking the effects of other loci. Genes that buffer or reveal phenotypic variation may affect an organism's evolvability, defined here as the capacity to generate heritable phenotypic variation that can be acted upon by natural selection. We perform a screen to detect quantitative trait loci that influence levels of phenotypic variation within clonal populations and between segregating genotypes. To maximize QTL detection, we use wild, genetically divergent yeast strains as the parents of our mapping family as they may contain genetic variation that has not been previously studied in laboratory strains. We use the Calmorph software package to quantify 220 morphological phenotypes per cell, reducing significantly redundant phenotypes via principal component analysis. We report a set of QTLs that contribute to cell morphology, as well as a set of QTLs that influence the degree of morphological variability within and between the segregants of this cross.

P-2131

Genetic response of *Daphnia pulicaria* to historic pH changes in lakes

Billy Culver[1,2], Philip Morton[2], Dagmar Frisch[2], Lawrence Weider[1,2]
[1]*University of Oklahoma, Norman, OK, USA,* [2]*University of Oklahoma Biological Station, Kingston, OK, USA*

Organisms are subjected to a variety of environmental stresses in which they must respond in order to survive and reproduce. Some are able to adjust to these stresses, while others do not and are extirpated. Although it is known that organisms can respond to environmental stress, the underlying physiological and genetic mechanisms are often not well understood. Using "resurrection ecology" and genomic tools from the recently sequenced and annotated genome of the model aquatic organism, *Daphnia pulex*, this study is attempting to find the link between historical pH changes in Madison Lake, Minnesota and long-term temporal genetic variation in *D.pulicaria*, a sister species of *D. pulex.* Furthermore, we want to elucidate the genetic mechanisms that allow *D. pulicaria* to maintain a stable population through time when confronted with periods of pH change.

One meter sediment core samples using Loss-on-Ignition analysis indicate that calcium carbonate (i.e. $CaCO_3$) levels increased between the 25-cm to 47-cm depths. Using a conservative estimate of sedimentation rate of 1-cm yr^{-1}, this corresponds to 25-47 years ago (recent Pb^{210} dating of a sister lake suggests that these dates may be a 2-fold underestimate; i.e. actual ages of 50-94 years ago). $CaCO_3$ was used as an indicator of pH level during the period of time retrieved by the core sample. During the 25-cm to 47-cm period, mean pH was at a higher level when compared to present day conditions. Using 17 *D. pulex/pulicaria*-specific microsatellite loci from "resurrected" hatchlings from all layers of the sediment, population genetic structure was analyzed through time. Additionally, genetic structure was analyzed for correlation with pH history.

Evidence suggests that acid-base regulation within an organism is controlled by carbonic anhydrase (CA). In studies of fish, three isoforms of CA are known to play a role in acid-base regulation. However, 31 isoforms have been resolved in *D. pulex.* In order to ascertain which isoforms are analogous to those in fish and daphniids, a maximum-likelihood phylogeny of the CA genes was constructed using Mr.Bayes. The results of this tree show that CA1, CA2, and CA5 are analogous to those genes in fish that control acid-base regulation. Primers were designed for these genes based on the *D. pulex* genome. The CA genes were sequenced and characterized for this study. CA genes may be informative in understanding the mechanisms that allow organisms like *Daphnia* to persist through time when faced with changing pH conditions in their aquatic habitat.

P-2132

Evolving on phenotype landscapes

Jose A Cuesta[1], Susanna C Manrubia[2]
[1]Universidad Carlos III de Madrid, Leganes, Madrid, Spain,[2]Centro de Astrobiología (INTA-CSIC), Torrejon de Ardoz, Madrid, Spain

Genotype landscapes are a very useful tool upon which most evolutionary models rely. But they also are a source of potential misunderstandings. The reason is that the genotype-to-phenotype map is highly degenerated. Huge patches of the genotype landscape (so-called neutral networks) correspond to just a single phenotype and therefore natural selection is blind to genotypic differences within the same patch. If genotype landscapes are patchworks of neutral networks, the spreading of a molecular quasi-species over this landscape by mutations will exhibit quite an unusual behavior that should reflect the topology of neutral networks as well as the accessibility of alternative phenotypes. In this work we develop a simplified model of phenotype landscape inspired by quantitative studies of neutral networks of RNA. In a first approximation, this landscape can be viewed as a network of interconnected phenotypes, each defining a node of the network. Individuals with the same phenotype reproduce at the same rate. Besides, the population may jump from a given phenotype to a neighbor one, but the rate at which this transition occurs is derived from the intrinsic properties of the corresponding neutral network. Using our current knowledge on RNA secondary structure neutral networks, it turns out that the jumping rate is determined both by the size of the neutral network and by the time the quasi-species has spent on it. Remarkably, this latter feature renders the evolutionary process non-Markovian. In this sense, a more accurate representation of each phenotype is not as a node in a network, but as a series of layers, each characterized by a transition probability to other layers in the same phenotype and to other phenotypes that depends on the current state of each individual in the quasi-species. We explore the implications of this phenotype-based evolutionary model for the adaptability of quasi-species as well as for phylogeny.

P-2133

Why do so many fungal open reading frames contain repeats?

Jan Schmid[1], Matt Wilkins[2], Ningxin Zhang[1], Rosie Bradshaw[1], Murray Cox[1], Richard Cannon[3], Chris Schardl[4]
[1]Massey University, Palmerston North, New Zealand, [2]Scientific Consulting, Palmerston North, New Zealand, [3]University of Otago, Dunedin, New Zealand, [4]University of Kentucky, Lexington, KY, USA

Tandem repeat-containing DNA mutates orders of magnitude faster than normal DNA by insertion and deletion of repeat units. In microbes, hypermutable DNA repeats, when located in open reading frames (ORFs) or promoters in so-called contingency genes, can serve as an 'insurance'. Their hypermutation constantly replenishes a pool of variants in the population, from which the fittest are chosen. If circumstances change suddenly, for instance if the host launches an immune response against the predominant allele of a contingency gene of a microbial pathogen, the very survival of the population may depend on the presence of rare variant alleles not targeted by this response. Hypermutability is a cost however when the environment is not changing, by constantly generating inferior alleles. Research by us and others has shown that the pathogenic yeast *Candida albicans* has 2600 hypermutable tandem repeat-containing ORFs (TR-ORFS), but that most of these may not be contingency genes used in short-term adaptation: Isolates of similar genetic background tend to have similar or identical TR-ORF alleles, even when isolated from different patient types and anatomical locations. Furthermore a genome- wide survey revealed that in the majority of TR-ORFs synonymous mutations reduce repetitiveness and thus mutability of the DNA, while retaining the encoded amino acid repeat. A survey of a set of other fungi (*Saccharomyces Epichloë festucae, Cladosporium, Dothistroma),* suggests that TR-ORFs are frequent in fungal genomes and that reduction of mutability by synonymous mutations is common. This suggests that TR-ORFs in fungi may often function to speed up adaptation by increasing the rate of evolution of novel proteins or for generating genetic background- specific alleles that work well under a variety of circumstances rather than as contingency genes. However, in *C. albicans* we observe that the association of specific alleles and specific genetic backgrounds often breaks down for a small number of TR-ORFs with high predicted mutability (VAR scores) and for which predicted mutability is not diminished by synonymous mutations, an indication that these TR-ORFs might be contingency genes. The latter result also suggest that combining in silico predictions of mutability and the degree of its reduction by synonymous mutations might be a useful tool for detecting contingency genes in fungal genomes.

P-2134

Rapid evolution in lipid and metabolic composition of the human brain

Kasia Bozek[1], Patrick Giavalisco[2], Philipp Khaitovich[1]
[1]Partner Institute for Computational Biology MPG-CAS, Shanghai, China,[2]Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

Despite strong phenotypic differences between humans and other primates, these species show only minor genetic divergence. Genetic variability among primates has been an object of numerous studies. Metabolic and lipid composition differences between humans and other primates remained largely unexplored. In this study we analyzed metabolite and lipid concentration levels in approximately 400 samples from humans, chimpanzees, macaques and mice from five tissues: prefrontal, visual and cerebellar cortexes of the brain, as well as muscle and kidney. Based on liquid chromatography and gas chromatography coupled with mass spectrometry distinguishing over 100 000 metabolite and lipid peaks, we identified significant excess of human-specific changes in the brain regions going beyond the genetic divergence of the human lineage. Specifically, we observed a pronounced human-specific divergence of metabolite and lipid levels in the prefrontal cortex that might contribute to the evolution of human cognitive abilities. We additionally assessed the environmental effects on the lipid and metabolite composition of the tissues by including control measurements of macaques subjected to a fat-rich diet and limited physical activity. Finally, we assessed overall rates of metabolic evolution in different tissues and identified a number of significant changes in metabolic and lipid networks on the primate and on the hominoid evolutionary lineages.

P-2135

Medaka (*Oryzias latipes*) as a population model for studying functional differences between the alleles of human polymorphisms

Takafumi Katsumura[1,2], Shoji Oda[1], Hiroshi Mitani[1], Shoji Kawamura[1], Hiroki Oota[1,2]
[1]*University of Tokyo, Kashiwa, Japan,* [2]*Kitasato University School of Medicine, Sagamihara, Japan*

Genome-wide polymorphism data have been rapidly growing for humans in recent years. However, there have been potential difficulties to assess how such polymorphisms cause any functional differences. Here we would propose a new model animal system for human population genetics/genomics using medaka fish (*Oryzias latipes*) population as analogy of human population, especially for conducting functional assay of their alleles of polymorphic sites. The reasons why madaka is the best for the system are (1) well-studied diversity in wild and lab-stock populations, (2) complete genome sequence data available, and (3) experimental feasibility enabling transgenic technologies.

In the present study, we focused on the alleles in the *cytochrome P450* (*CYP*) gene family as a test case because its biochemical feature has been well characterized, and each gene in humans has highly polymorphic involving functional differences between the alleles. First, for an efficient screening method for genetic variation, we applied population genetic approaches for medaka populations. We tested deme- and grid-based sampling methods for the mitochondrial genome (mtDNA) D-loop region to examine neutral variations. The statistical values obtained from population genetic tests showed that medaka populations were highly diverged among local habitats, and that between-population diversity based on the grid sampling was larger than within-population diversity based on the deme sampling, suggesting the grid-based sampling method will more efficiently give us polymorphic data than the deme-based sampling method. To compare medaka and human CYP polymorphisms, we determined 20 *CYP* genes to be with apparent orthology between medaka and human. Out of 20 *CYP* genes, we chose four *CYP* genes (*CYP1A*, *CYP1B1*, *CYP5A1*, and *CYP20A1*) harboring non-synonymous single nucleotide polymorphisms (SNPs) with different allele frequencies among human populations, and conducted SNPs screening for the medaka *CYP* genes using the grid-based samples. We found allelic variations in each *CYP* gene among medaka populations. We compared the medaka CYP1B1 enzyme activity between the alleles using Luciferin-CEE, human CYP1B1 metabolizing substrate, in the *Drosophila* S2 cells, and found significant differences in the relative activities of CYP1B1 between the alleles as found in human CYP1B1. Four alleles that Niigata, Maegok, (*O.latipes*), Luzon (*O.luzonensis*) and Hainan (*O.curvinotus*) harbor showed significantly lower enzyme activities than the highest frequency allele in medaka populations (30.4%) that Tanabe does ($p<0.05$).

Thus, the results showed the usefulness of medaka as a population model animal for estimating functional differences between alleles of human polymorphisms.

P-2136

Amino acid substitution regime is dependent upon the protein interaction mode with the GroEL chaperone

Judith Ilhan, Giddy Landan, Tal Dagan
Institute for Molecular Evolution, Duesseldorf, Germany

Many proteins require the assistance of molecular chaperones in order to fold efficiently. Chaperones have been shown to buffer the effects of slightly deleterious mutations, probably because they can compensate for the deficiency in spontaneous folding. One of the best- studied chaperones is the eubacterial GroEL/GroES system. In *Escherichia coli*, three classes of proteins have been distinguished based on their degree of dependency on GroEL for folding: I) casual interactors do not require GroEL to gain functionality, II) partially-dependent substrates require GroEL in a temperature-dependent manner, and III) obligate interactors require GroEL in order to fold into a functional conformation. Proteins in the three classes have been recently shown to evolve in significantly different substitution rates. Whether they differ also in their amino acid substitution regimes is still unknown. Here we compare amino acid substitution matrices among the three classes in increasing phylogenetic depth. The frequency of amino acid substitutions was obtained from multiple sequence alignments as well as reconstructed ancestral character states. To detect significant differences in the substitution matrices, we conducted permutation tests where class membership was assigned randomly. These permuted datasets provided empirical distributions of differences that were then used to assign significance levels to the observed differences between the classes. We found that the substitution matrices of proteins from the three classes are significantly different. Analyzing the structure of the substitution matrices revealed that the three classes significantly differ in their amino acids composition, but that different amino acids do not differ in their propensity to undergo substitutions. However, the conditional probability of specific substitutions was found to be significantly different between the three classes. These observations remain unchanged when the substation matrices are derived from alignments of sequences in four levels of phylogenetic depth. The significant differences between substitution matrices in the three GroEL-dependency classes may be attributed either to the degree of buffering by that chaperone, or to the evolution of protein interaction with the GroEL. Our results suggest that GroEL/GroES chaperones are a biological mechanism that generates variable amino acids substitution regimes.

P-2137

Topological structure of the space of phenotypes: The case of RNA neutral networks

Jacobo Aguirre[1], Javier M. Buldú[2], Michael Stich[1], Susanna C. Manrubia[1]
[1]Centro de Astrobiología (INTA-CSIC), Torrejón de Ardoz, Madrid, Spain,[2]Complex Systems Group, Universidad Rey Juan Carlos, Fuenlabrada, Madrid, Spain

The evolution and adaptation of molecular populations is constrained by the diversity accessible through mutational processes. RNA is a paradigmatic example of biopolymer where genotype (sequence) and phenotype (approximated by the secondary structure fold) are identified in a single molecule. The extreme redundancy of the genotype-phenotype map leads to large ensembles of RNA sequences that fold into the same secondary structure and can be connected through single-point mutations. These ensembles define neutral networks of phenotypes in sequence space. Here we analyze the topological properties of neutral networks formed by 12-nucleotides RNA sequences, obtained through the exhaustive folding of sequence space. That genome space fragments into 645 subnetworks that correspond to 57 different secondary structures. The topological analysis reveals that each subnetwork is far from being random: it has a degree distribution with a well-defined average and a small dispersion, a high clustering coefficient, and an average shortest path between nodes close to its minimum possible value, i.e. the Hamming distance between sequences. RNA neutral networks are assortative due to the correlation in the composition of neighboring sequences, a feature that together with the symmetries inherent to the folding process explains the high modularity observed. Several topological relationships can be analytically derived attending to structural restrictions and generic properties of the folding process. The average degree of these phenotypic networks grows logarithmically with their size, such that abundant phenotypes have the additional advantage of being more robust to mutations. This property prevents fragmentation of neutral networks and thus enhances the navigability of sequence space. In summary, RNA neutral networks show unique topological properties, unknown to other networks previously described.

P-2138

SIGNIFICANT CLUSTERING ON GEOMETRIC MORPHOMETRICS DATA: DEFINING EVOLUTIONARY MODULES IN COMPLEX BIOLOGICAL DATASETS

Jose Sergio Hleap[1,2], Christian Blouin[1]
[1]*Dalhousie University, Halifax,NS, Canada,* [2]*Fundación SQUALUS, Cali, Valle, Colombia*

There is a need to define expressive representations of protein structures to enable new kinds of analysis of evolution in 3D. In this work, we are interested in the evolutionary module. A module is a subset of the protein's shape that is self consistent across homologs. The challenge is to define this concept in a way where significance can be evaluated. We present here an approach inspired from the analysis of shapes in animal datasets, where landmarks are points in space defined by the center of mass of side chains. We demonstrate that classic methods for inferring modularity are inadequate for optimization when there are many landmarks. A method of inference is described here based on a graph theoretic approach drawn from the literature on community detection. The significance testing of the resulting cluster is presented and validated on simulated data. Further testing on animal shapes data is presented and preliminary work on protein structure data. The protein structure data comes from curated alignments from the HOMSTRAD database. The analysis of the *Lycalopex* dataset, showed sensical results for the fox skull, giving support to the zygomatic arch development in the fox ontogeny. At the protein level, our results suggest that there exist significant modularity at finer and/or coarser level than domains. We are currently testing the hypothesis that modularity in protein structures is defining groups of sites that following a complementary mutation regimen.

P-2139

Estimating the rate of irreversibility in protein evolution

Onuralp Soylemez[1], Fyodor Kondrashov[1,2]
[1]Center for Genomic Regulation, Barcelona, Spain, [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

In the course of evolution novel phenotypes and genotypes are created. However, to what degree the acquired phenotypes in an evolving lineage can revert back to the forms realized in its direct ancestors remains an open and debated question. Dollo has formulated what is now known as the Dollo's law stating that "an organism is unable to return, even partially, to a previous stage already realized in the ranks of its ancestors." (Dollo, 1893) Dollo's statement refers to irreversibility on the level of phenotypes. On the genotype level Dollo's law can be adapted to describe a situation when a substantial fraction of substitutions that revert the genotype state to the ancestral state are under strong negative selection. Many nucleotide changes, both synonymous and non-synonymous, are expected to be entirely neutral because they never lead to any form of phenotypic change in the organism. There must also be a number of genotypic changes that have profound effects on a phenotypic level, and the issue of whether or not such changes may be reversed in the course of further evolution is the issue at hand when considering Dollo's Law on the genotypic level. There are a handful of examples where the irreversibility of specific amino acid changes in proteins has been tested empirically. Nevertheless, the overall rates of irreversibility of evolutionary change on the molecular level and the underlying genetic mechanisms of genotype irreversibility have not been comprehensively addressed. Here, we test the rate of irreversibility of phenotypically relevant amino acid changes by comparing data on human genetic disease mutations with sequence divergence data between human genes and orthologs from placental mammalian species. We find that 4% of all amino acid substitutions are irreversible, such that a return to the ancestral amino acid state would lead to an unequivocally deleterious phenotype. Clearly, the pathogenic phenotypes were not ancestral. Therefore, the 4% rate of evolutionary irreversibility is the contribution of compensatory evolution. Thus, the nature of Dollo's law on the molecular level is perhaps best explained through compensatory changes.

P-2140

Molecular evolution of immune genes in socially diverse bees

Lumi Viljakainen[1], Brielle J. Fischman[2], S. Hollis Woodard[2], Gene E. Robinson[2], Andrew G. Clark[3]
[1]University of Oulu, Oulu, Finland,[2]University of Illinois at Urbana-Champaign, Urbana, IL, USA,[3]Cornell University, Ithaca, NY, USA

The evolution of sociality in ants, bees, wasps and termites has been substantially influenced by the elevated pathogen pressure in densely inhabited colonies. Social insects have evolved many collective defenses, i.e. social immunity, to complement the physiological individual-level immune responses. These include the use of antimicrobial compounds in the nest building material and collective recognition and removal of infected individuals. The effect pathogens have had on the evolution of genes involved in physiological immune system genes in social insects can be studied by comparing evolutionary patterns in immune genes of closely related social and solitary organisms. We have done this by studying sequence evolution of a large number of immune genes in bees with varying levels of sociality (highly eusocial, primitively eusocial and non-eusocial) using codon-based likelihood models of nucleotide substitution. There is surprisingly little evidence for positive selection in the contrasts done to date. This suggests that social immunity may act as a buffering mechanism and reduce selection pressure on the immune genes.

P-2141

Mapping the evolutionary fitness landscape of C4 photosynthesis

David Heckmann, Martin J. Lercher
Institute for Computer Science, HHU Düsseldorf, Düsseldorf, North Rhine-Westphalia, Germany

C4 photosynthesis is an add-on to the ancestral C3 photosynthesis, enabling plants to avoid the deleterious fixation of oxygen by RuBisCO (Ribulose-1,5-bisphosphate carboxylase oxygenase). The evolution of C4 photosynthesis from a C3 ancestor required a considerable amount of changes in expression of metabolic genes, leaf development and biochemical properties of participating enzymes. Despite these barriers, C4 photosynthesis evolved independently in over 60 lineages. This high level of convergence suggests a low evolutionary trough towards expression of the trait, which indicates that each intermediate step during the evolution was likely adaptive or at least neutral. Studies on plant genera that include C3, C4, and C3-C4 intermediate species, like Flaveria, are a common way of building hypotheses on C4 evolution.

We utilize in silico metabolic modeling to estimate the fitness landscape lying between plants using C3 and C4 photosynthesis. This fitness landscape depends on multiple parameters and is thus of high dimensionality. We calculate a set of evolutionary paths using a Monte Carlo approach based on the mutation and fixation probabilities of the intermediate steps. This set of predicted paths that lead to C4 photosynthesis allows us to build hypotheses on the most probable sequence realised in evolutionary transitions. We present our findings in comparison to experimental data from C3-C4 intermediate genera like Flaveria.

P-2142

Impact of selection on genes involved in regulatory network: a modelling study.

Frederic Austerlitz[1,2], Bénédicte Rhoné[2], Jean-Tristan Brandenburg[1,2]
[1]*Laboratoire Eco-Anthropologie et Ethnobiologie, UMR 7206 CNRS/MNHN/Université Paris 7, Paris, France,*
[2]*Laboratoire Ecologie, Systématique et Evolution, UMR 8079 CNRS/Université Paris-Sud/AgroParisTech, Orsay, France*

Complex phenotypes are often controlled by many interacting genes. One question emerging from such organisation is how selection, acting at the phenotypic level, shapes the evolution of genes involved in regulatory networks controlling the phenotypes. We studied this issue through a matrix model of such networks. In a population submitted to selection, we simulated the evolution of a quantitative trait controlled by a set of loci that regulate each other through positive or negative interactions. Investigating several levels of selection intensity on the trait, we studied the evolution of regulation intensity between the genes and the evolution of the genetic diversity of those genes as an indirect measure of the strength of selection acting on them. We show that an increasing intensity of selection on the phenotype leads to an increased level of regulation between the loci. Moreover, we found that the genes responding more strongly to selection within the network were those evolving toward stronger regulatory action on the other genes and/or those that are the less regulated by the other genes. This observation is strongest for an intermediate level of selection. This may explain why several experimental studies have shown evidence of selection on regulatory genes inside gene networks.

P-2143

Studying the indirect impact of intrachromosomalrearrangements on the genome structure of microbes

Stephan Fischer[1,3], Carole Knibbe[1,4], Samuel Bernard[2,4], Guillaume Beslon[1,3]
[1]Inria Beagle Team, Lyon, France,[2]Inria Dracula Team, Lyon, France,[3]INSA-Lyon, Lyon, France,[4]Université Lyon I, Lyon, France

Experimental evolution has shown that rearrangements within the chromosomes of bacteria occur at a very frequent rate. Indeed, several experiments show that fitness increases in a step-wise manner and that the first and most important steps of evolution are linked with chromosomal rearrangements, mostly deletions or DNA amplification. These deletions are observed when experiments are repeated with independent populations, indicating that rearrangements must occur frequently spontaneously but also that they are essential for adaptation [Cooper et al., Journal of Bacteriology, 2001 ; Kugelberg et al., PNAS, 2006]. Obviously, such a high rate of rearrangement has a direct impact on the genome structure - e.g. gene order. Yet, it is also likely to strongly impact genome robustness and evolvability, leading to indirect effects that may be difficult to predict.

To quantify these effects, we developed computational and mathematical models to understand the impact of rearrangements on the evolution of genome structure. These models included local mutations but also inversions, translocations, duplications and large deletions, as well as competition for reproduction. Taking into account all these factors allowed the genome to evolve in length and in coding ratio, enabling a wide diversity at a structural level.

Both simulations and theoretical analysis show that rearrangements not only play a major role in evolvability (by allowing for duplication/divergence or deletion of fragments of chromosomes), but also have an indirect impact on the genome structure. Indeed, we have shown with a mathematical model that duplication and deletion rates impose a maximal genome size, even though there is no direct selective cost on the genome size or non-coding sequences. Simulations confirm that genomes converge towards a specific finite size and a coding ratio that are strongly linked with rearrangement rates.

As a result, we infer a trade-off comparable to the 'error-threshold' that has been proposed for local mutation rates [Eigen, Naturwissenschaften, 1971]. Rearrangements are crucial for evolvability but, on the other hand, they limit the quantity of sequences that can be maintained through evolution. However, this 'rearrangement-threshold' effect is likely to be stronger than the 'error-threshold' effect because the maximal genome sizes it imposes is more restrictive for comparable rearrangement and local mutation rates.

P-2144

How evolvable are polarization machines?

Liedewij Laan, Andrew Murray
Fas Institute for Systems Biology, Harvard University, Cambridge, MA, USA

In many different cell types, ranging from yeast to human epithelial cells, proper polarization is essential for cell function. The polarization mechanisms however, differ in different cell types and even closely related species use a variety of polarization machines. How much do different mechanisms really differ between species and how many mutations are necessary to switch from one mechanism to another? Our approach to study the evolvability of polarization machineries is to destroy polarization in budding yeast and to subsequently try to restore polarization by evolution experiments.

We "destroyed" polarization by deleting one or two genes that are essential for its function. After the destruction of polarization the growth rate of cells was highly decreased (2-5 fold) and their cell shape and size was highly variable and perturbed, indicating severe problems in the establishement of polarity. Subsequently, we evolved these cells by serial dilutions for 10 days. Surprisingly, the evolved cells rapidly overcame most of their polarity defects. The evolved cells grew at ~ 90% wildtype growth rate and their cell shape and size were signifigantly less perturbed. The next step will be to shed light on how these cells rescued polarization.

P-2145

Functional implications of a polymorphism in a primate-specific microRNA

Maria Lopez-Valenzuela[1], Oscar Ramírez[1], Ignasi Torruella-Loran[1], Antonio Rosas[2], Samuel García-Vargas[2], Marco de la Rasilla[3], Carles Lalueza-Fox[1], Yolanda Espinosa-Parrilla[1]
[1]Institut de Biologia Evolutiva (UPF-CSIC), Barcelona, Catalunya, Spain,[2]Paleoanthropology Group, Department of Paleobiology, Museo Nacional de Ciencias Naturales (CSIC), Madrid, Madrid, Spain,[3]Área de Prehistoria, Departamento de Historia, Universidad de Oviedo, Oviedo, Asturias, Spain

MicroRNAs are small non-coding RNAs that act as postranscriptional regulators. MicroRNA mediated regulation heavily depends on perfect complementarity between the 7nt seed region on the microRNA and the target site of the messenger RNA to be repressed. Hence, the smallest change in the seed region can be deleterious as it may affect target recognition. Given that a single microRNA can potentially target hundreds of genes, a change in the seed may lead to deregulation of entire biological pathways. According to this, purifying selection has strongly acted on human microRNAs and particularly on their seed regions that exhibit a significantly lower SNP density than the genome average, particularly on their seed regions where SNPs are rare. Sequencing of the Neanderthal genome prompted us to mir-1304, a microRNA that differs from the human reference genome in a nucleotide change on the seed region. Further analysis revealed that most humans carry the derived form of this microRNA (der-mir-1304) while non-human primates and Neanderthal shared the ancestral allele (anc-mir-1304). Interestingly, anc-mir-1304 is still present at low frequency in modern Asian populations (~5%) and is very rare in Africans. Mir-1304 is a novel primate specific microRNA whose function and targets are still unknown. To gain insight into the implications of a differential gene regulation by the two alleles, we predicted their targets using TargetRank and TargetScan algorithms and found a tenfold difference in the number of target genes of the two alleles (36 for anc-mir-1304 versus 515 for der-mir-1304) suggesting an important functional evolution for this microRNA. Among the few target genes predicted for anc-mir-1304 there are two, enamelin and amelotin, involved in teeth development. By means of luciferase assays we proved that the anc-mir-1304 indeed exerts a strong down-regulation of both genes while the derived version has no effect on their expression. As for der-mir-1304, analysis of the putative targets indicates an association with behavior and nervous system development and function, this association is further supported by the finding of mir-1304 expression in human brain. Involvement of miR-1304 in brain function regulation raises the possibility of particular neurological phenotypes being associated with one mir-1304 allele or the other. Future functional analysis on the differential gene repression by the two mir-1304 alleles would be of great interest to gain insights into primate evolution and in the possible implication of mir-1304 in phenotypic differences among human populations and its relationship with complex disorders.

P-2146

Robustness to stress in antibiotic resistant *Escherichia coli*

Ana Sousa, Sandra Trindade, Isabel Gordo
Instituto Gulbenkian de Ciencia, Oeiras, Portugal

The role of mutation in evolution depends upon the distribution of their effects on fitness. This distribution is likely to depend on the environment. Indeed genotype-by- environment interactions are key for the process of local adaptation and ecological specialization. An important trait in bacterial evolution is antibiotic resistance, which presents a clear case of change in the direction of selection in environments with and without antibiotics. Here we study the distribution of fitness effects of mutations, which confer antibiotic resistance in Escherichia coli, in benign and stressful environments in the absence of drugs. We interpret the distributions in the light of a fitness landscape model that assume a single fitness peak. We find that mutation effects (s) are well described by a shifted gamma distribution, with a shift parameter that reflects the distance to the fitness peak and that varies across environments. Consistent with the theoretical predictions of Fisher's geometrical model, with a Gaussian relationship between phenotype and fitness, we find that the main effect of stress is to increase the variance in s. Our findings are in agreement with the results of a recent meta-analysis, which suggest that a simple fitness landscape model can capture the variation of mutation effects across species and environments.

P-2147

Melanesian Blond Hair is Caused by an Amino Acid Change in TYRP1

Eimear Kenny[1], Nicholas Timpson[2], Martin Sikora[1], Muh-Ching Yee[1], Andres Moreno Estrada[1], Celeste Eng[3], Scott Huntsmann[3], Esteban Gonzalez Burchard[3], Mark Stoneking[4], Carlos Bustamante[1], Sean Myles[1]
[1]Stanford University, Stanford, CA, USA, [2]University of Bristol, Bristol, UK, [3]University of California San Francisco, San Francisco, CA, USA, [4]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Human pigmentation varies considerably within and among populations and is a function of both variation in exposure to ultraviolet radiation (UVR) and melanin production. We examined the genetic basis of blond hair in the Solomon Islands, a population that breaks from the trend of darker hair and skin pigmentation near the equator where there is higher UVR. Strikingly, while individuals from the Solomon Islands and Oceania have the darkest skin pigmentation outside of Africa, they also have the highest prevalence of blond hair (5-10%) outside of Europe.

We performed a genome-wide association study for hair color and observed a single strong association signal (top SNP rs13289810; P=1.11x10^{-19}) on chromosome 9p23. The mapping interval contained one known gene, tyrosinase-related protein 1 (*TYRP1*), which encodes a melanosomal enzyme involved in mammalian pigmentation, and for which null alleles are known to cause rufous albinism in humans. Resequencing of *TYRP1* exons detected a single novel polymorphism that corresponds to a predicted arginine-to-cysteine mutation in exon 2 (R93C). We genotyped R93C in 918 Solomon Islanders for whom we had measured hair pigmentation and found that a recessive model provided the best fit for the association (P = 2.19x10^{-90}), explaining 46.4% of overall variance in hair color. The frequency of the 93C allele in the Solomon Islands is 0.26, and genotyping of the R93C SNP in an additional 941 individuals from 52 worldwide populations revealed that the allele is absent outside of Oceania. Furthermore, we found no evidence for gene flow from Europe nor convincing signatures of recent positive selection at the 9p23 locus.

R93C shows similarity with the functional allele in brown^{light} mice, which also exhibit lightened hair. The molecular alteration in both cases is an R->C substitution in the 15-Cys epidermal growth factor domain of *TYRP1,* which results in reduced *TYRP1* stability and catalytic function and decreased melanin contain in mouse hair. It is likely that the human 93C mutation operates via a similar mechanism.

This study realizes some of the postulated benefits for mapping complex human traits in isolate populations; namely, how genetic drift (and/or local positive selection) may allow otherwise rare alleles that influence such traits to rise to an appreciable frequency. More generally, this work strongly supports the growing notion that population-specific variation is important in accounting for the heritable variance in physiological phenotypes, and underscores the importance of enabling medical genomics in diverse worldwide populations.

P-2148

Phenotypic plasticity of body size in Daphnia pulex

France Dufresne
Université du Québec à Rimouski, Rimouski, Canada

Body size is an extremely plastic trait that has been widely studied since it is correlated with many physiological characteristics and to fitness. Genetic analyses of model species have revealed that the insulin and rapamycin pathways play a major role in body size determination. The microcrustacean Daphnia is an ideal model species to examine genes involved in body size determination since its whole genome has been sequenced. This species has been a classical model for studies of phenotypic plasticity since it produces impressive defensive structures in response to predators and body size changes when exposed to a variety of environmental factors. Annotation of the insulin signaling pathway (ISP) in *Daphnia pulex* has revealed the presence of four insulin-like receptors (InR) as opposed to a single one in the majority of invertebrates. This study aimed to examine expression of the four InR in *Daphnia* clones exposed to factors known to influence body size. *D. pulex* clones were grown under : 1) low and high temperatures, 2) *Chaoborus* (invertebrate predators) kairomones. QPCR analyses of the four InR along with three reference genes were conducted in a LC480 from Roche. Somatic growth rates in the three treatments were also measured and compared with gene expression data. These results will provide a better understanding of the importance of the insulin signalling pathway for the modulation of body size changes in response to short term changes in the environment.

P-2149

Genotyping Candidate Loci for Colorectal Cancer: A Pointer to Ancestral Susceptibility.

Stefanie Huhn[1], Melanie Bevier[1], Barbara Pardini[2], Alessio Naccarati[2], Ludmila Vodickova[2,3], Jan Novotny[4], Pavel Vodicka[2,3], Kari Hemminki[1,5], Asta Försti[1,5]
[1]Department of Molecular Genetic Epidemiology, German Cancer Research Center Heidelberg, Heidelberg, Germany, [2]Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Prague 4, Czech Republic, [3]Institute of Biology and Medical Genetics, 1st Faculty of Medicine, Charles University, Prague 2, Czech Republic, [4]Department of Oncology, General Teaching Hospital, U Nemocnice 2, Prague 2, Czech Republic, [5]Center of Primary Health Care Research at the Clinical Research Center, Lund University, Malmö, Sweden

Colorectal cancer (CRC) is a complex disease, caused by both genetic and non-genetic risk factors. It is one of the most common cancers worldwide. However, the incidence rates vary depending on age, gender and country, with the highest incidence rates in the modern, industrialized societies. The differences are mainly attributed to differences in diet and other environmental factors that are also known to differ significantly among worldwide populations. To gain insights into the mechanisms behind the global differences, we asked whether genetic variants contributing to CRC susceptibility show signatures of selection and local adaptations in industrialized countries as such patterns are already known for risk traits related to diet and pathogen load. Under this premise, genetic variants may possess ancestral susceptibility to CRC with protective derived adaptations.

In 2006 and 2007 two studies described the genomic landscape of colorectal cancers1,2. We now asked whether common variants in the genes know to be somatically mutated in CRC contribute to CRC susceptibility and if so, whether these variants show ancestral susceptibility. We selected the 10 genes that were reported to be mutated in more than 10% of the tumour samples, excluding APC, K-Ras, TP53 and ABCA1 that are already well described in CRC. All 10 candidate genes were analyzed for single nucleotide polymorphisms (SNPs) that may possess functional power (e.g. missense SNPs). Additionally, genes and SNPs were analyzed for signatures of selection (iHS, Fay Wu's H, global allele frequency pattern) and linkage disequilibrium. We selected 36 SNPs that were functional and/or showed signatures of selection for an association study on a Czech population containing 1412 hospital-based CRC cases, 787 colonoscopy negative controls and 853 healthy blood donors.
Preliminary results indicated four SNPs to be associated with CRC risk, with two of them showing ancestral susceptibility to CRC and protective derived adaptations, respectively.

An analysis of associations of ancestral alleles instead of minor alleles of SNPs on CRC risk provides insights into development of common, risk modifying variants and into the interplay of genes and environment. Furthermore, this approach provides an opportunity to discuss and compare populations with different evolutionary backgrounds.

Reference:
1 Sjoblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. Science 314, 268-74 (2006).
2 Wood, L.D. et al. The genomic landscapes of human breast and colorectal cancers. Science 318, 1108-13 (2007).

P-2150

Evolutionary origin and process, and genetic diversity of goldfish

Tomoyoshi Komiyama, Atsushi Ogura, Kazuho Ikeo, Takashi Gojobori
National Institute of Genetics, Mishima, Shizuoka, Japan

With the aim of evolutionary origin and process, and genetic diversity of goldfish, we studied mitochondrial DNA sequences as well as nuclear DNA sequences for a variety of goldfish. Goldfish have been domesticated through strong artificial selection for at least hundreds of years, and probably more than a thousand years, because historical records of breeding of goldfish have been found in China from almost 1,500 years ago. Artificial selection has been conducted for a wide variety of traits of goldfish such as Celestial and Telescopic eyes, fancy but uncontrollable shapes of tail fins, an unfittingly fat body, and a multitude of other qualities. In Japan in particular, there are many kinds of goldfish that exhibit many unique phenotypes such as body color, and fin and eye shapes. Japanese goldfish have been derived from traditional breeding stemming from the Japanese aesthetic appreciation for goldfish particularly with unique characteristics. We have sequenced mitochondrial DNAs of Japanese goldfish, and have traced back their breeding history with help of historical documents. In a phylogenetic tree obtained, we found that three traits of typical characteristics of Japanese goldfish, dorsal finless and Celestrial and Telescope eyes, must have emerged independently of each other at different times. It suggests that goldfish were not imposed by a systematically well-designed way of strong artificial selection, but only to meet diversified needs of human preferences in a rather unsystematic way. To examine if this is the case for nuclear DNAs, we are now in process of sequencing nuclear DNAs of Japanese goldfish. In the meeting, we make the report of all the data we obtained, in order to elucidate the evolutionary origin and process, and genetic diversity of goldfish.

P-2151

Can the evolutionary dynamics of structural disorder contribute to biological divergence and neostructuralization?

Madolyn MacDonald[2], Juan Felipe Ortiz Quinonez[1], Patrick Masterson[1], Jessica Siltberg-Liberles[1]
[1]University of Wyoming, Laramie, WY, USA,[2]Rochester Institute of Technology, Rochester, NY, USA

Proteins with disordered structure are found as interconverting protein conformations over a flattened energy landscape, as compared to the more traditional energy landscape with the one or two defined wells that we commonly see for globular structured proteins. For globular proteins the stability of its fold after a mutation is a major determinant for whether a mutation is fixed or lost, and the structures of homologous globular proteins are generally assumed to be conserved. Structurally disordered proteins likely evolve following a different scenario. These proteins are already less stable and if the stability for a mutation is considered for all conformations, it may be stabilizing for some, destabilizing for some, and neutral for others. Further, as the structurally disordered proteins often are functionally promiscuous, there may also be a functional trade off as to which functions benefit from the altered conformational ensemble following a certain mutation or not. We call this mutation driven conformational selection, where mutations alter the equilibrium of the conformational ensemble. Subtle changes in the conformational ensemble could provide a route to biological divergence and innovation.

In order to improve our understanding of the evolution of the structurally disordered regions in proteins further, a comparative phylogenetic (un)structural genomics screen was performed in order to investigate the evolutionary dynamics of structural disorder in the flaviviruses. Flaviviruses, e.g. Dengue virus, are small single strand RNA viruses. The four different strains of Dengue virus can cause Dengue fever, a disease that infects 50 to 100 million people per year. Using antibody dependent enhancement, the second infection with a different Dengue virus strain is often more severe than the first infection, due to proteins with high conformational flexibility. Our results indicate high evolutionary dynamics in some structurally disordered regions and a great variety in the extent of structural disorder among homologous positions in proteins from different flaviviruses.

P-2152

Genome-wide genetic architecture of morphological traits in yeast

Wei-Chin Ho, Jianzhi Zhang
University of Michigan, Ann Arbor, MI, USA

Understanding the genetic basis of phenotypic variation, or the genotype-phenotype map, is a major goal of evolutionary biology. We analyzed a large dataset of the budding yeast *Saccharomyces cerevisiae* in which 220 morphological traits were measured in multiple cells of each of the wild-type strain and 4718 single-gene-deletion strains. We found that many traits are affected by hundreds to thousands of genes, with the mean and median numbers of genes affecting a trait being 1720 and 1746, respectively. The effect sizes of all genes on a trait approximately follow a normal distribution, with different dispersions of the distribution for different traits. Effects tend to be small on traits that have small phenotypic variations among wild-type cells, indicating a positive correlation between environmental canalization and genetic canalization. The extent of this correlation varies among genes. Genes with weak correlations have large effects on all traits, while genes with strong correlations have lowered effects on traits with low phenotypic variations, suggesting that a subset of genes have been canalized in terms of their effects on traits with small phenotypic variations, which are typically important for fitness. We discuss the implications of these findings on phenotypic evolution.

P-2153

GENETIC ARCHITECTURE AND EVOLUTION OF HUMAN SKIN AND EYE PIGMENTATION: GENOME-WIDE ASSOCIATION STUDY IN THE ADMIXED POPULATION OF CAPE VERDE

Sandra Beleza[1,2], Nick Jonhson[1], Sophie Candille[1], Isabel Inês Araújo[3], António Correia e Silva[3], Mark Shriver[4], Jorge Rocha[5,6], Greg Barsh[1,7], Hua Tang[1]
[1]Genetics Department, Stanford University, Stanford, CA, USA, [2]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal, [3]Universidade de Cabo Verde (UNI-CV), Praia, Santiago Island, Cape Verde, [4]Department of Anthropology, Penn State University, University Park, PA, USA, [5]Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO), Vairão, Portugal, [6]Departamento de Zoologia e Antropologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal, [7]HudsonAlpha Institute for Biotecnhology, Huntsville, AL, USA

Pigmentary variation is one of the most striking aspects of human morphologic diversity. Quantitative genetic studies of human pigmentation have demonstrated a high heritability for both skin and eye colour. More recently, efforts to study human pigmentation have focused on discovering the genes affecting these traits. But our knowledge of skin and eye colour genetics is incomplete, based primarily on candidate gene studies in African Americans, or on the limited range of phenotypic diversity found in Europeans.

Admixed populations offer unique opportunities for understanding the genetic basis of traits that exhibit phenotypic difference between populations, such as pigmentary traits. We report a genome-wide association study of skin and eye colours in the European-African admixed Cape Verdean population, in which extensive phenotypic variation is observed. Genotype-based and genetic ancestry-based association scans were consistent in identifying four major loci for skin colour and two major loci for eye colour, which act in an additive or dominant manner, and together with ancestry explain about half of the overall estimated phenotypic variation. Both coding and regulatory variants are involved in the genetic determination of skin and eye colour. Our results unravel a more complex nature for skin colour than previously thought, indicate that there is a common genetic basis for skin and eye colour and provide some understanding about the evolution of these traits in human populations.

P-2154

Difference in Gene expression between dry and wet thalli of lichen *Usnea bismolliuscula*

Mieko Kono[1], Yoshihito Ohmura[2], Yoko Satta[1]
[1]*Department of Evolutionary Studies of Biosystems, the Graduate University for Advanced Studies [SOKENDAI], Hayama, Japan,* [2]*Department of Botany, National Museum of Nature and Science, Tsukuba, Japan*

Lichens are symbiotic organisms between fungi and algae and/or cyanobacteria found in almost all terrestrial habitats. Despite symbiosis is drawing increasing attention due to its importance in the ecosystem and evolution, the molecular basis of lichen symbiosis remains unknown. Physiological studies suggested that lichens subjected to environment with dry and wet cycle result in better growth. Thus alternating drying and wetting is a prerequisite for functioning symbiosis in lichens. In this study, in order to find genes related to symbiosis we identified genes differentially expressed between dry and wet thalli of lichen *Usnea bismolliuscula*. For the dry state, thalli kept in a natural dry condition were used, and thalli rehydrated by distilled water were used for the wet state. The thalli in the both states were exposed to white fluorescent lamp (4.3µmol $m^{-2}s^{-1}$) for 1 hour. Total RNAs were extracted from the thalli of each state and converted into cDNA. The nucleotide sequences of the cDNAs were determined by next generation sequencer. cDNA sequences present in both states were excluded as genes commonly expressed. 2,568 and 1,630 sequences were left in the dry and the wet state respectively and were considered as candidates for differentially expressed genes. BLASTX searches of the candidate genes were then carried out. The result showed genes related to fungal mannitol synthesis and algal photosynthesis. This is consistent with physiological studies indicating effects of drying and wetting cycle on polyol metabolisms and photosynthetic activities in lichens. For genes related to photosynthesis, the quantitative analysis was performed by RT-PCR. This analysis revealed that although the expression of photosynthetic genes changes according to the water content of thalli, the difference between the two states is not as high as expected from physiological observation. This may due to non mRNA regulation of metabolism in dry and wet thalli of lichen or the condition of our experiment. To elucidate this problem, we will perform experiments with different light and water conditions.

P-2155

Revaluating duplication gene evolution from the perspective of their phylogenetic origin

Krishna Swamy[1,2], Chien-Hao Su[2], Daryi Wang[3], Huai-Kuang Tsai[1,2]
[1]Bioinformatics Program, Taiwan International Graduate Program,Academia Sinica, Taipei, Taiwan,[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan,[3]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan

Background
The chromosomal localization of duplicate genes within genomes has attracted considerable attention due to their importance in function. It has been found that selection can constraint groups of physically linked duplicates, which are functionally related by ordered structuring of eukaryotic genomes. Studies have also examined the contribution of sequence structure, evolution of their transcription start sites among other factors to divergence of gene expression. However, the understanding on duplicated genes added at different evolutionary times, which could have different evolutionary history is still limited. In this study, we have tried to envisage the constraints on physical organization of duplicated genes in Drosophila melanogaster, while considering their emergence time i.e., at the same and different stages of Drosophila evolution.
Results
Using a genomic phylostatigraphic approach, we studied the evolutionary emergence of duplicated genes at 12 phylogenetic levels, spanning the full spectrum of Drosophila evolution. The duplicate genes were categorized according to duplication events at the same phylogenetic level and at different phylogenetic levels (as the emergence of genes), which includes reduplication events. We looked for association between the categorized duplicate pairs with previously emphasized features such as gene order, coding sequence heterogeneity, functional similarity and regulatory evolution respectively. Our results indicate that, for genes from the two categories, there is a significant difference in preference for each of these features. Duplicate genes that arose due to duplication events occurring at the same phylogenetic level tend to have a significantly higher tendency to be organized in tandem and lower coding sequence heterogeneity. In addition, there was also a difference in their preference for regulators. Duplicate genes that have emerged at the same phylogenetic level are bound by more number of regulators and also have higher fraction of shared regulatory binding motifs. Furthermore, such duplicated genes are also constrained to have similar functional traits than genes due to duplication events at different phylogenetic levels.
Conclusions
These results reveal that evolutionary history of duplicated genes can play a pivotal role in the evolution of their sequence structure, their function and their localization across the genome. The organization of regulatory elements in their promoters and preference of regulators can also differ considerably with their time of emergence.

P-2156

Investigating Two Different Mechanisms of Multiple-Host Recognition in the Bacterial Transferrin Receptor

Anastassia Pogoutse[1], Charles Calmettes[1], Rong-hua Yu[2], Estela Costa[2], Anthony Schryvers[2], Trevor Moraes[1]
[1]University of Toronto, Toronto, Ontario, Canada, [2]University of Calgary, Calgary, Alberta, Canada

Pathogenic bacteria have developed various iron acquisition systems in order to survive in the iron limiting environments within their hosts. One such receptor system, found in the outer membranes of bacteria from the *Neisseriaceae* and *Pasteurellaceae* families that cause diseases such as meningitis and gonorrhea, directly binds and removes iron from the iron-carrying serum glycoprotein, transferrin. This bipartite bacterial receptor consists of a TonB-dependent gated pore, called transferrin binding protein A (TbpA), and a lipid-anchored protein, called transferrin binding protein B (TbpB).

Studies have shown that although transferrin is highly conserved between organisms, Tbps are not. The variability of the Tbp receptor, which may contribute to immune system evasion, is pronounced in the receptor's transferrin binding sites, which reside in the hypervariable loop regions of both receptor proteins. Furthermore, while TbpB recognizes conserved structural elements in transferrin, there appears to be sequence variation in these regions.

Although most Tbp receptors are host-specific, there are some cases in which Tbp-containing bacteria are known to recognize the transferrin of more than one host. One of these, *Mannheimia haemolytica*, can infect and recognize the transferrins of domestic cattle, sheep, and goats. In contrast, strains of *Haemophilus somnus*, another pathogen that shares the same niche, have been shown to solely recognize bovine transferrin. However, an exception to this has been found. One strain has been shown to bind all three transferrins and possesses a third transferrin binding protein that has previously been identified in only a few bacterial species that can also infect the three aforementioned hosts. This protein, named TbpA2, belongs to the same family as TbpA, and binds a different, more highly conserved region of trasferrin.

This discovery has prompted us to investigate how the transferrin receptors in each pathogen can produce the same phenotype and whether the differences in receptor composition and receptor structure may provide selective advantages in certain environments. Towards this, we have solved several TbpB structures, including the x-ray crystal the structure of *Mannheimia haemolytica* TbpB, and performed affinity capture experiments to study the composition of each pathogen's receptor system. Furthermore, we have performed preliminary mutational analyses and binding studies to identify binding determinants in TbpB and to define TbpB specificities for the three ruminant transferrins.

P-2157

Phenol metabolisms in yeasts: identification and evolutionary analysis of genes involved in gentisate and 3-oxoapidate pathways in Candida parapsilosis

Leszek Pryszcz[1], Josef Nosek[2], Toni Gabaldón[1]
[1]Department of Bioinformatics and Genomics, CRG-Centre for Genomic Regulation, Barcelona, Spain, [2]Department of Biochemistry, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia

Candida parapsilosis is an emerging fungal human pathogen. It is able to degrade various hydroxy derivatives of benzenes and benzoates through two pathways: gentisate and 3-oxoapidate. However, genes involved in these pathways remain unknown. We analyzed transcriptomes (RNA-Seq) of Candida parapsilosis cultivated on three carbon sources: glucose, 3- and 4-hydroxybenzoate. Among highly up-regulated genes, we identified seven possibly coding for enzymes that are involved in these pathways. Recently, some of these were confirmed1. We proposed transcription factors regulating switch of metabolism from glucose to hydroxy benzoates. Moreover, we identified putative membrane transporters required to import 3- and 4-hydroxybenzoate.
Subsequently, the evolutionary history of gentisate and 3-oxoapidate pathways in yeasts was analyzed. We didn't find any evidence of horizontal gene transfer. Complete pathways were found only in 5 out of the 25 sequenced Saccharomycotina species. Among these, Pichia stipitis is known to metabolize phenols. The remaining four, Candida parapsilosis and another three Candida species recently sequenced by us, are emerging human pathogens. This suggests, gentisate and 3-oxoapidate pathways may play important role in pathogenesis.

1 Holesova, Z.; Jakubkova, M.; Zavadiakova, I.; Zeman, I.; Tomaska, L.; Nosek, J. Microbiology 2011, 157: 2152-63

P-2158

A multiplex platform for continuous culture and experimental evolution in yeast

Aaron Miller, Mei Huang, Annie Young, Bryony Lynch, Maitreya Dunham
University of Washington, Seattle, WA, USA

Chemostats are a continuous culture system in which cells are grown in a tightly controlled, chemically constant environment where culture density is constrained by limiting specific nutrients. Unlike batch cultures, the selective pressure imposed in a chemostat is constant, which has made them a desirable tool for evolution, competition, and physiology experiments. Traditionally, chemostat experiments with *S. cerevisiae* have been constrained by cost and throughput. In order to solve this problem we have developed an inexpensive continuous culture system that allows 32 chemostats to be run simultaneously using entirely off the shelf parts.

Using this system we have evolved dozens of yeast cultures under constant sulfate, glucose or phosphate limitation for 300 generations. We are now using a variety of techniques including whole genome sequencing of evolved populations to characterize mutations that have been selected for during the course of these evolution experiments. Changes in the genome can be correlated to changes in relative fitness, which we assayed through competition with a GFP expressing ancestral strain every 50 generations.

Population fitness estimates suggest that there may be complex interactions between subpopulations that arise throughout the course of a given evolution. We have begun to characterize the nature of these subpopulation interactions through population sequencing and through competitive fitness estimates for clones derived across different time points for a given population.

Ultimately It is our hope that higher replicate number linked to changes in fitness will better reveal the spectrum of genes important for adaptive growth in nutrient-limited environments. Furthermore we would like to use these methods to understand what genes are important in a wide variety of selective environments.

P-2159

Genomewide association study for melanoma susceptibility

Susanne Horn
German Cancer Research Center, Heidelberg, Germany

Melanoma with its propensity to metastasize is associated highest lethality of all skin cancers. It is fairly common in Caucasian populations with around 70,000 new cases reported in the United States in 2011 (Howlader et al. 2011). The development of metastatic melanoma is characterized by a poor prognosis with overall 5-year survival being only 15 percent. In order to determine genetic determinants of the disease susceptibility and outcome we carried out a genome wide association study (GWAS). 1218 melanoma patients of European ancestry were genotyped for about 300,000 SNPs using the Illumina HumanCytoSNP array. Two sets of 909 and 1223 healthy control individuals were genotyped on microarrays (HumanOmniExpress and HumanOmni1-Quad) with 730,000 and 1,000,000 markers, respectively. After quality control of the data the genotype information of the cases was compared with that of the controls. The data were corrected for population structure using eigenstrat adjustment. We present preliminary results from the initial genome wide scan as well as the subsequent validation phase. Previously unknown loci were followed up for validation in two additional populations of melanoma cases and controls. We used HapMap linkage data to select SNPs for validation and to obtain haplotype information for the respective regions. Genotyping for validation was performed using allelic-discrimination based PCR and Sanger sequencing. The identification of new genetic loci associated with melanoma will provide additional information about the influence of genetic variation on the risk of melanoma for understanding the disease and eventual prognosis.

P-2160

Adaptive evolution of Mediterranean pines

Delphine Grivet[1], Zaida Lorenzo[1], José Climent[1], Mario Zabal-Aguirre[1], Sara Torre[4], David B. Neale[2,3], Giovanni G. Vendramin[4], Santiago C. González Martínez[1]
[1]INIA-CIFOR, Madrid, Spain,[2]Department of Plant Sciences,University of California, Davis, Davis, California, USA, [3]Center for Population Biology, University of California, Davis, Davis, California, USA,[4]Plant Genetics Institute, Division of Florence, National Research Council, Sesto Fiorentino, Florence, Italy

Mediterranean pines represent an extremely heterogeneous and interesting assembly. Although they have evolved under similar environmental conditions, they diversified long ago, back to the Miocene, and present distinct biogeographic and demographic histories. Therefore, it is of special interest to understand if and to what extent they have developed specific strategies of adaptive evolution through time and space. In order to explore evolutionary patterns, we first establish their phylogeny analyzing a new set of 21 low-copy nuclear genes with some Bayesian tree reconstruction methods that integrate multilocus approaches. Secondly, to gain some insights on Mediterranean pines adaptive evolution, we look for footprint of natural selection in candidate genes as well as at the evolution of phenotypic traits across the phylogeny.

P-2161

Duplications in Streamlined Apicomplexan Genomes

Jeremy DeBarry, Jessica Kissinger
The University of Georgia, Athens, Georgia, USA

We have characterized the distribution, divergence, expression, and putative function of two-copy paralogs and segmental duplications in a phylum of parasitic protists, the Apicomplexa. Apicomplexans are obligate intracellular parasites responsible for a wide range of human and veterinary diseases (e.g. malaria, toxoplasmosis, and cryptosporidiosis). Gene loss has proven to be a major force in the phylum. Genome sizes are small (~9 - 63 Mb) and protein-encoding gene repertoires are highly reduced (~3400 - 8000/genome). Despite genomic streamlining, duplications and gene family amplifications are detected. The fate and potential for innovation introduced by duplications is of particular interest. We clustered orthologs and paralogs from 12 species in 6 genera. We then identified all two-copy paralogs and segmental duplications, treating each as a discreet class of genes to investigate their evolutionary and adaptive histories. We investigated the patterns of duplicate-gene distribution and show that species have biases toward either tandem or non-tandem locations. The observed distributions may be indicative of the method of duplication and/or the degree of rearrangement since duplication. Not surprisingly, there is little overall nucleotide sequence conservation between paralogs. In some species, higher conservation is found when pairs are distributed in tandem, while in others, distribution appears to play little role in the level of conservation. When pairs have many orthologs in multiple species, they tend to have greater similarity to their orthologs than to each other. Overall, gene pairs with few orthologs have the highest similarity to each other. Preliminary analyses show that most paralogs for which we have expression data do not have similar levels of expression over the same time points. This could mean that duplicates have new roles, or they may be exploited for temporal regulation. There is not a clear link between chromosomal distribution and the conservation of expression levels and timing, though in at least one species, it appears that tandem pairs have similar patterns for both. Enrichment analyses of GO terms show that two-copy paralogs in most species are enriched for multiple functions. An enrichment common to many genera is cysteine-protease activity. Cysteine proteases are known to play key roles in host immune evasion and cell invasion and their duplication is likely beneficial.

P-2162

Evaluating the Ortholog Conjecture using the Structure Function Linkage Database

Kimmen Sjolander[1], Patricia Babbitt[2]
[1]University of California Berkeley, Berkeley, CA, USA, [2]University of California San Francisco, San Francisco, CA, USA

The "ortholog conjecture" - that orthologs are more likely to share a common function than paralogs -- has recently been called into question. Given the widespread use of orthology relationships as a basis for functional annotation (1), examination of the actual agreement in function across orthologs is clearly warranted.

In brief, orthology is a phylogenetic term; two genes are orthologs if they are related by speciation from a common ancestral gene. Orthology subtypes exist, complicating matters. Two genes are each other's super-orthologs if and only if every node on a path joining the two genes in a gene tree corresponds to a speciation event. Two genes are each other's ultraparalogs if all nodes on a path joining them in an inferred gene tree correspond to duplication events. Moreover, the propensity of homology searches to include sequences sharing only partial homology can result in proteins having different multi-domain architectures being called orthologs.

Given these orthology and paralogy subtypes, the complications of promiscuous domains, and the prevalence of functional annotation errors (2), we anticipate that the majority of cases of "orthologs" having lower functional similarity than paralogs can be attributed to the use of the inclusive definitions of orthology, changes in domain architecture, or errors in functional annotation. In brief, we expect that if the most stringent definition of orthology is used - requiring orthologs to agree at the multi-domain architecture and using super-orthology instead of simple orthology - that the "ortholog conjecture" will be shown to be largely supported.

We will present detailed analyses of a mechanistically diverse set of enzyme superfamilies in the Structure Function Linkage Database (3) to evaluate the agreement between orthology and function. While exceptions to the ortholog conjecture are likely to be relatively infrequent, they should provide valuable insights to guide the development of novel bioinformatics methods making use of sequence, structural and phylogenetic signals to predict functional properties across orthologs.

1. Sjolander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics, 20, 170-179.
2. Schnoes, A.M., Brown, S.D., Dodevski, I. and Babbitt, P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS computational biology, 5, e1000605.
3. Brown, S.D., Gerlt, J.A., Seffernick, J.L. and Babbitt, P.C. (2006) A gold standard set of mechanistically diverse enzyme superfamilies. Genome biology, 7, R8.

P-2163

Ortho-Gen: The ortholog dataset constructing tool for phylogenetic analysis

Tokumasa Horiike[1], Ryoichi Mimai[1], Daisuke Miyata[1,2], Yoshio Tateno[1,3]
[1]Shizuoka University, Shizuoka, Japan,[2]Chiba University of Commerce, Ichikawa, Japan,[3]Pohang University of Science and Technology, Pohang, Republic of Korea

Phylogenetic trees of prokaryotes based on the sequence data of single genes are often inconsistent to one another. To overcome this problem, supertree (concatenated tree) was developed, and it showed plausible species phylogeny. However, it tend to yield inaccurate relationships particularly for distantly related species due to disturbing factors such as horizontal gene transfer and gene loss in out-paralogs. We developed new method named **"Ortho-Gen"** to construct ortholog dataset for phylogenetic analysis. We introduced the following four ideas to the method to decrease the influences of horizontal gene transfer and out-paralogs.

HGT filter: Deletion of genes that are derived from horizontal gene transfer from initial sequence dataset.

Out-paralog filter: Deletion of out-paralogs from result data of BLAST using the similarity score. Remaining out-paralogs are deleted in the following steps.

Tree split: This program split a tree into two to remove the out-paralogs from the candidates of ortholog dataset with the information of phylogenetic tree's topology, such as monophyletic or polyphyletic states in the group and species level.

Changing threshold: This program cut the candidate of ortholog group (including out-paralog) into two ortholog groups by the difference of evolutionary distance among true ortholog members and that among out-paralogs.

P-2164

A Transcript Perspective on Evolution

Yann Christinat, Bernard Moret
EPFL, Lausanne, Switzerland

Gene duplication, loss, and mutation are the main driving forces for transcriptome and proteome diversity. However, alternative splicing—a greatly underestimated mechanism twenty years ago—has now be shown to play a major role for diversity in higher eukaryotes and especially in humans. Some researchers indeed conjecture that 90% of human multi-exon genes are alternatively spliced.

In spite of the importance of alternative splicing, the study of evolution from a transcript perspective has not seen much work; the evolution of the mechanism itself is also poorly understood. Alternative splicing yields differences in isoforms that can cover entire domains and thus allows complete removal of functions. However, due to poor transcriptome annotation, the link between transcripts and functions—be it gene ontology terms, tissue localization, or developmental stages—is not well understood.

We will present a new tool and model to help in understanding the evolution of alternative splicing.
The method, given a set of transcripts and an associated gene tree, reconstructs the most parsimonious forest of transcript trees. Each tree represents the evolution, along the gene tree, of an ancestral transcript towards its present-day children.

As an application of the method, we followed the case study in a recent article by Nerht *et al.* (PloS Comput. Biol. 2011) and ran our algorithm on the MAP4K2 homologs in human and mouse. Human MAP4K2 transcripts were better clustered with their paralogs than their mouse orthologs, a result in line with the conclusions of Nerht *et al.*. Interestingly, the human transcripts were clustered with noncoding transcripts from its mouse ortholog and vice-versa. We also noted that human MAP4K1 transcripts were consistently well associated with their mouse orthologs. The ortholog conjecture may thus hold true for MAP4K1 but not for MAP4K2.

P-2165

Building Structure-based Classification from Multiple Sequence Alignment with the T-RMSD method

Cedrik Magis[1,2], Cedric Notredame[1,2]
[1]*Center for Genomic Regulation (CRG), Barcelona, Catalunya, Spain,* [2]*Universitat Pompeu Fabra (UPF), Barcelona, Catalunya, Spain*

Phylogenetic reconstructions are usually based on multiple sequence alignments. These analyses require aligning the sequences and estimating their time of divergence on the basis of their similarity. Due to significant variations of the selective pressure under which they evolve, all positions in a DNA sequence do not mutate at the same rate. For instance in proteins coding regions, the most constrained positions are those whose variation results in an amino acid substitution. These substitutions are therefore more informative to estimate the relationship between distantly related sequences, though they may eventually saturate when considering sequences distant enough. By contrast, corresponding three-dimensional structures appear to be evolving slower as a consequence of stronger signal resilience; meaningful evolutionary similarities can sometimes be estimated by comparing the structures of protein sequences having diverged beyond recognition. We present here T-RMSD, a novel method able to derive a tree-like classification based on the structure-based comparison of a group of related sequences. Its main advantage is to allow a precise quantification of the structural support for each individual node.
T-RMSD is a distance RMSD (dRMSD) based method. It estimates the topology of a tree by comparing intra-molecular distances between equivalent C-alphas. Equivalences are derived from a structure-based multiple sequence alignment produced using 3D-Coffee. On each ungapped column, differences of distances are combined in order to fill-up a distance matrix further resolved into a Neighbor-Joining tree. The collection of trees thus produced (one per column) is eventually turned into a consensus tree using CONSENSE from PHYLIP. For each node, the number of supporting trees is estimated, and used to quantify the structural support of the considered split. We applied the T-RMSD procedure to several well-characterized protein domain families. Our results show that on all cases, we managed to recover and sometimes refine existing classifications. Of course, the question of whether these reconstructions are phylogenetically meaningful remains open, with convergent evolution being a major confounding factor when dealing with structures. Yet, the method thus provided is now ready to be evaluated by the community with respect to its capacity of reconstructing ancient evolutionary relationships. T-RMSD is incorporated within the T-Coffee package an open-source freeware available from www.tcoffee.org.

P-2166

Gene function prediction in plant genomes from large scale phylogenomics analyses

Mathieu Rouard[1], Jean-François Dufayard[2], Valentin Guignon[1], Mathieu Conte[4], Anne-Muriel Arigon[3], Vincent Berry[3]
[1]Bioversity International, Commodities for Livelihood Programme, Montpellier, France,[2]Cirad, Department BIOS, UMR DAP, Montpellier, France,[3]MAB, LIRMM, CNRS, Université Montpellier, Montpellier, France,[4]SYNGENTA Seeds SAS, St Sauveur, France

With the increasing number of plant genomes being sequenced, a major work is to accurately transfer annotation from characterized sequences to newly obtained sequences. GreenPhylDB is a database designed for comparative and functional genomics based on complete genomes (Conte et al, 2008, Rouard et al, 2011). The database currently comprises gene families of protein sequences from 16 plant species, including socio-economically important crops like rice, sorghum and maize that. Gene families are manually annotated (i.e. properly named and classified) and then analyzed phylogenetically in order to elucidate evolutionary relationships (e.g. orthologs, super-orthologs, in/out-paralogs) between sequences. The GreenPhyl's pipeline relies on RapGreen, a new version of the RAP reconciliation tool (Dufayard et al, 2005) that allows us root gene trees and infer orthology relationships between sequences of a family. This relationship, together with new information sources and methods, are currently investigated to explore gene function prediction. For instance, cross-references to gene expression databases are available in order to compare expression profiles; orthologous genes predicted by both phylogenetic-based and similarity-based methods are summarized per gene family. For each pair of orthologous sequences, GreenPhylDB computes functional similarities between associated Gene Ontology terms. Overall, this bioinformatic resource available at http://greenphyl.cirad.fr is particularly helpful for gene annotation.

P-2167

Functional divergence and convergent evolution in the plastid-targeted glyceraldehyde-3-phosphate dehydrogenases of Archeasplastida and Chromalveolata. Gene duplication, endosymbiositic gene transfer, and two independent transitions to Calvin cycle function

Daniel Gaston, Edward Susko, Andrew Roger
Dalhousie University, Halifax, Nova Scotia, Canada

The functional divergence of genes (whether orthologs, paralogs, or xenologs) is a major driving force of molecular innovation and evolutionary novelty. In order to characterize the mechanisms driving this divergence, and to identify the key amino acid residues in proteins contributing to differences in function, we are often interested in monophyletic assemblages of sequences (sub-families) with some shared molecular or biological role which can be compared to background sequences or other sub-families. Under these conditions, the power of phylogenetic methods can be brought to bear.

To predict the amino acid residues contributing to the functional divergence of two or more monophyletic groups of sequences in a protein family, we have developed a maximum-likelihood based phylogenetic mixture model called FunDi. This two-component mixture model captures both the evolutionary dependence and common ancestry of sequences in the standard ‘dependent' component, as well as the evolutionarily ‘uncoupled' nature of functional divergence under a new ‘independent' component. FunDi is developed as a flexible framework for maximum-likelihood computation to allow for the easy integration of new methods and models of protein evolution.

We compare FunDi to several existing classifiers from the literature on a set of reasonably well described biological datasets. In addition, we construct two novel strategies for simulating functional divergence and compare these classifiers under 470 simulated cases with a variety of phylogenetic tree shapes, sizes, and numbers of taxa. Finally we use FunDi, and two other high performing classification methods, to characterize the functional divergence of two independent groups of plastid-targeted glyceraldehyde-3-phosphate dehydrogenases, those of the Archeaplastida (cyanobacterial -derived, GapA/B), and the Chromalveolata (GapC1). GapC1 represents a novel, independent transition of a eukaryotic cytosol-localized GAPDH to plastid and Calvin cycle function after duplication while the ‘canonical' plastid-GAPDH is of endosymbiotic origin. GapC1 and GapA/B sequences therefore represent two distinct cases of functional divergence from cytosolic GAPDHs as well as a case of functional convergence in plastid-targeting, Calvin cycle function, and enzymatic regulation. We find that FunDi performs comparably to existing classifiers on several biological datasets, but performs better under our wide array of simulated conditions. Additionally FunDi outperforms two other high-performing methods in our GAPDH analyses, identifying more putative cases of convergent evolution under functional divergence and in identifying sites known to be functionally divergent based on experimental data.

P-2168

The evolutionary history of the eukaryotic ribosome biogenesis pathway from a yeast perspective.

Ingo Ebersberger[1], Stefan Simm[2], Enrico Schleiff[2]
[1]CIBIV MFPL, Vienna, Austria, [2]Institute for Molecular Biosciences, Goethe University, Frankfurt, Germany

The identification of orthologous proteins as functional equivalents is a central concept in evolutionary research. Still, examples exist where functional equivalents are not homologous or where orthologous proteins have diverged in their function. Only recently it has been suggested that in many cases proteins whose genes split by duplication, i.e. paralogs, are more similar in function than their respective orthologs. This renders already the tracing of an individual protein—or better of the function it conveys—in a phylogenetic tree complex. The situation becomes substantially worse when analyzing the evolution of entire metabolic pathways. Ribosome biogenesis is present in all living organisms. In eukaryotes it has been mainly investigated in *Saccharomyces cerevisiae*, and currently 254 yeast proteins are integrated into this pathway. Much less is known in animals or plants. Meanwhile, individual components of this pathway have been identified in species outside the fungi, some of which lack a counterpart in yeast. Thus, although the evolutionary age of ribosome biogenesis and its functional relevance suggest that major parts of this pathway are conserved among the eukaryotes, it may have a hitherto unrecognized functional plasticity.

Here, we will investigate the evolutionary history of ribosome biogenesis by tracing its components identified in yeast in 268 eukaryotic genomes and 12 archaea. We set out by showing that the recently proposed limitation of the orthology conjecture can be, at least partially, explained by community specific differences in the GO annotation of gene products. We continue by delineating the evolutionary stable scaffold of ribosome biogenesis from the phylogenetic profile of the 254 yeast ribosome biogenesis factors. About a quarter of these can be traced back to the common ancestor of eukaryotes and archaea. Eventually, we concentrate on the interpretation of gaps in the phylogenetic profile. We present a novel approach to assess whether an ortholog may be present but is likely to have diverged to an extent that it has been missed by the search and cases of true absence. In the latter case we investigate whether paralogs or even non-homologous proteins may have overtaken the missing functionality or if it is indeed absent. As a result a comprehensive evolutionary picture of ribosome biogenesis in eukaryotes will emerge.

P-2169

Understanding the exceptional amplification of the STS gene family involved in resveratrol synthesis in grapevine

Claire Parage[1], Raquel Tavares[2], Stéphane Réty[4], Raymonde Baltenweck[1], Anne Poutaraud[1], Dimitri Heintz[5], Raphaël Lugan[5], Gabriel Marais[2], Sébastian Aubourg[3], Philippe Hugueney[1]
[1]University f Strasbourg / INRA, Colmar, France,[2]University of Lyon1 / CNRS, Villeurbanne, France,[3]INRA, Evry, France,[4]CNRS, Paris, France,[5]CNRS, Strasbourg, France

Stilbenes are a small family of phenylpropanoids produced in a number of unrelated plant species including grapevine (*Vitis vinifera*). Stilbenes are involved in constitutive and inducible defense mechanisms in plants. One of the stilbenes, resveratrol, present in red wine, has been shown to slow down the progression of a number of diseases (e.g. cancer, cardiovascular diseases) and is involved in the increased lifespan of south-west French population known as “the French paradox”. Stilbene synthases (STS) are part of the chalcone synthases (CHS) gene family, and seem to have emerged several times independently from the CHS in stilbene-producing plants. The grapevine genome includes an exceptionally large STS gene family (37 complete functional paralogues). Here we used the recently released 12X genome sequence of *V. vinifera* inbred Pinot Noir PN40024 to work on the STS gene family. We combine molecular evolution, structural and functional analyses to understand the remarkable amplification of the STS family in grapevine.

P-2170

Evolution of BCL-2 homologous proteins: a trip to Indels

Abdel AOUACHERIA
CNRS LBMC UMR5239, Lyon, France

BCL-2 is the prototypical member of a large protein family positively or negatively influencing apoptotic cell death, an evolutionarily conserved process in metazoans. In mammals, the repertoire of proteins homologous to BCL-2 comprises about fourteen members that share a similar 3D structural fold and supposedly common ancestry. There is no clear explanation as yet of how these structurally homologous proteins evolved opposite functions with respect to apoptosis regulation, making these regulators an extreme instance of functional divergence. The availability of experimental data and molecular phylogenetic tools, and the increasing size of sequence databases provide the potential to establish orthologous and paralogous relationships, and make this family of proteins a good model for phylogenomic analysis. At a basic level, Bcl-2 homologous proteins can be considered as consisting of two halves: (i) a variable N-terminal part which shows divergence both in sequence and in length between paralogs and which contains multiple types of regulatory motifs; (ii) a more constrained C-terminal moiety made out of alpha-helices packed together. Multiple lines of evidence indicate that insertions or deletions (indels) of amino acids in the N-terminal region gave rise to a significant degree of structural variability among paralogs, which appears to convert into functional versatility. Just recently attention has been paid to indels as a source of evolutionary change between Bcl-2 family orthologous proteins as well. As an example, we recently reported that BCL2L10, a BCL-2 paralog expressed in eggs and oocytes, has evolved in the lineage leading to humans a calcium binding site in its C-terminal region, a property that may be of adaptive value in the context of maternal expression. We suggest that the study of indels combined to that of amino acid substitutions can improve resolution of the BCL-2 family tree and give hints to the development of paralog-selective drugs.

References

Guillemin Y, Cornut-Thibaut A, Gillet G, Penin F, Aouacheria A. Characterization of unique signature sequences in the divergent maternal protein Bcl2l10. Mol Biol Evol. 2011 Dec;28(12):3271-83.

Guillemin Y, Lalle P, Gillet G, Guerin JF, Hamamah S, Aouacheria A. Oocytes and early embryos selectively express the survival factor BCL2L10. J Mol Med (Berl). 2009 Sep;87(9):923-40.

Blaineau SV, Aouacheria A. BCL2DB: moving 'helix-bundled' BCL-2 family members to their database. Apoptosis. 2009 Jul;14(7):923-5.

Phylogenomics of life-or-death switches in multicellular animals: Bcl-2, BH3-Only, and BNip families of apoptotic regulators. Aouacheria A, Brunet F, Gouy M. Mol Biol Evol. 2005 Dec;22(12):2395-416.

P-2171

MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score.

Leszek Pryszcz, Jaime Huerta-Cepas, Toni Gabaldón
CRG-Centre for Genomic Regulation, Barcelona, Spain

Reliable prediction of orthology is central to comparative genomics. Approaches based on phylogenetic analyses closely resemble the original definition of orthology and paralogy and are known to be highly accurate. However, the large computational cost associated to these analyses is a limiting factor that often prevents its use at genomic scales. Recently, several projects have addressed the reconstruction of large collections of high-quality phylogenetic trees from which orthology and paralogy relationships can be inferred. This provides us with the opportunity to infer the evolutionary relationships of genes from multiple, independent, phylogenetic trees. Using such strategy, we combine phylogenetic information derived from different databases, to predict orthology and paralogy relationships for 12.1 million proteins in 1906 fully-sequenced genomes. We show that the number of independent sources from which a prediction is made, as well as the level of consistency across predictions, can be used as reliable confidence scores. A webserver has been developed to easily access these data (http://orthology.phylomedb.org), which provides users with a global repository of phylogeny-based orthology and paralogy predictions.

P-2172

Asymmetric evolution of *Bmp16* and its sister genes *Bmp2* and *Bmp4*

Nathalie Feiner[1], Gerrit Begemann[1], Axel Meyer[1], Shigehiro Kuraku[1,2]
[1]*University of Konstanz, Konstanz, Germany,* [2]*Center for Developmental Biology RIKEN, Kobe, Japan*

Recently, a novel relative of the well-characterized developmental genes *Bmp2* and *Bmp4*, designated *Bmp16*, was first identified in zebrafish. The notably distinct properties of *Bmp16* compared to those of the highly conserved *Bmp2* and *-4* make it an interesting system to study patterns of functional diversification after gene duplication. Our molecular phylogenetic analysis of the vertebrate *Bmp2/4/16* subfamily revealed an origin of *Bmp16* at the earliest phase of vertebrate evolution. Further comparisons of gene orders in the flanking genomic regions between teleost *bmp2*, *-4* and *-16* demonstrated conserved synteny suggesting their origins during the two-rounds of whole genome duplications (2R-WGD). A careful database survey showed that *Bmp16* orthologs are present in sharks, teleosts, coelacanth and reptiles and indicated secondary gene losses of its orthologs in the amphibian, avian and mammalian lineages for which whole genomes have been sequenced. *In situ* hybridization in embryos of zebrafish, anole lizard and a shark showed that expression patterns of *Bmp16* varies a lot between these animals: while widely expressed in the zebrafish (swim bladder, heart, tail bud, ectoderm of pectoral and median fin folds and gut epithilium), it is only expressed in the limb buds and the ventral part of the tail in the anole lizard and no expression domain at all was found in shark. This variation in *Bmp16* expression is in sharp contrast to the high level of conservation of developmental roles within the groups of *Bmp2* and *Bmp4* genes. To understand the embryonic function of *bmp16* in zebrafish and to disentangle its redundancy with other *Bmp* relatives, we performed morpholino-based knock-down experiments. Despite an overlap of expression domains in the tail bud of *bmp16* with *bmp2a* and *-7*, *bmp16* knock-down showed an impaired tail bud outgrowth. The low level of conservation of *Bmp16* expression patterns across vertebrates together with multiple secondary gene losses indicate an elevated evolutionary rate of *Bmp16* compared to its sister genes *Bmp2* and *-4*. A relative-rate test comparing substitution rates ruled out the possibility that this is caused by hypermutability of the *Bmp16* locus and thus the cause for its elevated evolutionary rate remains in question. This study highlights the necessity to combine gene expression analysis and molecular phylogenetics to shed light on patterns of diversification that act on paralogs upon genome duplications. Comprehensive case studies like this allow us to gain insights into discrepancies in gene repertoires, and ultimately understand their evolutionary impact on developmental programs.

P-2173

Evolution of dFOXO binding sites in the Inslulin-like Receptor (*InR*) gene of Drosophila

Dorcas J. Orengo, Montserrat Aguadé, Elvira Juan
Dept. Genètica, Univ. Barcelona, Avda. Diagonal 645, Spain

The *InR* gene encodes the first component of the Insulin receptor signaling pathway, whose role includes control of growth, development, metabolic homeostasis, reproduction, lifespan and body size. There is some evidence for the action of positive selection in the coding region of this gene. The question arises of whether the regulatory region has also evolved adaptively. To address this question it is necessary to precisely define the binding sites for transcription factors. When nutrient conditions are limiting, *InR* transcription is activated by the binding dFOXO upstream of the P2 promoter. Since dFOXO has a DNA binding domain highly conserved in the 12 sequenced species of Drosophila, we used dFOXO protein from *D. melanogaster* to characterize dFOXO binding sites (BS) in a 1,3kb fragment upstream of *InR* P2 promoter in 5 species of Drosophila. We have precisely characterized dFOXO BS by automatic DNase I footprinting and have obtained the position weight matrix (PWM) for dFOXO. In order to get some insight into the evolutionary forces underlying the evolution of binding sites, we have also surveyed nucleotide variation in both *D. melanogaster* and *D. simulans*.

P-2174

Evolutionary dynamics of MADS-box transcription factor binding sites in plants

Suzanne de Bruijn[1], Jose M Muino[2], Gerco Angenent[2], Kerstin Kaufmann[1]
[1]Wageningen University, laboratory for molecular biology, Wageningen, The Netherlands,[2]Plant Research International, department Bioscience, Wageningen, The Netherlands

It is known that changes in *cis*-regulatory elements (CREs) which are binding sites of transcription factors (TFs) play an important role during evolution. Mutations in CREs mediate changes in gene expression. These changes can be either spatial, temporal, or in the level of expression. Although there are several examples of changes in CREs and their consequences for plant evolution reported, basic mechanisms and dynamics of cis-regulatory evolution are still largely unexplored. To study general mechanisms of CRE evolution, we studied the *in vivo* DNA-binding sites of a TF in two closely related plant species.

The goal of our project is to get a better understanding of the evolutionary dynamics of TF binding sites in plants and the potential consequences for gene regulation. We use a combination of ChIP-seq and RNA-seq experiments to study binding sites of a MADS-box TF with a conserved role in flower development and consequences for floral gene expression divergence in two closely related Arabidopsis species. Here, we will provide a general overview of our recent findings.

P-2175

Selection on noncoding DNA in the Brassicaceae

Annabelle Haudry[1,2], Adrian Platts[3], Stephen I. Wright[2], Robert Williamson[2], Emilio Vello[3], Paul Harrison[3], Thomas Bureau[3], Alan Moses[2], Mathieu Blanchette[3]
[1]University of Lyon, Lyon, France,[2]University of Toronto, Toronto, ON, Canada,[3]McGill University, Montreal, QC, Canada

Despite the central role of noncoding DNA in gene regulation and evolution, our understanding of the genomic extent and nature of selection on plant noncoding regions remains limited. Taking advantage of newly sequenced genomes in the Brassicaceae, we combined comparative genomics and population genomics approaches to analyze patterns of molecular evolution and the strength of selection of noncoding DNA. We conducted whole genome alignments of nine Brassicaceae species, including two new genome assemblies from the DOE Joint Genome Institute (*Capsella rubella* and *Eutrema halophila*) and de novo Illumina assemblies of three additional species by the VEGI project (*Aethionema arabicum*, *Sisymbrium irio*, and *Leavenworthia alabamica*). Using a systematic computational approach we identified over 90,000 conserved noncoding regions (CNSs) that show a degree of conservation across species similar to average coding exons and no evidence of expression or coding potential. Gene ontology analysis revealed a significant enrichment of CNSs near transcription factors and CNSs were found to be enriched in known transcription factor binding sites. These results support the potential regulatory function of CNSs. Population genomic analyses in *A. thaliana* and *C. grandiflora* showed reduced levels of diversity and excess of rare variants for CNSs compared to neutral expectation, confirming that most CNSs evolve under purifying selection. Our analysis indicates the strength of selection acting on plant noncoding regions, highlighting the likely importance of these regions for functional variation and adaptive evolution.

P-2176

In pursuit of functional nucleotides: Molecular mechanisms and fine-scale functional testing of noncoding sequences that contribute to pigmentation divergence in *Drosophila*.

Arielle Cooley, Lisa Sramkoski, Wesley McLaughlin, Brad Lankowsky, Patricia Wittkopp
University of Michigan, Ann Arbor, MI, USA

An important yet challenging goal in molecular evolutionary biology is to understand how variation at the level of individual nucleotides shapes phenotypic variation. We are investigating this question for a case of adaptive pigmentation divergence between two closely related Drosophila species, *Drosophila americana* and *D. novamexicana*. Genetic mapping has shown that interspecific body color divergence is largely explained by genomic regions containing the candidate pigmentation genes *tan* and *ebony*. Fine-scale genetic mapping has further localized part of the pigmentation effect to the first intron of *tan*.

Using a series of chimeric transgenic alleles, we are functionally testing small collections of candidate nucleotides within the first intron of *tan*, to examine their contribution to species divergence in body color. To complement this work, we have examined patterns of *cis*- and *trans*-regulatory gene expression divergence at a series of developmental time points, for both *ebony* and *tan*. Together, the data suggest that temporally specific shifts in *cis*-regulatory expression likely contribute to, but may not entirely explain, the recently evolved light yellow body color of *D. novamexicana*. By identifying single nucleotides or small groups of nucleotides responsible for the *cis*-regulatory component of pigmentation divergence, we hope to improve our mechanistic understanding of how noncoding changes contribute to the evolution of diversity.

P-2177

Sequence Features That Determine Transcriptional Enhancers

Jeffrey Chuang[1], Deborah Ritter[1], Ningtao Shi[1], Kourosh Zarringhalam[1], Zhiqiang Dong[2], Su Guo[2]
[1]Boston College, Chestnut Hill, MA, USA,[2]University of California, San Francisco, CA, USA

Distal enhancers play a critical role in the regulation of gene expression in higher eukaryotes, notably during development. However, the sequence features that control enhancer activity, and the mechanisms by which enhancers function, remain poorly understood. We have analyzed multiple large-scale datasets of positive and negative enhancers originally identified via embryonic studies, p300 binding, and histone modification. In particular, these include multiple vertebrate datasets on comparative enhancer activity across mouse, human and zebrafish, and such datasets show significant divergence in activity even for orthologous sequences. Using an SVM and lasso-based classification approach, we have identified the features (sequence motifs, properties of overlapping RNA transcripts, etc) which most strongly explain the evolution of enhancer activity for orthologous sequences. We demonstrate a novel enhancer classifier based on this analysis, and we present results on the relative importance of sequence features versus histone modification. In addition, we present an analysis of coding sequences which we have experimentally verified to have enhancer activity in zebrafish embryos.

P-2178

Modeling expression evolution with varying mutational effect sizes

Rori Rohlfs[1], Patrick Harrigan[3], Rasmus Nielsen[1,2]
[1]University of California, Berkeley, Berkeley, CA, USA, [2]University of Copenhagen, Copenhagen, Denmark, [3]University of California, San Francisco, San Francisco, CA, USA

With the increasing reliability of expression typing methods, particularly RNA-Seq, large comparative expression datasets are now becoming available. These datasets hold great investigatory potential for long-standing hypothesis of changing expression as the mechanism underlying phenotype divergence and adaptation. However, adaptive expression level shifts can be mistaken for non-adaptive expression changes due to large-effect size mutations. Based on the Ornstein-Uhlenbeck process model for expression evolution, we have developed a discretized expression evolution model. The discrete method enables likelihood calculations with arbitrary models for variable mutation effect sizes, including conditioning on expression level. Using this framework, we compare the likelihoods of different models of expression evolution (specifically non-heritable expression, neutral expression drift, stabilizing expression selection, and selective shift in expression) and infer an evolutionary scenario.
We will show both the expected power of these tests for expression level selection using simulations, and results when applied to a diverse panel of mammals. These results highlight gene candidates for conservation of expression level and for recent selection on expression level.

P-2179

Placental endogenous retroviruses facilitate rapid evolution of core trophoblast regulatory network.

Edward Chuong[1], Mohammad Rumi[2], Michael Soares[2], Julie Baker[1]
[1]Stanford University, Stanford, CA, USA,[2]University of Kansas Medical Center, Kansas City, KS, USA

Following uterine implantation, the mammalian embryo undergoes global de novo DNA methylation to silence endogenous retroviruses (ERVs) and other transposons, as their activity poses a severe threat to genome stability. However, the fetal placenta remains hypomethylated relative to somatic tissue, coinciding with high placenta-specific ERV transcription which has been observed across diverse mammalian taxa. Why the placenta seems to permit ERV activity is poorly understood, but an emerging trend is that placental ERVs are commonly coopted for host function, both as proteins (e.g. Syncytins, derived from env proteins) and as regulatory elements (e.g. hCYP19 placental promoter, derived from an LTR). Placental ERV cooptions observed thus far are predominantly lineage-specific, which has given rise to the intriguing hypothesis that species-specific ERV activity may be partly responsible for the dramatic placental morphological variation seen across modern eutherian mammals.

To directly investigate the impact of ERV activity on placenta evolution, we conducted a comparative epigenomic analysis of trophoblast stem cells (TSCs) derived from mouse and rat. Consistent with the placenta's hypomethylated state, we observed widespread colocalization of LTRs and histone modifications associated with active regulation. By utilizing H3K4me1 and H3K27ac as signatures of enhancers, we identified hundreds of enhancers derived from copies of RLTR13, which amplified in mouse 15-25 million years ago and is not present in rat. Analysis of the consensus RLTR13 sequence predicted binding sites for three major TSC-specific regulators: elf5, cdx2, and eomes. Subsequent ChIP-Seq in TSCs revealed extensive co-localization of all three proteins throughout the genome, with half (~300/600) of all multi-bound sites derived from RLTR13. Finally, we compared our data with a dataset of 17 embryonic tissues (mouse ENCODE) and found that genome-scale recruitment of ERVs as enhancers is uncommon in the embryo and is restricted to other hypomethylated tissues including embryonic stem cells and testis.

In summary, we found that ERVs otherwise silenced in the embryo can extensively remodel the regulatory network of placenta development. Given that regulatory mutations are a major force in developmental evolution, genome-wide enhancer cooption of species-specific placental ERVs represents a potential mechanism linking placental hypomethylation and ERV activity with the rapid morphological diversification of the placenta in mammalian evolution.

P-2180

Personal and population genomics of human regulatory variation

Benjamin Vernot, Andrew Stergachis, John Stamatoyannopoulos, Joshua Akey
Department of Genome Sciences, University of Washington, Seattle, WA, USA

The characteristics and evolutionary forces acting on regulatory variation in humans remains elusive because of the difficulty in defining functionally important non-coding DNA. Here, we combine genome-scale maps of regulatory DNA marked by DNaseI hypersensitive sites from 138 cell and tissue types with whole-genome sequences of 53 geographically diverse individuals in order to better delimit the patterns of regulatory variation in humans. We estimate that individuals contain up to seven times as many functional variants in regulatory DNA compared to protein-coding regions, although they are likely to have, on average, smaller effect sizes. Moreover, we demonstrate that there is significant heterogeneity in the level of functional constraint in regulatory DNA among different cell types, with a striking difference between normal and immortal cells. We also find marked variability in functional constraint among transcription factor motifs in regulatory DNA, with sequence motifs for major developmental regulators, such as HOX proteins, exhibiting more constraint than protein-coding regions. Finally, we perform a genome-wide scan of recent positive selection, and identify hundreds of novel substrates of adaptive regulatory evolution that are enriched for biologically interesting pathways such as melanogenesis and adipocytokine signaling. These data and results provide new insights into patterns of regulatory variation in individuals and populations, and demonstrate that a large proportion of functionally important variation lies beyond the exome.

P-2181

Adaptive *cis*-regulatory evolution of age-related differentially expressed genes in the human brain

Michael McGowen[1], Kirstin Sterner[2], Jennifer Baker[3], Chet Sherwood[3], Christopher Kuzawa[4], Harry Chugani[1], Leonard Lipovich[1], Lawrence Grossman[1], Derek Wildman[1]
[1]Wayne State University School of Medicine, Detroit, MI, USA, [2]University of Oregon, Eugene, OR, USA, [3]The George Washington University, Washington, DC, USA, [4]Northwestern University, Evanston, IL, USA

The evolution in humans of an extended juvenile period, altered patterns of brain glucose utilization, and neuroplasticity are associated with the emergence of encephalization, but the genetic basis of these features remains unclear. To investigate this basis, we used oligonucleotide microarrays to examine expression in surgically resected human neocortical tissues spanning ages from infancy to adulthood, and identified 40 transcripts with significant age-related changes in expression. Given the increasing evidence of the central role that regulatory regions of genes have played in human evolution, we investigated whether these observed developmental patterns of gene expression might be related to divergent evolution in non-coding DNA sequences (i.e. promoters, UTRs, introns). Therefore, we downloaded sequences of noncoding regions of these genes in 21 recently sequenced human genomes a subset of data derived from the UCSC Genome Browser and the 1000 Genomes Project. Human genomes were compared with available sequences from other primate genomes (chimpanzee, gorilla, orangutan, gibbon, macaque, marmoset) and fixed differences were identified and quantified. We used an extension of the McDonald-Kreitman test that compares neighboring coding and non-coding sequences to examine adaptive evolution in four distinct regions that usually contain *cis*-regulatory elements: 5' and 3' UTRs, first introns, and promoter regions, defined here as sequences 1000 bp upstream from the transcription start site. We found that multiple genes in our scan show significant evidence of adaptive evolution in at least one *cis*-regulatory region. Among these genes, brevican (*BCAN*), a gene implicated in stabilizing synapses during postnatal brain development, shows evidence of adaptive evolution in its promoter region. In addition, glutamate receptor, ionotropic, NMDA subunit 3A (*GRIN3A*), shows evidence of adaptive evolution in its promoter region as well as its 5' and 3' UTRs. *GRIN3A* as part of the NMDA receptor is directly involved in synaptic transmission and plasticity. Extensive adaptive evolution of *cis*-regulatory features in these genes with temporal variation in expression add to the number of specific genomic changes that have occurred in the human lineage and can be implicated in processes leading to increased cognitive function.

P-2182

A coherent theory of ultraconserved non-coding element evolution

Philipp Bucher[1,2]
[1]*EPFL, Lausanne, Switzerland,* [2]*Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Ultraconserved non-coding elements (UCNEs) are the most conserved sequences in vertebrates. They are typically more than 200 bp long and evolve at rates lower than 1% base substitutions per 100 million years. It is believed that most of them act as tissue-specific enhancers. Besides this, little is understood about their molecular function.

In will present and discuss a coherent theory of vertebrate UNCE evolution which accounts for the hallmark evolutionary properties of these element:

1. very high conservation,
2. strong clustering within the genome,
3. absence of UCNE families,
4. no detectable modularity (absence of UCNE domains).

Properties 2-4 clearly distinguish UCNEs from conserved coding regions. I will argue that all hallmark features can be explained by two assumptions:

1. UCNEs of the same cluster cooperate through a dense interaction network.
2. UCNEs are de novo generated by a series of point mutations.

Briefly, cooperativity explains high conservation and clustering, whereas de novo generation explains the absence of UCNE families and domains.

The cooperativity scenario works as follows: Individual UCNEs interact with many other UCNEs in order to fine-tune the expression of their target genes. This model contrasts with the conventional view that enhancers directly interact with promoters without interference of additional regulatory elements. The idea that many interaction partners implies strong sequence conservation is well accepted for proteins, take histones as an example. The same trend is likely to hold for UCNEs. Interactions between UCNEs would also explain the clustering. Since UCNEs are supposed to cooperate via cis-acting mechanisms, they have to be kept together on the chromosome. The cooperativity hypothesis further explains the observation that UNCE cluster size and degree of conservation are correlated.

De novo generation relates to the hypothesis that UCNEs are arrays of transcription factor binding sites (TFBS).TFBS have low information content and thus are readily generated by point mutations. Of course, a de novo generated single TFBS is not yet a UCNE. However, it may well have a weak phenotype upon which natural selection can act, thereby initiating an adaptive process leading to a full-fledged UCNE. Note the difference to coding regions. Protein domains, the functional units of regulatory proteins, are too complex to be generated by single point mutations. Therefore, innovation has to proceed via gene duplication, followed by reshuffling and diversification, leading to the family structure and modular design seen in the protein universe.

P-2183

A Mammalian Conserved Element Derived from SINE Displays Enhancer Properties Recapitulating Satb2 Expression in Early-Born Callosal Projection Neurons

Hidenori Nishihara[1], Anne Teissier[2], Naoki Kobayashi[1], Kensuke Tashiro[1], Akiko Nakanishi[1], Takeshi Sasaki[3], Masaki Iwakiri[1], Kazuki Izawa[1], Kuo Yan[4], Victor Tarabykin[4], Kenta Sumiyama[5], Mika Hirakawa[6], Alessandra Pierani[2], Norihiro Okada[1]

[1]Tokyo Institute of Technology, Yokohama, Japan,[2]Université Paris Diderot, Paris, France,[3]Tokyo University of Agriculture, Kanagawa, Japan,[4]Max Planck Institute for Experimental Medicine, Göttingen, Germany,[5]National Institute of Genetics, Mishima, Japan,[6]JT Biohistory Research Hall, Osaka, Japan

Transposable elements (TEs) are highly repeated sequences that account for a significant proportion of many eukaryotic genomes. Although TEs are usually considered "junk DNA", a part of mammalian TEs have been recently reported to act as *cis*-regulatory elements (exaptation) such as enhancers and insulators, suggesting that they may have acquired significant functions involved in controlling mammalian-specific traits. However detailed analysis of the exapted TEs remains insufficient in terms of its molecular mechanisms such as identification of transcription factors and of its contribution to the development and evolution of the mammalian traits. We aimed to demonstrate how TEs such as short interspersed elements (SINEs) are involved in the development of mammalian brain. We previously discovered that many AmnSINE1 loci are evolutionarily conserved across mammalian genomes and that one of them (the AS021 locus) located 390 kbp upstream of Satb2 act as a distal enhancer in mammalian cerebral cortex. The transcription factor Satb2 is expressed by cortical neurons extending axons through the corpus callosum and is a determinant of callosal versus subcortical projection. Here we show that the activity of the AS021 enhancer recapitulates the expression of Satb2 at later embryonic and postnatal stages in deep-layer neurons. In addition, we demonstrate that the AS021 enhancer is activated in neurons projecting through the corpus callosum, as described for Satb2$^+$ neurons. Notably, AS021 drives specific expression in axons crossing through the ventral portion of the corpus callosum. By analyzing transcription factor binding sites within conserved AmnSINE1 sequences, we propose a possible regulatory network in which AS021 enhancer might be involved. These results suggest that exaptation of the AS021 SINE locus might have been involved in the establishment of interhemispheric communication via the corpus callosum, a eutherian-specific brain structure.

P-2184

SINE-derived element as a regional determinant for the hypothalamic specificity of mammalian diencephalic *Fgf8* enhancer

Akiko Nakanishi[1], Naoki Kobayashi[1], Asuka Hirano-Suzuki[2], Hidenori Nishihara[1], Takeshi Sasaki[1,5], Mika Hirakawa[3,6], Kenta Sumiyama[4], Tomomi Shimogori[2], Norihiro Okada[1]
[1]Graduate School of Biotechnology, Tokyo Institute of Technology, 4259-B-21, Nagatsuta-cho, Midori-ku,Yokohama, Kanagawa, Japan, [2]RIKEN Brain Science Institute, 2-1 Hirosawa, Wako City, Saitama, Japan, [3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, Japan, [4]National Institute of Genetics, Mishima, Shizuoka, Japan, [5]Department of Human and Animal-Plant Relationships, Faculty of Agriculture, Tokyo University of Agriculture, Atsugi-shi, Kanagawa, Japan, [6]Science Communication and Pruduction, JT Biohistory Research Hall, Takatsuki, Osaka, Japan

Transposable elements, including short interspersed repetitive elements (SINEs), compose nearly half of the mammalian genome, and are a major source of conserved non-coding elements (CNEs). CNEs are considered to possess important roles on the regulation of development genes, which serve for diversification of morphological and physiological features among species. We previously reported a novel SINE family, AmnSINE1 that make up a part of mammalian-specific CNEs. Among these AmnSINE1 loci, the AS071 locus showed the enhancer property in the developing mouse diencephalon, which appears to recapitulate a part of the expression of *Fgf8*.

In this study, by comparing the expression pattern of AS071-enhanced lacZ with that of *Fgf8* for wide embryonic stages, we confirmed that the AS071 locus is a distal enhancer that directs diencephalic *Fgf8* expression. Furthermore, by the enhancer assays using partially-deleted AS071 constructs, we revealed its unique modular organization, in which the AS071-enhancer contains at least three functionally distinct sub-elements that cooperatively regulate enhancer activity in three diencephalic domains including hypothalamus. Interestingly, AmnSINE1-derived sub-element was found to be specialized in the determination of regional specificity to hypothalamus. To our knowledge, this is a novel discovery in which whole enhancer element could be separated into the respective sub-elements that determine the regional specificity and/or the core enhancing activity.

In addition, we found that the *Fgf8* expression in mouse diencephalon is much stronger than that in chick, suggesting that the AS071-enhancer might contribute to the formation of mammalian specific brain through the increase of the *Fgf8* expression in diencephalon.

This is the first report for the enhancer that activates the expression of *Fgf8* in mammalian diencephalon, and its unique modular organization will provide a novel insight for understanding the evolution of *cis*-regulatory element and complexity of the gene regulation during development.

P-2185

Identifying long range cis-regulation in the human genome using evolutionary co-segregation

Magali Naville, Alexandra Louis, Hugues Roest Crollius
ENS, Paris, France

The identification of non-coding regulatory sequences along with their target genes is one of the current challenges in genomics. Due to the complex 3D structure of the vertebrate chromosome, enhancers often target genes distant from several hundreds of base pairs, and simply hypothesizing the nearest gene as being the regulatory target might be wrong. As the validation of target genes still needs heavy experimental procedures, computational prediction of such functional pairs is of great interest. It represents a preliminary step for the better understanding of tight spatio-temporal controls of gene expression.
Our bioinformatic method allows the identification of evolutionary co-segregation between putative enhancers and specific genes, using the information on a large range of genomes.

Putative enhancers are in first approximation assimilated to Conserved Non-coding Elements (CNEs). CNEs are defined based on their conservation in a range of vertebrate species, using an algorithm that scans the UCSC 46 species multiZ alignment and looks for conserved regions of a minimal length and identity. It led to the delimitation of 168.899 CNEs on the X chromosome, covering 3.8% of its total length. These CNEs overlap 61% of the elements identified by the more conservative Siphy algorithm; both types of predictions were finally fused in a global set covering 4.4% of the X chromosome. As the simple conservation criteria do not stand exclusively for enhancers, we further use different annotations to characterize them: ENCODE or literature data such as coactivator p300 binding sites.
The target prediction first consists in collecting immediate neighbour genes of each given CNE in the human reference genome. A score of co-segregation measures the frequency of each CNE-gene pair to be retained in proximity during evolution, taking into account the rearrangement rate of the different genomes. Genes showing the highest values of this score are considered as the most likely targets. Strikingly, CNEs with high scores are enriched in functional annotations, which seems to indicate that elements segregating with genes are more likely to be de facto enhancers.

In collaboration with several human genetics groups, we are working on applying our procedure to genes involved in genetic disorders that could be explained by mutations in such non-coding regions. In a more theoretical way, this study will allow to reconstruct the evolution of chromosomal regulatory circuits, and to analyze the role of enhancers in negative selection of genomic rearrangements.

P-2186

Using Transcription Factor Binding Site Motifs to Predict Enhancer Function for Noncoding Sequences

Dennis Kostka[1,4], Katherine S. Pollard[2,3]
[1]Department of Developmental Biology, University of Pittsburgh, Pittsburgh, USA, [2]Gladstone Institute of Cardiovascular Disease, University of California San Francisco, San Francisco, USA, [3]Division of Biostatistics and Institute of Human Genetics, University of California San Francisco, San Francisco, USA, [4]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, USA

Transcriptional enhancers are short, non-coding sequences that influence the expression of genes. They play crucial roles during development, and the combinatorial binding of transcription factors to these sequences is believed to be a mechanism underlying their function. One step in deciphering the circuitry of gene regulation is the annotation of non-coding sequences as enhancers. We take a discriminative learning approach to predict whether a DNA sequence may function as a transcriptional enhancer. Utilizing sets of positive and negative examples from the VISTA enhancer browser (i.e., DNA sequences that consistently drive expression in transgenic mice), we designed a support vector classifier. To quantify similarity between two sequences for a given transcription factor, we use a model based on correlated Bernoulli trials to calculate p-values for the difference of observed binding sites. This pairwise embedding then forms the basis of the kernels we employ.

With this approach we can accurately classify sequences in the VISTA enhancer browser, for instance as consistently driving gene expression in heart. Further, it is straightforward to assess the importance of single transcription factors. For example, we find that binding sites for NKX2.5, a transcription factor essential to heart development, are amongst the best predictors for heart enhancers. To find sets of transcription factors that perform particularly well, we use the framework of multiple kernel learning. This allows us to explore combinatorial effects of transcription factor binding sites in tissue specific enhancers of different types. This approach can be applied to other types of non-coding sequences, and will be helpful to elucidate the roles various groups of transcription factors play in gene regulation.

P-2187

The effects of alternative splicing and structural disorderness on exonic evolutionary rates

Feng-Chi Chen[1,2], Chia-Lin Pan[1], Hsuan-Yu Lin[1]
[1]National Health Research Institutes, Zhunan, Miaoli County, Taiwan,[2]China Medical University, Taichung, Taiwan

Alternative splicing is known to significantly affect exon-level protein evolutionary rates in mammals. Particularly, alternatively spliced exons (ASEs) have a higher nonsynonymous-to-synonymous substitution rate (*dN/dS*) ratio than constitutively spliced exons (CSEs), possibly because the former are required only occasionally for normal biological functions. Meanwhile, intrinsically disordered regions (IDRs), the protein regions lacking fixed three-dimensional structures, are also reported to have an increased evolutionary rate due to lack of structural constraint. Interestingly, IDRs tend to be located in alternative protein regions. Yet which of these two factors is the major determinant of the increased *dN/dS* in mammalian ASEs remains unclear. By comparing human-macaque and human-mouse one-to-one orthologous genes, we demonstrate that alternative splicing and protein structural disorder have independent effects on mammalian exon evolution. We performed analyses of covariance to demonstrate that the slopes of the (*dN/dS*-percentage of IDR) regression lines differ significantly between CSEs and ASEs. In other words, the *dN/dS* ratios of both ASEs and CSEs increase with the proportion of IDR (PIDR), whereas ASEs have higher *dN/dS* ratios than CSEs when they have similar PIDRs. Since IDRs may less frequently overlap with protein domains (which also affect *dN/dS*), we also examined the correlations between *dN/dS* ratio and PIDR by controlling for the density of protein domain. We found that the effects of PIDR on *dN/dS* are independent of domain density. Our results imply that nature can select for different biological features with regard to ASEs and IDRs, even though the two biological features tend to be localized in the same protein regions.

P-2188

A model for the evolution of a uniquely mammalian T cell receptor

Robert Miller, Zuly Parra
Center for Evolutionary Immunology, University of New Mexico, Albuquerque, NM, USA

T cell receptors (TCR) are a type of antigen receptor that is diversified by somatic DNA recombination in developing T cells resulting in individual antigen specificity. There are four different, conventional TCR chains (α, β, γ, and δ) and their expression as cell-surface heterodimers define the two classes of T cells, αβT and γδT cells. The genes encoding these four chains are found in all jawed vertebrates and are conserved in sequence and organization. Recently a novel subset of TCR has been discovered in the form of TCRδ chains that use variable (V) domains indistinguishable from those used by antibody heavy chains (VH). These TCRd have been found in amphibians, birds, and monotremes (*e.g.* platypus). In addition, the non-eutherian mammals (marsupials and monotremes) have an additional TCR chain, called TCRμ, which also uses VH-like V genes. TCRμ would have been present in the last common mammalian ancestor but was lost in the eutherian lineage. A hypothesis is these atypical TCR chains endow T cells the ability to bind native antigen directly, like B cells, rather than processed-antigen like conventional T cells. TCRμ clearly evolved from duplication and translocation of TCRδ after the divergence of mammals and birds. Comparative analyses of the genomic content and organization of the TCRα/δ and TCRμ genes in amphibians, birds, monotremes, and marsupials has provided a model for understanding both the evolution of TCRδ chains using VH genes, as in amphibians, birds, and monotremes, and the origins and evolution of TCRμ in early mammals. This model involves a clear pattern of gene duplication within the TCRα/δ locus, followed by the insertion of VH genes among the TCRδ genes early in the evolution of the tetrapods. In the galliform birds (chicken and turkey) TCRδ genes including the VH were translocated to a separate chromosome creating a second TCRδ locus producing TCR chains using VH exclusively. TCRμ is the product of a separate, mammal specific duplications/translocations, independent from that which occurred in galliforms. TCRμ later underwent additional internal duplications generating the current form capable of encoding atypical TCR chains that contain double V domains, analogous to TCRd chains found in cartilaginous fish. The independent evolution of these atypical TCR forms in mammals and sharks suggests selection on a common function.

P-2189

The Evolutionary Consequences of Meiotic Recombination:New Insights from Double Strand Breaks in Mouse

Yves Clement[1], Laurent Duret[2], Peter Arndt[1]
[1]Max Planck Institute for Molecular Genetics, Berlin, Germany,[2]UMR CNRS 5558 Laboratoire de Biométrie et Biologie Evolutive, Lyon, France

Meiotic recombination is an important process for sexually reproducing organisms, as it ensures the correct pairing and migration of chromosomes. This process starts with a double strand break which will then be repaired by gene conversion, the copy and paste of one DNA fragment into another. This process will finally lead to either chromosomal crossover or non-crossover. There is evidence that meiotic recombination is a possible major effector of GC-content evolution through a neutral process called GC-biased gene conversion. Current genetic maps available in mammals only have enough power to detect crossovers; non-crossovers, however, cannot be investigated yet. The recent mapping of double strand break hotspots in the mouse genome allows us to study both outcomes and their effect on substitution patterns. To do so, we computed substitution patterns using a genome-wide comparative approach. We first analyzed substitution patterns around double strand breaks and observed a specific and easily identifiable evolutionary signature consistent with GC-biased gene conversion in their close vicinity. Since this signature is not observable in the sister species, we conclude that the location of double strand breaks evolve rapidly. Furthermore, in a genome-wide analysis we see that double strand breaks better predict substitution patterns than crossover rates do and show that both crossovers and non-crossover events influence genome evolution. Finally, we observe a genome-wide recent shift in mouse GC-content evolution to lower GC-content values, interpreted as a decrease of GC-biased gene conversion strength in mouse. Overall, these results give us a much more detailed picture of the influence of recombination on genome evolution in mouse.

P-2190

Large genomic domains resulting from replication-associated mutational asymmetries are conserved since amniota divergence

Chun-Long Chen[1], Antoine Baker[2], Benjamin Audit[2], Arach Goldar[3], Yves d'Aubenton-Carafa[1], Guillaume Guilbaud[4], Aurélien Rappailles[4], Olivier Hyrien[4], Alain Arneodo[2], Claude Thermes[1]
[1]Centre de Génétique Moléculaire, CNRS, Gif-sur-Yvette, France, [2]Laboratoire Joliot Curie, ENS Lyon, CNRS, Lyon, France, [3]CEA, iBiTecS, Gif-sur-Yvette, France, [4]ENS Paris, UMR CNRS 8541, Paris, France

Human genome studies showed that replication induces different mutation rates on the leading and lagging replicating strands; this can generate during evolution strong asymmetries of genome nucleotide composition (Chen et al. MBE 2011). Analysis of this asymmetry, S=(G-C)/(G+C)+(T-A)/(T+A), revealed large ~1Mb-domains exhibiting characteristic N-shaped pattern covering >1/3 of the genome (Huvet et al. Genome Res. 2007). Sharp upward jumps of skew profile at N-domain borders are associated with early replication initiation zones identified as sharp peaks of replication timing profiles. The striking linear decrease of skew S between these initiation zones (N-domain borders) raises three questions: Does this pattern result from replication-associated mutational asymmetries acting over evolutionary periods? Does this reflect a specific spatio-temporal replication organization? Is this N-shaped pattern conserved among vertebrate species?
We showed that replication associated mutational asymmetries decrease from maximum values at left ends of N-domains to zero at centers, and to opposed values at right ends. The skew resulting from these rates acting over evolutionary times strikingly reproduces N-shaped pattern. **This strongly suggests that this N-shaped pattern results from replication-associated mutational asymmetries** and suggests a progressive inversion in replication fork polarity from one end of N-domain to the other. Analyses of replication timing profiles allowed us to demonstrate that mean replication fork polarity can be extracted from the derivative of the timing profile. This profile along N-domains presents a "U-shape" and its derivative is an "N" strongly supporting that **the N-shaped pattern of compositional asymmetries results from specific replication program associated with fork polarity gradients**.
Chromatin structure studies showed that N-domains correspond to chromatin interaction domains and suggested that the replication program within these domains is mediated by gradients of open chromatin conformation. We propose that replication first initiates at N-domain extremities and secondary origins fire coordinately from borders to centers in a domino-like manner. Computational simulations of this model generate linear gradients of replication fork polarity and N-shaped skew profile.
N-domains are observed in all studied mammals, birds and reptile, but not in amphibians and fishes. N-domain extremities are located at homologous regions amongst different amniota genomes. It seems that this replication program has been conserved since amniota divergence, in agreement with the age (~300 Million years) estimated from simulations of skew values resulting from the observed substitution rates. **This indicates that specific spatio-temporal replication program associated with gradients of chromatin structure have driven the genome compositional skew evolution since amniota divergence.**

P-2191

On the track of horizontally transferred sequences between two closely related species Drosophila melanogaster and Drosophila simulans.

Laurent MODOLO, Emmanuelle LERAT
UMR CNRS 5558 Université Lyon1, VILLEURBANNE, France

Transposable elements (TEs) are DNA sequences that can move around their host genome. Long been viewed as junk or selfish DNA, they are nowadays recognized as having a significant part in the evolution of genomes. The genome of the model species D. melanogaster contains around 15% of TEs, whereas its sibling species D. simulans genomes consists of only 5% of these sequences. However, recent studies have shown that D. simulans contains more TE copies than D. melanogaster, but that these copies are more damaged and shorter in D. simulans than in D. melanogaster. Moreover, the high degree of sequence identity of some TEs between D. simulans and D. melanogaster and the fact that D. melanogaster displays more active copies of these TEs can be explained by massive horizontal transfers (HT) of TEs from D. simulans to D. melanogaster, followed by bursts of transposition. In order to explore this hypothesis, we have developed a method to search for HT between these two species.

Horizontal transfers correspond to the transfer of genetic material between species. There are different ways of detecting HT events. Here we have developed a bioinformatics pipeline to detect horizontally transferred regions between two species based on the nucleotide identity between sequences. This approach consists mainly in four steps. 1) The definition of an identity threshold for each chromosomes under study. 2) The pruning of all n x n pairs of sequences identifiable between the two species in order to obtain 1 x 1 pairs of sequences. 3) The test of the identity of these pairs according to the threshold obtained in the first step. 4) The correction of the p-values obtained in the previous step for multiple testing, accounting for the spacial dependency between the pairs of sequences.

We have detected sequences likely to have been horizontally transferred between the two species that can be analyzed to identify CDS, TEs, and intergenic sequences. The coding sequences found were analyzed in order to determine if their high identity between the two species is due to particular evolutionary pressures or reflect true HT. Our approach allowed us to detect all the TEs suspected or proved to have been horizontally transferred between D. simulans and D. melanogaster in the literature, but also new ones. Globally this method, that can be applied to any pair of genomes, allows the accurate identification of horizontally transferred sequences without any a priori concerning their type.

P-2192

Genome-wide copy number variation in African Pygmy hunter-gatherers and Bantu-speaking farmers.

Eddie Loh[1], Etienne Patin[1], Katie Siddle[1], Hélène Quach[1], Christine Harmant[1], Noémie Becker[2], Béatrice Régnault[3], Laure Lemée[3], Jean-Marie Hombert[4], Alain Froment[2], Paul Verdu[5], Luis Barreiro[6], Nathaniel J. Dominy[7], George Perry[8], Evelyne Heyer[2], Lluís Quintana-Murci[1]
[1]Human Evolutionary Genetics, CNRS URA3012, Institut Pasteur, Paris, France, [2]Ecoanthropology and Ethnobiology, CNRS/MNHN/Université P7 UMR 5145, Paris, France, [3]Genotyping Platform, Institut Pasteur, Paris, France, [4]Département de Dynamique du Langage, CNRS UMR 5596, Université Lumière-Lyon 2, Lyon, France, [5]Department of Biology, Stanford University, Stanford, CA, USA, [6]Université de Montréal, Centre de recherche CHU Sainte-Justine, Montréal, Canada, [7]Anthropology department, Dartmouth College, Hanover, NH, USA, [8]Department of Human Genetics, University of Chicago, Chicago, IL, USA

Copy number variations (CNVs), a class of genomic structural variation resulting from the deletion or duplication of genomic DNA segments, is increasingly being recognized as a substantial source of phenotypic variation in human populations. We are interested in studying the extent and effects of CNVs in different African populations, namely the Pygmy populations from central and central-west Africa, which maintain a mostly forest-dwelling hunter-gatherer itinerant lifestyle, and other Bantu-speaking populations, which adopted an agriculture-based sedentary lifestyle ~5,000 years ago. These differences in lifestyles and ecological/environmental pressures experienced by the different populations would have led to different adaptation processes and outcomes, influencing the variability of the human genome. Using the Illumina HumanOmni1-Quad BeadChip genotyping platform, we genotyped about 300 individuals from 8 populations of Pygmy hunter-gatherers and Bantu farmers. CNV detection was performed using various bioinformatic software packages, after which we studied the patterns of CNV diversity among different populations. We found CNVs that displayed significant differentiation in allele distribution between populations, which allowed for the identification of genes and biological pathways that may have experienced differential selective pressures in different populations over recent human evolutionary history.

P-2193

Analyses and computer simulations of hotspot distributions in human genome

Dorota Mackiewicz[1], Paulo Murilo Castro de Oliveira[2], Suzana Moss de Oliveira[2], Stanisław Cebrat[1]
[1]Faculty of Biotechnology, University of Wroclaw, Wroclaw, Poland,[2]Instituto de Física, Universidade Federal Fluminense, Niterói, Brazil

Recombination process is the main cause of genetic diversity and its errors can lead to chromosomal abnormalities. The recombination events are confined to narrow chromosome regions called hotspots in which characteristic DNA motifs were found. We have compared results of analyses of distribution of recombination hotspots and the DNA motifs in human with results of Monte Carlo simulation of genome evolution. DNA motifs along real human chromosomes are distributed unevenly similarly to recombination. Clusters of these motifs roughly follow the distribution of recombination events while single motifs show rather negative correlation with recombination distribution. Computer simulation predicted that selection against defective alleles arisen by mutation were strong enough to distribute recombination hotspots unevenly along virtual chromosomes – higher density in the sub-telomeric regions and lower in the central region of chromosomes mimicking the distribution of recombination events in eukaryotic chromosomes. After simulations long enough, structure of chromosomes reached dynamic equilibrium – numbers and global distribution of hotspots and defective alleles stayed stochastically unchanged, while their precise localization changed. That resemble the dynamic structure of human and chimpanzee genomes where hotspots change there precise localisation leaving the global distribution of recombination events unchanged.

P-2194

Yeast evolution: comparative genomics of 10 Lachancea species

Véronique Sarilar[1], Jean-Philippe Meyniel[2], Cécile Neuvéglise[1]
[1]INRA UMR1319, AgroParisTech, Institut Micalis, Biologie intégrative du métabolisme lipidique microbien, 78850 Thiverval-Grignon, France,[2]ISoft, Bioinformatic Team, 91190 Gif-sur-Yvette, France

Genome sequencing advances in the last few years have made it possible to explore clades distant from model species and to have a broader view of genome evolution. We focus on the *Lachancea* yeast clade, closely related to *Saccharomyces* species, but which diverged before the *Saccharomycetaceae* whole genome duplication. Up to now, 10 *Lachancea* species have been isolated worldwide in association with plants (leaves, roots, tree exudates), soil, or fermentation processes. Three of them have been described in the last three years and one is still not named, we called it *L. fantastica*. Among the 10 genomes, three have been already sequenced and annotated. Obtaining the last 7 genomes has enabled us to perform comparative genomic analysis to explore this recently defined clade.

We first developed a pipeline based on the Amadea software (ISoft, France) to transfer the genome annotation from two reference genomes (*Lachancea kluyveri* and *L. thermotolerans*) annotated by the Génolevures consortium, to the seven new genomes. These annotations were then manually curated. By comparing their protein coding genes, tRNA genes and transposable element content, as well as synteny conservation, we sought to answer several evolutionary questions. Two of them will be discussed here.

Firstly, we investigated genome size evolution. In two groups of 3 species, nuclear genomes were subjected to two independent losses of 0.6 and 1Mbp respectively, for a genome of about 11 Mbp. We aimed at characterizing these losses and their putative impact on metabolic pathways.

Secondly, we investigated if transposable elements, which are known to be involved in chromosome rearrangements, had an impact on the *Lachancea* genome architecture. Notably, the *L. kluyveri* retrotransposon *Tsk1* is suspected to have been acquired by horizontal transfer. We analyzed the sequence evolution of this family of transposable elements by phylogenetic methods, in comparison with other retrotransposons from different *Saccharomycetaceae* species.

This study has to be set in the broader context of yeast evolution including other yeast clades which evolution could be highly different.

P-2195

Estimating the rate of meiotic gene conversion rate in humans

Amy Williams[1,2], Nick Patterson[1,2], Thomas Dyer[3], John Blangero[3], David Reich[1,2]
[1]Harvard Medical School, Boston, MA, USA, [2]Broad Institute of Harvard and MIT, Cambridge, MA, USA, [3]Texas Biomedical Research Institute, San Antonio, Texas, USA

Meiotic gene conversion is a process whereby short segments of DNA are replaced by the homologous chromosome sequence during gamete formation and is related to meiotic crossovers, with a fundamental difference being the size of homologous sequence incorporated into the resulting chromosome. Whereas crossovers introduce homologous sequence spanning several megabases, gene conversion tracts are estimated to span between 20-1000 bp. Gene conversions affect sequence variation by breaking down allelic association within a localized region and can have a substantial impact on haplotype diversity. To date, there has been no genome-wide study of the rate of de novo gene conversion, little is known about gene conversion tract lengths, and most models of haplotype variation do not explicitly account for gene conversion.

We obtained a preliminary estimate of the rate of gene conversions at 7.91×10^{-6} per bp per generation using 14 three-generation families genotyped with Illumina arrays. This estimate includes information from a total of 38 meioses (19 paternal, 19 maternal) at a total of over 3.79×10^{6} informative SNPs. We phased these families to obtain minimum-recombinant haplotype transmissions from parents to offspring and located de novo gene conversions, identifiable as single-SNP double recombinations in the resulting haplotypes. Because genotypes are subject to errors, we verified each putative gene conversion by ensuring that the gene converted allele is transmitted to the third generation. We verified the genotype of the first generation individual that transmitted the gene conversion by ensuring that this individual transmits its alleles to separate second generation offspring in the expected manner, as dictated by the haplotype segregations at surrounding loci. This design ensures that the gene conversions we identified are either real or the result of at least two genotyping errors at the same site. These results are preliminary, and we are planning to carry out further work in an additional 1679 genotyped individuals and to confirm the discovered gene conversions by re-genotyping using a different genotyping technology.

We have obtained whole genome sequence for 568 of the genotyped samples and will use these data to identify gene conversion events. These additional data provide the opportunity to (a) identify a larger number of gene conversion events and thus obtain a more precise estimate of the gene conversion rate, (b) estimate gene conversion tract lengths in regions where SNP variants are occur at a high density, and (c) analyze the extent to which biased gene conversion occurs.

P-2196

Expanded phylogenetic analysis of the telomere-associated transposable element, *HeT-A,* in *Drosophila.*

Jonathan Clark, Haylie Cox
Weber State University, Ogden, UT, USA

In most eukaryotes, telomeres are formed by a short nucleotide sequence that is repeated many times at the chromosome end. In contrast, the chromosome ends of *Drosophila* consist of at least two different transposable elements, *HeT-A* and *TART*, which are tandemly arrayed in multiple copies. In *D. melanogaster*, these transposable elements are confined to the ends of chromosomes and are not found at any other sites in the genome. This is the most striking example of a eukaryotic transposable element that performs an essential cellular function. A molecular study has been initiated that examines the phylogeny of the *HeT-A* transposable element among eight species within the *melanogaster* species subgroup, which includes *D. melanogaster*. Multiple *HeT-A* sequences were obtained from each species and these sequences are compared to an expanded dataset of *HeT-A* sequences available from the *Drosophila* genome projects. The phylogeny of the *HeT-A* sequences is compared to the phylogeny of the host species, determined by *ADH* gene sequences. For some comparisons, the extent of *HeT-A* nucleotide divergence exceeds 50%. The phylogeny reveals that multiple sequences from each species are not always monophyletic. This suggests that multiple subfamilies, each with their own evolutionary history, exist in all genomes examined. Alternative explanations, including lateral transfer of *HeT-A* elements between species, are discussed. Additional comparisons of the rate of synonymous and nonsynonymous nucleotide substitutions suggest that there is no selection operating on the *HeT-A* coding region, surprising finding given the importance of telomeres for cellular stability.

P-2197

The Plant Proteome Folding Project : Structure and Positive Selection in Plant Protein Families

Melissa Pentony[1], Patrick Winters[1], Duncan Penfold-Brown[1], Kevin Drew[1], Apurva Narechania[2], Rob DeSalle[2], Richard Bonneau[1], Michael Purugganan[1]
[1]New York University, New York, New York, USA, [2]American Museum of Natural History, New York,New York, USA

Despite its importance, relatively little is known about the relationship between the structure, function and evolution of proteins, particularly in land plant species. We have developed a database with predicted protein domains for five plant proteomes (http://pfp.bio.nyu.edu), and used both protein structural fold-recognition and *de novo* Rosetta-based protein structure prediction to predict protein structure for Arabidopsis and rice proteins. Based on sequence similarity, we have identified ~15,000 orthologous/paralogous protein family clusters among these species, and used codon-based models to predict positive selection in protein evolution within 175 of these sequence clusters. Our results show that codons that display positive selection appears to be less frequent in helical and strand regions, and are over-represented in amino acid residues that are associated with a change in protein secondary structure. Like in other organisms, disordered protein regions also appear to have more selected sites. Structural information provides new functional insights into specific plant proteins and allows us to map positively selected amino acid sites onto protein structures and view these sites in a structural and functional context.

P-2198

Relative rates of evolution among the three genetic compartments of the red alga *Porphyra* differ from those of land plants and green algae

David Smith[1], Jimeng Hua[2], Robert Lee[2], Patrick Keeling[1]
[1]*University of British Columbia, Vancouver, Canada,* [2]*Dalhousie University, Halifax, Canada*

Data on relative silent-site nucleotide substitution rates (which can be used to approximate relative mutation rates) among mitochondrial, plastid, and nuclear DNAs (mtDNAs, ptDNA, and nucDNAs) are restricted to green plants. In green algae, all three genetic compartments appear to have similar mutation rates, whereas in most seed plants the average mutation-rate estimates of mtDNA vs ptDNA vs nucDNA are 1:3:10. Here we investigate relative rates of evolution within the model red algal genus *Porphyra*. Using complete organelle genome sequences and various nuclear loci from *Porphyra purpurea* and *Porphyra umbilicalis*, we find that the relative levels of silent-site divergence for the mtDNA, ptDNA, and nucDNA are approximately 1:0.3:0.2. This implies that for *Porphyra*, the mitochondrial mutation rate is between 3-5-times greater than that of the plastid and nucleus, which is opposite to the trend in most seed plants. These data are discussed in context to hypotheses on genome evolution as well as the horizontal transfer of red plastids to other eukaryotic lineages.

P-2199

Genomic Modification in Bdelloid Rotifers Following Full-Body Desiccation and DNA Repair

Norian Caporale-Berkowitz[1], Bette Hecox-Lea[2,3], David Mark Welch[1,2]
[1]Brown University, Providence, RI, USA, [2]Marine Biological Laboratory, Woods Hole, MA, USA, [3]Northeastern University, Boston, MA, USA

Bdelloid rotifers are a strictly asexual clade of freshwater microinvertebrates whose unique 40-million year history of chastity seems to pose an exception to the general rule that the absence of sex is a dead end in evolution. Bdelloids are also distinctive in their ability to recover from full-body desiccation at any life stage. The genomes of bdelloids bear the hallmarks of extensive double-strand break repair in the form of gene conversion tracks consistent with synthesis-dependent strand annealing and microhomology consistent with non-homologous end joining. Our research probes the link between desiccation and DNA repair by examining transcriptomes of clonal lineages taken through multiple rounds of desiccation and rehydration. By measuring the loss of heterozygosity in post-desiccation transcriptomes and mapping individual transcripts to reference genomes to locate breakpoints of DNA repair we will characterize the amount and type of DNA repair that occurs from desiccation. A central question is whether the type and extent of repair effects such as gene conversion could be sufficient to act as a surrogate for sexual reproduction. This would account for the bdelloids' striking evolutionary success in the absence of sex and suggest an alternate strategy for creating the genetic variation that acts as the substrate for natural selection.

P-2200

Global characterization of polymorphic inversions in the human genome

Mario Cáceres[1,2], Alexander Martínez-Fundichely[2], Cristina Aguado[2], David Vicente[2], Meritxell Oliva[2], David Izquierdo[2], Sergi Villatoro[2], Marta Puig[2], José Ignacio Lucas-Lledó[2], Xavier Estivill[3], Juan R. González[4], Sònia Casillas[2]
[1]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [2]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain, [3]Center for Genomic Regulation (CRG-UPF), Barcelona, Spain, [4]Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

Inversions were one of the first types of genomic rearrangements studied and for a long time have been a paradigm in evolutionary biology. Recent advances in genomic techniques have spurred a renovated interest in all kinds of structural variants, especially in humans, and have revealed an extraordinary degree of structural variation. However, due to the difficulty of their study, inversions are still relatively overlooked. Here we present the current progress of our work as part of a larger project towards the complete characterization of inversions in the human genome. First, we have developed a new algorithm and analysis strategy to identify and map as accurately as possible inversion breakpoints based on available paired-end mapping data. This has allowed us to determine that a big fraction of recently defined inversions are false positives and to identify the main causes of these errors, ranging from problems in the genome assembly to sequence differences between individuals. Second, by reanalyzing the data and merging the inversions predicted from different published studies, we have generated the most reliable catalogue of human polymorphic inversions, comprising a few hundred non-redundant inversions. Third, by using different PCR techniques we have experimentally validated more than 30 of the inversions and genotyped 20 of them in multiple HapMap individuals from different populations, with frequencies between 1% and 60%. Finally, the use of an accurate inversion set, not obscured by a high error rate, has made possible to uncover the main characteristics of inversion polymorphism in humans, including its genomic distribution, length and generation mechanisms. Moreover, we can start making valid predictions on the effect of the inversions on nearby genes, and both cases of gene breakage and exchange of parts of genes have been observed. Thus, these results represent an important first step towards determining the evolutionary and functional impact of inversions in the human genome.

P-2201

Indels and the evolution of genome architecture: The long and short of it

Erika Kvikstad, Laurent Duret
University Lyon 1, Lyon, France

Elucidating the mechanisms of mutation accumulation and fixation is critical to understand the nature of genetic variation and its contribution to genome evolution. For example, a well-known feature of mammalian genomes is the strong correlation between GC content and gene density: GC-rich regions of the genome contain more genes, which are in turn more compact. Changes in the amount of genomic DNA via insertions and deletions (indels) of multiple bases are presumed to contribute to this pattern, although this hypothesis has not been directly confirmed. Here we take advantage of recent population-scaled sequencing efforts, together with multiple primate whole-genome sequence alignments, to test hypotheses concerning the physical and temporal scales at which indels have played a role in establishing genome architecture.

To ascertain if there is indeed a bias in indel patterns that correlates with the local genomic base composition, we first investigated short (1-50 base pairs) indel variation identified in the 1000 Genomes Pilot 1 data. Interestingly, we find that indel rates are highly dependent on the allele nature in terms of event type (insertion/deletion), length, and context. Indeed, we observe a strong deletion to insertion bias that varies from ~1x to 5x depending on the inserted/deleted nucleotides. After factoring for this rate heterogeneity, we next performed analyses of the site frequency spectra of human polymorphic indel alleles, in conjunction with comparisons of indel intra-species (diversity) vs. inter-species (divergence) to reveal potential fixation biases and the relative contributions of selection and biased gene conversion to indel evolution. Additionally, we developed a protocol to infer long (>50 base pairs) indel events from analysis of changes in lengths of orthologous introns between closely related primate species. We identified approximately 7,000 potential indel-containing introns since human divergence from the common ancestor with chimpanzee. Of these introns, 1963 and 532 were inferred to evolve due to a single long human-specific expansion or contraction, respectively.

We further analyzed the genomic distribution of these indels to determine the relative contributions of local GC content, crossover rate, and presence of constrained sequences to the evolution of intron size. The trends observed here will thus allow us to infer the forces contributing to genome-wide variation in indel rates and patterns, and reveal the extent to which small and large indels have shaped the evolution of mammalian genome architecture.

P-2202

Copy Number Variation in the Porcine Genome

Yogesh Paudel, Hendrik-Jan Megens, Ole Madsen, Mirte Bosse, Laurent Frantz, Richard Croojimans, Martien A.M. Groenen
Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands

One of the most astonishing results from the recent genomic studies is the importance of structural variations (SV) in mammalian genome . However, little is known about the phenotypical consequences of SVs. Structural variation, including copy number variation (CNV) of genomic regions often containing complete genes, can result in alteration of gene expression and phenotypic variation. The recent completion of a draft genome of the pig (Sus scrofa) is a landmark for genetic studies of this important agricultural species. The large number of domesticated varieties across the world, and its use as a biomedical model, makes Sus scrofa a candidate to study the evolution of SV and its phenotype. Since its domestication (~10,000 years ago) pigs have been selected for different phenotypical trait across the world. Due to these different selective regime we expect to find differences in structural variation among breeds and wild animals, that reflects biogeography, domestication and selection.
We have studied copy number variation based on whole-genome re-sequencing data from 50 individuals representing a variety of domestic and wild population from different parts of the world. Copy number variation between and within breeds was examined by using a read-density based approach, where number of copies present is inferred from sequenced depth of whole genome sequence data. In total, we identified around 20,000 CNVs that are at least 4kb in length, of which ~5000 copy number gains, per individual with an average size of 6 kb. In order to understand the evolutionary dynamics of SVs in both domestication and adaptation process, we compared the genomic location and intensities of those CNVs between domestic and wild individuals. Around 500 genes were found to overlap with inferred CNVs. Functional analysis reveals that most of the CNV regions are enriched in gene related to sensory perception (p-value <0.001), neurological process (p-value <0.001) and response to stimulus (p-value <0.001). We found that CNVs common in the wild population were overlapping with genes related to immune system whereas CNVs common in domestic populations were overlapping with genes related to metabolism as well as immune system. This difference in CNV gene content between wild and domesticated animals show that CNV were probably the target of selection by breeders. Further studies will show how widespread this process is and shed light on the importance of CNVs in domestication and artificial selection.

P-2203

Structured variations in intron/exon architecture according to gene intron number in two angiosperm species

Adrienne Ressayre[1,2], Pierre Montalent[1,2], Christine Dillmann[2], Johann Joets[1,2]
[1]INRA, Gif sur Yvette, France, [2]Université Paris-Sud, Orsay, France

Rice and Arabidopsis genomes are composed of several tens of thousands of protein coding genes. These genes are often interrupted by one or several introns. Introns are non-coding regions removed after transcription to produce mature mRNAs. In order to document the intra-genome dynamics in both species, we considered the subsample of genes displaying evidences for expression.

We first studied the distribution of intron number of the genes in the two species and classified the genes according to their number of introns. In both species, the distribution of gene number per classes of intron number is remarkably similar. It follows a geometric distribution, compatible with the hypothesis that most of the intron dynamics could be due to a stochastic intron birth and death process.

Second, we investigate the existence of patterns for organization of the exon-intron structure according to the degree of gene segmentation (i.e. intron number). More precisely, we compute the average length and GC content for each of the elements composing a gene (exons and introns) and per position of the element along the gene. Again we find clear patterns that are strickingly regular in both species. Thoses patterns appear to be shaped by (i) intron/exon structure, (ii) the position along the gene and (iii) the number of intron of the gene. These results suggest that at least part of the constraints on genome organization take place at the gene level.

P-2204

The *Drosophila buzzatii* Genome Project

Yolanda Guillén[1], Nuria Rius[1], Alejandra Delprat[1], Marta Puig[1], Alfredo Ruiz[1], Sònia Casillas[2], Miquel Ràmia[2], Antonio Barbadilla[2], Gisela Mir[3], Jordi Garcia-Mas[3], Ivo Gut[4], Jordi Camps[4], David Torrens[5], Valentí Moncunill[5], Andrew G. Clark[6], Robert L. Unckless[6], Cedric Feschotte[7], Aurelie Kapusta[7]
[1]Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain, [2]Plataforma Bioinformàtica de la UAB, Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain, [3]Centre for Research in Agricultural Genomics (CRAG), Campus UAB, Edifici CRAG, 08193 Bellaterra (Barcelona), Spain, [4]Parc Científic de Barcelona, Centro Nacional de Análisis Genómico (CNAG), Torre I, Baldiri Reixac 4, 08028 Barcelona, Spain, [5]Barcelona Supercomputing Center (BSC), Edifici TG (Torre Girona), Jordi Girona 31, 08034 Barcelona, Spain, [6]Department of Molecular Biology and Genetics, Cornell University, Ithaca (New York), USA, [7]Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA

The *Drosophila* genus is large and diverse with about 2,000 known species. Phylogenetic analyses indicate that two main lineages exist, which diverged 40-60 myr ago. One lineage led to the *Sophophora* subgenus comprising some 300 species (one of them being *D. melanogaster*), whereas the other one led to the subgenus *Drosophila*, with about 1,700 species (including the Hawaiian *Drosophila*). Out of the 20 *Drosophila* genomes already sequenced and available in FlyBase, only three belong to the *Drosophila* subgenus species: *D. virilis*, *D. mojavensis* and *D. grimshawi*. The remaining seventeen species belong to the *Sophophora* subgenus.
Here, we present our collective effort to sequence and annotate the genome of *D. buzzatii*, a community resource for comparative genomics within the *Drosophila* subgenus. *D. buzzatii* belongs to the *repleta* species group of the *Drosophila* subgenus. This species group includes >100 species, many of them cactophilic species living in the deserts and arid zones of the American continent, and has been used for studies of ecological adaptation and speciation for over sixty years. So far only one other member of the *repleta* group, *D. mojavensis*, has a sequenced genome. We built a genomic library with ~18,000 BAC clones with an average insert size of ~150 kb (an ~18× expected representation of the *D. buzzatii* euchromatic genome) and generated a BAC-based physical map covering ~90% of *D. buzzatii* chromosomes. Five of these BAC clones have been entirely sequenced using Sanger technology (total ~800 kb). We also sequenced the ends of 1,152 BAC clones from chromosome 2 to generate > 2000 BAC end-sequences (BES). We used the 454/Roche sequencing platform to generate 1.7 Gb of shotgun reads plus 788 Mb of paired reads (6-8 kb fragments). We also used Illumina/Solexa platform to generate 28.8 Gb of short fragment (500 bp) paired reads and 4.5 Gb of long fragment (8 kb) paired reads. Sequences from different sources have been assembled using Newbler, CABOG and SOAPdenovo. Finally, we have also produced 15 Gb of RNA sequences from five life- stages (embryos, larvae, pupae, adult males and adult females) in order to help in gene annotation. Protein-coding genes have been annotated using a strategy that combines not only ab initio predictions (N-SCAN, GENSCAN) but also annotation by homology (Exonerate, Blast, Tophat and Cufflinks), whereas transposable elements have been identified, annotated and masked using RepeatScout, REPCLASS, RepeatModeler and RepeatMasker.

P-2205

Simulating evolutionary scenarios to test whether they can induce reductive evolution

Berenice Batut[1,2], Mathilde Dumond[1], Gabriel Marais[2], Guillaume Beslon[1], Carole Knibbe[1]
[1]*INRIA Beagle Team, Lyon, France,* [2]*BGE Team (LBBE), Lyon, France*

Some bacterial lineages seem to have undergone significant genome shrinkage over the last millions of years, a process called reductive evolution. For example, the genome of *Buchnera aphidicola* is only one-seventh the size of the genome of its close relative *Escherichia coli* (Moran and Mira, Genome Biol. 2001). This reductive evolution was initially thought to be a signature of intracellular lifestyle. Thus, explanatory mechanisms related to this lifestyle were proposed, like a smaller effective population size, a lower exposure to horizontal transfer, or the uselessness of some genes. However, reduced genomes were also found in some free-living cyanobacteria like *Prochlorococcus marinus* (Rocap et al., Nature 2003) or *Pelagibacter ubique* (Giovannoni et al., Science 2005). This questions the evolutionary mechanisms underlying reductive evolution: Are they shared by endosymbionts and cyanobacteria?

Predicting the effect of these mechanisms would shed light on the minimal combination of factors required to observe reductive evolution. We performed *in silico* evolutionary experiments to test the effect of a smaller population size, a higher mutation rate, a less demanding environment, and a less varying environment. We used the individual-based model *aevol* (Knibbe et al., Mol. Biol. Evol. 2007), where virtual genomes undergo small-scale mutations, rearrangements, and reproduce differentially based on their performance at a curve-fitting task. Experiments began with ancestral populations that evolved with standard parameters during 150,000 generations. Then the value of the tested parameter was changed and evolution continued for 150,000 additional generations.

These simulations showed that a higher mutation rate, as well as a less demanding environment, led to both fewer genes and shorter intergenic sequences. Environmental stabilization triggered reduction in non-coding sequences but not in gene number. Finally, with a smaller population size, some genes were lost but the genome expanded because intergenic sequences lengthened, even with an enforced deletional bias. To explain this surprising effect, one hypothesis is that the sudden reduction in population size leaves a less fit population. The selection, now directional, indirectly favors the most evolvable genomes, that is, genomes with longer non-coding sequences undergoing larger rearrangements.

This study is thus a first step towards predicting the genome structure expected for a given evolutionary scenario. The next steps are to (i) further analyze the evolutionary trajectories in these experiments, (ii) test the effect of combined scenarios with multiple changed parameters, and (iii) find the best scenarios for intracellular and free-living species with reduced genomes.

P-2206

Evolutionary Genomics of The Hedgehog Gene Family in Metazoans: Identification of The Desert Hedgehog Gene on Avian Genomes

Joana Pereira[1], Warren E. Johnson[2], Stephen J. O'Brien[2], Vitor Vasconcelos[1,3], Agostinho Antunes[1,2]
[1]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal, [2]Laboratory of Genomic Diversity, National Cancer Institute, Frederick, USA, [3]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal

The Hedgehog (Hh) gene family codifies a class of secreted proteins composed of two active domains that act as signalling molecules during embryo development (e.g. development of the nervous and skeletal systems and the formation of the testis cord). The vertebrate Hh genes evolved by ancient duplications of the ancestral invertebrate Hh gene: while only one Hh gene is found typically in invertebrate genomes, vertebrates have three, each with different functions in distinct tissues (Sonic hedgehog – Shh; Indian hedgehog – Ihh; and Desert hedgehog – Dhh), which likely favoured the increased complexity of vertebrates and their successful diversification. However, when the vertebrate members of the Hh genes family are browsed over the avian genomes available to date on the GenBank and Ensemble databases no evidence of the Dhh gene is found.

In this study, we used detailed comparative genomic and synteny analyses to characterize the Hh family on avian genomes and understand why no Dhh gene is found in the bird genome assemblies available to date. The analysis of the Dhh gene synteny on the lizard genome using the Genomicus 64.01 server shows that this gene forms a conserved cluster with the LMBR1L and RHEBL1 genes, which are also not found on the current avian genomes assemblies. When the scaffold where these three genes are found on the lizard genome is compared with the avian genomes, we find that it has strong homologies with varied regions of the avian ChUn_random sequence, suggesting the putative presence of the Dhh, LMBR1L and RHEBL1 genes on avian genomes, which are likely located on regions difficult to sequencing.

P-2207

Nucleomorph genome sequence of the cryptophyte alga *Chroomonas mesostigmatica* reveals lineage-specific gene loss and genome complexity.

Christa Moore, John Archibald
Dalhousie University, Halifax, Nova Scotia, Canada

The process of endosymbiosis has had a profound impact on the evolution of the eukaryotic cell. The engulfment of a cyanobacterium by a non-photosynthetic eukaryote resulted in the evolution of the plastid, a landmark event that paved the way for the direct and indirect spread of photosynthesis across the eukaryotic tree. Cryptophytes and chlorarachniophytes are two unrelated eukaryotes that acquired their plastids secondarily by assimilating red and green algal endosymbionts, respectively. Unlike all other secondary plastid-bearing algae, cryptophytes and chlorarachniophytes retain the nuclei of their algal endosymbionts in a miniaturized form. These relict nuclei - 'nucleomorphs' - harbor the smallest nuclear genomes known (~300 to 1,000 Kbp in size), and are useful models for studying the architecture and evolution of reduced genomes in both free-living and parasitic organisms. To gain insight into nucleomorph genome diversity within cryptophytes, we used Illumina and 454 'next-generation' technologies to completely sequence the 702.9 Kbp nucleomorph genome of *Chroomonas mesostigmatica* CCMP1168. The genome possesses several features found in other nucleomorph genomes, including a three-chromosome architecture, low G-C content (~26%), sub-telomeric rDNA operons, few introns, and a high gene density (505 protein-coding genes). Unsurprisingly, 'house-keeping' genes predominate. However, the *C. mesostigmatica* nucleomorph genome is unusual in being the first to harbor repetitive elements and numerous multi-copy genes. A four-way comparison of cryptophyte nucleomorph genomes provides insight into the authenticity of nucleomorph 'ORFan' genes, hypothetical protein-coding genes with no homologs in current databases, which constitute up to 23% of the genes in these reduced genomes. While homology is undetectable at the sequence level, protein size and gene order conservation across the cryptophyte nucleomorph genomes suggest that many of these ORFan genes are in fact homologous. Transcriptome data from *C. mesostigmatica* confirms that these genes are expressed. Proteosomal subunit genes, which are abundant in all other cryptophyte nucleomorph genomes, are completely absent in *C. mesostigmatica*, suggesting recent transfer to the host cell nucleus or outright loss. Differences in the degree and patterns of synteny are observed and suggest a recombination-mediated mechanism for gene loss. The absence of different gene families in the nucleomorph genomes of different cryptophyte lineages examined suggests that gene loss may still be occurring, and that nucleomorphs have yet to reach an endpoint to their reduction.

P-2208

Evolution of the 20-Gb genome of Norway Spruce (*Picea abies*)

Douglas Scofield[1], The Norway Spruce Genome Project[1,2]
[1]*Umeå Plant Science Centre, Umeå, Sweden,* [2]*SciLifeLab, Stockholm, Sweden*

Using a variety of next.-generation sequencing (NGS) technologies, we have sequenced the nuclear and organellar genomes and transcriptome of Norway Spruce. Traits shared with other members of the pine family Pinaceae, including an extremely large nuclear genome (~100x Arabidopsis) with low gene content (~0.1%), presented particular challenges for assembly and annotation.

We sequenced genomic libraries prepared from diploid leaf tissue, haploid megagametophytes extracted from seed, and ~1x coverage in haploid fosmid pools prepared from diploid leaf tissue. Libraries included 180-, 300- and 650-bp paired-end preparations along with ~2Kb mate pairs and ~40Kb fosmid ends. We assembled contigs using CLC Genome Analyzer, and scaffolded and merged genome assemblies using existing and custom software tools. The current set of scaffolds sums to ~75% of total genome length. For comparative analyses, we also shallow-sequenced 5 other gymnosperm species at 5-20x coverage. Heterozygosity is pervasive and, in assemblies based on diploid libraries, has resulted in haplotypes being split into separate contigs at many locations throughout the genome.

The mitochondrial genome has increased greatly in size (to ~5 Mb), with large intergenic regions containing complex sequence features. The chloroplast genome is 122Kb and highly conserved in comparison to other gymnosperms with respect to gene composition and gene order, with the major exception of some structural variation around the inverted repeat.

Gene content is indeed low, with large regions of assembled sequence completely lacking genes. In other regions, we are able to identify a large number of intact genes, and mapped nearly 9,000 full-length cDNAs from white spruce (*P. glauca*) to single contigs. Within genes, many introns are larger than introns in homologous genes in angiosperms (1Kb in length or greater).

We examined repeat content, including scans for known mobile element families and novel repeats. Analyses to date indicate 77% of the genome is contained in repetitive sequences, with 41% in LTR retrotransposons (primarily Copia and Gypsy elements) and fully 34% in unclassified repeats. There were at least two bursts of mobile element activity in the distant past, including many insertions that predate the divergence of *Picea* from *Pinus* (~90MYA), but there is no evidence of significant numbers of insertions within the last 6MY. Paradoxically, there is little evidence for loss of repeat elements, which may help explain the extremely large genome size of this and other gymnosperms.

P-2209

A global view of mutation patterns across eukaryotes.

Jean-François GOÛT, Michael LYNCH
Lynch Lab, Indiana University, Bloomington, IN, USA

All genomes evolve as a consequence of new variation caused by mutations and subsequent modification of allele frequency through selection, recombination and drift. Therefore, characterizing the mutation patterns of living organisms is important for understanding the evolutionary processes that have shaped genomes.

Two main trends emerge from our current knowledge of mutation patterns. First, there seems to be a universal bias for mutations from Gs and Cs towards As and Ts that is often countered by selection in favor of Gs and Cs. Second, relative to transversions, transitions are more frequent than expected, which has strong implications for the models of evolution commonly used. However, this view of mutational patterns comes from only a handful of well-studied organisms and needs to be expanded to a more diverse range of eukaryotes.

Mutation patterns are usually obtained by one of three strategies: reporter genes, mutation accumulation experiments or using regions of genomes that are not under selective constraint. Reporter genes have been widely used in the past but are limited to a small fraction of the genome and therefore can give biased results. With the drop in sequencing costs, mutation accumulation experiments have become a powerful way to investigate mutation patterns. However, the experiments themselves remain labor intensive and costly. Using regions of the genome that are not under selection provides the advantages of a genome wide view (as long as these regions are not limited to small parts of the genome) and does not require long and expensive experiments. Such regions can be found in pseudogenes: dead copies of functional genes. Mutation patterns can be inferred simply by comparing the sequences of functional genes to their corresponding pseudogenes that are relieved from selection and are therefore free to accumulate mutations.

In this study, we have used an annotation pipeline (PseudoPipe, Zhang et al. (2006) Bioinformatics 22:1437-9) to find pseudogenes in 150 fully sequenced eukaryotic genomes. We then used this dataset of over a million pseudogenes to obtain mutation patterns across a wide range of eukaryotes. For species with previously published mutation patterns, we compare our results to those obtained by different methods. By extending this analysis to many new species, we provide the first global view of mutation patterns across the tree of eukaryotes and gain insight on the evolutionary processes that are shaping the genomes of eukaryotes.

P-2210

Large variations in the rate of DNA loss among vertebrates

Stephane Boissinot
Queens College, the City University of New York, Flushing, NY, USA

Genome size is an extremely variable trait among eukaryotes. Understanding the causes of this variation has been the subject of much research in the past two decades yet it remains an unresolved and controversial issue. The size of a given genome results from two opposite processes: the addition of DNA either by transposition or segmental duplication, and the removal of DNA by deletion. If the factors affecting the increase of genome size have been investigated in detail, the factors affecting the rate of DNA loss have not received the same level of attention. By examining the decay of transposons, we found that large (>500bp) deletions seem to occur in vertebrates much more often than previously thought. In particular, large deletions seem to occur more frequently in the genome of non-mammalian vertebrates, such as fish and reptiles, relative to mammalian genomes. We demonstrate that variations in the rate of DNA loss through large deletions are largely sufficient to account for genome size differences among vertebrates.

P-2211

Influence of mating system on genome evolution in Caenorhabditis

Janna Fierst, John Willis, Timothy Ahearne, Rose Reynolds, Kristin Sikkink, William Cresko, Patrick Phillips
University of Oregon, Eugene, OR, USA

Self fertilization is predicted to reduce levels of standing genetic variation at neutral loci by one half. More importantly, however, self fertilization reduces the effective recombination rate, which should dramatically increase the opportunity for linkage disequilibrium, thus making the genomes from self-fertilizing species more susceptible to selective sweeps and/or background selection. These effects in turn should have important consequences for distribution of genomic variation and the evolution of genomic structure. Nematodes of the genus Caenorhabditis display mating system variation in which selfing hermaphrodites have evolved independently multiple times from an outcrossing male-female ancestor. Using a combination of next generation sequencing and genetic mapping approaches, we have assembled the genome of *C. remanei*, a close outcrossing relative of the selfing *C. elegans* model system. We can compare these genomes, together with that from *C. briggsae*, which has also independently evolved hermaphrodites, to gain insights into how the evolution of self fertilization influences genome structure, including the evolution of gene families and the evolution of genome size.

P-2212

Genomic response to drastic environmental change

Thomas Cuypers, Paulien Hogeweg
Utrecht University, Utrecht, The Netherlands

Background: Phylognetic analysis suggests that whole genome duplication events may have enhanced evolvability during drastic changes in the environment (Fawcett et al., 2009; Mayrose 2011). Here we take a modeling approach to overcome the inherent difficulty in using phylogenetic method to identify a potential causal relationship and to get a deeper understanding of genome evolution in the face of environmental change.

Previously, we studied genome evolution in a model of virtual cells that evolve homeostasis in a fluctuating environment (Cuypers, 2012). We found a pattern of rapid genome expansion in early evolution. Here we extend our model to study how evolved, fit cells can readapt to novel circumstances. As a proxy for major environmental change, we make parameter shifts on several axes that were not experienced by cells before, such as the level of membrane permeability, the yield in energy from resource, and system wide protein degradation as well as the homeostatic setpoint that cells have to reach.

Results: The gene regulatory networks of evolved cells are tuned to respond to a range of environmental resource concentrations to maintain homeostasis. Given this highly tuned state of the network it was surprising to find that lineages were able to rapidly readapt to an imposed extensive change in environmental conditions.
This readaptation appears however to be 'easier' in large genomes, i.e. genomes which have not yet undergone streamlining in long term evolution is the same environment.
Two main modes of readaptation are generally observed. Depending on the extent of the shift, cells may adapt using mostly point mutations while preserving genome structure and size. However, more severe types of change are often associated with whole genome duplications. We observe fixation of whole genome duplication to occur almost exclusively at the start of the simulation and around the point of drastic environmental change.

Conclusion: Our simulations show that readaptation of fit lineages to drastic change can be surprisingly quick. Moreover, mimicking patterns observed in phylogenetic studies, whole genome duplications, although ongoing in terms of mutational events only go to fixation in the line of descent during the first stages of adaption towards novel conditions.

P-2213

Comparative proteomic characterization of protein disorder distribution across Eukaryota

Catherine Mooney, Denis C. Shields, Therese A. Holton
UCD, Dublin, Ireland

Many proteins, or protein regions, do not adopt three dimensional structures, instead remaining disordered or unstructured. There has been a great increase in interest in these proteins over recent years. With more than 150 fully sequenced genomes of Eukaryota available for analysis we now have an opportunity to analyse these proteins, their distribution across species and their evolution.

There is evidence to support an increase in the number of disordered proteins, or the number of disordered protein regions, in a proteome, from prokaryotes to Eukaryotes. However, it is often suggested that there is an increase in disorder with increased organism complexity (Tompa, 2003), but there appears to be little, if any, evidence to support this in Eukaryota. In fact, a recent study by Schad et al. (2011) found no proteome-wide correlation between protein disorder and organism complexity. We explore the correlation, if any, between genome properties such as organism complexity, predicted percentage disorder and proteome size.

Recently, disorder has been observed to fall into three distinct classes: conserved-flexible disorder, conserved-constrained disorder and non-conserved disorder (Bellay, 2011). We examine the distribution of these three classes of disorder across the known supergroups of Eukaryota, and the relationship with organism complexity. We also investigate if there is any evidence of correlation between the evolution of disorder in apicomplexa, parasites which are known to be enriched in intrinsically unstructured proteins (Feng, 2006), and their mammalian hosts.

Results:

Our initial analysis of 42 proteomes selected from Excavata, Unikonts, Plante and Chromalveolates supergroups, covers most of the eukaryote complexity. We found a correlation between the percentage of predicted disorder in a proteome and the percentage GC content of the genome (r=0.54) and the average length of protein sequences in the proteome (r=0.69). We found a weak correlation with the number of residues in the proteome (r=0.30), the number of sequences in the proteome (r=0.15) and the organism complexity (r=0.23). We found that there is a correlation between organism complexity and the number of residues in a proteome (r=0.69).

Bellay, J. et. al. (2011) Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biology.

Feng, Z.P. et. al. (2006) Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. Molecular and biochemical parasitology.

Schad, E. et. al. (2011) The relationship between proteome size, structural disorder and organism complexity. Genome Biology.

Tompa, P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. Bioessays.

P-2214

Testing the “suppressed recombination” model of chromosomal evolution during human-chimpanzee speciation.

Marta Farré[1], Aurora Ruiz-Herrera[1,2]
[1]Departament de Biologia Cel•lular, Fisiologia i Immunologia, Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès, Barcelona, Spain, [2]Institut de Biotecnologia i Biomedicina (IBB), Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès, Barcelona, Spain

Reorganization (shuffling) of the genomic landscape plays an important role in the evolutionary processes as well as in the development of inherited diseases and carcinogenesis. Traditionally, it has been argued that chromosomal reorganizations may contribute to speciation due to the underdominant fitness effects associated with meiotic abnormalities, but this model has important limitations and is difficult to test in natural populations. More recently it has been proposed that chromosomal rearrangements could reduce gene flow and potentially contribute to speciation by the suppression of recombination. According to this "suppressed recombination" model, chromosome rearrangements could have a minimal influence on fitness, but would suppress recombination leading to the reduction of gene flow across genomic regions and to the accumulation of genetic incompatibilities. However, few empirical data are available that address the mechanisms by which new chromosomal variants are fixed in populations of mammalian species and how recombination influences chromosomal speciation and vice versa. Here we test whether the suppression recombination model of chromosomal evolution can be applied to the human-chimpanzee speciation event. To this end, we have constructed a high-refined map of the reorganizations and evolutionary breakpoint regions between human and chimpanzee and analyzed them in relation to a high-resolution genome-wide map of recombination rates in the human genome. Our data suggest the existence of a reduction in the recombination within genomic regions that have been implicated in the chromosomal evolution between human and chimpanzee. We propose a model where inversions might have persisted in the heterozygous state long enough for the impact on recombination rates still be detected in the human genome.

P-2215

High variation of mitochondrial gene order in basidiomycetes

Gabriela Aguileta[1], Elsa Petit[2], Tatiana Giraud[3], Toni Gabaldon[1], Michael Hood[2]
[1]*Centre for Genomic Regulation (CRG), Barcelona, Spain,* [2]*Department of Biology, Amherst College, Amherst, USA,* [3]*Ecologie, Systematique et Evolution, Universite Paris Sud, Orsay, France*

Mitochondrial inheritance provides extraordinary examples of genome evolution. From their alpha proteobacterial origin to extant cellular organelles, an amazing genomic reorganization has taken place in the mitochondria of different groups of organisms. Mitochondria have been studied mostly in animals and plants but fungi provide new opportunities to study highly diverse mitochondrial genomes . To date, most studies have focused on ascomycete mitochondria but we now have several complete basidiomycete mitochondrial genomes. Here we take advantage of these data to investigate the evolution of gene content and gene order. We also look for evidence of recombination and the dynamics of mobile elements, more specifically the distribution of homing endonucleases. We focus on the comparison of previously unpublished mitochondrial genomes of two *Microbotryum* species that show striking differenes in gene order despite their recent divergence. All the species analyzed contain the standard fungal mitochondrial gene set, plus *rps3* (encoding ribosomal protein S3). Also, we identified *rnpB* (encoding the RNA subunit of mitochondrial RNase P; mtP-RNA) in the mitochondrial genomes of *Microbotryum*, *Schizophyllum* and *Ustilago*, which has not been described before in basidiomycetes. Intriguingly, *rnpB* has so far only been recognized in mtDNAs of a few zygomycete and ascomycete fungi, two protists, and never in animals and plants. Basidiomycetes feature impressive variation in mitochondrial gene order (even between species of the *Microbotryum* genus), mitochondrial genome size, composition of intergenic regions, and the presence of introns and other insertion elements. These results provide strong evidence for the presence of extensive mitochondrial recombination in basidiomycetes, a feature shared with other fungi but conspicuously lacking in most Metazoa. The complex pattern of observed rearrangements may be explained by the combined roles of recombination, mobile element dynamics, and mitochondrial inheritance in relation to reproductive mode.

P-2216

Structural evolution of primate genomes mediated by highly similar duplicated sequences undergoing gene conversion

Jeffrey Fawcett, Hideki Innan
Graduate University of Advanced Studies, Hayama, Kanagawa, Japan

Duplicated sequences have played a large role in the structural evolution of primate genomes by inducing structural changes, such as duplications, deletions, and inversions, by mediating non-allelic homologous recombination (NAHR). These changes can have adaptive consequences but are also responsible for various genomic disorders. One might imagine that recently duplicated sequences that are nearly identical are responsible for such changes. However, many duplicated sequences responsible for genomic disorders appear to be shared with other primate species, indicating their rather early origin. In addition, a number of examples have emerged where recurrent structural changes are caused by duplicated sequences with high similarity not due to their recent origin but due to long-term gene conversion throughout evolution. Here, to better understand the role of gene conversion between duplicated sequences in facilitating the structural evolution of primate genomes, we examined copy-number variations (CNVs) and rearrangements associated with genomic disorders reported in the literature that are likely caused by NAHR between duplicated sequences. We found several cases where such duplications are shared with other primate species and show traces of homogenization by gene conversion. In addition, in some cases, the same shared duplicates appear to be generating CNVs not only in humans but also in chimp and/or macaque probably due to the high similarity maintained by gene conversion. These results suggest that duplicated sequences with high similarity maintained by gene conversion are frequently involved in generating structural changes in primate genomes. We also suggest that high gene conversion rate might be sufficient to preserve such highly similar sequences in genomic disorder regions despite their obvious negative effects. We argue that gene conversion has a significant and widespread role in preserving genomic configurations that facilitate the structural evolution of primate genomes

P-2217

Abundance of ultramicro inversions within whole genome alignments between closely related species

Yuichiro Hara, Tadashi Imanishi
National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

Chromosomal inversion is one of the most important mechanisms of genome evolution. While recent comparative genome analysis revealed more than one thousand inversions between the human and chimpanzee genomes, the impact of inversions in human evolution is still obscure. Minimal or nearly minimal inversion, which we call ultramicro inversion, has been hitherto-untouched because most of the ultramicro inversions are such short that the local alignments completely cover them and, as a consequence, may be falsely characterized as mutations and/or indels-rich regions. In this study, we developed a method for identifying ultramicro inversions by scanning of local genome alignments. This technique achieved a high sensitivity and a very low rate of false positives. We identified 3,648 ultramicro inversions, ranging from five base pair to 125 bp, within the orthologous nucleotide alignments between the entire human and chimpanzee genomes. Based on phylogenetic profiles using the primate outgroups, 714 and 1,789 ultramicro inversions were inferred to have specifically occurred in the human and chimpanzee lineages, respectively. Ultramicro inversions were preferably found in the region correlated with genomic structural instability. For example, a large part of the ultramicro inversions, 937 inversions (25.7%), exclusively involved adenine and thymine, and the density of the ultramicro inversions in chromosome Y was 4.0 times higher than those in autosomes and chromosome X. In addition, 98 ultramicro inversions were identified within the exons of the human protein-coding genes, suggesting that a part of the ultramicro inversions may have contributed to gene evolution. In order to examine whether ultramicro inversion is a universal evolutionary event of organisms' genomes, we collected genome alignments of various closely related species and identified ultramicro inversions in insects, fungi, and plants. While the ultramicro inversions in *Saccharomyces paradoxus* strains and rice plants are as frequent as primates, those in *Drosophila* species are three times as frequent as those in primates. These observations suggest that ultramicro inversion is a shared characteristic among various lineages, and that the frequencies of them vary among the lineages. Our new findings of ultramicro inversions would be helpful in understanding the evolutionary mechanisms of genomic structure of minimal size.

P-2218

Structural evolution in high-grade serous ovarian cancer and its influence on patient outcome and chemotherapy resistance

Roland Schwarz[1,2], Susanna Cooke[3], Charlotte Ng[1,2], Jill Temple[1], James Brenton[1,2], Florian Markowetz[1,2]
[1]Cancer Research UK Cambridge Research Institute, Cambridge, UK, [2]Department of Oncology, University of Cambridge, Cambridge, UK, [3]The Wellcome Trust Sanger Institute, Cambridge, UK

High-grade serous ovarian cancer (HGSOC) is known to exhibit a significant degree of genomic heterogeneity within a patient at presentation. Due to reduced DNA repair capabilities the metastatic branching process accumulates structural variations and point mutations. These changes, such as secondary BRCA2 mutations that restore homologous recombination, have been proposed as a potential source of resistance to platinum-based chemotherapy (Sakai2009). It is hypothesized that chemotherapy selects for minor resistant subclones embodied in presentation disease leading to short progression-free survival and resistant relapse (Cooke2011). So far no systematic effort has been made to quantify intra-patient structural variation in HGSOC and map events to its evolutionary tree which would allow us to date such events and characterize the impact of selective pressure by chemotherapy and immune response. Here we perform such an analysis, show how structural heterogeneity and the degree of clonal expansion predict patient response and survival, and point out potential mechanisms for increased tumor resistance after chemotherapy.

We infer evolutionary trees of cancer within patients from copy number changes using a tailored transducer-based inference algorithm that models horizontal dependencies. We use a parsimony inspired criterion of a minimum event distance to quantify spatial and temporal heterogeneity in a patient cohort with multiple samples per patient. We show how heterogeneity predicts CA125 tumor marker reduction after chemotherapy and how it together with the degree of clonal expansion is highly correlated with progression free survival. Our results point at loss-of-heterozygosity events around 9q34 as potentially conferring drug resistance; this region includes members of the TRAF family of proteins and a set of genes found as growth-inhibiting by a shRNA-based vulnerability study (Project Achilles, Cheung2011). Reconstruction of ancestral genomes allows relative dating of these events and indicates that they happen late in tumorigenesis, underlining the hypothesis that the resistant cells are available as a minor subclone at presentation.

Our results show that structural evolution is an important factor in the adaptive processes that shape chemotherapy-resistant ovarian cancer. Our work demonstrates how modern approaches to evolutionary inference can increase our understanding of the degree and distribution of genomic structural variation. This in turn has important implications for cancer genomics studies, because current efforts like The Cancer Genome Atlas (TCGA Network 2011) only contain single samples per patient and might systematically under-estimate heterogeneity of disease.

P-2219

Horizontal gene transfer and gene conversion as a novel mechanism of intron loss

Nancy Hepburn, Derek Schmidt, Jeff Mower
University of Nebraska-Lincoln, Lincoln, NE, USA

Intron loss is often thought to occur through retroprocessing, the reverse transcription and genomic integration of a spliced transcript. In plant mitochondria, the evidence for intron loss via retroprocessing is generally weak and indirect, usually supported by the parallel loss of a few adjacent RNA edit sites. To evaluate mechanisms of intron loss, we designed a PCR-based assay to detect recent intron losses from the mitochondrial *cox2* gene within genus *Magnolia*, which was previously suggested to have variability in *cox2* intron content. Our assay showed that all 22 examined species have a *cox2* gene with two introns. However, one species contains an additional *cox2* gene that lacks both introns. Both *cox2* versions are present in the mitochondrial genome, but the distribution of RNA editing is inconsistent with retroprocessing models. Instead, our analyses indicate that the intronless gene was horizontally acquired from a eudicot and then underwent gene conversion with the native intron-containing gene. Models are presented to summarize the roles of horizontal gene transfer and gene conversion as a novel mechanism of intron loss.

P-2220

UNEARTHING THE FOSSIL RECORD OF ENDOGENOUS RETROVIRUSES WITHIN THE PRIMATE LINEAGE

Ravi Subramanian, John Coffin
Tufts University, Boston, MA, USA

Integration of retroviral DNA into a germ cell may lead to a provirus that is transmitted vertically to that host's offspring as an endogenous retrovirus (ERV). In humans, ERVs (HERVs) comprise about 8% of the genome, the vast majority of which are truncated and/or highly mutated and no longer encode functional genes. The most recently active retroviruses known to integrate into the human germ line are the Betaretrovirus-like HERV-K (HML-2) group, many of which contain intact open reading frames (ORFs). We compared endogenous retrovirus integrations in the primate lineage as a tool to study the host-virus evolution throughout the Catarrhine primates. We have identified and characterized the relationships of many HML-2 proviruses integrations that occurred in the last 30 million years (MY), identifying new proviruses that are unique to orangutans, rhesus macaques, and chimpanzees. Non-human primate specific integrations suggest a possibility for cross-species transmission of HML-2 viruses, including a potential reservoir of infectious betaretroviruses existing in non-human primates. Furthermore, comparative genomics of ERVs in multiple primate species gives us insights into the history of virus evolution against host restriction factors. Finally, our analysis identified 9 human-specific proviruses that are incorrectly labeled as being present in the publicly available gorilla genome, underlining the concern for inappropriate identification of HERVs in re-sequenced human genomes.

P-2221

Conservation of the Human Integrin-Type Beta-Propeller Domain in Bacteria

Bhanupratap Chouhan[1,2], Alexander Denesyuk[1], Jyrki Heino[3], Mark S. Johnson[1], Konstantin Denessiouk[2]
[1]Department of Biosciences, Åbo Akademi University, Turku, Finland,[2]Turku Center for Biotechnology (University of Turku and Åbo Akademi University), Turku, Finland,[3]Department of Biochemistry and Food Chemistry, University of Turku, Turku, Finland

Integrins are heterodimeric cell-surface receptors with key functions in cell-cell and cell-matrix adhesion. Integrin α and β subunits are present throughout the metazoans, but it is unclear whether the subunits predate the origin of multicellular organisms. Several component domains have been detected in bacteria, one of which, a specific 7-bladed β-propeller domain, is a unique feature of the integrin α subunits. Here, we describe a structure-derived motif, which incorporates key features of each blade from the X-ray structures of human αIIbβ3 and αVβ3, includes elements of the FG-GAP/Cage and Ca^{2+}-binding motifs, and is specific only for the metazoan integrin domains. Separately, we searched for the metazoan integrin type β-propeller domains among all available sequences from bacteria and unicellular eukaryotic organisms, which must incorporate seven repeats, corresponding to the seven blades of the β-propeller domain, and so that the newly found structure-derived motif would exist in every repeat. As a result, among 47 available genomes of unicellular eukaryotes we could not find a single instance of seven repeats with the motif. Several sequences contained three repeats, a predicted transmembrane segment, and a short cytoplasmic motif associated with some integrins, but otherwise differ from the metazoan integrin α subunits. Among the available bacterial sequences, we found five examples containing seven sequential metazoan integrin-specific motifs within the seven repeats. The motifs differ in having one Ca^{2+}-binding site per repeat, whereas metazoan integrins have three or four sites. The bacterial sequences are more conserved in terms of motif conservation and loop length, suggesting that the structure is more regular and compact than those example structures from human integrins. Although the bacterial examples are not full-length integrins, the full-length metazoan-type 7-bladed β-propeller domains are present, and sometimes two tandem copies are found

P-2222

Evolution of gene structure in conifers.

Juliana Sena[1], Isabelle Giguère[1], Brian Boyle[1], Kermit Ritland[2], Jean Bousquet[1], John Mackay[1]
[1]Institute for Systems and Integrative Biology and Center for Forest Research, Université Laval, Québec/Québec, Canada, [2]Department of Forest Sciences, University of British Columbia, Vancouver/British Columbia, Canada

Conifer trees represent a group of ecologically dominant, long-lived, undomesticated and ancient species belonging to the group of gymnosperm plants. As a group, they have very large genomes ranging from 18 to 30 Mb. The impacts of their large genome size on gene structure and evolution are poorly understood. To address these questions and gain insights into the factors that have shaped modern day conifer genomes, we isolated 20 BAC clones (128 Kb on average) containing single copy genes of *Picea glauca* and submitted them to GS-FLX shotgun sequencing. We delineated the gene structure of introns and exons based on full length cDNAs from *P. glauca*. A comparative analysis of gene structures was carried out with the angiosperm *Arabidopsis thaliana, Zea mays* and one other conifer, *Pinus taeda*. Overall, gene structures and exons were well conserved but the total length of the genes varied considerably. We calculated the total length of introns and found that the ratio between *P. glauca* and *A. thaliana* was 4.1 on average (median 2.4), showing a considerable intron size increase in *P. glauca*. Comparison of *P. glauca* and *Z. mays* homologous genes gave an average ratio of 1.8 (median 1.8) for total intron lengths in favour of *P. glauca*; however, six of the genes had a shorter total intron length in *P. glauca*, indicating that the evolution of intron size is heterogeneous. We were also able to compare seven of the *P. glauca* genes with BACs from *Pinus taeda* available in the NCBI database. Their gene structures were largely conserved and the ratio of intron lengths was 1.1 on average. A comparative sequence analysis of 50 introns from these two conifer species indicated that introns smaller than 450bp are well conserved (sequence similarity up to 90%), suggesting that shorter introns may be under selection and that evolution of intron size may not be completely random. We investigated potential causes of large intron sizes in conifers by searching for transposable elements in the gene structure of 1006 genes from *Picea glauca* (sequences obtained from NimbleGen Sequence Capture Technology). Only 0.44% and 0.46% of the exonic and intronic sequences, respectively, were composed of TE fragments (average 150bp, ranging from 30bp to 850bp). These results indicate that conifers tend to accumulate more long introns. Known transposable elements may have a role in the evolution of expressed coding sequences but seem to play a minor role in intron size.

P-2223

Genomic structure and evolution of multigene families: "Flowers" on the human genome

Hie Lim Kim[1,4], Mineyo Iwase[1], Satoko Kaneko[2], Yukako Katsura[1], Takeshi Igawa[3], Tasuku Nishioka[1], Naoyuki Takahata[1], Yoko Satta[1]
[1]The Graduate University for Advanced Studies (Sokendai), Hayama, Kanagawa, Japan,[2]Kyoto University, Kyoto, Japan,[3]Hiroshima University, Higashihiroshima, Hiroshima, Japan,[4]Pennsylvania State University, State College, PA, USA

We report the results of an extensive investigation of genomic structures in the human genome, with a particular focus on relatively large repeats (>50 kb) in adjacent chromosomal regions. We named such structures "Flowers" because the pattern observed on dot plots resembles a flower. We detected a total of 291 Flowers in the human genome. They were predominantly located in euchromatic regions. Almost half of the Flowers (42%) have both tandem and inverted repeats, and show reticulated patterns. Flowers are gene-rich compared to the average gene density of the genome. Genes involved in systems receiving environmental information, such as immunity and detoxification, were overrepresented in Flowers. Within a Flower, the mean number of duplication units was approximately four. The maximum and minimum identities between homologs in a Flower showed different distributions; the maximum identity was often concentrated to 100% identity, while the minimum identity was evenly distributed in the range of 78% to 100%. Using a gene conversion detection test, we found frequent or recent gene conversion events within the tested Flowers. Interestingly, many of those converted regions contained protein-coding genes. Computer simulation studies suggest that one role of such frequent gene conversions is the elongation of the life span of gene families in a Flower by the resurrection of pseudogenes.

P-2224

EVOLUTION OF COPY NUMBER VARIATION IN THE RHESUS MACAQUE B-DEFENSIN REGION

Barbara Ottolini, Edward J. Hollox
University of Leicester, Leicester, UK

Beta defensins are multifunctional secreted short peptides: they present antibacterial and antiviral action in many species, possess immune cell signal activity in humans, recruiting immature dendritic cells to infection sites, and control coat color in dogs and presumably in cattle. In humans the β-defensin region is known to be copy number variable (CNV) and contains six genes repeated as a block, with a diploid copy number between 1 and 12 and an approximate repeat length of 2.5 Mb. Although genome-wide comparative genomic hybridisation arrays (aCGH) provided evidence of this region also being CNV in chimpanzee and macaque, the extent and nature of CNV in different mammals remains unknown. The rhesus macaque (*Macaca mulatta*) is the most widespread non-human primate and represents a good model for immunity studies. Its genome has been sequenced, although there is poor assembly quality in repeated segments such as the β-defensin region.

For all these reasons, we studied the genomic architecture of the rhesus macaque β-defensin region using a variety of methods. We performed high-resolution aCGH on 16 non-related macaque individuals, PCR-based methods (Paralogue Ratio Test and microsatellite assay) on a cohort of 70 samples, and metaphase spread FISH and fibre-FISH on lymphoblastoid cell lines from 5 related individuals. Combining these approaches, we aimed to overcome the limitations of the assembly and to define an absolute copy number for this region in rhesus macaque.

We present here preliminary results showing that only the region containing the human DEFB4 (DEFB2L) orthologue is CNV, while the rest of the genes show no variation. The repeated block encompasses only 8.5kb, considerably smaller than that found in humans, and this will be discussed in an evolutionary context. Also, we will present how the combined use of PCR-based and cytogenetic approaches allows absolute copy number calling and may be applied to other mammalian species to give insight into the genomic evolution of the β-defensin region.

P-2225

Detecting change in speciation and extinction rates with phylogenetic trees

Sacha Laurent[1,2], Marc Robinson-Réchavi[1,2], Nicolas Salamin[1,2]
[1]*University of Lausanne, Lausanne, Switzerland,* [2]*Swiss Institute of Bioinformatics, Lausanne, Switzerland*

In order to test for the efficiency of the available methods for estimating macro-evolutionary rates on a phylogeny, simulations were conducted using the standard birth and death model. Besides, designs were created in order to allow changes in rates at particular points of the phylogeny. Methods such as MEDUSA or BayesRates were then applied in order to verify whether the signal was recovered accurately.

P-2226

And another paradigm bites the dust: rooted global phylogeny

Ajith Harish, Anders Tunlid, Charles Kurland
Lund University, Lund, Sweden

Rooted global phylogeny has been reconstructed using compact protein domains (Fold Super Families) data from 47 fully sequenced genomes from each of the three superkingdoms. The archaea and bacteria emerge as sister clades that diverge from a last akaryote common ancestor (LACA). The eukaryote clades diverge independently from a different last common ancestor, (LECA). LACA and LECA diverge independently from the last universal common ancestor (LUCA). All three ancestors are quite complex and share a majority of their fold super families (FSFs). From a total of 1732 unique FSFs present in extant genomes, LUCA, LACA and LECA contain according to one reconstruction algorithm 1138, 871 and 1088 unique FSFs, respectively. The independent divergences of akaryotes and eukaryotes are not consistent with the suggestion that the evolution of the latter occurred by fusion of cells from the former. The great complexity of the reconstructed ancestral genomes is consistent with the paleontological record which suggests that at least seventeen mass extinction events during the last billion years have efficiently expunged simpler ancestral clades from the tree of extant organisms.

P-2227

Identifying species trees from probabilities of clades or splits.

James Degnan[1,2], Elizabeth Allman[3], John Rhodes[3]
[1]University of Canterbury, Christchurch, New Zealand,[2]National Institute of Mathematical and Biological Synthesis, Knoxville, TN, USA,[3]University of Alaska Fairbanks, Fairbanks, AK, USA

Many methods for inferring species trees have been developed in the last five years. As new methods are developed, it is important to know what types of data can be used to reconstruct species trees. Here we show that probabilities of clades on rooted trees (i.e., subsets of taxa that are monophyletic) can theoretically be used to recover the topology of the species tree for trees of any size. We also show that probabilities of splits (analogous to clades but for unrooted trees) can be used to recover the rooted species tree for the 5-taxon case by using invariants (linear combinations of split probabilities that evaluate to 0 only for certain topologies) and linear inequalities. The results are especially of interest for users of the software BUCKy that use clade and split probabilities to generate concordance trees.

P-2228

Green and red algal phylogenetic signals in nuclear genes shared by eukaryotes bearing secondary plastids of green algal origin: looking beyond endosymbiotic versus lateral gene transfer

Shinichiro Maruyama, John Archibald
Dalhousie University, Halifax, Nova Scotia, Canada

Primary and secondary endosymbiosis has given rise to a myriad of photosynthetic eukaryotes that have driven the evolution of the biosphere. It is widely accepted that the ancestors of two distantly related eukaryotic lineages, euglenophytes and chlorarachniophytes, independently acquired plastids (chloroplasts) by phagocytosing green algal prey cells. However, unlike the straightforward evolutionary history that can be inferred from plastid genome sequences, analyses of host nuclear genomes have been much less clear. Although green algal phylogenetic signals from the euglenophyte and chlorarachniophyte nuclear genomes are often deemed evidence of endosymbiotic gene transfer, distinguishing between gene transfers from endosymbionts and those from environmental sources with similar phylogenetic affinities (e.g., prey cells) has proven difficult. To address this, we conducted a phylogenomic analysis of algal-like genes in the host nuclear genomes of euglenophytes and chlorarachniophytes. As expected, a significant number of these genes are of green algal ancestry. Given that the euglenophyte and chlorarachniophyte host lineages are only distantly related, we were surprised to find a number of genes shared among these 'green' secondary plastid-bearing eukaryotes and red algae, as well as organisms bearing secondary plastids of red algal origin. One interpretation of the data is that the shared 'red' phylogenetic signals reflect a high rate of gene uptake from phagocytosed prey cells by ancestrally phagotrophic host lineages, which also could have contributed to the 'green' signal in these genomes. Collectively, our results suggest that the presence of algal genes in the nuclear genomes of these and other secondary plastid-bearing organisms may not necessarily be the result of endosymbiotic gene transfer. It is often difficult to tell which green algal-like genes in the euglenophyte and chlorarachniophyte nuclear genomes are derived from endosymbionts and which are not. The implications of these findings on current models of nuclear genome evolution in complex algae are discussed.

P-2229

Recombinations associated with Gene Transfer Events Reveal a Complex Evolutionary History of Aminoacyl-tRNA Synthetases

Greg Fournier[1,2]
[1]*MIT, Cambridge, MA, USA,* [2]*NASA Astrobiology Institute, Cambridge, MA, USA*

The phylogenies of aminoacyl-tRNA synthetases are complex, with numerous identified horizontal gene transfers between distantly related groups, as well as ancient duplications of "homeoalleles" with subsequent lineage-specific losses and transfers within groups. Here, we show that, in many cases, these transfers are associated with intragenic recombination events between donor and recipient homologs, which are subsequently inherited by descendant lineages. So far, TyrRS, TrpRS, ValRS, and LeuRS are all shown to contain several of these events, involving different groups of organisms, and different regions of the conserved protein structure. We also show that the specific regions undergoing recombination potentially reveal the selective pressure driving the fixation of the transferred gene, as well as the preservation of recipient regions. Finally, we suggest that these events complicate phylogenetic inference of horizontal transfer, through fallacious averaging between disparate signals contained within gene regions, resulting in gene trees that do not accurately reflect the evolutionary history of any of the constituent regions.

P-2230

Quantifying beta diversity over phylogenetic networks

Donovan Parks, Robert Beiko
Dalhousie University, Halifax, Nova Scotia, Canada

Phylogenetic beta-diversity measures use gene trees to compare the taxonomic or metabolic diversity of communities. Such analyses yield univariate diversity estimates that can be related to environmental and geographic factors. Currently, measures of phylogenetic beta diversity are restricted to operating over trees. We have generalized the widely used UniFrac statistic so it can be applied to phylogenetic networks. This allows beta diversity to be calculated even when there is a degree of uncertainty in the reference phylogeny. We demonstrate that the proposed method can be interpreted as an assessment of phylogenetic beta diversity averaged over phylogenetic uncertainty. Application of our method to a phylogenetic network inferred from 16S rRNA sequences from 19 marine metagenomes collected off the Atlantic coast of North America revealed that environmental conditions strongly influence the composition of these communities. This pattern persists despite high levels of phylogenetic uncertainty. When applied to a haplotype network of human mitochondrial DNA we recovered patterns of migration similar to those obtained with a classical fixation index (F_{st}) analysis. Nonetheless, we demonstrate that the incorporation of phylogenetic information can produce results which differ from those of F_{st} and discuss the merits of these two types of analyses.

P-2231

Genetic variance-control in genome-wide network expression analyses.

Ronald Nelson[1], Xidan Li[1], Xia Shen[2], Örjan Carlborg[1]
[1]Swedish University of Agricultural Sciences, Uppsala, Sweden,[2]Uppsala University, Uppsala, Sweden

By combining genome-wide genotypic data and expression data it is possible to directly associate polymorphisms on the DNA level to biochemical pathways. The usefulness of the approach has been shown in earlier association studies, where large effects of individual genes have been implied. The nature of biochemical pathways is, however, a complex interplay of many molecular components that are likely to act in both an additive and a non-additive fashion. Although the traditional association and linkage studies are powerful in identifying additive effects, they are less suitable for identifying complex interactions involving multiple loci. Using a newly developed method (vGWAS) where the variance of the phenotype (here, gene expression) is analysed instead of the mean, we are able to find multiple loci that regulate the variability in expression phenotypes rather than the mean. Using expression data from a cross of two yeast strains (BY4716 and RM11-1a), we show that the vGWAS is useful for identifying gene-gene interactions on both expression and DNA levels. In this way, it presents an approach to fill the gaps between the genotype and expression phenotype in populations by assigning new functions to known biochemical pathways and expand our understanding of incomplete gene networks and biochemical pathways. Also, once the markers or genes that significantly affect the phenotype are found, the differences in response between crossed strains can be investigated. Of particular interest are mutations that lead the different evolutionary trajectories. vGWAS is thus emerging as a useful tool to supplement current network analysis methods, that use genomic and expression data to understand the role of genetics in biochemical pathways and their evolution.

P-2232

Interdomain transfer of an intein residing within the archaeal-type ATP synthase catalytic subunit.

J. Peter Gogarten, Shannon M. Soucy, Kristen S. Swithers, Pascal Lapierre, David Williams
University of Connecticut, Storrs,Connecticut, USA

Inteins are self-splicing parasitic mobile genetic elements found in all three domains of life. The vacuolar and archaeal ATPase catalytic subunits are a protein family often invaded by these parasites. Two distinct sites within the *vma-1* gene have been targeted by inteins: the “a” insertion site in yeasts, and the “b” insertion site in Archaea. The two insertion sites correspond to the most conserved regions in the ATPase catalytic subunits [1]. The inteins targeting the two insertion sites are not closely related. To date thirteen “b”-type inteins were identified in the NCBI and JGI sequence databases. In case of *Ferroplasma acidarmanus* strains coexisting in the acid mine drainage community in the Richmond Mine at Iron Mountain [2], only one strain (fer2) has been invaded by the intein. Similarly, *Candidatus* Nanosalinarum does contain the intein; whereas *Candidatus* Nanosalina, characterized in the same environment [3], does not. The coexistence of these two states suggests the intein is still actively moving through the population. Surprisingly, a *vma1*-b intein is present in *Mahella australiensis* [4], a moderately thermophilic bacterium, grouping with the Thermoanaerobacteriaceae. In light of the unexpected finding of an archaeal intein in a bacterium, we performed phylogenetic analyses to determine the evolutionary history of this intein family. Breakpoint analysis using the Genetic Algorithm for Recombination Detection (GARD) [5]suggests points of recombination at the extein/intein boundaries. Further phylogenetic reconstructions, mapping of intein presence/absence onto the host protein phylogeny, and analysis of the pairwise divergence rate ratios between extein and intein sequences reveals multiple transfers of the intein, including a transfer between the archaeal and bacterial domains.

1. Swithers KS, et al. (2009) Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. BMC Evol Biol 9: 303.

2. Baker BJ, et al. (2006) Lineages of acidophilic archaea revealed by community genomic analysis. Science 314: 1933-1935.

3. Narasingarao P, et al. (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. The ISME journal 6: 81-93.

4. Sikorski J, et al. (2011) Complete genome sequence of Mahella australiensis type strain (50-1 BON). Standards in genomic sciences 4: 331-341.

5. Kosakovsky Pond SL, et al. (2006) Automated phylogenetic detection of recombination using a genetic algorithm. MBE 23: 1891-1901.

This research was supported through NSF DEB 0830024 and NASA Exobiology Program (NNX08AQ10G).

P-2233

Understanding the transcriptional gene regulation networks: a quantitative genetics approach

Murat Tugrul, Tiago Paixao, Gasper Tkacik, Nick Barton
IST-Austria, Klosterneuburg, Austria

The relation between genotype and phenotype is a central challenge in biology. Transcriptional gene expression levels can be considered as a highly heritable molecular phenotype, and which can be analysed as a set of quantitative traits. We adapt a thermodynamic approach to modelling transcription that allows us to map from sequence to expression level in a mechanistic manner. Our research focuses on the evolutionary dynamics of transcriptional gene interactions. We analyze the evolutionary dynamics of different gene networks (including a single promoter) by assuming stabilizing selection on gene expression, and mutation and recombination of DNA sequence. This research framework promises to help us address many questions, from molecular (e.g. the number and motif distribution of transcription factor binding sites) to evolutionary (e.g. the maintenance of genetic variation in quantitative traits, through a balance between mutation and stabilising selection).

P-2234

Evolution of the Omp85 protein family and implications for outer membrane protein diversity

Eva Heinz, Trevor Lithgow
Monash University, Clayton, Victoria, Australia

Members of the Omp85 protein family can be found across eukaryotes and bacteria, and are characterized by a membrane-embedded domain (bacterial surface antigen) and a variable number of polypeptide-transport associated (POTRA) domains. An intensely studied protein of this family is BamA, the key member of the multi-protein BAM complex, which is conserved across bacteria and is responsible for the assembly of proteins into the outer membrane. Mostly based on work in *Escherichia coli* and *Neisseria meningitidis*, BamA is known to insert several essential porins, and mutants compromised in BamA function have drastic outer membrane defects. The eukaryotic members of the Omp85 family, Sam50 and Toc75, are located in mitochondria or plastids, respectively. Here, they perform similar functions to transport and insert proteins into the outer membrane of the respective organelles. A recent analysis has shown that there is, however, even more diversity - the newly described TAM complex (Selkrig et al, *Nature Struct. Mol. Biol.*, in press), which also includes a member of the Omp85 family (TamA) is responsible to insert bacterial autotransporter proteins into the outer membrane, a process previously assumed to be performed by the autotransporters themselves. In addition to this diverse distribution of BamA homologues, several organisms have more than one copy of BamA, begging the question *why*? We are working to decipher the relationships of the different members of the Omp85 family (BAM, SAM, TAM) – with a view to a better understanding of their function(s) and evolution. These analyses promise exciting insights into the evolutionary mechanisms and extent of dependency between membrane translocation complexes and their client proteins. In addition, the bacterial outer membrane comprises a high amount of known virulence factors, and a comparison with closely related virulent and non-virulent strains will be investigated to see the potential impact of the finding of an additional BamA copy on pathogenicity.

P-2235

EVOLUTION OF PROTEIN COMPLEXES BETWEEN SPECIES AND THEIR HYBRID

Jean-Baptiste Leducq, Guillaume Charron, Guillaume Diss, Alexandre Dubé, Christian Landry
PROTEO – Institut de Biologie Intégrative et des Systèmes - Université Laval, Québec, QUEBEC, Canada

Understanding how Darwinian evolution results in contrasted patterns of molecular evolution remains a fundamental question in evolutionary biology. Recent advances in whole-genome sequencing now allow to dissect the mechanisms responsible of such divergences in many model taxa. However, knowledge is still lacking about how complex key protein interaction networks (PINs) are affected during evolution to confer the same or different phenotypic traits. Moreover, few genomes have been studied at the proteome level, and even less at the interactome level since investigating PINs in vivo is made difficult in many organisms for practical purposes. Recent progress using protein-fragment complementation assay (PCA) technology allowed to identify thousands of protein-protein interactions (PPIs) of the yeast Saccharomyces cerevisiae (S.c), many of them involving fundamental cellular functions. Here we adapted the PCA to S. kudriavzevii that diverged from S. cerevisiae 10-20 My ago. We examined whether we could readily detect PPIs in these species and their hybrids using the PCA in two well-known and -conserved protein complexes: the Nuclear Pore Complex (NPC) and the RNA polymerase II (RNApII). We found that the PCA method was highly powerful to detect and compare PPIs in S. cerevisiae and S. kudriavzevii since most of interactions were conserved and the PCA signal strongly correlated among both species and their hybrids. In one hand, our results support the expectation that the structure of essential protein complexes is highly conserved during evolution. They also suggest that this structure is not perturbed in hybrids, despite the high molecular divergence observed between orthologous proteins. We are extending the PPIs screen to the whole interactome of S. kudriazevii and hybrids in order to detect changes of PPIs associated with species-specific traits, even involved in speciation and hybrid sterility.

P-2236

Reverse engineering the gap gene network in dipterans.

Anton Crombach[1], Karl Wotton[1], Damjan Cicin-Sain[1], Maksat Ashyralyyev[2], Johannes Jaeger[1]
[1]Centre for Genomic Regulation (CRG), Barcelona, Spain, [2]Bahçesehir Üniversitesi, Istanbul, Turkey

We are performing a comparative systems-level study of the gap gene network, involved in patterning in the early embryo, across three species of diptera: the vinegar fly Drosophila melanogaster, the scuttle fly Megaselia abdita, and the moth midge Clogmia albipunctata. I will focus on the methodology we have developed to elucidate multiple gene network architectures in a relatively short amount of time.

We use a reverse engineering approach, also known as the gene circuit method, to infer regulatory interactions from mRNA-based, spatial, gene expression data. Originally developed for protein expression data, we adapted the gene circuit approach to work with whole-mount enzymatic mRNA in-situ hybridization.
In this manner, we avoid various labour-intensive stages of the experimental protocol, and we are able to simplify the image processing of the fly embryos substantially.

To verify our method works correctly, we use Drosophila as a test case. We extract the boundaries of gap gene mRNA expression domains from embryos in the late blastoderm stages C10-C14A. We show that from this mRNA boundary data one can reliably infer the gene regulatory network of the gap genes. Furthermore, we explore the minimal requirements of a data set in order to successfully perform the reverse engineering.

Concluding, we have developed a general method for the reverse engineering of gene regulatory networks. We apply this method to Drosophila and other flies, and we are convinced it is more widely applicable, potentially opening various fields of investigation to a larger scale analysis of their favourite gene networks.

P-2237

Efficient constraint-based metabolic modeling and optimization in R

Gabriel Gelius-Dietrich, C. Jonathan Fritzemeier, Abdelmoneim Desouki, Martin J. Lercher
Heinrich-Heine-University, Düsseldorf, Germany

Whole organism metabolic networks contain on the order of 1000 chemical reactions; only for a small fraction of these reactions do we know detailed dynamical parameters. However, if we assume that natural selection has adjusted reaction rates in order to optimize a fitness-related metabolic output of the network (e.g., by adjusting cellular enzyme counts), we can still simulate network function. This is the basis of flux-balance analysis (FBA) [Edwards et al., 2001] and related constraint-based approaches. Within this framework, it is possible to simulate genetic defects as well as network changes during evolution [Pál et al. 2005, 2006; Szappanos et al., 2011].

We developed SyBiL, an open source Systems Biology Library for R, which implements the corresponding algorithms (e.g., FBA, linear MOMA, robustness analysis, flux variability). R is a freely available object oriented programming environment, and is widely used in biological data analysis and modeling.

Additional R packages (glpkAPI, clpAPI and cplexAPI) we developed to communicate with standard libraries to perform linear and mixed integer optimization. These packages can be used as 'stand alone' software or in conjunction with SyBiL. The API packages are designed in particular for usage in efficient large scale analysis, e.g. when analysing systematic perturbations, such as all single or multi-gene knockouts [Szappanos et al., 2011].

Input files for SyBiL containing metabolic networks can be SBML formated or in csv format. The input parsing functions provide basic methods to check model consistency like dead end pathways. Metabolites involved in such pathways can be removed from the model.

Future work will include the integration of dynamic and regulatory FBA. We particularly welcome collaborations with groups interested in applying constraint-based methods to the study of metabolic network evolution.

References

Edwards, J. S., Ibarra, R. U. & Palsson, B. Ø. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 2001, 19:125.

Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 2005, 37:1372.

Pál, C., Papp, B., Lercher, M. J., Csermely, P., Oliver, S. G. & Hurst, L. D. Chance and necessity in the evolution of minimal metabolic networks. Nature 2006, 440:667

Szappanos, B. et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet 2011, 43:656.

P-2238

A comprehensive census of horizontal gene transfers from prokaryotes to unikonts

Pere Puigbo, Sergei Mekhedov, Yuri I. Wolf, Eugene V. Koonin
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Horizontal gene transfer (HGT) is a dominant factor in the evolution of prokaryotes. In eukaryotes, the impact of HGT is generally assumed to be much lower but has not been thoroughly analyzed, with the exception of the massive HGT from endosymbionts. Here we report a comprehensive census of likely HGT events in different groups of unikonts and an in depth analysis of a conservatively defined minimal set of transfer events. We analyzed complete proteomes from 36 species of unikonts (1 Archamoeba, 1 Mycetozoa, 18 Fungi, 13 Metazoa and 1 Choanoflagellida). These proteomes were manually selected to widely represent the Unikont supergroup. Candidate horizontally transferred genes were initially identified by analyzing each proteome using the DarkHorse method. Subsequently these candidates were analyzed in detail using taxonomic breakdown of database search results and maximum likelihood phylogenetic analysis followed by statistical tests on tree topology. Several tests were performed to rule out contamination of eukaryotic genome sequences with prokaryotic ones including analysis of the genomic neighbors for each HGT candidate and analysis of exon-intron structures of the candidate genes. Using this methodology, we detected 1816 highly probable HGT events from prokaryotes to unikonts. Approximately 90% of the detected transfers were from bacteria and the remaining 10% from Archaea. The transfer events are non-uniformly distributed in the evolution of unikonts: for example, almost all gene transfers detected in Amoebozoa (202 events) occurred after the divergence of Archamoeabe and Mycetozoa. In addition, there are many more apparent HGT events in Fungi (1050) than in Metazoa (369) or Choanoflagellates (153). Moreover, the distributions of the probable bacterial donors of transferred genes were substantially different for different Unikont taxa. Examination of the known and predicted functions of the genes acquired from prokaryotes reveals a clear preponderance of genes encoding various enzymes. We conclude that, although HGT is not as pervasive in eukaryotes as it is in prokaryotes, the amount of HGT detected in this study implies that acquisition of genes from bacteria played a major role in the evolution of the unikonts.

P-2239

A system-level, molecular evolutionary analysis of mammalian phototransduction

Brandon Invergo, Ludovica Montanucci, Hafid Laayouni, Jaume Bertranpetit
IBE-Institute of Evolutionary Biology (UPF-CSIC), CEXS-UPF-PRBB, Barcelona, Spain

Visual perception is initiated in the photoreceptor cells of the retina via the phototransduction system. This system has shown marked evolution during mammalian divergence in such complex attributes as activation time and recovery time. We have performed a molecular evolutionary analysis of proteins involved in mammalian phototransduction in order to unravel how the action of natural selection has been distributed throughout the system to evolve such traits. We found selective pressures to be non-randomly distributed according to both a simple protein classification scheme and a protein-interaction network representation of the signaling pathway. Proteins which are central to the signaling pathway, such as the G proteins, as well as retinoid cycle chaperones and proteins involved in photoreceptor cell-type determination, were found to be more constrained in their evolution. Proteins peripheral to the pathway, such as ion channels and exchangers, as well as the retinoid cycle enzymes, have experienced a relaxation of selective pressures. Furthermore, signals of positive selection were detected in two genes: the short-wave (blue) opsin (OPN1SW) in hominids and the rod-specific Na+/Ca2+,K+ ion exchanger (SLC24A1) in rodents.

P-2240

Homology Network Articulation Points Reveal Rosetta Stone Proteins

Leanne Haggerty, James McInerney
NUI Maynooth, Kildare, Ireland

In this study we have looked for 'Rosetta Stone' proteins that provide homology links between protein sequences that otherwise do not find each other in database searches. In other words, we look for protein B that has homology with A and C, but neither A nor C have homology with each other. These Rosetta Stone proteins can be fusions or hybrids of other proteins. They might be proteins that retain homology to two *de facto* homologous subfamilies that otherwise have diverged so much that there is no recognizable homology between them.

Homology data can be efficiently represented and explored using network structures. These networks can be traversed using standard graph-theory approaches. We propose an algorithm, based on network analysis, that produces a dataset that is enriched in Rosetta Stone genes. We perform an all-versus-all similarity search of a collection of genes. We then represent the output of this analysis as a network where each gene in our sample is a node and each edge is a statement of homology between the two nodes it connects. Using the Bron-Kerbosch algorithm we find all maximal cliques in the network. The Rosetta Stone genes are those that lie in the overlap of two maximal cliques, *i.e.* part of the gene is homologous to one maximal clique and a different part is homologous to the other clique.

We tested the algorithm on different datasets, from a variety of different prokaryote 'phyla'. We found pervasive homology linkages that spanned very small segments of the genes. In some cases the two cliques are highly divergent versions of the same homolog and being a member of both cliques really represents a slow evolutionary rate where a gene retains similarity to both divergent subfamilies. Network articulation points reveal a story of homology that is untold by pairwise sequence comparison.

P-2241

The origin and evolution of the melanopsin gene family: the *OPN4x* and *OPN4m* paralogs

Rui Borges[1,2], Vítor Vasconcelos[1,2], Agostinho Antunes[1,2]
[1]CIIMAR, Porto, Portugal,[2]FCUP, Porto, Portugal

In 1998 additional photosensitive cells have been identified in the mammalian retina. Melanopsins (*OPN4*) constitutes a photopigment found in specialized photosensitive ganglion cells involved in the regulation of circadian rhythms and pupillary light reflex. Differently from their close evolutionary relatives, the invertebrate visual opsins (*InRHO*) that are specialized in image forming visual functions, melanopsins have an intrinsic relationship with cone and rod opsins, playing an important role in non-image forming functions. Thus, the study the genomic and proteomic variability of melanopsins in vertebrate genomes is of particular interest to assess the main evolutionary events that shape circadian rhythms regulation at the genetic level. Moreover, melanopsins underwent a duplication event that produced two variants - the *OPN4m* and *OPN4x* paralogs - and their relative structural and functional differences are yet to be properly described.

Here, we performed a comprehensive gene/protein evolutionary study of the melanopsin gene family (*OPN4m* and *OPN4x*). We used detailed phylogenetic, synteny and selective pressure analyses to unravel the *OPN4* evolutionary history. We show that melanopsins paralogs are present in all groups of vertebrates, excluding mammals that lost the *x*-type. Furthermore, we found that melanopsins evolve under negative selection although some minor episodes of positive selection and functional divergence could be described. We also correlated the syntenic information about the genomic segment where melanopsins are confined with the whole genome duplication events to explain the great melanopsin protein variability on vertebrate genomes.

Our melanopsins evolutionary analyses are insightful to understand the origin of the rhabdomeric opsins on the context of animal photoreception emergence and to disentangle the molecular and genetic mechanisms by which the circadian reception stimuli are received and signalize, especially on fishes where many copies are found.

P-2242

Networks of Polynesian Language Evolution Reveal Frequent Word Borrowing Across the Islands

Shijulal Nelson-Sathi[1], Simon Greenhill[2], Russell D. Gray[2], Tal Dagan[1], William F. Martin[1]
[1]Institute of Molecular Evolution, Duesseldorf, NRW, Germany, [2]Department of Psychology, Auckland, New Zealand

Languages, like genomes, evolve by vertical inheritance. Understanding language evolution under different geographical and social-cultural barriers is still a challenge. Homologous to lateral gene transfer (LGT) in prokaryotes, words can also be borrowed between languages during evolution. One of the important factors that influence word borrowing among languages is the contact between them. To address the frequency of word borrowing rates over marine barriers, we studied the lateral evolutionary component of 33 recently diverged Polynesian languages spoken in different islands. The settlement and expansion phases during population expansion in the Polynesian islands brought them new technological and social innovations and maintained a cultural contact across the islands. Analysing both basic vocabulary and whole lexicon (4,751 cognates) using networks approach reveals a high frequency of word borrowing among Polynesian languages during evolution. Up to 96.2 % of basic vocabulary and 98% of lexicon cognates have experienced one or more borrowing events. Our inference of mean borrowing frequencey for basic vocabulary and total lexicon yielded similar estimates of 2.7 and 2.8 borrowing per cognate respectively. Hence counter to the widely accepted view, words of basic vocabulary in the Polynesian languages are not more resistant to borrowing than the total lexicon. The borrowing frequency during Polynesian language evolution is five-fold higher than the borrowing estimates for Indo-European languages using the same approach. Hence, counterintuitive, the marine boundaries between the islands pose a lower barrier for word transfer in comparison to the continental barriers across Europe. This may be due to the lower cultural barriers among the Polynesian settelments as the Austronesians have developed new social-cultural strategies for dealing with greater geographical isolation.

P-2243

Are changes in lifestyle associated with bursts of HGT in proteobacteria?

Thomas Laubach, Martin J. Lercher
Heinrich-Heine-University Duesseldorf, Duesseldorf, Germany

The members of the family Enterobacteriaceae live as commensals or pathogens of mammals but can also be found in freshwater and soil. They are genetically endowed to survive in extreme environments and can adapt to changing living conditions. Horizontal gene transfer (HGT) plays a major role in prokaryotic evolution and is a putative mediator of newly acquired skills among different bacterial species.

There is evidence that successfully transferred genes often represent adaptations to changed or new environments. This suggests that changes in lifestyle may be associated with bursts of HGT events.

To test this hypothesis, we compiled data for 12 lifestyle traits across 51 Salmonella, Escherichia coli and Shigella strains. Using parsimony, we reconstructed ancestral lifestyles. Based on gene presence and absence patterns, we also identified putative horizontal gene transfers in the same species. We then compared the distribution of horizontal gene transfer events to the distribution of lifestyle changes across the branches of the phylogeny.

P-2244

Protein-Chaperone Connectivity in Yeast and Protein Ancestry

David Bogumil[1], David Alvarez-Ponce[2], Giddy Landan[1], James McInerney[2], Tal Dagan[1]
[1]*Institute of Molecular Evolution, Heinrich-Heine University, Düsseldorf, Germany,* [2]*Department of Biology, National University of Ireland, Maynooth, County Kildare, Ireland*

Eukaryotes originated from an endosymbiosis event of an α-proteobacterium, the ancestor of mitochondria, within an archaeon host. During the evolution of the mitochondrion, many of the α-proteobacterial endosymbiont genes were transferred into the host nuclear genome. Imprints of this ancient event have been recently recognized within eukaryotic protein-protein interaction networks, where interactions between proteins of the same ancestry (eubacterial or archaebacterial) are more frequent than expected by chance. Here we study the evolution of the protein-folding pathway in eukaryotes in light of the endosymbiotic origin of eukaryotes by examining a binary, bipartite network of chaperones and their interacting partners in *S. cerevisiae*. We find that substrate-chaperone interactions are independent of the substrate and chaperone ancestries. Substrate connectivity, however, depends upon the protein ancestry, with archaeal substrates being highly connected to chaperones in comparison to eubacterial substrates. Yet, a comparison of eubacterial and archaeal substrate physiochemical properties revealed that archaeal substrates do not fit the typical profile of proteins that depend on chaperones for folding. They are shorter and contain a lower proportion of beta-sheets and coiled coil structures, as well as less hydrophobic and aromatic residues in comparison to eubacterial substrates. Our results suggest that proteins of archaeal origin do not interact with chaperones more frequently than eubacterial proteins, but rather that they are less specific in their interaction partners.

P-2245

The molecular basis of rapid temperature-driven divergence in European grayling (Thymallus thymallus) as revealed by shotgun proteomics

Spiros Papakostas[1], Mei Ning[2], Asbjørn Vøllestad[3], Craig Primmer[1], Erica Leder[1], Matthieu Bruneaux[1]
[1]University of Turku, Turku, Finland, [2]Institute for Nutritional Sciences, Shanghai, China, [3]University of Oslo, Oslo, Norway

Rates of evolutionary change in a population can be high but our understanding about the molecular mechanisms that may enhance or constrain rapid adaptation in natural populations is still limited. Thermal adaptation in particular is important for every species as internal temperature can affect normal biochemical reaction rates, physiology and life-history traits. This is especially true for aquatic ectotherms, like fish, which are sensitive to changes in water temperature. In this study, we investigate the molecular basis of rapid thermal adaptation in subpopulations of a salmonid fish species, the European grayling, in an introduced population in Norway. These subpopulations have common ancestors from a colonization event about 22 generations ago and earlier research has revealed distinct developmental patterns consistent with temperature-driven adaptation predictions. We used embryos from a common-garden experiment in which four sympatric grayling subpopulations, two cold-spawning and two warm-spawning, were each grown at natural cold and warm conditions (6 oC and 10 oC water temperature, respectively). Using isobaric tags, iTRAQ, we quantified more than 700 proteins in developing embryos from the four subpopulations. Three biological and two technical replicates per population were used. Not surprisingly, we found that temperature had a profound impact on protein expression with over 50 proteins being significantly affected by temperature in both cold and warm subpopulations. Yolk proteins were negatively correlated with water temperature indicating lower conversion efficiency in the colder environment. On the other hand, proteins associated with muscle development had a positive correlation with temperature, a finding that supports previous knowledge regarding trade-offs between myoskeletal growth and environmental temperature in these subpopulations. Surprisingly, there was little difference between the proteomes of cold and warm subpopulations raised in common water temperatures. This suggests a high degree of plasticity in protein expression among subpopulations. Overall, our findings may disentangle the interplay between plastic and adaptive gene expression patterns in association with ontogenic thermal adaptation in European grayling.

P-2246

Evolutionary analysis of innate immune system network in human and mouse marries phenotype and genotype.

Andrew Webb, Claire Morgan, Thomas Walsh, Mary O'Connell
Dublin City University, Dublin, Ireland

In modern immunological studies, the mouse has become the predominant model organism - an unsurprising fact considering the reported similarities between mouse and human in both physiology and genetics. However, there are an increasing number of instances in the scientific literature where mouse models of human disease have resulted in unexpected phenotypes that do not mimic the human condition. This raises an important question about the widespread suitability of mouse to modeling the innate immune system. To better understand the cause of the observed disparity between human and mouse innate immunity, we have formulated a bioinformatic pipeline to analyze the molecular evolution of these genes. We have generated a dataset containing 725 innate immune-related human genes from the highly curated database InnateDB. We then identified homologs across 21 high-coverage (>6X) vertebrate genomes. Utilizing these gene datasets we performed two stages of analysis. The first stage of the analysis assayed for functional shift whereby the identified homologs were placed into protein families for alignment, model selection, phylogenetic signal testing, and phylogenetic reconstruction where appropriate. Protein families were analyzed for lineage specific selective pressures, as lineage specific positive selection in key regions could lead the unexpected phenotypes in immune response due to functional shifts. The second phase of the analysis assesses network structure in two distinct ways. First**,** we will model the evolutionary history of the innate immune network in human and mouse using the MCL (Markov Clustering) algorithm. Secondly, we will assay for the presence of domain sharing and its potential affect on the innate immune network. Here we present the findings of our evolutionary analysis of the innate immune system network to date.

P-2247

Towards a Timeline in Networks of Lateral Gene Transfer during Prokaryote Evolution

Ovidiu Popa, Giddy Landan, Tal Dagan
Institute of Molecular Evolution, Duesseldorf, Germany

Lateral gene transfer (LGT) is an important mechanism for natural variation in prokaryotes where multiple mechanisms for gene acquisition have evolved including transformation, conjugation, transduction, and gene transfer agents. Accumulating evidence shows that LGT, a distinctly non-treelike evolutionary process, plays a major role in prokaryote evolution. Virtually all prokaryotic gene families have been affected by LGT and only few gene families are resistant to it. Consequently the reconstruction of microbial phylogenomics is slowly shifting towards using network approaches. Studying the structural properties of LGT networks revealed hub species that frequently donate or receive genes, as well as community structure of densely connected donors and recipients. How such network properties evolve over time is so far unknown. Here we use a network of directed LGT events among genomes, where donors and recipients are connected by directed edges of LGT events reconstructed from phylogenetic trees. Using a likelihood based divergence time analysis, we estimate the relative age of LGT events in the network. Incorporating these relative time estimates into a cumulative network of LGTs yields snapshots of the network in different time-points during evolution. This enables us to study temporal dynamics in LGT during microbial evolution. Our results reveal the formation and development of communities of highly interacting species, including merging of separate communities and splitting into sub-communities. We also detect that while some species serve as LGT hubs at relatively constant levels, other species' role as LGT hubs changes over time. Some of the trends observed can be correlated with species habitat, symbiotic relationships and pathogenicity.

P-2248

Large scale analysis of plasmids relationships through gene sharing networks and composition analyses

Manu Tamminen[2], Emanuele Bosi[1], Marko Virta[2], Renato Fani[1], Marco Fondi[1]
[1]University of Florence, Florence, Italy,[2]Department of Food and Environmental Sciences, Helsinki, Finland

Plasmids are vessels of gene exchange in microbial communities and the paradigmatic example of the network-like structure that sometimes arises in bacterial evolution. They are known to transfer between different host organisms and acquire diverse genetic elements from chromosomes and/or other plasmids. Therefore, they constitute an important element in microbial evolution by rapidly disseminating various genetic properties among different communities. A key example of this is the dissemination of antibiotic resistance genes that has resulted in the emergence of multiresistant pathogenic bacterial strains. To globally analyze plasmids evolutionary dynamics, we built a large graph in which 2343 plasmids (nodes) are connected according to the proteins shared by each other and applied graph theory measures to analyze its global properties. The analysis of this gene sharing network revealed an overall coherence between network clustering and phylogenetic affiliation of the corresponding micro-organisms, likely resulting from genetic barriers to horizontal gene transfer. Furthermore, analyses of networks metrics revealed a statistically significant correlation between plasmid mobility and their centrality within the network, providing support to the observation that mobile plasmids are particularly important in spreading genes in microbial communities. Our study also reveals an extensive (and previously undescribed) sharing of antibiotic resistance genes from Actinobacteria to Gammaproteobacteria, suggesting that the first might represent an important reservoir of antibiotic resistance genes for the latters.

Finally, this approach was integrated with data coming from composition analyses of (probably) recently transferred plasmids genes (plasmids atypical genes, PAGs), providing additional clues on the timing and the importance of plasmids-mediated HGT within the complex bacterial evolutionary network and in the dissemination of important biological traits.

P-2249

The molecular basis of speciation in Drosophila.

Nitin Phadnis, Harmit Malik
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Speciation - the process by which pne species splits into two - involves the evolution of reproductive barriers that stop gene flow between populations. A fundamental goal in evolutionary biology is to identify the genes that cause reproductive isolation between species and to understand the particular biological forces that drive the evolution of such genes. There is now growing evidence that genetic conflict between segregation distorters and their supp.ressors might drive the evolution of hybrid sterility. Previously, I showed that a single gene - *Overdrive* - causes both segregation distortion and hybrid sterility between the Bogota and USA subspecies of *Drosophila pseudoobscura*. We performed a genome-wide dissection of the genetic architechture of all components that interact in a single Dobzhansky-Muller incompatibility and cause male sterility in Bogota-USA hybrids. We also dissect the genetic basis of segregation distortion and its suppression in Bogota. Surprisingly, the genetic bases of both, sterility and segregation distortion, involve only a handful large effect loci that are largely, but not completely, overlapping. Here we describe how the identification of these genes, along with a comprehensive understanding the molecular and cellular properties of *Overdrive* are providing a much richer understanding of how segregation distorters can drive the evolution of new species.

P-2250

Sequence similarity networks reveal gene fusion evolution.

Pierre-Alain Jachiet[1], Romain Pogorelcnik[1], Eric Bapteste[1], Philippe Lopez[1]
[1]*Université Pierre et Marie Curie, Paris, France,* [2]*LIMOS UMR CNRS 6158, Aubière, France*

Gene fusion events are important evolutionary phenomenons. They are major contributors to the evolution of multi-domain proteins, generating novel genes and functions. They also represent valuable 'Rosetta stone' information for the identification of potential protein-protein interactions and metabolic or regulatory networks.

We developed a new method to identify fused genes families, based on the structure of sequence similarity networks. A fused gene exhibits similarity to its components, even though these components show no similarity to each other. Therefore in a sequence similarity network, a fused gene family forms a dense group of vertices that links otherwise unrelated groups of vertices. We formalized this pattern as a minimal clique separator. This well-studied graph topology provides a robust and fast method of detection, well suited for automatic analyses of big datasets. We implemented this method into a piece of software that additionally indicates potential fusion points. Its results are excellent for recent fusion events on both biological and simulated data.

We used this software to study the evolution of fused genes at the scale of Life, based on a large sample of 865.000 sequences, from fully sequenced cellular organisms - bacteria, archaea, eukaryotes - and mobile genetic elements - viruses and plasmids. We found that fusion is a universal process. It occurs in every type of entities, but more so in eukaryotes, and concerns every biological function. It appears especially important for the evolution of amino acid metabolism and transport proteins, and comparatively less so for information storage and processing proteins.

P-2251

An initial survey of the protein-protein interaction network of the plant pathogen *Phytophthora infestans* and its dynamics during infection

Michael F Seidl, Adrian Schneider, Berend Snel
University Utrecht, Utrecht, The Netherlands

Phytophthora infestans, which caused the Irish potato famine, is a model organism for pathogenic oomycetes. In contrast to our continuously growing knowledge on molecular mechanisms involved in the interaction between this pathogen and its hosts, core biological functions have so far received less attention. Protein-protein interactions (PPI) create a network that reflects essential pathways and protein complexes. In a few but relevant eukaryotic organisms e.g. human and yeast, the protein-protein interaction networks (PIN) have been experimentally defined using yeast-two hybrid or tandem affinity purification techniques. Despite the economical and ecological importance of *P. infestans*, an experimental determined PIN will not be available in the near future. To survey the *P. infestans* PIN, we used *in silico* methods to predict its composition and integrate gene expression data to study the dynamics of this network during infection.

We obtained the core *P. infestans* PIN by transferring high-confidence PPIs from yeast and human to *P. infestans* based on the prediction of a substantial number of orthologs between these species. We predicted 3,300 orthologous groups and subsequently transferred PPIs that originated from three different data sources; ranging from 13% successfully transferred interactions for BioGRID up to 39% for CORUM. The union of the three predicted interaction networks formed the *P. infestans* PIN: This network consists of 17,582 PPIs involving 3,584 proteins in 1,862 orthologous groups resulting in 19 tightly connected components. We integrated gene expression data to identify differentially expressed modules during infection and to study their dynamic over time. The observed modules display coherent behaviours validating our ortholog predictions and the PPI mapping. We exemplified the observed dynamics by studying the eIF3 complex, which is surprisingly upregulated during the early, biotrophic phase of the infection.

We demonstrated the feasibility of PPI mapping in *P. infestans*. By applying comparative genomics, we are able to generate hypothesis on fundamental biological processes guided by experimentally derived PINs. Furthermore, we will highlight how we can extend this initial survey of the *P. infestans* PIN by using co-occurrence to predict functional association between genes, e.g. genes encoding proteins involved in host-pathogen interaction that are not part of the experimentally defined PINs. Moreover, we discuss how the addition of comprehensive expression data and their comparison between closely related species enables us to predict associations and to study the dynamics of the network in greater temporal resolution during different stages of the *P. infestans* lifecycle.

P-2252

Evidence for Network Evolution in an *Arabidopsis* Interactome Map

Benoit Charloteaux[1,2], Anne-Ruxandra Carvunis[2,3], Matija Dreze[2,3], Samuel J. Pevzner[2,3], Murat Tassan[2,5], Mary Galli[6], Sabrina Rabello[1,7], Gourab Ghoshal[1,7], Jean Vandenhaute[4], Michel Georges[1], Joseph R. Ecker[6], Michael E. Cusick[2,3], David E. Hill[2,3], Frederick P. Roth[1,5], Albert-Laszlo Barabasi[1,7], Pascal Braun[2,3], Marc Vidal[2,3]
[1]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium, [2]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA, USA, [3]Department of Genetics, Harvard Medical School, Boston, MA, USA, [4]Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium, [5]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA, [6]Genomic Analysis Laboratory and Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, USA, [7]Center for Complex Network Research, Northeastern University, Boston, MA, USA

Genotype-to-phenotype relationships are partly mediated through physical interactions among gene products. The interplay between natural selection and interactome networks remains under-explored. We investigated the contribution of protein-protein interaction rewiring to the evolution of duplicated genes in a high quality protein-protein interactome network recently mapped for *Arabidopsis* (AI-1). This dataset contains ~10 times more pairs of proteins encoded by duplicates (paralogous pairs) than previous networks of similar quality. We were able to estimate the degree of interaction rewiring, despite experimental incompleteness, by using an empirically controlled quantitative framework. Having verified that interaction rewiring is predictive of the functional divergence of paralogous pairs, we showed that paralogous pairs share on average less than half of their interactors (~41%) soon after duplication, although they exhibit retention of ancestral history. The extent of interaction rewiring is largely independent of paralog family size but is slightly influenced by the duplication mechanism. Rather than following an exponential decay, as expected for a random rewiring, the decline follows a trend akin to a power law decay, mirroring the sequence divergence of paralogous pairs. This rapid-then-slow divergence and the magnitude of the rewiring with respect to sequence divergence are consistent with interactome network rewiring being a primary driver of evolution.

P-2253

Hierarchical modularity and the evolution of genetic interactomes across species

Colm Ryan[1,2], Assen Roguev[1], Kristin Patrick[1], Jiewei Xu[1], Harlizawati Jahari[1], Nevan J. Krogan[1]
[1]University of California, San Francisco, San Francisco, CA, USA, [2]University College Dublin, Dublin, Ireland

Genetic interactions emerge when the phenotype of a perturbation of a gene is affected by the presence or absence of another gene. These interactions can be identified in a high throughput fashion and offer insight into cellular organization at multiple levels. To date, cross-species comparisons of genetic interactomes have been restricted to small or functionally related gene sets, limiting our ability to infer evolutionary trends. To facilitate a more comprehensive analysis, we constructed a genome-scale genetic interaction map for the fission yeast *Schizosaccharomyces pombe*, containing ~1.6 million pairwise measurements and providing phenotypic signatures for ~60% of the non-essential genome. Based on comparison with a newly integrated genetic interactome from the budding yeast *Saccharomyces cerevisiae*, we propose a hierarchical model for the evolution of genetic interactions, with conservation highest within protein complexes, lower within biological processes, and lowest across distinct biological processes. We combine information from both species using a novel clustering algorithm and show how this approach can be used to identify conserved functional modules. Finally we show that, despite the large evolutionary distance and extensive rewiring of individual interactions, both networks retain conserved features and display similar levels of functional cross-talk between biological processes, suggesting general principles of genetic interactome design.

P-2254

Genomes of saprophytic *Thraustotheca clavata* and parasitic *Achlya hypogyna* shed light on the pathogenicity of oomycetes

Ian Misner[1], Cedric Bicep[3], Eric Bapteste[3], Sebastien Halary[4], Philippe Lopez[3], J Craig Bailey[2], Christopher Lane[1]
[1]University of Rhode Island, Kingston, RI, USA, [2]University of North Carolina at Wilmington, Wilmington, NC, USA, [3]Universite Pierre et Marie Curie, Paris, France, [4]Universite de Montreal, Montreal, Canada

Parasitism, as an evolutionary strategy, has evolved independently numerous times throughout the tree of life. The majority of parasite research focuses on biochemistry, infection process, and disease prevention. Despite all that we have learned about how parasites work we have a very poor understanding of how parasites evolve. Oomycetes present a unique opportunity to study the mechanisms behind parasite evolution. With saprobic members, facultative parasites, and obligate parasites in closely related taxa oomycetes provide an ideal framework for investigating parasite evolution. As organisms adapt from free-living to parasitic the selective pressures on genes and gene families will change. Ultimately, these adaptations will result in novel gene families evolving while genes and gene families required only for a free-living lifestyle are lost. In order to identify the genomic consequences of adapting a parasitic lifestyle we have sequenced the genomes of free-living *Thraustotheca clavata* and the facultative parasite *Achlya hypogyna*, two Saprolegnian oomycetes. Combining our data with the available genomes from parasitic Peronosporelan taxa we have assembled a comparative library containing a set of diverse lifestyles from closely related taxa. Evolutionary Gene Networks (EGN) combined with secretome analysis were used to identify genes or gene families under selective pressure, relative to lifestyle. Using these methods we have identified unique gene family expansions and contractions that are potentially key to the evolution of parasitism within this diverse group of organisms.

P-2255

Epistatis and the order of mutations in real and model fitness landscapes

Guillaume Achaz[1], Dan Weinreich[2], Yutaka Kobayashi[3], Atsushi Yamauchi[4], Fumio Tajima[0]
[1]Univerité Pierre et Marie Curie, Paris, France,[2]Brown University, Providence, USA,[3]Tokyo University, Tokyo, Japan, [4]Kyoto University, Kyoto, Japan

At the genome scale, most of the genetic locus have interactions with other locus within the same genome. We studied several "real" fitness landscapes and show that they could be classified broadly into 3 categories i) landscapes with no sign epistasis, ii) landscapes with "random" interactions (ie similar to kaufman NK landscapes) and landscapes where interactions are best described as a network of compatibility. Indeed, we show that allelic replacement at one locus can drive the allelic replacement at another interacting locus, which itself can drive another replacement at a third locus, etc. creating a "chain" of allelic replacements. This chain is a very strong constraint on the mutations order, since it predicts the order of several successive replacements.

P-2256

Evolution of novel traits in the plant metabolic network through gene and genome duplication

Michaël Bekaert[4], Corey Hudson[1], Patrick Edger[2], Emily Puckett[2], Chris Pires[2,1], Gavin Conant[1,3]
[1]Informatics Institute, University of Missouri, Columbia, MO, USA, [2]Division of Biological Sciences, University of Missouri, Columbia, MO, USA, [3]Division of Animal Sciences, University of Missouri, Columbia, MO, USA, [4]Institute of Aquaculture, University of Stirling, Stirling, UK

Network biology can link genetic changes to the phenotypic alterations that are the raw material for natural selection. In plants, both gene and especially genome duplication may be sources of evolutionary innovation. However, the identifying their contributions to phenotypic change is challenging. Using a combination of metabolic modeling, comparative genomics and sequence analysis, we have shown a link between metabolic enzymes retained in duplicate after recent genome duplications in *Arabidopsis* and enzymes that are predicted to carry high biochemical flux. Since this effect is coupled to a tendency for products of genome duplications to cluster in the metabolic network, we propose a time-dependent relaxation of metabolic gene dosage post-WGD. We broaden these conclusions with an analysis of gene duplication and metabolic flux across the flowering plants, showing that selection for increased dosage in high flux reactions operates more globally. Finally, we expanded our model to include a particular innovation in *Arabidopsis* secondary metabolism, where *both* gene and genome duplication have expanded the repertoire of glucosinolates, an important class of defensive compounds. Using metabolic models, we are able to show that the duplications that allowed this expansion occurred despite metabolic costs of glucosinolate production, suggesting that the compensating benefit of protection from insect herbivores is significant. Collectively, we argue that comparative genomics, network biology and sequence evolution are beginning to illuminate some of the dimmer corners of evolution.

P-2257

Dynamics and adaptive benefits of emergence and modular rearrangements of protein domains

Erich Bornberg-Bauer
Insitute for Evolution and Biodiversity, Huefferstr.1, Germany

Modularity is a characteristic of molecular evolution as it helps reuse autonomous module in different context and thus expedite evolutionary innovation. Protein domains are the evolutionary units of proteins and their rearrangements generate a rich molecular variety on which selection can act. These rearrangement processes can be well studies using HMM-based methods because they have a high accuracy and rearrangements occur relatively rarely. Accordingly, rearrangement events become amenable to study using the ever growing wealth of available genomes.

We use insects genomes, well resolved and diverse taxonomic group. We find high diversity in domain arrangements in even very closely related organisms and provide -- to the best of our knowledge for the first time -- branch specific rates for domain gain and loss and fusion and fission. The majority of all new domain arrangements can be explained by a just one step of modular rearrangement events but frequency of fission and terminal deletions increase over time. We find a particularly high rate of rearrangements in signaling molecules. Furthermore, emerging domains are predominantly single domain and have a high degree of disorder probability. They thus most likely result from neighbouring genomic regions. Most strikingly and remniscent of the tenet that paralogs evolve fastest shortly ofter creation, we also find that novel domains establish higher copy numbers within their genomes than older domains did and are predominantly associated with environmental adaptation such as biotic defence, abiotic stress response, reproduction and development. These results also demonstrate how easily domain based analyses can analyse adaptive changes and complement other, more established methods such as site based methods or gene family growth.

P-2258

Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations.

Hafid Laayouni[1], Pierre Luisi[1], David Alvarez-Ponce[2], Martin Sikora[3], Jaume Bertranpetit[1]
[1]Inuversitat Pompeu Fabra, Barcelona, Spain,[2]National University of Ireland Maynooth, Maynooth, Ireland,[3]Stanford University School of Medicine, Stanford, USA

Genes and proteins rarely act in isolation, but they rather operate as components of complex networks of interacting molecules. Therefore, for understanding their evolution it may be helpful to take into account the interaction networks in which they participate. It has been shown that selective constraints acting on genes depend on the position that they occupy in the network. Less understood is how the impact of local adaptation at the intraspecific level is affected by the network structure. Here we analyzed the patterns of molecular evolution of 67 genes involved in the insulin/TOR signal transduction pathway. For that purpose, we combined genotype data from worldwide human populations with current knowledge on the structure and function of the pathway. We identified the footprint of recent positive selection in nine of the studied genomic regions. Most of the adaptation signals were observed among Middle East and North African, European, and Central South Asian populations. We found that positive selection preferentially targets the most central elements in the pathway, in contrast to previous observations in the whole human interactome. Furthermore, genes evolving under positive selection tend to encode proteins that physically interact to each other in the pathway. These observations indicate that the impact of positive selection on genes involved in the insulin/TOR pathway is affected by the pathway structure.

P-2259

Photoactivation of zebrafish rhodopsin: insights into the molecular evolution of retinal release.

James Morrow, Belinda Chang
University of Toronto, Toronto, Ontario, Canada

Rhodopsin is the visual pigment responsible for mediating the critical first step of dim-light vision in vertebrates. Upon activation by a photon of light, the visual pigment undergoes a series of conformational changes, eventually leading to the biologically active metarhodopsin II, known to activate the G protein transducin. The decay of metarhodopsin intermediates leads to release of retinal from opsin, a crucial step in the visual cycle. As a non-mammalian vertebrate, zebrafish presents an intriguing contrast to the bovine visual system, living in an aquatic photic environment, and lacking the self-regulation of physiological temperature provided by endothermy. Additionally, zebrafish rhodopsin contains a number of interesting substitutions relative to bovine rhodopsin at transmembrane helices 5 and 6, which are implicated in various aspects of rhodopsin function. We measured the half-life of retinal release of expressed zebrafish rhodopsin to be more than twice as fast as that of bovine rhodopsin using fluorescence spectroscopy. This difference may allow bovine rhodopsin to compensate for the higher physiological temperature at which it must operate. Therefore, retinal release rates were also measured at their respective physiological temperatures, where rates became more similar. We also performed mutagenesis studies and revealed three key sites in zebrafish rhodopsin able to modulate retinal release rates: F213, V266, and W273. All three sites are located near an opening between transmembrane helices 5 and 6, thought to be used as an exit for retinal release following photoactivation. While sites 213 and 266 are variable across vertebrates, site 273 is conserved as tryptophan in all fish, and phenylalanine in all non-fish vertebrates, including bovine rhodopsin. We hypothesize that variation at this motif, especially at site 273, could represent adaptive evolution aimed at adjusting the kinetics of retinal release during the evolution of endothermy. This rate is important in vision as it is thought to contribute to *in vivo* processes such as visual pigment regeneration and visual response recovery.

P-2260

Reconstructing functional networks of the last common ancestor

Aaron Goldman, Laura Landweber
Princeton University, Princeton, NJ, USA

The last universal common ancestor (LUCA) is a statistical construct, not necessarily a single organism, representing the stage in the development of life just prior to the divergence of the Bacteria, Archaea, and Eukarya. Features of LUCA can be inferred through bioinformatic surveys of characters that span the tree of life. Taken together, these approaches depict a complex, sophisticated organism with a moderate-sized genome, a cell membrane, and a nearly complete translation system. But little is known of LUCA's metabolic characteristics. Here we infer ancient protein superfamilies by a unique machine learning survey of 1024 species with completely sequenced genomes. The machine learning simulation identifies the percentage of taxa and the percentage of phylogenetic depth for each domain of life, which, in combination, chooses the most ancient set of protein superfamilies. The results of this survey were superimposed onto a database of metabolic networks in order to infer ancient enzymes and pathways. This network approach to reconstructing early metabolism also provides an opportunity to objectively test origin of life hypotheses. For example, our approach provides strong support for the development of protein-based metabolism from RNA-derived functions, whereas this analysis explicitly rejects several alternative hypotheses.

P-2261

The Evolution of Cold Tolerance in the Endemic New Zealand Stick Insect Fauna using RNA-Seq: Counts, Pathways and Selection

Luke Thomas Dunning[1,2], Alice Dennis[2,3], Richard Newcomb[1,4], Thomas Buckley[1,2]
[1]Auckland University, Auckland, New Zealand,[2]Landcare Research, Auckland, New Zealand,[3]Allan Wilson Centre for Molecular Ecology and Evolution, Palmeston North, New Zealand,[4]Plant and Food, Auckland, New Zealand

The continued struggle for survival in the mutable environment is a major driving force of adaption and ultimately evolution. Temperature is a critical variable affecting the geographic distribution of many organisms. The relatively young New Zealand alpine habitat provides the perfect opportunity to elucidate the mechanisms that species adapt to a novel environment. Cold shock experiments coupled with RNA-Seq have been employed to ascertain which genes are differentially expressed between populations and species as a result of low temperature. These candidate cold tolerant genes have then been validated through qPCR. The large amount of sequence data produced by the 454 and Illumina sequencing has also allowed for primer design and targeted sequencing of specific genes and pathways. All 23 described NZ species have been shown to have the enzymes in the initial phases of glycolysis under selection when compared to their New Caledonian and Australian outgroups. We hypothesise that this selection is associated with the dispersal from subtropical and tropical habitats to their present temperate and cool environments.

P-2262

Comparative transcriptomics without the aid of reference genomes.

Peter Harrison[1], Marie Pointer[1,2], Judith Mank[1]
[1]University College London, London, UK,[2]University of Oxford, Oxford, Oxfordshire, UK

It is now possible to reconstruct whole transcriptomes from multiple species without the aid of reference genomes. However, it is not yet clear how useful this wealth of data is for conducting comparative transcriptomics. We have carried out de novo assembly of the American Wild Turkey utilising 104Gb of Illumina Hiseq data and validated this against the available reference genome. Using the lessons learnt from this validation, we then applied de novo transcriptome assembly approaches on a number of other species of bird for which no reference genome is available. Combining a further 501Gb of Illumina Hiseq data we identified gene expression differences across the clade allowing insights into the evolution of gene expression and alternative splicing in a similar fashion to the well-documented changes in gene sequence evolution. De novo assembly whilst essential in many cases is also becoming increasingly important considering the number of genomes that are only partially sequenced or too erroneous for effective comparative transcriptomics and in these instances de novo transcriptome assembly could prove more effective than conventional mapping.

P-2263

Patterns of gene expression divergence and protein evolution in mammals

Maria Warnefors[1], Henrik Kaessmann[1]
[1]*University of Lausanne, Lausanne, Switzerland,*[2]*Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Divergence of protein sequences and gene expression patterns are two fundamental mechanisms that generate organismal diversity. Here we use genome and transcriptome data from eight mammalian species and one bird outgroup to study the evolutionary relationships between these two processes. We explore to what extent the degree of gene expression and protein divergence correlate across genes and how this correlation changes over different evolutionary timescales and between tissues. We further consider how the rate of protein divergence varies for genes that are expressed at different levels and with different degrees of tissue specificity. The presence of multiple species in our dataset allows us to test whether changes in gene expression and tissue specificity are reflected in the evolutionary rates of individual proteins. Finally, we investigate the strength of selection on gene expression and protein sequences for different types of genes and identify functional groups that have primarily diverged in terms of gene expression, protein sequence or both. We also assess whether genes that show signs of positive selection on gene expression are more likely to also be candidates for positive selection on their protein-coding sequences, and vice versa.

P-2264

Conservation and function of noncoding RNAs in primate evolution

Courtney Babbitt, Lisa Pfefferle, Olivier Fedrigo, Gregory Wray
Duke University, Durham, NC, USA

Changes in gene expression underlie many important differences between species, however, protein-coding DNA is not the only DNA to be transcribed. Large portions of the non-coding genome are actively transcribed. Yet, it is unclear what fraction of these transcripts are biologically relevant. Evolutionary conservation provides an approach for distinguishing functional non-coding transcripts. Here we use directional paired-end RNA-Seq data to assess changes in global transcript abundance in five metabolically important tissues (cortex, cerebellum, liver, muscle, and white adipose tissue) of humans, chimpanzees, and rhesus macaques. We assay expression in both genic and intergenic regions and distinguish between sense and antisense (relative to nearby genes) transcripts. We find that an abundance of noncoding transcripts, including many never previously annotated, are conserved in both genomic location and expression level between species, suggesting an important functional role for these poorly characterized transcripts.

These conserved intergenic RNA transcripts are enriched in 5' and 3' flanking regions of protein-coding genes. Interestingly, we find a negative correlation between 5' flanking antisense transcripts and the expression of the downstream gene, suggesting that these antisense transcripts are playing a regulatory, possibly repressive, role for nearby genes. Looking at the evolutionary differences between multiple tissues, we find that many of these noncoding transcripts are playing tissue-specific roles. We find that these noncoding transcripts may be differentially regulating very specific pathways in homologous tissues between humans and non-human primates. Comparative analysis can provide important insights into transcripts responsible for differences in tissue functions between humans and non-human primates, as well as highlighting novel candidate noncoding transcripts for further detailed functional studies.

P-2265

Birth and functional evolution of mammalian microRNA genes

julien meunier[1], frédéric lemoine[1], magali soumillon[1], angelica lietchi[1], manuela weier[1], katerina guschanski[1], haiyang hu[1], philipp khaitovitch[2], henrik kaessmann[2]
[1]University of Lausanne, Lausanne, Switzerland,[2]Chinese Academy of Sciences, Shanghai, China

MicroRNAs (miRNAs) are major post-transcriptional regulators of gene expression, yet their origins and functional evolution in mammals remain little understood due to the lack of appropriate comparative data. Using RNA sequencing, we have generated extensive miRNA data for five organs across six species that represent all main mammalian lineages and birds (the evolutionary outgroup). Our analyses of these data reveal an overall expansion of miRNA repertoires in mammals, with three-fold accelerated birth rates of miRNA families in placentals and marsupials, facilitated by the de novo emergence of miRNAs in host gene introns. New miRNA genes gradually evolved higher expression levels and optimized target gene pools during evolution and apparently mainly contributed to the evolution of complex expression networks in the brain. However, through selectively driven duplication-divergence processes, X-chromosomal miRNAs evolved high expression levels and potentially diverse functions during spermatogenesis, in spite of meiotic sex chromosome inactivation.

P-2266

Accurate tracing of a transcriptome changes in the course of hybrid speciation

Till Czypionka[1], Jie Cheng[1], Alexander E. Pozhitkov[1,2], Arne W. Nolte[1]
[1]Max-Planck-Institute for Evolutionary Biology, Ploen, Germany, [2]Department of Periodontics, University of Washington, Seattle, USA

Next generation sequencing was used to develop a microarray to study evolutionary change in a recently evolved hybrid lineage of fish (Cottus). A central goal was to test whether transgressive patterns of gene expression in hybrids generate increased phenotypic variance and if such variance is utilized to invade a new adaptive peak in nature. From a technical perspective, previous studies had documented that results from RNAseq and microarray experiments can only be partially replicated which illustrates that artefacts prevail. For oligonucleotide microarrays, the binding behaviour of probes and target transcripts has been shown to vary widely but is not considered in data analyses to date. We have implemented a normalization procedure that establishes a common hybridization-signal to gene-expression ratio for all probes and fundamentally improved the quality of our inference. Levels of gene expression divergence were found to correlate with genetic divergence at neutral markers and, accordingly, hybrid sculpins were intermediate between the parental species. There was no excess of biological functions among the genes that are differentially expressed between the parental species, but the hybrid lineage is distinguished through unique patterns of gene expression that are enriched for specific biological functions. Since the rise of invasive sculpins is very recent, the Cottus system is suitable to recreate initial steps that have led to the formation of invasive sculpins. We compare independent F2 crosses with natural invasive sculpins to show that transgressive patterns of gene expression that distinguish invasives can be observed in F2 crosses and that the invasive transcriptome was subject to secondary changes after admixture.

P-2267

CHROMEVALOA: A MOLECULAR DATABASE OF CHROMATIN-ASSOCIATED PROTEINS USEFUL FOR THE EVALUATION OKADAIC ACID GENOTOXICITY IN BIVALVE MOLLUSCS

Victoria Suarez-Ulloa[1], Vanessa Aguiar-Pulido[2], Rodrigo Gonzalez-Romero[1], Ciro Rivera-Casas[1], Juan Fernandez-Tajes[1], Juan Ausio[3], Josefina Mendez[1], Julian Dorado[2], Jose M. Eirin-Lopez[1]
[1]CHROMEVOL-XENOMAR Group, Department of Cellular and Molecular Biology, University of A Coruna, A Coruna, Spain, [2]RNASA-IMEDIR Group, Department of Information and Communication Technologies, University of A Coruna, A Coruna, Spain, [3]Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada

Among the different marine biotoxins produced by harmful algae blooms, okadaic acid (OA) stands out due to its principal role in causing diarrheic shellfish poisoning episodes across the European coasts. Therefore, a great effort has been devoted to the biomonitoring of OA, mainly using bivalve molluscs as sentinel organisms. OA has been widely described as a potent tumor promoter capable of inducing extremely severe damage on the hereditary material, including DNA double strand breaks (DSBs) triggering, among other repair mechanisms, a process of chromatin remodeling involving different histone variants, most notably H2A.X and H2A.Z. Consequently, the study of the expression profiles displayed by these variants in organisms exposed to OA constitute a potential chromatin-based genotoxicity test. In this work we introduce a new molecular database, searchable through keyword queries or a BLAST-based web application, designed to improve the study of OA genotoxicity in the marine environment using the mussel *Mytilus galloprovincialis* as sentinel organism. The database provides differential expression data focused on chromatin-associated proteins, obtained from transcriptome libraries (subtractive and normalized) generated *de novo* from individuals exposed/non-exposed to OA. The web application has been built using Bioperl modules in Object-Oriented Perl (OO Perl) language, allowing an easy parameter control of the sequence alignments. Overall, this resource will contribute to unravel the mechanisms by which chromatin-associated proteins participate in the maintenance of genome integrity in molluscs, evaluating their potential as genotoxicity biomarkers. In addition, the viability of future objectives applying Knowledge Discovery in Database (KDD) techniques will be evaluated in order to further characterize the role of chromatin-associated proteins in protostomes.

P-2268

Rapid ecological speciation driven by selection on a small set of ‘adaptation' genes

Mark Chapman, Dmitry Filatov
University of Oxford, Oxford, UK

The *Senecio* hybrid zone on Mt. Etna (Sicily) is a classic example of ecological speciation due to adaptation to contrasting conditions of high and low altitudes. *Senecio aethnensis* is adapted to the challenging environment of high altitudes with UV-radiation, sulphur-rich volcanic soils and sharp changes in temperature, while *Senecio chrysanthemifolius* grows in drier and hotter conditions at the base of Mt. Etna. Despite their very recent split (only ~50,000 years ago) and extensive interspecific gene flow, the phenotypes of the two species are very different. To identify the genes responsible for phenotypic differentiation we sequenced transcriptomes of multiple individuals of the two species grown in a common environment. The analysis of expression data for almost 20,000 loci identified a list of 151 'adaptation genes'; i.e. loci that show significantly different expression patterns between the two species. Putative functions of these genes suggest a variety of roles, including freezing and UV tolerance, flowering time and sugar transport as well as a number of genes with no, or only indirect, functional evidence. Interestingly, the adaptation candidates show strong differentiation with low or zero amounts of allele sharing between the species. This contrasts with no or very little differentiation in the rest of the genome. Our data support the view that strong disruptive selection on a few genes can lead to rapid speciation despite on-going gene flow. These ‘adaptation genes' are unable to introgress and are responsible for the maintenance of morphological (and therefore genetic) species identity, with little or no differentiation in the rest of the genome.

P-2269

Quantitative trait mapping of *trans*-regulatory factors associated with opsin gene expression in cichlid fishes

Kelly O'Quin, Jane Schulte, Zil Patel, Matthew Conte, Karen Carleton
University of Maryland, College Park, MD, USA

Recent studies of regulatory evolution have emphasized a central role for *cis*-regulatory elements in the evolution of gene expression and phenotypic adaptation. Mutations within *trans*-regulatory factors also contribute to regulatory divergence, but these factors have been harder to study since their genomic position relative to the genes they regulate is unknown. African cichlid fishes vary adaptively in the expression of six cone opsin genes that are necessary for color vision: *SWS1* (ultraviolet), *SWS2B* (violet), *SWS2A* (blue), *RH2B* (blue-green), *RH2A* (green) and *LWS* (red). Previous work has shown that several protein-coding and regulatory mutations tune color vision in these fishes. To identify potential regulatory factors, we used experimental crosses of cichlids that differ in the expression of their cone opsin genes to scan for Expression Quantitative Trait Loci (eQTL). We generated 115 individuals from three hybrid F2 families and genotyped them at over 1,000 differentially-fixed single nucleotide polymorphisms (SNPs) using next generation restriction site associated DNA sequencing (RAD-seq). We then generated a genetic map from these markers and performed a regression analysis of opsin expression and marker genotypes. Our results identified four eQTL: two associated with *SWS2B* and *SWS2A* expression, one associated with *RH2B* expression, and one associated with *RH2A* and *LWS* expression. Only one of these eQTL potentially occurred in cis (on the same linkage group) to the six opsins genes, although none of the opsins were part of the associated region. The remaining three eQTL all occurred in trans (on different linkage groups) to the opsins. As part of this study, we also included 13 candidate genes that are known to regulate vertebrate opsin expression in *trans*. None of these candidates fell within the four eQTL, though comparative mapping of the associated markers to a draft assembly of the cichlid genome reveals other excellent candidates in these regions. Our results demonstrate that *trans*-regulatory factors can also play an important role in the evolution of gene regulation and phenotypic adaptation. Future work will fine-map these eQTL to further resolve what role mutations within protein-coding genes, *cis*-regulatory elements, and *trans*-regulatory factors play in the adaptive evolution of cichlid color vision.

P-2270

Bgee, a database for the study of gene expression evolution

Frederic Bastian[1,2], Julien Roux[1,2], Anne Niknejad[1,2], Aurelie Comte[1,2], Sebastien Moretti[1,2], Gilles Parmentier[1,2], Marc Robinson-Rechavi[1,2]
[1]*University of Lausanne, Lausanne, Switzerland,* [2]*Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Gene expression patterns (where and when genes are expressed) are a key feature in understanding gene function and evolution. To apply compare results between different model organisms and human, or to study gene expression evolution, a comparative approach must be used, but no tools allow to easily compare gene expression across species. We have thus developed Bgee (Base for Gene Expression Evolution), a database designed to automatically compare expression patterns between animals. This is achieved by i) the aggregation and curation of expression data from different types and sources, to map them to formal representations of anatomies and developments of different species; Bgee release 10 contains curated data for 13,560 Affymetrix chips and 3,364 EST libraries annotated by our curators, as well as 231,992 in situ hybridizations. ii) the analysis of these data by dedicated statistical tests to define high confidence gene expression patterns. iii) the definition of comparison criteria between anatomies of different species; to date, Bgee curators have designed relationships between 5,192 species-specific terms, which map to 1,175 homologous organ groups; the latter are organized in multi-species ontologies (the HOG and vHOG ontologies). Bgee is available at: http://bgee.unil.ch/

P-2271

Evolutionary patterns of regulatory divergence in *Drosophila*

Joseph Coolon[1], C. Joel McManus[2], Kraig Stevenson[1], Brenton Graveley[3], Patricia Wittkopp[1]
[1]University of Michigan, Department of Ecology and Evolutionary Biology, Ann Arbor, MI, USA, [2]Carnegie Mellon University, Department of Biological Sciences, Pittsburgh, PA, USA, [3]University of Connecticut Health Center, Department of Genetics and Developmental Biology, Farmington, CT, USA

Proper organismal function requires fine-tuned regulation of gene expression and yet variation exists within and between species. Despite its importance, the genetic mechanisms underlying regulatory evolution are not yet well understood. Regulatory divergence is governed by changes in *cis*- and *trans*-regulatory elements that affect the level of expression of genes. *cis*-regulatory changes have effects on allele-specific expression, while *trans*-regulatory variation contributes to the expression of both alleles in a diploid cell. The effects of *cis*- and *trans*-regulatory elements can be disentangled using measures of allele-specific gene expression in pairs of species and hybrids made by crossing them. A previous study of 78 genes in *Drosophila* showed that *cis*-regulatory divergence explained a greater proportion of expression differences between species than within, suggesting that *cis*-regulatory variants may preferentially accumulate over time. To test this hypothesis genome-wide we used RNA-Seq for enumeration of allele-specific expression to determine the contribution of *cis*- and *trans*-regulatory divergence to expression differences between strains and species of *Drosophila* ranging in divergence times from 10,000-4 million years ago. We found that the proportion of regulatory divergence explained by *cis*-regulatory change (%*cis*) is significantly greater between species than within, however there is not a linear relationship between %*cis* and evolutionary time.

P-2272

Evolutionary Genomics of Plant Reproductive Isolation

Toni Gossmann, Karl Schmid
University of Hohenheim, Stuttgart, Germany

Numerous hypotheses regarding mechanisms underlying reproductive isolation in plants have been established. However the molecular basis that operate on the level of genes is poorly understood. Here we present a systematic analysis on a genome wide context to identify candidate genes responsible for reproductive isolation in Arabidopsis. We analyze cell specific expression data (RNAseq and Microarray) of nine different male and female developmental stages of gametophytic cell types in Arabidopsis thaliana. Based on these results we extract genes with gametophytic specific expression patterns and compare sequence diversity and divergence to orthologs of Arabidopis lyrata and Arabidopsis halleri. For this we use polymorphism data from more than 100 A. thaliana genomes as well as 20 genomes of A.lyrata and A.halleri each. We find that the number of differentially expressed genes during male reproductive tissues varies substantially which is strikely different to female cell types. We also find evidence for positive selection among reproductive male genes but not in female genes based on an extension of the McDonald Kreitman test. We identify more than 20 candidate genes which are likely to be involved in reproductive isolation as a consequence of natural selection and aiming for a functional characterization of those. Taken together our results suggest that similar to the evolution of the human and fly Y chromosomes, the main driver of reproductive isolation in plants are male specific genes.

P-2273

Genetic variants count for gene expression variability in human lymphoblastoid cells

James Cai
Texas A&M University, College Station, Texas, USA

Expression QTL (eQTL) analyses have established convincing relationships between genetic variants and gene expression levels. To date, however, genetic variants influencing the variance of gene expression levels are poorly understood. Here, we adapt the double generalized linear regression model to identify genome-wide loci associated with gene expression variability, which we designate expression variability QTLs (evQTLs), in the human genome. We identify 218 *cis*-acting evQTL loci, among which 8 (*ADCY1*, *CTNNA2*, *DAAM2*, *FERMT2*, *IL6*, *PLOD2*, *SNX7*, and *TNFRSF11B*) are cross-validated using two data sets of gene expression assessed in lymphoblastoid cells from HapMap individuals. Furthermore, a total of 167 *trans*-acting evQTLs are identified from 500 representative genes and more than 13,000 SNPs across all autosomes. This is the first systematic analysis of variability of gene expression levels being potentially controlled by genetic variants in humans.

P-2274

Transcriptomic resilience to global warming in the seagrass *Zostera marina*, a marine foundation species

Susanne U. Franssen[1], Jenny Gu[1], Nina Bergmann[2], Gidon Winters[2], Ulrich C. Klostermeier[3], Philip Philip Rosenstiel[3], Erich Bornberg-Bauer[1], Thorsten B.H. Reusch[2]
[1]Institute for Evolution and Biodiversity, University of Münster, Muenster, Germany, [2]Evolutionary Ecology of Marine Fishes, Leibniz-Institute for Marine Sciences, Kiel, Germany, [3]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany

RNA-seq offers the opportunity to perform global transcriptome profiling of key-ecological species, predicting evolutionary and ecologically relevant responses under global warming. The seagrass *Zostera marina*, occurring along a thermal cline, provides the unique opportunity to assess temperature effects on gene expression as a function of their long term adaptation to heat stress.

Here, we exposed natural southern and northern European populations of *Zostera marina* to a realistic heat wave scenario in a common stress garden setup. In a fully crossed experiment the transcriptomic responses were obtained by RNA-seq of eight cDNA libraries during and after the heat wave along with unstressed controls, each comprising ~125 000 reads. The expression profiles were assessed subsequent to transcriptome de novo assembly and gene identification via orthologous plant genes. Expression profiles revealed similar acute heat stress responses of locally adapted populations, with a focus on heat shock proteins. Population differences, however, became apparent during heat recovery. Gene-expression patterns in southern genotypes returned to control values immediately, but genotypes from the northern site failed to recover and revealed the induction of genes involved in protein degradation, indicating failed metabolic compensation to high sea-surface temperature.

The results suggest microevolutionary adaptation of the seagrass population to different thermal environments. We conclude that the return of gene-expression patterns during recovery provides critical information on thermal adaptation in aquatic habitats under climatic stress. As a unifying concept for ecological genomics, we propose transcriptomic resilience, analogous to ecological resilience, as an important measure to predict the tolerance of individuals and hence the fate of local populations in the face of global warming (Franssen et al. 2011).

Franssen SU, Gu J, Bergmann N, Winters G, Klostermeier UC, Rosenstiel P, Bornberg-Bauer E, Reusch TB (2011) Transcriptomic resilience to global warming in the seagrass Zostera marina, a marine foundation species. *PNAS* 108(48):19276-81.

P-2275

The origins and evolution of long non-coding RNAs in tetrapods

Anamaria Necsulea[1,2], Magali Soumillon[1,2], Angélica Liechti[1,2], Manuela Weier[1,2], Henrik Kaessmann[1,2]
[1]University of Lausanne, Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Long non-coding RNAs (lncRNAs) have long been known to carry out essential functional roles, such as X-chromosome inactivation (mediated by the Xist lncRNA in placental mammals), or allele-specific gene silencing at parentally imprinted loci. The importance of lncRNAs in mammalian transcriptomes has recently become indisputable, with the first genome-wide studies that revealed the presence of thousands of lncRNA genes, in human and mouse. However, the origin and evolution of lncRNAs are still largely unexplored.

Here, we present the first large-scale evolutionary study of lncRNAs, by surveying with RNA-Seq the transcriptomes of 11 tetrapod species (great apes, macaque, mouse, opossum, platypus, chicken and xenopus) and 8 tissues (cerebellum, cortex, heart, kidney, liver, ovary, testis and placenta). Using this extensive transcriptome dataset, we detected several tens of thousands of lncRNAs in each of the studied species. We found that lncRNAs evolve considerably faster than protein-coding genes, both at the level of the DNA sequence and in terms of expression patterns. Nevertheless, we were able to identify many lncRNAs which were preserved during more than 350 million years of tetrapod evolution, and which are thus likely to be involved in important functions.

While many lncRNAs appear to be derived from protein-coding genes, through pseudogenization, we found that lncRNAs also often originate by exploiting the enhancer functions of endogenous retroviruses. This pattern is particularly strong for lncRNAs that are specifically expressed in the placenta, not only for eutherian mammals, but also in the marsupial yolk sac placenta.

Finally, we investigated sex-biased expression patterns, and showed that extreme female-specific expression (over all somatic organs) is the unique signature of Xist in placental mammals. We reveal the presence of an opossum X-linked lncRNA, which shows the same female-biased expression pattern as Xist. This lncRNA is not a homologue of Xist, but shares many of its characteristics, such as its large size, repetitive sequence content and potential to form stem-loop structures. This locus is highly conserved in marsupials, but only marginal sequence similarity could be found in placental or monotreme mammals. We propose that this lncRNA may act as a mediator of X-chromosome inactivation in marsupials, which would represent an extraordinary case of evolutionary convergence in the underlying mechanism of an essential mammalian process.

P-2276

Next-generation insights into a classic evolutionary system and the development of *Senecio*DB

Owen Osborne[1], Thomas Batstone[2], Simon Hiscock[2], Dmitry Filatov[1]
[1]University of Oxford, Oxford, UK,[2]University of Bristol, Bristol, UK

What kind of genomic changes underlie evolutionary processes such as speciation and phenotypic adaptation? Understanding the genomic basis for evolutionary change is one of the greatest challenges in modern biology; a goal in which natural, non-model systems will play a central role. On the slopes of Mount Etna, Sicily, two closely related Senecio species form an altitudinal cline, with Senecio aethnensis adapted to colder UV-exposed conditions of high altitude, while S. chrysanthemifolius inhabits the hotter drier environment at the bottom of the mountain. The species differ substantially in morphological, physiological and life-history traits and are regarded as a classic example of recent ecological speciation. Like many of the most interesting and tractable evolutionary systems, however, they have not been well characterised genomically, and this has stifled progress. We used Illumina sequencing to sequence transcriptomes from both species, as well as from an outgroup S. vernalis. We assembled the reads into 57873 contigs, which represented 19291 unique contigs present in all three species. 47% of these were putative full-length transcripts. We used the data to assemble tripartite alignments of these species and undertake the first transcriptome-wide analysis of sequence divergence in the system. Median synonymous divergence between the homologous genes of the two Mount Etna species was 0.012 ± 0.017[SD] and from either to the outgroup *S. vernalis* was 0.029 ± 0.033[SD]. Contrary to our expectations, synonymous substitution rate for the low altitude S. chrysanthemifolius was significantly higher than for S. aethnensis that is exposed to higher levels of UV at higher altitude. Also, we detected no significant difference between the species in rates of mutations that are caused by UV radiation. The non-synonymous to synonymous rate ratio was similarly low in both species (Ka/Ks 0.248 for both species), indicating the prevalence of purifying selection in the majority of genes. Finally, we identified putative duplicate gene pairs and found a significant excess of several Gene Ontology terms in those which were retained in the two species. The datasets were used to develop SenecioDB (http:www.seneciodb.org/). The database now represents a useful resource for Asteraceae researchers and will serve as the starting point for population genomic scale analyses of this remarkable system.

P-2277

A comparison of brain transcriptomes in domesticated and wild animals

Frank W. Albert[1,2], Mehmet Somel[3], Miguel Carneiro[4], Ayinuer Aximu[1], Michel Halbwax[1], Olaf Thalmann[5], Jose A. Blanco-Aguiar[4,7], Irina Plyusnina[6], Lyudmila Trut[6], Rafael Villafuerte[7], Nuno Ferrand[4], Sylvia Kaiser[8], Per Jensen[9], Svante Pääbo[1]
[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, [2]Princeton University, Princeton, NJ, USA, [3]University of California Berkeley, Berkeley, CA, USA, [4]University of Porto, Porto, Portugal, [5]University of Turku, Turku, Finland, [6]Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia, [7]Instituto de Investigación en Recursos Cinegéticos, Ciudad Real, Spain, [8]University of Münster, Münster, Germany, [9]Linköping University, Linköping, Sweden

Domestication has greatly altered the appearance and behavior of several wild animal species, often in strikingly similar ways. Little is known about the genetic basis and molecular correlates of these phenotypic differences. To better understand what molecular changes are associated with animal domestication, we used mRNA sequencing to compare gene expression patterns in brain frontal cortex between three domesticated species and their close wild relatives: dogs and wolves, pigs and boars, and domesticated and wild rabbits. To provide a context for the differences, we compared them to those between 1) domesticated guinea pigs (*C. porcellus*) and a more distantly related wild cavy (*C. aperea*), and 2) a tame and an aggressive line of rats that had been selected for their behavioral response to humans. Over-all, gene expression differences scaled well with DNA sequence divergence and domestication is found to be associated with modest changes in gene expression in dogs/wolves, pigs and rabbits. We are currently analyzing whether the gene expression differences affect common biological pathways in these four domesticated species.

P-2278

Shifting fitness consequence of variation in *LCB2* gene expression in yeast

Joshua Rest[1], Christopher Morales[1], John Waldron[1], Dana Opulente[1], Julius Fisher[1], Seungjae Moon[1,2], Kevin Bullaughey[3], Lucas Carey[4], Demitri Dedousis[1]
[1]Stony Brook University, Stony Brook, NY, USA, [2]Weill Cornell Graduate School of Medical Science, New York, NY, USA, [3]University of Chicago, Chicago, IL, USA, [4]Weizmann Institute of Science, Rehovot, Israel

Gene expression levels vary by genotype and environment, but there is little detailed or fine-scale mapping of the fitness consequences of this variation. To explore prospects for such a map, we assayed fitness for many levels of up-regulated and down-regulated expression of a single essential gene, *LCB2*, involved in sphingolipid synthesis in *Saccharomyces cerevisiae*. Reduced *LCB2* expression rapidly decreased cellular fitness, yet increased expression had little effect, indicating that the wild-type expression level is perched on the edge of a fitness cliff. We also found that in a different environment (stress) the entire fitness function is shifted upward to higher expression, although the overall shape remains the same. Lower expression levels of *LCB2* in wild yeast strains suggests that the higher levels in the experimental lab strain are the result of recent selection, possibly associated with the disruption of amino acid synthesis. Our results demonstrate that fitness functions can be highly nonlinear, but are specific to genetic background and environment. Exploration of gene expression fitness functions using this system will provide insights into the meaning of expression variation.

P-2279

How to survive over the winter on northern latitudes? Cold acclimation in two Drosophila species with different distributions - a comparative genomic approach

Maaria Kankare[1], Laura Vesala[1], Tiina Salminen[1], Asta Laiho[2], Anneli Hoikkala[1]
[1]*University of Jyvaskyla, Jyvaskyla, Finland,* [2]*Finnish Microarray and Sequencing Centre, Turku, Finland*

Adaptation to fluctuating temperature conditions especially on northern latitudes requires a capability to cope with notably lower than optimal temperatures. Short term exposure to a relatively low temperature (cold acclimation) has been found to increase cold tolerance and consequently also the overwintering survival of several insect species, but much less is known about the genetic pathways involved in this process.

Drosophila montana, a malt fly species with a wide northern distribution range (30-70°N) is well-adapted to survive in seasonally varying environmental conditions with a mean temperature below 0°C for about half of the ye ar and occasional drops close to -30°C. *Drosophila virilis*, which is not able to overwinter on high latitudes, is a close relative of *D. montana* found from market places and breweries mainly in eastern Asia (south from 35°N). Despite the obvious differences in their biology and distribution range, these two species appear to possess genetically very similar cold acclimation processes. Candidate gene microarray study with a custom design (219 genes) revealed several shared genes after six days of cold acclimation in +5°C, i ncluding genes involved in heat shock response, circadian rhythm and oxidation and carbohydrate metabolism.

To gain more information on flies' cold acclimation at genomic level, a transcriptome run with Solid next generation sequencing technique was carried out for cold acclimated and control samples of the two above-mentioned species. In *D. virilis* data the genes showing highest downregulation during cold-acclimation are involved in transporter and transferase activity, and oxidative stress defence. Upregulated genes, on the other hand, include mainly unannotated genes requiring further investigation. Information gained from this research can be used to study the evolution of seasonality and overwintering strategies in northern insect species.

P-2280

Detection of dual coding regions in mRNAs using ribosomal profiling data.

Audrey M. Michel[1], Kingshuk Roy Choudhury[1], Andrew E. Firth[2], Nicolas T. Ingolia[3], John F. Atkins[1], Pavel V. Baranov[1]
[1]University College Cork, Cork, Ireland,[2]University of Cambridge, Cambridge, UK,[3]Carnegie Institution for Science, Baltimore, USA

Ribosome profiling (also known as *ribo-seq*), a technique developed recently by Ingolia et al. (2009), yields genome-wide information on protein synthesis (GWIPS) *in vivo* (Weiss and Atkins, 2011). The technique is based on the measurement of ribosomal density on translated mRNAs by sequencing and quantifying ribosome protected fragments. The sub-codon resolution of this technique yields a triplet periodicity whose phase is dependent on the reading frame that is being translated. We have developed a computational method that exploits this characteristic triplet periodicity to detect transitions in the translated reading frame. Application of this technique to ribosomal profiling data available for humans (Guo et al 2010) has allowed us to identify dual coding instances where the same genomic segment is decoded in alternative frames. In addition to known cases of ribosomal frameshifting, examples also include dually coded regions as a result of upstream and non-upstream ORFs overlapping the main protein coding ORF, as well as dually coding exons translated in more than one frame in alternative transcripts. Comparative sequence analysis on the vertebrate orthologs of our top scoring candidates confirms evolutionary constraints on these dual coding regions.

References:

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science.* **324**:218-23.

Guo H, Ingolia NT, Weissman JS, Bartel DP. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature.* **466:**835-40.

Weiss RB, Atkins JF. (2011) Molecular biology. Translation goes global. Science. **334**:1509-1510

P-2281

Long term evolutionary adaptation of yeast to a fluctuating environment

Riddhiman Dhar, Rudolf Sägesser, Christian Weikert, Andreas Wagner
Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

Adaptation to a fluctuating environment is one of the most common challenges that confronts yeast in its natural habitat. Three mechanisms, namely cross-protection, stochastic switching and anticipatory mechanism, that could potentially aid in adaptation of yeast cells to fluctuating environment have been described in literature. However, our understanding of the relative importance of these mechanisms for adaptation to fluctuating environment is limited. In addition, we do not know if there exists other mechanisms for adaptation to fluctuating environment in yeast. Uncovering these novel mechanisms would require evolution of yeast cells to fluctuating environment in the lab and no study to date has attempted to do so. Such a study would also allow us to find whether yeast cells can evolve any of the already known mechanisms when exposed to such environment. To address this gap in our understanding, we evolved yeast cells in the laboratory under periodic environmental changes for 300 generations. In brief, we exposed yeast cells to salt stress for 10 generations followed by exposure to oxidative stress for the subsequent 10 generations. We analyzed the fitness of the evolved lines using competition assays. The yeast cells adapted rapidly to the fluctuating environmental stressors. We investigated mechanisms of adaptation through analysis of gene expression using microarrays. We observed that long-term exposure of yeast cells to oxidative stress protected them against salt stress. The evolved yeast cells also showed basal decrease in expression of signaling genes even in the absence of the stressors.

P-2282

Studying patterns of genetic variation in long non coding RNAs along the Great Ape lineage

Marta Melé[1,2], Roderic Guigó[2], Tomàs Marquès[1]
[1]IBE, Barcelona, Spain,[2]CRG, Barcelona, Spain

There is increasing evidence that long non-coding RNA (lncRNA) may play a relevant role in the regulation of several cellular processes. Studies at the patterns of conservation comparing vertebrate genomes have found that lncRNAS are indeed under purifying selection. However, except for a few well-known cases such as Xist or HotAir, most of the annotated lncRNA have an unclear function. To date, there has not yet been any comprehensive attempt to assess events of positive selection in the recent evolution of lnRNAs. For this purpose we have compared levels of intraspecific variation of lncRNAS within the human lineage and compared similar intraspecific patterns to their closest living relatives: the non-human great Apes. Here we will present the study the patterns of genetic variation of around 10,000 annotated lncRNAs in the human genome across all the major human and Great Ape populations (including full genome data for ~100 great ape genomes--great Ape diversity Consortium). We believe that studying how natural selection has targeted lncRNAs along the Great Ape lineage will allow shedding some light in our understanding of the relevance and functionality of these emerging elements of the genome.

P-2283

Sox duplicate evolution and constraints from molecular properties

Emilien Voldoire, Frédéric Brunet, Jean-Nicolas Volff, Delphine Galiana-Arnoux
The Institute of Functional Genomics of Lyon, ENS de Lyon, Lyon, France

The *sox* genes (for "*SRY*-related high-mobility-group box" genes) constitute a large family of transcription factors involved in key processes. This family, first described in mammals, is based on sequence similarity with the HMG box of the *SRY* gene. HMG domains appear to have arisen early in metazoan evolution. In chordate, *sox* gene family evolution is characterized by duplication events (two at the basis of vertebrates and one specific of teleosts) and divergence. In tetrapod, twenty *sox* genes are described and divided into eight groups according to their conserved structure and function.

We first realized a global bioinformatic approach. Obviously, teleosts have more than twenty *sox* genes due to the teleost specific whole genome duplication. They also present an interestingly high degree of duplicate retention (50% versus 20% observed in systematic studies). Then, we focused our interest on two *sox* groups (C and E), which show the most important degree of retention, and present loss and/or differential evolution depending on the taxa. For example, in group E, *sox8a* is specifically lost in Percomorpha, while *sox10a* is absent from Cypriniformes. Within *soxC*, we observed in some species a conserved but highly divergent *sox11b* duplicate. Interestingly, expression analyses show that, despite this different evolutionary history, ancestral expressions (tetrapods are used as a benchmark) of these *sox* groups appear conserved among the three teleost species analyzed (zebrafish, medaka and platyfish). For instance, all *soxC* are mainly expressed in brain and ovaries, but depending on species, *sox4* or *sox11* paralogs have different expression pattern.

An hypothesis could explain the observed retention pattern: *sox* duplicate retention could be due to evolutionary constraints from biochemical (interaction rates in particular) and pleiotropic properties of Sox protein. These molecular properties could induce purifying selection on both duplicates (network dynamics' constraints) eventually associated with an increase of the subfunctionalization rate. In agreement, expression analyses suggest that *soxC* and *soxE* have partitioned the putative pleiotropic ancestral function between the "descendant" duplicates. Otherwise, purifying selection on both duplicates could increase the temporal window during which new function can occur thanks to degenerative mutation (as it could be the case for some teleost *sox11b*). Altogether, this suggests a coevolution not only between paralogs but also between all genes from a same *sox* group coming from successive duplication events.

P-2284

Population and sex differences in *Drosophila* brain gene expression

Ana Catalan, John Parsch
LMU, Munich, Germany

We used Illumina next-generation RNA-sequencing (RNA-seq) to investigate gene expression differences in adult brain between two populations of *Drosophila melanogaster*, one from the ancestral species range in sub-Saharan Africa and one from the derived species range in Europe. Males and females of each population were examined separately. Overall, we identified a modest number of genes with sex-biased expression in the brain, but a greater number that differed in expression between populations. Most genes that differed in expression between populations did so in both males and females, although there were also cases of sex-dependent expression divergence between populations. Our data allowed us to examine specific transcript isoforms that differed in expression between the sexes or the populations. For example, an inter-populational difference in expression of the foraging gene was predominantly caused by transcripts originating from a distal upstream promoter.

P-2285

The Evolutionary Significance of Y-linked Regulatory Variation in Drosophila

Timothy Sackton[1], Jun Zhou[1], Bernardo Lemos[2], Daniel Hartl[1]
[1]Harvard University, Cambridge, MA, USA,[2]Harvard School of Public Health, Cambridge, MA, USA

The Drosophila Y chromosome is degenerated, heterochromatic, and gene-poor, with very little polymorphism in unique sequence; it has long been thought to have little role in phenotypic variation beyond a few Y-linked male fertility factors. Recently, our lab has described a new phenomenon termed Y-linked regulatory variation (YRV) that affects the level of expression of hundreds of genes across the genome. We have shown (1) that these effects are largely, if not exclusively, epigenetic effects of microsatellites and other repeated sequences on the Y, (2) that Y chromosomes from natural populations are polymorphic for their effects on autosomal and X chromosomal gene expression, and (3) that Y chromosomes from closely related species show Y-linked regulatory divergence (YRD) and that these effects are associated with differences in male reproductive fitness. However, the evolutionary significance of these findings is unclear. Two competing hypotheses can explain our findings and previous results of other investigators. In one hypothesis, the epigenetic regulatory effects of YRV and YRD could be positively selected because of their adaptive value in local populations. The alternative hypothesis is that the epigenetic regulatory effects of YRV are neutral or deleterious and maintained by mutation-selection balance, driven by high mutation rates in repetitive sequences.

Here, we present the results of RNA-seq experiments designed to test these hypotheses. Using a unique set of Y chromosome mutation-accumulation lines, we quantify the baseline rate of accumulation of Y-linked variation in gene expression using RNA-seq and for a subset of Y-linked microsatellites. By comparing this mutation-accumulation data to gene expression data collected from Y replacement lines from temperate and tropical populations of *D. melanogaster*, we parameterize models of gene expression evolution to test evolutionary hypotheses. Furthermore, we link these gene expression results to phenotypic measures of male reproductive function across Y replacement lines. Taken together, these results offer a rare opportunity to connect epigenetic regulation to specific adaptive traits, and to test evolutionary hypotheses about segregating heterochromatic variation and the evolution of the Drosophila Y chromosome.

P-2286

Molecular analysis of the mechanisms involved in *THBS4* differential gene-expression in the human brain

Raquel Rubio-Acero[1], James W. Thomas[2], Todd M. Preuss[3], Mario Cáceres[1,4]
[1]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain,[2]NIH Intramural Sequencing Center, National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), Bethesda, Maryland, USA,[3]Division of Neuroscience and Center for Behavioral Neuroscience, Yerkes National Primate Research Center, Emory University, Atlanta, Georgia, USA,[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

The last decades have seen a growing interest in what makes us humans and how the human brain differs from that of our closest relatives at the molecular level. Hundreds of genes with expression differences between human and non-human primates have been identified so far. However, it is important to study these genes in more detail to see if they are really involved in human brain characteristics. Thrombospondins (THBSs) are multimeric extracellular calcium-binding glycoproteins that modulate cell-cell and extracellular matrix interactions and have been implicated in the control of synaptogenesis. Within the THBSs family, *THBS2* and *THBS4* show, respectively, a ~2-fold and ~6-fold upregulation in human cerebral cortex compared to chimpanzees and macaques. To analyze the causes of these expression differences, we have carried out a comparative and functional analysis of the *THBS4* promoter region in humans and chimpanzees. First, we have identified and validated an alternative transcription start site (TSS) for *THBS4* that is located ~44 kb upstream from the previously known TSS and generates a new mRNA isoform encoding a 91 amino acid shorter protein. Second, we used quantitative RT-PCR to compare expression levels of both *THBS4* mRNAs in different human tissues and in cortical regions of 11 humans and 11 chimpanzees. Interestingly, the new isoform of *THBS4* is expressed predominantly in brain tissues. Moreover, consistent with the observed differences for total *THBS4* expression, it shows ~6-fold higher expression in human than in chimpanzee cortex. Third, we have evaluated the activity of the two *THBS4* promoter sequences from humans and chimpanzees in different human cell lines using reporter assays, and we have found significant differences between both promoters, but not between species. Finally, the comparison of the DNA methylation in a CpG island upstream the new isoform in 5 humans and 5 chimpanzees detected similar low methylation levels in all of them. Increased *THBS4* expression in the human brain therefore appears to be related to higher transcription from the alternative promoter and understanding its regulation could be relevant to the functional consequences of *THBS4* expression differences during human brain evolution.

P-2287

Life at Extremes: Molecular Adaptations of the Protist Halocafeteria to Hypersaline Environments

Tommy Harding, Alastair G. B. Simpson, Andrew J. Roger
Dalhousie University, Halifax, Nova Scotia, Canada

Microbes living in extremely hypersaline habitats (near salt saturation) have evolved to withstand the osmotic stress that would otherwise kill the cells. Typically, Halobacteriaceae (the famous haloarchaea) and Halanaerobiales (anaerobic eubacteria), preferentially import some ions into their intracellular environment to equilibrate the osmotic balance. This adaptation has led to a strong molecular signature that includes a highly acidic and hydrophilic proteome. Most other microbes, including protists like the alga Dunaliella salina, cope with the osmotic stress by exporting the salt, coupling it with import or synthesis of compatible solutes like glycerol. However, virtually nothing is known about heterotrophic halophilic eukaryotes (halophilic protozoa) that thrive in hypersaline habitats all over the world. We recently characterized the obligately halophilic stramenopile Halocafeteria seosinensis, a bacterivorous nanoflagellate. To unravel Halocafeteria's molecular adaptations, transcriptomic investigations were conducted using Illumina next-generation sequencing under optimal and maximal salt concentrations for growth. After clustering of reads, ~15,000 genes were found to be present in both conditions. Preliminary analysis of the amino acid composition at the fastest evolving sites of 160 house-keeping genes revealed a molecular signature different from that usually observed in extreme halophilic Archaea. Even though the acidic and hydrophilic residue glutamate was overrepresented as in haloarchaea, the borderline hydrophobic residue threonine, normally overrepresented in halophiles, was underrepresented in Halocafeteria. Also glutamine, a neutral but polar residue, was strongly overrepresented and alanine, a hydrophobic amino acid, was underrepresented. Ongoing work, including description of gene content and differential gene expression, will shed light on the evolutionary path and innovations used by Halocafeteria while being subject to the strong selection of hypersaline environments.

P-2288

Whole transcriptome sequencing to identify candidate genes in the response of the water flea, Daphnia magna, to infection with its parasite, Pasteuria ramosa.

Seanna McTaggart[1], Desiree Allen[1], Timothee Cezard[2], Carolyn Riddell[1], Marian Thomson[2], Urmi Trivedi[2], Mark Blaxter[1,2], Tom Little[1]
[1]University of Edinburgh, Edinburgh, UK,[2]The GenePool, Edinburgh, UK,[3]Center for Immunity, Infection and Evolution, Edinburgh, UK

@font-face { font-family: "Arial"; }@font-face { font-family: "Cambria"; }p.MsoNormal, li.MsoNormal, div.MsoNormal { margin: 0cm 0cm 0.0001pt; font-size: 12pt; font-family: "Times New Roman"; }div.Section1 { page: Section1; }

The water flea *Daphnia magna* and its co-evolving bacterial parasite *Pasteuria ramosa* both show extensive genetic variation for susceptible and infectivity respectively. Extensive trials have demonstrated that the host's infection status following exposure depends on the specific combination of host and parasite genotypes. However, little is known about *Daphnia*'s immune response to *P. ramosa* infection, or the genes generating the resistant and susceptible phenotypes. In the first attempt to identify the genes and pathways responsible for these interactions, we exposed three *Daphnia* host genotypes to two naturally infecting bacterial strains in a fully factorial design that included unexposed host controls. We sequenced the complete transcriptome of the host using RNA-seq. Quality controlled reads were mapped to the genome and differential gene expression and allelic specific expression were determined. We show that host genotypes differ in the genes they express in response to the invading pathogen, both in terms of absolute gene expression, as well as allelic specific response. We identified several gene families, and in particular chitinases, which change in expression upon pathogen challenge, which we discuss in the context of possible mechanisms of host defense.

P-2289

Estimating selection on codon usage in the face of noisy gene expression.

Edward Wallace[1,2], Edo Airoldi[2], D. Allan Drummond[1]
[1]University of Chicago, IL, USA,[2]Harvard University, MA, USA

The key quantities of molecular evolution -- estimates of selection, drift and mutation -- were historically limited by the availability of data; now, they are limited by our ability to properly extract these estimates from vast quantities of data.

As a classic example, in most organisms some codons are used at higher frequencies than their synonyms, particularly in high-expression genes. The pattern is consistent with mutational biases and selection for increased translational speed and accuracy. The contribution of each of these forces remains a major open question in molecular evolution. Increasingly powerful methods have been proposed which attempt to make quantitative inferences of these contributions. Either mode of selection, accuracy or speed, leads to a selection coefficient that is intuitively multiplied by gene expression level, giving gene expression measurements a central role in these inferences.

However, gene expression measurements are noisy. We show that noise in gene expression, if left unaddressed, creates havoc in these inferences: dramatic underestimation of selection, and overestimation of mutational bias. If noise is substantial enough, selection is inferred to be absent. We present an analytical framework which explicitly accounts for measurement noise, in which, given moderate amounts of noise, we can recover unbiased estimates of the true selection coefficients. In budding yeast these selection coefficients, while still compatible with weak selection on synonymous sites, are several-fold higher than those inferred using noise-blind approaches.

Our analysis provides a case study for a general theme: measurement noise -- an unavoidable part of any experiment -- is a shadowy force which does not merely obscure the signal we wish to see, but frequently biases the signal in ways we are only beginning to appreciate.

P-2290

Determination of the mechanisms of regulatory divergence across a large collection of *Saccharomyces cerevisiae* strains

Brian Metzger, Patricia Wittkopp
University of Michigan, Ann Arbor, MI, USA

Proper gene regulation is required for the correct function of organisms and differences in regulation are expected to underlie much of the phenotypic diversity within and differences between species. Determining the molecular and genetic mechanisms of gene regulatory divergence is thus a crucial component for a full understanding of the evolutionary process. Previous research on the evolution of gene regulation has explored divergence both within and between species, finding that *trans*-regulatory changes contribute more to divergence within species than between. These studies have primarily been genome surveys utilizing only a couple of strains and species. Here, we explore a different dimension of variation, focusing on the mechanisms of divergence for a single gene in a collection of 64 *Saccharomyces cerevisiae* strains. For each strain, we analyzed *TDH3* gene expression of co-cultured diploids and hybrids formed with a standard laboratory strain and found that both *cis* and *trans*-regulatory changes exist and that *trans*-regulatory changes are more common. In addition, we asked whether regulation in mRNA levels is also found at the protein level. While there is a well-known correlation between mRNA abundance and protein level across different genes, there is little evidence that differences in mRNA abundance result in meaningful protein level differences for the same gene. We thus created hybrids between all 64 strains and a reference strain carrying a copy of the *TDH3* promoter driving YFP expression. We found that strains with a significant *trans* effect on *TDH3* mRNA also had significant effects on YFP fluorescence, resulting in a significant positive correlation between *TDH3* mRNA levels and YFP fluorescence. Current work is focused on identification of these *trans*-regulators in each strain.

P-2291

Modelling copper transcriptional reaction norm evolution in *Saccharomyces cerevisiae*

Andrea Hodgins-Davis[1], Aleksandra Adomas[2], Daniel Rice[3], Jeffrey Townsend[1]
[1]Yale University, New Haven, CT, USA, [2]National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA, [3]Harvard University, Cambridge, MA, USA

Genetic variation for plastic phenotypes has the potential to contribute a large proportion of the phenotypic variation available to selection for novel adaptation. Models for linking population variation with plasticity make predictions about how selection may constrain the evolution of gene expression variation across transcriptional reaction norms. To empirically test these models, we have characterized population variation for transcriptional reaction norms in *Saccharomyces cerevisiae* for an ecologically relevant gradient of copper concentrations from starvation to toxicity. We find that although the vast majority of transcriptional variation is small in magnitude, not just some, but most genes demonstrate variable expression across environments or genetic backgrounds. Functionally, the most highly expressed genes defined three distinct cellular states across the copper reaction norm consistent with previous results characterizing eukaryotic responses to copper starvation, copper-replete fermentation, and copper overdose. These cellular states included direct transcriptional responses to intracellular levels of copper ions and diverse indirect metabolic adjustments to the consequences of changed copper levels. Indirect homeostatic changes of expression were more variable among genotypes in their direction of response than were the reaction norms of genes directly regulated by copper-binding transcription factors. To interpret this variability in transcriptional reaction norms in the context of the processes of natural selection and neutral drift, we account for variation in mutation and regulatory degree across the genome by parameterizing classic models of phenotypic evolution. Empirically estimating population genetic variance, mutational variance, and regulatory degree, we infer the strength of stabilizing selection operating on gene expression levels. We present evidence consistent with either weak or infrequent stabilizing selection on most gene expression phenotypes, and discuss the implications of the lack of constraint for the evolvability of transcriptional reaction norms.

P-2292

Gene expression patterns in two closely related crow taxa from a common garden RNA-seq experiment.

Jelmer Poelstra[1], Nagarjun Vijay[1], Vittorio Baglione[3], Inge Müller[2], Martin Wikelski[2], Jochen Wolf[1]
[1]Uppsala University, Uppsala, Sweden, [2]Max Planck Institute for Ornithology, Radolfzell, Germany, [3]University of Valladolid , Spain, Valladolid, Spain

The advent of next-generation sequencing has enabled accurate and high-throughput quantification of gene expression in non-model organisms. Gene expression divergence may be an important and potentially rapidly evolving component of speciation. Moreover, transcriptomics can be effectively employed to track down the genetic basis of phenotypic divergence. Carrion and hooded crows are two Eurasian corvids that hybridize in a narrow contact zone across Europa, yet show no significant neutral genetic divergence. They differ markedly in plumage colouration, which likely represents the main reproductive isolating barrier. However, the genetic basis of these colour differences remains unknown. A recent study of wild-caught crows from two populations surprisingly found that carrion and hooded crow gene expression profiles formed separate clusters, suggesting that in the early stages of speciation, observable divergence in gene expression may precede that among nucleotide sequences. Here, we describe a common garden gene expression experiment of carrion and hooded crows in a 2x2 population design. Using this set-up, we exclude confounding environmental differences between individuals, and are able to separate population and taxon-specific effects on gene expression. We perform RNA-seq Illumina HiSeq to test whether gene expression profiles cluster by population and/or taxon, and perform differential expression analysis using a suite of recently developed methods. We pay special attention to differential expression in melanin-producing skin tissue (from which feathers of different colours were growing) to elucidate the genetic basis of the colouration differences between these crows. Our results illustrate the feasibility of RNA-seq experiments in non-model organisms in order to answer questions of prime evolutionary relevance.

P-2293

Genetics of transcriptome-wide responses to ocean acidification in sea urchins

Daniel Runcie[1], David Garfield[2], Narimane Dorey[3], Sam Dupont[3], Gregory Wray[1]
[1]Duke University, Durham, NC, USA, [2]EMBL, Heidelberg, Germany, [3]University of Gothenburg, Kristineberg, Sweden

In addition to the global effects of climate change, marine species face a direct consequence of increasing atmospheric CO2 levels: ocean acidification (OA). Effects of OA are expected to be particularly dire for calcifying species, such as most marine invertebrates, because calcium carbonate crystals become harder to form in more acidic seawater. Current efforts to predict consequences of OA on marine communities have focused on identifying which species are most sensitive to low pH conditions. However, little is known about how the genetic structure of marine populations may affect their ability to adapt to acidified seawater. Here, we tested for genetic variation in the tolerance of larvae of the sea urchin Strongylocentrotus droebachiensis to a 0.4 pH unit decrease in seawater acidity. We integrated assays of growth, development and survival with RNAseq-based transcriptome measurements in a quantitative genetics breeding design. We found that more acidic seawater slowed larval growth rates, but had little effect on larval survival in this species. While we did not observe genetic differences in growth rate or survival in our population sample, we did find dramatic genetic differences in gene expression throughout the transcriptome, including in candidate pathways that may be involved in buffering the effects of pH in the larvae. This approach of linking life history measurements with gene expression profiling in a genetic experiment provides novel insights into how organismal buffering mechanisms may influence evolution.

P-2294

Comparative transcriptome analysis of increased encephalization in mormyrid fishes

Olivier Fedrigo[1], William J. Nielsen[1], Bruce A. Carlson[2]
[1]Duke University, Durham, NC, USA, [2]Washington University in St. Louis, St. Louis, MO, USA

The Mormyridae are African fresh-water fish that have become an important model in studies on electrophysiology and behavior because of their remarkable electrogenic and electrolocation capabilities. Interestingly, they also have an exceptionally large cerebellum. For instance, the elephant nose fish, *Gnathonemus petersii*, has a brain that accounts for ~2.7% of total body mass, comparable to that of the human brain. Brains are energetically costly, and this high degree of encephalization carries with it significant metabolic requirements. Indeed, mormyrid brain oxygen consumption accounts for ~60% of total resting body oxygen consumption, compared to ~20% in humans and 3-10% in other mammals. One hypothetical way to evolve increased encephalization is through an energy trade-off. By decreasing the energy requirements of other energetically expensive tissues, such as skeletal muscle and intestines, greater resources can be allocated to brain development and maintenance.

There are approximately 200 species of mormyrid that vary in their degree of encephalization. The experimental advantages of mormyrids and their high level of diversity, combined with recent advances in quantitative genomics, place us in a unique position to study whole body energetics with molecular characterization of gene expression between tissues and across species to identify the fundamental forces shaping the evolution and development of enlarged brains. We focused on two species with significantly different brain sizes: *G. petersii* (~2.7% of total body mass) and *Brienomyrus brachyistius* (~1.5% of total body mass). We first sequenced, assembled, and annotated the whole transcriptome of these two species using directional RNA-Seq. From this data, we established homologous gene models. We then performed RNA-Seq on these same species for four individuals each across 6 tissues (heart, liver, skeletal muscle, telencephalon, cerebellum, and intestines). We looked for differentially expressed genes between tissues and across species and contrasted these results with ontological categories, tissue-specificity, and signatures of positive selection in coding regions. We report here a set of candidate genes and pathways for the evolution of encephalization via an energy trade-off. Our work not only contributes genomic resources for the study of this fish, but also provides general insight into the genetic mechanisms that may be involved in the evolution of brain energetics.

P-2295

A transcriptomics view on speciation - gene expression studies in *Ficedula* flycatchers

Severin Uebbing, Axel Künstner, Linnéa Smeds, Hannu Mäkinen, Páll Ólasson, Hans Ellegren
Institute for Ecology and Genetics, Uppsala University, Uppsala, Sweden

Collared flycatcher and pied flycatcher represent a young pair of sister species that diverged less than 1 million years ago. Hybridization occurs in sympatry in, for example Central Europe, and given the ease by which they can be studied, the two flycatchers form an interesting ecological model for speciation and hybridisation, which has been studied intensively in the field.

Genetics divergence between species may be due to changes in gene expression patterns or changes in protein-coding coding sequences. While the latter has been in main focus in previous research, far less is known about the former. In order to gain insight into the levels of gene expression divergence in the flycatcher model, we have performed Illumina high throughput transcriptome sequencing using samples from 9 different tissues and 10 individuals from each species, 5 males and 5 females each. Sequencing resulted in more than 3 billion paired-end reads and we were able to recover the transcriptome for both species. This dataset also provides insight into patterns of sex-biased gene expression and dosage compensation, topics which are of particular interest in ZW sex chromosome systems.

P-2296

The Evolution of Reproductive Complexity in Male-Pregnant Fishes

Tony Wilson[1], Marie Gauthier[1], Kai Stoelting[1,2], Camilla Whittington[1]
[1]University of Zurich, Zurich, Switzerland,[2]University of Fribourg, Fribourg, Switzerland

The evolutionary origin of complex reproductive traits has attracted considerable theoretical interest, but empirical research in this area has been limited, due to the frequent lack of transitional forms in extant taxa and the methodological challenges inherent in assigning homology in distantly-related organisms. Syngnathid fishes (seahorses and pipefishes) are a group of close to 300 marine species known for their exceptional form of reproduction, male pregnancy. While male pregnancy occurs in all members of this group, the complexity of brooding structures varies among species, from the simple ventral attachment of eggs in some species of pipefish to the fully-enclosed pouch of the seahorse, in which males aerate, osmoregulate and provision embryos during their development.

We have designed a custom oligonucleotide microarray, derived from 454-based transcriptome sequencing, in order to quantify changes in gene expression in the brood pouch of the seahorse Hippocampus abdominalis during embryonic incubation. Strikingly, many of the differentially expressed transcripts present in the male brood pouch during pregnancy exhibit similar expression profiles in other forms of vertebrate reproduction, suggesting a common genetic basis for components of the reproductive machinery in divergent evolutionary lineages. A cross-species microarray approach is currently being used to investigate the evolution of reproductive complexity during the radiation of syngnathid fishes, in order to illuminate the genetic changes underlying functional innovations associated with the evolution of this unique form of reproduction.

P-2297

Cryptic genetic variation underlying stress resistance in evolved lines of *Caenorhabditis remanei*

Kristin Sikkink, Catherine Ituarte, Rose Reynolds, Patrick Phillips, William Cresko
University of Oregon, Eugene, OR, USA

When organisms become isolated in novel environments they must acclimate to new and often stressful conditions. If the population is to persist, then subsequent adaptive evolution must also occur. But what is the relationship between individual acclimation and adaptation to novel environments? Many organisms can acclimate to novel environments through phenotypic plasticity, which is the ability of a genotype to produce different phenotypes that match the environment. Underlying the exposure of novel plastic phenotypes is cryptic genetic variation, which does not contribute to an organism's phenotype until it is exposed by a stressful environment. Thus, variation in phenotypic plasticity can be heritable, subject to selection, and can evolve. However, an open question is how important the evolution of plasticity is generally for adaptation to novel environments, and the molecular basis of cryptic genetic variation is poorly understood. To address this question, we used experimental evolution of outbred *Caenorhabditis remanei* in a stressful environment. Lines were consistently evolved under high temperature or a control environment. The lines also were subject to one of three acute selective regimes – a heat shock, oxidative shock, or no acute stress. Replicate lines were evolved under all combinations of long-term and acute stress, allowing us to isolate the effects of each factor. After 20-30 generations of selection in the novel environment, we observed adaptation to acute stress in lines selected for either oxidative or heat shock. Furthermore, the response to selection was greater in the stressful heat environment, consistent with selection on cryptic genetic variation for acute stress resistance in the novel environment. To understand the molecular basis of the uncovering of cryptic genetic variation underlying acute stress resistance, we used RNA-seq to evaluate changes in gene expression resulting from the exposure of cryptic genetic variation in the novel environment, and to identify the key pathways that harbor cryptic genetic variation. These results provide a comprehensive understanding of the molecular basis of adaptation to a stressful environment.

P-2298

Comparative Transcriptomics for Eye Diversification of Cephalopod: Evolution of Camera eye and Pinhole eye

Atsushi Ogura
Ochanomizu University, Tokyo, Japan

The eye is a part of nervous system and connects the internal and external ecologies of organisms by processing visual information. It consists of photoreceptors, neural networks, and parts of the brain related to visual reception. A large number of studies on the developmental biology and molecular biology of eye evolution have revealed that there is a shared developmental process and a common gene regulatory network. Recent work on evolutionary genomics in various types of eyes, together with comparative analyses of gene expression comparison among closely related species, has led to the hypothesis of a dynamic mechanism for the diversification of eyes. Molluscs provide a good example of the application of evolutionary genomics studies, as all eye types have evolved in one lineage. We have compared gene expressions in eyes of pygmy squid (camera eye), and nautilus (pinhole eye) by RNA-seq. We explore the transcriptome of the developing embryonic eye of the nautilus and squid using an Illumina GAII sequencer and Roche FLX sequencers. First, to assess the RNA-seq coverage to find lowly expressed but important genes, such as transcription factors, we searched Transcription factors that are already known to be involved in eye development in molluscs. As a result, all genes were found to be expressed with a FPKM of 1.7~2.5. These FPKMs are not significantly different between the nautilus and squid. We also searched the nr protein database for homologous genes, and confirmed that variation of gene expression in the nautilus is richer than that in the squid. We, then, searched transcription factors related to eye development and found that key TF for lens, lhx, was expreesed in Nautilus, but loss of six3/6 in Nautilus might be essential to the evolution of pinhole eye.

P-2299

Divergent gene expression patterns produce highly similar aggressive phenotypes in male and female cichlid fish (*Julidochromis*)

Suzy Renn
Reed College, Portland OR, USA

Julidochromis marlieri and *Julidochromis transcriptus* are two closely related Tanganyikan cichlids that have evolved different behavior and mating strategies since they diverged from their common ancestor. While *J. transcriptus* follows the ancestral pattern of male dominance, male biased sexual size dimorphism, and territoriality, the pattern is reversed in *J. marlieri*. In *J. marlieri,* females show all of these behavioral and morphological characteristics. This raises the question of whether female *J. marlieri* achieve the dominant phenotype by expressing the same genes as *J. transcriptus* males, or whether novel brain gene expression patterns have evolved to produce a similar behavioral phenotype in the females of *J. marlieri* . This study used cDNA microarrays to investigate the evolution of gene expression and whether female *J. marlieri* and male *J. transcriptus* show conserved or divergent patterns of brain gene expression. Analysis of microarray data in both species showed that there are certain gene expression patterns associated with sex-role independent of gonadal sex, and to a lesser extent, gene expression patterns associated with sex independent of sex-role. Overall, the data suggest that *J. marlieri* females have undergone significant changes in gene expression patterns potentially linked to the evolution of female-biased aggressive behavior.

P-2300

Evidence for positive selection on a number of microRNA regulatory interactions during recent human evolution including regulation of a pigmentation gene

Jingjing Li[1], Yu Liu[1], Xiaofeng Xin[1], TaeHyung Kim[1], Eduardo Aguiar Cabeza[1], Jie Ren[1], Rasmus Nielsen[1,2], Jeffrey Wrana[1], Zhaolei Zhang[1]
[1]University of Toronto, Toronto, Canada, [2]UC Berkeley, Berkeley, USA

MicroRNA (miRNA) mediated gene regulation is of critical functional importance in animals and is thought to be largely constrained during evolution. However, little is known regarding evolutionary changes of the miRNA network and their role in human evolution. Here we show that a number of miRNA binding sites display high level of population differentiation in humans and thus are likely targets of local adaptation. In a subset we demonstrate that allelic differences modulate miRNA regulation in mammalian cells, including an interaction between miR-155 and TYRP1, an important melanosomal enzyme associated with human pigmentary differences. We identify alternate alleles of TYRP1 that induce or disrupt miR-155 regulation and demonstrate that these alleles are selected with different modes among human populations, causing a strong negative correlation between the frequency of miR-155 regulation of TYRP1 in human populations and their latitude of residence. We propose that local adaptation of microRNA regulation acts as a rheostat to optimize TYRP1 expression in response to differential UV radiation. Our findings illustrate the evolutionary plasticity of the microRNA regulatory network in recent human evolution.

P-2301

Large-Scale Transcriptome Sequencing and Gene Analyses in the Crab-Eating Macaque (*Macaca fascicularis*) for Biomedical Research

Young-Hyun Kim[1,2], Sang-Je Park[1,3], Jae-Won Huh[1], Kyu-Tae Chang[1,2]
[1]National Primate Research Center, Korea Research Institute of Bioscience and Biotechnology, Chungbuk, Republic of Korea, [2]University of Science & Technology, National Primate Research Center, KRIBB, Chungbuk, Republic of Korea, [3]Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan, Republic of Korea

As a human mimic, the crab-eating macaque (*Macaca fascicularis*) is an invaluable non-human primate model for biomedical research including translational studies for drug safety and efficacy testing. However, their genetic information concerning gene structures and gene expression profiles has not been well investigated. Here, we sequenced the transcriptome of 16 tissues and identified genes to resolve the main obstacles for understanding the biological response of the crab-eating macaque. From 4 million reads with 1.4 billion base sequences, 31,786 isotigs containing genes similar to those of humans, 12,672 novel isotigs, and 348,160 singletons were identified using the GS FLX sequencing method. Approximately 86% of human genes were represented among the genes sequenced in this study. Additionally, 175 tissue-specific genes were identified, 81 of which were experimentally validated. In total, 4,314 alternative splicing (AS) events were identified and analyzed. Intriguingly, 10.4% of AS events were associated with transposable element (TE) insertions. Finally, investigation of TE exonization events and evolutionary analysis were conducted, revealing interesting phenomena of human-specific amplified trends in TE exonization events. This report represents the first large-scale transcriptome sequencing and genetic analyses of *M. fascicularis* and could contribute to its utility for biomedical research and basic biology.

P-2302

Identification and characterization of rhesus monkey genes from placenta full-length cDNA libraries: comparative analysis with human genes

Sang-Je Park[1,2], Young-Hyun Kim[1,3], Jae-Won Huh[1], Sang-Rae Lee[1], Kyu-Tae Chang[1,3]
[1]National Primate Research Center, Korea Research Institute of Bioscience and Biotechnology, Chungbuk 363-883, Republic of Korea, [2]Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Republic of Korea, [3]University of Science & Technology, National Primate Research Center, KRIBB, Chungbuk 363-883, Republic of Korea

Rhesus monkeys (*Macaca mulatta*) are widely-used as experimental animals in biomedical research and are closely related to other laboratory macaques, such as cynomolgus monkeys (*Macaca fascicularis*), and to humans, sharing a last common ancestor from about 25 million years ago. However, their gene information are insufficient to understand the specific experimental conditions. Therefore, we constructed the cDNA libraries of rhesus monkey using placenta tissue and sequenced for the identification and characterization of full-length genes. Totally, 1835 cDNA sequences longer than 100 bp were collected and among these sequences, 106 mRNA transcripts sequence having a structural variation in comparison with humans were determined. Among them, six genes (BCS1L, CRYBA2, GSTT1, IER3, MDK, and OXSM) are selected and experimentally validated using 5' and 3' RACE and RT-PCR methods. Especially, BCS1L gene shows the Macaca lineage specific transcripts derived by integration events of AluY family. And the results of CRYBA2 gene indicated various alternative splicing events including differential start sites and exon skipping. Although, we could not sequence and analyze the 1835 cDNA clones, our results indicates that there were differential alternative transcripts and different coding sequences between human and rhesus monkey derived by lineage specific integration event of transposable element, alternative start sites, exon skipping, single nucleotide substitution, etc.

P-2303

Bioinformatic Analysis and Experimental Validation of Fusion Gene in Human, Rhesus Monkey, and Crab-Eating Monkey

Jae-Won Huh[1], Sang-Je Park[1,2], Young-Hyun Kim[1,3], Kyu-Tae Chang[1,3], Sang-Rae Lee[1]
[1]National Primate Research Center, Korea Research Institute of Bioscience and Biotechnology, Chungbuk 363-883, Republic of Korea, [2]Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Republic of Korea, [3]University of Science & Technology, National Primate Research Center, KRIBB, Chungbuk 363-883, Republic of Korea

Fusion gene is a hybrid gene transcript created by the merging of two separate genes. Therefore, most of fusion gene have been identified and characterized as oncogenes, specifically in the leukemia patients through the chromosome translocation events. However, few fusion genes are have been identified in the normal condition, rarely. In our study, we tried to identify the fusion gene in normal condition using bioinformatic tools. And also we investigated the evolutionary conservation of fusion gene for the understanding the biological roles. In our bioinformatic analysis, twenty-eight fusion genes were identified, and experimental validation confirmed the presence of nine fusion genes (SIAE-SPA17, HS2ST1-LOC339524, IDS-LOC100131434, NHEJ1-SLC23A3, MIA-RAB4B, RLN3-IL27RA, PSD-FBXL15, PAOX-MTG1, and DONSON-CRYZL1) in normal human tissue samples. And we applied the nine human fusion genes in the rhesus and crab-eating monkeys for the validation of evolutionary conservation. Remarkably five and four fusion genes are experimentally confirmed in rhesus and crab-eating monkey, respectively. From our results of the presence and conservation of normal fusion gene, we could assume that fusion gene is one of the intriguing evolutionary factors which could enhance the transcriptome diversity under limited genomic sources.

P-2304

Age-related LincRNAs in human brain

Zhisong He[1], Haiyang Hu[1], Philipp Khaitovich[1,2]
[1]CAS-MPG Partner Institute for Computational Biology, Shanghai, China,[2]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

In the past few years, thousands of long intergenic non-coding RNAs (lincRNAs) have been discovered in humans and other species. Although functions of lincRNAs remain largely unknown, several lincRNAs have been shown to act as developmental regulators. To gain an insight into the roles of lincRNAs in human brain development and aging, we sequenced human brain transcriptome at 14 stages of development and aging distributed over the entire human lifespan using high throughput transcriptome sequencing (RNA-seq).

More than 400 annotated human lincRNAs showing evident expression in human brain were identified, over half of which changed their expression significantly with age. Based on their expression patterns, we inferred lincRNAs functions and potential regulators. Further, we identified both cis- and trans- regulatory effects of these age-related lincRNAs on expression of protein-coding genes. Using transcriptome sequencing data from chimpanzee and macaque brains, we further tested evolutionary conservation on the lincRNA repertoire and age-related expression patterns across primate species. Overall, our results demonstrate functionality of human lincRNA in brain and show their evolutionary dynamics and conservation.

P-2305

Positively charged residues affect the speed of translation much more than non-optimal codons

Catherine Charneski, Laurence Hurst
University of Bath, Bath, UK

Changes in the rate of translation are important for understanding protein synthesis, folding, error attenuation and protein localization. Using data from a ribosomal footprinting assay in yeast, we examined changes in ribosomal density across and within mRNAs and found that positive charges added to a protein significantly slow ribosomes in an additive manner. This is thought to be due to the positively-charged amino acids interacting with the negatively-charged ribosomal exit tunnel (Lu, Deutsch 2008). We also find that codon usage bias is a much poorer predictor of ribosomal density than charge, which is at odds with the common conception that it is an important predictor of translation speed. The slowing of ribosomes by positive charges may have implications for the evolution of the poly-A tail.

P-2306

Adaptive evolution of *Pseudomonas aeruginosa* revealed by archetypal analysis of gene expression data

Juliane Thoegersen, Lars Jelsbak, Soeren Molin
Technical University of Denmark, Lyngby, Denmark

Analysis of global gene expression by DNA microarrays can be used to study adaptive evolution of bacteria. However the complexity of such high-dimensional data sets makes it difficult to fully understand the underlying biological features present in the data. In order to cope with this high complexity different dimension reduction techniques have been applied to find patterns in such data sets. Principal component analysis is often used for unsupervised pattern recognition in gene expression data but the results can be difficult to interpret biologically.

The aim of this study is to introduce a method for DNA microarray analysis that provides an intuitive interpretation of data through dimension reduction and pattern recognition. Thus we present the first "Archetypal Analysis" of global gene expression. Archetypal analysis finds characteristic patterns between samples in a high-dimensional data set by introducing a few representative prototypes called "archetypes" in the data set. Each sample is then described as a combination of these archetypes. Archetypal analysis is an unsupervised method that benefits both from the strength of clustering as in k-means clustering and from matrix factorization as in principal component analysis. The analysis is based on microarray data from four integrated studies of *Pseudomonas aeruginosa* in the cystic fibrosis lung.

Our analysis clusters samples into distinct groups with comprehensible characteristics since the archetypes representing the individual groups are closely related to samples present in the data set. Significant changes in gene expression between different groups can identify adaptive changes of the bacteria to the cystic fibrosis lung. The analysis suggests a similar gene expression pattern between isolates with a high mutation rate (hypermutators) despite accumulation of different mutations for these isolates. This suggests positive selection in the cystic fibrosis lung environment, and changes in gene expression for these isolates are therefore most likely related to adaptation of the bacteria. Genes showing significantly changed expression in these isolates include several metabolic genes. This pattern for hypermutators could neither be identified by k-means clustering nor by principal component analysis.

Archetypal analysis succeeded in identifying adaptive changes of *P. aeruginosa.* The combination of clustering and matrix factorization made it possible to reveal minor similarities between different groups of data, which other analytical methods failed to identify. We suggest that this analysis could be used to complement current methods used to analyze DNA microarray data.

P-2307

EVOLUTION OF MRNA EXPRESSION AMONG MICE, MEN, FISH AND FLIES

Samuel Loftus, Douglas Crawford
University of Miami, Miami, FL, USA

Complex 1 (NADH: ubiquinone dehydrogenase) is made of 45 separate subunits: 38 encoded by the nuclear and 7 by the mitochondrial genomes. This enzyme is the in the Oxidative Phosphorylation pathway and plays a crucial role in ATP synthesis. Defects in the enzyme complex are implicated in 70-80 percent of mitochondrial diseases. We examine the patterns of mRNA expression among mammals, teleost fish and the arthropod insect *Drosophila melanogaster.* For Complex I mRNA expression patterns, there is a surprising amount of significant correlations among subunit mRNA expression. Unlike our expectation, there are too many negative correlations among Complex I subunits within human and fish. Although statistically there are too many significant correlations, the pairs of correlated genes are not often common among species, except for the phylogenetically oldest subunits which are often correlated with each other regardless of species. These data suggest that if the variation in mRNA expression is important for the functional variation in Complex I, it is most likely to occur in the basal subunits.

P-2308

Mitochondrial genotypes alter the nuclear transcriptional response to varied levels of hypoxia in Drosophila: Mito-Nuclear Epistasis meets Systems Genetics

David Rand, Patrick Flight, Nick Jourjine, Lei Zhu
Brown University, Providence, Rhode Island, USA

The function of the eukaryotic genome requires proper coordination between three dozen mitochondrial genes and over 1000 nuclear genes. This partnership has coevolved for over 2 billon years, and this intergenomic cross talk provides a powerful context to study the evolution of gene networks. We present an empirical approach to this problem using Drosophila and hypoxia. Exposure to reduced oxygen tension, or hypoxia, is a common occurrence in a wide array of environmental conditions ranging from cancer to stressed microhabitats in nature. One response to hypoxia is to reduce cellular demand for oxygen by down-regulating mitochondrial functions. Despite the central role mitochondria play in oxygen consumption, the effect of alternative mitochondrial genotypes on the hypoxic response has not been examined in flies. Here we use mtDNA introgression strains of Drosophila to examine the effects of alternative mtDNA-encoded genes on the nuclear transcriptional response to varied hypoxia. Flies carrying mtDNA from either D. melanogaster OreR, D. melanogaster Zimbabwe, D. simulans siI, or D. simulans siII on D. melanogaster OreR nuclear chromosomal background were constructed using balancer substitutions and maternal cytoplasm from these four genotypes. Replicate cultures of adult males of each genotype were exposed to four different oxygen tensions (normoxia, 6%, 3%, and 1.5%) for 2 hours, and flash frozen. Expression profiles were determined using Affymetrix 2.0 arrays. The mtDNA genotype design allows for partitioning effects to alternative mtDNAs within a species, or fixed differences between Dmel and Dsim mtDNAs. MtDNA has subtle effects on gene expression under normoxia (~10 genes altered) and strong hypoxia (1.5%; ~25 genes altered), but had pronounced effects at 3% (>200 genes altered) and 6% (>500 genes altered). These results provide strong evidence for mitochondrial retrograde signaling in the nuclear transcriptional response to different levels of hypoxia. Gene ontology analyses reveal that alternative mtDNAs within species alter genes associated with ‘oxidation-reduction’ and ‘mitochondrion’, while the effects of Dmel vs. Dsim mtDNAs alter very different classes of genes (‘ribonuclear protein’; ‘ribosome biogenesis’). These studies offer the first evidence that genes in mtDNA play a critical role in modulating the nuclear transcriptional response to hypoxia, and the different scales of mtDNA divergence have distinct effects on the transcriptional network underlying this response.

P-2309

The evolution of gene expression in numerous cancer lines

Corey M. Hudson[1], Michaël Bekaert[0,2], Gavin C. Conant[1]
[1]University of Missouri, Columbia, MO, USA, [2]University of Stirling, Stirling, Scotland, UK

The evolution of gene expression provides a valuable frame of reference for explaining the development and progression of cancer. Many tissue types radically alter their gene expression profile after becoming oncogenic. We evaluate this change in gene expression in several cancer lines in order to disentangle the influence of differentiated gene imprinting, proliferative reprogramming, as well as cancer-specific expression profiles. We do this by comparing these cancer lines to the gene expression of the associated differentiated cell types as well as the gene expression profiles for proliferative human embryonic stem cells. Given that in tissue cancer lines are in competition with one another, as well as with the host cells for space and resources, we adopt an explicitly evolutionary approach. We simulate the relative fitness of various genotypes and expression profiles for different cell types using multiple metabolic flux prediction algorithms, including Flux Balance Analysis, Flux Variability Analysis and Minimization of Metabolic Adjustment. These approaches ultimately allow us to computationally evaluate the relative metabolic fitness for each of the potentially numerous empirically determined cell types in a given tumor.

P-2310

Exploring the roles of molecular evolution, expression patterns and function of posterior HoxA and HoxD genes in teleost paired-fin diversity: novel observations from non-model taxa.

Sophie Archambeault, Karen Crow-Sanchez
San Francisco State University, San Francisco, CA, USA

There are over 27,000 species of teleost fishes with an astounding array of fin shapes, sizes, positions and specialized modifications, yet the molecular evolution and development of paired fins in teleost fishes is not well understood. There are few well-characterized systems for fin development; the best example is that of the larval pectoral fin bud of the zebrafish, *Danio rerio*, a basal teleost fish. However, many fishes, including zebrafish, form a larval pectoral fin, which transforms into an adult fin later in development. This morphogenesis has not been fully characterized for gene expression, yet must be involved in the growth and patterning of the diversity of adult fin structures observed in teleosts. Hox genes are associated with the evolution of novelty and diversity, and are known to play crucial roles in the growth and development of fins and limbs. Hox gene expression has been well characterized in the development of the limb buds in lobe-finned fishes, the early pectoral fin bud of zebrafish and *Polyodon spathula*, a basal ray-finned fish. Comparisons of this work indicate that the early development of the fin and limb buds share many similarities across vertebrates, while the autopod is a novel feature of tetrapods, and corresponds to a novel late-phase hox gene expression pattern. In addition, mutant limb phenotypes have been linked to modifications in Hox gene sequences in mice and humans. We compare the molecular evolution, gene expression patterns and functional divergence of Hox genes during pectoral fin morphogenesis and pelvic fin development, two processes that have not been well characterized. We sequenced the posterior HoxA and HoxD paralogs of 25 teleost fishes and tested for evolutionary trends correlating to adult fin morphology. In addition, we characterized Hox gene expression patterns during fin morphogenesis and pelvic fin development in zebrafish and compared it to several derived percomorph fishes with modified pectoral and/or pelvic fins. We conclude that both molecular evolution and divergence of expression patterns have aided in the evolution of diversity of teleost paired fins.

P-2311

Gene expression analysis in neural tissues of the Chordate, *Ciona intestinalis* by the use of a customized microarray

Hiromi Matsumae[1], Mayuko Hamada[2], Manabu Fujie[2], Hiroshi Tanaka[1], Yoshihito Niimura[1], Takeshi Kawashima[2]
[1]Department of Bioinformatics, Tokyo Medical and Dental University, Tokyo, Japan, [2]Marine Genomics Unit and Genome Sequencing Center, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan

Ciona intestinalis is a key organism to study evolution of chordates. Draft genome of *C. intestinalis* becomes avaibale in 2002 (Dehal et al. 2002). To analyze gene expression profiles for non-model organisms, microarray technology is still an available tool due to costs and accessibility of analysis. Now we designed a new customized microarray for *C. intestinalis* on the NimbleGen 135k platform (NCBI GEO: GPL14686) based on newly published gene models, KH models(Satou et al. 2008). We will discuss design strategy of a custimzed microarray for non-model organisms. Moreover, I will demosntrate gene expression profiles in neural tissues of mature adult *C. intesitnalis*.

P-2312

Characterizing Gene Expression Variation Across Seven Diverse Human Populations

Alicia Martin[1], Helio Costa[1], Jeffrey Kidd[2], Brenna Henn[1], Muh-Ching Yee[1], Stephen Montgomery[1], Howard Cann[3], Michael Snyder[1], Carlos Bustamante[1]
[1]Stanford University School of Medicine, Department of Genetics, Stanford, CA, USA, [2]University of Michigan School of Medicine, Department of Human Genetics, Ann Arbor, MI, USA, [3]Foundation Jean Dausset, Centre d'Etude du Polymorphisme Humain, Paris, France

Genetic variation has been studied across diverse human populations, but our understanding of its impact on phenotypic variation is limited without extending these studies to determine the effect of variation on gene expression. Genome-wide mRNA sequencing (RNAseq) studies in individual populations have yielded insights into natural variation in mRNA levels, isoform diversity, and novel transcripts—connecting gene expression differences to complex phenotypes. However, a complete understanding of human transcriptome variation requires examining many populations from a wide range of biogeographic ancestries. In order to elucidate how expression changes occurred throughout historical human migrations, we have integrated the genome and transcriptome sequences of 45 lymphoblastoid cell lines from seven populations within the Human Genome Diversity Project that represent the full spectrum of human migration history. These populations include the San Bushmen of southern Africa, Mbuti Pygmies of central Africa, Mozabites of north Africa, Pathans of central Asia, Cambodians of east Asia, Yakut of Siberia, and Mayans of Mexico. This approach allows us to perform a comparative study using the single-nucleotide resolution of RNAseq to assess rare transcripts, novel gene structures, alternative splicing, allele-specific expression, and differential expression within and among populations. We have quantified reads for known exons, transcripts and whole genes and have employed a novel statistical approach for identifying systematically differentially expressed genes among populations. Preliminary results suggest that on average, over 7,000 are expressed in each population. Further, we have identified several transcripts that are differentially expressed by population. In the future, we aim to analyze population specific splice variants, which we believe will be consistent with drift and/or selection. For example, the Mbuti Pygmies are shorter in stature on average than other populations. Eighty percent of the variation in height among individuals is due to genetic factors, so we expect to find differential expression among candidate height genes in Pygmies compared to other populations. We also expect that under a neutral model, the number of population-specific transcripts will be proportional to divergence time as measured by time to most recent common ancestor (TMRCA), and that deviations from this model will be suggestive of selection. Our dataset will allow a detailed investigation of the landscape of human transcriptome variation in diverse human populations.

P-2313

Stress and polyphenism: an RNA-Seq analysis in the butterfly *Bicyclus anyana*

Christopher Wheat[1,2]
[1]*University of Helsinki, Helsinki, Finland,* [2]*Stockholm University, Stockholm, Sweden,* [3]*Leiden University, Leiden, The Netherlands,* [4]*University of California, Irvine, Irvine, USA*

Many organisms are adapted to their local conditions through phenotypic plasticity, allowing them to modify their phenotypes in response fluctuating environments. However, little is known about the mechanistic basis of such plastic phenotypes, and how they are related to heritable differences still segregating within populations for related phenotypes. Using deep RNA-Seq data from individual abdomens and thoraces, we assessed the genetic variation underlying significant phenotypic differences among larvae to nutritional limitation in: development time, pupal mass, relative allocation to abdomen, fat percentage, and resting metabolic rate. Our primary focus was to quantify stress responses in dry and wet seasonal morphs, and the potential family level differences in these responses. We find a range of gene expression differences associated with the observed phenotypic differences and interpret our findings in relation to previous gene expression investigations of polyphenism in this species.

P-2314

Transcriptome profiling studies in spruce trees reveal gene family evolution, high levels of diversity in gene expression and a novel insect resistance mechanism

Jukka-Pekka Verta[1,2], Nathalie-Suzett Delvas[2], Elie Raherison[1,2], Isabelle Giguère[1,2], Sébastien Caron[1,2], Claude Bomal[1,2], Éric Bauce[2], Christian Landry[1], John MacKay[1,2]
[1]Institute of integrative and systems biology, Laval University, Québec, Québec, Canada, [2]Center for forest research, Laval University, Québec, Québec, Canada

Studying species with life habits and evolutionary histories that are distinct from model organisms may afford untapped opportunities for uncovering unique adaptations and novel mechanisms, in addition to broadening our understanding of evolutionary processes.
Our research uses spruce trees (Picea spp.) as representatives of the conifers, which also include pines, firs, and other northern hemisphere forest trees. Conifers are long-lived and outbred species, which dominate many terrestrial ecosystems. They are evolutionary ancient and have among the largest genomes (18 to 30 gigabases). A gene catalogue comprised of 27,700 unique Picea glauca sequences and large-scale transcriptome profiling methods were developed to study the evolution of gene families, expressional diversity and adaptive mechanisms.
We present an approach for genetic analyses of expression variation that directly measures the segregation of gene expression in the haploid meiotic products of single diploid individuals. We discovered abundant Mendelian variation in gene expression, surpassing two fold the levels found in other organisms. We associated this variation with putative cis and trans variants. We analyzed associations of expression variation with specific gene characteristics and observed that expression variation was biased towards genes involved in interacting with the environment, and duplicated genes. Our results suggest that analyses of wild species such as P. glauca may reveal new parameters that affect the nature of genetic variation in gene expression networks.
A naturally-occurring resistance phenotype was discovered for a major defoliating insect, the spruce budworm Choristoneura fumiferana (Clem.). Resistant trees accumulate two phenolic compounds to relatively high concentrations while non-resistant trees accumulate glycosylated forms (Delvas et al. 2011 Entomol. Exper. Applic. 141:35). Transcriptome profiling showed that transcripts encoding a putative glycosyl hydrolase accumulate to high levels specifically in resistant trees. Further evidences indicated the compounds are toxic to the insect and that gene product is likely required for their synthesis. These findings are important because little knowledge has been developed of the tree's defense mechanisms owing to the cyclical nature and variable spatio-temporal scale of insect outbreaks. This approach has enabled the discovery of a novel insect resistance mechanism directly in a wild population of trees and opens the doors to investigations of chemical ecology.

P-2315

Characterizing the X-autosome difference in regulation of testes-specific genes in *Drosophila melanogaster*

Emily Landeen, Colin Meiklejohn, Daven Presgraves
University of Rochester, Rochester, NY, USA

Heteromorphic sex chromosomes have evolved independently numerous times across taxa, often accompanied by the evolution of sex-chromosome specific content, organization and regulation. Our previous work showed that, in *Drosophila*, sex chromosome dosage compensation is absent in the testis, and there is little evidence for meiotic sex chromosome inactivation. There is however evidence for a X-autosome difference in the regulation of testis-specific genes. Transgenic reporters driven by testis-specific promoters inserted onto the X and autosomes reveal a strong average difference in regulation: sex-linkage affects the magnitude of expression achieved by testis-specific promoters, a difference that appears to be established before meiosis. As these findings depend on gene expression from transpon-vectors, we sought to assay the expression of endogenous genes. Therefore, to begin to understand how the X chromosome is regulated in the *Drosophila* male germline, we compared expression of testes-specific and non-specific endogenous genes in the testes of males bearing synthetic X-autosome transpositions and X-autosome translocations. We consider the evolutionary significance of our findings for sex chromosome evolution and the evolution of gene expression in the male germline.

P-2316

Characterization of the olfactory transcriptome in the European eel *Anguilla anguilla*

Allison Churcher[1], Weiming Li[2], Scot Libants[2], Alessandro Coppe[3], Lorenzo Zane[3], Adelino Canário[1], Mar Huertas[1]
[1]Centre of Marine Science, University of Algarve, Faro, Portugal,[2]Michigan State University, Michigan, USA,[3]University of Padova, Padova, Italy

The European eel *Anguilla anguilla* is a teleost with a complex lifecycle that includes transoceanic and freshwater migrations in separate life stages. Several aspects of eel biology combined with their highly developed sense of smell suggest that chemical communication is involved at key life stages (e.g. the migration over the Atlantic Ocean and complex reproductive cycle of the eel). While a few behavioral studies have addressed the role of chemoreception in eel biology much remains to be understood about the molecular, cellular and physiological mechanisms involved in eel olfaction. The first objective of this study was to identify chemosensory receptors and specific components of the olfactory signal transduction pathway in *A. anguilla*. For this, RNA was isolated from the olfactory epithelium of animals from three life stages. Samples were then used to construct 454 (MGE Technology Platform at MPI Berlin) and Illumina (Beijing Genome Institute) sequence libraries. We also used a combination of bioinformatics approaches to search available transcriptome data for *A. anguilla*. Our preliminary results indicate that the *A. Anguilla* genome encodes at least 18 odorant receptors from the rhodopsin-like GPCR family and 10 chemosensory receptors from the glutamate GPCR family, one type 1 vomeronasal receptor as well as other genes involved in olfactory signal transduction (e.g. cyclic nucleotide-gated channels, adenylate cyclase and phospholipase C). The next step is to construct an assembly for the olfactory epithelium transcriptome. This assembly will be used as a reference to look for differential expression in pre-spermiating freshwater and seawater males and sexually mature males. We expect that this approach will help us to identify receptors and other genes that facilitate this transition from saline to freshwater environments and will direct physiological investigations on eel communication by identifying genes involved in olfaction (e.g. odorant receptors) and by helping to characterize the routes of integration in the nervous system from a functional and anatomical perspective. Furthermore, the results from this study will expand our knowledge of both receptors and genes involved in eel olfaction and will facilitate the identification of orthologous genes in other basal teleosts. This research was supported by the Portuguese Fundação para a Ciência e a Tecnologia.

P-2317

Muller's ratchet and hybrid origins of apomixis: high quality SNP discovery in the Ranunculus auricomus complex

Marco Pellino, Diego Hojsgard, Thomas Schmutzer, Elvira Hörandl, Timothy Sharbel
IPK institute, gatersleben, Germany

Asexual reproduction has evolved repeatedly and independently from sexual ancestors in many species of animals and plants. Apomixis (asexual seed formation) is found naturally in more than 400 plant species, and is characterized by three developmental steps: the production of a meiotically unreduced egg cell (*apomeiosis*), *parthenogenetic* development of this egg cell without fertilization, and production of a functional endosperm with (*pseudogamy*) or without (*autonomous*) fertilization of the binucleate central cell of the ovule.

The *Ranunculus auricomus* complex is an interesting model system for comparative evolutionary genomic analyses of sexual and asexual species. It comprises hundreds of agamospecies that are divided into two main subcomplexes which constitute morphologically distinct groups (*R. auricomus* and *R. cassubicus*). The "cassubicus" subcomplex is a well defined subgroup of sexual diploid and apomictic polyploid cytotypes, and two sexual and one apomictic taxa have been studied in detail. *Ranunculus carpaticola* is a sexual diploid species widespread in the Carpathians and central Slovakia, and the closely related sexual autotetraploid *R.cassubicifolius* is distributed in lower Austria and Switzerland. The apomictic taxon is hexaploid and present in central Slovakia, and shows an intermediate morphology between the 2 sexual taxa. Population level analyses of DNA markers suggest that diploid *R. carpaticola* and tetraploid *R. cassubicifolius* hybridized to form the hexaploid apomicts in Slovakia.

In the present work we describe SNP discovery and comparative analyses of Illumina-sequenced ESTs from sexual and apomictic genotypes of the *R. auricomus* complex. Using a bioinformatics pipeline developed in our institute, we have identified a large number of high-quality SNPs which are being used to analyse: (1) mutation accumulation and its correlation with asexual reproduction (Muller's ratchet); and (2) signatures of hybridization in the apomictic genome.

P-2318

Accelerated Recruitment of New Brain Development Genes into the Human Genome

Yong Zhang[1,2], Patrick Landback[1], Maria Vibranovski[1], Manyuan Long[1]
[1]Department of Ecology and Evolution, University of Chicago, Chicago, USA, [2]Institute of Zoology, Chinese Academy of Sciences, Beijing, China

How the human brain evolved has attracted tremendous interests for decades. Motivated by case studies of primate-specific genes implicated in brain function, we examined whether or not the young genes, those emerging genome-wide in the lineages specific to the primates or rodents, showed distinct spatial and temporal patterns of transcription compared to old genes, which had existed before primate and rodent split. We found consistent patterns across different sources of expression data: there is a significantly larger proportion of young genes expressed in the fetal or infant brain of humans than in mouse, and more young genes in humans have expression biased toward early developing brains than old genes. Most of these young genes are expressed in the evolutionarily newest part of human brain, the neocortex. Remarkably, we also identified a number of human-specific genes which are expressed in the prefrontal cortex, which is implicated in complex cognitive behaviors. The young genes upregulated in the early developing human brain play diverse functional roles, with a significant enrichment of transcription factors. Genes originating from different mechanisms show a similar expression bias in the developing brain. Moreover, we found that the young genes upregulated in early brain development showed rapid protein evolution compared to old genes also expressed in the fetal brain. Strikingly, genes expressed in the neocortex arose soon after its morphological origin. These four lines of evidence suggest that positive selection for brain
function may have contributed to the origination of young genes expressed in the developing brain. These dat a demonstrate a striking recruitment of new genes into the early development of the human brain.

P-2319

Detection of WGDs by modeling duplication dynamics

Kevin Vanneste[1,2], Yves Van de Peer[1,2], Steven Maere[1,2]
[1]VIB, Ghent, Belgium, [2]UGent, Ghent, Belgium

Whole Genome Duplications (WGDs) have received a lot of attention because they have been documented in many different species and are hypothesized to provide massive amounts of raw genetic material that can be employed for evolutionary innovations and adaptations[1]. The exact number and timing of WGDs, and whether there is a true causal link with evolutionary innovations, is however still heavily debated. Our goal is to present a method to accurately infer the occurrence of WGDs from age distributions of duplicated genes, where they manifest themselves as peaks against a small-scale duplication (SSD) background[2], and study their effects on paranome expansion. The interpretation of duplicate age distributions is however complicated by the use of K_S, the number of synonymous substitutions per synonymous site, as a proxy for the age of paralogs. Our group therefore constructed a K_S simulation model that simulated both the SSD and WGD components of duplicate age distributions allowing for better inference of WGDs, but still lacking accuracy in its tail[3]. Of particular concern are the stochastic nature of synonymous substitution, leading to increasing uncertainty in K_S with age since duplication; and K_S saturation caused by the inability of evolutionary models to fully correct for the occurrence of multiple substitutions at the same site. Increasing K_S uncertainty is expected to erode the signal of older WGDs, while K_S saturation may lead to artificial peaks in the distribution. We therefore investigated the consequences of these effects on K_S-based age distributions and WGD inference. We adapted a popular codon model of evolution[4] and employed an artificial evolution approach to simulate species-specific sequences according to predefined real age distributions after which we re-estimated the corresponding K_S-based age distributions. We demonstrate that, although K_S estimations can be used for WGD inference beyond the commonly accepted K_S threshold of 1, K_S saturation effects indeed can cause artificial peaks at higher ages. We argue that for detecting WGDs from age distributions, these effects need to be properly accounted for, and the failure to do so could lead to false WGD inferences. We are currently working to incorporate these effects into our K_S simulation model to allow more accurate inferences of WGDs.

[1] Fawcett et al. 2009. PNAS 106, p5737.

[2] Blanc and Wolfe 2004. Plant Cell 16, p1667.

[3] Maere et al. 2005. PNAS 102, p5454.

[4] Goldman and Yang 1994. MBE 11, p725.

P-2320

Evidence for birth-and-death evolution of a secondary metabolite biosynthetic gene cluster and its relocation within and between genomes of the filamentous fungus *Fusarium*

Robert Proctor[1], François Van Hove[2], Antonia Susca[3], Gaetano Stea[3], Mark Busman[1], Theo van der Lee[4], Cees Waalwijk[4], Todd Ward[1], Antonio Moretti[3]
[1]U.S. Department of Agriculture, Peoria, Illinois, USA, [2]Université catholique de Louvain, Louvain-la-Neuve, Belgium, [3]National Research Council, Bari, Italy, [4]Plant Research International, Wageningen, The Netherlands

In fungi, genes required for synthesis of secondary metabolites are often clustered. The *FUM* gene cluster is required for synthesis of a family of toxic secondary metabolites, fumonisins, produced by some fungi of the *Gibberella fujikuroi* species complex (GFSC). Among GFSC species, the *FUM* cluster is discontinuously distributed but uniform in gene order and orientation. Here, analyses of phylogenetic relationships and synonymous site divergence provide evidence for amplification of the cluster within the ancestor of the GFSC and subsequent loss and sorting of paralogous clusters in a manner consistent with the birth-and-death model of multigene family evolution. The results also indicate that the cluster has relocated multiple times within GFSC genomes and has undergone horizontal transfer from GFSC to another *Fusarium* lineage. Thus, despite conservation of gene organization within the *FUM* cluster, the evolutionary history of the cluster in *Fusarium* has been complex.

P-2321

Experimental Characterization of Urzymes Supports a Sense/Antisense Origin of Amino Acid Activation by Class I and II Aminoacyl-tRNA Synthetases

Charles Carter, Li Li, Srinivas Niranj Chandrasekharan, Martha L. Collier
University of North Carolina at Chapel Hill, Chapel Hill, USA

Uncatalyzed amino acid activation by ATP is by far the slowest chemical step in protein synthesis. ATP mobilization by early aminoacyl-tRNA synthetases (aaRS) was thus a key requirement for codon-directed protein synthesis. We have therefore investigated the polypeptide ancestors of contemporary aminoacyl-tRNA synthetases. Superposing structures for 10 Class I aaRS onto that of *B. stearothermophilus* TrpRS allowed us to identify and assemble a disjoint, 130 residue subset containing the active site (1, 2). This construct has 65% of the transition state stabilization free energy of the full-length dimeric enzyme. We call it an "Urzyme" from Ur = "primitive, original" plus enzyme (1, 2).

Rodin & Ohno, observing that class-defining peptide signatures in the two aaRS Classes are encoded by nearly complementary sequences, proposed that the earliest Class I and II aaRS were encoded by complementary strands of the same gene (3). Their proposal seems difficult to test experimentally. However, long stretches of contemporary aaRS that cannot be aligned sense/antisense are incompatible with the hypothesis, suggesting *similar, testable predictions with respect to the two ancestral aminoacyl-tRNA synthetases:* deleting non-aligned domains and fusing disjoint segments should produce stable molecules that accelerate cognate amino acid activation. We have now shown that the observed catalytic activities for both TrpRS and HisRS (4) Urzymes are authentic. Because *the most highly conserved sequences and secondary structures in these enzyme superfamilies also retain considerable enzymatic activity and specificity,* these Urzymes are valid models for ancestral enzymes, providing the first experimental evidence favoring the hypothesis of sense/antisense origins.

Recapitulating evolutionary ascent by successively restoring modules deleted in the Urzymes shows that catalytic proficiency increases monotonically with mass. As structural hierarchies from 3D alignments lead to comparable hierarchies of catalytic activities, successive evolutionary innovations could have entailed *parallel* increases in proficiency, ($k_{cat}/K_M/k_{Non}$), assuring comparable chemical reaction rates throughout evolution (5). Amino acid specificity, however, requires subtler synergies between restored modules. Supported by GM78227.

References

1. Pham Y*, et al.* (2007) *Mol. Cell* 25:851-862.

2. Pham Y*, et al.* (2010) *J. Biol. Chem.* 285.

3. Rodin SN & Ohno S (1995) *Orig. Life Evol. Biosph.* 25:565-589.

4. Li L*, et al.,* (2011) *J. Biol. Chem.* 286:10387-10395.

5. Radzicka A & Wolfenden R (1995) *Science* 267:90-93.

P-2322

Extensive intron gain in the ancestor of placental mammals

Dusan Kordis
Josef Stefan Institute, Department of Molecular and Biomedical Sciences, Ljubljana, Slovenia

Background

Genome-wide studies of intron dynamics in mammalian orthologous genes have found convincing evidence for loss of introns but very little for intron turnover. Similarly, large-scale analysis of intron dynamics in a few vertebrate genomes has identified only intron losses and no gains, indicating that intron gain is an extremely rare event in vertebrate evolution. These studies suggest that the intron-rich genomes of vertebrates do not allow intron gain. The aim of this study was to search for evidence of *de novo* intron gain in domesticated genes from an analysis of their exon/intron structures.

Results

A phylogenomic approach has been used to analyse all domesticated genes in mammals and chordates that originated from the coding parts of transposable elements. Gain of introns in domesticated genes has been reconstructed on well established mammalian, vertebrate and chordate phylogenies, and examined as to where and when the gain events occurred. The locations, sizes and amounts of *de novo* introns gained in the domesticated genes during the evolution of mammals and chordates has been analyzed. A significant amount of intron gain was found only in domesticated genes of placental mammals, where more than 70 cases were identified. *De novo* gained introns show clear positional bias, since they are distributed mainly in 5' UTR and coding regions, while 3' UTR introns are very rare. In the coding regions of some domesticated genes up to 8 de novo gained introns have been found. Intron densities in Eutheria-specific domesticated genes and in older domesticated genes that originated early in vertebrates are lower than those for normal mammalian and vertebrate genes. Surprisingly, the majority of intron gains have occurred in the ancestor of placental mammals.

Conclusions

This study provides the first evidence for numerous intron gains in the ancestor of placental mammals and demonstrates that adequate taxon sampling is crucial for reconstructing intron evolution. The findings of this comprehensive study challenge the current view on the evolutionary stasis in intron dynamics during the last 100 – 200 My. Domesticated genes could constitute an excellent system on which to analyse the mechanisms of intron gain in placental mammals.

P-2323

Molecular evolutionary analysis of human *AQP7* and the pseudogenes

Koyuru Inui, Yoko Satta
The Graduate University for Advanced Studies (Sokendai), Hayama, Kanagawa, Japan

Water is essential to living organisms. “AQUAPORIN” (AQP) works as a water channel or glycerol channel (molecules of similar small size as glycerol were included) in the membrane, and almost all living organisms possess it. A single AQP molecule can express in various tissue or cells and it plays slightly different roles depending on its expressed site. In humans, the *AQP* gene family consists of 13 members (*AQP0~AQP12*). AQP7 is involved in transportation of water and glycerol. AQP7, therefore, has been believed to plays an important role of lipid metabolism on adipose cells. Human *AQP7* subfamily is composed of five members: one is a functional gene and four are pseudogenes (*AQP7Ps*:*AQP7P1, AQP7P2, AQP7P3, and AQP7P4*). A previous study has reported that pseudogenization in *AQP7* occurred after *AQP7* duplication but the detailed process of this pseudogenization is still unclear. In the present study, using the molecular evolutionary analysis we found that human *AQP7* pseudogenization has two independent causes. One is frameshift in exon7 of *AQP7P1* and the other is nonsense mutation in exon2 of *AQP7P2~P4*. Although a functional *AQP7* and four pseudogenes are all located on chromosome 9, neighboring genes of them are different. For *AQP7* they are *AQP3* and *NFX1*, while those for pseudogenes are *BMS1* and *MTHFD1L*. Both copies of *BMS1* and *MTHFD1L* are located in various chromosomes in the genome. Phylogenic analysis of *AQP7, AQP7Ps, BMS1* and *MTHFD1L* revealed that *AQP7* duplicated and generated proto-*AQP7P* 17 million years ago (mya) and duplication (translocation) to chromosome 9 of *BMS1* occurred around 9 mya, and duplication of *MTHFD1L* occurred almost simultaneously. Moreover, the result shows that the proto-*AQP7P* duplicated again with *BMS1* and *MTHFD1L* 3.1 mya and diverged into proto-*AQP7P1* and proto-*AQP7P2~P4.* The loss of function occurred by a mutation independently in these proto-genes 3.1~1.5 mya. The timing of gene duplication and pseudogenization might be related to emergence of genus *Homo*.

P-2324

Genome-scale analyses of non-bilaterian metazoans clarify opsin evolution and the origin of vision in animals

Roberto Feuda, Sinead Hamilton, James O. McInerney, Davide Pisani
N.U.I Maynooth, Kildare, Ireland

Opsins are universally used by the Eumetazoa to detect light. Accordingly, there is great interest in the evolution of the opsin family, as this will improve our understanding of the evolution of vision as well. However, there is no consensus on the relationships of the main opsin subfamilies: the C-, R- and Go/RGR-opsins, particularly with reference to known, cnidarian-specific opsins. Here we studied opsin evolution in the non-bilaterian metazoan: the Cnidaria, the Placozoa and the sponges, and considered also the still unpublished genome of the homoscleromorph sponge *Oscarella carmela*.

We show that the sister group of the eumetazoan opsins is composed by a group of placozoan sequences of unknown function. Among the functionally characterised GPCRs, the closest opsin outgroups are the Melatonin (MLT) receptors. Taxonomic distribution within the MLT and the opsin families implies that the placozoan sequences are, from an evolutionary point of view, true opsins (irrespective of their function). The closest sponge paralogues of the eumetazoan opsins are members of a large and poorly resolved group, which includes also the MLT + Opsin clade. The cnidarian opsins split in three groups representing the orthologues of the eumetazoan C-, R-, and Go/RGR-opsins. Overall, our results imply a novel, maximally parsimonious and highly plausible scenario of opsin evolution: the Opsin family originated from the duplication of the MLT plus opsins ancestor in the Eumetazoa plus Placozoa stem lineage. Two further duplications in the eumetazoan stem lineage separated the R-opsins from the C- plus Go/RGR lineage, and the C- from the Go/RGR-opsins, respectively. This implies the first animal with at the least one opsin belonging to each of the three main subfamilies was a stem eumetazoan. Ancestral character state reconstruction showed that the Last Common Opsin Ancestor (LCOA) did not have a Lysine at position 269 (K269). Instead, it most likely had a Methionine. K269 is ubiquitously used by the eumetazoan opsins to link retinal: the photoreceptor, and the LCOA might not have been a light-sensitive protein. Light detection, or at least K269 mediated light detection, most likely evolved through a process of neofunctionalisation in the stem eumetazoan lineage.

P-2325

Evolutionary causes and consequences of losing key chaperone genes

Tobias Warnecke
Centre for Genomic Regulation, Barcelona, Spain

Some genes families are highly dynamic, frequently gaining and losing members over the course of evolution. Others, in contrast, are remarkably static. In fact, some genes are almost never lost, to the extent that they are considered universal for a certain domain of life and attributed key roles in genome function.

Finding such putatively universal genes to be absent from a particular lineage is intriguing, not least because it reshapes our ideas about what is minimally required to build a working, independently replicating organism. Losses of quasi-universal genes, for example of certain tRNA synthetases, have mostly been documented for highly reduced genomes of endosymbionts and obligate pathogens, i.e. in a context where missing functionality might be provided by the host. Here, I consider a case where a core genetic module has been lost in a free-living organism.

Surveying complete bacterial genome sequences, I show that the Dnak-DnaJ-GrpE (KJE) chaperone system has been lost – likely in a single ancestral event - from the genomes of *Thermovibrio ammonificans* and *Desulfurobacterium thermolithotrophum*, two chemolithotrophic bacteria isolated from deep-sea hydrothermal vents.

KJE functions at the fulcrum of several protein folding pathways in bacteria, being involved in the folding of nascent proteins that emerge from the ribosome as well as in the refolding of denatured proteins. Its loss in these two hyperthermophiles is puzzling given that DnaK in particular is highly conserved in their closest relatives, some of which grow at even higher temperatures.

What might be the genomic changes that enabled KJE loss? Did specific client proteins or the proteome as a whole evolve to become less dependent on chaperone-assisted folding? Or is there, perhaps, evidence for a radical re-organization of protein folding pathways that might compensate for the absence of KJE?

Analyzing KJE substrates in a phylogenetic context, I find little support for systematic changes in protein features that might facilitate or mitigate the loss of KJE. However, comparing genomic chaperone repertoires across the order *Aquificales*, some intriguing difference emerge. Notably, the *T. ammonificans* genome codes for prefoldin (alpha subunit), a protein of archaeal origin, which is implicated in the folding of nascent proteins and might functionally substitute for DnaK.

Based on this and further evidence, I suggest that gene exchange with archaea might have enabled a radical deviation from the archetypal bacterial folding organization towards a KJE-less state that is reminiscent of chaperon systems in hyperthermophilic archaea.

P-2326

A non-cording DNA region is an origin of three primate orphan gene families on the sex chromosome.

Mineyo Iwase[1], Hielim Kim[2], Naoyuki Takahata[1], Yoko Satta[1]
[1]*The Graduate University for Advanced Studies (Sokendai), Hayama,Knagawa, Japan,* [2]*The Pennsylvania State University, 417 Old Main University Park,Pennsylvania, USA*

Eukaryote genomes contain a large number of “orphan” genes that do not have recognizable homologues in other species and that are likely to be involved in important species-specific adaptive processes. Orphan gene might arise from duplication and rearrangement processes although the sequence similarity at this moment is too low to detect a homologous relationship with a parental gene. Recently, non-coding genomic regions has been noticed as an important mechanism for de novo generation of a new gene.

To investigate the processes and mechanism for how the primate orphan gene family originates from non-cording DNA region we performed the evolutionary study of primate Variable Charge *X/Y* (*VCX/Y*), the primate orphan gene on sex chromosomes.

The phylogenetic analysis of the retrieved sequences reveal that *VCX* sequences are divided into two distinct region concerning of the evolutionary origin. In the stem lineage of mammals, there was a segment homologous to the 5’ half of the extant *VCX* gene (proto-*VCX)*. However, this proto-*VCX* sequence is non-coding and was duplicated several times on an X-chromosome. Later, one of these duplicated sequences acquired the 3’ half part of *VCX* from somewhere on the genome. This particular fused gene is an ancient *VCX* and this fusion occurred before the simian and prosimian divergence. Then species-specific gene duplication of *VCX* occurred on X chromosome in the lineage leading to macaque, orangutan, chimpanzee and humans, respectively.

Surprisingly, proto-*VCX* genes (the 5’ part of present *VCX*) are included in *SPANX* (The sperm protein associated with nucleus in the X chromosome) and CSAG (Chondrosarcoma Associated Gene) genes. Both *SPANX* and *CSAG* consist of a relatively large gene family. Furthermore, these analyses also revealed that the boundary of 5’ and 3’ part in three genes included commonly NAHR (non-allelic homologous recombination hot spot common motif; 5’-CCTCCCT-3’). These data support the idea these orphan gene families are derived from ancestral non-coding sequence and NAHR may play a key role in the emergence of novel gene.

P-2327

Chemosensory protein gene family evolution in ants

Jonna Kulmuni[1,2], Yannick Wurm[3], Pekka Pamilo[4]
[1]*University of Oulu, Oulu, Finland,* [2]*Biocenter Oulu, Oulu, Finland,* [3]*Queen Mary, University of London, London, UK,* [4]*University of Helsinki, Helsinki, Finland*

Ants and other social insects have sophisticated chemosensory abilities because they use chemical communication in social context. In addition to signals used by other animals, such as those involved in mate and habitat choice, ants are able to communicate for a mass of individuals in order to co-ordinate the actions of the whole colony. We studied how this complexity is reflected in the evolution of one gene family involved in peripheral olfactory processing. More specifically, we looked at the chemosensory protein (CSP) gene family repertoires and their evolution in seven sequenced ant species. CSP's are similar to odorant binding proteins (OBP) in that they are small, soluble binding proteins and secreted into the lymph of insect chemosensory sensilla, but are also expressed elsewhere in the body. In ants, one of the CSP's has been shown to bind the substance used in nest-mate recognition. We found that the CSP gene family has expanded in ants (11 to 21 genes) compared to honey bee (6 genes) and *Drosophila* (4 genes). CSP gene family contains several orthologous genes among ants as well as with honey bee. The fire ant *Solenopsis invicta* represents one extreme with 21 intact genes and 4 pseudogenes, where the increase in gene number results from lineage-specific expansions. Especially interesting is an expansion of six genes, where one of the members is the major protein expressed in *S. invicta* antennae and could be the protein used in binding nest-mate recognition substances. Preliminary estimates suggest that the birth-and-death rate (λ) for the CSP gene family is higher (0.0033) than the genome average for example in *Drosophila* (0.0012) but lower than that of OBP (0.005), OR (0.006) or GR (0.011) gene family. Birth-and-death rate of the CSP gene family is higher in social than in other insects. Both the ant-specific expansions and the higher birth-and-death rate possibly reflect the importance of the CSP gene family in the extended chemosensory abilities of ants.

P-2328

The fate of *S-RNase* gene duplications in Rosaceae

Bruno Aguiar[1], Jorge Vieira[1], Ana Cunha[1], Nuno Fonseca[2], Cristina Vieira[1]
[1]IBMC, Porto, Portugal,[2]CRACS-INESC, Porto, Portugal

Self-incompatibility systems allow the pistil to recognize and reject pollen from genetically related individuals (known as self-pollen). The most common variant, S-RNase-based gametophytic self-incompatibility (GSI), in which the pollen tubes from a self-pollen grain are degraded by the action of an S-RNase, is thought to have evolved only once before the split of Asteridae and Rosidae, about 120 million years ago. By analyzing the S-RNase lineage sequences from different families, amino acid patterns characteristic of this gene lineage have been identified, which allowed the identification of S-RNase lineage sequences in the Cucurbitaceae, Fabaceae, Euphorbiaceae, Malvaceae, Rubiaceae, Solanaceae, Plantaginaceae and Rosaceae families. However, molecular studies on the GSI have only been performed in the latter four, with different evolutionary patterns being observed within species of the same family, like in *Prunus* (Amygdaloideae) and *Malus* or *Pyrus* (Pyrinae), all belonging to the Rosaceae family. Moreover, different mechanisms have also been proposed for the interaction between the *S*-pistil specificity (an S-RNase) and the *S*-pollen specificity (an F-box). In *Prunus*, there is only one F-box protein determining *S*-pollen specificity, which is thought to protect the self S-RNases from being inhibited by a general inhibitor, thus remaining active. In *Malus* and *Pyrus* however, there are several F-box proteins determining the *S*-pollen specificity, that are proposed to interact with a subset of non-self S-RNases marking them for degradation by the ubiquitin-26S-proteasome system. It is thus possible that the system, although ancestral, has evolved from different gene duplications in the two genes.
When considering the amino acid patterns that are typical of the *S-RNase* lineage genes, a total of 15 and 5 genes were found in the *Malus domestica* and *Prunus persica* genomes respectively. Phylogenetic analyses of these genes as well as subsets of *S-RNases* from the different plant families support the hypothesis that different *S*-duplicates have been recruited for the GSI function in Rosaceae. Similar analyses of the flanking genes also support this hypothesis. Moreover, since *S-RNase* genes are expressed in pistils only, tissue expression patterns of these genes are being addressed in order to understand the fate of an *S*-duplicate.

P-2329

Role of low-complexity regions in protein formation and functional diversification

Núria Radó-Trilla[1], Macarena Toll-Riera[1], Florian Martys[1], M. Mar Albà[1,2]
[1]Evolutionary Genomics Group, Fundació Institut Municipal d'Investigació Mèdica (FIMIM)-Universitat Pompeu Fabra (UPF), Barcelona, Spain, [2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

A large number of eukaryotic proteins contain regions of very simple amino acid composition, known as low-complexity regions (LCRs). Several neurological disorders, including Huntington's disease, are caused by an uncontrolled expansion of triplet repeats encoding LCRs. The high abundance of LCRs is striking considering their high pathogenic potential. Why are they so abundant?

One possibility is that they facilitate the formation of novel functions in existing proteins. Experimental data has shown that some LCRs can modulate transcriptional activity (Gerber et al., 1994) or are important for protein subcellular localization (Salichs et al, 2009). We have tested this idea by examining gene duplicates originated at the base of the vertebrates. We have measured the LCR content of 154 human transcription factor gene families. Interestingly, 31% of the duplicated genes in our dataset contain LCRs, but only 17% of the non-duplicated genes do. In addition, LCRs in gene families are often copy-specific and show a significant tendency to have been formed soon after the duplication event. This strongly suggests that LCRs contribute to the rapid gain of gene copy-specific functions (neofunctionalization). We are currently performing experiments to quantify the effects of copy-specific LCRs on transcriptional activation.

Another interesting possibility is that LCRs play a key role in the formation of novel protein-coding genes. The reason is that the expansion of triplets (or any multiple of 3 nucleotides unit) by slippage may help generate novel long coding sequences. We have tested this hypothesis by examining the LCR content of proteins of different age. We have found that, recently emerged, mammalian-specific proteins contain double the amount of LCRs than older proteins (Toll-Riera et al., 2012). In addition, some LCRs in young proteins extend over the complete length of the protein and play important lineage-specific functions. These data supports the idea that LCRs play a key role in the emergence of novel functional genes.

Gerber, H-P. et al. "Transcriptional activation modulated by homopolymeric glutamine and proline stretches". Science 1994, vol. 263, pp- 808-811.
Salichs, E., Ledda, A., Mularoni, L., Albà, M.M. and de la Luna, S. "Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment."

P-2330

Context-dependent evolution of protein domains and their repeats

Dan Reshef[1], Liran Carmel[2], Ora Schueler-Furman[1]
[1]Department of Microbiology and Molecular Genetics, Institute for Medical Research Israel-Canada, Hadassah Medical School, The Hebrew University of Jerusalem, Jerusalem, Israel,[2]Department of Genetics, The Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Jerusalem, Israel

Many proteins are built of several modular building blocks called domains. In particular, some domains or domain combinations may appear in repeats. In this large-scale study, we compare the sequence conservation at both the DNA and protein level of domains within the context of single domain proteins, multi-domain proteins, and multi-repeat proteins, we identify different constraints that act upon each context and thus shape the evolution of a domain in a distinct way. We find that domains that form part of a functional unit often show increased levels of conservation, among others at the interface with the other domains, suggesting that the interaction of this domain with the other domains within the functional context constraints the sequence divergence more than in single domain proteins. Such differences in conservation can improve functional annotation of domains within different contexts, and shed light on evolutionary forces that act on the DNA and protein sequence level to form higher order unit, as well as unit repeats.

P-2331

Duplication dynamics and functional innovation in the Rab family of small GTPases

Yoan Diekmann, Jose B. Pereira Leal
Instituto Gulbenkian de Ciência, Oeiras, Portugal

The eukaryotic endomembrane system is believed to have evolved its extant complexity by expansion of gene families already present in the last common ancestor. These include the Rab family of small GTPases, essential proteins that regulate all steps of vesicular trafficking.
Despite considerable efforts to understand the structure and cell biology of Rabs and their binding partners called effectors, little is known about how Rab functions evolve after gene duplication.

Here, we analyze duplication dynamics and the functional diversification of the Rab gene family, and its impact on intracellular trafficking and organization in Metazoa.

We developed and validated a tool for accurate identification and classification of Rabs, and applied it to perform a detailed annotation and analysis of the Metazoan Rab family. Both our tool and dataset are available at www.RabDB.org. We find for example an important expansion and specialization of the secretory pathway at the base of Metazoa.

Crossing phylogenetic analysis and data on tissue specificities, intracellular localization and biochemical function, we show that the animal Rab family expands frequently by neofunctionalization.

We pinpoint key mutations for functional divergence by detection of site-specific rate shifts, and analyze their effects on protein stability, enzymatic activity and protein interactions using bioinformatic methods. Finally, we scan Rab-effector pairs for coevolving sites. Our findings suggest that early disruption of interactions with effectors is key to evolution of new Rab functions in Metazoa.

In conclusion, using Rabs as a model system this work contributes to establish a mechanistic basis for the process of intracellular evolutionary innovation.

P-2332

Non-neutral evolution of *fatty acid desaturase* F pseudogenes in *Drosophila*

Shun-Chern Tsaur[1], Kuang-Hsi Chu[2], Wei-Chin Ho[2], Wan-Ju Shen[3], Kevin Wei[3], Chau-Ti Ting[2,4], Shu Fang[3]
[1]Department of Life Sciences & Institute of Genome Sciences, National Yang-Ming University, Taipei, Taiwan,[2]Institute of Zoology, National Taiwan University, Taipei, Taiwan,[3]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan,[4]Department of Life Science & Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan

Pseudogenes have long been recognized to evolve neutrally due to the lack of functional constraints. Yet, a pseudogene may be subject to selection if a loss-of-function allele results in adaptive changes. In *Drosophila,* loss-of-function of *fatty acid desaturase F* (*desatF*) has been demonstrated to be adaptive, and the pseudogenization happened independent in multiple lineages. It provides a means for understanding how selection acting on the pseudogenes at the molecular level. To reveal the mechanism of selection during pseudogenization, we focused on the nucleotide polymorphism of 5 *desatF* pseudogenes in *D. melanogaster* species subgroup. The result showed that *desatF* exhibited a high pseudogenization rate indicated by frameshift-causing indel polymorphism in *D. mauritiana* and *D. yakuba* whereas *desatF* accumulated very few indel polymorphisms in *D. santomea, D. simulans*, and *D. teissieri*. This difference cannot be explained simply by different degeneration time because low nucleotide polymorphisms in the coding region of 3 *desatF* pseudogenes are deviated from the neutral expectation. The reduced polymorphism was further confirmed to be as a result of recent selective sweeps. Our findings demonstrate that genetic hitchhiking would shape the evolutionary pattern of pseudogenization.

P-2333

Adaptation and evolution of the olfactory subgenome in mammals

Graham Hughes, Des Higgins, Emma Teeling
University College Dublin, Dublin, Ireland

Olfaction is a critical form of sensory perception in mammals, used to detect food, avoid predators and as a means of social signaling. Mammalian smell is governed by olfactory receptors (ORs) - G-coupled protein receptors containing no introns that are up to 1kb in length and typically account for 6% of the protein coding genes in mammals (Malnic *et al,* 2004). ORs constitute the largest gene family across mammals and can be split into two Classes. Class I dominates fish OR repertoires while Class II dominates terrestrial animals. These classes can be further divided into 5 subfamilies and 14 subfamilies respectively.

A potential link between adaptation to environment and the loss of function has previously been established (Hayden et al, 2010). In order to further explore the effects of environmental adaptations on the evolution and pseudogenization of various OR genes, we have generated a number of data from aerial, terrestrial and aquatic mammalian species using sequenced genomes, Illumina sequencing, 454 Roche technologies and gene cloning. By combining *de novo* sequence assemblers with gene mining and read mapping, we have determining that in terms of yield and quality, 454 sequencing coupled with long reads is the best method to amplify a large subset of the OR subgenome.

Various paradigm shifts in the history of Mammalia have given rise to new ecological niches. Examples of these include the evolution of frugivory, the evolution of flight and the adaptation to an aquatic environment. Using a variety of phylogenetic tools for selection tests, we aim to understand the selective pressures such adaptations may have had on olfaction. We will explore the loss of OR function in a number of terrestrial and aquatic mammals to elucidate if the pseudogenization of OR genes is a random event or a result of adaptation to new environmental conditions.

Malnic, B., Godfrey, P.A. & Buck, L.B. (2004) The human olfactory receptor gene family. *PNAS*, **101**, 2584 – 2589.

Hayden, S., Bekaert, M., Crider, T.A., Mariani, S., Murphy, W.J. & Teeling, E.C. (2010). Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Research*, **20**, 1-9.

P-2334

Identification of potential chemosensory gene families in funnel web spiders using NGS

Cristina Frias-Lopez[1], Francisca C. Almeida[1], Sara Guirao-Rico[2], Miquel Arnedo[3], Julio Rozas[1]
[1]Universitat de Barcelona, Departament de Genètica, Barcelona, Spain,[2]University of Edinburgh, Institute of Evolutionary Biology, Edinburgh, UK,[3]Universitat de Barcelona, Departament de Biologia Animal, Barcelona, Spain

Chemoreception is a biological process involved in essential aspects of an organism's life, such as the detection of food, egg-laying substrates, mates, and predators. In comparison to their arthropods counterparts, chelicerata chemosensory system is still poorly known. This is an interesting subject because chelicerates and insects colonized lands independently. This raises the question of whether the same set of genes involved in the olfactory system of insects are found in the chelicerata olfactory system, or, alternatively, that a entirely different gene family has evolved to fulfill this function in spiders and allies. In order to identify candidate chemosensory gene families in chelicerata, we sequenced organ specific cDNA libraries of the funnel web spider *Macrothele calpeiana* (Hexathelidae), which incidentally is the only spider protected under European legislation. Subtractive cDNA libraries were obtained from the palps and the first two legs, which according to the literature bear the chemosensitve organs in spiders. We also build a library from ovaries, for the sake of comparison. Using the Roche 454 GS-FLX Titanium technology, we obtained a little over 50,000 reads for each library. The average read length across all 50.8 Mb sequenced, was 352 bp. *De novo* assemblies yielded 1300, 1680, and 833 potential genes for legs, palps, and ovaries, respectively. BLAST searches (blastx, e-value below 1e-3) revealed similarity of about 50 % of the sequences of each library with other chelicerates proteins available in NR database. Identification of gene families was carried out using reciprocal BLAST searches within each library and expression specificity was assessed with searches between libraries. Functional annotation of transcripts was done with the software Blast2GO to identify candidate receptor and ligand-binding genes that could play a role in the chemosensory system of *M. calpeiana.*

P-2335

High prevalence of positive selection in rodent gene duplicates evolving both symmetrically and asymmetrically

Cinta Pegueroles[1], M. Mar Albà[1,2]
[1]*Evolutionary Genomics Group, Fundació Institut Municipal d'Investigació Mèdica (FIMIM)-Universitat Pompeu Fabra (UPF), Barcelona, Spain,* [2]*Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain*

Gene duplication is a major mechanism modeling genomes, contributing to genome evolution and adaptation. Three main evolutionary models have been proposed to explain the preservation of the two gene copies after a duplication event: neofunctionalization (gain of a new function in one of the copies), subfunctionalization (accumulation of degenerate mutations and split of functions) and, increased gene dosage advantage. In principle we should be able to distinguish between these different models by the analysis of the changes in the ratio of non-synonymous versus synonymous substitutions (dN/dS) with respect to the pre-duplication. Accelerated dN/dS in only one of the copies, accompanied by signs of positive selection, would be expected under neofunctionalization, whereas approximately symmetric dN/dS in the two copies would fit the other scenarios.

We have used an exhaustive dataset of recent gene duplicates in Mus musculus and Rattus norvegicus (90 and 218 duplicated genes, respectively) to investigate the relative frequency of asymmetric versus symmetric gene copy evolution, and to determine which is the impact of positive selection in the evolution of duplicates following different models. Human and cow orthologs have been used to be able to estimate dN/dS in both pre-duplication and post-duplication mouse and rat branches. Using a Fisher test and correcting for multiple testing we have found that asymmetric evolution is the most prevalent mechanism (57% asymmetric and 43% symmetric). In most cases of symmetric gene copy evolution we find no dN/dS acceleration with respect to the pre-duplication branch, suggesting sub-functionalization is rare.

Positive selection appears to be very pervasive in both asymmetrically and symmetrically evolving duplicated genes. In the first case the accelerated branch shows significant positive selection in 51% of the cases. Families in which dN/dS in both post-duplication branches is not significantly different to dN/dS in the pre-duplication branch show significant positive selection in 37% of the cases. Interestingly, many of these genes are rapidly evolving genes, as they show higher than average dN/dS not only in the post-duplication branches but in the pre-duplication branches as well. Our findings disagree with the classical view of non accelerated duplicates being strongly constrained by purifying selection.

P-2336

Positive selection and diversification of the V1R gene family in strepsirrhine primates

Nick Mundy[1], Philipp Kappel[1,2], Ute Radespiel[2]
[1]*University of Cambridge, Cambridge, UK,* [2]*University of Hannover, Hannover, Germany*

The V1R (vomeronasal-1 receptor) gene family encodes receptors for pheromones (intraspecific signalling) and kairomones (interspecific signalling) and has undergone spectacular expansion in some mammalian lineages, with positive selection a common feature. Here we use strepsirrhine primates, and in particular mouse lemurs (*Microcebus*), to investigate patterns of selection in V1Rs and their relationship to V1R diversity. Mouse lemurs are small nocturnal primates in which olfactory communication is important and they have an estimated 200 V1R loci. We find that positive selection is common among V1R loci in mouse lemurs, occurring in most lemur-specific V1R clades. In addition, we demonstrate ongoing positive selection in 5 out of 7 single loci resequenced from up to 10 mouse lemur species. Positively selected sites are concentrated in the extracellular region where they might influence ligand binding. However, there is no relationship between strength of selection and V1R clade size, ruling out simple models of diversification and functional conservation. Thus we find pervasive adaptive evolution of V1Rs in mouse lemurs, but variation in strength of selection and factors affecting rate of diversification remain to be explained and this will be an interesting area for further research.

P-2337

Molecular evolution and divergence of vertebrate heat shock factors

Kalle Rytkönen
University of Turku, Turku, Finland

Heat shock factors (HSF) are central regulators of stress responses, cell differentiation and development in all metazoans. HSFs have been studied in invertebrates (Drosophila and Caenorhabditis) that have only one HSF and in vertebrates that have three or more HSF paralogs. To date the evolution and duplication pattern of the vertebrate HSF paralogs has not been studied in detail. Here we analyze the molecular evolution of vertebrate HSFs, predict which specific amino acid changes may have been responsible for the subfunctionalization of the paralogs and when during vertebrate evolution functional changes have occurred. HSF paralogs are tested for the signatures of directional natural selection with maximum likelihood codon substitution models and for functional divergence of specific amino acid sites with distance based protein level models. We will specifically investigate the paralog specific divergence in amino acid sites that are known (in HSF1) to be responsible for post-translational regulation.

P-2338

Evolutionary analysis and natural selection in betaine homocysteine methyltransferase (*BHMT*) and *BHMT2* genes

Radhika Ganu, Yasuko Ishida, Alfred Roca, Timothy Garrow, Lawrence Schook
University of Illinois at Urbana-Champaign, Urbana, IL, USA

The enzymes betaine homocysteine methyltransferase (BHMT) and BHMT2 convert homocysteine to methionine using, respectively, the substrates betaine and S-methylmethionine. Increased homocysteine levels are associated with cardiovascular disease and developmental defects. We compared available sequences of *BHMT* and *BHMT2* genes for 38 species of deuterostomes. *BHMT* was present in the genomes of the sea urchin, amphibians, reptiles, birds and mammals; *BHMT2* was present only across all mammals. *BHMT2* and *BHMT* were present in tandem in the genomes of all monotreme, marsupial and placental species examined. Gene duplication likely occurred in an ancestor to all extant mammals, followed by two deletions in *BHMT2* that encode differences between the two enzymes in oligomerization and in substrate specificity. Evolutionary rates were accelerated for BHMT2 relative to BHMT. Nine codons were found to display signatures of positive selection. Selective pressure varied across lineages, with the highest dN/dS ratios for *BHMT* and *BHMT2* occurring immediately following the gene duplication event.

P-2339

The evolutionary dynamics of Ppp1 catalytic subunits in the Metazoan

Susana Pereira[1], Vítor Vasconcelos[1,2], Agostinho Antunes[1,2]
[1]*CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal,*
[2]*Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal*

Protein phosphatases catalyze the dephosphorylation of proteins in specific residues, counteracting the action of protein kinases and contributing to the regulation of virtually all physiological processes.
Ppp1 is a serine/threonine-specific protein phosphatase encoded by the PPP gene family. These enzymes consist of highly conserved small catalytic subunits that diversified their biological function through interaction with novel regulatory subunits encoded by separate genes. The combinatorial subunit principle of PPP enzymes generates a great amount of diversity and flexibility as the mutually-exclusive binding of different regulatory subunits allows the same catalytic subunit to specifically dephosphorylate a large number of varied substrates.
In this study, we evaluated the molecular evolution of different PPP catalytic subunits, and in particular of Ppp1c, in Metazoan organisms. The study comprehended both gene and protein-level analyses, and clarified the correlation of the catalytic subunits evolutionary dynamics with their interaction with the regulatory subunits.

P-2340

New tools for studying gene family evolution show changes in a single yeast species that correlate with the environment

Simon Lovell, Simon Whelan, Ryan Ames, Daniel Money
University of Manchester, Manchester, UK

Genome-wide variation in copy number in gene families is common. There is increasing evidence that this variation in gene copy number can give rise to substantial phenotypic effects, and in some cases these variations have been shown to be adaptive. These observations show that a full understanding of the evolution of biological function requires an understanding of gene gain and gene loss.

We have developed weighted parsimony and maximum likelihood methods for inferring gene gain and loss events. To test these methods we have used Markov models of gain and loss to simulate data with known properties. We have examined three models: a simple birth-death model, a single rate model, and a birth-death-innovation model. We find that for all simulations maximum likelihood-based methods are very accurate for reconstructing the number of duplication events on the phylogenetic tree, and that maximum likelihood and weighted parsimony have similar accuracy for reconstructing the ancestral state.

We have applied these tools to data from the Saccharomyces Genome Resequencing Project in which the genomes of 38 S. cerevisae strains were sequenced. We find that yeast show an abundance of duplicate genes that are lineage specific, leading to a large degree of variation in gene content between individual strains. There is a detectable bias for specific functions, indicating that selection may be acting to preferentially retain certain duplicates. Most strikingly we find that sets of over and under-represented duplicates correlate with the environment from which they were isolated.

We have also analyzed variation in transcription factor binding sites of these lineage-specific duplicates. We find a large degree of variation both between these closely-related strains and between pairs of duplicated genes. There is a correlation between changes in promoter regions and changes in coding sequences, indicating a coupling of changes in expression and function. We show that (i) the types genes with diverged promoters vary between strains from different environments and (ii) that a subset of duplicated promoters have accelerated divergence that may indicate positive selection.

We conclude that, even within a single species, we can detect duplicates that have the potential for substantial phenotypic differences between strains. In addition, we can identify early signs of selection from the environment acting to alter expression patterns. Thus, functional innovation arises in part from duplicated genes, even within one species, and may reflect adaptation to each strain's natural environment, offering a shortcut to evolutionary adaptation.

P-2341

An evolutionary analysis of miRNA binding sites suggests an important role for genetic drift

Alfred Simkin[1], Fen-Biao Gao[1], Jeffrey Jensen[2,1]
[1]University of Massachusetts Medical School, Worcester, MA, USA,[2]Ecole polytechnique federale de Lausanne, Lausanne, Switzerland

The deep conservation of a large number of miRNAs suggests a persistent and essential role for these small RNA molecules. Some miRNAs, such as let-7, have been demonstrated to be essential and conserved in the regulation of organismal development, governing the same processes in the same targets across virtually all bilaterians. In seeming contradiction with these observations, a large number of similarly well-conserved miRNAs have tissue expression patterns and predicted targets with very little overlap in their respective species. These findings highlight the remaining uncertainty in determining the evolutionary dynamics governing miRNAs. Much of what is known about miRNA-target interactions is based on experimental work conducted on a small subset of targets that are often initially chosen based on a high degree of conservation, while the great majority of potential miRNA targets identified based on sequence complementarity to a 6 basepair predicted binding site are nonconserved and have yet to be examined. In order to better understand the dynamics governing the evolution of miRNA targets, we have assessed the inferred rate of nucleotide substitution in the computationally predicted binding sites for several well-conserved miRNAs in a number of species, compared against substitution rates in all other 6 basepair sequences. We find no significant difference between miRNA binding sites and other 6 basepair sequences. These findings suggest that the evolutionary dynamics governing the turnover of miRNA binding sites in the genome as a whole are not strongly constrained, and that the creation or loss of individual miRNA binding sites is driven largely by genetic drift. As an individual miRNA may regulate hundreds of targets, we speculate that miRNAs may be conserved in spite of these largely neutral dynamics through an ever changing functional subset of drifting targets.

P-2342

Convergent evolution of concerted evolution in a master sex determiner

Eyal Privman[1], Sanne Nygaard[2], Alexandra von Siebenthal[1], Laurent Keller[1], Yannick Wurm[1,3]
[1]University of Lausanne, Lausanne, Switzerland,[2]University of Copenhagen, Copenhagen, Denmark,[3]Queen Mary, University of London, London, UK

Duplication of a gene results in two initially redundant gene copies (paralogs) that may either diverge in sequence and in function or alternatively retain high similarity because of inter-locus recombination known as concerted evolution. While it is well established that divergence of duplicates can lead to phenotypic innovation, little is known about what drives concerted evolution. If there is an adaptive advantage to concerted evolution it can be predicted to independently evolve several times. We provide the first evidence that concerted evolution is adaptive because it evolved twice following independent duplications in ants and bees of the ancestrally single *transformer* (*tra*) locus, a conserved upstream component of the insect sex determination pathway. The novel molecular function of one of the paralogs in the honeybee as the *complementary sex determiner* (*csd*) is a possible source for selection for diversification of tra paralogs. Positive selection is evident at the *csd* of honeybees as well as in some of the ant paralogs. Concomitant with this selection for point mutations and micro indel mutations, concerted evolution may operate to enrich the allele pool of the *csd* locus with the segments of its paralog's diverged sequence. These observations provide a novel type of evidence for an adaptive role of concerted evolution.

P-2343

Evolution and Diversification of Small Heat Shock Protein in Plants

Elizabeth Waters, Bharath Bharadwaj
San Diego State University, San Diego, CA, USA

The small heat shock proteins are a large and diverse family of chaperones. These proteins interact with misfolded proteins during stress and at certain developmental stages and prevent irreversible protein aggregation. The small heat shock proteins are part of the larger chaperone network. It is well established that small heat shock protein structure and function has been conserved across all domains of life (Archaea, Bacteria and Eukarya). It is of interest that the small heat shock proteins underwent a significant diversification in plants. There are 11 subfamilies in angiosperms, none of which are present in green algae or in other groups. It is notable that comparable diversifications did not occur in the other chaperone families (ex. HSP70s) which are part of the same folding network. The goals of our analysis are to identify all small heat shock protein homologs in plant genomes and to understand the patterns of diversification and sequence evolution within this protein family. We have examined all of the available land plant (Embryophyta) complete genomes and have conducted extensive phylogenetic, and structural analysis of this complex family. We evaluate this data (over 500 protein sequences) in light of expression analysis of these genes in well-studied angiosperms. We have identified novel sHSP subfamilies and have determined the times of origin of these and the other well-studied sHSP subfamilies. We have found that while the moss *Physcomitrella patens* has over 20 different sHSP genes only 3 of the 11 sHSP subfamilies known in angiosperms are present in this moss. In addition, we have found that some but not all of the subfamilies have undergone lineage specific expansions. We discuss these results in relation to evolution at the organismal level. Finally we discuss our analysis of the small heat shock proteins and relate this to the more general challenges of determining orthology and paralogy in complex gene families. To this end we discuss our use of structural and expression analysis to distinguish types of homologous proteins.

P-2344

The details in the Devils: horizontal transfer of transposable elements in a carnivorous marsupial

Sarah Schaack
Reed College, Portland, OR, USA

We recently identified several interesting cases of horizontal transfer of transposable elements (TEs) involving the exchange of DNA from 4 different TE families among up to 11 divergent eukaryotic lineages on multiple continents. This added to a growing number of cases of horizontal transfer of transposable elements (HTT) that have now been documented in eukaryotes, indicating this type of DNA exchange is not unique to prokaryotes. The mechanisms underlying transfer among eukaryotes, however, are likely to be more diverse and complicated. The identification of each new case sheds light on how and with what frequency such exchange occurs and is incorporated into the germline. Here, we report a new case of horizontal transfer in the Tasmanian Devil (*Sarcophilus harrisii), including a characterization of the expansion of the TE family after HTT occurred, estimates of the timing of the event, and a discussion of possible vector.*

P-2345

Frequent subcellular relocalization after gene duplication during the evolution of the Brassicaceae family

Shao-Lun Liu, Angel Pan, Keith Adams
University of British Columbia, Vancouver, BC, Canada

Gene duplication during plant evolution, by successive rounds of whole genome duplication as well as by smaller scale duplication events, has provided a large reservoir of new genes for the evolution of new functions. Duplicated genes can diverge by changes in sequences, expression patterns, and functions. We studied subcellular localization of gene pairs duplicated during the evolution of the Brassicaceae family, as a study system for evolutionarily recent gene duplicates. We analyzed experimental localization data from fluorescent protein experiments for 125 duplicate pairs in *Arabidopsis thaliana*. We found that 18% of them showed evidence for differential localization, indicating relocalization of one or both duplicates. Many of the cases showed an accelerated rate of amino acid sequence evolution, and a VAMP gene showed evidence for positive selection, consistent with functional diversification to the new subcellular location or neofunctionalization. We identified sequence mutations resulting in neolocalization of four gene products, including loss of an N-terminal localization signal in a serine O-acetyltransferase, mutations at the C-terminus of a protein phosphatase, two amino acid sequence changes near the N-terminus of a calcium dependent protein kinase, and deletions near the C-terminus of a potassium channel protein. We show that two cases of neolocalization have undergone regulatory neofunctionalization in pollen, and two other pairs have undergone functional diversification. This study indicates that relocalization is a common outcome of gene duplication in plants, that it can occur on a relatively short evolutionary time scale during the evolution of a family, that various types of sequence mutations can lead to neolocalization, and that neolocalization can be accompanied by neofunctionalization. Neolocalization after gene duplication within a family is an important mode of duplicate gene divergence that could play a role in family-specific functions and phenotypes.

P-2346

Phylogeny and Nomenclature of Histone Variants

Paul B. Talbert[1], Kami Ahamd[2], Geneviève Almouzni[3], Juan Ausio[4], Frederic Berger[5], Prem L. Bhalla[6], William M. Bonner[7], W. Zacheus Cande[8], Brian P. Chadwick[9], Simon W. L. Chan[10], George A. M. Cross[11], Liwang Cui[12], Stefan I. Dimitrov[13], Detlef Doenecke[14], José M. Eirin-López[15], Martin A. Gorovsky[16], Sandra B. Hake[17], Barbara A. Hamkalo[18], Sarah Holec[5], Steven E. Jacobsen[19]
[1]Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [2]Department of BCMP, Harvard Medical School, Boston, MA, USA, [3]CNRS, Institut Curie, Centre de Recherche, Paris, France, [4]University of Victoria, Victoria, British Columbia, Canada, [5]Temasek Lifesciences Laboratory, National University of Singapore, Singapore, Singapore, [6]Plant Molecular Biology and Biotechnology Group, The University of Melbourne, Parkville, Victoria, Australia, [7]Laboratory of Molecular Pharmacology, CCR, NCI, NIH,, Bethesda, MD, USA, [8]Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA, [9]Department of Biological Science, Florida State University, Tallahassee, FL, USA, [10]Department of Plant Biology, UC Davis / HHMI, Davis, CA, USA, [11]Laboratory of Molecular Parasitology, The Rockefeller University, New York, NY, USA, [12]Department of Entomology, Pennsylvania State University, University Park, PA, USA, [13]Laboratoire de Biologie Moléculaire et Cellulaire de la Différenciation, Institut Albert Bonniot, Grenoble, France, [14]Department of Biochemistry, University of Goettingen, Goettingen, Germany, [15]Department of Cellular and Molecular Biology, University of A Coruna, A Coruna, Spain, [16]Department of Biology, University of Rochester, Rochester, NY, USA, [17]Center for Integrated Protein Science Munich at the Adolf-Butenandt Institute, Department for Molecular Biology, Ludwig-Maximilians-University Munich, Munich, Germany, [18]Department of Molecular Biology & Biochemistry, University of CA, Irvine, CA, USA, [19]Howard Hughes Medical Institute, Department of Molecular Cellular and Developmental Biology, University of California, Los Angeles,, Los Angeles, CA, USA

Histone variants are non-allelic protein isoforms that play key roles in diversifying chromatin structure. The known number of such variants has greatly increased in recent years, but the lack of naming conventions for them has led to a variety of naming styles, multiple synonyms, and misleading homographs that obscure variant relationships and complicate database searches. We propose here a unified nomenclature for variants of all five classes of histones that uses consistent but flexible naming conventions to produce names that are informative and readily searchable. The nomenclature builds on historical usage and incorporates phylogenetic relationships, which are strong predictors of structure and function. A key feature is the consistent use of punctuation to represent phylogenetic divergence, making explicit the relationships among variant subtypes that have previously been implicit or unclear. We recommend that by default new histone variants be named with organism-specific paralog-number suffixes that lack phylogenetic implication, while letter suffixes be reserved for structurally distinct clades of variants. For clarity and searchability, we encourage the use of descriptors that are separate from the phylogeny-based variant name to indicate developmental and other properties of variants that may be independent of structure.

P-2347

140 million years of multigene family evolution: Origin and recombination history of the avian MHC class IIB

Reto Burri[1,2], Marta Promerová[3], Luca Fumagalli[2]
[1]Department of Ecology & Genetics, Uppsala University, Uppsala, Sweden, [2]Department of Ecology & Evolution, University of Lausanne, Lausanne, Switzerland, [3]Institute of Vertebrate Biology, AS CR, Brno, Czech Republic

The relative importance of the birth-and-death process versus recombination-driven concerted evolution and their impact on the reconstruction of the origins of genes are long-standing question in the study of multigene families. Seminal studies by Nei and coworkers have established the major histocompatibility complex (MHC) of mammals as a paradigm of birth-and-death evolution. Meanwhile the avian MHC was suggested to be subject to strong concerted evolution, and to date its long-term evolutionary history largely remains in the dark. Here we present unprecedented insights into the long-term duplication and recombination history of the avian MHC class IIB (MHCIIB). New data from all over the avian phylogeny provide conclusive evidence for a pre-avian origin of two major avian MHCIIB lineages, demonstrating that concerted evolution had far less impact on this class of avian MHC than assumed so far. Nevertheless, both recombination and gene conversion are major players in the evolution of the avian MHCIIB lineages, masking orthologous relationships (i) to various amounts, (ii) in various sequence regions, and (iii) to various extents in different species. It thus appears that birth-and-death and phylogenetically unconstrained concerted evolution have both been acting on the avian MHCIIB, but that their relative importance varies in space, in time, and between species. Together with previous results the present study suggests that both neutral and adaptive evolutionary forces are important in the evolution of the avian MHCIIB, and their interplay remains to be investigated.

P-2348

Determining the evolutionary history of gene families

Ryan Ames[1], Daniel Money[2], Vikramsinh Ghatge[1], Simon Whelan[1], Simon Lovell[1]
[1]University of Manchester, Manchester, UK, [2]University of Kansas, Kansas, USA

Recent large-scale studies of individuals within a population have demonstrated that there is widespread variation in copy number in many gene families. In addition, there is increasing evidence that the variation in gene copy number can give rise to substantial phenotypic effects. In some cases these variations have been shown to be adaptive. These observations show that a full understanding of the evolution of biological function requires an understanding of gene gain and gene loss. Specifically, both duplication and gene loss events must be mapped to the underlying phylogenetic tree if we are to correlate genotypic change with phenotypic change or understand the effects of selection. Accurate, robust evolutionary models of gain and loss events are, therefore, required.

We have developed weighted parsimony and maximum likelihood methods for inferring gain and loss events. To test these methods we have used Markov models of gain and loss to simulate data with known properties. We examine three models: (i) a simple birth-death model in which the rate of birth or death is proportional to the current size of a gene family and where families that fall to 0 members are extinct in that lineage, (ii) a single rate model where gain and loss events of single genes are equiprobable, and (iii) a birth-death-innovation model where birth and death represent natural gain and loss of gene in a family, and innovation represents the (re)gain of a gene family from other sources. The parameters for the birth-death-innovation model were estimated from *Drosophila* genome data.

We find that for all simulations maximum likelihood-based methods are very accurate for reconstructing the number of duplication events on the phylogenetic tree, and that maximum likelihood and weighted parsimony have similar accuracy for reconstructing the ancestral state. Our implementations are robust to different model parameters and provide accurate inferences of ancestral states and the number of gain and loss events. For ancestral reconstruction we recommend weighted parsimony because it has similar accuracy to maximum likelihood, but is much faster. For inferring the number of individual gene loss or gain events maximum likelihood is noticeably more accurate, albeit at greater computational cost. Overall, we find that the accuracy of maximum likelihood is dependent on the underlying probabilistic model used to infer gain and loss and that more work is required to accurately describe the process of gene family evolution.

P-2349

Biological meaning of pseudogenization in human evolution

Yoko Satta
The Graduate University for Advanced Studies (Sokendai), Hayama, Kanagawa, Japan

By definition any pseudogene does not have biological function, and according to the neutral theory they can evolve at the fastest rate of nucleotide substitutions. However, recent studies show that some pseudogenes may not be completely non-functional. For example, a pseudogene acts as a source of siRNA, as found in mice. In addition, there has been an argument for beneficial effect of gene loss on organismal evolution. Olson (1999), in particular, proposed the significant role of function-less (pseudogenization: becoming pseudogene) in evolution in his less-is-more hypothesis. The hypothesis is based on the observation that a large fraction of genetic functions of a genome are dispensable and on the speculation that selection may permit the emergence of pseudogenes in genomes.

The number of so-called human specific-pseudogenes (HSPs) in the genome is more than 100. Interestingly, among those HSPs, about one-forth (25) of cases show the independent pseudogenization in both humans and non-human primates. On the other hand, the number of HSPs for a single copy gene, which does not have closely related paralogs, is 38. In this presentation, I will summarize process of pseudogenization in the human genome and argue their functional contribution to human evolution.

P-2350

Gene conversion as both a boost and drag on diversity in a gene family across genetically divergent populations of the copepod *Tigriopus californicus*

Christopher Willett
University of North Carolina, Chapel Hill, NC, USA

The effect of gene conversion is often to homogenize genes and retard divergence between duplicated genes, but in some contexts it can actually increase the amount of variation in genes. If gene conversion between paralogous genes is a rare enough event on an evolutionary timescale and divergence between these paralogs has occurred, rare gene conversion can introduce variation from one paralogous gene to the other. Here it is shown that in a gene family of aspartate transaminase (GOT) loci in the copepod *Tigriopus californicus* that just such a scenario has occurred. The examination of one moderately divergent pair of paralogous genes across a set of genetically divergent populations of this copepod species shows clearly that gene conversion has led to the unidirectional introduction of a great deal of genetic variation as polymorphism from one paralogous gene to the other gene. This process has occurred in only a subset of the populations examined in this study. This has led to a great deal of variation in the levels of polymorphism and heterozygosity among individuals within and between populations. The heterogeneous levels of gene conversion across this pair of paralogs are also evident in the patterns of divergence among paralogs along the gene. Some regions of the two gene copies show complete identity and others have no discernable identity for some stretches. The evolution of four other paralogous copies of GOT are also examined across this same set of genetically divergent *T. californicus* populations. One other pair that is highly divergent from the pair described above is likely to have undergone gene conversion within the pair but not with the other pair of genes. An examination of the GOT homologs across populations of *T. californicus* has revealed both the creative and constraining roles of gene conversion in the evolutionary process.

P-2351

Evolution of olfactory and gustatory receptor repertoires in Drosophila.

Eri Kudo, Mao Nakamura, Atsushi Ogura, Rumi Kondo
Ochanomizu University, Bunkyo-ku, Tokyo, Japan

Chemosensory stimuli play a crucial role for host selection in insects, including fruit flies. The evolutionary dynamics of chemosensory receptors in *Drosophila* were examined through phylogenetic analyses of chemosensory receptor gene family in 12 *Drosophila* species and comparison of expressed chemosensory receptor repertoires among 8 *Drosophila* species. Most chemosensory receptors originated before the divergence of *Drosophila* genus. Although the number of receptor genes is relatively similar, considerable amount of lineage specific gene gains and losses occurred during the divergence of each species, giving rise to highly diverse receptor repertoires between species. Expansion and contraction of repertoire size might relate to genome size and adaptation to host plant.

P-2352

Contrasted patterns of selective pressures in three recent paralogous gene pairs in the *Medicago* genus (L.)

Joan Ho-Huu[1,2], Joëlle Ronfort[1], Stéphane De Mita[3,5], Thomas Bataillon[4], Jean-Marie Prosperi[1], Nathalie Chantret[1]
[1]*INRA, Montpellier, France,* [2]*CNRS/Université Lyon1, Lyon, France,* [3]*IRD, Montpellier, France,* [4]*BiRC, Aarhus, Denmark,* [5]*INRA, Nancy, France*

Gene duplication is considered as a major evolutionary force, allowing for functional innovations. Several models predicting the evolutionary fate of duplicated genes have been proposed. One way to explore experimentally these models is to identify the evolutionary forces exerted on duplicated genes through molecular evolution studies. Here we report a molecular evolution analysis of three recent pairs of duplicated genes (two pairs of polygalacturonases, *Pg11-Pg3 and Pg11a-Pg11c*, and one pair of auxine transporters like, *Lax2-Lax4*) whose sequences were obtained on 17 species belonging to the *Medicago* genus. Codon evolution models based on ratio between non-synonymous and synonymous rates of nucleotide substitution were applied. We searched for divergent levels of constraints between paralogous genes, for the occurrence of positive selection in either copy and for signatures of transient relaxed constraints following the duplications. Our main result is that selective pressures were different between both paralogs for each studied pair of genes. We found sites under positive selection in *Pg11* but *Pg3* is mainly under purifying selection with one quarter of sites evolving neutrally. The two most recent paralogs, *Pg11a* and *Pg11c* harbour sites under positive selection, but with a different intensity, indicating a faster evolution and potentially the possibility of acquiring new functions. The *Lax2* and *Lax4* genes are both under purifying selection but *Lax4* harbours some sites less constrained. Finally, we also detected an increase of the rate of evolution on the *Lax2-Lax4* pair just after the duplication, which corresponds to the hypothesis of a temporary release of selective pressure after duplication.

P-2353

Evolution of the Rh family genes and the AVPR-OXTR family genes expressing in the kidney

Akinori Suzuki, Hironobu Ikeya, Mari Sugaya, Kaoru Yamashita, Kouhei Endo, Takashi Kitano
Ibaraki University, Hitachi, Japan

RhBG and RhCG belong to the Rhesus (Rh) blood group gene family and are mainly expressed in the kidney. They are transmembrane proteins with 12 membrane spanning domains and are predicted to act as ammonium transporter. AVPR1A and AVPR2 belong to the G-protein coupled receptor family which includes vasopressin receptors and oxytocin receptor (OXTR) and are expressed in the kidney. In the study, we analyzed evolution of the Rh gene and the AVPR-OXTR gene families. Four Rh family genes (RH, RHAG, RHBG, and RHCG) in vertebrates arose from two rounds of genome duplication. An acceleration of the evolutionary rate had been observed in the chicken RHBG. Our analysis suggested that the period of the acceleration predated a common ancestral lineage of birds. Four AVPR-OXTR family genes (AVPR1A, AVPR1B, AVPR2, and OXTR) in vertebrates also arose from two rounds of genome duplication. In the common ancestral lineage of the placental mammals, oxytocin hormone, which consists of nine peptides, had evolved by one amino acid change (I8L) from mesotocin hormone. Subsequently a series of nonsynonymous substitutions were accumulated in its receptor gene to arise the OXTR before the mammalian radiation. Since the oxytocin-OXTR system plays an important role in the uterus during parturition, these amino acid changes might contribute the evolution of the placental mammals.

P-2354

Investigating the expansion of gene families in otherwise reduced microsporidian genomes

Sirintra Nakjang[1], Tom Williams[1], Andrew Watson[1], Peter Foster[2], Eva Heinz[1], Robert Hirt[1], Martin Embley[1]
[1]Newcastle University, Newcastle, Tyne & Wear, UK,[2]Natural History Museum, London, UK

Microsporidia are a group of eukaryotic obligate intracellular parasites related to fungi. Genome sequencing projects have revealed high levels of genome reduction in these organisms, with genome sizes varying between 2.3Mb and 24Mb. Microsporidia possess a reduced protein coding capacity, with the loss of many important metabolic genes. This pattern of gene loss is typical of obligate intracellular parasites, which instead rely on host cells for nutrient acquisition. Despite this background of reductive genome evolution; a limited number of examples of gene acquisition have also been identified in Microsporidia. These include cases of acquisition by novel innovation, lateral gene transfer and lineage specific expansion of gene families by duplication. These acquisitions and expansions are surprising in what are otherwise minimal parasite genomes; and could represent important adaptations, such as interactions with the host cell, that are required for the parasitic lifestyle of Microsporidia. For example, all currently sequenced Microsporidia encode homologues of nucleotide transport proteins that were originally acquired via lateral gene transfer, and have undergone subsequent lineage specific duplication events. Characterisation of the *Encephalitozoon cuniculi* nucleotide transporters has indicated they could play a key role in supplying the parasite with host ATP. We investigated the translated protein sequences of seven Microsporidian species to systematically identify cases of lineage specific genome family expansion and gene aquisition. Such searches may aid in identifying proteins which, similar to the nucleotide transport proteins, could play an important role in the interaction of the parasite and its host cell.

P-2355

The enigma of cADPR/NAADP calcium messenger biosynthesis and the limited taxonomic distribution of the ADP-ribosyl cyclases by computational multi-genome survey

Enza Ferrero, Nicola Lo Buono, Fabio Malavasi
University of Torino, Torino, Italy

Calcium signalling is a universal feature of eukaryotic cells. As cognate ligands of calcium-releasing channels, the twin NAD^+-derived second messengers cADPR (cyclic ADP ribose) and NAADP (nicotinic acid adenine dinucleotide phosphate) are important determinants of intracellular Ca^{2+} mobilization, together with the archetypal messenger IP_3. The ADP-ribosyl cyclases (ARCs) are currently the only enzymes known to synthesize cADPR and NAADP but little is known of their phylogenetic distribution and evolutionary history. We carried out a multi-genome survey of over 1500 genomes from the three domains of life. Our survey greatly expanded the small body of less than 20 published ARCs having retrieved 163 nonredundant ARC protein sequences characterized by the presence of the *ribosyl_hydrolase* protein domain (Pfam02267). Of these proteins, 131 are robustly supported and gene-based. Although high sequence divergence complicated protein-based family tree reconstruction, congruent ARC trees and classification into ARC subfamilies was possible by intron evolution analysis. Unlike the widespread second messengers they purportedly synthesize, the taxonomic distribution of the ARCs was limited and suggested that they are a metazoan innovation with patchy distribution even among animal taxa. This raises interesting questions about the origins and functions of cADPR and NAADP in the three domains of life.

P-2356

Patterns of *KLK2* and *KLK3* diversity in Cercopithecoidea

Patrícia Marques[1,2], Rui Bernardino[3], Teresa Fernandes[3], Víctor Quesada[2], Susana Seixas[1], Belen Hurle[4]
[1]Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal, [2]Department of Biochemistry and Molecular Biology, University of Oviedo, Oviedo, Spain, [3]Lisbon Zoo Veterinary Hospital, Lisbon, Portugal, [4]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

The human *kallikrein* (*KLK*) cluster, located at chromosome 19q13.4, comprises 15 paralogue *KLK* genes. *KLK3* and *KLK2* encode serine proteases with key roles in the semen liquefaction cascade, which is physiologically an important step in male fertility. In primates, previous studies provided evidences for KLK subtracts (semenogelins) being preferred targets of natural selection through mechanisms linked to male fertility and sperm competition. Here we address the evolution of *KLK2* and *KLK3,* by undertaking a systematic analysis of their patterns of intra- and inter-specific diversity in Cercopithecoidea. This taxon was selected given its generally polygamous mating systems that favor sperm competition and natural selection.

Our intra-specific diversity study included the sequencing of all coding exons and selected intronic regions of *KLK2* and *KLK3* in 15 *Macaca fuscata* specimens, to a total of 49kb. Regarding *KLK2* it was possible to identify five polymorphic sites, including two non-synonymous substitutions (G51R and L54P) that might affect protein activity and a frame-shift mutation (L171fsX181). Concerning *KLK3*, we identified eight polymorphic sites, of which one was a non-synonymous benign substitution (D220N). The summary statistics of polymorphism levels for *KLK2* and *KLK3* were dissimilar but not unusual for a standard neutral model.

For an inter-specific diversity survey, we analyzed the orthologue *KLK3-KLK2* sequences for the following species: *Macaca mulatta*, *Macaca fascicularis*, *Papio anubis*, *Chlorocebus aethiops* and *Colobus guereza*. Regarding *KLK2*, we identified two changes expected to affect protein structure and activity. These include a premature stop codon in *M. mulatta* (R109X) and a frameshift mutation, in *P. anubis*, causing an mRNA increment (V247fsX337) which is unlikely to be translated into a KLK2. The previously described mutation of the KLK2 catalytic triad (D120A) in *M. mulatta* was not detected and no evidence for the accumulation of deleterious mutations was observed in *M. fascicularis* and *C. aethiops*. In Cecopithecoidea, the single example of *KLK* deletion was found in *C. guereza*, with a complete loss of *KLK2* and *KLK3*, possibly by two deletions events.

Although preliminary, our results suggest a relaxed evolution of *KLK2* supported by the accumulation of several deleterious mutations in different Cercopithecoidea. Such hypothesis would be consistent with a polygamous mating system, in which the hydrolysis of the semen coagulum may be delayed as part of a post-copulatory competition strategy. The possible loss of function of KLK2, which proteolytically activates KLK3, could delay the release of spermatozoa, in several species, resulting in an adaptive advantage.

P-2357

Evolution of the Kelch domain-containing family of protein phosphatases from plants and alveolate protists

Claudio Slamovits[1], Santiago Mora-García[2]
[1]Dalhousie University, Dept. of Biochemistry and Molecular Biology, Halifax, NS, Canada, [2]Instituto de Investigaciones Bioquímicas "Luis F. Leloir", Buenos Aires, Argentina

Protein phosphatases are involved in essential cellular processes including signal transduction, enzymatic regulation, gene expression control and many others. Genomes of plants, green algae and alveolate protists encode one or more proteins with a C-terminal serine/threonine catalytic domain and an N-terminal domain containing several kelch repeats known as Kelch domain containing protein phosphatase (KPP). BSL1, a KPP from *Arabidopsis thaliana* has been studied experimentally and found to be involved in signaling by brassinosteroid hormones. However, presence of homologs in unicellular eukaryotes suggests that these proteins are also involved in more basic cellular functions. In addition to their relevance for the molecular biology of plant cells, this family constitute a potential chemotherapeutic target for apicomplexan parasites such as *Plasmodium* and *Toxoplasma*. KPP is restricted to two eukaryotic lineages, the viridiplantae and alveolates, and appear to be highly conserved and ubiquitous in each of the two lineages. However, it is not clear if the KPP genes from the two lineages have a common origin or if they resulted from independent fusions of a PPC-type catalytic domain with Kelch-type motifs. In addition, KPP genes experienced several independent events of gene duplication at least in plants, apicomplexans and ciliates. In order to clarify the evolutionary history of this intriguing protein family and help determine their possible roles in plants and apicomplexan parasites we performed a phylogenetic analysis of the KPP family using all available genomic and expressed data from plants, green algae, ciliates, apicomplexans and dinoflagellates. Analysis of protein domain structure, indels and intron conservation in conjunction with the gene phylogeny indicates that the KPP genes from the two lineages have a common origin, possibly moving to alveolates by endosymbiotic or lateral gene transfer. In addition, detailed analysis of selection on the four known plant KPP paralogs shows that duplication and neofunctionalization is driving functional diversification in the KPP family.

P-2358

The inhibition of the *S-RNase* cytoxicity in Rosaceae: the *Prunus* 'one *S*-pollen gene' model and the Pyrinae 'non-self recognition by multiple S-pollen factors' model

Bruno Aguiar[1], Jorge Vieira[1], Nuno Fonseca[2], Olivier Raspé[3], Cristina Vieira[1]
[1]Instituto de Biologia Molecular e Celular (IBMC), University of Porto, Porto, Portugal,[2]CRACS-INESC Porto, Porto, Portugal,[3]National Botanic Garden of Belgium, Meise, Belgium

Self-incompatibility (SI) is a major genetic barrier to self-fertilisation in which the female reproductive cells discriminate between genetically related and non-related pollen, and reject the former. In gametophytic SI (GSI) the pollen is rejected when it expresses a specificity that matches either of those expressed in the style. The *S*-pistil gene product in Rosaceae, Rubiaceae, Solanaceae and Plantaginaceae is an extracellular ribonuclease, called *S-RNase*. Phylogenetic analyses suggest that *RNase*-based GSI has evolved only once, about 120 million years ago. Therefore, similarities are expected when comparing the GSI players in these plant families. In Rosaceae, between 22% (*Prunus)* to 30% (Pyrinae) of the exposed protein surface is made of positively selected amino acid sites. These are, in principle the amino acids that determine the specificity of the GSI reaction. The oldest *S-RNase* specificity lineages, both in Pyrinae and Prunus seem to be about 20 million years old. The number of ancestral lineages and the degree of specificity sharing between closely related species is, however, different in Pyrinae and *Prunus*.
The pollen component, always an F-box protein, has been identified as one gene in *Prunus* (Rosaceae, called *SFB*), but multiple genes in *Malus*, *Pyrus*, (Rosaceae, called *SFBBs*), *Petunia*, and *Nicotiana* (Solanaceae). The *SFB* gene presents the expected evolutionary signatures for the *S*-pollen, namely expressed in pollen only, linkage with the *S-RNase* gene, high levels of synonymous and non-synonymous diversity, as well as positively selected amino acid sites that account for the many specificities present in natural populations. *SFBB* genes are also expressed in pollen only, are linked to the *S-RNase*, but present levels of diversity 10 times lower than those at the *S-RNase* gene. They were suggested as putative *S*-pollen genes, in a system of 'non-self recognition by multiple factors'. Subsets of *SFBB* allelic products interact with non-self *S-RNases*, marking them for degradation, and allowing compatible pollinations.
We performed a detailed characterization of *SFBB* genes in *Sorbus aucuparia* (Pyrinae) to verify three predictions of the 'non-self recognition by multiple factors' model: the number of *SFBB* genes is large to account for the many *S-RNase* specificities; *SFBB* genes are old; and amino acids under positive selection are identified, located in specificity determination regions. We also address how such system could have evolved.

P-2359

A formal definition for syntenic blocks

Cristina G. Ghiurcuta, Bernard Moret
EPFL, Lausanne, Switzerland

Rapid and inexpensive high-throughput sequencing is making available more and more complete genome sequences. Analyzing these genomes presents formidable challenges: even simple pairwise comparisons are hard, since we lack good models for genome structure and evolution. Current approaches are based on the identification of so-called syntenic blocks (genome fragments that present identical or highly similar collections of markers in most of the genomes under study). The identification of such blocks is the first step in comparative studies, yet its effect on final results has not been well studied, nor has any formal, biologically meaningful definition of syntenic blocks been proposed. At present, a syntenic block is simply construed as the output of the synteny tool used, of which there are many -- FISH, Cinteny, ADHoRe and DRIMM-Synteny, to name a few.
Syntenic blocks are in many ways analogous to genes -- in many cases, the markers used in constructing them are genes (as in OrthoCluster). Like genes, they can exist in multiple copies, in which case we could define analogs of orthology and paralogy. However, whereas genes are typically studied at the sequence level, syntenic blocks are too large for that level of detail. It is their arrangement within the entire genome that is the main object of study. Thus the definition and construction of syntenic blocks involves both large- and small-scale evolutionary models.
We focus on an abstract framework and definition that applies to any type of markers and their collection of homology statements. Homology statements come from sequence similarity analysis (such as BLAST); but they can also come from other types of analyses as well as from existing databases.
Whereas most other tools rely on collinearity, that is, a series of markers in the same order and orientation in the genomes, our framework only requires that the set of markers in a syntenic block in one genome must not contain any marker that has no homolog in the set of markers in the corresponding syntenic block in any other genome, regardless of ordering and orientation. Collinearity unnecessarily excludes well documented evolutionary events such as rearrangements, but our framework supports them. Moreover, whereas most existing tools are restricted to pairwise comparisons, our framework explicitly supports multi way comparisons.
Our framework provides a means of comparison for a wide range of implementations as well as for existing tools.

P-2360

Not another alignment program paper: How to improve multiple alignments by selecting the appropiate isoform combination.

José Luis Villanueva-Cañas[2], M Mar Albà[1,2]
[1]Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia, Spain,[2]volutionary Genomics Group, Fundació Institut Municipal d'Investigació Mèdica (FIMIM)-Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

The number of transcripts for each gene stored in the databases grows continuously. These different transcripts are mainly the result of alternative splicing and multiple transcription start sites and they often result in different protein isoforms. In large-scale evolutionary analysis we often wish to automatize processes such as the alignment of coding sequences of orthologous genes or gene families for estimating evolutionary rates. Constructing all possible alignments using the different protein isoform combinations is often not feasible and greatly complicates subsequent analysis. Therefore, first we need to choose one protein isoform for each gene. Which criteria should we use? By default, ENSEMBL, one of the most widely used genome databases, uses the transcript with the longest coding sequence (CDS) from each species to make multiple alignments. However, this may result in the alignment of sequences with very different length and including non-homologous regions. One of the main questions we want to address is if this default method is working reasonably well and if it can be improved.

In our evaluation we defined 3 different methods for transcript selection: LONG, where the longest transcript for each species was chosen, our method PALO (Protein Alignment Optimizer) and a random combination (RAND). The PALO heuristics consists in finding the isoform combination with the smallest difference in CDS length. We obtained alignments for all possible isoform combinations in 3 different datasets (mammalia, vertebrata and eumetazoa one-to-one orthologs) and identified the one with the highest percentage identity (BEST) for each gene in each dataset. We then quantified how often the combination containing the longest sequences (LONG), the combination selected by PALO and the RAND combination, resulted in an alignment with the same scores as BEST. In the about 30% of cases in which LONG and PALO chose different transcript combinations, LONG selected the BEST combination in 16-19% of cases, PALO in 60-71% of cases and RAND in 16-21% of cases.

The LONG method is widely used and has performed reasonably well for a long time, but its performance is decreasing as the number of known isoforms per gene increases. The heuristics employed by PALO resulted in alignments with higher scores than when using the LONG approach. We also found that the use of the LONG method resulted in higher branch-specific non-synonymous to synonymous (dN/dS) substitution rate values, suggesting that it might increase the number of false positives in tests of positive-selection based on divergence data.

P-2361

Indel Reliability in Phylogenetic Inference

Haim Ashkenazy[1], Ofir Cohen[1], Dorothée Huchon[2], Tal Pupko[1]
[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel-Aviv, Israel, [2]Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel-Aviv, Israel

Insertion and deletions (indels) in protein sequences are considered to be rare genomic events and thus it is often assumed that finding the same indel independently in two evolutionary lineages is unlikely. This suggests that indel-based inference of phylogeny should be less subject to homoplasy compared to standard inference based on substitution events. Indeed, indels were recently successfully used to solve debated evolutionary relationships among Metazoa. However, indel based phylogeny may suffer from biases and artifacts that can impede accurate phylogenetic inference. For example, we hypothesized that since indels are never directly observed but rather inferred from the alignment, indel-based inference may be sensitive to the alignment algorithm used. To test this hypothesis, we first determined the level of agreement among different alignment methods (MAFFT, PRANK, and ClustalW). We show that differences among alignments obtained from various methods suffice to generate different indel-based trees. We next developed a method to quantify the reliability of indel characters by measuring how often they appear in a set of alternative multiple sequence alignments. Our approach is based on the assumption that indels, which are consistently present in most alternative alignments are more reliable compared to indels that appear only in a small subset of these alignments. Specifically, for each indel character we assign a reliability score, which is the fraction of sub- and co-optimal alignments in which it is present out of 100 such alignments. We further show that filtering unreliable indels increases the accuracy of indel-based phylogenetic reconstruction. Finally, we conducted a genome scale analysis, in which we characterize unreliable indels (e.g., in terms of the distribution of their length, their position in the protein coding sequence). Taken together, our results show that indel-based inference is sensitive to biases stemming from uncertainty in alignment and that filtering unreliable indels is critical for accurate indel-based phylogeny reconstruction. Our indel reliability program is freely available.

P-2362

Disentangling the phylogenetic signal carried by alignment gaps and guide trees

Salvador Capella-Gutíerrez, Toni Gabaldón
Centre for Genomic Regulation (CRG), Barcelona, Spain

Multiple Sequence Alignment (MSA) plays a central role in modern molecular biology, being used in a broad set of applications from evolutionary studies to the identification of conserved motifs. Alignment optimization is generally based on scoring residue pairings according to a given empirical matrix, and on applying certain gap penalties that are usually rather arbitrary. In the context of phylogenetic reconstruction highly gapped regions have been shown to be unreliable and is common practice to eliminate them prior to the analysis. However, recent results claim that patterns of gaps in alignments are phylogenetically informative. One possible explanation for this apparent contradiction may be that current approaches do not sufficiently exploit information contained in gaps. Alternatively, we hypothesized that patterns of gaps in current alignments tend to reflect information already present in a distance-based guide tree and are thus carrying little additional information. We here develop a methodology to evaluate the strength of such guide-tree dependency and show that most gaps are erroneously placed in patterns that tend to follow the guide tree. Different aligners are affected by this effect to different degrees but all suffer from it. Thus, most gaps in an alignment may not contain additional phylogenetic information per se but rather reflect information obtained from the pairwise comparison of the sequences in the alignment. Our results indicate that better modelling of gaps insertion is needed before alignment gaps can directly be used for phylogenetic inference.

P-2363

Measuring the distance between multiple sequence alignments

Benjamin Blackburne, Simon Whelan
University of Manchester, Manchester, UK

Multiple sequence alignment (MSA) is a core method in bioinformatics and phylogenetic analyses. The accuracy of such alignments may influence the results of phylogenetic inference, as well as other downstream analyses such as protein structure prediction or functional prediction. The importance of MSA has lead to the proliferation of MSA methods, with different objective functions and heuristics to search for the optimal MSA. Recent methods have begun to explcitly incorporate evolutionary models. Different methods of inferring MSAs produce different results in all but the most trivial cases. By measuring the differences between inferred alignments, we may be able to develop an understanding of how these differences (i) relate to the objective functions and heuristics used in MSA methods, and (ii) affect downstream analyses, including phylogenetic inference.

We introduce four metrics to compare MSAs, which may include the position in a sequence where a gap occurs or the location on a phylogenetic tree where an insertion or deletion (indel) event occurs. Our metrics allow a wider range of analyses than existing scores such as the Sum of Pairs or Total Column scores. Results across many aligners can be compared against each other without need for a reference alignment, for instance by plotting the distances between aligners in two dimensions using Principal Coordinate Analysis (multidimensional scaling).

We use both real and synthetic data to explore the information given by these metrics and demonstrate how the different metrics in combination can yield more information about MSA methods and the differences between them. For the synthetic analysis, different phylogenies are used to generate sequences and the effect of their toplogy on alignment accuracy is assessed.

A free software implementation of these metrics is available from http://kumiho.smith.man.ac.uk/whelan/software/metal/.

P-2364

SuiteMSA and improving MSAs

Catherine Anderson, Etsuko Moriyama
University of Nebraska-Lincoln, Lincoln, Nebraska, USA

Multiple sequence alignment (MSA) plays a central role in nearly all bioinformatics and molecular evolutionary applications. Sequence analysis often starts with building MSAs. Therefore, the quality of MSAs directly affects the outcomes and the quality of downstream analysis. Due to its significant impact on a wide range of biological analysis, MSA reconstruction is one of the most actively studied fields in bioinformatics, and numerous MSA methods have been developed. However, studies have shown that no single MSA method can handle all possible alignment problems consistently better than others. Different methods perform with varying degrees of accuracy depending on the type of alignments. Users are often left in the dark. Practical usage of MSA methods appears to be influenced simply based on accessibility of a program, *e.g.,* if a user-friendly Web interface is available or if a method is included in any sequence analysis package on hand. This is certainly not a good practice considering that low quality alignments could impact directly on the direction of the proceeding research. The accuracy of MSAs has been shown to be critical, for example, when we identify positively selected genes and gene regions. We are trying to rectify this problem from two sides. First we have developed SuiteMSA, a java-based application that provides unique MSA editor/viewers (Anderson et al. 2011, BMC Bioinformatics 12:184). SuiteMSA allows users to compare multiple MSAs directly and to evaluate where MSAs are consistent or inconsistent. It displays secondary structure and transmembrane prediction in alignment format, which is useful for visual inspection of MSA quality and its adjustment. SuiteMSA also provides a GUI for a sequence evolution simulator, which users can use to generate a reference MSA and evaluate the accuracy of any MSA. For the true alignment of simulated sequences, insertions and deletions are individually labeled, which aids evaluation of MSAs. Several alignment statistics (e.g., SPS and CP) are provided to aid the quantitative evaluation. It is available from: http://bioinfolab.unl.edu/~canderson/SuiteMSA/ . Our second aim is to develop a method to improve MSA quality and to integrate this new function in our easy-to-use SuiteMSA. Preliminary results show that our MSA improvement strategy by combining different methods can improve the MSA quality significantly compared to the original MSA methods ($P<0.0001$ by both paired t-test and Wilcoxon signed rank test).

P-2365

PHYRN: A Robust Method for Phylogenetic Analysis of Highly Divergent Sequences

Gaurav Bhardwaj[1], Sree Chintapalli[1], Yoojin Hong[0,2], Zhenhai Zhang[0,2], Edward Holmes[0,2], Damian van Rossum[0,2], Randen Patterson[1]
[1]University of California, Davis, Davis, CA, USA, [2]Pennsylvania State University, University Park, PA, USA

Both multiple sequence alignment and phylogenetic analysis are problematic in the “twilight zone” of sequence similarity (≤25% amino acid identity). We explored the accuracy of phylogenetic inference at extreme sequence divergence using a variety of synthetic data sets. We evaluated leading multiple sequence alignment (MSA) methods (*MAFFT, T-COFFEE, CLUSTAL, and MUSCLE*) and six commonly used programs of tree estimation (*Distance-based: Neighbor-Joining; Character-based: PhyML, RAxML, GARLI, Maximum Parsimony, and Bayesian*) against a novel MSA-independent method (PHYRN) described here. PHYRN is a profile-driven approach, wherein with the use of PSSM (Position Specific Scoring Matrix) libraries and a PHYRN composite scoring function, homology is encoded as a matrix of evolutionary distances. Strikingly, at “midnight zone” genetic distances (~7% pairwise identity and 4.0 gaps per position), PHYRN returns high-resolution phylogenies that outperform traditional approaches. We reason this is due to PHYRN's capability to amplify informative positions, even at the most extreme levels of sequence divergence. We also assessed the applicability of the PHYRN algorithm for inferring deep evolutionary relationships in the recA/RAD51 protein family and RNA-dependent RNA Polymerases (RdRP) from RNA viruses. In these biological data sets, PHYRN provides more robust mega-phylogenies than traditional methods. Further, meta-analysis of PHYRN vectors provides statistical measures of phylogenetic signal, and a quantifiable metric of signal-to-noise. Taken together, these results demonstrate that PHYRN represents a powerful mechanism for mapping uncharted frontiers in highly divergent protein sequence data sets.

*PHYRN is available for download at http://code.google.com/p/phyrn/

P-2366

Molecular evolution of the genes encoding the PRC2 and PhoRC Polycomb complexes in Drosophila.

Juan Manuel Calvo, Montserrat Aguadé, Montserrat Papaceit, Carmen Segarra
Universitat de Barcelona, Barcelona, Spain

The Polycomb group (PcG) proteins are epigenetic repressors highly conserved from yeast to plants and animals. Clustered in at least three main functional complexes (PRC1, PRC2 and PhoRC) Polycomb proteins are involved in many biological processes, being the silencing of the Hox genes during the embryonic development their most studied role. Here, we analyze the divergence of PcG genes in different Drosophila lineages with the aim of getting new insights into their molecular evolution. First, we sequenced the 10 genes which constitute the PRC2 and PhoRC complexes in the species *D. subobscura*, *D. madeirensis* and *D. guanche*, and we aligned the obtained sequences with those of the Drosophila 12 Genomes Consortium. The phylogenetic trees and relative rate tests performed show a significant acceleration of the evolutionary rate in the lineage leading to the obscura group species in five of the ten genes studied. Finally, maximum likelihood analysis performed with PAML indicates that this acceleration could have been driven by positive selection, which suggests, in turn, a coevolution process of interacting proteins within the Polycomb complexes.

P-2367

Genome-wide usage of accurately translated codons tracks a tRNA modification across the Drosophila phylogeny

John Zaborske[1], Vanessa Bauer DuMont[2], Tao Pan[1], Charles Aquadro[2], D. Allan Drummond[1]
[1]*University of Chicago, Chicago, IL, USA,* [2]*Cornell University, Ithaca, NY, USA*

Illuminating the interplay between evolutionary sequence patterns and biochemical features is a core aim of the post-genomic era. Prominent among evolutionary patterns is the unequal use of alternate synonymous codons, so-called biased codon usage, which appears in organisms as diverse as bacteria, yeast, fruit flies and humans. Yet the evolutionary reason for the most dramatic and consistent trend, the use of codons corresponding to abundant tRNA molecules in high-expression genes, remains obscure. Here, we first show that across 12 sequenced species of the Drosophila/Sophophora genus, codon usage for four amino acids shifts from the 3' C-ending synonym to the U-ending synonym, in a previously unrecognized genome-wide and phylogenetically coherent trend linked to systematic changes in the translational accuracy of these codons. Strikingly, the strongest shifts involve four amino acids for which a single genomically encoded tRNA species reads both codons. These tRNAs are known in D. melanogaster to be modified from guanosine (G) to queuosine (Q) at the anticodon's 5' nucleotide, corresponding to the 3' codon nucleotide. Acryloyl gel electrophoresis reveals that Q modification parallels the frequency of C-ending codons, from the dominant molecular species in D. melanogaster to a rarity in D. virilis, exactly opposite predictions made from preexisting biochemical data. These results link a codon-usage shift to a developmentally regulated epigenetic state, and illustrate the pitfalls and opportunities in connecting patterns of selection to the biochemical states of diverged organisms.

P-2368

PRDM9 binding targets occur within most human recombination hotspots

Nudrat Noor[2], Alexandru Aricescu[1], Julian Knight[2], Gil McVean[1,2], Simon Myers[1,2]
[1]Department of Statistics, Oxford, UK, [2]Wellcome Trust Centre for Human Genetics, Oxford, UK, [3]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Oxford, UK

In humans, meiotic recombination and crossing over is an essential process required for correct segregation of chromosomes. Recombination clusters within 1 to 2kb regions of the genome called “hotspots”, enriched for the presence of a 13-bp sequence motif, CCNCCNTNNCCNC. Recent work by several groups has shown PRDM9 binds to this motif, and so controls human recombination hotspot positioning. However, a close motif match occurs in only 40% of hotspots, while recent work shows that PRDM9 somehow controls hotspot activity even for hotspots lacking a close motif match. To address how this occurs, we used EMSAs to explore binding characteristics of human PRDM9. In each of five previously identified hotspots without a close motif match, we observe strong *in vitro* PRDM9 binding to more degenerate, but recognisably 13-bp motif like sequences near the hotspot centre, offering an explanation for hotspot activity in these cases. Although we also show PRDM9 binds DNA in a highly selective manner, with certain single base pair changes essentially abolishing binding, it is capable of binding considerably degenerate sequences, raising the question of why only a subset of *in vitro* binding targets correspond to recombination hotspots. A likely explanation is a role for nucleosome positioning and nucleosome modifications. We report the results of the first genome-wide survey revealing the key role of these features in interacting with PRDM9 binding to shape the human recombination landscape.

P-2369

Rapid diversification of Drosophila telomere capping genes.

Raphaëlle Dubruille[1], Gabriel Marais[2], Benjamin Loppin[1]
[1]CGphiMC - University of Lyon - CNRS, Lyon, France,[2]LBBE - University of Lyon - CNRS, Lyon, France

Drosophila capping proteins localize to telomeres in a sequence-independent manner and are essential to prevent chromosome end-to-end fusions. In somatic cells, the protective capping complex involves HipHop, HOAP and HP1a. We had previously shown that the duplication of the hiphop ancestor gene before the radiation of the *melanogaster* subgroup of species gave birth to K81, a unique paternal effect gene specifically expressed in the male germline. K81 is a telomere capping protein specialized in the protection of sperm telomeres. In the absence of K81, the fusion of paternal telomeres results in the aberrant division of paternal chromosomes during the first zygotic mitosis. Although HipHop and K81 seem to share a similar function in their respective expression domains, we have shown, by reciprocal genetic rescue experiments, that these proteins are not interchangeable. Here, we will present an phylogenetic analysis of the *hiphop* and *hoap* gene families in *Drosophila*. Interestingly, the recent publication of new sequenced *Drosophila* genomes representative of several other subgroups within the *melanogaster* group revealed a broader distribution of K81 and a highly dynamic repertoire of *hiphop*-related genes, with multiple gene losses and gains observed at all levels of the phylogeny. We also observed, to a lesser extent, a significant diversification of the *hoap* family. The functional consequences of the rapid diversification of these essential capping gene families will be discussed.

P-2370

A genomic study of the contribution of DNA methylation to regulatory evolution in primates

Julien Roux, Yoav Gilad
University of Chicago, Chicago, USA

A long-standing hypothesis is that changes in gene regulation play an important role in adaptive evolution, notably in primates. Yet, in spite of the evidence accumulated in the past decade that regulatory changes contribute to many species-specific adaptations, we still know remarkably little about the mechanisms of regulatory evolution. In this study we focused on DNA methylation, an epigenetic mechanism whose contribution to the evolution of gene expression remains unclear.

To interrogate the methylation status of the vast majority of cytosines in the genome, we performed whole-genome bisulfite conversion followed by high-throughput sequencing across 4 tissues (heart, kidney, liver and lung) in 3 primate species (human, chimpanzee and macaque). Because the 4 tissues are from the same individuals, we are able to monitor methylation differences between individuals, tissues and species.

In parallel, we collected gene expression profiles using RNA-seq from the same tissue samples, allowing us to perform a high resolution scan for genes and pathways whose regulation evolved under natural selection.

We integrated these datasets to characterize better the genome features whose methylation status leads to expression changes, and we developed a statistical model to quantify the proportion of variation in gene expression levels across tissues and species which can be explained by changes in methylation.

Globally, our study leads to a better understanding of the molecular basis for regulatory changes and adaptations in primates.

P-2371

Integrating Functional and Population Genomics in *Drosophila melanogaster*

David Castellano[1,2], Antonio Barbadilla[1,3]
[1]Universitat Autónoma de Barcelona, Bellaterra, Barcelona, Spain, [2]Institut de Biotecnologia i Biomedicina, Bellaterra, Barcelona, Spain, [3]Departament de Genètica i Microbiologia, Bellaterra, Barcelona, Spain

To gain insight into how the large-scale organization of chromatin landscape is related with polymorphism, functional constraint (the ratio of selective to neutral polymorphism) and the fraction of adaptive substitutions, we have related two impressive data sets that have recently released in *Drosophila melanogaster*; (1) the model organism Drosophila Encyclopedia of DNA Elements (modENCODE) project, which maps chromatin states among other kind of genomic information and (2) the Drosophila Genetic Reference Panel (DGRP) project, where 158 lines of a natural population of *D. melanogaster* has been sequenced to the subsequent association QTL mapping and population genomics studies. We have discovered that different chromatin states are related with different chromosome regions and recombination contexts within the genome, affecting differentially the selective and neutral polymorphism throughout the genome of Drosophila. We conclude that chromatin function affects constraint and adaptive evolution for every functional site class (intergenic regions, UTRs, coding sequence and introns) in the *D. melanogaster* genome. The consequences of these finding on the interpretation of the patterns of genetic diversity are discussed.

P-2372

Epigenetic inheritance versus mobility in fluctuating environments

Luca Ferretti[1], Michele Cortelezzi[2]
[1]CRAG, Barcelona, Spain,[2]Universita' di Pisa, Pisa, Italy

Epigenetic inheritance has been shown to be common in plants and animals. However, epigenetic phenomena in plants can affect important phenotipic traits and appear to be more relevant for adaptive responses than the corresponding phenomena in animals. One plausible explanation for this observation is the difference in the available responses to fluctuations of the environment. Recent papers show that in an environment with only time-dependent fluctuations, epigenetic inheritance or similar phenomena like stochastic switching do not provide a significant fitness advantage. This result appears to be valid under reasonable assumptions on the behaviour of the fluctuations and on the effect of the epigenetic response on fitness.
However, spatial fluctuations or spatial heterogeneity of the environment could also play a significant role. In fact, it has been suggested that animals could respond to spatio-temporal heterogeneities exploiting their nervous system and their mobility. Since plants have a reduced mobility, they have to exploit different mechanisms, like epigenetic inheritance, to adapt to variable environmental conditions.
In this work, we study a model of epigenetic inheritance in fluctuating environment that includes adaptation to changes in the external conditions and short-term heritability of adaptive changes. This model is simple but compatible with several mechanisms of inheritance. We consider both spatial and temporal fluctuations and spatial heterogeneity. We use numerical simulations and some analytic results to obtain the conditions (environmental fluctuations, adaptability and heritability of epigenetic changes, mobility and offspring dispersal) under which epigenetic inheritance mechanisms are positively selected.

P-2373

The role of epigenetics in the evolution of sexual attraction in moths

Astrid T. Groot
University of Amsterdam, Amsterdam, The Netherlands

The evolution of sexual attraction and its role in speciation is an outstanding evolutionary question. Especially in species where signal and response are governed by unlinked genes residing on different chromosomes, the apparent co-ordinated evolution of signal and response poses a dilemma for evolutionary theory. Generally, signal and response are hypothesized to be under stabilizing selection. A mutation in either the signal or the response thus faces a direct evolutionary hurdle. A phenotype-first scenario does not face such a hurdle: if a new environment induces epigenetic modifications, it may affect many individuals at the same time, so that the new phenotype appears at a nonnegligible frequency from the start.

Variation in sexual attraction has been observed in many species, but how much of this variation is genetic and how much is environmentally induced has been hardly explored. Moths are ideal animals to disentangle these factors, because sexual attraction is well-defined and purely chemical: females produce a species-specific sex pheromone that attracts males from a distance.

In two closely related noctuid moth species that we have studied in the past decade, Heliothis subflexa and Heliothis virescens, we found two types of natural variation: in Hs we found geographic variation, while in Hv we found a polymorphism in the sex pheromone in every population studied so far. In both species, we found that much of this natural variation is genetic. Excitingly, we also found phenotypic plasticity in the sex pheromone of both species, which was in the same direction as the natural variation that we found within each of the species. This plasticity cannot be explained by genetics but by epigenetic modifications. Therefore, my current research focuses on how environmental factors affect signal and response epigenetically and how important are these effects in the evolution of sexual attraction. In this presentation I will give an overview of our progress to answer this question.

P-2374

Intron positional conservation: a new paradigm of evolutionary conservation

Michal Chorev, Liran Carmel
The Hebrew University of Jerusalem, Jerusalem, Israel

A fundamental supposition in comparative genomics is that evolutionary conservation is indicative of biological function. This makes the identification of highly conserved genomic regions a chief strategy in looking for function. Yet, many functional elements, especially noncoding, evade detection by standard evolutionary conservation methods (like sequence, structure, etc.). In particular, we focus here on intron functions, which are among the most elusive to detect. While previously regarded as neutral or slightly deleterious selfish genomic elements, it is now recognized that many introns are critical to the normal function of the cell, and participate in a wide variety of functions, and in almost any step of mRNA processing [1].

To detect intron function, we suggest a novel type of evolutionary conservation that is based on analyzing the *gene architecture*, which is the intron-exon structure of the gene. Specifically, we look on the conservation of intron positions, defined as the points of intron insertion along the mRNA sequence. The idea is that an intron in a particular gene and in a particular lineage has a characteristic gain and loss rates [2, 3], but once it becomes associated with a function, of whatever type, its chances to be lost will substantially decrease. Detecting such introns with very low loss rate would be, therefore, indicative of function of any type, even if the function is unrelated to the intron position.

In order to make inference on intron positional conservation, there is a need to reconstruct the evolutionary history of introns, and to estimate the rates of intron gains and losses. To this end we use EREM [2], a comprehensive model of gene architecture evolution. EREM allows for intron gain and loss rates heterogeneity between lineages, as well as between individual genes. We also built an intron-exon database, with more than 70 eukaryotes, and with introns functional annotations (like hosting miRNAs, snoRNAs, enhancers, etc.). Applying our model to this dataset, we are able to obtain an evolutionary reconstruction of gene architecture with great precision, and to show a link between intron positional conservation and intron function.

1. Lynch, M., *The Origins of Genome Architecture*. 2007: Sinauer Associates Inc.

2. Carmel, L., et al., *Patterns of intron gain and conservation in eukaryotic genes.* BMC Evol Biol, 2007. **7**: p. 192.

3. Carmel, L., et al., *Three distinct modes of intron dynamics in the evolution of eukaryotes.* Genome Res, 2007. **17**(7): p. 1034-44.

P-2375

Promoter regions of duplicated *Schistosoma mansoni* genes crucial for infection success undergo rapid epigenetic modifications in varying environments while remaining genetically stable

Cecile Perrin[1], Julie MJ Lepesant[1], Emmanuel Roger[2], David Duval[1], Virginie Thuillier[1], Jean Francois Alliene[1], Guillaume Mitta[1], Christoph Grunau[1], Celine Cosseau[1]
[1]UMR5244 CNRS UPVD, Perpignan, France,[2]CNRS UMR 8204, Lille, France

The digenetic trematode *Schistosoma mansoni* is a human parasite that uses the mollusc *Biomphalaria glabrata* as intermediate host. Specific *S. mansoni* strains can infect efficiently only certain *B. glabrata* strains (compatible strains C) while others are incompatible (IC). Differences in transcription of polymorphic mucins (*Sm*PoMuc) between strains of *S. mansoni* are one of the principle determinants for compatibility with *B. glabrata*. *Sm*PoMuc are polymorphic glycoproteins that interact with *B. glabrata* Fibrinogen Related Proteins (FREPs). FREPS are involved in the immune response of the mollusc to the parasite. *Sm*PoMuc are encoded by a multi-gene family. Importantly, the high degree of protein variability is derived from a limited number of *Sm*PoMuc genes. The *Sm*PoMuc coding sequences of C and IC strains are very similar. To generate high variability and differences between strains based on a relatively low number of *Sm*PoMuc genes, *S. mansoni* has evolved a complex cascade of mechanisms, a "controlled chaos", acting at the transcriptional, translational and post-translational level. In the present study, we investigated the bases of the control of *Sm*PoMuc expression that evolved to evade *B. glabrata* diversified antigen recognition molecules in the promoter of *Sm*PoMuc genes. We compared two strains of different geographic origin and either compatible (C) or incompatible (IC) with a reference snail host. We reveal that although sequence differences are observed between active promoter regions of *Sm*PoMuc genes, the sequences of the promoters are not diverse and are conserved between IC and C strains, suggesting that genetics alone cannot explain the evolution of compatibility polymorphism. In contrast, promoters carry epigenetic marks that are significantly different between the C and IC strains. Moreover, we show that these modifications of the structure of the chromatin of the parasite lead to enhanced transcription of *Sm*PoMuc in the IC strain compared to the C strain and correlate with the presence of additional combinations of *Sm*PoMuc transcripts only observed in the IC phenotype. These results indicate that epigenetic marks change before genetic modifications occur, and suggest that epigenetic changes may be important for the early steps in the evolution of an adaptive trait in *S. mansoni*. This validate theoretical hypothesis that predict that in changing environments, in particular parasite-host interactions with fluctuating hosts, epigenetic inheritance system is important and allows for rapid adaptive evolution.

P-2376

DNA methylation in *Biomphalaria glabrata* the intermediate host of the human parasite *Schistosoma mansoni*: role of epigenetics in the coevolution of host / parasite interactions.

Sara Fneich[1,2], Julie Lepesant[1,2], Céline Cosseau[1,2], Michael Reichelt[3], Christoph Grunau[1,2]
[1]Université de Perpignan Via Domitia, Perpignan, F-66860, France,[2]CNRS, UMR 5244, Ecologie et Evolution des Interactions (2EI), Perpignan, F-66860, France,[3]Max-Planck-Institut fuer Chemische Oekologie, D-07743 Jena, Germany

In parasite-host interactions, parasites exert selective pressures on their hosts and *vice versa*, leading to a genuine arms race between both partners. To adapt, each partner must evolve the capacity to express new phenotypic variants. We propose that epigenetic variations play an important role in the genesis of phenotypic variability without changing the nucleotide sequence. The parasite of our study model *Schistosoma mansoni* is characterized by high capacity for adaptation. Its life cycle requires passage though an intermediate host (snail) and a definitive host (man). Our previous work has demonstrated that epigenetic changes in *S. mansoni* increase its phenotypic variability. In the same context of interactions, we suggest an interdependence of epigenomes during the schistosome / snail interaction. DNA methylation is a carrier of epigenetic information in many, but not all species. We show here that DNA methyation occurs in the snail *Biomphalaria glabrata*, the intermediate host of the parasite. Currently, we use different techniques (MS-AFLP, LC-MS and bisulfite sequencing), to analyse in more detail covalent DNA modifications across the genome. Our results will contribute to a better understanding of the mechanisms of host / parasite coevolution.

P-2377

Y-by-background effects on global gene expression in *Drosophila melanogaster*

Pan-Pan Jiang, Bernardo Lemos, Lene Martinsen, Daniel Hartl
Harvard University, Cambridge, MA, USA

The Y chromosome, inherited without meiotic recombination from father to son, carries relatively few genes in most species, including *Drosophila melanogaster*. This is consistent with predictions from evolutionary theory that non-recombining chromosomes lack variation and degenerate rapidly. However, some recent work has suggested a more dynamic role for the Y, implicating it in critical evolutionary pathways such as spermatogenesis, thermal regulation, divergence, and male fitness. The Y chromosome, however, may not work alone, and work in *D. melanogaster* suggests that there are significant Y-by-background interaction effects on male fitness. We then asked two questions: are there Y-by-background effects on global gene expression, and, if so, what are the molecular mechanisms responsible?

We studied Y chromosomes from two populations collected from two geographically diverse populations of *D. melanogaster,* one from France (F) and the other from India (I). We expressed these Y chromosomes in three different backgrounds, and contrasted males carrying the two Y chromosomes (Y_I vs Y_F) in each background using ~18,000-feature cDNA microarrays.

Our results present novel evidence for significant regulatory interactions between the chromosome and the genetic background: genes that are highly expressed in some Y-by-background combinations may be lowly expressed in other combinations. Affected genes show an association with immune response and pheromone detection, suggesting a functional cohesion to Y-by-background effects.

Finally, the molecular mechanism for the large Y-effect remains elusive, since Y-linked genes are almost entirely devoid of polymorphisms. One hypothesis is that balanced polymorphisms in non-gene regions, such as in Y-heterochromatic repeats, may provide the necessary variation for differential competitive binding of transcription factors. We provide preliminary data suggesting that microsatellite regions on the Y, which have previously been shown to act as an on-off transcriptional activator of reporter gene lacZ and to also bind GAGA transcription factors, may also harbor enough inter-population variation to differentially affect genome-wide transcription on a more subtle continuum.

P-2378

Transgenerational epigenetic inheritance in *Drosophila melanogaster*

Jaime Grace, Urban Friberg
Uppsala University, Uppsala, Sweden

Epigenetic inheritance encompasses several categories of non-Mendelian transmission of information from parent to offspring. These epigenetic factors can influence not only the developmental environment of the offspring, but may have effects that persist throughout the offspring's lifetime and even the generations beyond. The persistence of these effects demands a greater understanding of how epigenetic factors act to regulate gene expression and which genes might be the targets of epigenetic influence. Several recent studies have shown specific genes to be epigenetically regulated across generations; however, the generality and the magnitude of epigenetic effects under normal conditions have not been established. What is missing is a genome-wide perspective on the pervasiveness of these effects and the extent to which the whole genome is influenced by transgenerational epigenetic inheritance. In this study, we will compare gene expression profiles across the entire genome of four divergent populations of *D. melanogaster* to determine which genes are differentially expressed due to paternal epigenetic inheritance. We have constructed males that differ with respect to half of their genomes (hemiclones). When mated to genetically uniform females, they produce sons that are genetically uniform, identifiable by several phenotypic markers. Since these genetically identical offspring are produced from fathers having different genetic backgrounds, the gene expression differences among lines can be attributed to epigenetic inheritance. We also determine whether the effects of epigenetic inheritance on gene expression are widespread throughout the *Drosophila* genome, and which genes are most affected. Genes that are differentially expressed through epigenetic means will be identified as candidate genes for further exploration.

P-2379

Inter-and Intra-specific morphological variation and scaling of microtubule derived-organelles

Beatriz F. Gomes, José B. Pereira-Leal
Instituto Gulbenkian de Ciência, Oeiras, Portugal

The analysis of morphological variation has a long tradition in evolutionary biology. The onset of molecular genetics and the genomics age allowed us to begin exploring the genetic and developmental determinants of morphological diversity and plasticity. However, morphological diversity at the cellular level, and in particular of intracellular structures remains largely unexplored, and we understand little of the constrains and paths of evolution within the subcellular morpho-space.

We are studying the evolution of microtubule-derived organelles in eukaryotic cells, the cilia/flagella, centrosomes, nuclear associated bodies, spindle pole bodies, etc, and using them as a model system to address the evolution of morphology at the subcellular level. Morphological variation across species has been documented in these structures, but its molecular, environmental and functional determinants are unclear. Morphological variation is also observed in many human diseases which are collectively designated ciliopathies and centrosomopathies, and more recently in several human cancers. It is however unclear to what extent the underlying genetic causes of these diseases allow the organelles to explore new regions of the morpho-space, or instead, they just move to other states that are 'normal' in other tissues or organisms.

As part of a large community effort (mtocdb.org) we have annotated hundreds of electron microscopy images of these structures in multiples species according to a controlled vocabulary, and correlated the presence and morphological details of these structures with the presence and properties of genes and also with the tree of life. We used this resource to investigate the scaling relationships for the cylindrical structure of the centriole/basal body within and across species, and our preliminary results suggest that the longitudinal axis of the centriole represents the direction of larger variation, while the diameter of the cylinder varies considerably less. We have also observed that these two traits are coupled with a quasi linear scaling law. Furthermore, our analysis revels size variations across species that exceed the intra-specific variation. We are currently expanding the scope of our analysis (more species, more tissues per species), and characterizing "abnormal" morphologies in human diseases using primary data from human cancer samples. We are also seeking to understand the molecular variations that underlie both intra- and inter-specific variability.

Our approach and results fit into an emerging area of evolutionary cell biology, which like evo-devo, aims to open up a "black box" that has not received much attention by evolutionary biologists.

P-2380

Multiple sequence alignment over-optimization bias impact downstream on population statistics and inferred phylogeny.

Abhijeet Shah, Dirk Metzler
University of Munich, Munich, Germany

Sequence alignment is a fundamental component of sequence based molecular evolutionary and comparative studies. Errors in alignments lead to errors downstream in the interpretation of evolutionary information and population statistics. Most MSA methods rely on dynamic programming, heuristic approaches, profile analysis and statistical approaches to build optimized alignments. However, it is unknown how these approaches may bias the inferred hypotheses though over-optimization of the alignment. Here, we provide evidence for a previously unknown alignment over-optimizations bias in the branch lengths estimates of the inferred phylogenies in distantly related sequences, and in common population statistics estimates of highly similar sequences. Our pilot study suggested that there is a trend for over-estimating bias in estimating branch length ratios in distantly related sequences across most tested packages. Interestingly, the simultaneous Bayesian estimation of phylogeny and alignment package, BAli-Phy, showed the most consistently accurate results. BAli-Phy uses a novel joint estimation approach to increase the resolving power by making use of information in shared indels while avoiding bias and overconfidence in inferred topologies resulting from inaccurate alignments. While estimating the Fst population statistic on over-optimized alignments of highly similar sequences, we observed over-estimating bias for packages such as MAFFT and ProbCons whereas under-estimating bias was observed for the phylogeny aware PRANK package. Whereas estimating the Tajima's D population statistic, we observed a trend of under-estimated values with the exception of ClustalW2, which showed the most accurate. However, we also observed that ClustalW2 under-estimated the number of gaps by the largest margin, thereby suggesting an inherent systemic bias.

P-2381

A Search For Fusions Of Unrelated Genes In The Yeast Genome

Leanne M. Murphy, Leanne S. Haggerty
National University of Ireland Maynooth, Co. Kildare, Ireland

Gene fusion refers to the genetic recombination of two or more genes resulting in the generation of a new gene. Alterations to gene frequency and structure allows for variation and adaptation in a population of various organisms. This study uses a novel method for gene fusion detection to investigate the role that gene fusions play in yeast species, such as Saccharomyces cerevisiae, and how they may contribute to their evolutionary success. We have carried out an

all-versus-all BLAST search of the yeast genome and using the resulting homology statements, we have built a

homology network. Then, using a novel method for detecting fusion proteins, we have identified a number of genes that appear to be the fusion of two unrelated genes. We have characterised these fusion genes in terms of their function.

P-2382

The asymptotic null distribution of certain likelihood ratio test statistics for detecting natural selection in codon sequences

Jean Kai Ling Lim
National University of Singapore, Singapore, Singapore

The likelihood ratio test statistic between two nested hypotheses often has an asymptotic chi-square distribution under the null hypothesis, with degree of freedom equal to the difference in the number of parameters between the alternative and the null hypotheses. However, in the Goldman-Yang frame- work for detection of adaptive evolution in coding sequences, the asymptotic null distribution may not be chi-square, because the parameters specifying the null hypothesis lie on the boundary of the parameter space associated with the alternative hypothesis. Ignoring this boundary problem tends to result in lower statistical power.
We use the simulation approach to explore the null distribution of the likelihood ratio tests for three nested hypotheses: (a) H0: $\omega = 1$, (b) H1: $\omega>0$,(c)H2: Mixtureofthreecomponents: $\omega<1$,$\omega=1$,and$\omega>1$. Codon sequence pairs, of length 1,000, 5,000 and 10,000 and with similar- ity 90% and 70%, were simulated 500 times from a simple Goldman-Yang model in H0, using the statistical software R. The likelihood ratio test statis- tic L10 (between H1 and H0) and L21 (between H2 and H1) were evaluated using PAML and PyCogent. The empirical distribution of L10 fits the ex- pected $\chi2(1)$ very well, and the agreement gets better as length increases. Literature suggests the asymptotic null distribution of L21 to be a 50:50 mix- ture of 0 and $\chi2(1)$. This is somewhat supported, but is less convincing. Several possible explanations are apparently ruled out by the following ob- servations: the statistics from PAML and PyCogent are almost identical, the sequence lengths and number of repetitions are considered large enough, a probability-probability plot is used to check the distribution, instead of the biased quantile-quantile plot.
Once a clear understanding of the puzzle is attained, the methodology can be used to obtain more accurate P values in real data analysis, including the more complicated selection models involving mixture distributions.

P-2383

The genetic impact of infectious disease on indigenous southern African populations

Katharine Owers
Uppsala University, na, Sweden

Humans have been impacted by infectious diseases throughout their history. While contemporary diseases can be studied with modern methods, allowing rapid collection and dissemination of information about their effects on populations, studies of historical diseases do not benefit from those advantages. Some historical diseases, such as the plague that struck Europe in the 14th century, are relatively well-understood, but in other cases we have little information on the diseases and their impacts. Such is the case for the wave of disease caused by European colonization of southern Africa. Societies present in the area at that time did not keep written records, so we must reconstruct their pre-contact history from European reports, oral histories, and information from archeology and linguistics. Recent advances in genetics, however, have provided new sources of information on population history, structure, and selection. I analyze SNPs from several indigenous southern African populations with differing levels of contact with Europeans to look for evidence of selection due to pressure from introduced infectious diseases. Two general approaches are used. First, I locate regions of the genome likely under selection according to various test statistics based on haplotype homozygosity and population differentiation and then check those regions for infectious disease-related gene enrichment. Second, I create a list of infectious disease genes and examine their genomic regions for indicators of selection. This dual approach allows both detection of areas of strongest selection and an overall idea of selection on infectious disease genes in these populations.

P-2384

Pairwise Evolutionary Distance Estimation Under Alignment Uncertainty

Ge Tan
Department of Computer Science ETH Zuerich, na, Switzerland

Computing accurate evolutionary distances lies at the heart of the phylogenetics and molecular evolution. Typically, distances are estimated from an optimal alignment followed by a maximum likelihood procedure. However, the underlying sequence alignment can be a large source of uncertainty, particular among divergent sequences. In this work, we focus on the simplest possible scenario: how to best compute an evolutionary distance between two sequences. We developed a new implementation based on the TKF91 and TKF92 model, which consider all the possible paths of transforming from one sequence to another. Computer simulations were conducted to validate this implementation. Next, we compared the performance comparison of 6 different state-of-the-art types of distance estimation methods on real biological data. Our results indicate a great potential for two distance estimations: kernel distance and TKF92 model. Alignment-based methods are still useful for fast computation of evolutionary distance. This study provides a comprehensive comparison of commonly used distance estimation approaches and recommendations on practical distance estimation procedure.

P-2385

Birth, death, and replacement of karyopherins in Drosophila

Emily Hsieh, Nitin Phadnis, Harmit S. Malik
Fred Hutchinson Cancer Research Center, Seattle, na, USA

The nuclear transport pathway performs the fundamental function of moving cargo between the cytoplasm and nucleus. Nuclear transport is an essential function and is carried out through a highly conserved mechanism across all eukaryotes. Yet, in Drosophila, several components of the nuclear transport apparatus evolve rapidly under positive selection. Genetic conflict with selfish elements, such as segregation distortion, is a possible cause for this pattern of rapid evolution.

Here, we performed a comprehensive phylogenomic analysis of importin gene evolution in Drosophila. Importins are adapter molecules that directly mediate the transport of cargo into the nucleus. Our analysis reveals a recurrent pattern of gain and loss of importin paralogs in Drosophila across independent lineages. Interestingly, we discovered that almost all new copies of importins have acquired a testes-specific expression pattern since their birth through gene duplication. This pattern of repeated gains of testes-specific copies of importins and signatures of episodic lineage-specific positive selection suggests a function in suppressing genetic conflicts in the male germline such as segregation distortion. Segregation distorters such as SD in Drosophila melanogaster act by impairing nuclear transport in the testes. We are currently performing functional tests for the hypothesis that an increased dosage of these non-canonical importins in the testes may serve a role in suppressing segregation distortion in males by restoring nuclear transport during spermatogenesis.

P-2386

Emergence of multicellular novelty

Chelsea Du Fresne
University of Minnesota- Twin Cities, na, USA

The origin of biological novelty is a central question in evolutionary biology. Previously we observed the de novo formation of multicellularity during selection for settling, using Saccharomyces cerevisiae to experimentally investigate a major evolutionary transition. Initially, the evolution of multicellularity was beneficial because it increased settling rate. Here we investigated if this transition promoted subsequent adaptive novelties. Populations of clonally propagated multicellular Saccharomyces cerevisiae were subjected to a daily selection regime for fast settling. Settling selection treatments ranged from zero to 300. Adaptation to settling was observed, and the responses to selection were not uniform across treatments or lineages within treatments. In particular, one lineage evolved an extreme phenotype after 164 transfers. Individuals from this lineage settle at a rate 47% faster than the mean for all 15 lineages and no other lineage differs by more than 15%. This dramatic increase in settling did not evolve as a consequence of increased body size, unlike all previous responses to selection. Increasing settling rate by means other than increasing size is a novel adaptation observed during the transition to multicellularity. We conclude that even after a major evolutionary transition, the emergence of novel modes of adaptation are strongly affected by chance mutational events.

P-2387

Genetic variation of Symbiodinium spp. of the coral host Agaricia lamarcki between shallow and mesophotic habitats

Ramon Rivera Vicens
University of Puerto Rico, na, Puerto Rico

The Caribbean mesophotic coral ecosystems (reef habitats found between 50 and 100 m depth) and their shallow water counterparts provide a unique system to examine the patterns of genetic connectivity for scleractinian corals and their algal symbionts (Symbiodinium spp.). The coral Agaricia lamarcki harbors symbionts and inhabits both shallow and mesophotic habitats. Patterns of genetic connectivity of Symbiodinium populations from A. lamarcki were estimated with the internal transcribed spacer of rDNA (ITS2) among shallow (< 30 m) and mesophotic populations (50-70 m) of Mona Island, southwestern Puerto Rico, and St. Thomas, USVI. Preliminary data from a 280 bp region of ITS2 rDNA show that A. lamarcki populations > 50 m harbor "deep specialists" symbionts with unique haplotypes not previously described. Additionally, phylogenetic analysis of 112 cloned sequences from 15 individuals of A. lamarcki revealed two clades consisting of predominantly Clade C, and five individuals of A. lamarcki harboring Clade D, that were only found in shallow waters of St. Thomas, USVI. Continuing work will include tests for population genetic structure among different depth habitats within a region and between regions. These data are important since coral-symbiont associations are hypothesized to provide an evolutionary plasticity (adaptability) for corals responding to future environmental degradation and climate change.

P-2388

Prevalence and Genetic Diversity of Microsporidia intracellular parasites across fire ant species in South America

Gebreyes Kassu
University of Florida, na, USA

Two species of fire ants: Solenopsis invicta, the red imported fire ant, and S. richteri, the black imported fire ant were accidentally introduced to the United States from South America almost 100 years ago. Kneallhazia solenopsae and Varimorpha invictae are microsporidia that infect fire ants that cause queens to have low weight and reduced fertility. Thus, these microsporidia were identified as potential biological control agents against fire ants. In this study, the presence of microsporidia was detected using a polymerase chain reaction (PCR) assay that amplifies a portion of the 16S ribosomal RNA (16S rRNA) gene of the microsporidian genome. Prevalence of both genera of microsporidia was determined in 323 S. richteri colonies and 14 colonies of S. invicta. In addition, 221 of the S. richteri colonies were infected with an ant that acts as a parasite on other ants, Solenopsis daguerrei. These social parasitic ants were also screened. Among the 323 S. richteri colonies screened from nine geographic sites, eight sites (89%) were infected with microsporidia with 40 (12.4%) colonies positive for K. solenopsae and 4 (1.2%) colonies positive for V. invictae. The parasitic ant, S. daguerrei, also showed infection from microsporidia with seven ants from 3 sites being infected with K. solenopsae and one ant from a single site containing V. invictae. S. invicta ants were collected in one site where none of the colonies were infected. Phylogenetic analyses of 16S sequences revealed that K. solenopsae haplotypes found in S. richteri (the host) and S. daguerrei (the social parasite) were identical to each other, but different from those reported from S. invicta. Further studies analyzing the molecular diversity of K. solenopsae among different hosts will be discussed.

Plenary 3

Sex, flies and (possibly) conflict: the molecular evolution of Germline Stem Cell genes

Charles “Chip” Aquadro[1,2], Heather A. Flores[1,2], Vanessa Bauer DuMont[1,2], Jae Choi[1,2], Daniel Barbash[1,2]
[1]Cornell Center for Comparative and Population Genomics, Cornell University, Ithaca, New York, USA,[2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA

Genes involved in many aspects of reproduction show strong statistical evidence of natural selection driving amino acid diversification. For many sperm and egg proteins, as well as seminal fluid proteins, in insects and mammals, much available data are consistent with evolutionary conflict associated with sperm competition, cryptic sexual selection, or male-female conflict. The involvement of pathogens has been only hinted at. Our group, working collaboratively with Daniel Barbash, has studied numerous genes that regulate the maintenance and early differentiation of germline stem cells (GSCs) and find that a majority of those genes show evidence of departures from neutrality, in many cases suggesting strong persistent selection for the same patterns of amino acid diversification reminiscent of the "downstream" reproductive genes. The striking difference is that at the GSC and early gametogenesis stages, cells are genetically identical, divisions are only mitotic, and there is no interaction between males and females, thus ruling out many of the classically hypothesized drivers of reproductive gene molecular evolution. We have also shown that oogenesis and spermatogenesis processes appear conserved among D. melanogaster and closely related species, ruling out a major interspecific shift in GSC biology. We further show that while these diverged genes work normally in their "resident" species, when transformed into D. melanogaster, the D. simulans gene shows defects in GSC differentiation, but primarily in females. Following this lead, we discovered a direct interaction between the key GSC "switch" gene and the maternally inherited bacteria Wolbachia, well known for it's manipulation of reproduction in insects. These results make clear that male-female conflict, and/or sperm-competition alone do not drive positive selection at reproductive genes in at least Drosophila, and raise in prominence the hypothesis that pathogens may be causing these genes to change rapidly.