Department of Energy
Joint Genome Institute

GENOMICS OF ENERGY AND ENVIRONMENT

User Meeting Abstracts

March 18-20
2014
Walnut Creek
CALIFORNIA

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

U.S. DEPARTMENT OF ENERGY
Office of Science

All information current as of March 12, 2014
JGI Contact: **Denise Yadon**
DOE Joint Genome Institute
dyadon@lbl.gov

The Joint Genome Institute is a user facility of
the Department of Energy Office of Science

DOE Joint Genome Institute: **www.jgi.doe.gov**
DOE Office of Science: **science.energy.gov**

Ninth Annual
DOE Joint Genome Institute
User Meeting

Sponsored By

U.S. Department of Energy

Office of Science

March 18-20, 2014

Walnut Creek Marriott

Walnut Creek, California

Contents

Table of Contents

Speaker Presentations

Abstracts alphabetical by speaker

A single cell perspective on bacterial interactions

Martin Ackermann (martin.ackermann@env.ethz.ch)

ETH-Zurich

KBase Overview

Adam P. Arkin* (aparkin@lbl.gov)[1], Jason Baumohl[1], Aaron Best[2], Jared Bischof[2], Ben Bowen[1], Tom Brettin[2], Tom Brown[2], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Ric Colasanti[2], Neal Conrad[2], Brian Davison[3], Matt DeJongh[6], Paramvir Dehal[1], Narayan Desai[2], Scott Devoid[2], Terry Disz[2], Meghan Drake[3], Janaka Edirisinghe[2], Gang Fang[7], José Pedro Lopes Faria[2], Mark Gerstein[7], Elizabeth M. Glass[2], Annette Greiner[1], Dan Gunter[1], James Gurtowski[5], Nomi Harris[1], Travis Harrison[2], Fei He[4], Matt Henderson[1], Chris Henry[2], Adina Howe[2], Marcin Joachimiak[1], Kevin Keegan[2], Keith Keller[1], Guruprasad Kora[3], Sunita Kumari[5], Miriam Land[3], Folker Meyer[2], Steve Moulton[3], Pavel Novichkov[1], Taeyun Oh[8], Gary Olsen[9], Bob Olson[2], Dan Olson[2], Ross Overbeek[2], Tobias Paczian[2], Bruce Parrello[2], Shiran Pasternak[5], Sarah Poon[1], Gavin Price[1], Srivdya Ramakrishnan[5], Priya Ranjan[3], Bill Riehl[1], Pamela Ronald[8], Michael Schatz[5], Lynn Schriml[10], Sam Seaver[2], Michael W. Sneddon[1], Roman Sutormin[1], Mustafa Syed[3], James Thomason[5], Nathan Tintle[6], Will Trimble[2], Daifeng Wang[7], Doreen Ware[5], David Weston[3], Andreas Wilke[2], Fangfang Xia[2], Shinjae Yoo[4], Dantong Yu[4], Robert Cottingham[3]* , Sergei Maslov[4] , Rick Stevens[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Argonne National Laboratory, Argonne, Illinois; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Brookhaven National Laboratory, Upton, New York; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [6]Hope College, Holland, Michigan; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

Systems biology is driven by the ever-increasing wealth of data resulting from new generations of genomics-based technologies. With the success of genome sequencing, biology began to generate and accumulate data at an exponential rate. In addition to the massive stream of sequencing data, each type of technology that researchers use to analyze a sequenced organism adds another layer of complexity to the challenge of understanding how different biological components work together to form a functional living system. Achieving this systems-level understanding of biology will enable researchers to predict and ultimately design how biological systems will function under certain conditions. A collaborative computational environment is needed to bring researchers together so they can share and integrate large, heterogeneous datasets and readily use this information to develop predictive models that drive scientific discovery.

The advancement of systems biology relies not only on sharing the results of projects through traditional methods of peer review and publication, but also on sharing the datasets, workflows, software, models, best practices, and other essential knowledge that made those published results possible. Establishing a common framework for managing this knowledge could save time, reduce duplicative effort, and

increase the scientific return on investment in systems biology research. This framework for precisely tracking what was done to achieve a certain outcome also will empower researchers to reproduce published results and review projects more effectively. This ability to reproduce experiments described in publications is a key principle of the growing Open Science movement.

The KBase team is developing an open-source, open-architecture framework for reproducible and collaborative computational systems biology. KBase's primary scientific aim is to push multiple types of functional data towards increasingly specific models of metabolic and regulatory behavior of microbes, plants and their communities. We have brought together data and tools that allow probabilistic modeling of gene function, which can be used in turn to produce experimentally testable models of cellular metabolism and gene regulation. The system will soon allow models to be compared to observations and dynamically revised.

One of the new components that will enable these complex, iterative analyses is KBase's new prototype Narrative Interface, which provides a transparent, reproducible, and persistent view of the computational steps and thought processes leading to a particular conclusion or hypothesis. These "active publications" enable researchers anywhere in the world to re-use a workflow, follow chains of logic, and experiment with changing input data and parameters. Augmented by the KBase Workspace, users can upload their own data, integrate and analyze them with information from the public data stores, and share their data and results with colleagues. As more researchers are able to test the algorithms, datasets, and workflows that are shared with the larger community, they can begin to apply these diverse approaches to their own datasets and provide more informed feedback for improvement. In addition to sharing the information and tools essential to reproducing results, KBase's computational environment is open and extensible, allowing users to add tools and functions to meet their particular research needs.

Pectinases link early fungal evolution to the land plant lineage

Mary Berbee* (mary.berbee@gmail.com)[1], Ying Chang[1], Sishuo Wang[1], Satoshi Sekimoto[1,2], Joseph Spatafora[3], and the AFTOL2 working group

[1]University of British Columbia Vancouver British Columbia; [2]Biological Resource Center, NITE (NBRC), Chiba Japan; [3]Oregon State University, Corvallis, Oregon

Almost all terrestrial fungi in the Ascomycota (molds and yeasts) and Basidiomycota (mycorrhizal and wood decay species) absorb nutrients from plants. The plant/fungal relationship is ancient and recorded in fossils from 400 million years ago. However, fungi first evolved before land plants, and unlike the Asco- and Basidiomycota, the earlier diverging fungal lineages variously extract nutrients from algae, plants or animals. In our community sequencing project with JGI, we sequenced genomes from three early diverging fungi including *Gonapodya prolifera*, an aquatic species that grows on decaying plant material in stagnant water. We hypothesized that if fungi evolved in association with green algae from the land plant lineage, then the *Gonapodya* genome would share fungus-specific genes for the breakdown of the pectins that are specific to the land plants and their algal allies. We reconstructed fungal phylogeny through analysis of 106 'housekeeping' loci. The branching order shows *Gonapodya* in a clade of aquatic flagellated fungi (Chytridiomycota) that diverged early from terrestrial fungi. Analyzing the distribution of pectinases, the *Gonapodya* genome has 27 genes representing 6 of the 7 classes of pectin-specific enzymes known from fungi. While pectinase gene genealogy is difficult to reconstruct due low sequence similarity across taxa, most of the *Gonapodya* pectinases fall into clades with other fungal genes. The gene genealogies do not signal horizontal gene transfer of the *Gonapodya* pectinases.

Indicating functional as well as sequence similarity, *Gonapodya* seems able to use pectin as a carbohydrate source for growth in pure culture. The shared pectinases of Ascomycota, Basidiomycota and *Gonapodya* provide evidence that even ancient fungi were already extracting nutrients from the plants in the green lineage.

Single cell isotope probing of microbes via Raman microspectroscopy: A new way of in situ functional analyses and cell sorting

David Berry* (berry@microbial-ecology.net) and Michael Wagner

Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, Faculty of Life Sciences, University of Vienna, Austria

Using Raman microspectroscopy, microbiologists can within seconds obtain an "optical fingerprint" of individual living microbial cells that contains an astonishing amount of information regarding cellular chemical composition. When combined with the use of stable isotope labeled substrates or deuterated water, this technique offers a direct means to identify microbial cells performing a defined physiological activity within complex samples, which can subsequently be sorted for cultivation or single cell genomics. Using examples from environmental and medical microbiology, the potential of single cell stable isotope probing by Raman microspectroscopy as well as Raman-activated cell sorting in capillaries and microfluidic devices will be demonstrated.

Metagenomic and metaproteomic analyses of symbioses between bacteria and gutless marine worms

Nicole Dubilier (ndubilie@mpi-bremen.de)

Max-Planck Institut

Genomics and proteomics of Zeta proteobacteria

Katrina Edwards (kje@usc.edu)[1], Roman Barco[1], Esther Singer[2], and Gustavo Ramírez[1]

University of Southern California; USC/JGI

Nuetrophilic iron-oxidizing bacteria (FeOB) have been recognized in the environment for well over a century. These bacteria have important ecological and biogeochemical roles mediated through the catalysis of Fe^{2+} -> Fe^{3+}, ultimately resulting in the precipitation of often-copious quantities of iron oxyhydroxide minerals as a by-product of their metabolism. Despite their prevalence and biogeochemical importance, little is known about their physiology in general and specifically, on the molecular mechanism of iron oxidation. In 2011 a draft genome for one iron oxidizing isolate, PV-1, a *Zetaproteobacterium* isolated from a marine environment, was produced. In tandem, tools have been developed for large-scale growth of this bacterium and other procedures necessary for conducting proteomic analysis. Proteomic profiling of PV-1 illuminates several key genes and operons likely to be involved in iron oxidation and electron transport. We present a model for this novel electron transport chain and then use these data to generate qPCR primers targeting key genes from a marine

environmental setting, demonstrating their expression *in situ* during iron oxidation. These are the first functional gene targets developed for screening of the process of biological iron oxidation.

A gene-to-molecule approach to the discovery and characterization of natural products

Michael Fischbach (fischbach@fischbachgroup.or)

University of California at San Francisco, San Francisco, California

Genetic control of mosquitoes

Kelly Matzen* (kelly.matzen@oxitec.com)[1], **Luke Alphey**[1,2]

[1]Oxitec Ltd., Abingdon, United Kingdom; [2]The Pirbright Institute, Woking, United Kingdom

Oxitec Ltd uses synthetic biology approaches to design new methods of pest insect control. Here, we summarize the theoretical design of our genetic systems and the results of both contained and open field trials.

Oxitec's pest population control technology closely resembles the Sterile Insect Technique (SIT), a species-specific and environmentally benign method that has been used successfully against various agricultural pests. The SIT uses the mass release of radiation-sterilised insects to suppress a targeted insect population; any wild-type insect mating with sterile partner produces no offspring, and over time populations are reduced. Our technology uses a genetics-based approach to achieve the same population suppression. Insect strains of the target pest insect, homozygous at one or more loci for a dominant lethal genetic system, are released in large numbers to mate with the wild population; any progeny inherit a copy of the lethal system and therefore die. Release of this type have resulted in the reduction of the target population by over 90% in field trials. Oxitec's genetic approach maintains the advantages of the SIT, but by using engineered insects we can make a range of improvements, including the ability to track our released insects and avoiding the fitness effects that radiation-sterilisation induces.

Through the development of our engineered insects, we and others have begun to develop a modest tool-kit, a set of molecular components, for pest insect synthetic biology. The lethal systems must be conditional as they need to be 'switched off' to allow the strain to be reared prior to release. Thus, we have designed and optimised tetracycline-switchable systems for pest insects (e.g., fruit flies, mosquitoes and moths), using conventional bipartite and also 'positive feedback' circuitry. We have developed female-specific expression systems utilising modular, sex-specific alternative splicing systems as well as employing female-specific promoters, lethal effectors, and fluorescent markers. Combined into single molecular constructs, they constitute the complete circuitry necessary for a genetic-based SIT control system. Our design philosophy, progress and limitations will be discussed, as well as the practical significance of these engineered insect strains.

Uncovering signatures of domestication using genomic resequencing and association mapping

Phil McClean (mcclean@beangenes.cws.ndsu.nodak.edu)

North Dakota State University

Common bean (*Phaseolus vulgaris*) is the most important food legume and is a critical source of protein in many emerging and developing countries. Using the recently released draft sequence of common bean as a reference, a detailed data analysis of DNA pools representing wild germplasm of this major societal crop has again confirmed the existence of ancestral wild germplasm pools located in Mesoamerican (MA) and the Andes. These appeared ~165kya from a common ancestral population. While the MA wild germplasm maintained its size, the wild Andean population underwent a bottleneck (n=3590) that reduced its diversity to 23% of MA wild germplasm. Asymmetric gene flow was strong from the wild MA into the wild Andean population. During domestication, MA landrace diversity was reduced slightly during domestication while Andean landrace diversity increased slightly. Domestication sweeps were observed for ~15% for the MA and Andean landraces genomes. Yet sweeps covering only 7.2 Mb of the genome were shared. As defined by a combined statistic utilizing diversity ratios and F_{ST}, 74Mb underwent a sweep during MA domestication and 60Mb during Andean domestication. Only 7.2Mb were shared. Based on a combined statistic using diversity and F_{ST} data, 1424 and 748 genes were defined as MA and Andean domestication candidates, respectively. Of these, only 59 domestication candidate were in common between the two landraces. Multiple genes with in pathways critical to growth and development were discovered to be domestication candidates. Association mapping confirmed a number of the seed weight candidates.

Making better plants: Synthetic approaches in plant engineering

June Medford (June.Medford@colostate.edu)

Colorado State University

Synthetic biology: Helping the good guys and stopping the bad

Maria Mercedes-Roca (mmroca@zamorano.edu)

Zamorano Panamerican Agriculture School, Honduras

Our understanding of how microbes behave and interact in our labs, our bioreactors, our bodies and in the natural world is exploding exponentially. So is our acknowledgement of the formidable challenges our planet faces, caused in large part by a growing human population and the attendant environmental degradation. These challenges are also incredible opportunities for changing paradigms. Our understanding on how we may use and enhance microorganisms and plants can help us manage human, plant and animal diseases caused by pathogens. We want to stop pathogens. On the other hand, using and enhancing good microbes may help us stay healthier and produce more biomass for food, bioenergy, and materials with less land, water, and oil, and in the face of climate change and rapid loss of biodiversity. Humanity is embarking on a new powerful genomic revolution, coupled with the digital, nanotechnology and other revolutions. The pace of technological change is dizzying and society needs to keep up with it. Society is not, and may become increasingly fearful of advancing technology, especially in the biological sciences. Scientists and policy makers need to acknowledge this.

The field of synthetic biology (SB) offers the promise of revolutionary new products to face these challenges, but also the threat of misusing this technology for nefarious purposes: we must find a way to stop them. However, scientists in the public sector and especially the young generation being initiated through iGEM and other science education initiatives, can bring solutions and ensure that the benefits of these technologies are fairly distributed across the globe: we must not stop and discourage them with unnecessarily onerous and costly regulation. Public perception, activism and science communication play important roles in shaping international and national policy. Scientists and policy makers must acknowledge both the potential of SB and the public's deep mistrust for new, untested technologies they feel are outside their control. Helping the "good guys and stopping the bad", and the future success of SB, depends to a large extent on whether public policy is well-crafted. This paper explores the great challenges and opportunities that lie ahead and the need to involve different disciplines and stakeholders in developing the science, the application, and the public understanding and acceptance.

How humans will survive a mass extinction

Annalee Newitz (annalee@gmail.com)

io9

Unlocking plant metabolic diversity

Anne Osbourn (anne.osbourn@jic.ac.uk)

John Innes Centre, Norwich Research Park, Norwich, United Kingdom

Plants produce a tremendous array of natural products, including medicines, flavours, fragrances, pigments and insecticides. The vast majority of this metabolic diversity is as yet untapped, despite its huge potential value for humankind. So far research into natural products for the development of drugs, antibiotics and other useful chemicals has tended to focus on microbes, where genome sequencing has revolutionised natural product discovery through mining for gene clusters for new metabolic pathways. Identifying novel natural product pathways in plants is extremely difficult because plant genomes are much larger and more complex than those of microbes. However, the recent discovery that genes for some types of plant natural product pathways are organised as physical clusters is now enabling systematic mining of plant genomes in the quest for new pathways and chemistries. Improved understanding of the genomic organization of different types of specialized metabolic pathways will shed light on the mechanisms underpinning pathway and genome evolution. It will further open up unprecedented opportunities for exploiting Nature's chemical toolkit by providing grist for the synthetic biology mill.

Single cell genomics

Stephen Quake (quake@stanford.edu)

Stanford University

Whole genome sequencing of rice mutants to identify genes controlling response to stress and cell wall saccharification

Pamela Ronald (pcronald@ucdavis.edu)

University of California at Davis

Cassava genomics: Applying genomic technologies to benefit smallholder farmers in Africa

Steve Rounsley (rounsley@email.arizona.edu)

University of Arizona

New technologies overview and user feedback

Eddy Rubin, Len Pennacchio (LAPennacchio@lbl.gov**), Jim Bristow**

DOE Joint Genome Institute, Walnut Creek, California

Genome biology of *Miscanthus*

Kankshita Swaminathan (kanksw@gmail.com)

Energy Biosciences Institute

Rich genomics resources facilitate progress in understanding wood formation

Gerald Tuskan (tuskanga@ornl.gov)

Oak Ridge National Laboratory

Understanding the forest microbiome: A fungal perspective

Rytas Vilgalys (fungi@duke.edu)

Duke University

Precise engineering of genomes with sequence-specific nucleases

Dan Voytas (voytas@umn.edu)

Department of Genetics, Cell Biology and Development and Center for Genome Engineering, University of Minnesota

Methods for precisely altering DNA sequences in living cells enable detailed functional analysis of genes and genetic pathways. In plants, targeted genome modification has applications ranging from understanding plant gene function to developing crop plants with new traits of value. Our group has enabled efficient methods for targeted genome modification of plants using sequence-specific nucleases. With zinc finger nucleases (ZFNs), TAL effector nucleases (TALENs), and the CRISPR/Cas9 system, we have achieved targeted gene knockouts, replacements and insertions in a variety of plant species. Current work is focused on optimizing delivery of nucleases and donor DNA molecules to plant cells to more efficiently achieve targeted genetic alterations.

Poster Presentations

Posters alphabetical by first author. *Presenting author

Insights into the global diversity and physiology of '*Aigarchaeota*' through synergistic analysis of single-cell genomes and metagenomes

Timothy W. Alba* (albat@unlv.nevada.edu),[1] Senthil K. Murugapiran,[1] Paul C. Blainey,[2] Jeremy A. Dodsworth,[1] Scott Thomas,[1] Tanja Woyke,[3] Christian Rinke,[3] Iwijn DeVlaminck,[2] Ramunas Stepanauskas,[4] Stephen R. Quake,[2] and Brian P. Hedlund[1]

[1]School of Life Sciences, University of Nevada, Las Vegas, Las Vegas, Nevada; [2]Department of Bioengineering, Stanford University, and Howard Hughes Medical Institute, Stanford, California; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine

The '*Aigarchaeota*' is a candidate phylum in the domain Archaea that is globally distributed in thermal environments. However, the phylogeny and distribution of the '*Aigarchaeota*' has not been studied in detail and, due to a lack of cultivated representatives, little is known about the physiological potential of this lineage. Data from 27 single-cell genomes and four metagenomes obtained from terrestrial geothermal springs in the U.S. Great Basin and Tengchong County, China, were analyzed synergistically to obtain five nearly-complete genomes, each representing a genus-level group of the archaeal phylum '*Aigarchaeota*'. Single-cell genomes were obtained via two distinct methods involving amplification of femtogram quantities of DNA obtained via standard fluorescence-activated sorting, and using a novel technique involving microfluidics. These single-cell genomes were used to enable '*Aigarchaeota*'-targeted binning of metagenome datasets. 16S rRNA gene phylogenetic analysis showed that the genomes described here cover the majority of known diversity within the '*Aigarchaeota*', including one genus-level group that appears to be ubiquitous in both marine and terrestrial habitats. Phylogenomic and whole genome analysis supported the positioning of '*Aigarchaeota*' within the recently proposed 'TACK' superphylum. Metabolic reconstruction suggests a diversity of anaerobic lifestyles within the '*Aigarchaeota*', including both chemoorganotrophic and chemolithotrophic activities and both anaerobic respiration via dissimilatory sulfate reduction and Embden-Meyerhof glycolysis with production of hydrogen, acetate, and ethanol. Single '*Aigarchaeota*' cells visualized during microfluidic cell sorting had a long, thin, rod-shaped or filamentous morphology, and all '*Aigarchaeota*' genomes encode the capacity for flagellar motility. This study demonstrates some metabolic diversity around a common metabolic theme, anaerobic chemoheterotrophy, imparting roles in both carbon and sulfur cycling in geothermal habitats.

KBase Overview

Adam P. Arkin* (aparkin@lbl.gov)[1], Jason Baumohl[1], Aaron Best[2], Jared Bischof[2], Ben Bowen[1], Tom Brettin[2], Tom Brown[2], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Ric Colasanti[2], Neal Conrad[2], Brian Davison[3], Matt DeJongh[6], Paramvir Dehal[1], Narayan Desai[2], Scott Devoid[2], Terry Disz[2], Meghan Drake[3], Janaka Edirisinghe[2], Gang Fang[7], José Pedro Lopes Faria[2], Mark Gerstein[7], Elizabeth M. Glass[2], Annette Greiner[1], Dan Gunter[1], James Gurtowski[5], Nomi Harris1, Travis Harrison[2], Fei He[4], Matt Henderson[1], Chris Henry[2], Adina Howe[2], Marcin Joachimiak[1], Kevin Keegan[2], Keith Keller[1], Guruprasad Kora[3], Sunita Kumari[5], Miriam Land[3], Folker Meyer[2], Steve Moulton[3], Pavel Novichkov[1], Taeyun Oh[8], Gary Olsen[9], Bob Olson[2], Dan Olson[2], Ross Overbeek[2], Tobias Paczian[2], Bruce Parrello[2], Shiran Pasternak[5], Sarah Poon[1], Gavin Price[1], Srivdya Ramakrishnan[5], Priya Ranjan[3], Bill Riehl[1], Pamela Ronald[8], Michael Schatz[5], Lynn Schriml[10], Sam Seaver[2], Michael W. Sneddon[1], Roman Sutormin[1], Mustafa Syed[3], James Thomason[5], Nathan Tintle[6], Will Trimble[2], Daifeng Wang[7], Doreen Ware[5], David Weston[3], Andreas Wilke[2], Fangfang Xia[2], Shinjae Yoo[4], Dantong Yu[4], Robert Cottingham[3]* , Sergei Maslov[4] , Rick Stevens[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Argonne National Laboratory, Argonne, Illinois; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Brookhaven National Laboratory, Upton, New York; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [6]Hope College, Holland, Michigan; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

Systems biology is driven by the ever-increasing wealth of data resulting from new generations of genomics-based technologies. The advancement of systems biology relies not only on sharing the results of projects through traditional methods of peer review and publication, but also on sharing the datasets, workflows, software, models, best practices, and other essential knowledge that made those published results possible. Establishing a common framework for managing this knowledge could save time, reduce duplicative effort, and increase the scientific return on investment in systems biology research. This framework for precisely tracking what was done to achieve a certain outcome also will empower researchers to reproduce published results and review projects more effectively. This ability to reproduce experiments described in publications is a key principle of the growing Open Science movement.

The KBase team is developing an open-source, open-architecture framework for reproducible and collaborative computational systems biology. KBase's primary scientific aim is to push multiple types of functional data towards increasingly specific models of metabolic and regulatory behavior of microbes, plants and their communities. We have brought together data and tools that allow probabilistic modeling of gene function, which can be used in turn to produce experimentally testable models of cellular metabolism and gene regulation.

One of the new components that will enable these complex, iterative analyses is KBase's new prototype Narrative Interface, which provides a transparent, reproducible, and persistent view of the computational steps and thought processes leading to a particular conclusion or hypothesis. These “active publications” enable researchers anywhere in the world to re-use a workflow, follow chains of logic, and experiment with changing input data and parameters.

Reconstruction of food webs in biological soil crusts using metabolomics

Richard Baran* (RBaran@lbl.gov),[1] Jazmine Mayberry-Lewis,[2] Ulisses Nunes Da Rocha,[3] Benjamin P. Bowen,[1] Eoin L. Brodie,[3] Hinsby Cadillo-Quiroz,[2] Ferran Garcia-Pichel,[2] and Trent R. Northen[2]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [2]School of Life Sciences, Arizona State University, Tempe, Arizona; [3]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Biological soil crusts (BSCs) are communities of organisms inhabiting the upper layer of soil in arid environments. BSCs persist in a dessicated dormant state for extended periods of time and experience pulsed periods of activity facilitated by infrequent rainfall. *Microcoleus vaginatus*, a non-diazotrophic filamentous cyanobacterium, is the key primary producer in BSCs in the Colorado Plateau and is an early pioneer in colonizing arid environments. Over decades, BSCs proceed through developmental stages with increasing complexity of constituent microorganisms and macroscopic properties. Metabolic interactions among BSC microorganisms probably play a key role in determining the dynamics of constituent populations and cycling of carbon and nitrogen. However, these metabolic interactions have not been studied systematically. We performed untargeted mass spectrometry-based metabolite profiling of cell extracts and spent media of *Microcoleus* and six heterotrophic bacterial isolates from BSCs. The growth media of these bacteria were also supplemented with either the metabolite extract of *Microcoleus* or a pooled metabolite extract of the six heterotrophs. Metabolite profiles of spent media were compared to those of control media to detect uptake or release of specific metabolites by the respective bacteria. *Microcoleus* was found to release a very broad range of metabolites, while it uptakes a small set of central metabolites. These metabolites were also taken up by the heterotrophs, but were not released into any media type by any of the bacteria. This observations points to a competition for a small set of central metabolites by all analyzed bacteria if the metabolites are released into the environment due to cell lysis. Surprisingly, the studied heterotrophs were found to utilize only a subset of metabolites released by *Microcoleus* suggesting a specialization of individual heterotrophs towards limited sets of organic nutrients. To our knowledge, this study is the first broad scale molecular level analysis of potential metabolic interactions among bacteria in microbial communities.

Fungal perception of nutritional cues: A comparative transcriptomics analysis of resource utilization in *Neurospora crassa*

J. Philipp Benz* (ph.benz@berkeley.edu)[1], John E. Dueber[1,2], Chris R. Somerville[1,3], and N. Louise Glass[1,3]

[1]Energy Biosciences Institute, [2]Department of Bioengineering, [3]Department of Plant and Microbial Biology, University of California at Berkeley, Berkeley, California

The need for a stable and sustainable source of energy is one of the grand challenges of our generation. Using lignocellulosic plant biomass as a renewable feedstock e.g. for biofuel production can contribute substantially on our way to meet this goal. However, for highest efficiency the utilization of all major constituent sugars is required. To this end, enzyme cocktails adapted to the individual composition of given feedstocks would be desirable as well as the engineering of downstream organisms able to ferment all major monosaccharides. In particular fungi are among the most promising tools to deconstruct plant cell wall polysaccharides due to their importance in global carbon recycling, but much

remains to be learned about their enzyme production and sugar utilization mechanisms. A more profound knowledge of the fungal interaction with different plant substrates will therefore be highly informative for efforts to achieve a cost-effective biofuel production process.

So far, we performed systematic analyses of how *Neurospora crassa* perceives the presence of the three major plant-derived polysaccharides cellulose, xylan and pectin. Although still low in resolution, the comparison of the individual responses allowed us to deduce fundamental knowledge on fungal adaptation strategies to the carbon composition in its environment at a genome-wide level. Our study moreover revealed new factors involved in specific sugar utilization pathways. For example, we identified the first eukaryotic high-affinity MFS-type transporter for the uptake of galacturonic acid (GAT-1), which was used for pathway engineering in yeast to generate a transporter-dependent uptake and catalysis system into downstream products with high potential for utilization as platform chemicals.

As part of the JGI Community Sequencing Program "The Fungal Nutritional ENCODE Project" (lead-PI: N. Louise Glass), we are planning to dramatically increase the resolution of our gene regulation models as of today. By transcriptional analysis of the fungal response to several hundred nutritional and light regimes using RNA-Seq, and by leveraging the genomic resources of *N. crassa*, such as the transcription factor deletion strain library, we will create a nutritional regulatory map important for the discovery of the function of many additional factors and regulons associated with resource utilization. These data will be informative for the rational design of fungal strains for biomass degradation and conversion, and at the same time serve as a community tool for hypothesis-driven research on *N. crassa* as well as a wide spectrum of other filamentous fungi.

Elimination of hydrogenase active site assembly blocks H_2 production and increases ethanol yield in *Clostridium thermocellum*

Ranjita Biswas,[1,2] Charlotte M. Wilson,[1,2] Tianyong Zheng,[2,4] Richard J. Giannone,[2,3] Dawn M. Klingeman,[1,2] Robert L. Hettich,[2,3] Daniel G. Olson,[2,4] Lee R. Lynd,[2,4] Steven D. Brown,[1,2] and **Adam M. Guss*** (gussam@ornl.gov)[1,2]

[1]Biosciences Division, [2]BioEnergy Science Center, [3] Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4] Thayer School of Engineering at Dartmouth College, Hanover, New Hampsire

The native ability of *Clostridium thermocellum* to rapidly consume cellulose and produce ethanol makes it a leading candidate for a consolidated bioprocessing (CBP) biofuel production strategy. *C. thermocellum* also synthesizes lactate, formate, acetate, H_2, and amino acids that compete with ethanol production for carbon and electrons. Elimination of H_2 production could redirect carbon flux towards ethanol production by making more electrons available for acetyl-CoA reduction to ethanol. *C. thermocellum* encodes four hydrogenases and rather than delete each individually, we targeted a hydrogenase maturase gene (*hydG*), involved in converting the three [FeFe] hydrogenase apoenzymes into holoenzymes. Further deletion of the [NiFe] hydrogenase (*ech*) resulted in a mutant that functionally lacks all four hydrogenases. H_2 production in *ΔhydGΔech* was undetectable and ethanol yield increased nearly 2-fold compared to wild type. Interestingly, mutant growth improved upon the addition of acetate, which led to increased expression of genes related to sulfate metabolism, suggesting these mutants may use sulfate as a terminal electron acceptor to balance redox reactions. Genomic analysis of *ΔhydG* revealed a mutation in *adhE*, resulting in a strain with both NADH- and NADPH-dependent alcohol dehydrogenase activities. While this same *adhE* mutation was found in ethanol tolerant *C. thermocellum* strain E50C, *ΔhydG* and *ΔhydG Δech* are not more ethanol tolerant

than wild type, illustrating the complicated interactions between redox balancing and ethanol tolerance in *C. thermocellum*. The dramatic increase in ethanol production here suggests that targeting protein post-translational modification is a promising new approach for inactivation of multiple enzymes simultaneously for metabolic engineering.

Learning direct microbial interaction networks from perturbations on artificially constructed bacterial communities

Surojit Biswas* (sbiswas@live.unc.edu),[1] Nathan P. McNulty,[2] Derek S. Lundberg,[3] Jeffrey I. Gordon,[2] Jeffery L. Dangl,[3,4] and Vladimir Jojic[5]

[1]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; [2]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri; [3]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; [4]Howard Hughes Medical Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; [5]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Microbes that associate with higher eukaryotes often perform metabolic functions that underpin host vitality. Importantly, they coexist with various symbiotic relationships. Understanding the structure of their interactions may simplify the list of microbial targets that can be modulated for host benefit.

While correlation based methods are a popular approach to study interrelationships between microbes, two correlated microbes need not interact if, for example, they are co-regulated by a third. Gaussian graphical models remedy this concern by inferring direct associations; however, they assume that the data in hand is normally distributed. Finally, neither correlation nor Gaussian graphical modeling offer a systematic way to control for confounding covariates. We therefore developed a Poisson-multivariate normal hierarchical model that controls for confounding predictors at the Poisson layer, and parsimoniously captures direct microbe-microbe interactions at the multivariate normal layer using an L1 penalized precision matrix.

To test the model with real data, we inoculated the sterilized and inert calcined-clay soil of each of 46 *Arabidopsis thaliana* plants with 9 bacterial soil-isolates in varying relative abundances in order to perturb any underlying interaction structure. Four weeks post inoculation, we harvested and 16S profiled each root and derived a 46-samples x 9-isolates count matrix. Application of our model to this count table, while entering starting input abundance, log-sequencing depth, and experimenter as predictors, revealed strong predicted antagonisms between isolate pairs (181,105), (105,50), and (303,8). *In vitro* co-plating experiments corroborated these predictions in direction and also semi-quantitatively.

Additionally, we applied our model to data from three independent artificial community experiments previously published by the Gordon lab. Each study had an intellectually unique focus on a specific host-microbe dynamic in gnotobiotic mice. Surprisingly, in all three datasets our model predicted a strong, positive interaction between *Clostridium scindens* and *Collinsella aerofaciens*, and the presence of an inert strain, *Bacteroides uniformis*. In collaboration with the Gordon lab, we are presently validating these predictions using a series of single-member drop-out and drop-in experiments.

We conclude that our Poisson-multivariate normal hierarchical demonstrates potential to infer true, direct microbe-microbe interactions in artificially constructed communities. Importantly, explicitly modeling counts and accounting for confounding predictors is necessary to detect these interactions.

Our ongoing collaboration with the Gordon lab will help to further establish our method's accuracy for detecting direct microbe-microbe interactions. We envision using our model to help engineer stable microbial consortia that are deliverable as a probiotics to plant or human hosts.

Molecular dissection of stomatal infection in the maize foliar pathogen *Cercospora zeae-maydis* through comparative and functional genomics

Burt H. Bluhm* (bbluhm@uark.edu)[1], R. Louis Hirsch[2], John Ridenour[1], Jonathon Smith[1], Ahmad Fakhoury[3], Shantae Wilson[1], and Kara Troglin[1]

[1]Department of Plant Pathology, University of Arkansas, Division of Agriculture, Fayetteville, Arkansas; [2]Department of Mathematics and Sciences, Coker College, Hartsville, South Carolina; [3]Department of Plant, Soil, and Agricultural Systems, Southern Illinois University, Carbondale, Illinois

Cercospora zeae-maydis causes gray leaf spot, a ubiquitous and often devastating foliar disease of maize. The infection biology underlying gray leaf spot is complex: *C. zeae-maydis* senses host stomata, which it then grows towards and utilizes as portals of entry into maize leaves, followed by a latent, hemi-biotrophic phase preceding the induction of necrosis. Our focus is to understand the molecular basis of infection in *C. zeae-maydis* and related Dothidiomycete fungi, with the overarching goal of using this information to improve disease management. To dissect pathogenesis, we have employed a combination of techniques in molecular genetics and functional genomics, which was substantially accelerated by the public release of the *C. zeae-maydis* genome sequence by DOE-JGI in 2011. Recently, through reverse genetics, we established a genetic linkage between light sensing, the innate circadian clock of *C. zeae-maydis*, and stomatal perception by the fungus. Additionally, we identified putative epigenetic regulators governing the transition to hemi-biotrophy through a forward genetic screen (tagged mutants with deficiencies in appressorium formation) combined with differential expression profiling via RNA-seq (genes highly expressed in *C. zeae-maydis* during apressorium formation). The forward genetic component was facilitated by whole-genome resequencing of selected mutants to precisely characterize the site(s) of disruption. Furthermore, we are utilizing comparative genomics to identify putative regulators of pathogenesis in *C. zeae-maydis*. To this end, we have recently obtained high-quality draft genome sequences for *C. zeina* (a sibling species of *C. zeae-maydis* that also causes gray leaf spot of maize), *C. sojina* (causal agent of frogeye leaf spot of soybean), and *C. kikuchii* (causal agent of Cercospora leaf blight of soybean). Of particular interest are genes undergoing diversifying selection, displaying patterns of presence/absence that correlate with host range or other relevant parameters, or are unique at the genus or species level. Although many questions remain to be answered, our findings to date have provided the most complete understanding thus far of stomatal infection by a plant pathogenic Ascomycete. Importantly, our findings have led to the formulation of new hypotheses regarding how *C. zeae-maydis* senses stomata, the genetic and morphogenic basis of the transition to hemi-biotrophy, and mechanisms through which necrosis is induced during the final stages of pathogenesis.

JAMO: A scaleable metadata, data provenance and file organizer

Alex E. Boyd* (aeboyd@lbl.gov)[1], Chris Beecroft[1], and Kjiersten Fagnan[2]

[1]DOE Joint Genome Institute, Walnut Creek, California; NERSC, Lawrence Berkeley National Laboratory

As data ages and old file systems become obsolete, data files and information become increasingly difficult to locate. The problem is compounded as data loses ownership or when institutional knowledge about the data becomes lost. Furthermore these "lost" files continue to consume space on local disks reducing space available for new data.

The JGI Archiver and Metadata Organizer (JAMO) was designed to solve these problems. JAMO stores metadata about every file and/or collection in a flexible key value store using MongoDB that allows queries across every record against any key. In addition, JAMO's hierarchical data management system addresses data life-cycle as older, less-used data is regulated to higher-latency secondary storage systems freeing up space for newer and often used data. JAMO is also language agnostic as all communication to the system is done though the RESTful interface standard.

Recent divergence in fungal populations

Sara Branco* (sara.mayer.branco@gmail.com), John Taylor, and Tom Bruns

Department of Plant and Microbial Biology, University of California at Berkeley, Berkeley, California

Genetic differentiation between populations, signaling the effects of reduced gene flow and adaptation that are the main drivers of speciation, is essential to understanding biodiversity. Very little is currently known about how fungal populations differentiate in nature. Here, we investigate the divergence of a coastal and a mountain population of *Suillus brevipes*, a widespread mycorrhizal fungus associated with pine forests. Whole genome assembly and resequencing of 28 *S. brevipes* individuals revealed 567,192 shared SNPs that supported two recently diverged populations, each with low genetic diversity. As expected, mean genome-wide Fst is very low, indicating little overall population differentiation. However, genome-wide scans for regions of high differentiation using 5kb windows, reveal several highly differentiated regions, that harbor eight genes coding for five known proteins associated with cellular processes and signaling. Demographic inference suggests the two populations are isolated, with no migration between the coast and mountain. Population genetic differentiation and adaptation can occur by natural selection or genetic drift. Tests of the ratio of non-synonymous to synonymous mutations and identity of genes in the highly diverged regions may provide evidence of natural selection and suggest modes of adaptation. Further research including other mountain and coastal *S. brevipes* populations will be needed to identify the sources of the highly diverged regions.

Metagenome-enabled investigations of carbon and hydrogen fluxes within the serpentinite-hosted subsurface biosphere

William J. Brazelton* (william.brazelton@utah.edu)[1] and Matthew O. Schrenk[2]

[1]Department of Biology, University of Utah, Salt Lake City, Utah; [2]Department of Geological Sciences, Michigan State University, East Lansing, Michigan

Ultramafic rocks in the Earth's mantle represent a tremendous reservoir of carbon and reducing power. Upon tectonic uplift and exposure to fluid flow, serpentinization of these materials generates copious

energy, sustains abiogenic synthesis of organic molecules, and releases hydrogen gas (H_2). Microbial communities hosted within serpentinites may be important mediators of carbon and energy exchange between the deep Earth and the surface biosphere. Actively serpentinizing rocks are present on all of the world's continents, comprise significant portions of the deep seafloor, generate large quantities of geochemical energy, and yet are some of the most poorly understood portions of the biosphere. Our team is involved in a series of ongoing interdisciplinary investigations aimed at defining the global serpentinite microbiome in the context of detailed chemical and physical data.

Our JGI community sequencing project includes complete genome sequencing of cultivated isolates and metagenome and metatranscriptome sequencing of subsurface fluid samples collected from several sites of active serpentinization. The intent of this project is to improve our understanding of both the taxonomic and functional diversity of microbial communities in the serpentinite subsurface. Specifically, we intend to target catabolic processes and microbial interactions with carbon pools (autotrophy, fermentation, methanogenesis, respiration). By studying community genomes in the context of detailed environmental data, we will also be able to resolve physiological adaptations to the serpentinite microbiome, with implications for both culturing approaches and practical applications (*e.g.* carbon capture and storage, extremozymes, alternative energy). By targeting the metagenome and metatranscriptome of such samples in parallel, we will assess both the functional potential and the activities of serpentinite-hosted subsurface microbial communities.

Our previous results have revealed genes involved in lithotrophy, fermentation, and hydrogen oxidation, suggesting that the dominant organisms are supported by serpentinization-related processes (Brazelton *et al*., 2012). More specifically, our results point to H_2-utilizing Betaproteobacteria thriving in shallow, oxic-anoxic transition zones and anaerobic Clostridia thriving in anoxic, deep subsurface habitats. These data demonstrate the feasibility of metagenomic investigations into novel subsurface habitats via surface-exposed seeps and indicate the potential for H_2-powered primary production in serpentinite-hosted habitats. Analyses of our new JGI CSP data are ongoing, and preliminary results indicate extraordinary taxonomic similarities throughout the geographic range of serpentinite microbiomes we have sampled.

Brazelton, W.J., B. Nelson and M.O. Schrenk (2012) Metagenomic evidence for H_2 oxidation and H_2 production by serpentinite-hosted subsurface microbial communities. *Frontiers in Extreme Microbiology* 2:268. doi: 10.3389/fmicb.2011.00268.

Developing for KBase

Thomas Brettin* (brettin@cels.anl.gov),[1] Rick Stevens,[1] Jason Baumohl,[2] Aaron Best,[3] Jared Bischof,[1] Ben Bowen,[2] Tom Brown,[1] Shane Canon,[1] Stephen Chan,[2] John-Marc Chandonia,[2] Dylan Chivian,[2] Ric Colasanti,[1] Neal Conrad,[1] Brian Davison,[4] Matt DeJongh,[3] Paramvir Dehal,[2] Narayan Desai,[1] Scott Devoid,[1] Terry Disz,[1] Meghan Drake,[4] Janaka Edirisinghe,[1] Gang Fang,[7] José Pedro Lopes Faria,[1] Mark Gerstein,[7] Elizabeth M. Glass,[1] Annette Greiner,[2] Dan Gunter,[2] James Gurtowski,[6] Nomi Harris,[2] Travis Harrison,[1] Fei He,[5] Matt Henderson,[2] Chris Henry,[1] Adina Howe,[1] Marcin Joachimiak,[2] Kevin Keegan,[1] Keith Keller,[2] Guruprasad Kora,[4] Sunita Kumari,[6] Miriam Land,[4] Folker Meyer,[1] Steve Moulton,[4] Pavel Novichkov,[2] Taeyun Oh,[8] Gary Olsen,[9] Bob Olson,[1] Dan Olson,[1] Ross Overbeek,[1] Tobias Paczian,[1] Bruce Parrello,[1] Shiran Pasternak,[6] Sarah Poon,[2] Gavin Price,[2] Srivdya Ramakrishnan,[6] Priya Ranjan,[4] Bill Riehl,[2] Pamela Ronald,[8] Michael Schatz,[6] Lynn Schriml,[10] Sam Seaver,[1] Michael W. Sneddon,[2] Roman Sutormin,[2] Mustafa Syed,[4] James Thomason,[6] Nathan Tintle,[3] Will Trimble,[1] Daifeng Wang,[7] Doreen Ware,[5,6] David Weston,[4] Andreas Wilke,[1] Fangfang Xia,[1] Shinjae Yoo,[5] Dantong Yu,[5] Robert Cottingham,[4] Sergei Maslov,[5] and Adam P. Arkin[2]

[1]Argonne National Laboratory, Argonne, Illinois; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Hope College, Holland, Michigan; [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Brookhaven National Laboratory, Upton, New York; [6]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

The KBase team is developing an open-source, open-architecture framework for reproducible and collaborative computational systems biology. One of the key operating principles of KBase is to allow the scientific community to incorporate their own algorithms into the system to make them available to others easily; to avail themselves of the KBase computational architecture; and to make use of the KBase data sources. The KBase team aims to make this process simple and to provide an easy route for dissemination of new tools and comparison to existing tools in a common framework.

KBase system design is firmly rooted in our service-oriented architecture and in our commitment to open source development and distribution models. These aspects of the project have been with the system since its inception, and will continue to be a guiding force for future work. The KBase system design is based on several sound best practice principles including consistent code use, code re-use, and the decoupling of modular system components. We have established standard software engineering processes for version control, software and data builds, testing, QA/QC, deployments and releases. These enable the deployment of a large number of services by a relatively small release engineering team.

To prepare for potential future services contributed by the community, we provide developer training materials on our website as well in hands-on developer training sessions called bootcamps. In the future, we plan to offer a wider range of bootcamps and webinars to target different types of developers and different scientific focus areas. In the meantime, prospective developers and computational biologists can find information about KBase service design at http://kbase.us/developer-zone/.

Niche adaptation of Methanomicrobiales methanogens revealed by comparative genomics

Patrick Browne and **Hinsby Cadillo-Quiroz*** (hinsby@asu.edu)

School of Life Sciences, Arizona State University, Tempe, Arizona

The Methanomicrobiales order of methanogens are the dominant methanogens in a a broad range of terrestrial and freshwater anaerobic habitats. Their contribution of methane (CH_4) to the atmosphere is important to consider in light of global climate change. Yet they remain understudied relative to the class I methanogens and the Methanosarcinales. We examine the genomes of nine Methanomicrobiales strains isolated from a diversity of habitats including Northern peatlands, tar pits and anaerobic sludges digesting a variety of organic wastes. Although all nine strains are capable of reducing carbon dioxide (CO_2) to CH_4, via the same basic biochemical steps, differences in reducing equivalent trafficking and energy conservation during methanogenesis were identified and are likely implicated in niche adaptation. Our results imply that electron bifurcation between heterodisulfide reductase and formylmethanofuran dehydrogenase occurs via a different complex of proteins to that characterized in the class I methanogens, highlighting the need for more biochemical studies focused on the Methanomicrobiales order. Adaptations to the peatland environment were focused on coping with the low availability of inorganic nutrients. The peatland strains, *Methanoregula boonei* and *Methanosphaerula palustris*, shared a complement of inorganic ion transporters not found in the other strains, including an ATP-dependent potassium uptake system not commonly found in the Archaea. The requisite genes to fix atmospheric nitrogen were also found in both peatland strains and in *Methanoplanus petrolearius*, a strain isolated from an offshore oil field. *Methanocorpusculum labreanum*, isolated from a tar pit showed a clear trend of genome downsizing leading to niche specialization. *M. labreanum* was the only strain that lacked genes encoding an archaeal flagellum as well as chemotaxis related genes. Despite having the smallest genome, *M. labreanum* had the highest inferred rate of horizontal gene acquisition, and acquired far more genes from the Firmicutes and far fewer genes from the Proteobacteria in comparison with the other eight strains.

Anchored assembly: An algorithm for large structural variant detection using NGS data

Jeremy Bruestle, **Becky Drees**, Niranjan Shekar* (niranjan@spiralgenetics.com)

Spiral Genetics, Seattle, Washington

Characterization of large indels, inversions, and multi-nucleotide variants is important for microbial genomics and agrigenomics studies. These are often undetected by standard pipelines. Spiral Genetics has developed Anchored Assembly, a novel method using direct, *de novo* read overlap assembly to accurately detect variants from next-generation sequence reads. We detect, on average, over 90% of indels and structural variants up to 30 kbp in non-repetitive regions. The ability to detect deletions and structural variants is undiminished by variant size, and the ability to accurately detect and assemble insertions continues well into the 30 kbp range.

We assembled 100 bp Illumina reads from *E. coli* MG1655 using an *E. coli* DH10B reference. Anchored Assembly reproduced over 95% of the structural differences between DH10B and MG1655 *E. coli* strains, including a 56 kbp insertion and 144 kbp deletion, and detected 99% of SNVs.

Anchored Assembly was evaluated against Pindel and BWA + GATK using simulated read data. Datasets were generated by populating chromosome 22 of the human genome reference sequence with a set of SNPs, insertions, deletions, inversions, and tandem repeats. Overall, Anchored Assembly detected over 90% of indels and structural variants up to 50 kbp and SNPs with false discovery rates well below 1%. In comparison, Pindel and BWA + GATK had overall false discovery rates of 10% and 9%, respectively.

Anchored Assembly's extensive detection range and low false discovery rates may benefit a number of applications. The ability to accurately detect structural differences will be useful for characterizing microorganism strains and drafting genomes using a reference of a sufficiently closely related species.

Validation of molecular targets associated with reduced biomass recalcitrance in naturally occurring *Populus trichocarpa*

Anthony C. Bryan* (bryanac@ornl.gov)[1], Wellington Muchero[1], Jianjun Guo[1], Hengfu Yin[1], Sara S. Jawdy[1], Olaf Czarnecki[1], Lee E. Gunter[1], Jennifer L. Morrell-Falvey[1], Stephen P. DiFazio[2], Gancho T. Slavov[3], Gerald A. Tuskan[1], and Jin-Gui Chen[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Department of Biology, West Virginia University, Morgantown, West Virginia; [3]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Gogerddan, Aberystwyth Ceredigion, United Kingdom

Renewability, sustainability and affordability are attributes fueling the development of alternative energy sources. With a specific emphasis on biofuel development, *Populus* represents a viable energy alternative through the conversion of lignocellulosic biomass into biofuel. The high cost of extracting sugars from the recalcitrant biomass through chemical or enzymatic treatments has necessitated the identification of poplars with higher sugar release (e.g. reduced recalcitrance) potential. Our objective is to develop *Populus* feedstocks with favorable traits that can be exploited by the biofuel industry.

To identify favorable genetic backgrounds and loci regulating desirable traits, a genome-wide association study was conducted utilizing naturally occurring *P. trichocarpa* variants genotyped with 34,160 single nucleotide polymorphisms (SNPs). Several loci were identified with high association to cell wall chemistry phenotypes including sugar release, S/G ratio and lignin content. Through this analysis we identified several proteins with diverse molecular functions including transcription factors, kinases and metabolic enzymes that have previously not been associated with cell wall chemistry phenotypes. These loci were verified based on sequence analysis which identified alternative allelic variants segregating in the population. Variant alleles of individual loci were transfected into *Populus* protoplasts. We determined the effects of overexpressing the individual allelic variants on cell wall chemistry in protoplasts using an expression analysis of specific genes known to function in the biosynthetic pathways of cellulose (*CESA8*), hemicellulose (*GT43B*) or lignin (*CCOAOMT1*). We identified distinct effects on these markers for the cell wall biosynthesis pathways for the individual allelic variants. One locus, representing a potential transcription factor that may also function outside the nucleus, had an allelic variant in the form of extra poly glutamine repeats (PolyQ). This variant showed a reduced amount of CCOAOMT1 expression. Additional functional analyses of identified target genes were tested using T-DNA disruptions of homologous loci in *Arabidopsis thaliana*. Loss-of-function mutant lines of the *Arabidopsis* homolog of the *Populus* PolyQ locus showed a reduction in lignin content. The extra PolyQ repeat in this variant locus may disrupt protein function, similar to PolyQ repeats disruption in animals, leading to reduced lignin accumulation in *Populus* lines harboring variant. This PolyQ locus may represent a novel regulator of cell wall chemistry and is a potential target for engineering reduced recalcitrance of lignocellulosic biomass.

Fungal host-association as a driver of metabolic diversity

Kathryn E. Bushley* (bushleyk@science.oregonstate.edu),[1] C. Alisha Owensby,[2] and Joseph W. Spatafora[2]

[1]Department of Plant Biology, University of Minnesota, St. Paul, Minnesota; [2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon

Fungal metabolism plays a key role in pathogenesis of fungi on both plants and animals as well as adaption of fungi to particular ecological niches. Drawing on examples from the fungal order Hypocreales, which includes pathogens of plants, insects/animals, and other fungi, we examine changes in metabolic potential associated with a shift from plant to either insect/animal hosts. Using a combination of comparative genomics, phylogenomics, and transcriptomics we examine expansions and contractions in gene families (CAZymes, P450s, proteases) involved in degradation of distinct host carbon and nitrogen sources in plant versus insect or fungal associated lineages. We also identify gene clusters involved in production of secondary metabolites that are either unique to insect pathogens or upregulated in media containing insect cuticle or insect hemolymph. Focusing on nonribosomal peptide synthetases (NRPSs), a class of secondary metabolite multimodular enzymes which produce small bioactive peptides (NRPs) without the aid of ribosomes, we examine expansions and contractions of phylogenetic clades of NRPS adenylation domains which correlate with host. We also address the extent to which these correlations result from independent expansions versus acquisition by horizontal transfer and address evolutionary and genetic mechanisms that contribute to diversification of both NRPS genes and their corresponding chemical products.

BBMap: A fast, accurate splice-aware aligner for short and long reads

Brian Bushnell* (bbushnell@lbl.gov)

DOE Joint Genome Institute, Walnut Creek, California

http://sourceforge.net/projects/bbmap/

Current alignment software is capable of running fast enough to process the huge volume of 3rd-generation sequencer output, or sensitively enough to map one species or strain to another or tumor to normal tissue, but not both. The commonly used fast aligners have severe limitations regarding indel length, number of mismatches, maximum reference genome size, and ability to process RNA-seq data that limit their usefulness outside of examining SNPs in complete genomes such as Homo Sapiens.

BBMap is a mapping and alignment program designed to surpass these limits, processing reads from 3rd-generation or single-molecule sequencing technologies (such as PacBio) at greater speed, sensitivity, and specificity than any other high-throughput alignment software. It is capable of processing RNA-seq data without a splice database, achieving greater speed and mapping rates than Tophat. It can index far more rapidly than Burrows-Wheeler based algorithms, which is very useful for examining multiple assemblies of the same organism. It also has several novel features designed to enhance JGI's quality control and metagenomic assembly, such as binning reads by mapping to multiple references simultaneously and generating insert size and defect histograms internally, rather than as a second pass on a sam file.

This presentation will describe how BBMap works, and show some results comparing the speed and accuracy of BBMap to other software including bwa, Bowtie2, and TopHat.

BBMap is written in pure Java and is platform independent, with no recompilation needed, making distribution and installation trivial. It is available open source under a BSD license.

Assessment of mercury resistance in Fosmid library from mangrove sediment

Lucélia Cabral* (luc.g.cabral@gmail.com), Sanderson Tarciso Pereira de Sousa, Gileno Vieira Lacerda Júnior, Júlia Ronzella Ottoni, Daniela Ferreira Domingos, and Valéria Maia de Oliveira

Divisão de Recursos Microbianos, Research Center for Chemistry, Biology and Agriculture (CPQBA), Campinas University – UNICAMP, São Paulo. Brazil.

Brazil holds one of the largest mangrove extensions in the world, an ecosystem very rich in biodiversity, including aquatic animals, birds, reptiles, mammals and microorganisms. In Brazil, mangroves are areas of permanent environmental protection. Despite of this, mangroves have been highly exposed to anthropic activities, including oil spills and industrial wastes that carry heavy metals. Microorganisms found in the environment can adapt to the presence of pollutants, thus developing survival mechanisms. In spite of the great diversity and relevance of the microorganisms that inhabit mangroves, scarce research has been carried out about this subject in Brazil. Although improvements on in vitro techniques have been made in order to recover microorganisms from these sites, the knowledge of the mangrove microbiota remains incipient. In this context, the aim of this study was to assess the presence of mercury resistance in one fosmid library constructed using sediment samples collected from a mangrove area located in Bertioga, State of São Paulo, Brazil. The fosmid library comprised 13,000 clones and the sampling site was moderately impacted by anthropogenic activities. The functional screening of clones resistant to mercury was carried out according to methodology described by Freeman et al. (2005) with few modifications. Briefly, clones were grown in microplates (96 wells) containing LB liquid medium and chloramphenicol (12.5 µg/ml), followed by incubation at 37^{o} C in a rotary shaker (150 rpm) for 18 h. After incubation, aliquots of the clone culture (10 µl) were transferred to another microplate filled with 150 µl LB broth containing chloramphenicol (12.5 µg/ml) and mercury. *E. coli* EPI 300 was used as negative control. After 48 h, the growth was measured using Elisa spectrophotometer. From 1,920 clones tested, 5 positive clones presented higher optical density (O.D.) in comparison to the others in the presence of 0.016 mM of mercury. These clones were selected for further studies on metal tolerance and degradation. Fosmid DNA sequencing of the positive clones for structural analysis of the operon responsible for metal resistance is currently being undertaken.

Analysis and comparison of metagenomes from six North American forests: Giants on shoulders of giants

Erick Cardenas* (carden24@mail.ubc.ca), Kendra Maas, Steven Hallam, and William W. Mohn

Department of Microbiology and Immunology, Life Sciences Institute, University of British Columbia, Vancouver, Canada

Forest are important ecosystems that play a critical role in the global carbon cycle contributing to almost half of the terrestrial net primary productivity and storing 45% of the terrestrial carbon. In the forest, the soil microbial communities provide many of the critical ecosystem functions including transport of water and nutrients, and nutrient cycling by decomposing woody biomass, making the nutrients trapped in the biomass available for primary productivity. Thus, improved understanding of the soil system is

essential to improve forest productivity and predict the impact of forest harvesting on the microbial communities and the functions they provide.

We used metagenomics to understand the functional and phylogenetic diversity of soil microbial communities under six North American forest sites that are part of the Long Term Soil Productivity experiment. Over a hundred shotgun metagenomes were generated by Illumina sequencing for soil samples which include both organic and mineral soil layers. Additionally, community structure was determined by pyrotag sequencing of the 16S rRNA gene (Bacteria) and the ITS2 region (Fungi) via 454 sequencing. After quality control, we obtained over 940 Gb that ware later assembled and analyzed. Each sample was assembled separately, and also a combined assembly was generated for each site (21 to 24 samples together). Combining samples greatly improved he overall assembly which produced 8.5 Gb of assembled contigs. Assembled contig were annotated using IMG, and unassembled reads were compared against the CAZy database to determine the diversity of enzymes involved in lignocellulose degradation. Partial genomes were binned, and recovered from the assemblies using Metawatt and later analyzed for completeness (using a set of widely conserved genes) and classification (using MEGAN and marker genes).

The microbial communities were consistently dominated by *Alphaproteobacteria* followed by *Actinobacteria* and *Acidobacteria* on the Bacterial side while *Ascomycota* and *Agaromycota* dominated the fungal side. Clustering based on phylogeny was strongly influence by the ecozone and secondarily by the soil layer, while clustering based on functional genes (COGS, and CAZy gene families) showed the opposite trend where the soil layer was the strongest separating factor. Recovered genomes were mostly associated with *Alphaproteobacteria* (*Bradyrhizobium, Rhodopseudomonas*), *Acidobacteria* , and *Actinobacteria* in agreement with our pyrotags results. Few partial fungal genomes were recovered.

This research contributes to the understanding of the soil forest environment as a strongly structured environment and as a potential source enzymes of biotechnological application.

Helitrons, active tools shaping Basidiomycete genomes. The *Pleurotus ostreatus* case.

Castanera R., Pérez G., Santoyo F., Alfaro M., Parenti A., Muguerza E., Meijueiro M.L., Borgognone A., Varas L., Sancho R., Oguiza J., Pisabarro A.G., and Ramírez L.* (lramirez@unavarra.es)

Genetics and Microbiology Research Group, Department of Agrarian Production, Public University of Navarre, Pamplona, Spain

Eukaryotic rolling-circle transposons, also known as helitrons, are a type of class-II transposons recently discovered by computational analyses in plant and animal genomes (1). They bear conserved structural hallmarks such as a 5' - TC terminus; a subterminal palindromic hairpin followed by a 3'- CTRR terminal end, and displays an insertion preference at AT dinucleotides. Putative autonomous elements previously described in plants, animals and fungi encode a protein (RepHel) containing a motif conserved in the replication initiator (Rep) of plasmid rolling-circle replicons, as well as a DNA helicase (Hel) domain. The presence of these domains as well as the absence of target site duplications flanking helitron boundaries suggest that these elements could transpose by a rolling circle mechanism [1]. Previous studies have demonstrated their capacity to capture and mobilize gene fragments and complete genes in plant and animal genomes [1]. We have used a bio-informatics approach for the identification of fungal helitrons based on de novo detection of their structural features combined with homology-based searches of helitron-specific helicases. We have found two helitron families in PC15 and PC9 *P. ostreatus* strains,

accounting for 0.21% and 0.07% of their respective genome sizes. These families (*HELPO1* and *HELPO2*) bear putative autonomous helitrons, non autonomous elements as well as truncated copies. Interestingly, every intact helitron of *HELPO1* family carries at least one captured gene of unknown function downstream the rephel helicase domain in the 5' to 3' direction. RNAseq analyses revealed that both the rephel helicase and the captured genes of *HELPO1* helitrons are actively transcribed in glucose-based media and lignocelluse-based cultures. By contrast, members of the *HELPO2* family didn´t show transcriptional activity. Homology-based searches of helitron-like helicases in other fungi (284 genomes [2]) showed a wider colonization of these elements in basidiomycotina (87% of the genomes) *versus* ascomycotina (30% of the genomes) division. Phylogenetic analyses of plant, animal and fungal helitron-like helicases uncovered a polyphyletic origin of these proteins, supporting the possibility of ancient mobilizations between species belonging to different kingdoms through horizontal transfer.

1. Kapitonov, V.V. and J. Jurka, *Rolling-circle transposons in eukaryotes.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(15): p. 8714-8719.

2. Grigoriev, I.V., et al., *The Genome Portal of the Department of Energy Joint Genome Institute.* Nucleic Acids Research, 2012. **40**(D1): p. D26-D32.

A systems biology characterization of the biotechnological potential stored in the wood-feeding beetle *Odontotaenius disjunctus*

Javier A. Ceja-Navarro* (JCNavarro@lbl.gov),[1] Ulas Karaoz,[1] Mari Nyssoonen,[1] Mary S. Lipton,[2] Joshua N. Adkins,[2] Kristin E. Burnum-Johnson,[2] Markus Bill,[1] Timothy Filley,[3] Mark Conrad,[1] Xavier Mayali,[4] Peter K. Weber,[4] Meredith Blackwell,[5] Jennifer Pett-Ridge,[4] and Eoin Brodie[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Pacific Northwest National Laboratory, Richland, Washington; [3]Purdue University, West Lafayette, Indiana; [4]Lawrence Livermore National Laboratory, Livermore, California; [5]Louisiana State University, Baton Rouge, Louisiana

The passalid beetle (*Odontotaenius disjunctus*) is a subsocial insect that survives on a low-nutrient diet by feeding on large amounts of decaying wood. The morphologically differentiated gut regions of these insects represent a complex of subunits with stratified microbial communities that degrade lignocellulosic materials. Our goal is to characterize the potential stored in the microbiome of the passalid beetle for the optimization of lignocellulosic-dependent energy production processes.

The transformation of lignocellulosic materials by the passalid beetle was first tested measuring the fermentation products of plant polymer decomposition (H_2 and CH_4) with the use of microelectrodes and gas chromatography-isotope ratio mass spectrometry (GC-IRMS). The transformation of lignin after its passage through the gut was determined by ^{13}C-labeled tetramethylammonium hydroxide thermochemolysis. Fosmid libraries were constructed using DNA from different beetle gut regions and screened in high-throughput for lignin, cellulose, and hemicellulose degrading activity. Metagenomic libraries were prepared and sequenced on the Illumina platform and annotated as well as the positive clones from the fosmid libraries. Proteomics and metabolomics were applied to detect expressed proteins and produced metabolites in each gut region.

Lignin side chain oxidation was confirmed by thermochemolysis with acid/aldehyde ratios increasing in the beetle frass. Hydrogen gradients were observed using microelectrodes with concentrations as high as 140 μmol/L in the anterior hindgut (AHG). GC-IRMS analyses of C and H stable isotope fractionation indicated that produced CH_4 was primarily hydrogenotrophic. Fosmid library screening yielded a high

number of clones with activity for the decomposition of cellulose, hemicellulose, and lignin – with the highest potential detected in the AHG. The annotation of metagenomic libraries allowed the identification of the likely contributors to cellulose, xylose and lignin modification. Sequences from hydrogenotrophic methanogenic archaea were more abundant in the anterior hindgut, confirming our previous phylogenetic studies of compartmentalization in the passalid beetle gut. A filtered isolate database and predicted protein sequences from the metagenomes were used to search peptide spectra for proteome reconstruction – preliminary results indicate a variation in the protein expression patterns among the different gut segments potentially indicating a compartmentalization of function.

Our multi-scale approach demonstrates that the passalid beetle harbors and expresses the functional potential to deconstruct lignocellulosic materials and produce H_2, CH_4 and potentially other biofuels. Identifying the microbial contributors to polymer deconstruction and fermentation, and determining their spatial arrangement in the beetle gut will improve our understanding of the ecology of these beetles and inform the design of lignocellulosic fuel production processes.

Functional phylogenomics analysis of bacteria and archaea using consistent genome annotation with UniFam and improved metagenome assembly with Omega

Juanjuan Chai[1], Tae-Hyuk Ahn[1], Bahlul Haider[1], Jillian F. Banfield[2], **Chongle Pan*** (panc@ornl.gov)[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]University of California at Berkeley, Berkeley, California

Cellular functions in microorganisms can be acquired via prevalent horizontal gene transfers. However, if all types of cellular functions can be easily transferred across the entire Bacteria and Archaea domains, this will raise the fundamental question of what defines a microbial species and whether the speciation of a microorganism should be considered as a process of descending from an ancestor (a phylogenetic tree) or combining genes from multiple sources (a phylogenetic network). To shed light on this question, we re-annotated ~12,000 genomes in GenBank using a new protein family database, UniFam, available at http://unifam.omicsbio.org.

UniFam derived rich and uniform annotation from SwissProt (recommended full names, gene name, enzyme classification (EC) number, and gene ontology (GO) terms), obtained high family-wide sequence diversity using TREMBL, and constructed families at the whole protein level. In comparison to the commonly used InterPro annotation pipeline, UniFam provided a comprehensive coverage of protein sequence space to obviate the need to integrate multiple databases and enabled fast annotation of a genome in hours. UniFam is to be updated in sync with the UniProt database to take advantage of the increasing sequence coverage in TREMBL and the improving annotation in SwissProt. The comprehensive genome annotation by UniFam allowed metabolic network reconstruction using MetaCyc.

We represented cellular functions with GO biological processes and MetaCyc pathways and compared cellular functions across 1,191 genera of the GenBank bacterial and archaeal genomes. Cellular functions were characterized based on the consistency between their presence/absence in genomes and the phylogeny and taxonomy of the genomes. Many cellular functions, such as antibiotic resistance, were found to be dispersed across the phylogenetic tree in many distant taxa, probably as a result of horizontal gene transfer. However, many other cellular functions, such as methanogenesis and CO_2

fixation, were found to be consistent with the phylogeny and concentrated in a few taxa. Therefore, we hypothesize that a microbial species can be defined by a set of cellular functions that are generally passed by vertical gene transfer and not by horizontal gene transfer. Microbial species may obtain certain types of modular cellular functions via horizontal gene transfer to adapt to certain environments.

A more accurate functional phylogenomics analysis requires more comprehensive genome coverage of the phylogenetic tree, particularly the novel taxa that cannot be readily cultured in laboratory for isolate genome sequencing. To address this challenge, we are developing strategies to recover complete or near-complete microbial genomes directly from metagenomic sequencing data. We developed a new overlap-graph metagenome assembler (Omega), available at http://omega.omicsbio.org. Omega was optimized for long Illumina reads (overlapping paired-end 300-bp MiSeq or 250-bp HiSeq sequencing). The progress in sequencing technologies and assembly algorithms will soon allow reconstruction of high-quality genomes from many novel taxa.

The KBase narrative user interface

Stephen Chan* (sychan@lbl.gov),[1] Jason Baumohl,[1] Aaron Best,[2] Jared Bischof,[2] Ben Bowen,[1] Tom Brettin,[2] Tom Brown,[2] Shane Canon,[1] John-Marc Chandonia,[1] Dylan Chivian,[1] Ric Colasanti,[2] Neal Conrad,[2] Brian Davison,[3] Matt DeJongh,[6] Paramvir Dehal*,[1] Narayan Desai,[2] Scott Devoid,[2] Terry Disz,[2] Meghan Drake,[3] Janaka Edirisinghe,[2] Gang Fang,[7] José Pedro Lopes Faria,[2] Mark Gerstein,[7] Elizabeth M. Glass,[2] Annette Greiner,[1] Dan Gunter,[1] James Gurtowski,[5] Nomi Harris,[1] Travis Harrison,[2] Fei He,[4] Matt Henderson,[1] Chris Henry,[2] Adina Howe,[2] Marcin Joachimiak,[1] Kevin Keegan,[2] Keith Keller,[1] Guruprasad Kora,[3] Sunita Kumari,[5] Miriam Land,[3] Folker Meyer,[2] Steve Moulton,[3] Pavel Novichkov,[1] Taeyun Oh,[8] Gary Olsen,[9] Bob Olson,[2] Dan Olson,[2] Ross Overbeek,[2] Tobias Paczian,[2] Bruce Parrello,[2] Shiran Pasternak,[5] Sarah Poon,[1] Gavin Price,[1] Srivdya Ramakrishnan,[5] Priya Ranjan,[3] Bill Riehl,[1] Pamela Ronald,[8] Michael Schatz,[5] Lynn Schriml,[10] Sam Seaver,[2] Michael W. Sneddon,[1] Roman Sutormin,[1] Mustafa Syed,[3] James Thomason,[5] Nathan Tintle,[6] Will Trimble,[2] Daifeng Wang,[7] Doreen Ware,[5] David Weston,[3] Andreas Wilke,[2] Fangfang Xia,[2] Shinjae Yoo,[4] Dantong Yu,[4] Robert Cottingham,[3] Sergei Maslov,[4] Rick Stevens,[2] and Adam P. Arkin[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Argonne National Laboratory, Argonne, Illinois; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Brookhaven National Laboratory, Upton, New York; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [6]Hope College, Holland, Michigan; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois;[10]University of Maryland, College Park, Maryland

Currently, there are several ways for researchers familiar with command-line computing environments to use KBase's analysis tools and datasets. To serve the broader biological community, KBase is developing a new graphical user interface called the Narrative interface.

The Narrative Interface, which is now operational as an alpha release (http://demo.kbase.us/), provides a novel, more user-friendly way to access KBase's tools and datasets to tackle problems in systems biology, while recording every step of the workflow and thereby capturing a researcher's thought process. Thedatasets, analyses, and thought processesunderlying the execution of these actionsare captured in the form of an interactive,dynamic publication called a Narrative.Within each Narrative, users can interleavetext and commentary with workflows, sothat hypotheses and conclusions can be captured as well as raw results andprocedural notes. The Narrative, with thehelp of Workspaces, maintains provenance and metadata for all data, thereby providing a virtual reference list for all

Narratives. A sophisticated social framework will soon allow members of a research team to share, execute, modify, and comment on Narratives at multiple levels of granularity enabled by KBase infrastructure.

Intrinsically reproducible, Narratives will serve as a new type of publication by (1) explicitly capturing the parameters associated with various algorithms, (2) clearly showing how users applied the algorithms to selected datasets, and (3) transparently recording the process by which resulting output was used to derive biological insight and conclusions. Users who access a Narrative that another researcher has created and shared not only will be able to see a complete inventory of the data and algorithms underlying a conclusion, they will be able to repeat the computational experiment with the push of a button, even altering parameters to achieve different or improved results.

RNAseq-enabled expression quantitative trait locus (eQTL) mapping in *Populus*

Jay Chen* (chenj@ornl.gov), Wellington Muchero, Xiaohan Yang, Priya Ranjan, and David Weston

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee

Fuels developed from lignocellulosic biomass offer a potential renewable and clean alternative to conventional fossil-fuel-based energy sources. *Populus* is one of DOE's "flagship" plant species that is of special interest as a biofuel feedstock. Hybrids among different *Populus* species that specifically exploit heterosis for enhanced productivity are widely used for biomass production. However, there are still significant biological and technological barriers that need to be overcome in order to achieve cost-effective, sustainable production and conversion systems for *Populus* biomass into biofuels. In particular, plant cell wall recalcitrance poses tremendous challenges for *Populus* feedstock to become economically competitive. Although a wealth of research has been directed towards modifying plant cell walls by targeting specific genes involved in cell wall biosynthesis, it has become clear that the complex biomass quantity and quality traits require the regulation and coordinated interactions of many genes. Identifying such genes and genetic networks remains unaccomplished. This JGI Community Science Program project aims to perform gene expression Quantitative Trait Locus (eQTL) analysis based on RNAseq data from 400 biological samples representing 200 progeny derived from a *Populus trichocarpa* × *Populus deltoides* pseudo-backcross pedigree. These individuals have been intensively characterized for phenotypic traits related to cell wall chemistry, biomass productivity and sustainable biomass production. These phenotypic data are complemented by fully resequenced parental genotypes of the pedigree and an ultra-dense genetic map with 3,568 single nucleotide polymorphism markers with an average marker distance of less than 0.75 cM. With the cumulative resources, we have identified large effect QTLs related to cell wall chemistry and biomass productivity. However, molecular characterization and functional validation of these QTLs have proven to be the major limiting factor. The resource generated in this JGI CSP project will enable an unprecedented RNAseq-based eQTL analysis in *Populus*. The resulting data will facilitate the rapid characterization of high-value QTLs already mapped in the same pedigree and the identification of both genetic loci and genetic networks contributing to complex phenotypic traits of biomass quality, overall productivity and cell wall chemistry. In the short term, this project will provide novel knowledge to inform genetic improvement of *Populus* for cost-effective, sustainable biomass production, a topic that has become increasingly important in research areas highly relevant to DOE's missions.

String graph assembly for diploid genomes with long reads

Jason Chin*(jchin@pacificbiosciences.com), David Rank, Joseph Karalius, and Paul Peluso

Pacific Biosciences, Menlo Park, California

We use PacBio® data from whole-genome shotgun sequencing of multiple strains of *Arabidopsis thaliana* to evaluate the feasibility of using a string graph to assemble diploid heterozygous genomes. We first demonstrate the generation of high-quality assemblies (N50 ~ 4 Mbp) for inbred strains with a string graph assembly method. Then, we perform a few computational experiments to understand the properties of the string graph for polymorphic or heterozygous diploid genomes. Distinct diploid structures in the string graph constructed with long reads are observed and a simple layout algorithm is developed to resolve homologous copies from genomic repeats. This allows constructing long primary contigs from a diploid heterozygous sample. The structural variations between the haplotypes can be captured by constructing "associated contigs," which represent alternative paths other than those of the primary contigs in the string graph. A full set of structure variations (SVs) can be then categorized. These SVs and the primary contigs can be used as the foundation for the downstream process of phasing SVs and SNPs together for full reconstruction of the haplotype sequences. We successfully applied this strategy to generate heterozygous diploid assemblies (N50 > 2 Mbp) and evaluate the correctness from a data set of mixed inbredArabidopsis sequencing data. Finally, we propose a general bioinformatics workflow for diploid assembly with long sequence reads.

MetaMaize: Aboveground maize-endophyte associations as influenced by host genotype, climatic region, seed source, and tissue type

Alice C.L. Churchill* (acc7@cornell.edu),[1] Santiago X. Mideros,[1,5] Peter Balint-Kurti,[2] Surya Saha,[1,3] Jenny B. Cornell,[4] and Rebecca J. Nelson[1,4]

[1]Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, New York; [2]USDA-ARS Plant Science Research Institute, Department of Plant Pathology, North Carolina State University, Raleigh, North Carolina; [3]Boyce Thompson Institute for Plant Research, Ithaca, New York; [4]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York; [5]Current address: BASF Plant Science, Research Triangle Park, NC 27709

Several diseases affect maize production globally and regionally. The maize ear rot and stalk rot pathogens are particularly important for their production of mycotoxins that adversely affect human and livestock health, biomass yield, and quality of grain and biofuel feedstock. The MetaMaize project is aimed at characterizing endophytic microbiota from aboveground tissues of a set of diverse, genetically well-characterized maize lineages. We are using comprehensive microbiota survey approaches as a first step to identify unique maize-associated endophytes that potentially can improve plant growth traits and resistance to mycotoxin-producing fungi. Field-grown inbred lines are the source of tissues for characterizing culturable and unculturable fungal and bacterial/archaeal endophyte communities. For each of five lines, we planted seed sourced from two geographically distinct locations during the 2012 field season at two climatically distinct sites in the U.S., each with different disease and pest pressures - NY with a temperate climate and NC with a subtropical climate. Kernels harvested from the 2012 sites were the source of seeds planted in 2013 at the same locations. Asymptomatic tissues assayed over the two field seasons were from kernels of the same source as those planted, adventitious roots, stems, leaves, immature ears, and kernels produced after self-pollination. A total of 540 maize DNA samples

will be submitted to JGI for PCR amplicon production and barcoding of fungal ITS and bacterial 16S rRNA sequences. Maize DNA isolation for MiSeq sequencing by JGI is in progress. Twelve maize-associated endophytic bacteria and six fungi will be targeted for genome sequencing, based on novelty of taxonomy and/or biological activity. To date, three temperate and two tropical maize lines show significant differences in diversity and number of culturable endophytes from asymptomatic tissues. Generally, tropical lineages contain a greater diversity than temperate lines. Different tissue types generally harbor phenotypically distinctly communities, suggesting tissue as a determinant of endophyte habitat. Leaves are rich in fungi (often heavily pigmented) and bacteria; immature ears are deficient in fungi, with bacteria isolated from only a subset of lines; more lightly pigmented fungi are prevalent in kernels, especially those from NC, with bacteria rarely cultured from kernels to date. Identification of cultured endophytes is in progress. The MetaMaize project will allow us to begin to evaluate genetic and environmental factors effecting the composition and complexity of endophyte communities in aboveground tissues of diverse maize lines. It is the basis for future studies to understand whether beneficial endophytic microbes are maize generalists or genotype-specific, if they are heritable, and, if so, which maize loci contribute to endophyte heritability.

Single cell genomics and transcriptomics for unicellular Eukaryotes

Doina Ciobanu*(dgciobanu@lbl.gov),[1] Alicia Clum,[1] Vasanth Singan,[1] Asaf Salamov,[1] James Han,[1] Alex Copeland,[1] Igor Grigoriev,[1] Timothy James,[2] Steven Singer,[3] Tanja Woyke,[1] Rex Malmstrom,[1] and Jan-Fang Cheng[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]University of Michigan, Ann Arbor, Michigan; [3]Joint BioEnergy Institute, Emeryville, California

Despite their small size, unicellular eukaryotes have complex genomes with a high degree of plasticity that allow them to adapt quickly to environmental changes. Unicellular eukaryotes live with prokaryotes and higher eukaryotes, frequently in symbiotic or parasitic niches. To this day their contribution to the dynamics of the environmental communities remains to be understood. Unfortunately, the vast majority of eukaryotic microorganisms are either uncultured or unculturable, making genome sequencing impossible using traditional approaches. We have developed an approach to isolate unicellular eukaryotes of interest from environmental samples, and to sequence and analyze their genomes and transcriptomes. We have tested our methods with six species: an uncharacterized protist from cellulose-enriched compost identified as *Platyophrya*, a close relative of *P. vorax*; the fungus *Metschnikowia bicuspidate,* a parasite of water flea *Daphnia*; the mycoparasitic fungi *Piptocephalis cylindrospora,* a parasite of *Cokeromyces* and *Mucor*; *Caulochytrium protosteloides*, a parasite of *Sordaria; Rozella allomycis*, a parasite of the water mold *Allomyces*; and the microalgae *Chlamydomonas reinhardtii*. Here, we present the four components of our approach: pre-sequencing methods, sequence analysis for single cell genome assembly, sequence analysis of single cell transcriptomes, and genome annotation. This technology has the potential to uncover the complexity of single cell eukaryotes and their role in the environmental samples.

Exploring *Arabidopsis thaliana* endophytes via single-cell genomics

Scott Clingenpeel* (srclingenpeel@lbl.gov)[1], Derek Lundberg[2], Tanja Woyke[1], Susannah Tringe[1], and Jeff Dangl[2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Land plants grow in association with microbial communities both on their surfaces and inside the plant (endophytes). The relationships between microbes and their host can vary from pathogenic to mutualistic. Colonization of the endophyte compartment occurs in the presence of a sophisticated plant immune system, suggesting finely tuned discrimination of pathogens from mutualists and commensals. Despite the importance of the microbiome to the plant, relatively little is known about the specific interactions between plants and microbes, especially in the case of endophytes. The vast majority of microbes have not been grown in the lab, and thus one of the few ways of studying them is by examining their DNA. Although metagenomics is a powerful tool for examining microbial communities, its application to endophyte samples is technically difficult due to the presence of large amounts of host plant DNA in the sample. One method to address these difficulties is single-cell genomics where a single microbial cell is isolated from a sample, lysed, and its genome amplified by multiple displacement amplification (MDA) to produce enough DNA for genome sequencing. We have applied this technology to study the endophytic microbes in *Arabidopsis thaliana* roots. Extensive 16S gene profiling of the microbial communities in the roots of multiple inbred *A. thaliana* strains has identified 164 OTUs as being significantly enriched in all the root endophyte samples compared to their presence in bulk soil. Approximately 14,000 single microbial cells were isolated from these samples and ~500 of these were identified as being members of the enriched OTUs. The genomes of 131 of these single cells, representing 29 of the target OTUs, have been sequenced. Due to amplification bias, the genomes obtained from single-cell technology are incomplete. The genomes recovered for this project have an estimated completeness of 2 – 99%. Since the amplification bias occurs in random locations, combining multiple single-cell genomes from the same OTU can produce more complete genomes. Thus, 99 single-cell genomes were coassembled into 22 combined assemblies. The single-cell genomes obtained from this project along with genomes of cultured isolates from the same samples are being compared to those from close relatives that are not plant-associated to identify genes that are likely involved in the plant-microbe interaction.

Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways

R. Cameron Coates* (rccoates@ucsd.edu),[1] Sheila Podell,[1] Anton Korobeynikov,[3,4] Alla Lapidus,[3,5] Pavel Pevzner,[3,6] David H. Sherman,[7] Eric E. Allen,[1] Lena Gerwick,[1] and William H. Gerwick[2]

[1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California; [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, California; [3]Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg, Russia; [4]Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia; [5]Theodosius Dobzhansky Center for Genome Bionformatics, St. Petersburg State University, St. Petersburg, Russia; [6]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, California; [7]Life Sciences Institute and Department of Medical Chemistry, University of Michigan, Ann Arbor, Michigan

Cyanobacteria possess the unique capacity to naturally produce hydrocarbons from fatty acids. Hydrocarbon compositions of thirty-two strains of cyanobacteria were characterized to reveal novel structural features and insights into hydrocarbon biosynthesis in cyanobacteria. This investigation revealed new double bond (2- and 3-heptadecene) and methyl group positions (3-, 4- and 5-methylheptadecane) for a variety of strains. Additionally, results from this study and literature reports indicate that hydrocarbon production is a universal phenomenon in cyanobacteria. All cyanobacteria possess the capacity to produce hydrocarbons from fatty acids yet not all accomplish this through the same metabolic pathway. One pathway comprises a two-step conversion of fatty acids first to fatty aldehydes and then alkanes that involves a fatty acyl ACP reductase (FAAR) and aldehyde deformylating oxygenase (ADO). The second involves a polyketide synthase (PKS) pathway that first elongates the acyl chain followed by decarboxylation to produce a terminal alkene (olefin synthase, OLS). Sixty-one strains possessing the FAAR/ADO pathway and twelve strains possessing the OLS pathway were newly identified through bioinformatic analyses. Strains possessing the OLS pathway formed a cohesive phylogenetic clade with the exception of three *Moorea* strains and *Leptolyngbya sp.* PCC 6406 which may have acquired the OLS pathway via horizontal gene transfer. Hydrocarbon pathways were identified in one-hundred-forty-two strains of cyanobacteria over a broad phylogenetic range and there were no instances where both the FAAR/ADO and the OLS pathways were found together in the same genome, suggesting an unknown selective pressure maintains one or the other pathway, but not both.

Connecting organismal physiology to community structure in assembling phototrophic communities

Jessica K. Cole, Janine R. Hutchison, William C. Nelson, Margie F. Romine, Ryan S. Renslow, William B. Chrisler, L. Meng Markillie, Ronald C. Taylor, H. Steven Wiley, James K. Fredrickson, and **Stephen R. Lindemann*** (stephen.lindemann@pnnl.gov)

Pacific Northwest National Laboratory, Richland, Washington

PNNL Foundational Scientific Focus Area

Microbial mats are compact ecosystems composed of highly interactive organisms in which complete energy and element cycles occur. Because these communities frequently experience significant environmental variation over both short (e.g., diel) and long (e.g., seasonal) time scales, they must be resistant to environmental change and resilient after perturbation. We hypothesize that these higher-order community properties emerge in microbial communities in predictable ways and arise from the interspecies interactions operating between members.

Hot Lake is a hypersaline lake that seasonally harbors a phototrophic mat that assembles each spring and disassembles each fall. The mat is subject to greater than tenfold variation in salinity (primarily Mg^{2+} and SO_4^{2-}) and irradiation over the annual cycle. Despite this extreme environmental variation, the composition of the mat community exhibited relative stability in its key autotrophic and heterotrophic membership throughout 2011. By mid-autumn the mat began to disassemble, with mat community richness and evenness declining precipitously under conditions of increasing salinity and decreasing irradiance and temperature. However, few OTUs displayed a pattern of variation that correlated with salt concentrations; changes in mat community structure were driven more strongly by variation in light availability (energy input) than by salinity.

Elucidation of the principles governing interspecies interactions in natural systems is challenging because of their diversity, complexity and intractability to *in situ* manipulation. Therefore, we isolated four filamentous cyanobacteria and their co-isolated heterotrophic consorts from the Hot Lake mat, in

two of which studies of primary succession were conducted. Both unicyanobacterial consortia retained essentially the same suite of ~20 heterotrophic species after sequential passage, representing ten of the forty most-abundant OTUs within the mat community. The abundances of heterotrophic species exhibited substantial differences depending upon which cyanobacterium served as primary producer. Gammaproteobacterial members were initially abundant in consortium UCC-O but dropped precipitously over the month-long growth period. The most abundant heterotroph at the beginning of this period, *Idiomarinaceae* sp. HL-53, was succeeded by two species of *Bacteroidetes* and one *Rhizobiales* sp. as dominant heterotrophs. Transcriptional analysis of the UCC-O consortium revealed significant changes in the gene regulatory patterns of HL-53 compared with other consortium members, which were suggestive of nutrient stress midway through the growth period and transitions in flagellum and transporter expression.

Future genome-enabled studies will involve an iterative field-lab approach in which observations of the mat exposed to natural variation in the field are paired with laboratory-based environmental perturbation studies of consortia. These studies will employ integrated global molecular techniques (e.g., transcriptomics, proteomics, and metabolomics) to define the niches occupied by members and detect interactions between members that may contribute to a community's higher-order properties.

Amplicon, genomic and metagenomic sequencing were provided for this study by JGI under CSP 701.

Creating a multi-functional library of grass transcription factors for the energy crop model system *Brachypodium distachyon*

Joshua H. Coomey* (jcoomey@cns.umass.edu)[1], John Gierer[2], Samuel Deutsch[3], Steven A. Kay[4], **Todd C. Mockler**[2], and **Samuel P. Hazen**[1]

[1]University of Massachusetts Amherst; [2]Donald Danforth Plant Science Center, St. Louis; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]University of Southern California, Los Angeles, California

Project Goals: The overall goal of this project is to develop BradiTORFL, a comprehensive *Brachypodium distachyon* Transcription Factor ORF Library.

Comprehensive collections of full-length transcription factor cDNAs (fl-cDNA) have proven to be an extraordinary reagent for advanced research systems such as human, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*. One of the seven DOE-JGI Flagship Plant Genome species, *Brachypodium distachyon*, is a model for potential energy crops such as switchgrass, sorghum, and *Miscanthus*, as well as for the cereal crops that constitute a large part of the world's diet. We are constructing a complete grass transcription factor collection in an entry vector that will be of value for numerous functional studies. To accelerate reagent development we have included gene synthesis as a technique to capture fl-cDNAs. From expression profile data we identified high priority candidates for the regulation of biofuel feedstock traits, such as growth, cell wall biogenesis, and abiotic stress tolerance. The DOE-JGI synthesized 143 unique transcription factors from families that include bHLH, bZIP, CCAAT, GRAS, Homeodomain, HSF, MADS box, MYB, NAM, WRKY, and several classes of zinc fingers. Specific subfamilies include putative Aux/IAA-ARF auxin response factors, bHLH factors and G-box binding proteins implicated in light signaling, CBF-like genes with predicted abiotic stress response function, ethylene-associated factors, MYBs and CCT domain genes predicted to function in cell wall, circadian clock, and light signaling. Including these candidates in BradiTORFL will significantly enhance our ability to build regulatory models of gene expression and function. The collection will be transferred into multiple destination vectors for downstream applications including protein-DNA and protein-protein interaction platforms in yeast. Genes that yield positive interactions can then be shuttled from

their pENTR vector to a variety of other constructs for continued study, including expression *in planta* to further characterize their functions. We are presently evaluating *B. distachyon* protein-DNA interactions in yeast using two approaches. The first is "gene-centric" where a promoter is tested for interactions with all transcription factor proteins in the library. The second approach, developed by the Mockler Lab, is a "protein-centric" interrogation of each transcription factor's capacity to interact with a collection of 768 synthetic 250 bp promoters. These synthetic promoters maximize potential binding motif sequence diversity with all possible 8 nt DNA motifs occurring in at least 4 independent promoters. Proof-of-concept experiments demonstrate the utility of this approach and we are currently expanding the analyses to infer the binding specificities of all the transcription factors synthesized by DOE-JGI in this project.

This work was supported by the Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy grant DE-SC0006621 and the Joint Genome Institute Community Sequencing Program grants CSP-667 and CSP-1431.

Molecular dissection of amylolysis in *Fusarium verticillioides*

John C. Crandon*, John R. Ridenour, and Burt H. Bluhm (bbluhm@uark.edu)

Department of Plant Pathology, University of Arkansas, Division of Agriculture, Fayetteville, Arkansas

Maize is a globally important crop that serves as a food staple throughout the world. Fusarium ear rot, caused by *Fusarium verticillioides*, is a devastating maize disease due to the accumulation of fumonisins within infected kernels. Fumonisins are polyketide-derived mycotoxins associated with various mycotoxicoses of humans and livestock. Although previous work has established a link between starch metabolism and fumonisin biosynthesis in *F. verticillioides*, little is known about the mechanisms underlying starch sensing or the genetic regulation of amylolysis within infected maize kernels. A combination of forward and reverse genetic approaches is being applied to dissect the molecular mechanisms of starch sensing and metabolism in *F. verticillioides*. The forward genetic component utilizes random insertional mutagenesis to identify novel genes involved in amylolysis. Two separate forward genetic screening methods have been utilized to assess virulence in insertional mutants: a live ear assay has identified mutants impaired in kernel colonization, while a cracked kernel assay is being utilized to quantify amylolysis and fumonisin biosynthesis. A preliminary live ear screen of 490 insertional mutants identified 36 strains displaying reduced kernel colonization. The reverse genetic component consists of targeted deletion of six candidate regulatory genes to dissect signal transduction pathways involved in amylolysis. Targeted deletion mutants are being evaluated for impaired starch hydrolysis, abnormalities in growth and development, and impaired fumonisin biosynthesis. The combination of forward and reverse genetic approaches will provide new insight into amylolysis and kernel colonization in *F. verticillioides*. An improved understanding of these processes will lead to a new working model explaining kernel colonization and the genetic linkage between starch metabolism and fumonisin biosynthesis.

Defining the functional diversity of the *Populus* root microbiome

W.N. Cude,[1] T-Y.S. Lu,[1] M.S. Robeson II,[1] Z.K. Yang,[1] C. Timm,[1] D.J. Weston,[1] S.L. Allman,[1] M. Podar,[1] C.W. Schadt,[1] C.S. Harwood,[2] A.L. Schaefer,[2] Y. Oda,[2] J. Morrell-Falvey,[1] A. Bible,[1] D.W. Ussery,[1] G.A. Tuskan,[1] M.J. Doktycz,[1] and **D.A. Pelletier*** (pelletierda@ornl.gov)[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Department of Microbiology, University of Washington, Seattle, Washington

The beneficial association between plants and microbes exemplifies a complex system shaped by the participating organisms and environmental forces acting upon it. *Populus* are leading biomass woody species that host a variety of microorganisms within their endosphere and rhizosphere that influence host phenotypes. Therefore, *Populus* and its associated microbial community are being studied as part of our Plant-Microbe Interfaces project (http://pmi.ornl.gov).

Our goal is to understand the phylogenetic and functional diversity within the *Populus* microbiome and to elucidate metabolic and molecular mechanisms responsible for shaping the *Populus*-microbial interface. To detangle this complex ecosystem, we have applied cultivation dependent and cultivation independent techniques to capture and characterize the *Populus* root microbiome. We have isolated and begun to characterize approximately 3000 bacterial strains belonging to >427 OTUs from *Populus* rhizosphere and endosphere compartments. We have successfully sequenced the genomes of 43 of these strains. With this JGI-CSP project we are currently sequencing the genomes of an additional 400 phylogenetically and functionally diverse isolates. We are concurrently utilizing select *Populus*-derived isolates chosen based on phenotypic screens and genomic inventory data to dissect the signaling pathways (specifically cyclic-di-GMP and quorum sensing) involved in plant-microbe interactions using directed analytical approaches in laboratory based experiments. This is being completed in order to determine molecular mechanisms of host signaling and colonization.

Additionally, metagenomic sequencing of native root endosphere communities will provide insight into functions enriched in this compartment. Microbial metagenomes have been difficult to obtain from *Populus* root tissues due to low microbial biomass relative to plant biomass, metagenomes, have historically been difficult to obtain from *Populus* root tissues. We are developing a method of enriching endophytes from *Populus* root homogenates for isolation and metagenomic investigation to address these issues. Endophytic bacterial communities were enriched using differential and density gradient centrifugation. Total DNA was extracted from enriched and unenriched samples and the endophytic bacterial community composition was determined by 16S rDNA sequencing. The enrichment protocol reduced the number of contaminating chloroplast DNA reads by approximately 10X while having only minor effects on bacterial community diversity and abundance. Additionally we applied flow-cytometry to bacterial endosphere-enriched samples for isolation of single cells for subsequent genomics and potential growth of as yet uncultivated *Populus* associates. The ability to perform single-cell multi-omic analyses will allow for in-depth characterization of un-cultured or rare endophytic bacteria residing within the natural root system of *Populus.* This joint endeavor with the JGI will provide additional genomic insights into the ways in which these beneficial plant-microbe interactions might occur.

ETOP Project: Development and implementation of high throughput methods for fungal culturing and nucleic acid isolation

David Culley[1], Beth Hofstad[1], Mark Butcher[1], James Collett[1], Igor Grigoriev[2], and **Jon Magnuson*** (Jon.Magnuson@pnnl.gov)[1]

[1]Pacific Northwest National Laboratory, Richland, Washington; [2]DOE Joint Genome Institute, Walnut Creek, Califorenia

As part of the Emerging Technologies Opportunity Program (ETOP; www.jgi.doe.gov/programs/ETOP/) at the Joint Genome Institute, we are developing high throughput methods for culturing fungi and isolating quality DNA and RNA. This ETOP project will also support the goals of the 1000 Fungal Genomes Project (genome.jgi.doe.gov/programs/fungi/1000fungalgenomes.jsf) by providing DNA and RNA for sequencing the diverse fungi represented in this whole Kingdom sequencing project. The ETOP project has three aspects. 1. Method development for high throughput fungal culturing: culturing filamentous fungi in a flexible and efficient microtiter plate format to rapidly obtain phenotypic growth data, optimize culture conditions for gDNA isolation, and provide information rich samples from a wide variety of culture conditions for mRNA sequencing to support genome annotation. 2. Method development for high throughput nucleic acid isolation: lysing fungi and extracting DNA and RNA of high quality and quantity suitable for downstream sequencing and analysis at JGI. 3. Fungal nucleic acid delivery: delivering DNA and RNA and associated growth data from 145 fungi to JGI. Progress from the first half-year of this project has mainly been in the area of fungal culturing and growth monitoring. This has been done in 24 well standard footprint plates using the resorufrin/resazurin growth reporting reagent on an array of media types representing different carbon and nitrogen sources. Initial extractions of RNA have shown it to be of high quality. Future directions will be in the area of higher quality DNA and higher efficiency in streamlining the process, but the method development aspect of the project is quickly maturing. In the remainder of this two year project, DNA/RNA isolation will become the dominant focus and we will likely experience new challenges in growing very diverse fungi from different phyla and isolating the nucleic acids for sequencing.

Analysis of microbes

Paramvir Dehal* (pdehal@lbl.gov),[1] Chris Henry,[2] Jason Baumohl,[1] Aaron Best,[1] Jared Bischof,[2] Ben Bowen,[1] Tom Brettin,[2] Tom Brown,[2] Shane Canon,[1] Stephen Chan,[1] John-Marc Chandonia,[1] Dylan Chivian,[1] Ric Colasanti,[2] Neal Conrad,[2] Brian Davison,[3] Matt DeJongh,[6] Paramvir Dehal,[1] Narayan Desai,[2] Scott Devoid,[2] Terry Disz,[2] Meghan Drake,[3] Janaka Edirisinghe,[2] Gang Fang,[7] José Pedro Lopes Faria,[2] Mark Gerstein,[7] Elizabeth M. Glass,[2] Annette Greiner,[1] Dan Gunter,[1] James Gurtowski,[5] Nomi Harris,[1] Travis Harrison,[2] Fei He,[4] Matt Henderson,[1] Chris Henry,[2] Adina Howe,[2] Marcin Joachimiak,[1] Kevin Keegan,[2] Keith Keller,[1] Guruprasad Kora,[3] Sunita Kumari,[5] Miriam Land,[3] Folker Meyer,[2] Steve Moulton,[3] Pavel Novichkov,[1] Taeyun Oh,[8] Gary Olsen,[9] Bob Olson,[2] Dan Olson,[2] Ross Overbeek,[2] Tobias Paczian,[2] Bruce Parrello,[2] Shiran Pasternak,[5] Sarah Poon,[1] Gavin Price,[1] Srivdya Ramakrishnan,[5] Priya Ranjan,[3] Bill Riehl,[1] Pamela Ronald,[8] Michael Schatz,[5] Lynn Schriml,[10] Sam Seaver,[2] Michael W. Sneddon,[1] Roman Sutormin,[1] Mustafa Syed,[3] James Thomason,[5] Nathan Tintle,[6] Will Trimble,[2] Daifeng Wang,[7] Doreen Ware,[5] David Weston,[3] Andreas Wilke,[2] Fangfang Xia,[2] Shinjae Yoo,[4] Dantong Yu,[4] Robert Cottingham,[3] Sergei Maslov,[4] Rick Stevens,[2] and Adam P. Arkin[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]Argonne National Laboratory, Argonne, Illinois; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Brookhaven National Laboratory,

Upton, New York; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [6]Hope College, Holland, Michigan; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

The Microbial Sciences component of the KBase project has three overall goals: 1) to enable the generation of predictive models for metabolism and gene regulation to facilitate the manipulation of microbial function; 2) to vastly increase the capability of the scientific community to communicate and utilize existing data; and 3) to enable the planning of effective experiments and to maximize our understanding of microbial system functions. To achieve these goals we have focused on unifying existing 'omics datasets and modeling toolsets within a single integrated framework that will enable users to move seamlessly from the genome assembly and annotation process through to a reconciled metabolic and regulatory model that is linked to all existing experimental data for a particular organism. The results are hypotheses for such things as gene-function matching and the use of comparative functional genomics to perform higher quality evidence-based annotations. KBase offers tools for reconciling the models against experimental growth phenotype data, and using them to predict phenotypes in novel environments or under genetic perturbations.

To prioritize the development of the microbial science area and enable new science, we are focusing on developing prototype analysis workflows that will be most useful to microbial scientists. To date we have developed KBase capabilities and demonstrations workflows for: (1) genome annotation and metabolic reconstruction, (2) regulon reconstruction, (3) metabolic and regulatory model reconstruction, and (4) reconciliation with experimental phenotype and expression data.

Phylogeny and physiological potential of the candidate phylum "Atribacteria" inferred from single-cell genomes and binned metagenomes

Jeremy A. Dodsworth* (jeremy.dodsworth@unlv.edu),[1] Nobu Masaru,[2] Senthil K. Murugapiran,[1] Christian Rinke,[3] Patrick Schwientek,[3] Esther A. Gies,[4] Gordon Webster,[5] Peter Kille,[5] Andrew Weightman,[5] Wen-Tso Liu,[2] Steven J. Hallam,[4] George Tsiamis,[6] Tanja Woyke,[3] and Brian P. Hedlund[1]

[1]University of Nevada, Las Vegas, Nevada; [2]University of Illinois, Champaign-Urbana, Illinois; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]University of British Colombia, Vancouver, BC, Canada; [5]Cardiff University, Cardiff, United Kingdom; [6]University of Patras, Agrinio, Greece

The "*Atribacteria*" is a candidate phylum in the Bacteria recently proposed to include members of the JS1 and OP9 lineages. JS1 and OP9 are globally distributed, and in some cases abundant, in anaerobic marine waters and sediments as well as geothermal environments, petroleum reservoirs, and anaerobic digestors/reactors. However, the monophyly of JS1 and OP9 has been questioned and their physiology and ecology remain enigmatic due to a lack of cultivated representatives. Here we use recently published and newly obtained OP9 and JS1 single-cell genomes to define bins corresponding to these lineages in metagenomes from several environments, including biofilm from a terephthalate-degrading reactor and anaerobic, sulfidic waters of Sakinaw Lake. Multiple binning techniques were employed and the resulting bins were compared to assess the robustness of each technique. The single-cell genomes and metagenome bins represent two major lineages within OP9 and at least four lineages within JS1. Phylogenomic analyses of these combined datasets confirm the monophyly of the "*Atribacteria*" inclusive of OP9 and JS1. Based on comparison of metabolic potential inferred from these datasets,

members of the "*Atribacteria*" appear to be heterotrophic anaerobes that lack respiratory capacity. Identification and analysis of conserved genes, including a bacterial microcompartments gene cluster, suggest potential roles in catabolism of carbohydrates that are common to members of the "*Atribacteria*".

Deciphering the lexicon of *cis*-regulatory elements in the spatiotemporal control of gene expression in plants

Shahram Emami* (shahrame@stanford.edu), Rui Wu, Muh-ching Yee, and José R. Dinneny

Carnegie Institution for Science, Department of Plant Biology, Stanford, California

We seek to gain a systematic understanding of the role that *cis*-regulatory elements (CREs) play in orchestrating the regulation of spatiotemporal gene expression in the *Arabidopsis* root tissue. We have assembled a pipeline for discovery and validation of CREs and their corresponding network of transcription factors (TFs) using bioinformatics, synthetic biology and genomics tools and the existing high-resolution root expression data. Based on the data gathered, a computational predictive model of the effects of the validated CREs on the gene expression will be generated, which will: 1) advance our understanding of the mechanisms controlling gene expression in multicellular organisms 2) be used to generate synthetic promoters inducing gene expression to suit a particular application.

Multi-dimensional study of microbial community behavior using nanoDESI mass spectrometry, SIMS and fluorescence microscopy

Yigang Fang* (yigangfang@lbl.gov),[1] Manfred Auer,[2] James Berleman,[2] Marcin Zemla,[2] Megan Danielewicz,[2] Trent Northen,[2] and Musahid Ahmed[1]

[1]Chemical Sciences Division, [2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Territorial disputes between bacterial species are common in most habitats, and are of great interest to a large number of energy-relevant topics such as bioremediation, carbon cycling and lignocellulose degradation. Here we study model microbial community structure and metabolism (*Myxococcus xanthus* and *Bacillus subtilis*) to investigate how and why a community responds to an external threat in a localized, regional or global manner. A high throughput workflow is being developed, by which multi-dimensional information can be acquired from a single sample. This includes Nanospray Desorption Electrospray Ionization (nanoDESI) mass spectrometry, secondary ion mass spectrometry (SIMS), and fluorescence microscopy. Data obtained from different techniques will be pipelined to the super computer center for computational analysis and mathematical modeling.

We have developed the novel techniques of correlative ambient environment nanoDESI ion and fluorescence microscopy imaging. These are automated using LabVIEW programs dedicated written for this purpose. The sample is examined under a fluorescence microscope to acquire optical and fluorescence images. Then it is transferred to a XYZ linear stage to obtain nanoDESI/MS images. Acquired data is pipelined to the super computer center for data processing and mathematical modeling. Results can be accessed via a web-browser based user interface for data demonstration and visualization. Multidimensional images can be stacked on the basis of a standard marker, and images for each species can be extracted and presented individually. Comparison between mass spec and optical/fluorescence images facilitates results elucidation and explanation, providing valuable insights of

mechanisms for microbial colonies. We have identified mass signatures specific for *M. xanthus* DZ2 and *B. subtilis* 3610 using SIMS and nanoDESI/MS. Time-dependent imaging is underway for studying kinetics of colonies.

Ligninolytic peroxidases in the oyster mushroom genome: Differential regulation in lignocellulose medium

Elena Fernández-Fueyo,[1] Raúl Castanera,[2] José M. Barrasa,[1] Francisco J. Ruiz-Dueñas,[1] María F. López-Lucendo,[1] Lucía Ramírez,[2] Antonio G. Pisabarro,[2] and Angel T. Martínez* (ATMartinez@cib.csic.es)[1]

[1]CIB, CSIC, Madrid, Spain; [2]Department of Agrarian Production, UPN, Pamplona, Spain

Pleurotus ostreatus is an important edible mushroom and a model lignin degrading organism, whose genome contains nine genes of ligninolytic peroxidases, characteristic of white-rot fungi. These genes encode six manganese peroxidase (MnP) and three versatile peroxidase (VP) isoenzymes that differ in their catalytic and stability properties. Using liquid chromatography coupled to tandem mass spectrometry, secretion of four of these peroxidase isoenzymes (VP1, VP2, MnP2 and MnP6) was confirmed when *P. ostreatus* grows in a lignocellulose medium (pH 5.5) at 25 ºC (three more were identified by only one unique peptide). Then, the effect of environmental parameters on the expression of the above nine genes was studied by reverse transcription-quantitative PCR (RT-qPCR) by changing the incubation temperature and medium pH of the *P. ostreatus* cultures pre-grown under the above conditions (using two reference genes for normalization of the RT-qPCR results). The cultures maintained at 25 ºC provided the highest levels of peroxidase transcripts and the highest total activity on Mn^{2+} (a substrate of both MnP and VP) and Reactive Black 5 (a VP specific substrate). After global analysis of the expression patterns, peroxidase genes were divided into three main groups according to the level of expression at optimal conditions (*vp1/mnp3* > *vp2/vp3/mnp1/mnp2/mnp6* > *mnp4/mnp5*). Adjusting the culture pH to acidic or alkaline conditions (pH 3 and 8) or decreasing/increasing the incubation temperature (to 10 ºC/37 ºC) led to downregulation of most of the peroxidase genes (and decrease of the enzymatic activity) in most of the cases. The analysis also reveals differences in the transcription levels of the peroxidase genes when the culture temperature and pH parameters were changed, suggesting a possible adaptive expression according to environmental conditions. pH modification produced more dramatic effects than temperature modification, with *vp* expression resulting more affected than *mnp* expression. While *mnp3* was the less affected gene under temperature modified conditions, *mnp4* and *mnp5* were the only peroxidase genes being slightly upregulated under alkaline pH conditions.

Ligninolytic peroxidases in the oyster mushroom genome: Molecular structure, catalytic/stability properties and lignin-degrading ability

Elena Fernández-Fueyo,[1] Francisco J. Ruiz-Dueñas,[1] María Jesús Martínez,[1] Antonio Romero,[1] Kenneth E. Hammel,[2] Francisco Javier Medrano,[1] and Angel T. Martínez* (ATMartinez@cib.csic.es)[1]

[1]CIB, CSIC, Madrid, Spain; [2]Forest Products Laboratory, USDA, Madison, Wisconsin

The genome of *Pleurotus ostreatus* has been completed, as an important edible mushroom and as a model lignin-degrading organism. Heterologous expression of the nine class-II peroxidase genes,

followed by kinetic studies, confirmed the preliminary structural-functional classification for all but one of them. The resulting inventory revealed the presence of three versatile peroxidases (VPs) and six manganese peroxidases (MnPs), two of which were solved at 1.0-1.1 Å, and the absence of lignin peroxidases (LiPs). Gene number expansion supports the importance of both peroxidase types in the white-rot life-style of this fungus. Using ^{14}C-labeled model dimer and synthetic lignin we showed that *P. ostreatus* VP is able to degrade lignin. Moreover, the dual Mn-mediated and Mn-independent activity of *P. ostreatus* MnPs justifies their inclusion in a new peroxidase subfamily, with a role oxidizing phenolic degradation products. The availability of the whole POD repertoire enabled to investigate the existence of duplicated genes at a biochemical level. Interestingly, the isoenzymes not only differ in kinetic constants. Comparison of temperature and pH stabilities revealed surprising differences in activity T_{50} (43-63 ºC, after 10 min at pH 5), and residual activity at both acidic (0-96% after 4 h at pH 3) and alkaline pH (0-57% after 4 h at pH 9). Crystal structures and homology models, together with CD and UV-visible spectroscopy results, explained some of the stability differences, including the highest β-turn proline numbers in thermostable VP1, and the tightly-coordinated structural Ca^{2+} ions and number of exposed lysines in MnP4 being stable at both acidic and alkaline pH. The analysis of *P. ostreatus* genome reveals a lignin-degrading system where the role generally played by LiP has been assumed by VP. Moreover, it enabled the first comprehensive characterization of peroxidase isoenzymes in a basidiomycete, providing some clues on the high stability properties of some of them.

High-density linkage map of lowland *P. virgatum* (switchgrass) using single nucleotide polymorphisms from a reduced representation library

Fiedler, J.* (Jason.fiedler@ars.usda.gov),[1] Lanzatella, C.,[1] Okada, M.,[2] Jenkins, J.,[3] Schmutz, J.,[3] and Tobias, C.[1]

[1]USDA-Agricultural Research Service, Albany, California; [2]University of California at Davis, Davis, California; [3]HudsonAlpha Institute of Biotechnology, Huntsville, Alabama

Switchgrass is an allotetraploid C4 warm-season perennial grass with great potential as a dedicated bioenergy crop in the United States. However, the lack of genomic resources has inhibited the development of plant lines with optimal characteristics for sustainable feedstock production. We established a mapping population from a cross between two high-yielding lowland cultivars (i.e. Alamo A4 and Kanlow K5). We generated reduced representation libraries of 234 individuals and sequenced barcoded libraries on the Illumina Hi-Seq and My-Seq platforms. Approximately 150,000 single nucleotide polymorphisms (SNPs) were identified and 5,532 remained after filtering out SNPs with low-quality and abundant missing genotypes. SNP markers were integrated with previously reported microsatellite markers to produce 18 linkage groups of both parents via a *pseudo-test cross* strategy. Linkage maps generated from the linkage groups have high -marker density and agree with previously reported maps. The early release draft genome assembly of switchgrass (v1.1) is anchored to linkage maps resulting from analysis of an Alamo AP13 and an upland cultivar (Summer VS16) mapping population. Alignment A4xK5 linkage maps with the draft genome assembly results in collinearity and provides an independent metric of map and assembly quality. This methodology has proven to be a rapid and cost effective way to generate high quality linkage maps of heterozygous parents.

Identification of a multicopper oxidase from the microbial community of the Red Sea Atlantis-II brine pool

Mohamed A. Ghazy* (mghazy@aucegypt.edu), Mohamed Maged, and Ahmed Sayed

Department of Biology, The American University in Cairo, New Cairo, Egypt

Red Sea harbors a number of deep-sea brine pools with diverse chemical, physical, and geological conditions. The Atlantis II Brine Pool, is unique to the Red Sea, the deepest part of Atlantis II Brine Pool - the Lower Convective Layer (LCL) – is 2200 meters below the sea surface with an area of about 60km^2. It is characterized by extreme conditions, 26% salinity, temperature of about 68°C, low pH, it is almost anoxic, and contains a very high concentration of heavy metals. Microbial communities inhabiting this harsh environment are expected to have enzymes and proteins that are catalytically adapted to such extreme conditions. This work focused on identifying, isolating and characterizing extremophilic metal detoxifying enzymes and bioremediation catalysts such as multicopper oxidases from such environment, which is important group of enzymes with a broad range of biotechnological applications. They have a variety of applications in polymer synthesis, dye-bleaching in textiles and more importantly bioremediation. Toward this goal, and in collaboration with King Abdullah University for Science and Technology (KAUST), water samples from this environment was collected, DNA of the microbial community trapped on different size filters was isolated efficiently followed by direct 454-pyrosequencing.

We generated an ATII-LCL metagenomic dataset in our laboratory, the dataset composed of 4,104,966 454-pyrosequencing reads, assembled in 40,693 contigs. We mined the ATII-LCL metagenomic dataset for Multicopper oxidase (MCO) sequences, and a small MCO database was constructed using the predicted sequences. BLASTX analysis was performed on the predicted MCO sequences against the NCBI nr database. Based on the BLASTX results, positive MCO sequences were further analyzed to recognize the signal peptide and the regulatory elements. One full-length sequence was selected and PCR primers were designed to isolate the gene from the LCL genomic DNA. The gene was successfully isolated and the candidate MCO enzymes will be subjected to several biochemical analysis under several salt concentrations and temperatures to determine the optimum conditions for the new identified ATII-LCL Multicopper oxidase.

KBase outreach and partnership

Elizabeth M. Glass* (marland@mcs.anl.gov)[1], Brian Davison[2], Jason Baumohl[3], Aaron Best[4], Jared Bischof[1], Ben Bowen[3], Tom Brettin[1], Tom Brown[1], Shane Canon[3], Stephen Chan[3], John-Marc Chandonia[3], Dylan Chivian[3], Ric Colasanti[1], Neal Conrad[1], Matt DeJongh[4], Paramvir Dehal[3], Narayan Desai[1], Scott Devoid[1], Terry Disz[1], Meghan Drake[2], Janaka Edirisinghe[1], Gang Fang[7], José Pedro Lopes Faria[1], Mark Gerstein[7], Annette Greiner[3], Dan Gunter[3], James Gurtowski[6], Nomi Harris[3], Travis Harrison[1], Fei He[5], Matt Henderson[3], Chris Henry[1], Adina Howe[1], Marcin Joachimiak[3], Kevin Keegan[1], Keith Keller[3], Guruprasad Kora[2], Sunita Kumari[6], Miriam Land[2], Folker Meyer[1], Steve Moulton[2], Pavel Novichkov[3], Taeyun Oh[8], Gary Olsen[9], Bob Olson[1], Dan Olson[1], Ross Overbeek[1], Tobias Paczian[1], Bruce Parrello[1], Shiran Pasternak[6], Sarah Poon[3], Gavin Price[3], Srivdya Ramakrishnan[6], Priya Ranjan[2], Bill Riehl[3], Pamela Ronald[8], Michael Schatz[6], Lynn Schriml[10], Sam Seaver[1], Michael W. Sneddon[3] , Roman Sutormin[3], Mustafa Syed[2], James Thomason[6], Nathan Tintle[4], Will Trimble[1], Daifeng Wang[7], Doreen Ware[5,6], David Weston[2], Andreas Wilke[1], Fangfang Xia[1], Shinjae Yoo[5], Dantong Yu[5], Robert Cottingham[2], Sergei Maslov[5], Rick Stevens[1], and Adam P. Arkin[3]

[1]Argonne National Laboratory, Argonne, Illinois; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Lawrence Berkeley National Laboratory, Berkeley, California; [4]Hope College, Holland, Michigan; [5]Brookhaven National Laboratory, Upton, New York; [6]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

The Department of Energy (DOE) Systems Biology Knowledgebase (KBase) is an emerging computational environment that enables researchers to bring together the diverse data, algorithms, analytical tools, and workflows needed to achieve a predictive understanding of biological systems. KBase's Outreach team works to engage a diverse user community and maximize successful use of KBase to advance predictive biology. To accomplish this requires developing relationships to better understand researchers, their scientific objectives, and how KBase could benefit their research, as well as training users and developers. KBase is a community-driven project, which makes getting feedback and priorities from users very important. Feedback helps us improve KBase capabilities and workflows, improve documentation and training and understand the community's bioinformatics needs.

Besides, online tutorials, Outreach produces webinars and holds tutorials and developer bootcamps to provide education on the use of KBase and how to contribute. The Outreach team has conducted webinars to demonstrate and instruct users on how to use the system and services, with a total attendance to date of 129 people. This past year we have given 67 KBase presentations at meetings and conferences and held 9 face-to-face workshops and 8 KBase Developer bootcamps. To accompany such efforts and to promote independent learning, we develop education materials to assist users in navigating and using KBase effectively.

The KBase team is building partnering relationships with other synergistic projects. For important stakeholders--such as JGI, the BRCS, EMSL and iPlant--we have been co-designing science and software milestones, sharing infrastructure, defining routes for users to migrate between software and data systems, and developing cross-training programs. This allows KBase to maximize impact and relevance to the community and to prioritize its development goals.

Tight, inducible gene expression with an alternatively spliced suicide exon

Tania L. Gonzalez* (tlgonzalez@berkeley.edu),[1] Geoffrey Liou,[1,2] Scott Hickey,[1] Brian J. Staskawicz,[1] and Ming C. Hammond[1]

[1]University of California at Berkeley, Berkeley, California; [2]Massachusetts Institute of Technology, Cambridge, Massachusetts

Alternative splicing coupled to nonsense mediated decay is a gene regulation strategy common in nature, but up until now there has been no precedent to reliably and simply design splicing-regulated transgenes. We have engineered HyP5SM, a plant-specific splicing cassette that can be inserted directly and tracelessly into a variety of open reading frames to regulate protein expression in dicot plants. HyP5SM is called a "suicide exon" because its retention results in nonsense mediated decay of mRNA. HyP5SM exon skipping is induced upon co-expression with the monocot OsL5 protein. We demonstrate that HyP5SM can regulate the hypersensitive response (HR), a defensive programmed cell death response initiated by disease resistance plants upon detection of specific pathogen effector proteins. When effector proteins are regulated by inducible promoters only, the leaky protein expression is sufficient to trigger HR; transgenic plants with these genes are not viable. The addition of HyP5SM splicing regulation renders the leaky transcript non-productive, thus eliminating leaky protein detected by Western blot or leaky HR phenotype. Furthermore, plants inducibly recover both effector protein expression and the HR phenotype. The HyP5SM cassette exon can be generally applied to regulate other genes in dicot plants as well. We are interested in regulating many genes at once with HyP5SM. To avoid potential gene silencing, we have engineered new HyP5SM cassettes with different sequences. With the help of JGI, we will be testing these divergent HyP5SM cassettes in *N. benthamiana*.

The genus *Brachypodium* as a tool to study perenniality and polyploidy

Sean Gordon,[1] Jeremy Schmutz,[2,3] Kerrie Barry,[2] Pilar Catalan,[4] Bruno Contreras,[5] Robert Hasterok,[6] John Doonan,[7] Luis Mur,[7] Hikmet Budak,[8] Boulos Chalhoub[9] and John Vogel* (john.vogel@ars.usda.gov)[1]

[1]USDA-ARS-WRRC, Albany, California; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama; [4]Universidad de Zaragoza, Huesca, Spain; [5]Estación Experimental de Aula Dei (CSIC), Zaragoza, Spain; [6]University of Silesia, Katowice, Poland; [7]Aberystwyth University, Aberystwyth, United Kingdom; [8]Sabanci University, Istanbul, Turkey; [9]Unité de Recherches en Génomique Végétale (URGV-INRA), Évry, France

Polyploid genomes are characteristic of most of the grasses being developed as biomass crops and many grain crops. Therefore, a deeper understanding of gene regulation and genome evolution in polyploid genomes would be useful for developing improved crop varieties for both food and fuel. Unfortunately, biomass and grain crops are difficult experimental subjects because of their large, complex genomes, outbreeding nature and/or their large physical size. Thus, a simple model system to study polyploid genome regulation and evolution would be very useful. The diploid grass *Brachypodium distachyon* has emerged as a powerful model to study various aspects of grass biology and numerous experimental resources and tools have been developed. Another member of the genus, *B. hybridum*, is an allotetraploid with subgenomes derived from ancestors similar or identical to the current diploid species *B. distachyon* and *B. stacei*. All three species have very compact genomes, small stature and are easily grown and manipulated in the laboratory. We are sequencing the genomes of *B. hybridum* and *B. stacei*

and are conducting extensive comparative RNA-Seq experiments among the three species to understand the evolution of gene regulation in the subgenomes of *B. hybridum*. In addition, we have crossed *B. distachyon* and *B. stacei* to create a synthetic *B. hybridum*. When the genomes are completed, we will have in hand a powerful system to explore many questions about polyploid genome regulation and evolution. This system will have great utility in defining functional gene regulatory elements that can be translated into grass crops grass for applied applications.

Genetic, phenotypic and expression variation in highly diverse *Brachypodium distachyon* lines

Sean P. Gordon* (Sean.Gordon@ars.usda.gov),[1] David L. Des Marais,[2] Henry Priest,[3] Wendy Schackwitz,[4] Melania Figueroa,[5] Joel Martin,[4] Jennifer N. Bragg,[1] Ludmila Tyler,[6] Cheng-Ruei Lee,[7] Doug Bryant,[3] Wenqin Wang,[8] Joachim Messing,[8] Kerrie Barry,[4] David Garvin,[9] Hikmet Budak,[10] Metin Tuna,[11] Thomas Mitchell-Olds,[7] William F. Pfender,[12] Jennifer Bragg,[1] Richard Poire,[13] Robert Furbank,[13] Liang Chen,[14] Sun Choi,[14] Hoi-Ying Holman,[14] Thomas Juenger,[15] Todd C. Mockler,[3] and John P. Vogel[1]

[1]USDA-ARS-WRRC, Albany, California; [2]Department of Organismic and Evolutionary Biology, Harvard University, Boston, Massachusetts; [3]Donald Danforth Plant Science Center, Saint Louis, Missouri; [4]DOE Joint Genome Institute, Walnut Creek, California; [5]University of Minnesota, St. Paul, Minnesota; [6]University of Massachusetts, Amherst, Massachusetts; [7]Duke University, Durham, North Carolina; [8]Waksman Institute, Rutgers University, Piscataway, New Jersey; [9]USDA-ARS, PSRU, St. Paul, Minnesota; [10]Sabanci University, Istanbul, Turkey; [11]Namik Kemal University, Tekirdag, Turkey; [12]USDA-ARS, Corvallis, Oregon; [13]Australian Phenomics Facility, Canberra, Australia; [14]Lawrence Berkeley National Laboratory, Berkeley California; [15]The University of Texas at Austin, Austin, Texas

Natural variation is a powerful resource for studying the genetic basis of biological traits. Brachypodium distachyon is a model grass with a small genome and a large collection of diverse, inbred, diploid lines. As a step towards understanding the genetic basis for this variation, we sequenced the reference accession, Bd21, and 53 divergent lines from 22-160-fold depth with paired-end Illumina reads. We are using these data to generate assemblies and gene annotations for the divergent lines as well as a comprehensive set of genetic variants. We are using genetic variants in combination with RNA-Seq information from the reference line and 30 divergent lines to identify genome-wide expression quantitative trait loci, presence-absence variation, and alternative poly-adenylation events. To correlate genetic and expression variation to phenotypic variation among the lines we are collaborating with Lawrence Berkeley Laboratories and applying high-throughput Synchrotron FTIR to measure differences in cell wall chemistry. In addition, the High Resolution Plant Phenomics Centre (Canberra, Australia) has conducted high-throughput phenotyping of many of these lines. These data will be used to connect genetic variation to gene expression and control of traits relevant for grass biomass feedstocks.

Current developments in prokaryotic single cell whole genome amplification

Danielle Goudeau* (dgoudeau@lbl.gov), Nandita Nath, Doina Ciobanu, Jan-Fang Cheng, and Rex Malmstrom

DOE Joint Genome Institute; Walnut Creek, California

Our approach to prokaryotic single-cell Whole Genome Amplification at the JGI continues to evolve. To increase both the quality and number of single-cell genomes produced, we explore all aspects of the process from cell sorting to sequencing. For example, we now utilize specialized reagents, acoustic liquid handling, and reduced reaction volumes eliminate non-target DNA contamination in WGA reactions. More specifically, we use a cleaner commercial WGA kit from Qiagen that employs a UV decontamination procedure initially developed at the JGI, and we use the Labcyte Echo for tip-less liquid transfer to set up 2uL reactions. Acoustic liquid handling also dramatically reduces reagent costs. In addition, we are exploring new cell lysis methods including treatment with Proteinase K, lysozyme, and other detergents, in order to complement standard alkaline lysis and allow for more efficient disruption of a wider range of cells. Incomplete lysis represents a major hurdle for WGA on some environmental samples, especially rhizosphere, peatland, and other soils. Finding effective lysis strategies that are also compatible with WGA is challenging, and we are currently assessing the impact of various strategies on genome recovery.

Photosynthetic production of a cyclic hydrocarbon by an engineered cyanobacterium

Charles Halfmann* (charles.halfmann@sdstate.edu), Liping Gu, and Ruanbao Zhou

South Dakota State University, Biology and Microbiology, Brookings, South Dakota

Terpenes are a large class of organic molecules with important applications in pharmaceuticals, cosmetics, and biofuels. Current technologies for harvesting terpenes from plant biomass require arable land, impose high energy demands from industrial processing, and contribute to global climate change by emitting CO_2 into the atmosphere. Here, we focus on a direct photosynthesis-to-product approach in terpene product by metabolically engineering a cyanobacterium as cellular machinery to over-produce and excrete valuable compounds into a photo-bioreactor using CO_2, H_2O, and light. As a proof of concept, we have engineered the filamentous cyanobacterium *Anabaena* 7120 to synthesize the cyclic hydrocarbon limonene ($C_{10}H_{16}$) by introducing a limonene synthase gene (*lims*) into its genome. To improve the product yield, we created a synthetic *lims-dxs-ipphp-gpps* operon to drive carbon flux towards limonene synthesis. Our study revealed that limonene produced by the genetically engineered cyanobacterium was excreted across the cell membrane and volatilized into the culture headspace, allowing for quick separation of the target compound from the culture biomass. We envision that the platform of using cyanobacteria as a cellular factory and CO_2 as a sustainable feedstock can be applicable for the production of a wide range of commodity chemicals and drop-in-fuels.

Metagenomic and functional analysis of microbial communities associated with natural crude oils that seep into the Santa Barbara Channel offshore Southern California

Erik R. Hawley*,[1] Alexander Sczyrba,[2] Nicole M. Scott,[3] Stephanie Malfatti,[4] Ioanna Pagani,[5] Marcel Huntemann,[5] Amy Chen,[5] Tijana Glavina del Rio,[5] Brian Foster,[5] Alex Copeland,[5] Janet Jansson,[5,6] Amrita Pati,[5] Susannah Tringe,[5,6] Jack A. Gilbert,[3,7] Thomas D. Lorenson,[8] and Matthias Hess (matthias.hess@tricity.wsu.edu)[1,5,9,10]

[1]Washington State University Tri-Cities, Richland, Washington; [2]University of Bielefeld, Bielefeld, Germany; [3]Argonne National Laboratory, Lemont, Illinois; [4]Lawrence Livermore National Laboratory, Livermore, California; [5]DOE Joint Genome Institute, Walnut Creek, California; [6]Lawrence Berkeley National Laboratory, Berkeley, California; [7]University of Chicago, Chicago, Illinois; [8]U.S. Geological Survey, Menlo Park, California; [9]Pacific Northwest National Laboratory, Richland, Washington; [10]Washington State University, Pullman, Washington

Hydrocarbons are main constituents of crude oil, major contaminants of the marine ecosystem, and many processes are involved in their degradation. To improve our understanding of the microbiology associated with oil that enters the marine ecosystem, we analyzed DNA extracted from three distinct crude oils that seeped into the Santa Barbara Channel (SBC). The V4 region of the 16S rRNA gene was amplified and ~133,000 high-quality reads were generated for each sample. The generated reads were classified into 51 bacterial and 2 archaeal phyla – with ~97% of the reads classified at the genus level. Sequences were clustered into 2,106 operational taxonomic units (OTUs), of which 283 OTUs were shared between all three oils. The most abundant OTU of Oil-1 and Oil-3 was identified as uncultured member of the *Campylobacterales* and of the candidate phylum *Atribacteria* respectively. The most abundant OTU of Oil-2 was classified as a member of the *Paenisporosarcina*. Interestingly, members of the *Oceanospirillales*, a bacterial order which recruited more than 60% of the microbial community sequence data generated from the oil plume that formed in the aftermath of the Deepwater Horizon (DWH) blowout, recruited only 1.4%, 0.5% and 11.6% of the reads generated from Oil-1, Oil-2 and Oil-3 respectively. To determine the metabolic potential of the microbiomes associated with the SBC oils, we generated ~52Gb of metagenomic sequence data from each sample. On average ~500Mb of sequence, representing ~1 million protein coding genes, were assembled for each metagenome. Among the SBC oils, Oil-1 and Oil-2 were most closely related (ρ=0.55) followed by Oil-2 and Oil-3 (ρ=0.41). Correlations between the SBC oils and the DWH oils ranged between ρ=0.05 and ρ=0.17. This suggests that the functions of the microbial communities associated with the SBC oils are more similar to each other than to those from the microbial communities associated with the DWH oils. It is likely that this is caused by the fact that the DWH oil plume was oxygen-saturated, whereas the SBC oils are mostly anaerobic. To test this hypothesis the assembled metagenomes will be analyzed for key pathways of anaerobic and aerobic methane oxidation, microbial processes that play a significant role in the marine carbon cycle. Results of this analysis will be presented.

Synthesis of *S*-adenosylmethionine (SAM) analogs and development of high-throughput screens to target SAM-I riboswitches

Scott F. Hickey* (sfhickey1@gmail.com,[1,2] Khiem Nguyen,[4] Andrew J. Cullen,[1] and Ming C. Hammond[1,2,3]

[1]Department of Chemistry, [2]Synthetic Biology Institute, [3]Department of Molecular & Cell Biology, University of California at Berkeley, Berkeley, California; [4]Department of Chemistry, University of Utah, Salt Lake City, Utah

S-adenosylmethionine (SAM) is the second most widely utilized cofactor inside the cell, where it acts as a universal methyl donor for protein, DNA, RNA, and small molecule methyltransferase reactions. SAM-dependent methyltransferases have been the targets of diverse therapeutic efforts in cancer and infectious disease, including the use of SAM analogs as methyltransferase inhibitors. Recently, SAM-binding riboswitch RNAs have been described that control gene expression in bacteria. These riboswitches are found exclusively in bacteria and represent promising targets for novel antibacterial development. Riboswitches are typically found in the 5'-UTR of mRNAs and control downstream gene expression by sensing cellular metabolites through an aptamer domain. Although riboswitch sequences are conserved within a specific family, there are significant differences between riboswitches in distinct bacterial species. For example, SAM-I class riboswitches have been identified in both pathogenic (*Staphylococcus*) and commensal (*Lactobacillus*) bacteria, representing a potential opportunity to develop selective antibacterial compounds for targeting only a specific SAM-I riboswitch RNA. Using structure-guided design, we have synthesized stable and fluorescent SAM analogs in which the fluorophore is introduced through a functionalizable linker to the ribose. A Cy5-labeled SAM analog was shown to bind several SAM-I riboswitches via in-line probing and fluorescence polarization assays, including one from *Staphylococcus aureus* that controls the expression of SAM synthetase in this organism. A fluorescent ligand displacement assay was developed and validated for high-throughput screening of compounds to target the SAM-I riboswitch class. In parallel, we have developed a microfluidic screening platform for SAM-I riboswitches that is well suited to high-throughput screening due to its low resource consumption, rapid measurment times, and high reproducibility. We are using these assays to screen large compound libraries and validate selective hit compounds for the identification of pathogen-specific antibacterials. Finally, we have developed a practical and efficient synthesis of unnatural SAM analogs for use as rationally-designed methyltransferase and riboswitch inhibitors, and for directed evolution of orthogonally-selective methyltransferases and riboswitches.

Optimization of a thermotolerant and ionic liquid-tolerant cellulase cocktail, Jtherm, for the saccharification of lignocellulosic biomass using metagenome derived glycoside hydrolases

Jennifer Hiras* (jhiras@lbl.gov)[1,2], Blake A. Simmons[1,3], and Steven W. Singer[1,4*]

[1]Deconstruction Division, Joint BioEnergy Institute, Emeryville, California; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]Biofuels and Biomaterials Science and Technology Department, Sandia National Laboratories, Livermore, California; [4]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Development of thermotolerant cellulase cocktails may advance the production of sugars from lignocellulosic biomass for downstream fermentation to biofuels. Previously, a thermophilic and ionic liquid (IL)-tolerant bacterial cellulase cocktail, Jtherm, was developed at the Joint BioEnergy Institute.

Jtherm was derived from the secretome of a thermophilic bacterial community cultivated on microcrystalline cellulose and supplemented with two recombinant enzymes, a CBM3-GH5 from *Caldicellulosiruptor saccharolyticus* (cellobiohydrolase) and GH3 from *Thermotoga petrophila* (β-glucosidase). This mixture released glucose from IL-pretreated switchgrass at up to 80 ^{o}C and in up to 20% of the ionic liquid 1-ethyl-3-methylimidazole acetate. Here, we describe efforts to improve efficiency of Jtherm while decreasing enzymatic loading. One potential area for improvement involves xylan, a large component of lignocellulosic biomass, which is known to inhibit cellulase activity. We hypothesize that the addition of xylanases will improve the hydrolysis of xylan and thereby increase cellulase activity. A β-xylanase (EC 3.2.1.55) and two endoxylanases (EC 3.2.1.8) belonging to *Rhodothermus marinus* were identified from the secretome of the original Jtherm base cocktail by metaproteomic measurements. These three proteins were overexpressed in *E. coli*, subsequently purified, and demonstrated high activity at 70 ^{o}C. Hemicellulase (endoxylanase and beta-xylanase) activities were measured using artificial substrates and 3,5-dinitrosalysilic acid (DNS) *p*-nitrophenol (pNP) enzyme assays. Using high-throughput robotics techniques, we are supplementing Jtherm with combinations of these purified accessory proteins and screening for increased sugar production.

Measurements and modeling atmospheric gene flow from GR horseweed (*Conyza canadensis*)

Haiyan Huang.[1], Rongjian Ye[2], Yanhui Peng[2], C. Junming Wang* (wangjim@illinois.edu),[1] Neal Stewart, Jr.[2], and David DuBois[3]

[1]Illinois State Water Survey, Prairie Research Institute,University of Illinois at Urbana-Champaign, Champaign, Illinois; [2]University of Tennessee, Knoxville, Tennessee; [3]New Mexico State University, Las Cruces, New Mexico

Horseweed has become increasingly problematic in agricultural fields throughout the United States due to its glyphostate resistance (GR). Like many wind-dispersed species, even a single mature horseweed plant can produce significant populations of lightweight seed/pollen, which can stay in the atmosphere for a long period and may travel a long distance and transfer (outcross) genes to other non-GR horseweeds. Quantifying horseweed seed/pollen production and dispersal is of great importance to agricultural management strategies.

For quantifying the gene flow, GFP pollen horseweed is being made for GR horseweed; the pollen can be discriminated from the glyphosate susceptible (GS) receptor horseweed's and outcrossed seed seedlings can be discriminated from that pollinated by GS receptor horseweed by fluorescence.

Two field experiments were conducted to measure horseweed seed/pollen production and horizontal and vertical dispersal and collect atmospheric data. Local and regional models/online tools were built that can be used to quantify pollen/seed dispersion and deposition under different atmospheric conditions, vegetation covers, and topographies. (http://rsetserver.sws.uiuc.edu/horseweed/)

The JGI pipeline for annotation of microbial genomes and metagenomes

Marcel Huntemann* (mhuntemann@lbl.gov),[1] Konstantinos Mavrommatis,[1] Natalia Ivanova,[1] Natalia Mikhailova,[1] Galina Ovchinnikova,[1] Andrew Schaumberg,[1] Jim Tripp,[1] Krishna Palaniappan,[2] Ernest Szeto,[2] I-Min Chen,[2] Nikos Kyrpides,[1] and Amrita Pati[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Lawrence Berkeley National Laboratory, Berkeley, California

Here we present JGI-PAMM v1.0, the JGI Pipeline for Annotation of Microbial genomes and Metagenomes. All microbial genomes and metagenomes sequenced at the JGI or submitted to IMG are processed by the JGI-PAMM on a compulsory basis. The pipeline, which performs quality analysis, screening, sequence dereplication, structural annotation, and functional annotation, has been deployed in an integrated manner since the beginning of 2012 and has processed more than 16 Billion genes till date. The models for *de novo* structural annotation used in JGI-PAMM were derived after exhaustive benchmarking and comparison with best practices. Similarly, functional annotation results in genes predicted by structural annotation being associated with Pfams, COGs, KOGs, KO numbers, KEGG pathways, IMG terms and pathways, EC numbers, TC numbers, MetaCyc pathways, phenotypes, and putative biosynthetic activity/associated natural products, and homologs in other genomes. These associations are then used in several specific contexts as part of various comparative analysis tools in IMG and IMG-M. Constant monitoring by on-site scientists and a user base of over 4,000 ensures that the pipeline is up to date with the latest technologies in the field and delivers accurate products. The pipeline is available to users submitting their genomes via IMG's submission system.

JGI-PAMM is deployed on computational systems at NERSC (www.nersc.gov) and the functional annotation tools are implemented within the Hadoop framework. It can handle large unassembled metagenome datasets as efficiently as smaller assembled datasets.

Microbial community structure and function in acidic geothermal Fe(III) mineralizing systems

W.P. Inskeep* (binskeep@montana.edu),[1] J. Beam,[1] H. Bernstein,[2,3] R. Brown,[3] R. Carlson,[2] J. Fredrickson,[3] Z. Jay,[1] R. Jennings,[1] M. Kozubal,[1] H. Kreuzer,[3] M. Lipton,[3] J. Moran,[3] M. Romine,[3] S. Tringe,[4] and T. Woyke[4]

[1]Department of Land Resources and Environmental Sciences and Thermal Biology Institute, Montana State University, Bozeman, Montana; [2]Department of Chemical and Biological Engineering, Montana State University, Bozeman, Montana; [3]DOE-Pacific Northwest National Laboratory, Richland, Washington; [4]DOE Joint Genome Institute, Walnut Creek, California

Microbial growth and activity are responsible for the mineralization of Fe(III)-oxides in chemotrophic communities found within oxygenated zones of low pH (2.5 – 3.5) geothermal outflow channels. Long-term research on several types of high-temperature Fe(III)-microbial mats in Yellowstone National Park (YNP) has shown that these communities are dynamic and accrete Fe(III)-oxide at rates ranging from ~ 0.1 - 2 mm per month, depending on seasonal and/or geochemical attributes. We hypothesize that Fe(III)-oxide mats form as a result of constant interaction among primary colonizers including *Hydrogenobaculum* spp. (bacterial order Aquificales) and *Metallosphaera* spp. (archaeal order Sulfolobales), and subsequent colonization by archaeal heterotrophs, which vary in abundance as a function of oxygen, pH and temperature. We are integrating a complementary suite of geochemical,

genomic, proteomic and modeling approaches to study the role of microorganisms in Fe(III)-oxide mat development, and to elucidate the primary microbial interactions that are coupled with key abiotic events, and in turn, drive community succession. Curated *de novo* assemblies of major phylotypes were obtained from random sequence data and are being used to analyze additional random sequencing, transcriptomic and proteomic measurements. *Hydrogenobaculum* spp. are the dominant bacterial population(s) present, and have been shown to predominate during early stages of mat development (< 30 d). Although not thought to be responsible for Fe(II)-oxidation, these autotrophic populations are important initial colonizers of Fe(III)-oxide microbial mats. Other Sulfolobales populations known to oxidize Fe(II) and fix carbon dioxide (e.g., *Metallosphaera* spp.) require a longer time to develop (e.g., 14 – 30 d), and it is not clear whether these populations may be auxotrophs, which are dependent on co-factors produced by *Hydrogenobaculum* spp. Although *Hydrogenobaculum* spp. are early colonizers, their cell surfaces appear to be excellent sites for the mineralization of Fe(III)-oxides, which likely has an effect on overall growth and turnover rates. Several heterotrophic organisms colonize Fe(III)-oxide mats over time (> 30 d), including several novel archaeal lineages (Candidate phylum Geoarchaeota, Novel Archaeal Groups 2 and 3, Thaumarchaeota) as well as representatives within the Sulfolobales, Desulfurococcales and Thermoproteales. Several of these phylotypes appear to be more adapted to low oxygen environments, which have been shown to exist within Fe(III)-oxide mats of Norris Geyser Basin (YNP) at depths below approximately 0.5 -1 mm. Physicicochemical factors controlling the distribution of these heterotrophs are not fully understood, but may include oxygen, pH, temperature and key interactions among community members. Oxygen consumption measurements indicate that steep gradients occur within the top 1 mm of mat surface, and suggest that different organisms may occupy different niches in these microenvironments. Proteomic data support the hypothesis that *Hydrogenobaculum* and *Metallosphaera* are predominant autotrophs in these systems, and may supply other heterotrophs with a source of fixed carbon. Isotopic (^{13}C) evidence shows that Fe(III) mats contain a significant amount (> 40 %) of fixed C from dissolved inorganic C. Consequently, autotroph-heterotroph associations may control the dynamics and succession of thermal Fe(III)-oxide mineralizing communities. Future work will elucidate functional attributes of specific phylotypes and their spatiotemporal role in community function and succession.

Studies on terrestrial deep biosphere of Fennoscandian Shield

Merja Itävaara* (merja.itavaara@vtt.fi)[1], Ilmo Kukkonen[2], Lasse Ahonen[3], Riikka Kietäväinen[3], Malin Bomberg[1], Lotta Purkamo[1], Hanna Miettinen[1], Heikki Salavirta[1], Minna Vikman[1], Juho Rousu[4]

[1]VTT, Finland; [2]University of Helsinki, Department of Geophysics; [3]Geological Survey of Finland; [4]Aalto University, Finland

Microorganisms are dominant life forms on Earth, and they have a great impact on Earth's geobiological cycles and climate change. Especially under intensive investigation is carbon cycle, which is especially connected to gases evolving from the mantle of earth and can be used as sources for energy and carbon. Our research is connected to solve basic scienctific questions concerning the existence of deep life and it's different forms as well as studying geomicrobial processes. On the hand we also focus on studies concerning nuclear waste disposal safety issues, where also anaerobic methane oxidation is considered as an important topic area.

Our multdisciplinary team consists of geologists, microbiologists and bioinformatician aiming to develop sampling techniques and studying deep life based on several molecular methods, sequencing and connecting this data to physicochemical parameters such as geology, geochemistry and gases.

We are also connected to DCO (Deep carbon observatory (Deep Life project) and have been focusing on methane oxidizing microbial communities and sulphate reduction. Microorganisms in deep groundwaters are metabolizing very slowly and therefore addition of electron acceptors induces metabolic activitation which we used to separate the cells by staining the active cells for sequencing. Fluorescent activated cell sorting (FACS) was performed at Bigelow laboratory and sequencing at JGI, USA.

Nyyssönen, M. Hultman, J., Ahonen, L., Kukkonen, I., Paulin, L., Laine, P., Itävaara, M. and Auvinen, P. 2014.. Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield" - ISME J, 8: 126–138; doi:10.1038/ismej.2013.125.

Purkamo L, Bomberg M, Nyyssönen M, Kukkonen I, Ahonen L, Kietäväinen R, Itävaara M. 2013. Dissecting the deep biosphere: retrieving authentic microbial communities from packer-isolated deep crystalline bedrock fracture zones. –FEMS Microbiology Ecology, Volume 85, Issue 2, pages 324–337, August 2013, DOI: 10.1111/1574-6941.12126

Kietäväinen, R., Ahonen, L-, Kukkonen, I.T., Hendriksson, N., Nyyssönen, M. Itävaara, M. 2013.

Characterisation and isotopic evolution of saline waters of the Outokumpu Deep Drill Hole, Finland – Implications for water origin and deep terrestrial biosphere. Applied Geochemistry 32, 37-51. doi.org/10.1016 j.apgeochem.22012.10.013,

Nyyssönen, M., Bomberg, M., Kapanen, A., Nousiainen, A., Pitkänen, P., Itävaara, M. 2012. Methanogenic and sulphate-reducing microbial communities in deep groundwater of crystalline rock fractures in Olkiluoto, Finland. - Geomicrobiology Journal, 29:863–878, 2012, doi-link: 10.1080/01490451.2011.635759

Itävaara, M., Nyyssönen, M., Kapanen, A.,Nousiainen, A., Ahonen, L., Kukkonen, I. 2011. Characterization of bacterial diversity down to a depth of 1500 m of the Outokumpu deep borehole. FEMS Microbiology 2011, 1-15,. DOI:10.1111/j.1574-6941.2011.01111.x

Itävaara, M., Nyyssönen, M., Bomberg, M., Kapanen, A., Nousiainen, A., Ahonen, L., Hultman, J., Paulin, L., Auvinen, P., Kukkonen, I.T. 2011. Microbiological sampling and analysis of the Outokumpu Deep Drill Hole biosphere in 2007-2009. 'Special Paper' series of the Geological Survey of Finland. The Outokumpu Deep Drilling Project 2004-2010.

Stop codon reassignments in the wild

Natalia N. Ivanova,[1] **Patrick Schwientek**,[1] H. James Tripp* (HJTripp@lbl.gov),[1] Christian Rinke,[1] Amrita Pati,[1] Marcel Huntemann,[1] Axel Visel,[1,2,3] Tanja Woyke,[1] Nikos C. Kyrpides,[1] and Edward M. Rubin[1,2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]School of Natural Sciences, University of California at Merced, Merced, California

The genetic code appears largely conserved across all domains of life, but some deviations from the canonical genetic code have been reported, mainly reassignment of stop codons to amino acids. One of the drivers for stop codon reassignment is thought to be an escape from predation by phage, predicated on the assumption that phage are constrained to use the translation machinery of the host. In order to challenge this assumption and gain insight into the phenomenon of stop codon reassignment in the wild, we undertook a comprehensive survey of all detectable stop codon reassignments in 1726

metagenomic samples totaling 5.6 trillion base pairs. Our survey uncovered circumstantial evidence in healthy human oral microbiomes of a phage with reassigned amber codons infecting a host with reassigned opal codons. Phylogenetic analysis of the hundreds of observations of stop codon reassignment identified primarily opal reassignment in Bacteria and amber reassignment in phages. We saw no evidence of stop codon reassignment in Archaea despite substantial sampling depth. There was a bacterial opal-to-glycine reassignment in the uncultured phyla Gracilibacteria and SR1 that appeared to be inherited from their common ancestor. Our comprehensive compilation of non-canonical stop codon usage in nature shows that while stop codon reassignment has clear phylogenetic boundaries, these boundaries are not insurmountable by phages carrying their own copies of translation machinery.

KitBase: Towards a fully indexed database for 2000 sequenced rice mutant lines

Rashmi Jain* (rsjain@ucdavis.edu)[1,2], Guotian Li[1,2], Mawsheng Chern[1,2], Liangrong Jiang[1], Miguel Vega-Sanchez[1,2], Patrick E. Canlas[1,2], Wendy Schackwitz[3], Joel Martin[3], Nicholas Santoro[4], and Pamela Ronald*[1,2]

[1]Department of Plant Pathology and the Genome Center, University of California at Davis, Davis, California; [2]Joint BioEnergy Institute, Emeryville, California; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan

Background: Rice is a model plant species for research in monocots, which include cereal crops and candidate bioenergy crops. The rice varieties Nipponbare, Dongjin and Hwayoung, often used in rice genetic studies, have long life cycles and can be difficult to grow. For these reasons, the rice variety, Kitaake, has become popular because this variety has a smaller stature and a much shorter life cycle (ca. 9 weeks).

Description: We have generated a mutant population for Kitaake using fast-neutron mutagenesis. From 7,000 M1 mutant lines, we selected 2,000 lines for whole genome sequencing and analysis, in collaboration with The Joint Genome Institute (JGI). Sequencing of 178 lines is complete. We carried out saccharification assays on 5,000 M2 lines and identified 95 candidate mutants altered in Saccharification. Our goal is to identify the genes corresponding to the mutation through whole genome sequence analysis. To optimize use of this data we are developing KitBase, a database that uses MySQL and PHP. The goal of the database is to make seed information, phenotypes and genotypes easily accessible to the research community.

Microbial dark matter: Unusual intervening sequences in the 16S rRNA genes of candidate phyla from the deep subsurface

Jessica Jarett* (jkjarett@lbl.gov),[1] Ramunas Stepanauskas,[2] Thomas L. Kieft,[3] Tullis Onstott,[4] Lily Momper,[5] Jan Amend,[5] and Tanja Woyke[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine; [3]New Mexico Tech, Soccoro, New Mexico; [4]Princeton University, Princeton, New Jersey; [5]University of Southern California, Los Angeles, California

The Microbial Dark Matter project has sequenced genomes from over 200 single cells from candidate phyla, greatly expanding our knowledge of the ecology, inferred metabolism, and evolution of these

widely distributed, yet poorly understood lineages. The second phase of this project aims to sequence an additional 800 single cells from known as well as potentially novel candidate phyla derived from a variety of environments. In order to identify whole genome amplified single cells, screening based on phylogenetic placement of 16S rRNA gene sequences is being conducted. Briefly, derived 16S rRNA gene sequences are aligned to a custom version of the Greengenes reference database and added to a reference tree in ARB using parsimony. In multiple samples from deep subsurface habitats but not from other habitats, a large number of sequences proved difficult to align and therefore to place in the tree. Based on comparisons to reference sequences, structural alignments using SSU-ALIGN, and manual inspection and editing of alignments, many of these "difficult" sequences appear to originate from candidate phyla, and contain intervening sequences (IVSs) within the 16S rRNA genes. These IVSs are short (approximately 60 to 100 nt) and do not appear to be self-splicing or to contain open reading frames. Several different taxonomic groups contain IVSs, and these are placed in a similar location in the gene in all of the members of each group. Comparison of the phylogenetic placement of these groups with the IVSs both included and removed is ongoing. Based on data from samples screened in this project, IVSs appear to be more common in microbes occurring in deep subsurface habitats, although the reasons for this remain elusive.

Identification of genes for saccharification in rice by high throughput mutant screening and whole-genome resequencing

Liangrong Jiang*(lrjiang@ucdavis.edu)[1,2], Guotian Li[1,2], Mawsheng Chern[1,2], Rashmi Jain[2], Miguel Vega-Sanchez[1,2], Patrick Canlas[1,2], Wendy Schackwitz[3], Joel Martin[3], Nicholas Santoro[4], and Pamela Ronald[1,2]

[1]Feedstocks Division, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Department of Plant Pathology and the Genome Center, University of California at Davis, Davis, California; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Department of Energy Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan

The rice variety, Kitaake, is a model for grass genetic studies. As a first step towards identifying genes that control cell wall saccharification, we screened 3,000 M_2 kitaake mutants generated with fast neutron irradiation. Fifty-two lines showed an increased accumulation of reducing sugars and five showed a decrease compared to the wild type. Another 2,000 M_2 lines were screened using a robotic high-throughput saccharification platform that separately measured glucose and pentose yields. Outliers were selected based on a Z factor which is greater than 10 or smaller than -10. One hundred eighty nine outliers were identified that had alterations in in glucose yields and tow hundred seventy-seven outliers for alterations in pentose yields. The correlation coefficient between glucose and pentose yields ($r_{G\,\&P}$) is 0.5923** and the determination coefficient (r^2) is 0.3508. A secondary screen is underway to verify these candidates.

One candidate (*rcs60*) that showed reduced cell wall saccharification has been further characterized. The *rcs60* mutant segregates for reduced saccharification efficiency in the M_2 generation. The progeny plants of the *rcs60* mutant are all dwarfed and impaired in fertility. W60-4 is a heterozygous sibling of *rcs60* and its progeny segregated in plant height. The average height of the normal group in the population is 86.5 cm, which is similar to that of the parent. The average height of the dwarf group is 27.0 cm. The segregation ratio between normal and dwarf plants conforms to the expected 3:1 ratio for the recessive inheritance. Another candidate, *ecs786,* grows normally, but showed a 75.6% increase in glucose yield and 61.5% increase in pentose yield. In collaboration with JGI, we are using a whole-genome re-sequencing approach to identify all mutations in each mutant line. The re-sequencing data

indicate that *ecs786* carries eight mutations that affect 8 genes. Four of these affected genes are annotated to be associated with polysaccharide metabolism. In summary, our mutant screen approach provides an efficient way to identify novel rice genes that affect cell wall saccharification.

Deep surveys of biological modules: K-biclustering functional genomics data

Marcin P. Joachimiak* (MJoachimiak@lbl.gov),[1] Cathy Tuglus,[2] Mark van der Laan,[2] and Adam P. Arkin[1,3]

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Department of Biostatistics, [3]Department of Bioengineering, University of California at Berkeley, Berkeley, California

New algorithms that statistically associate features of genomic data to predict phenotypic outcome are becoming increasingly important in precision medicine, synthetic biology, and microbial ecology. One of the first critical steps is the detection of patterns in the data which group together features across measured molecules that are similar over a subset of conditions. Many such algorithms are based on a biclustering procedure wherein sets of molecular features such as genetic sequence, molecular abundance or molecular interactions co-occur in certain patterns across a set of conditions or outcomes. While current approaches have proven useful it has also become clear that they suffer from limitations that reduce their predictive power: inability to identify overlapping patterns, discovery of biclusters that are too small or too large, and low coverage of the available signal in datasets. Moreover, it has proven difficult to compare the strengths and weakness of different algorithms to one another because of a lack of gold-standard data sets and criteria for evaluation.

Here we develop a new method, Massive Associative K-biclustering (MAK, http://genomics.lbl.gov/MAK), addressing the key limitations in other biclustering approaches. In a formal evaluation against new simulated data sets with more realistic properties MAK improves recovery of biclusters exhibiting common behavior across conditions, with the highest F1scores, a measure of a test's accuracy, and lowest false positive rates, the probability of falsely rejecting the null hypothesis, compared to established methods. We show how MAK can be applied to a popular data set from *Saccharomyces cerevisiae* to uncover new properties not captured by other methods. MAK biclusters showed increased statistical enrichments for known regulatory associations and gene functions including both types of enrichments for nearly 30% of the 174 novel MAK biclusters not overlapping biclusters from other methods. Compared to other methods, MAK results had at least two-fold higher coverage of differential expression in the overall dataset, an overlapping bicluster network with distinct modules for protein expression and energy metabolism, and biologically relevant overlapping biclusters including cases of combinatorial regulation. MAK is especially suited for large datasets with many potentially noisy and overlapping patterns of varying sizes.

Culture-based assay for screening microbial lignin deconstruction

Chijioke Joshua* (cjjoshua@lbl.gov), Blake A. Simmons, and **Steven W. Singer**

Joint BioEnergy Institute, Emeryville, California

Lignin is a recalcitrant component of lignocellulose and a potential source of high value chemicals. Lignin consists of three main aromatic subunits: p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) subunits

linked by complex array of C-C and C-O bonds. Microbial lignin deconstruction is a very challenging primarily due to the absence of validated routine screening methods. Previous studies have focused on lignin fungal breakdown that employ free radical mechanisms involving enzymatic mechanisms (white-rot) and/or small molecule (brown rot) catalysts. Bacterial lignin deconstruction has been documented but mechanistic models are lacking. We have developed a culture-based assay for screening lignin deconstructing species that was based on the clearance of the dark brown color from minimal media supplemented with colloidal lignin-rich substrates within 3 – 5 days and the appearance of large cell-lignin aggregates. As a proof of concept, we demonstrated clearance of the assay broth by *Streptomyces viridosporus* T7A*,* a model lignin depolymerization species, compared to an uninoculated control, while no change was observed with *Escherichia coli*. Using this assay, we identified potential lignin degrading species, including *Lysinbacillus*, a previously unidentified lignin deconstruction species. Lysinibacillus, rapidly cleared cultures with lignin and forming cell-lignin aggregates characterized by microscopy. This lignin-clearing assay has been demonstrated with lignin from Ionic-liquid pre-treated and saccharified eucalyptus, as well as commercially available forms of insoluble lignin. Efforts are currently focused on understanding the detailed mechanisms of the assay and validating lignin deconstructing capabilities of species that were identified using this assay.

MetaBAT: Metagenome Binning based on Abundance and Tetranucleotide frequency

Dongwan D. Kang* (ddkang@lbl.gov)[1,2], Jeff Froula[1,2], Rob Egan[1,2], and Zhong Wang[1,2]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California

Grouping large fragments assembled from shotgun metagenomic sequences to deconvolute complex microbial communities, or metagenome binning, enables the study of individual organisms and their interactions. Here we developed automated metagenome binning software, called MetaBAT, which integrates empirical probabilistic distances of genome abundance and tetranucleotide frequency. On synthetic datasets MetaBAT on average achieves 98% precision and 90% recall at the strain level with 281 near complete unique genomes. Applying MetaBAT to a human gut microbiome data set we recovered 176 genome bins with 92% precision and 80% recall. Further analyses suggest MetaBAT is able to recover genome fragments missed in reference genomes up to 19%, while 53 genome bins are novel. In summary, we believe MetaBAT is a powerful tool to facilitate comprehensive understanding of complex microbial communities.

Single cell genomic study of *Dehalococcoidetes* species from deep-sea sediments of the Peruvian margin

Anne-Kristin Kaster* (annekristinkaster@gmail.com),[1] Koshlan Mayer-Blackwell,[2] Ben Pasarelli,[3] and Alfred M. Spormann[1,2]

[1]Department of Chemical Engineering, [2]Department of Civil and Environmental Engineering, [3]Department of Bioengineering, Stanford University, Stanford, California

The phylum *Chloroflexi* is one of the most frequently detected phyla in the subseafloor of the Pacific Ocean margins. Dehalogenating *Chloroflexi* (*Dehalococcoidetes*) were originally discovered as the key microorganisms mediating reductive dehalogenation via their key enzymes reductive dehalogenases

(Rdh) as sole mode of energy conservation in terrestrial environments. The frequent detection of *Dehalococcoidetes*-related 16*S* rRNA and *rdh* genes in the marine subsurface implies a role for dissimilatory dehalorespiration in this environment, however the two genes have never been linked to each other. In order to provide fundamental insights into the metabolism, genomic population structure, and evolution of marine subsurface *Dehalococcoidetes sp*., we analyzed a non-contaminated deep-sea sediment core sample from the Peruvian Margin Ocean Drilling Program (ODP) site 1230, collected 7.3 meters below the seafloor by a single cell genomic approach. We present for the first time single cell genomic data on three deep-sea *Chloroflexi* (Dsc) single cells from a marine sub-surface environment. Two of the single cells were considered to be part of a local *Dehalococcoidetes* population and assembled together into a 1.38 Mb genome, which appears to be at least 85% complete. Despite a high degree of sequence-level similarity between the shared proteins in the Dsc and terrestrial *Dehalococcoidetes,* no evidence for catabolic reductive dehalogenation was found in Dsc. The genome content is however consistent with a strictly anaerobic organotrophic or lithotrophic lifestyle.

A reference set of rumen microbial genomes: The Hungate1000 project

William J. Kelly* (bill.kelly@agresearch.co.nz),[1] Graeme T. Attwood,[1] Peter H. Janssen,[1] Adrian L. Cookson,[1] Gemma Henderson,[1] Suzanne C. Lambie,[1] Rechelle Perry,[1] Kenneth The,[1] Nikola Palevich,[1] Samantha Noel,[1] Lynne A. Goodwin,[2] Nicole Shapiro,[2] Tanja Woyke,[2] Christopher J. Creevey,[3] and Sinead C. Leahy[1]

[1]AgResearch Ltd., Grasslands Research Centre, New Zealand, [2]DOE Joint Genome Institute, Walnut Creek, California; [3]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, United Kingdom

The Hungate1000 project (www.Hungate1000.org.nz) seeks to produce a reference set of rumen microbial genome sequences from cultivated rumen bacteria and archaea, together with representative cultures of rumen anaerobic fungi and ciliate protozoa. The resulting data will be used for two main purposes; 1) to support international efforts to develop methane mitigation and rumen adaptation technologies and, 2) to initiate genome-based research aimed at understanding rumen function, feed conversion efficiency, methanogenesis and plant cell wall degradation in order to find a balance between food production and greenhouse gas emissions. It will underpin the analysis and comprehension of rumen metagenomic and metatranscriptomic sequence datasets and broaden our knowledge of rumen microbial phylogenetic diversity through genome sequencing of novel microorganisms.

Ruminants have evolved a symbiotic relationship with a complex microbiome located in their fore-stomach (reticulo-rumen) that allows these animals to use the lignocellulose component of plant material as their main energy source. Bacteria are the most numerous organisms in the rumen microbiome, being present at 10^{10} to 10^{11}/g of content and making up more than 50% of the cell mass. Studies of rumen bacteria have revealed a bacterial community that includes both generalists and specialists able to convert large plant polymers to a variety of smaller organic compounds that are used by the ruminant for growth.

To guide the selection of bacteria for sequencing through the Hungate1000, we performed a meta-analysis of rumen 16S rRNA-based studies overlaid with cultured representative data. We identified cultivated rumen bacteria belonging to 88 different bacterial genera while a survey of international culture collections gave 146 bacterial cultures of rumen origin. While these culture collection isolates

cover all the major taxonomic groups and include several well described organisms known to have key roles in rumen function, they do not represent the full diversity of the rumen microbiome. Consequently, studies focused on bringing additional rumen organisms into cultivation are important to this project.

The Hungate1000 project currently has genome sequencing in progress for more than 220 bacterial cultures, with sequence information already available for 90 cultures. These belong in 14 different bacterial families from five phyla and include several organisms that are not yet taxonomically described and only identified to the family level. When combined with genomes sequenced by other research groups ~120 genomes of rumen bacteria are now available. While not yet comprehensive, this collection will be a valuable resource for researchers interested in the microbiology of gut environments.

Comparative genomic analysis of *Megasphaera* sp. BS-4 and *Megasphaera elsdenii* ATCC 25940

Hyunjin Kim,[1] Byoung Seung Jeon* (jbstrust@gmail.com),[1] and Byoung-in Sang* (biosang@hanyang.ac.kr)[1,2]

[1]Department of Chemical Engineering, Hanyang University, Seoul, Republic of Korea; [2]Department of Fuel Cells and Hydrogen Technology, Hanyang University, Seoul, Republic of Korea

Megaspharera sp. BS-4 producing medium chain fatty acids was isolated from cow rumen. Fermentative products of the isolate were pentanoic acid, hexanoic acid, heptanoic acid, and octanoic acid. Interestingly, the isolate produced pentanoic acid or hexanoic acid up to 10 g/L using short chain fatty acid such as acetic acid, propionic acid, and butyric acid within 24 hours. When the 16s RNA of this strain was analyzed, this strain showed 96 % similarity with *Megasphaera elsdenii* ATCC 25940. As well, physiological characteristics related to carbon utilization and fatty acid production of the isolates displayed the difference with *M. elsdenii* ATCC 25940. In order to investigate features of isolates in genomic level, whole genome of the isolate was analyzed using pyrosequencing equipment, GS FLX 454. Genome size and total gene number of the isolates were 2,877,851 bp and 2,892 genes, respectively. All genes related to fatty acid production and carbon utilization of isolate were selected and metabolic pathways were postulated. Compared with genes related to fatty acid production of the isolate and *M. elsdenii* ATCC 25940, while *M. elsdenii* BS-4 had 4 acetyl CoA transferases related to synthesis of fatty acids and CoA molecules, the isolate had 8 acetyl CoA transferases. In addition to this, COG distribution (Clusters of Orthologous Groups) of the isolate and *M. elsdenii* ATCC 25940 was compared. While F(Nucleotide transport and metabolism), J(Translation, ribosomal structure and biogenesis), and Q(Secondary metabolites biosynthesis, transport and catabolism) of isolate were much higher than them of *M. elsdenii* ATCC 25940, H(Coenzyme transport and metabolism), I(Lipid transport and metabolism), L(Replication, recombination and repair), and P(Inorganic ion transport and metabolism) of *M. elsdenii* ATCC 25940 were higher than it of isolates.

Halophilic communities as a source for novel lignocellulolytic enzymes

Jeffrey A. Kimbrel* (jakimbrel@lbl.gov)[1,2], Nicholas Ballor[1,2], Thomas Ruegg[1,3], Hannah Woo[1,4], Terry C. Hazen[1,4], Michael P. Thelen[1,3], Blake A. Simmons[1,5], Steven W. Singer[1,6], and Janet K. Jansson[1,6]

[1]Deconstruction Division, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, California; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California; [4]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Biological and Materials Sciences Center, Sandia National Laboratories, Livermore, California; [6]Earth Sciences Division. Lawrence Berkeley National Laboratory, Berkeley, California

Lignocellulose presents a challenge to next generation biorefineries due to its recalcitrance to microbial degradation. Ionic liquid (IL)-based pretreatment has been successful in preparing biomass for enzyme saccharification, but the most common ILs used for pretreatment inhibit many downstream enzymatic and microbial processes derived from mesophiles. Certain halophiles are adapted to relevant salt enriched environments and are thus a potential source for IL-tolerant enzymes. Here we sought to discover & recover novel lignocellulolytic enzymes from environmental and feedstock-enriched halophilic bacterial communities. We sampled different saline environments from Puerto Rico and San Francisco; including salt flats, saltern ponds, turtle grass beds and sediments. For each of the environmental samples we obtained 16S rRNA gene sequences, metagenomes and metaproteomes. The data revealed an increase in relative abundance of haloarchaea and genes and proteins implicated in a hypersaline lifestyle with increasing salinity.

Samples from a turtle grass bed (3.5% salinity) and a high salinity saltern pond (33.2% salinity) were selected for enrichment on the potential biofuels feedstocks: miscanthus (M), eucalyptus (E) or pine (P) under aerobic and anaerobic conditions and followed through three 2-week passages. At the end of each passage cells were harvested, specific enzyme activities were measured and DNA, RNA and proteins were extracted. We found that enzyme activity was typically highest after the first passage, with the aerobic turtle grass enrichments having consistent activity on each feedstock. After the third passage, metagenomes were constructed and binned using MaxBin, a binning algorithm developed at JBEI. Bins were subsequently searched against the CAZY/dbCAN HMMs. In addition, expressed transcripts from 11 metatranscriptomes were identified by either alignment to the reference metagenomes or *de novo* assembled. To date we have identified over 1000 expressed candidate carbohydrate active enzymes from the enrichments and obtained reconstructed genomes for >100 feedstock-enriched archaea/bacteria. Heterologous expression of a diverse collection of 29 putative glycoside hydrolases is ongoing. The next step will be to validate and incorporate these candidate enzymes into a halophilic deconstruction enzyme mixture with high activity and IL-tolerance.

A pyrophosphate-dependent, more energy efficient version of the Calvin cycle

Manuel Kleiner* (mkleiner@mpi-bremen.de),[1] Arren Bar-Even,[2] Jillian M. Petersen,[1] and Nicole Dubilier[1]

[1]Max Planck Institute for Marine Microbiology, Department of Symbiosis, Bremen, Germany; [2]Department of Plant Sciences, The Weizmann Institute of Science, Rehovot, Israel

The Calvin-Benson-Bassham (CBB) cycle is the metabolic pathway used by plants, algae and a variety of microorganisms to fix CO_2 into biomass and thus it is responsible for the majority of biological CO_2 sequestration from the atmosphere. A recent metaproteomic study of autotrophic symbionts in a marine worm led us to propose that these symbionts use a modified version of the CBB cycle that is 10 to 30 % more energy efficient than the classical version of the CBB cycle. In this modified cycle, energy is conserved through the tight coupling of a multifunctional, pyrophosphate-dependent enzyme (PPi-PFK) and a membrane bound proton-translocating pyrophosphatase (HPPase). Using comparative genomics to probe for the presence of this PPi-dependent CBB cycle in other bacteria we found that more than 14 other symbiotic and free-living bacteria potentially possess this pathway and that it may have evolved as an adaptation to habitats that are in some way energy limited. Additionally, our comparative genomic analyses revealed that a large range of bacteria with the CBB cycle lack two or three of the classical enzymes for this pathway suggesting that modifications of the CBB cycle are widespread.

Thermodynamical modeling showed that the PPi-dependent CBB cycle would function well under normal physiological conditions within the cell and that its rate would be only minimally reduced as compared to the classical version of the CBB cycle. On the long term bioengineering may allow to transfer the PPi-dependent CBB cycle to other microorganisms such as cyanobacteria to increase yields in the production of biofuels and other CO_2 neutral, renewable resources.

Assembly of repeat content using next generation sequencing data

K. LaButti* (klabutti@lbl.gov), A. Copeland, and A. Kuo

Lawrence Berkeley National Laboratory, Berkeley, California; DOE Joint Genome Institute, Walnut Creek, California

Short read technology has enabled efficient and high quality assembly of prokaryotic genomes. However, highly repetitive organisms still pose a challenge for short read assembly, and typically only unique regions and repeat regions shorter than the read length, can be accurately assembled. Recently, we have been investigating the use of Pacific Biosciences long reads for de novo fungal assembly. We will present an assessment of the quality and degree of repeat reconstruction possible in a fungal genome using long read technology. We will also compare differences in assembly of repeat content using short read and long read technology.

New frontiers of genome assembly with SPAdes 3.0

Alla Lapidus* (piterlabs@gmail.com),[1,2] Dmitry Antipov,[1] Anton Bankevich,[1] Alexey Gurevich,[1] Anton Korobeynikov,[1,2] Sergey Nurk,[1] Andrey D. Prjibelski,[1] Yana Safonova,[1] Irina Vasilinetc,[1] and **Pavel Pevzner**[1,3,3]

[1]Algorithmic Biology Lab, Saint Petersburg Academic University of RAS, Saint Petersburg, Russia; [2]Saint Petersburg State University, Saint Petersburg, Russia; [3]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California

Despite all the efforts high quality genome assembly is a complex task that so far remains unsolved. It is very well known that majority of problems caused by repeats present in all genomes of any nature. The usage of multiple methods of genomic DNA (gDNA) isolation, different sequencing technologies and different types of genomic libraries for research projects introduces additional levels of complication to the genome assembly. The assembler tool SPAdes was originally developed at the St Petersburg Academic University (St. Petersburg, Russia) for the purpose of overcoming the complications associated with single-cell microbial data (uneven coverage, increased level of errors and chimerical reads). The tool was able to successfully resolve these issues for Illumina reads and was recognized by the scientific community as one of the best assemblers working with both isolates and single-cell data. Even though the assembler was specifically designed to work solely with microbial genomes, scientists have tested the tool on a large number of different types of data produced with a variety of different sequencing platforms. Their efforts and feedback have inspired us to extend the capabilities of SPAdes to include additional platforms (Ion Torrent, PacBio, Sanger), combinations of platforms, and to work with both paired end and mate pair libraries of different insert sizes. The latest release of SPAdes provides the following new features: IonTorrent-specific error correction and assembly; hybrid assemblies including the combination of Illumina/IonTorrent with PacBio (or other long reads technologies) data and, finally, special algorithm tailored for assembly of polymorphic diploid genomes producing the consensus contigs of haplomes.

Activation tagging in poplar for gene function discovery

Arianna Latini,[1] Denis Lachance,[1] Marie-Josée Morency,[1] Gervais Pelletier,[1] Don Stewart,[1] Lee Gunter,[2] Gerald A. Tuskan,[2] Jin-Gui Chen,[2] Sharon Regan,[3] and **Armand Seguin*** (armand.seguin@nrcan.gc.ca)[1]

[1]Laurentian Forestry Centre, Natural Resources Canada, Québec, Canada; [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Department of Biology, Queen's University, Kingston, Canada

Activation tagging is an approach by which a T-DNA fragment containing an engineered tetramer of the CaMV35S enhancer randomly integrates into the plant genome to activate the transcription of flanking genes. A total of 1164 activation tagged lines of hybrid *Populus* (*P. tremula x P. alba*) were evaluated at the Laurentian Forestry Centre research station starting in August 2006. The trees were grown for four years to monitor appearance of specific phenotypes. Close to 100 mutants were identified and a large proportion showed dwarfism. However, many trees were lost during the second growth season, leaving 14 mutants for further analyses. Another group of about 20 mutants showed early or late bud break, winter marcescent foliage and change in tree or leaf architecture phenotypes. In December 2010, trees were cut and 727 stem sections were collected for wood chemistry analysis. A total of 18 lines were selected (from 400 analysed) based on important variation of cell wall chemistry (e.g. lignin percentage, S/G ratio).

Identification of T-DNA insertions in the poplar genomes was achieved and genes located within the 15 kb region flanking the T-DNA insertion site were selected for further investigations. For each mutant we gathered information on gene annotation and expression profiling (database mining and our own quantitative RT-PCR analysis) on the selected candidate genes. Our analysis already revealed several potentially interesting proteins related to wood chemistry (e.g. glucanase, hydrolase, F-box protein and HRGP) and bud development (e.g. bHLH transcription factors and protein kinases) for instance. Association of other candidate genes, (e.g. stress related genes; wound-inducible, dehydrin) with specific mutant phenotypes such as tree architecture and dwarfism could however be more challenging. We are pursuing further characterization of those genes and related proteins to improve our understanding of their roles in key physiological processes.

Acknowledgment: This work is supported by the Genomics R & D Initiative of Canada and the Arborea project (www.arborea.ca funded by Genome Canada and Genome Quebec). Support from the United States Department of Energy (DOE) BioEnergy Science Center project is also acknowledged. The BioEnergy Science Center is a Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

Towards Raman-activated cell sorting (RACS) for single bacteria genomics

Kang Soo Lee* (ks_lee@mit.edu),[1] Márton Palatinszky,[2] Michael Wagner,[2] and Roman Stocker[1]

[1]Ralph M. Parsons Laboratory, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; [2]Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, University of Vienna, Wien, Austria

Raman microspectroscopy is a non-destructive technique for chemical measurements that has long been used on solids, liquids, and gases. In recent years, Raman approaches have been increasingly used to measure the chemical composition of microbial cells, largely because there is no need for pretreatment (e.g., fluorescence labeling), because water does not interfere with the acquisition of Raman spectra, and because Raman spectra show sharp fingerprints. Here we report on our work to couple Raman measurements with microfluidics, aimed at developing a Raman-activated cell sorting platform. Our goal is the high-throughput sorting of microbial cells based on functional traits of interest, for subsequent single-cell genomic analysis. Bacteria labeled with deuterated water are employed as the target for sorting, based on the easily detectable signiture peak of the carbon-deuterium bonds of their lipids. To minimize background signal, we (i) used quartz coverslips as the bottom of our microchannels, (ii) located the measurement volume near the quartz bottom, and (iii) ensured that cells would travel through the measurement volume by a three-dimensional hydrodynamic focusing method, which was optimized with the help of Computational Fluid Dynamics (CFD) simulations. Our next steps will consist in implementing a cell-displacement approach, which we envisage will be based on optical forces, and then activate it based on desired properties of the Raman spectrum analyzed in real time. Ultimately, this system will make it possible to directly connect phenotypic with genotypic characteristics of individual cells, and the microfluidic engineering discussed here will be a fundamental step in enabling this unprecedented level of resolution in microbial ecology.

This work is supported by the Emerging Technologies Opportunity Program (RFP No. ETOP2013) of the US Department of Energy Joint Genome Institute (DOE-JGI).

Genome-wide analysis of mutations in rice mutants generated by fast-neutron irradiation to study cell wall recalcitrance

Guotian Li* (gtli@ucdavis.edu),[1,2] **Mawsheng Chern*** (mschern@ucdavis.edu)[1,2] Rashmi Jain,[1,2], Liangrong Jiang,[2] Miguel Vega-Sanchez,[1,2] Patrick Canlas,[1,2] Wendy Schackwitz,[3] Joel Martin,[3] Nicholas Santoro,[4] and Pamela Ronald[1,2]

[1]Feedstocks Division, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, California; [2]Department of Plant Pathology and the Genome Center, University of California at Davis, Davis, Califoria; [3]U.S. Department of Energy Joint Genome Institute, Walnut Creek, California; [4]Department of Energy Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan

As cost of sequencing technology becomes more affordable, large-scale comparative analysis of mutants and wild-type genomes to quickly identify genes corresponding to mutant phenotypes has become feasible. In this study, we used whole-genome re-sequencing to analyze genome-wide mutations in 45 rice mutants that were identified from a screen of 5,000 rice lines for alterations in cell wall recalcitrance.

These 45 mutant lines were re-sequenced with Illumina HiSeq2500 at the Joint Genome Institute (JGI) as part of a project aimed at sequencing 2,000 mutants from a fast-neutron mutagenized rice mutant population in a model rice variety, Kitaake. On average, aligned reads for each mutant line are equal to 49-fold sequence coverage of the genome. From this analysis we detected 215,000 single nucleotide polymorphisms (SNP) and 38,000 small insertions or deletions (INDEL) between the genomes of Kitaake and the reference Nipponbare. In these 45 re-sequenced lines, 300 changes are predicted to affect gene/protein function when compared to the Kitaake parent. Among these, 69% are caused by INDELs and 31% belong to SNPs. On average, there are seven mutations in each M_3 mutant line (two generations after the seeds were irradiated). In summary, this study provides a comprehensive genome-wide examination of the effects caused by fast-neutron irradiation in rice genome. This whole-genome sequence approach provides a cost-efficient approach to identify novel saccharification-related genes in rice. The knowledge gained from this research will eventually facilitate the development of strategies for cost efficient conversion of lignocellulosic biomass into next-generation biofuels.

The Global Peatland Microbiome Project: Concept and initial results from PEATcosm

Erik Lilleskov* (elilleskov@fs.fed.us)[1], L. Jamie Lamit[2], Karl Romanowicz[2], Susannah Tringe[3], Evan Kane[2], Rod Chimner[2], Jay Lennon[4], Chris Schadt[5], Mark Waldrop[6], and Nathan Basiliko[7]

[1]U.S. Forest Service, Northern Research Station, Houghton, Michigan; [2]Michigan Technological University, Houghton, Michigan; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Indiana University, Bloomington, Indiana; [5]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [6]USGS, Menlo Park, California; [7]Laurentian University, Sudbury, Ontario, Canada

While peatland ecosystems represent only ~3% of the land surface, they sequester up to 30% of soil organic carbon (C). This large pool of C is vulnerable to climate change. The balance of methane and carbon dioxide produced by peatland systems has potential to dramatically alter global atmospheric greenhouse gas and temperature stabilization points. Despite this important position in the biosphere much less is known about microbial communities and their functions in peatlands compared to most other terrestrial and ocean habitats. Our goal is two-fold: to create a resource for understanding the

large-scale patterns of microbial (here defined as bacterial, archaeal, and fungal) community structure and function in peatland ecosystems, and to understand how these communities regulate C cycling in response to environmental change. Our project has three parts: 1) a global survey of microbial communities in peatland ecosystems; 2) metagenomics and metatranscriptomics in two peatland climate change experiments (SPRUCE and PEATcosm); and 3) single cell genomics for species of interest. To address our first objective we have developed a consortium of scientists and study sites from around the world to leverage existing global change experiments and natural gradients to characterize these microbial communities. As a proof of concept for our extraction, sequencing and analytical protocols, and to provide insights into regulation of bog microbial communities, we analyzed the microbial communities in the PEATcosm experiment, a large mesocosm experiment manipulating water table and plant functional groups (sedges, Ericaceae) to understand the consequences for *Sphagnum* bog carbon cycling. Two months after vegetation manipulation only was initiated we isolated and analyzed microbial communities from two depths from all replicates. We found that there was a strong and significant effect of depth on microbial communities, with weaker effects of plant community. In the Bacteria, Acidobacteria dominated at both depths and increased under sedges, and Bacteroidetes declined with depth and decreased under sedges. In the Archaea, methanobacteriales and Methanomicrobiales declined with depth, and unclassified Euryarchaeota increased. In the Fungi, Helotiales dominated at both depths, whereas Leotiales declined with depth. There was clear depth differentiation among different putative *Rhizoscyphus* OTUs. Which represent ericoid symbionts vs. free-living heterotrophs awaits further investigation. We hypothesize that significant community inertia exists because of persistent residual roots and dormant microbial propagules. We will reanalyze this experiment this summer after four years of vegetation treatments and three years of water table treatments, and hypothesize we will find a much greater separation, with a strong effect of the divergent symbiomes of Ericaceae and sedges leading to clearly differentiated fungal and prokaryote communities. These methods will also be applied to the Global Microbiome Project beginning this summer.

Applying functional metagenomics towards the development of an ethanologenic CBP microorganism

Inés Loaces,[1] Vanesa Amarelle,[1] Elena Fabiano,[1] Lonnie Ingram,[2] Alfredo Martínez-Jiménez,[3] and Francisco Noya* (fnoya@iibce.edu.uy)[1]

[1]Departamento de Bioquímica y Genómica Microbianas, Instituto de Investigaciones Biológicas "Clemente Estable", Montevideo, Uruguay; [2]Florida Center for Renewable Chemicals and Fuels, Department of Microbiology and Cell Science, University of Florida, Gainesville, Florida; [3]Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico.

Lignocellulose is a promising feedstock for ethanol production. The conversion involves three mayor steps: pretreatment, saccharification and fermentation of released sugars. Several enzymes are involved in the saccharification step such as glucanases, glycosidases and xylanases. Nowadays, these enzymes have to be added to the process and are usually produced and purified by a dedicated supplier. Consolidated bioprocess (CBP) is the name given to an ideal process in which one microorganism is able to produce these enzymes and ferment the sugars in the same bioreactor. In this work, we applied functional metagenomics to screen for enzymes encoding genes and regulatory elements that could confer to industrial *E. coli* strains the ability to grow on cellulose as the sole carbon source while producing ethanol in the process.

Two fosmid libraries of 30.000 and 100.000 clones were constructed from metagenomic DNA isolated from the microorganisms inhabiting the rumen of Uruguayan cows and the sludge of an anaerobic digester, respectively. The library was screened for the expression of cellulases and xylanases using carboxymethyl-cellulose (CMC), avicel and xylan from birch wood as substrates. Twenty seven clones were positive for cellulolytic activity, 11 showed xylanase activity while 11 more showed both activities.

Some of these clones conferred *E. coli* the ability to grow on minimal media supplemented with cellobiose, filter paper (FP) or sugarcane bagasse as sole carbon sources. It was also found that the cellulolytic activity was inhibited by glucose and xylose, but not by cellobiose or avicel. All of these fosmids were transferred into the ethanologenic *E. coli* strains LY180 and MS04, and ethanol production was confirmed under anaerobic conditions using CMC as the sole carbon source.

Full sequencing was performed on a subset of these fosmids. Some of them showed no obvious genes for glycolytic enzymes while other contained full operons that combined endoglucanases and xylosidases with other enzymes also involved in lignocellulose deconstruction such as laccases.

Current work is aimed at the detailed biochemical characterization of these genes as well as at their optimized expression in the ethanologenic strains.'

FASTERp: A feature array search tool for estimating resemblance of protein sequences

Derek Macklin* (dmacklin@lbl.gov), Rob Egan, and Zhong Wang

DOE Joint Genome Institute, Walnut Creek, California

Metagenome sequencing efforts have provided a large pool of billions of genes for identifying enzymes with desirable biochemical traits. However, homology search with billions of genes in a rapidly growing database has become increasingly computationally impractical. Here we present our pilot efforts to develop a novel alignment-free algorithm for homology search. Specifically, we represent individual proteins as feature vectors that denote the presence or absence of short kmers in the protein sequence. Similarity between feature vectors is then computed using the Tanimoto score, a distance metric that can be rapidly computed on bit string representations of feature vectors. Preliminary results indicate good correlation with optimal alignment algorithms (Spearman *r* of 0.87, ~1,000,000 proteins from Pfam), as well as with heuristic algorithms such as BLAST (Spearman *r* of 0.86, ~1,000,000 proteins). Furthermore, a prototype of FASTERp implemented in Python runs approximately four times faster than BLAST on a small scale dataset (~1000 proteins). We are optimizing and scaling to improve FASTERp to enable rapid homology searches against billion-protein databases, thereby enabling more comprehensive gene annotation efforts.

Analysis capabilities for microbial communities in KBase

Folker Meyer* (folker@anl.gov),[1] Dylan Chivian,[2] Jason Baumohl,[2] Aaron Best,[3] Jared Bischof,[1] Ben Bowen,[2] Thomas Brettin,[1] Tom Brown,[1] Shane Canon,[1] Stephen Chan,[2] John-Marc Chandonia,[2] Ric Colasanti,[1] Neal Conrad,[1] Brian Davison,[4] Matt DeJongh,[3] Paramvir Dehal,[2] Narayan Desai,[1] Scott Devoid,[1] Terry Disz,[1] Meghan Drake,[4] Janaka Edirisinghe,[1] Gang Fang,[7] José Pedro Lopes Faria,[1] Mark Gerstein,[7] Elizabeth M. Glass,[1] Annette Greiner,[2] Dan Gunter,[2] James Gurtowski,[6] Nomi Harris,[2] Travis Harrison,[1] Fei He,[5] Matt Henderson,[2] Chris Henry,[1] Adina Howe,[1] Marcin Joachimiak,[2] Kevin Keegan,[1] Keith Keller,[2] Guruprasad Kora,[4] Sunita Kumari,[6] Miriam Land,[4] Steve Moulton,[4] Pavel Novichkov,[2] Taeyun Oh,[8] Gary Olsen,[9] Bob Olson,[1] Dan Olson,[1] Ross Overbeek,[1] Tobias Paczian,[1] Bruce Parrello,[1] Shiran Pasternak,[6] Sarah Poon,[2] Gavin Price,[2] Srivdya Ramakrishnan,[6] Priya Ranjan,[4] Bill Riehl,[2] Pamela Ronald,[8] Michael Schatz,[6] Lynn Schriml,[10] Sam Seaver,[1] Michael W. Sneddon,[2] Roman Sutormin,[2] Mustafa Syed,[4] James Thomason,[6] Nathan Tintle,[3] Will Trimble,[1] Daifeng Wang,[7] Doreen Ware,[5,6] David Weston,[4] Andreas Wilke,[1] Fangfang Xia,[1] Shinjae Yoo,[5] Dantong Yu,[5] Robert Cottingham,[4] Sergei Maslov,[5] Rick Stevens,[1] and Adam P. Arkin[2]

[1]Argonne National Laboratory, Argonne, Illinois; [2]Lawrence Berkeley National Laboratory, Berkeley, California; [3]Hope College, Holland, Michigan; [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Brookhaven National Laboratory, Upton, New York; [6]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

KBase provides the infrastructure and tooling for in-depth metagenome analysis, facilitating the annotation of microbial communities and the quest for identification of key players in a microbial community or the identification of trends. By automatically transforming microbial communities into abundance profiles, KBase is enabling users to drill down so that trends, specific taxa or functions can be identified. Combining metagenomic and environmental data makes it possible to correlate information about organism or function abundance to metadata that describe a variety of biologically intriguing characteristics of the samples, such as the biome the samples were collected from, the pH of the samples, etc.

Metabolic modeling can help elucidate the roles played by individual taxa in microbial communities by providing a detailed characterization of their functional repertoire. Once obtained this knowledge can be used for a number of purposes, such as the prediction of cultivation conditions for functionally important taxa. While the current state of modeling and our ability to annotate microorganisms has advanced greatly in recent years, the emerging models should be viewed as a first approximation rather than the final answer to these questions.

With KBase the functionality is in place to perform comparisons of multiple strategies for deriving metabolic models from microbial community data. Those strategies include the use of PCR primer amplified ribosomal genes as a reporter for the organisms informing PICRUSt predictions, the use of shotgun metagenomics to obtain functional information, the use of EMIRGE to extract complete ribosomal sequences from shotgun metagenomics sequences and the use of taxonomic information obtained from metagenomic sequences to inform PiCRUST predictions.

Structural genomics of sedimentary Archaea in postgenomic era

Karolina Michalska[1,2], Andrew Steen[3], Gekleng Chhor[1], Katlyn Fayman[1], Michael Endres[1], Gyorgy Babnigg[1], Karen Lloyd[3], Robert Jedrzejczak[1], and Andrzej Joachimiak* (andrzejj@anl.gov)[1,2]

[1]Midwest Center for Structural Genomics, [2]Structural Biology Center, Biosciences Division, Argonne National Laboratory; [3]Department of Microbiology, University of Tennessee

Traditional microbiology has established a foundation for studies of microorganisms easily accessible to pioneer experimenters. Therefore, the developed laboratory cultivation techniques are suitable for microbes, mostly bacteria, originating predominantly from terrestrial ecosystems with moderate temperature ranges. With the development of modern molecular biology methods, these organisms have also served as model systems to investigate the roles of individual proteins and RNAs encoded in their genomes, which eventually linked their sequences to specific functions. Recently, it has become evident that these well-understood prokaryotes represent only a minute fraction of the Earth's microbiome. Functional annotation of Microbial Dark Matter sequences derived from metagenomics and single-cell genome sequencing projects is primarily generated through similarity with characterized proteins. However, given the size discrepancy between the "known" and "unknown," the available "reference library" of sequences and biochemical functions is not sufficiently representative. Database-driven constraints may hamper our ability to discover novel functions in uncultivated microorganisms from thus far unexplored, and often extreme, environments. Nevertheless, information about microbial activities in these ecosystems is crucial for understanding the circulation of nutrients on a global scale. Significant contribution to biogeochemical cycles comes from marine sediments. This huge biomass reservoir is inhabited by recently discovered heterotrophic Archaea, which encode putative extra- and intra-cellular proteases that presumably facilitate utilization of proteinaceous organic matter deposits. To validate this sequence-based hypothesis we have undertaken a structural genomics approach to investigate the structure and function of a recombinant, cytosolic protease named AP TA1 from *Thaumarchaeota archaeon* SCGC AB-539-E09 phylogenetically assigned to the miscellaneous crenarchaeotal group (MCG) of Archaea. The protein is a homotetrameric, self-compartmentalizing enzyme related to bacterial α-amino acid ester hydrolases/transferases (AEH). It shows α-aminopeptidase activity towards dipeptides with a preference for a small, L-configured hydrophobic residue at the N-terminus, including aromatic amino acids. Thus, AP TA1 may be involved in degradation of peptides generated by extracellular proteases and subsequently imported to the cell. Close homology to AEHs does not enable us to preclude the synthetic function of AP TA1, however different substrate specificity and structural features of the enzyme indicate that a likely secondary metabolite is different from AEH products.

This work was supported by the following funds: National Institutes of Health grant GM094585 (AJ), the U. S. Department of Energy, Office of Biological and Environmental Research, under contract DE-AC02-06CH11357 (AJ), NSF Center for Dark Energy Biosphere Investigations grant NSF-157595 (KGL) and the U.S. Department of Energy contract DE-AC02-06CH11357.

Genomic investigation of long-term effects of forest harvesting on soil microbial communities

William W. Mohn* (wmohn@mail.ubc.ca)**,** Erick Cardenas Poire, Kendra R. Mass, Roland Wilhelm, Hilary Leung, and Steven J. Hallam

Department of Microbiology & Immunology, Life Sciences Institute, University of British Columbia

Forest soil microbial communities perform critical ecosystem functions. Improved understanding of these communities is needed to enable forest management to ensure sustainability and to promote carbon sequestration. We are employing multiple genomic approaches (1) to characterize forest soil communities across North America, (2) to understand how forest harvesting affects those communities and their capacity to decompose biomass and (3) to discover novel enzymes with applications in biomass conversion. Samples were collected at 18 sites in 6 ecozones, 10 to 15 years after experimental treatments testing four levels of organic matter removal during forest harvesting. Pyrotag analysis of 740 samples yielded 6.5M bacterial and 5.8M fungal tags. Variability in community structure was largely explained by ecozone:, bacterial variability, 48%; fungal variability, 74%. Within ecozones, harvesting had very inconsistent effects on community structure, ranging from no detectable effect to explaining 10% of fungal community variability. From one site in British Columbia where harvesting altered community structure substantially, 21 shotgun metagenomes totaling 71 Gb were generated. Harvesting reduced the metagenomic potential for biomass degradation, decreasing the overall abundance of the carbohydrate active enzyme (CAZY) and fungal oxidation of lignin enzyme (FOLY) genes by 16%. Harvesting substantially altered profiles based on relative abundances of CAZY and FOLY gene families and depleted gene families associated with unharvested forest. Preliminary analyses of soil metagenomes from additional ecosystems confirm that effects of harvesting on community structure are accompanied by changes in potential for biomass decomposition and that these effects differ among ecozones. From the metagenomes, we also identified genes encoding novel variants of enzymes with potential for commercially useful transformations of lignin. Stable isotope probing was employed to investigate bacterial and fungal populations putatively responsible for hemicellulose and cellulose degradation. Profiles of ^{13}C-enriched phospholipid fatty acids, indicate that forest harvesting alters populations of hemicellulose- and cellulose-degrading bacteria and fungi. Based on distributions of pyrotag OTUs, hemicellulose- and cellulose-degrading bacterial and fungal populations were disproportionality altered by harvesting. However, effects of harvesting on these populations again differed substantially among ecozones. Overall, our results demonstrate that forest harvesting can have significant long-term impacts, altering the structure of the soil community and reducing the potential of that community for biomass decomposition. However, these impacts vary greatly among ecozones, and it is necessary to further investigate how these impacts are affected by edaphic and climactic factors. Soil communities appear to be sensitive indicators of the effects of disturbance on forest ecosystems.

Finishing the *Thermus aquaticus* genome

Scott A. Monsma* (smonsma@lucigen.com), Michael Lodes, Svetlana Jasinovica, Erin Ferguson, David Mead, and **Phillip Brumm**

Lucigen Corporation, Middleton, Wisconsin

The discovery of *Thermus aquaticus* by Brock 35 years ago launched the modern era of molecular biology and microbial genomics. Technical limitations at the time prevented closing the genome and determining the presence or absence of plasmids in *Thermus aquaticus* Y51MC23, resulting in a permanent

draft genome (ACCESSION NZ_ABVK00000000) containing 22 unordered contigs. Lucigen's long range, clone-free mate pair library construction technology was used to assemble the *Thermus aquaticus* genome into a single scaffold plus three contigs left to close. A 350 bp insert fragment library was sequenced using Illumina MiSeq 2 x 300 chemistry. Mate pair libraries of *T. aquaticus* random-sheared genomic DNA were prepared with size-selected 10kb and 20kb fragments using Lucigen's NxMate protocol. Initial assembly of 5.78 million paired-end fragment reads (average read length 237 nt, average coverage ~600x) was performed with Ray version 2.3.0, resulting in 54 contigs of length >500 bp and N50 of 79,749 bp, and total length of 2.26 Mbp. The 54 contigs were then assembled along with 303,709 NxMate™ mate-pair reads (10kb and 20kb libraries) by DNAStar SeqMan NGen resulting in a single scaffold of length 2,161,286 bp, along with 3 unconnected contigs of 5614, 17,991 and 72,959 bp. Further analysis of the 3 unconnected contigs revealed overlapping ends resulting in circular sizes of 5,719, 14,447 and 71,044 bp. Analysis of the proteins encoded by the circular contigs found repA, transposase, helicase, DNA methylase, and other DNA modification/repair enzymes to be present. Additionally, the largest circular contig encodes peptidases, lytic transglycosylase, ATPase AAA, plasmid partioning proteins ParA and ParB, and TraC. Therefore we propose that these three contigs are plasmids independently replicating in the *Thermus aquaticus* host cells. Annotation of the genome will be discussed as well as evidence for the closure of this genome.

High efficiency long insert mate pair library preparation for NGS platforms

Scott Monsma* (smonsma@lucigen.com), David Mead, Svetlana Jasinovica, Erin Ferguson, and **Michael J. Lodes**

Lucigen Corporation, Middleton, Wisconsin

Next generation DNA sequencing (NGS) instruments produce gigabases per run, but the short read lengths and small size of sequenced fragments result in gaps, misassembled contigs, collapsed repeats and missing sequences, leaving these regions to be finished manually, if at all. A technology that provides long range sequence linkage from short reads is needed for accurate, economical *de novo* assembly of genomes. We developed a >90% efficient mate pair library construction technology that incorporates Chimera Codes™ to distinguish true mate pairs from false junctions. NxMate™ NGS libraries were constructed using a reference *E. coli* strain, *Thermus aquaticus* and two repeat rich mouse BACs. Without mate pair libraries the BAC and genome assemblies contained numerous unordered contigs. The addition of >90% efficient NxMate data allowed accurate de novo assembly and closing of the BACs and genomes. For comparison, several Illumina Nextera mate pair libraries were constructed and yielded <5% mate pair data. Ion Torrent "mate pair reads without paired end sequencing" permits economical sequence assembly of BAC clones and small genomes, while NxMate libraries for the Illumina platform allows for long range paired end sequencing and assembly of larger genomes.

Genome re-sequencing reveals a species-specific whole-gene deletion associated with *Populus-Laccaria* mycorrhizal symbiosis

Wellington Muchero* (mucherow@ornl.gov)[1], Juan Wang[1], Jessy Labbe[1], Olaf Czarnecki[1], Jin-Gui Chen[1], Priya Ranjan[2], Xiaohan Yang[1], Luke Evans[3], Stephen DiFazio[3], Anthony Bryan[1], Sara Jawdy[1], Lee Gunter[1], Wendy Schackwitz[4], Joel Martin[4], Christopher W. Schadt[1], and Gerald A. Tuskan[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]University of Tennessee; Knoxville, Tennessee; [3]West Virginia University; Morgantown, West Virginia; [4]DOE Joint Genome Institute; Walnut Creek, California

Mycorrhizal symbiosis between perennial plants and fungal associates has critical implications for diverse phenomena including global carbon, water and nutrient cycling, as well as agricultural and forestry productivity with limited inputs on marginal croplands. As such, characterizing the molecular genetics underlying such interactions holds tremendous potential in engineering biological systems for enhanced carbon sequestration and sustainable biomass production.

In the perennial bioenergy feedstock *Populus*, numerous studies have demonstrated the species-dependent colonization efficiency by the fungal symbiont *Laccaria bicolor*, with *P. trichocarpa* exhibiting high levels of mycorrhization compared to *P. deltoides*. This highly species-specific attribute of the interaction presents an opportune platform for the discovery of host genetic factors governing mycorrhizal interactions using inter-specific hybrids. To this end, we identified a major quantitative trait locus (QTL) contributing up to 60% of the phenotypic variance explained (PVE) in colonization of *P. trichocarpa* x *P. deltoides* F_1 hybrids by *Laccaria.* Genome anchoring of this QTL using single nucleotide polymorphism (SNP) markers with known physical positions revealed its co-location with a region harboring tandemly repeated lectin-type receptor kinases. Alignment of the *P. trichocarpa* and *P. deltoides* re-sequenced parental genomes suggested major structural differences in this region including a whole-gene deletion event in *P. deltoides* involving a d-mannose lectin receptor kinase. Analysis of allelic effects of the indel revealed that individuals carrying a full copy of the gene exhibited 2X more colonization by the fungal symbiont compared with individuals missing segments of the same gene. Further, we screened pure *P. trichocarpa* and *P. deltoides* natural variants to assess penetrance of the indel in the species' natural habitats. We could not detect a full copy of the gene in any of the 60 *P. deltoides* genotypes collected from diverse geographical origins in eastern United States whereas the gene was highly conserved in 673 re-sequenced *P. trichocarpa* genomes evaluated. Since d-mannose receptor kinases have been implicated in innate immunity and self-incompatibility responses which require highly specific recognition of cells and microorganisms, we hypothesize that this indel polymorphisms contributes substantially to the species-specificity observed in *Populus* interaction with *Laccaria*. Transgenic validation of putative effects of the receptor kinase on mycorrhization is currently underway and results of these analyses will be discussed.

Sequencing and assembly of DNA from "Aigarchaeota" single cells using Pacific Biosciences, Illumina and 454 sequencing platforms, and integration with metagenomic analyses

Senthil K. Murugapiran* (senthil.murugapiran@unlv.edu),[1] Jeremy A. Dodsworth,[1] Iwijn DeVlaminck,[2] Paul C. Blainey,[2,3] Jad Kanbar,[2] Charles de Bourcy,[2] Christian Rinke,[4] Tanja Woyke,[4] Stephen R. Quake,[2] and Brian P. Hedlund[1]

[1]School of Life Sciences, University of Nevada, Las Vegas, Nevada; [2]Department of Bioengineering, Stanford University, and Howard Hughes Medical Institute, Stanford, California; [3]Current address: Massachusetts Institute of Technology, Department of Biological Engineering, Cambridge, Massachusetts; [4]DOE Joint Genome Institute, Walnut Creek, California

Sequencing of 16S rRNA gene tags and metagenomes from microbial populations in geothermal environments in China and the U.S. revealed the presence of several lineages within the "Aigarchaeota", a globally-distributed candidate phylum of Archaea with no cultivated representatives. Recent advances in single-cell genomics techniques have expanded our ability to access the genomes of Bacteria and Archaea that are recalcitrant to laboratory cultivation. Because of its novelty, however, robust bioinformatic pipelines for assembling single-cell genomes using data from a variety of next-generation sequencing techniques, and for synergistic analysis of single-cell and metagenome datasets, are still under development. We assessed the advantages of using PacBio single molecule, real-time (SMRT) continuous long read (CLR) data to complement Illumina and 454 short reads in single-cell genomics studies, to better understand "Aigarchaeota". An optifluidic technique enabled morphology-based isolation and genome amplification of single "Aigarchaeota" cells from sediments. Amplified genomic DNA of seven "Aigarchaeota" cells was sequenced using Roche 454 FLX Titanium, Illumina MiSeq (2 x 250 paired-end) and PacBio methodologies. Short read sequences from individual single amplified genomes were assembled using Velvet (v1.2.10), SPAdes (v2.5.1), GSAssembler (v2.6). The average nucleotide identity (ANI%) between the seven individual "Aigarchaeota" genomes was determined to be > 98% by BLASTN comparison, hence all short reads and long reads were combined and PacBio sequencing oriented assembly methods (HGAP, AHA and ALLORA) were tested in addition to short-read assemblers. The assembly statistics of the combined Illumina reads from the seven sorted cells were significantly better in the case of SPAdes compared to Velvet, yielding a genome of ~3 Mb. The genome coverage was estimated to be 96% based on 162 conserved, single-copy phylogenetic markers identified in >90% sequenced Archaea (141 of 162 markers identified, with 9 present in >1 copy). Furthermore, the SPAdes assembly that used either Illumina or both Illumina and 454 data was further improved when PacBio long-reads were used to scaffold SPAdes contigs using AHA, resulting in a somewhat more contiguous assembly (increased N50, average contig length). However, the size of the assembly as well as the number of single-copy marker remained unchanged in the AHA assembly. These single-cell genomes were used to inform binning of metagenomes from LHC and other geothermal springs using various methods including Metawatt, emergent self-organizing maps (ESOM) and TNF-PCA, which led to the identification of related lineages of "Aigarchaeota". In conclusion, incorporation of long-read PacBio sequence data improved single-cell genome assembly statistics marginally and, in comparison to either technique on its own, the combined application of single-cell and metagenomics promises to yield a more complete view of the phylogeny, diversity, and metabolic potential of "Aigarcheota".

Building a genome-scale technology platform for predictive design and optimization of bacterial systems

Vivek K. Mutalik* (vkmutalik@lbl.gov),[1] Emily F. Freed,[2] Ryan T. Gill,[2] and Adam Arkin[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, California; [2]University of Colorado, Boulder, Colorado

The ability to comprehensively understand, predict, and modify biological systems at the genome-scale requires the development of new technologies for the iterative search and optimization of mutational-design space on laboratory timescales. One critical barrier to achieve desired traits is the huge combinatorial search space between genetic parts, conditions and genome level modifications that the rapidly scales to un-addressable sizes. Recently developed multiplex genome engineering methods (MAGE, TRMR and cas9) in combination with a library of reliable pathway expression platforms may aid in bridging predictive designs and rational genetic alterations to desired traits. Towards this goal, by using standard biological parts we are devising methods to rationally sample and optimize metabolic pathways; and improving TRMR technology to express every gene in *E coli* to any desired level. As a demonstration framework, we are applying these methods for Isobutanol production in E coli, an attractive biofuel candidate. The genome-scale technology platform developed here will serve as a knowledgebase for other biological systems.

Estimating average genome size from metagenomic data and its application to the human microbiome

Stephen Nayfach* (snayfach@gmail.com)[1,2] and Katherine S. Pollard[1,2,3,4]

[1]UCSF Graduate Program in Bioinformatics; [2]Gladstone Institute of Cardiovascular Disease; [3]UCSF Division of Biostatistics; [4]UCSF Institute for Human Genetics

The microbial communities living in the human body, known as the human microbiome, have been implicated in a number of disease processes. Next-generation sequencing has the potential to shed light on the functional role of the microbiome in these processes through identification of microbial genes differentially abundant between host phenotypes. However, identification of these candidate genes can be confounded by differences in the average size of genomes between microbial communities, leading to increases in both false positives and false negatives. Furthermore, due to a lack of tools designed for short-read metagenomic data, it has not been possible to characterize and correct for genome size variation in the human microbiome.

Here, we have developed a tool to rapidly and accurately estimate the average size of microbial genomes in a microbiome from shotgun sequence data by aligning short reads to a set of 30 universally distributed single copy genes. Using this tool, we show that average genome size varies significantly both within and between body sites in the human microbiome and that average genome size variation can be largely explained by species level taxonomic differences. Lastly, we explore whether normalizing for average genome size is able to reveal a more biologically relevant set of candidate genes associated with inflammatory bowel disease and type-II diabetes in the gut microbiome.

Simplified microbial communities as models for studying functional compartmentalization

William C. Nelson* (William.Nelson@pnnl.gov), Stephen R. Lindemann, Margie F. Romine, and James K. Fredrickson

Pacific Northwest National Laboratory, Richland, Washington

Microbial mats are laminated biofilm communities of bacteria, archaea, and microeukaryotes embedded in extracellular polymeric substances. The complex and compact spatial structure promotes metabolic networks among community members, leading to efficient energy utilization, nutrient cycling, and obligate mutualistic relationships. Assembly of these communities is driven by a combination of stochastic forces (e.g., colonization order) and deterministic forces (energy acquisition, nutrient availability, resource competition, microbial interactions), while spatial arrangement is constrained by gradients of physical and chemical parameters formed by environmental forces and microbial function. We hypothesize that functional compartmentalization promotes metabolic interaction and interdependencies between community members. These interactions have a stabilizing effect upon diversity, and thereby lead to community resilience to environmental variation.

Studies of microbial ecology generally suffer from three common difficulties: 1) high community diversity complicates identification of rare but important members; 2) in field studies, environmental variation is uncontrolled and stochastic, clouding causative relationships; 3) in laboratory studies, constructed communities rely on cultured organisms, which can call into question the ecological relevance of the interactions. We have developed a system that addresses all three problems.

To examine metabolic functional compartmentalization and the effects of environmental variables upon community composition, we have generated two unicyanobacterial consortia (UCC-A and UCC-O) derived from the hypersaline Hot Lake phototrophic mat. Community composition is simple, being composed of a single autotroph and <20 associated heterotrophs. The consortia are stable to serial dilution passage in culture and thus tractable to controlled manipulation and, being derived from an environmental sample, interactions and dependencies between members are likely to have ecological relevance.

Metagenomic sequence data collected from the UCC-A and UCC-O cultures have been assembled and segregated into taxonomic bins resulting in 18 distinct, near-complete genome sequences, 15 of which are shared between the two communities. Each consortium contains a single cyanobacterium that is the sole autotroph and a heterotrophic assemblage comprised of members of Bacteroidetes, Gammaproteobacteria, and Alphaproteobacteria. Genome analysis indicates most organisms in the cultures can use urea as a nitrogen source, and the Gammaproteobacteria can use cyanate, a toxic product of urea decomposition. The alphaproteobacterium *Oceanicola* possesses the most diverse set of carbohydrate catabolic genes and is capable of degrading a range of mono- and disaccharides, organic acids and sugar alcohols. Other community members contain varying subsets of carbohydrate catabolism genes, with the gammaproteobacterium *Idiomarina* and alphaproteobacterium *Oceanicaulis*, putative amino acid fermenters, having fewest. Isolate strains were tested for growth on various carbon sources to test predicted carbon sources.

These examples highlight potential niche specialization and functional compartmentalization among community members that will be empirically validated using our consortia and isolate cultures. Interactions identified in the consortia cultures will be tested through field observations to evaluate the consortia as model ecological systems.

Ochratoxin A biosynthetic pathway elucidation in the fungus *Penicillium verrucosum* by de novo genome assembly, genome annotation, and transcriptome analysis

Hai D.T. Nguyen* (hai.nguyen.1984@gmail.com),[1,2] Michael Li,[1] Ekaterina Ponomareva,[1] Ravpreet Setia,[1] Jeff Cullis,[1] Iyad Kandalaft,[1] David McMullin,[3] J. David Miller,[3] and Keith A. Seifert[1,2]

[1] Biodiversity (Mycology and Microbiology), Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada; [2] Department of Biology, University of Ottawa, Ottawa, Ontario, Canada; [3]Department of Chemistry, Carleton University, Ottawa, Ontario, Canada

Ochratoxin A (OTA) is a polyketide mycotoxin produced by *Penicillium verrucosum*, a fungus that grows on cereals such as wheat and barley in storage. OTA is nephrotoxic and is a possible human carcinogen (IARC 2B). OTA is regulated in a variety of products in the EU at the ng/kg level. Previous studies indicated that the polyketide synthase gene (otapksPV), a putative transporter gene (otaT) and a putative oxidoreductase gene (otaE) are involved in OTA biosynthesis in *P. verrucosum*. However, the full biosynthetic pathway has yet to be elucidated. The objectives of this study are to characterize the full set of genes involved in OTA biosynthesis and to find evidence of gene expression for other secondary metabolites, e.g. for verrucolone, verrucin A and B. The genome of an OTA producing strain of *P. verrucosum* DAOM 242724 was sequenced. The transcriptome was sequenced from a time course experiment, lasting 6 days, as OTA accumulated in the production medium measured by LC/MS. The genomic and transcriptomic reads were assembled followed by annotation of the genome. Differentially expressed genes were determined by comparing the transcriptome of the time points before OTA production, which occurred at day 2, and during OTA production, which occurred at day 4. The genome assembly size of *P. verrucosum* DAOM 242724 was 30.66 Mbp and 10389 genes were annotated. During OTA production, 2345 genes were down regulated, 2286 genes were up regulated and otapksPV was up regulated. These preliminary results provide additional insight into the secondary metabolite genes and their expression.

Ecogenomics sheds light on synergistic networks of microbial dark matter in a methanogenic bioreactor

Masaru Nobu* (nobu2@illinois.edu), Takashi Narihiro,[1,2] Christian Rinke,[3] Yoichi Kamagata,[2,4] Susannah Tringe,[3] Tanja Woyke,[3] and Wen-Tso Liu[1]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois; [2]Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan; [3]DOE Joint Genome Institute, Walnut Creek, California; [4]Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Sapporo, Japan

In methanogenic environments, fermenters, syntrophic metabolizers (syntrophs), and methanogens interact to facilitate anaerobic degradation of organic compounds to CH_4 and CO_2, an essential component of both the natural global carbon cycle and biotechnology treating anthropogenic waste. However, besides these microbial guilds, many other uncharacterized and uncultivated microorganisms have been consistently observed in natural methanogenic environments and in methanogenic bioprocesses treating wastewater with simple or complex chemical composition. Without characterizing such "microbial dark matter" (MDM), we cannot fully understand the ecology of the global carbon flux

through methanogenic environments. To address this, we apply single-cell genomics, metagenomics, and metatranscriptomics to a methanogenic bioreactor degrading terephthalate (TA), in which MDM dominate half of bacterial community. This ecogenomic effort generated 34 draft genomes and six pangenomes spanning 23 putative genera over 15 phyla and successfully characterized over 90% of the microbial community. *Pelotomaculum* and *Syntrophoharbdus* were identified as TA-degrading syntrophs that to syntrophically partner with methanogens to accomplish primary TA decomposition to CH_4/CO_2. Alongside methanogens, we identify several MDM that also syntrophically support the TA-degrading syntrophs by metabolizing inhibitory TA catabolism byproducts (CBP) (acetate, butyrate, and H_2). In parallel, MDM were also found to degrade biomass-bound biosynthesis products (BBP) (lipids and protein) anabolically produced by these syntrophic guilds. While lipolytic glycerol fermentation could directly produce methanogen utilizable substrates (MUS), protein degradation required multiple chaining syntrophic interactions to degrade amino acids, branched-chain fatty acids, and propionate. Therefore, the holistic TA mineralization to CH_4/CO_2 not only requires primary degraders and methanogens, but also CBP-degrading and BBP-scavenging MDM to form syntrophic networks beyond the conventionally studied binary syntrophy.

The Genome Portal of the Department of Energy Joint Genome Institute

Henrik Nordberg* (hnordberg@lbl.gov)**,** Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Tatyana Smirnova, and Inna Dubchak

DOE Joint Genome Institute, Walnut Creek, California

The JGI Genome Portal (http://genome.jgi.doe.gov) provides unified access to all JGI genomic databases and analytical tools. A user can search, download and explore multiple data sets available for all DOE JGI sequencing projects including their status, assemblies and annotations of sequenced genomes. Genome Portal in the past 2 years was significantly updated, with a specific emphasis on efficient handling of the rapidly growing amount of diverse genomic data accumulated in JGI.

A critical aspect of handling big data in genomics is the development of visualization and analysis tools that allow scientists to derive meaning from what are otherwise terrabases of inert sequence. An interactive visualization tool developed in the group allows us to explore contigs resulting from a single metagenome assembly. Implemented with modern web technologies that take advantage of the power of the computer's graphical processing unit (gpu), the tool allows the user to easily navigate over a 100,000 data points in multiple dimensions, among many biologically meaningful parameters of a dataset such as relative abundance, contig length, and G+C content.

Comparative genomics of *Nitrosospira* , ammonia-oxidizing bacteria from diverse environments

Jeanette M. Norton* (jeanette.norton@usu.edu)[1], Marlen C. Rice[1], Yang Ouyang[1], Hidetoshi Urakawa[2], Annette Bollmann[3], Lisa. Y. Stein[4], Luis Sayavedra-Soto[5], Martin G. Klotz[6], Chee Kent Lim[6], Christopher J. Sedlacek[3], Jessica Kozlowski[4], Jeppe L. Nielsen[7], Nicole Shapiro[8], and Lynne A. Goodwin[9]

[1]Utah State University, Logan, Utah; [2]Florida Gulf Coast University, Fort Myers, Florida; [3]Miami University, Oxford, Ohio; [4]University of Alberta, Edmonton, Alberta, Canada; [5]Oregon State University,

Corvallis, Oregon; [6]University of North Carolina, Charlotte, North Carolina; [7]Aalborg University, Aalborg, Denmark; [8]DOE Joint Genome Institute, Walnut Creek, California; [9]Los Alamos National Laboratory, Los Alamos, New Mexico

Bacteria in the genus *Nitrosospira* (*Bacteria; Proteobacteria; Betaproteobacteria; Nitrosomonadales; Nitrosomonadaceae*) are autotrophic, obligate chemolithotrophic ammonia-oxidizing bacteria (AOB) that have adapted to diverse habitats. The nitrosospiras exhibit a variety of morphologies (spiral, lobate and vibriod), which historically led to the assignment of different genera designations that are no longer valid (*Nitrosolobus*, *Nitrosovibrio*). Starting with *Nitrosospira multiformis* in 2008, several genomes of AOB in the genus *Nitrosospira* including the genome of the type species for the genus, *Nitrosospira briensis*, have been sequenced to variable degrees of closure, most under the umbrella of the DOE Community Sequencing Program at the Joint Genome Institute. Both *N. multiformis* and *N. briensis* were isolated from soil environments. Additional isolates from soil (*Nitrosospira* sp. strain AV, *Nitrosospira tenuis* and *Nitrosospira* sp. strain B6) and freshwater habitats (*Nitrosospira* sp. strains APG3 and IS-148), have been sequenced or are in progress. We comparatively examined the structure and evolutionary relationships of several gene clusters that encode modules key to their metabolic lifestyle including ammonia oxidation (*amo, hao*), carbon fixation (*cbb*), urea hydrolysis (*ure*), NAD-reducing hydrogen dehydrogenase (*hox*) and nitrite reduction (*nir, nor*). In addition, the genomes of these nitrosospiras were analyzed to discover inventory attributable to niche differentiation as well as differences in morphological and functional characteristics. Challenges to genome closure include the multiple near-identical copies of operons coding for the key metabolic modules for ammonia oxidation, ammonia monooxygenase and hydroxylamine oxidoreductase. The results are interpreted in the context of evolutionary descent within the betaproteobacterial ammonia-oxidizing bacteria. The contribution of ammonia oxidizing bacteria versus ammonia oxidizing archaea to nitrification in an agricultural soil treated with contrasting nitrogen sources is also described. Our observations suggest that although nitrosospiras are lower in abundance than archaeal ammonia oxidizers (i.e. *Nitrososphaera* spp.) they remain functionally dominant players in the nitrification process of agricultural soils.

Support from DOE-JGI Community Sequencing Program 2010, Utah Agric. Expt. Sta. #UTA00371, NSF:0541797 and the Nitrification Network http://nitrificationnetwork.org/

Mapping soil carbon from cradle to grave: Using comparative transcriptomics and proteomics analyses to identify the microbial blueprint for root-enhanced decomposition of organic matter

Erin E. Nuccio* (nuccio1@llnl.gov)[1], Shengjing Shi[2,3], Jizhong Zhou[3], Mary Lipton[4], Mary K. Firestone[2], Eoin L. Brodie[2,5], and **Jennifer Pett-Ridge[1]**

[1]Lawrence Livermore National Laboratory, Livermore, California; [2]University of California at Berkeley, Berkeley, California; [3]University of Oklahoma, Norman, Oklahoma; [4]Pacific Northwest National Laboratory, Richland, Washington; [5]Lawrence Berkeley National Laboratory, Berkeley, California

Plant roots and associated soil microbiomes comprise the primary nexus of belowground carbon cycling in terrestrial systems. Roots provide the source of C for stabilized organic matter in soil and it is now understood that most plant C is utilized or transformed by soil microorganisms en route to stabilization. Rhizosphere microbial taxa provide precursors to stabilized soil organic matter by transforming plant root exudates into microbial biomass and accelerating breakdown of plant tissues and root detritus. However, the soil microbial communities and carbohydrate/lignolytic gene transcripts mediating root-accelerated decomposition are largely unidentified. The primary goal of our recently initiated CSP and

JGI/EMSL projects is to determine how organic C transformation and decomposition in soil is altered by interactions between plant roots and the soil microbial community (bacteria, archaea, fungi, microfauna). We hypothesize that root-exudate stimulation of soil microbial populations results in the elevated expression of transcripts and proteins for the decomposition of macromolecular C compounds. We are currently using genomic, transcriptomic, and proteomic approaches (included stable-isotope techniques) to identify key metabolic pathways responsible for C transformation and mineralization during root in-growth and root death/decay. By integrating stable isotopes as tracers of natural resource utilization (i.e. root litter), and analysis of the functional properties of the communities that respond to those resources, we plan to identify what molecular pathways are stimulated in the soil microbiome in response to root litter, living roots, and their intersection. We will use this information to identify key traits of grassland soil microorganisms that impact their fitness as members of the rhizosphere and detritosphere, and will develop and parameterize trait-based models of microbial community function to interpret and predict carbon stabilization and turnover.

Predominance and adaptation of SAR11 bacteria in the planktonic microbial community of Lake Qinghai as revealed by metagenomics

Seungdae Oh* (seungdae@illinois.edu),[1] Rui Zhang,[2,3] Qinglong Wu,[4] and Wen-Tso Liu[1]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois; [2]State Key Laboratory of Marine Environmental Science, [3]Institute of Marine Microbes and Ecospheres, [4]Xiamen University, Fujian, China, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China

The SAR11 clade is a member of heterotrophic *Alphaproteobacteria* that are abundant and play a key role in the carbon cycle in seawater. Our previous 16S rRNA gene clone library analysis discovered novel freshwater SAR11 bacteria (subclade IIIa) in Lake Qinghai, a high-altitude mountain lake on the Tibetan Plateau; however, differences in gene content between the freshwater bacteria and other SAR11 ones and what functional genomic potential relates to the ecological niche of the bacteria remain unknown. To address these issues, we performed terminal restriction fragment length polymorphism (T-RFLP) analysis of planktonic samples over the whole lake area at different depths (0.5 to 22 meter). The T-RFLP analysis suggested that all the microbial communities (n > 10) sampled are closely related with each other (>80% similarity), as assessed by occurrence (presence/absence) and abundance of phylotypes. Based on the T-RFLP analysis, a representative microbial community sample was selected and further analyzed by metagenomics. The results revealed the predominance (14% of the total) of the SAR11 subclade IIIa species, which showed significant genomic divergence (Avg. 80% amino acid sequence identity) to its closest marine counterpart, despite the high 16S rRNA gene sequence identity (>98%). Comparative analysis of the freshwater species genome revealed i) conservation of carbon and energy storage metabolism gene operons (biosynthesis of polyphosphate and polyhydroxyalcanoate) that are rarely found among other SAR11 bacteria genomes and ii) amino acid substitutions of a gene (exodeoxyribonuclease) related to DNA repair system, implying the genetic footprint of the species adaptation to the unique environment (e.g., nutrient limitation and strong UV radiation). Collectively, our results provide insights into the predominance, divergence, and adaptive evolution of the SAR11 bacteria inhabiting the unique freshwater lake ecosystem.

Functional and comparative genomics of lignocellulose degradation by *Schizophyllum commune*

Robin A. Ohm* (raohm@lbl.gov),[1] Hanbyul Lee,[1,2] Hongjae Park,[1,2] Heather M. Brewer,[3] Akiko Carver,[1] Alex Copeland,[1] Jane Grimwood,[4] Erika Lindquist,[1] Anna Lipzen,[1] Joel Martin,[1] Samuel O. Purvine,[3] Wendy Schackwitz,[1] Martin Tegelaar,[5] Andrew Tritt,[1] Scott Baker,[3] In-Geol Choi,[1,2] Luis G. Lugones,[5] Han A.B. Wösten,[5] and Igor V. Grigoriev[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Korea University, Seoul, South Korea; [3]Pacific Northwest National Laboratory, Richland, Washington; [4]Hudson Alpha Institute for Biotechnology, Huntsville, Alabama; [5]Utrecht University, Utrecht, The Netherlands

The Basidiomycete fungus *Schizophyllum commune* is a wood-decaying fungus and is used as a model system to study lignocellulose degradation. Version 3.0 of the genome assembly filled 269 of 316 sequence gaps and added 680 kb of sequence. This new assembly was reannotated using RNAseq transcriptomics data, and this resulted in 3110 (24%) more genes.

Two additional *S. commune* strains with different wood-decaying properties were sequenced, from Tattone (France) and Loenen (The Netherlands). Sequence comparison shows remarkably high sequence diversity between the strains. The overall SNP rate of > 100 SNPs/kb is among the highest rates of within-species polymorphisms in Basidiomycetes. Some well-described proteins like hydrophobins and transcription factors have less than 70% sequence identity among the strains. Some chromosomes are better conserved than others and in some cases large parts of chromosomes are missing from one or more strains.

Gene expression on glucose, cellulose and wood was analyzed in two *S. commune* strains. Overall, gene expression correlated between the two strains, but there were some notable exceptions. Of particular interest are CAZymes (carbohydrate-active enzymes) that are regulated in different ways in the different strains.

In both strains the transcription factor Fsp1 was strongly up-regulated during growth on cellulose and wood, when compared to glucose. Over-expression of Fsp1 using a constitutive promoter resulted in higher cellulose and xylose-degrading enzyme activity, which suggests that Fsp1 is involved in regulating CAZyme gene expression.

Two CAZyme genes (of family GH61 and GH11) were shown to be strongly up-regulated during growth on cellulose, compared to glucose. Proteomics on the secreted proteins in the growth medium confirmed this. A promoter analysis revealed the shortest active promoters for these two genes, as well as putative transcription factor binding sites.

Geneious R7: A bioinformatics platform for biologists

Olsen, C.* (christian@biomatters.com),[1] Qaadri, K.,[1] Moir, R.,[2] Kearse, M.,[2] Buxton, S.,[2] Cheung, M.,[2] Milicevic, B.,[2] Hengjie, W.,[2] Kuhn, J.,[2] Stones-Havas, S.,[2] and Duran, C.[2]

[1]Biomatters Inc., Newark, New Jersey; [2]Biomatters Ltd., Auckland, Zealand

Biomatters' Geneious R7 is a bioinformatics software platform that allows researchers the command of industry-leading algorithms and tools for their genomic and protein sequence analyses. Using a glass-box approach for software design, Geneious R7 offers a comprehensive suite of peer-reviewed tools that enable researchers to be more efficient with their bioinformatic workflows. Researchers at all levels can easily manage, analyze, and share their sequence data via a single intuitive software application. R7

provides tools for next-generation sequence analysis, sequence alignment, molecular cloning, chromatogram assembly, and phylogenetics. New features for this major version release include tools for Gibson assembly and TOPO cloning along with algorithms for RAxML, FastTree, Garli, LastZ, Bowtie2, as well as a number of new plug-ins. R7 affords real-time dynamic interaction with sequence data and empowers biologists to produce stunning publication quality images to increase the impact of their research. By utilizing Geneious R7, biologists can easily improve their genomic workflow efficiencies to free up more time for their research. This poster aims to demonstrate the new features and benefits of the highly integrated Geneious R7 tool-suite.

Haplotyping assembly refinement and improvement II

Olsen, C.* (Christian@biomatters.com),[1] Qaadri, K.,[1] Hsiau, T.,[2] Rudenko, G.,[2] Huynh, T.,[2] Somanchi, A.,[2] Zhao, X.,[2] and **Brubaker, S.**[2]

[1]Biomatters Inc., Newark, New Jersey; [2]Solazyme Inc., South San Francisco, California

Haplotyping is the assignment of polymorphisms to the correct allele(s) in a diploid or polyploid organism. Haplotyping is an especially complex problem that has been little addressed by current technologies. In particular, current Next-Generation Sequencing (NGS) technologies that use short reads make it difficult to haplotype over long distances. A related complex problem is that of diploid or polyploid assembly, which requires the accurate extension of reads to assemble distinct haplotypes for each allele. Current sequence assemblers are designed to try and collapse/merge two or more alleles that may be present in an organism. Applying these assemblers to sequences from diploid or highly polymorphic organisms leads to many problems and misassemblies, which have been well documented. However an effective solution to this problem has not yet been presented. In this poster using Geneious R7 we present an efficient scalable strategy to assemble and haplotype a very high-GC, polymorphic, diploid organism.

Command line environment in the DOE Systems Biology Knowledgebase

Bob Olson (olson@mcs.anl.gov),[1] James Thomason,[2] Thomas Brettin*,[1] Jason Baumohl,[1,4] Aaron Best,[3] Jared Bischof,[1] Ben Bowen,[4] Tom Brown,[1] Shane Canon,[4] Stephen Chan,[4] John-Marc Chandonia,[4] Dylan Chivian,[4] Ric Colasanti,[1] Neal Conrad,[1] Brian Davison,[5] Matt DeJongh,[3] Paramvir Dehal,[4] Narayan Desai,[1] Scott Devoid,[1] Terry Disz,[1] Meghan Drake,[5] Janaka Edirisinghe,[1] Gang Fang,[7] José Pedro Lopes Faria,[1] Mark Gerstein,[7] Elizabeth M. Glass,[1] Annette Greiner,[4] Dan Gunter,[4] James Gurtowski,[2] Nomi Harris,[4] Travis Harrison,[1] Fei He,[6] Matt Henderson,[4] Chris Henry,[1] Adina Howe,[1] Marcin Joachimiak,[4] Kevin Keegan,[1] Keith Keller,[2] Guruprasad Kora,[5] Sunita Kumari,[2] Miriam Land,[5] Folker Meyer,[1] Steve Moulton,[5] Pavel Novichkov,[4] Taeyun Oh,[8] Gary Olsen,[9] Dan Olson,[1] Ross Overbeek,[1] Tobias Paczian,[1] Bruce Parrello,[1] Shiran Pasternak,[2] Sarah Poon,[4] Gavin Price,[4] Srivdya Ramakrishnan,[2] Priya Ranjan,[5] Bill Riehl,[4] Pamela Ronald,[8] Michael Schatz,[2] Lynn Schriml,[10] Sam Seaver,[1] Michael W. Sneddon,[4] Roman Sutormin,[4] Mustafa Syed,[5] James Thomason,[2] Nathan Tintle[3] Will Trimble,[1] Daifeng Wang,[7] Doreen Ware,[2] David Weston,[5] Andreas Wilke,[1] Fangfang Xia,[1] Shinjae Yoo,[6] Dantong Yu,[6] Robert Cottingham,[5] Sergei Maslov,[6] Rick Stevens,[1] Adam P. Arkin[4]

[1]Argonne National Laboratory, Argonne, Illinois; [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [3]Hope College, Holland, Michigan; [4]Lawrence Berkeley National Laboratory, Berkeley,

California; [5]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [6]Brookhaven National Laboratory, Upton, New York; [7]Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9]University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

KBase provides a variety of user experiences targeted to meet the needs of a diverse user-base in the systems biology community. A rich programmatic API is provided in a wide variety of languages for maximum performance and flexibility for computationally advanced users. The graphical Narrative interface provides an intuitive visual experience to facilitate hands-on interactive systems biology. Between these is the command line interface, providing a powerful but simple command-driven interface capable of high-throughput interaction with KBase tools and data. The online IRIS web application (http://iris.kbase.us) permits users to run commands in KBase from any computer with web access without installing any local software. The installable KBase clients for Mac and Ubuntu Linux enable users to run KBase commands from a shell on their own computers, supporting seamless integration with locally installed data and tools. The installable client also includes API libraries in numerous programming languages (e.g., Java, JavaScript, Perl, Python, JSON RPC), enabling users to program directly against the KBase API from their own computers and workstations. In this way, the existing computational infrastructure in any academic lab can seamlessly integrate KBase tools and data into local data analysis pipelines.

The KBase command line environmentnow offers a powerful suite of over 800 commands that encompass genome assembly, annotation, metabolic modeling, expression data analysis, phenotype analysis, and GWAS. Since the initial release in February 2013, the KBase command line environments have collectively been applied by over 300 unique users to: (i) annotate or import over 34K genomes, (ii) construct over 40K metabolic models, and (iii) run over 110K FBA simulations.

Multi-platform characterization of the root microbiome in immune altered hosts

Sur Herrera Paredes* (sur00mx@gmail.com)[1,2], **Sarah L. Lebeis[1,4]**, Derek S. Lundberg[1,3], Stephanie Malfatti[5], Tijana Glavina del Rio[5], Susannah Tringe[5], and Jeffery L. Dangl[1,3,6]

[1]Department of Biology, [2]Curriculum in Bioinformatics and Computational Biology, [3]Curriculum in Genetics and Molecular Biology, The University of North Carolina at Chapel Hill; [4]Department of Microbiology, University of Tennessee Knoxville; [5]DOE Joint Genome Institute; [6]Howard Hughes Medical Institute

Plants, like other multicellular organisms, assemble microbial communities from their surrounding environment. In light of the complexity of bacterial communities in the root, a natural question is how does the plant immune system distinguish and tolerate this bacterial population, while still being able to mount a defense response in the presence of a pathogen. To understand the contribution of the plant-immune system to the assembly of the root microbiome, we characterized the microbial composition in the roots of plants grown in a natural soil in a number of mutants in several phytohormone biosynthesis and signaling pathways which show altered immune response. We found that a distinct set of taxonomic groups associates differentially with each of these mutants, and the bacterial enrichment profile of these mutants correlates with their underlying genetic architecture. To validate our results, we re-sequenced a subset of our samples with a different technology and scanned two different variable regions from the 16S ribosomal gene. Despite clear platform specific biases; we were able to recapitulate the main patterns in bacterial alpha and beta diversity, as well as the major enrichments in

roots. On the other hand, fungal profiles obtained by sequencing of the ITS2 region showed no differences in the taxonomic profiles of roots and surrounding soils. In summary, our results indicate that the plant immune system modulates microbiome composition and provides a robust, sequencing platform independent, set of hypothesis to be tested on a more controlled setting.

Characterization of U(VI) toxicity and detoxification in the uranium tolerant organism, *Caulobacter crescentus*

Dan M. Park* (park36@llnl.gov) and Yongqin Jiao

Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California

The oxidized form of uranium (U(VI)) predominates in oxic environments and poses a major threat to living organisms due to its high toxicity and bioavailability. The obligate aerobe, *Caulobacter crescentus*, has a ubiquitous presence in fresh water lakes and streams and is known to tolerate high concentrations of U(VI). The physiological basis for U(VI) tolerance by *C. crescentus* is unclear. In this study, we demonstrated that exposure to U(VI) caused a temporary growth arrest in *C. crescentus* and three other bacterial species, although the duration of growth arrest was significantly shorter for *C. crescentus*. Upon recovery from growth arrest, *C. crescentus* proliferated with normal growth kinetics but remained inherently susceptible to U(VI), suggesting that recovery did not occur through the selection of U(VI) insensitive mutants. Characterization of U(VI) toxicity in the growth media throughout the duration of growth arrest with growth assays and a functional reporter assay that employed a U(VI)-inducible promoter (P_{urcA}), indicated that *C. crescentus* reduced U(VI) toxicity to non-growth inhibitory levels. Measurements of U(VI) solubility revealed that this reduced toxicity was not due to decreased U(VI) solubility, in contrast to previously reported U(VI) detoxification strategies (e.g., U(VI) reduction, precipitation or surface biosorption). Instead, a slight increase in medium pH, resulting from metabolic activity during growth arrest, was shown to be sufficient to alleviate growth arrest and abolish P_{urcA} expression. The simplest explanation is that the pH change altered U(VI) speciation in a manner that reduced toxicity. However, a similar increase in pH was not sufficient to alleviate growth arrest in the other three species tested, suggesting that *Caulobacter* confers higher U(VI) tolerance. RNA-seq analyses during the course of growth arrest and after recovery are currently underway to better understand the causes of U(VI)-induced growth arrest. Preliminary scanning electron microscopy (SEM) imaging revealed that U(VI) inhibited swarmer to stalk cell differentiation, a pre-requisite step for cell division during the *Caulobacter* cell cycle. qRT-PCR and β-galactosidase assays suggested that cell cycle arrest is not likely a result of oxidative stress since *sodA* (superoxide dismutase) and *katG* (catalase-peroxidase) are not induced during growth arrest. However, DNA damage may play a role in growth arrest since a *recA* mutant exhibited increased sensitivity to U(VI) and SEM imaging of cells that had recently transitioned from growth arrest into exponential phase appeared elongated in comparison to cells without U(VI) exposure. Together, this data provides insight into the mechanisms of U(VI) toxicity, which remain poorly understood in microorganisms.

Evaluation of acyl-CoA transferases from *Megasphaera* sp. BS-4 for alcohol production

Hyojung Park* (suya8259@naver.com),[1] Byoung Seung Jeon,[1] and Byoung-In Sang* (biosang@hanyang.ac.kr)[1,2]

[1]Department of Chemical Engineering, Hanyang University, Seoul, Republic of Korea; [2]Department of Fuel Cells and Hydrogen Technology, Hanyang University, Seoul, Republic of Korea

Megasphaera sp. BS-4, producing various carboxylic acids, was isolated from cow rumen and produced pentanoic acid, hexanoic acid, heptanoic acid, and octanoic acid when several short-chain carboxylic acids, such as propionic acid and butyric acid, were added. Based on our analysis of whole-genome sequencing and the metabolic pathway, nine kinds of acyl-CoA transferase(ACT) were the key enzyme of carboxylic acids production from the added short-chain carboxylic acids. The ACT can transfer CoA molecule of acetyl-CoA into carboxylic acid anion and relates to various CoA dependent pathways. Each ACT was expressed in Escherichia coli BL21 (DE3) and then their reaction characteristics were investigated in vivo. E. coli containing the ACTs had ability to convert carboxylic acids to the corresponding alcohols.

Long-read sequencing and assembly of the Landsberg erecta strain of *Arabidopsis* with the PacBio® RS II

Paul Peluso* (ppeluso@paciifcbiosiences.com), **David Rank*** (drank@pacificbiosciences.com)**,** Jason Chin, and Joey Karalius

Pacific Biosciences, Menlo Park, California

Significant advances in single-molecule-sequencing read lengths enable highly contiguous genome assemblies for bacteria and model genomes. Improvements in raw-read error correction, assembly algorithms, and assembly polishing of Pacific Biosciences' long-read data have led to high-quality assemblies that rival pre-second generation clone-by-clone sequencing efforts. As a demonstration, we have whole-genome shotgun sequenced the Landsberg erecta (Ler-0) strain of *Arabidopsis thaliana* and produced a highly contiguous assembly in which many full chromosome arms are represented as single contigs. The contig N50 for the assembly is 6.2 Mbases with the longest contig, exceeding 12.5 Mbases. The average base quality of this assembly is ~Q50. Library preparation methods including shearing and size selection, sequencing chemistry and assembly methods will be presented. Additionally, we will present data using the most recent chemistry release (P5-C3).

Genome editing in *Escherichia coli* with Cas9 and synthetic CRISPRs

Ze Peng* (zpeng@lbl.gov), Sarah Richardson, David Robinson, Samuel Deutsch, and Jan-Fang Cheng

DOE Joint Genome Institute, Walnut Creek, California

Recently, the Cas9-CRISPR system has proven to be a useful tool for genome editing in eukaryotes, which repair the double stranded breaks made by Cas9 with non-homologous end joining or homologous recombination. Escherichia coli lacks non-homologous end joining and has a very low homologous recombination rate, effectively rendering targeted Cas9 activity lethal. We have developed a heat curable, serializable, plasmid based system for selectionless Cas9 editing in arbitrary E. coli strains

that uses synthetic CRISPRs for targeting and λ-red to effect repairs of double stranded breaks. We have demonstrated insertions, substitutions, and multi-target deletions with our system, which we have tested in several strains.

Identification of carbohydrate active enzymes from rumen fungi by targeted metatranscriptomics

Hailan Piao*,[1] Stephanie Malfatti,[2] Alexander Sczyrba,[3] Kristin Tennessen,[4] Andrew Tritt,[4] Erika Lindquist,[4] Kerrie Barry,[4] Chew Yee Ngan,[4] James Han,[4] Amrita Pati,[4] Susannah Tringe,[4] Igor Grigoriev,[4] Roderick Mackie,[5] and Matthias Hess (matthias.hess@tricity.wsu.edu)[1,4,6,8]

[1]Washington State University Tri-Cites, Richland, Washington; [2]Lawrence Livermore National Laboratory, Livermore, California; [3]University of Bielefeld, Bielefeld, Germany; [4]DOE Joint Genome Institute, Walnut Creek, California; [5]University of Illinois, Urbana-Champaign, Illinois; [6]Pacific Northwest National Laboratory, Richland, Washington; [7]Washington State University, Pullman, Washington

The cow's rumen is one of the most efficient biomass-degrading ecosystems and it contains billions of microorganisms (i.e. archaea, bacteria, protozoa, and fungi) that work synergistically [1,2]. Many of these microbes are tightly attached to solid rumen content [3] and have attracted great interest by the scientific community due to their ability to convert complex lignocellulosic biomass into less complex biofuel precursors [4]. Rumen fungi contribute only between 8 to 20% of the total rumen microbial biomass [5,6] but they play a key role in the anaerobic degradation of lignocellulosic plant material [5,7]. To overcome the paucity to fungal isolates from the rumen and to identify fungal enzymes capable of biomass degradation, we generated a total of ~2.7 Gb of raw metatranscriptome data, representing ~16 million high quality reads. The assembled data amounted to a total of almost 5.2 Mb and gene prediction yielded approximately 16,000 protein coding genes, with 2.5% of the genes assigned to be of fungal origin. Of the protein coding genes ~76% lacked significant similarity to already known genes to be classified phylogenetically, which is most likely due to the small number of reference genomes from anaerobic fungi that are presently available. The assembled contigs were screened for regions specific for glycoside hydrolases (GHs) and carbohydrate binding modules (CBMs), resulting in the identification of 22 and 18 of the currently known GH and CBM families, respectively. Contigs representing GH family 1, 3 and 13 were selected for targeted gene amplification subsequent gene expression and detailed physicochemical analysis of the corresponding recombinant proteins. Expression constructs are currently being generated in the Hess Lab and results of the enzymatic assays will be presented.

Functional Encyclopedia of Bacteria and Archaea: Evidence-based annotation of the microbial Tree of Life

Morgan N. Price[1], Kelly M. Wetmore[1], Robert J. Waters[2], Jacob S. Lamson[1], Cindi Hoover[2], Gareth Butland[3], James Bristow[2], Matthew J. Blow[2], Adam P. Arkin[1], and **Adam Deutschbauer*** (AMDeutschbauer@lbl.gov)[1*]

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley California; [2]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, California; [3]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

There is a great discrepancy between our ability to sequence microbial genomes and to characterize the genes encoded by these genomes. To meet this challenge, we have initiated a large microbial functional

genomics initiative with the JGI, the Functional Encyclopedia of Bacteria and Archaea (FEBA). Overall, our aim is to generate evidence-based annotations of gene structure and gene function for 50 poorly characterized and genetically diverse microorganisms including those with relevance to DOE missions of bioenergy, carbon sequestration, and bioremediation. Here, we describe our microbial functional annotation pipeline consisting of high-throughput culturing, RNAseq, and transposon mutagenesis, and how these tools are widely applicable to diverse microorganisms. In particular, to enable the inference of gene function on a global scale in diverse bacteria, we developed a method, random barcode TnSeq (RB-TnSeq), to increase the throughput of mutant fitness profiling by incorporating random DNA barcodes into transposons and monitoring mutant fitness using barcode sequencing (BarSeq). Here, we present RB-TnSeq results demonstrating the reproducibility and biological meaningfulness of the method using *Shewanella oneidensis* MR-1, *Shewanella amazonensis* SB2B, *Phaeobacter inhibens* BS107 and *Pseudomonas stutzeri* RCH2 as pilot organisms. We identified phenotypes for 5,157 genes across 150 diverse metabolic and stress conditions from 640 genome-wide fitness experiments. For 723 genes that are not annotated with a specific function, we detected both a strong phenotype and a significant correlation in fitness with another gene, thereby providing an associative functional annotation for these poorly characterized genes.

UDP-Rhamnose and UDP-Galactose in plants are transported by bi-functional nucleotide sugar transporters localized in Golgi membranes

Carsten Rautengarten[1], Berit Ebert[1,2], Ignacio Moreno[3], Henry Temple[3,] Thomas Herter[1], Bruce Link[4], Jennifer C. Mortimer[5], Alex Schultink[6], Wolf-Dieter Reiter[4], Paul Dupree[5], Markus Pauly[6], Ariel Orellana[3], Joshua L. Heazlewood[1], and Henrik Vibe Scheller* (hscheller@lbl.gov)[1,6]

[1]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, California; [2]Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Denmark; [3]Centro de Biotecnología Vegetal, Universidad Andrés Bello, Santiago, Chile; [4]Department of Molecular and Cell Biology, University of Connecticut; [5]Department of Biochemistry, University of Cambridge, UK; [6]Department of Plant and Microbial Biology, University of California at Berkeley, Berkeley, California

Plant cell walls are composed of polysaccharides most of which are synthesized in the Golgi apparatus from nucleotide sugars that are transferred from the cytosol into the Golgi lumen by nucleotide sugar transporters (NSTs). By using a novel approach that combines reconstitution of NSTs into liposomes and subsequent LC-MS/MS analysis of nucleotide sugar uptake, we identified six bi-functional UDP-rhamnose / UDP-galactose transporters. All six transporters are localized in the Golgi. Mutants in URGT1 have reduced galactose in the cell wall whereas overexpressors accumulate up to 50% more galactose than wild-type plants, mainly in the pectin-rich fraction. In contrast, galactose in xyloglucan is unaffected in mutants and overexpressors. Mutants in URGT2 exhibit lower levels of rhamnose and galacturonic acid in seed mucilage, a structure rich in rhamnogalacturonan I while URGT2 overexpressors have no change in leaf cell wall composition. Our results suggest that some NSTs may be channeling nucleotide sugars to specific glycosyltransferases. Furthermore, the results show that nucleotide sugar transporters can be limiting for biosynthesis of cell wall polysaccharides, and therefore they are interesting targets for engineering of plants with improved properties as feedstocks for biofuel production.

This work conducted by the Joint BioEnergy Institute was supported by the Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Whole metagenome profiling of microbial community evolution in a tetrachloroethene-contaminated aquifer responding to enhanced reductive dechlorination protocols

Rebecca A. Reiss* (reiss@nmt.edu**),**[1] Peter Guerra[2], and Oleg Makhnin[1]

[1]New Mexico Institute of Mining and Technology, Socorro, New Mexico; [2]AMEC Environment & Infrastructure Inc., Albuquerque, New Mexico

Chlorinated solvent contamination of potable water supplies is a worldwide problem. Although biostimulation protocols can successfully remediate chlorinated solvent contamination by facilitating enhanced reductive dechlorination (ERD) pathways, this process is poorly understood and it sometimes stalls creating a more serious problem. Here we compare the metagenomes of a tetrachloroethene contaminated EPA Superfund site in Española, New Mexico before and after ERD biostimulation protocols were applied. Environmental DNA (eDNA) was extracted from uncultured microbes harvested by on-site filtration of groundwater one month prior to and five months after the injection of emulsified vegetable oil (EVO) and nutrient bioamendments into the contaminated aquifer. Pair-end libraries were prepared for high-throughput DNA sequencing and 90 basepairs from both ends of randomly fragmented 400 bp DNA fragments were sequenced. Over 31 millions reads were uploaded to the Metagenome Rapid Annotation using Subsystem Technology (MG-RAST) site and annotations representing 32 prokaryotic phyla, 869 genera, and 3,181 species were obtained. An 11-$\log_2$ fold increase in biomass as measured by DNA yield per mL water was observed, but 79 out of the 869 genera were undetectable following biostimulation. Members of the Archaea domain increased 4.7 $\log_2$ fold, dominated by methanogens. Plotting the rank abundance of each genus against log-fold-change facilitates the visualization of the changes in the microbial community in response to ERD. Marker shape denotes different taxonomic groups plot and demonstrates that α- and β-proteobacteria dominated the community prior to ERD but exhibit significant decreases five months after biostimulation. *Geobacter* and *Sulfurospirillum* replace *Sideroxydans* and *Burkholderia* as the most abundant genera. Marker color indicates metabolic classification and δ- and ε-proteobacteria capable of dehalogenation, iron and sulfate reduction, and sulfur oxidation increase as a result of biostimulation. These metabolic classifications track the geochemical data collected during remediation as the increase in ferrous iron is consistent with the decrease in iron oxidizing bacteria. Additionally, the increase in sulfate reducing bacteria correlates to the decrease in sulfate. Matches to thermophilic Archaea genera suggest that there are novel species that are important in dechlorination. Additionally, potentially pathogenic OTUs increase, indicating that the process of bioremediation may contribute to the emergence of pathogenic bacteria.

Towards improved methods for riboswitch identification

Matan Drory Retwitzer* (matandro@bgu.ac.il),[1] Ilona Kifer,[2] Zohar Yakhini,[2] and Danny Barash[1]

[1]Department of Computer Science, Ben Gurion University, Israel; [2]Agilent Laboratories, Tel Aviv, Israel

Riboswitches are regulatory segments of an mRNA that provide a unique mechanism of gene regulation without the intervention of proteins. They were originally discovered in bacteria where they are abundant and only very few were found in eukaryotes. A riboswitch functions by binding a small molecule, resulting in a change in structure which then leads to altered production of proteins encoded by the mRNA. Discovering new riboswitches, especially in eukaryotic systems, where they are scarce, is

highly significant not only for the general understanding of the principles governing gene expression regulation but also for potential applications in synthetic biology and in medical science. Currently, riboswitches are being tested as antibacterial drug targets, for example as a way to treat human bacterial infections caused by the bacterium Clostrodium difficile. However, the discovery of riboswitches in higher organisms, and most ambitiously in humans, will have further implications on drug discovery. Thus, the development of additional bioinformatics methods, which are more specific towards the detection of riboswitch patterns and can be applied in a variety of genomes, is valuable. We are developing more flexible methods than the ones currently used, as well as new structure-based methods, intended to improve riboswitch pattern detection. As a model organism, we selected fungi where our methods will be tested on the 1000 fungal genomes project sequenced at the DOE Joint Genome Institute (JGI). More ambitiously, we will apply these methods to human genomic sequence data for the purpose of riboswitch identification.

* The Frankel Center for Computer Science at Ben-Gurion University is acknowledged for partially funding this work.

FUG1, a novel fungal-specific gene, is a key regulator of morphogenesis, conidiation, and pathogenesis in multiple *Fusarium* spp.

John B. Ridenour* and Burt H. Bluhm (bbluhm@uark.edu)

Department of Plant Pathology, University of Arkansas, Division of Agriculture, Fayetteville, Arkansas

Filamentous fungi are responsible for numerous plant diseases and can be detrimental to human and animal health directly or through the production of toxic metabolites. The filamentous ascomycete *Fusarium verticillioides* is a ubiquitous and destructive pathogen of maize. *F. verticillioides* can cause severe yield reduction and economic loss through seedling disease, stalk rot, and ear and kernel rot. Also of substantial concern is the contamination of infected kernels with fumonisins, a group of polyketide-derived toxins linked to chronic and acute toxicoses in humans and livestock. Although the number of sequenced fungal genomes is rapidly increasing, the functional characterization of novel genes has not progressed at a comparable rate. In pathogenic fungi, uncharacterized genes represent a source of novel virulence factors or anti-fungal targets. To fully understand the mechanisms underlying fungal pathogenesis and identify new strategies for control, characterizing genes with no known or predicted function is of significant importance in the post-genomic era. Recently, we identified a novel gene, provisionally named *FUG1* (Fungal Unknown Gene 1), in *F. verticillioides*. Despite broad conservation among phylogenetically diverse groups of fungi, putative *FUG1* orthologs have not been studied and have no known or predicted biological function. The overarching goal of this project is to functionally characterize the role *FUG1* plays in growth and development, secondary metabolism, and virulence in *F. verticillioides*. Targeted deletion of *FUG1* in *F. verticillioides* resulted in a pleiotropic phenotype, including reduced radial growth and conidiation and altered colony morphology. Additionally, the deletion strain was more sensitive to oxidative stress, providing evidence that *FUG1* may have an important role in responding to host defense mechanisms. Deletion of *FUG1* impaired the colonization of maize kernels and stalks and reduced the accumulation of fumonisins in the kernel environment. RNA-seq of infected maize kernels revealed numerous categories of fungal genes were differentially expressed under *FUG1* perturbation, particularly genes with putative binding, catalytic, or metabolic activity. Moreover, deletion of *FUG1* in the related pathogen *Fusarium graminearum* resulted in reduced colonization of maize silks and wheat heads, suggesting that *FUG1* may play a conserved role

in pathogenesis in plant pathogenic fungi. Taken together, the results from this work directly implicate *FUG1* in growth, reproduction, and pathogenesis in two *Fusarium* spp. and represent the first step in creating a working model to explain how *FUG1* functions at the molecular level to regulate multiple biological processes.

Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white rot/brown rot paradigm for wood decay fungi

Robert Riley* (rwriley@lbl.gov),[1] Asaf Salamov,[1] Daren W. Brown,[2] Laszlo G. Nagy,[3] Dimitris Floudas,[3] Benjamin Held,[4] Anthony Levasseur,[5] Vincent Lombard,[6] Emmanuelle Morin,[7] Robert Otillar,[1] Erika Lindquist,[1] Hui Sun,[1] Kurt LaButti,[1] Jeremy Schmutz,[1] Dina Jabbour,[8] Hong Luo,[8] Scott E. Baker,[9] Antonio Pisabarro,[10] Jonathan D. Walton,[8] Robert Blanchette,[4] Bernard Henrissat,[6] Francis Martin,[7] Dan Cullen,[11] David Hibbett,[3] and Igor V. Grigoriev[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]USDA, Peoria, Illinois; [3]Department of Biology, Clark University, Worcester, Massachusetts; [4]University of Minnesota, St. Paul, Minnesota; [5]Institut National de la Recherche Agronomique, Aix-Marseille Université, Marseille, France; [6]Centre National de la Recherche Scientifique, Aix-Marseille Université, Marseille, France; [7]Institut National de la Recherche Agronomique, Université Henri Poincaré, Interactions Arbres/Micro-organismes, Champenoux, France; [8]DOE Great Lakes Bioenergy Research Center, Michigan State University, E. Lansing, Michigan; [9]Pacific Northwest National Laboratory, Richland, Washington; [10]Depto. de Producción Agraria, Univ. Pública de Navarra, Pamplona, Spain; [11]USDA Forest Products Laboratory, Madison, Wisconsin

Basidiomycota (basidiomycetes) make up 32% of the described fungi and include most wood decaying species, as well as pathogens and mutualistic symbionts. Wood-decaying basidiomycetes have typically been classified as either white rot or brown rot, based on the ability (in white rot only) to degrade lignin along with cellulose and hemicellulose. Prior genomic comparisons suggested that the two decay modes can be distinguished based on the presence or absence of ligninolytic class II peroxidases (PODs), as well as the abundance of enzymes acting directly on crystalline cellulose (reduced in brown rot). To assess the generality of the white rot/brown rot classification paradigm we compared the genomes of 33 basidiomycetes, including four newly sequenced wood decayers, and performed phylogenetically-informed Principal Components Analysis (PCA) of a broad range of gene families encoding plant biomass-degrading enzymes. The newly sequenced *Botryobasidium botryosum* and *Jaapia argillacea* genomes lack PODs, but possess diverse enzymes acting on crystalline cellulose, and they group close to the model white rot species *Phanerochaete chrysosporium* in the PCA. Furthermore, laboratory assays showed that both *B. botryosum* and *J. argillacea* can degrade all polymeric components of woody plant cell walls, a characteristic of white rot. We also found expansions in reducing polyketide synthase genes specific to the brown rot fungi. Our results suggest a continuum rather than a dichotomy between the white rot and brown rot modes of wood decay. A more nuanced categorization of rot types is needed, based on an improved understanding of the genomics and biochemistry of wood decay.

Understanding the factors shaping microbial community structure within root and rhizosphere microbiome of *Populus* Species

M.S. Robeson II,[1] M. Shakya,[2] Z.K. Yang,[1] M.Kerley,[1] S.R. Jun,[1] A.A. Gorin,[1] M. Podar,[1] G.M. Bonito,[3] J. Labbe,[1] S. Tringe,[4] W. Muchero,[1] D.A. Pelletier,[1] G.A. Tuskan,[1] R. Vilgalys,[5] and **C.W. Schadt*** (schadtcw@ornl.gov)[1]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [2]Department of Biological Sciences, Dartmouth College Hanover, New Hampshire; [3]National Herbarium, Royal Botanic Gardens, Melbourne, Victoria/Australia; [4]Joint Genome Institute, Walnut Creek, California; [5]Biology Department, Duke University, Durham, North Carolina

Populus spp. are a genetically diverse genus of tree species, broadly distributed across temperate environments of North America, that have become important model species. Fast growth rates, clonal propagation, and the ease with which clones and hybrids can be grown on land unsuitable for food production, have also made them candidate feedstocks for pulp and bioenergy. It is thus important to understand the role of the microbiome as it relates to the health and productivity of *Populus*. As part of the Plant Microbe Interfaces project (PMI – http://PMI.ornl.gov), we have been examining the root, rhizosphere and soil communities of *Populus deltoides* in riparian habitats in the eastern U.S., as well as common garden populations of *P. trichocarpa* in the western U.S. Microbiome data for both bacteria and fungi are analyzed against corresponding soil, geographic, tree phenotype, and tree genotype data in order to understand how such properties influence microbiome structure.

In *P. deltoides* we have shown that the rhizosphere and endosphere environments feature highly developed, diverse and often exclusive communities of bacteria and fungi. Endophytic bacterial diversity is highly variable, but approximately tenfold lower than the rhizosphere, suggesting root tissues provide distinct environments supporting less species. The endosphere is more heavily dominated by Actinobacteria and γ-Proteobacteria when compared with the rhizosphere. *Populus* spp. roots appear to be highly enriched for *Pseudomonas fluorescens*-like species/OTUs when compared to the *Populus* rhizosphere as well as surrounding (non-*Populus*) tree species. Fungal endophytic species are more numerous than bacterial endophytes, but also less diverse than rhizosphere fungi. Both fungal and bacterial rhizosphere samples show distinct phylogenetic composition patterns compared to the more variable endophyte samples. Contrary to expectations, *Populus* spp. have low levels of colonization by ectomycorrhizal fungi (EMF) and arbuscular mycorrhizal fungi (AMF), but high levels of presumed fungal endophytic taxa such as *Nectria*, *Mortierrella*, and *Atractiella*. Several isolates of *Mortierrella* and *Atractiella* have been obtained and shown to have growth promoting effects.

Overall, the *P. trichocarpa* soil and rhizosphere rRNA communities studied to date in the western U.S. separate based on the two Oregon common gardens (Clatskanie and Corvallis) were clearly divergent. Conversely, the endophyte communities between the common gardens were more similar to each other but highly variable from sample to sample. *Pseudomonas* OTUs also dominated both the rhizosphere and endophyte samples, while being rare within bulk soil. Finally, the effects of *P. trichocarpa* genotype on the composition of its rhizosphere appears to be limited when compared to the effects of local soil environment; while endophyte influences are less clear. Additionally we have recently projected functional community profiles based on the KEGG ontology terms of sequenced representatives most closely related to our OTUs. These projected functional profiles also show similar, but often stronger, site and habitat divergence patterns.

Future research plans will 1) expand both the diversity and geographic range of species of *Populus* examined, 2) move beyond the rooting zone to total microbiome studies of *Populus,* as well as 3) include additional functional analyses enabled by developments in metagenomics.

pH stability improvement of a high redox-potential fungal peroxidase based on the analysis of stable peroxidases identified in genomes

V. Sáez-Jiménez, E. Fernández-Fueyo, F.J. Medrano, A. Romero, A.T.Martínez, and F.J. Ruiz-Dueñas* (FJRuiz@cib.csic.es)

CIB, CSIC, Ramiro de Maeztu 9, E-28040 Madrid, Spain

Versatile peroxidase (VP) from the white-rot fungus Pleurotus eryngii is a high-redox potential peroxidase of biotechnological and environmental interest due to its ability to oxidize a wide range of recalcitrant substrates, including phenolic and non-phenolic aromatic compounds and dyes. However, its relatively low stability is a drawback for its industrial application. During the screening of the sequenced Pleurotus ostreatus genome, the highly pH-stable manganese peroxidase (MnP) isoenzyme-4 was identified and crystallized (Fernández-Fueyo et al., 2014). The analysis of the MnP4 crystal structure revealed the presence of some H-bonds, salt bridges and solvent exposed charged residues that could have a stabilizing effect and contribute to the high pH stability of this enzyme. With the aim of improving the pH-stability of versatile peroxidase, the possible stabilizing motifs found in MnP4 were introduced in VP through site-directed mutagenesis, giving rise to three VP variants. The VP150 variant included eight residues found in MnP4 responsible for creating in this enzyme H-bonds and salt bridges that reinforce its structure. VP150 showed a wider range of pH in which it is stable, compared with native VP, both at acid and alkaline pH. In parallel, seven basic residues exposed to the solvent and putatively contributing to stabilize MnP4 at acid pH were also identified. We introduced these basic residues in VP150, obtaining the VP153 variant. Finally, the VP154 variant was designed by introducing the double A49C/A61C mutation in VP150. These two new cysteines formed a new disulfide bridge near the distal calcium binding site stabilizing a long external loop at this region. The VP153 and VP154 variants exhibited improved stability at acidic pH, retaining 85% of the initial activity at pH 3.5 after 25 h incubation at 25ºC, whereas VP150 only retained 60%, and native VP was fully inactivated. The structural motifs found in MnP4, such as H-bonds and salt bridges, as well as the extra disulfide bridge introduced in VP reinforced the structure of the enzyme improving its pH-stability. Moreover, the introduction in VP of solvent exposed basic residues found in MnP4 also enhanced its acid stability. These results show the possibility of using structural motifs identified in large genomic analyses for the design of biocatalysts of interest based on high-redox potential peroxidases. This work was supported by the PEROXICATS and INDOX grants of the European Union and by the HIPOP grant of the Spanish MINECO.

Fernández-Fueyo, E., Ruiz-Dueñas, F.J., Martínez, M.J., Romero, A., Hammel, K.E., Medrano, F.J., Martínez, A.T., 2014. Ligninolytic peroxidase genes in the oyster mushroom genome: Heterologous expression, molecular structure, catalytic and stability properties and lignin-degrading ability. Biotechnol. Biofuels 7:2.

dipSPAdes: Assembler for highly polymorphic diploid genomes

Yana Safonova* (yana.safonova@gmail.com),[1] Anton Bankevich,[1] and Pavel A. Pevzner[1,2]

[1]Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg, Russia; [2]Department of Computer Science and Engineering, University of California at San Diego, San Diego, California

Assembly of highly polymorphic (HP) diploid genomes is a complex computational problem. When two haplomes are very similar, e.g., as human haplomes that differ from each other by only ≈ 0.1% of

nucleotides, both haplomes are usually assembled as a single reference genome (with further analysis of SNPs). Assembling found SNPs into human haplomes is a difficult but well studied problem. We address an even more challenging problem of assembling haplomes that differ from each other by 0.4–10%. The standard assembly approaches fail to reconstruct individual haplomes in HP genomes; moreover, it is not clear whether the algorithms proposed for human haplome assembly can contribute to assembling HP genomes.

Assembly of a diploid genome can result in two types of contigs: haplocontigs (contigs representing both haplomes) and consensus contigs (representing a consensus of both haplomes for the orthologous regions). Consensus contigs do not adequately represent haplomes but are rather a mosaic of segments from both haplomes. Thus, in each polymorphic site of a diploid genome, the alleles present in a consensus contig are somewhat randomly chosen from one of haplomes. In practice, since some regions of HP genome are less polymorphic than others, conventional assemblers generate a mixture of haplocontigs and consensus contigs while assembling HP genomes.

We present dipSPAdes, a new algorithm for assembling HP genomes. dipSPAdes uses the de Bruijn graph constructed by SPAdes assembler to generate both consensus and haplocontigs. Instead of analyzing contigs or long read alignments (as in the previous approaches), we use the de Bruijn graph to mask polymorphism in contigs and to produce a more comprehensive representation of the genome by both consensus contigs and haplocontigs. The question how accurate are these assemblies constructed in the pre-NGS era remains open since there is no gold standard for checking the validity of HP assemblies. To provide the first comprehensive benchmarking of HP assemblies, we took advantage of a unique dataset generated in the course of a recent massive effort to sequence 37 genomes of S. commune conducted in Dr. Alexey Kondrashov laboratory at Moscow State University. S. commune is a model organism (wood-degrading mushroom) whose genome is ideally suited for benchmarking HP genome assemblers. The unique feature of the haploid S. commune is that two different organisms differ by ≈ 10%. Thus, combining reads from two S. commune genomes perfectly models an HP genome, yet allowing one to test the quality of assembly, the bottleneck in previous studies of assembly algorithms for diploid genomes. Benchmarking of dipSPAdes on both simulated and real fungi datasets demonstrated that dipSPAdes significantly improves assemblies of HP genomes. dipSPAdes is also an excellent comparative assembler that can be used to generate a consensus assembly of multiple similar genomes.

Sympathy for the ligule (*sol*) dominant modifier of liguleless narrow (*lgn*)

Brian St. Aubin* (staubinb@berkeley.edu)[1] and Sarah Hake[1,2]

[1]University of California at Berkeley, Berkeley, California; [2]PGEC ARS-USDA

Changes in the development of leaves in maize can have drastic impacts on the viability of the entire plant. We are interested in understanding the genes which regulate leaf development. Heterozygous *Liguleless narrow* mutants have shorter, narrower leaves, however in some Inbred lines the heterozygous mutants significantly recover. Learning what kind of gene is causing the recovery could also shed light on the developmental process of leaf formation in the grasses. Monitoring differential gene expression in an expected region can lead to localizing a new gene faster than by traditional mapping alone. Performing this investigation will also reveal other genes that may be affected by this developmental pathway making future research faster and better directed. Identification of which gene is causing the recovery could have impacts on the ability to produce hardy crops.

Along with traditional mapping of recombinants, RNAseq data from the four genotypes using illumina sequencing will be used to assess the impacts of the mutation and modifier on the whole plant. By selecting tissue from the shoots and roots for plants with and without the modifier, and with and without *lgn,* a profile of gene expression based on genotype will be created. Roots and shoots were studied based on previous expression data for sister of liguleless narrow (sln) a homolog for lgn, and the fact that many developmental genes are differentially expressed between the shoots and roots.

Genomic and transcriptomic definition of a methylglucuronoarabinoxyan-utilization regulon in *Paenibacillus* sp. JDR-2

Neha Sawhney[1], Franz J. St. John[2], and James F. Preston* (jpreston@ufl.edu)[1]

[1]Department of Microbiology and Cell Science, University of Florida, Gainesville, Florida; [2]U.S. Forest Service, Forest Products Laboratory, Madison, Wisconsin

Paenibacillus sp. JDR-2 (Pjdr2), an aggressively xylanolytic bacterium, secretes a multi-modular cell-associated GH10 endoxylanase (XynA1) which catalyzes depolymerization of methylglucuronoxylan ($MeGX_n$) and rapidly assimilates the products of depolymerization (St. John et al., 2006, http://aem.asm.org/content/72/2/1496.full). Coordinate regulation of gene expression has led to the description of a xylan-utilization regulon encoding proteins, including an intracellular GH67 α-glucuronidase, GH10 endoxylanase and GH43 xylosidase that contribute to these processes (Chow et al., 2007, http://jb.asm.org/content/189/24/8863.full). The efficient utilization of the polysaccharide compared to the oligosaccharide products of depolymerization has been postulated to result from the coupling of the processes of extracellular depolymerization, assimilation, and intracellular metabolism (Nong et al., 2009, http://aem.asm.org/content/75/13/4410.full). *Paenibacillus* sp. JDR-2 shows growth and substrate-utilization patterns with sorghum methylglucurono-arabinoxylan ($MeGAX_n$) as a carbon source similar to those observed with $MeGX_n$ as substrate. The preferred utilization of polymeric $MeGAX_n$ compared to arabinoxyloligosaccharides supports a process in which the depolymerization and assimilation processes are coupled. The co-expression of a cluster of genes encoding an intracellular GH51 α-L-arabinofuranosidase and transcriptional regulator genes is coordinate with the genes encoding the xylan-utilization regulon , supporting the role of these in $MeGAX_n$ utilization. The sequenced genome of *Paenibacillus* sp. JDR-2 (http://www.standardsingenomics.org/index.php/sigen/article/view/sigs.2374349/651) has identified candidate gene clusters for this process that encode a secreted multimodular GH10 endoxylanase containing SLH domains for surface anchoring along with genes potentially encoding additional enzymes for the utilization of $MeGAX_n$. Transcriptomic analysis (RNA seq) performed by JGI has identified these genes as collectively up-regulated for the utilization of $MeGAX_n$, allowing their functional assignments to a $MeGAX_n$-utilization regulon. This GH10/GH67 system for efficient xylan processing has now been expanded to include an intracellular α-L-arabinofuranosidase as well as transcriptional regulators for the assimilation of the products generated by a cell-associated multi-modular endoxylanase. This system, the genes for which occur in certain bacteria, e.g. *Clostridium cellulolyticum* H10, *Caldicellulosipruptor saccharolyticus* DSM 8903, *Thermoanaerobacterium saccharolyticum* JW/SL-YS485, *Thermotoga maritima* MSB8, *Geobacillus sp*. Y412MC10 and other species, may be one best suited for developing biocatalysts for the efficient conversion of lignocellulosics of monocots as well as dicots to biofuels and chemicals.

Discovery of functional toxin/antitoxin systems in bacteria by shotgun cloning

Hila Sberro* (hsberro@gmail.com)[1], **Azita Leavitt**[1], **Ruth Kiro**[2], Eugene Koh[1], Yoav Peleg[3], Udi Qimron[2], and Rotem Sorek[1]

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel; [2]Department of Clinical Microbiology and Immunology, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel; [3]Israel Structural Proteomics Center, Faculty of Biochemistry, Weizmann Institute of Science

To survive in the face of constant phage attacks, bacteria have developed a variety of anti-phage defense mechanisms. One of the documented defense strategies is abortive infection (Abi), where the infected bacterial cell commits "suicide", thus preventing phage spread and enabling colony survival. Several reports show that Abi can work via a pair of stable "toxin" and unstable "antitoxin" that prevents the toxic effect of the toxin but is degraded upon phage infection. Although several families of toxin/antitoxin (TA) modules were found in multiple bacterial genomes, the extent and diversity of TA module types and their roles in phage defense are not yet clear. To systematically discover new types of defense TA modules in an unbiased manner we analyzed the experimental insertion of 1.5 million genes from 388 microbial genomes into an E. coli host using over 8.5 million random clones. This analysis detected genes (toxins) that could only be cloned when the neighboring gene (antitoxin) was also present on the same clone. Clustering of these genes revealed TA families widespread in bacterial genomes, some of which deviate from the classical characteristics previously described for such modules. Introduction of these genes into *E. coli* validated that the toxin toxicity is mitigated by the antitoxin. Infection experiments with T7 phage showed that two of the new modules can provide resistance against phage. Moreover, our experiments exposed an 'anti-Abi' mechanism in T7 that neutralizes bacterial suicide by inhibiting a protease essential for antitoxin degradation.

Preliminary predictive models of microbial community interactions during the Deepwater Horizon Oil Spill

Nicole M. Scott* (nicolescott1@gmail.com; nscott1@uchicago.edu)[1,2]*, Tingting Yang[3], Jarrad Marcell[1], Tinfani Anton[1], Andreas P. Teske[3], and Jack A. Gilbert[1,2]

[1]Ecology and Evolutionary Biology, University of Chicago, Chicago, Illinois; [2] Institute for Genomic and Systems Biology, Argonne National Laboratory, Lemont, Illinois; [3]Department of Marine Sciences, University of North Carolina, Chapel Hill, North Carolina

The Deepwater Horizon oil spill, with its complex hydrocarbon plume, created spatiotemporal shifts in microbial community structure as microorganisms dynamically responded to heterogeneous and transient nutrient sources. The relationship between microbial community structure and geochemical markers allows us to model these dynamical relationships through time. Using Microbial Assemblage Prediction (MAP), we have previously modeled the web of community interactions as an artificial neural network (ANN) derived from metabolic predictions, taxonomic data, and environmental data to create models that predict the relative abundance of microbial communities through time and space as a function of environmental conditions. Here we modify the MAP heuristics to create models that can be used predict environmental or physical-chemical parameters using the relative abundance profile of microbial taxa. We hypothesize that our models will validate previously published spatiotemporal trends and may better explain the ecology of these trends.

To test our hypothesis, a matrix of the relative abundance of operational taxonomic units (OTU) and associated metadata (detailing oil chemistry and basic nutrient measurements) from 65 contained and un-contaminated sediment samples was used to generate a Bayesian network capturing taxa-environment conditional dependencies. Mathematical descriptions of the conditional dependencies between nodes are calculated based on these data, and used to train a set of ANNs. These models capture the niche relationships between bacterial taxa and the changing environmental conditions in the GoM. From these models we are able to predict oil chemistry and nutrient parameters as a function of the observed microbial community structure, which we validated using measured chemistries from a subset of samples not used to train the model.

This novel research leverages the successive states of the microbial community degradation of polluting hydrocarbons to predict environmental and functional parameters. Such models are necessary to determine how changes in microbial community assemblage can influence remediation and nutrient cycles (e.g. N and C) in the GoM and other regions affected by oil pollution.

Multi-kb Illumina reads reveal significant strain variation and rare organisms in aquifer sediment microbial communities

Itai Sharon* (itaish@berkeley.edu),[1] Laura A. Hug,[1] Michael Kertesz,[2] Dmitry Pushkarev,[2] Timothy A. Blauwkamp,[2] Kyle R. Frischkorn,[1] Mojgan Amirebrahimi,[3] Brian C. Thomas,[1] Kenneth H. Williams,[4] Susannah G. Tringe,[3] and Jillian F. Banfield[1,4]

[1]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, California; [2]Illumina Inc., Hayward, California; [3]Department of Energy, Joint Genome Institute, Walnut Creek, California; [4]Lawrence Berkeley National Laboratory, Berkeley, California

Genomic analysis of complex microbial communities in environments such as soils and sediments remains challenging, despite recent advancements in DNA sequencing technologies. Here we used metagenomic approaches to study microbial communities in organic-rich sediments collected from three depths (4, 5 and 6 m) in an aquifer adjacent to the Colorado River, USA. The communities contain thousands of different species, the vast majority of which have no close relatives with sequenced genomes. In addition to using the Illumina HiSeq platform we explored the use of the new Moleculo platform, which generates long multi-kb reads. Overall, ~200 Gb HiSeq and ~1.5 Gbp Moleculo data were sequenced for the three samples at median read length of 150 and ~8,000 bps, respectively. We assembled the Moleculo data and used it to scaffold the short read assemblies. Relatively low levels of assembly and limited scaffolding indicates extremely high species richness and community evenness.

To evaluate the quality of the Moleculo data and genomes assembled from metagenomic short read data we compared the assembled Moleculo reads to the 2.1 Mb genome of RBG-1, previously assembled from the short read data of the 6 m sample. Inconsistencies were found for 24/210 Moleculo reads and contigs assigned to the RBG-1 genome, of which 22 were the result of either strain variation or errors in the Moleculo data, and 2 resulted from local mis-assemblies in the RBG-1 genome.

Clustering of the Moleculo reads based on read overlap revealed large clusters of reads in the 4 and 5 m samples that represent closely related species and strains. By evaluating the synteny of genes on these reads we could reconstruct most of the genome architecture for the most abundant operational taxonomic unit (OTU) in the two samples. The OTUs belong to Deltaproteobacteria (5 m) and candidate phylum OP8 (4 m), and have no closely related assembled genomes. Intriguingly this abundant, heterogeneous strain mixture had very little representation in the short read assemblies.

Between 40 % and 65 % of the Moleculo reads had less than 3x coverage by the short read data, therefore represent genomes that could not be assembled using these data. Taxonomic evaluation of the relatively low abundance microbes revealed "more of the same": almost all genomes were from lineages represented by more abundant genotypes, e.g., from the Chloroflexi, OD-1, and OP11 phyla and Deltaproteobacteria class. Overall, the long read data provide insight into the biology of both rare and abundant organisms that currently cannot be achieved using short read data.

RNA-Seq gene expression analysis at the Joint Genome Institute

Vasanth Singan* (VRSingan@lbl.gov), Anna Lipzen, James Han, and Erika Lindquist

DOE Joint Genome Institute, Walnut Creek, California

Transcriptome sequencing and analysis spans all scientific programs at the JGI. Gene expression profiling of transcriptomes remains a core competency at the JGI. Several open source Differential Gene Expression (DGE) tools (Cufflinks, DESeq and EdgeR) were analyzed in an effort to improve the RNA-Seq gene expression analysis pipeline for Eukaryotic projects. The outcome of the analyses, current and new deliverables to users and the schematic of the analysis pipeline are covered in this poster.

The current pipeline maps reads to a reference transcriptome and uses an in-house tool based on bwa as the aligner to perform gene counts. The new pipeline utilizes the genome sequence as the reference and a splice-aware aligner (TopHat) to map the reads to the genome. DGE analysis was performed using the DESeq2 package, which has shown improved capability to detect certain classes of differentially expressed genes and increased ability to utilize replicate information compared to other tools including Cufflinks (cuffdiff) and EdgeR. Cufflinks seemed to miss genes that have an on/off state between conditions, which DESeq2 was able to identify. Similarly DESeq2 showed superior performance in use of replicate variability to identify differential express genes, thereby influencing our choice of tools. The format of the new summary table providing the log2 fold-change, p-value and differential expression significance for each pair of conditions across all the annotated genes is discussed in detail. The new pipeline also provides additional outputs to visualize library contamination and sample replicate correlations, which is discussed in this poster.

High-resolution phylogenetic microbial community profiling

Esther Singer* (esinger@lbl.gov),[1] Devin Coleman-Derr,[1] Brett Bowman,[2] Patrick Schwientek,[1] Alicia Clum,[1] Alex Copeland,[1] Doina Ciobanu,[1] Jan-Fang Cheng,[1] Susannah Tringe,[1] and Tanja Woyke[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]Pacific Biosciences, Menlo Park, California

The representation of bacterial and archaeal genome sequences is strongly biased towards cultivated organisms, which belong to merely four phylogenetic groups. Functional gene information and inter-phylum level relationships are still largely underexplored for candidate phyla, which are often referred to as 'microbial dark matter'. Furthermore, a large portion of the 16S rRNA gene records in the GenBank database are labeled as "environmental samples" and "unclassified", which is in part due to low read accuracy, potential chimeric sequences produced during PCR amplifications and the low resolution of short amplicons. In order to improve the phylogenetic classification of novel species and advance our knowledge of the ecosystem function of uncultivated microorganisms, high throughput full-length 16S rRNA gene sequencing methodologies with reduced biases are needed. We evaluated the performance

of PacBio single-molecule real-time (SMRT) sequencing in high-resolution phylogenetic microbial community profiling. For this purpose, we compared PacBio and Illumina metagenomic shotgun and 16S rRNA gene sequencing of a mock community as well as of an environmental sample from Sakinaw Lake, British Columbia. Sakinaw Lake is known to encompass a high abundance of microbial species from candidate phyla. Sequencing results show that community structure based on PacBio shotgun and 16S rRNA gene sequences is highly similar in both, the mock and the environmental community. Resolution power and community representation accuracy from SMRT sequencing data appeared to be independent of %GC content of microbial genomes and was significantly higher when compared to Illumina-based metagenome shotgun and 16S rRNA gene (iTag) sequences, *e.g.* full-length sequencing resolved all 23 OTUs in the mock community, while iTags could not resolve closely related species. SMRT sequencing hence offers various potential benefits when characterizing uncharted microbial communities.

Genomics of Candidatus *Synechococcus spongiarum*, a cyanobacterial sponge symbiont

Beate M. Slaby* (beate.slaby@uni-wuerzburg.de),[1] Alex Copeland,[2] Tanja Woyke,[2] and Ute Hentschel[1]

[1]Department of Botany II, Julius-von-Sachs Institute of Botanical Sciences, University of Würzburg, Germany; [2]DOE Joint Genome Institute, Walnut Creek, California

Marine cyanobacteria of the genus *Synechococcus* are important contributors to the global carbon cycle and are major primary producers in the world's oceans. In addition to various free-living *Synechococcus* species, also termed 'sea water ecotype', there is a 'sponge ecotype', namely Candidatus *Synechococcus spongiarum*, which is a widespread and abundant symbiont of marine sponges. These ecotypes are closely related, but phylogenetically distinct as a result of occupying different ecological niches. The question arises, which genomic differences account for the existence of two different ecotypes, one free-living, and one marine sponge-associated. A combination of single-cell and metagenomics is implemented to access the genomic repertoire of Candidatus *Synechococcus spongiarum*, the so far uncultivated cyanobacterial symbiont of the Mediterranean sponge *Aplysina aerophoba*. At the DOE JGI sequences (Illumina HiSeq) of FACS sorted single-cells and metagenomes of Mesohyl as well as Pinacoderm, inner and outer sponge tissue respectively, are created. Different cell lysis methods (bead beating, freeze-thaw cycles, and proteinase K digestion) are implemented for metagenomic DNA extraction to gain differential coverage, resulting in a total of 6 metagenome and iTag datasets. By binning of the metagenomic datasets combined with mapping onto the single cell data, we are aiming to reconstruct a consensus genome of Candidatus *Synechococcus spongiarum*. It's comparison with genomes of planktonic *Synechoccous* sp. will give insights towards cyanobacterial adaptations to an existence within marine sponges.

Relating the abundance, activity and physiological diversity of nitrifying microorganisms to rates of nitrogen transformation in the coastal ocean

Jason M. Smith* (jsmith@mbari.org)[1] and Christopher A. Francis[2]

[1]Biological Oceanography Group, Monterey Bay Aquarium Research Institute, Moss Landing, California; [2]Environmental Earth System Science, Stanford University, Stanford, California

Nitrification, the microbial oxidation of ammonium to nitrate, is a central part of the oceanic nitrogen cycle. In the ocean's surface layer, the process alters the distribution of inorganic nitrogen species available to phytoplankton and produces N_2O. There is widespread consensus that the first step in nitrification, the oxidation of ammonium to nitrite, is carried out by the ammonia oxidizing archaea (AOA) throughout the ocean. Our previous work in Monterey Bay (MB) and the contiguous California Current System (CCS) has demonstrated a strong relationship between the abundance of archaeal ammonia monooxygenase (*amoA*) genes and transcripts and rates of nitrification. Our community sequencing project, currently in progress, seeks to determine whether variability in the the abundance, genomic diversity or cellular activity of ammonia oxidizing archaea relates to observed changes in rates of nitrogen transformation, or the prevailing 'metabolic state' of a given water mass in MB or the CCS. Here, we present an initial analysis of the relationship between the abundance and diversity of Marine Group I (MGI) 16S rRNA gene abundances in relation to observed environmental variability and rates of nitrification rates in the coastal ocean. Across a composite library of 94 samples and more than 9 million 16S rRNA reads, we observed a stinking level of agreement between the relative abundance of MGI 16S rRNA reads (as % of total) and the absolute abundance of AOA *amoA* genes, ascertained from quantitative PCR. Further, relative changes in the abundance of MGI 16S rRNA genes both with depth and between stations were in close agreement with changes in nitrification rates. Together with our previous work, these data suggest tag sequence data to be a powerful tool for the study of AOA abundance and diversity, and the distribution of nitrification rates, in the ocean.

Controlling mycotoxin contamination in maize via host-induced gene silencing

Jonathon E. Smith*, Yamunarani Ramegowda, and Burt H. Bluhm (bbluhm@uark.edu)

Department of Plant Pathology, University of Arkansas, Division of Agriculture, Fayetteville, Arkansas

Aspergillus flavus and *Fusarium verticillioides* are important ear rot pathogens of maize. Although *A. flavus* and *F. verticillioides* reduce yields and grain quality, the principal concern associated with these pathogens is the production of mycotoxins. Aflatoxins (produced by *A. flavus*) and fumonisins (produced by *F. verticillioides*) are distinct groups of polyketide-derived mycotoxins that pose significant health risks to humans and livestock. Despite the health and economic concerns associated with aflatoxins and fumonisins, few effective tools are available to prevent grain contamination. Host-induced gene silencing (HIGS) was recently demonstrated as a means of manipulating gene expression in pathogens through the expression of pathogen-specific hairpin RNA (hpRNA) in the host plant. The goal of this research was to generate transgenic maize plants expressing pathogen-specific hpRNA to inhibit mycotoxin accumulation. To this end, a pipeline was created to develop and evaluate HIGS vectors. First, fungal genes that regulate mycotoxin biosynthesis were selected for silencing. Second, a high-throughput cloning process was developed in which a single ligation step was used to generate constructs containing a sense and antisense fragment of the target gene separated by an intron from

the *Magnaporthe oryzae* cutinase gene. The constructs were then transformed into *A. flavus* or *F. verticillioides* and evaluated for reduced toxin accumulation. Finally, hpRNA constructs that significantly lowered mycotoxin levels were used to generate plant expression vectors that were then transformed into maize. Currently, vectors have been created to silence expression of aflatoxin biosynthesis-specific genes in *A. flavus*, including the polyketide synthase gene (*aflC*), a fatty acid synthase gene (*aflA*), and a regulatory gene (*aflS*). A vector was also created to silence expression of the polyketide synthase gene (*FUM1*) required for fumonisin biosynthesis in *F. verticillioides*. In addition, silencing vectors have been generated to target non-cluster genes that regulate mycotoxin biosynthesis, such as genes encoding hexokinase (*hxkA*) in *A. flavus* and α-amylase (*AMY1*) in *F. verticilliodes*. Through the creation and evaluation of transgenic maize plants expressing pathogen-specific hpRNA, this research will advance the current understanding of HIGS and potentially provide a novel means of controlling mycotoxin contamination in maize.

Identification and regulation of novel cellulases within anaerobic gut fungi

Kevin V. Solomon* (ksolomon@engineering.ucsb)[1], **John K. Henske[1*]**, Charles H. Haitjema[1], Diego Borges-Rivera[2], Dawn A. Thompson[2], Aviv Regev[2], and Michelle O'Malley[1] (momalley@engineering.ucsb.edu)

[1]Department of Chemical Engineering, University of California at Santa Barbara, Santa Barbara, California; [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts

The economical breakdown of plant biomass into simple sugars remains a significant bottleneck in the production of renewable fuels via microbial fermentation. Existing technologies, however, are insufficient for industrial-scale production of biofuels due to the recalcitrance of lignin-rich biomass, and the high cost/poor performance of known cellulolytic enzymes. Therefore, there is a critical need to develop new technologies to break down crude biomass into fermentable sugars for downstream fuel development. Towards this goal, much can be learned by studying how anaerobic gut fungi depolymerize lignocellulose in biomass-rich environments, such as the digestive tract of large herbivores. Anaerobic gut fungi are native to the gut and rumen of these animals, where they have evolved unique abilities to break down lignocellulosic biomass through invasive growth and the secretion of powerful enzymes and enzyme complexes (cellulosomes). Towards engineering gut fungi as novel platform organisms for biofuel production, we have isolated a panel of unique specimens and are developing bioinformatics pipelines that have characterized the biomass-degrading machinery of a novel *Piromyces* isolate from transcriptomic and proteomic studies. We are collaborating with the JGI and EMSL through the Joint EMSL/JGI user program to extend this approach to novel *Neocallimastix* and *Anaeromyces* isolates. Each isolate thrives on lignocellulosic substrates, and secretes multi-protein cellulosome complexes of cellulases, hemicellulases, and cellulose binding domains that can be elucidated by transcriptomic and genomic sequencing. To determine the basic metabolic networks that govern biomass hydrolysis within anaerobic fungi, we are employing RNAseq at the JGI to quantify transcript abundance when simple sugars repress lignocellulosic degradation. Parallel proteomics efforts are underway at EMSL to compare the regulation of secreted fungal enzymes with those regulated at the transcriptional level. Through these efforts, we will determine how important enzyme groups are coordinated during biomass breakdown across fungal genera. Collectively, this information will establish the molecular framework for anaerobic fungal hydrolysis, and will guide in the development of lignocellulosic biofuels.

Identification of novel genes associated with pathogenicity and mycotoxin biosynthesis in *Fusarium verticillioides*

Sreenivasa, M.Y.* (smarikun@uark.edu),[1,2]John B. Ridenour,[1] and Burt H. Bluhm[1] (bbluhm@uark.edu)

[1]Department of Plant Pathology, University of Arkansas, Division of Agriculture, Fayetteville, Arkansas; Department of Microbiology, University of Mysore, Manasagangotri, Karnataka, India

Fusarium verticillioides is an important pathogen of maize and sorghum in all stages of development, from the early hours of seed germination to the time of harvest, and also causes post-harvest decay of grain. When *F. verticillioides* colonizes grain, there is a significant risk of contamination with fumonisins, a group of polyketide-derived toxins. Fumonisins cause chronic and acute diseases in humans and animals and are classified as group 2B carcinogens by the International Agency for Research on Cancer (IARC). With the complete genome of *F. verticillioides* and an expression sequence databases on hand, the next major target is to identify novel genes involved in pathogenesis. In this study, forward and reverse genetic approaches are being applied to identify novel genes associated with pathogenicity and mycotoxin biosynthesis in *F. verticillioides*. For the forward genetic approach, a collection of insertional mutants is being created via Restriction Enzyme Mediated Integration (REMI) with a novel activation-tagging cassette. Mutants that display altered levels pathogenicity and/or mycotoxin biosynthesis will be selected for further analysis. For the reverse genetic component, four putative regulatory genes are being targeted for deletion, which will elucidate their role in growth, morphology, pathogenesis, stress responses and mycotoxin biosynthesis. The combination of forward and reverse genetic approaches provides an efficient approach to clarify the genetic network underlying infectious development, virulence and pathogenicity. The significance of this work is that identification of novel fungal genes involved in pathogenesis will inform future efforts to improve genetic and biological control of *F. verticillioides* in maize and sorghum.

Population dynamics of sequence-discrete bacterial populations inferred using metagenomes

Sarah Stevens* (sstevens2@wisc.edu),[1] Matthew Bendall,[2] Dongwan Kang,[2] Jeff Froula,[2] Leong-Keat Chan,[2] Susannah Tringe,[2] Rex Malmstrom,[2] and Katherine McMahon[1]

[1]University of Wisconsin-Madison, Department of Bacteriology, Madison, Wisconsin; [2]Department of Energy Joint Genome Institute, Walnut Creek, California

From a multi-year metagenomic time series of two dissimilar Wisconsin lakes we have assembled dozens of genomes using a novel approach that bins contigs into distinct genome based on sequence composition, e.g. kmer frequencies, and contig coverage patterns at various times points. Next, we investigated how these genomes, which represent sequence-discrete bacterial populations, evolved over time and used the time series to discover the population dynamics. For example, we explored changes in single nucleotide polymorphism (SNP) frequencies as well as patterns of gene gain and loss in multiple populations. Interestingly, SNP diversity was purged at nearly every genome position in some populations during the course of this study, suggesting these populations may have experienced genome-wide selective sweeps. This represents the first direct, time-resolved observations of periodic selection in natural populations, a key process predicted by the 'ecotype model' of bacterial diversification.

Resource allocation in hypersaline microbial mats: The role of cyanobacterial Extracellular Polymeric Substances (EPS)

R.K. Stuart* (stuart25@llnl.gov)[1], Stannard, W.[1], Bebout, B.[2], Pett-Ridge, J.[1], and Thelen, M.[1]

[1]Lawrence Livermore National Laboratory, Livermore, California; [2]Exobiology Branch, NASA Ames Research Center, Moffett Field, California

Hypersaline laminated cyanobacterial mats are thought to represent some of the earliest forms of life and orchestrate essentially all the major biogeochemical cycles in a few millimeters. These self-sustaining ecosystems are characterized by high phylogenetic and functional diversity and steep physiochemical gradients on the millimeter scale. The mats at Elkhorn Slough, CA have relevance to biofuels systems biology because they produce significant nighttime fluxes of hydrogen gas and other potential biofuels as fermentation byproducts. To gain a more comprehensive understanding of energy flow within these complex communities will require elucidating the distribution of energy from light absorption into biomass. Extracellular polymeric substances (EPS), which can be composed of proteins, polysaccharides, lipids and DNA are a major component of these mats. We hypothesize that the extracellular matrix plays several important roles in the transfer of energy and metabolites between primary producers and other members of the mat community. Recently, our group identified, isolated and sequenced (CSP # 701) the dominant N-fixing organism from the mats, filamentous cyanobacterium ESFC-1. With the genome for this strain, we are now able to use this organism as a model culture to study mat-building cyanobacteria. We compared EPS from complex natural mat communities from Elkhorn Slough, CA and the model-mat building culture. We used MS-proteomics to interrogate different subcellular fractions, and have identified hundreds of outer membrane and extracellular proteins. The most abundant of the matrix proteins are predicted to be involved in protein degradation, sugar degradation and oxidative stress response. We have also begun to look at changes in abundance and activity of these proteins over a diel cycle. Our findings suggest that cyanobacteria may facilitate carbon transfer to other groups, as well as uptake of their own carbon products through degradation of EPS components.

A fully automated single-cell decontamination pipeline

Kristin Tennessen* (ktennessen@lbl.gov) and **Amrita Pati**

DOE Joint Genome Institute, Walnut Creek, California

Recent technological advancements in single-cell genomics have encouraged the classification and functional assessment of microorganisms from a wide span of the biosphere's phylogeny.[1,2] Environmental processes of interest to the DOE, such as bioremediation and carbon cycling, can be elucidated through the genomic lens of these unculturable microbes.

Contamination can occur at various stages of the single-cell sequencing process. Sources of contamination are from reagents and free DNA that make it into the well, along with the single cell. Screening single-cell datasets for contaminants is currently a very manually-intensive procedure.[3] The processing time for one highly-trained individual to decontaminate one single cell dataset is several hours. A fully automated decontamination tool is necessary to increase throughput of the single-cell sequencing process. Currently, no fully automated decontamination methods exist for single-cells.

We propose the Single Cell Decontamination Pipeline, a fully-automated software tool which classifies unscreened contigs from single cell datasets through a combination of homology and feature-based

methodologies using the organism's nucleotide sequences and known NCBI taxonomy. On average, ninety percent of contaminated contigs are removed. The software is freely available to download and install, and can be run on any system.

1 Rinke C, Schwientek P, Sczyrba A, Ivanova NN,Anderson IJ, Cheng J-F, Darling A, Malfatti S, SwanBK, Gies EA et al. 2013.Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431-437

2 Rex R Malmstrom, Sébastien Rodrigue, Katherine H Huang, Libusha Kelly, Suzanne E Kern, Anne Thompson, Sara Roggensack, Paul M Berube, Matthew R Henn and Sallie W Chisholm. Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis

The ISME Journal (2013) 7, 184–198.

3 Clingenpeel, Scott. JGI Microbial Single Cell Program Single Cell Data Decontamination. http://img.jgi.doe.gov/w/doc/SingleCellDataDecontamination.pdf

Microbial diversity and carbon cycling in San Francisco Bay wetlands

Susanna Theroux* (stheroux@lbl.gov),[1] Wyatt Hartman[1], Shaomei He[2], and Susannah Tringe[1]

[1]DOE Joint Genome Institute, Walnut Creek, California; [2]University of Wisconsin-Madison, Madison, Wisconsin

Wetland restoration efforts in San Francisco Bay aim to rebuild habitat for endangered species and provide an effective carbon storage solution, reversing land subsidence caused by a century of industrial and agricultural development. However, the benefits of carbon sequestration may be negated by increased methane production in newly constructed wetlands, making these wetlands net greenhouse gas (GHG) sources to the atmosphere. We investigated the effects of wetland restoration on belowground microbial communities responsible for GHG cycling in a suite of historic and restored wetlands in SF Bay. Using DNA and RNA sequencing, coupled with real-time GHG monitoring, we profiled the diversity and metabolic potential of wetland soil microbial communities. The wetland soils harbor diverse communities of bacteria, archaea and fungi whose membership varies with sampling location, proximity to plant roots and sampling depth, closely correlating with gradients of electron acceptor availability and methane production. Our results also highlight the dramatic differences in GHG production between historic and restored wetlands and allow us to link microbial community composition and GHG cycling with key environmental variables including salinity, soil carbon and plant species.

Promoting, understanding, recording, and utilizing metadata in genomic/metagenomic studies

Alex Thomas* (alexanderthomas@lbl.gov), T.B.K. Reddy, Michelle Isbandi, Jyothi Mallajosyula, Dimitrios Stamatis, Jonathon Bertsch, and Nikos Kyrpides

DOE Joint Genome Institute, Walnut Creek, California

It is critical to ensure the quality of metadata documentation for genome and metagenome projects to facilitate comparative analysis and hypothesis testing. A project's metadata includes descriptions of taxonomy, physical characteristics, environment, and sequencing. The Genomes OnLine Database (GOLD) is a resource to store and search genome and metagenome projects from around the

world. Specifying a project in GOLD is also required to utilize the analysis tools in the Integrated Microbial Genomes (IMG) data warehouse. The ability to search projects in GOLD depends on the quantity and quality of the project's metadata. Metadata omissions and errors unfortunately occur and require active curation to correct. The omission of metadata can cause searches for sequences to fail, even when the desired sequence is available. Metadata errors can provide misleading information, causing confusion and requiring unnecessary work to resolve. Furthermore, metadata omissions and errors can negatively impact the results of analyses in IMG. Examples of some common metadata omissions and errors are presented, as well as some consequences. Complete and correct metadata examples are also presented, including successful analytical results.

Rapid binning and metabolic profiling of complex microbial community metagenomic data

Brian C. Thomas[1], Andrea Singh[1], Itai Sharon[1], Cindy J. Castelle[1], Laura A. Hug* (laura.hug@berkeley.edu)[1], Chongle Pan[2], and **Jillian F. Banfield**[1,3]

[1]University of California at Berkeley, Berkeley, California; [2]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [3]Lawrence Berkeley National Laboratory, Berkeley, California

Our goal is to develop bioinformatics tools for the management and rapid analysis of microbial community data, including metagenomic and transcriptomic sequences and proteomic profiles. We have developed and tested methods for genome binning, metabolic pathway curation, genome curation, and community composition analysis. These tools guide analysis and interpretation of complex datasets and will be useful for researchers interested in the rapid binning, genome curation, strain analysis, and metabolic assessment of metagenomic data.

The scientific focus of our research is to develop a predictive understanding of microbial ecology and biogeochemistry across a range of environment types. The questions metagenomic datasets can address are multiple: what organisms are present in a community, what is their relative abundance, and how do these change over time and space or in relation to changing physiological conditions. Beyond taxonomic profiling, metagenomics allows prediction of the metabolic potential of the community, including which processes may be occurring and through which intermediates. Transcriptomics and proteomics can be used to confirm the active metabolic processes and identify the key members within a community for nutrient cycling. Datasets are growing in size in line with increases in sequencing capacity, and the communities being examined are likewise scaling in complexity and diversity. Thus, there are major challenges relating to efficiency and accuracy of data analysis. We will report results from JGI CSP, DOE carbon cycle, and JGI ETOP projects. Methods are being tested on microbial communities that vary dramatically in complexity levels and novelty, including acid mine drainage biofilms, infant gut consortia, hydrocarbon-contaminated aquifer samples, riparian zone aquifer groundwater and sediments, and adult human gut microbiome samples.

Unveiling the complex fungal transcriptome with the PacBio® Iso-Seq™ method

Elizabeth Tseng* (etseng@pacificbiosciences.com),[1] Sean Gordon,[2] Jason G. Underwood,[3] Xiandong Meng,[4] Feng Chen,[4] Zhiying Zhao,[4] and Zhong Wang[4]

[1]Pacific Biosciences, Menlo Park, California; [2]USDA-ARS, Western Regional Research Center, Albany, California; [3]Current address: Department of Genome Sciences, University of Washington, Seattle, Washington; [4]DOE Joint Genome Institute, Walnut Creek, California

Alternative isoforms of mRNA greatly increase the proteome diversity that can be encoded by a limited number of genes. Understanding the role of alternative isoforms leads to insights in disease, biodiversity, and evolution. However, isoform discovery with short reads is very challenging leading to underestimation of the true diversity in transcriptomes.

Here we describe Iso-Seq™ method for *de novo* characterization of full-length isoforms using long reads generated by the PacBio RS II. This method produces full-length reads originating from single molecule polyadenylated transcripts that eliminate the need for assembly. We sequenced the fungus *Plicaturopsis Crispa* using the Clontech SMARTer® kit for cDNA library preparation. Size selection was done to create three library sizes: 1-2 kb, 2-3 kb, and 3-6kb. Sequencing was performed on the PacBio RS II using P4-C2 chemistry. After filtering for reads with both the adapters and polyA tail signals, we obtained over 2 million putative full-length transcript sequences. To improve consensus accuracy, we developed an isoform-level clustering algorithm called ICE (Iterative Clustering for Error Correction). Finally, we polished the full-length consensus sequences from ICE using the consensus-calling algorithm Quiver.

We show that using the Iso-Seq method followed by our bioinformatics pipeline outputs full-length transcripts with ≥ 99% accuracy. The longest high-quality, full-length transcript achievable is mainly determined by effectiveness of size selection. In the *P. crispa* sample, we generated full-length transcripts up to 5.5 kb with an average sequence length of 1,657 bp. We were able to detect hundreds of isoforms for some genes, a task not feasible using short reads. Furthermore, we observed widespread transcriptional readthrough, resulting in transcripts containing two or more open reading frames. Independent support from both genome-based predictions and short read data suggest that readthrough transcription is widespread in *P. crispa.*

To conclude, we describe the Iso-Seq method for full-length transcriptome sequencing and demonstrate its several advantages over short-read assemblies: (1) full-length transcript sequencing up to 6 kb, and possibly longer; (2) no assembly required; (3) no reference genome required. We believe high quality sets of transcripts identified by Iso-Seq method are critical for understanding transcriptome complexity and discovering novel biological insights.

Integrated OMICs analysis in IMG/M

Neha Varghese,[1] **I-Min Amy Chen,**[2] **Marcel Huntemann,**[1] Krishna Palaniappan,[2] Anna Ratner,[2] Manoj Pillay,[2] Ken Chu,[2] Victor Markowitz,[2] Natalia Ivanova,[1] Nikos Kyrpides,[1] and Amrita Pati* (apati@lbl.gov)[1]

[1]DOE Joint Genome Institute, Walnut Creek California; [2]Biological Data Management and Technology Center, CRD, Berkeley, California

The DOE Joint Genome Institute has grown from being one of the largest genomics facilities in the world to a science-driven generator of various types of functional genomics data. The IMG family of systems,

which functions as the most downstream component of the sequencing pipeline, integrates both JGI-generated and external datasets into a large-scale comparative genomics context. Newly integrated data-types include transcriptomics, metatranscriptomics, methylomics, and transposomics datasets. We describe new developments in IMG to integrate various OMICs datasets in the context of the comparative and statistical analysis tools in IMG. Analysis scenarios and visualization tools for each type of dataset are explored in this presentation.

Additionally, the IMG production pipeline has been enhanced to predict secondary metabolite biosynthetic clusters and natural products for both genomes and assembled metagenomes. IMG now serves as a repository for natural products and both experimentally validated and predicted biosynthetic clusters. Analysis scenarios for these clusters using the frameworks in IMG/M are presented in this work.

The evolution of microbial species: A view through the genomics lens

Neha J. Varghese* (njvarghese@lbl.gov)[1], Supratim Mukherjee[1], Natalia N. Ivanova[1], Konstantinos T. Konstantinidis[2], Kostas Mavrommatis[3], Nikos C. Kyrpides[1], and Amrita Pati[1].

[1]Microbial and Metagenome Superprogram, DOE Joint Genome Institute, Walnut Creek, California;[2]Department of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia; [3]Celgene Corp., San Francisco, California

For a long time prokaryotic species definition has been under debate and a constant source of turmoil in microbiology. This has recently prompted the ASM to call for a scalable and reproducible technique, which uses meaningful commonalities to cluster microorganisms into groups corresponding to prokaryotic species. Whole-genome Average Nucleotide Identity (gANI) was previously suggested as a measure of genetic distance that generally agrees with prokaryotic species assignments based on the accepted best practices (DNA-DNA hybridization and 16S rDNA similarity). In this work, we prove that gANI is indeed the meaningful commonality based on which microorganisms can be grouped into the aforementioned clusters. By analyzing 1.76 million pairs of genomes we find that identification of the closest relatives of an organism via gANI is precise, scalable, reproducible, and reflects the evolutionary dynamics of microbes. We model the previously unexplored statistical properties of gANI using 6,000 microbial genomes and apply species-specific gANI cutoffs to reveal anomalies in the current taxonomic species definitions for almost 50% of the species with multiple genome sequences. We also provide evidence of speciation events and genetic continuums in 17.8% of those species. We consider disagreements between gANI-based groupings and "named" species and demonstrate that the former have all the desired features to serve as the much-needed "natural groups" for moving forward with taxonomy. Further, the groupings identified have been presented in detail at http://ani.jgi-psf.org to facilitate comprehensive downstream analysis for researchers across different disciplines.

Increasing C6 cell wall sugar content by engineering the accumulation of a low recalcitrance polysaccharide in plants

Miguel E. Vega-Sánchez* (mevega-sanchez@lbl.gov),[1] Dominique Loqué,[1] Jeemeng Lao,[1] Thomas Herter,[1] Michela Catena,[1] Joshua L. Heazlewood,[1] Henrik V. Scheller,[1] and Pamela C. Ronald[2]

[1]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Walnut Creek, California; [2]University of California at Davis, Davis, California

The production of biofuels from lignocellulose remains a challenging process due primarily to the inherent resistance of plant cell walls to deconstruction (commonly referred to as recalcitrance). In addition, the presence of significant amounts of 5 carbon (C5) sugars in plant biomass that are not readily fermented by microorganisms is another obstacle to efficient conversion to fuels. For these reasons, reduced cell wall recalcitrance and increased C6 monosaccharide content are desirable traits in potential biofuel crops, as long as these biomass modifications do not significantly alter normal growth and development. Mixed-linkage glucan (MLG), a cell wall polysaccharide comprised of glucose monomers linked by both β-1,3 and β-1,4 bonds, is a good example of a low recalcitrance and C6-containing polymer. As opposed to cellulose, MLG is non-linear and cannot aggregate into microfibrils due to the presence of the β-1,3 bonds, which makes it more soluble in aqueous environments. We have previously shown that *Nicotiana benthamiana* plants transiently over-expressing the rice mixed-linkage glucan synthase CslF6 using a 35S promoter accumulate large amounts of MLG (over 10% of dry weight). However, previous data in barley and our own results in *N. benthamiana* have shown that constitutive production of MLG in plants severely compromises growth and development. In order to bypass these negative effects, we have successfully developed a strategy to engineer *Arabidopsis* plants to accumulate significant amounts of MLG in the cell wall without growth defects. The strategy involved the expression of the MLG synthase driven by a senescence promoter and resulted in an increase of over 40% in sugar release following enzymatic biomass saccharification. This approach offers a new engineering alternative to enhance cell wall properties of lignocellulosic biofuel crops.

The DOE Systems Biology Knowledgebase: Plant Science Domain

Doreen Ware* (ware@cshl.edu),[1,2] David Weston,[3] Jason Baumohl,[4] Aaron Best,[5] Jared Bischof,[6] Ben Bowen,[4] Tom Brettin,[6] Tom Brown,[6] Shane Canon,[4] Stephen Chan,[4] John-Marc Chandonia,[4] Dylan Chivian,[4] Ric Colasanti,[6] Neal Conrad,[6] Brian Davison,[4] Matt DeJongh,[5] Paramvir Dehal,[4] Narayan Desai,[6] Scott Devoid,[6] Terry Disz,[6] Meghan Drake, [3] Janaka Edirisinghe,[6] Gang Fang,[7] José Pedro Lopes Faria,[6] Mark Gerstein,[7] Elizabeth M. Glass,[6] Annette Greiner,[4] Dan Gunter,[4] James Gurtowski,[2] Nomi Harris,[4] Travis Harrison,[6] Fei He,[1] Matt Henderson,[4] Chris Henry,[6] Adina Howe,[6] Marcin Joachimiak,[4] Kevin Keegan,[6] Keith Keller,[4] Guruprasad Kora,[3] Sunita Kumari*,[2] Miriam Land,[3] Folker Meyer,[6] Steve Moulton,[3] Pavel Novichkov,[4] Taeyun Oh,[8] Gary Olsen,[9] Bob Olson,[5] Dan Olson,[5] Ross Overbeek,[5] Tobias Paczian,[5] Bruce Parrello,[5] Shiran Pasternak,[2] Sarah Poon,[4] Gavin Price,[4] Srivdya Ramakrishnan,[2] Priya Ranjan,[3] Bill Riehl,[4] Pamela Ronald,[8] Michael Schatz,[2] Lynn Schriml,[10] Sam Seaver,[6] Michael W. Sneddon,[4] Roman Sutormin,[4] Mustafa Syed,[3] James Thomason,[2] Nathan Tintle,[5] Will Trimble,[6] Daifeng Wang,[7] Andreas Wilke,[6] Fangfang Xia,[6] Shinjae Yoo,[1] Dantong Yu,[1] Robert Cottingham,[3] Sergei Maslov,[1] Rick Stevens,[6] and Adam P. Arkin[4]

[1]Brookhaven National Laboratory, Upton, New York; [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [4]Lawrence Berkeley National Laboratory, Berkeley, California; [5]Hope College, Holland, Michigan; [6]Argonne National Laboratory,

Argonne, Illinois; [7] Yale University, New Haven, Connecticut; [8]University of California at Davis, Davis, California; [9] University of Illinois at Champaign-Urbana, Champaign, Illinois; [10]University of Maryland, College Park, Maryland

KBase's primary scientific aim is to push multiple types of functional data towards increasingly specific models of metabolic and regulatory behavior of microbes, plants and their communities. The plant team is currently focused on reconstruction and modeling of genotype-to-phenotype relationships in plant species relevant to the DOE mission. Our workflows, which are accessible via the Narrative and command line interfaces, provide interactive, data-driven analysis and exploration across multiple experiments and diverse data types. KBase allows our users to process next generation sequencing data to identify novel genomic variation and to quantify genome-wide expression levels. Expression data can then be used to calculate co-expression networks and to identify and annotate densely interconnected functional modules within those networks. We also provide a comprehensive computational workflow allowing our users to carry out Genome-Wide Association (GWA) analysis to identify SNPs and candidate genes significantly correlated with important plant phenotypes. Genes identified by the GWA can be filtered and validated using a large collection of public and user-generated networks, expression data and metabolic models in KBase.

Comparative genomics of highly abundant *Bradyrhizobium spp.* isolated from bulk forest soils from across North America reveals extensive intra-genus diversity

Roland Wilhelm* (rwilhelm@mail.ubc.ca), David VanInsberghe, Erick Cardenas Poire, and William W. Mohn

Department of Microbiology and Immunology, Life Sciences Institute, University of British Columbia, Vancouver, Canada

The ecology of *Bradyrhizobium sp.,* a ubiquitous soil bacteria, ranges from endosymbiont to non-nodulating plant symbiont and to free-living species (incapable of nodulation), the latter two of which are poorly understood. Comprehensive profiling the microbial communities of bulk soil from softwood forests across North America, using 16S rRNA gene pyrotag and whole-extract metagenomics, demonstrate that a clade of closely related *Bradyrhizobium spp.* dominate the microbial community, constituting between 8-25% of all 16S reads and ~4% of total metagenomic reads (4-fold greater than the next most abundant taxa). Draft genomes have been assembled from 31 cultured strains in order to assess the metabolic capacity and possible ecological role of this group. Genome size and genetic diversity varied substantially between strains, with genome size spanning 4.0 – 11.4 Mb ($\bar{x}$ = 7.8 Mb), while only 1,711 out of 50,000 total COGs were part of the "core" set of genes found in all genomes, while 2,340 genes were common to >95% of genomes compared. Common traits within the family *Bradyrhizobiaceae*, such as indole acetic acid synthesis and salicylic acid degradation, were found in all genomes, but not canonical genes associated with symbiosis, like *nif* and *nod* genes. There is evidence that the core metabolic function of these organisms can be both chemolithoautrophy, producing energy via sulfite oxidation and fixing carbon via phosphoenol pyruvate carboxylase, or chemoorganoheterotrophy using low energy organics like carbon monoxide, formate, formaldehye, ethanol, and pyruvate. The decomposition of xylan, micro-crystalline cellulose and, lignin-alkali was assayed for a subset of isolates, yielding negative results for all but two strains capable of xylan decomposition. In addition to presenting a complete metabolic overview, we will examine strain-level abundance patterns across six forest sites in North America by mapping metagenomic data to draft

genomes. We will focus on the distribution of non-core genes across sites in an attempt to explain some of the extensive diversity observed between strain genomes. This research contributes valuable understanding of the diversity within a dominant taxonomic group in forest soils, as well as across most soil environments, providing a foundation for further investigations into the ecology of free-living *Bradyrhizobium spp.*

DOE JGI Emerging Technologies Opportunity Program (2013-2015) – Resources and protocols for the generation of high-quality genome assemblies

Rod A. Wing* (rwing@mail.arizona.edu)**,** Dave Kudrna, Yeisoo Yu, Jianwei Zhang, and Jose Luis Goicoechea

Arizona Genomics Institute, University of Arizona, Tucson, Arizona

The Arizona Genomics Institute (AGI) has played significant roles in numerous genome projects over the past 15 years, including rice and its 20 wild relatives, *maize*, *Brachypodium*, sugarcane, soybean and its wild relatives, tomato, Eucalyptus, Amaranthus, barley, cacao, cotton, *Amborella*, and *Eutrema salsugineum*. AGI's expertise is not limited to plants, and also includes model species like *Drosophila* (19 genomes), zebra finch, *Biomphalaria*, and nurse shark, as examples. The majority of AGI's roles in these genome projects have centered on the generation of high-quality physical maps and BAC libraries with the ultimate goal of producing the highest quality reference genome sequences possible. AGI's philosophy is that the first genome of every important species under investigation should be of the highest quality (i.e. reference sequence (RefSeq), or near-RefSeq quality) which significantly enhances/facilitates all subsequent downstream functional, population, and evolutionary analyses.

Our ETOP grant with JGI was funded in September 2014 to fulfill the following objectives:

A) Produce high-quality high-molecular weight genomic DNA and RNA from plants and other organisms under investigation at JGI and its CSP partners.

B) Development of robust protocols to generate large-insert Illumina mate-paired libraries (20, 40kb in size), and cost effective BAC-end sequencing.

C) Continue to generate large-insert high-quality BAC libraries for JGI flagship genomes and CSP partners.

D) Evaluate the cost effectiveness of the generation of sequence-based physical maps for JGI flagship genomes and CSP partners.

E) Evaluate the use of a new software program developed at AGI that can be used to edit next- gen genome assemblies with that aid of physical and genetic maps.

Progress in all of the areas will be discussed.

RNA-seq analysis the haptophyte *Emiliania huxleyi* reveals novel oleosin-like proteins related to production of neutral lipids for biofuels

Gordon Wolfe* (GWolfe2@csuchico.edu)[1] and Adarsh Jose[2]

[1]Department of Biological Sciences, California State University at Chico, Chico, California; [2]NSF Engineering Research Center for Biorenewable Chemicals (CBiRC), Biorenewables Research Laboratory Iowa State University, Ames, Iowa

Introduction: *Emiliania huxleyi* and some related haptophyte algae produce as neutral lipids a set of PolyUnsaturated Long-Chain (C_{37-39}) Alkenones, Alkenoates, and Alkenes (PULCA). These hydrocarbons are package into cytoplasmic lipid bodies [LBs, (Eltgroth et al 2005)], and have potential for biofuels (O'Neil et al 2012), but their biosynthesis and mobilization is still poorly understood.

Methods: We studied cellular transcriptomics of *E. huxleyi* CCMP 1516, previously sequenced by DOE-JGI (Read et al 2013), using RNA-seq. We designed a time-series experiment where lipid production was increased by a combination of phosphate limitation and bicarbonate boost, and lipid catabolism increased by placing cells in continuous darkness.

Results: We detected 30,399 proteins as transcripts, 78% of which showed significant ($p<0.05$) change for at least one time point. We used the Short Time-series Expression Miner [STEM; (Ernst & Bar-Joseph 2006)] to cluster genes with >2-fold change in expression by expression pattern vs. time. In the two most highly down-expressed clusters, we observed a set of proteins that, although unique to *Emiliania*, somewhat resemble oleosins, <25 kDa proteins that coat LBs in plant seeds and chlorophytes, and regulate LB trafficking and mobilization. The *E. huxleyi* genome appears to contain a large set of these 'oleosin-like' proteins, most near 12 kDa, but a few with insertions or deletions. While a previous LB proteomics screen found other LB structural proteins such as PAP-fibrillin homologs, the oleosin-like proteins were missed because of their low mass. The *E. huxleyi* proteins are also structurally distinct: they are much less hydrophobic, lack the proline knot motif, and their expression goes down, not up, during LB accumulation. The hypothesis that these proteins may keep LBs from coalescing is supported by observations that LBs grow larger when cells become phosphate limited, when these show reduced transcription.

Conclusion: This study reveals how RNA-seq approach can detect novel proteins with a strong transcriptional response that correlates with neutral lipid production in this globally-important alga.

References Cited:

Eltgroth ML, Watwood RL, Wolfe GV. 2005. Production and cellular localization of long-chain neutral lipids in the haptophyte algae *Isochrysis galbana* and *Emiliania huxleyi*. J. Phycology 41: 1000-09

Ernst J, Bar-Joseph Z. 2006. STEM: a tool for the analysis of short time series gene expression data. *BMC* Bioinformatics 7: 191

O'Neil GW, Carmichael CA, Goepfert TJ, Fulton JM, Knothe G, et al. 2012. Beyond fatty acid methyl esters: expanding the renewable carbon profile with alkenones from *Isochrysis* sp. Energy & Fuels 26: 2434-41

Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, et al. 2013. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature 499: 209-13

Does Neurospora know what it's eating?: Characterizing transcriptional response of *N. crassa* on different nutrient sources

Vincent Wu* (vwu104@berkeley.edu),[1] David Kowbel,[1] Anna lipzen,[2] Vasanth Singan,[2] Igor Gregoriev,[1,2] and N. Louise Glass[1]

[1]Department of Plant and Microbial Biology University of California at Berkeley, Berkeley California; [2]DOE Joint Genome Institute, Walnut Creek, California

Conversion of lignocellulosic plant biomass to biofuels holds great potential for alleviating our reliance on fossil fuels. A major goal in this area of research is to engineer strains of fungi that can reliably secrete effective hydrolytic enzymes for deconstruction of plant biomass. To better engineer fungi, an understanding of how filamentous fungi respond to and degrade this compositionally diverse material is required. Using Neurospora crassa as a model, we have begun to characterize the transcriptional regulation and function of genes involved in carbon metabolism. First we have looked at transcriptional change through time as Neurospora digests and consumes plant material. We have also collected and sequenced mRNA transcripts from N. crassa grown on a wide variety of specific plant cell wall components from mono and disaccharides to complex polysaccharide components, such as cellulose, hemicellulose and pectin. By comparative analyses of plant biomass deconstruction pathways, we hope to define novel hypotheses regarding function of unannotated genes, potential regulators of transcription of different subsets enzymes and signaling molecules that induce these regulatory circuits. Initial results show clear differences in transcriptional response between different carbon sources, as well as plant biomass through time. These data also point to specific mono and dissacharides that are able to induce transcription of cellulolytic genes.

MaxBin: Accurate, automatic, and fast binning software for genome reconstruction from metagenomic datasets using an expectation-maximization algorithm

Yu-Wei Wu* (ywwei@lbl.gov),[1,2] Yung-Hsu Tang,[1,3] Susannah G. Tringe,[4,5] Blake A. Simmons,[1,4] and Steven W. Singer[1,5]

[1]Joint BioEnergy Institute, Emeryville, California; [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California; [3]City College of San Francisco, San Francisco, California; [4]Biological and Materials Science Center, Sandia National Laboratories, Livermore, California; [5]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

Genomic reconstructions of microbial populations obtained by metagenomic sequencing allow access to uncultivated microbial populations that may have important roles in natural and engineered ecosystems. Understanding the roles of these uncultivated populations has broad application in ecology, evolution, biotechnology and medicine. Rapid and accurate binning of assembled metagenomic sequences is a required step in reconstructing the genomes and understanding microbial function. We have developed a binning algorithm, MaxBin, which automates the binning of metagenomic sequences using an expectation-maximization algorithm. Binning of simulated metagenomic datasets demonstrated that MaxBin had high levels of accuracy in binning microbial genomes. MaxBin was used to reconstruct genomes from metagenomic data obtained through the Human Microbiome Project, which demonstrated its ability to reconstruct genomes from complex metagenomic datasets with shallow sequencing coverage. Application of MaxBin to metagenomes obtained from microbial consortia adapted to grow on cellulose allowed the genomic reconstruction of new uncultivated cellulolytic

bacterial populations, including an abundant myxobacterial population distantly related to Sorangium cellulosum that possessed a much smaller genome (5MB versus 13-14MB) but has a more extensive set of genes for biomass deconstruction. For the cellulolytic consortia, the MaxBin results were compared to binning using Emergent Self-Organizing Maps (ESOMs) and differential coverage binning, demonstrating that it performed comparably to these methods but had distinct advantages in automation, resolution of related genomes and sensitivity.

Synthesis of a eukaryotic chromosome

Zhuwei Xu* (Zhuwei.Xu@nyumc.org)[1,3], Sam Deutsch[2], and Jef Boeke[3]

[1]Johns Hopkins School of Medicine, Baltimore, Maryland; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]New York University Langone Medical Center, New York, New York

The Sc2.0 project aims to synthesize the genome of *S.cerevisiae* from DNA oligos, and my work is to build the largest synthetic yeast chromosome, *synIV* (~1.45 Mb in Sc2.0). The synthetic strain features: I) wild-type comparable fitness; II) a stable genome with destabilizing elements deleted, including transposons and tRNAs; and III) genetic flexibility and features that allow novel genetic approaches for future research. The new genetic features include: I) an inducing evolution system, SCRaMbLE (Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution); II) all UAG codons recoded to UAA stop codons; and III) removal of repetitive and destabilizing repeats and tRNA genes. The SCRaMbLE system is a novel genetic tool that shuffles genes, and induces genome-wide gene duplications, deletions, inversions and translocations. The spared UAG codon allows *in vivo* incorporations of the designed amino acid into protein. The synthetic yeast also has a neochromosome that facilitates the integration of novel pathways in yeast.

The synthetic yeast genome has a hierarchical structure to guide the construct process from chemically synthesized DNA oligos. The hierarchy is: chromosome-"megachunk"-"chunk"-"minichunk"-"building block"-DNA oligo. The *synIV* left arm chunks are being synthesized at JGI using the Community Synthesis Program. The right arm chunks are being assembled at NYU. The megachunk (~30 to 60 kb) is the integration unit. The megachunks contain autotrophic markers and replace the wild-type chromosome to construct synthetic strains. We utilize an *in vivo* assembly-based approach to obtain the chunks to minimize risk of DNA mutations. A fitness assay is performed for the synthetic strains competed against the wild type strain to verify that intermediate strains show high fitness. Novel genetic tools like SCRaMbLE will be applied to *synIV* in future studies.

Integrative biology of a fungus: Using PacBio® SMRT® sequencing to interrogate the genome, epigenome, and transcriptome of *Neurospora crassa*

Jane Yeadon[1], **Kristi Kim**[2], Elizabeth Tseng[2], Susana Wang[2], Joan Wilson[2], David Catchside[1], Jane Landolin* (jlandolin@pacificbiosciences.com)[2]

[1]Flinders University, Adelaide, Australia; [2]Pacific Biosciences, Menlo Park, California

PacBio SMRT® Sequencing has the unique ability to directly detect base modifications in addition to the nucleotide a sequence of DNA. Because eukaryotes use base modifications to regulate gene expression, the absence or presence of epigenetic events relative to the location of genes is critical to elucidate the function of the modification. Therefore an integrated approach that combines multiple omic-scale

assays is necessary to study complex organisms. Here, we present an integrated analysis of three sequencing experiments: 1) DNA sequencing, 2) base-modification detection, and 3) Iso-SeqTM analysis, in Neurospora crassa, a filamentous fungus that has been used to make many landmark discoveries in biochemistry and genetics. We show that de novo assembly of a new strain yields complete assemblies of entire chromosomes, and additionally contains entire centromeric sequences. Base-modification analyses reveal candidate sites of increased interpulse duration (IPD) ratio, that may signify regions of 5mC, 5hmC, or 6mA base modifications. Iso-Seq method provides full-length transcript evidence for comprehensive gene annotation, as well as context to the base-modifications in the newly assembled genome. Projects that integrate multiple genome-wide assays could become common practice for identifying genomic elements and understanding their function in new strains and organisms.

Targeted mutagenesis of CENH3 in switchgrass and *Brachypodium* using TALEN

Sangwoong Yoon* (tellite7@gmail.com),[1] Christina Craig,[2] Eduardo Blumwald,[1] and Christian Tobias[2]

[1]Department of Plant Sciences, University of California, Davis, Davis, California; [2]Genomics and Gene Discovery Unit, USDA-ARS, Albany, California

Biofuels are one form of sustainable energy which can partially substitute for petroleum based transportation fuels and can be produced by breakdown and fermentation of plant cell wall carbohydrates. Switchgrass (*Panicum virgatum* L.) is one important bioenergy crop thanks to its high yield, adaptability and perennial nature. Moreover, there is huge potential for improvement of this energy crop through breeding approaches. Because it is difficult to fix beneficial alleles within outcrossing switchgrass populations, we have initiated an approach to enable doubled haploid production through a process referred to as centromere-mediated genome elimination. For this purpose we are generating targeted mutations in CENH3 using TALEN. Six TALEN pairs were designed to knock out endogenous swichgrass genes PviCENH3-1 and PviCENH3-2. Moreover, another 6 pairs of TALEN were constructed to knock out the CENH3 and COMT genes of the model grass *Brachypodium distachyon*. Each left and right TALEN pair was inserted in the pZHY013 vector which contains a T2A translational skipping sequence and FokI nuclease. The whole expression cassettes were then cloned into the plant expression vector pSY32. These transformation vectors will enable us to assess rates of DNA-cleavage and non-homologous end joining in transformed plant callous via direct sequencing or PCR approaches. The goal is to produce stably inherited cenH3 mutant alleles in whole plants.

The functional potential of *Arabidopsis thalina* root endophytic communities

Scott Yourstone* (scott.yourstone81@gmail.com)**,[1]** Hunter Cameron,[1] Derek Lundberg,[1] Sur Herrera Paredes,[1] Scott Clingenpeel,[2] Susannah Tringe,[2] Corbin Jones,[1] and Jeff Dangl[1,3]

[1]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, North Carolina; [2]DOE Joint Genome Institute, Walnut Creek, California; [3]Jeffrey Dangl is an HHMI-GBMF Plant Science Investigator

Bacterial communities living inside of plant roots (endophytes) are of particular agronomic interest because of their potential to provide beneficial phenotypes such as disease resistance, increased

growth, or resistance to abiotic stress. A primary interest is the functional potential of endophytic communities as revealed by whole metagenome sequencing. To circumvent the challenge of contamination by host plant DNA, we are combining a variety of approaches. These include 1) whole genome sequencing of cultured clonal bacterial isolates via HiSeq, MiSeq and Pacific Biosciences RS sequencing platforms, 2) cell sorting and single-cell sequencing of uncultured cells, 3) *En masse* sequencing of all culturable material via metagenomic DNA made from "plate scrapes", and 4) metagenomic DNA made from pools of bacteria separated from host material via cell sorting. Here we describe these sequencing efforts along with the pan- and core-genome characteristics of several taxonomic groups of isolates and the prevalence of plasmid DNA.

Identification of genomic elements required for uranium resistance by *Caulobacter crescentus*

Mimi C. Yung* (yung6@llnl.gov),[1] Dan M. Park,[1] Cindi A. Hoover,[2] Matthew J. Blow,[2] and Yongqin Jiao[1]

[1]Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California; [2]DOE Joint Genome Institute, Walnut Creek, California;

(Work at LLNL conducted under Contract DE-AC52-07NA27344. LLNL-ABS-650802.)

Uranium (U), particularly in its water soluble form U(VI), has been shown to be a significant environmental toxin and a major contaminant at DOE legacy sites. One potential method to remediate U contamination is to use microbes that are highly tolerant of U and are able to decrease its mobility in the environment. *Caulobacter crescentus* is an aerobic, aquatic bacterium that is highly tolerant of U and has the potential to be used for U bioremediation. In order to gain an understanding of how *C. crescentus* resist U, we employed a transposon mutagenesis screening approach (Tn-seq) to identify essential genomic elements that are required for U resistance. In our method, highly saturated transposon (Tn) mutagenesis was first performed to generate a library of 10^6 mutants. The library was then grown on solid agar plates containing U, Cd, or no metal control. Mutants surviving each exposure were subsequently sequenced via high-throughput Illumina sequencing to identify Tn insertion sites. Genes that accumulated fewer Tn insertions under U stress compared to the Cd or no stress controls were identified as genes specific for U tolerance. Using this method, we identified 18 genes potentially involved in U tolerance, which were subsequently tested through mutational analysis. Twelve of the 18 mutants were confirmed to have a growth defect under U stress compared to the wild-type strain. Several genes involved in cell-cycle progression were identified, such as SpoT, ClpA and FilM, suggesting a close relationship between cell-cycle progression and U toxicity. Additionally, TolC-like transporters $RsaF_a$ and $RsaF_b$, previously identified to be involved in S-layer protein transport, as well as stress factors CztR and CztA were found to help confer resistance to U. Together our results demonstrate a complex cellular response to U by *C. crescentus* and provide important insight into the various resistance mechanisms employed for survival.

Transcriptomic analysis of four switchgrass genotypes toward revealing the molecular basis of cell wall recalcitrance

Chengcheng Zhang* (chengcheng2012@ou.edu),[1] Presenjit Saha,[1] Desalegn Serba,[2] Ajay Jha,[1] Kangmei Zhao,[1] Malay Saha,[2] and **Laura E. Bartley**[1]

[1]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma; [2]Forage Improvement Division, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma

Switchgrass (*Panicum virgatum*) is a promising feedstock for lignocellulosic biofuel production due to its stress tolerance and high yield. However, as with most species being considered for terrestrial biomass production, cell wall recalcitrance hinders biochemical conversion of plant biomass to biofuels. Improving understanding of the molecular mechanisms leading to cell wall recalcitrance will facilitate genetic engineering of switchgrass to produce more usable biomass. In this study, we compared the transcriptomes of four full-sibling switchgrass genotypes with differing digestibility (two low digestibility (LD); two high digestibility (HD)) to identify differentially expressed genes involved in cell wall synthesis and regulation. We used RNA Seq to estimate gene expression in four different samples from the elongation 4 (E4) developmental stage, including mature (stem hard, SH) and less mature halves (stem soft, SS) of the second internode, the leaf (L), and the whole tiller (W), generating ~80 million 90 base pair, paired-end reads for each sample. For our initial analysis with CLC Genomics Workbench, we uniquely mapped ~84% of reads to the newest switchgrass genome reference (v.1.1) from the Joint Genome Institute. Due to potentially high heterozygosity between the reference genome and the genotypes analyzed, we allowed up to 30% mismatches and about 5% of reads were unmapped. The replicates show correlation coefficients >0.65 (0.87 ± 0.10), consistent with reasonably high-quality data. Using the R package, DESeq, we found that 571 genes are differentially expressed

between the high digestibility group (HD) and the low digestibility group (LD) in the four sample types (mean normalized counts between SH, SS, L and W samples for each genotype pair, at FDR<0.01). Searching by BLAST with 111 putative phenylpropanoid pathway genes from the swithgrass unique transcript database in the newest genome, we found 236 lignin biosynthesis candidate loci. We found that 34 of 236 putative lignin biosynthesis genes showed significant differential expression between the HD and LD groups across the sample types, 14 of which are more abundant in the recalcitrant group (LD). In leaves, MYB46 and MYB83, two mid-level transcription factors regulating secondary cell wall synthesis, are expressed more highly in LD than HD. We identified organ and development-specific gene expression changes in the dataset. Compared with leaves, 2615 genes are differentially expressed in stems (SH and SS combined) in every genotype. A number of grass-diverged BAHD CoA acyltransferase genes that have been implicated grass cell wall cross-linking, and many putative secondary cell wall regulators are more highly expressed in stems than leaves. Comparing SH and SS from each genotype, we found only 6 genes differentially expressed in every genotype. Combined with cell wall composition and microscopic tissue structure analyses, we expect that these data will contribute to deciphering the molecular mechanism of cell wall recalcitrance in switchgrass toward altering gene expression to produce easily digestible lignocellulosic biomass.

Rapid microbial feedbacks mediates vulnerability of permafrost carbon to climate warming

Jizhong Zhou* (jzhou@ou.edu),[1] Kai Xue,[1] Mengting Yuan,[1] Liyou Wu,[1] Zhili He,[1] Jim Cole,[2] James M. Tiedje,[2] Rosvel Bracho-Garrillo,[3] Edward A.G. Schuur,[3] Chengwei Luo,[4] Konstantinos Konstantinidis,[4] and Yiqi Luo[1]

[1]Institute for Environmental Genomics and Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma; [2]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan; [3]Department of Biology, University of Florida, Gainesville, Florida; [4]Center for Bioinformatics and Computational Genomics and School of Biology, Georgia Institute of Technology, Atlanta, Georgia

Microbial decomposition of permafrost carbon (C) is one of the most likely potential positive feedbacks from terrestrial ecosystems to atmosphere in a warmer world, and thus understanding microbial mechanisms controlling permafrost C decomposition is critical to project feedbacks to future global climate warming. However, microbial responses (speed, direction and magnitude) in permafrost to climate warming remain elusive. Here, using integrated metagenomic technologies (amplicon sequencing, shotgun sequencing, GeoChip), we demonstrated the rapid feedbacks of microbial communities to climate warming in a tundra warming experiment in Alaska. Consistent with the changes in soil temperature, moisture and ecosystem respiration (Reco), microbial functional community structure mainly measured by GeoChip was shifted dramatically just after one and half year of warming, indicating rapid response and high sensitivity of tundra ecosystem to climate warming. Also, warming stimulated not only functional genes involved in aerobic respiration for degrading both labile and recalcitrant C, as supported by 19.1% higher Reco and 56.3% higher decomposition rate on standard substrate under warming than control, but also genes for anaerobic decomposition via denitrification, sulfate reduction and methanogenesis. Nitrous oxide and methane emissions from anaerobic processes are likely to further amplify positive C feedbacks to warming. In addition, warming greatly enhanced nutrient cycling processes such as nitrogen (N) mineralization, N fixation and phosphorus utilization, which promotes increased plant growth and contribute to the 35.9% increase in gross primary productivity (GPP). As a result of simultaneously increased Reco and GPP, net ecosystem exchange did not change by warming. Together, our results demonstrate the vulnerability of permafrost C to climate warming and the importance of microbial feedbacks in mediating such vulnerability.

Attendees

Martin Ackermann
ETH Zurich
martin.ackermann@env.ethz.ch

Catherine Adam
JGI
cadam@lbl.gov

Musahid Ahmed
LBL
mahmed@lbl.gov

Timothy Alba
UNLV
albat@unlv.nevada.edu

Ed Allen
JGI
eaallen@lbl.gov

Dilara Ally
Bayer Crop Sciences - Biologics
dilara.ally@bayer.com

John Anderson
UC Berkeley
rfrawley@berkeley.edu

Todd (Robert) Anderson
U.S. Department of Energy
todd.anderson@science.doe.gov

Bill Andreopoulos
DOE JGI
wandreopoulos@lbl.gov

Massie Ballon
DOE Joint Genome Institute
mlballon@lbl.gov

Anton Bankevich
St.Petersburg Academic University
anton.bankevich@gmail.com

Richard Baran
LBNL
RBaran@lbl.gov

José María Barrasa
CIB
josem.barrasa@uah.es

Kerrie Barry
JGI
kwbarry@lbl.gov

Michae l Barton
Joint Genome Institute
mbarton@lbl.gov

Gopala Battu
University Of Illinois
gbattu@illinois.edu

Diane Bauer
JGI
dmbauer@lblb.gov

Chris Beecroft
JGI
cjbeecroft@lbl.gov

Matthew Bendall
JGI
mlbendall@lbl.gov

Cole Benjamin
JGI
bjcole@lbl.gov

David Benton
Great Lakes Bioenergy Research Center, UW Madison
dbenton@glbrc.wisc.edu

J Philipp Benz
Energy Biosciences Institute; UC Berkeley
ph.benz@berkeley.edu

Mary Berbee
University of British Columbia
mary.berbee@gmail.com

Randy Berka
Novozymes, Inc.
ramb@novozymes.com

David Berry
University of Vienna
berry@microbial-ecology.net

Craig Betts
Clontech Laboratories
craig_betts@clontech.com

Surojit Biswas
University of North Carolina at Chapel Hill
sbiswas@live.unc.edu

Matthew Blow
JGI
mjblow@lbl.gov

Burt Bluhm
University of Arkansas
bbluhm@uark.edu

Alexander Boyd
JGI
aeboyd@lbl.gov

Sara Branco
UC Berkeley
sara.mayer.branco@gmail.com

William Brazelton
University of Utah
william.brazelton@utah.edu

Thomas Brettin
Argonne National Laboratory
brettin@anl.gov

Eoin Brodie
Lawrence Berkeley National Lab
elbrodie@lbl.gov

Natasha Brown
DOE Joint Genome Institute
nzvenigorodsky@lbl.gov

Anthony Bryan
ORNL
bryanac@ornl.gov

Jim Burrows
Pacific Biosciences
jburrows@pacificbiosciences.com

David Burstein
UC Berkeley
davidbur@berkeley.edu

Joshua Burton
University of Washington, Genome Sciences
jnburton@uw.edu

Posy Busby
University of Washington
busby@uw.edu

Kathryn Bushley
University of Minnesota
kbushley@umn.edu

Kristin Butcher
Joint Genome Institute
kdbutcher@lbl.gov

Lucélia Cabral
University of Campinas - UNICAMP
luc.g.cabral@gmail.com

Julia Calderone
UC Santa Cruz Science Communication
juliacalderone@gmail.com

Saori Campen
Pacific Northwest National Lab/Joint BioEnergy Institute
sacampen@lbl.gov

Shane Canon
LBNL - NERSC
scanon@lbl.gov

Erick Cardenas poire
University of British Columbia
carden24@mail.ubc.ca

Joe Carlson
DOE JGI
JWCarlson@lbl.gov

Akiko Carver
JGI
acarver@lbl.gov

Linda Castle
self
lacastle_2000@yahoo.com

Dee Catino
JGI
gdcatino@lbl.gov

David Cavalier
Great Lakes Bioenergy Center
cavalie8@msu.edu

Shannon Ceballos
UC Davis
sjceballos@ucdavis.edu

Javier A. Ceja-Navarro
Lawrence Berkeley National Laboratory
jcnavarro@lbl.gov

Patrick Chain
LANL
pchain@lanl.gov

Patricia Chan
Maverix Biomics
pchan@maverixbio.com

Naveen Chandramohan
JGI
nchandramohan@lbl.gov

Jay Chen
Oak Ridge National Laboratory
chenj@ornl.gov

Cindy Chen
USDA-ARS
cindy.chen@ars.usda.gov

Cindy Chen
JGI
cindychen@lbl.gov

Jan-Fang Cheng
Joint Genome Institute
jfcheng@lbl.gov

HAOYU CHENG
JGI
haoyucheng@lbl.gov

Mawsheng Chern
University of California
mschern@ucdavis.edu

Hari Chhetri
West Virginia University
harichhetri@gmail.com

Jason Chin
Pacific Biosciences
jchin@pacificbiosciences.com

Jennifer Chiniquy
DOE Joint Genome Institute
JLChiniquy@lbl.gov

Dylan Chivian
Lawrence Berkeley National Lab
DCChivian@lbl.gov

Mansi Chovatia
JGI
mrchovatia@lbl.gov

Julianna Chow
JGI
jchow@lbl.gov

Victor Chubukov
JBEI
vachubukov@lbl.gov

Doina Ciobanu
JGI
dgciobanu@lbl.gov

Scott Clingenpeel
DOE Joint Genome Institute
srclingenpeel@lbl.gov

Alicia Clum
JGI
aclum@lbl.gov

Robert Cameron
Coates Scrips institution of Oceanography, UCSD
rccoates@ucsd.edu

Devin Coleman-Derr
JGI
dacoleman-derr@lbl.gov

Joshua Coomey
University of Massachusetts Amherst
jcoomey@cns.umass.edu

Alex Copeland
JGI
accopeland@lbl.gov

Robert Cottingham
Oak Ridge National Laboratory
cottinghamrw@ornl.gov

John Crandon
University of Arkansas
jccrando@uark.edu

Byron Crump
Oregon State University
bcrump@coas.oregonstate.edu

Chris Daum
DOE Joint Genome Institute
cgdaum@lbl.gov

Matt Davenport
UC Santa Cruz
mr.matt.davenport@gmail.com

Anne Dekas
LLNL
adekas@gmail.com

Ratnakar Deole
Northeastern State University
deole@nsuok.edu

Shweta Deshpande
JGI
SDeshpande@gmail.com

Sam Deutsch
JGI
sdeutsch@lbl.gov

Adam Deutschbauer
Lawrence Berkeley National Lab AM
Deutschbauer@lbl.gov

Anthony Devine
FMC Corporation
Anthony.Devine@fmc.com

Jeremy Dodsworth
University of Nevada, Las Vegas
jeremy.dodsworth@unlv.edu

Mitchel Doktycz
Oak Ridge National Laboratory
doktyczmj@ornl.gov

Marie Donnelly
EBI UC Berkeley
mkdonnelly@berkeley.edu

Meghan Drake
Oak Ridge National Laboratory
drakemm@ornl.gov

Daniel Drell
DOE OBER
daniel.drell@science.doe.gov

Matan Drory Retwitzer
Ben-Gurion University
matandro@bgu.ac.il

Inna Dubchak
JGI
ildubchak@lbl.gov

Nicole Dubilier
Max Planck Institute for Marine Microbiology
ndubilie@mpi-bremen.de

Kecia Duffy
JGI
kmduffy@lbl.gov

Sebastien
Duplessis INRA
duplessi@nancy.inra.fr

Katrina Edwards
University of Southern California
kje@usc.edu

Rob Egan
JGI
rsegan@lbl.gov

Emiley Eloe-Fadrosh
Gordon and Betty Moore Foundation
emiley.eloe-fadrosh@moore.org

Shahram Emami
Carnegie Institution for Science, Department of Plant Biology
shahrame@stanford.edu

Cassandra Ettinger
UC Davis
clettinger@ucdavis.edu

Kjiersten Fagnan
JGI/NERSC
kmfagnan@lbl.gov

Stephen Fairclough
JGI
srfairclough@lbl.gov

Luisa Falcon
UNAM, Institute of Ecology
LFalcon@lbl.gov

Yigang Fang
Lawrence Berkeley National Laboratory
yigangfang@lbl.gov

Cat Ferguson
Freelance
carolinetaylorferguson@gmail.com

Elizabeth Fergusson
Bigelow Laboratory for Ocean Sciences
edmitrieff@bigelow.org

Jason Fiedler
ARS/USDA
jason.fiedler@ars.usda.gov

MichaelFischbach
UCSF
MFischbach@lbl.gov

Bryce Foster
JGI
brycefoster@lbl.gov

Rashmi Fotedar
Ministry of Environment
rfotedar@hotmail.com

Chris Francis
Stanford University
caf@stanford.edu

Levi Gadye
Berkeley Science Review (UC Berkeley)
lgadye@gmail.com

Mohamed Ghazy
The American University in Cairo
mghazy@aucegypt.edu

Amit Ghosh
Joint BioEnergy Institute
itsamit79@gmail.com

Donald Gibson
UC Davis
dongibson@ucdavis.edu

David Gilbert
DOE Joint Genome Institute
degilbert@lbl.gov

Lori Giver
Calysta Energy
lgiver@calystabio.com

John Gladden
JBEI
jmgladden@lbl.gov

Elizabeth Glass
Argonne National Laboratory
drstrawberry@aol.com

Tijana Glavina del Rio
Joint Genome Institute
tglavina@lbl.gov

Eugene Goltsman
Joint Genome Institute
egoltsman@lbl.gov

Tania Gonzalez
University of California, Berkeley
tlgonzalez@berkeley.edu

David Goodstein
JGI
dmgoodstein@lbl.gov

Sean Gordon
USDA
seangordon07@gmail.com

Danielle Goudeau
JGI
dgoudeau@lbl.gov

Dave Greenshields
Novozymes BioAg
dvgs@novozymes.com

Igor Grigoriev
JGI
ivgrigoriev@lbl.gov

Jenny Gu
Pacific Biosciences
jgu@pacificbiosciences.com

Adam Guss
Oak Ridge National Laboratory
gussam@ornl.gov

Alba Gutierrez
Joint Genomic Institute
AAGutierrez@lbl.gov

Michalis Hadjithomas
JGI
mhadjithomas@lbl.gov

Charles Halfmann
South Dakota State University
cthalfmann@jacks.sdstate.edu

Cyd Hamilton
ORISE/ORNL
CEHDOEwork@gmail.com

Matthew Hamilton
JGI
mghamilton@lbl.gov

Nancy Hammon
JGI
nmhammon@lbl.gov

James Han
JGI
jkhan@lbl.gov

Sajeet Haridas
JGI
sharidas@lbl.gov

Miranda Harmon-Smith
JGI
mlharmon-smith@lbl.gov

Wyatt Hartman
DOE Joint Genome Institute
whhartman@lbl.gov

Erik Hawley
WSU Tri-Cities
ehawley.amge@gmail.com

Richard Hayes
DOE Joint Genome Institute
rdhayes@lbl.gov

David Hays
JGI
dehays@lbl.gov

Brian Hedlund
Univ. of NV, Las Vegas, School of Life Sciences
Brian.Hedlund@unlv.edu

Monica Heger
GenomeWeb
mheger@genomeweb.com

Uffe Hellsten
DOE Joint Genome Institute
uhellsten@lbl.gov

Matt Henderson
LBNL/KBase
mhenderson@lbl.gov

John Henske
University of California - Santa Barbara
john.henske@gmail.com

Sur Herrera Paredes
The University of North Carolina
sur00mx@gmail.com

Matthias Hess
WSU Tri-Cities/PNNL
matthias.hess@tricity.wsu.edu

Luke Hickey
Pacific Biosciences
lhickey@pacificbiosciences.com

Scott Hickey
UC Berkeley
sfhickey1@gmail.com

Nathan Hillson
JGI
njhillson@lbl.gov

Aditya Hindupur
NASA Ames Research Center
aditya.hindupur@nasa.gov

Jennifer Hiras
Joint BioEnergy Institute/LBNL
jhiras@lbl.gov

Cindi Hoover
DOE Joint Genome Institute
cahoover@lbl.gov

Chiaki Hori
RIKEN
chori718@jcom.home.ne.jp

Leila Hornick
JGI
lahornick@lbl.gov

Susan Hua
Joint Genome Institute
shua@lbl.gov

Corey Hudson
Sandia National Labs
cmhudso@sandia.gov

Laura Hug
UC Berkeley
laura.hug@berkeley.edu

Marcel Huntemann
JGI
mhuntemann@lbl.gov

Miriam Hutchinson
University of New Mexico
yesterdaymail@gmail.com

Patrik Inderbitzin
UC Davis Department of Plant Pathology
prin@ucdavis.edu

Bill Inskeep
Montana State University
binskeep@montana.edu

Michelle Isbandi
JGI
misbandi@lbl.gov

Merja Itävaara
VTT
merja.itavaara@vtt.fi

Natalia Ivanova
Joint Genome Institute
nnivanova@lbl.gov

Niranjani Iyer
Monsanto
nnjamb@monsanto.com

Douglas Jacobsen
JGI/NERSC
dmjacobsen@lbl.gov

Rashmi Jain
UC-Davis
rsjain@ucdavis.edu

Jessica Jarett
Joint Genome Institute
jkjarett@lbl.gov

Erin Jarvis
UC Berkeley
erinjarvis@berkeley.edu

ByoungSeung Jeon
Hanyang University
jbstrust@gmail.com

Liangrong Jiang
Plant Pathology Department, UC Davis
lrjiang108@xmu.edu.cn

Marcin Joachimiak
LBNL
MJoachimiak@lbl.gov

Nicole Johnson
JGI
NicoleVJohnson@lbl.gov

Lorrie Johnson
DOE/OSTI
JohnsonL@osti.gov

Richard Jorgensen
Carnegie Institution for Science
rajorgensen@mac.com

Chijioke Joshua
Joint BioEnergy Institute
cjjoshua@lbl.gov

Nicholas Justice
LBNL
nbjustice@lbl.gov

Don Kang
JGI
ddkang@lbl.gov

Jenifer Kaplan
JGI
jkaplan@lbl.gov

Ulas Karaoz
LBNL
ukaraoz@lbl.gov

Anne-Kristin Kaster
Stanford University
annekristinkaster@gmail.com

Lisa Kegg
JGI
lrkegg@lbl.gov

William Kelly
AgResearch
bill.kelly@agresearch.co.nz

Megan Kennedy
JGI
mckennedy@lbl.gov

Shintaro Kikuchi
Muroran Institute of Technology
shintaro@mmm.muroran-it.ac.jp

Jeff Kimbrel
Joint BioEnergy Institute/LBNL
jakimbrel@lbl.gov

GM King
LSU
gkingme@gmail.com

Manuel Kleiner
Max Planck Institute for Marine Microbiology
mkleiner@mpi-bremen.de

Sunita Kumari
CSHL
kumari@cshl.edu

Govindarajan Kunde Ramamoorthy
Joint Genome Institute
gkramamoorthy@lbl.gov

Victor Kunin
Taxon Biosciences
kunivi@gmail.com

Rita Kuo
JGI
rckuo@lbl.gov

Alan Kuo
JGI
akuo@lbl.gov

Cheryl Kuske
LANL
kuske@lanl.gov

Nikos Kyrpides
DOE Joint Genome Institute
nckyrpides@lbl.gov

Kurt LaButti
DOE JGI
klabutti@lbl.gov

David Lai
JGI
dlai@lbl.gov

Kathleen Lail
JGI
klail@lbl.gov

Jane Landolin
PacBio
jlandolin@pacificbiosciences.com

Christina Lanzatella
USDA-ARS
christina.craig@ars.usda.gov

Alla Lapidus
JGI
alapidus@lbl.gov

Debbie Laudencia
USDA-ARS
debbie.laudencia@ars.usda.gov

Jackson Lee
NASA Ames Research Center
jackson.z.lee@nasa.gov

Janey Lee
DOE JGI
jlee2@lbl.gov

Kang Soo Lee
MIT
ks_lee@mit.edu

Guotian Li
University of California,Davis
gtli@ucdavis.edu

Xiang Li
Novozymes BioAg
xl@novozymes.com

Shoudan Liang
JGI
sliang@lbl.gov

Marc Libault
University of Oklahoma
libaultm@ou.edu

Erik Lilleskov
US Forest Service R&D, NRS
elilleskov@fs.fed.us

Steve Lindemann
Pacific Northwest National Laboratory
stephen.lindemann@pnnl.gov

Erika Lindquist
JGI
ealindquist@lbl.gov

Gregory Linshiz
JGI
glinshiz@lbl.gov

Anna Lipzen
Joint Genome Institute
alipzen@lbl.gov

Derek Lundberg
University of North Carolina at Chapel Hill
derek.lundberg@gmail.com

Jed Lynn
JBEI
jlynn@lbl.gov

Sally Madden
JGI
smadden@lbl.gov

Jon Magnuson
Pacific Northwest National Laboratory
jon.magnuson@pnnl.gov

Jyothi Mallajosyula
JGI
jmallajosyula@lbl.gov

Joel Martin
JGI
j_martin@lbl.gov

Angel T Martinez
CIB, CSIC
atmartinez@cib.csic.es

Claudia Martinez-Schwarz
JGI
cmartinez-schwarz@lbl.gov

Rebekah Mathews
DOE Joint Genome Institute
mathews14@llnl.gov

Peter Matlock
PY Matlock Consulting
peter@biomatlock.com

Kelly Matzen
Oxitec Ltd
kelly.matzen@oxitec.com

Mary Maxon
LBNL
memaxon@lbl.gov

Phillip McClean
North Dakota State University
phillip.mcclean@ndsu.edu

Jack McFarland
US Geological Survey
jmcfarland@usgs.gov

June Medford
Colorado State University
june.medford@colostate.edu

Xiandong Meng
JGI
Xiandongmeng@lbl.gov

Maria Mercedes-Roca
Zamorano University Honduras
mmroca@zamorano.edu

Folker Meyer
Argonne National Laboratory
folker@anl.gov

Sirma Mihaltcheva
JGI
smihaltcheva@lbl.gov

Bonnie Millenbaugh
JGI/LLNL
bamillenbaugh@lbl.gov

William Mohn
University of British Columbia
wmohn@mail.ubc.ca

Scott Monsma
Lucigen Corp.
smonsma@lucigen.com

William Moore
JBEI/UC Berkeley
wmoore@lbl.gov

Emmanuelle Morin
INRA Nancy-Lorraine
emmanuelle.morin@nancy.inra.fr

Annika Mosier UC Berkeley
annika.mosier@gmail.com

Wellington Muchero
Oak Ridge Nattional Laboratory
mucherow@ornl.gov

Supratim Mukherjee
LBNL/JGI
supratimmukherjee@lbl.gov

Monica Munoz-Torres
JGI
mcmunozt@lbl.gov

Senthil Murugapiran
University of Nevada, Las Vegas
senthil.murugapiran@unlv.edu

Vivek Mutalik
Berkeley National Lab
vkmutalik@lbl.gov

Christine Naca
JGI
clnaca@lbl.gov

Kishore Nannapaneni
Bayer CropScience
kishore.nannapaneni@bayer.com

Nandita Nath
JGI
nnath@lbl.gov

Sangeeta Nath
JGI
snath@lbl.gov

Don Natvig
1952
dnatvig@gmail.com

Ali Navid
Lawrence Livermore National Laboratory
navid1@llnl.gov

Stephen Nayfach
UCSF
snayfach@gmail.com

Simona Necula
JGI
sfnecula@lbl.gov

William Nelson
Pacific Northwest National Laboratory
william.nelson@pnnl.gov

Annalee Newitz
io9.com
annalee@gmail.com

Vivian Ng
JGI
vng@lbl.gov

Chew Yee Ngan
Joint Genome Institute
cyngan@lbl.gov

Hai Nguyen
University of Ottawa
hai.nguyen.1984@gmail.com

Nancy Nichols
USDA Natl. Ctr. for Agricultural Utilization Res.
nancy.nichols@ars.usda.gov

Masaru Nobu
Univ. of Illinois, Urbana-Champaign
nobu2@illinois.edu

Matt Nolan
JGI
mpnolan@lbl.gov

Angela Norbeck
EMSL at Pacific Northwest National Lab
angela.norbeck@pnnl.gov

Alex Nord
JGI/LBNL
asnord@lbl.gov

Henrik Nordberg
Joint Genome Institute
hnordberg@lbl.gov

Trent Northen
LBNL/JGI
trnorthen@lbl.gov

Jeanette Norton
Utah State University
jeanette.norton@usu.edu

Francisco Noya
Instituto "Clemente Estable"
fnoya@geocom.com.uy

Erin Nuccio
Lawrence Livermore National Lab
nuccio1@llnl.gov

Steve Oh
Clontech Laboratories, Inc
Steve_Oh@Clontech.com

Seungdae Oh
Univ. of Illinois at Urbana-Champaign
seungdae@illinois.edu

Robin Ohm
DOE Joint Genome Institute
raohm@lbl.gov

Ai Oikawa
JBEI
AOikawa@lbl.gov

Christian Olsen
Biomatters
christian@biomatters.com

Brian Olson
Lawrence Berkeley National Labratory
bsolson@lbl.gov

Jeff Osborne
Manchester University
jposborne@manchester.edu

Anne Osbourn
John Innes Centre
anne.osbourn@jic.ac.uk

Robert Otillar
JGI
RPOtillar@lbl.gov

Antonio David Paez Espino
JGI
adpaezespino@lbl.gov

Tatiana Paley
JGI
tlpaley@lbl.gov

Chongle Pan
Oak Ridge National Laboratory
panc@ornl.gov

Jasmyn Pangilinan
JGI
jlpangilinan@lbl.gov

Dan Park
Lawrence Livermore National Lab
park36@llnl.gov

Hyojung Park
Hanyang University
suya8259@naver.com

Charlotte Park
CSU Chico
cpark8@mail.csuchico.edu

Amrita Pati
DOE JGI
apati@lbl.gov

Dale Pelletier
Oak Ridge National Laboratory
pelletierda@ornl.gov

Paul Peluso
Pacific Biosciences, Inc.
ppeluso@pacificbiosciences.com

Ze Peng
DOE Joint Genome Institute
zpeng@lbl.gov

Yi Peng
JGI
ypeng@lbl.gov

Len Pennacchio
JGI
lapennacchio@lbl.gov

Christa Pennacchio
JGI
cppennacchio@lbl.gov

Rene Perrier
DOE Joint Genome Institute
raperrier@lbl.gov

JenniferPett-Ridge
Lawrence Livermore National Lab
pettridge2@llnl.gov

Pallavi Phatale
Joint Bio Energy Institute
pphatale@lbl.gov

Jeremy Phillips
JGI
jlphillips@lbl.gov

Hailan Piao
WSU Tri-Cities
hailan.piao@tricity.wsu.edu

Dominic Pinel
University of California, Berkeley
dpinel@berkeley.edu

Antonio Pisabarro
Public University of Navarre
gpisabarro@unavarra.es

Kunal Poorey
Sandia National Labs
kpoorey@sandia.gov

James Preston
University of Florida
jpreston@ufl.edu

Stephen Quake
Stanford University
quake@stanford.edu

Pablo Rabinowicz
U.S. Department of Energy
pablo.rabinowicz@science.doe.gov

Preethi Ramaiya
Novozymes
pira@novozymes.com

Gerardo Ramirez
JGI
gjramirez@lbl.gov

David Rank
Pacific Biosciences
drank@pacificbiosciences.com

Nicole Rapicavoli
Pacific Biosciences
nrapicavoli@pacificbiosciences.com

Igor Ratnere
DOE Joint Genome Institute
iratnere@lbl.gov

Tatiparthi Reddy
JGI
tbreddy@lbl.gov

Rebecca Reiss
New Mexico Tech
reiss@nmt.edu

Rebecca Reiss
JGI
rreiss@lbl.gov

Sarah Richardson
JGI
SMRichardson@lbl.gov

John Ridenour
University of Arkansas
jridenou@uark.edu

Frank Roberto
Newmont Mining Corporation
frank.roberto@newmont.com

Aaron Robinson
Department of Biology, University of New Mexico
arobin01@unm.edu

Dmitry Rodionov
Sanford-Burnham Medical Research Institute
rodionov@sanfordburnham.org

Pamela Ronald
UC Davis
pcronald@ucdavis.edu

Marisa Rosa
UC Berkeley/PGEC
massrosa@berkeley.edu

Steve Rounsley
Dow Agrosciences
steve.rounsley@gmail.com

Francisco Javier Ruiz-Dueñas
CIB, CSIC
fjruiz@cib.csic.es

Douglas Rusch
IN Univ. Center for Genomics and Bioinformatics
drusch@cgb.indiana.edu

Elizabeth Ryan
JGI
eryan@lbl.gov

Yana Safonova
Saint-Petersburg Academic University
safonova.yana@gmail.com

Asaf Salamov
DOE JGI
aasalamov@lbl.gov

Navaneetha Santhanam
Self
teenavan@gmail.com

Hila Sberro-Livnat
Weizmann Institute of Science
hserro@gmail.com

Daniel Schachtman
University of Nebraska
daniel.schachtman@unl.edu

Wendy Schackwitz
JGI
wsschackwitz@lbl.gov

Christopher Schadt
Oak Ridge National Laboratory
schadtcw@ornl.gov

Andrew Schaumberg
JGI
aschaumberg@lbl.gov

Alicia Scheffer
Agilent Technology
alicia_scheffer@agilent.com

Henrik Scheller
LBNL
hscheller@lbl.gov

Robert Schmitz
University of Georgia
schmitz@uga.edu

Jeremy Schmutz
JGI-HA
jschmutz@hudsonalpha.org

Patrick Schwientek
Bayer CropScience
patrick.schwientek@bayer.com

Nicole Scott
University of Chicago
nicolescott1@gmail.com

Armand Seguin
Natural Resources Canada
armand.seguin@nrcan.gc.ca

Rekha Seshadri
JGI
rseshadri@lbl.gov

Igor Shabalov
JGI
ishabalov@lbl.gov

Atif Shahab
JGI
ashahab@lbl.gov

Nicole Shapiro
JGI
nrshapiro@lbl.gov

Aditi Sharma
JGI/ LBNL
aditisharma@lbl.gov

Itai Sharon
UC Berkeley
itaish@berkeley.edu

Niranjan Shekar
Spiral Genetics
niranjan@spiralgenetics.com

Louis Sherman
Purdue University
lsherman@purdue.edu

Sisarie Sherry
CSU Chico
ssherry1@mail.csuchico.edu

Shengqiang Shu
JGI
sqshu@lbl.gov

Bryan Siepert
JGI
bsiepert@lbl.gov

Blake Simmons
JBEI
basimmons@lbl.gov

Vasanth Singan
JGI
VRSingan@lbl.gov

Esther Singer
Joint Genome Institute
esinger@lbl.gov

Kanwar Singh
JGI
ksingh@lbl.gov

Beate Slaby
University of Wuerzburg, Dept. of Botany II
beate.slaby@uni-wuerzburg.de

Steven Slater
Great Lakes Bioenergy Research Center
scslater@glbrc.wisc.edu

Tatyana Smirnova
DOE Joint Genome Institute
tsmirnova@lbl.gov

Jacquelyn Smith
Berkeley National Lab
jesmith@lbl.gov

Jason Smith
MBARI
jsmith@mbari.org

Jonathon Smith
University of Arkansas
jes002@uark.edu

Jason Smith
Pacific Biosciences
jrsmith@pacificbiosciences.com

Kevin Solomon
UCSB
ksolomon@engineering.ucsb.edu

Kristi Spittle Kim
Pacific Biosciences
kspittle@pacificbiosciences.com

Alexander Spunde
JGI
ayspunde@lbl.gov

Marikunte Yanjarappa Sreenivasa
University of Arkansas
smarikun@uark.edu

Brian St. Aubin
UC Berkeley
staubinb@berkeley.edu

Shelagh Steckler
Novozymes BioAg
sghs@novozymes.com

Sarah Stevens
Univ. of Wisconsin-Madison/McMahon Lab
sstevens2@wisc.edu

Rhona Stuart
Lawrence Livermore Natl Lab
stuart25@llnl.gov

Seung-Jin Sul
JGI
ssul@lbl.gov

Kankshita Swaminathan
University of Illinois
kank@illinois.edu

Yi Tang
University of California Los Angeles
yitang@ucla.edu

Tootie Tatum
JGi
oltatum@lbl.gov

John Taylor
University of California
jtaylor@berkeley.edu

Kristen Taylor
JGI
kmtaylor@lbl.gov

Susanna Theroux
JGI
stheroux@lbl.gov

Alex Thomas
DOE Joint Genome Institute
alexanderthomas@lbl.gov

Hope Tice
JGI
HNTice@lbl.gov

Don Tjandra
JGI
datjandra@lbl.gov

Christian Tobias
USDA-ARS
christian.tobias@ars.usda.gov

Kathleen Treseder
University of California Irvine
treseder@uci.edu

Susannah Tringe
DOE Joint Genome Institute
sgtringe@lbl.gov

Jim Tripp
JGI
hjtripp@lbl.gov

Andrew Tritt
Joint Genome Institute
ajtritt@lbl.gov

Stephan Trong
JGI
strong@lbl.gov

Elizabeth Tseng
Pacific Biosciences
etseng@pacificbiosciences.com

Ray Turner
JGI
crturner@lbl.gov

Gerald Tuskan
Oak Ridge National Laboratory
gtk@ornl.gov

FNU Vaishnavi Nagesh
JGI
fvaishnavinagesh@lbl.gov

Daniel Van der Lelie
FMC
daniel.vanderlelie@fmc.com

Jean VanderGheynst
UC Davis
jsvander@ucdavis.edu

Neha Varghese
DOE Joint Genomic Institute
njvarghese@lbl.gov

Miguel Vega-Sanchez
LBNL
mevega-sanchez@lbl.gov

Ophelia Venturelli
LBL
osventurelli@lbl.gov

Rytas Vilgalys
Duke University
fungi@duke.edu

Axel Visel
JGI / LBNL
avisel@lbl.gov

John Vogel
USDA-ARS Western Regional Research Center
john.vogel@ars.usda.gov

Dan Voytas
University of Minnesota
voytas@umn.edu

Momo Vuyisich
Los Alamos National Lab
vuyisich@lanl.gov

Carey Waage
JGI
cvwaage@lbl.gov

Soichi Wakatsuki
SLAC
Soichi.wakatsuki@stanford.edu

Mark Waldrop
USGS
mwaldrop@usgs.gov

Zhong Wang
JGI
zhongwang@lbl.gov

I-Ching Wang
NuGEN
icwang@nugen.com

Junming Wang
Univ. of Illinois at Urbana-Champaign
wagjim@illinois.edu

Mei Wang
Joint Genome Institute
mwang@lbl.gov

Jordan Waters
JGI
jwaters@lbl.gov

Peter Weber
LLNL
weber21@llnl.gov

Kelly Wetmore
LBNL
kmwetmore@lbl.gov

Nick Wigginton
Science / AAAS
nwiggint@aaas.org

Roland Wilhelm
University of British Columbia
rwilhelm@mail.ubc.ca

Andreas Wilke
ANL
wilke@mcs.anl.gov

Timothy Williams
JGI/LBL
tcwilliams@lbl.gov

Rod Wing
University of Arizona
rwing@mail.arizona.edu

Gordon Wolfe
California State Univ. Chico
gwolfe2@csuchico.edu

Chee-Hong Wong
DOE Joint Genome Institute
chwong@lbl.gov

Kelsey Wood
UC Davis Genome Center
klsywd@ucdavis.edu

Tanja Woyke
DOE Joint Genome Institute
twoyke@lbl.gov

Yu-Wei Wu
LBNL
ywwei@lbl.gov

Gang Wu
Algenol Biofuels
gang.wu@algenol.com

Vincent Wu
UC Berkeley PMB
vwu104@berkeley.edu

G. Albert Wu
DOE JGI
gwu@lbl.gov

Dongying Wu
University of California Davis
dygwu@ucdavis.edu

Zhuwei XU
NYULMC
wszwei@gmail.com

Denise Yadon
JGI
dyadon@lbl.gov

Juying Yan
JGI
juyingyan@lbl.gov

Shinjae Yoo
BNL
sjyoo@bnl.gov

Sangwoong Yoon
USDA ARS WRRC
tellite7@gmail.com

Yuko Yoshinaga
JGI
yyoshinaga@lbl.gov

Scott Yourstone
UNC-Chapel Hill
scott.yourstone81@gmail.com

Mimi Yung
LLNL
yung6@llnl.gov

Matthew Zane
JGI
mczane@lbl.gov

Jinxiang Zhang
Joint Bioenergy Institute
jinxiangzhang@lbl.gov

Chengcheng Zhang
University of Oklahoma
chengcheng2012@ou.edu

Jean Zhao
JGI
zyzhao@lbl.gov

Jizhong Zhou
University of Oklahoma
jzhou@ou.edu

Naupaka Zimmerman
University of Arizona
naupaka@gmail.com

Gerald Zon
TriLink BioTechnologies
gzon@trilinkbiotech.com

Author Index

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

U.S. DEPARTMENT OF ENERGY
Office of Science

CSO 0245