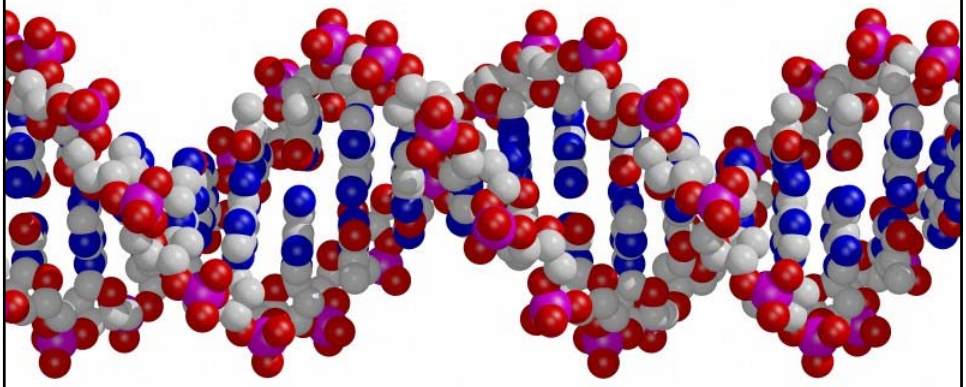
Evolutionary Genomics



Antonis Rokas
Department of Biological Sciences
Vanderbilt University

http://as.vanderbilt.edu/rokaslab

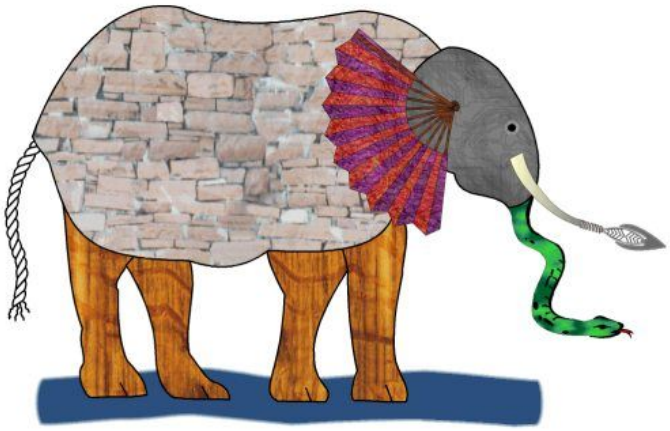
Lecture Outline

- Introduction to evolutionary genomics
- Benchmarking *de novo* transcriptome sequencing for functional and evolutionary genomics
- Phylogenomics

----------Coffee Break----------

- Comparative, Population & Functional Genomics

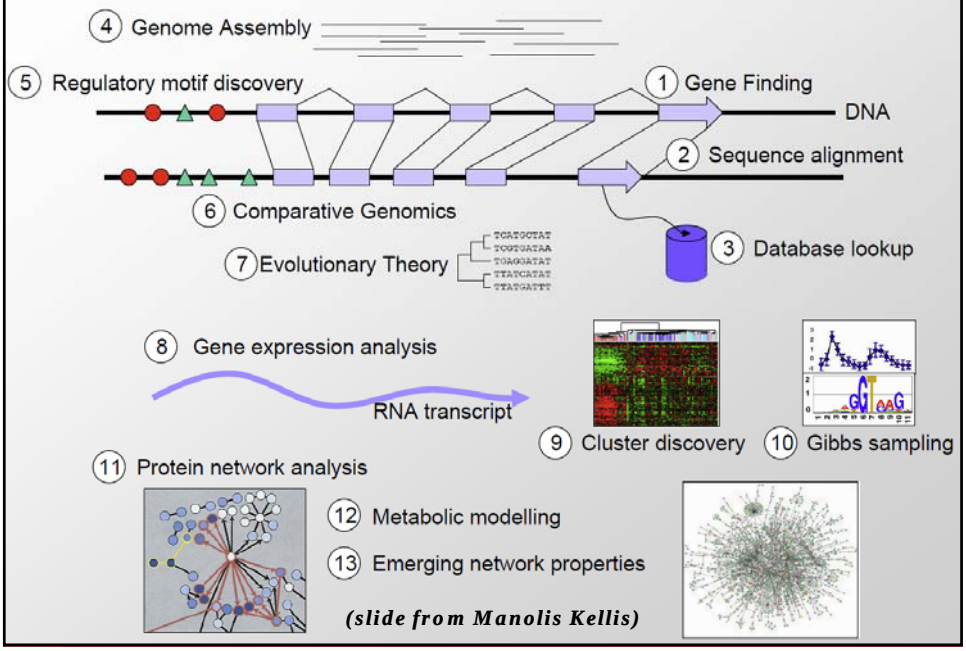
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

Understanding Sequences Requires Tools and Evolution

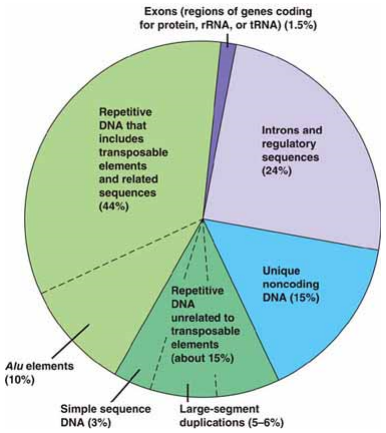


What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

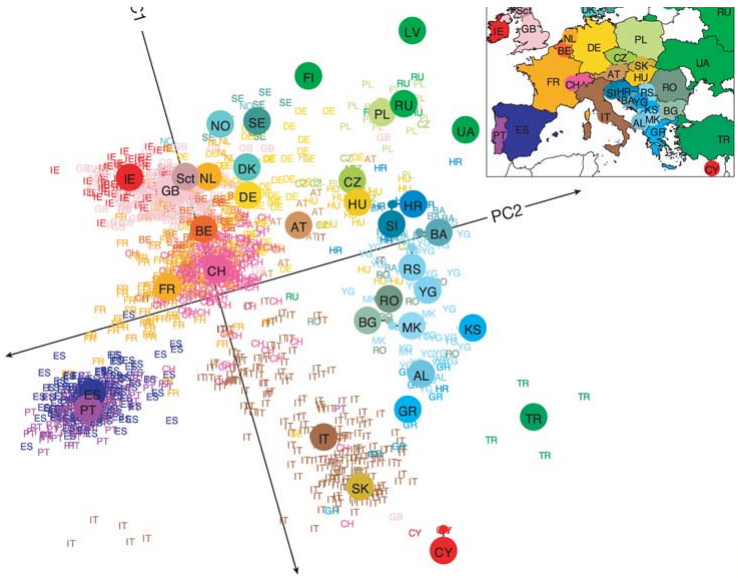
Transposon Protein Binding Site Exon Intron

Organization of the Human Genome



http://www.bio.miami.edu/~cmallery/150/gene/c7.19.14.human.genome.jpg

Human Genes Mirror Geography



Novembre et al. (2008) Nature

Recent Positive Selection in Human Populations

in the Asian Population, involved in hair follicle development

The twenty-two strongest candidates for natural selection

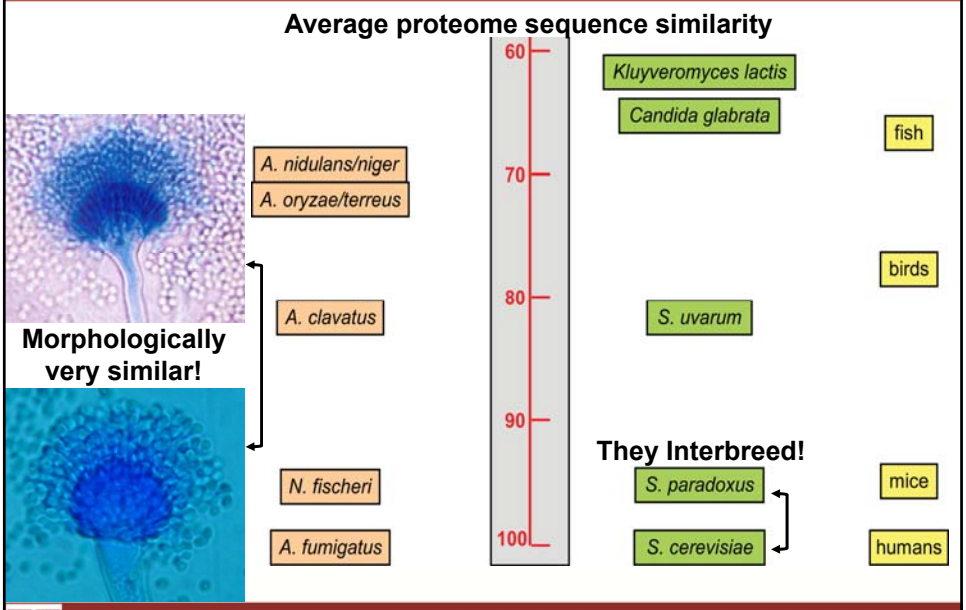
Chr:position (MB, HG17)	Selected population	Long Haplotype Test	Size (Mb)	Total SNPs with Long Haplotype Signal	Subset of SNPs that fulfil criteria 1	Subset of SNPs that fulfil criteria 1 and 2	Subset of SNPs that fulfil criteria 1, 2 and 3	Genes at or near SNPs that fulfil all three criteria
chr1:166	CHB + JPT	LRH, iHS	0.4	92	39	30	2	*BLZF1, SLC19A2*
chr2:72.6	CHB + JPT	XP-EHH	0.8	732	250	0	0	
chr2:108.7	CHB + JPT	LRH, iHS, XP-EHH	1.0	972	265	7	1	*EDAR*
chr2:136.1	CEU	LRH, iHS, XP-EHH	2.4	1,213	282	24	3	*RAB3GAP1, R3HDM1, LCT*
chr2:177.9	CEU,CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	*PDE11A*
chr4:33.9	CEU,YRI, CHB + JPT	LRH, iHS	1.7	413	161	33	0	
chr4:42	CHB + JPT	LRH, iHS, XP-EHH	0.3	249	94	65	6	*SLC30A9*
chr4:159	CHB + JPT	LRH, iHS, XP-EHH	0.3	233	67	34	1	
chr10:3	CEU	LRH, iHS, XP-EHH	0.3	179	63	16	1	
chr10:22.7	CEU, CHB + JPT	XP-EHH	0.3	254	93	0	0	
chr10:55.7	CHB+JPT	LRH, iHS, XP-EHH	0.4	735	221	5	2	*PCDH15*
chr12:78.3	YRI	LRH, iHS	0.8	151	91	25	0	
chr15:46.4	CEU	XP-EHH	0.6	867	233	5	1	*SLC24A5*
chr15:61.8	CHB + JPT	XP-EHH	0.2	252	73	40	6	*HERC1*
chr16:64.3	CHB + JPT	XP-EHH	0.4	484	137	2	0	
chr16:74.3	CHB + JPT, YRI	LRH, iHS	0.6	55	35	28	3	*CHSTS, ADAT1, KARS*
chr17:53.3	CHB + JPT	XP-EHH	0.2	143	41	0	0	
chr17:56.4	CEU	XP-EHH	0.4	290	98	26	3	*BCAS3*
chr19:43.5	YRI	LRH, iHS, XP-EHH	0.3	83	30	0	0	
chr22:32.5	YRI	LRH	0.4	318	188	35	3	*LARGE*
chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	
	Total SNPs		16.74	9,166	2,898	480	41	

In the European population, involved in skin pigmentation

In the West African population, related to Lassa virus infection

Sabeti et al. (2007) Nature

Genomes Provide a Common Yardstick for Comparison



Fedorova et al. (2008) PLoS Genet.

What Makes Us Sick Is the Stuff of Life

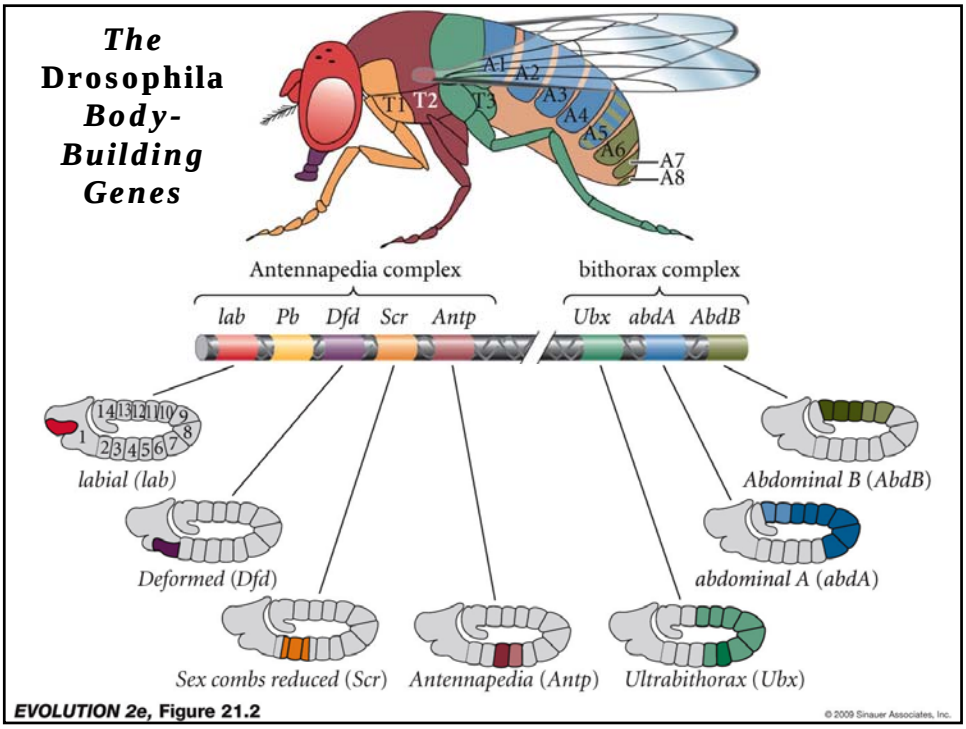
F	W	Y	Cancer
+			ABL1
+			Acute Myeloid Leukemia-DEK
+			Adenomat. Polyposis Coli-APC
+			AKT2
+			Ataxia Telangiectasia-ATM
-			BRCA1
-			BRCA2
+			Basal Cell Nevus-PTC
+			B-Cell Lymphoma 2-BCL2
-			B-Cell Lymphoma 3-BCL3
+			Bloom-BLM
+			Burkitt's Lymphoma-MYC
-			CDKN2C
-			CSF1R/C-Fms
+			Chk2 Protein Kinase
-			PDGFB
+			CML-BCR
+			Cyclin D1-CCND1
+			Cyclin Dep. Kinase 4-CDK4
+			EGFR
+			ERBB2
-			ETS
+			E-Cadherin-CDH1
+			Ewing Sarcoma-FLI-1
-			FGF3
-			Fanconi's Anemia A-FANCA
-			Fanconi's Anemia C-FANCC
-			Fanconi's Anemia G-FANCG
+			HNPCC*-MSH2
+			HNPCC*-MSH3
+			HNPCC*-MSH6
+			HNPCC*-MLH1
+			HNPCC*-PMS2
-			KIT

F	W	Y	Neurological
+			Adrenoleukodystrophy-ABCD1
+			Alzheimer-PS1
+			Alzheimer-APP
+			Amyotrophic Lat. Sclero.-SOD1
+			Angelman-UBE3A
+			Aniridia-PAX6
+			Best Macular Dystrophy-VMD2
+			Ceroid-Lipofuscinosis-PPT
+			Ceroid-Lipofuscinosis-CLN3
-			Ceroid-Lipofuscinosis-CLN2
-			Charcot-Marie-Tooth 1A-PMP22
-			Charcot-Marie-Tooth 1B-MPZ
+			Choroideremia-CHM
-			Creutzfeldt-Jakob-PRNP
+			Deafness, Hereditary-MYO15
+			Deafness, X-Linked-TIMM8A
+			Diaphanous 1-DIAPH1
+			Dementia, Multi-Infarct-NOTCH3
+			Duchenne MD*-DMD
-			Emery-Dreifuss MD*-EMD
+			Emery-Dreifuss MD*-LMNA
+			Familial Encephalopathy-PI12
+			Fragile-X -FRAXA
+			Friedreich Ataxia-FRDA
+			Frontotemporal Dement.-TAU
-			Fukuyama MD*-FCMD
+			Huntington-HD
+			Limb Girdle MD* 2A-CAPN3
+			Limb Girdle MD* 2B-YSF
-			Limb Girdle MD* 2E-BSG
+			Lissencephaly, X-Linked-DCX
+			Lowe Oculocerebroren.-OCRL
-			Machado-Joseph-MJD1
+			Miller-Dieker Lissen.-PAF

F	W	Y	Malformation Syndromes
-			Aarskog-Scott-FGD1
+			Achondroplasia-FGFR3
+			Alagille-JAG1
+			Barth-TAZ
-			Beckwith-Wiedemann-CDKN1C
-			Cerebral Cavern. Malf.-CCM1
+			Chondrodyspl. Punct. 1-ARSE
+			Cleidocranial Dysplasia-OFC1
-			Cockayne I-CKN1
+			Coffin-Lowry-RPS6KA3
+			Diastrophic Dyspl.-SLC26A2
+			EEC 3-Ket. P63
+			Greig Cephalopolysynd.-GLI3
-			Hand-Foot-Genital-HOXA13
+			Holoprosencephaly 3-SHH
+			Holoprosencephaly-SIX3
+			Holt-Oram-TBX5
-			ICF-DNMT3B
+			Kallman-KAL1
-			Laterality, X-Linked-ZIC3
+			Melnick-Fraser-EYA1
+			Nail Patella-LMX1B
-			Opitz-MID1
+			Renal Coloboma-PAX2
+			Rieger, Type 1-PITX2
-			Rubinstein-Taybi-CREBBP
+			Saethre-Chotzen-TWIST
-			Septooptic Dysplasia-HESX1
+			Simpson-Golabi-Behmel-GPC3
+			Townes-Brockes-SALL1
-			Treacher-Collins-TCOF1
-			VMCM-TEK
+			Wardenburg-PAX3
+			Zellweger-PEX1

Human disease-associated genes shared with flies (F), worms (W), and Yeast (Y); from Rubin et al. (2000) Science

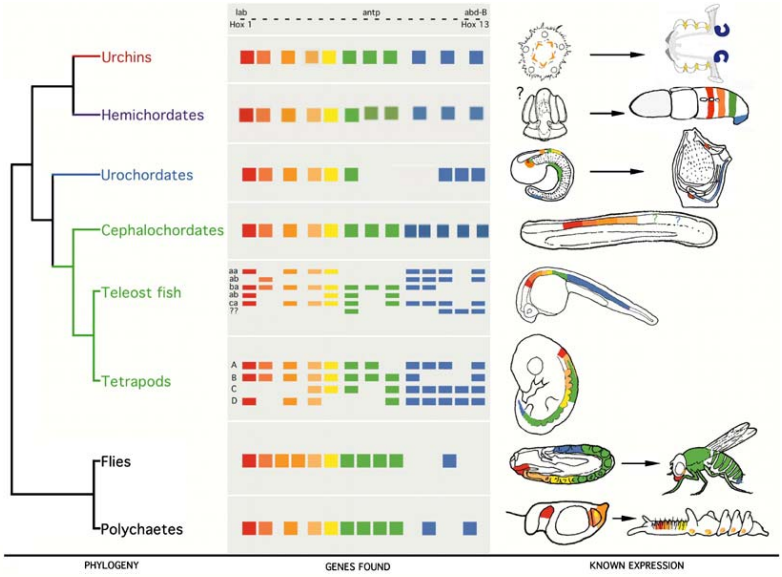
The Drosophila *Body-Building Genes*



EVOLUTION 2e, Figure 21.2

© 2009 Sinauer Associates, Inc.

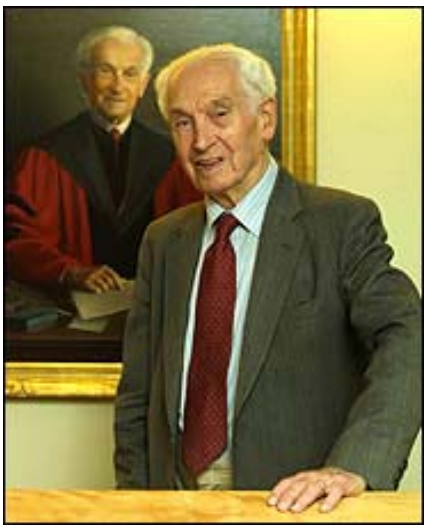
Animal Bodies are Built from the Same Genetic Toolkit



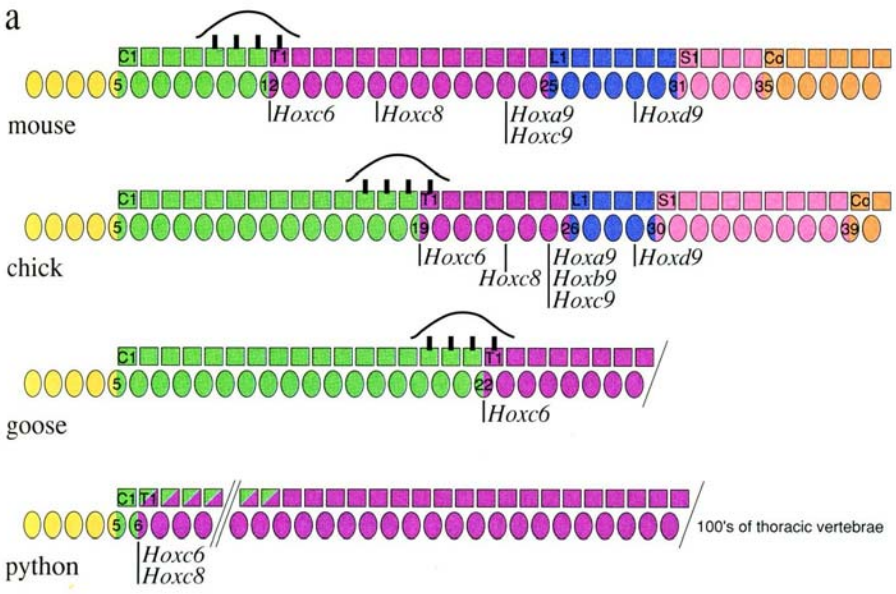
Swalla (2006) Heredity

"...the search for homologous genes is quite futile except in very close relatives. If there is only one efficient solution for a certain functional demand, very different gene complexes will come up with the same solution, no matter how different the pathway by which it is achieved. The saying 'Many roads lead to Rome' is as true in evolution as in daily affairs"

Ernst Mayr, 1963

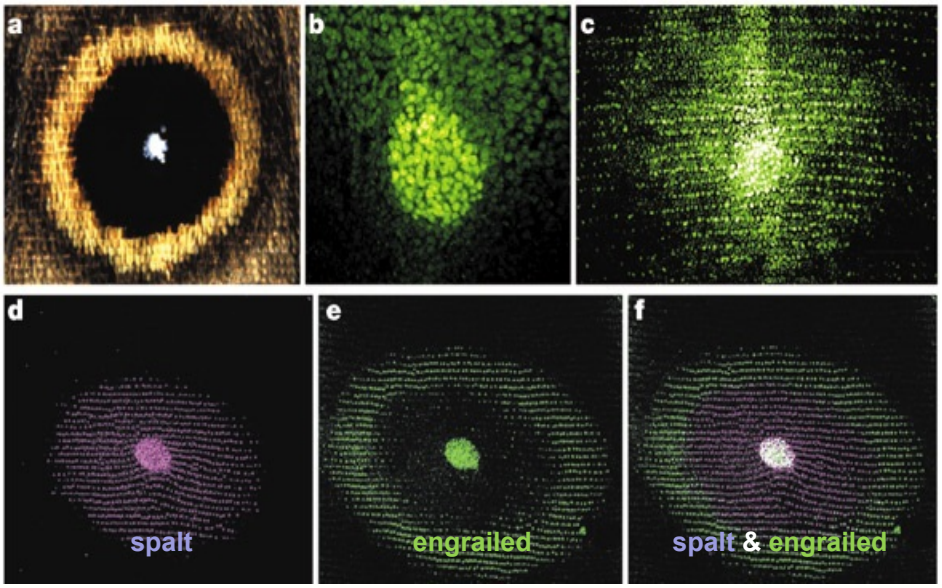


Altering the Expression of Old Genes



Carroll et al. (2005) From DNA to Diversity, 2nd ed., Fig. 5.6

Teaching Old Genes New Tricks (Gene Co-Option)



Beldade & Brakefield (2002) Nature Rev. Genet.

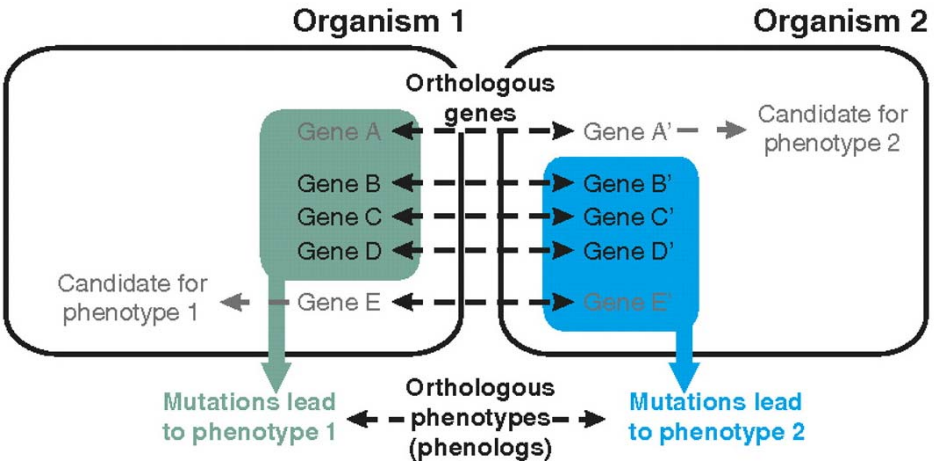
"Our method suggests a yeast model for angiogenesis defects, a worm model for breast cancer, mouse models of autism, and a plant model for the neural crest defects […], among others"

McGary et al. (2010) PNAS

"Modelling neurodegeneration in *Saccharomyces cerevisiae*"

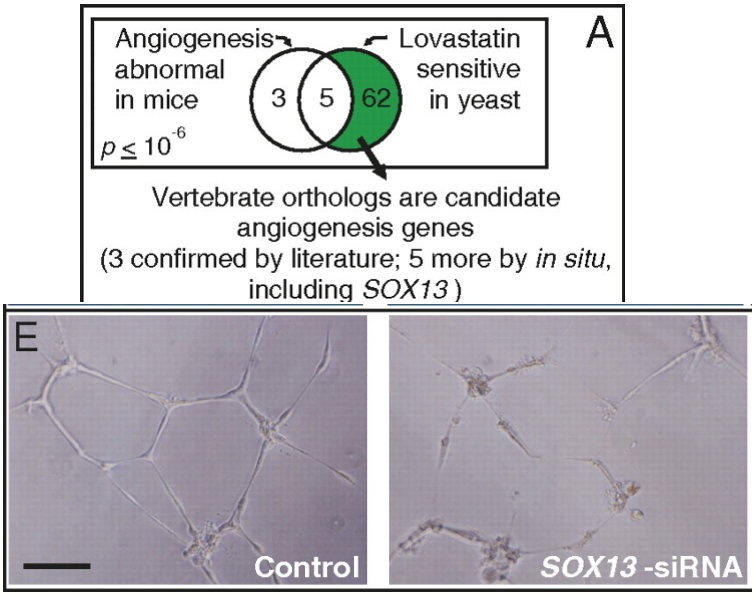
Khurana & Lindquist (2010) *Nat. Rev. Neurosci.*

Deep Homology Extends To Gene Networks



McGary et al. (2010) PNAS

A Yeast Model for Angiogenesis



McGary et al. (2010) PNAS

Genomics: "Big Science" Driven by a Few Centers

wellcome trust sanger institute

BCM Baylor College of Medicine

BROAD INSTITUTE

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

J. Craig Venter INSTITUTE

THE Genome CENTER AT WASHINGTON UNIVERSITY

High-Throughput DNA Sequencing Technologies

Helicos		
2 x 55 bp	~8 days	35 Gb

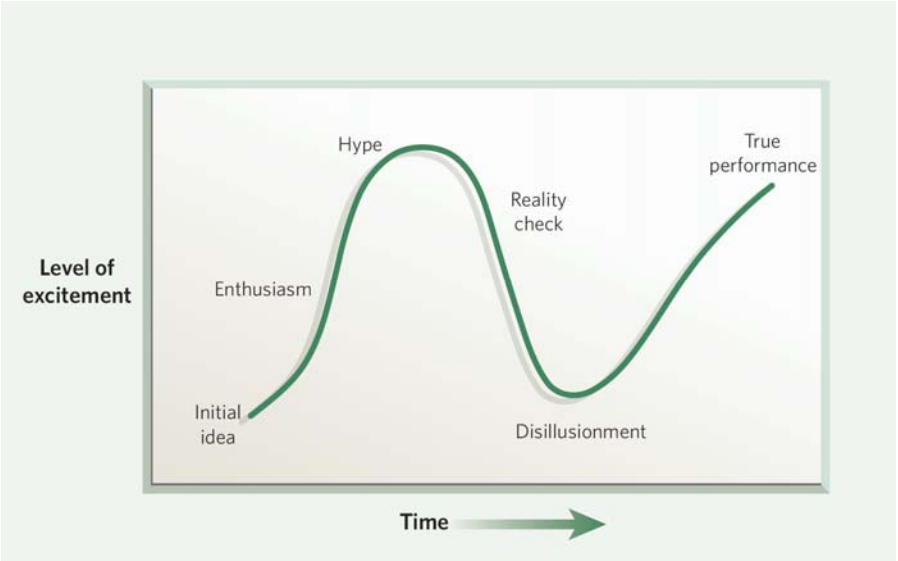
Illumina		
2 x 100 bp	~8 days	200 Gb

454 / Roche		
2 x 400 bp	~0.4 days	0.6 Gb

SOLiD ABI		
2 x 75 bp	~14 days	300 Gb

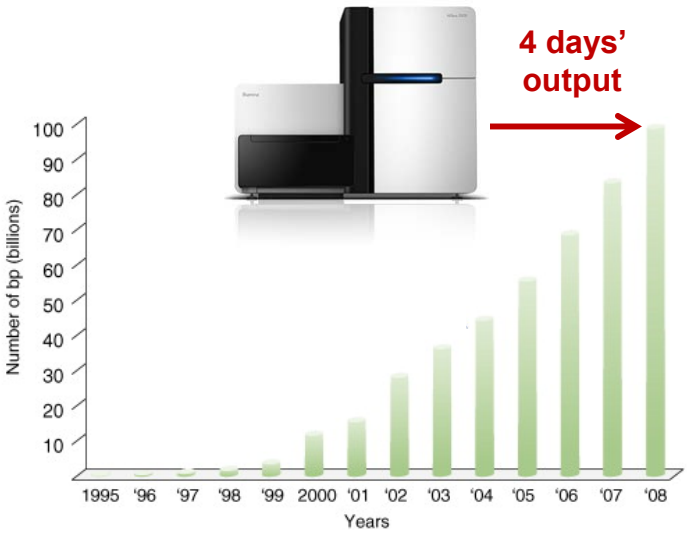
For the latest specs consult: http://ngsbuzz.blogspot.com/

"New Technology Level of Enthusiasm" Curve



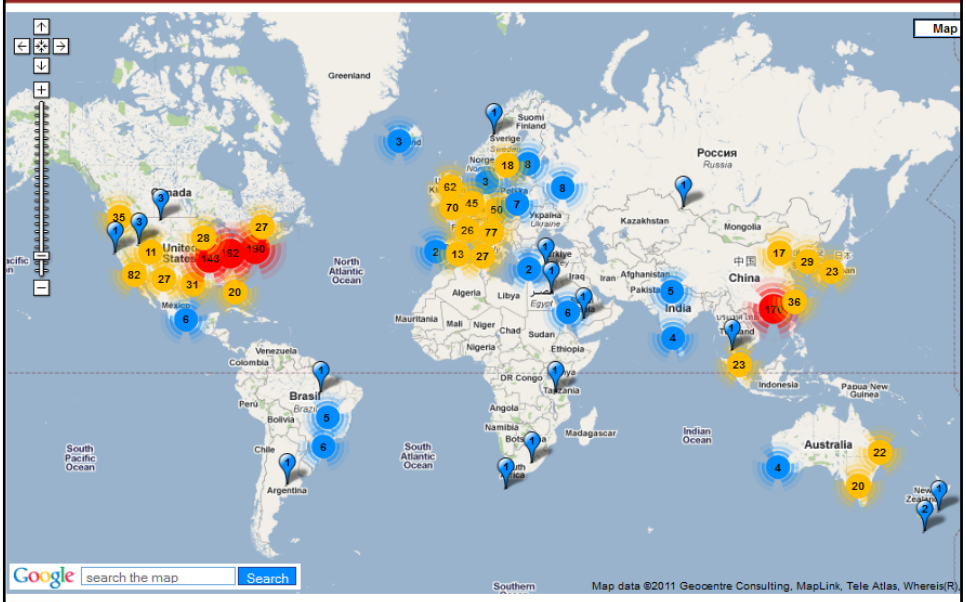
Collins (2006) Nature

Why Is Next-Generation DNA Sequencing So Exciting?



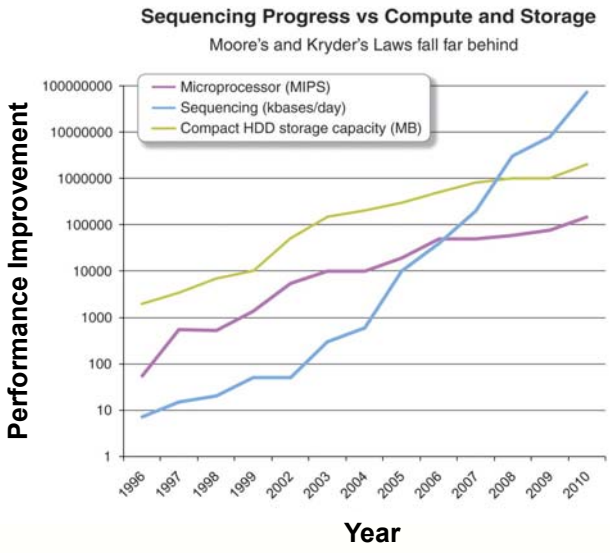
Gilad et al. (2009) Trends Genet.

World-Map of High-Throughput Sequencers



http://pathogenomics.bham.ac.uk/hts/

Sequencing is Outpacing Storage and Computation



Kahn (2011) Science

Meeting report

Genome informatics: taming the avalanche of genomic data

Erik LL Sonnhammer

Address: Center for Genomics and Bioinformatics, Karolinska Institutet, 171 77 Stockholm, Sweden. E-mail: Erik.Sonnhammer@cgb.ki.se

Comment

Tsunami

Gregory A Petsko

Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA.
E-mail: petsko@brandeis.edu

PERSPECTIVES

Preparing for the flood: evolutionary biology in the age of genomics

James R. Brown

Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data

CHRISTOPHER SOUTHAN
GRAHAM CAMERON
EMBL-European Bioinformatics Institute

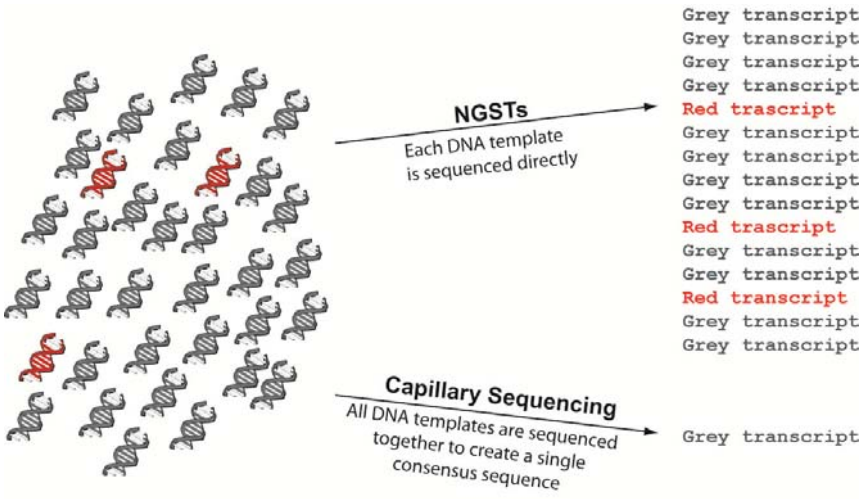
Lecture Outline

- **Introduction to evolutionary genomics**
- **Benchmarking *de novo* transcriptome sequencing for functional and evolutionary genomics**
- **Phylogenomics**

----------Coffee Break----------

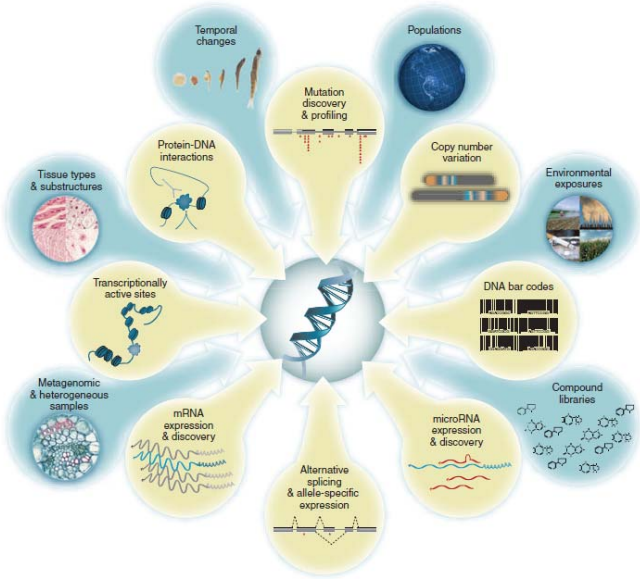
- **Comparative, Population & Functional Genomics**

Next-Gen Sequencing is Qualitative and Quantitative



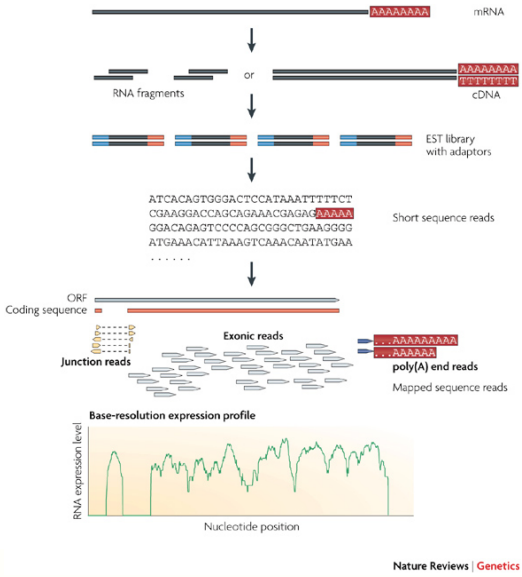
Rokas & Abbot (2009) Trends Ecol. Evol.

Sequence Census Assays



Kahvejian et al . (2008) Nature Biotech.

A Typical RNA-Seq Experiment

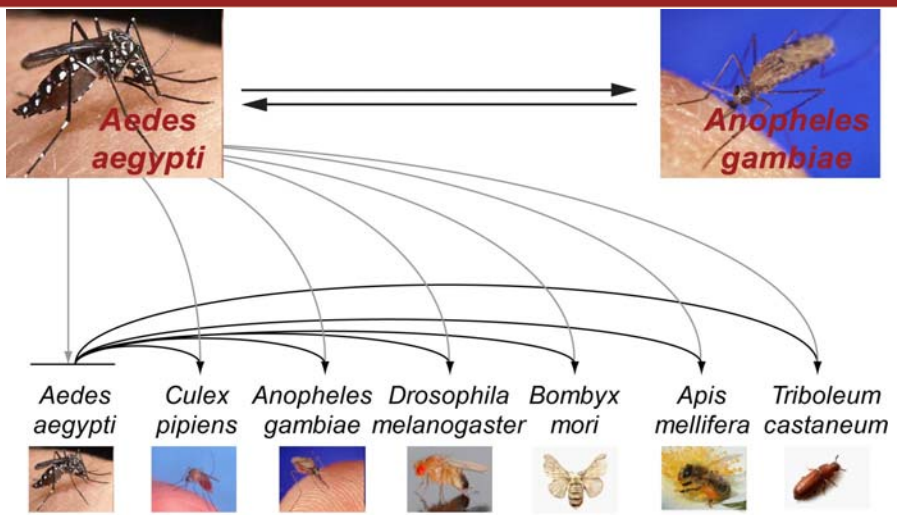


Wang et al. (2009) Nature Rev. Genet.

RNA-Seq for Non-Model Organisms: What's in it for You?

John Gibbons

The Next-Gen Revolution: What's in it for You?



How much? How good? How useful for functional studies? How useful for evolutionary studies?

Gibbons et al. (2009) Mol. Biol. Evol.

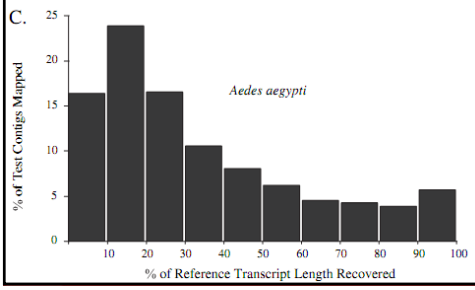
How Much?

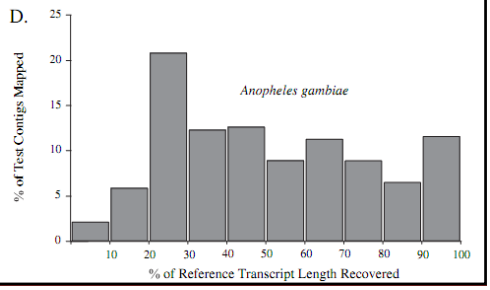
Aedes aegypti

27% of the transcriptome
34% of each transcript

Anopheles gambiae

20% of the transcriptome
50% of each transcript





RNA-Seq Data is Qualitatively Accurate

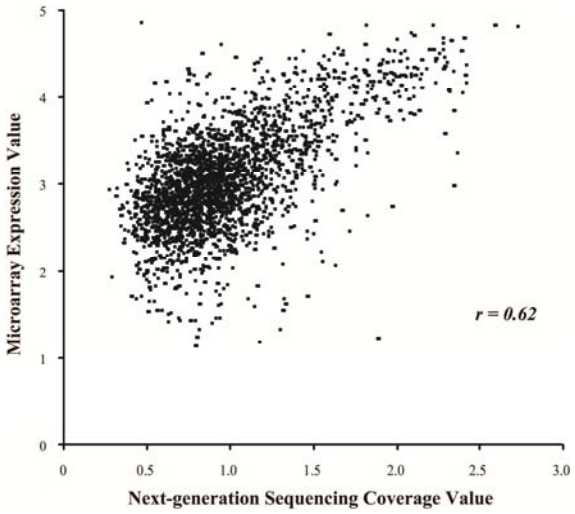
```
                Reference CAG CCA ATG AGG TCT CAG TCG
Substitutions               ||| ||| ||  ||| ||  ||| |||   3x10-3
                     Test CAG CCA ATC AGG TCA CAG TCG

                Reference CAG CCA ATG --- --- CAG TCG
Indels                      ||| ||| |||         ||| |||   4x10-5
                     Test CAG CCA ATG AGG TCA CAG TCG

                Reference CAG CCA ATG AGG TCT CAG TCG
Assembly Breaks               |       | ||| ||| ||| |||   2x10-4
                     Test ATG GGG TAC AGG TCT CAG TCG
```

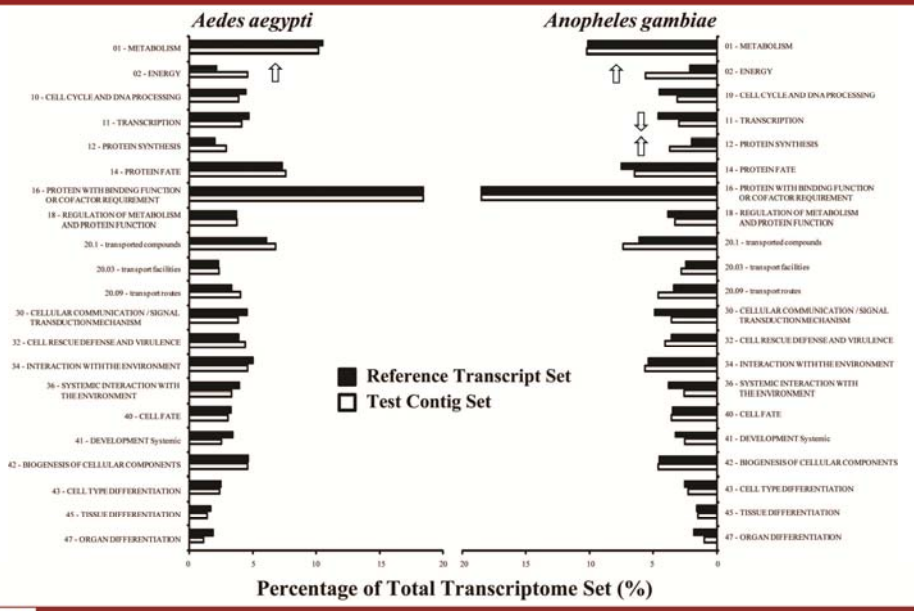

Gibbons et al. (2009) Mol. Biol. Evol.

RNA-Seq Data is Quantitatively Accurate



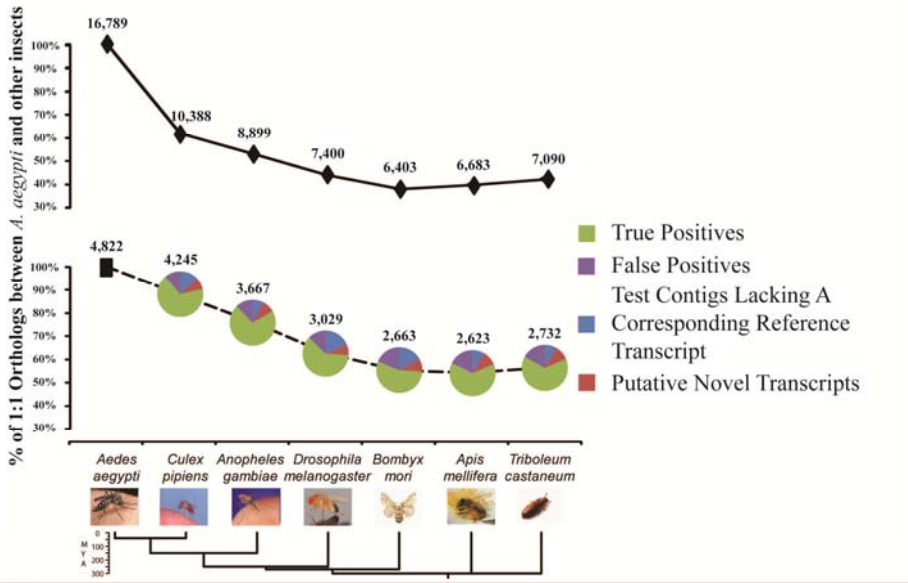
Gibbons et al. (2009) Mol. Biol. Evol.

RNA-Seq Data is Functionally Diverse



Gibbons et al. (2009) Mol. Biol. Evol.

Our RNA-Seq Data is a Goldmine for Molecular Markers



Gibbons et al. (2009) Mol. Biol. Evol.

What Can We Assemble de novo with Current Data?

Velvet & Oases *de novo* assembly of 10 million 100bp paired-end reads from *Anopheles albimanus*

	A. gambiae	*A. albimanus*
Transcriptome size	19.3 Mb	17.7 Mb
Gene number	13,683	16,952
Average gene length	4,542 bp	1,042 bp

Data from collaboration with Dinglasan lab (Johns Hopkins Univ.)

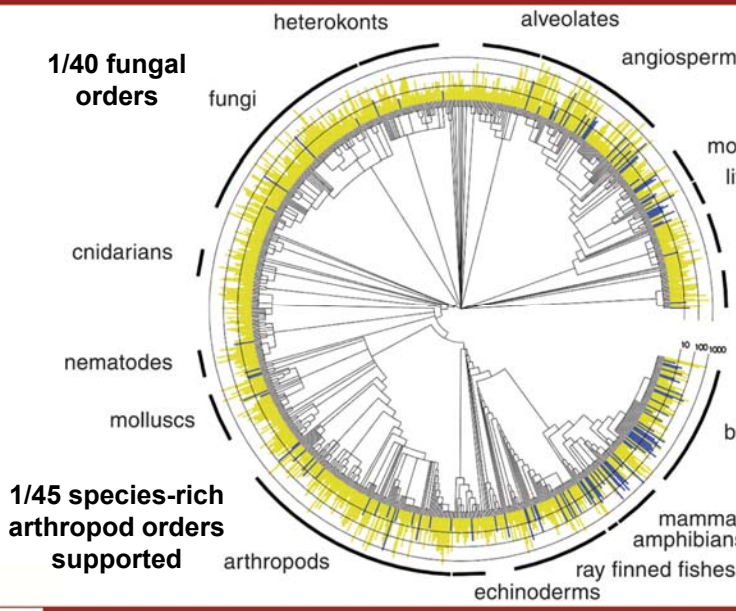
Lecture Outline

- Introduction to evolutionary genomics
- Benchmarking *de novo* transcriptome sequencing for functional and evolutionary genomics
- Phylogenomics

----------Coffee Break----------

- Comparative, Population & Functional Genomics

Taxonomic Breadth versus Genomic Depth



1/40 fungal orders

3/45 angiosperm orders

16/57 vertebrate orders

1/45 species-rich arthropod orders supported

Sanderson (2008) Science

Can we Use RNA-Seq to Increase Genomic Depth?

Targeting the transcriptome is a good idea because:

1. **Smaller than genome (coding parts cover 7% of *Anopheles gambiae* genome)**
2. **Fewer repetitive and transposable elements**
3. **Unequal representation (> 5 orders of magnitude)**
   - **Enriched for housekeeping and energy genes (they tend to be conserved)**
4. **The overwhelming majority of sequence evolution models have been developed for and tested in coding sequences**

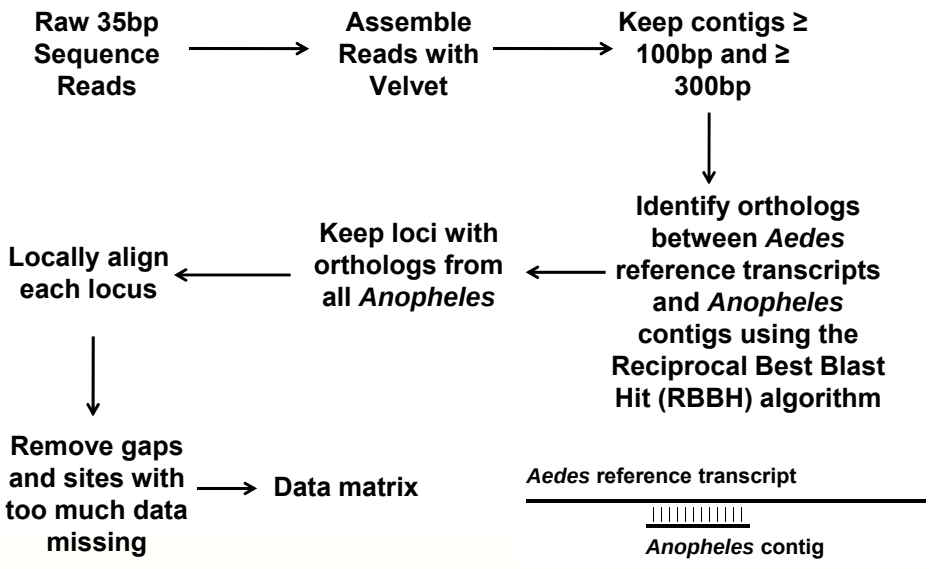
Hittinger et al. (2010) PNAS

Can we Use RNA-Seq to Increase Genomic Depth?

Species	Stock No.	Collection Location
Anopheles albimanus(Nyssorhynchus)	MRA-126	El Salvador
Anopheles arabiensis Cellia)	MRA-339	Zimbabwe
Anopheles dirus (Cellia)	MRA-700	Thailand
Anopheles farauti (Cellia)	MRA-489	Papua New Guinea
Anopheles freeborni (Anopheles)	MRA-130	USA
Anopheles gambiae (Cellia)	MRA-765	Liberia
Anopheles quadriannulatus (Cellia)	MRA-761	South Africa
Anopheles quadrimaculatus (Anopheles)	MRA-139	USA
Anopheles stephensi (Cellia)	MRA-128	India
Aedes aegypti (Stegomyia)	MRA-735	West Africa

Hittinger et al. (2010) PNAS

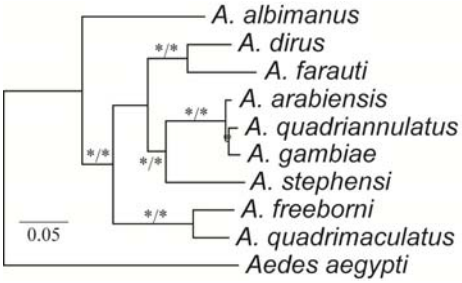
Data Matrix Construction: The "Singlecontig" Strategy



Hittinger et al. (2010) PNAS

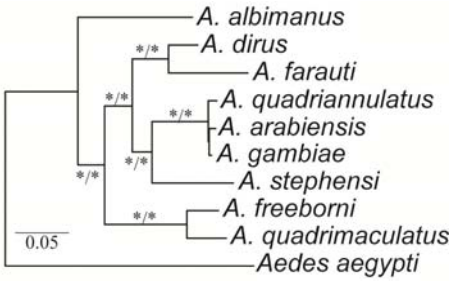
Robust Phylogenetic Inference from RNA-Seq Data

Using ≥ 100bp contigs



 # Loci = 553
Aln Length = ~390 Kb
% Missing data = 51

Using ≥ 300bp contigs

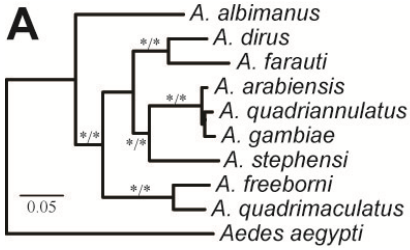


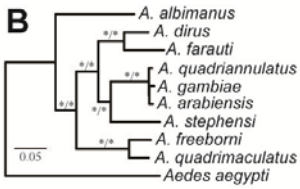
 # Loci = 69
Aln Length = ~73 Kb
% Missing data = 44

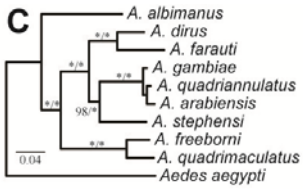
Hittinger et al. (2010) PNAS

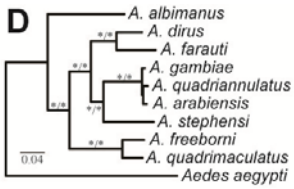
Accurate Phylogenetic Inference From our Data

553 loci
Aln L: ~390 Kb
Missing data: 51%









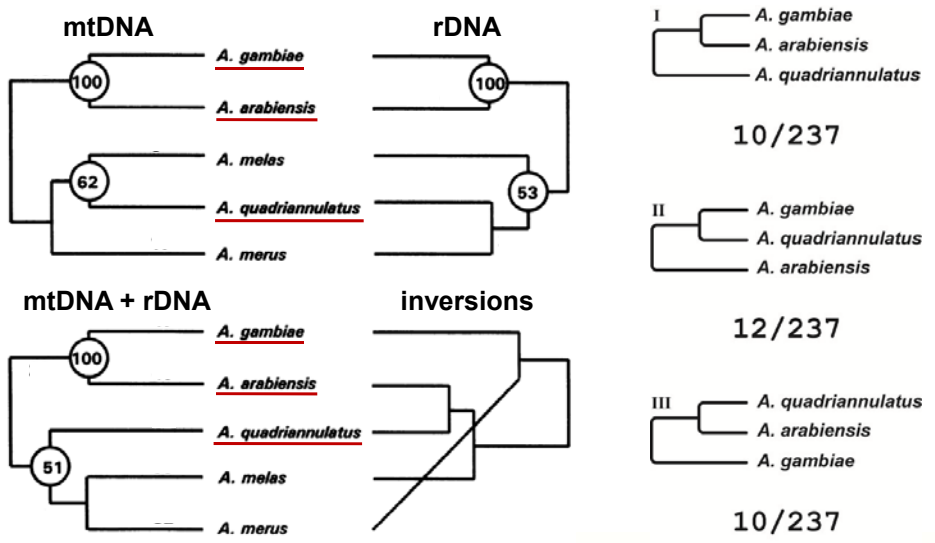
<u>Exclude erroneous loci</u>
491 loci
Aln L: ~329 Kb
Missing data: 50%

<u>Use only sites without data missing</u>
Aln L: ~15 Kb
Missing data: 0%

<u>Use *A. gambiae* as ref.</u>
634 loci
Aln L: ~472 Kb
Missing data: 50%

Hittinger et al. (2010) PNAS

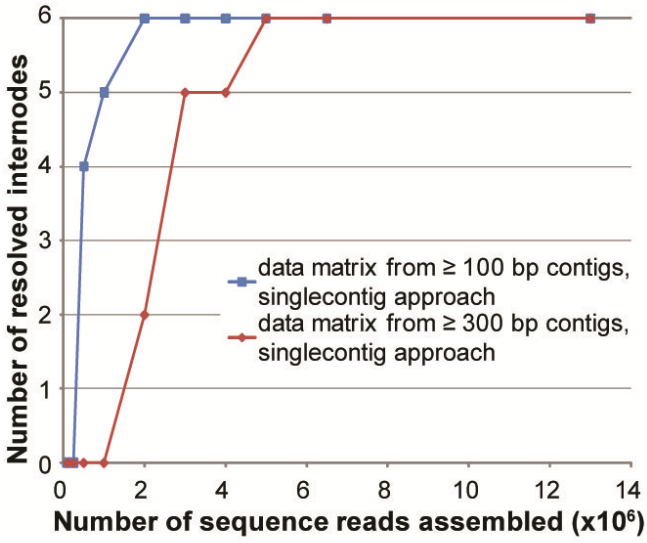
Our Data Matrices Can Detect Population-Level Events



Besansky et al. (1994) PNAS

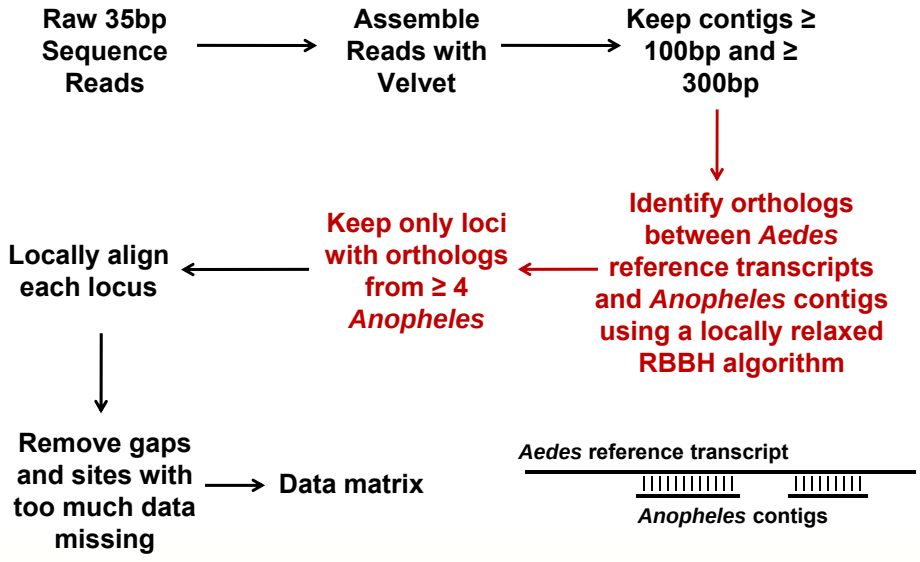
Hittinger et al. (2010) PNAS

Robust Phylogenetic Inference From Few Sequence Reads



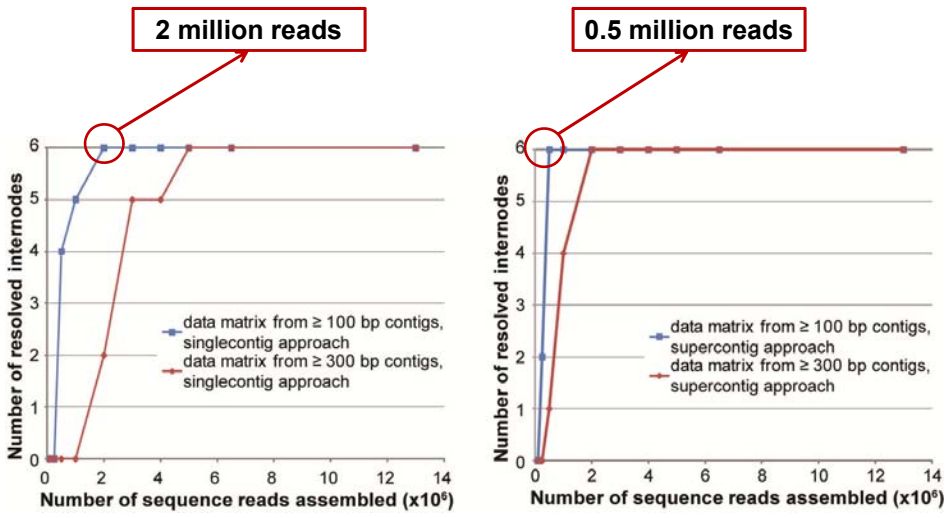
Hittinger et al. (2010) PNAS

Experimental Design: The "Supercontig" Strategy



Hittinger et al. (2010) PNAS

Robust Phylogenetic Inference From Very Few Reads



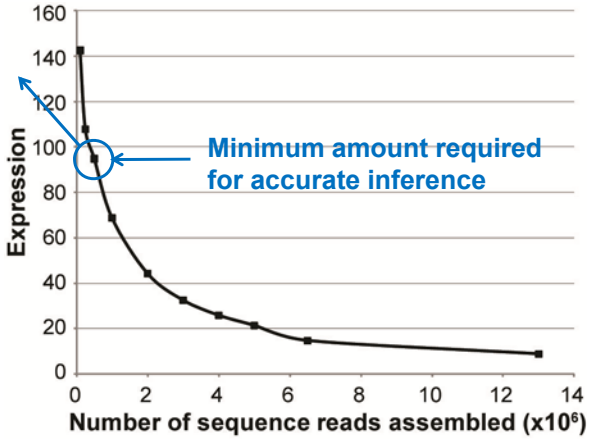
553 loci, AlnL: ~390Kb, %miss: 51

2,661 loci, AlnL: ~971Kb, %miss: 62

Hittinger et al. (2010) PNAS

Our Sequences are from Highly-Expressed Transcripts

2008 cost: ~$50
2011 cost: <$10



Hittinger et al. (2010) PNAS

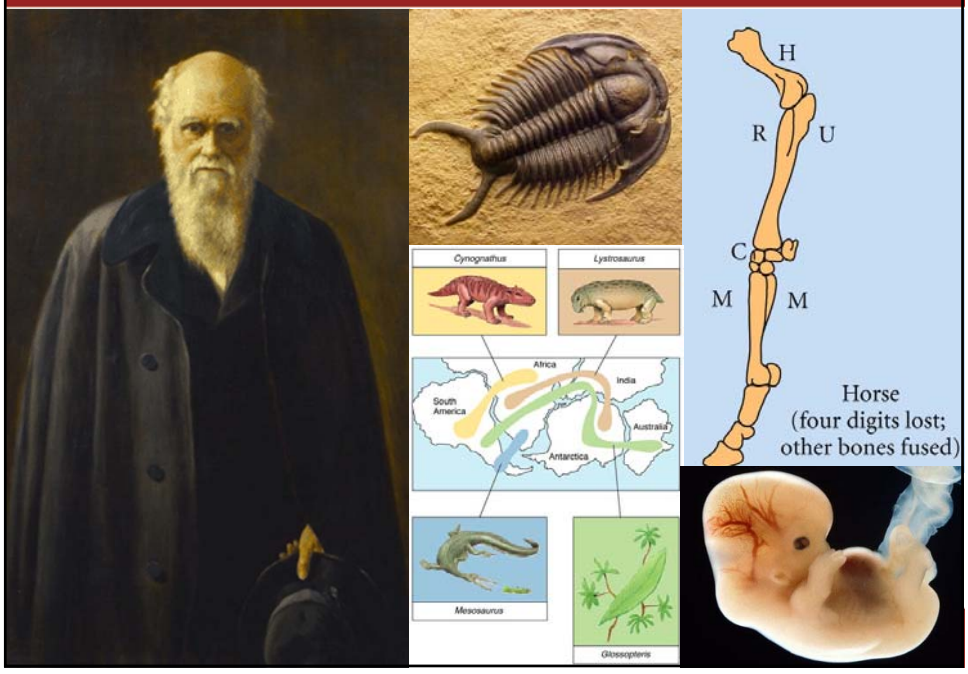
Coffee Break

The Molecular Foundations of the Fungal Lifestyle

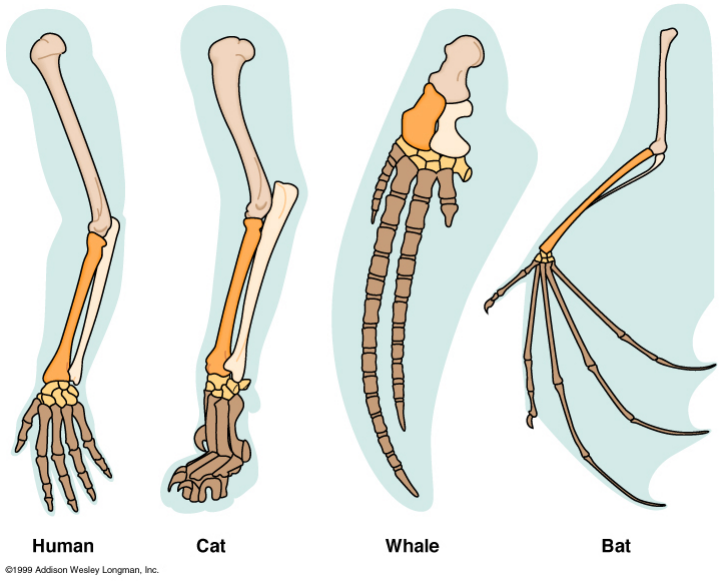
Antonis Rokas
Department of Biological Sciences, Vanderbilt University

http://as.vanderbilt.edu/rokaslab

Darwin's Data

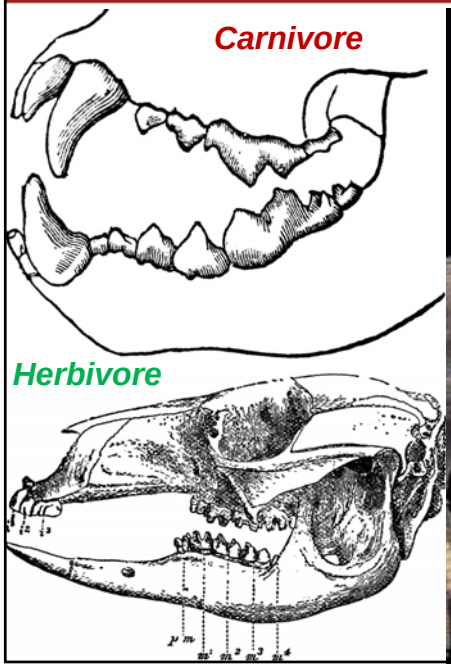


Similarity in Anatomy Suggests Common Origins



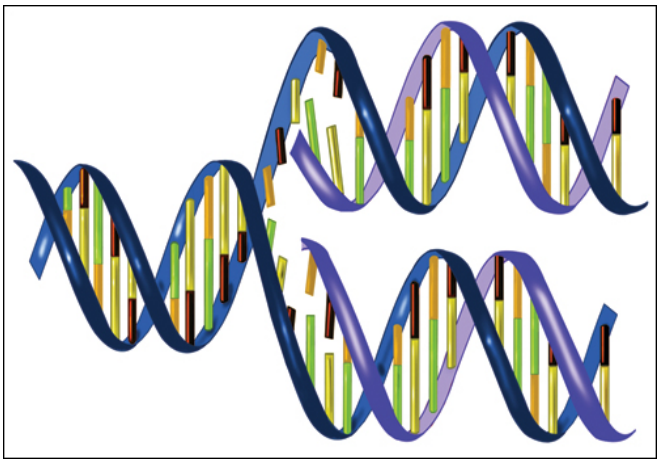
http://www.mun.ca/biology/scarr/139393_forelimb_homology.jpg

Differences in Anatomy Suggest Adaptations



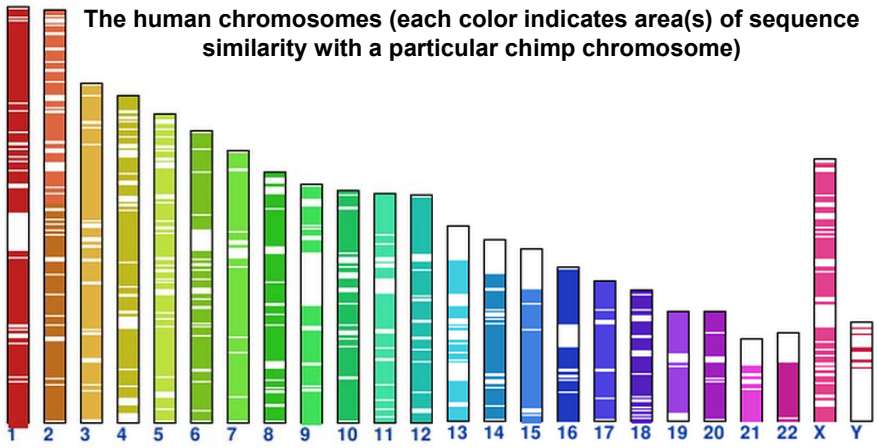
The DNA Record

The DNA record contains important clues about organisms' biological past, and their history of change and adaptation



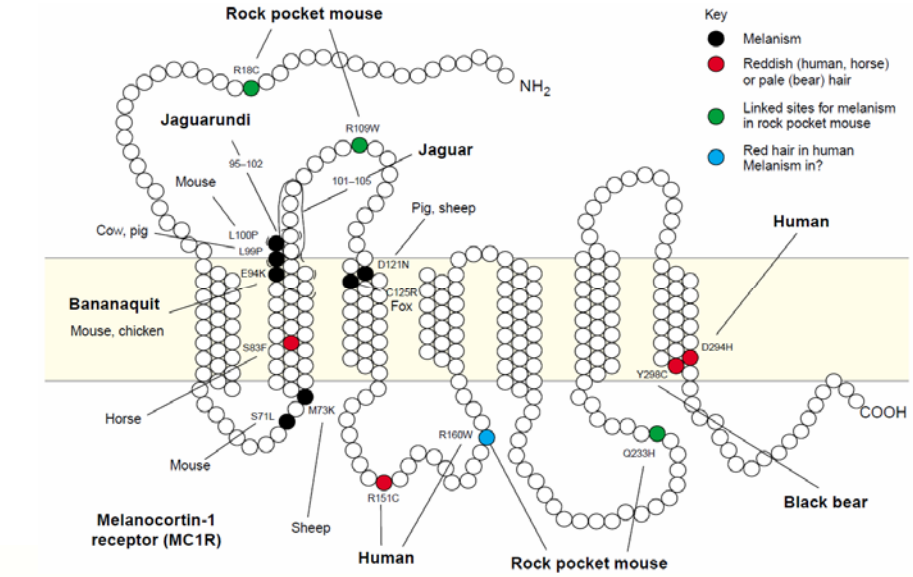
Similarity in the DNA Record Suggests Common Origins

The human chromosomes (each color indicates area(s) of sequence similarity with a particular chimp chromosome)



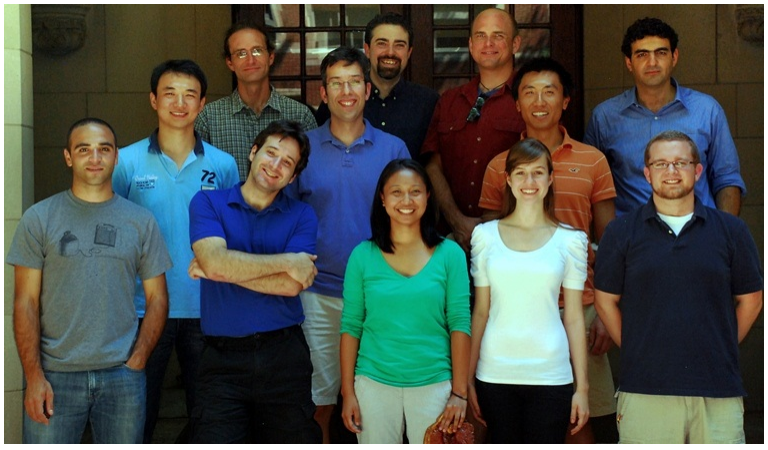
http://cinteny.cchmc.org/

Differences in the Record (Might) Suggest Adaptations



Majerus & Mundy (2003) Trends Genet.

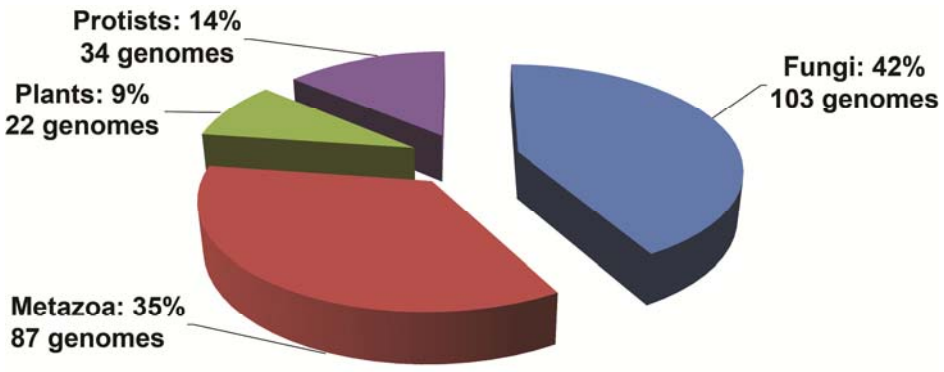
The Rokas Lab



We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches

Fungi: A Model for the Study of the DNA Record

Fungi: the Most Sequenced Eukaryotic Lineage

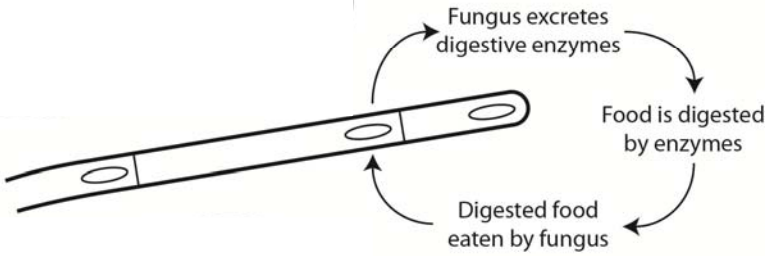


fungi: 1000genomes

http://www.jgi.doe.gov/sequencing/cspseqplans2012.html

Data from GOLD 3.0 (www.genomesonline.org), March 2011

Fungi Eat by Absorption



- **Fungi have evolved powerful "weapons" to defend their food**
- **Different fungi have specialized to "eating" different foods**

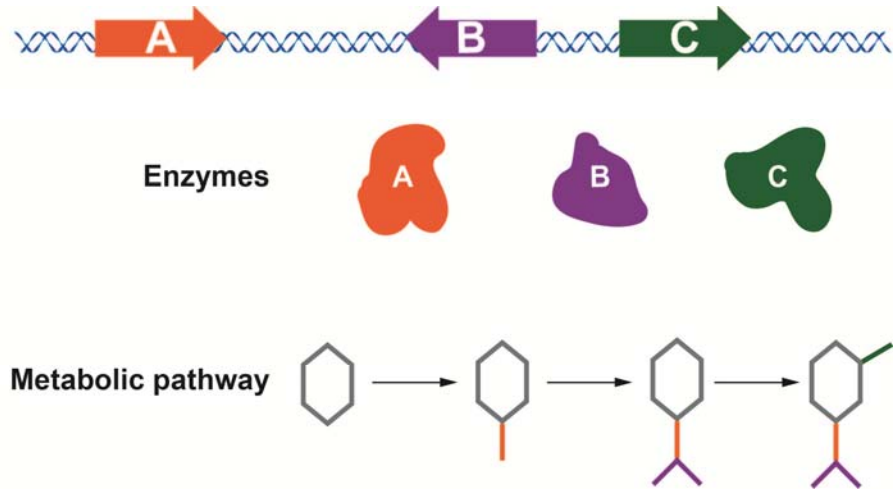
Fungi "Eat" Almost Everything, Grow on Any Surface

Wood, leaves, nails, leather, cloth, manure, animal carcasses, live hosts, ink, syrup, paint, glue, hair

What Fungi Cannot Eat!

Diner fries McDonald's fries

Fungal Metabolic Genes Are Often Physically Clustered



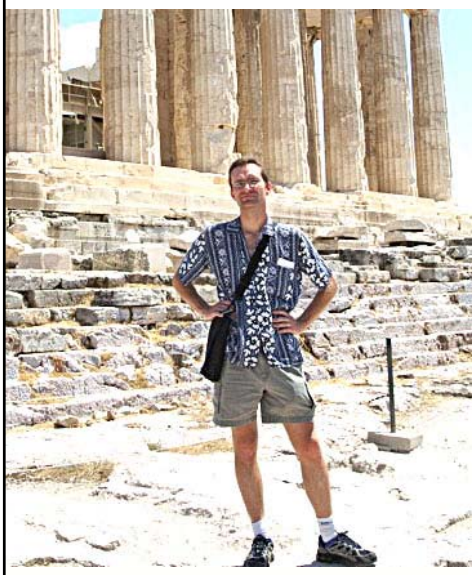
~1,000 metabolic genes in a typical fungal genome

~6% of metabolic genes reside in clusters

Questions

- **How are fungal metabolic clusters formed?**
- **Which are the evolutionary drivers of cluster formation?**
- **What are the evolutionary and functional implications of this clustering?**

The CRAP (Cluster Reconstruction And Phylogeny) Pipeline



Jason Slot

Kyoto Encyclopedia of Genes and Genomes

Retrieve genes from 173 pathways from 35 KEGG-annotated genomes (23 fungi)

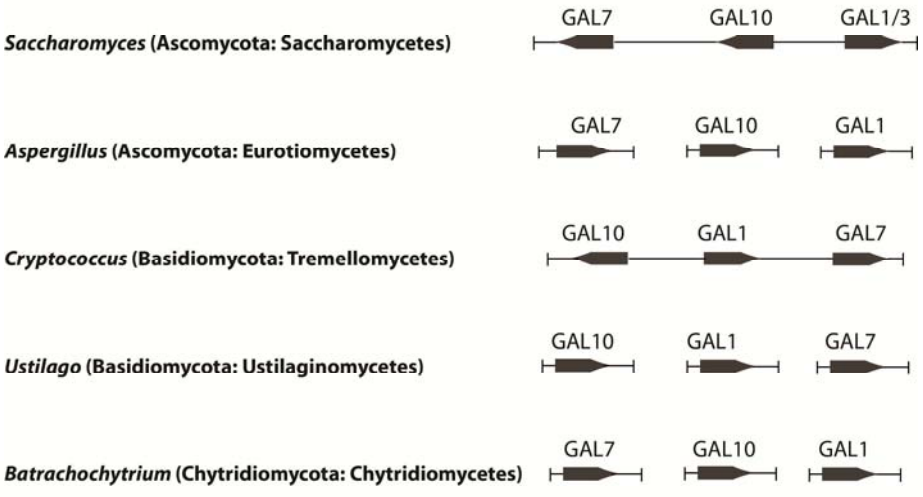
Search for presence of pathway homologs in ~130-200 genomes (80-150 fungi) by BLAST and OrthoMCL

Infer phylogenetic relationships of pathway homologs

Search for physical linkage (clustering) of pathway homologs

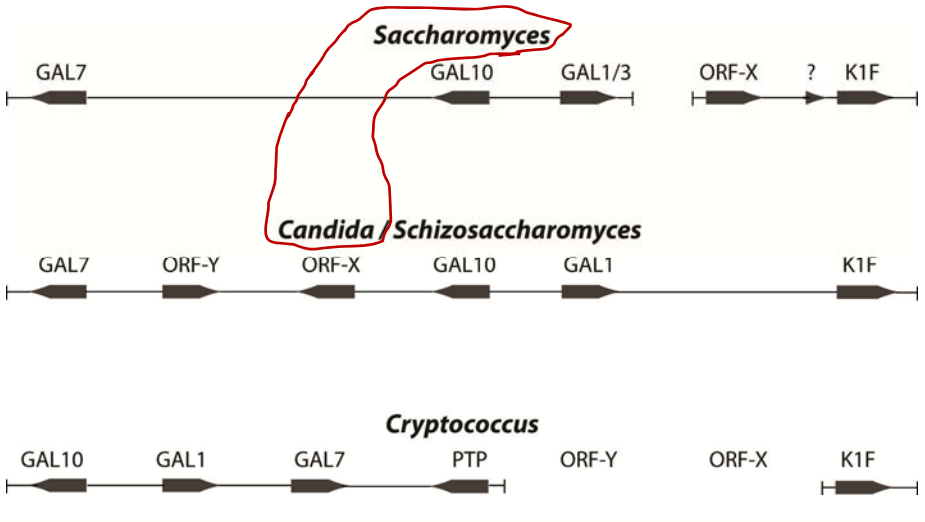
Variation in GAL Gene Clustering Across Fungi

Galactose —GAL1→ Galactose-1-Phosphate —GAL7→ UDP-Galactose —GAL10→ UDP-Glucose



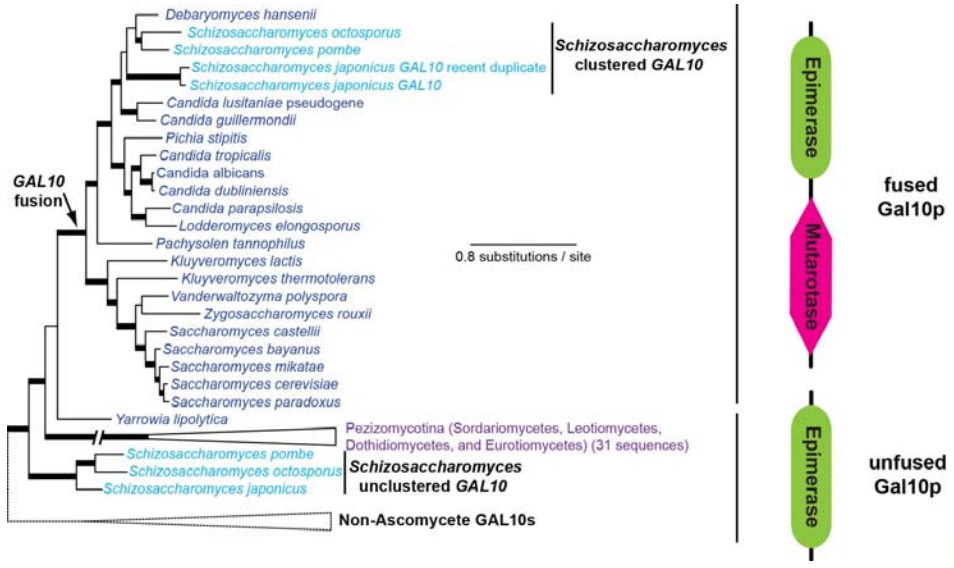
Slot & Rokas (2010) PNAS

Three Distinct Types of Fungal GAL Gene Clusters



Slot & Rokas (2010) PNAS

Schizosaccharomyces Gained Its GAL Cluster by HGT from Candida



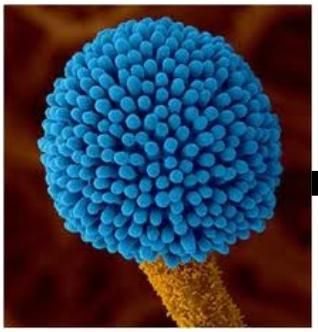
Slot & Rokas (2010) PNAS

Even Very Large Gene Clusters are Subject to HGT

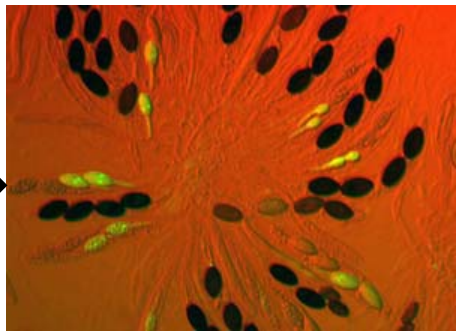
The Sterigmatocystin Gene Cluster

Structural genes
Regulators

In the US, agricultural loss due to aflatoxin contamination is ~ $270M



HGT



Aspergillus (Eurotiomycetes) *Podospora* (Sordariomycetes)

Slot & Rokas (2011) Curr. Biol.

Parasitol Res
DOI 10.1007/s00436-010-2098-1

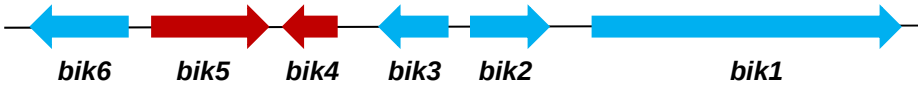
ORIGINAL PAPER

Larvicidal activity of metabolites from the endophytic *Podospora* sp. against the malaria vector *Anopheles gambiae*

Josphat C. Matasyoh · Birger Dittrich · Anja Schueffler · Hartmut Laatsch

Abstract In a screening for natural products with mosquito larvicidal activities, the endophytic fungus *Podospora* sp. isolated from the plant *Laggera alata* (Asteraceae) was conspicuous. Two xanthones, sterigmatocystin (**1**) and secosterigmatocystin (**2**), and an anthraquinone derivative (**3**) 13-hydroxyversicolorin B were isolated after fermentation on M_2 medium. These compounds were characterised using spectroscopic and X-ray analysis and examined against third instar larvae of *Anopheles gambiae*. The results demonstrated that compound **1** was the most potent one with LC_{50} and LC_{90} values of 13.3 and 73.5 ppm, respectively. Over 95% mortality was observed at a concentration 100 ppm after 24 h. These results compared

Does HGT of Metabolic Gene Clusters Contribute Anything Else?



The Bikaverin Gene Cluster

Structural genes
Regulators

HGT

Fusarium (Sordariomycetes)

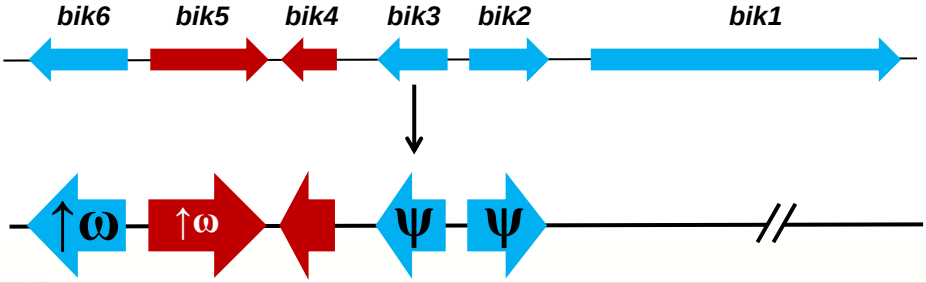
Botrytis (Leotiomycetes)

Campbell et al. (2012) Genome Biol. Evol.

HGT of Metabolic Clusters Also Contributes Novel Regulation

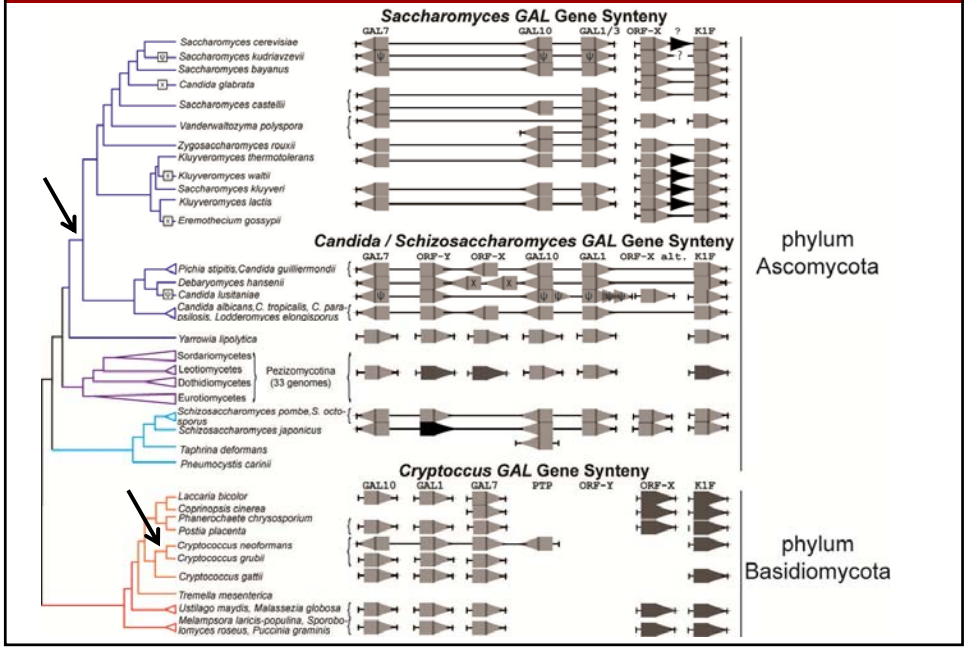
Botrytis (Leotiomycetes)

Structural genes
Regulators

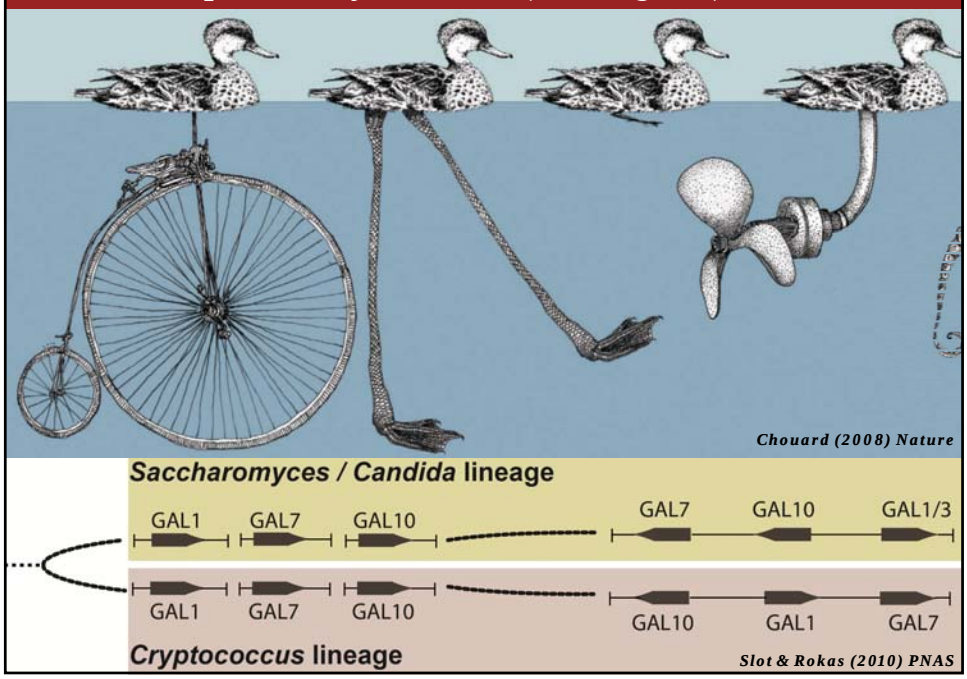


Campbell et al. (2012) Genome Biol. Evol.

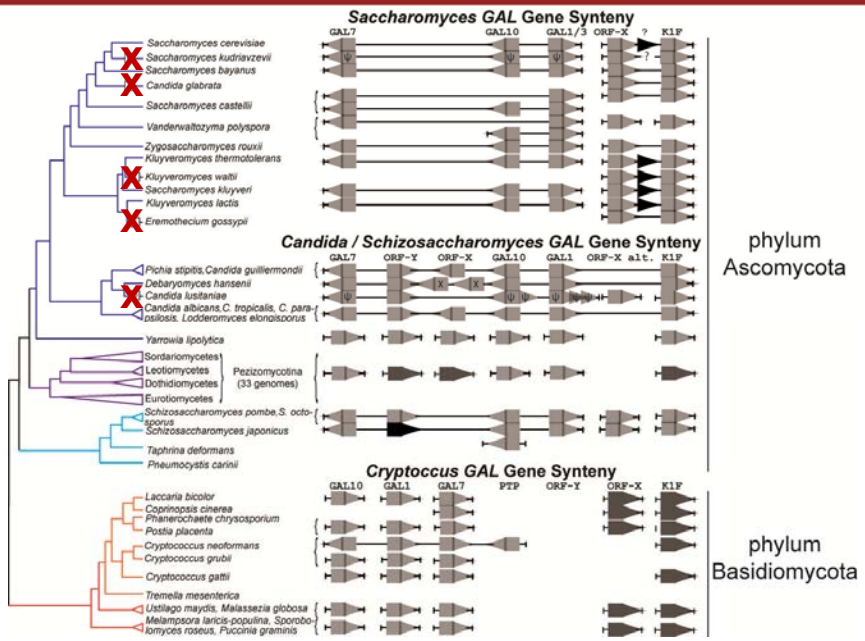
The Saccharomyces & Cryptococcus *Clusters Evolved Independently by Native Gene Relocation*



Independently Evolved (Analogous) Traits



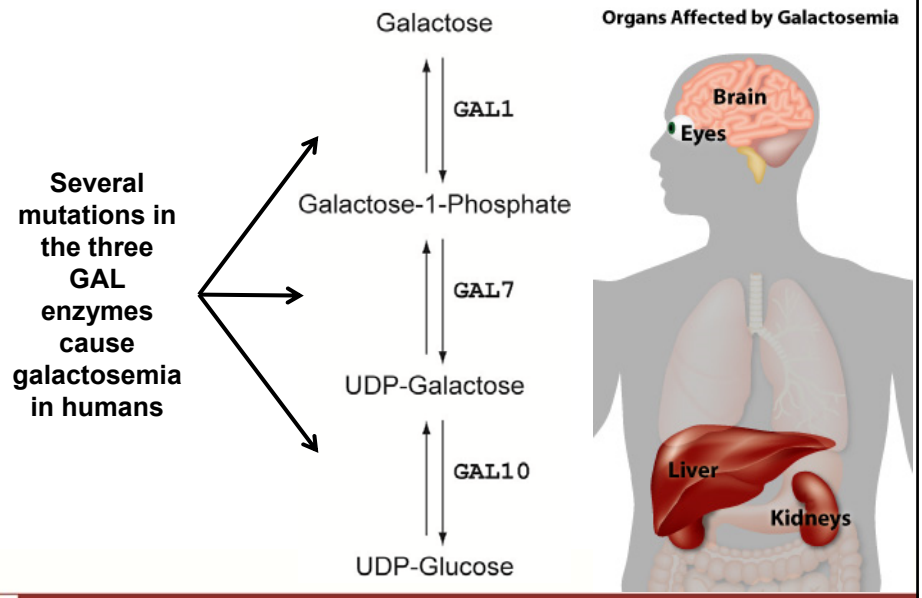
Clustered Pathways Are Easier Lost



Questions

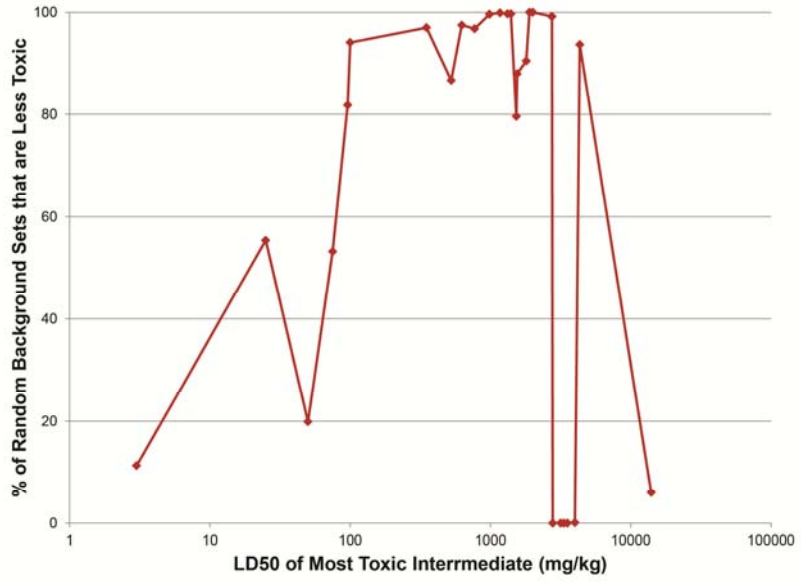
- How are fungal metabolic clusters formed?
  - Native gene relocation
  - Horizontal gene transfer
- Which are the evolutionary drivers of cluster formation?
- What are the evolutionary and functional implications of this clustering?
  - Genome remodeling
  - Novel metabolism and regulation
  - Avoidance of genetic addiction

Hypothesis: Clustering Reduces Impact of Toxic Intermediates



Fridovich-Keil (2006) J. Cell. Physiol.

Intermediates in Clustered Pathways Are More Toxic

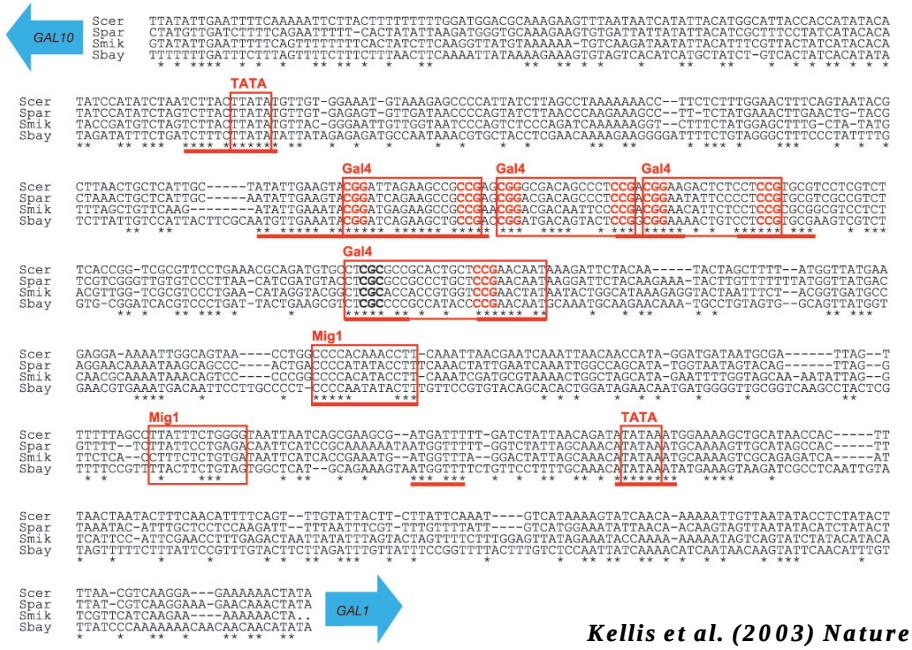


Kris McGary & Jason Slot (unpublished)

Mechanisms That Could Drive Clustering

- **Cluster selfishness**
- **Coordination of gene expression**
- **Genetic linkage**

Evidence for the Coordinated Expression Hypothesis



Kellis et al. (2003) Nature

Disrupting the GAL Cluster Doesn't Impair Fitness

~~Coordinated expression~~

Genetic linkage

Lang & Botstein (2011) PLoS ONE

But Selection Can Only Maintain, Not Bring Together

Theoretical Genetics of Batesian Mimicry
II. Evolution of Supergenes

D. CHARLESWORTH AND B. CHARLESWORTH†

Department of Genetics, University of Liverpool, Liverpool, England

(*Received* 24 *March* 1975)

In this paper, we describe an extension of the computer model explained in paper I to simulate the genetics of two loci, and have used it to study the selection of genes modifying the effects of mimicry genes, either by improving the mimicry towards the same model, or by giving some mimetic resemblance to a new model.

The results suggest a possible explanation of the phenomenon of imperfect mimicry. They also predict that, as appears to be the case in *Papilio dardanus*, populations will either become fixed for the mimicry gene or else eliminate any non-specific modifiers that are not extremely closely linked to the mimicry locus. This leads us to a consideration of the evolution of mimicry supergenes, which we conclude is unlikely to have occurred by means of unlinked or loosely-linked genes being brought together into one region of the chromosome.

Charlesworth & Charlesworth (1976) J. Theor. Biol.

S. kudriavzevii *is Polymorphic in Eating Galactose*

Strain	Nationality
IFO1802	Japan
IFO10990	Japan
IFO10991	Japan
IFO1803	Japan
ZP513	Portugal
ZP537	Portugal
ZP542	Portugal
ZP591	Portugal
ZP594	Portugal
ZP595	Portugal
ZP620	Portugal
ZP621	Portugal
ZP623	Portugal
ZP625	Portugal
ZP627	Portugal
ZP629	Portugal
ZP630	Portugal
ZP634	Portugal

2004: 4 Japanese GAL- strains

2007: 14 Portuguese GAL+ strains discovered

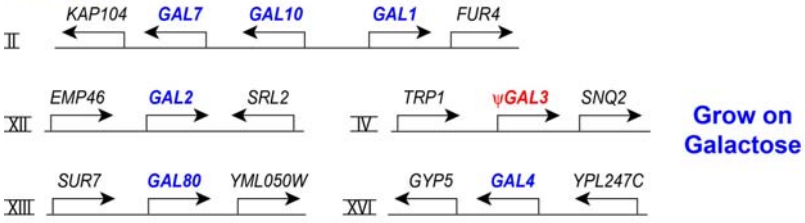
Illumina sequencing

80,000,000 reads
2,800,000,000 bp

Sampaio & Gonçalves (2008) Appl. Environ. Microbiol.

Two GAL *Gene Network States in* S. kudriavzevii

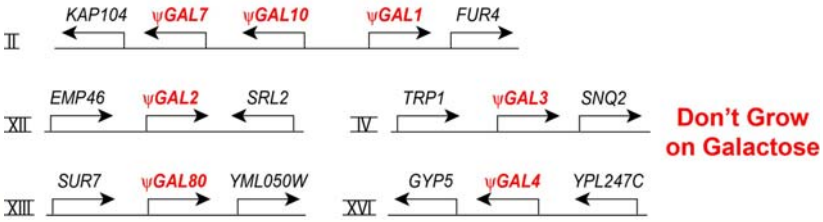
Portugal



Grow on Galactose

Not Hybrids / *GAL* genes Segregate Independently / No Post-zygotic Barriers

Japan

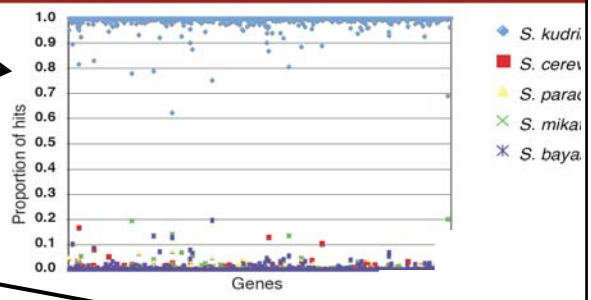


Don't Grow on Galactose

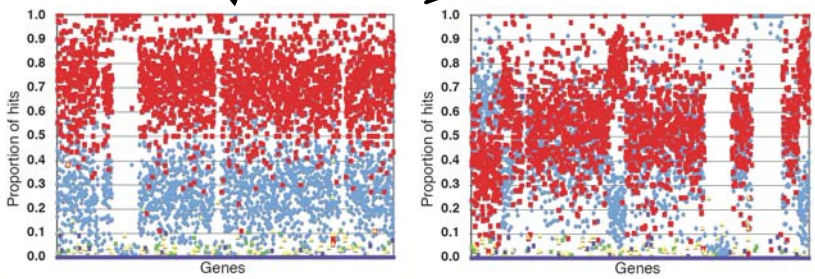
Hittinger et al. (2010) Nature

The GAL⁺ Isolates of S. kudriavzevii *are Not Hybrids*

Portuguese
S. kudriavzevii

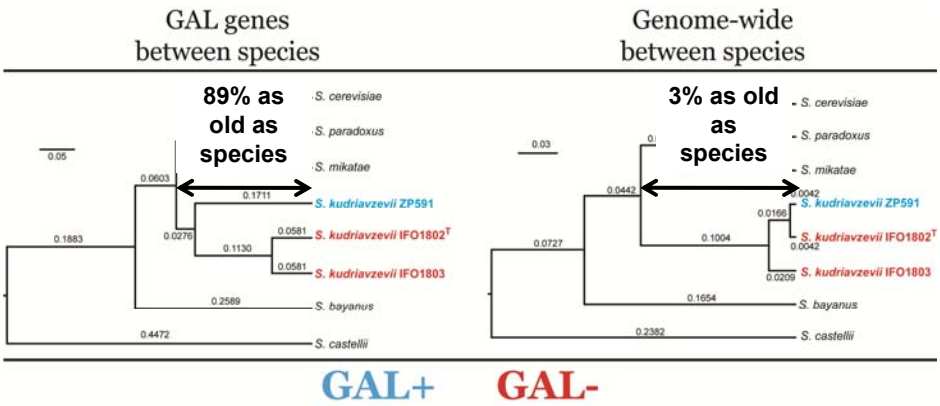


Known and wild-caught
S. cerevisiae X *S. kudriavzevii*
hybrids



Hittinger et al. (2010) Nature

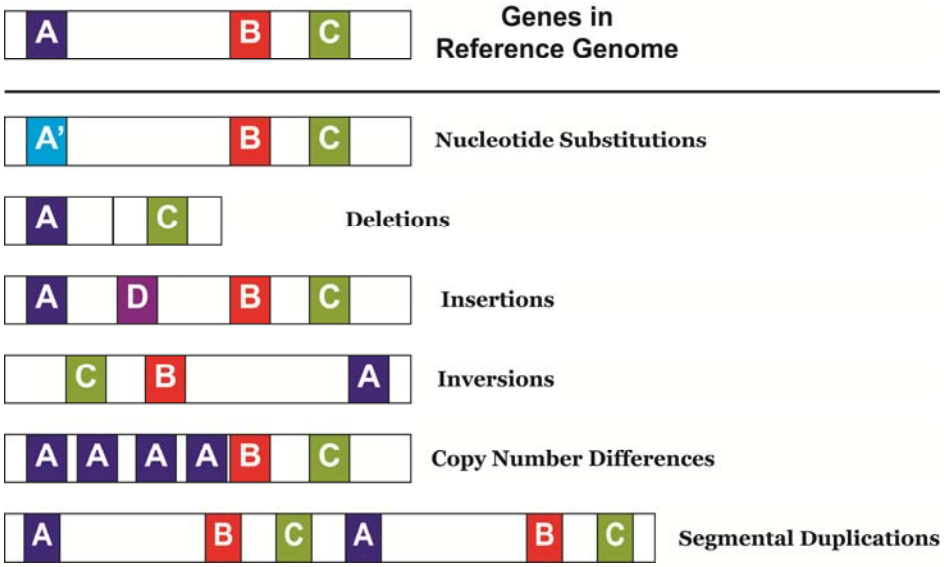
Ancient GAL Divergence Despite Recent Genome Divergence



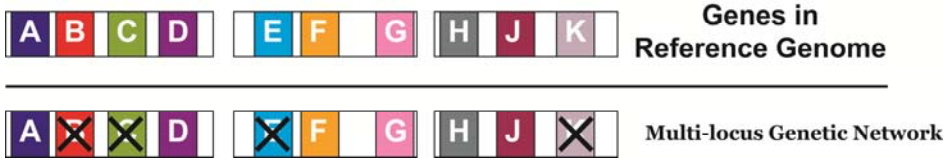
Only two non-*GAL* genes in the *S. kudriavzevii* genome significantly reject the lineage phylogeny in favor of the *GAL* gene phylogeny – both flank *GAL* loci

Ultrametric trees constructed using BEAST

Types of Genetic Variation Within Species



Balanced unlinked Gene Network Polymorphisms (BuGNPs)



How is this striking polymorphism in a multi-locus gene network maintained?

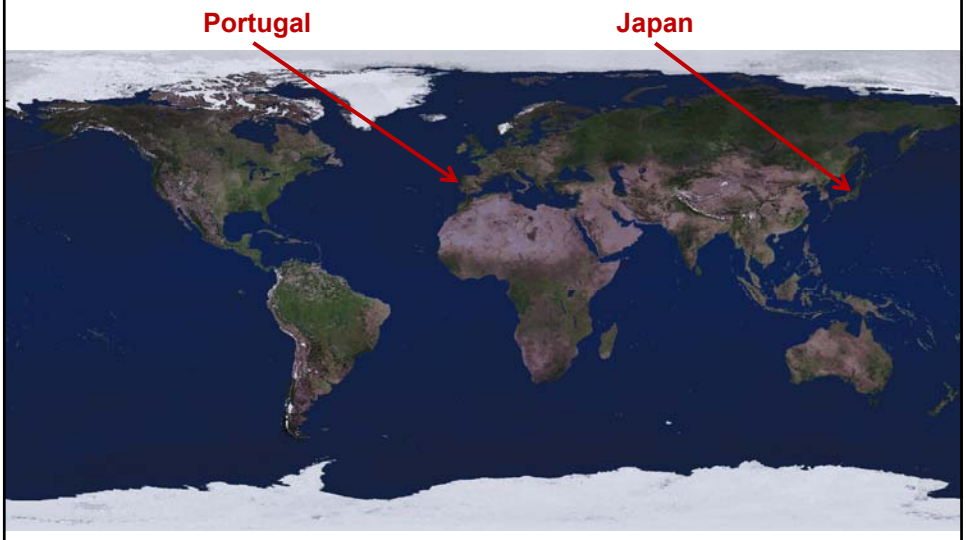
Hittinger et al. (2010) Nature

How Can We Explain BuGNPs?

"It is now generally understood that […] alleles at different loci may not be randomly associated with each other in a population. While this effect is generally regarded as a consequence of linkage, even *genes on different chromosomes* may be held temporarily or permanently out of random association by forces of *selection*, *drift* and *non-random mating*."

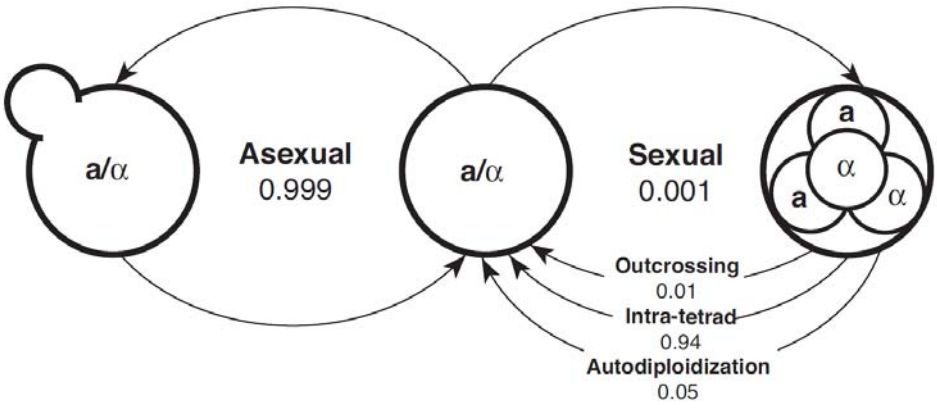
Lewontin (1998) Genetics

1. Maintaining BuGNPs Through Drift



http://nssdc.gsfc.nasa.gov/planetary/image/earth_day.jpg

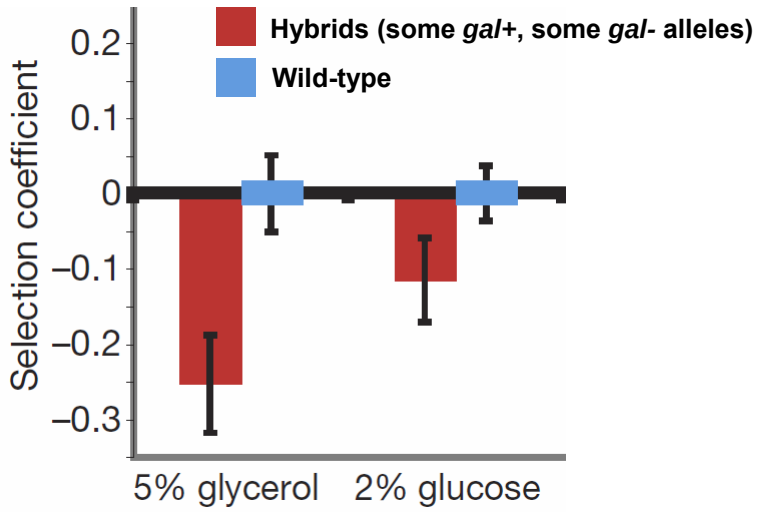
2. Maintaining BuGNPs by Selfing



The reproductive lifestyle of *Saccharomyces paradoxus*

Tsai et al. (2008) PNAS

3. Maintaining BuGNPs Through Selection Against Intermediates



Hittinger et al. (2010) Nature

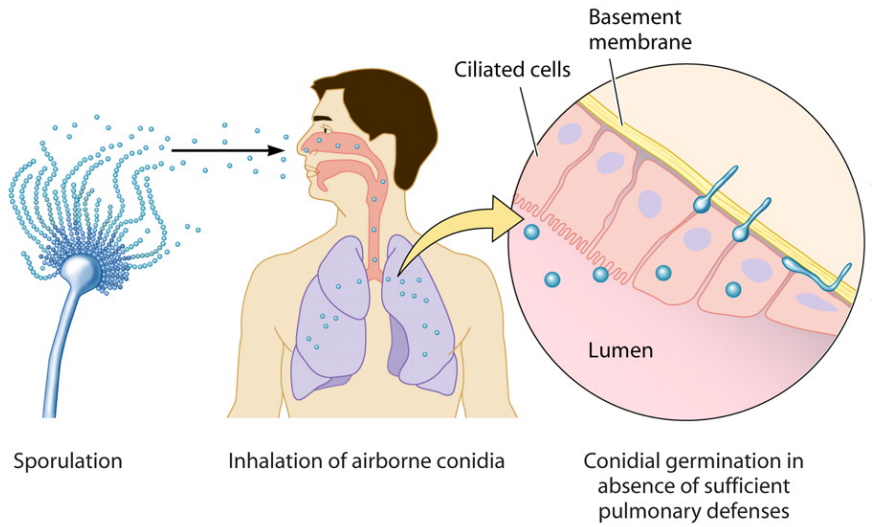
Questions

- **How are fungal metabolic clusters formed?**
  - Native gene relocation
  - Horizontal gene transfer
- **Which are the evolutionary drivers of cluster formation?**
  - Reducing impact of toxic intermediates
  - Genetic linkage > Coordinated expression > Selfishness
- **What are the evolutionary and functional implications of this clustering?**
  - Genome remodeling
  - Novel metabolism and regulation
  - Avoidance of genetic addiction

What Are the Functional Implications of Clustering?

John Gibbons

Aspergillus fumigatus *is an Opportunistic Pathogen*

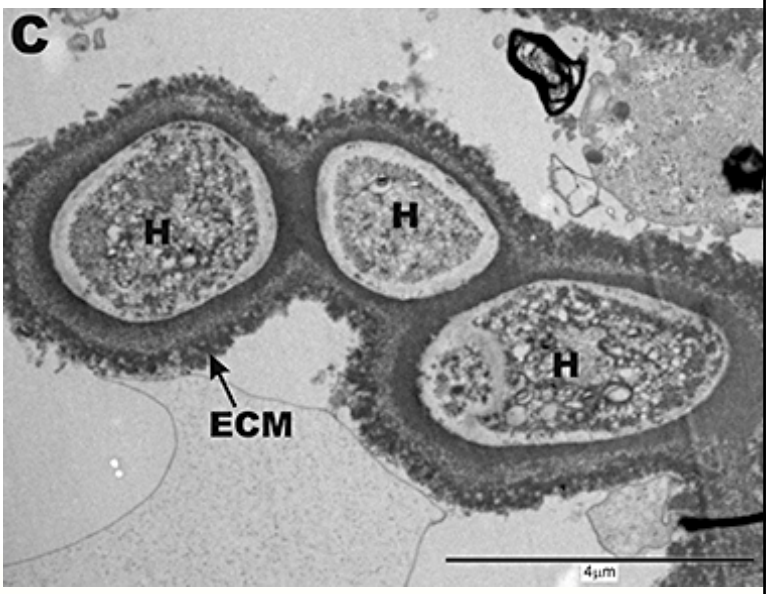


90% of aspergillosis cases, >50% mortality rate, multidrug resistant

Dagenais & Keller (2009) Clin. Microbiol. Rev.

Aspergillus fumigatus *Forms Biofilms In the Lung*

In the lung, the fungus usually forms a dense colony of filaments embedded in a polymeric ECM (i.e., a biofilm), which confers increased drug resistance



Loussert et al. (2010) Cell. Microbiol.; Beauvais et al. (2007) Cell. Microbiol.

Which Genes are Differentially Expressed During Biofilm Growth?

A. fumigatus
ATCC46645

Liquid growth

- Shaken
- Liquid BRIAN media
- 37^{0}C
- 16 hours

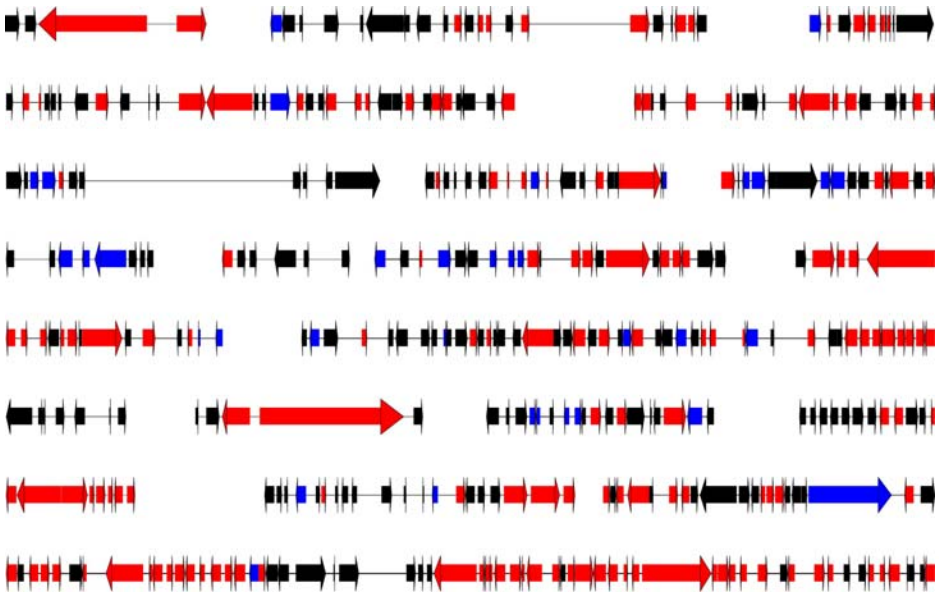
Biofilm growth

- Solid BRIAN agar
- 37^{0}C
- 16 hours

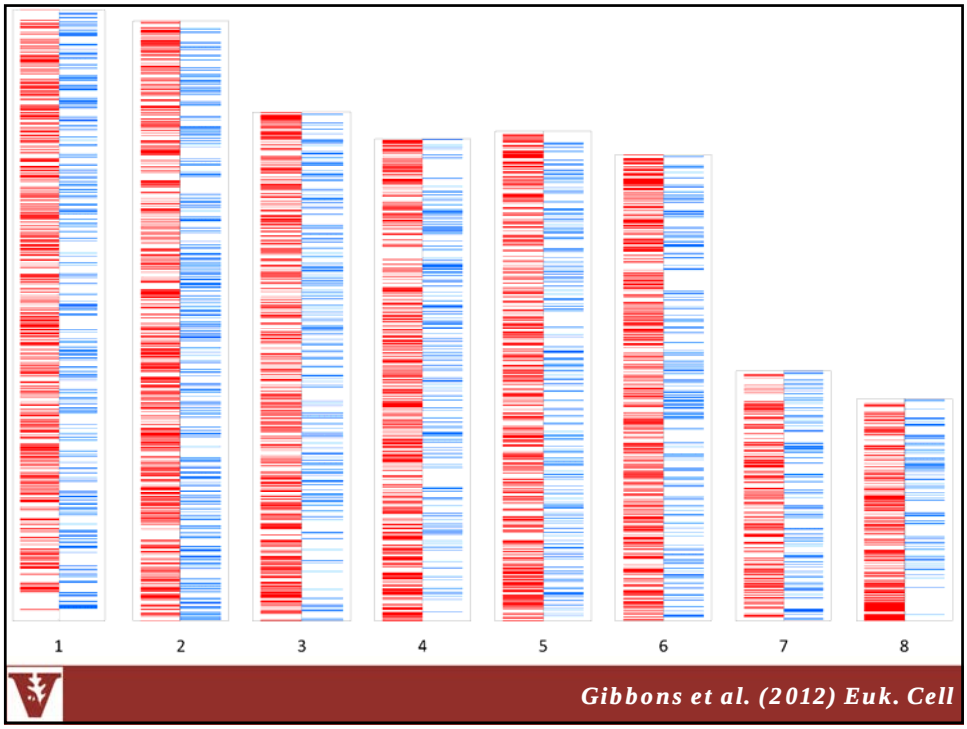
RNA-Seq using Illumina next-generation sequencing technology

Gibbons et al. (2012) Euk. Cell

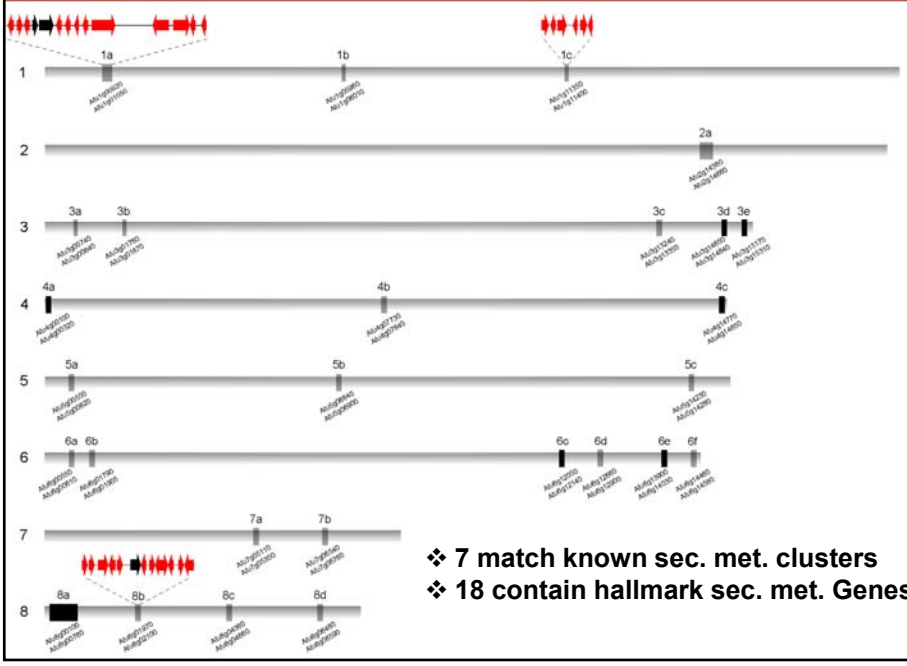
Up-regulated Sec. Met. Genes are Overrepresented in Biofilm



43% of genes up-regulated, 9% down-regulated (p-value< 10^{-17})

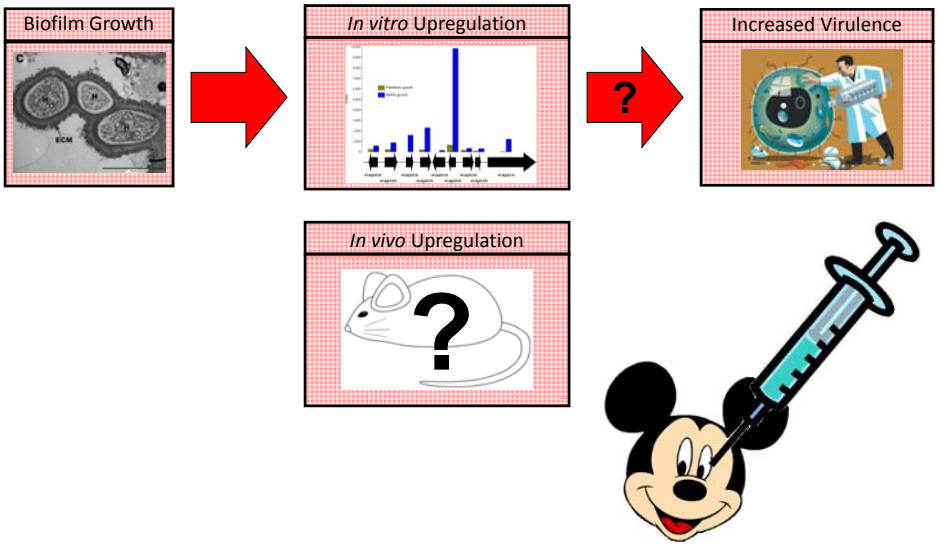


27 Up-regulated Gene Clusters in Biofilm



- 7 match known sec. met. clusters
- 18 contain hallmark sec. met. Genes

Are These Gene Clusters Involved in Virulence?



In collaboration with the Latgé lab @Institut Pasteur, France

The Birth, Evolution and Death of Gene Clusters

- **How are fungal metabolic clusters formed?**
  - Native gene relocation
  - Horizontal gene transfer
- **Which are the evolutionary drivers of cluster formation?**
  - Reducing impact of toxic intermediates
  - Genetic linkage > Coordinated expression > Selfishness
- **What are the evolutionary and functional implications of this clustering?**
  - Genome remodeling
  - Novel metabolism and regulation
  - Avoidance of genetic addiction
  - Virulence factors?

The Genomes of Non-Model Organisms are the New Frontiers

Epidemiology	What is the aetiology of emerging, uncharacterized infectious diseases? What are the core genomes of different microbes or pathogens occupying different niches or hosts?
Social Evolution	Are maternal care and sibling care behaviors regulated by similar patterns of gene expression?
Biodiversity	What is the extent of undocumented microbial diversity in different ecosystems? What is the microbial flora of the human gut?
Phylogenetics	What is the phylogeny of all vertebrate species or angiosperms using organelle or nuclear genome data?
Population Genetics	What is the variation in the mutation rate across organisms?
Experimental evolution	What is the genetic basis of phenotypes emerging during laboratory evolution?
Palaeontology	What are the evolutionary relationships of ancient organisms to extant taxa?
Evolution of development	How do regulatory networks evolve and rewire? What are the *cis*-regulatory targets of all transcription factors in a species and how do they evolve?

Rokas & Abbot (2009) Trends Ecol. Evol.

Personal High-Throughput Sequencers

Illumina MiSeq		**$125K**
2 x 150 bp	27 hours	1 Gb

454 / Roche GS Junior		**$100K**
2 x 400 bp	10 hours	35 Mb

Personal Genome Sequencer		**$49K**
1 x 100 bp	2 hours	10 Mb

Acknowledgements

Rokas Lab	Local Collaborations	Global Collaborations
John Gibbons	Abbot lab	Latge lab (Pasteur)
Jason Slot	Bordenstein lab	Johnston lab (WU/UCD)
Leonidas Salichos	Friedman lab	C. Klaassen (NL)
Kris McGary	Ganter lab (TSU)	Cummings lab (UM)
David Rinker	Muglia lab (CCH)	Geiser lab (PSU)
Patricia Soria	Patton lab	J. M. Lassance (Lund)
Xiaofan Zhou	Reynolds lab	King lab (Berkeley)
Matt Campbell	Sterling lab	Reynolds lab (UT)
Holly Elmore	Williams lab	Dinglasan lab (JHU)
Han Zhang	Zwiebel lab	Cunningham lab (Duke)
Rokas Lab Alumni		
Ioannis Stergiopoulos	John Tossberg	Natalie Christian
Chelsea Hamilton	Richard Minford	Douglas Denniston
Pad Mahadevan	Melissa Bridy	Rosalynne Korman

Chris Hittinger

Mark Johnston

Jean-Paul Latgé

SEARLE SCHOLARS PROGRAM

National Science Foundation WHERE DISCOVERIES BEGIN

National Institutes of Health The Nation's Medical Research Agency

Literature

1. Rokas & Abbot (2009) Harnessing genomics for evolutionary insights. *TREE 24:192-200.*
2. Gibbons et al. (2009) Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Mol. Biol. Evol. 26: 2731-2744.*
3. Hittinger et al. (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *PNAS107: 1476-1481.*
4. Hittinger et al. (2010) Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature 464: 54-58.*
5. Slot & Rokas A (2010) Multiple *GAL* pathway gene clusters evolved independently and by different mechanisms in fungi. *PNAS 107: 10136-10141.*
6. Slot & Rokas (2011) Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi. *Curr. Biol. 21: 134-139.*
7. Gibbons et al. (2012) Global transcriptome changes underlying colony growth in the opportunistic human pathogen *Aspergillus fumigatus*. *Eukaryotic Cell 11: 68-78*.

http://as.vanderbilt.edu/rokaslab/publications.html