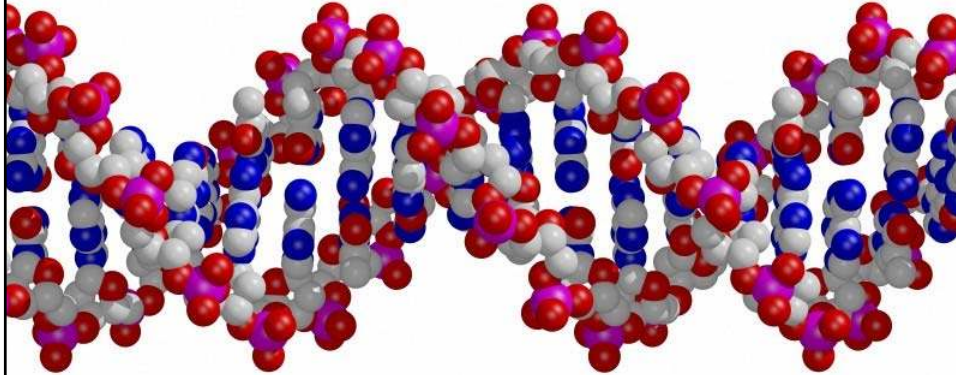


Evolutionary Genomics



**Antonis Rokas
Department of Biological Sciences
Vanderbilt University**



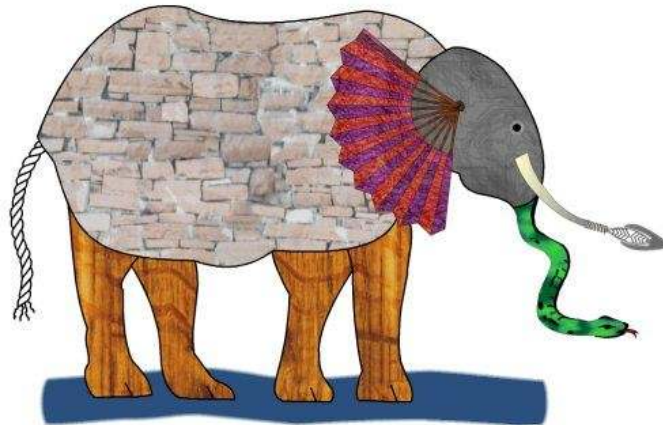
<http://as.vanderbilt.edu/rokaslab>

Lecture Outline

- ❖ **Introduction to evolutionary genomics**
- ❖ **Benchmarking *de novo* transcriptome sequencing for functional and evolutionary genomics**
- ❖ **Phylogenomics**
- Coffee Break-----
- ❖ **Comparative, Population & Functional Genomics**



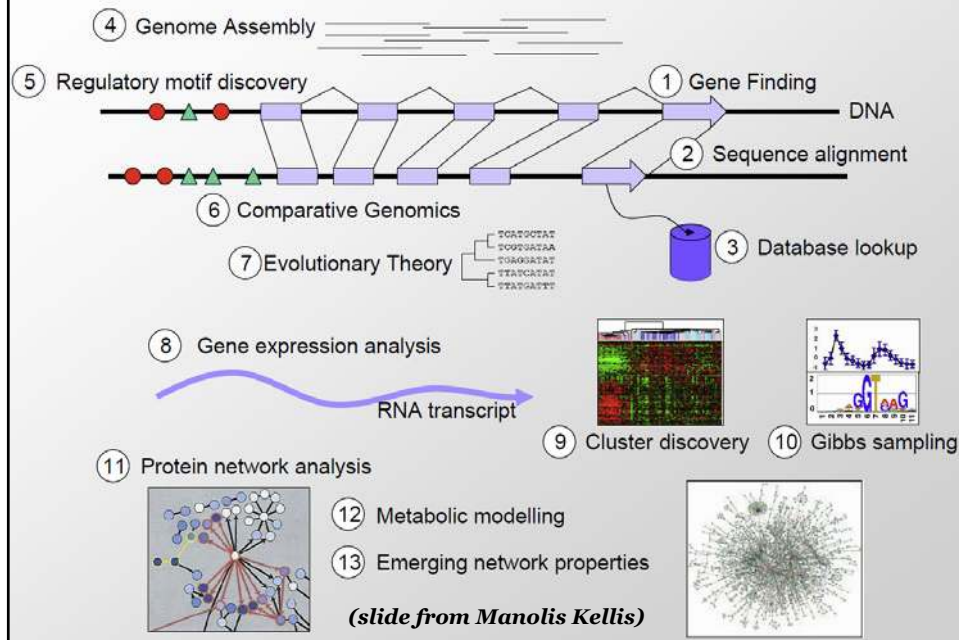
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
 GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCTTAAGAACGGAGCCTCTGTCAAGTCTCCAGCTGAACGTAGGCCGCGGG
 CCAGCCACTCATGAAATCGCCCTTTTCATTAGCATACTCTAGCGGCATTGATATCATCTTATACAGGAGCCATACATATATACTGACCTCA
 GCCGGCAATCACAAAAGGCACCCATCATACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAAGCGGTGCGACCCCGGGTCTCACC
 TATGTCGGAAAAAGATGGGCATTGGGCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
 ATCCTCTCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
 CTTACACTCCTGCCTGACTCTACTTTTCCAGCCTACCGTGCTTGGCGGCACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
 CATTGAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATGCTGCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
 ACGATAGAGGCAACGTTCTTGACCTCACTTTGCTCCTCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTAGAGT
 CAACATCAGATCATCGGCCACTCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
 GACCACCTCGCTTCTCTCACTACTCAGTTCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCTGGACAGTCTAGCT
 CATGGGTTAACCTTAGCAACTGCTAGTGCCTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
 GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGCTCTGTTCAAGAAACGACTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
 TGGCGAGATAAACTTACCAGTGACACAGATCAAGATGCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAAACCTT
 CCACTAAACGACCCCTTAAAGGCCAAACAGCCCTCCAGCAGGGGCTGGAATGAGAAACAGAGCTATTAGTCCGTAATCTTCTCAGAAT
 ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTTAGAGAAGCA
 TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCCTCAAGTCT
 GCGATATTCTGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAACT
 AACGCTCGACCCCGCAATTATGGCTCAAGGTTAGACTACGCTCCTGTGAGCCTTGATATGCAAGATTAGTTCTGCGATTGAAATCTAAG
 AGGATCTAATGGTAAGCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
 ATTTGCGGTTTATGCTGTCTTAAAAATGTTATACGCCCTCGGCGAAGAGAGGTCAACATTAATAGGCCCTCCTGGGATGTTTAAAGAT
 GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCTACATCAGGCGCTTAATACCAGAAATTTAAACAAGCGGAGGAGATCA
 GGACATGTTCTTGGTAACCATCAGCCAACGTATAGAGACCGACGACCAACATCCTGACATTGAGATATTTACCTCTAGTCAGGAAAA
 GGGAAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGCTATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAAATGAAAG
 AAGTTATTCGTAAGAGCTCTAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTACGCCAATCATCAACAC
 CCCCCACTGCTGGACAGAGGACTCTAAGGGGTTCTTCAAACCTTAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCAGGATCCTGCA
 ACAGTGTCTACTATGCAACGAAACAACAGCCGATCCCTACAAAATCTACCCAGTTACAGAACCCTCTGCACTGGAAGCAATTACTG
 ACAGTCCCCTGCTGGTGAAGCTTCTCAGGAGAACAGCCAAATTCGCGACTCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
 ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCTCAAGTGCAGAGGGCCACCAAGCAACATGTTT
 CAGTCTGACAACCTAAGAGCCTGGCATCACTGCGGAAGATCGTTCTGGATCTGTTTGAAGCCCTTGGGCCACGCGAAACAGGAA
 AGGGATTACAGTGGCGATTCTTACATTATATGGGCCAGCGATTGGAACCCCTCCGCTCGTAGATTTTCTGCTGGGGCAACTCTTTT
 TCGGATAGTGAACGATACCCGTTTTTACTTAGAAGGCTACGAATGGTATGATGATCATGTTTCAATGATAAGACATTCGTCAGT

Understanding Sequences Requires Tools and Evolution



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAACCCGGTTT
 GACGTTCTGTAGACGAACAACCTATGATATATCGACCCCGCTAAGAACGGAGCCCTGTGTCAGTGCTCCAGCTGAACCTAGGCCGCGGG
 CCAGCCACTCATGAAATCGCCCTTTTATTAGCATACTCTAGCGGCATTGATATCATCTTATACAGGAGCCATACATATATACTGACCTCA
 GCCGCAAAATCAGAAAAGGCCATCATACGAGTGCTTCTCCCAACAGACAGCTGGCTTGAAGCGGTCGACCCCGGGTCTCACC
 TATGTCCGGAAGATGGGCATTGCGGCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
 ATCCTCTCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
 CTTACACTCTGCTGACTCTACTTTTCCAGCCTACCGTGCTTCCGCGGCACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
 CATTGCGAGCCTACCACCTTTGCTGAGCCATTGTTGACTGGCTTGATCGCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
 ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTAGAGT
 CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
 GACCACCTCGCTTCTCTCACTACTCAGTTCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAGAGGGCTGGACAGTCTAGCT
 CATGGGTTAACTTAGCAACTGTAGTGCATATAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTGAGCCATGTTGGAATATT
 GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
 TGGCGAGATAAACTACCAGCAGTACACAGATCAAGATGCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCTC
 CCCTAAACGACCCCTTAAAGGCCAAACAGCCCTCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTCAGAAT
 ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCTGGGCTGTGTTGTCAGCATGCCCCCAACCATAGAACATTTAGAGAAGAACCA
 TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGATCTCTCTCTCCTCAAGTCT
 GCGATATCTGTTATACAGGATTGCCAGTTGAAACAATACTGCTTACGCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
 AACGCTCGACCCGGAATTTAGGCTCAAGGTTAGACTACGCTCTGTGATGCAAGATTAGTTCTGCGATTGGAATATCTAAG
 AGGATCTAATGTAAGCCCAAGGCTGCCATGGCTTTATTGATAGATTGATTTCTAGCTGACAATATGCAATTTGGGACAGGATCTGATG
 ATTGTCGGTTTATGCTGCTTCAAAATGTTATACGCCCTCGGCGAAGAGAGTCAACATTAATAGGCCCTCCTGGGATGTTTAAAGAT
 GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCTACATCAGGGCGCTTAATACCAGAATTTAAACAGCGAGGAGGATCAA
 GGACATGTTCTTGCATAAACCATCAGCCAACGTATAGAGACCGACGACGCAACATCTGACATTGAGATATTTACCTCTAGTCAGGAAAA
 GGGAACAGCACCCGCTATTTGGAGAGTGCTGCCAGCTCATAGCTACTGTCAGCTGTAGTAGCTGCTGACAGCACTCAATGAAAG
 AAGTATTCTGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTGACATCGGGTTACGCCAATCATCAACAC
 CCCCCTGCTGGACAGAGGACTCTAAGGGGTTCTTCAAACTTAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCAGGATCTGCA
 ACAGTGTCTACTATGCCAAGCAACCAACGCGCATCCCTACAAAATCTACCCAGTTACAGAACCTCTGCACTGGAAGCACTTACTG
 ACAGCTCCCGCTGGTGAAGCTTCTCAGGAGAACAGCCAAATTCGCGGACTCTACAGCTCCCGCTTACCCCAAAGCAATGATACTATT
 ATCGATCCCATTTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCTCAAGTGCAGAGGGCCACCAAGCAATGTTT
 CAGTCTGACAACATAAGAAGCCTGGCATCACTGCGGAAGATCGTTCTGGATCTGTTTGAAGCCCTTGGGCCAGCGGAAACCAAGGAAA
 AGGGGATACAGTGGCGATTCTCTACATTCATATGGGCCAGCGATTGGAACCCCTCCGCTCCGTAGATTTTCTGTCTGGGGCACTCTTTT
 TGGCATAGTGAACGATACCCGGTTTATACCTTAGAAGGCTACGAATGGTATGATGATCATGTTTCAATGATAAGACATTTCTGCAAGT



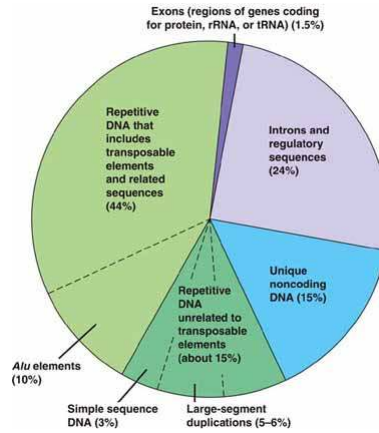
Transposon

Protein Binding Site

Exon

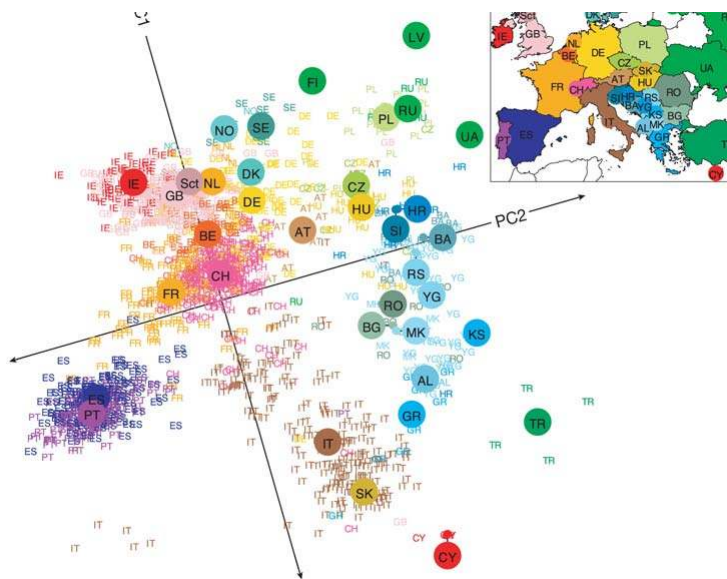
Intron

Organization of the Human Genome

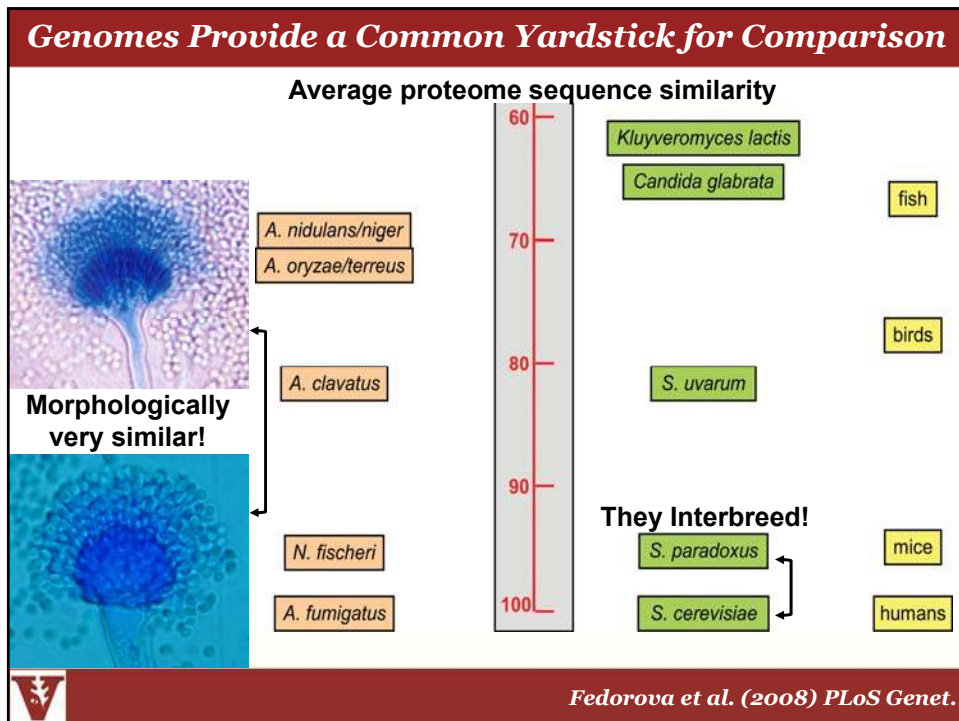
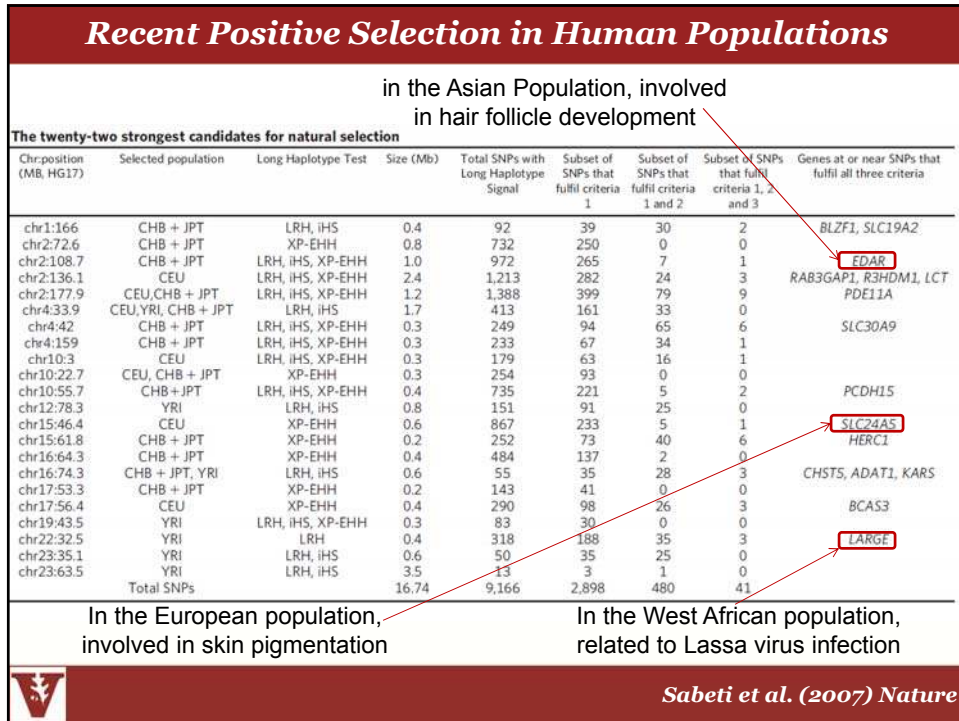


<http://www.bio.miami.edu/~cmallery/150/gene/c7.19.14.human.genome.jpg>

Human Genes Mirror Geography

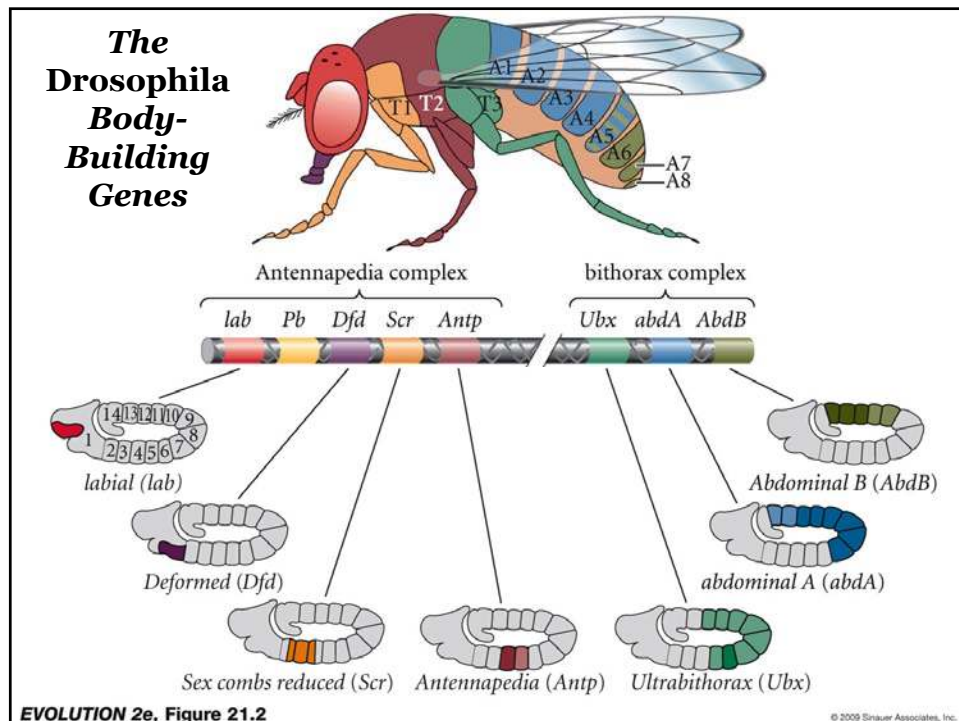


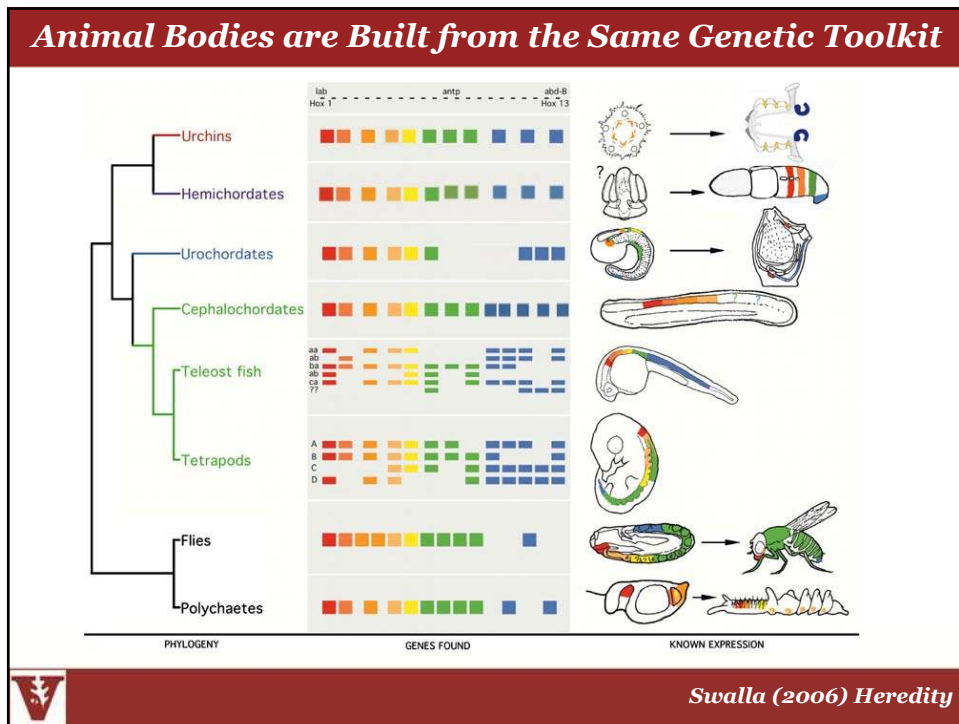
Novembre et al. (2008) Nature



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 1-CKN1
+	+	+	Coffin-Lowry-RPS6KA3
+	+	+	Diastraphic 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
+	+	+	Septo-optic 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

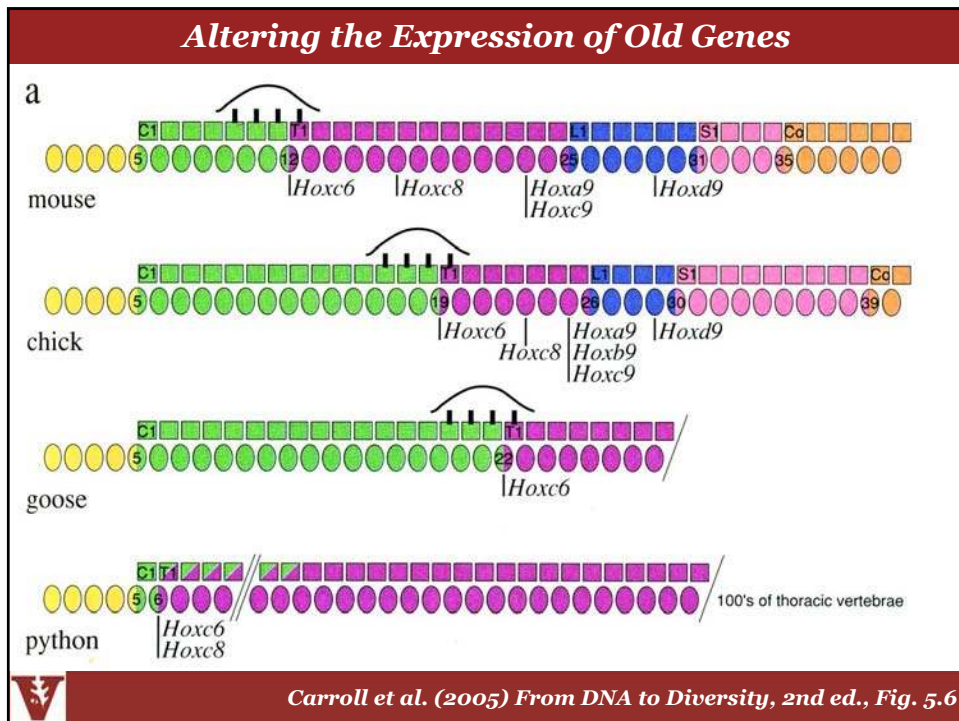
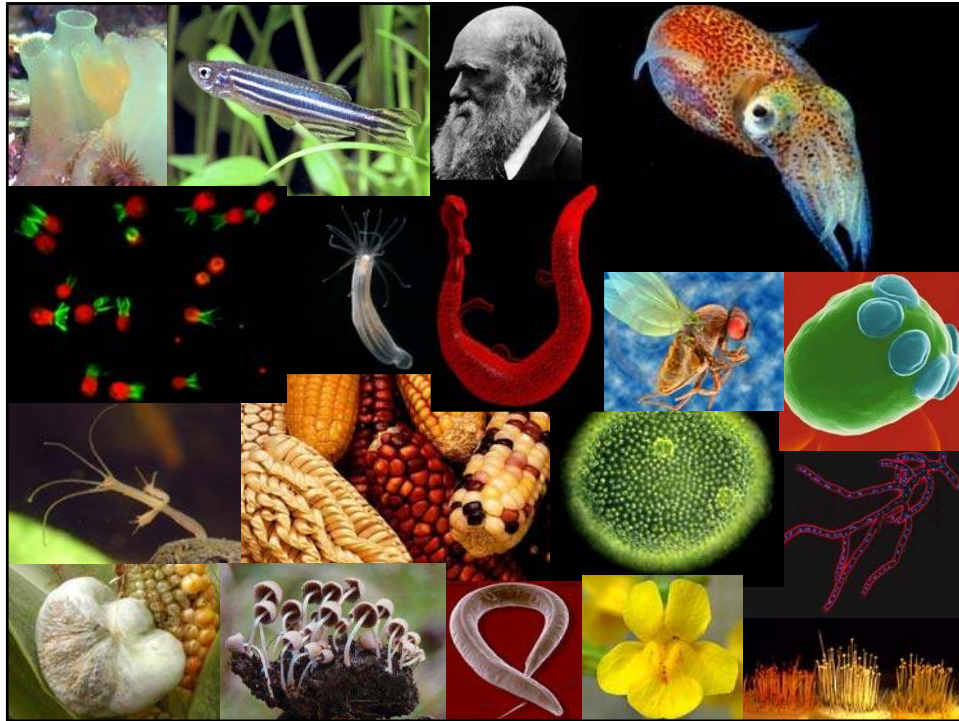


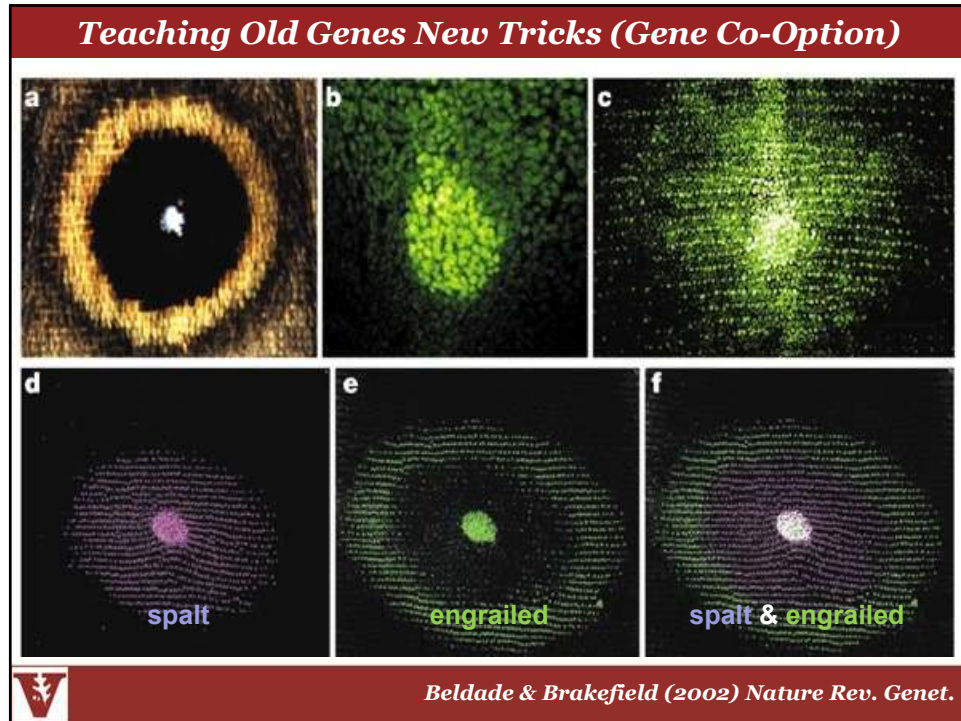


“...the search for homologous genes is quite futile except in very close relatives. If there is only one efficient solution for a certain functional demand, very different gene complexes will come up with the same solution, no matter how different the pathway by which it is achieved. The saying ‘Many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

Ernst Mayr, 1963





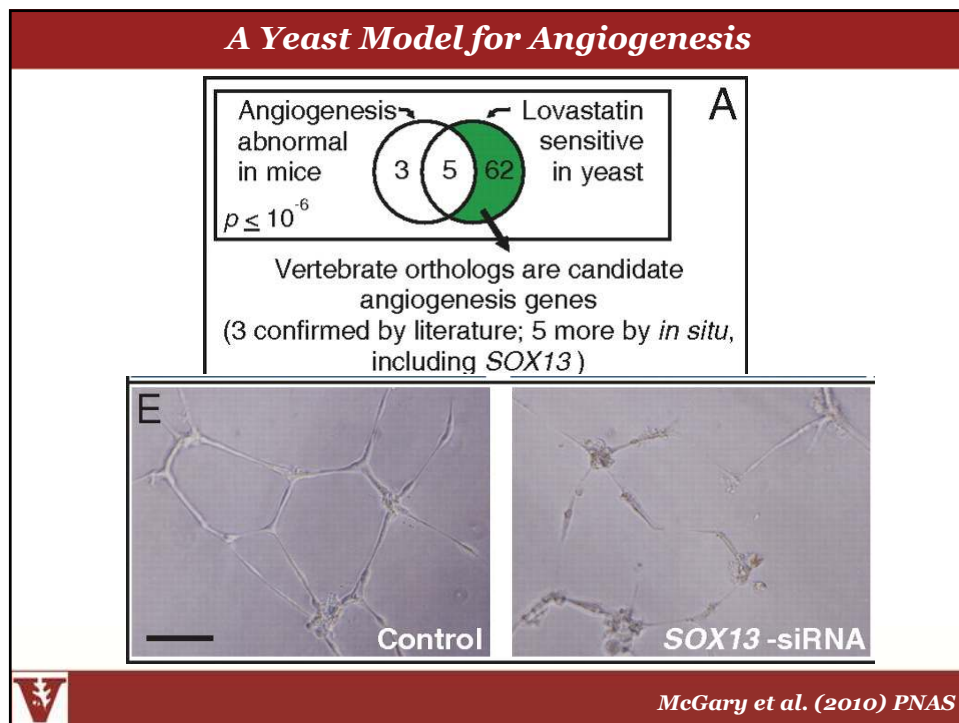
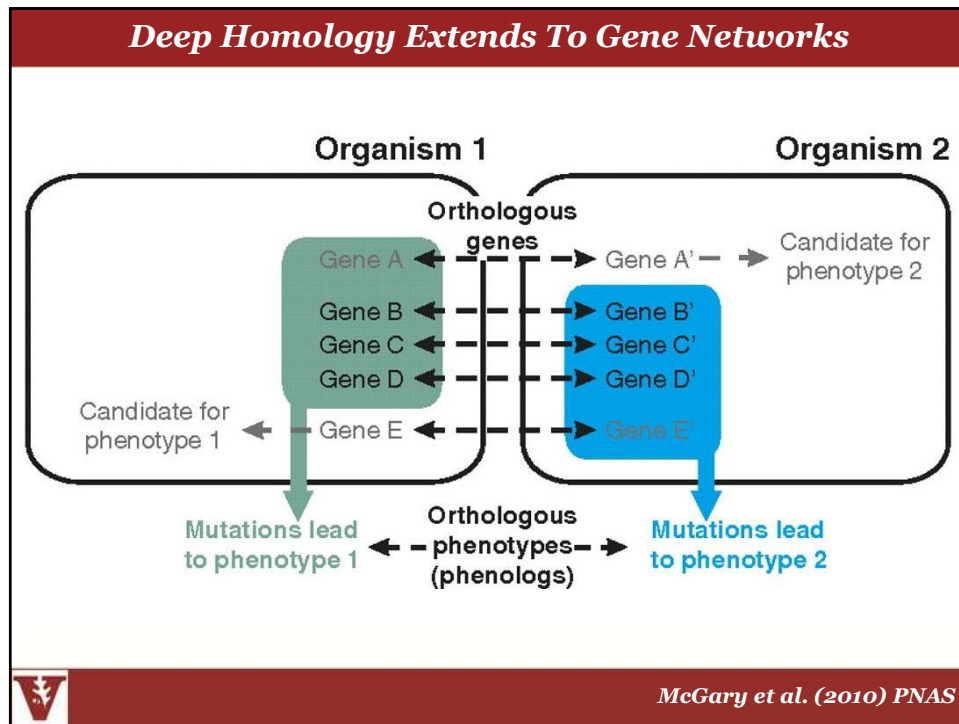


“Our method suggests a yeast model for angiogenesis defects, a worm model for breast cancer, mouse models of autism, and a plant model for the neural crest defects [...], among others”

McGary et al. (2010) PNAS

“Modelling neurodegeneration in *Saccharomyces cerevisiae*”

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



Genomics: "Big Science" Driven by a Few Centers



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

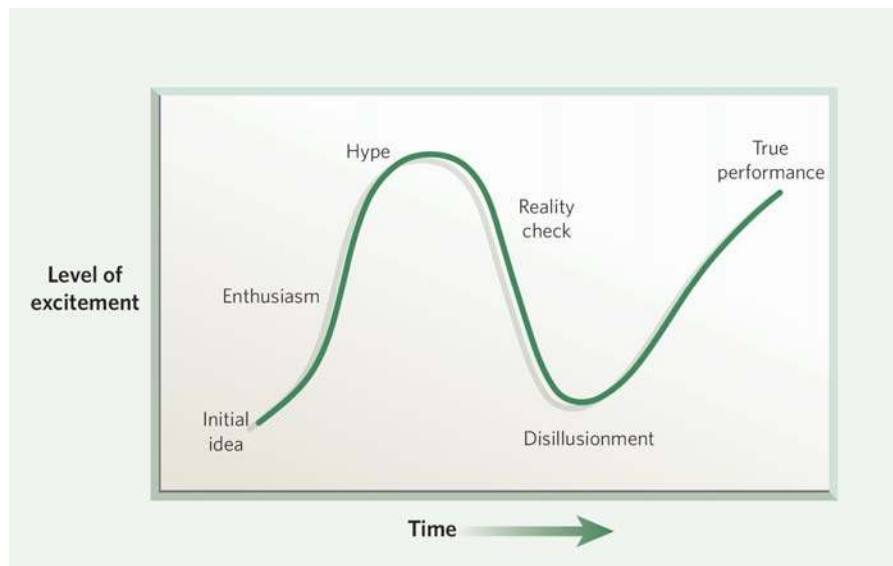
SOLID ABI

2 x 75 bp ~14 days 300 Gb



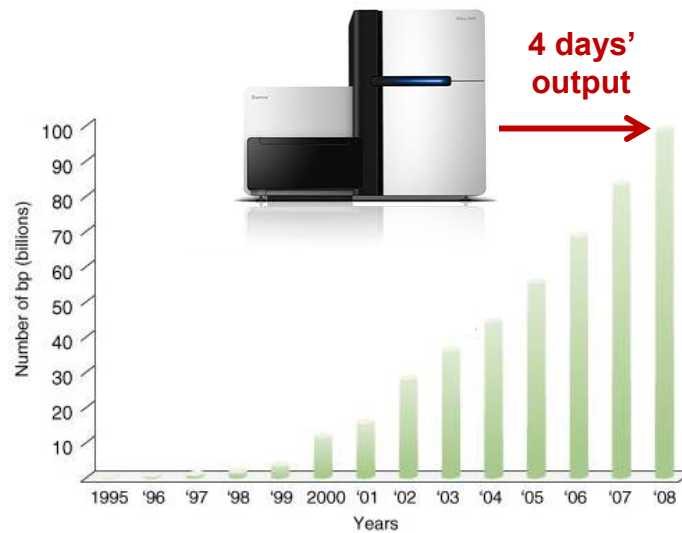
For the latest specs consult: <http://ngsbuzz.blogspot.com/>

“New Technology Level of Enthusiasm” Curve

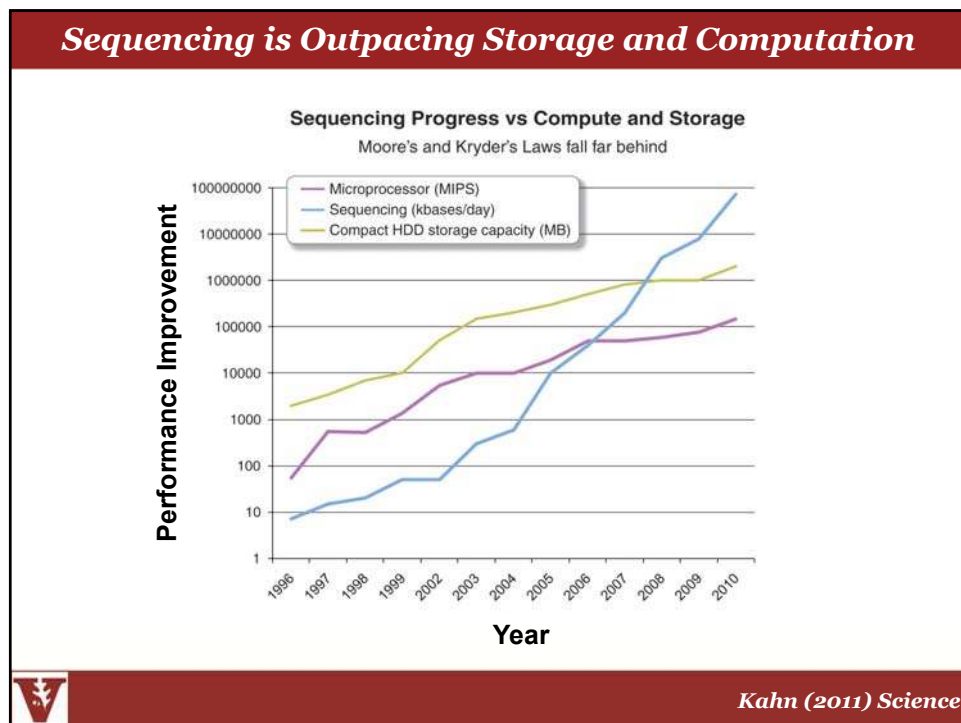
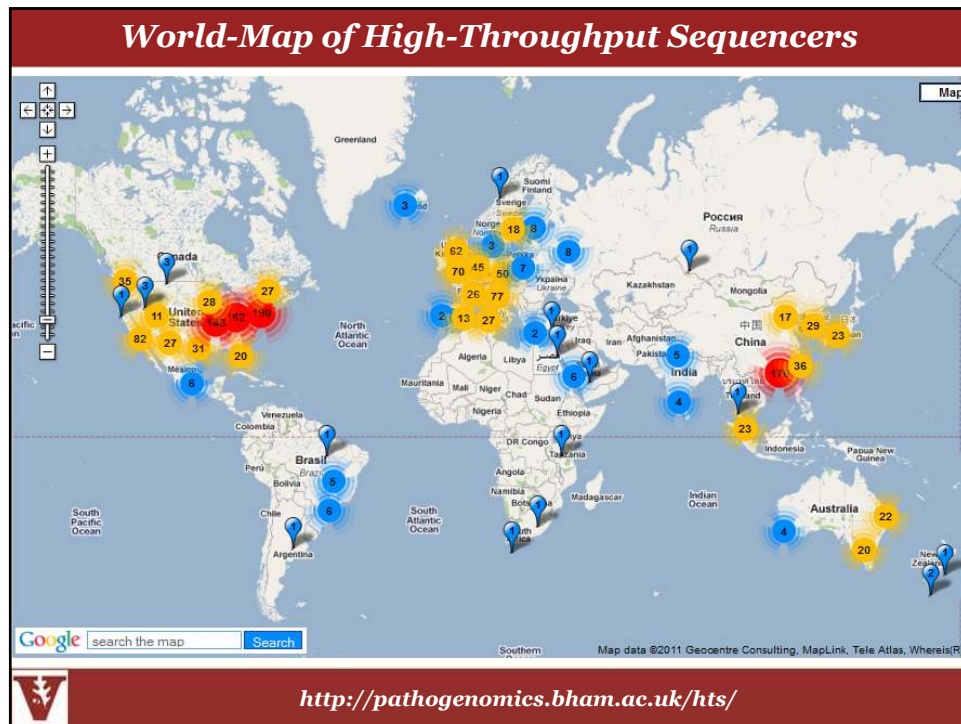


Collins (2006) Nature

Why Is Next-Generation DNA Sequencing So Exciting?



Gilad et al. (2009) Trends Genet.



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

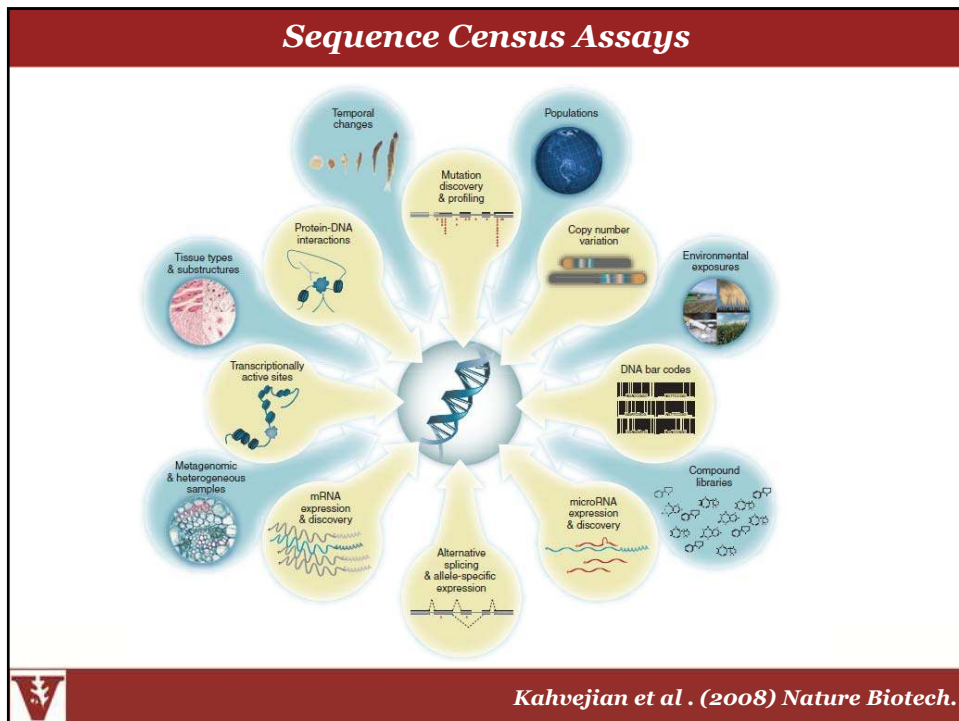
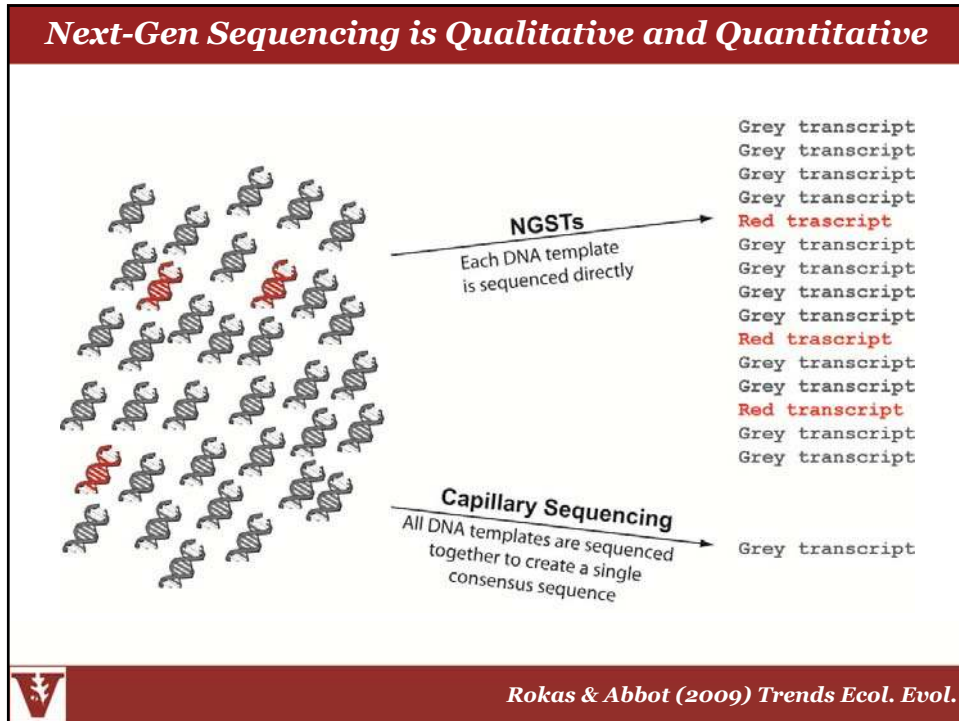
Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA.
E-mail: petsko@brandeis.edu



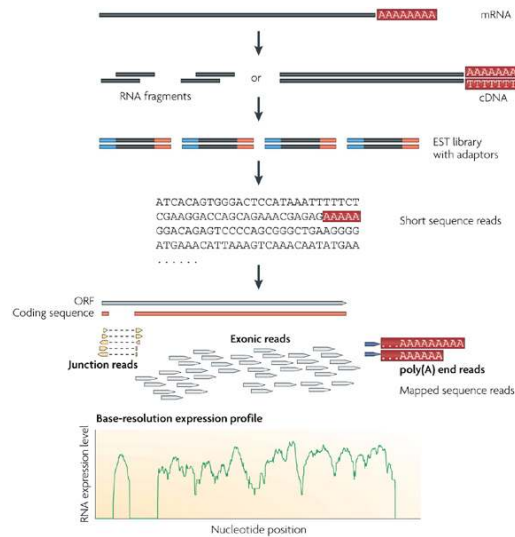
Lecture Outline

- ❖ Introduction to evolutionary genomics
- ❖ **Benchmarking *de novo* transcriptome sequencing for functional and evolutionary genomics**
- ❖ Phylogenomics
- Coffee Break-----
- ❖ Comparative, Population & Functional Genomics





A Typical RNA-Seq Experiment



Nature Reviews | Genetics

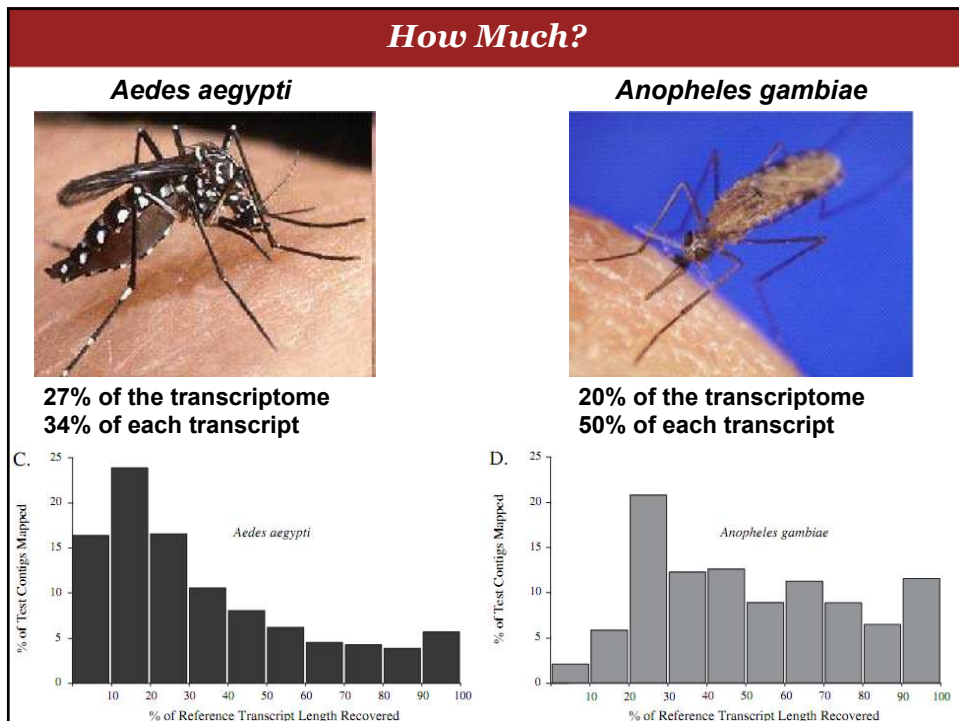
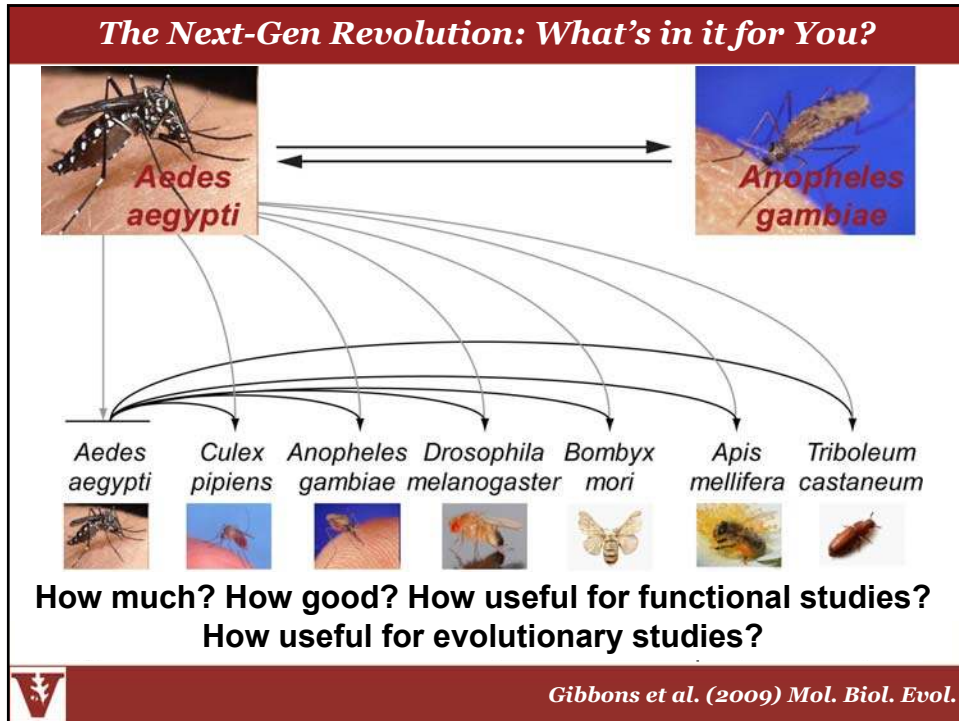


Wang et al. (2009) Nature Rev. Genet.

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



John Gibbons



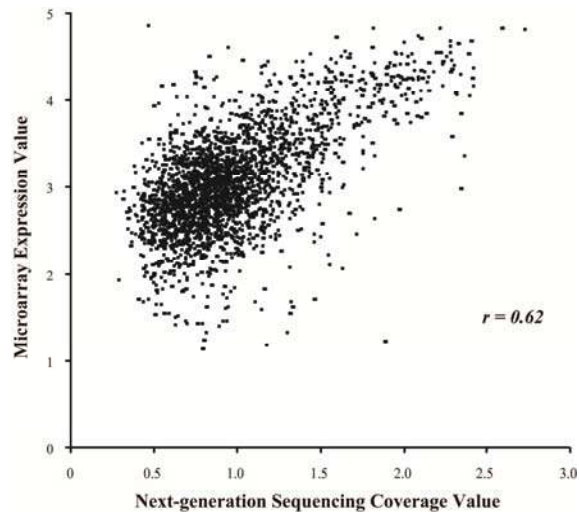
RNA-Seq Data is Qualitatively Accurate

Substitutions	Reference	CAG	CCA	ATG	AGG	TCT	CAG	TCG	3×10^{-3}
	Test	CAG	CCA	ATC	AGG	TCA	CAG	TCG	
Indels	Reference	CAG	CCA	ATG	---	---	CAG	TCG	4×10^{-5}
	Test	CAG	CCA	ATG	AGG	TCA	CAG	TCG	
Assembly Breaks	Reference	CAG	CCA	ATG	AGG	TCT	CAG	TCG	2×10^{-4}
	Test	ATG	GCG	TAC	AGG	TCT	CAG	TCG	

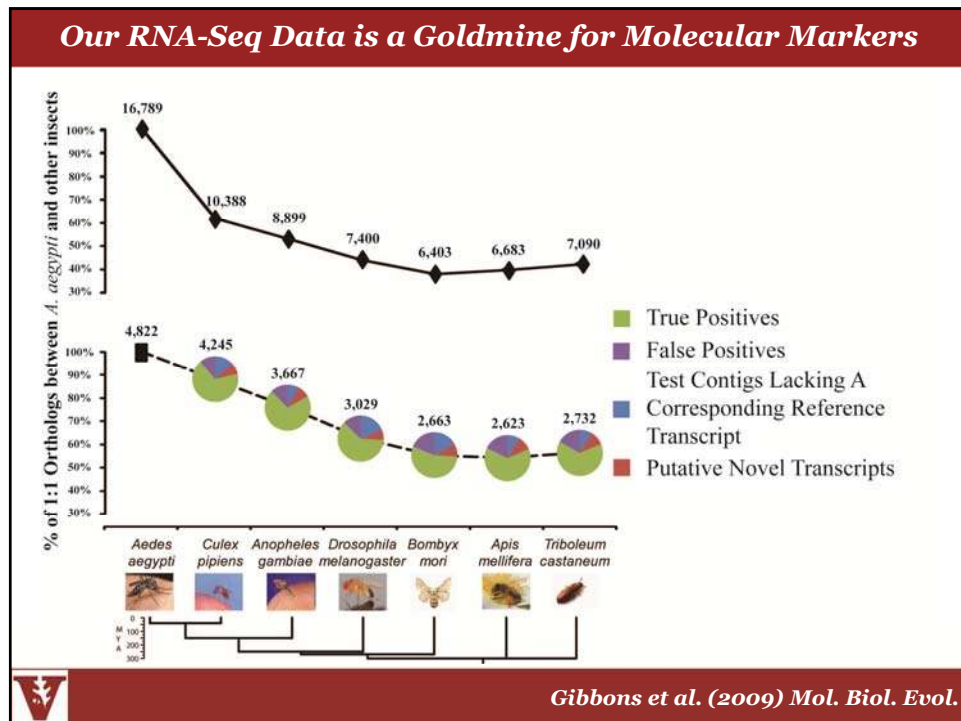
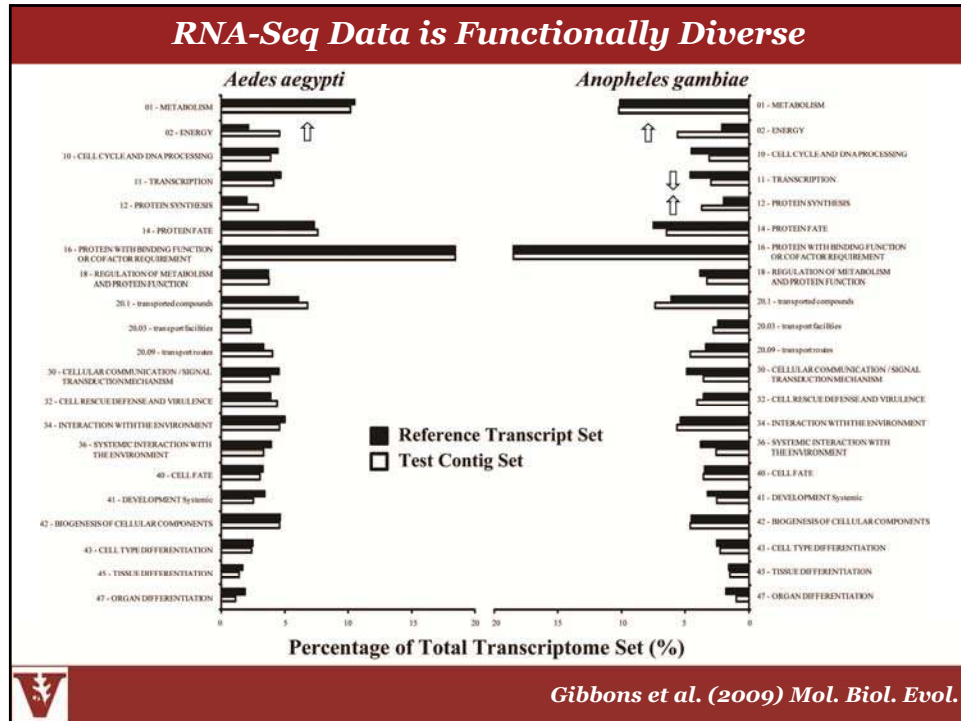


Gibbons et al. (2009) *Mol. Biol. Evol.*

RNA-Seq Data is Quantitatively Accurate



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



What Can We Assemble *de novo* with Current Data?

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

	<i>A. gambiae</i>	<i>A. albimanus</i>
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

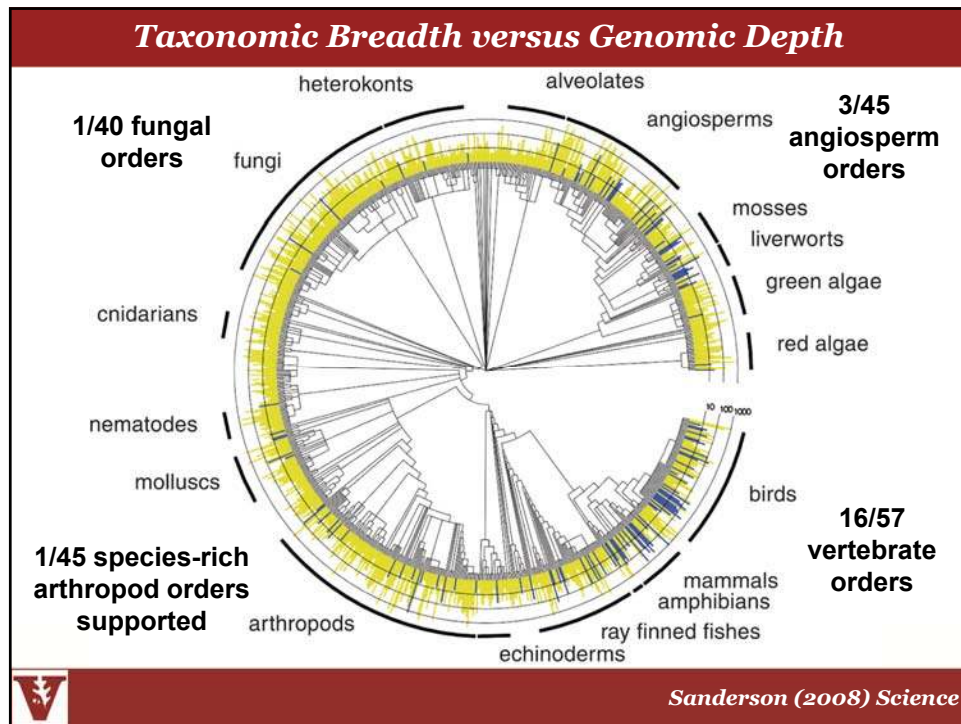
- ❖ Introduction to evolutionary genomics
- ❖ Benchmarking *de novo* transcriptome sequencing for functional and evolutionary genomics

❖ Phylogenomics

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

- ❖ Comparative, Population & Functional Genomics





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

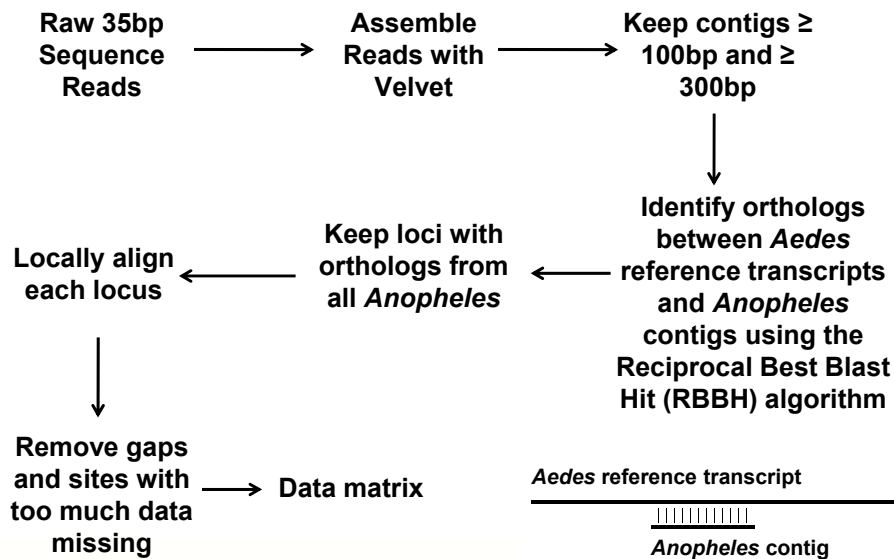
Can we Use RNA-Seq to Increase Genomic Depth?

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



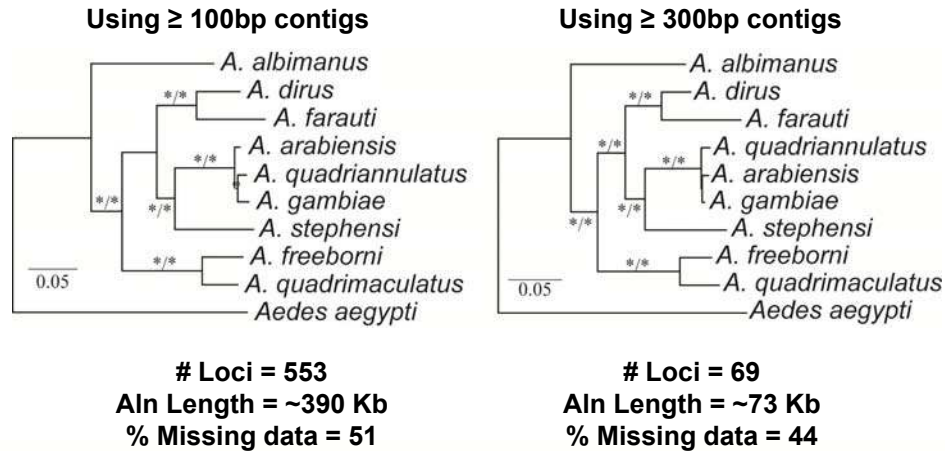
Hittinger et al. (2010) PNAS

Data Matrix Construction: The “Singlecontig” Strategy



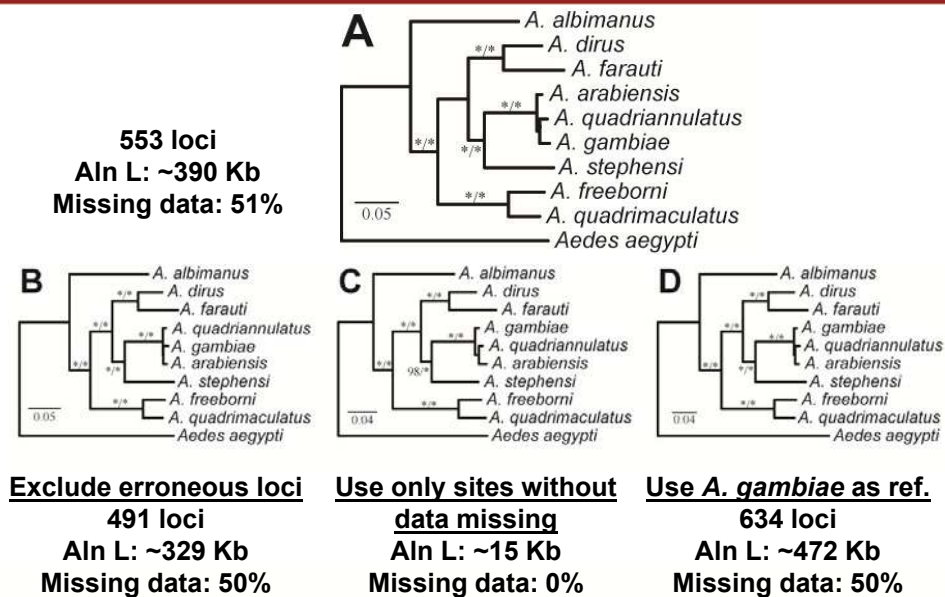
Hittinger et al. (2010) PNAS

Robust Phylogenetic Inference from RNA-Seq Data

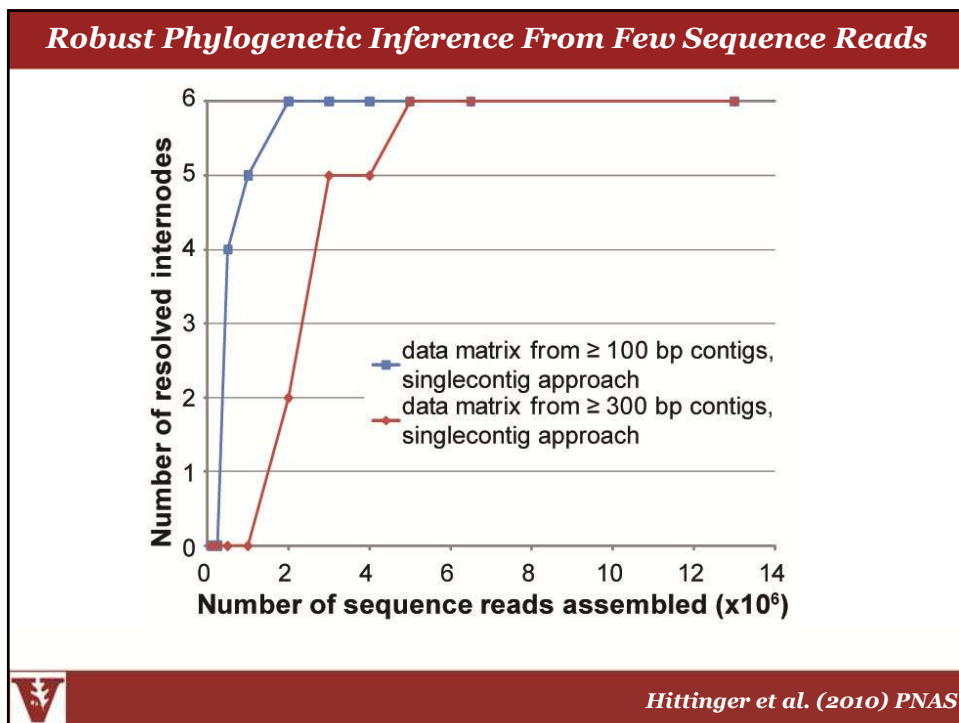
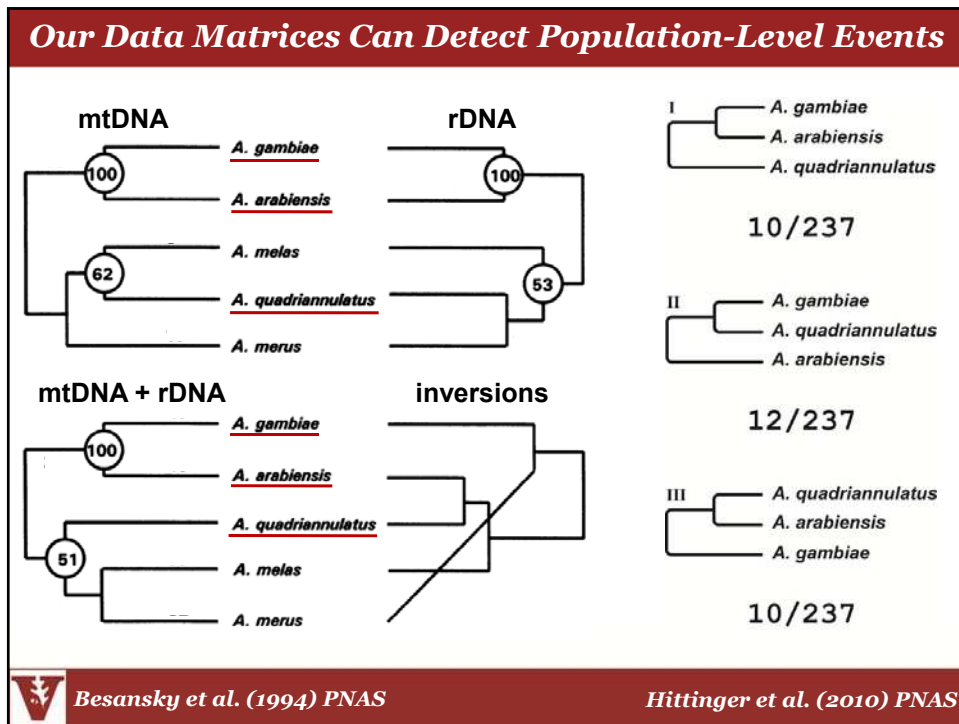


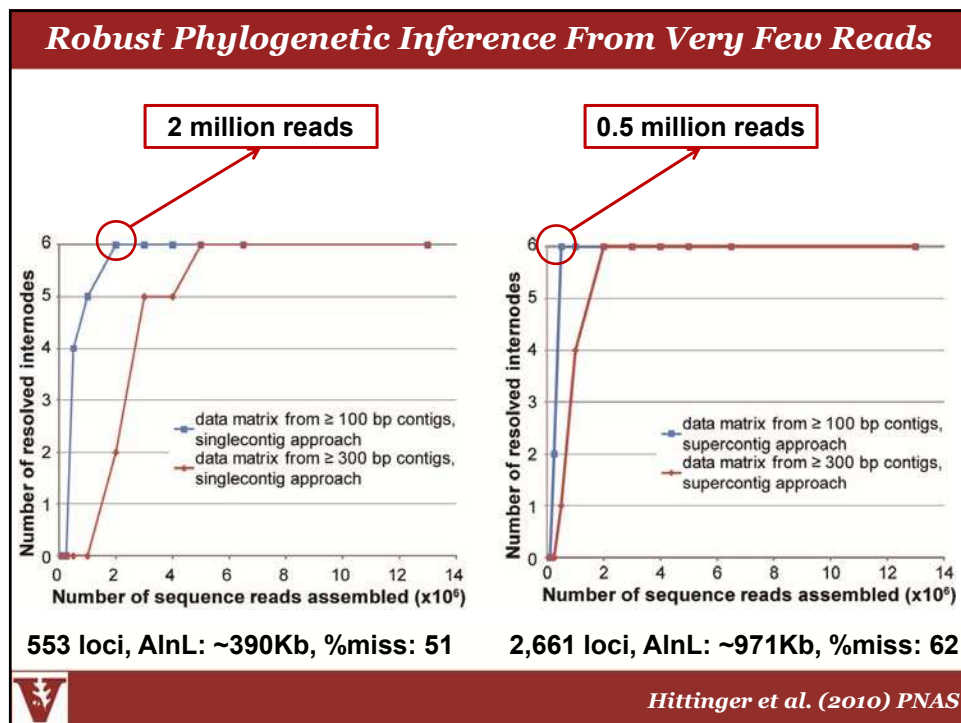
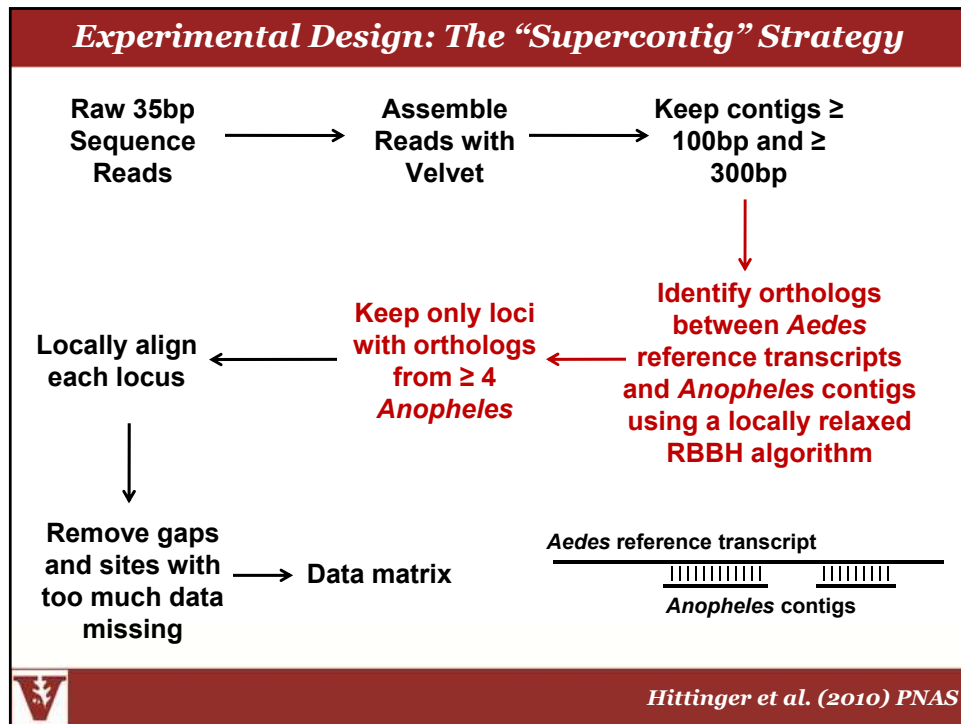
Hittinger et al. (2010) PNAS

Accurate Phylogenetic Inference From our Data



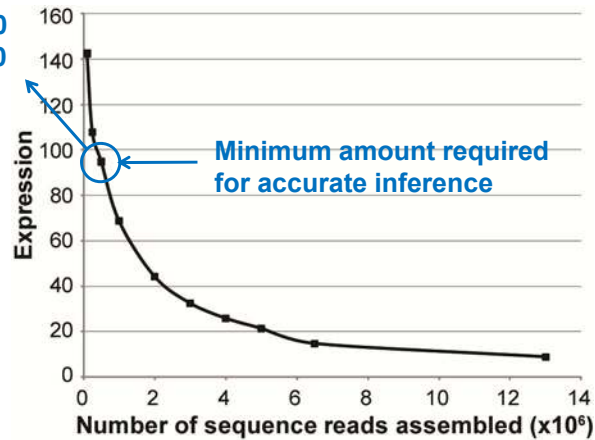
Hittinger et al. (2010) PNAS





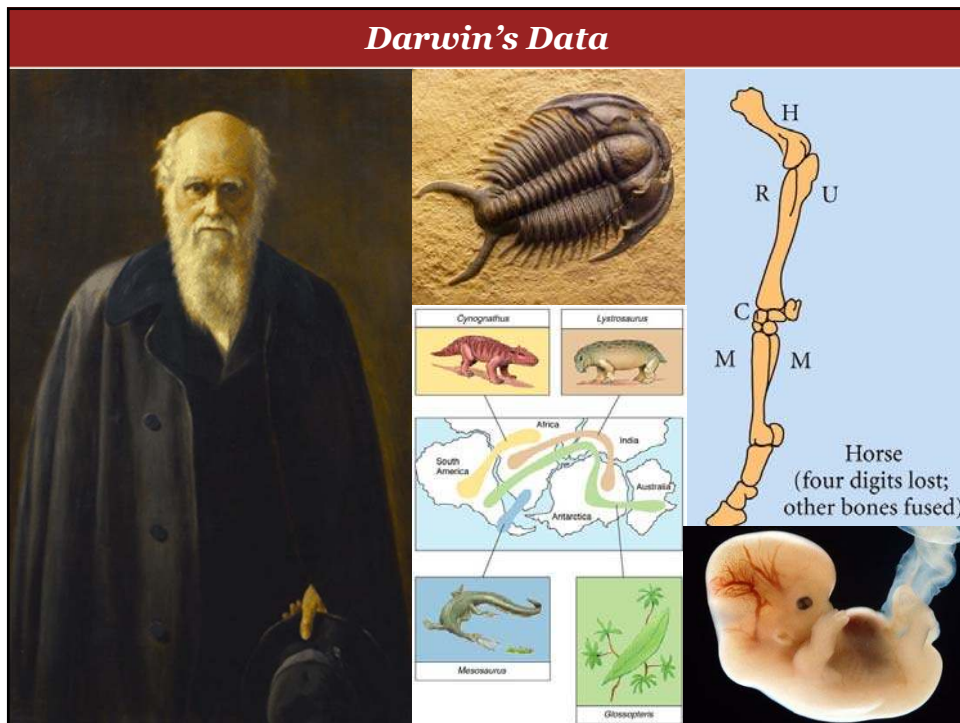
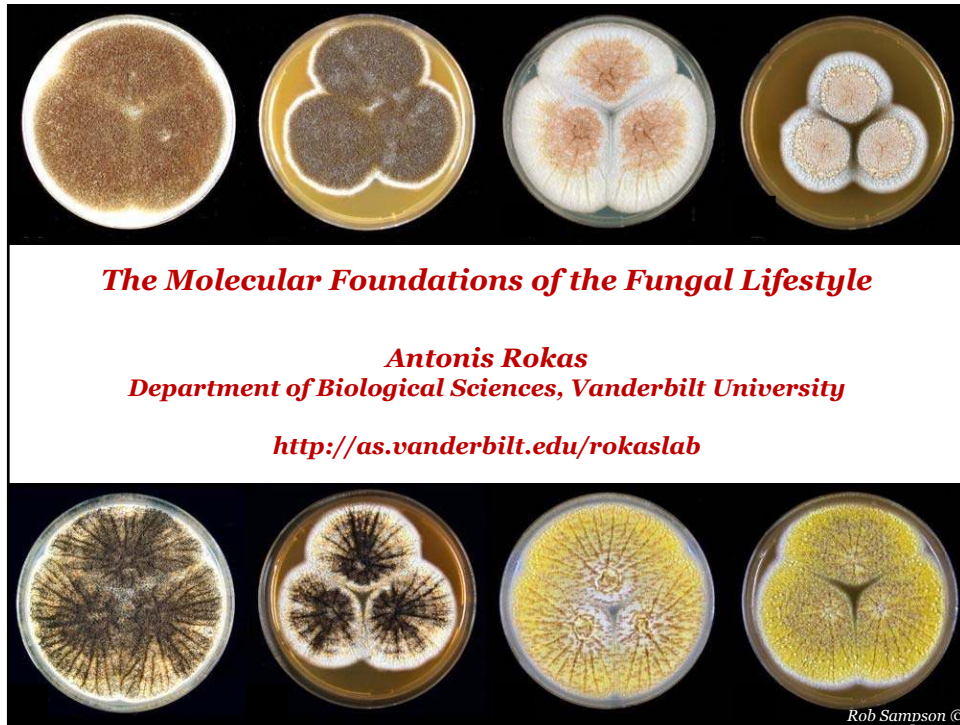
Our Sequences are from Highly-Expressed Transcripts

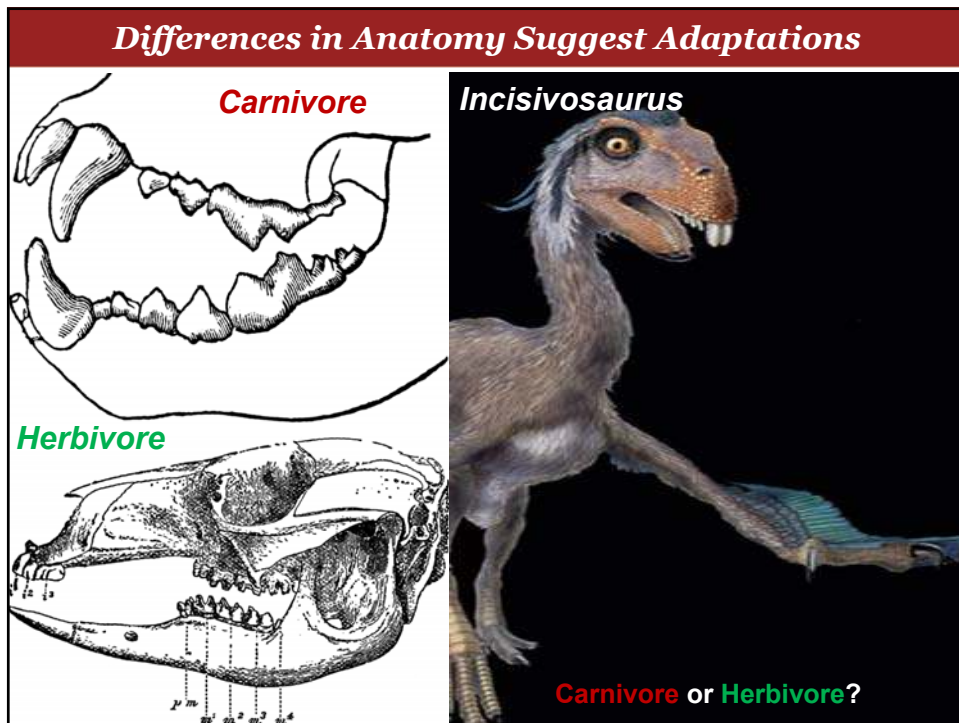
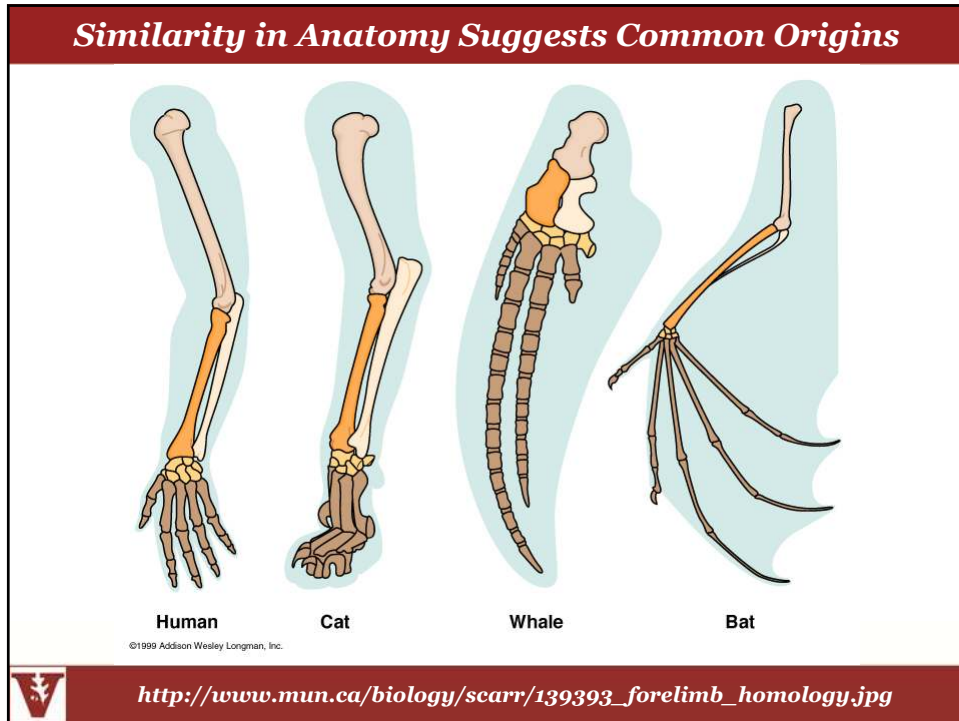
2008 cost: ~\$50
2011 cost: <\$10



Hittinger et al. (2010) PNAS

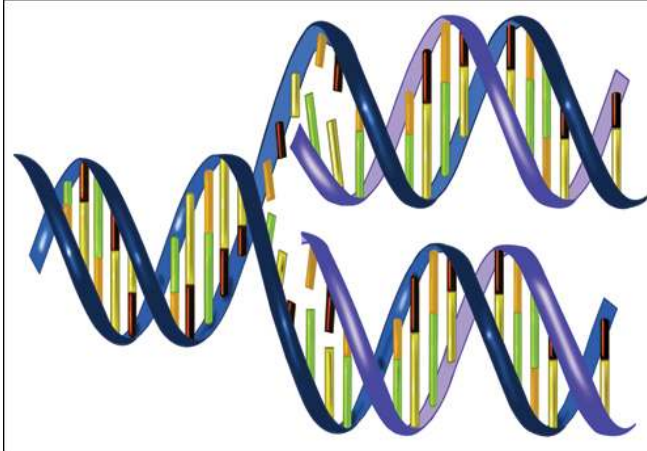
Coffee Break



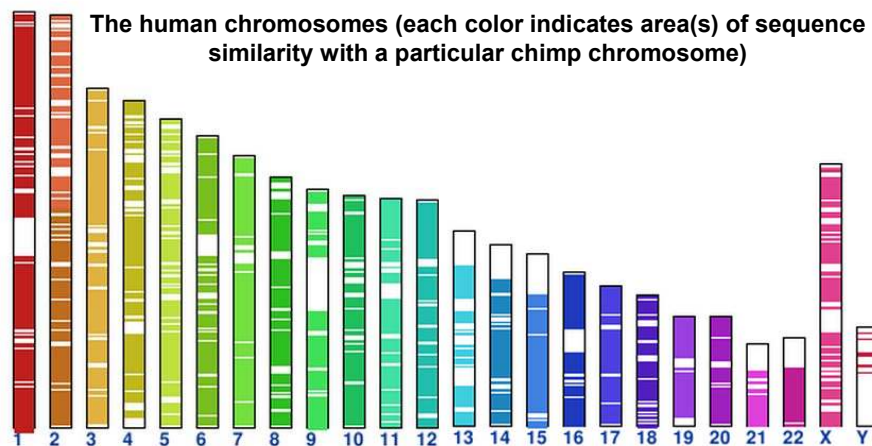


The DNA Record

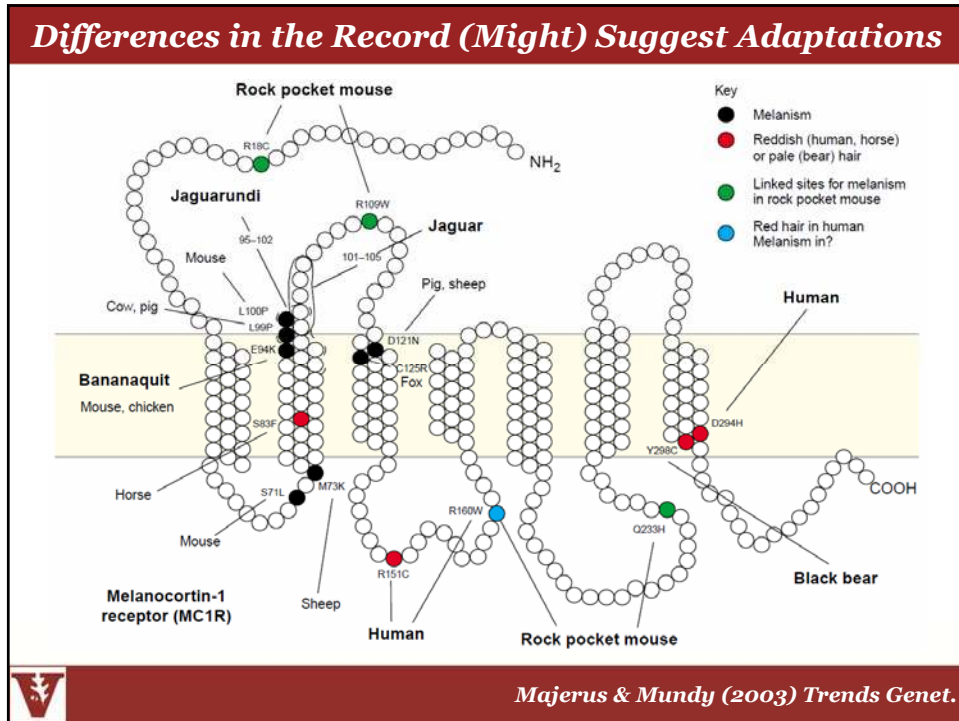
The DNA record contains important clues about organisms' biological past, and their history of change and adaptation



Similarity in the DNA Record Suggests Common Origins



<http://cinteny.cchmc.org/>



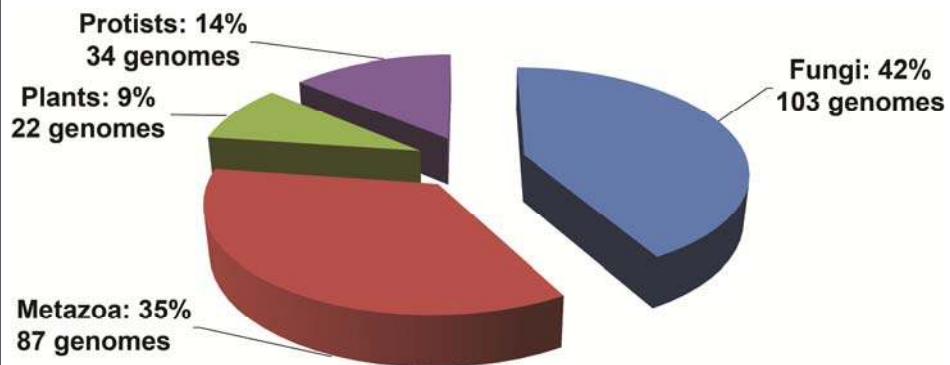
The Rokas Lab

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

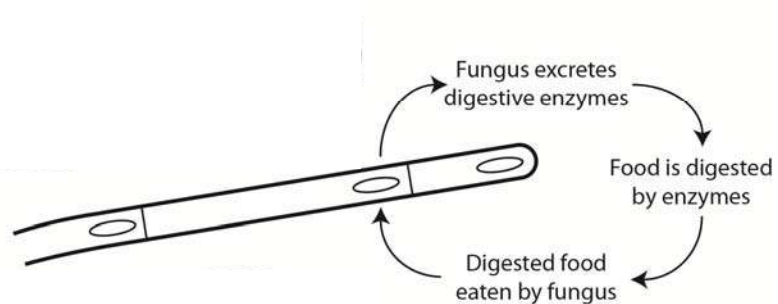
<http://www.jgi.doe.gov/sequencing/cspseqplans2012.html>



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

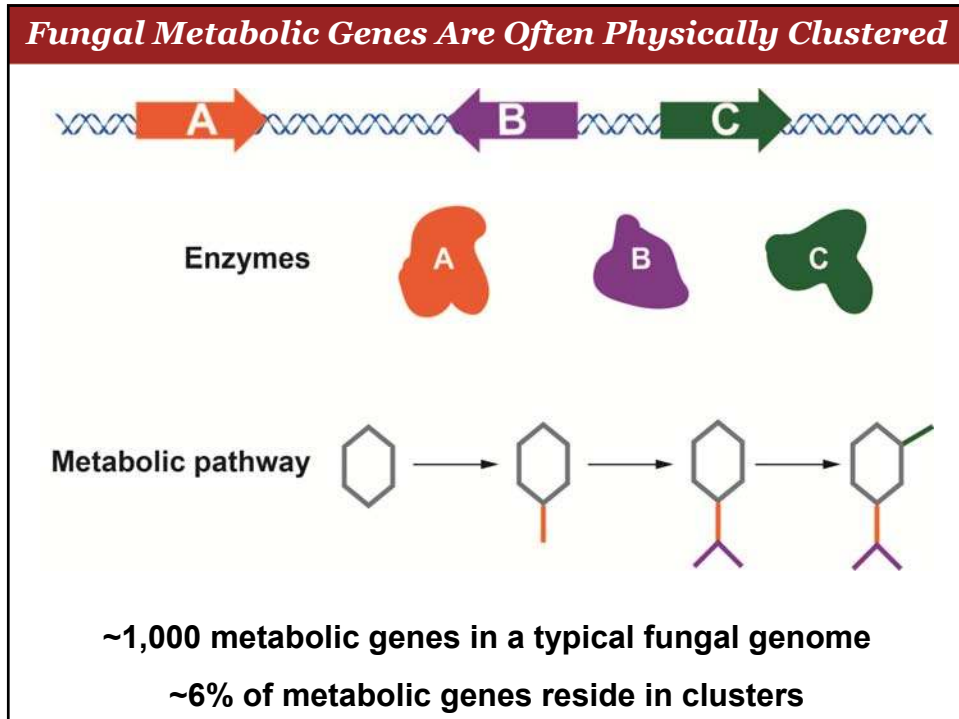


What Fungi Cannot Eat!



Diner fries

McDonald's fries



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



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



Saccharomyces (Ascomycota: Saccharomycetes)



Aspergillus (Ascomycota: Eurotiomycetes)



Cryptococcus (Basidiomycota: Tremellomycetes)



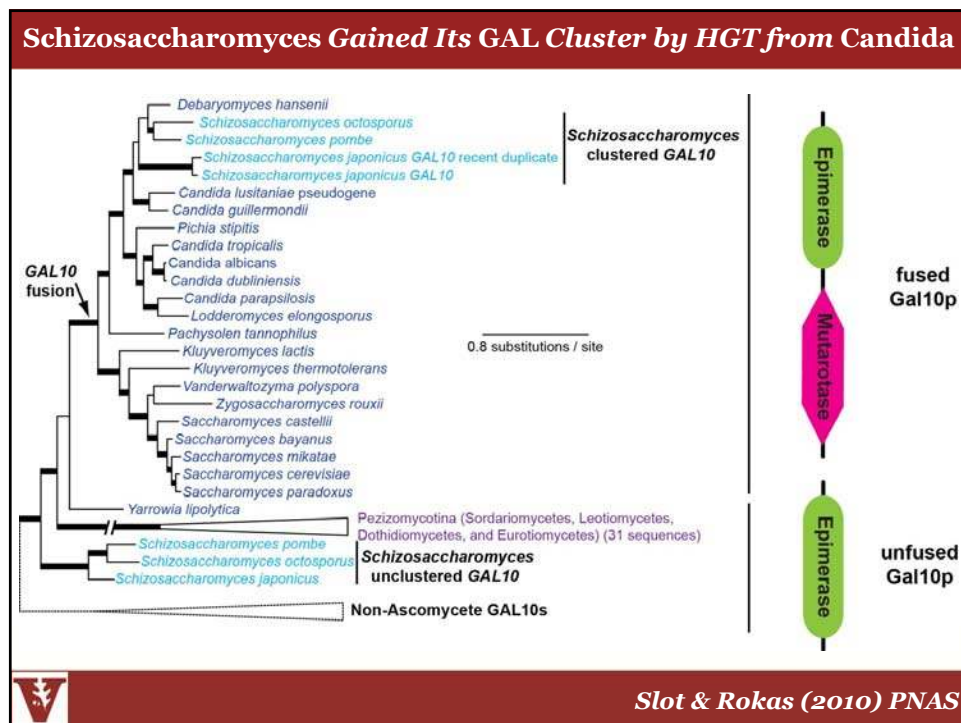
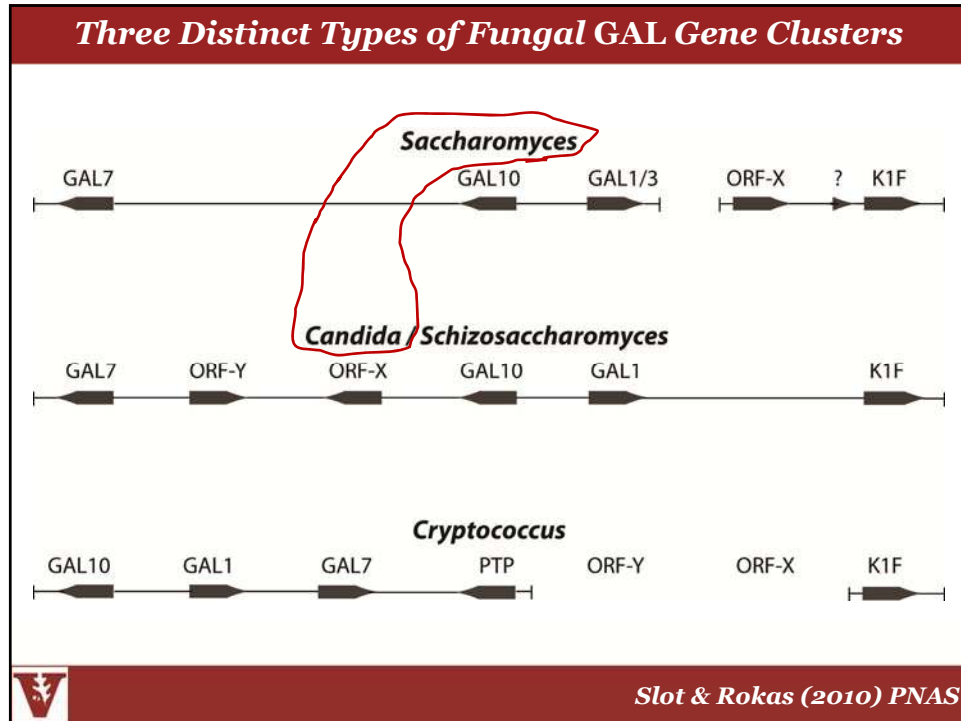
Ustilago (Basidiomycota: Ustilaginomycetes)

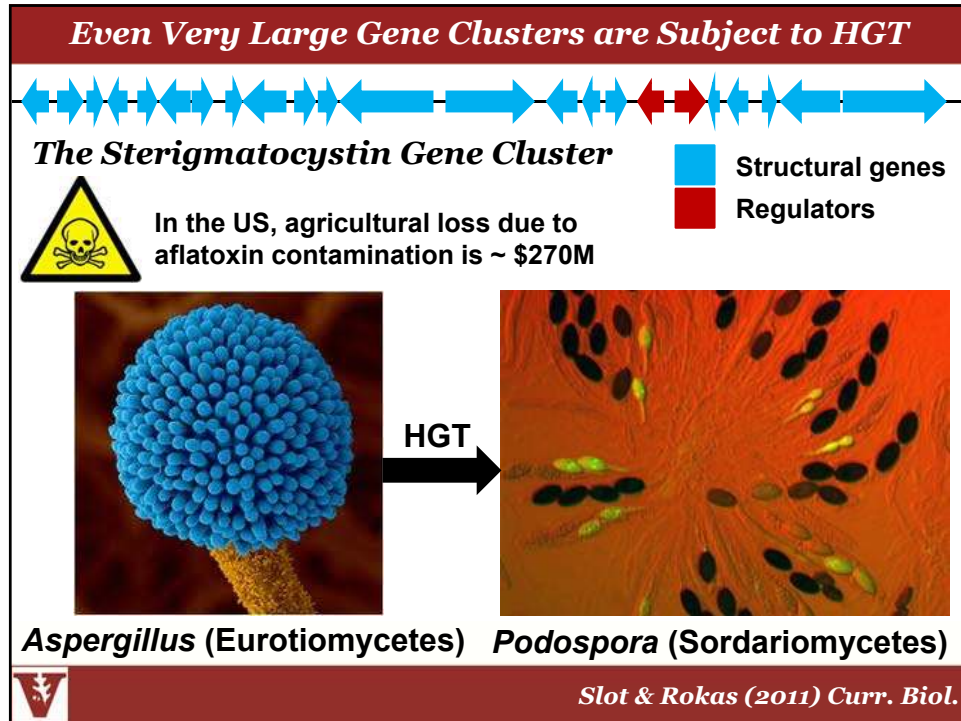


Batrachochytrium (Chytridiomycota: Chytridiomycetes)



Slot & Rokas (2010) PNAS





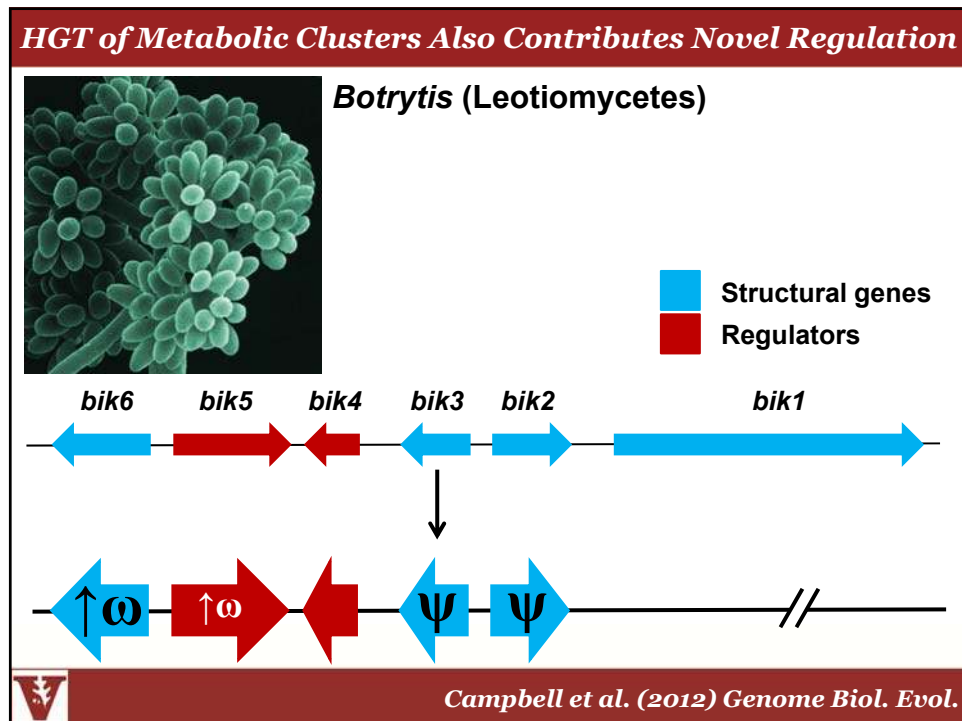
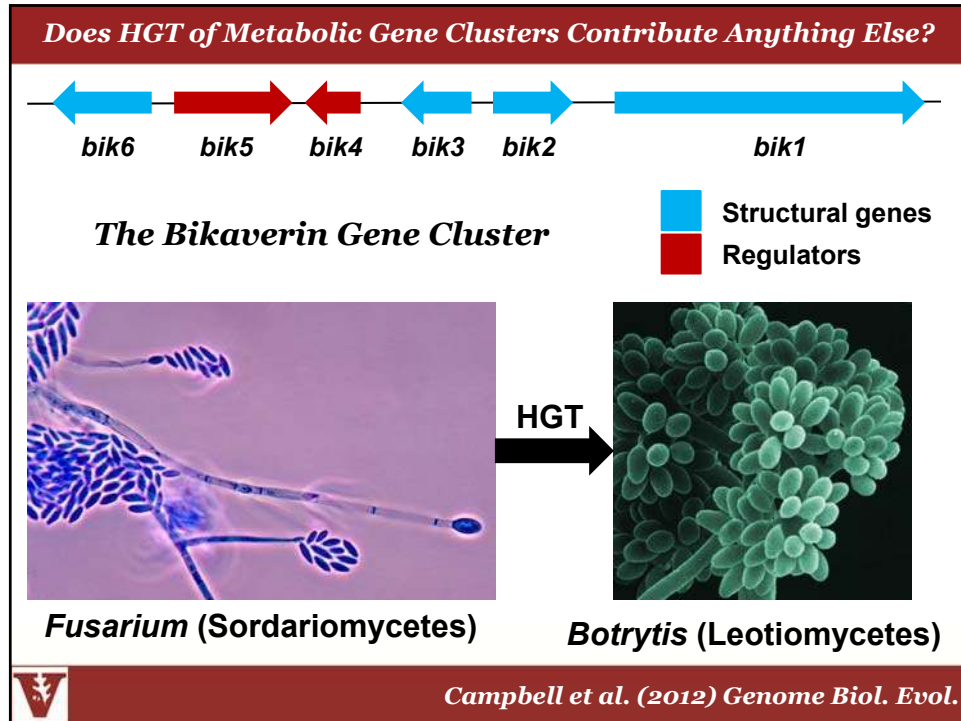
Parasitol Res
DOI 10.1007/s00436-010-2098-1

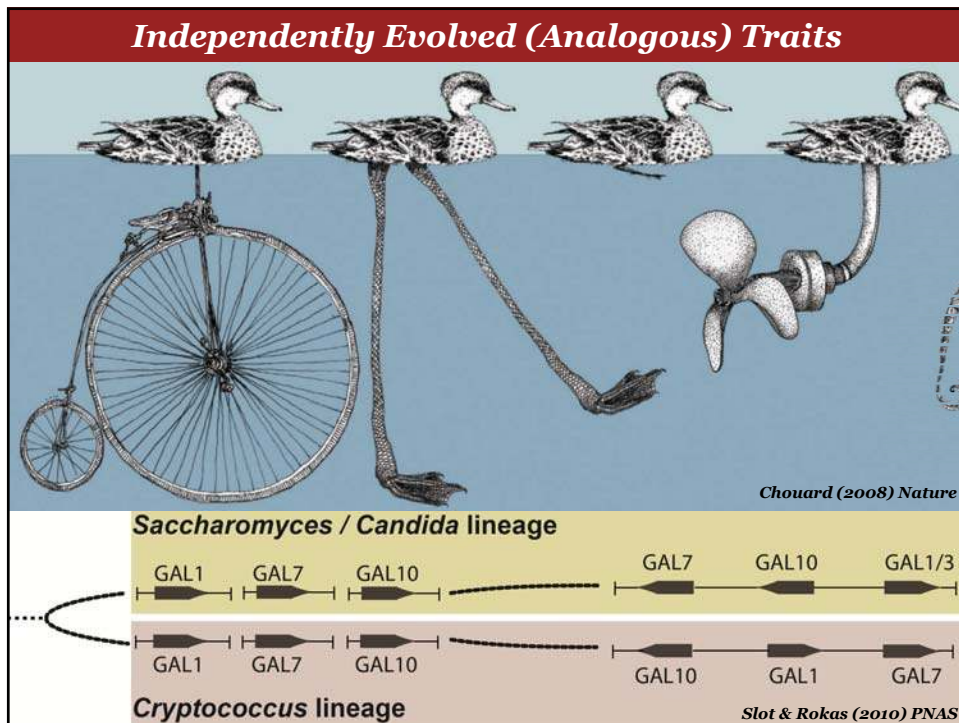
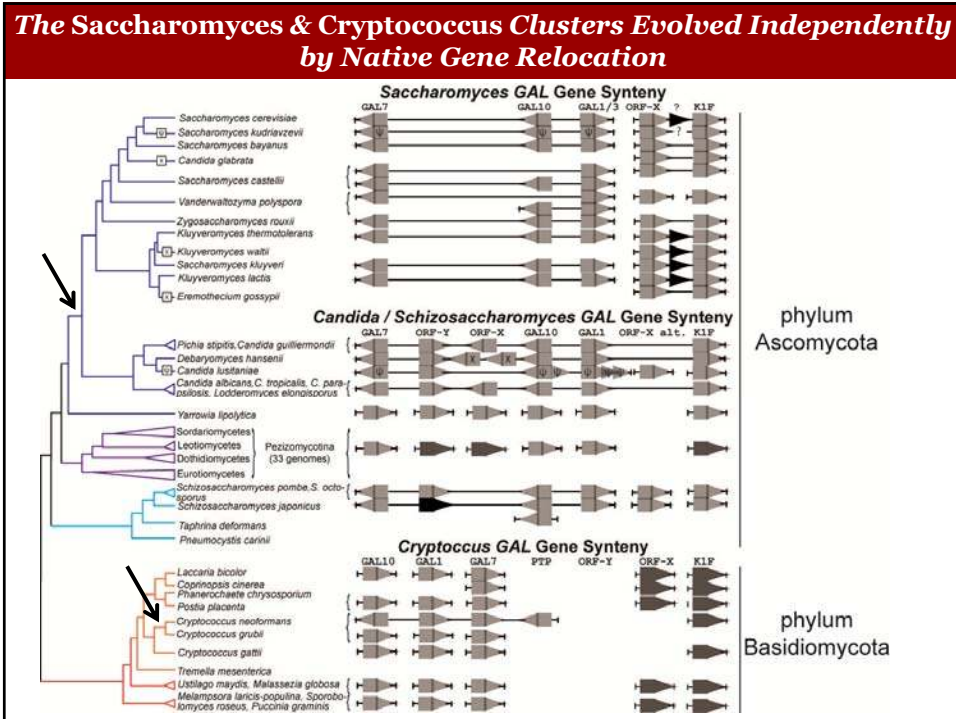
ORIGINAL PAPER

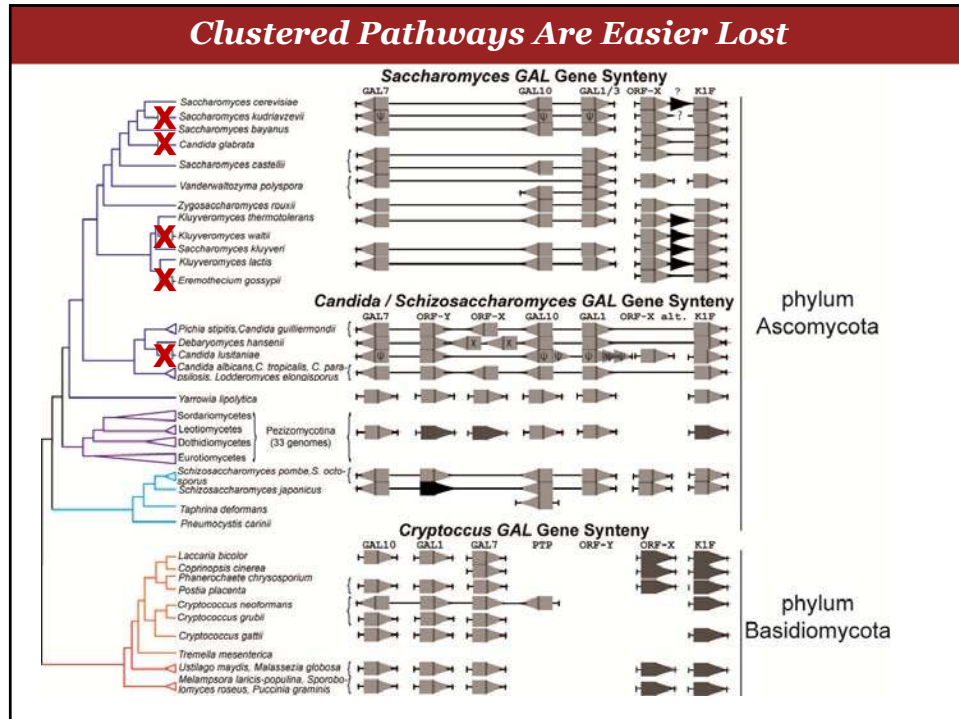
Larvicidal activity of metabolites from the endophytic *Podospora* sp. against the malaria vector *Anopheles gambiae*

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 xanthenes, sterigmatocystin (1) and secosterigmatocystin (2), and an anthraquinone derivative (3) 13-hydroxyversicolorin B were isolated after fermentation on M₂ 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 LC₅₀ and LC₉₀ 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







Questions

❖ How are fungal metabolic clusters formed?

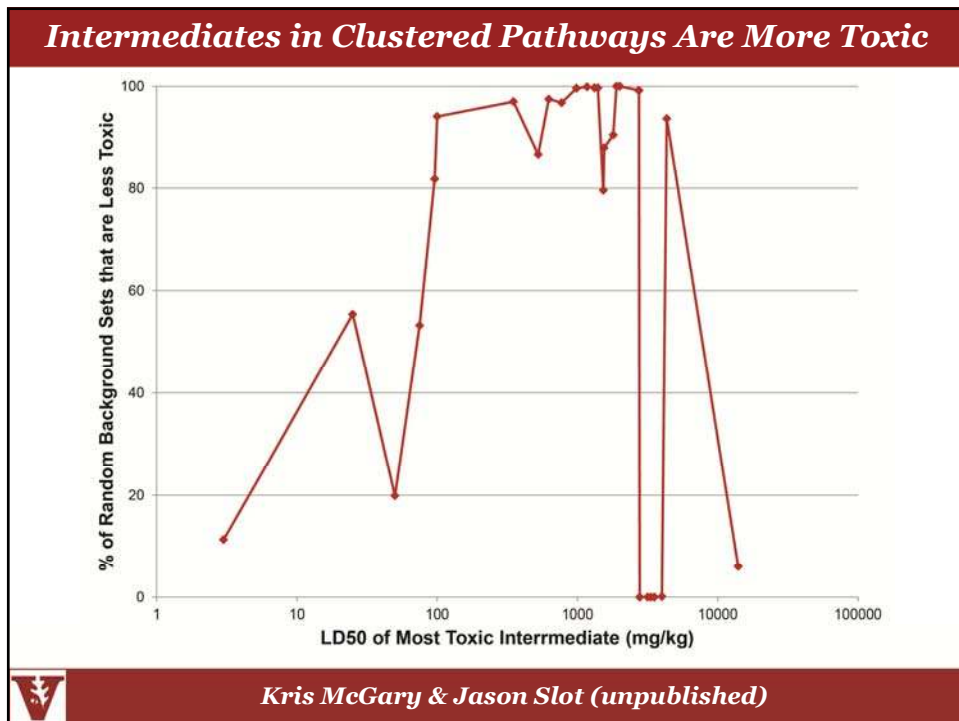
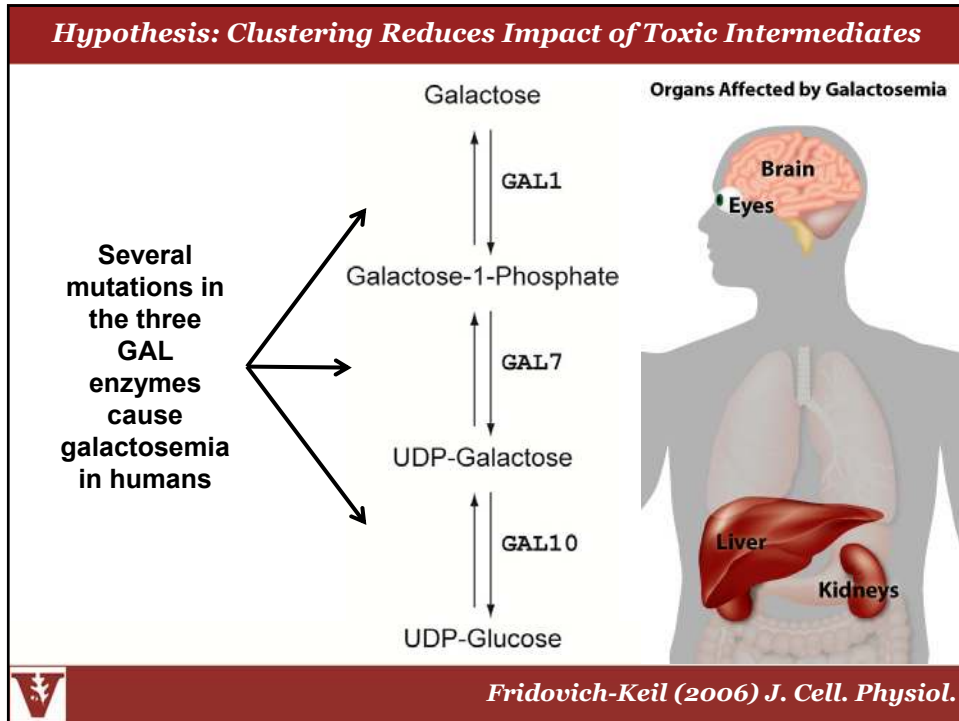
Native gene relocation
Horizontal gene transfer

❖ 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





Mechanisms That Could Drive Clustering

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



Evidence for the Coordinated Expression Hypothesis

← GAL10

Scer TTATATTGAATTTTCAAAATTTCTACTTTTTTTTGGATGGACGCAAGAAGTTTAATATCATATTACATGGCATTACCACCATATACA
 Spar CTATGTTGATCTTTTTCAGAAATTTT-CACTATATTAGATGGGTGCAAGAAGTGTATTATATATTACATCGCTTTCTCATCATACACA
 Smik GTATATTGAATTTTTCAGTTTTTTTCTACTATCTTCAAGTTATGTAAAAA-TGTCAAGATAATATTACATTTCGTACTATCATACACA
 Sbay TTTTTTGAATTTCTTAGTTTTCTTTTAACTTCAAAATTTATAAAGAAAGTGTAGTCACATCATGCTATCT-GTCACTATCATATATA

TATA

Scer TATCCATATCTAATCTTACTTTATATGTTGT-GGAAT-GTAAAGAGCCCCATATCTTAGCTTAAAAAACC--TTCTCTTGGAACTTTCAGTATATCG
 Spar TATCCATATCTAGTCTTACTTTATATGTTGT-GAGAGT-GTTGATAACCCAGTATCTTAACCAAGAAAGCC--TT-TCTATGAACTTGAAGCTG-TACG
 Smik TACCGATGCTAGTCTTACTTTATATGTTTAC-GGAATTTGTTGTAATCCAGTCTCCAGATCAAAAAGGT--CTTCTATGAGAGCTTTCG-CTA-TATG
 Sbay TAGATATTCTGATCTTTCTTATATATTATAGAGAGATGCCAATAAACGTGCTACCTCGAACAAAAAGGGGATTTCTGTAGGGCTTCCCTATTGTTG

Gal4

Scer CTTAATGCTCATTGC-----TATATTGAAGTTCGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAAGACTCTCCTCCGTCGCTCCTCGTCT
 Spar CTAATGCTCATTGC-----AATATTGAAGTTCGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAAATATCCCTCCGTCGCTCCTCGTCT
 Smik TTTAGCTCTCAAG-----AATTTGAAGTTCGGATCAGAAGCCGCCGAGCGGACGAGATTCCTCCGACGGAAATATCCCTCCGTCGCTCCTCGTCT
 Sbay TCTTATTGTCATTTACTTCGAATGTTGAAATTCGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGCGGAAATCTGCTCCGTCGGAAGTCGTCT

Gal4

Scer TCACCGG-TGCGGTTCTGAAAGCAGATGTGCTCCGCCGCACTGCTCCGAACAATAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
 Spar TCGTCGGGTTGTGTCCTTAA-CATCGATGTACTCCGCCGCCCTGCTCCGAACAATAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
 Smik ACGTTGG-TGCGGTTCTGAA-CATAGTGTGCTCCGCCGCACTGCTCCGAACAATAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
 Sbay GTG-CGGATCAGTCCCTGAT-TACTGAAGCGTTCGCCGCCGCACTGCTCCGAACAATAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA

Mig1

Scer GAGGA-AAAATTGGCAGTAA-----CTGGCCCCCAAACTTTC-CAAAATTAACGAATCAAAATTAACAACCATA-GGATGATAATGCGA-----TTAG--T
 Spar AGGAACAAAATAAGCAGCC-----ACTGACCCCATATACCTTTC-CAAACTATTGAATCAAAATGGCCAGCATA-TGTAATAGTACAG-----TTAG--G
 Smik CAACGCAAAATAAAGTCC-----CCCGCCCCACATACCTTTC-CAAACTGATGCGTAAACTGGCTAGCATA-GAATTTTGGTAGCAA-AATATTAG--G
 Sbay GAACGTGAATGCAATTCCTTGCCCTT-CCCAATATACCTTGTTCGGGTACAGCACTGGATGAACAATGATGGGGTGGGCTCAGGCTACTCG

Mig1

Scer TTTTATGCTTTTCTGGGGTAATTAATCAGCGAAGCG--ATGATTTT-TATCTATTACAGATATATAATGGAAAAGTCATACCCAC-----TT
 Spar GTTTT--TCTTATCTCTGAGCAATTCATCCGCAAAAATAATGTTT-TT-GGTCTATTAGCAAACTATAAATGCAAAAGTTGCATAGCCAC-----TT
 Smik TTCTCA--CTTTCTCTGTGATAATTCATCACCAGAAATG--ATGTTT-TA--GGACTATTAGCAAACTATAAATGCAAAAGTCGAGAGATCA-----AT
 Sbay TTTTCGGTTTACTCTCTAGTGGCTCAT--GCAGAAAGTAATGTTTCTGTTCTTTTGCACATATATAATATGAAAATGAAGTAAAGTCGCTCAATTTGT

TATA

Scer TAACATACTTTTCAACATTTTCAGT--TTGTATTACTT-CTTATTCAAAT----GTCATAAAGATATCAACA-AAAAATGTTTAATATACCTCTATACT
 Spar TAAATAC-ATTTGCTCTCCCAAGATT--TTAATTTCTG-TTTGTTTATT----GTCATGGAATATTAACA-ACAAAGTAGTTAATATACATCTATACT
 Smik TCATTC--ATTCGAACTTTGAGACTAATATATATTAGTACTAGTTTCTTTGGAGTTATAGAAATACCAAAA-AAAAATAGTCAGTATCTATACATACA
 Sbay TAGTTTTTCTTATTCGTTGTACTTCTAGATTGTATTTCGGTTTTACTTTGTCTCCAAATATCAAAACATCAATAACAGATTCTCAACATTGT

Gal1

Scer TTAA-CGTCAAGGA--GAAAAACTATA
 Spar TTAT-CGTCAAGGAA-GAACAACATA
 Smik TCGTTATCAAGAA--AAAAACTA...
 Sbay TTATCCAAAAACCAACACACATATA

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



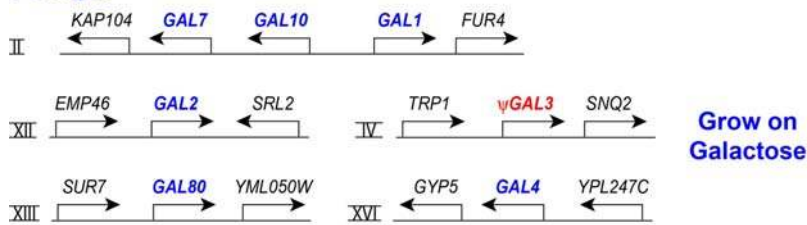
80,000,000 reads
2,800,000,000 bp



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

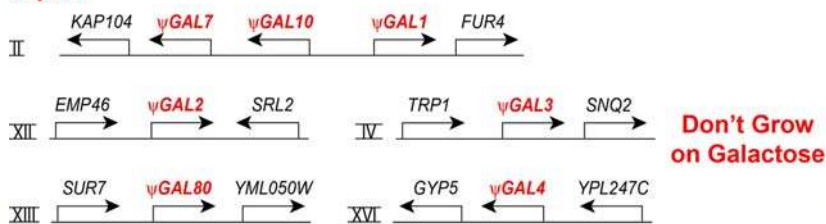
Two GAL Gene Network States in *S. kudriavzevii*

Portugal

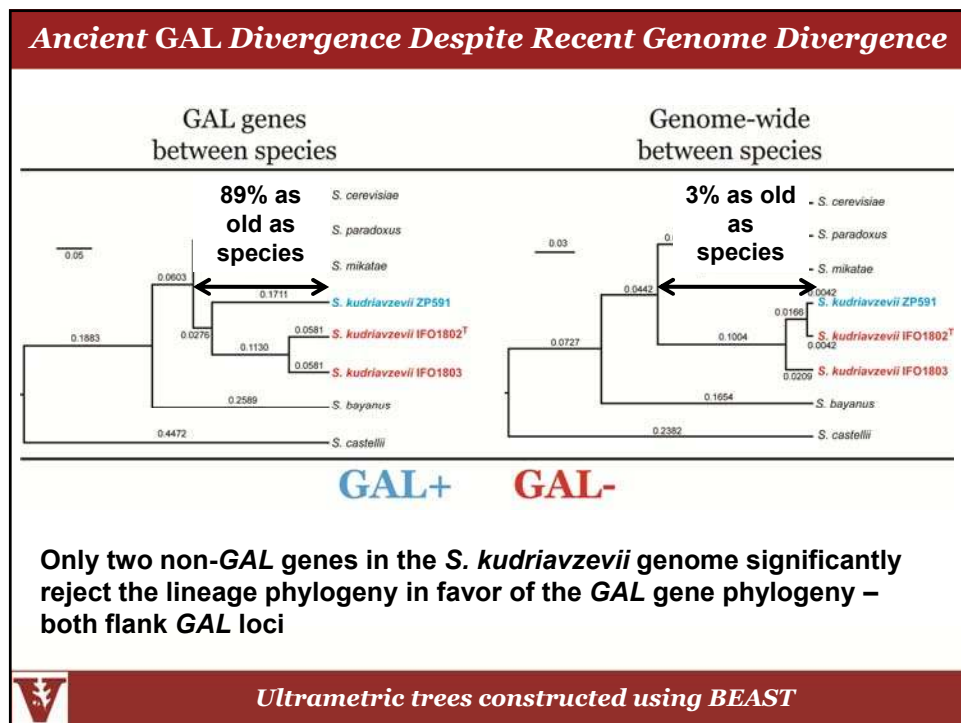
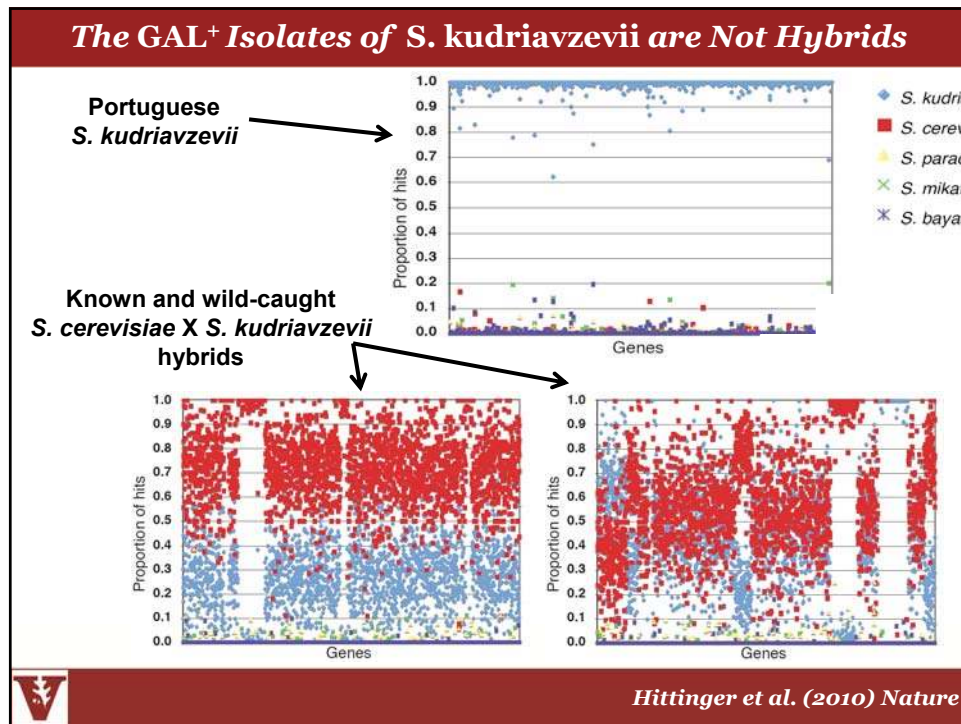


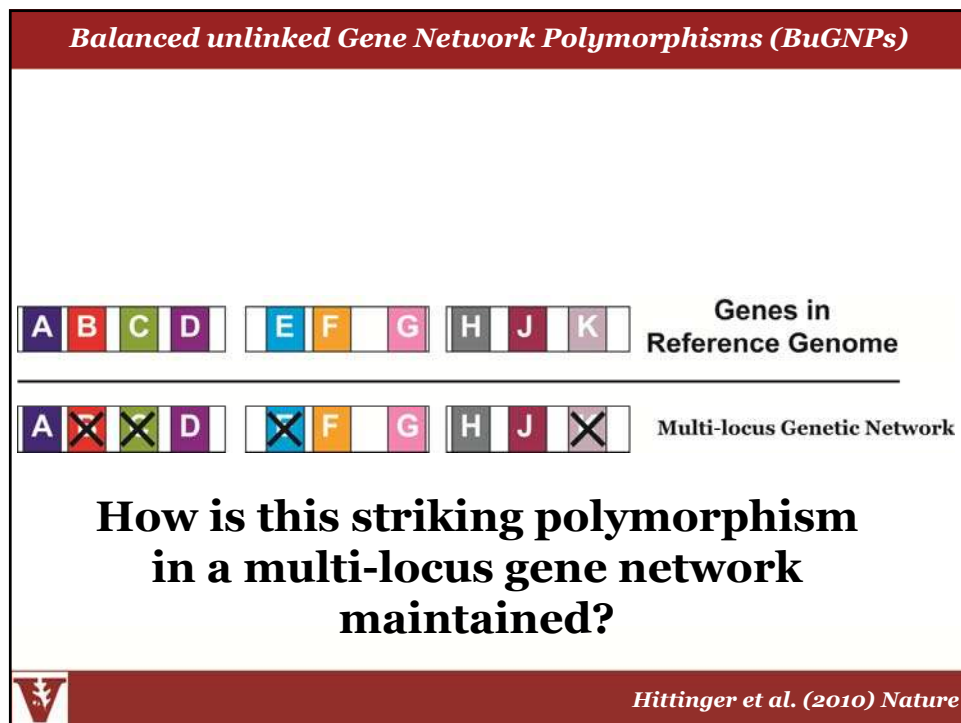
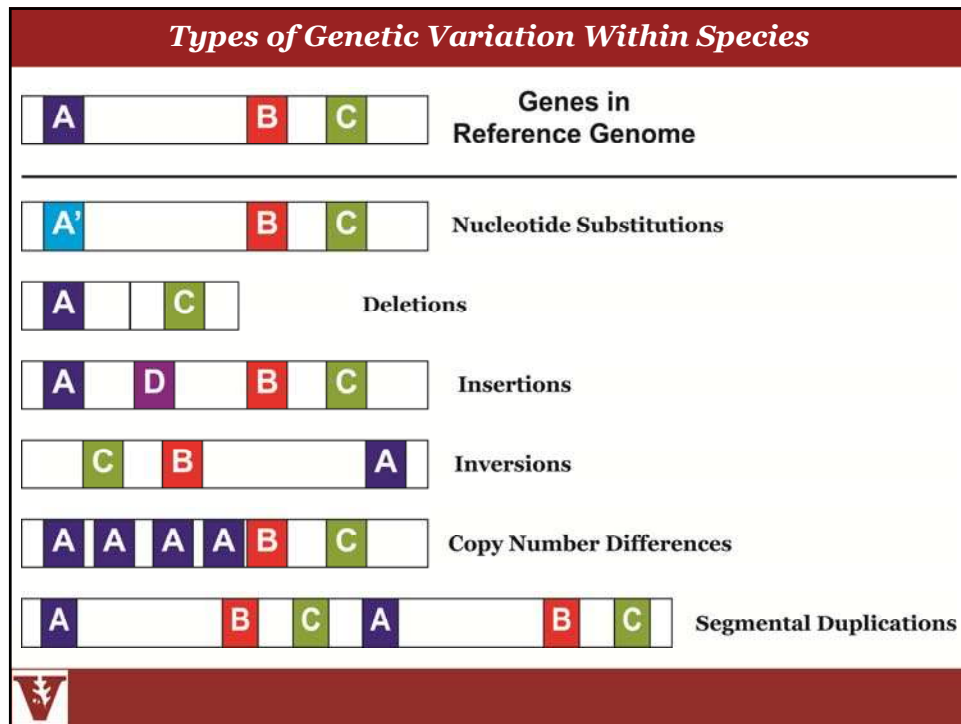
Not Hybrids / GAL genes Segregate Independently / No Post-zygotic Barriers

Japan



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**.”



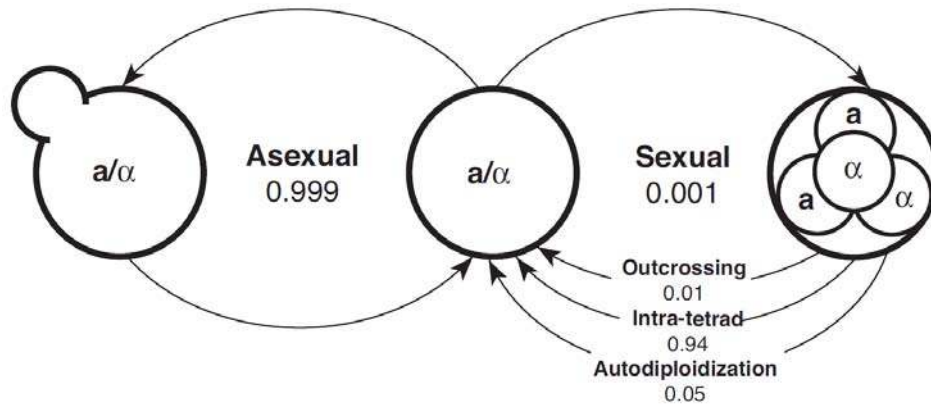
Lewontin (1998) *Genetics*

1. Maintaining BuGNPs Through Drift



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

2. Maintaining BuGNPs by Selfing

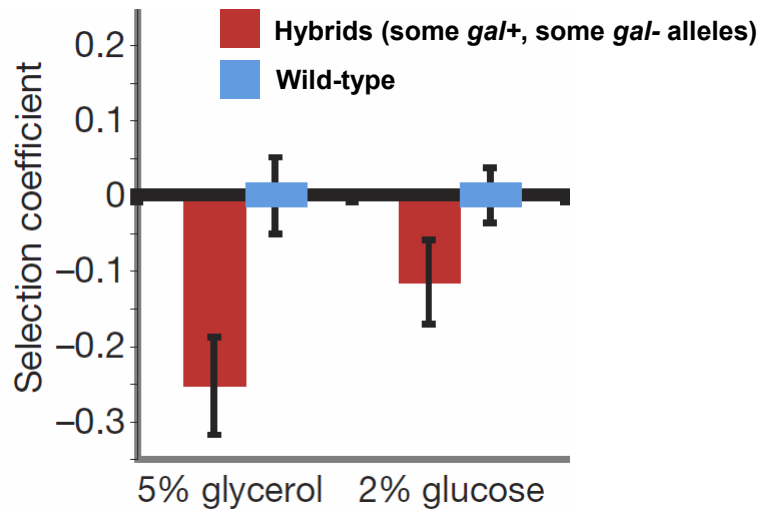


The reproductive lifestyle of *Saccharomyces paradoxus*



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?

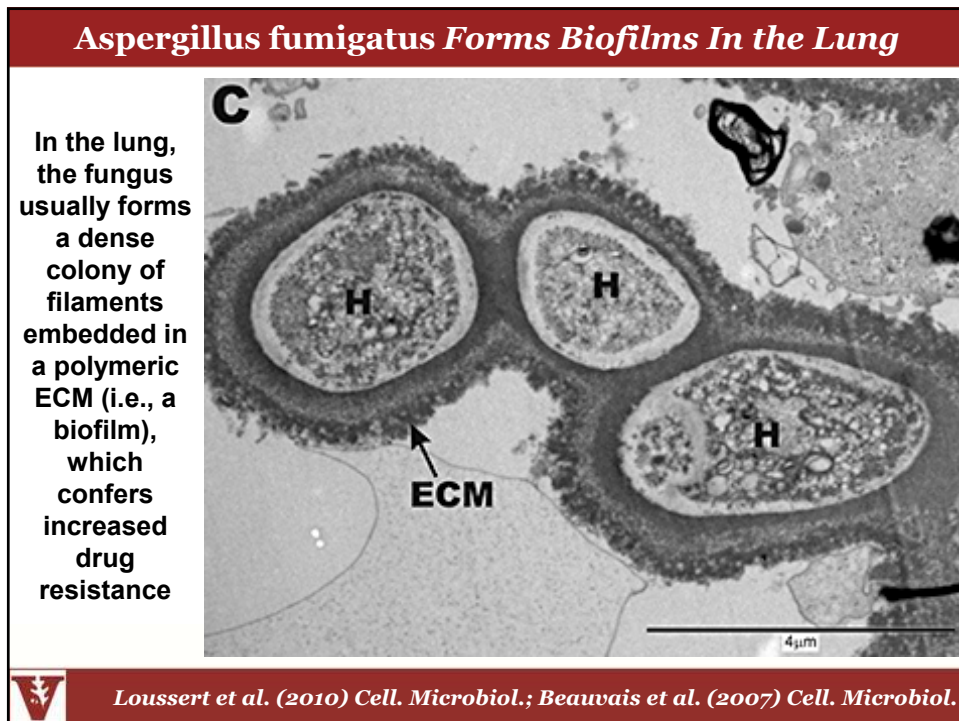
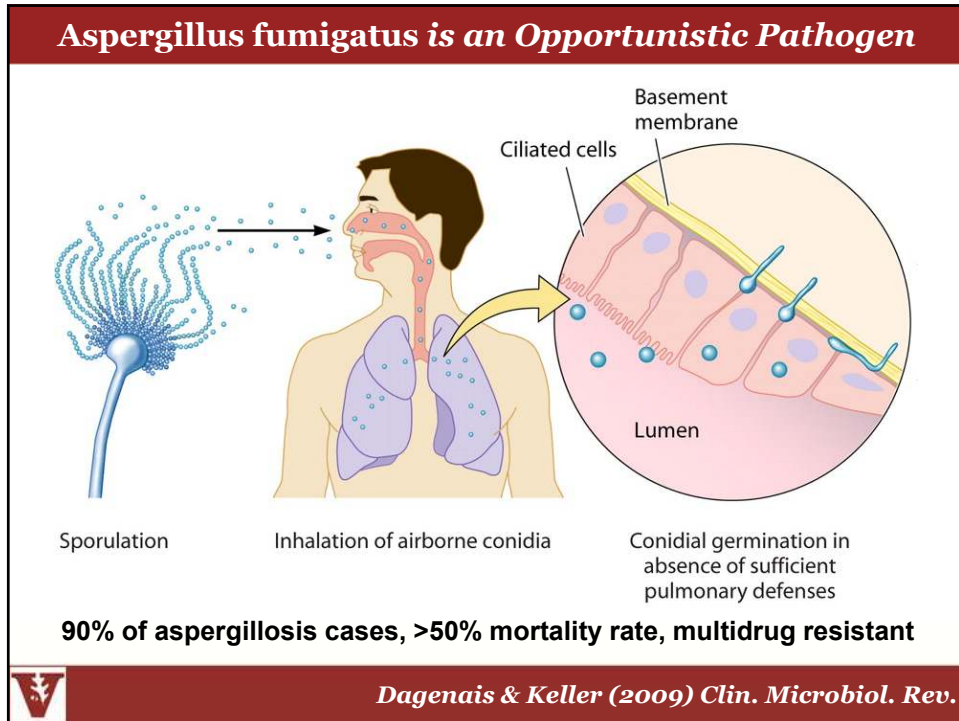
Genome remodeling
Novel metabolism and regulation
Avoidance of genetic addiction

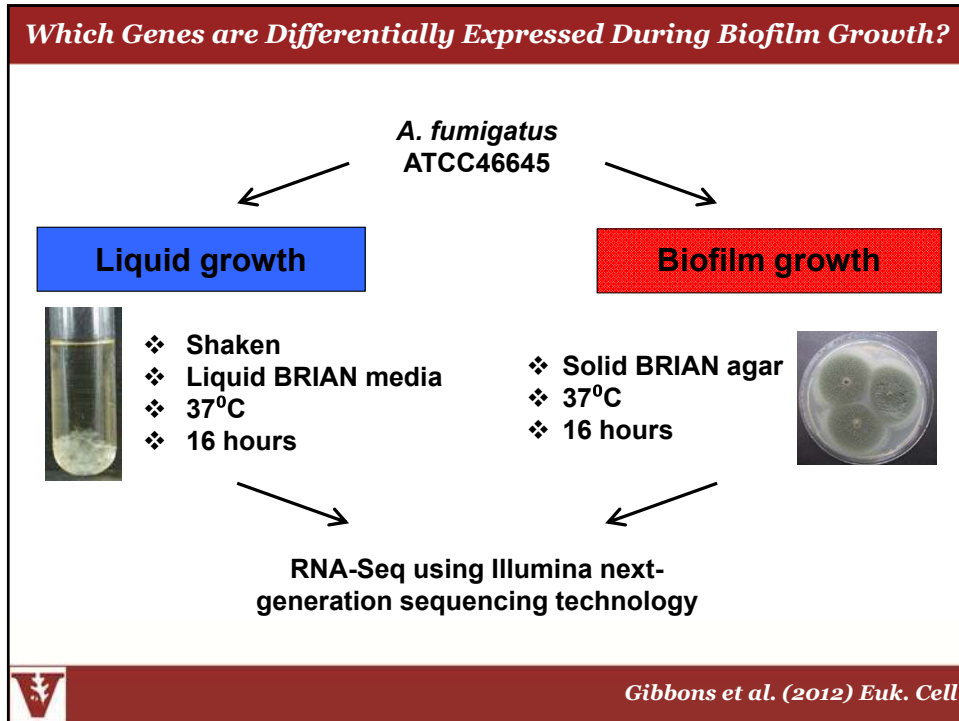


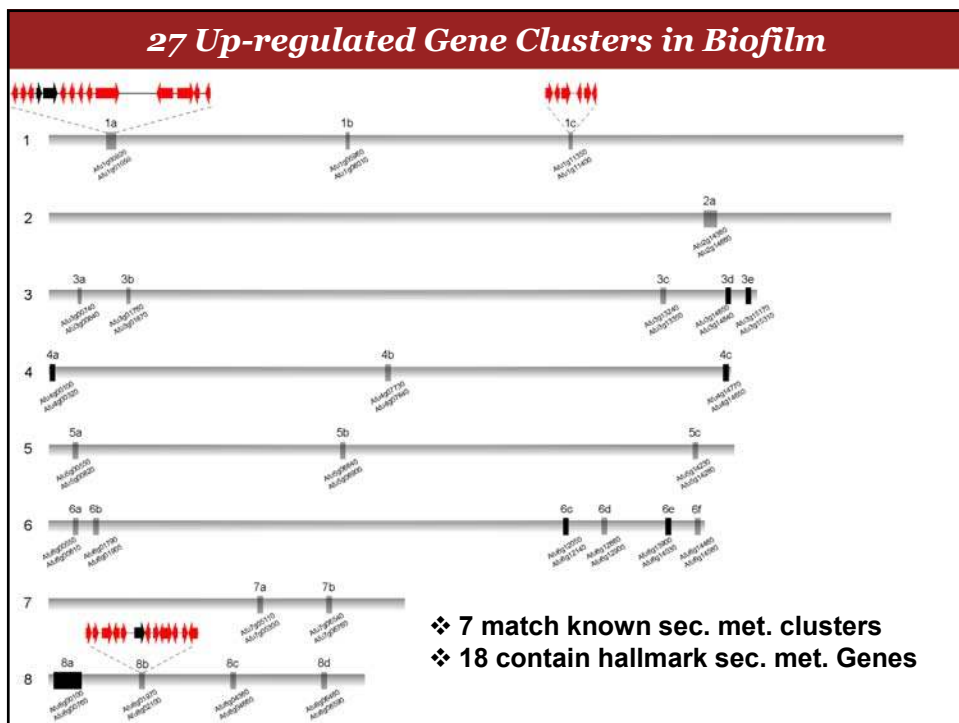
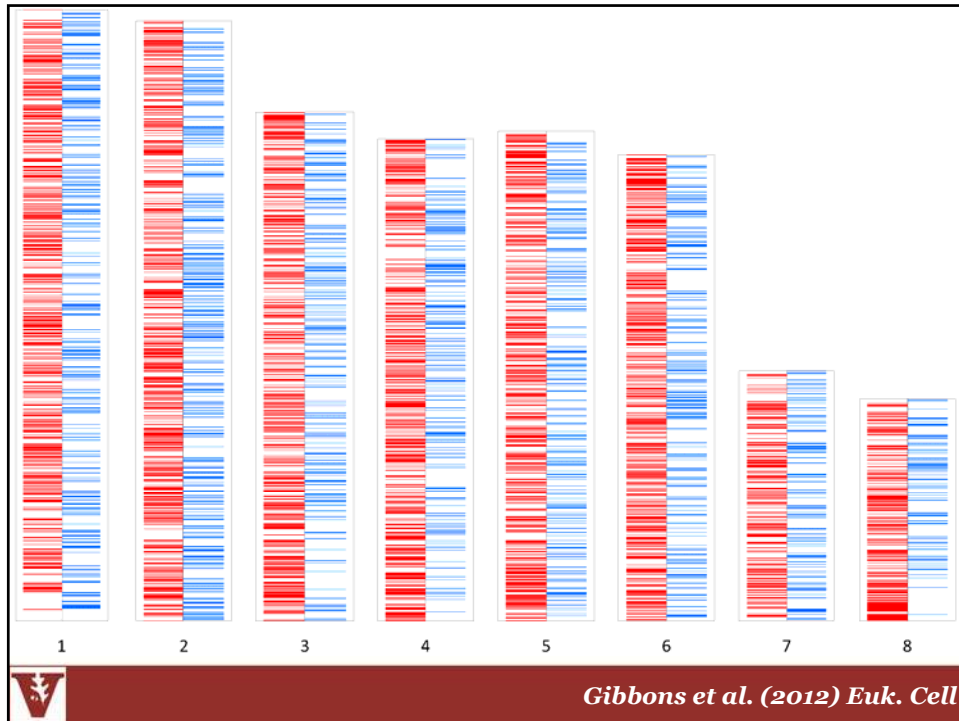
What Are the Functional Implications of Clustering?

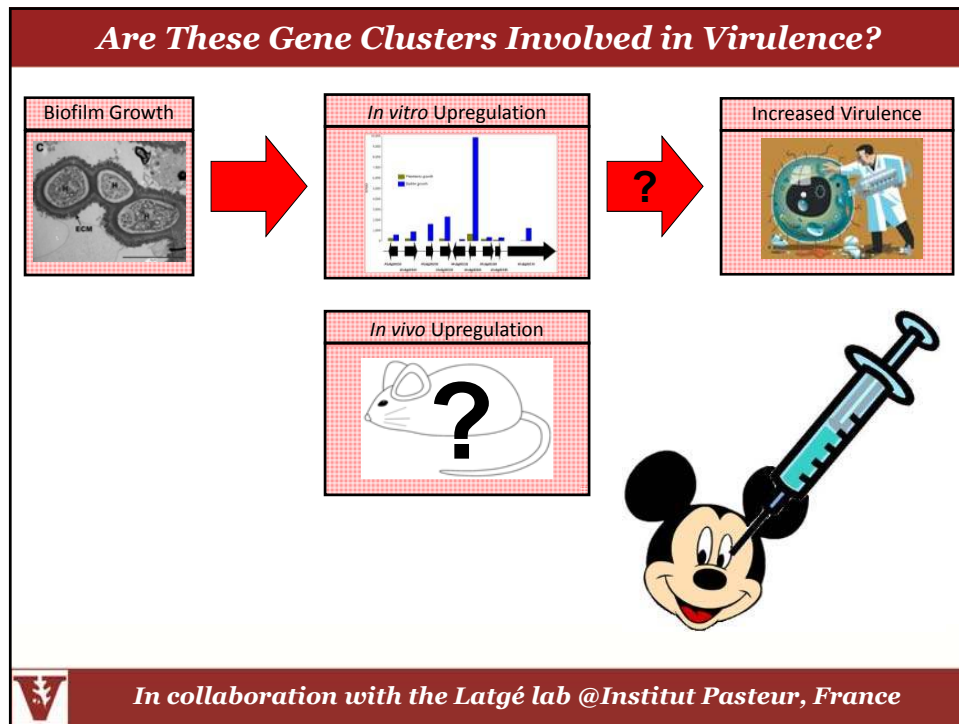


John Gibbons









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 <i>cis</i> -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	



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. *PNAS* 107: 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>