

***Evolutionary
Genomics /
Deciphering the
DNA Record***

Antonis Rokas

***Department of Biological Sciences
Vanderbilt University***

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

Lecture Outline

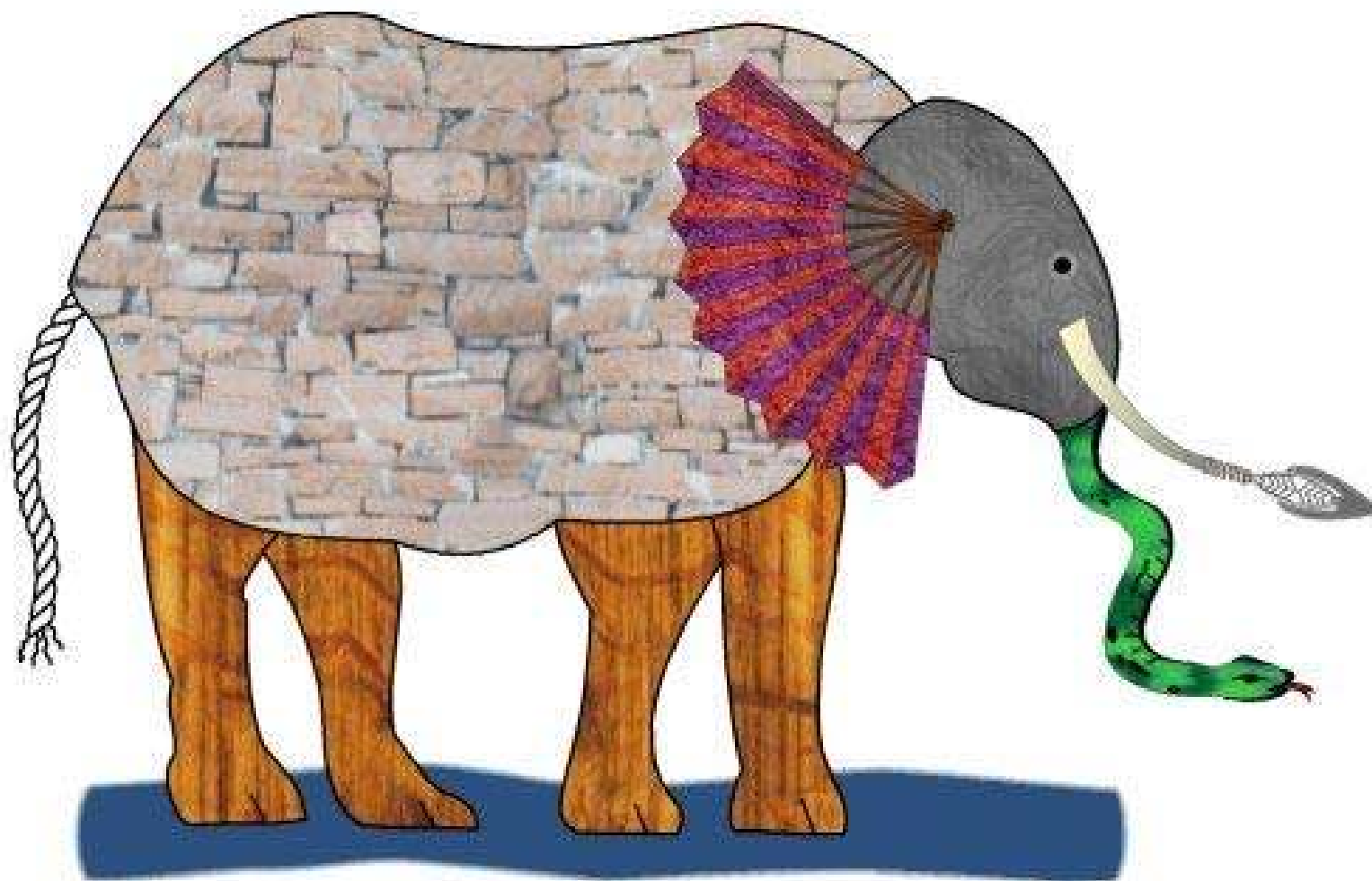
❖ **Introduction to Evolutionary Genomics**

❖ **Evolutionary and Functional Genomics**

----- **Coffee Break** -----

❖ **Phylogenomics**

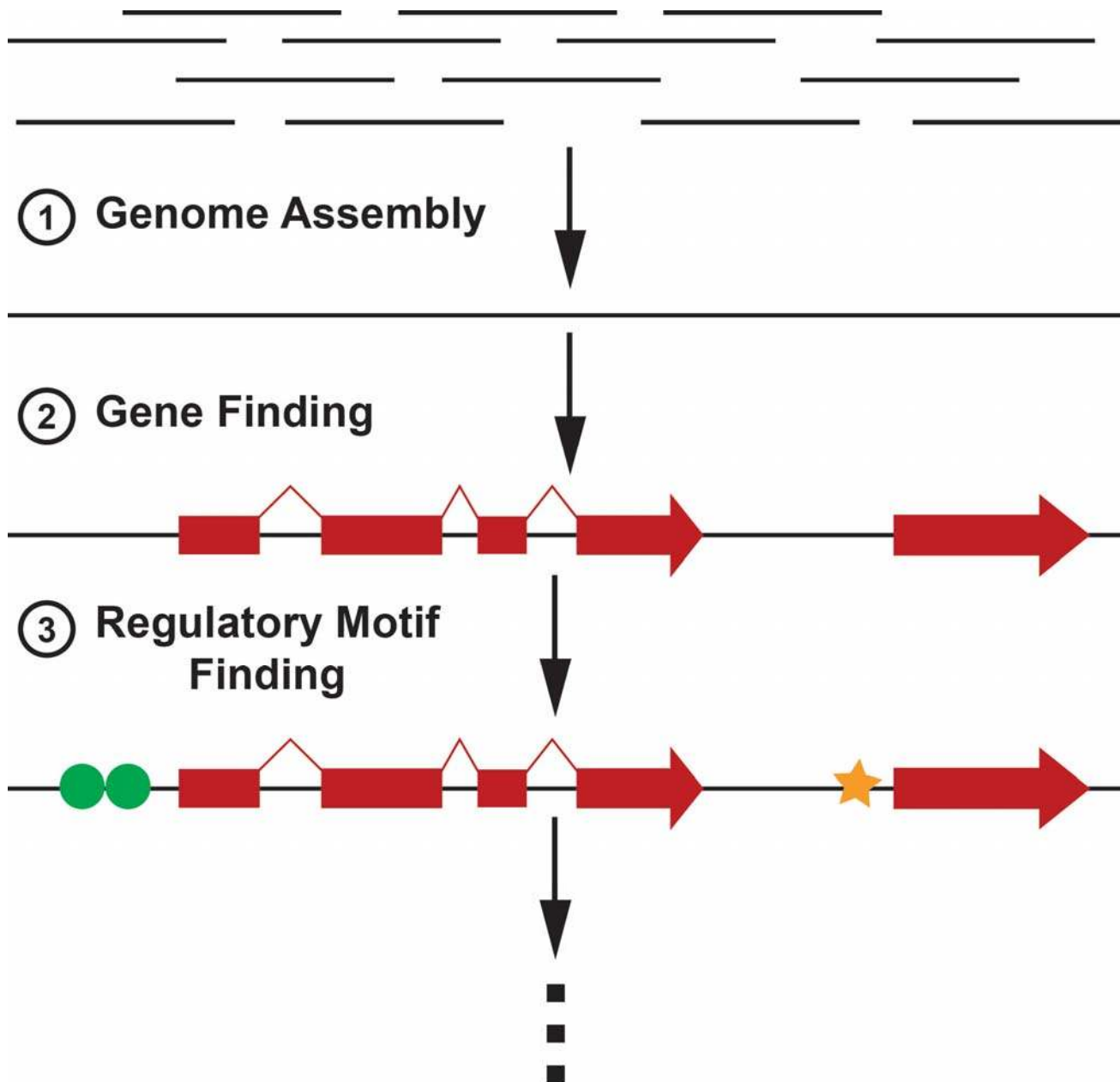
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTGAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAGGCACCCATCATACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAGCGGTGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTGCGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTGAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTGCGCTCCAGCTCCCTAGCACTGGCAGGGTTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTAAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCCTCACAGGTAAGTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGTTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACCTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACAGCCGCATCCCCTACAAAATCTACCCCAAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCGCGGACTCCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
ATCGATCCCATTGTGAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCGAGGGGCCACCAGGAACCATGTTT
CAGTCTGACAATAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAACCCCTTGGGCCCAGCGGAACAAGGAAA
AGGGGATACAGTGCGGATTTCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT

Understanding the Genome Requires Tools



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAAGTGTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATAACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAGCGGTGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGTTTCTCTCACTACTCAGTTCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTAAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGATCTGCTCTCTCCAAGTCT
GCGATATTGCTTATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCGTACAGGTAAGTCTGAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGTAAGCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTTCGTAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACCTTAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCTACAAAATCTACCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCGCGGACTCCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
ATCGATCCCATTGTGAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCGAGGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCTACATTCATATGGGCCAGCGATTGGAACCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT

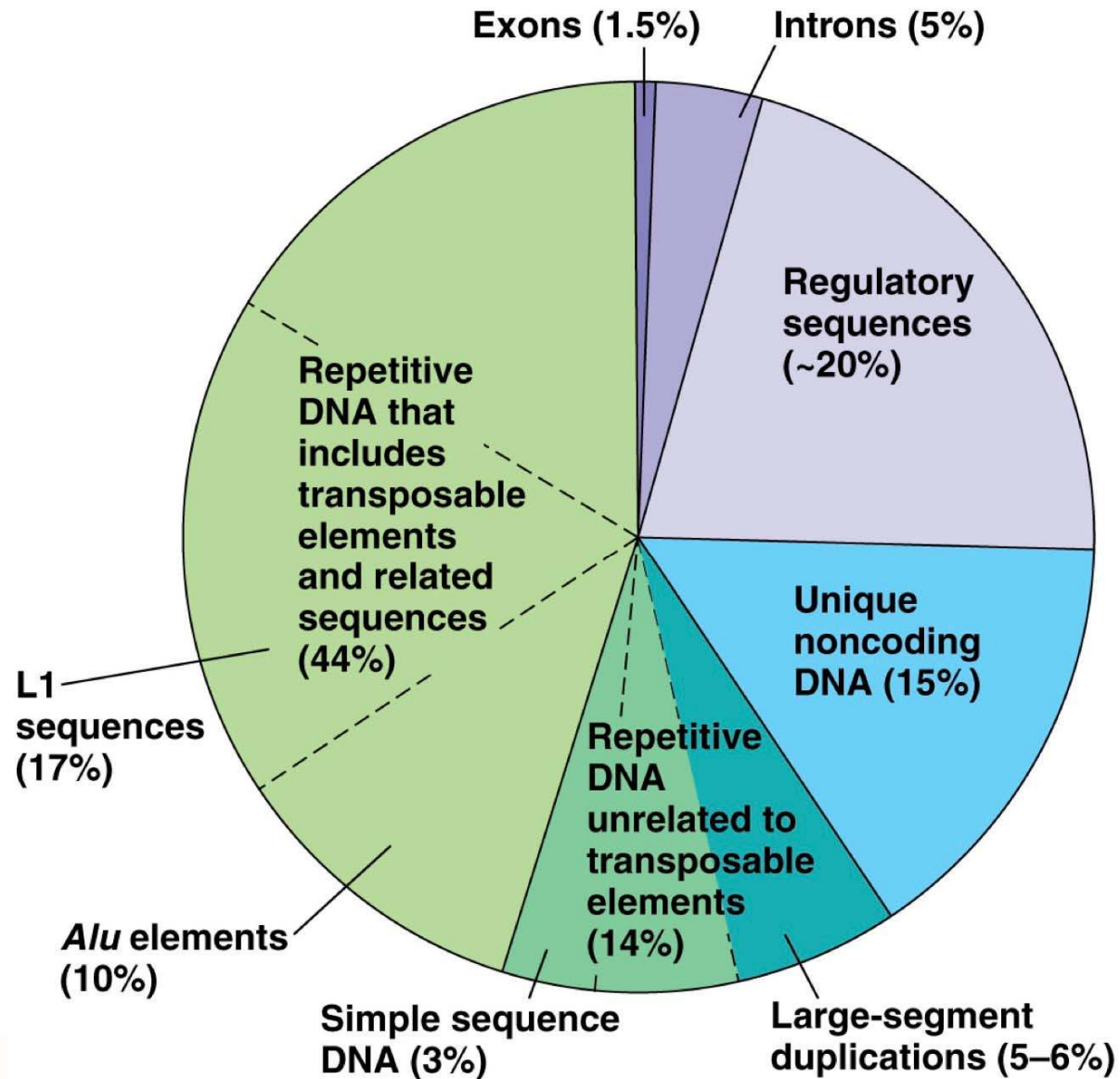
Transposon

Protein Binding Site

Exon

Intron

Organization of the Human Genome

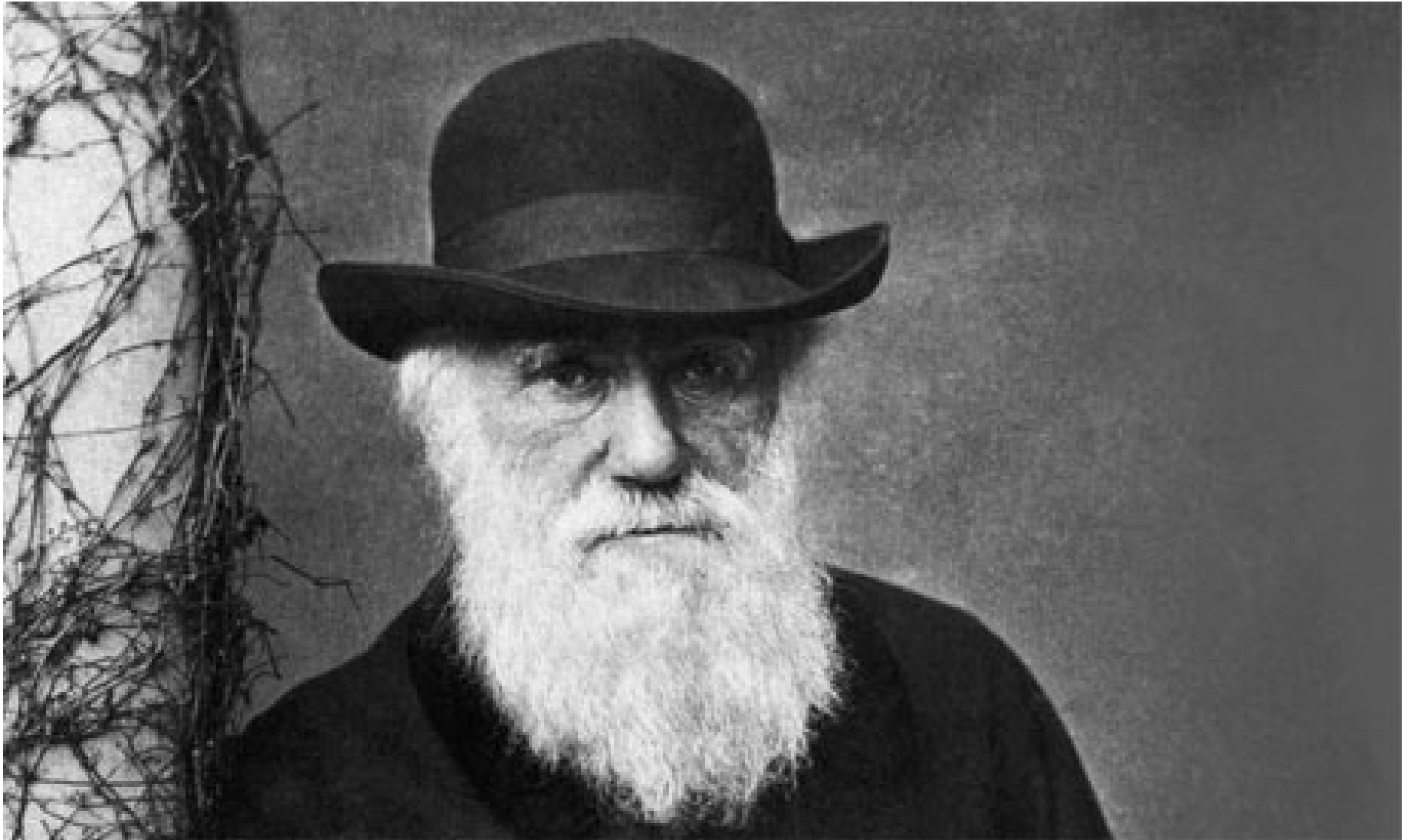


© 2011 Pearson Education, Inc.

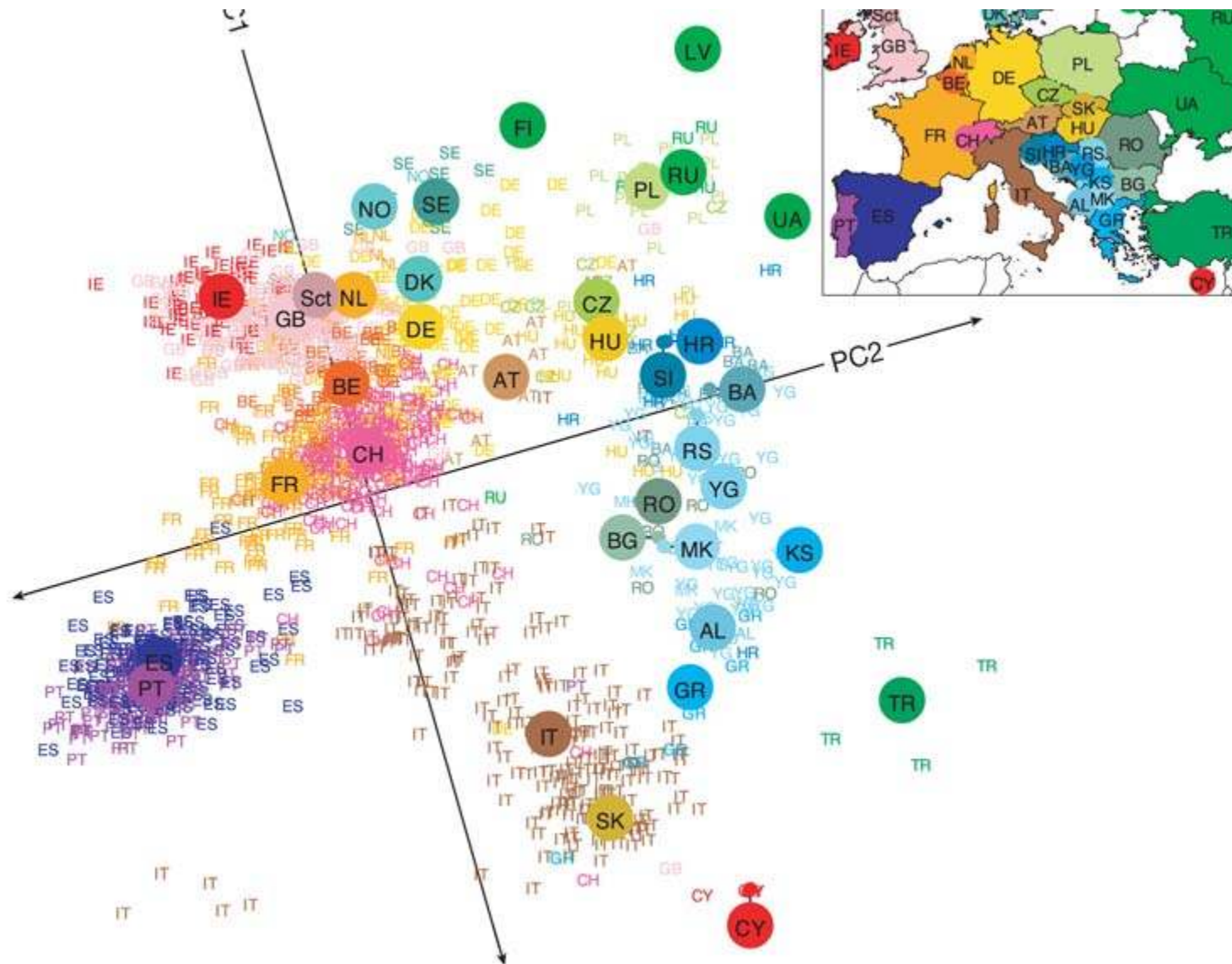


<http://todd.jackman.villanova.edu/HumanGenome.jpg>

Understanding the Genome Requires a Theory



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	<i>BLZF1, SLC19A2</i>
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	<i>RAB3GAP1, R3HDM1, LCT</i>
chr2:177.9	CEU, CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	<i>PDE11A</i>
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	<i>SLC30A9</i>
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	<i>PCDH15</i>
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	<i>HERC1</i>
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	<i>CHST5, ADAT1, KARS</i>
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	<i>BCAS3</i>
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	
chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	LARGE
Total SNPs			16.74	9,166	2,898	480	41	

In the European population,
involved in skin pigmentation

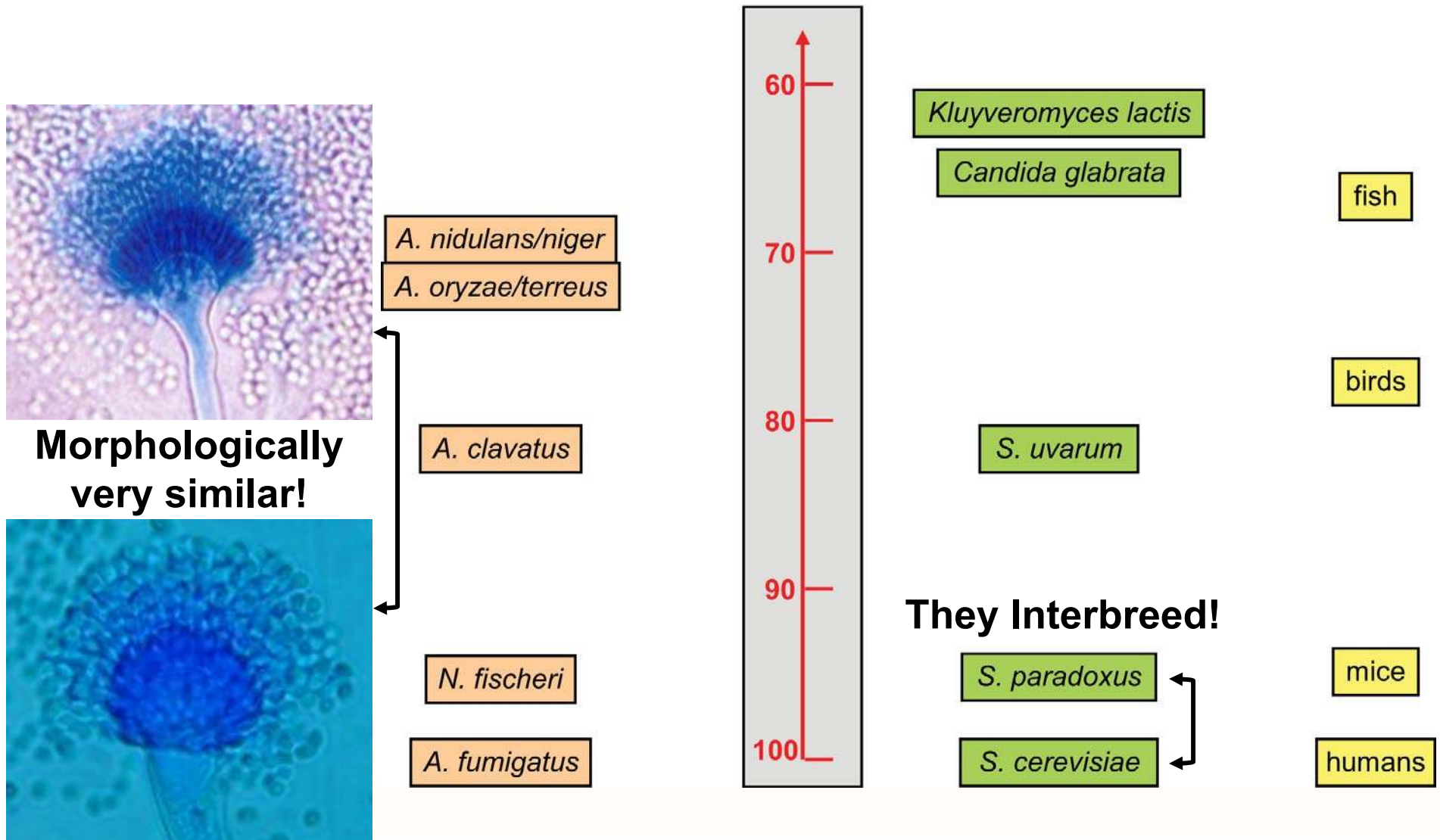
In the West African population,
related to Lassa virus infection



Sabeti et al. (2007) Nature

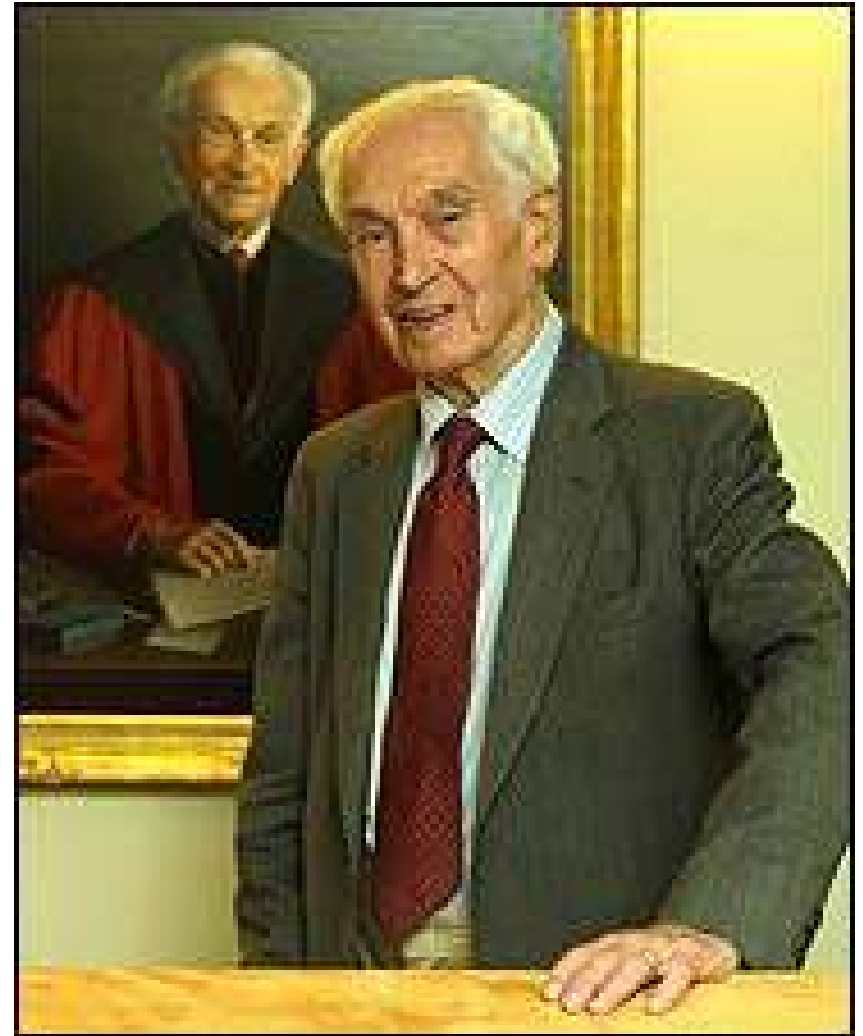
Genomes Provide a Common Yardstick for Comparison

Average proteome sequence similarity



**“...the search for homologous
genes is quite futile except in
very close relatives”**

Ernst Mayr, 1963



What Makes Us Sick Is the Stuff of Life

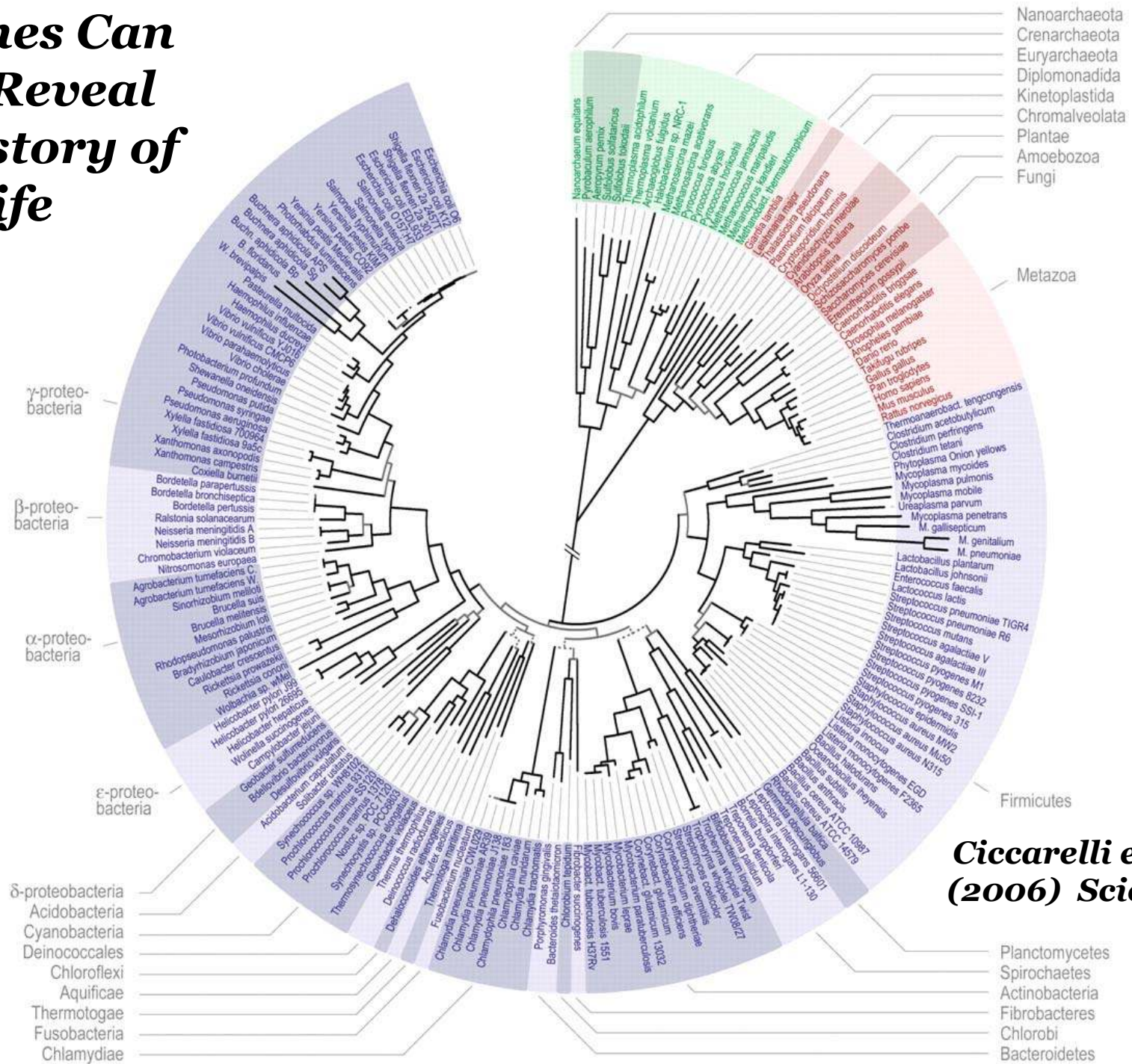
F	W	Y	Cancer	F	W	Y	Neurological	F	W	Y	Malformation Syndromes
+			ABL1	+			Adrenoleukodystrophy-ABCD1	-			Aarskog-Scott-FGD1
+			Acute Myeloid Leukemia-DEK	+			Alzheimer-PS1	+			Achondroplasia-FGFR3
+			Adenomat. Polyposis Coli-APC	+			Alzheimer-APP	+			Alagille-JAG1
+			AKT2	+			Amyotrophic Lat. Sclero.-SOD1	+			Barth-TAZ
+			Ataxia Telangiectasia-ATM	+			Angelman-UBE3A	-			Beckwith-Wiedemann-CDKN1C
-			BRCA1	+			Aniridia-PAX6	-			Cerebral Cavern. Malf.-CCM1
-			BRCA2	+			Best Macular Dystrophy-VMD2	+			Chondrodyspl. Punct. 1-ARSE
+			Basal Cell Nevus-PTC	+			Ceroid-Lipofuscinosis-PPT	+			Cleidocranial Dysplasia-OFC1
+			B-Cell Lymphoma 2-BCL2	+			Ceroid-Lipofuscinosis-CLN3	-			Cockayne I-CKN1
-			B-Cell Lymphoma 3-BCL3	-			Ceroid-Lipofuscinosis-CLN2	+			Coffin-Lowry-RPS6KA3
+			Bloom-BLM	-			Charcot-Marie-Tooth 1A-PMP22	+			Diastrophic Dyspl.-SLC26A2
+			Burkitt's Lymphoma-MYC	-			Charcot-Marie-Tooth 1B-MPZ	+			EEC 3-Ket. P63
-			CDKN2C	+			Choroideremia-CHM	+			Greig Cephalopolysynd.-GLI3
-			CSF1R/C-Fms	-			Creutzfeldt-Jakob-PRNP	-			Hand-Foot-Genital-HOXA13
+			Chk2 Protein Kinase	+			Deafness, Hereditary-MYO15	+			Holoprosencephaly 3-SHH
-			PDGFB	+			Deafness, X-Linked-TIMM8A	+			Holoprosencephaly-SIX3
+			CML-BCR	+			Diaphanous 1-DIAPH1	+			Holt-Oram-TBX5
+			Cyclin D1-CCND1	+			Dementia, Multi-Infarct-NOTCH3	-			ICF-DNMT3B
+			Cyclin Dep. Kinase 4-CDK4	+			Duchenne MD ⁺ -DMD	+			Kallman-KAL1
+			EGFR	-			Emery-Dreifuss MD ⁺ -EMD	-			Laterality, X-Linked-ZIC3
+			ERBB2	+			Emery-Dreifuss MD ⁺ -LMNA	+			Melnick-Fraser-EYA1
-			ETS	+			Familial Encephalopathy-PI12	+			Nail Patella-LMX1B
+			E-Cadherin-CDH1	+			Fragile-X-FRAXA	-			Opitz-MID1
+			Ewing Sarcoma-FLI-1	+			Friedreich Ataxia-FRDA	+			Renal Coloboma-PAX2
-			FGF3	+			Frontotemporal Dement.-TAU	+			Rieger, Type 1-PITX2
-			Fanconi's Anemia A-FANCA	-			Frontotemporal Dement.-TAU	-			Rubinstein-Taybi-CREBBP
-			Fanconi's Anemia C-FANCC	+			Fukuyama MD ⁺ -FCMD	+			Saethre-Chotzen-TWIST
-			Fanconi's Anemia G-FANCG	+			Huntington-HD	-			Septo-optic Dysplasia-HESX1
+			HNPCC*-MSH2	+			Limb Girdle MD ⁺ 2A-CAPN3	+			Simpson-Golabi-Behmel-GPC3
+			HNPCC*-MSH3	+			Limb Girdle MD ⁺ 2B-YSF	+			Townes-Brockes-SALL1
+			HNPCC*-MSH6	-			Limb Girdle MD ⁺ 2E-BSG	+			Treacher-Collins-TCOF1
+			HNPCC*-MLH1	+			Lissencephaly, X-Linked-DCX	-			VMCM-TEK
+			HNPCC*-PMS2	+			Lowe Oculocerebroren.-OCRL	+			Wardenburg-PAX3
-			KIT	-			Machado-Joseph-MJD1	+			Zellweger-PEX1
				+			Miller-Dieker Lissen.-PAF				



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

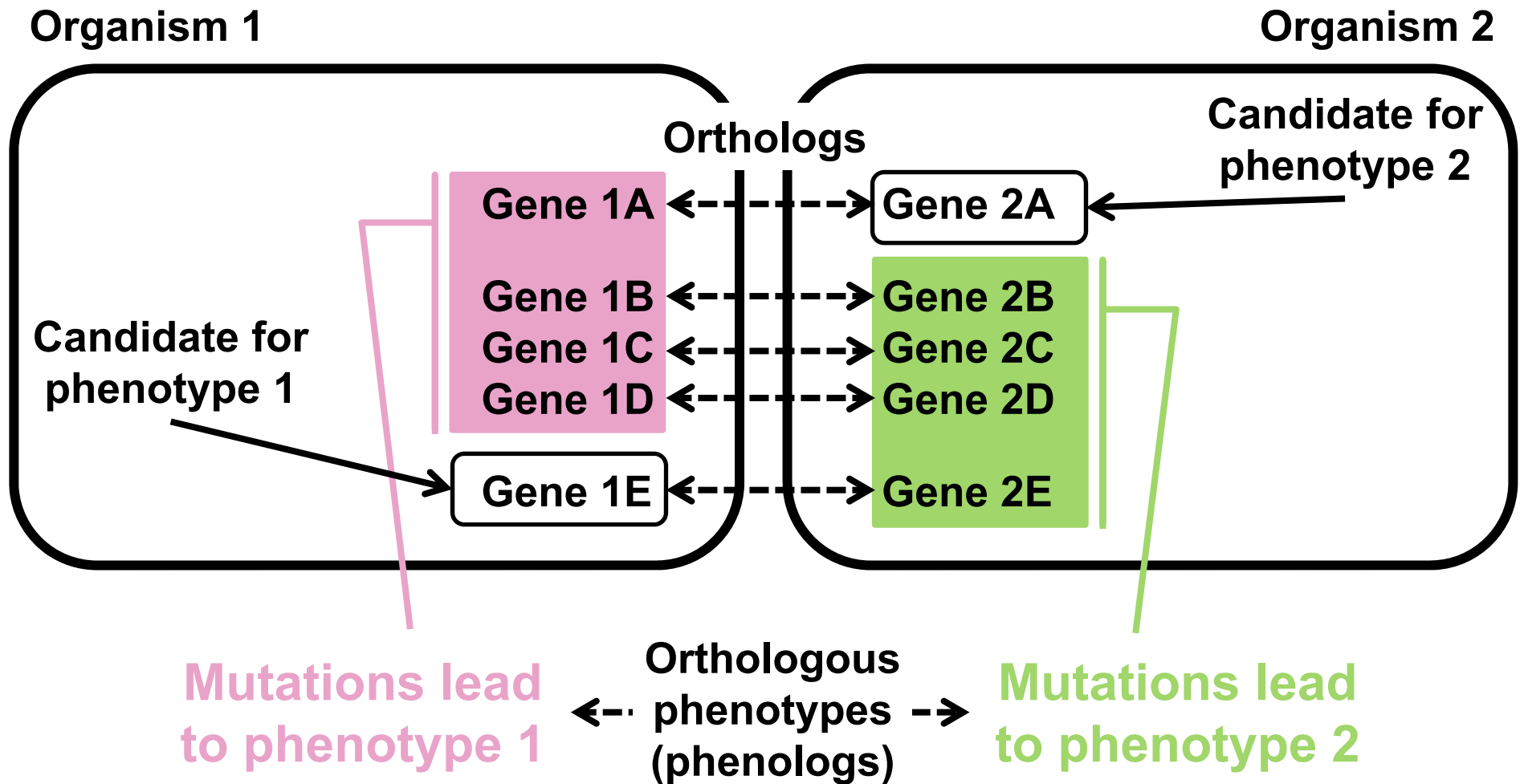


Genomes Can Help Reveal the History of Life



**Ciccarelli et al.
(2006) Science**

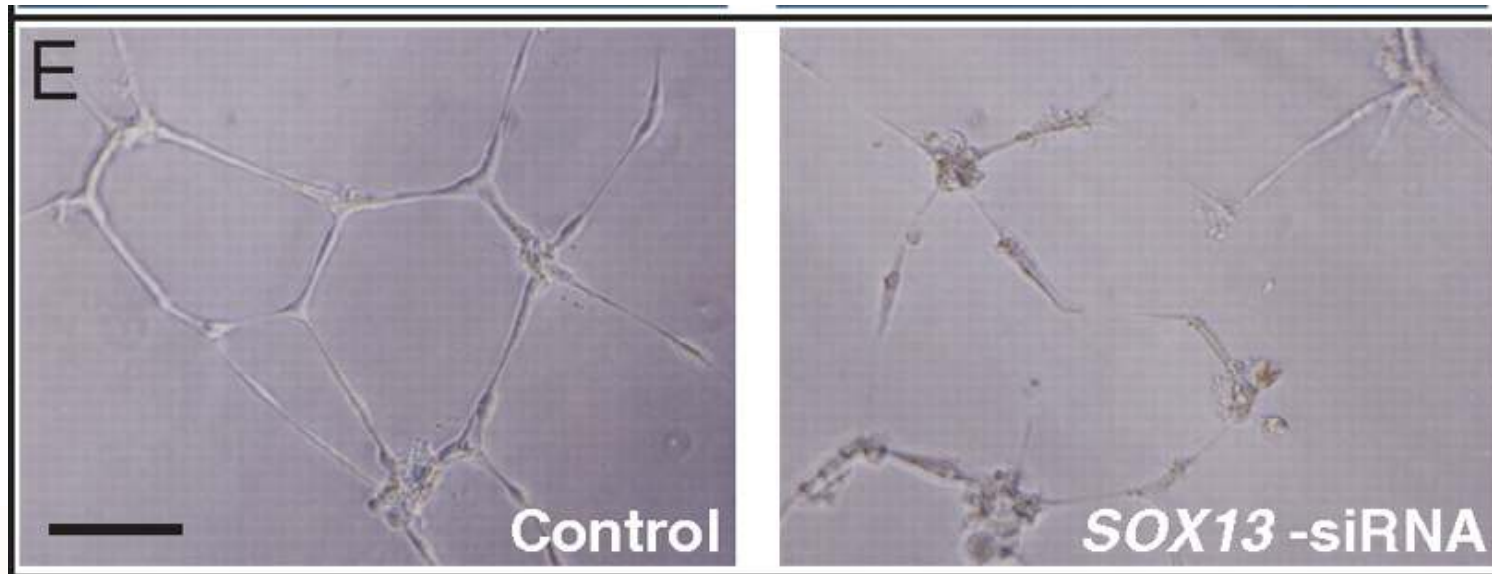
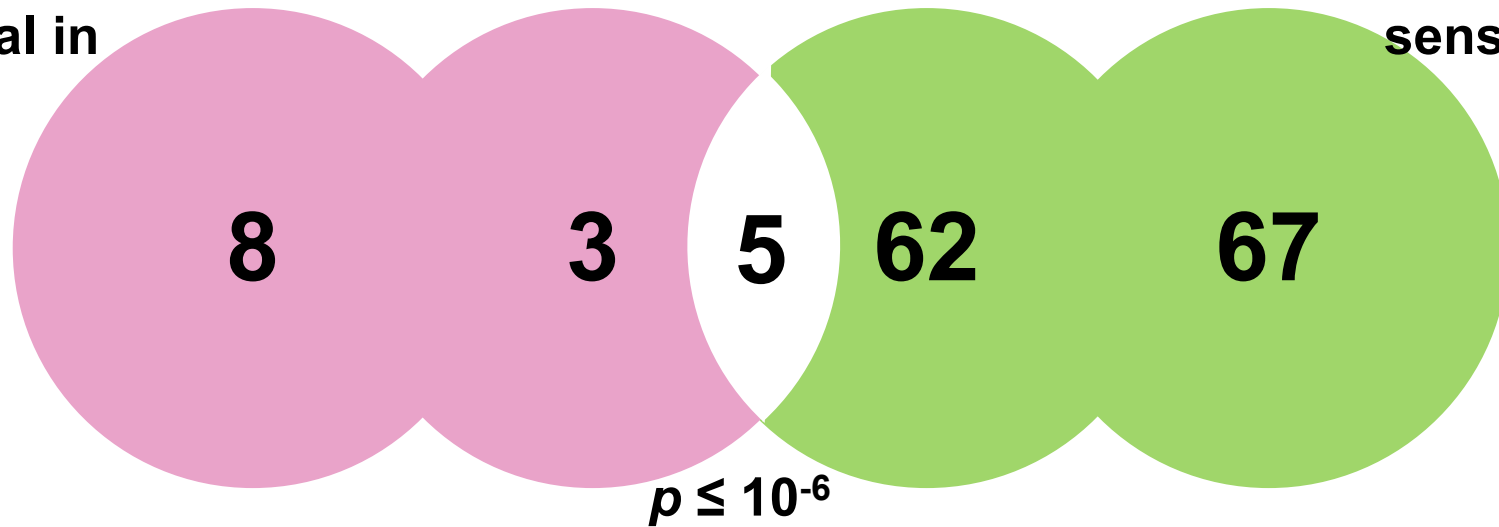
Evolution-Informed Analyses Have Great Predictive Power



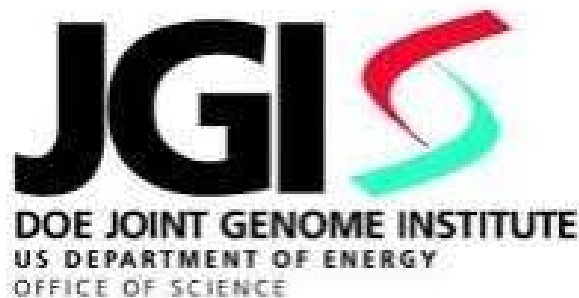
A Yeast Model for Angiogenesis

Angiogenesis
abnormal in
mice

Lovastatin
sensitive in
yeast



Genomics Used to Be “Big Science”



High-Throughput DNA Sequencing Technologies

ILLUMINA



454 / Roche

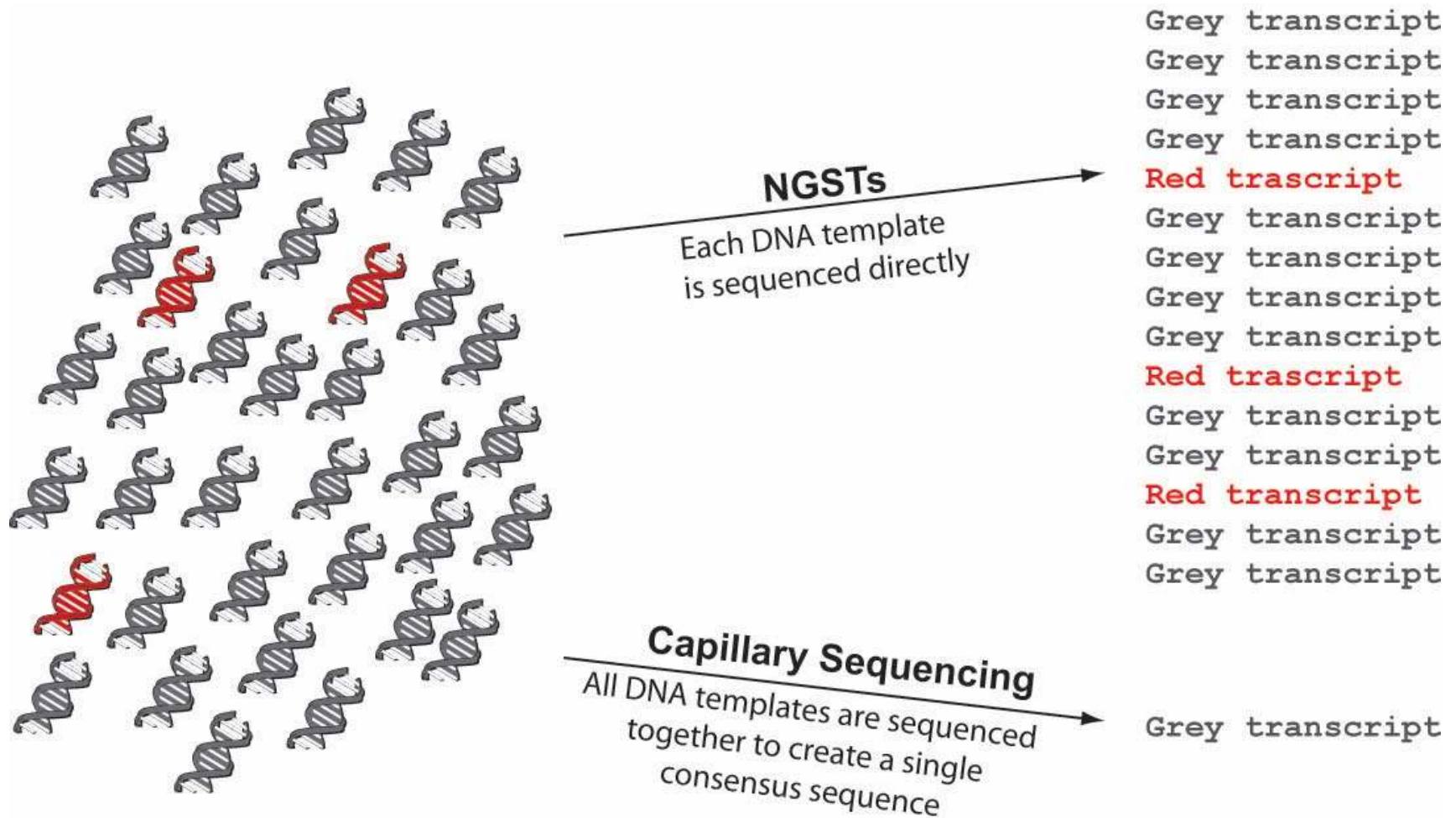


PacBio

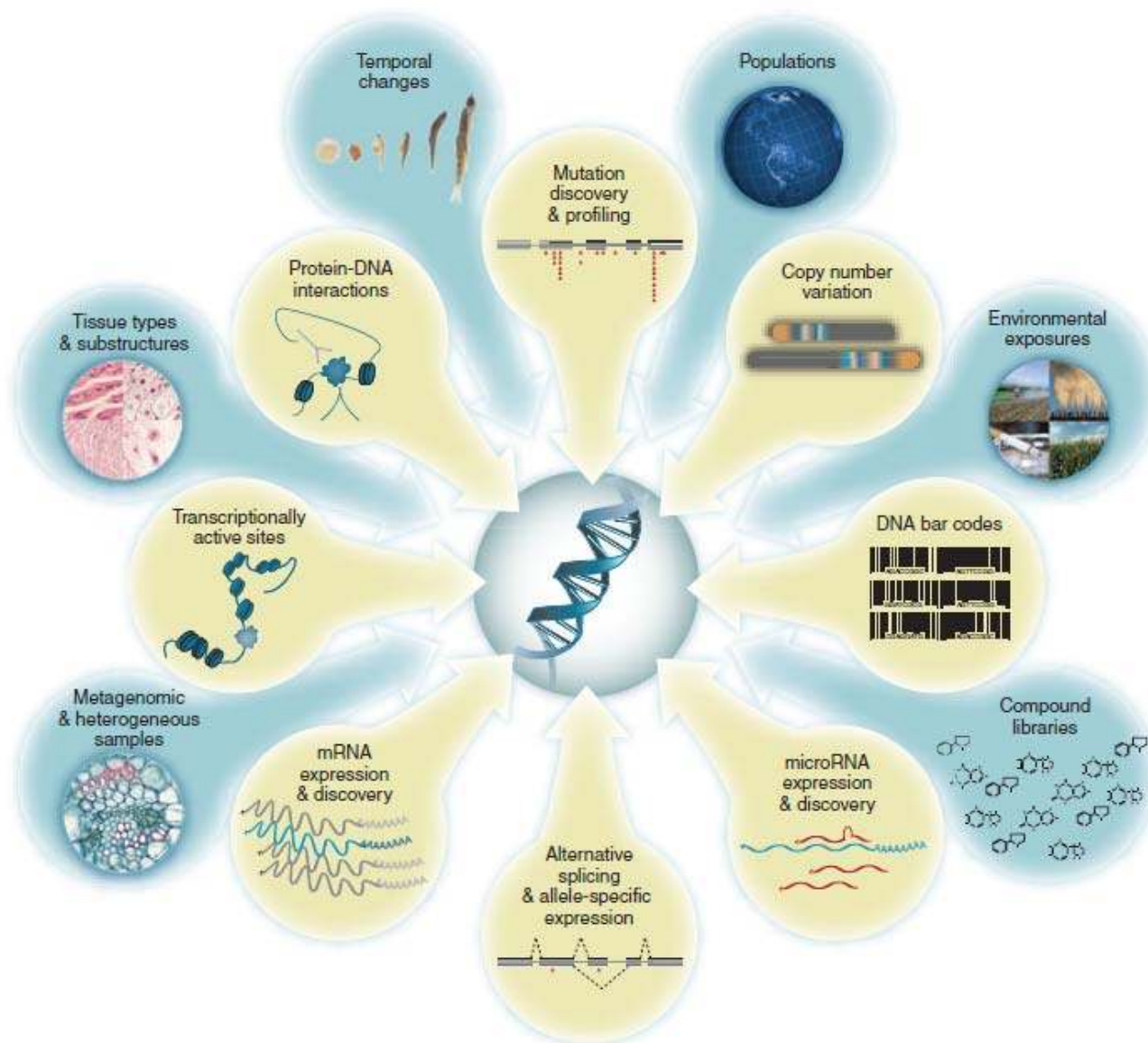


AB / Life Technologies

High Throughput Sequencing is Qualitative and Quantitative

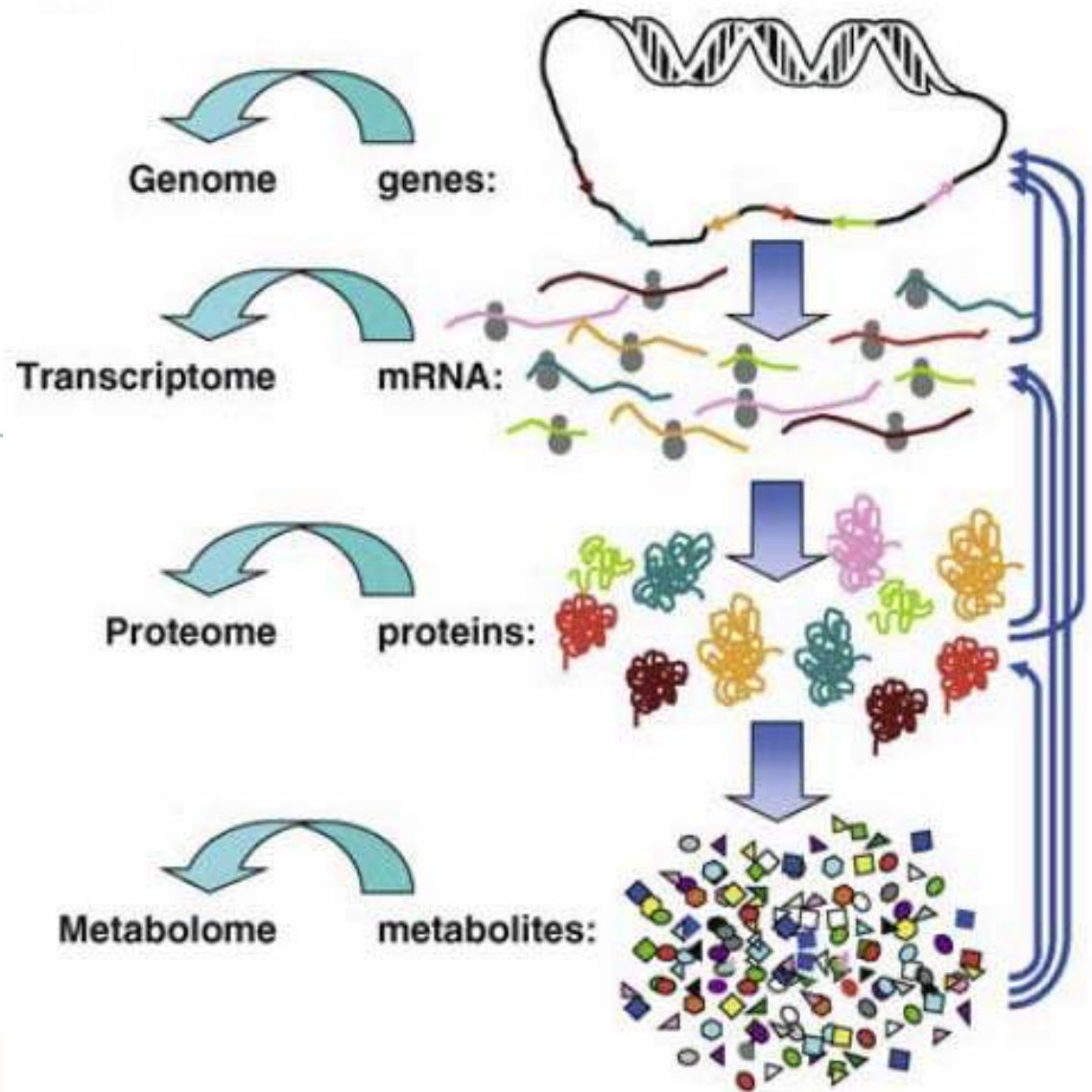


Sequence Census Assays



Kahvejian et al . (2008) Nature Biotech.

The Age of High Throughput Technologies

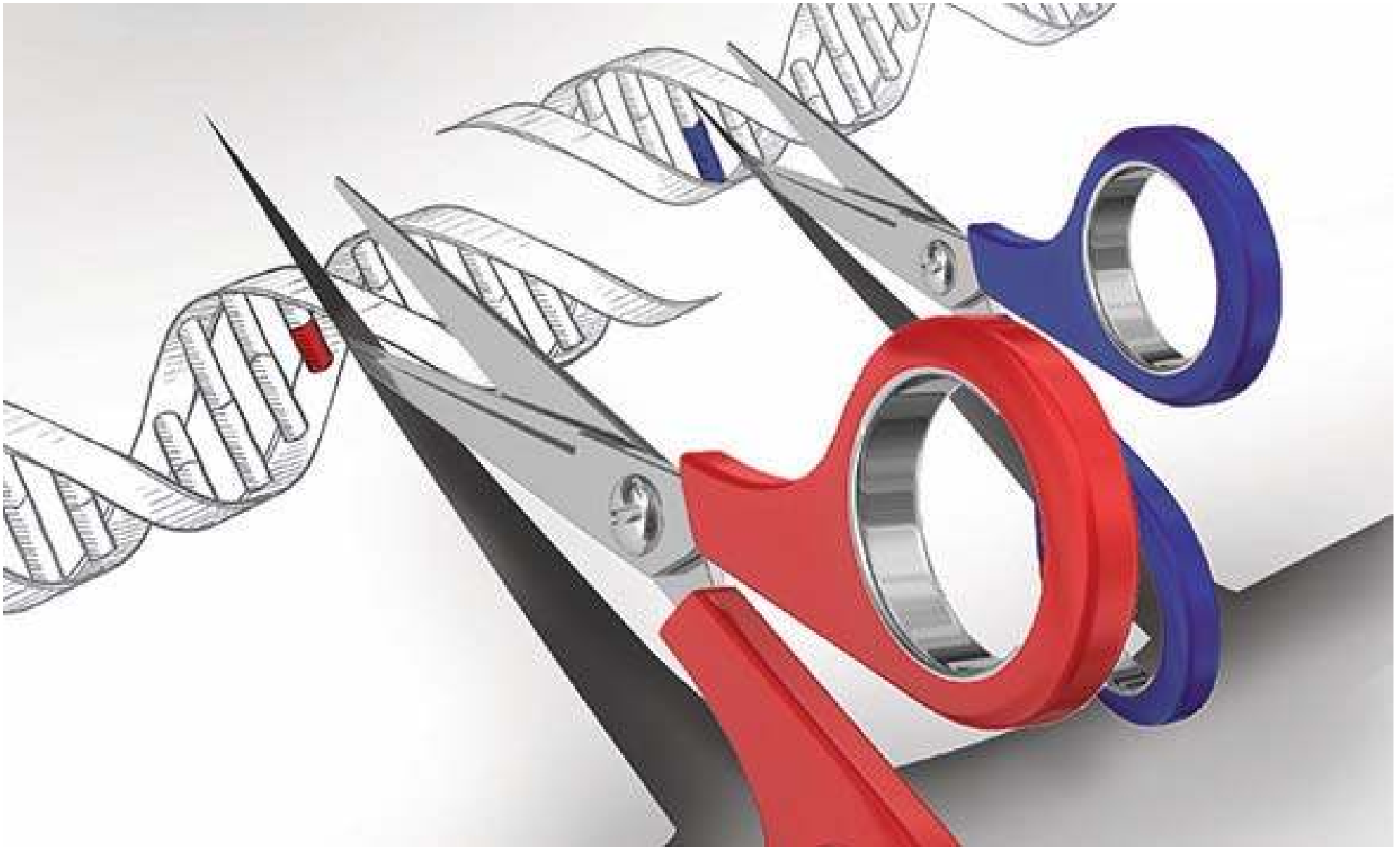


Goodacre (2005) Metabolomics

Novel Ways to Probe Gene Function in Any Organism

RNAi

TALENs / ZFNs and other nucleases / CRISPRs

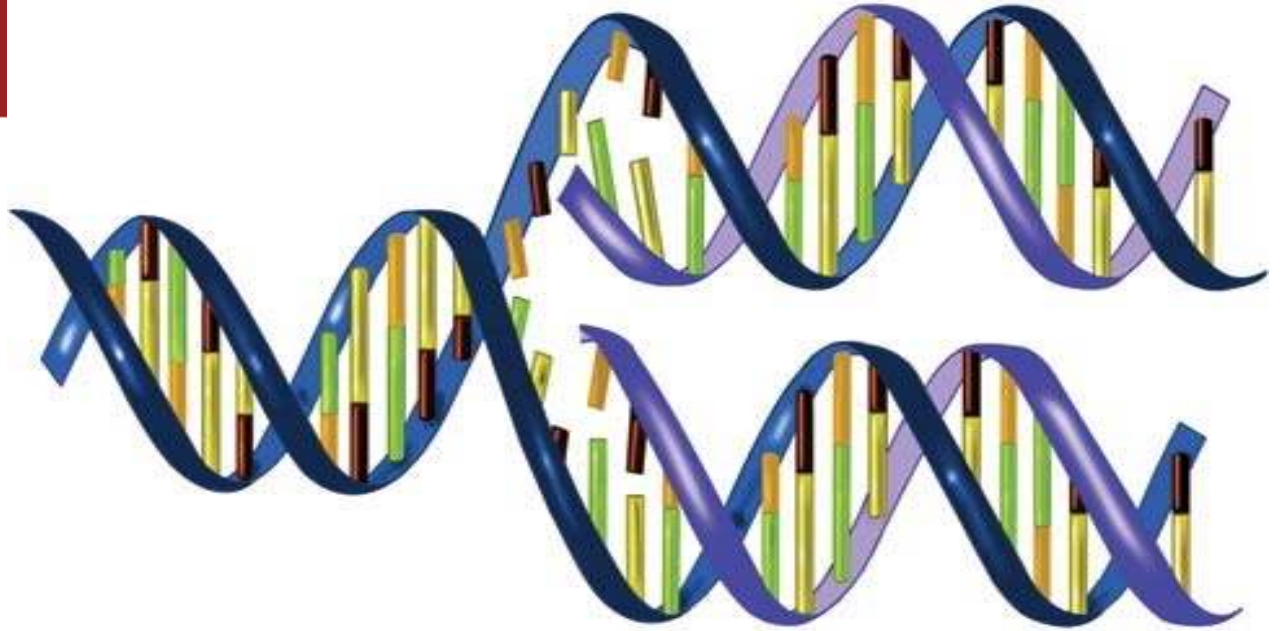


The Genomes of Non-Model Organisms are the New Frontiers



Rokas & Abbot (2009) Trends Ecol. Evol.

The DNA Record



“The genome is, it's a fossil record; the genome is a landscape; the genome is a whole geography of distributions. [...] The genome is a storybook that's been edited for a couple of billion years, and you could take it to bed, like *A Thousand and One Arabian Nights*, and read a different story, in the genome, every night.”

Eric Lander

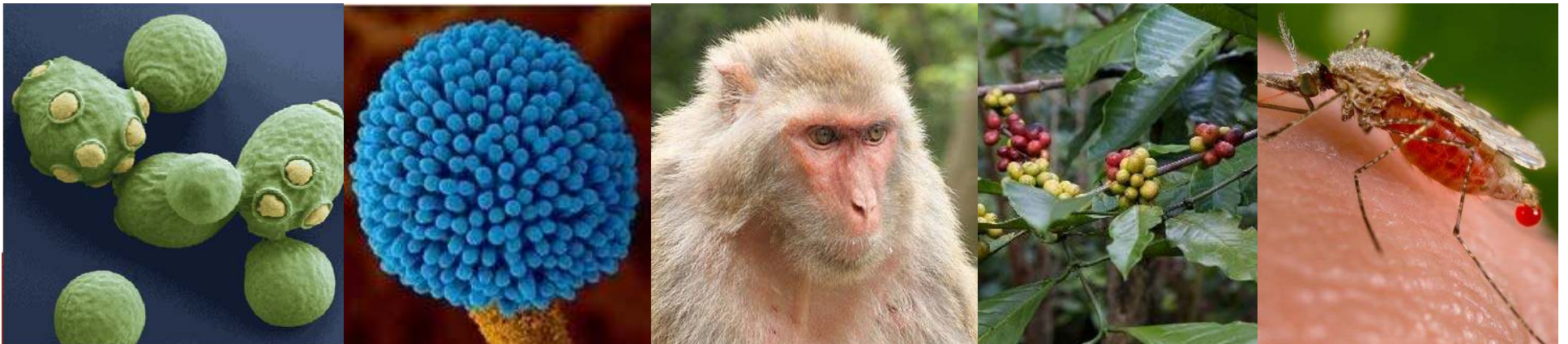
The Rokas Lab



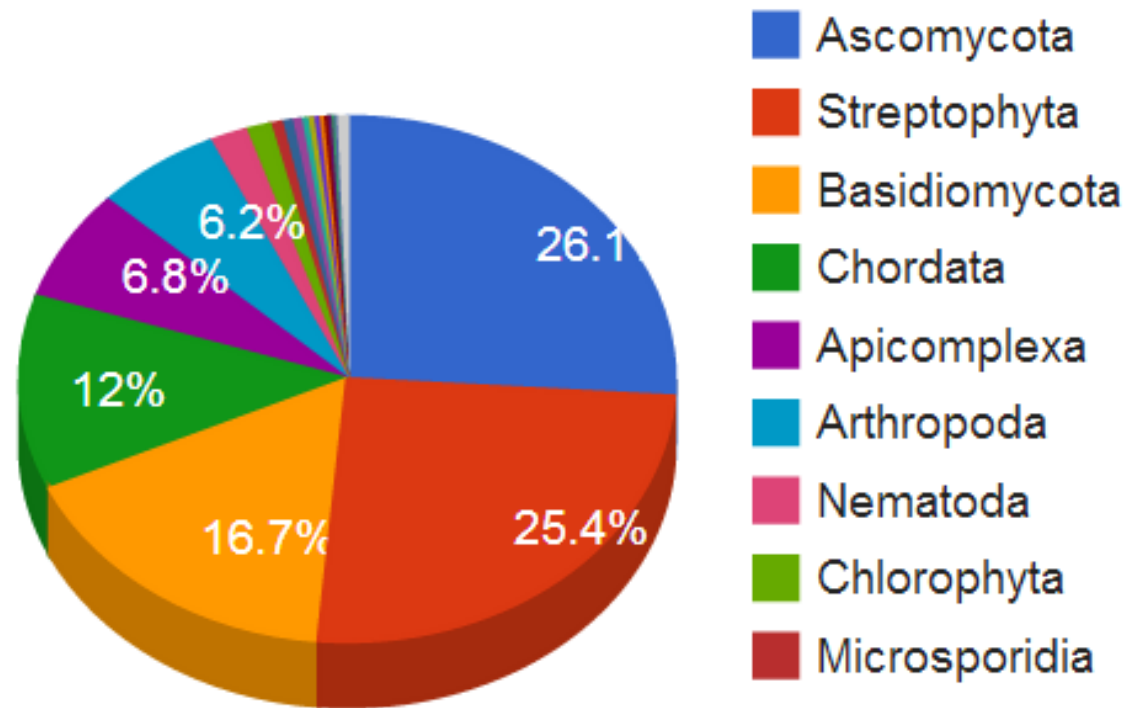
We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches

Rokas Lab Research Themes

- ❖ **Phylogenomics of ancient divergences**
- ❖ **The evolution of fungal chemodiversity**
- ❖ **The evolution of human pregnancy**



One in Two Eukaryotic Genomes Comes From a Fungus



What Functional Stories Can We Read in the DNA Record?

“Does the genome sequence *per se* now enable us to decode the language of life? The answer, of course, is an emphatic "no".

[...] the inactivation of most genes yields either no obvious phenotype or early death, neither of which is immediately informative of the gene's function. In any case, even a gene whose mutant phenotype is, say, a change in nose shape, cannot be called "a gene for the nose".

The genome does not say "make nose", it says "make serpentine receptors"—it speaks biochemistry, not phenotype.”





“Fungi characteristically live embedded in a food source or medium, in many cases excreting enzymes for external digestion, but in all cases feeding by absorption of organic food from the medium.”



Whittaker (1969) Science

Fungi “Eat” Almost Everything, Grow on Any Surface



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

“I put maybe a shot of whiskey in a liter of agar and filled the petri plates with it. That made [the fungus] grow a hell of a lot faster”



*Dr. James Scott, Mycologist
Wired Magazine, 06/11 issue*

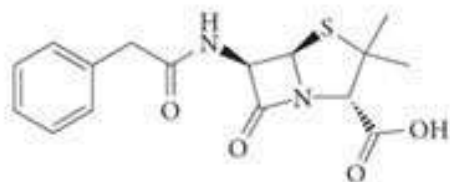
What Fungi Cannot Eat!



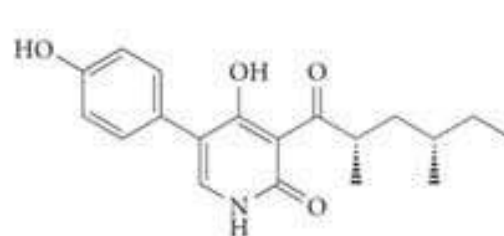
Diner fries

McDonald's fries

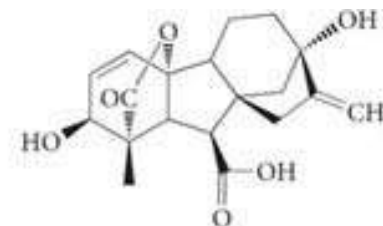
Fungi are Superb Chemical Engineers



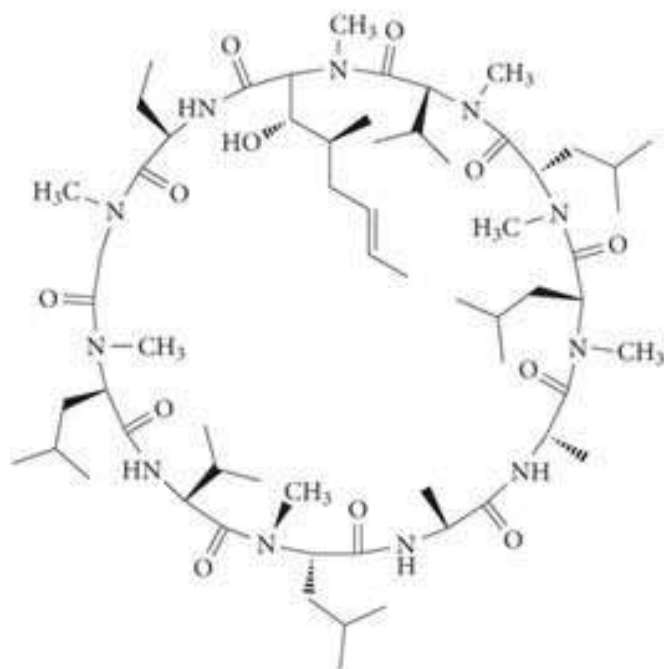
Penicillin G



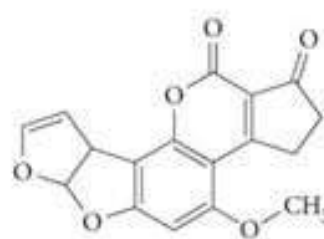
Aspyridone A



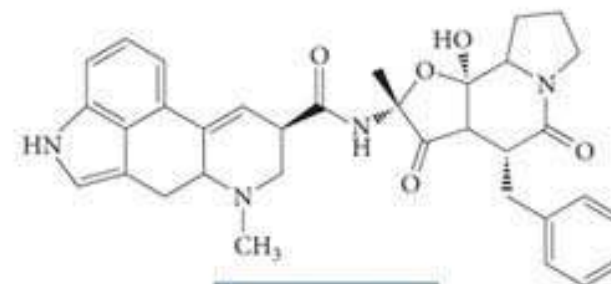
Gibberellin A3



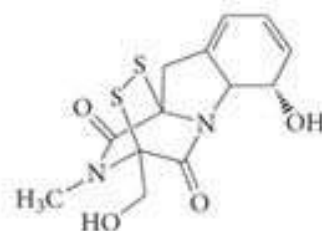
Cyclosporine A



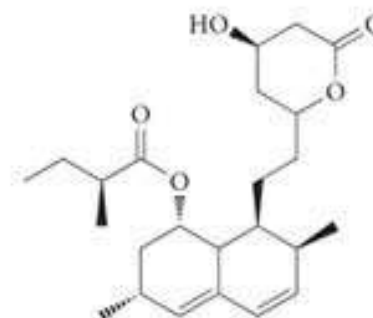
Aflatoxin B1



Ergotamine



Gliotoxin



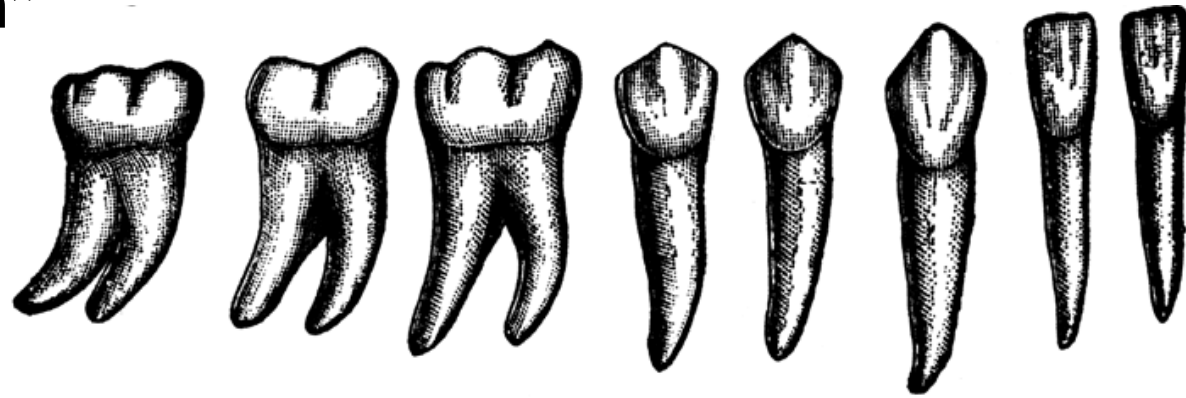
Lovastatin



Keller et al. (2005) Nature Rev. Microbiol.

Chemodiversity is Fundamental to the Fungal Lifestyle

- ❖ The genes involved in fungal primary metabolism are their “teeth”



- ❖ The genes involved in fungal secondary metabolism are their “horns”, “spines”, and “claws”



Beetle horns



Osprey claws



Echidna spines

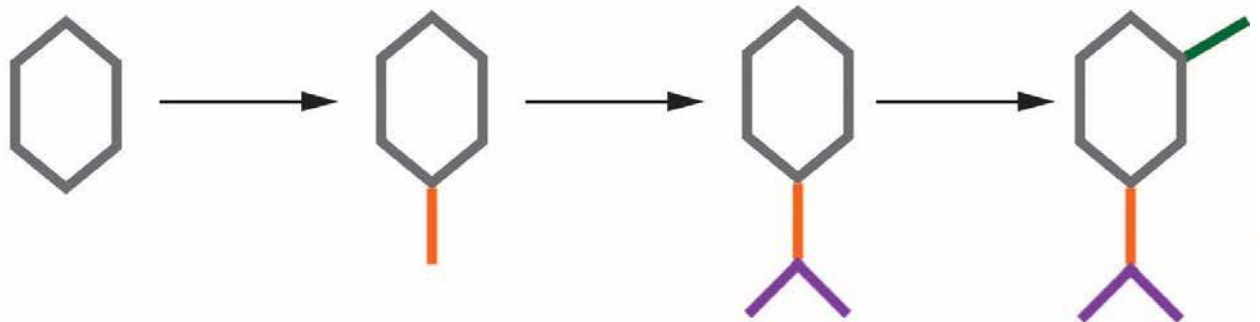
Fungal Metabolic Genes Are Often Physically Clustered



Enzymes



Metabolic pathway



The Evolution of Metabolic Gene Clusters in Fungi

How? (Slot & Rokas, 2010, *PNAS*)

- ❖ Gene clusters originate by native gene relocation
- ❖ Convergent evolution of gene clusters
- ❖ Gene clusters move by HGT

Why? (McGary et al., 2013, *PNAS*)

- ❖ Selection **for**: reducing impact of toxic intermediates
- ❖ Selection **of**: coordinating gene expression, genetic linkage

What?

- ❖ Clusters facilitate adaptation to diverse nutritional environments (Gibbons et al., 2012, *Curr. Biol.*, Gibbons et al., 2012, *Euk. Cell*)
- ❖ Clustering of fungal genes predicts tissue-specific coexpression of their human orthologs (Eidem et al., in revision)

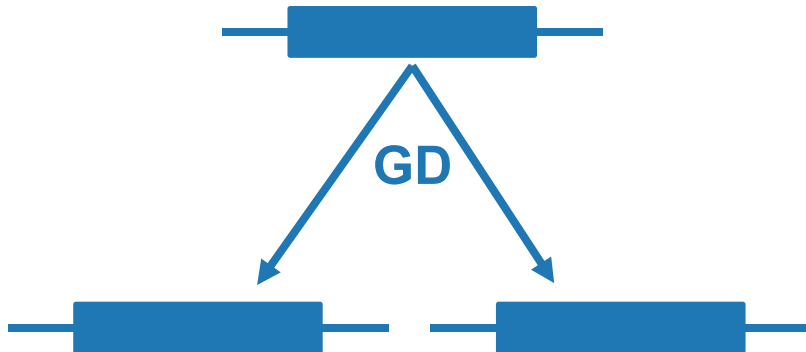
What contributed to the making of fungal chemodiversity?

- ❖ Genes and pathways**
- ❖ Regulatory networks**

Sources of Gene Innovation

Gene duplication (GD)

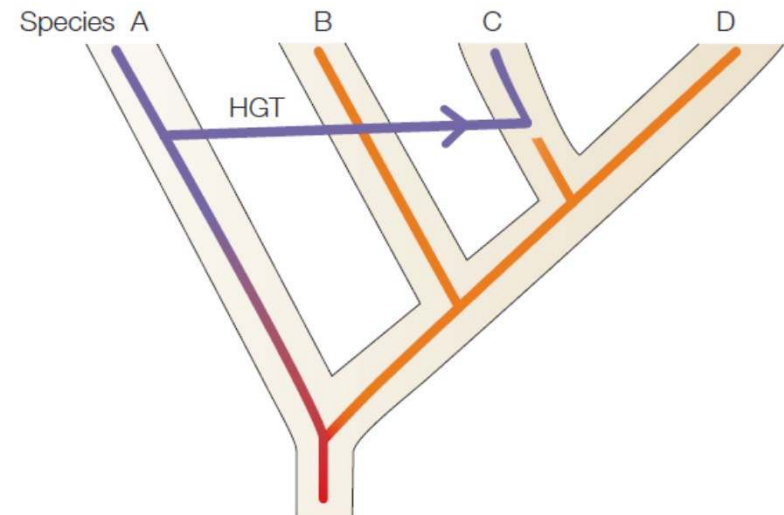
Any duplication of a region of DNA that contains a gene



- ❖ Plant organic material decay
- ❖ Starch catabolism
- ❖ Degradation of host tissues
- ❖ Toxin production

Horizontal gene transfer (HGT)

Exchange of genes between organisms other than through reproduction



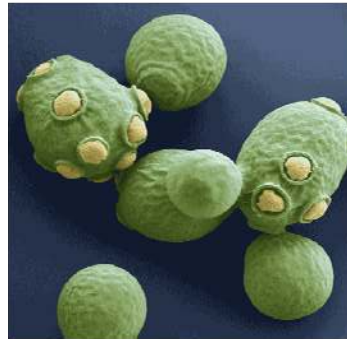
- ❖ Xenobiotic catabolism
- ❖ Toxin production
- ❖ Degradation of plant cell walls
- ❖ Wine fermentation

208 Genomes, 247,000 Genes, 875 Reactions, 130 Pathways

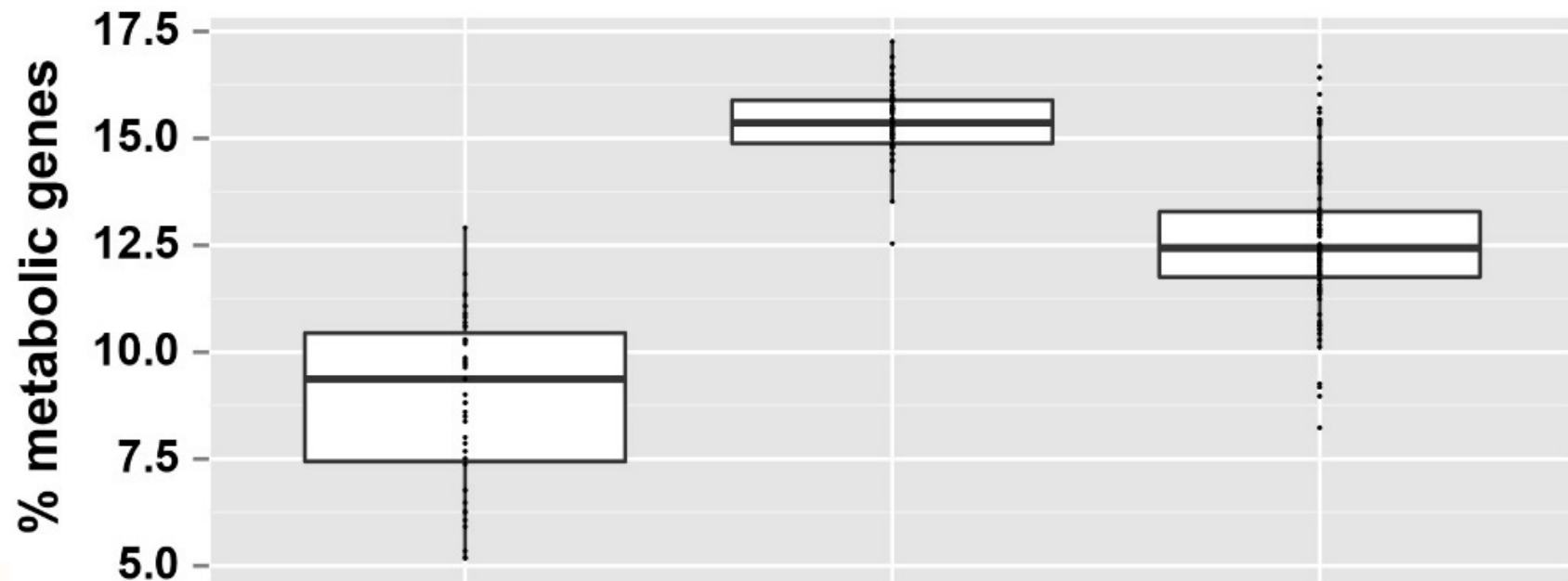
Agaricomycetes



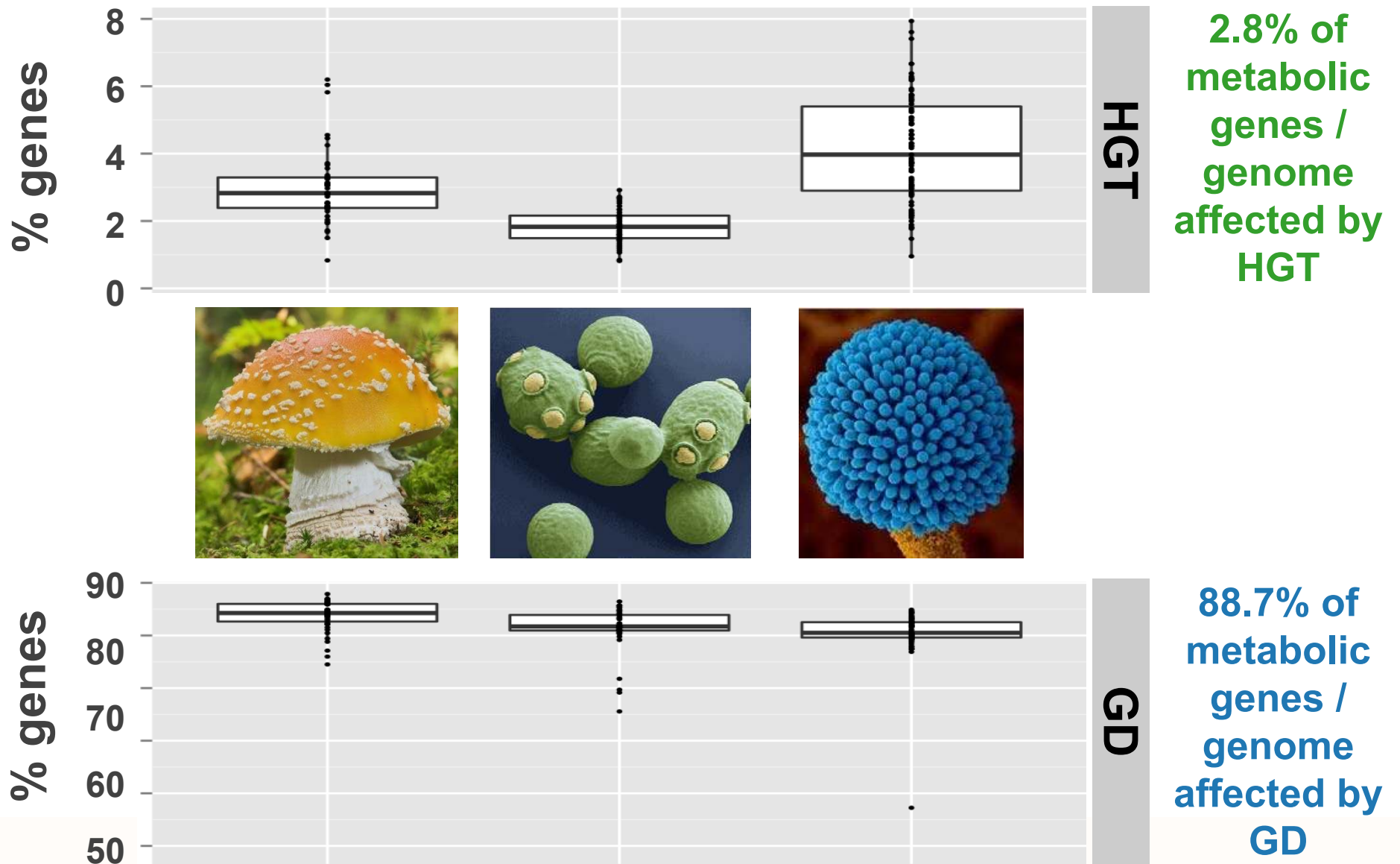
Saccharomycotina



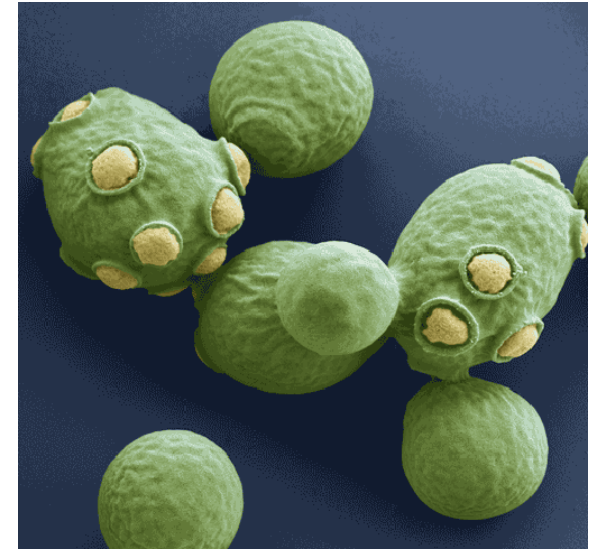
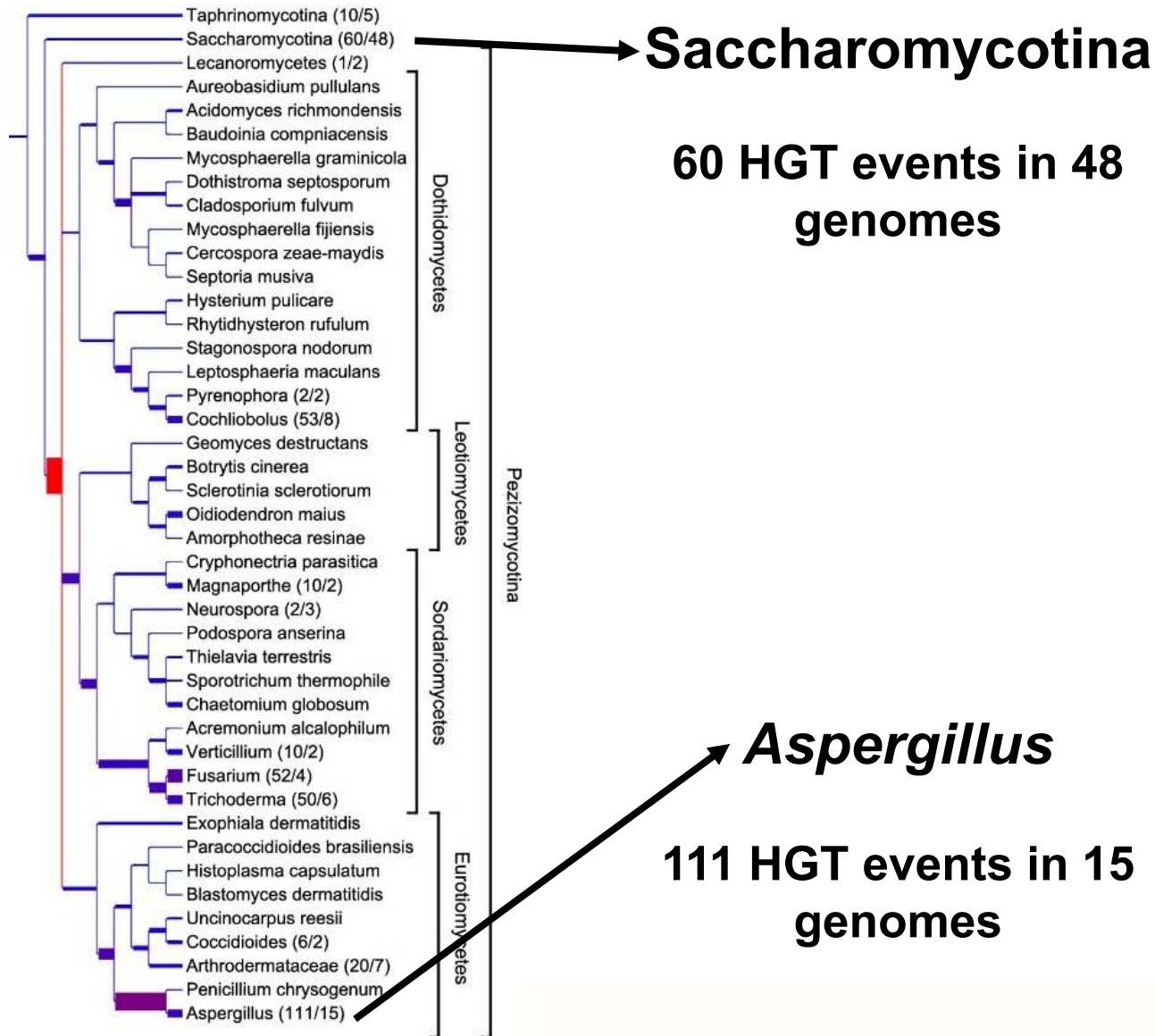
Pezizomycotina



The Impact of GD is Large but of HGT Small



HGT is Episodic and Acts in a Lineage-Specific Manner

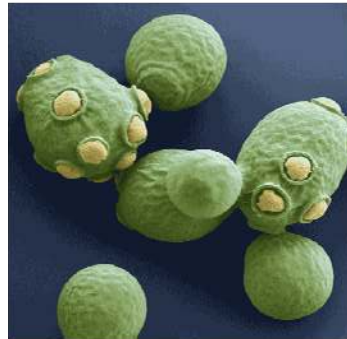


Lineages Vary With Respect to Gene Clustering

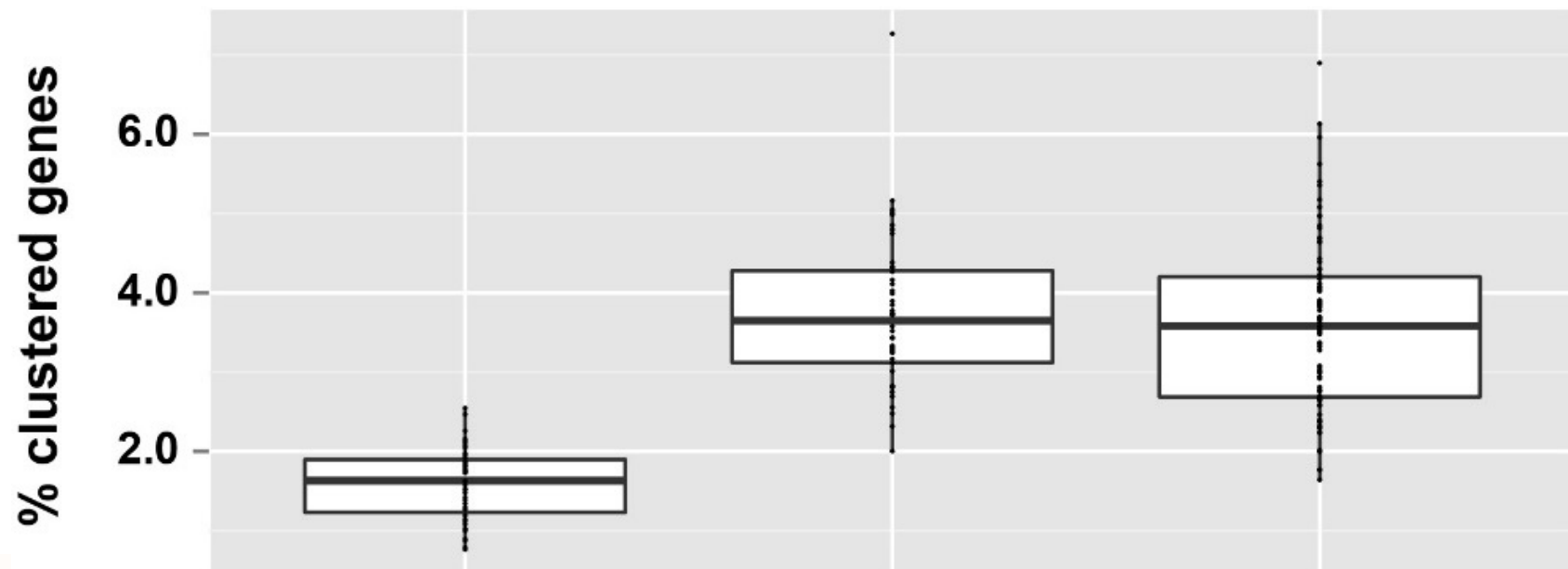
Agaricomycetes



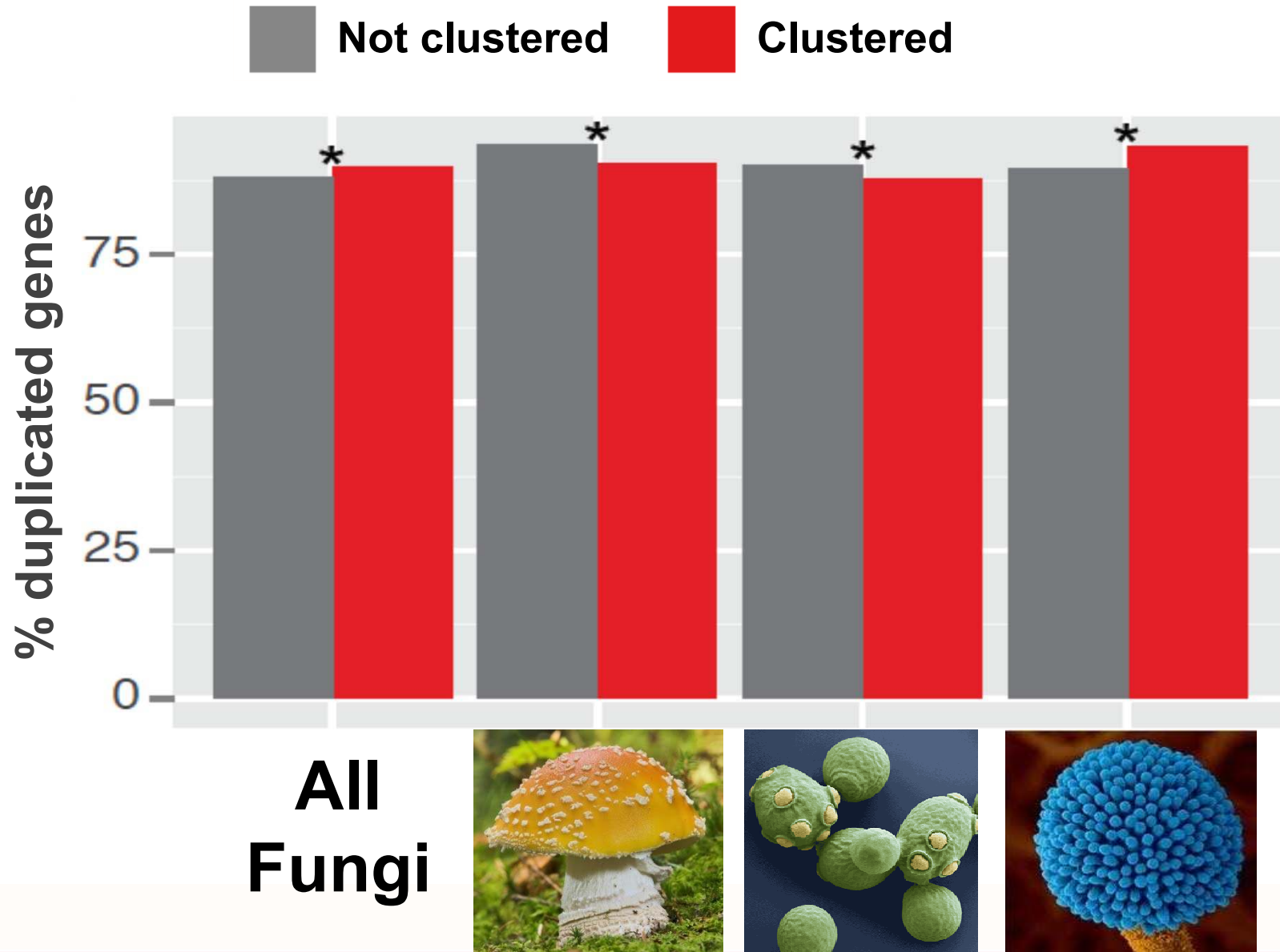
Saccharomycotina



Pezizomycotina

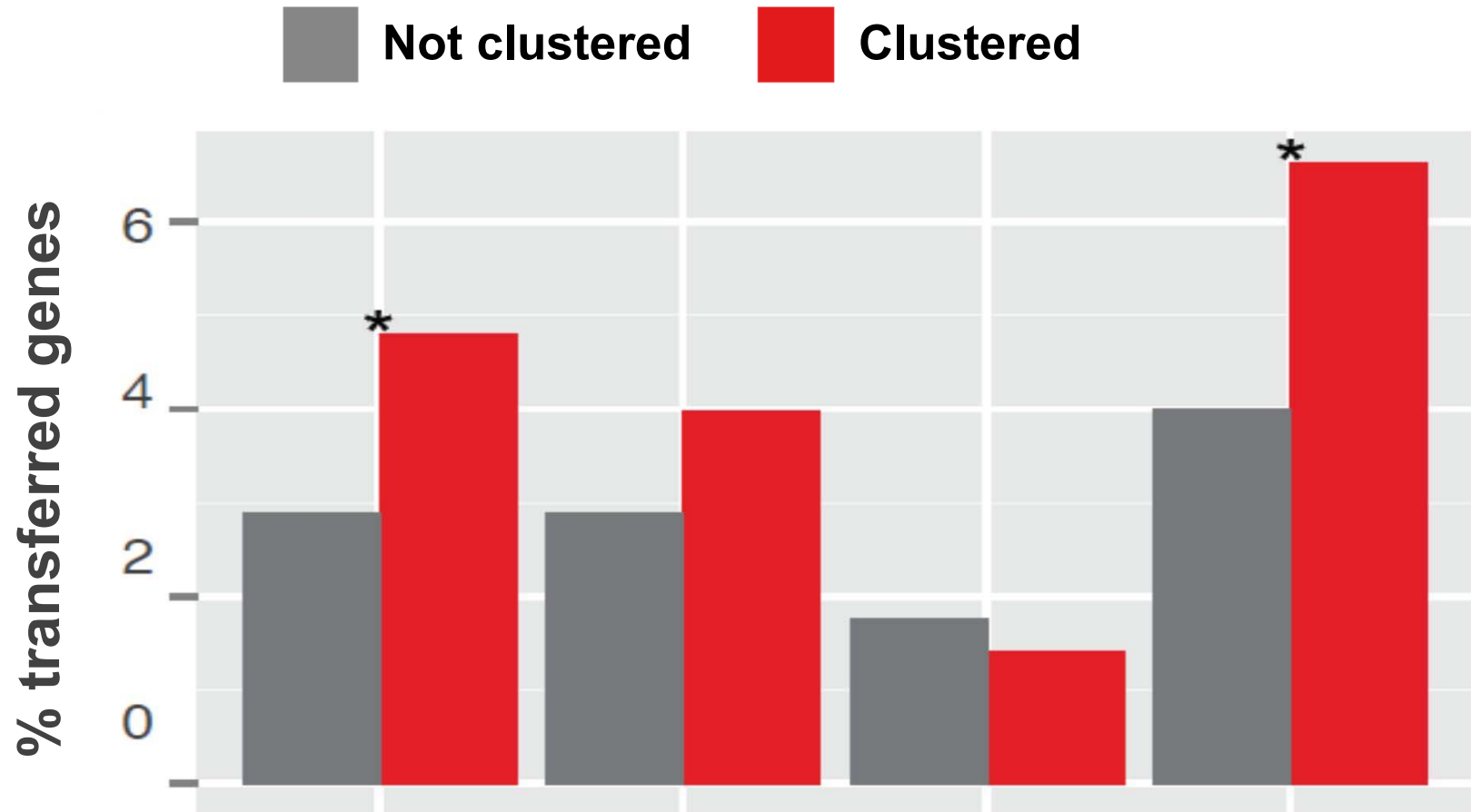


Does Clustering Correlate with GD?

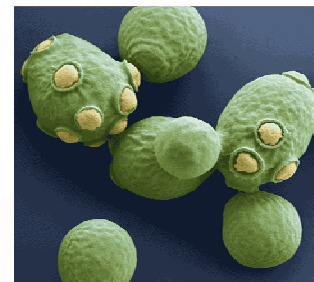


Wisecaver et al. (2014) PLoS Genet.

Does Clustering Correlate with HGT?

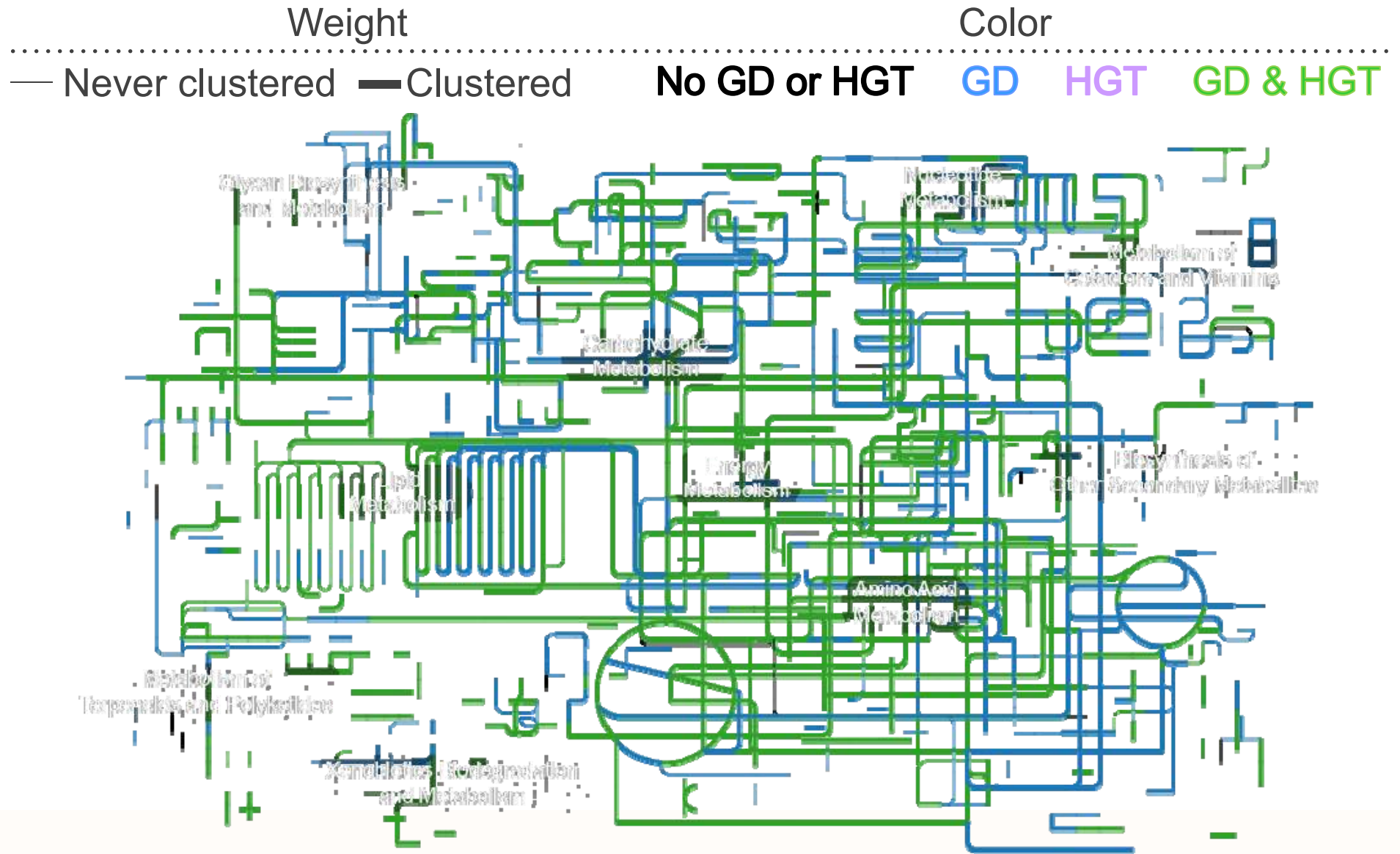


**All
Fungi**

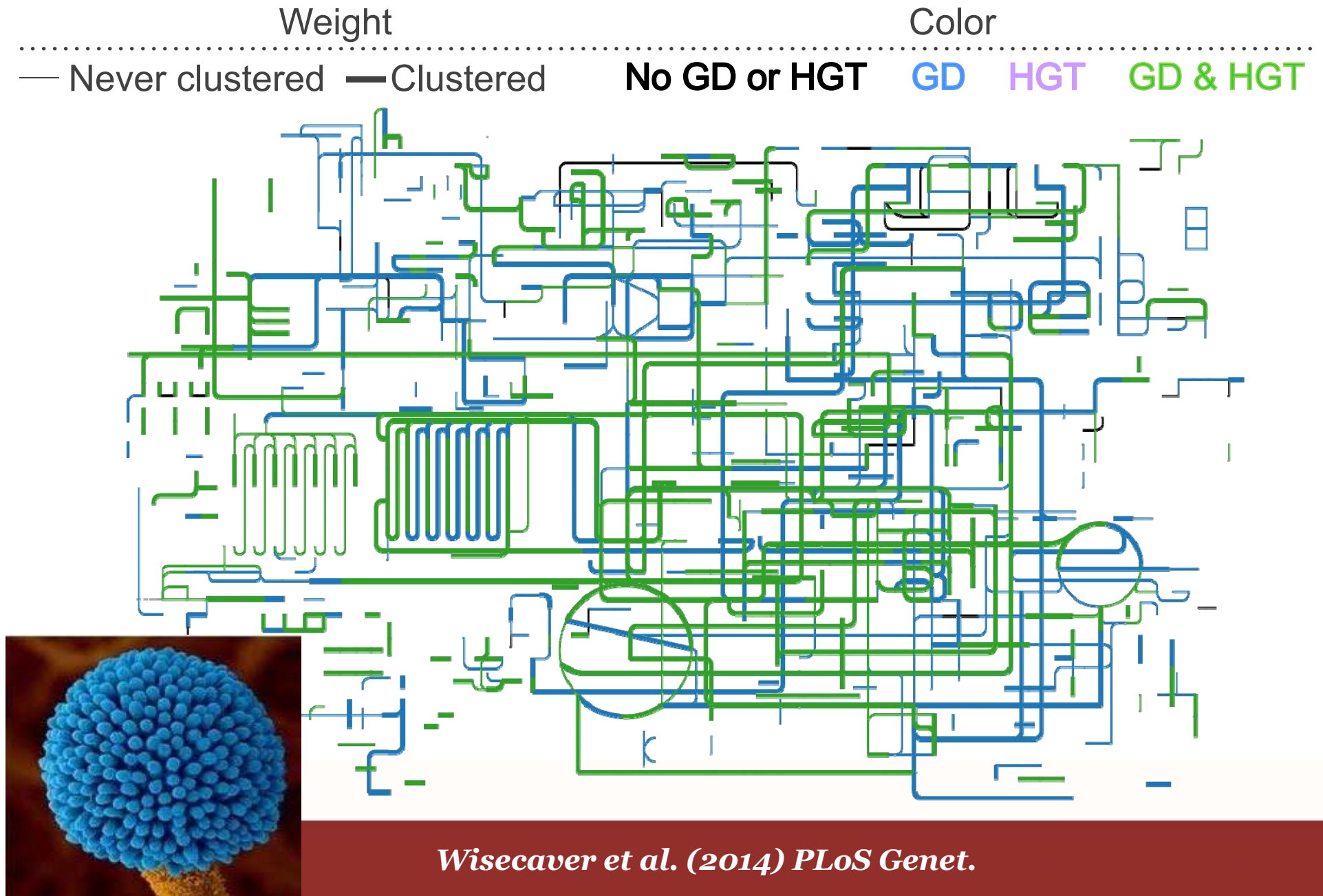


Wisecaver et al. (2014) PLoS Genet.

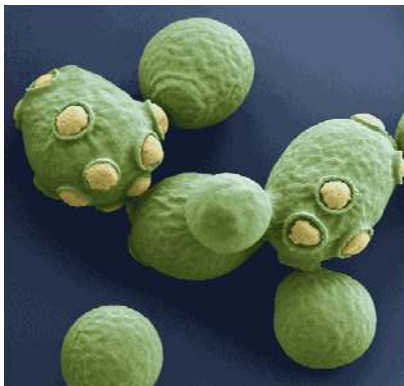
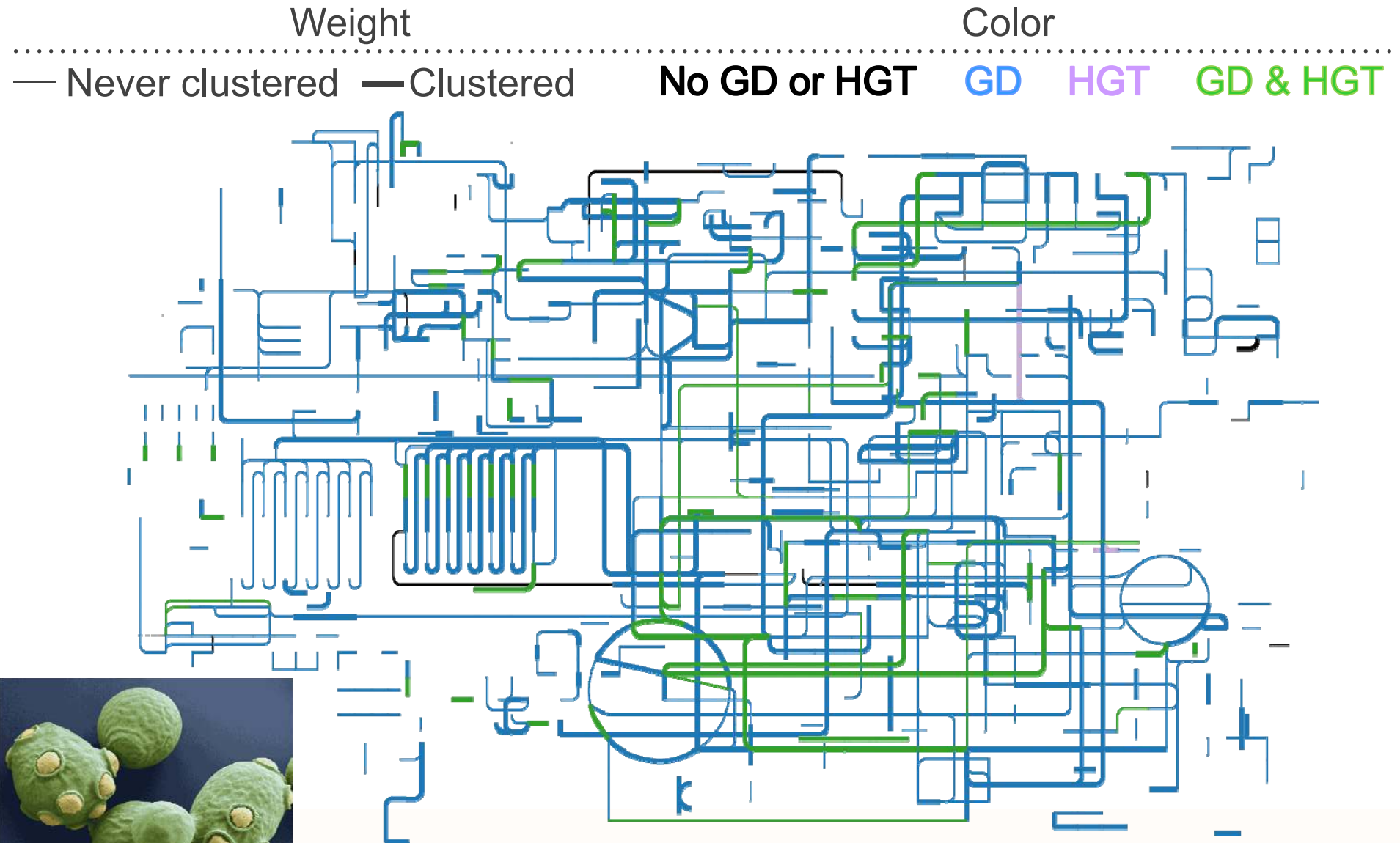
Gene Clustering, GD and HGT Across Fungi



Filamentous Fungi: Clustered Genes Undergo GD and HGT

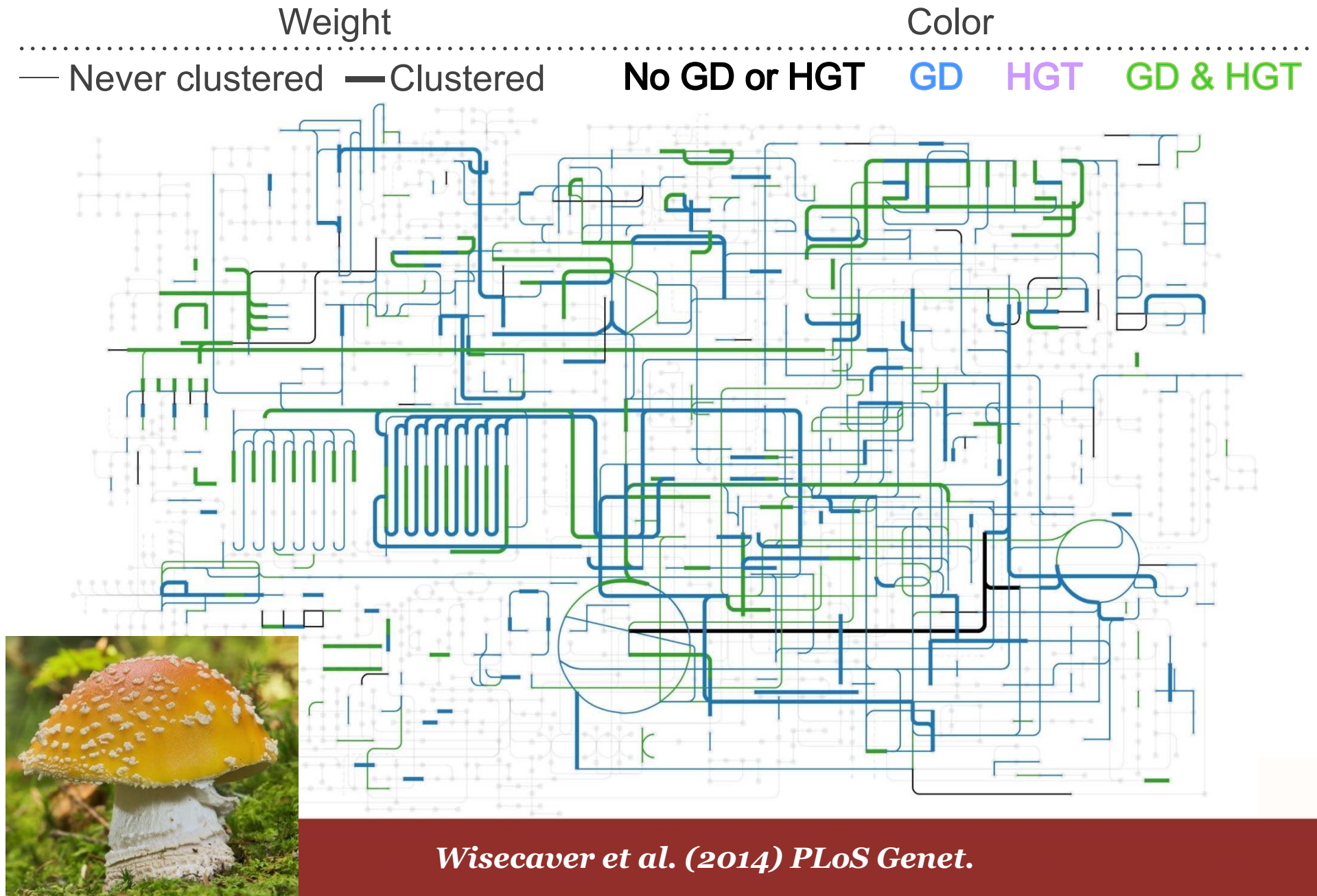


Yeasts: Genes are Clustered but Rarely Undergo HGT

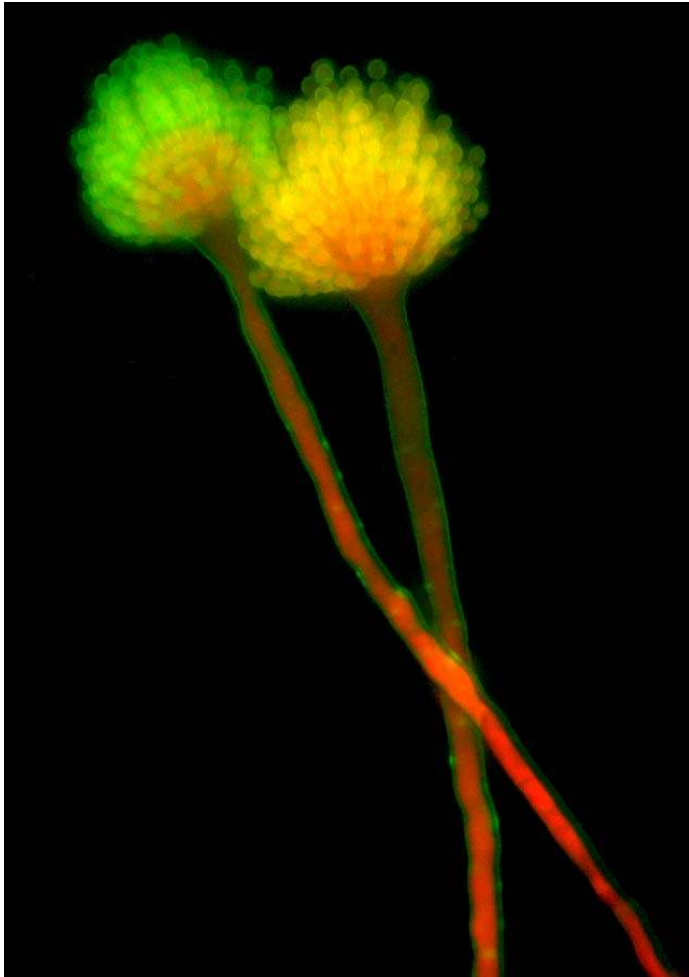


Wisecaver et al. (2014) PLoS Genet.

Basidios: Genes are not often Clustered or Transferred



***Aspergillus oryzae*: Cornerstone of Several Japanese Tasty Liquids**



Archaeological evidence suggests that mixed fermented alcoholic beverage of rice, honey and fruit was made in China as early as 7 – 9 millennia ago

***Aspergillus oryzae*, a filamentous fungus, is involved in the production of sake (rice wine), *miso* (soy bean paste), *su* (vinegar) and *shoyu* (soy sauce)**



Aspergillus oryzae, A Domesticated Microbe

A. flavus



- ❖ Agricultural pest
- ❖ Aflatoxin producer
- ❖ ~\$1 billion annually

DOMESTICATION RD

ONE WAY

A. oryzae



- ❖ Sake rice wine
- ❖ Non-aflatoxin producer
- ❖ USDA GRAS species

The *A. oryzae* and *A. flavus* genomes are nearly identical

Studying Microbial Domestication using Omics

A. oryzae

SRRC 302	Sake
RIB 331	Miso
RIB 333	Miso
RIB 537	Sake
RIB 632	Sake
RIB 642	Sake
RIB 949	Soy Sauce
RIB 40	Sake, Reference Strain

A. flavus

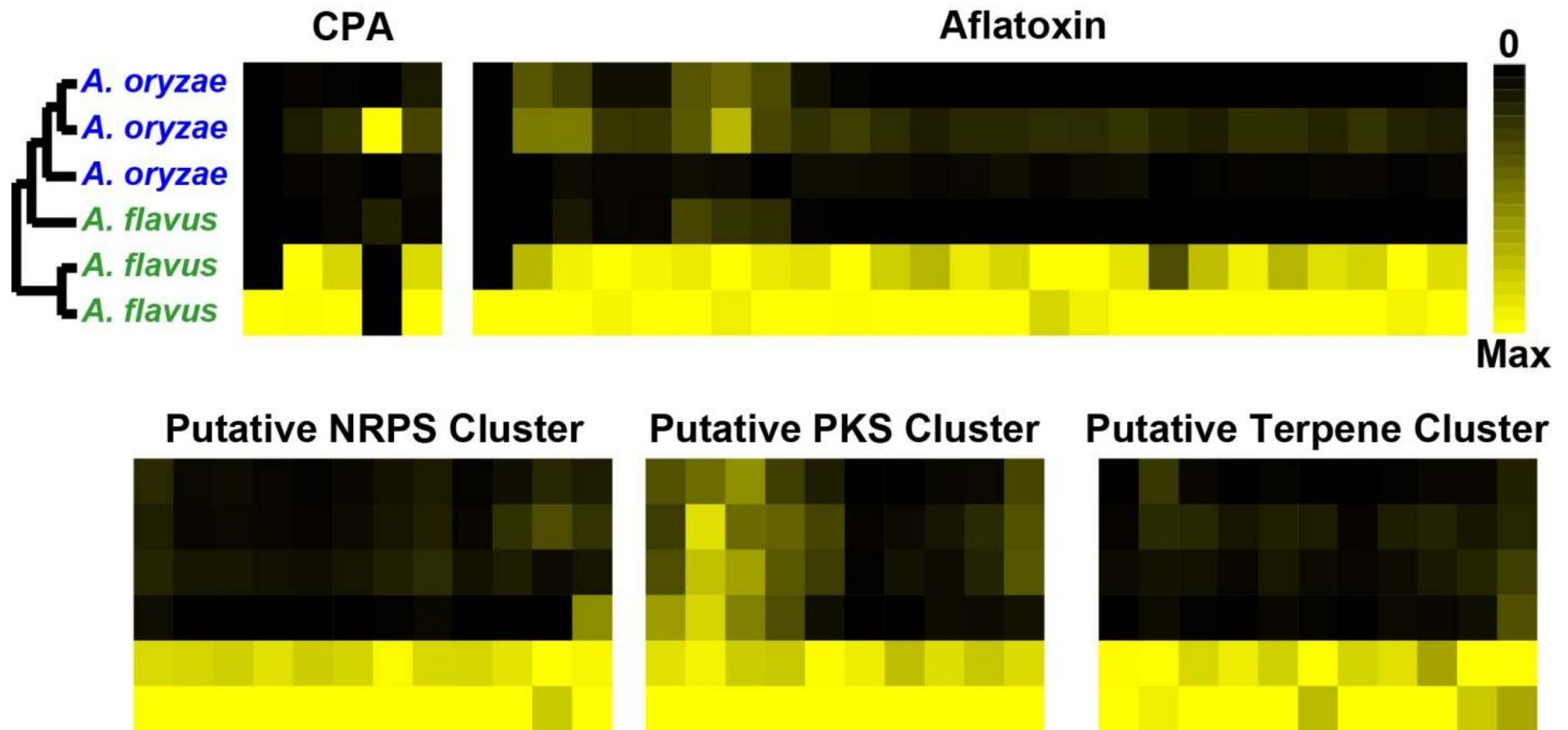
SRRC 1273	Soil, Arizona
SRRC 1357	Dried bacon, Croatia
SRRC 2112	Hazelnut, Turkey
SRRC 2114	Wheat, USA
SRRC 2524	Dead termites, China
SRRC 2632	Blood, Chicago, Illinois
SRRC 2653	Corneal ulcer, Miami, Florida
NRRL 3357	Peanut, Reference strain, USA

❖ Genome resequencing (14 isolates)

❖ RNA-Seq (6 isolates)

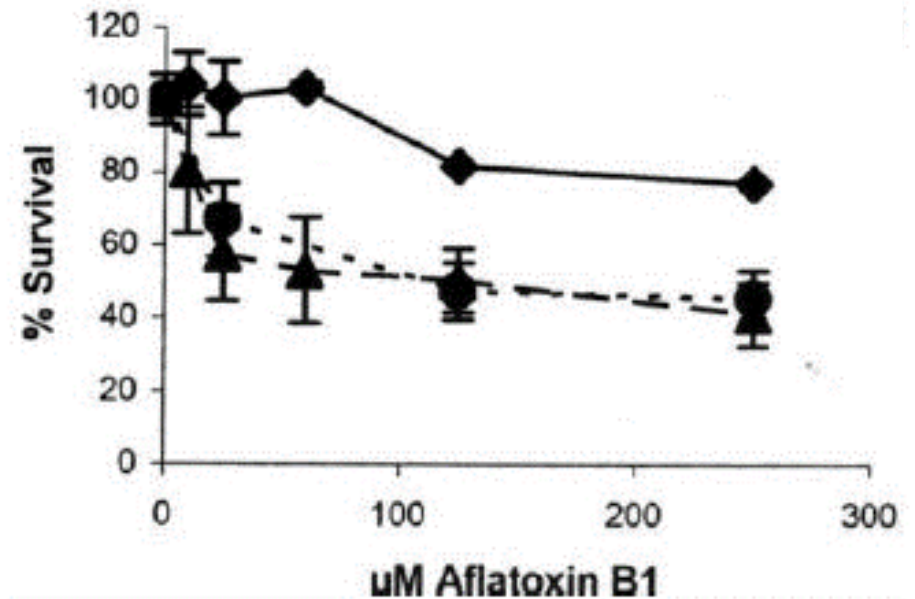
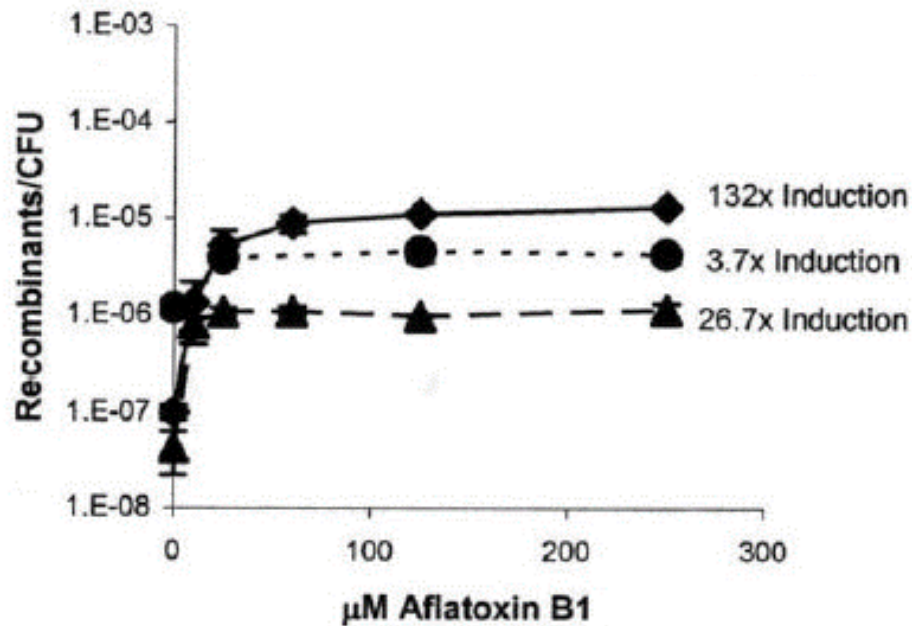
❖ Proteomics (2 isolates)

Down-Regulation of Secondary Metabolism in *A. oryzae*



Why is A. oryzae Atoxic?

Aflatoxin is genotoxic to *S. cerevisiae*



The atoxicity of *A. oryzae* might have been driven by its impact on yeast survival and, as a consequence, fermentation for making sake



Lecture Outline

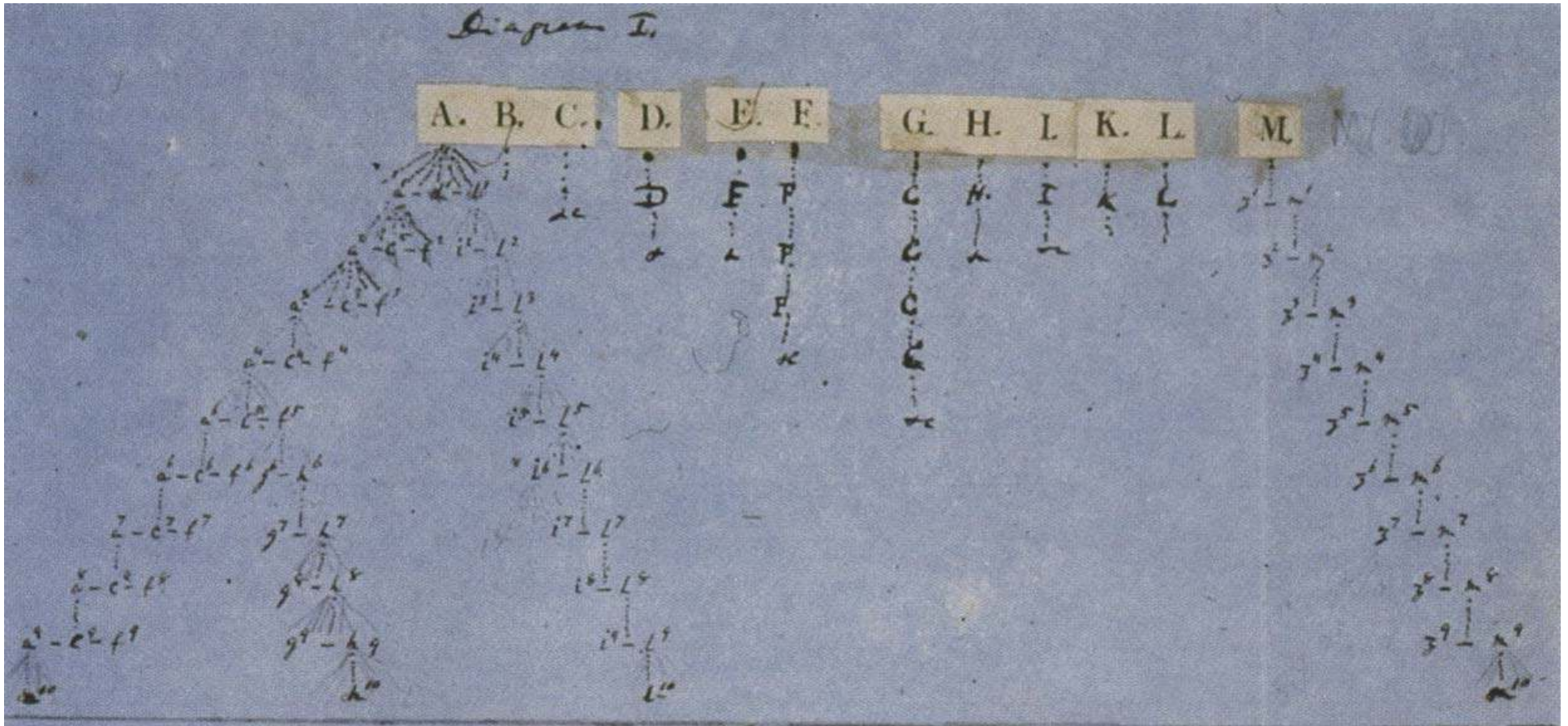
❖ **Introduction to Evolutionary Genomics**

❖ **Evolutionary and Functional Genomics**

----- **Coffee Break** -----

❖ **Phylogenomics**

Darwin's Tree



Darwin's hand-made proof of the famous diagram from his *Origin of Species*



Maderspacher (2006) *Curr. Biol.*

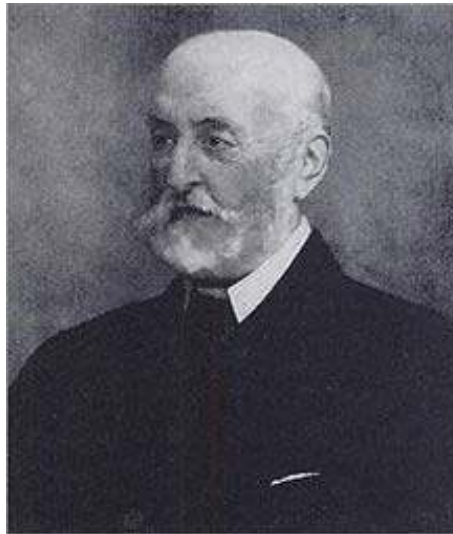
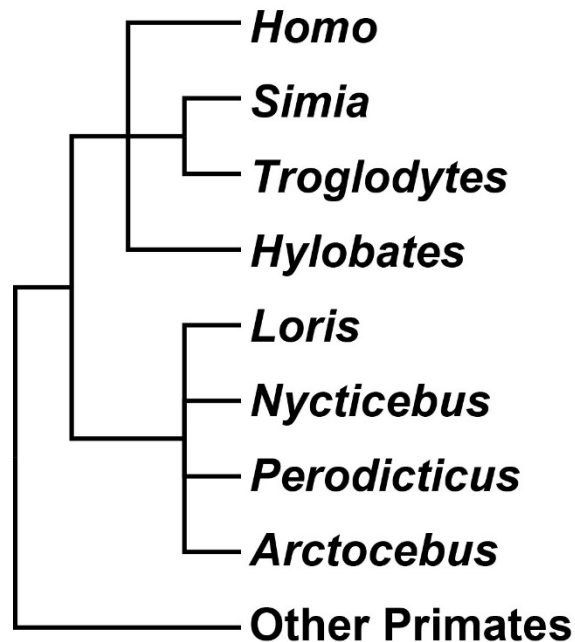
and instinct as the summing up of many contrivances, each useful to the possessor, nearly in the same way as when we look at any great mechanical invention as the summing up of the labour, the experience, the reason, and even the blunders of numerous workmen; when we thus view each organic being, how far more interesting, I speak from experience, will the study of natural history become!

A grand and almost untrodden field of inquiry will be opened, on the causes and laws of variation, on correlation of growth, on the effects of use and disuse, on the direct action of external conditions, and so forth. The study of domestic productions will rise immensely in value. A new variety raised by man will be a far more important and interesting subject for study than one more species added to the infinitude of already recorded species. Our classifications will come to be, as far as they can be so made, genealogies; and will then truly give what may be called the plan of creation. The rules for classifying will no doubt become simpler when we have a definite object in view. We possess no pedigrees or armorial bearings; and we have to discover and trace the many diverging lines of descent in our natural genealogies, by characters of any kind which have long been inherited. Rudimentary organs will speak infallibly with respect to the nature of long-lost structures. Species and groups of species, which are called aberrant, and which may fancifully be called living fossils, will aid us in forming a picture of the ancient forms of life. Embryology will reveal to us the structure, in some degree obscured, of the prototypes of each great class.

When we can feel assured that all the individuals of the same species, and all the closely allied species of most genera, have within a not very remote period de-

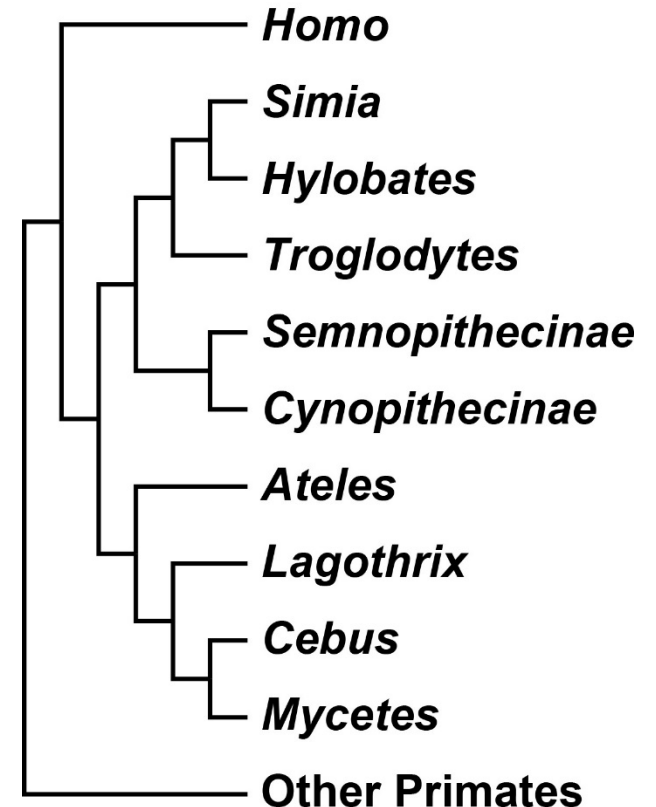
The Problem of Incongruence

1865: SPINAL COLUMN



St. George
Jackson Mivart

1867: LIMPS

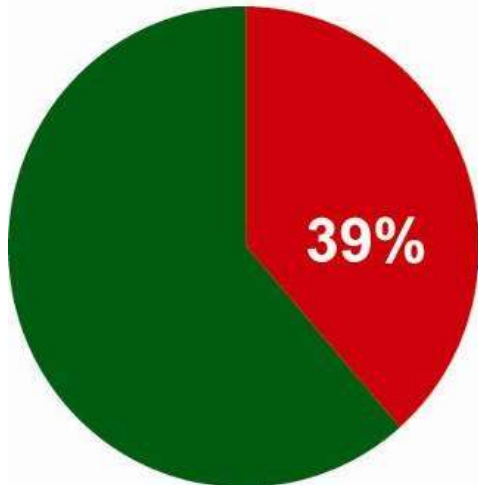


“The [1865] diagram [...] expresses what I believe to be the degree of resemblance as regards the spinal column *only*. The [1867] diagram expresses what I believe to be the degree of resemblance as regards the appendicular skeleton *only*”

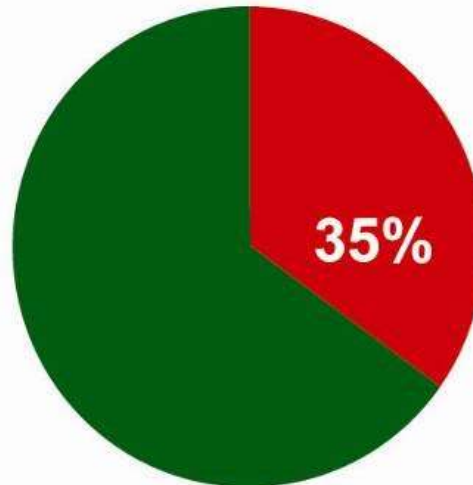
Darwin Correspondence Project letter 7170

The Problem of Incongruence

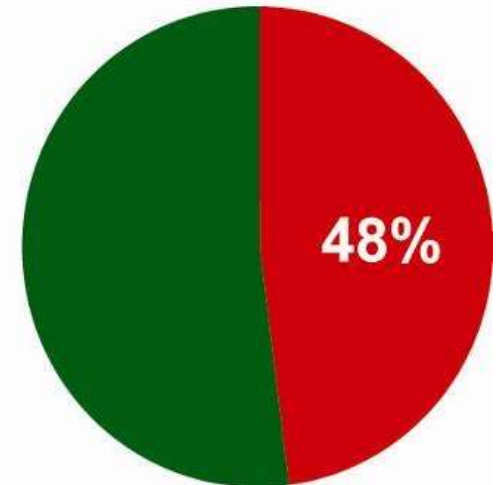
A: All organisms



B: Mammals



C: Insects

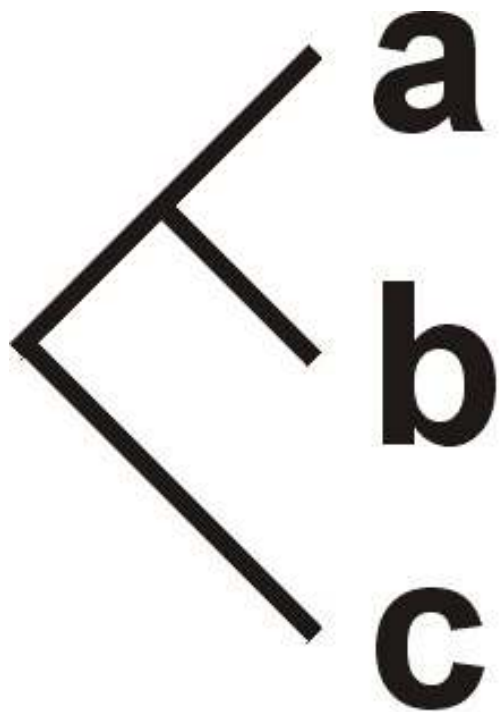


Incongruence is pervasive in the phylogenetics literature



Rokas & Chatzimanolis (2008) in Phylogenomics (W. J. Murphy, Ed.)

The Problem of Incongruence



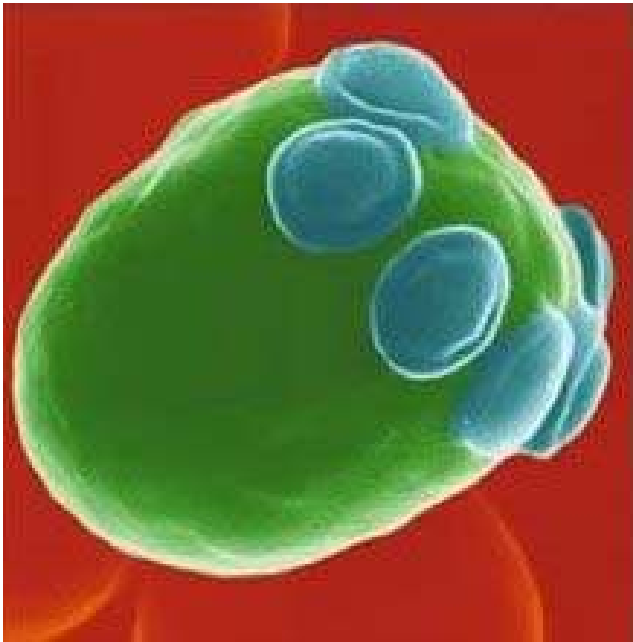
Gene X



Gene Y

**Species
tree?**

A Systematic Evaluation of Single Gene Phylogenies



S. cerevisiae

S. paradoxus

S. mikatae

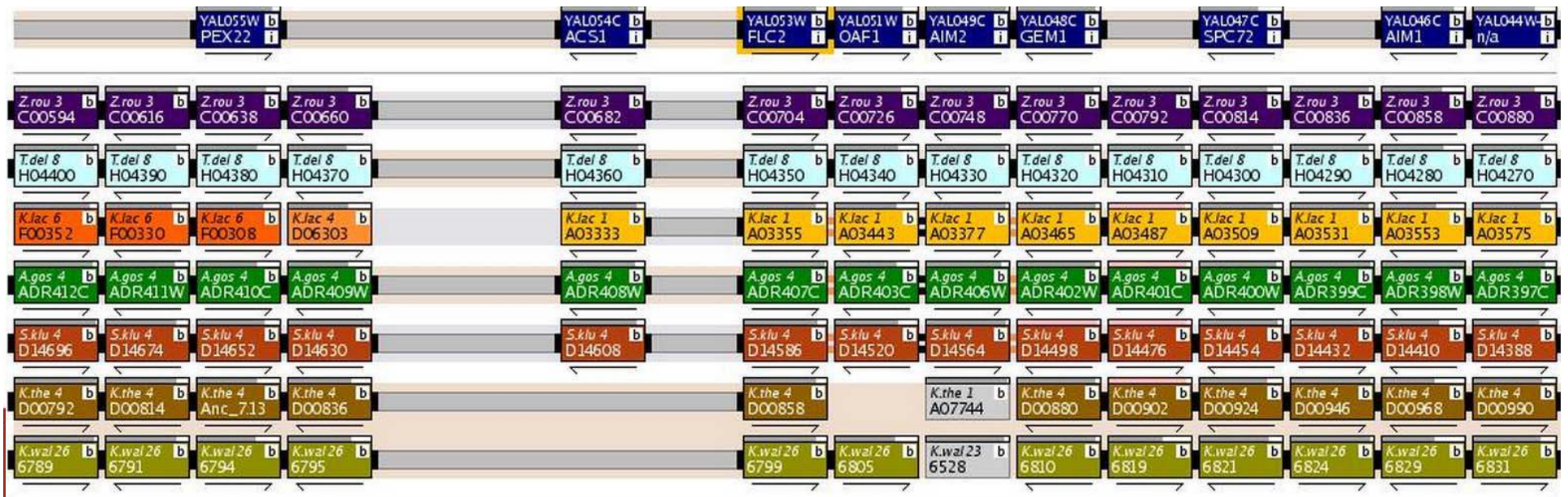
S. kudriavzevii

S. bayanus

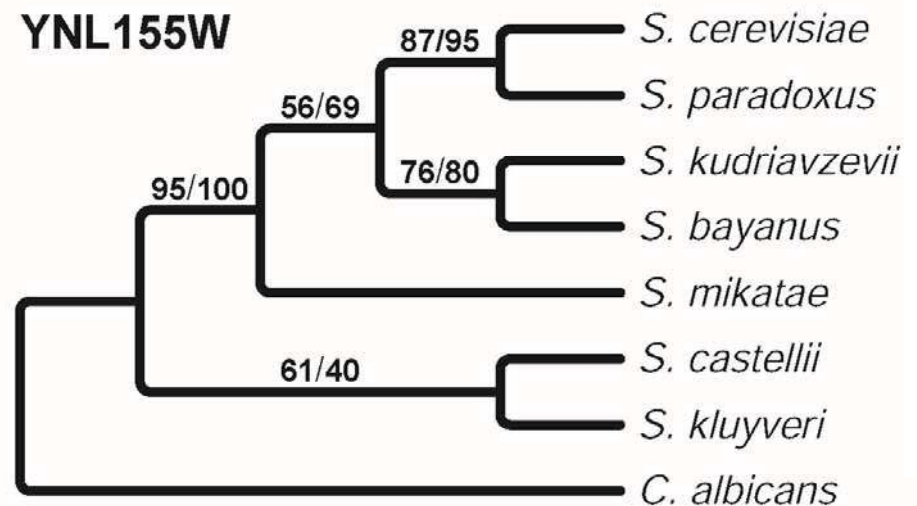
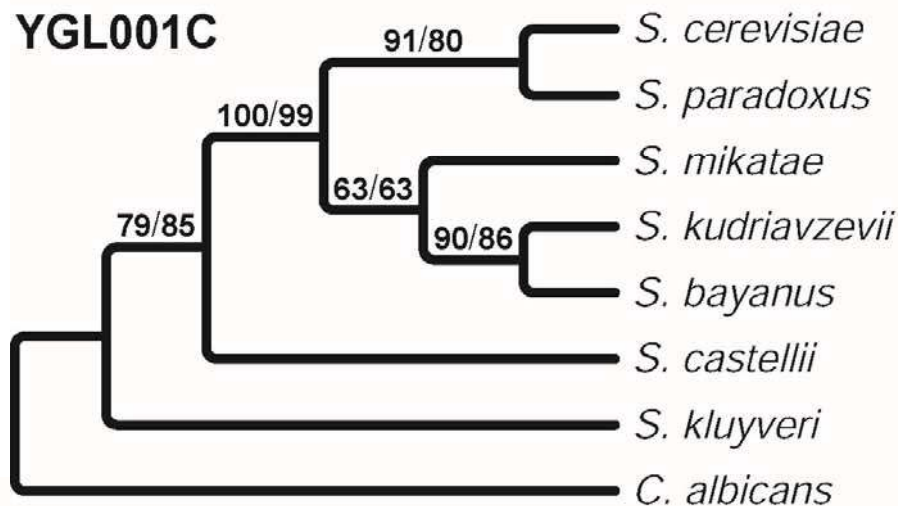
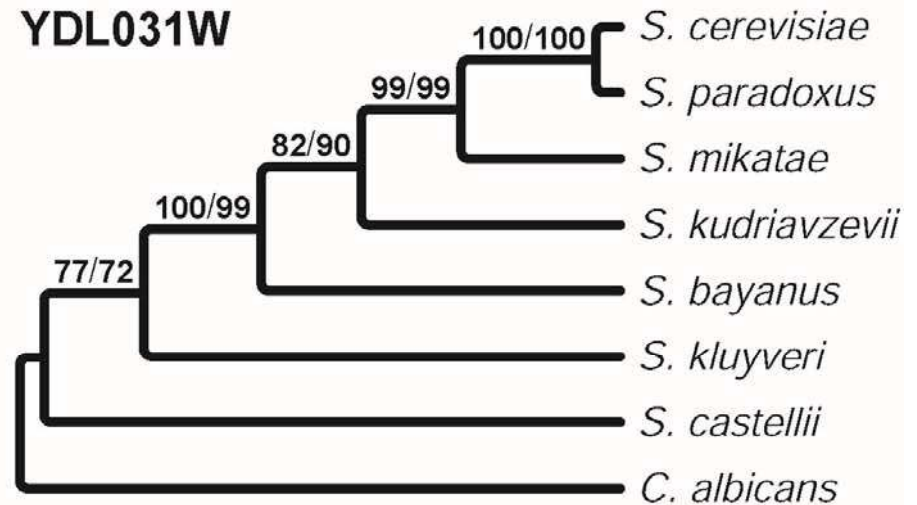
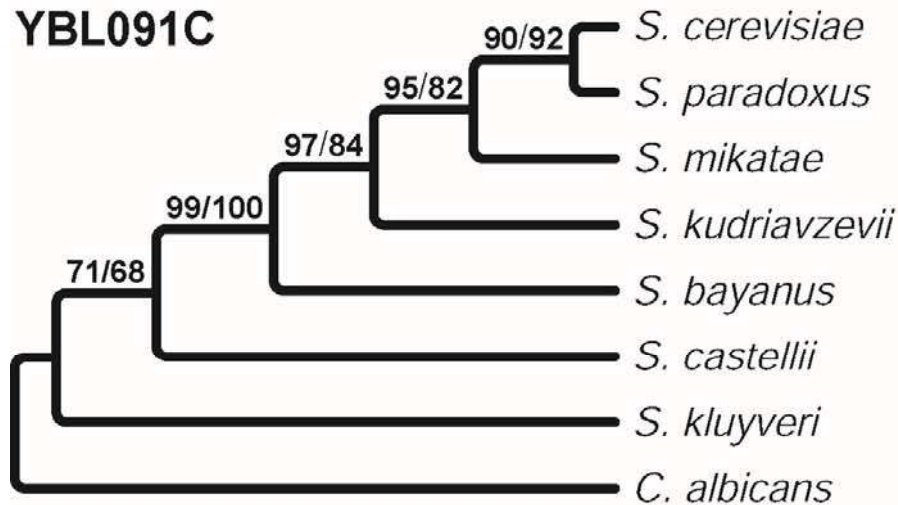
S. castellii

S. kluyveri

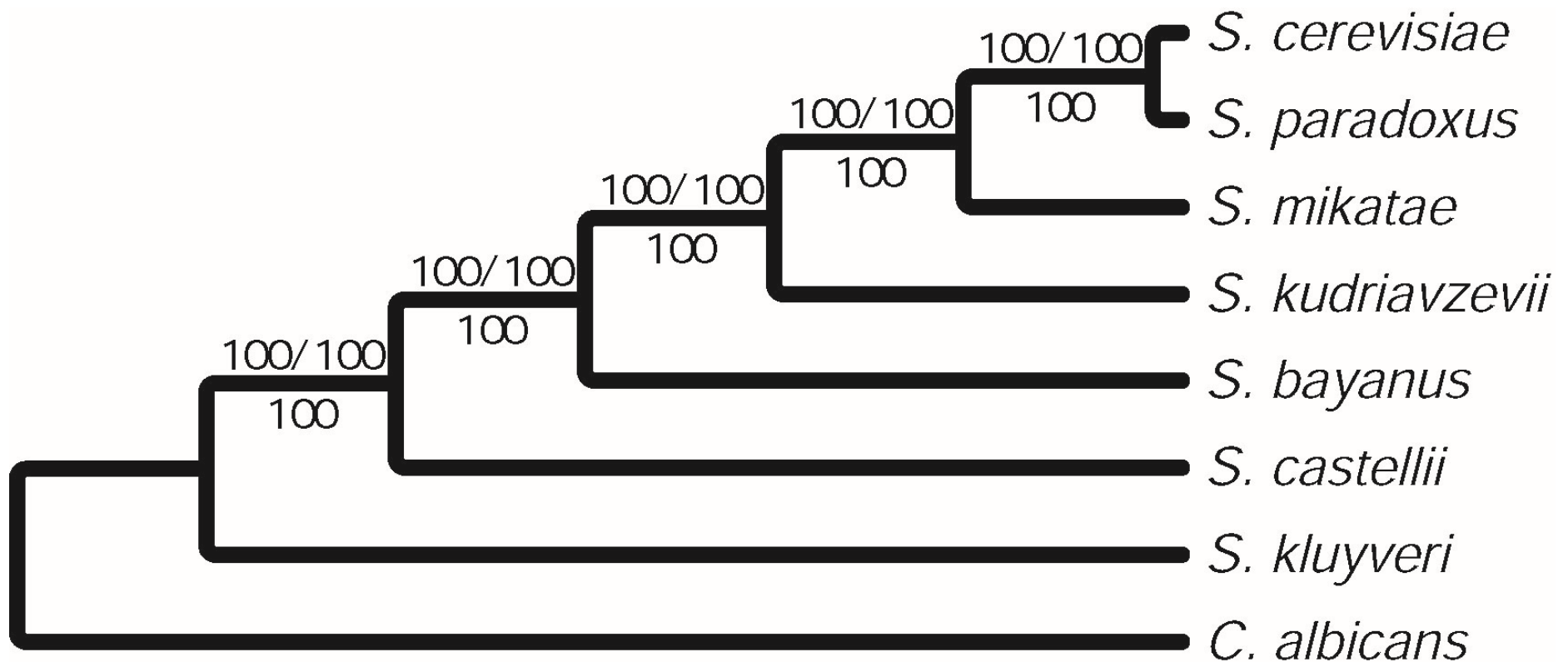
Candida albicans



Incongruence at the Single Gene Level



Concatenation of 106 Genes Yields a Single Yeast Phylogeny

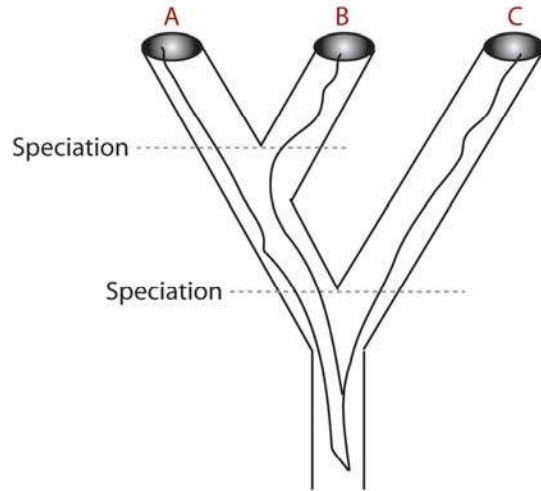


Clade support values = $\frac{\text{ML / MP on nucleotide alignments}}{\text{MP on amino acid alignments}}$

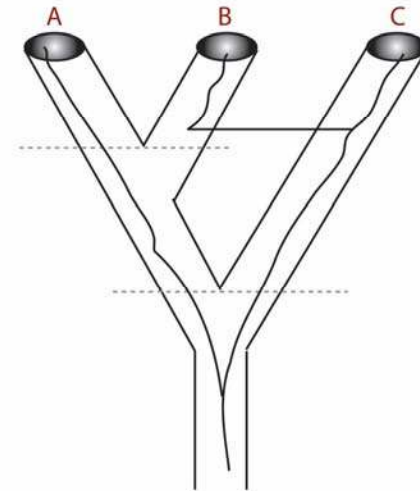


Gene Trees Can Differ from Species Trees

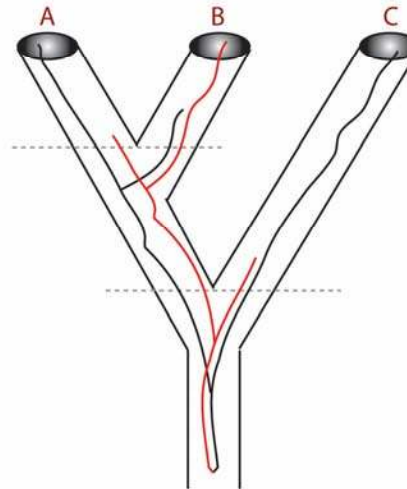
Lineage Sorting



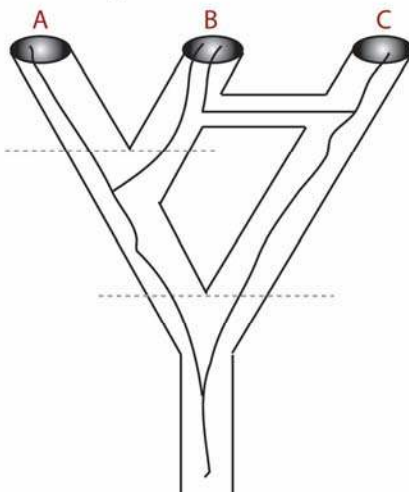
Horizontal Gene Transfer



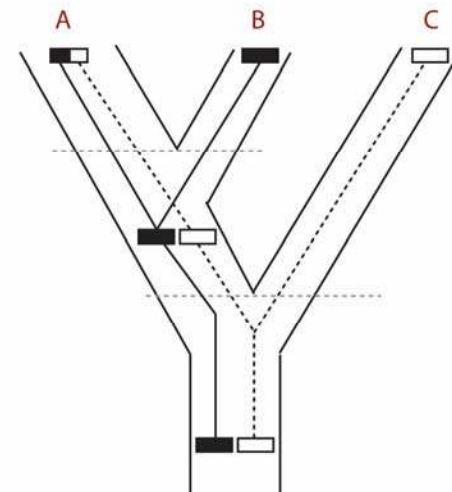
Gene Duplication and Loss



Hybridization

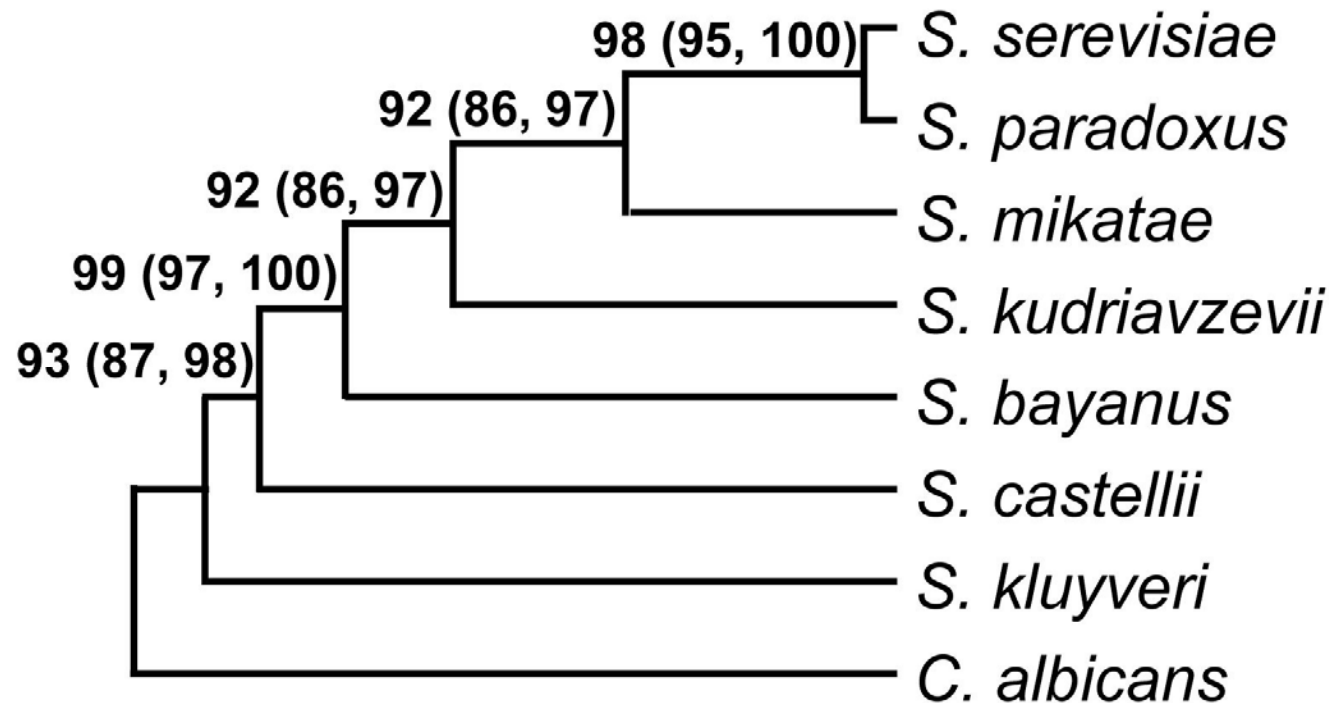


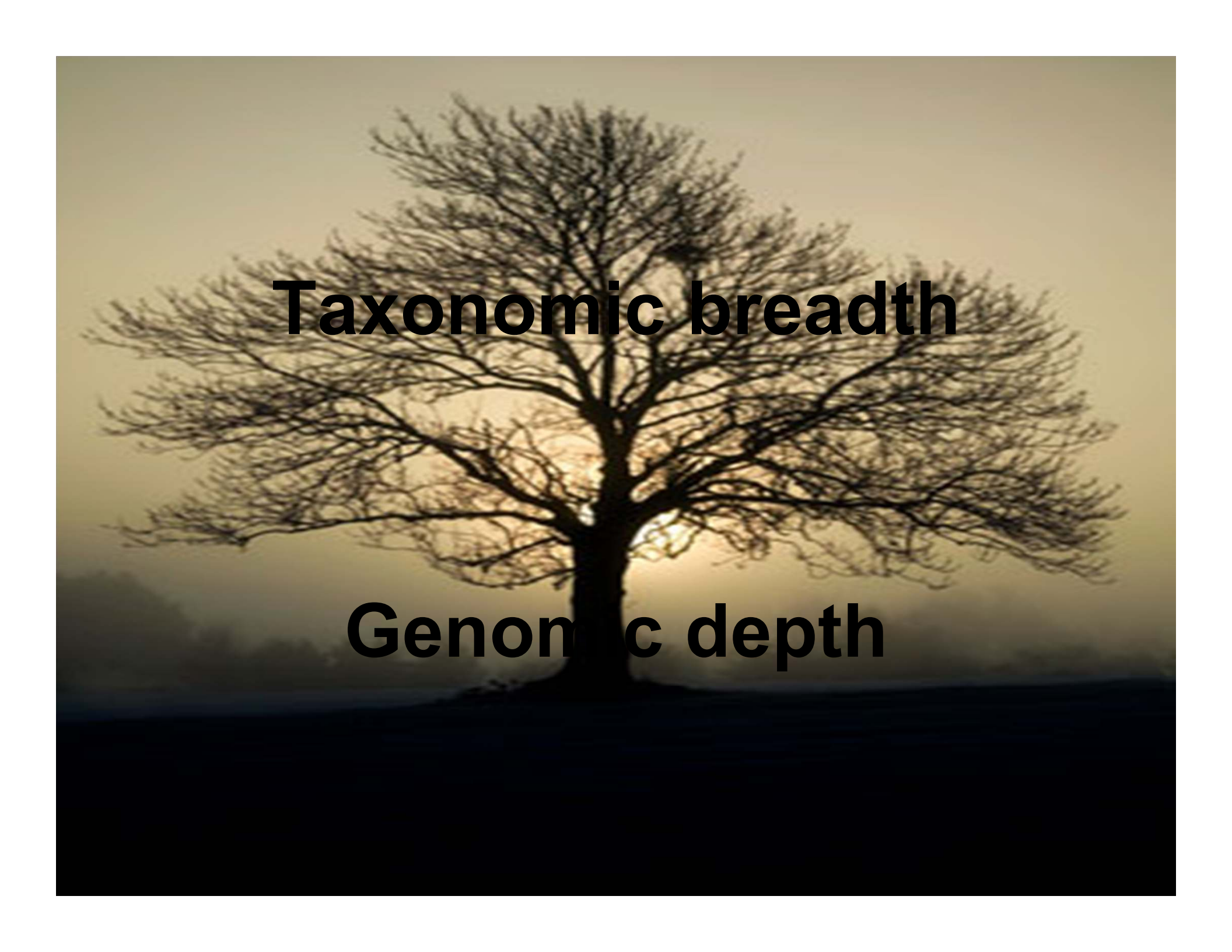
Recombination



Inferring the Species Tree from Individual Gene Histories

Concordance Factor: The proportion of the genome for which a clade is true

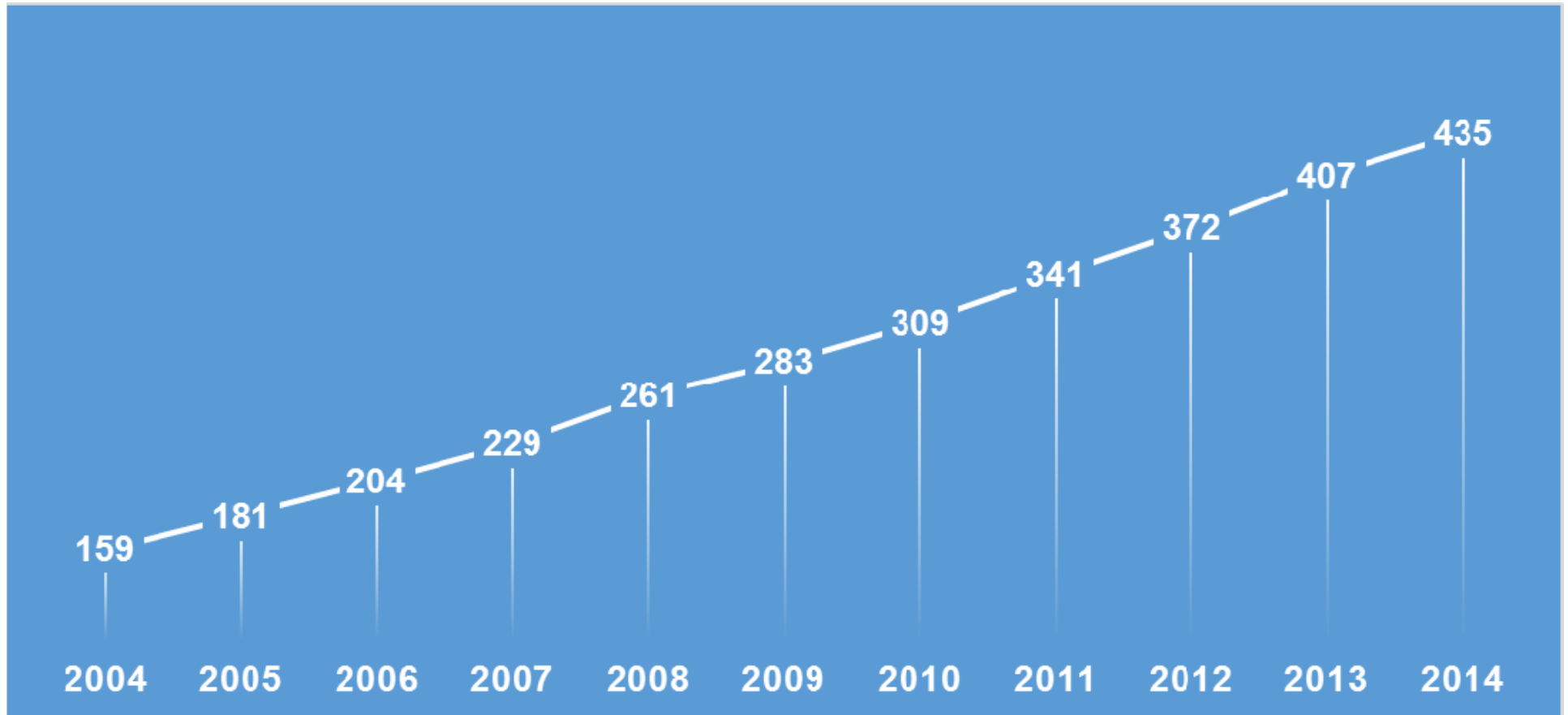




Taxonomic breadth

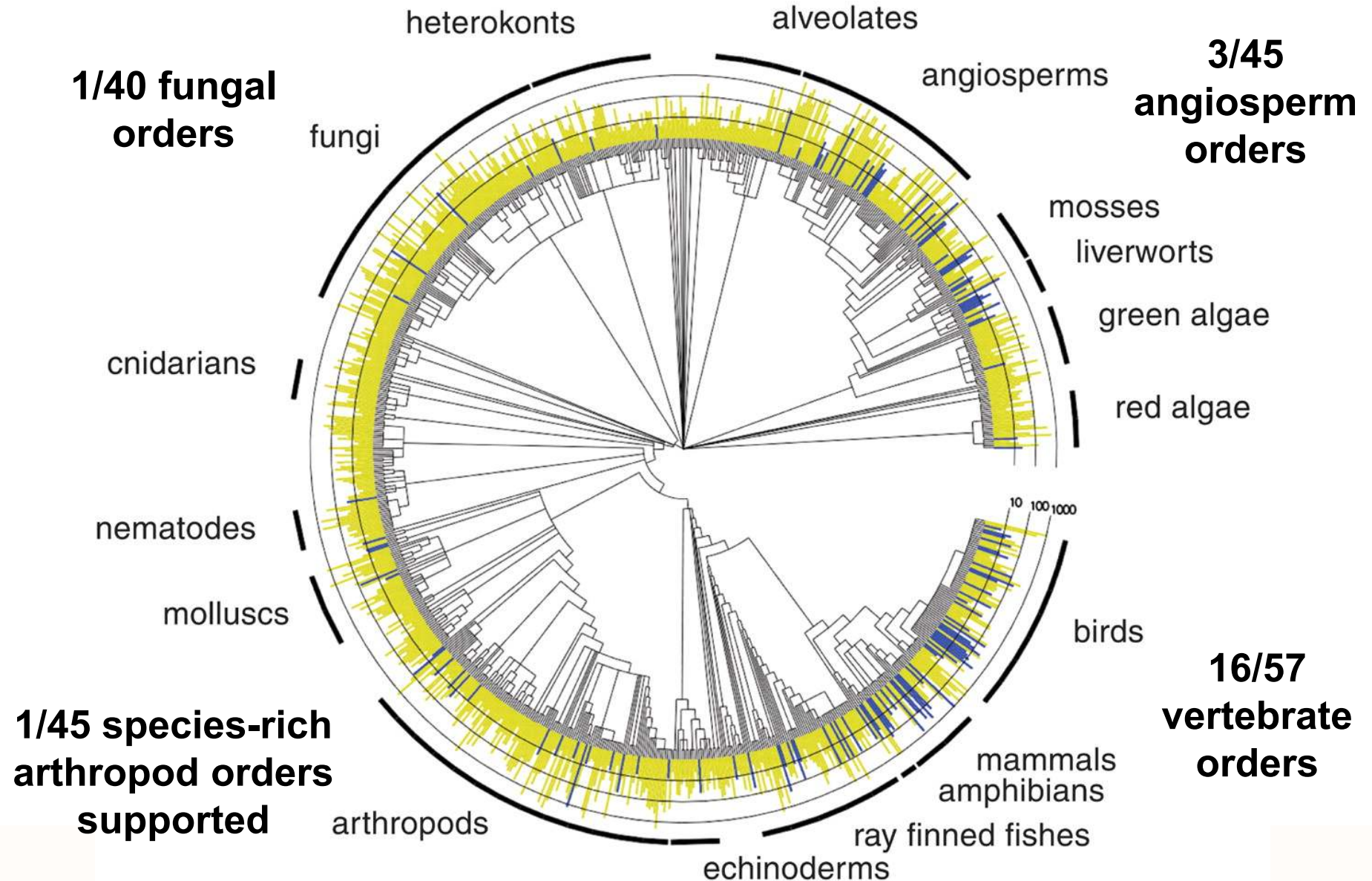
Genomic depth

Estimating the Taxonomic Breadth of the Tree of Life

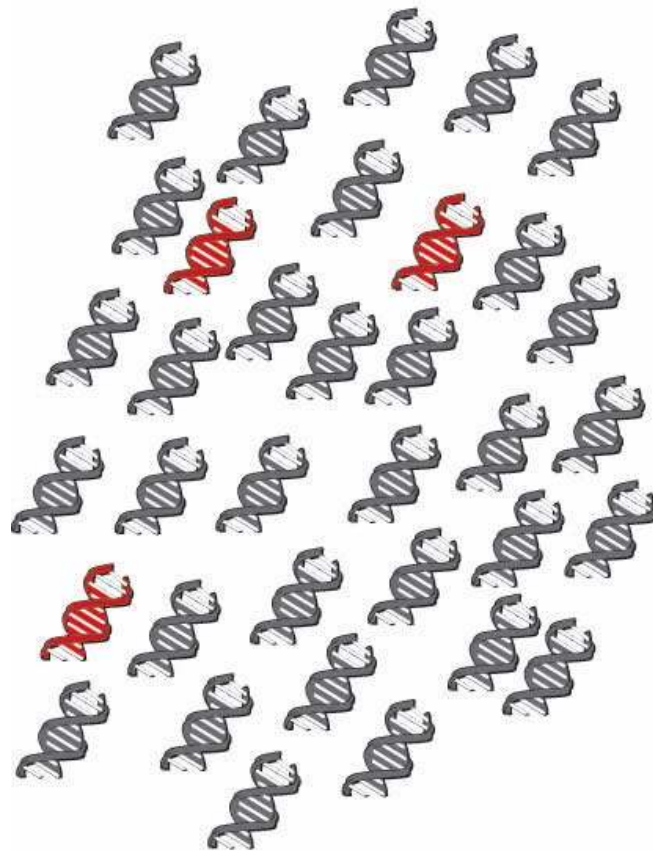


Number of new taxa added (x 1,000) in the GenBank Database

The Genomic Depth of the Tree of Life



Next-Gen Sequencing is Qualitative and Quantitative



NGSTs

Each DNA template
is sequenced directly

Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript

Capillary Sequencing

All DNA templates are sequenced
together to create a single
consensus sequence

Grey transcript

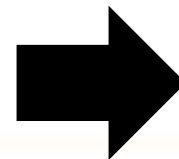


Can we Use RNA-Seq to Increase Genomic Depth?

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

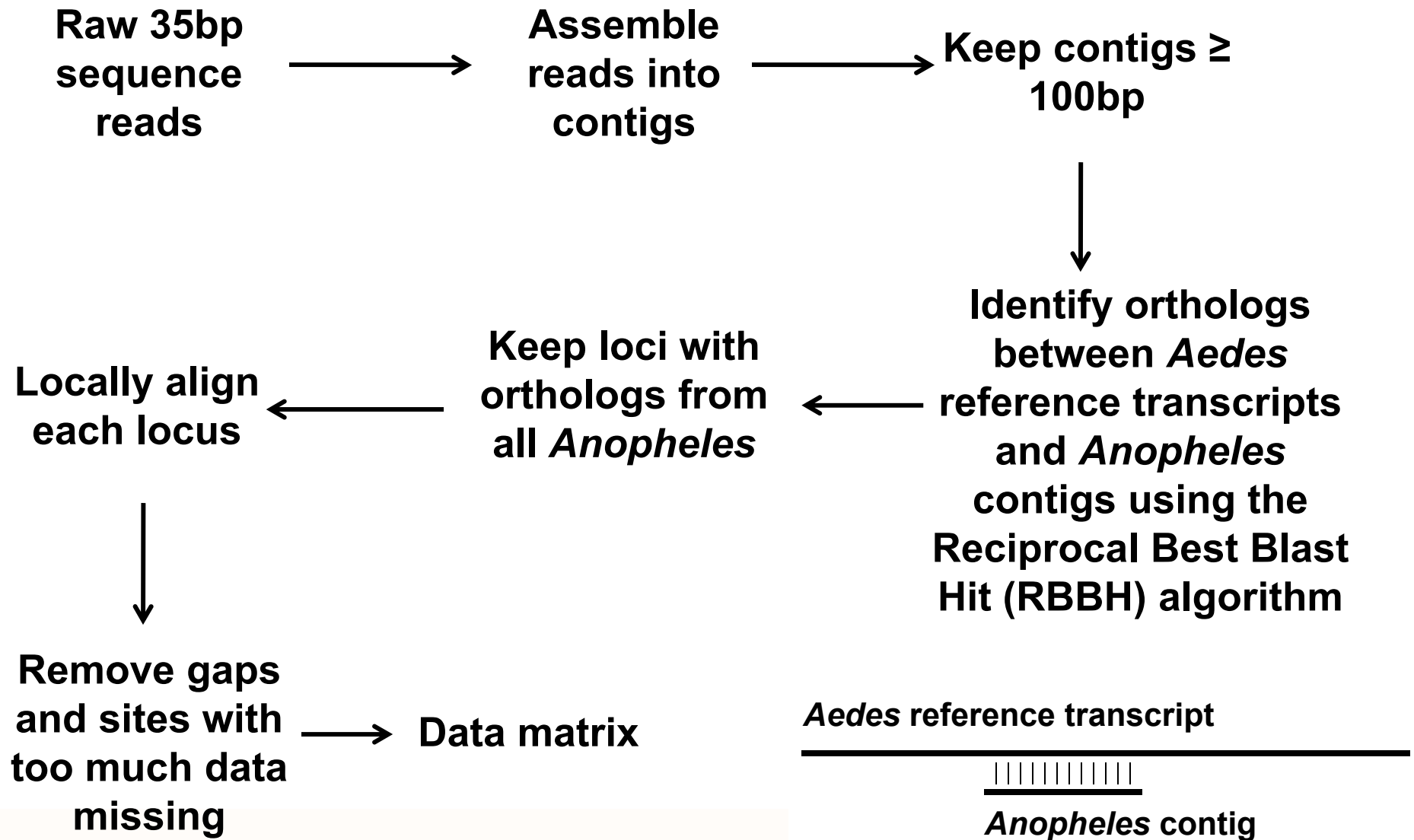


Illumina
RNA-Seq

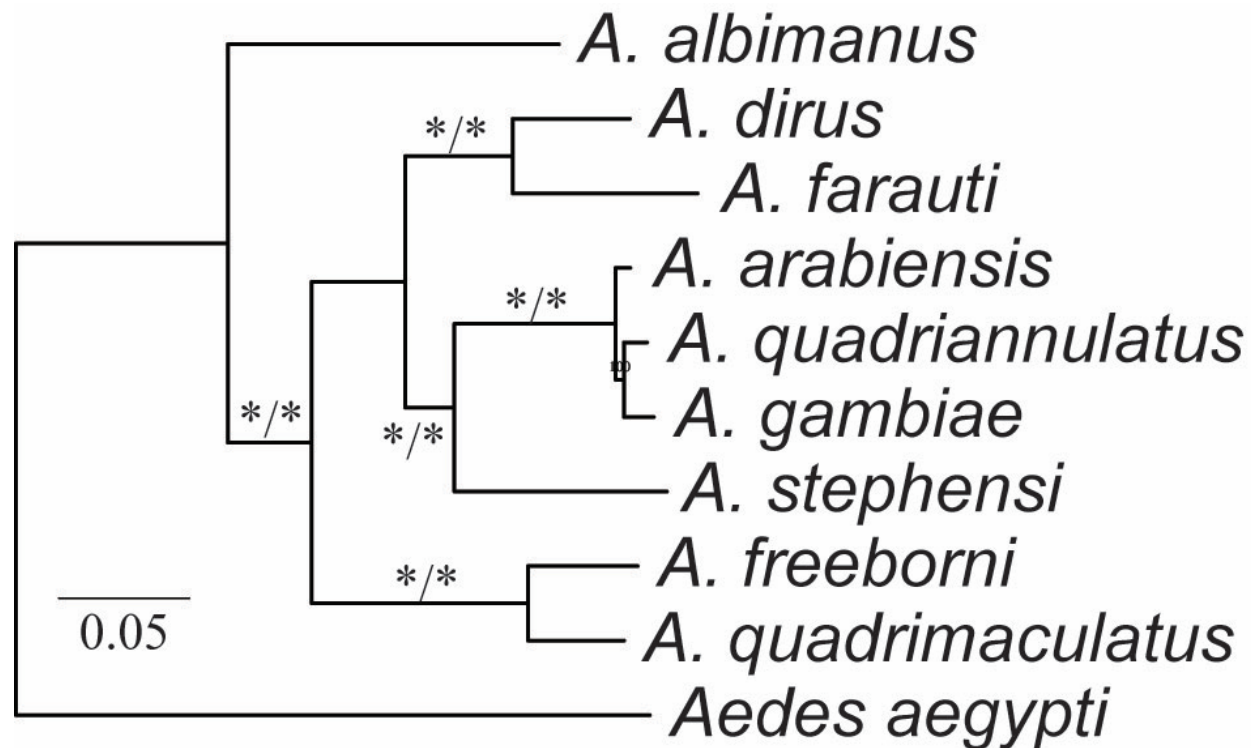


~150,000,000 reads
5,250,000,000 bp

Data Matrix Construction: The “Singlecontig” Strategy



Robust Phylogenetic Inference from RNA-Seq Data

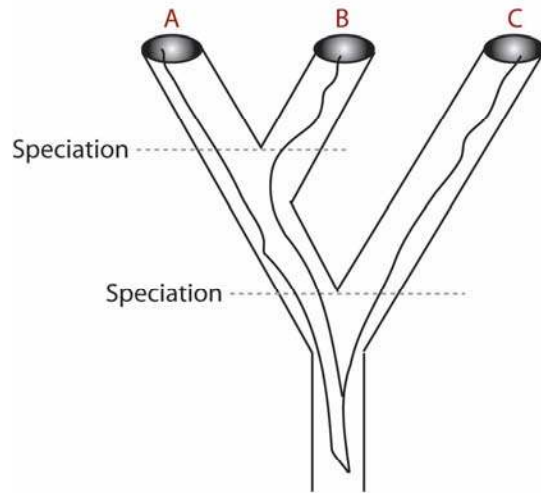


Loci = 553
Aln Length = ~390 Kb
% Missing data = 51

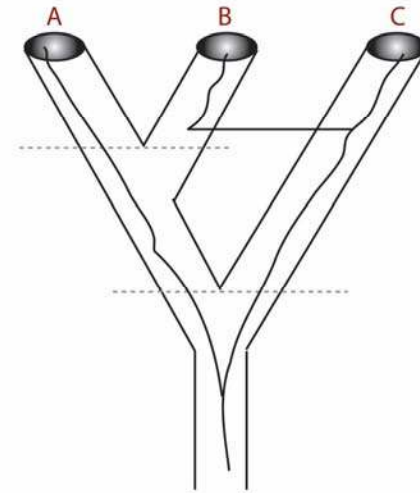


Gene Trees Can Differ from Species Trees

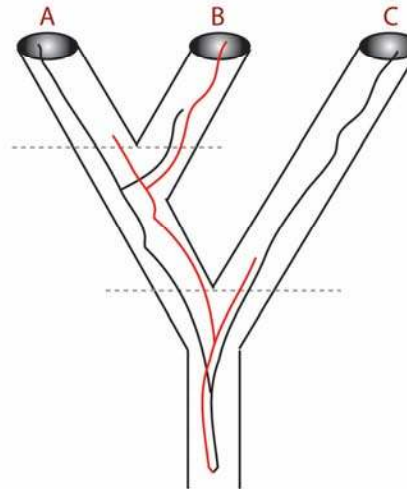
Lineage Sorting



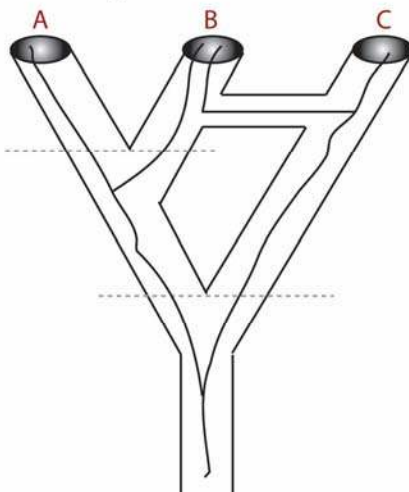
Horizontal Gene Transfer



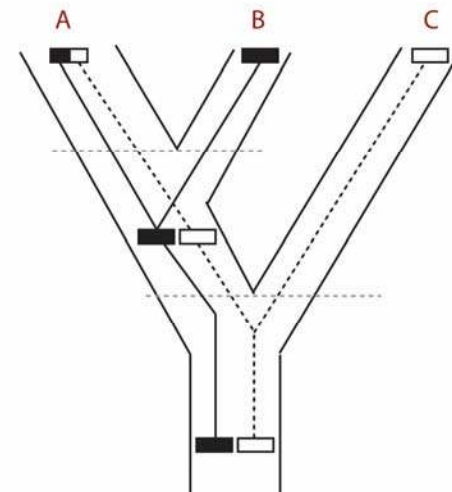
Gene Duplication and Loss



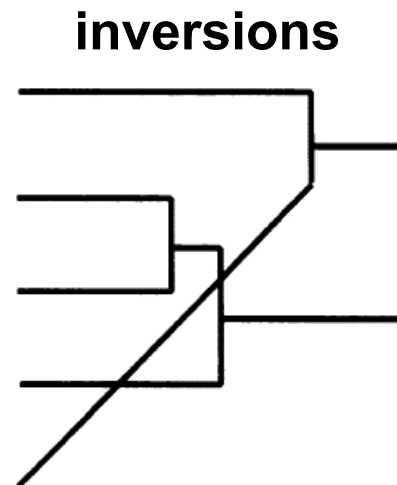
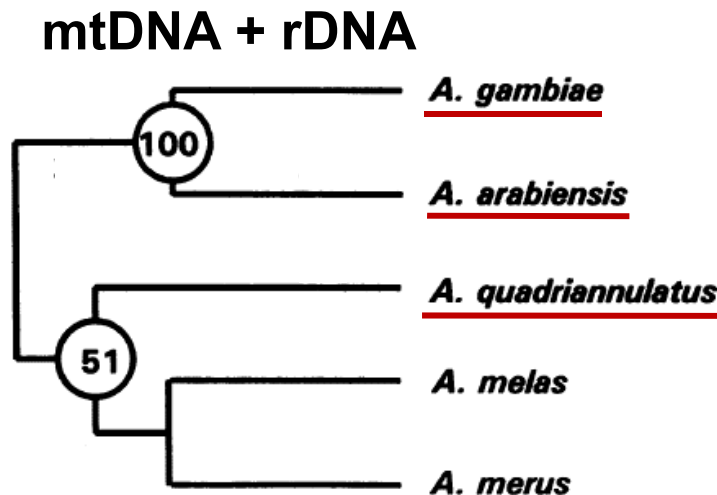
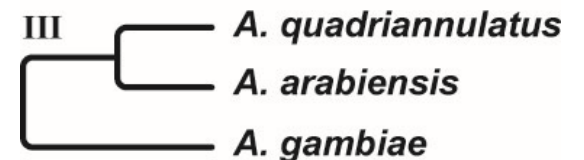
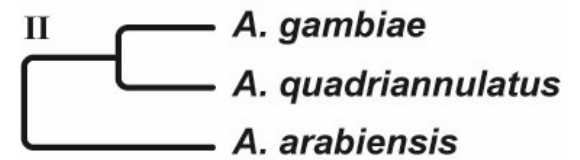
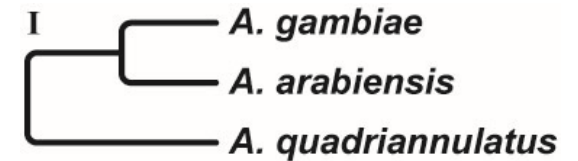
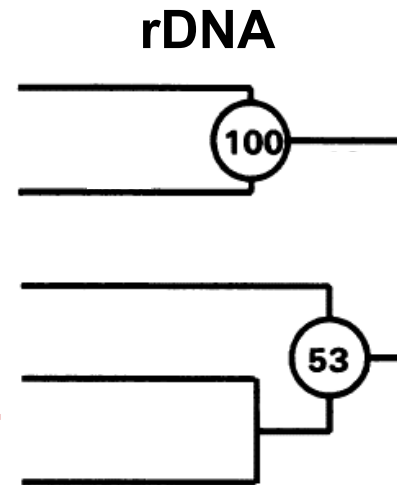
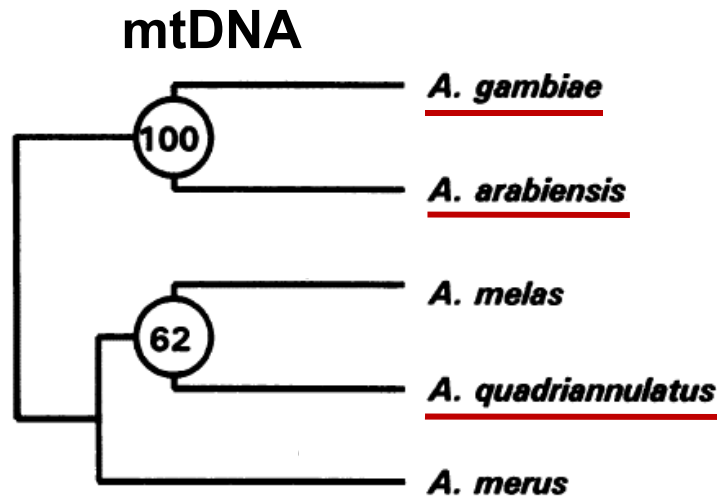
Hybridization



Recombination



Our Data Matrices Can Detect Population-Level Events



The Phylogenomics Era – “Resolving” the Tree of Life

Syst. Biol. 61(1):150–164, 2012

© The Author(s) 2011. Published by Oxford University Press on behalf of Society of Systematic Biologists.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI:10.1093/sysbio/syr089

Advance Access publication on September 7, 2011

LETT
LETT

Phylogenomic Analysis Resolves the Interordinal Relationships and Rapid Diversification of the Laurasiatherian Mammals

XUMING ZHOU, SHIXIA XU, JUNXIAO XU, BINGYAO CHEN, KAIYA ZHOU, AND GUANG YANG*

Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China;

*Correspondence to be sent to: Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China; E-mail: gyang@njnu.edu.cn.

Resolving the evolutionary relationships of molluscs with phylogenomic tools

nature

Stephen A. Smith^{1,2}, Nerida G. Wilson^{3,4}, Freya Gonzalo Giribet⁵ & Casey W. Dunn¹

Syst. Biol. 57(6):920–938, 2008

Copyright © Society of Systematic Biologists

ISSN: 1063-5157 print / 1076-836X online

DOI: 10.1080/10635150802570791

Resolving Arthropod Phylogeny: Exploring Phylogenetic Signal within 41 kb of Protein-Coding Nuclear Gene Sequence

JEROME C. REGIER,¹ JEFFREY W. SHULTZ,² AUSTEN R. D. GANLEY,^{3,6} APRIL HUSSEY,¹ DIANE SHI,¹ BERNARD BALL,³ ANDREAS ZWICK,¹ JASON E. STAJICH,^{3,7} MICHAEL P. CUMMINGS,⁴ JOEL W. MARTIN,⁵ AND CLIFFORD W. CUNNINGHAM³

Toward Resolving the Tree: The Phylogeny of Jakobids and Cercozoans

Yeast

An

Toward Resolving Priors

Prion-Like Proteins in the Fungal Kingdom

Edgar M. Medina · Gary W. Jones · David A. Fitzpatrick

OPEN ACCESS Free

Towards

Renee C. Pratt, Gillian C. Gibb,* Mary Morgan-Richards,* Matthew J. Phillips,† Michael D. Hendy,* and David Penny**

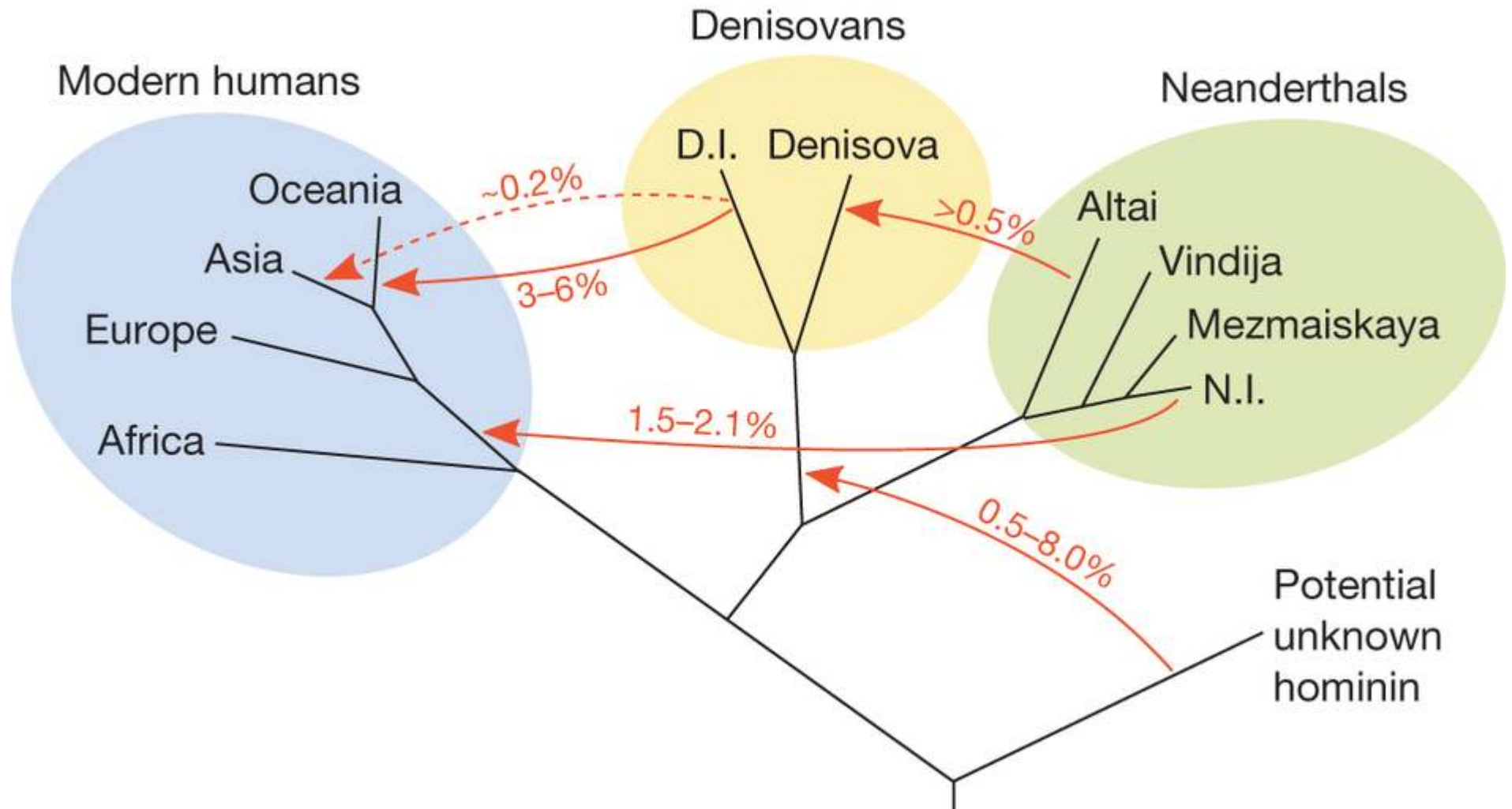
Samuli Lehtonen

Department of Biology, U

*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and †Centre for Macroevolution and Macroecology, School of Botany and Zoology, Australian National University, Canberra ACT, Australia

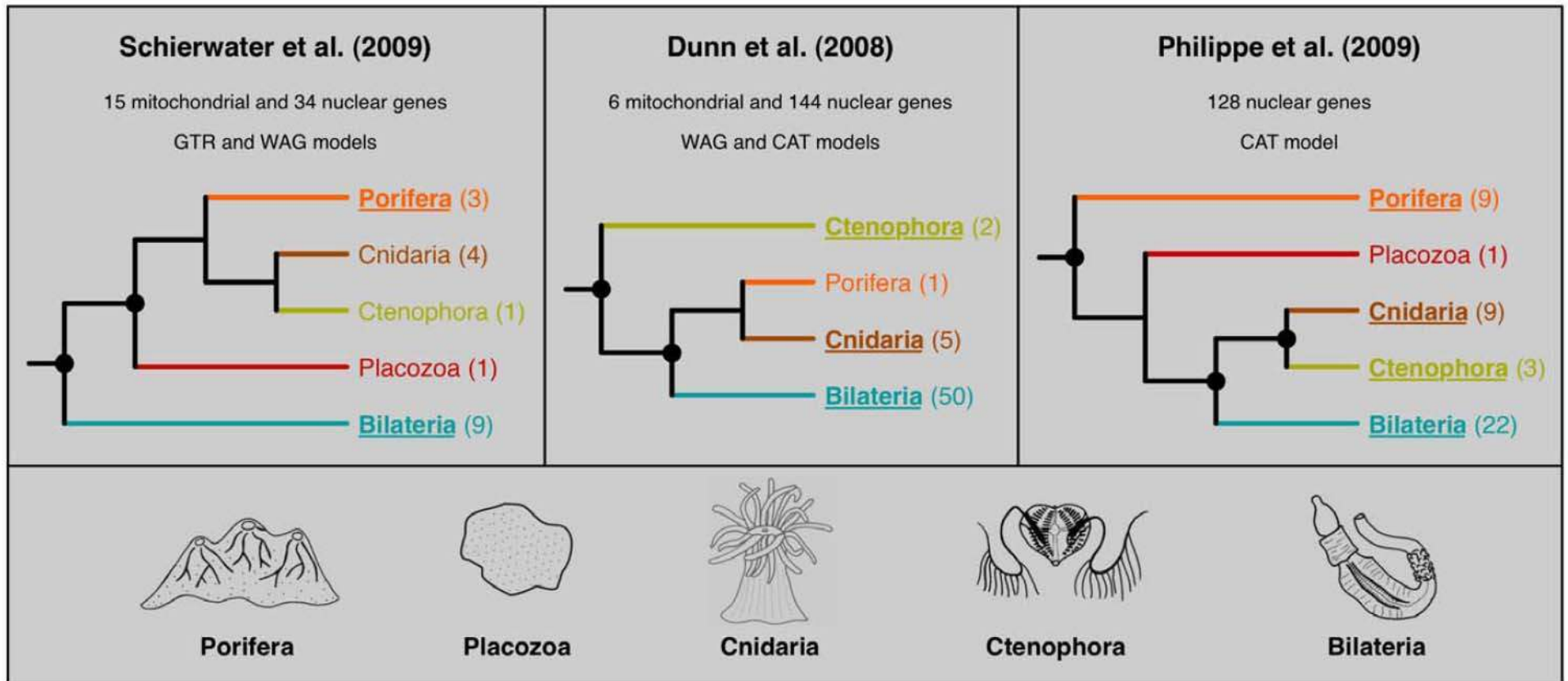
**Have we eliminated
incongruence?**

Phylogenomics Works Beautifully at Shallow Levels



Prufer et al. (2014) Nature

Incongruence in Deep Time



Incongruence in Deep Time



Why the disconnect?

An Expanded Yeast Data Matrix

Yeast Gene Order Browser (YGOB)



**Saccharomyces
lineage**

**1,070 genes
23 taxa
no missing data**

Candida Gene Order Browser (CJOB)



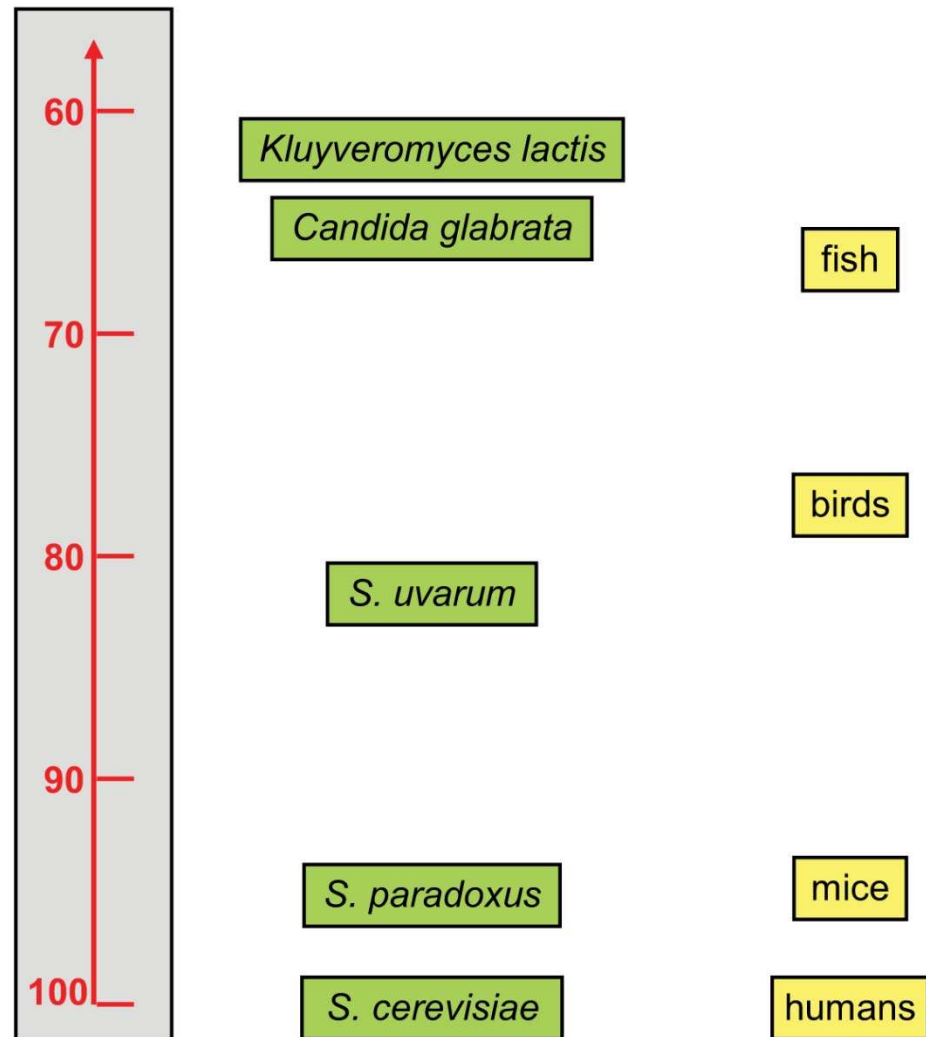
**Candida
lineage**



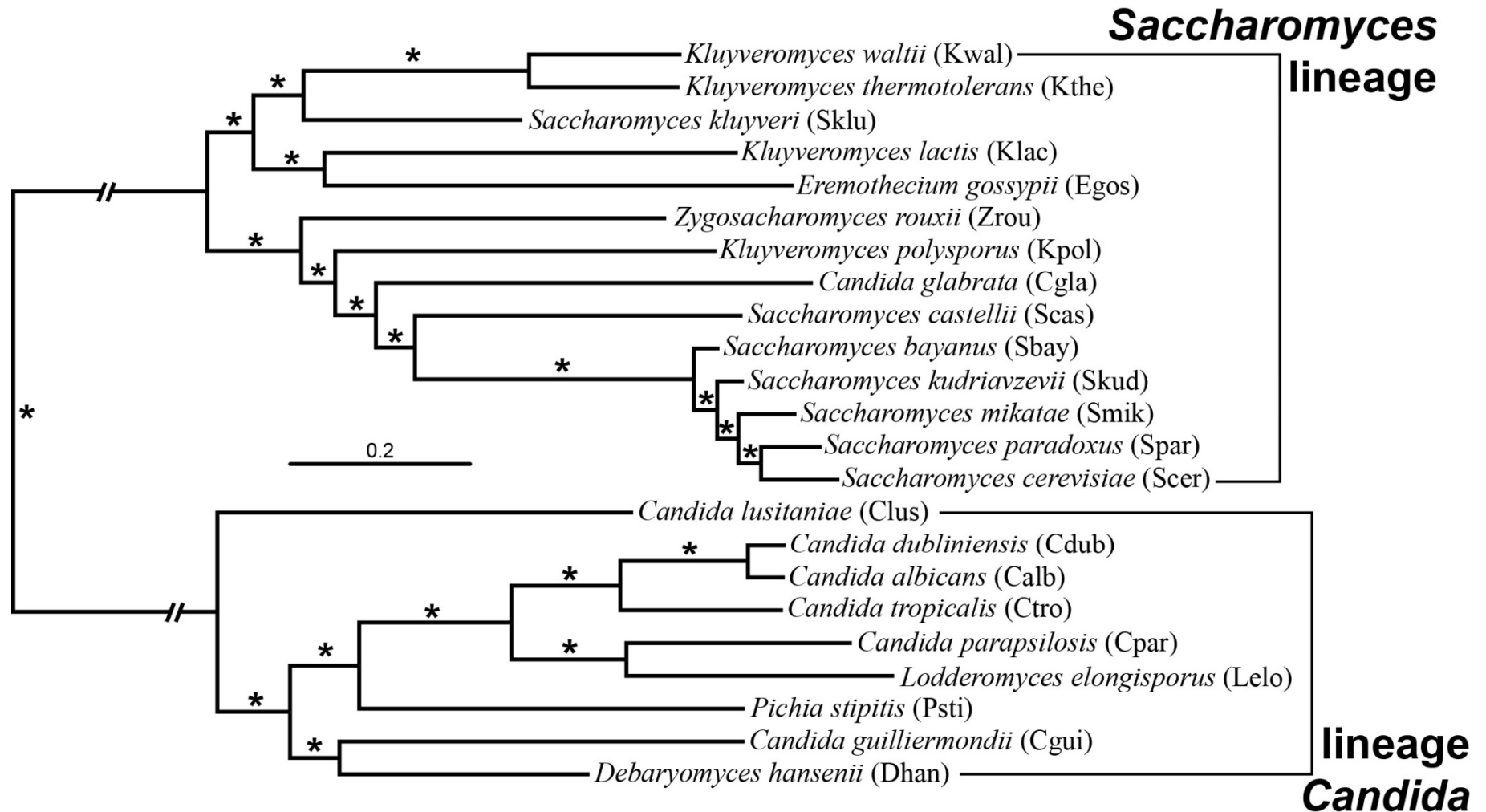
Fungal Genomes are Similar in Divergence to Animals

Proteome-wide average pairwise amino acid sequence similarity

***Saccharomyces*,
Candida, *Kluyveromyces*,
etc. are all polyphyletic
genera**



Concatenation Yields an Absolutely Supported Phylogeny

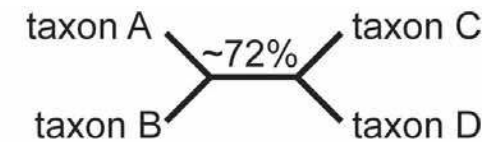


Bootstrap Support is Misleading When Used in Large Datasets

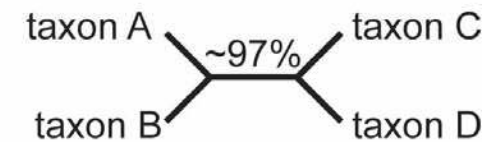
53% 47%

taxonA AAAAAAAAAATTTTTTTTT
taxonB AAAAAAAAAACCCCCCCCC
taxonC GGGGGGGGGTTTTTTTTT
taxonD GGGGGGGGGCCCCCCCCC

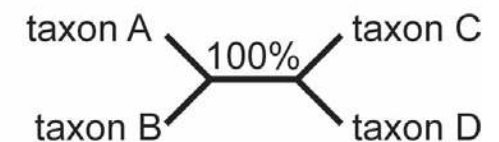
100 characters



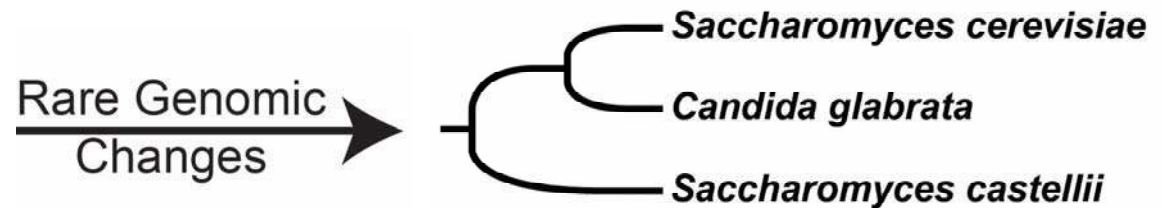
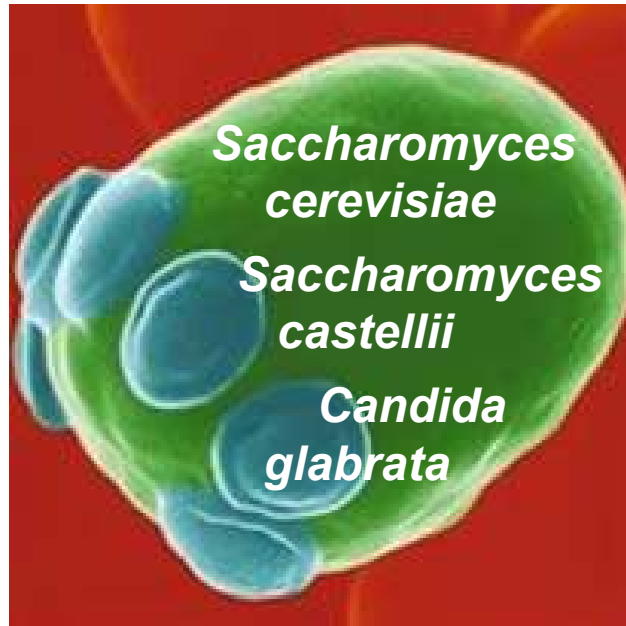
1,000 characters



10,000 characters

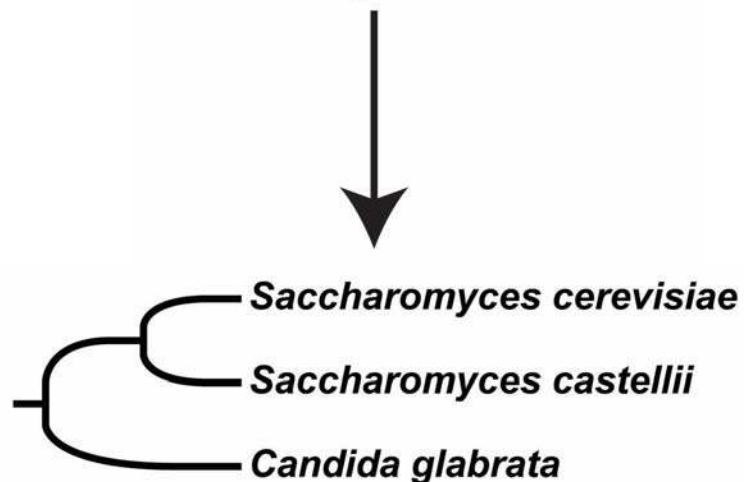


The Concatenation Phylogeny is at Least Partly Wrong

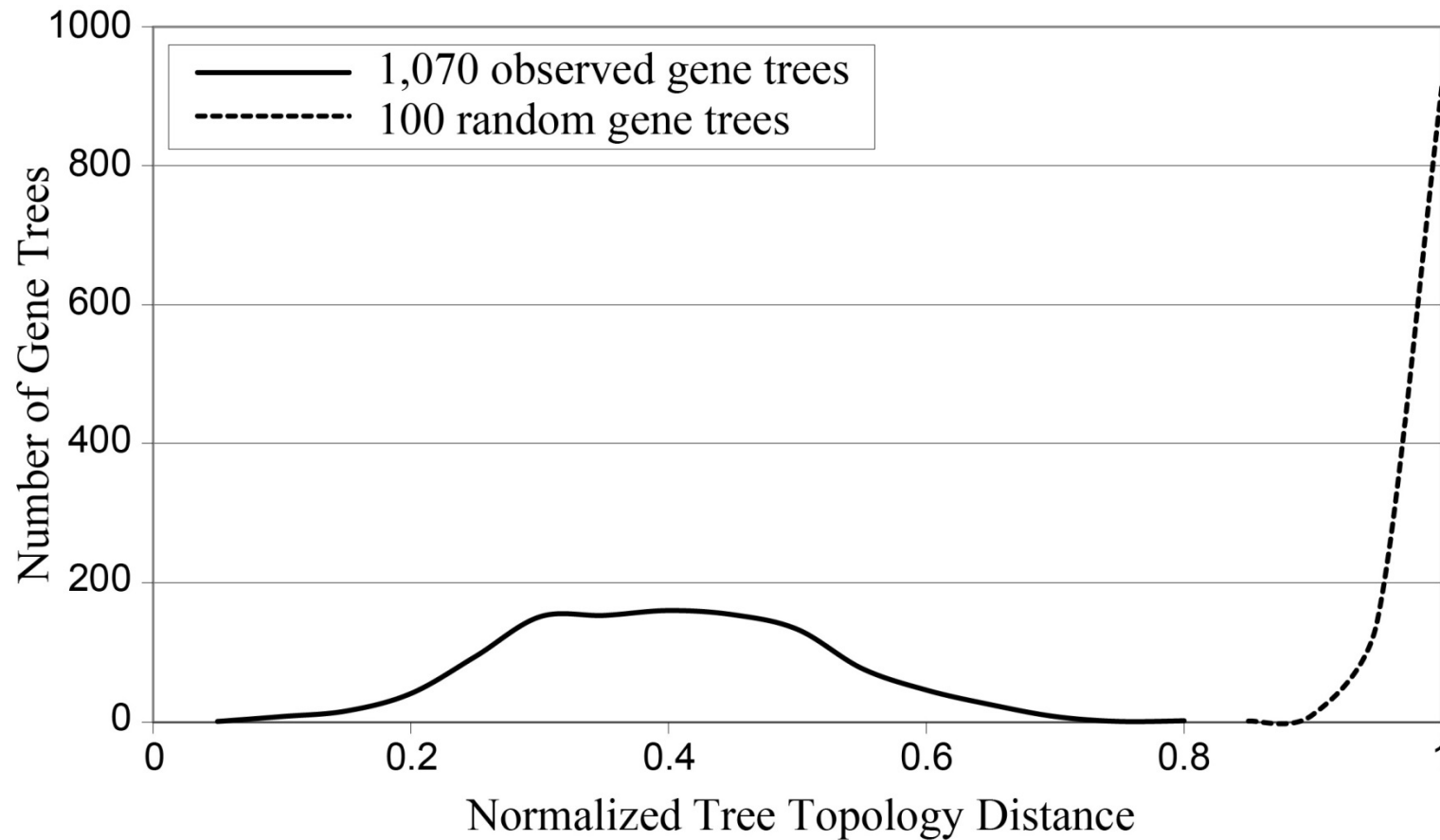


- ❖ 5 genomic rearrangements that are uniquely shared by *S. cerevisiae* and *C. glabrata*
- ❖ Much higher proportion of shared gene losses in *S. cerevisiae* and *C. glabrata*
- ❖ Bias in the placement of *C. glabrata* as an outgroup of *S. cerevisiae* and *S. castellii*

Linear Sequence Data



All Gene Trees Differ from the Concatenation Phylogeny



Gene Trees are Incongruent in Most Datasets



**182 / 184 gene
trees are distinct**



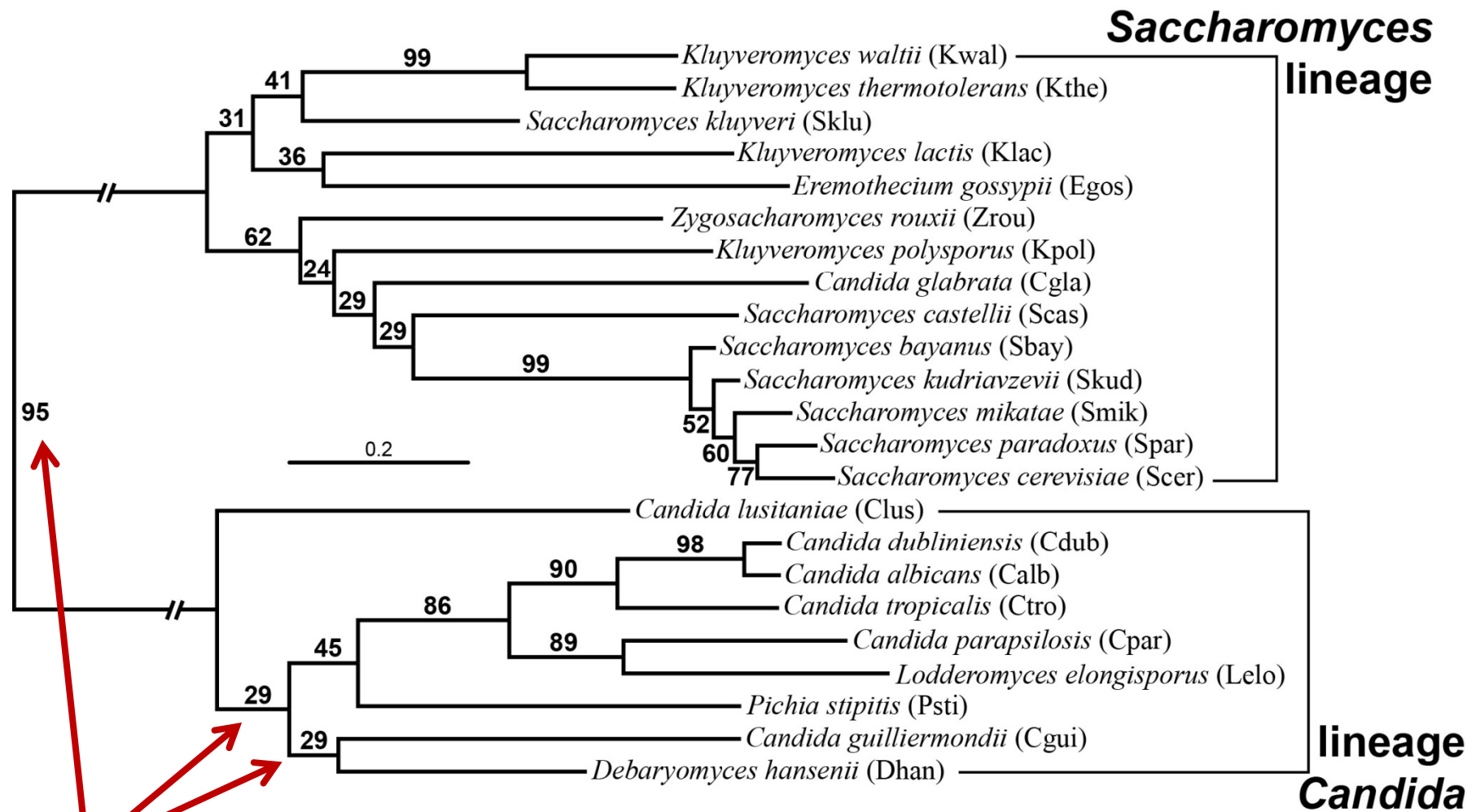
**440 / 447 gene
trees are distinct**



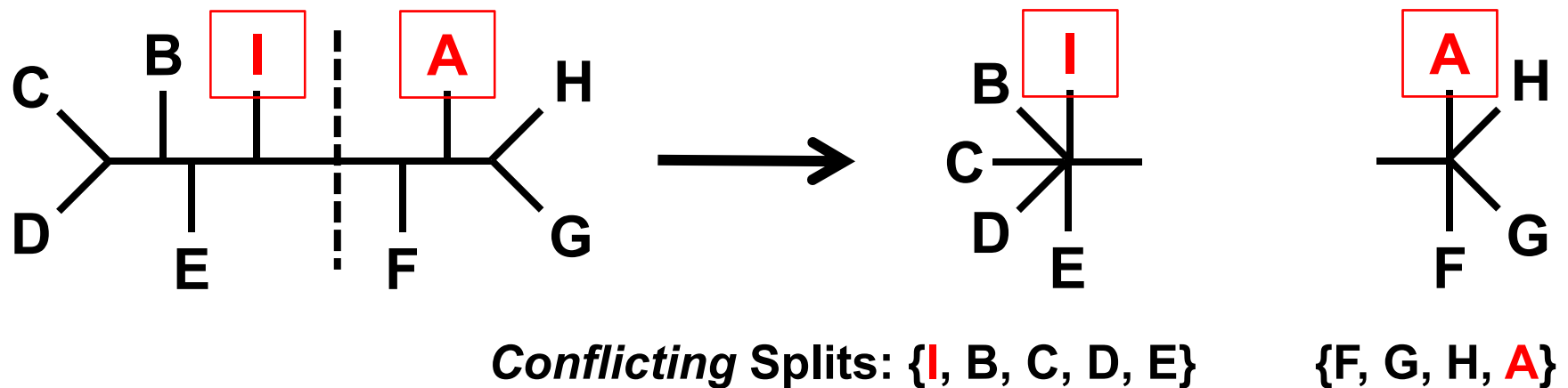
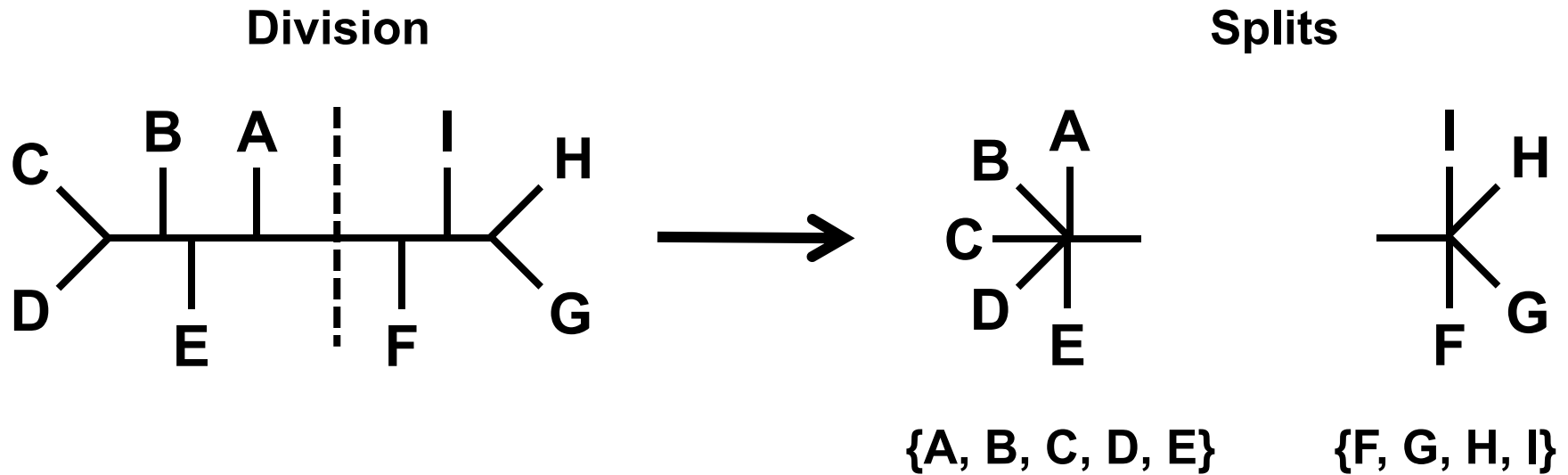
Zhong et al. (2013) Trends Plant Sci.

Song et al. (2012) PNAS

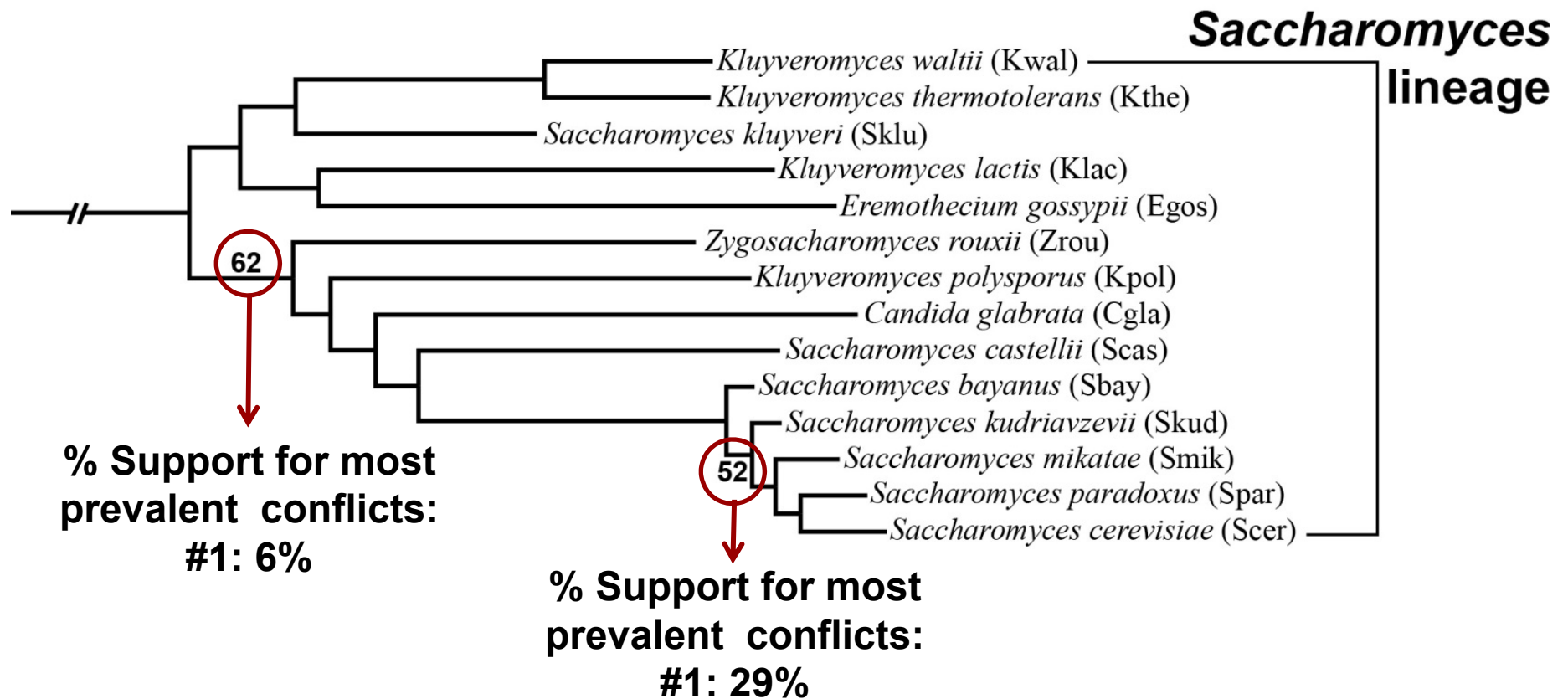
The Yeast Phylogeny Inferred by Majority-Rule Consensus



Phylogenetic Trees are Sets of Splits



Measuring the Degree of Conflict for each Internode

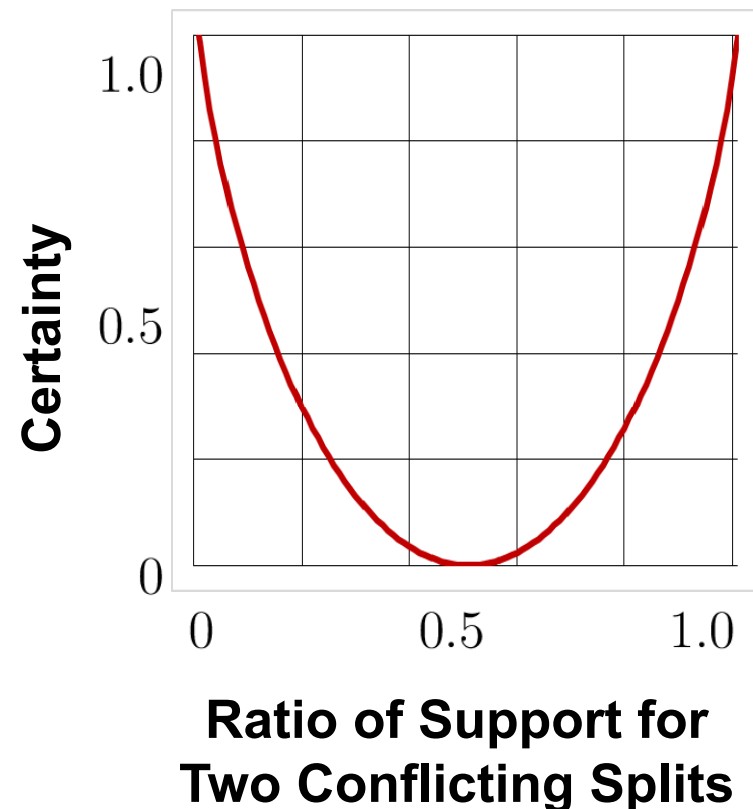


New Measures to Quantify Incongruence

Internode Certainty (IC): a measure of the support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting split in the same set of trees

Tree Certainty (TC): the sum of IC across all internodes

IC and TC are implemented in the latest versions of RAxML



Internode Certainty and Tree Certainty

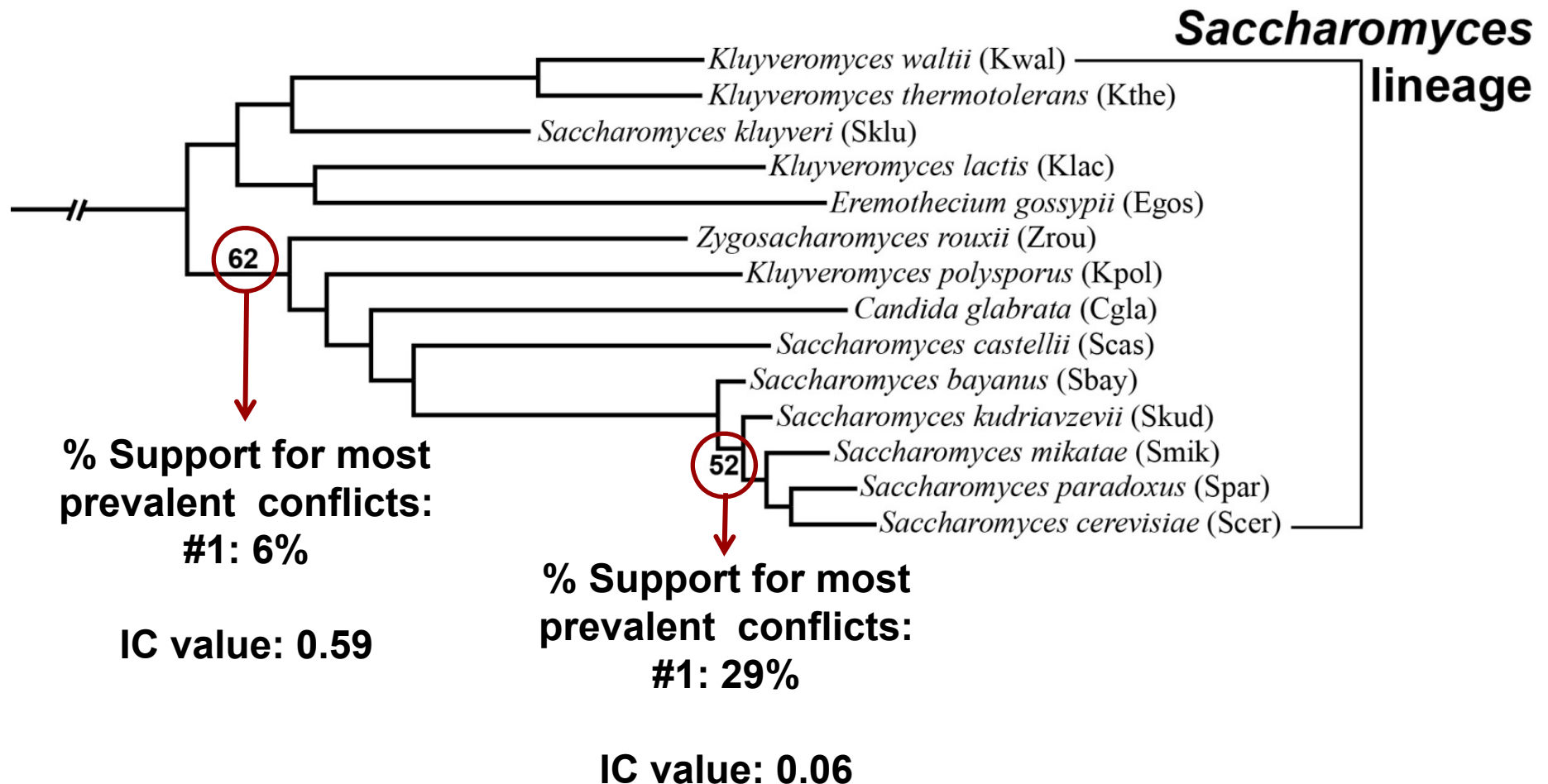
- ❖ Can be used on any set of data that either contain or define splits
 - Bootstrap replicate trees from a single gene
 - Gene trees
 - Sites in an alignment

	<u>abcdefghijkl</u>	
Taxon A	1010011111	#b: {A,B} / {C,D,E}
Taxon B	1010011110	
Taxon C	1110111110	#e: {A,B} / {C,D,E}
Taxon D	1110111110	
Taxon E	1110111111	#p: {A,E} / {B,C,D}

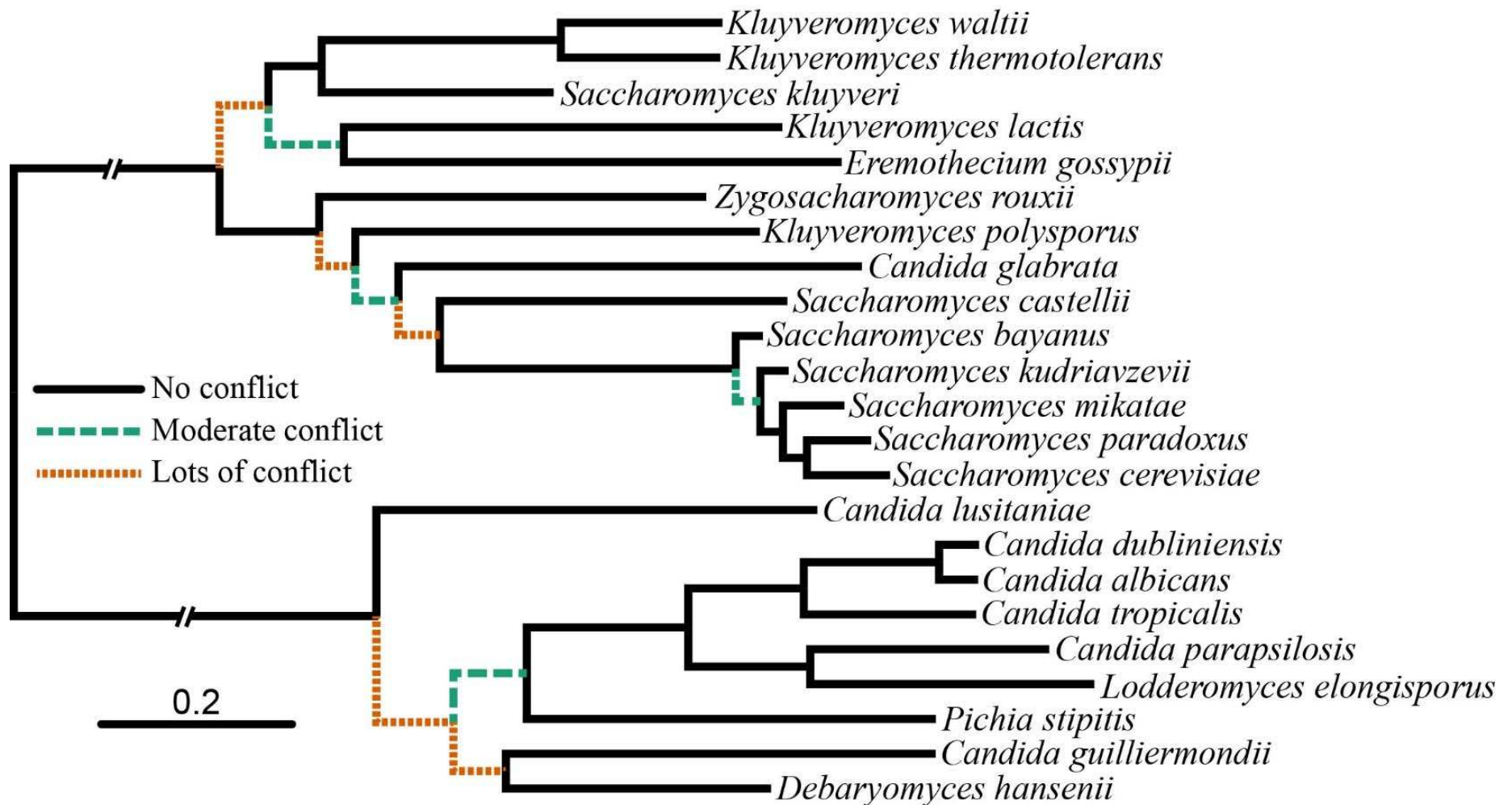
- ❖ Can be used to compare the effect of different treatments on a multigene data set



IC Can Be More Informative Measure of Internode Support

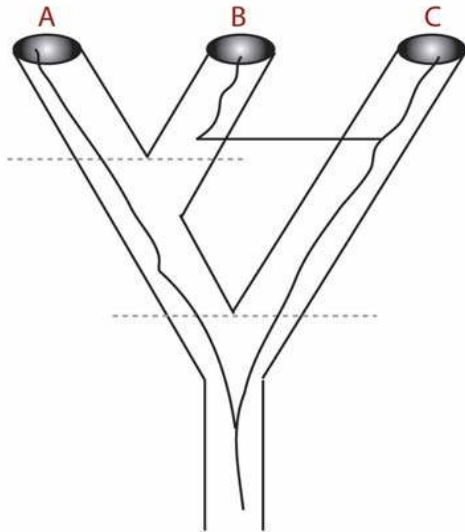


IC Reveals that There is Rampant Conflict

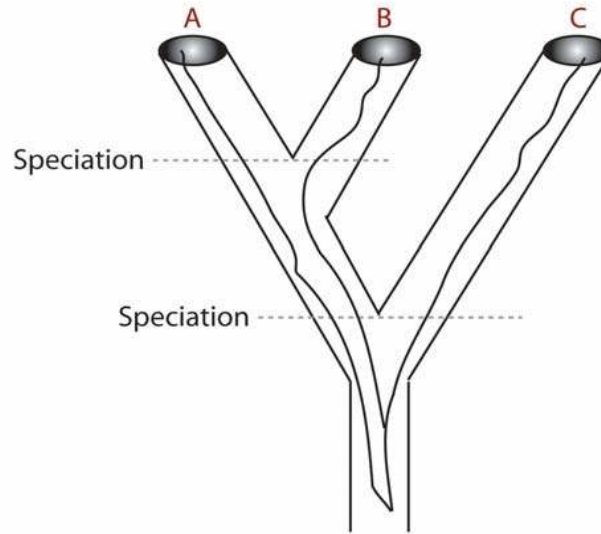


What So Much Incongruence? Biological Factors

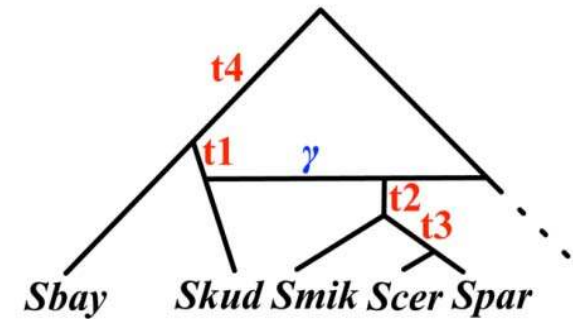
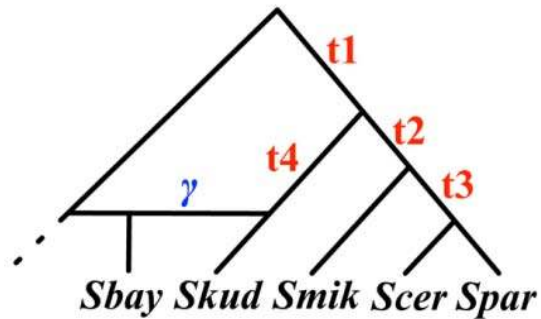
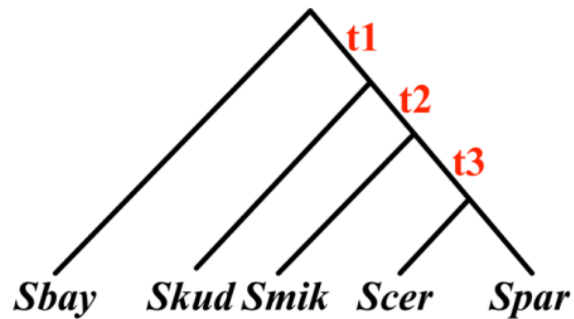
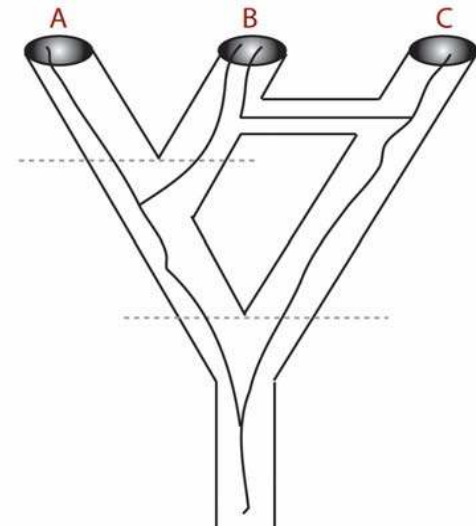
Horizontal Gene Transfer



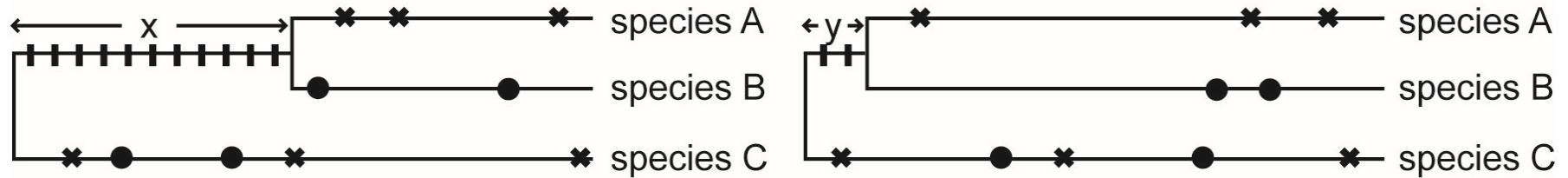
Lineage Sorting



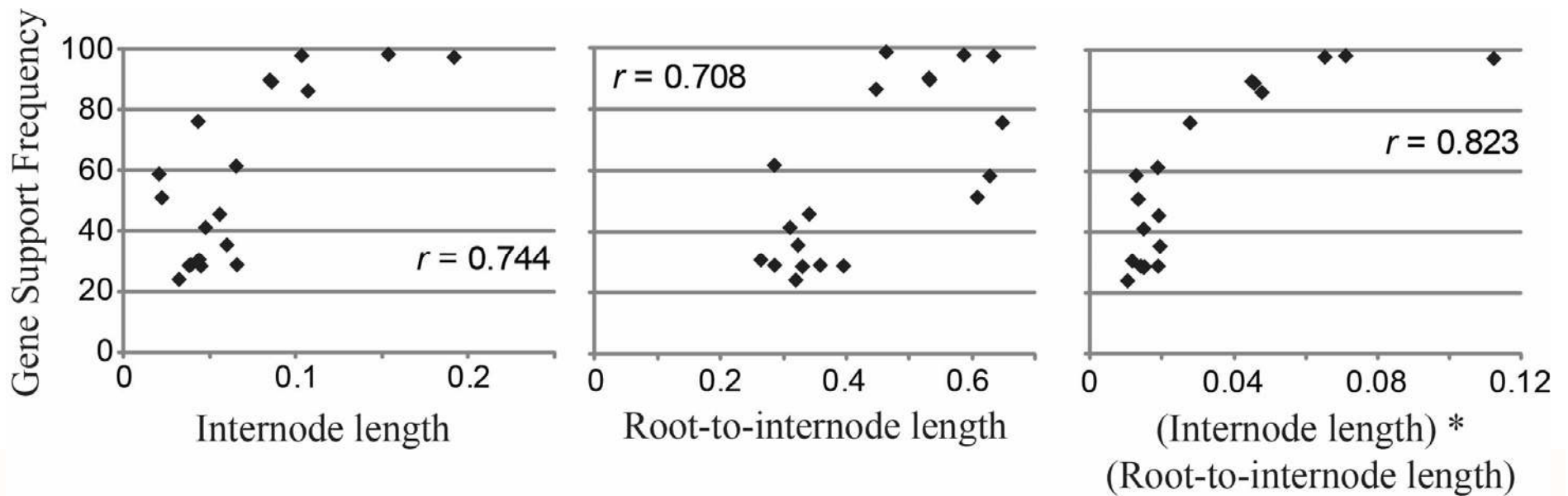
Hybridization



What So Much Incongruence? Analytical Factors



Internode length: influences amount of phylogenetic signal (I)
Homoplasy: independent evolution of identical characters (* , •)





Standard Recipes for Handling Incongruence Didn't Help

Treatment	Tree Certainty	# of Internodes where IC increased decreased
Default analysis	8.35	n/a
<i>Removing sites containing gaps</i>		
All sites with gaps excluded	7.91	0 7
<i>Removing fast-evolving or unstable species</i>		
<i>C. lusitaniae</i>	8.15	1 2
<i>C. glabrata</i>	8.30	2 2
<i>E. gossypii</i> , <i>C. glabrata</i> , <i>K. lactis</i>	7.88	1 3
<i>Selecting genes that recover specific clades</i>		
[<i>C. tropicalis</i> , <i>C. dubliniensis</i> , <i>C. albicans</i>]	8.62	0 0
<i>Selecting the most slow-evolving genes</i>		
100 slowest-evolving genes	6.76	2 9

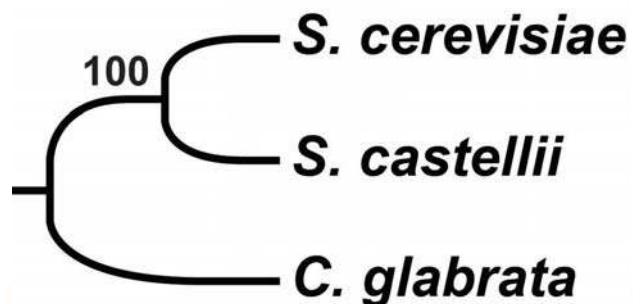




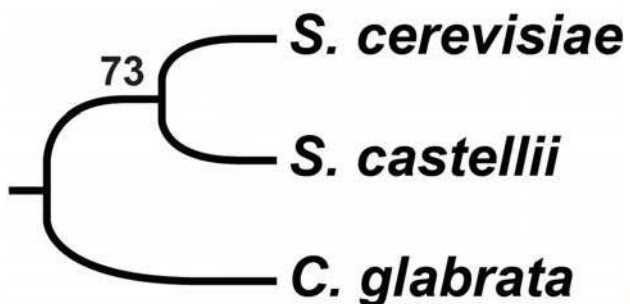
What Do We Do Then?

Treatment	Tree Certainty	# of Internodes where IC increased decreased
Default analysis	8.35	n/a
<i>Selecting genes whose bootstrap consensus trees have high average support</i>		
All genes with average BS $\geq 60\%$	8.59	4 0
All genes with average BS $\geq 70\%$	9.18	14 0
All genes with average BS $\geq 80\%$	9.92	15 0

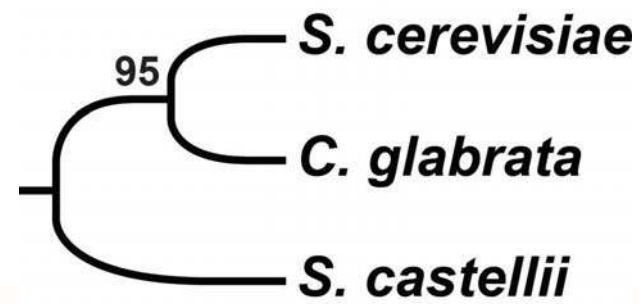
average BS $\geq 60\%$



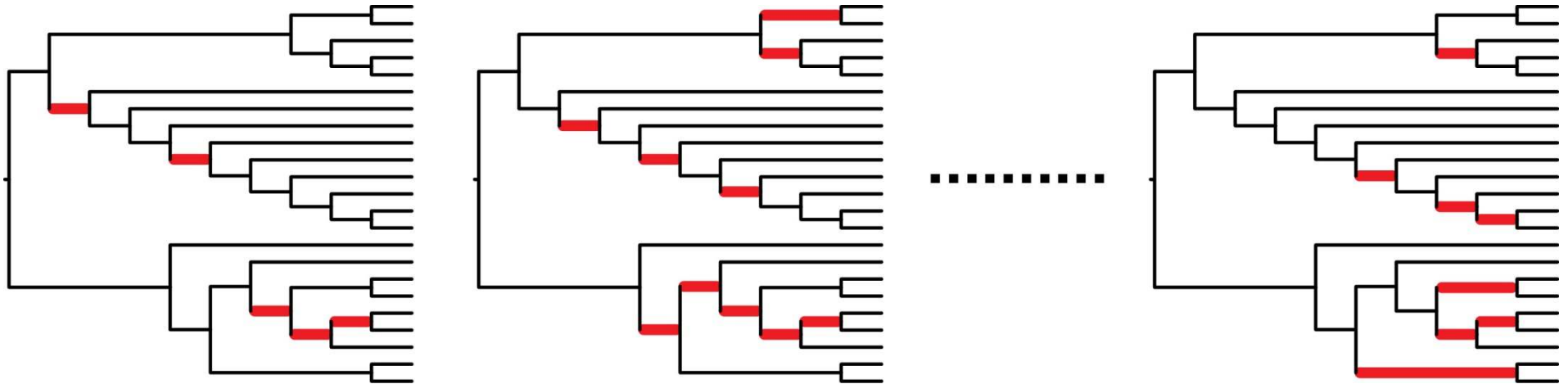
average BS $\geq 70\%$



average BS $\geq 80\%$



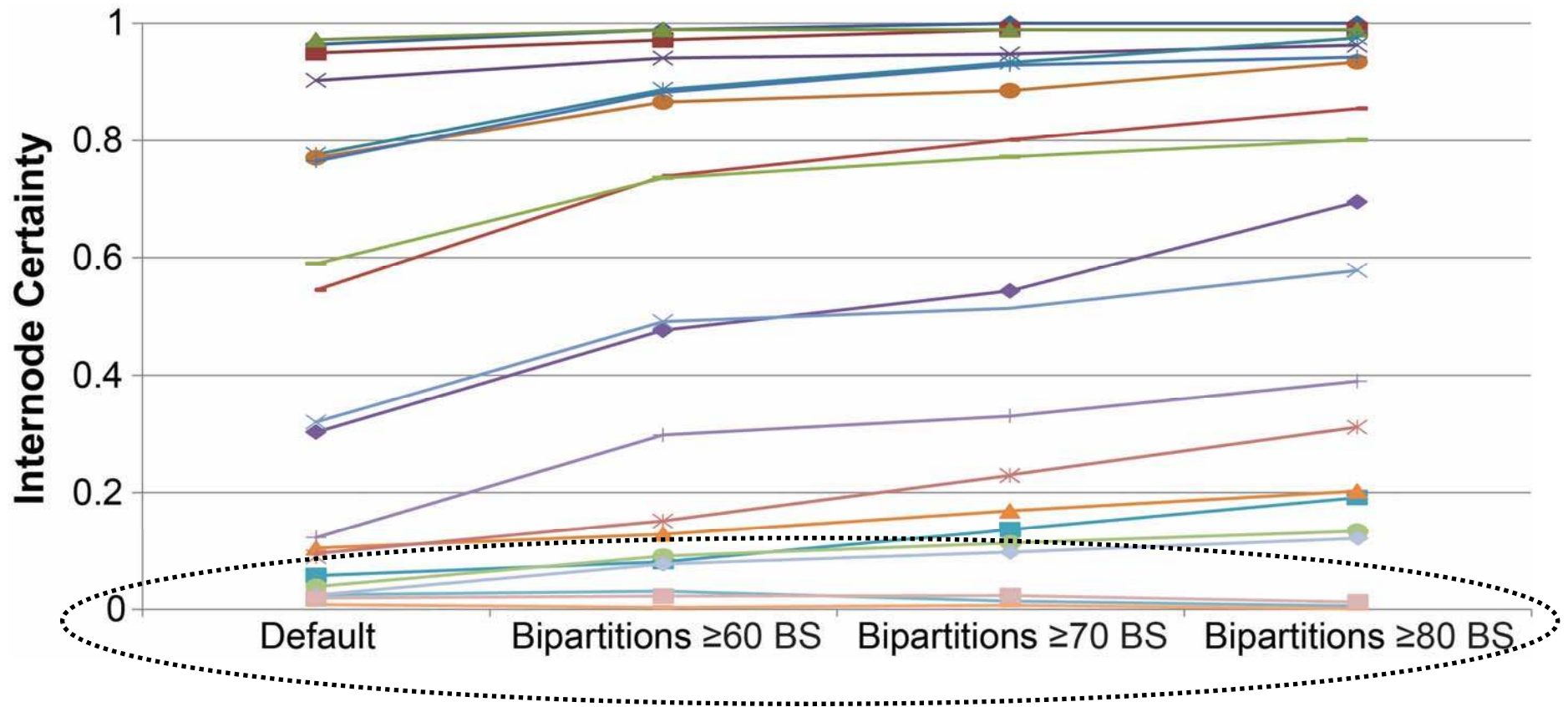
Selecting Specific Bipartitions Dramatically Improves Phylogeny



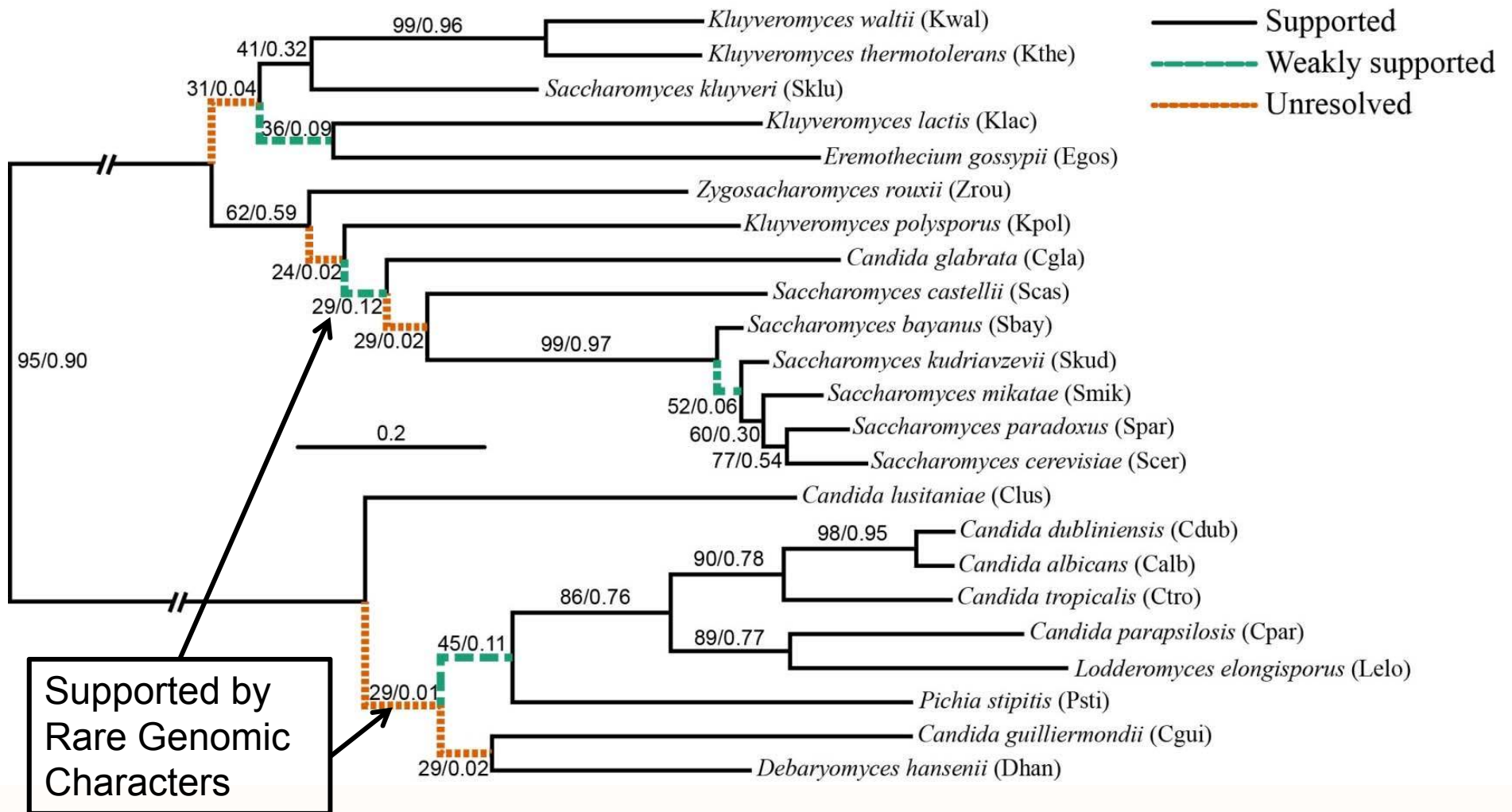
Treatment	Tree Certainty	# of Internodes where IC increased decreased
Default analysis	8.35	n/a
<i>Selecting genes whose bootstrap consensus trees have high average support</i>		
All bipartitions with BS $\geq$ 60%	10.11	14 0
All bipartitions with BS $\geq$ 70%	10.70	16 0
All bipartitions with BS $\geq$ 80%	11.32	15 0



Least Supported Internodes Harbor the Most Conflict

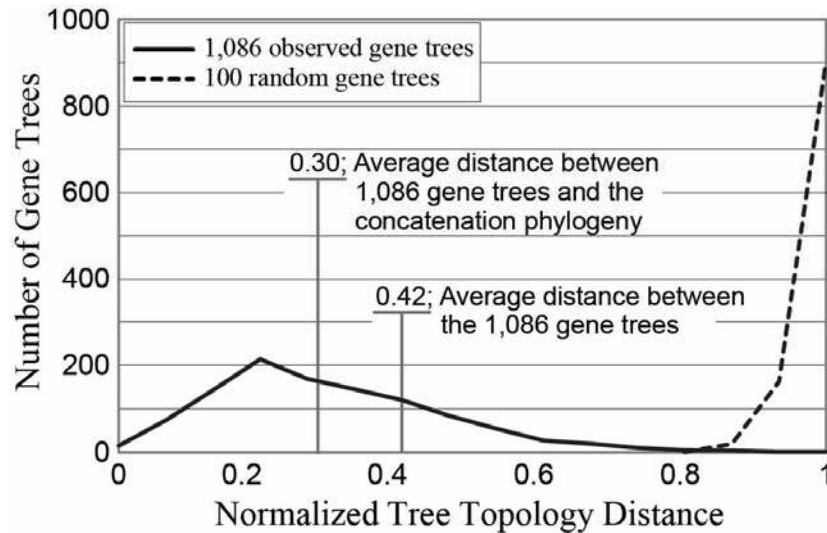


The Status of the Yeast Phylogeny

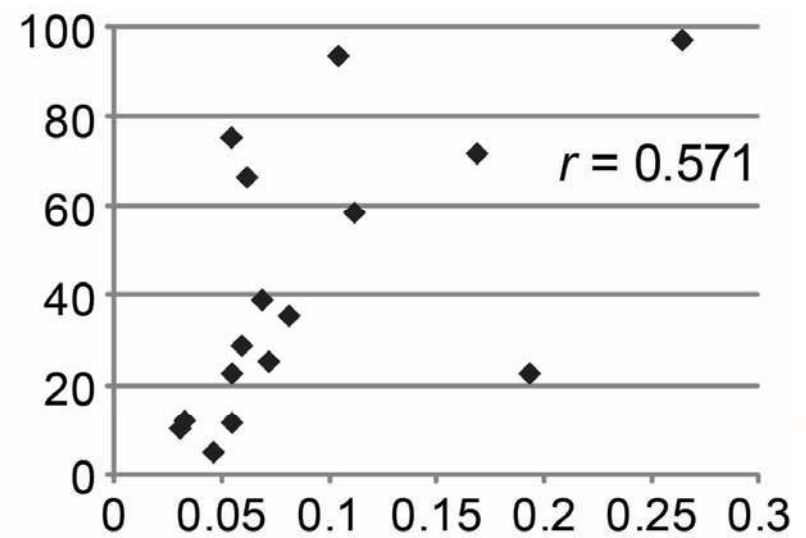
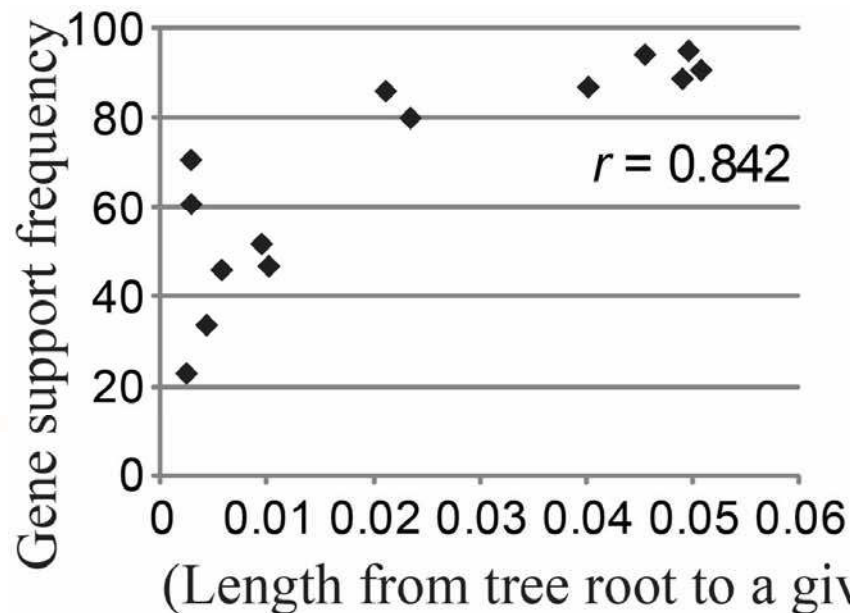
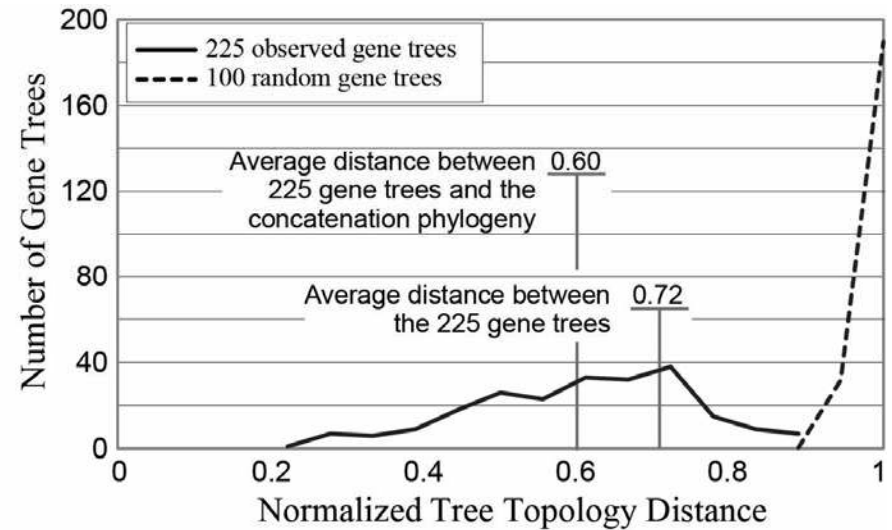


The Same is True for Vertebrate and Metazoan Datasets

Vertebrates (1,086 genes, 18 taxa)

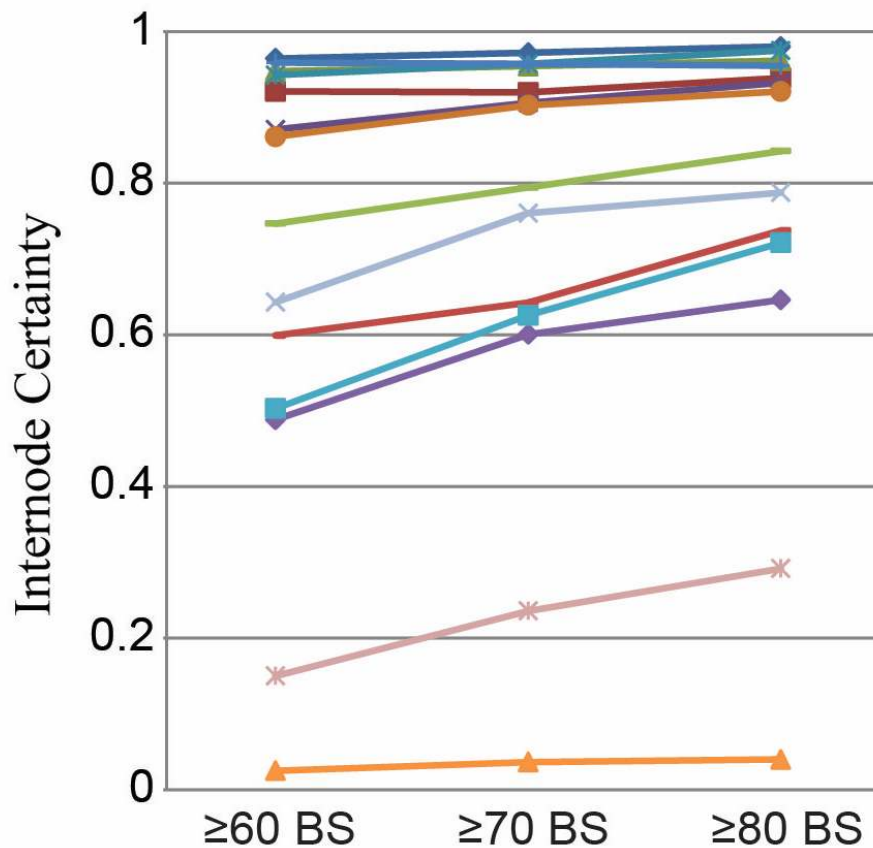


Animals (225 genes, 21 taxa)

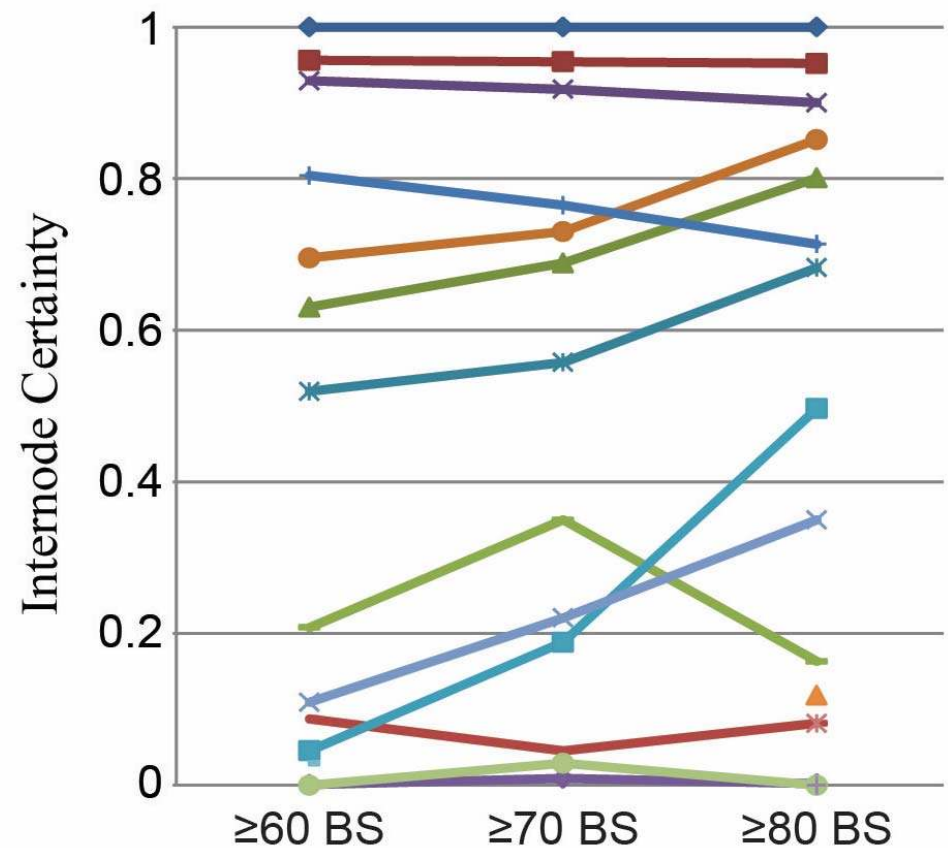


The Same is True for Vertebrate and Metazoan Datasets

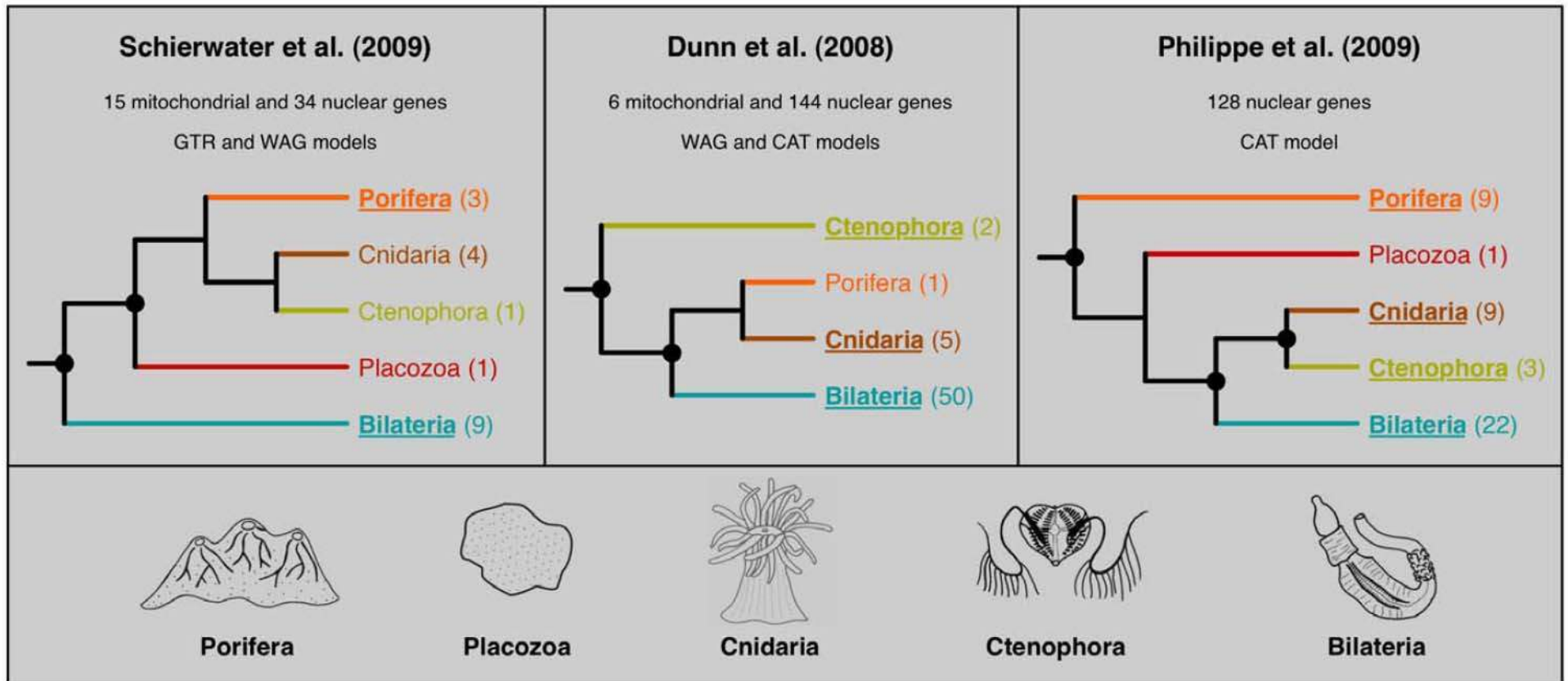
Vertebrates (1,086 genes, 18 taxa)



Animals (225 genes, 21 taxa)



Incongruence in Deep Time



Genomfart?

- ❖ **Parts of the tree of life are more likely to resemble a bush rather than a tree – do we expect that we can confidently infer every branch and twig?**
- ❖ **Bootstrap-based measures not useful in large data sets – desperate need for methods evaluating conflict**
- ❖ **Examining the signal present in individual genes and their trees offers promise**
- ❖ **Explicitly identify internodes that, despite the use of genome-scale data sets, robust study designs and powerful algorithms, are poorly supported**



“One can use the most sophisticated audio equipment to listen, for an eternity, to a recording of white noise and still not glean a useful scrap of information”

**Rodrigo et al. (1994) Chapter in:
Sponge in Time and Space; Biology, Chemistry, Paleontology**

Acknowledgements



Jen Wisecaver



Jason Slot



Abigail Lind



John Gibbons

Collaborating Labs:

Ana Calvo, Northern Illinois Univ.

Chris Hittinger, Univ. Wisconsin-Madison

Alexis Stamatakis, Heidelberg Inst. Theor. Studies



National Science Foundation
WHERE DISCOVERIES BEGIN



Leonidas Salichos



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