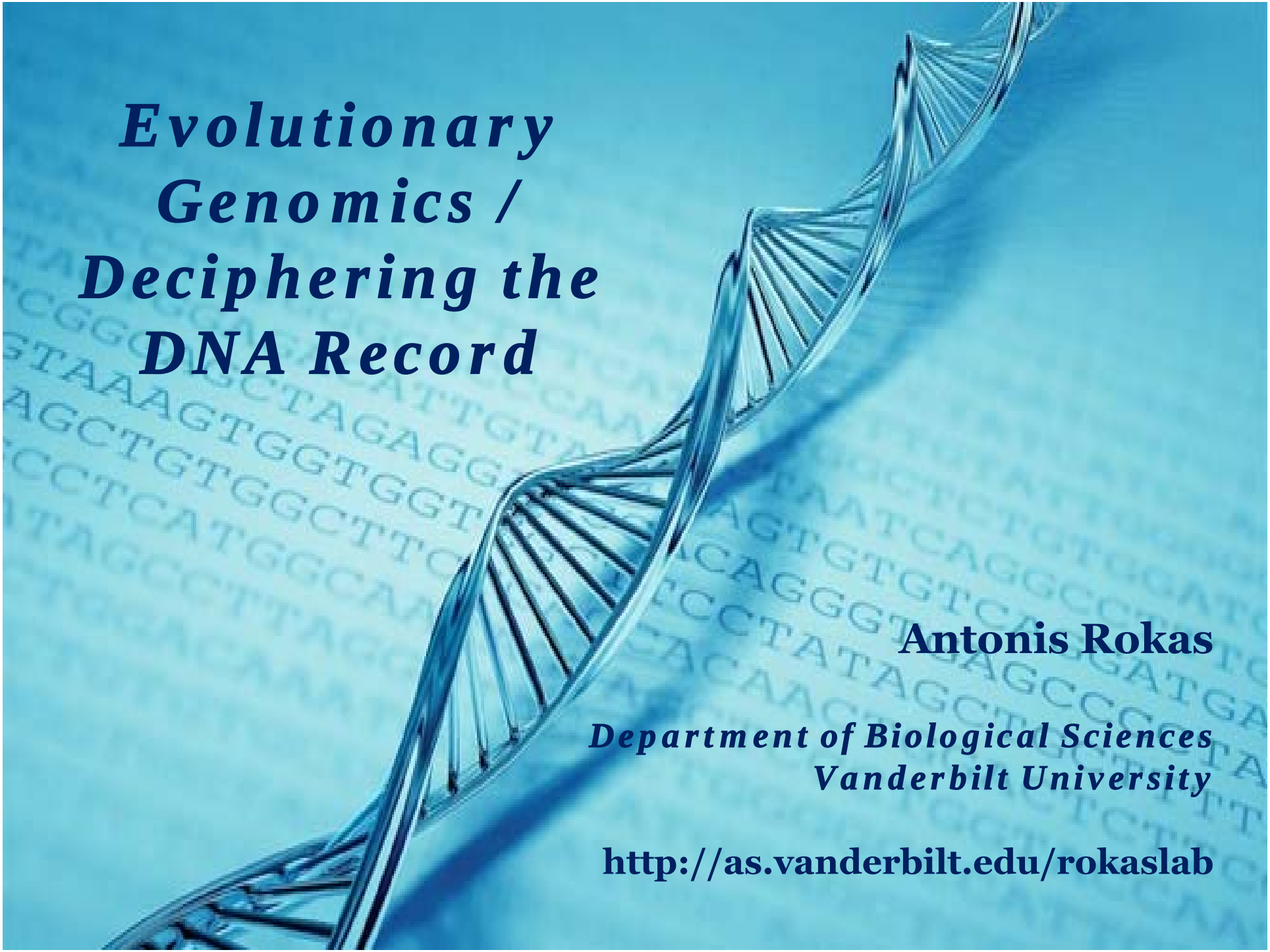


Evolutionary Genomics / Deciphering the DNA Record

Antonis Rokas

Department of Biological Sciences
Vanderbilt University

http://as.vanderbilt.edu/rokaslab

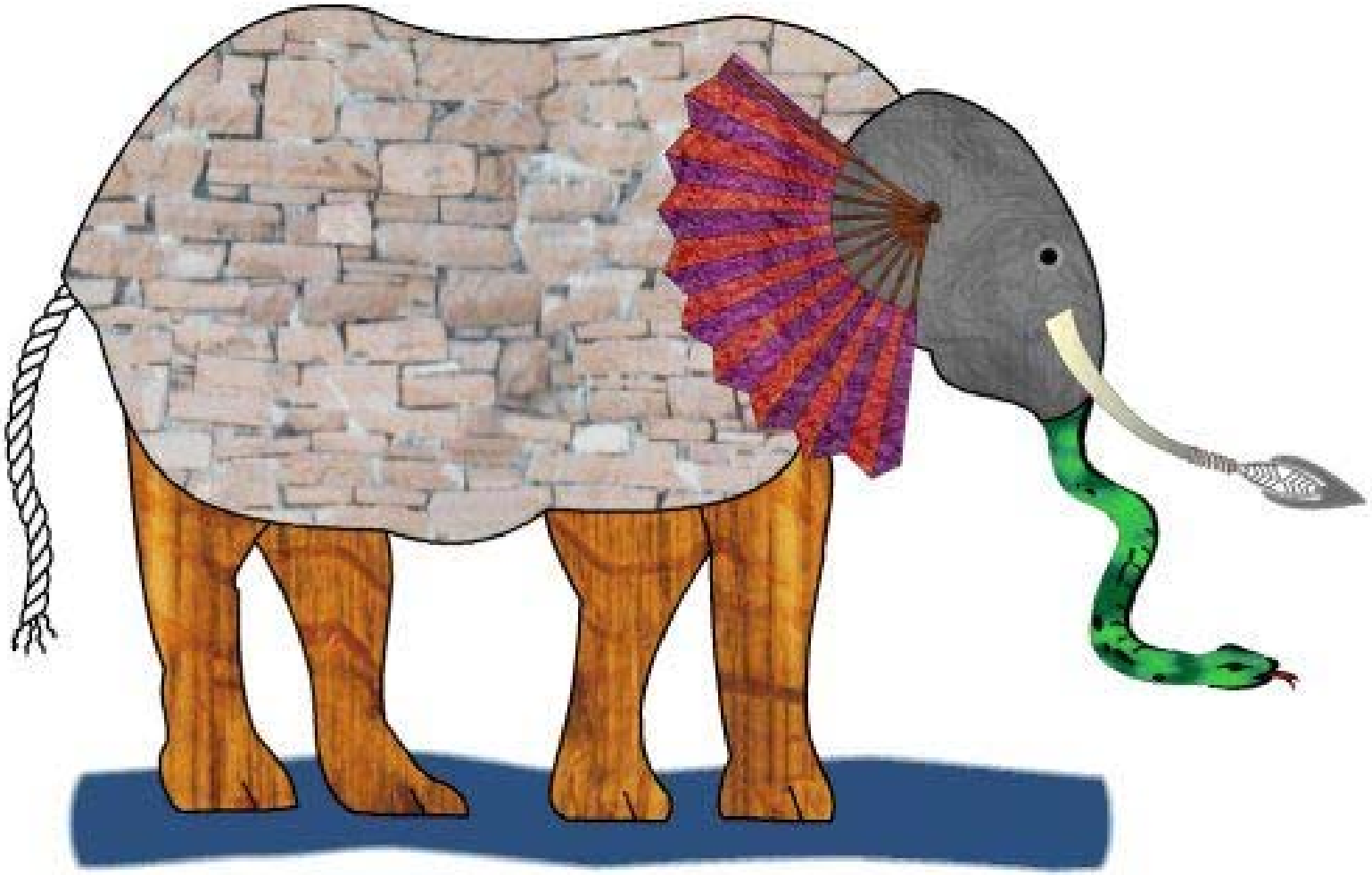
Lecture Outline

- **Introduction to Evolutionary Genomics**
- **Evolutionary and Functional Genomics**

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

- **Phylogenomics**

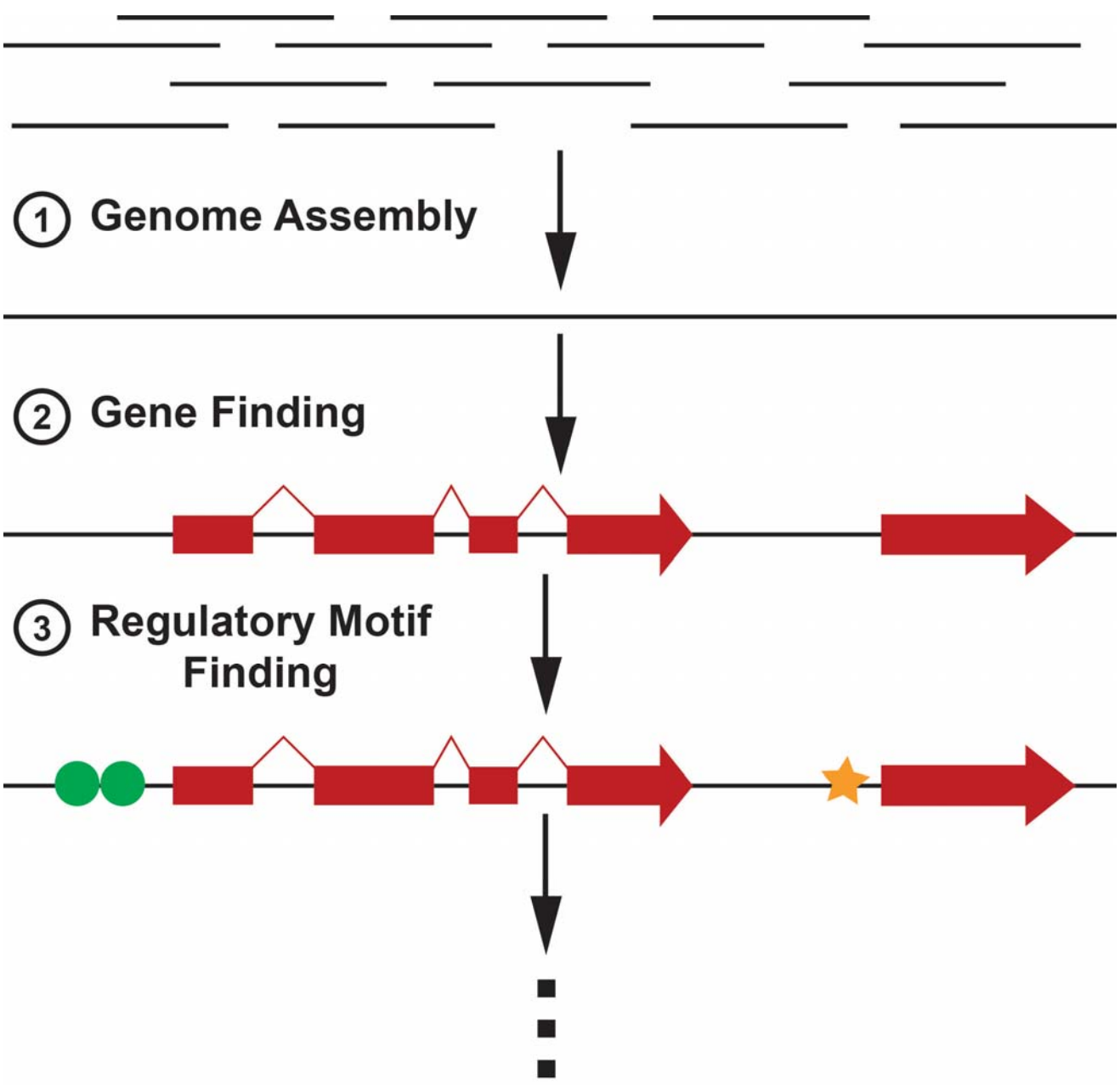
What is an Elephant Like?



What is a Genome Like?

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

Understanding the Genome Requires Tools

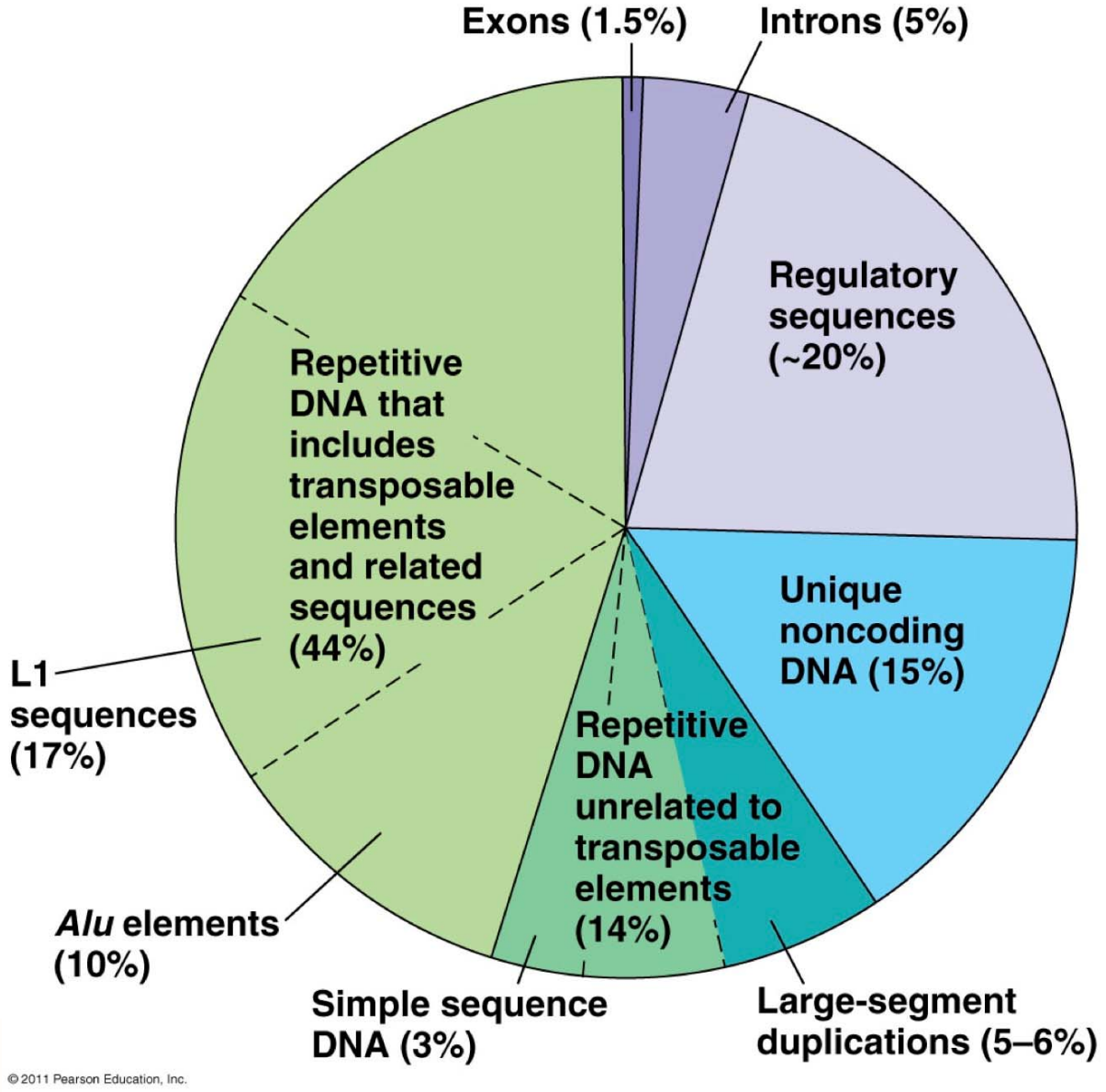


What is a Genome Like?

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

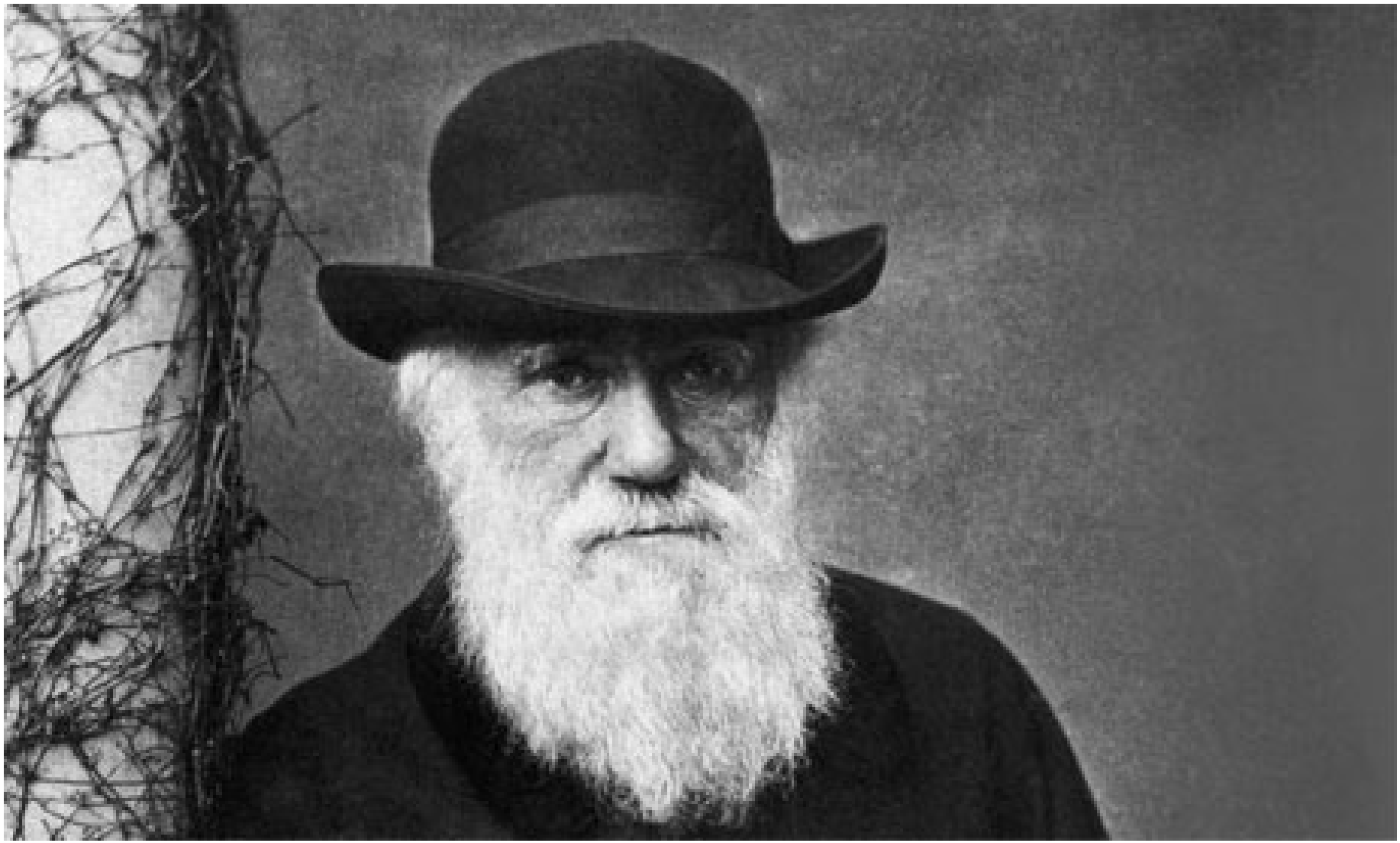
Transposon **Protein Binding Site** **Exon** **Intron**

Organization of the Human Genome

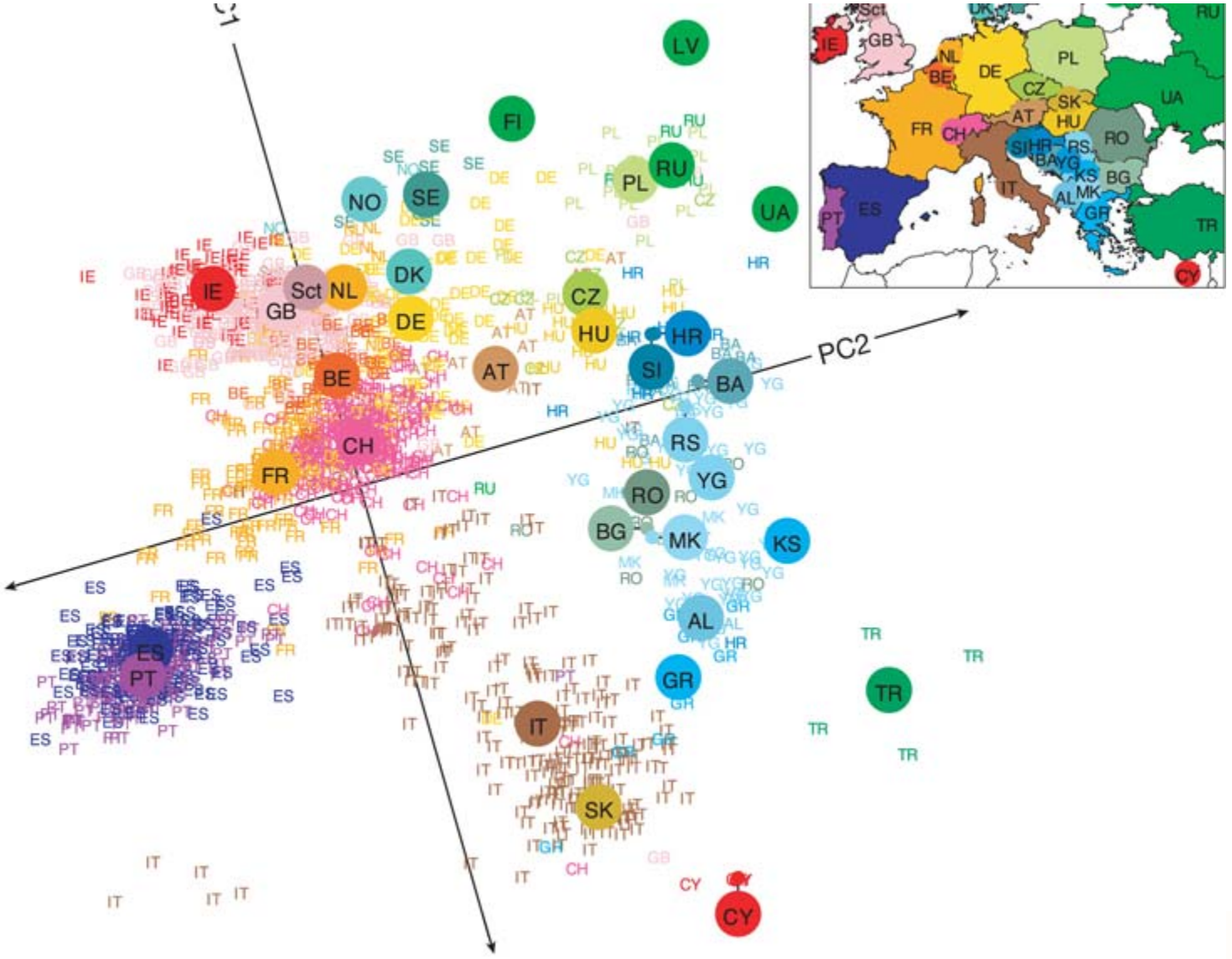


http://todd.jackman.villanova.edu/HumanGenome.jpg

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

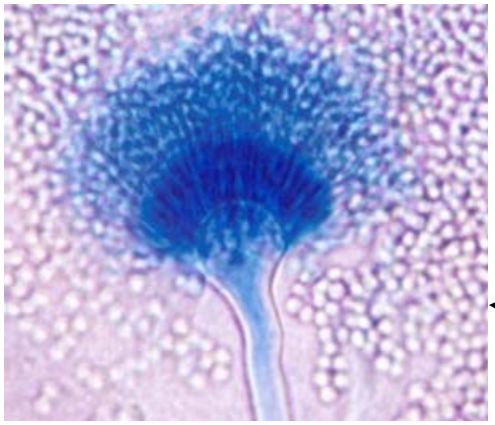
Chr:position (MB, HG17)	Selected population	Long Haplotype Test	Size (Mb)	Total SNPs with Long Haplotype Signal	Subset of SNPs that fulfil criteria 1	Subset of SNPs that fulfil criteria 1 and 2	Subset of SNPs that fulfil criteria 1, 2 and 3	Genes at or near SNPs that fulfil all three criteria
chr1:166	CHB + JPT	LRH, iHS	0.4	92	39	30	2	*BLZF1, SLC19A2*
chr2:72.6	CHB + JPT	XP-EHH	0.8	732	250	0	0	
chr2:108.7	CHB + JPT	LRH, iHS, XP-EHH	1.0	972	265	7	1	*EDAR*
chr2:136.1	CEU	LRH, iHS, XP-EHH	2.4	1,213	282	24	3	*RAB3GAP1, R3HDM1, LCT*
chr2:177.9	CEU,CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	*PDE11A*
chr4:33.9	CEU,YRI, CHB + JPT	LRH, iHS	1.7	413	161	33	0	
chr4:42	CHB + JPT	LRH, iHS, XP-EHH	0.3	249	94	65	6	*SLC30A9*
chr4:159	CHB + JPT	LRH, iHS, XP-EHH	0.3	233	67	34	1	
chr10:3	CEU	LRH, iHS, XP-EHH	0.3	179	63	16	1	
chr10:22.7	CEU, CHB + JPT	XP-EHH	0.3	254	93	0	0	
chr10:55.7	CHB+JPT	LRH, iHS, XP-EHH	0.4	735	221	5	2	*PCDH15*
chr12:78.3	YRI	LRH, iHS	0.8	151	91	25	0	
chr15:46.4	CEU	XP-EHH	0.6	867	233	5	1	*SLC24A5*
chr15:61.8	CHB + JPT	XP-EHH	0.2	252	73	40	6	*HERC1*
chr16:64.3	CHB + JPT	XP-EHH	0.4	484	137	2	0	
chr16:74.3	CHB + JPT, YRI	LRH, iHS	0.6	55	35	28	3	*CHST5, ADAT1, KARS*
chr17:53.3	CHB + JPT	XP-EHH	0.2	143	41	0	0	
chr17:56.4	CEU	XP-EHH	0.4	290	98	26	3	*BCAS3*
chr19:43.5	YRI	LRH, iHS, XP-EHH	0.3	83	30	0	0	
chr22:32.5	YRI	LRH	0.4	318	188	35	3	*LARGE*
chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	
	Total SNPs		16.74	9,166	2,898	480	41	

In the European population, involved in skin pigmentation

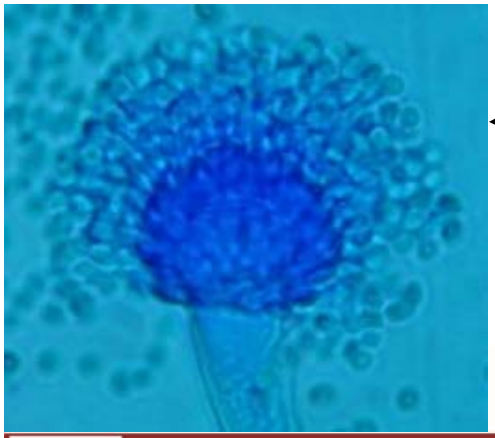
In the West African population, related to Lassa virus infection

Sabeti et al. (2007) Nature

Genomes Provide a Common Yardstick for Comparison



Morphologically very similar!



Average proteome sequence similarity

60
70
80
90
100

A. nidulans/niger
A. oryzae/terreus
A. clavatus
N. fischeri
A. fumigatus

Kluyveromyces lactis
Candida glabrata
S. uvarum

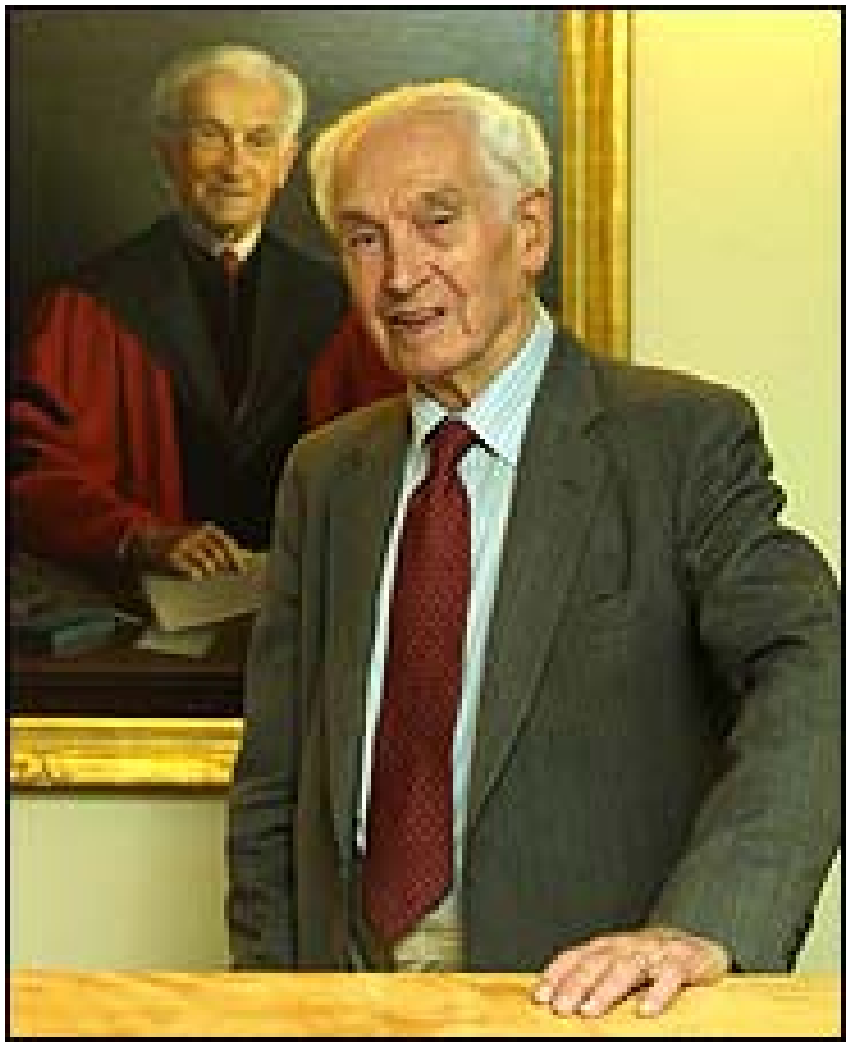
They Interbreed!

S. paradoxus
S. cerevisiae

fish
birds
mice
humans

“…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

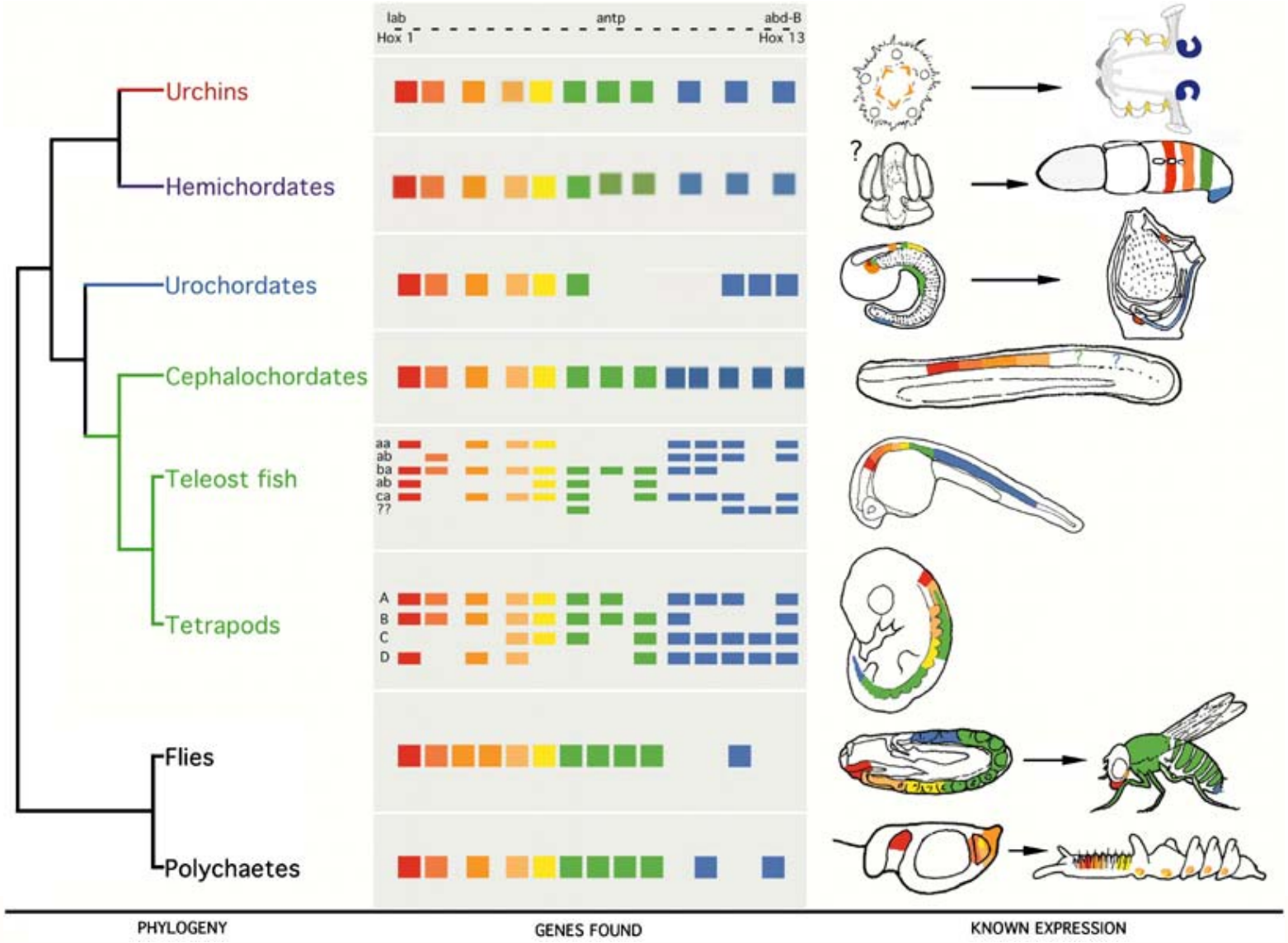
F	W	Y	Cancer
+	+	-	ABL1
+	-	-	Acute Myeloid Leukemia-DEK
+	-	-	Adenomat. Polyposis Coli-APC
+	+	+	AKT2
+	+	+	Ataxia Telangiectasia-ATM
-	-	-	BRCA1
-	-	-	BRCA2
+	+	-	Basal Cell Nevus-PTC
+	-	-	B-Cell Lymphoma 2-BCL2
-	-	-	B-Cell Lymphoma 3-BCL3
+	+	-	Bloom-BLM
+	-	-	Burkitt's Lymphoma-MYC
-	-	-	CDKN2C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

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

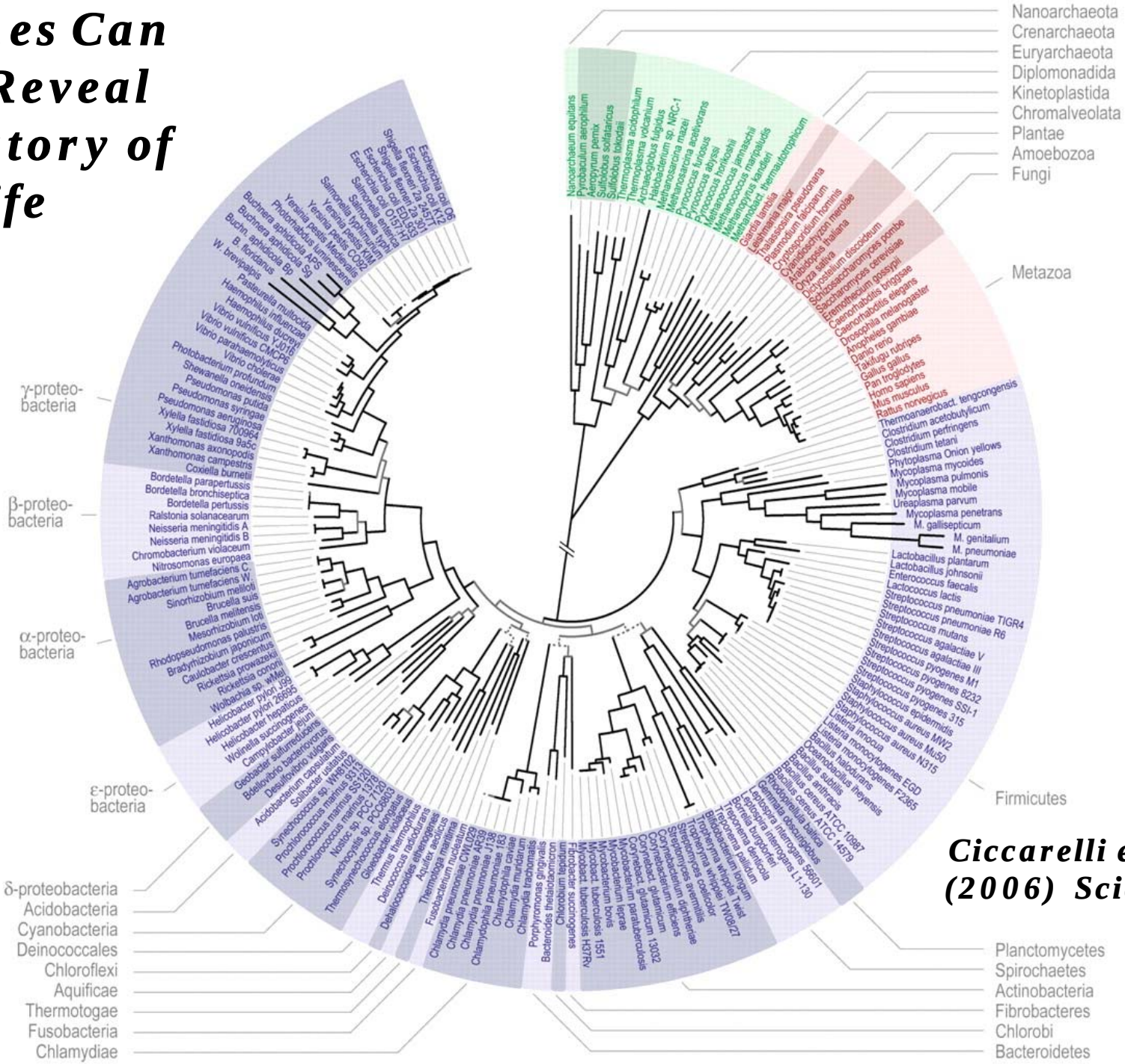
F	W	Y	Malformation Syndromes
-	+	-	Aarskog-Scott-FGD1
+	+	-	Achondroplasia-FGFR3
+	+	-	Alagille-JAG1
+	+	+	Barth-TAZ
-	-	-	Beckwith-Wiedemann-CDKN1C
-	-	-	Cerebral Cavern. Malf.-CCM1
+	-	-	Chondrodyspl. Punct. 1-ARSE
+	+	-	Cleidocranial Dysplasia-OFC1
-	-	-	Cockayne I-CKN1
+	+	-	Coffin-Lowry-RPS6KA3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

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

Animal Bodies are Built from the Same Genetic Toolkit

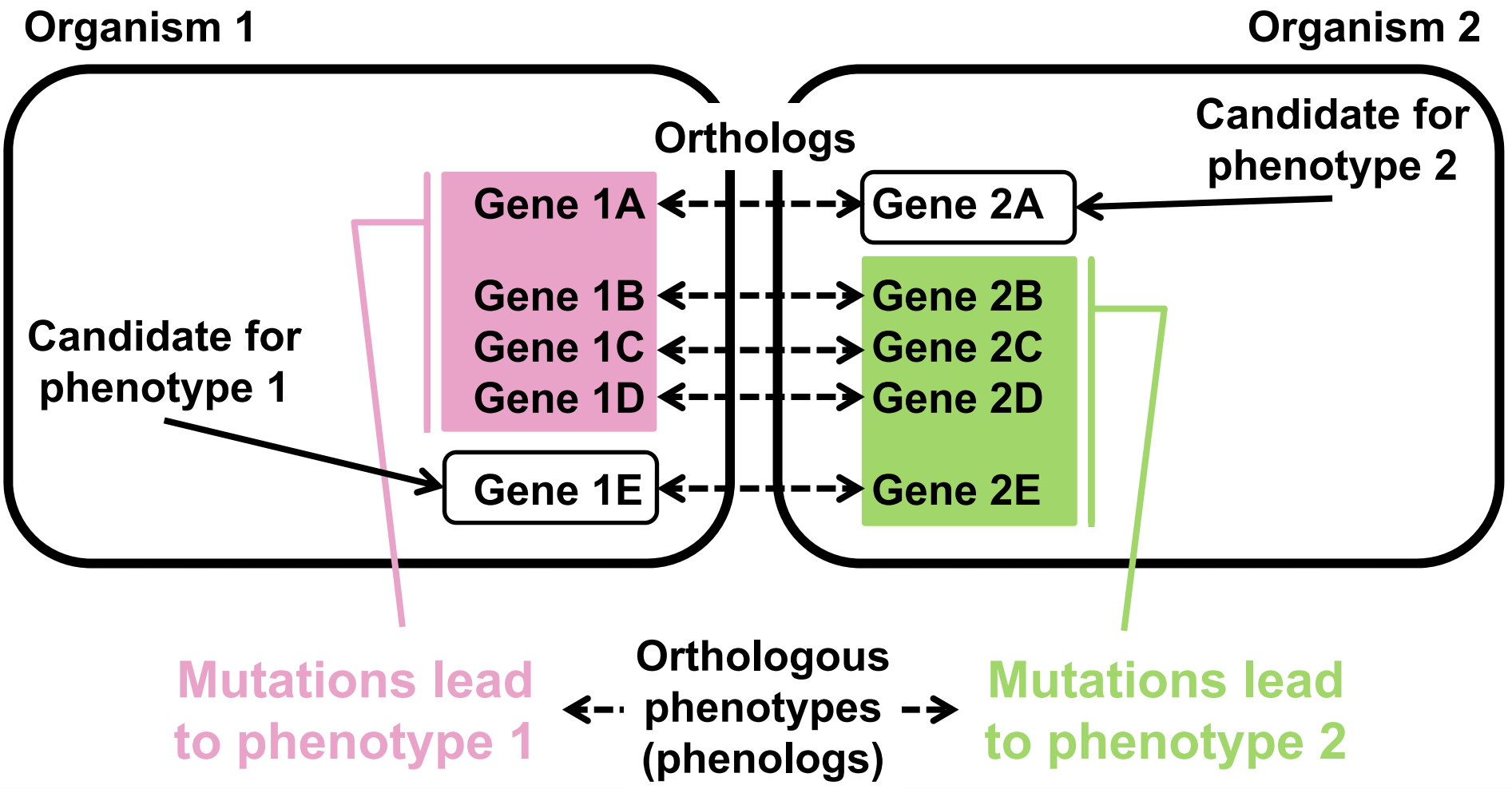


Genomes Can Help Reveal the History of Life



Ciccarelli et al. (2006) Science

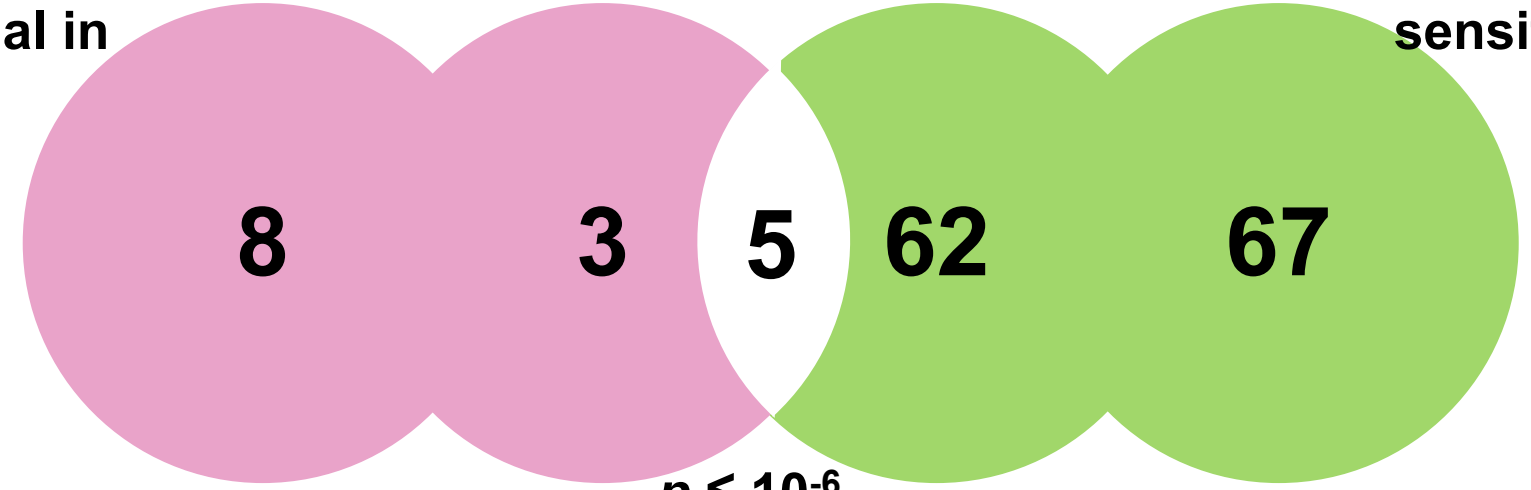
Evolution-Informed Analyses Have Great Predictive Power



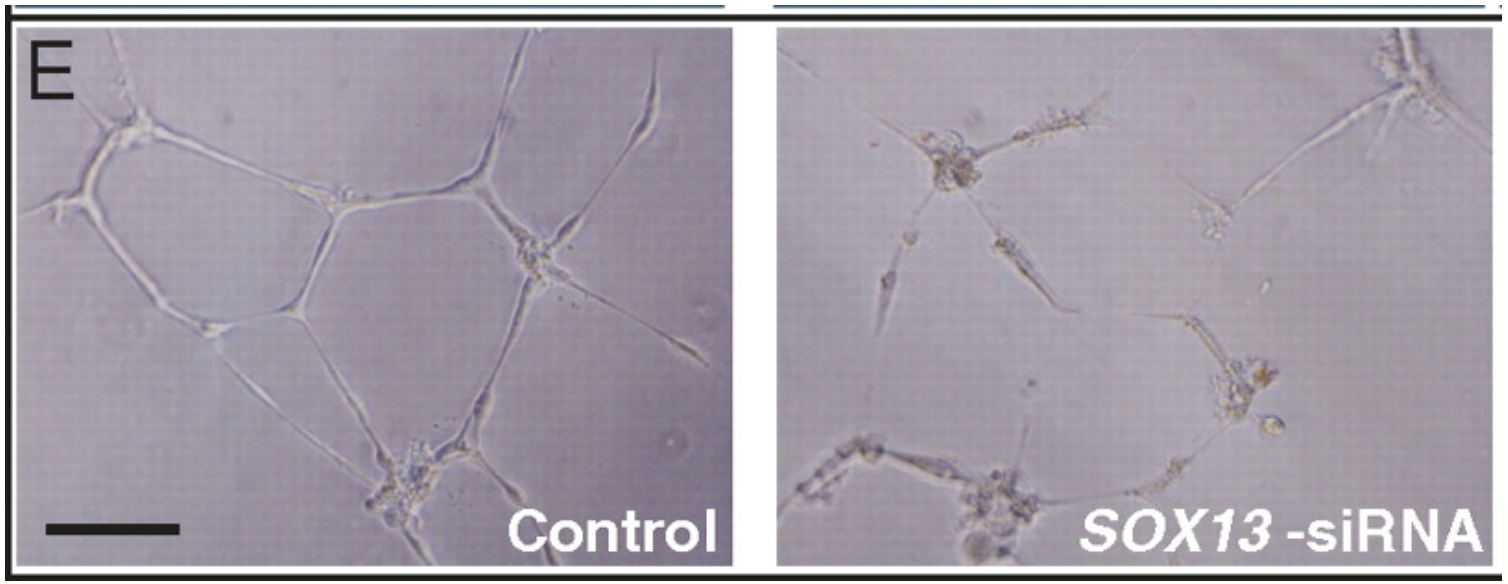
A Yeast Model for Angiogenesis

Angiogenesis abnormal in mice

Lovastatin sensitive in yeast



$p \leq 10^{-6}$



Genomics Used to Be "Big Science"

wellcome trust sanger institute

BCM
Baylor College of Medicine

BROAD INSTITUTE

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

J. Craig Venter INSTITUTE

THE Genome CENTER
AT WASHINGTON UNIVERSITY

High-Throughput DNA Sequencing Technologies

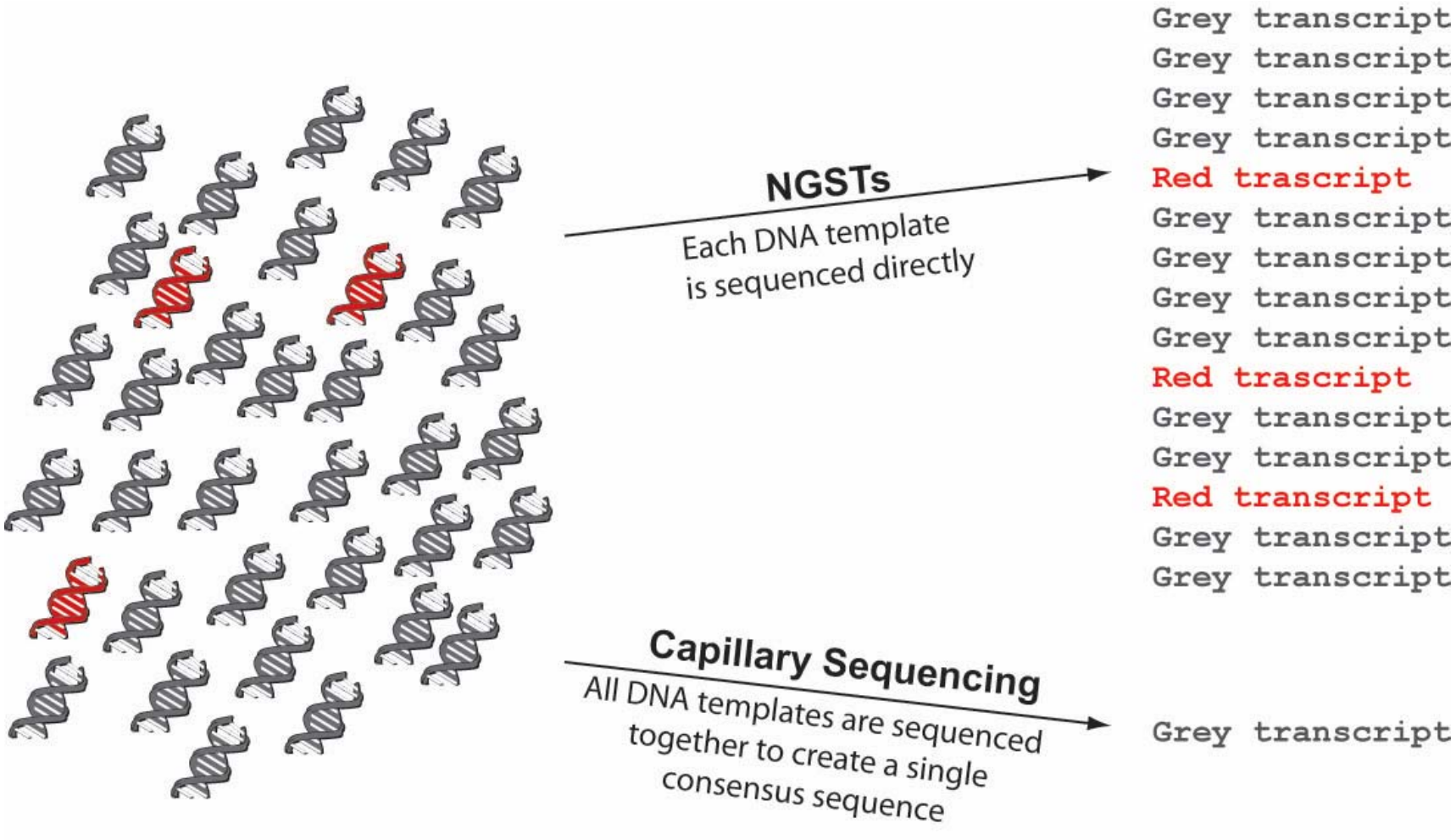
Illumina

454 / Roche

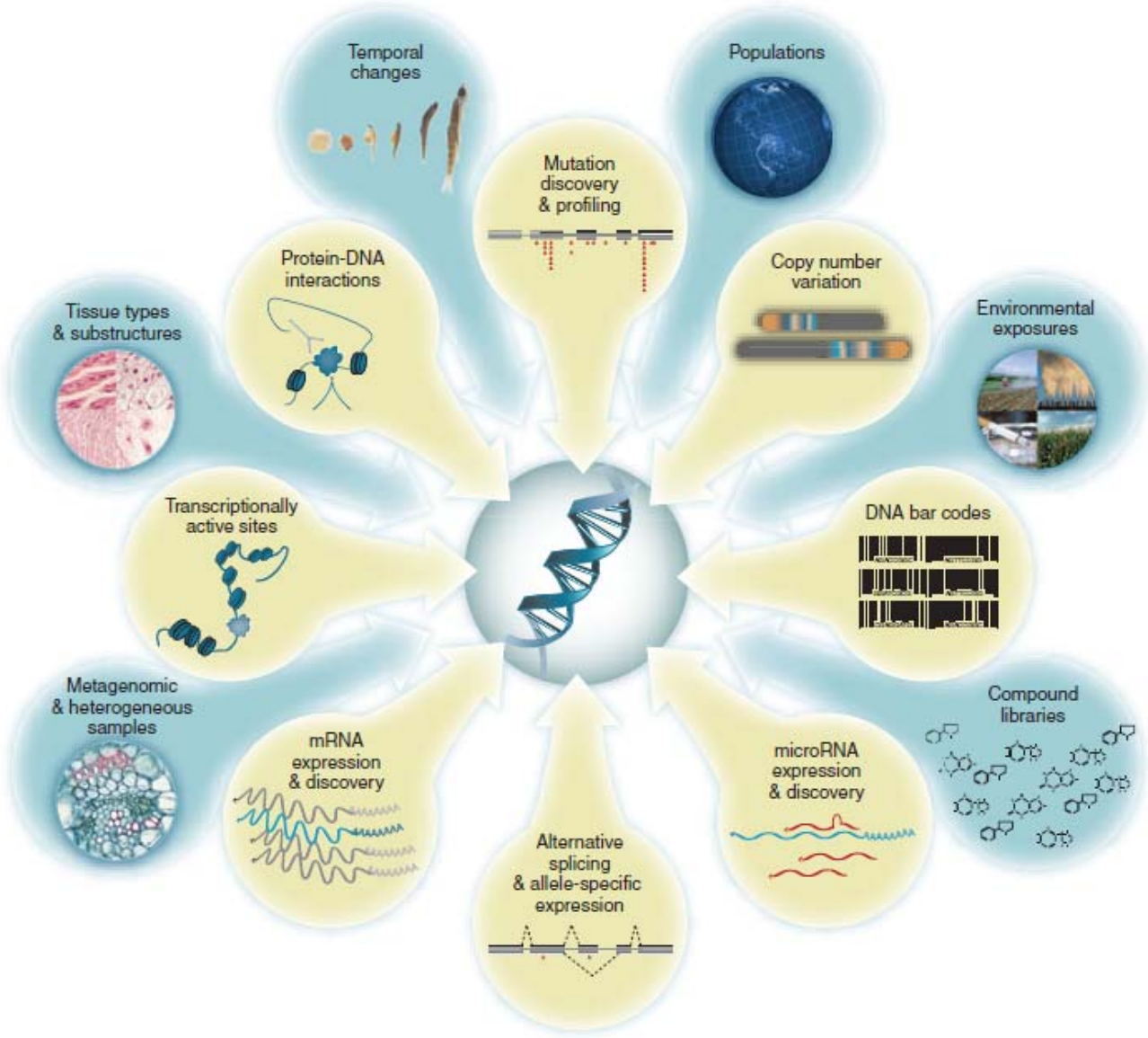
PacBio

AB / Life Technologies

High Throughput Sequencing is Qualitative and Quantitative

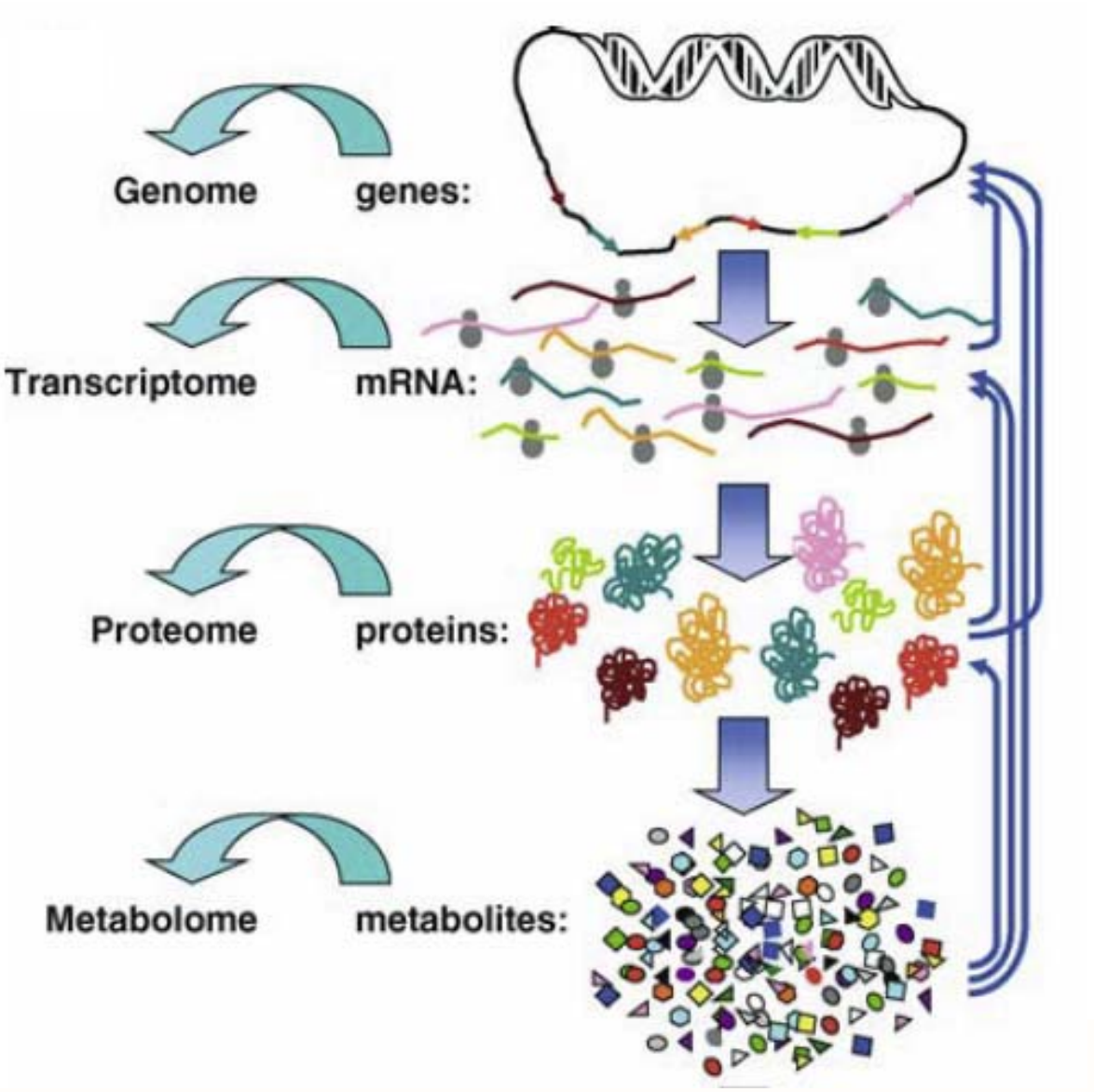


Sequence Census Assays



Kahvejian et al . (2008) Nature Biotech.

The Age of High Throughput Technologies



Goodacre (2005) Metabolomics

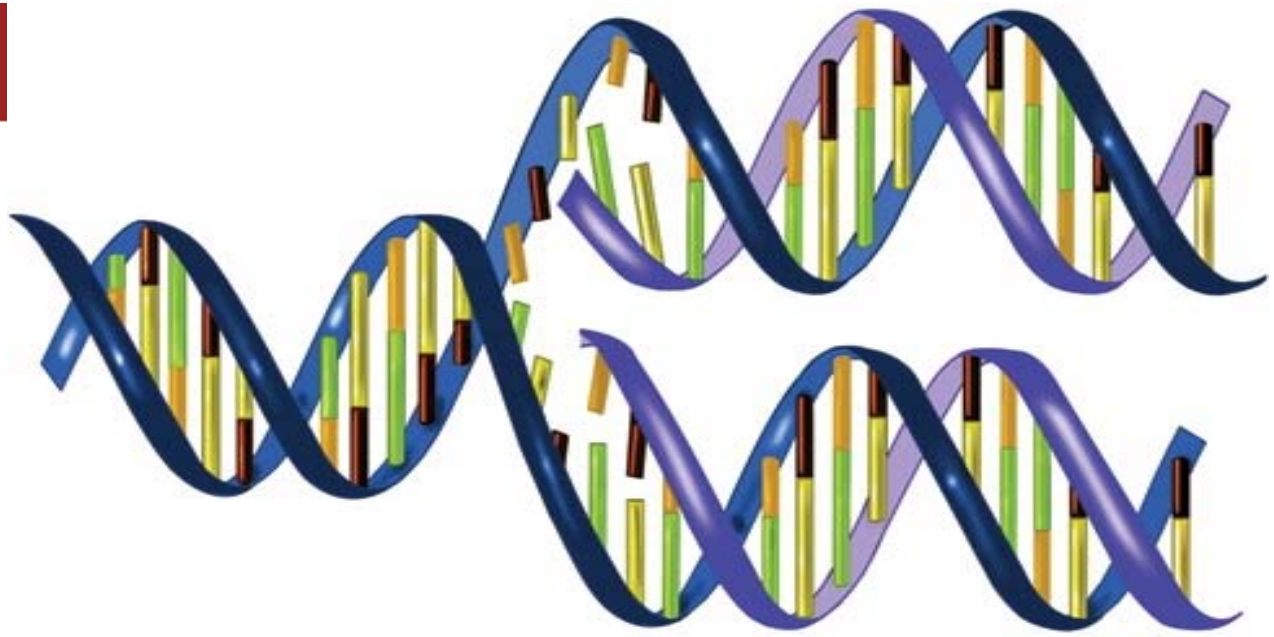
Novel Ways to Probe Gene Function in Any Organism

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.”

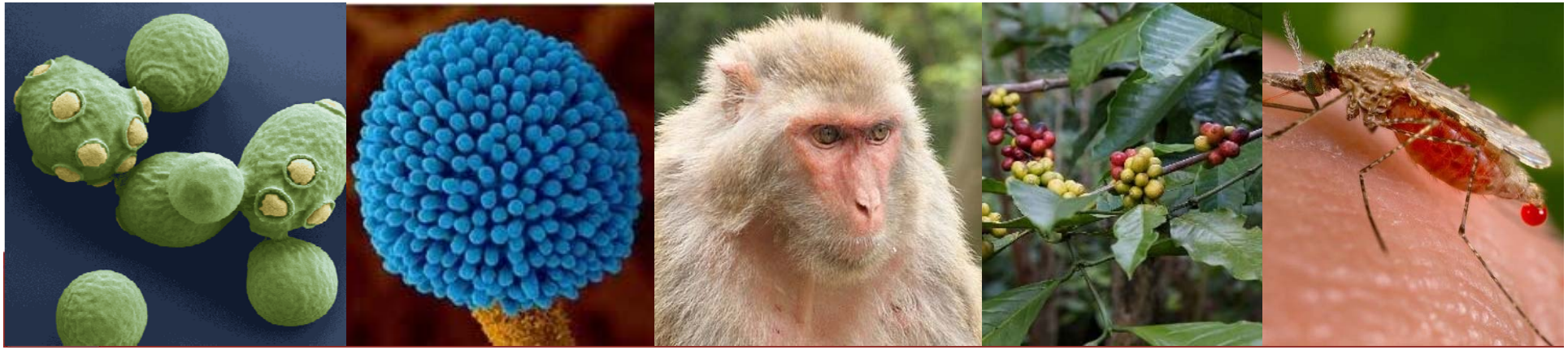
Eric Lander

The Rokas Lab

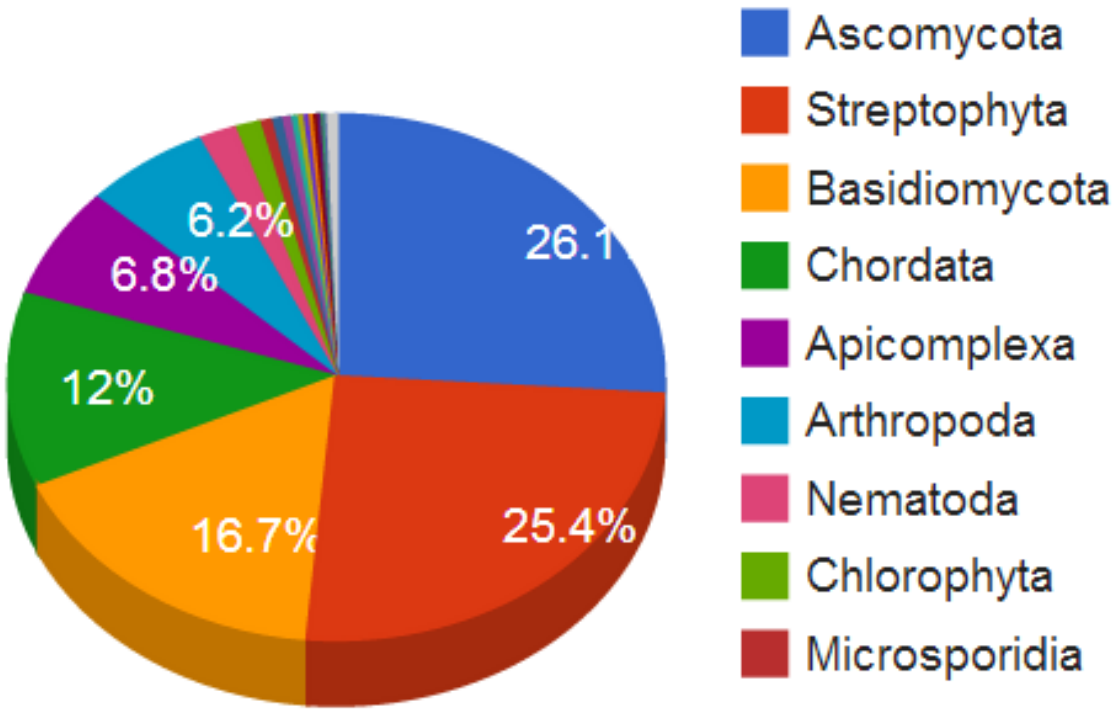
We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches

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



GOLD

Genomes Online Database

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."

Plasterk (1999) Nature Genet.

“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"

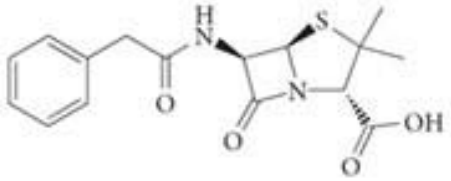
Dr. James Scott, Mycologist
***Wired* Magazine, 06/11 issue**

What Fungi Cannot Eat!

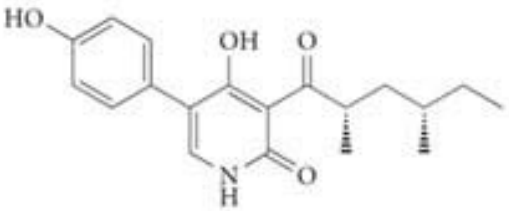
Diner fries

McDonald's fries

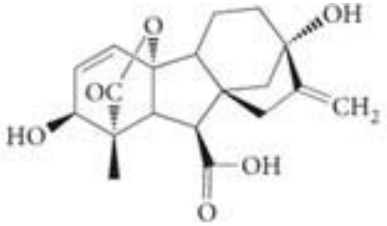
Fungi are Superb Chemical Engineers



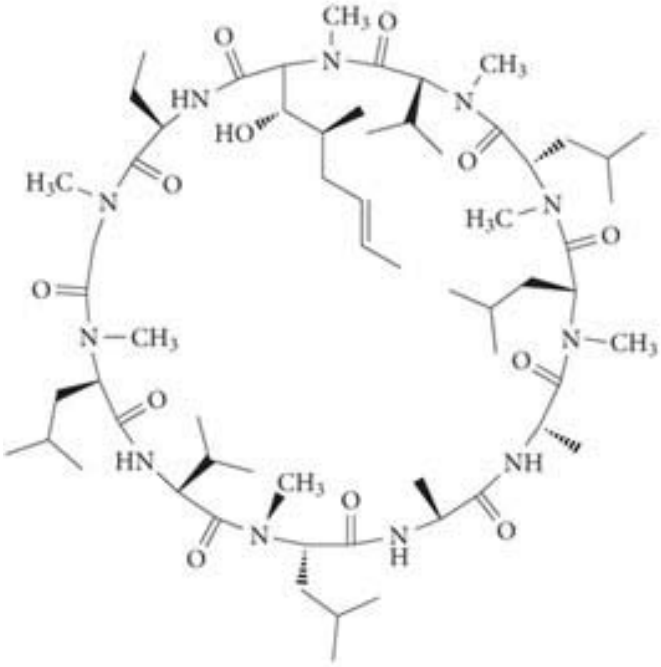
Penicillin G



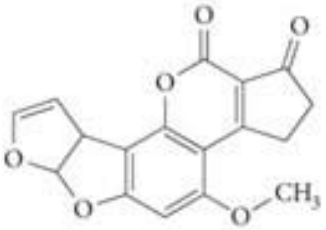
Aspyridone A



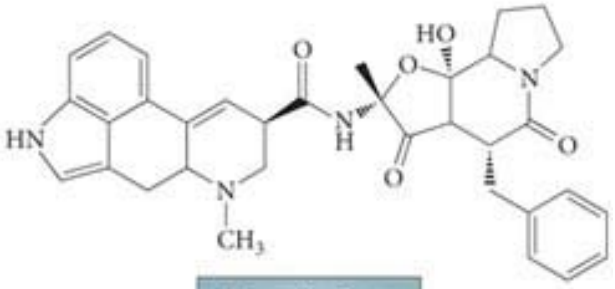
Gibberellin A3



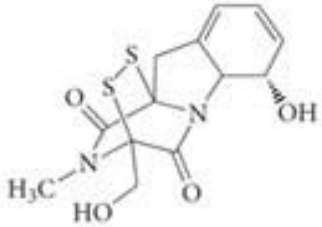
Cyclosporine A



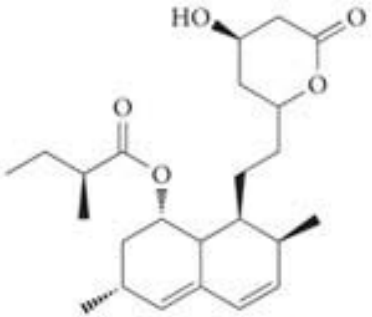
Aflatoxin B1



Ergotamine



Gliotoxin

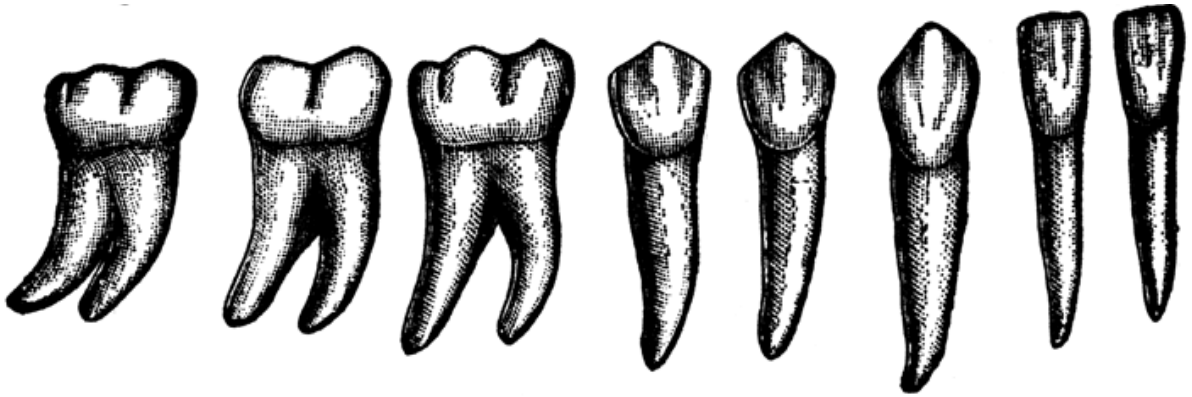


Lovastatin

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

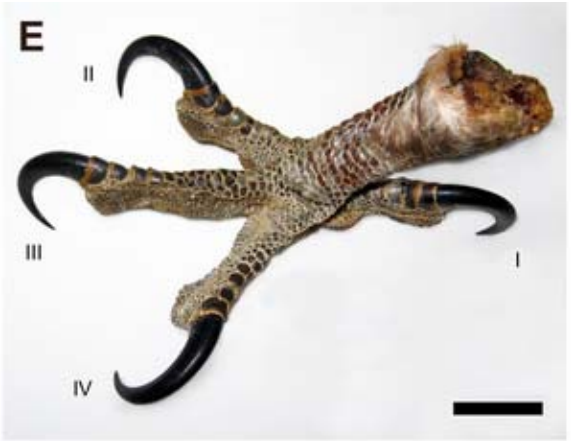
Chemodiversity is Fundamental to the Fungal Lifestyle

- **The genes involved in fungal primary metabolism are their "teeth"**



- **The genes involved in fungal secondary metabolism are their "horns", "spines", and "claws"**

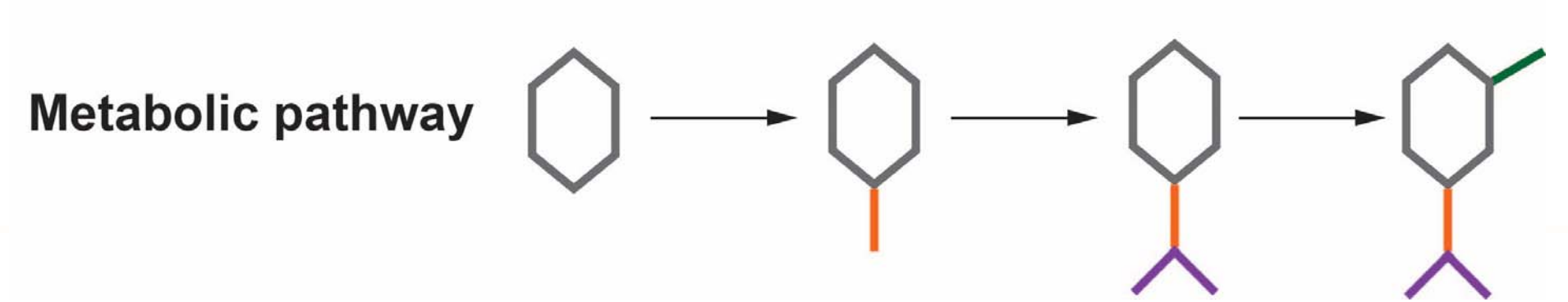
Beetle horns



Osprey claws

Echidna spines

Fungal Metabolic Genes Are Often Physically Clustered



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)

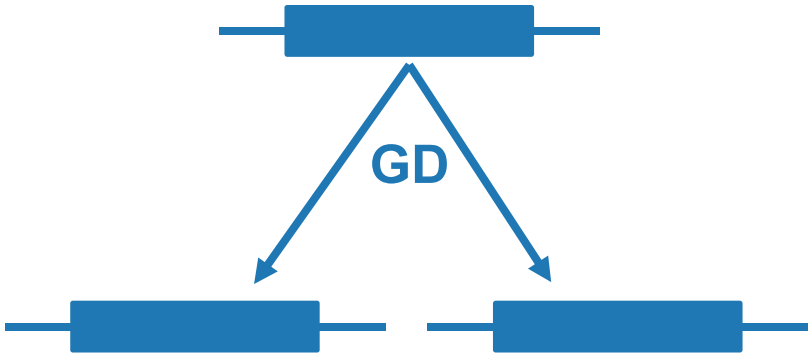
What contributed to the making of fungal chemodiversity?

- Genes and pathways
- Regulatory networks

Sources of Gene Innovation

Gene duplication (GD)

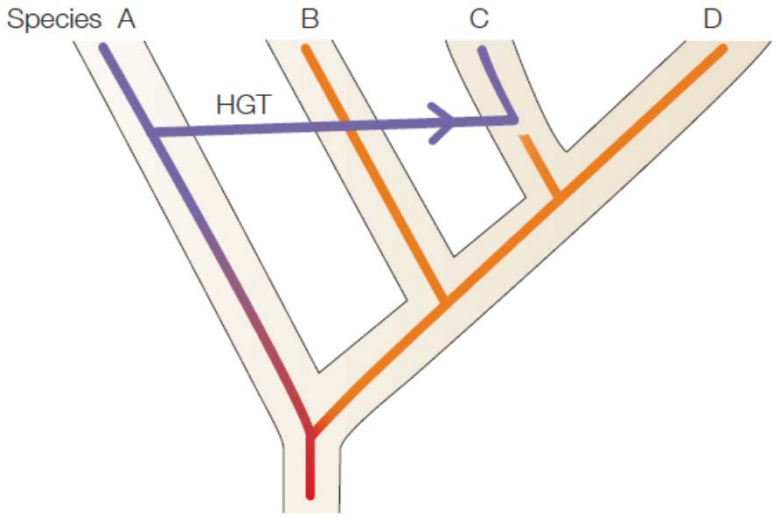
Any duplication of a region of DNA that contains a gene



- 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

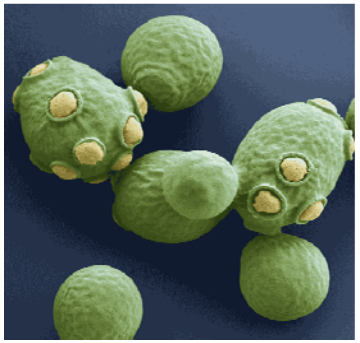


- Xenobiotic catabolism
- Toxin production
- Degradation of plant cell walls
- Wine fermentation

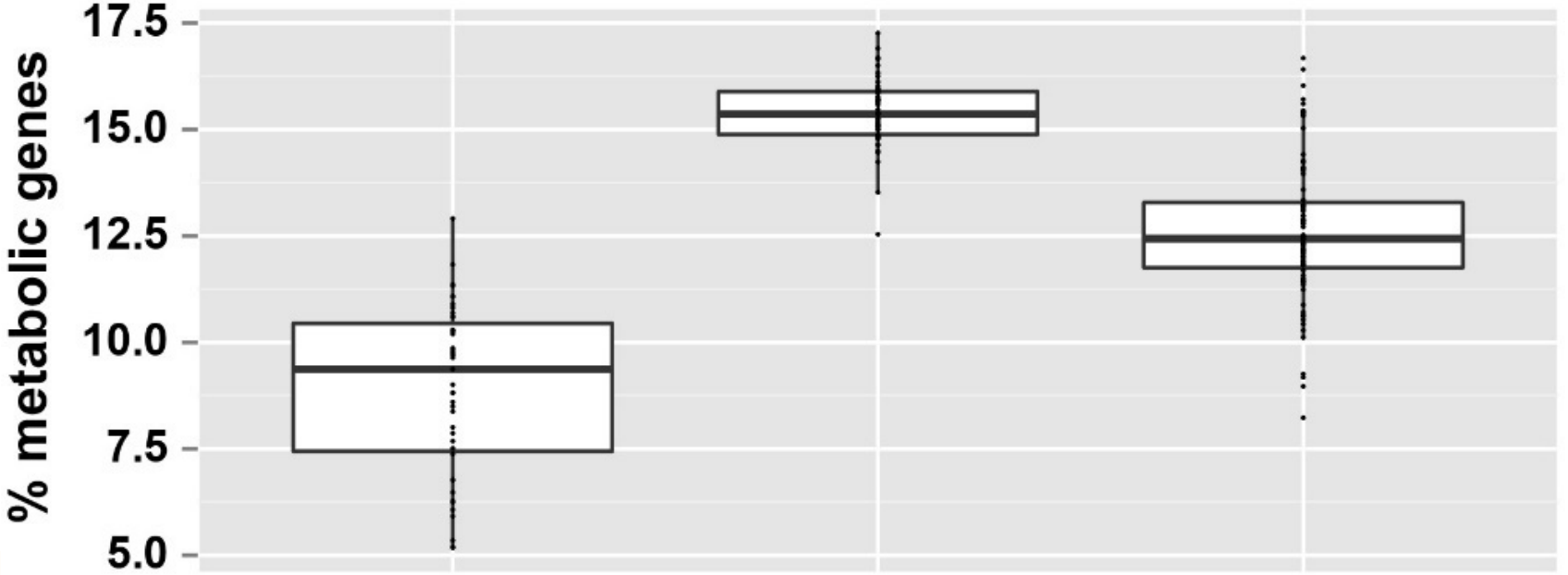
208 Genomes, 247,000 Genes, 875 Reactions, 130 Pathways

Agaricomycetes

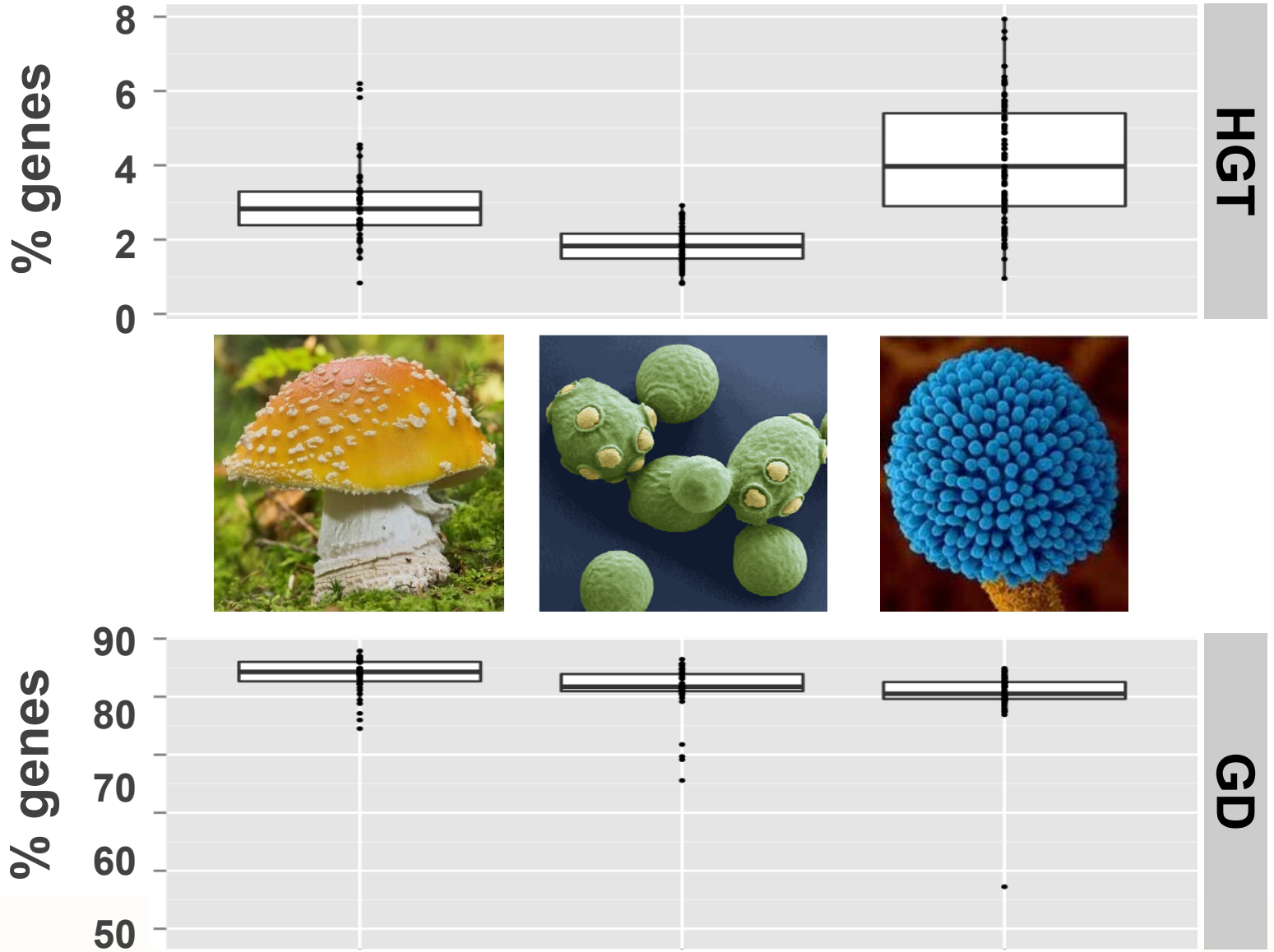
Saccharomycotina



Pezizomycotina



The Impact of GD is Large but of HGT Small

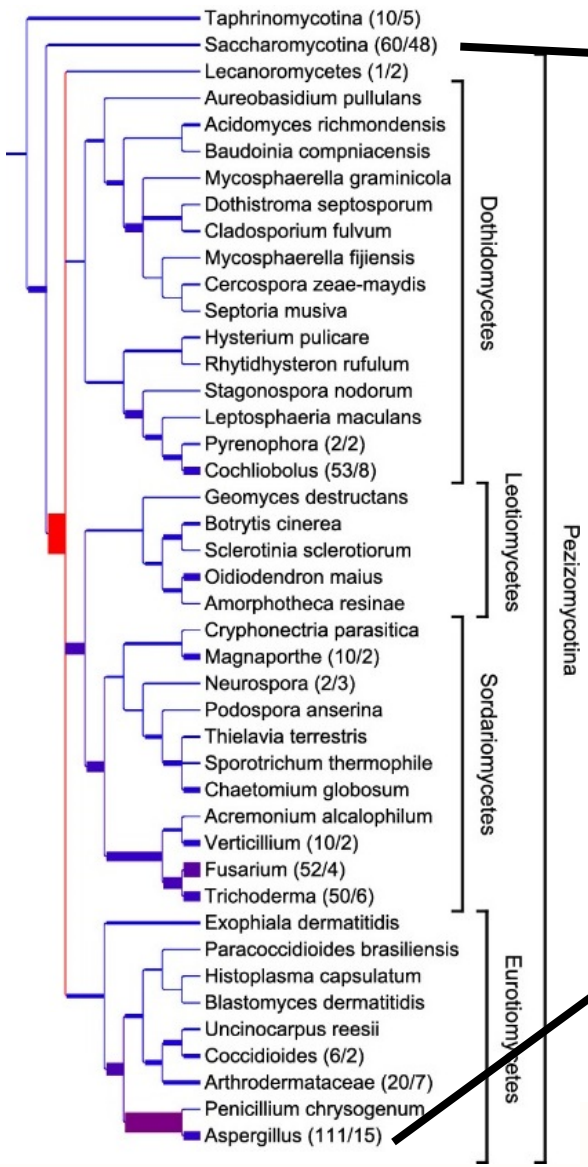


2.8% of metabolic genes / genome affected by HGT

88.7% of metabolic genes / genome affected by GD

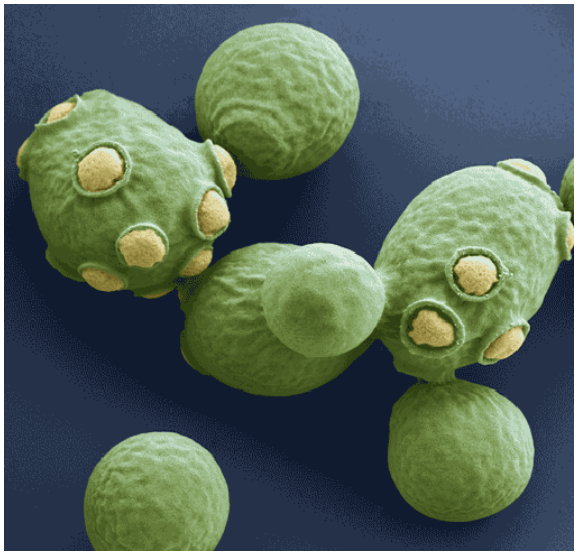
Wisecaver et al. (2014) PLoS Genet.

HGT is Episodic and Acts in a Lineage-Specific Manner



Saccharomycotina

60 HGT events in 48 genomes



Aspergillus

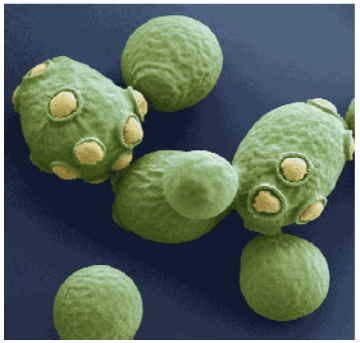
111 HGT events in 15 genomes

Wisecaver et al. (2014) PLoS Genet.

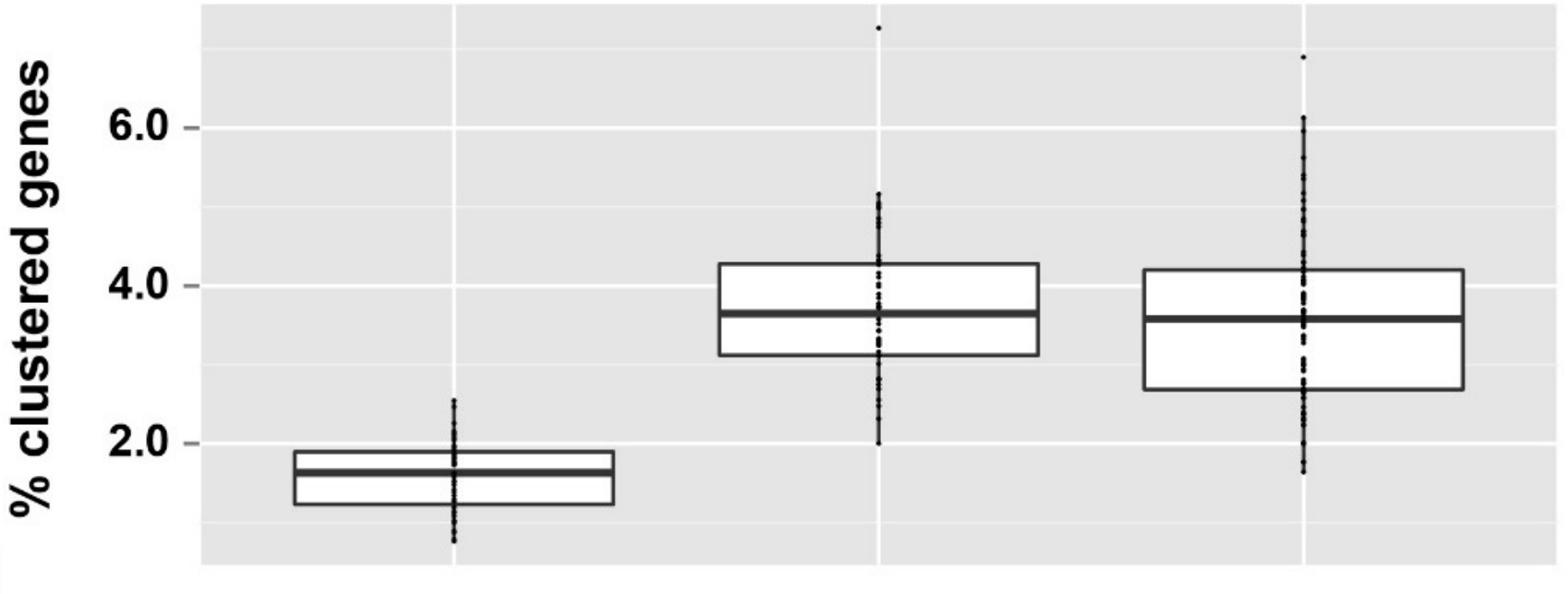
Lineages Vary With Respect to Gene Clustering

Agaricomycetes

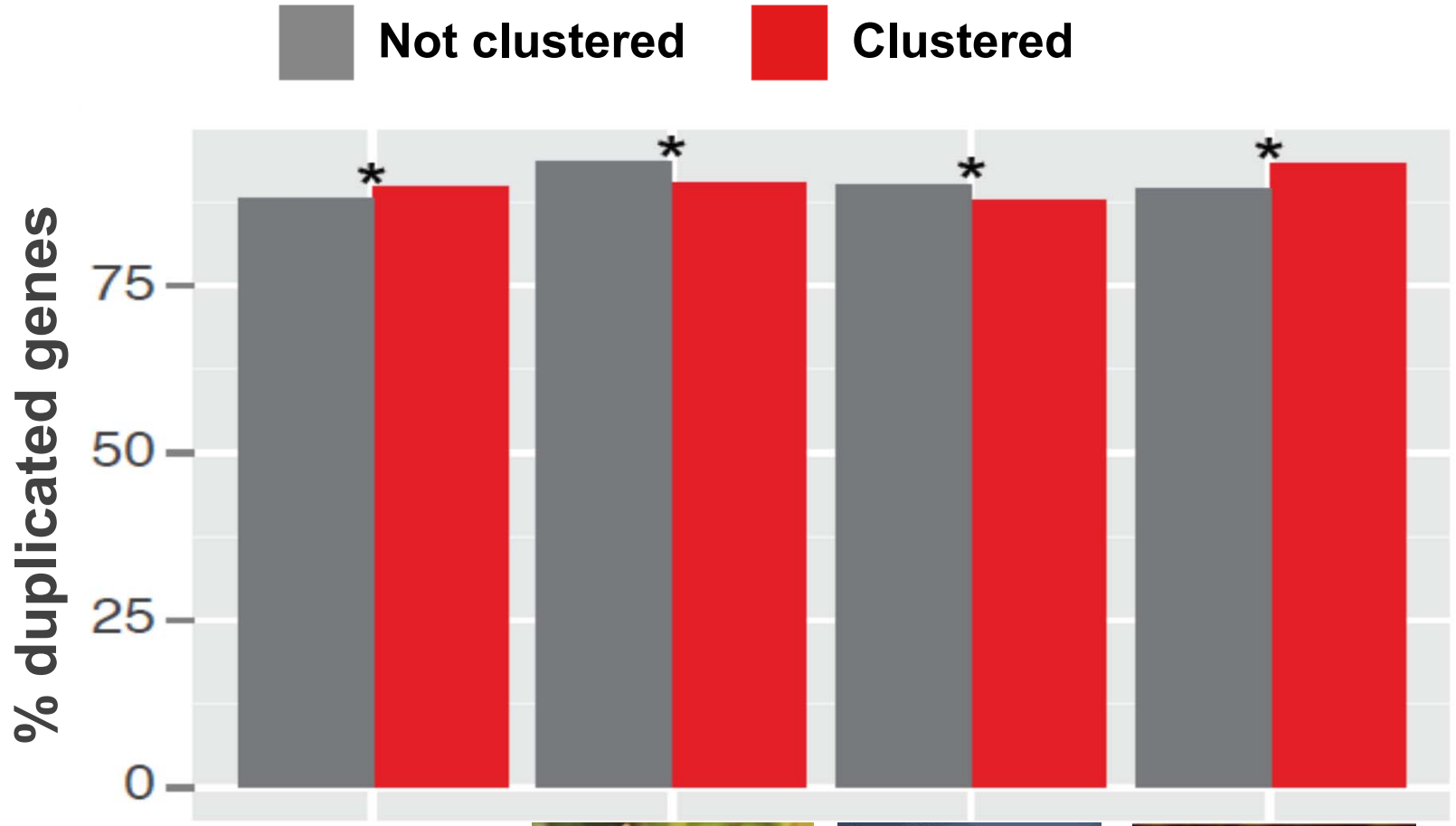
Saccharomycotina



Pezizomycotina



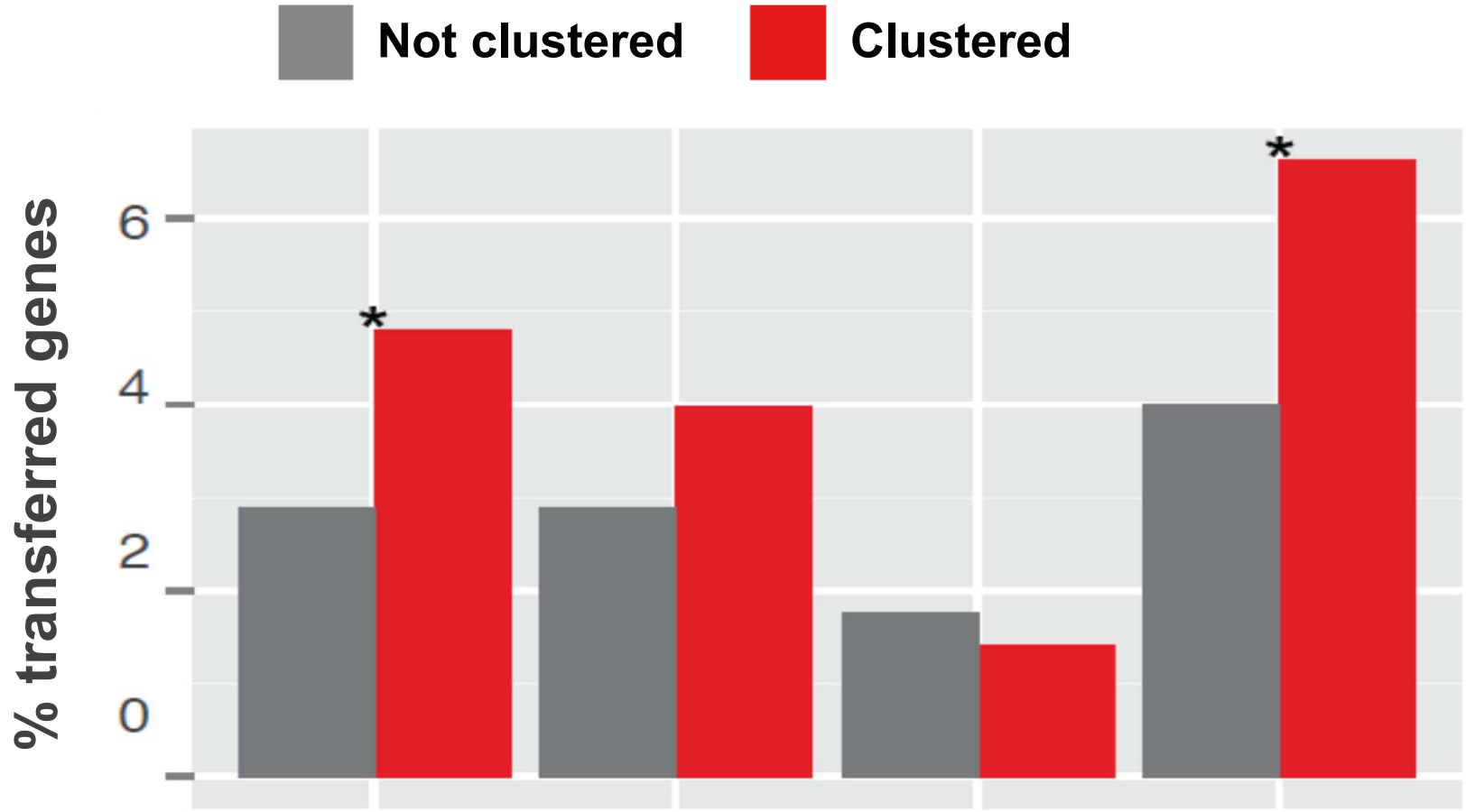
Does Clustering Correlate with GD?



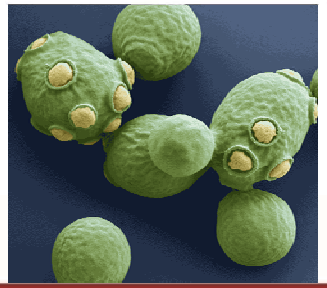
All Fungi

Wisecaver et al. (2014) PLoS Genet.

Does Clustering Correlate with HGT?



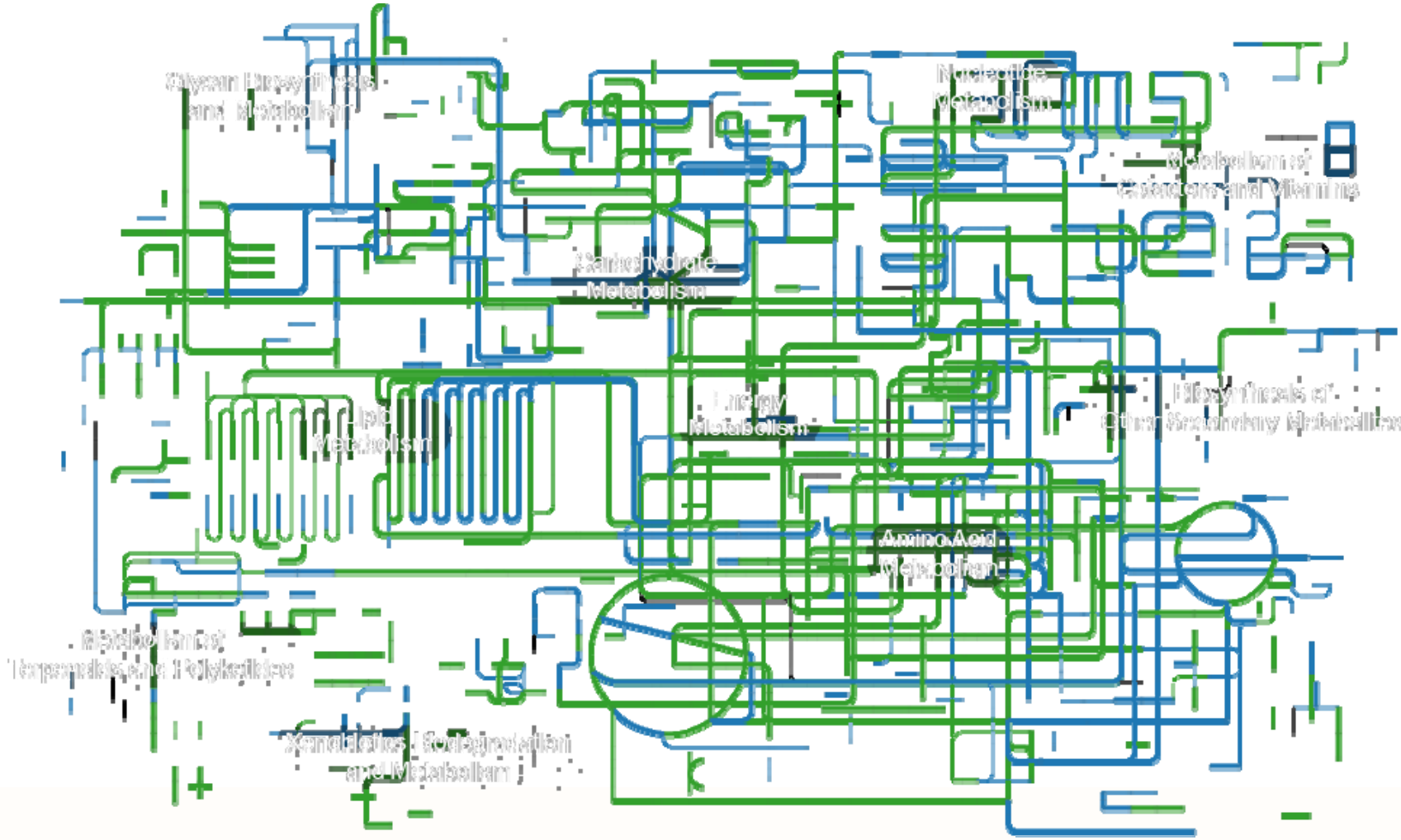
All Fungi



Gene Clustering, GD and HGT Across Fungi

Weight — Never clustered — Clustered

Color — No GD or HGT — GD — HGT — GD & HGT



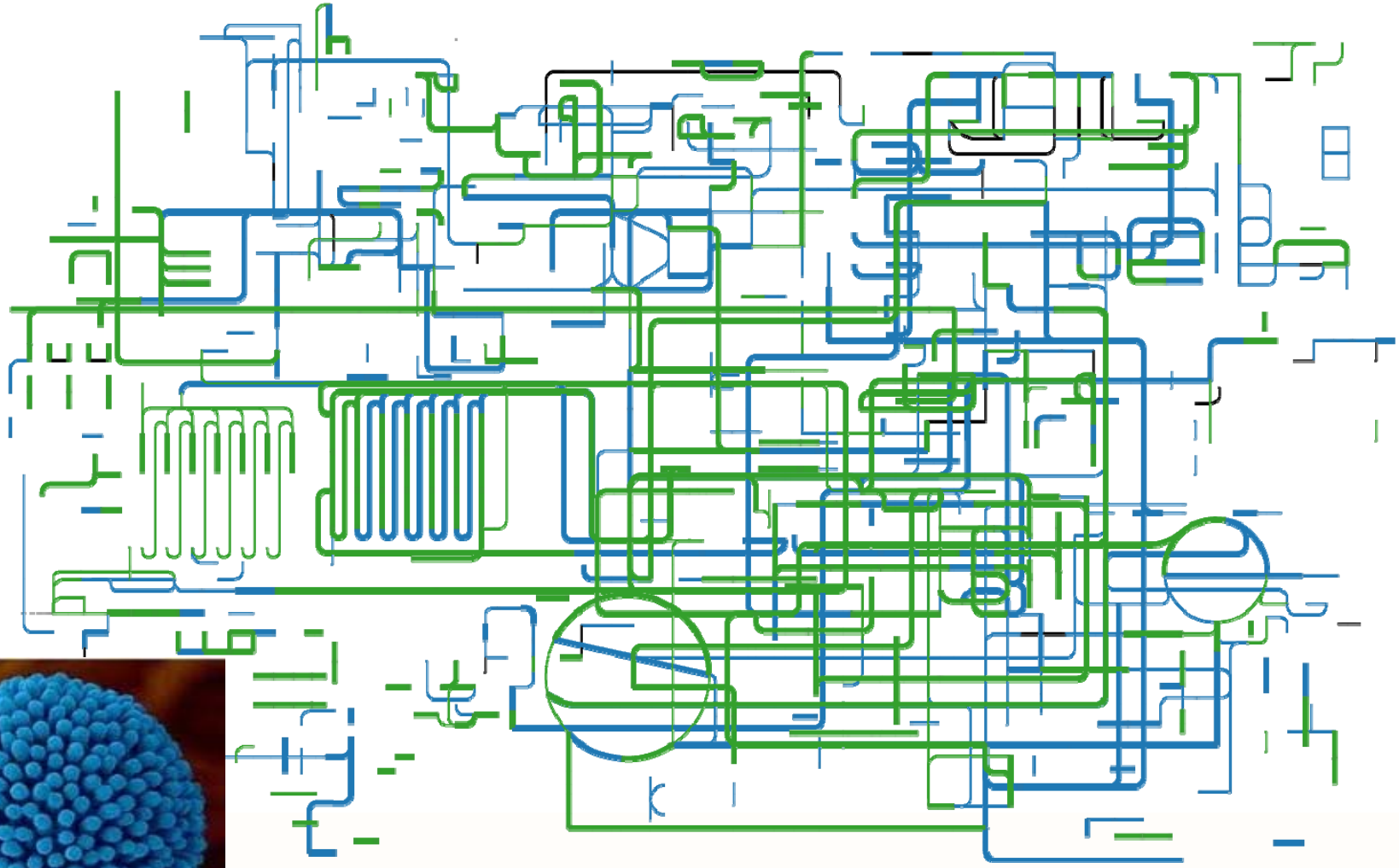
Filamentous Fungi: Clustered Genes Undergo GD and HGT

Weight

— Never clustered — Clustered

Color

No GD or HGT GD HGT GD & HGT



Wisecaver et al. (2014) PLoS Genet.

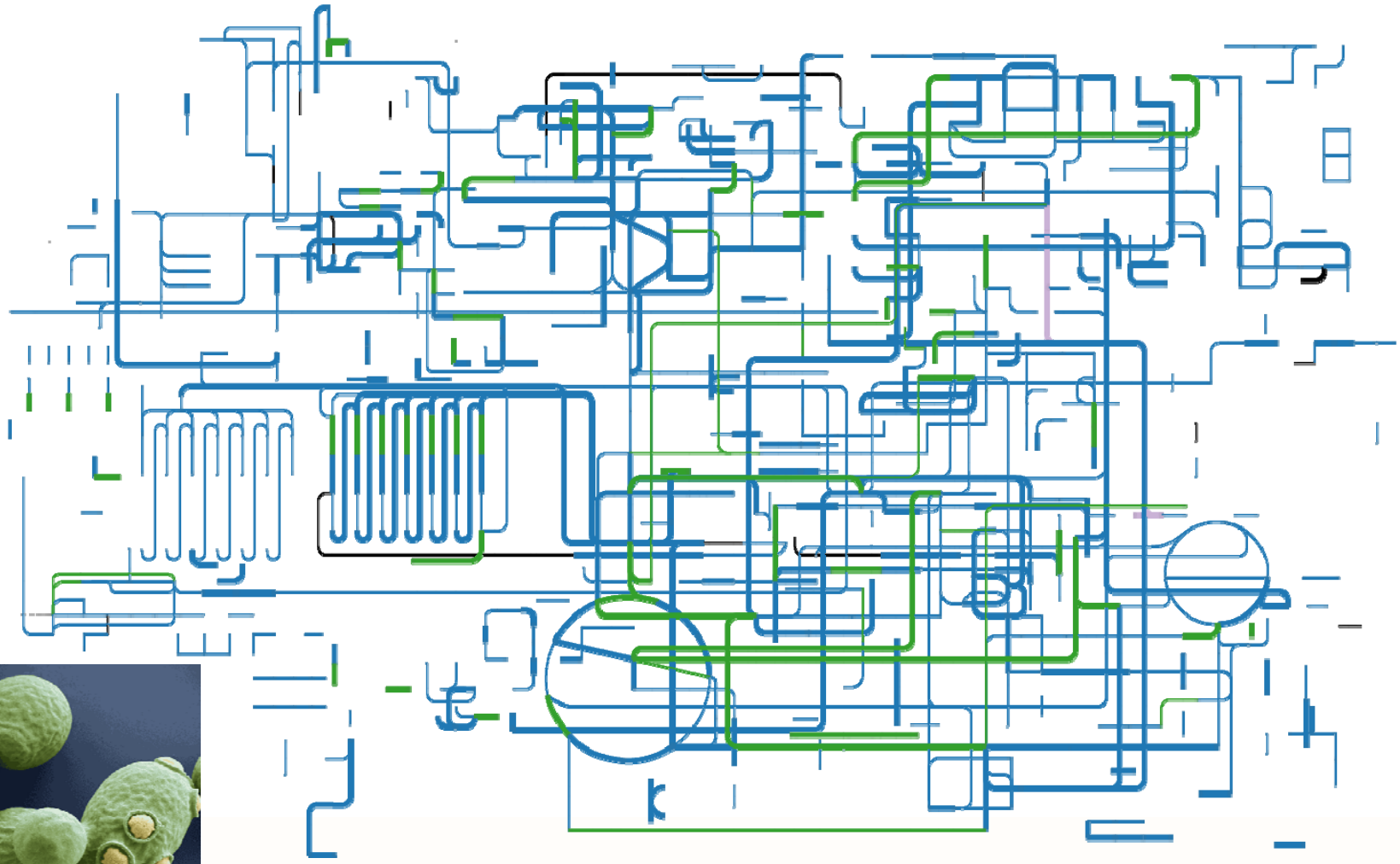
Yeasts: Genes are Clustered but Rarely Undergo HGT

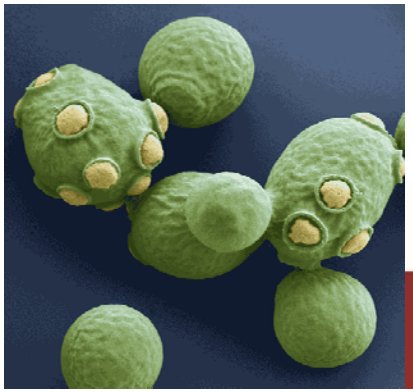
Weight

— Never clustered — Clustered

Color

No GD or HGT GD HGT GD & HGT





Wisecaver et al. (2014) PLoS Genet.

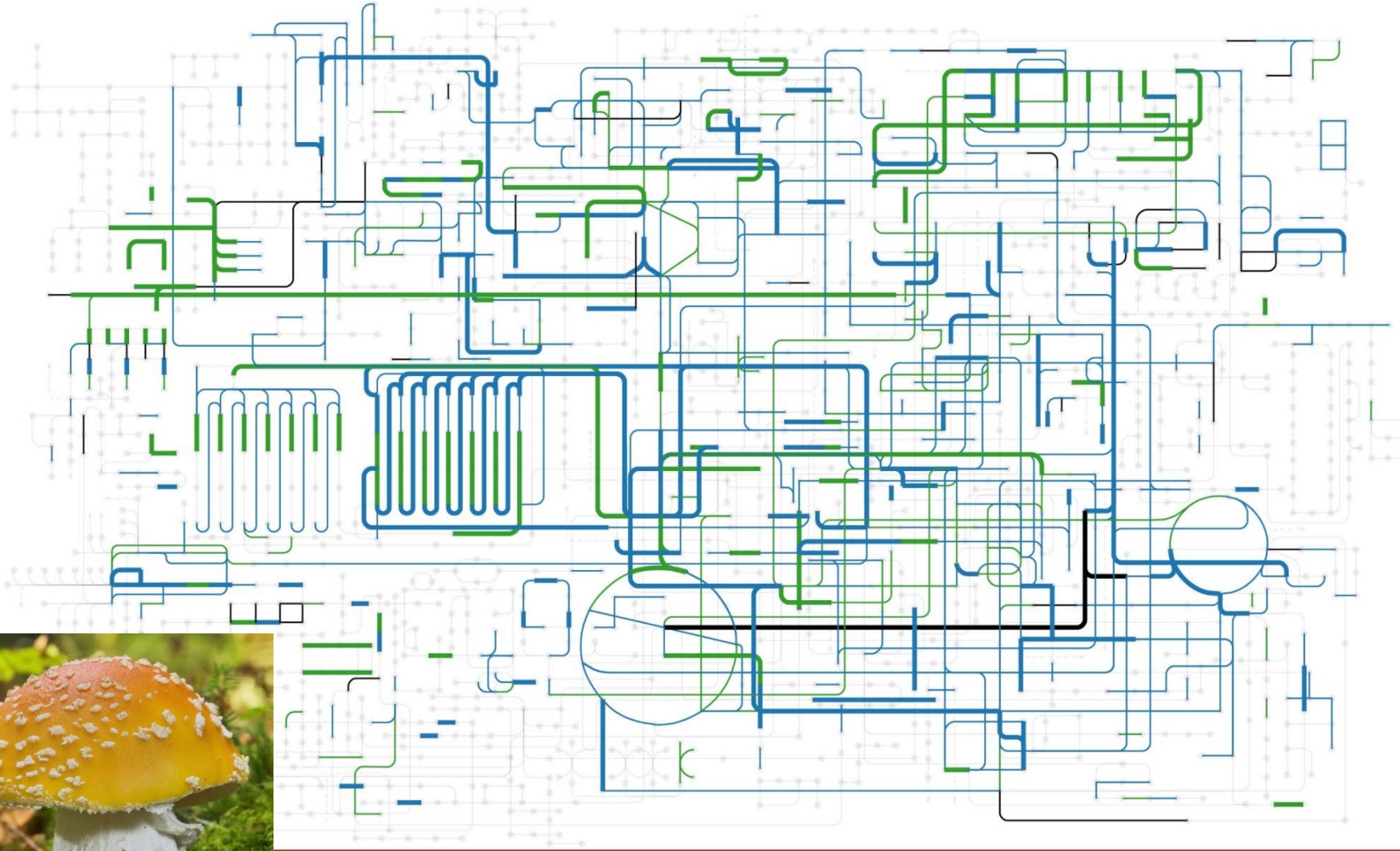
Basidios: Genes are not often Clustered or Transferred

Weight

— Never clustered ▬ Clustered

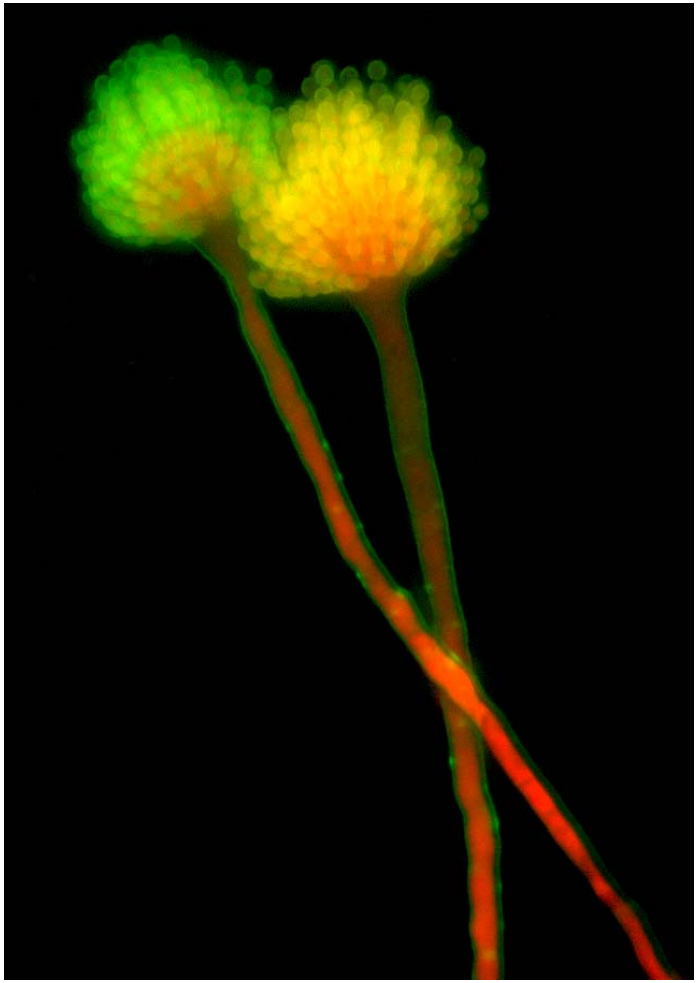
Color

No GD or HGT GD HGT GD & HGT



Wisecaver et al. (2014) PLoS Genet.

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)**

McGovern et al. (2004) PNAS

Aspergillus oryzae, *A Domesticated Microbe*

A. flavus

DOMESTICATION RD

ONE WAY

A. oryzae

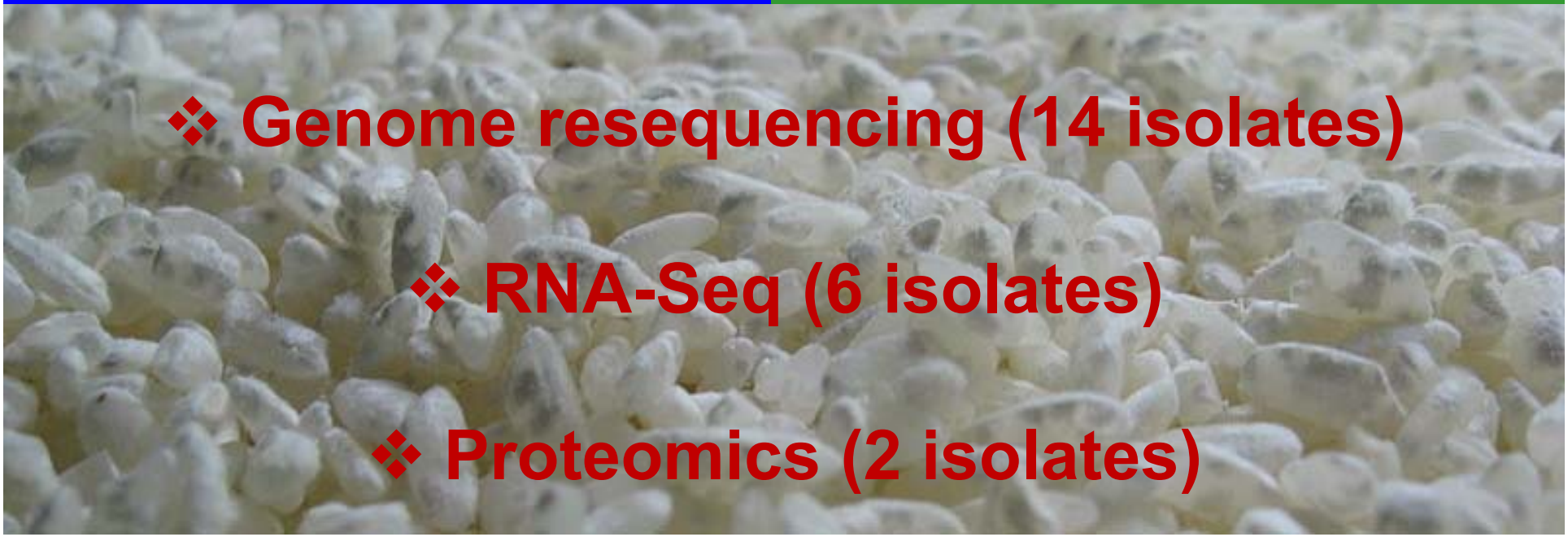
- ❖ Agricultural pest
- ❖ Aflatoxin producer
- ❖ ~$1 billion annually

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

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

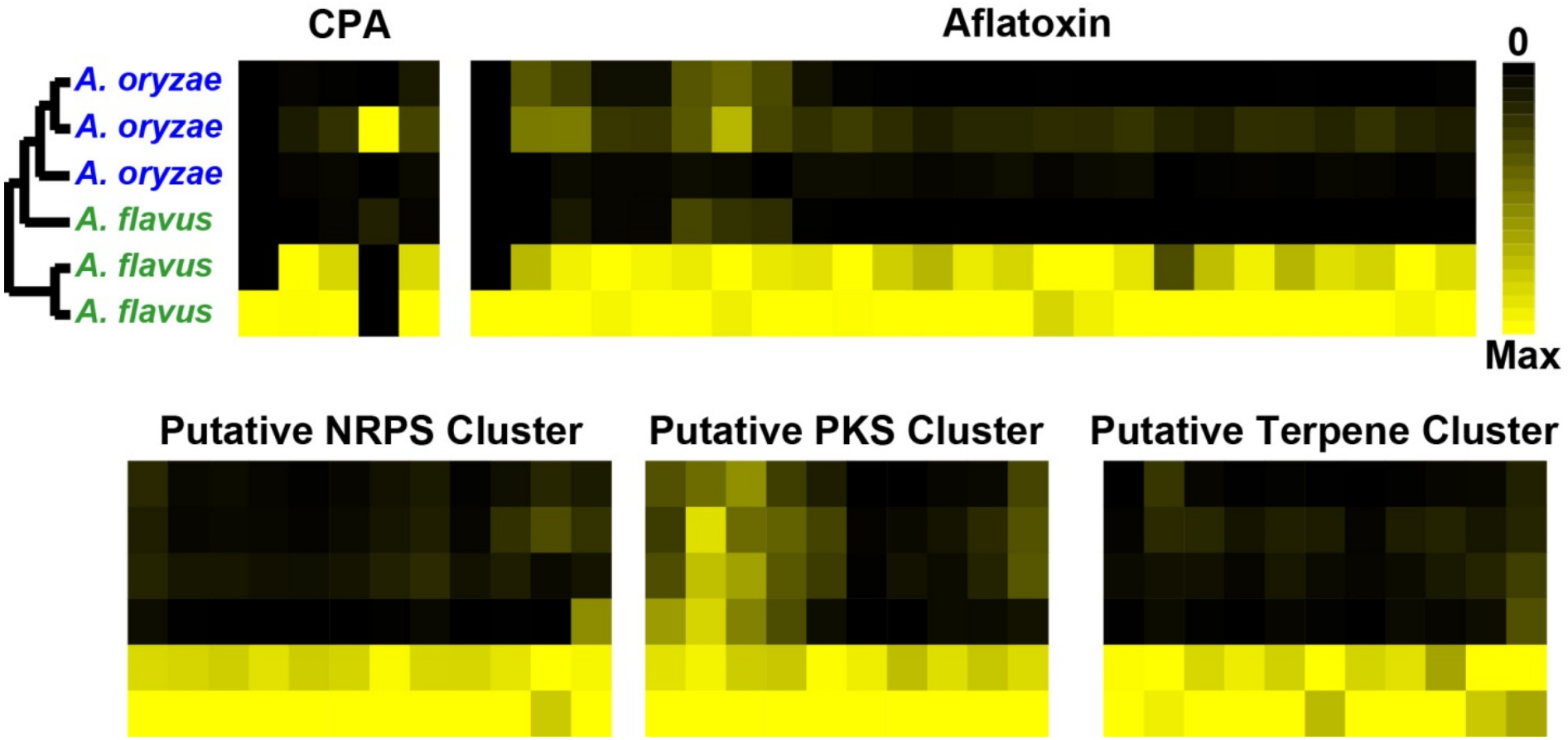
Studying Microbial Domestication using Omics

A. oryzae		*A. flavus*	
SRRC 302	Sake	SRRC 1273	Soil, Arizona
RIB 331	Miso	SRRC 1357	Dried bacon, Croatia
RIB 333	Miso	SRRC 2112	Hazelnut, Turkey
RIB 537	Sake	SRRC 2114	Wheat, USA
RIB 632	Sake	SRRC 2524	Dead termites, China
RIB 642	Sake	SRRC 2632	Blood, Chicago, Illinois
RIB 949	Soy Sauce	SRRC 2653	Corneal ulcer, Miami, Florida
RIB 40	Sake, Reference Strain	NRRL 3357	Peanut, Reference strain, USA



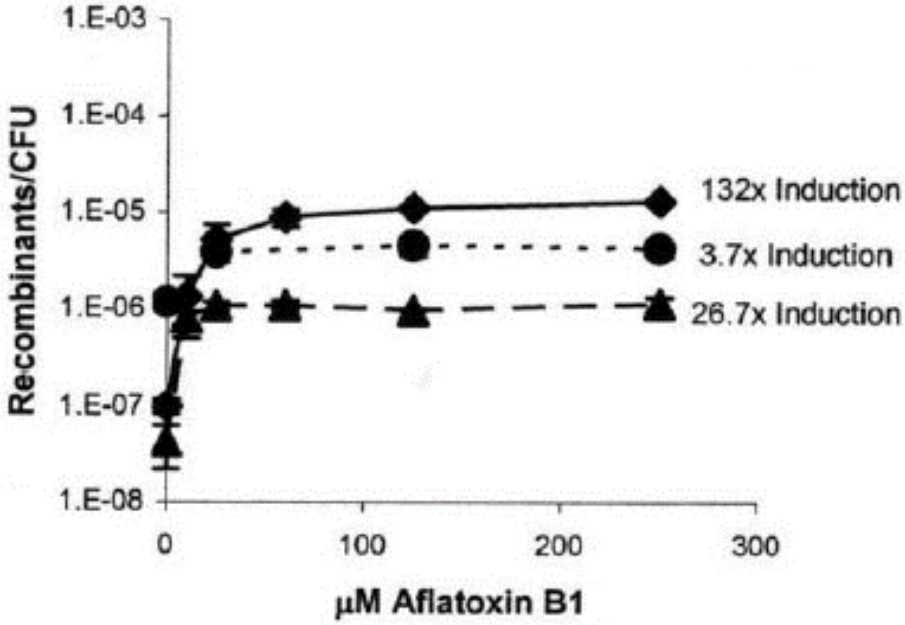
- ❖ Genome resequencing (14 isolates)
- ❖ RNA-Seq (6 isolates)
- ❖ Proteomics (2 isolates)

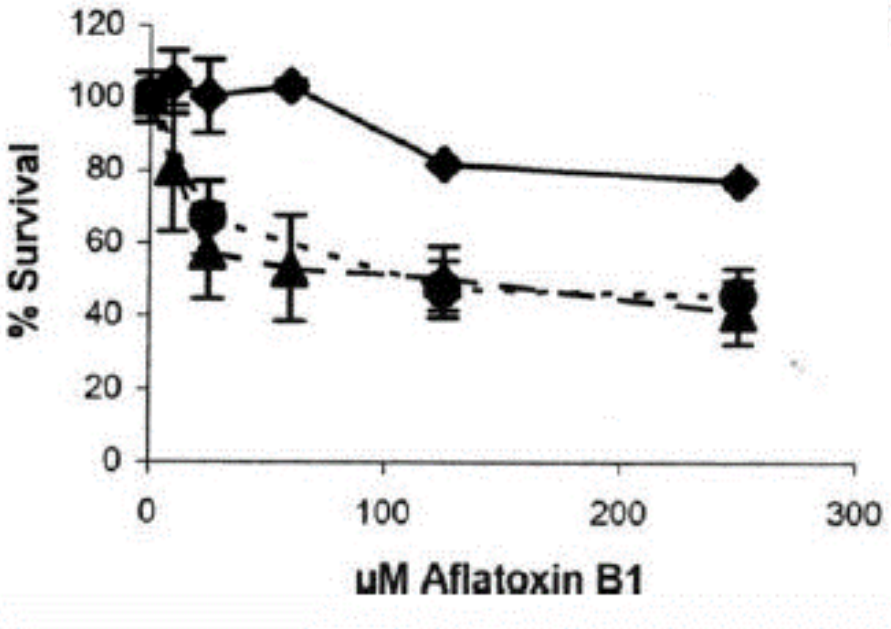
Down-Regulation of Secondary Metabolism in A. oryzae



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

Keller-Seitz et al. (2004) Mol. Biol. Cell

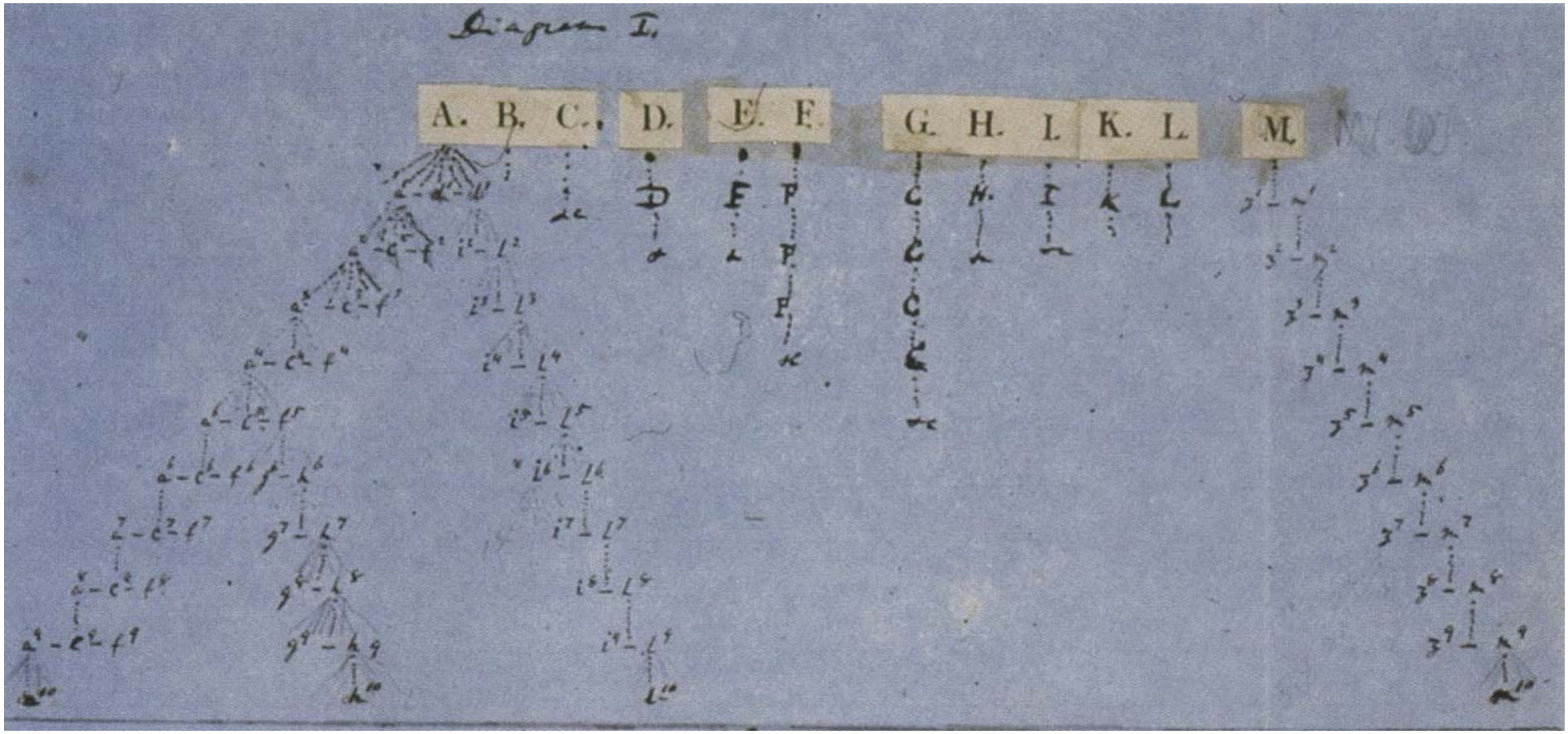
Lecture Outline

- **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*

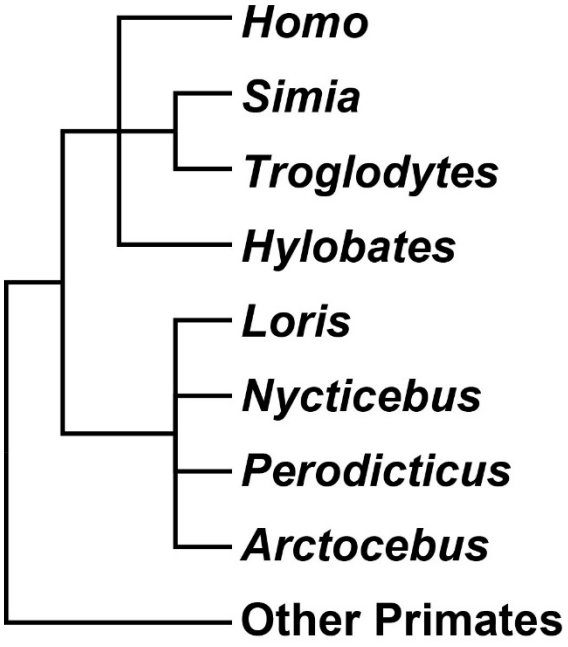
and instinct as the summing up of many contrivances, each useful to the possessor, nearly in the same way as when we look at any great mechanical invention as the summing up of the labour, the experience, the reason, and even the blunders of numerous workmen; when we thus view each organic being, how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

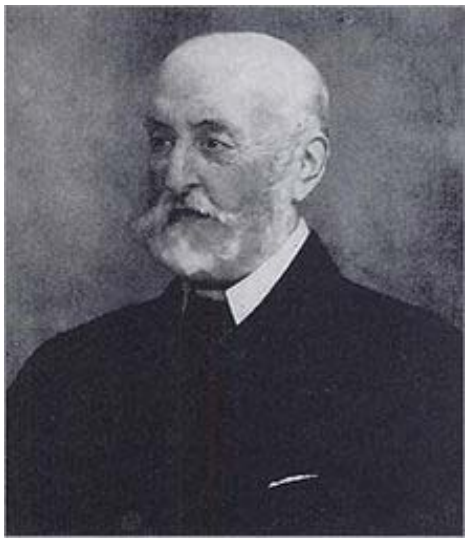
A grand and almost untrodden field of inquiry will be opened, on the causes and laws of variation, on correlation of growth, on the effects of use and disuse, on the direct action of external conditions, and so forth. The study of domestic productions will rise immensely in value. A new variety raised 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

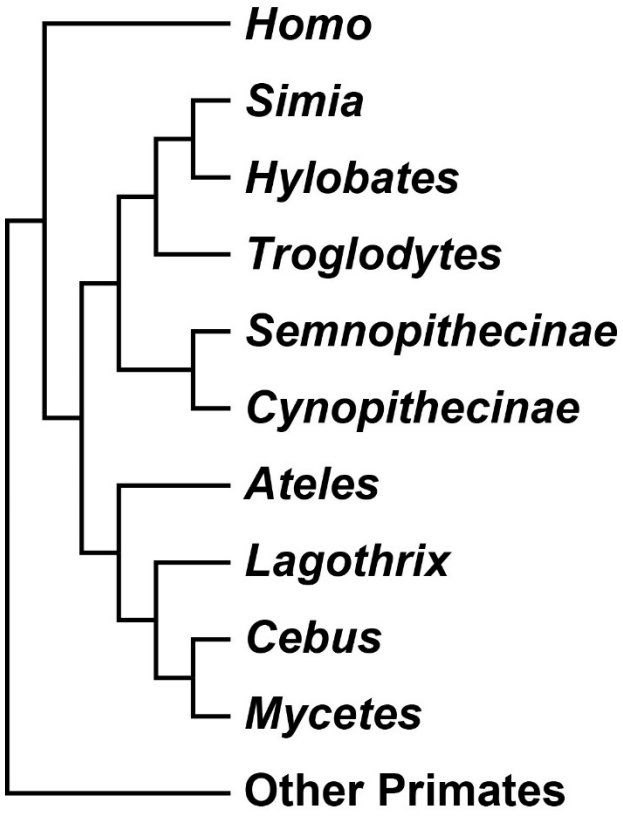
1865: SPINAL COLUMN





St. George Jackson Mivart

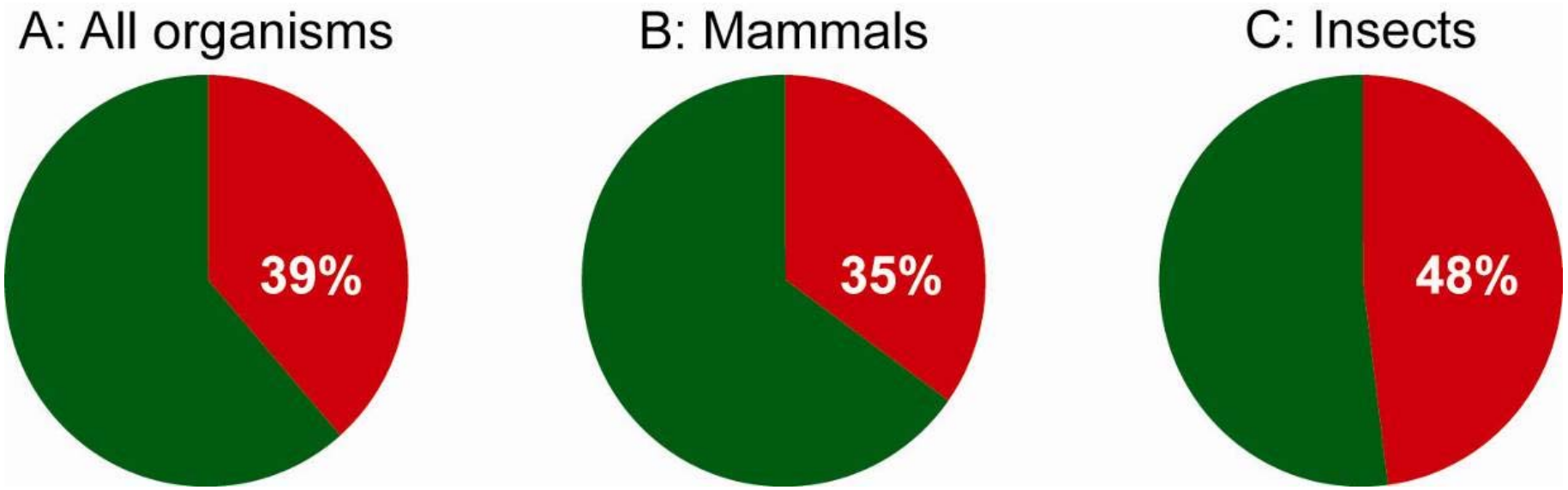
1867: LIMPS



"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*"

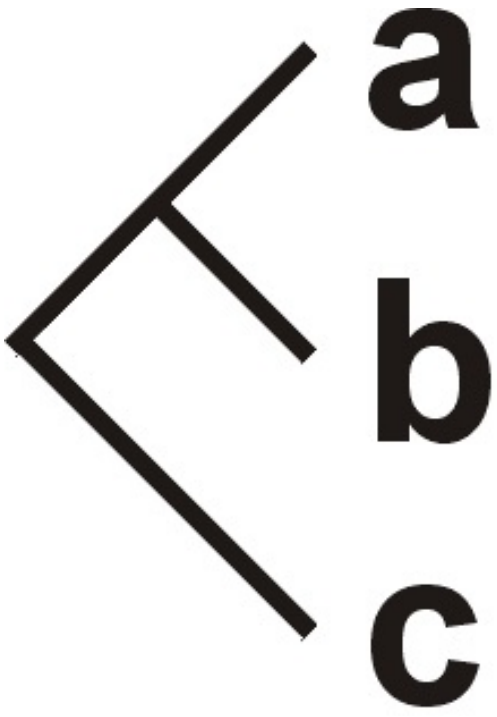
Darwin Correspondence Project **letter 7170**

The Problem of Incongruence

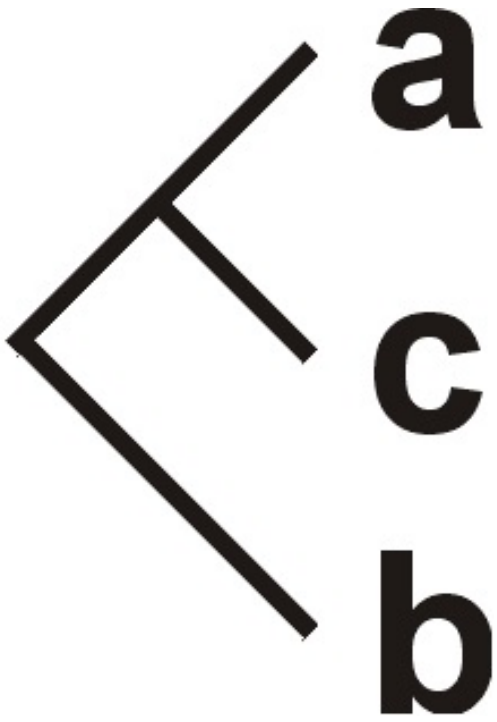


Incongruence **is pervasive in the phylogenetics literature**

The Problem of Incongruence



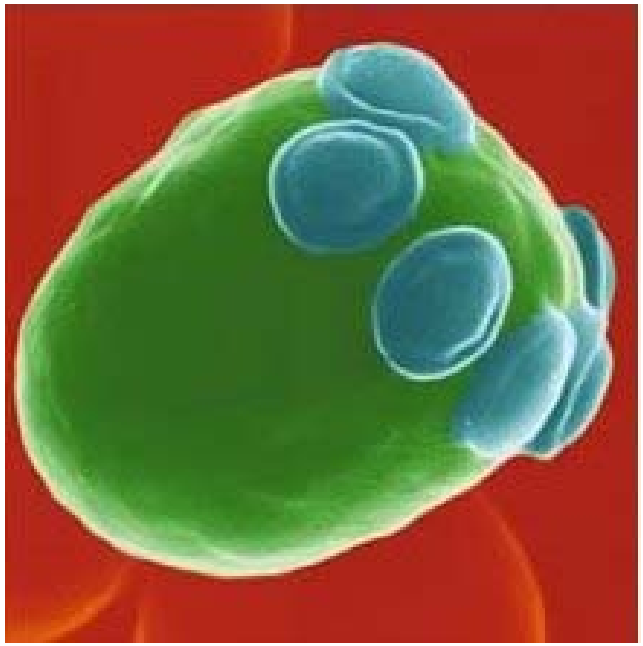
Gene X



Gene Y

Species tree?

A Systematic Evaluation of Single Gene Phylogenies



S. cerevisiae

S. paradoxus

S. mikatae

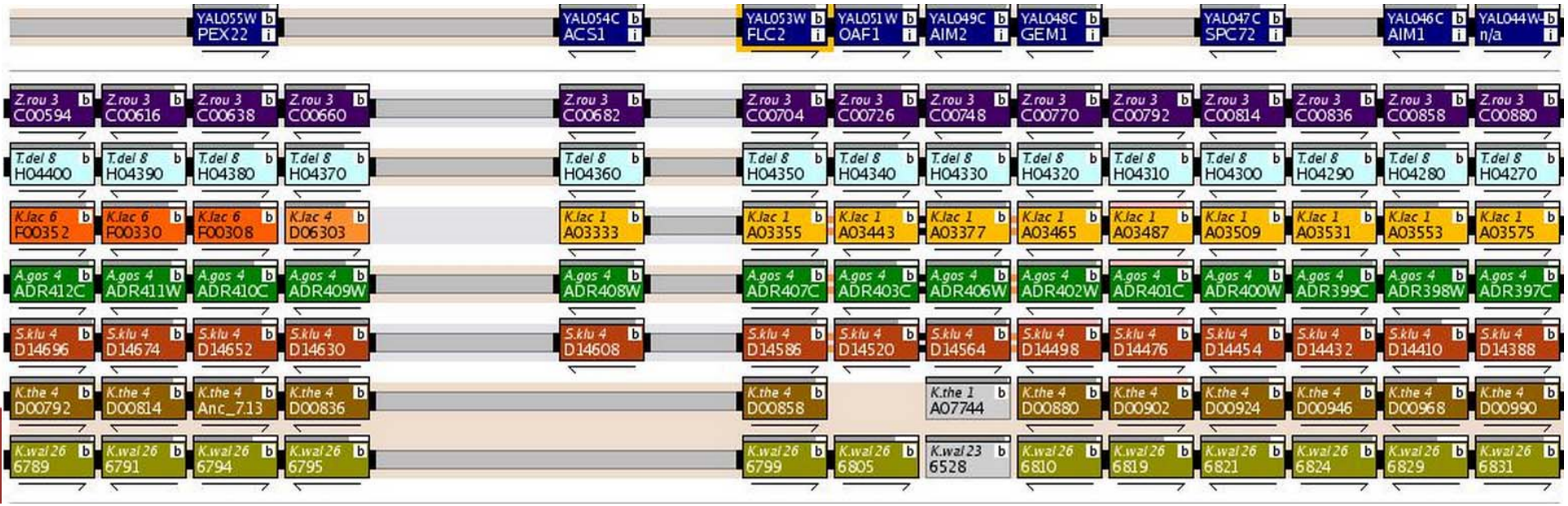
S. kudriavzevii

S. bayanus

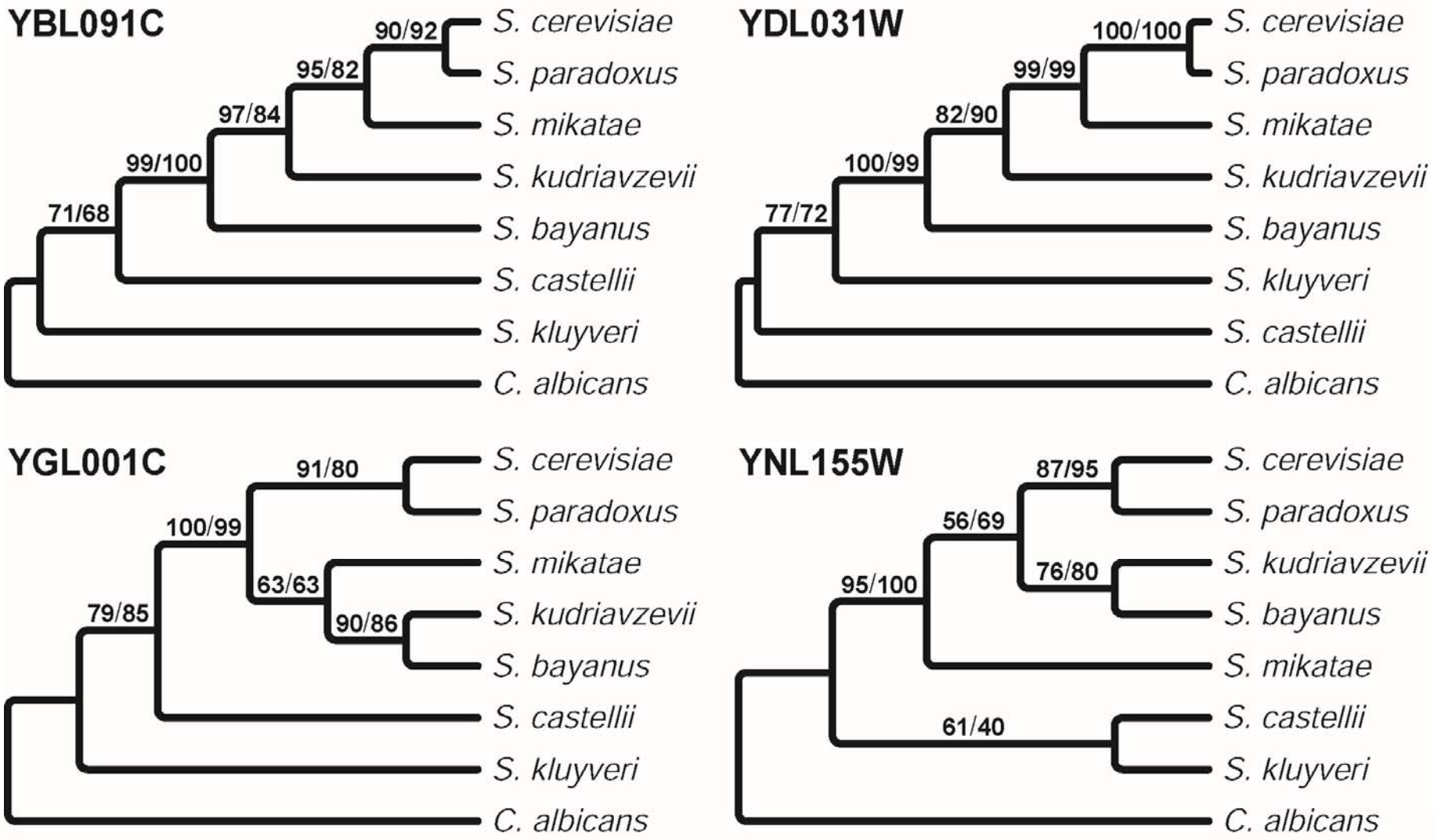
S. castellii

S. kluyveri

Candida albicans



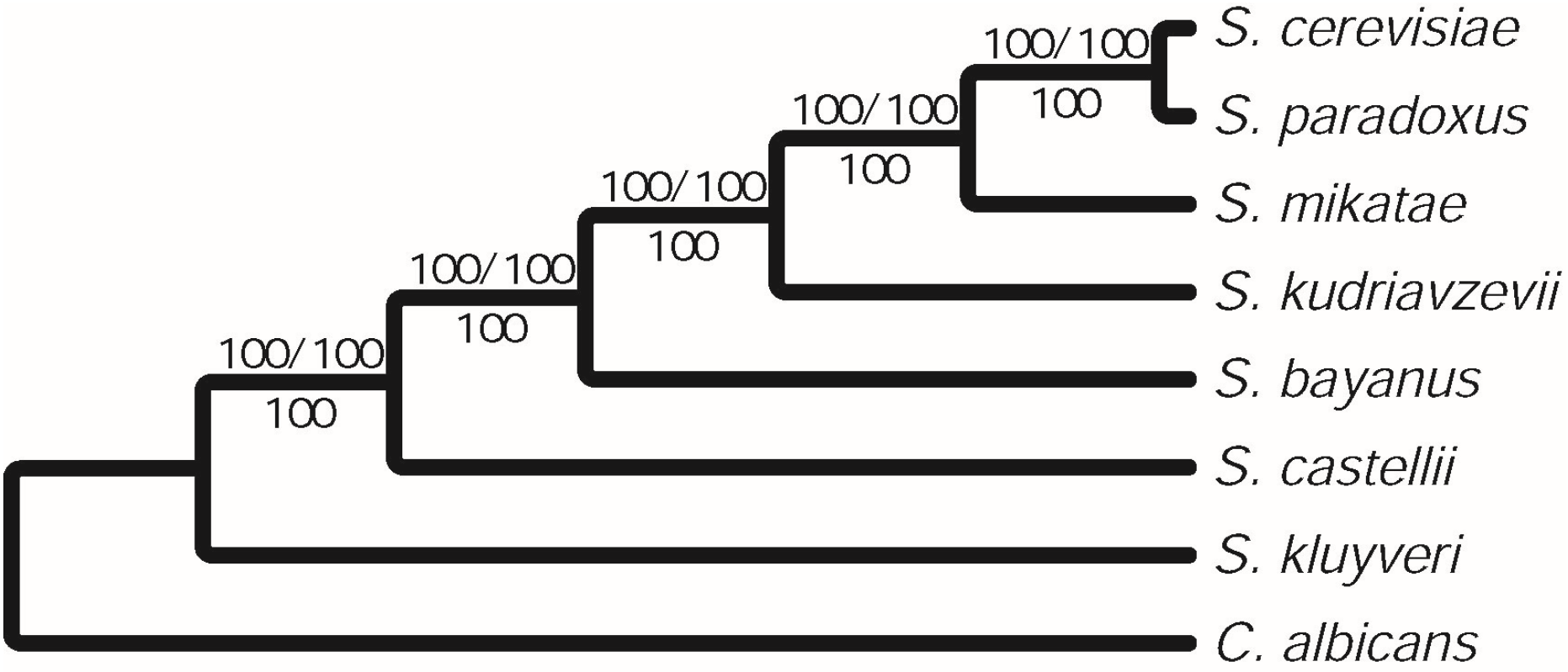
Incongruence at the Single Gene Level



ML / MP

Rokas et al. (2003) Nature

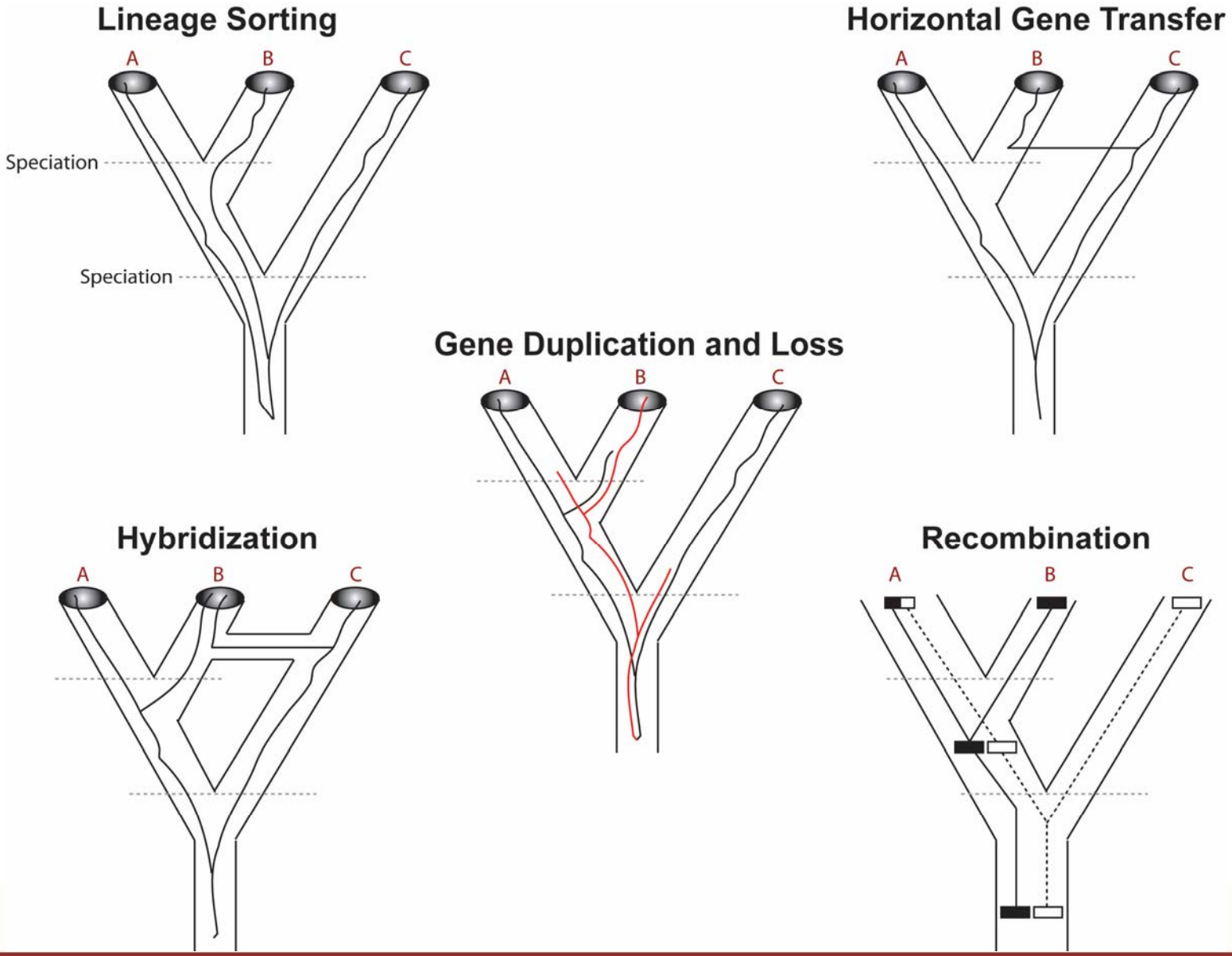
Concatenation of 106 Genes Yields a Single Yeast Phylogeny



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

Rokas et al. (2003) Nature

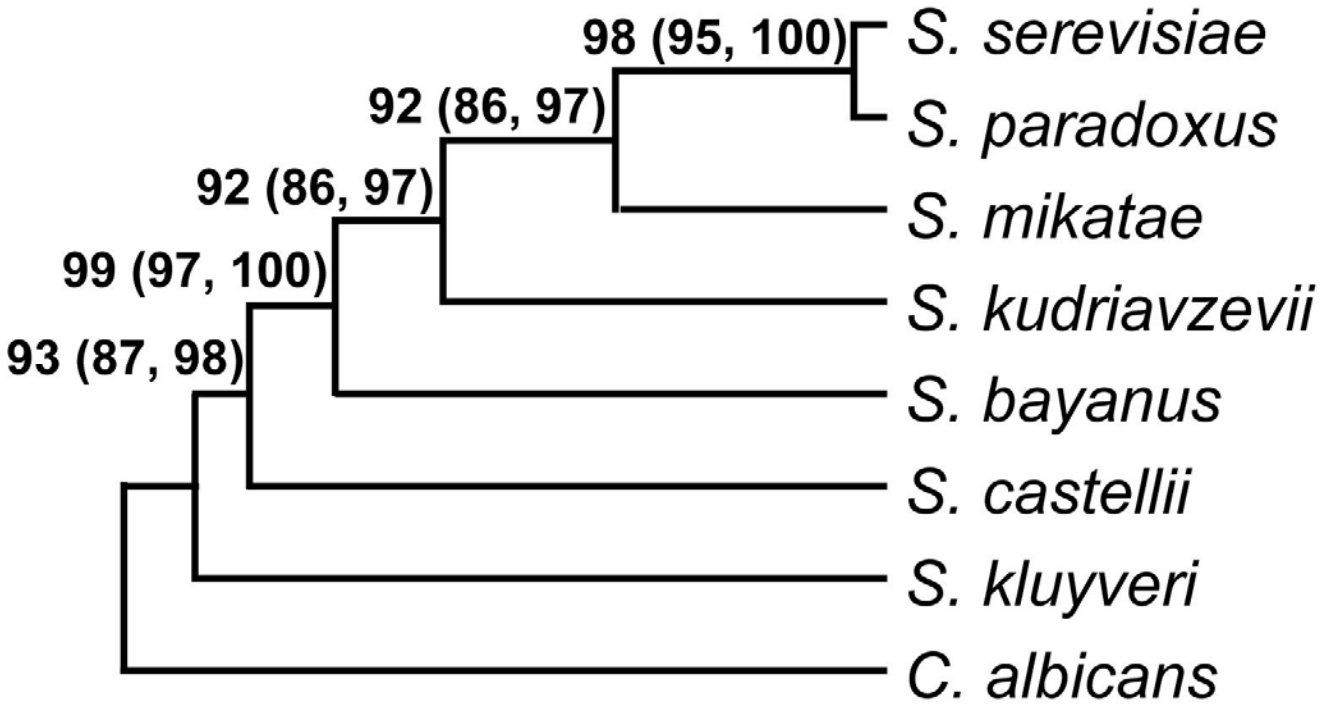
Gene Trees Can Differ from Species Trees



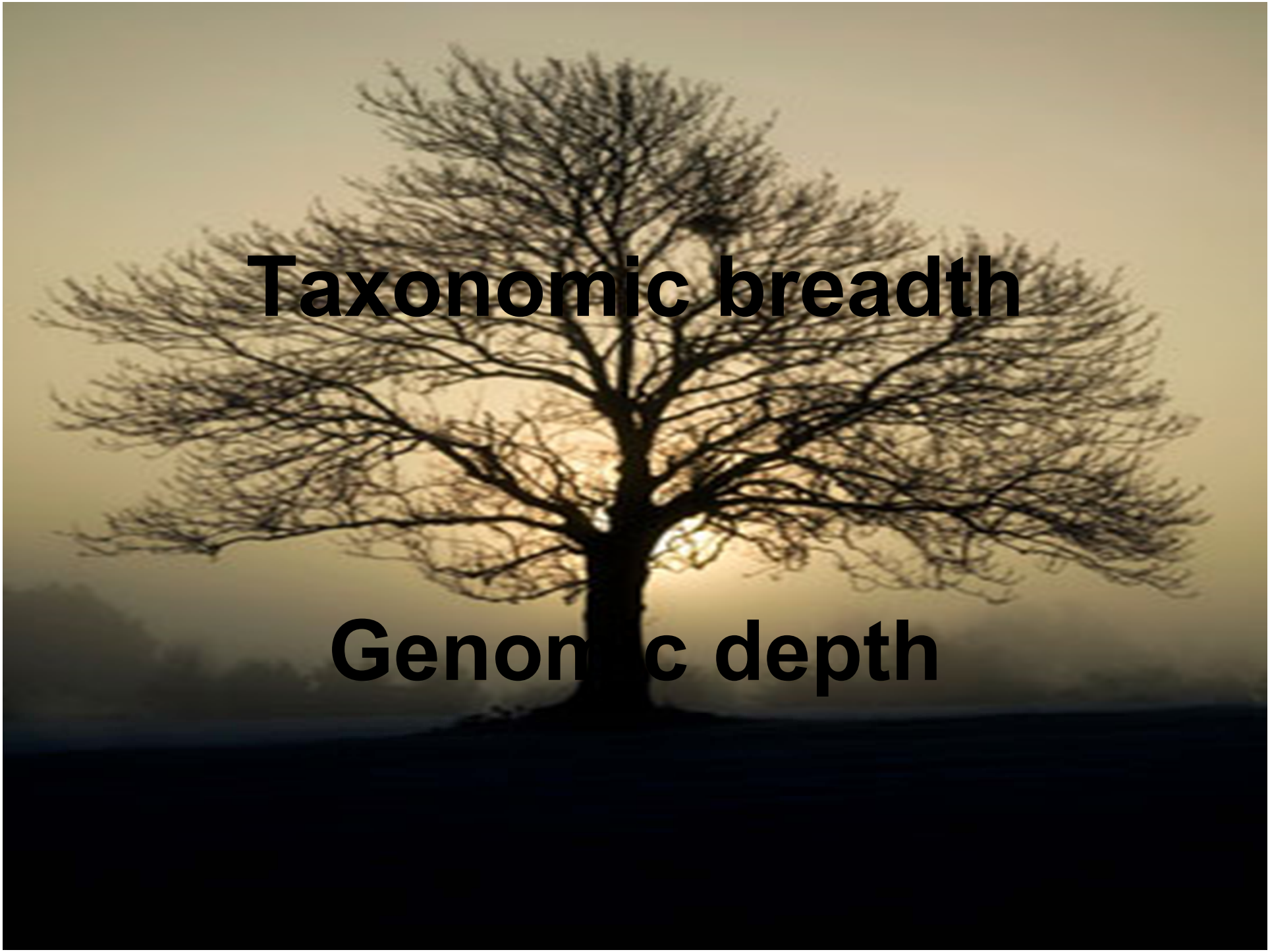
Degnan & Rosenberg (2009) Trends Ecol. Evol.

Inferring the Species Tree from Individual Gene Histories

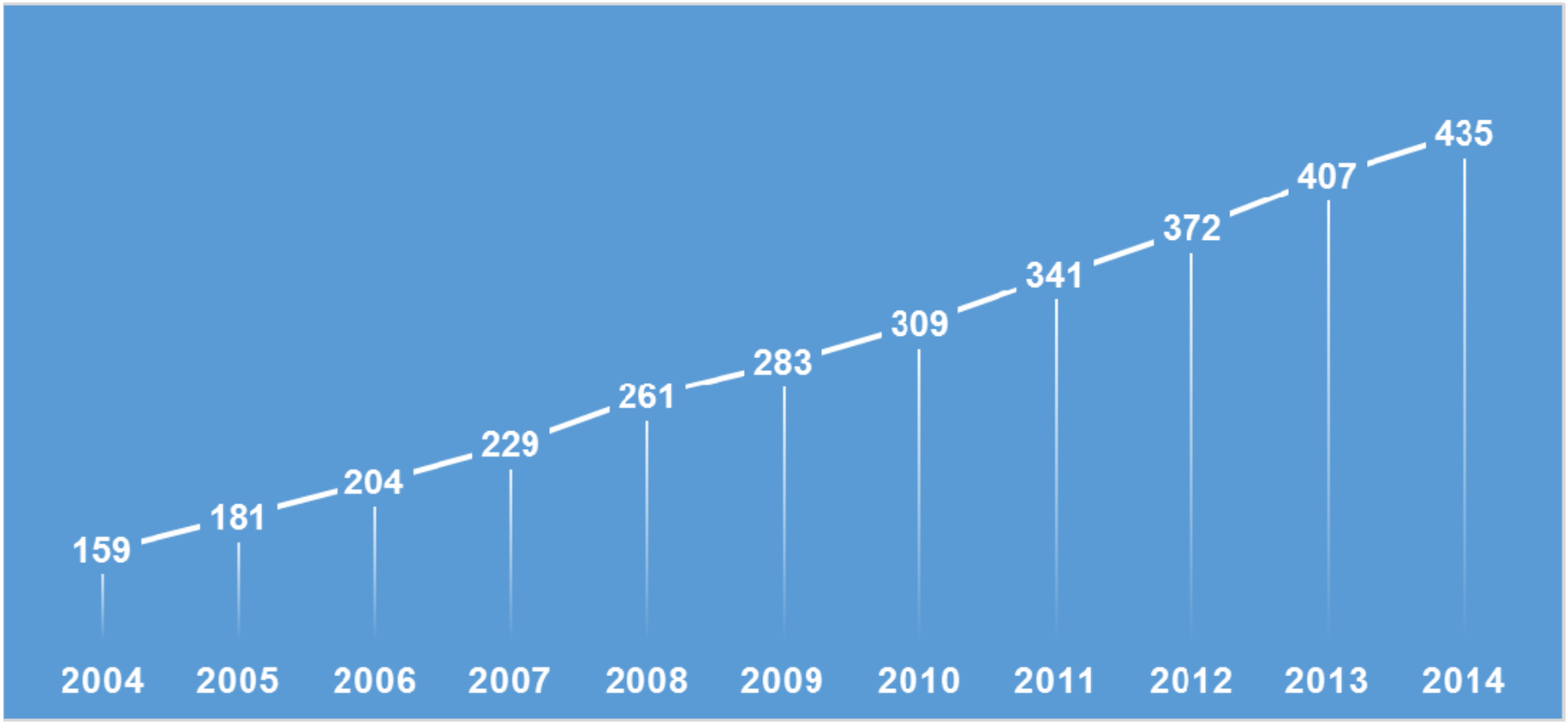
Concordance Factor: The proportion of the genome for which a clade is true



Ané et al. (2007) Mol. Biol. Evol.

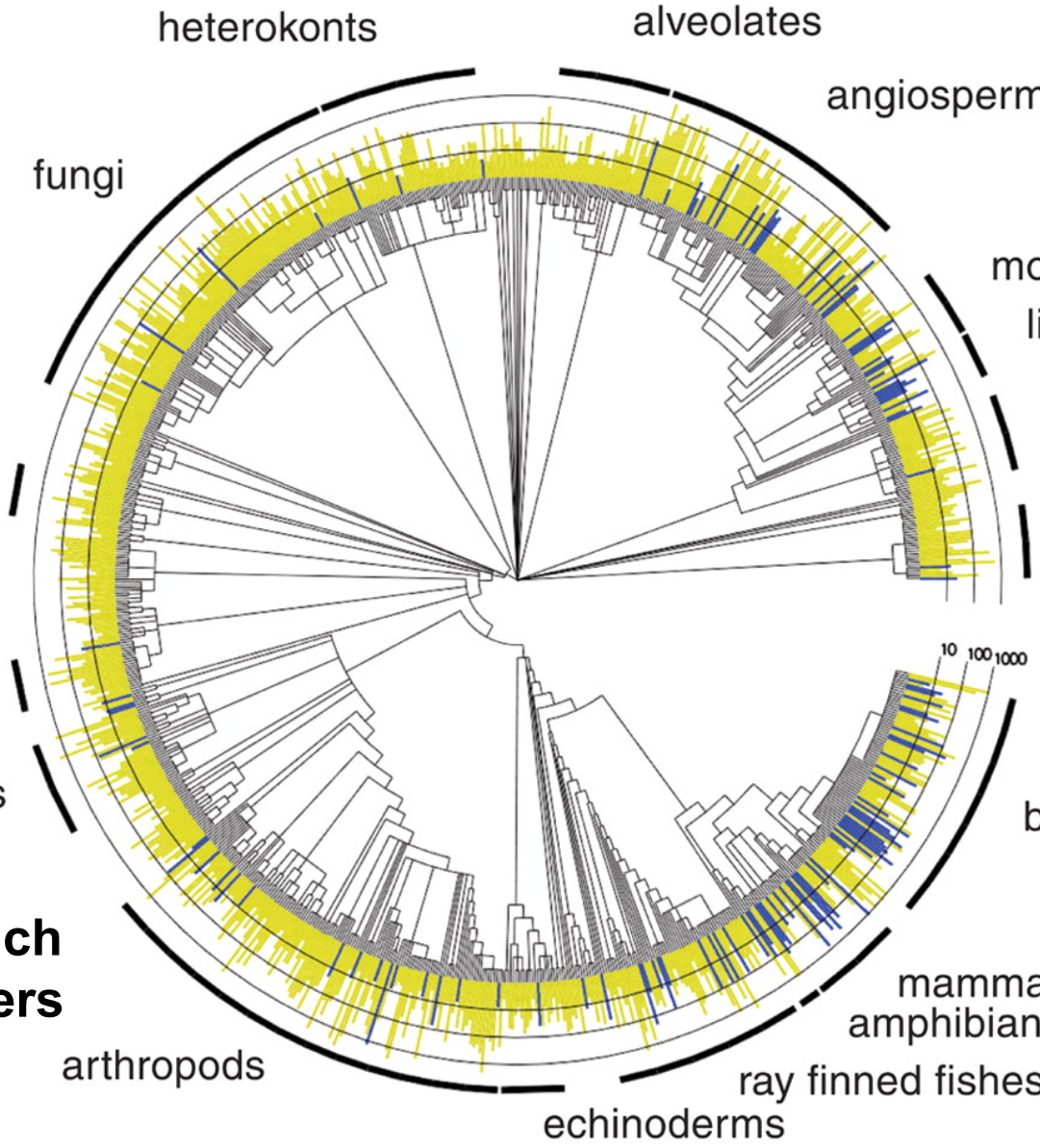

Taxonomic breadth
Genomic depth

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



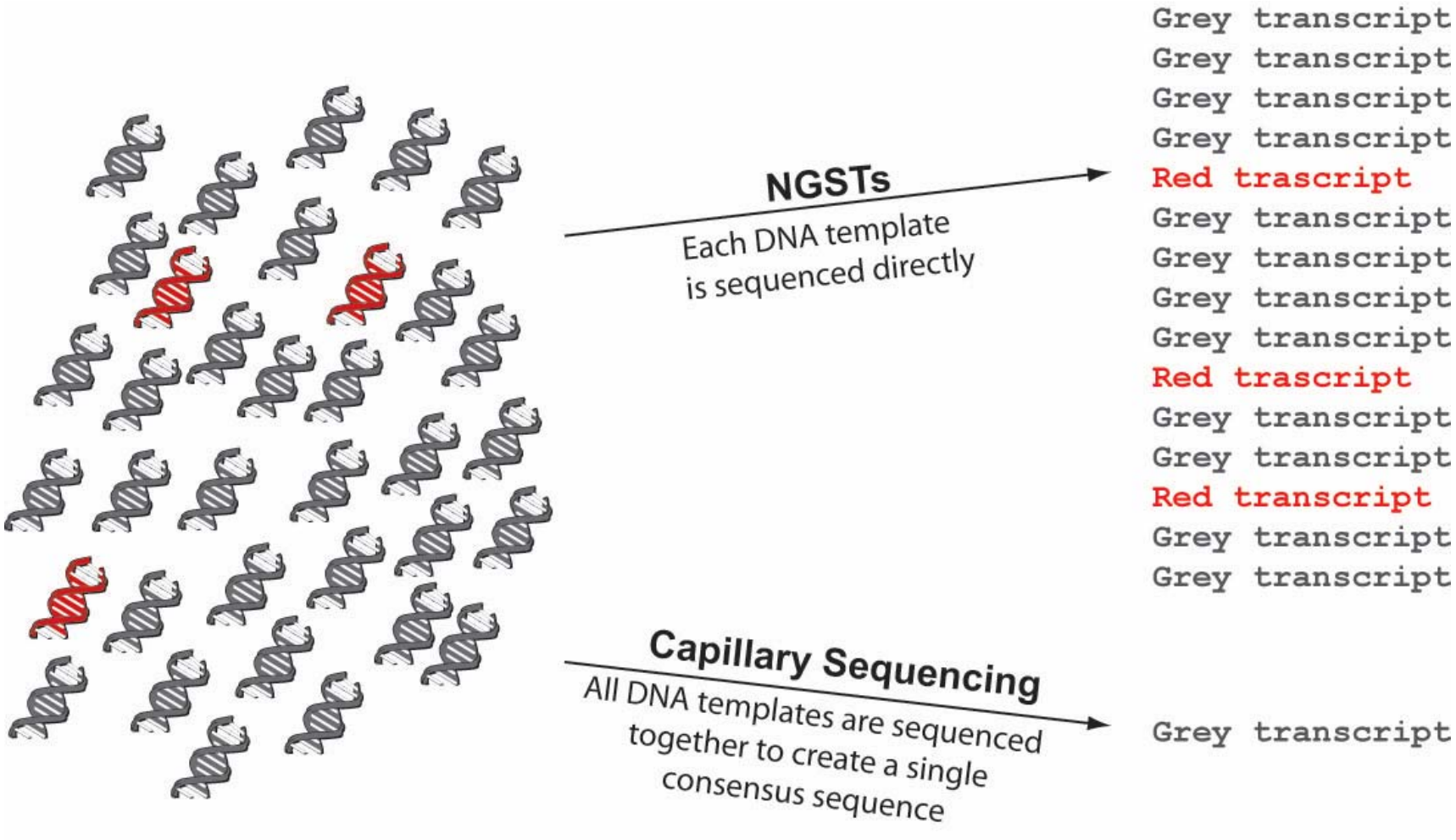
1/40 fungal orders

3/45 angiosperm orders

1/45 species-rich arthropod orders supported

16/57 vertebrate orders

Next-Gen Sequencing is Qualitative and Quantitative



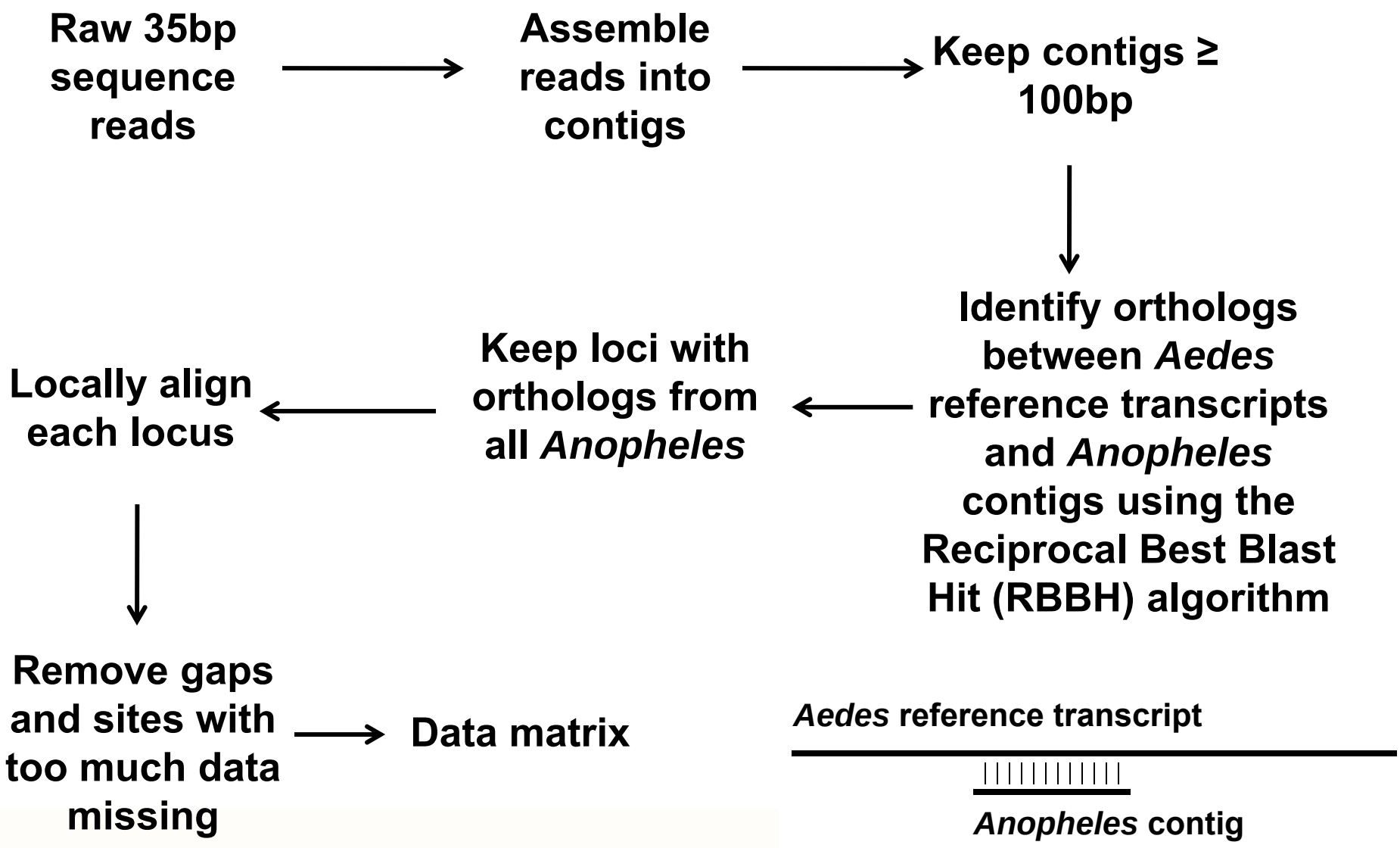
Rokas & Abbot (2009) Trends Ecol. Evol.

Can we Use RNA-Seq to Increase Genomic Depth?

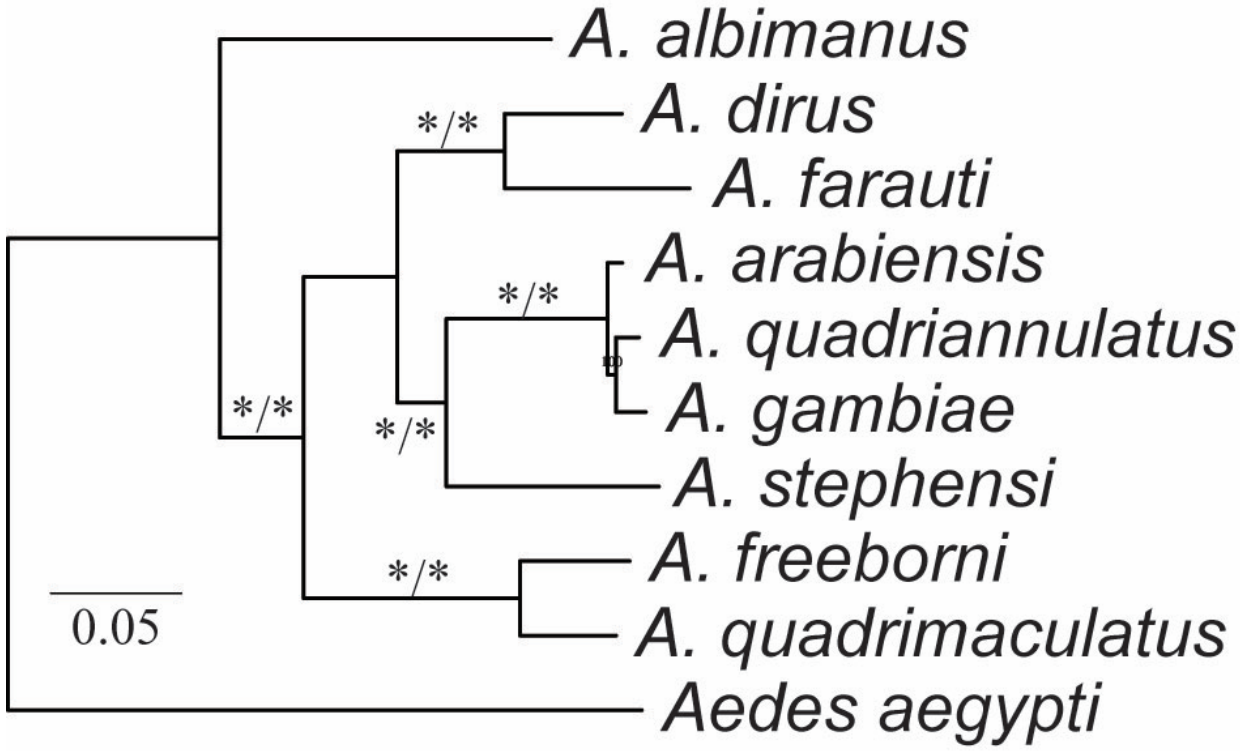
Species	Stock No.	Collection Location
Anopheles albimanus (Nyssorhynchus)	MRA-126	El Salvador
Anopheles arabiensis (Cellia)	MRA-339	Zimbabwe
Anopheles dirus (Cellia)	MRA-700	Thailand
Anopheles farauti (Cellia)	MRA-489	Papua New Guinea
Anopheles freeborni (Anopheles)	MRA-130	USA
Anopheles gambiae (Cellia)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

Illumina RNA-Seq → **~150,000,000 reads**
5,250,000,000 bp

Data Matrix Construction: The "Singlecontig" Strategy

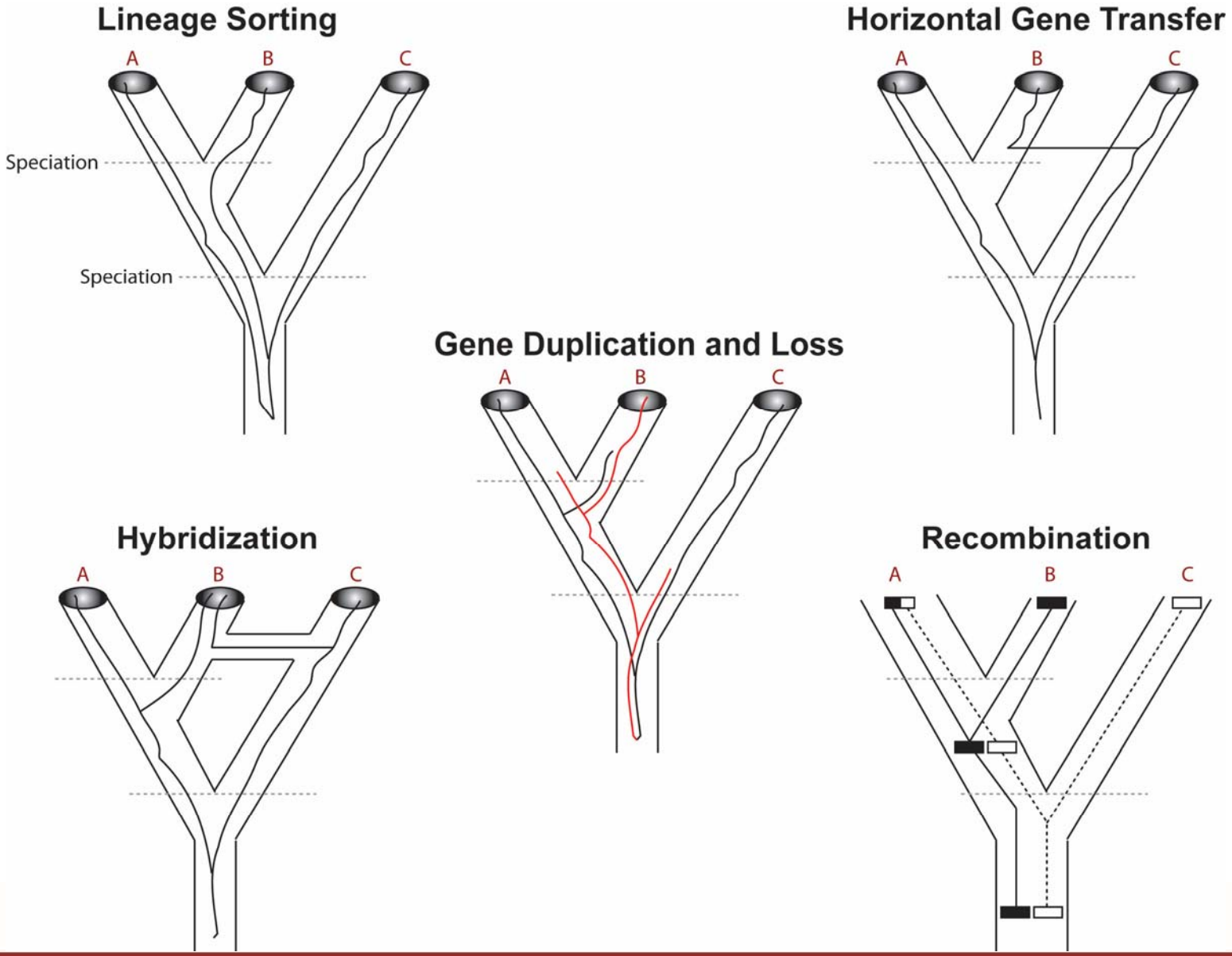


Robust Phylogenetic Inference from RNA-Seq Data



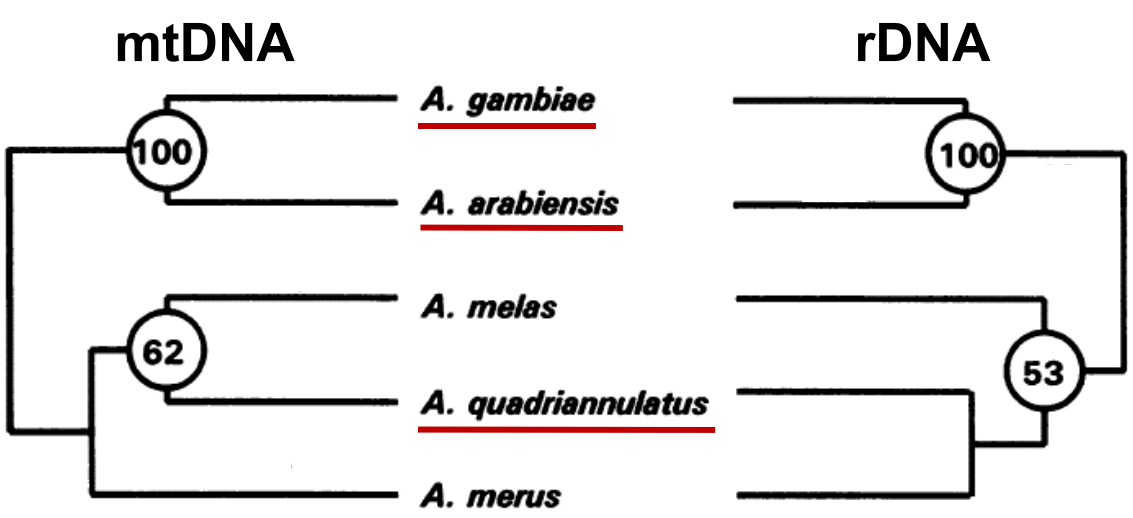
 # Loci = 553
Aln Length = ~390 Kb
% Missing data = 51

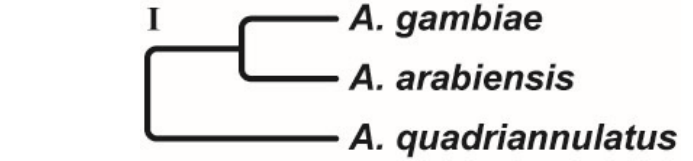
Gene Trees Can Differ from Species Trees

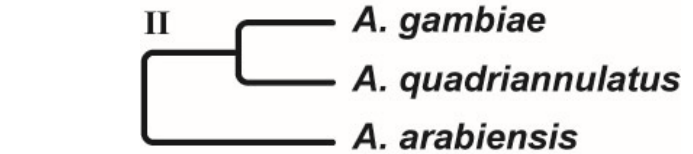


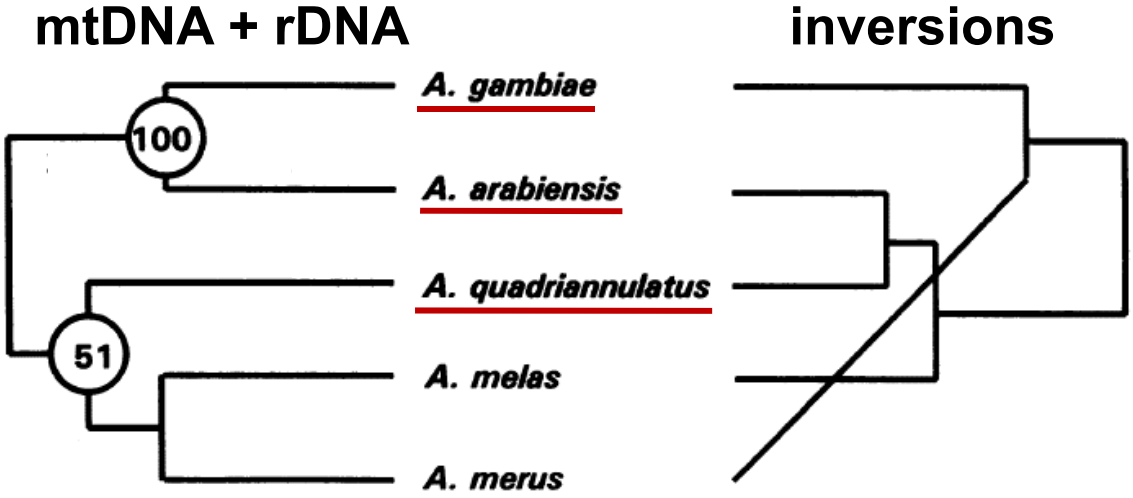
Degnan & Rosenberg (2009) Trends Ecol. Evol.

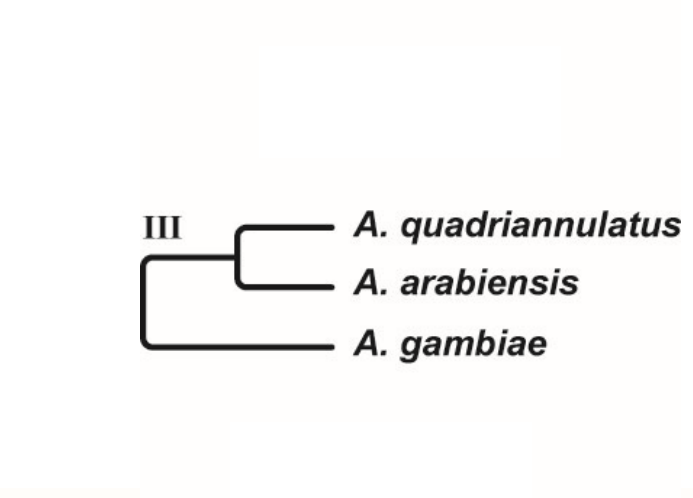
Our Data Matrices Can Detect Population-Level Events











Besansky et al. (1994) PNAS

Hittinger et al. (2010) PNAS

The Phylogenomics Era – "Resolving" the Tree of Life

Syst. Biol. 61(1):150–164, 2012

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

Stephen A. Smith[1,2], Nerida G. Wilson[3,4], Freya
Gonzalo Giribet[5] & Casey W. Dunn[1]

nature

LETTERS

Syst. Biol. 57(6):920–938, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802570791

Toward Resolving t
Tree: The Phylogen
of Jakobids and Cercozoans

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. CUNNINGHAM[3]

Yeast

An

OPEN ACCESS Free

Toward Resolving Priors

Prion-Like Proteins in the Fungal Kingdom

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

Towards

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

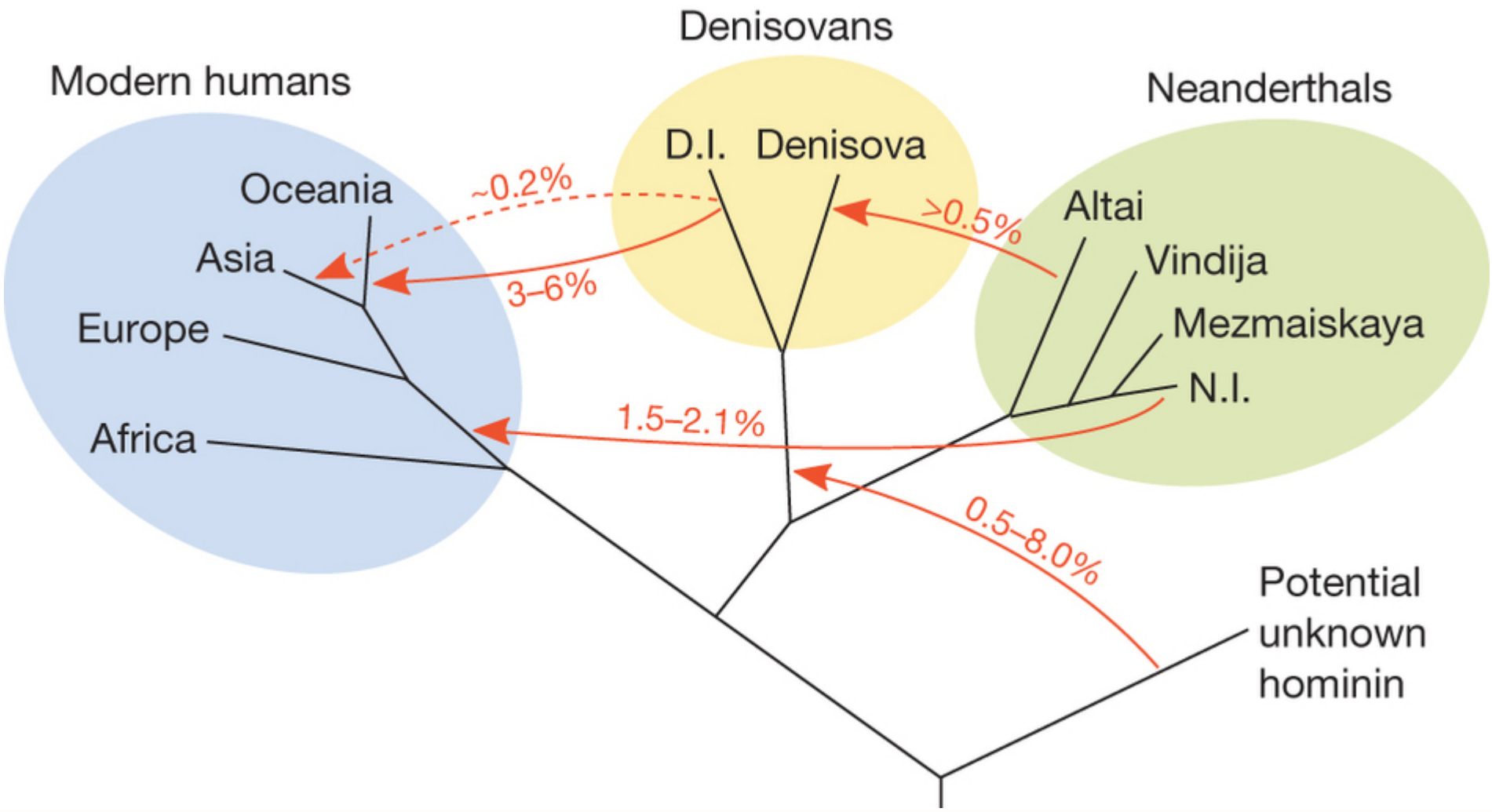
Samuli Lehtonen

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

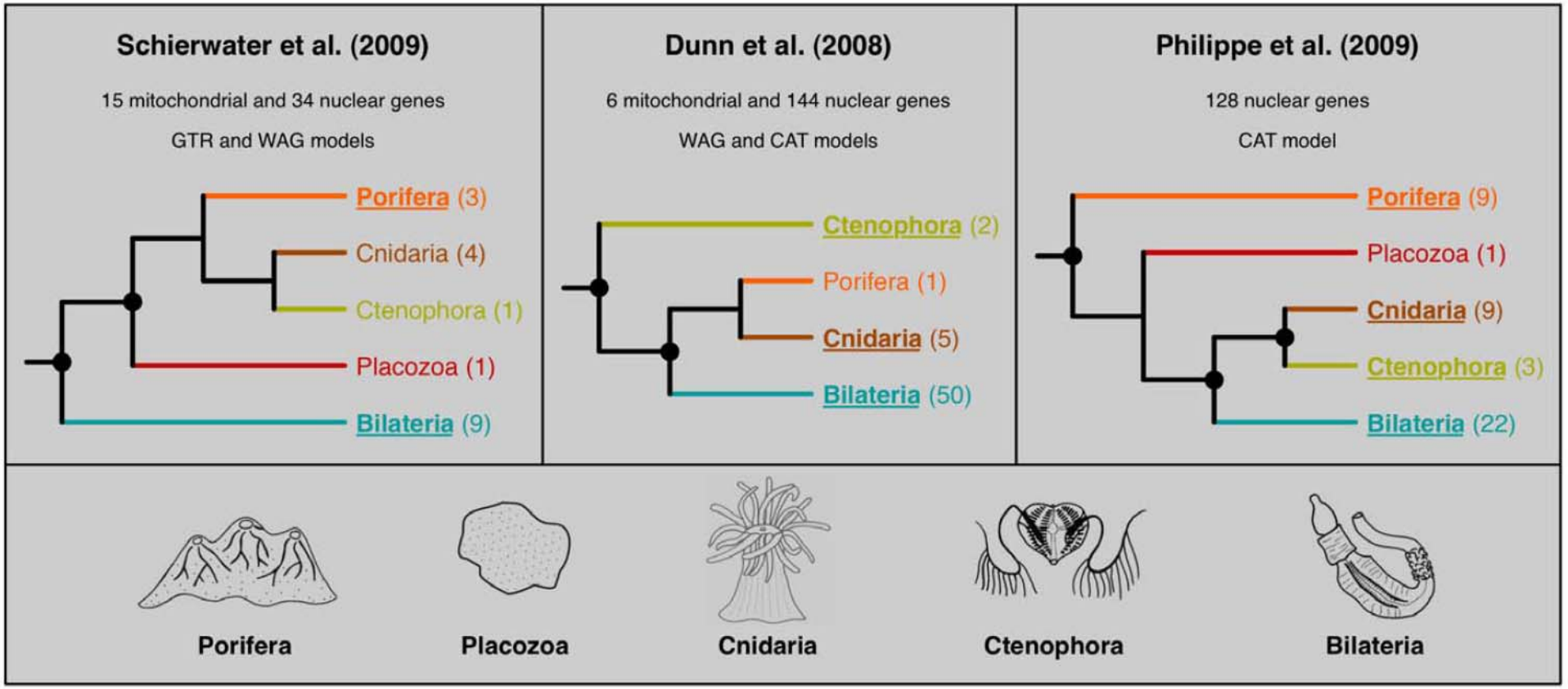
Have we eliminated incongruence?

Phylogenomics Works Beautifully at Shallow Levels



Prufer et al. (2014) Nature

Incongruence in Deep Time



Philippe et al. (2011) PLoS Biol.

Incongruence in Deep Time

Kocot et al. (2011) Nature

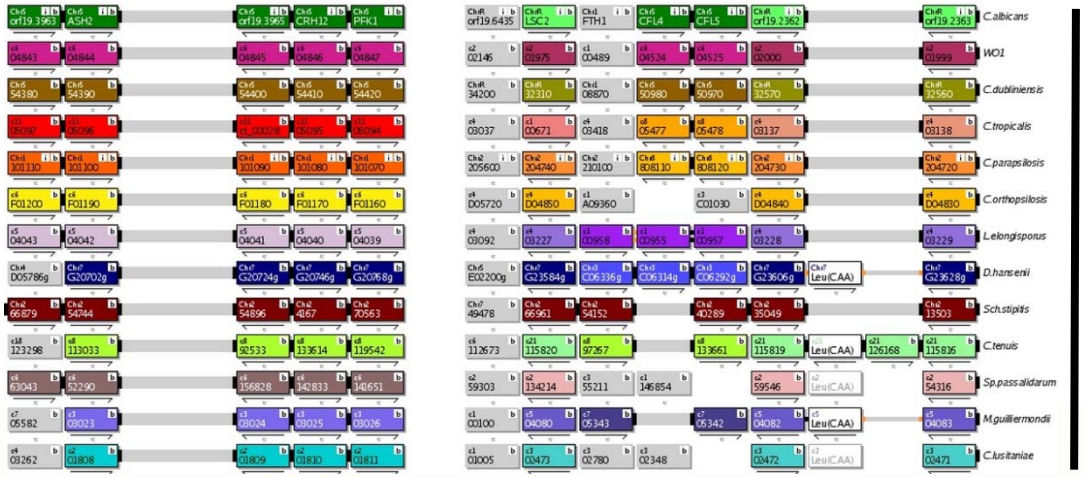
Smith et al. (2011) Nature

Why the disconnect?

An Expanded Yeast Data Matrix

Yeast Gene Order Browser (YGOB)

Candida Gene Order Browser (CGOB)



***Saccharomyces* lineage**

***Candida* lineage**

1,070 genes
23 taxa
no missing data

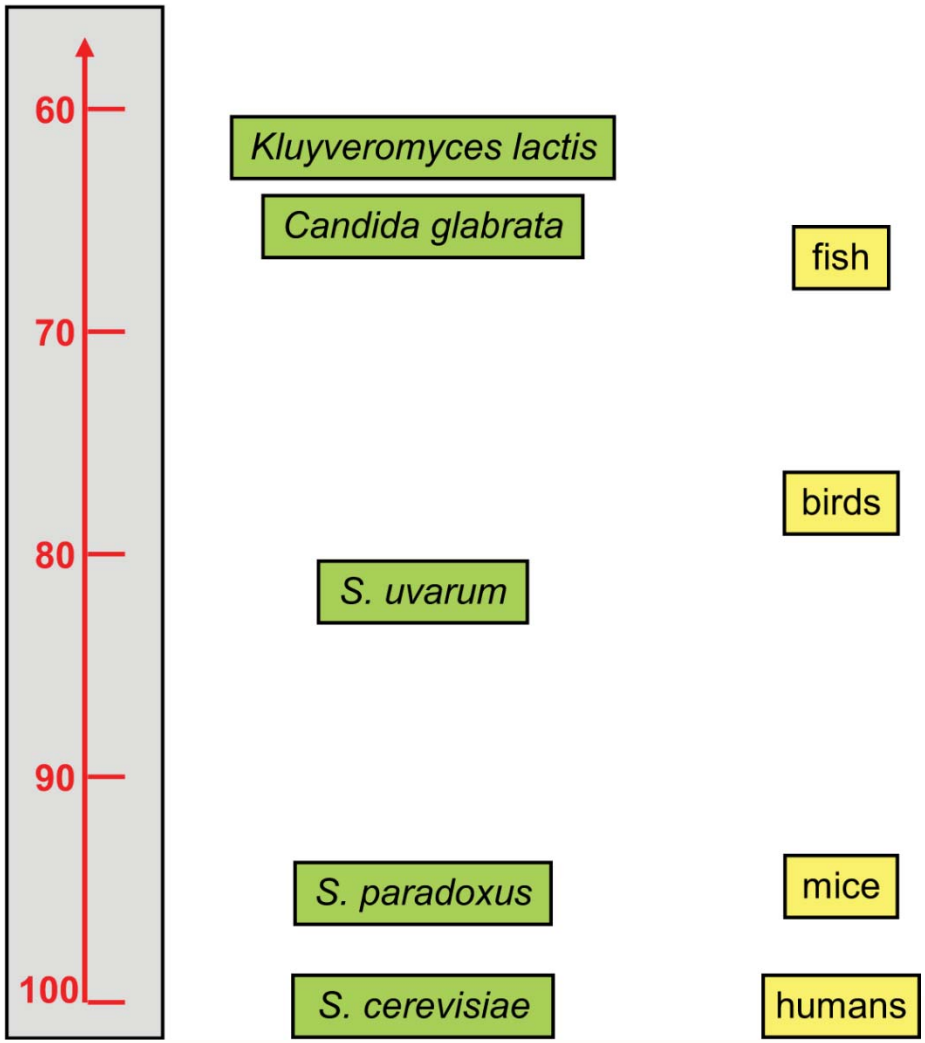
Byrne & Wolfe (2005) Genome Res.

Fitzpatrick et al. (2010) BMC Genom.

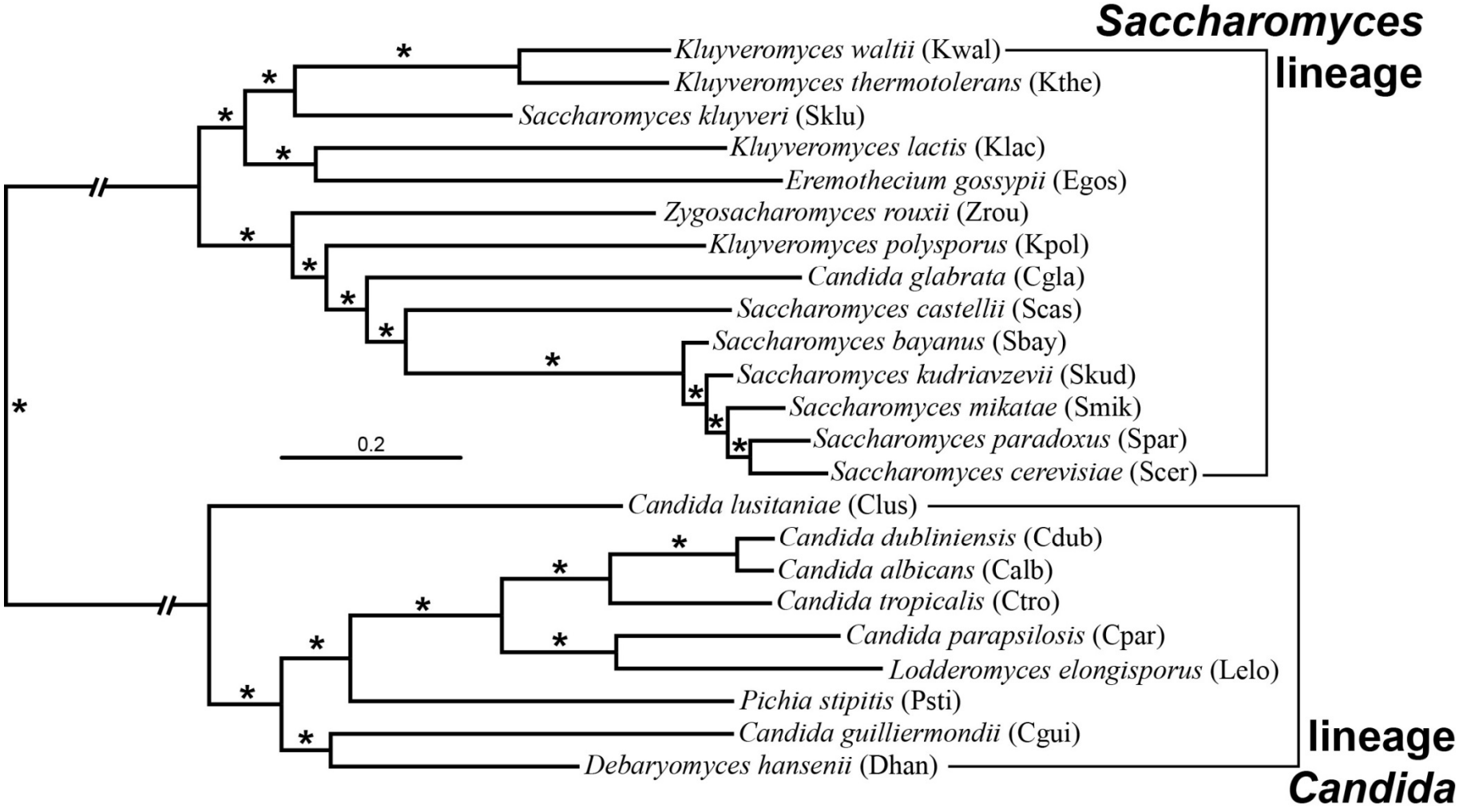
Fungal Genomes are Similar in Divergence to Animals

Proteome-wide average pairwise amino acid sequence similarity

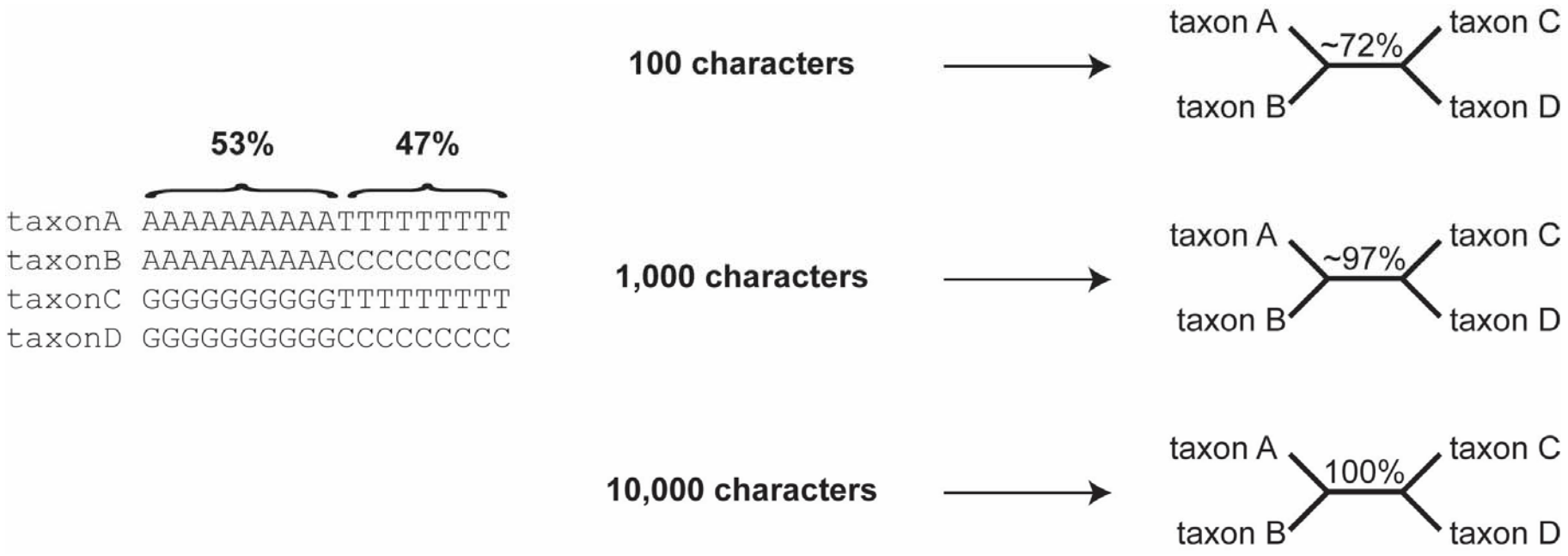
***Saccharomyces*, *Candida*, *Kluyveromyces*, etc. are all polyphyletic genera**



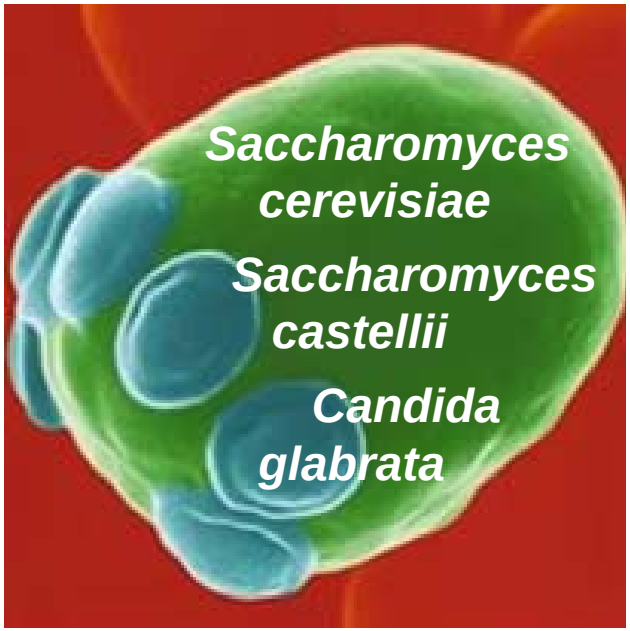
Concatenation Yields an Absolutely Supported Phylogeny



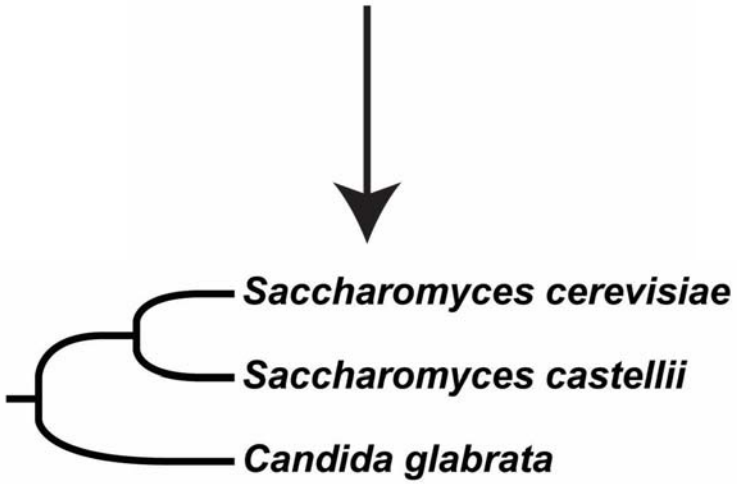
Bootstrap Support is Misleading When Used in Large Datasets



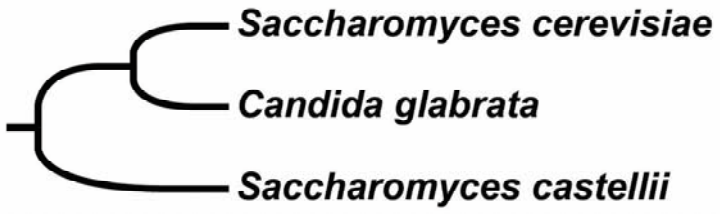
The Concatenation Phylogeny is at Least Partly Wrong



Linear Sequence Data

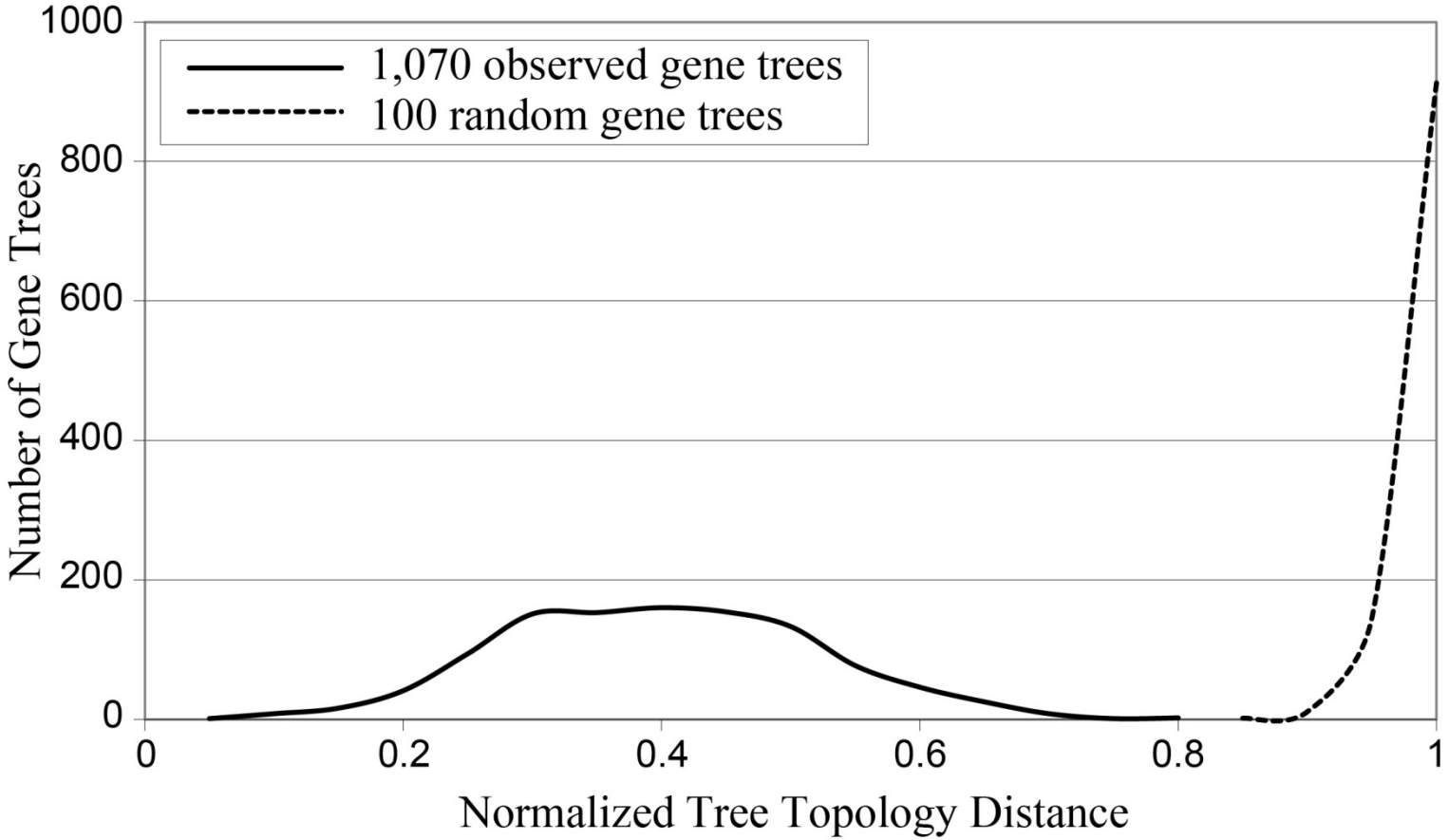


Rare Genomic Changes



- **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***

All Gene Trees Differ from the Concatenation Phylogeny



Gene Trees are Incongruent in Most Datasets

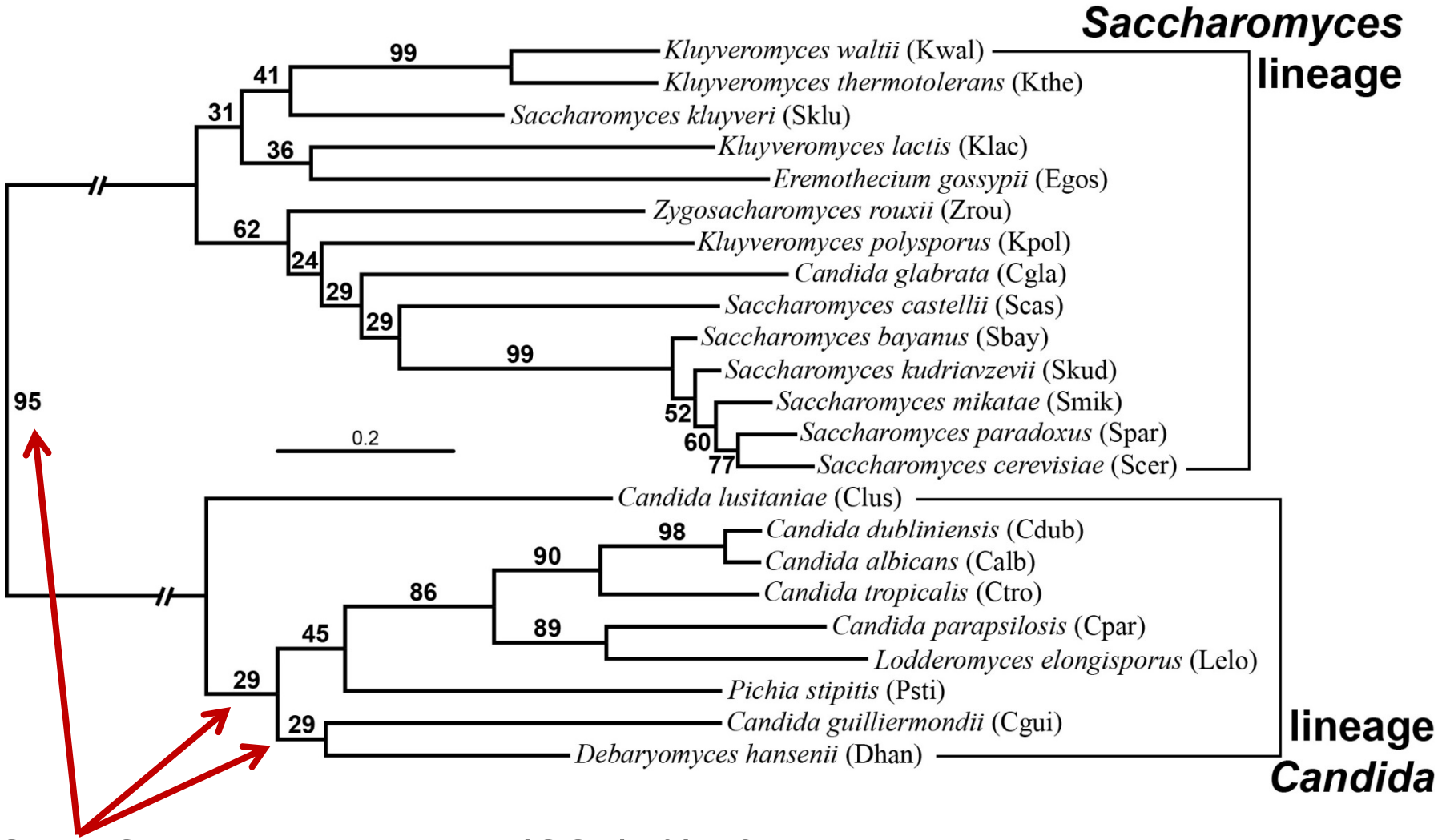
182 / 184 gene trees are distinct

440 / 447 gene trees are distinct

Zhong et al. (2013) Trends Plant Sci.

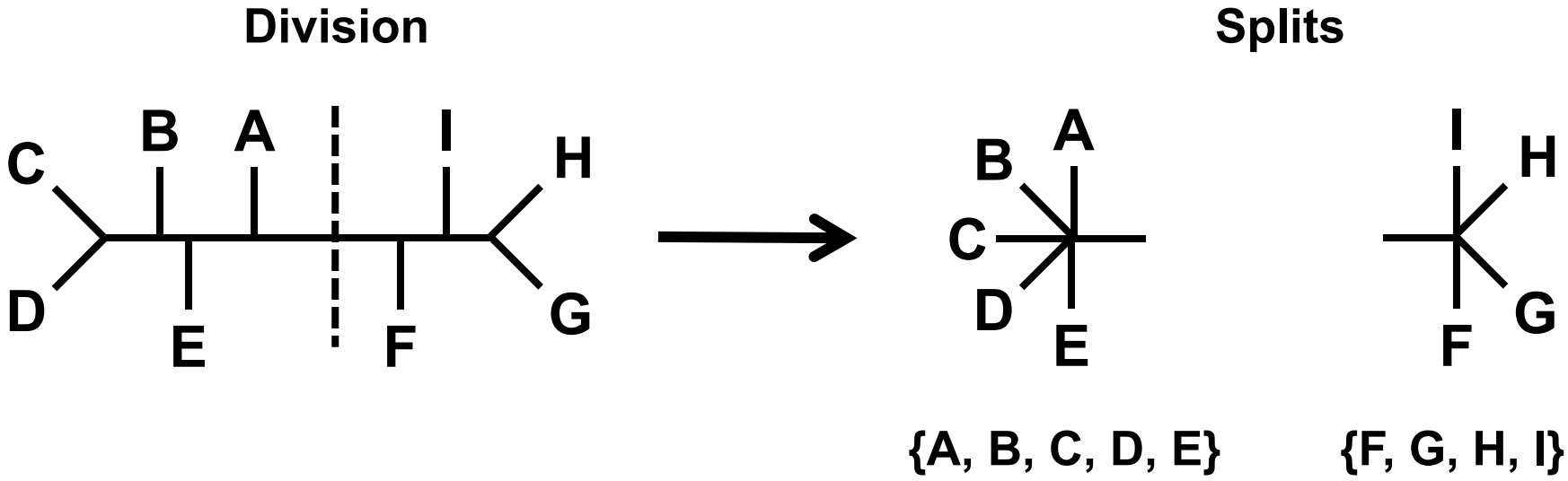
Song et al. (2012) PNAS

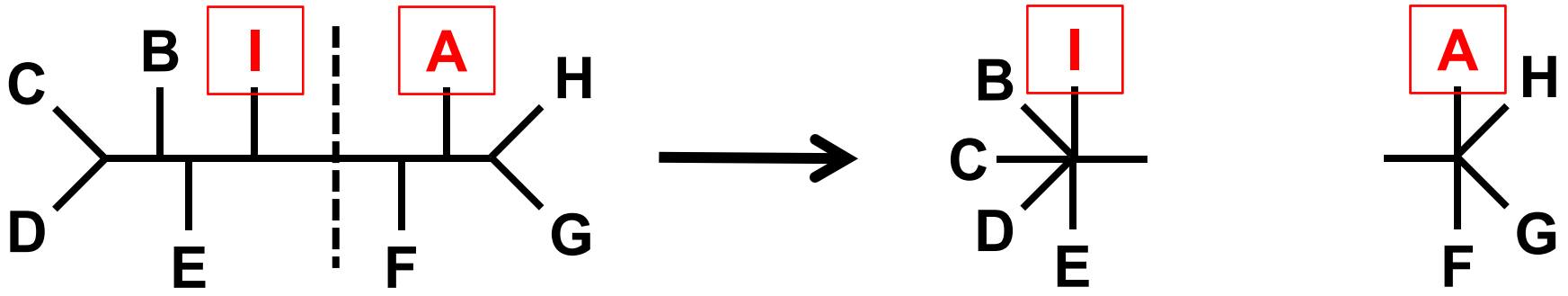
The Yeast Phylogeny Inferred by Majority-Rule Consensus



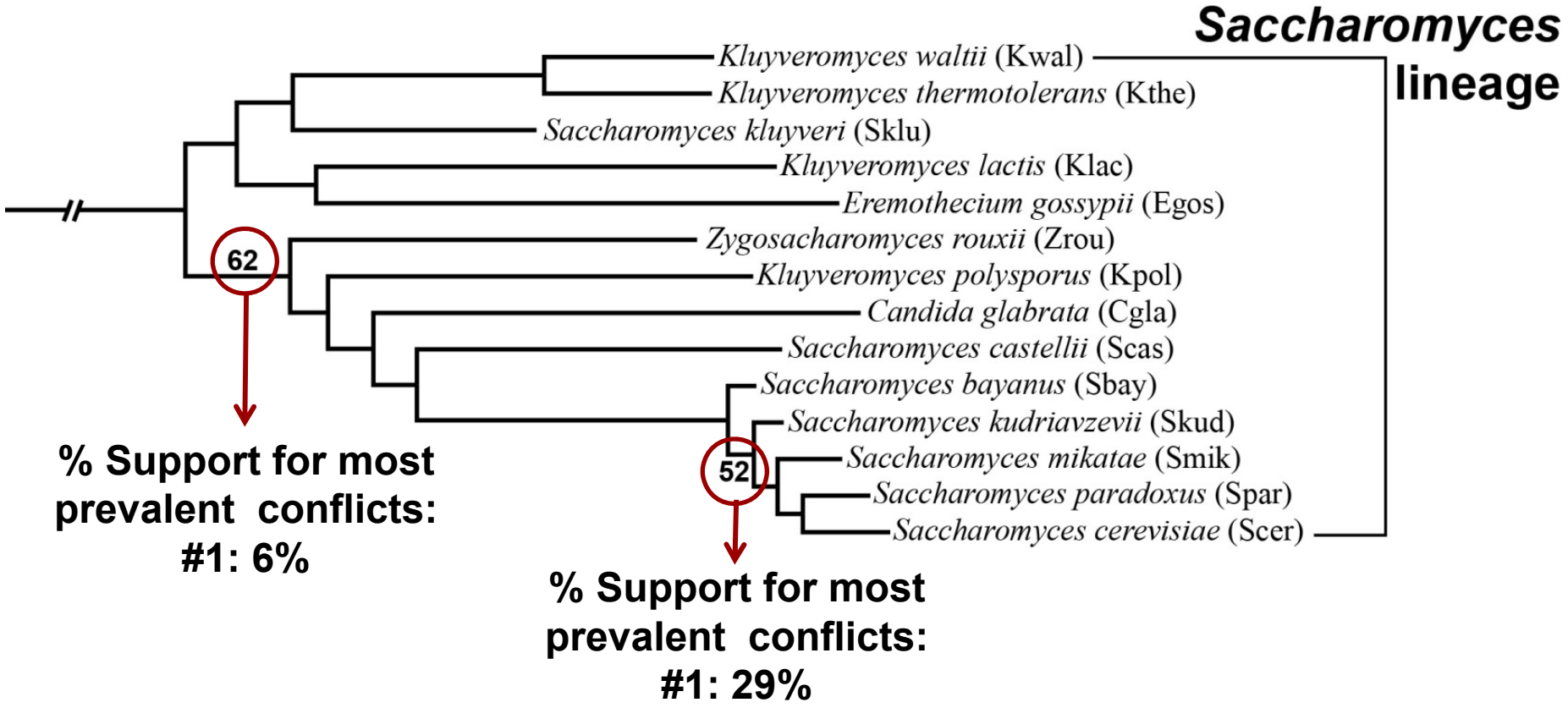
Gene Support Frequency (GSF): % of single gene trees supporting a given internode

Phylogenetic Trees are Sets of Splits





Measuring the Degree of Conflict for each Internode

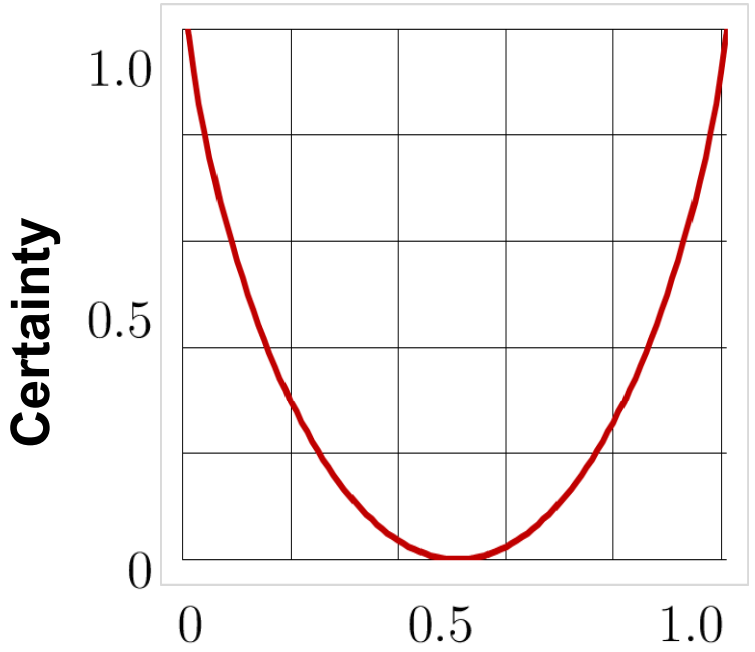


New Measures to Quantify Incongruence

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



Ratio of Support for Two Conflicting Splits

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**

```
          abcdefghij
Taxon A   1010011111
Taxon B   1010011110
Taxon C   1110111110
Taxon D   1110111110
Taxon E   1110111111
```

#b: {A,B} / {C,D,E}

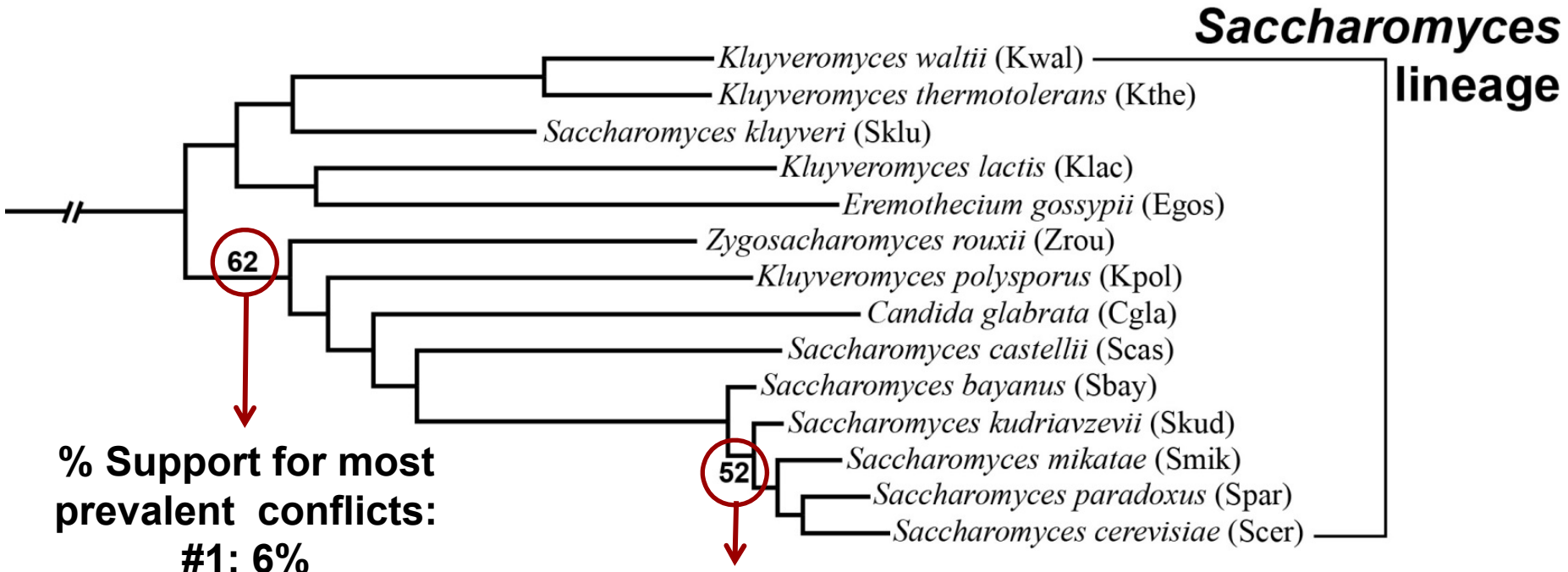
#e: {A,B} / {C,D,E}

#p: {A,E} / {B,C,D}

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

Salichos, Stamatakis & Rokas (2014) Mol. Biol. Evol.

IC Can Be More Informative Measure of Internode Support



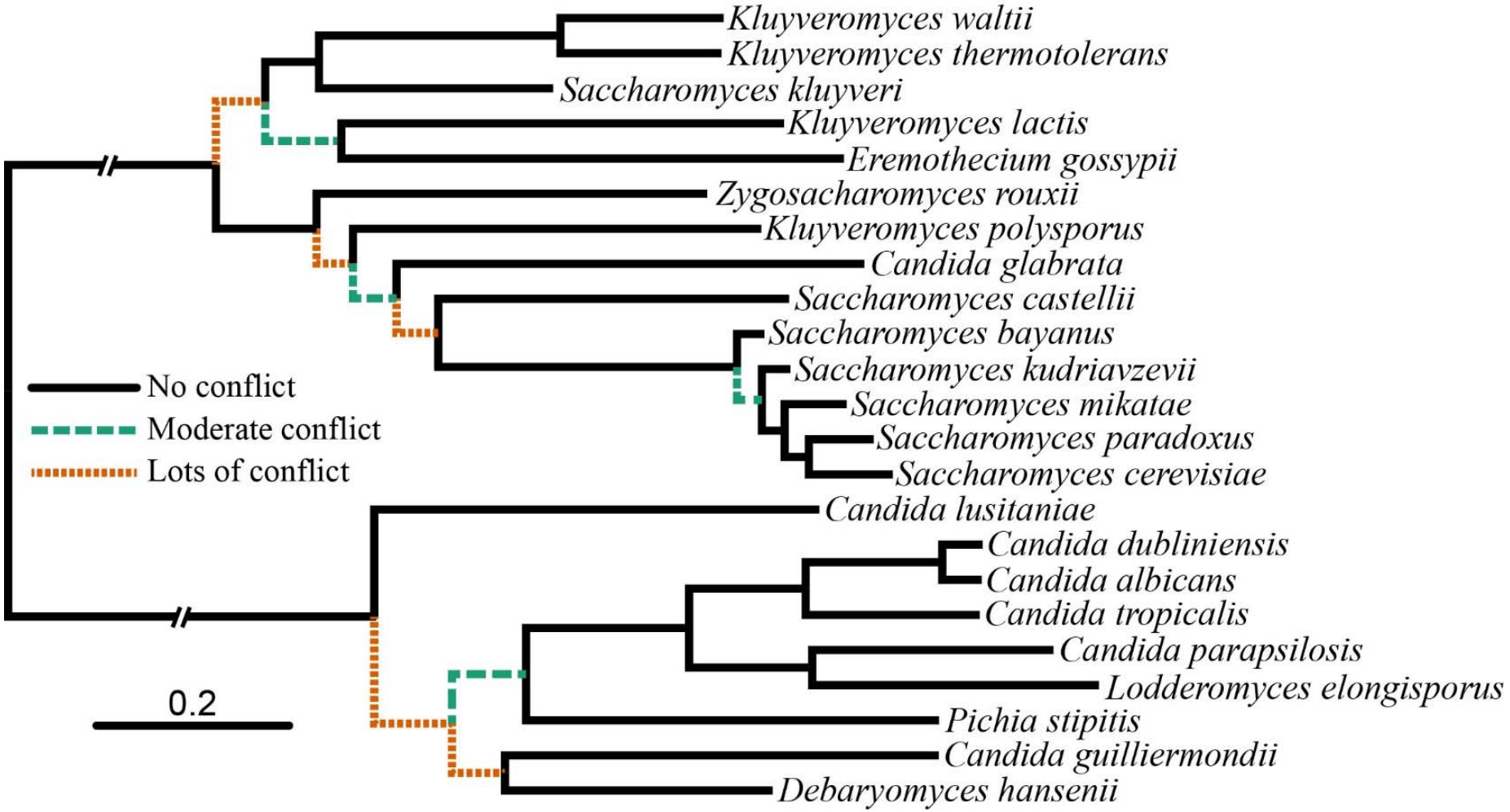
% Support for most prevalent conflicts: #1: 6%

IC value: 0.59

% Support for most prevalent conflicts: #1: 29%

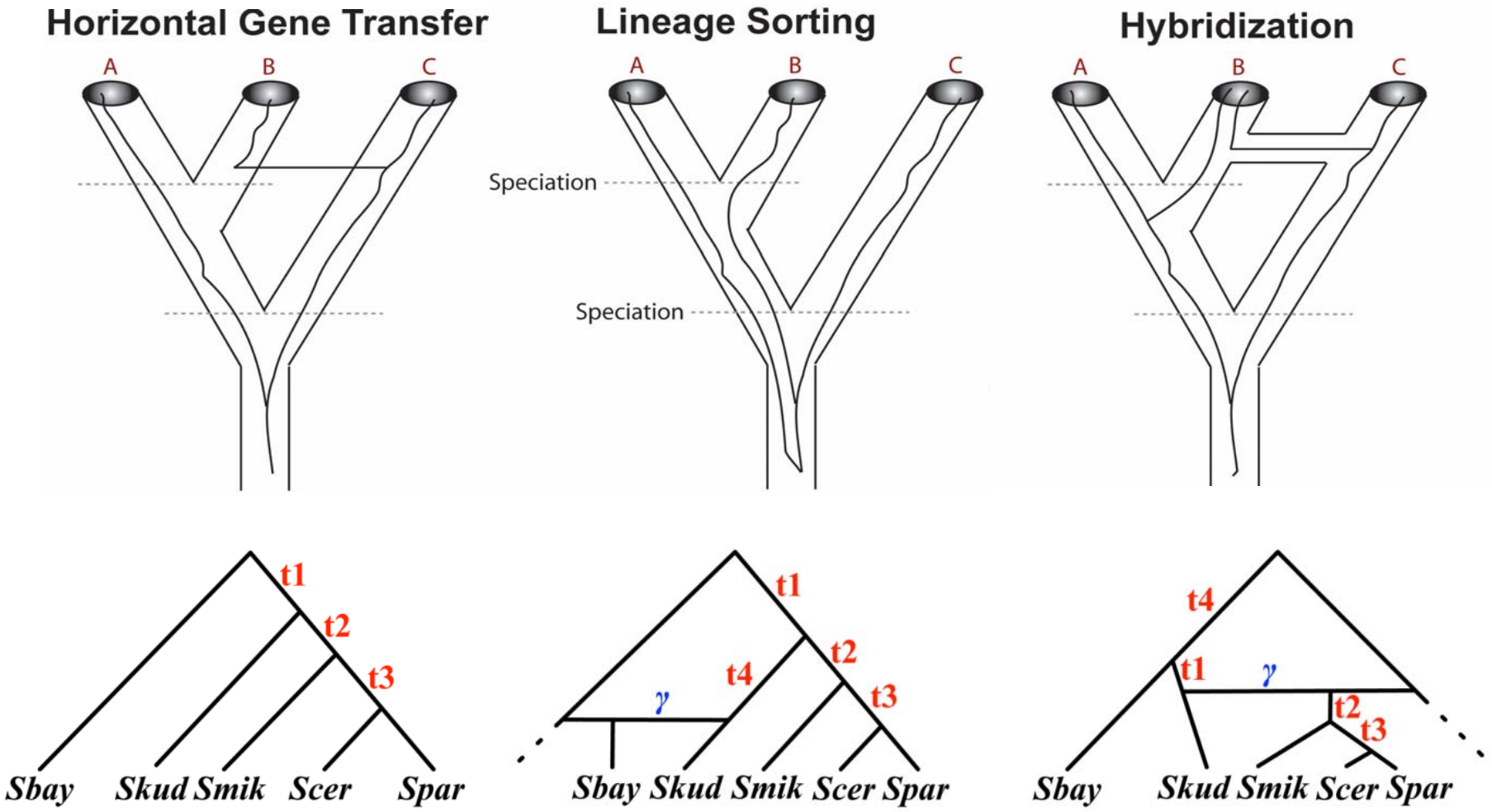
IC value: 0.06

IC Reveals that There is Rampant Conflict

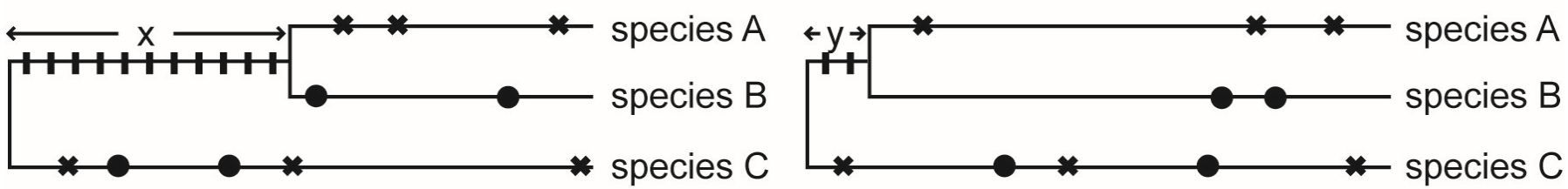


Salichos & Rokas (2013) Nature

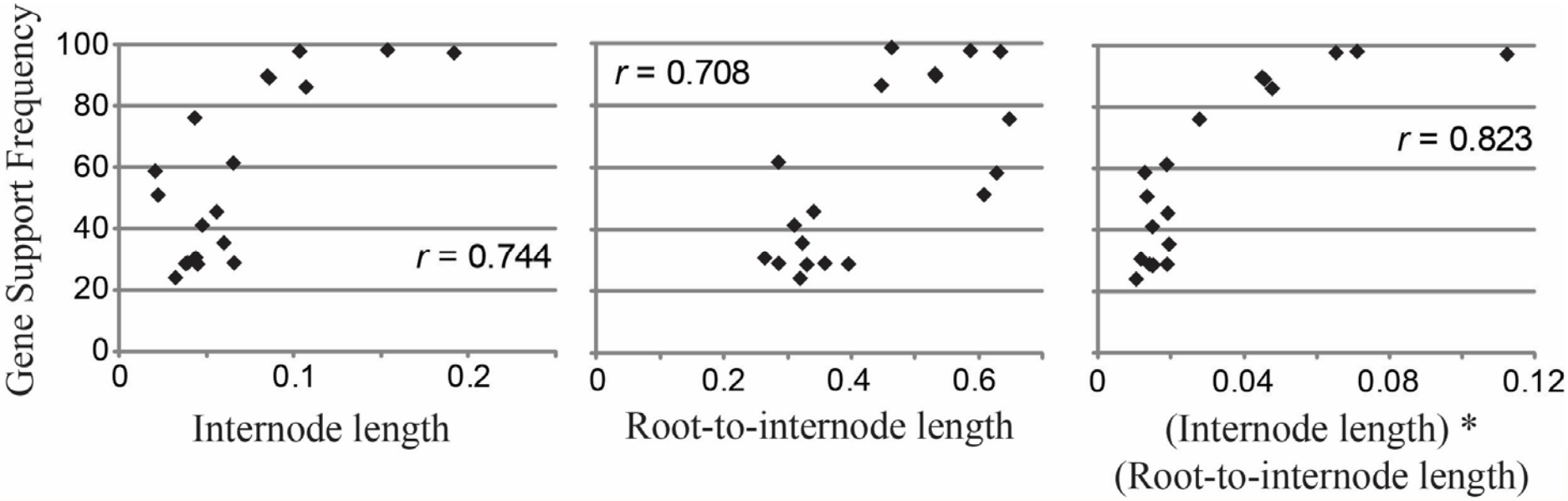
What So Much Incongruence? Biological Factors



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**
Removing sites containing gaps		
All sites with gaps excluded	**7.91**	**0 | 7**
Removing fast-evolving or unstable species		
C. lusitaniae	**8.15**	**1 | 2**
C. glabrata	**8.30**	**2 | 2**
E. gossypii*, *C. glabrata*, *K. lactis	**7.88**	**1 | 3**
Selecting genes that recover specific clades		
[*C. tropicalis*, *C. dubliniensis*, *C. albicans*]	**8.62**	**0 | 0**
Selecting the most slow-evolving genes		
100 slowest-evolving genes	**6.76**	**2 | 9**

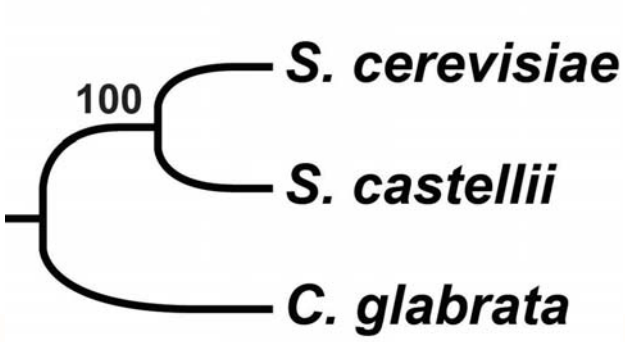
Salichos & Rokas (2013) Nature


WHAT CAN I DO?!!

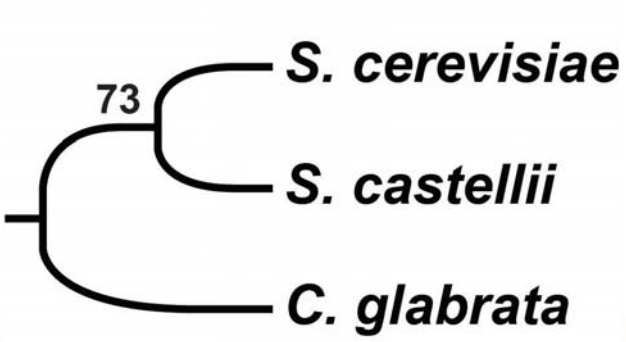
What Do We Do Then?

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

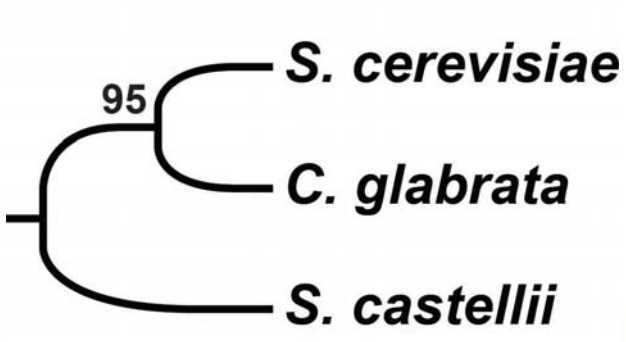
average BS ≥60%



average BS ≥70%

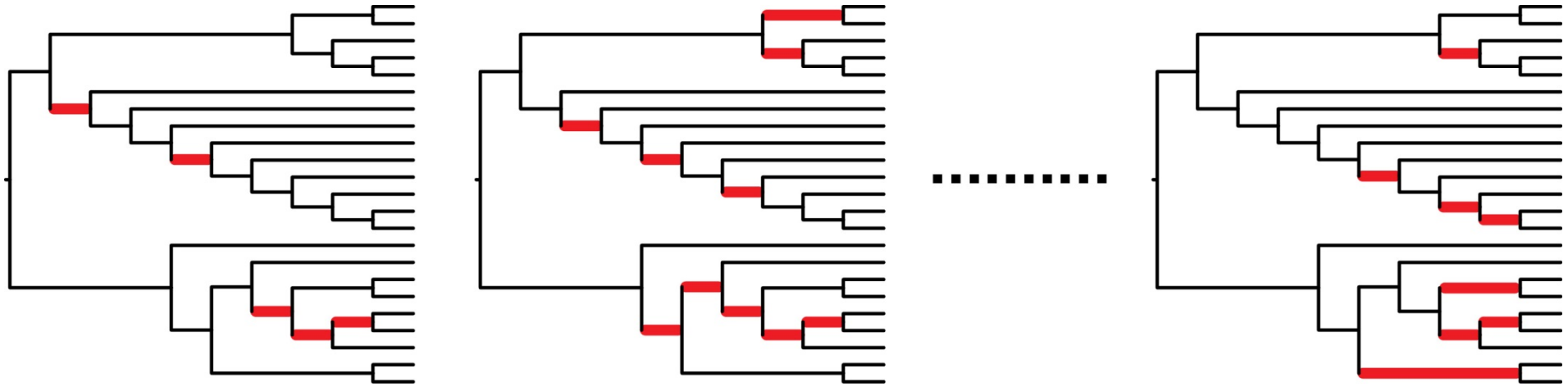


average BS ≥80%



Salichos & Rokas (2013) Nature

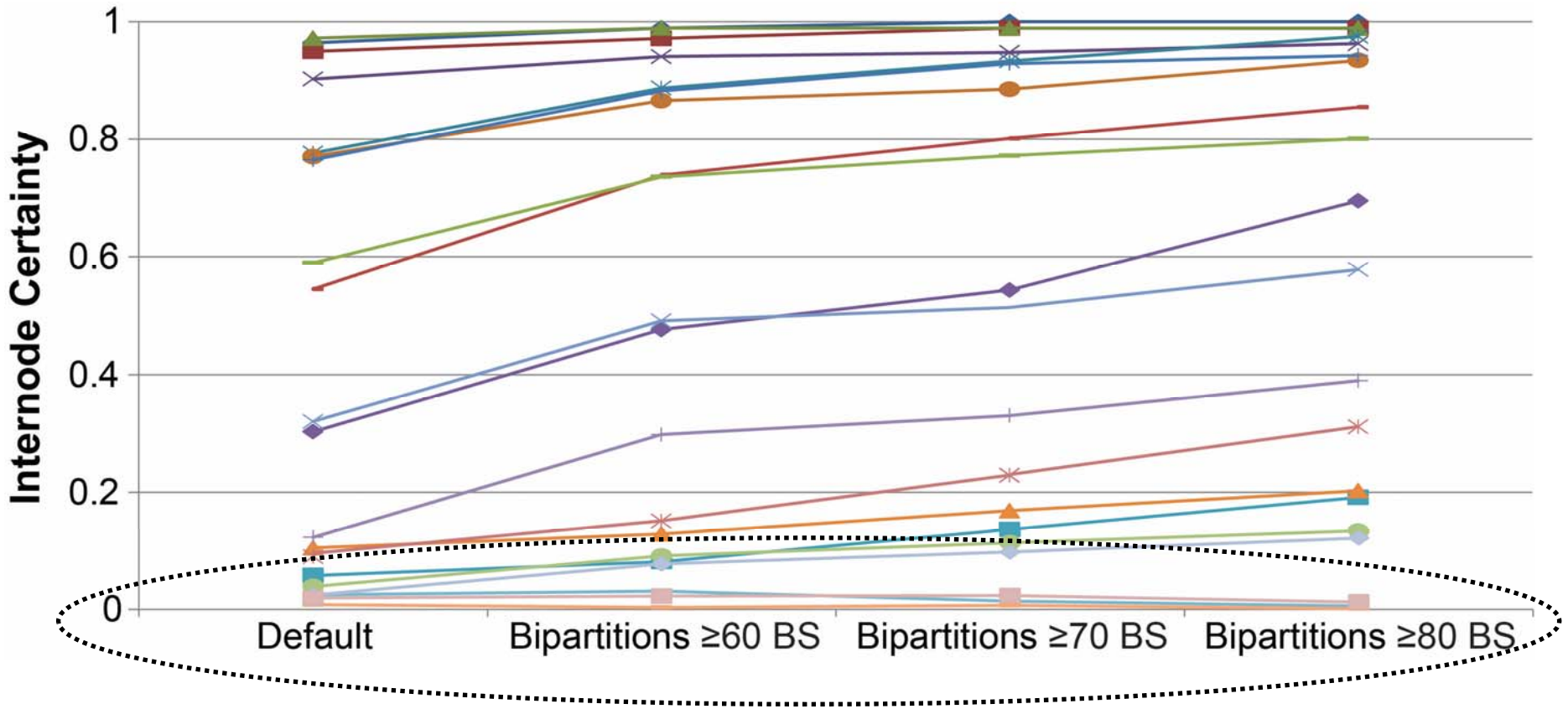
Selecting Specific Bipartitions Dramatically Improves Phylogeny



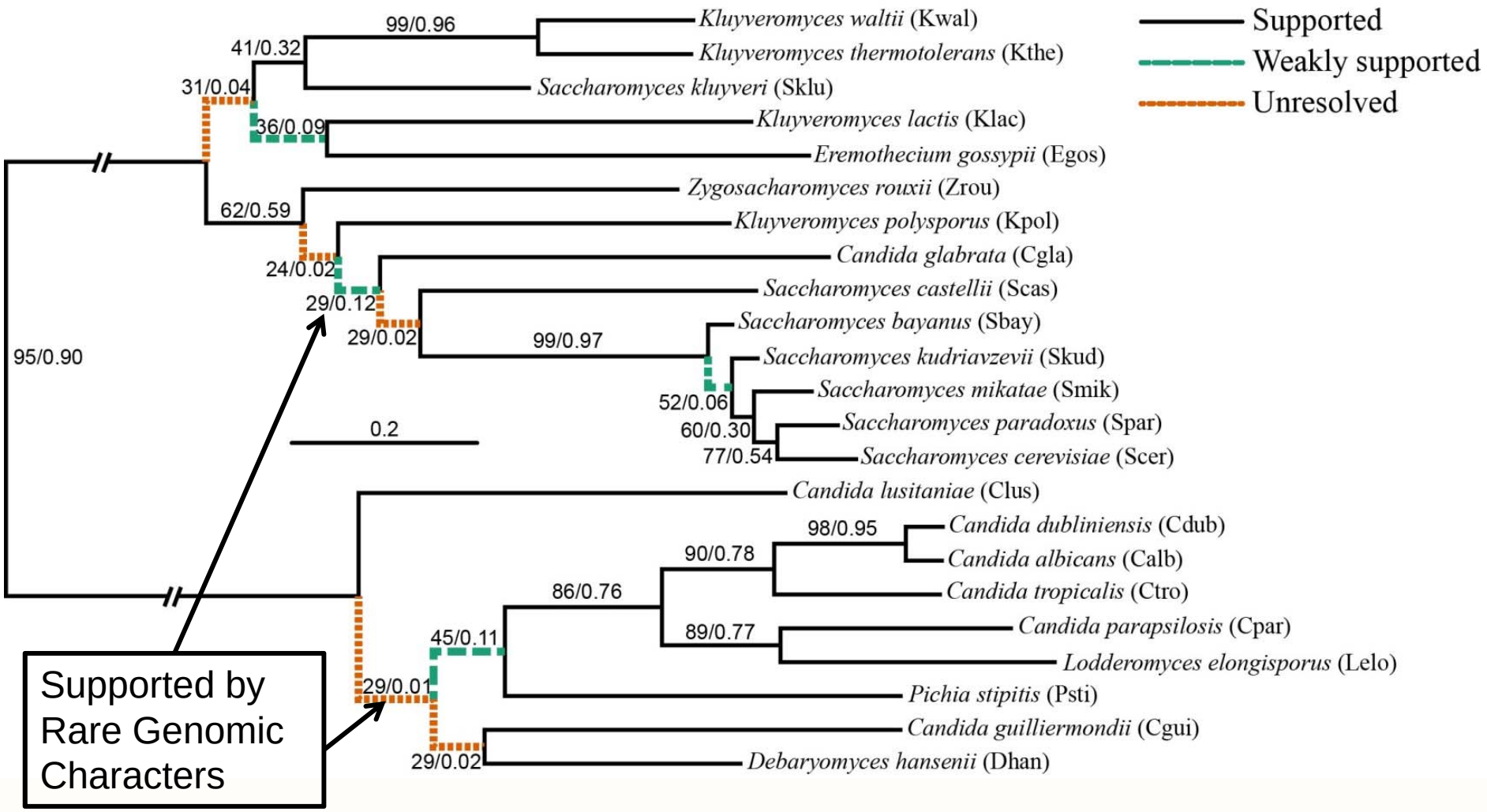
Treatment	Tree Certainty	# of Internodes where IC increased | decreased
Default analysis	**8.35**	**n/a**
Selecting genes whose bootstrap consensus trees have high average support		
All bipartitions with BS ≥ 60%	**10.11**	**14 | 0**
All bipartitions with BS ≥ 70%	**10.70**	**16 | 0**
All bipartitions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

Salichos & Rokas (2013) Nature

Least Supported Internodes Harbor the Most Conflict

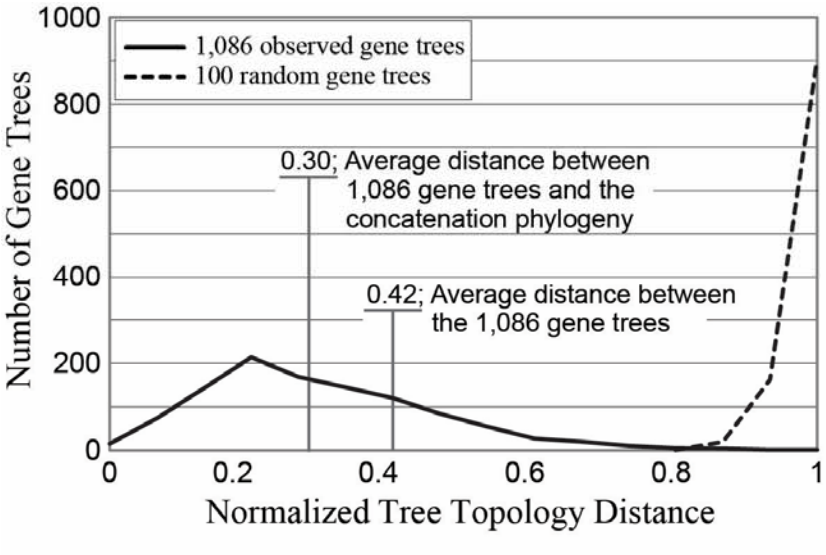


The Status of the Yeast Phylogeny

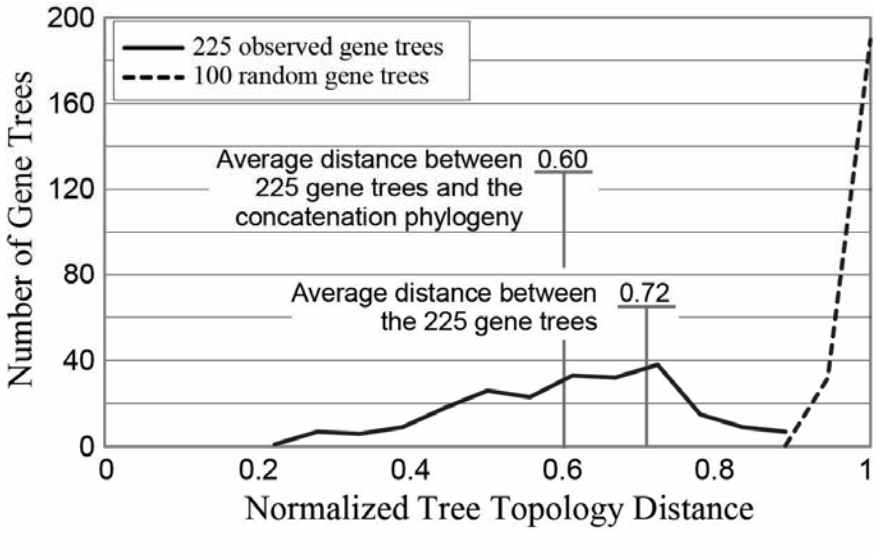


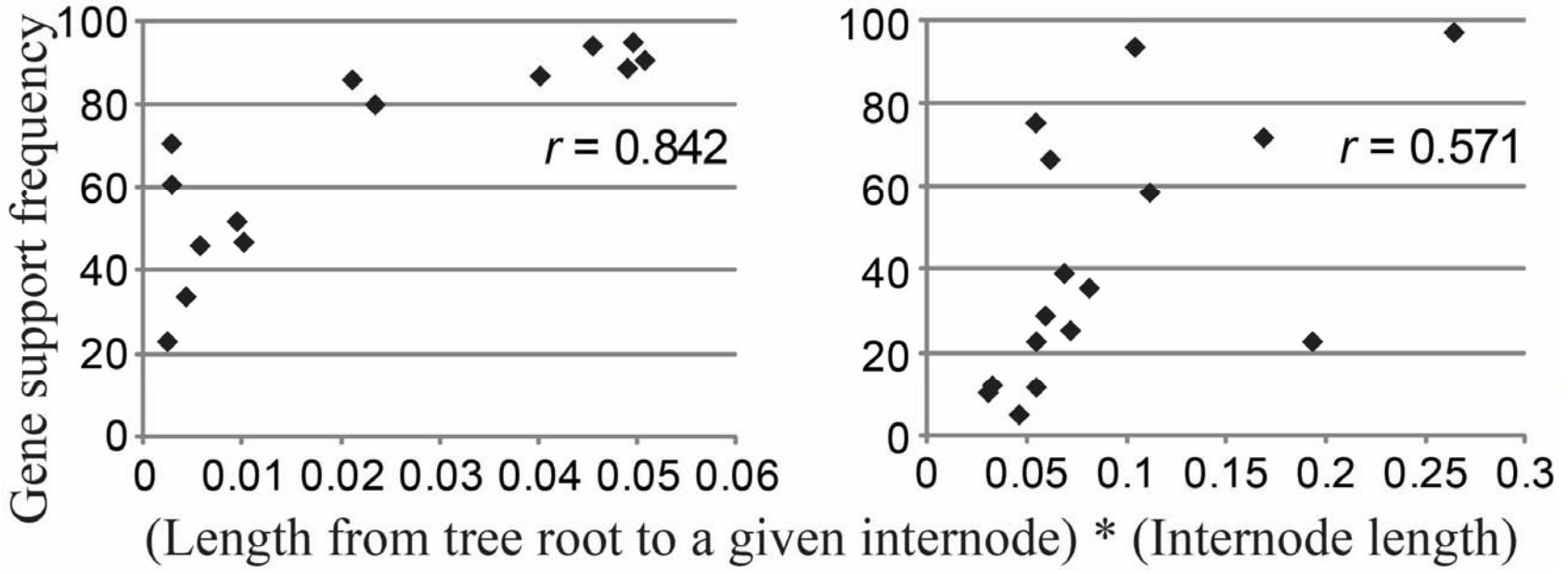
The Same is True for Vertebrate and Metazoan Datasets

Vertebrates (1,086 genes, 18 taxa)



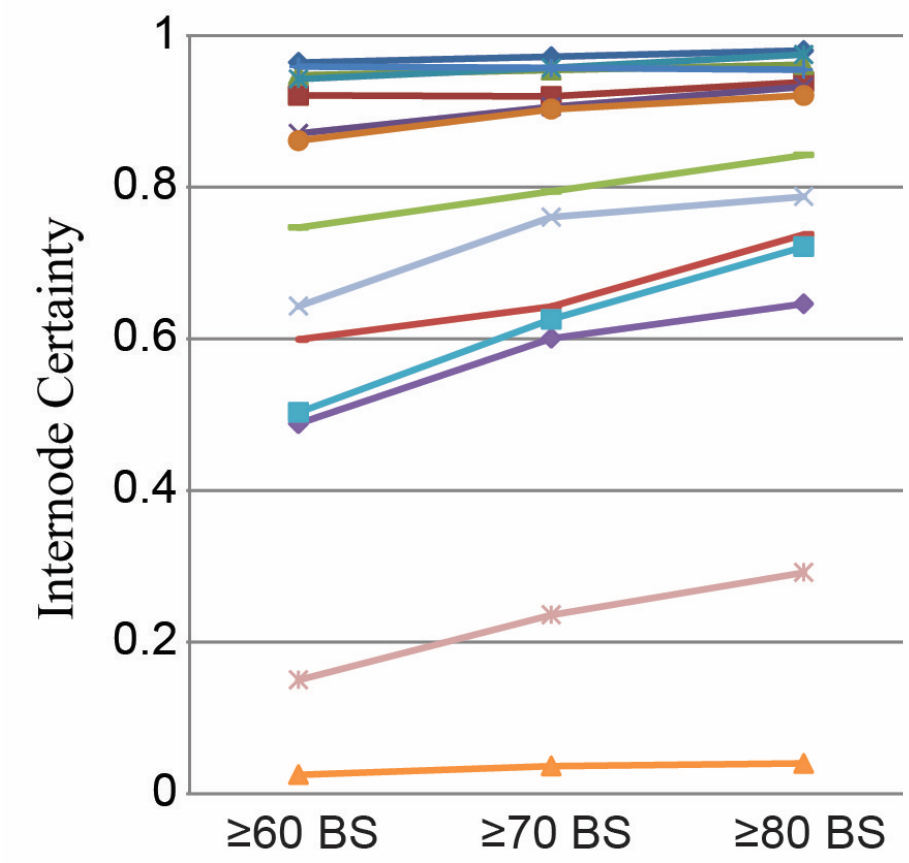
Animals (225 genes, 21 taxa)



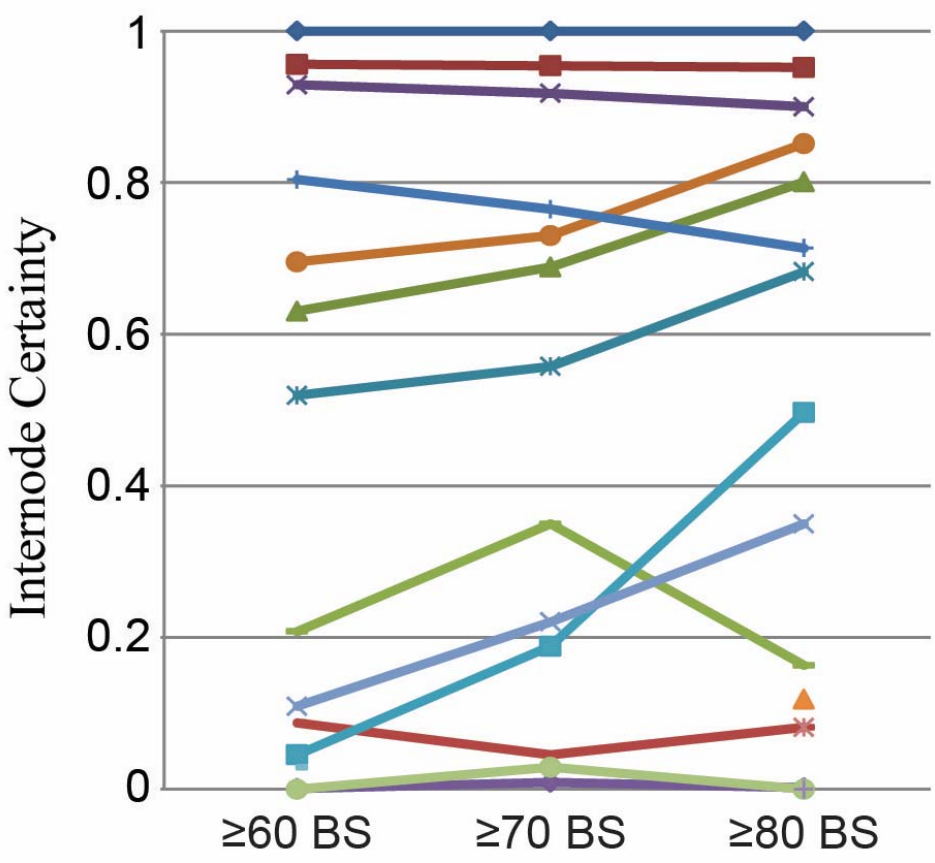


The Same is True for Vertebrate and Metazoan Datasets

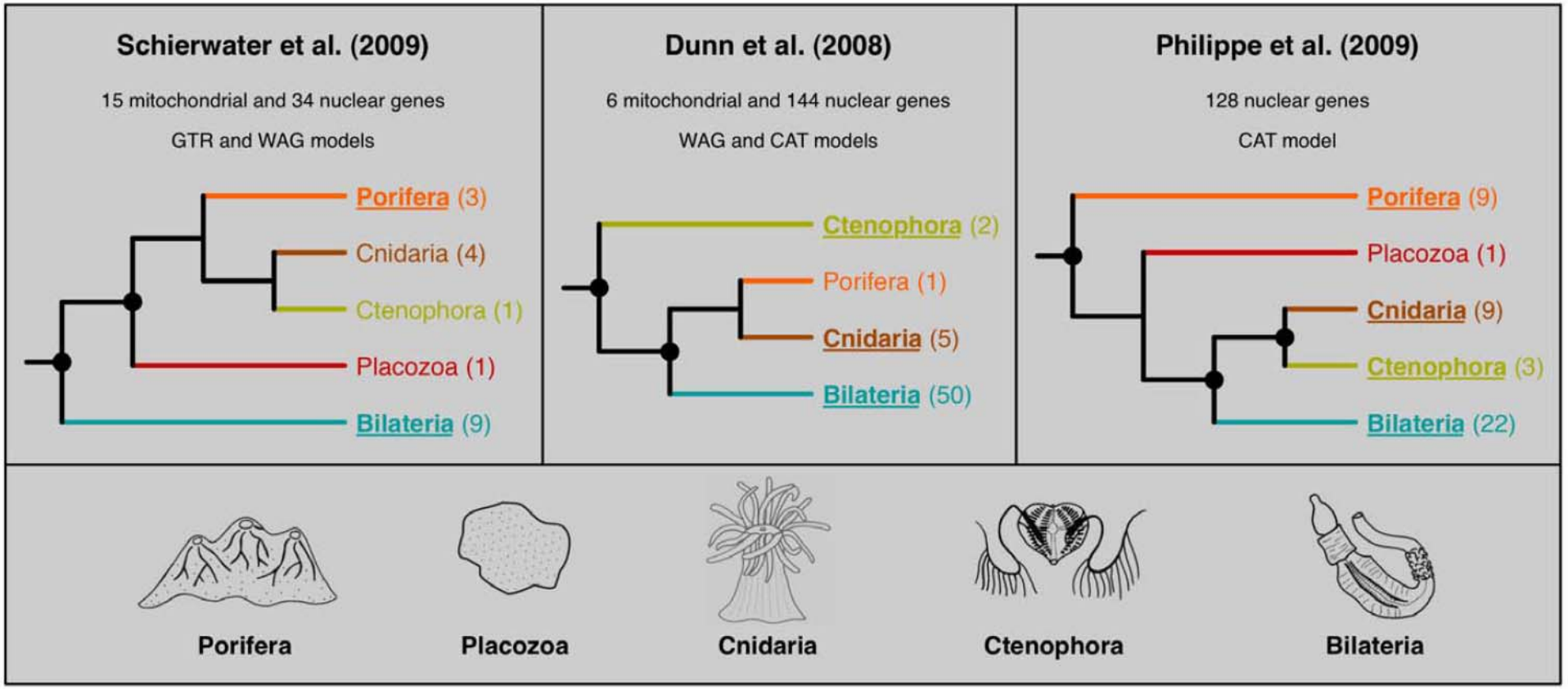
Vertebrates (1,086 genes, 18 taxa)



Animals (225 genes, 21 taxa)



Incongruence in Deep Time



Philippe et al. (2011) PLoS Biol.

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**

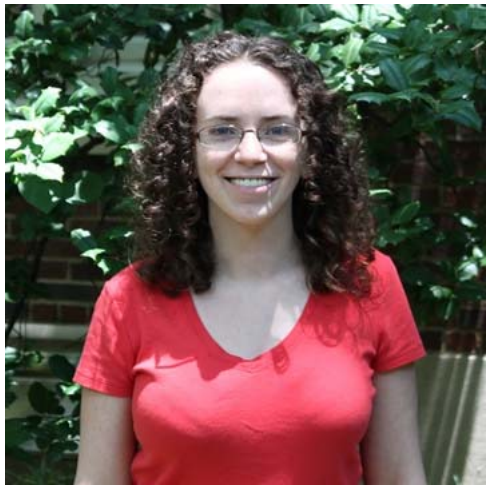
“One can use the most sophisticated audio equipment to listen, for an eternity, to a recording of white noise and still not glean a useful scrap of information”

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

Leonidas Salichos

National Science Foundation
WHERE DISCOVERIES BEGIN