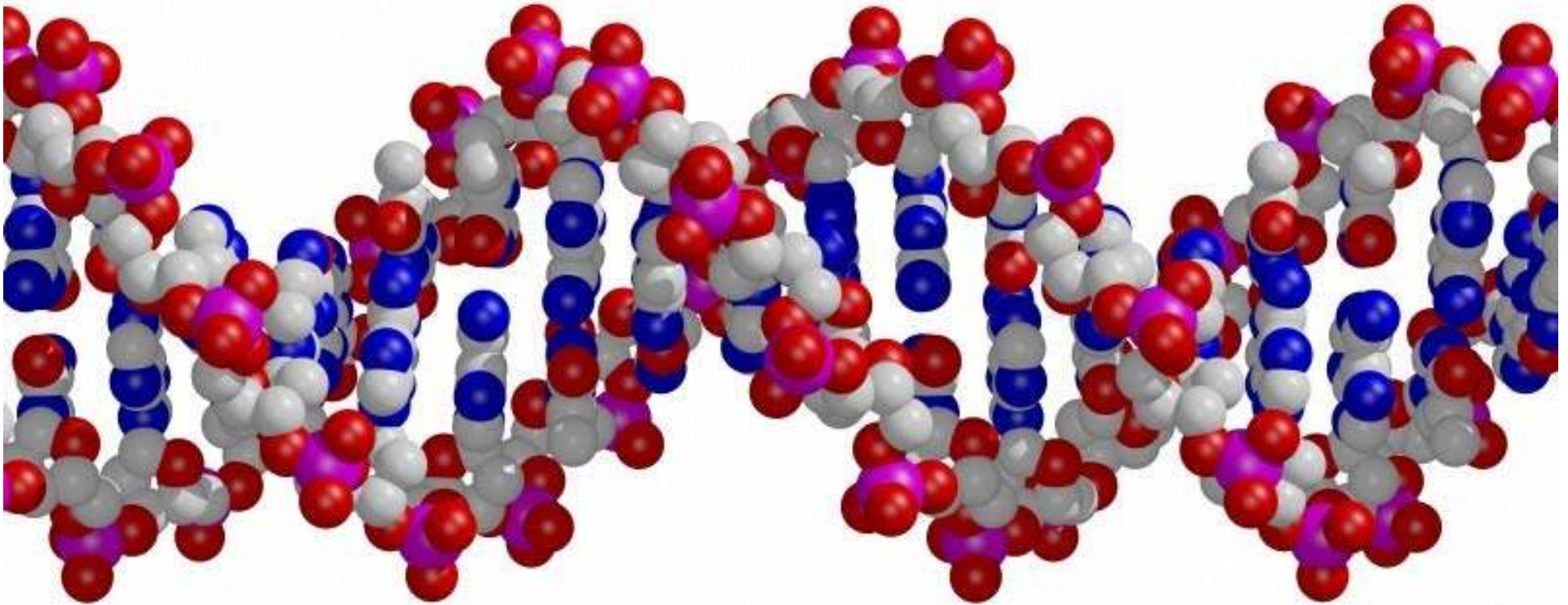


Evolutionary Genomics



Antonis Rokas
Department of Biological Sciences
Vanderbilt University



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

Lecture Outline

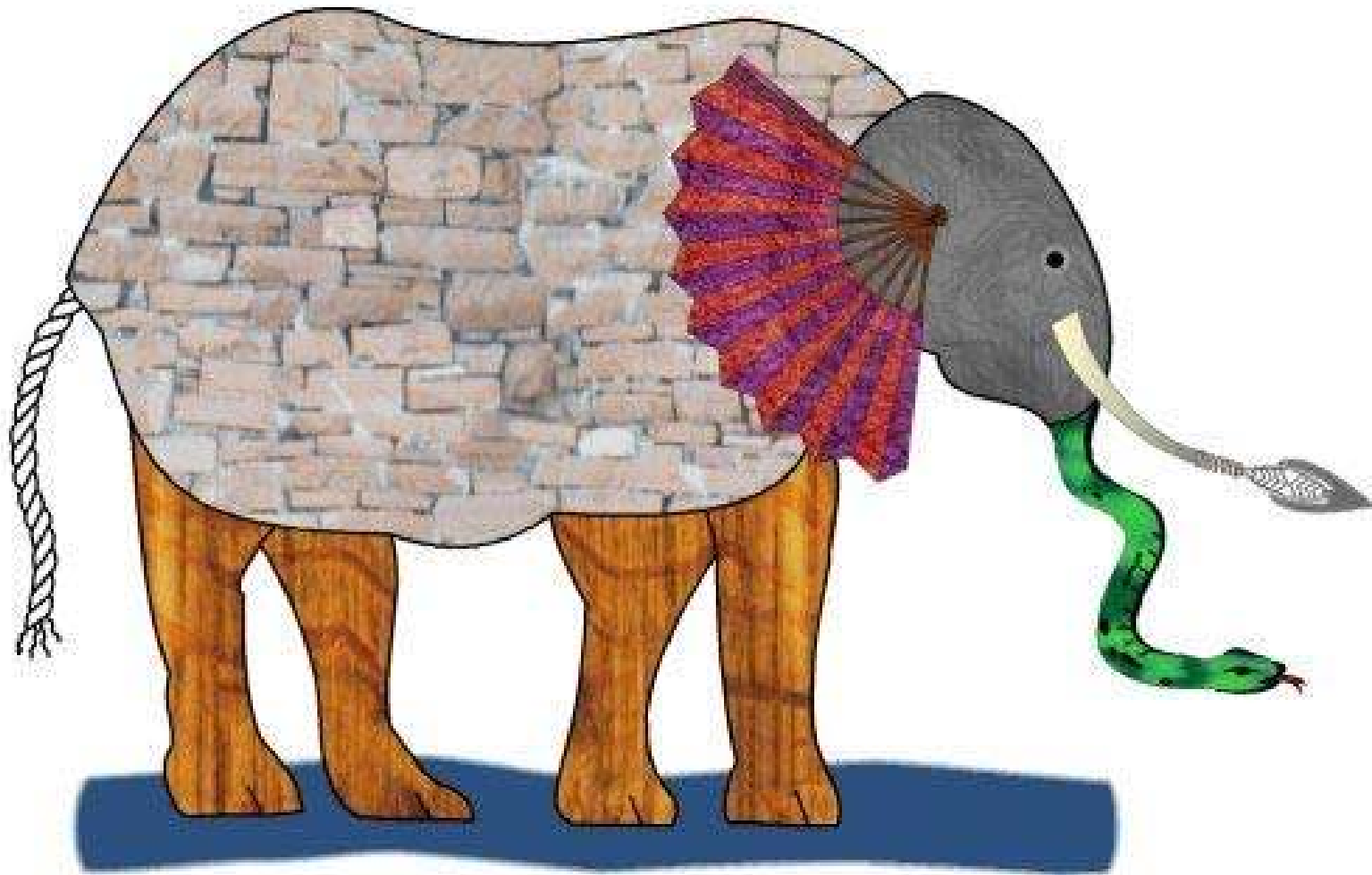
❖ **Introduction to Evolutionary Genomics**

❖ **Population Genomics**

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

❖ **Phylogenomics**

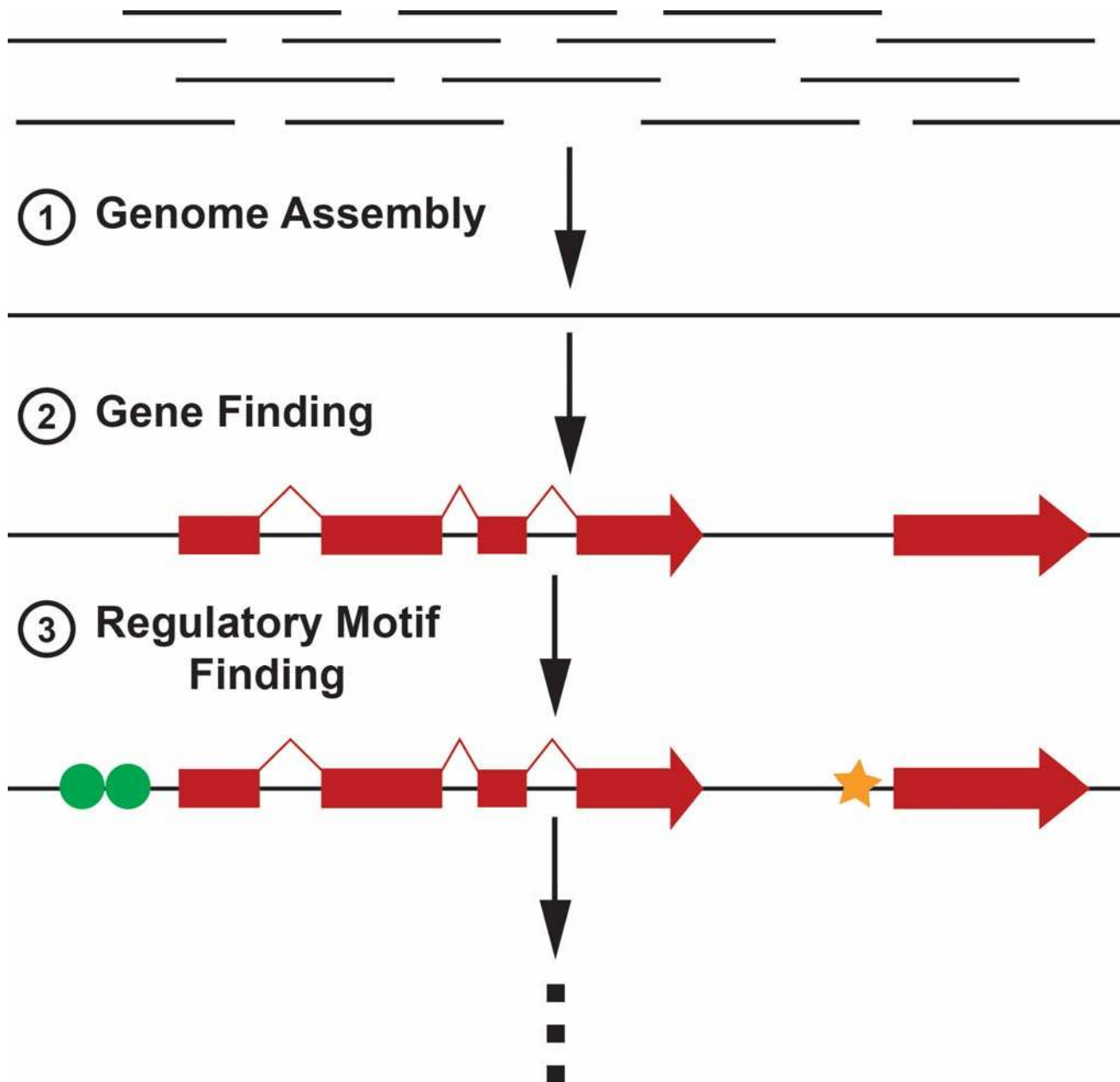
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACCTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTGAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAGGCACCCATCATACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAGCGGTGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTGCGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTGAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTGCGCTCCAGCTCCCTAGCACTGGCAGGGTTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTAAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCCTCACAGGTAAGTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGTTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACCTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACAGCCGCATCCCCTACAAAATCTACCCCAAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCGCGGACTCCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCGAGGGGCCACCAGGAACCATGTTT
CAGTCTGACAATAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAACAAGGAAA
AGGGGATACAGTGGCGATTTCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT

Understanding the Genome Requires Tools



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTGAGTGTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATAACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAGCGGTGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTGCGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTGCGCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGTTTCTCTCACTACTCAGTTCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTAAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGATCTGCTCTCTCCAAGTCT
GCGATATTGCTTATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCGTACAGGTAAGTCTGAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGTAAGCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTAATAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTTCGTAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACCTTAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCTACAAAATCTACCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCGCGGACTCCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
ATCGATCCCATTGTGAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCGAGGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCTACATTCATATGGGCCAGCGATTGGAACCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT

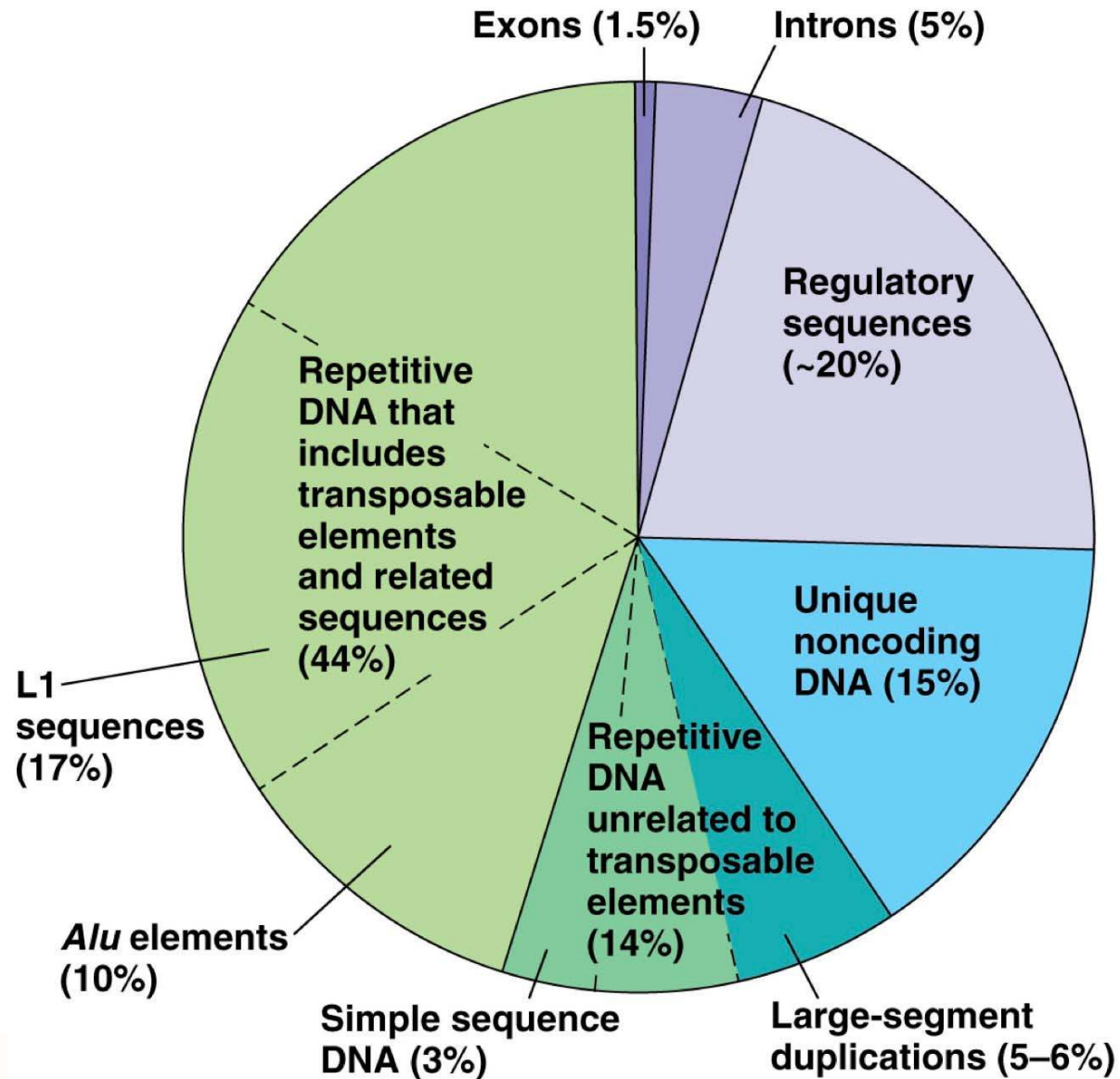
Transposon

Protein Binding Site

Exon

Intron

Organization of the Human Genome

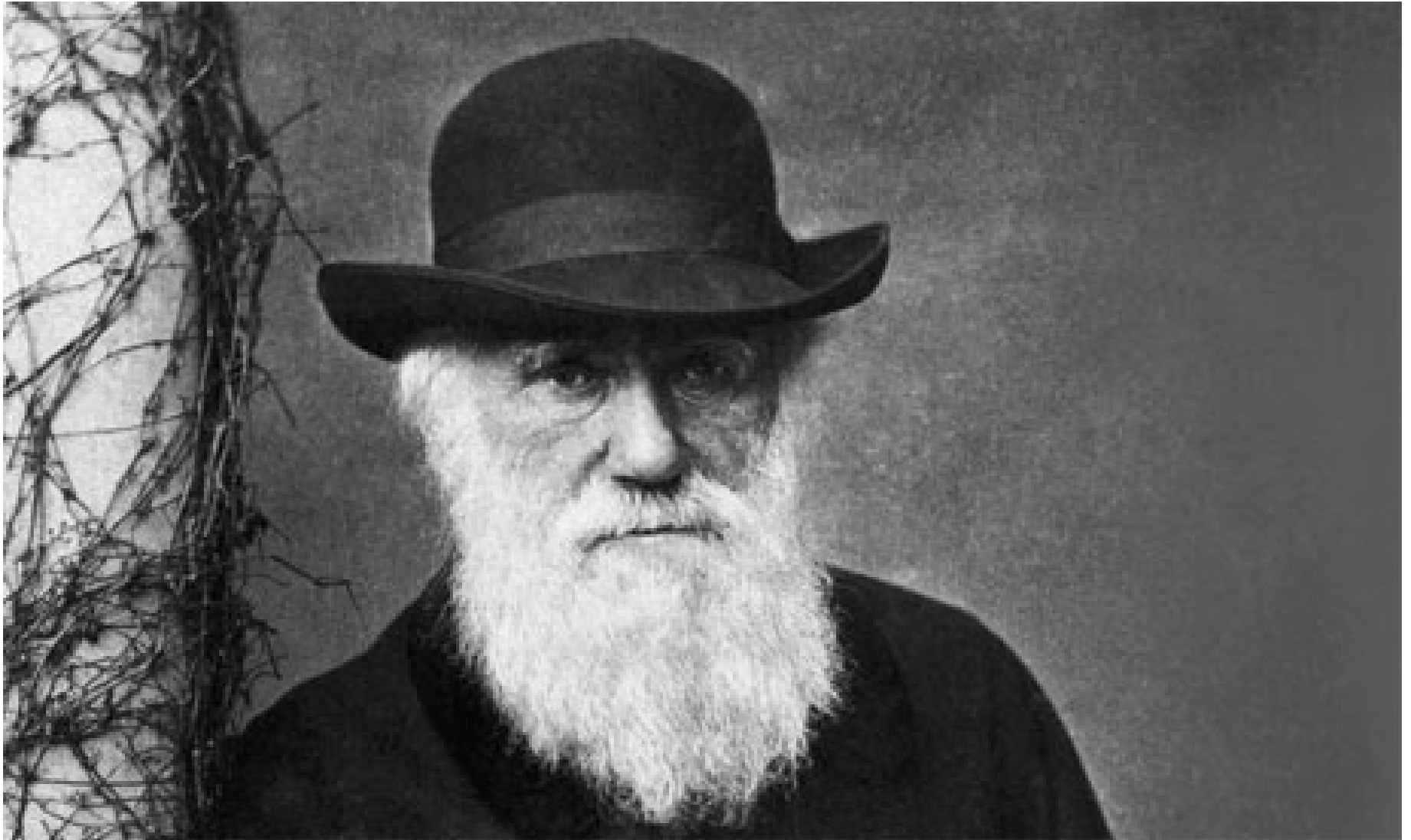


© 2011 Pearson Education, Inc.

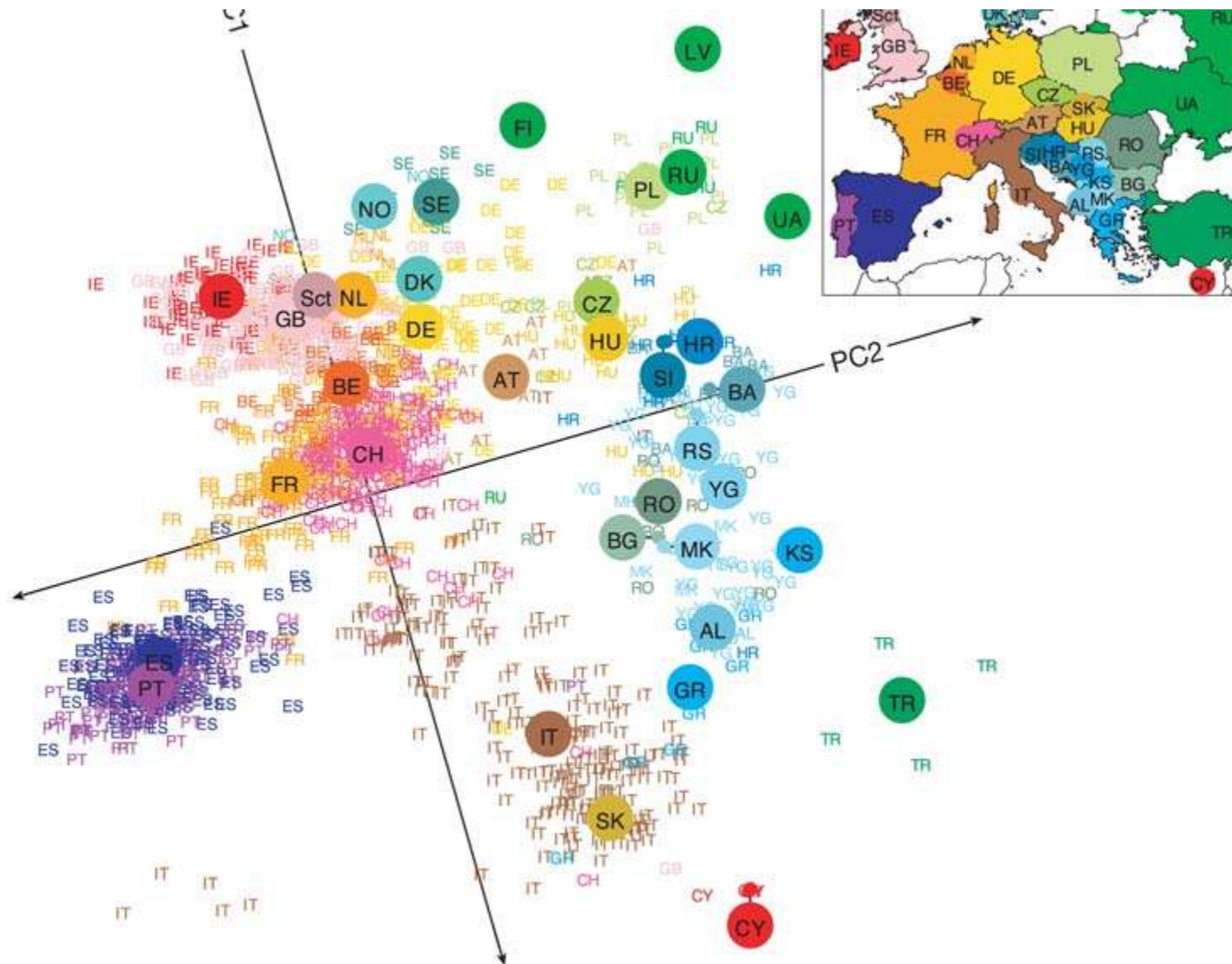


<http://todd.jackman.villanova.edu/HumanGenome.jpg>

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	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	
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	

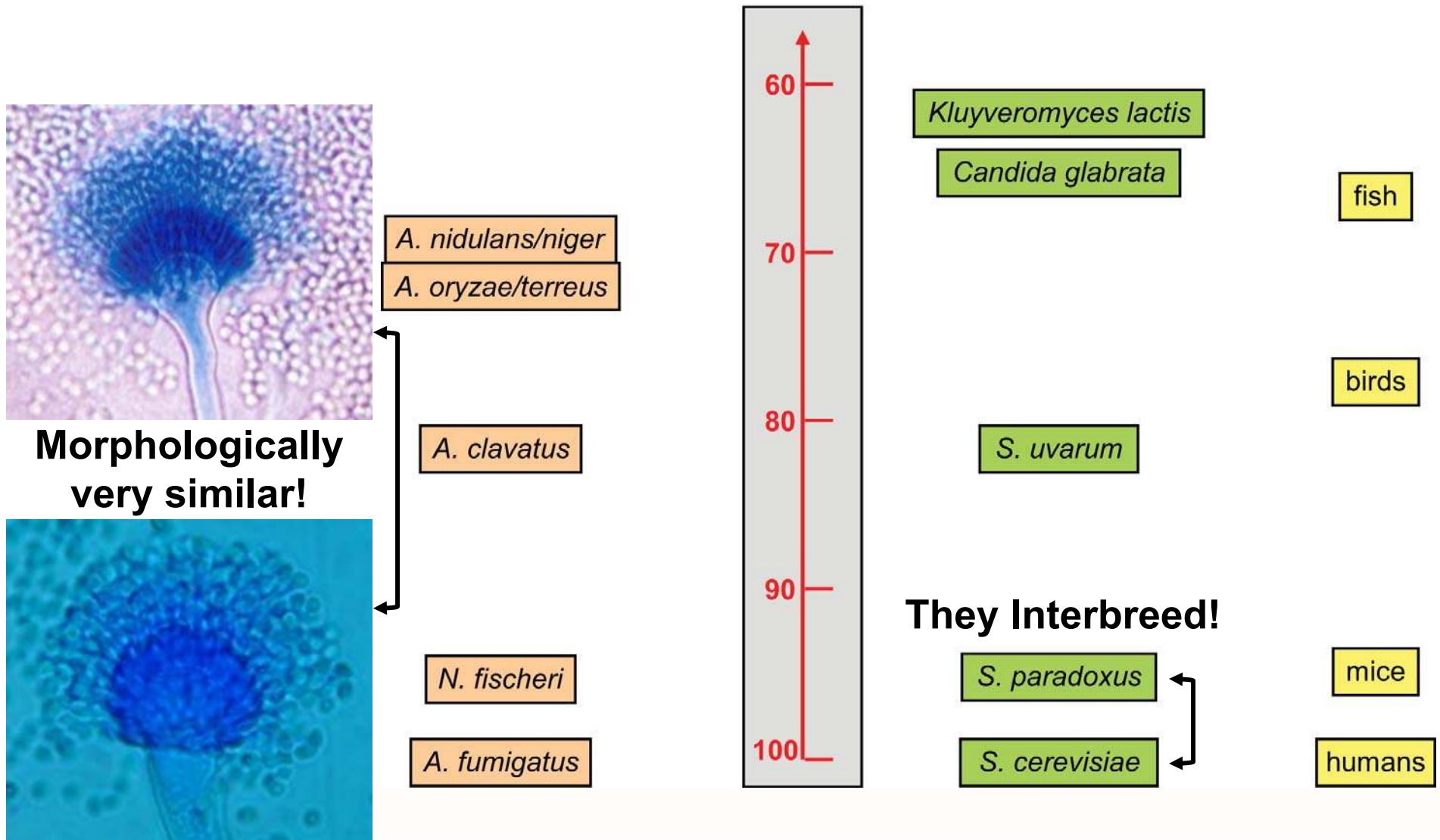
In the European population,
involved in skin pigmentation

In the West African population,
related to Lassa virus infection



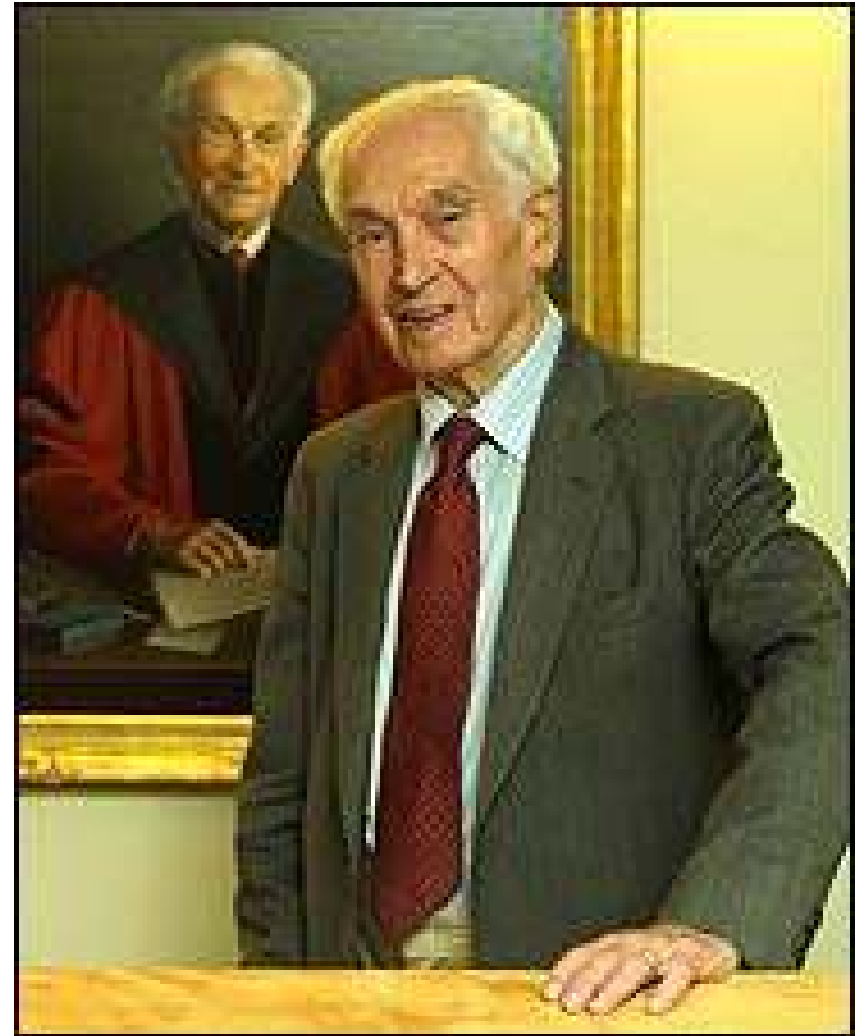
Genomes Provide a Common Yardstick for Comparison

Average proteome sequence similarity



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

Ernst Mayr, 1963



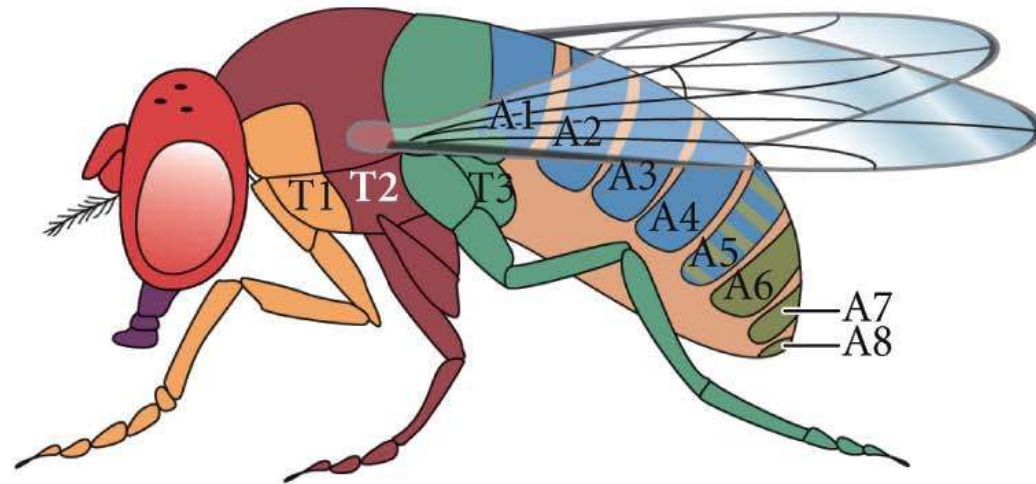
What Makes Us Sick Is the Stuff of Life

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	-			Fukuyama MD ⁺ -FCMD	-			Rubinstein-Taybi-CREBBP
-			Fanconi's Anemia C-FANCC	+			Huntington-HD	+			Saethre-Chotzen-TWIST
-			Fanconi's Anemia G-FANCG	+			Limb Girdle MD ⁺ 2A-CAPN3	-			Septo-optic Dysplasia-HESX1
+			HNPCC*-MSH2	+			Limb Girdle MD ⁺ 2B-YSF	+			Simpson-Golabi-Behmel-GPC3
+			HNPCC*-MSH3	-			Limb Girdle MD ⁺ 2E-BSG	+			Townes-Brockes-SALL1
+			HNPCC*-MSH6	+			Lissencephaly, X-Linked-DCX	-			Treacher-Collins-TCOF1
+			HNPCC*-MLH1	+			Lowe Oculocerebroren.-OCRL	+			VMCM-TEK
+			HNPCC*-PMS2	-			Machado-Joseph-MJD1	+			Wardenburg-PAX3
-			KIT	+			Miller-Dieker Lissen.-PAF	+			Zellweger-PEX1



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

The Drosophila Body- Building Genes



Antennapedia complex

bithorax complex

lab *Pb* *Dfd* *Scr* *Antp*

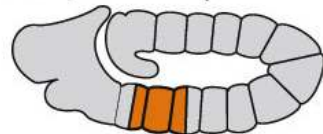
Ubx *abdA* *AbdB*



labial (*lab*)



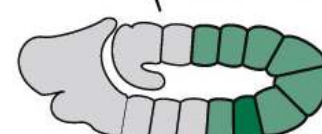
Deformed (*Dfd*)



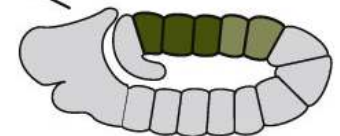
Sex combs reduced (*Scr*)



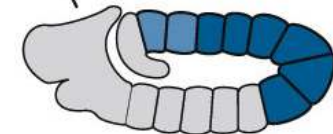
Antennapedia (*Antp*)



Ultrabithorax (*Ubx*)

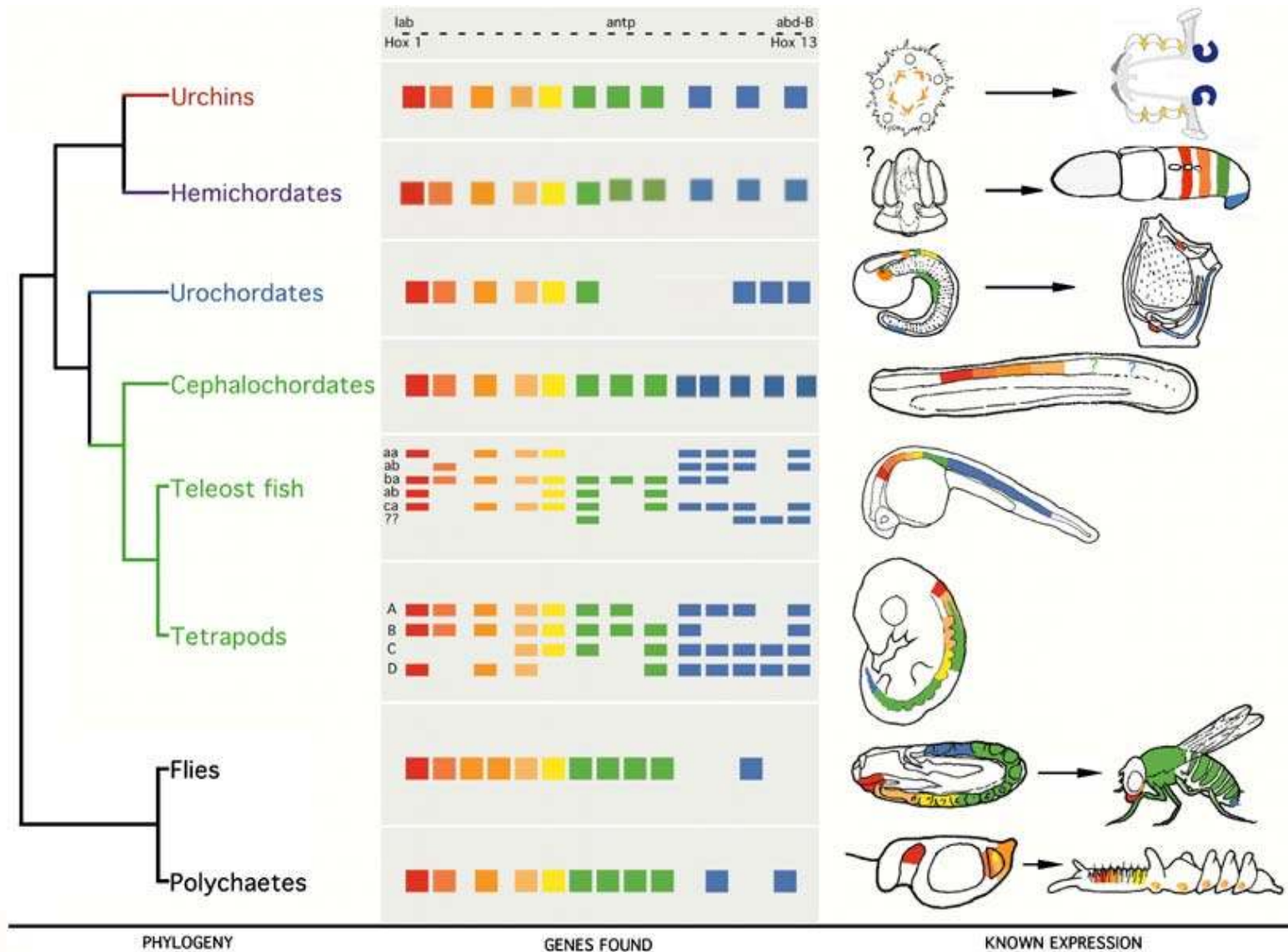


Abdominal B (*AbdB*)



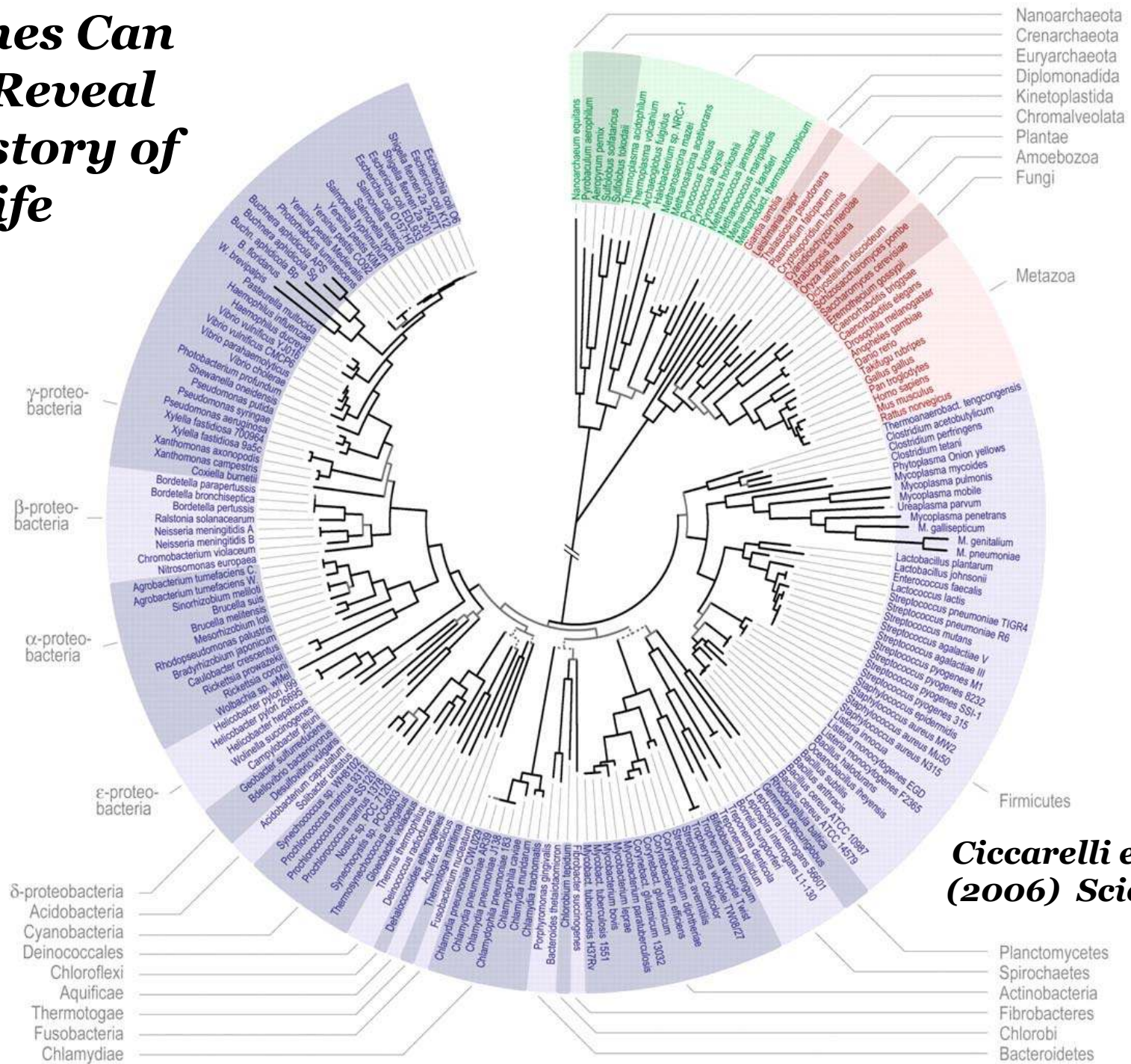
abdominal A (*abdA*)

Animal Bodies are Built from the Same Genetic Toolkit



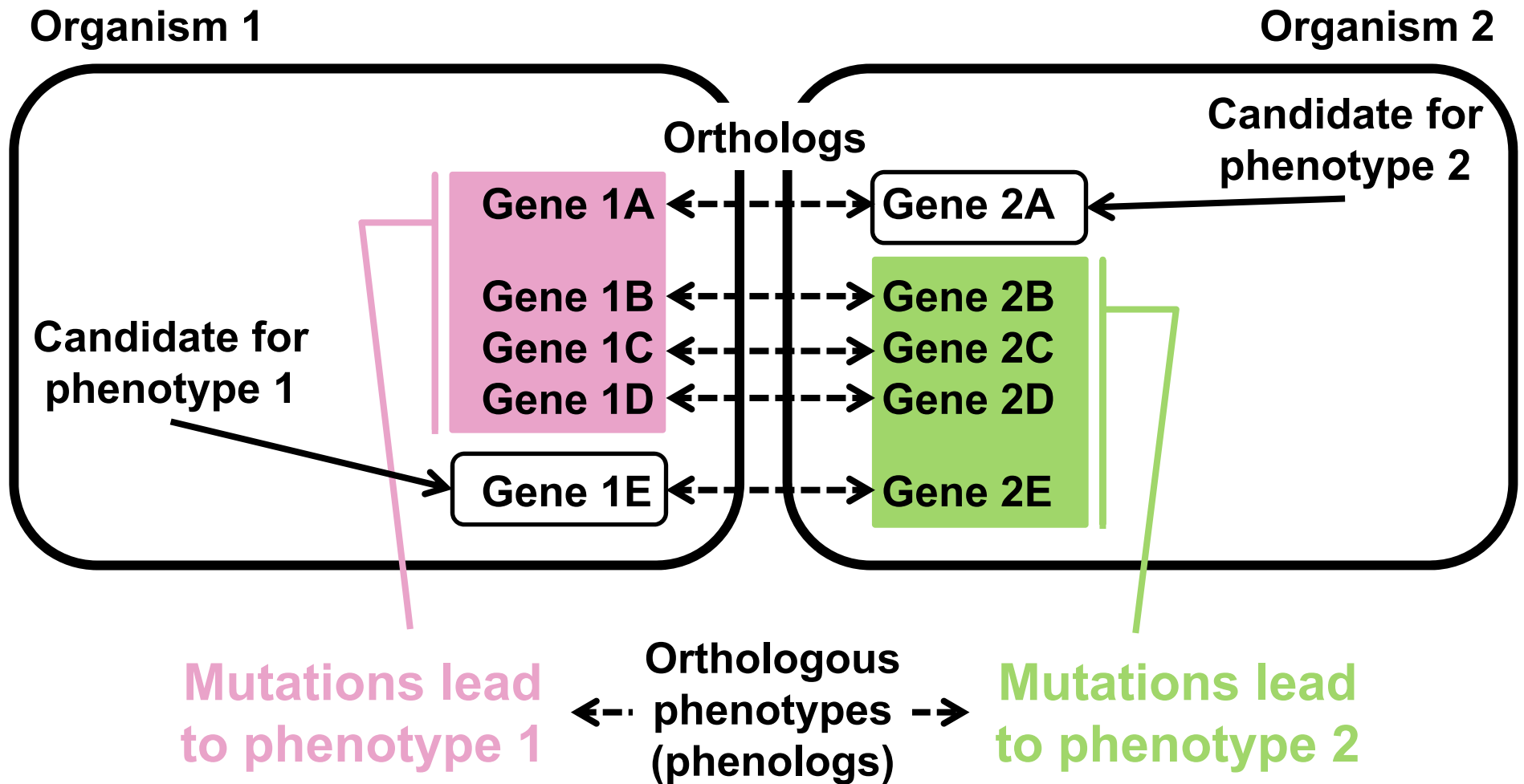
Swalla (2006) Heredity

Genomes Can Help Reveal the History of Life



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

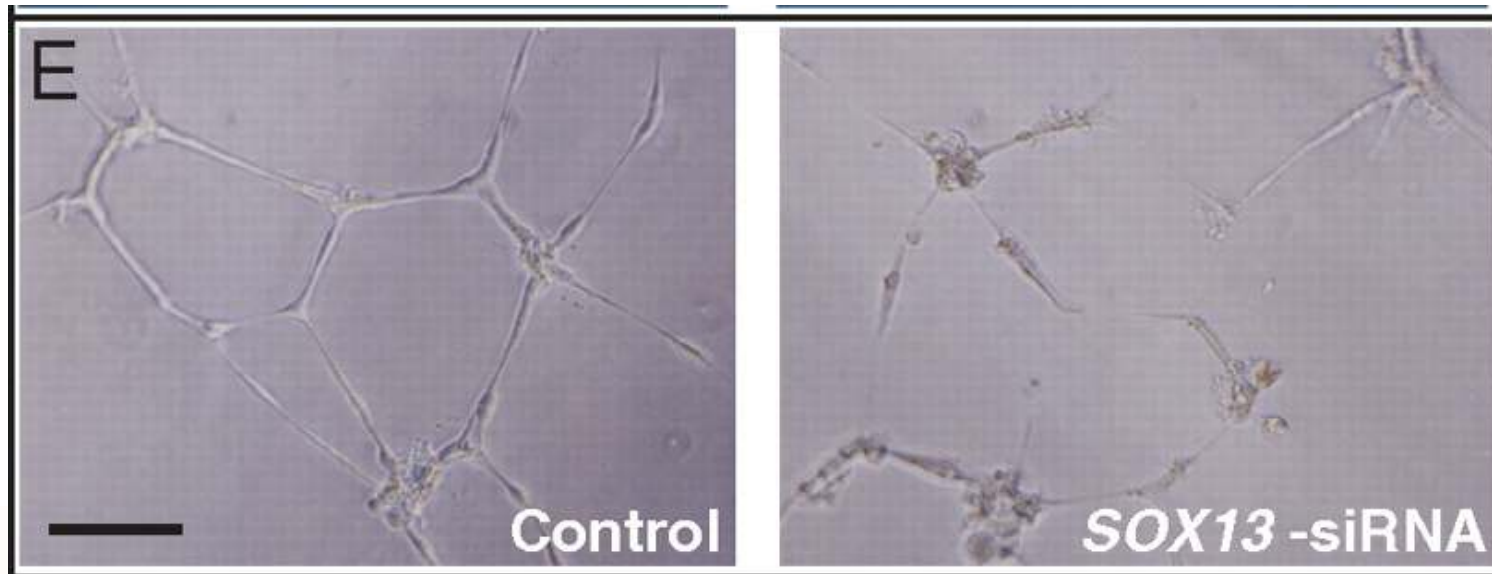
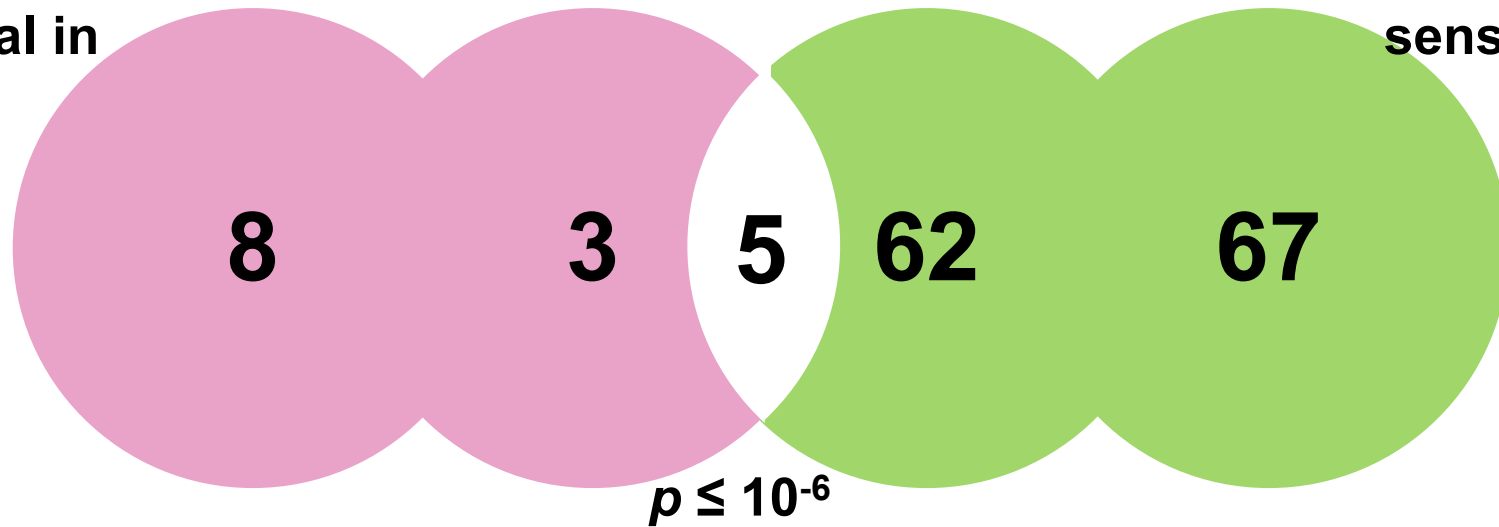
Evolution-Informed Analyses Have Great Predictive Power



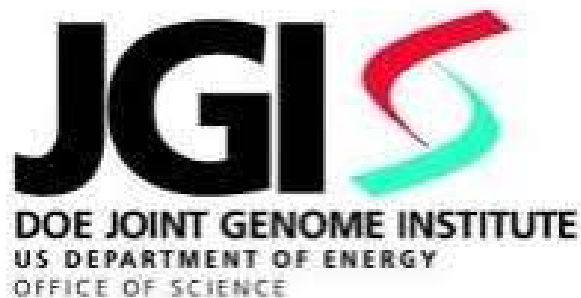
A Yeast Model for Angiogenesis

Angiogenesis
abnormal in
mice

Lovastatin
sensitive in
yeast



Genomics: “Big Science” Driven by a Few Centers



High-Throughput DNA Sequencing Technologies

454 / Roche

450 bp 1.5 Gbp / day



Illumina

150 bp 35 Gbp / day



Helicos

55 bp 4.5 Gb / day



SOLiD ABI

75 bp 22 Gbp / day



PacBio

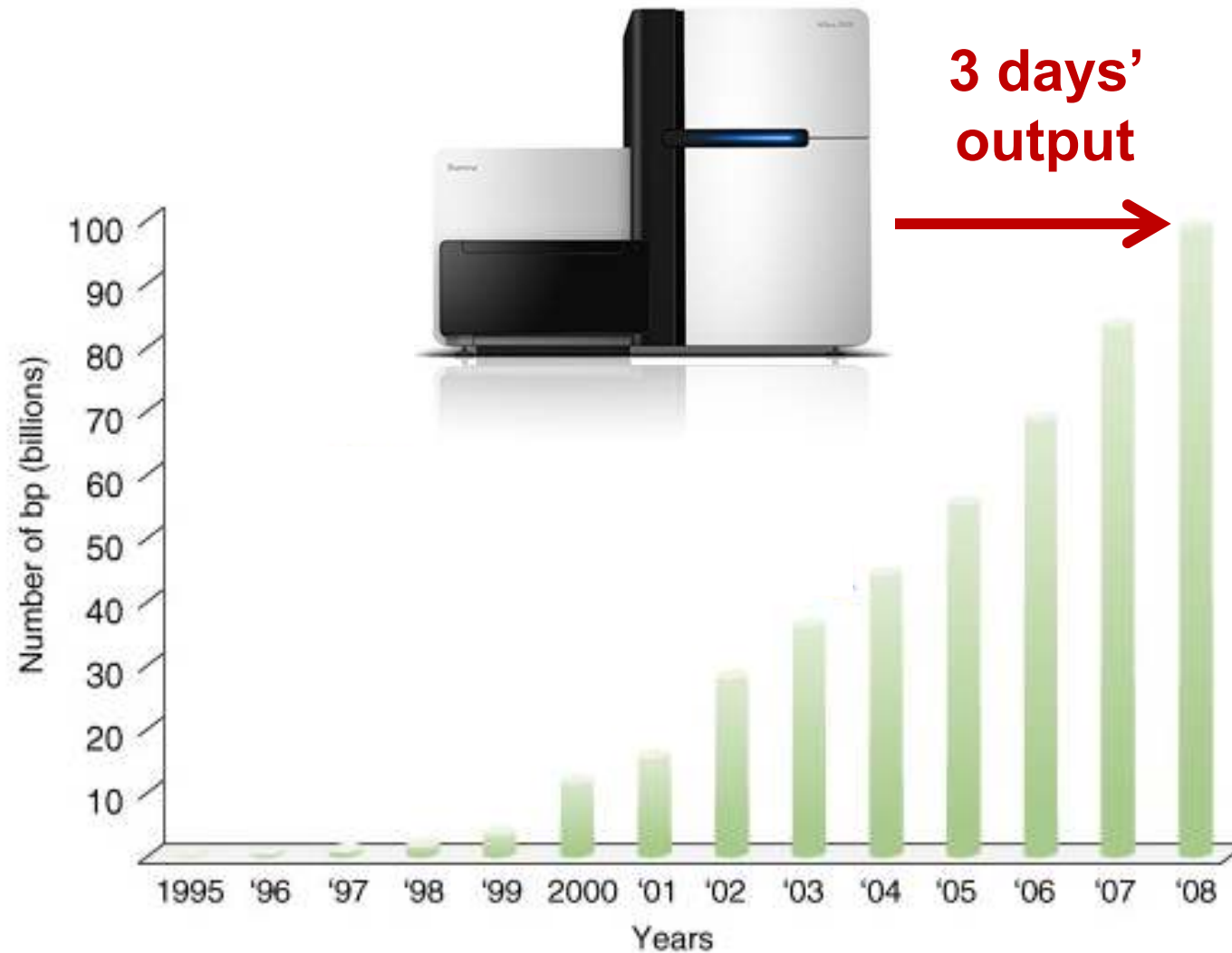
1000 bp 70 Gbp / day



Ion PGM

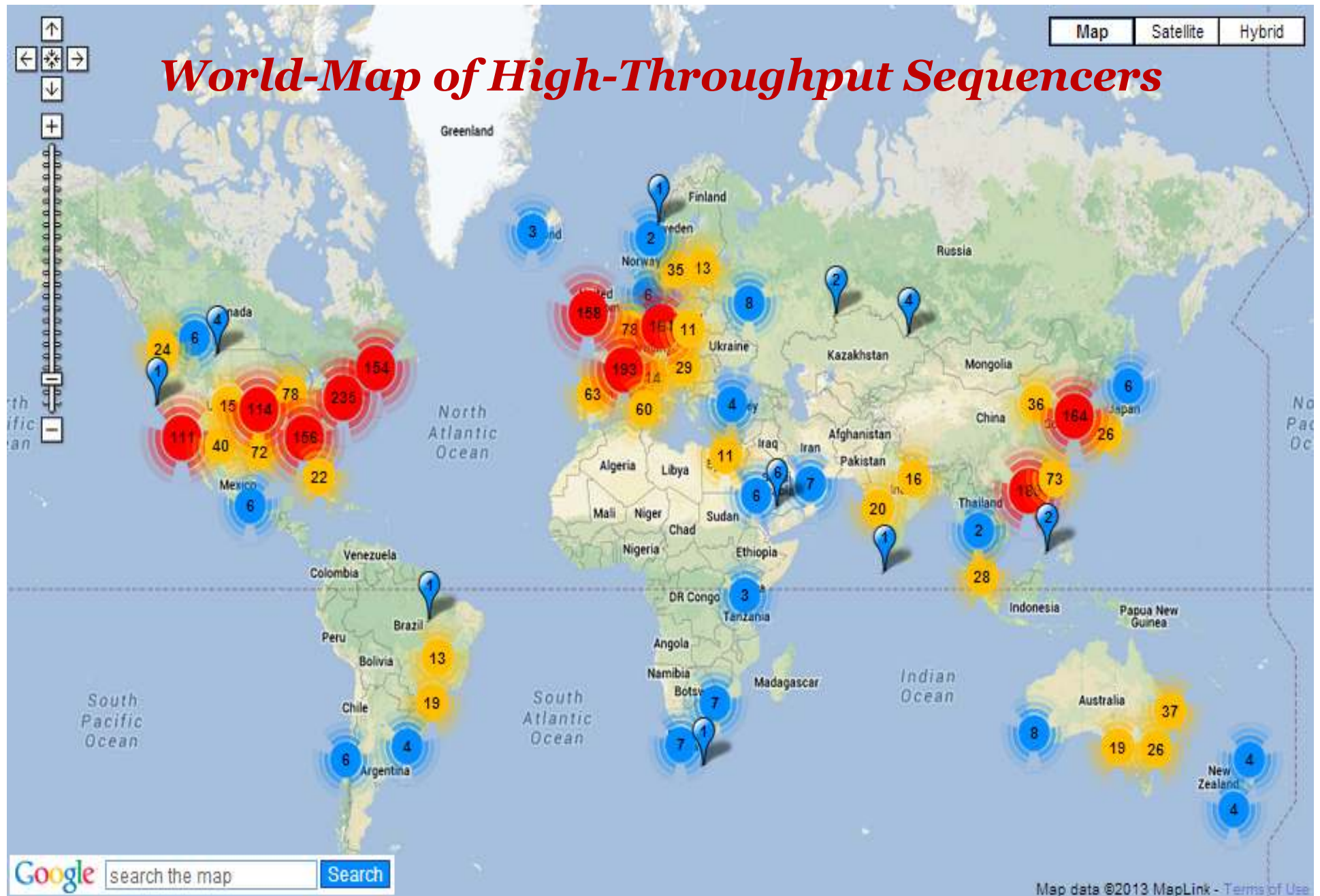
100 bp 120 Gbp / day

Why Is High-Throughput DNA Sequencing So Exciting?



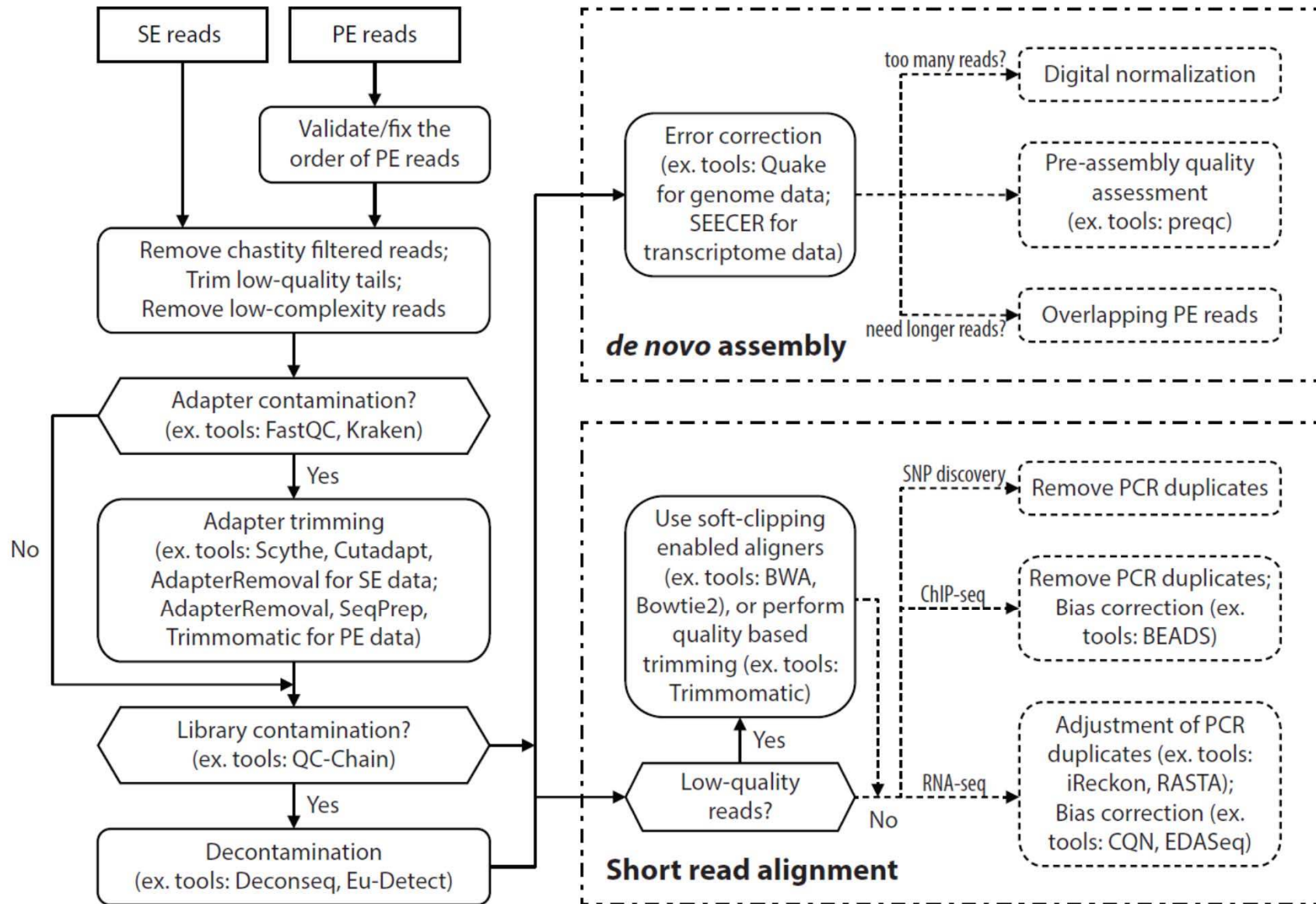
Gilad et al. (2009) Trends Genet.

World-Map of High-Throughput Sequencers

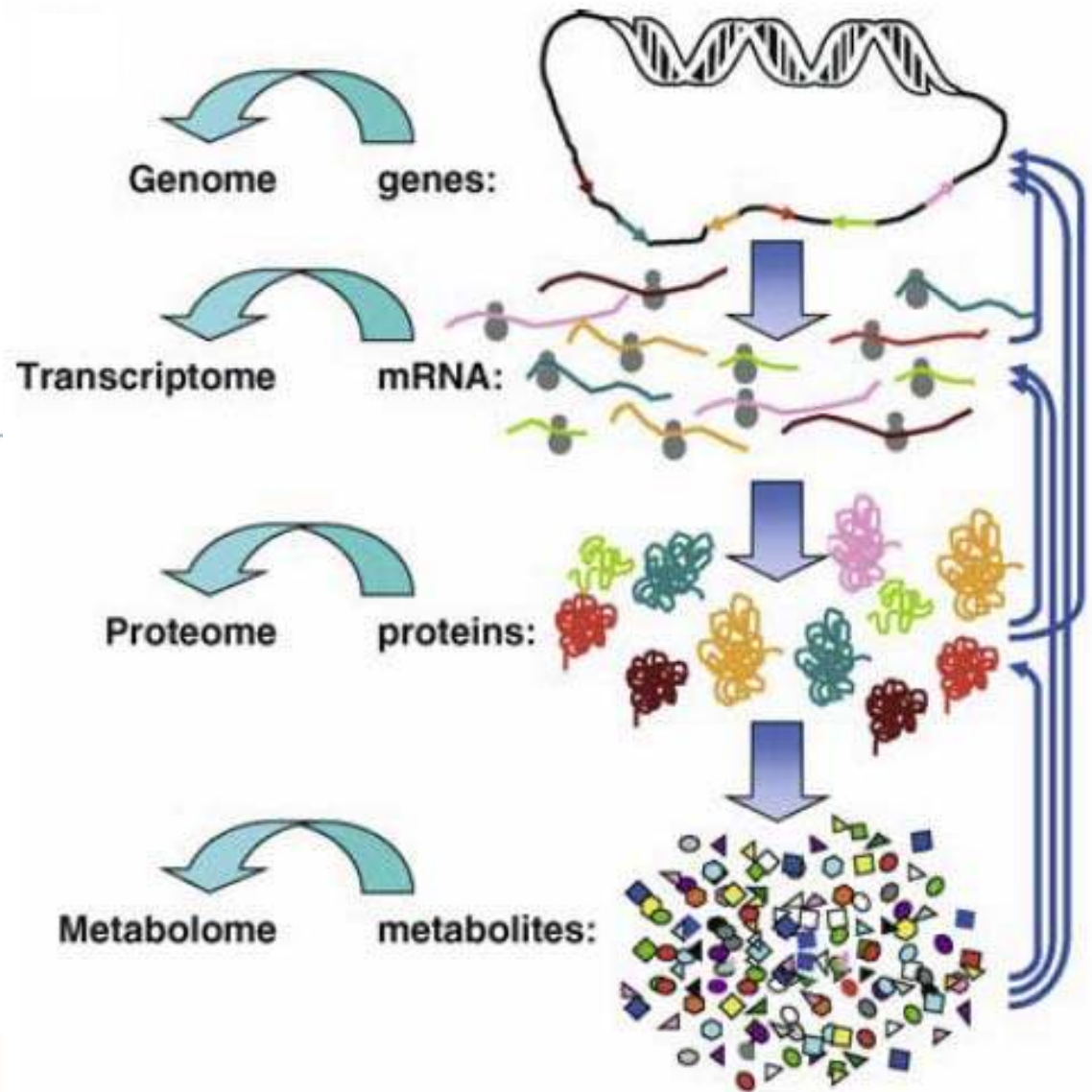


<http://omicsmaps.com/>

High Throughput Sequencing Data Pathologies



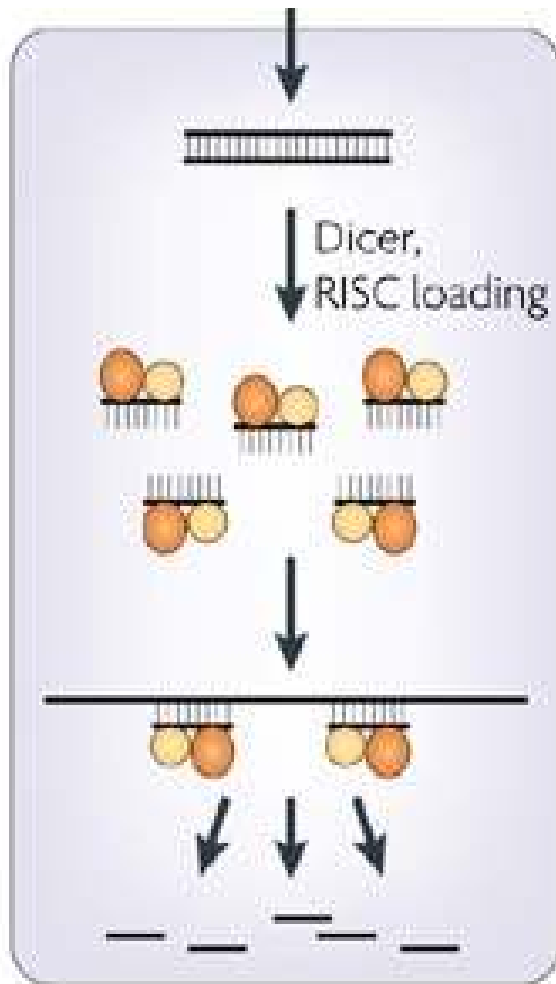
The Age of High Throughput Technologies



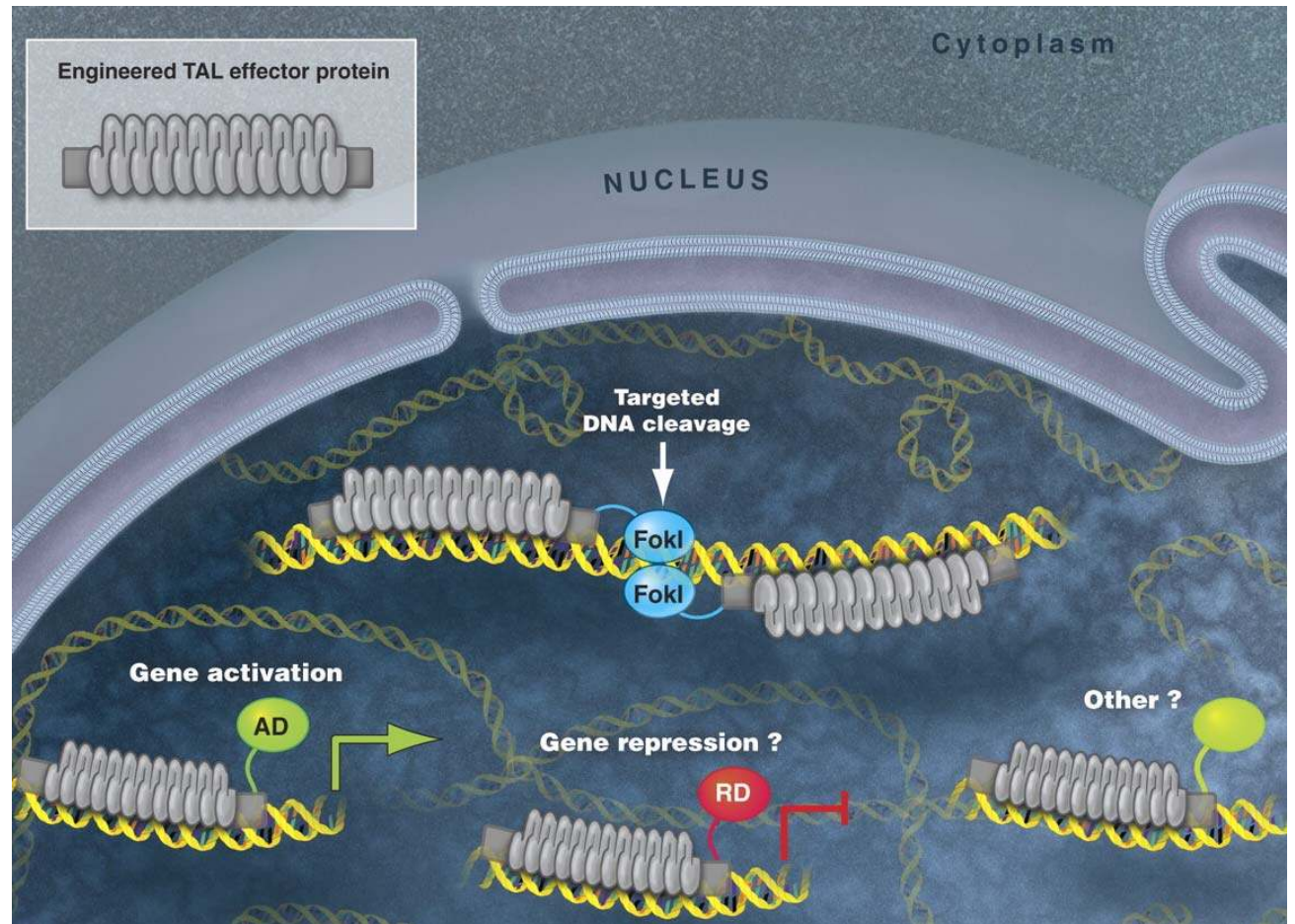
Goodacre (2005) Metabolomics

Novel Technologies for Probing Gene Function

RNAi



TAL Effectors



Boutros & Ahringer (2008) Nat. Rev. Genet.;
Bogdanove & Voytas (2011) Science

The Genomes of Non-Model Organisms are the New Frontiers



Rokas & Abbot (2009) Trends Ecol. Evol.

Lecture Outline

❖ **Introduction to Evolutionary Genomics**

❖ **Population Genomics**

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

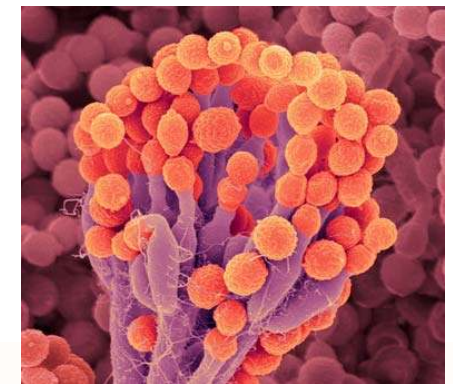
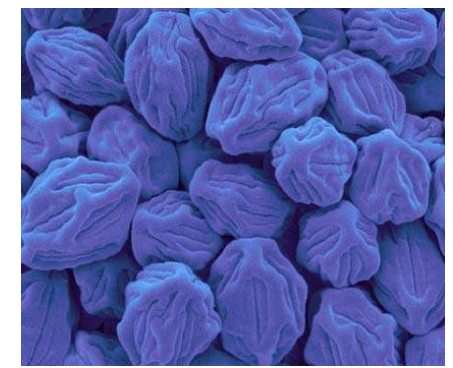
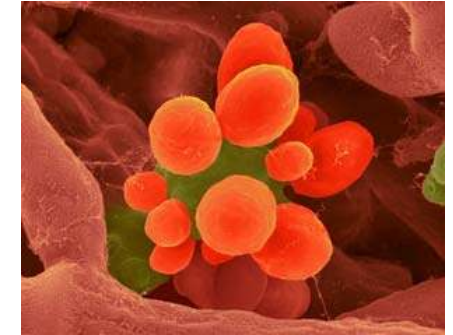
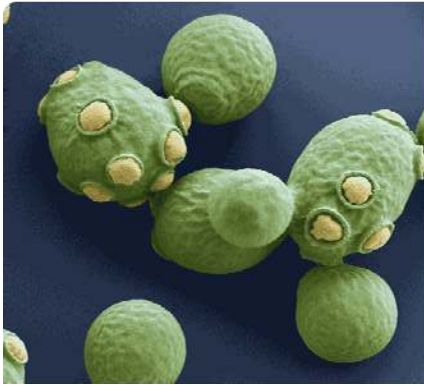
❖ **Phylogenomics**

The Rokas Lab

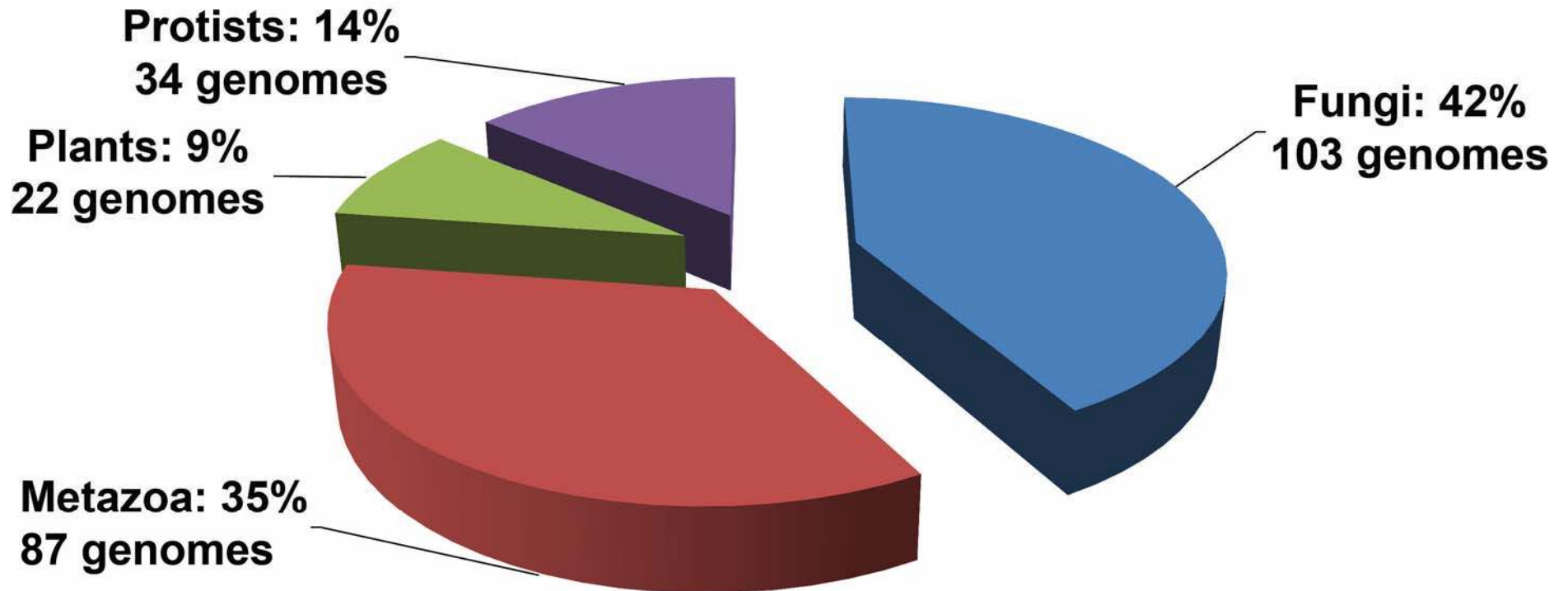


We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches

Fungi: A Model for the Study of the DNA Record



Fungi: the Most Sequenced Eukaryotic Lineage

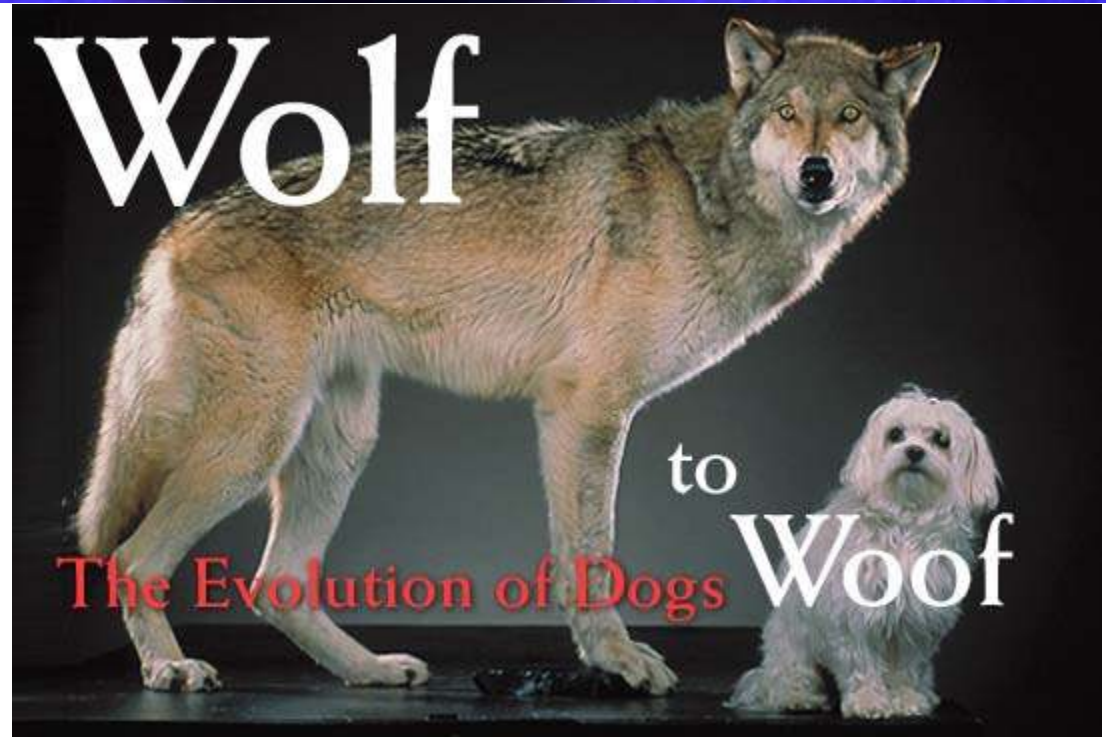


fungi: 1000genomes

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

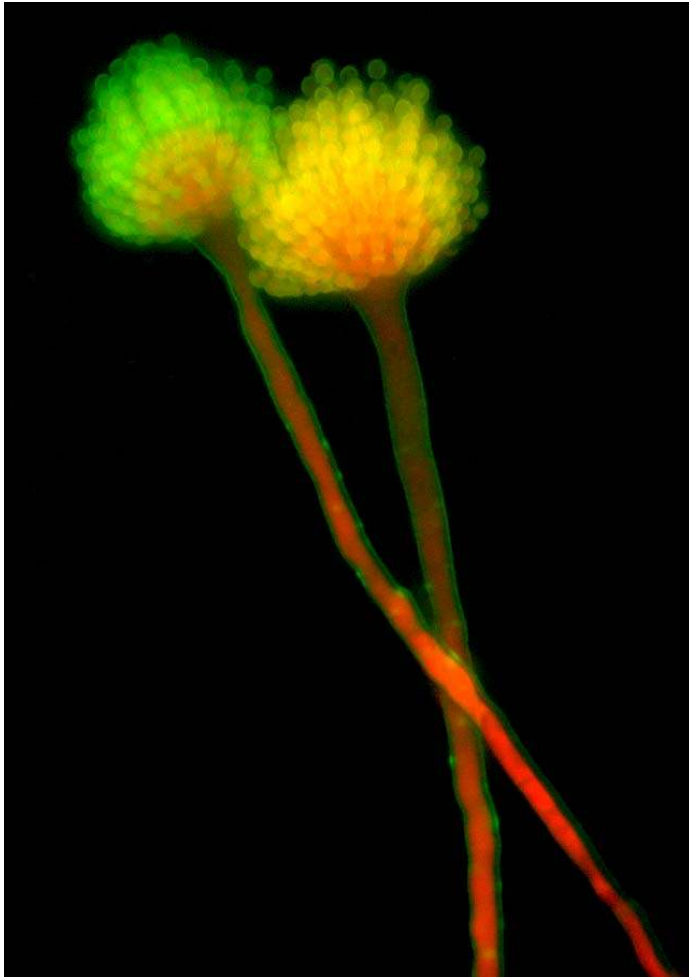


Data from GOLD 3.0 (www.genomesonline.org), March 2011





***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)**



How Sake is Made

1. Polishing



2. Washing



3. Steaming



4. Koji making



5. Yeast starter



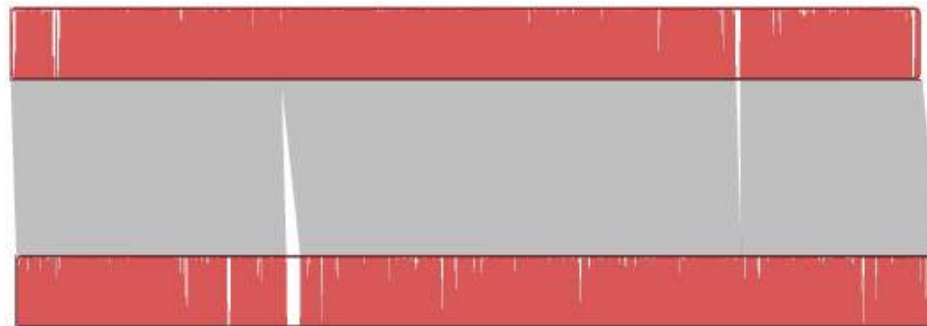
6. Pressing



http://www.sakejapan.com.au/whats_nihonshu.html#what

The *A. oryzae* and *A. flavus* Genomes are Nearly Identical

A. oryzae



A. flavus

6.5 Mb Chromosome 1

- ❖ 8 Chromosomes
- ❖ 37 Mb
- ❖ 12,000 genes
- ❖ 99.5% nt identity

A. oryzae

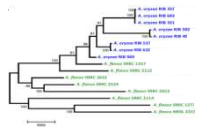
- ❖ Sake, soy sauce, miso
- ❖ Non-aflatoxin producer
- ❖ USDA GRAS species

A. flavus

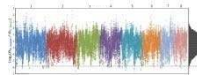
- ❖ Agricultural pest
- ❖ Aflatoxin producer
- ❖ ~\$1 billion annually



The Road to Domestication



Evolutionary Relationship?



Positively Selected Genes?



Functional Differences?

A. flavus



DOMESTICATION RD

ONE WAY

A. oryzae



Sequencing 16 Genomes

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

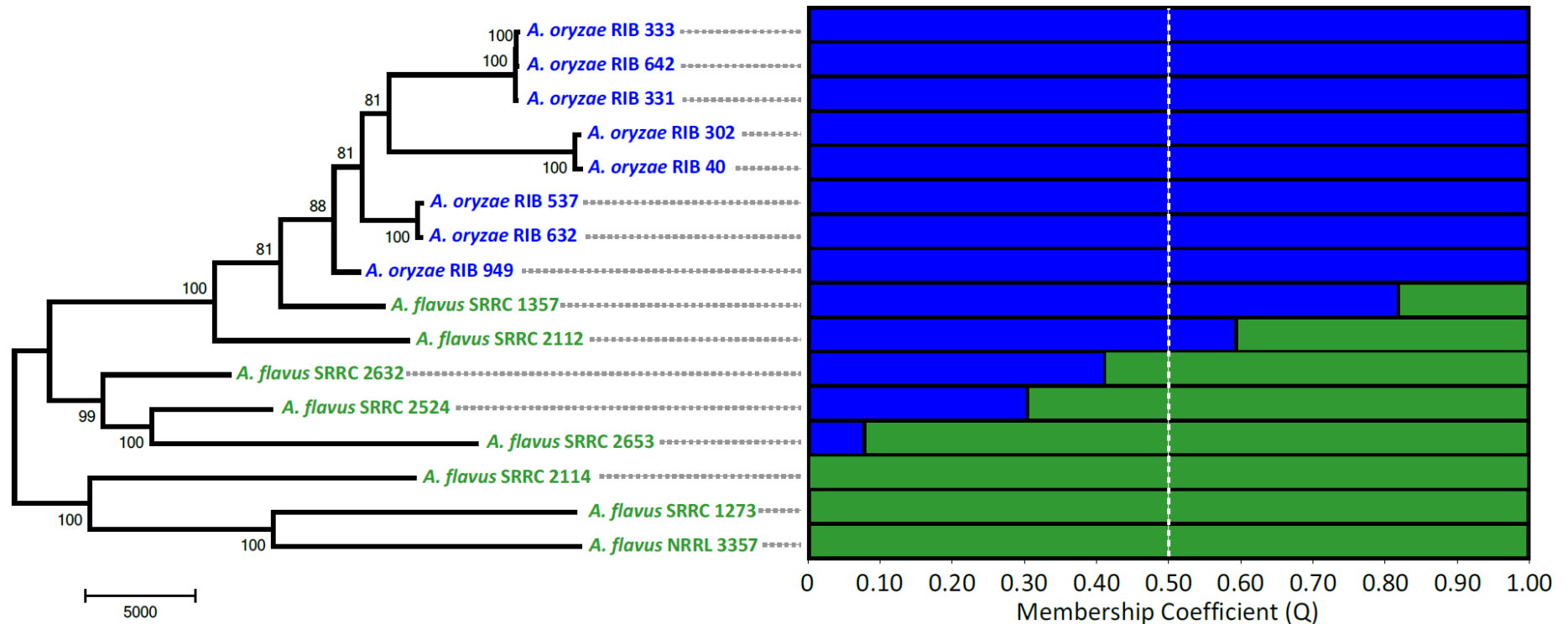
SRRC 1273	Soil, Arizona
SRRC 1357	Dried bacon, Croatia
SRRC 2112	Hazelnut, Turkey
SRRC 2114	Wheat, USA
SRRC 2524	Dead termites, China
SRRC 2632	Blood, Chicago, Illinois
SRRC 2653	Corneal ulcer, Miami, Florida
NRRL 3357	Peanut, Reference strain, USA

❖ Illumina Sequencing

❖ 12-30 million 80bp reads

❖ > 20x coverage

Aspergillus oryzae Isolates are Genetically Distinct

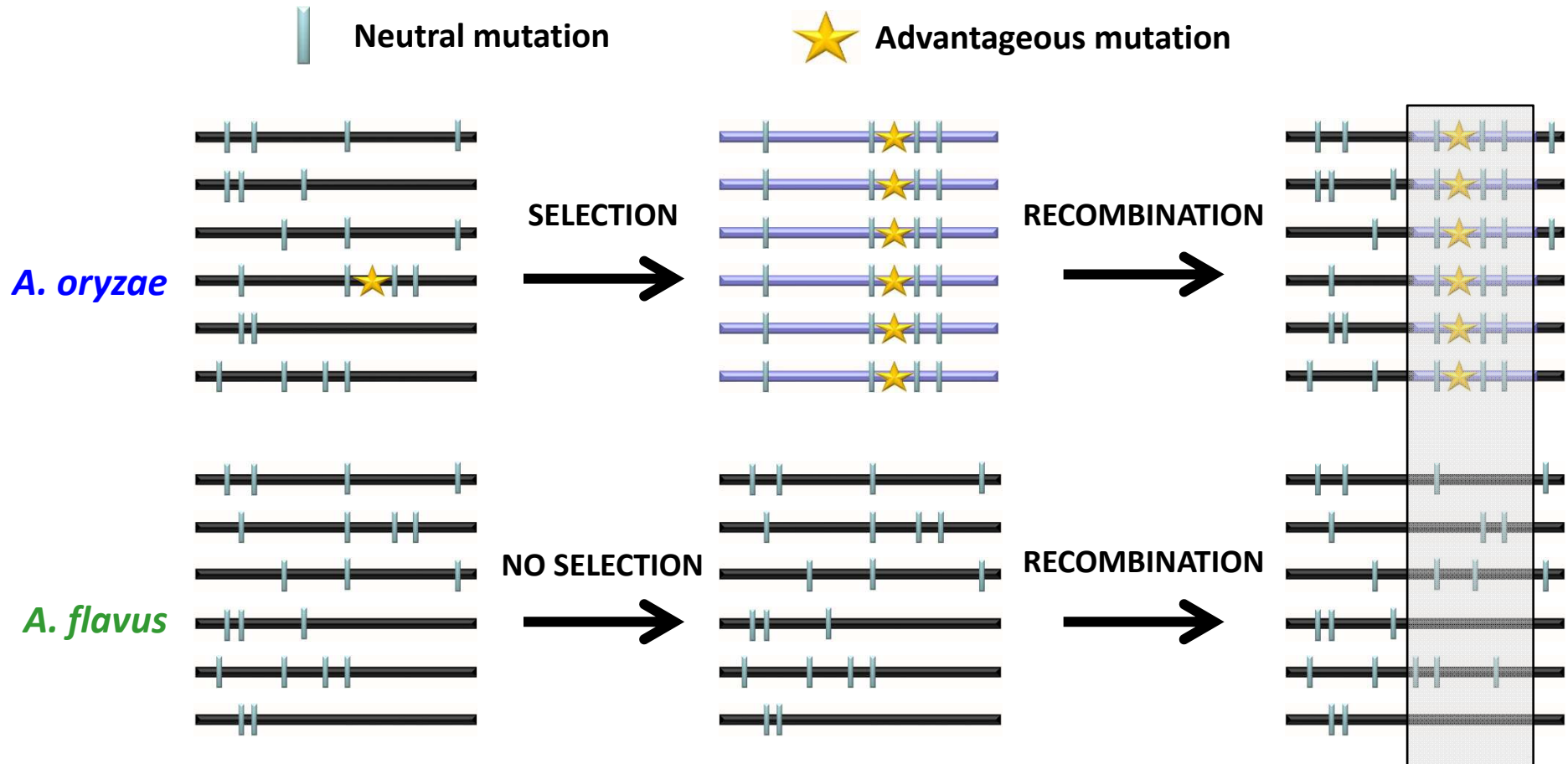


100,084 SNPs



Gibbons et al. (2012) Curr. Biol.

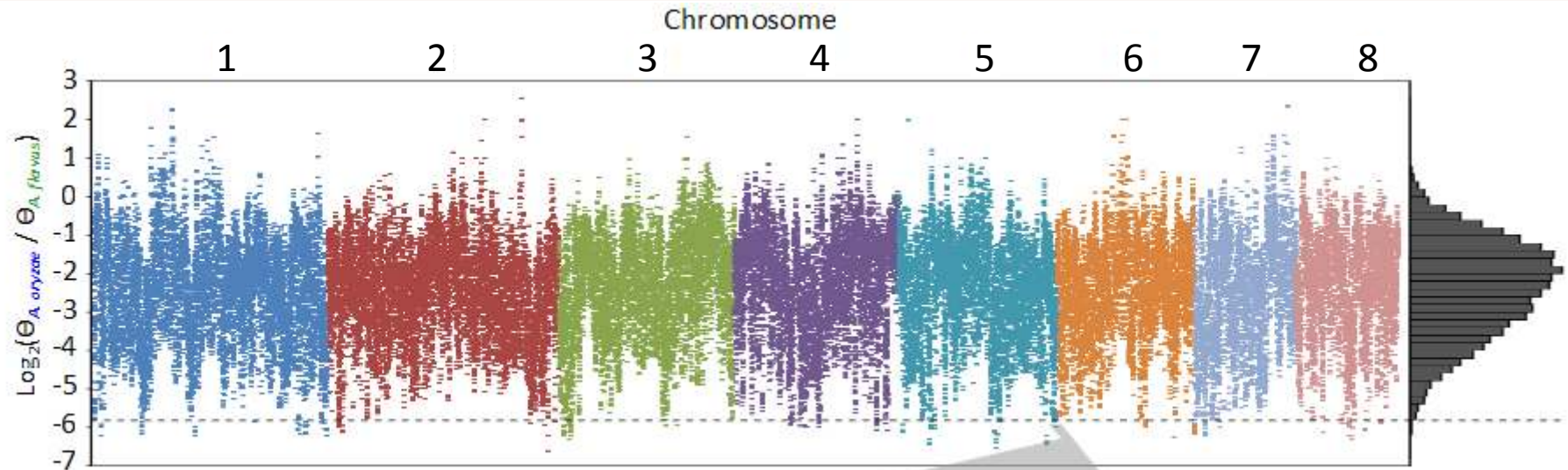
Detecting Recent Positive Selection in *A. oryzae*



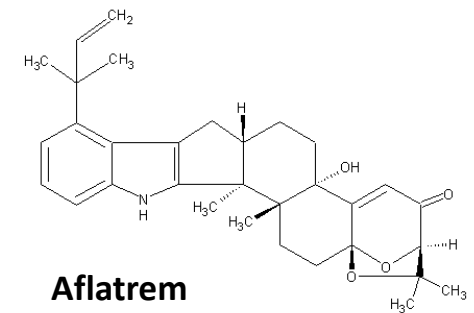
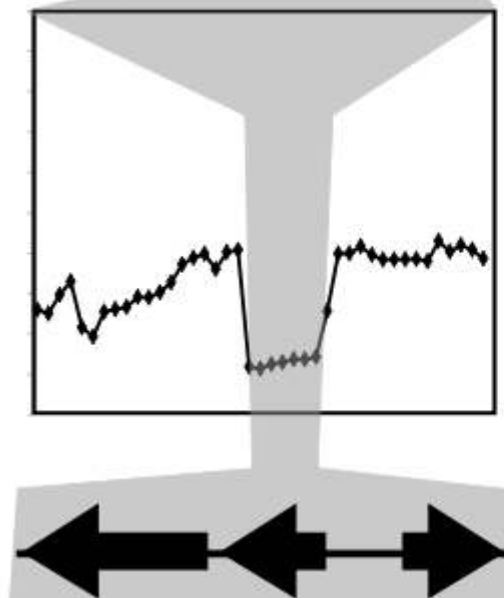
- ❖ Relative nucleotide diversity ($\Theta_{A.oryzae} / \Theta_{A.flavus}$)
- ❖ 5kb sliding windows, 500bp steps; >65,000 windows
- ❖ Overlapping regions of windows in the lowest 0.25%



Recent Positive Selection in the *A. oryzae* Genome



65,894 total windows
↓
164 candidate windows
↓
61 Putative Selective
Sweep Regions (PSSRs)
↓
148 Genes



**Genes in Sweep Regions
are Enriched in Secondary
Metabolism ($P < 0.0006$)**



The Complex Flavor of Sake

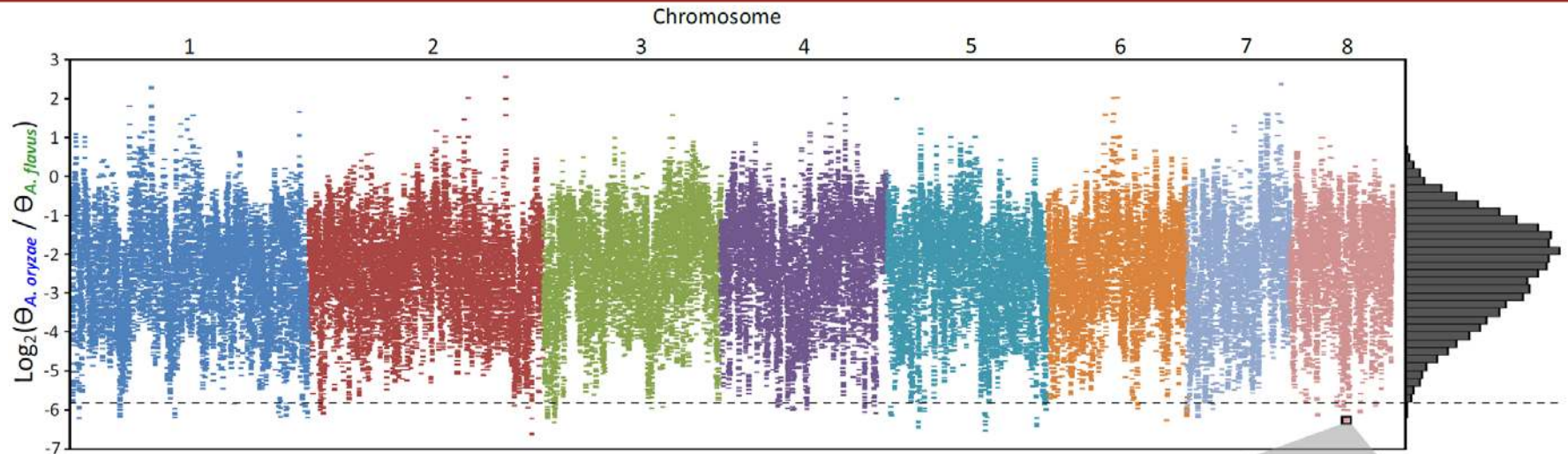


- ❖ Fragrance
- ❖ Impact
- ❖ Sweet/Dry
- ❖ Acidity
- ❖ Presence
- ❖ Complexity
- ❖ Earthiness
- ❖ Tail

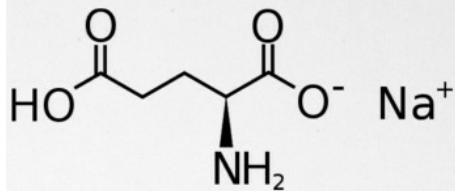
Have “flavor” genes been under selection?



Differences in a Flavor Associated Gene



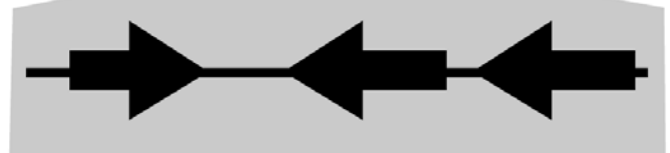
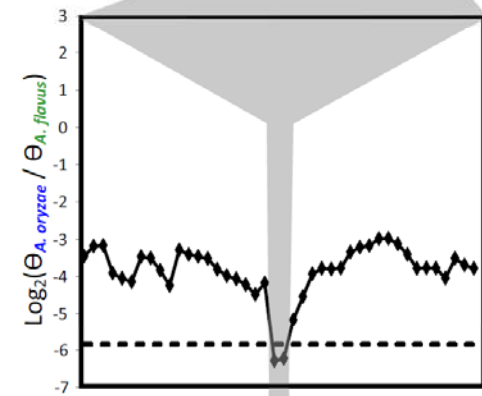
MONOSODIUM GLUTAMATE



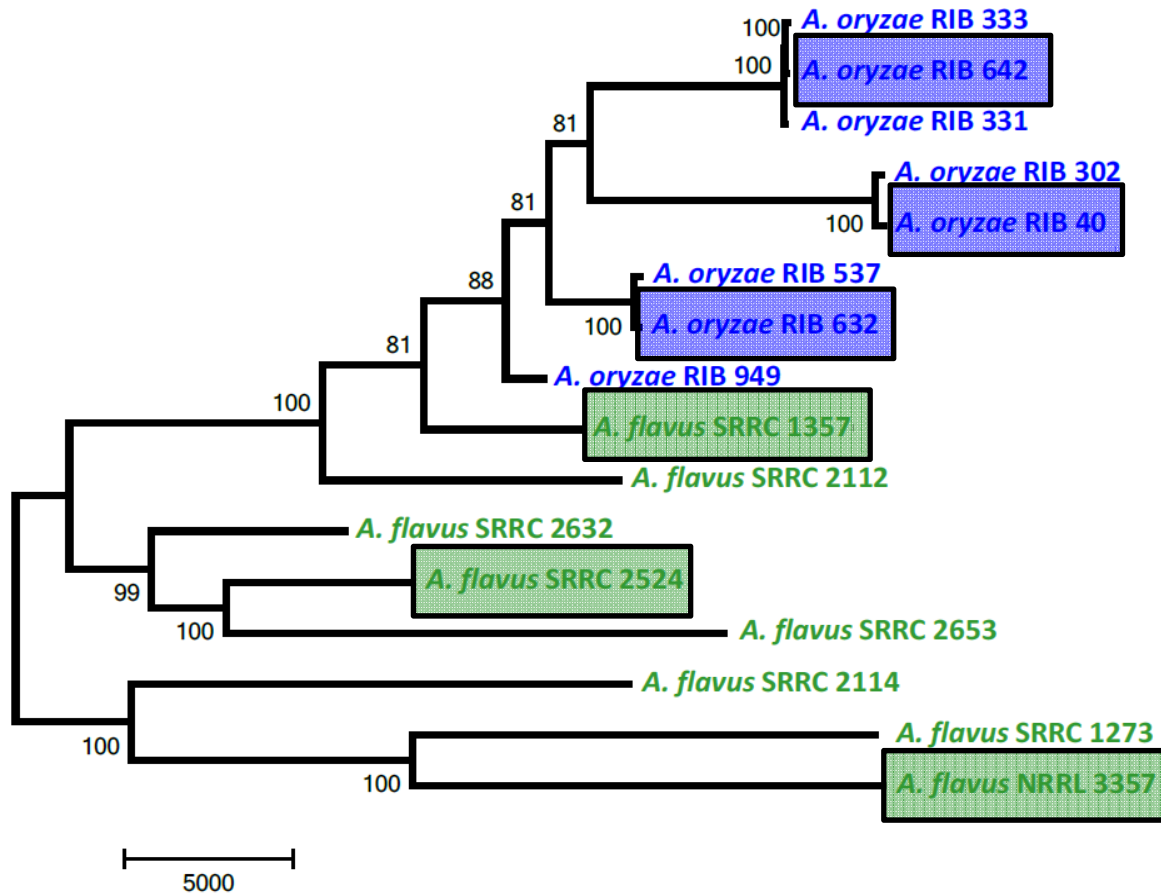
SODIUM SALT OF GLUTAMIC ACID



Glutamic Acid



Detecting Differences in Transcript and Protein Abundance

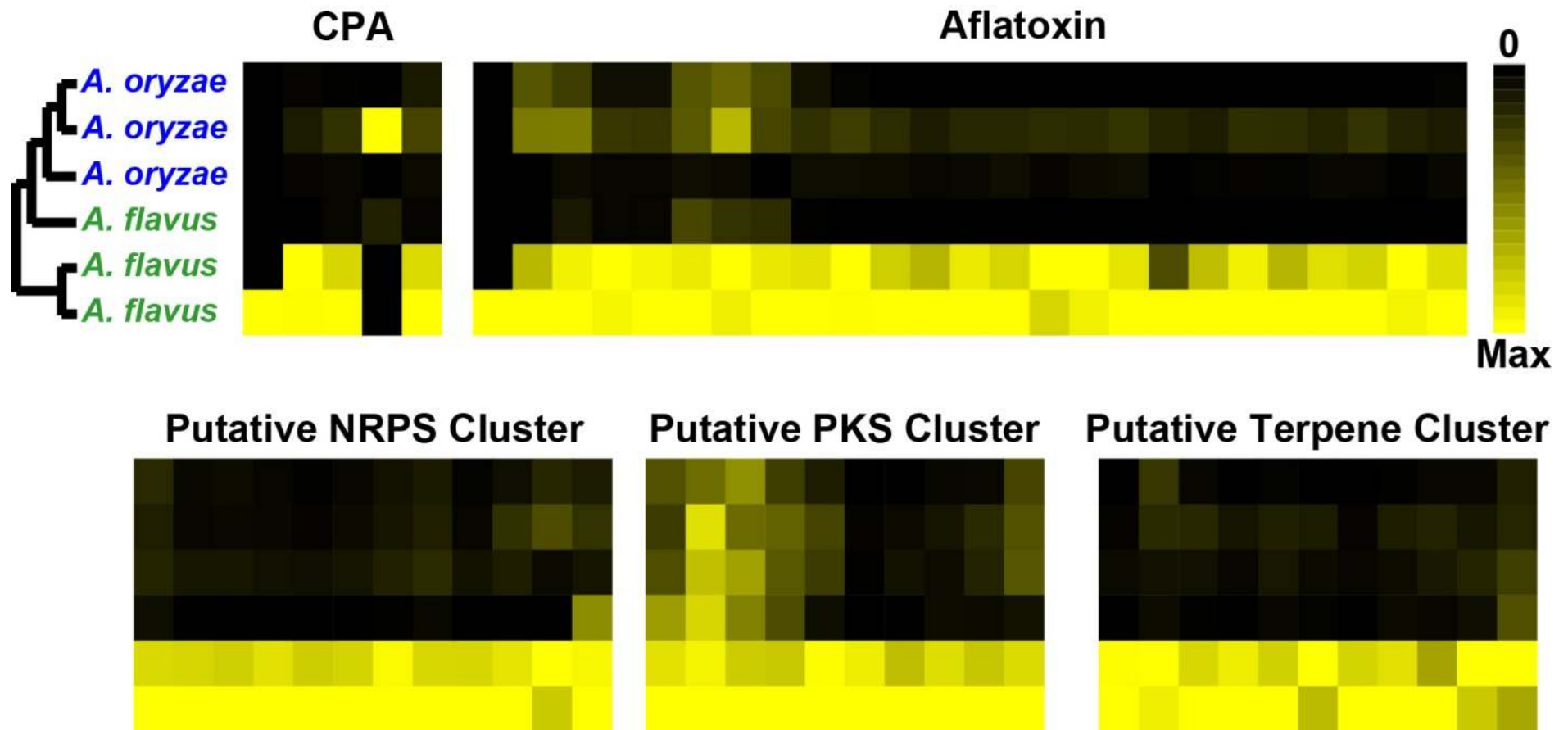


- ❖ Gene Expression (RNA-Seq)
- ❖ Protein Abundance (MudPIT)

- ❖ Sake strains
- ❖ 30° C
- ❖ 24 hours

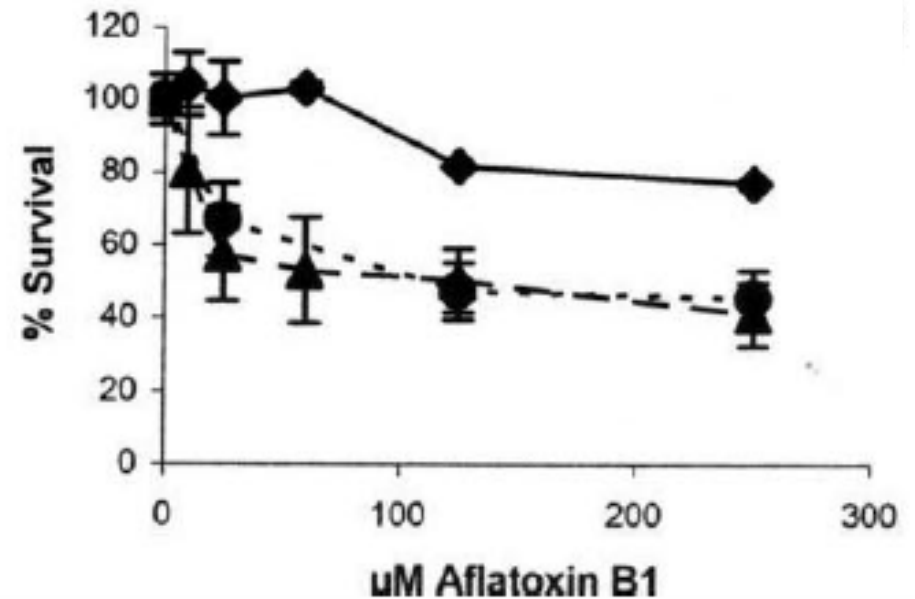
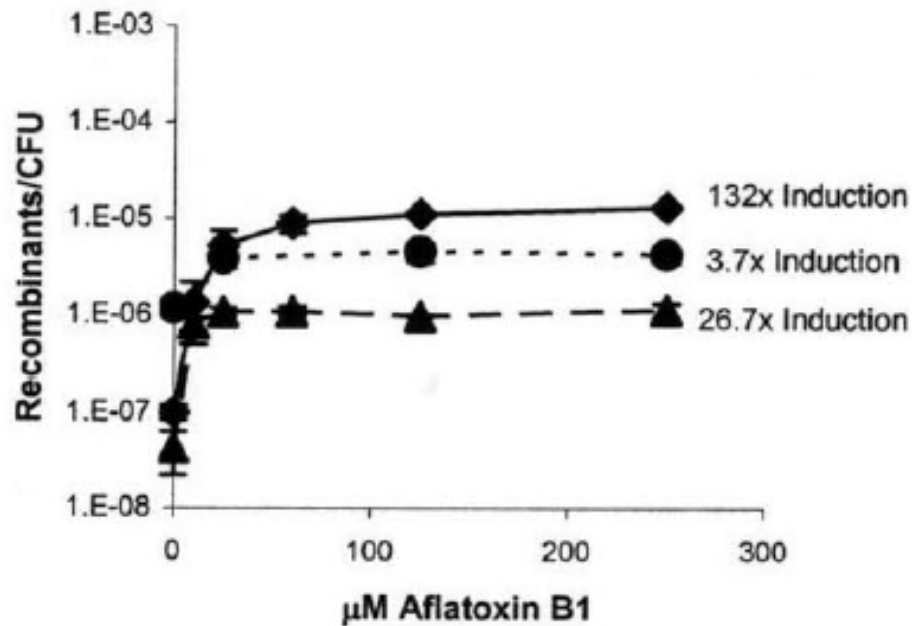


Down-Regulation of Secondary Metabolism in *A. oryzae*



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



A. oryzae Breaks Down Starch into Sugar

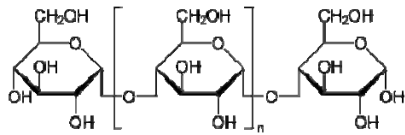
Rice Preparation



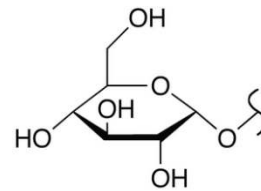
Saccharification



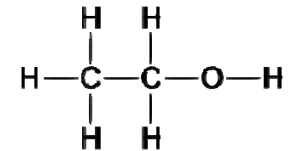
Fermentation



A. oryzae



Yeast



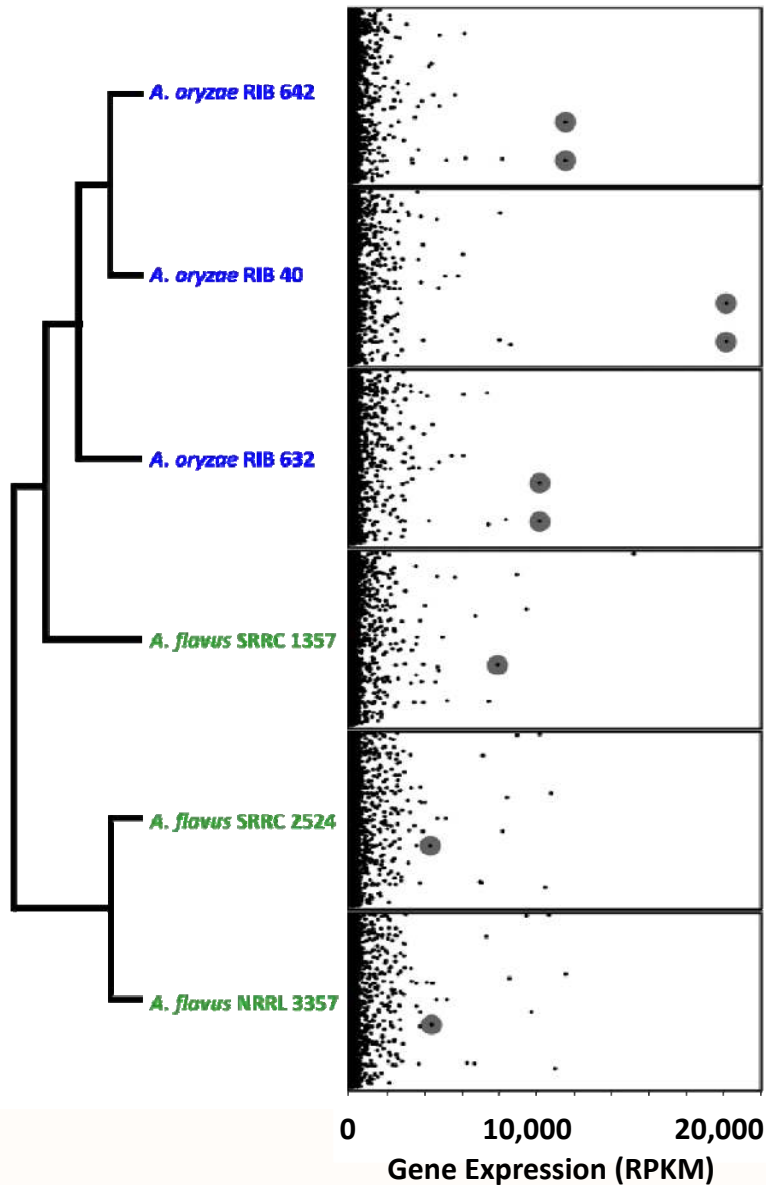
Starch

Sugar

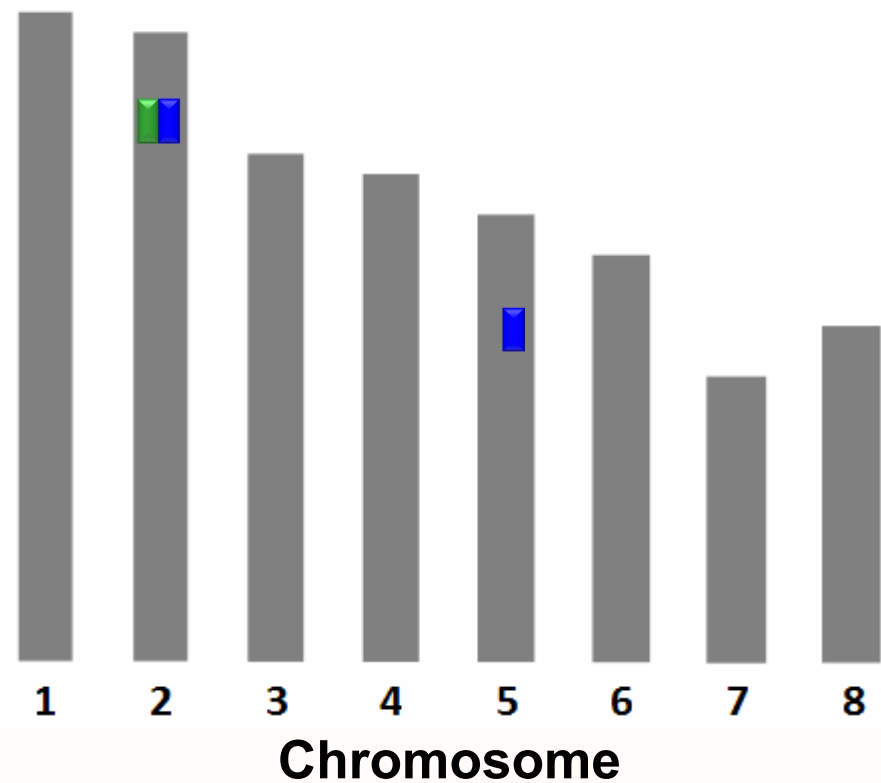
Alcohol



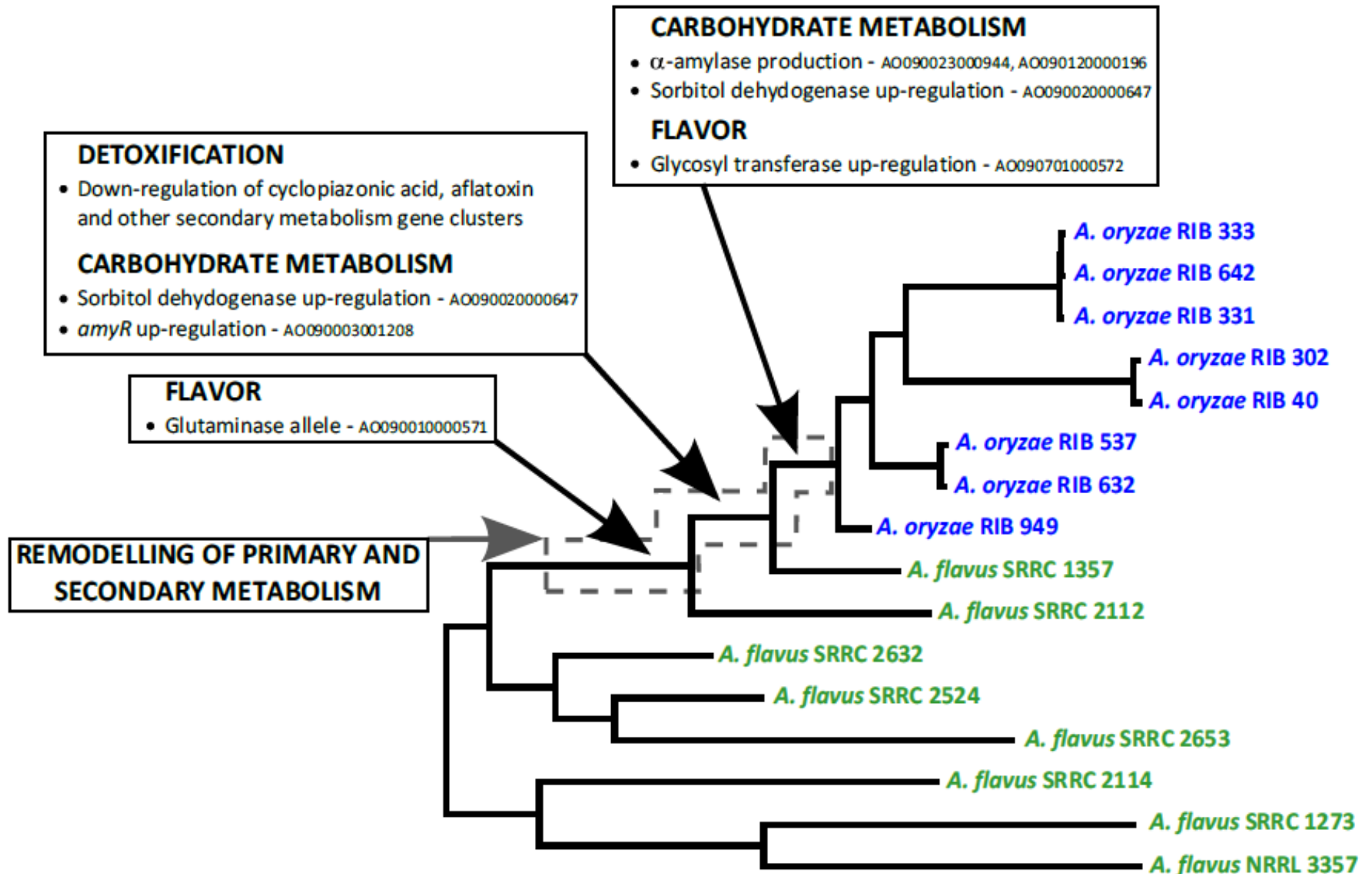
α -amylase is Highly Abundant



- ❖ Gene Expression $\uparrow P = 1e-300$
- ❖ Protein Enrichment $\uparrow P = 8e-63$
- ❖ Carbohydrate Metabolism $\uparrow P = 6e-5$



DOMESTICATION RD

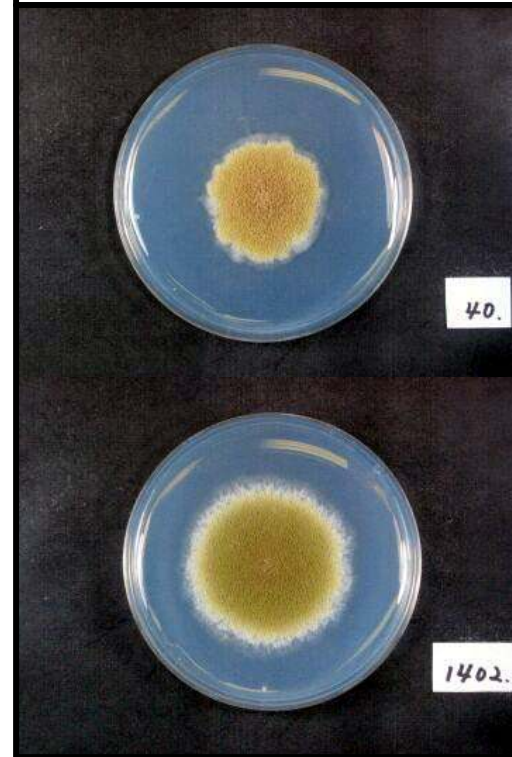


Contrasting Domestication Patterns Across Kingdoms

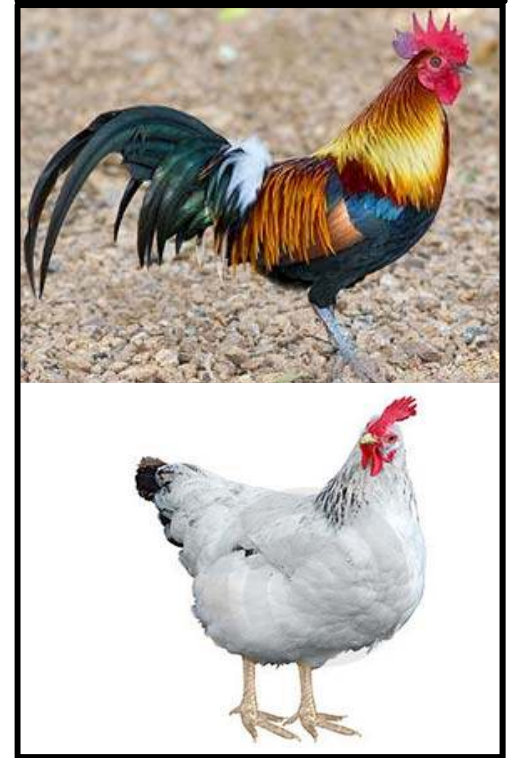
Plants



Microbes



Animals



Lecture Outline

❖ **Introduction to Evolutionary Genomics**

❖ **Population Genomics**

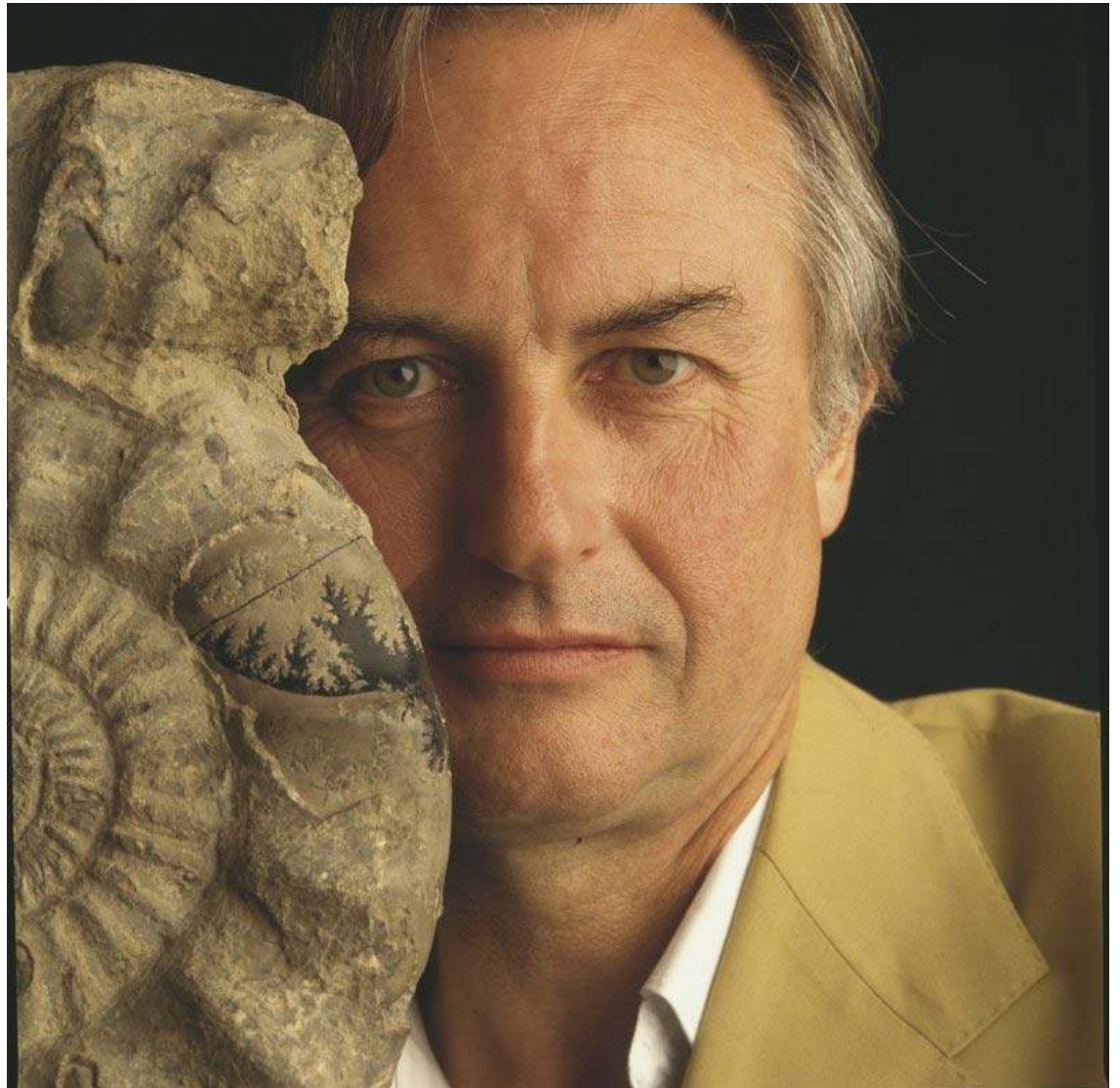
----- **Coffee Break** -----

❖ **Phylogenomics**

The Dawkins Delusion

“... there is, after all, one true tree of life [...]. It exists. It is in principle knowable. We don't know it all yet. By 2050 we should – or if we do not, we shall have been defeated only at the terminal twigs, by the sheer number of species.”

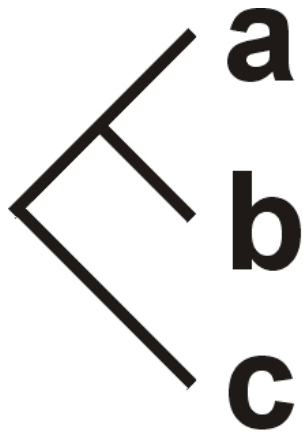
Richard Dawkins



Dawkins (2003), A Devil's Chaplain

The Problem of Incongruence

Gene X

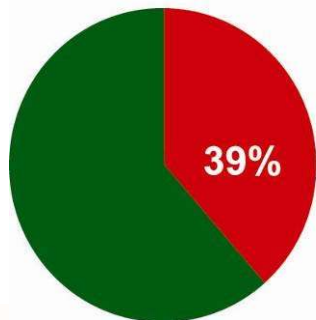


Gene Y

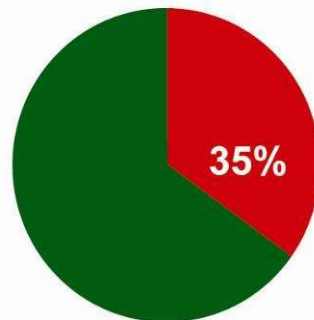


Species tree?

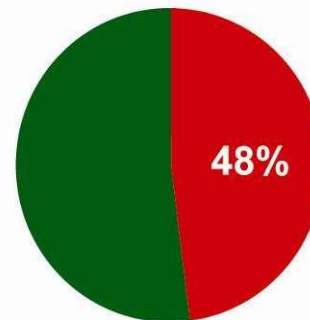
A: All organisms



B: Mammals



C: Insects



Incongruence is
pervasive in the
phylogenetics
literature



A Systematic Evaluation of Single Gene Phylogenies

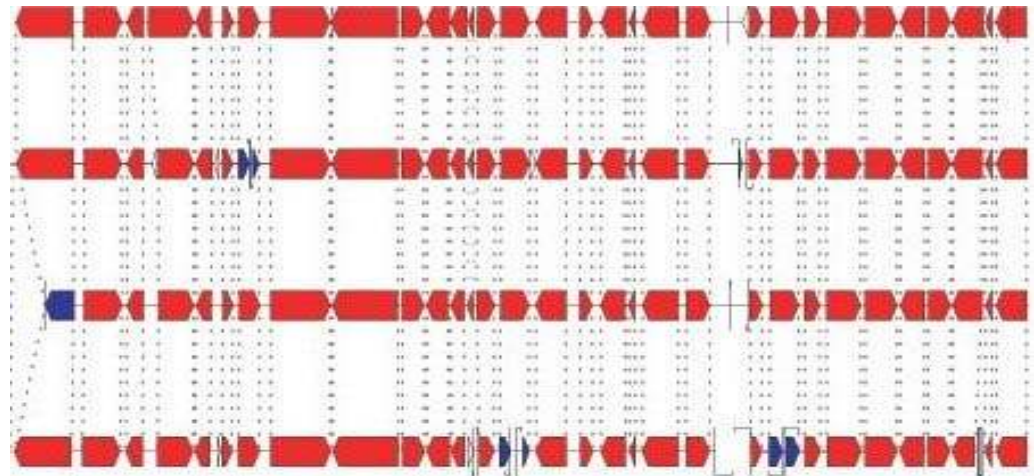


S. cerevisiae

S. paradoxus

S. mikatae

S. bayanus



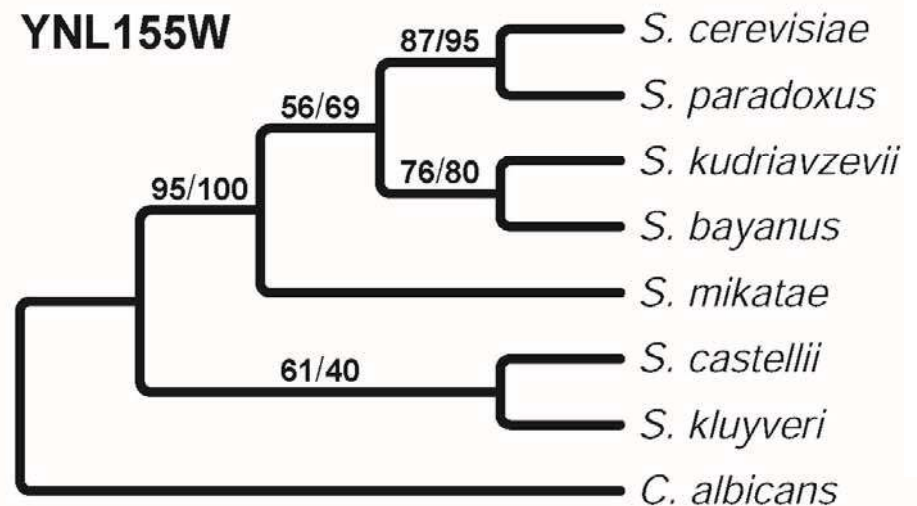
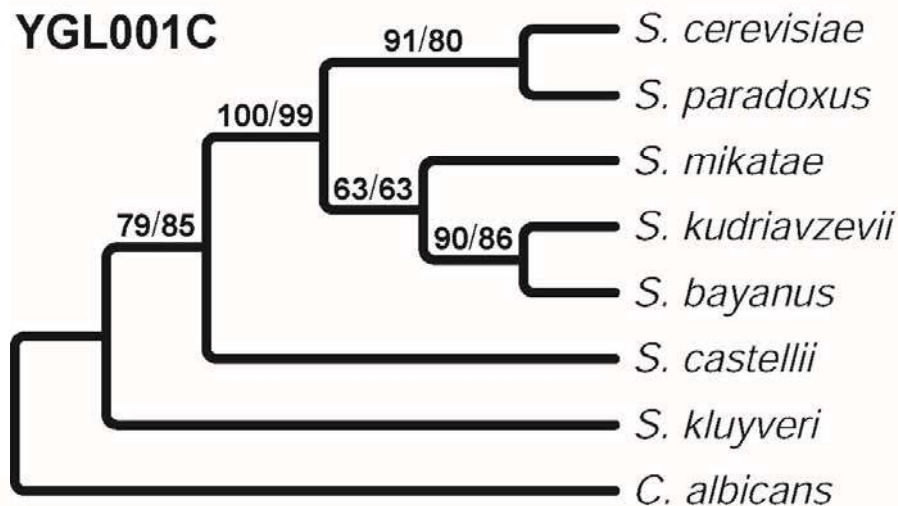
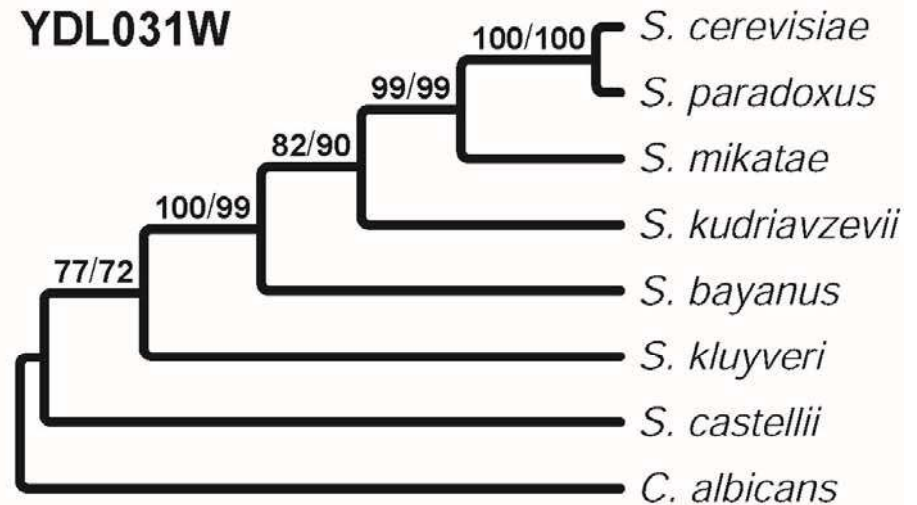
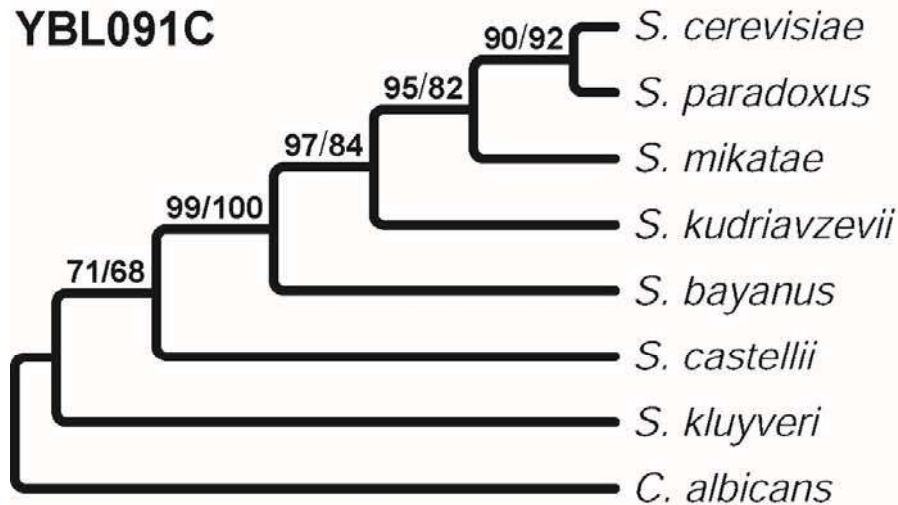
Dataset: 106 genes on all 16 chromosomes totaling 127kb corresponding roughly to 1% of the genomic sequence, 2% of genes

Analyses: Maximum Likelihood (ML) & Maximum Parsimony (MP) on nt data sets and MP on amino acid data sets

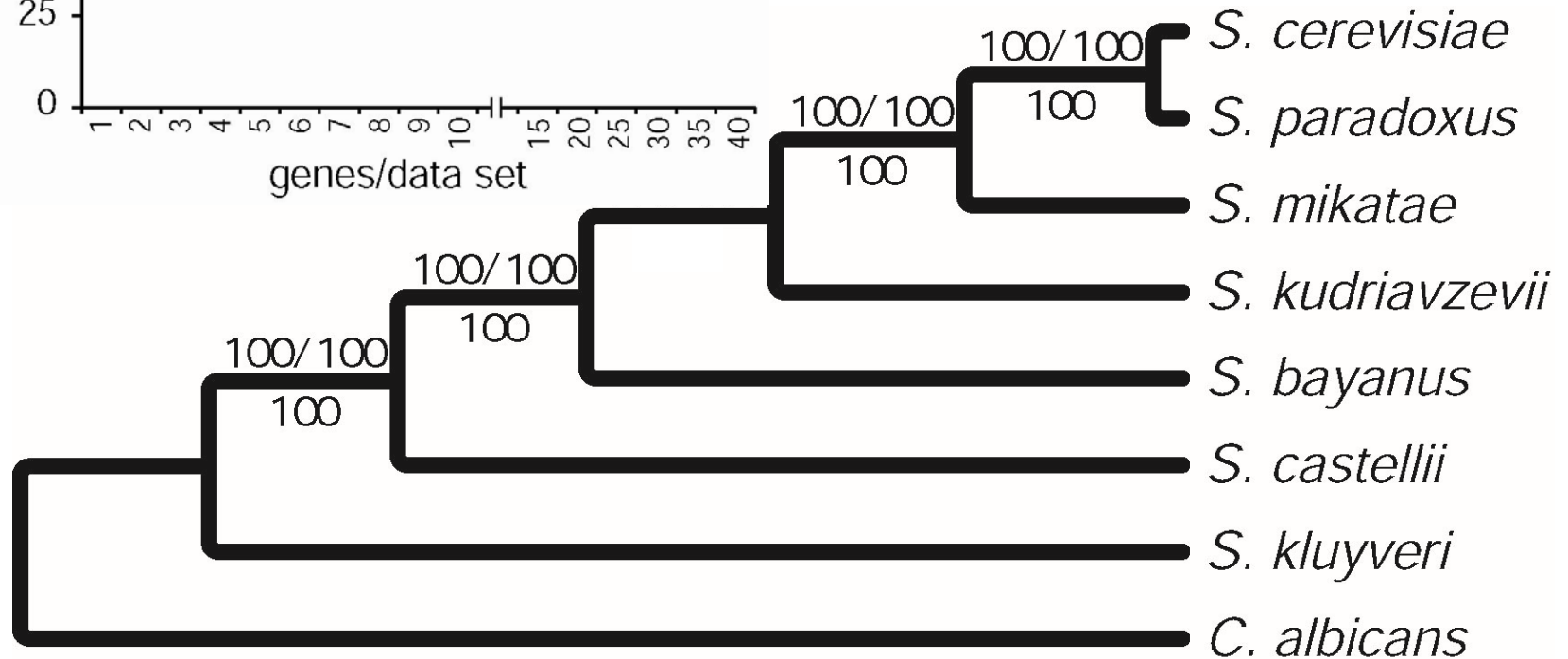
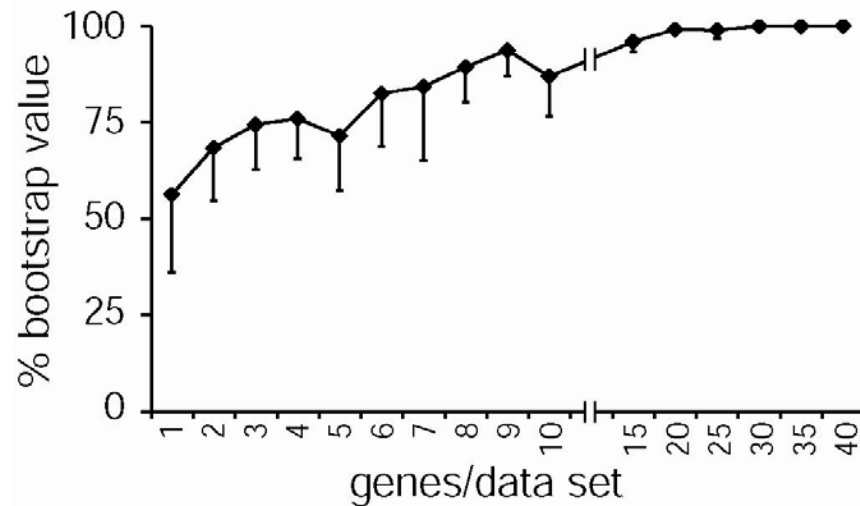


Kellis et al. (2003) Nature

Incongruence at the Single Gene Level



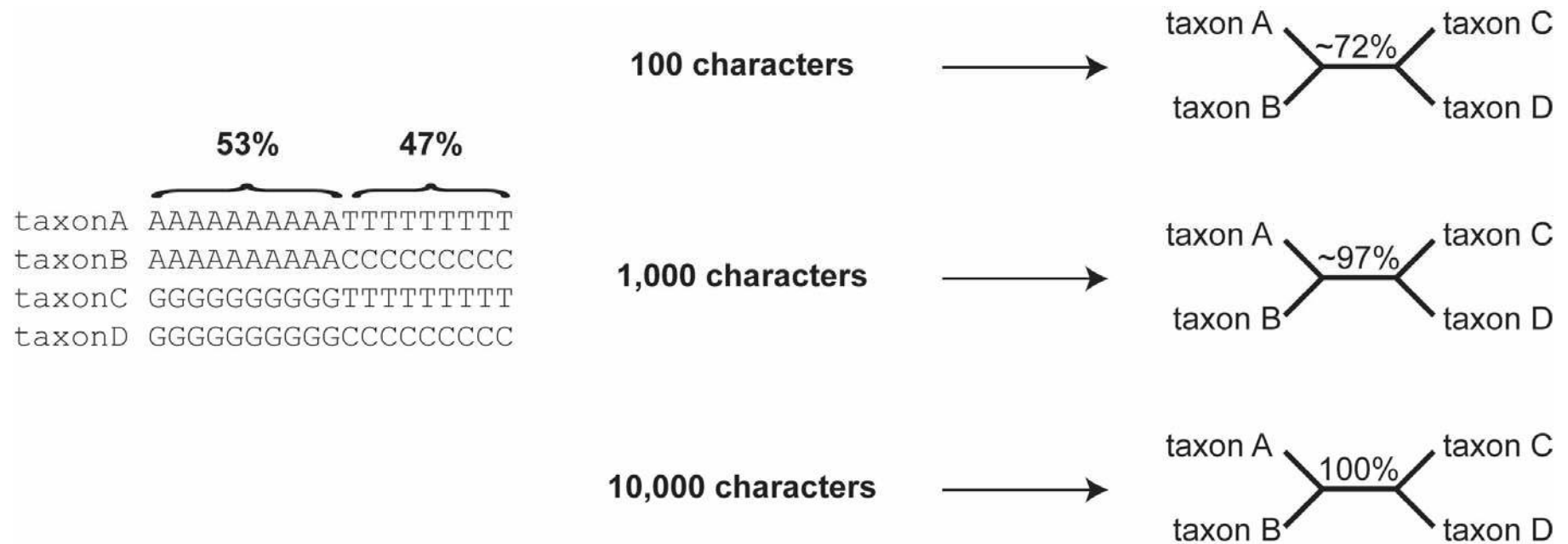
Concatenation of 106 Genes Yields a Single Yeast Phylogeny



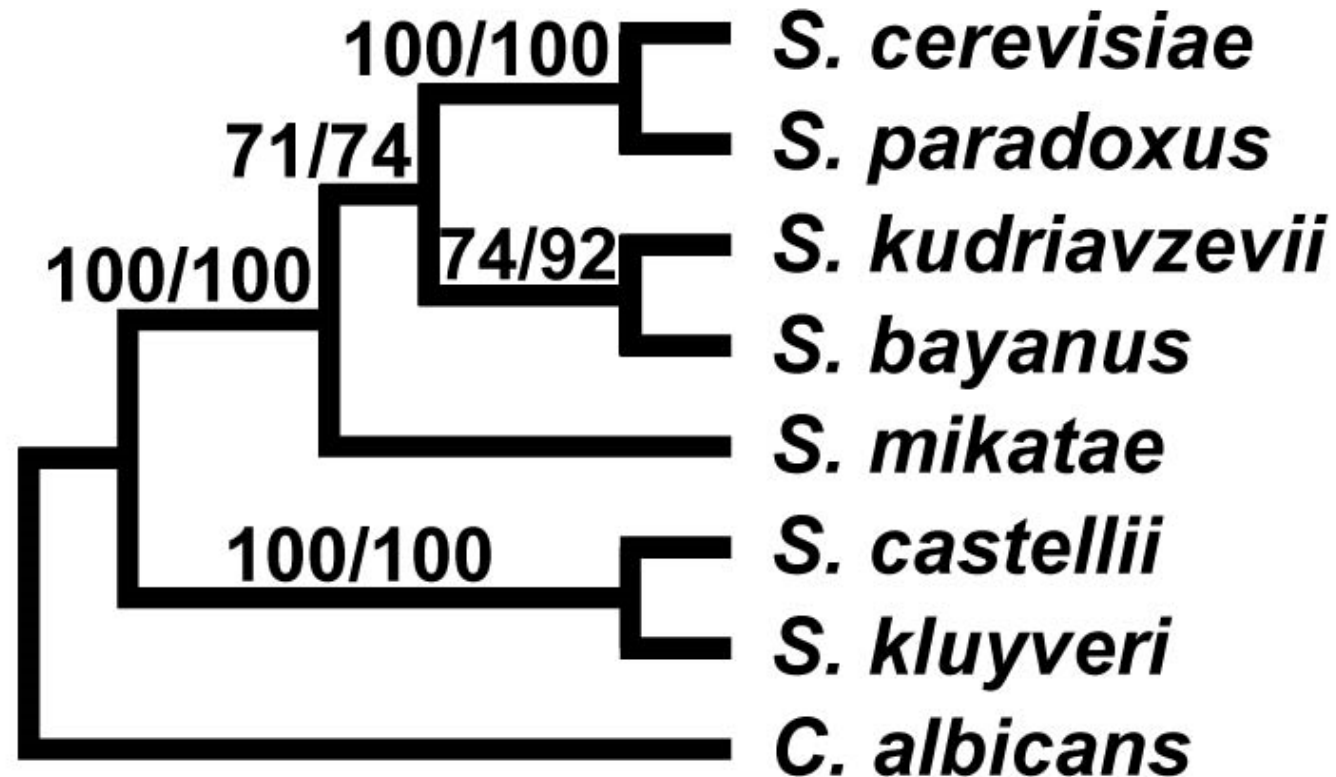
ML / MP on nt
MP on aa

Rokas et al. (2003) Nature

Bootstrap Support is Misleading When Used in Large Datasets

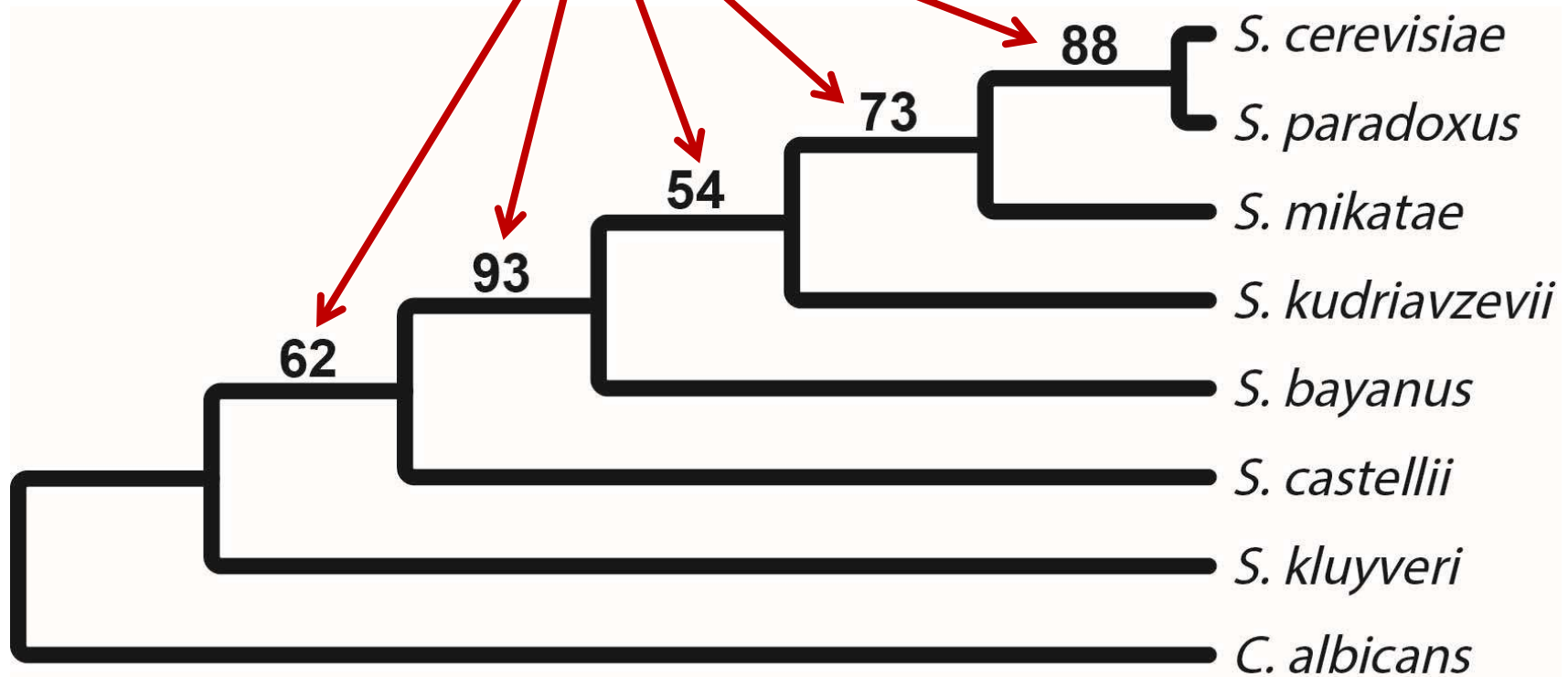


Bootstrap Support is Misleading When Used in Large Datasets



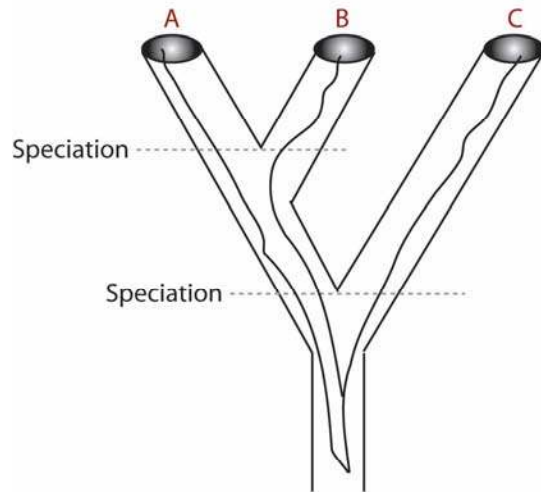
The Majority of Single Genes Support the Concatenation Phylogeny

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

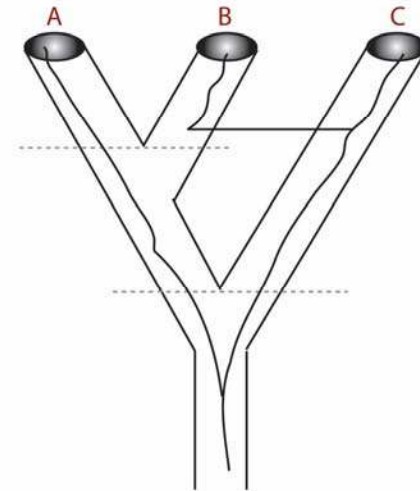


Gene Trees Can Differ from Species Trees

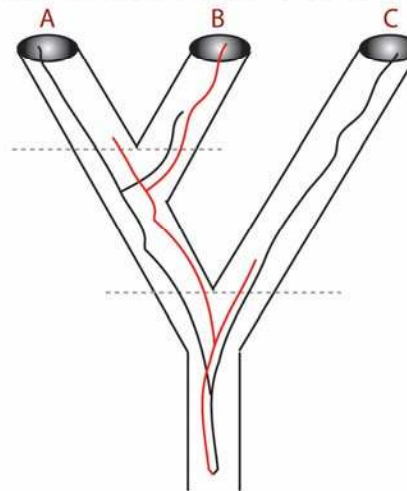
Lineage Sorting



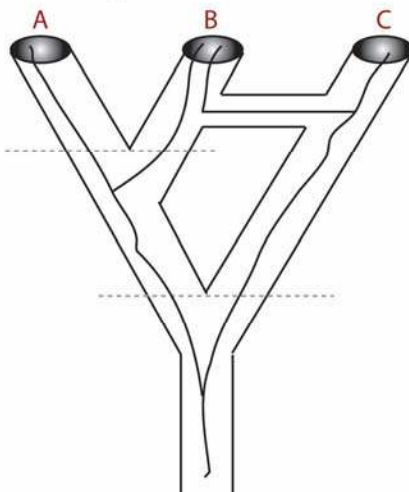
Horizontal Gene Transfer



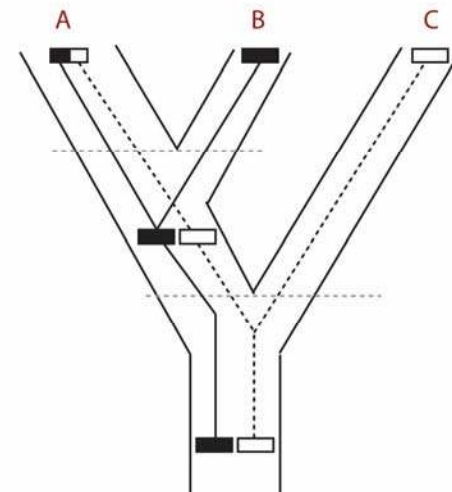
Gene Duplication and Loss



Hybridization

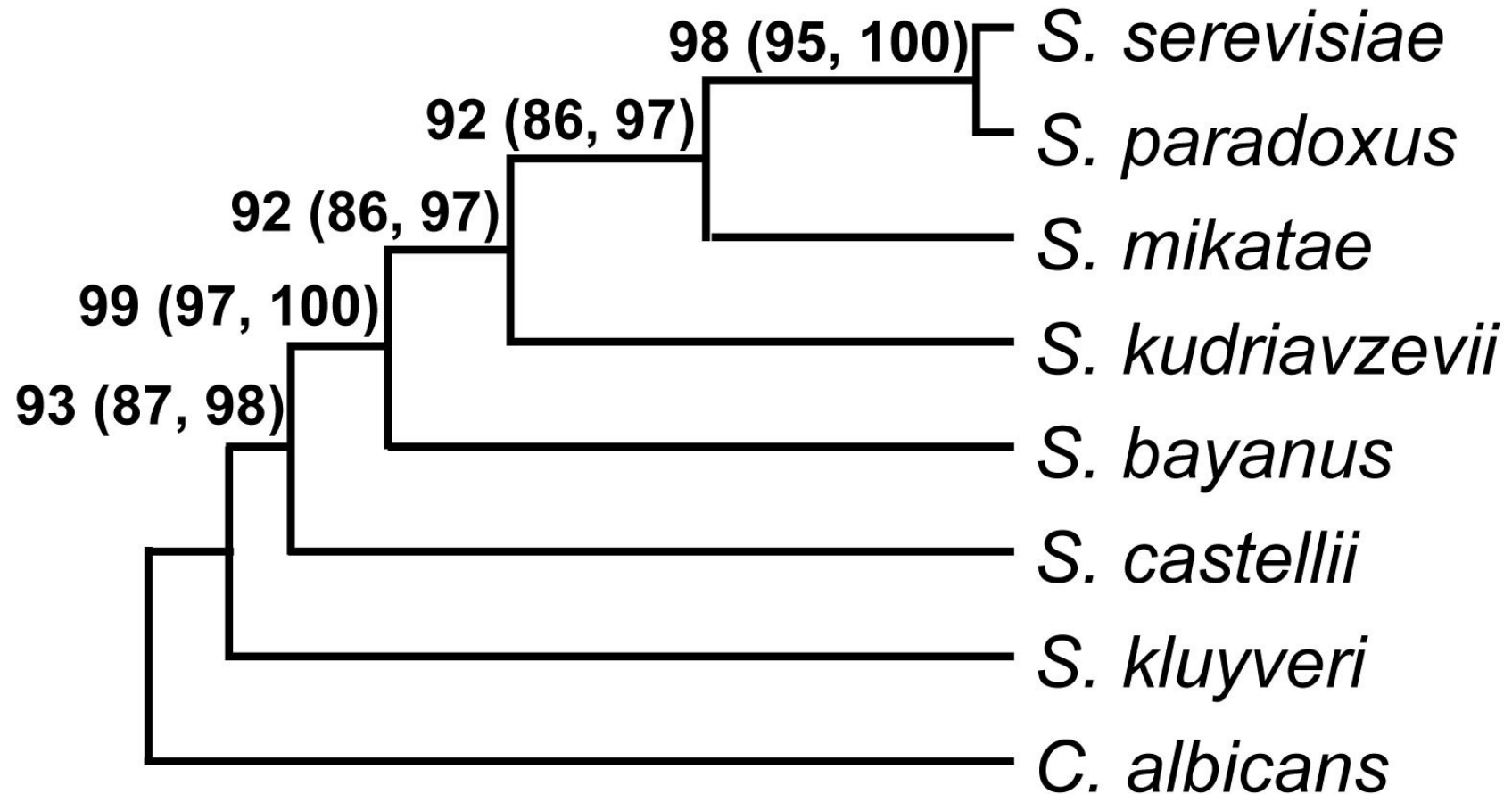


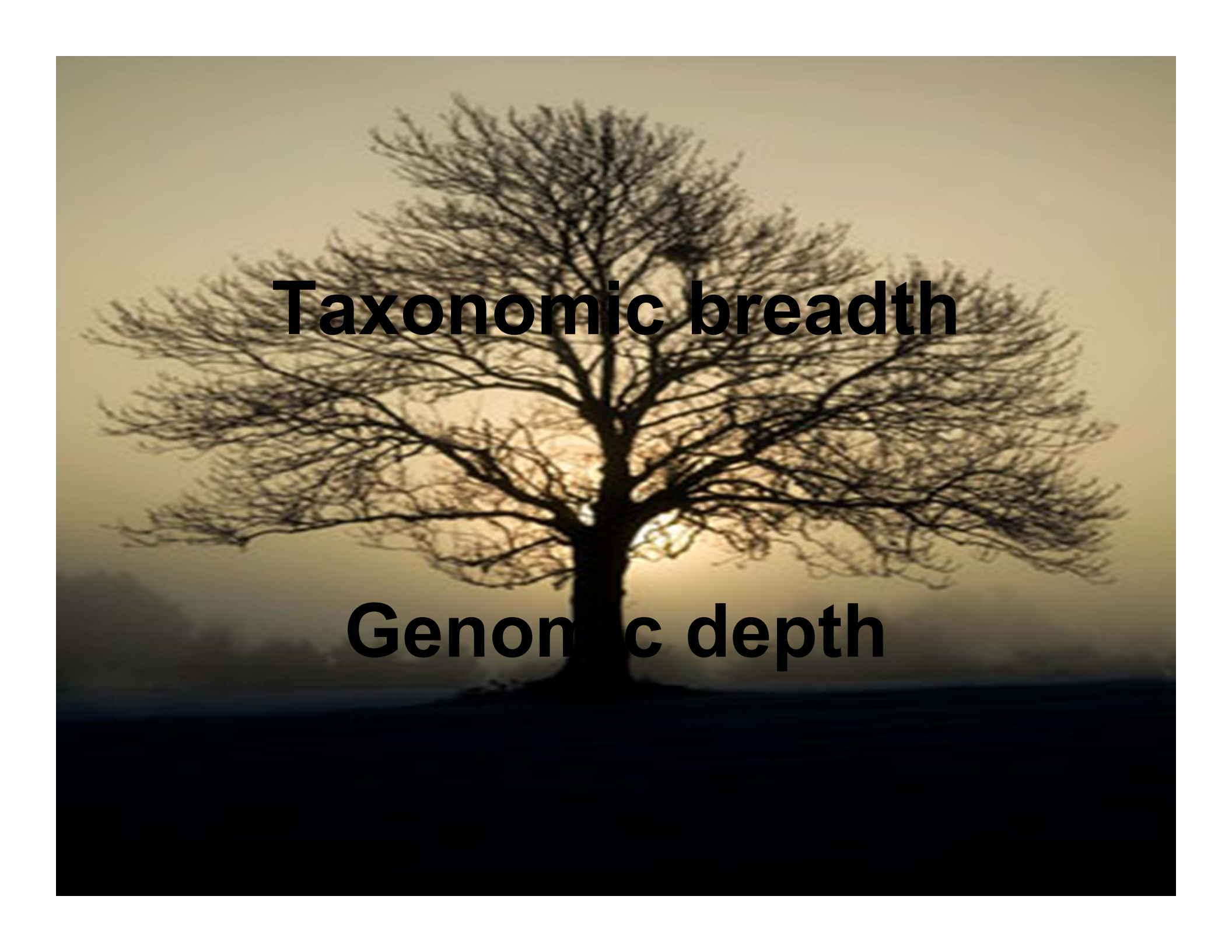
Recombination



Inferring the Species Phylogeny from Single Gene Histories

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





Taxonomic breadth

Genomic depth



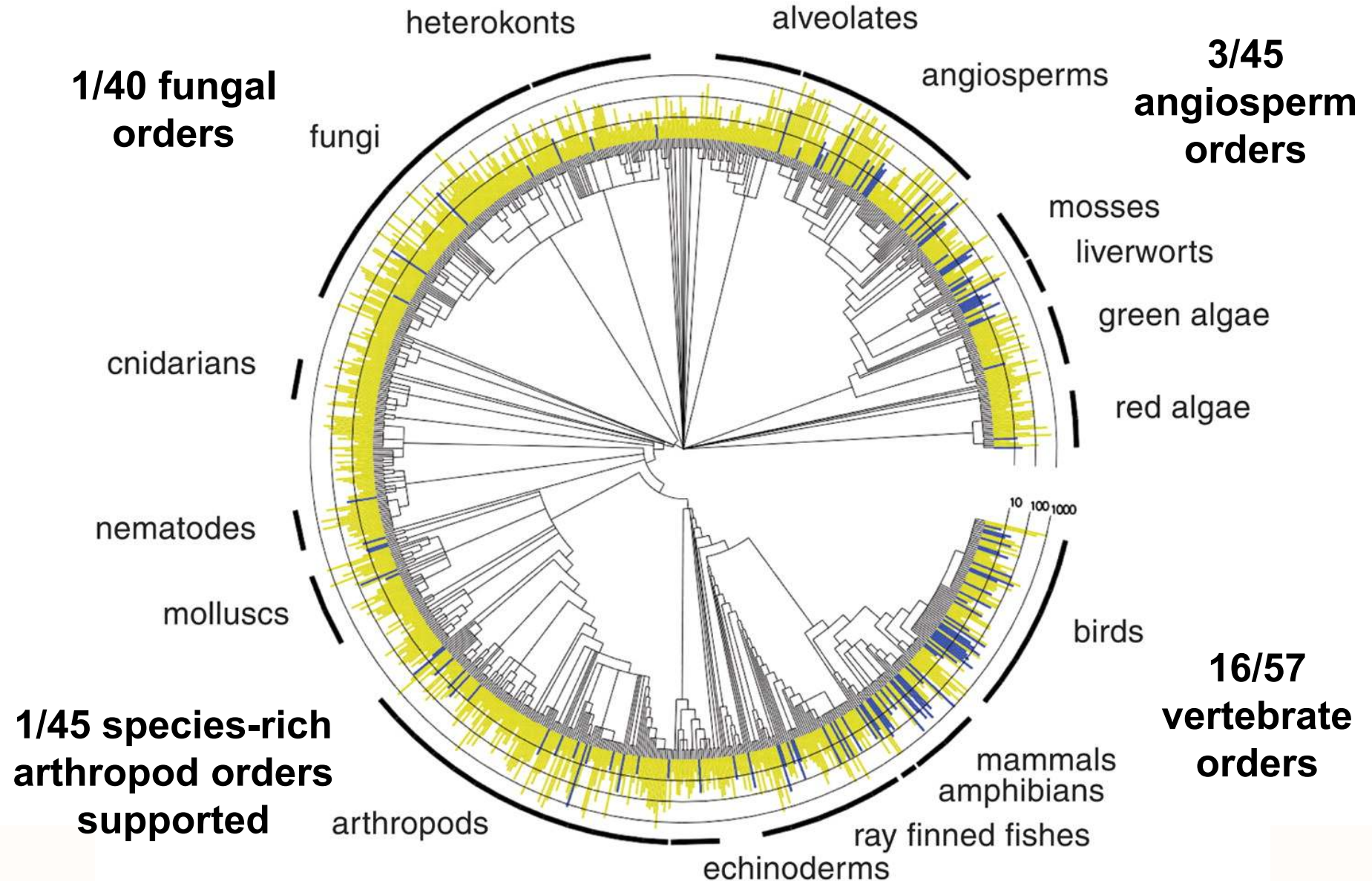
Estimating the Taxonomic Breadth of the Tree of Life

**2002: Cracraft
“guess-timates”
that 0.4% of all 2
million known
species have been
included in at least
one published
phylogeny**

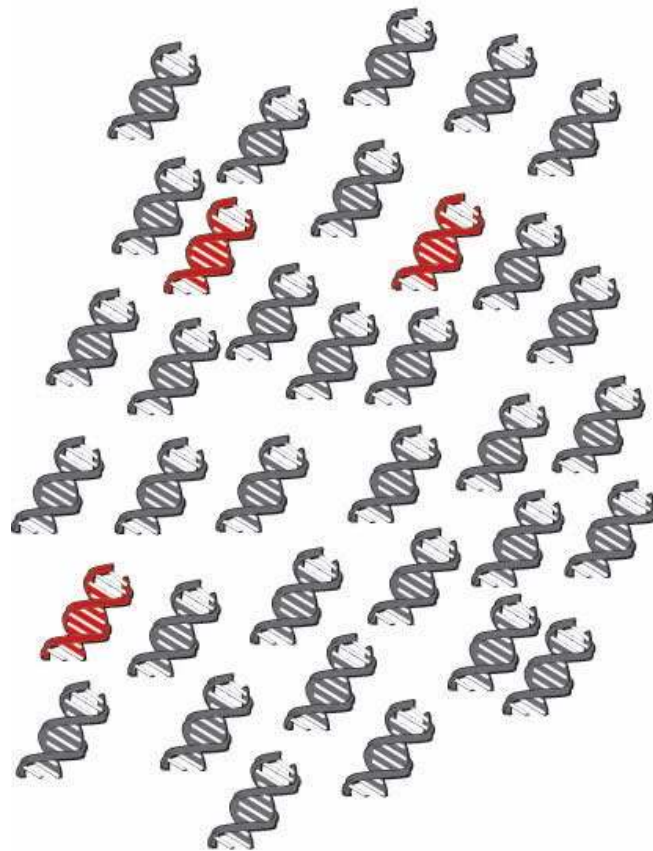
**2008: Sanderson
reports that
molecular
sequence data
have been sampled
from 10% of all 2
million known
species**



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

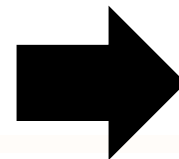


Can we Use RNA-Seq to Increase Genomic Depth?

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

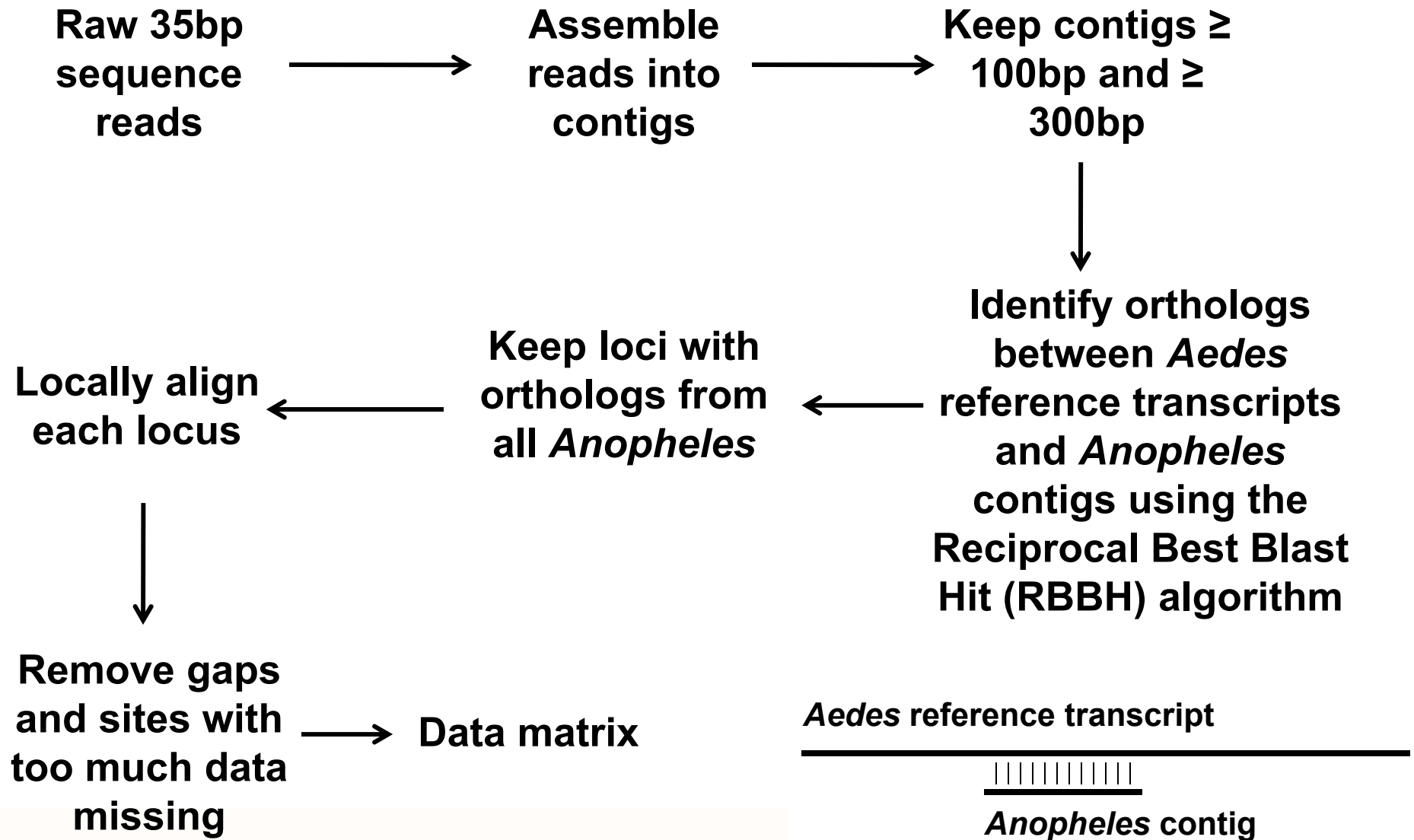


Illumina
RNA-Seq



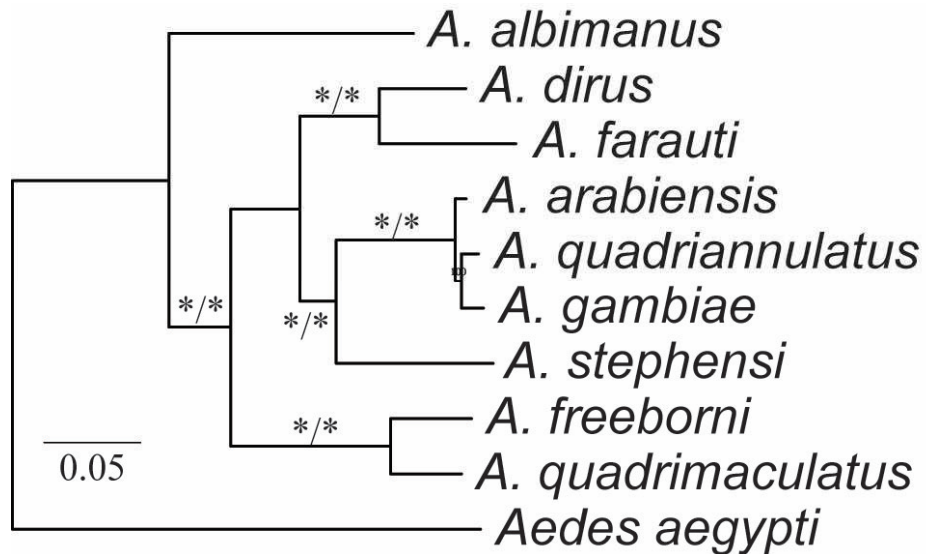
~150,000,000 reads
5,250,000,000 bp

Data Matrix Construction: The “Singlecontig” Strategy



Robust Phylogenetic Inference from RNA-Seq Data

Using ≥ 100 bp contigs

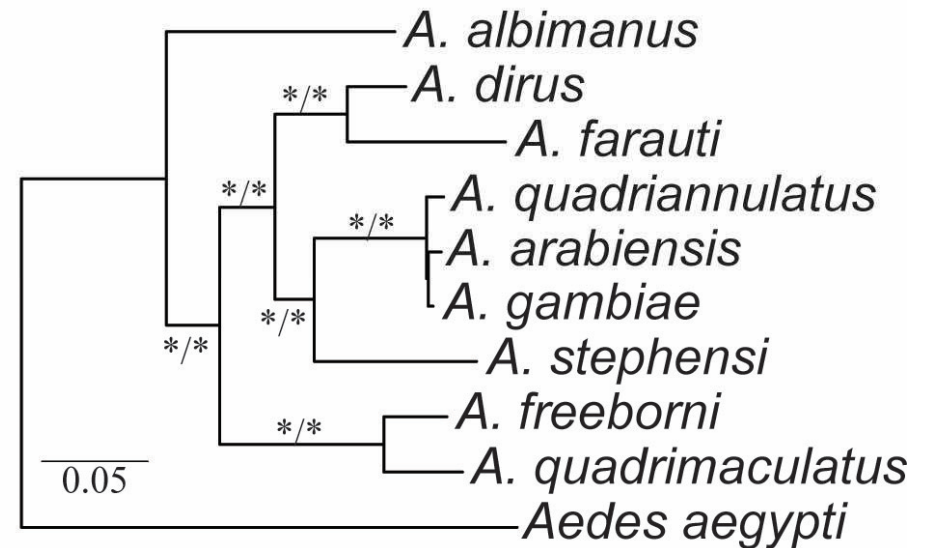


Loci = 553

Aln Length = ~390 Kb

% Missing data = 51

Using ≥ 300 bp contigs



Loci = 69

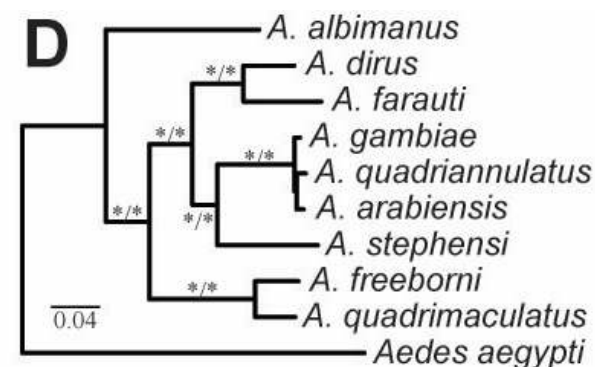
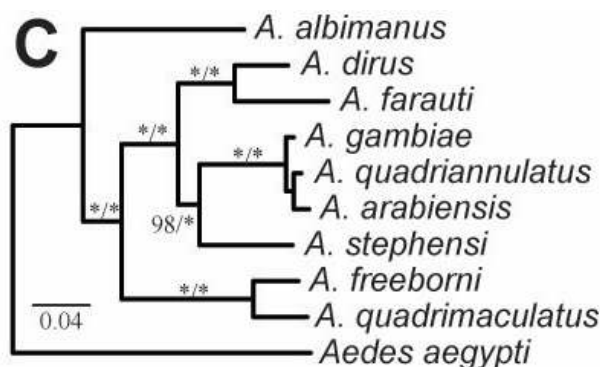
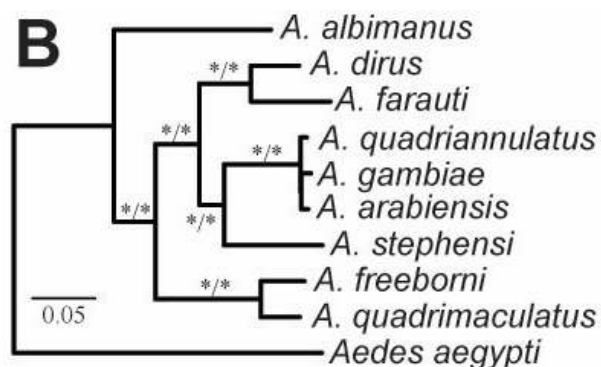
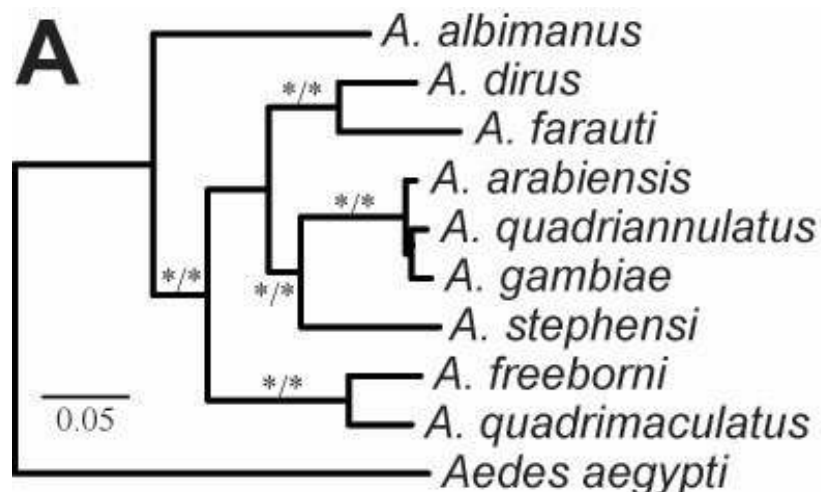
Aln Length = ~73 Kb

% Missing data = 44



Accurate Phylogenetic Inference From our Data

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



Exclude erroneous loci

491 loci

Aln L: ~329 Kb

Missing data: 50%

**Use only sites without
data missing**

Aln L: ~15 Kb

Missing data: 0%

Use *A. gambiae* as ref.

634 loci

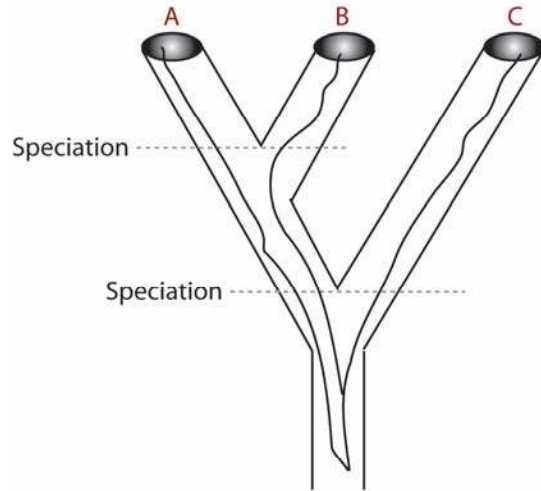
Aln L: ~472 Kb

Missing data: 50%

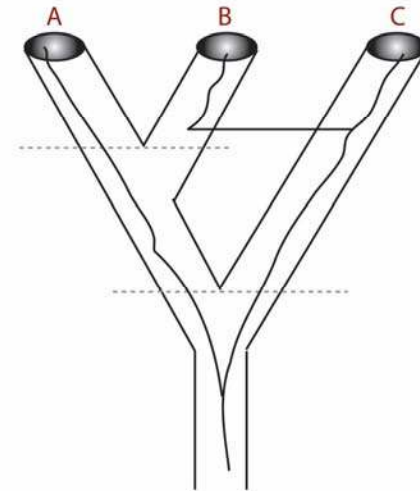


Gene Trees Can Differ from Species Trees

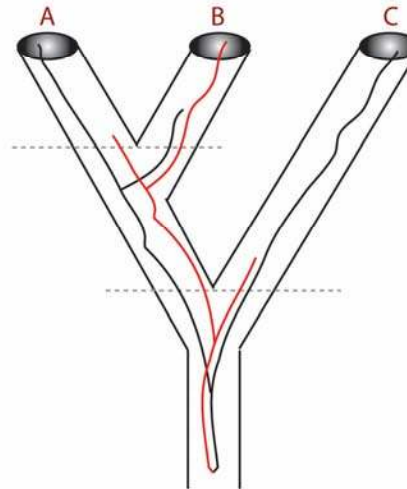
Lineage Sorting



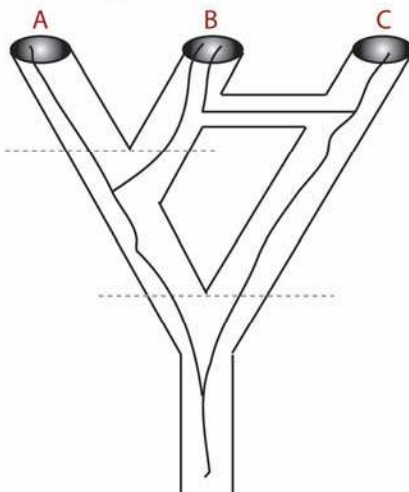
Horizontal Gene Transfer



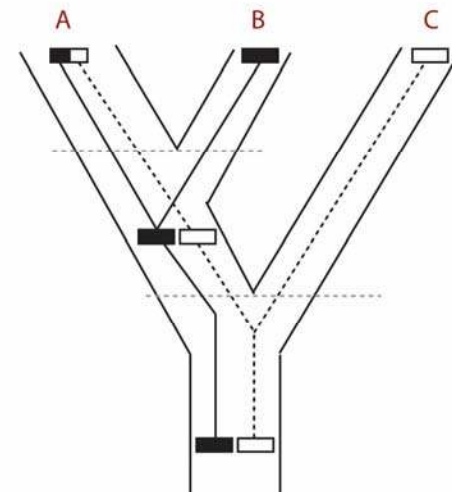
Gene Duplication and Loss



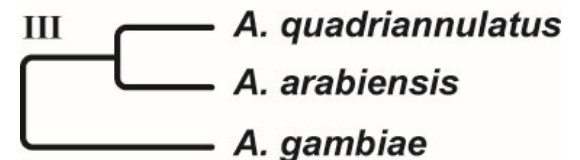
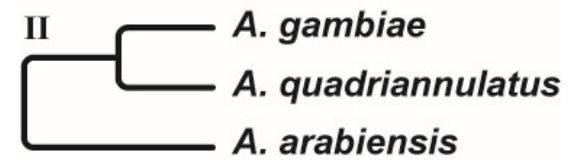
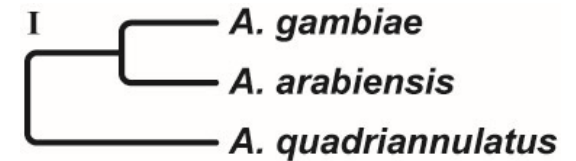
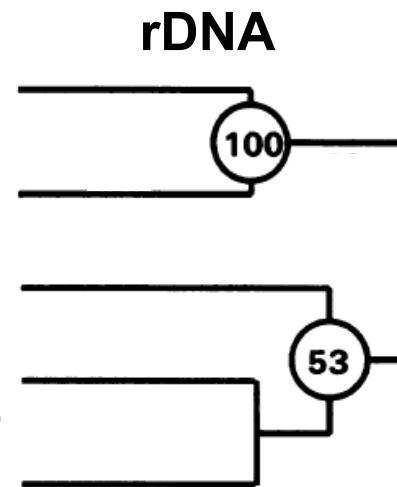
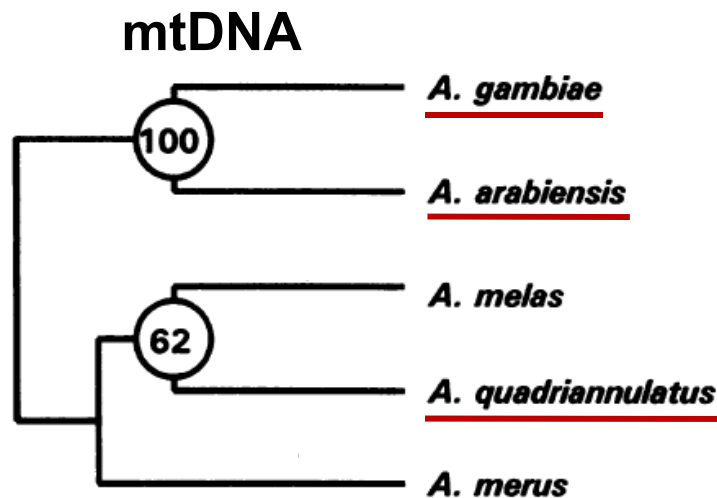
Hybridization



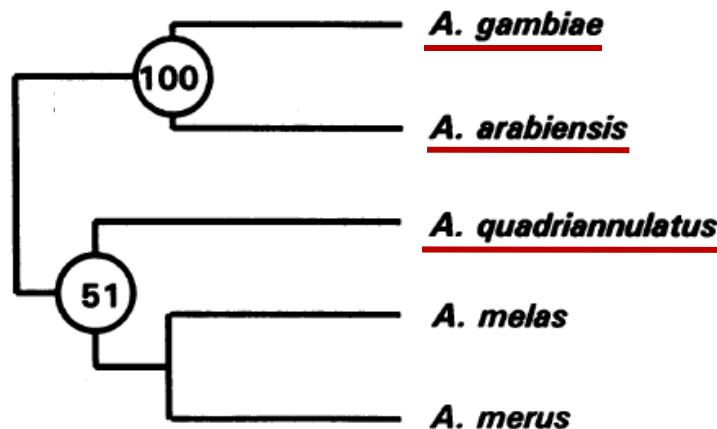
Recombination



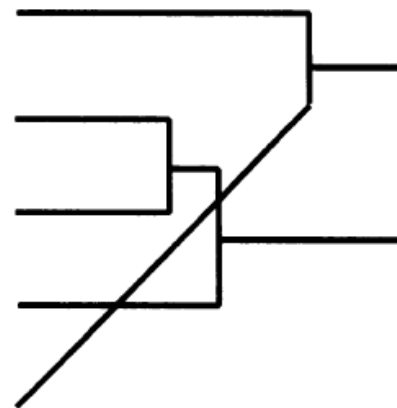
Our Data Matrices Can Detect Population-Level Events



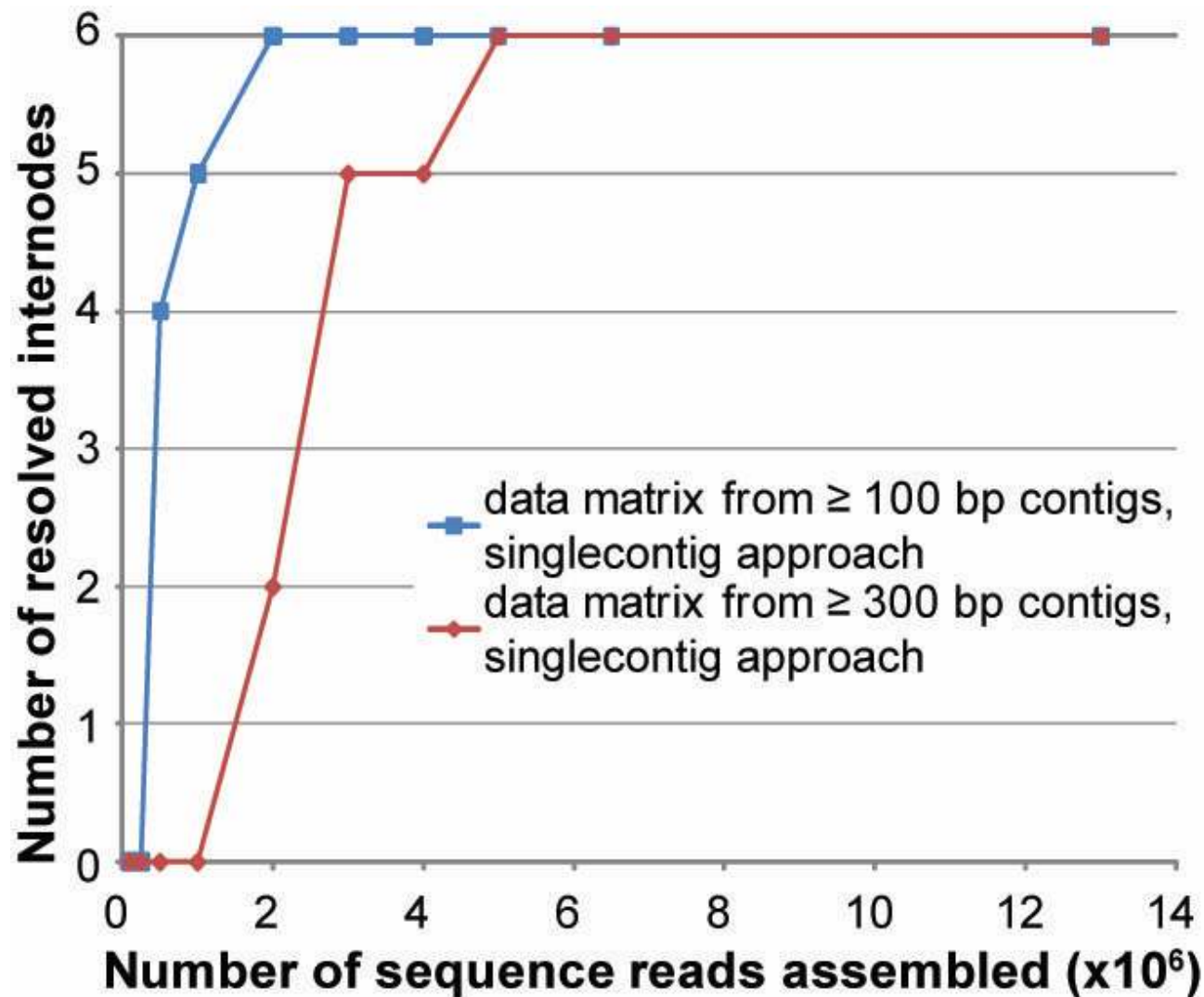
mtDNA + rDNA



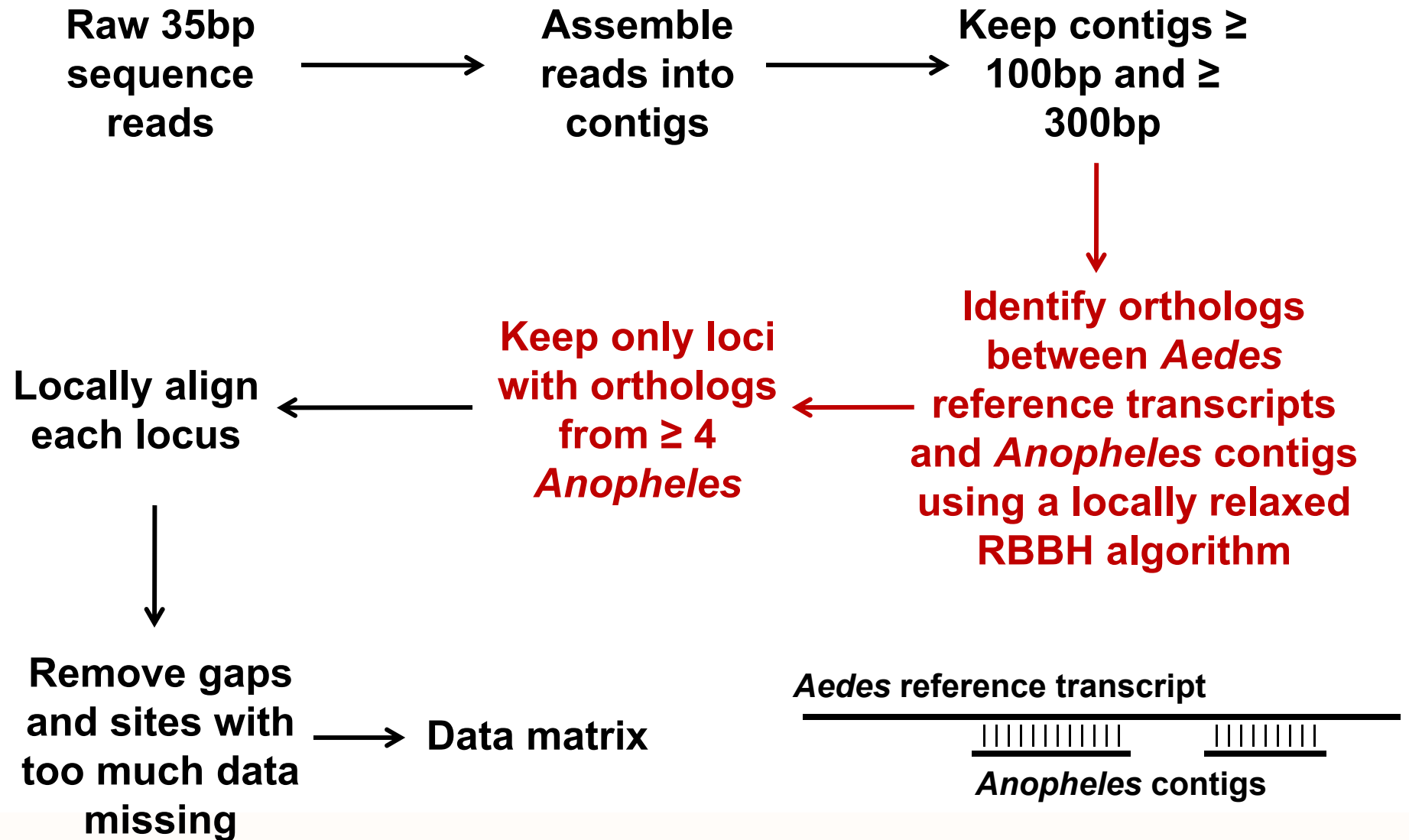
inversions



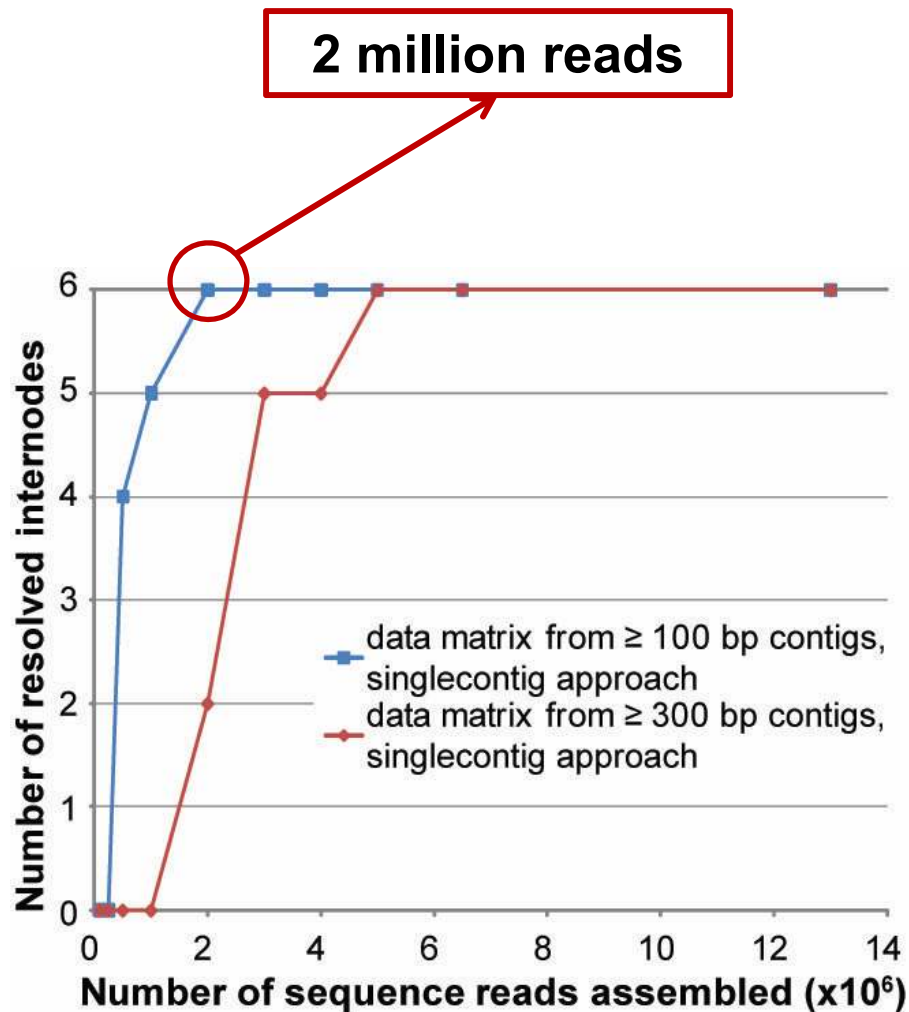
Robust Phylogenetic Inference From Few Sequence Reads



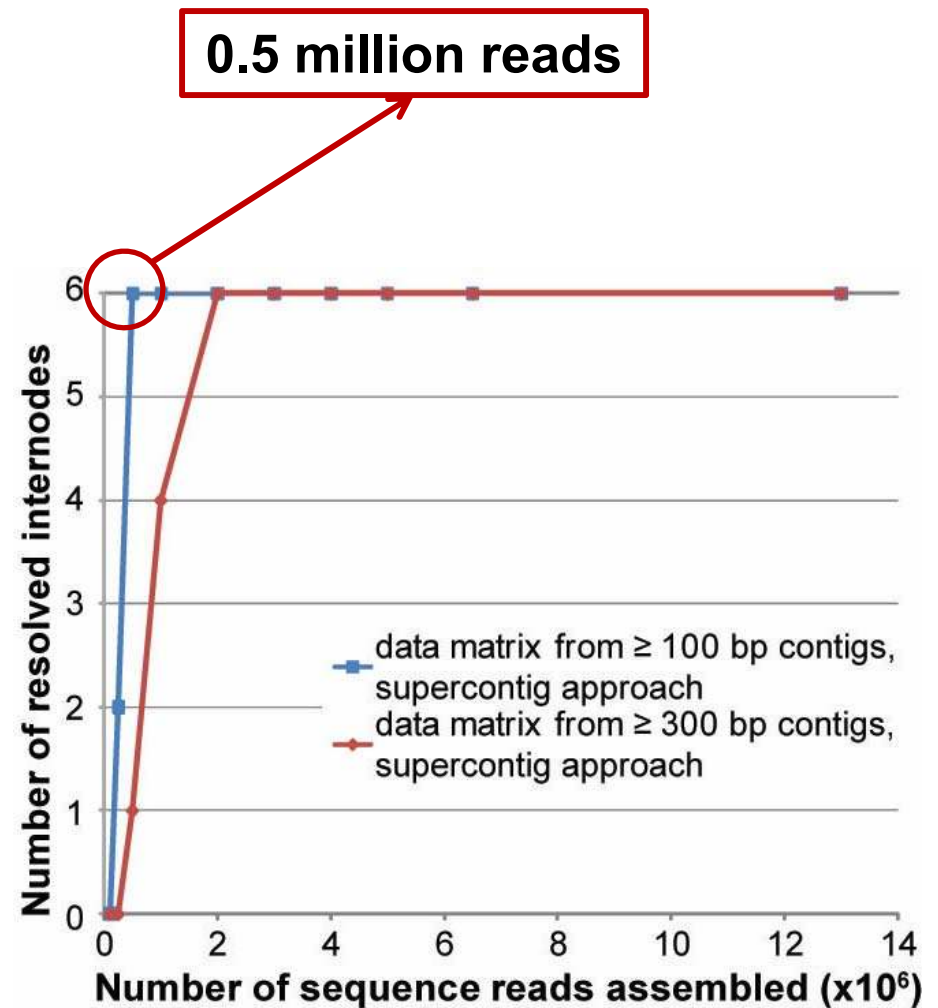
Experimental Design: The “Supercontig” Strategy



Robust Phylogenetic Inference From Very Few Reads



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



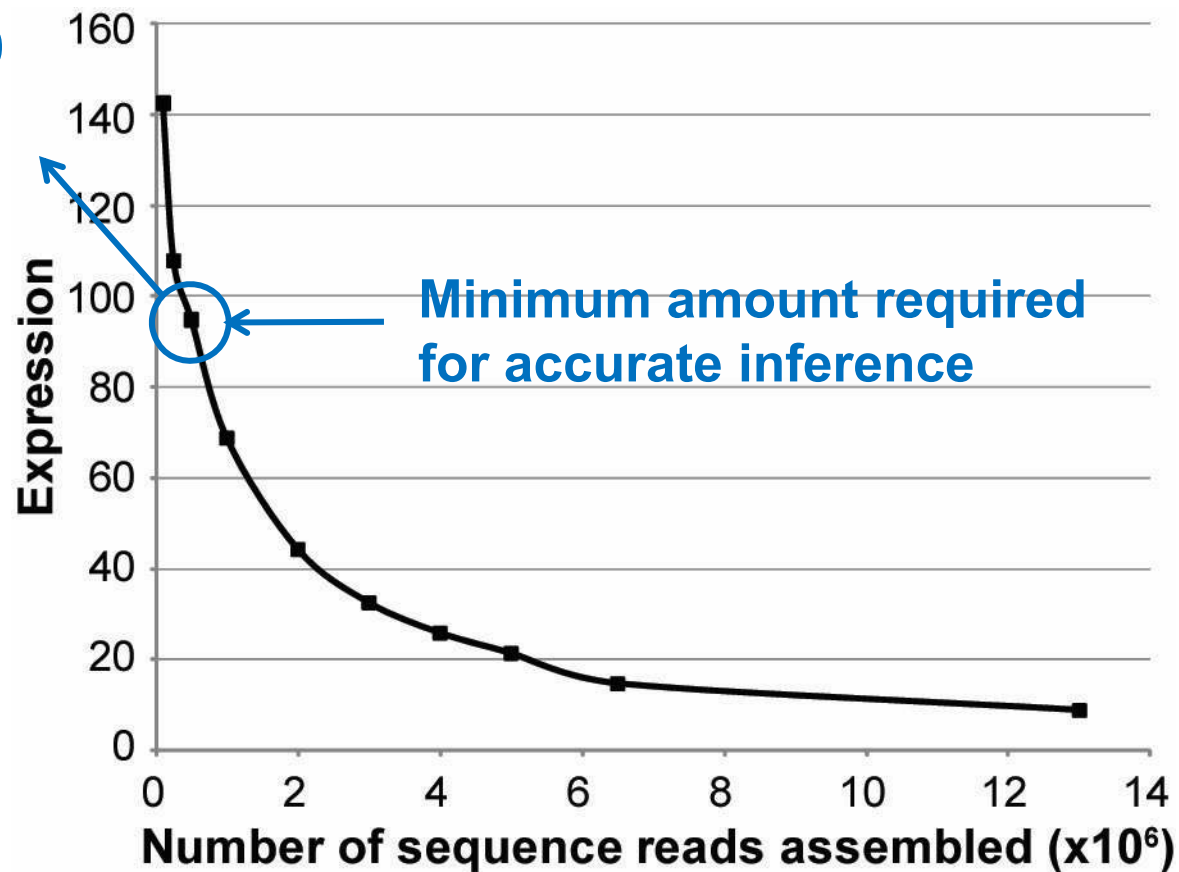
2,661 loci, AInL: ~971Kb, %miss: 62



Our Sequences are from Highly-Expressed Transcripts

2008 cost: ~\$50

2014 cost: < \$5



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

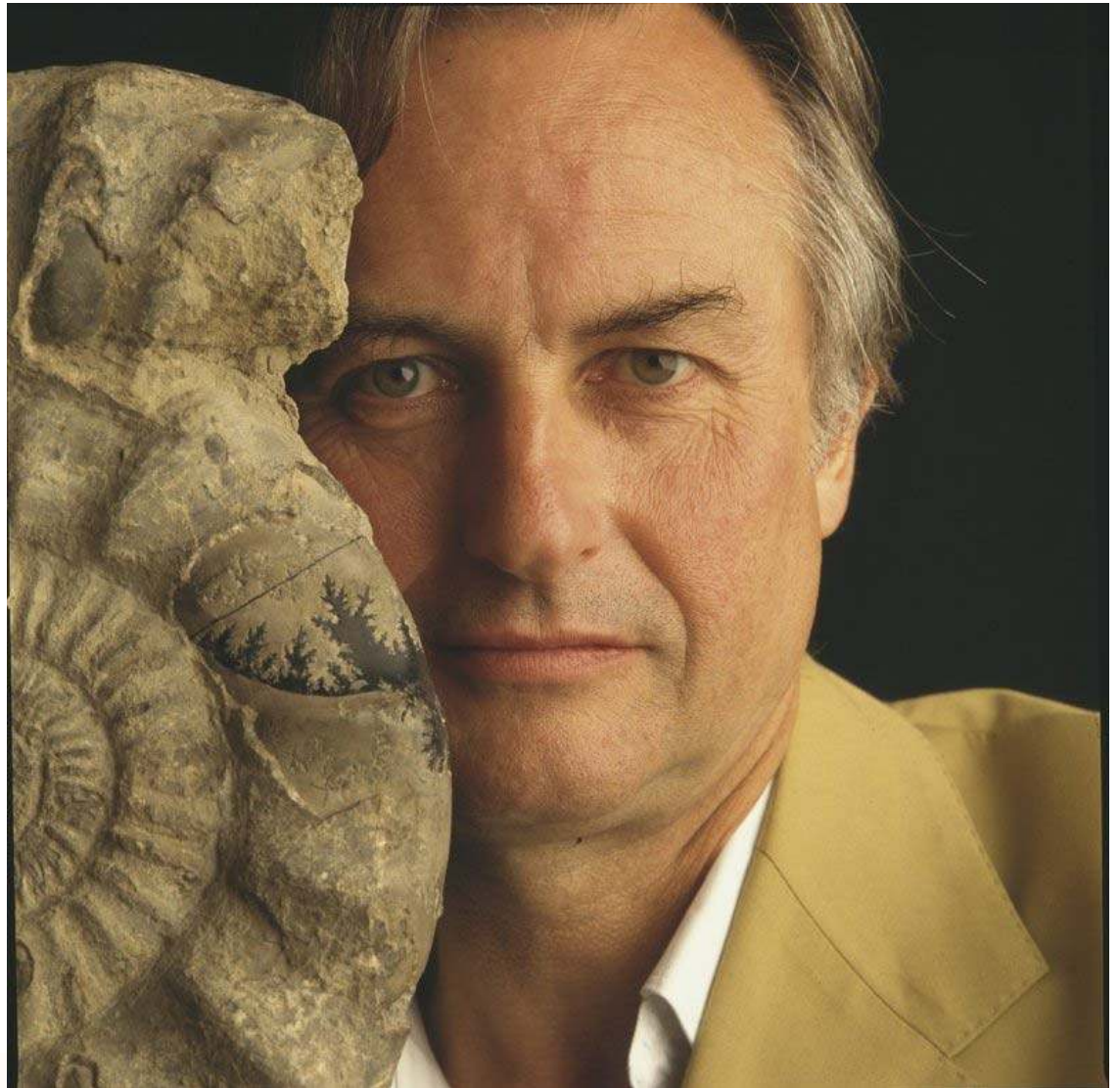
Department of Biology, U

*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

The Dawkins Delusion

“... there is, after all, one true tree of life [...]. It exists. It is in principle knowable. We don't know it all yet. By 2050 we should – or if we do not, we shall have been defeated only at the terminal twigs, by the sheer number of species.”

Richard Dawkins

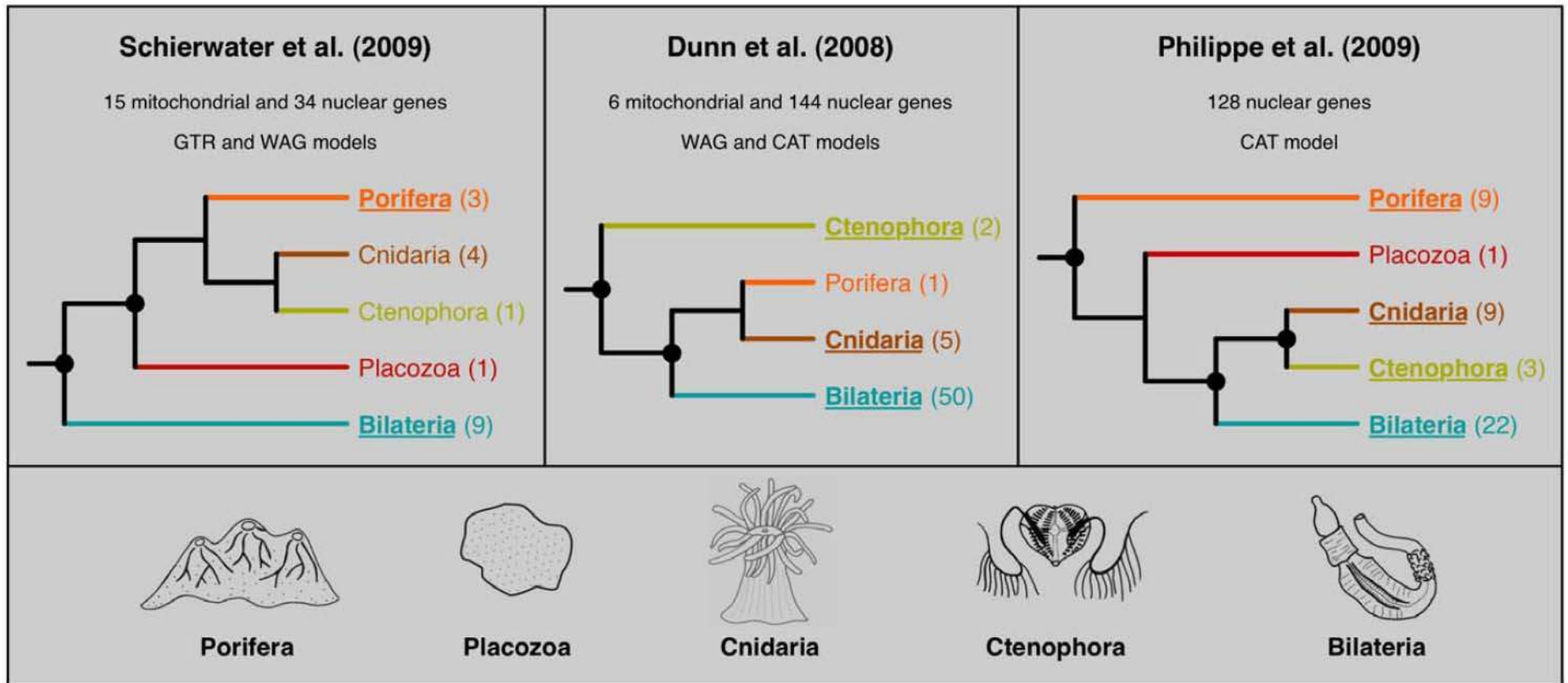


Dawkins (2003), A Devil's Chaplain

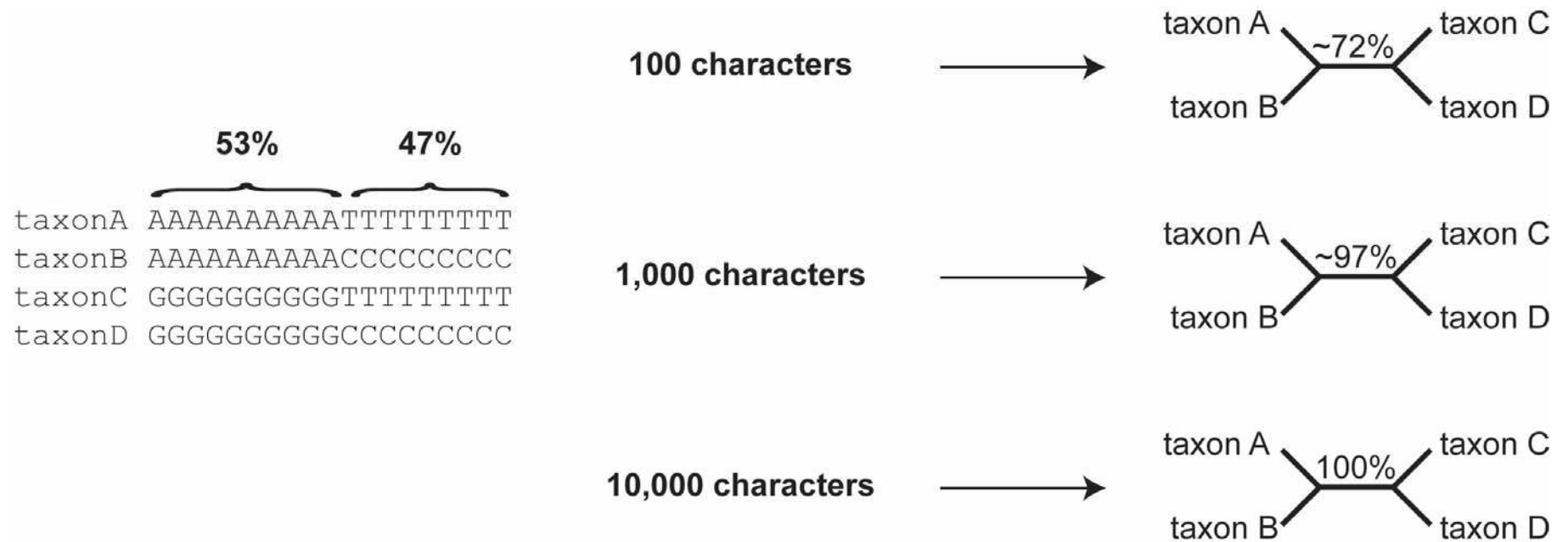
Incongruence in Deep Time



Incongruence in Deep Time



Bootstrap Support is Misleading When Used in Large Datasets



An Expanded Yeast Data Matrix

Yeast Gene Order Browser (YGOB)



Candida Gene Order Browser (CJOB)



**Saccharomyces
lineage**

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

**Candida
lineage**



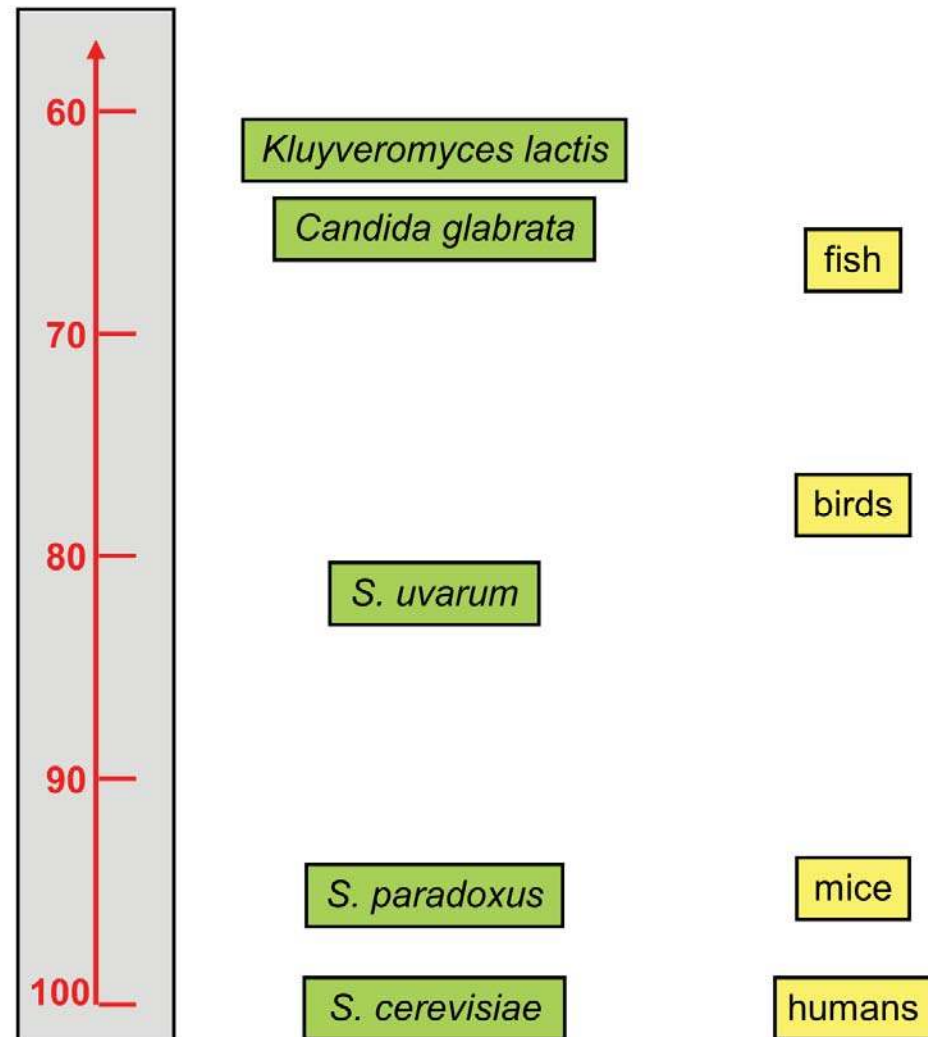
Byrne & Wolfe (2005) Genome Res.

Fitzpatrick et al. (2010) BMC Genom.

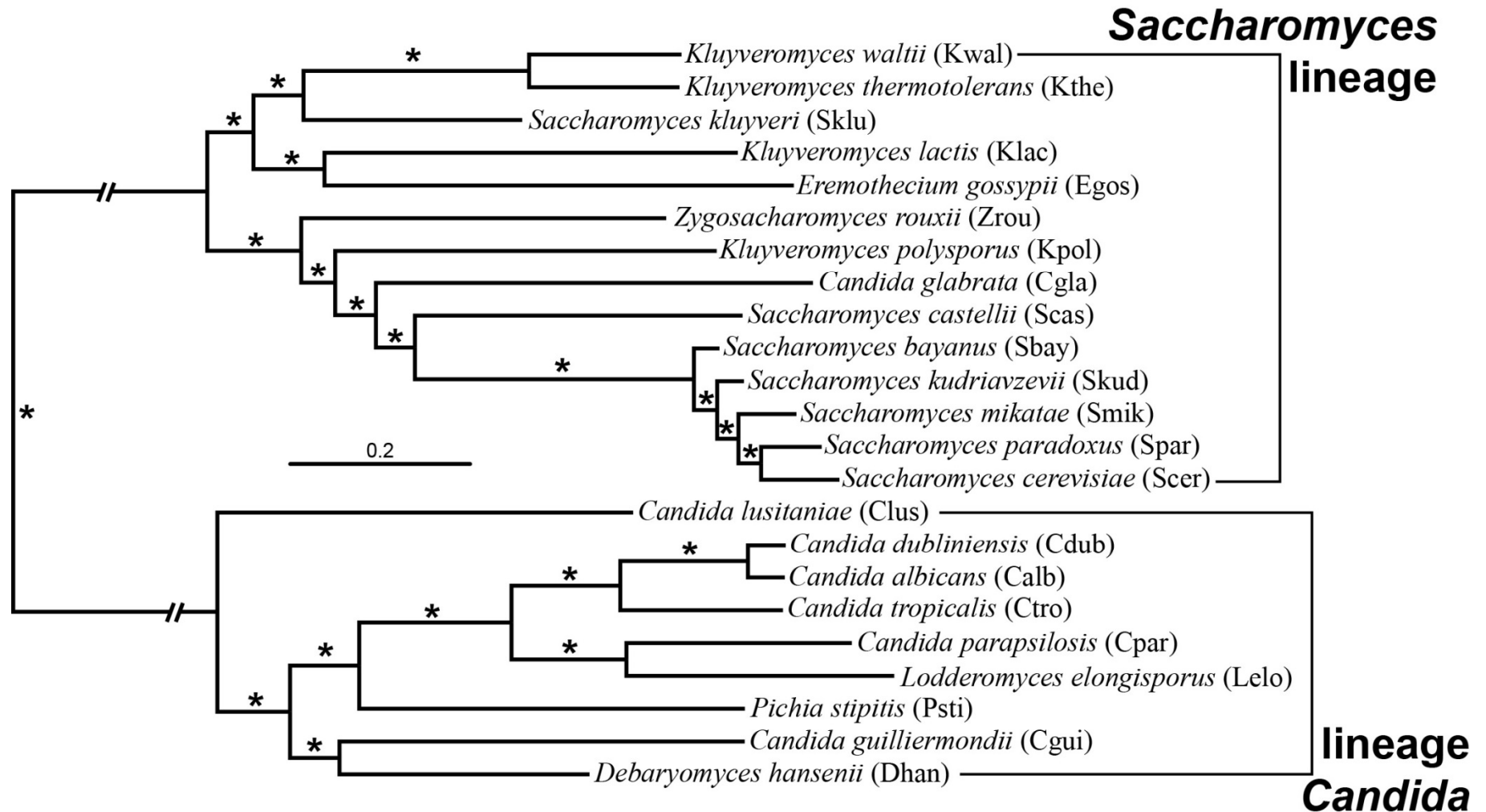
Yeast Divergence is Intermediate to Vertebrates and Animals

Proteome-wide average pairwise amino acid sequence similarity

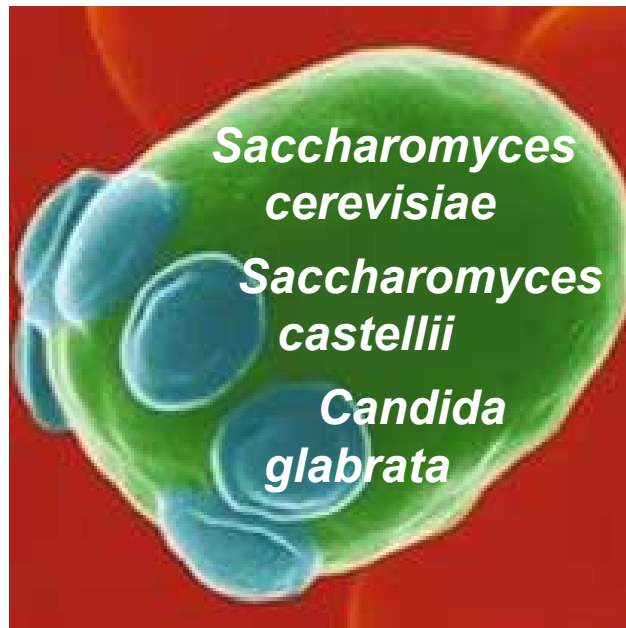
***Saccharomyces*,
Candida, *Kluyveromyces*,
etc. are all polyphyletic
genera**



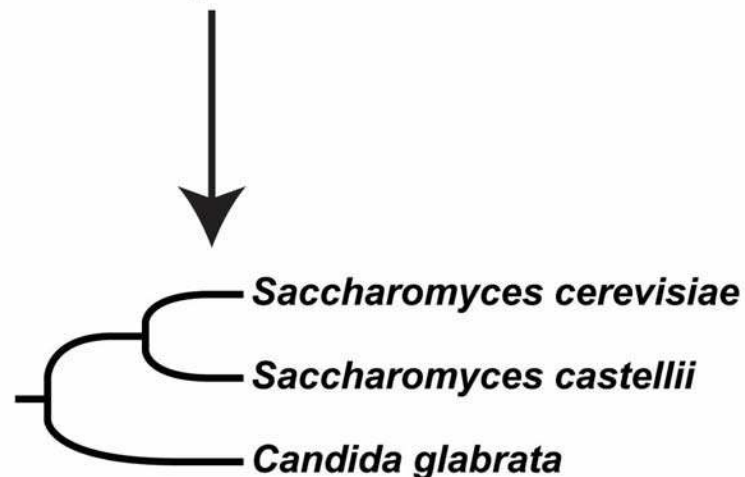
Concatenation Yields an Absolutely Supported Phylogeny



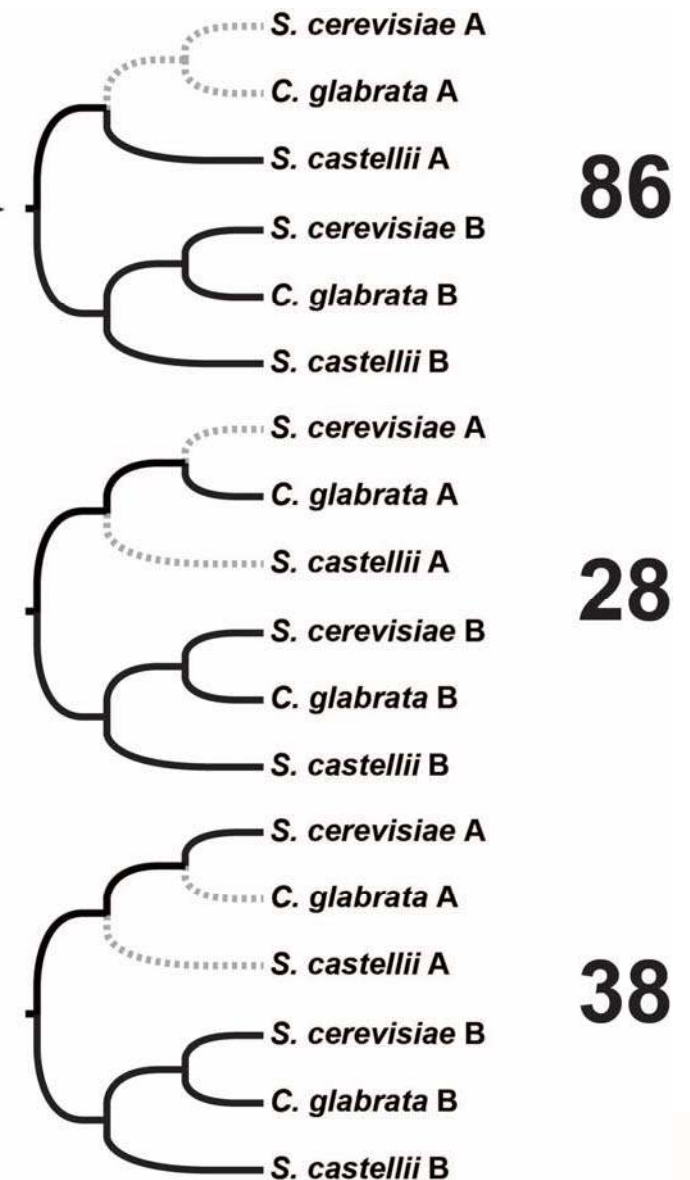
The Concatenation Phylogeny is at Least Partly Wrong



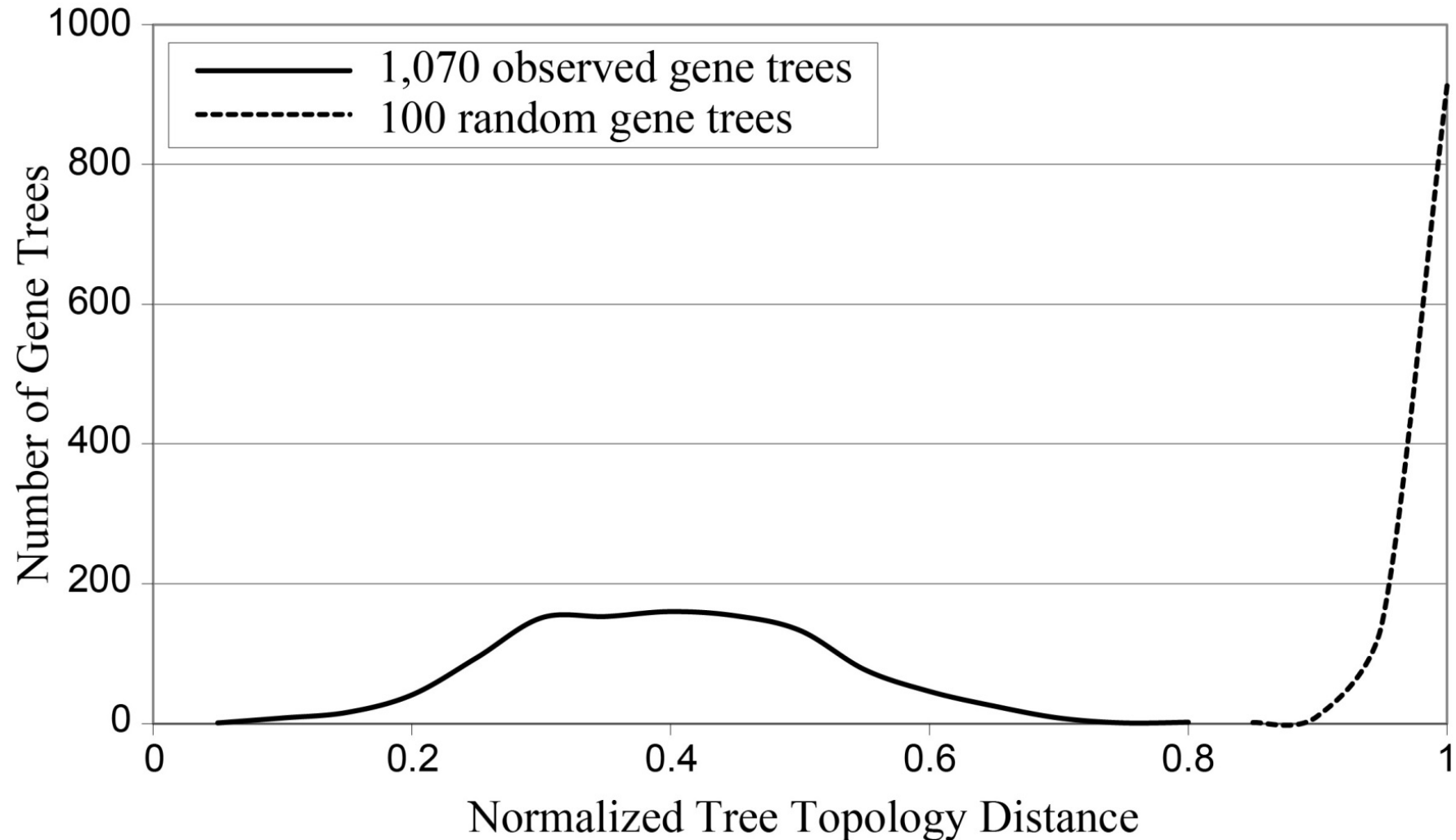
Linear Sequence Data



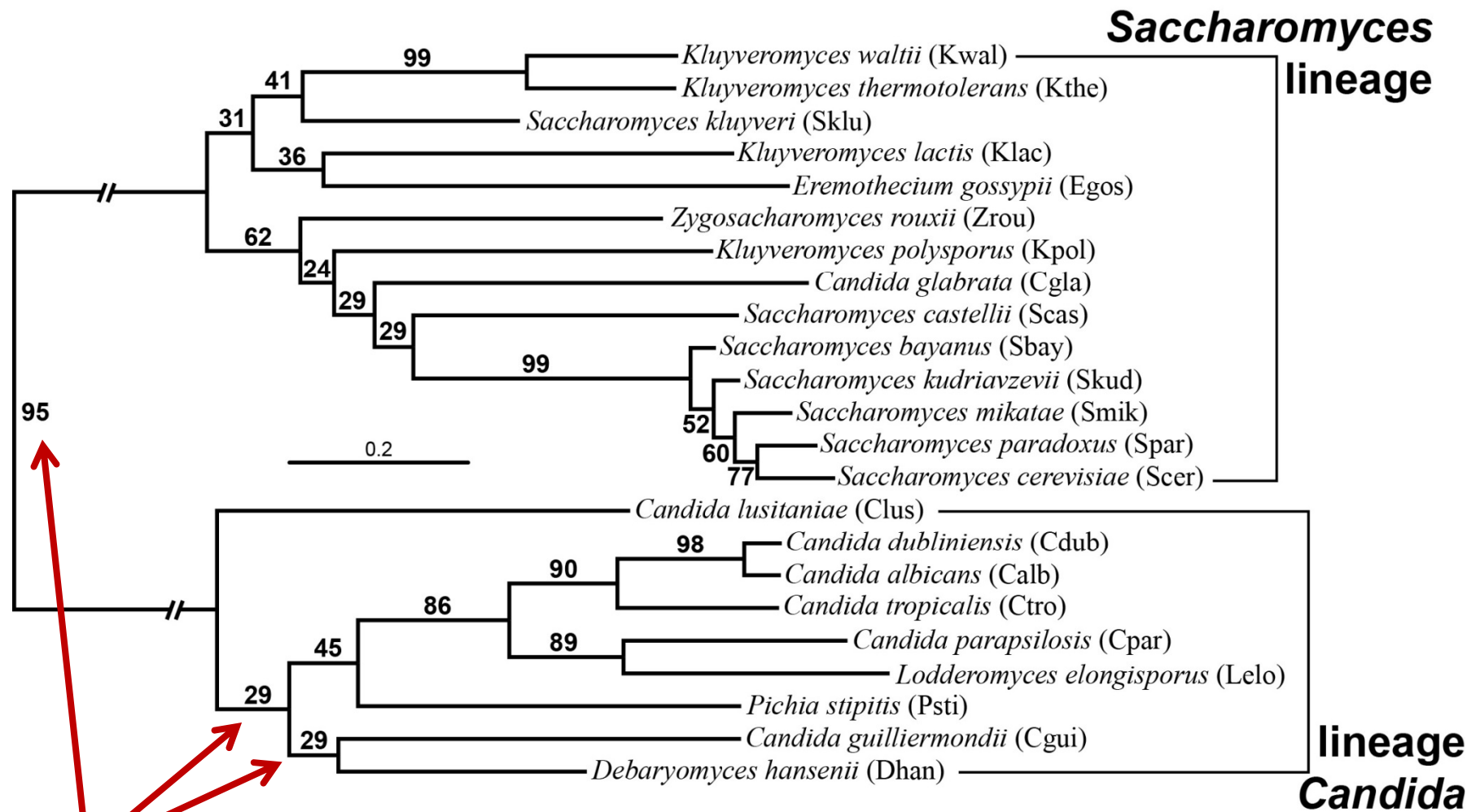
Rare Genomic Changes



All Gene Trees Differ from the Concatenation Phylogeny



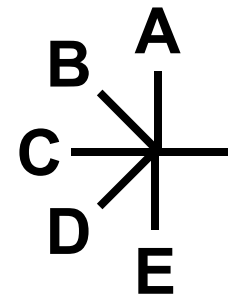
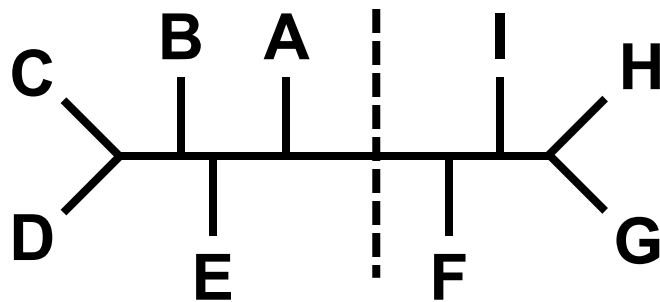
The Yeast Phylogeny Inferred by Majority-Rule Consensus



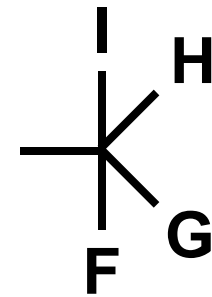
Phylogenetic Trees are Sets of Splits

Division

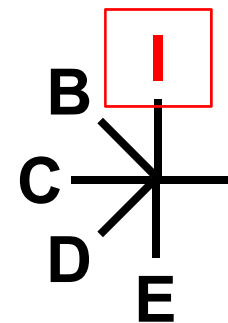
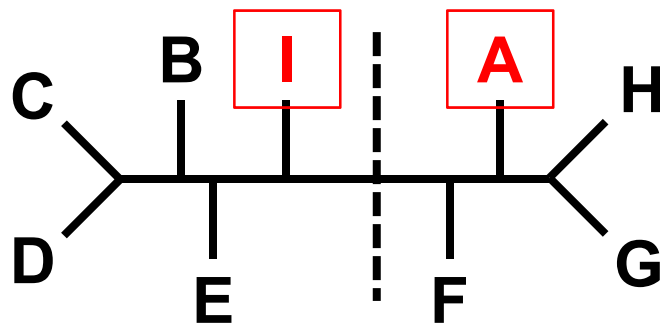
Splits



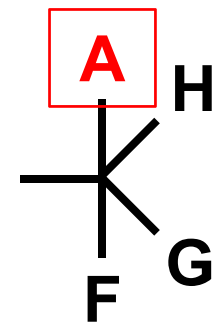
$\{A, B, C, D, E\}$



$\{F, G, H, I\}$



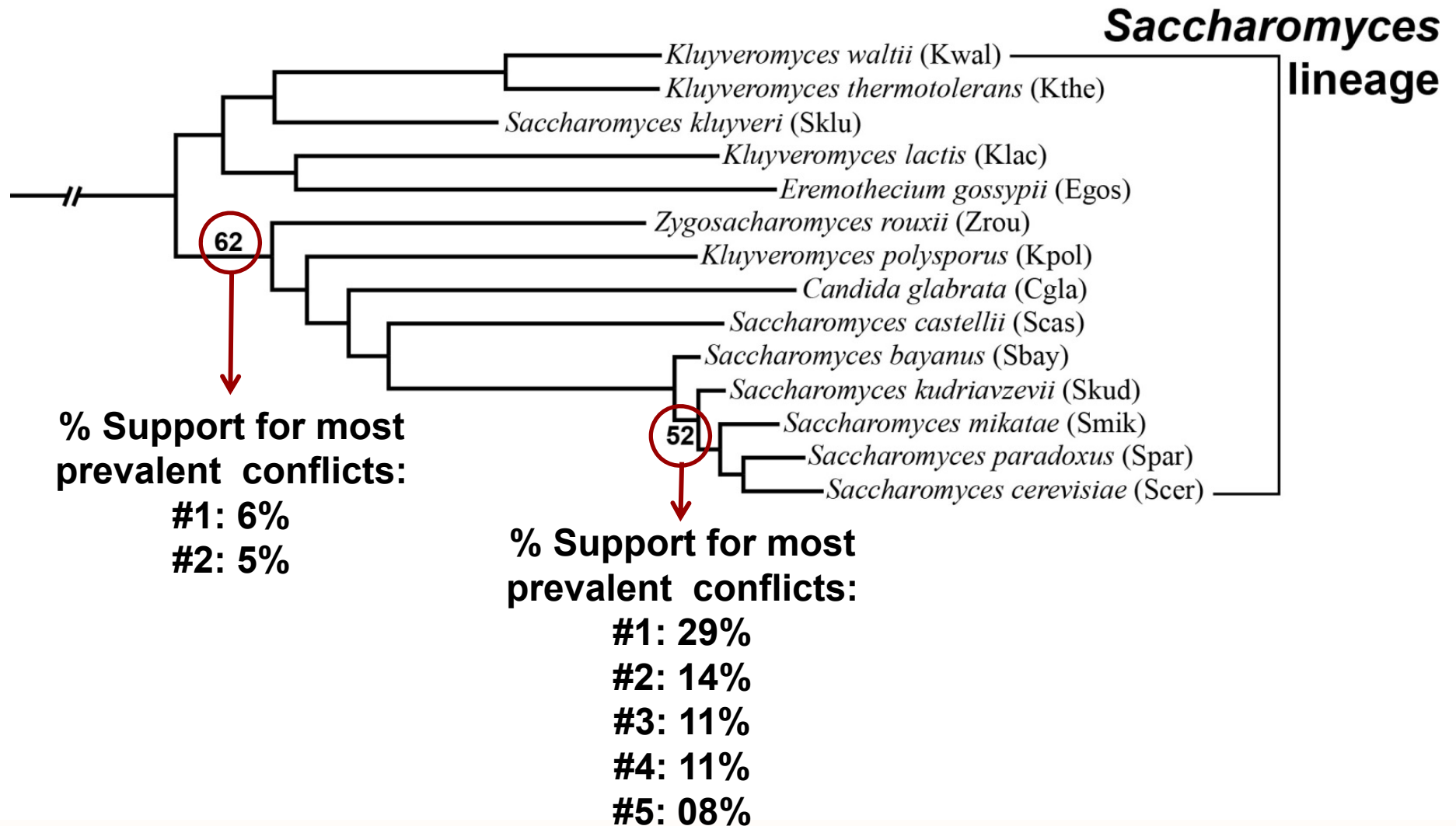
Conflicting Splits: $\{I, B, C, D, E\}$



$\{F, G, H, A\}$

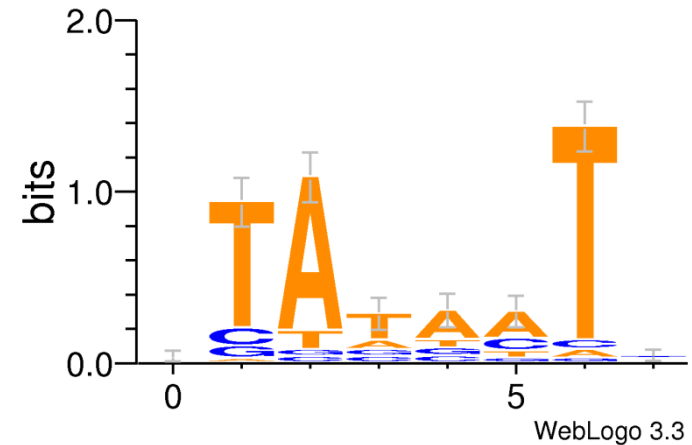
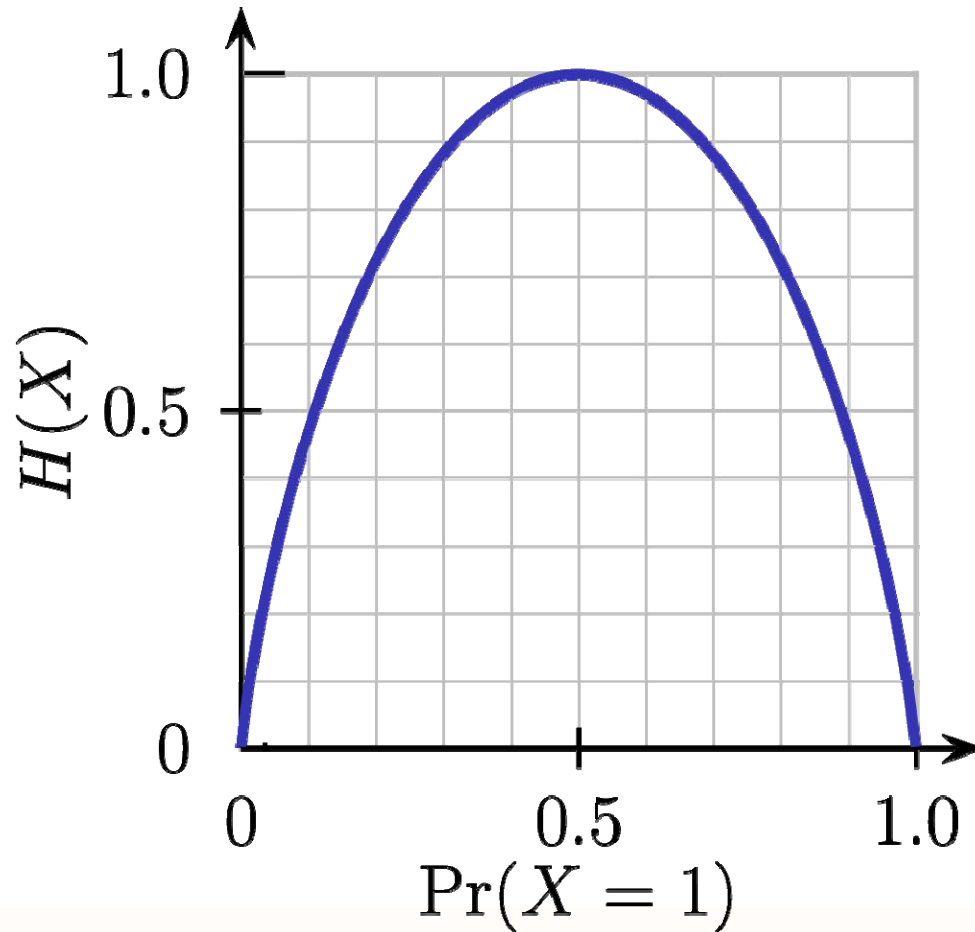


Measuring the Degree of Conflict for each Internode



Shannon's Entropy Measures the Uncertainty in a Variable

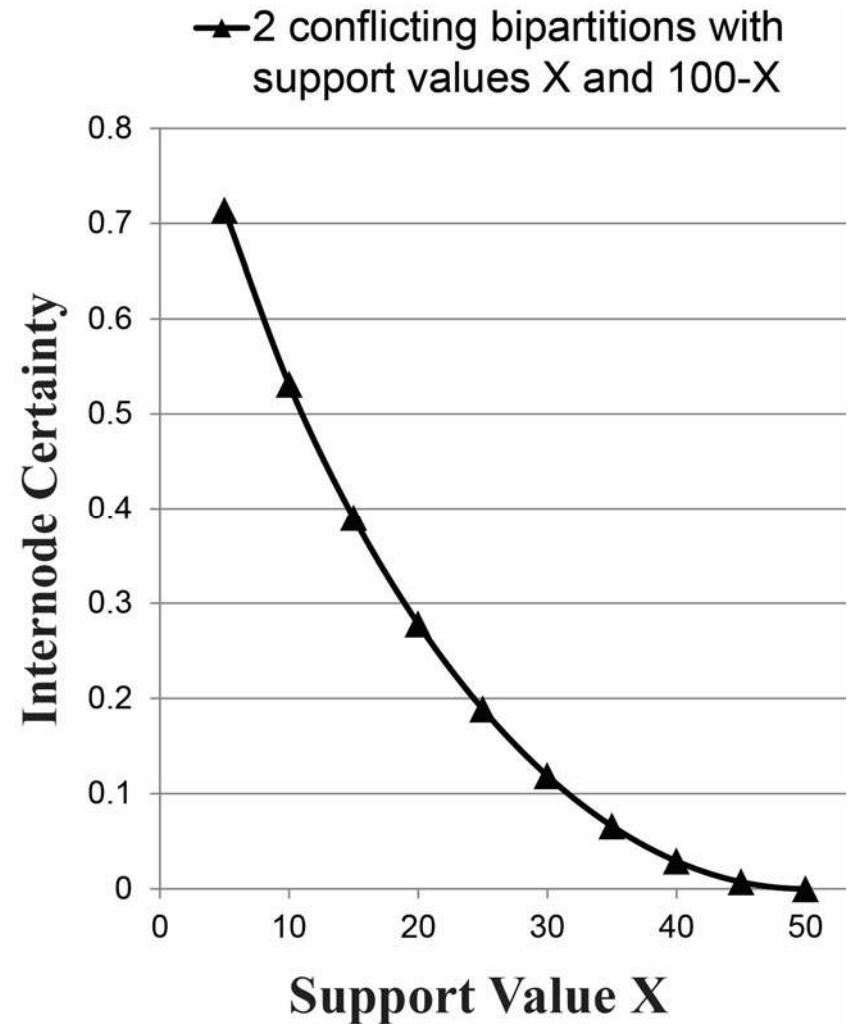
$$H(X) = - \sum_1^n p(X_n) \log_2(p(X_n))$$



Internode Certainty, a New Metric to Quantify Incongruence

Internode Certainty (IC): an evaluation 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. IC equal to 1 describes absolute certainty for a given internode.

Tree Certainty (TC): the sum of IC across all internodes. The max TC value is equal to the taxon number



$$IC = \log_2(2) + \left(\frac{x_1}{x_1 + x_2} \right) * \log_2 \left(\frac{x_1}{x_1 + x_2} \right) + \left(\frac{x_2}{x_1 + x_2} \right) * \log_2 \left(\frac{x_2}{x_1 + x_2} \right)$$

Internode Certainty in RAxML

IC and related measures are implemented in latest versions of RAxML
(<https://github.com/stamatak/standard-RAxML>)

```
[rokasa@vmeps65]$ more RAxML_verboseSplits.T6
```

```
Cint
```

```
Dmel
```

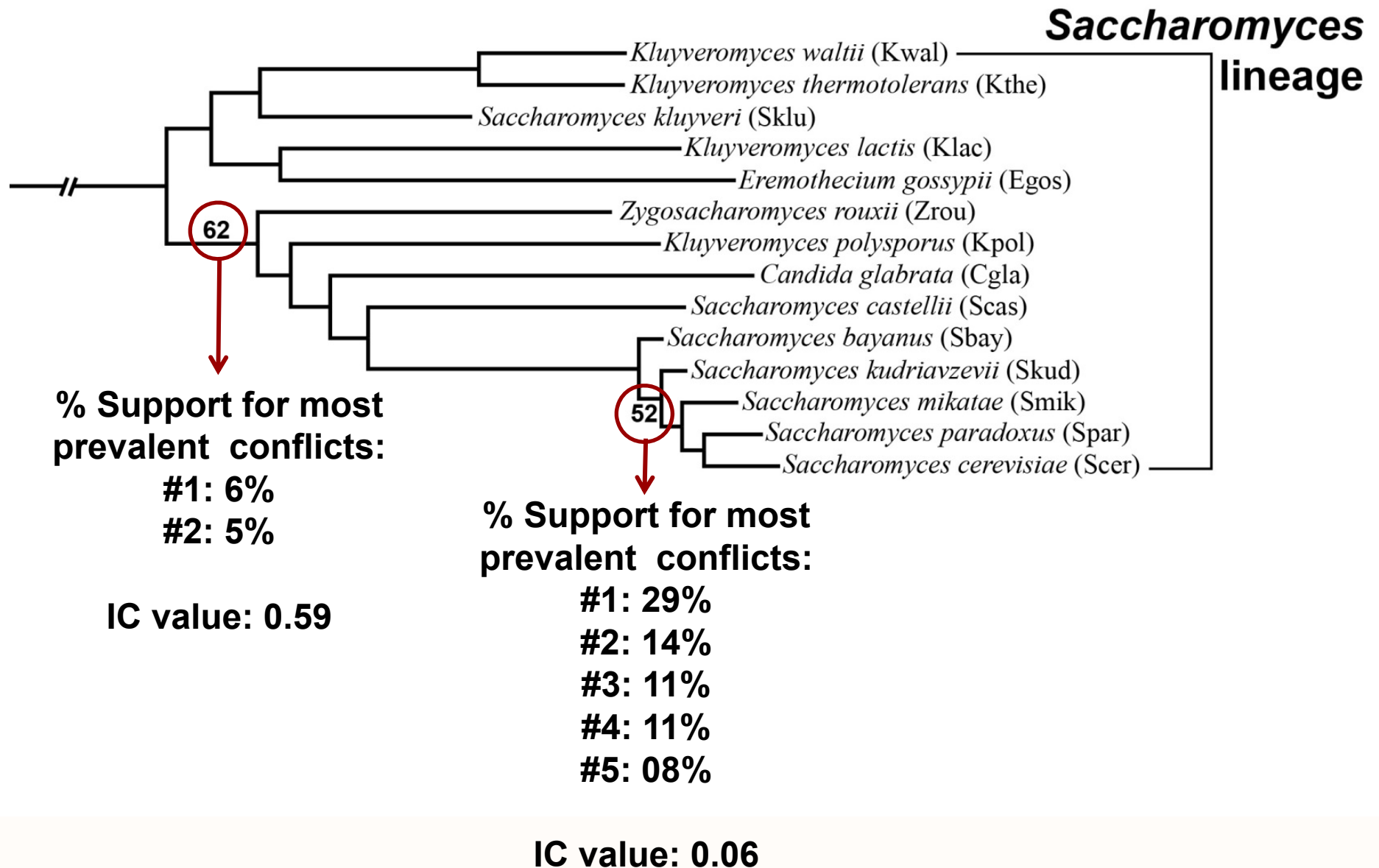
```
Sman
```

```
...
```

		Tree #	/	GSF	/	IC
partition:	-----**-----	211	/	93.77	/	0.82
	-----*-*----	6	/	2.66	/	0.82
partition:	-----**-*----	170	/	75.55	/	0.56
	-----*-**----	17	/	7.55	/	0.56
	-----***-----	12	/	5.33	/	0.56
partition:	-*-***-----	11	/	4.88	/	-0.26
	--*--*-----	42	/	18.66	/	-0.26
	-**-----	32	/	14.22	/	-0.26



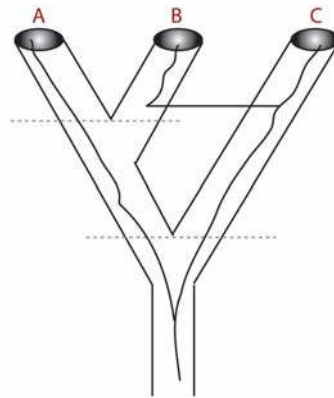
IC is a Much More Informative Measure of Internode Support



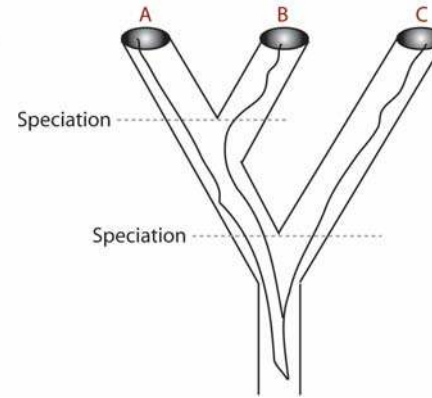
Why is the Yeast Phylogeny Hard to Resolve?

❖ Biological factors

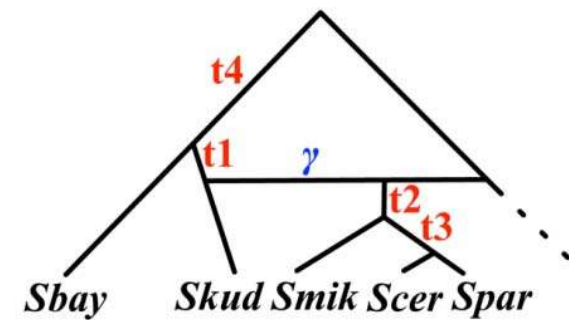
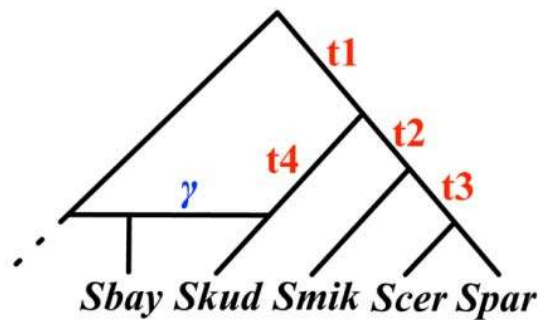
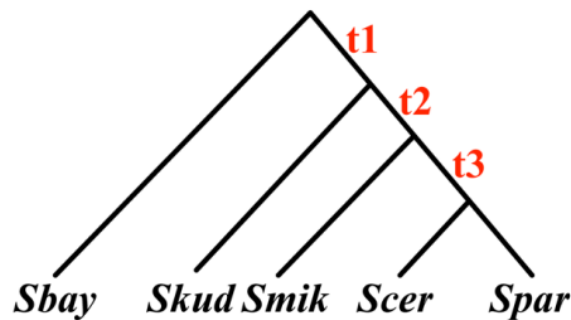
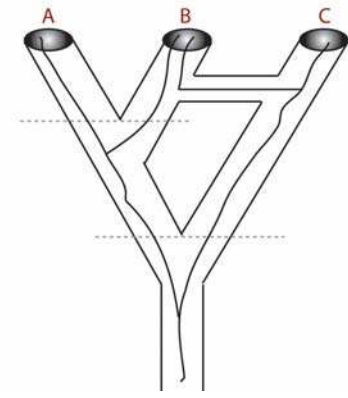
Horizontal Gene Transfer



Lineage Sorting

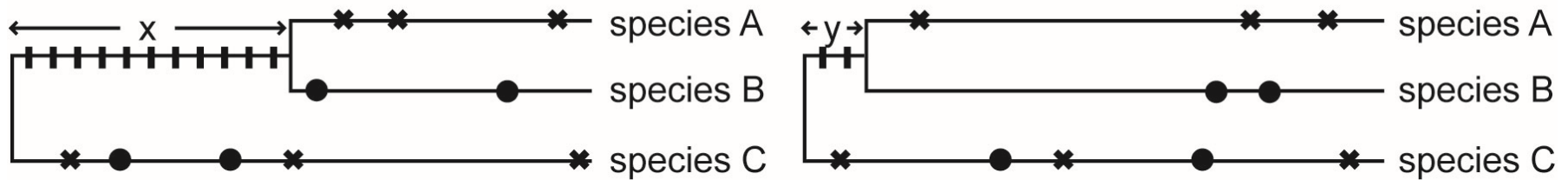


Hybridization



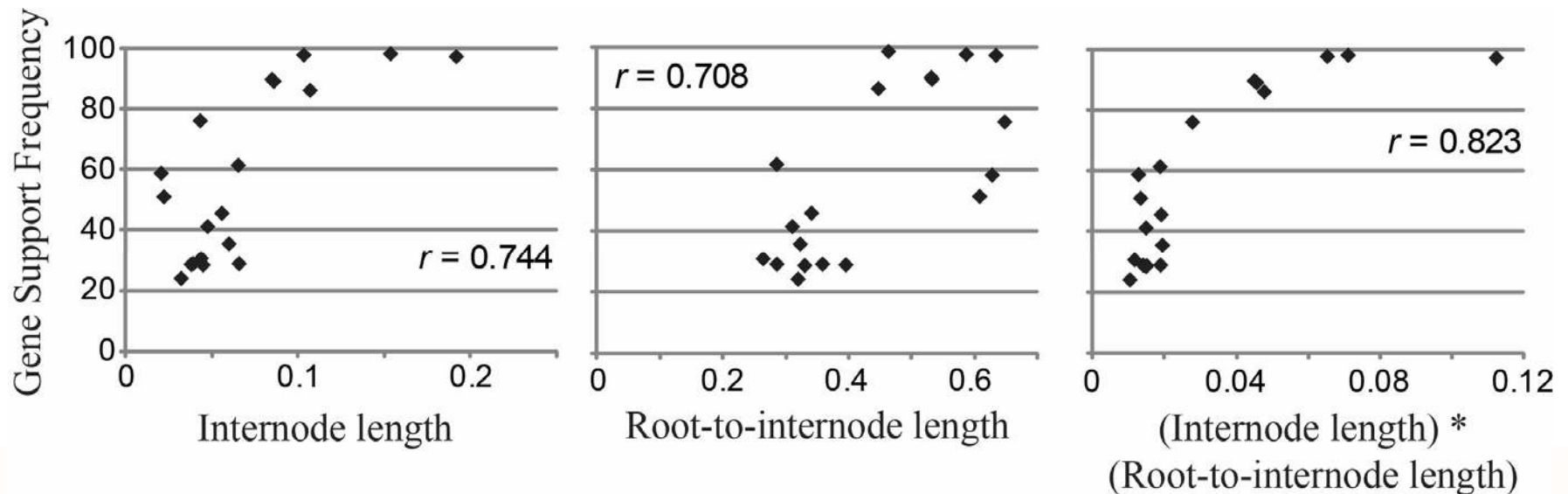
Why is the Yeast Phylogeny Hard to Resolve?

❖ 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

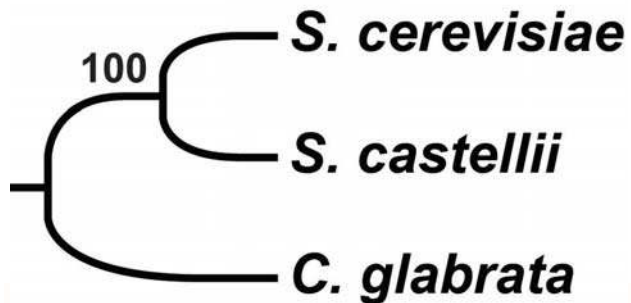




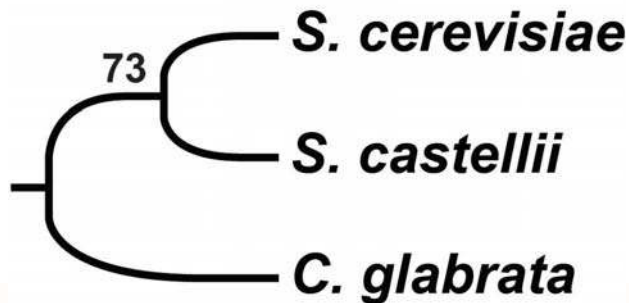
What Do We Do Then?

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

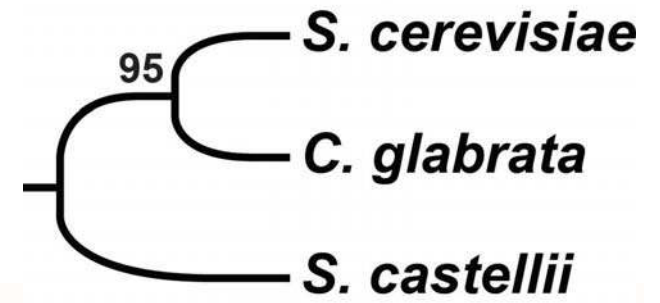
average BS $\geq 60\%$



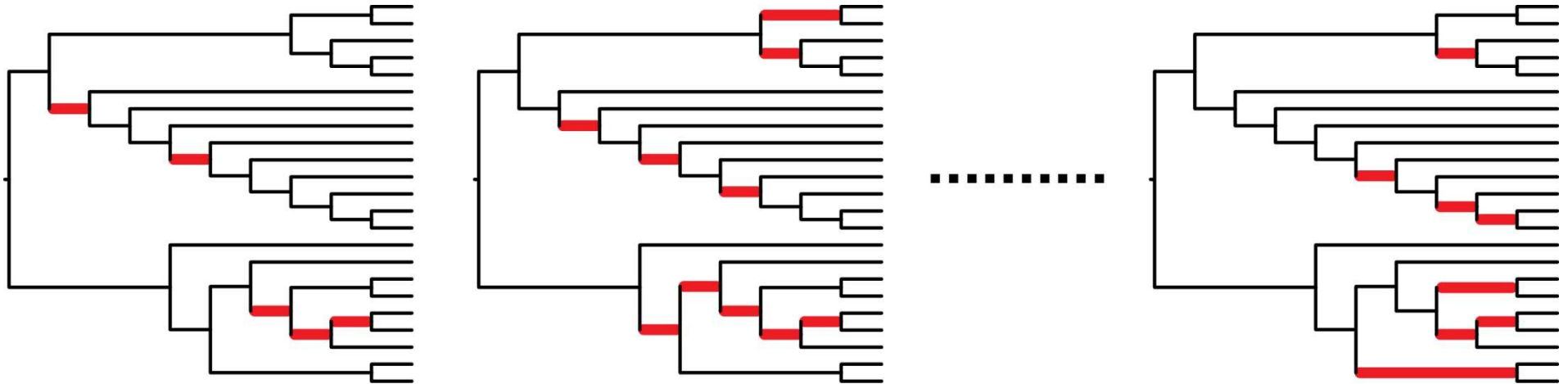
average BS $\geq 70\%$



average BS $\geq 80\%$



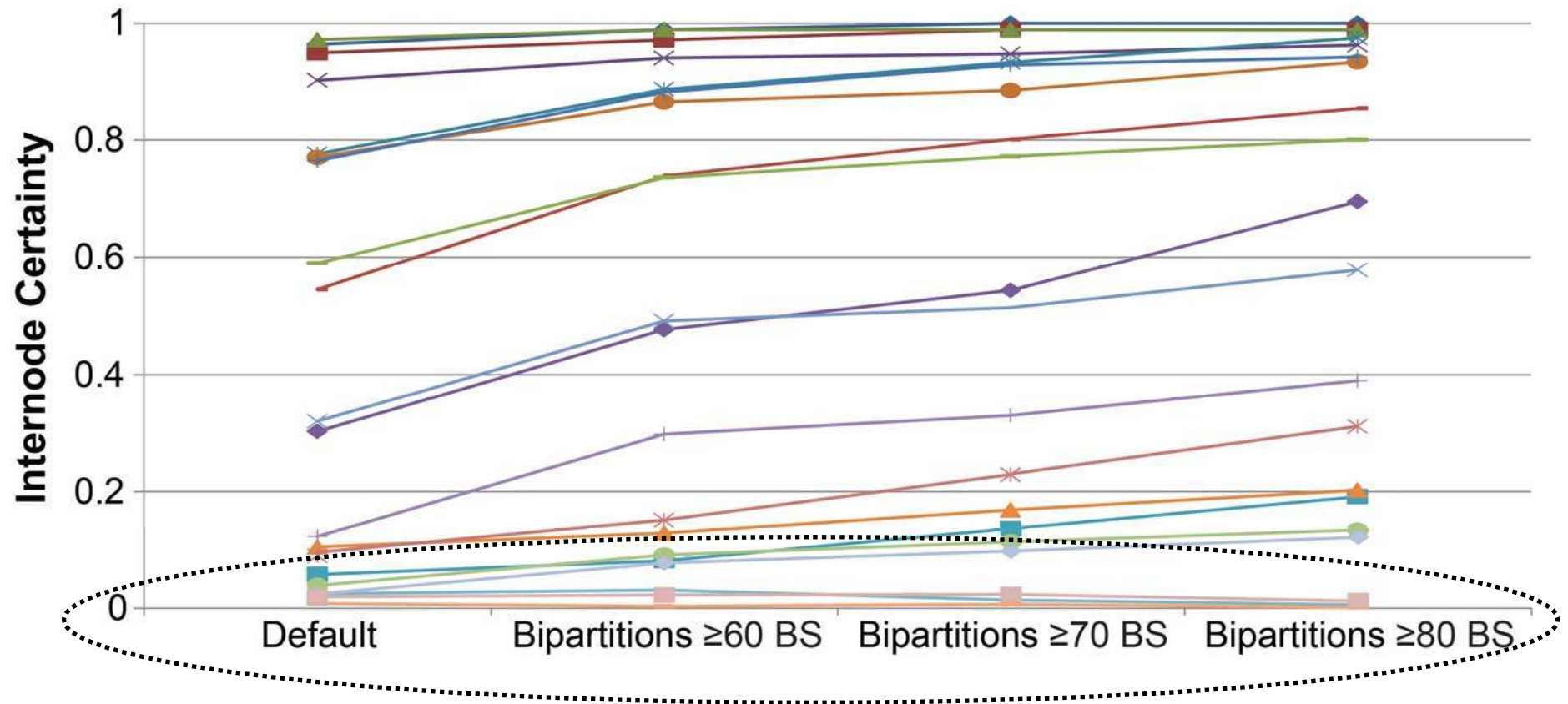
Selecting Specific Bipartitions Dramatically Improves Phylogeny



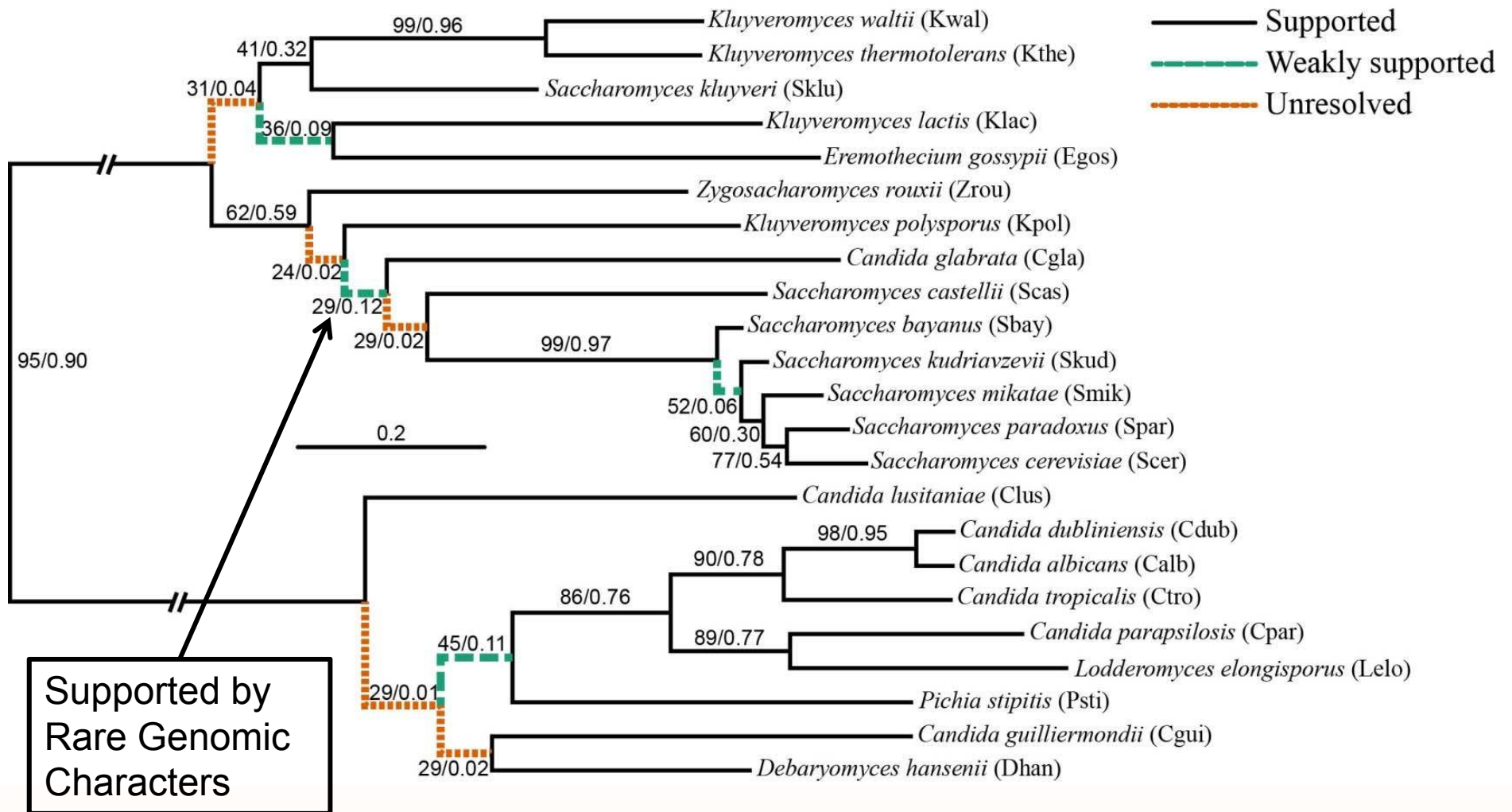
Treatment	Tree Certainty	# of Internodes where IC increased decreased
Default analysis	8.35	n/a
<i>Selecting genes whose bootstrap consensus trees have high average support</i>		
All bipartitions with BS $\geq$ 60%	10.11	14 0
All bipartitions with BS $\geq$ 70%	10.70	16 0
All bipartitions with BS $\geq$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



Least Supported Internodes Harbor the Most Conflict

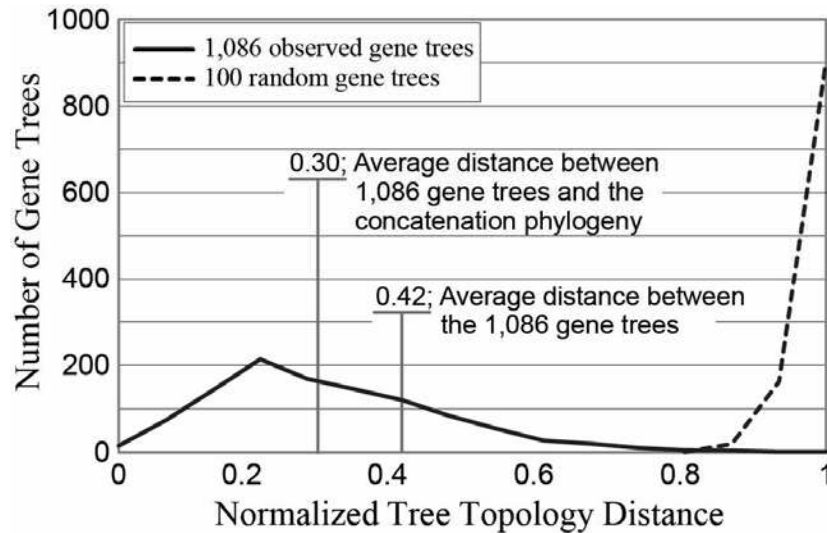


The Status of the Yeast Phylogeny

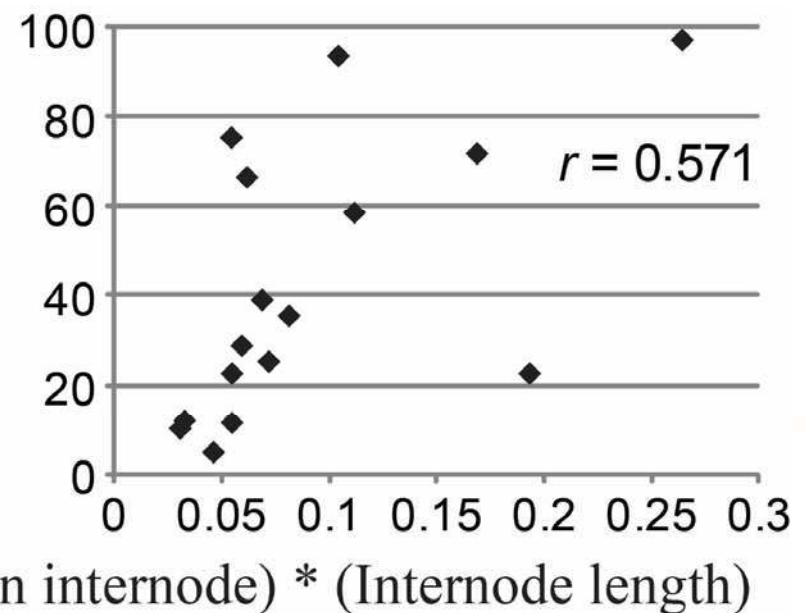
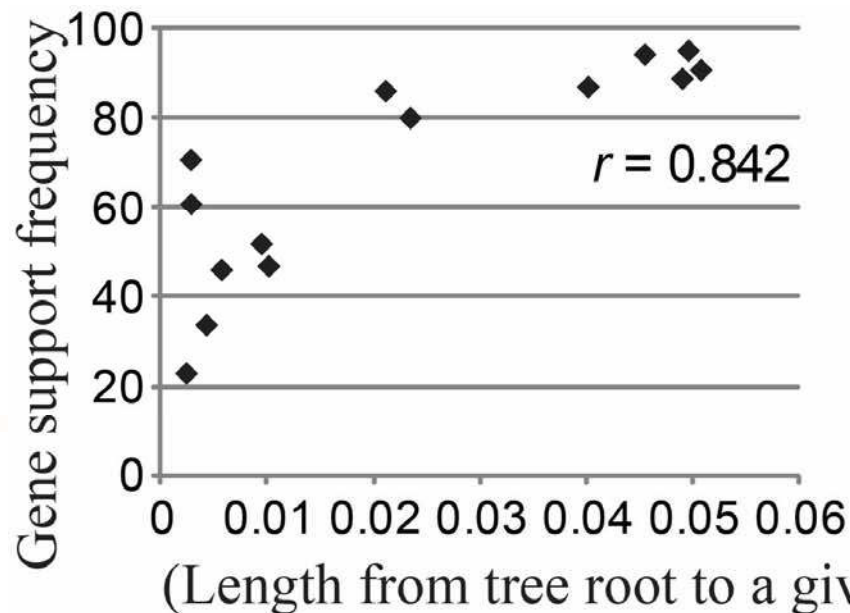
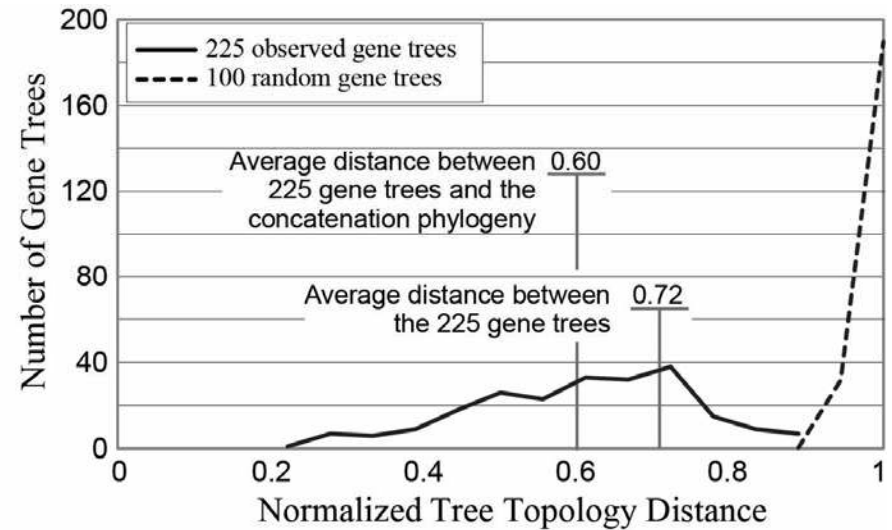


The Same is True for Vertebrate and Metazoan Datasets

Vertebrates (1,086 genes, 18 taxa)



Animals (225 genes, 21 taxa)

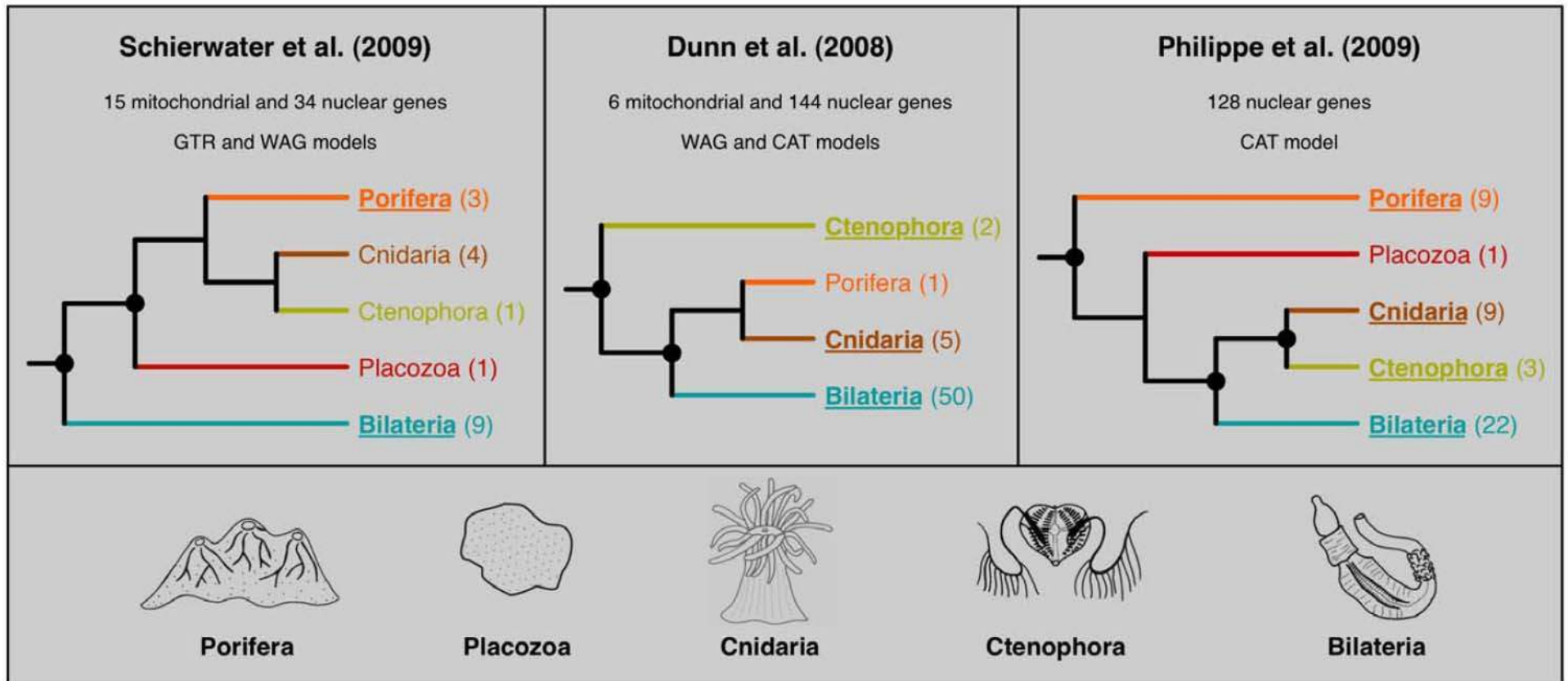


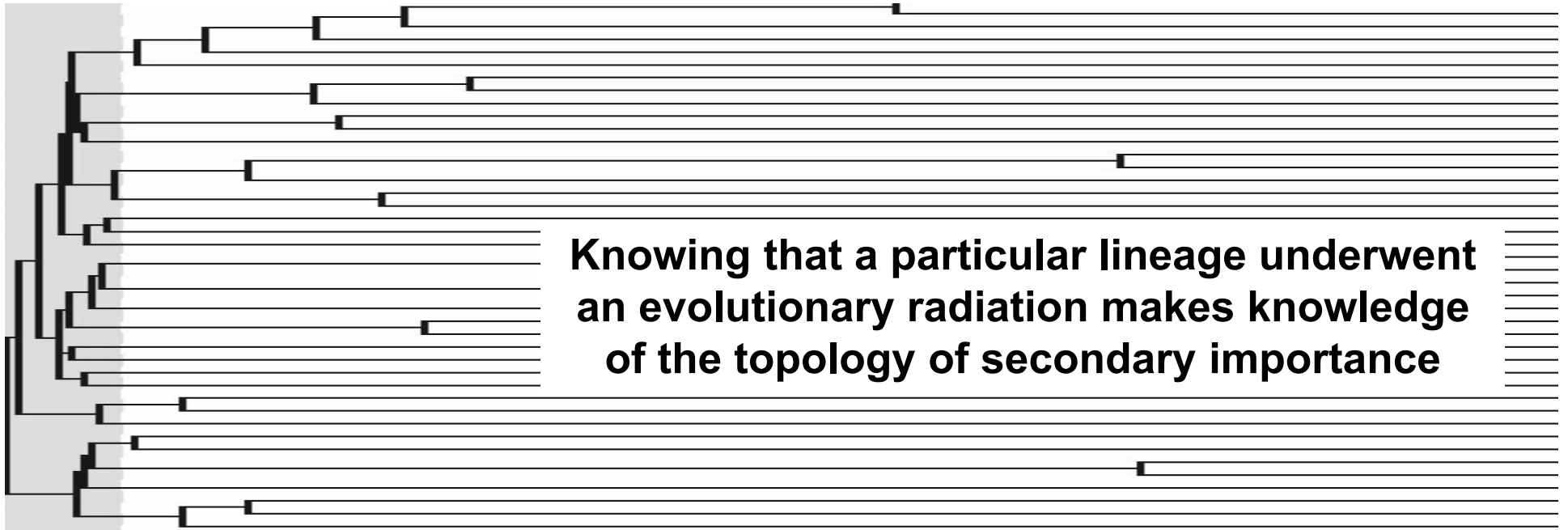
Lessons from Yeasts, Vertebrates & Animals

- ❖ **Few, if any, of the Gene Trees are Topologically Identical to Each Other or to the Phylogeny Inferred by Concatenation**
- ❖ **Concatenation analysis Overconfident and Can Mislead**
- ❖ **Internode Support is Inversely Correlated with Internode Length and Depth**
- ❖ **Selecting Genes or Gene Tree Bipartitions with Strong Signal Reduces Incongruence**



Incongruence in Deep Time





“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



Rokas & Carroll (2006) PLoS Biol.

Acknowledgements



John Gibbons



Chris Hittinger



Leonidas Salichos



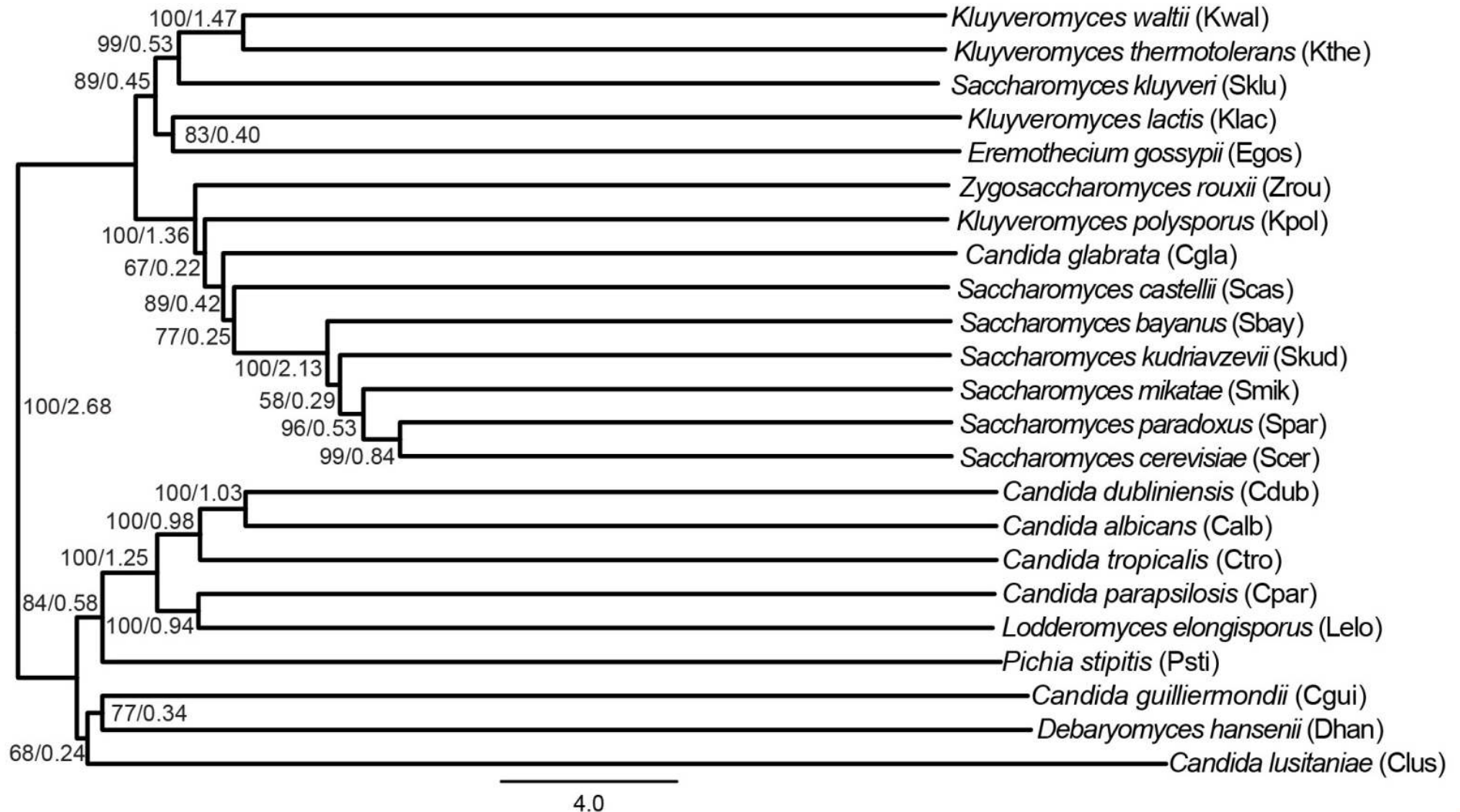
National Science Foundation
WHERE DISCOVERIES BEGIN

SEARLE SCHOLARS PROGRAM



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

The Yeast Phylogeny Using a Species Tree Method



Bootstrap support / Internode length in coalescent units

STAR Internode Lengths Correlate with GSF / IC

