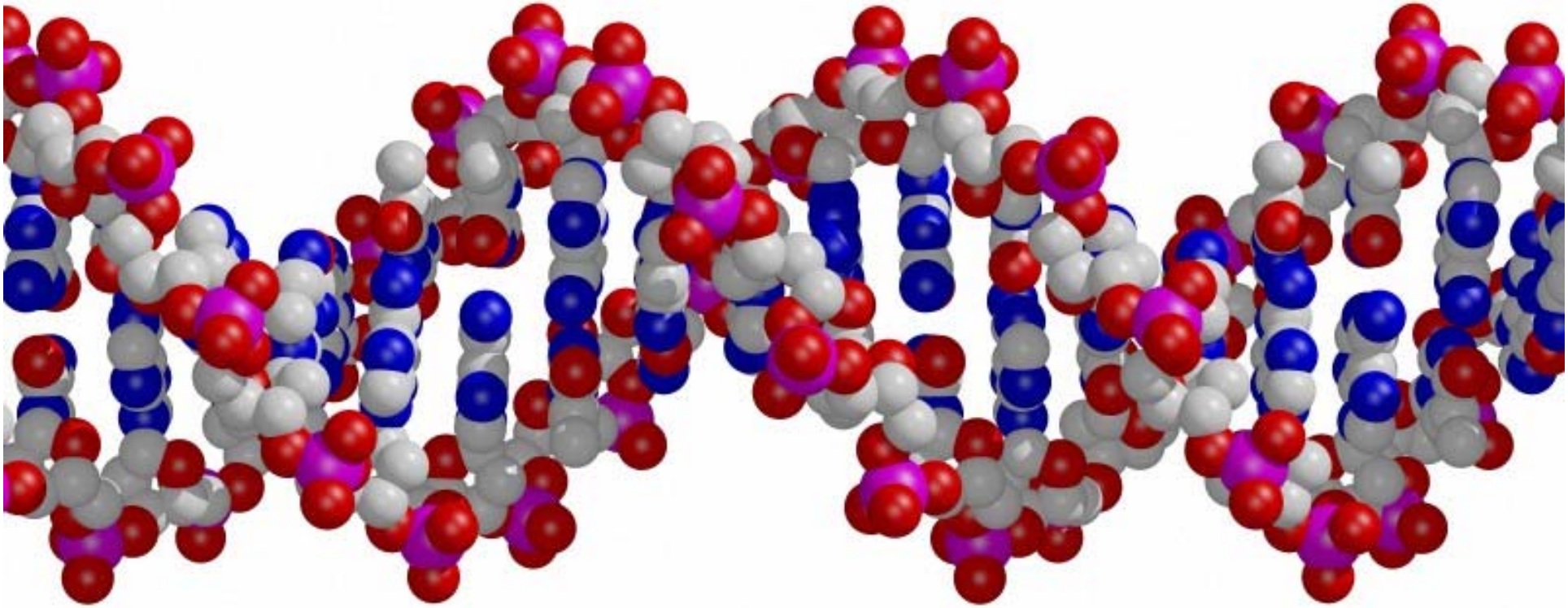
Evolutionary Genomics



Antonis Rokas
Department of Biological Sciences
Vanderbilt University

http://as.vanderbilt.edu/rokaslab

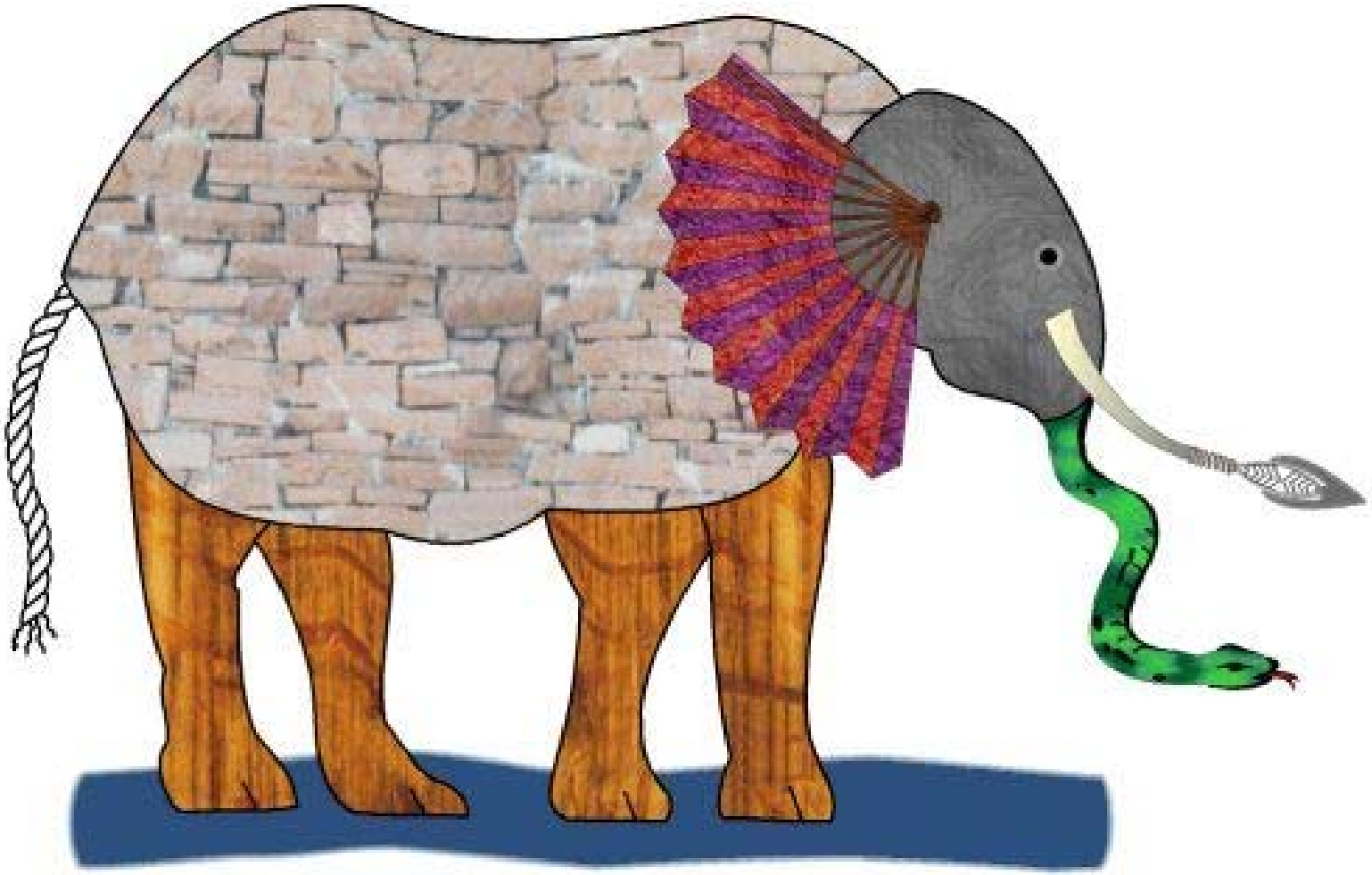
Lecture Outline

- **Introduction to Evolutionary Genomics**
- **Population Genomics**

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

- **Phylogenomics**

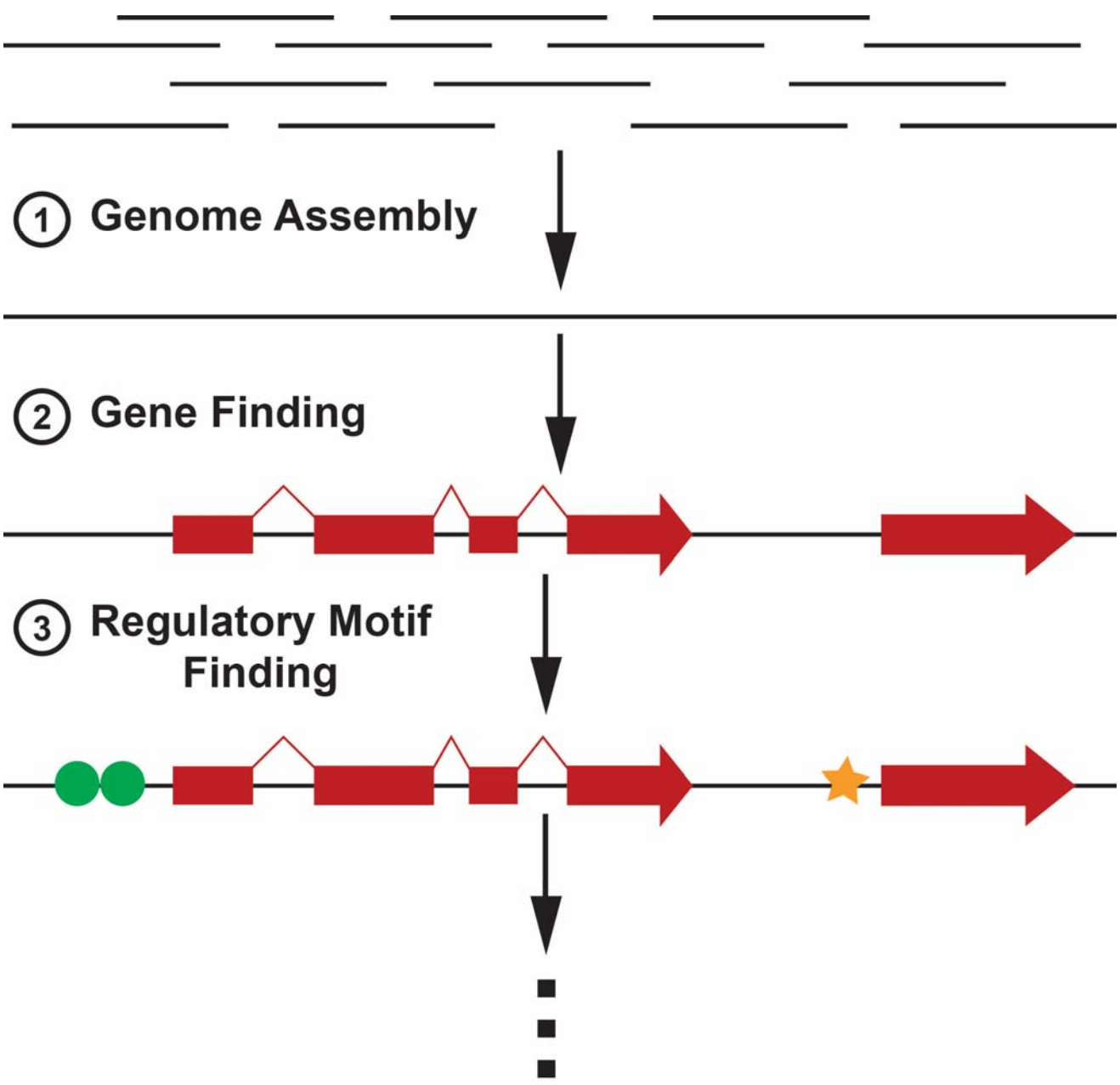
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
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

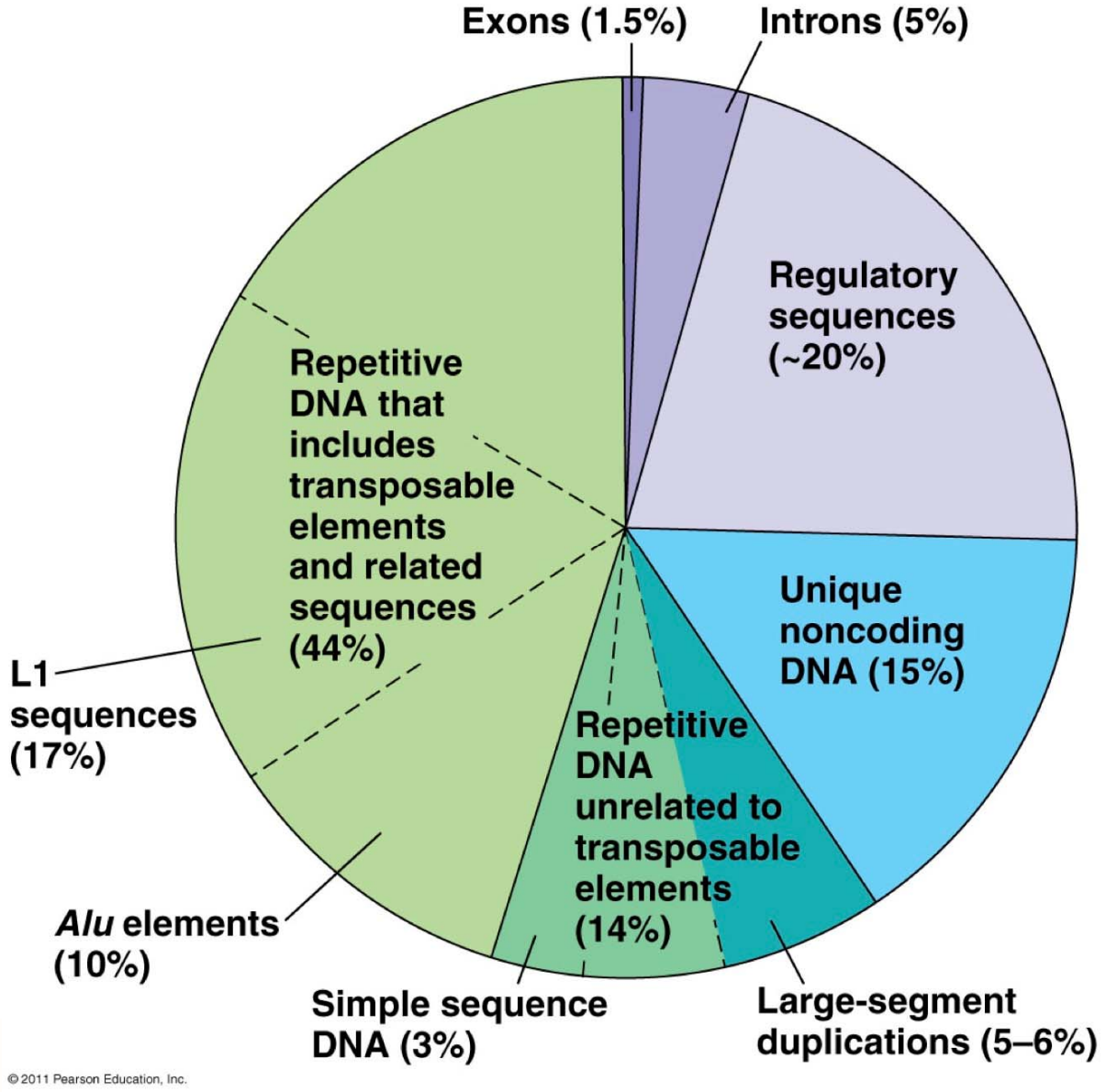


What is a Genome Like?

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

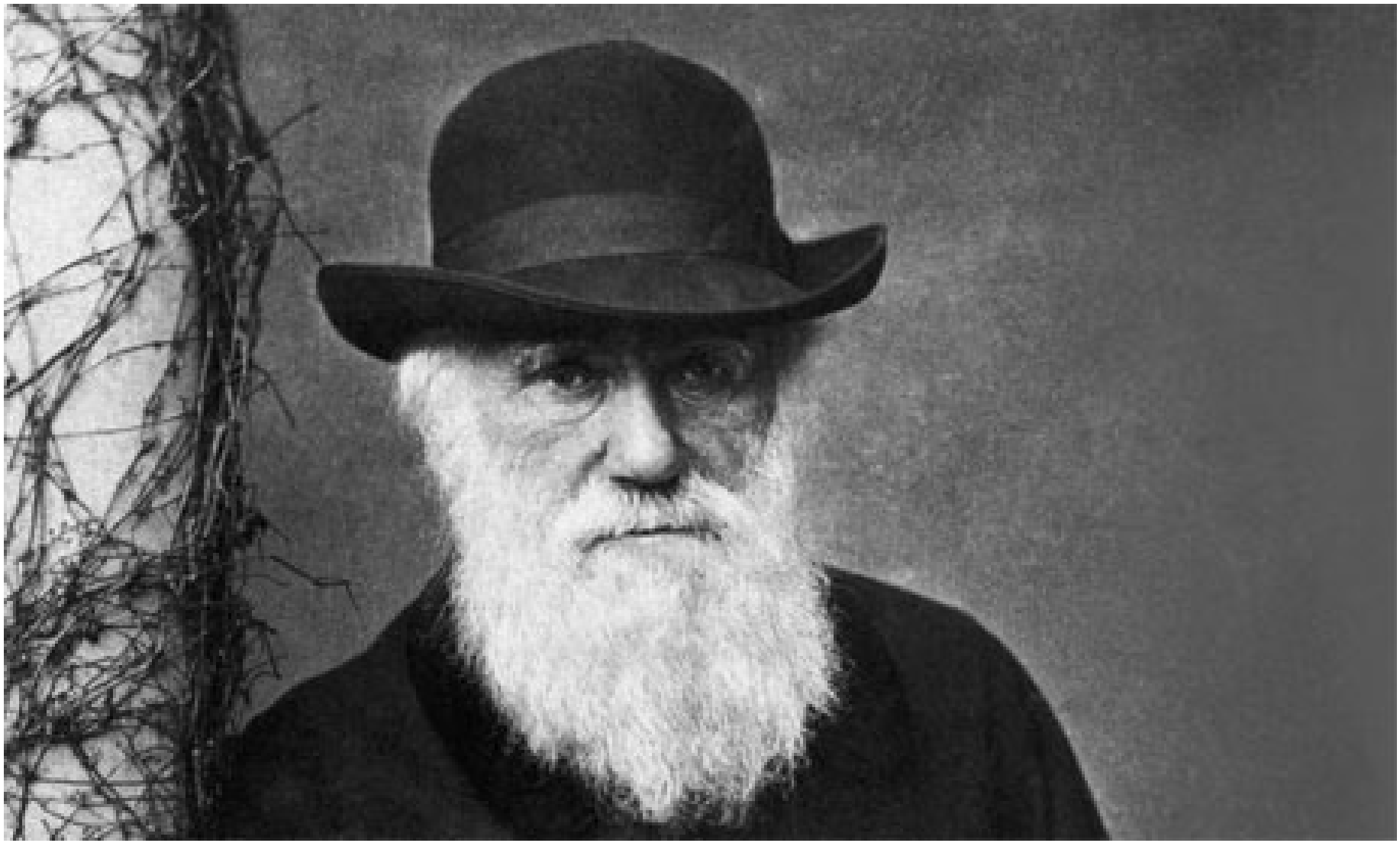
Transposon **Protein Binding Site** **Exon** **Intron**

Organization of the Human Genome

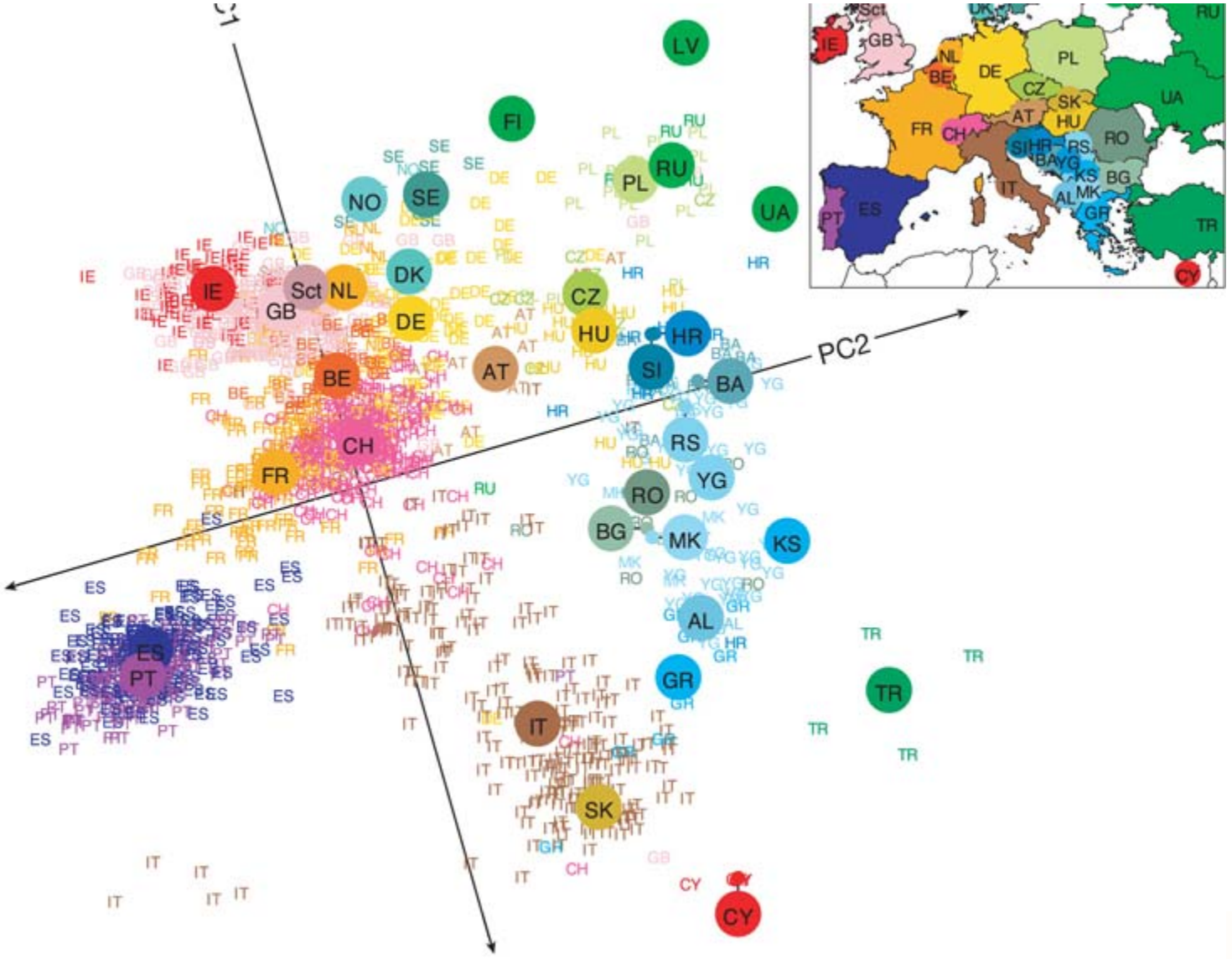


http://todd.jackman.villanova.edu/HumanGenome.jpg

Understanding the Genome Requires a Theory



Human Genes Mirror Geography



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	

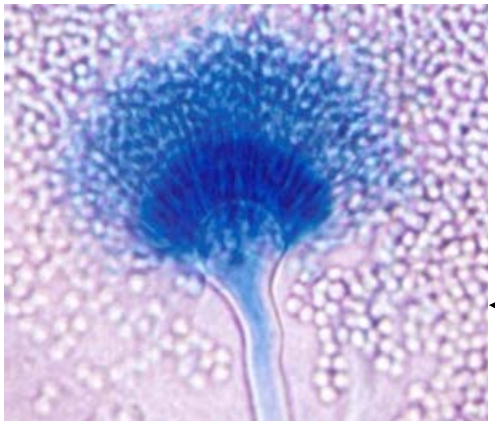
In the European population, involved in skin pigmentation

In the West African population, related to Lassa virus infection

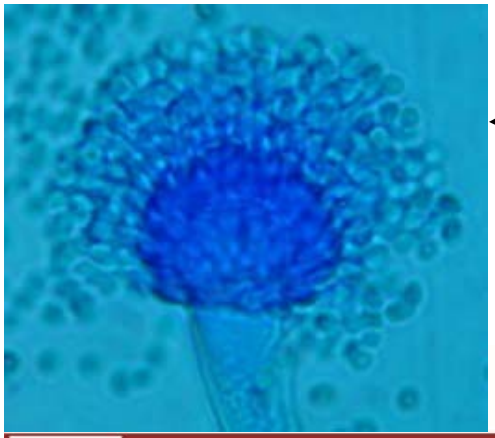
Sabeti et al. (2007) Nature

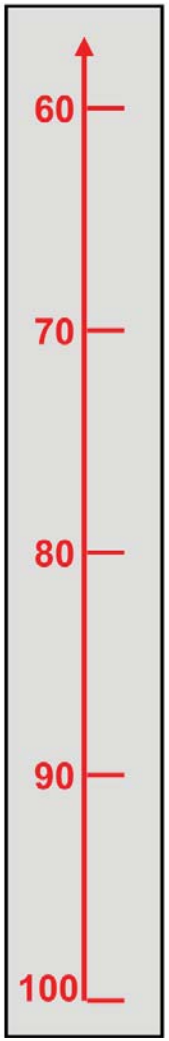
Genomes Provide a Common Yardstick for Comparison

Average proteome sequence similarity



Morphologically very similar!





A. nidulans/niger

A. oryzae/terreus

A. clavatus

N. fischeri

A. fumigatus

Kluyveromyces lactis

Candida glabrata

fish

birds

S. uvarum

They Interbreed!

S. paradoxus

S. cerevisiae

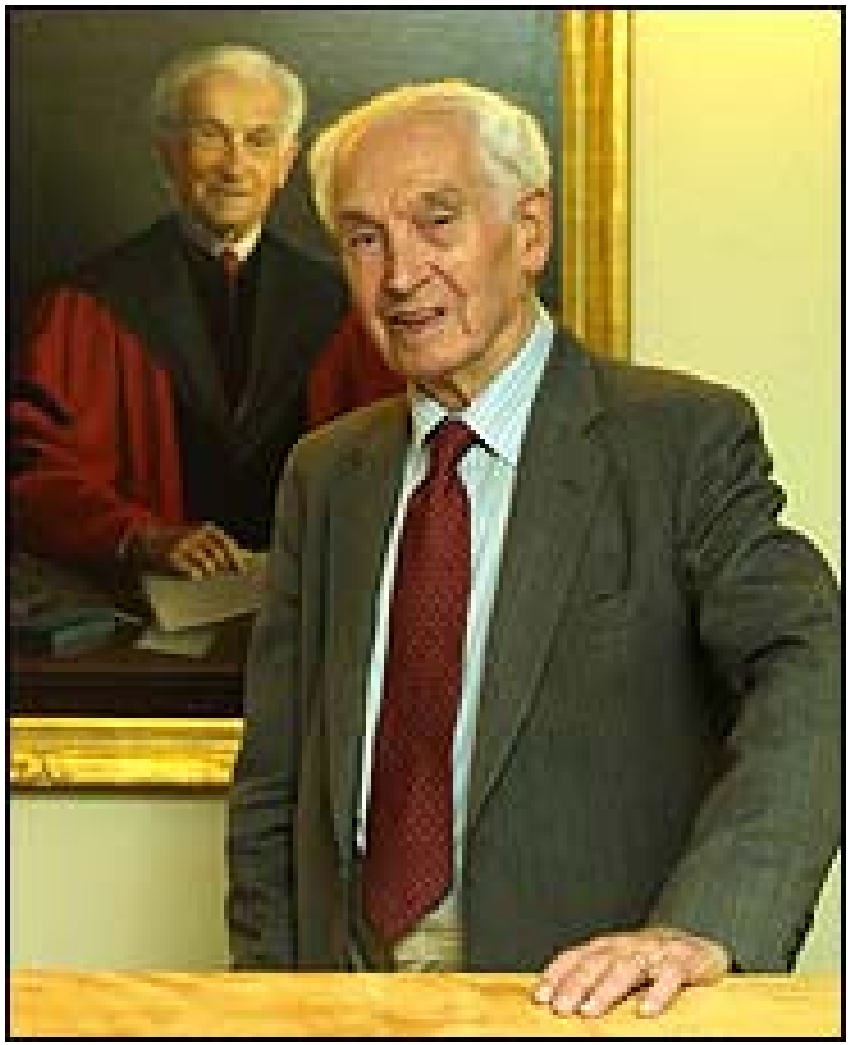
mice

humans

Fedorova et al. (2008) PLoS Genet.

“…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

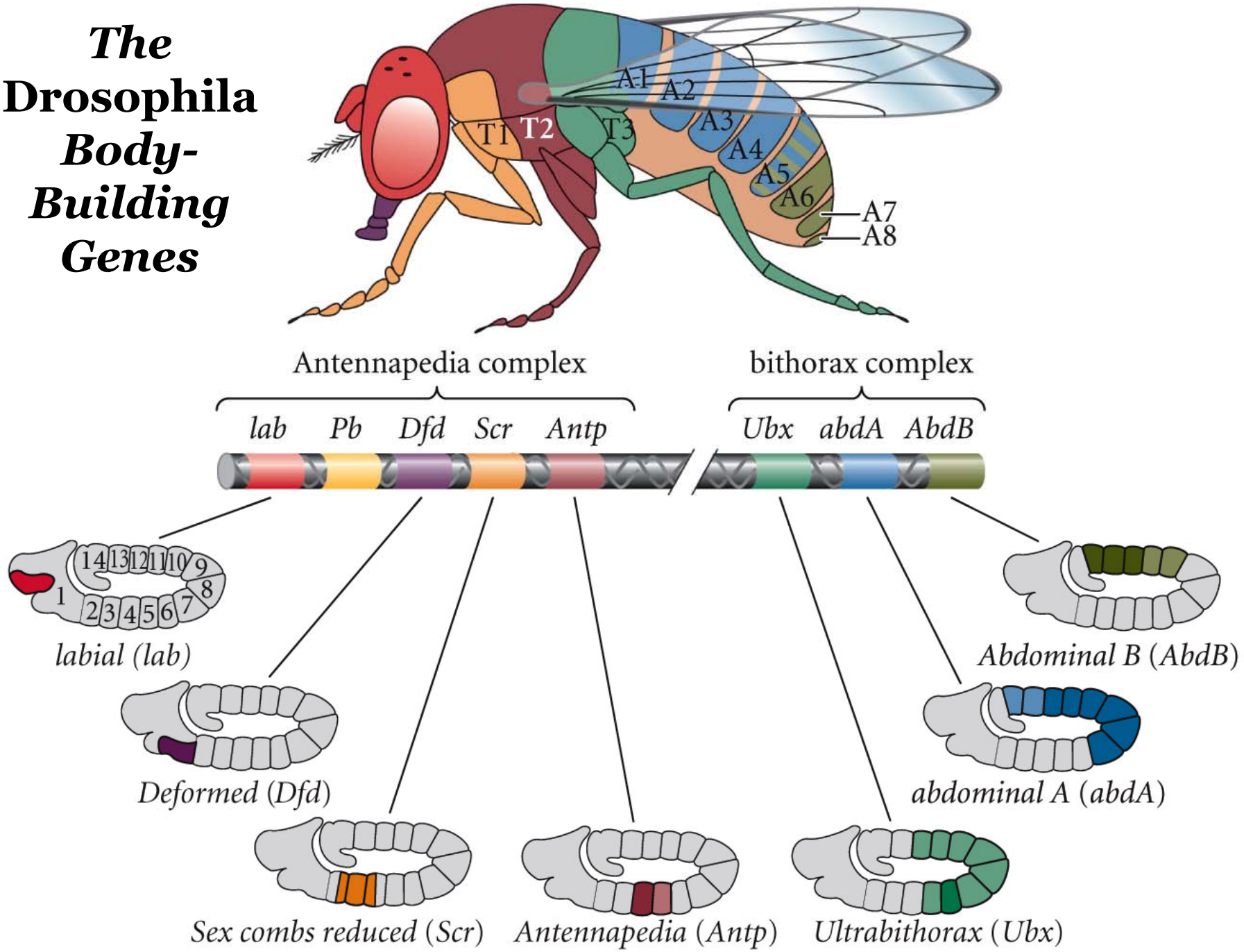
F	W	Y	Cancer
+	+	-	ABL1
+	-	-	Acute Myeloid Leukemia-DEK
+	-	-	Adenomat. Polyposis Coli-APC
+	+	+	AKT2
+	+	+	Ataxia Telangiectasia-ATM
-	-	-	BRCA1
-	-	-	BRCA2
+	+	-	Basal Cell Nevus-PTC
+	-	-	B-Cell Lymphoma 2-BCL2
-	-	-	B-Cell Lymphoma 3-BCL3
+	+	+	Bloom-BLM
+	-	-	Burkitt's Lymphoma-MYC
-	-	-	CDKN2C
-	+	-	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

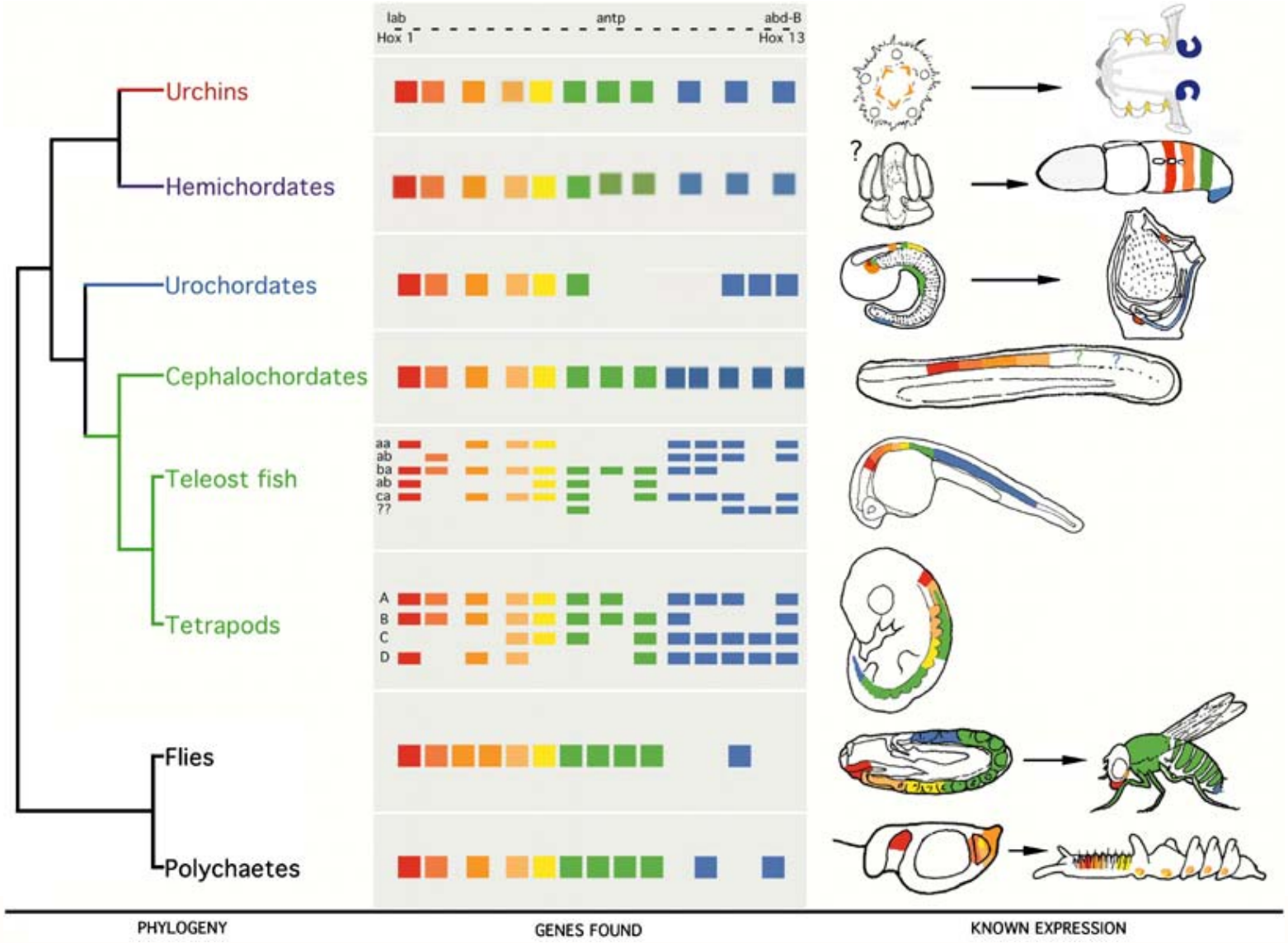
Human disease-associated genes shared with flies (F), worms (W), and Yeast (Y); from Rubin et al. (2000) Science

The Drosophila *Body-Building Genes*

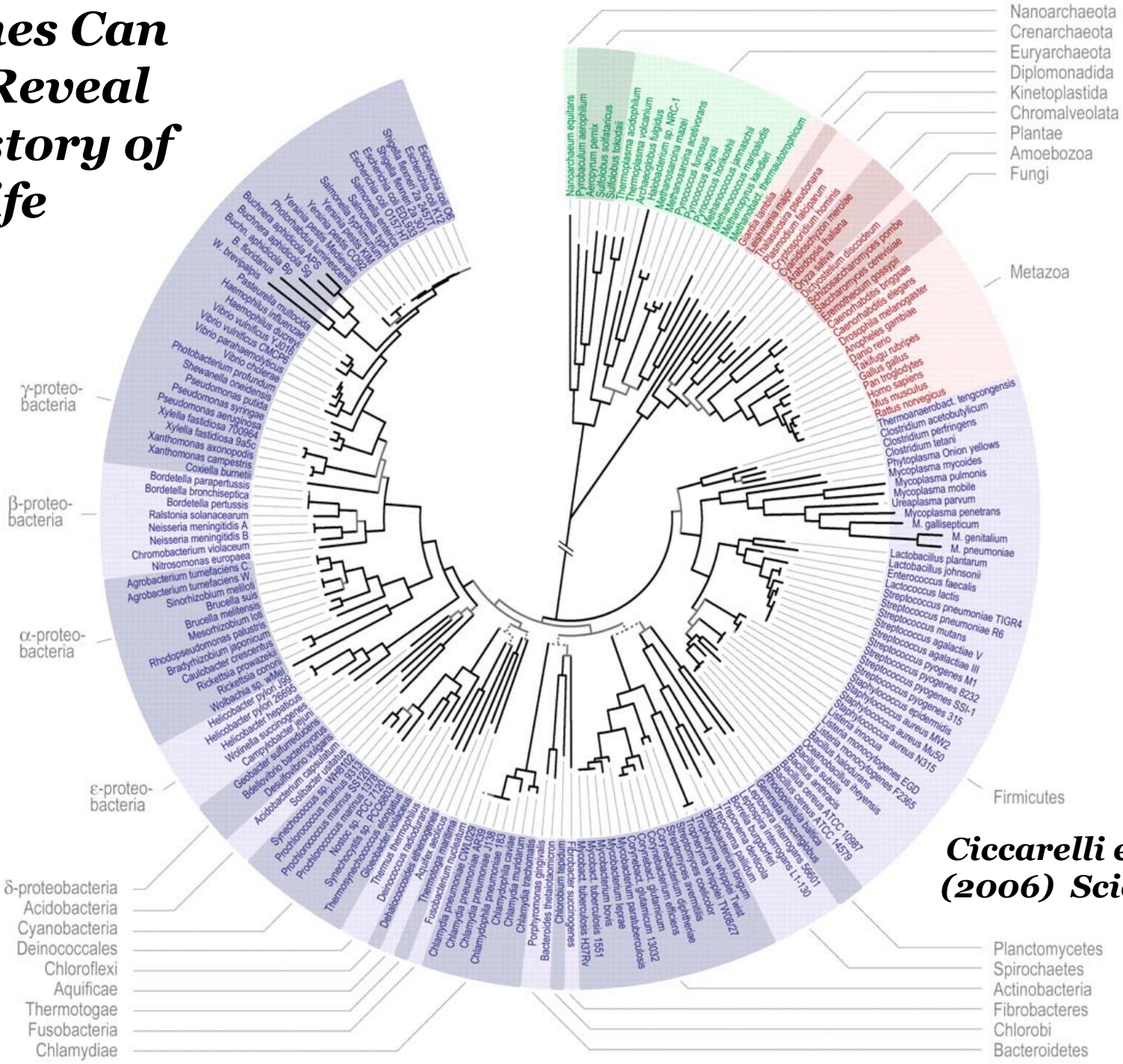


EVOLUTION 2e, Figure 21.2

Animal Bodies are Built from the Same Genetic Toolkit

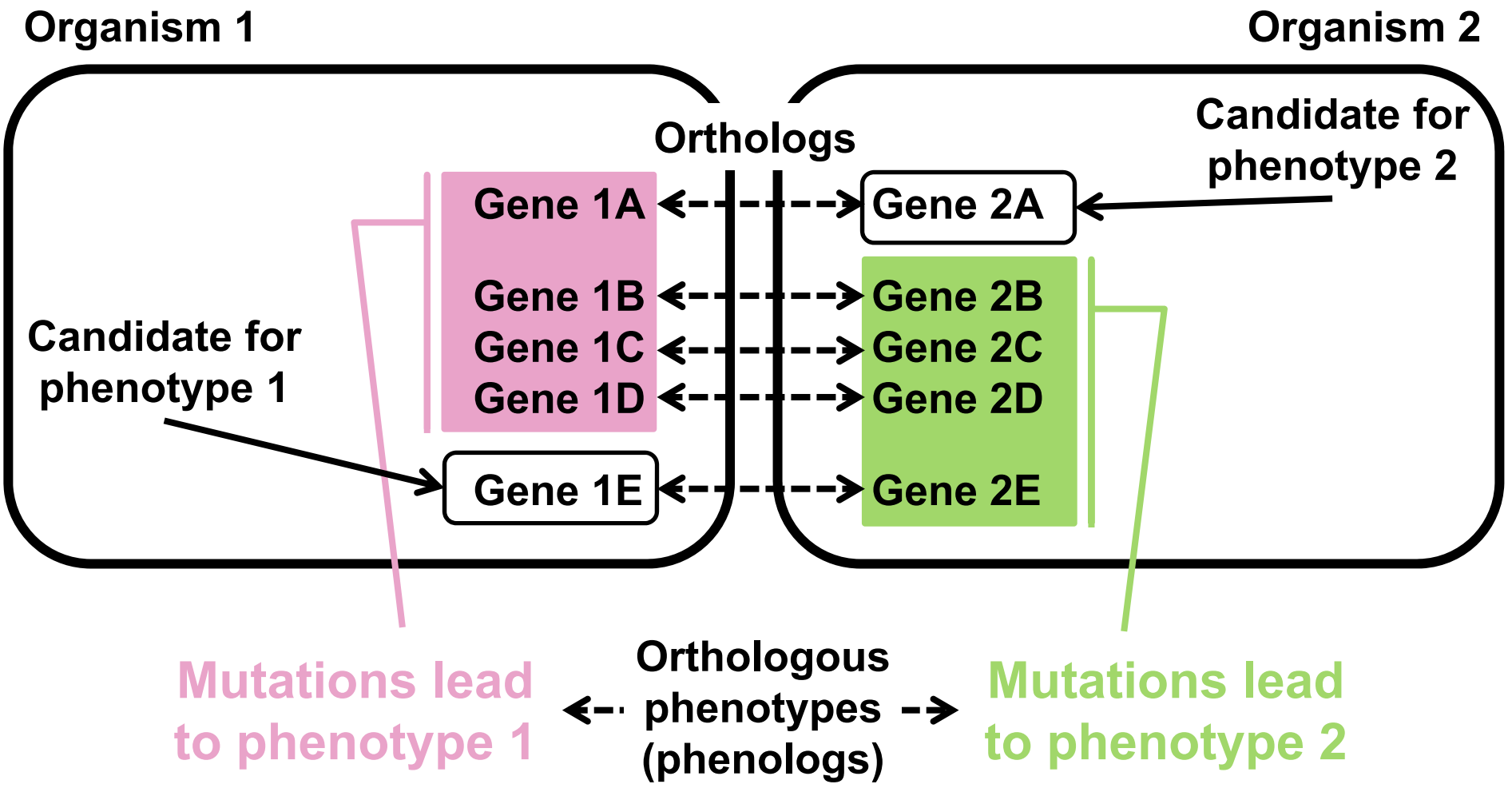


Genomes Can Help Reveal the History of Life



Ciccarelli et al. (2006) Science

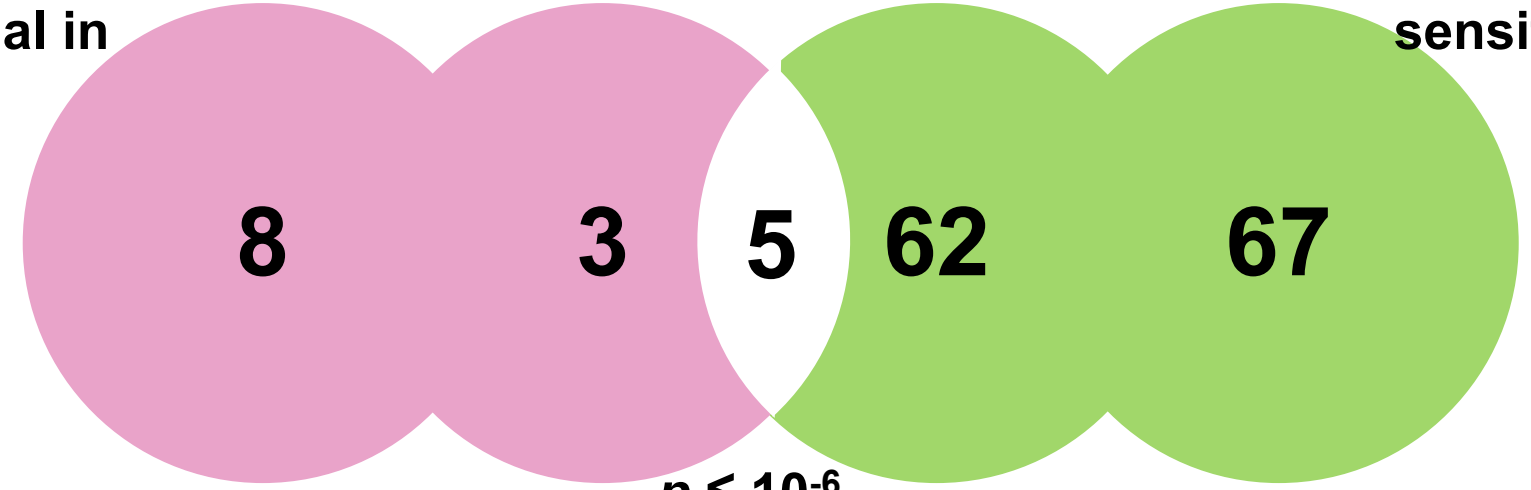
Evolution-Informed Analyses Have Great Predictive Power



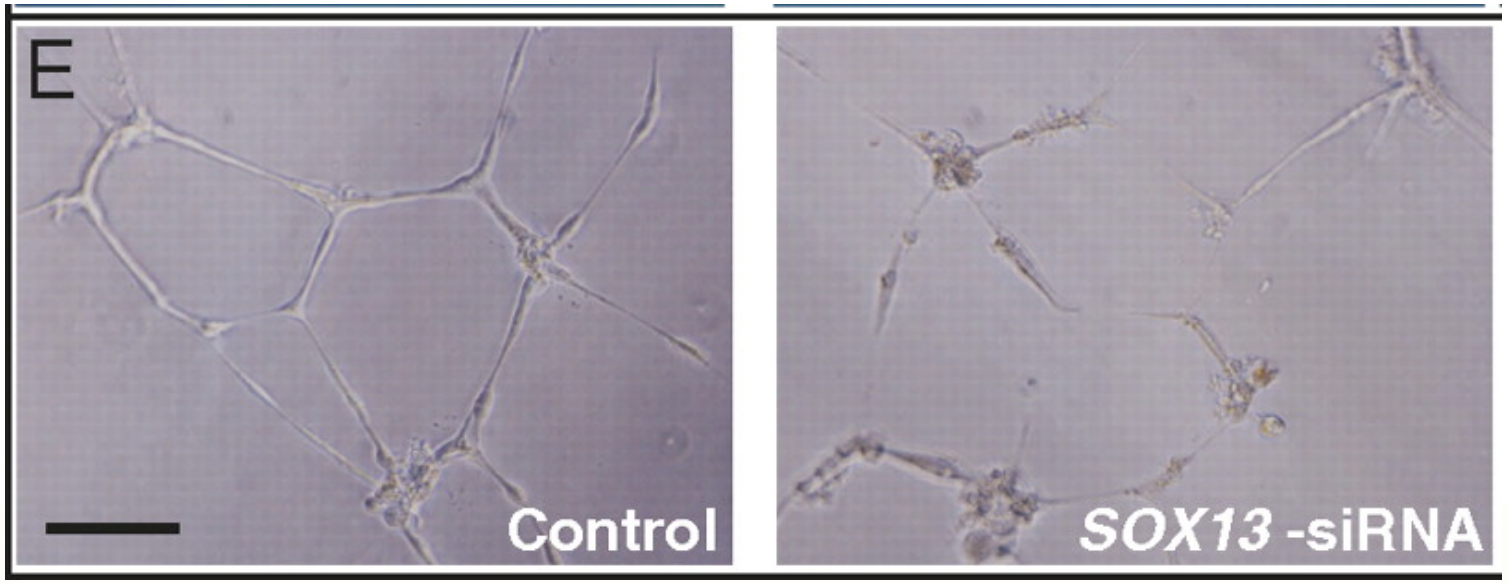
A Yeast Model for Angiogenesis

Angiogenesis abnormal in mice

Lovastatin sensitive in yeast



$p \leq 10^{-6}$



Genomics: "Big Science" Driven by a Few Centers

wellcome trust sanger institute

BCM
Baylor College of Medicine

BROAD INSTITUTE

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

J. Craig Venter INSTITUTE

THE Genome CENTER
AT WASHINGTON UNIVERSITY

High-Throughput DNA Sequencing Technologies

454 / Roche

450 bp 1.5 Gbp / day

EXTINCT BY 2016

Illumina

150 bp 35 Gbp / day

Helicos

55 bp 4.5 Gb / day

EXTINCT

SOLiD ABI

75 bp 22 Gbp / day

PacBio

1000 bp 70 Gbp / day

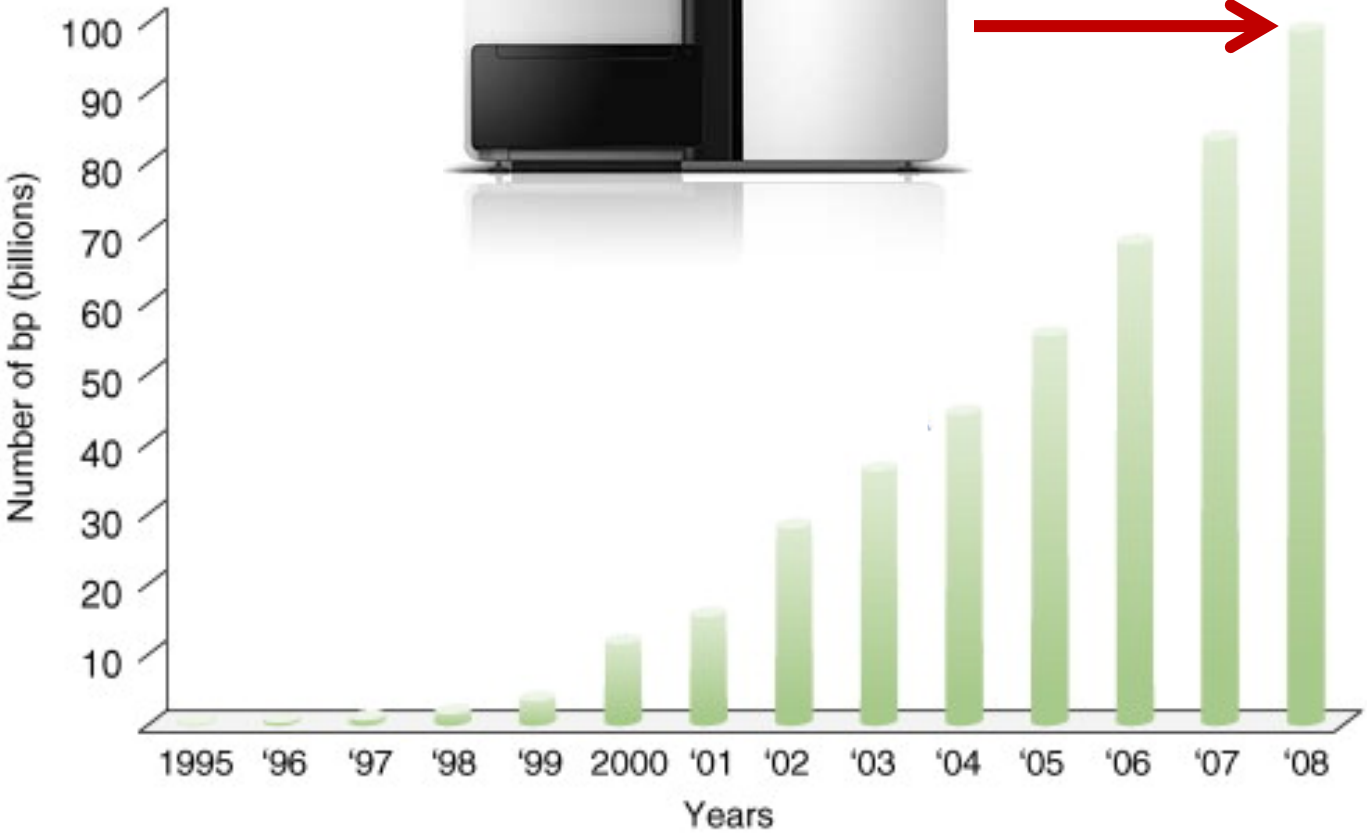
SOLD

Ion PGM

100 bp 120 Gbp / day

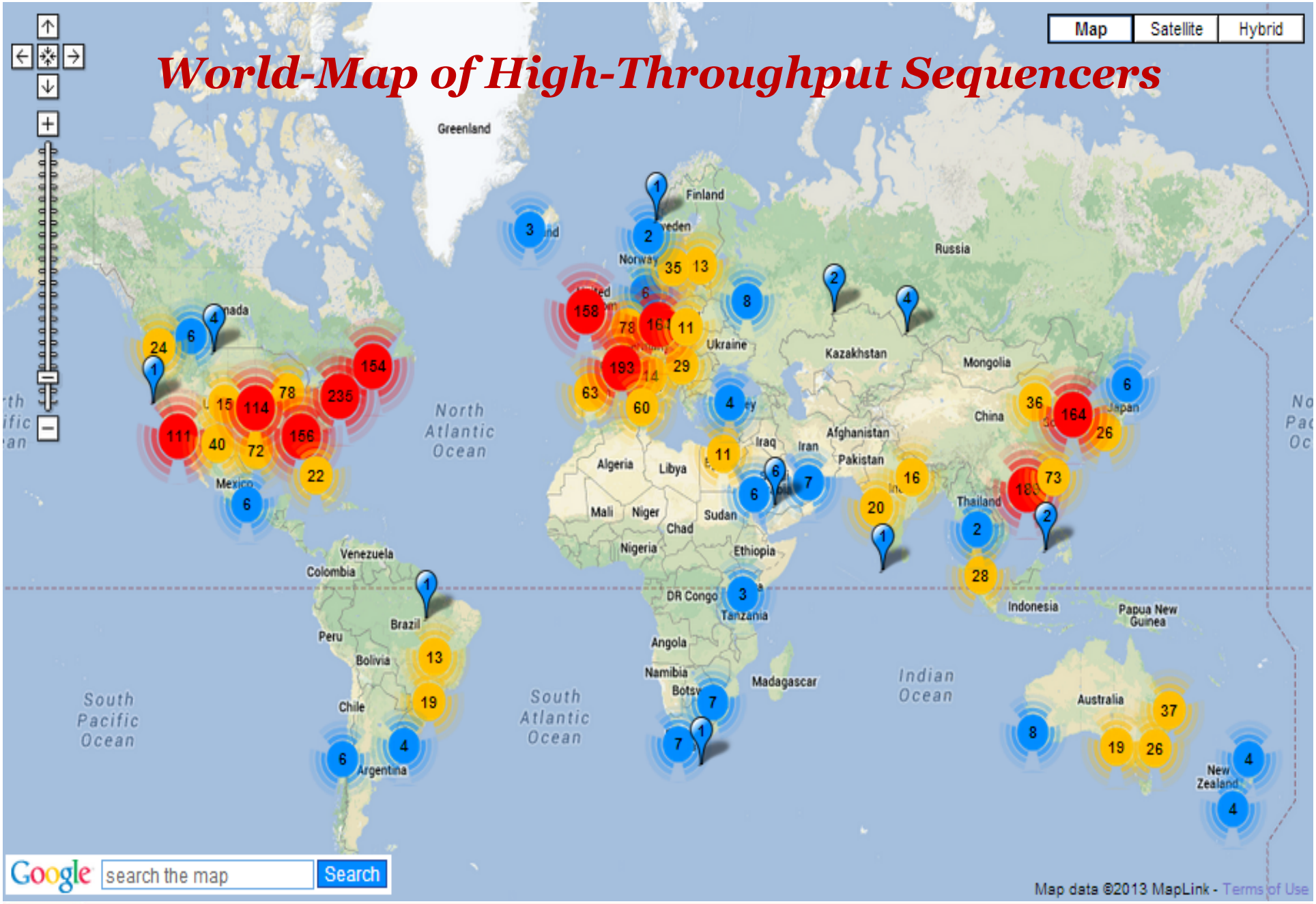
Why Is High-Throughput DNA Sequencing So Exciting?

3 days' output

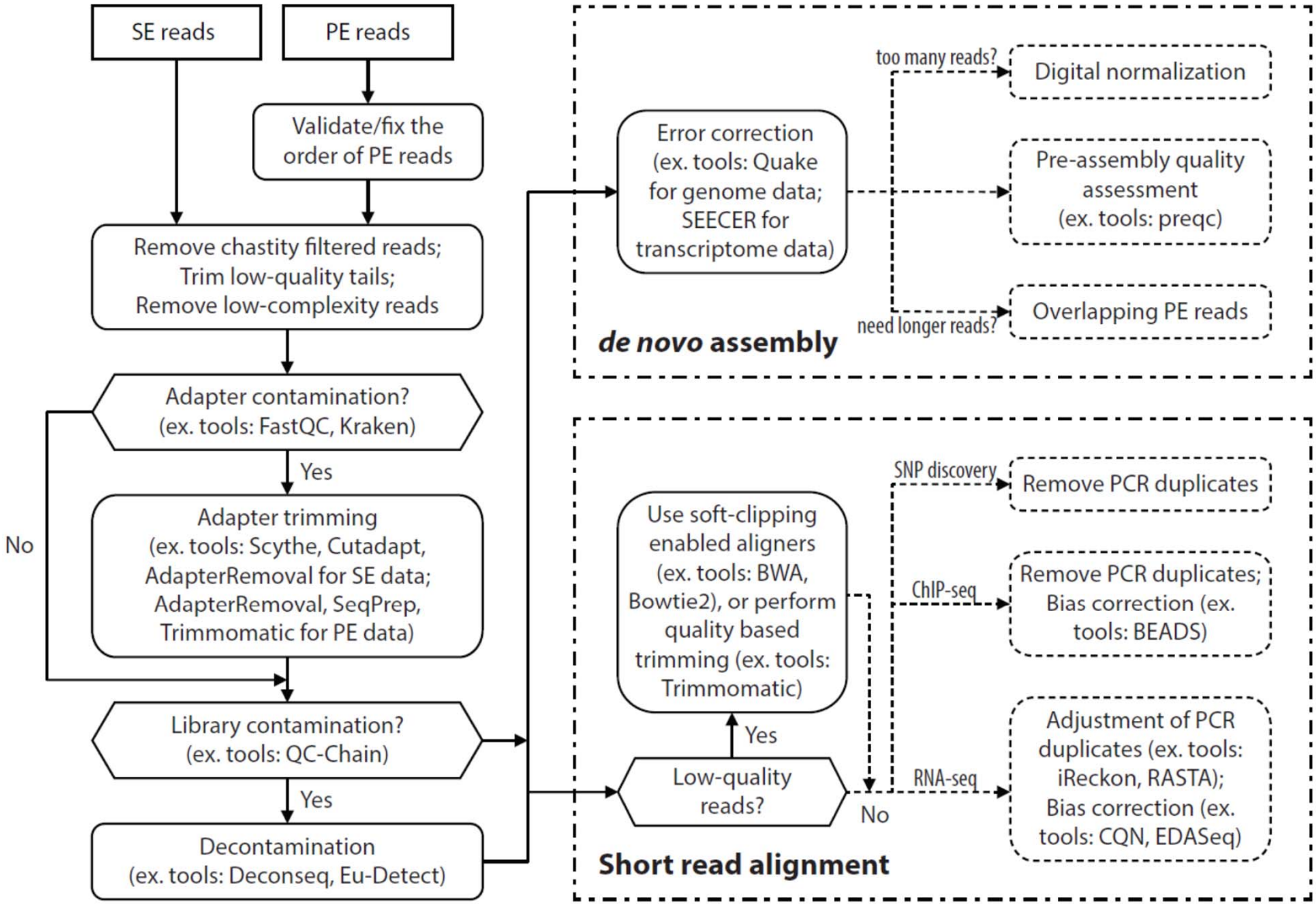


Gilad et al. (2009) Trends Genet.

World-Map of High-Throughput Sequencers

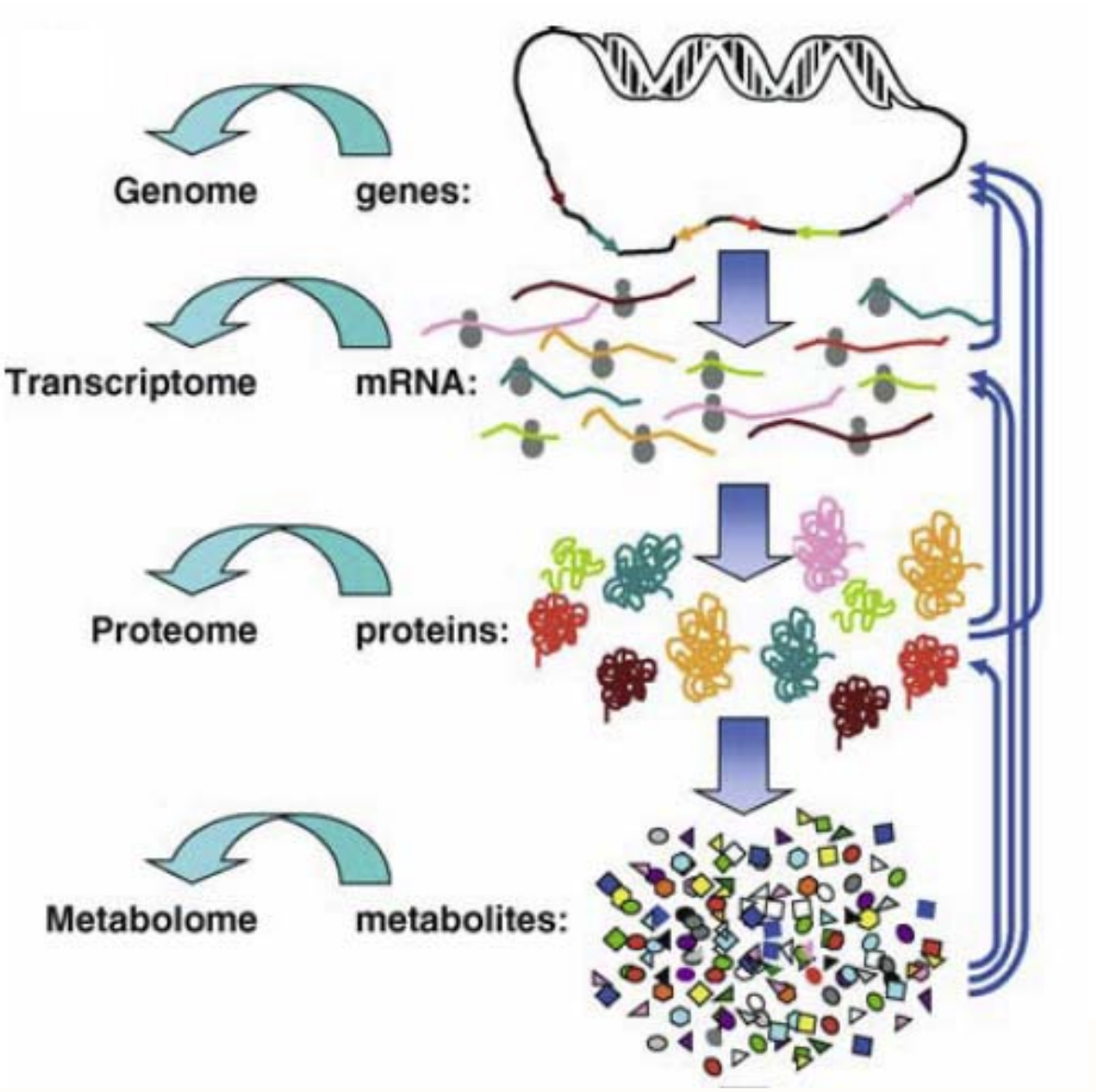


High Throughput Sequencing Data Pathologies



Zhou & Rokas (2014) Molecular Ecology

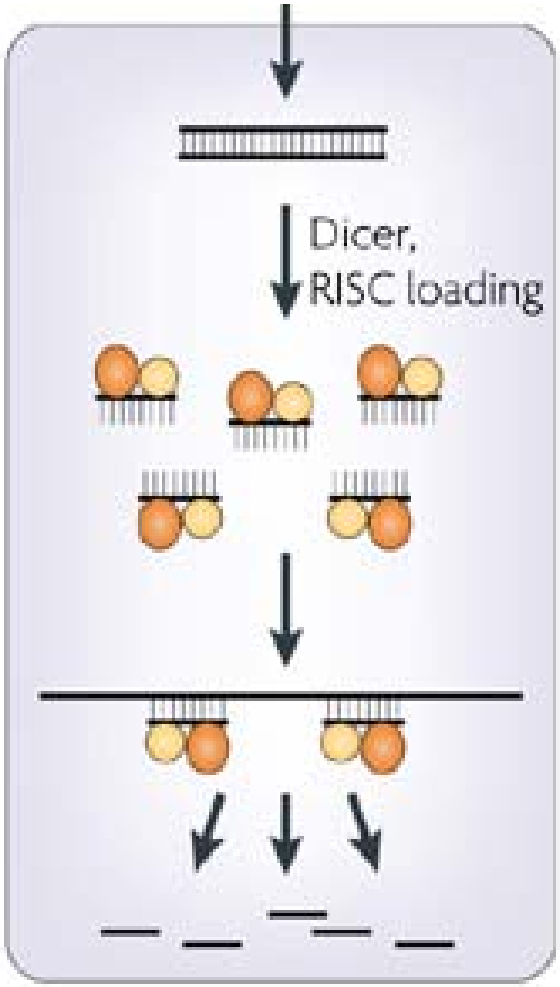
The Age of High Throughput Technologies



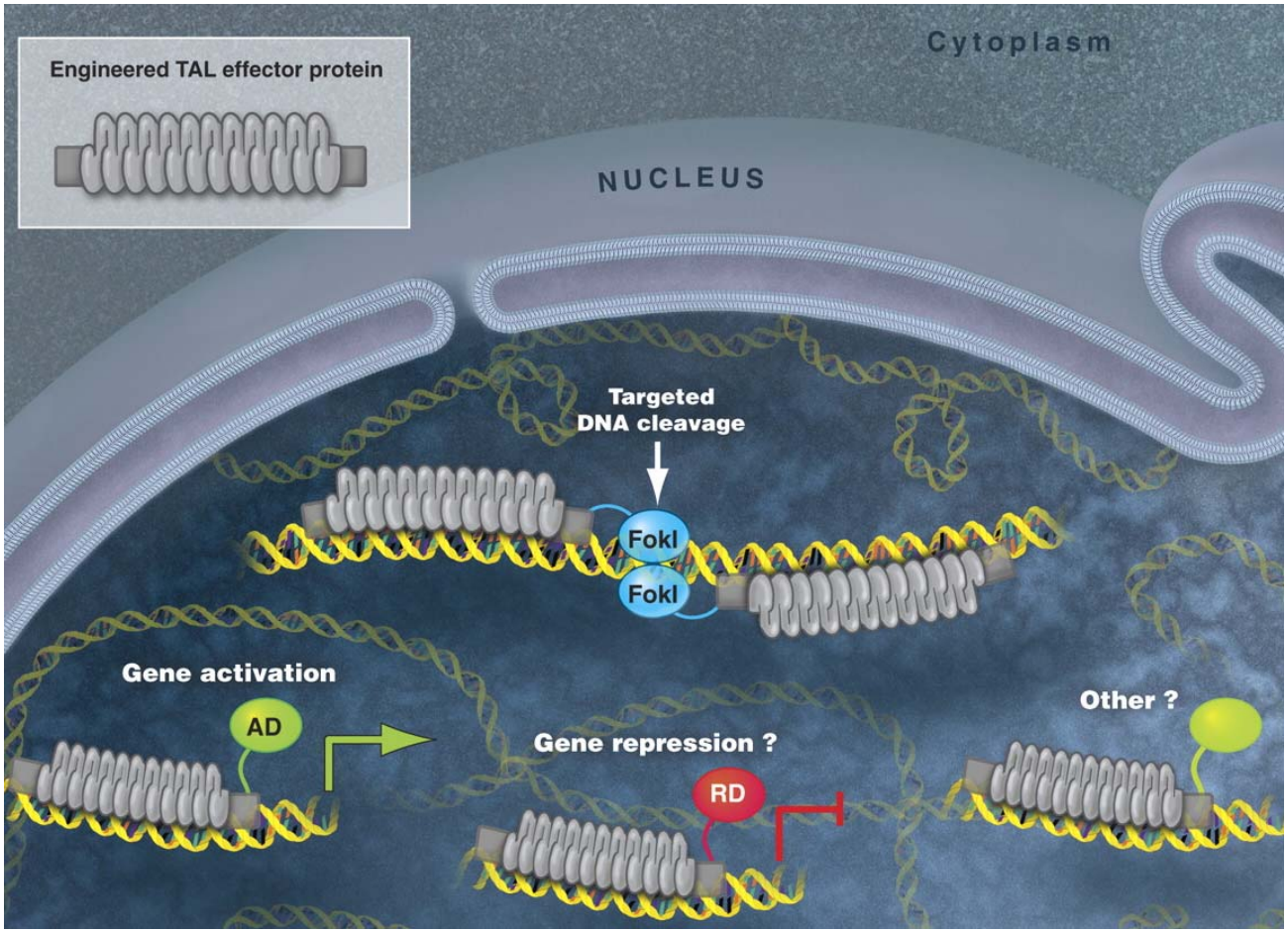
Goodacre (2005) Metabolomics

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

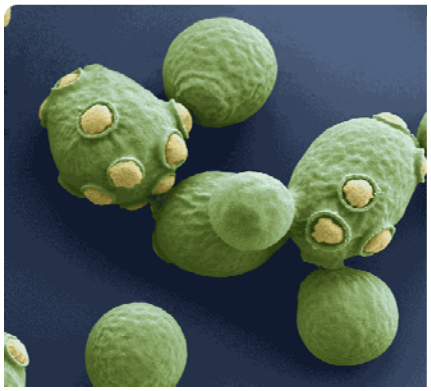
---------- Coffee Break ----------

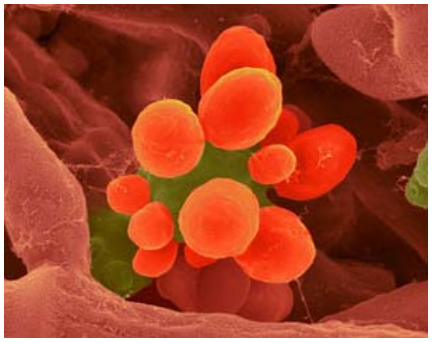
- Phylogenomics

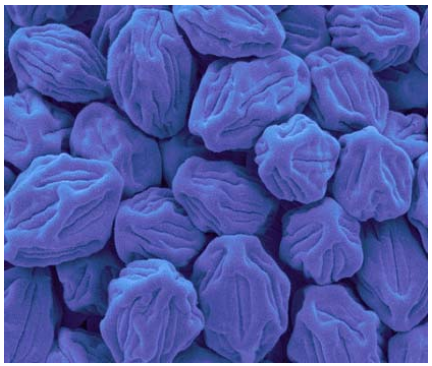
The Rokas Lab

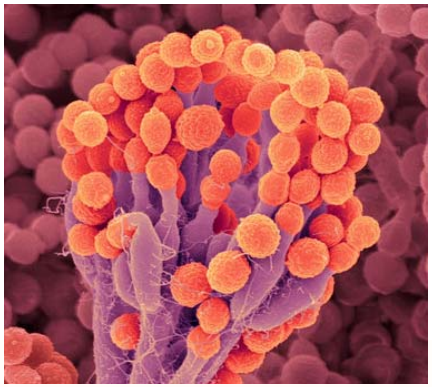
We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches

Fungi: A Model for the Study of the DNA Record

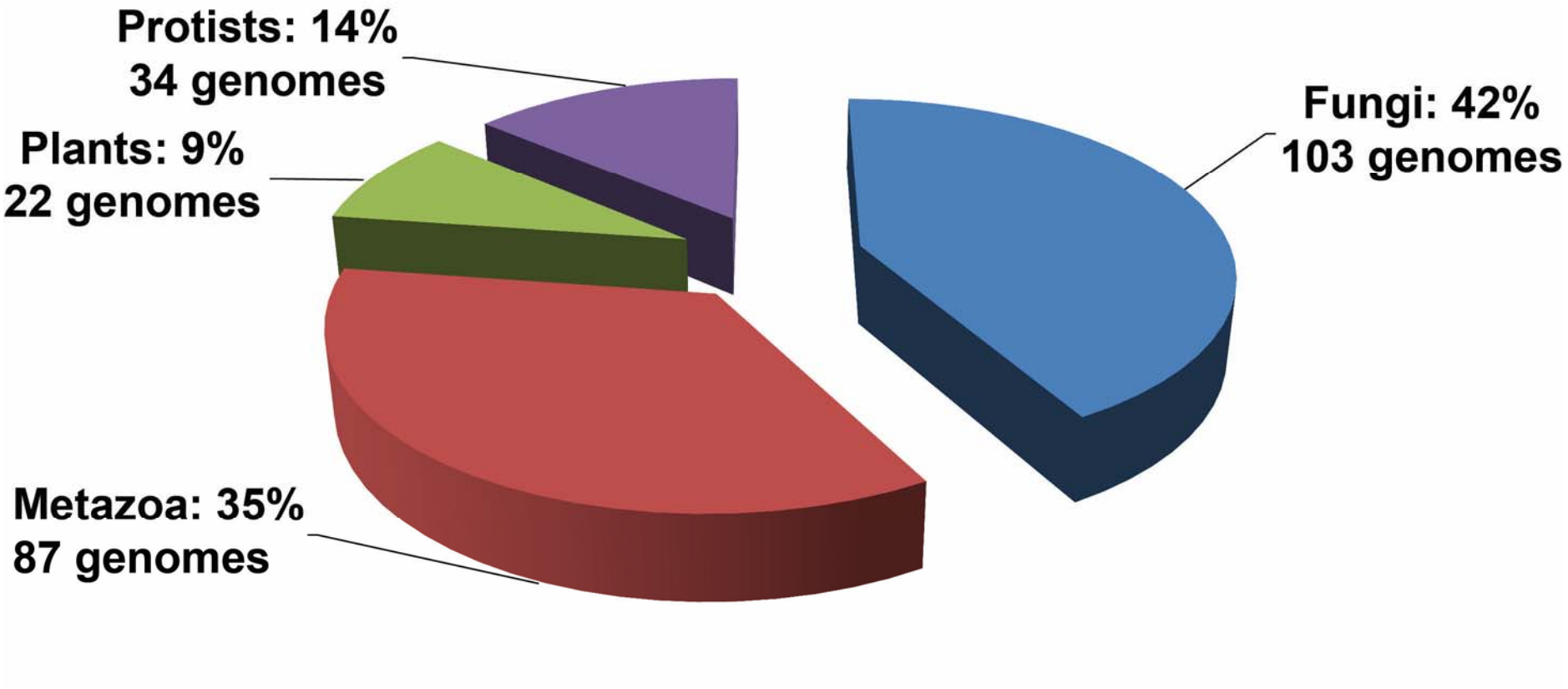








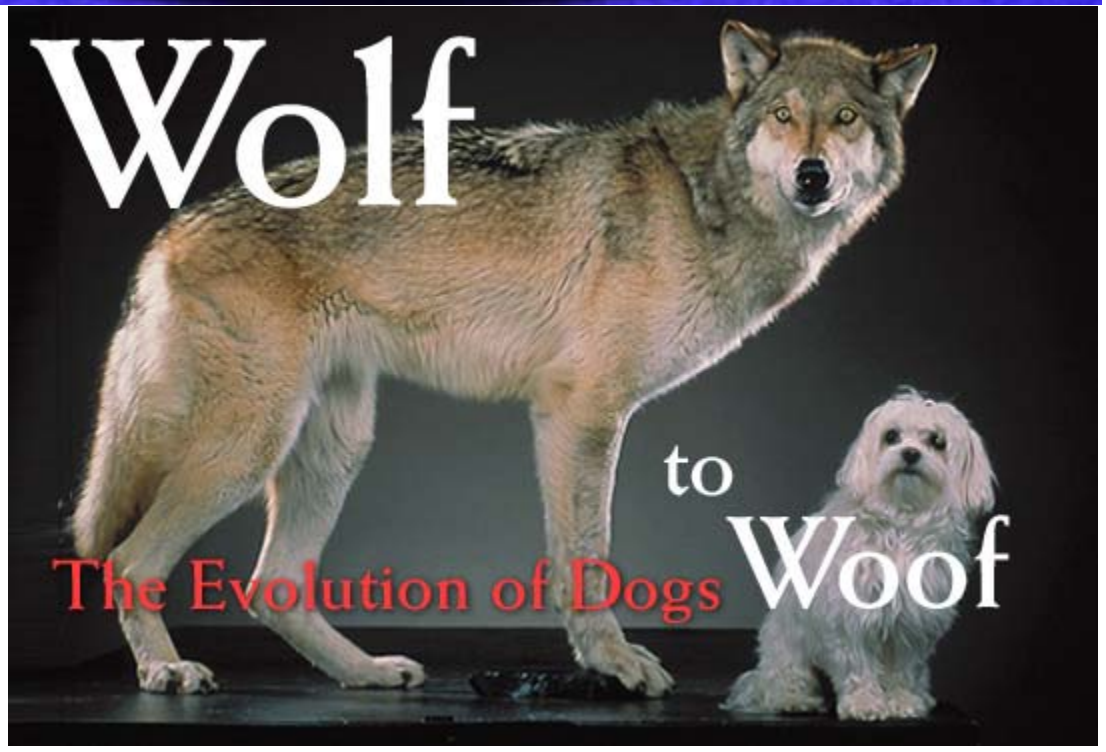
Fungi: the Most Sequenced Eukaryotic Lineage



fungi: 1000genomes

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

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


Wolf
to
Woof
The Evolution of Dogs

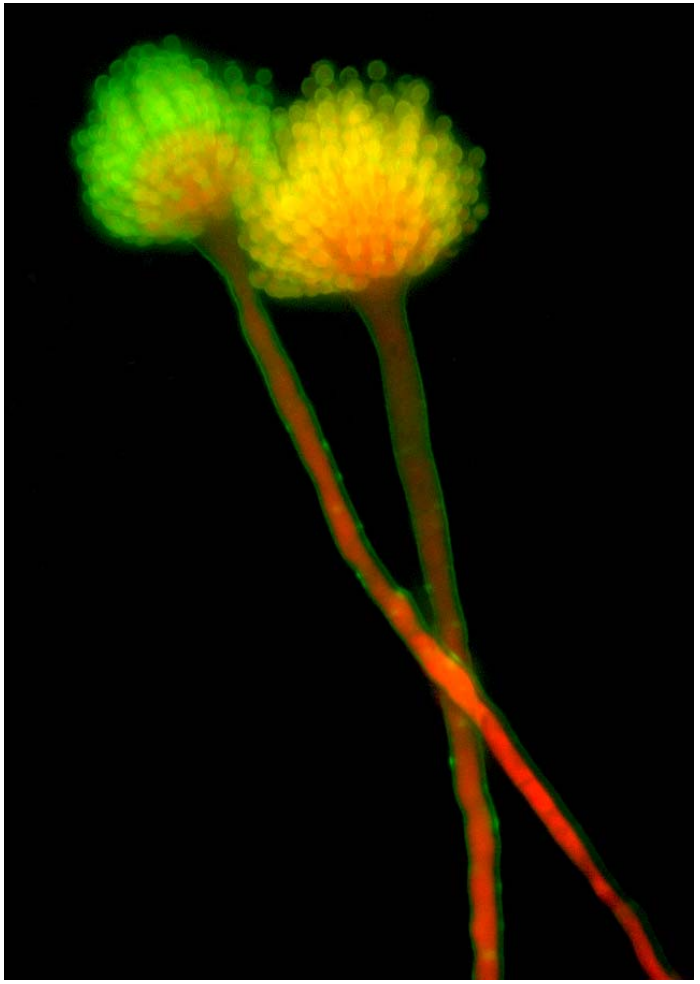

FAGE
Total
Net wt. 7oz (200g)
GREEK STRAINED
FAGE
Total 2%
GREEK STRAINED
FAGE
Total 0%
GRADE A
(pronounced: Fa-yeh!)
ALL NATURAL-NONFAT GREEK STRAINED YOGURT
change-diapers.com


特別純米樽酒
樽
にごり酒
鬼
WAKATA
DAIGINJO "ONIKOROSHI"


Gorgonzola


KIKKOMAN
NATURALLY BREWED
Soy Sauce
KIKKOMAN
NATURALLY BREWED
Soy Sauce

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

How Sake is Made

1. Polishing

2. Washing

3. Steaming

4. Koji making

5. Yeast starter

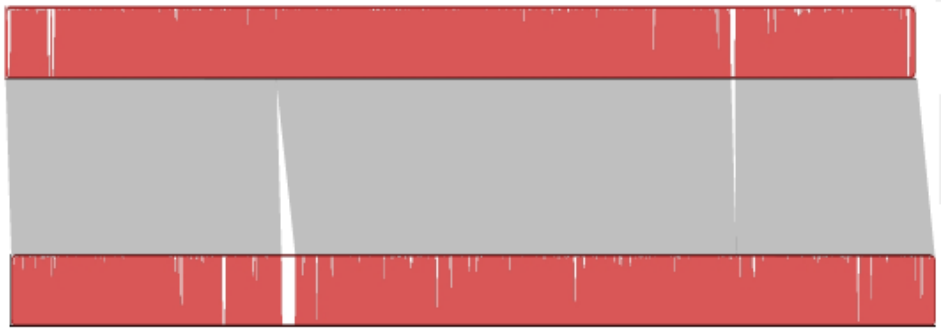
6. Pressing

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**

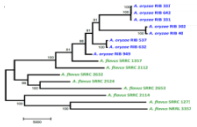
A. oryzae

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

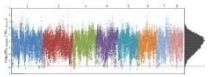
A. flavus

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

The Road to Domestication



Evolutionary Relationship?



Positively Selected Genes?

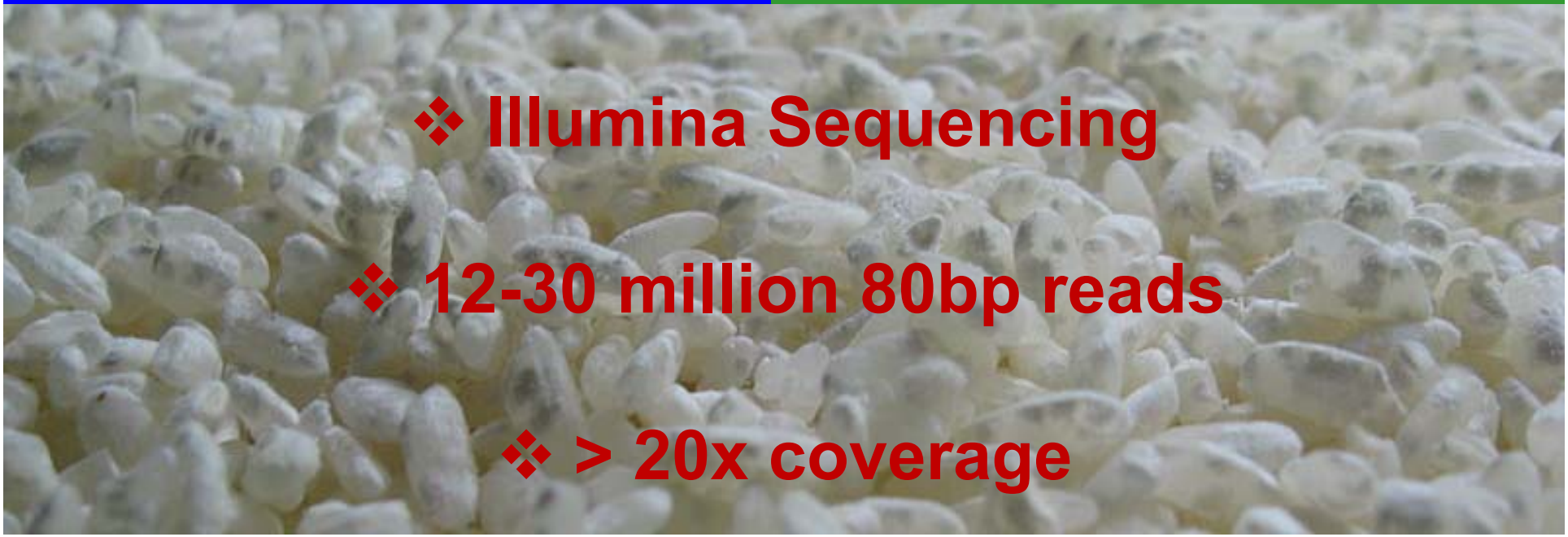
Functional Differences?

A. flavus

A. oryzae

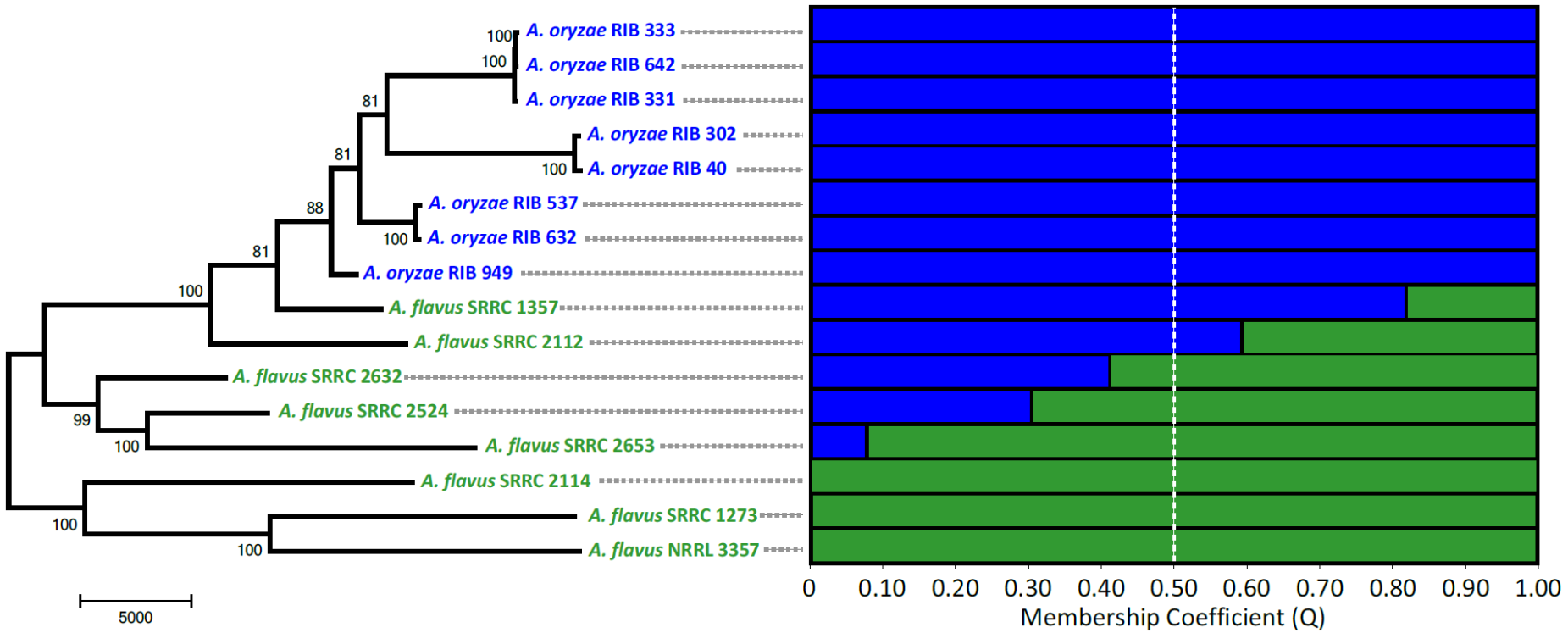
Sequencing 16 Genomes

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



- Illumina Sequencing
- 12-30 million 80bp reads
- > 20x coverage

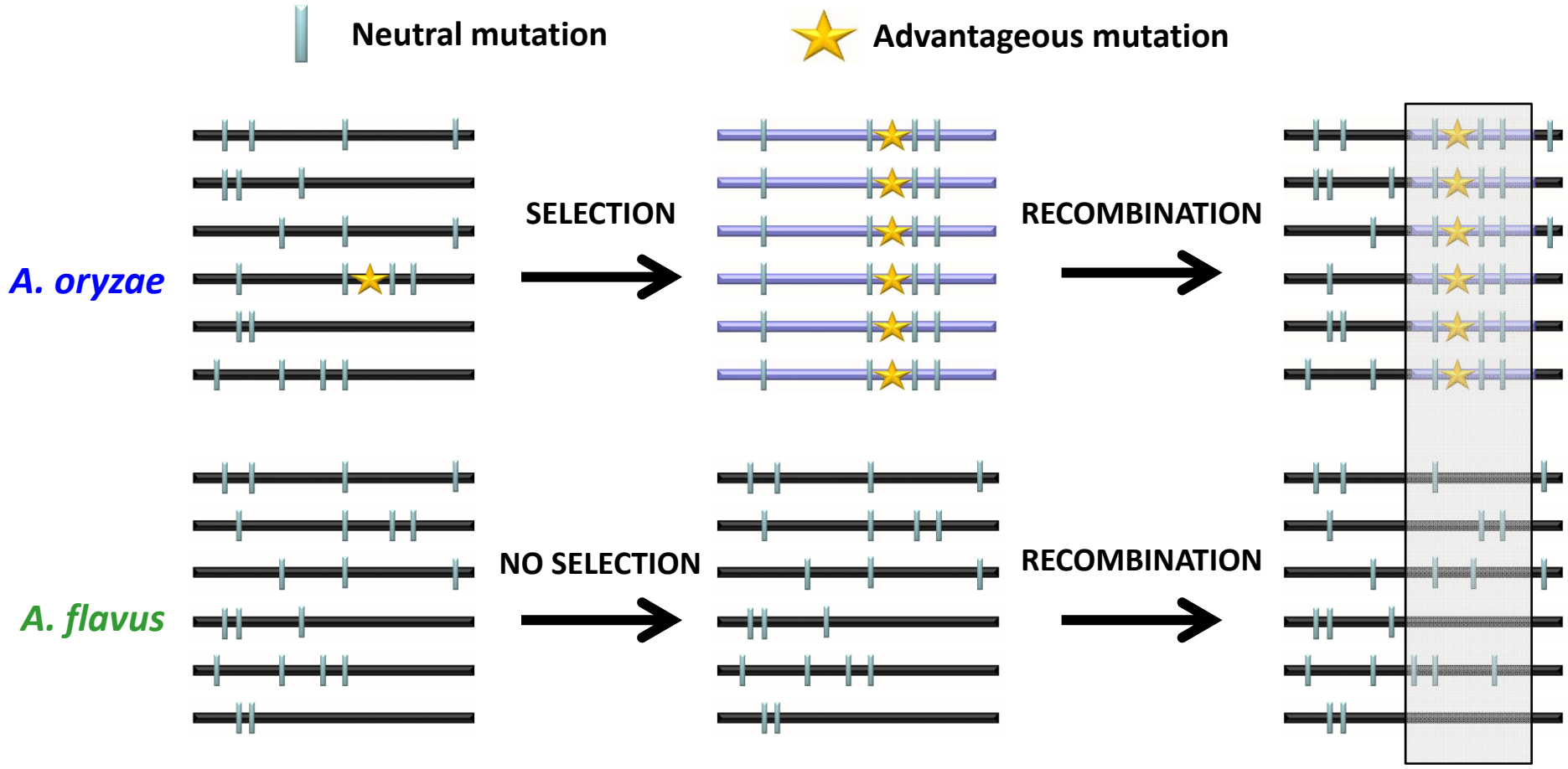
Aspergillus oryzae Isolates are Genetically Distinct



100,084 SNPs

Gibbons et al. (2012) Curr. Biol.

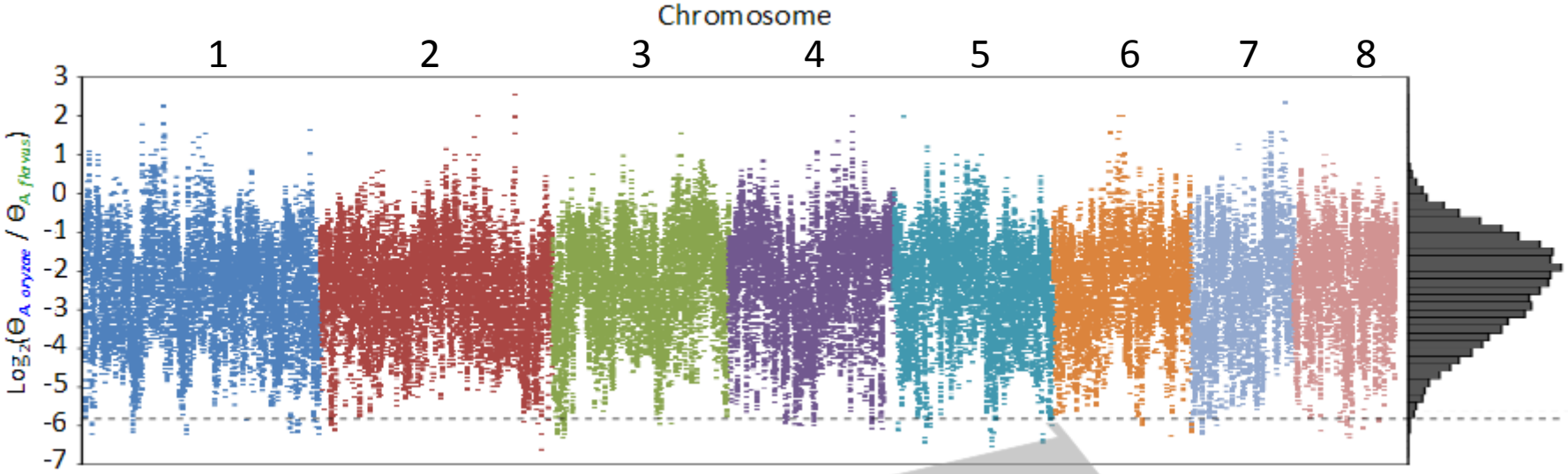
Detecting Recent Positive Selection in A. oryzae



- 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%

Gibbons et al. (2012) Curr. Biol.

Recent Positive Selection in the A. oryzae Genome



65,894 total windows

↓

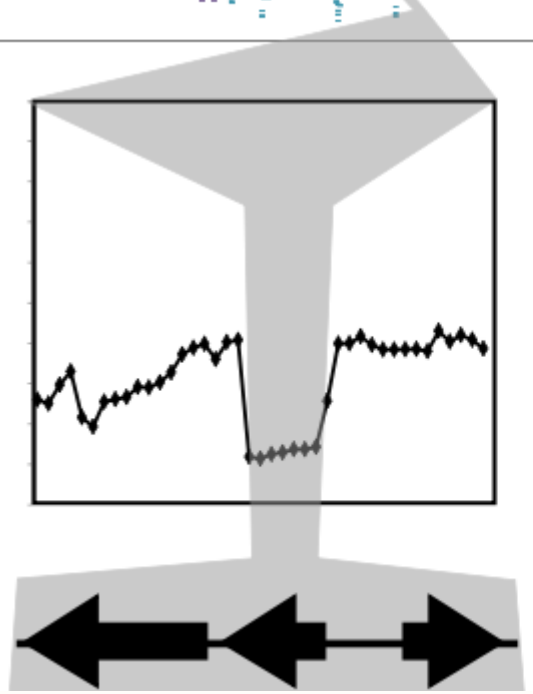
164 candidate windows

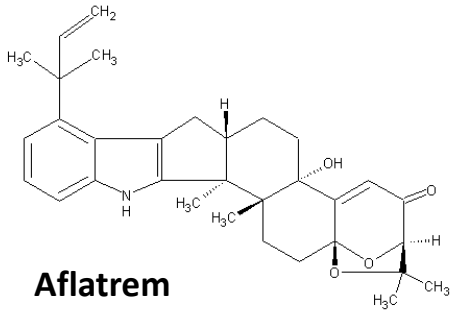
↓

61 Putative Selective Sweep Regions (PSSRs)

↓

148 Genes





Aflatrem

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

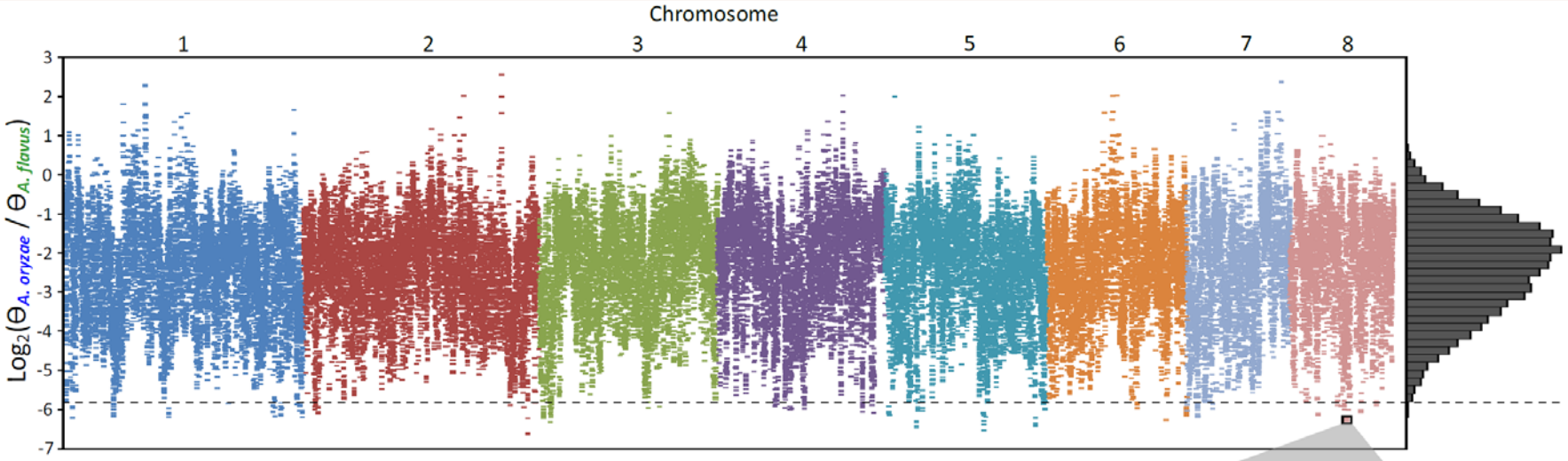
Gibbons et al. (2012) Curr. Biol.

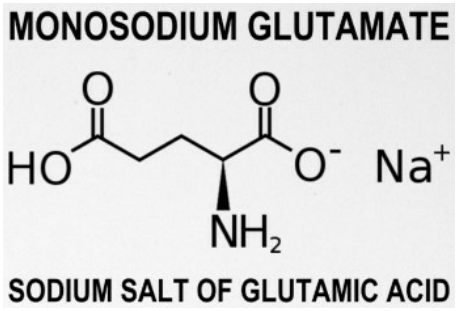
The Complex Flavor of Sake

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

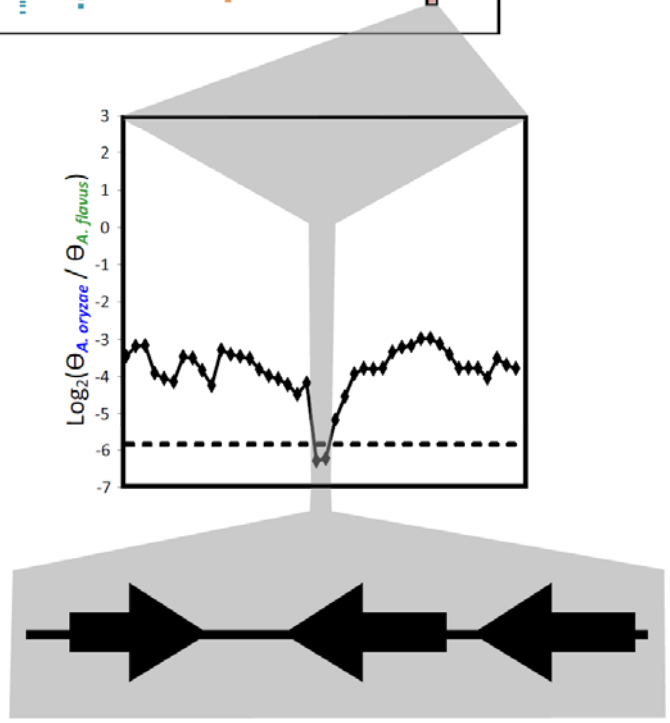
Have "flavor" genes been under selection?

Differences in a Flavor Associated Gene



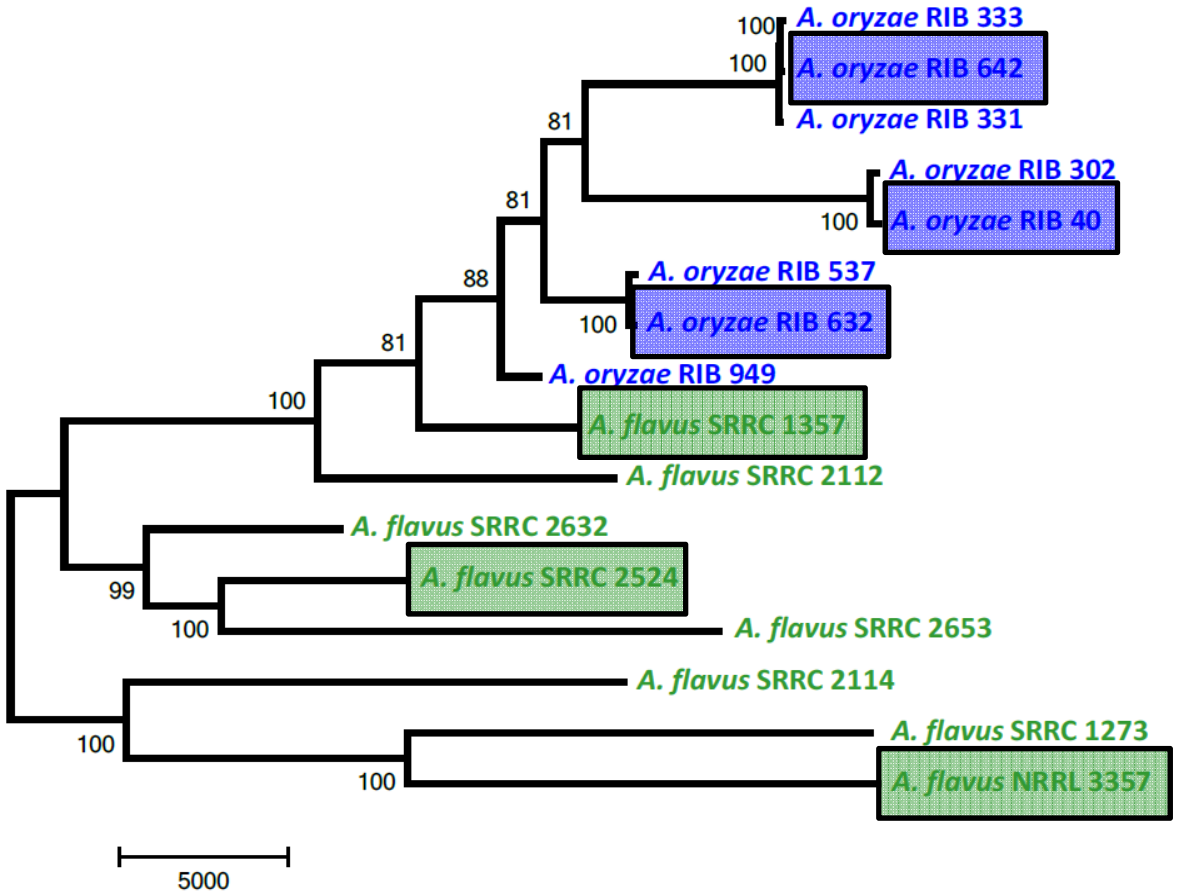


Glutamic Acid



Gibbons et al. (2012) Curr. Biol.

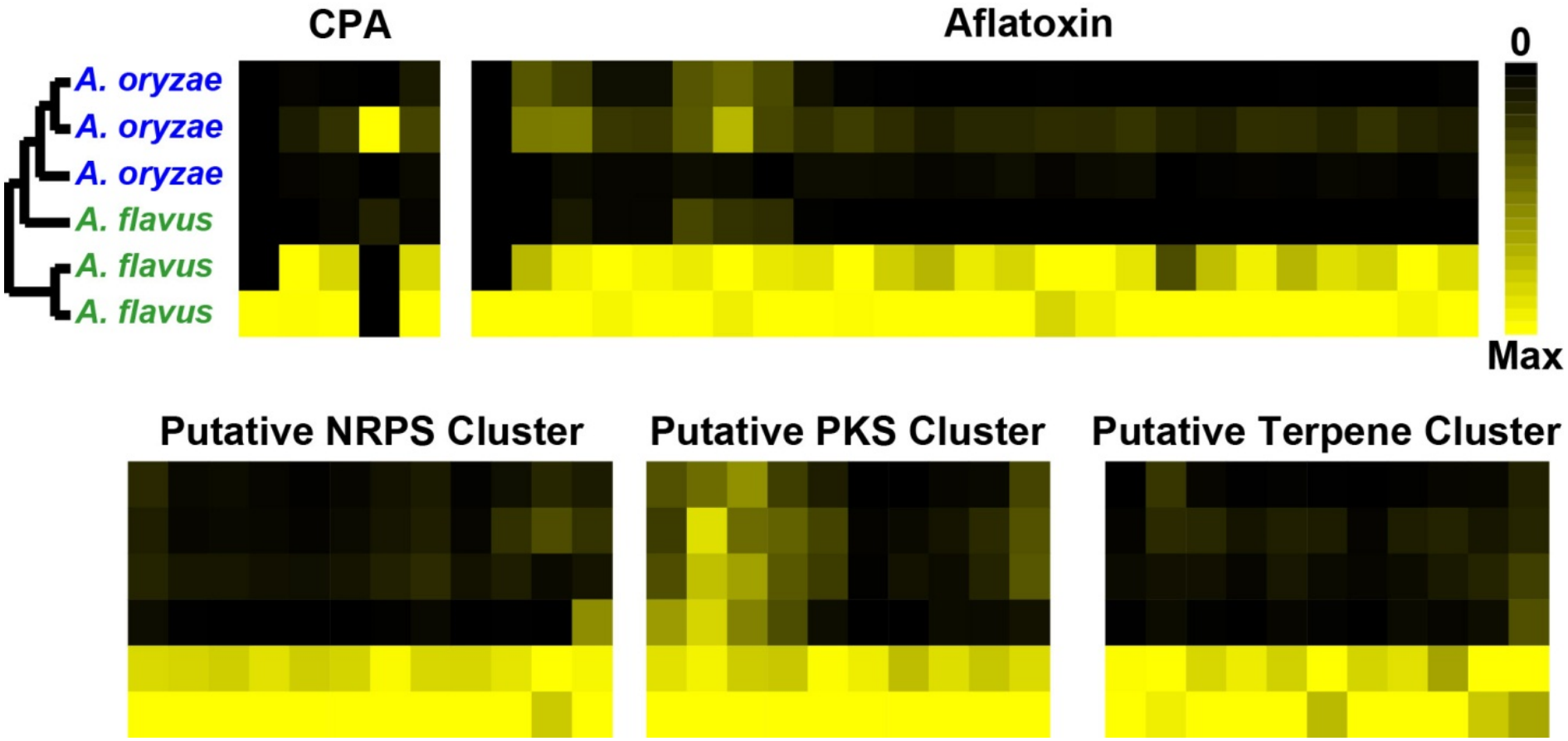
Detecting Differences in Transcript and Protein Abundance



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

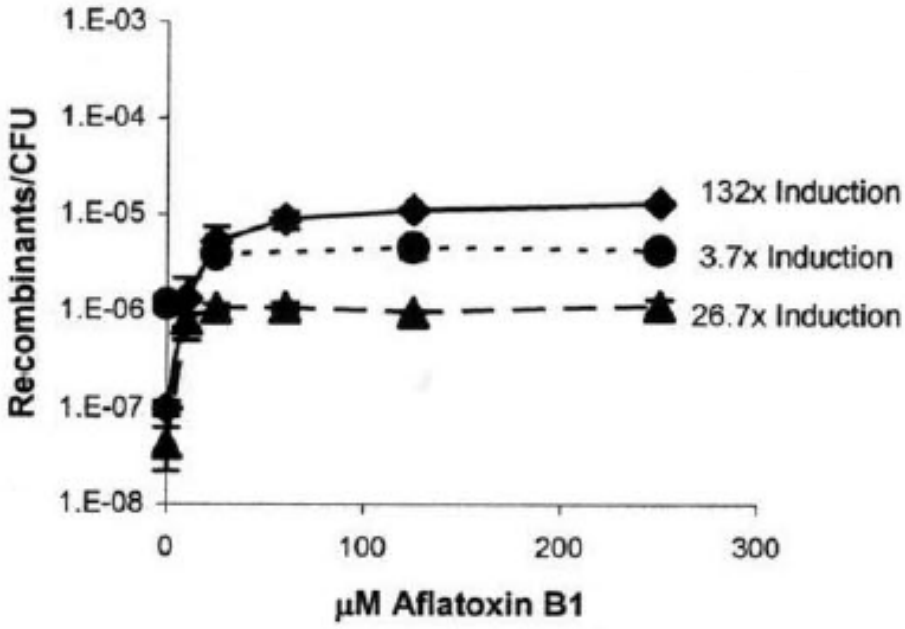
- Sake strains
- 30˚ C
- 24 hours

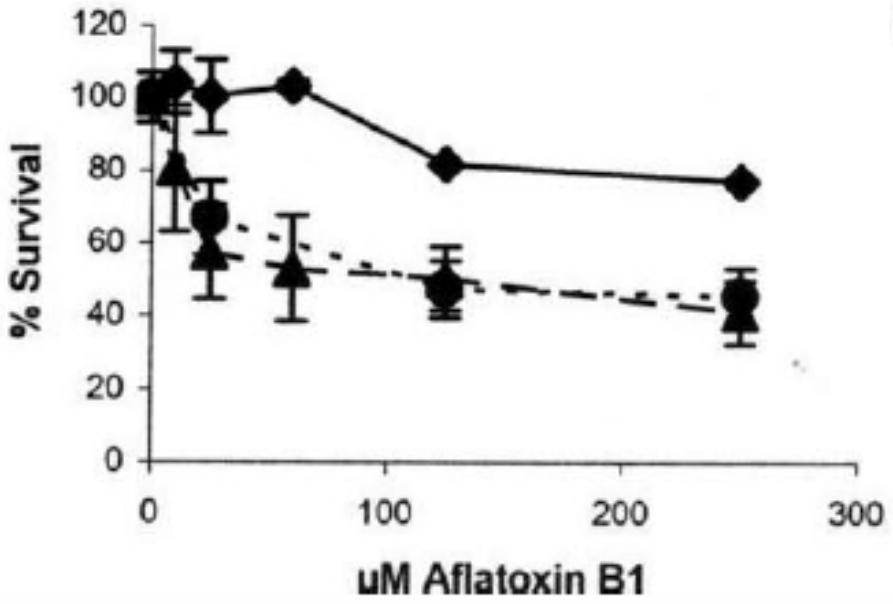
Down-Regulation of Secondary Metabolism in A. oryzae



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

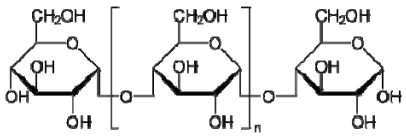
Keller-Seitz et al. (2004) Mol. Biol. Cell

A. oryzae *Breaks Down Starch into Sugar*

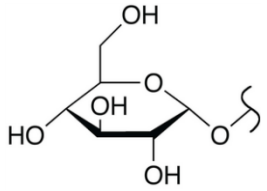
Rice Preparation

Saccharification

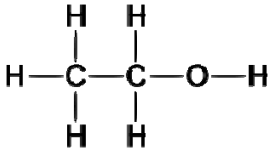
Fermentation



A. oryzae



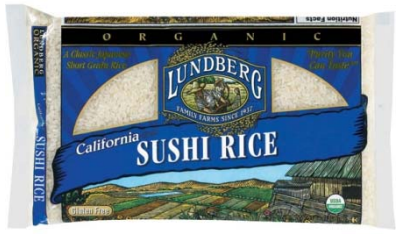
Yeast



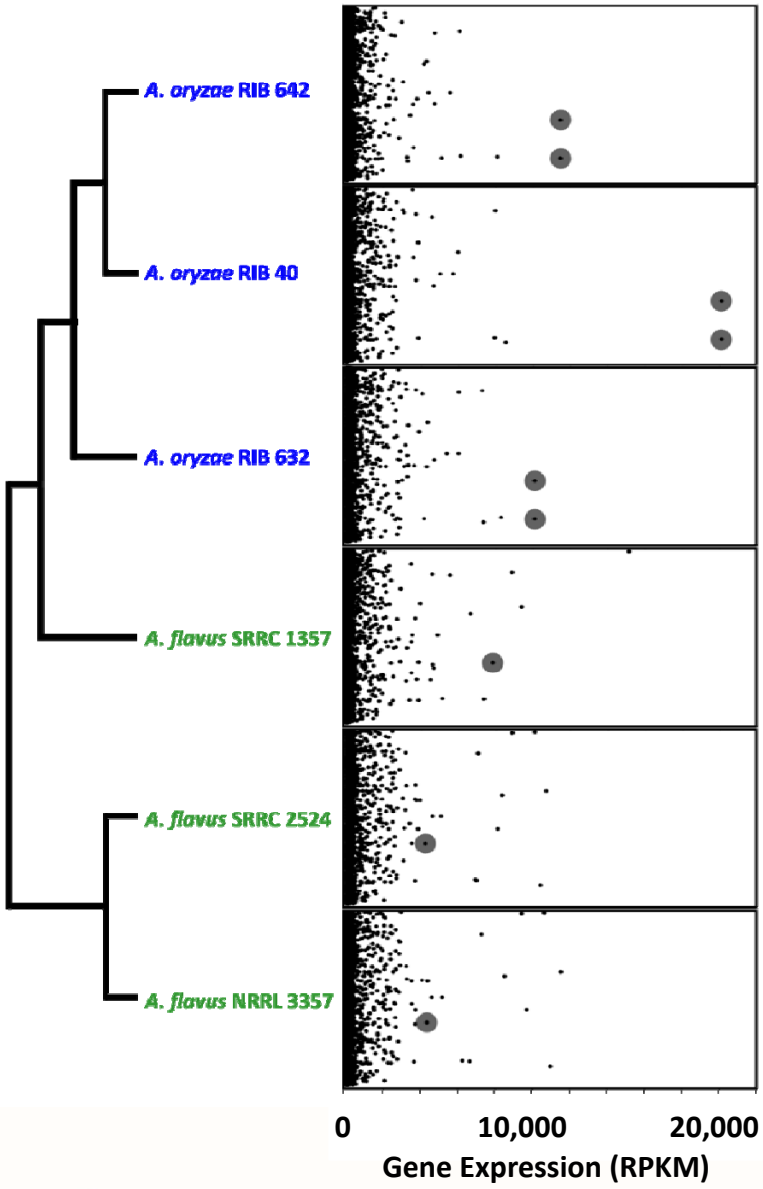
Starch

Sugar

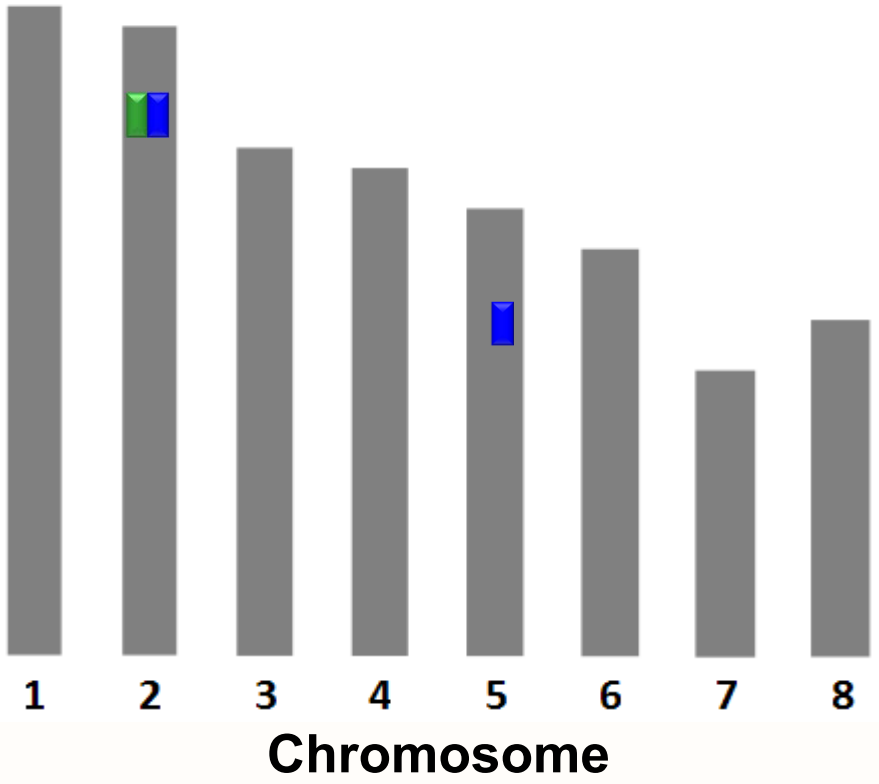
Alcohol



α-amylase is Highly Abundant

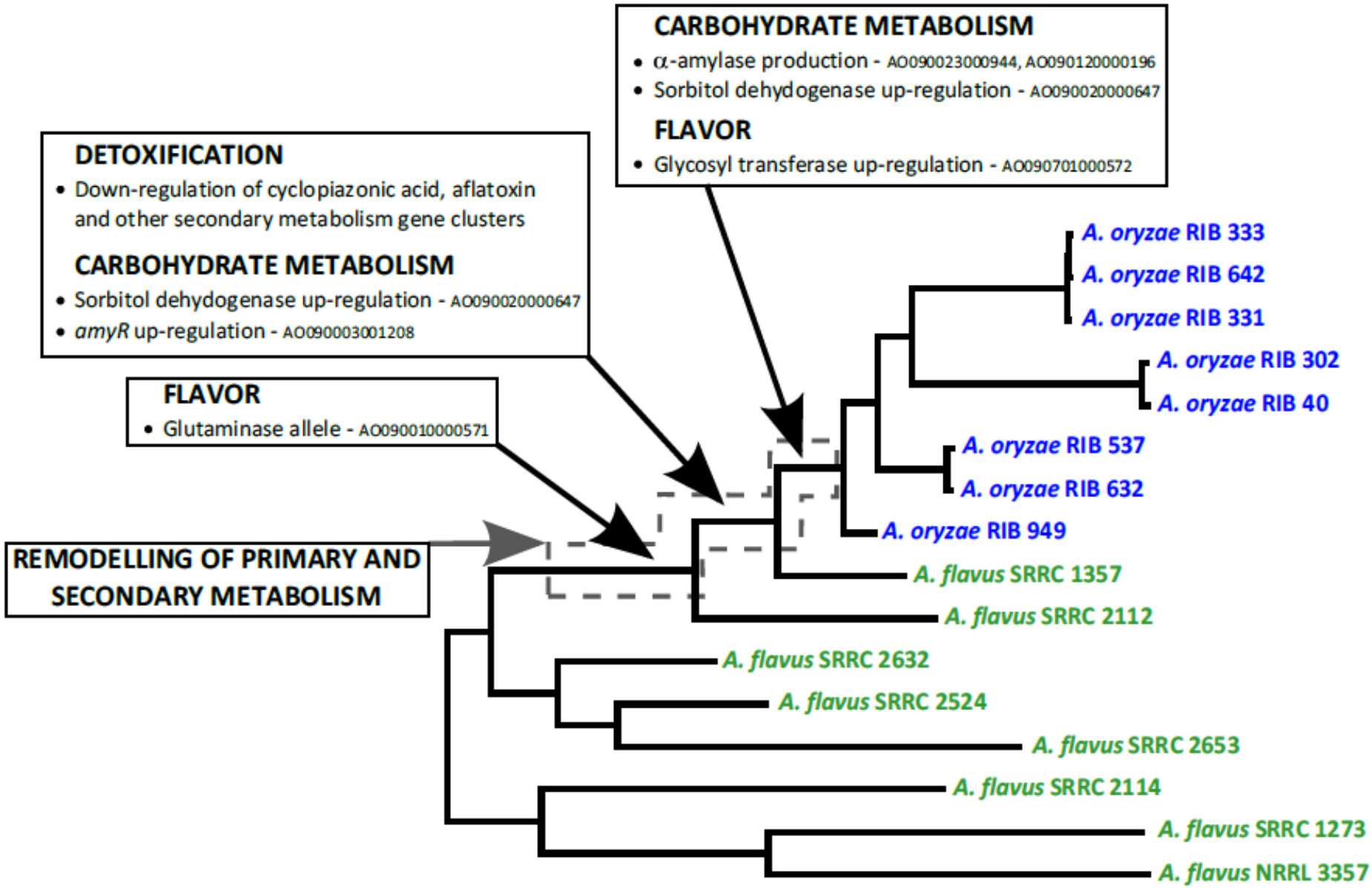


- Gene Expression ⬆ *P* = 1e-300
- Protein Enrichment ⬆ *P* = 8e-63
- Carbohydrate Metabolism ⬆ *P* = 6e-5



Gibbons et al. (2012) Curr. Biol.

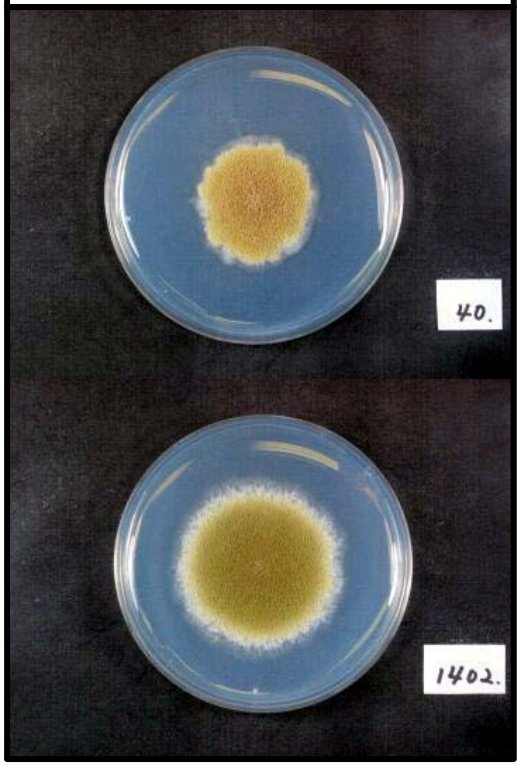
DOMESTICATION RD



Contrasting Domestication Patterns Across Kingdoms

Plants

Microbes



Animals

Lecture Outline

- Introduction to Evolutionary Genomics
- Population Genomics

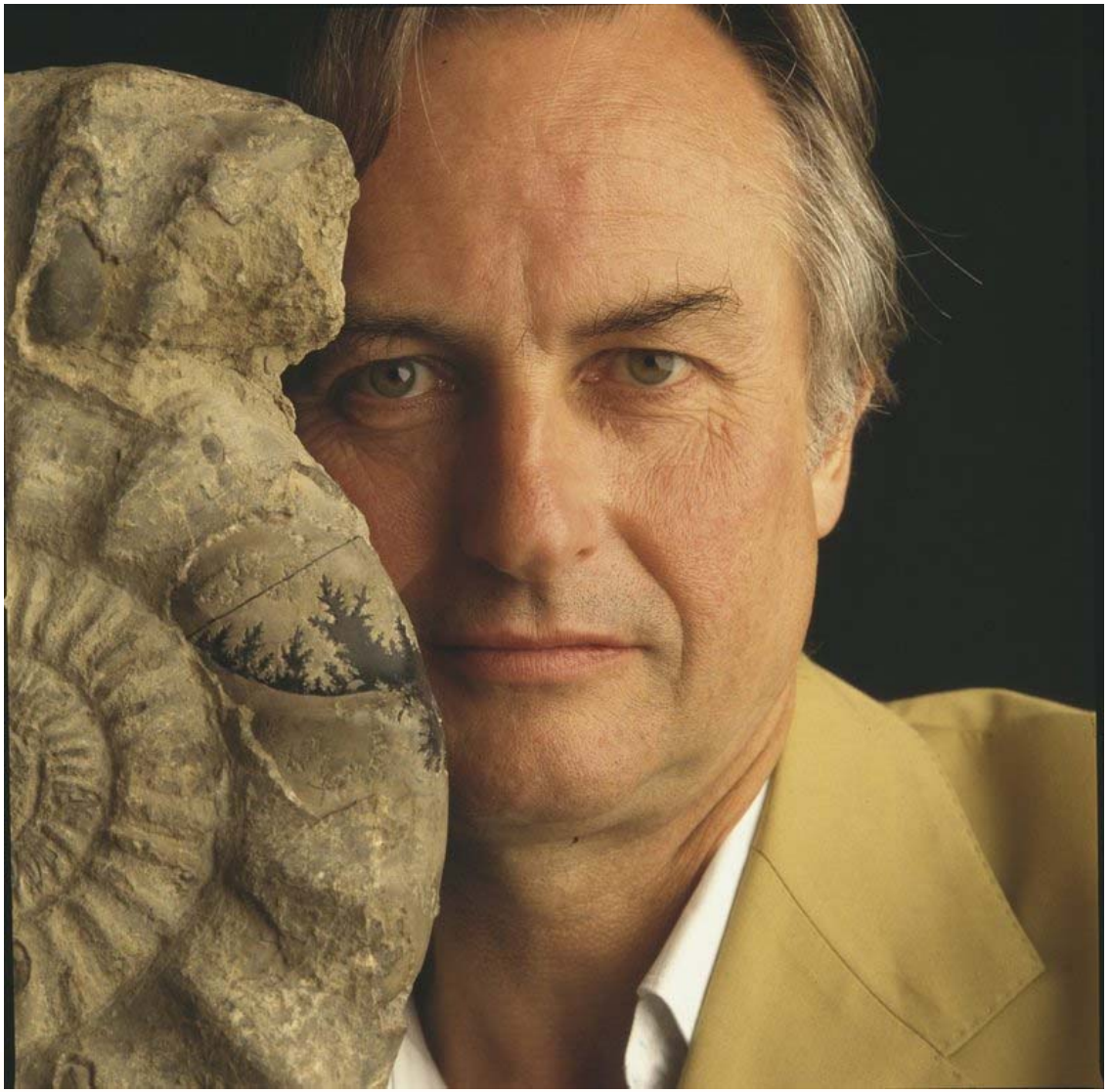
---------- Coffee Break ----------

- 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."

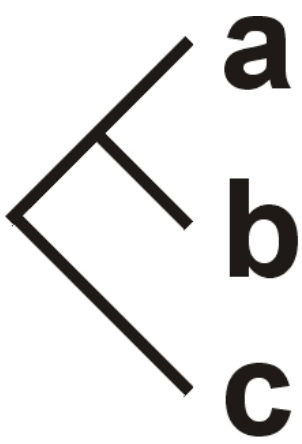
Richard Dawkins



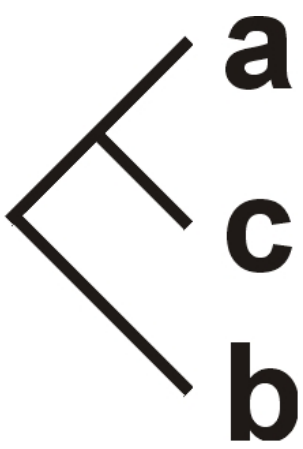
Dawkins (2003), A Devil's Chaplain

The Problem of Incongruence

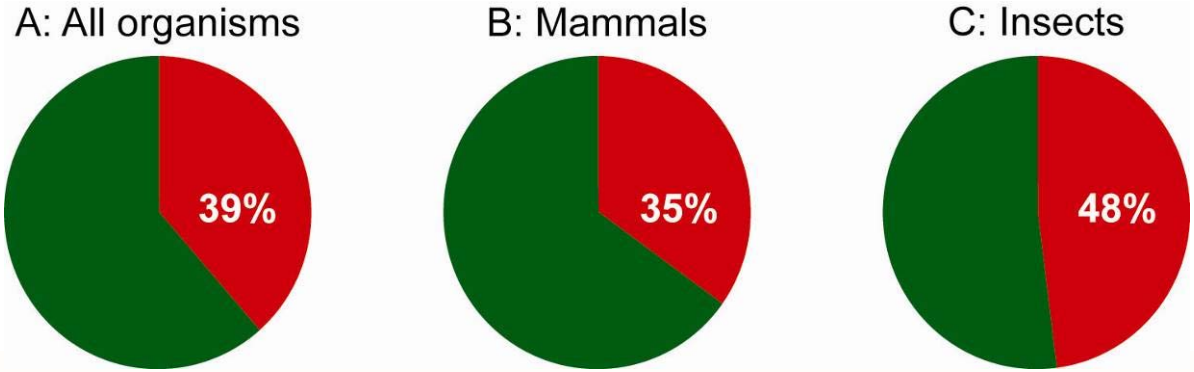
Gene X



Gene Y



Species tree?



Incongruence **is pervasive in the phylogenetics literature**

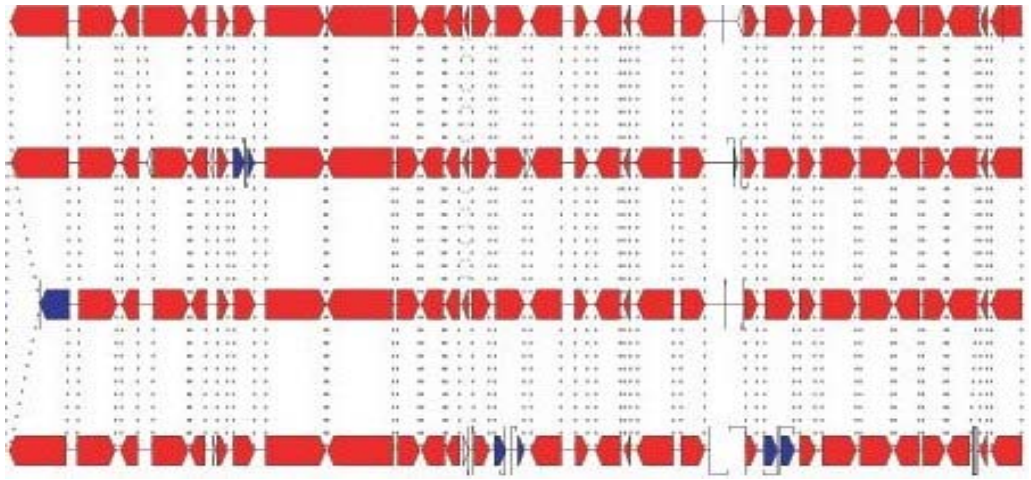
A Systematic Evaluation of Single Gene Phylogenies

S. cerevisiae

S. paradoxus

S. mikatae

S. bayanus

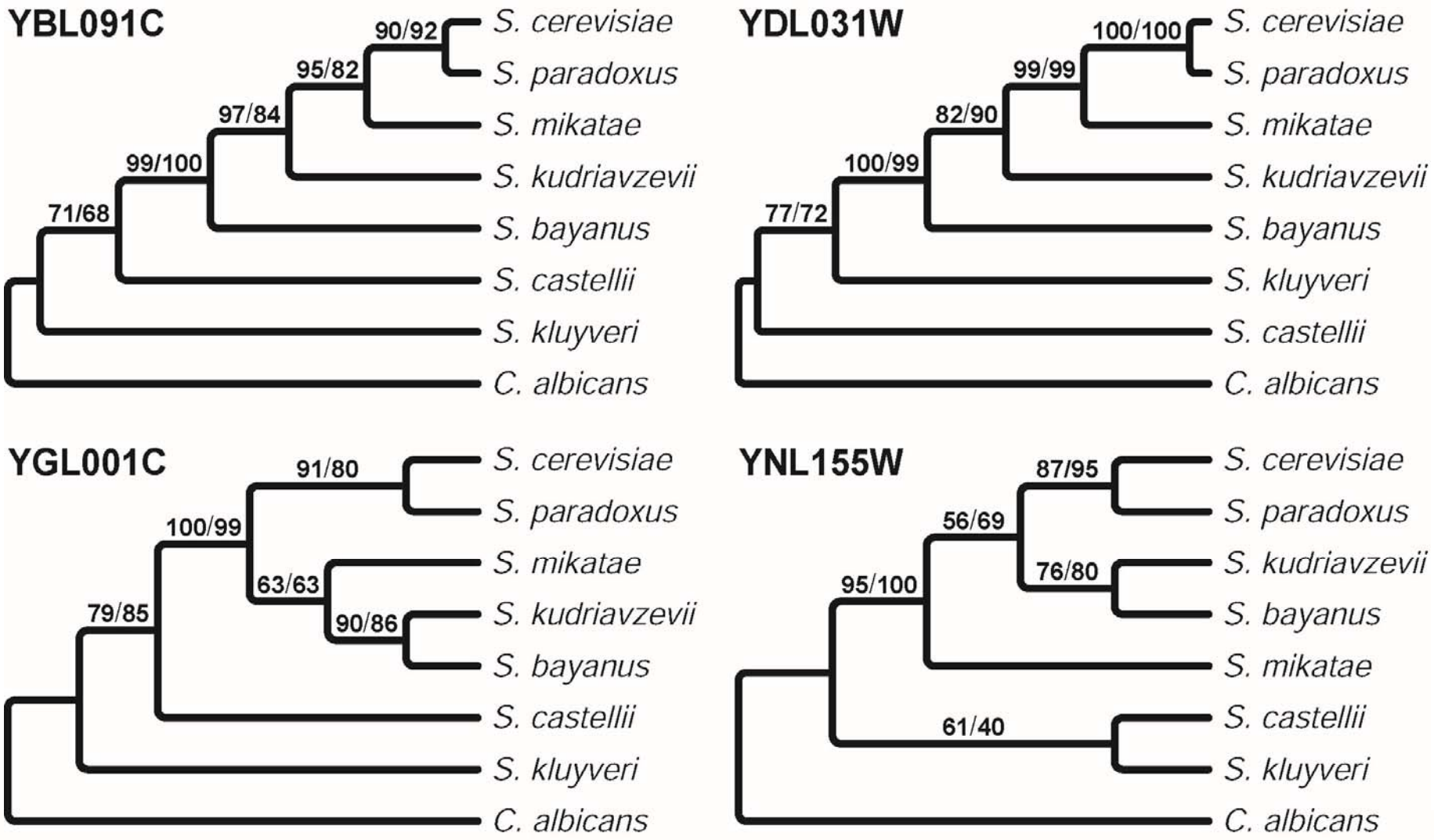


Dataset: 106 genes on all 16 chromosomes totaling 127kb corresponding roughly to 1% of the genomic sequence, 2% of genes

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

Kellis et al. (2003) Nature

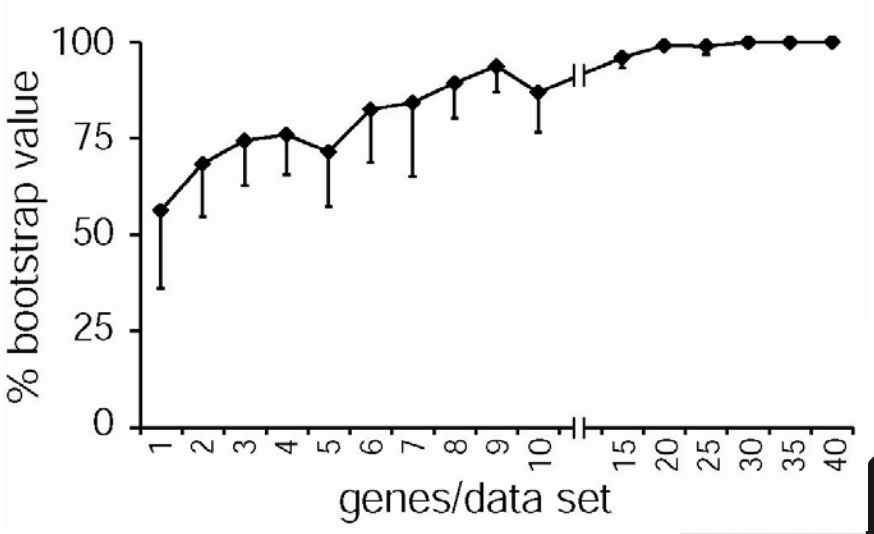
Incongruence at the Single Gene Level

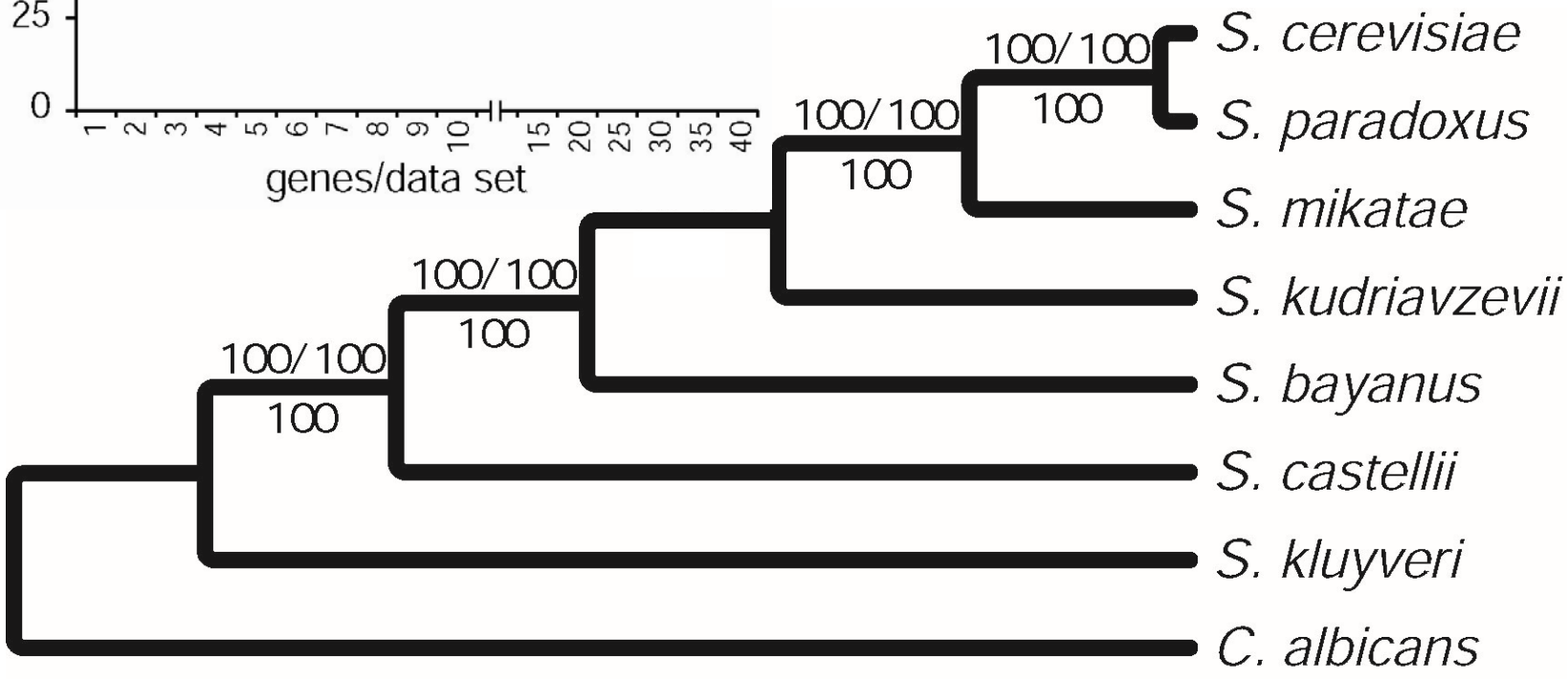


ML / MP

Rokas et al. (2003) Nature

Concatenation of 106 Genes Yields a Single Yeast Phylogeny

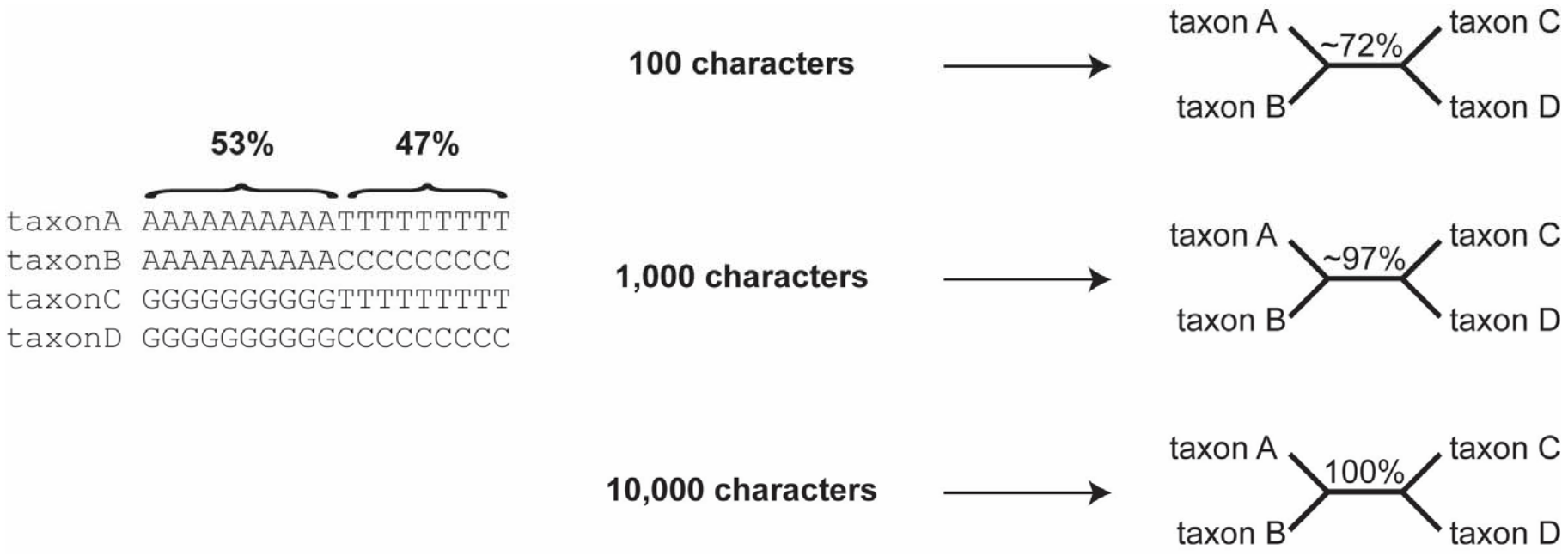




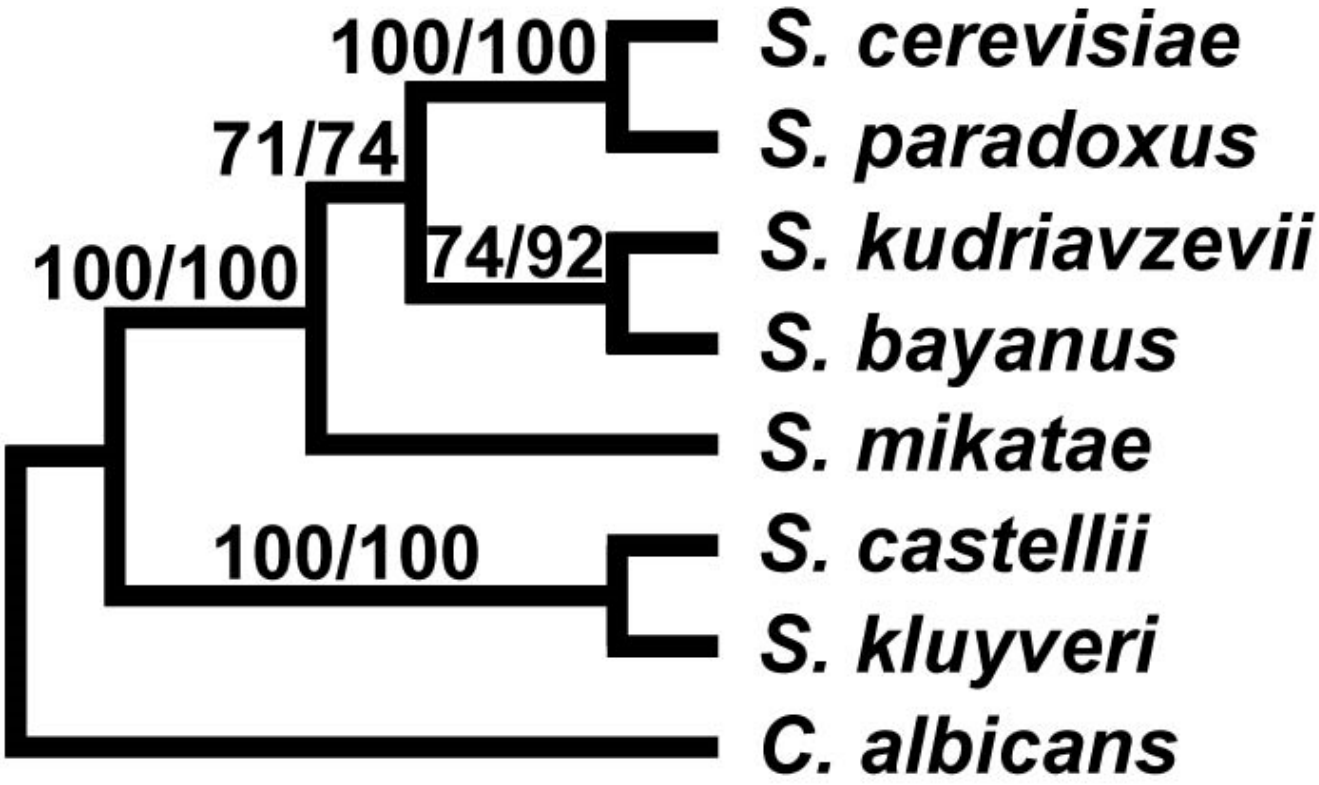
ML / MP on nt
MP on aa

Rokas et al. (2003) Nature

Bootstrap Support is Misleading When Used in Large Datasets



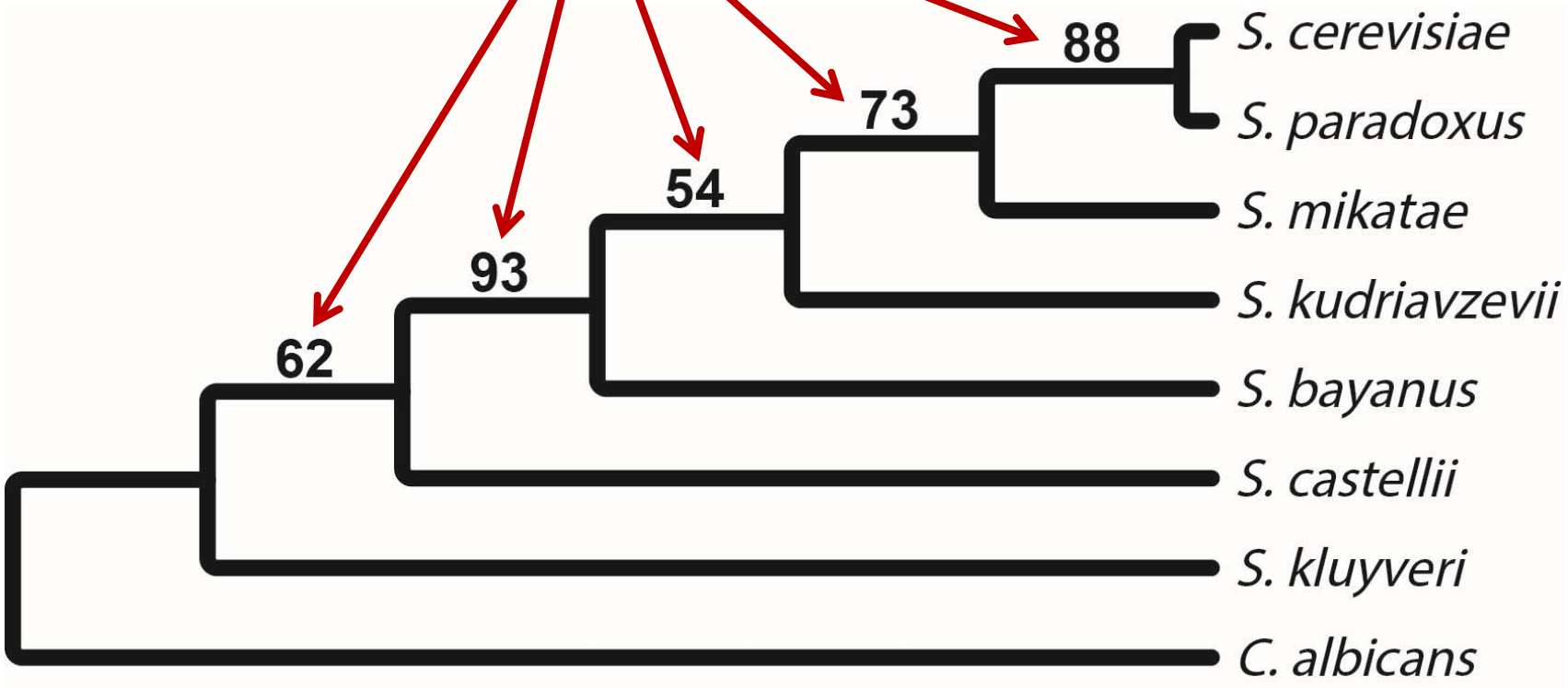
Bootstrap Support is Misleading When Used in Large Datasets



Rokas et al. (2003) Nature

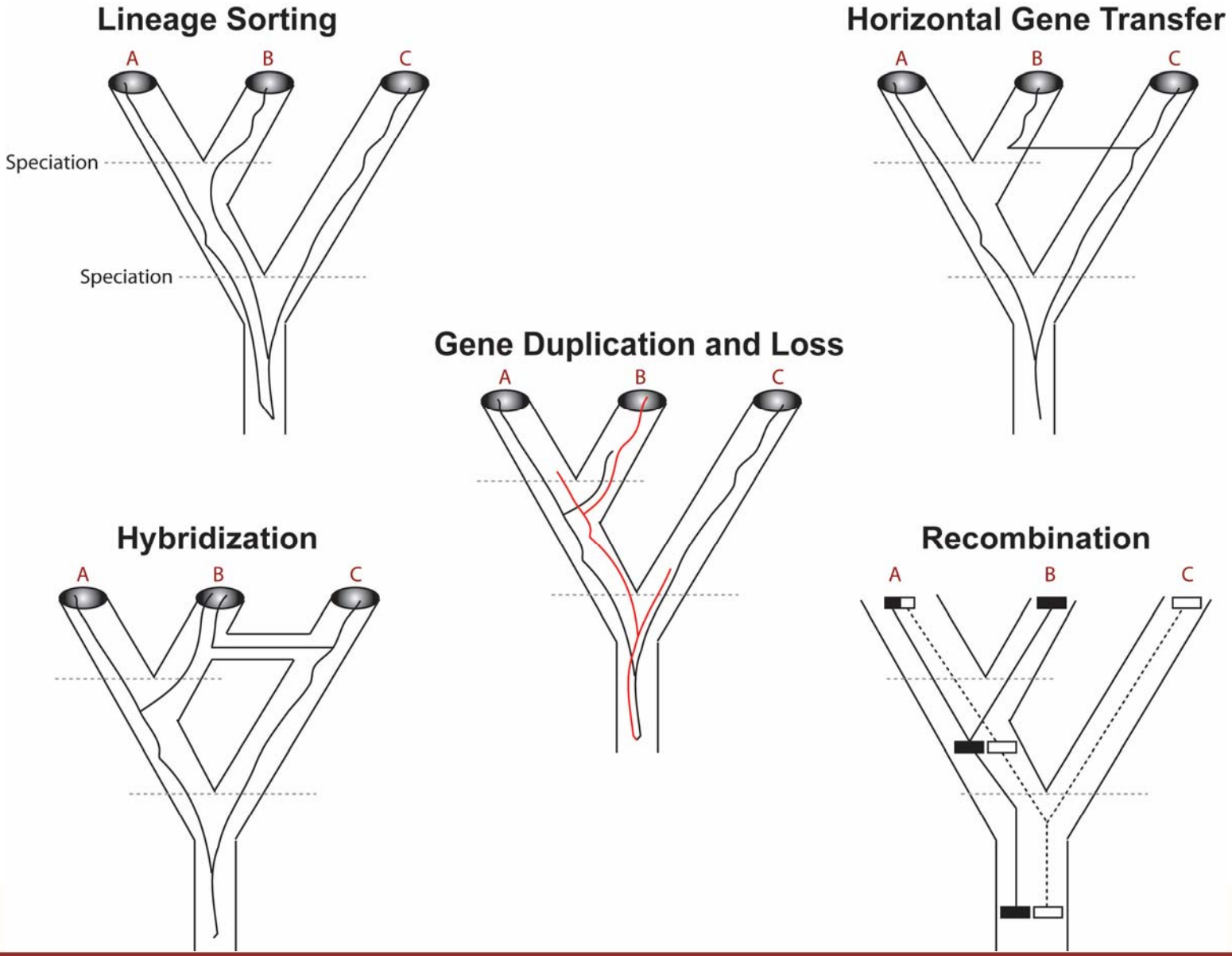
The Majority of Single Genes Support the Concatenation Phylogeny

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



Rokas et al. (2003) Nature

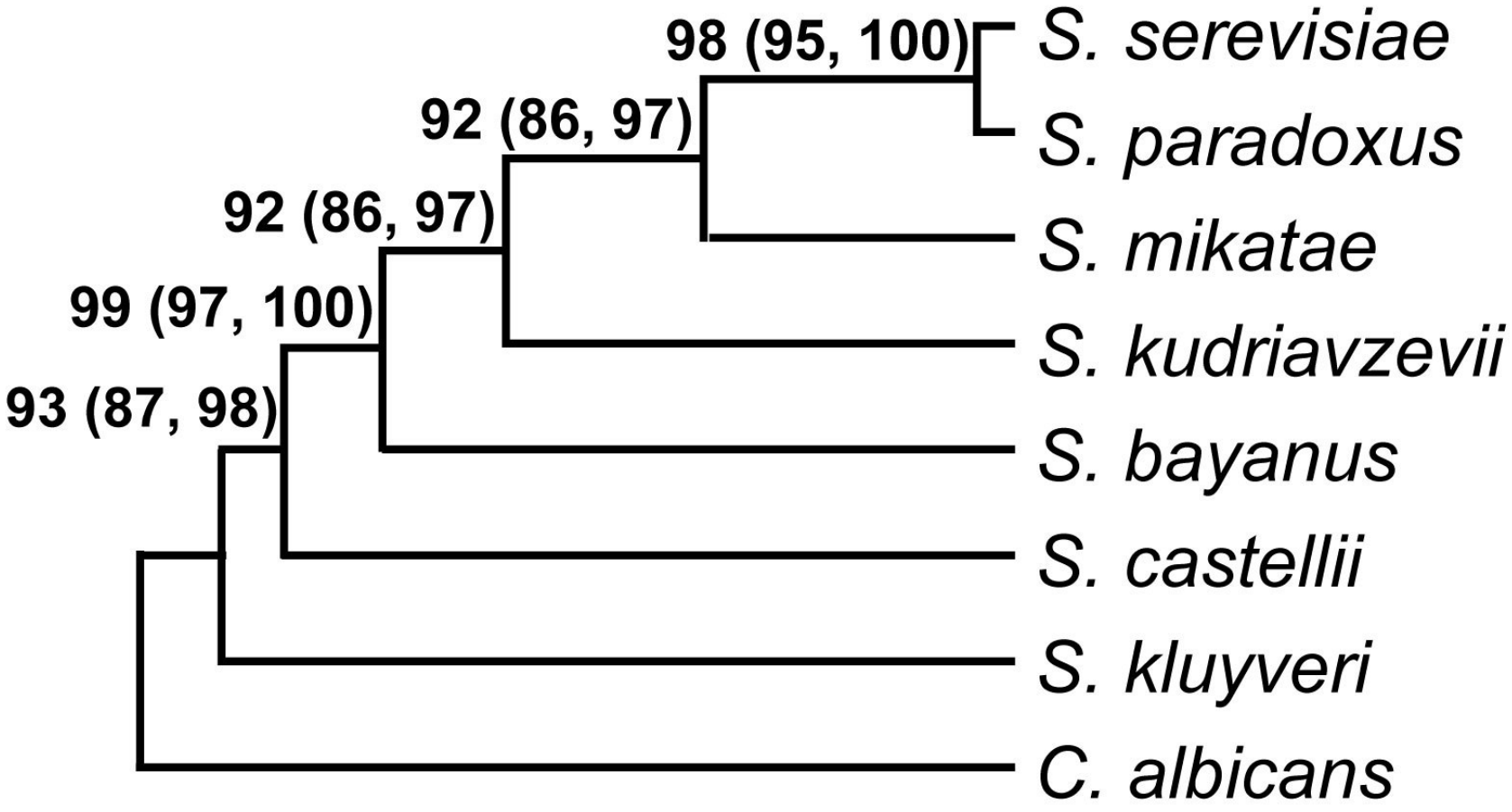
Gene Trees Can Differ from Species Trees



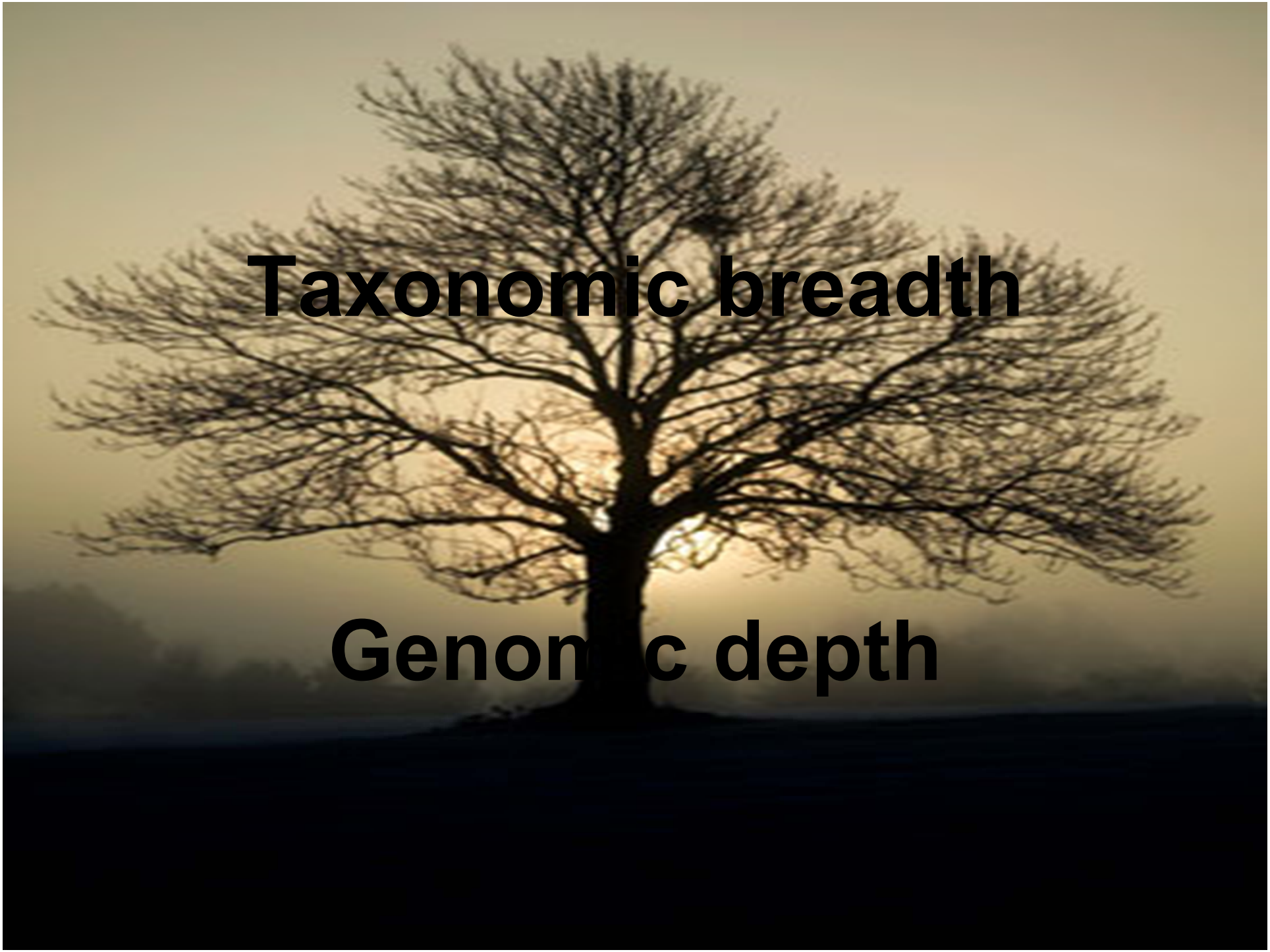
Degnan & Rosenberg (2009) Trends Ecol. Evol.

Inferring the Species Phylogeny from Single Gene Histories

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



Ané et al. (2007) Mol. Biol. Evol.

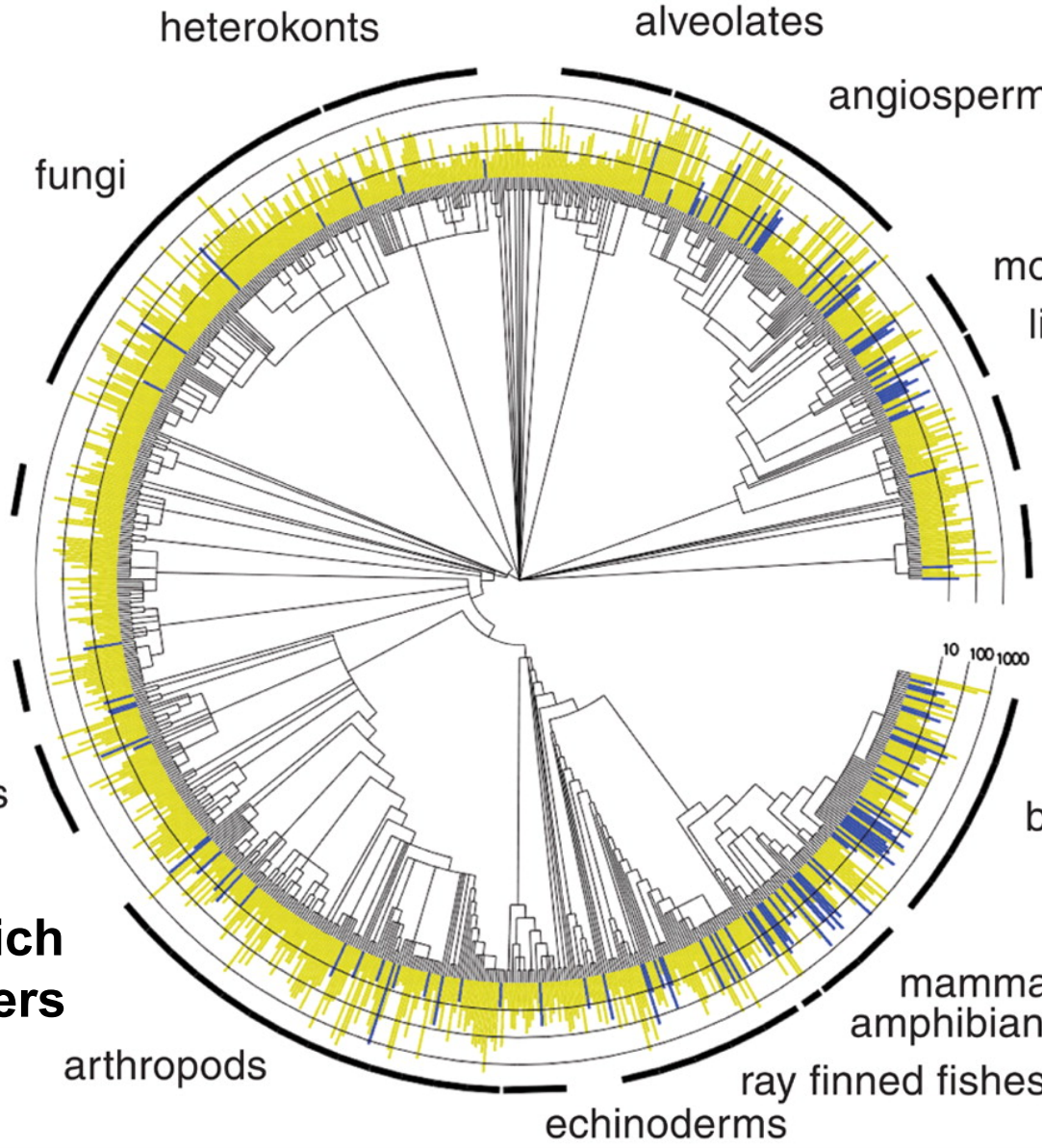

Taxonomic breadth
Genomic depth

Estimating the Taxonomic Breadth of the Tree of Life

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



1/40 fungal orders

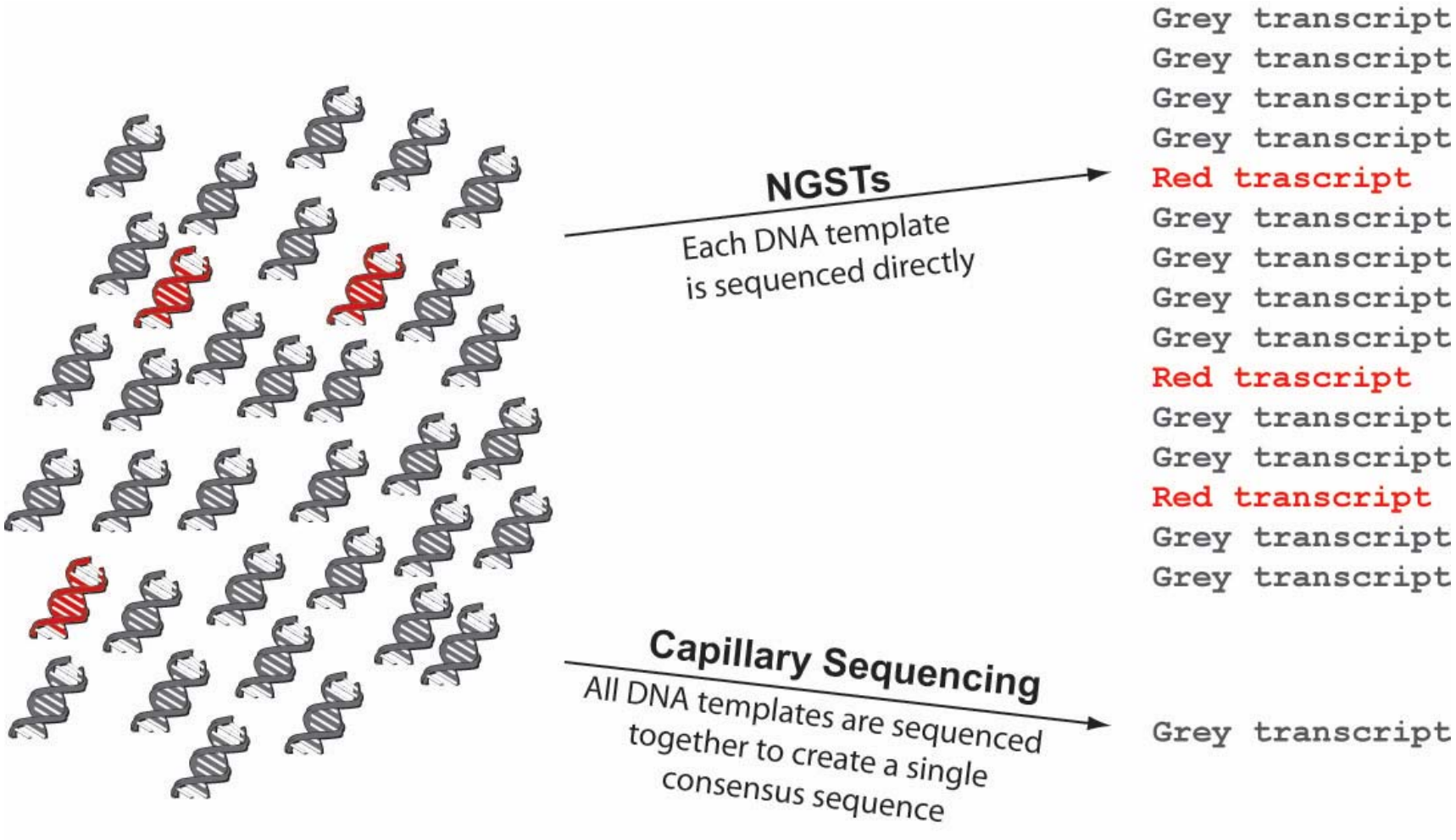
3/45 angiosperm orders

1/45 species-rich arthropod orders supported

16/57 vertebrate orders

Sanderson (2008) Science

Next-Gen Sequencing is Qualitative and Quantitative



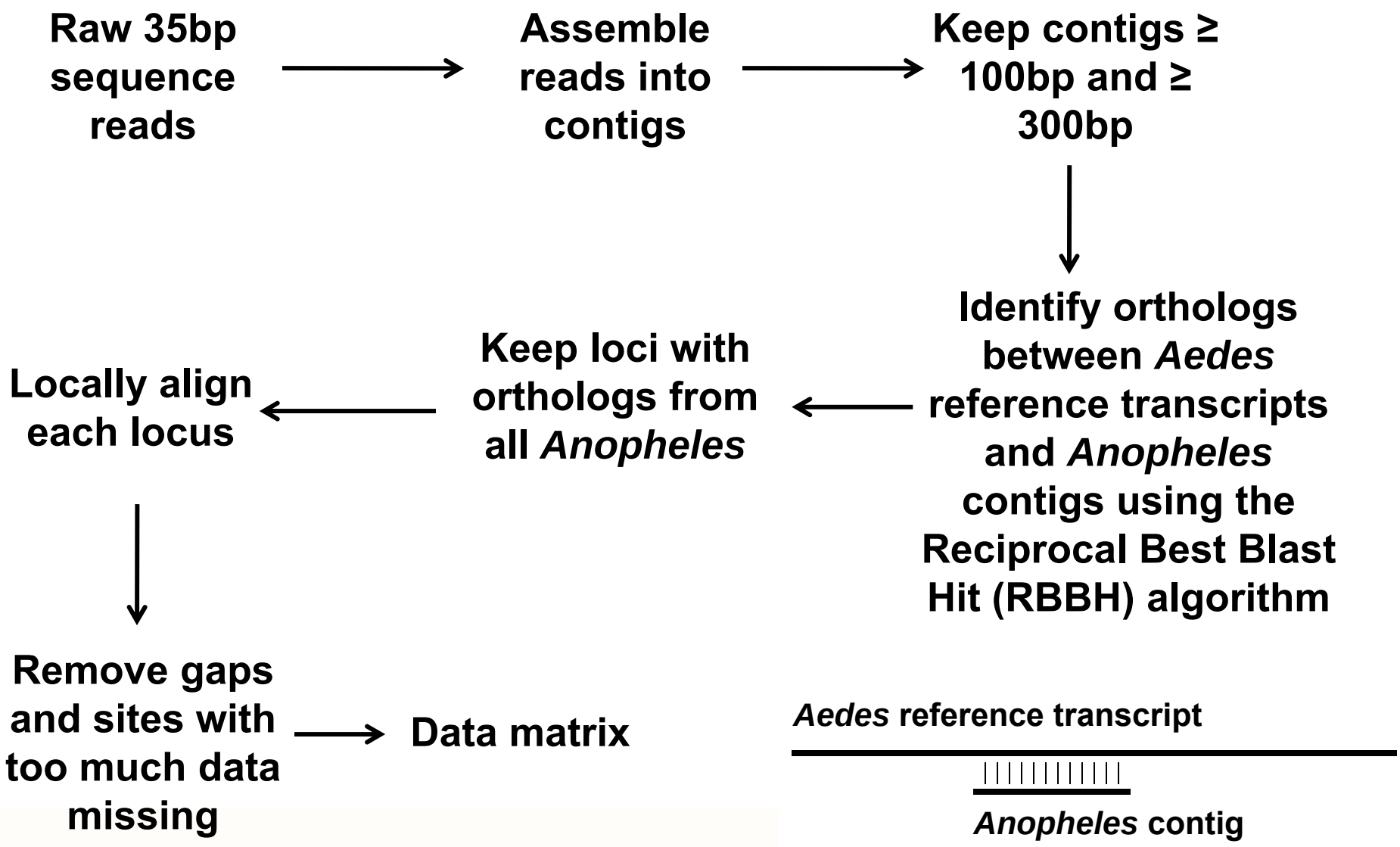
Rokas & Abbot (2009) Trends Ecol. Evol.

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

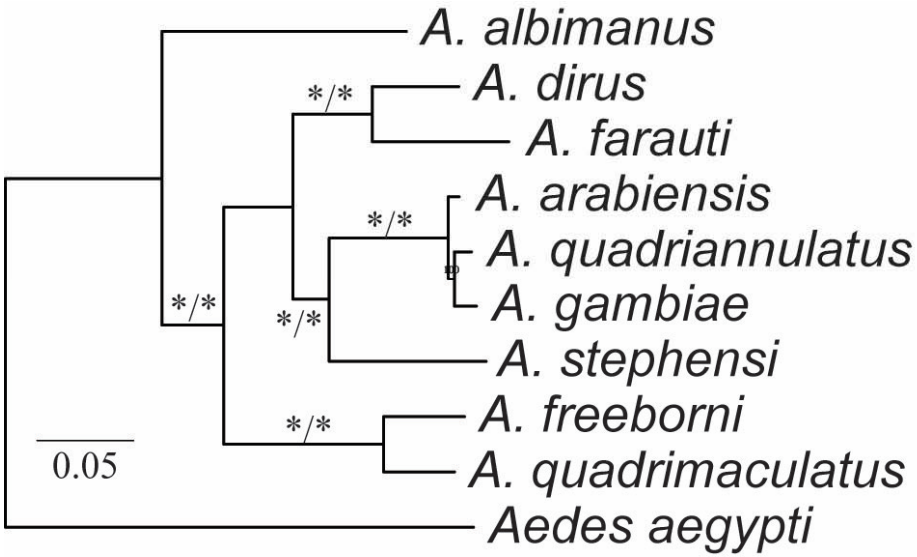
Illumina RNA-Seq → **~150,000,000 reads**
5,250,000,000 bp

Data Matrix Construction: The "Singlecontig" Strategy



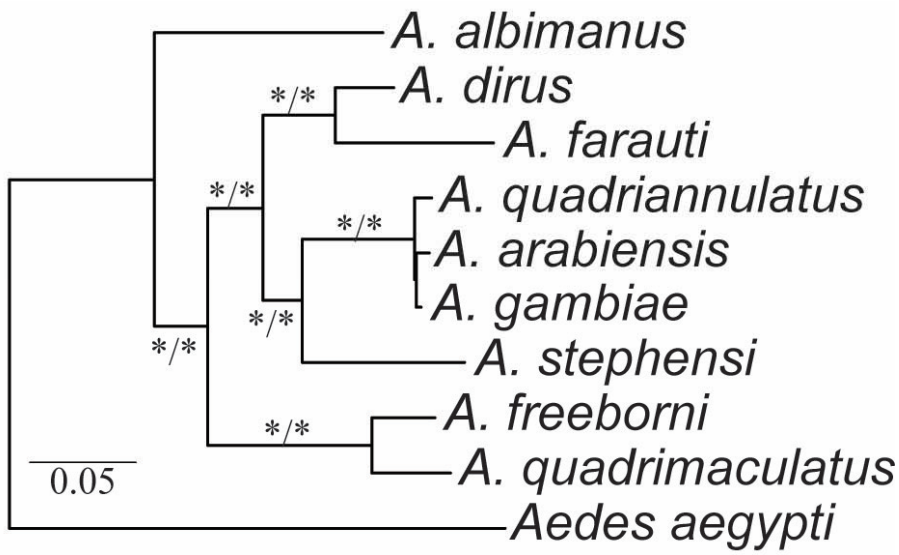
Robust Phylogenetic Inference from RNA-Seq Data

Using ≥ 100bp contigs



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

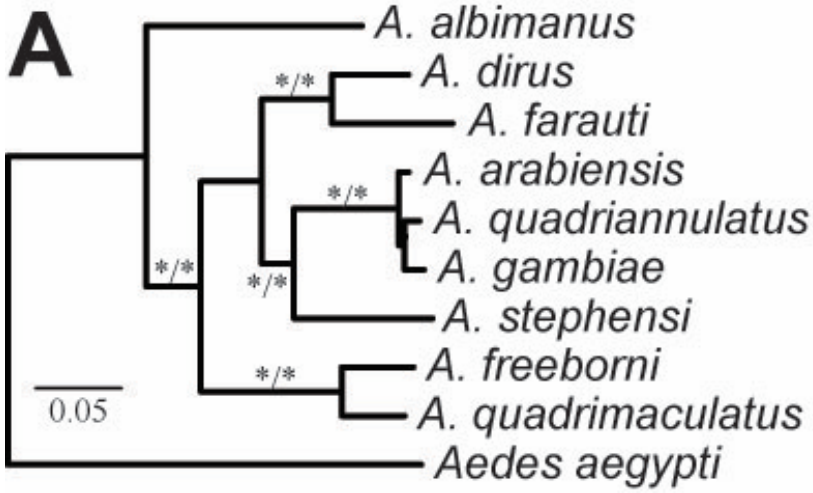
Using ≥ 300bp contigs

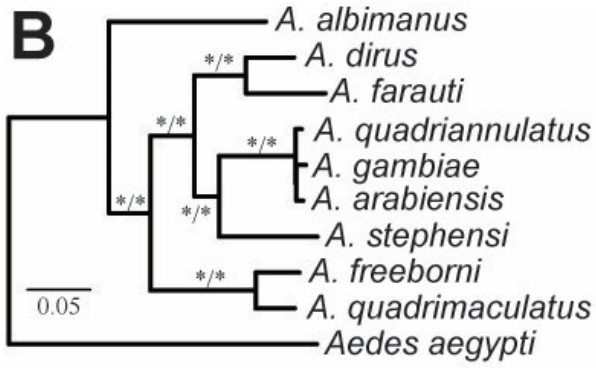


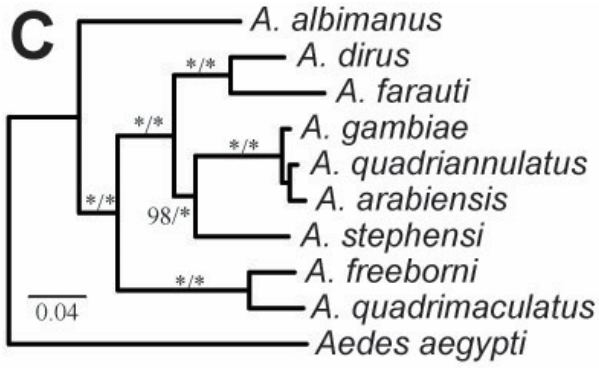
 # Loci = 69
Aln Length = ~73 Kb
% Missing data = 44

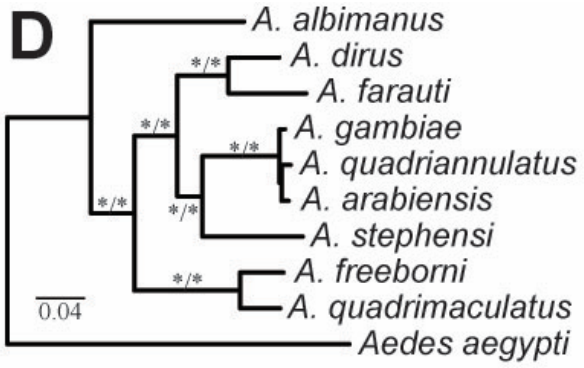
Accurate Phylogenetic Inference From our Data

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









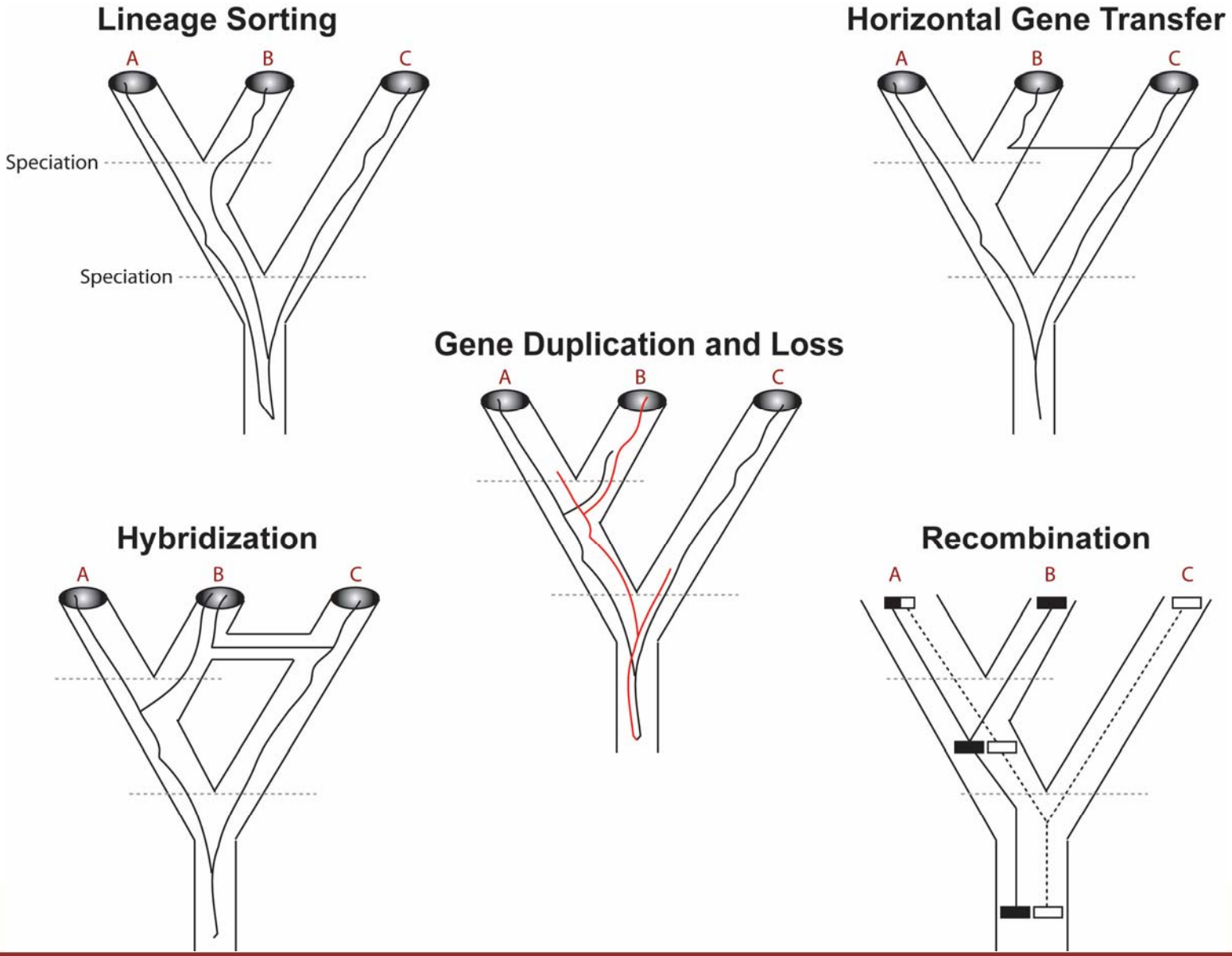
<u>Exclude erroneous loci</u>
491 loci
Aln L: ~329 Kb
Missing data: 50%

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

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

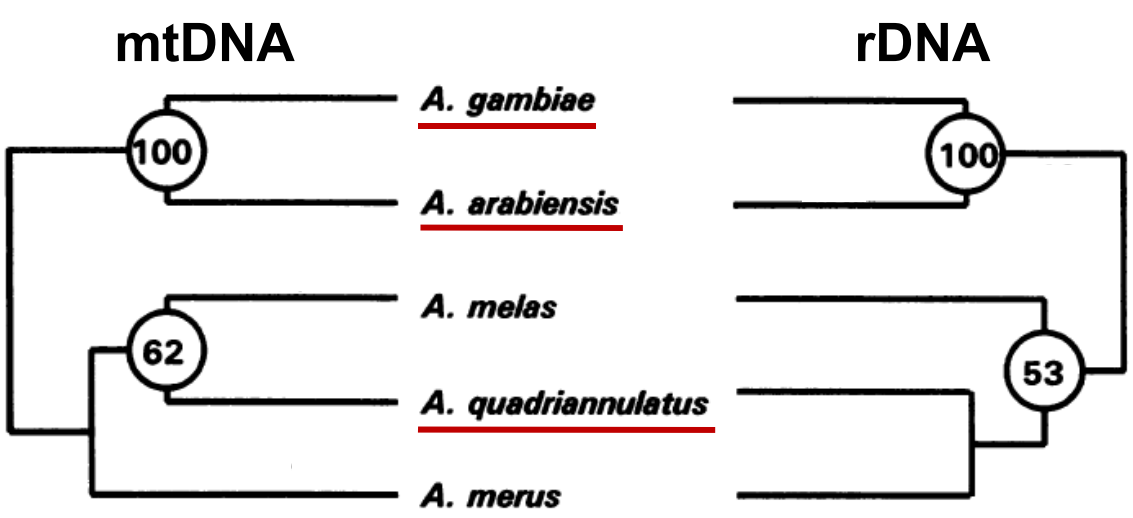
Hittinger et al. (2010) PNAS

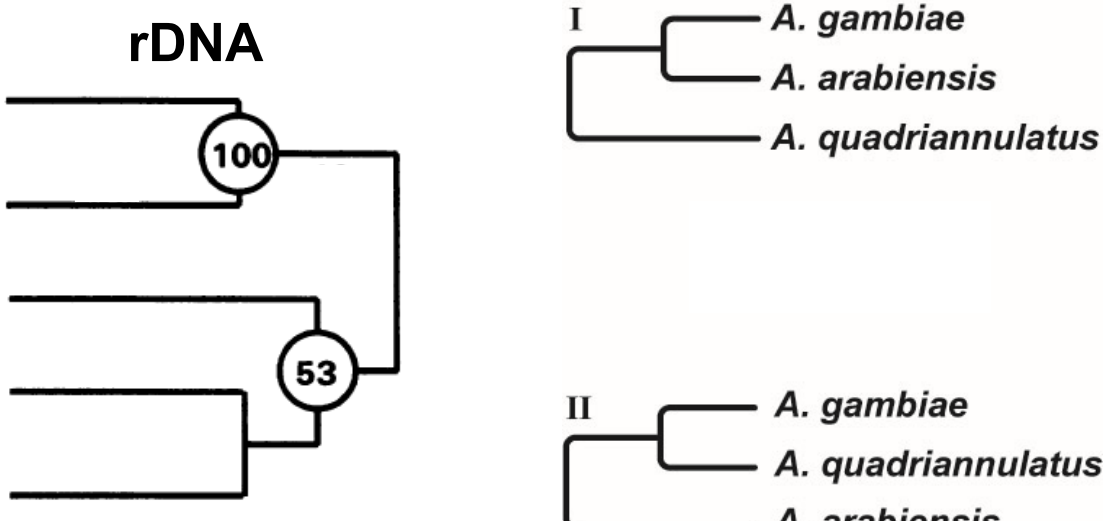
Gene Trees Can Differ from Species Trees

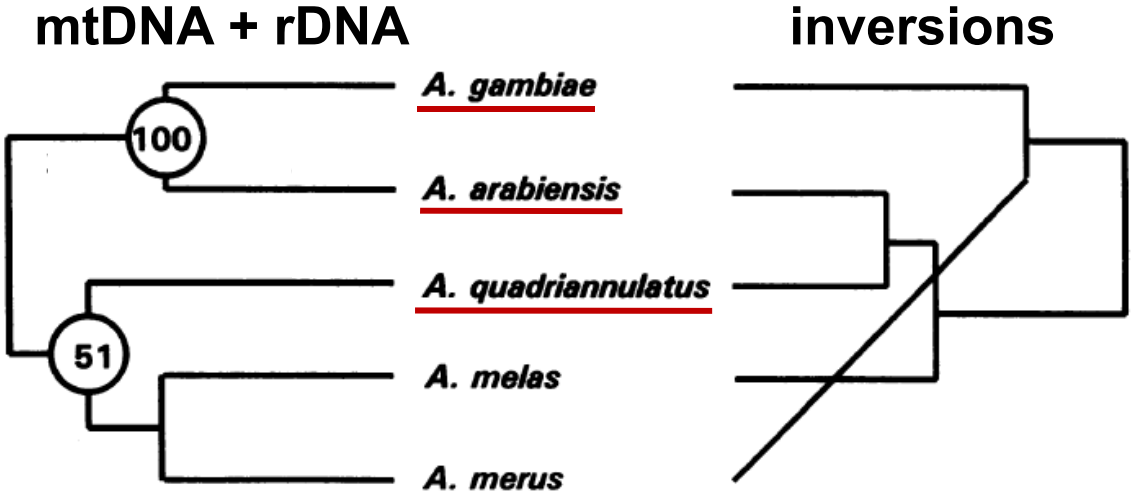


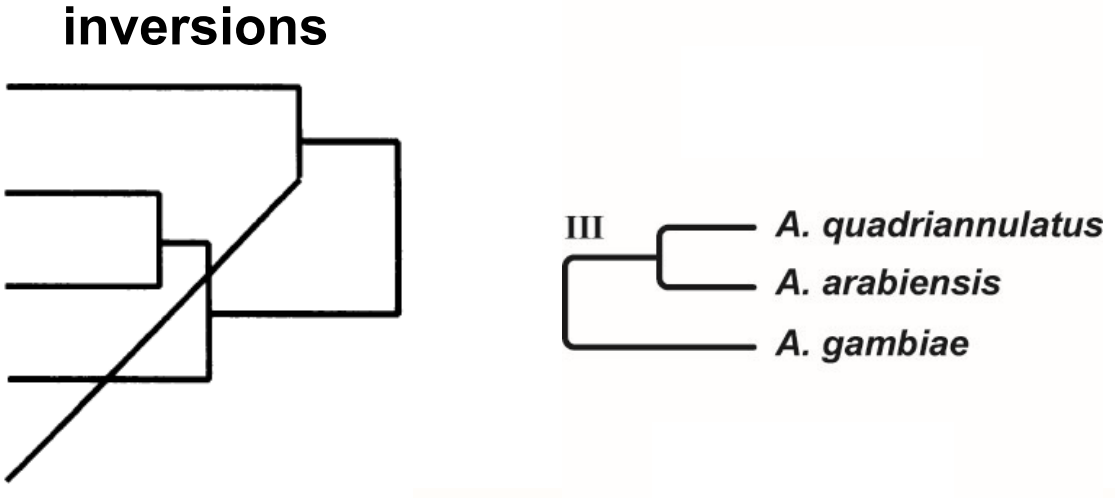
Degnan & Rosenberg (2009) Trends Ecol. Evol.

Our Data Matrices Can Detect Population-Level Events





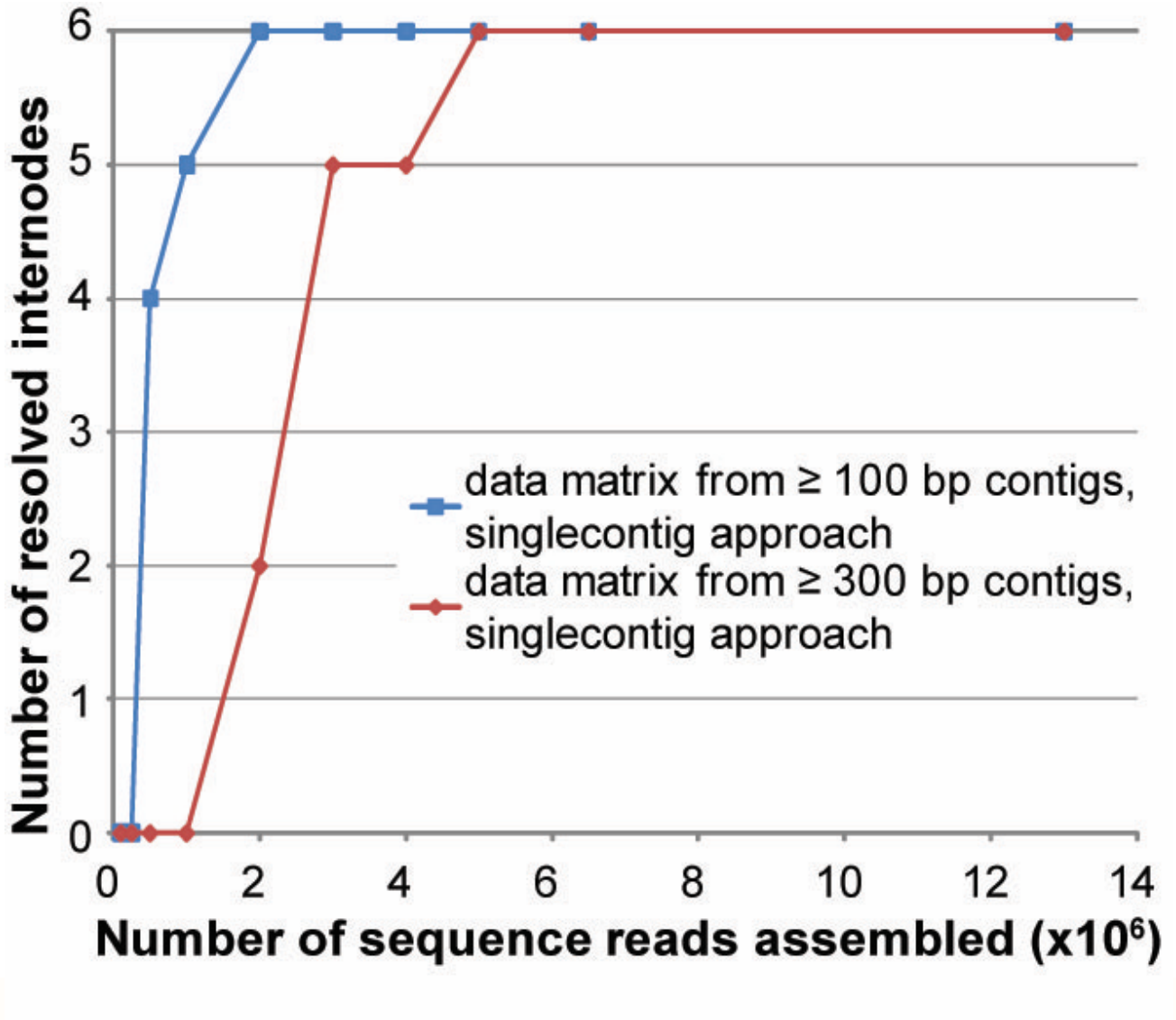




Besansky et al. (1994) PNAS

Hittinger et al. (2010) PNAS

Robust Phylogenetic Inference From Few Sequence Reads



Hittinger et al. (2010) PNAS

Experimental Design: The "Supercontig" Strategy

Raw 35bp sequence reads → Assemble reads into contigs → Keep contigs ≥ 100bp and ≥ 300bp

↓

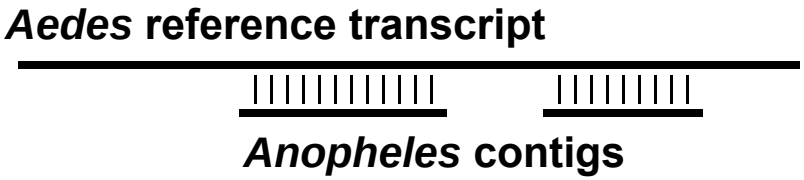
Identify orthologs between *Aedes* reference transcripts and *Anopheles* contigs using a locally relaxed RBBH algorithm

← Keep only loci with orthologs from ≥ 4 *Anopheles*

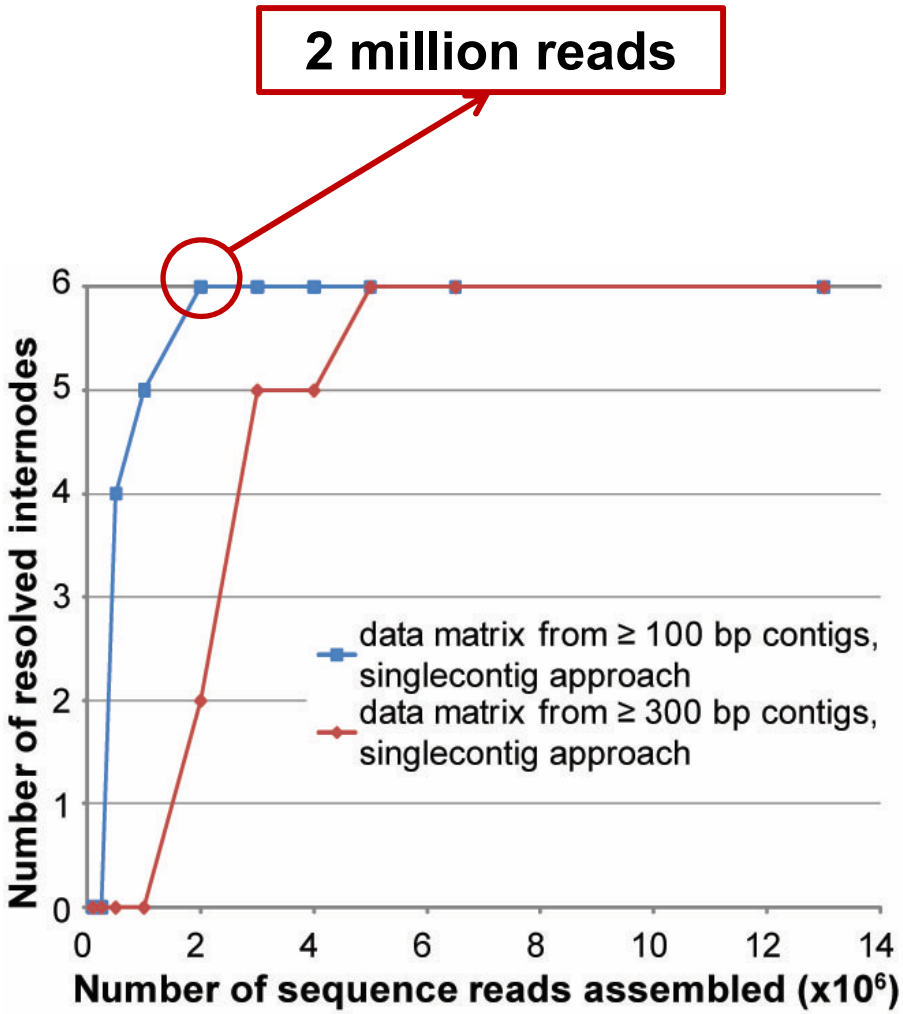
← Locally align each locus

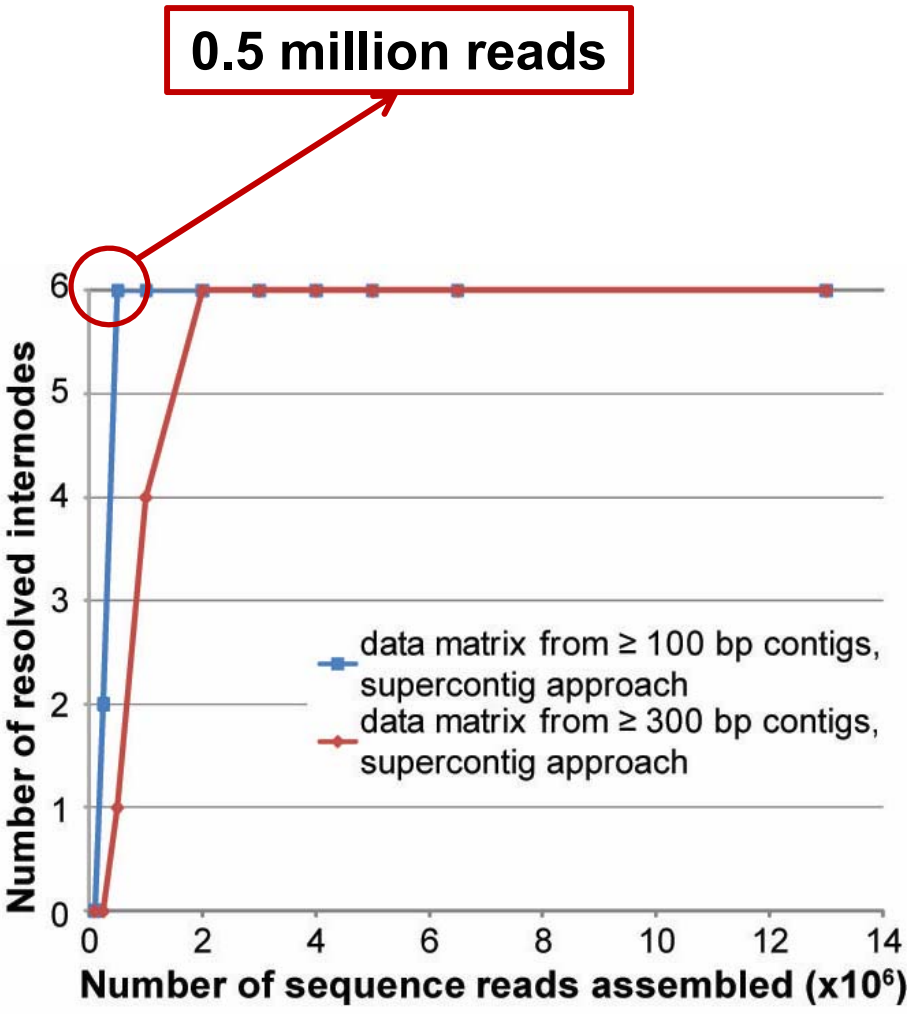
↓

Remove gaps and sites with too much data missing → Data matrix



Robust Phylogenetic Inference From Very Few Reads





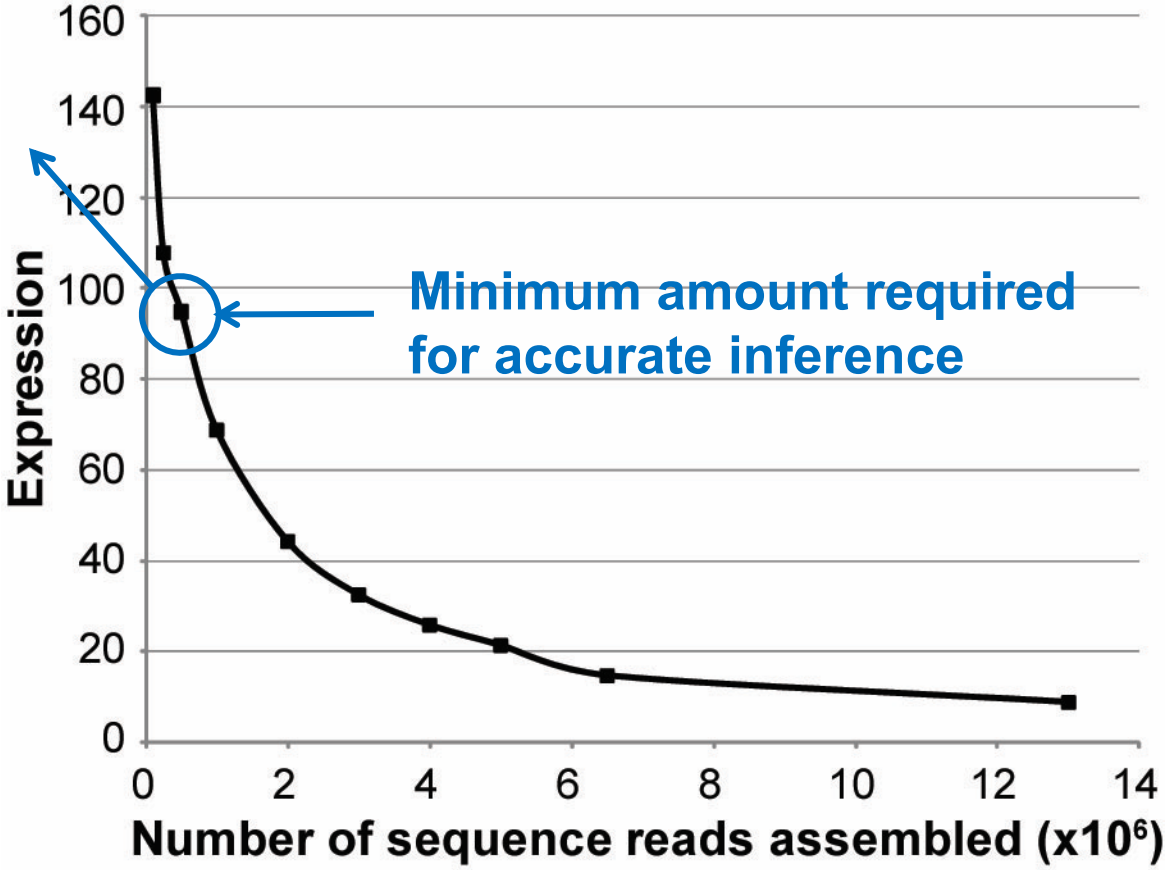
553 loci, AlnL: ~390Kb, %miss: 51

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

Hittinger et al. (2010) PNAS

Our Sequences are from Highly-Expressed Transcripts

2008 cost: ~$50
2014 cost: < $5



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

nature

LETTERS

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

Syst. Biol. 57(6):920–938, 2008

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. CUNNINGHAM[3]

Toward Resolving t
Tree: The Phylogen
of Jakobids and Cercozoans

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**

*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

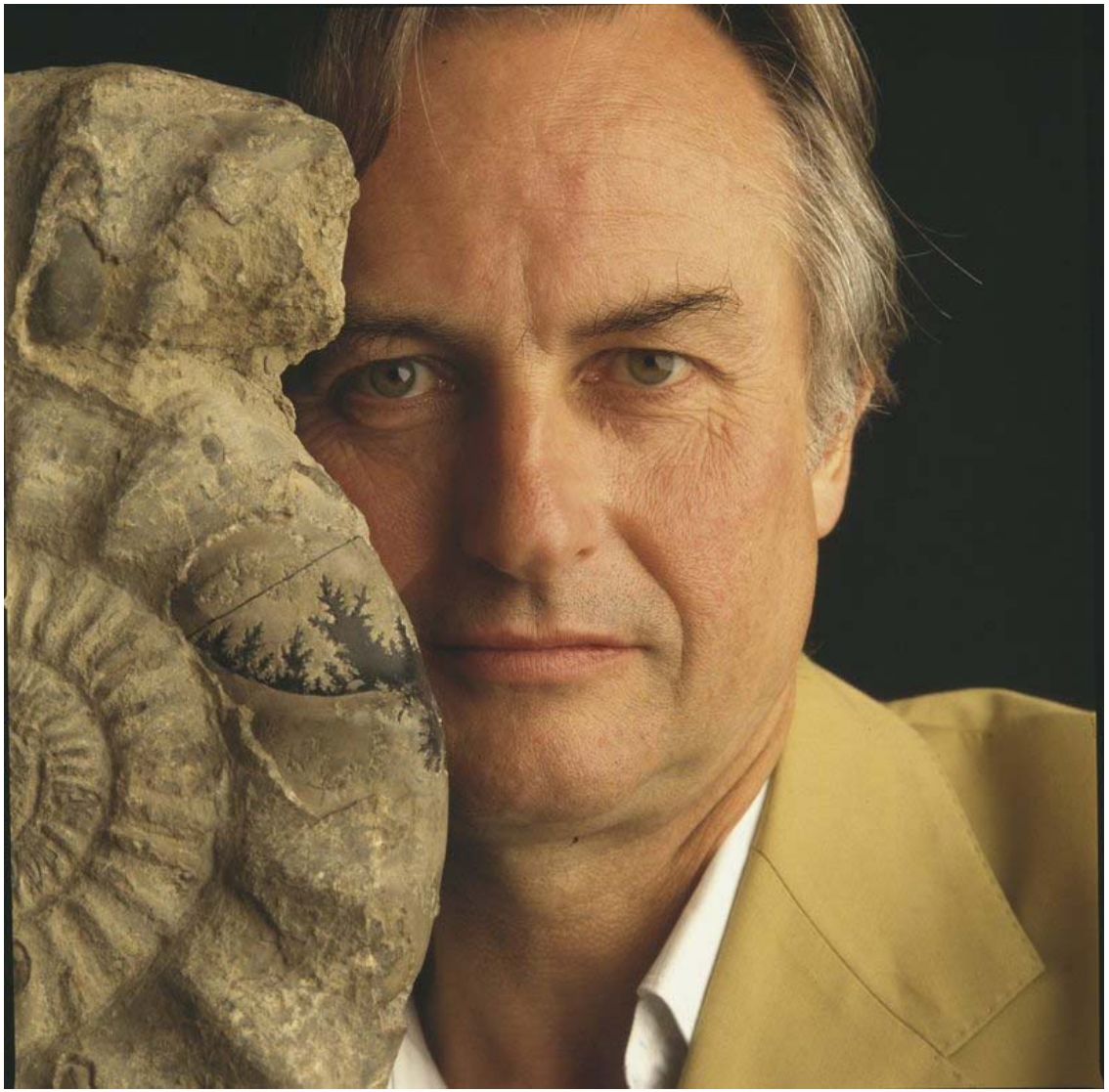
Samuli Lehtonen

Department of Biology, U

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



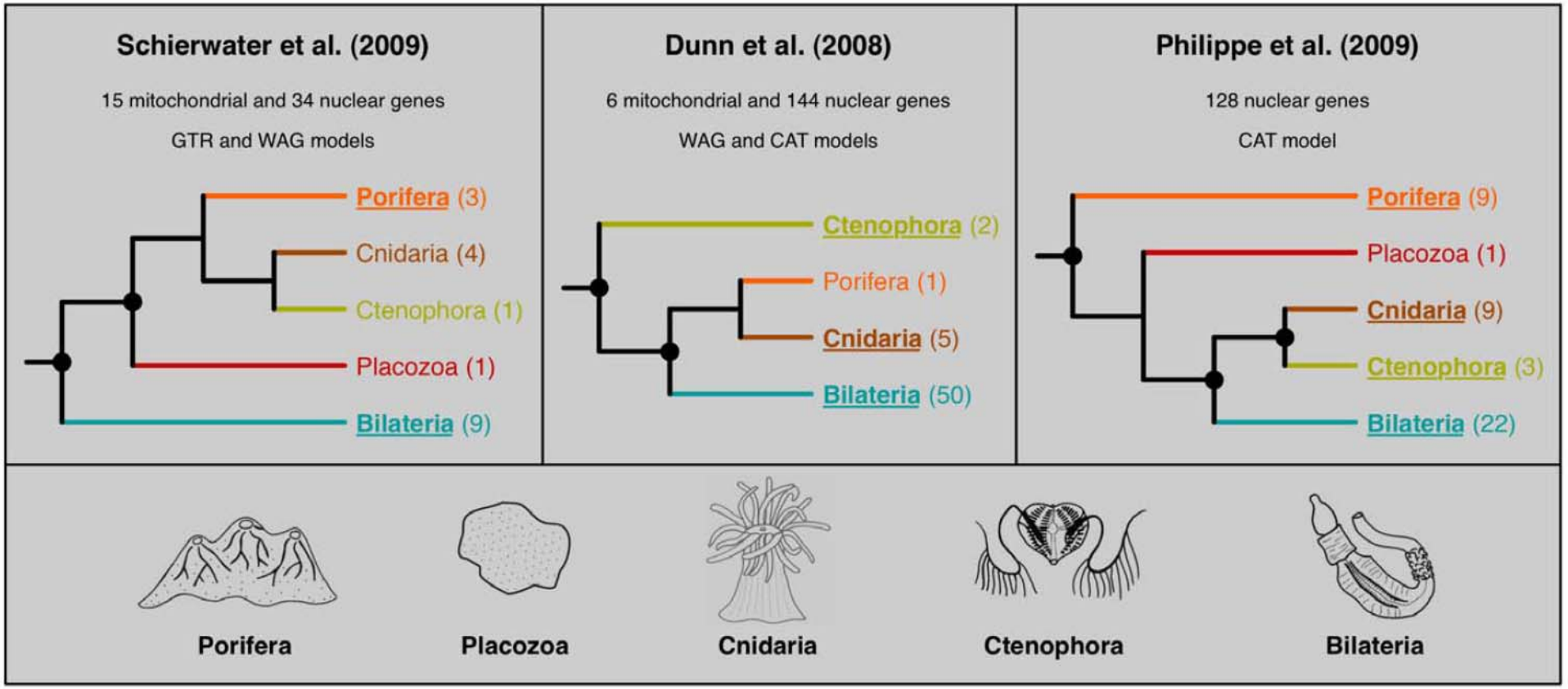
Dawkins (2003), A Devil's Chaplain

Incongruence in Deep Time

Kocot et al. (2011) Nature

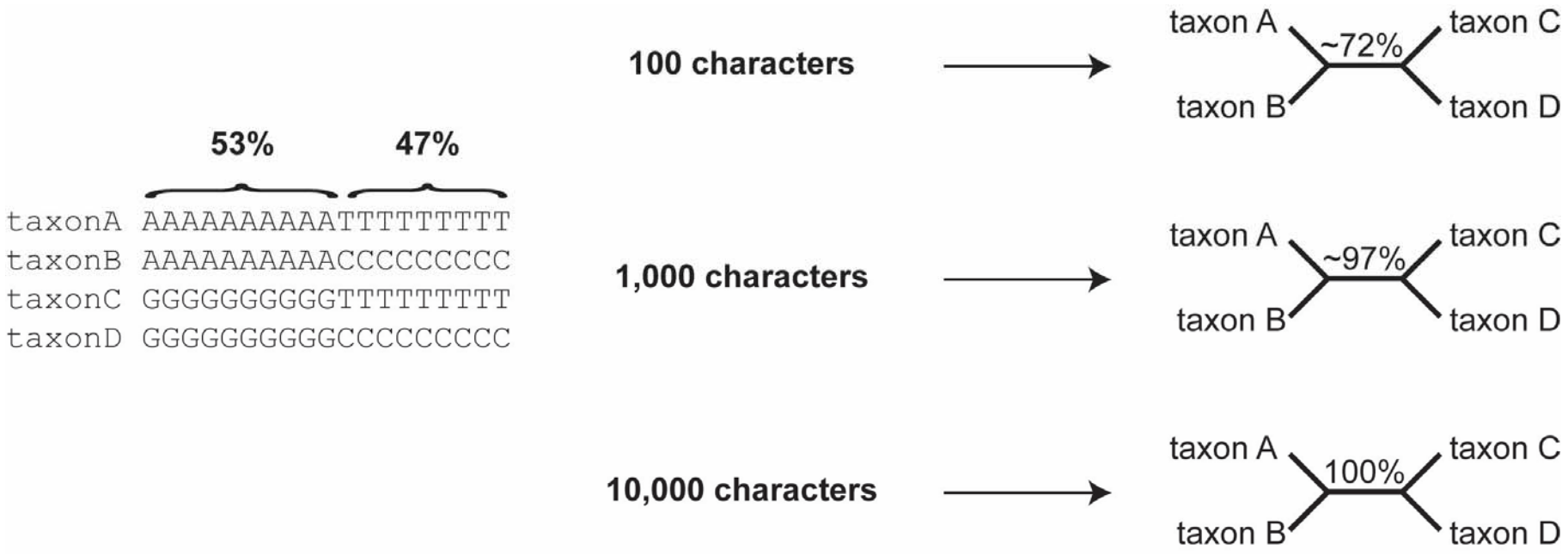
Smith et al. (2011) Nature

Incongruence in Deep Time



Philippe et al. (2011) PLoS Biol.

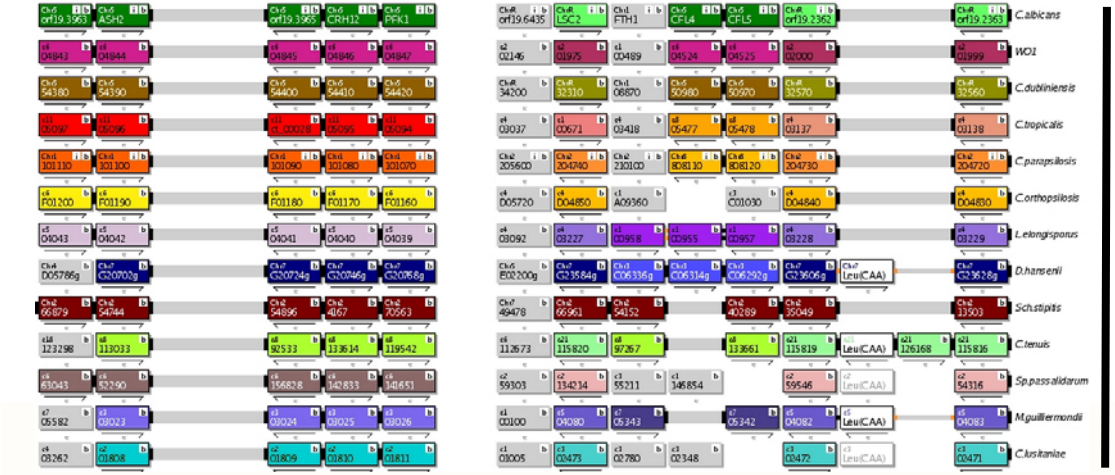
Bootstrap Support is Misleading When Used in Large Datasets



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

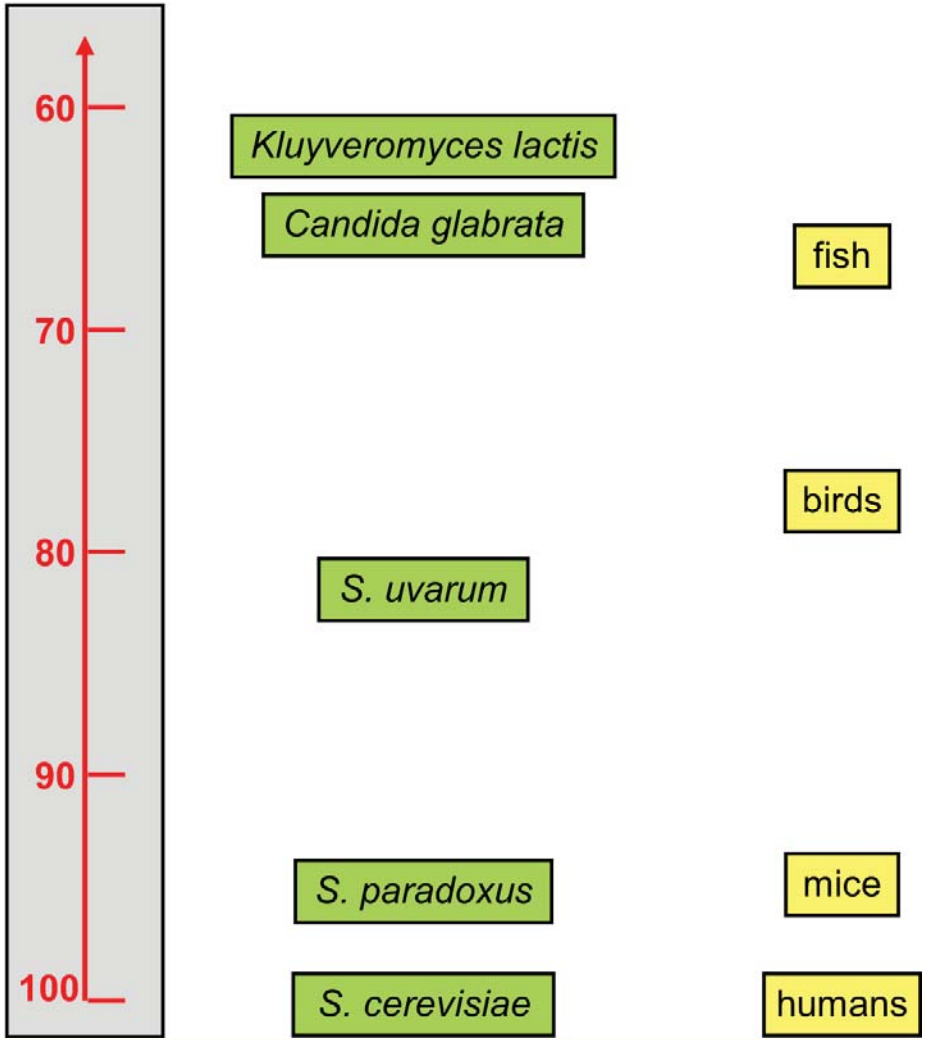
Byrne & Wolfe (2005) Genome Res.

Fitzpatrick et al. (2010) BMC Genom.

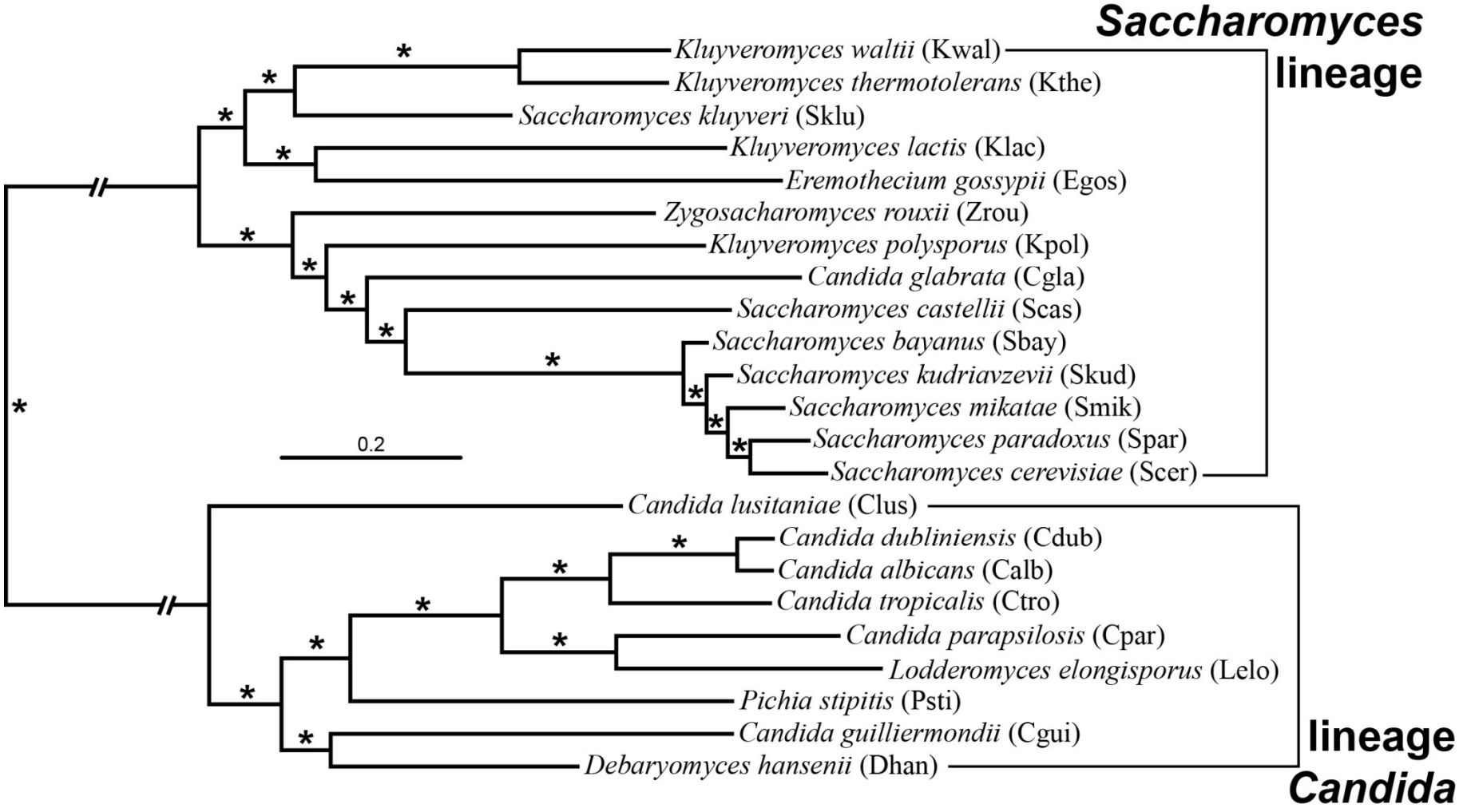
Yeast Divergence is Intermediate to Vertebrates and Animals

Proteome-wide average pairwise amino acid sequence similarity

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

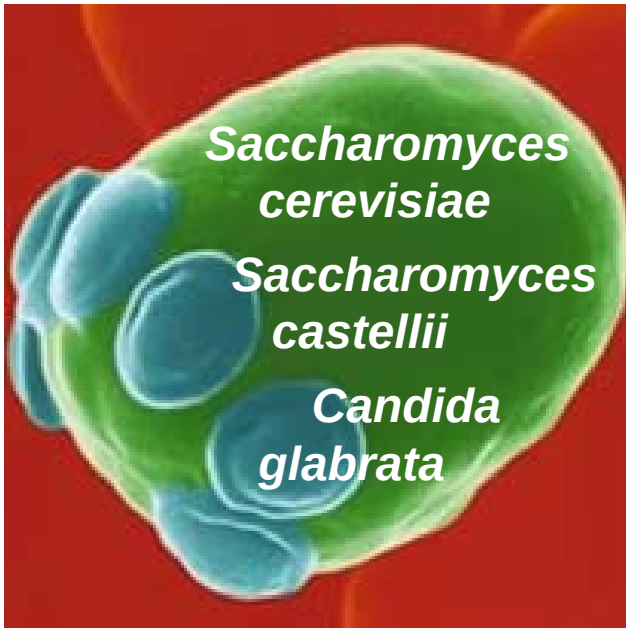


Concatenation Yields an Absolutely Supported Phylogeny

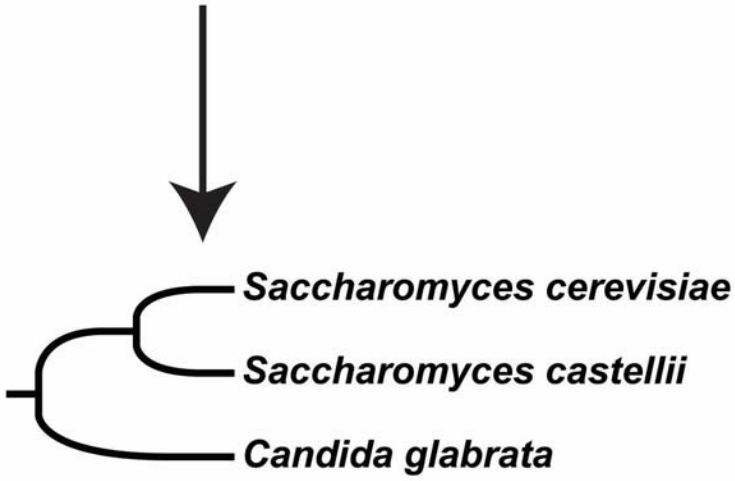


Salichos & Rokas (2013) Nature

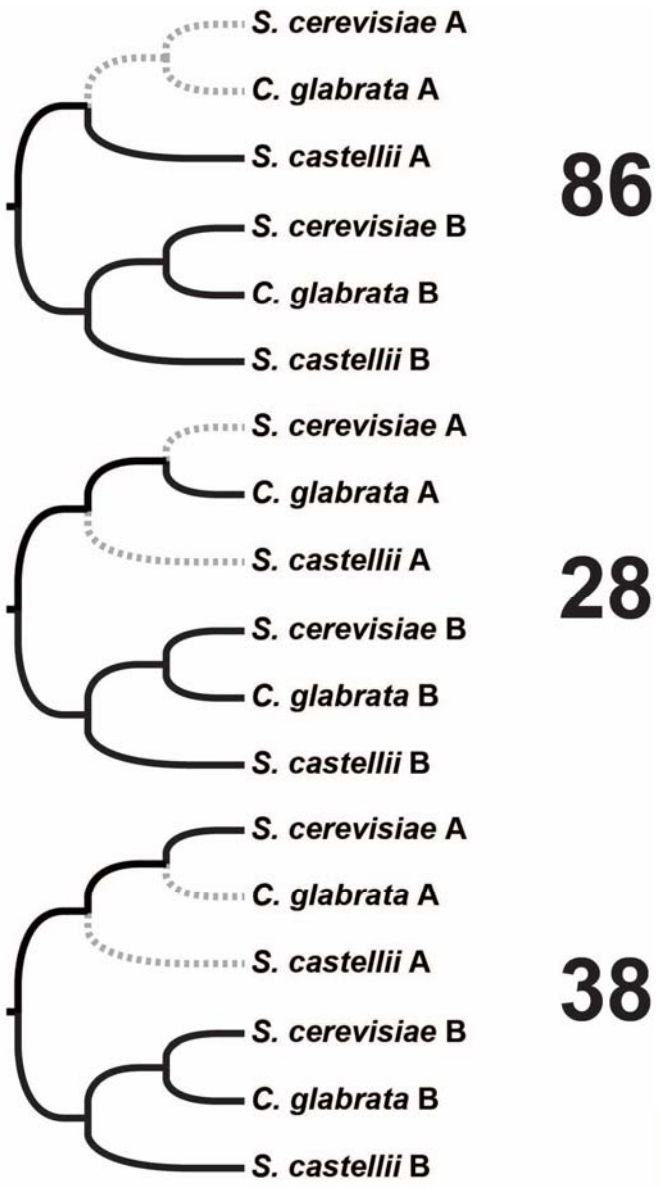
The Concatenation Phylogeny is at Least Partly Wrong



Linear Sequence Data

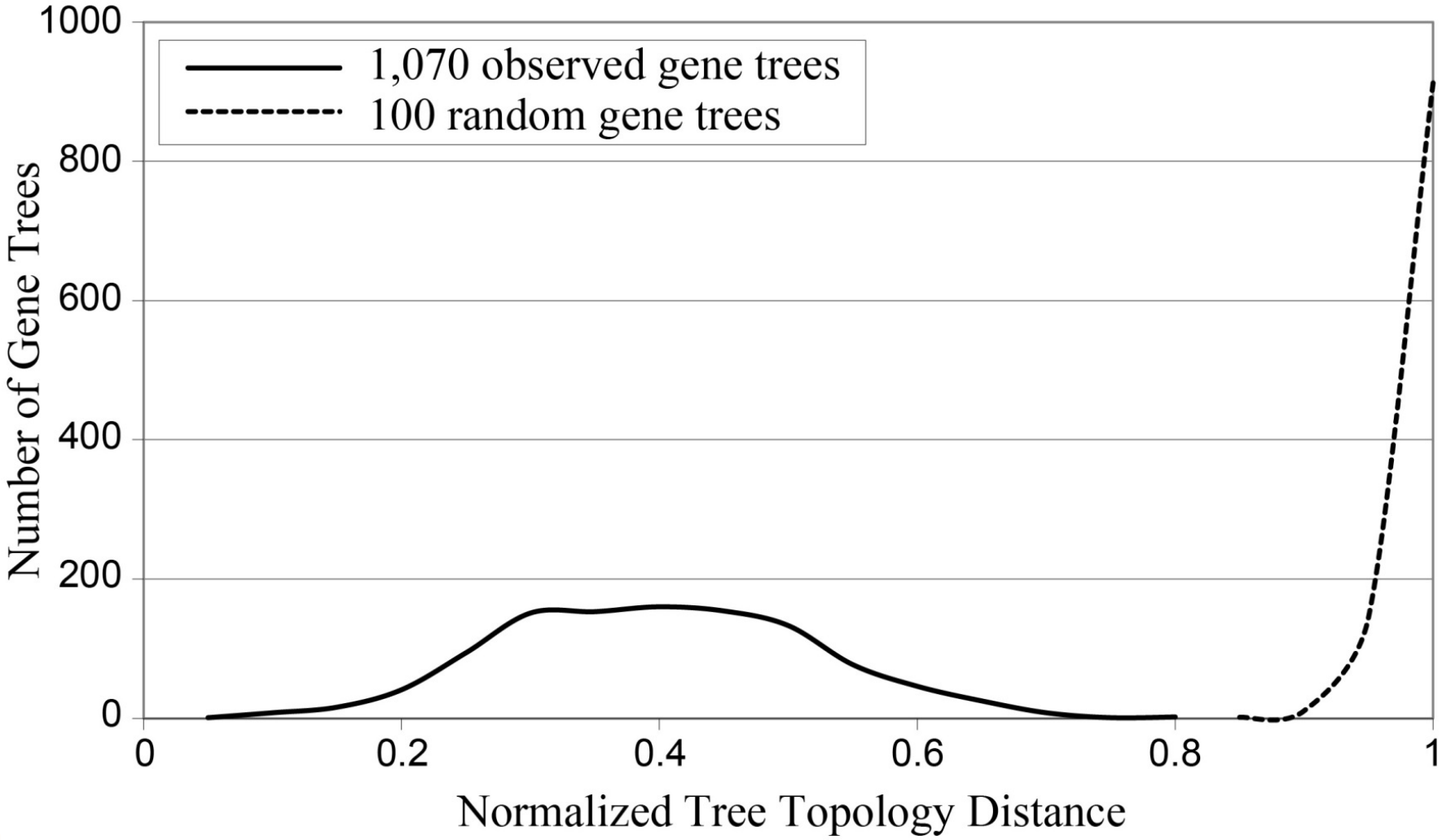


Rare Genomic Changes

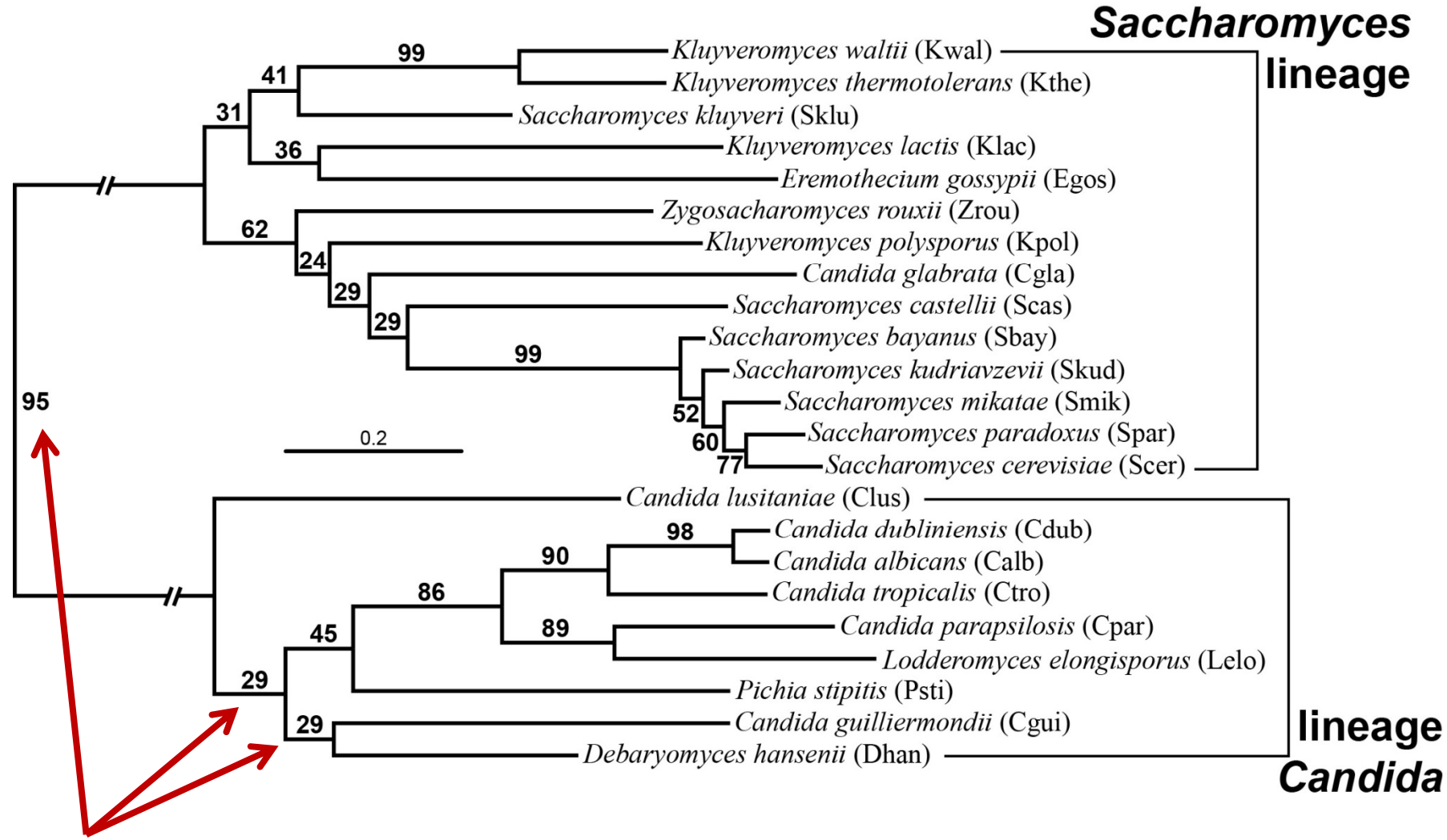


Scannell et al.(2006) Nature

All Gene Trees Differ from the Concatenation Phylogeny

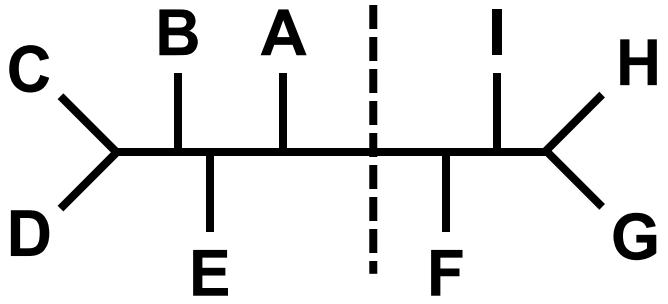


The Yeast Phylogeny Inferred by Majority-Rule Consensus

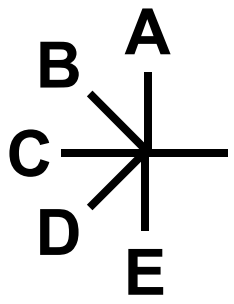


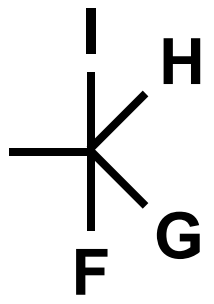
Phylogenetic Trees are Sets of Splits

Division

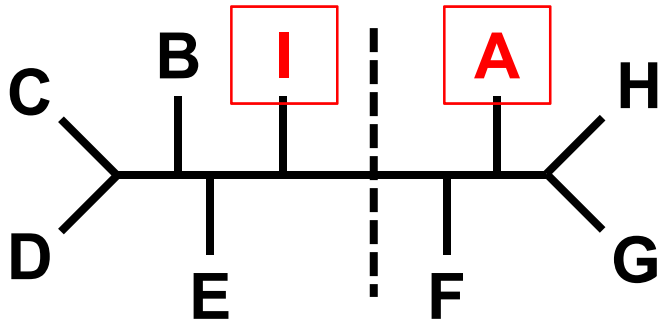


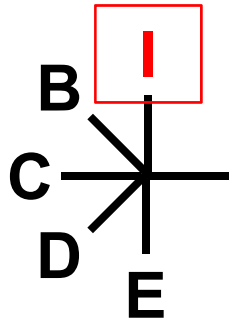
Splits

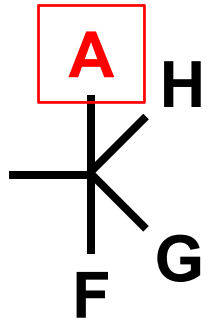




{A, B, C, D, E} **{F, G, H, I}**

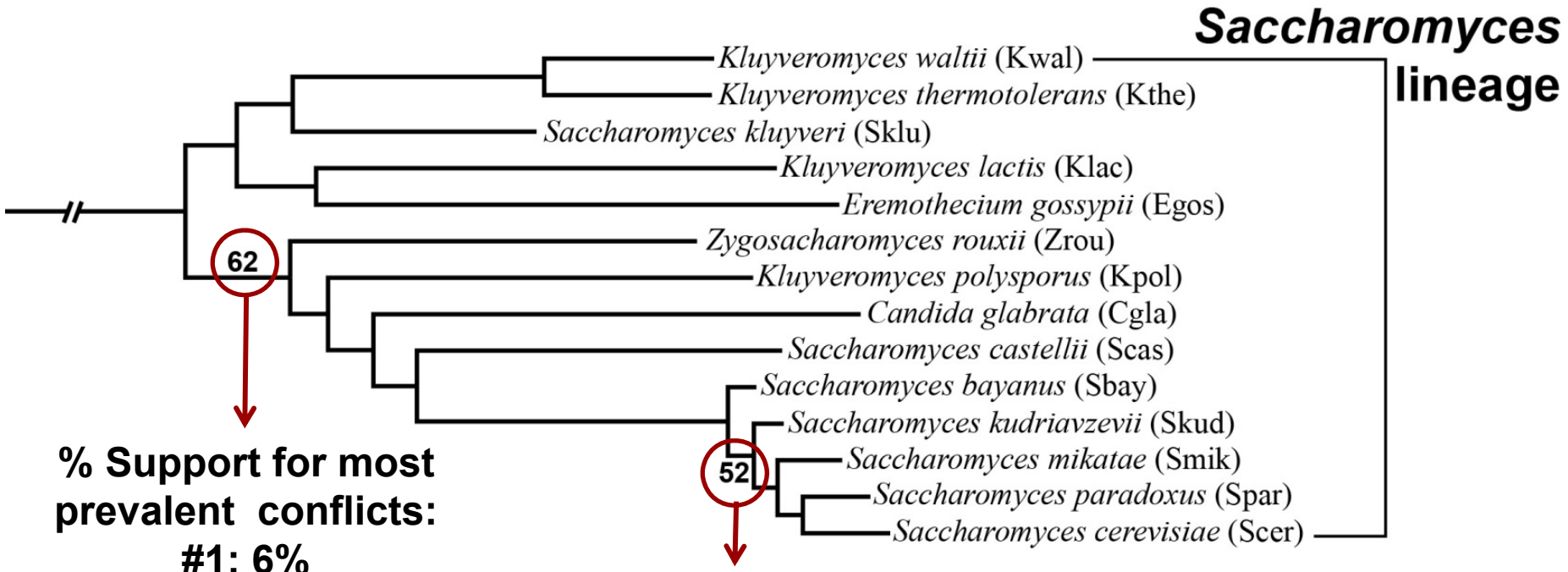






Conflicting **Splits: {I, B, C, D, E}** **{F, G, H, A}**

Measuring the Degree of Conflict for each Internode

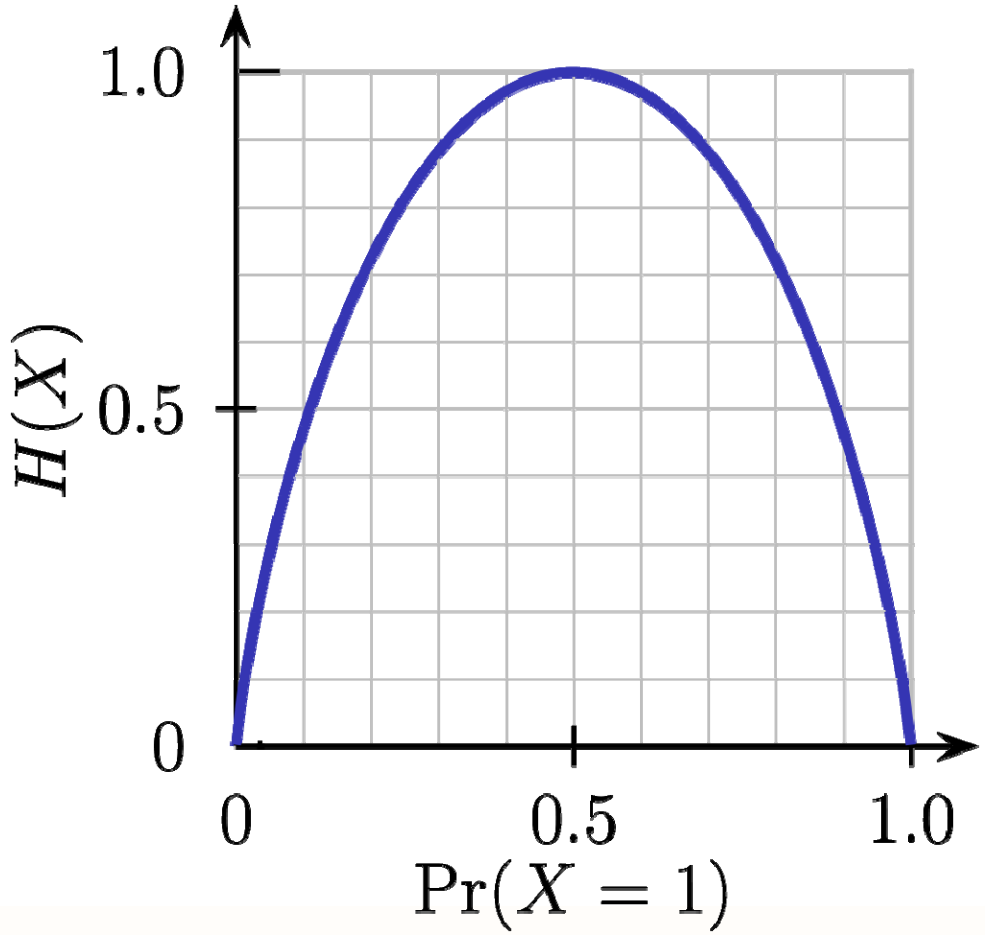


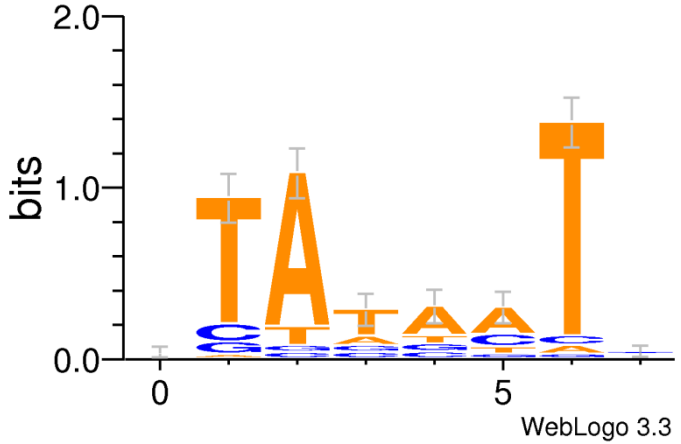
% Support for most prevalent conflicts:
#1: 6%
#2: 5%

% Support for most prevalent conflicts:
#1: 29%
#2: 14%
#3: 11%
#4: 11%
#5: 08%

Shannon's Entropy Measures the Uncertainty in a Variable

$$H(X) = -\sum_{1}^{n} p(X_n) log_2\big(p(X_n)\big)$$

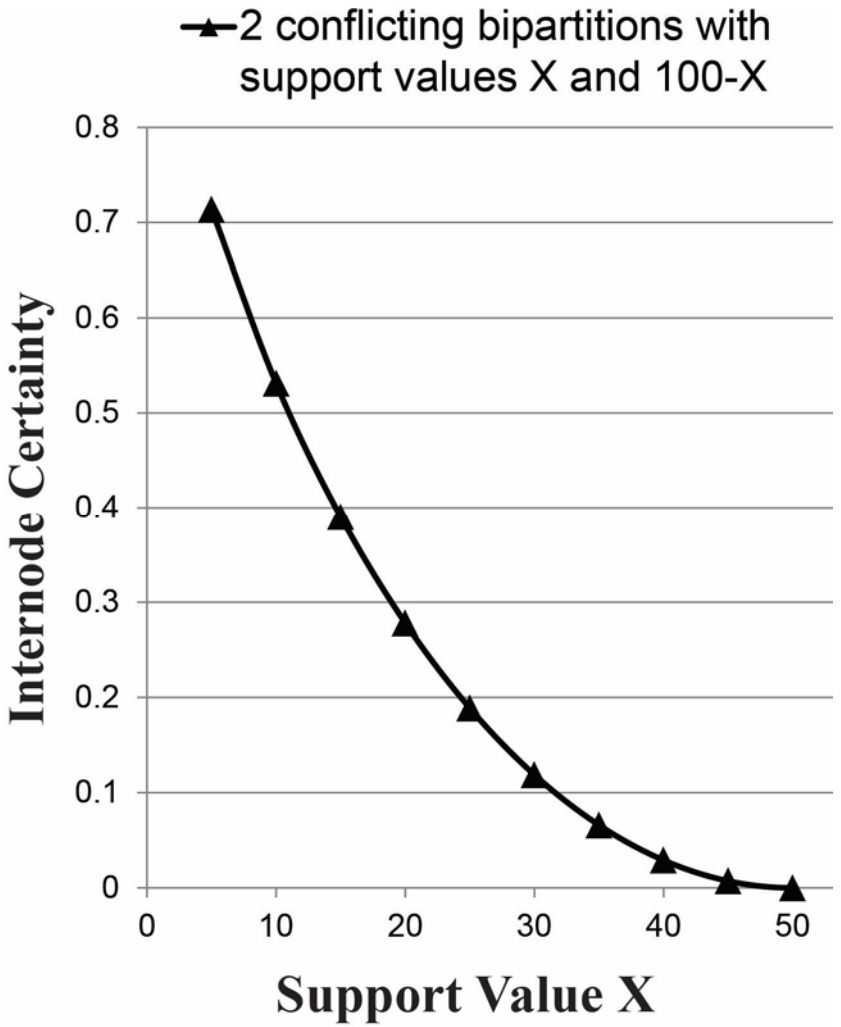




Internode Certainty, a New Metric to Quantify Incongruence

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)

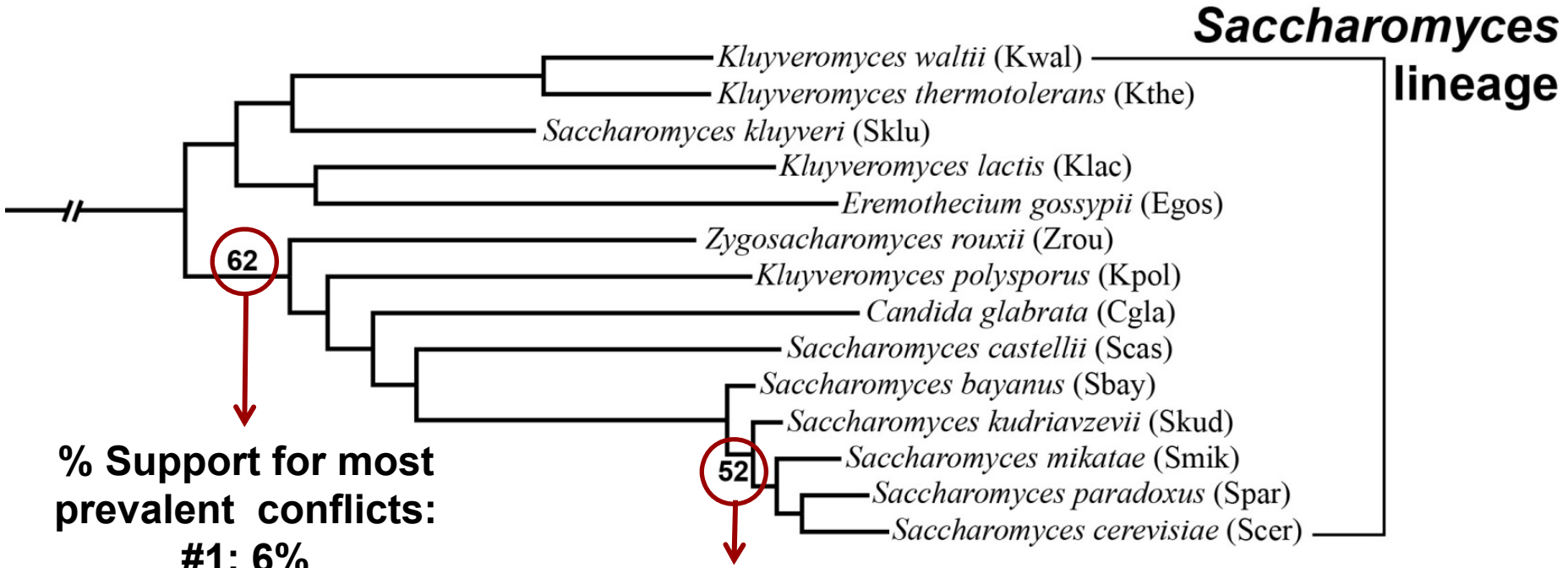
```
[rokasa@vmps65]$ more RAxML_verboseSplits.T6
Cint
Dmel
Sman
…

                                    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



% Support for most prevalent conflicts:
#1: 6%
#2: 5%

IC value: 0.59

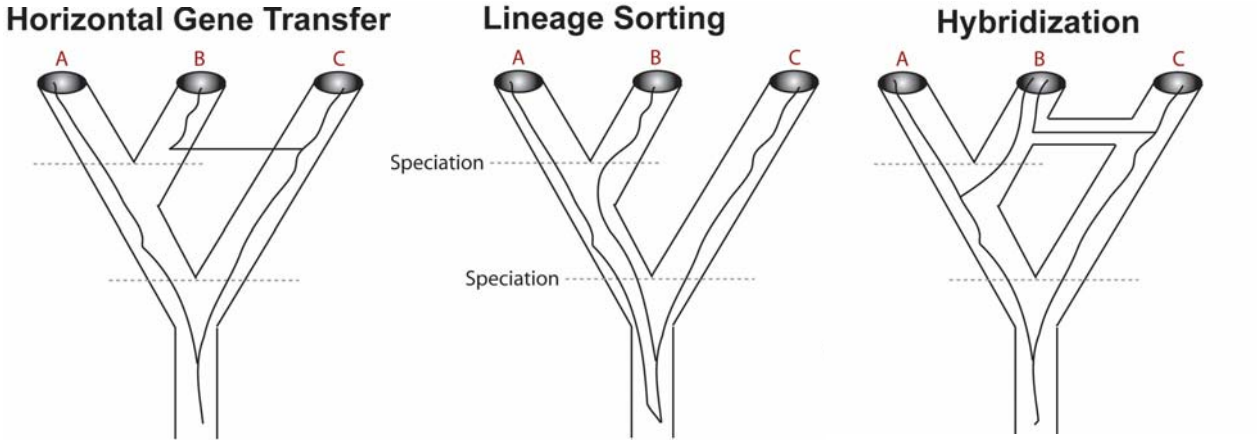
% Support for most prevalent conflicts:
#1: 29%
#2: 14%
#3: 11%
#4: 11%
#5: 08%

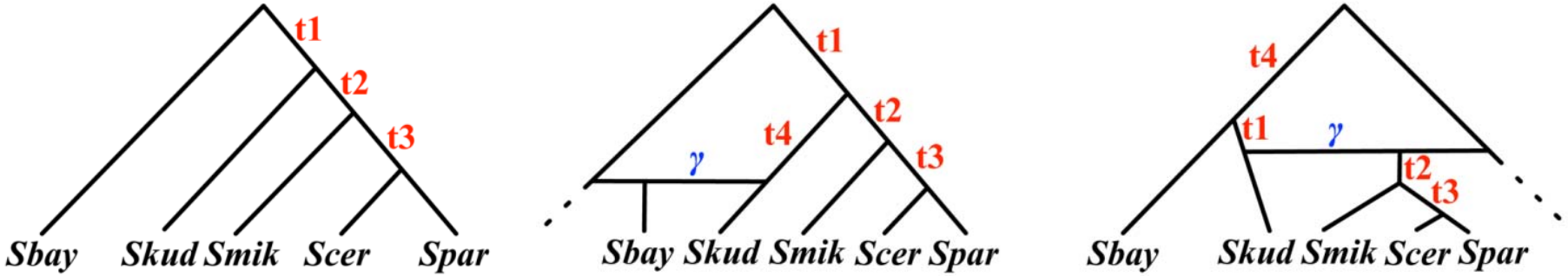
IC value: 0.06

Salichos & Rokas (2013) Nature

Why is the Yeast Phylogeny Hard to Resolve?

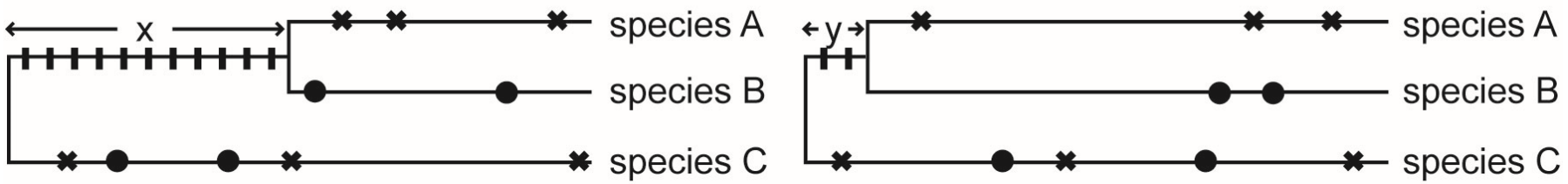
- ***Biological factors***



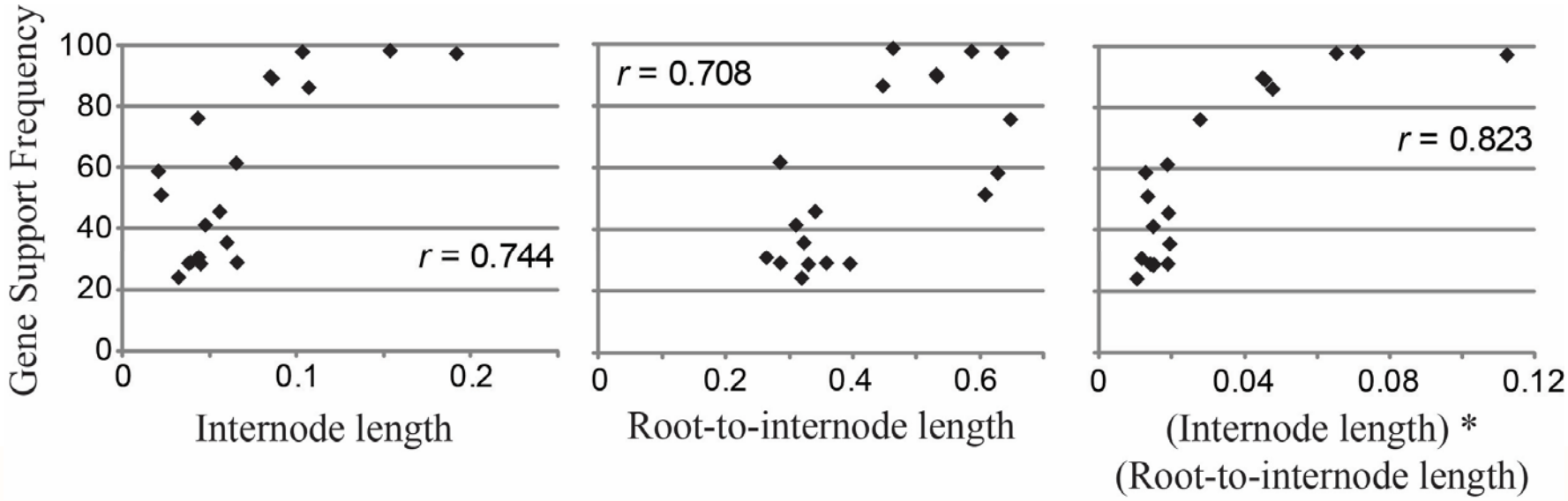


Why is the Yeast Phylogeny Hard to Resolve?

❖ *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**
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**

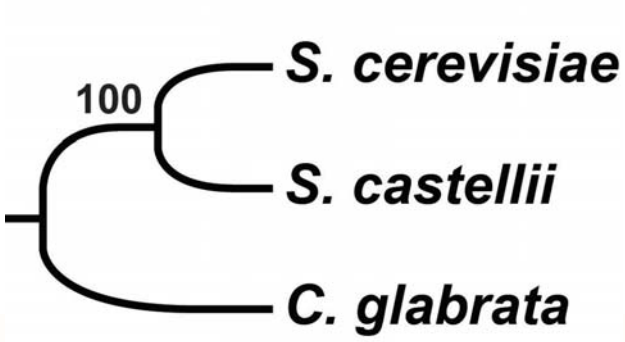
Salichos & Rokas (2013) Nature


WHAT CAN I DO?!!

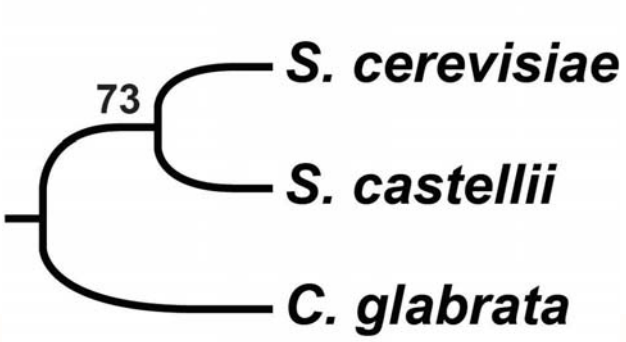
What Do We Do Then?

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**

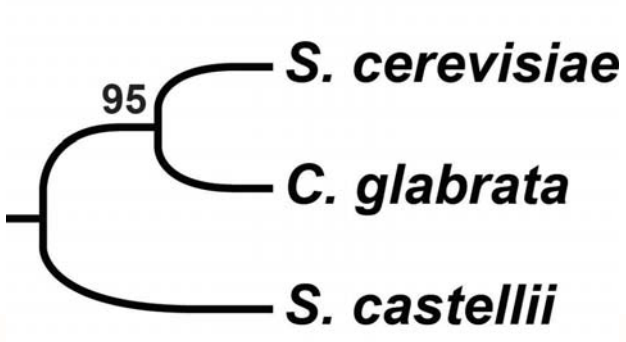
average BS ≥60%



average BS ≥70%

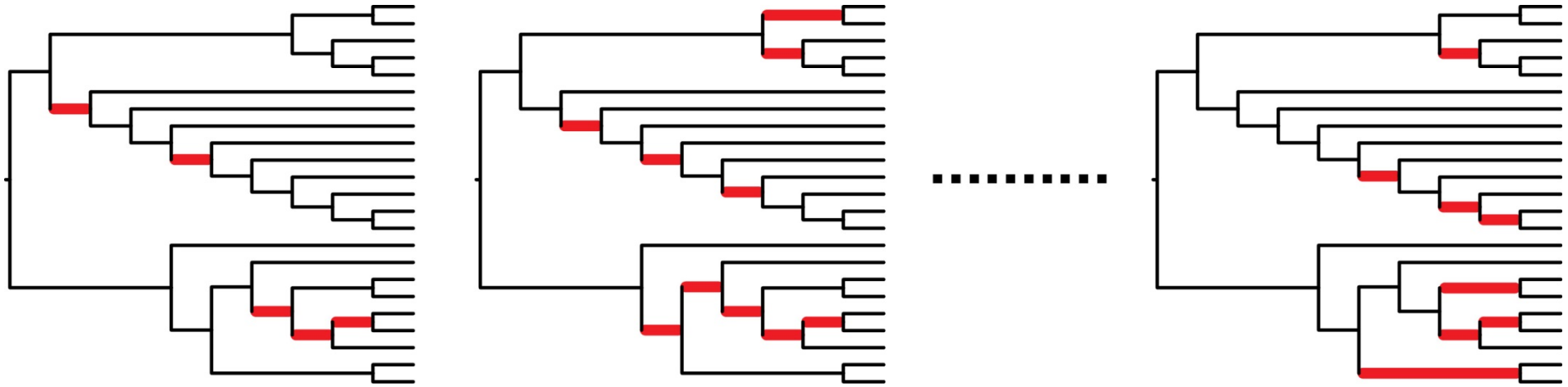


average BS ≥80%



Salichos & Rokas (2013) Nature

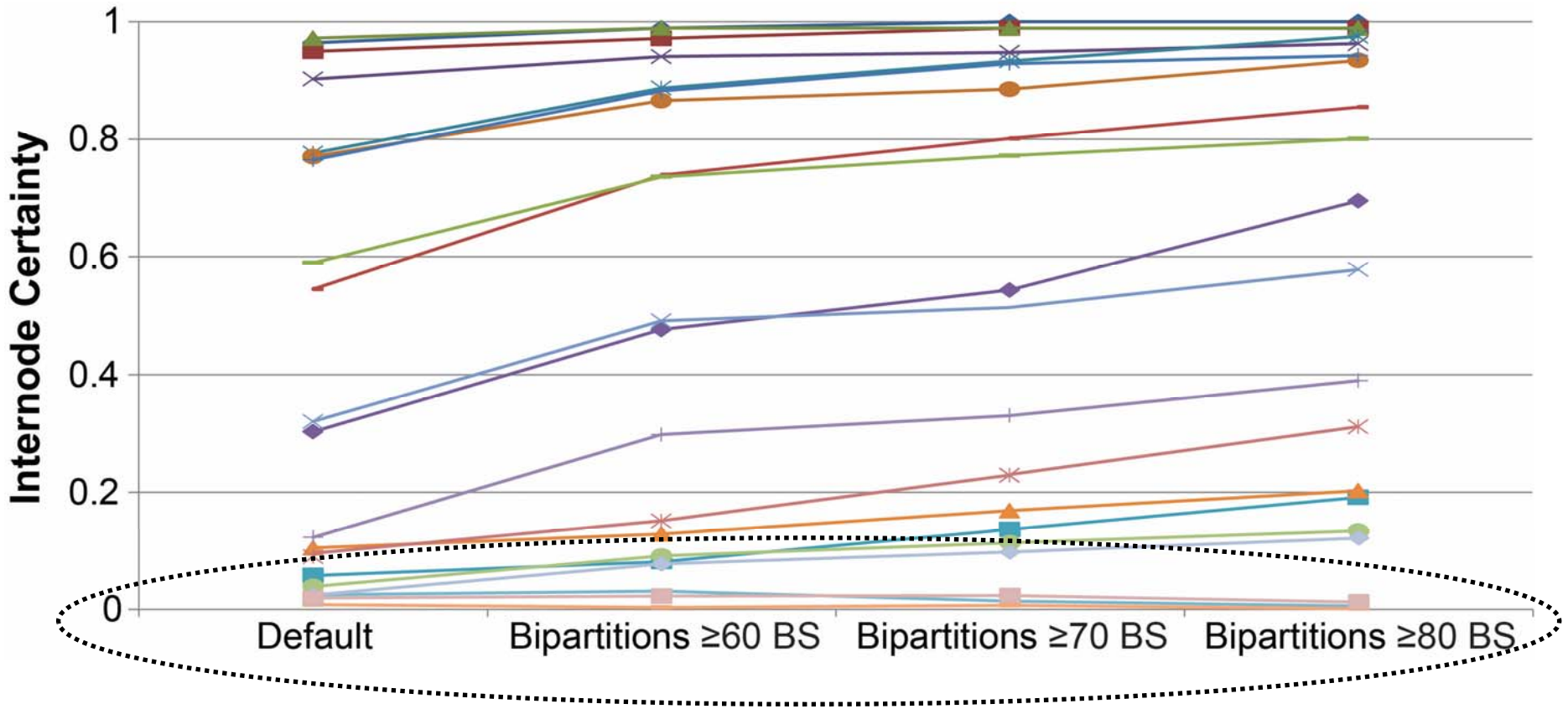
Selecting Specific Bipartitions Dramatically Improves Phylogeny



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**

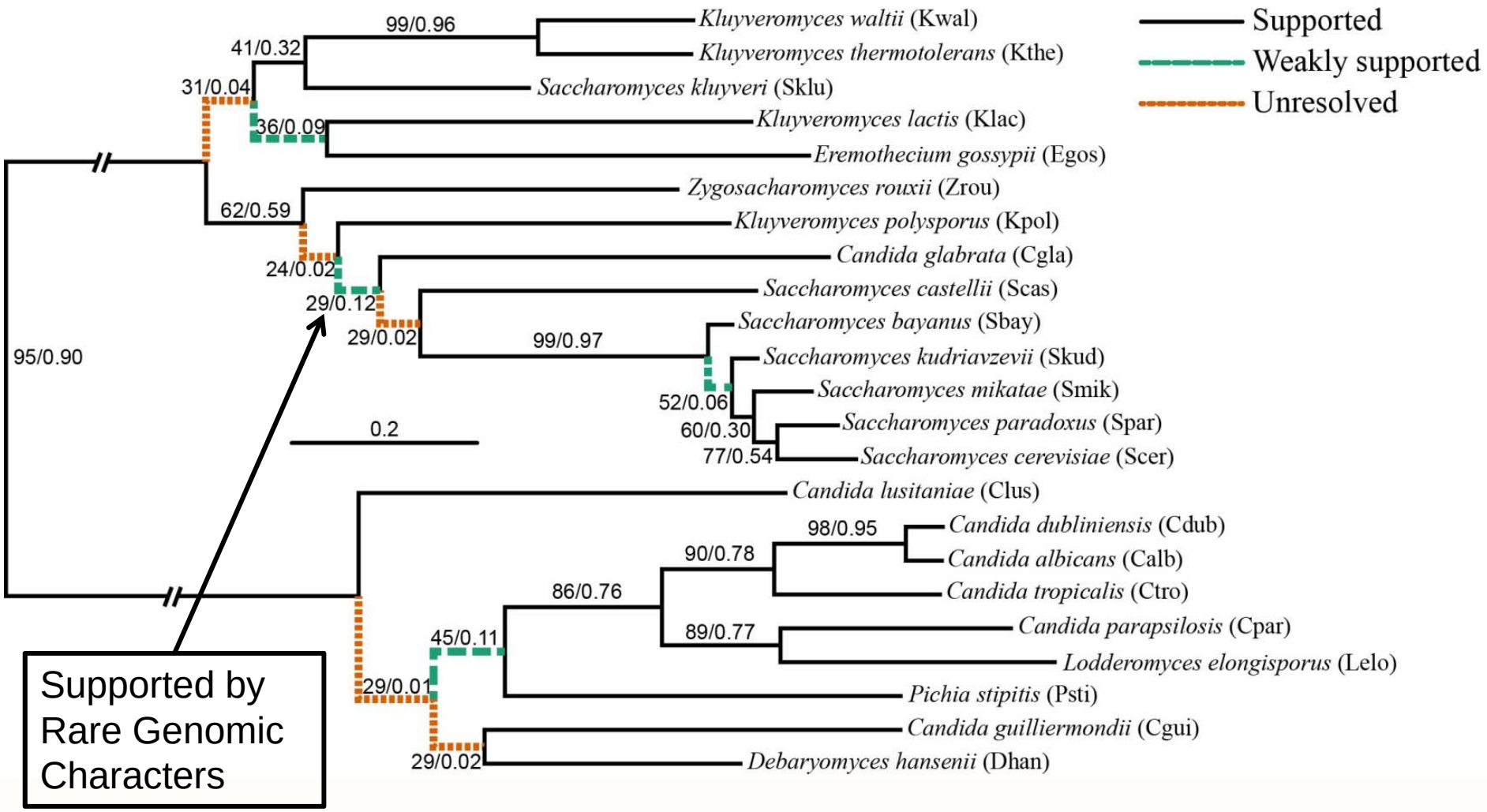
Salichos & Rokas (2013) Nature

Least Supported Internodes Harbor the Most Conflict



Salichos & Rokas (2013) Nature

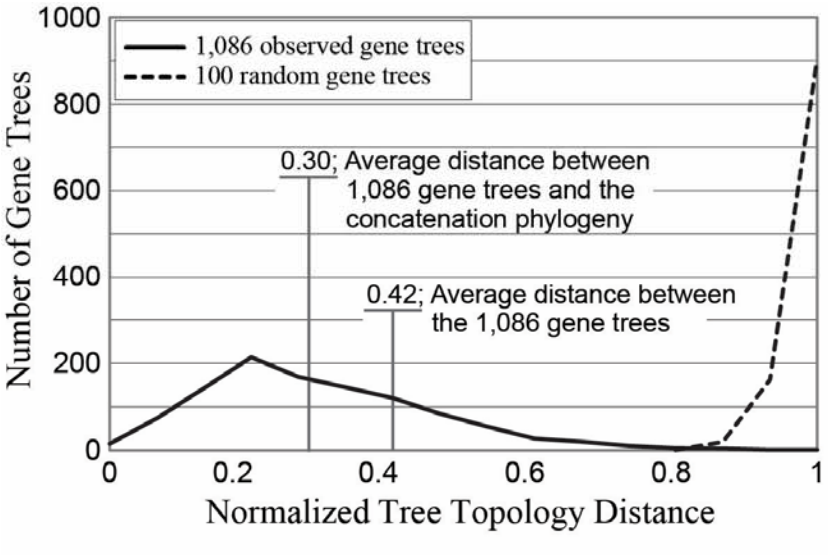
The Status of the Yeast Phylogeny



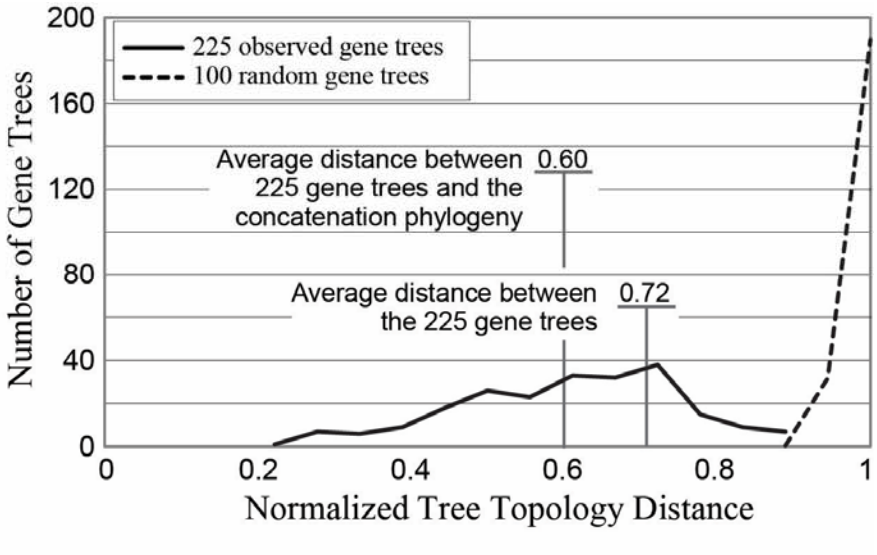
Gene Support Frequency / Internode Certainty

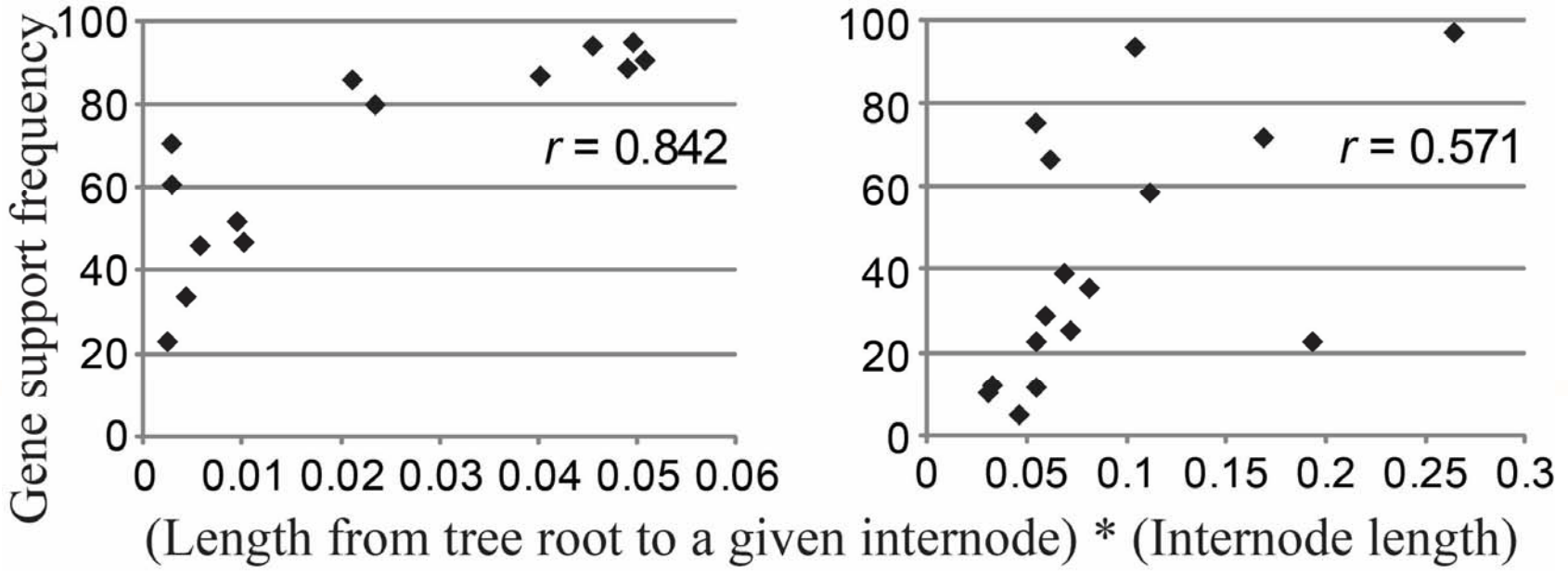
The Same is True for Vertebrate and Metazoan Datasets

Vertebrates (1,086 genes, 18 taxa)



Animals (225 genes, 21 taxa)



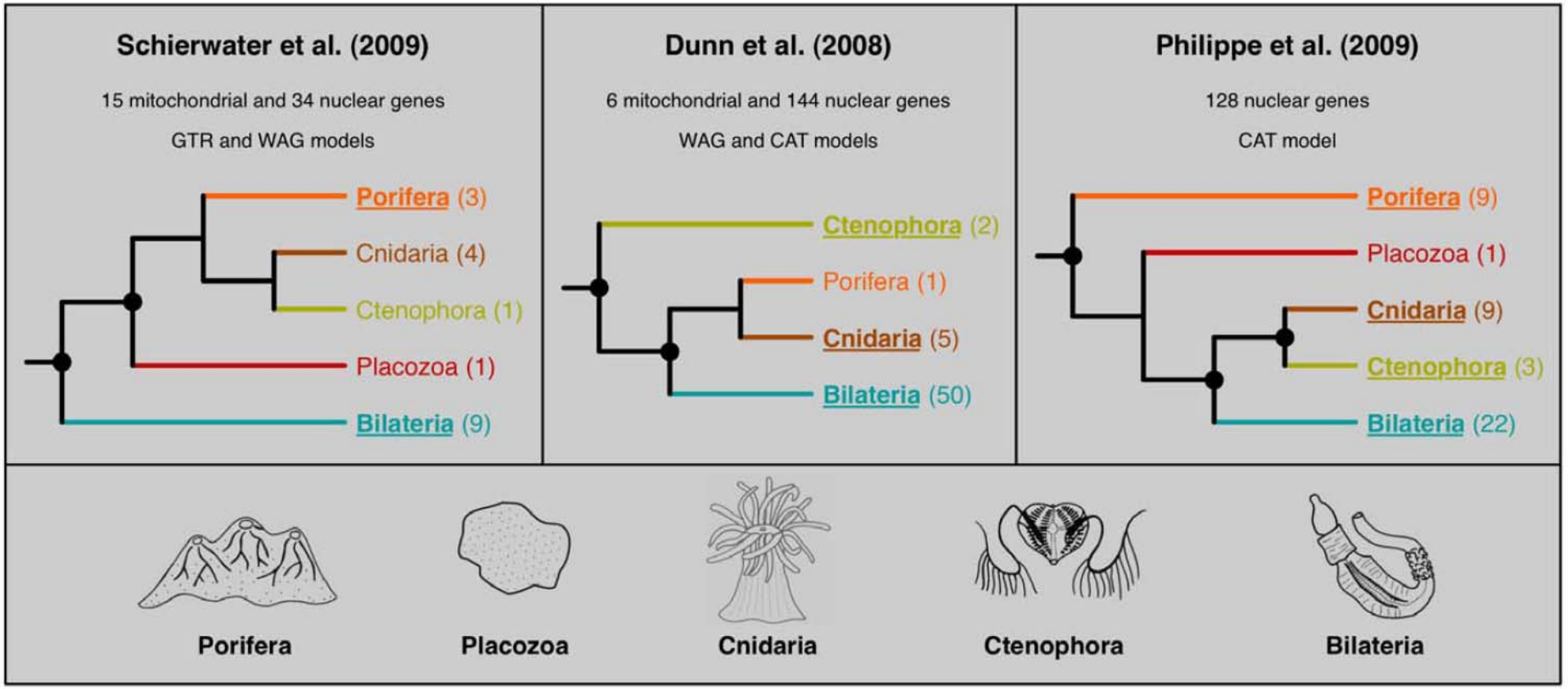


Lessons from Yeasts, Vertebrates & Animals

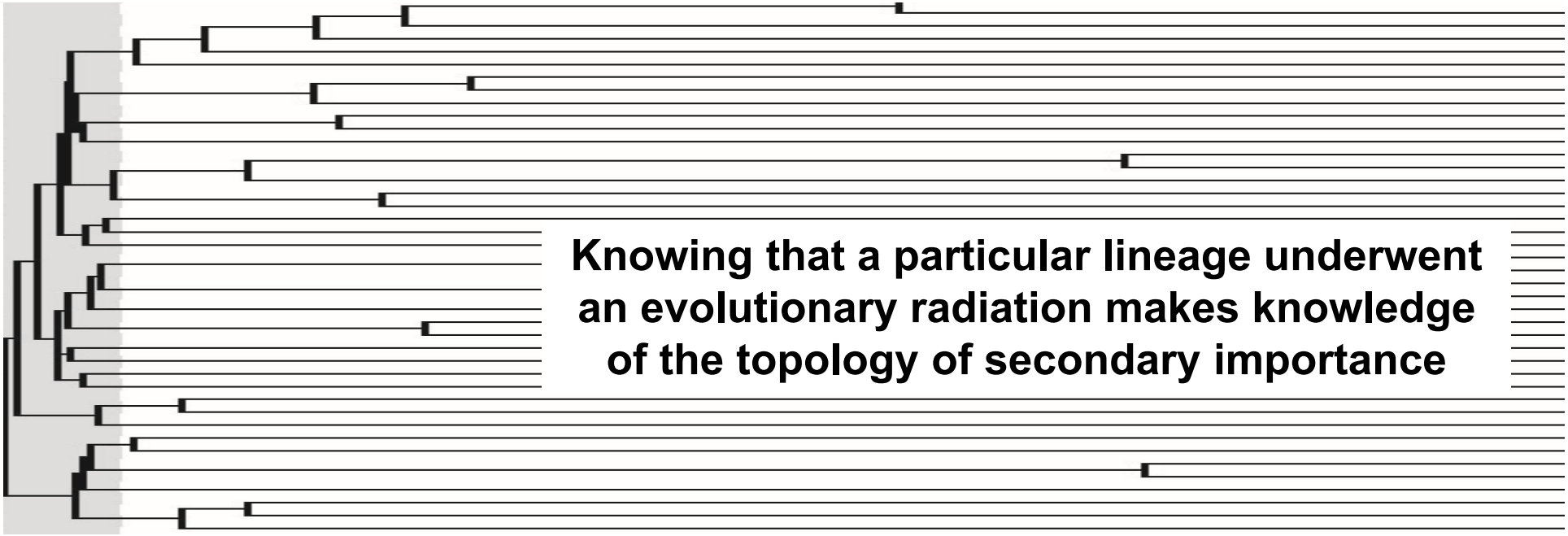
- **Few, if any, of the Gene Trees are Topologically Identical to Each Other or to the Phylogeny Inferred by Concatenation**
- **Concatenation analysis Overconfident and Can Mislead**
- **Internode Support is Inversely Correlated with Internode Length and Depth**
- **Selecting Genes or Gene Tree Bipartitions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

Salichos & Rokas (2013) Nature

Incongruence in Deep Time



Philippe et al. (2011) PLoS Biol.

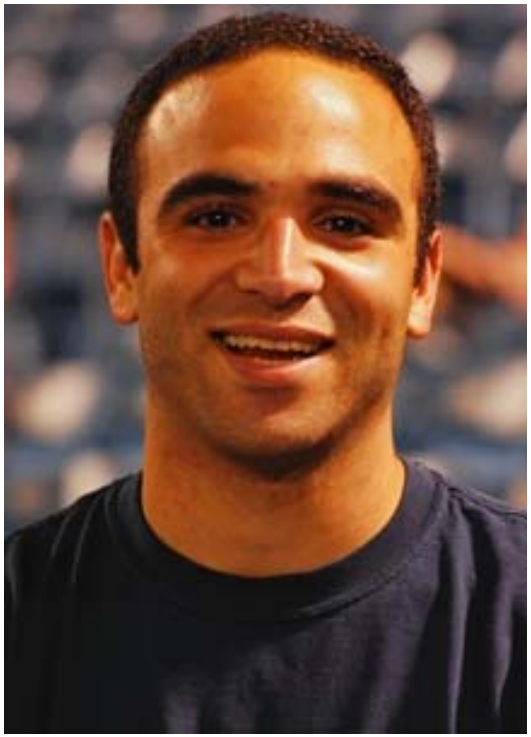


Knowing that a particular lineage underwent an evolutionary radiation makes knowledge of the topology of secondary importance

"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



John Gibbons

Chris Hittinger

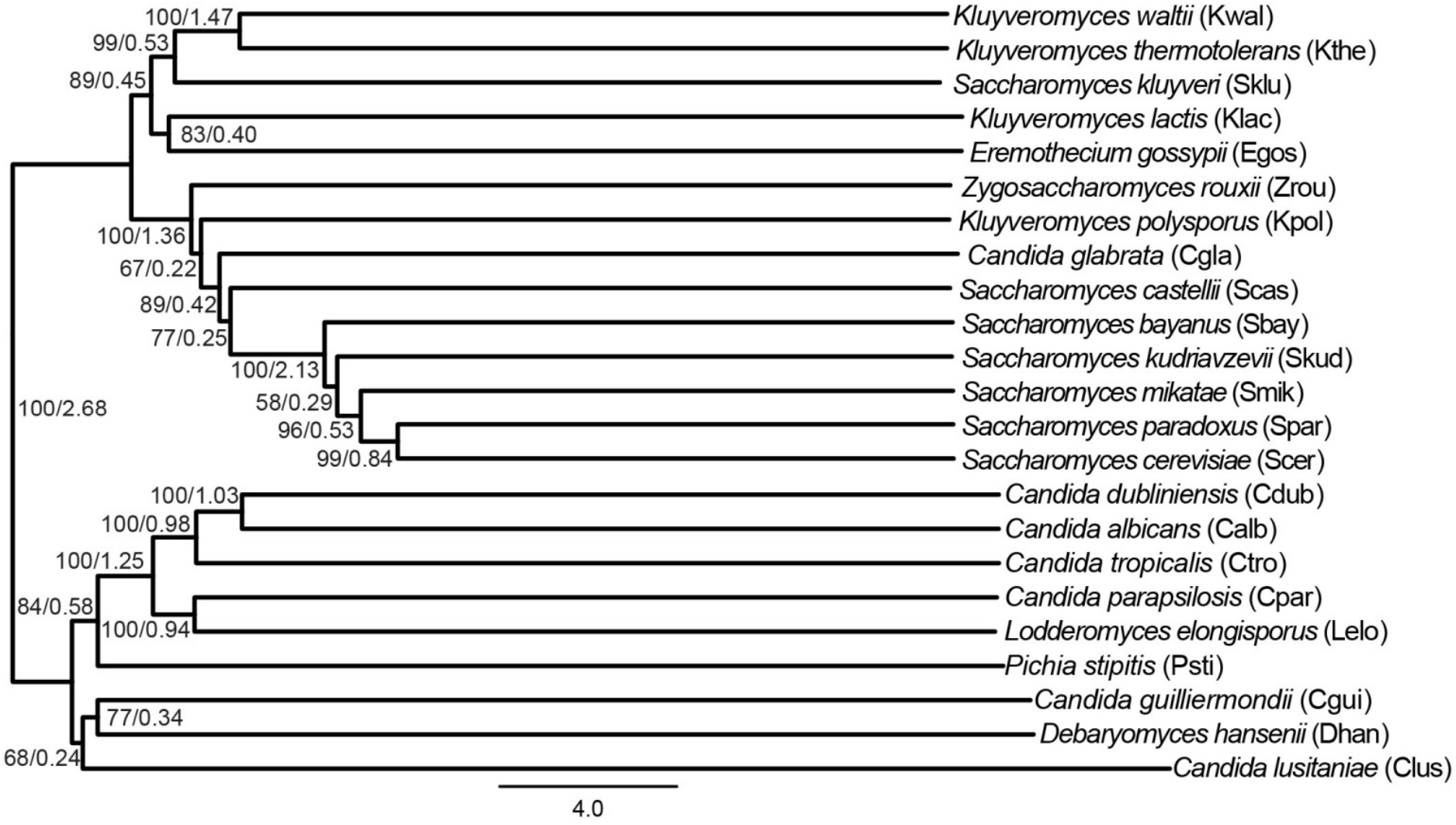
Leonidas Salichos

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