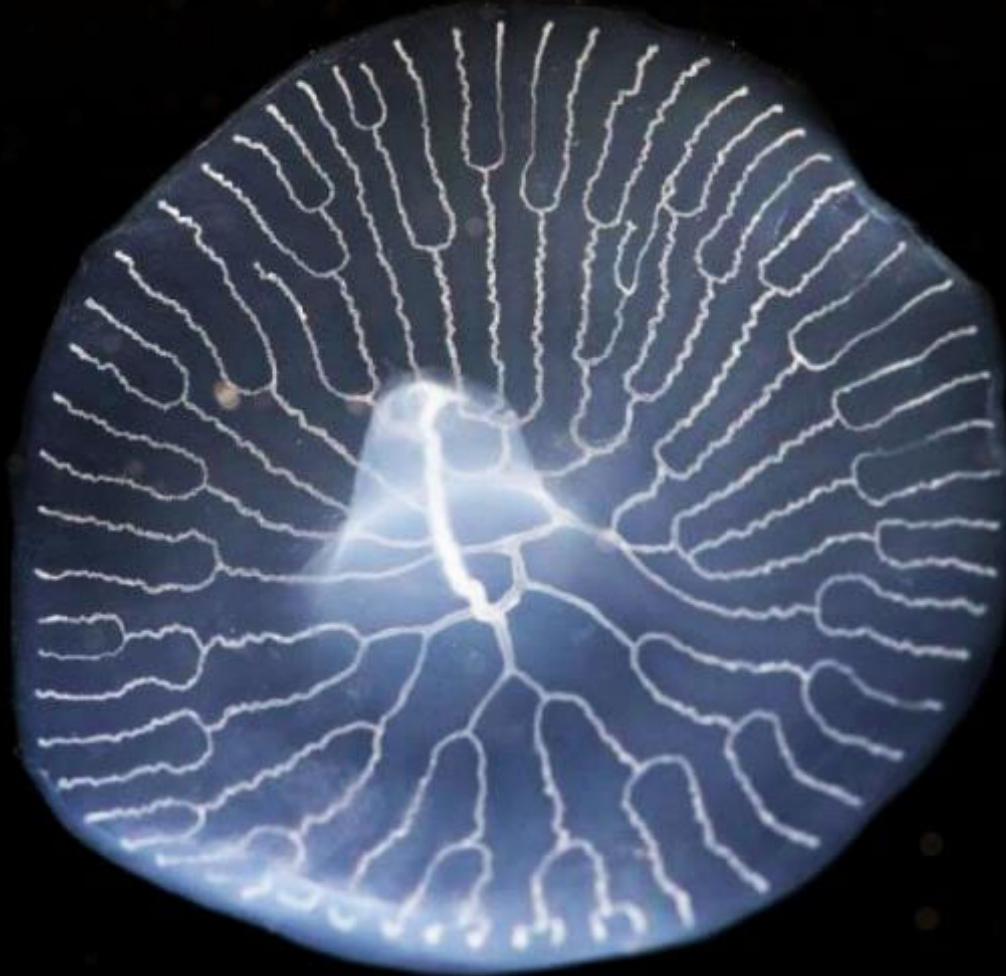


Evolutionary Genomics



Antonis Rokas

Department of Biological Sciences, Vanderbilt University

<http://www.rokaslab.org>

@RokasLab

Lecture Outline

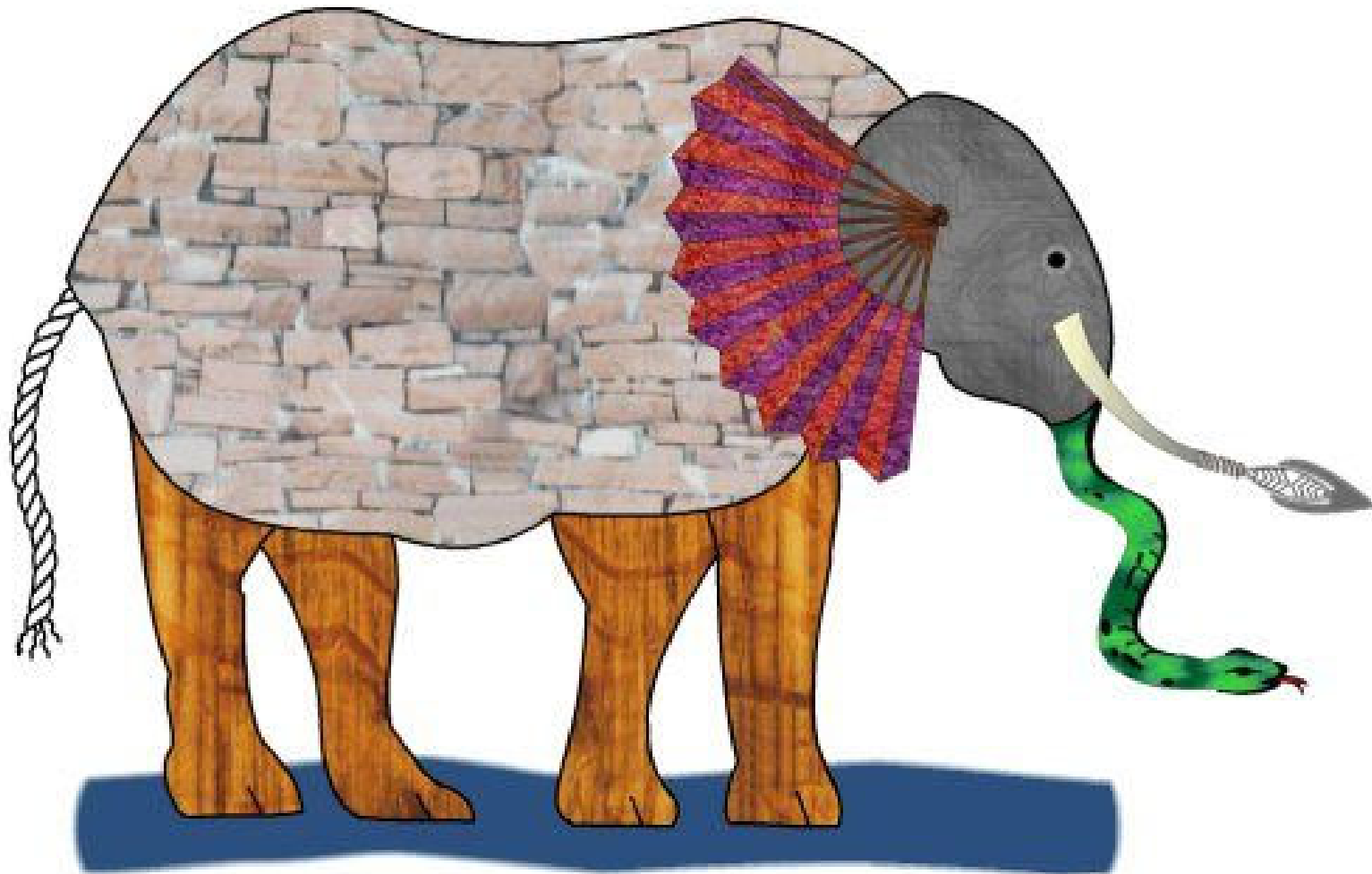
❖ **Introduction to evolutionary genomics**

❖ **Phylogenomics**

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

❖ **Using genomes to understand ecology and evolution**

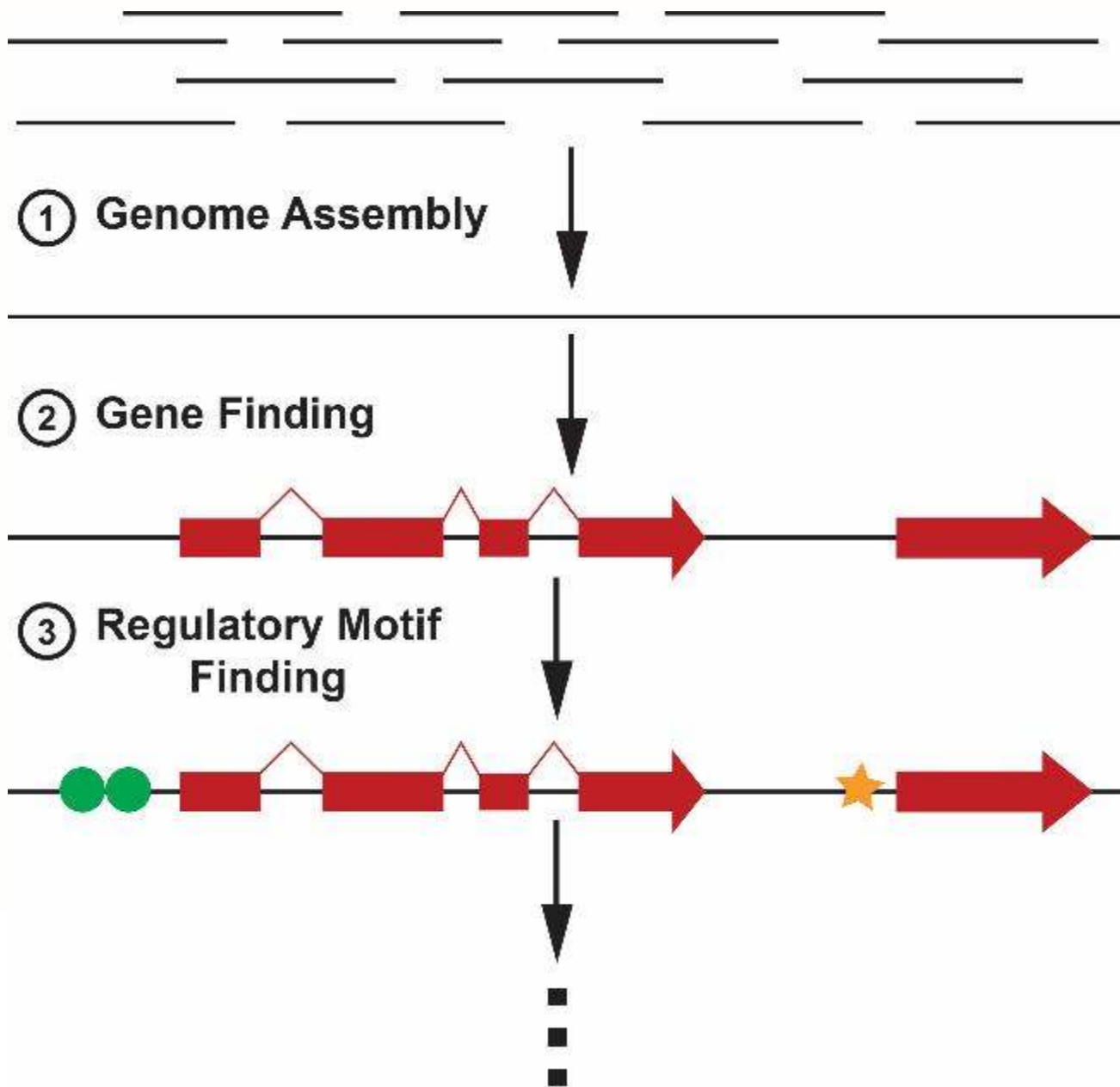
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTGAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATAACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAGCGGTGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTGAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTGCCTCCAGCTCCCTAGCACTGGCAGGGTTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTAAGTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGTTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACCTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACAGCCGCATCCCCTACAAAATCTACCCCAAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCGAGGGGCCACCAGGAACCATGTTT
CAGTCTGACAATAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAACAAGGAAA
AGGGGATACAGTGCGGATTTCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT

Understanding the Genome Requires Tools



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTGAGTGTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATAACGAGTGCTTCTCCCAACAGACAGCTGGCTTGTAAGCGGTGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTGCGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTGCGCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTACAGAGGCAGCTCAGAACTGAGATTTGATACATTA
GACCACCCTCGCTTCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTAAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGATCTGCTCTCTCCAAGTCT
GCGATATTGCTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTAAGTTCGAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGTAAGCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTAATAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACCTTAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCTACAAAATCTACCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCGCGGACTCCTACAGCTCCCGCTTACCCCCAAAGCAATGATACTATT
ATCGATCCCATTTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCATTCCCAAGTGCGAGGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCTACATTCATATGGGCCAGCGATTGGAACCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTTCGTCAAGT

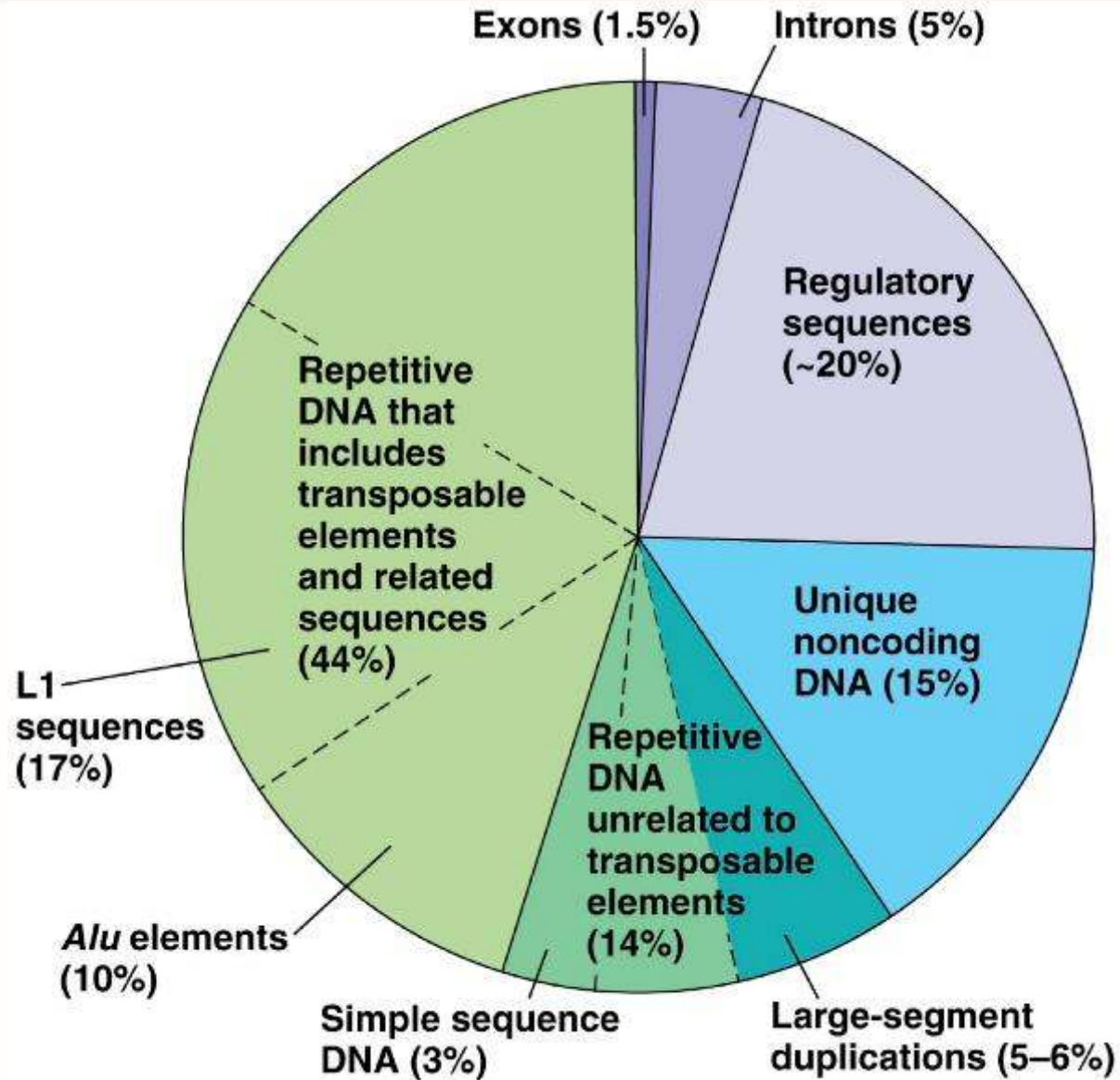
Transposon

Protein Binding Site

Exon

Intron

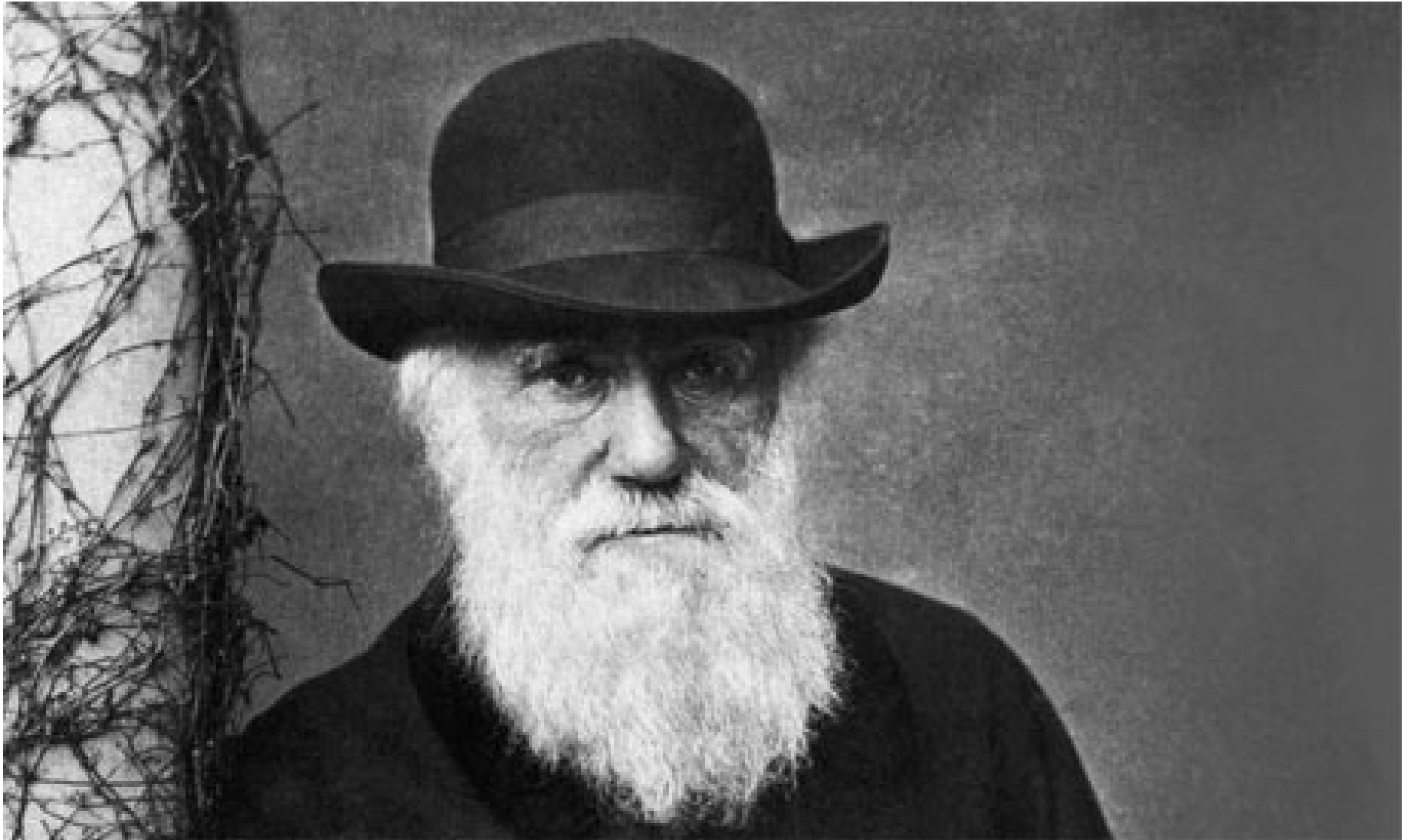
Organization of the Human Genome



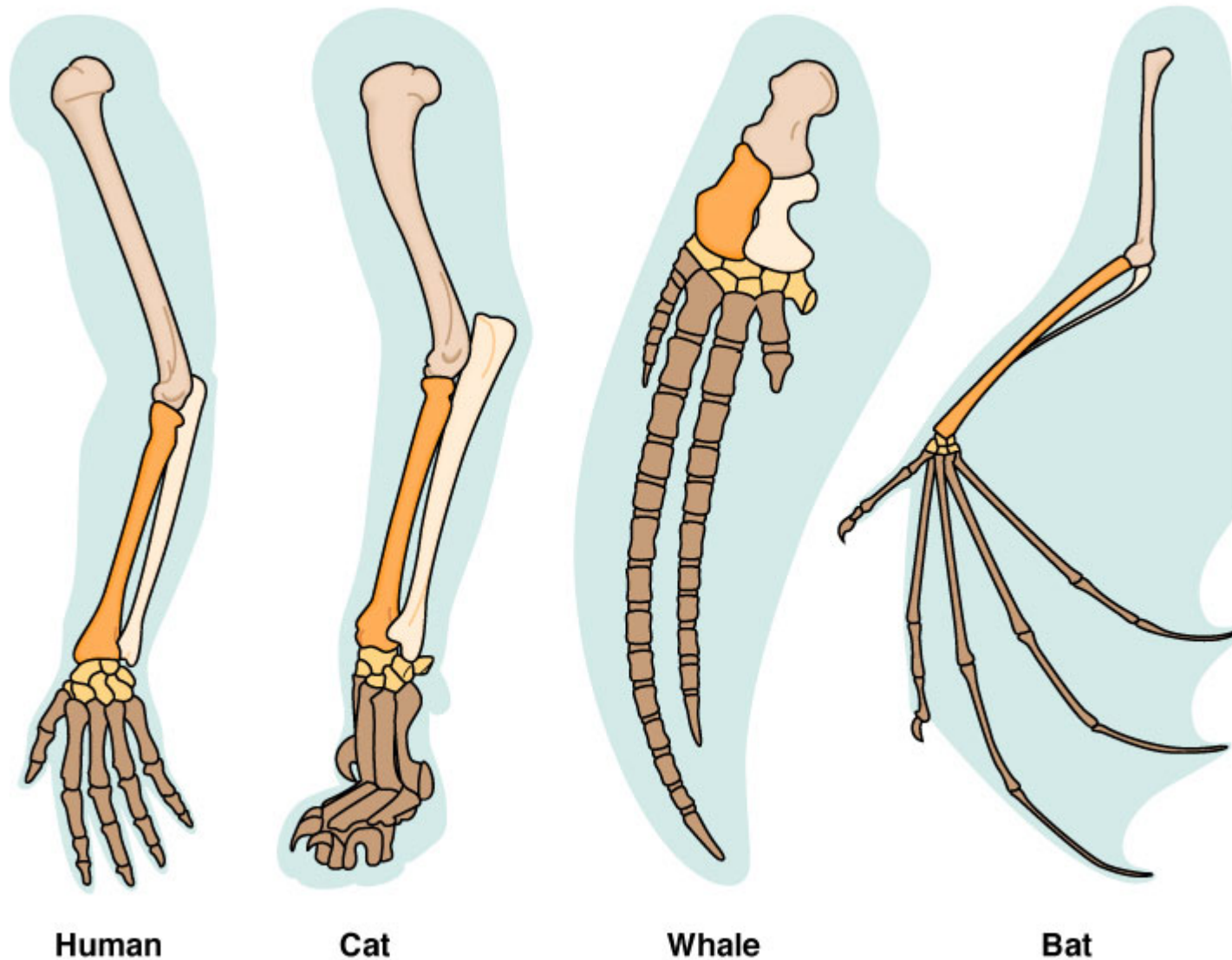
© 2011 Pearson Education, Inc.



Understanding the Genome Requires a Theory



Similarity in Anatomy Suggests Common Origins



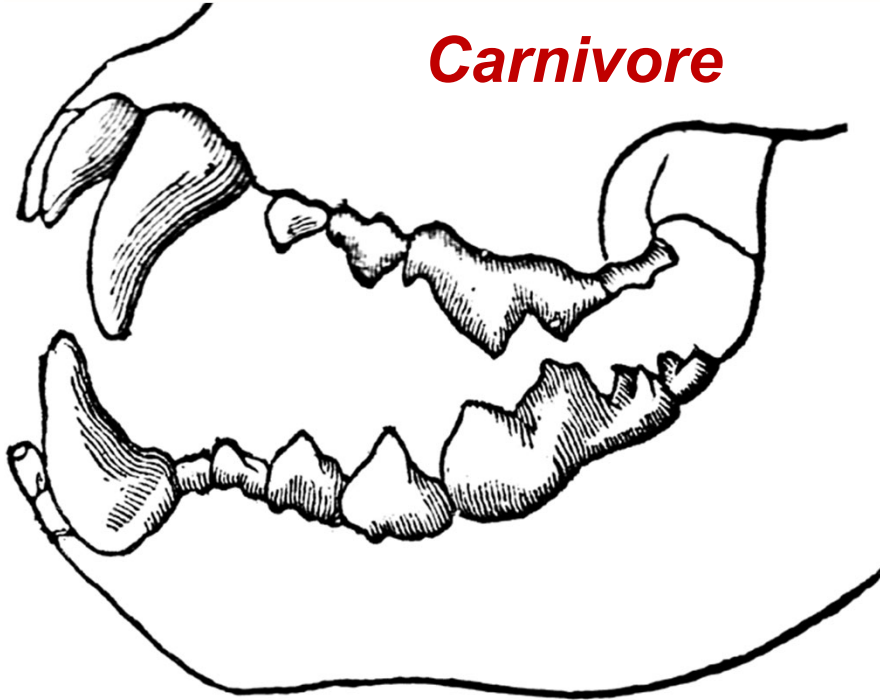
©1999 Addison Wesley Longman, Inc.



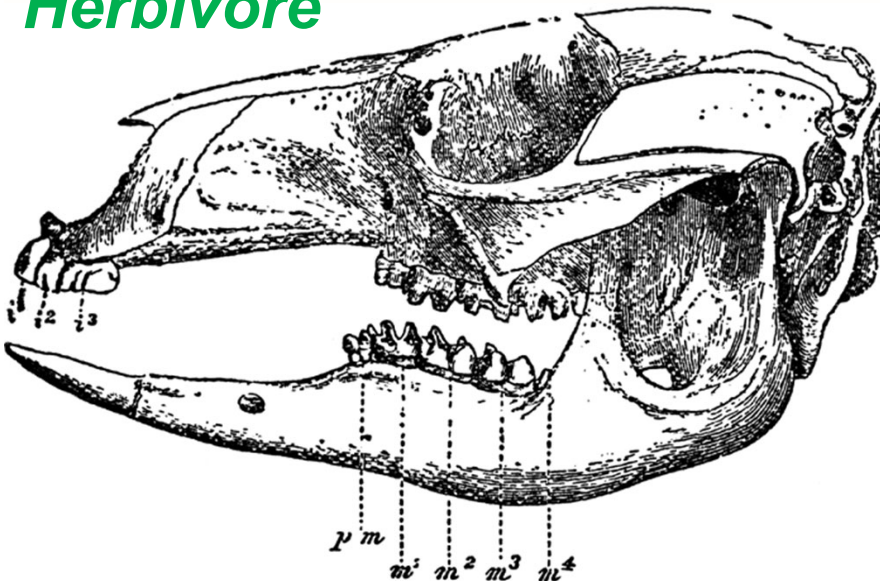
http://www.mun.ca/biology/scarr/139393_forelimb_homology.jpg

Differences in Anatomy Suggest Adaptations

Carnivore



Herbivore

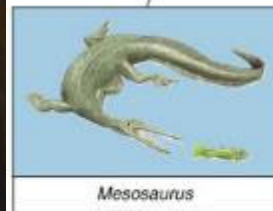
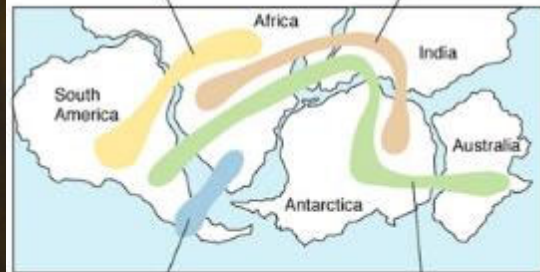
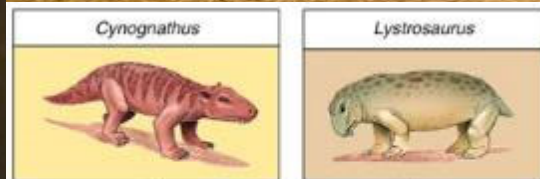
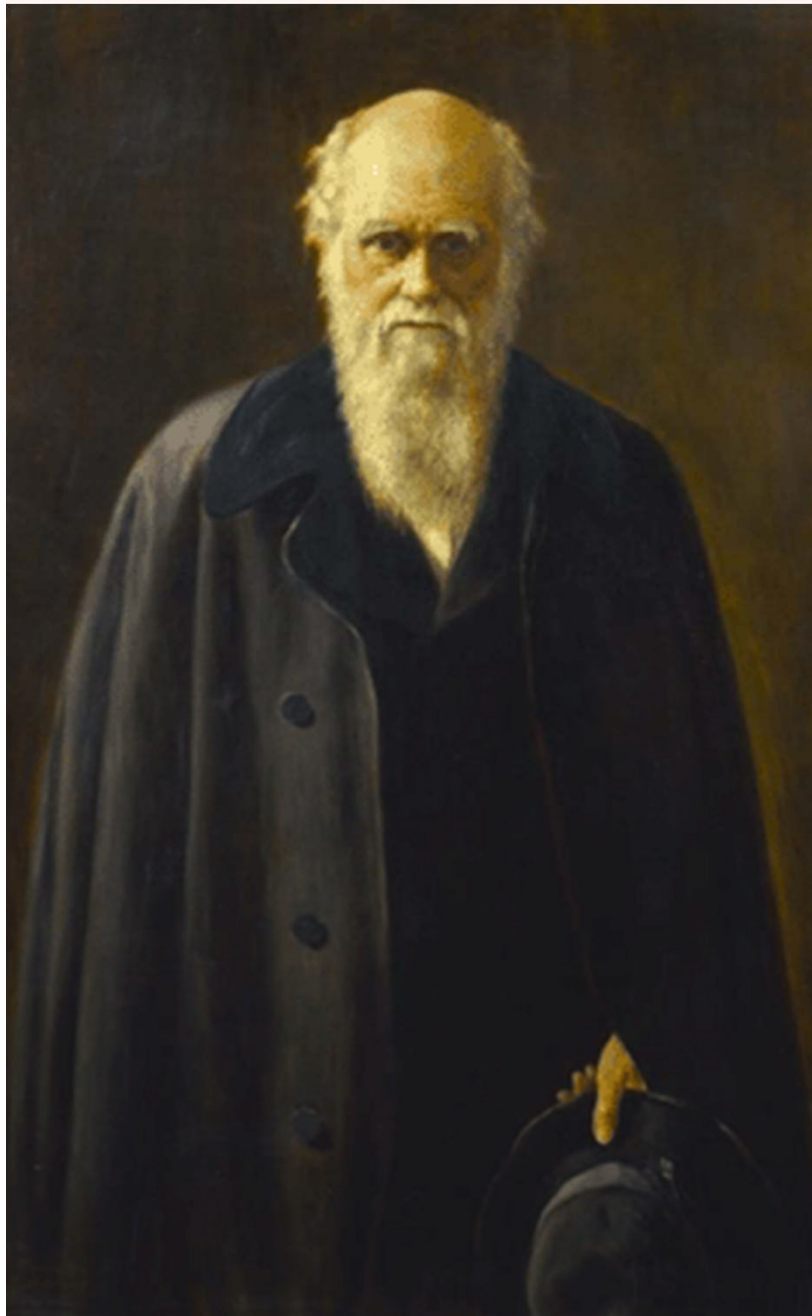


Incisivosaurus



Carnivore or Herbivore?

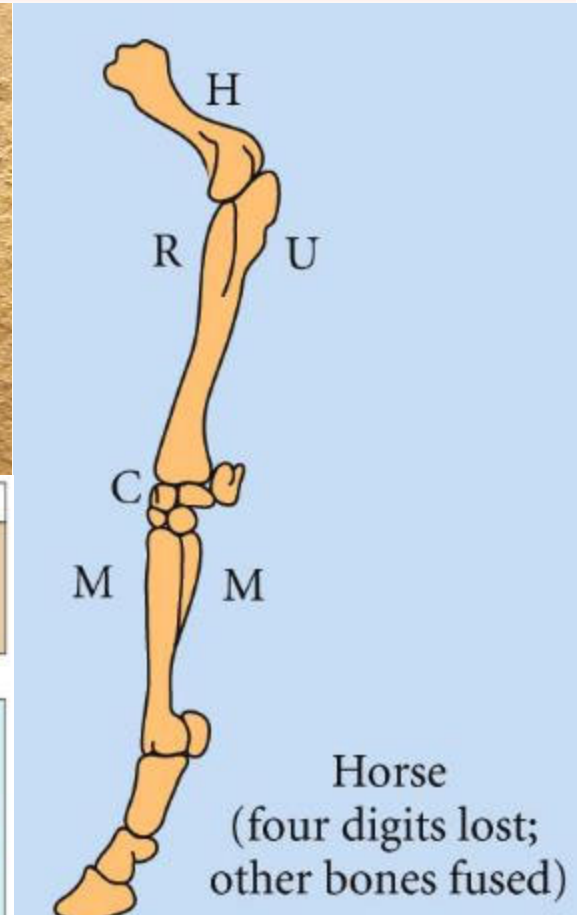
Darwin's Data



Mesosaurus

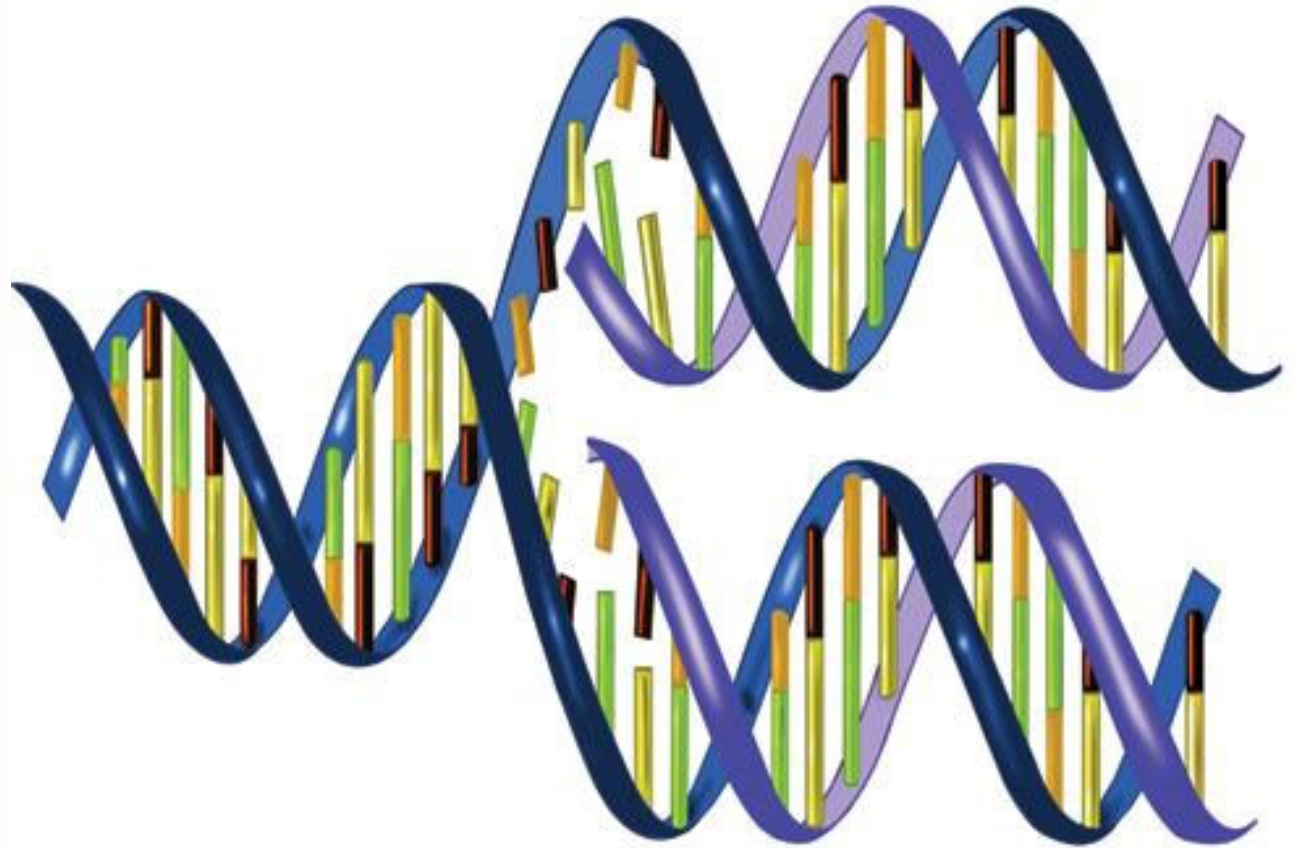


Glossopteris



The DNA Record

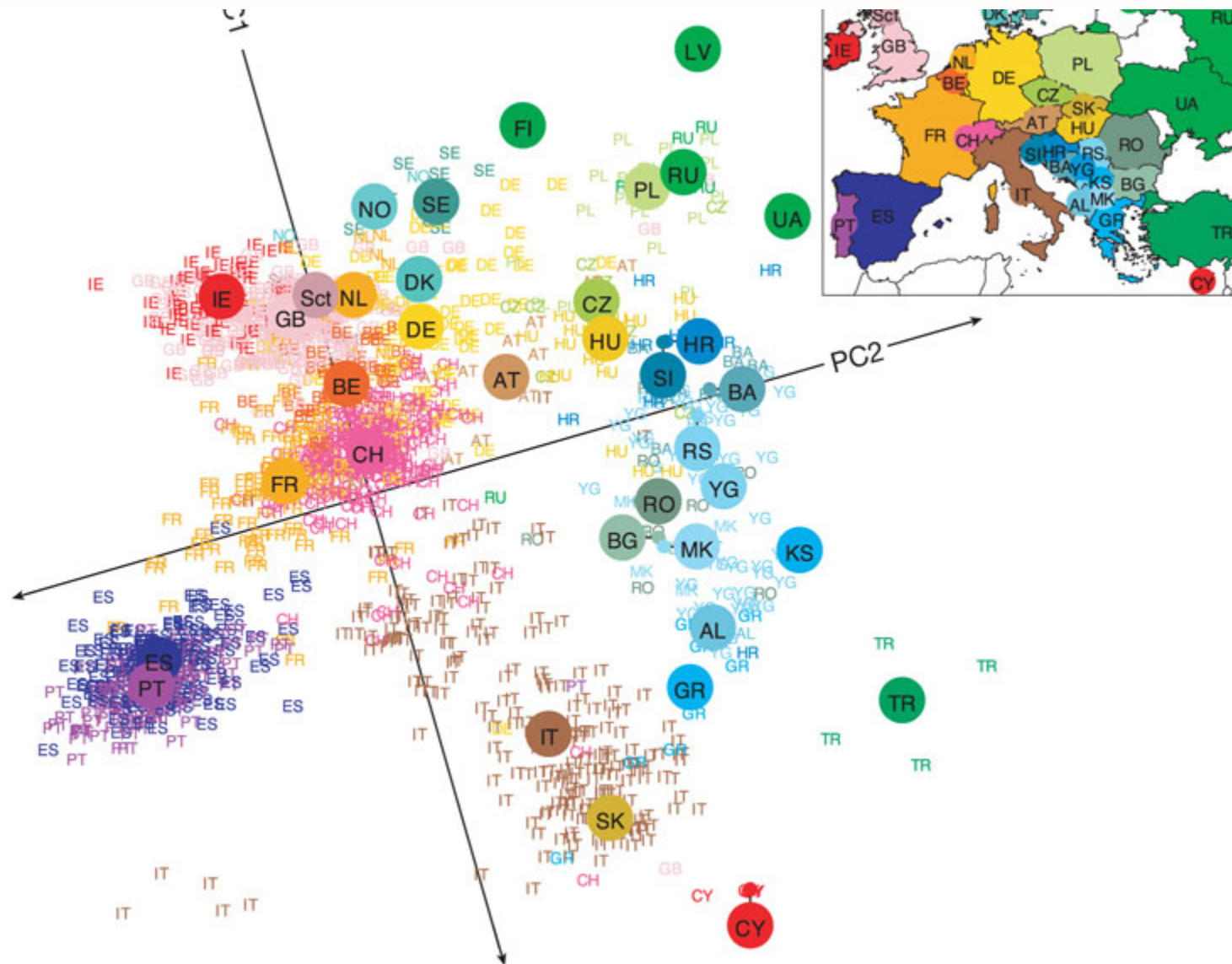
**The DNA record
contains
important clues
about
organisms'
biological past,
and their history
of change and
adaptation**



Similarities in the DNA record suggest common origin

Differences in the DNA record (might) suggest adaptations

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	

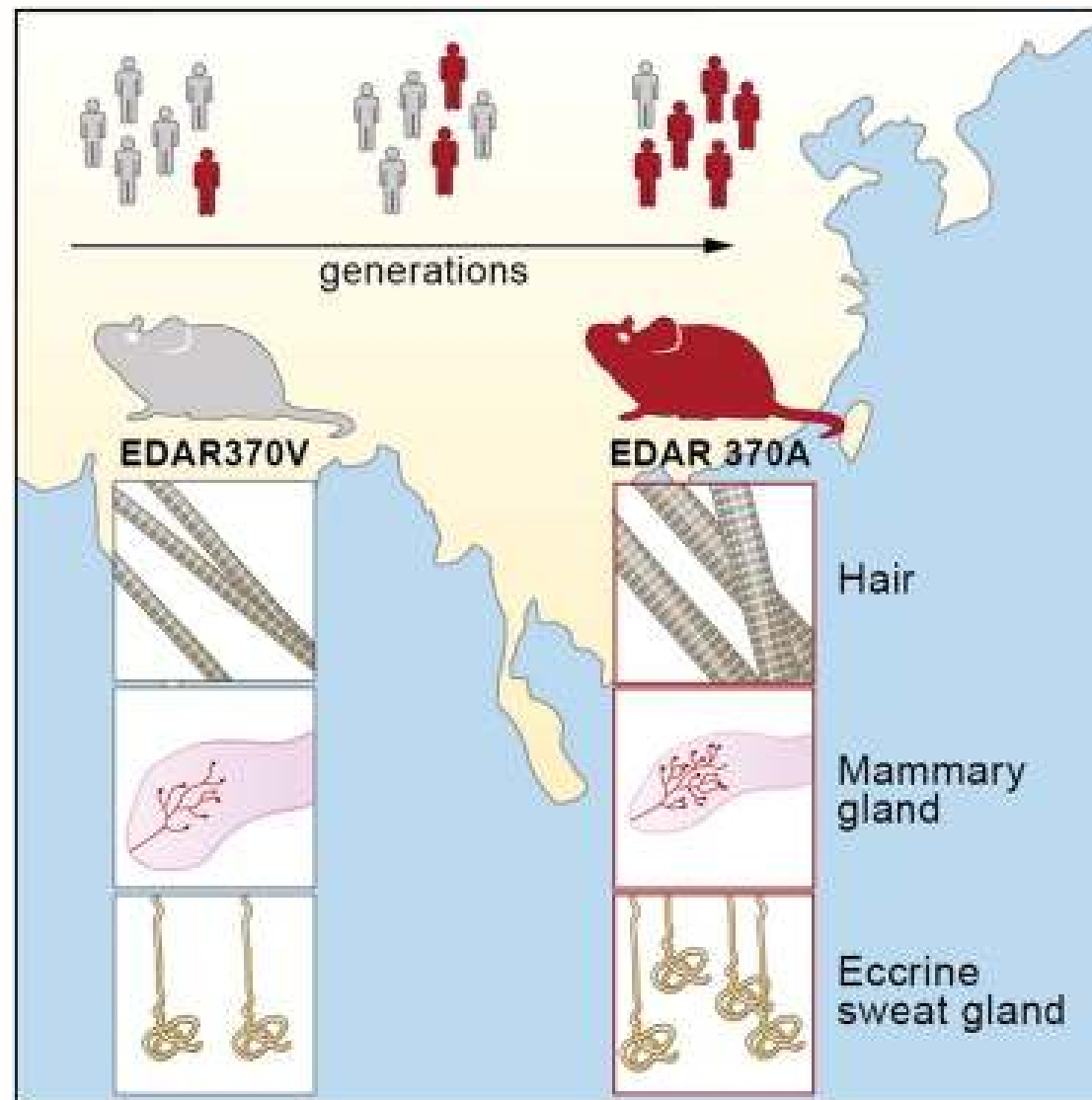
In the European population,
involved in skin pigmentation

In the West African population,
related to Lassa virus infection

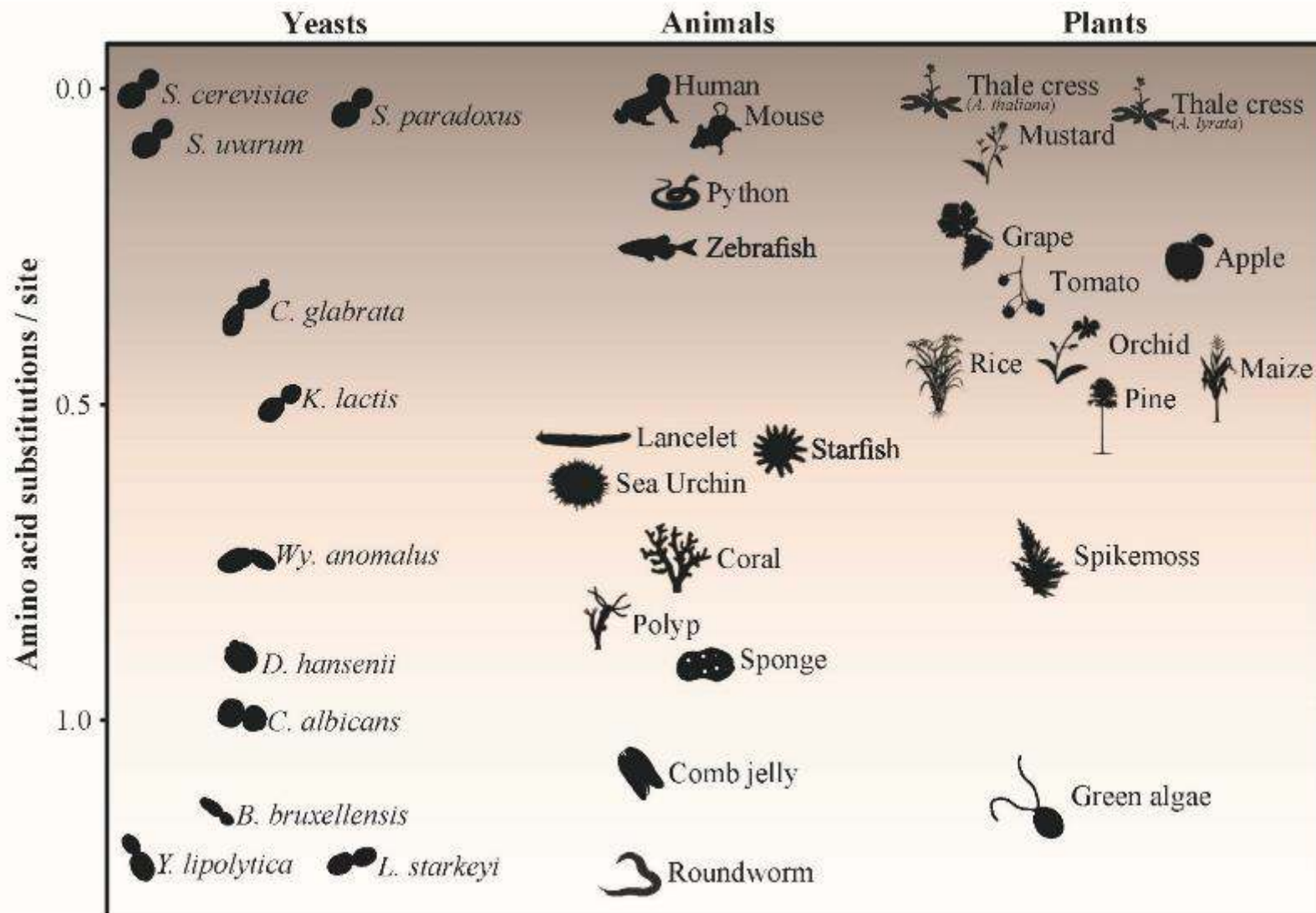


Sabeti et al. (2007) Nature

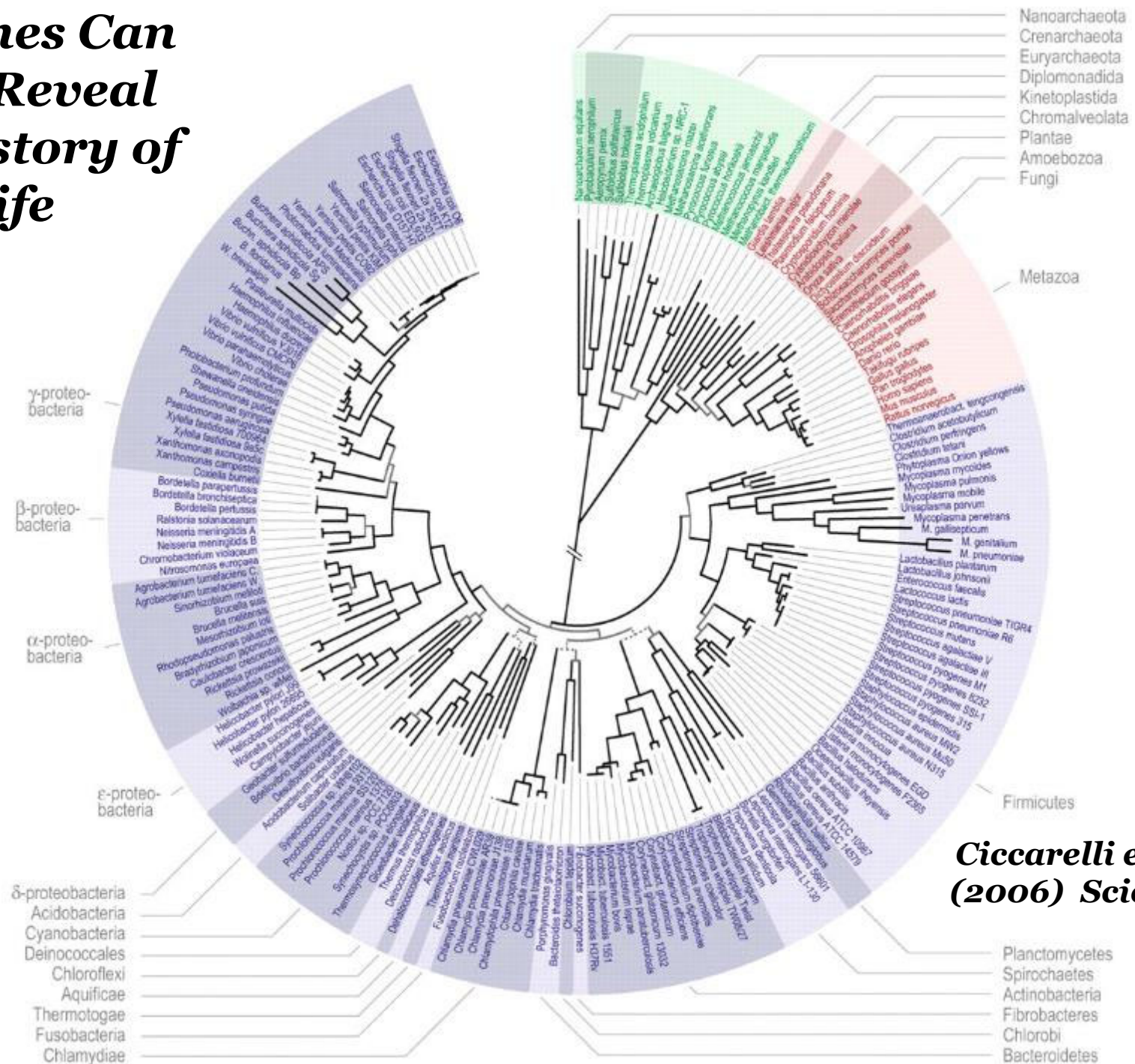
Phenotypic Effects of Recent Positive Selection



Genomes Provide a Common Yardstick for Comparison



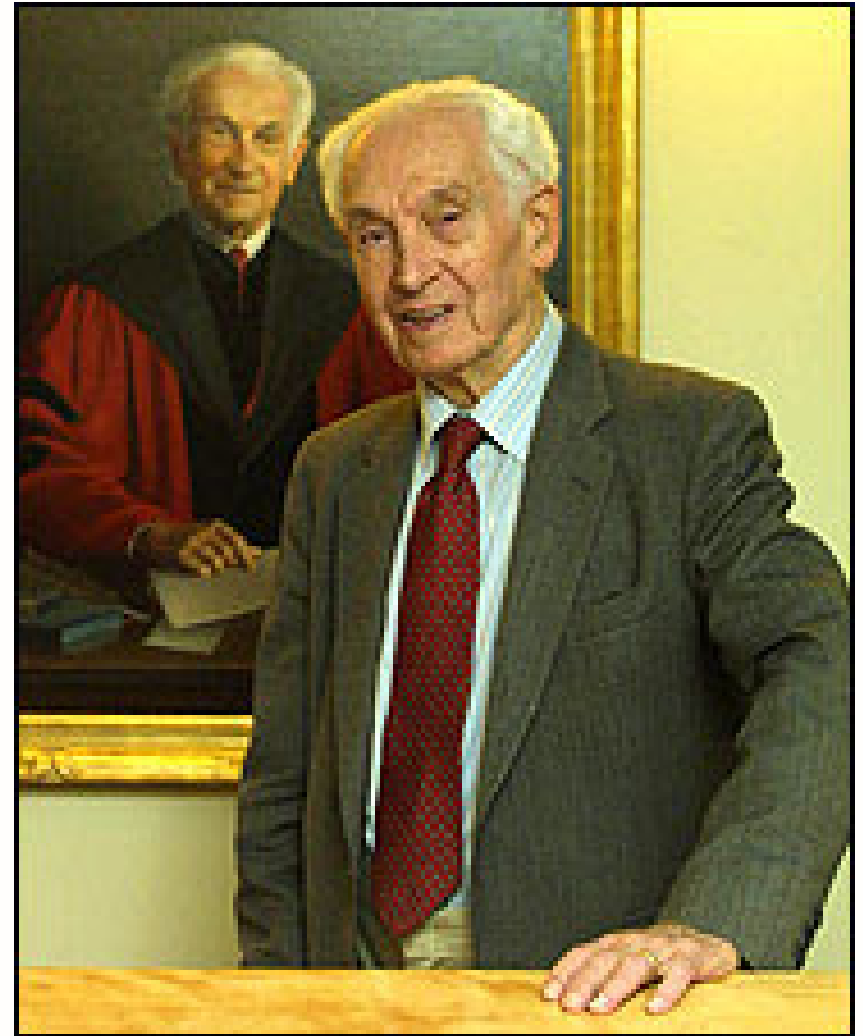
Genomes Can Help Reveal the History of Life



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

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

Ernst Mayr, 1963



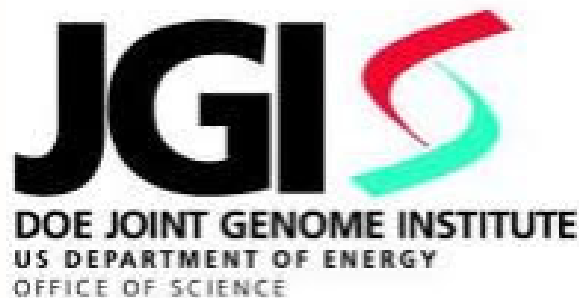
What Makes Us Sick Is the Stuff of Life

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*

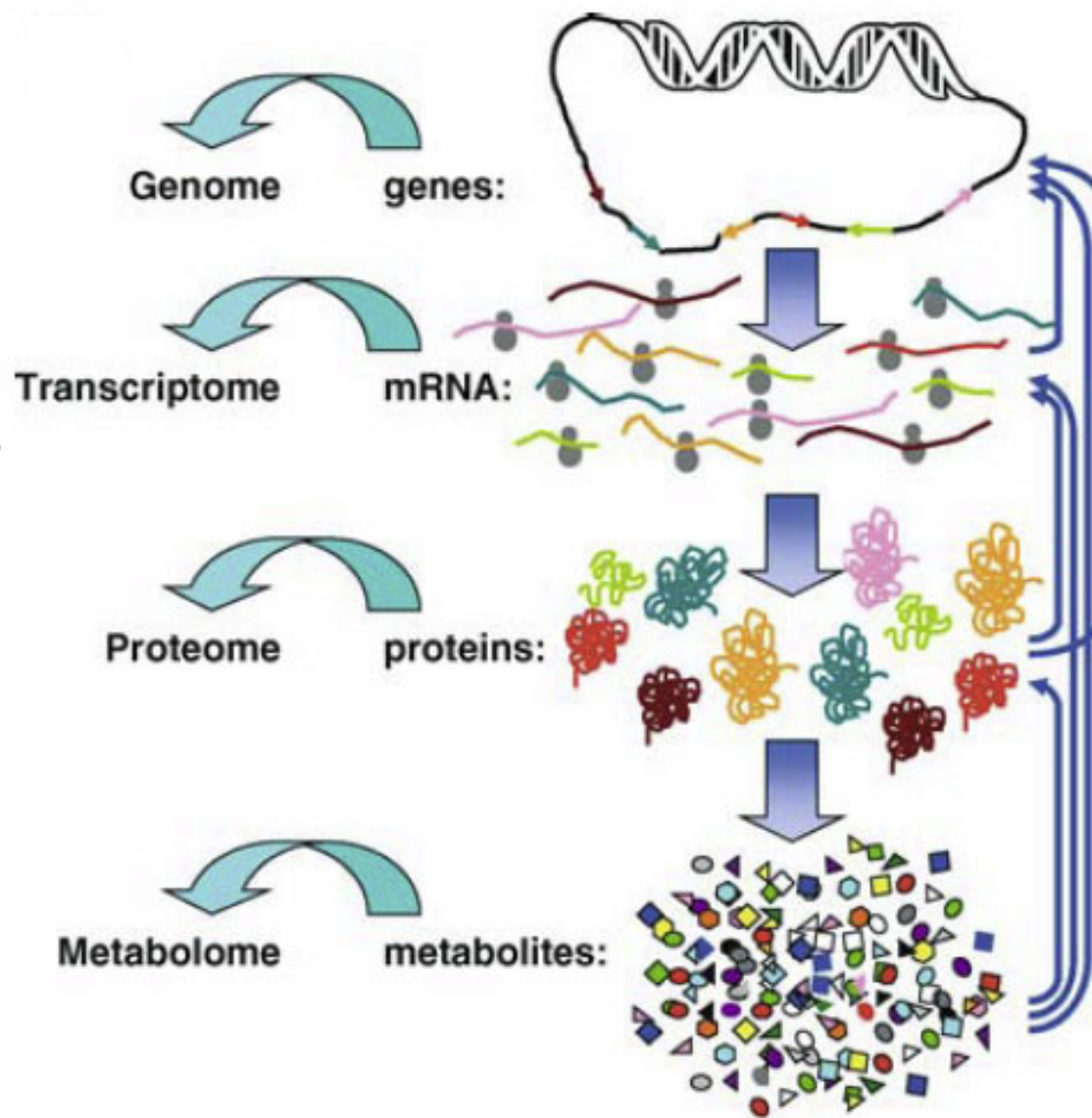
Genomics Used to Be “Big Science”...



... But is now Accessible to Every Lab



And So Are Many High Throughput Technologies



Goodacre (2005) Metabolomics

Novel Ways to Probe Gene Function in Any Organism

RNAi

TALENs / ZFNs and other nucleases / CRISPRs

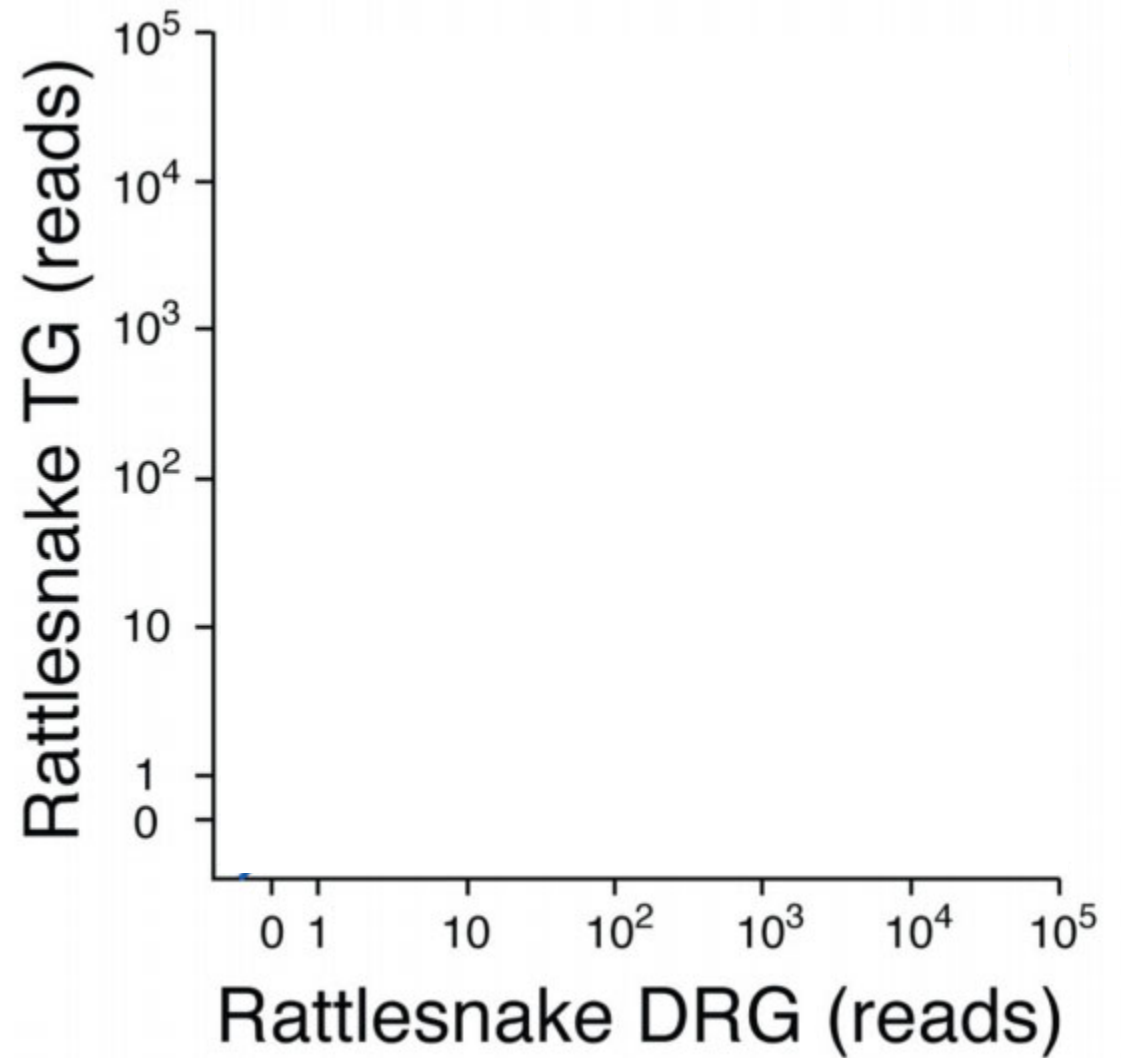
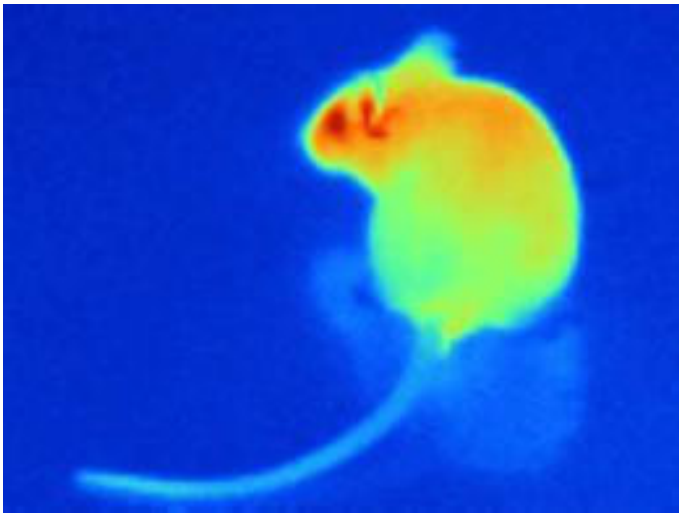


The Genomes of Non-Model Organisms are the New Frontiers



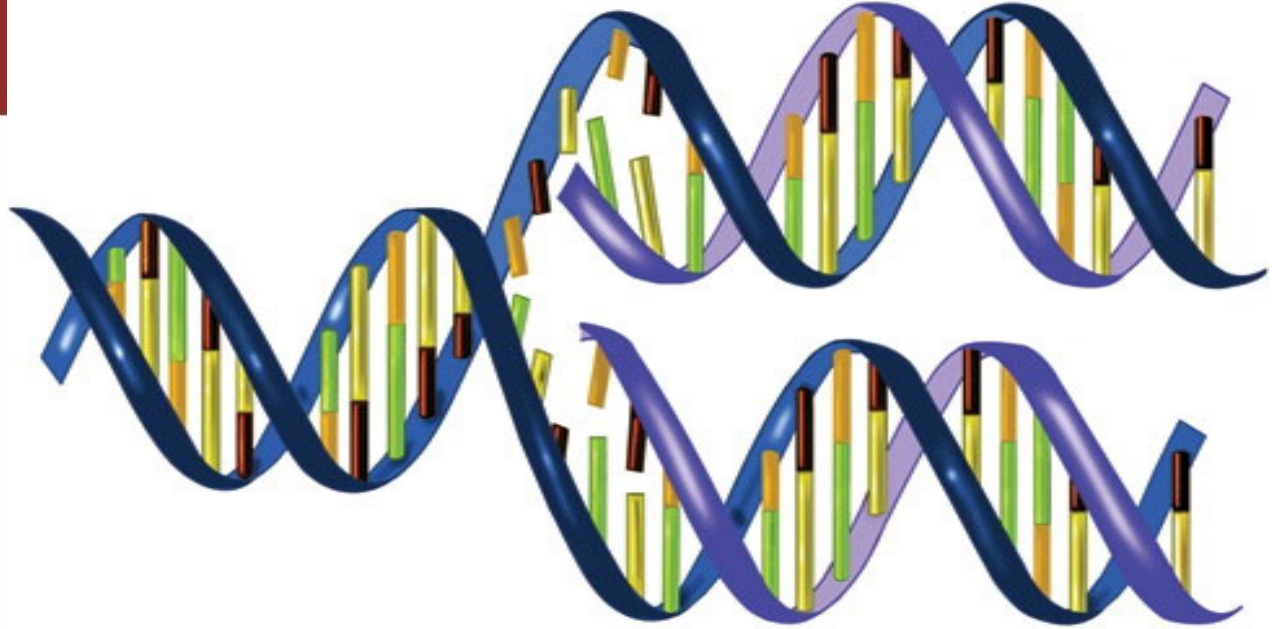
Rokas & Abbot (2009) Trends Ecol. Evol.

Snake Infrared Vision



Gracheva et al. (2010) Nature

The DNA Record



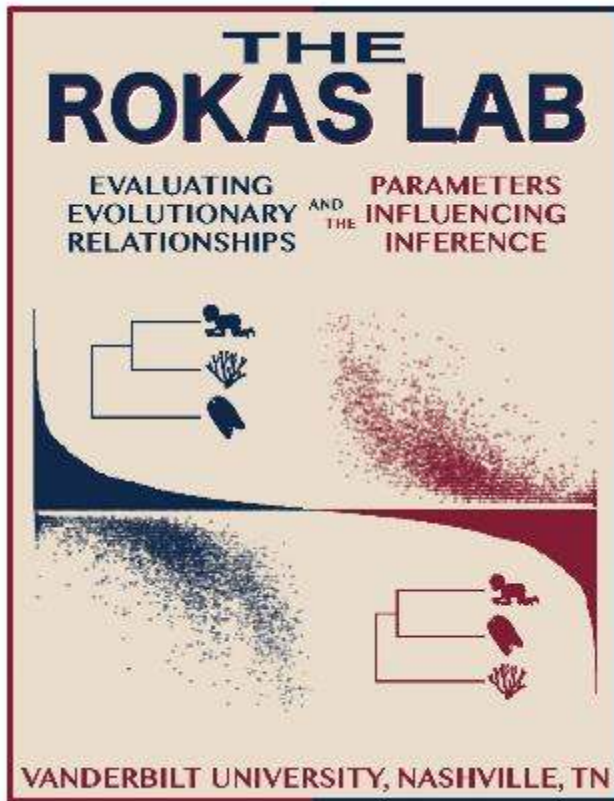
“The genome is, it's a fossil record; the genome is a landscape; the genome is a whole geography of distributions. [...] The genome is a storybook that's been edited for a couple of billion years, and you could take it to bed, like *A Thousand and One Arabian Nights*, and read a different story, in the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

Eric Lander

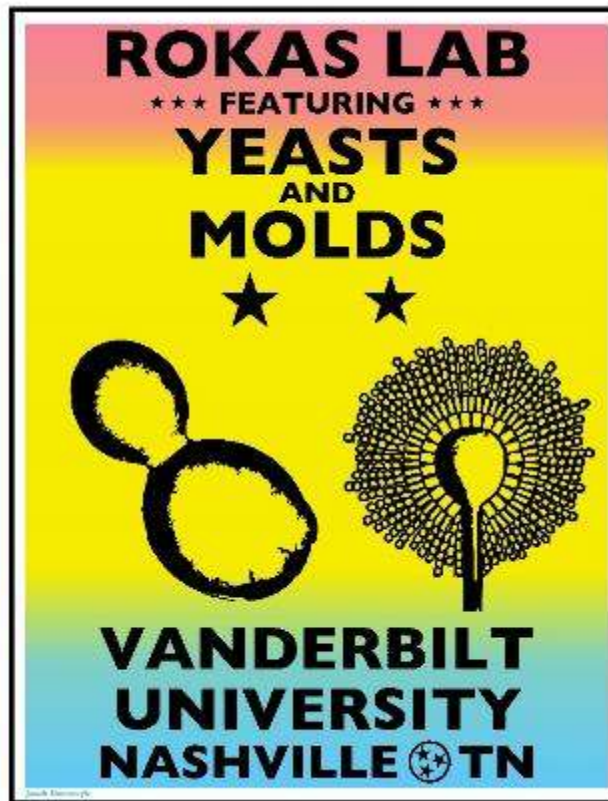
The Rokas Lab



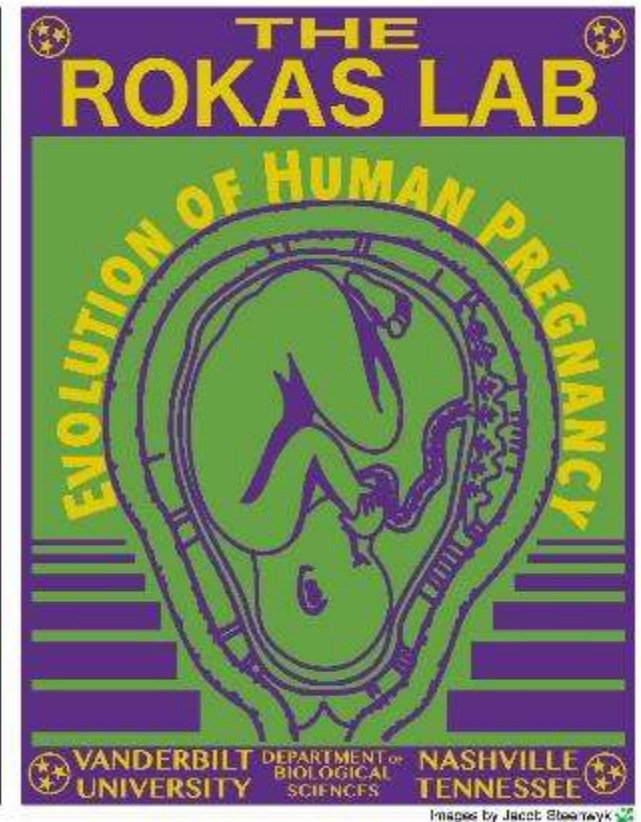
We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches



Phylogenomics

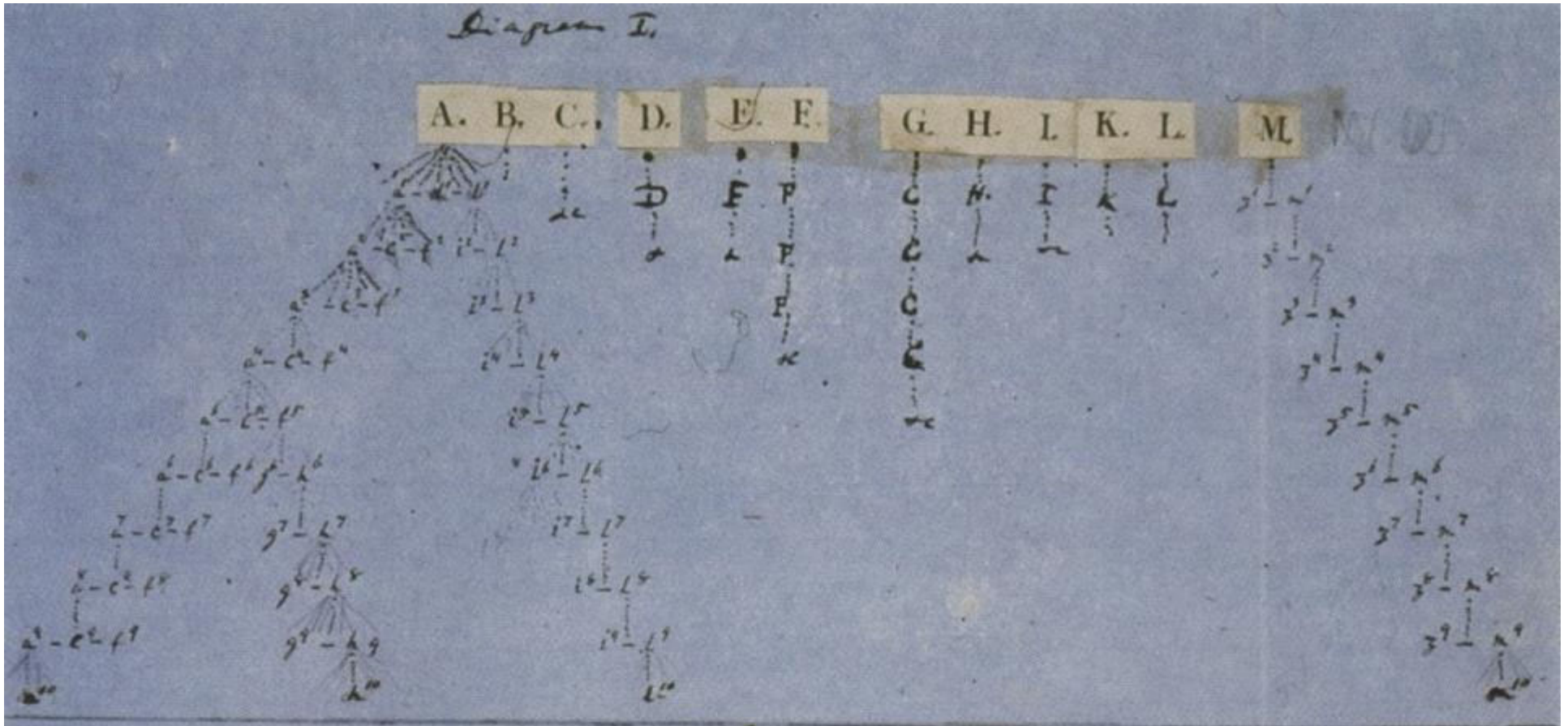


**The evolution of
primary &
secondary
metabolism in
fungi**



**The evolution
of mammalian
pregnancy**

Darwin's Tree



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



Maderspacher (2006) *Curr. Biol.*

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

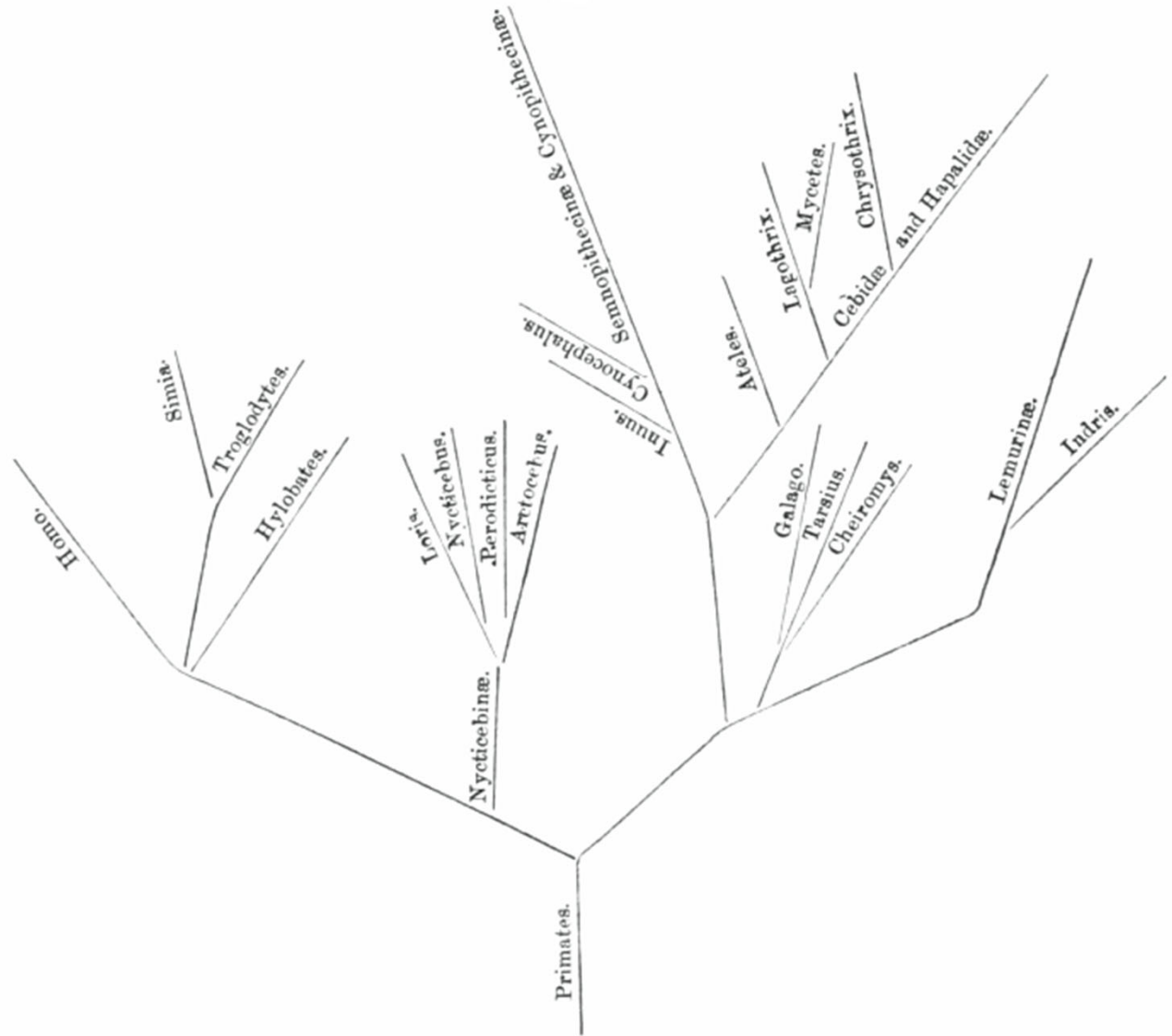
A grand and almost untrodden field of inquiry will be opened, on the causes and laws of variation, on correlation of growth, on the effects of use and disuse, on the direct action of external conditions, and so forth. The study of domestic productions will rise immensely in value. A new variety raised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

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

The First Published Phylogeny



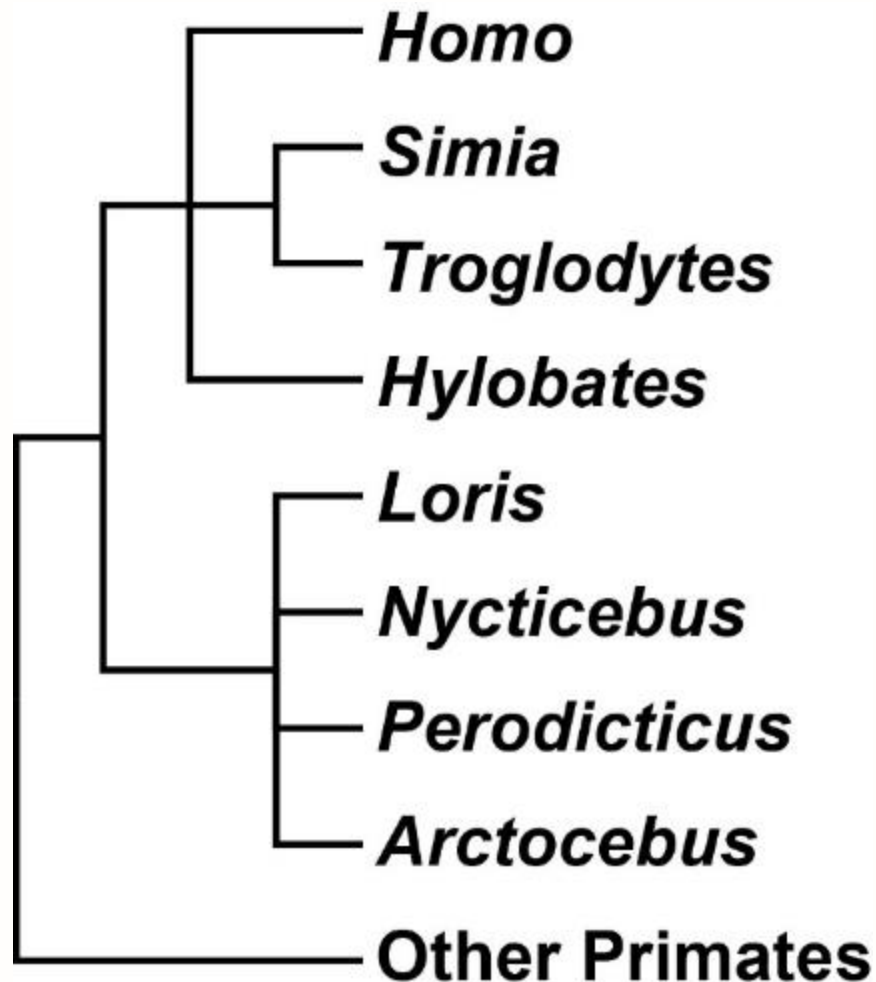
**St. George Jackson
Mivart**



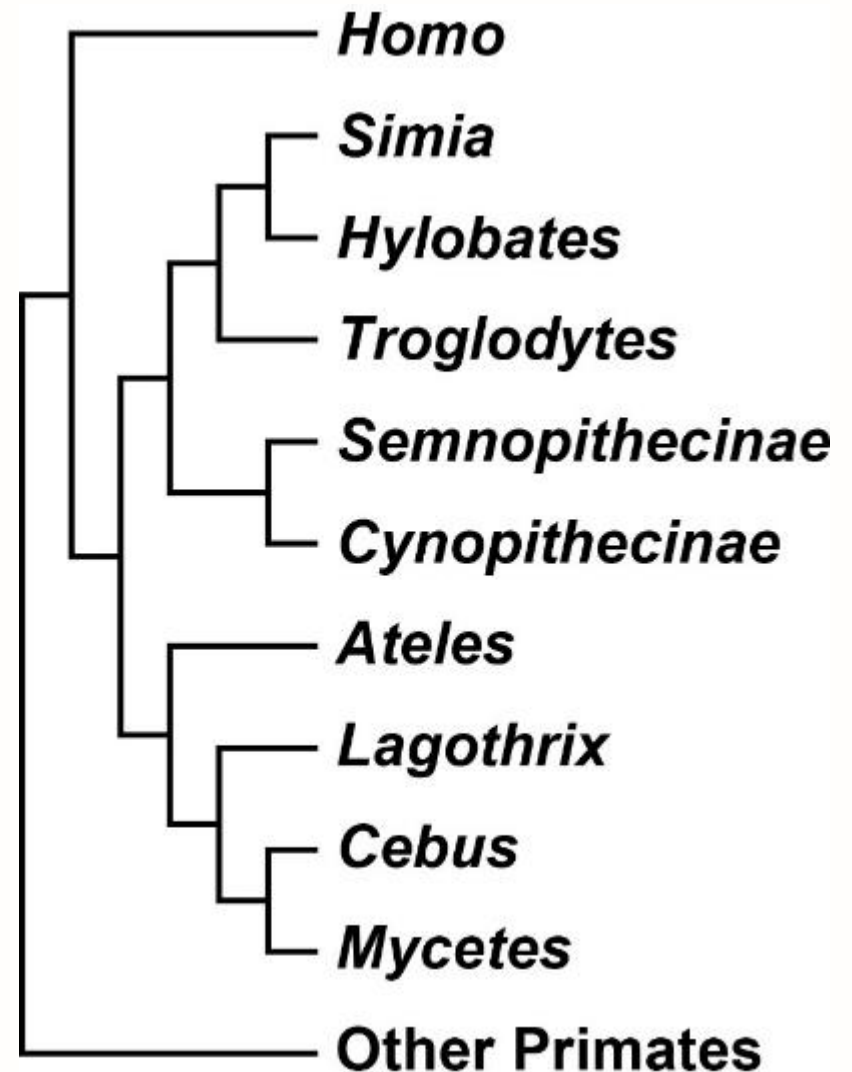
Mivart (1865) Proc. Zool. Soc. London

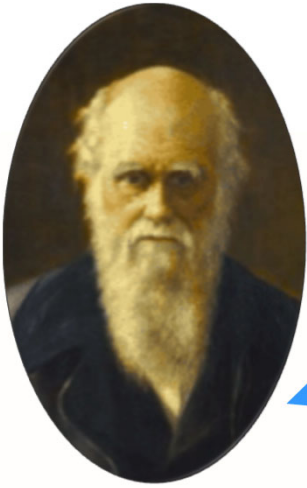
Discordance Between Trees There from the Beginning

1865: SPINAL COLUMN



1867: LIMBS



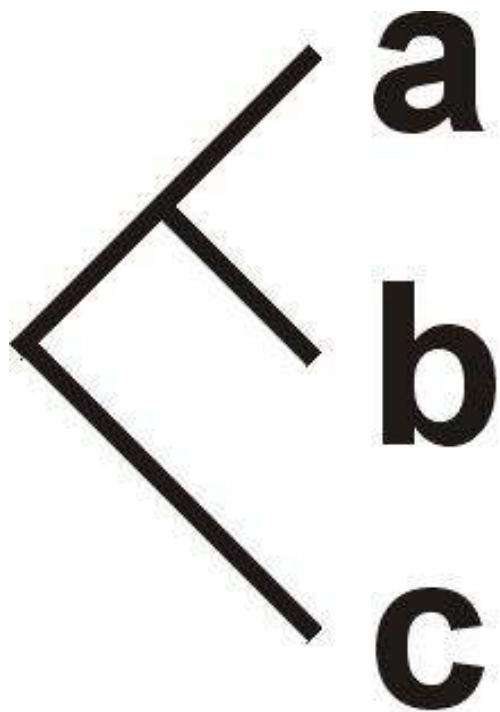


In some M.S. [... I say] that on genealogical principles alone, & considering whole organisation man probably diverged from the Catarhine stem a little below the branch of the anthropo:apes [...]. I have then added in my M.S. that this is your opinion [...]. Is this your opinion?

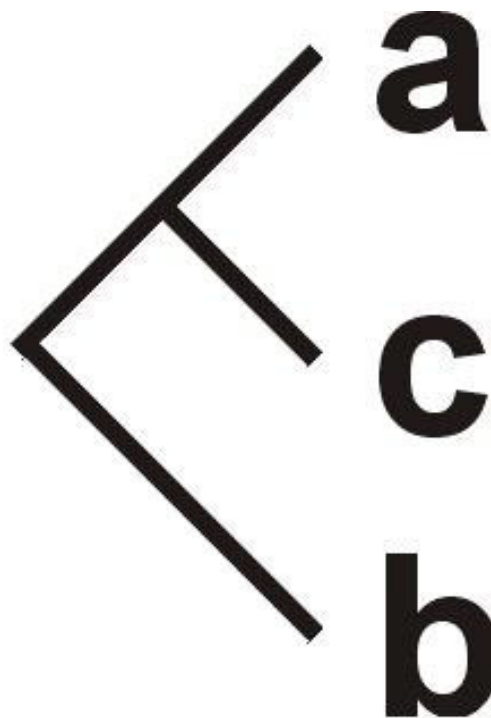
I have really expressed no opinion as to Man's origin nor am I prepared to do so at this moment. The [1865] diagram [...] expresses what I believe to be the degree of resemblance as regards the spinal column *only*. The [1867] diagram expresses what I believe to be the degree of resemblance as regards the appendicular skeleton *only*



The Problem of Incongruence



Gene X

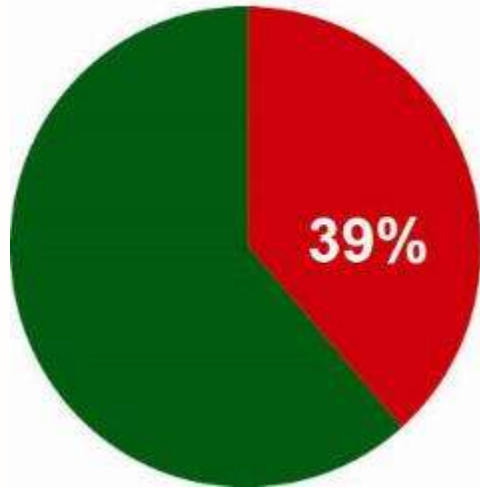


Gene Y

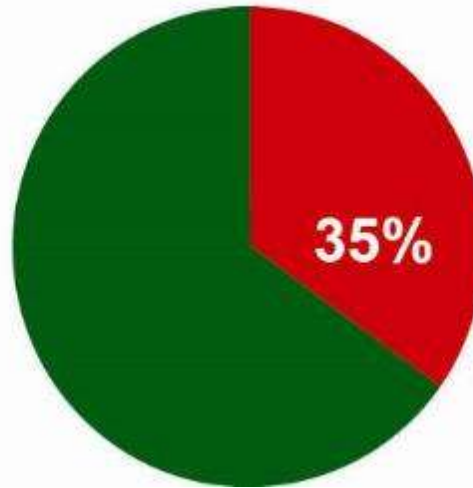
**Species
phylogeny?**

Incongruence is Pervasive in the Phylogenetics Literature

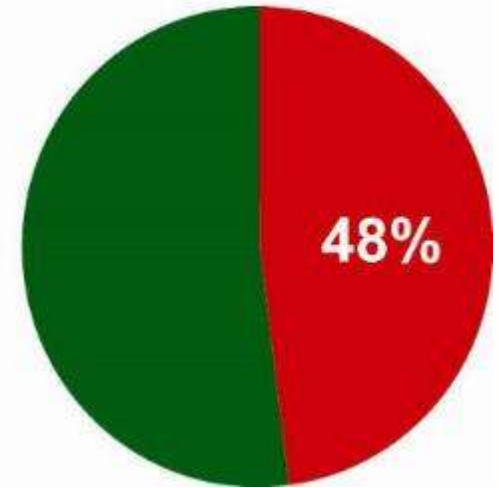
A: All organisms



B: Mammals



C: Insects



A Systematic Evaluation of Single Gene Phylogenies



S. cerevisiae

S. paradoxus

S. mikatae

S. kudriavzevii

S. bayanus

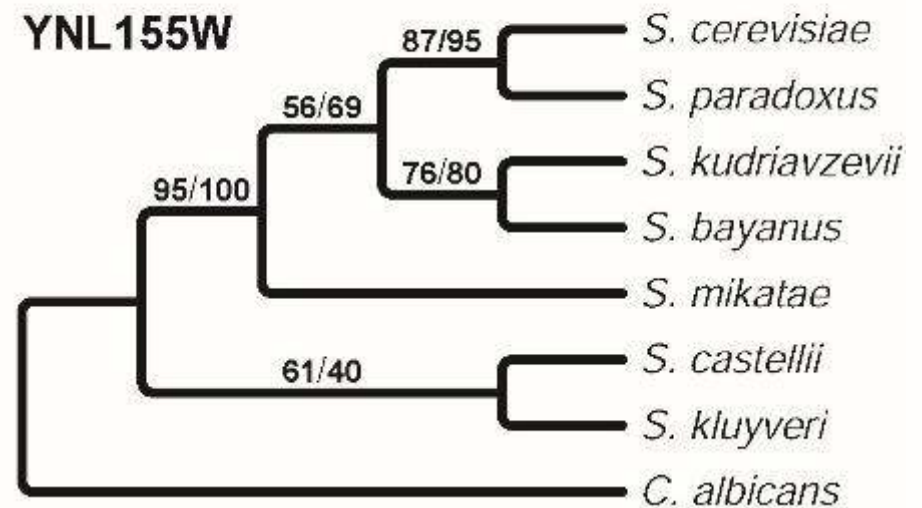
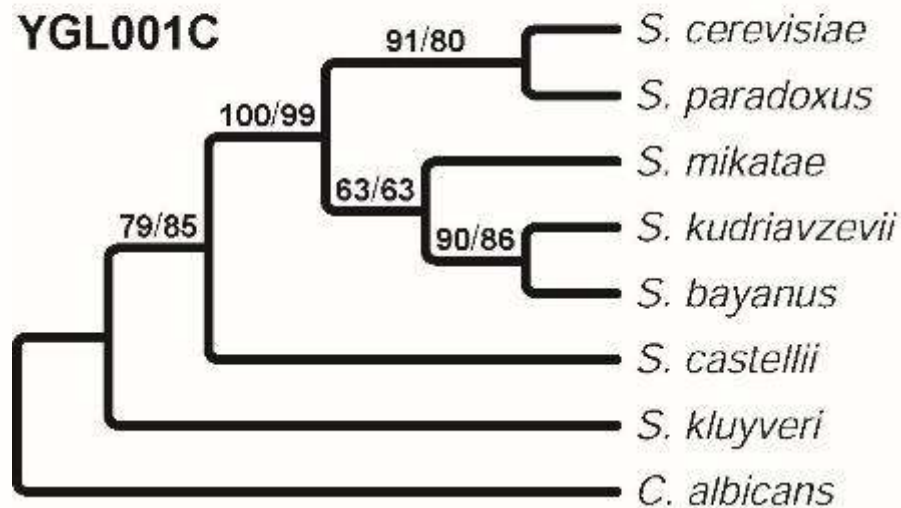
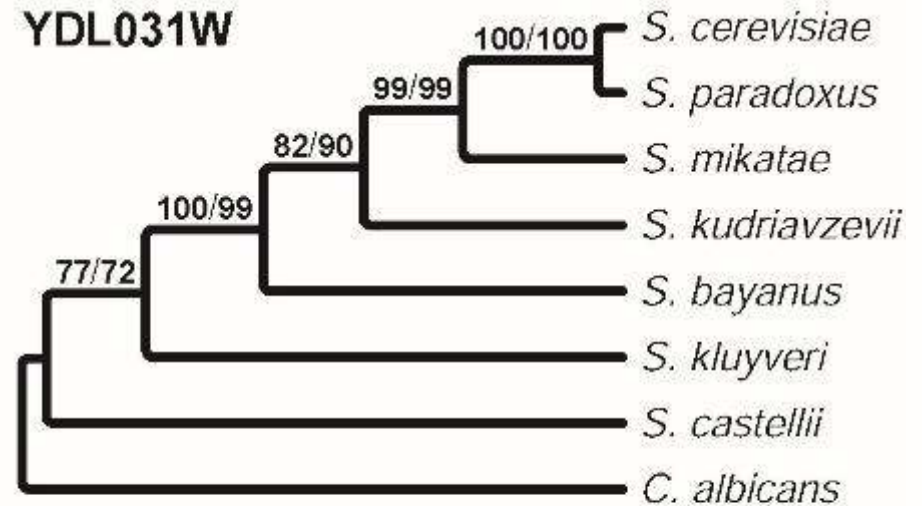
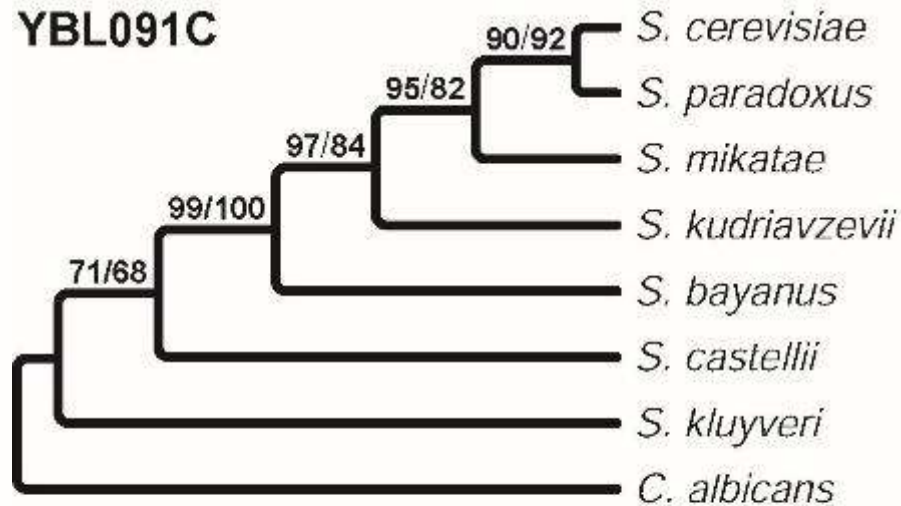
S. castellii

S. kluyveri

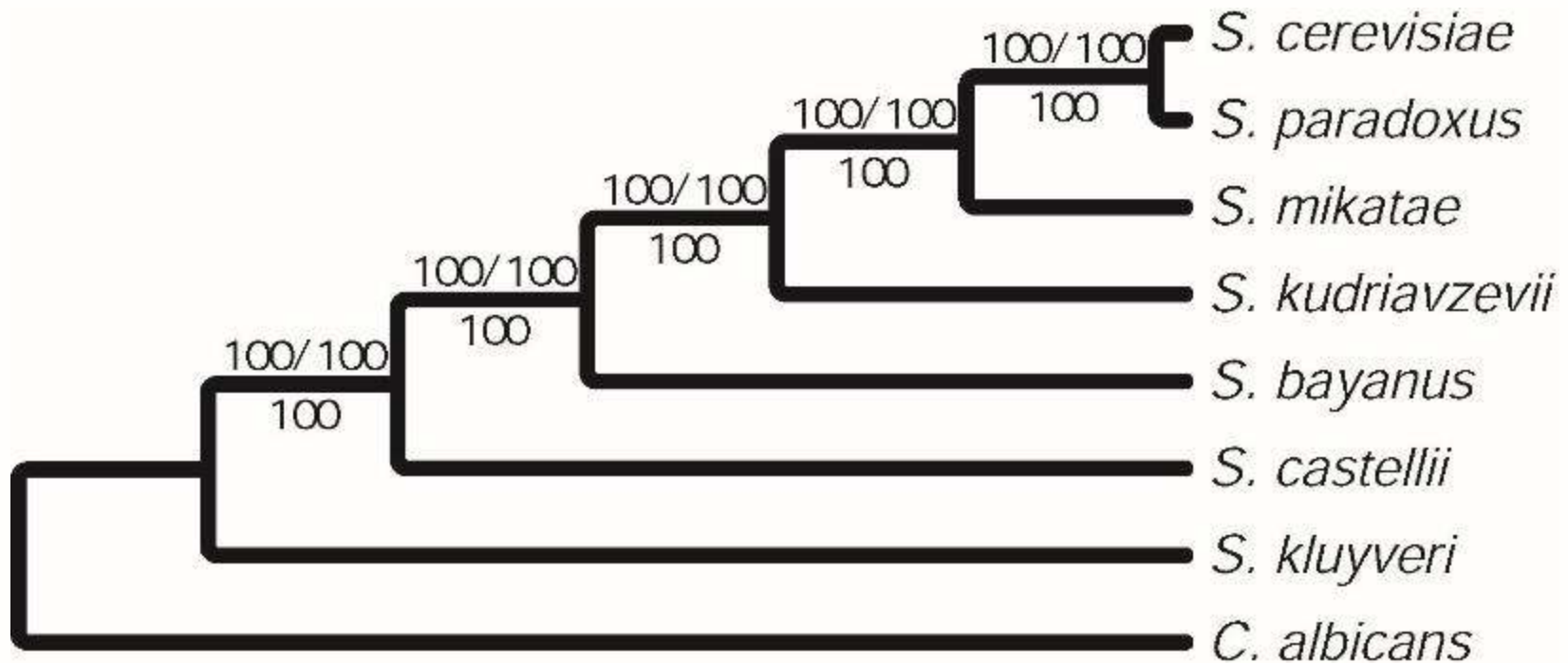
Candida glabrata



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

The Phylogenomics Era – “Resolving” the Tree of Life

Syst. Biol. 61(1):150–164, 2012

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

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

DOI:10.1093/sysbio/syr089

Advance Access publication on September 7, 2011

LETT
LETT

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

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

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

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

Resolving the evolutionary relationships of molluscs with phylogenomic tools

nature

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

Syst. Biol. 57(6):920–938, 2008

Copyright © Society of Systematic Biologists

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

DOI: 10.1080/10635150802570791

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

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

Toward Resolving the Tree: The Phylogeny of Jakobids and Cercozoans

Yeast

An

Toward Resolving Priors

Prion-Like Proteins in the Fungal Kingdom

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

OPEN ACCESS Free

Towards

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

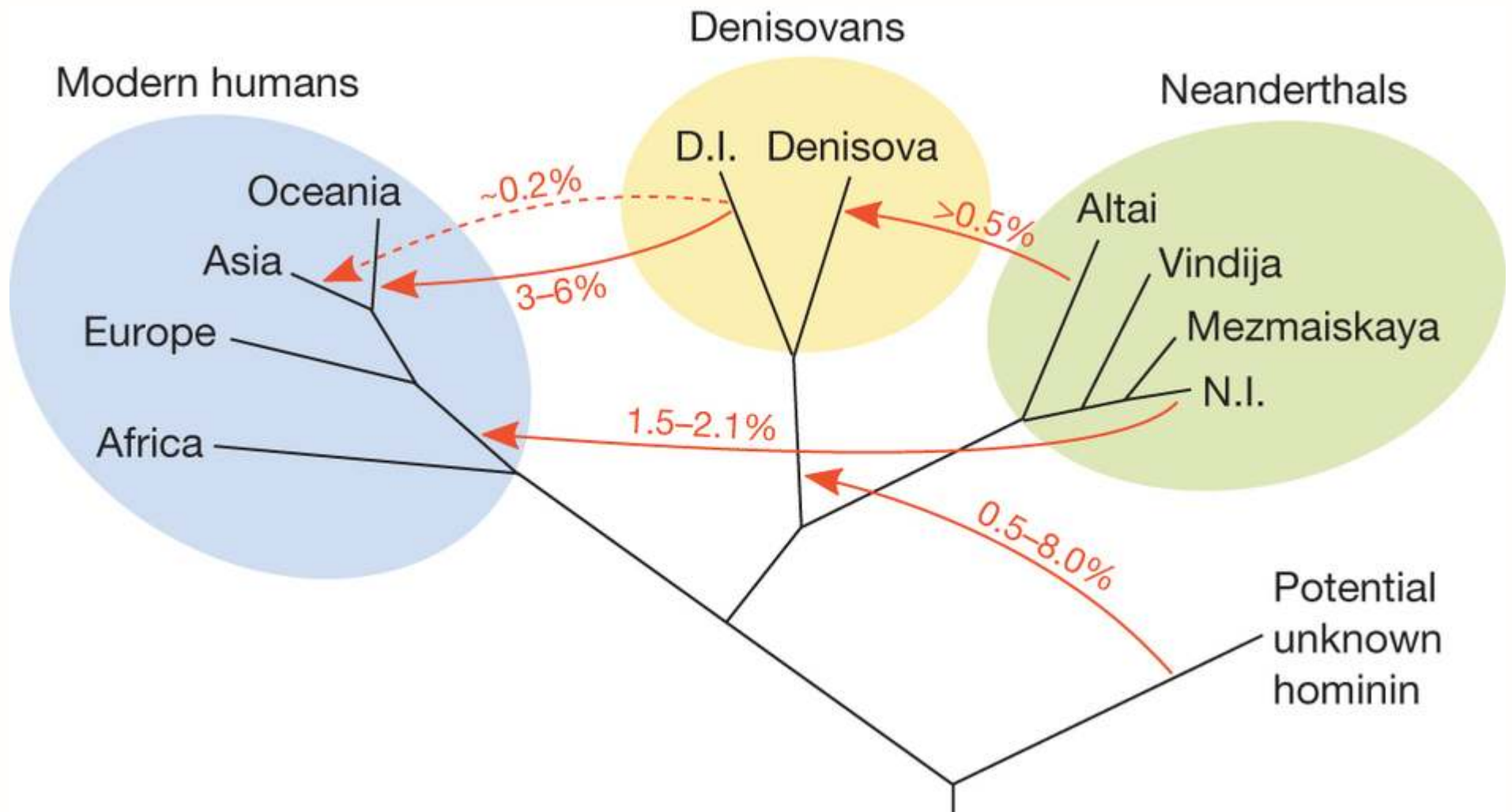
Samuli Lehtonen

Department of Biology, U

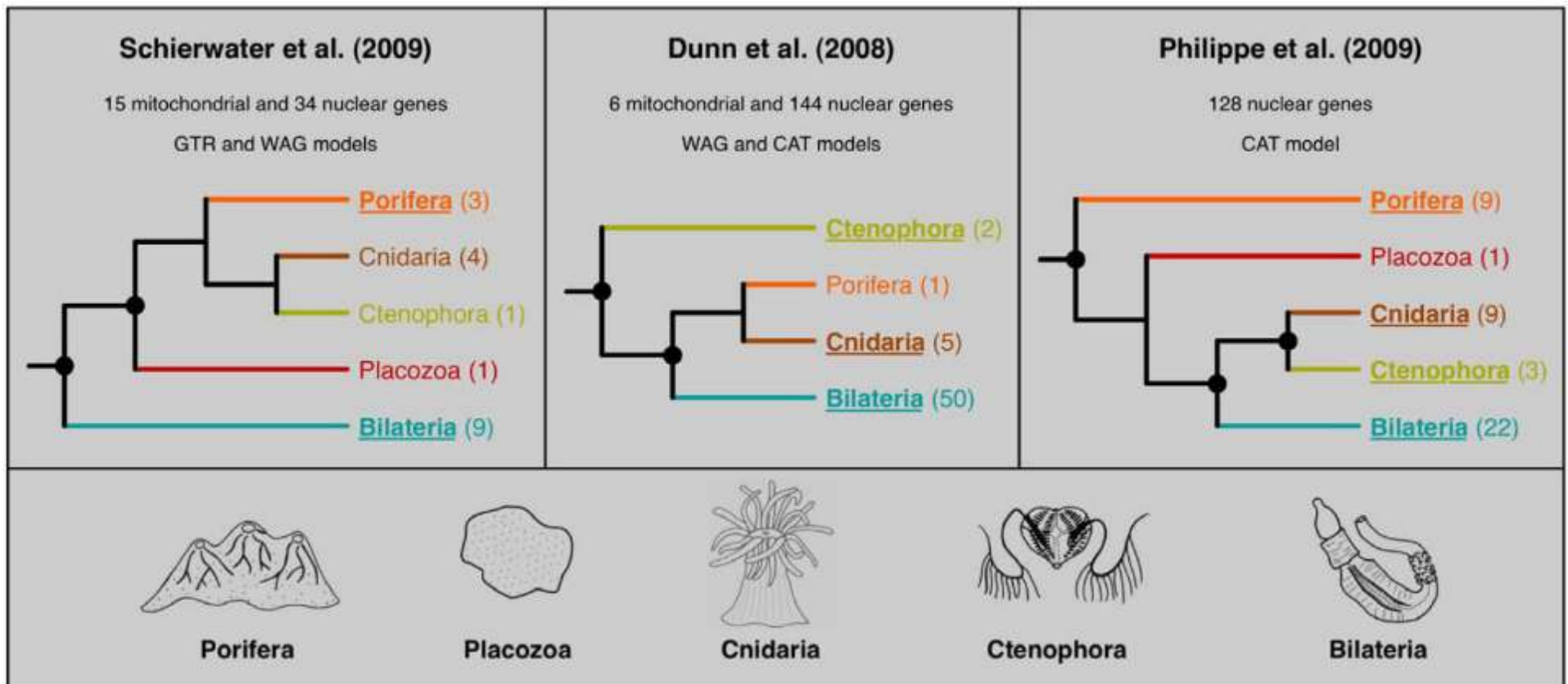
*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and †Centre for Macroevolution and Macroecology, School of Botany and Zoology, Australian National University, Canberra ACT, Australia

**Have we eliminated
incongruence?**

At “Shallow” Depths, True History is Easier Seen & Quantified



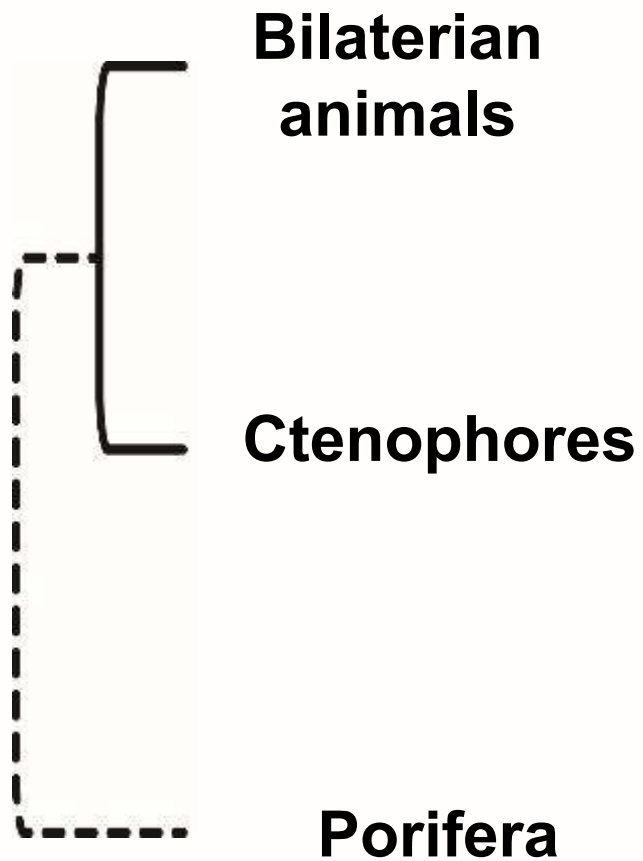
Incongruence in Deep Time is More Challenging



Incongruence in Deep Time is More Challenging

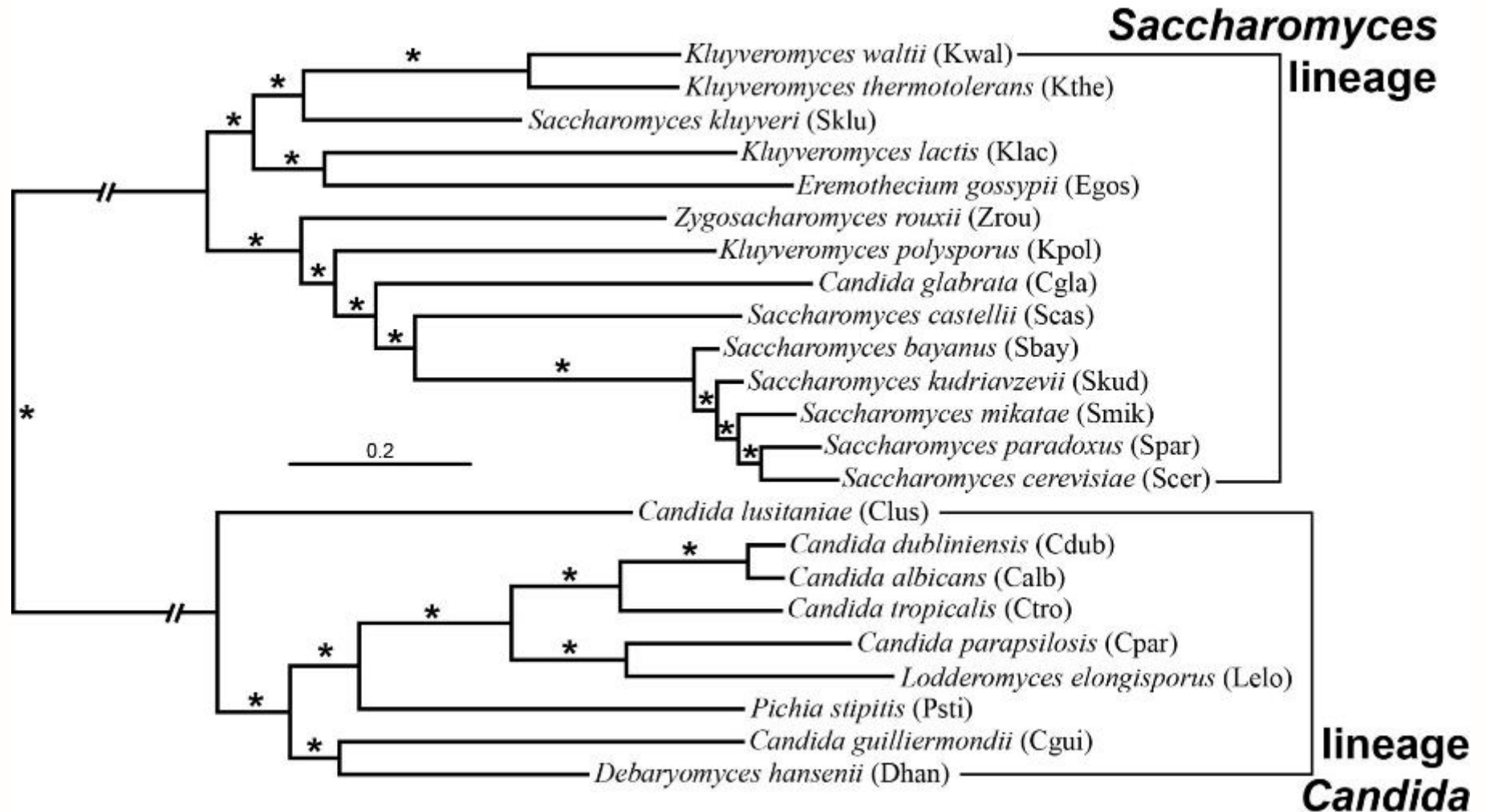


Incongruence in Deep Time is More Challenging

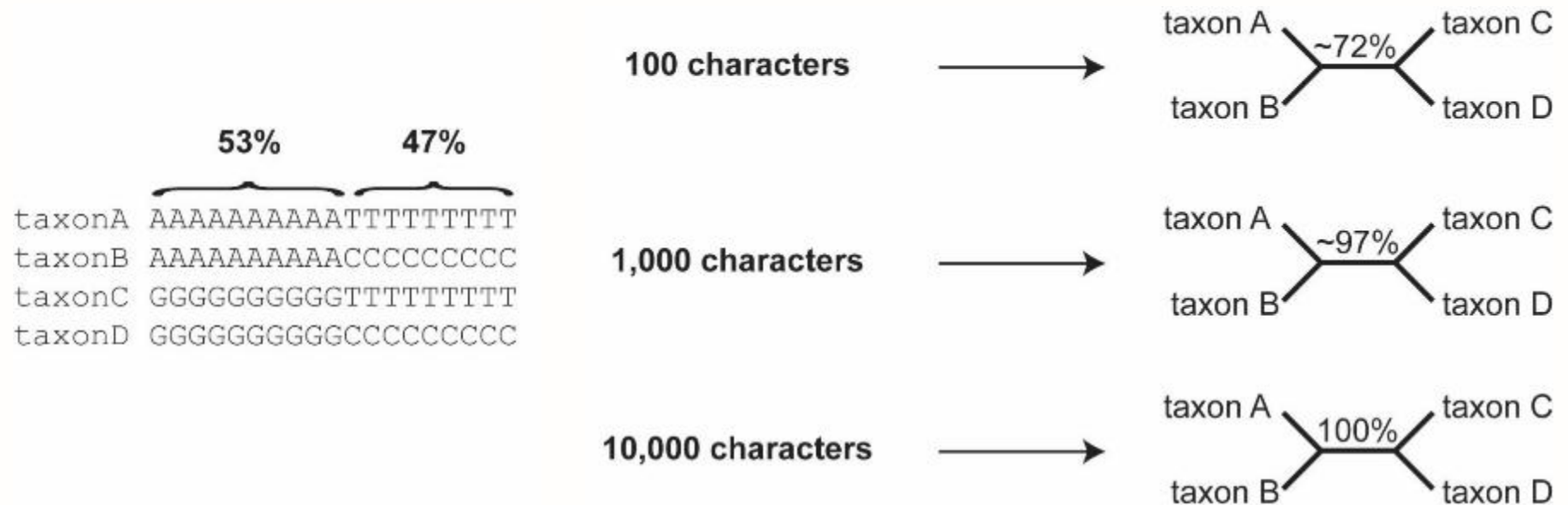


Why the disconnect?

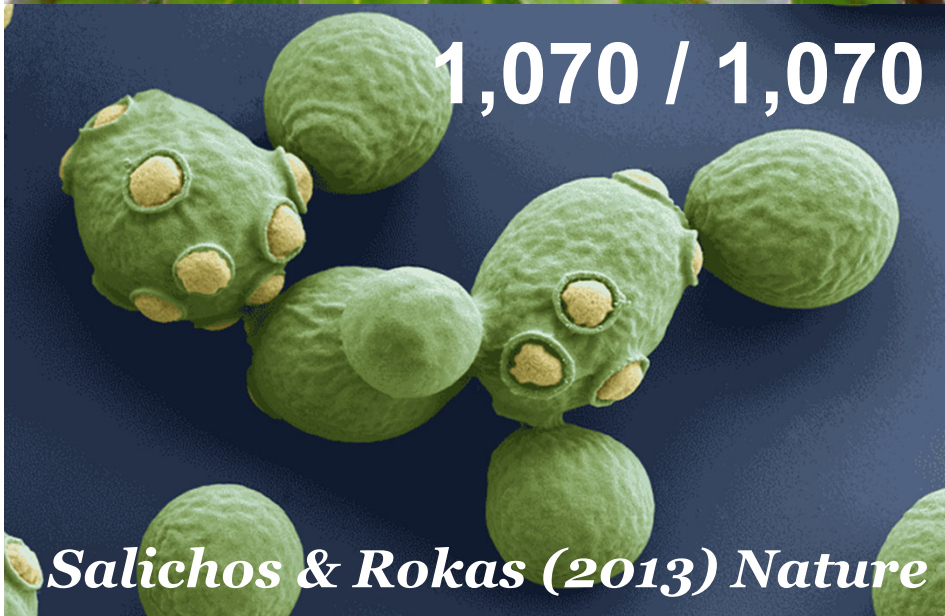
Concatenation Yields an Absolutely Supported Phylogeny



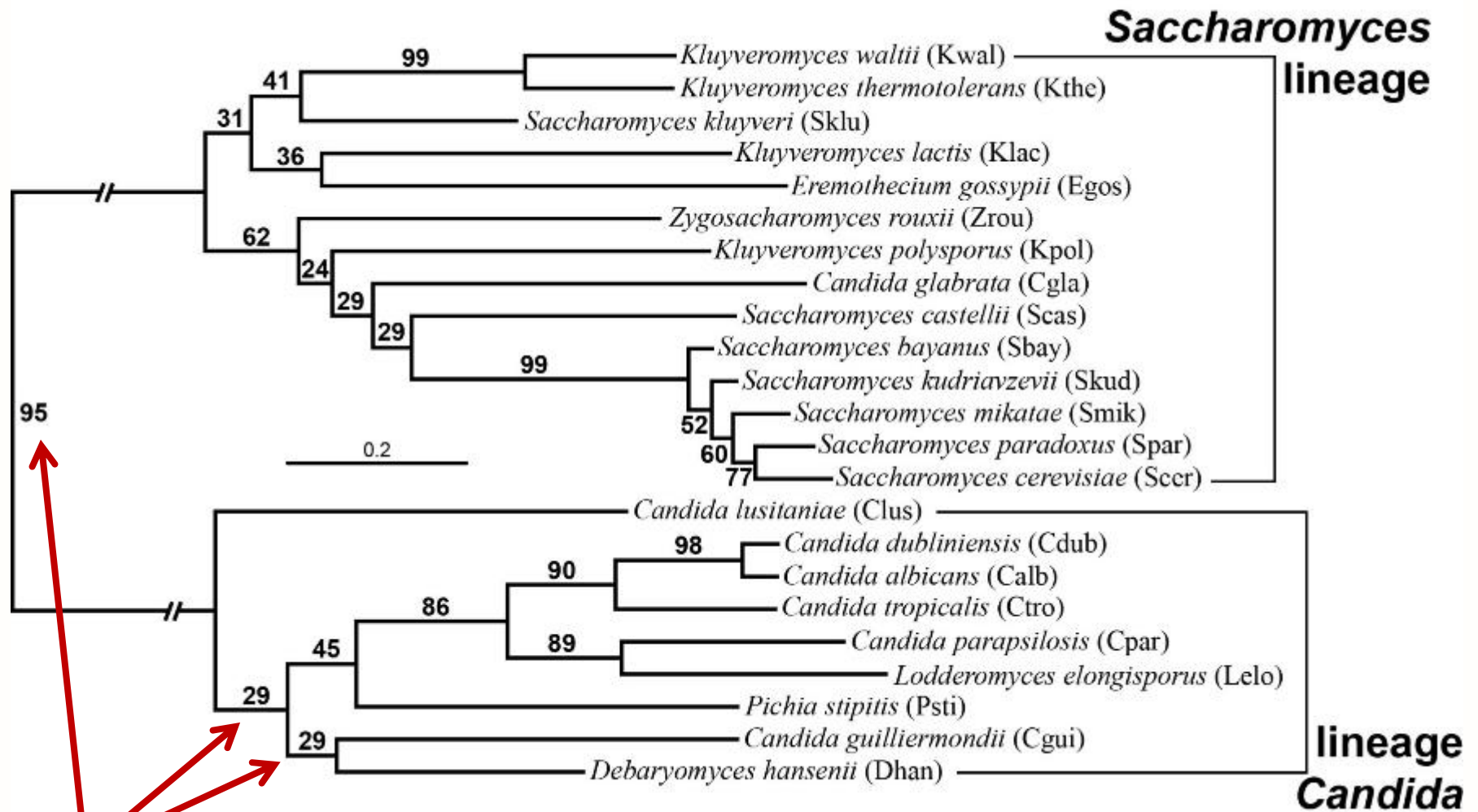
Bootstrap Support is Misleading When Used in Large Datasets



Gene Trees are Incongruent in Most Datasets



The Yeast Phylogeny Inferred by Majority-Rule Consensus

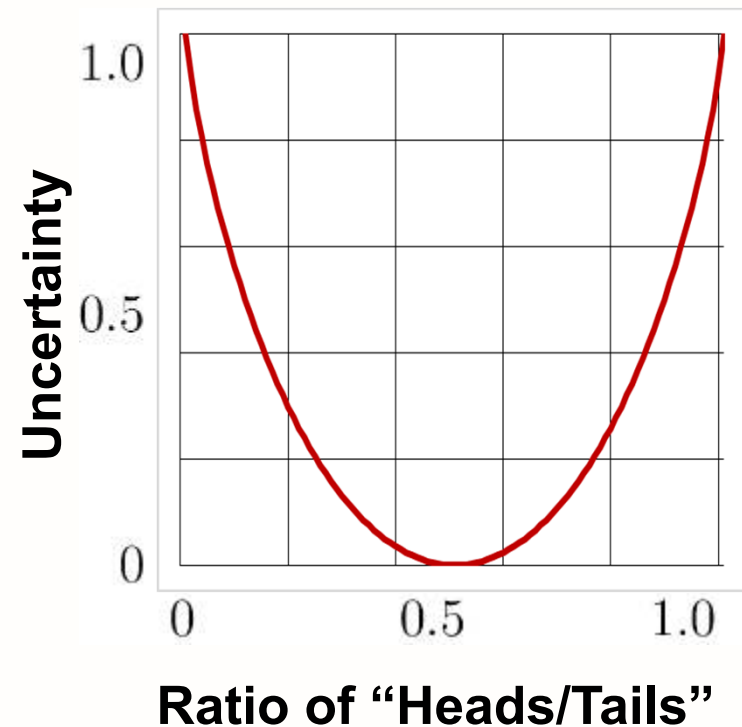


Quantifying Incongruence

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

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

IC and TC are implemented in the latest versions of RAxML

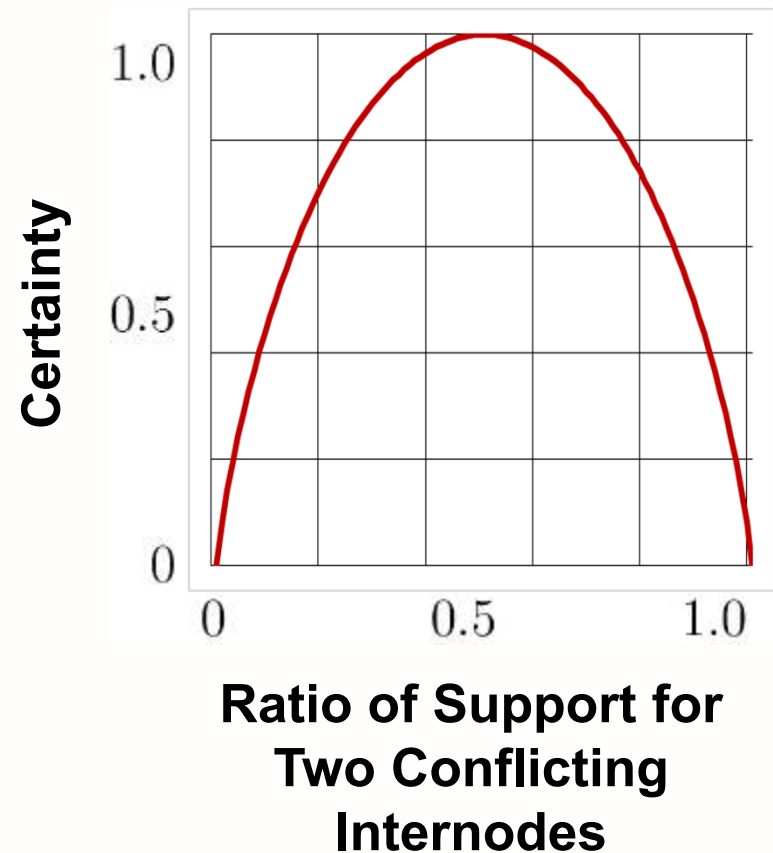


Quantifying Incongruence

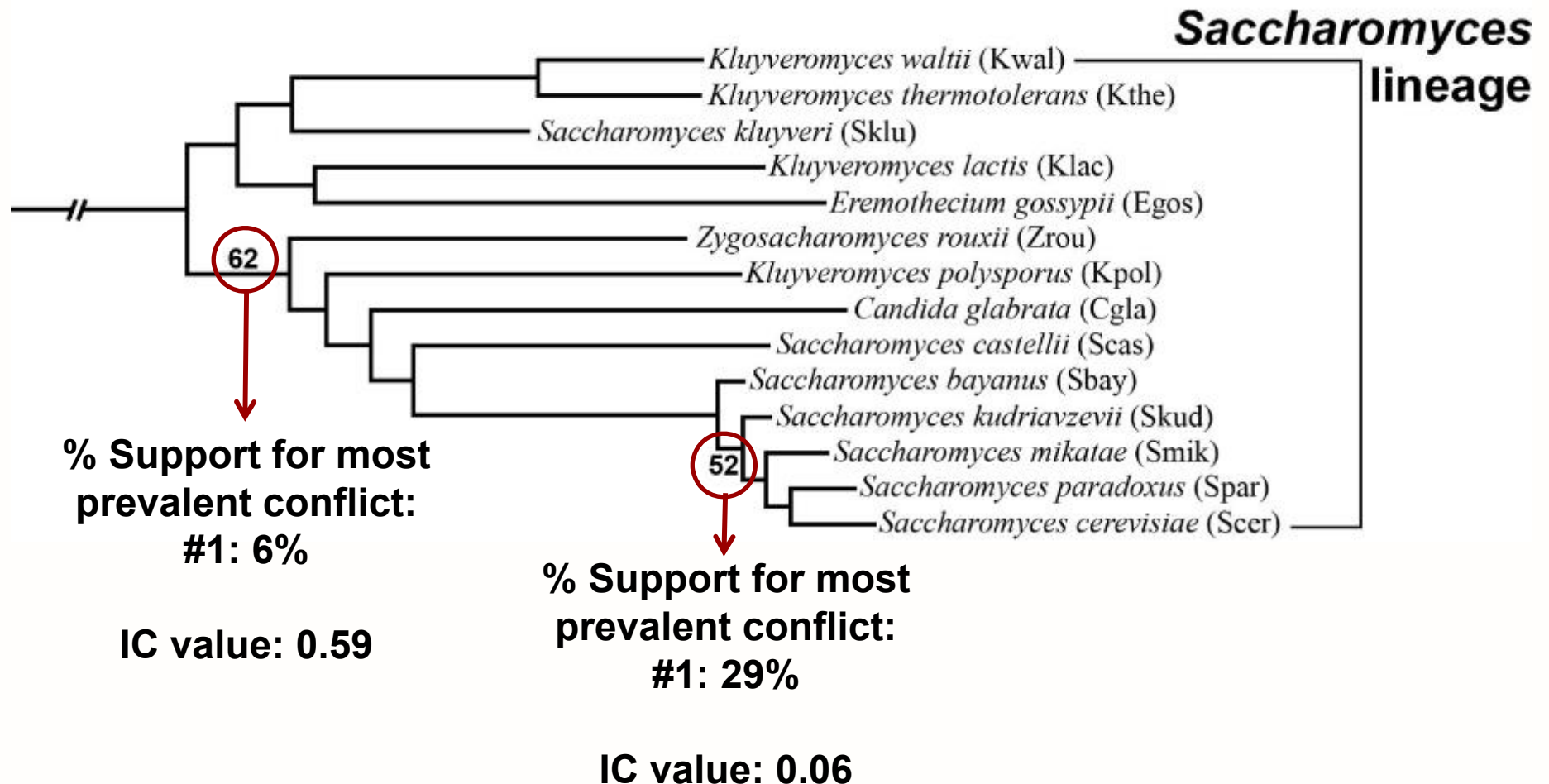
Internode Certainty (IC): a measure of the support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting internode in the same set of trees

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

IC and TC are implemented in the latest versions of RAxML



A More Informative Measure of Branch Support



Similar Results in Other Lineages

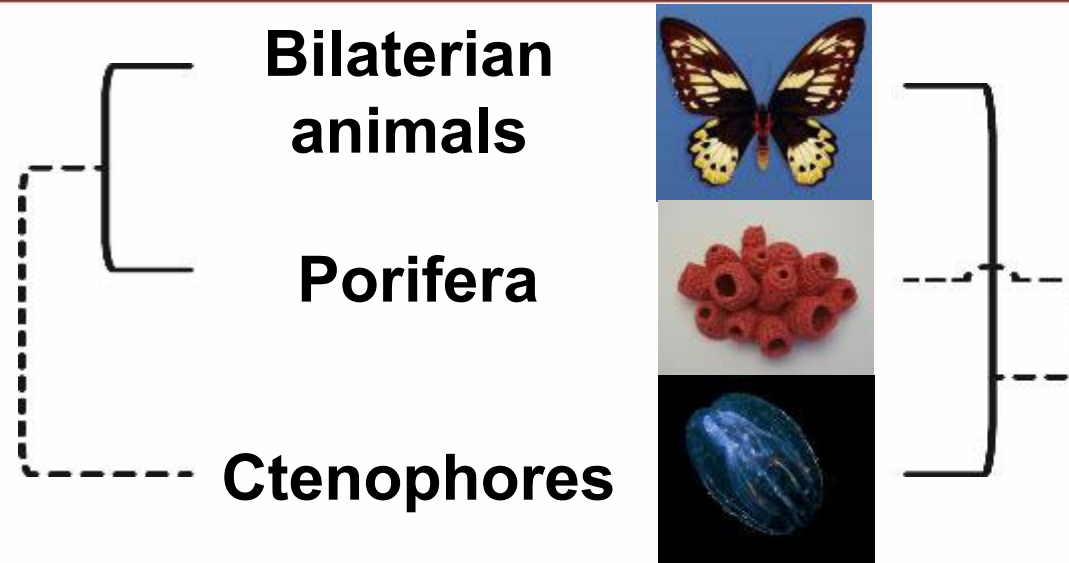
Vertebrates
(1,086 genes, 18 taxa)

Animals
(225 genes, 21 taxa)

Mosquitoes
(2,007 genes, 20 taxa)



Incongruence in Phylogenomic Datasets



These debates often concern relationships poorly supported by individual gene trees



**What is the phylogenetic
signal in branches of the tree
of life that are challenging to
resolve?**

Definitions of Phylogenetic Signal

A measure of the statistical dependence among species' trait values due to their phylogenetic relationships / the tendency of related species to resemble each other more than species drawn at random from the same tree

Revell et al. (2008) Syst. Biol.
Münkemüller et al. (2012) Methods Ecol. Evol.

The amount of support for a particular topology, e.g., the relative number of resolved internodes in a consensus tree

Sanderson (2008) Science

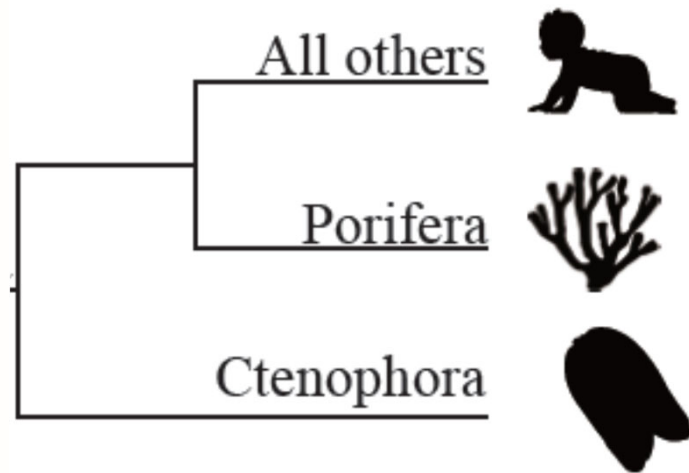
A measure of the substitutions occurring along a given branch of the evolutionary tree. In parsimony methods, the signal is encoded in shared derived characters. In probabilistic methods, the amount of phylogenetic signal actually extracted from a given dataset depends on the model and is expected to increase with the fit of the model to the data

Philippe et al. (2011) PLoS Biol.
Townsend et al. (2012) Syst. Biol.

Our Definition

Maximum Likelihood tree

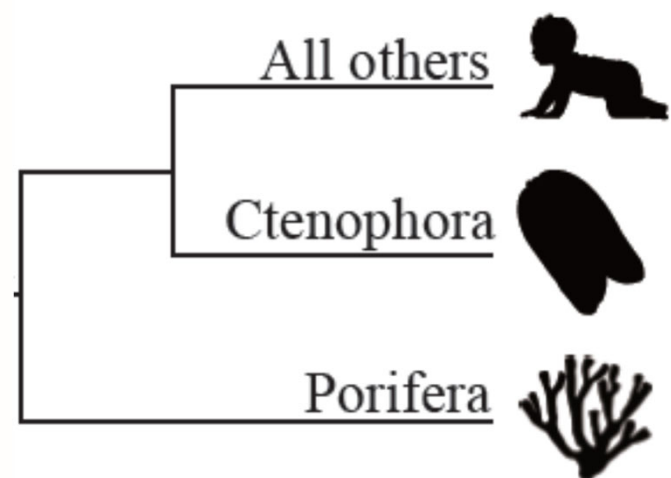
(T1)



$$\ln(T_1|X_i) = -100$$

Conflicting tree

(T2)



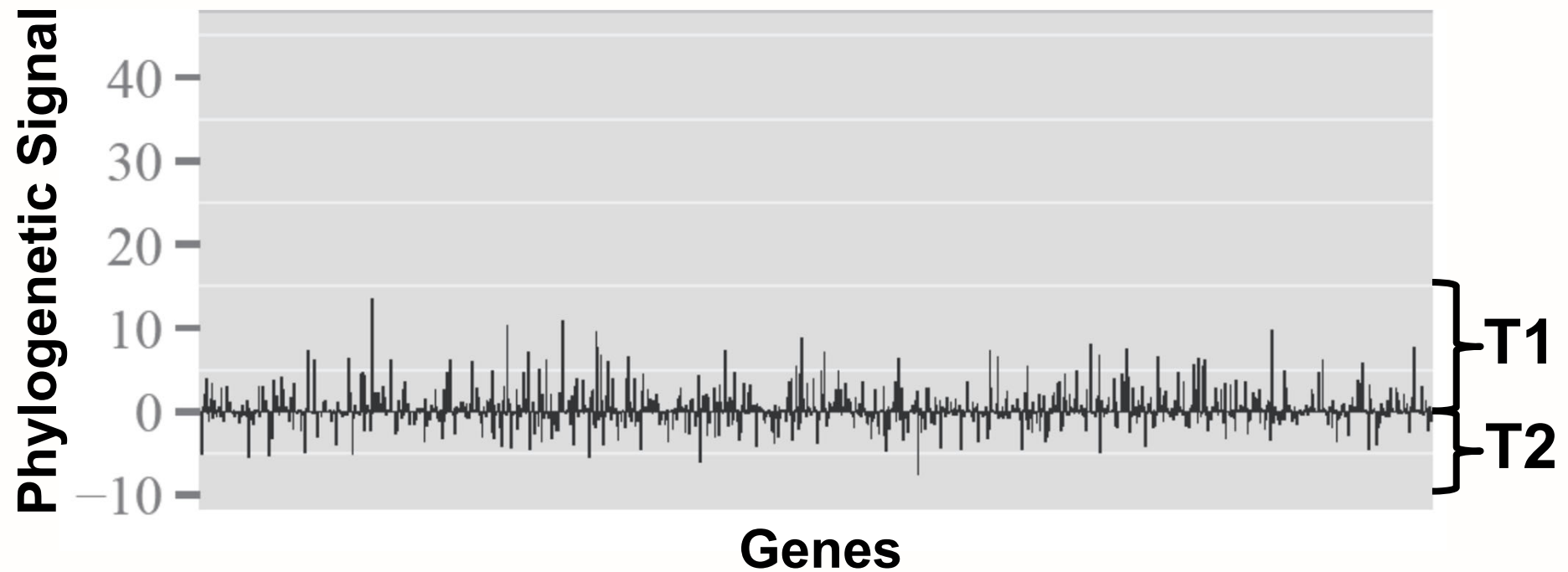
$$\ln(T_2|X_i) = -150$$

$$\text{Phylogenetic Signal} = -(\ln(T_1|X_i) - \ln(T_2|X_i))$$

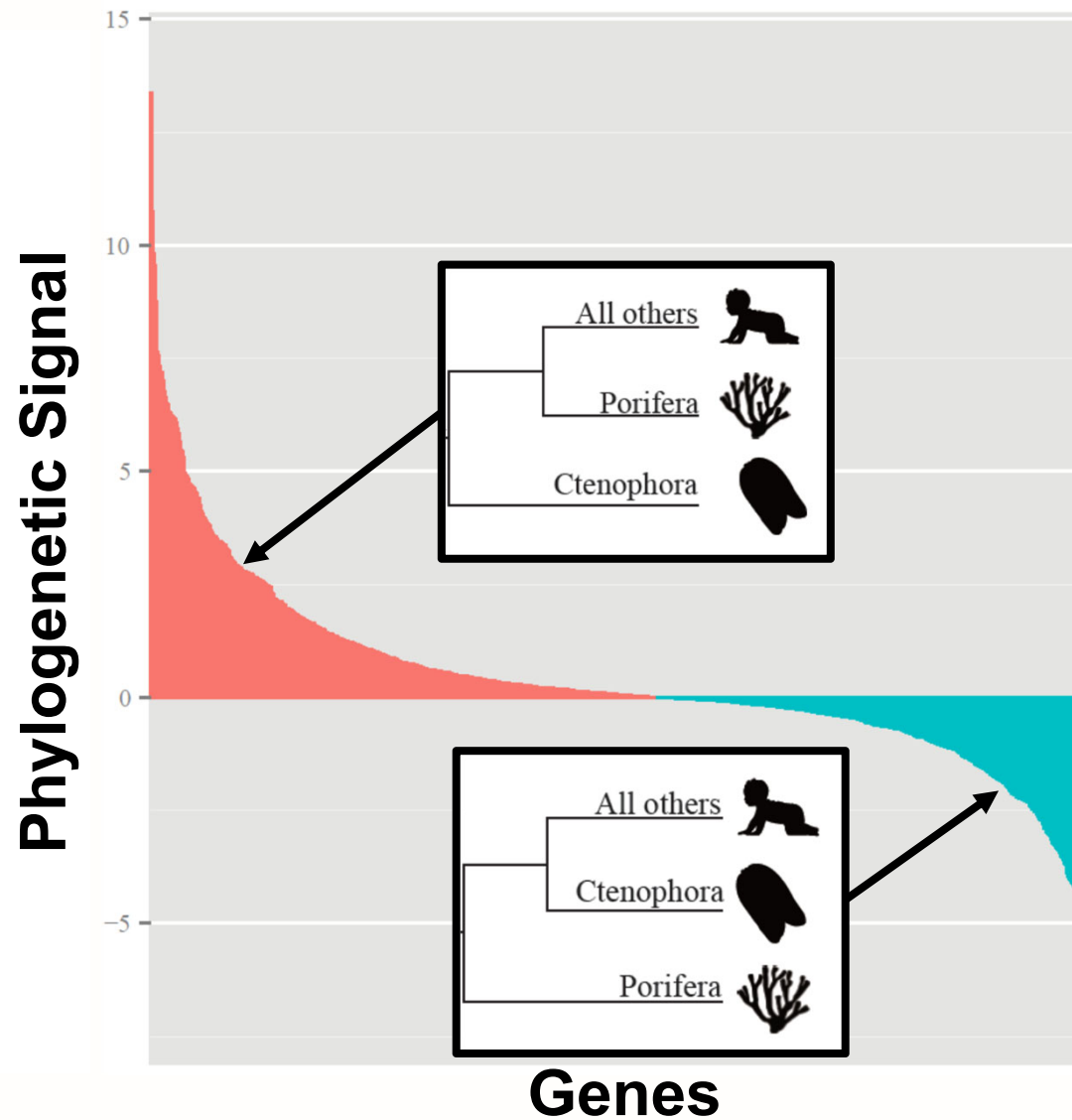


Signal of the Genes in a Phylogenomic Data Matrix

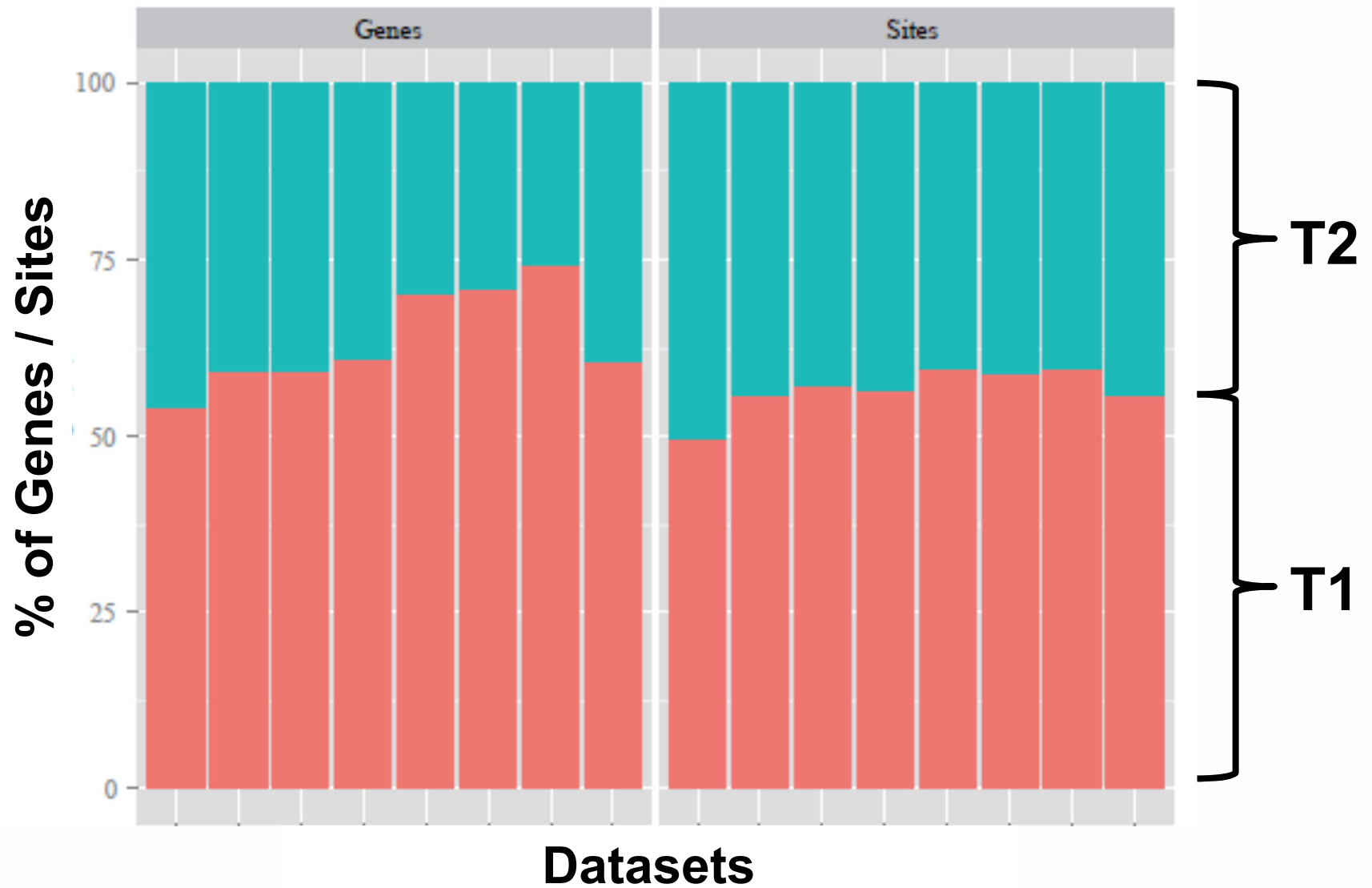
1,080 genes from 36 animal taxa



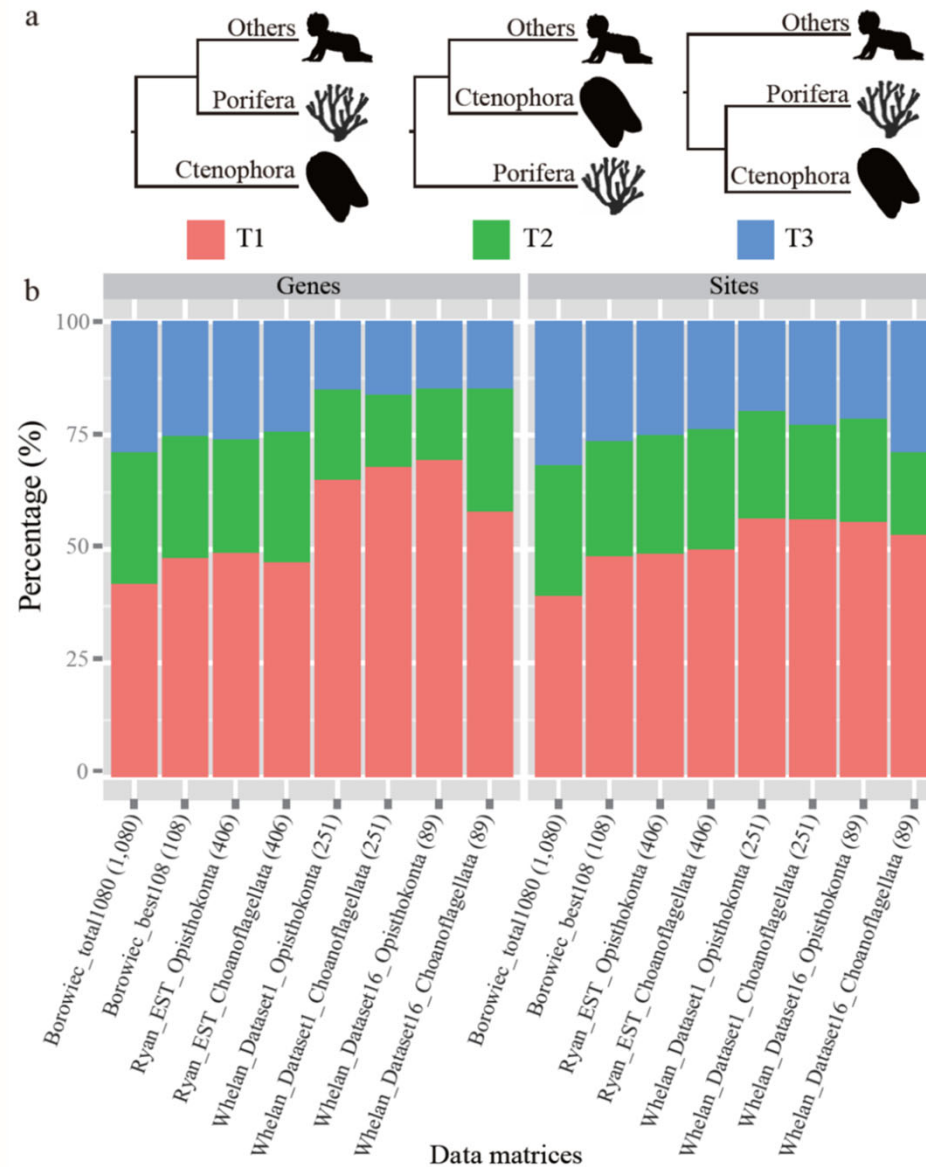
Signal of the Genes in a Phylogenomic Data Matrix



Summarizing Phylogenetic Signal Across Genes and Sites

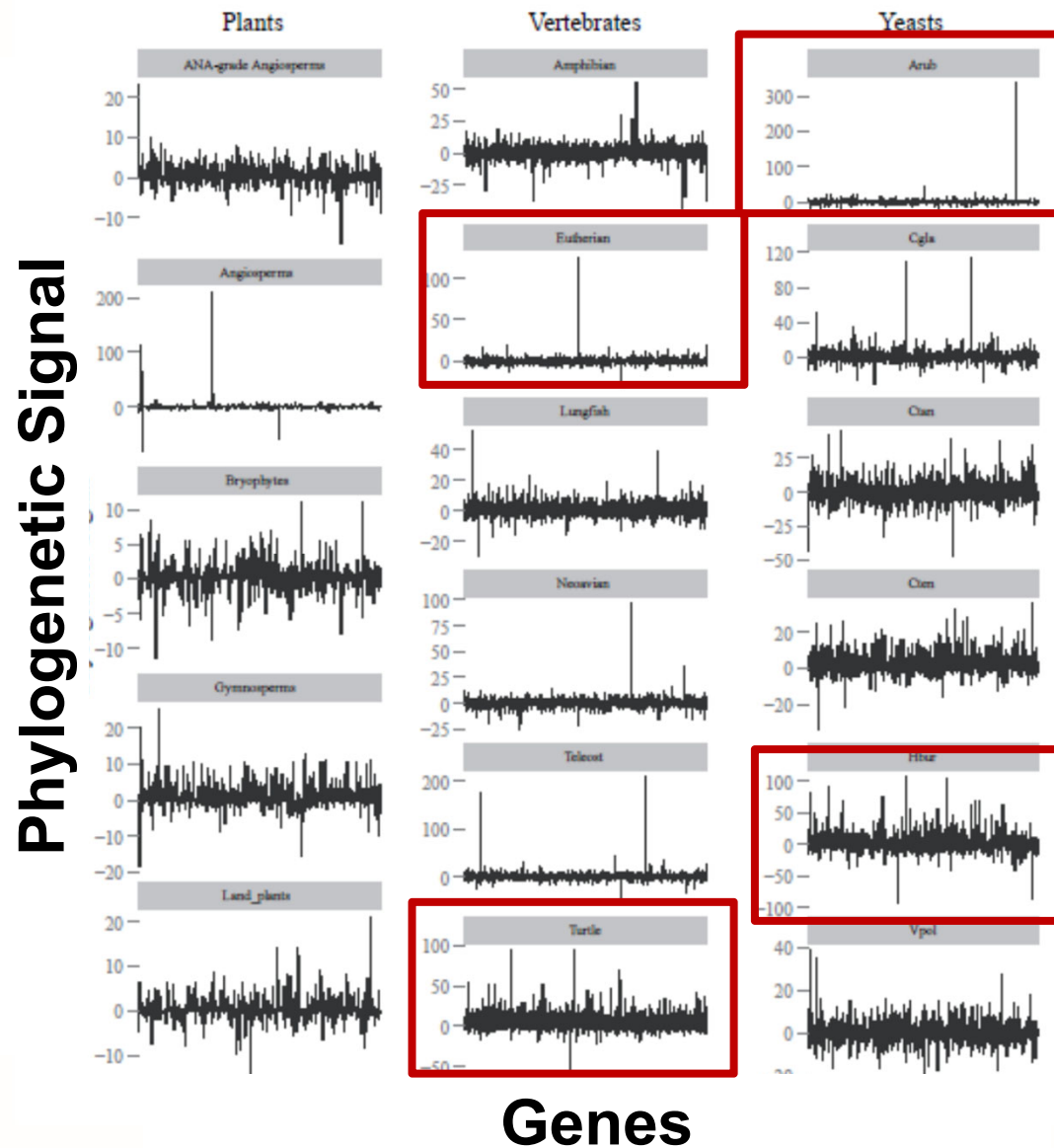


Summarizing the Signal Across All 3 Possible Topologies

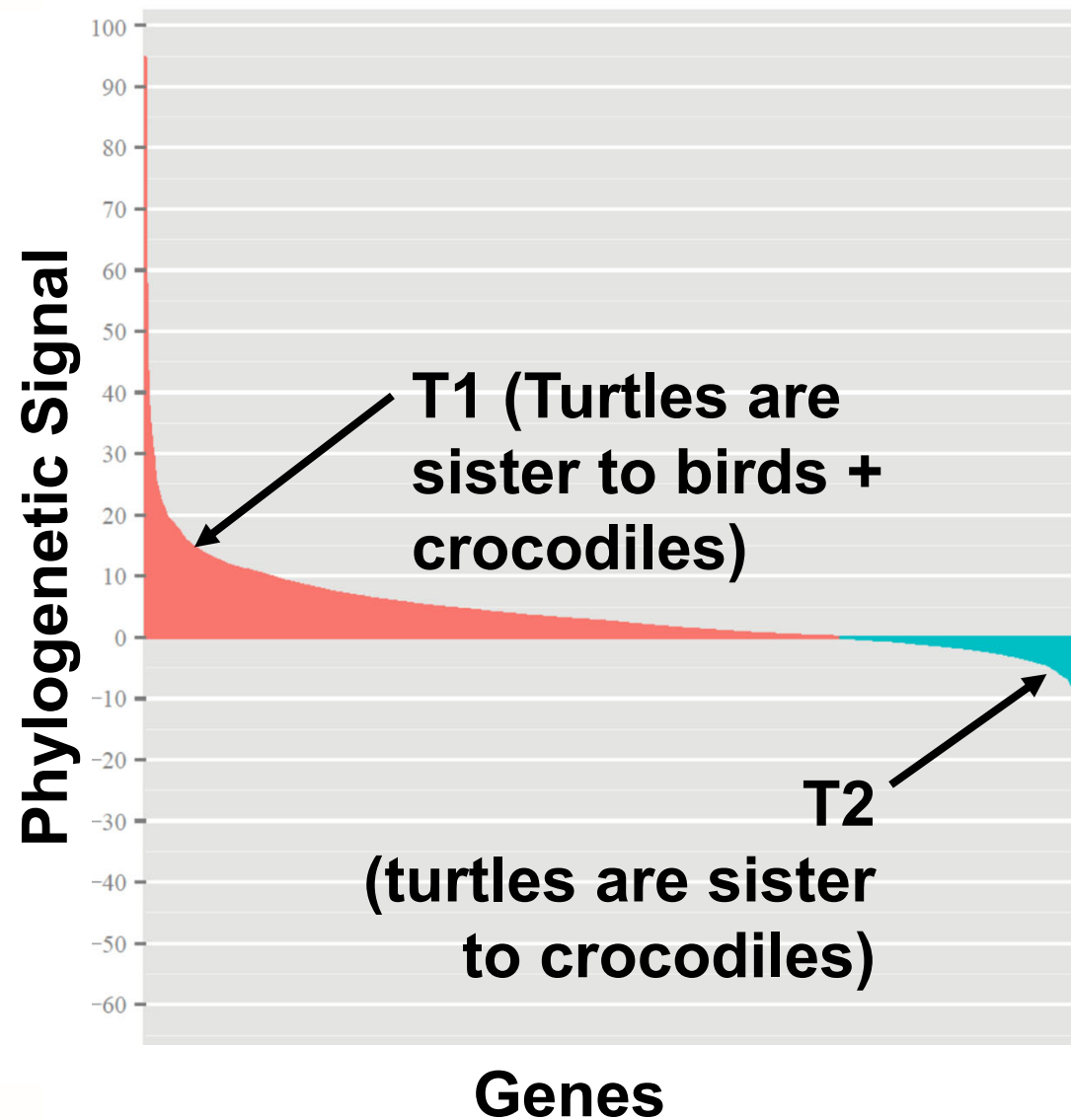


Shen et al. (2017) Nature Ecol. Evol.

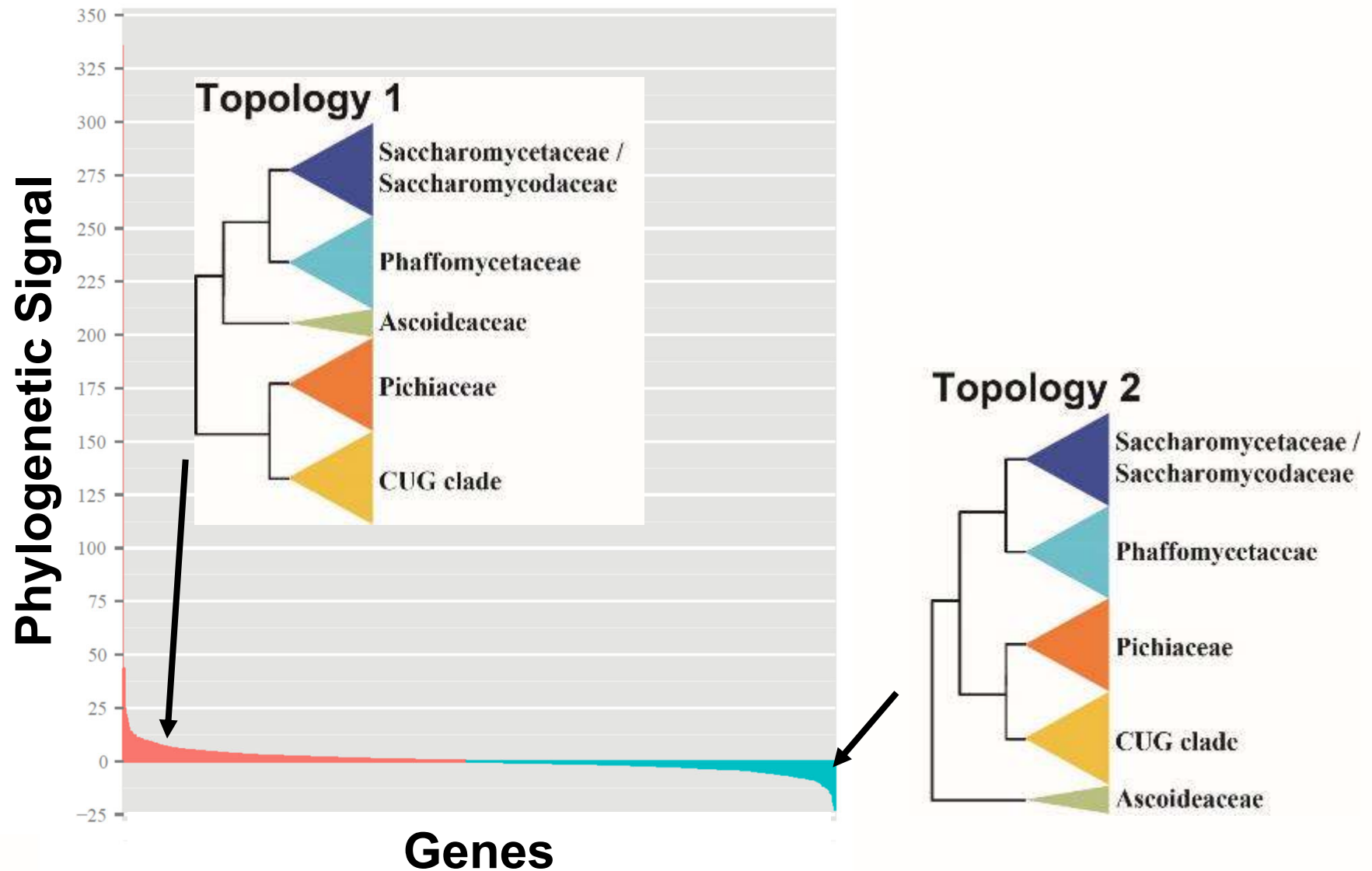
Phylogenetic Signal in Contentious Branches of the ToL



The Signal in Some Branches is Very Strong...

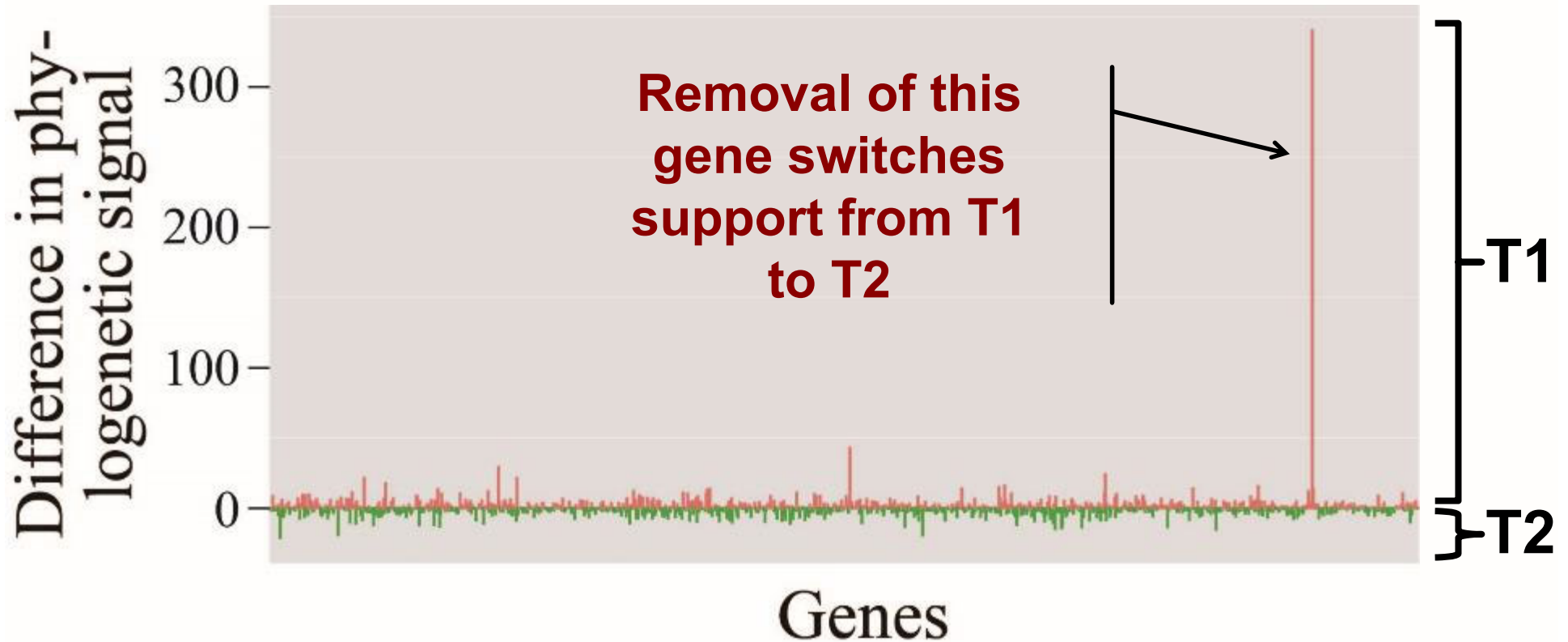


...But in Others It Stems from One or Two Genes

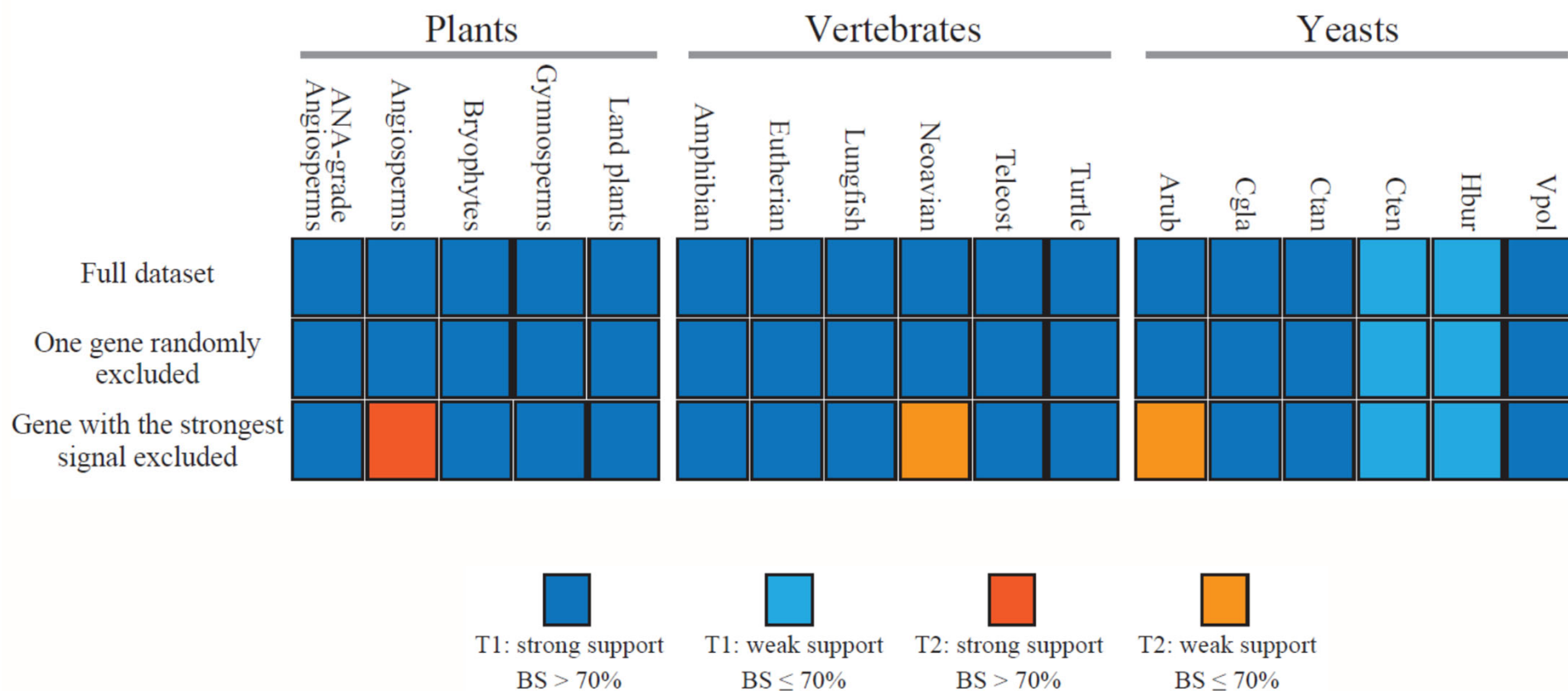


Phylogenetic Signal per Gene for the Two Hypotheses

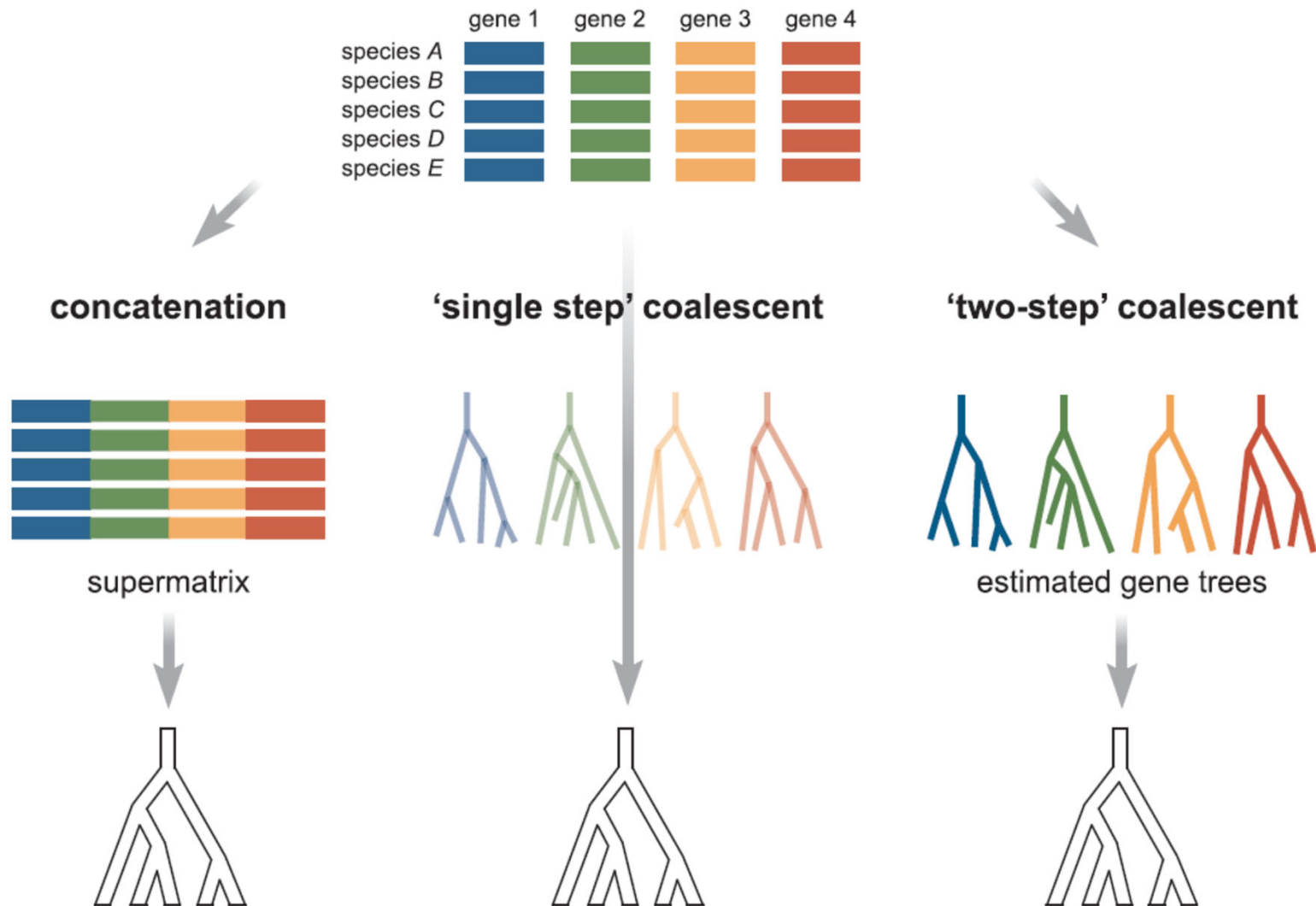
1233 genes, 86 yeast taxa



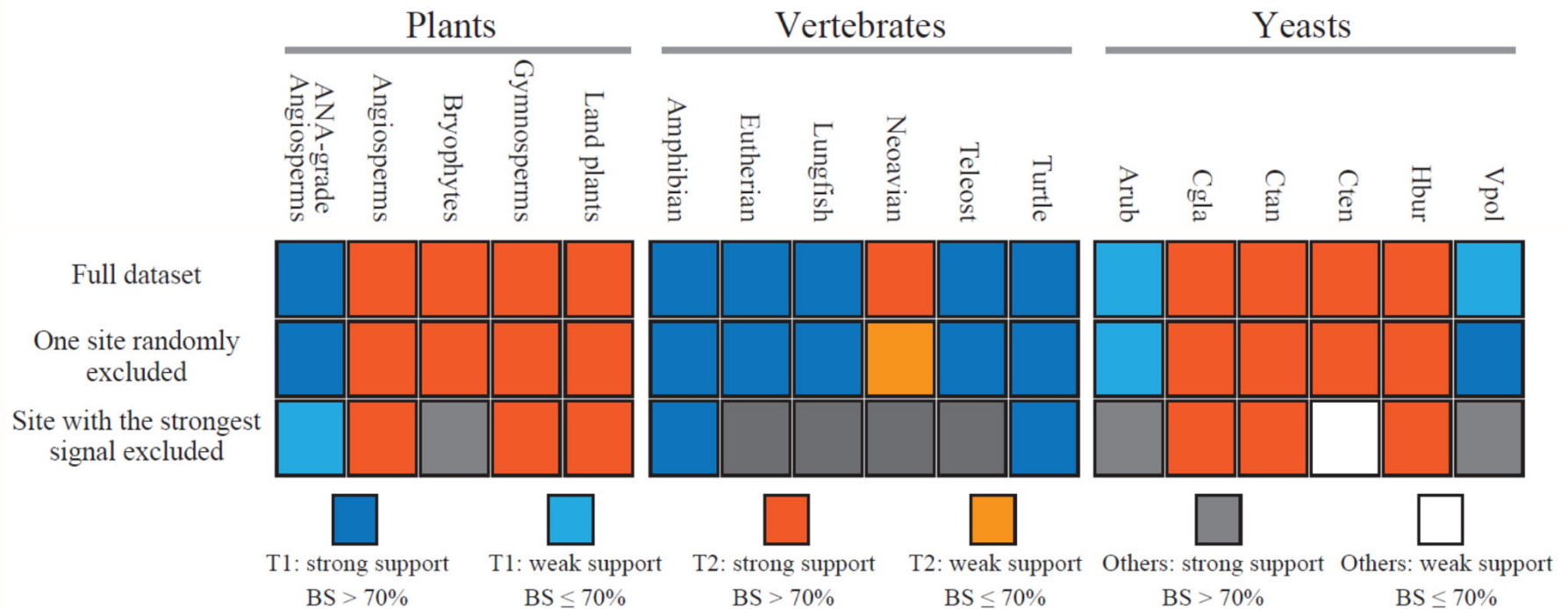
Removing One Gene Alters the Topology



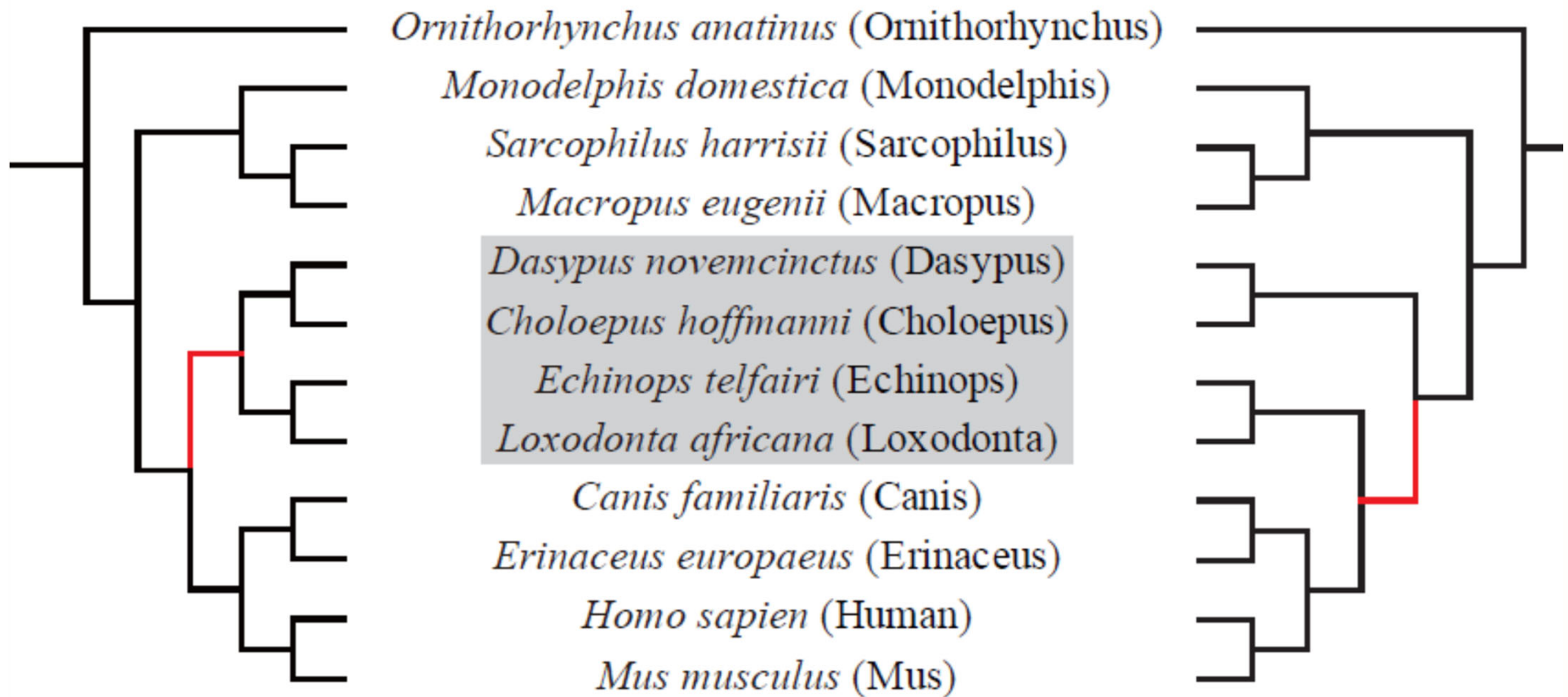
Methods for Phylogenomic Inference



What Happens if we Remove One Site from Every Gene?



Removing 1 Site Alters the Topology



What's Going On?

Explanation #1: Parts of the tree of life are bush-like / network-like (rather than tree-like)

Explanation #2: Our models don't fit well our data

The Phylogeny of Primate Genera

*Nomascus
leucogenys*



NLE

*Hoolock
leuconedys*



HLE

*Symphalangus
syndactylus*



SSY

*Hylobates
pileatus*



HPI

*Hylobates
moloch*

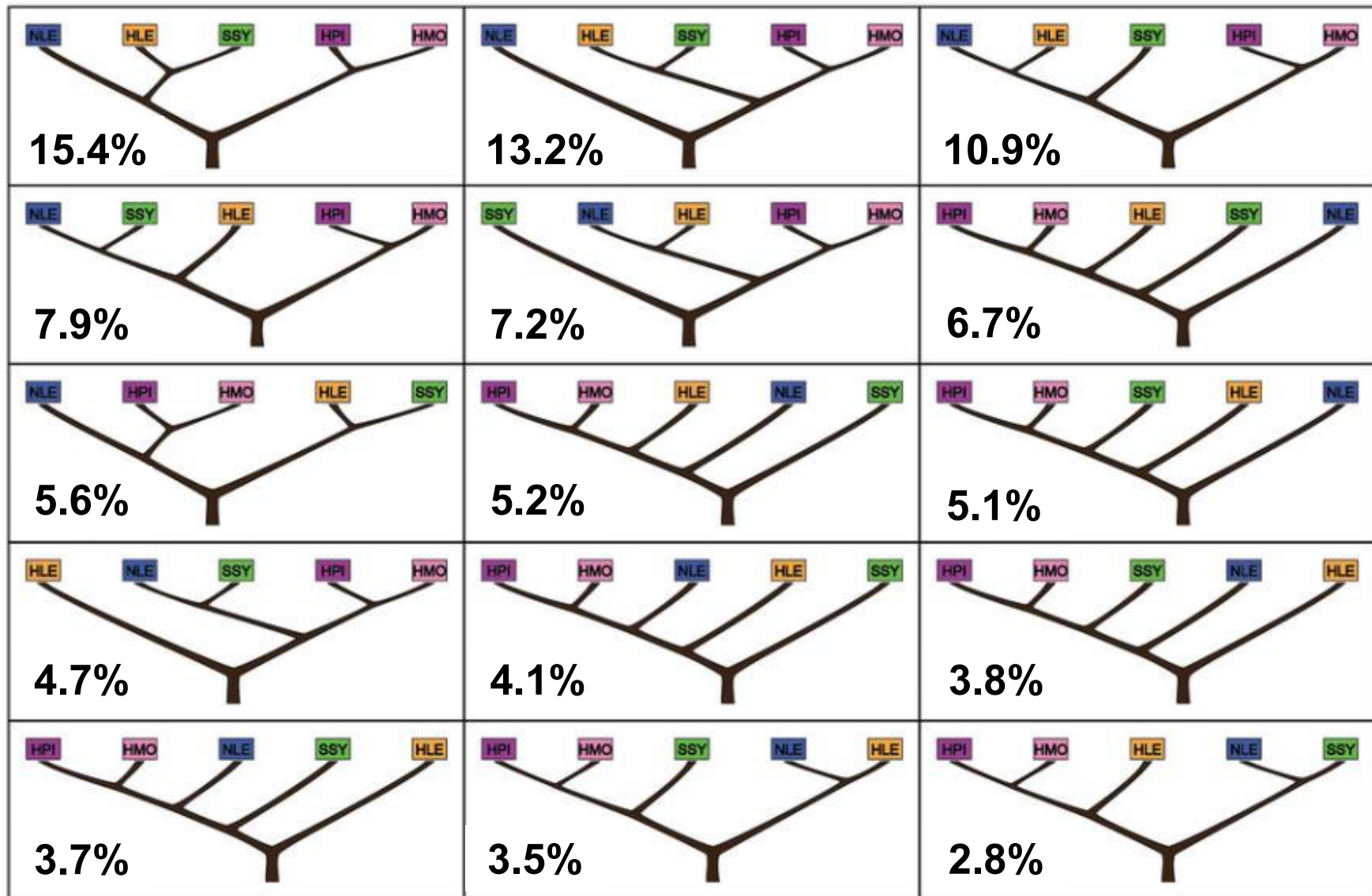


HMO



Carbone et al. (2014) Nature

Which is the Phylogeny of These 4 Genera?



Genomfart?

- ❖ **Parts of the tree of life are more likely to resemble a bush rather than a tree – do we expect that we can confidently infer every branch and twig?**
- ❖ **Bootstrap-based measures not useful in large data sets**
- ❖ **Methods evaluating conflict among data subsets (e.g., internode certainty among genes or sites) are preferable**
- ❖ **Explicitly identify internodes that, despite the use of genome-scale data sets, robust study designs and powerful algorithms, are poorly supported**

Coffee Break

Lecture Outline

❖ **Introduction to evolutionary genomics**

❖ **Phylogenomics**

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

❖ **Using genomes to understand ecology and evolution**

The Making of Biodiversity across the Yeast Subphylum



Hittinger lab



Kurtzman lab



Rokas lab

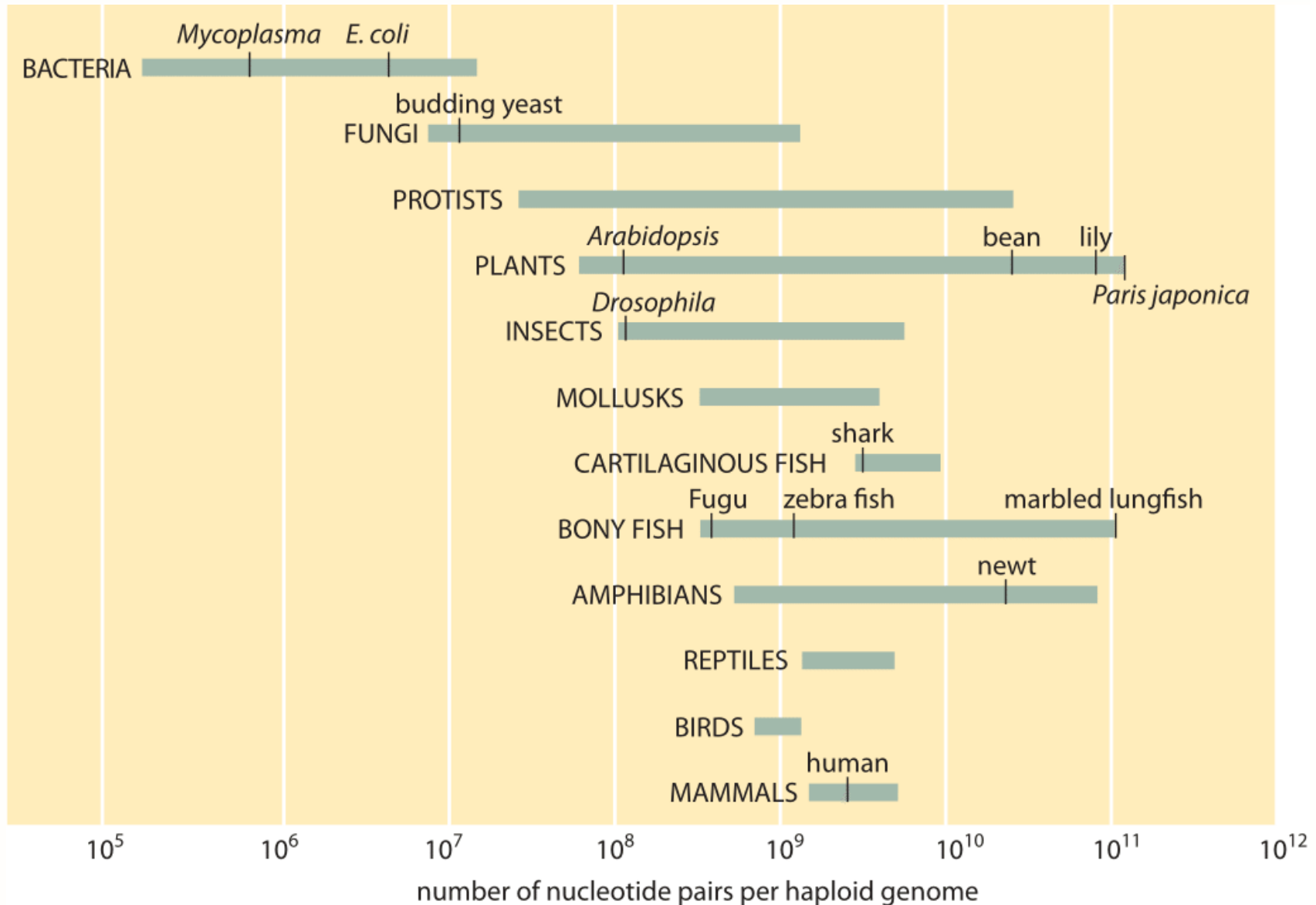
The Making of Biodiversity across the Yeast Subphylum

- ❖ **Sequence the genomes of all ~1,000+ known budding yeast species**
- ❖ **Construct their definitive phylogeny and timetree**
- ❖ **Examine the impact of metabolism on yeast diversification**
- ❖ **Revise their taxonomy**



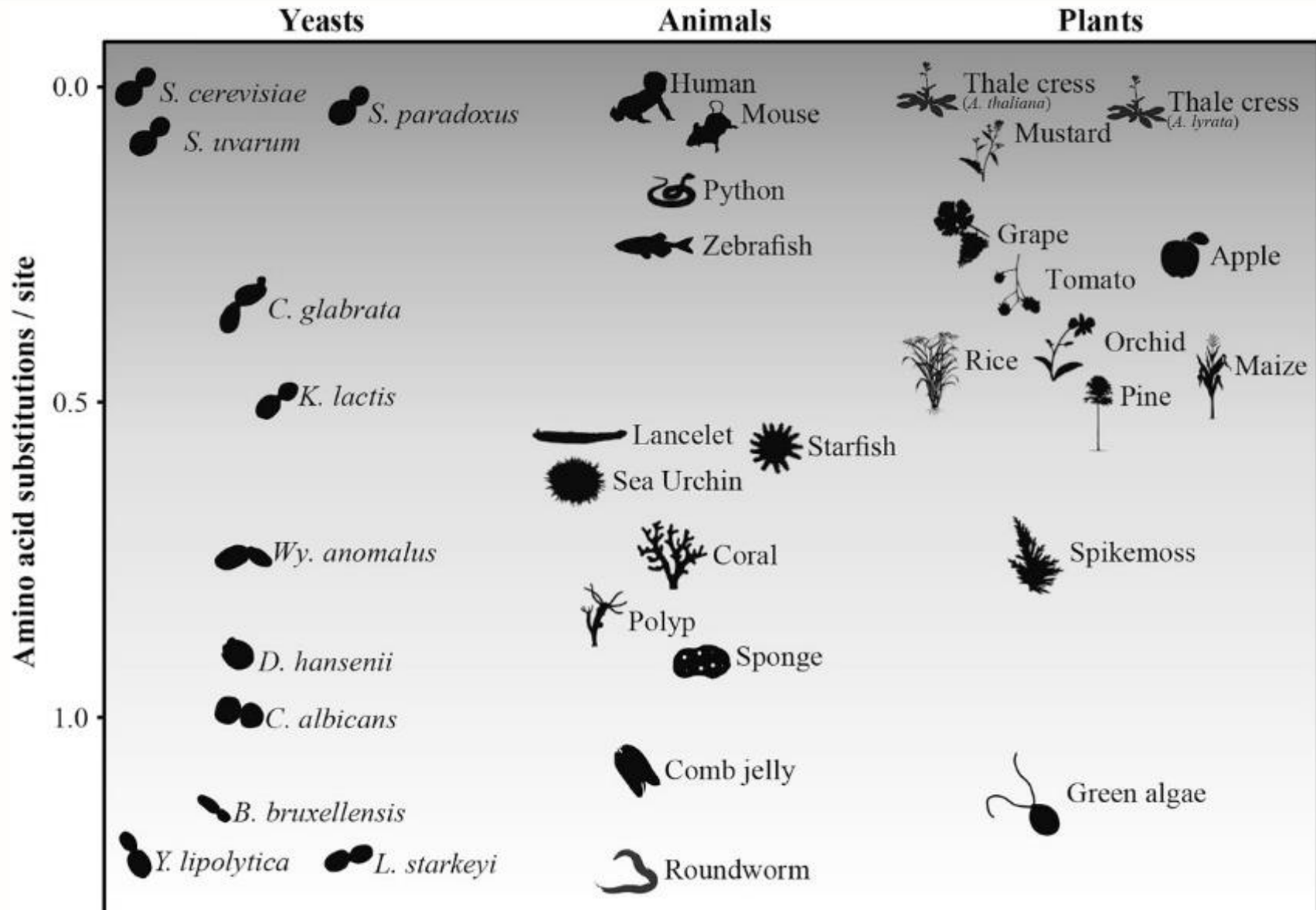
why budding yeasts?

Budding Yeasts Have Very Small Genomes



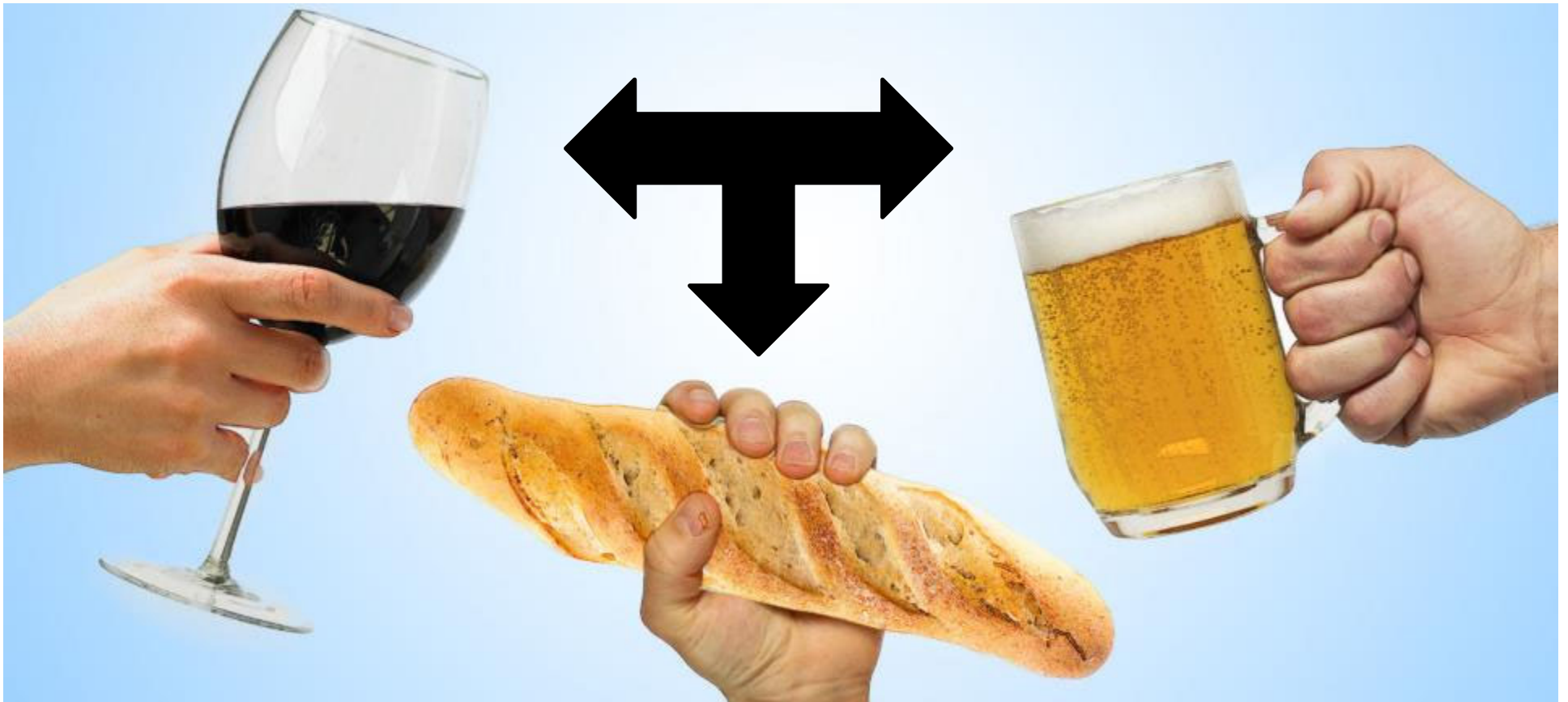
<http://book.bionumbers.org/how-big-are-genomes/>

Budding Yeasts Exhibit Striking Genomic Diversity



Shen, Opulente, Kominek, Zhou et al. (2018) Cell

S. cerevisiae, Cornerstone of Wine, Baking, and Brewing Industries



Several Other Genera Critical to the Food Industry

Kefir

Soy sauce

Sourdough

**Lambic
beers**

Kimchi

**Dietary
supplements**

Probiotics

Cheeses



Genera involved: *Saccharomyces*, *Kluyveromyces*, *Zygosaccharomyces*, *Candida*, *Kazachstania*, *Pichia*, and *Dekkera* (*Brettanomyces*)

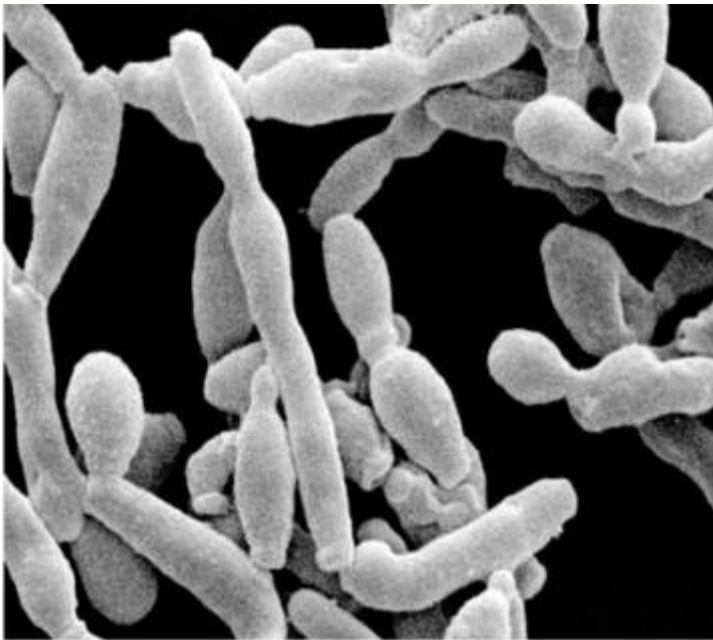


<http://y1000plus.org/>

The Metabolisms of Budding Yeasts Vary Widely

Xylose fermenters

(*Scheffersomyces (Pichia) stipitis*)

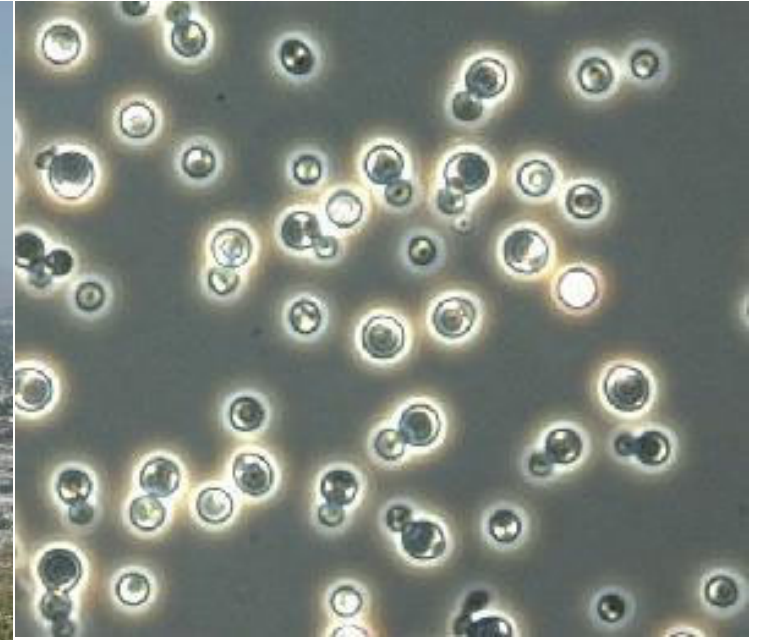


Cactophilic yeasts



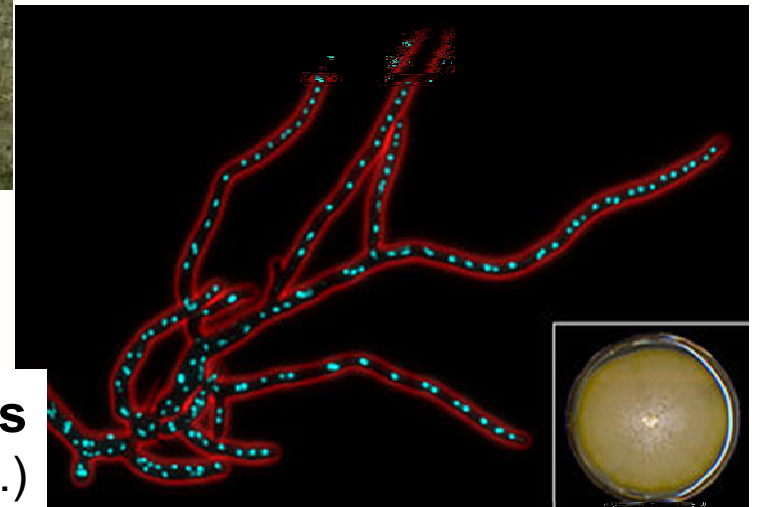
Oil producers

(*Lipomyces*, *Yarrowia*)



Animal pathogens
(*Candida*)

Plant pathogens
(*Eremothecium* sp.)



- i. Developing pipelines for genomic and phylogenomic data analyses**
- ii. Inferring a comprehensive & robust genome-scale phylogeny of budding yeasts
- iii. Using the phylogeny to gain insights into budding yeast evolution

Developing Pipelines for Genome Assembly

iWGS: *in silico* Whole Genome Sequencer & Analyzer

INPUT: Experimental Design + Reference Genome (Optional)



(Optional) Step 1: Simulation of Illumina / PacBio Data



Step 2: Quality Control (Quality / Adaptor Trimming; Error Correction)



Step 3: Assembly (Illumina-only (10), PacBio-only (3), Hybrid (4))



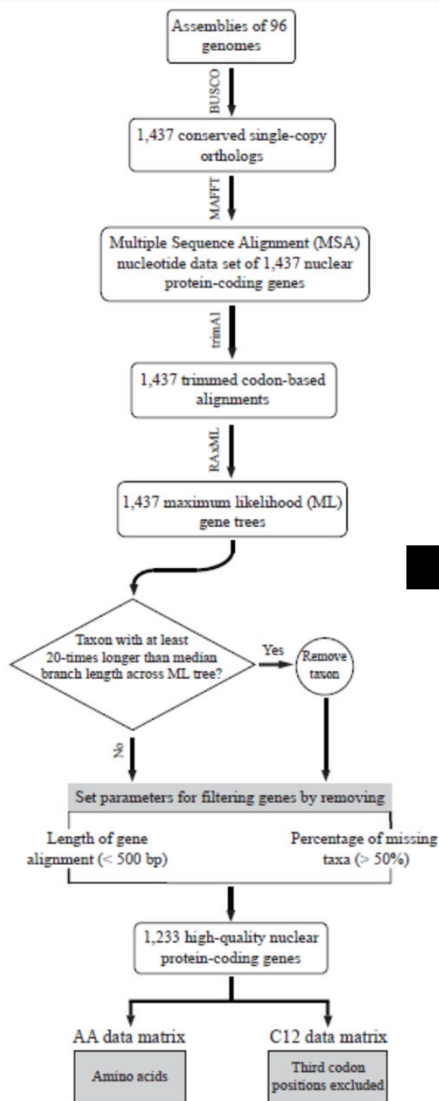
Step 4: QUAST Evaluation (Standalone or vs the Reference Genome)



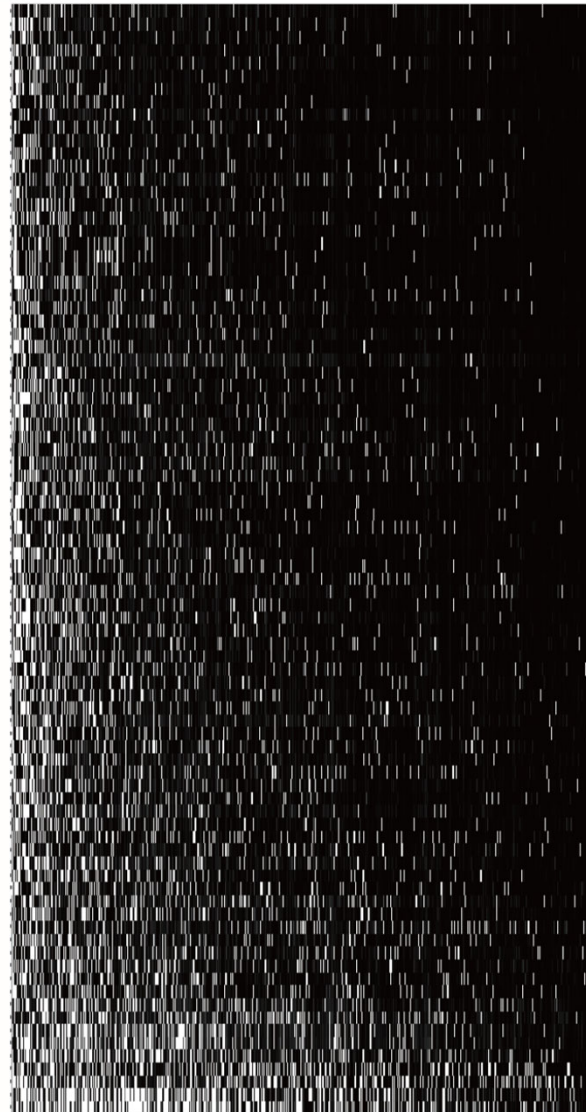
OUTPUT: Evaluation Report, Ranking of Experimental Designs Tested



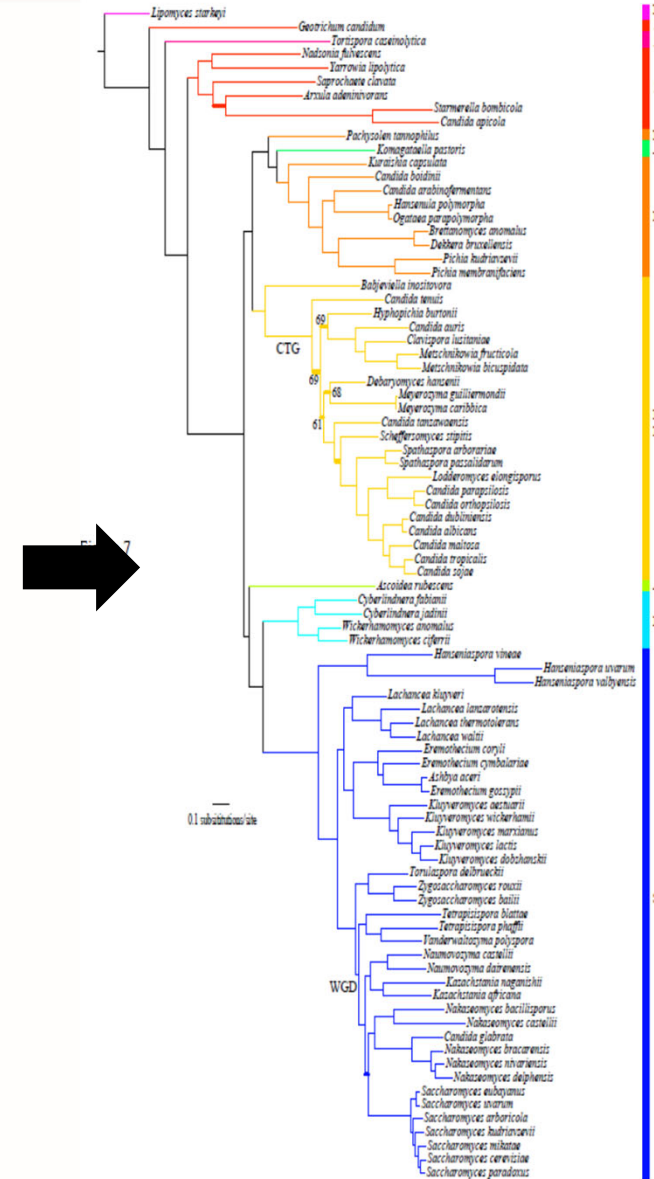
Developing Pipelines for Phylogenomic Inference



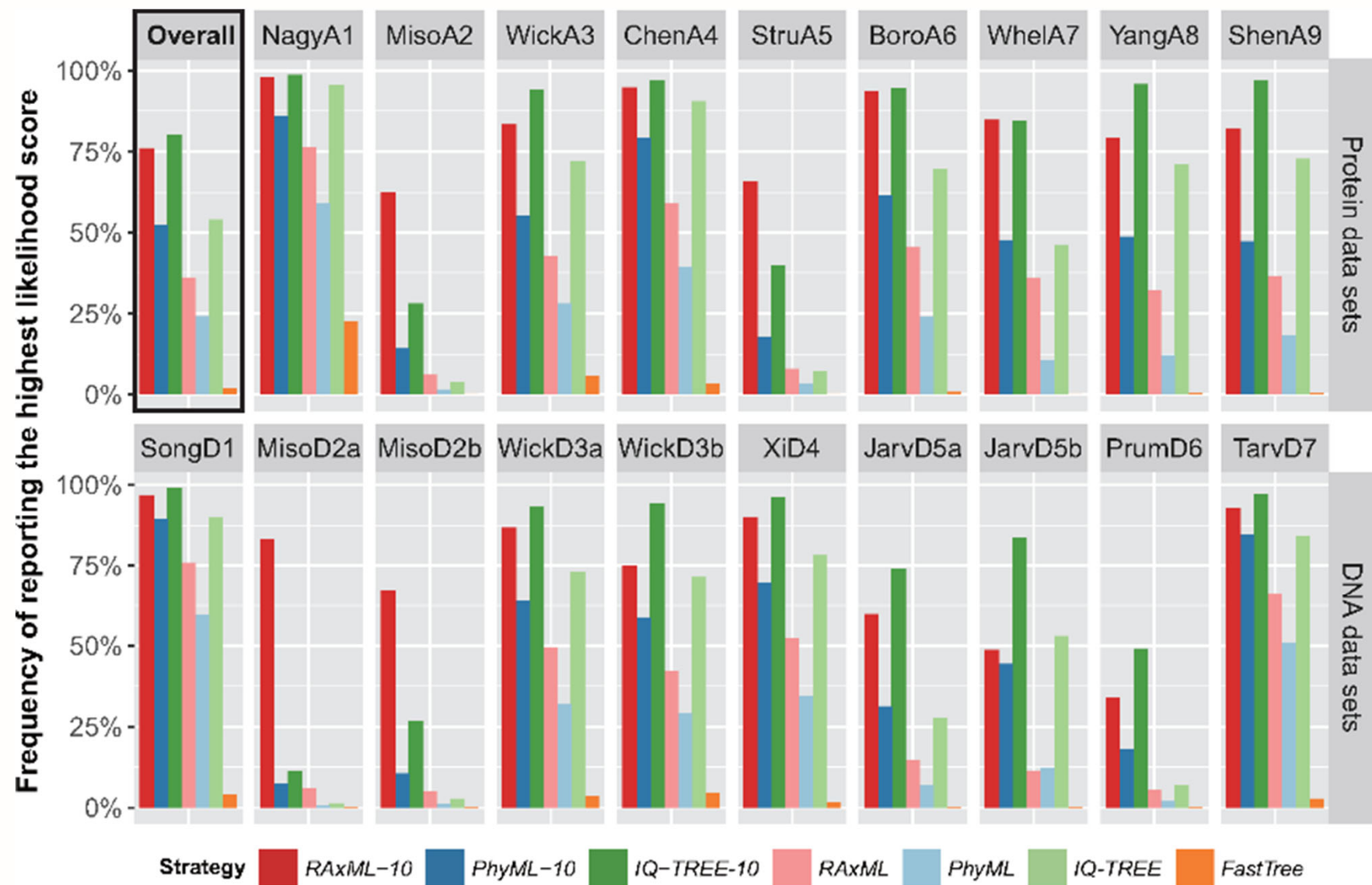
Taxa



Genes



Assessing Speed and Accuracy of Phylogenomic Software



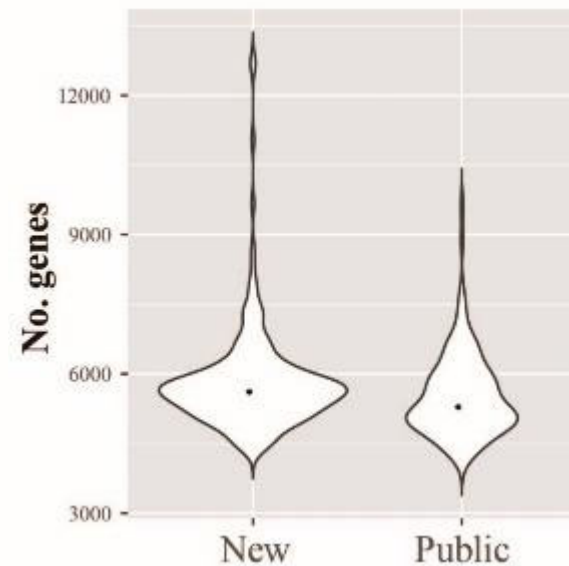
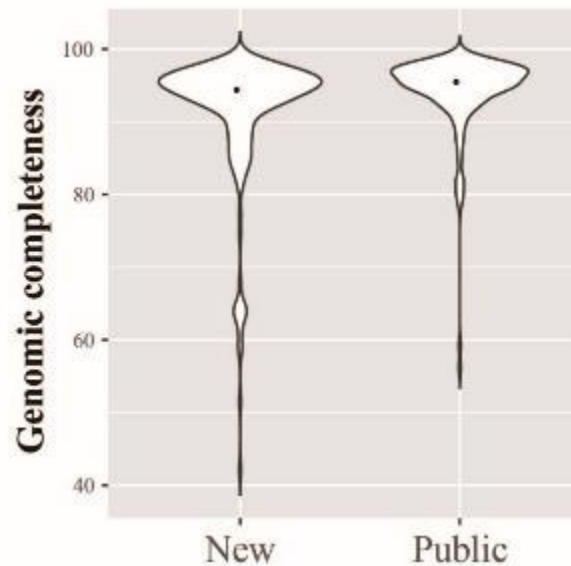
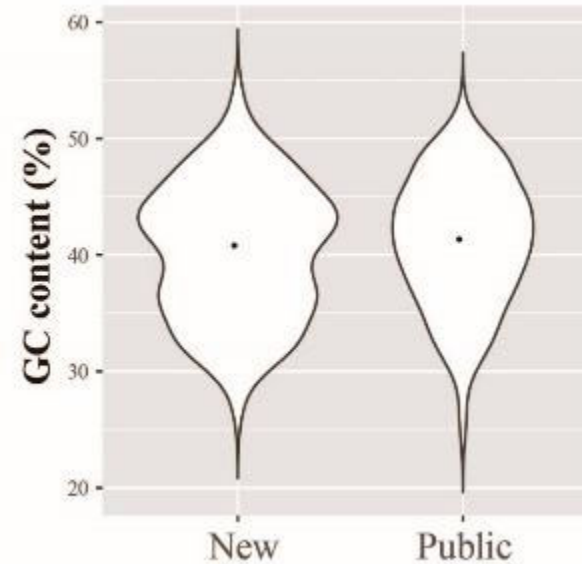
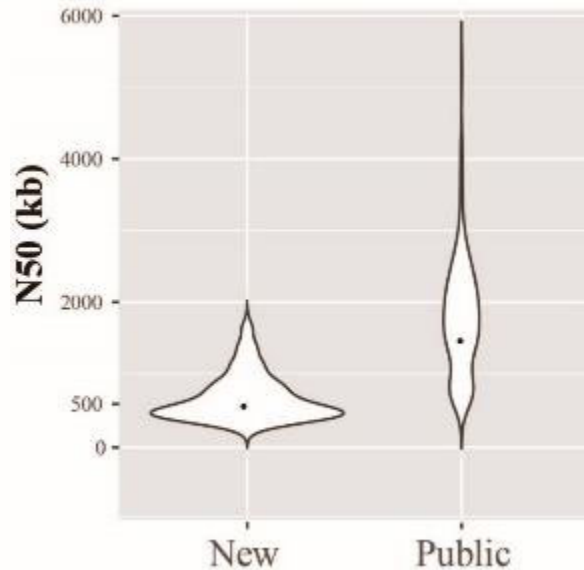
Zhou et al. (2018) Mol. Biol. Evol.

- i. Developing pipelines for genomic and phylogenomic data analyses
- ii. Inferring a comprehensive & robust genome-scale phylogeny of budding yeasts**
- iii. Using the phylogeny to gain insights into budding yeast evolution

- ❖ Sequenced the genomes of 220 species (196 Y1000+ species + 24 RIKEN genomes); most of them are from type strains
- ❖ + 112 publicly available genomes -> 332 genomes
- ❖ Sampled taxa from 79 / 92 genera (~85%)

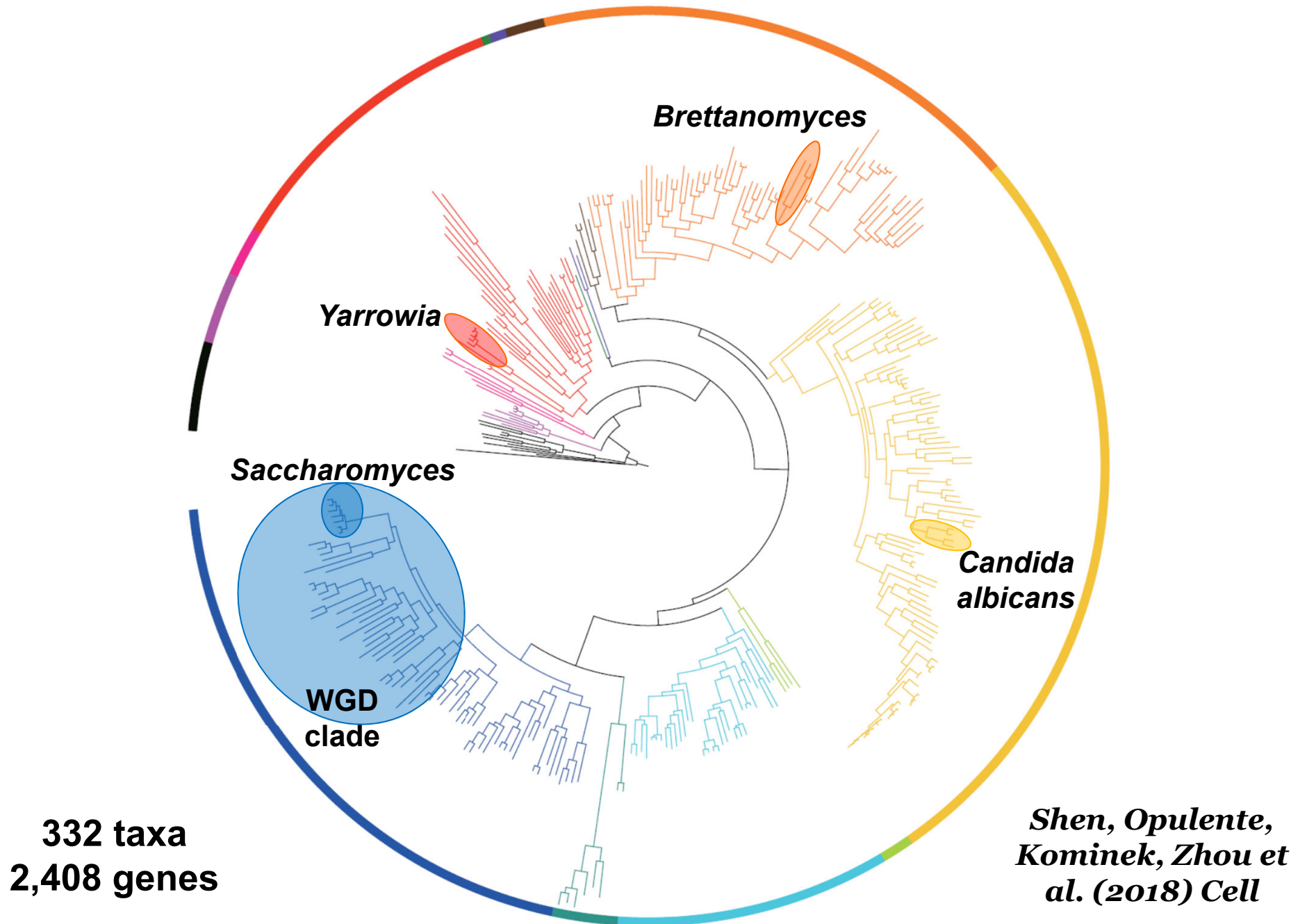
Shen, Opulente, Kominek, Zhou et al. (2018) Cell

The Quality of the New Genomes is High

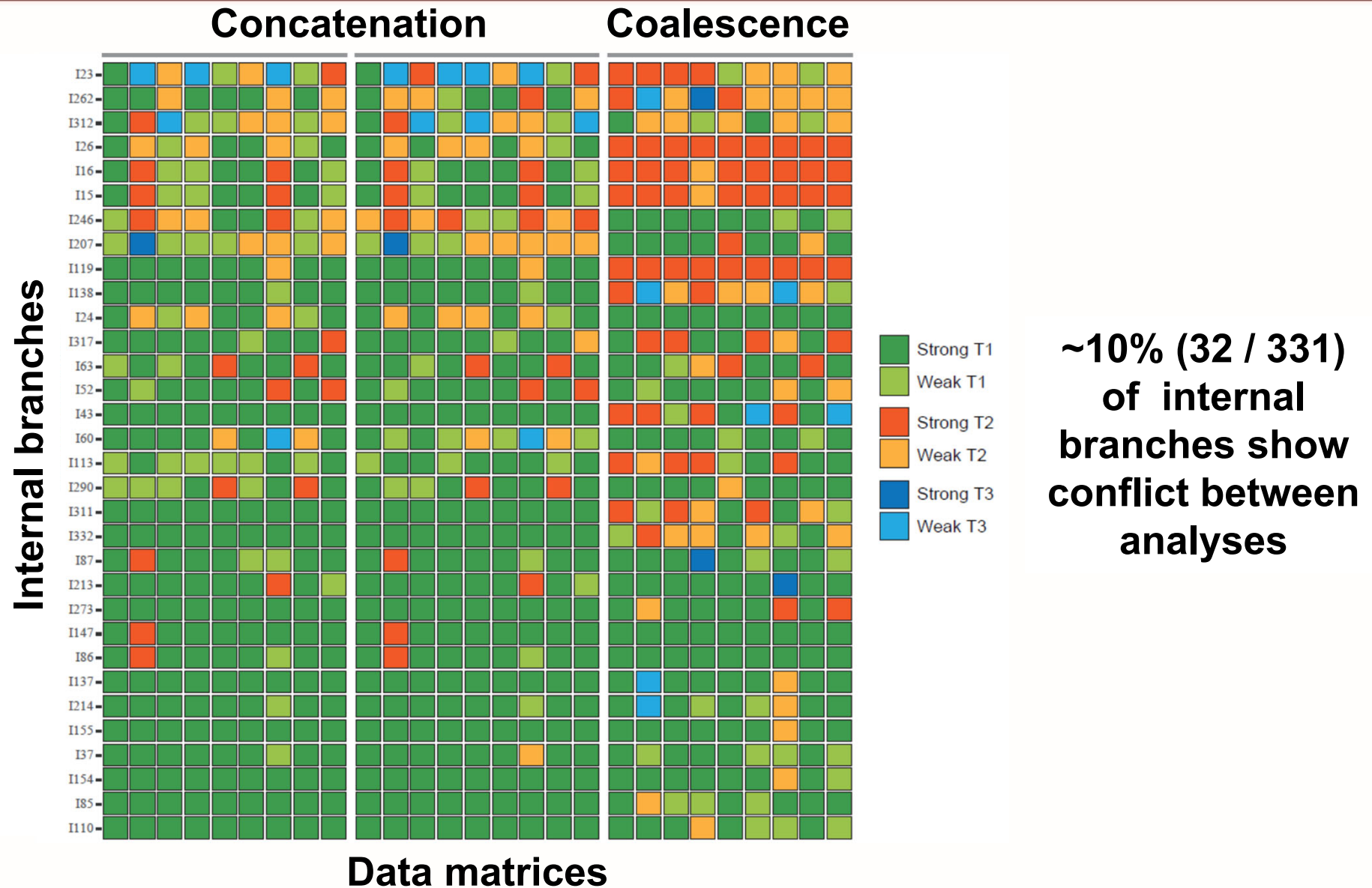


Shen, Opulente, Kominek, Zhou et al. (2018) Cell

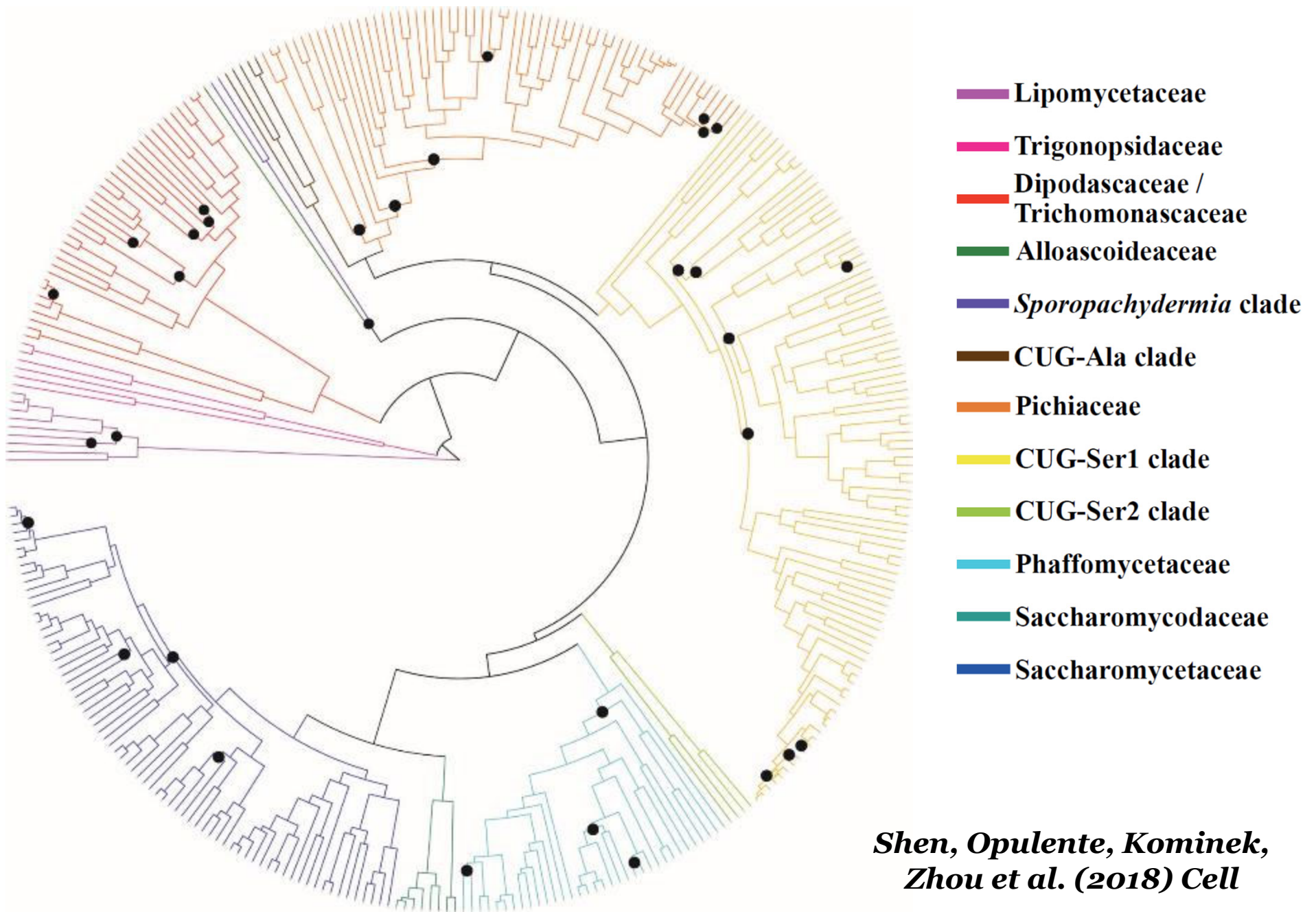
Genome-Scale Phylogeny of Budding Yeasts



The 32 Conflicting Branches in the Yeast Phylogeny



Distribution of Conflict on the Yeast Phylogeny



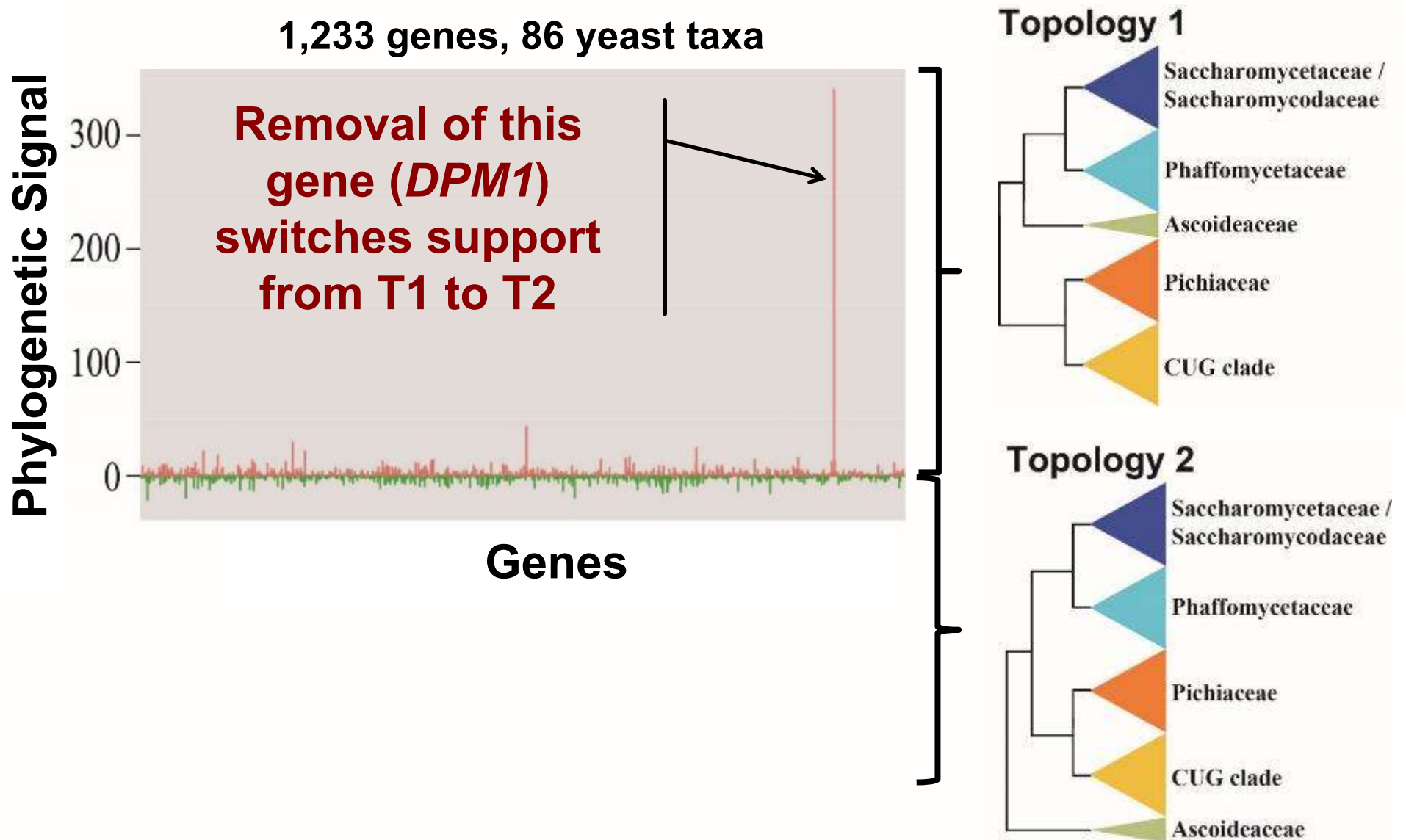
***Shen, Opulente, Kominek,
Zhou et al. (2018) Cell***



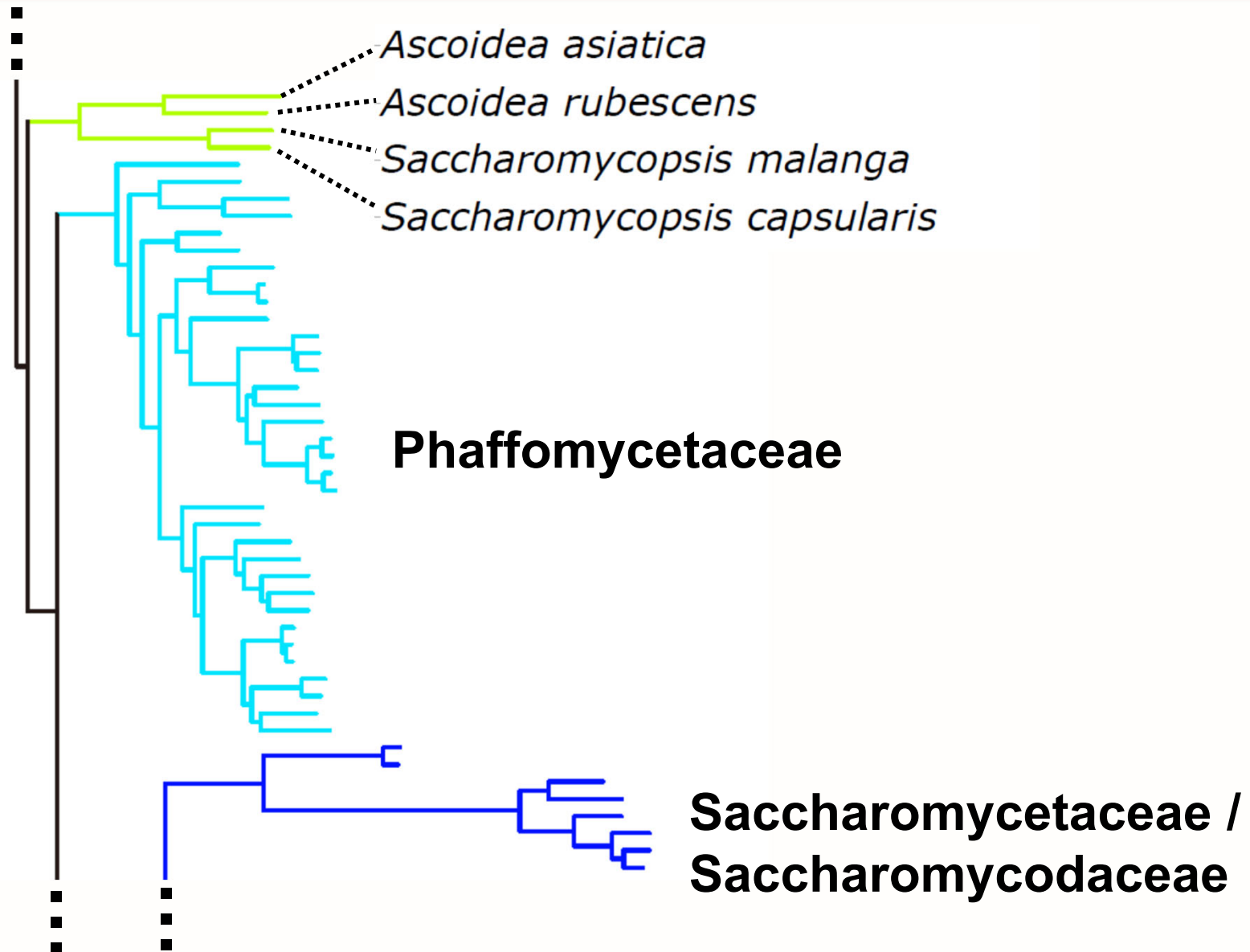
**Despite increasing
internal branches
~4X, (85 -> 331),
conflict decreased**

Shen et al. (2016) G3

A Single Gene Governs the Placement of Ascoideaceae

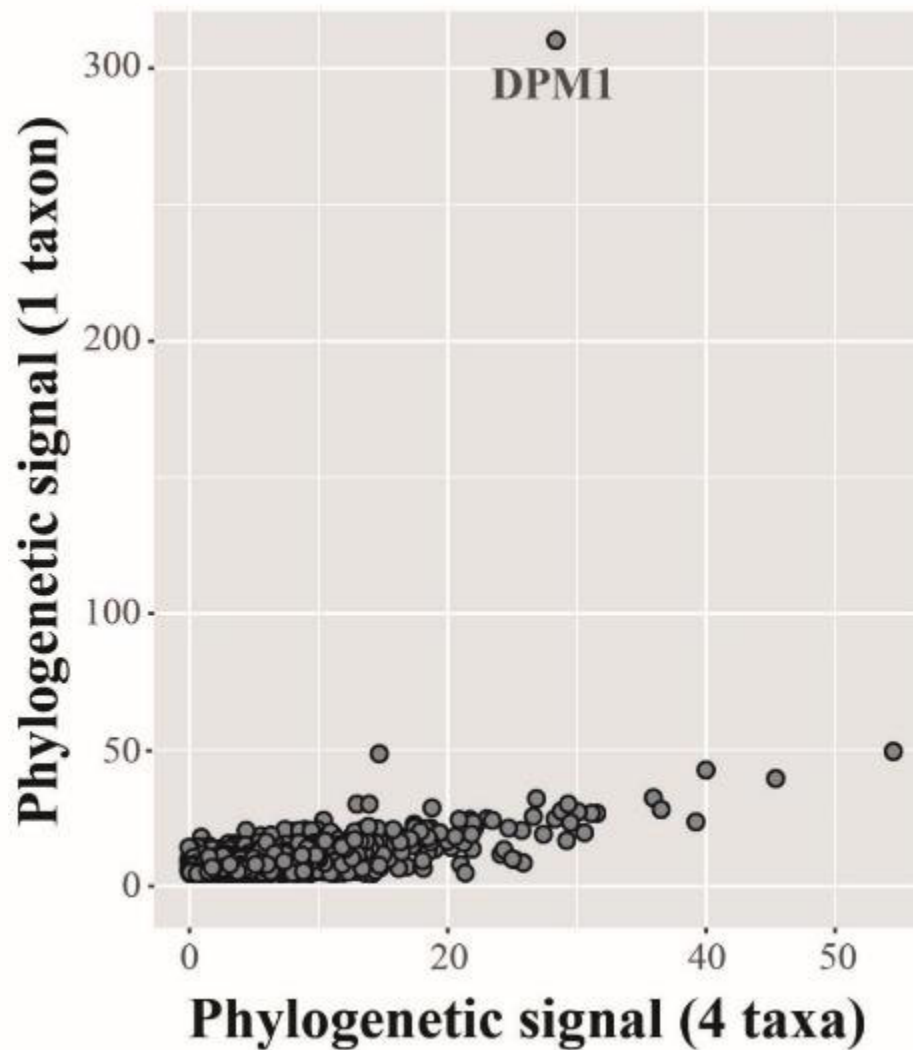


Sampling of 3 Additional Taxa “Breaks” the Long Branch



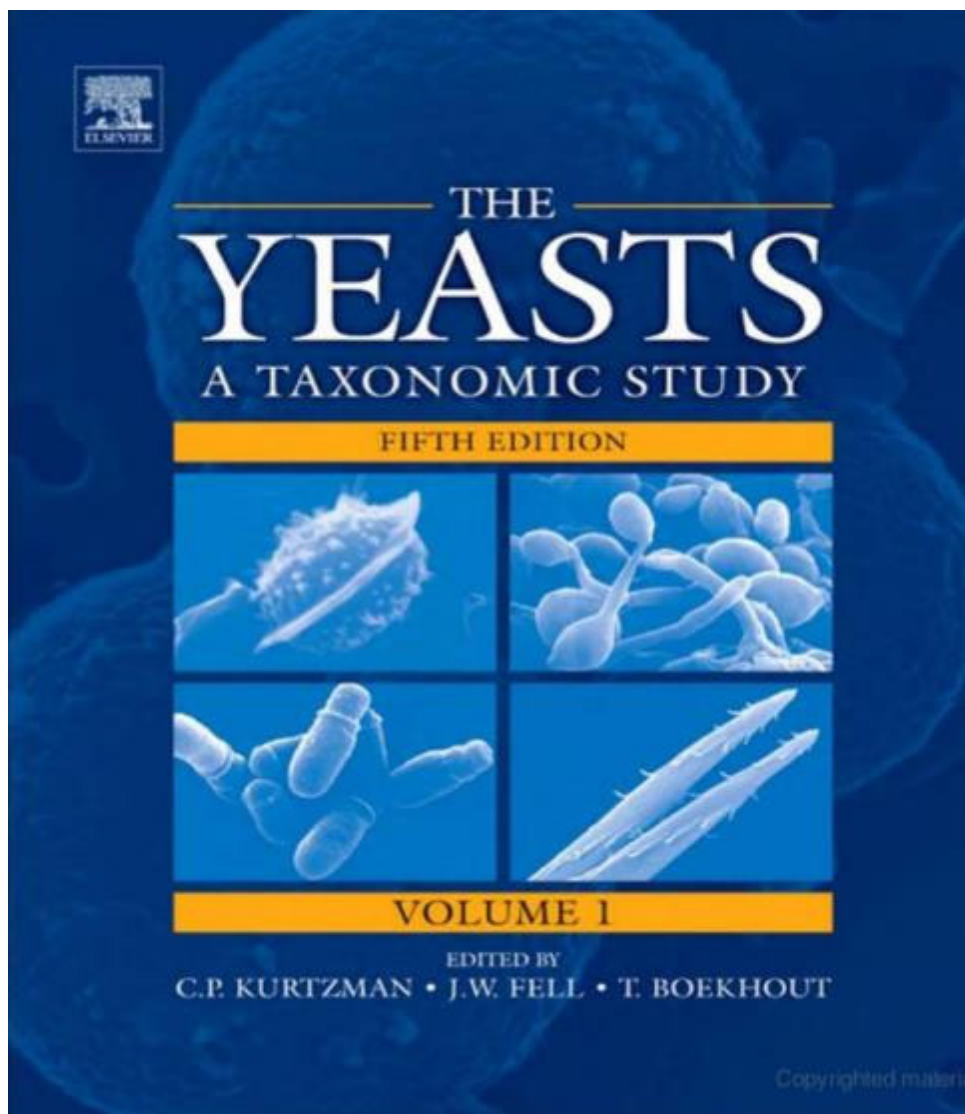
Sampling of 3 Additional Taxa Decreases Gene's Signal

2,408 genes, 329 – 332 yeast taxa



Shen, Opulente, Kominek, Zhou et al. (2018) Cell

- i. Developing pipelines for genomic and phylogenomic data analyses
- ii. Inferring a comprehensive & robust genome-scale phylogeny of budding yeasts
- iii. Using the phylogeny to gain insights into budding yeast evolution



Fermentation

Glucose	+	Lactose	—
Galactose	—	Raffinose	—
Sucrose	—	Trehalose	—
Maltose	—		

Growth (in Liquid Media)

Glucose	+	D-Ribose	—
Inulin	—	Methanol	—
Sucrose	—	Ethanol	—
Raffinose	—	Glycerol	—
Melibiose	—	Erythritol	—
Galactose	—	Ribitol	—
Lactose	—	Galactitol	—
Trehalose	—	D-Mannitol	—
Maltose	—	D-Glucitol	—
Melezitose	—	myo-Inositol	—
Methyl- α -D-glucoside	—	DL-Lactate	—
Soluble starch	—	Succinate	—
Cellobiose	+	Citrate	—
Salicin	+	D-Gluconate	+
L-Sorbose	—	D-Glucosamine	—
L-Rhamnose	—	N-Acetyl-D-glucosamine	n
D-Xylose	—	Hexadecane	n
L-Arabinose	—	Nitrate	—
D-Arabinose	—	Vitamin-free	—

Additional Growth Tests and Other Characteristics

2-Keto-D-gluconate	+	Growth at 25°C	+
Cycloheximide 0.01%	+	Growth at 30°C	—
Starch formation	—		

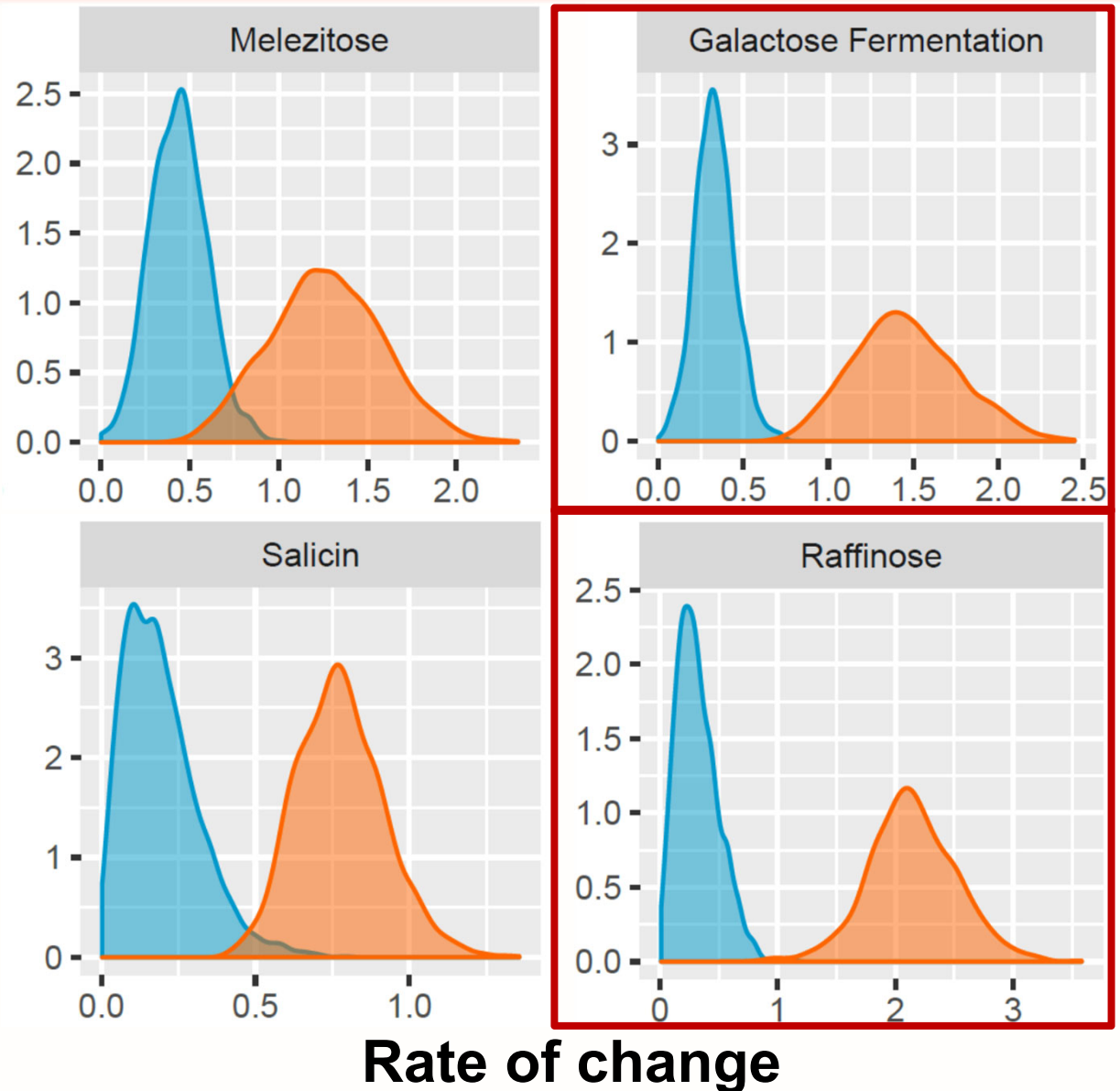
Loss Exceeds Gain for 38 / 45 Metabolic Traits

**Rate of trait
gain**

**Rate of trait
loss**

**For 17 / 38,
the difference
is statistically
significant**

Probability density

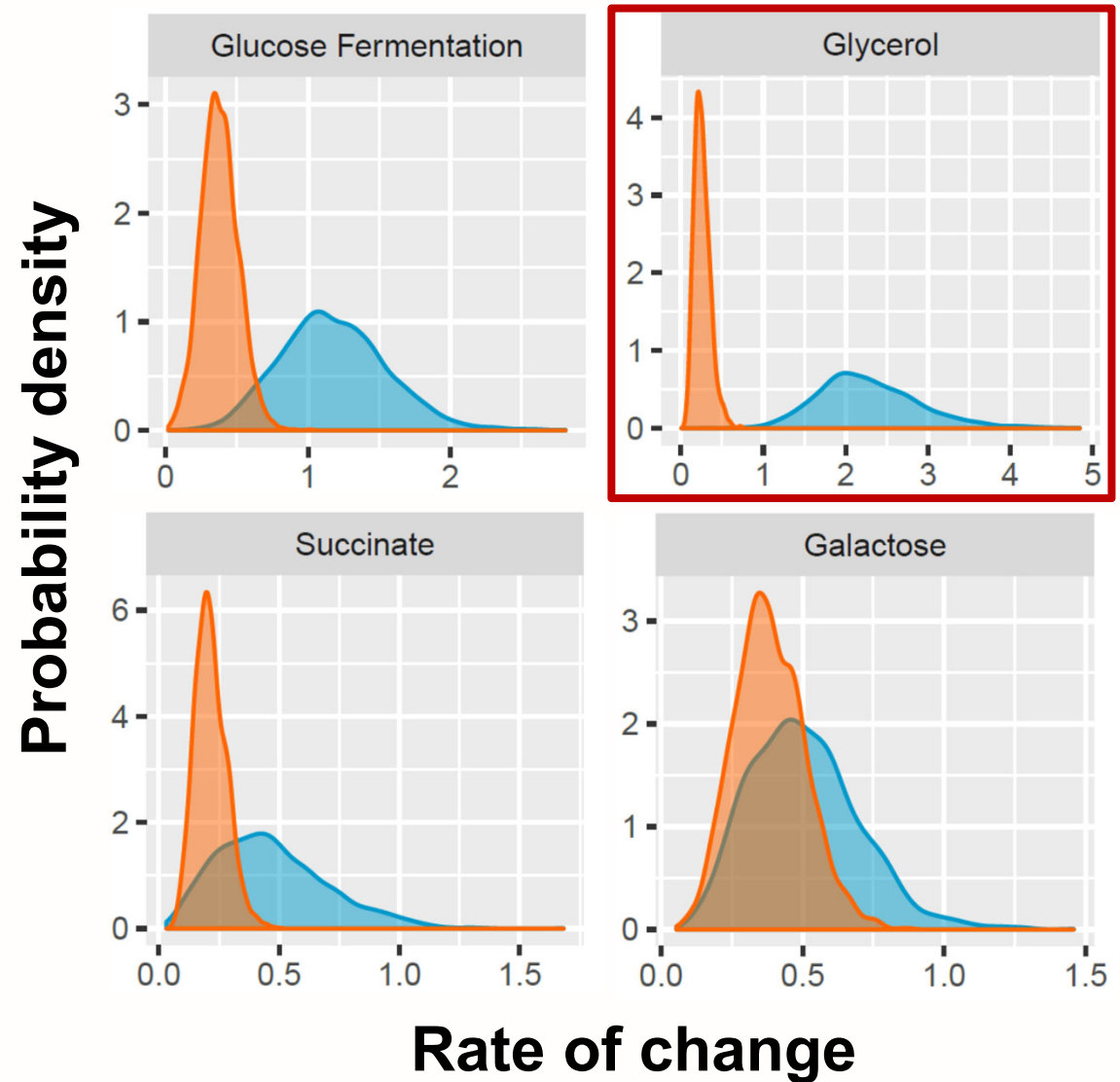


Gain Exceeds Loss for the Remaining 7 Traits

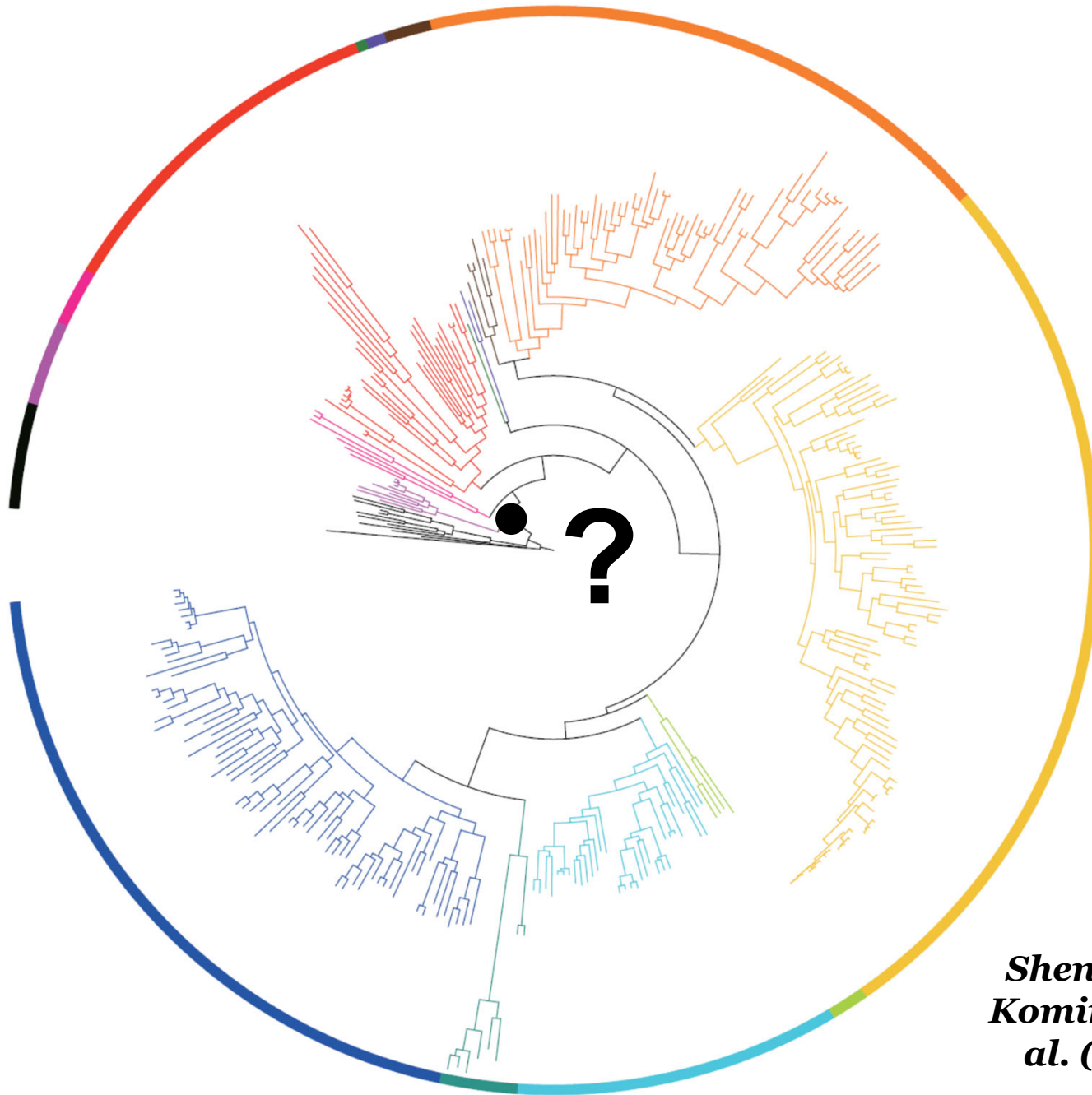
**Rate of trait
gain**

**Rate of trait
loss**

**For 1 / 7, the
difference is
statistically
significant**



Inferring the Metabolic Capabilities of Yeast Ancestors



***Shen, Opulente,
Kominek, Zhou et
al. (2018) Cell***

BYCA (Budding Yeast Common Ancestor) was a Generalist

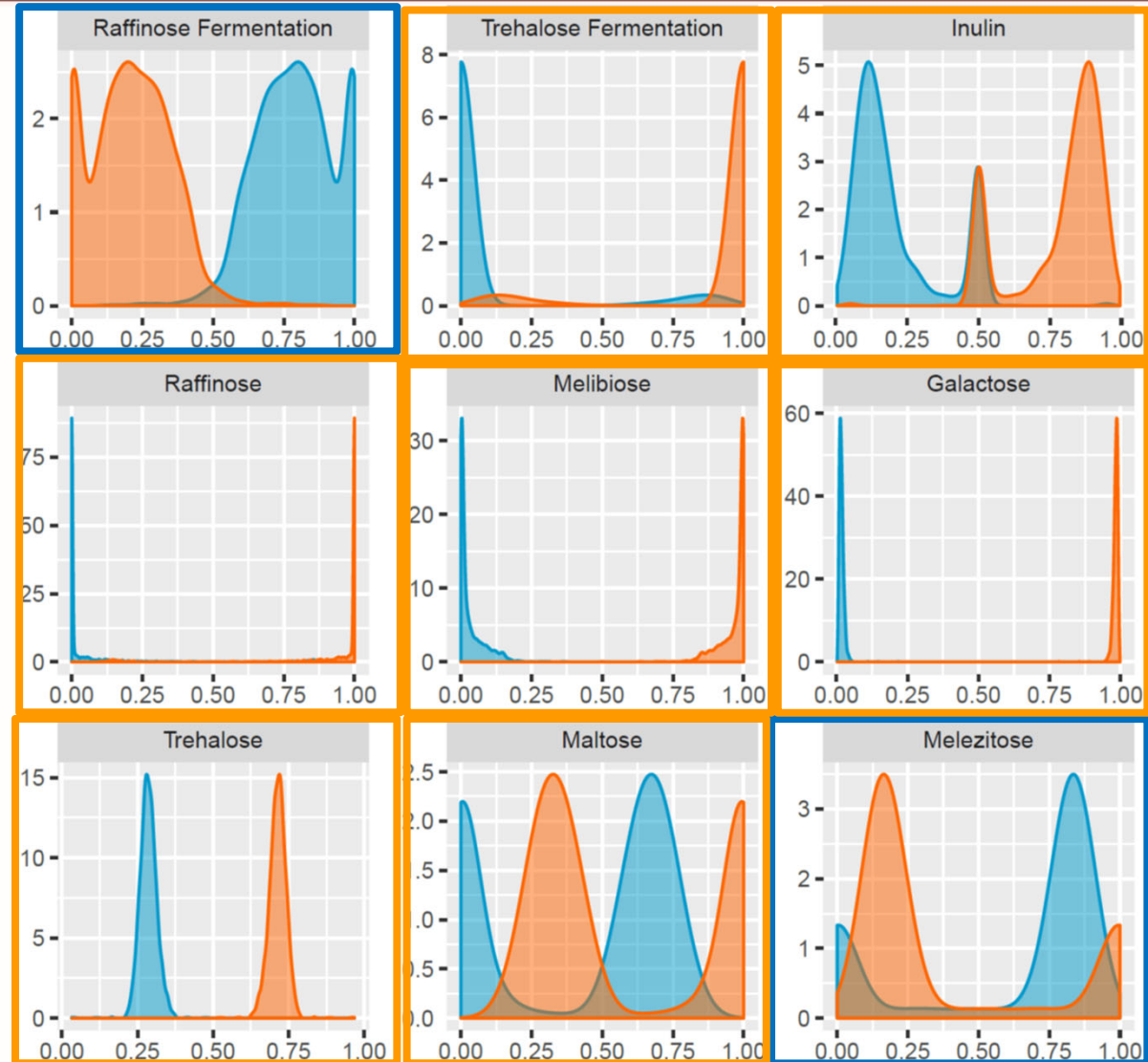
28 / 45 traits
likely present in
BYCA

Prob. of trait
presence

Prob. of trait
absence

17 / 45 traits
likely absent in
BYCA

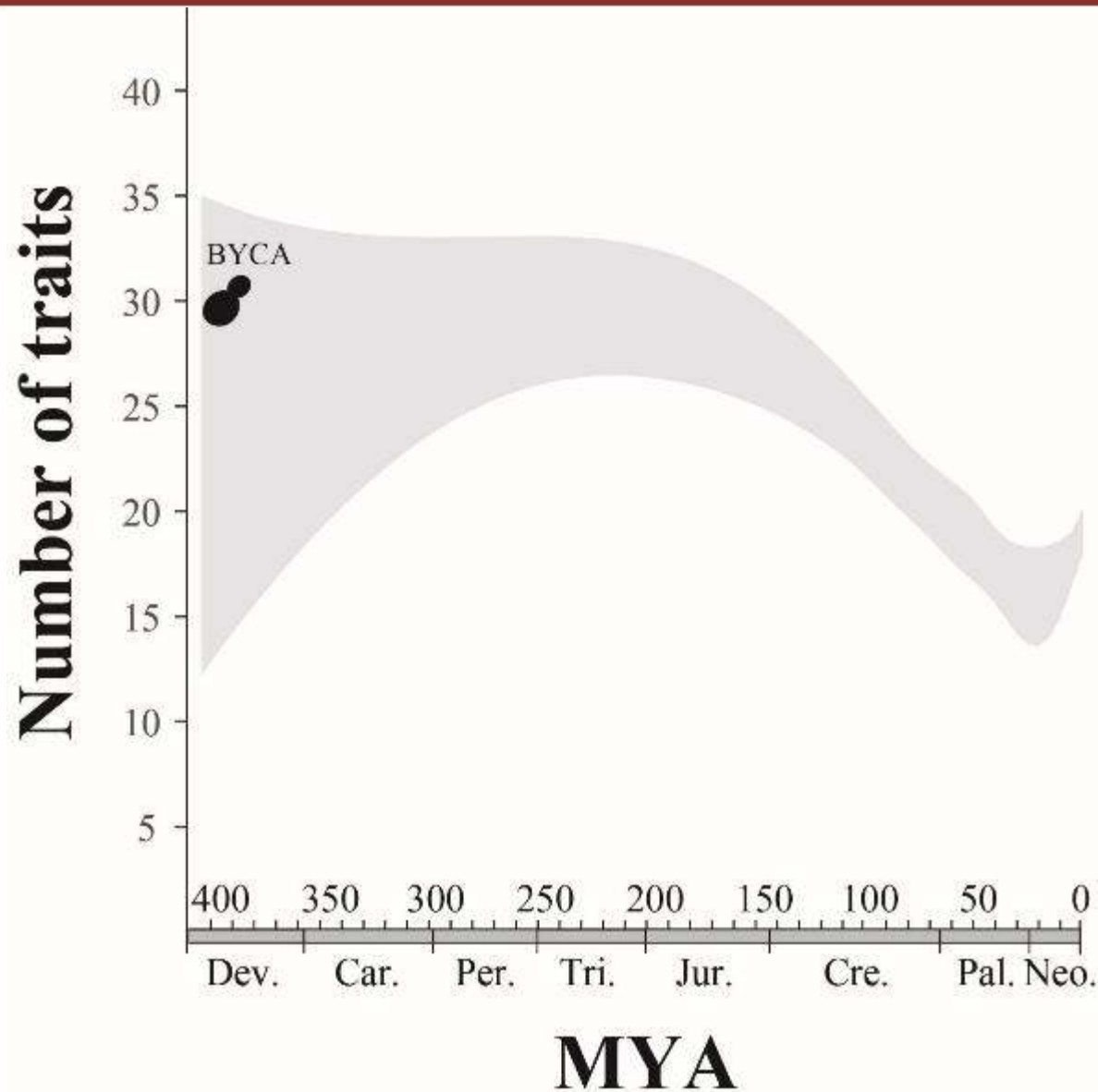
Probability density



Posterior Probability

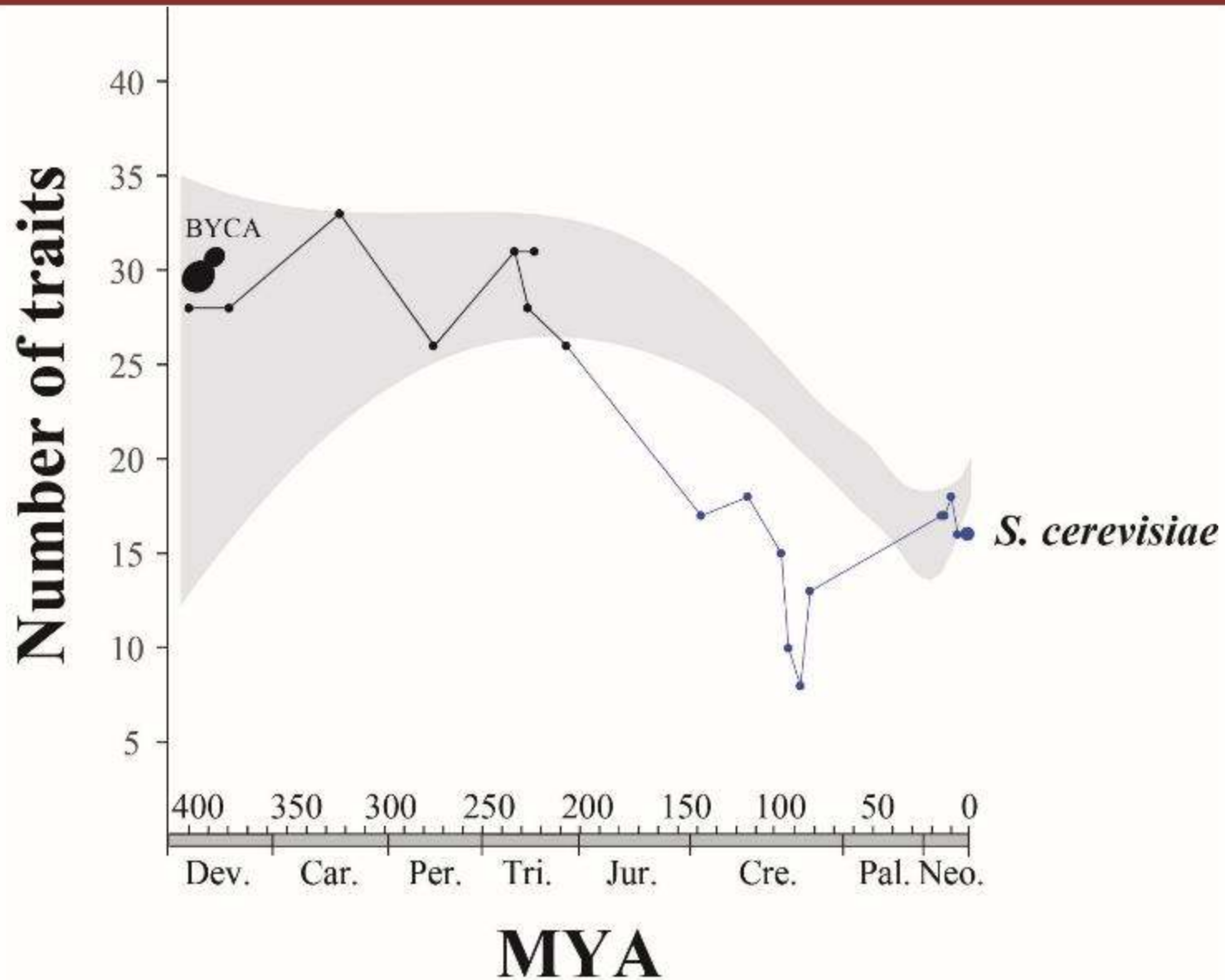


Widespread Loss of Traits



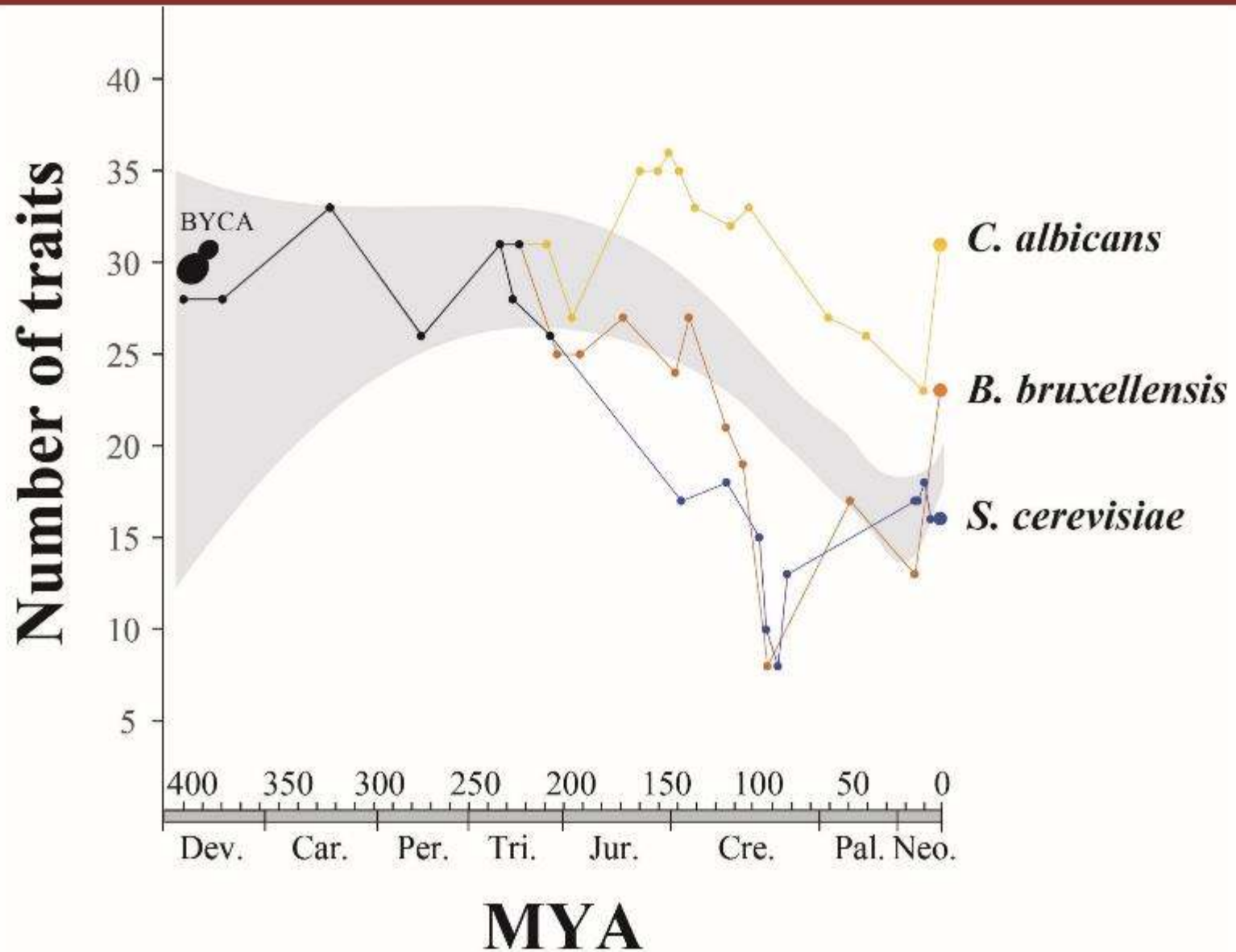
Shen, Opulente, Kominek, Zhou et al. (2018) Cell

Widespread Loss of Traits

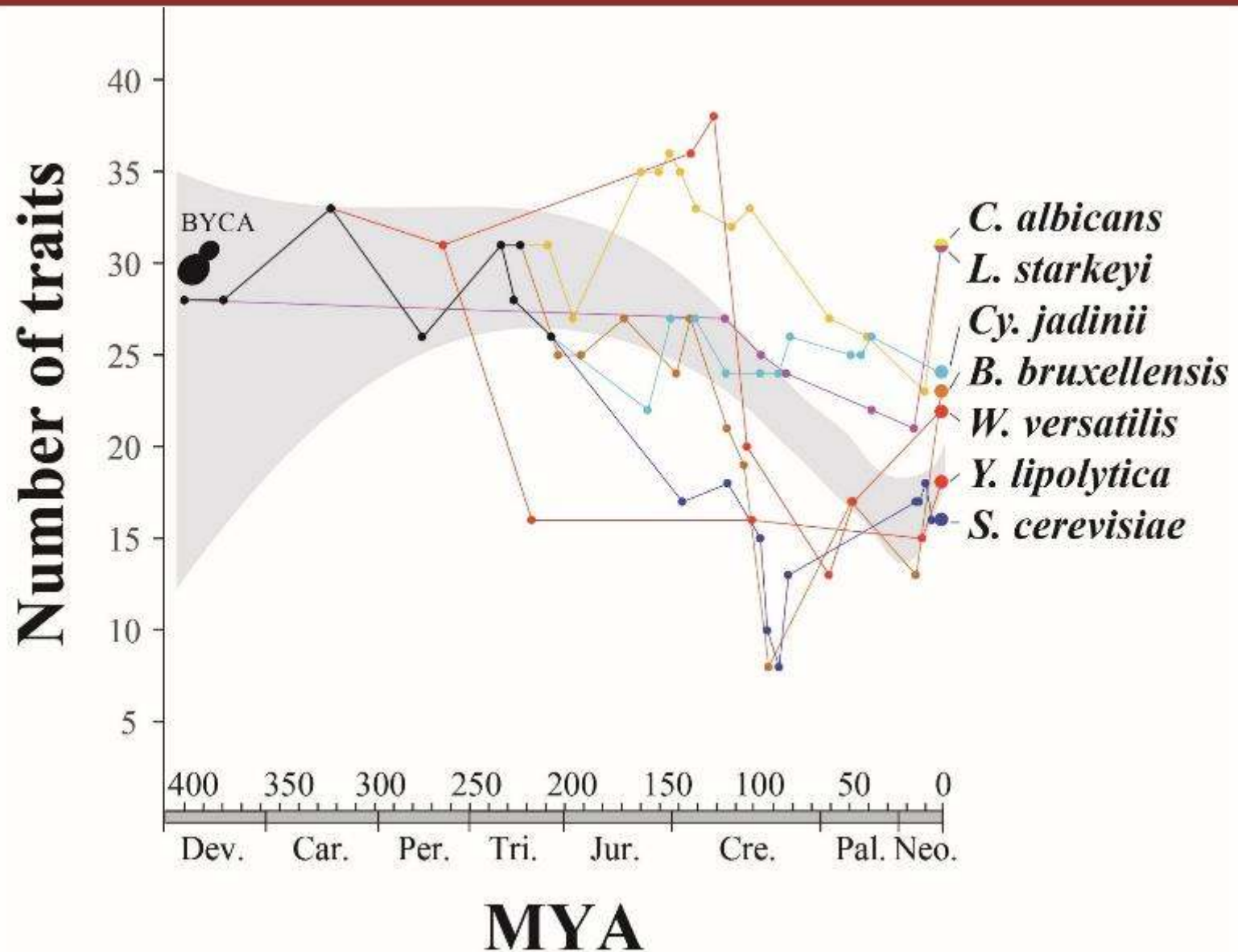


Shen, Opulente, Kominek, Zhou et al. (2018) Cell

Widespread Loss of Traits



Widespread Loss of Traits

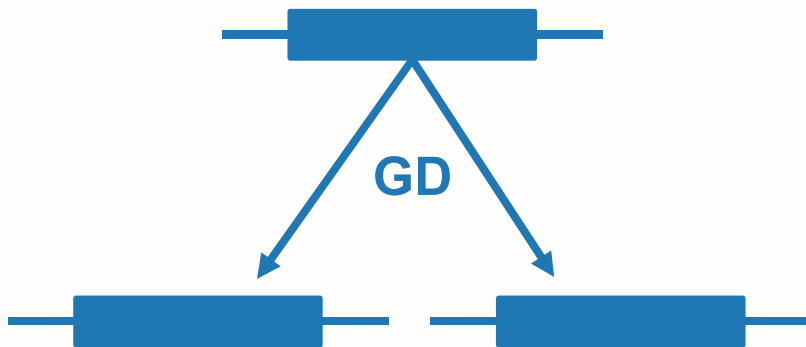


**with so much loss, how
did new metabolic
traits evolve in the
lineage?**

Two Major Sources of Gene Innovation

Gene duplication (GD)

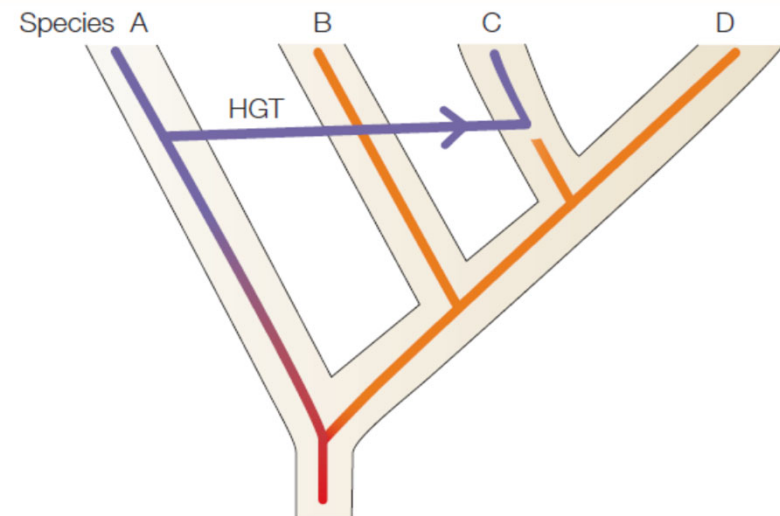
Any duplication of a region of DNA that contains a gene



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

Horizontal gene transfer (HGT)

Exchange of genes between organisms other than through reproduction

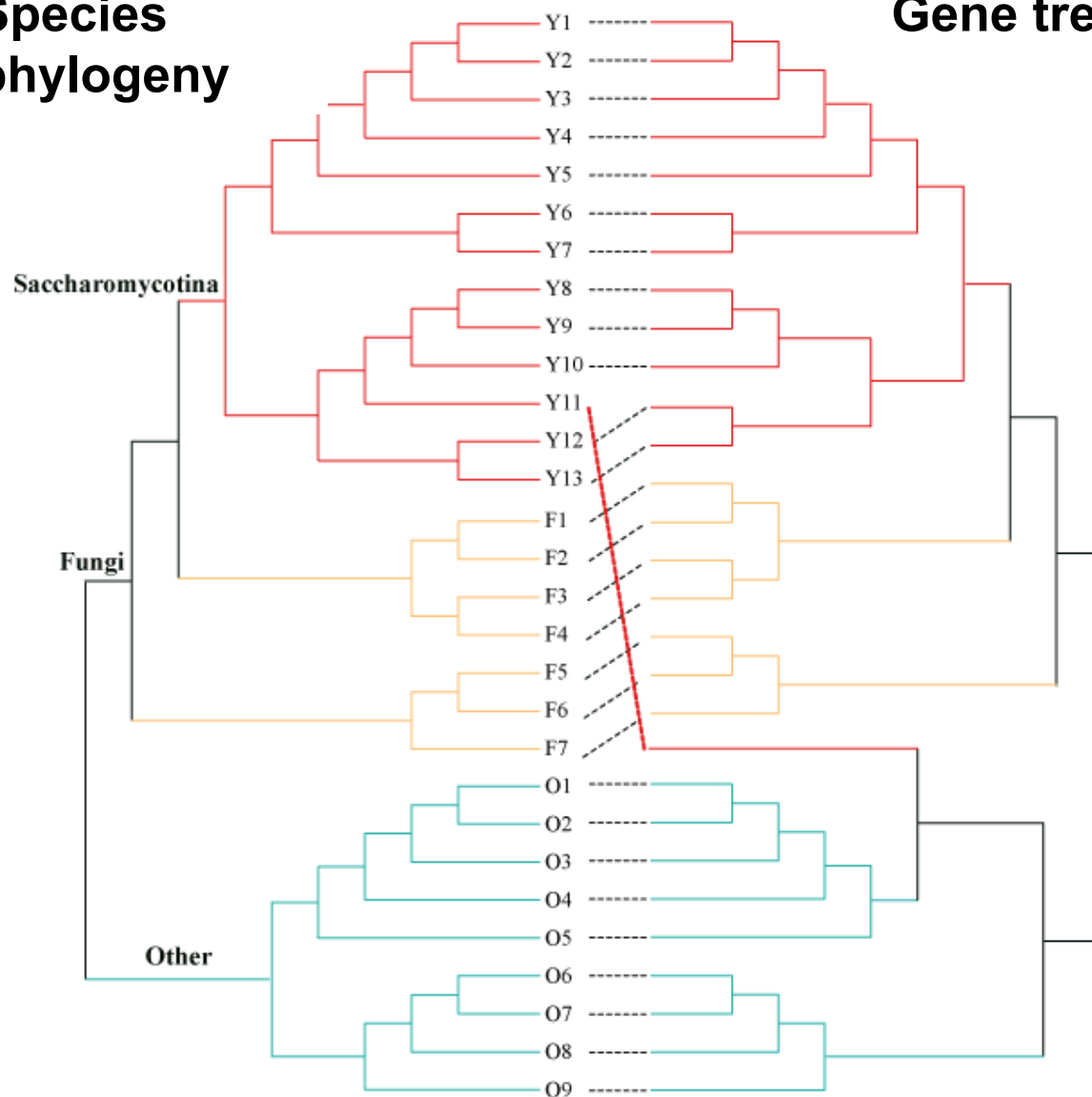


- ❖ Xenobiotic catabolism
- ❖ Toxin production
- ❖ Degradation of plant cell walls
- ❖ Wine fermentation

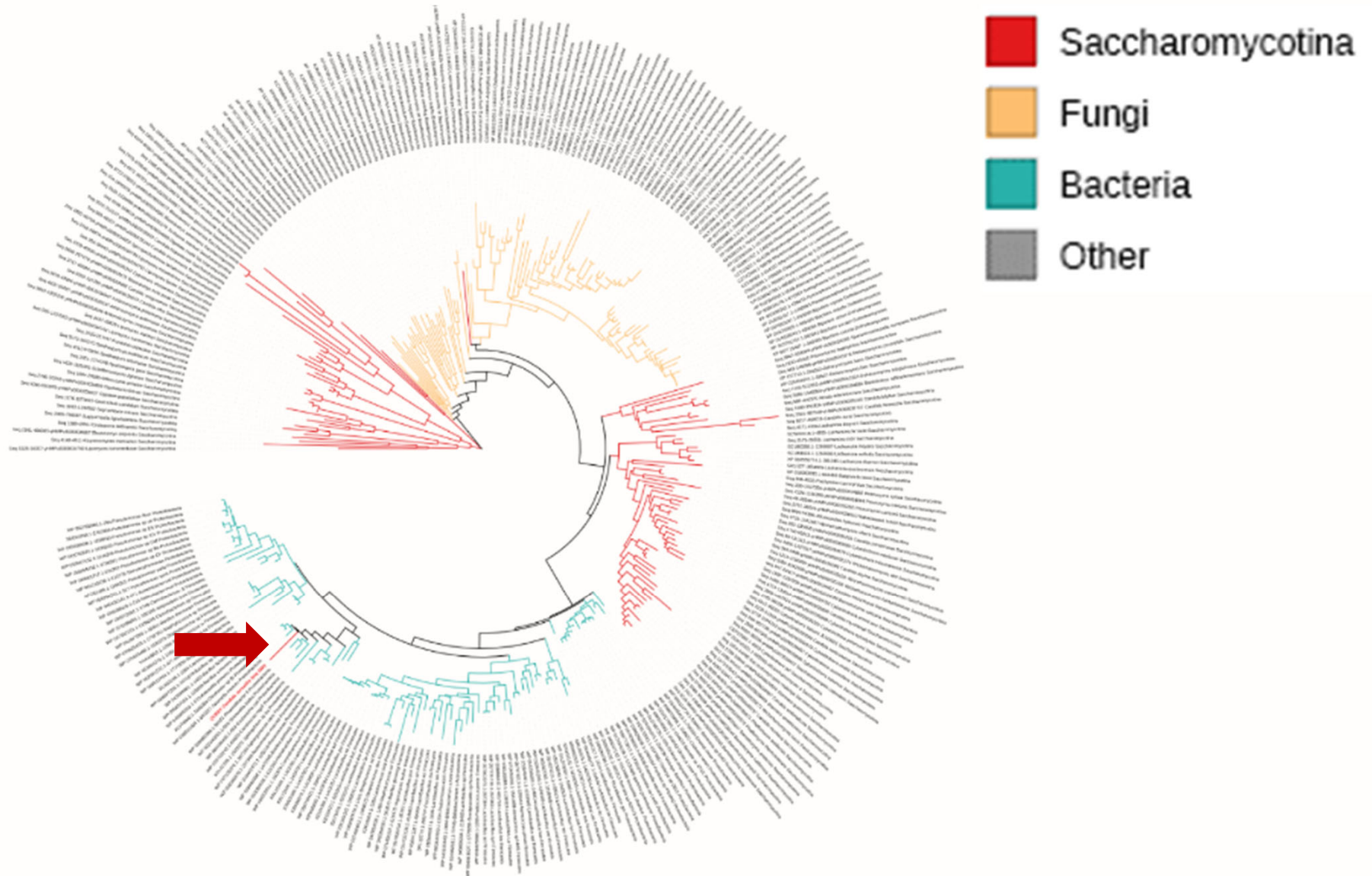
How Do We Search for Horizontal Gene Transfer?

**Species
phylogeny**

Gene tree

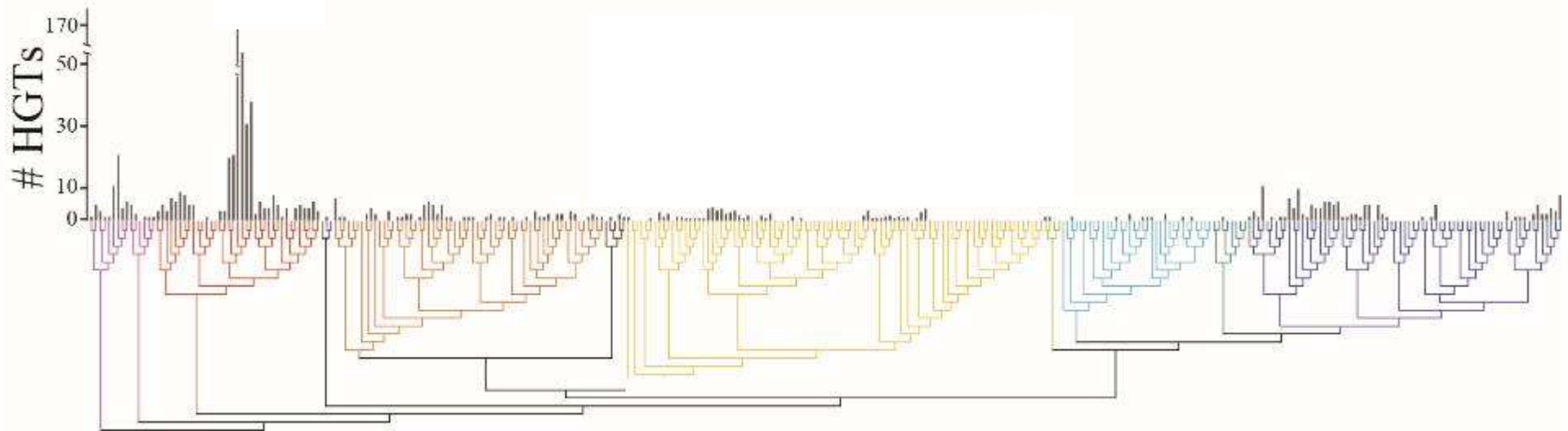


Example Horizontal Gene Transfer



Shen, Opulente, Kominek, Zhou et al. (2018) Cell

Distribution of 878 Horizontally Acquired Genes

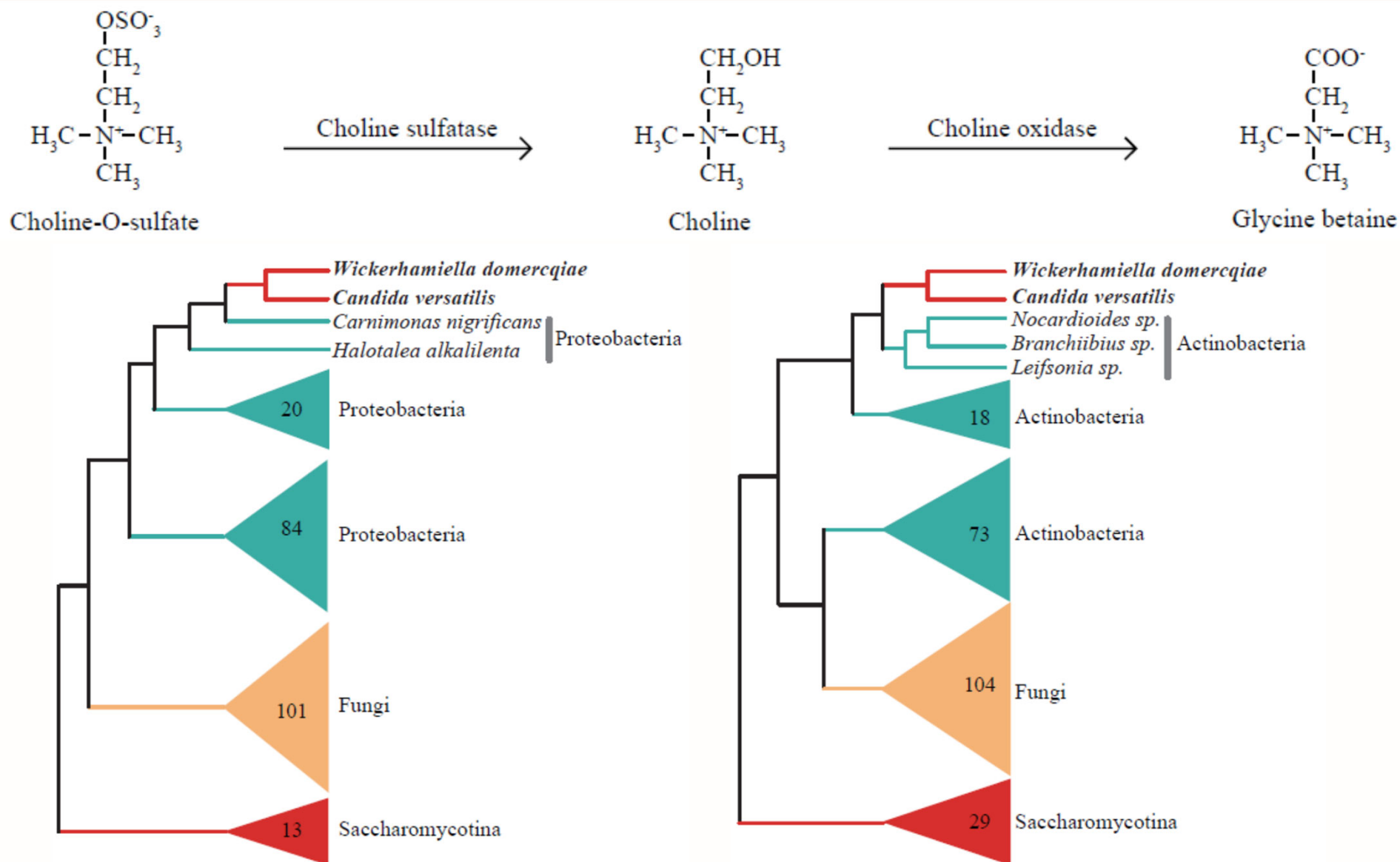


HGT in 226 yeasts with universal code: 0.071%

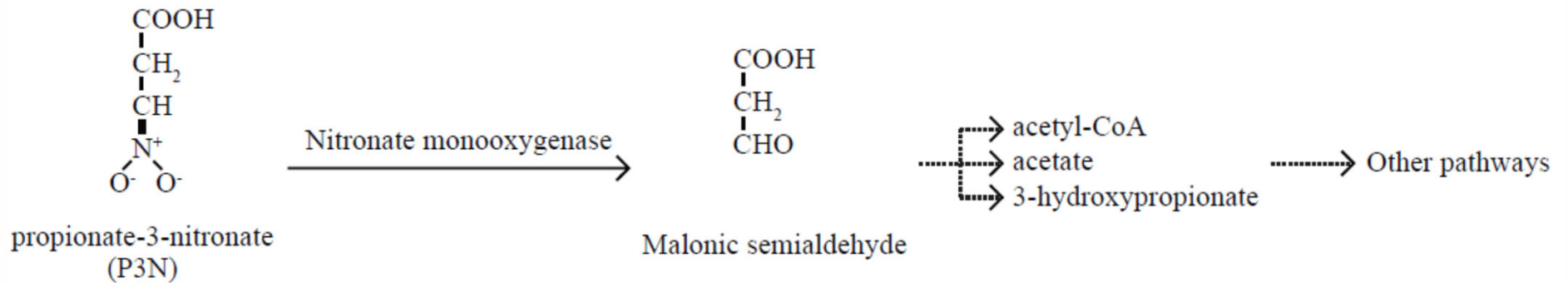
HGT in 103 yeasts with non-universal code: 0.025%



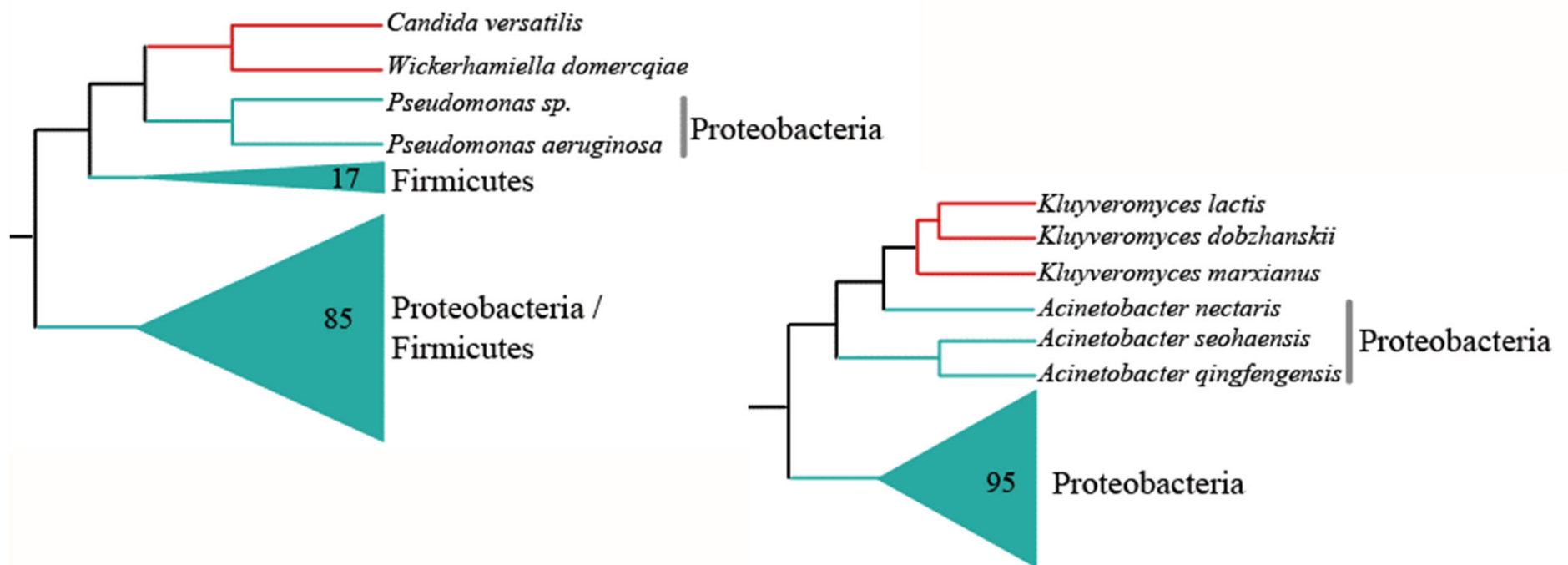
Osmotolerant Yeasts Acquired Enzymes for Osmoprotectant Betaine



Two Yeast Lineages Independently Acquired P₃N Detoxifier Genes



plant and
fungal toxin



Shen, Opulente, Kominek, Zhou et al. (2018) Cell

Loss is More

Specialization is often considered an evolutionary dead end

Starting 400 mya, a metabolically complex budding yeast gave (and keeps on giving) rise to organisms with highly diverse metabolisms & ecologies

- ❖ **Differential pathway loss**
- ❖ **Gene duplication and functional divergence**
- ❖ **HGT**

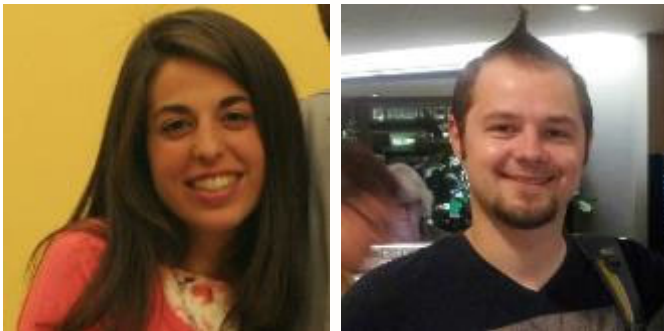
Acknowledgments

Hittinger lab – UW

Jacek Kominek

Dana Opulente

And many others



Kurtzman lab – USDA

Jeremy DeVirgilio



<http://y100plus.org>

Rokas lab – Vandy

Xing-Xing Shen

Xiaofan Zhou

Jacob Steenwyk

Leonidas Salichos

And many others



Budding Yeast Community

Moriya Ohkuma, Rikiya Endoh,

Masako Takashima, Ri-ichiroh

Manabe, Neža Čadež, Diego Libkind,

Carlos A. Rosa, Marizeth Groenewald



<http://www.rokaslab.org/>

@RokasLab