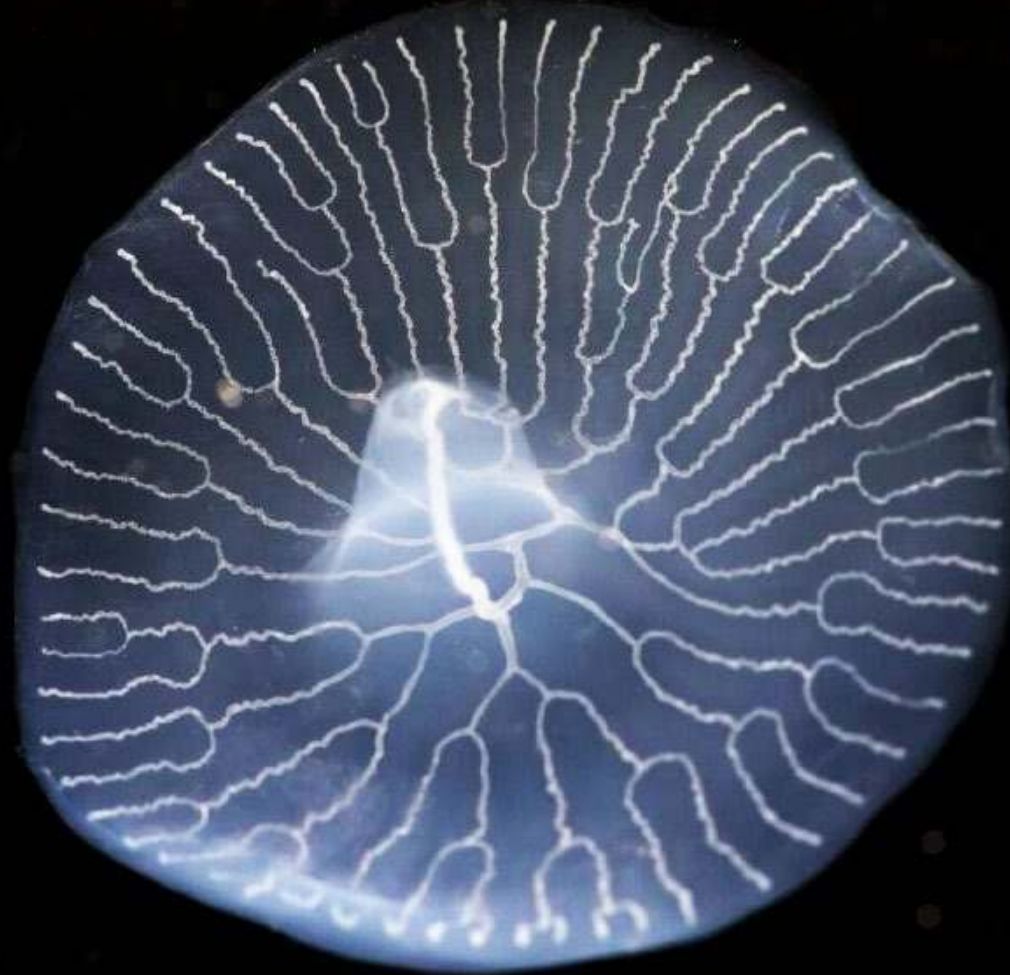


Incongruence



Antonis Rokas

Department of Biological Sciences, Vanderbilt University

<http://www.rokaslab.org>

@RokasLab

Brief Bio

93-98	B.Sc.	Biology	Univ. of Crete, Greece (Advisor: L. Zouros)
98-01	Ph.D.	Evolutionary Ecology	Edinburgh Univ., Scotland (Advisor: G. Stone)
02-05	PostDoc	Evolutionary Genomics	Univ. Wisconsin-Madison (Advisor: S. Carroll)
05-07	Res. Scientist	Fungal Genomics	Broad Institute
07-now:	Faculty	Evolutionary Biology	Vanderbilt University

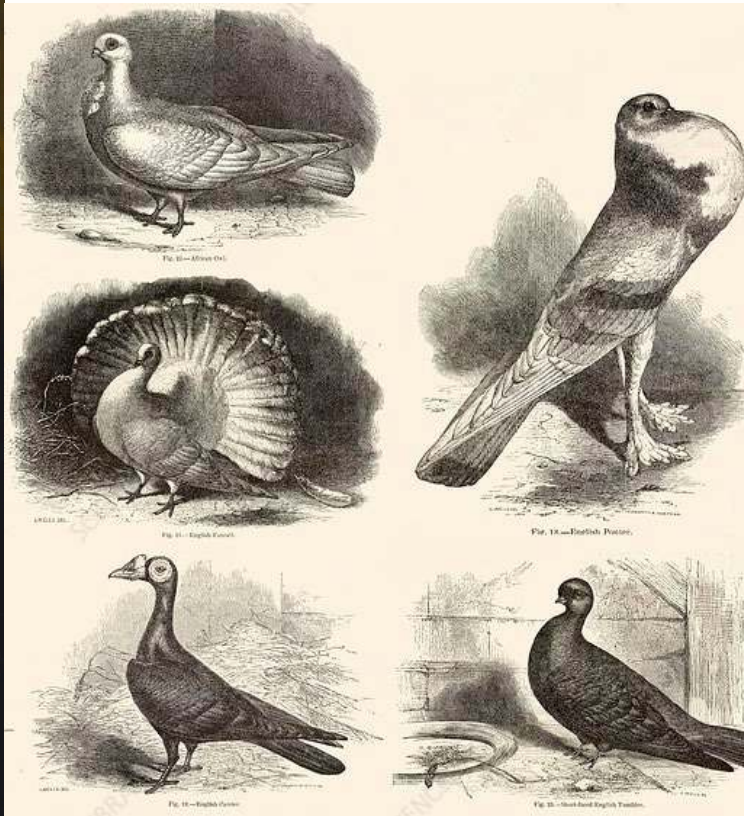
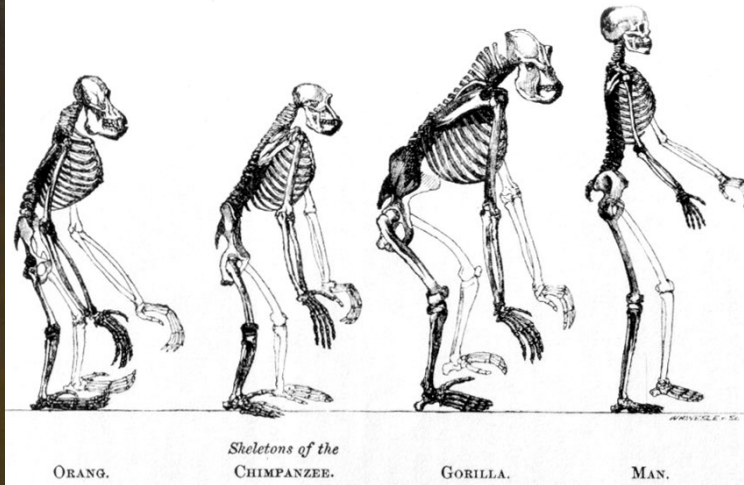


The Workshop on Molecular Evolution@Woods Hole



Antonis (me!) & Scott Handley

Darwin's data



The DNA record



“The genome is, it's a fossil record; the genome is a landscape; the genome is a whole geography of distributions. [...] you might think the genome's just a boring string of letters [...] 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

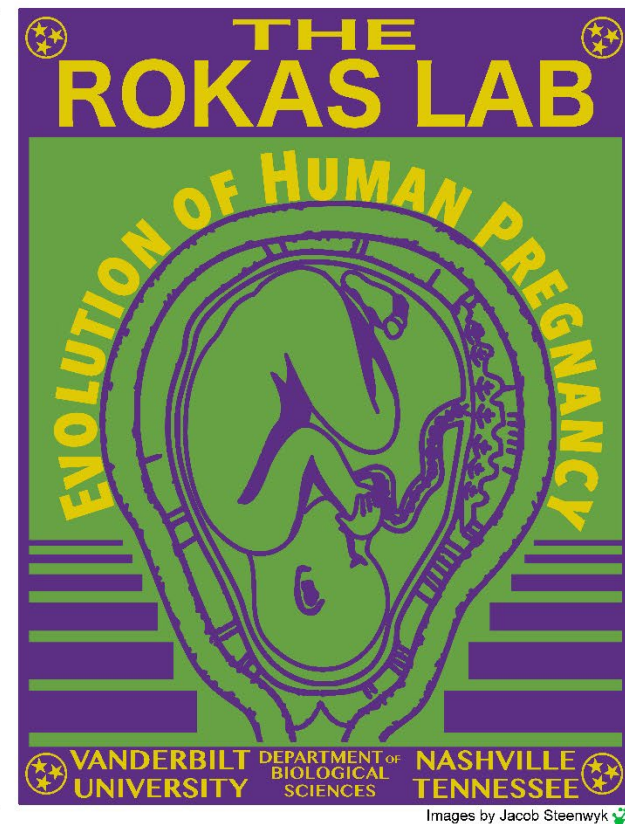
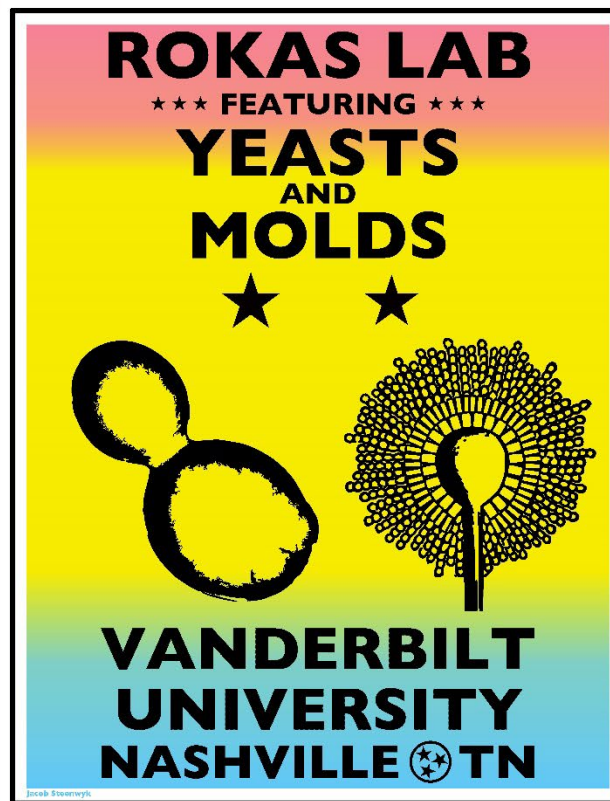
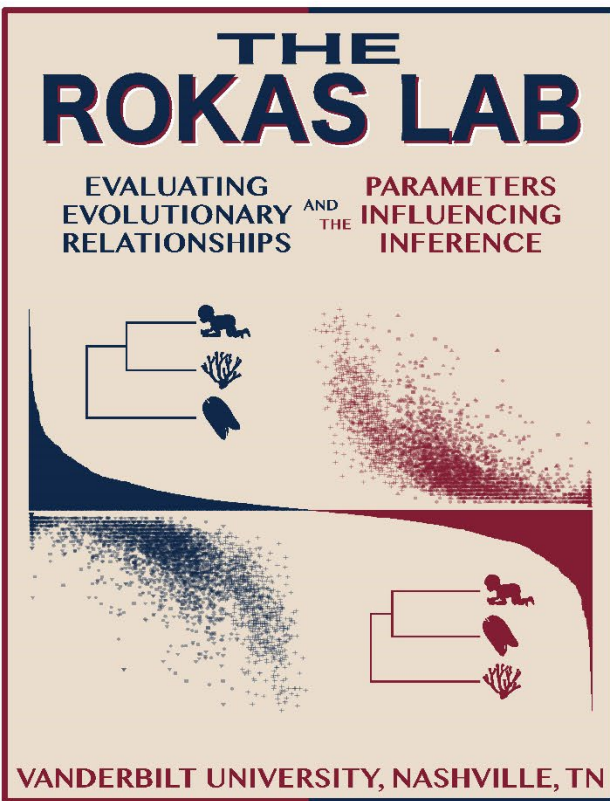


We study the DNA record to gain insight into evolutionary patterns and processes using computational and experimental approaches

Vanderbilt Univ. is in Nashville, Music City - USA





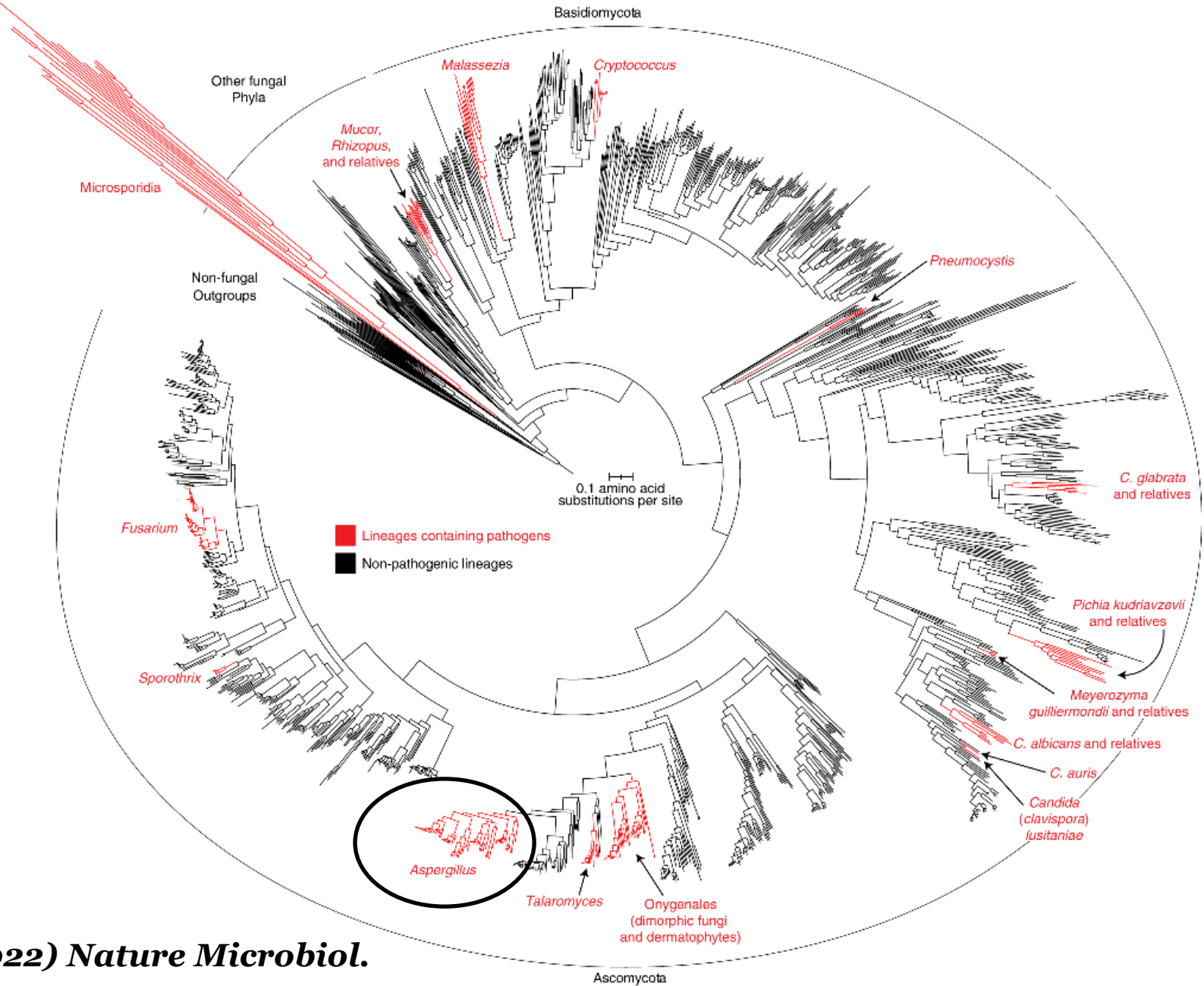


Phylogenomics
(NSF)

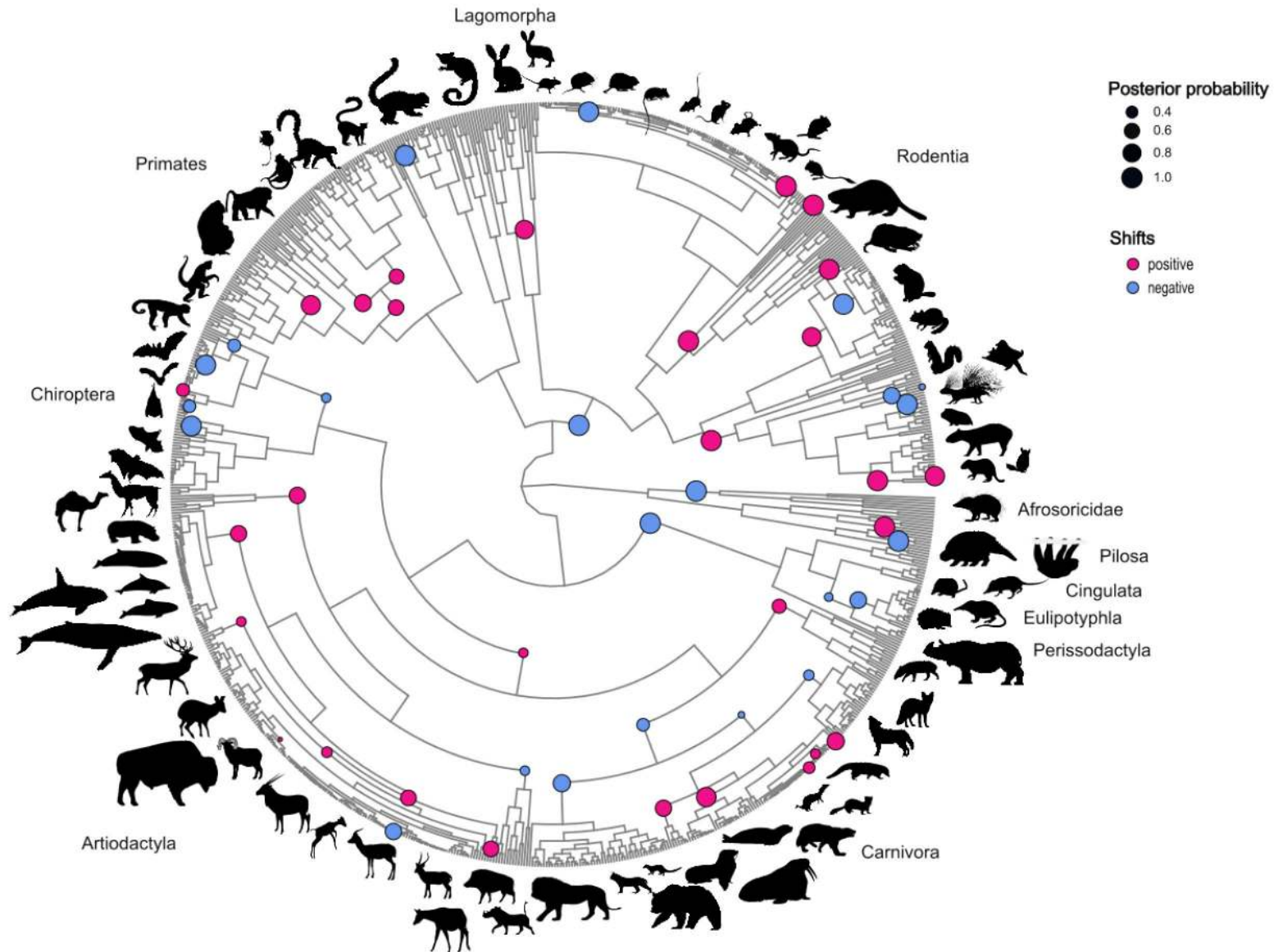
The molecular foundations of the fungal lifestyle
(NSF & NIH)

The evolution of mammalian pregnancy
(BWF & March of Dimes)

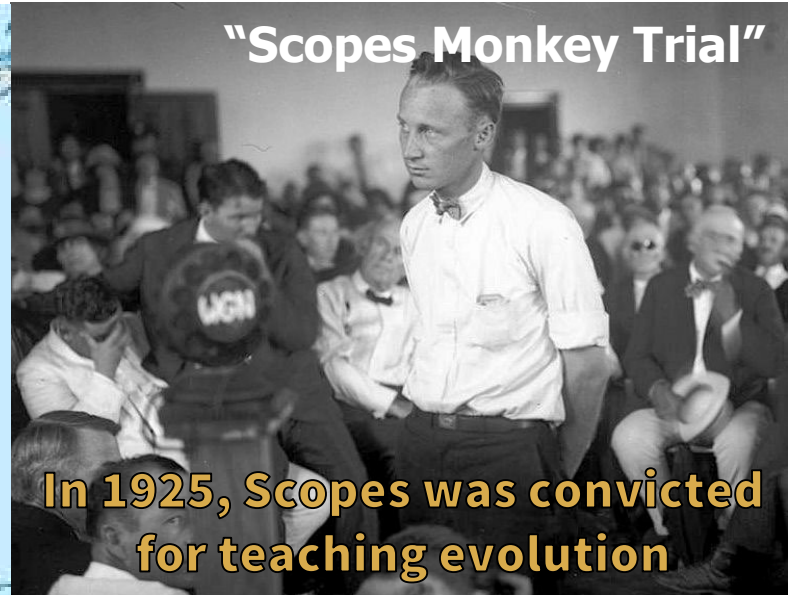
The repeated evolution of fungal pathogenicity



Evolution of variation in gestation length in mammals



...But I'm also an evolutionist in Tennessee (USA)



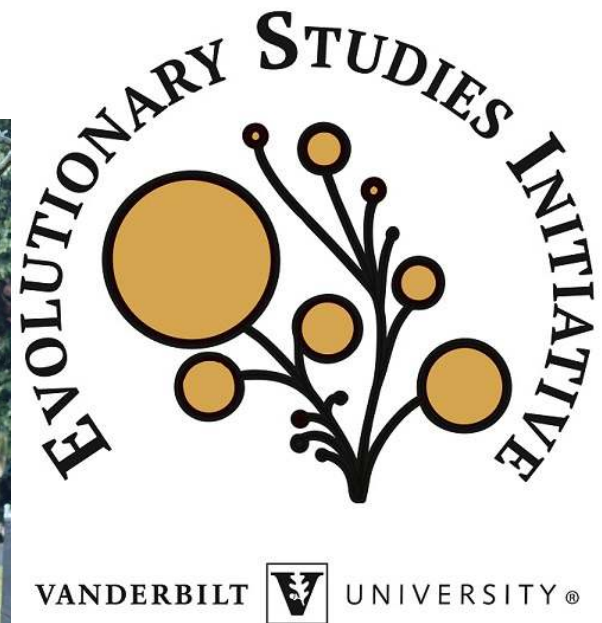
Foes of science faced ridicule at the Scopes trial.
We're paying the price 95 years later.

The Washington Post



Opinion by **Max Boot**
Columnist

July 8, 2020 at 1:01 p.m. CDT



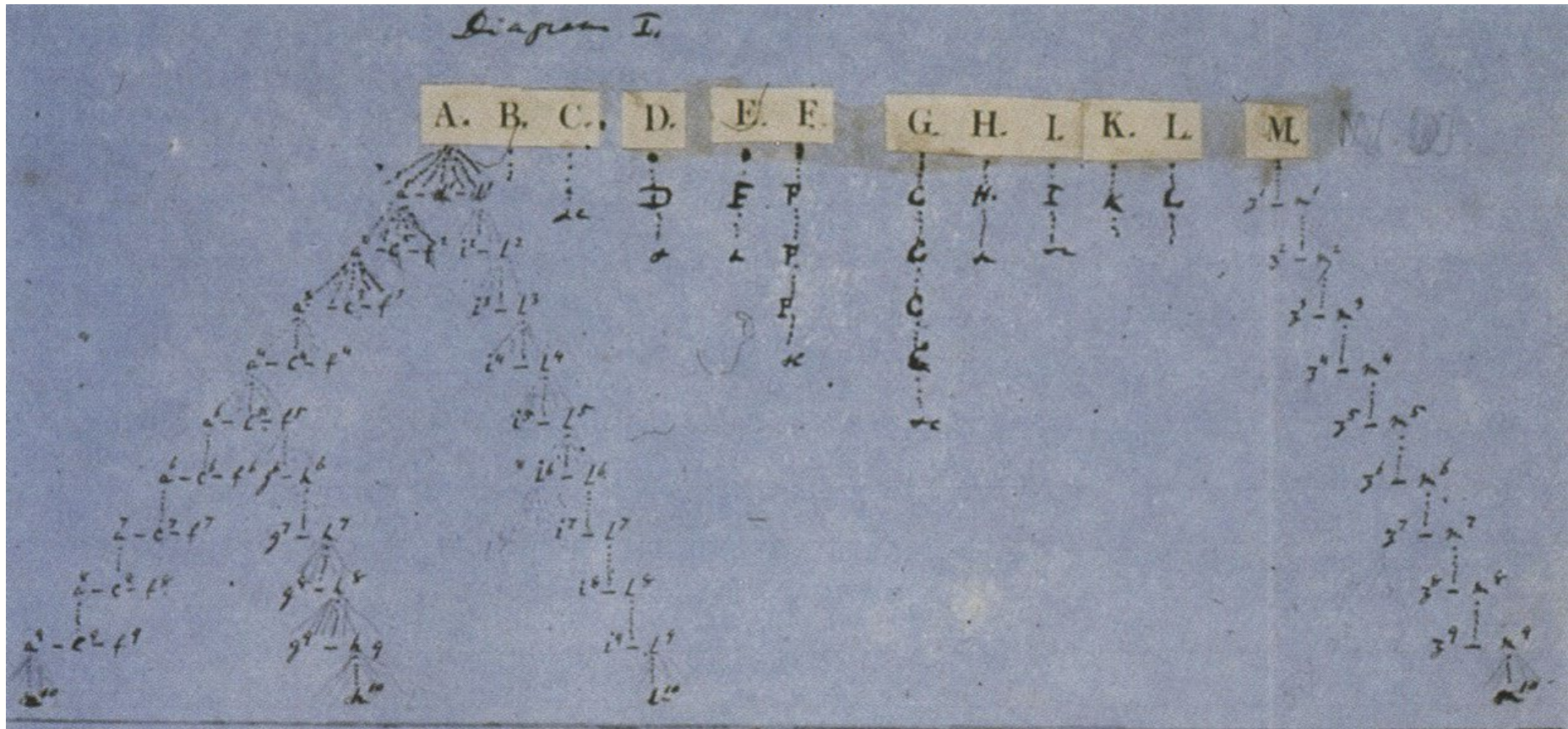
www.vanderbilt.edu/evolution/

❖ **Incongruence and its causes**

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

❖ **Handling incongruence in phylogenomic data**

Darwin's tree



Darwin's hand-made proof of the famous diagram in his Origin of Species;
Maderspacher (2006) Curr. Biol.

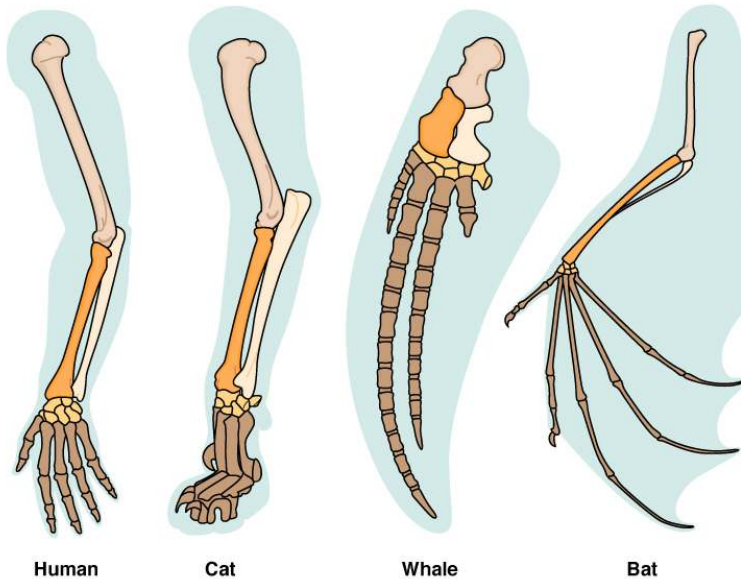


and instinct as the summing up of many contrivances, each useful to the possessor, nearly in the same way as when we look at any great mechanical invention as the summing up of the labour, the experience, the reason, and even the blunders of numerous workmen; when we thus view each organic being, how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

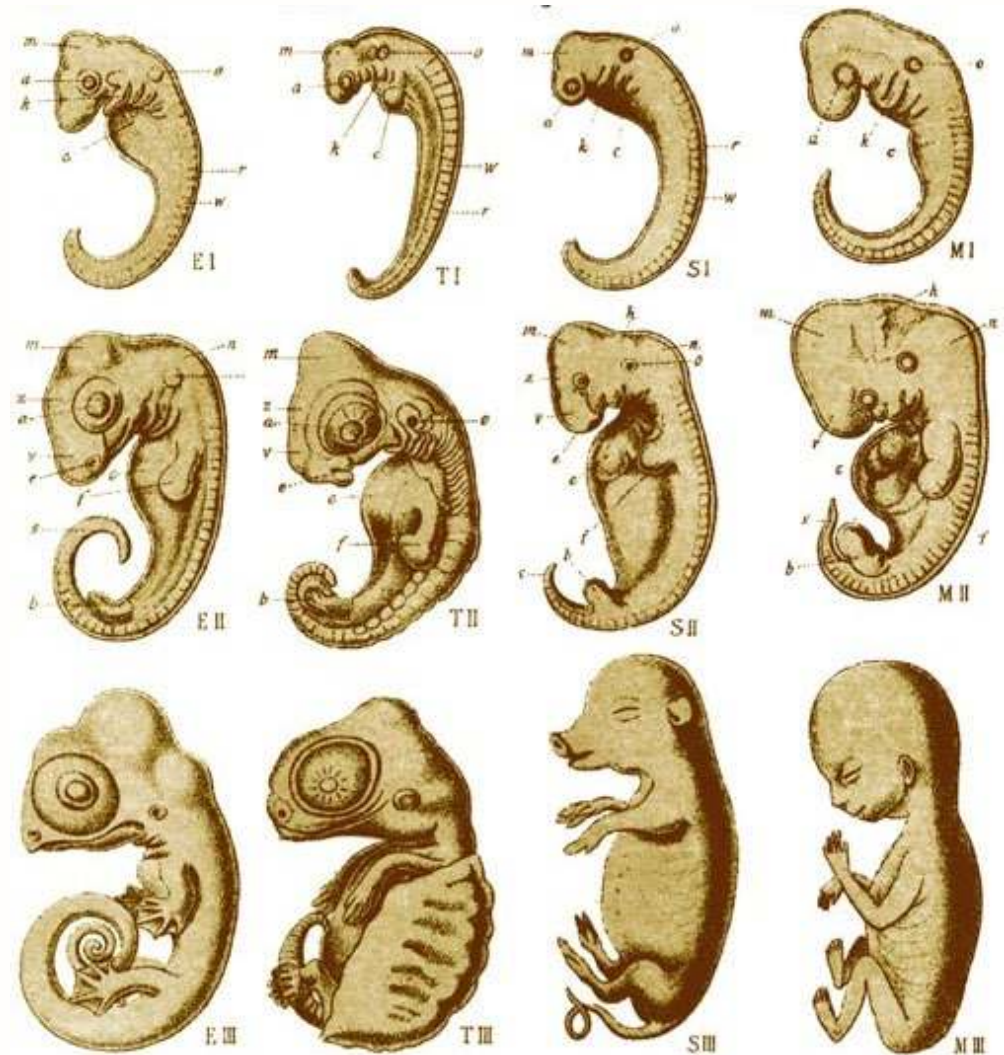
A grand and almost untrodden field of inquiry will be opened, on the causes and laws of variation, on correlation of growth, on the effects of use and disuse, on the direct action of external conditions, and so forth. The study of domestic productions will rise immensely in value. A new variety raised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

Data from comparative morphology and embryology



©1999 Addison Wesley Longman, Inc.



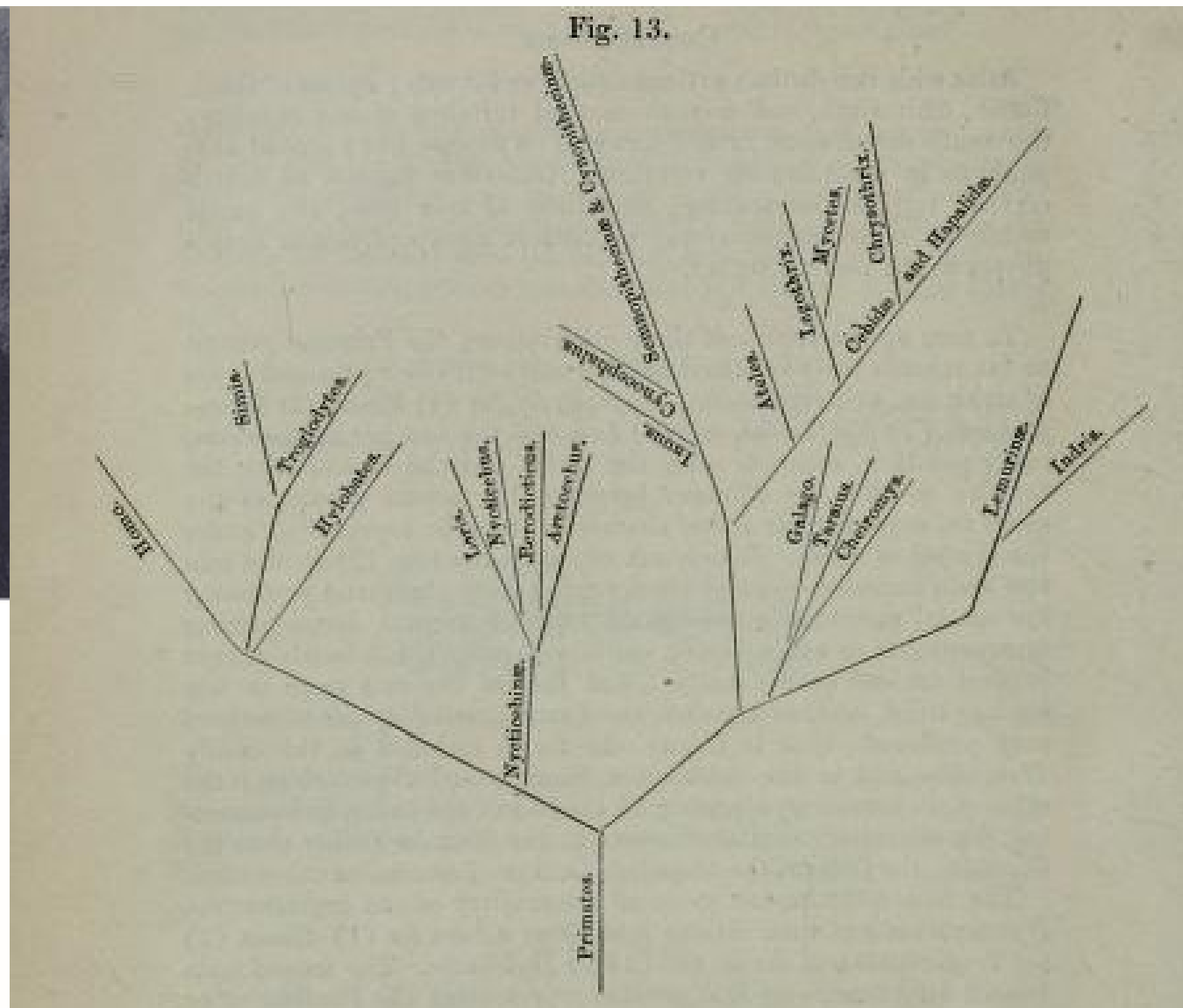
http://www.mun.ca/biology/scarr/139393_forelimb_homology.jpg &
<http://www.nature.com/nrg/journal/v7/n11/images/nrg1918-f2.jpg>



The first published phylogeny

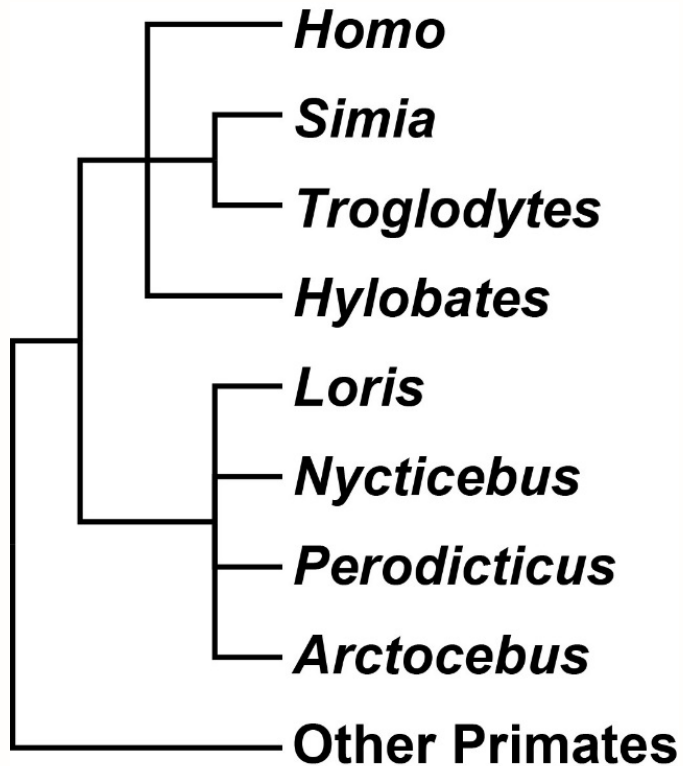


St. George Jackson Mivart

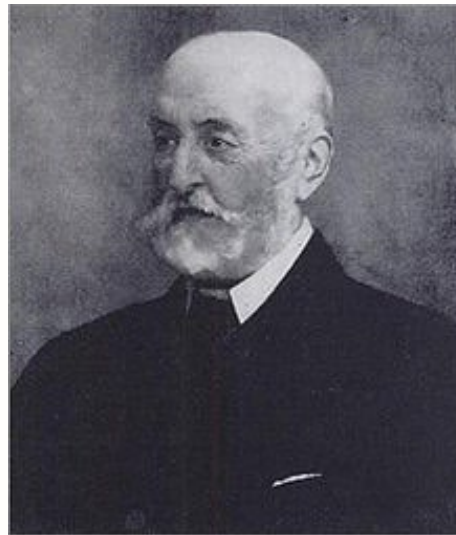
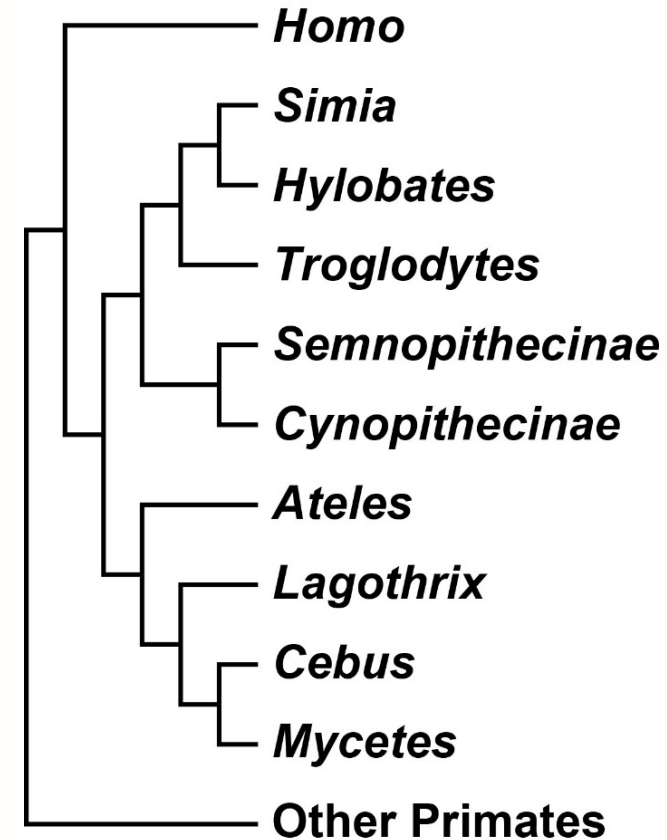


Disagreement between phylogenies

1865: SPINAL COLUMN



1867: LIMBS



St. George Jackson Mivart





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?

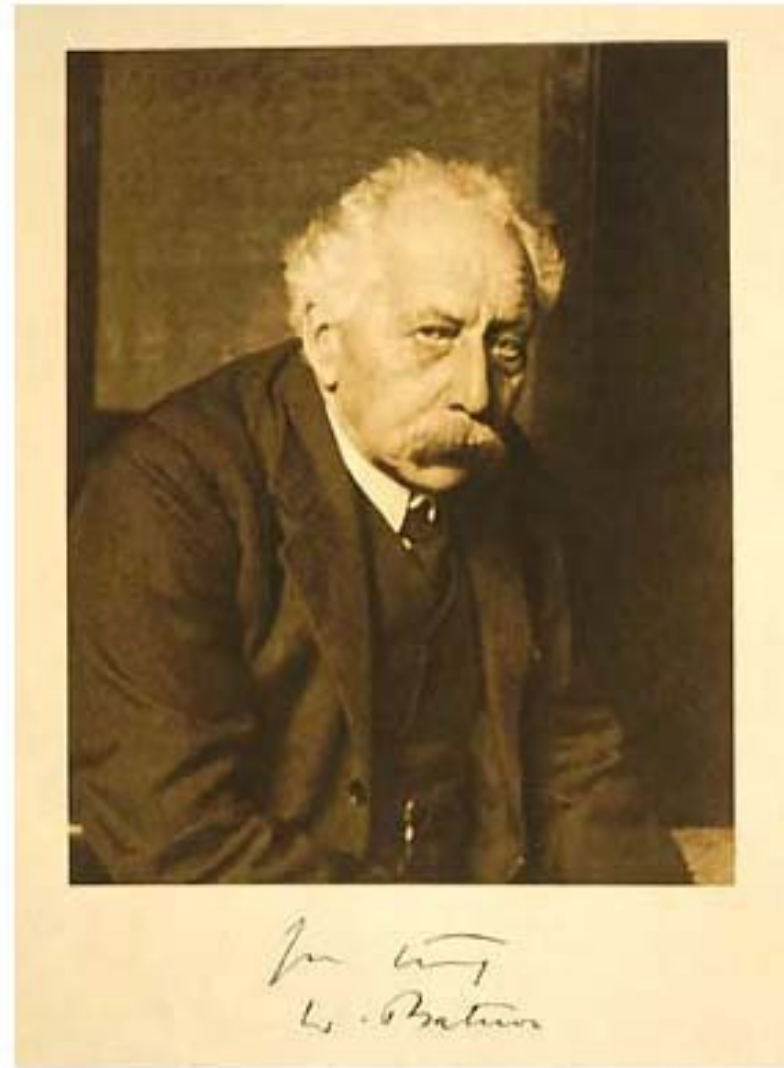
I have really expressed no opinion as to Man's origin nor am I prepared to do so at this moment. The [1865] diagram [...] expresses what I believe to be the degree of resemblance as regards the spinal column only. The [1867] diagram expresses what I believe to be the degree of resemblance as regards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



“From the same facts, opposite conclusions are drawn; facts of the same kind will take us no further. [...] Need we waste more effort in these vain and sophistical disputes. If facts of the old kind will not help, let us seek facts of a new kind.”

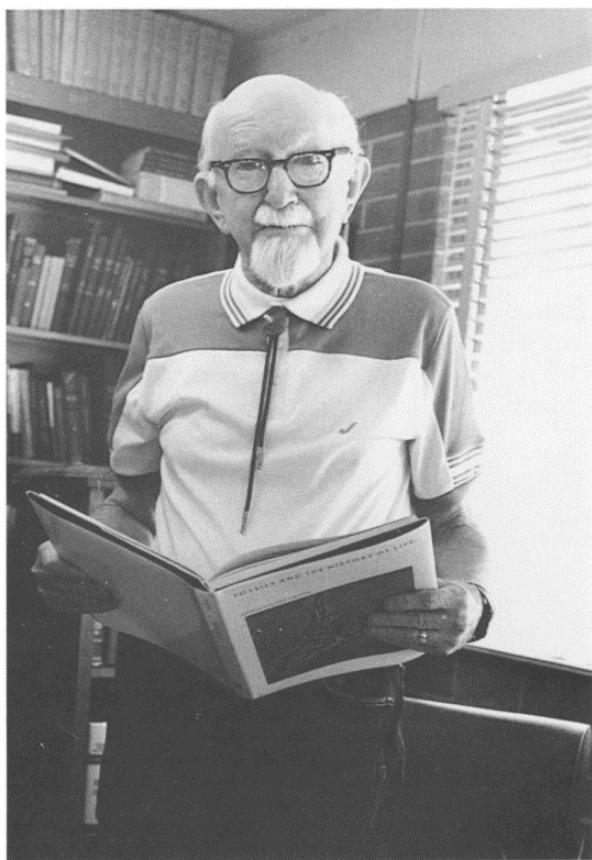
William Bateson (1894)

Materials for the Study of Variation

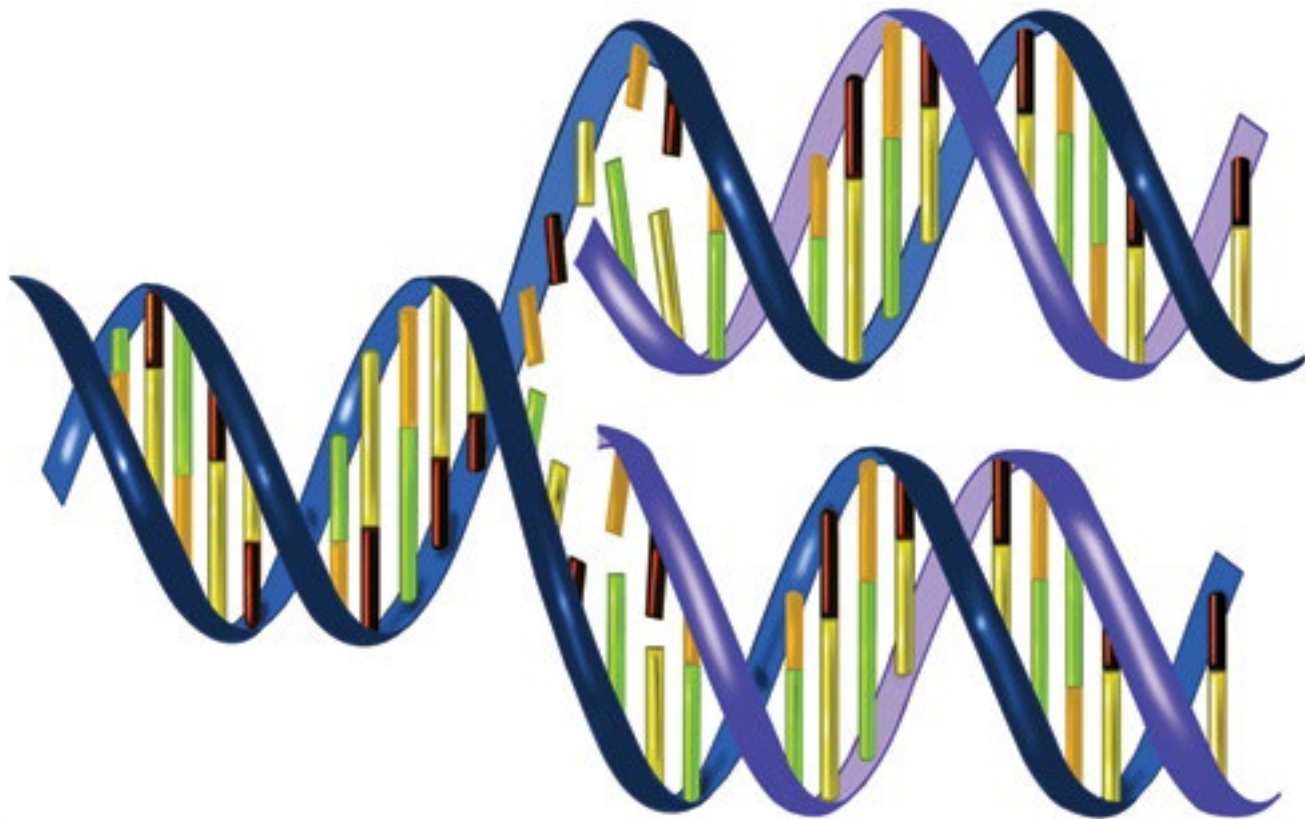


Courtesy of American Philosophical Society, Curt Stern Papers.
Noncommercial, educational use only.





George Gaylord Simpson



“The stream of heredity makes phylogeny; in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream”

G. G. Simpson, 1945

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

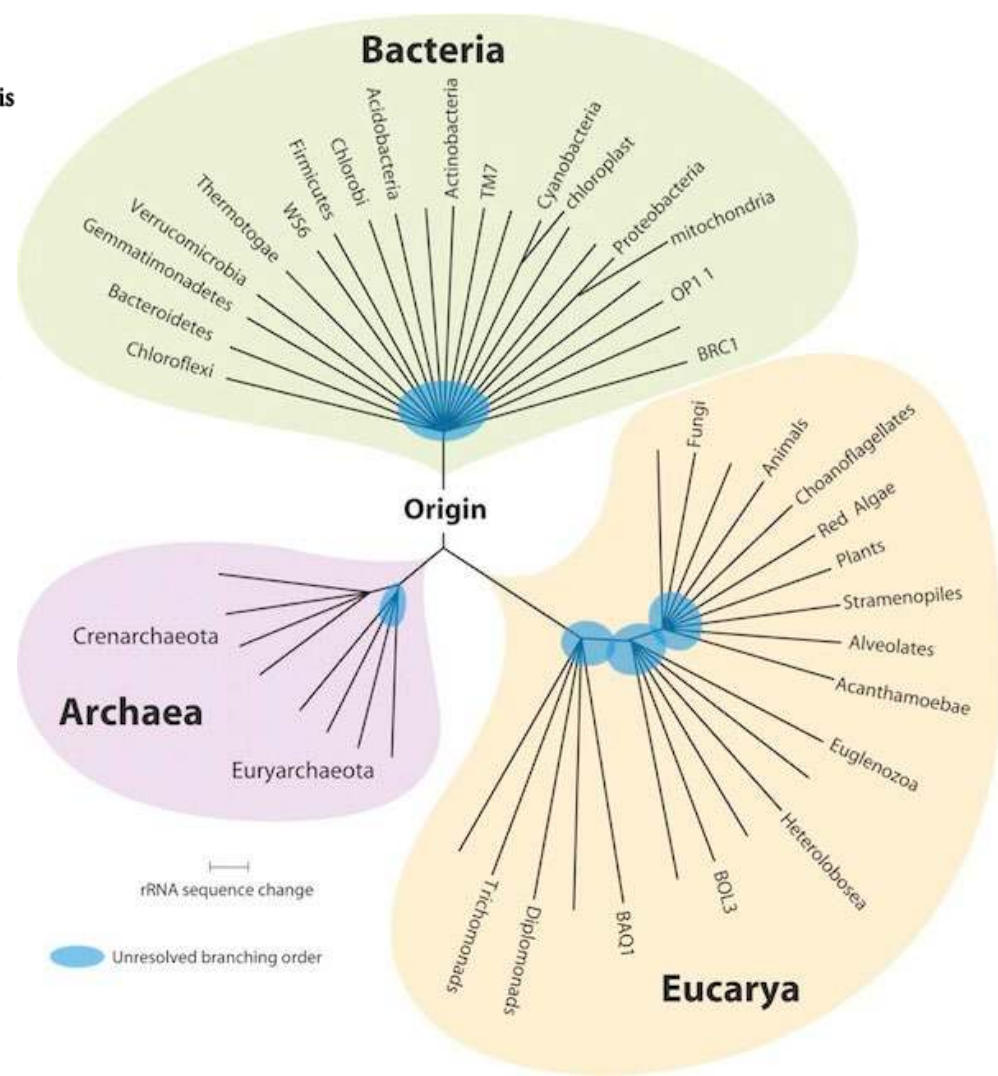
(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

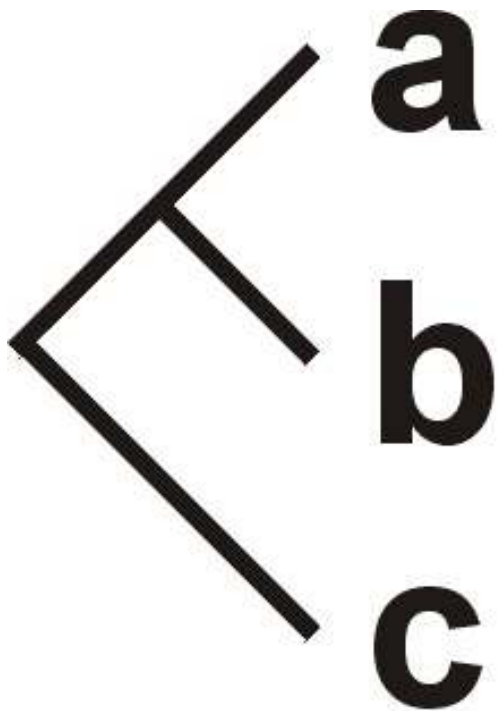
Department of Genetics and Development, University of Illinois, Urbana, Illinois

Communicated by T. M. Sonneborn, August 18, 1977

ABSTRACT A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (i) the eubacteria, comprising all typical bacteria; (ii) the archaebacteria, containing methanogenic bacteria; and (iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.



Incongruence / conflict / discordance



Gene X

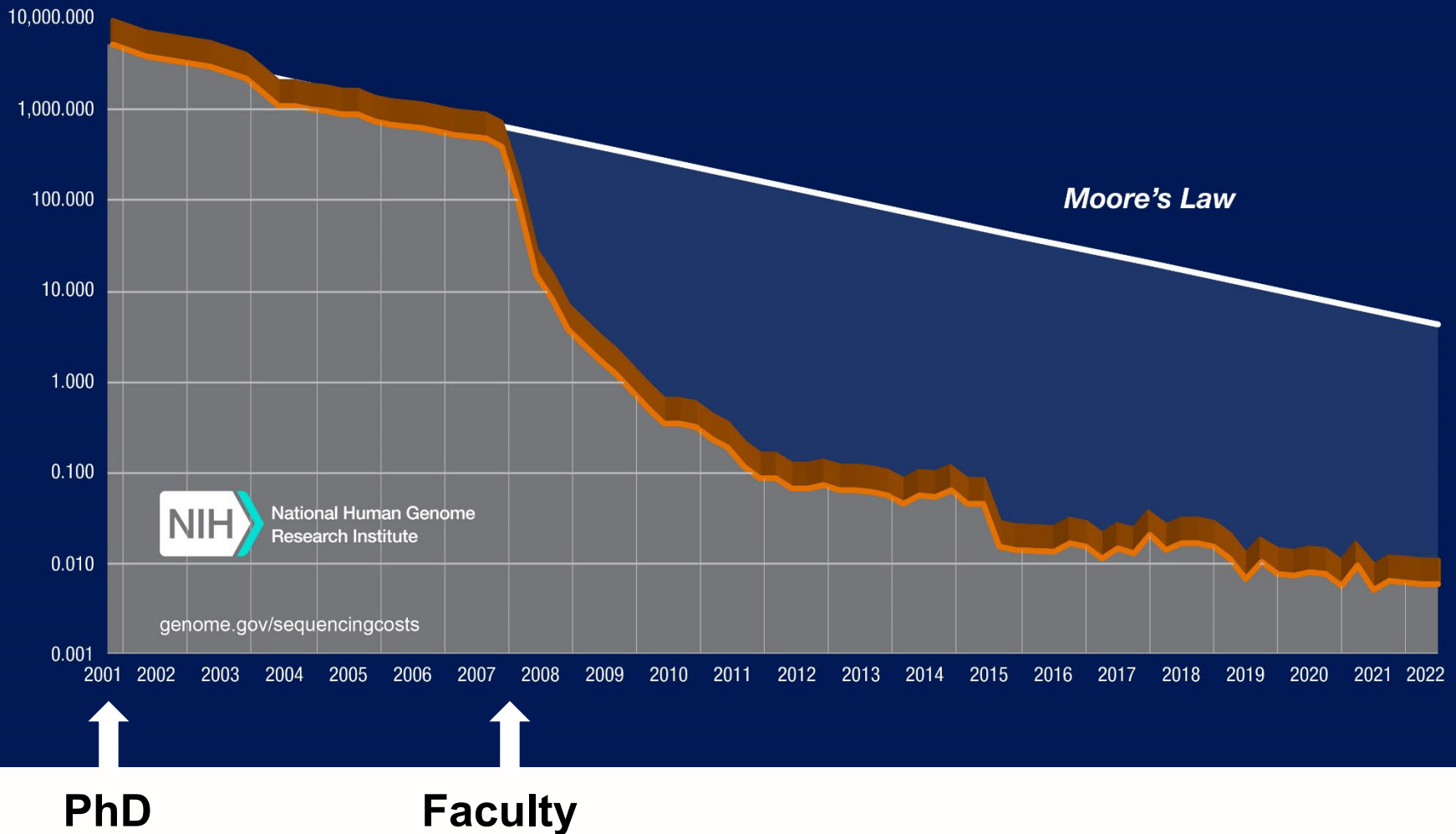


Gene Y

**Species
phylogeny?**

The genomics revolution

Cost per Raw Megabase of DNA Sequence



A systematic evaluation of single gene phylogenies



S. cerevisiae

S. paradoxus

S. mikatae

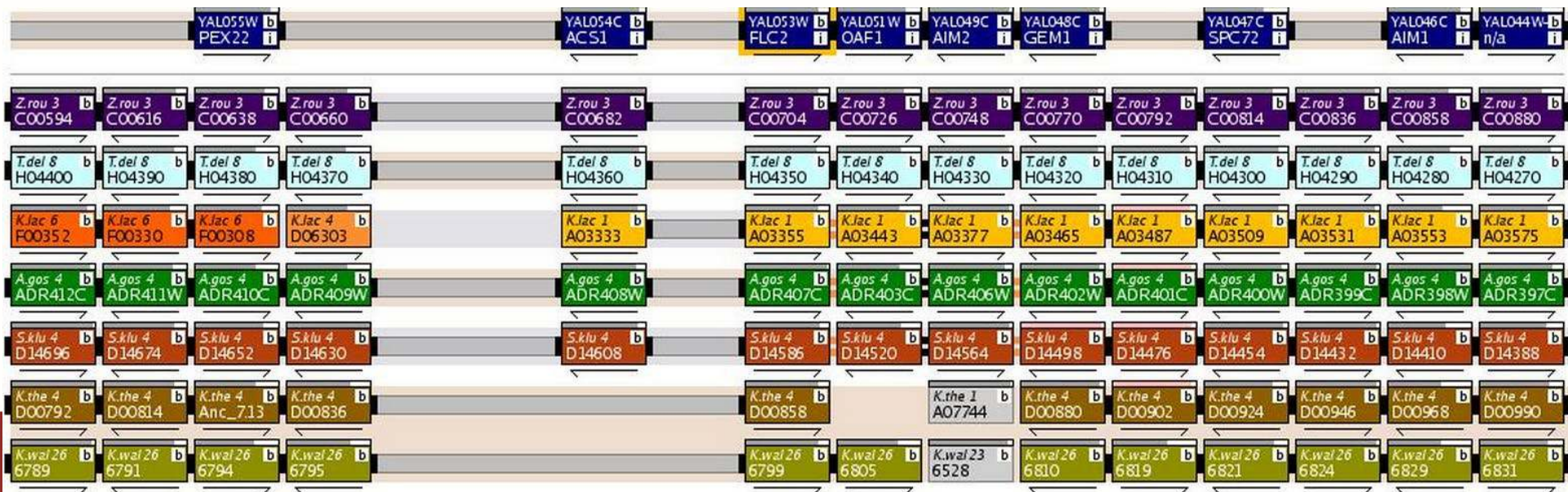
S. kudriavzevii

S. bayanus

S. castellii

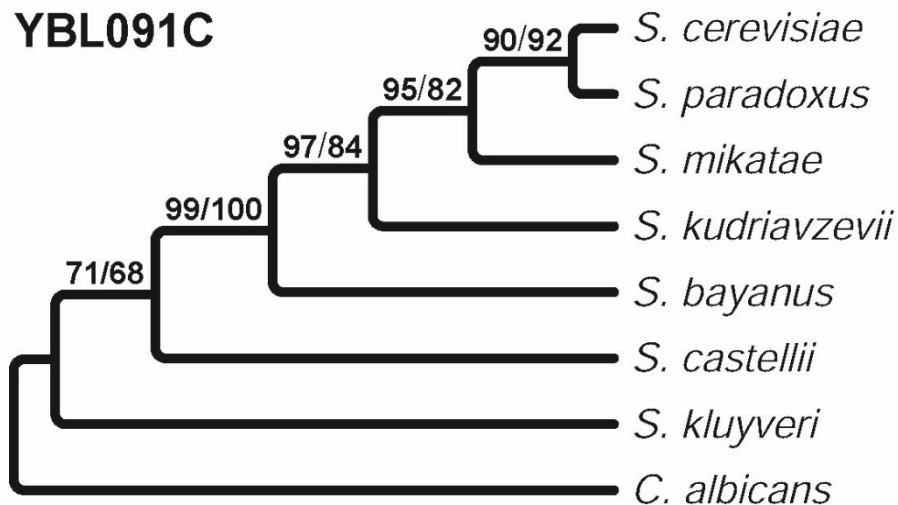
S. kluyveri

Candida glabrata

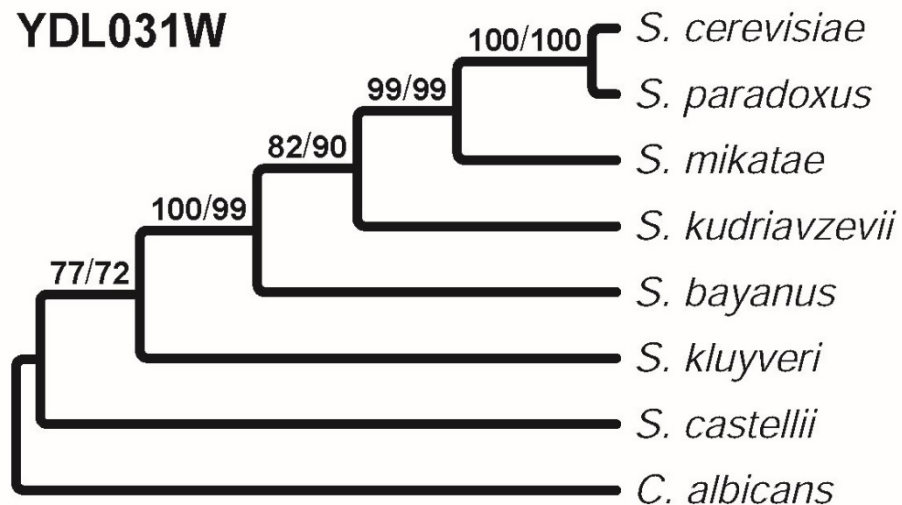


Incongruence at the single gene level

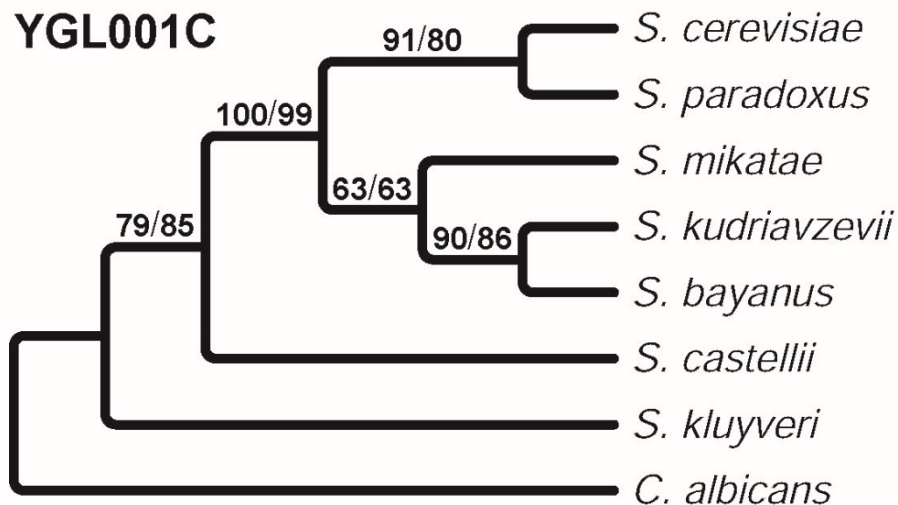
YBL091C



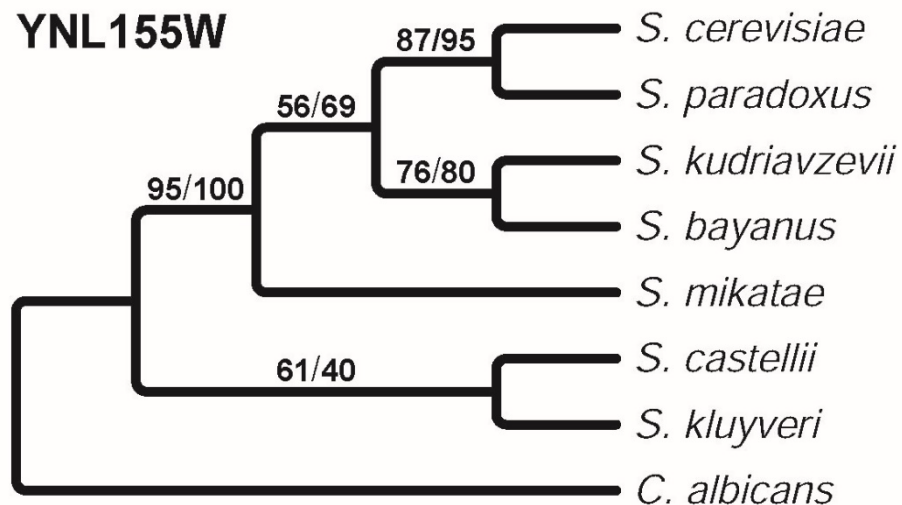
YDL031W



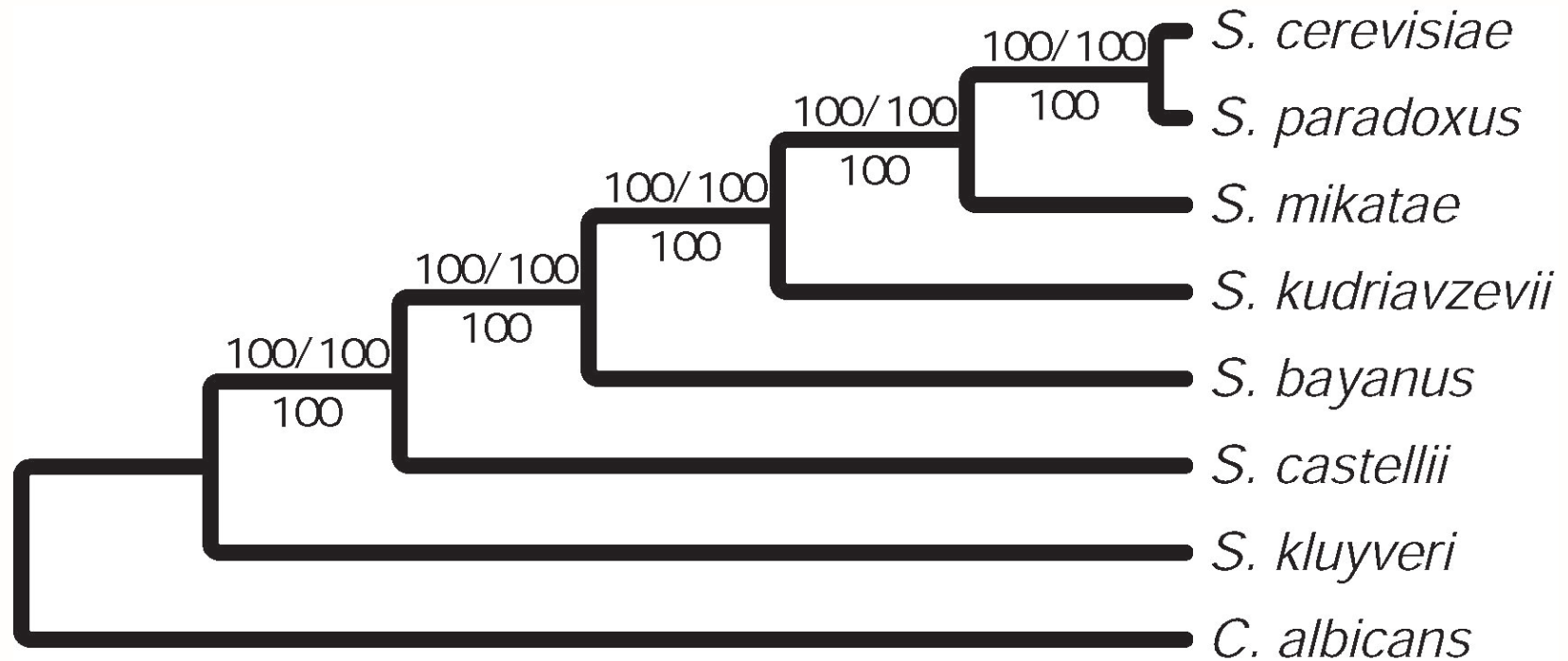
YGL001C



YNL155W



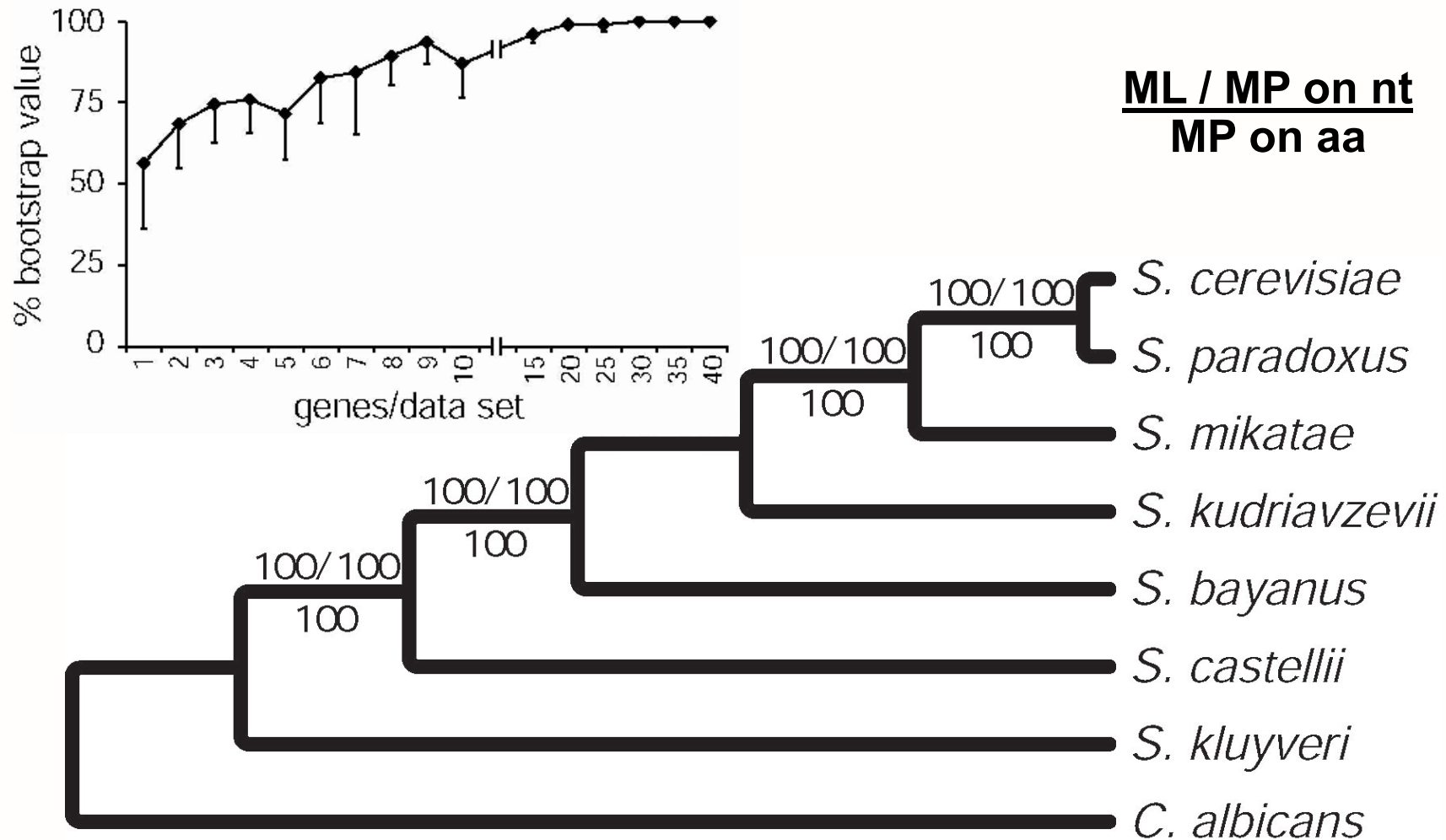
Concatenation of 106 genes yields a single species phylogeny



ML / MP on nt
MP on aa



The use of many genes eliminates incongruence



The dawn of the phylogenomics era

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

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 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

Yeast

Toward Resolving the Tree: The Phylogeny of Jakobids and Cercozoans

An
OPEN ACCESS Free
Toward Resolving Priors

Prion-Like Proteins in the Fungal Kingdom

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

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?**

2 sets of factors influence the gene tree – species tree relationship

Biological factors

They lead to gene trees whose histories may differ from each other and from the species tree. Known factors include **stochastic lineage sorting**, **hidden paralogy**, **horizontal gene transfer**, **recombination**, **hybridization / introgression**, and **natural selection**

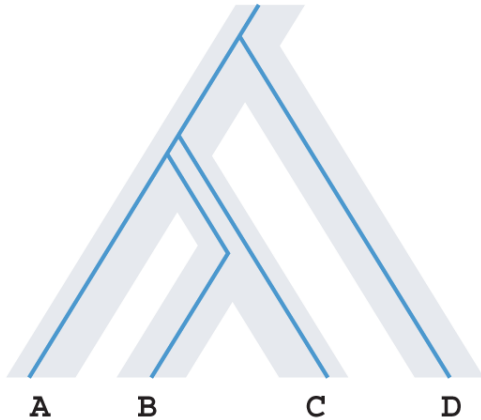
Analytical factors

They lead to failure in accurately inferring a gene tree; these can be either due to **stochastic error** (e.g., insufficient number of genes or taxa), **systematic error** (e.g., observed data deviate from model assumptions), or **treatment error** (e.g., excessive trimming)

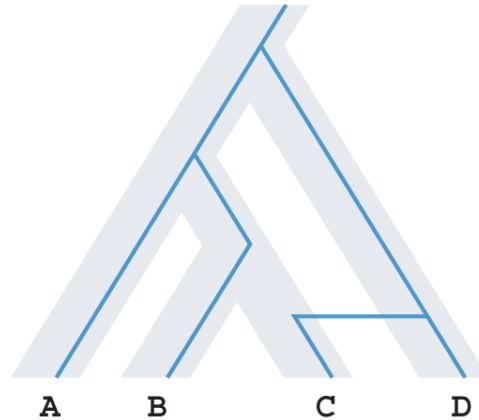


Biological factors

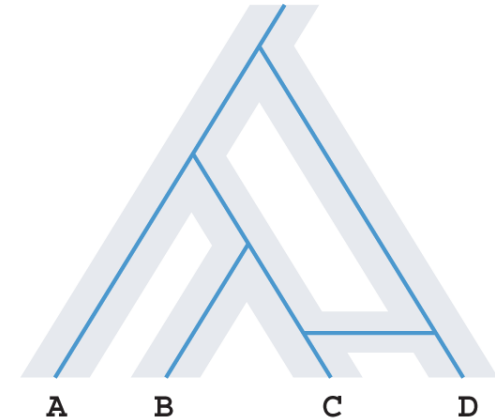
Lineage sorting



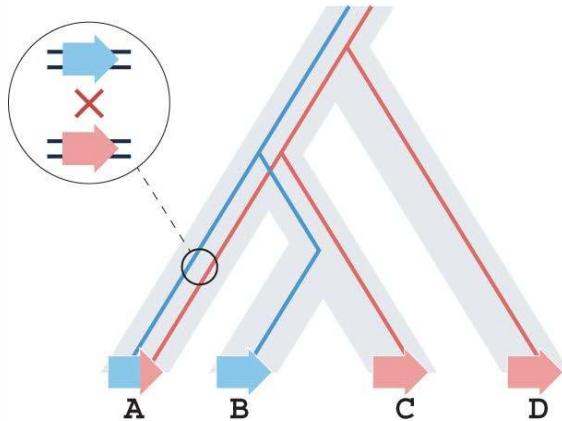
Horizontal transfer



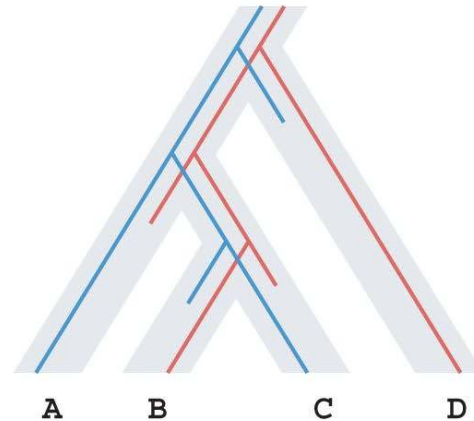
Hybridization / Introgression



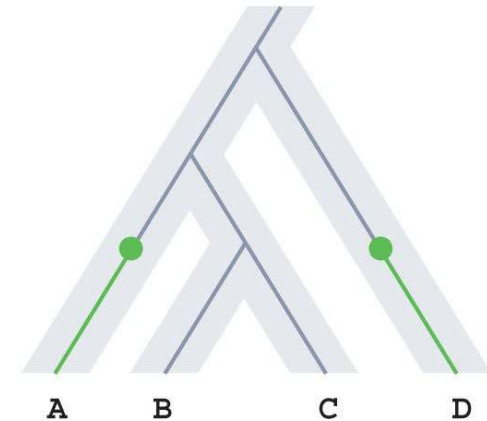
Recombination



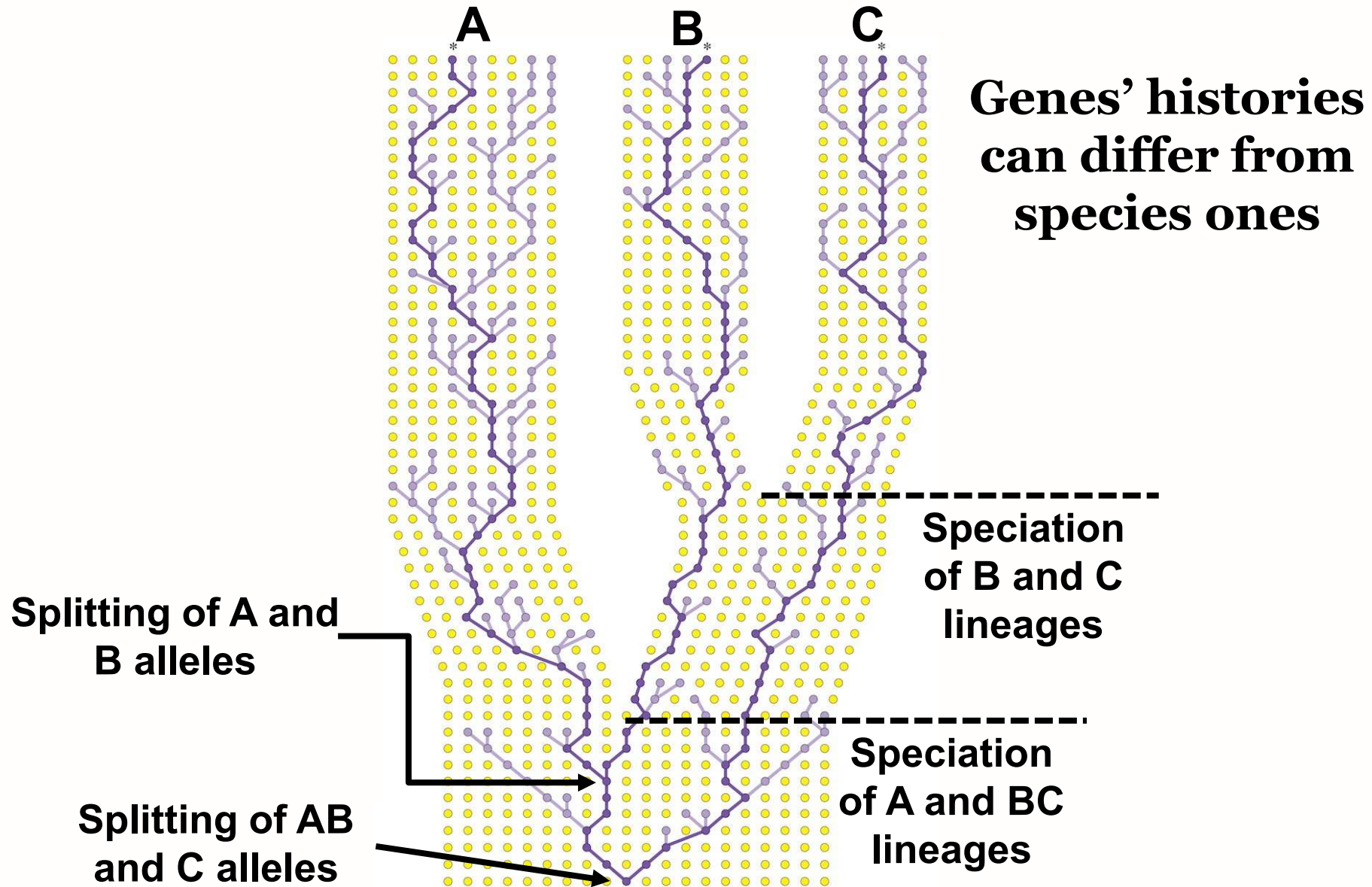
Duplication & Loss



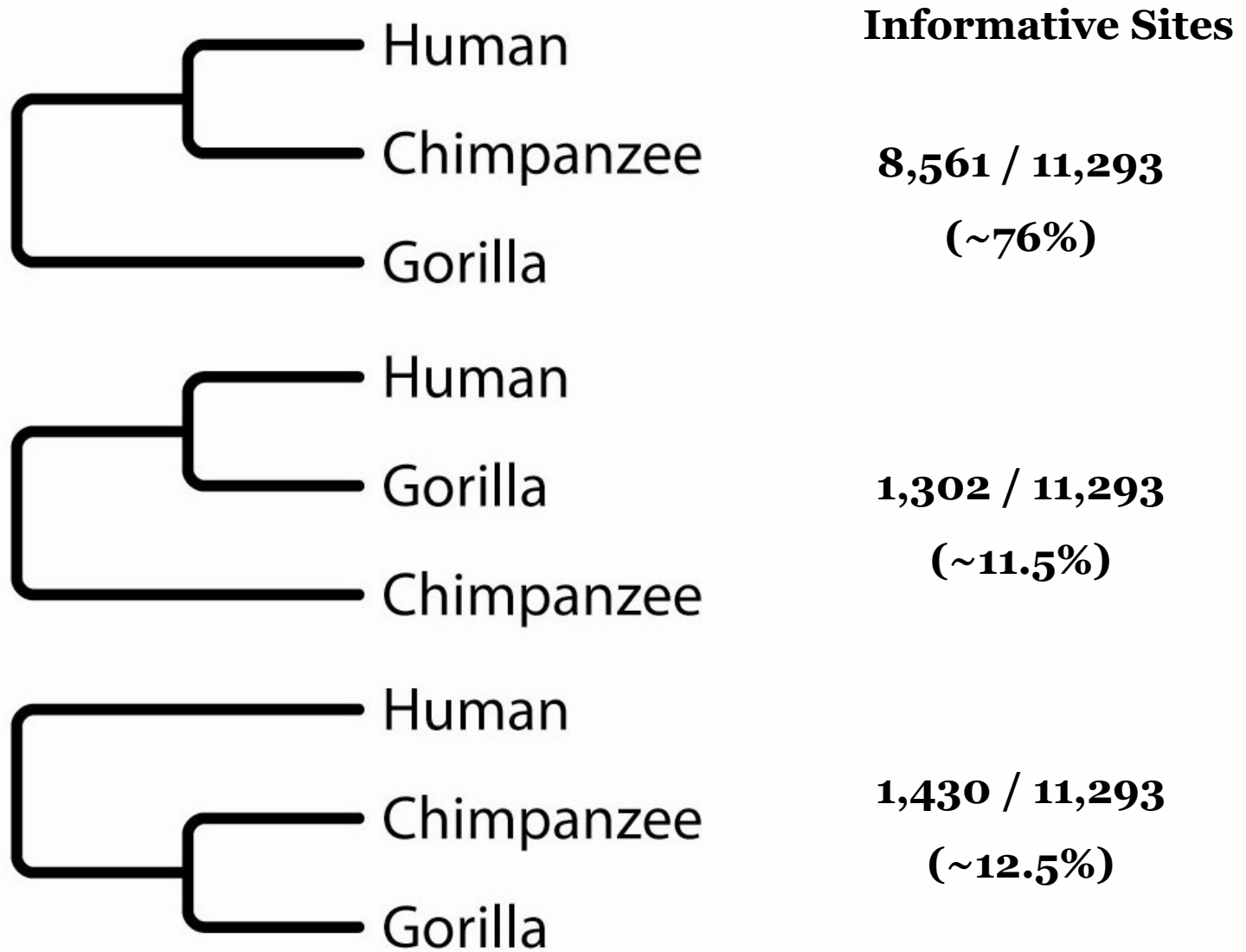
Selection



Stochastic lineage sorting of ancestral polymorphisms

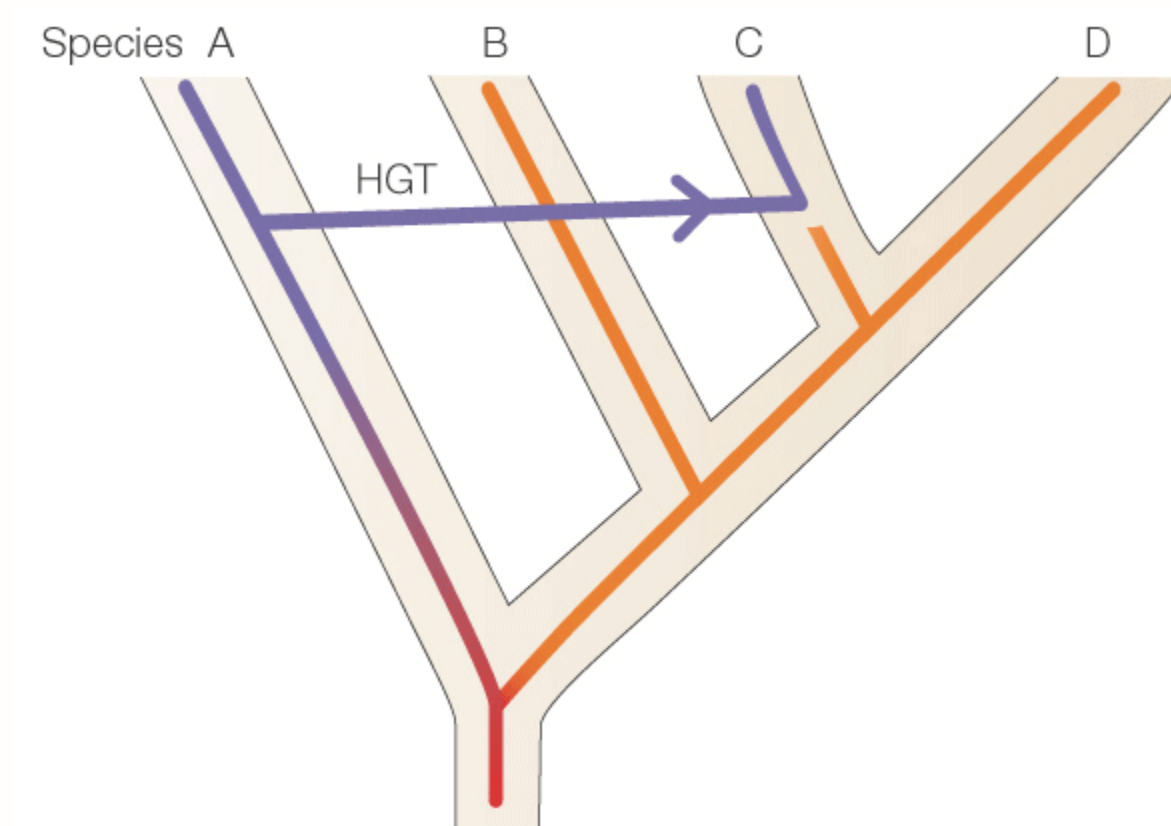


Lineage sorting in primates



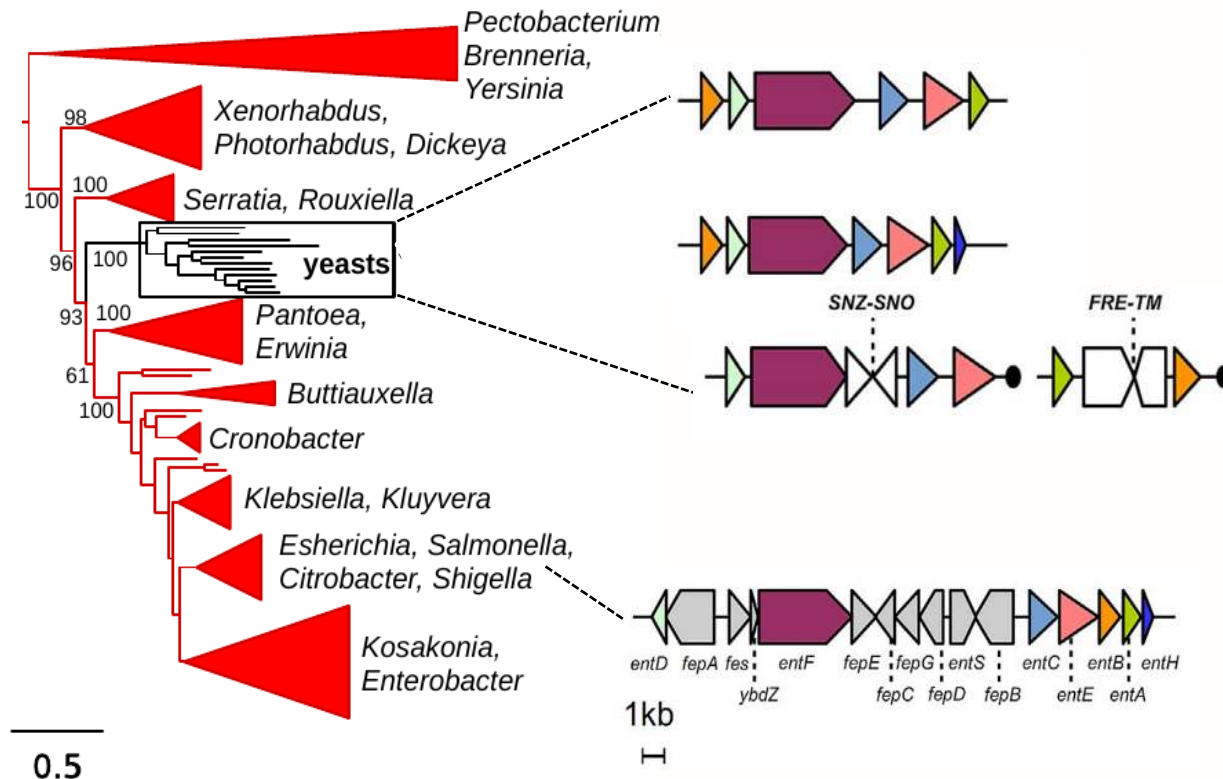
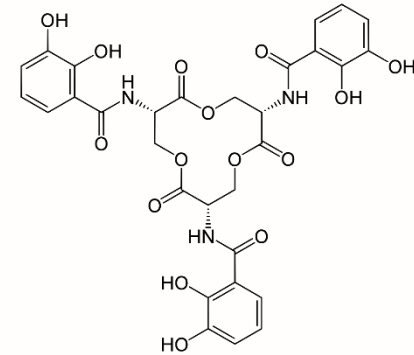
Horizontal transfer of genes

Exchange of genes between organisms other than through reproduction



Horizontal transfer of an entire operon in yeasts

A clade of yeasts acquired the enterobactin operon from **Enterobacteria** – organisms from both lineages co-occur in insect guts, where iron is a growth-limiting factor

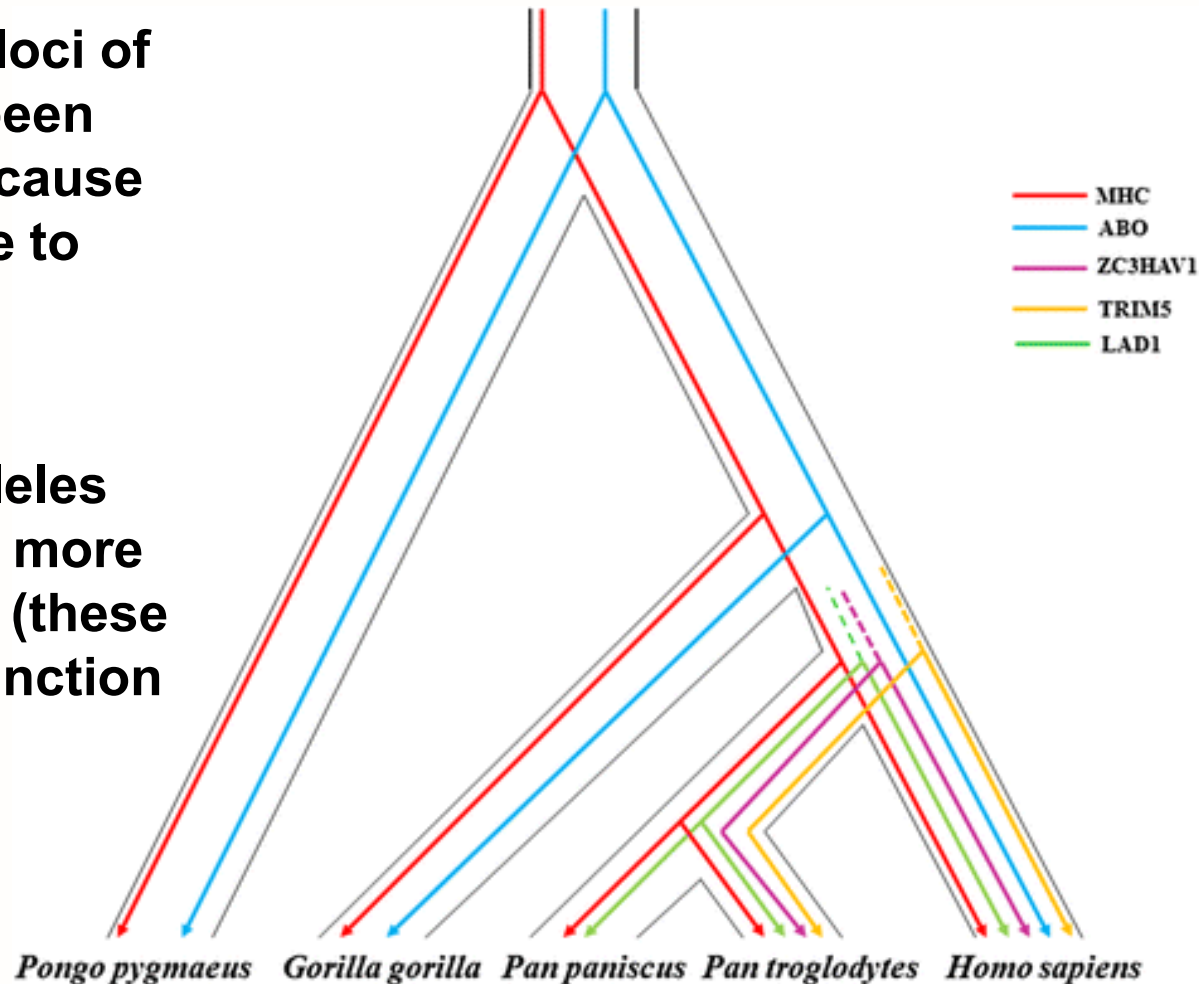


Balancing selection

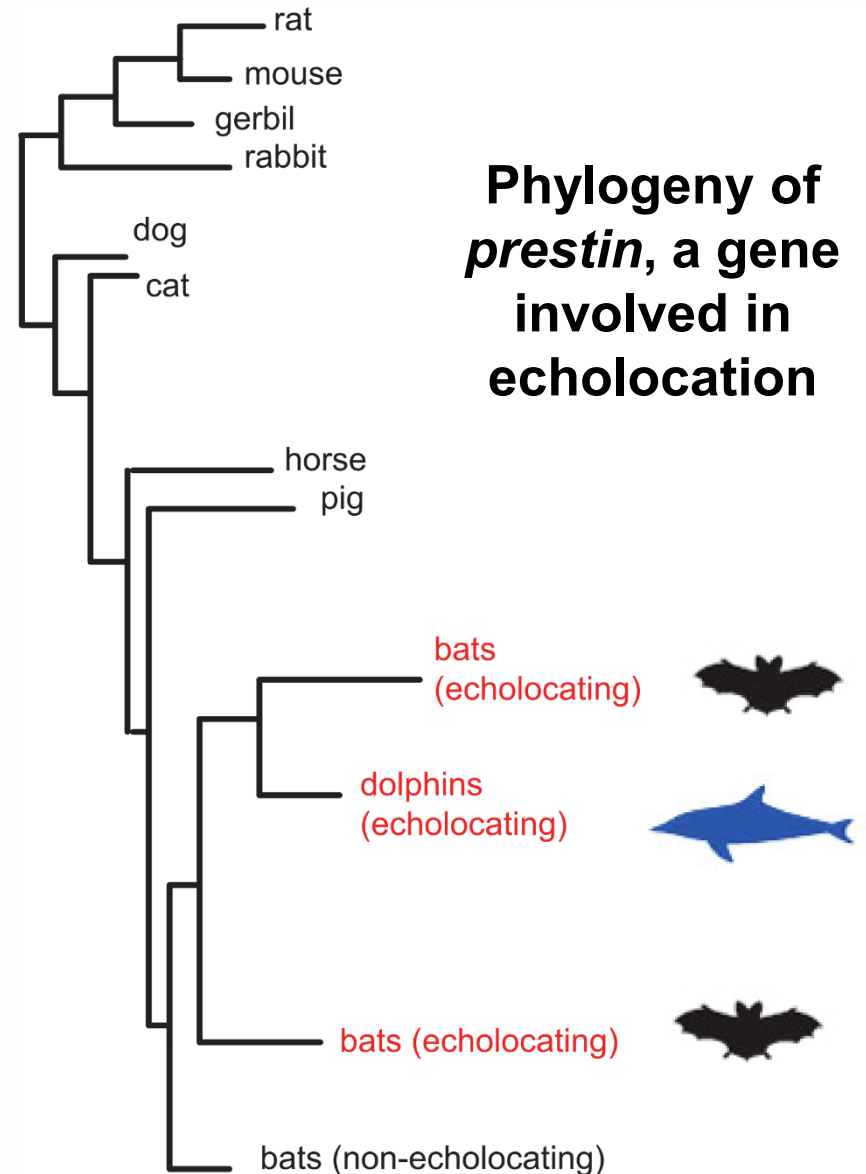
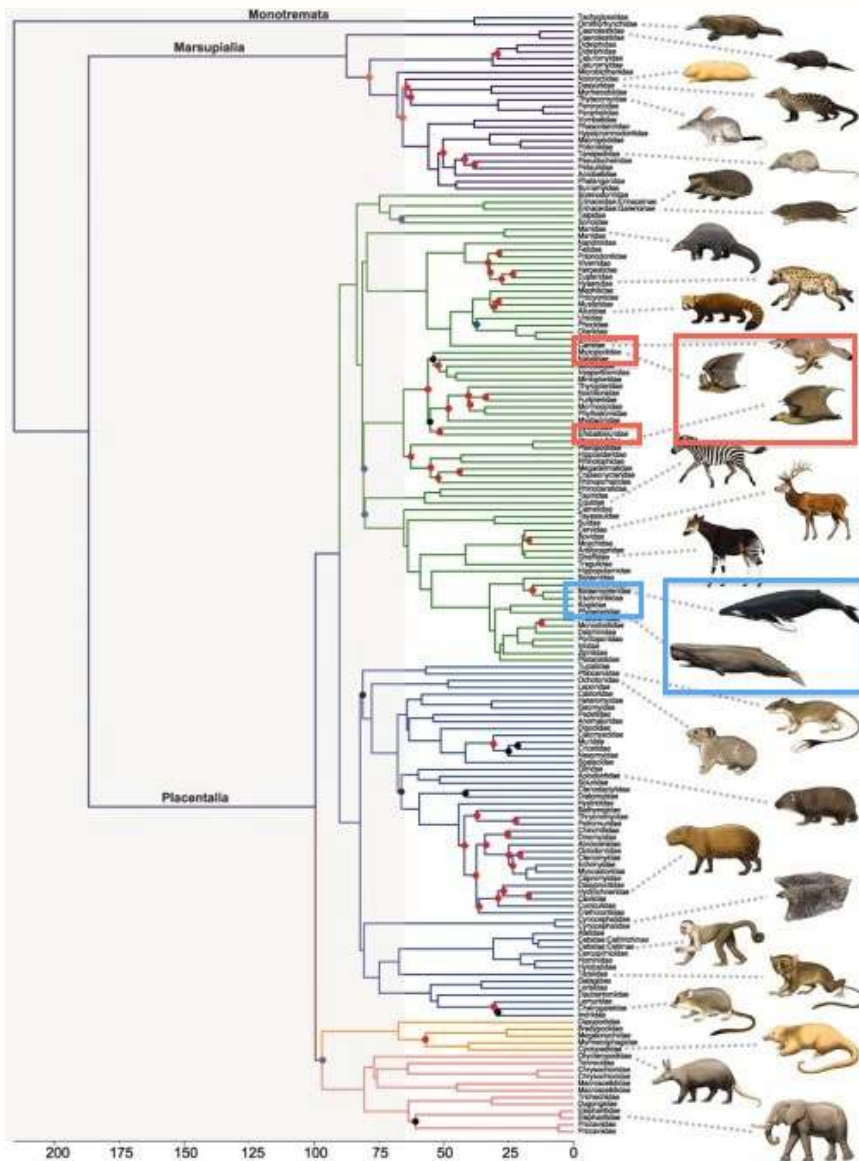
Balancing selection can maintain “trans-species polymorphisms”, in which the alleles are more ancient than the species

Best example: alleles at loci of the MHC – they have been retained by selection because they confer resistance to infection

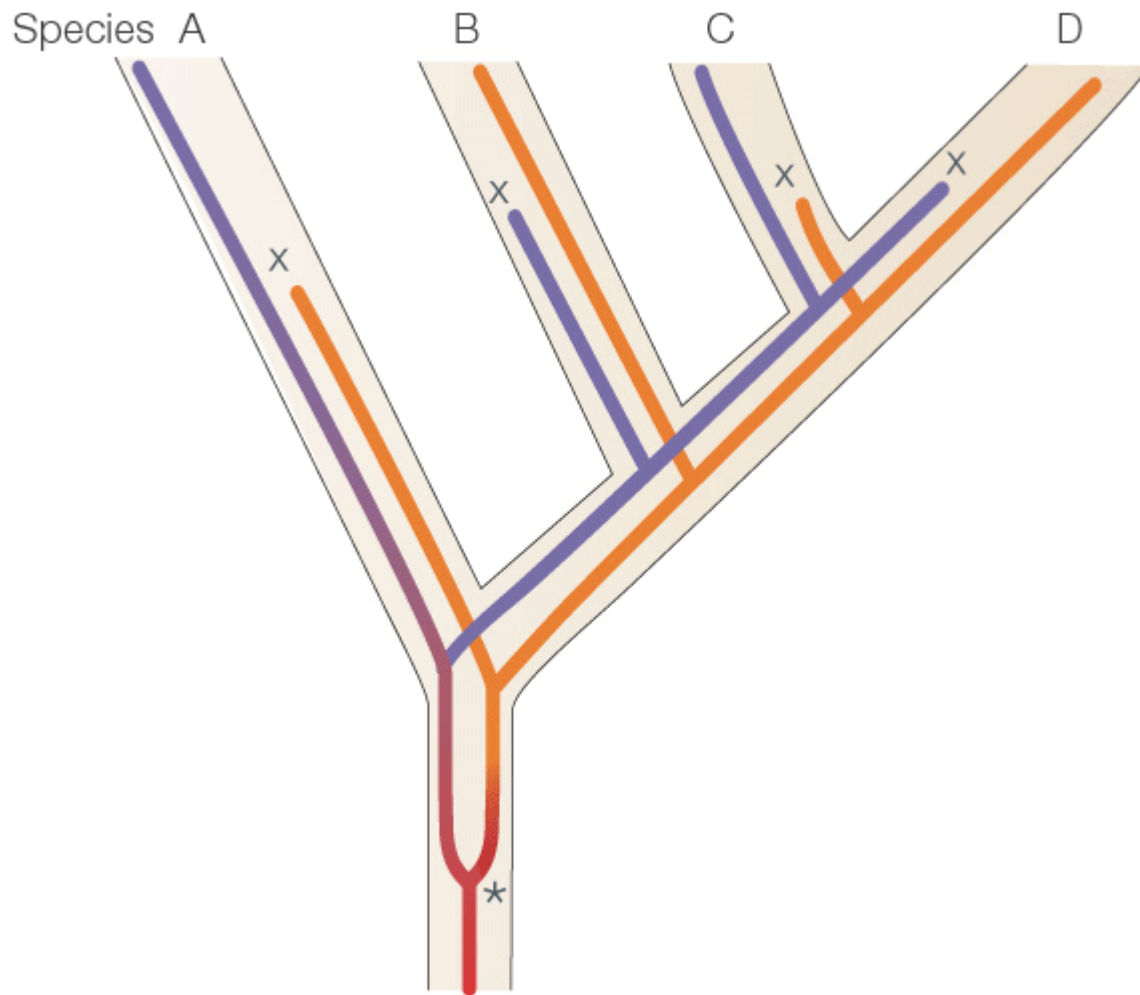
Certain human MHC alleles appear to have diverged more than 65 million years ago (these alleles witnessed the extinction of dinosaurs!!!)



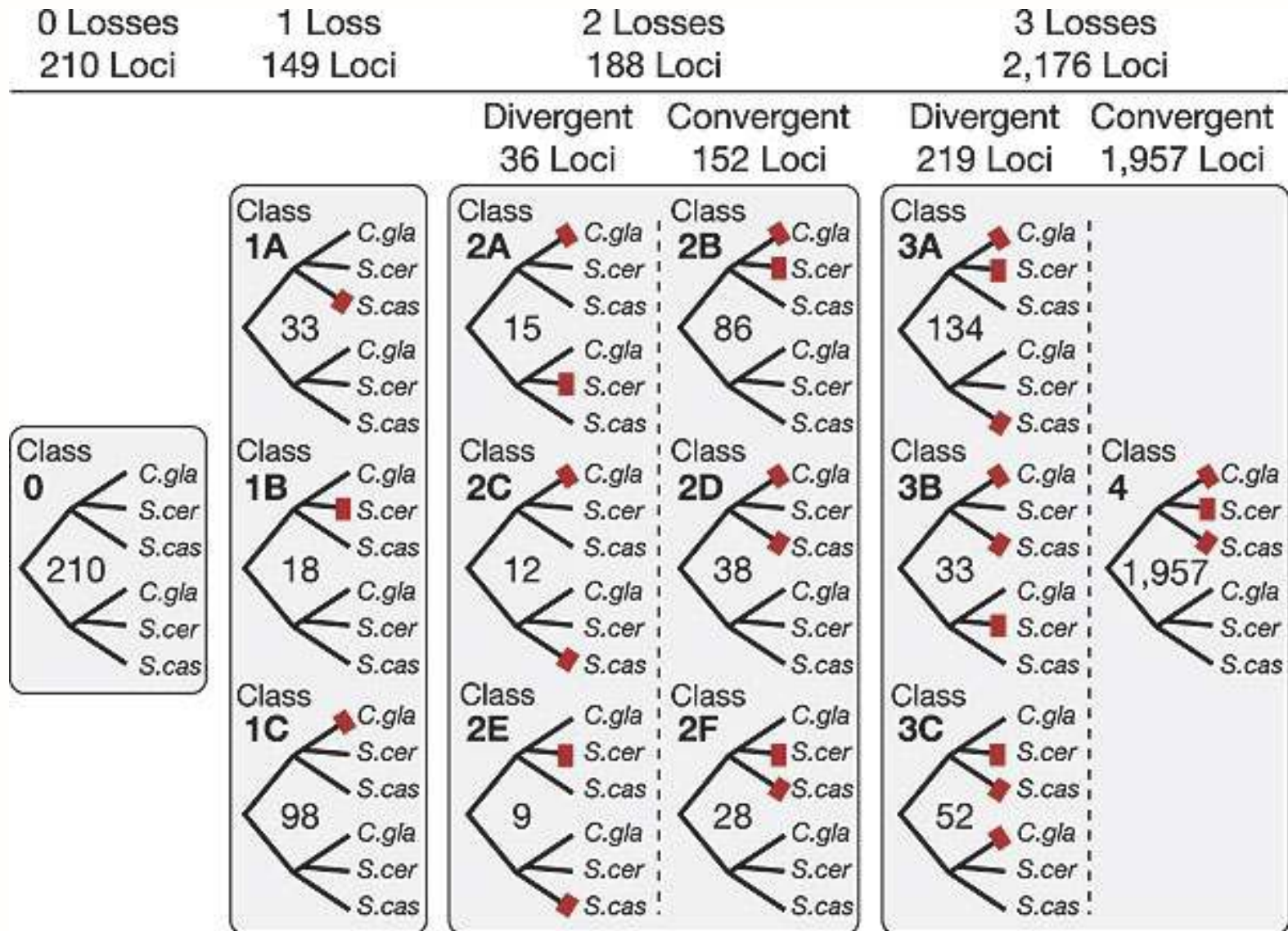
Positive selection



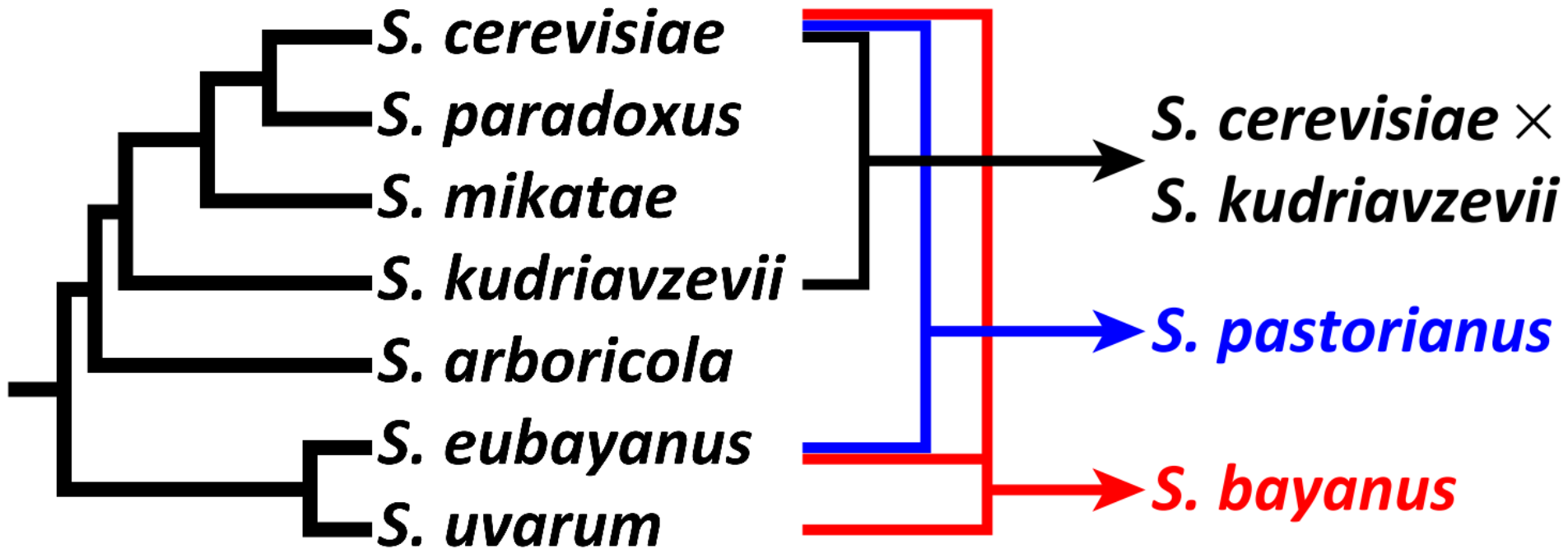
Gene duplication and loss



Gene duplication and loss



Hybridization and introgression



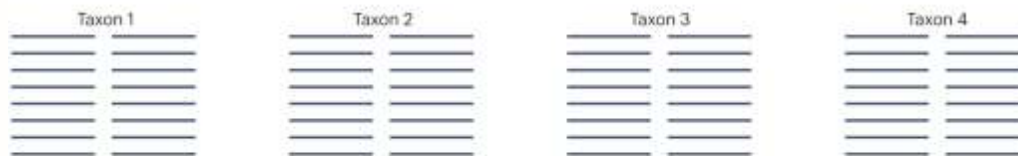
S. eubayanus was discovered in 2011 – until then, *S. bayanus* was thought to be a “pure” species

S. cerevisiae – *S. paradoxus* divergence $\approx$ human – mouse divergence
S. cerevisiae – *S. uvarum* divergence $\approx$ human – chicken divergence



Analytical factors

a Taxon selection



Contributor of incongruence

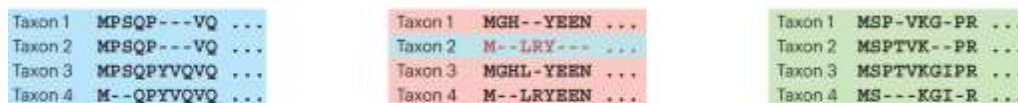
- Insufficient taxon sampling
- Insufficient locus sampling
- Fast-evolving lineages
- Rogue taxa
- Outgroup choice

b Orthology inference



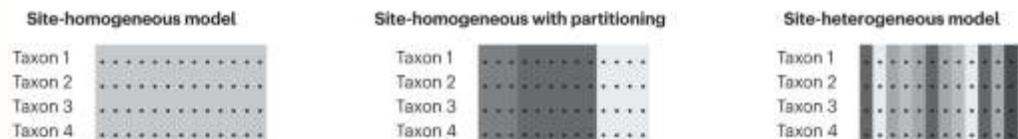
- Sequence length biases
- Erroneous orthologue inference (hidden paralogy and orthology)

c Alignment and site trimming



- Misalignment
- Excessive trimming
- Inappropriate recoding

d Selection of substitution model



- Long-branch attraction
- Model misspecification
- Inadequate model complexity

e Method of tree inference

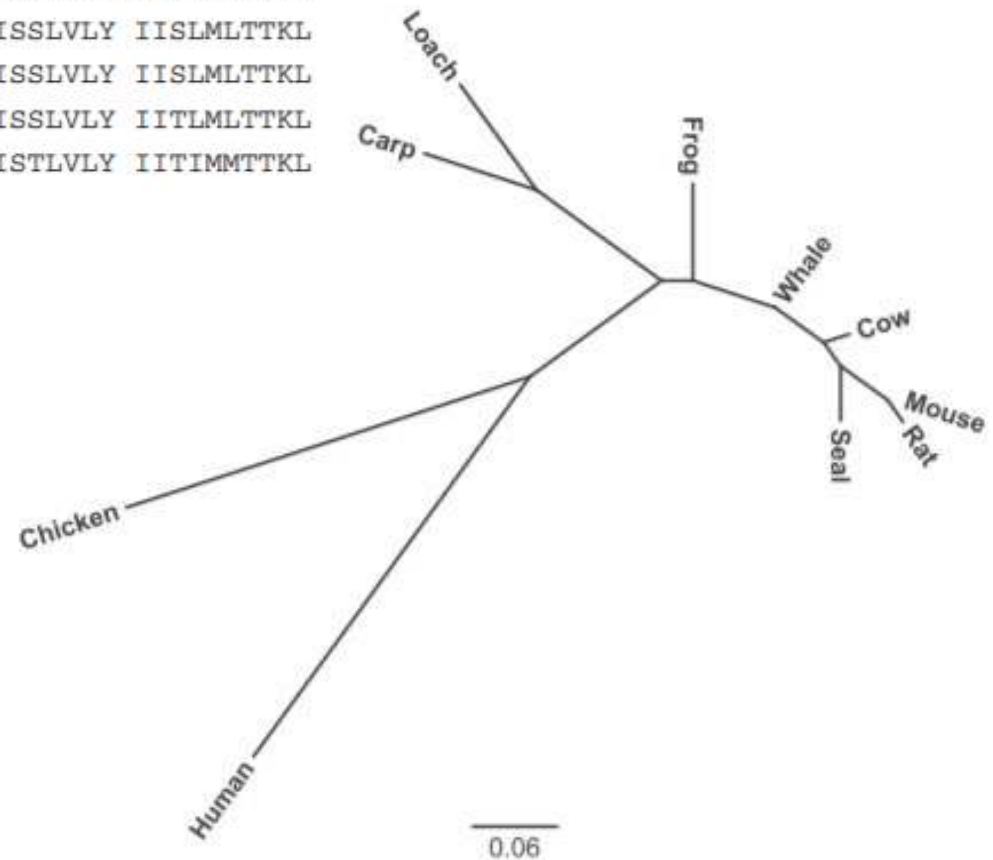


- Irreproducibility
- Single-locus accuracy

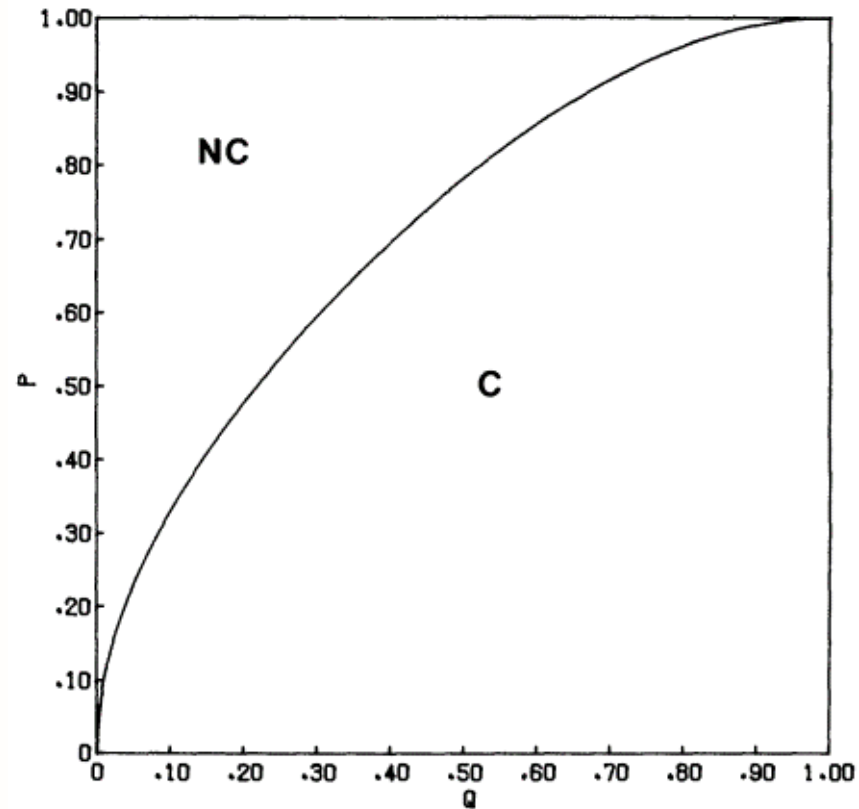
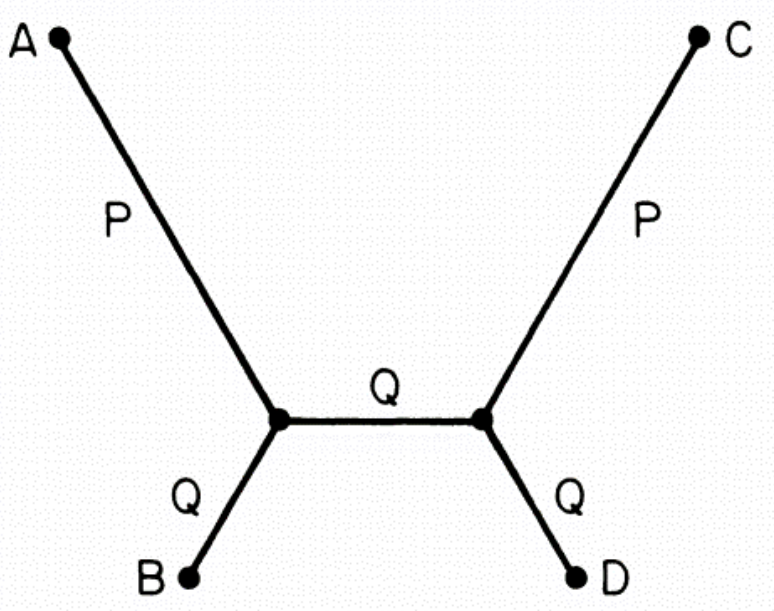


Sampling error

10 50					
Cow	MAYPMQLGFQ	DATSPIMEEL	LHFHDHTLMI	VFLISSLVLY	IISLMLTTKL
Carp	MAHPTQLGFK	DAAMPVMEEL	LHFHDHALMI	VLLISTLVLY	IITAMVSTKL
Chicken	MANHSQLGFQ	DASSPIMEEL	VEFHDHALMV	ALAICSLVLY	LLTLMLMEKL
Human	MAHAAQVGLQ	DATSPIMEEL	ITFHDHALMI	IFLICFLVLY	ALFLTTLTTKL
Loach	MAHPTQLGFQ	DAASPVMEEL	LHFHDHALMI	VFLISALVLY	VIITTVSTKL
Mouse	MAYPFQLGLQ	DATSPIMEEL	MNFHDHTLMI	VFLISSLVLY	IISLMLTTKL
Rat	MAYPFQLGLQ	DATSPIMEEL	TNFHDHTLMI	VFLISSLVLY	IISLMLTTKL
Seal	MAYPLQMGLQ	DATSPIMEEL	LHFHDHTLMI	VFLISSLVLY	IISLMLTTKL
Whale	MAYPFQLGFQ	DAASPIMEEL	LHFHDHTLMI	VFLISSLVLY	IITLMLTTKL
Frog	MAHPSQLGFQ	DAASPIMEEL	LHFHDHTLMA	VFLISTLVLY	IITIMMTTKL



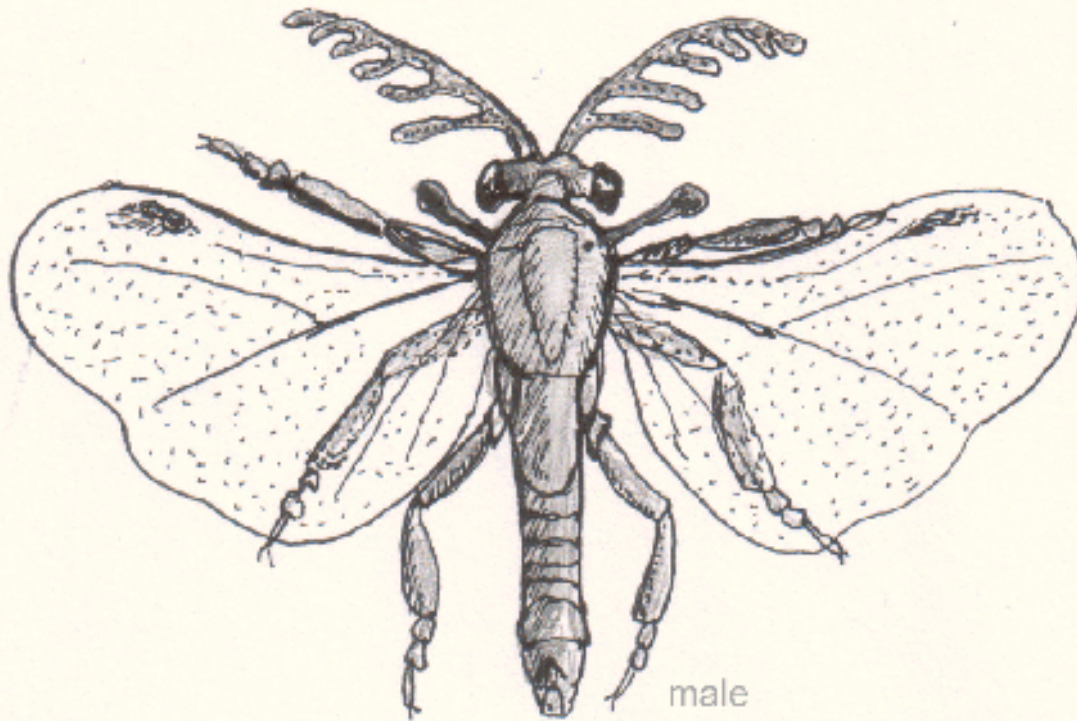
Systematic error



Long branch attraction



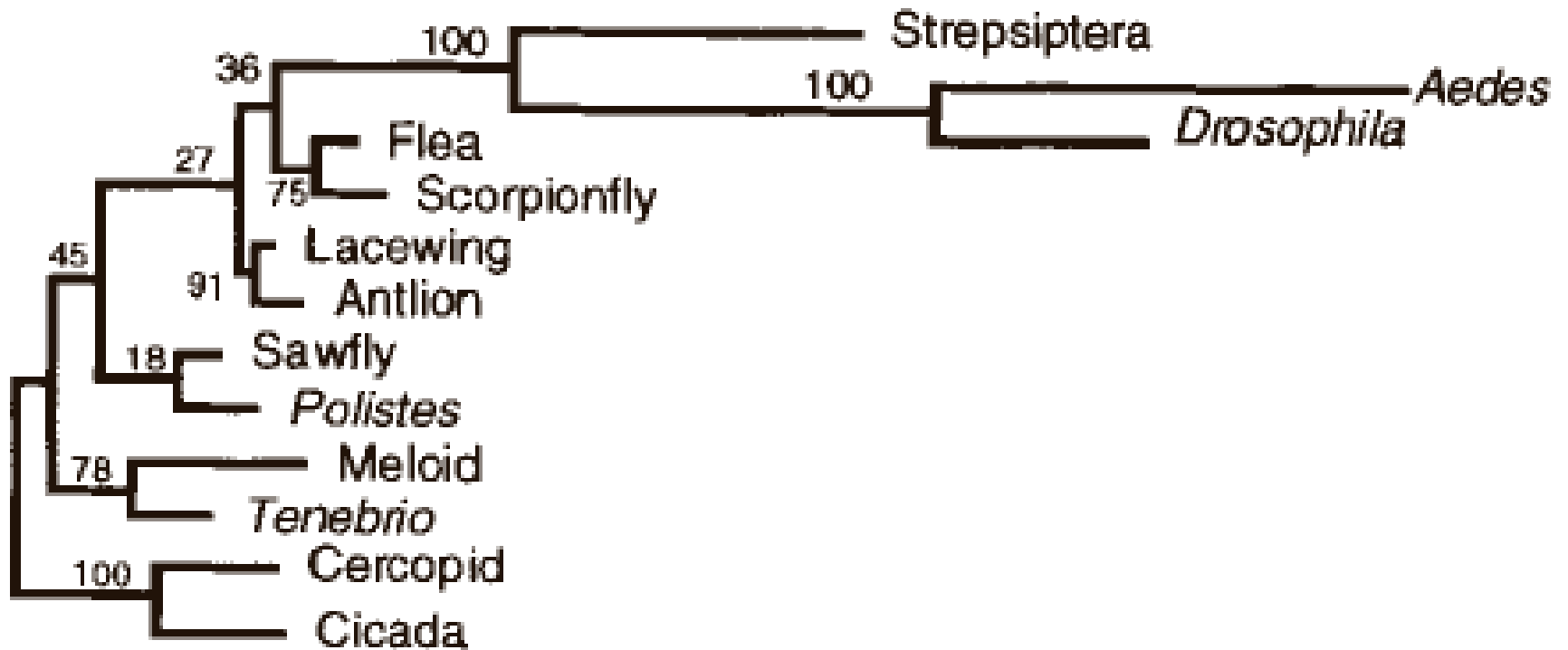
“The Strepsiptera Problem” is a classic example of LBA



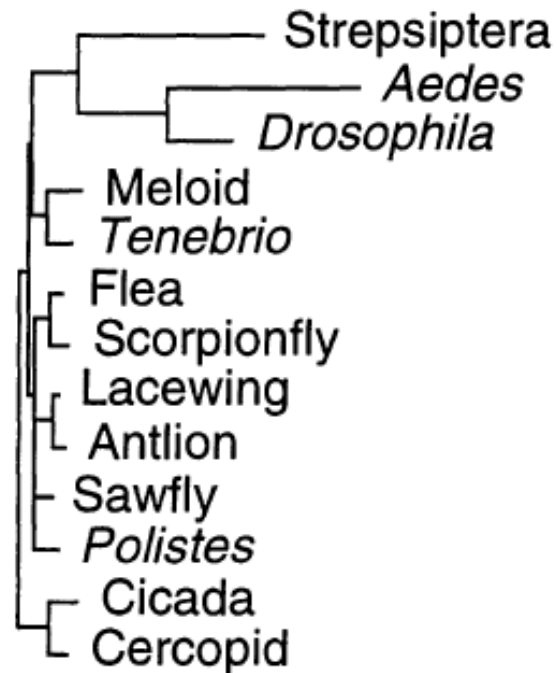
Halictophagidae (Strepsiptera)



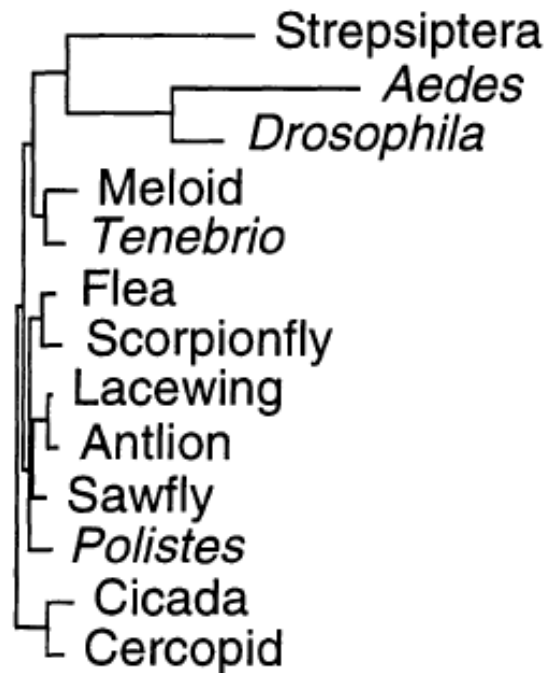
The Strepsiptera Problem



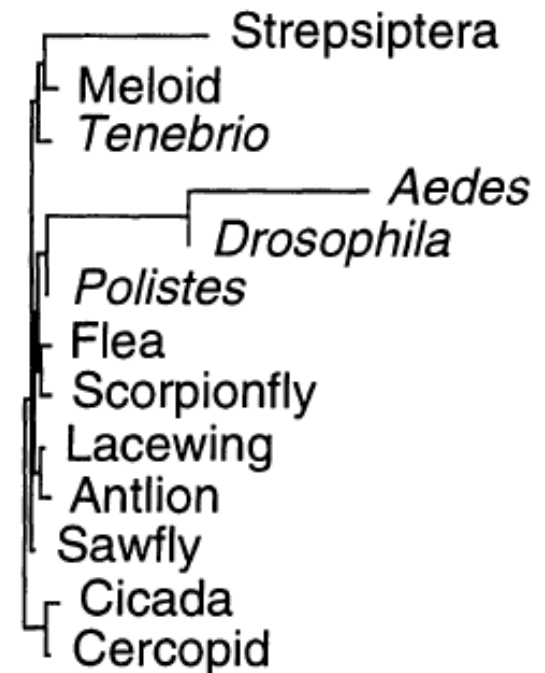
The Strepsiptera Problem



p distance



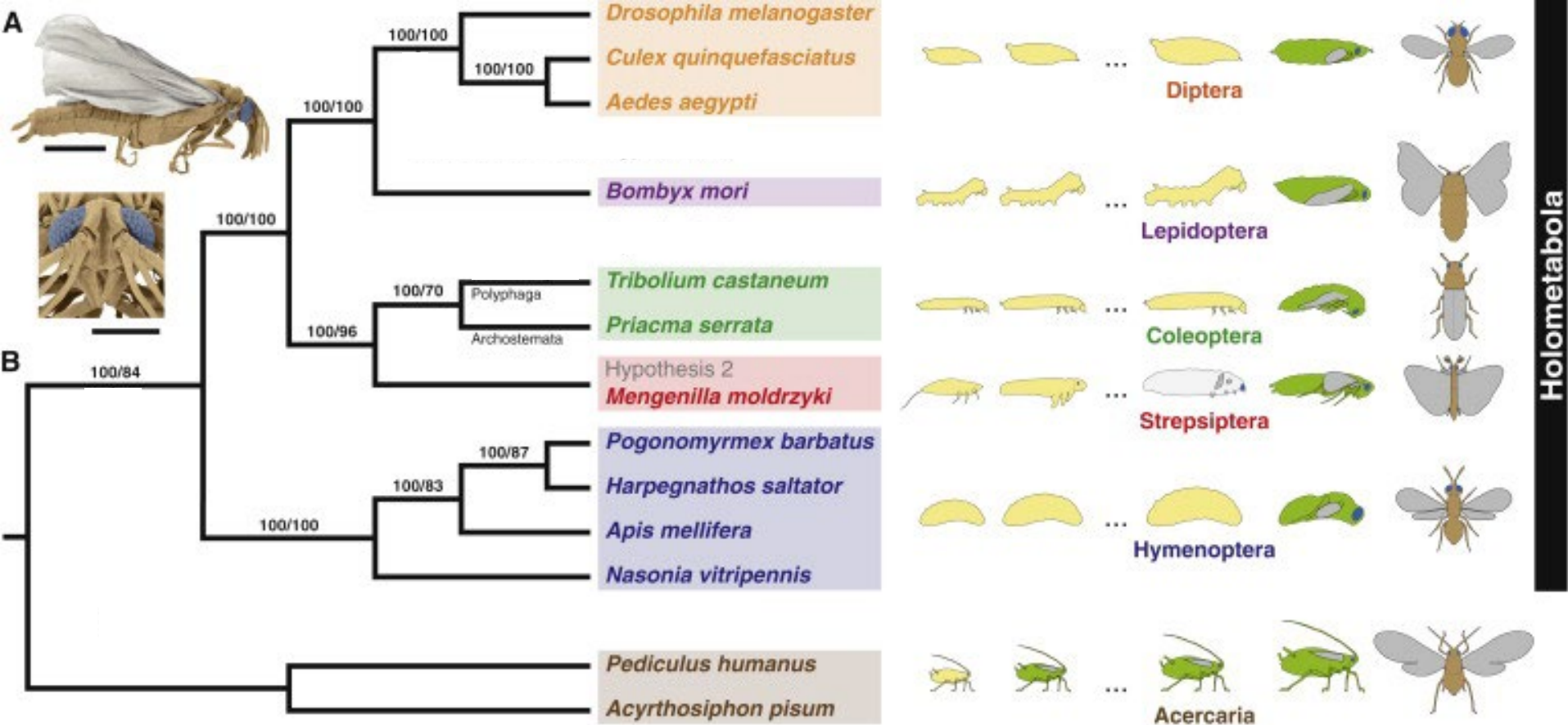
HKY85



HKY85+GAMMA



Solving the Strepsiptera Problem with more genes and better models



Multiple sequence alignment

Alignment trimming

Character recoding

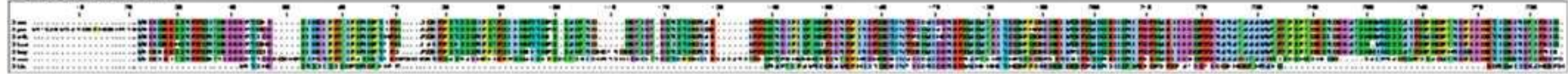
Irreproducibility

...

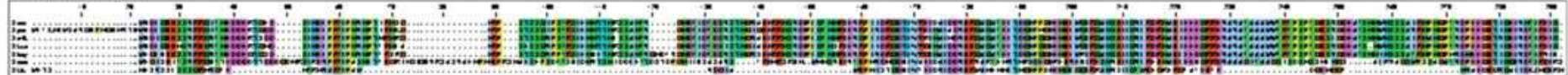


Multiple sequence alignment

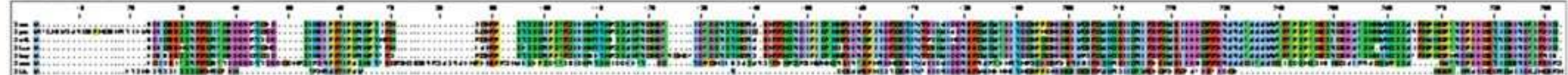
CLUSTAL W



MUSCLE



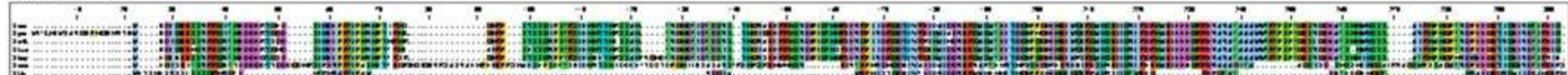
T-COFFEE



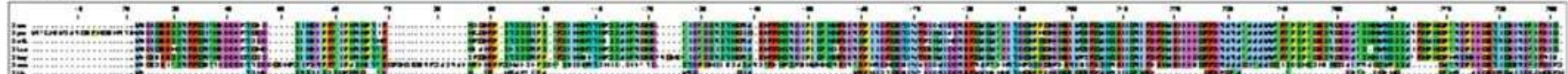
DIALIGN 2



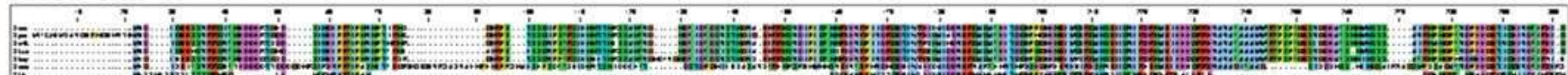
MAFFT



DCA

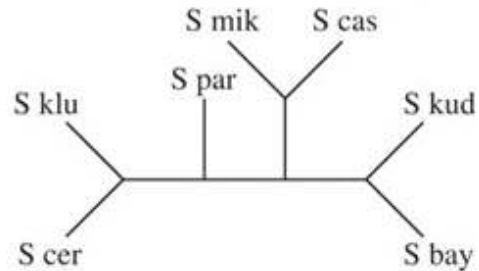


PROBCONS

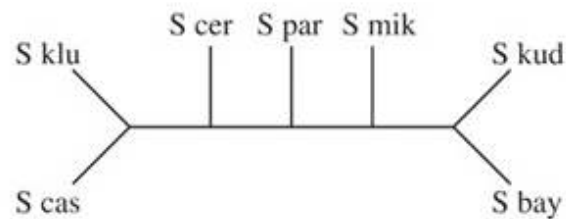


7 tools produce 6 different topologies

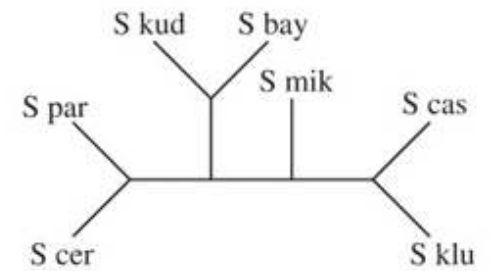
CLUSTAL/DIALIGN (0.24)



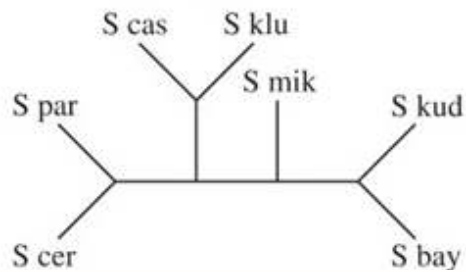
MUSCLE (0.25)



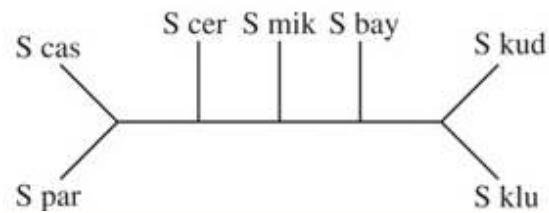
T-COFFEE (0.30)



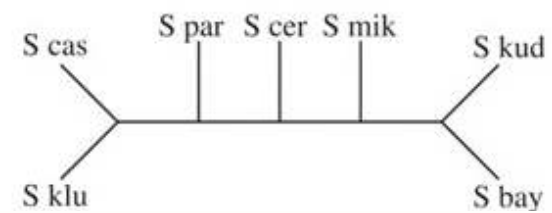
MAFFT (0.18)



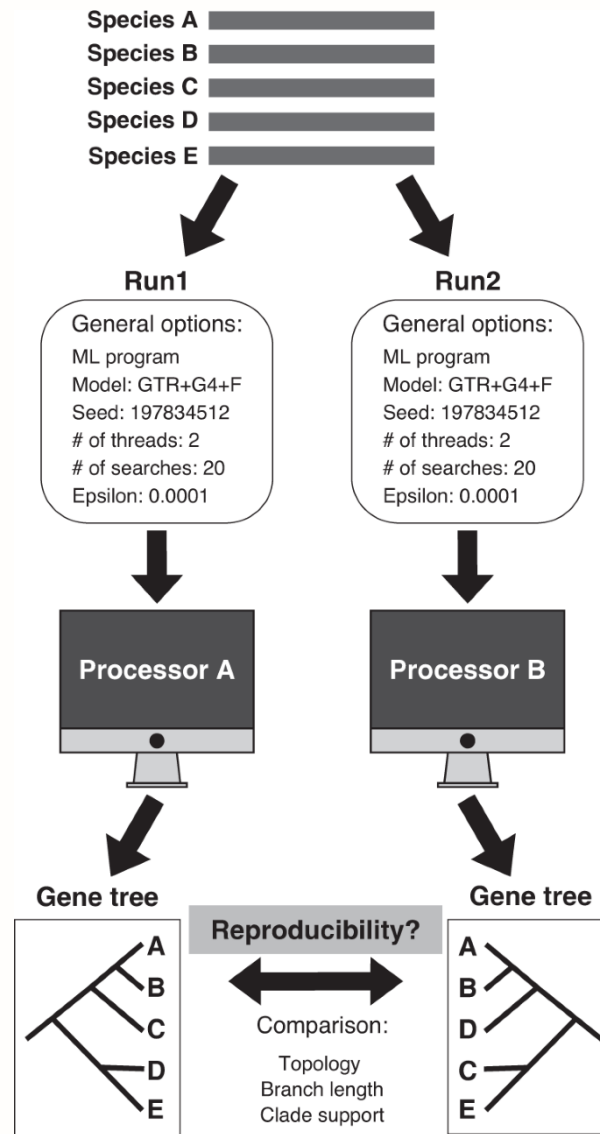
DCA (0.12)



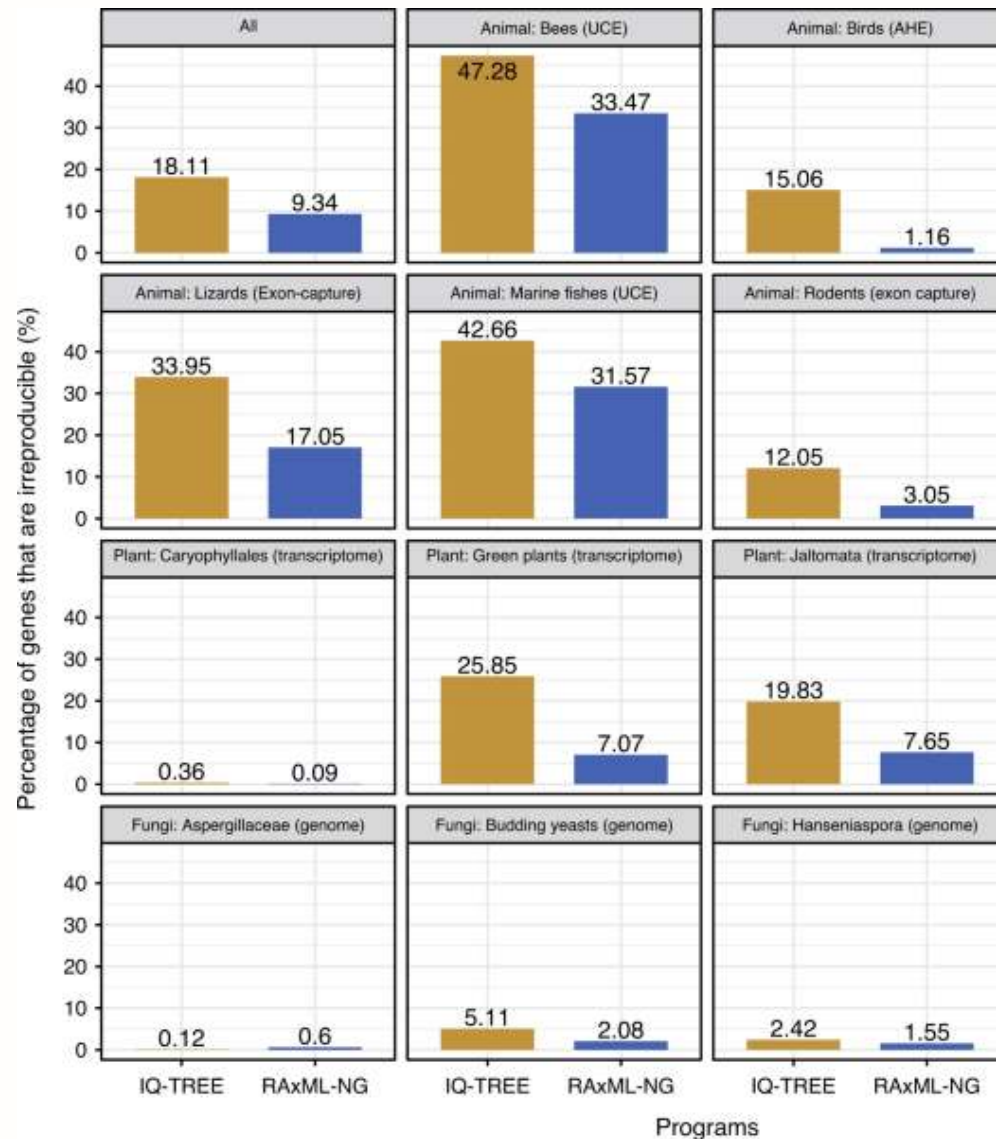
PROBCONS (0.05)



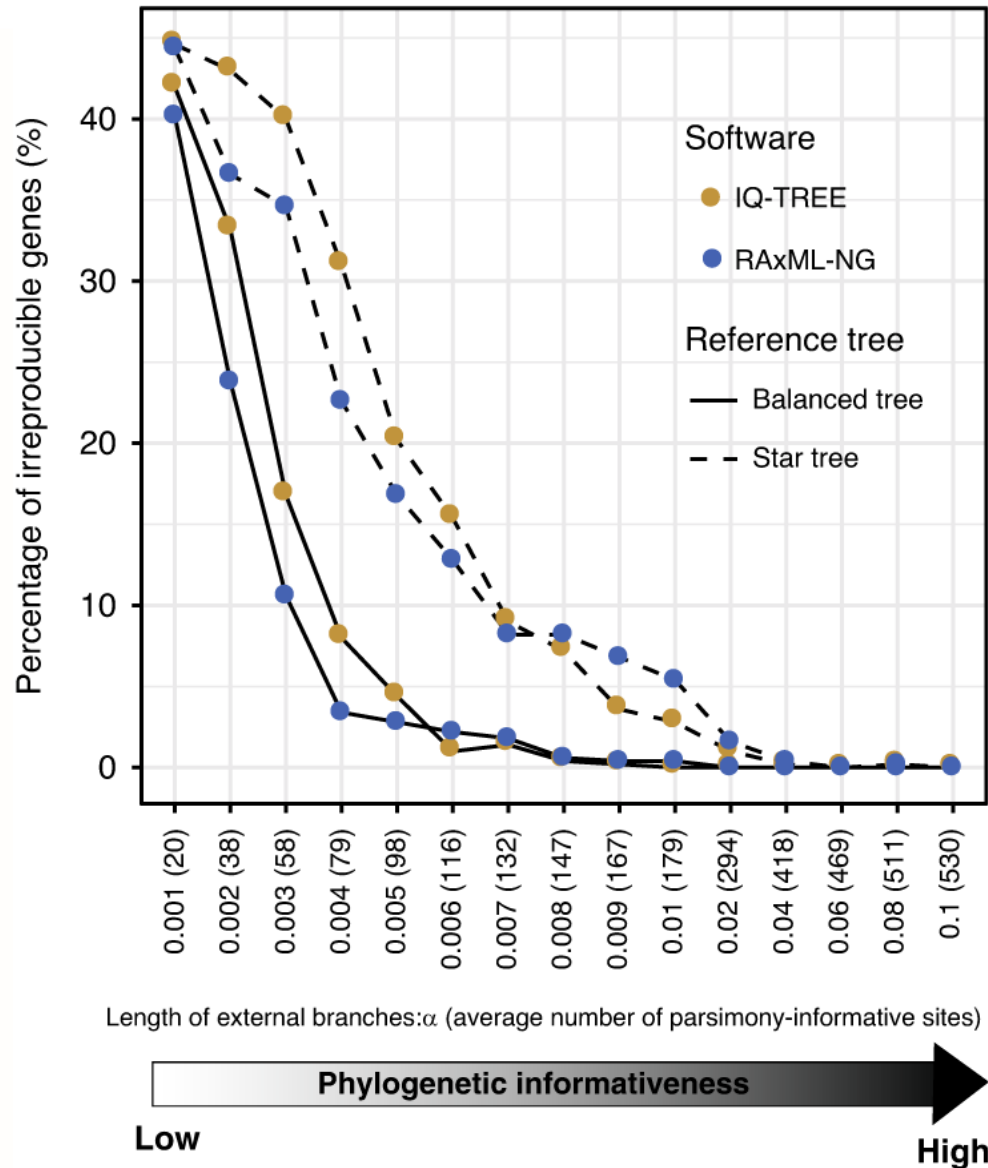
What is the reproducibility of our analyses?



9-18% of gene trees are irreproducible



Genes yielding irreproducible trees are less informative



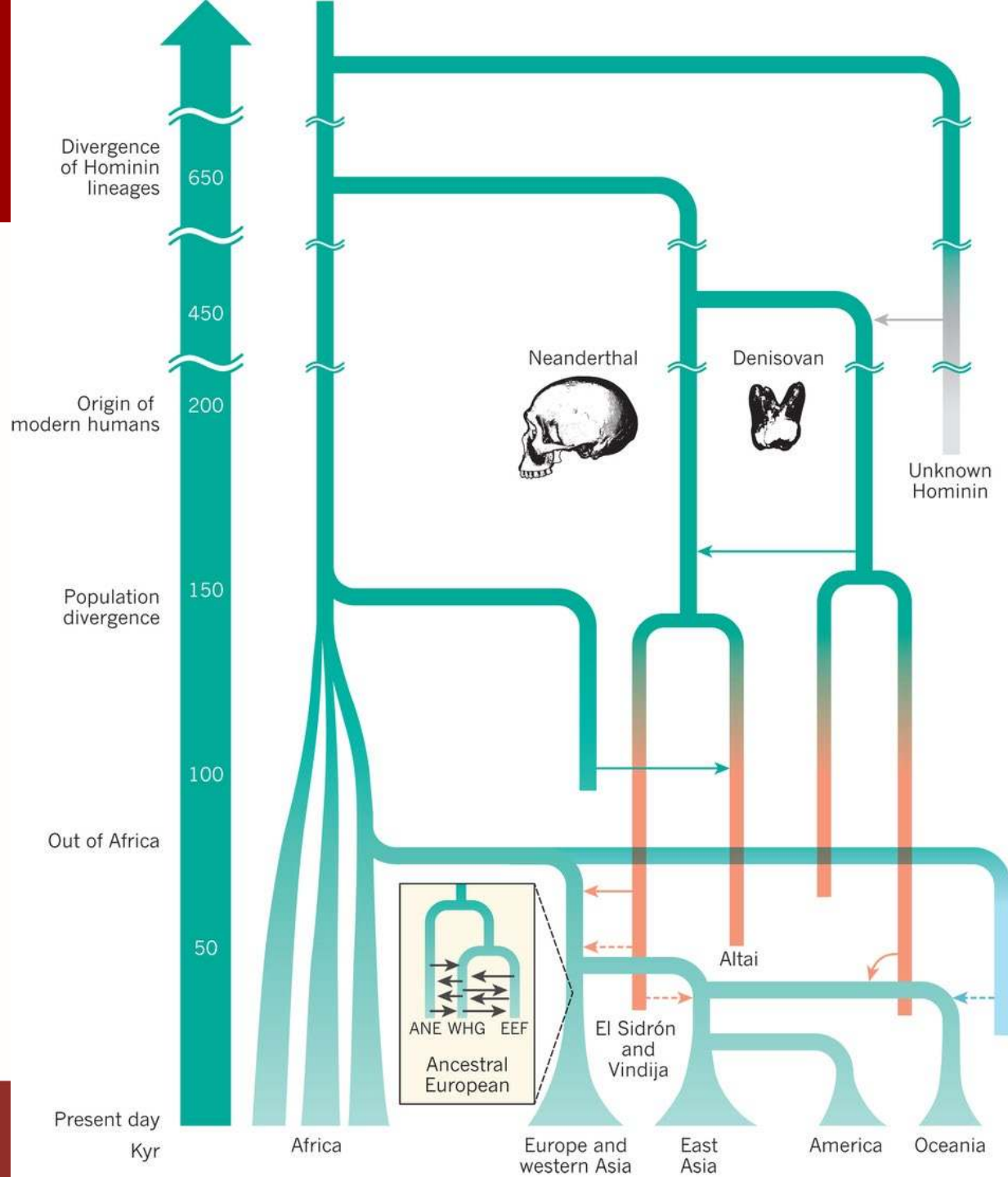
**We have not eliminated
incongruence but now better
understand the biological and
analytical contributing factors and
their impact; sources vary by lineage
and depth**

Inference at shallow depths is easier:

analytical factors

biological factors

The evolution of hominids



The phylogeny of primate genera

*Nomascus
leucogenys*



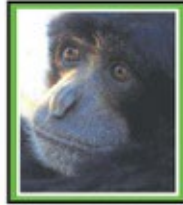
NLE

*Hoolock
leuconedys*



HLE

*Symphalangus
syndactylus*



SSY

*Hylobates
pileatus*



HPI

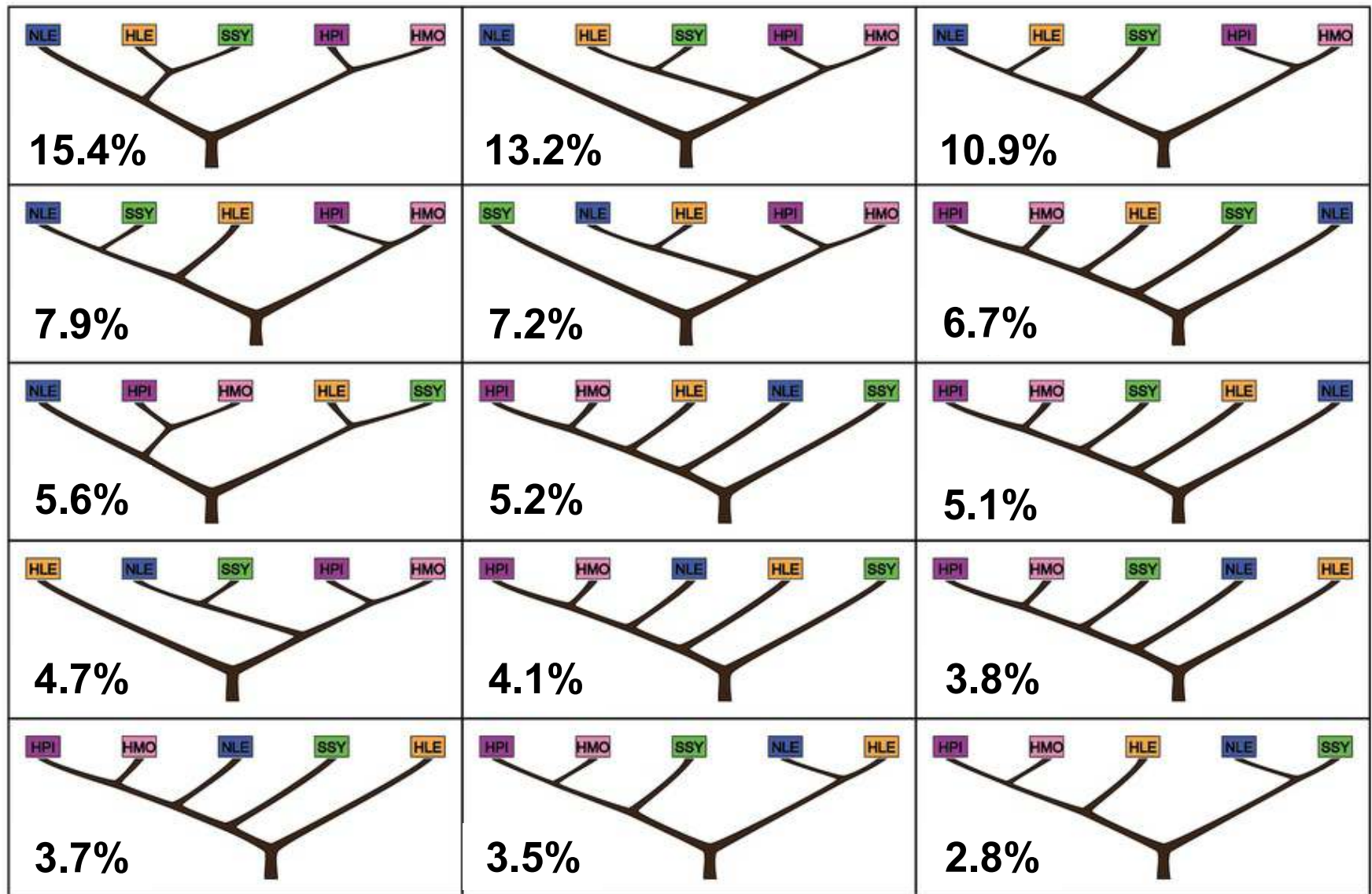
*Hylobates
moloch*



HMO



“Easier” doesn’t mean “easy”!



**Inference in deep time can be
more challenging:**

analytical factors

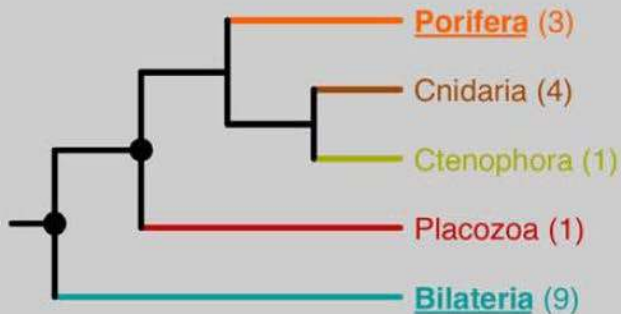
biological factors

Incongruence in deep time

Schierwater et al. (2009)

15 mitochondrial and 34 nuclear genes

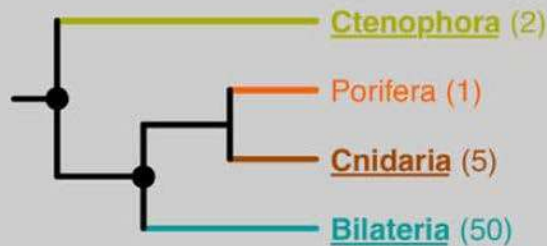
GTR and WAG models



Dunn et al. (2008)

6 mitochondrial and 144 nuclear genes

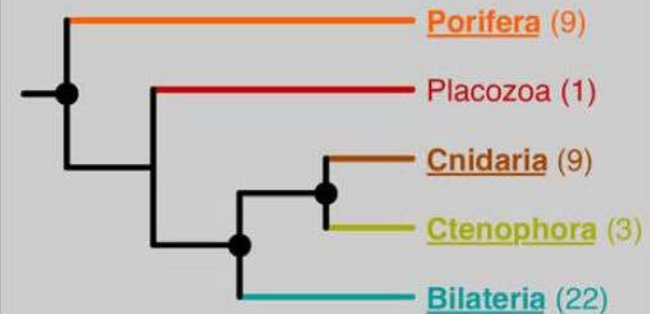
WAG and CAT models



Philippe et al. (2009)

128 nuclear genes

CAT model



Porifera



Placozoa



Cnidaria



Ctenophora



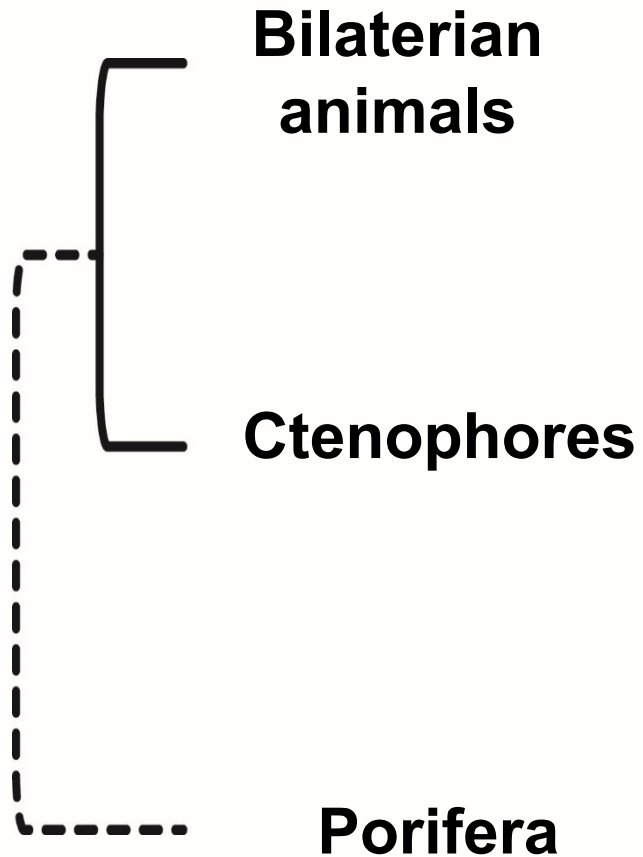
Bilateria



Incongruence in deep time



Incongruence in deep time



❖ **Incongruence and its causes**

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

❖ **Handling incongruence in phylogenomic data**

An expanded yeast data matrix

Yeast Gene Order Browser (YGOB)

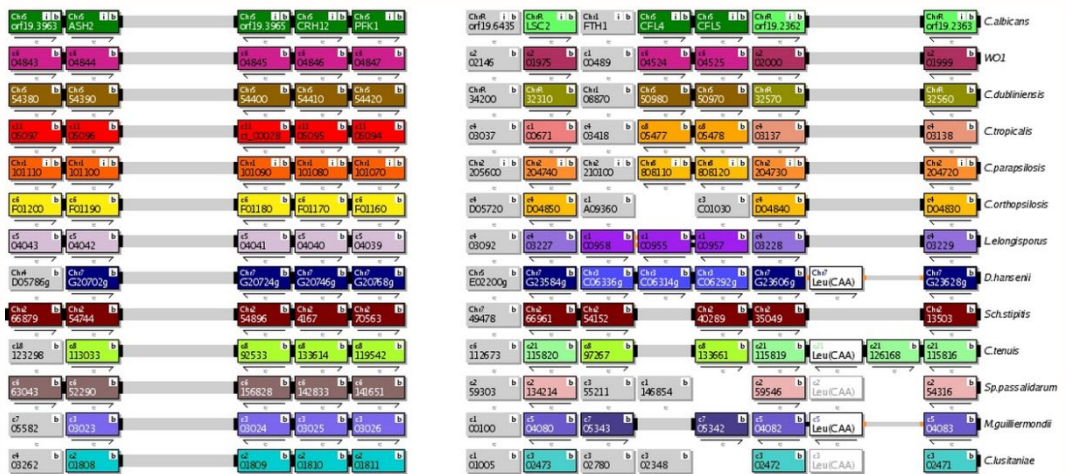


**Saccharomyces
lineage**

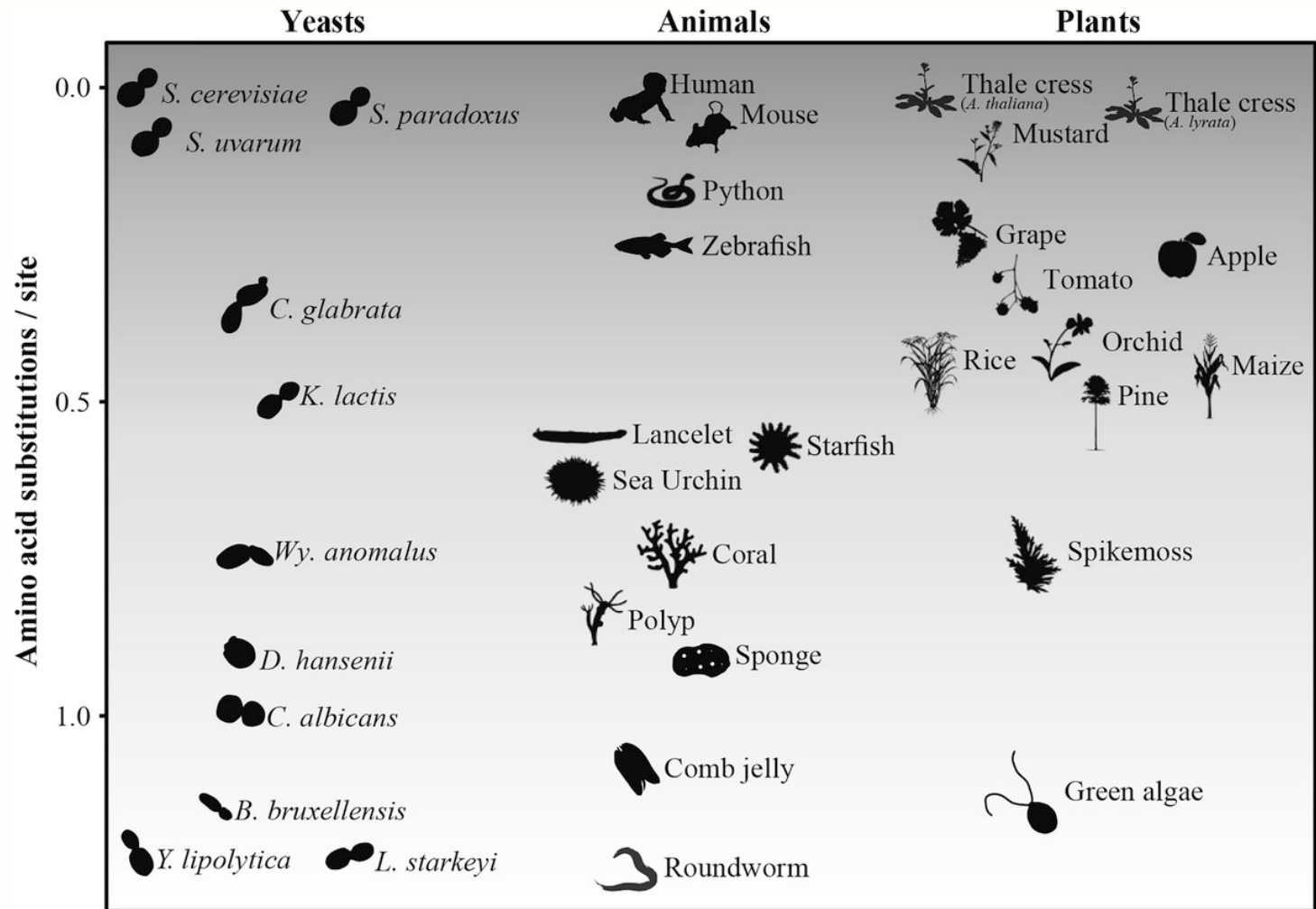
**1,070 genes
23 taxa
no missing data**

**Candida
lineage**

Candida Gene Order Browser (CJOB)



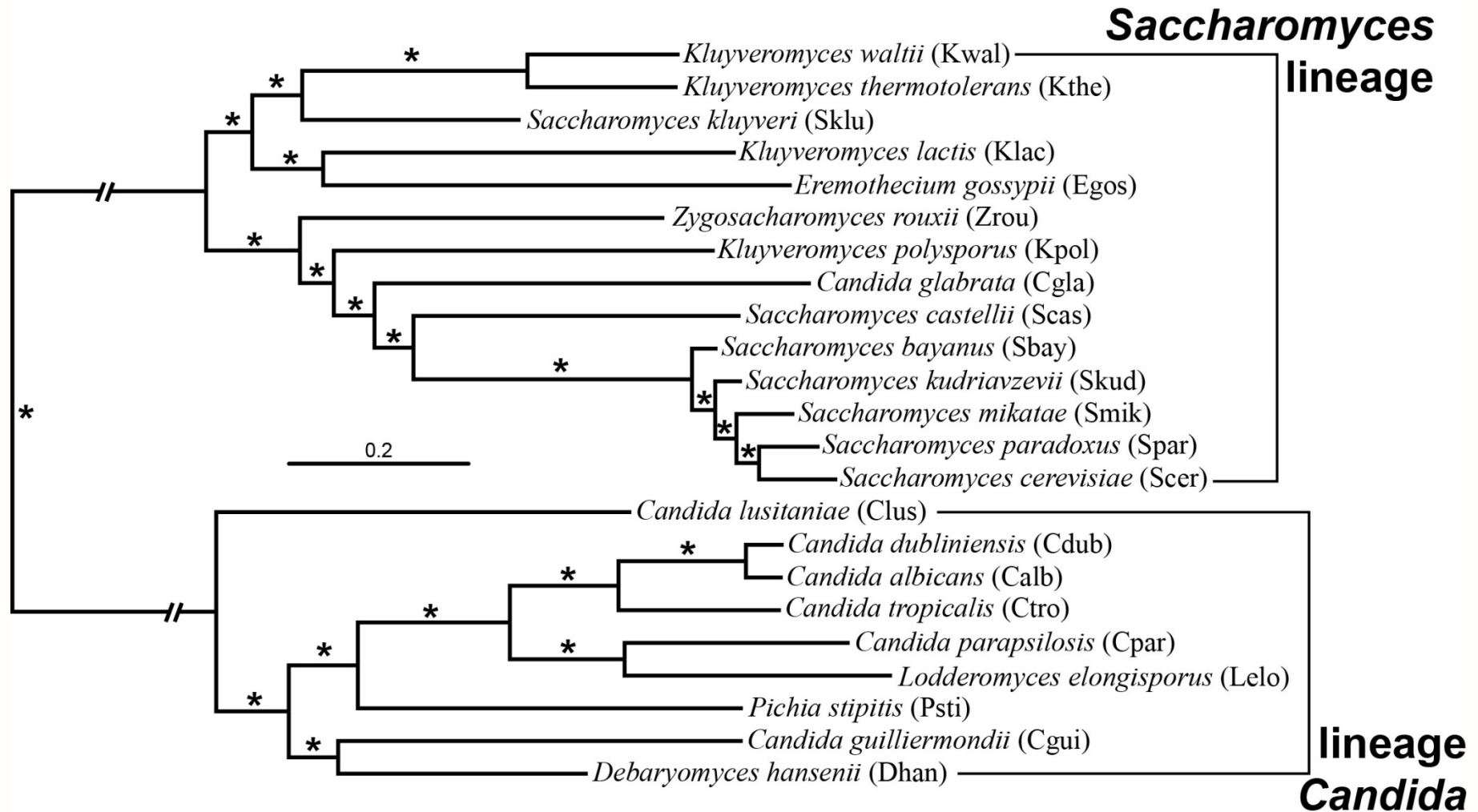
Yeasts exhibit striking genomic diversity



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



Concatenation yields an absolutely supported phylogeny

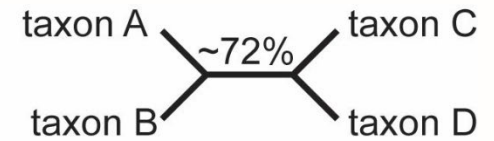


Bootstrap support is misleading when used in large datasets

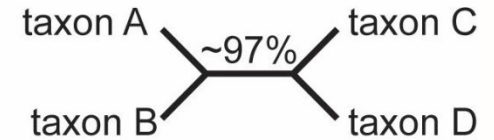
53% 47%

taxonA AAAAAAAAAATTTTTTTT
taxonB AAAAAAAAAACCCCCCCC
taxonC GGGGGGGGGTTTTTTTTT
taxonD GGGGGGGGGCCCCCCCCC

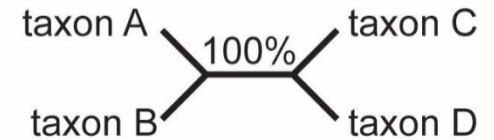
100 characters



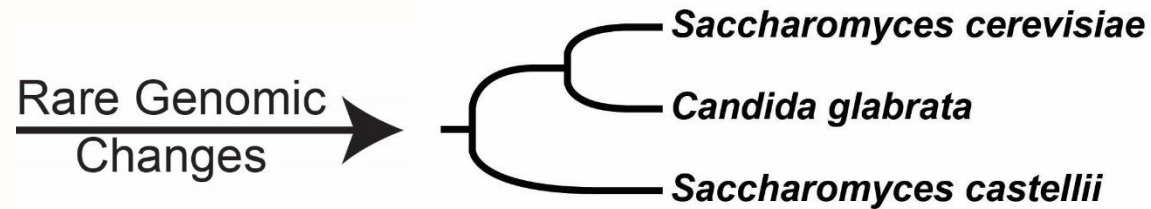
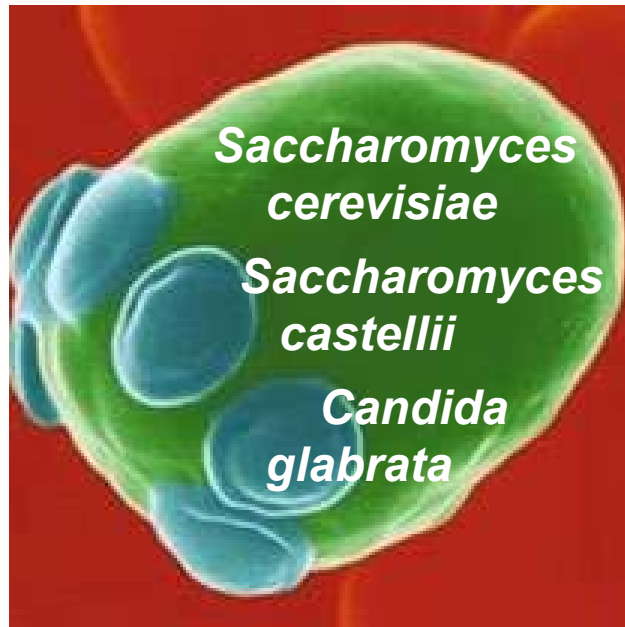
1,000 characters



10,000 characters



The concatenation phylogeny is at least partly wrong

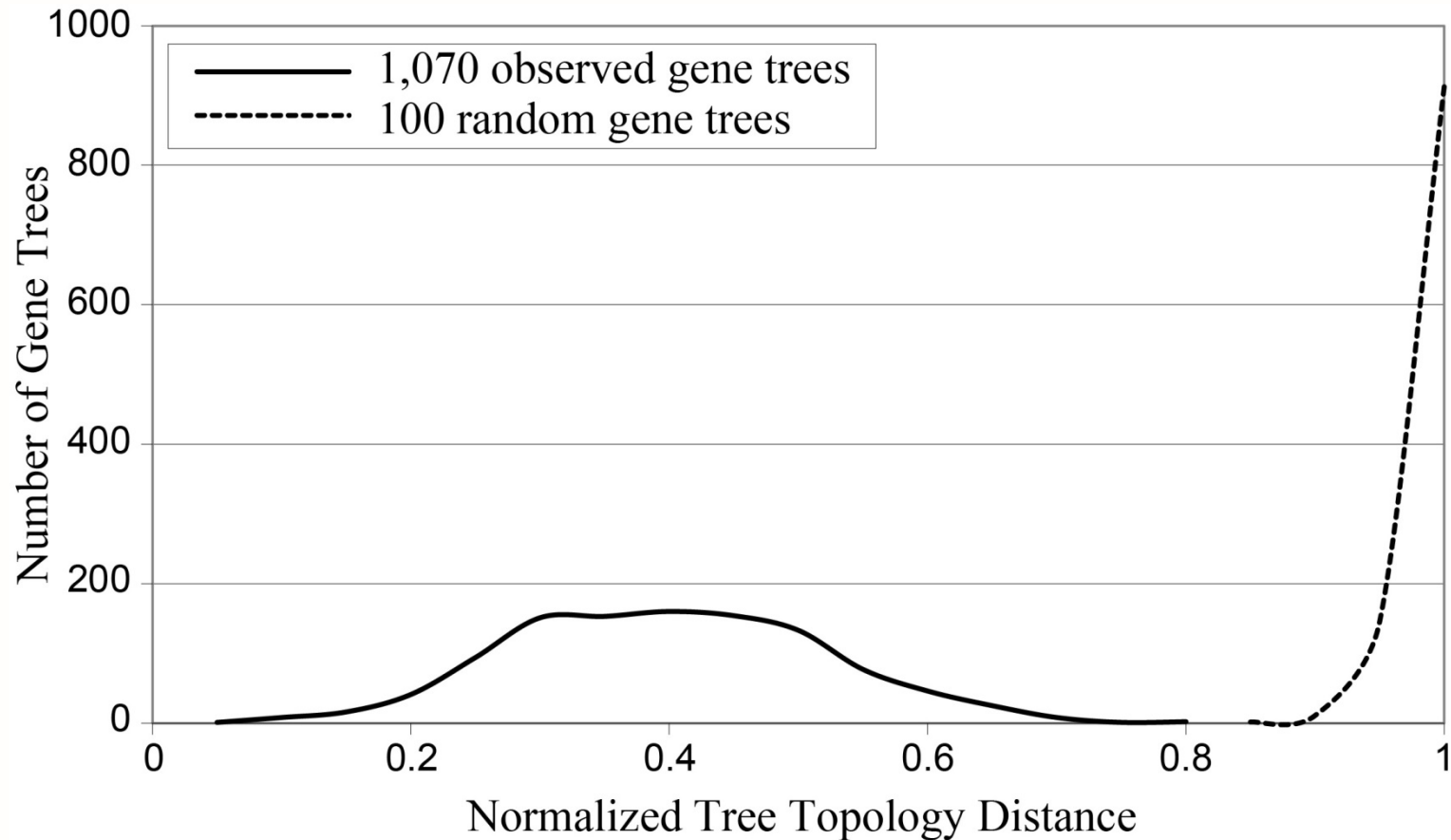


- ❖ 5 genomic rearrangements that are uniquely shared by *S. cerevisiae* and *C. glabrata*
- ❖ Much higher proportion of shared gene losses in *S. cerevisiae* and *C. glabrata*
- ❖ Bias in the placement of *C. glabrata* as an outgroup of *S. cerevisiae* and *S. castellii*

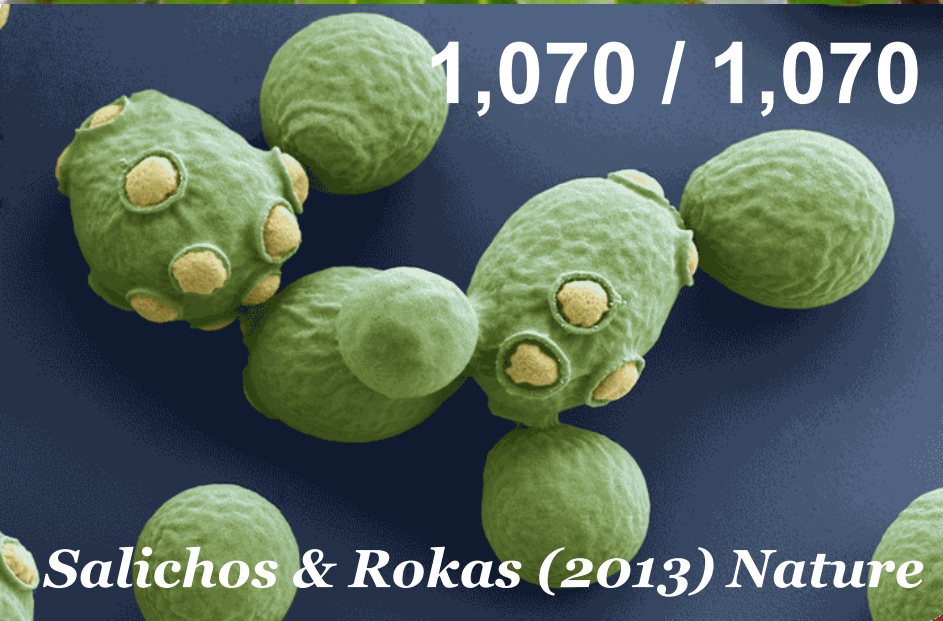
Linear Sequence Data



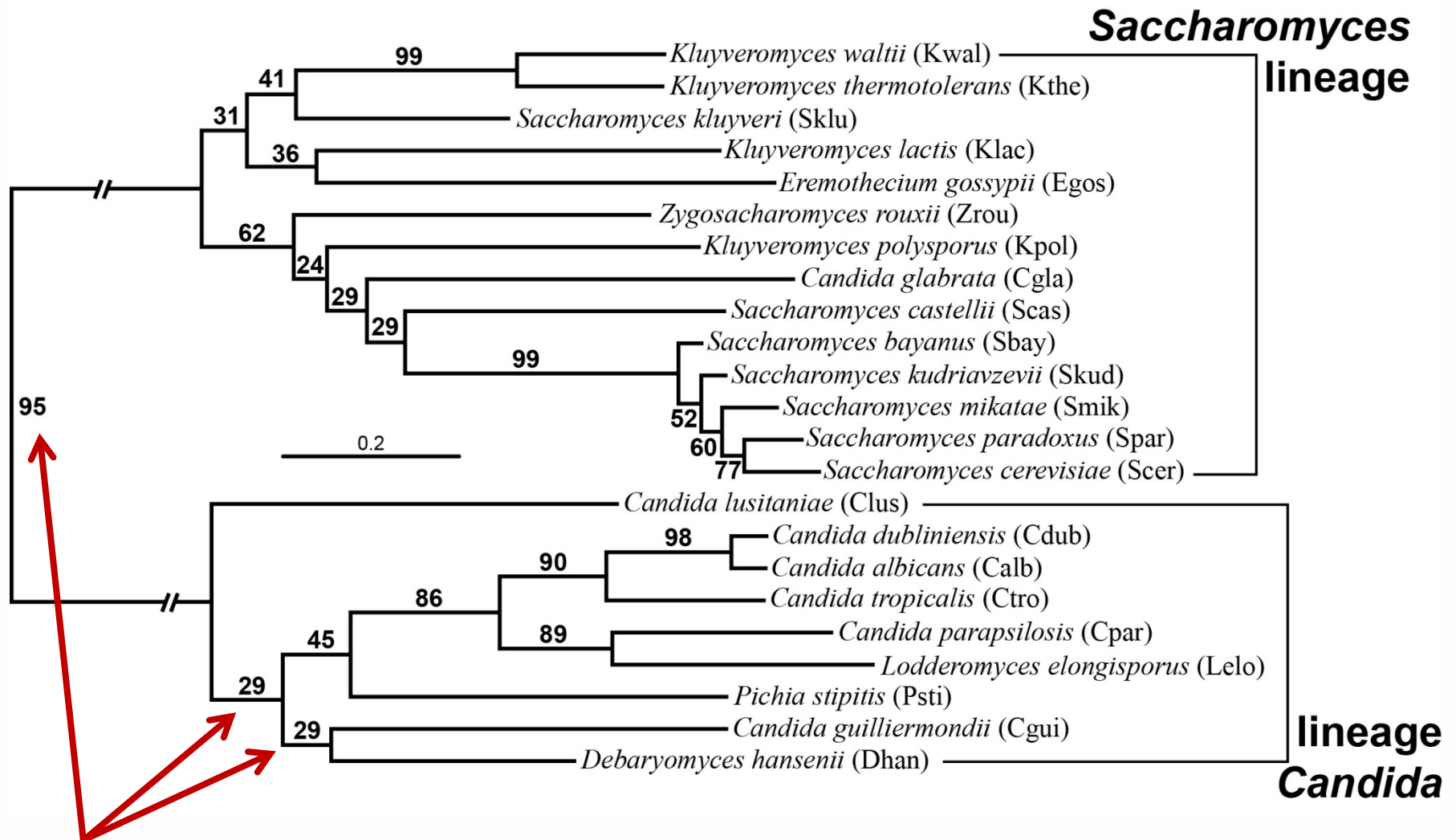
All gene trees differ from the concatenation phylogeny



Gene trees are incongruent in most datasets



The yeast phylogeny inferred by majority-rule consensus



New Methods to Calculate Concordance Factors for Phylogenomic Datasets

Bui Quang Minh ^{1,2}, Matthew W. Hahn,^{3,4} and Robert Lanfear^{*,2}

¹Research School of Computer Science, Australian National University, Canberra, ACT, Australia

²Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, ACT, Australia

³Department of Biology, Indiana University, Bloomington, IN

⁴Department of Computer Science, Indiana University, Bloomington, IN

Correspondence to: *Corresponding author: E-mail: rob.lanfear@anu.edu.au.

Associate editor: Michael Rosenberg

Abstract

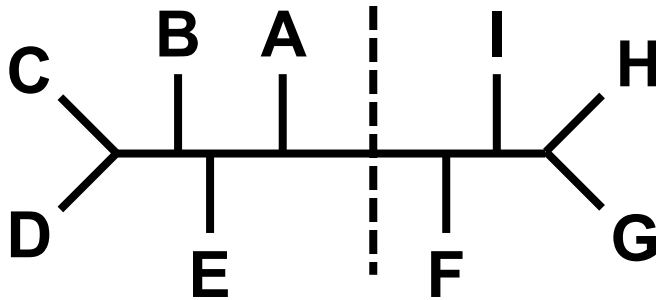
We implement two measures for quantifying genealogical concordance in phylogenomic data sets: the gene concordance factor (gCF) and the novel site concordance factor (sCF). For every branch of a reference tree, gCF is defined as the percentage of “decisive” gene trees containing that branch. This measure is already in wide usage, but here we introduce a package that calculates it while accounting for variable taxon coverage among gene trees. sCF is a new measure defined as the percentage of decisive sites supporting a branch in the reference tree. gCF and sCF complement classical measures of branch support in phylogenetics by providing a full description of underlying disagreement among loci and sites. An easy to use implementation and tutorial is freely available in the IQ-TREE software package (<http://www.iqtree.org/doc/Concordance-Factor>, last accessed May 13, 2020).

Key words: phylogenetic inference, concordance factor, phylogenomics.



Phylogenetic trees are sets of splits / bipartitions

Division

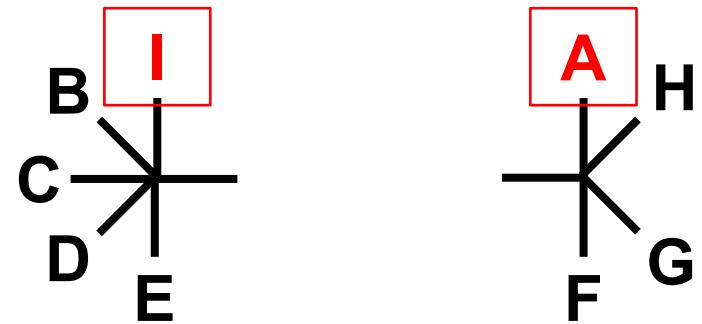
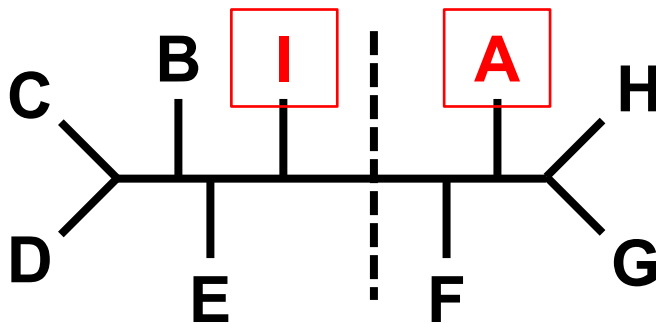


Splits / Bipartitions



Set of splits in reference tree: {A, B, C, D, E}

{F, G, H, I}



Conflicting set of splits: {**I**, B, C, D, E}

{F, G, H, **A**}

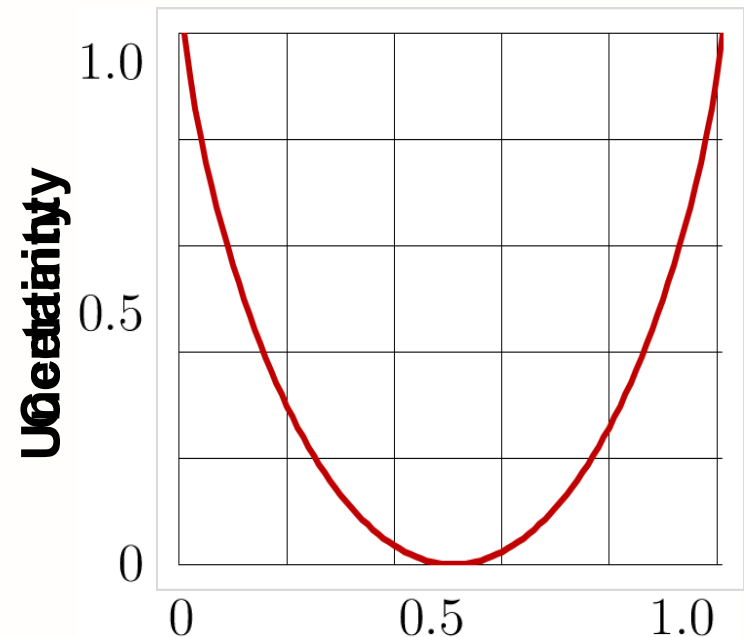
Quantifying incongruence

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

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

Implemented in RAxML and
QuartetScores

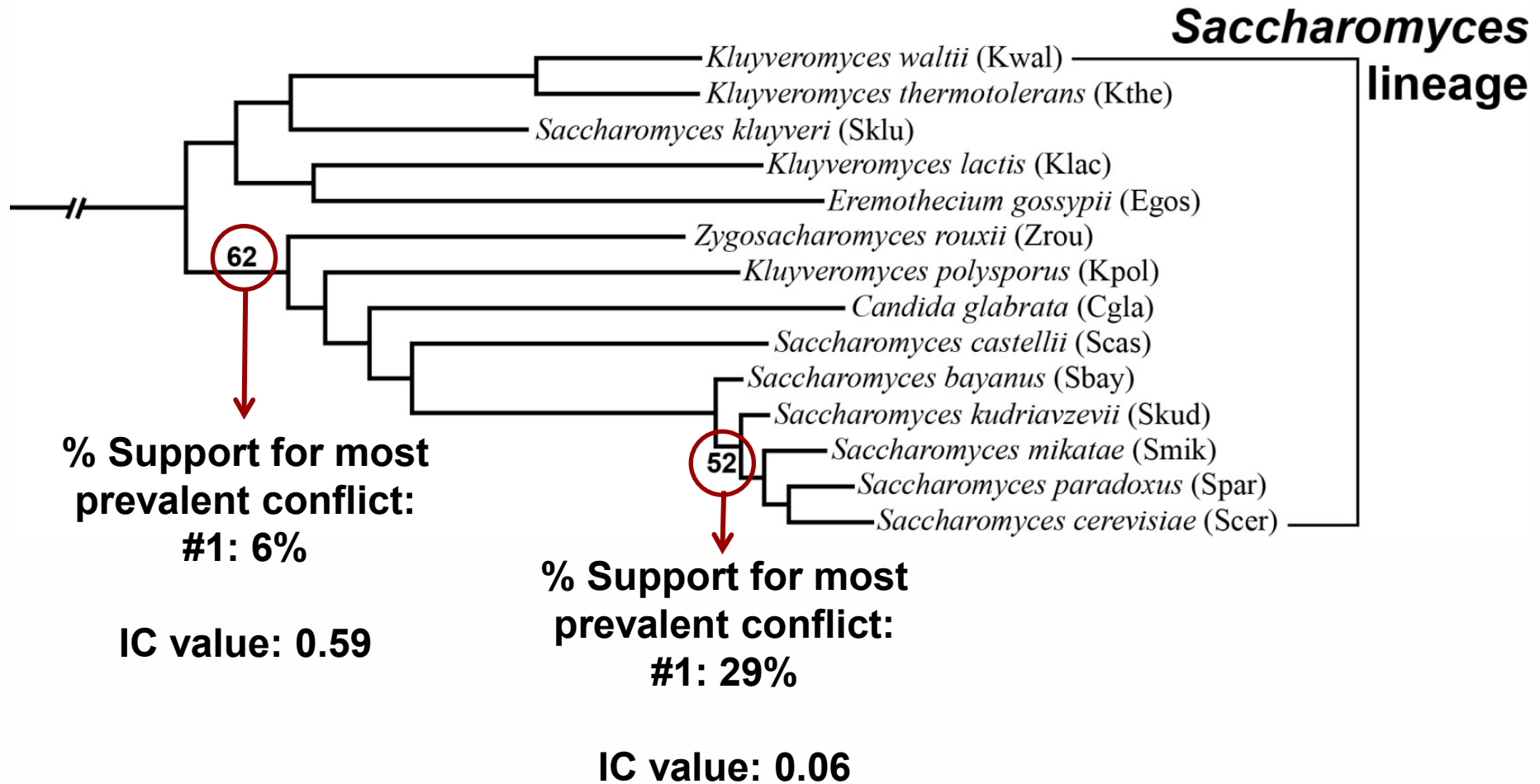
(<https://github.com/lutteropp/QuartetScores>)



Ratio of Supports for
Two Conflicting
Bipartitions

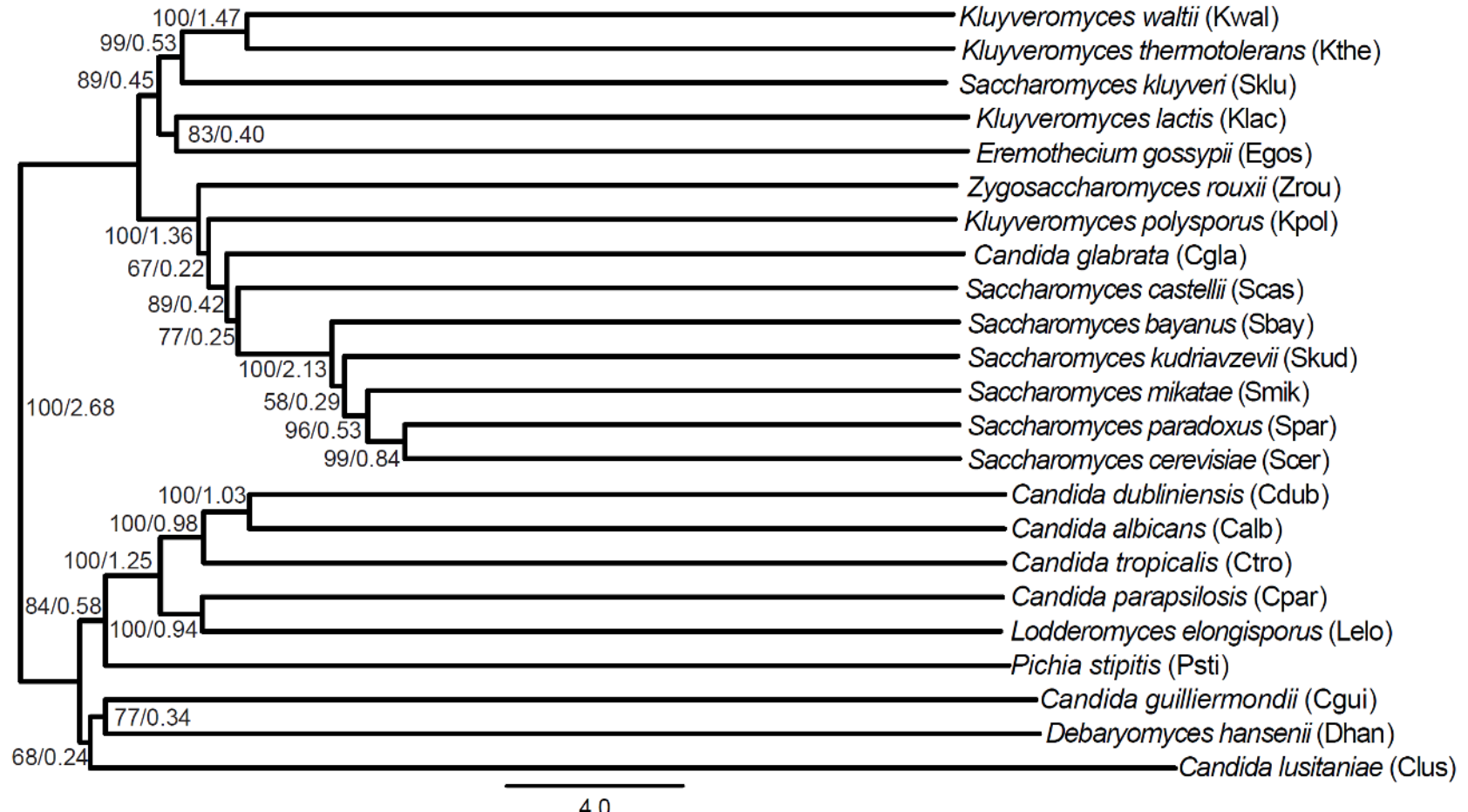


IC is a more informative measure of branch support



Coalescent-based inference of the yeast phylogeny

The yeast species phylogeny inferred using the STAR species tree method



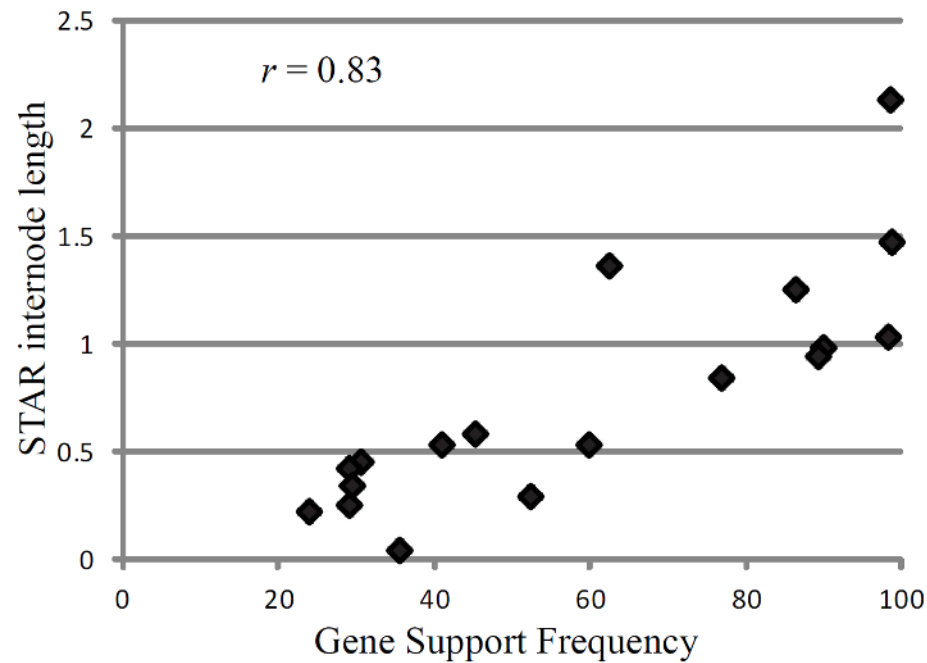
Coalescent units / IC



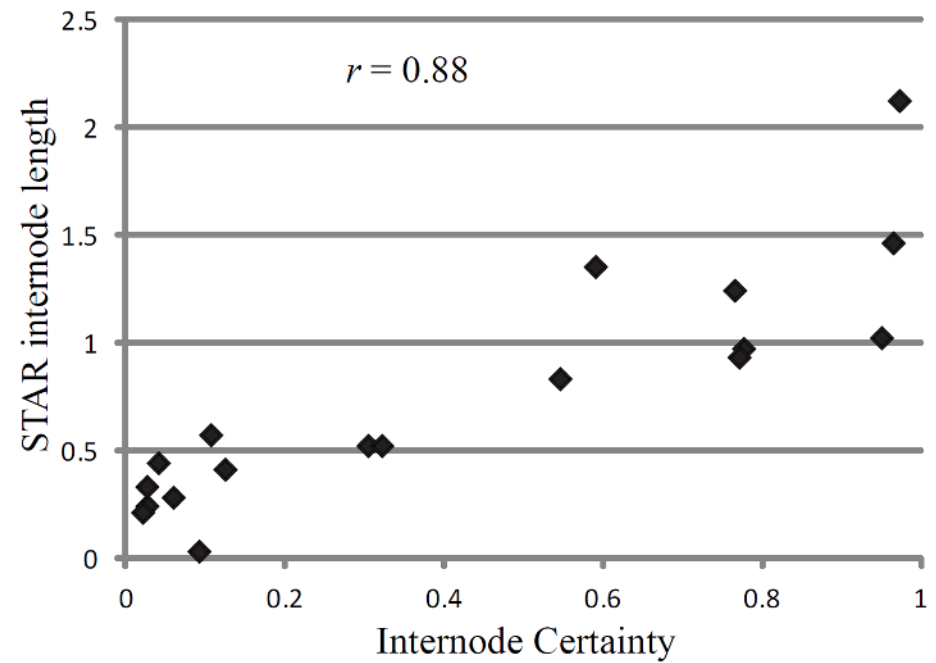
Coalescent branch lengths are correlated with GSF/IC

b

STAR internode length versus GSF

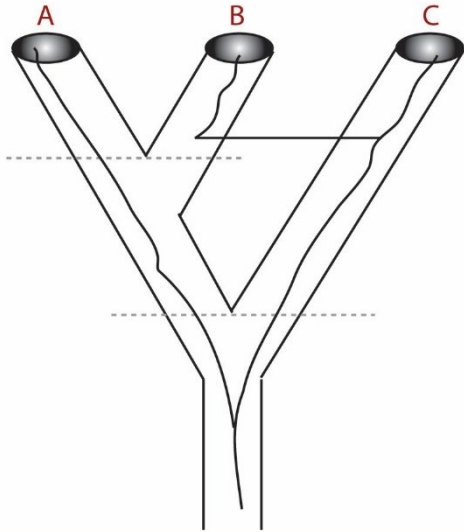


STAR internode length versus IC

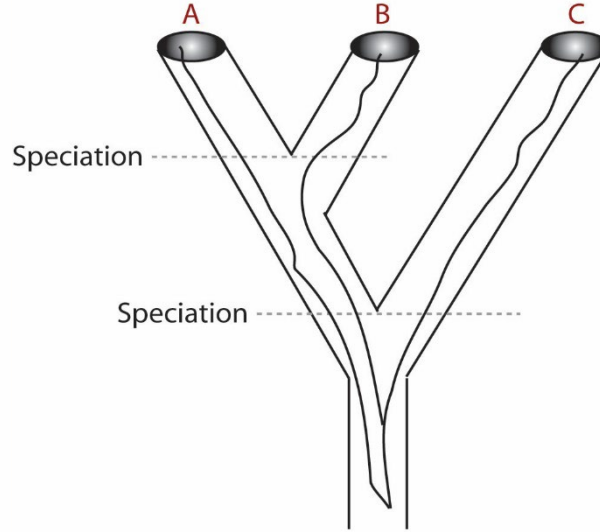


Why so much incongruence? Biological factors

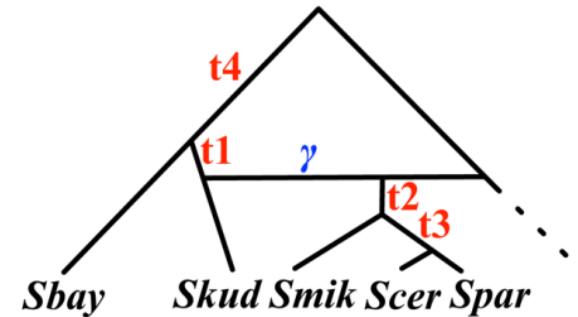
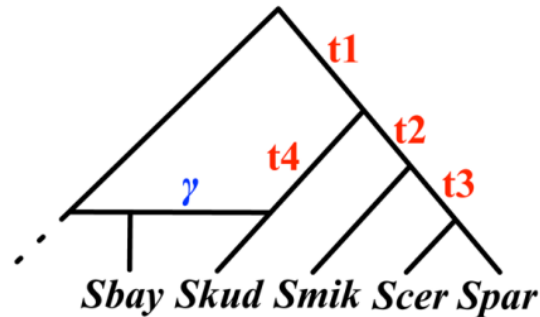
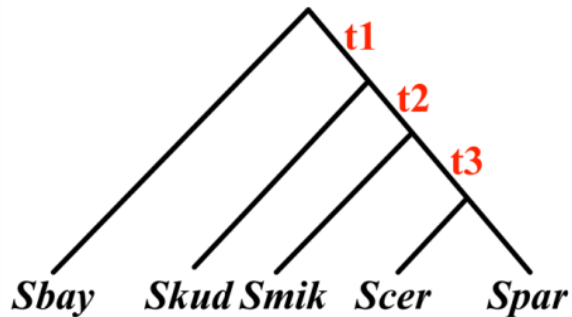
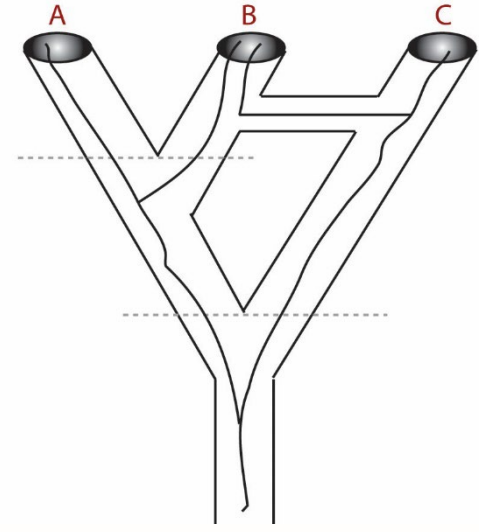
Horizontal Gene Transfer



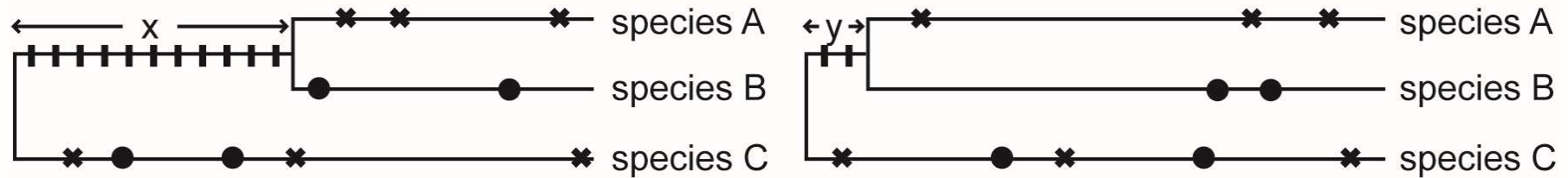
Lineage Sorting



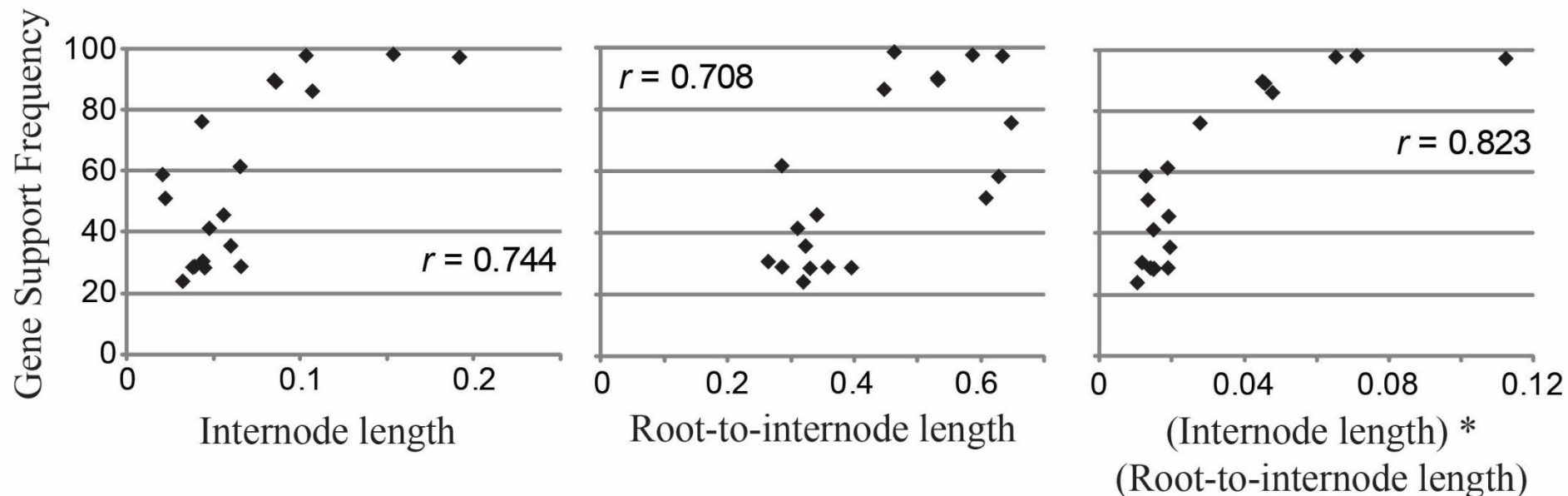
Hybridization



Why so much incongruence? Analytical factors



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



Certain recipes for handling incongruence didn't help

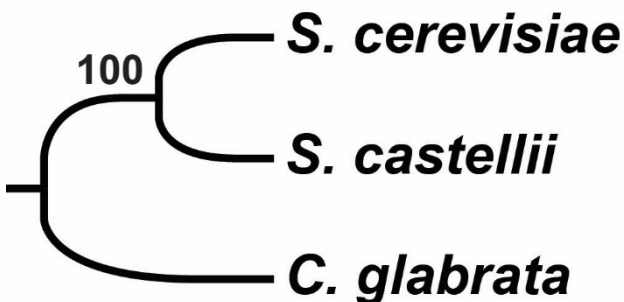
Treatment	Tree Certainty	# of Internodes where IC increased decreased
Default analysis	8.35	n/a
<i>Removing sites containing gaps</i>		
All sites with gaps excluded	7.91	0 7
<i>Removing fast-evolving or unstable species</i>		
<i>C. lusitaniae</i>	8.15	1 2
<i>C. glabrata</i>	8.30	2 2
<i>E. gossypii</i>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



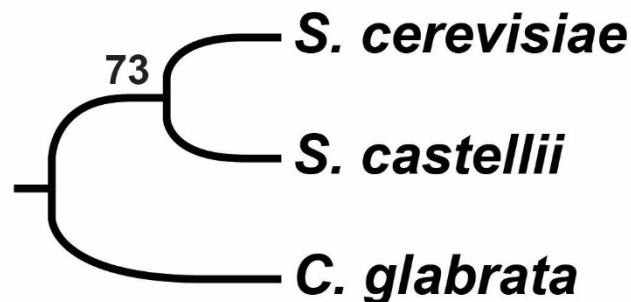
What do we do then?

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

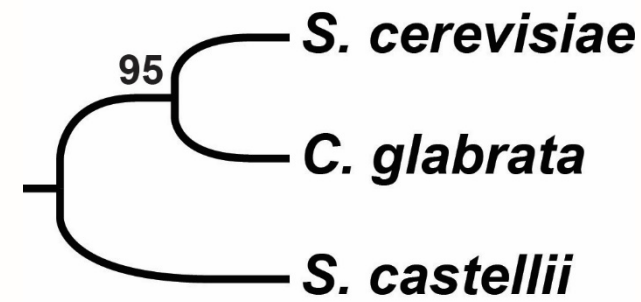
average BS $\geq 60\%$



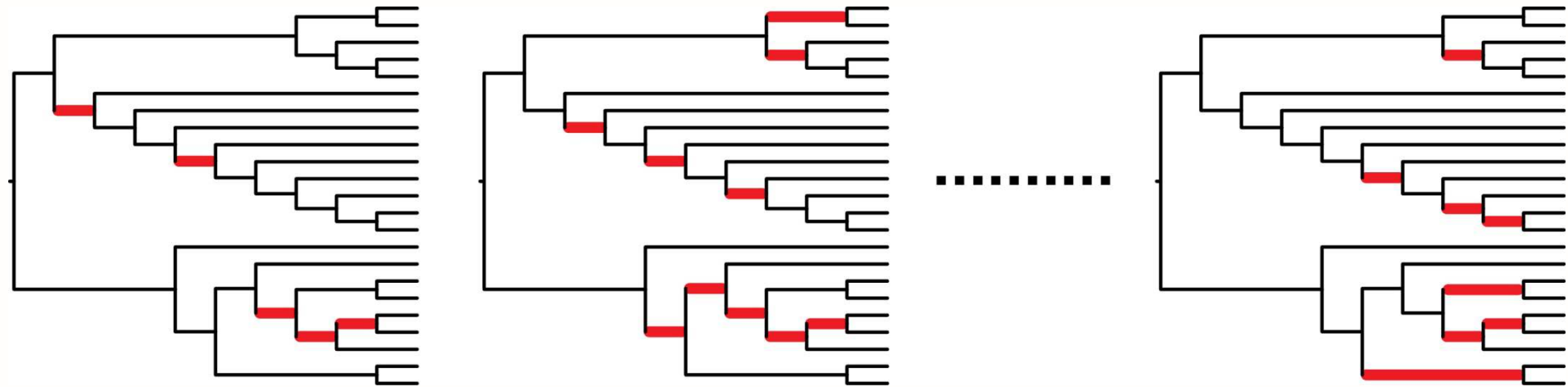
average BS $\geq 70\%$



average BS $\geq 80\%$



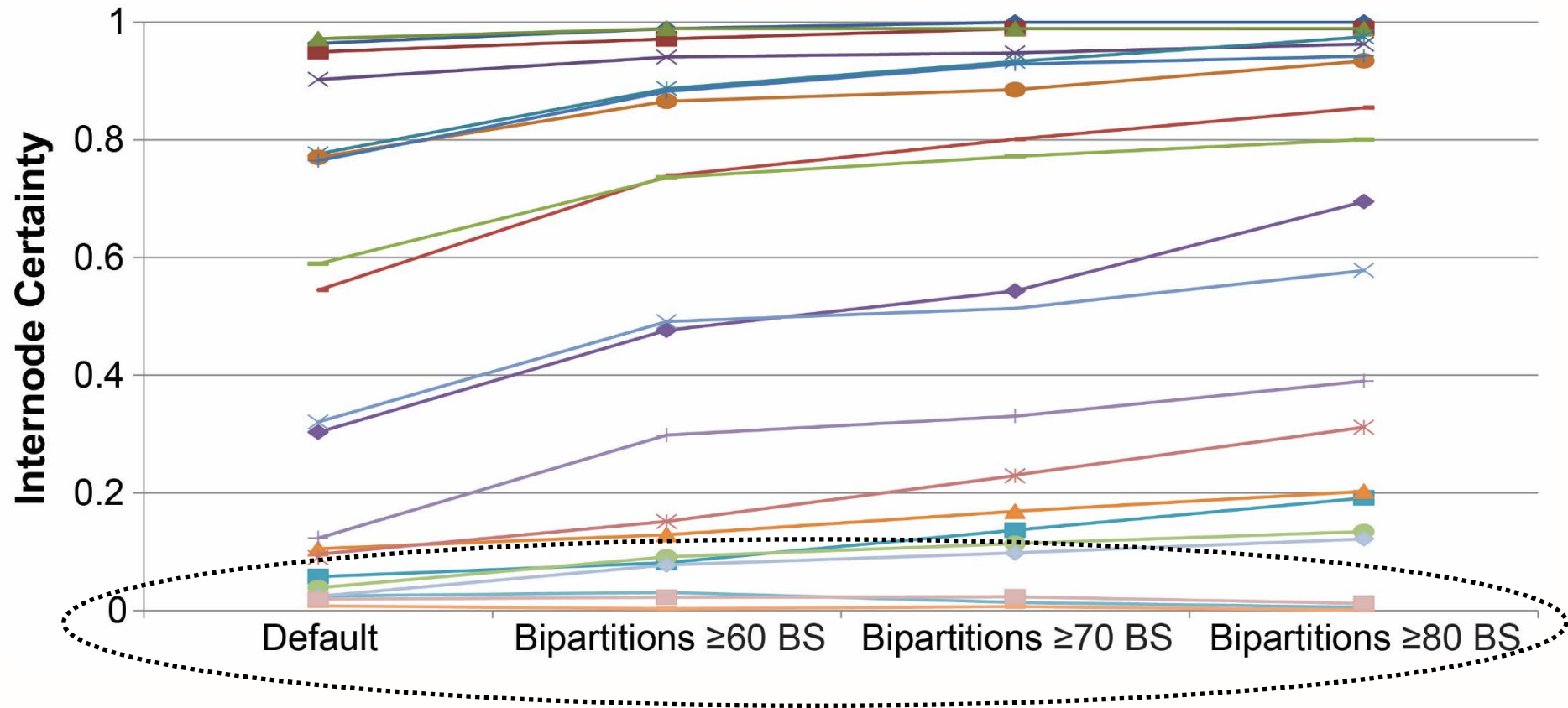
Selecting specific bipartitions dramatically improves phylogeny



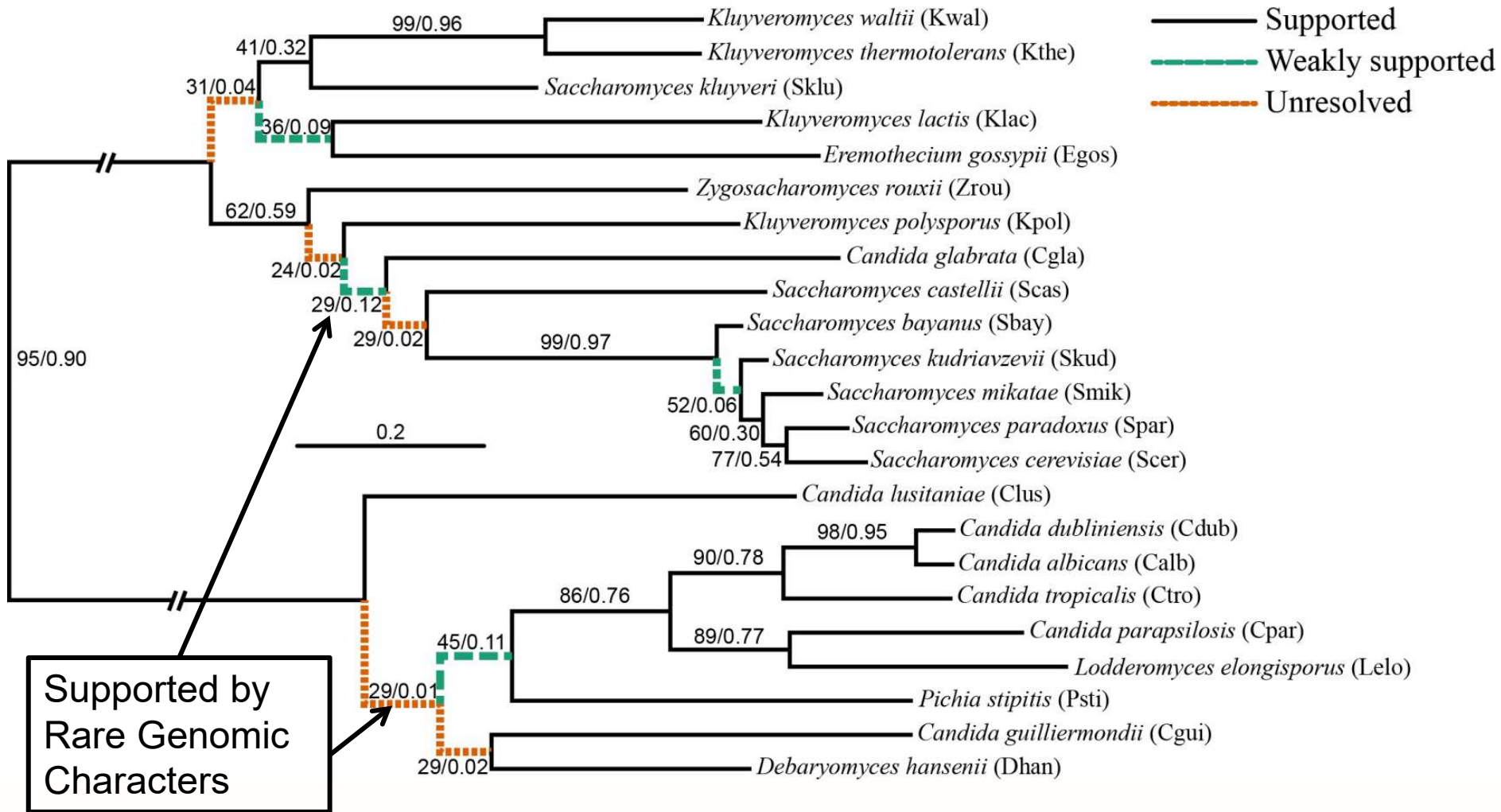
Treatment	Tree Certainty	# of Internodes where IC increased decreased
Default analysis	8.35	n/a
<i>Selecting genes whose bootstrap consensus trees have high average support</i>		
All bipartitions with BS $\geq$ 60%	10.11	14 0
All bipartitions with BS $\geq$ 70%	10.70	16 0
All bipartitions with BS $\geq$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



Least supported internodes harbor the most conflict



The status of the yeast phylogeny



Vertebrates

(1,086 genes, 18 taxa)

Animals

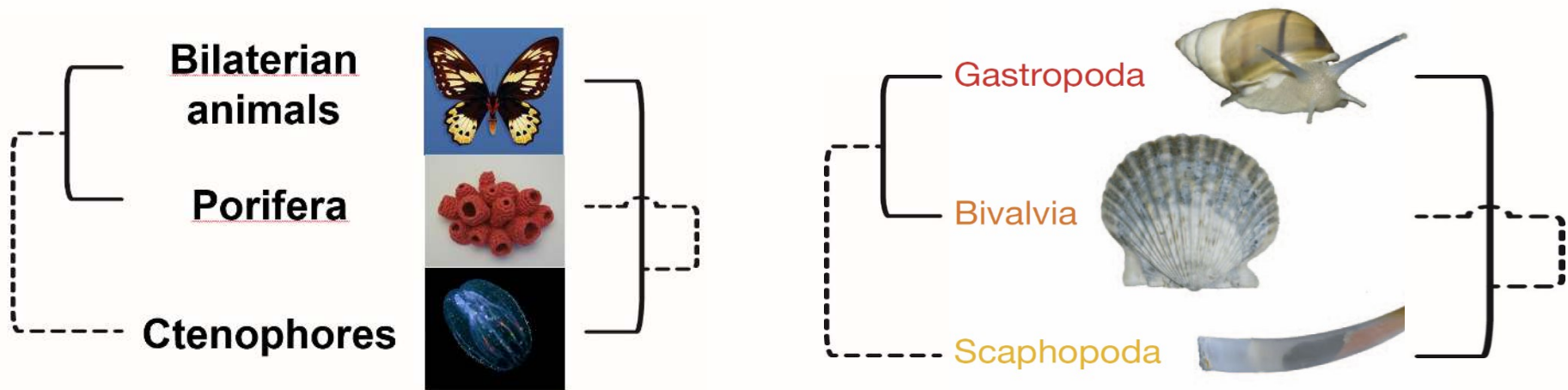
(225 genes, 21 taxa)

Mosquitoes

(2,007 genes, 20 taxa)



Incongruence in phylogenomic datasets



Hypothesis: these debates concern internodes that are poorly supported by individual gene trees

Test: measure the phylogenetic signal in contentious branches of the tree of life

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

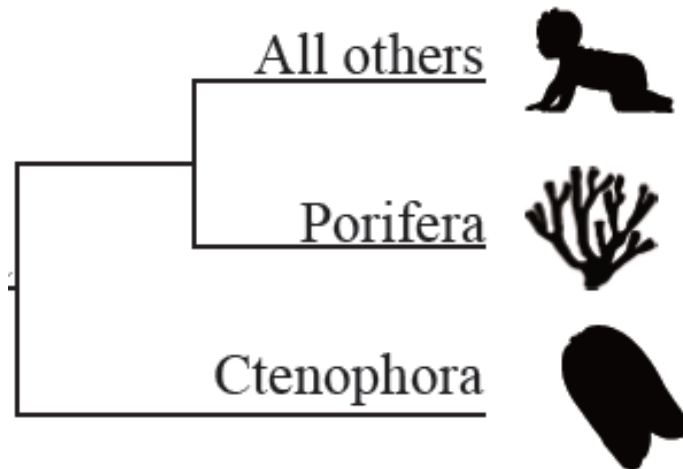
Philippe et al. (2011) PLoS Biol.

Townsend et al. (2012) Syst. Biol.

Our definition

Maximum Likelihood tree

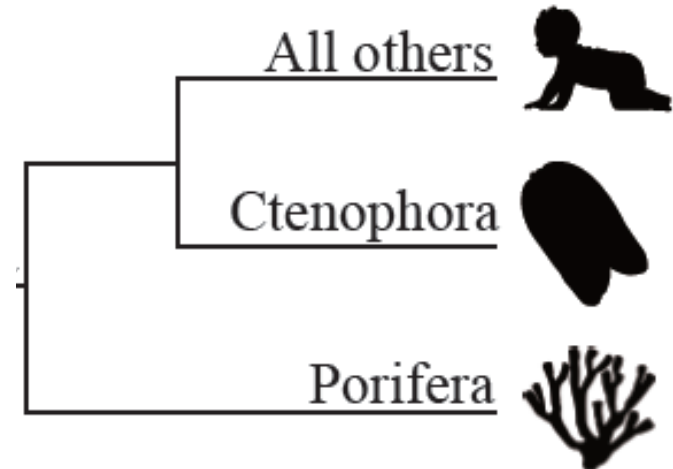
(T1)



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

Conflicting tree

(T2)



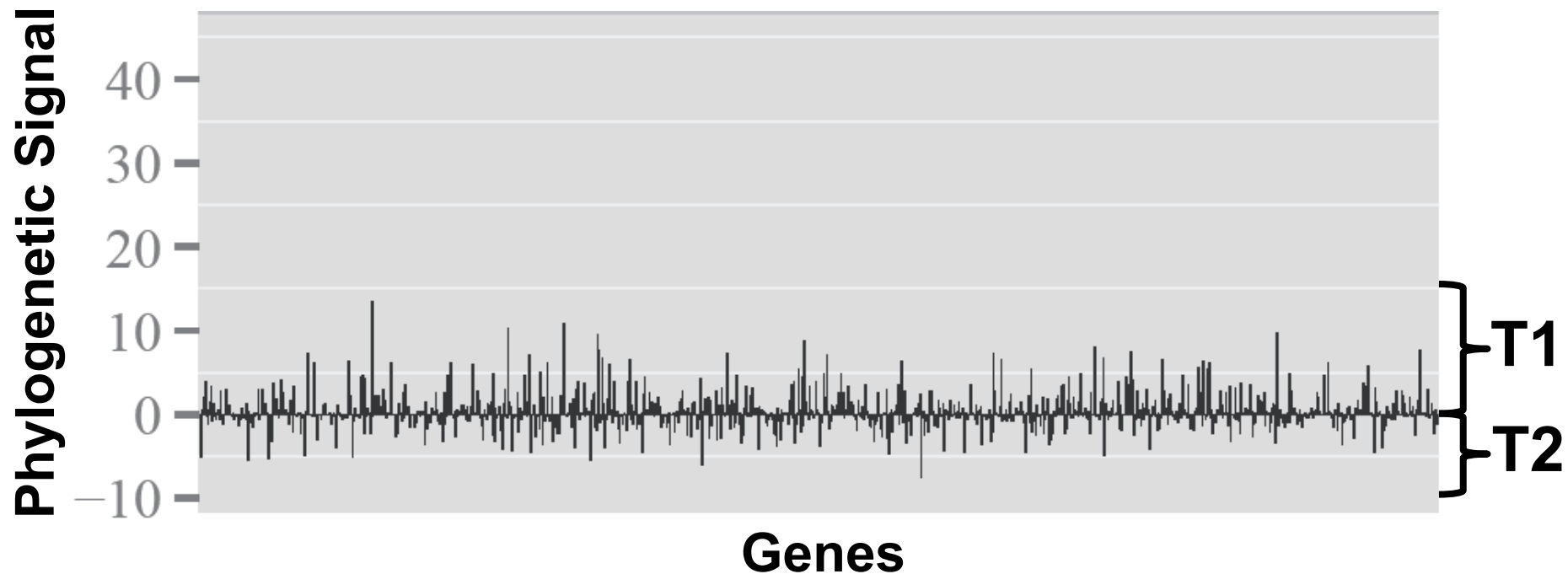
$$\ln(T_2|X_i) = -150$$

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

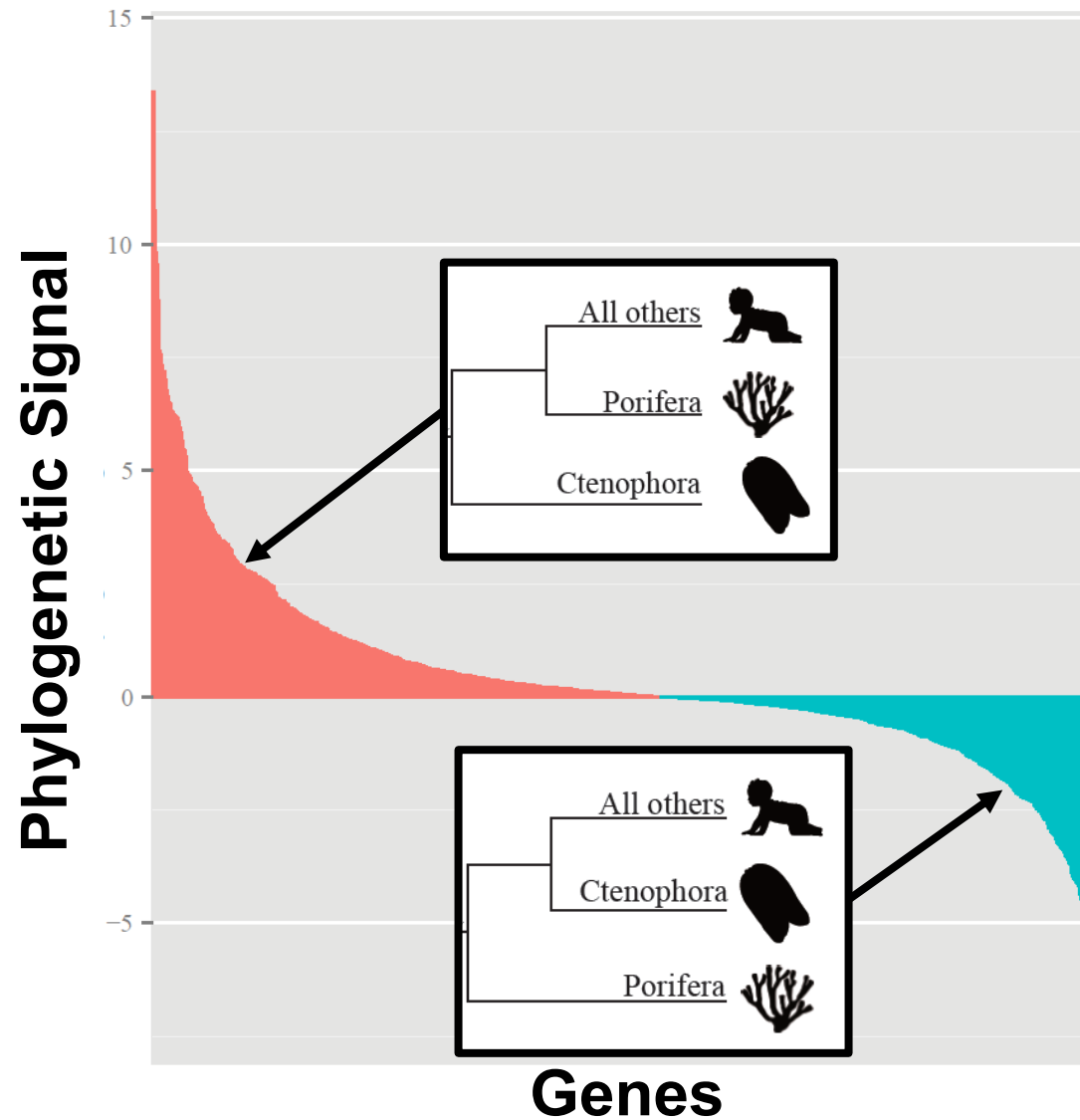


Signal of genes in a phylogenomic data matrix

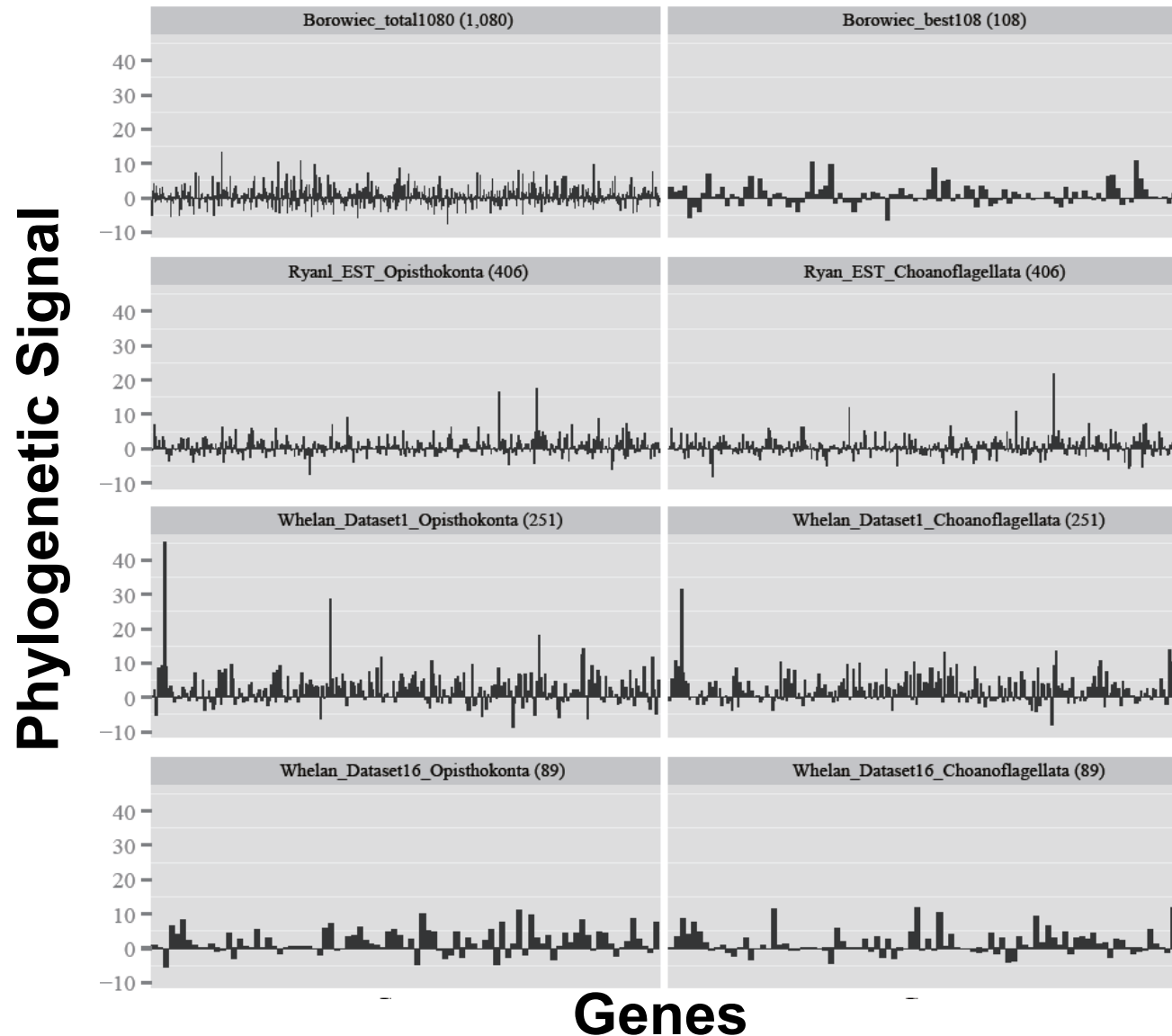
1,080 genes from 36 animal taxa



Signal of genes in a phylogenomic data matrix



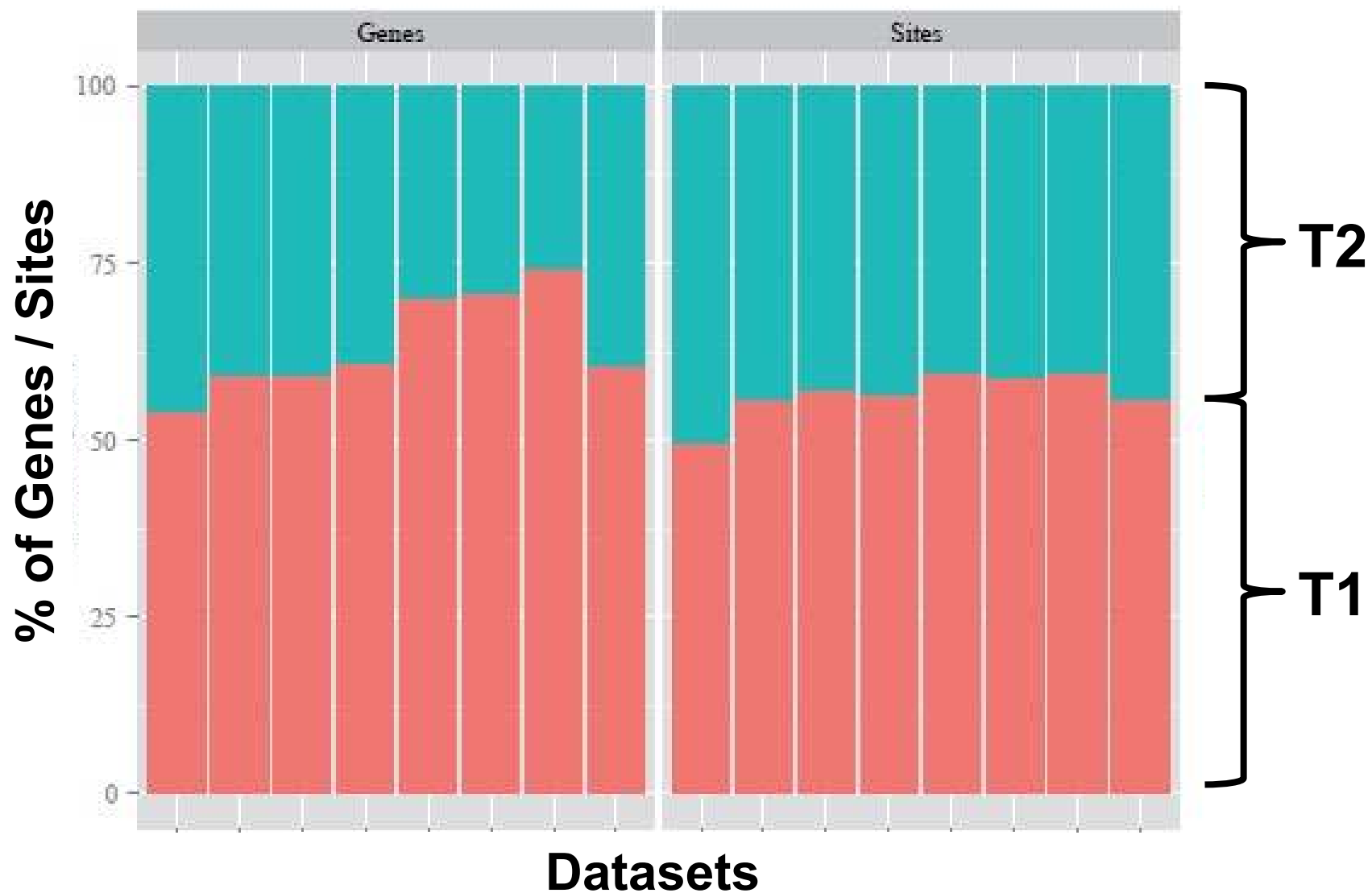
Signal of genes in multiple phylogenomic data matrices



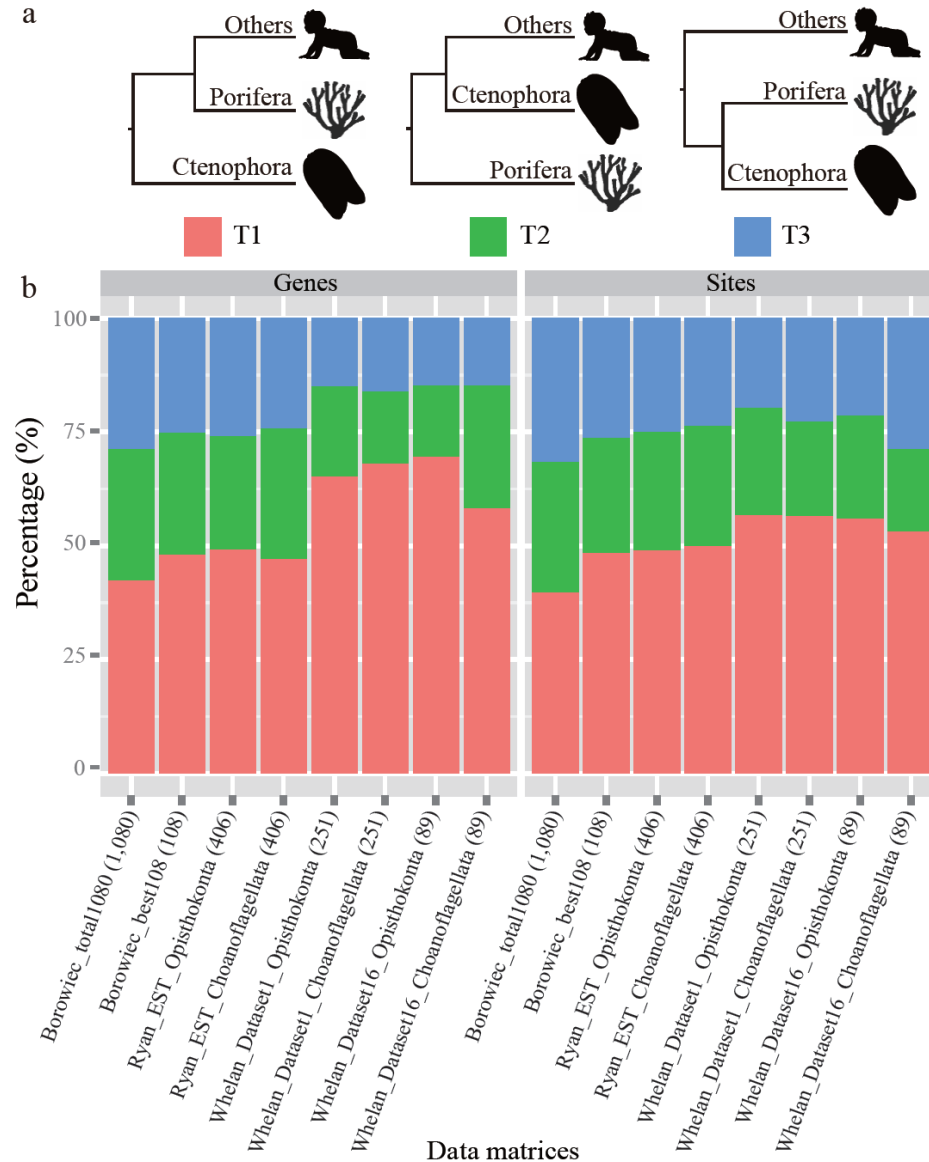
Shen et al. (2017) Nature Ecol. Evol.



Summarizing phylogenetic signal across genes and sites



Summarizing the signal across all 3 possible topologies

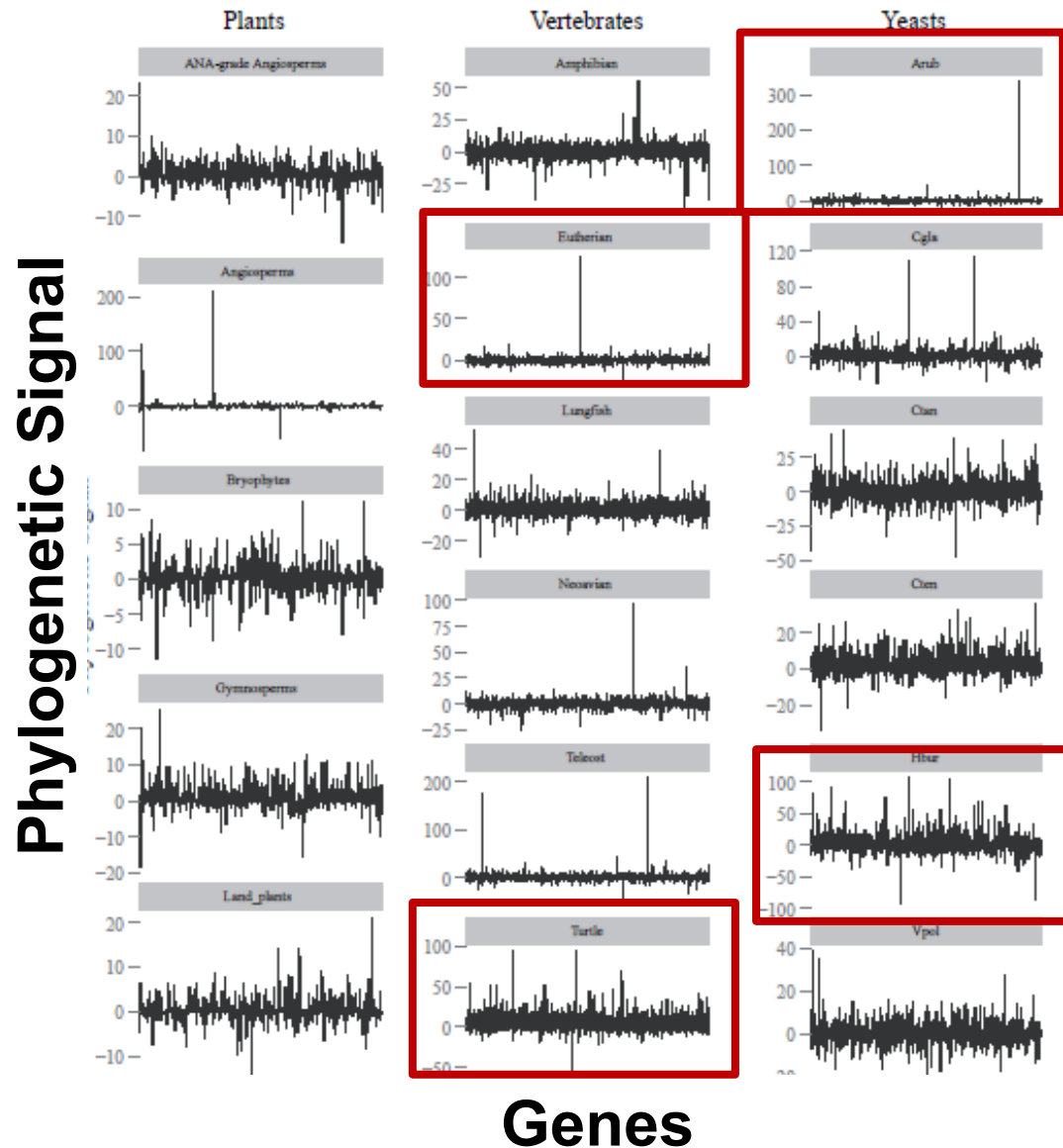


Testing several contentious branches of the tree of life

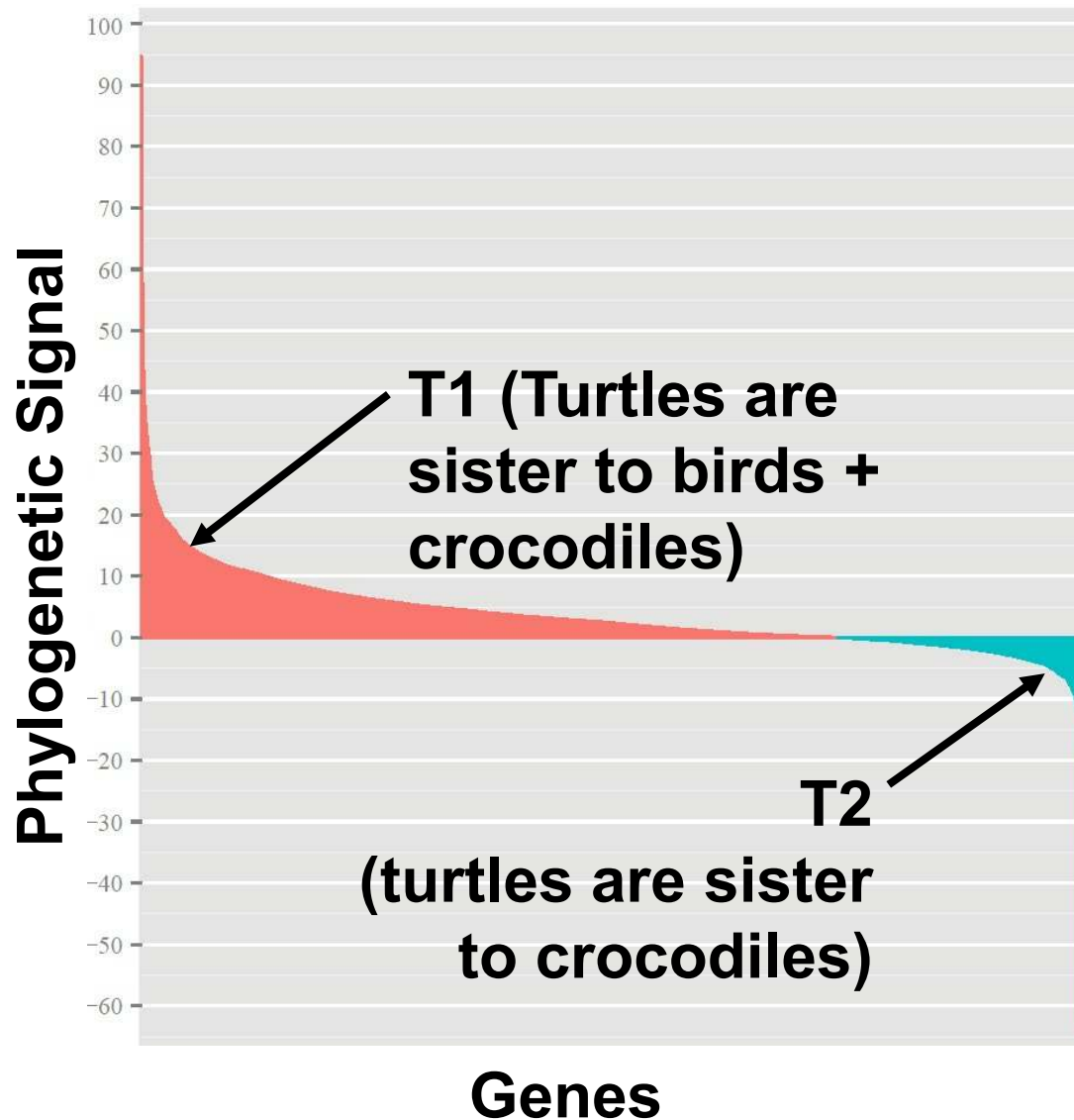
Clade	ML Tree (T1)	Conflicting Tree (T2)
Plants	<i>Amborella</i> as sister to all other flowering plants	<i>Amborella</i> + <i>Nuphar</i> as sister to all other flowering plants
	Magnoliids as sister to Eudicots + Chloranthales	Eudicots as sister to Magnoliids + Chloranthales
	Hornworts as sister to all other land plants, followed by a mosses + liverworts clade	Hornworts as sister to a mosses + liverworts clade
	Gnetales as sister to the Pinaceae, nested within the Coniferales	Gnetales as sister to the Coniferales
	Zygnematophyceae as sister to all land plants	Charales as sister to all land plants
Vertebrates	Gymnophiona as sister to all other amphibians	Anura as sister to all other amphibians
	Atlantogenata (Afrotheria + Xenarthra) as sister to all other placental mammals	Afrotheria as sister to all other placental mammals
	Lungfishes as sister to all tetrapods	Lungfishes + coelacanths as sister to all tetrapods
	Pigeons as sister to all other Neoaves	Falcons as sister to all other Neoaves
	Elopomorpha + Osteoglossomorpha as sister to all other teleosts	Osteoglossomorpha alone as sister to all other teleosts
Yeasts	Turtles as sister to archosaurs (birds + crocodiles)	Turtles as sister to crocodiles
	Ascoideaceae as sister to Phaffomycetaceae + Saccharomycetaceae	Ascoideaceae as sister to a clade comprising Pichiaceae, Debaryomycetaceae, Phaffomycetaceae, and Saccharomycetaceae
	<i>Candida glabrata</i> rather than <i>Naumovozya castellii</i> as sister to <i>Saccharomyces sensu stricto</i> yeasts	<i>Naumovozya castellii</i> rather than <i>Candida glabrata</i> sister to <i>Saccharomyces sensu stricto</i> yeasts
	<i>Hyphopichia burtonii</i> as sister to <i>Candida auris</i> + <i>Metschnikowia bicuspidata</i>	<i>Hyphopichia burtonii</i> as sister to <i>Debaryomyces hansenii</i>
	<i>Zygosaccharomyces rouxii</i> as sister to all other yeasts with occurring whole-genome duplication event	<i>Vanderwaltozyma polyspora</i> as sister to all other yeast with occurring whole-genome duplication event
	<i>Meyerozyma guilliermondii</i> as sister to <i>Debaryomyces hansenii</i>	<i>Meyerozyma guilliermondii</i> as sister to <i>Hyphopichia burtonii</i> + <i>Candida auris</i>
	<i>Candida tanzawaensis</i> as sister to <i>Pichia stipiti</i> + <i>Candida maltosa</i>	<i>Pichia stipiti</i> as sister to <i>Candida tanzawaensis</i> + <i>Candida maltosa</i>



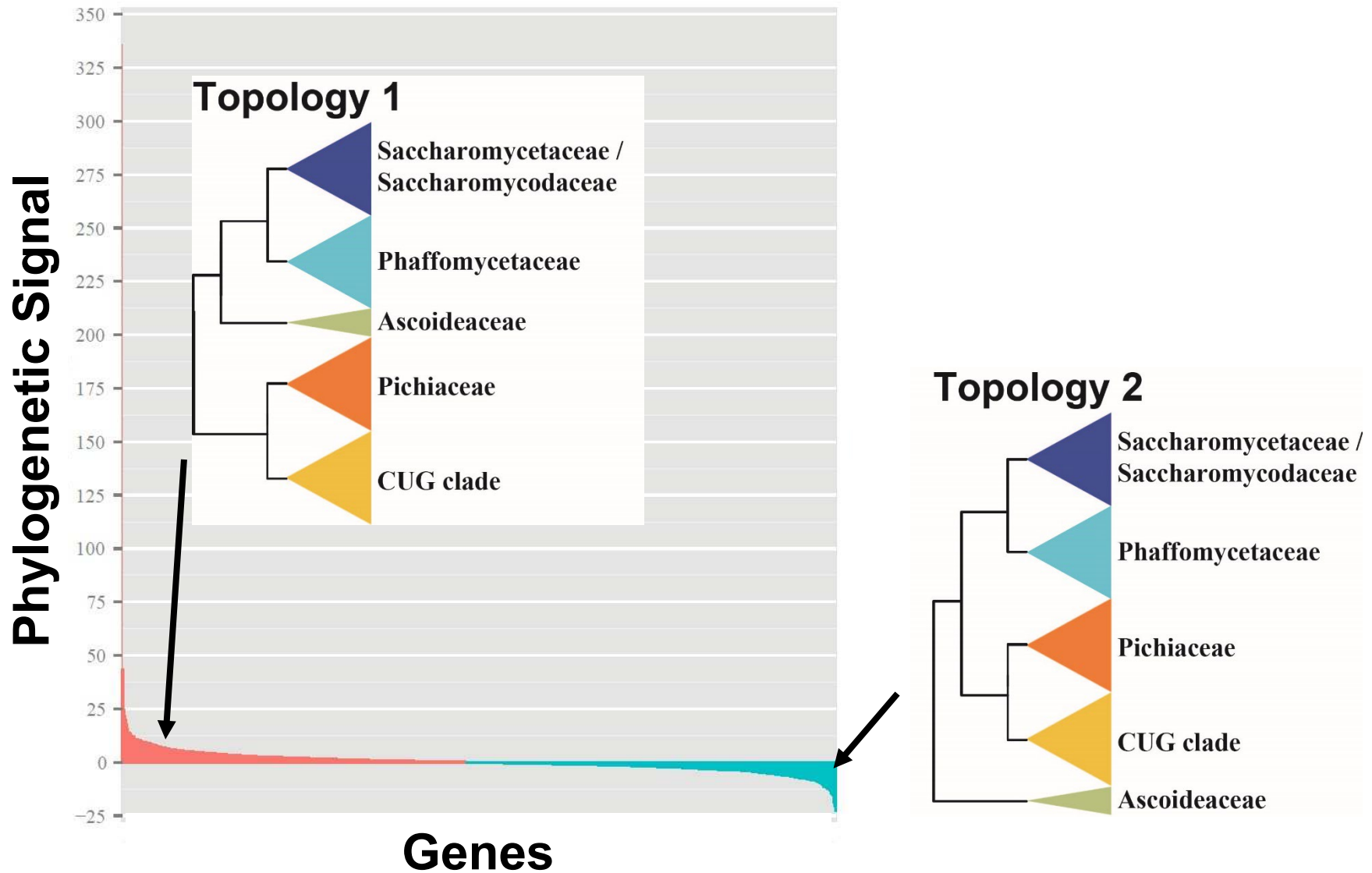
Phylogenetic signal in contentious branches of the ToL



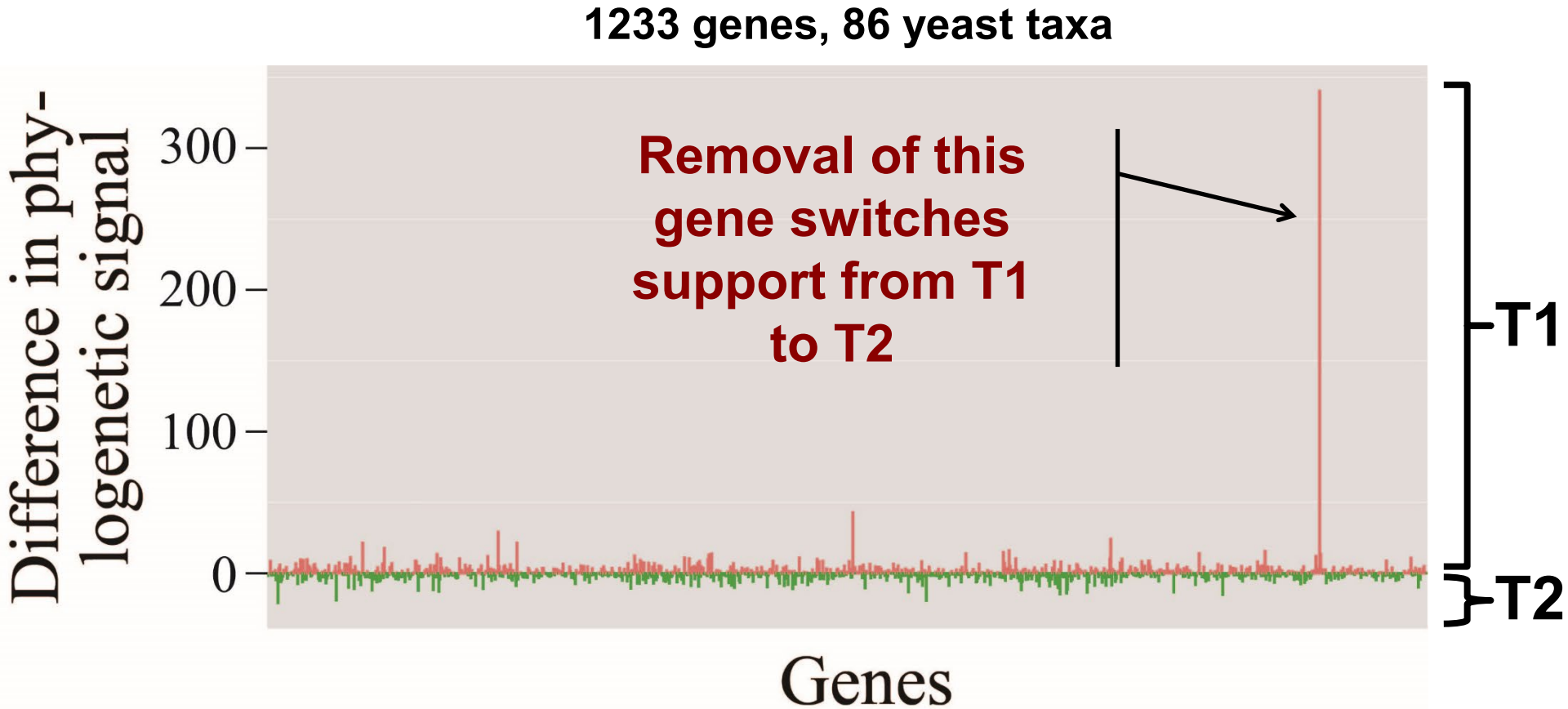
The signal in some branches is very strong...



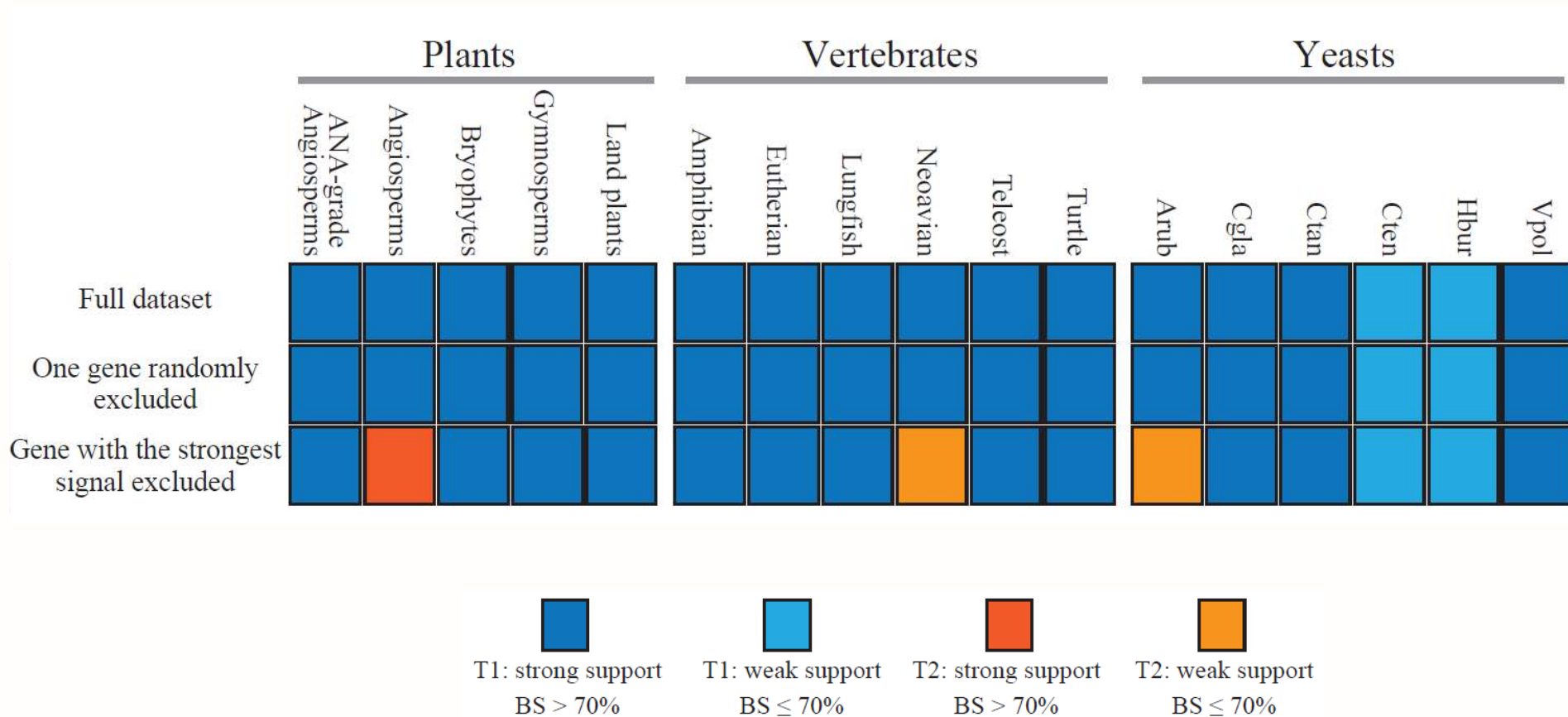
...But in others it stems from one or two genes

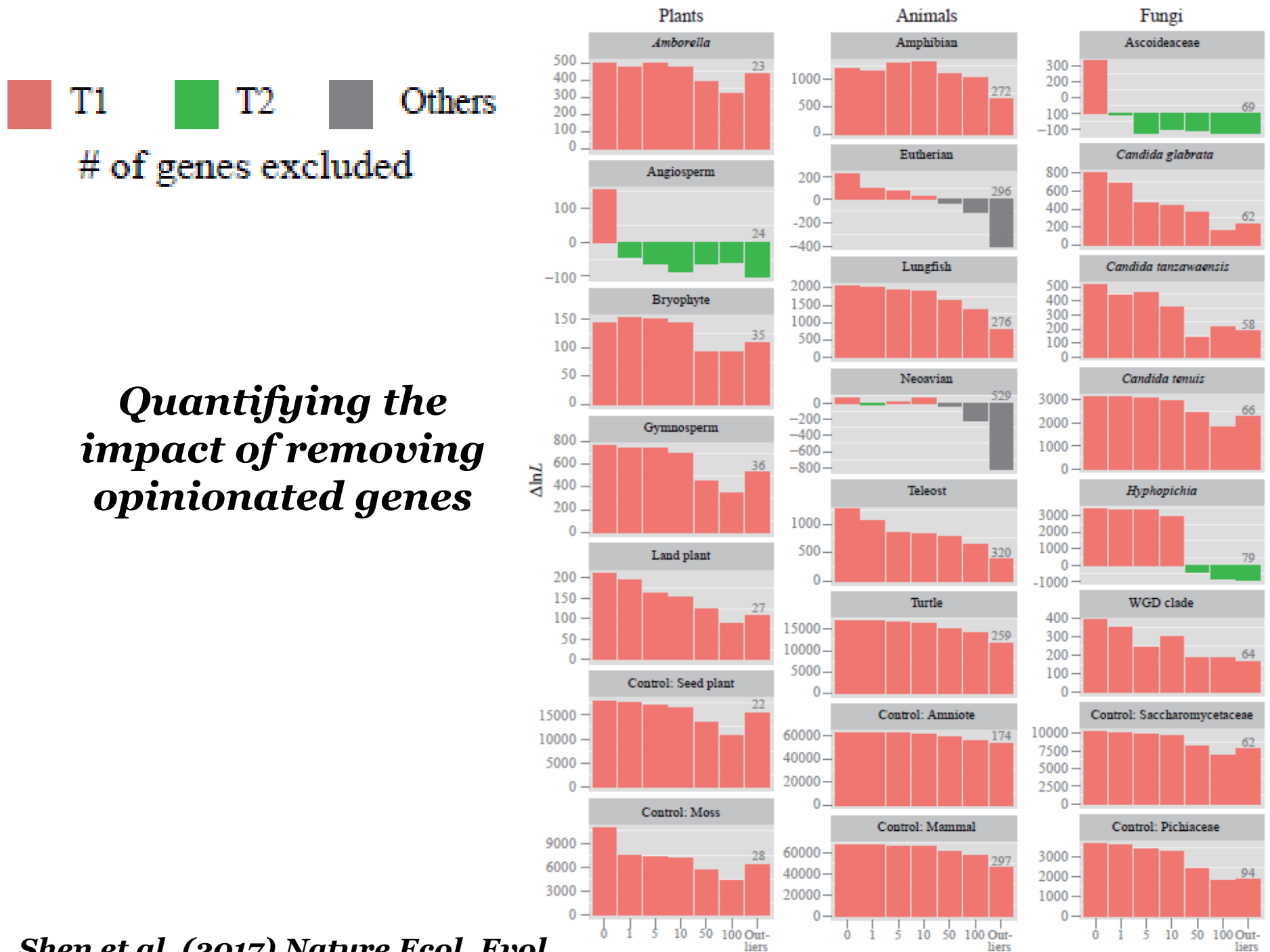


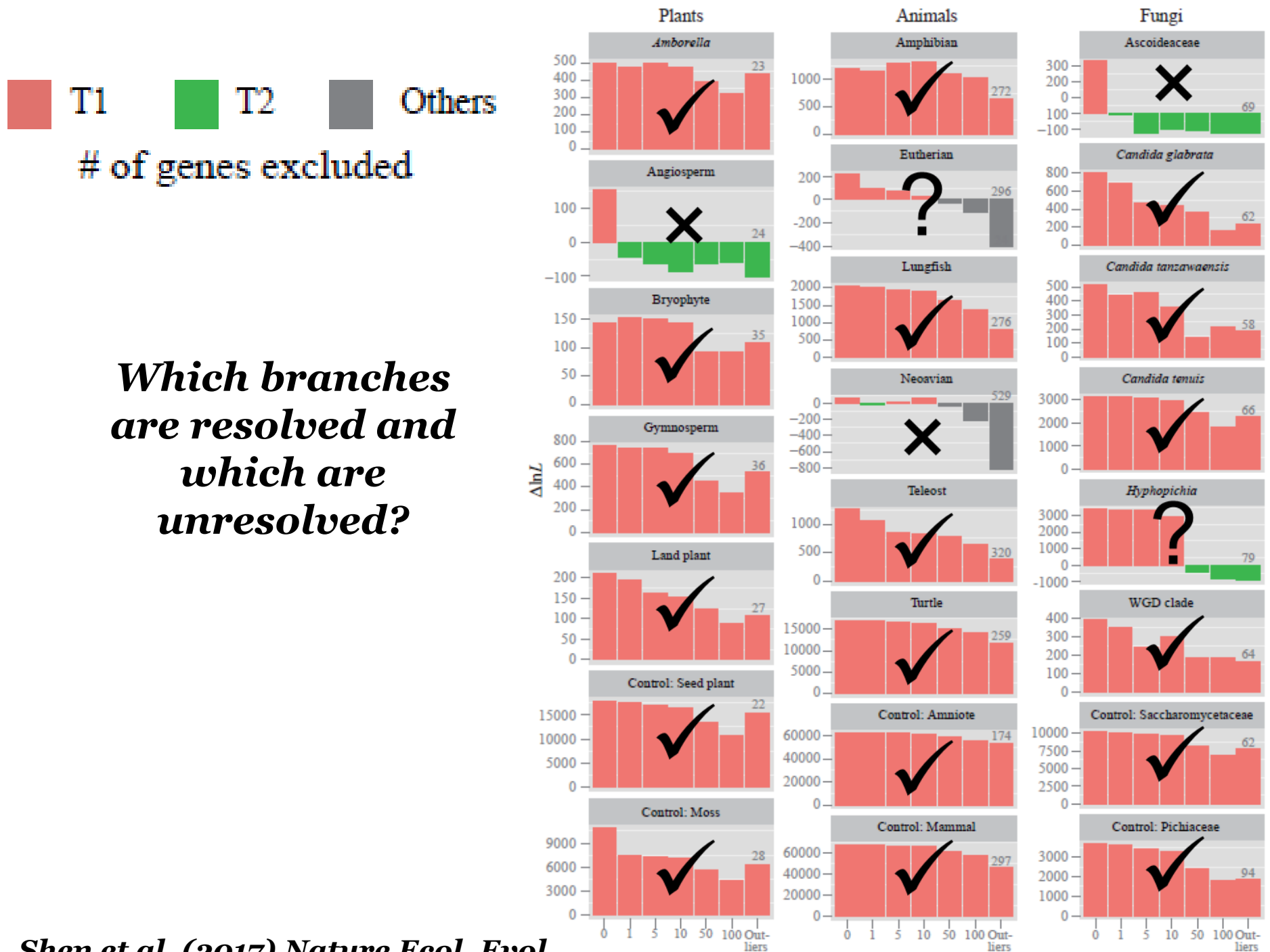
Phylogenetic signal per gene for the two hypotheses



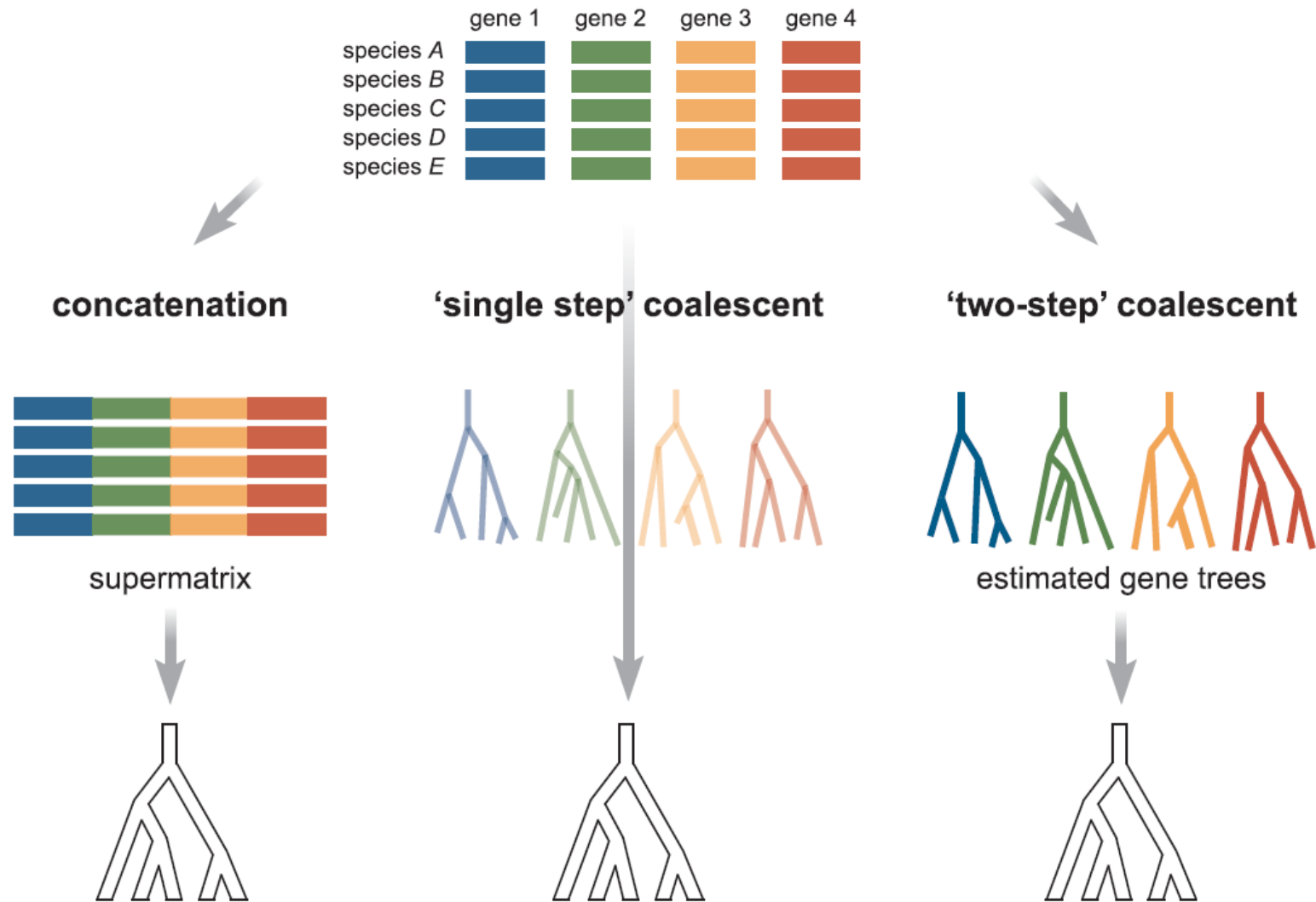
What happens if we remove that one gene?



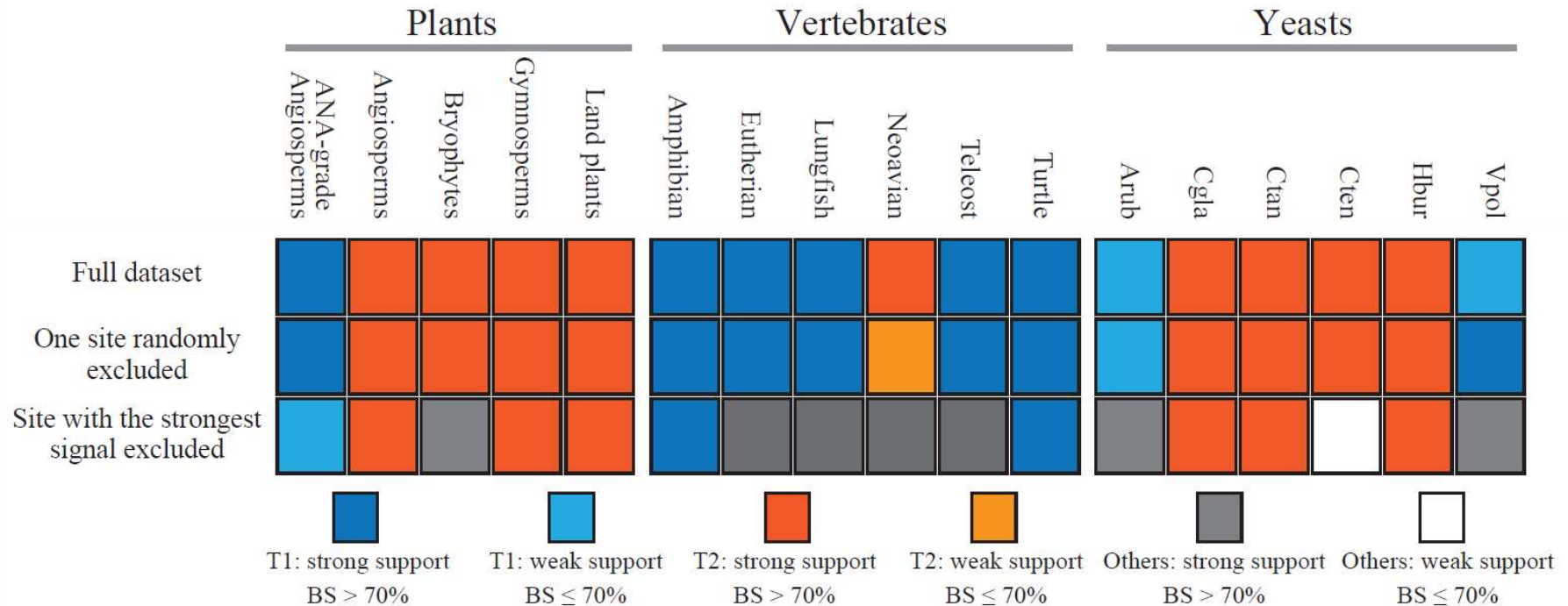




Methods for phylogenomic inference



What happens if we remove one site from every gene?

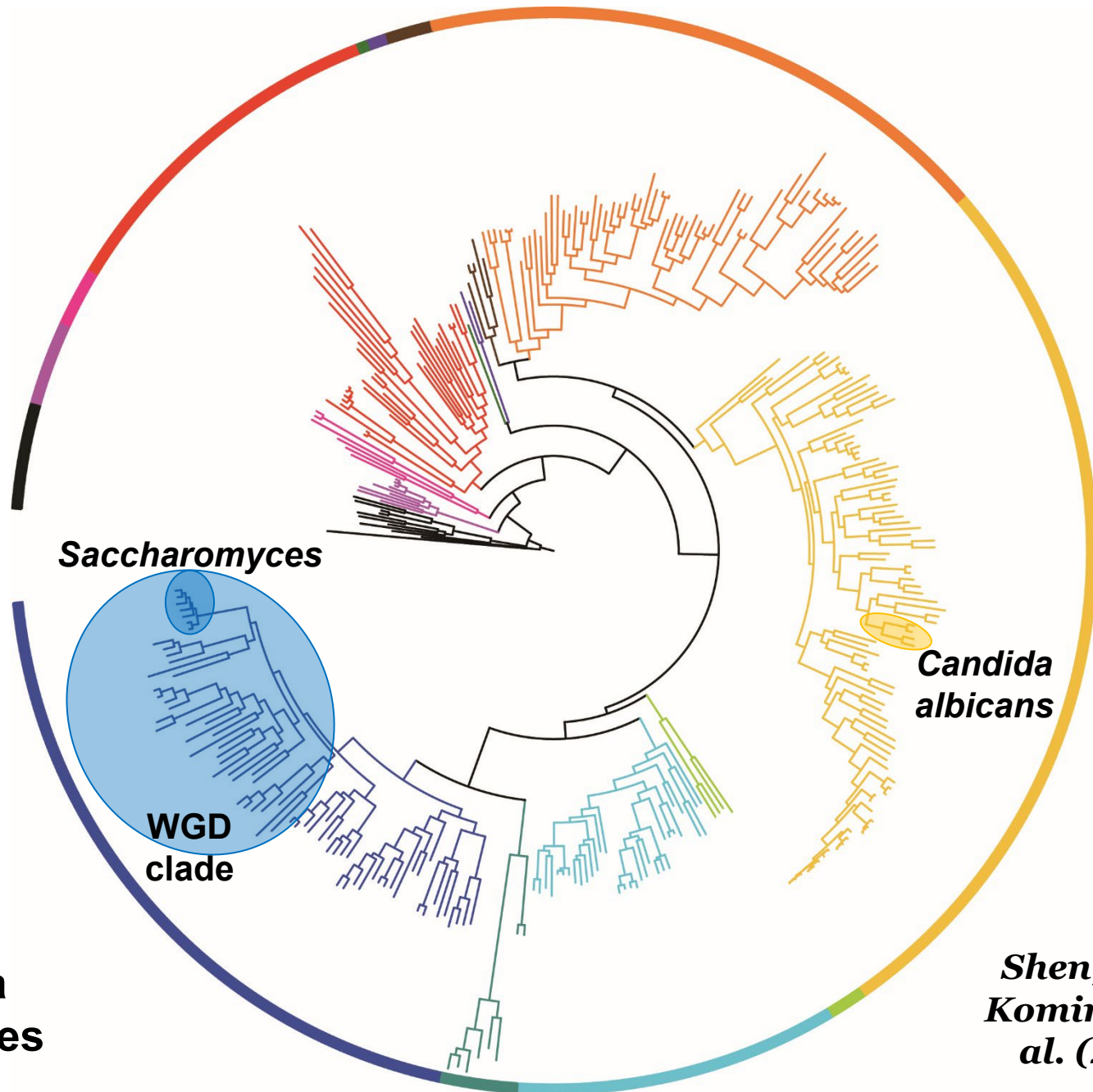


What's going on?

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

Explanation #2: Analytical factors (systematic error due to the bad fit of our models to our data)

Genome-scale phylogeny of 332 yeasts

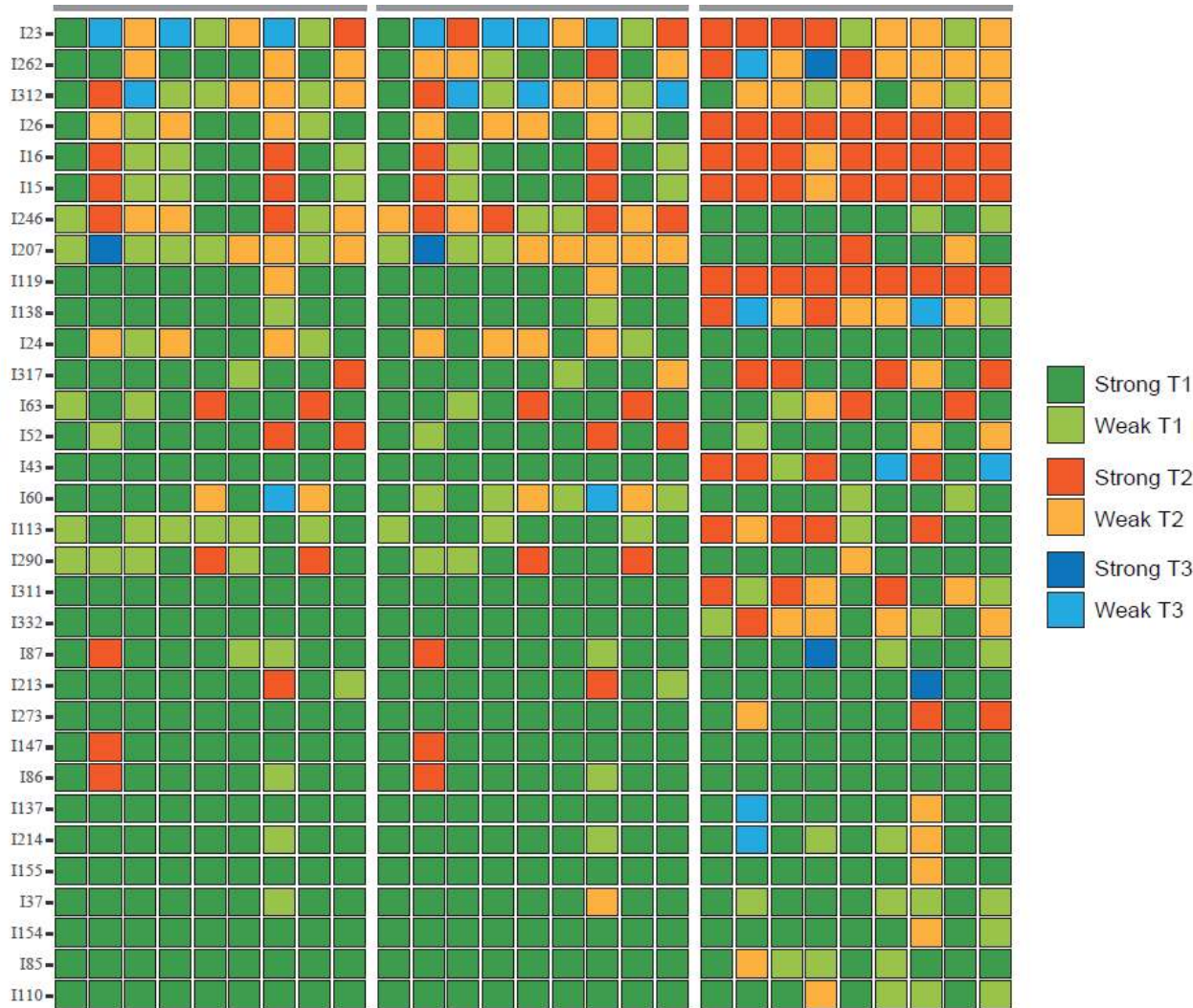


The 32 conflicting branches in the yeast phylogeny

Concatenation

Coalescence

Internal branches

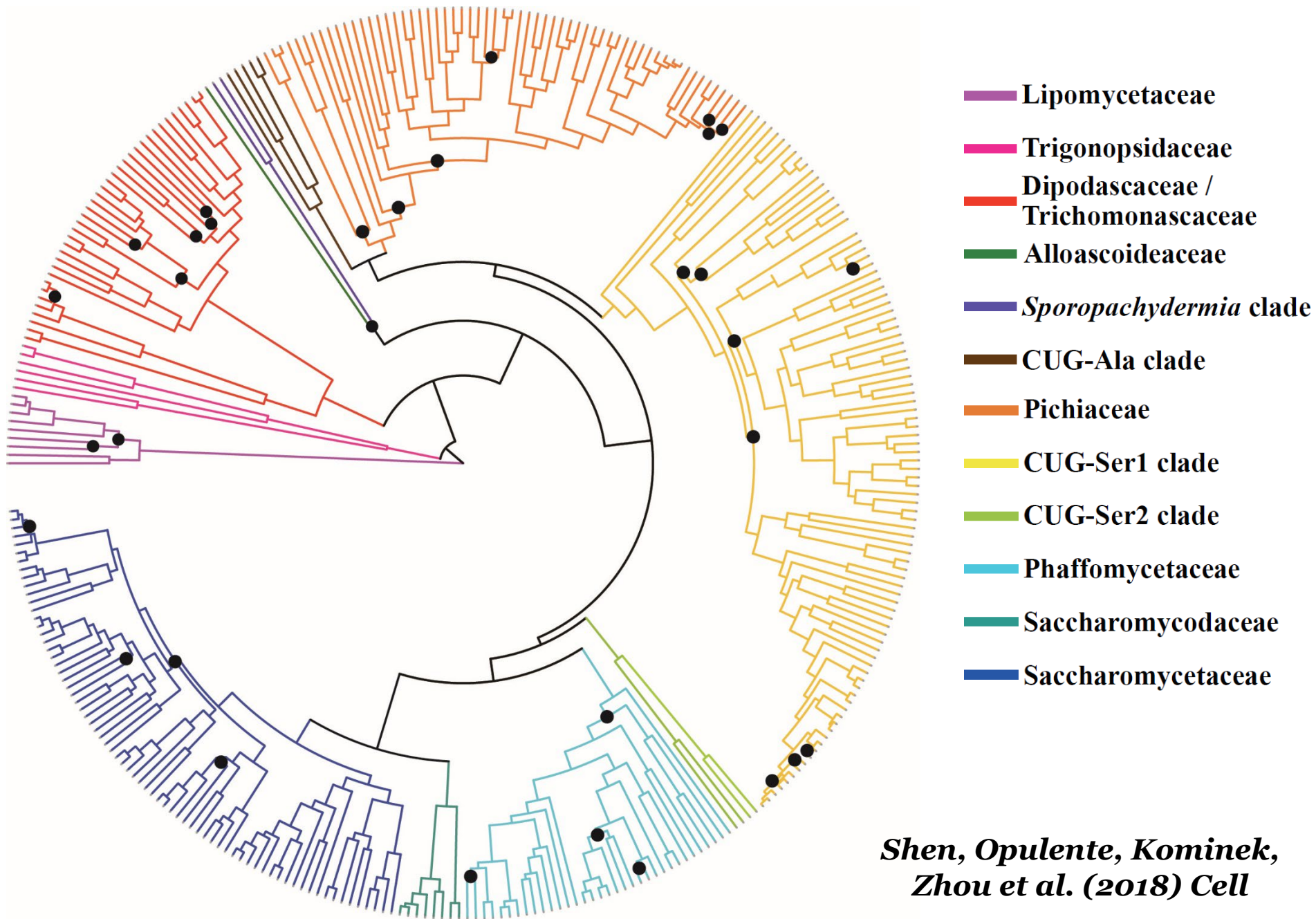


Data matrices

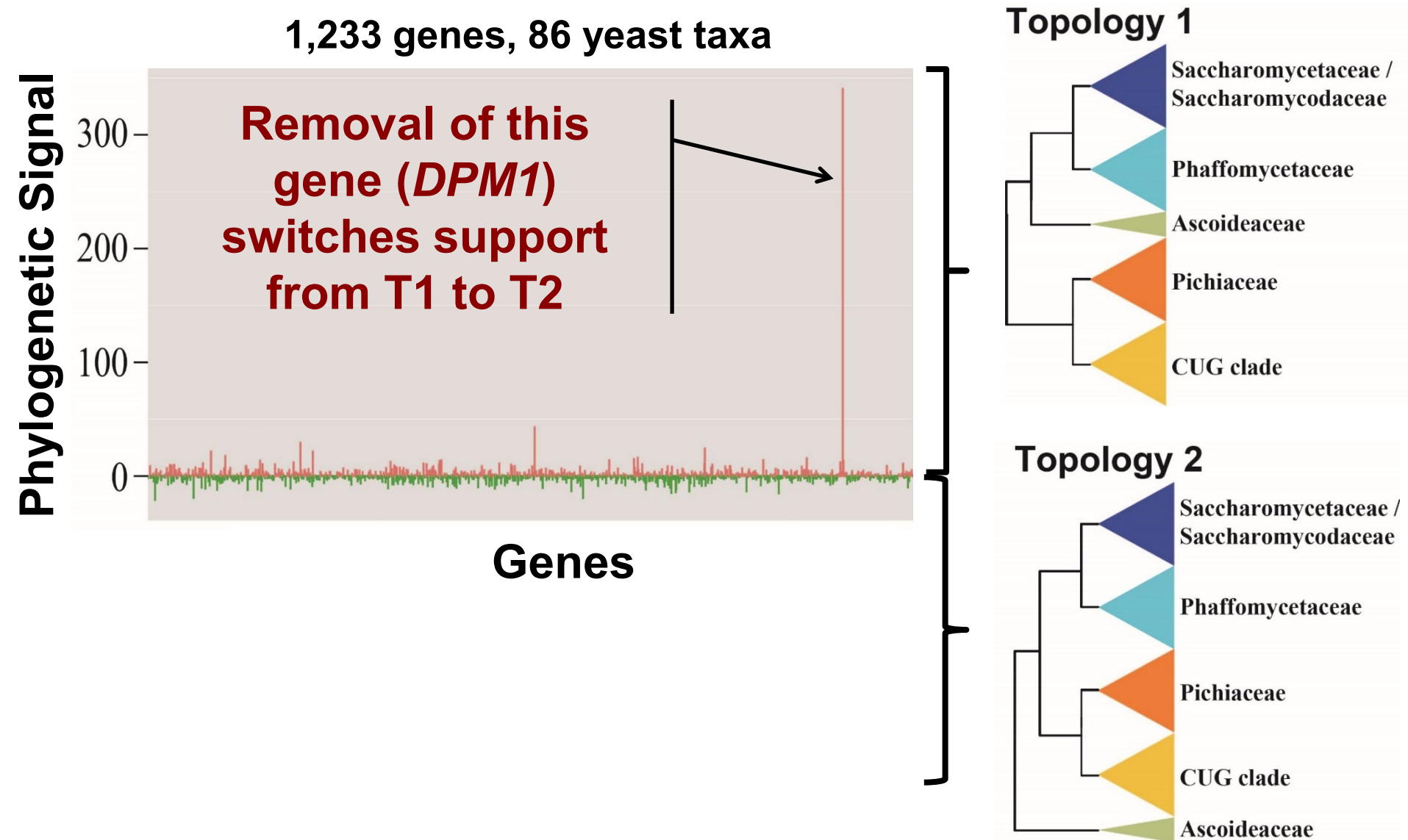
**~10% (32 / 331)
of internal
branches show
conflict between
analyses**



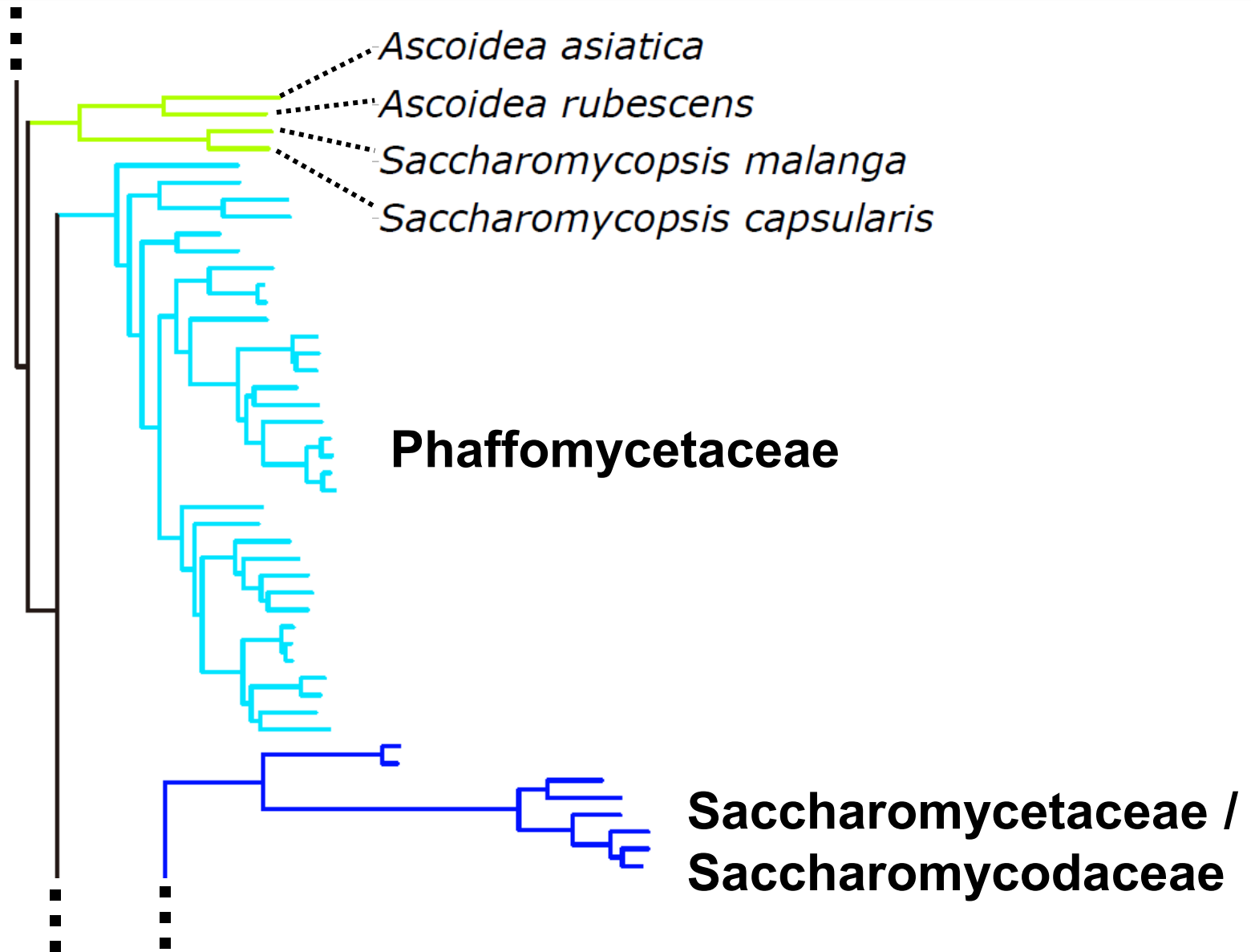
Distribution of conflict on the yeast phylogeny



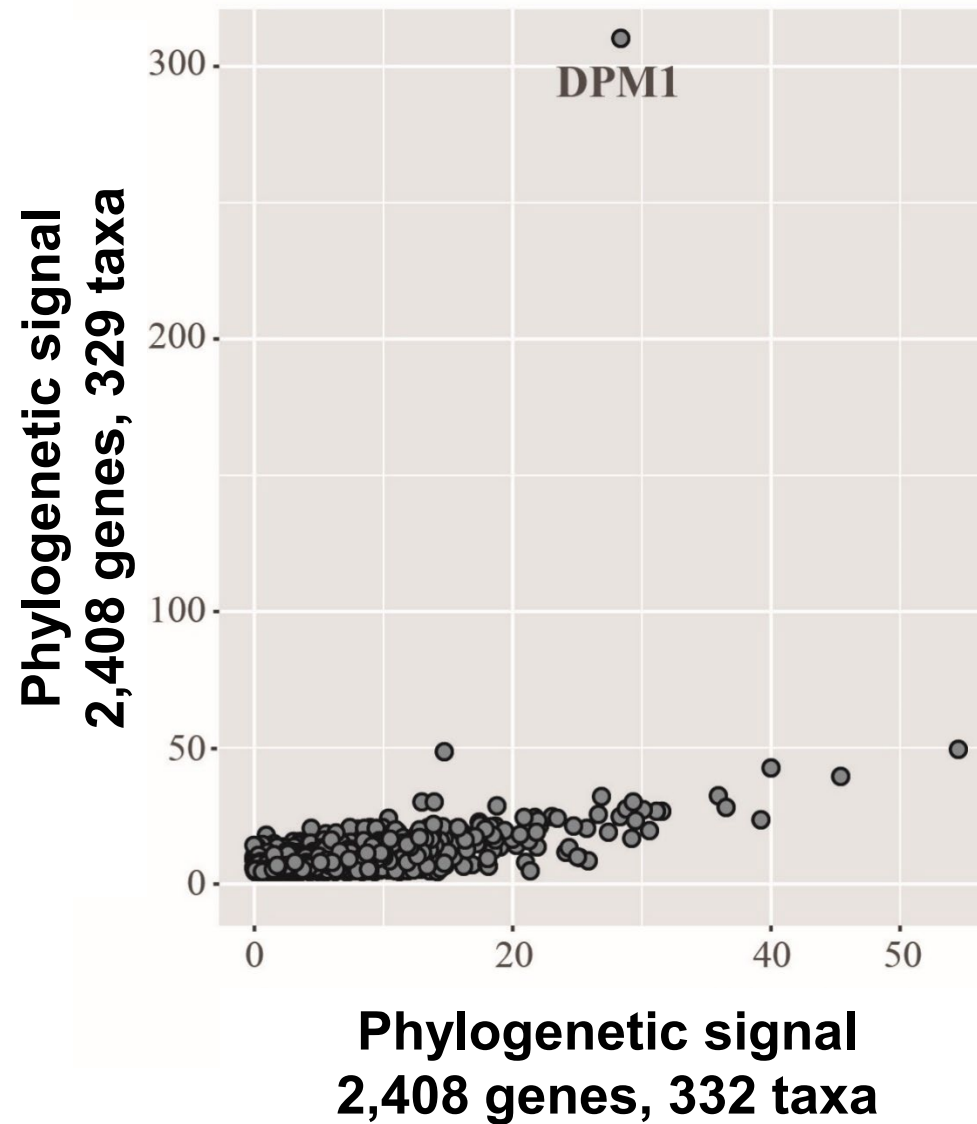
A single gene governs the placement of Ascoideaceae



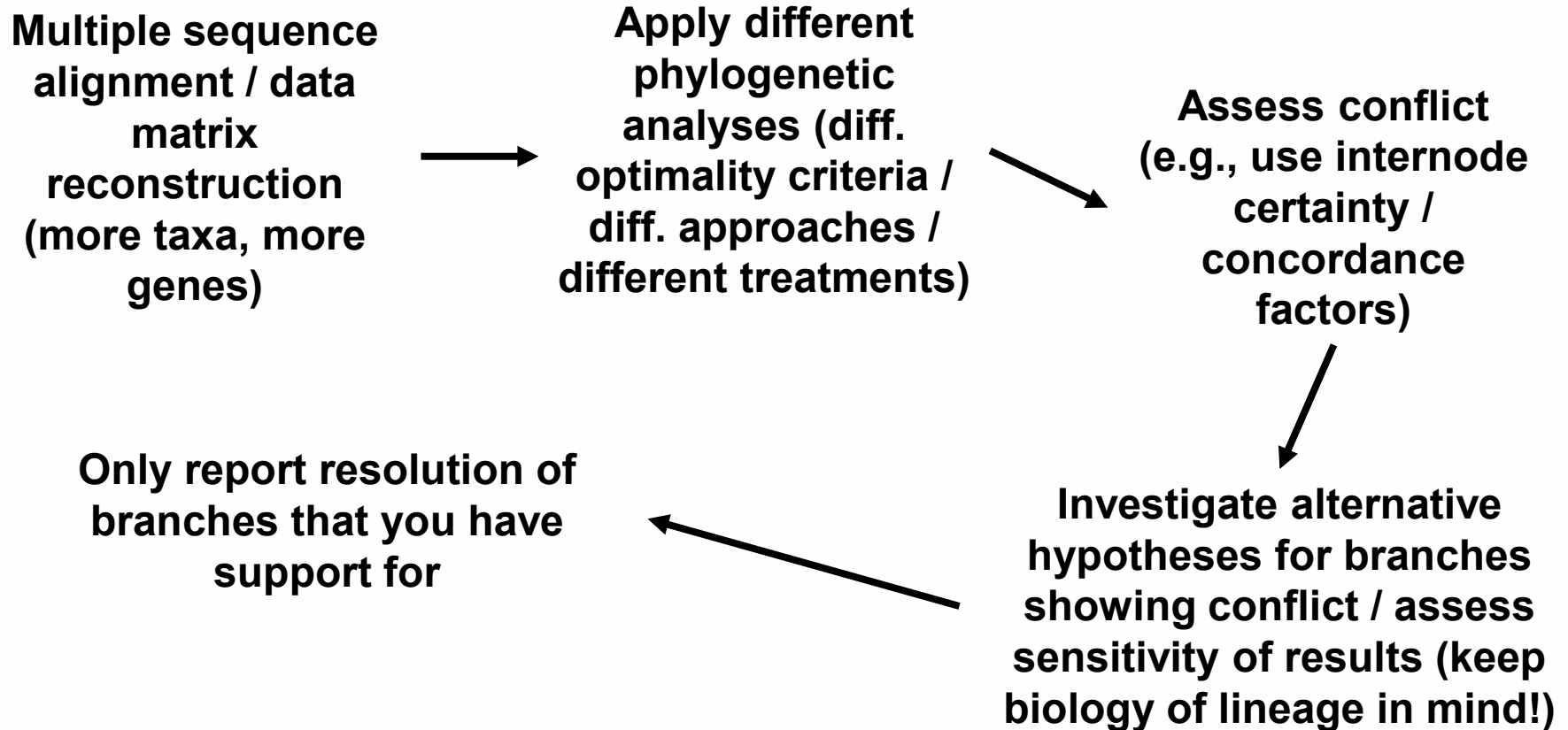
Sampling of 3 additional taxa “breaks” the long branch



Sampling of 3 additional taxa decreases gene's signal



Genomfart? The way forward?



Carefully analyze and listen to the data

“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



**Xing-Xing
Shen**



**Xiaofan
Zhou**



**Leonidas
Salichos**

NEW YORK INSTITUTE
OF TECHNOLOGY



**Jacob
Steenwyk**



**Alexis
Stamatakis**



National Science Foundation
WHERE DISCOVERIES BEGIN



**Chris Todd
Hittinger**



**Marizeth
Groenewald**



National Institute of
Allergy and
Infectious Diseases

BURROUGHS
WELLCOME
FUND 



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

@RokasLab