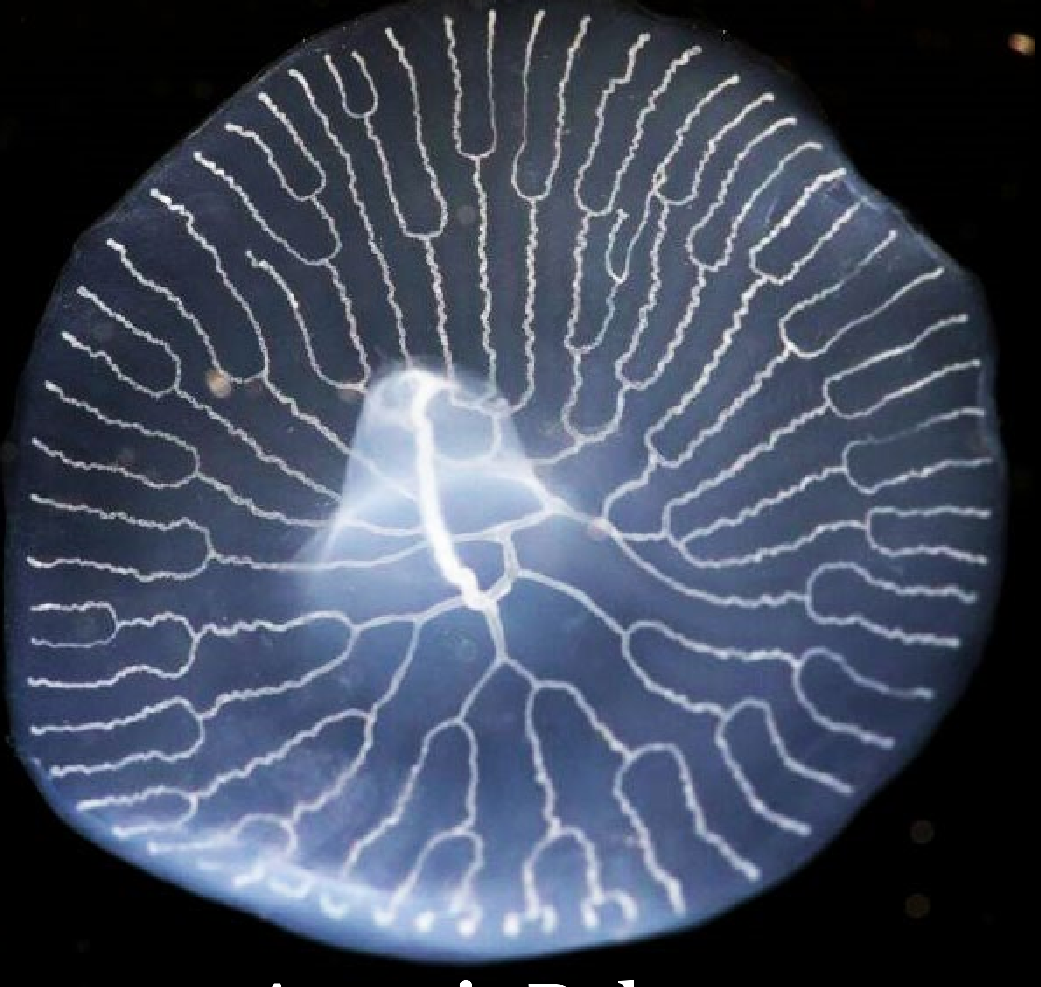
Incongruence



Antonis Rokas

Department of Biological Sciences, Vanderbilt University

http://www.rokaslab.org

@RokasLab

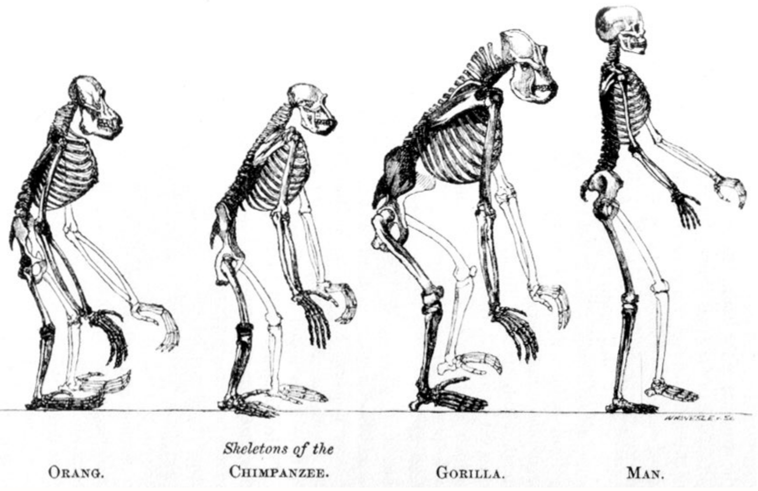
Brief Bio

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



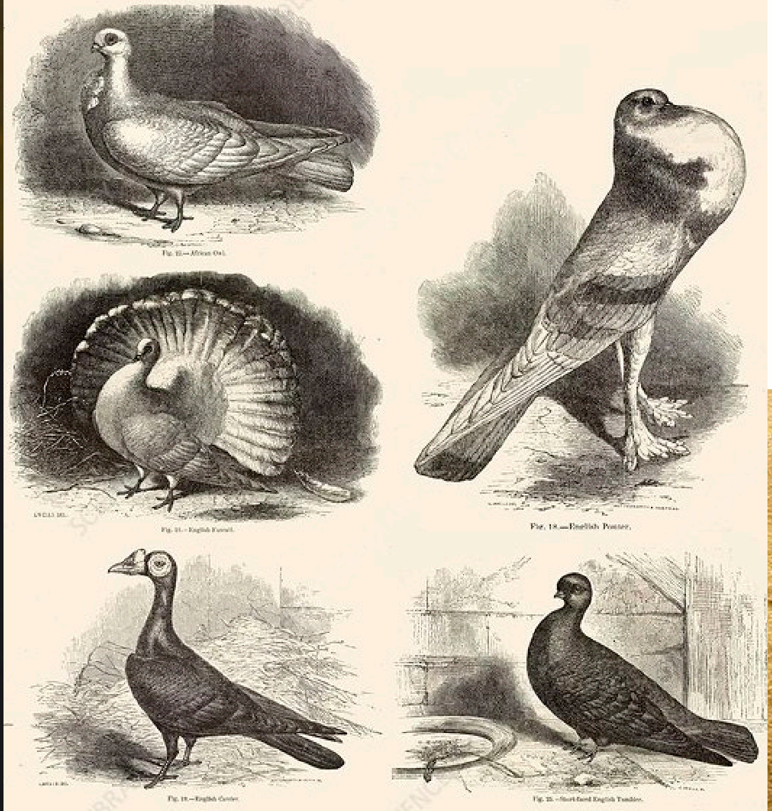
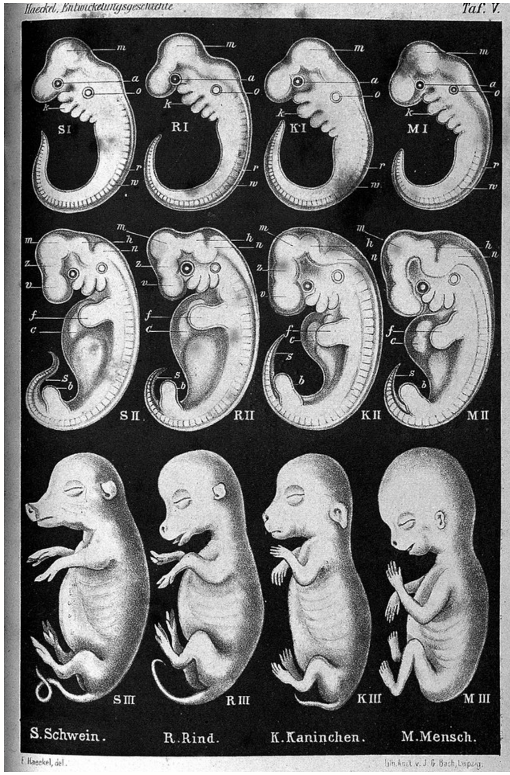
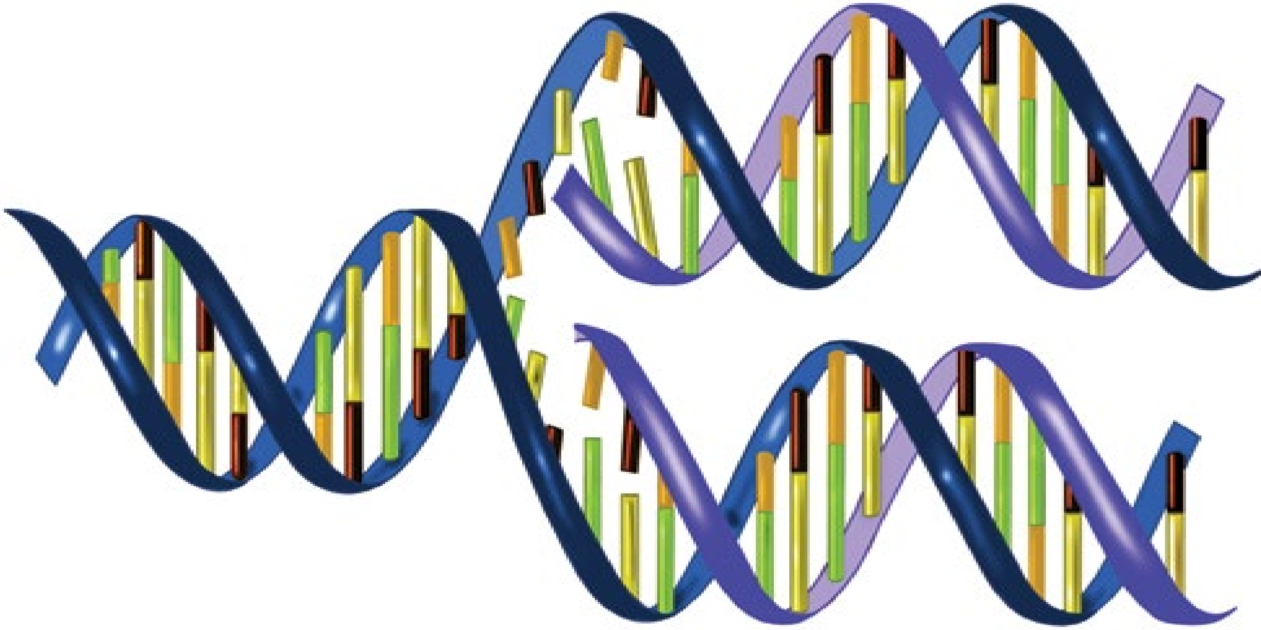


Fig. 12.—African Owl.
Fig. 21.—English Fantail.
Fig. 18.—English Pouter.
Fig. 19.—English Carrier.
Fig. 25.—Short-faced English Tumbler.



The DNA record



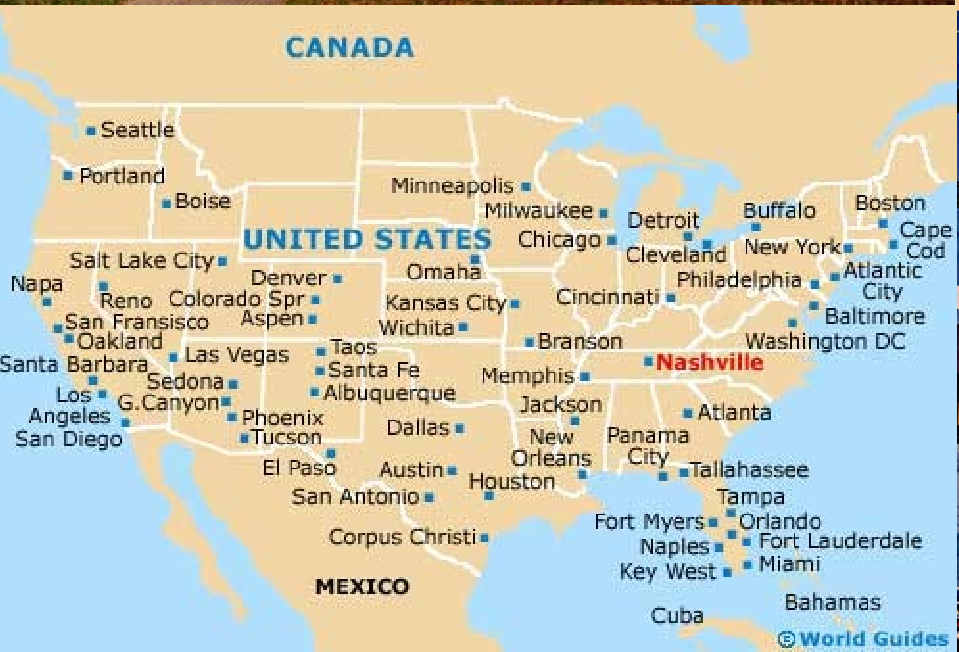
"The genome is, it's a fossil record; the genome is a landscape; the genome is a whole geography of distributions. […] you might think the genome's just a boring string of letters [...] The genome is a storybook that's been edited for a couple of billion years, and you could take it to bed, like *A Thousand 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



HATCH SHOW PRINT

224 5TH AVE. SOUTH - NASHVILLE, TENNESSEE - 615-256-2805

C.2015 HATCH SHOW PRINT

UNDER LICENSE WITH JOHN R. CASH REVOCABLE TRUST

JOHN PRINE & JASON ISBELL

WITH SPECIAL GUEST

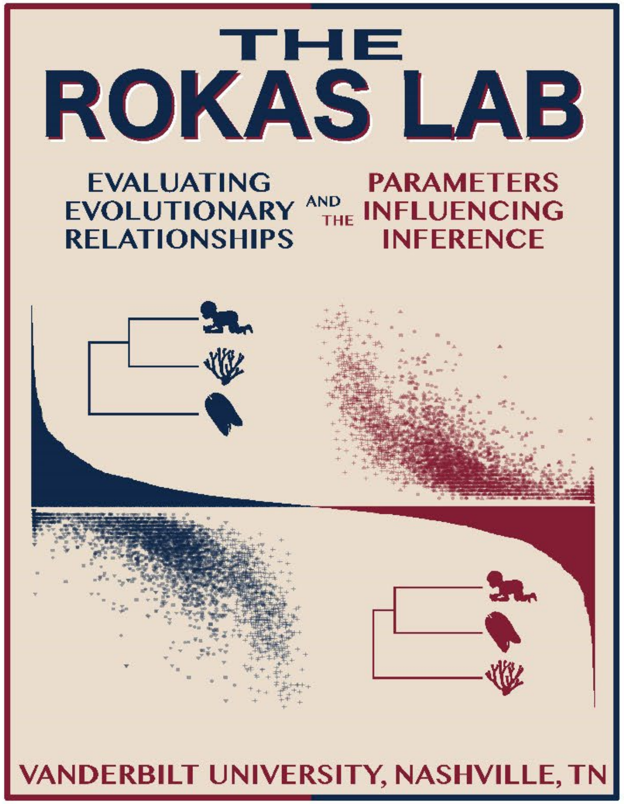
★ KACEY MUSGRAVES ★

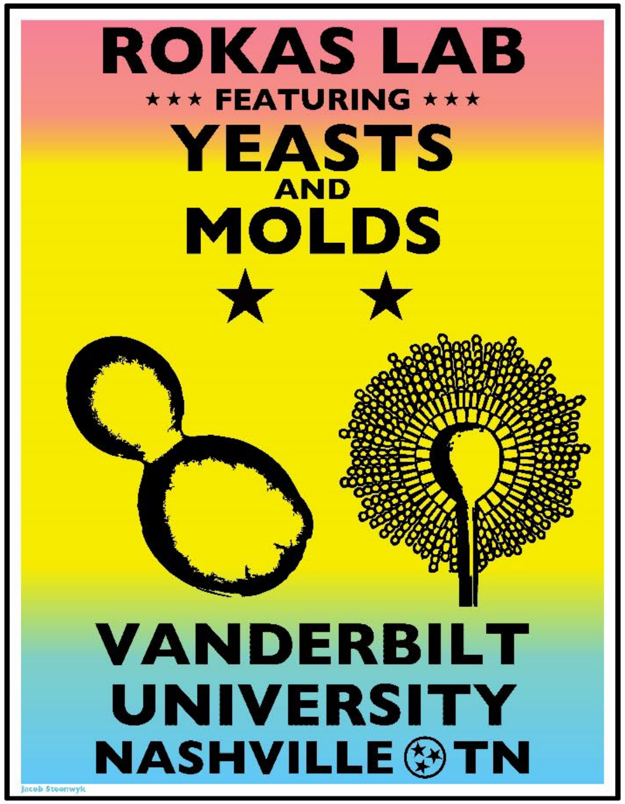
THE OPRY® HOUSE

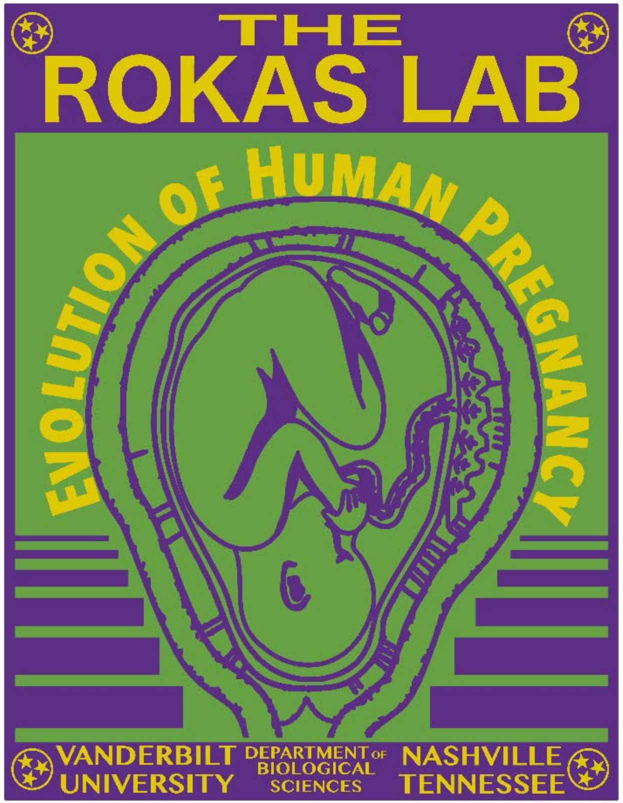
DECEMBER 31 2016

NASHVILLE TN

★ ©2016 Hatch Show Print ★


THE
ROKAS LAB
EVALUATING EVOLUTIONARY RELATIONSHIPS
AND THE
PARAMETERS INFLUENCING INFERENCE
VANDERBILT UNIVERSITY, NASHVILLE, TN




THE
ROKAS LAB
EVOLUTION OF HUMAN PREGNANCY
VANDERBILT UNIVERSITY
DEPARTMENT OF BIOLOGICAL SCIENCES
NASHVILLE TENNESSEE

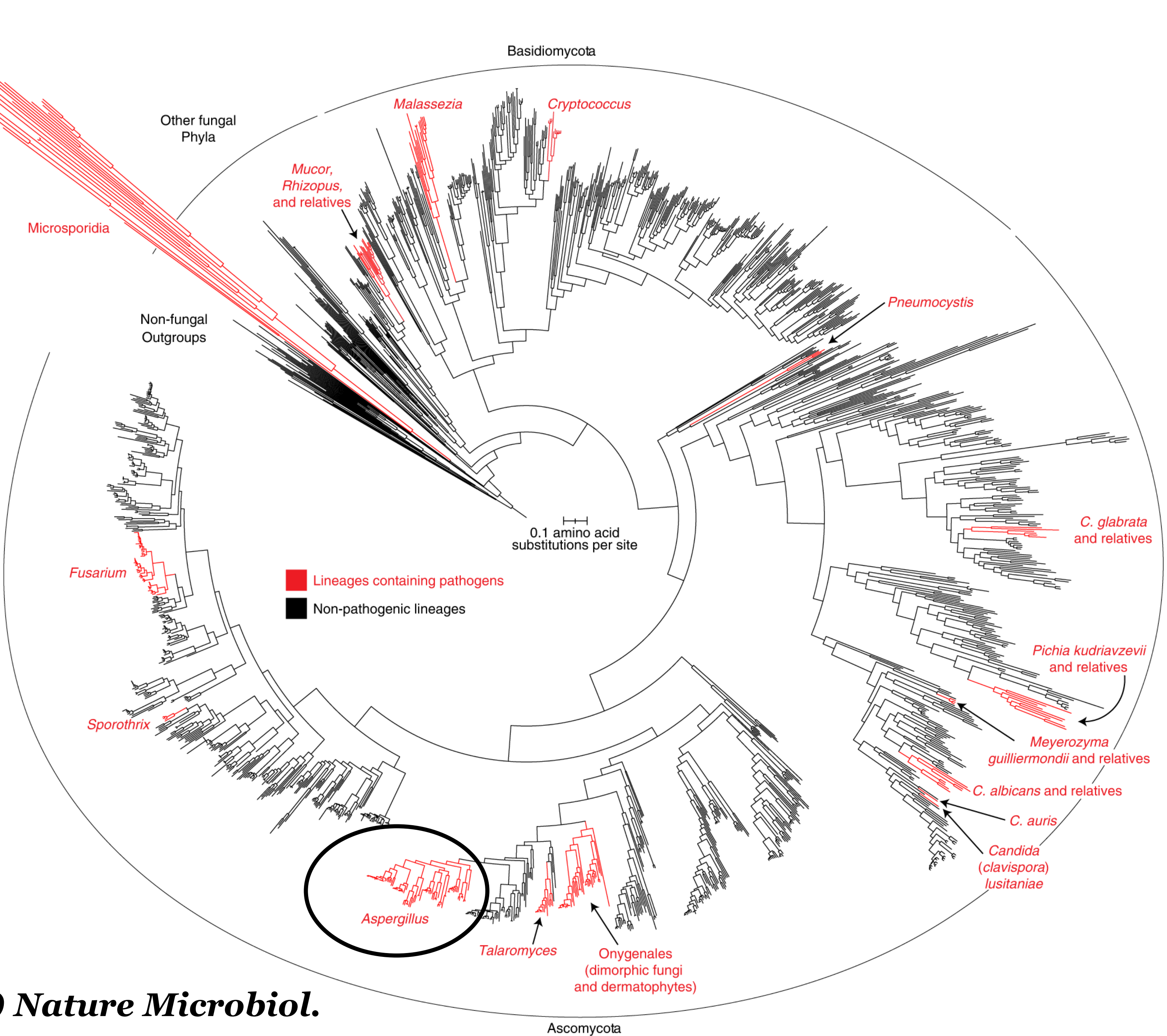
Images by Jacob Steenwyk

Phylogenomics
(NSF)

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

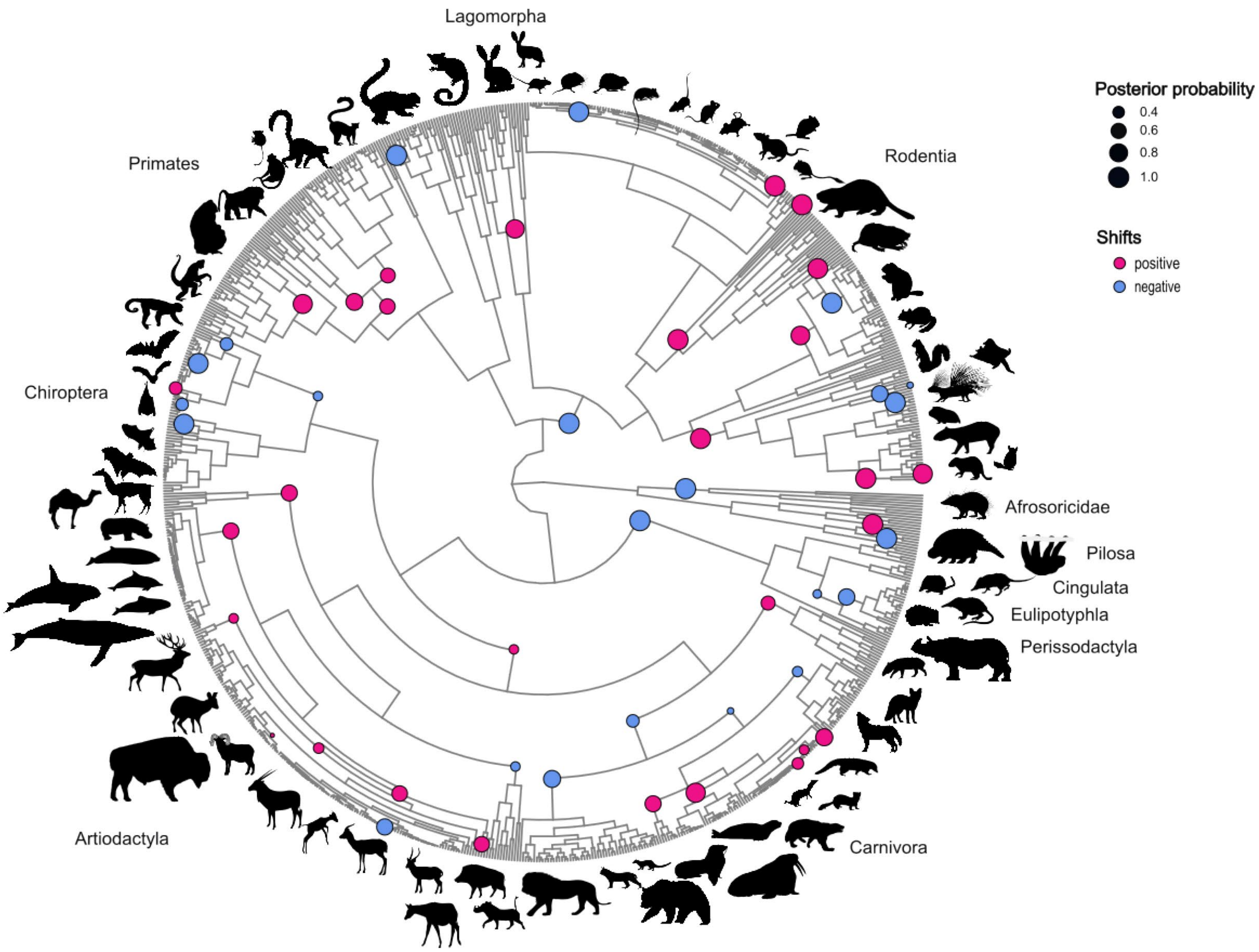
The evolution of mammalian pregnancy
(BWF & March of Dimes)

The repeated evolution of fungal pathogenicity


Basidiomycota
Other fungal Phyla
Malassezia
Cryptococcus
Mucor, Rhizopus, and relatives
Microsporidia
Non-fungal Outgroups
Pneumocystis
0.1 amino acid substitutions per site
C. glabrata and relatives
Fusarium
🟥 Lineages containing pathogens
⬛ Non-pathogenic lineages
Pichia kudriavzevii and relatives
Sporothrix
Meyerozyma guilliermondii and relatives
C. albicans and relatives
C. auris
Candida (clavispora) lusitaniae
Aspergillus
Talaromyces
Onygenales (dimorphic fungi and dermatophytes)
Ascomycota

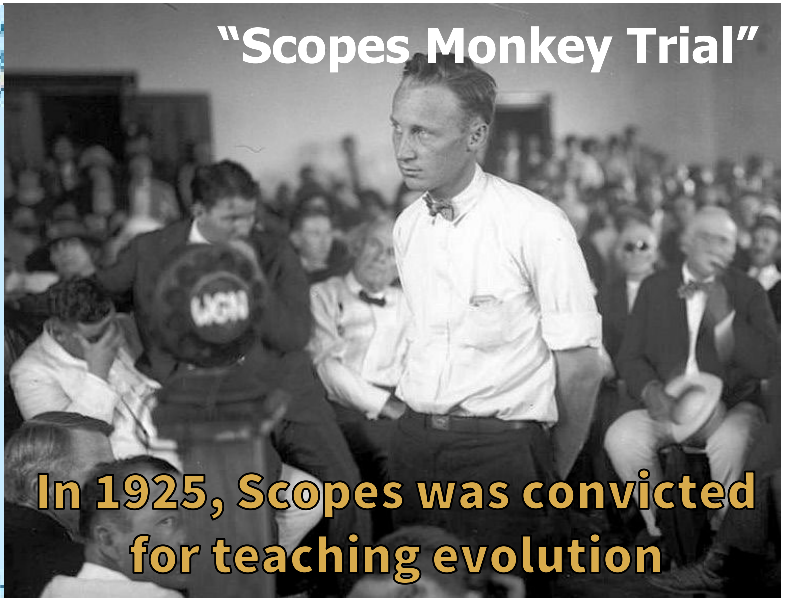
Rokas (2022) Nature Microbiol.

Evolution of variation in gestation length in mammals



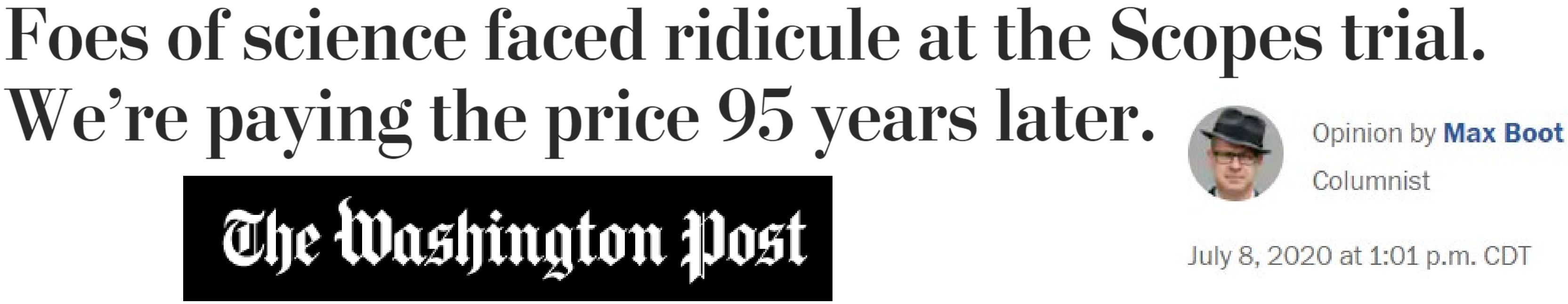
Danis & Rokas (2023) bioRxiv

…But I’m also an evolutionist in Tennessee (USA)



“Scopes Monkey Trial”

In 1925, Scopes was convicted for teaching evolution



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

EVOLUTIONARY STUDIES INITIATIVE

VANDERBILT UNIVERSITY®

www.vanderbilt.edu/evolution/

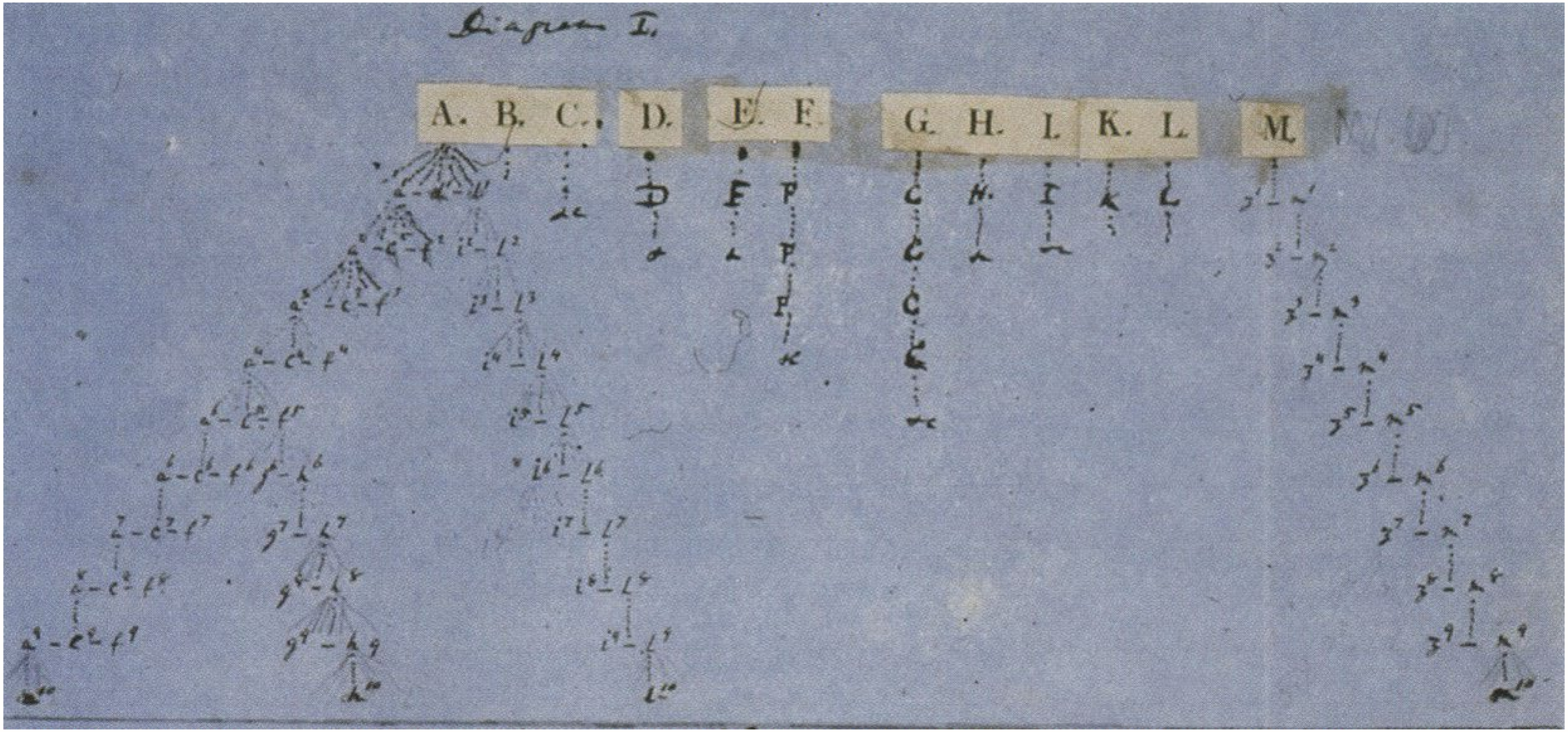
Lecture outline

- ❖ Incongruence and its causes

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

- ❖ Handling incongruence in phylogenomic data

Darwin's tree



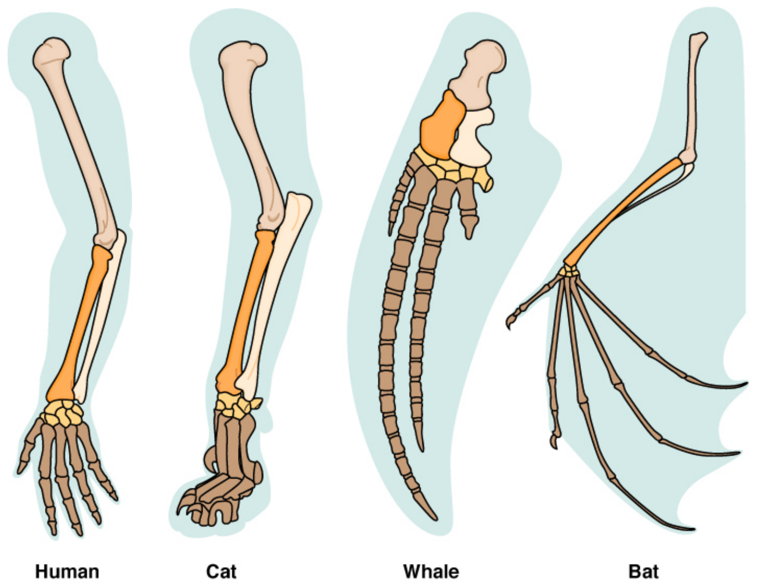
Darwin's hand-made proof of the famous diagram in his Origin of Species; Maderspacher (2006) Curr. Biol.

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!

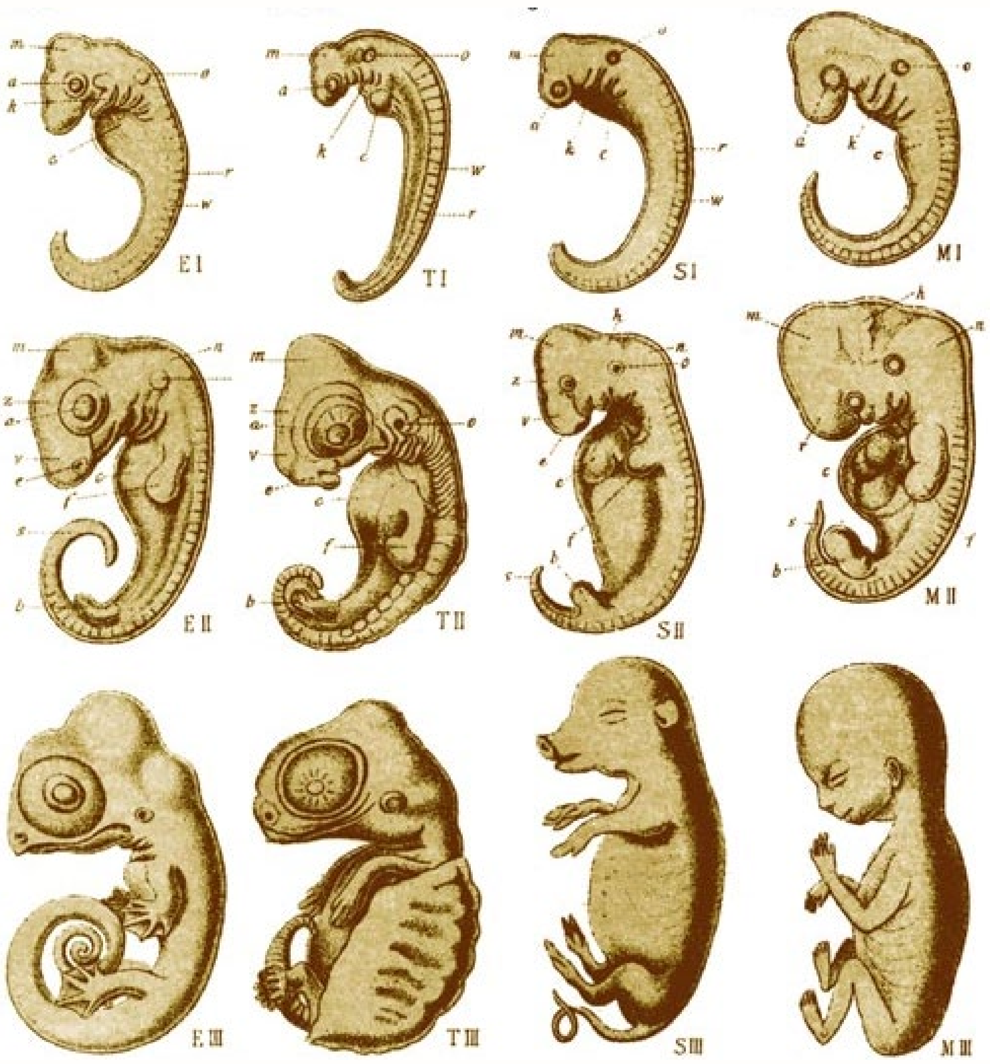
A grand and almost untrodden field of inquiry will be opened, on the causes and laws of variation, on correlation of growth, on the effects of use and disuse, on the direct action of external conditions, and so forth. The study of domestic productions will rise immensely in value. A new variety raised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

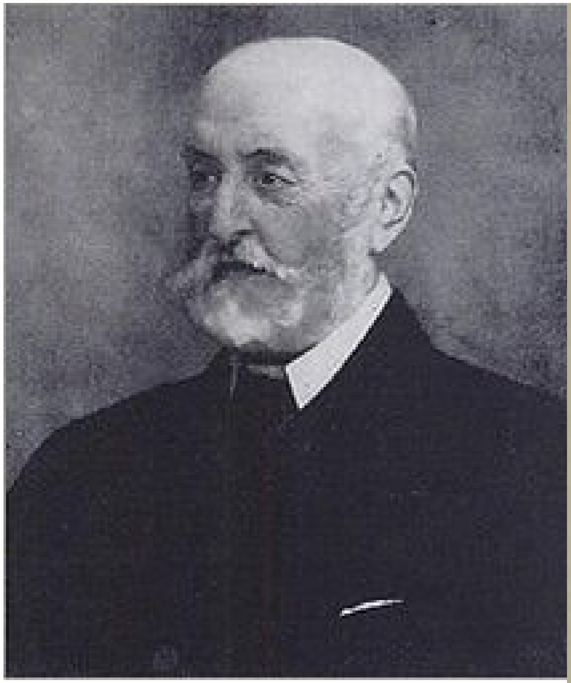


©1999 Addison Wesley Longman, Inc.



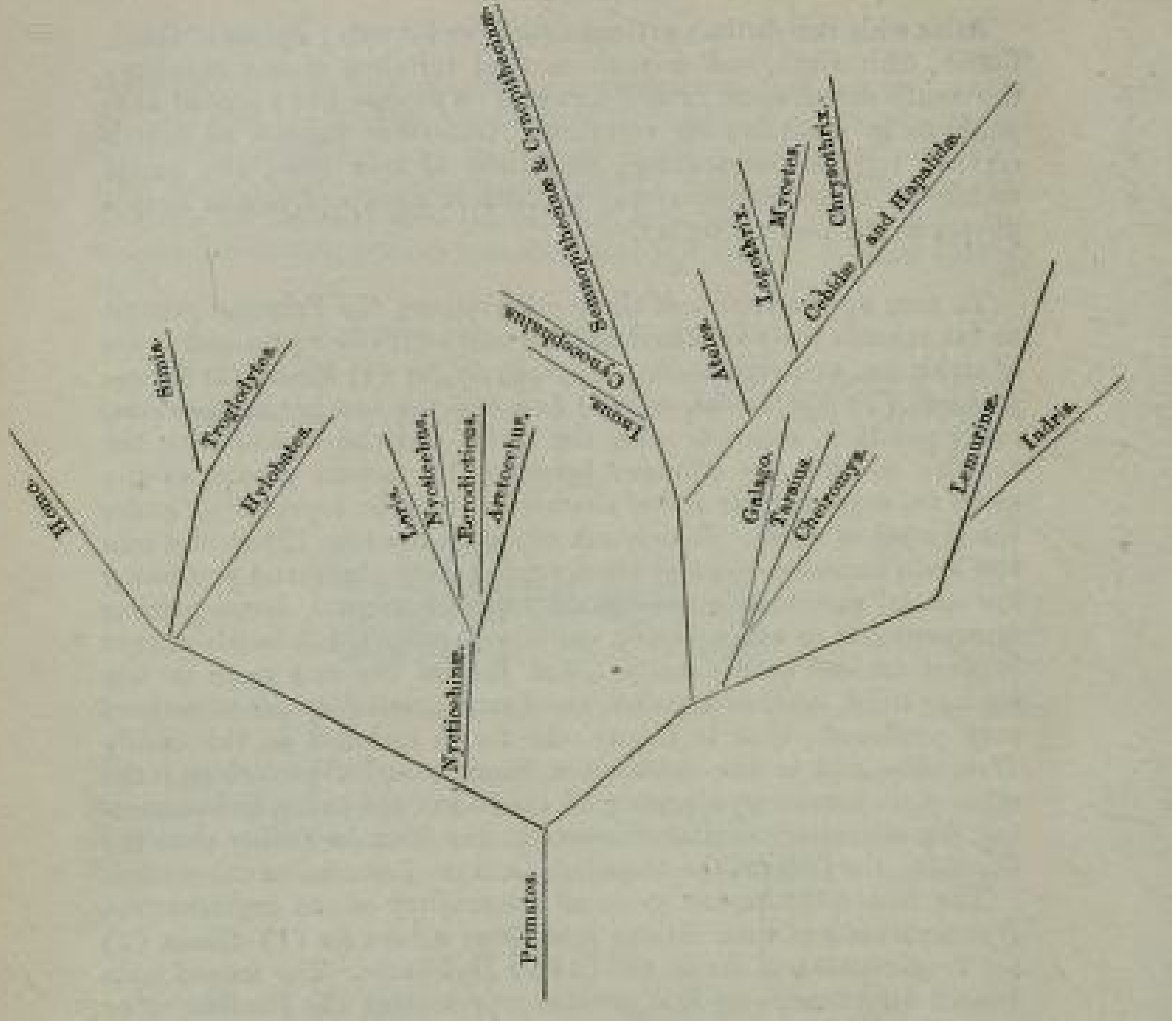
http://www.mun.ca/biology/scarr/139393_forelimb_homology.jpg &
http://www.nature.com/nrg/journal/v7/n11/images/nrg1918-f2.jpg

The first published phylogeny



St. George Jackson Mivart

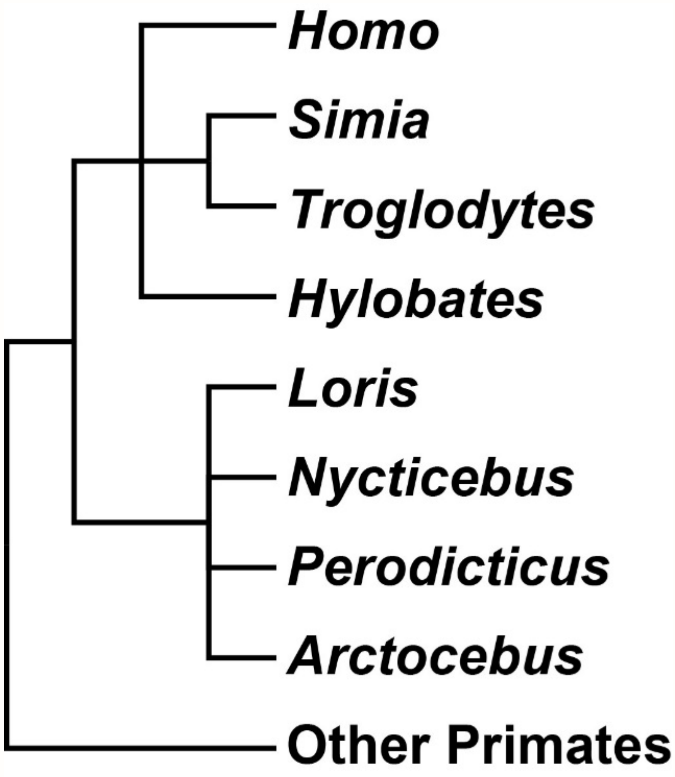
Fig. 13.



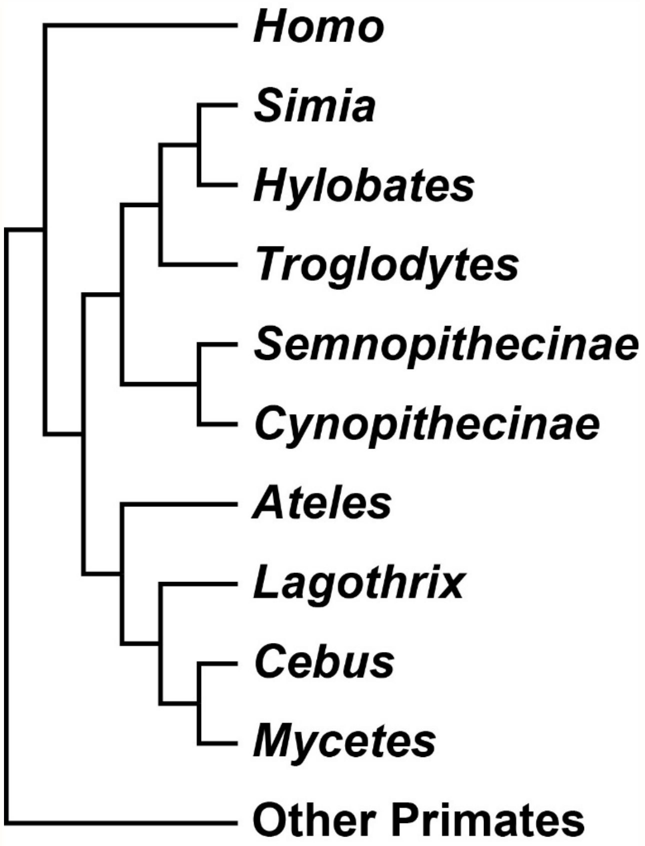
Mivart (1865) Proc. Zool. Soc. London

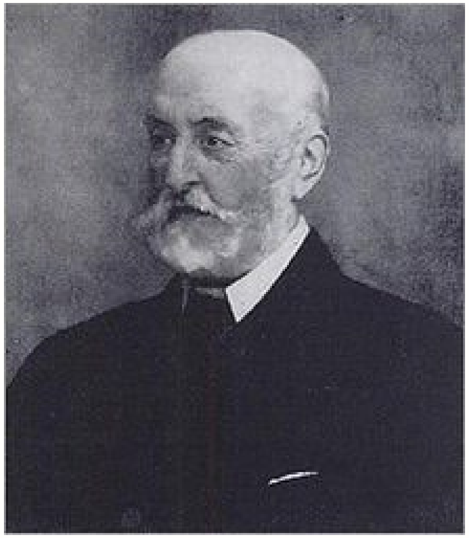
Disagreement between phylogenies

1865: SPINAL COLUMN



1867: LIMBS



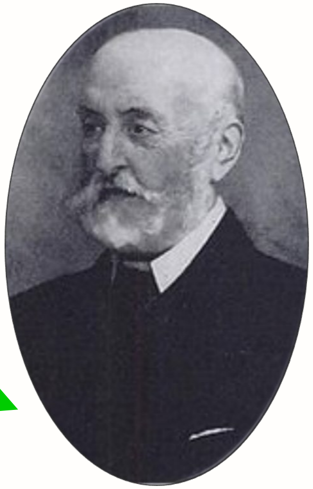


St. George Jackson Mivart

http://phylonetworks.blogspot.se/2012/09/the-first-network-from-conflicting.html

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*



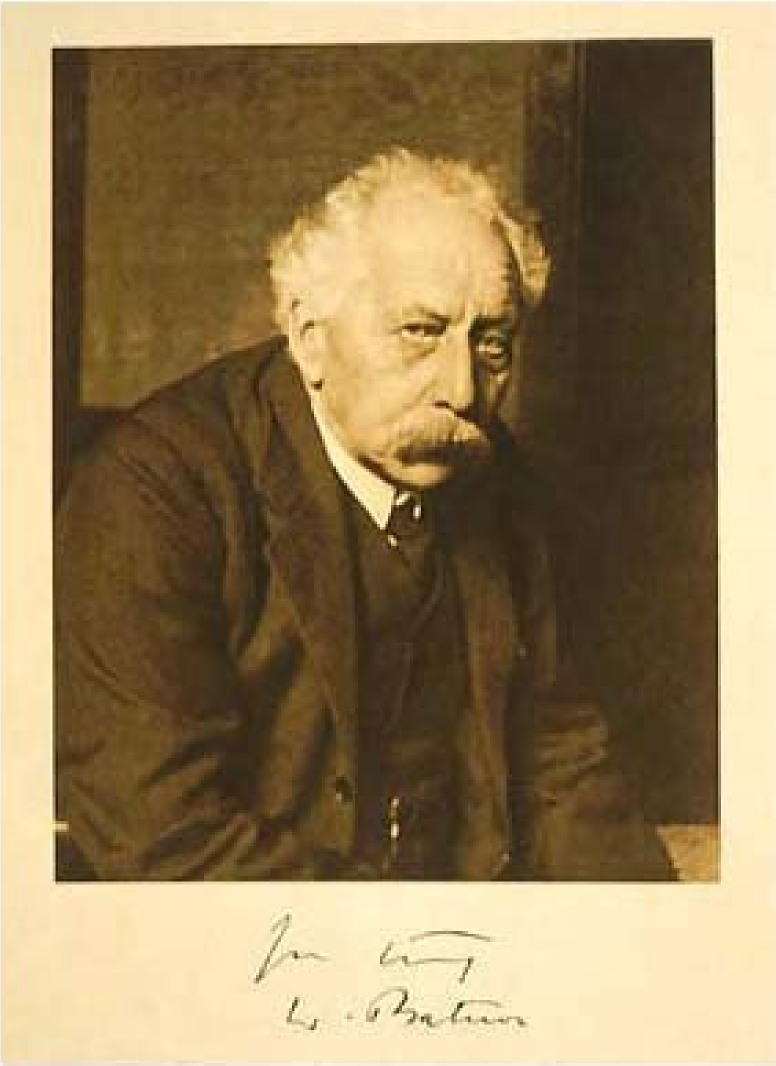
Darwin Correspondence Project letters 7718A & 7170

Inferences from comparative morphology / embryology data in trouble

"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."

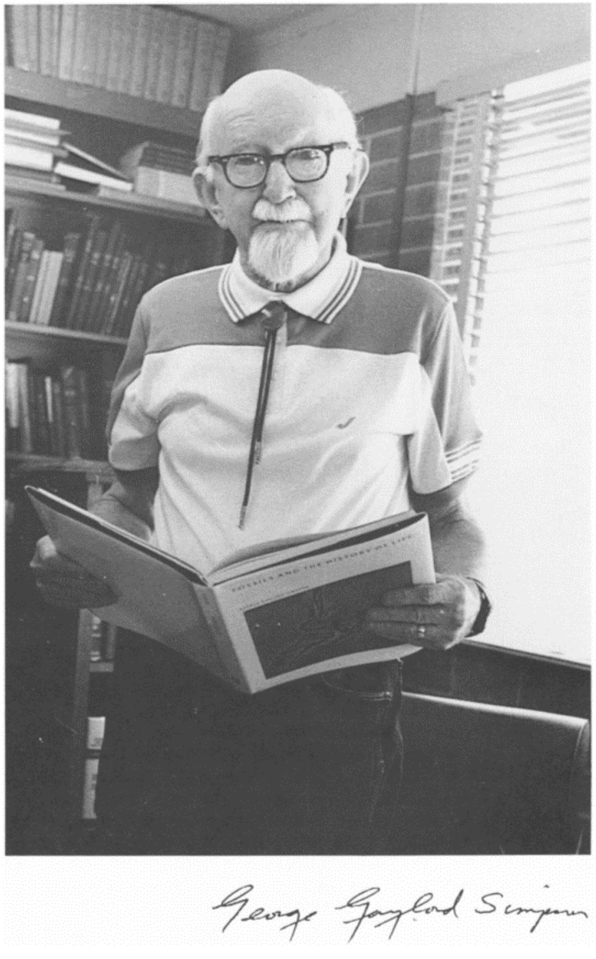
William Bateson (1894)

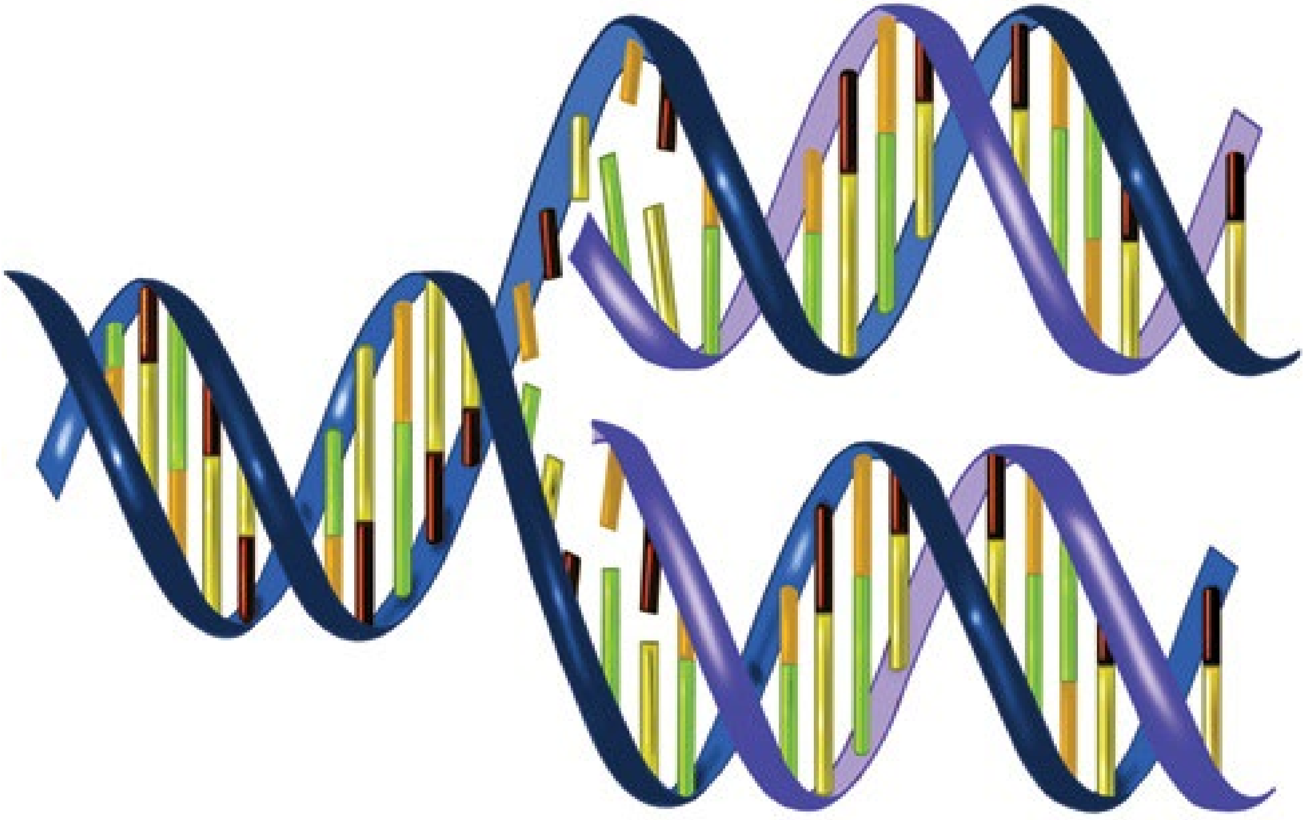
Materials for the Study of Variation



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

https://www.dnalc.org/view/16197-Gallery-5-William-Bateson-Portrait.html





“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

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

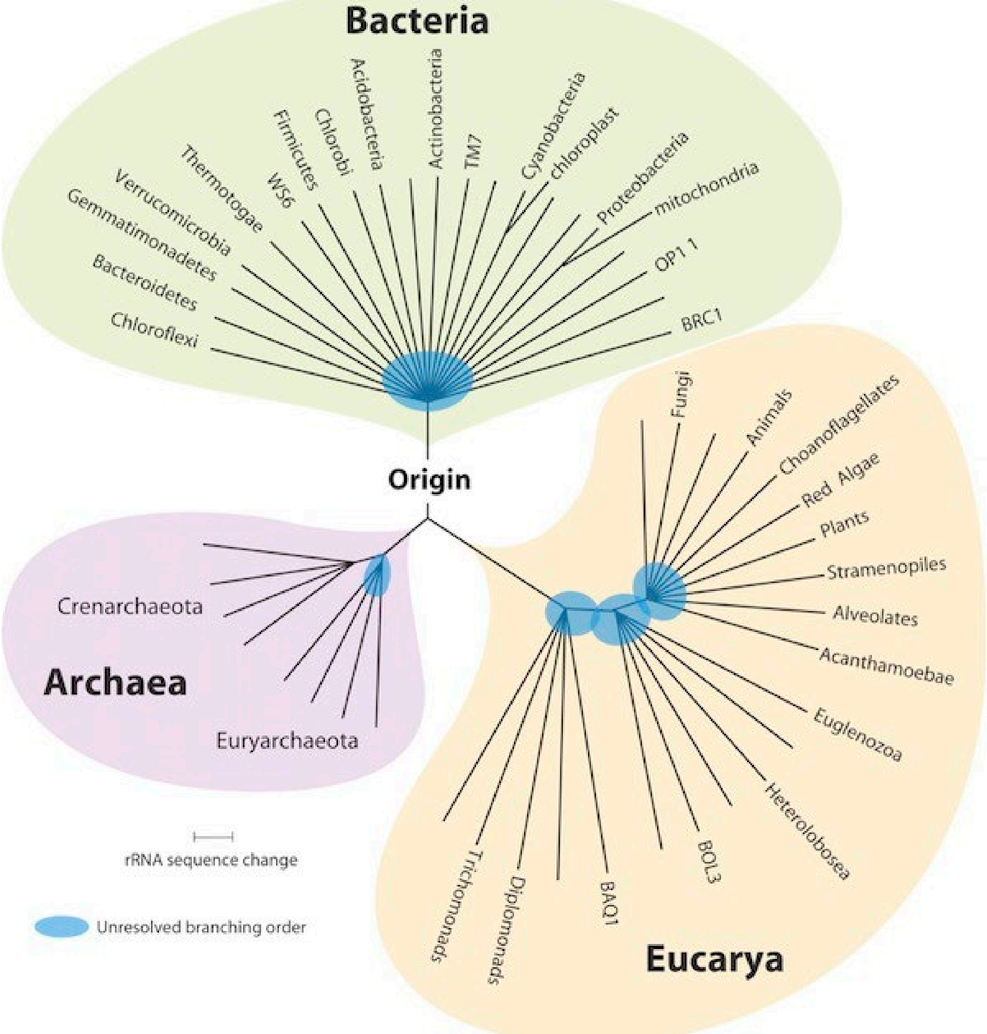
(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

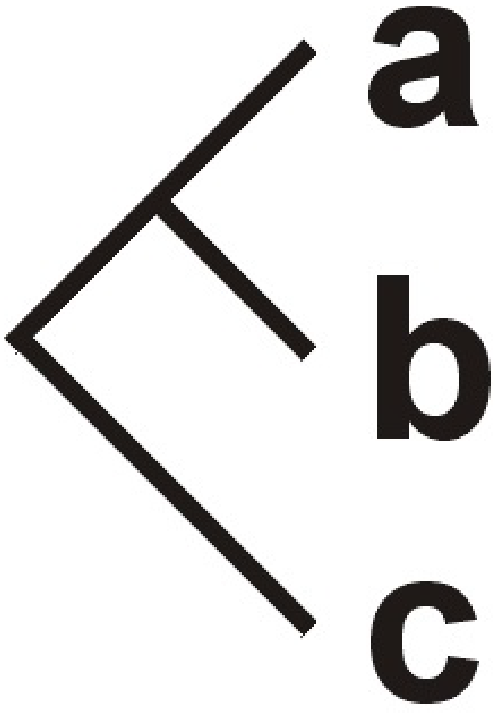
Department of Genetics and Development, University of Illinois, Urbana, Illinois

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

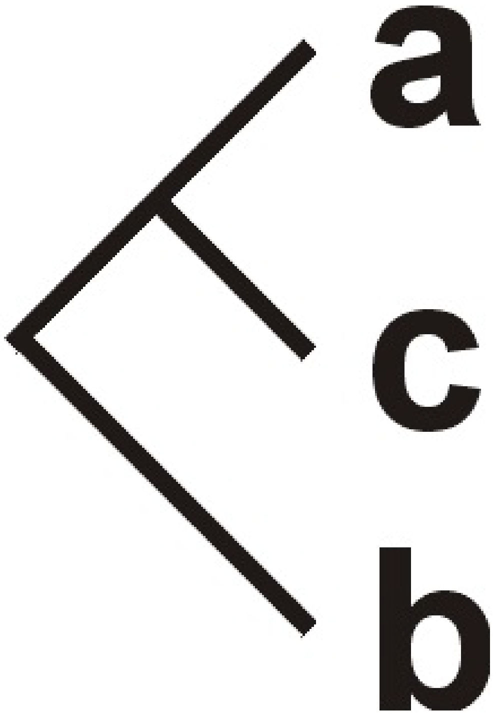
ABSTRACT A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (*i*) the eubacteria, comprising all typical bacteria; (*ii*) the archaebacteria, containing methanogenic bacteria; and (*iii*) the urkaryotes,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



Incongruence / conflict / discordance



Gene X

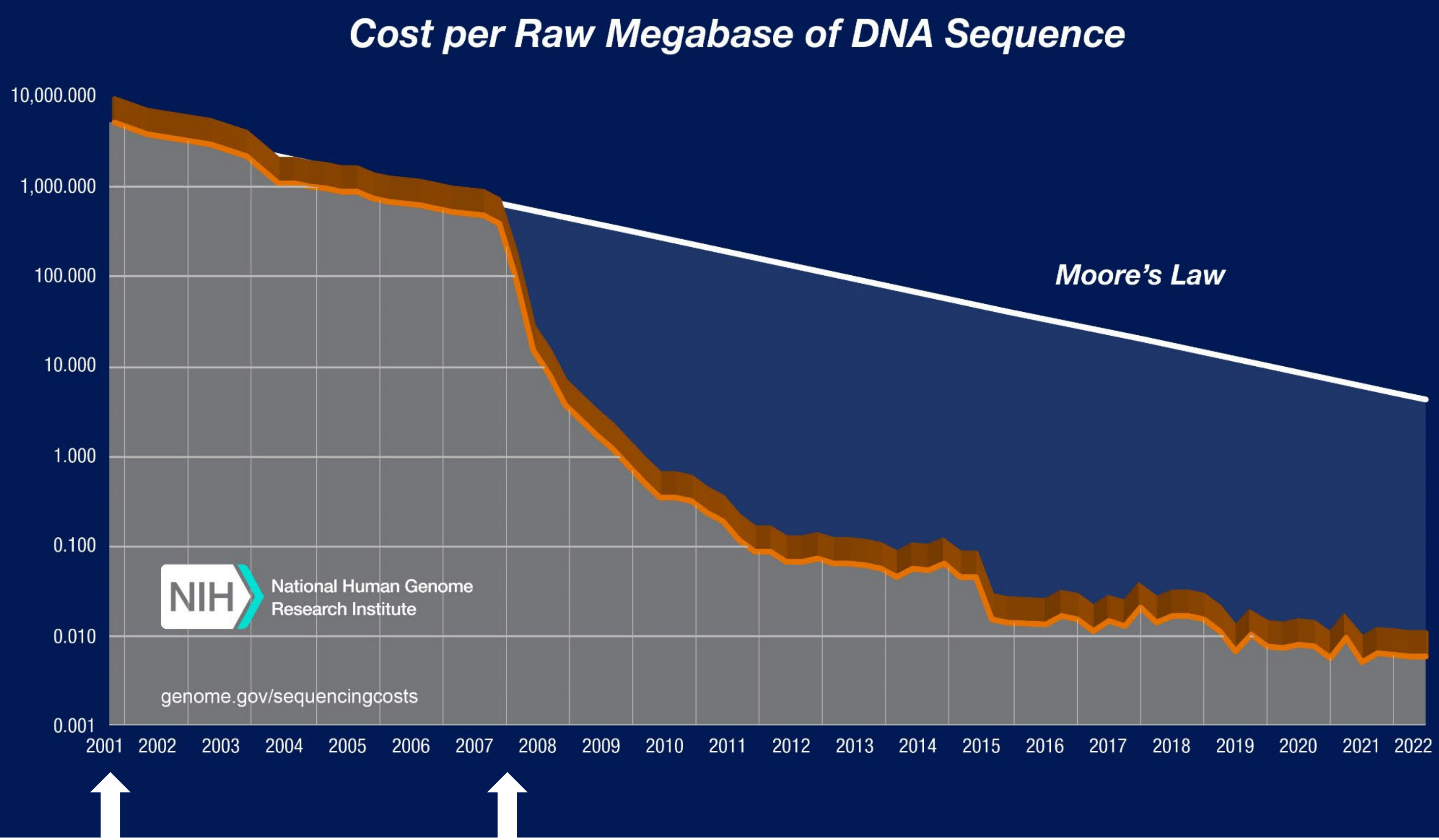


Gene Y

Species

phylogeny?

The genomics revolution



Cost per Raw Megabase of DNA Sequence

NIH National Human Genome Research Institute

genome.gov/sequencingcosts

PhD

Faculty

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

A systematic evaluation of single gene phylogenies

S. cerevisiae

S. paradoxus

S. mikatae

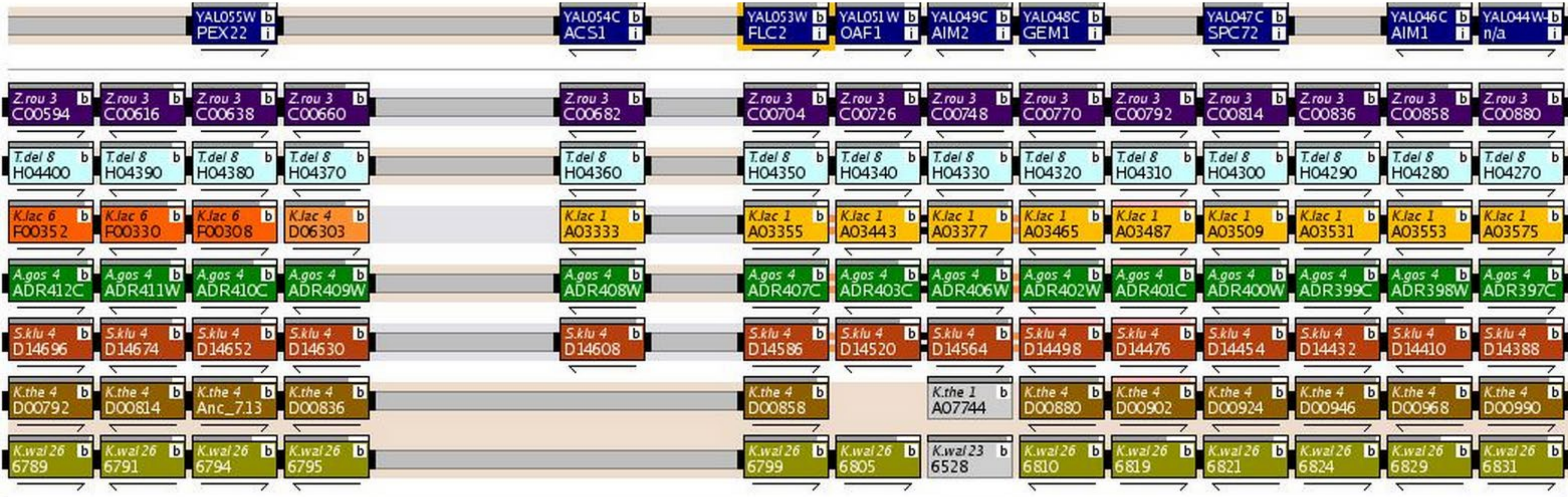
S. kudriavzevii

S. bayanus

S. castellii

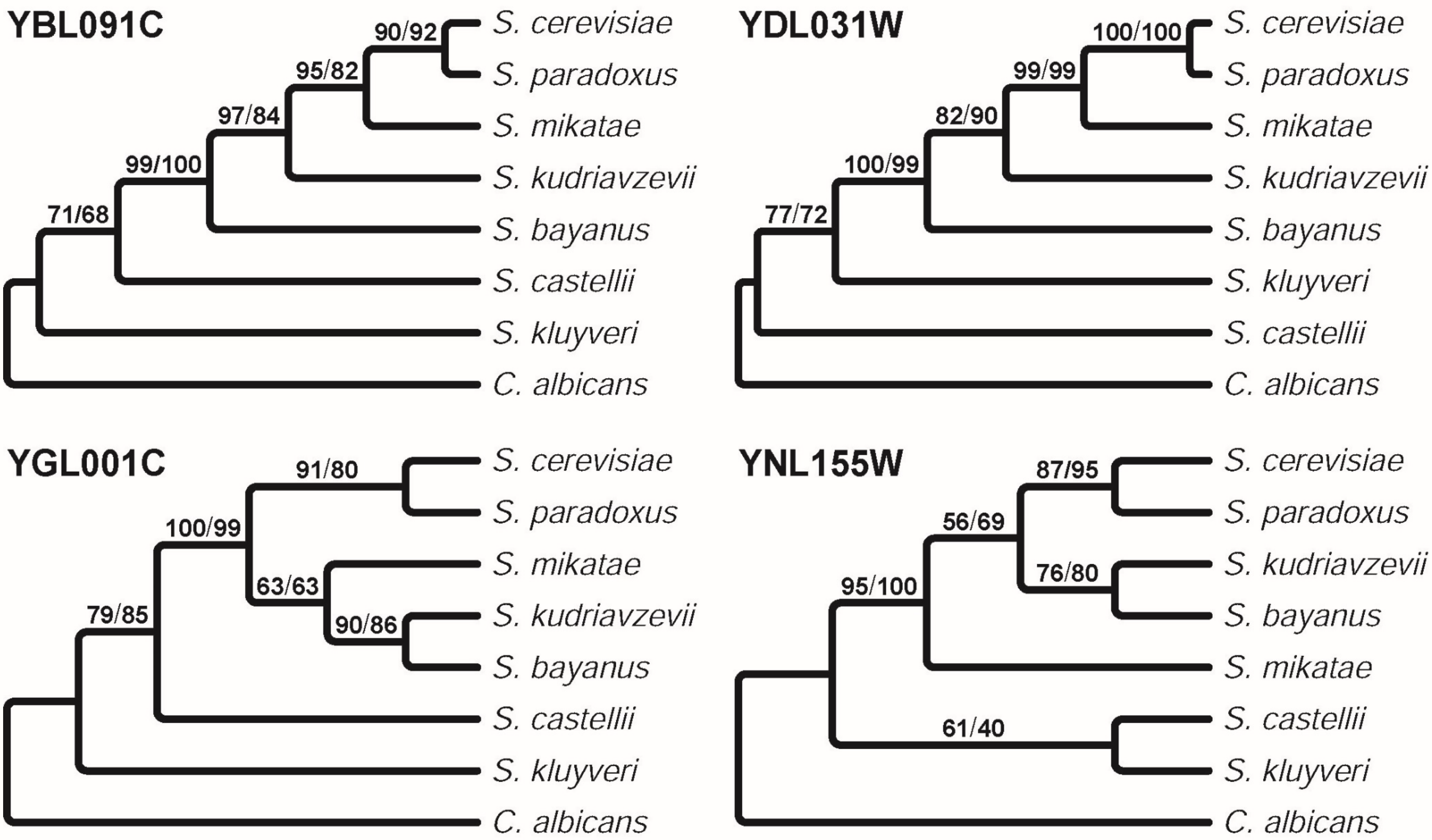
S. kluyveri

Candida glabrata



		YAL055W PEX22		YAL054C ACS1	YAL053W FLC2	YAL051W OAF1	YAL049C AIM2	YAL048C GEM1		YAL047C SPC72		YAL046C AIM1	YAL044W n/a
Z.rou 3 C00594	Z.rou 3 C00616	Z.rou 3 C00638	Z.rou 3 C00660	Z.rou 3 C00682	Z.rou 3 C00704	Z.rou 3 C00726	Z.rou 3 C00748	Z.rou 3 C00770	Z.rou 3 C00792	Z.rou 3 C00814	Z.rou 3 C00836	Z.rou 3 C00858	Z.rou 3 C00880
T.del 8 H04400	T.del 8 H04390	T.del 8 H04380	T.del 8 H04370	T.del 8 H04360	T.del 8 H04350	T.del 8 H04340	T.del 8 H04330	T.del 8 H04320	T.del 8 H04310	T.del 8 H04300	T.del 8 H04290	T.del 8 H04280	T.del 8 H04270
K.lac 6 F00352	K.lac 6 F00330	K.lac 6 F00308	K.lac 4 D06303	K.lac 1 A03333	K.lac 1 A03355	K.lac 1 A03443	K.lac 1 A03377	K.lac 1 A03465	K.lac 1 A03487	K.lac 1 A03509	K.lac 1 A03531	K.lac 1 A03553	K.lac 1 A03575
A.gos 4 ADR412C	A.gos 4 ADR411W	A.gos 4 ADR410C	A.gos 4 ADR409W	A.gos 4 ADR408W	A.gos 4 ADR407C	A.gos 4 ADR403C	A.gos 4 ADR406W	A.gos 4 ADR402W	A.gos 4 ADR401C	A.gos 4 ADR400W	A.gos 4 ADR399C	A.gos 4 ADR398W	A.gos 4 ADR397C
S.klu 4 D14696	S.klu 4 D14674	S.klu 4 D14652	S.klu 4 D14630	S.klu 4 D14608	S.klu 4 D14586	S.klu 4 D14520	S.klu 4 D14564	S.klu 4 D14498	S.klu 4 D14476	S.klu 4 D14454	S.klu 4 D14432	S.klu 4 D14410	S.klu 4 D14388
K.the 4 D00792	K.the 4 D00814	K.the 4 Anc_7.13	K.the 4 D00836		K.the 4 D00858		K.the 1 A07744	K.the 4 D00880	K.the 4 D00902	K.the 4 D00924	K.the 4 D00946	K.the 4 D00968	K.the 4 D00990
K.wal 26 6789	K.wal 26 6791	K.wal 26 6794	K.wal 26 6795		K.wal 26 6799	K.wal 26 6805	K.wal 23 6528	K.wal 26 6810	K.wal 26 6819	K.wal 26 6821	K.wal 26 6824	K.wal 26 6829	K.wal 26 6831

Incongruence at the single gene level

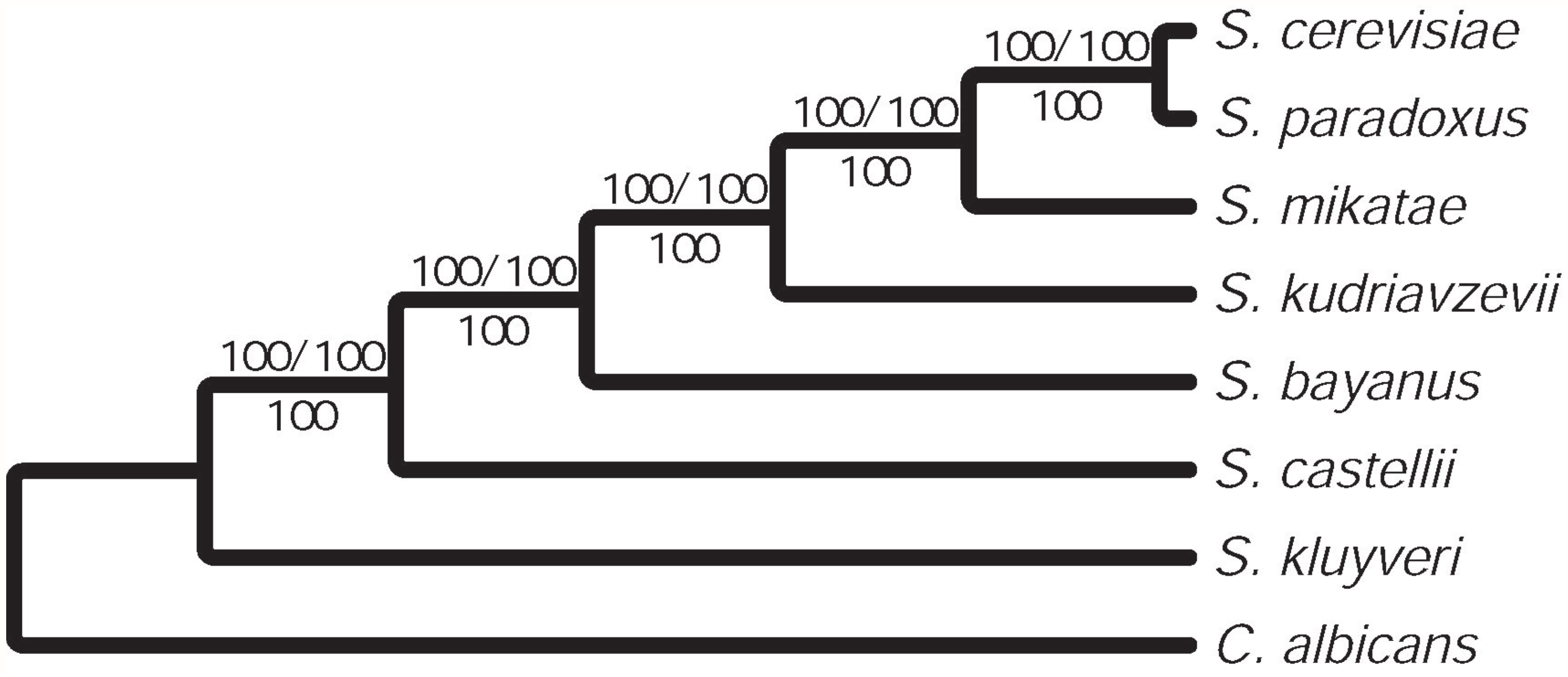


ML / MP

Anonymous Reviewer for Nature (2003)

Rokas et al. (2003) Nature

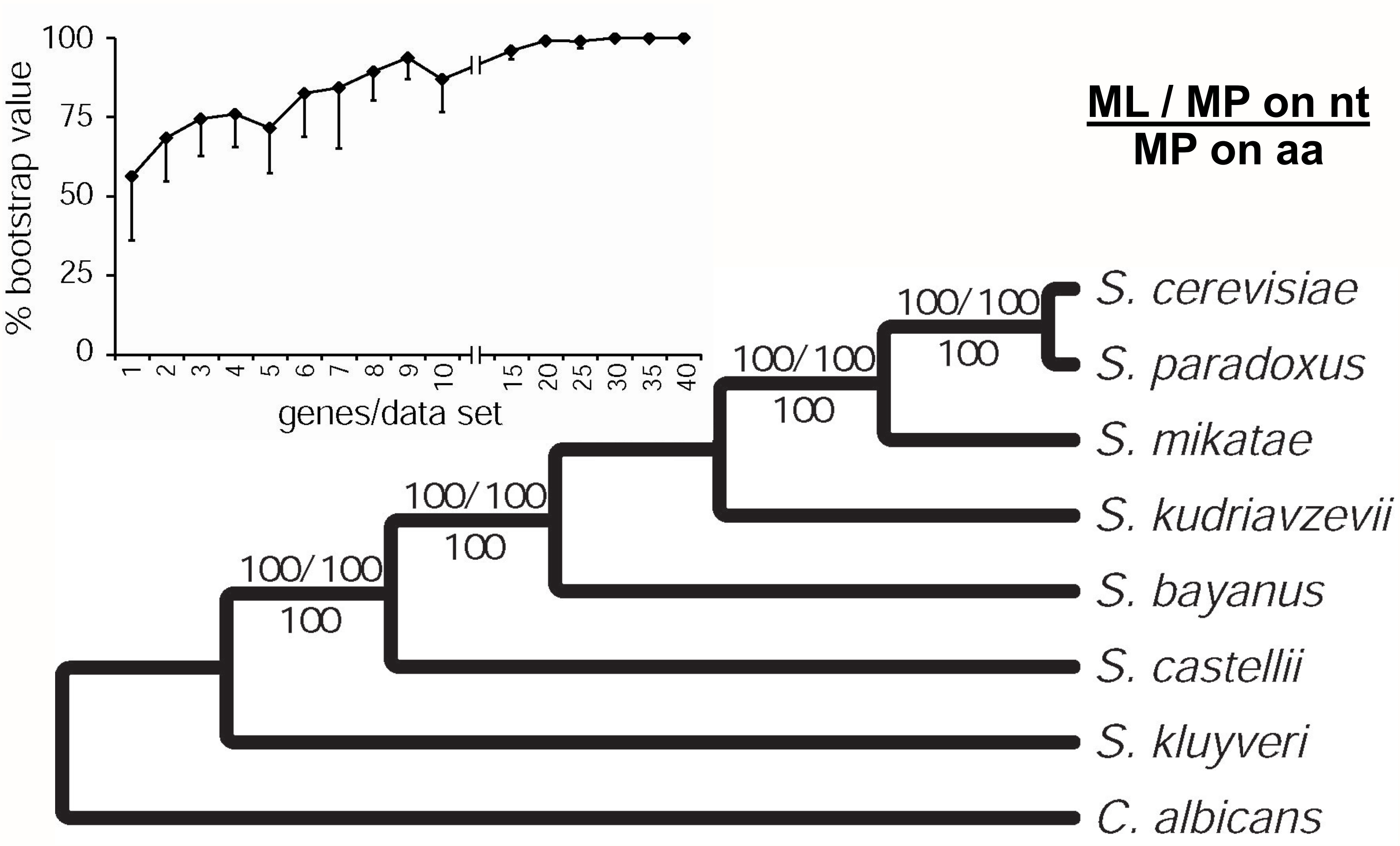
Concatenation of 106 genes yields a single species phylogeny



<u>ML / MP on nt</u>
MP on aa

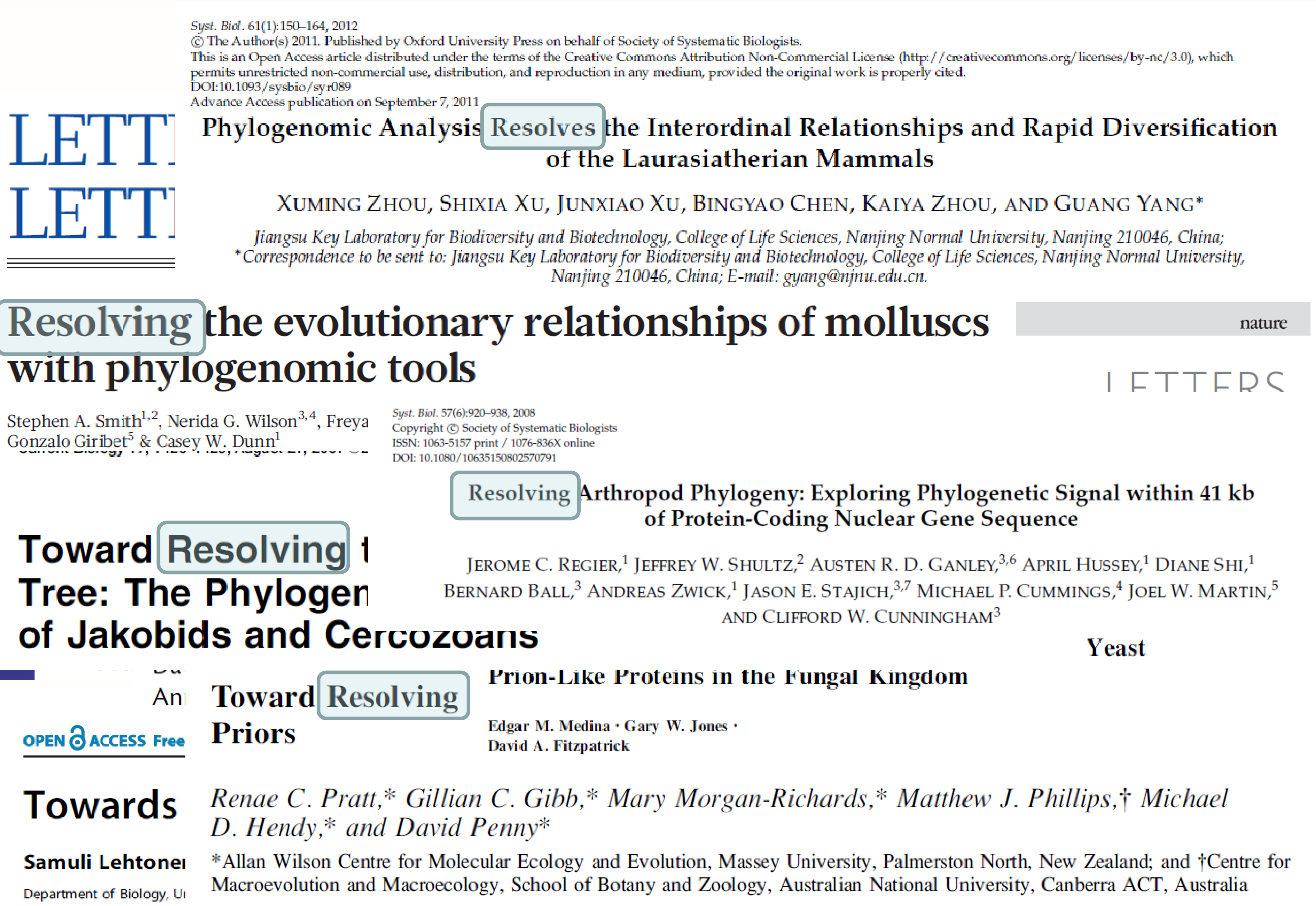
Rokas et al. (2003) Nature

The use of many genes eliminates incongruence



Rokas et al. (2003) Nature

The dawn of the phylogenomics era



Syst. Biol. 61(1):150–164, 2012

DOI:10.1093/sysbio/syr089
Advance Access publication on September 7, 2011

LETT

LETT

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

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

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

Resolving the evolutionary relationships of molluscs with phylogenomic tools

nature

LETTERS

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

Current Biology 17, 1420–1425, August 21, 2007 ©2

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. CUNNINGHAM[3]

Toward Resolving t
Tree: The Phylogen
of Jakobids and Cercozoans

Yeast

Prion-Like Proteins in the Fungal Kingdom

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

An

OPEN ACCESS Free

Toward Resolving Priors

Towards

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

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

Samuli Lehtonen

Department of Biology, U

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**

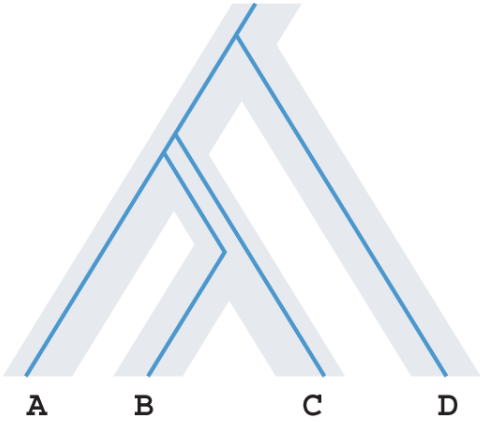
Analytical factors

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)

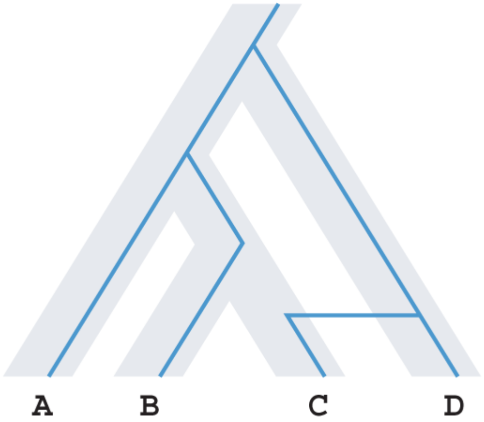
Steenwyk et al. (2023) Nature Rev. Genet.

Biological factors

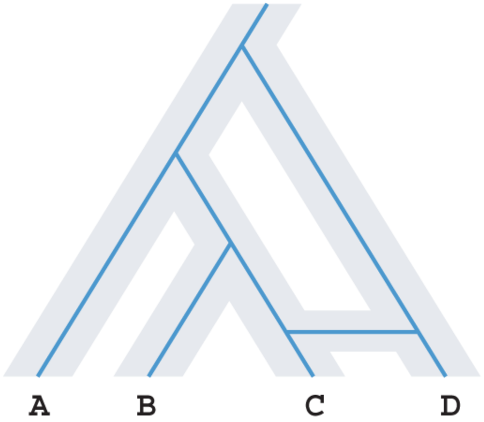
Lineage sorting



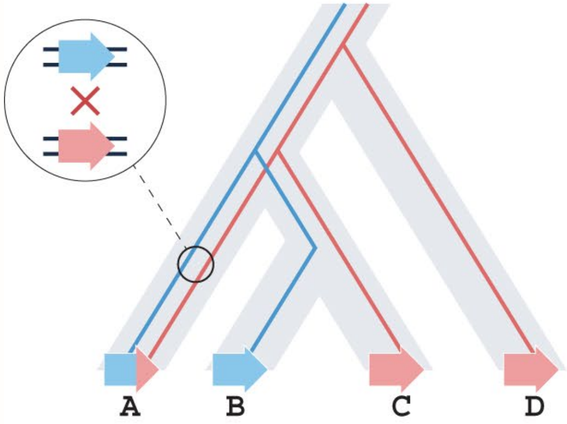
Horizontal transfer



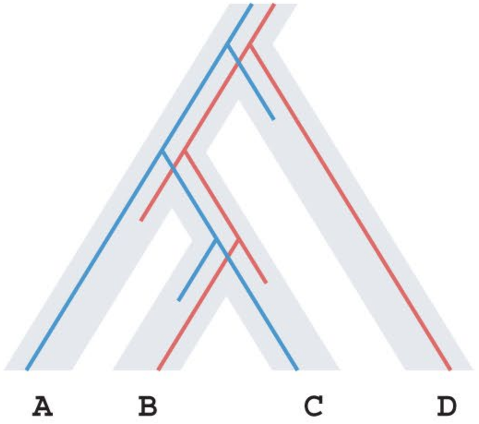
Hybridization / Introgression



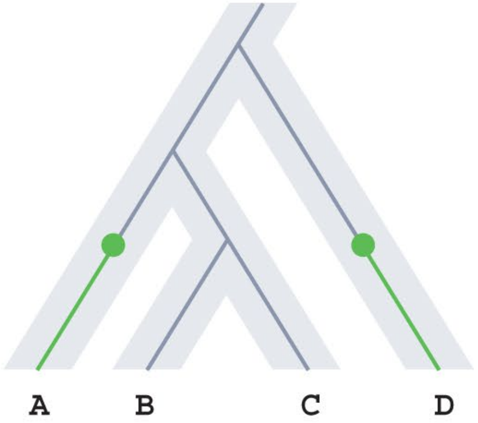
Recombination



Duplication & Loss

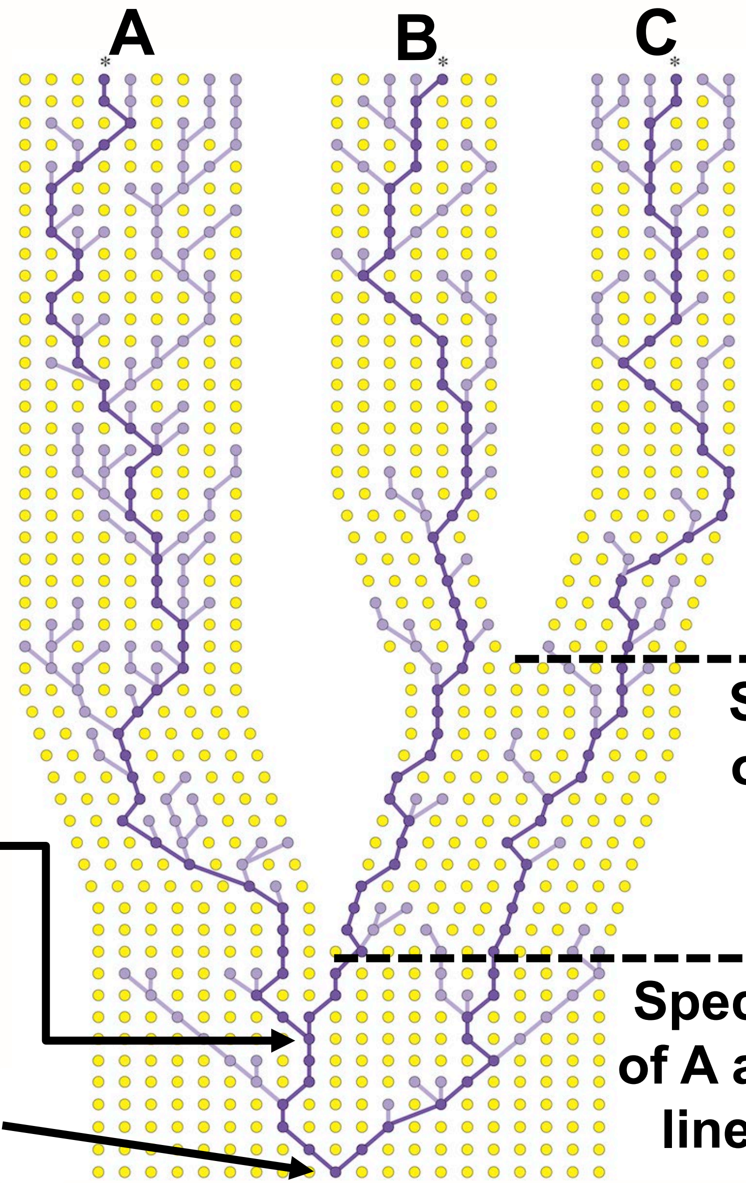


Selection



Steenwyk et al. (2023) Nature Rev. Genet.

Stochastic lineage sorting of ancestral polymorphisms



Genes' histories can differ from species ones

Speciation of B and C lineages

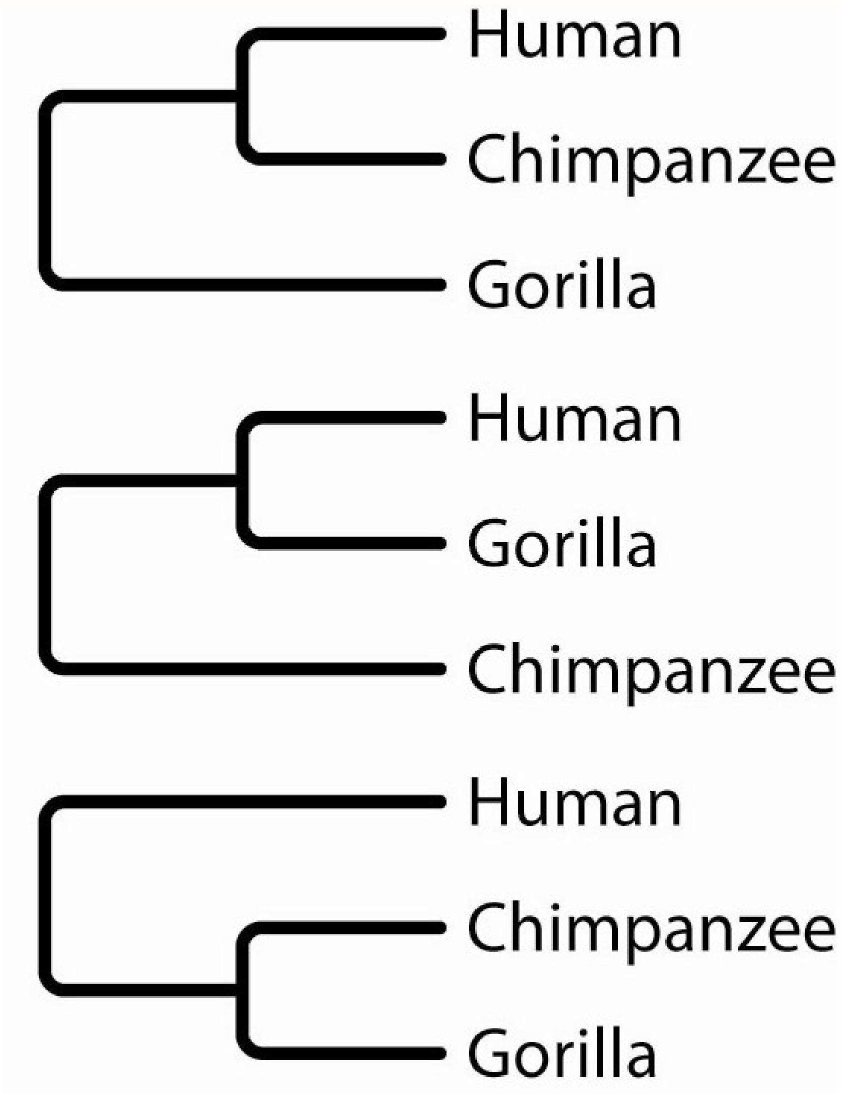
Speciation of A and BC lineages

Splitting of A and B alleles

Splitting of AB and C alleles

Nei (1987) Mol. Evol. Genet.; Maddison (1997) Syst. Biol.

Lineage sorting in primates



Informative Sites

8,561 / 11,293

(~76%)

1,302 / 11,293

(~11.5%)

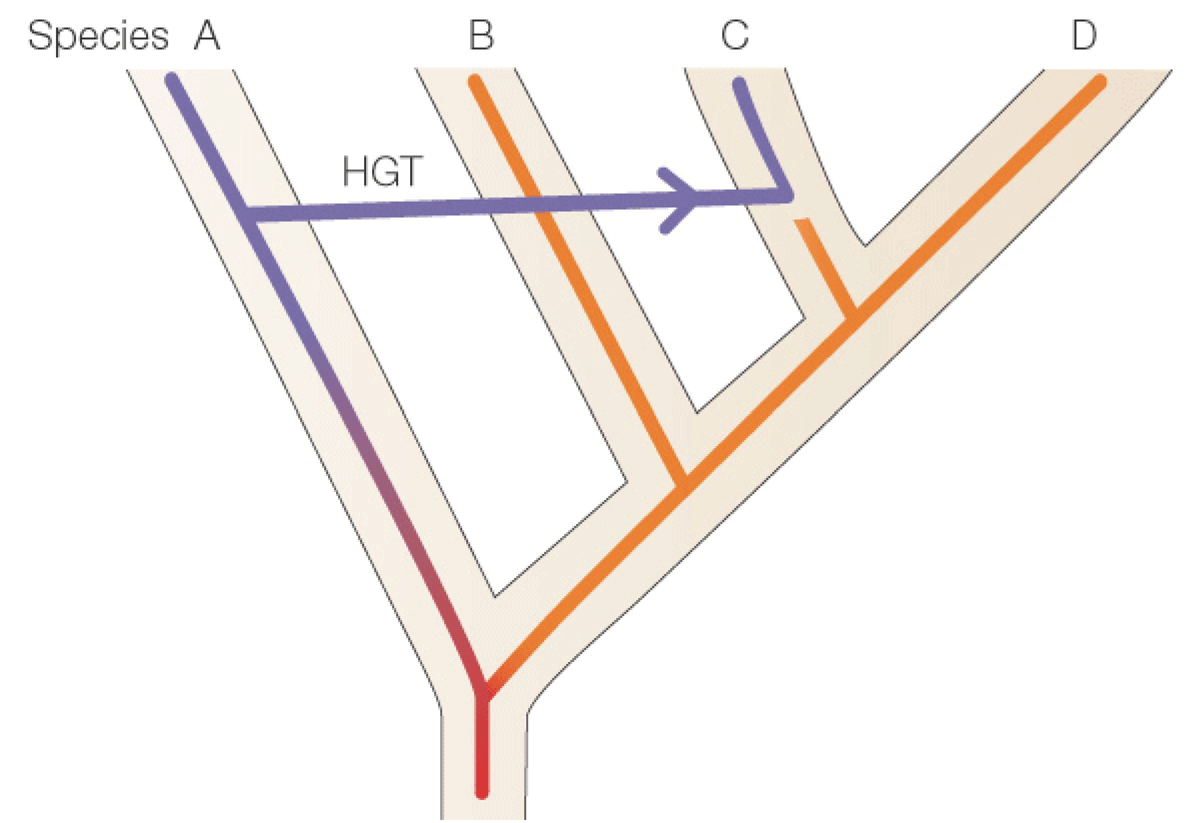
1,430 / 11,293

(~12.5%)

Patterson et al. (2006) Nature

Horizontal transfer of genes

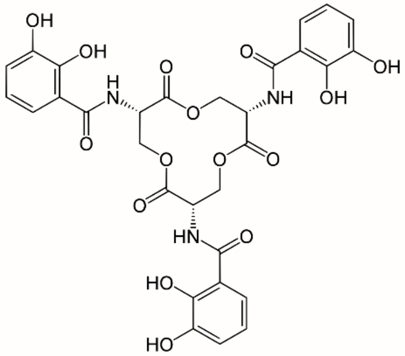
Exchange of genes between organisms other than through reproduction

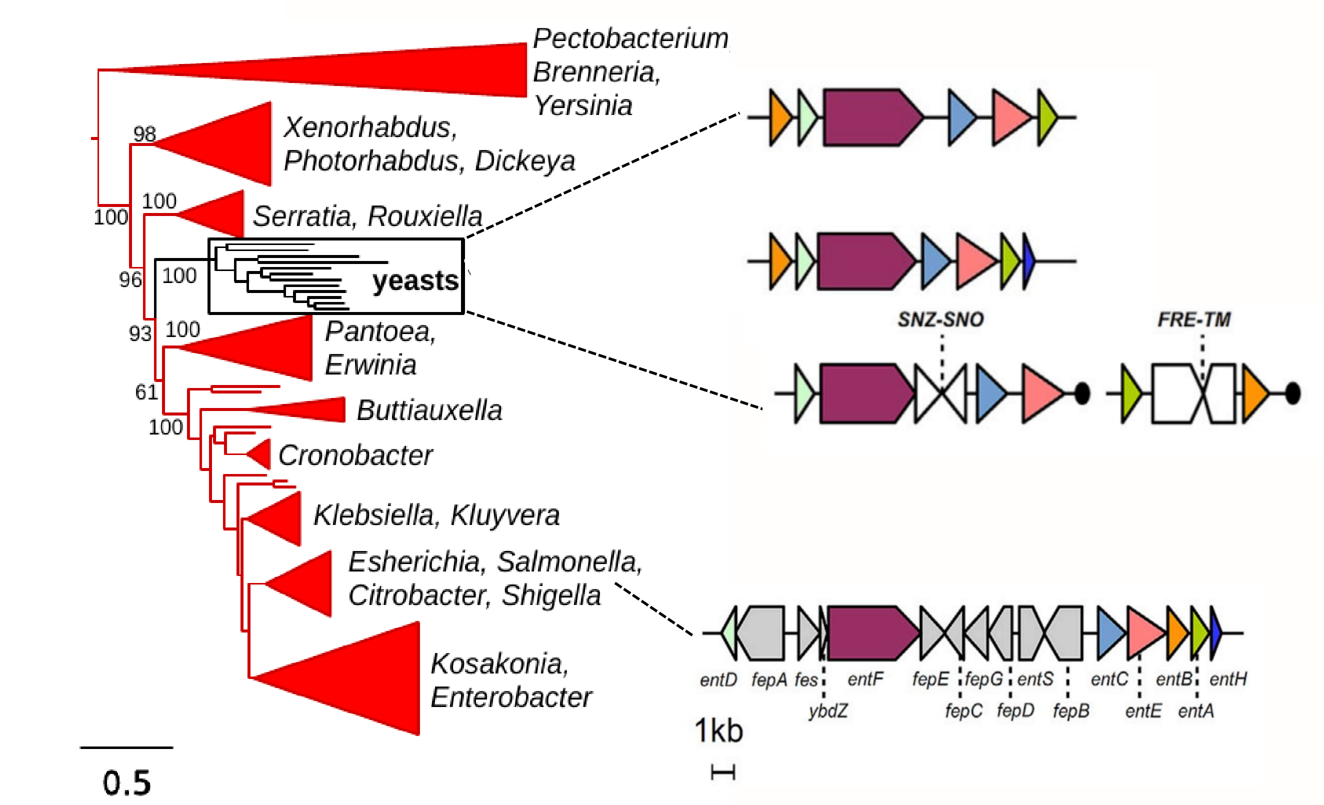


Gogarten & Townsend (2005) Nature Rev. Genet.

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





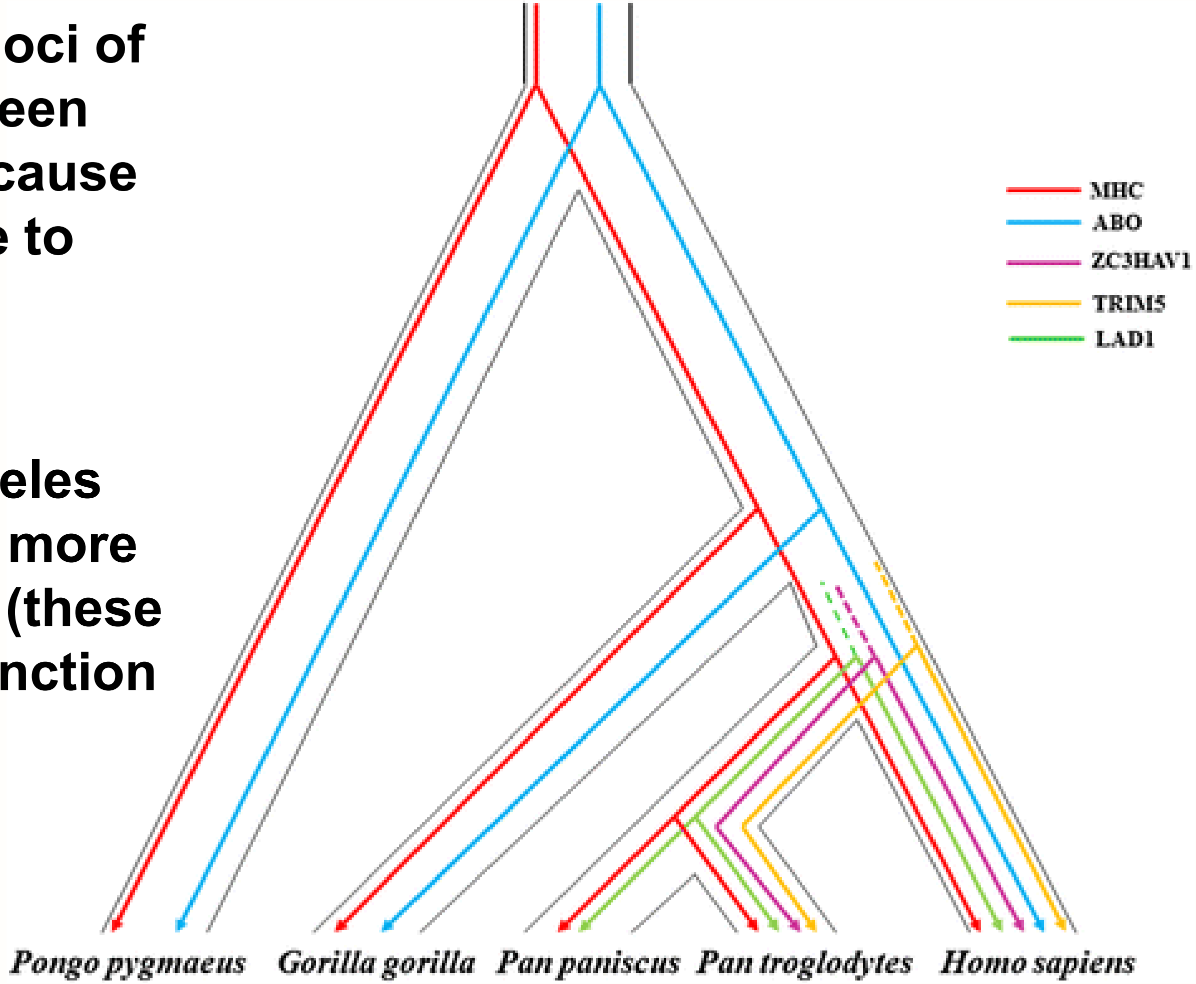
Kominek, Doering, et al. (2019) Cell

Balancing selection

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

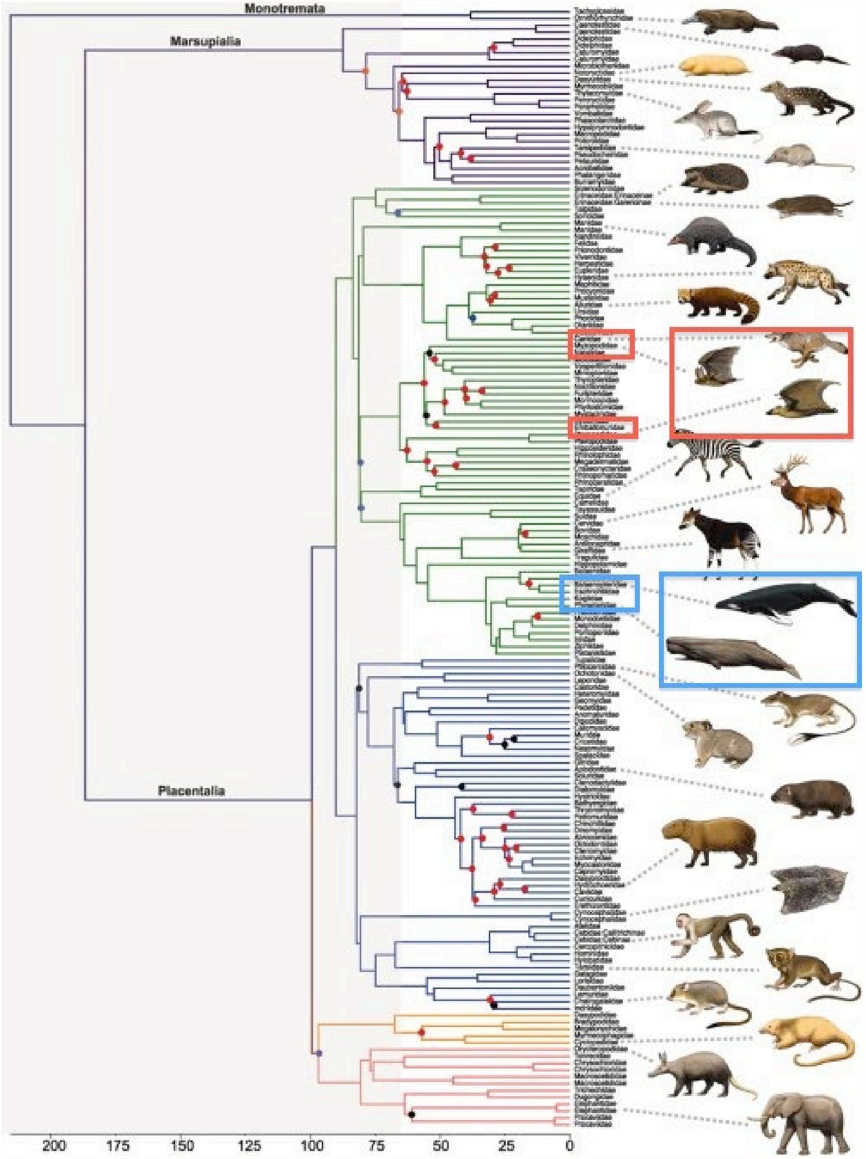
Best example: alleles at loci of the MHC – they have been retained by selection because they confer resistance to infection

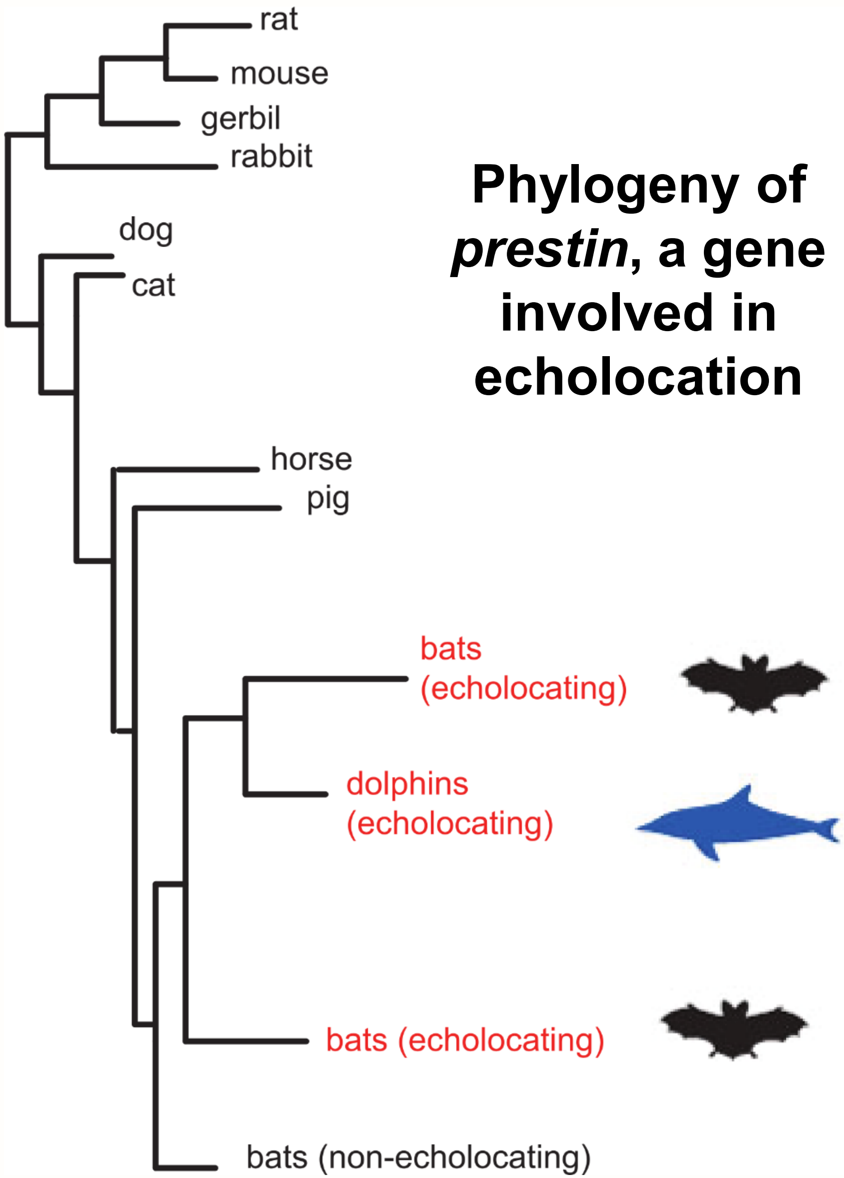
Certain human MHC alleles appear to have diverged more than 65 million years ago (these alleles witnessed the extinction of dinosaurs!!!)



Azevedo et al. (2015) Human Genomics

Positive selection

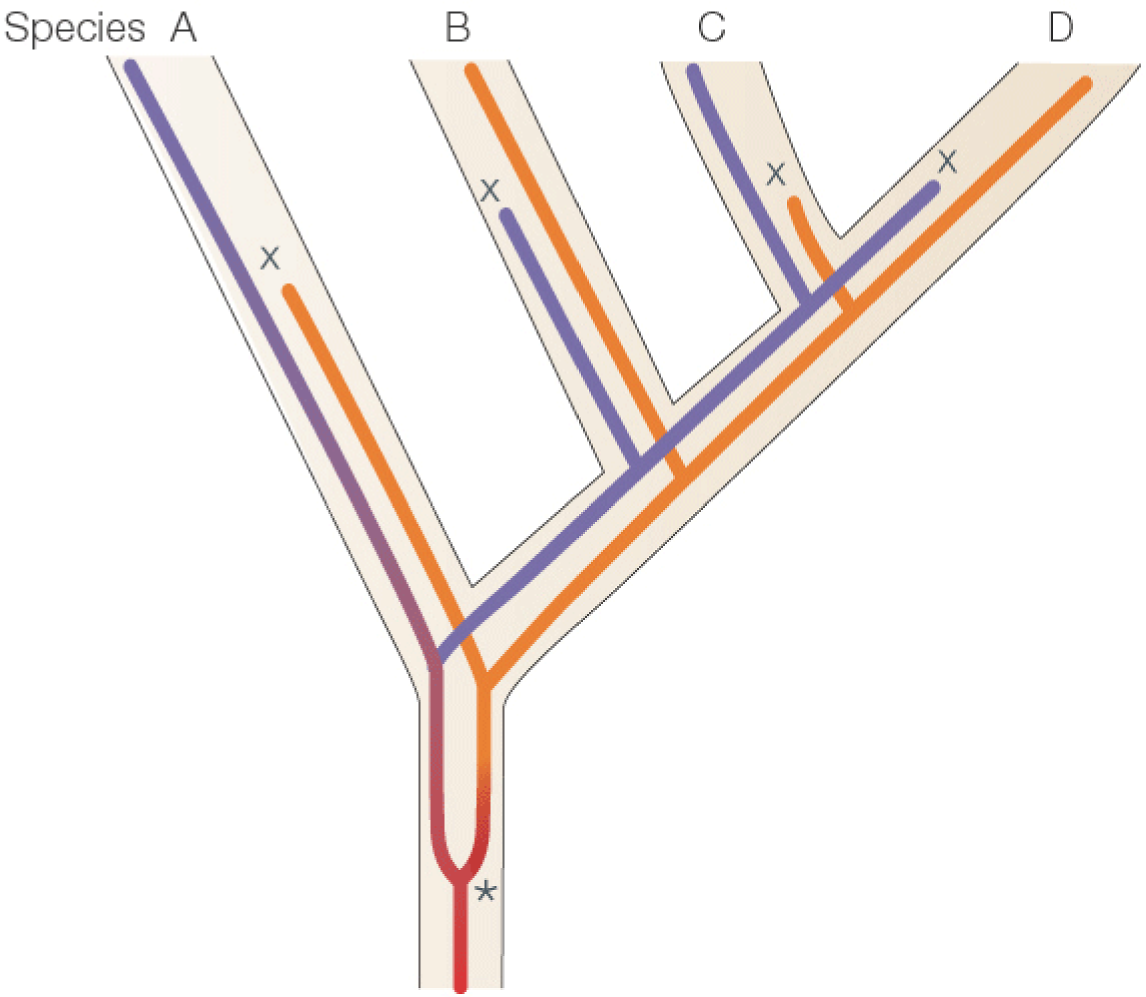




Phylogeny of *prestin*, a gene involved in echolocation

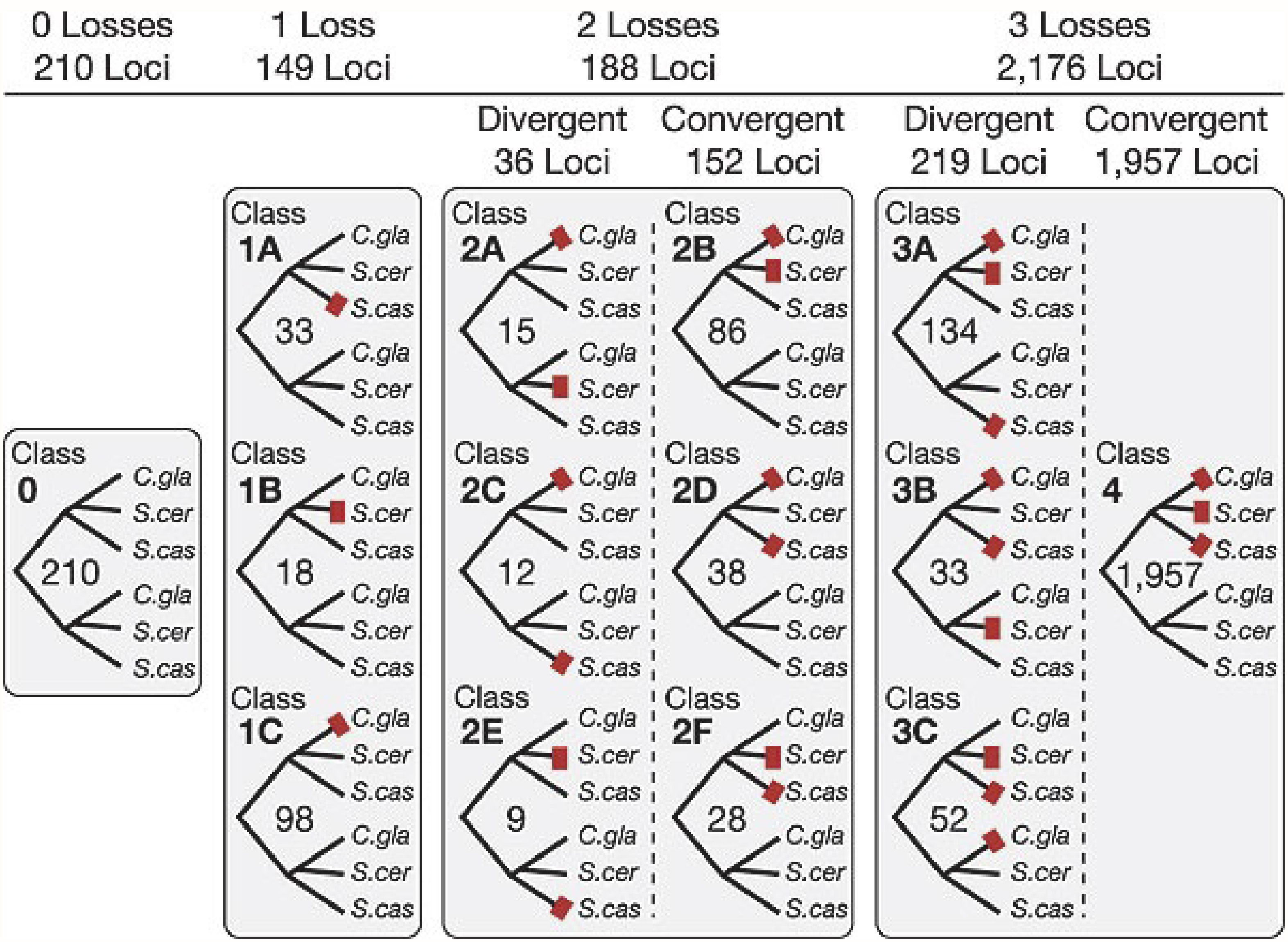
Li et al. (2010) Curr. Biol.

Gene duplication and loss



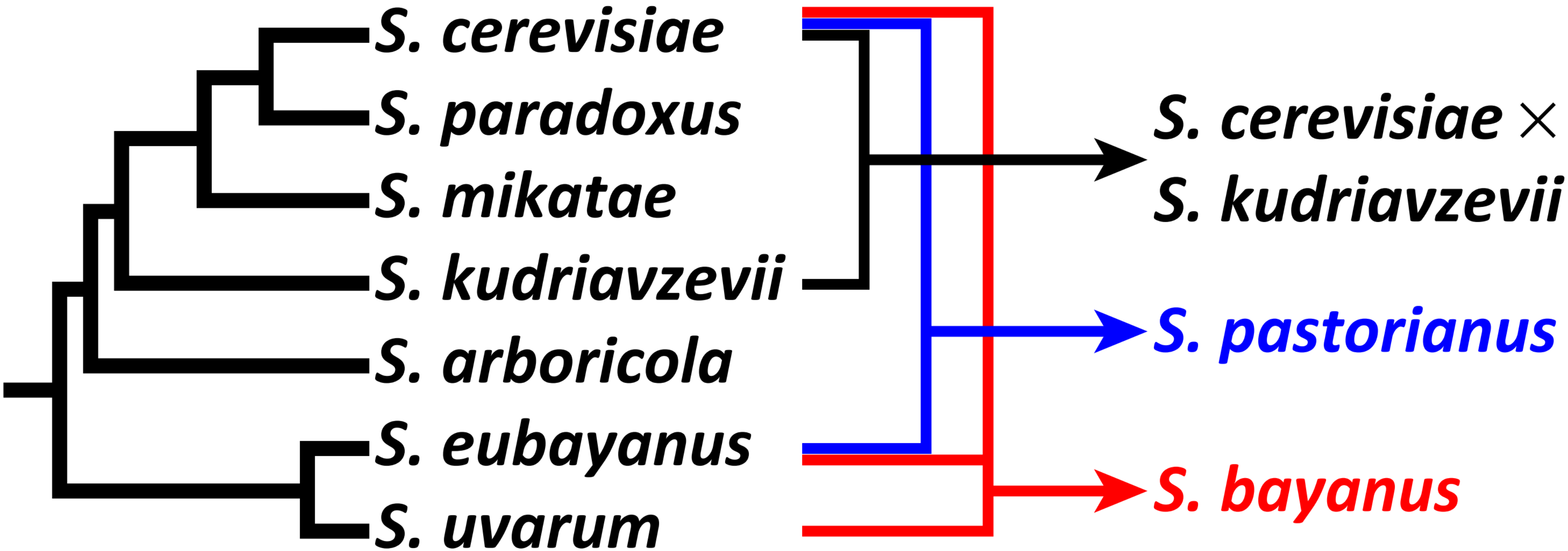
Gogarten & Townsend (2005) Nature Rev. Genet.

Gene duplication and loss



Scannell et al. (2006) Nature

Hybridization and introgression

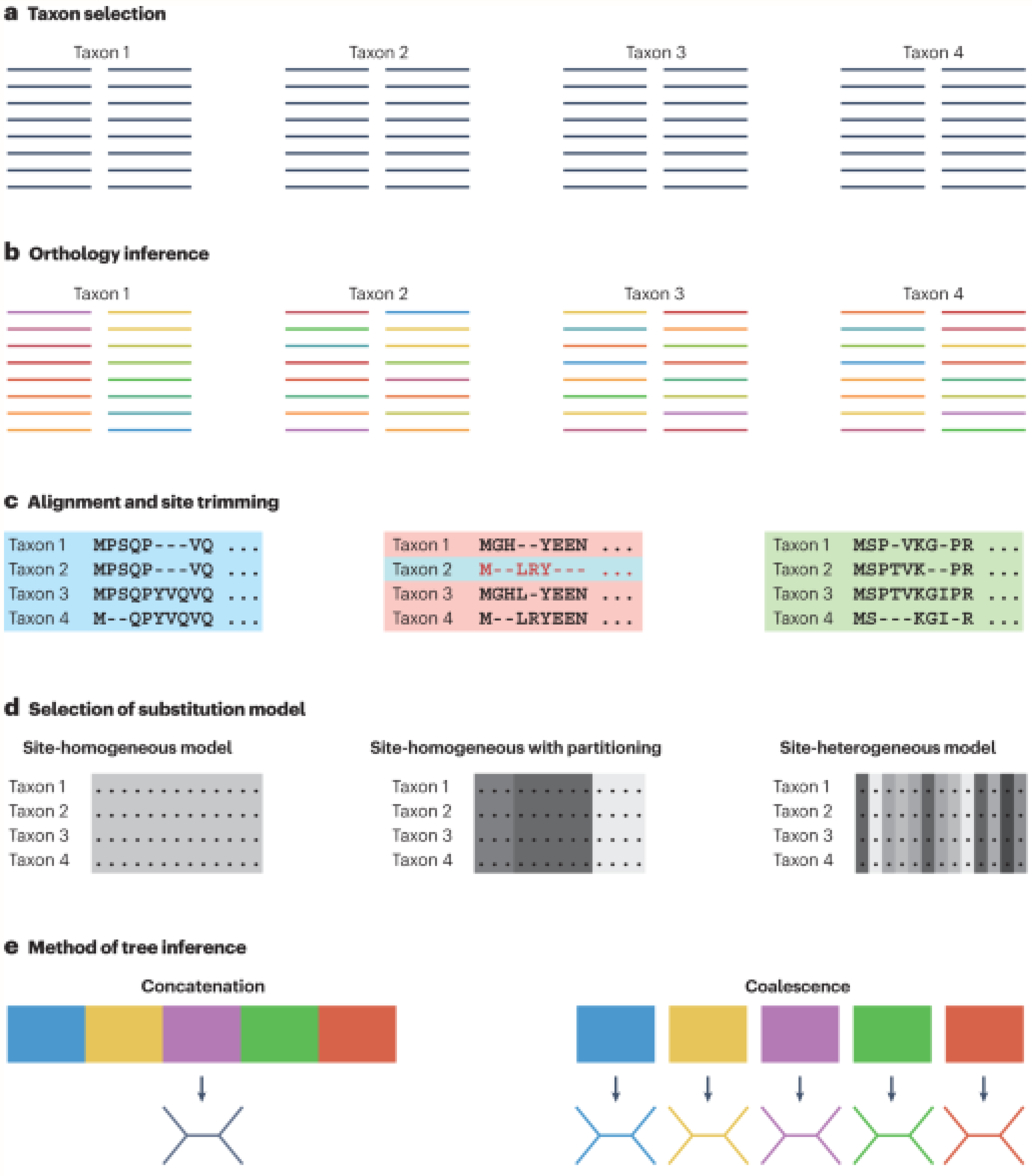


S. eubayanus **was discovered in 2011 – until then, *S. bayanus* was thought to be a "pure" species**

S. cerevisiae – S. paradoxus **divergence ≈ human – mouse divergence**
S. cerevisiae – S. uvarum **divergence ≈ human – chicken divergence**

Hittinger (2013) Trends Genet.

Analytical factors



```
Taxon 1   MPSQP---VQ  ...
Taxon 2   MPSQP---VQ  ...
Taxon 3   MPSQPYVQVQ  ...
Taxon 4   M--QPYVQVQ  ...
```

```
Taxon 1   MGH--YEEN  ...
Taxon 2   M--LRY---  ...
Taxon 3   MGHL-YEEN  ...
Taxon 4   M--LRYEEN  ...
```

```
Taxon 1   MSP-VKG-PR  ...
Taxon 2   MSPTVK--PR  ...
Taxon 3   MSPTVKGIPR  ...
Taxon 4   MS---KGI-R  ...
```

Contributor of incongruence

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

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

- Misalignment
- Excessive trimming
- Inappropriate recoding

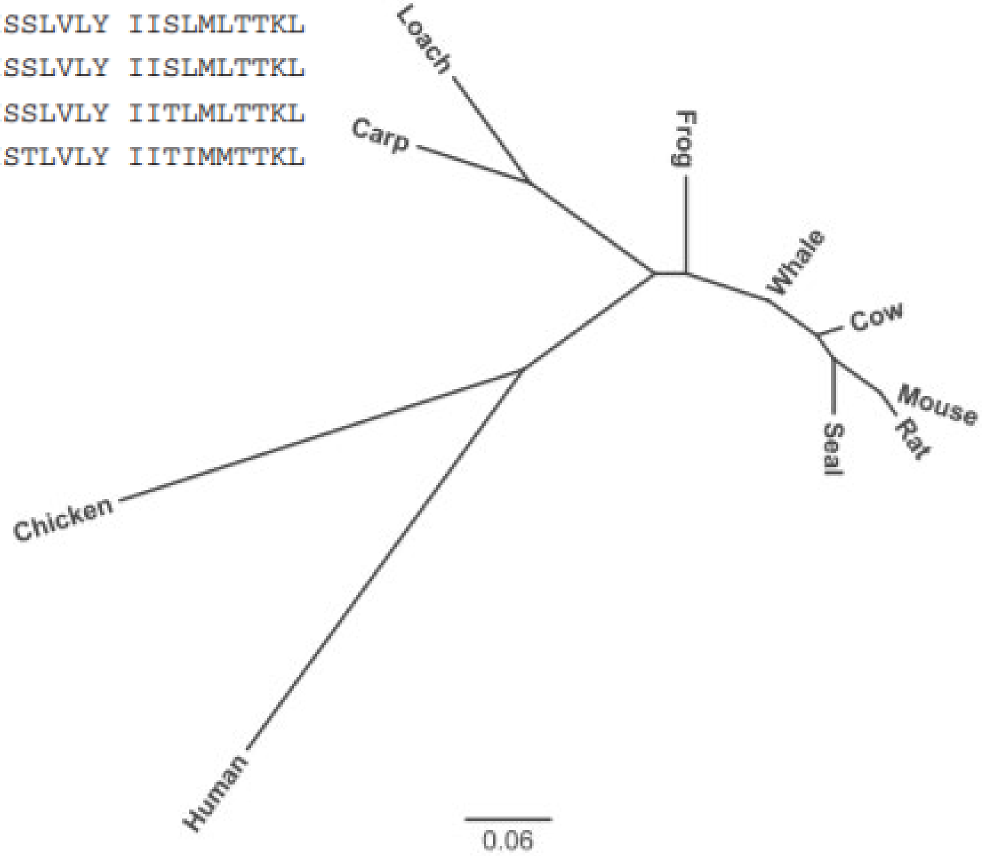
- Long-branch attraction
- Model misspecification
- Inadequate model complexity

- Irreproducibility
- Single-locus accuracy

Steenwyk et al. (2023) Nature Rev. Genet.

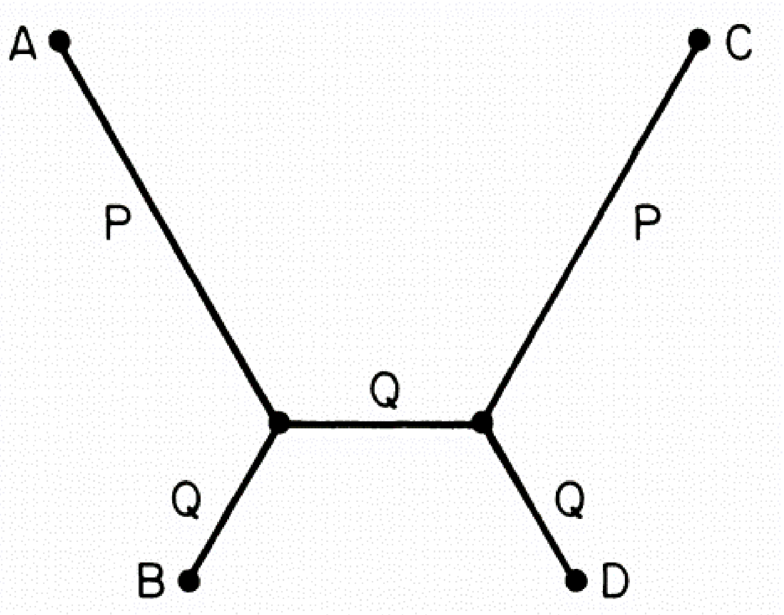
Sampling error

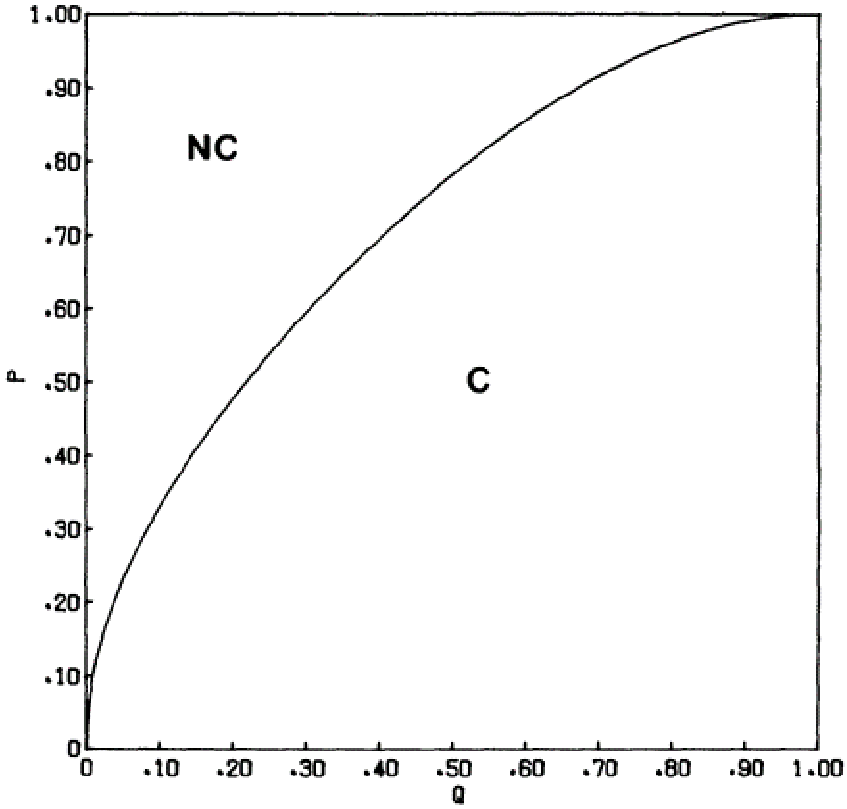
```
10 50
Cow          MAYPMQLGFQ DATSPIMEEL LHFHDHTLMI VFLISSLVLY IISLMLTTKL
Carp         MAHPTQLGFK DAAMPVMEEL LHFHDHALMI VLLISTLVLY IITAMVSTKL
Chicken      MANHSQLGFQ DASSPIMEEL VEFHDHALMV ALAICSLVLY LLTLMLMEKL
Human        MAHAAQVGLQ DATSPIMEEL ITFHDHALMI IFLICFLVLY ALFLTLTTKL
Loach        MAHPTQLGFQ 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
```



Rokas (2011) Curr. Prot. Mol. Biol.

Systematic error

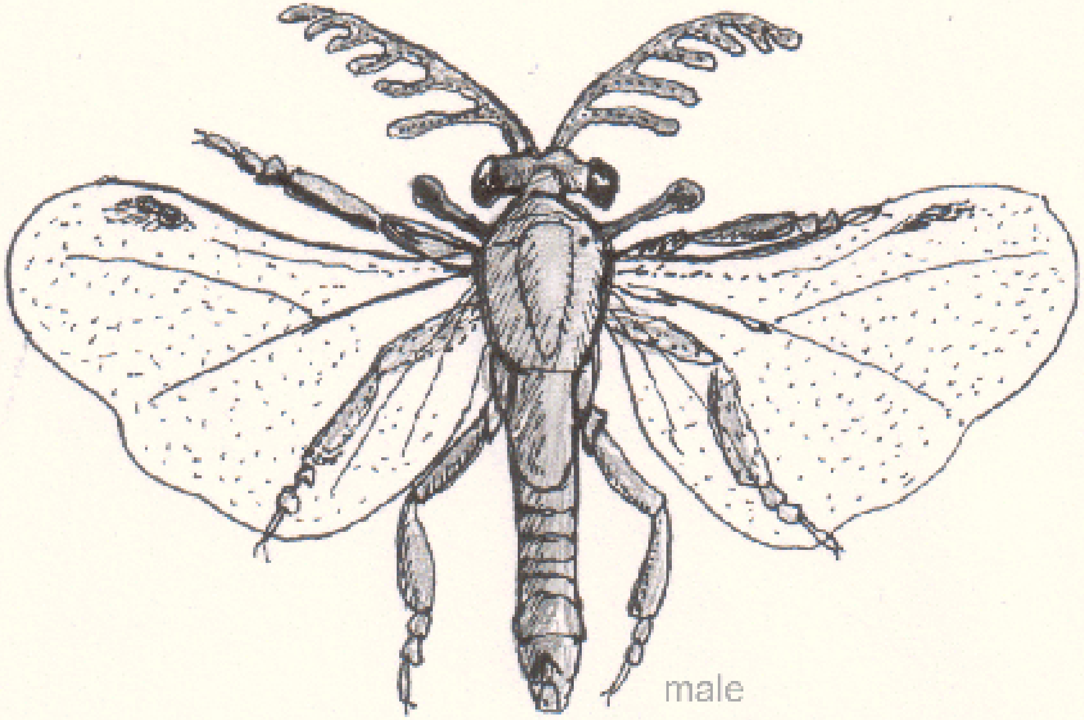




Long branch attraction

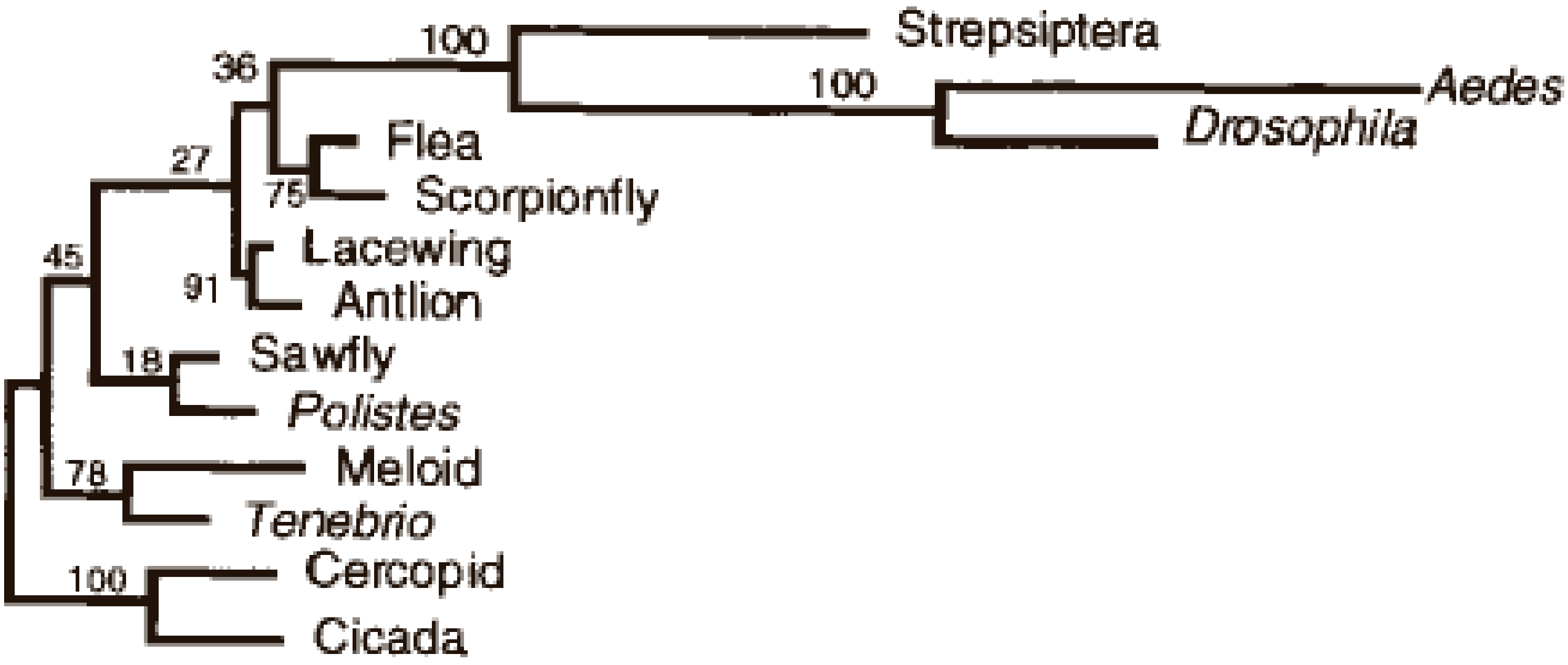
Felsenstein (1978) Syst. Zool.

“The Strepsiptera Problem” is a classic example of LBA



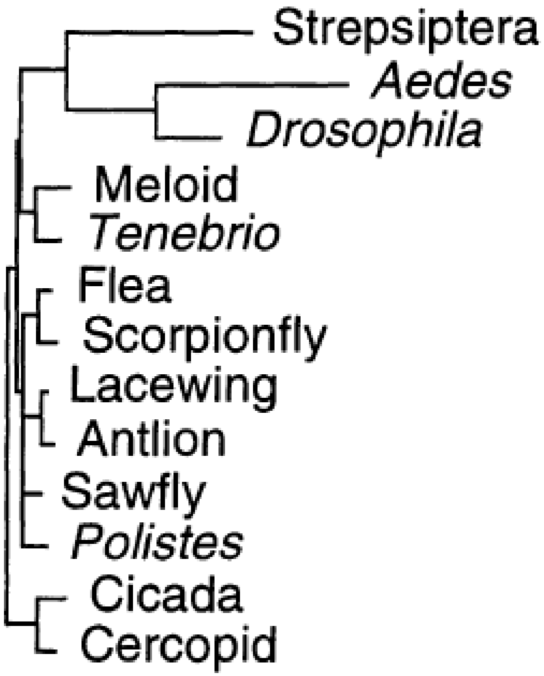
Halictophagidae (Strepsiptera)

The Strepsiptera Problem

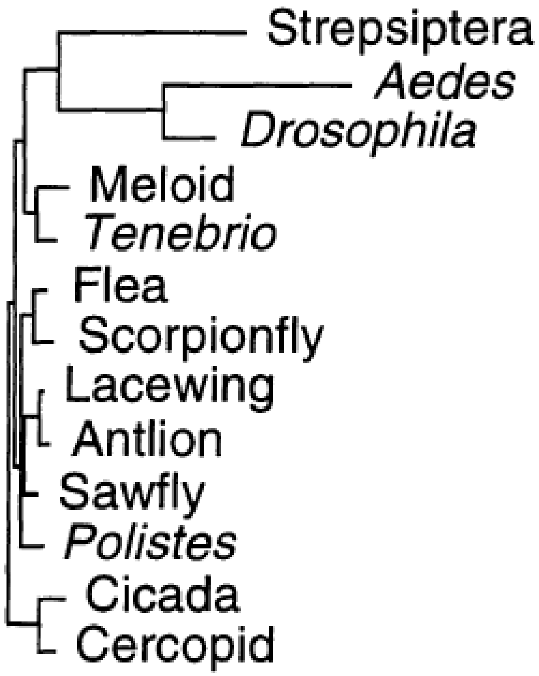


Carmean & Crespi (1995) Nature

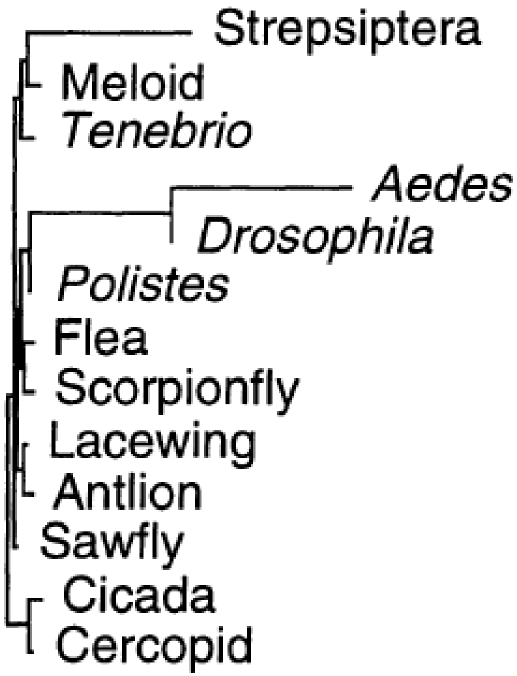
The Strepsiptera Problem



p distance



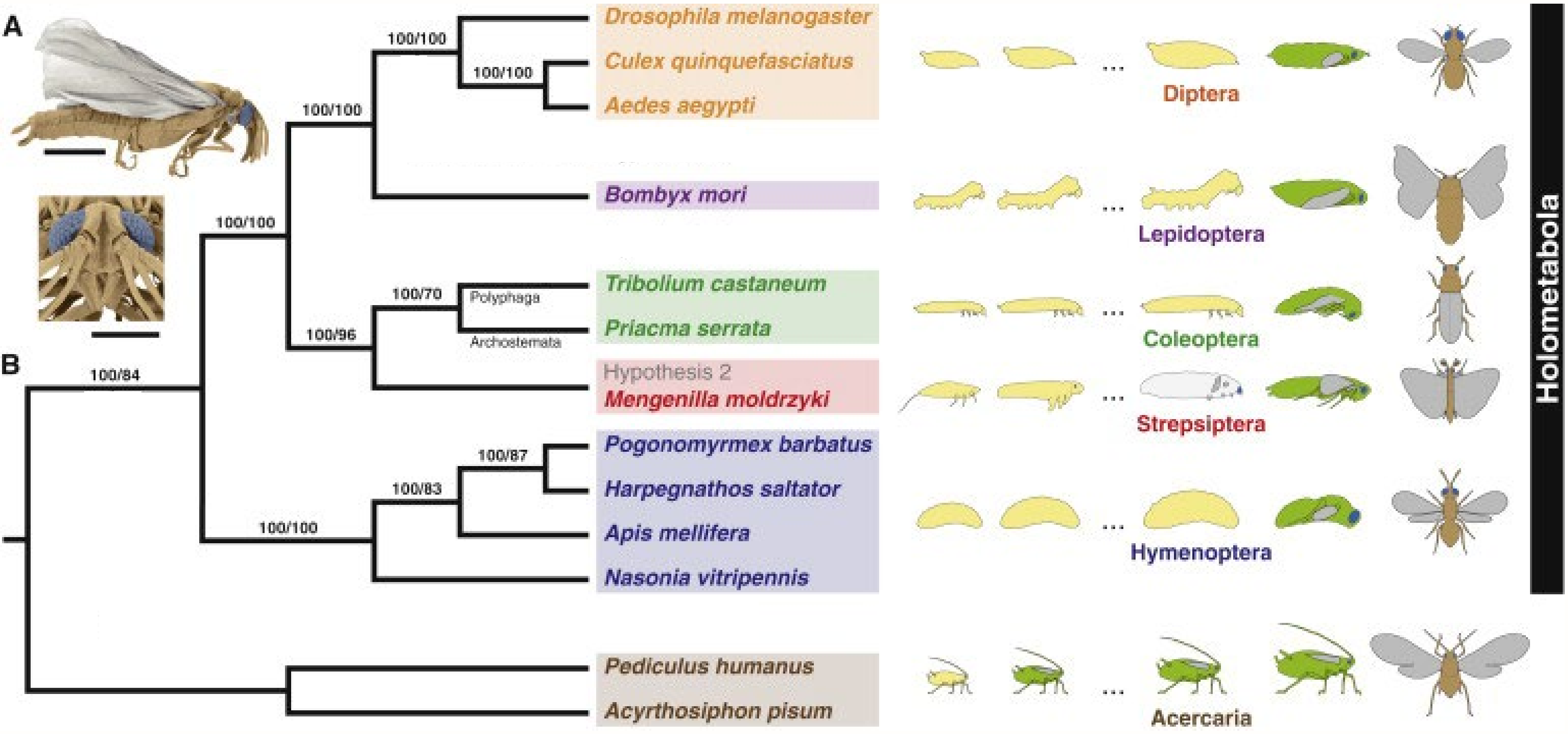
HKY85



HKY85+GAMMA

Huelsenbeck (1997) Syst. Biol.

Solving the Strepsiptera Problem with more genes and better models



Niehuis et al. (2012) Curr. Biol.

Treatment errors

Multiple sequence alignment

Alignment trimming

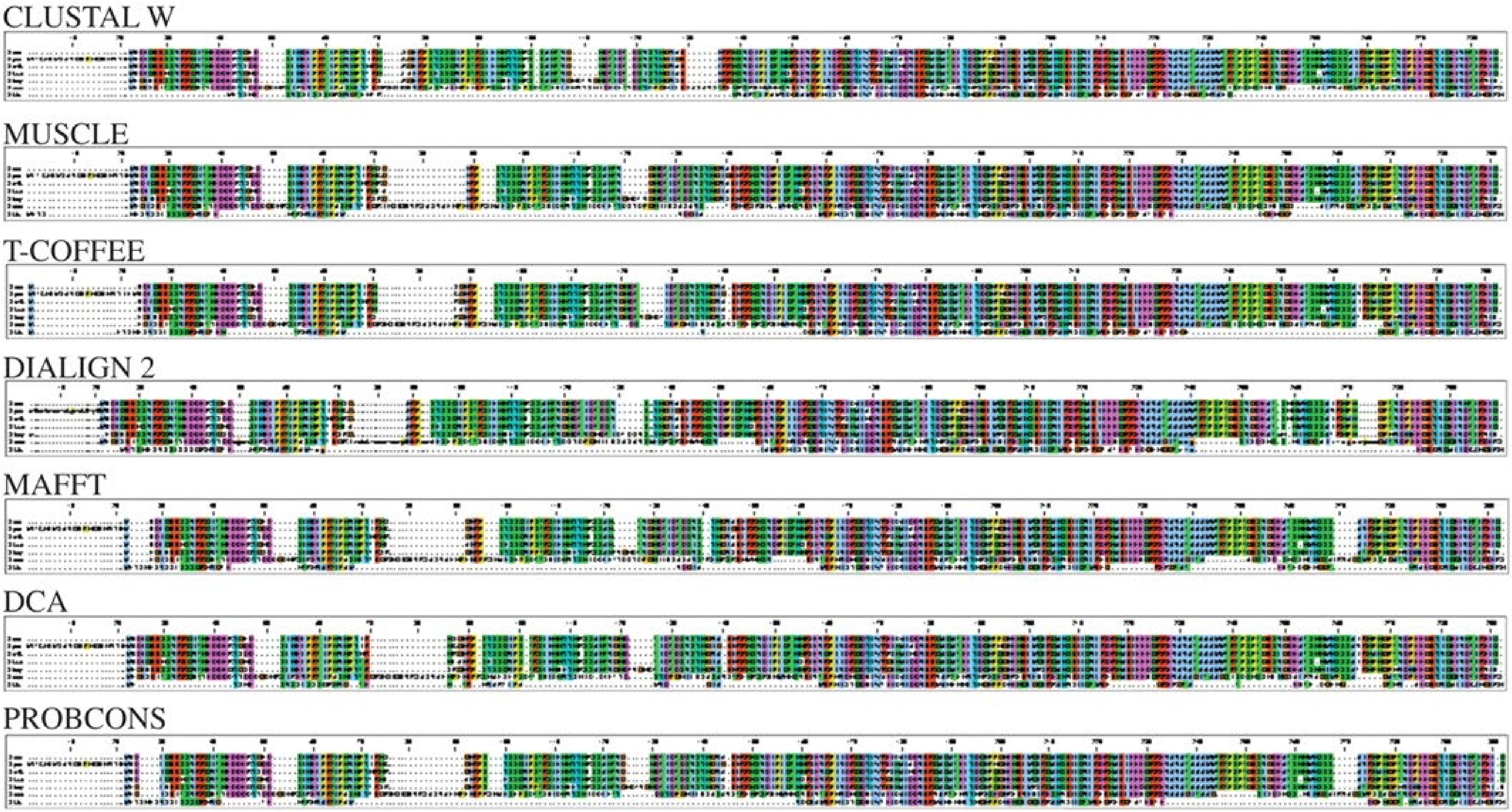
Character recoding

Irreproducibility

…

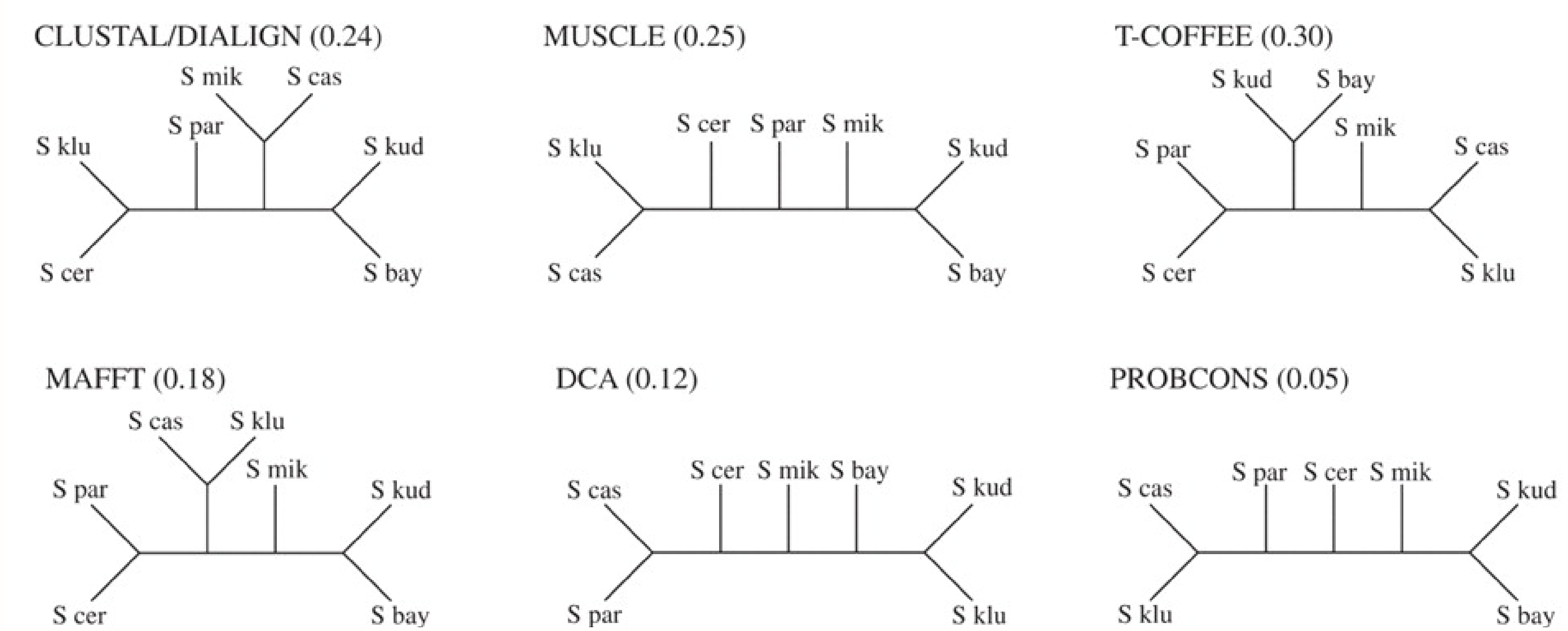
Steenwyk et al. (2023) Nature Rev. Genet.

Multiple sequence alignment



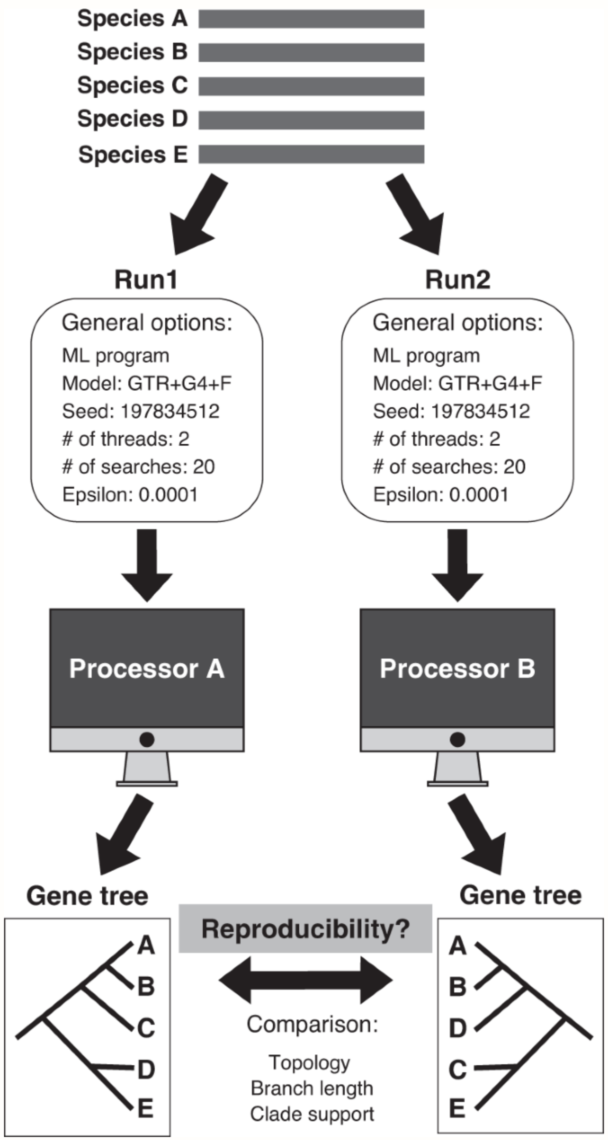
Wong et al. (2008) Science

7 tools produce 6 different topologies



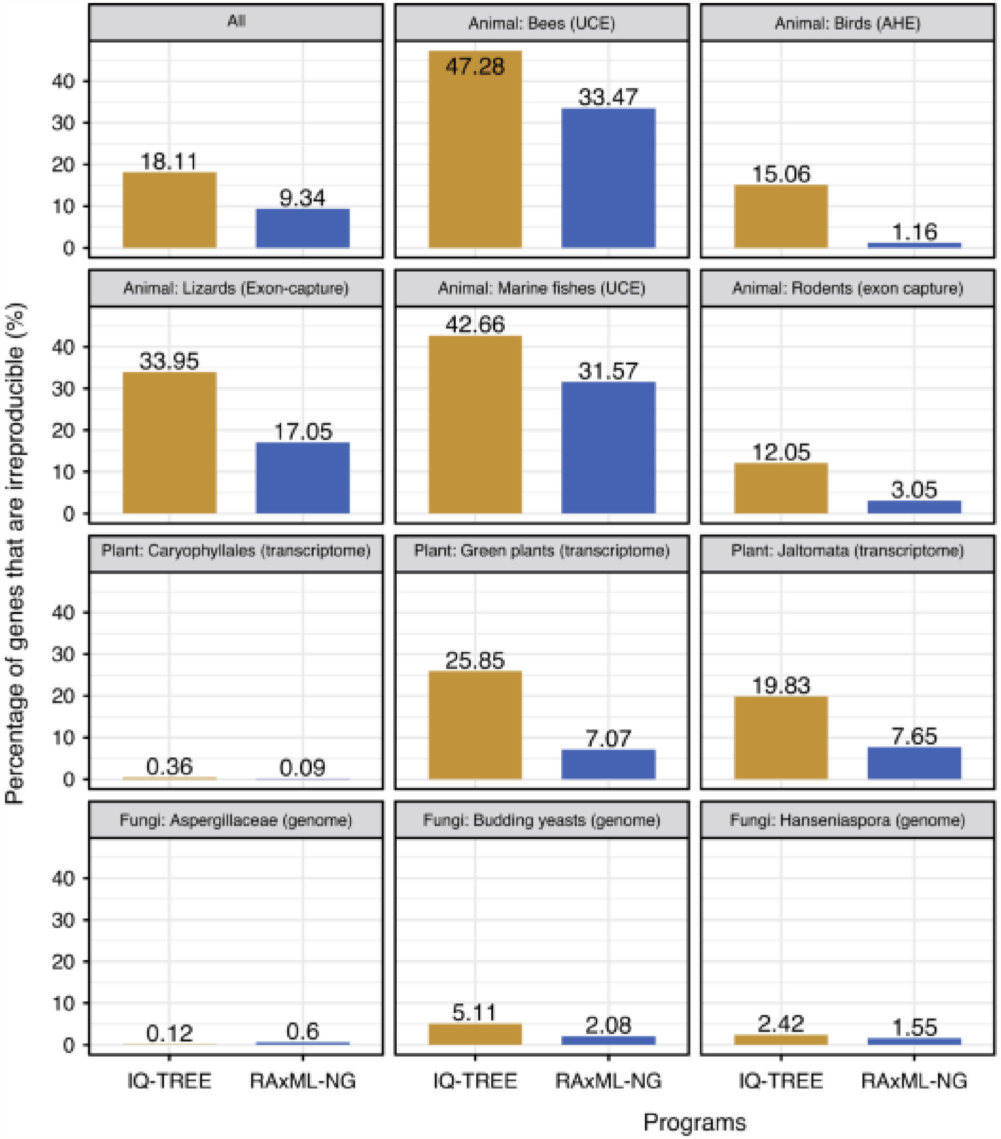
Wong et al. (2008) Science

What is the reproducibility of our analyses?



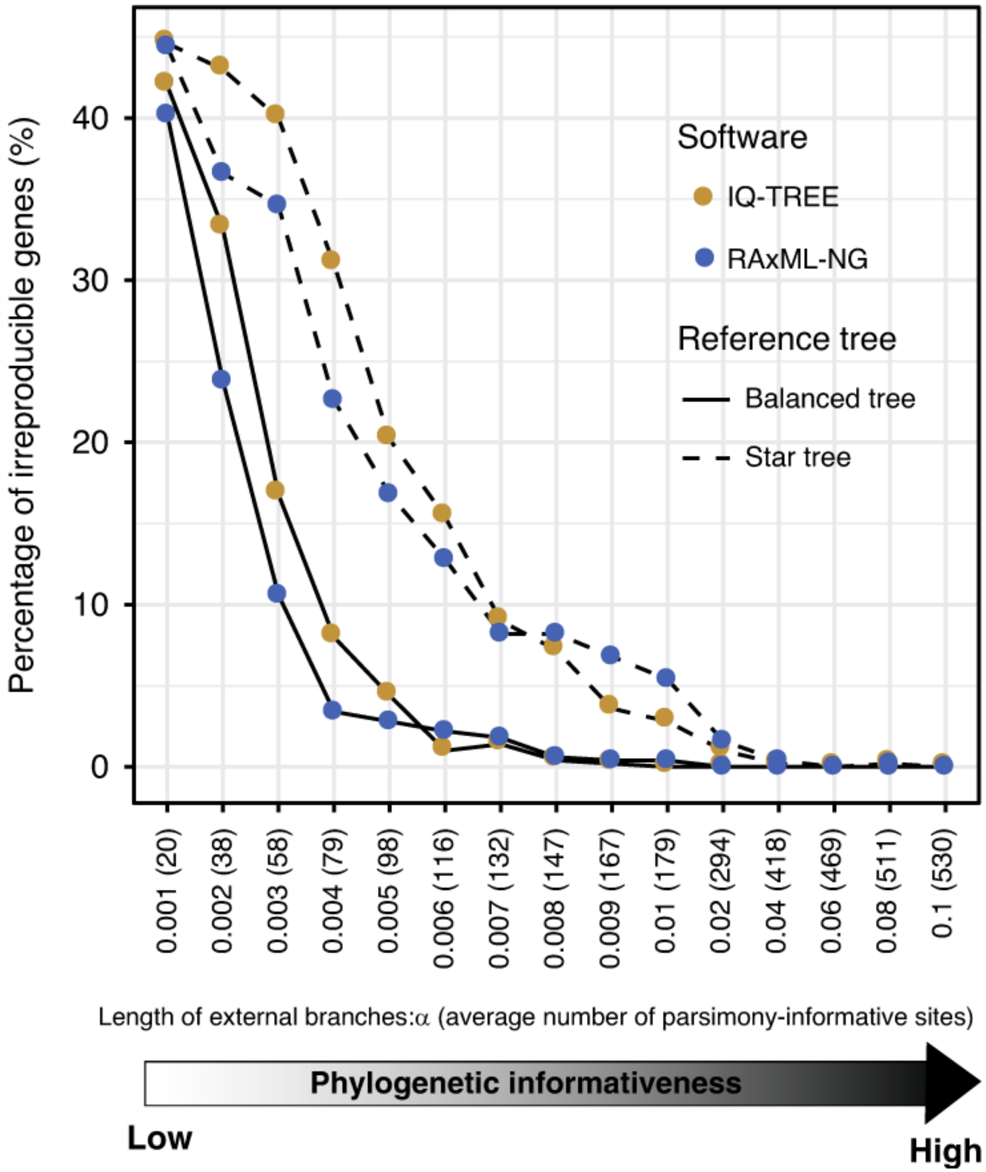
Shen et al. (2020) Nature Comm.

9-18% of gene trees are irreproducible



Shen et al. (2020) Nature Comm.

Genes yielding irreproducible trees are less informative



Shen et al. (2020) Nature Comm.

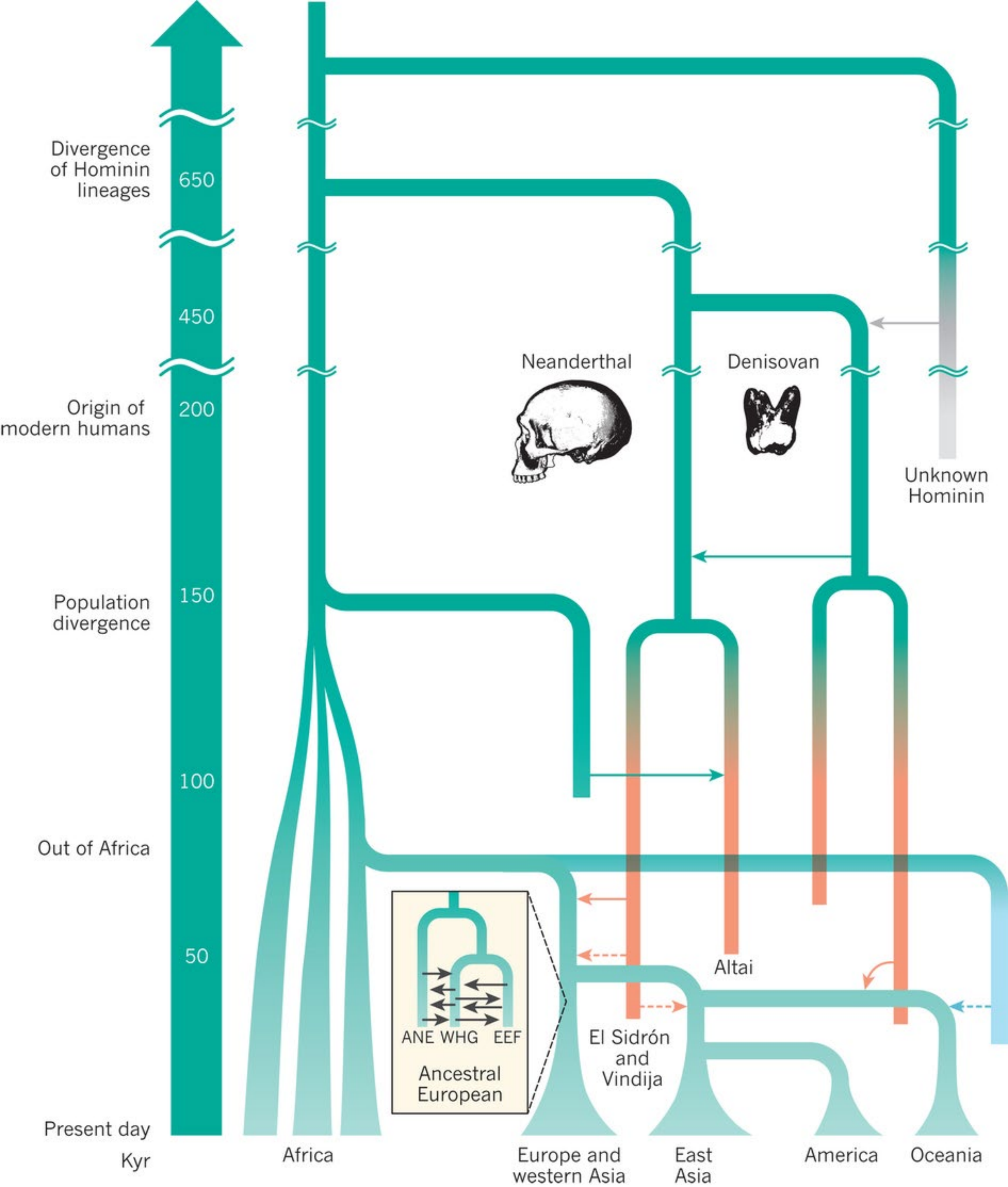
We have not eliminated incongruence but now better understand the biological and analytical contributing factors and their impact; sources vary by lineage and depth

Inference at shallow depths is easier:

analytical factors

biological factors

The evolution of hominids



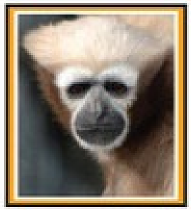
Nielsen et al. (2017) Nature

The phylogeny of primate genera

Nomascus leucogenys

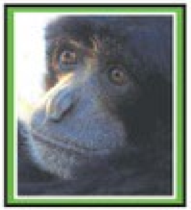
NLE

Hoolock leuconedys



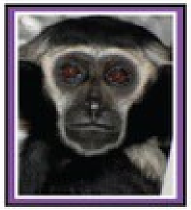
HLE

Symphalangus syndactylus



SSY

Hylobates pileatus



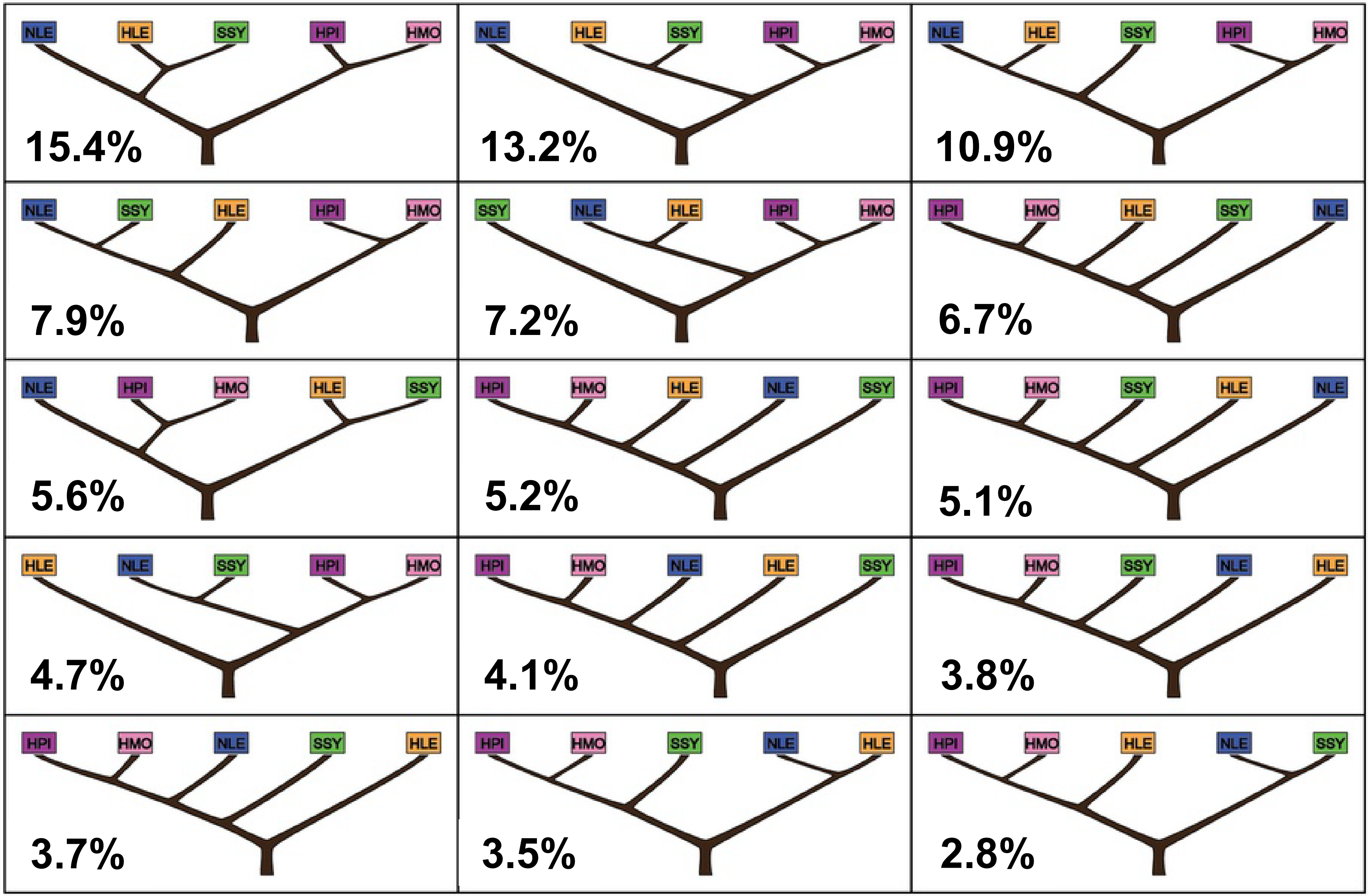
HPI

Hylobates moloch

HMO

Carbone et al. (2014) Nature

"Easier" doesn't mean "easy"!



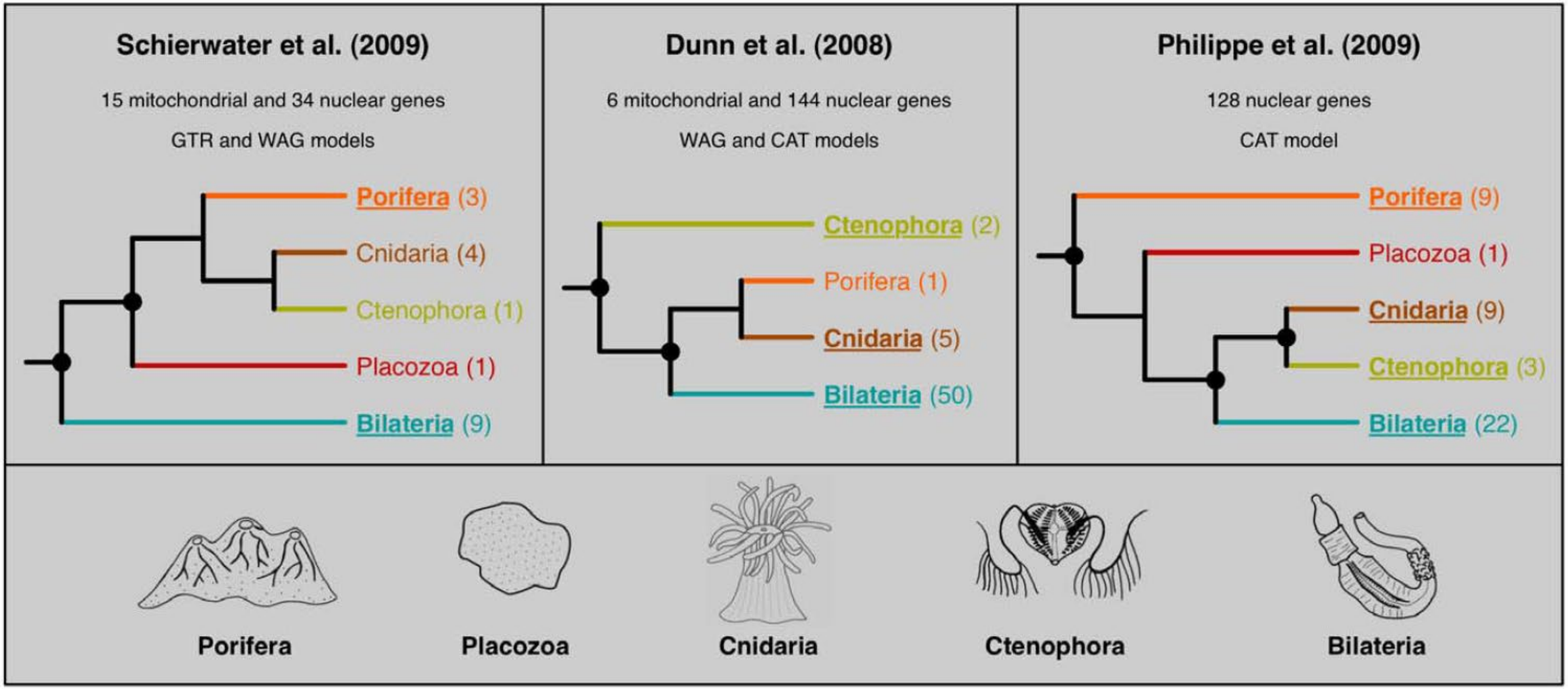
Carbone et al. (2014) Nature

Inference in deep time can be more challenging:

analytical factors

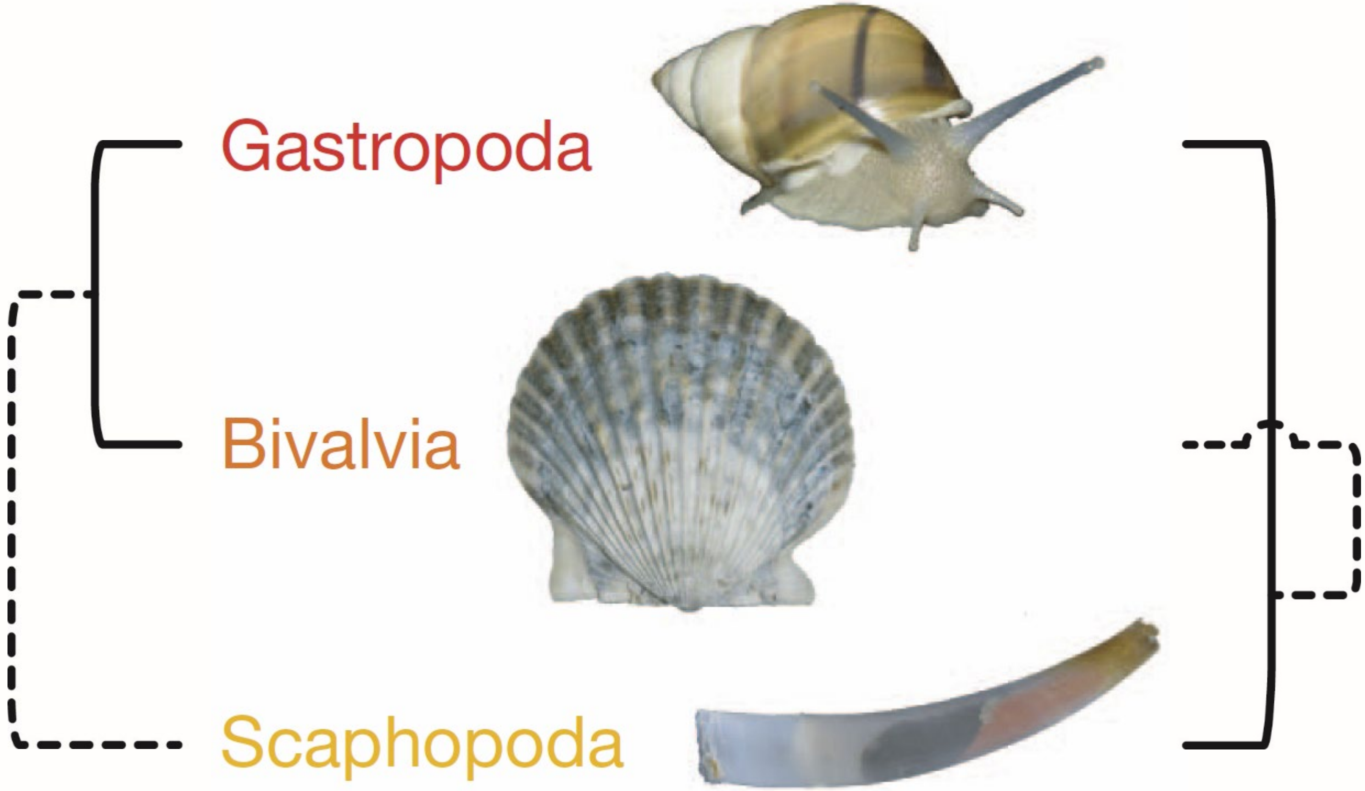
biological factors

Incongruence in deep time



Philippe et al. (2011) PLoS Biol.

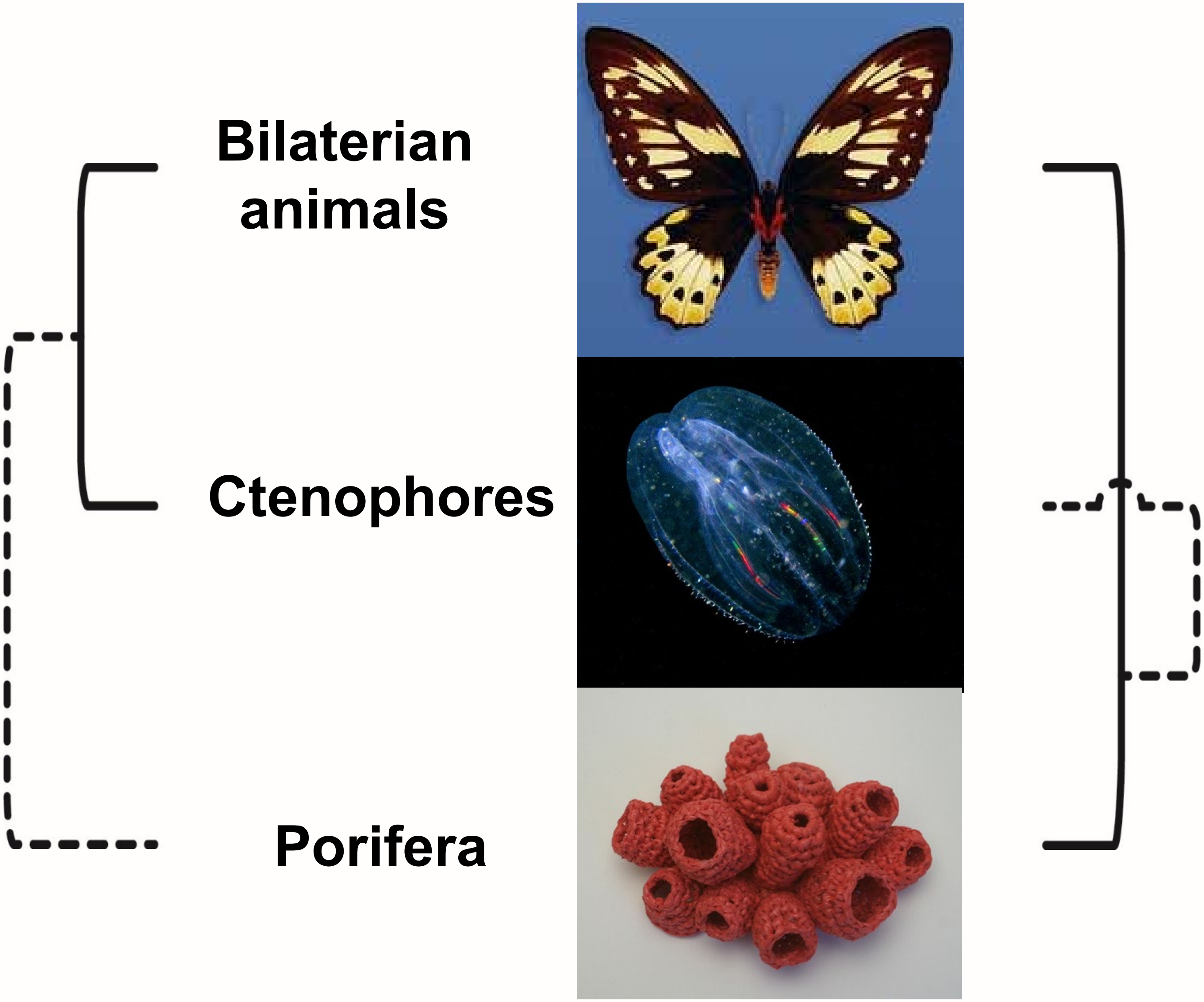
Incongruence in deep time



Kocot et al. (2011) Nature

Smith et al. (2011) Nature

Incongruence in deep time



Bilaterian animals

Ctenophores

Porifera

Pisani et al. (2015) PNAS

Chang et al. (2015) PNAS

Lecture outline

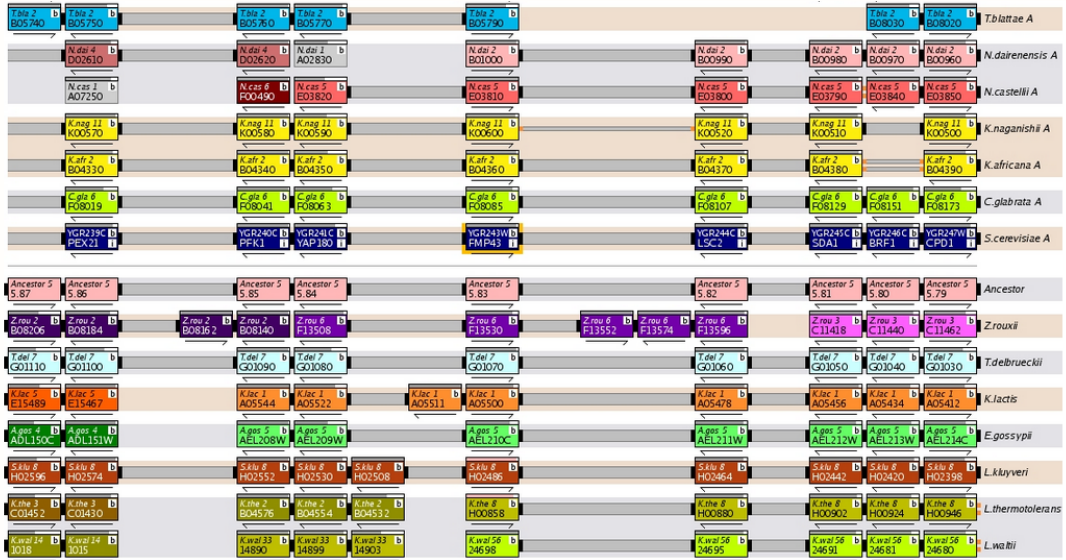
- Incongruence and its causes

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

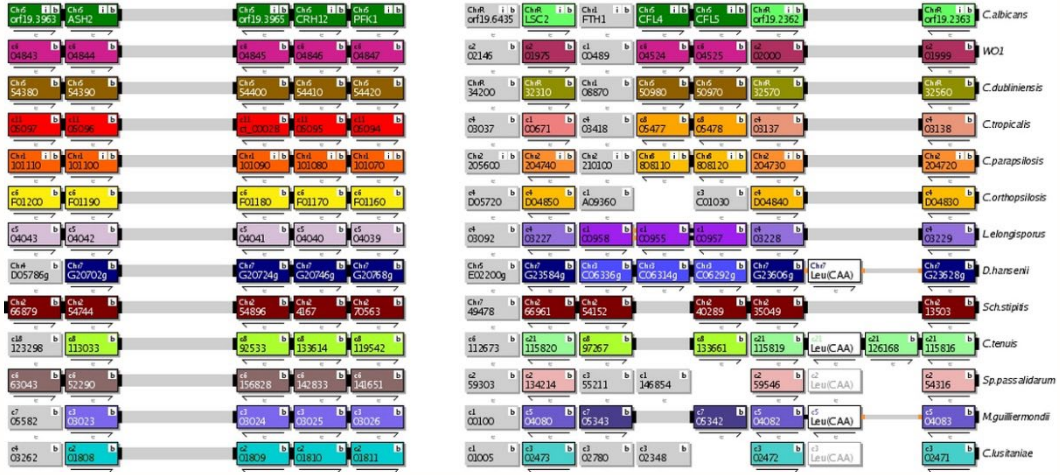
- Handling incongruence in phylogenomic data

An expanded yeast data matrix

Yeast Gene Order Browser (YGOB)



Candida Gene Order Browser (CGOB)



***Saccharomyces* lineage**

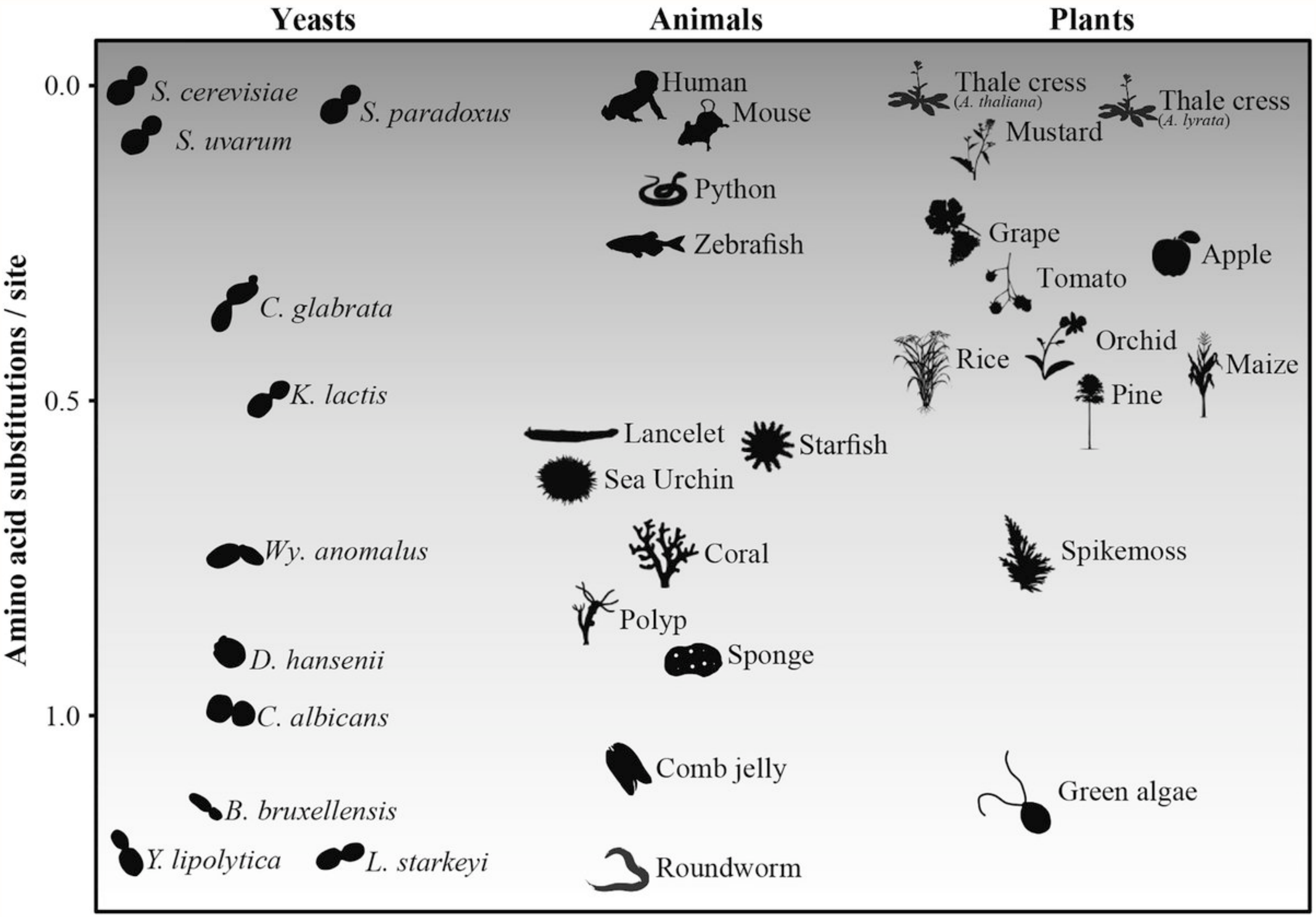
***Candida* lineage**

1,070 genes
23 taxa
no missing data

Byrne & Wolfe (2005) Genome Res.

Fitzpatrick et al. (2010) BMC Genom.

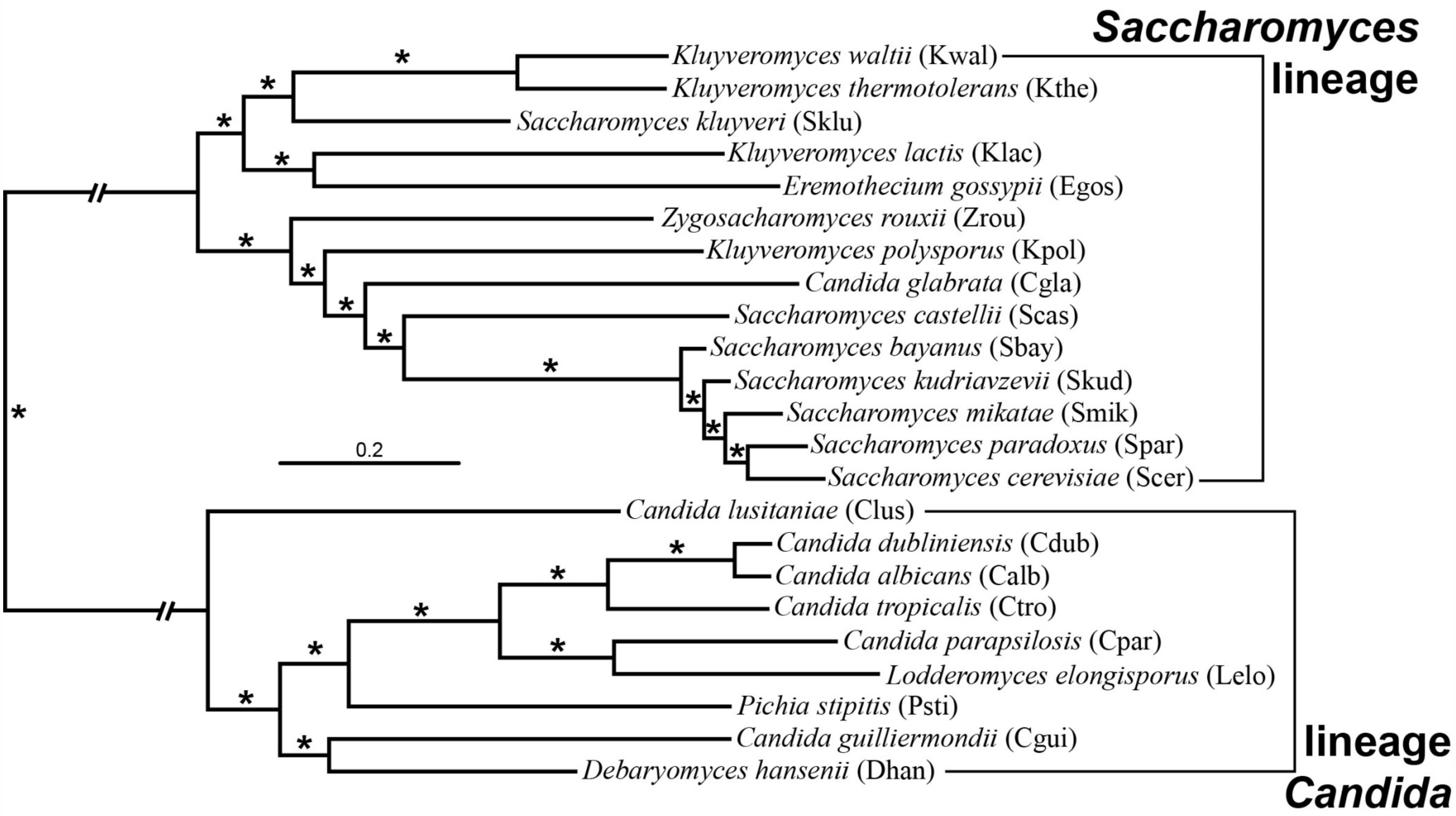
Yeasts exhibit striking genomic diversity



Saccharomyces, _Candida_, _Kluyveromyces_, etc. are all polyphyletic genera

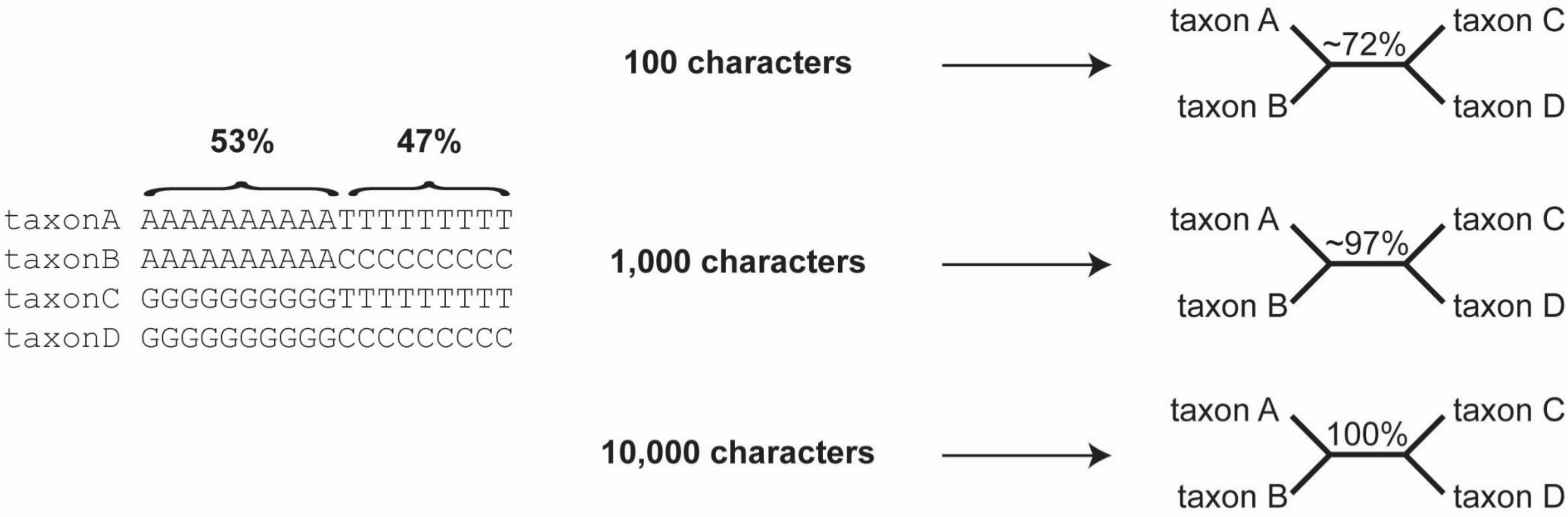
Shen, Opulente, Kominek, Zhou et al. (2018) Cell

Concatenation yields an absolutely supported phylogeny


Saccharomyces lineage
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosacharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
0.2
Candida lusitaniae (Clus)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
lineage Candida
*

Salichos & Rokas (2013) Nature

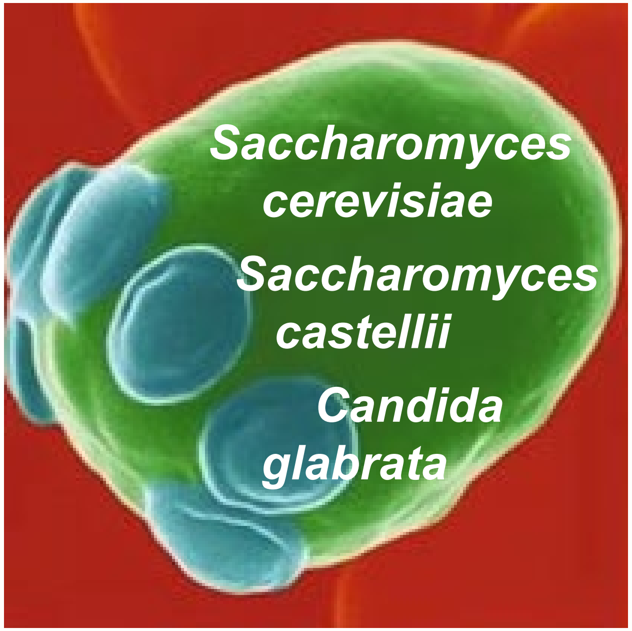
Bootstrap support is misleading when used in large datasets



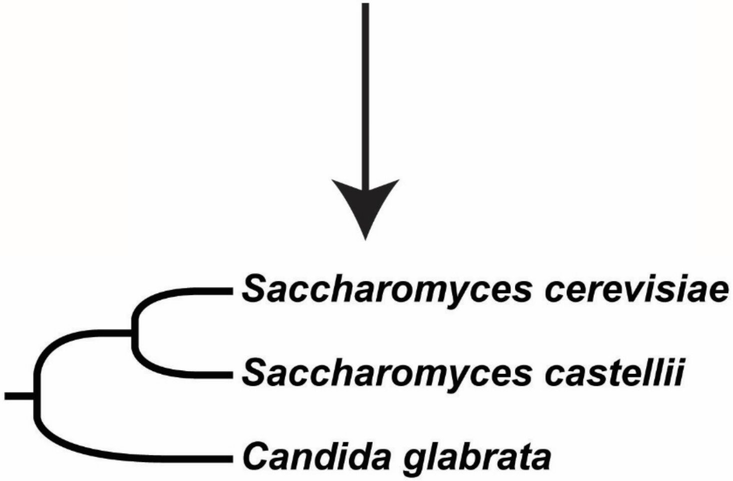
```
           53%        47%
taxonA AAAAAAAAATTTTTTTTT
taxonB AAAAAAAAACCCCCCCCC
taxonC GGGGGGGGGTTTTTTTTT
taxonD GGGGGGGGGCCCCCCCCC
```

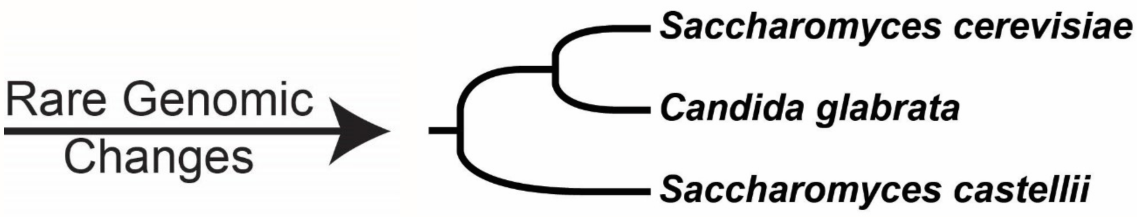
Rokas & Carroll (2006) PLoS Biol.

The concatenation phylogeny is at least partly wrong



Linear Sequence Data

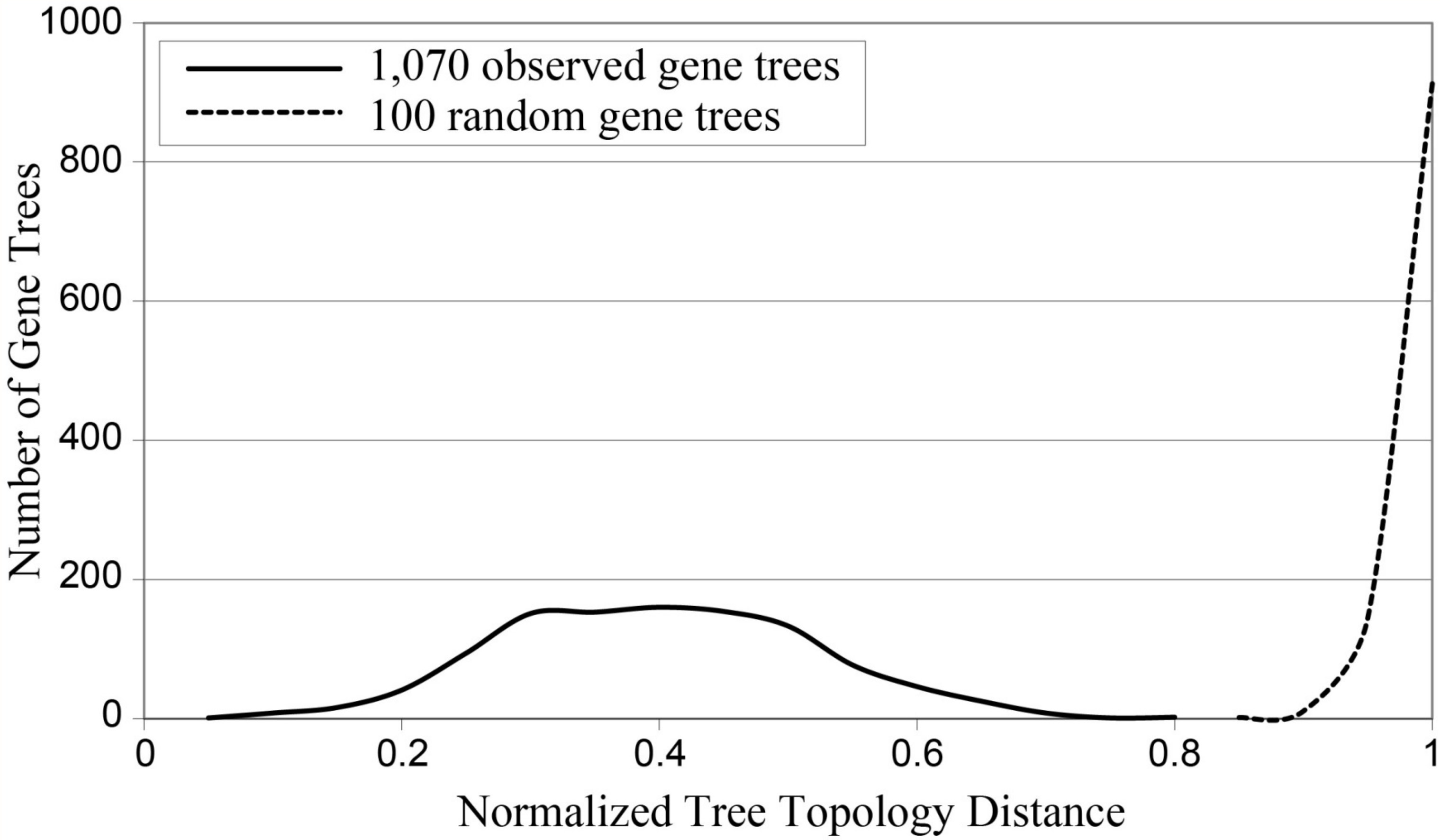




- **5 genomic rearrangements that are uniquely shared by *S. cerevisiae* and *C. glabrata***
- **Much higher proportion of shared gene losses in *S. cerevisiae* and *C. glabrata***
- **Bias in the placement of *C. glabrata* as an outgroup of *S. cerevisiae* and *S. castellii***

Scannell et al. (2006) Nature

All gene trees differ from the concatenation phylogeny



Salichos & Rokas (2013) Nature

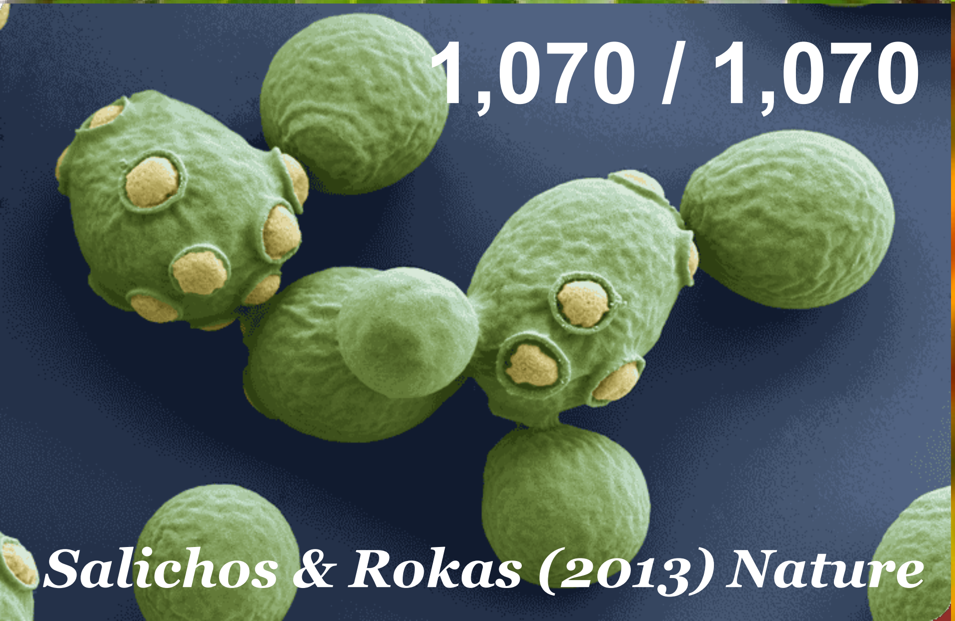
Gene trees are incongruent in most datasets

182 / 184

Zhong et al. (2013) Trends Plant Sci.

440 / 447

Song et al. (2012) PNAS



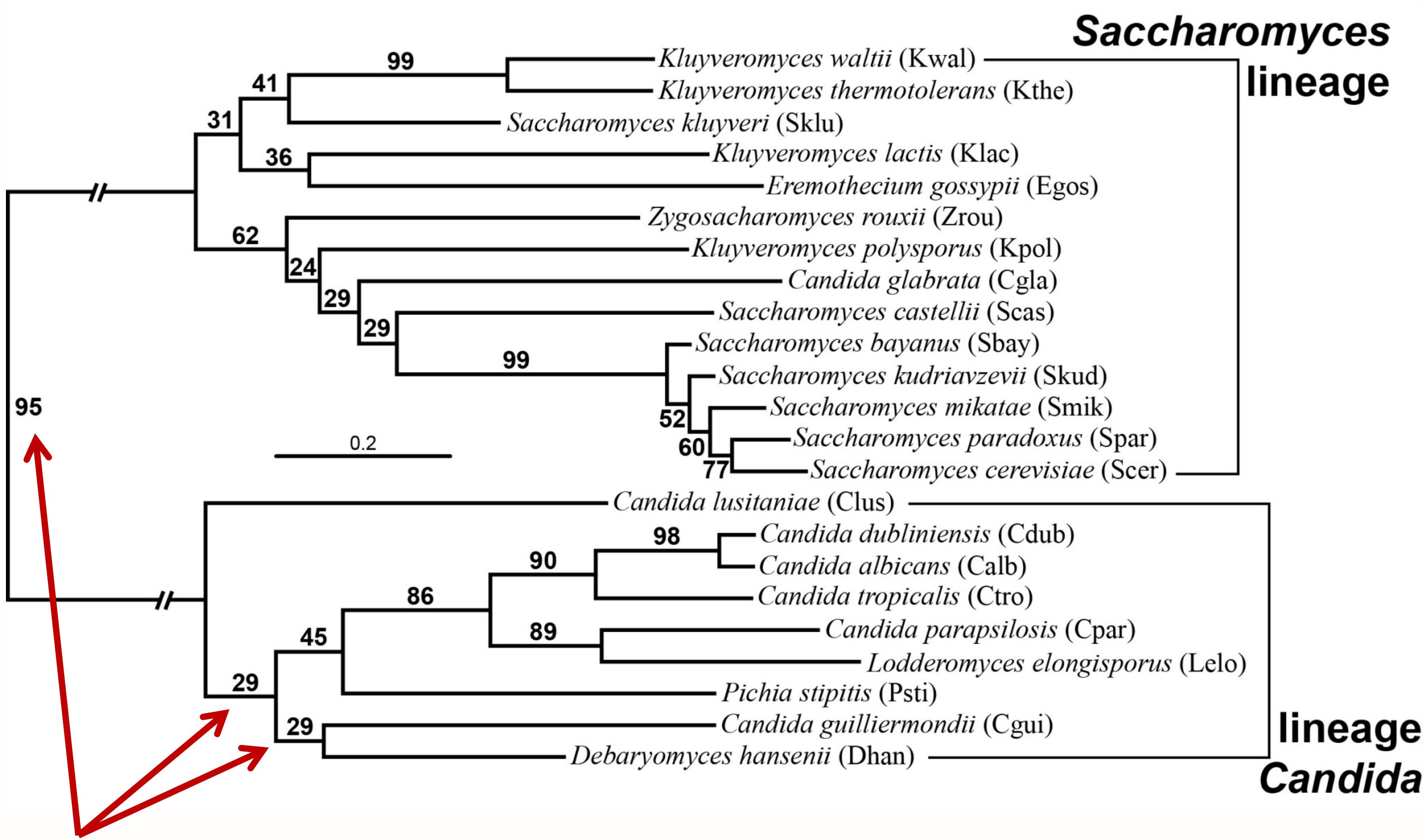
1,070 / 1,070

Salichos & Rokas (2013) Nature

14,536 / 14,536

Jarvis et al. (2014) Science

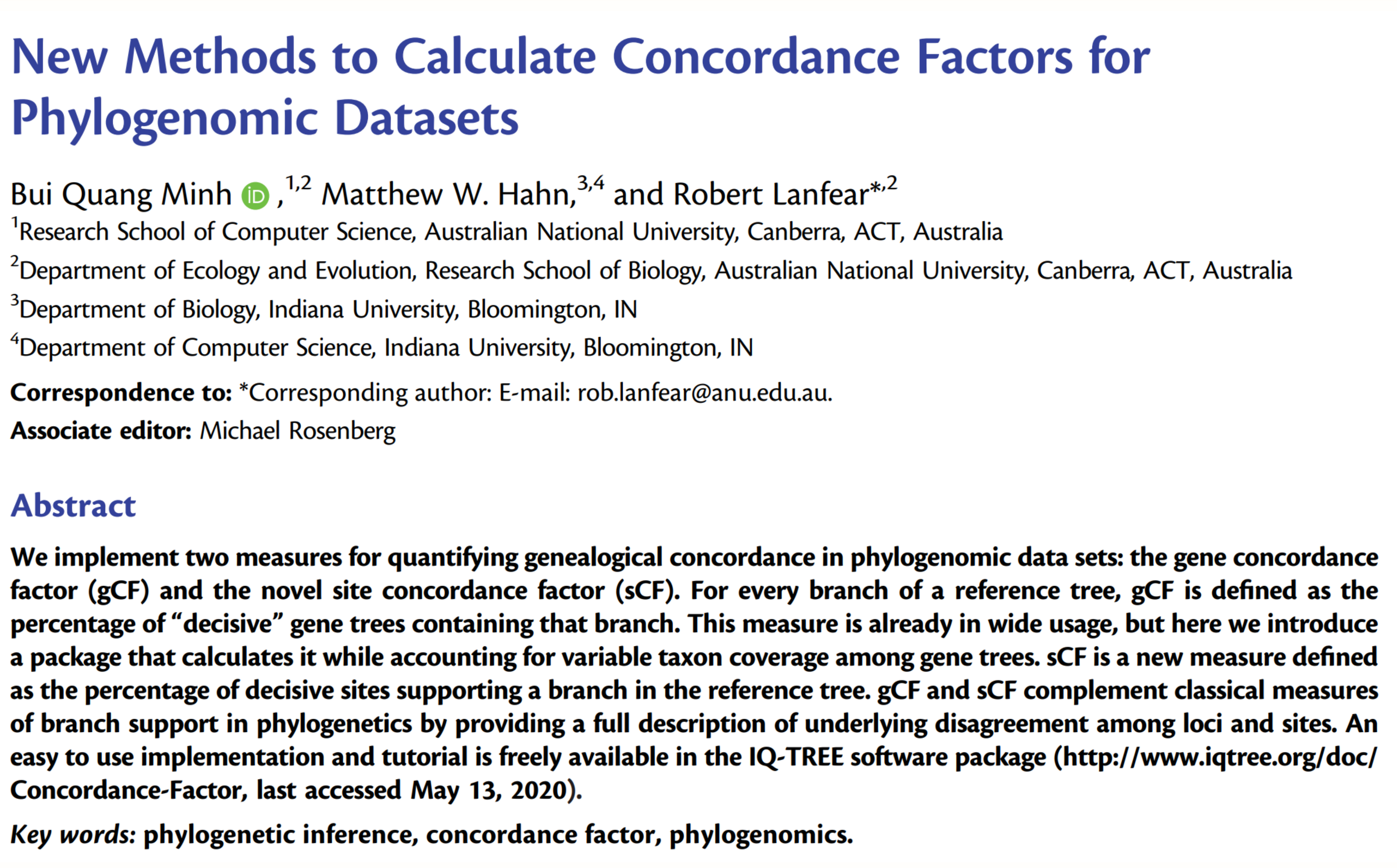
The yeast phylogeny inferred by majority-rule consensus


Saccharomyces lineage
99
41
31
36
62
24
29
29
99
52
60
77
95
0.2
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosacharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida lusitaniae (Clus)
98
90
86
89
45
29
29
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
lineage Candida

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

Salichos & Rokas (2013) Nature

Gene and site concordance factors



New Methods to Calculate Concordance Factors for Phylogenomic Datasets

Bui Quang Minh ,[1,2] Matthew W. Hahn,[3,4] and Robert Lanfear*,[2]

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

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

[3]Department of Biology, Indiana University, Bloomington, IN

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

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

Associate editor: Michael Rosenberg

Abstract

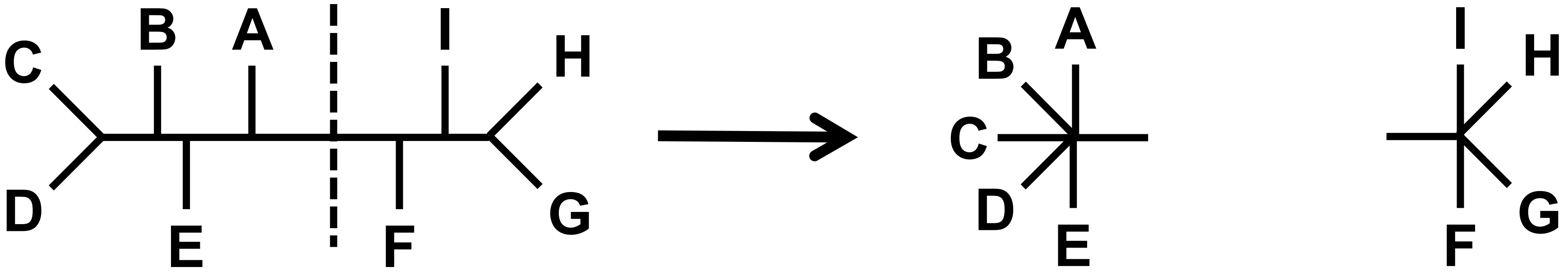
We implement two measures for quantifying genealogical concordance in phylogenomic data sets: the gene concordance factor (gCF) and the novel site concordance factor (sCF). For every branch of a reference tree, gCF is defined as the percentage of "decisive" gene trees containing that branch. This measure 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.**

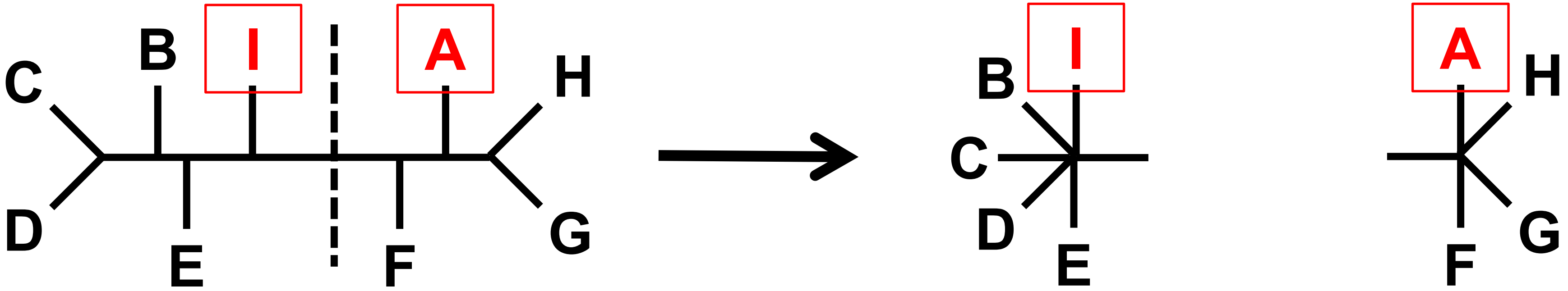
Minh et al. (2020) Mol. Biol. Evol.

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}



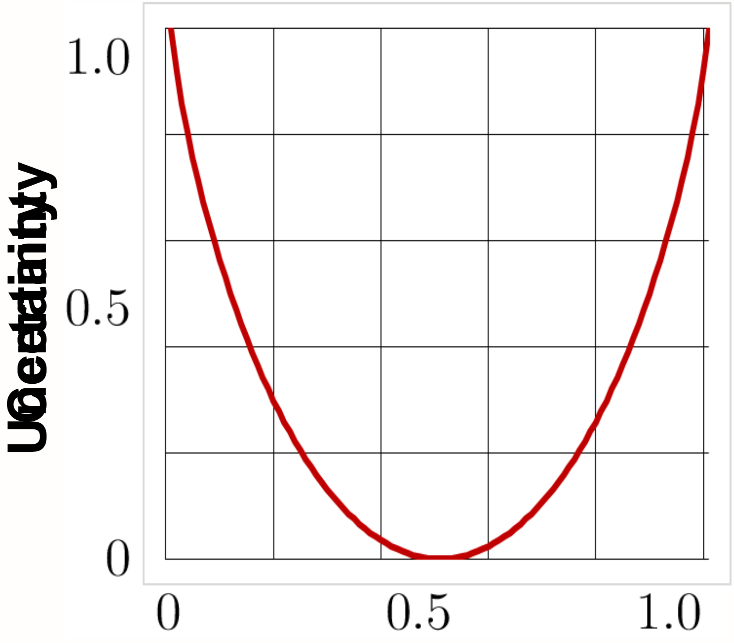
***Conflicting* set of splits: {I, B, C, D, E} {F, G, H, A}**

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)

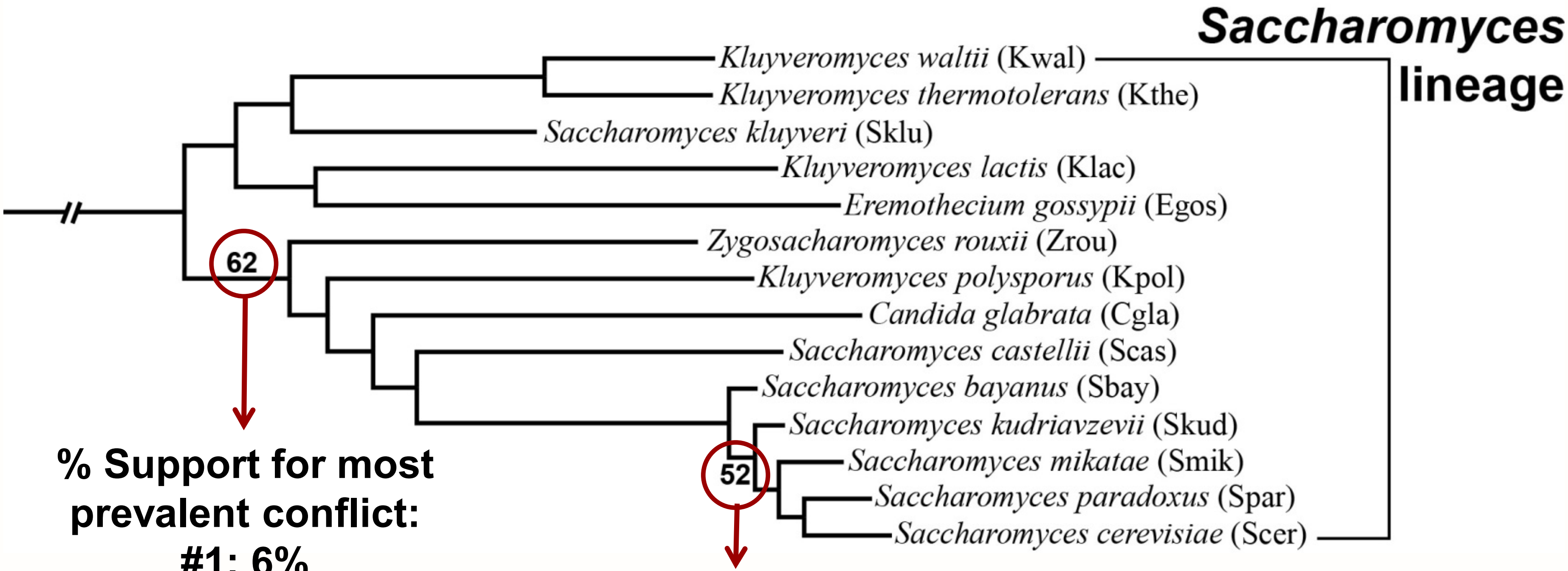


Ratio of 'Heads/Tails'

Ratio of Support for Two Conflicting Bipartitions

Salichos et al. (2014) Mol. Biol. Evol.; Kobert et al. (2016) Mol. Biol. Evol.; Zhou et al. (2020) Syst. Biol.

IC is a more informative measure of branch support


Saccharomyces lineage
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosacharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
62
52

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

IC value: 0.59

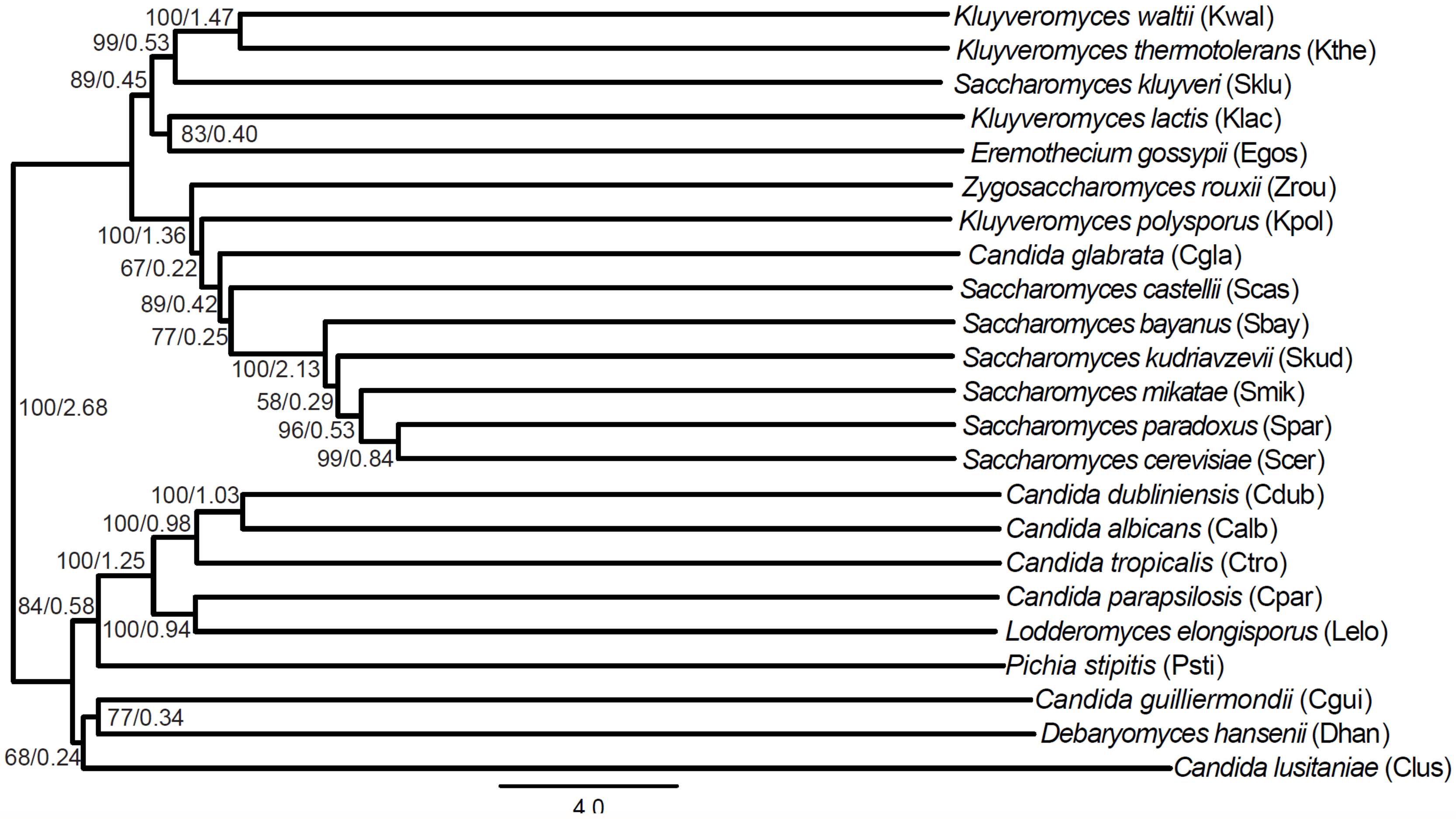
% Support for most prevalent conflict: #1: 29%

IC value: 0.06

Salichos & Rokas (2013) Nature

Coalescent-based inference of the yeast phylogeny

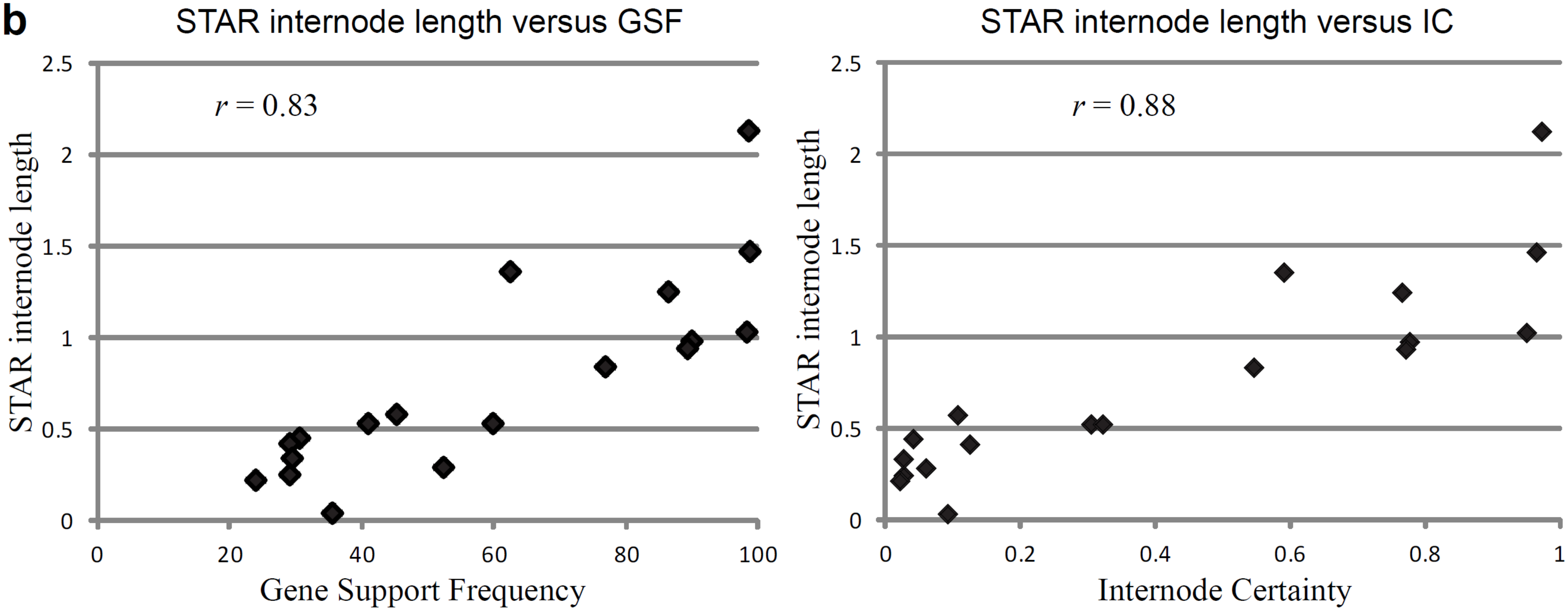
The yeast species phylogeny inferred using the STAR species tree method


100/1.47
99/0.53
89/0.45
83/0.40
100/1.36
67/0.22
89/0.42
77/0.25
100/2.13
58/0.29
96/0.53
99/0.84
100/2.68
100/1.03
100/0.98
100/1.25
84/0.58
100/0.94
77/0.34
68/0.24
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosaccharomyces rouxii (Zrou)
Kluyveromyces polysporus (Kpol)
Candida glabrata (Cgla)
Saccharomyces castellii (Scas)
Saccharomyces bayanus (Sbay)
Saccharomyces kudriavzevii (Skud)
Saccharomyces mikatae (Smik)
Saccharomyces paradoxus (Spar)
Saccharomyces cerevisiae (Scer)
Candida dubliniensis (Cdub)
Candida albicans (Calb)
Candida tropicalis (Ctro)
Candida parapsilosis (Cpar)
Lodderomyces elongisporus (Lelo)
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
Debaryomyces hansenii (Dhan)
Candida lusitaniae (Clus)
4.0

Coalescent units / IC

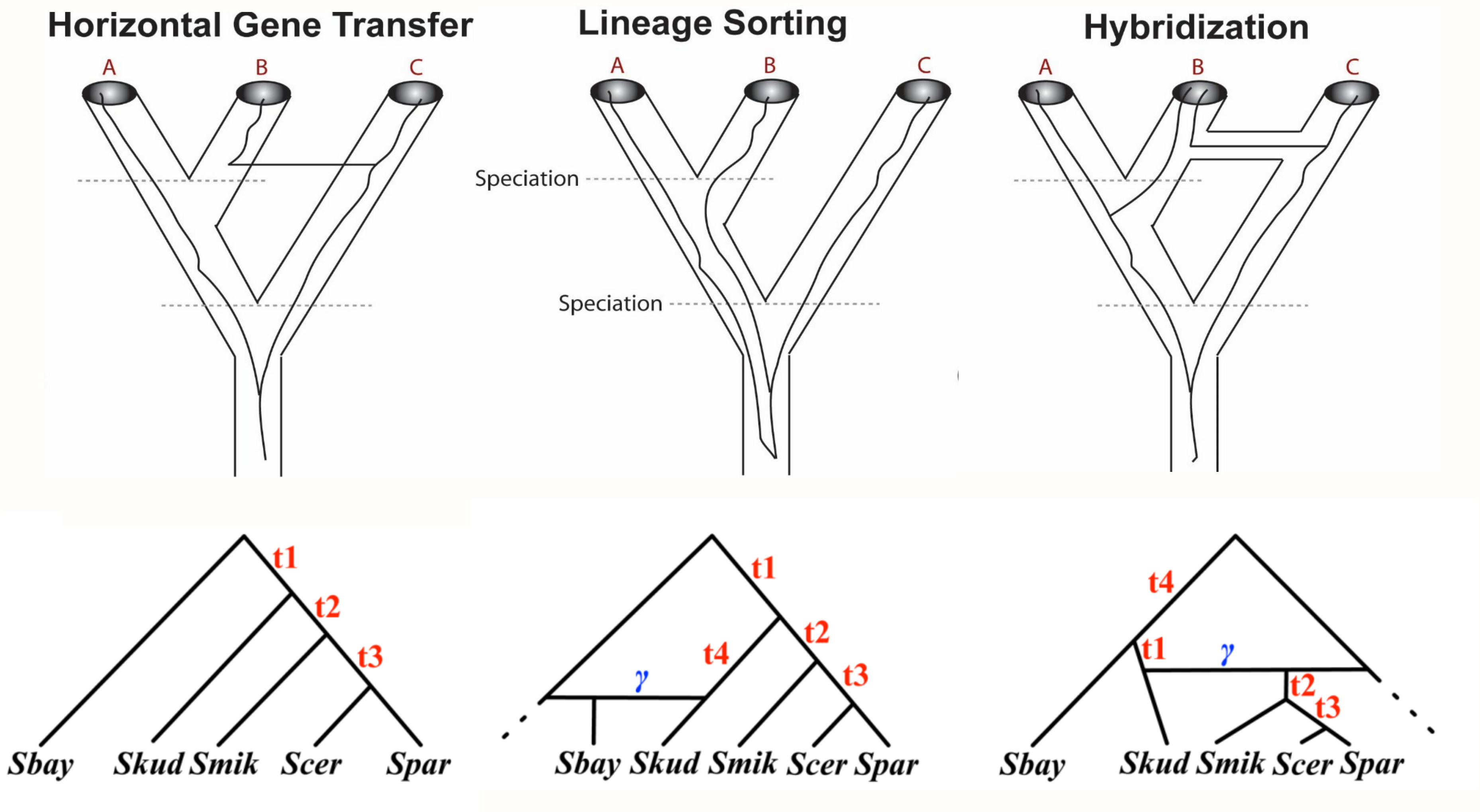
Salichos & Rokas (2013) Nature

Coalescent branch lengths are correlated with GSF/IC



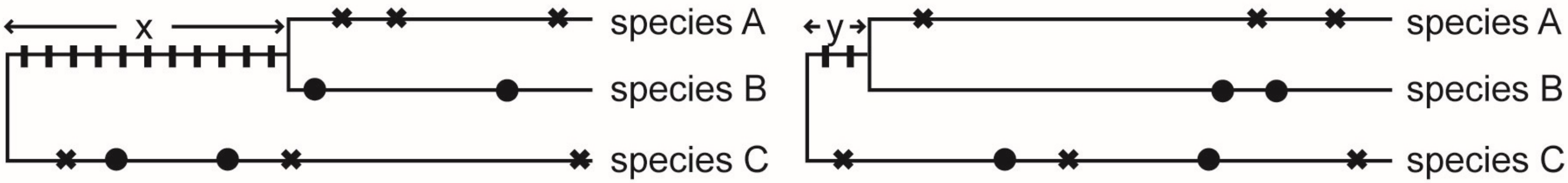
Salichos & Rokas (2013) Nature

Why so much incongruence? Biological factors

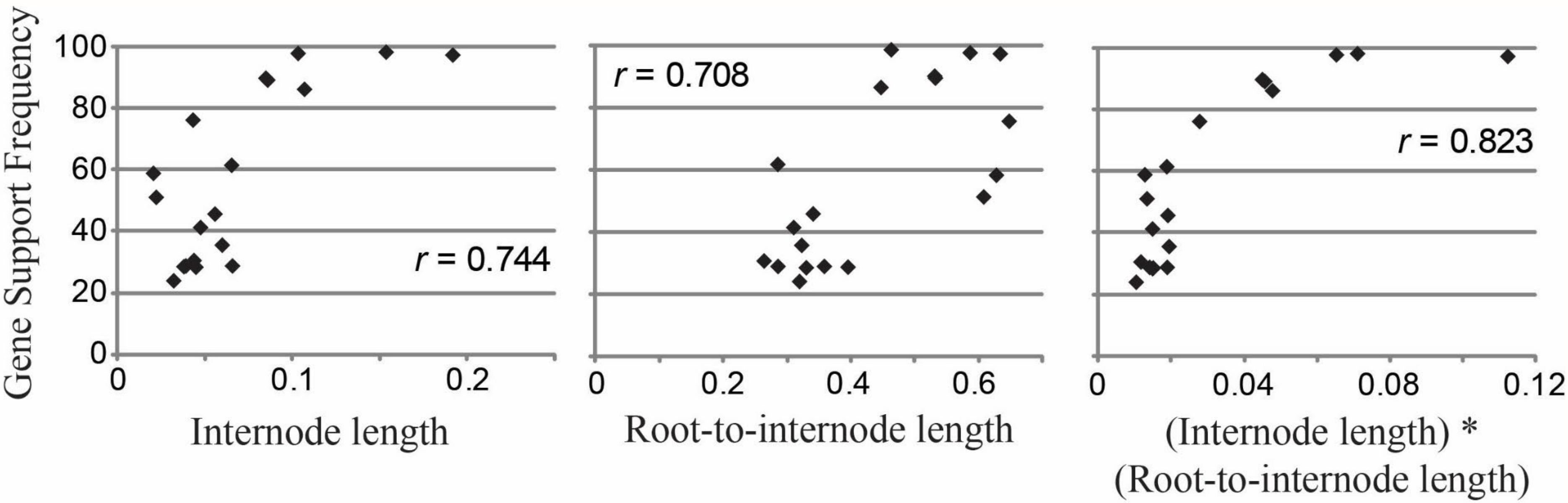


Yu et al. (2012) PLOS Genet.

Why so much incongruence? Analytical factors



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



Rokas & Carroll (2006) PLOS Biol. *Salichos & Rokas (2013) Nature*

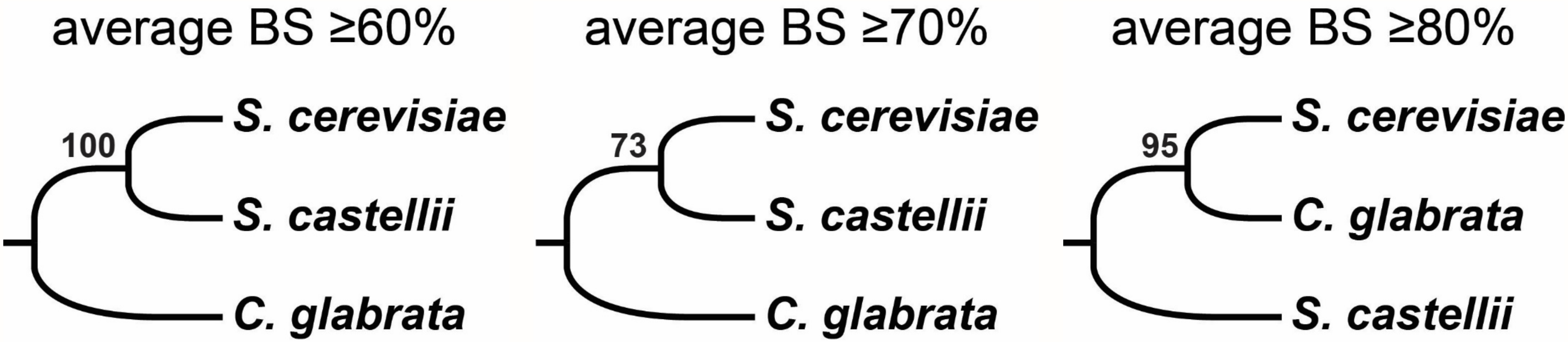
Certain recipes for handling incongruence didn't help

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

Salichos & Rokas (2013) Nature

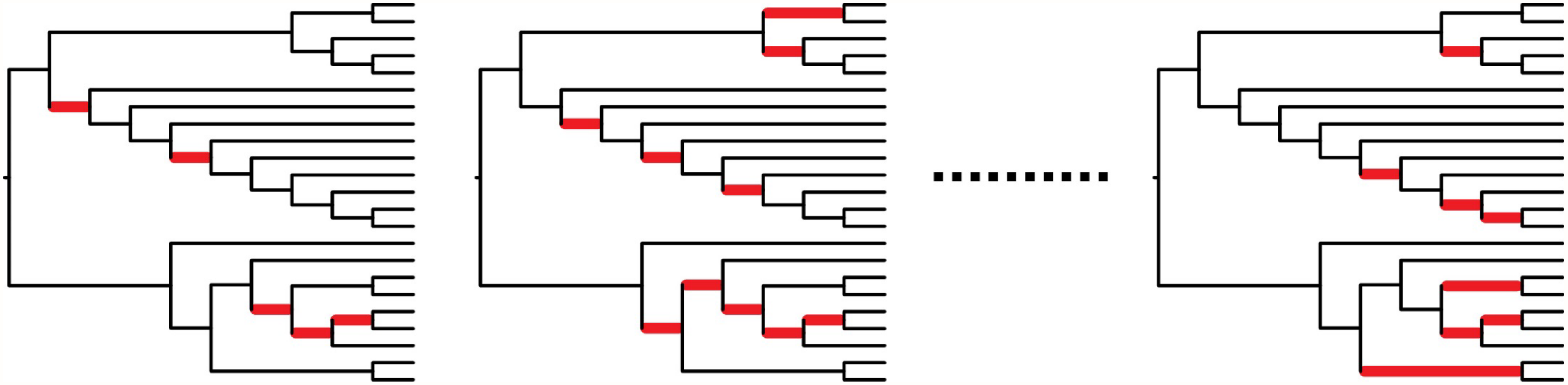
What do we do then?

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



Salichos & Rokas (2013) Nature

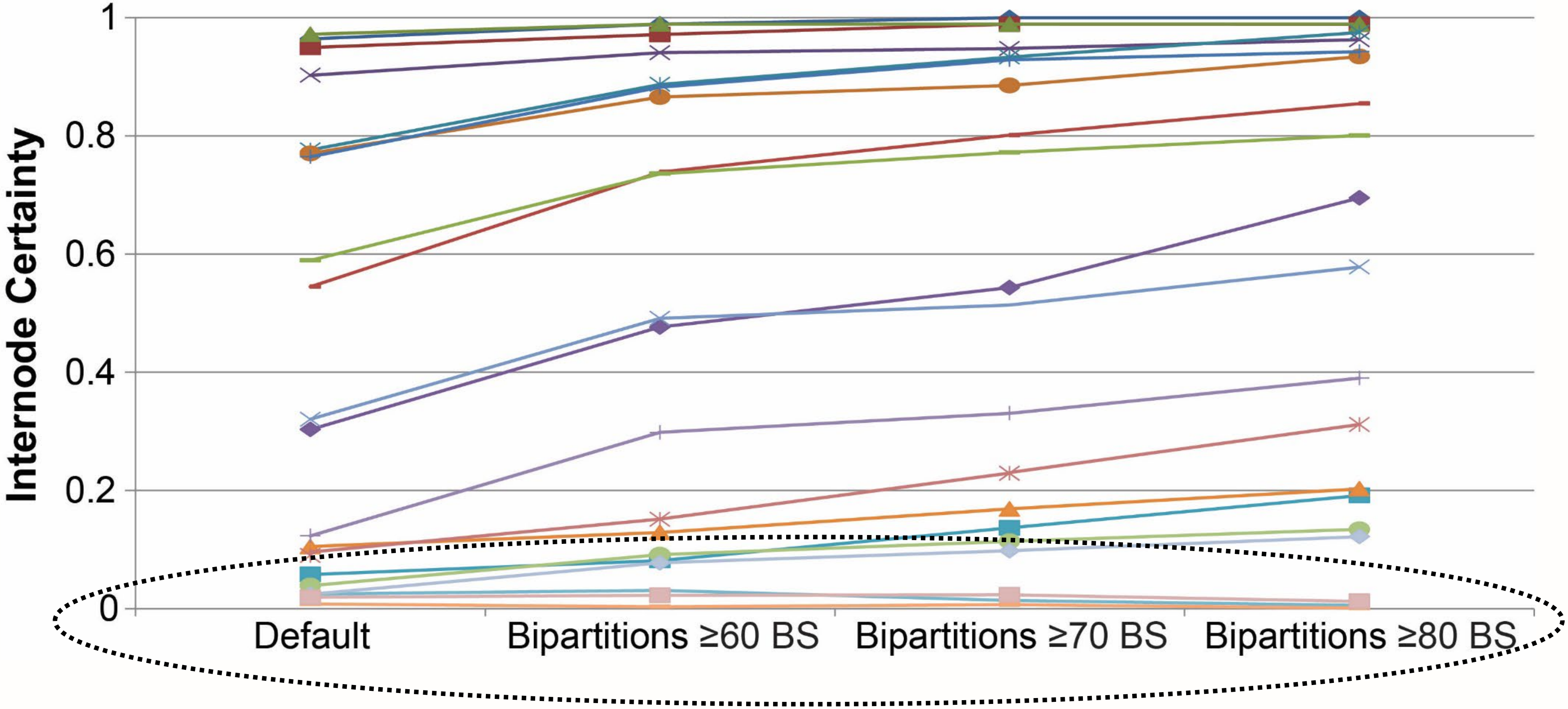
Selecting specific bipartitions dramatically improves phylogeny



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

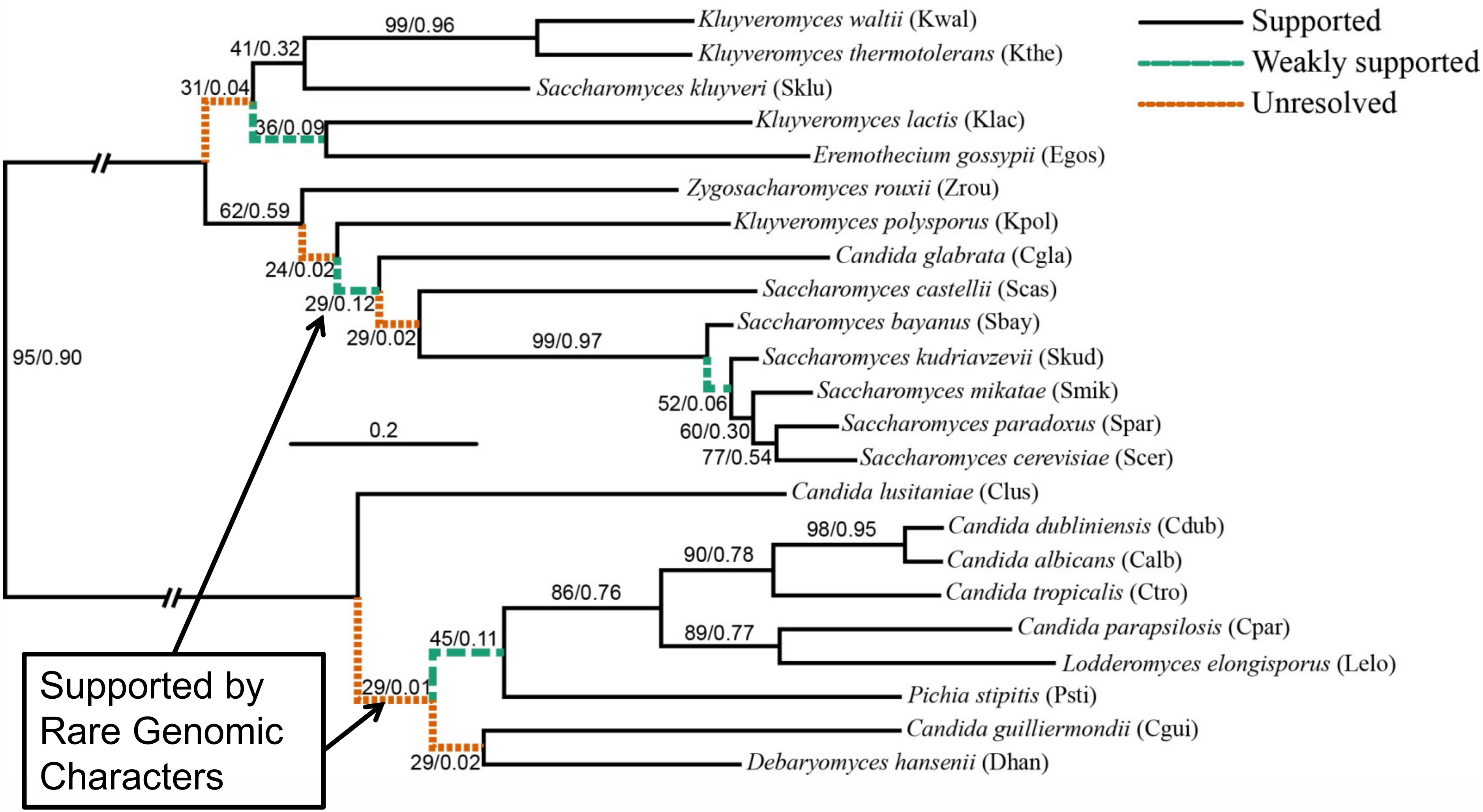
Salichos & Rokas (2013) Nature

Least supported internodes harbor the most conflict



Salichos & Rokas (2013) Nature

The status of the yeast phylogeny


Supported
Weakly supported
Unresolved
99/0.96
41/0.32
31/0.04
36/0.09
Kluyveromyces waltii (Kwal)
Kluyveromyces thermotolerans (Kthe)
Saccharomyces kluyveri (Sklu)
Kluyveromyces lactis (Klac)
Eremothecium gossypii (Egos)
Zygosacharomyces rouxii (Zrou)
62/0.59
Kluyveromyces polysporus (Kpol)
24/0.02
Candida glabrata (Cgla)
29/0.12
Saccharomyces castellii (Scas)
29/0.02
Saccharomyces bayanus (Sbay)
99/0.97
Saccharomyces kudriavzevii (Skud)
52/0.06
Saccharomyces mikatae (Smik)
60/0.30
Saccharomyces paradoxus (Spar)
77/0.54
Saccharomyces cerevisiae (Scer)
95/0.90
0.2
Candida lusitaniae (Clus)
98/0.95
Candida dubliniensis (Cdub)
90/0.78
Candida albicans (Calb)
86/0.76
Candida tropicalis (Ctro)
89/0.77
Candida parapsilosis (Cpar)
45/0.11
Lodderomyces elongisporus (Lelo)
29/0.01
Pichia stipitis (Psti)
Candida guilliermondii (Cgui)
29/0.02
Debaryomyces hansenii (Dhan)

Supported by Rare Genomic Characters

Gene Support Frequency (or gene Concordance Factor) / Internode Certainty

Similar results in other lineages

Vertebrates

(1,086 genes, 18 taxa)

Animals

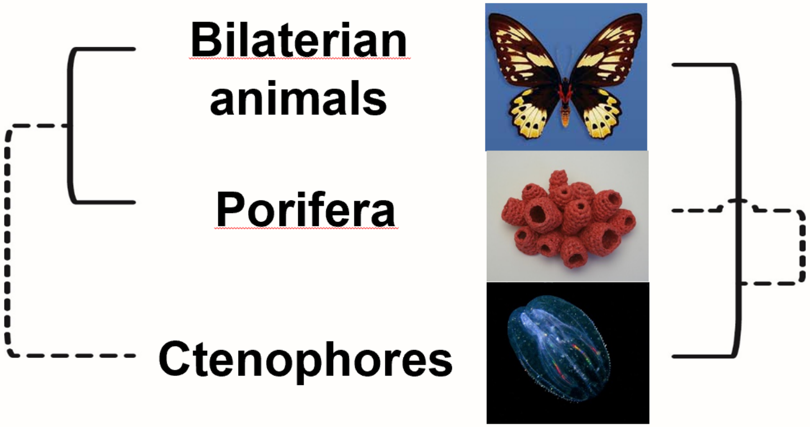
(225 genes, 21 taxa)

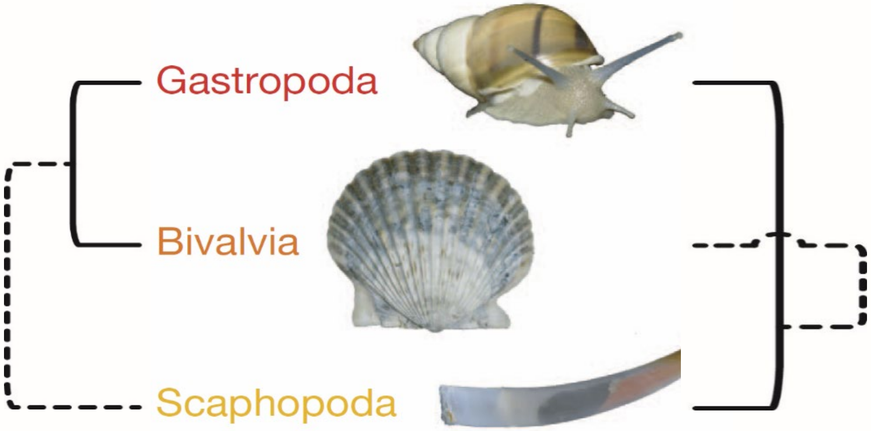
Mosquitoes

(2,007 genes, 20 taxa)

Salichos & Rokas (2013) Nature; Wang et al. (2015) Genome Biol. Evol.

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

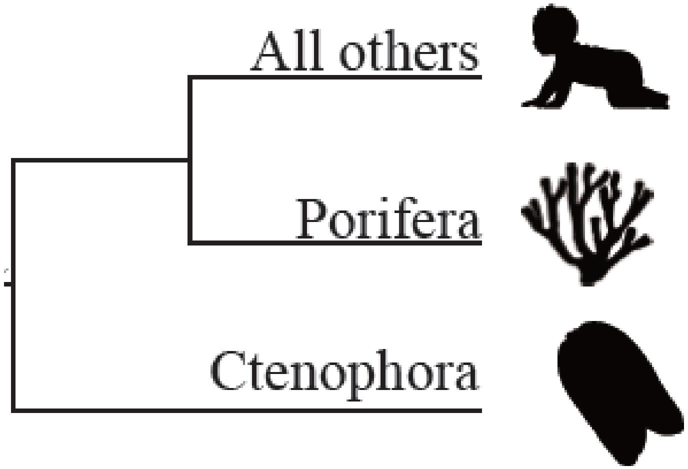
A measure of the substitutions occurring along a given branch of the evolutionary tree. In parsimony methods, the signal is encoded in shared derived characters. In probabilistic methods, the amount of phylogenetic signal actually extracted from a given dataset depends on the model and is expected to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

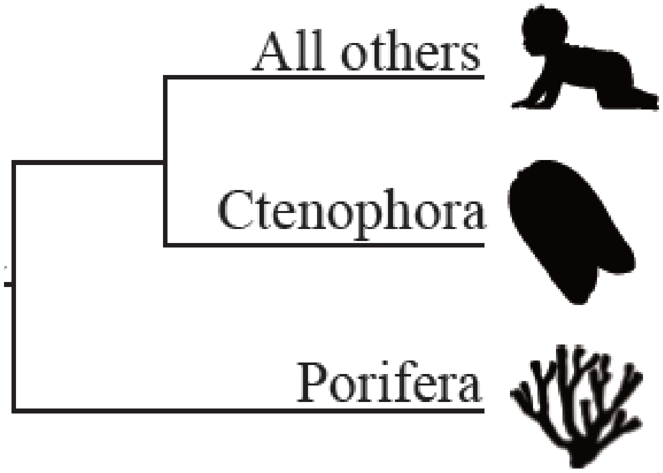
(T1)



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

Conflicting tree

(T2)



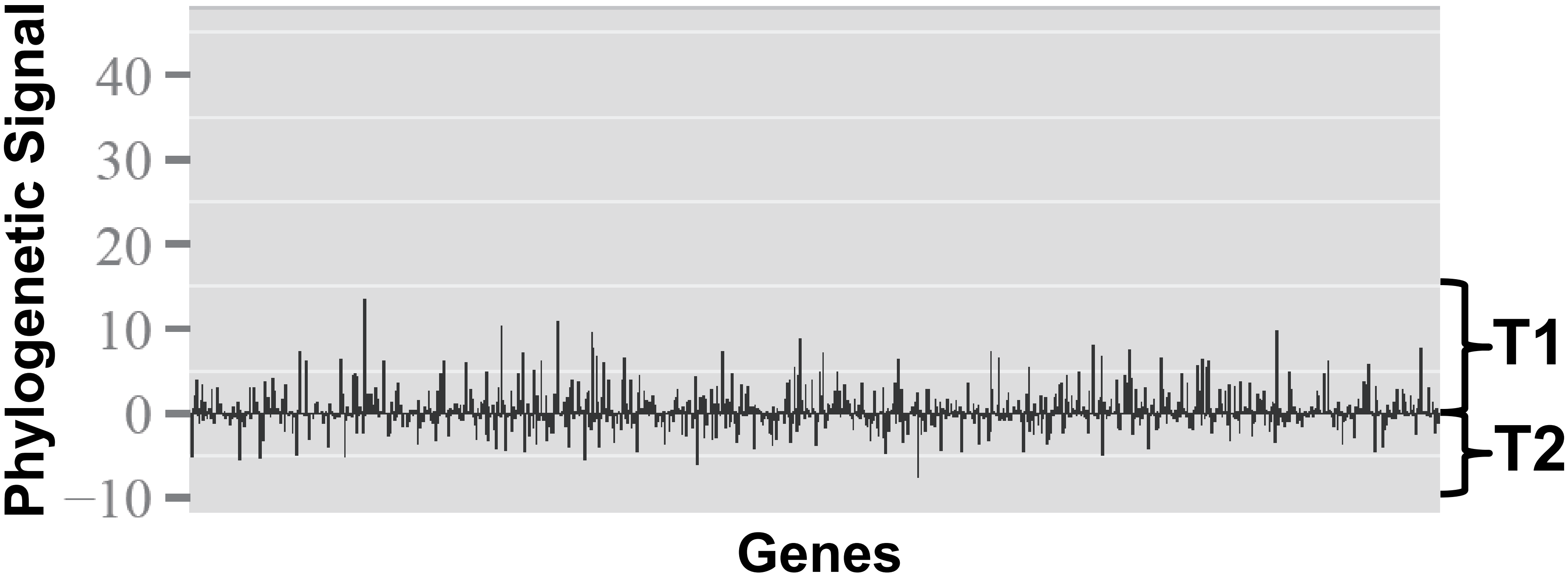
$$\ln(T_2|X_i) = -150$$

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

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

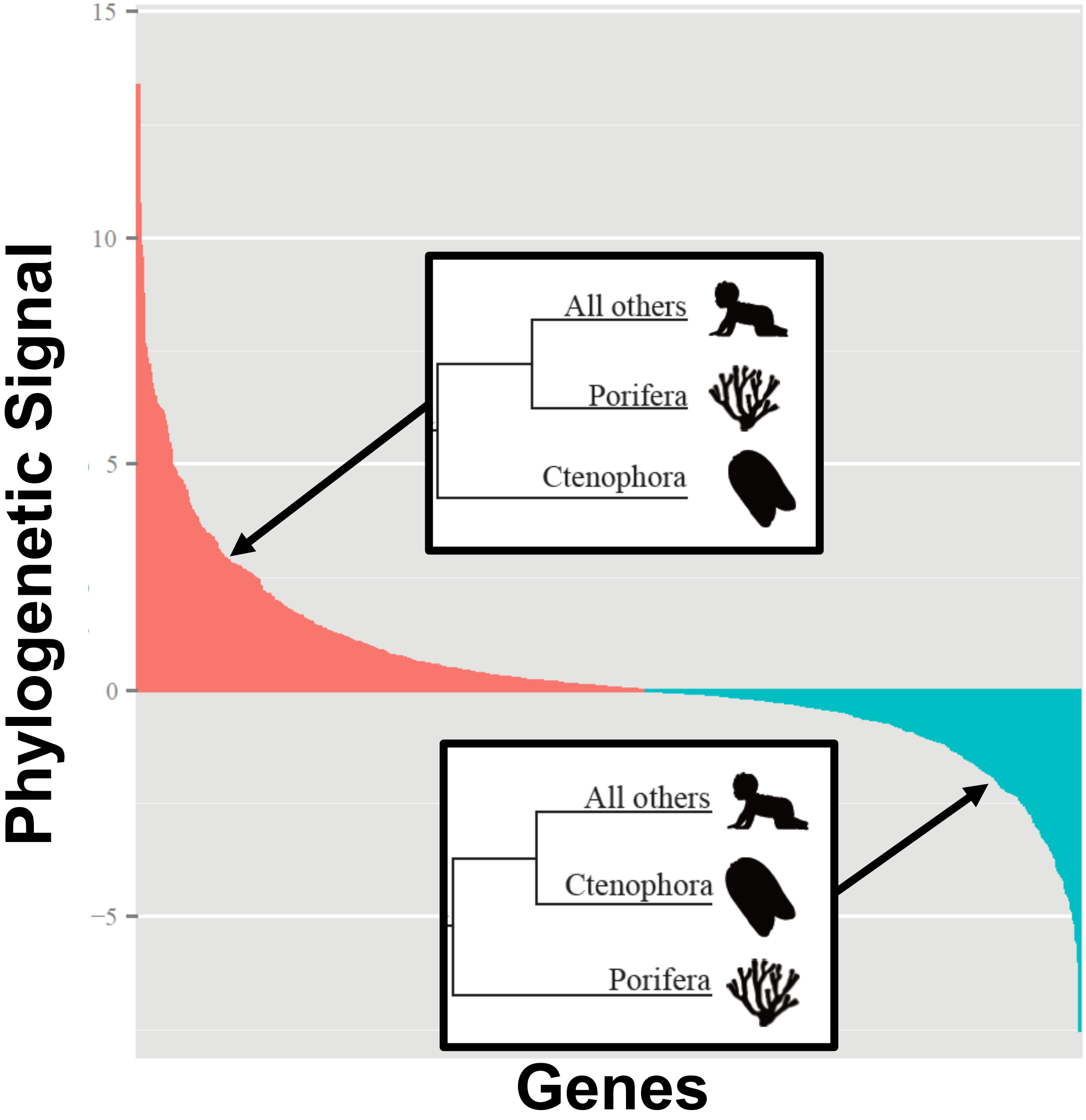
Signal of genes in a phylogenomic data matrix

1,080 genes from 36 animal taxa



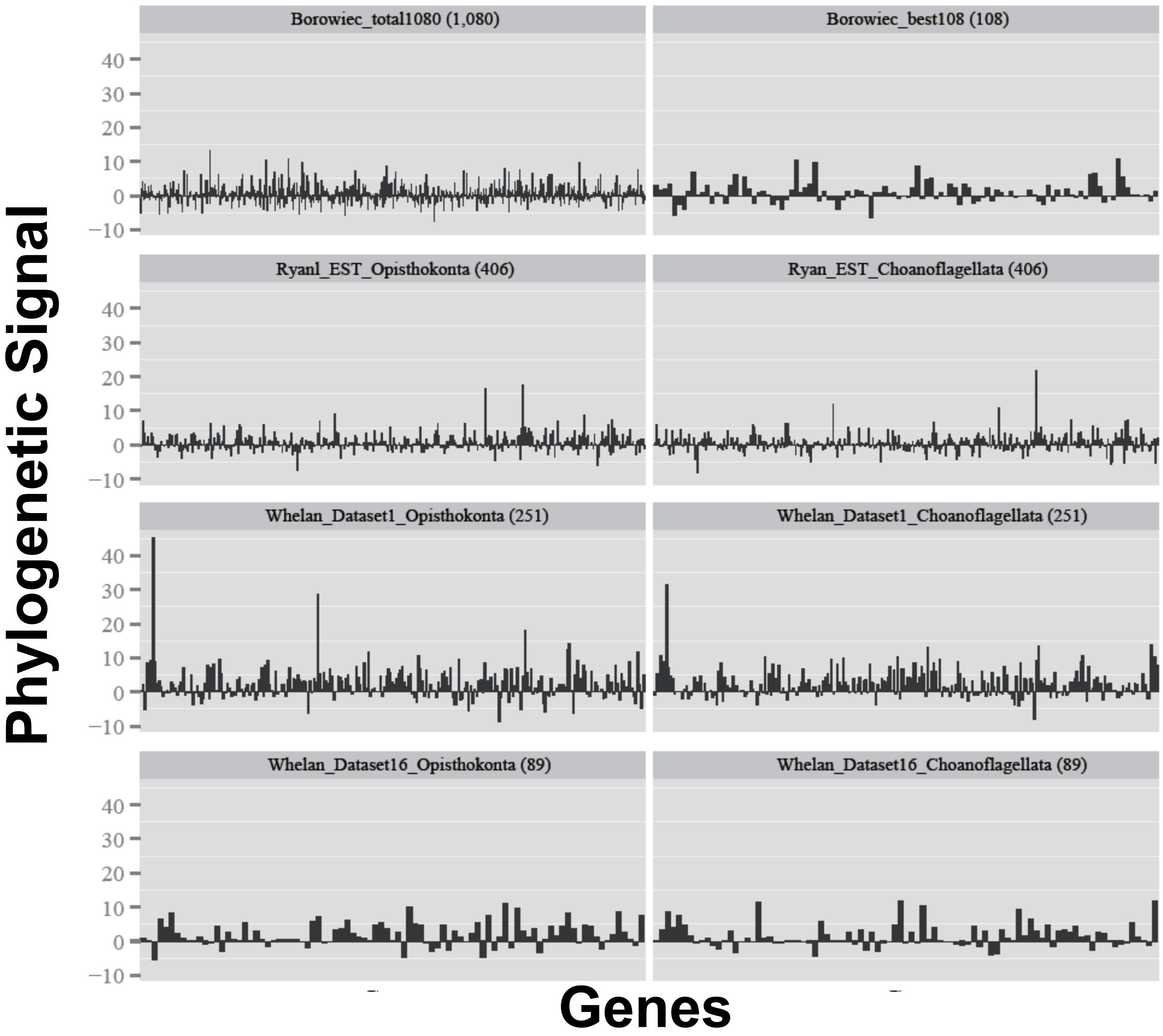
Shen et al. (2017) Nature Ecol. Evol.; data from Borowiec et al. (2015) BMC Genomics

Signal of genes in a phylogenomic data matrix



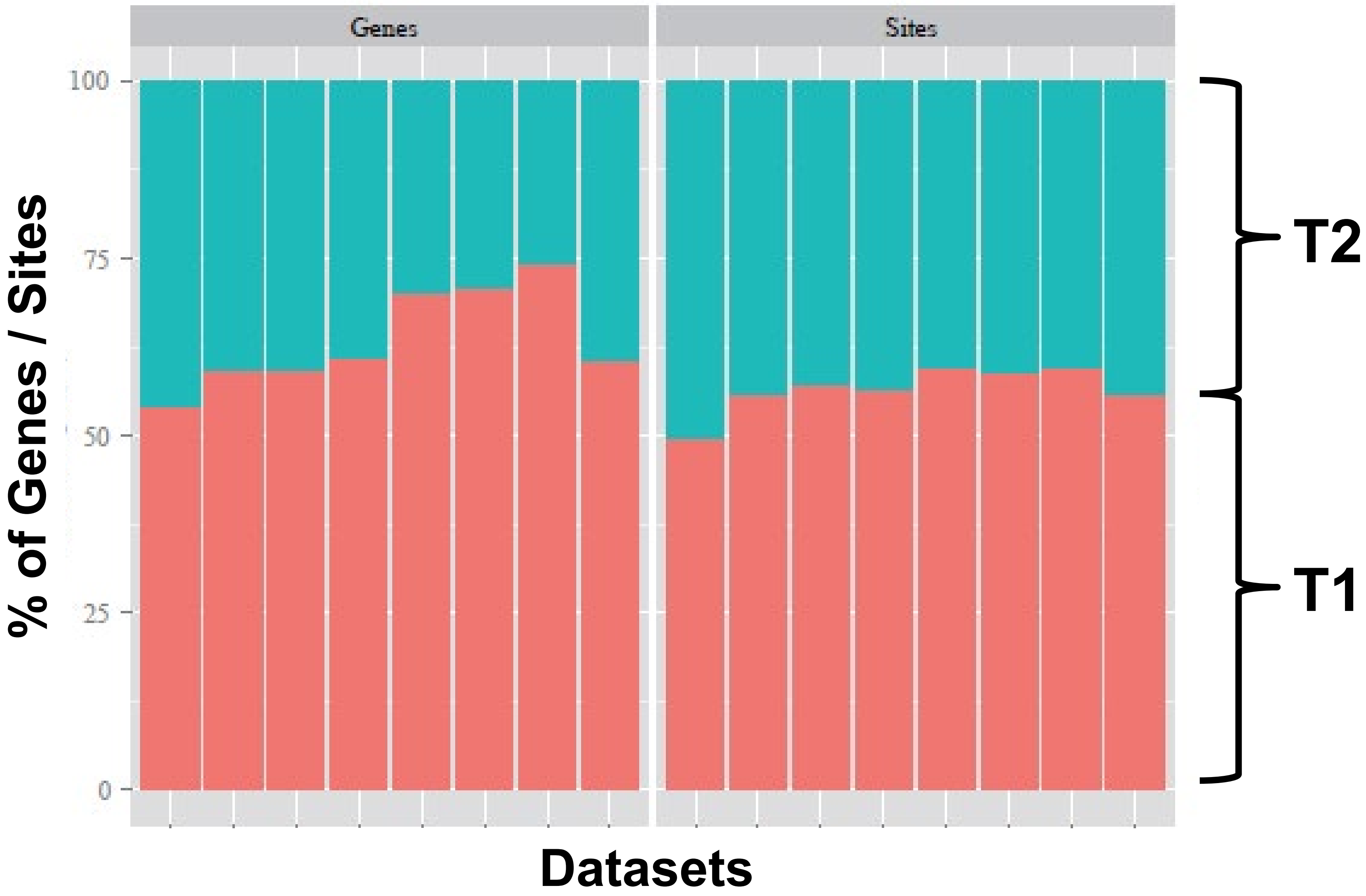
Shen et al. (2017) Nature Ecol. Evol.

Signal of genes in multiple phylogenomic data matrices



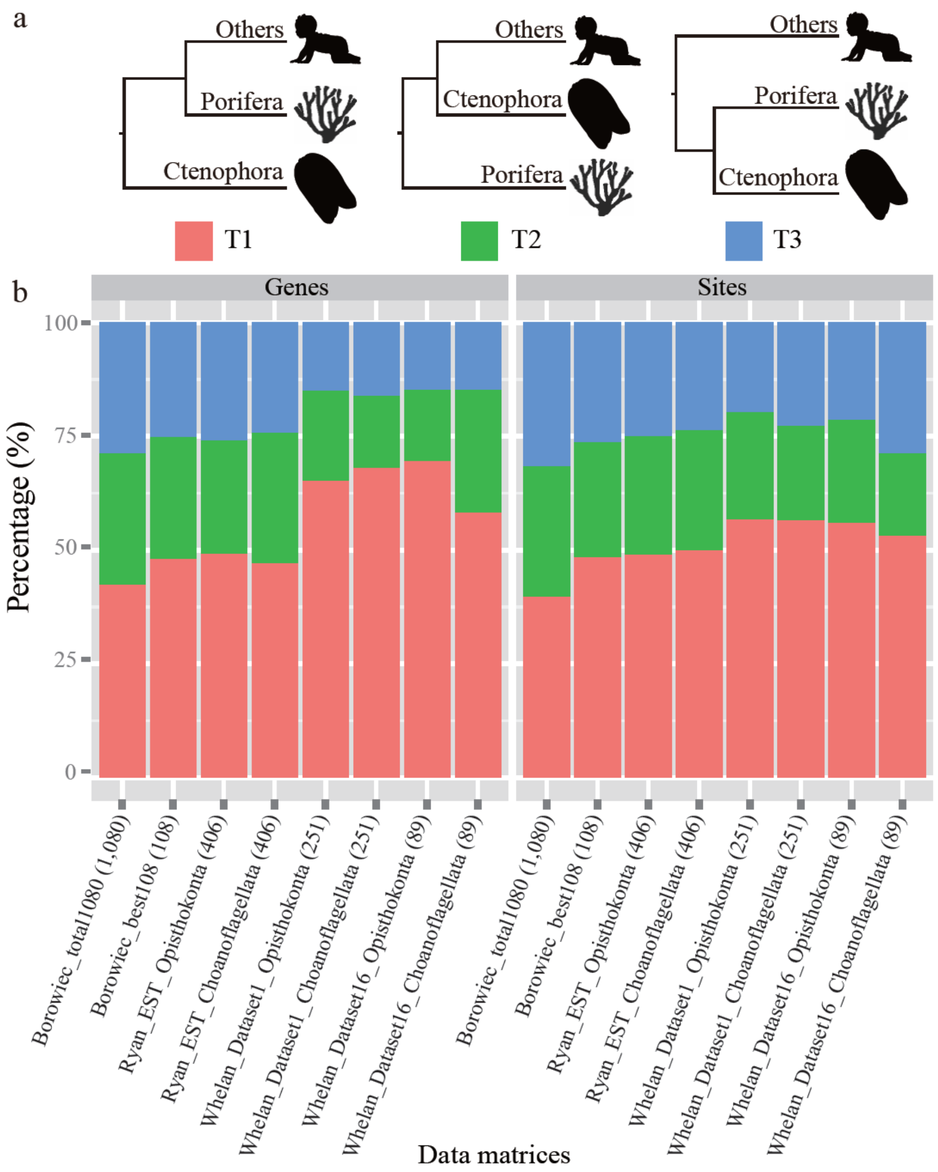
Shen et al. (2017) Nature Ecol. Evol.

Summarizing phylogenetic signal across genes and sites



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

Summarizing the signal across all 3 possible topologies



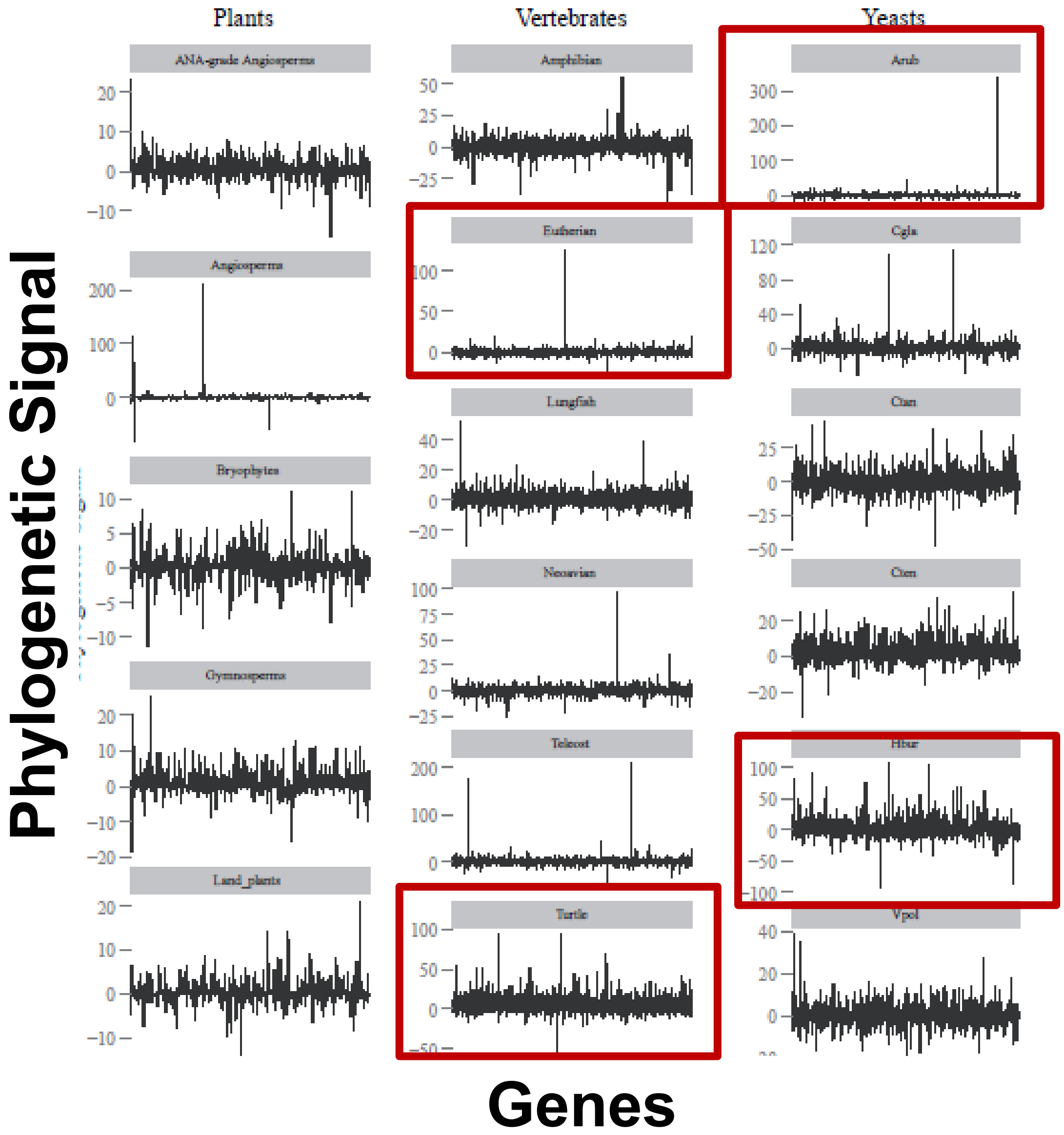
Shen et al. (2017) Nature Ecol. Evol.

Testing several contentious branches of the tree of life

Clade	ML Tree (T1)	Conflicting Tree (T2)
Plants	*Amborella* as sister to all other flowering plants	*Amborella* + *Nuphar* 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
	Turtles as sister to archosaurs (birds + crocodiles)	Turtles as sister to crocodiles
Yeasts	Ascoideaceae as sister to Phaffomycetaceae + Saccharomycetaceae	Ascoideaceae as sister to a clade comprising Pichiaceae, Debaryomycetaceae, Phaffomycetaceae, and Saccharomycetaceae
	Candida glabrata rather than *Naumovozyma castellii* as sister to Saccharomyces sensu stricto yeasts	*Naumovozyma castellii* rather than *Candida glabrata* sister to Saccharomyces sensu stricto yeasts
	Hyphopichia burtonii as sister to *Candida auris* + *Metschnikowia bicuspidata*	*Hyphopichia burtonii* as sister to *Debaryomyces hansenii*
	Zygosaccharomyces rouxii as sister to all other yeasts with occurring whole-genome duplication event	*Vanderwaltozyma polyspora* as sister to all other yeast with occurring whole-genome duplication event
	Meyerozyma guilliermondii as sister to *Debaryomyces hansenii*	*Meyerozyma guilliermondii* as sister to *Hyphopichia burtonii* + *Candida auris*
	Candida tanzawaensis as sister to *Pichia stipiti* + *Candida maltosa*	*Pichia stipiti* as sister to *Candida tanzawaensis* + *Candida maltosa*

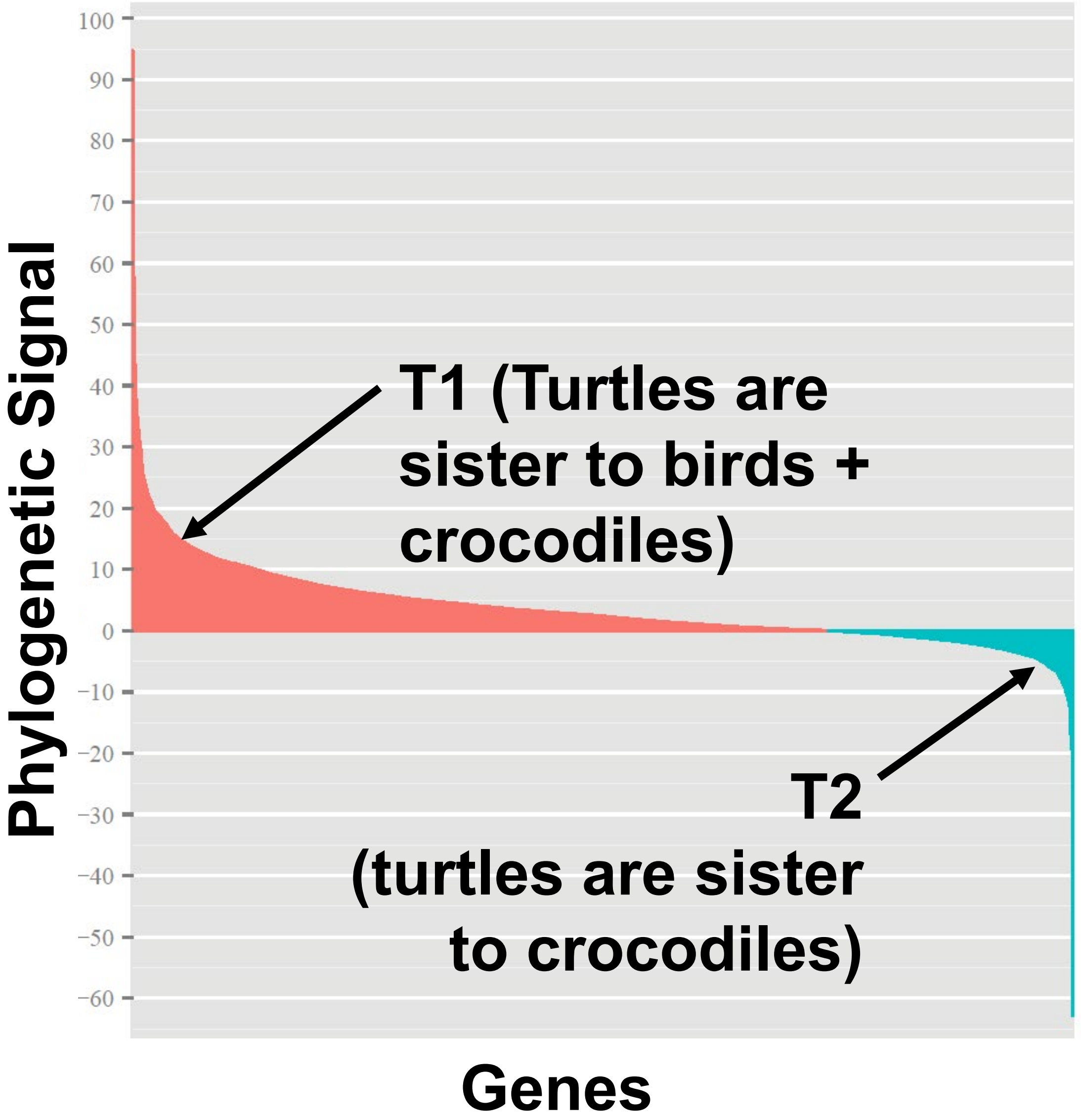
Shen et al. (2017) Nature Ecol. Evol.

Phylogenetic signal in contentious branches of the ToL



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

The signal in some branches is very strong…

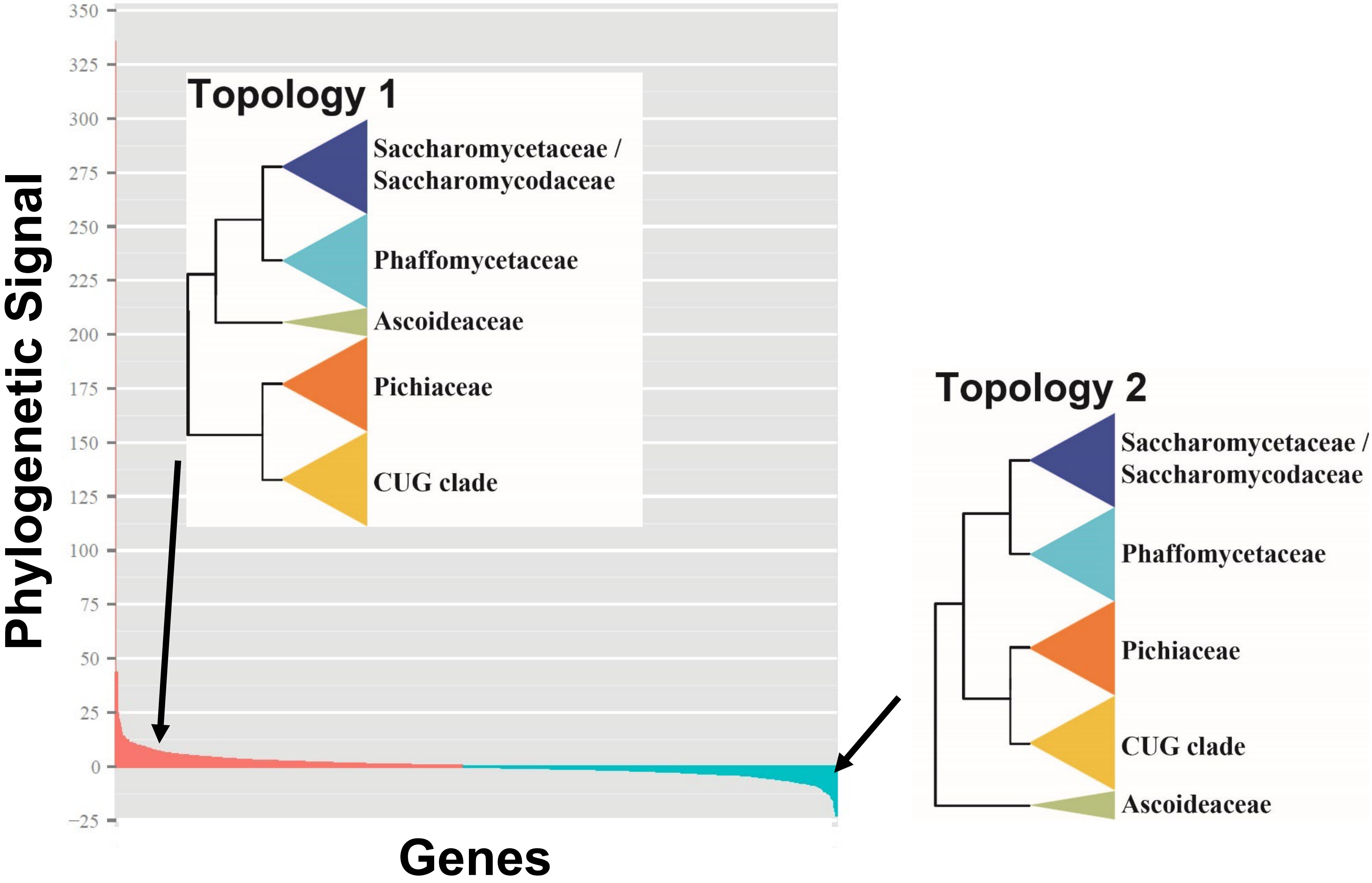


T1 (Turtles are sister to birds + crocodiles)

T2 (turtles are sister to crocodiles)

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

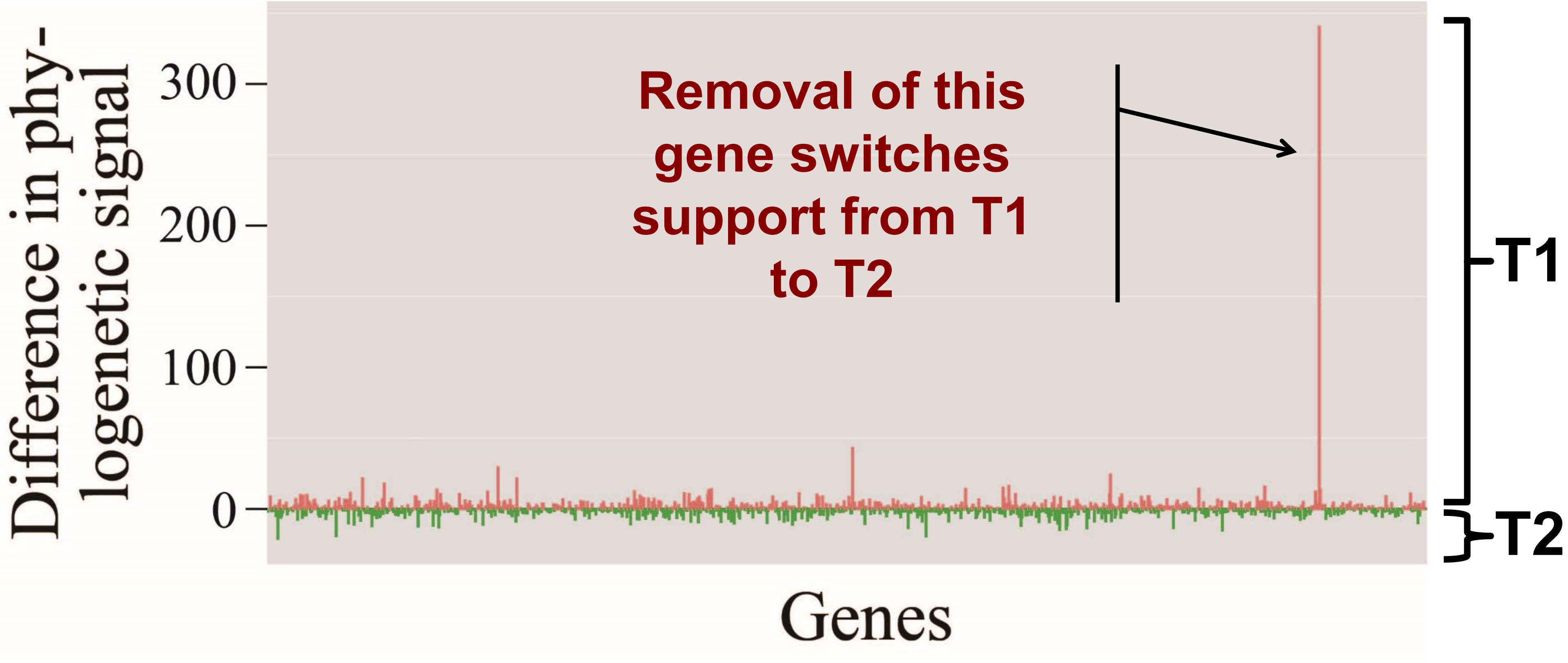
…But in others it stems from one or two genes



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

Phylogenetic signal per gene for the two hypotheses

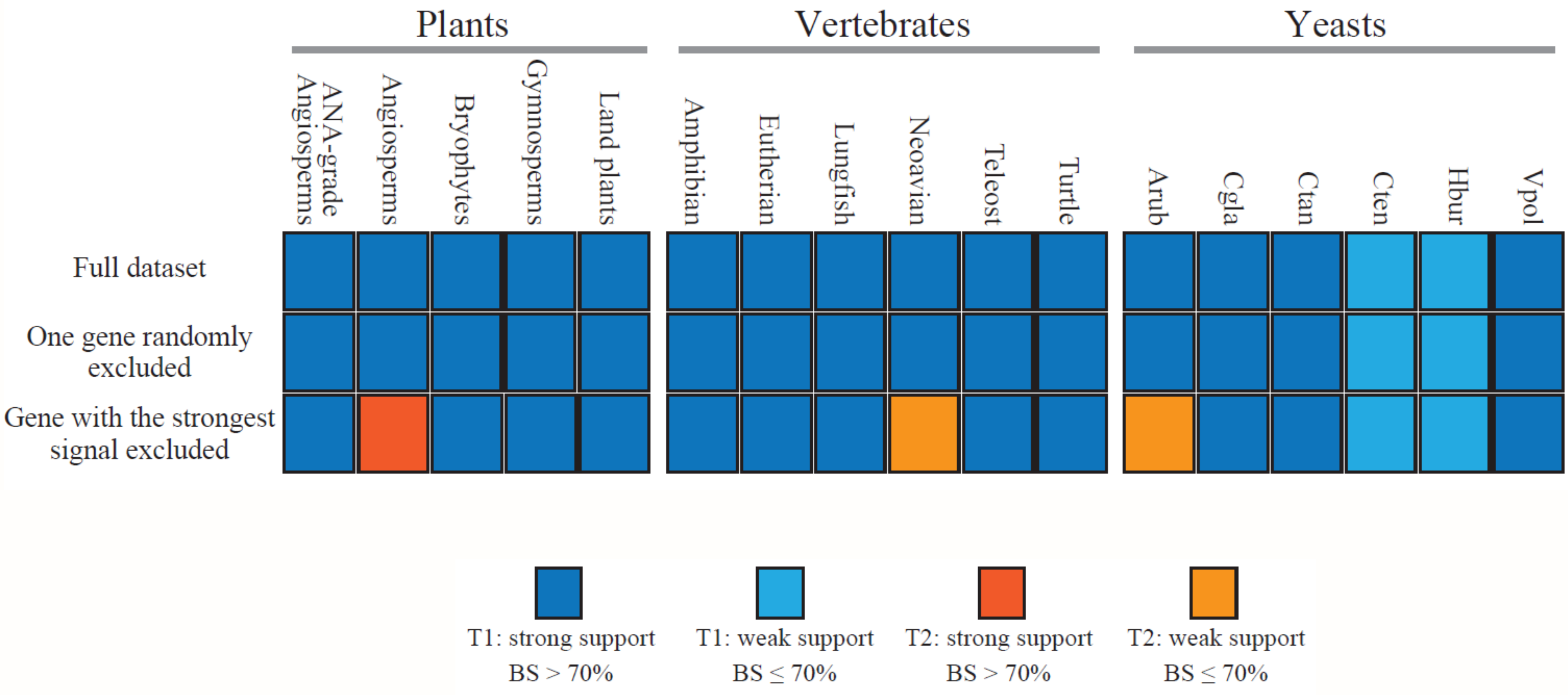
1233 genes, 86 yeast taxa



Removal of this gene switches support from T1 to T2

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

What happens if we remove that one gene?



	Plants					Vertebrates						Yeasts					
	ANA-grade Angiosperms	Angiosperms	Bryophytes	Gymnosperms	Land plants	Amphibian	Eutherian	Lungfish	Neoavian	Teleost	Turtle	Arub	Cgla	Ctan	Cten	Hbur	Vpol
Full dataset	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	▦	▦	🟦
One gene randomly excluded	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	▦	▦	🟦
Gene with the strongest signal excluded	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟧	🟦	🟦	▦	▦	🟦

🟦 T1: strong support BS > 70%

▦ T1: weak support BS ≤ 70%

🟥 T2: strong support BS > 70%

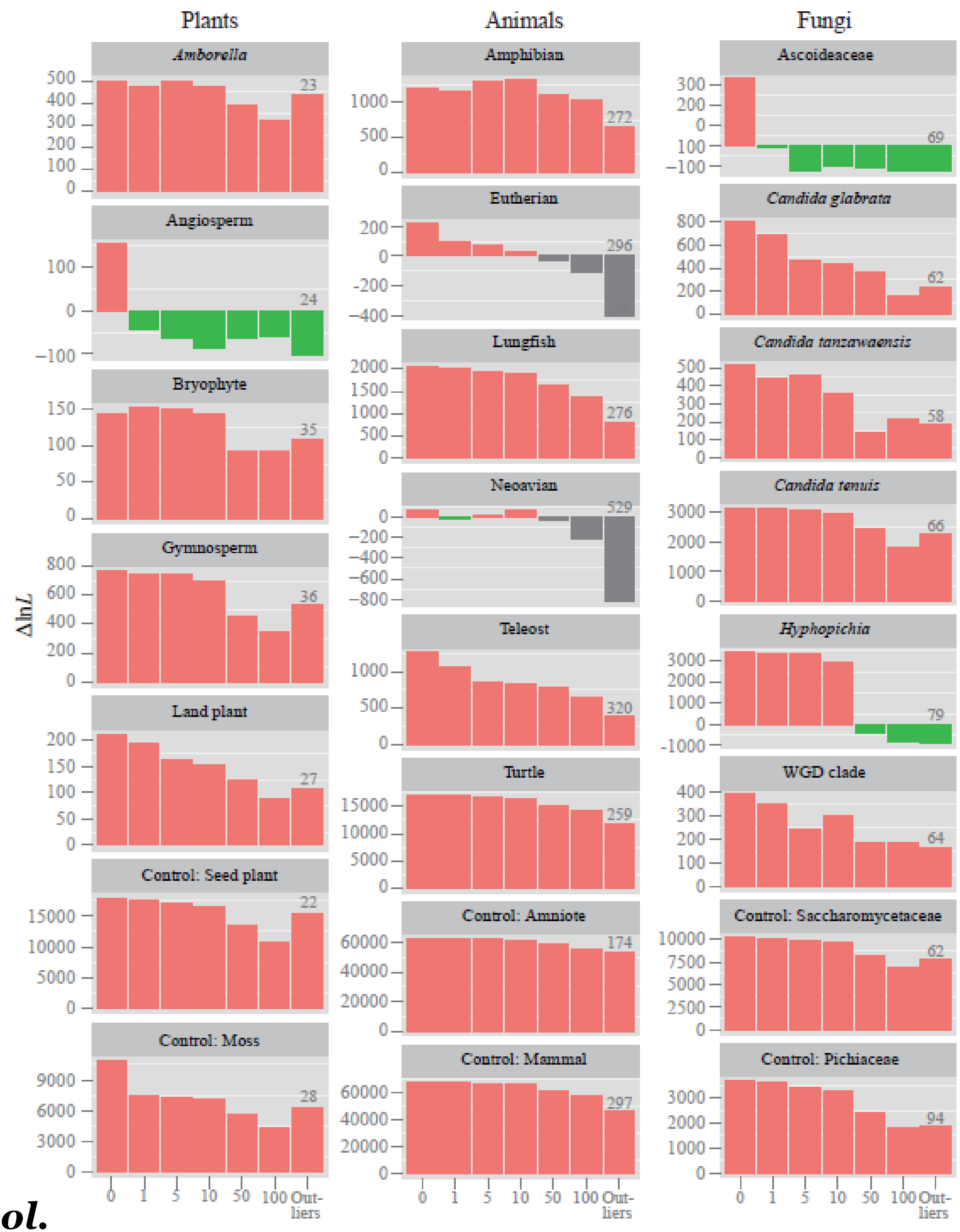
🟧 T2: weak support BS ≤ 70%

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

🟥 T1 🟩 T2 ⬛ Others

# of genes excluded

Quantifying the impact of removing opinionated genes

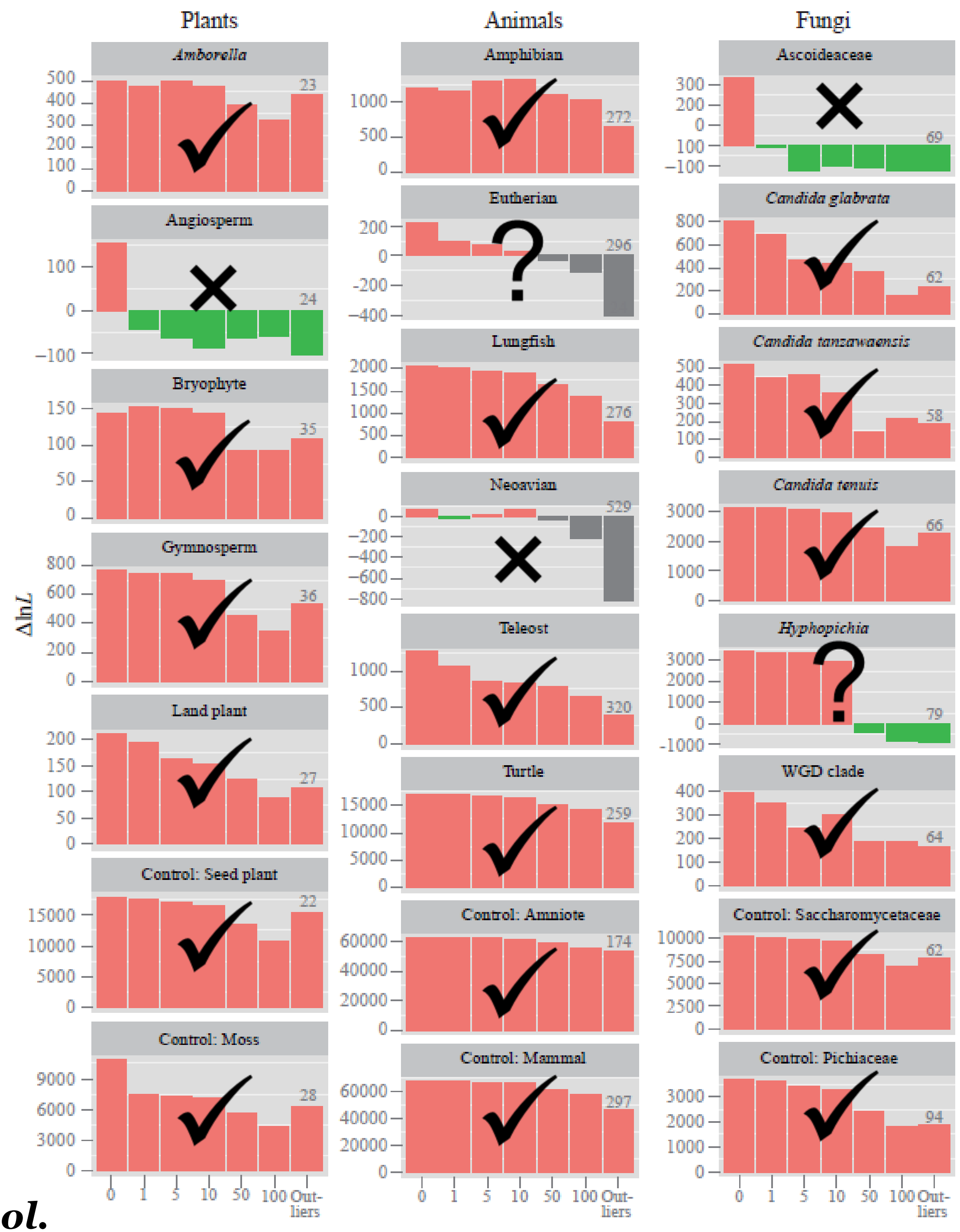


Shen et al. (2017) Nature Ecol. Evol.

🟥 T1 🟩 T2 ⬛ Others

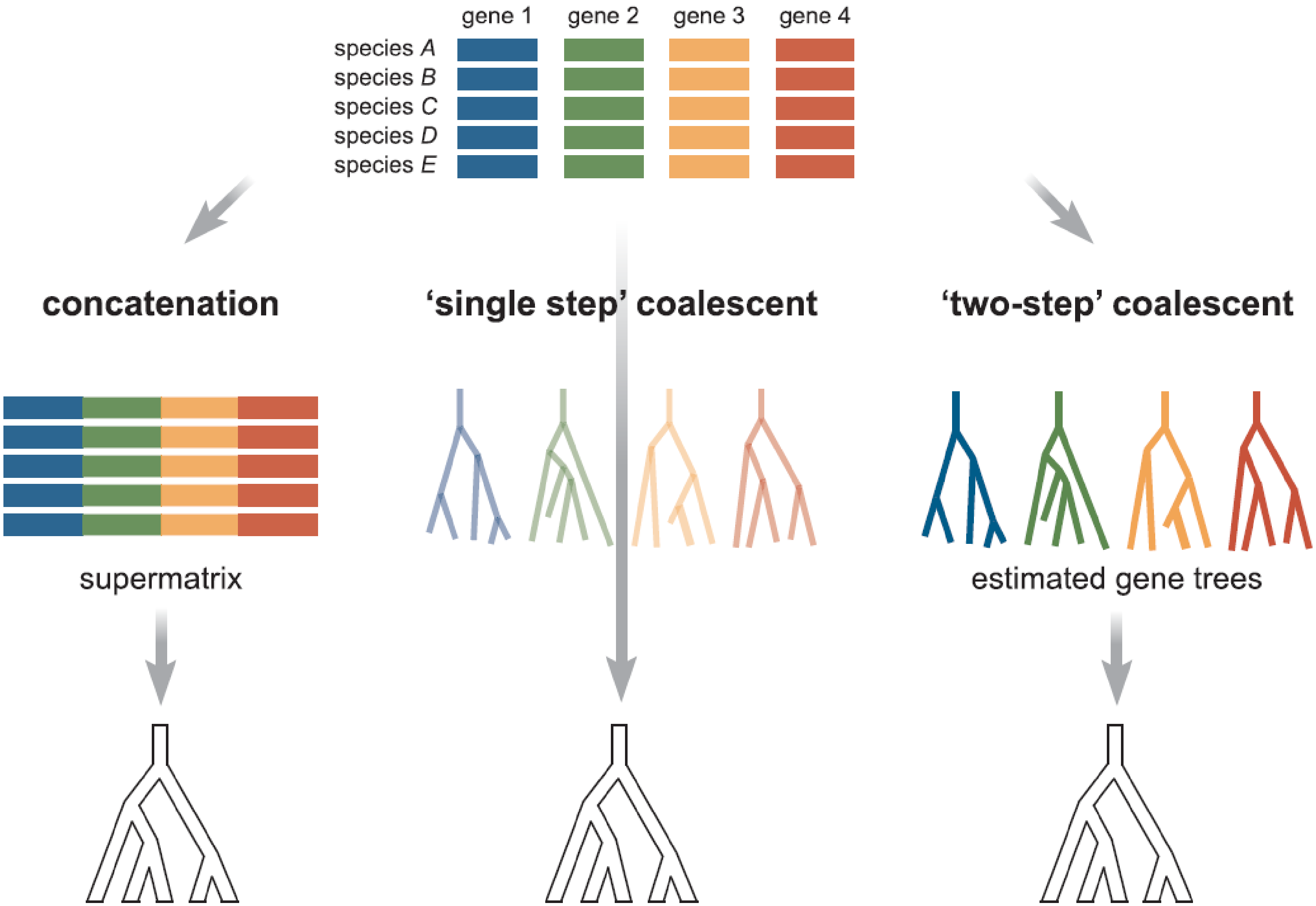
# of genes excluded

Which branches are resolved and which are unresolved?



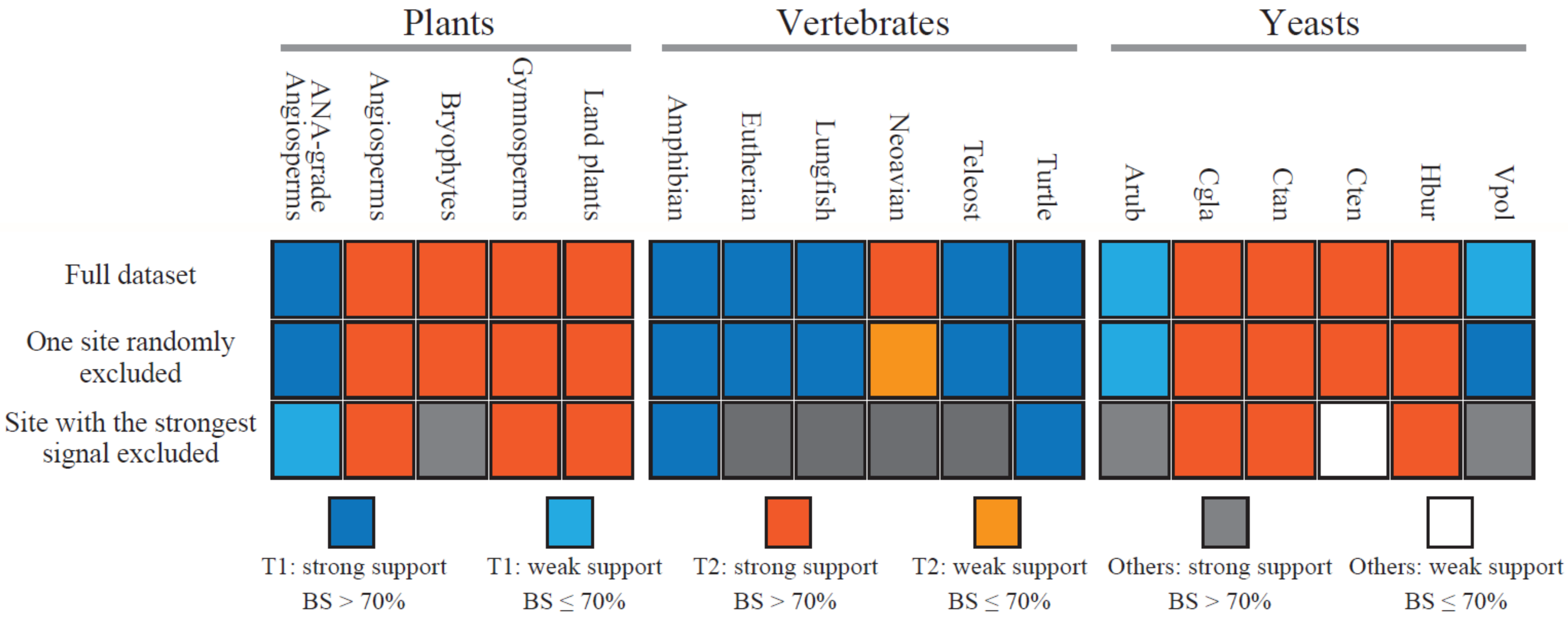
Shen et al. (2017) Nature Ecol. Evol.

Methods for phylogenomic inference



Liu et al. (2015) Ann. N. Y. Acad. Sci.

What happens if we remove one site from every gene?



	Plants					Vertebrates						Yeasts					
	ANA-grade Angiosperms	Angiosperms	Bryophytes	Gymnosperms	Land plants	Amphibian	Eutherian	Lungfish	Neoavian	Teleost	Turtle	Arub	Cgla	Ctan	Cten	Hbur	Vpol
Full dataset	🟦	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟥	🟦	🟦	🔷	🟥	🟥	🟥	🟥	🔷
One site randomly excluded	🟦	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟧	🟦	🟦	🔷	🟥	🟥	🟥	🟥	🟦
Site with the strongest signal excluded	🔷	🟥	⬛	🟥	🟥	🟦	⬛	⬛	⬛	⬛	🟦	⬛	🟥	🟥	⬜	🟥	⬛

🟦 T1: strong support BS > 70%

🔷 T1: weak support BS ≤ 70%

🟥 T2: strong support BS > 70%

🟧 T2: weak support BS ≤ 70%

⬛ Others: strong support BS > 70%

⬜ Others: weak support BS ≤ 70%

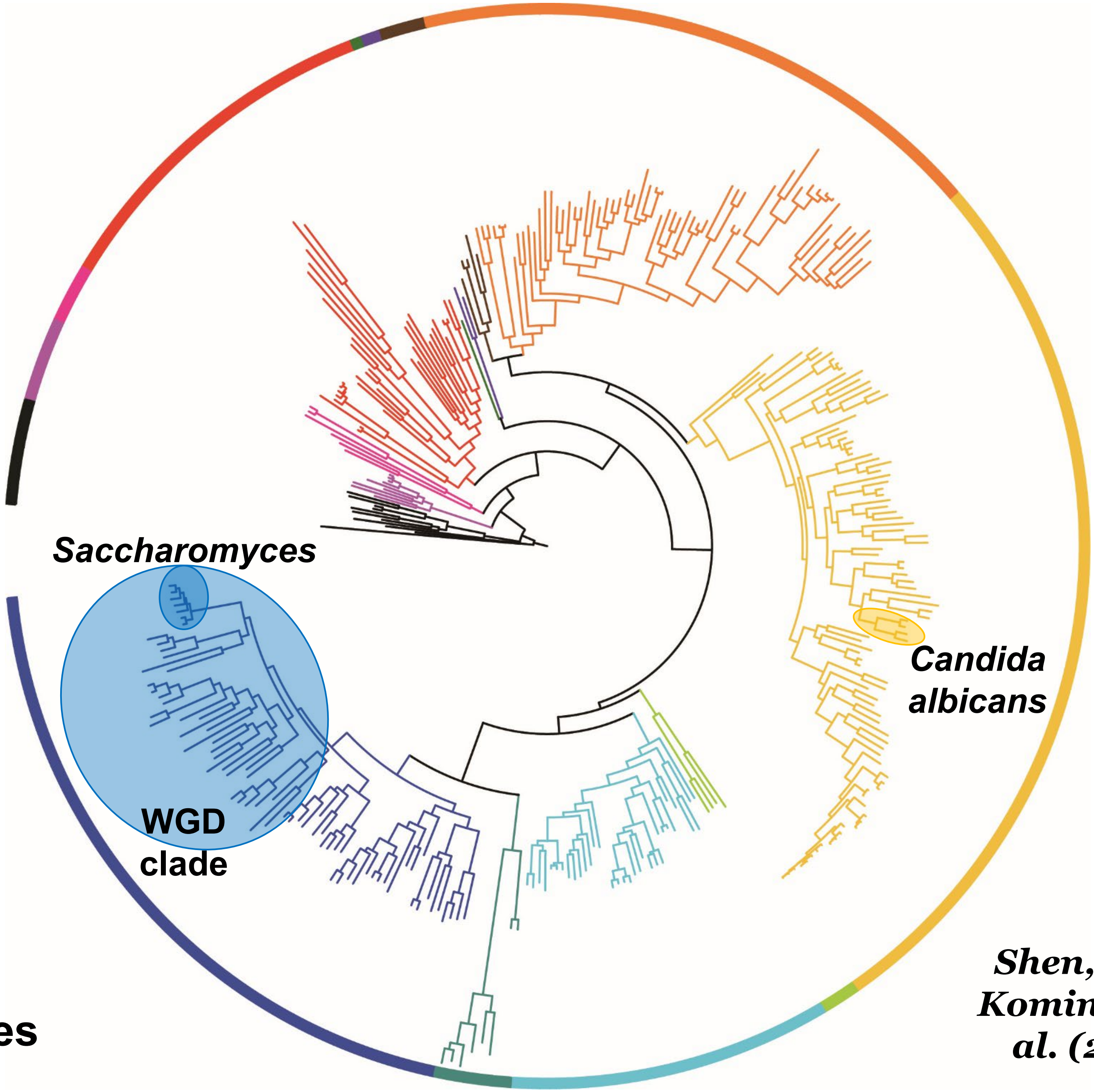
Shen et al. (2017) Nature Ecol. Evol.

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)

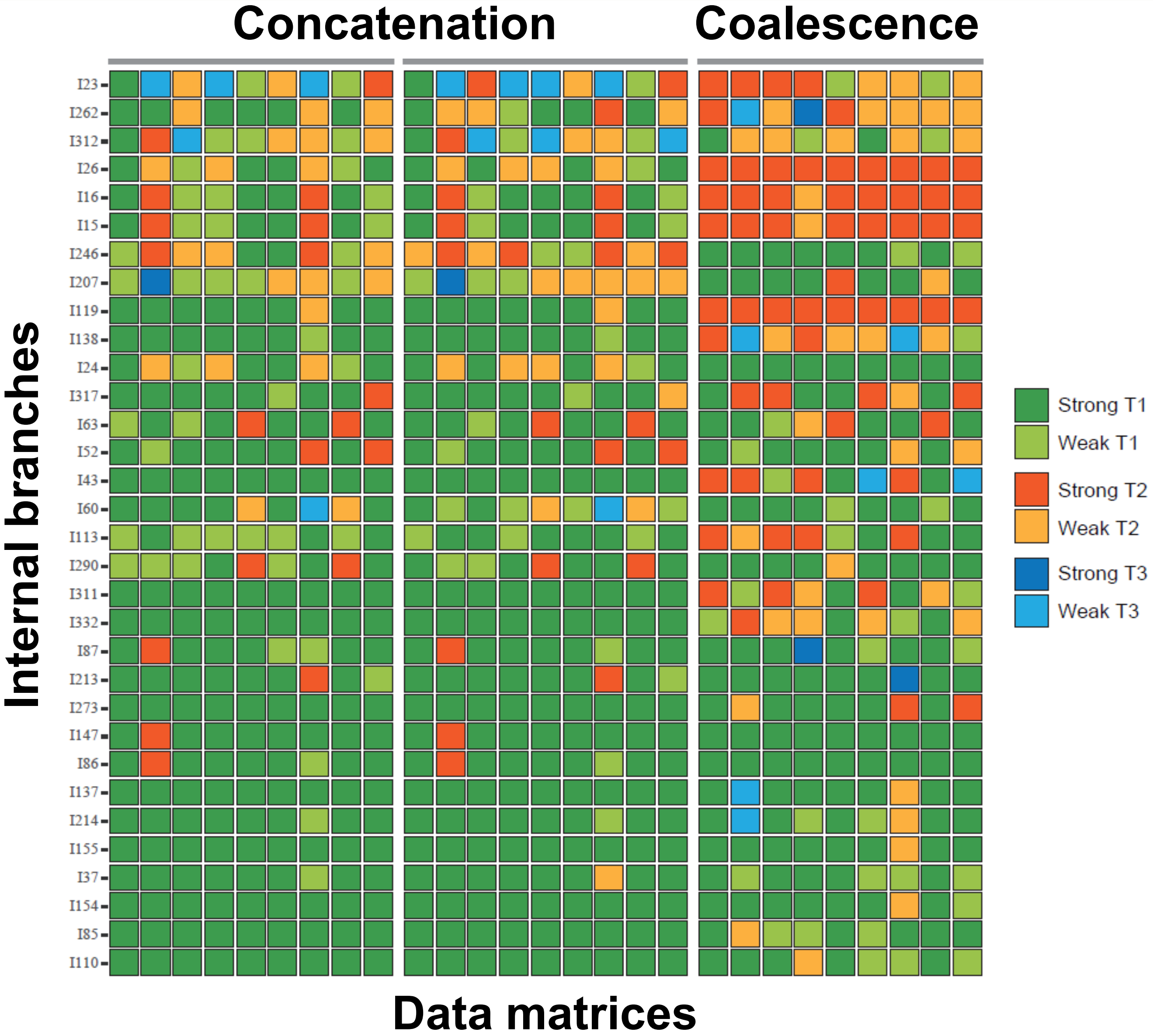
Genome-scale phylogeny of 332 yeasts



332 taxa
2,408 genes

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

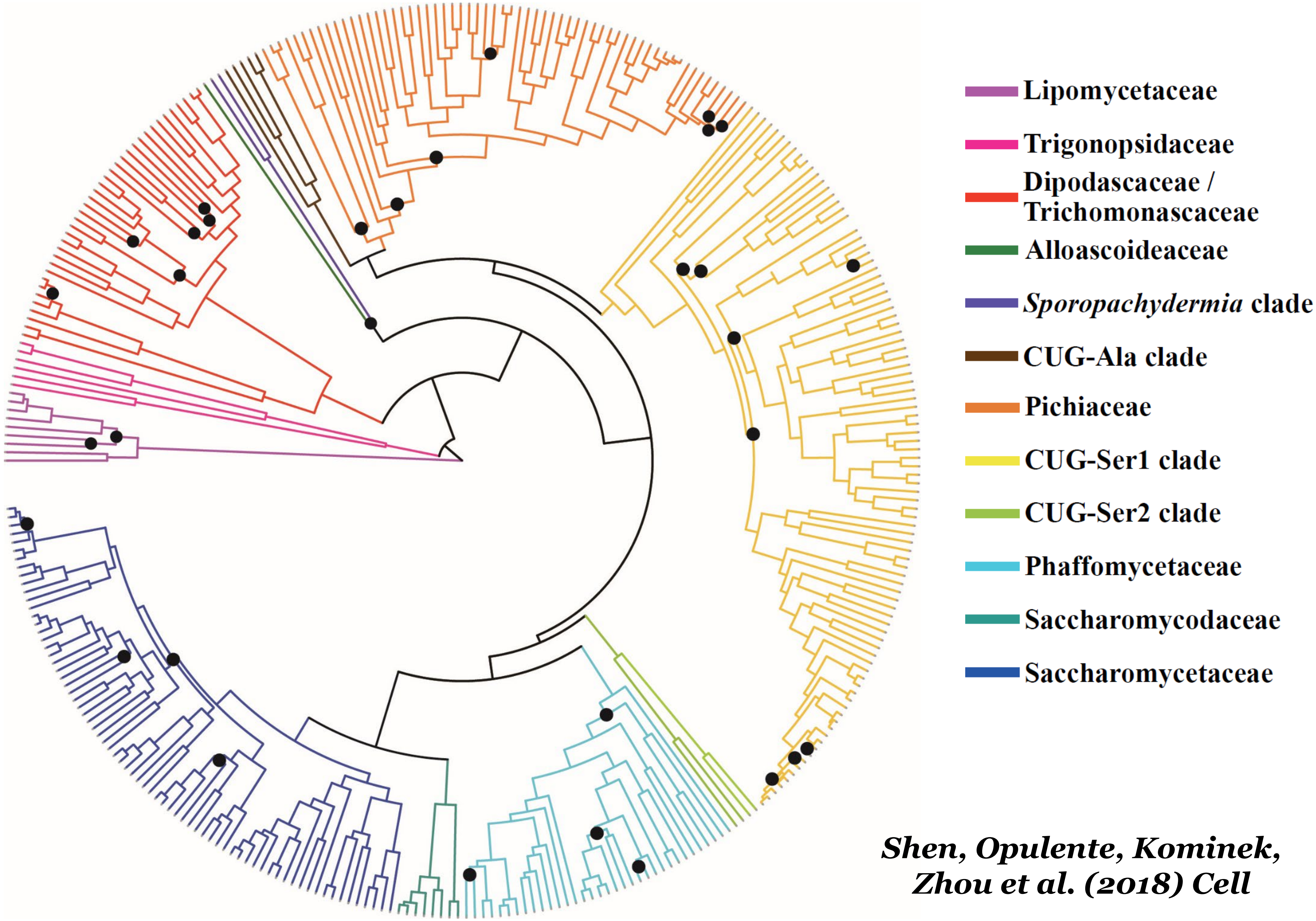
The 32 conflicting branches in the yeast phylogeny



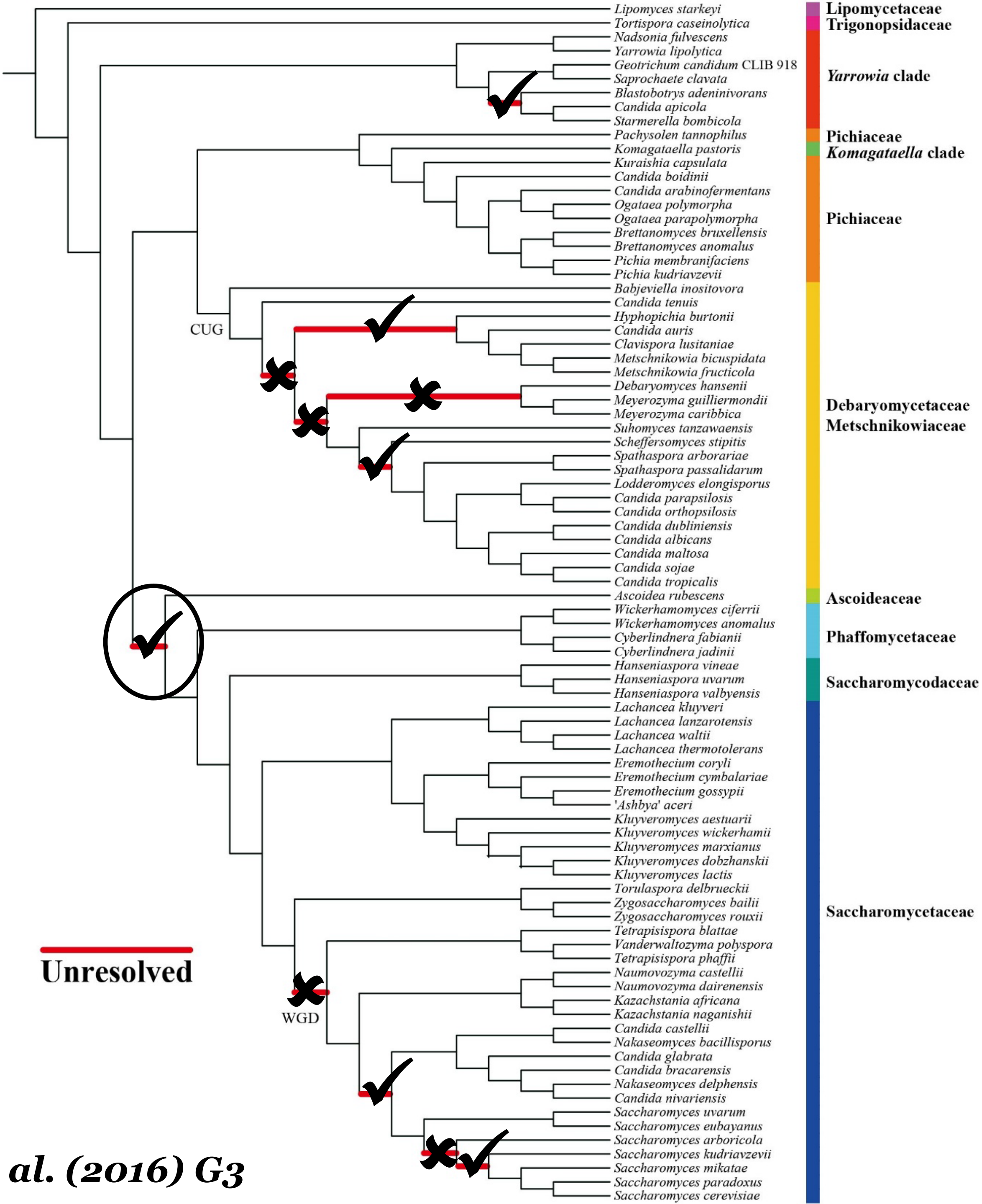
~10% (32 / 331) of internal branches show conflict between analyses

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

Distribution of conflict on the yeast phylogeny



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



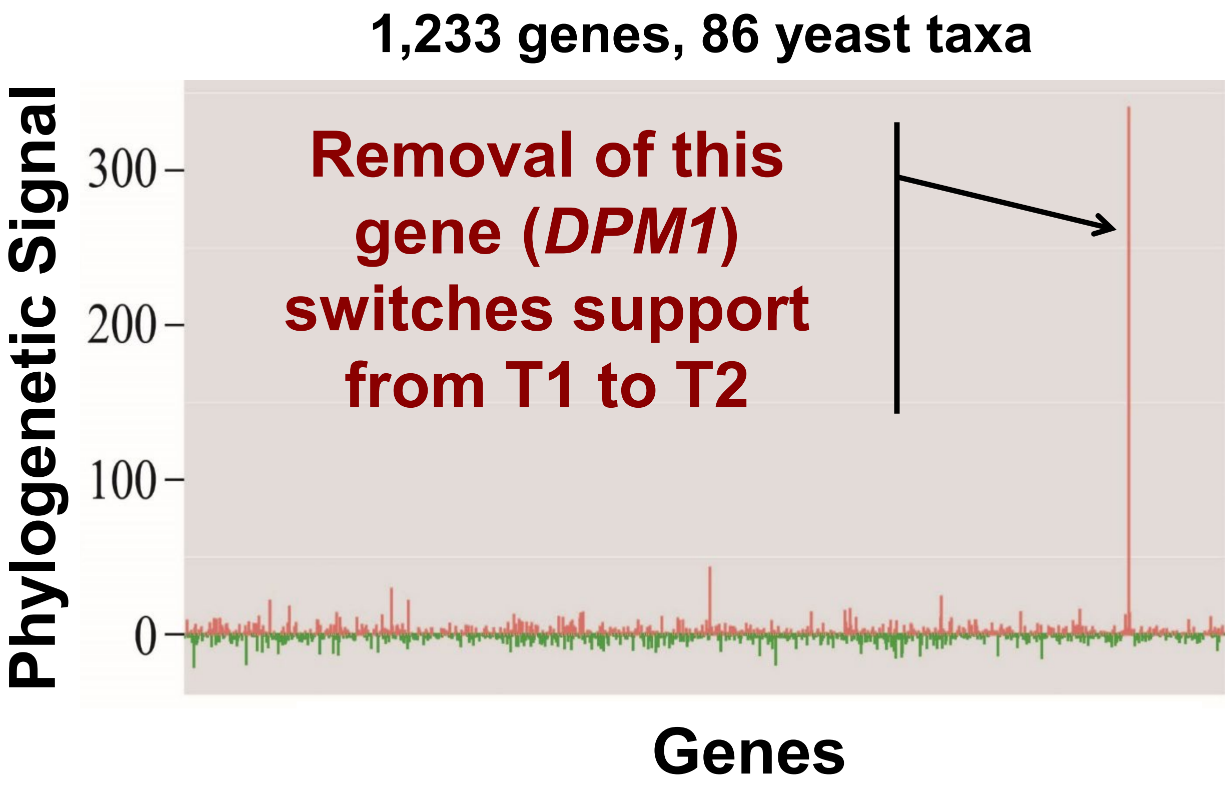
1,233-gene, 86-taxon data matrix

~13% (11 / 85) of internal branches conflict between analyses

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

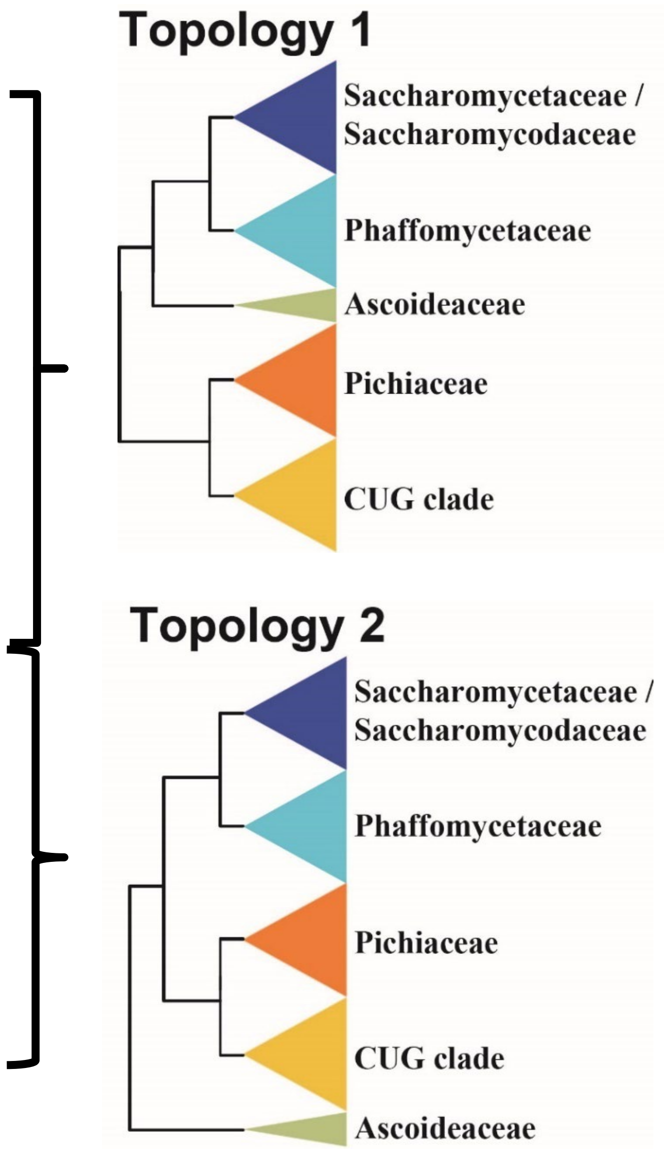
Shen et al. (2016) G3

A single gene governs the placement of Ascoideaceae



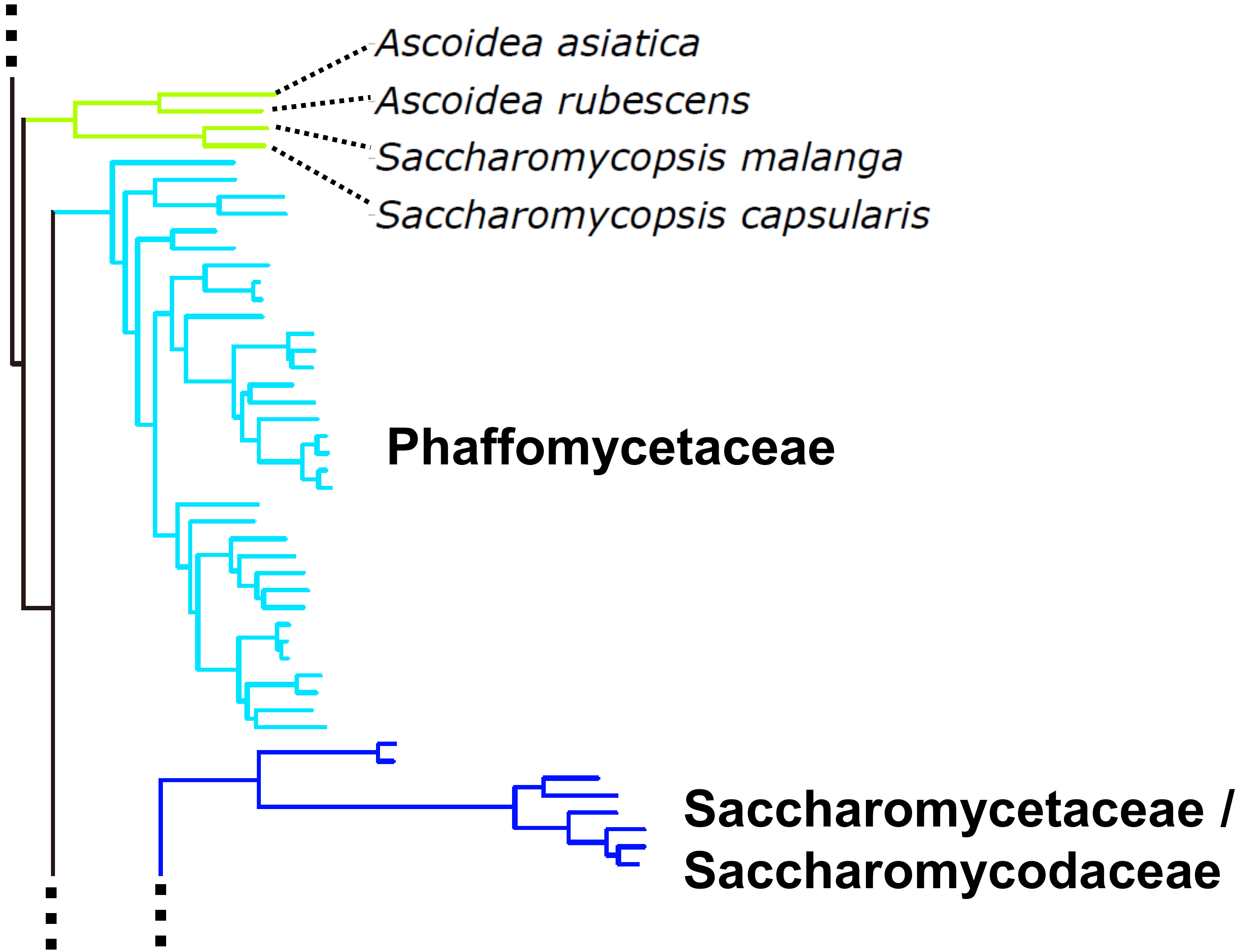
1,233 genes, 86 yeast taxa

Removal of this gene (*DPM1*) switches support from T1 to T2



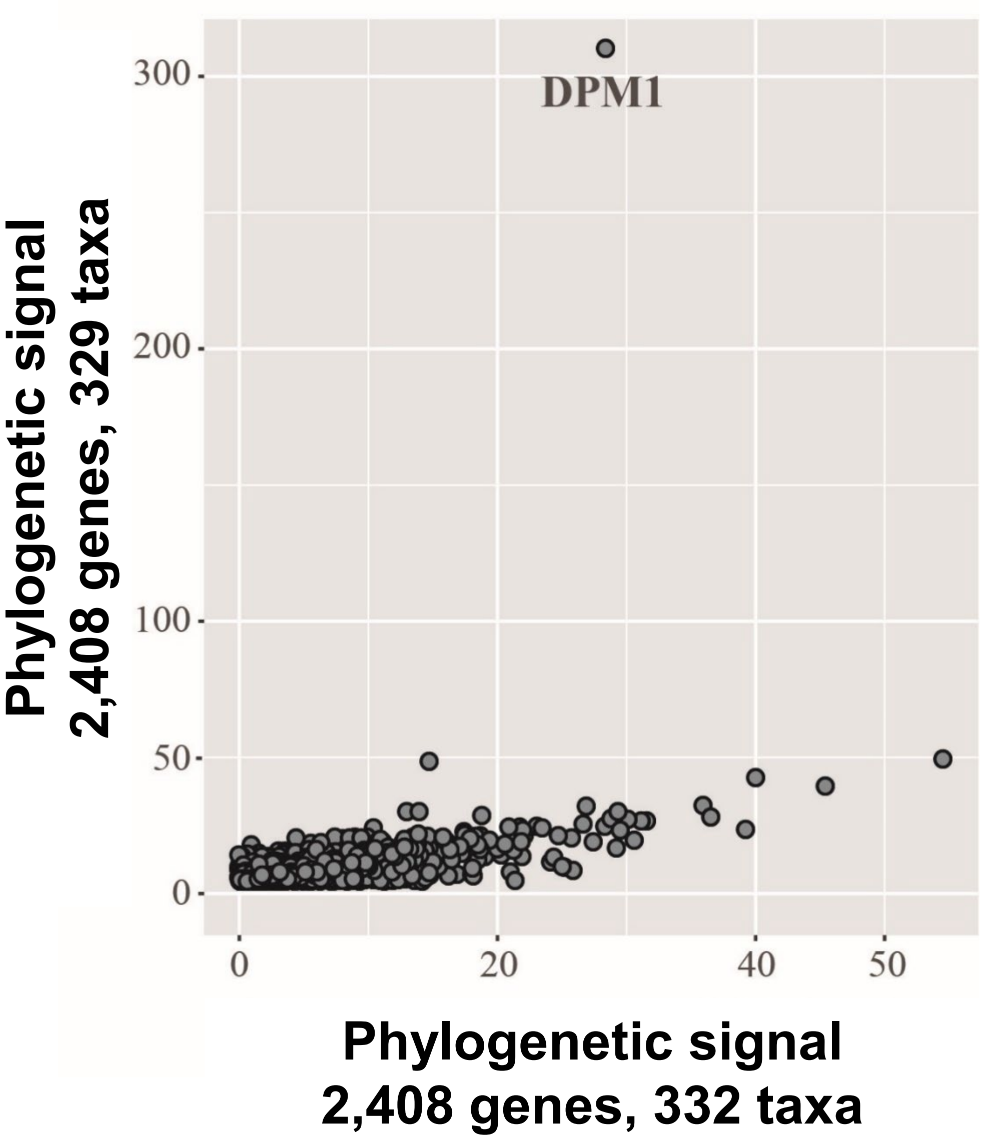
Shen et al. (2017) Nature Ecol. Evol.

Sampling of 3 additional taxa “breaks” the long branch


Ascoidea asiatica
Ascoidea rubescens
Saccharomycopsis malanga
Saccharomycopsis capsularis
Phaffomycetaceae
Saccharomycetaceae / Saccharomycodaceae

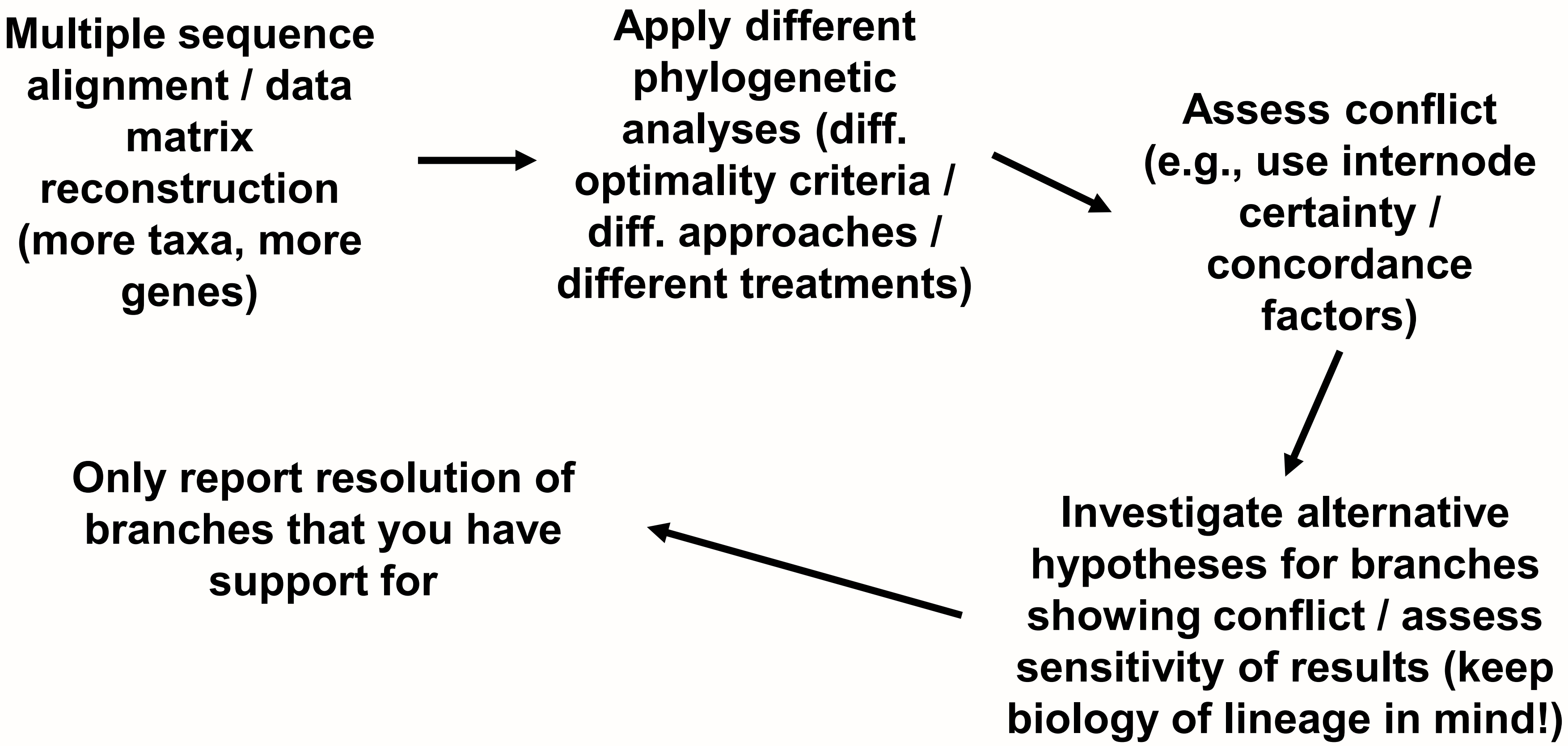
Shen, Opulente, Kominek, Zhou et al. (2018) Cell

Sampling of 3 additional taxa decreases gene's signal



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

Genomfart? The way forward?



Multiple sequence alignment / data matrix reconstruction (more taxa, more genes)

Apply different phylogenetic analyses (diff. optimality criteria / diff. approaches / different treatments)

Assess conflict (e.g., use internode certainty / concordance factors)

Investigate alternative hypotheses for branches showing conflict / assess sensitivity of results (keep biology of lineage in mind!)

Only report resolution of branches that you have support for

Carefully analyze and listen to the data

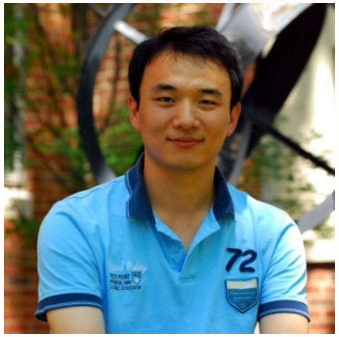
“One can use the most sophisticated audio equipment to listen, for an eternity, to a recording of white noise and still not glean a useful scrap of information”

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

Berkeley UNIVERSITY OF CALIFORNIA

HITS

Alexis Stamatakis

National Science Foundation

WHERE DISCOVERIES BEGIN

Chris Todd Hittinger

Marizeth Groenewald

National Institute of Allergy and Infectious Diseases

BURROUGHS WELLCOME FUND

WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

WESTERDIJK FUNGALBIO DIVERSITY INSTITUTE

http://www.rokaslab.org/ @RokasLab