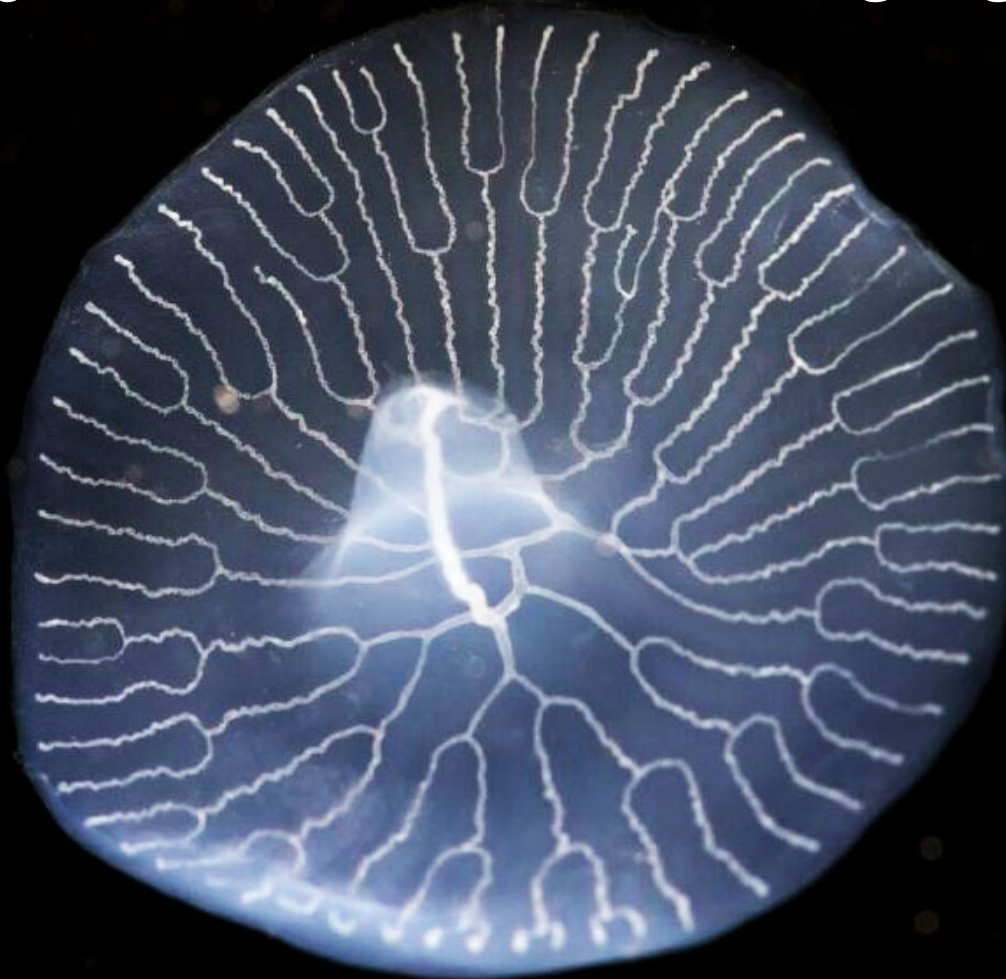


History of & Introduction to Phylogenomics



Antonis Rokas

Department of Biological Sciences, Vanderbilt University

<http://www.rokaslab.org>

@RokasLab

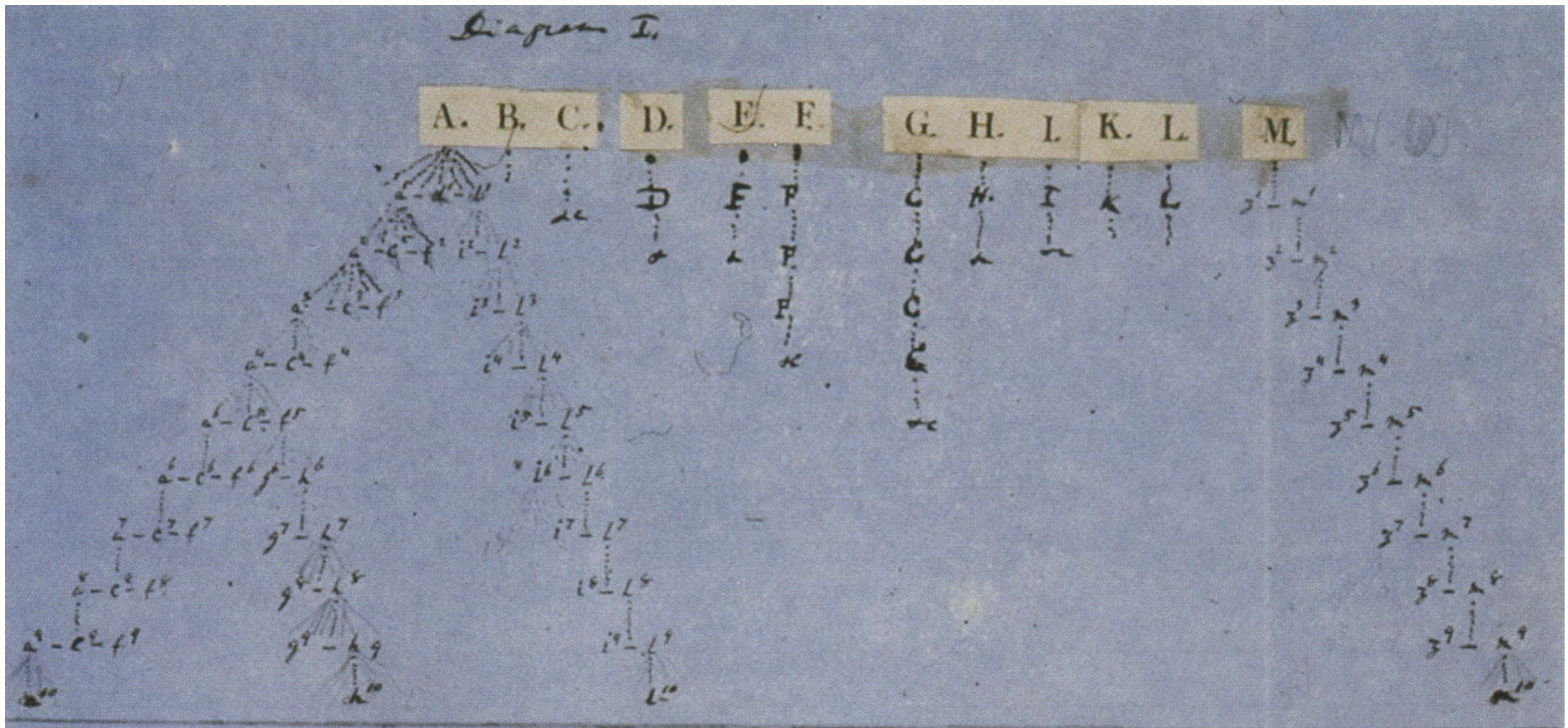
Lecture Outline

❖ From Darwin to Phylogenomics

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

❖ Prospects and Challenges of Phylogenomics

Darwin's Tree



“As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications”



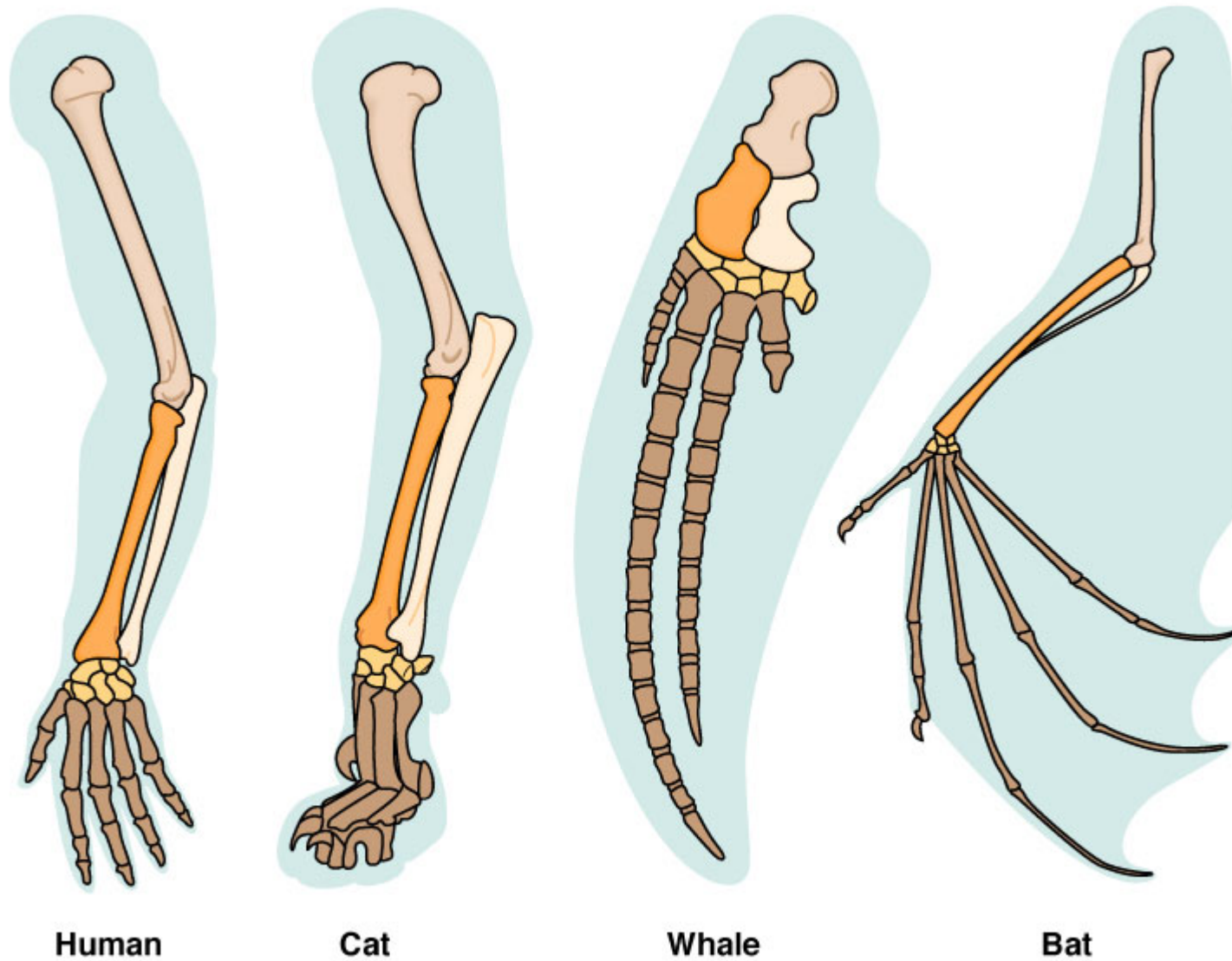
*Darwin's hand-made proof of the famous diagram in his Origin of Species;
Maderspacher (2006) Curr. Biol.*

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-

Comparative Morphology of Extant Organisms



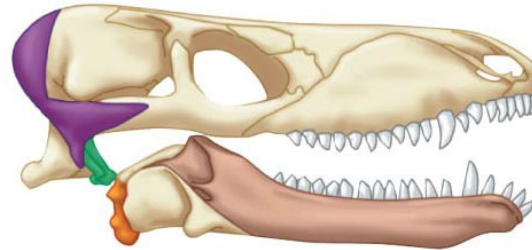
©1999 Addison Wesley Longman, Inc.



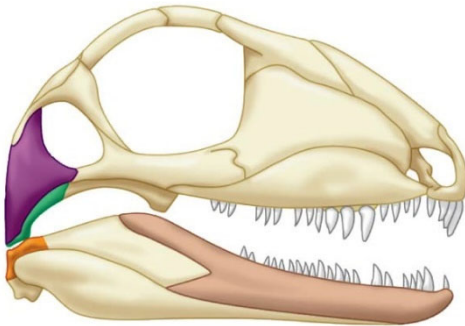
http://www.mun.ca/biology/scarr/139393_forelimb_homology.jpg

Comparative Anatomy of Fossils

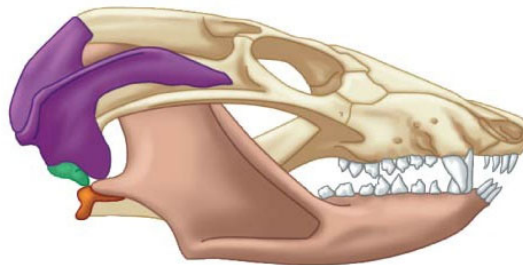
Early cynodont (260 mya)



Synapsid (300 mya)



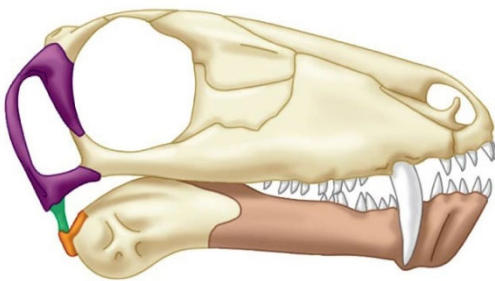
Later cynodont (220 mya)



Key to skull bones

- Articular
- Quadrate
- Dentary
- Squamosal

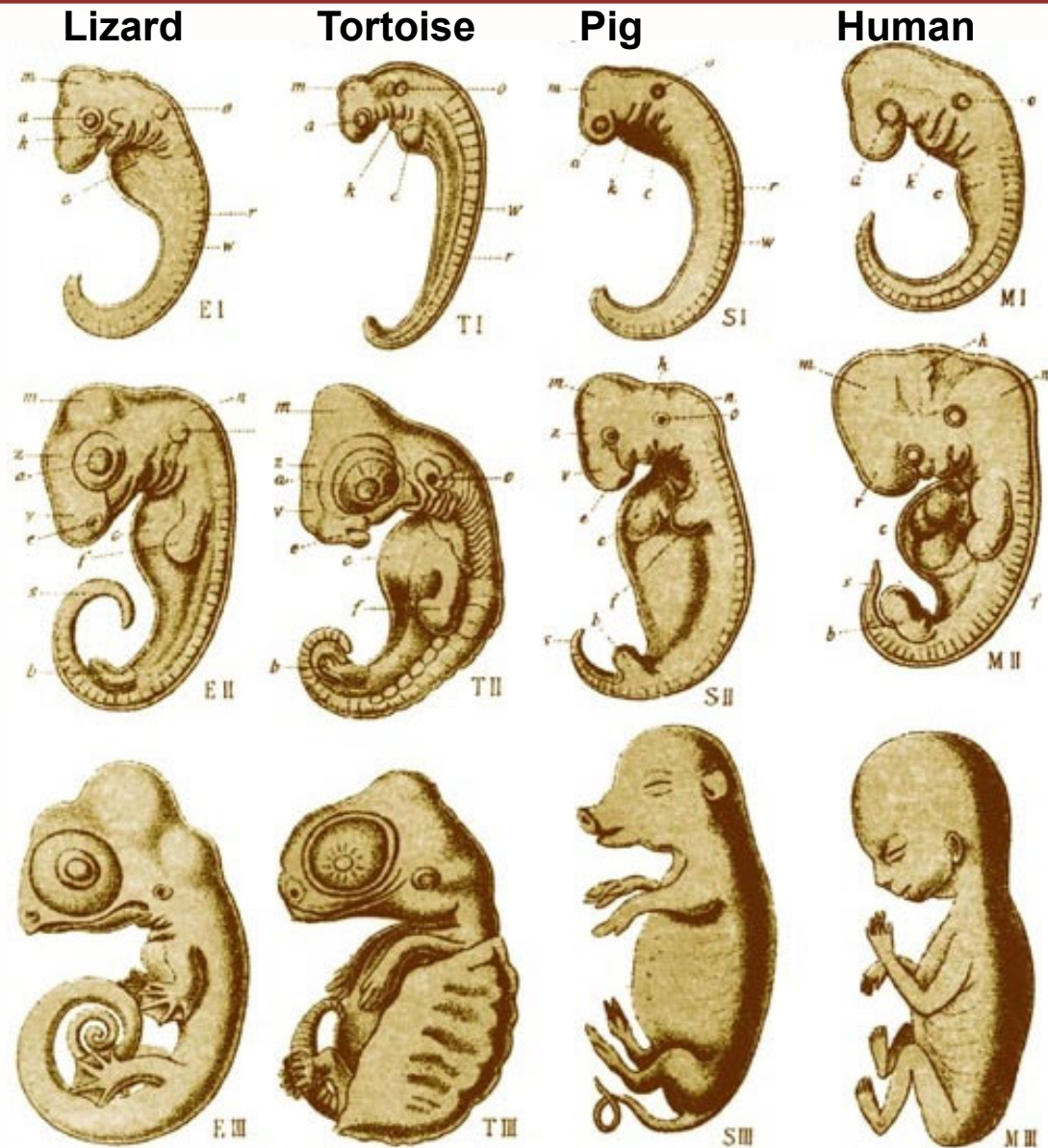
Therapsid (280 mya)



Very late cynodont (195 mya)



Comparative Embryology

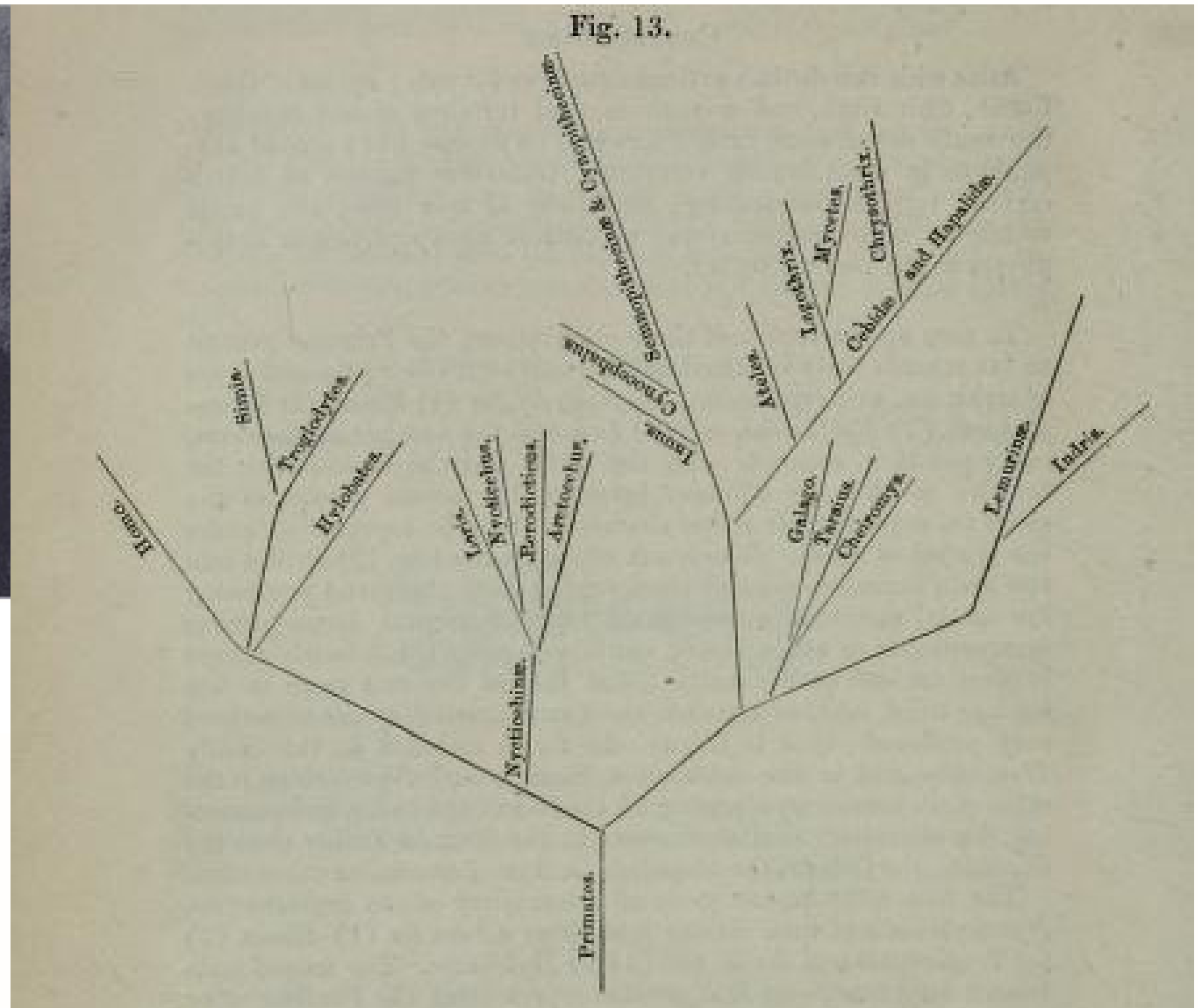


<http://www.nature.com/nrg/journal/v7/n11/images/nrg1918-f2.jpg>

The First Published Phylogeny



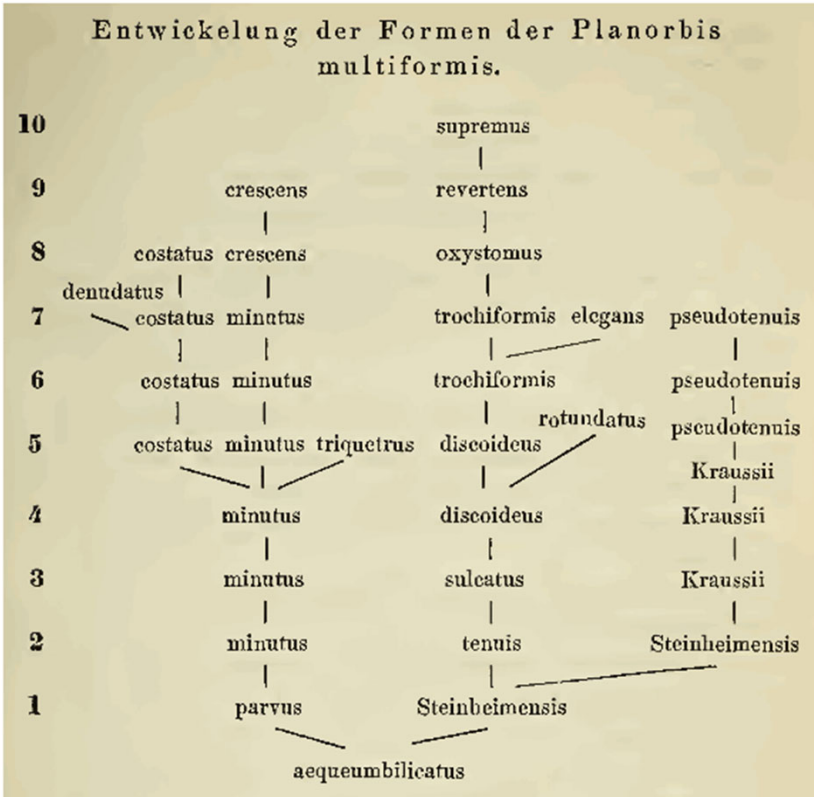
St. George Jackson Mivart



Mivart (1865) Proc. Zool. Soc. London

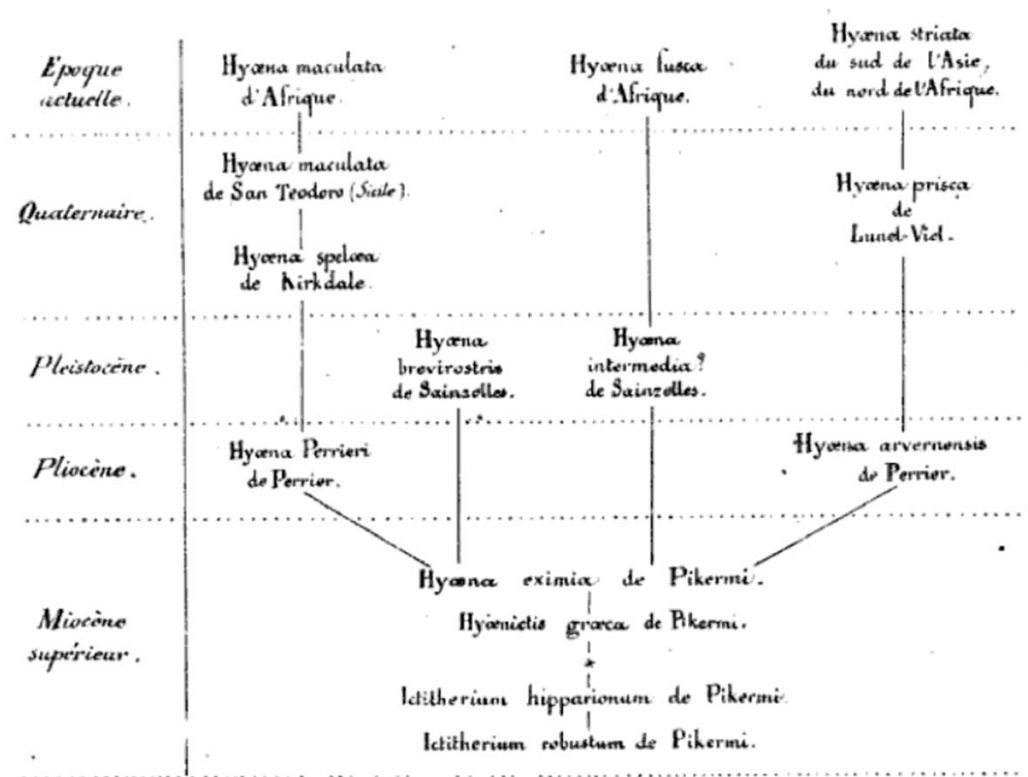
Inferring Phylogenies Becomes a Cottage Industry

Fossil gastropods



Hilgendorf, 1867

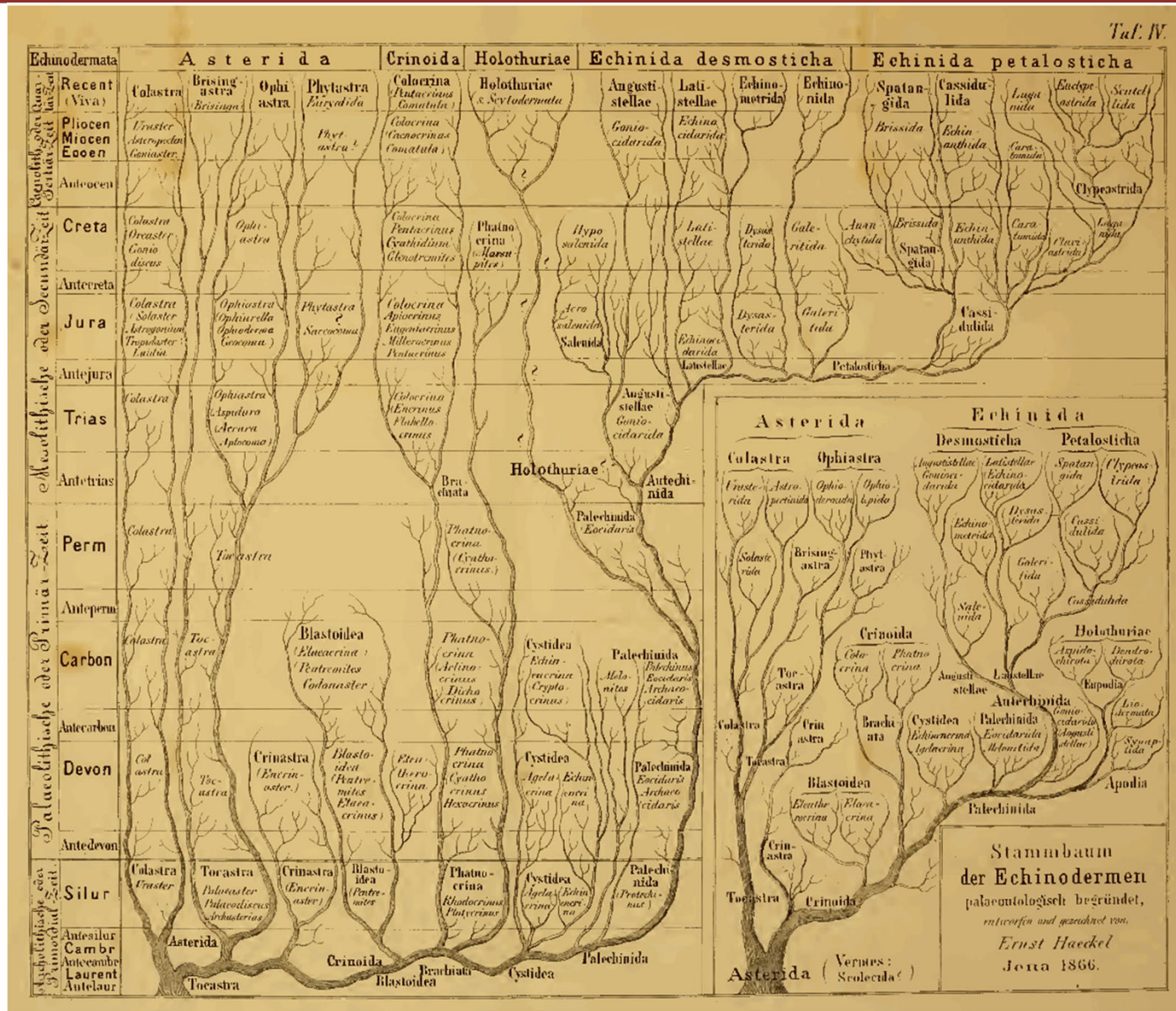
Extant and extinct mammals



Gaudry, 1866



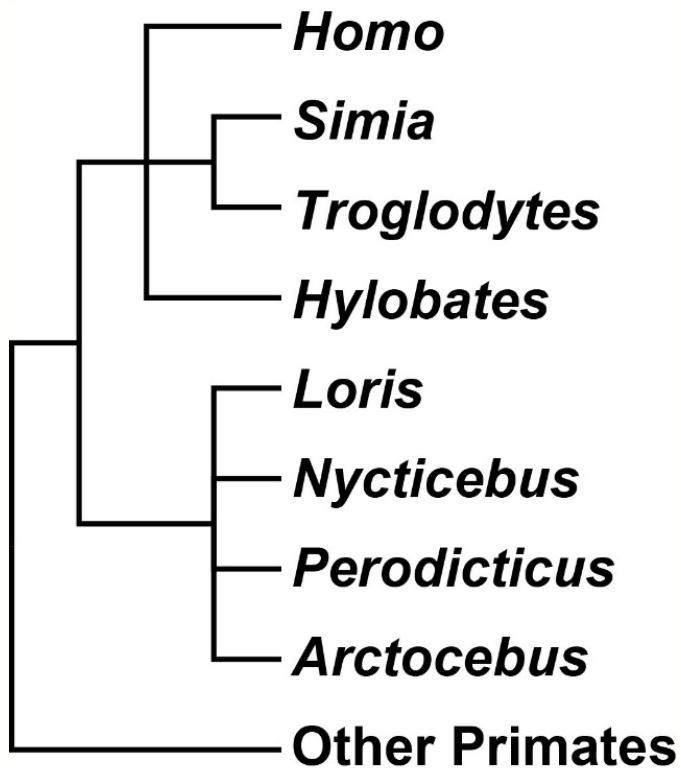
Haeckel's Phylogenies



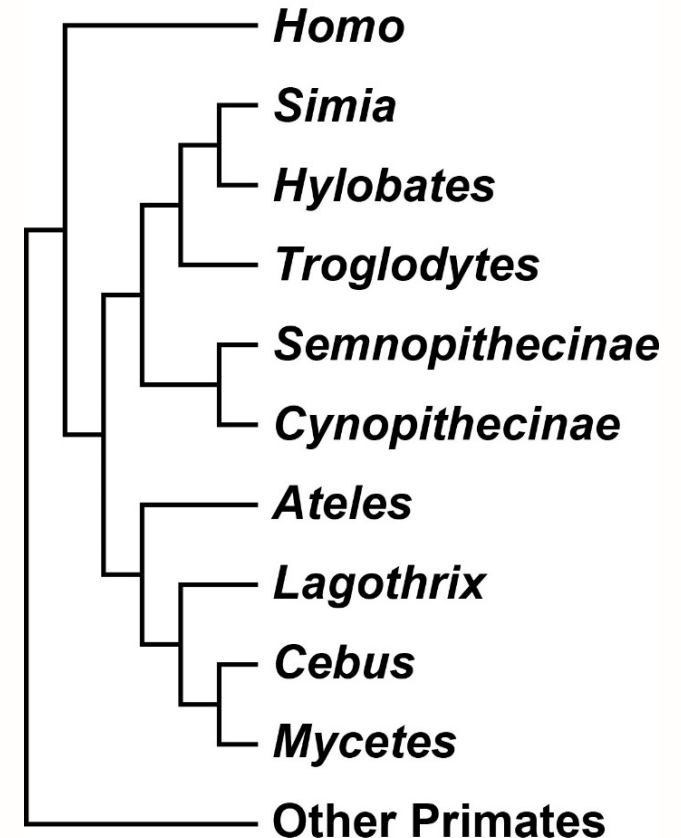
Haeckel (1866)

Disagreement Between Phylogenies

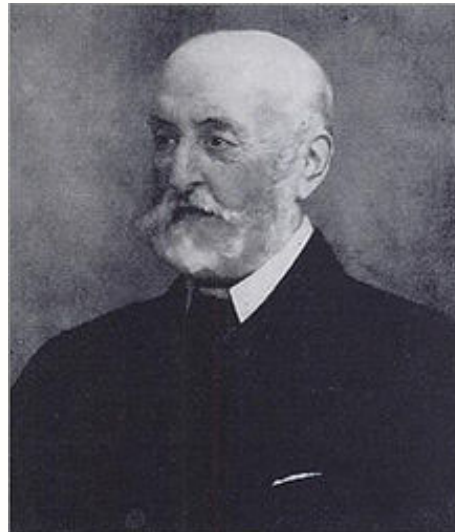
1865: SPINAL COLUMN



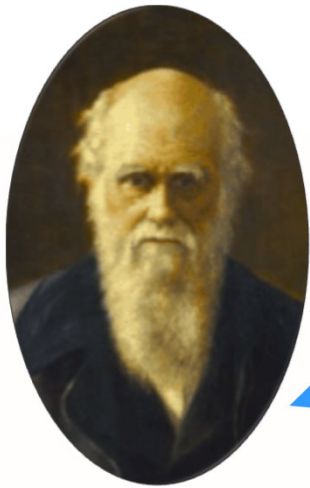
1867: LIMBS



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*



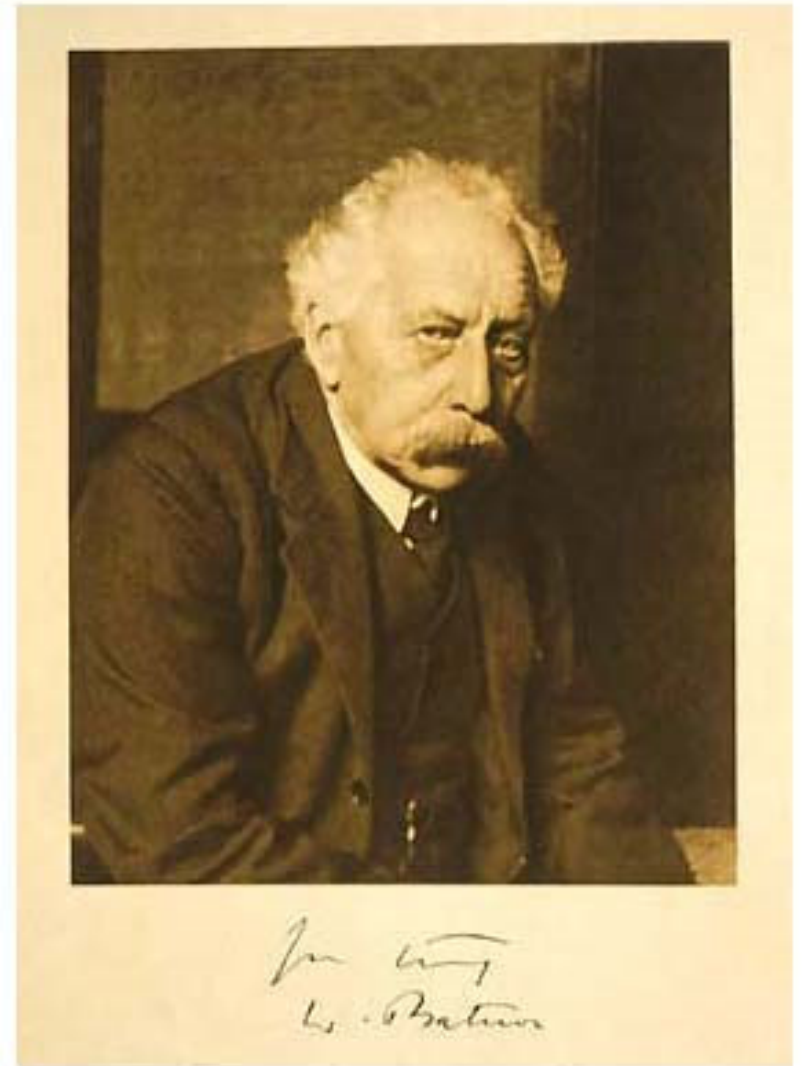
Comparative Morphology & Embryology in Trouble

By the turn of the century, the conflicting phylogenies produced by comparative morphology and embryology data have given rise to a sense of despair among the community

“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”

William Bateson (1894)

Materials for the Study of Variation



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



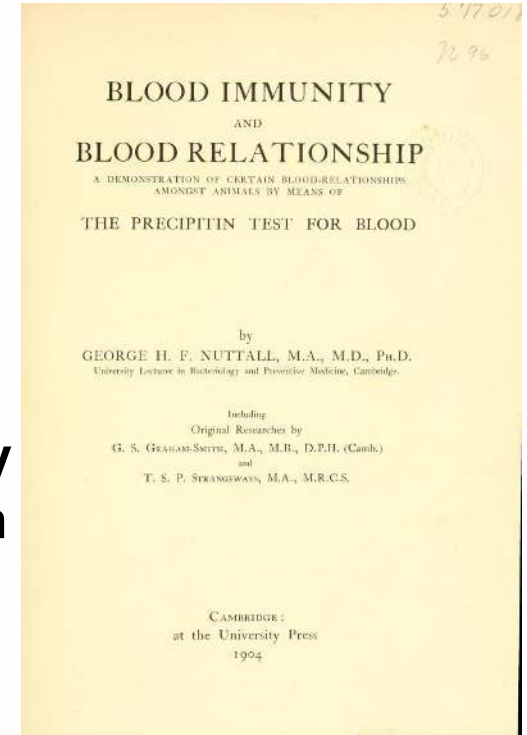
<https://www.dnalc.org/view/16197-Gallery-5-William-Bateson-Portrait.html>

The Origins of “Molecular” Phylogenetics



Studies in immunochemistry were showing that serological cross-reactions were stronger for more closely related organisms

Nuttall, realizing the evolutionary implications, used this approach to reconstruct the phylogenetic relationships among various groups of animals



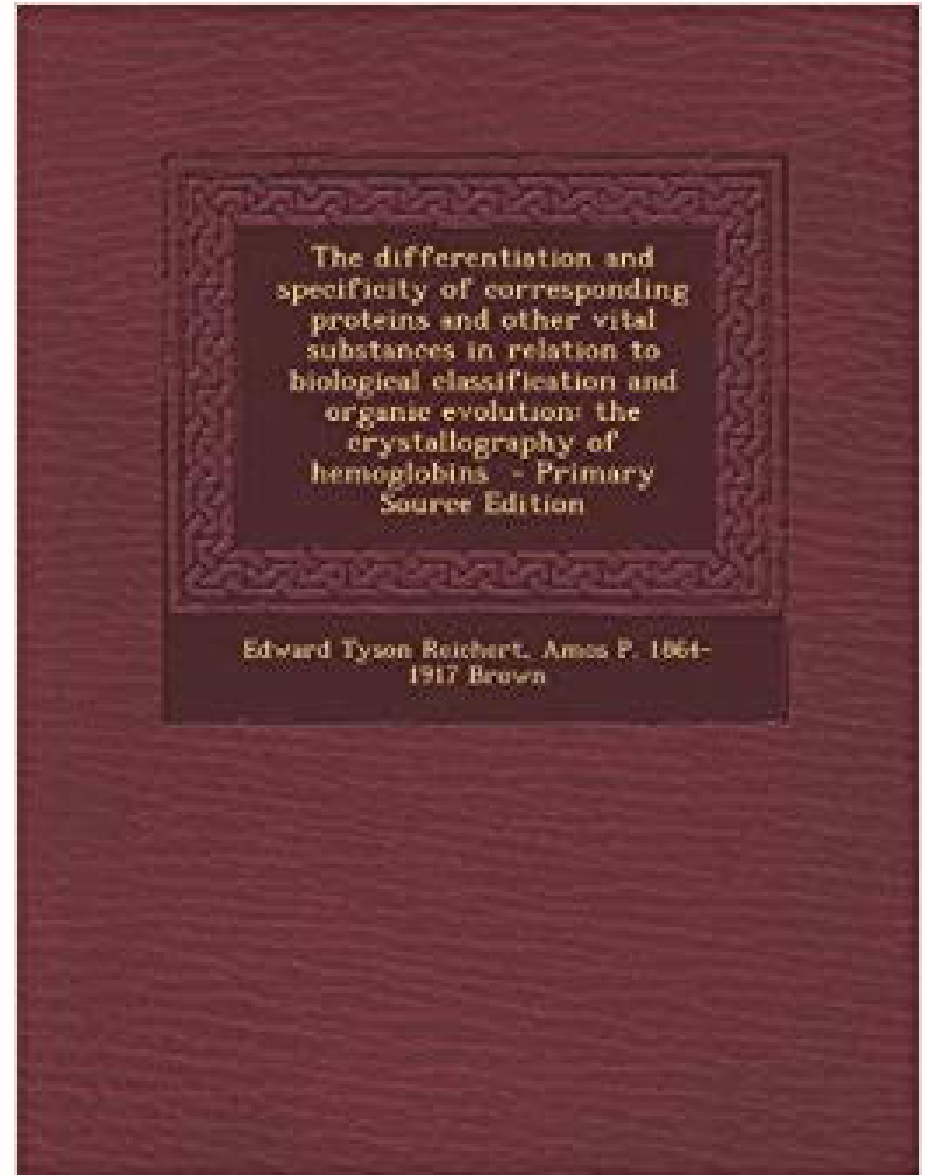
He determined that humans' closest relatives were the apes, followed, in order of relatedness, by Old World monkeys, New World monkeys, and prosimians (lemurs and tarsiers)!



Nuttall (1904) Blood Immunity and Blood Relationship

The First Large-Scale Molecular Investigation of Species Differences

- ❖ Different crystals of hemoglobin from the same species differed in size & shape, but angles between faces constant
- ❖ But interfacial angles differed from species to species -> similarities in angle values were consistent with taxonomy-based phylogeny
- ❖ 600 photomicrographs of crystals of hemoglobin from >100 species
- ❖ All this before discoveries of X-ray diffraction / protein sequencing / DNA & DNA sequencing



Reichert & Brown (1909) The Crystallography of Hemoglobins

“Molecular” Phylogenetics of Drosophila



Theodosius Dobzhansky

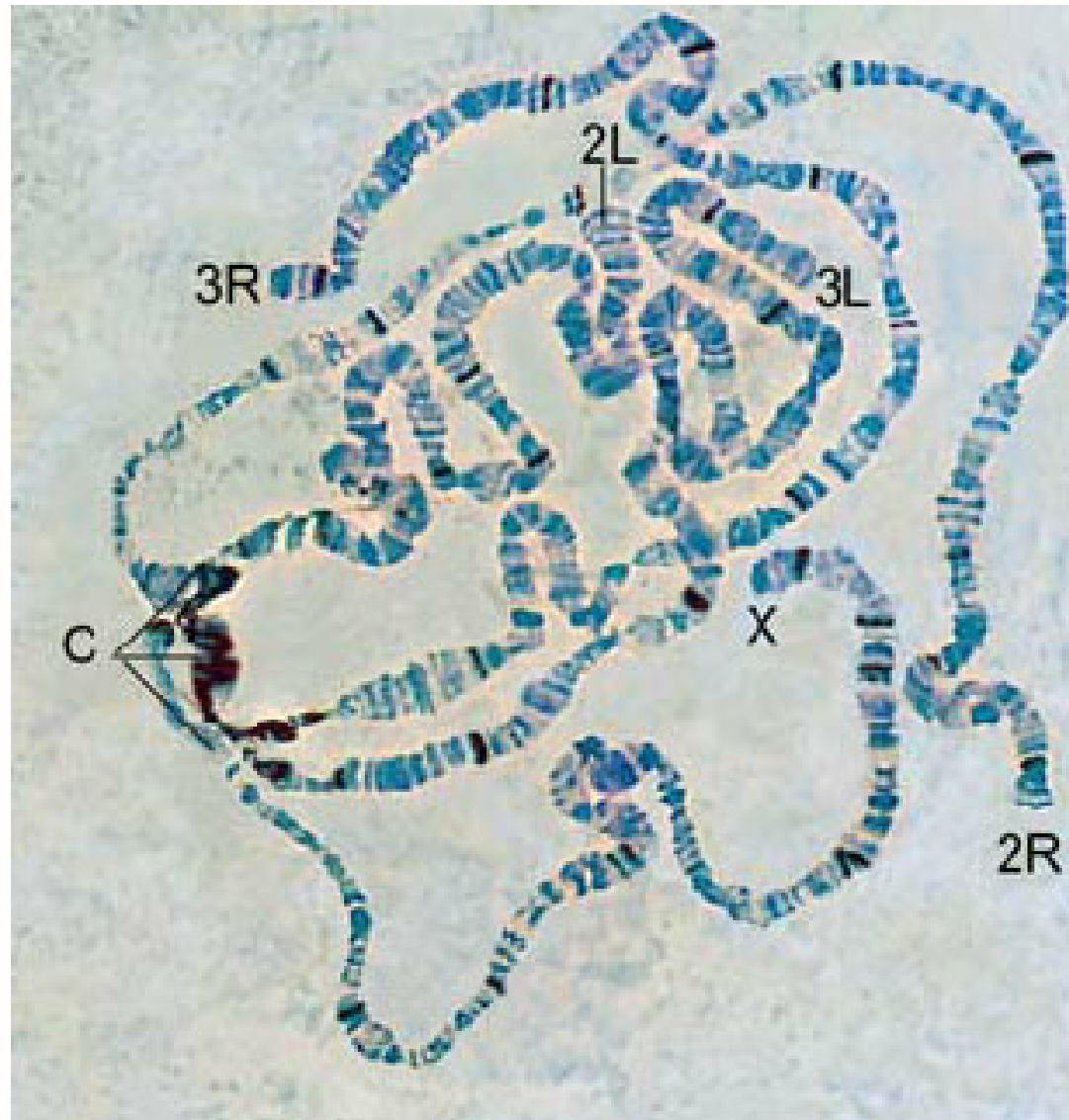


Alfred Sturtevant



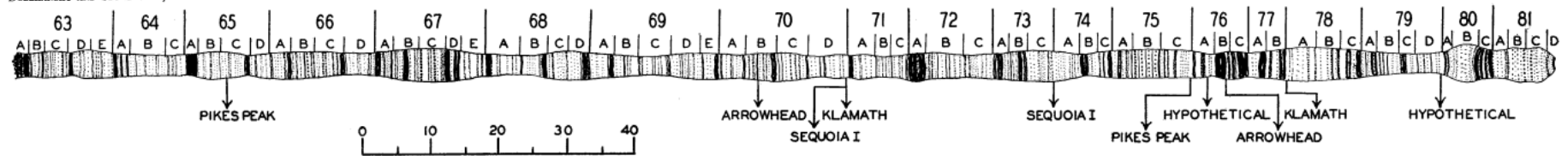
https://www-tc.pbs.org/wgbh/evolution/library/06/2/images/l_062_04_l.jpg;
<http://www.caltech.edu/news/first-genetic-linkage-map-38798>

Polytene Chromosomes

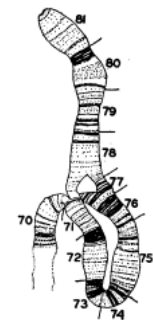


Using Chromosomal Rearrangements as Markers...

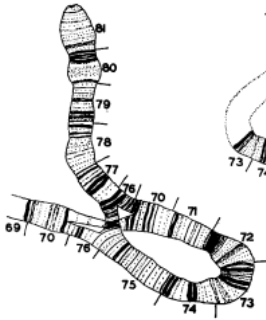
DOBZHANSKY AND STURTEVANT, CHROMOSOMES OF *DROSOPHILA PSEUDOORSURA*



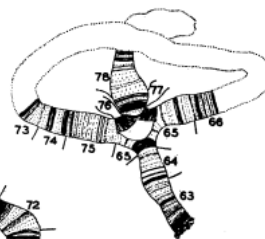
KLAMATH
STANDARD



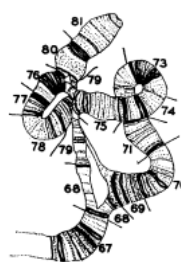
ARROWHEAD
STANDARD



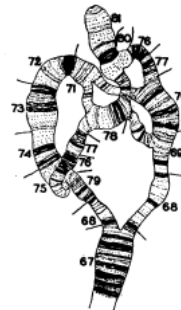
PIKES PEAK
STANDARD



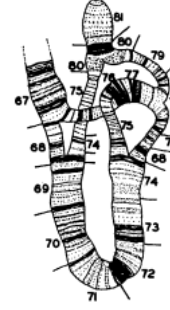
SANTA CRUZ
STANDARD



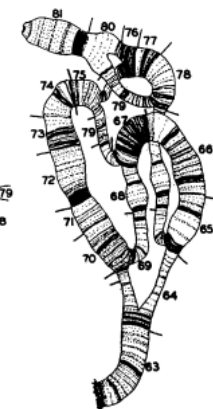
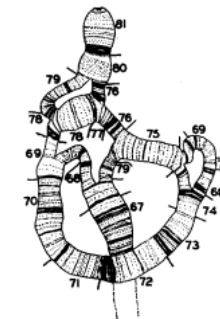
CHIRICAHUA
STANDARD



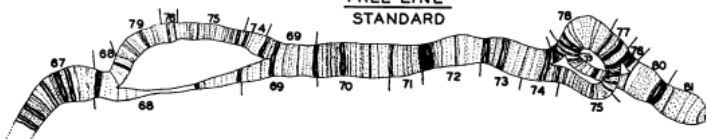
OLYMPIC
STANDARD



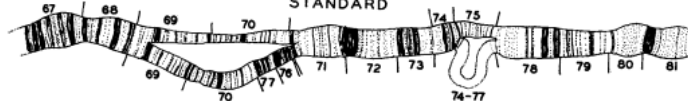
ESTES PARK
STANDARD



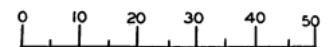
TREE LINE
STANDARD



WAWONA
STANDARD



CUERNAVACA
STANDARD



GENETICS 23: 28 Jan. 1938



Dobzhansky & Sturtevant (1938) Genetics

...To Infer the History of Species

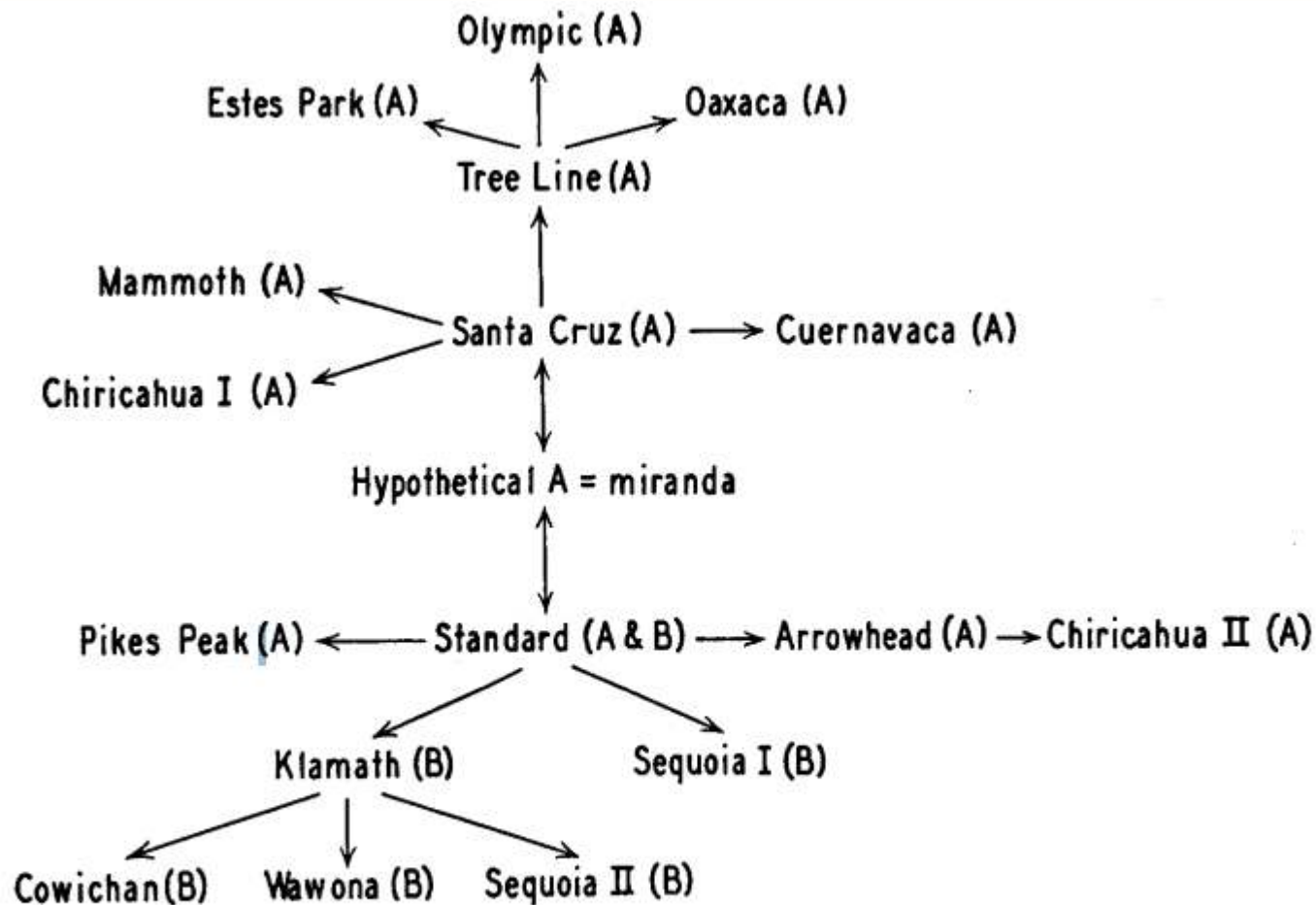
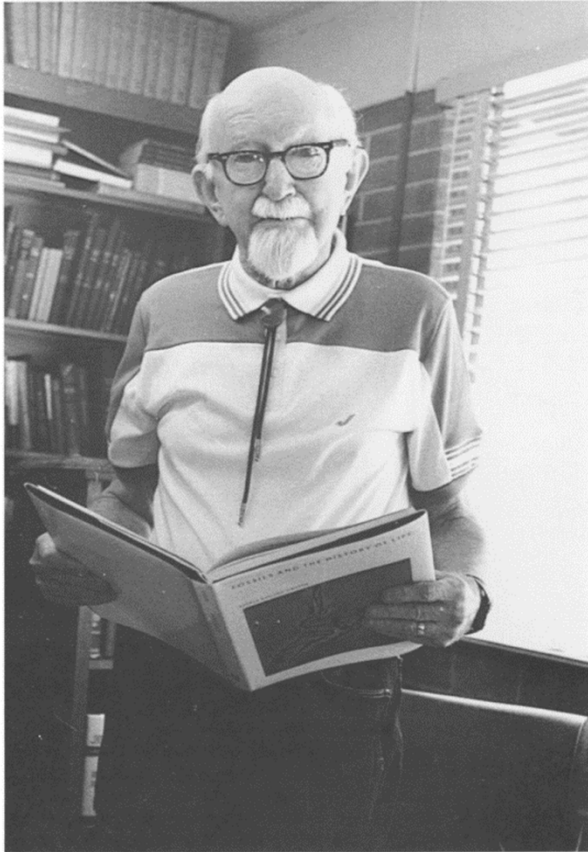
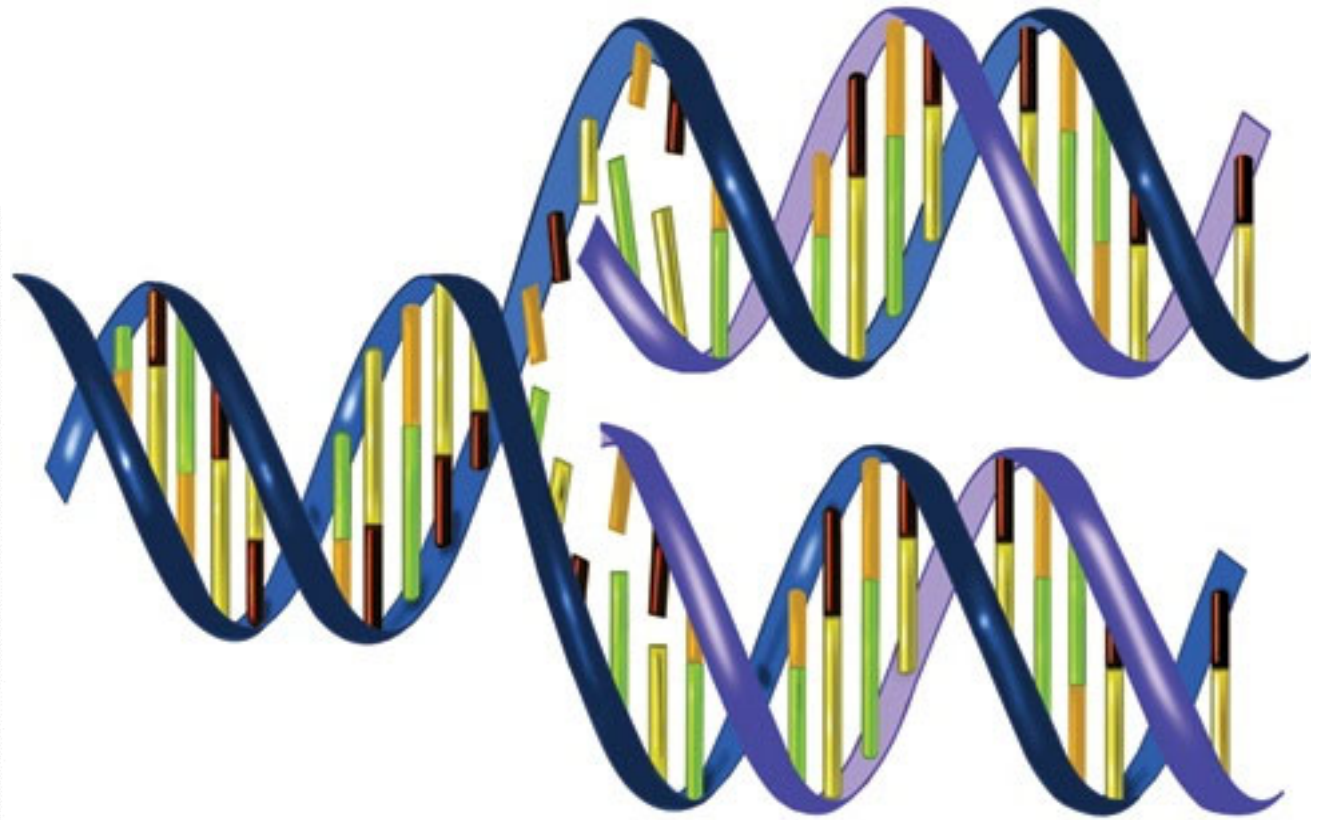


FIGURE 3.—Phylogeny of the gene arrangements in the third chromosome of *Drosophila pseudoobscura*. Any two arrangements connected by an arrow in the diagram differ by a single inversion. Further explanation in text.





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

Early '50s: Discovery of Protein Sequencing

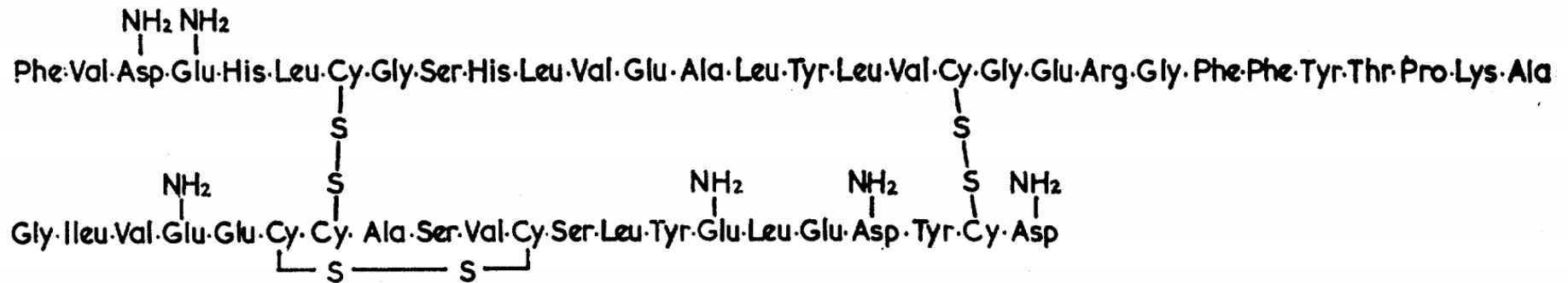


Fig. 2. Structure of insulin.



Frederick Sanger

The Nobel Prize in Chemistry 1958

Born: 13 August 1918, Rendcombe, United Kingdom

Died: 19 November 2013, Cambridge, United Kingdom

Affiliation at the time of the award: University of Cambridge, Cambridge, United Kingdom

Prize motivation: "for his work on the structure of proteins, especially that of insulin."

Prize share: 1/1



Sanger (1959) Science

“...before long we shall have a subject which might be called “protein taxonomy”; the study of amino acid sequences of the proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism and that vast amounts of evolutionary information may be hidden away within them”

Francis Crick (1957) Nature



Elucidating the Sequence of Proteins

NATURE

March 4, 1961 VOL. 189

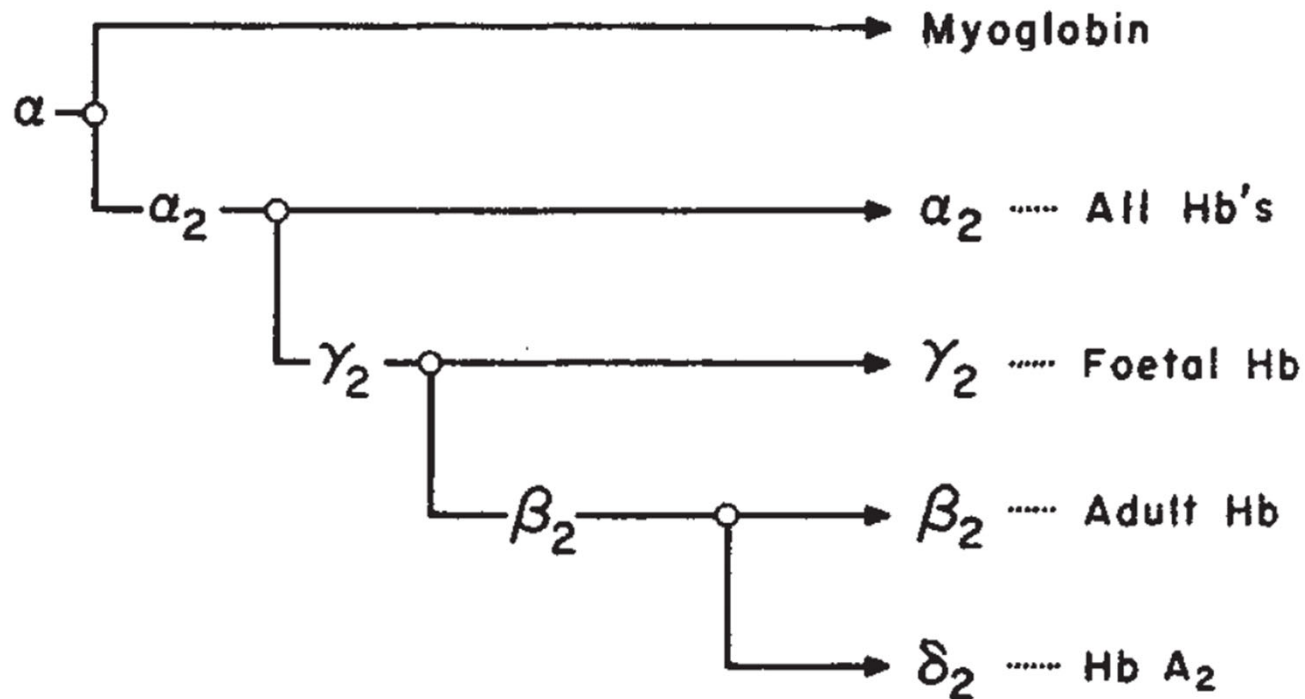
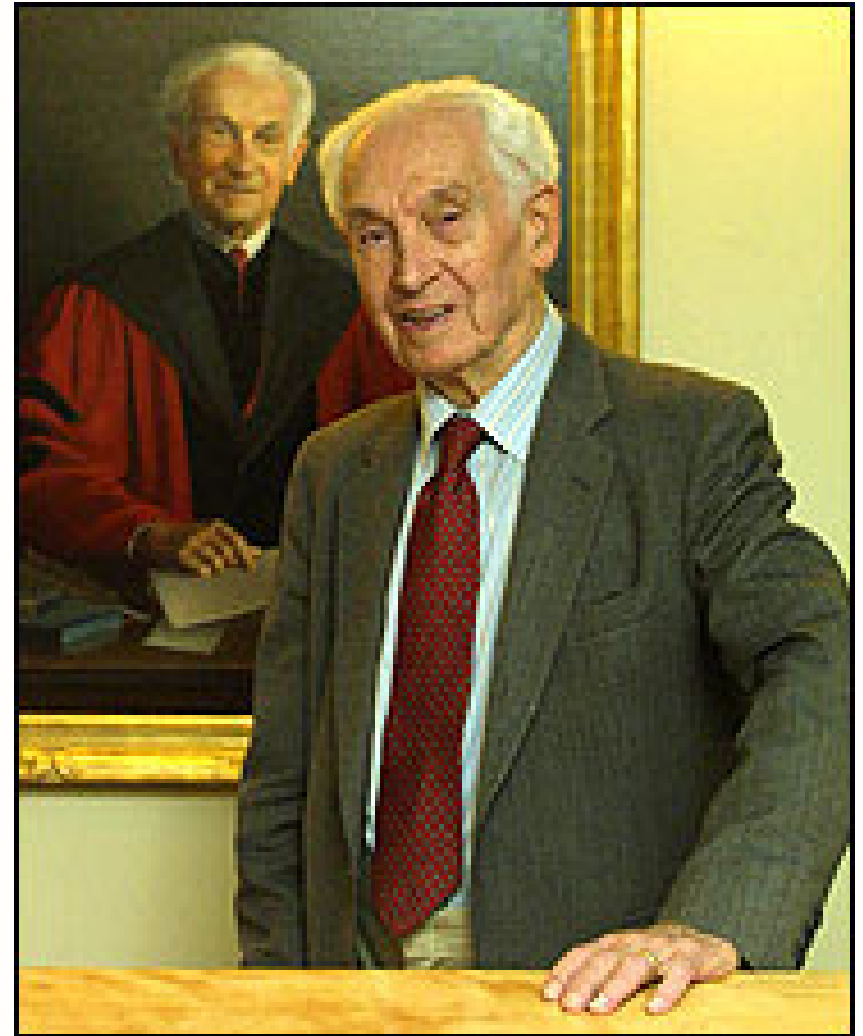


Fig. 1. Evolution of the hæmoglobin chains. The α -chain is the ancestral peptide chain. —○— indicates a point of gene duplication followed by translocation of the new gene



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

Ernst Mayr, 1963



DNA & Protein Sequences Record Evolutionary History

Molecules as Documents of Evolutionary History

EMILE ZUCKERKANDL AND LINUS PAULING

*Gates and Crellin Laboratories of Chemistry,
California Institute of Technology, Pasadena, California, U.S.A.*

(Received 17 September 1964)

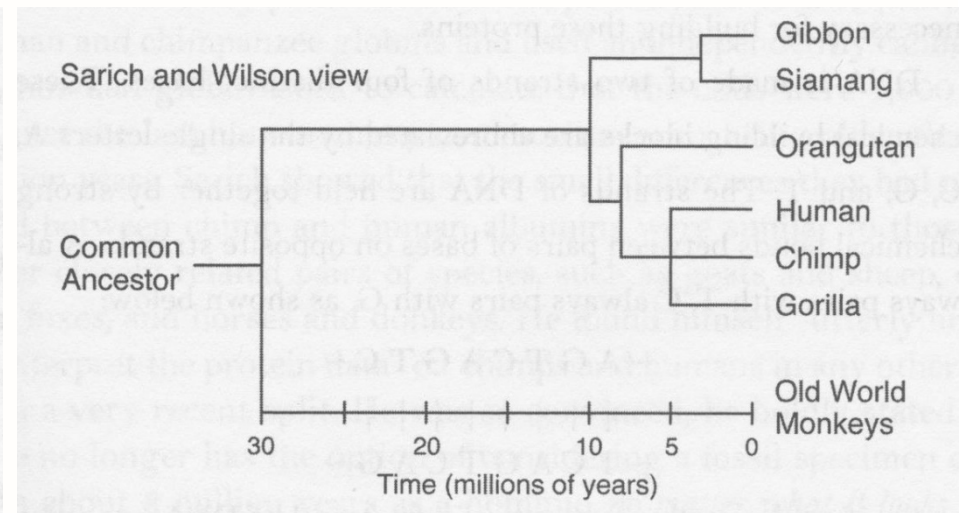
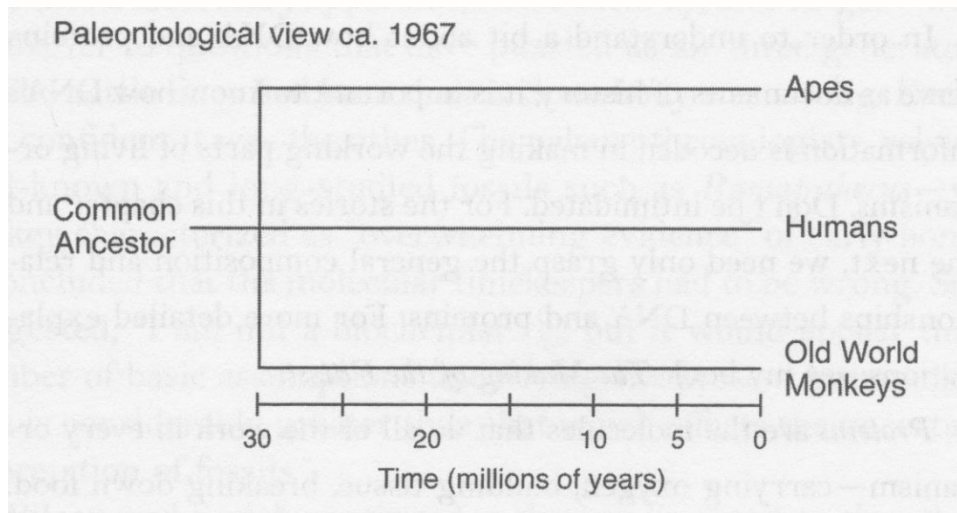
Different types of molecules are discussed in relation to their fitness for providing the basis for a molecular phylogeny. Best fit are the “semantides”, i.e. the different types of macromolecules that carry the genetic information or a very extensive translation thereof. The fact that more than one coding triplet may code for a given amino acid residue in a polypeptide leads to the notion of “isosemantic substitutions” in genic and messenger polynucleotides. Such substitutions lead to differences in nucleotide sequence that are not expressed by differences in amino acid sequence. Some possible consequences of isosemanticism are discussed.



Zuckerkandl & Pauling (1965) J. Theoret. Biol.

Estimating the Divergence of Humans and Chimps

Divergence times were estimated by measuring the immunological cross-reaction of blood serum albumin between pairs of primates



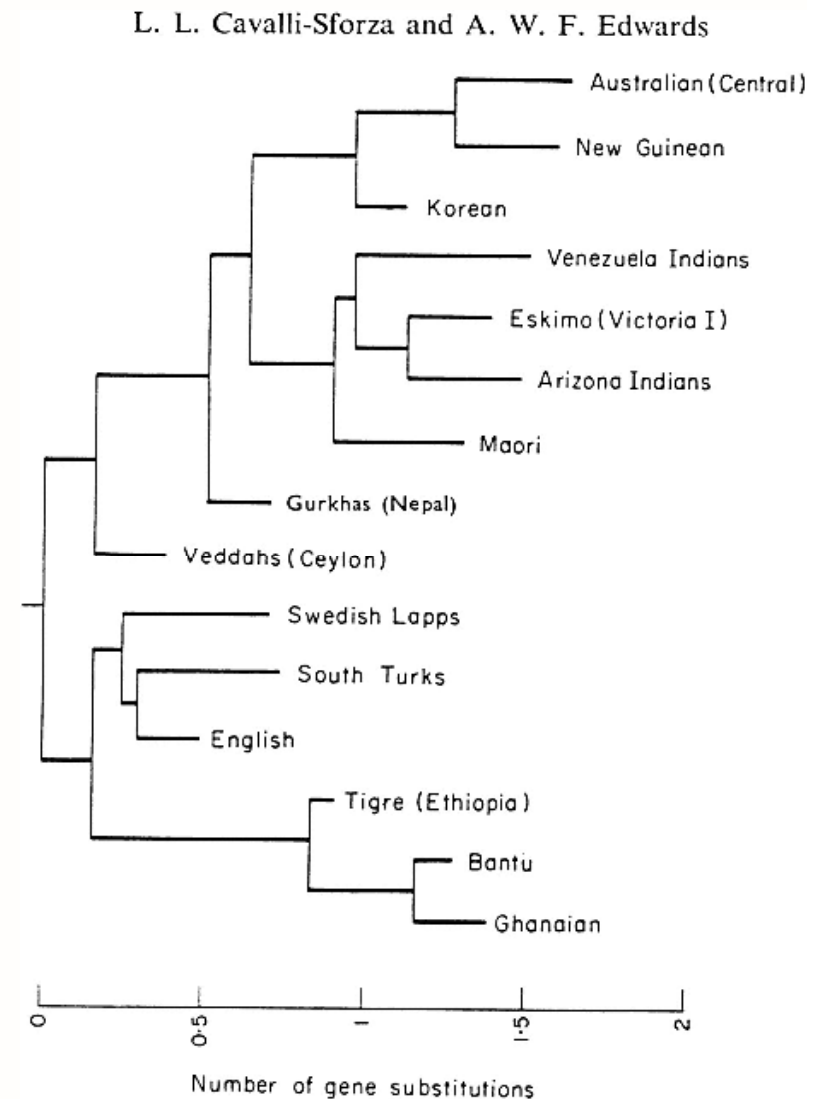
“no fuss, no muss, no dishpan hands. Just throw some proteins into a laboratory apparatus, shake them up, and bingo! – we have an answer to questions that have puzzled us for three generations.”



The Phylogeny of Human Populations



Phylogeny inferred from blood group allele frequencies from 15 populations

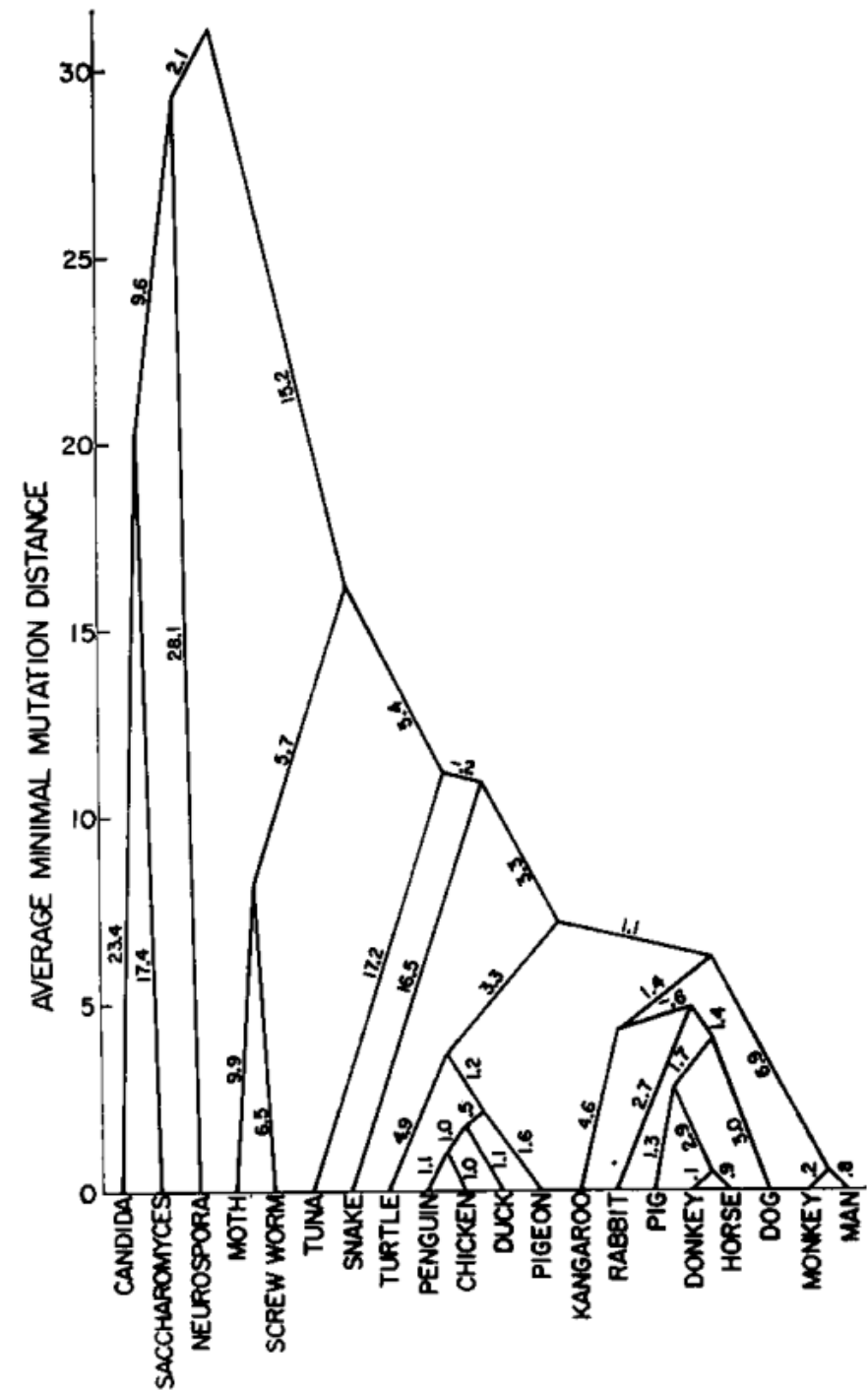


Cavalli-Sforza & Edwards (1965) Genetics Today

Sequence-based Phylogenies

Used protein sequences from a large number of organisms

Developed a computational approach for efficient analysis of large numbers of taxa (distance matrix method)



Fitch & Margoliash (1967) Science

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

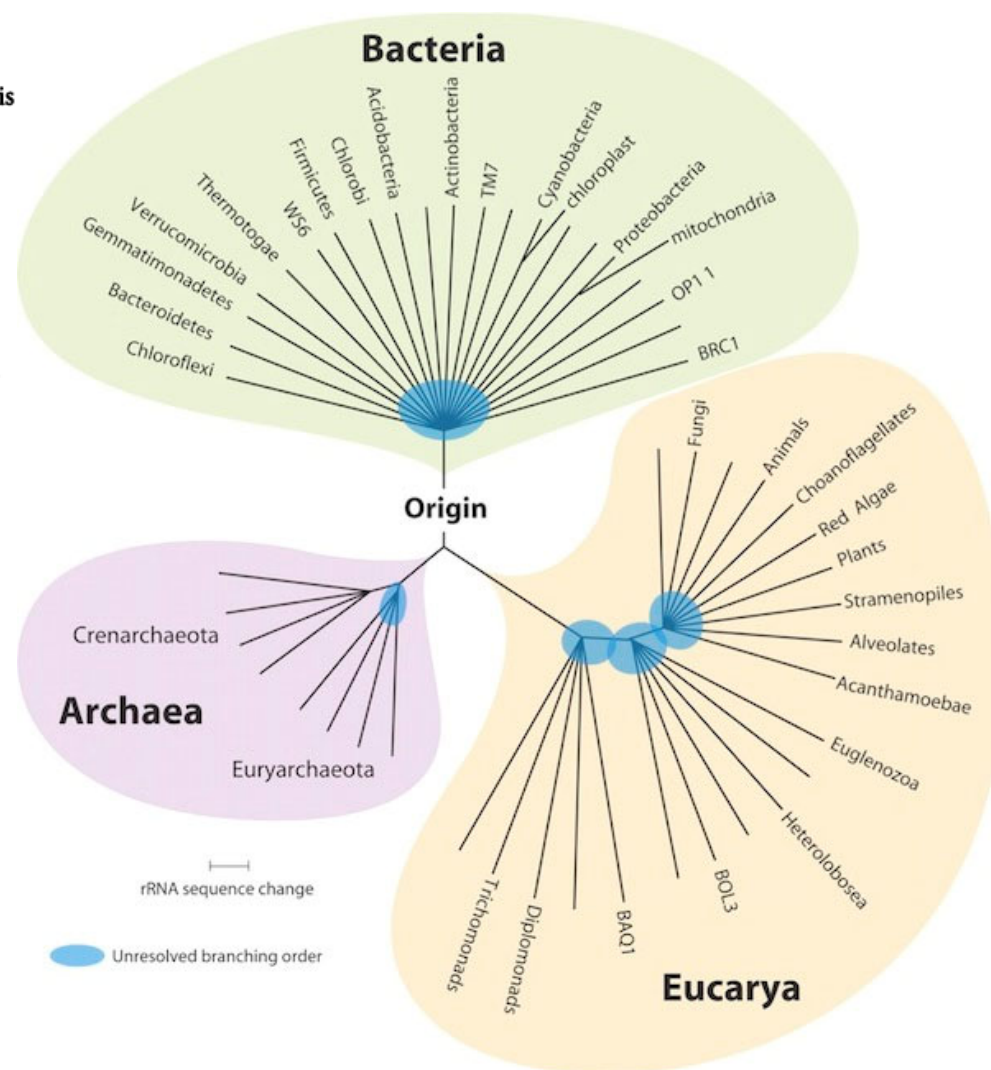
(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

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.



~~Gene tree $\approx$ Species phylogeny~~

Gene tree $\neq$ Species phylogeny

Two Types of Factors Influence the Relationship

Analytical factors

They lead to failure in accurately inferring a gene tree; these can be either due to **stochastic error** (e.g., insufficient sequence length or taxon samples) or due to **systematic error** (e.g., observed data far depart from model assumptions)

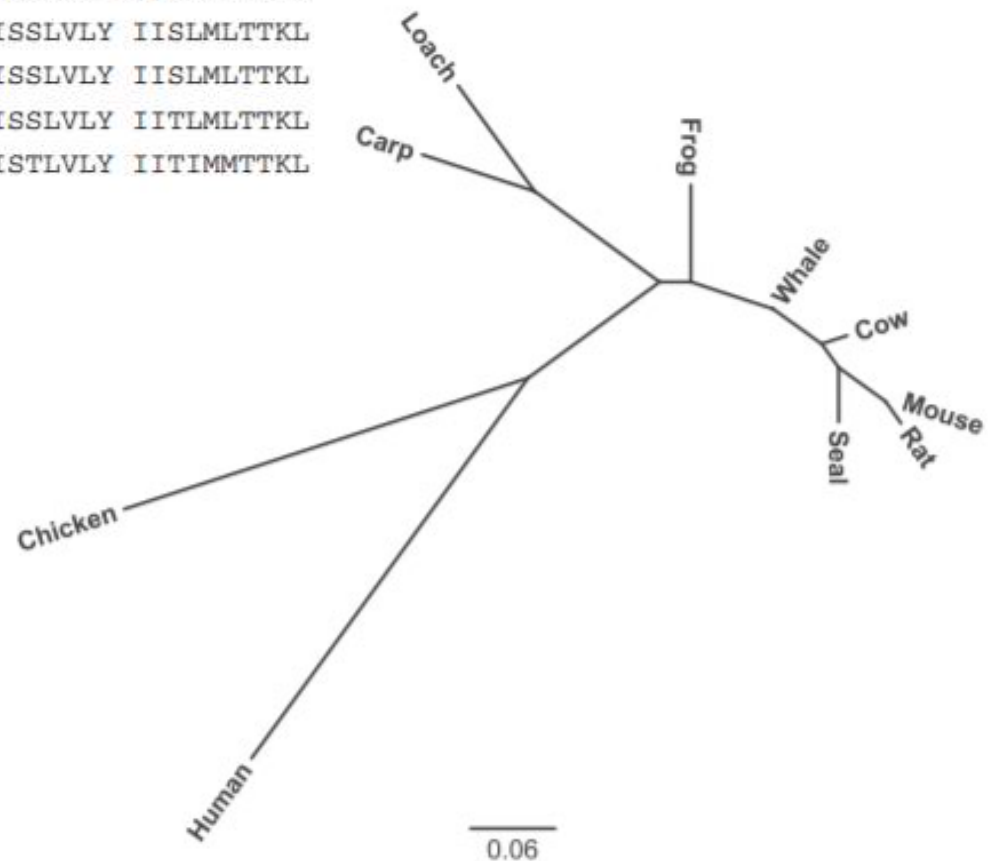
Biological factors

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 and natural selection**

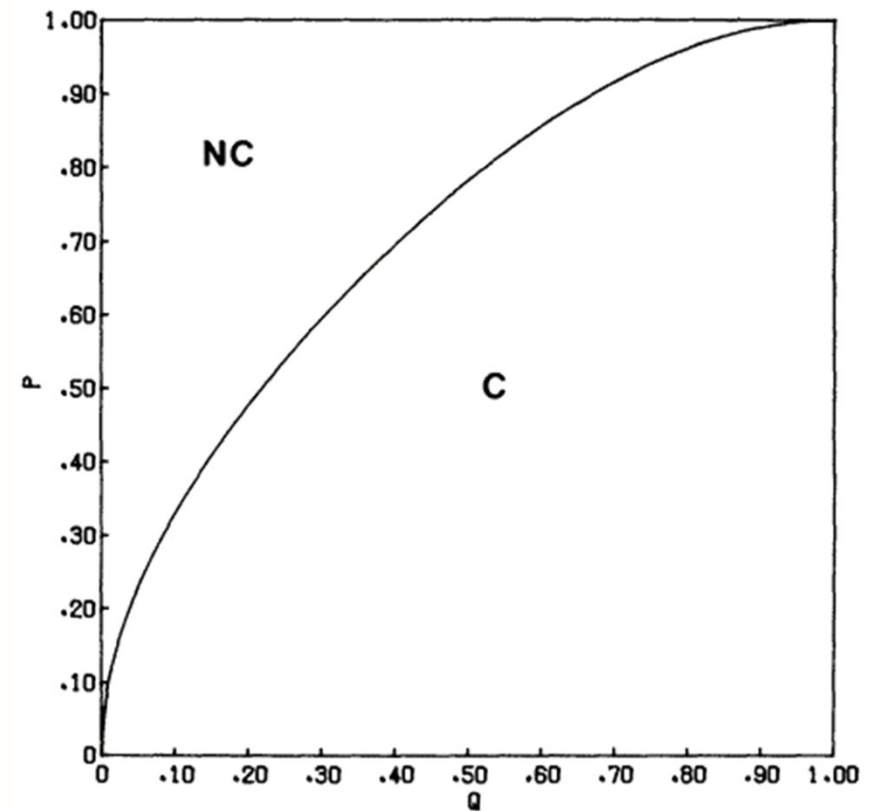
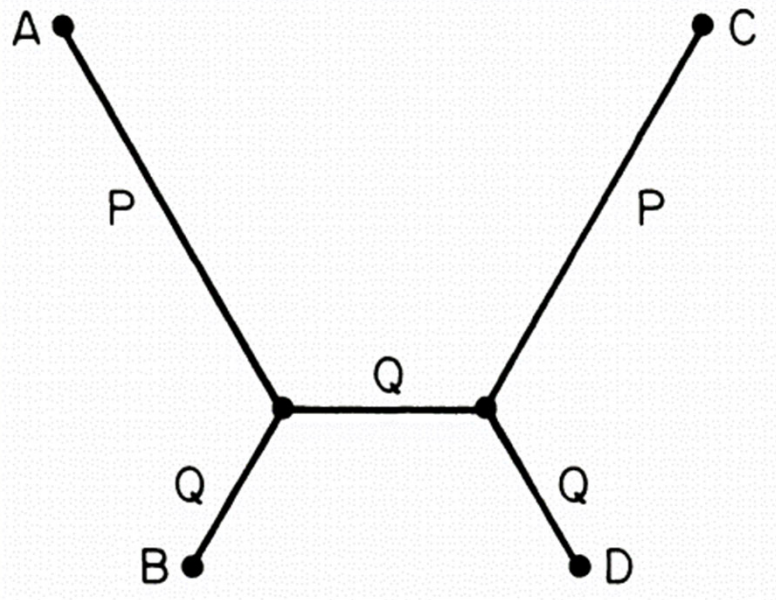
Sampling Error

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

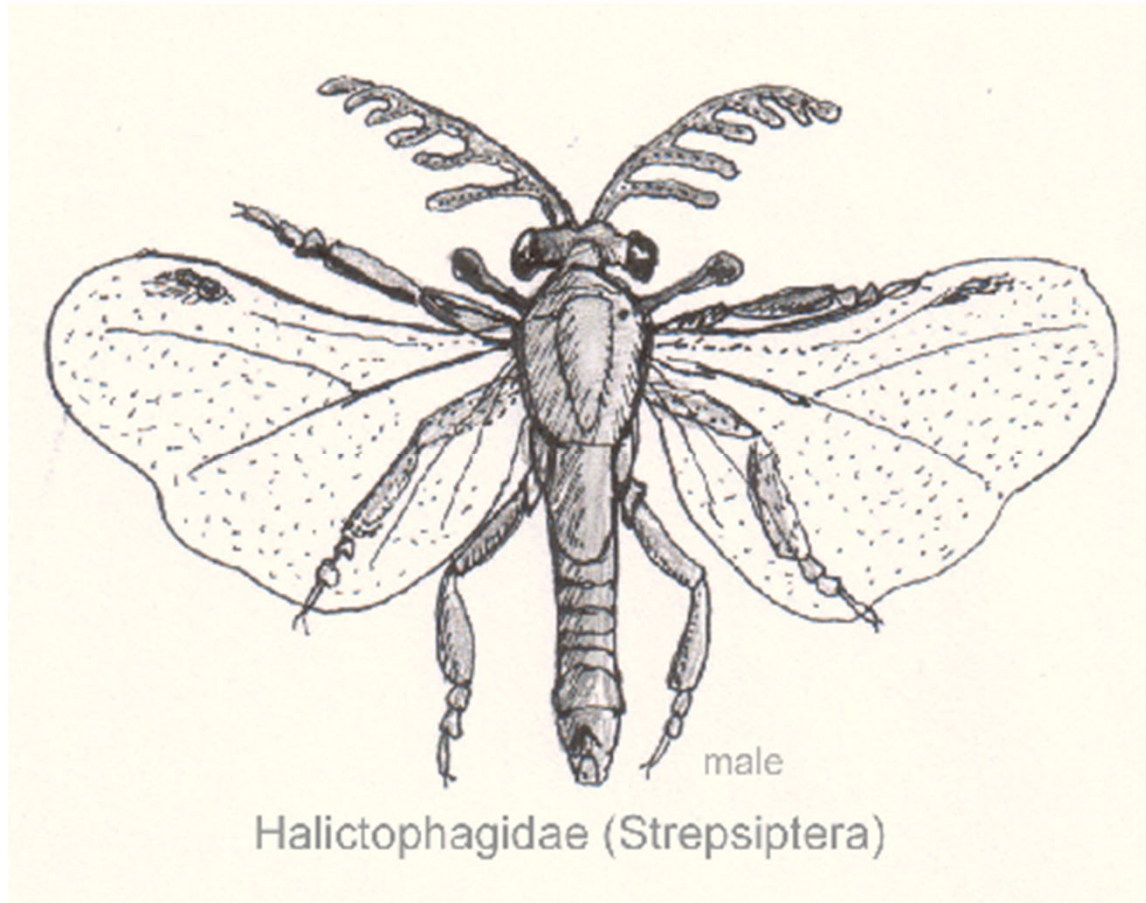


Long branch attraction

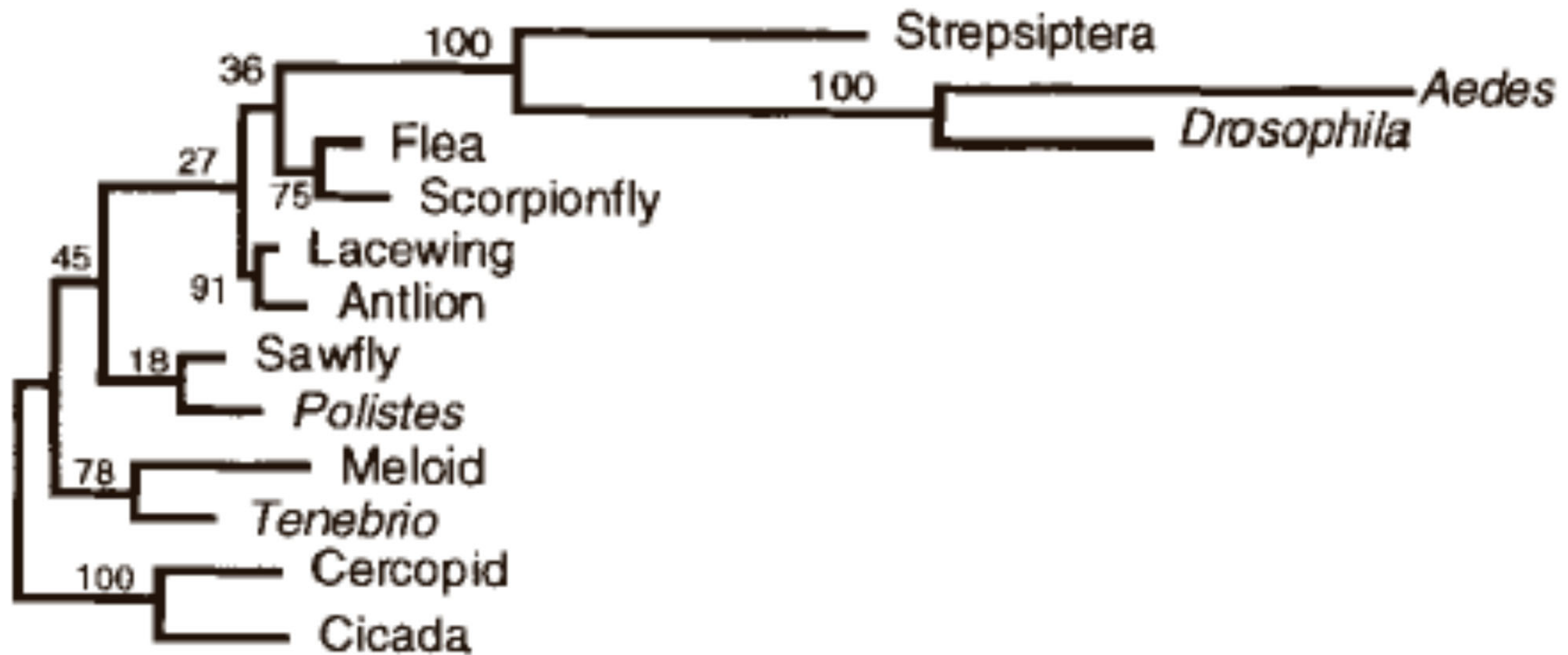


Felsenstein (1978) Syst. Zool.

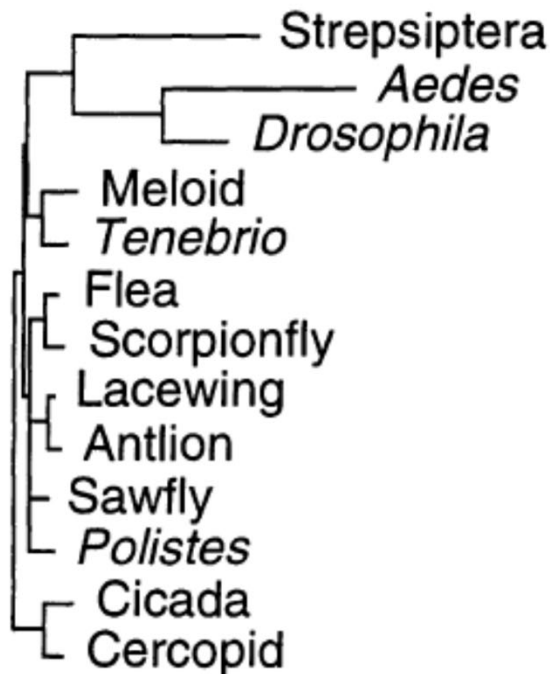
“The Strepsiptera Problem” is a Classic Example of LBA



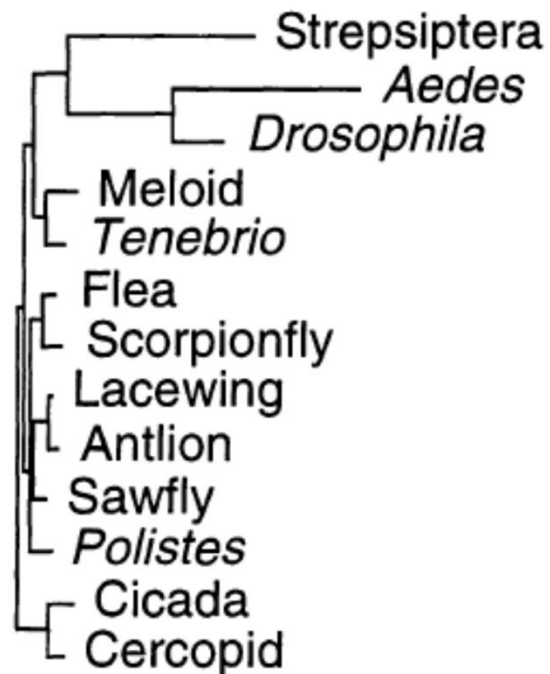
The Strepsiptera Problem



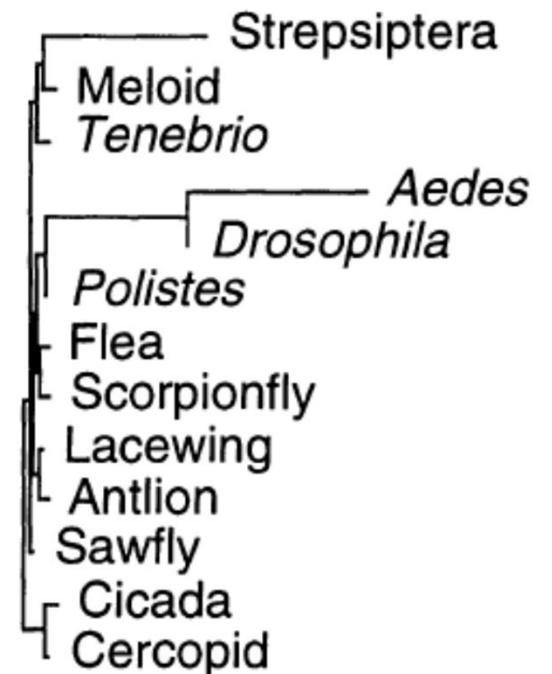
The Strepsiptera Problem



p distance



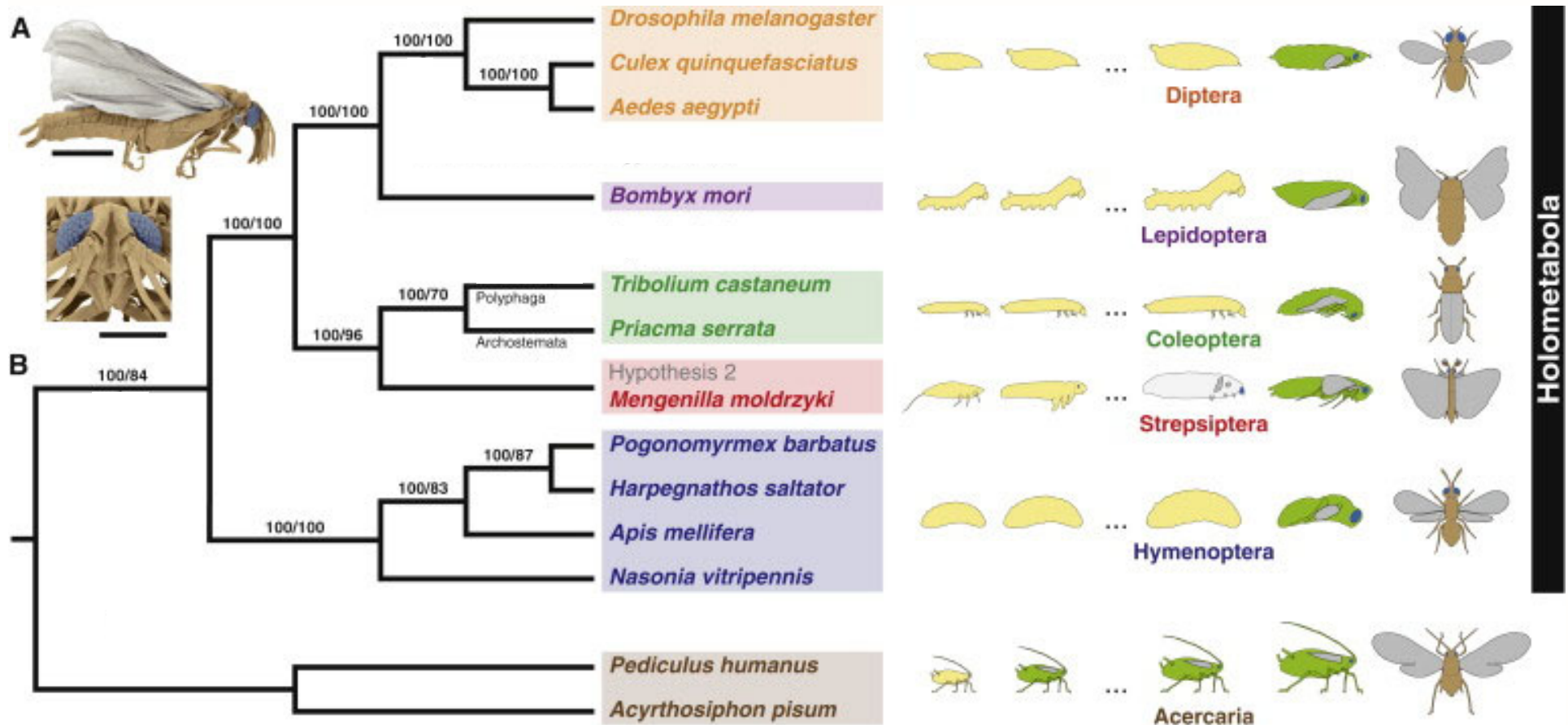
HKY85



HKY85+GAMMA

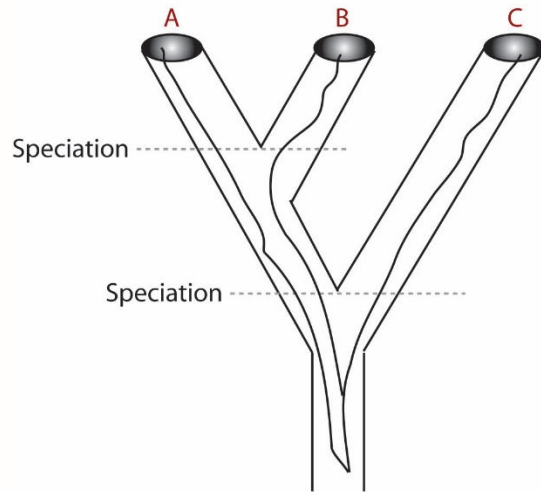


Solving the Strepsiptera Problem with More Genes and Better Models

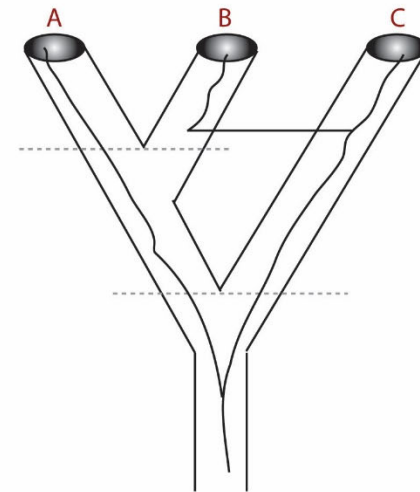


Biological Factors

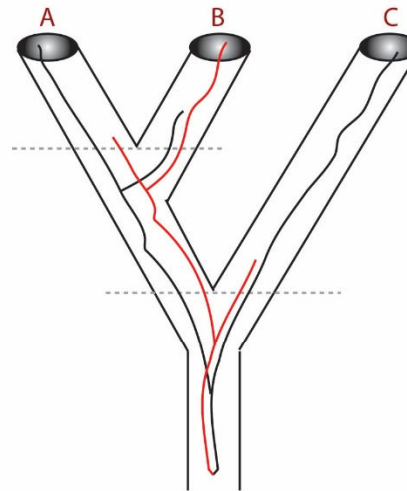
Lineage Sorting



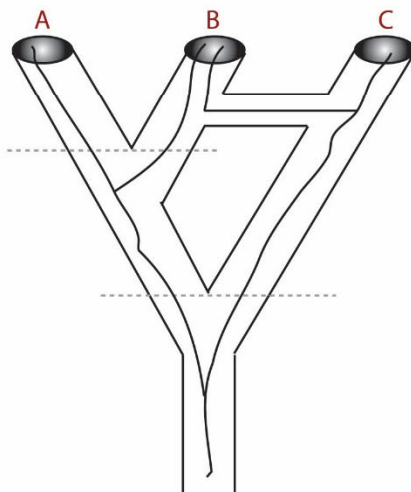
Horizontal Gene Transfer



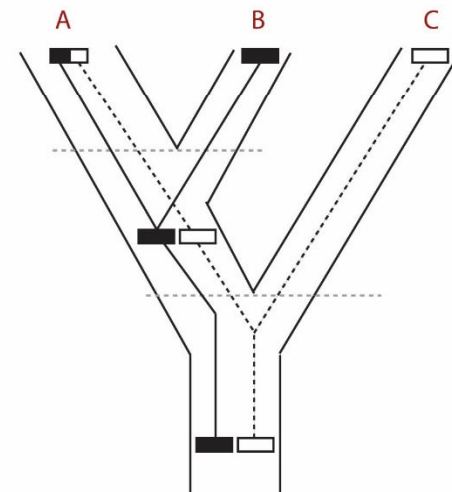
Gene Duplication and Loss



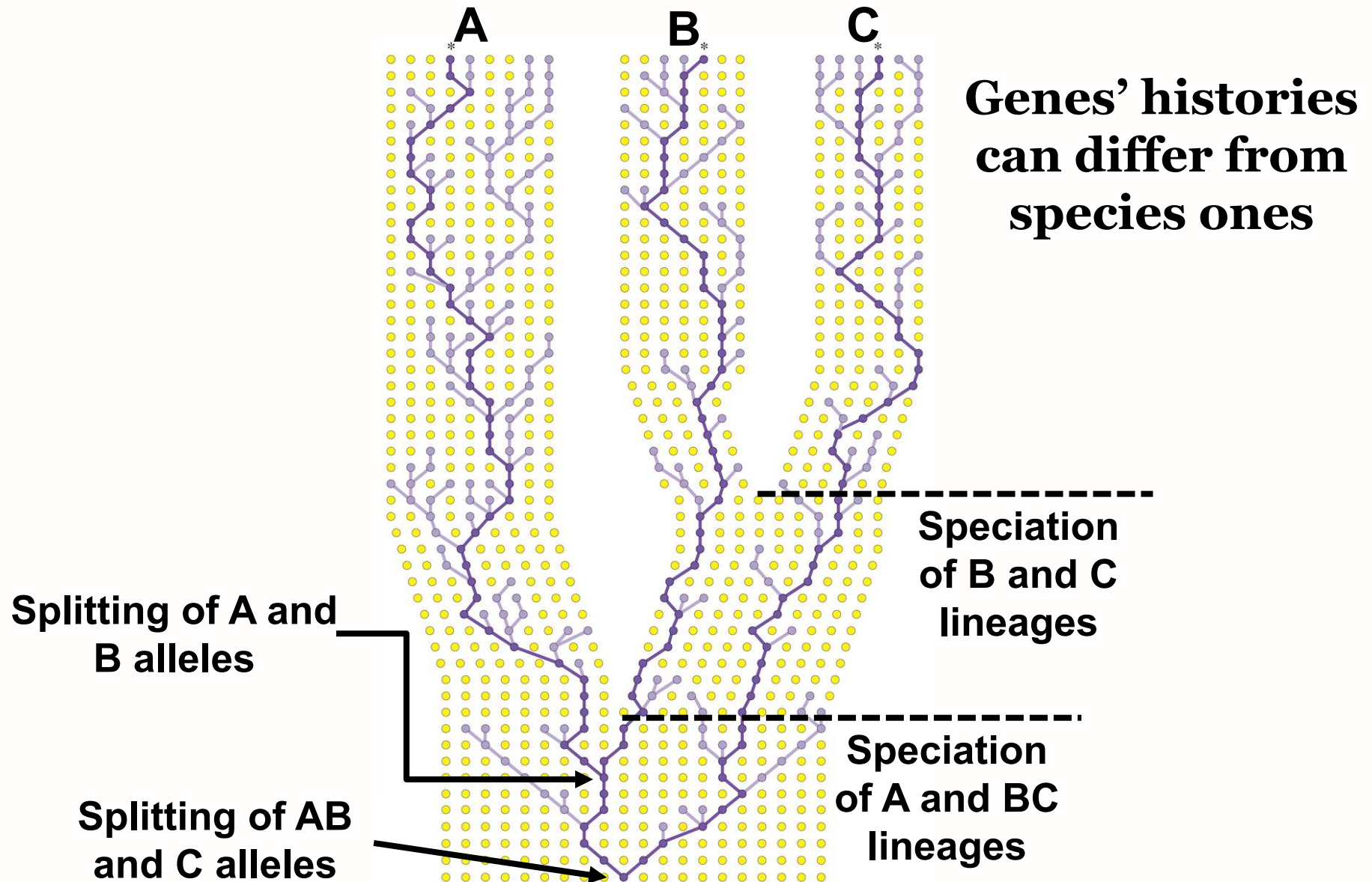
Hybridization



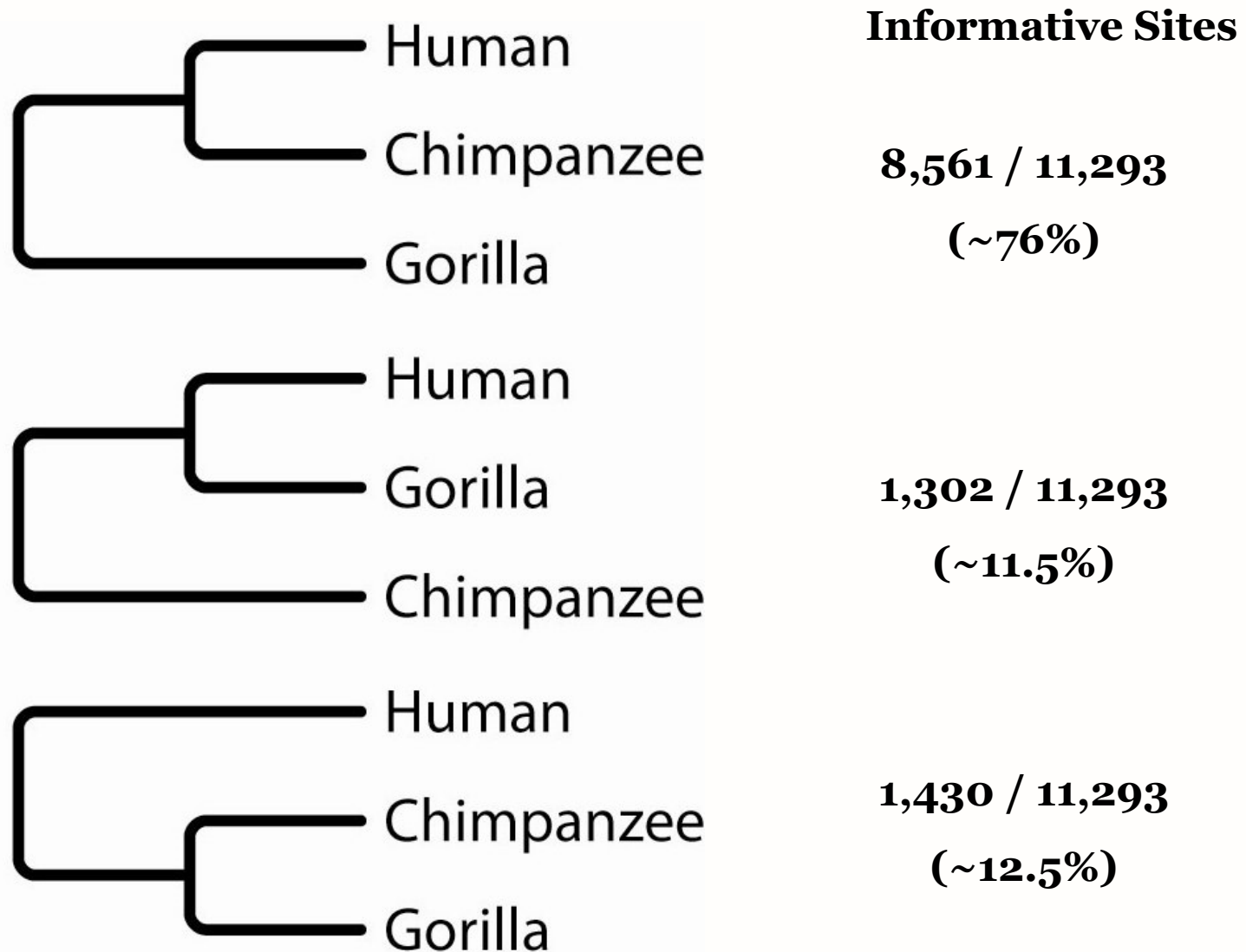
Recombination



Stochastic Lineage Sorting of Ancestral Polymorphisms

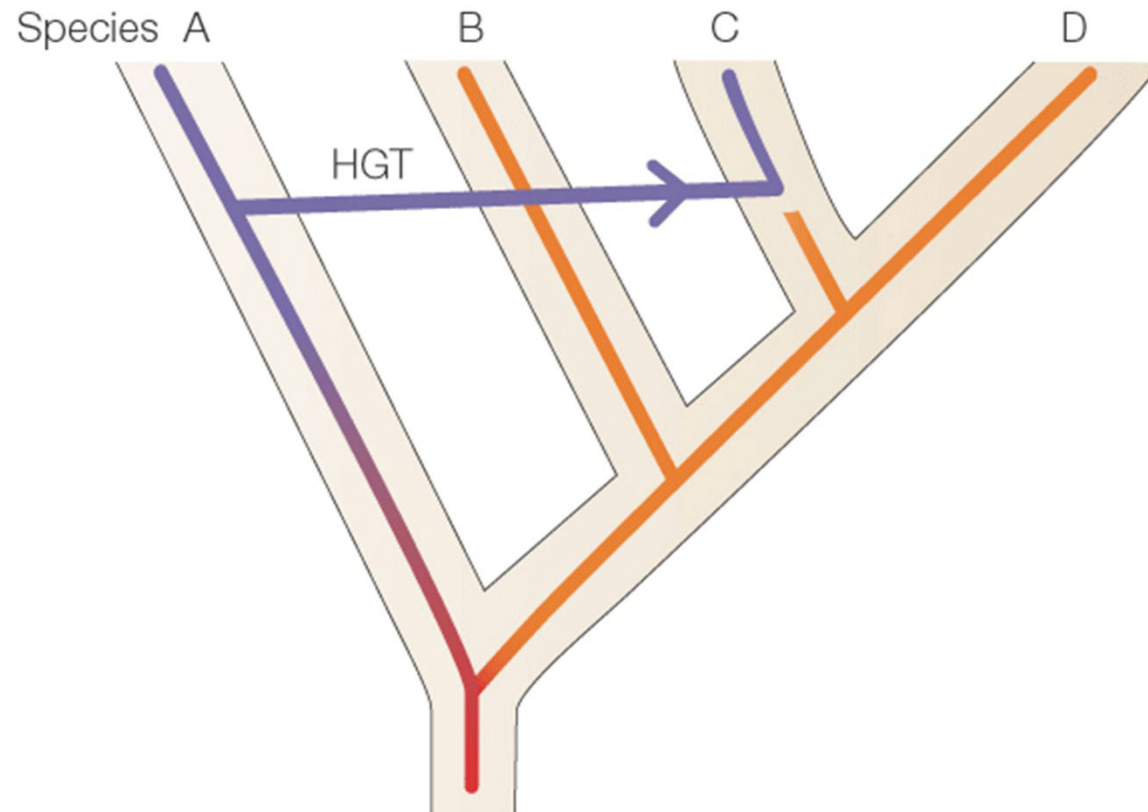


Lineage Sorting in Primates

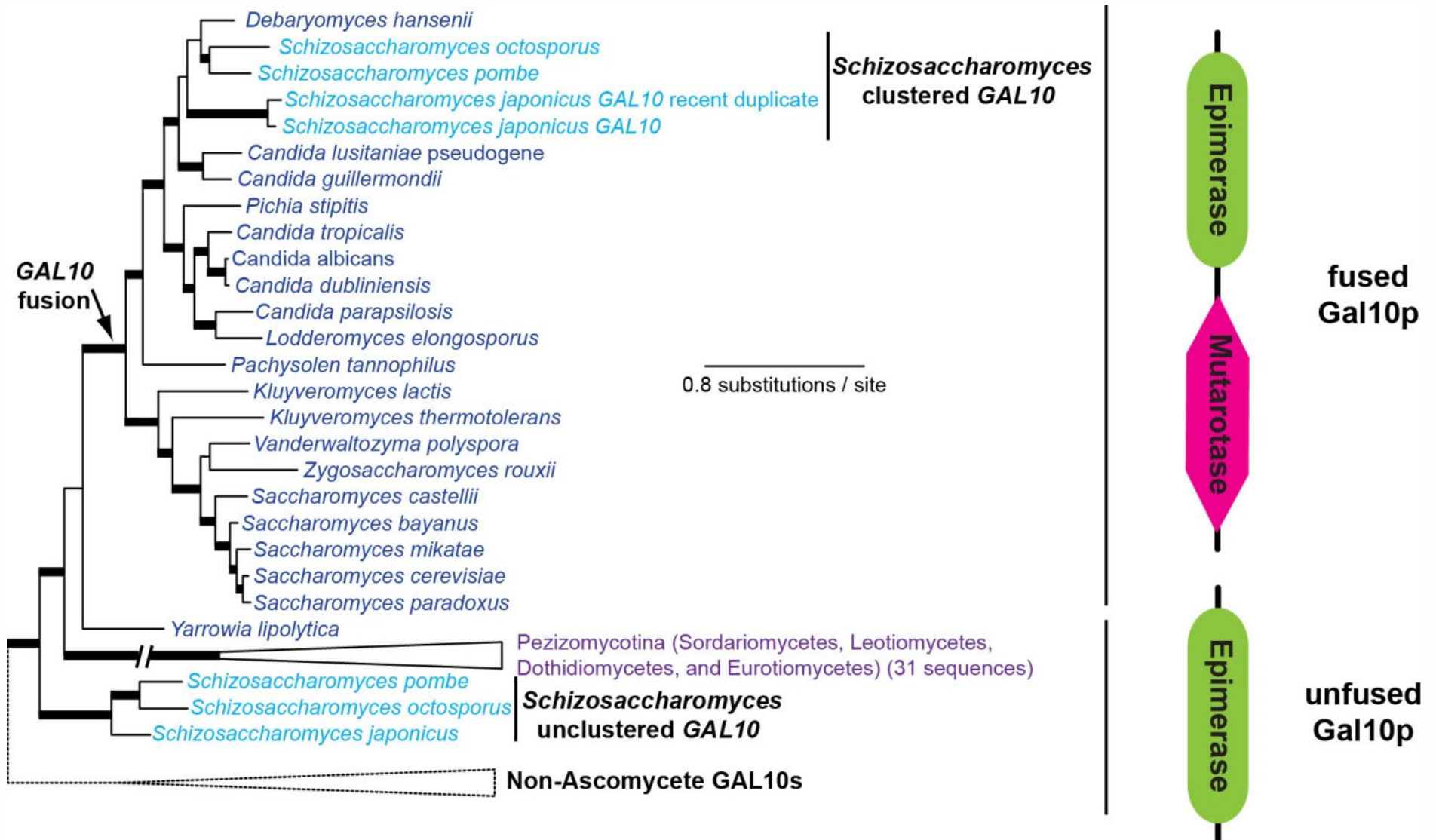


Horizontal Transfer of Genes

Exchange of genes between organisms other than through reproduction



Horizontal Gene Transfer in Fungi

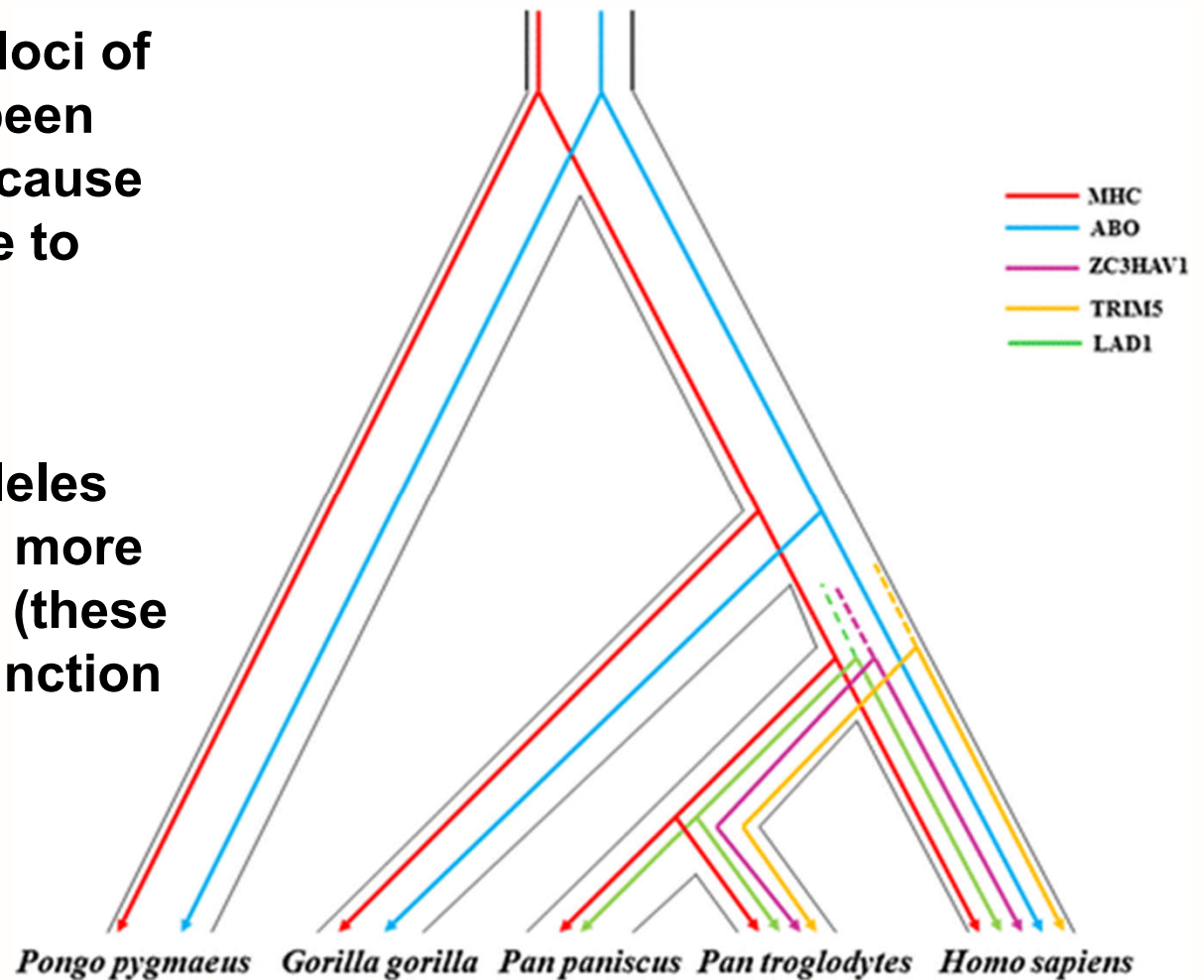


Balancing Selection

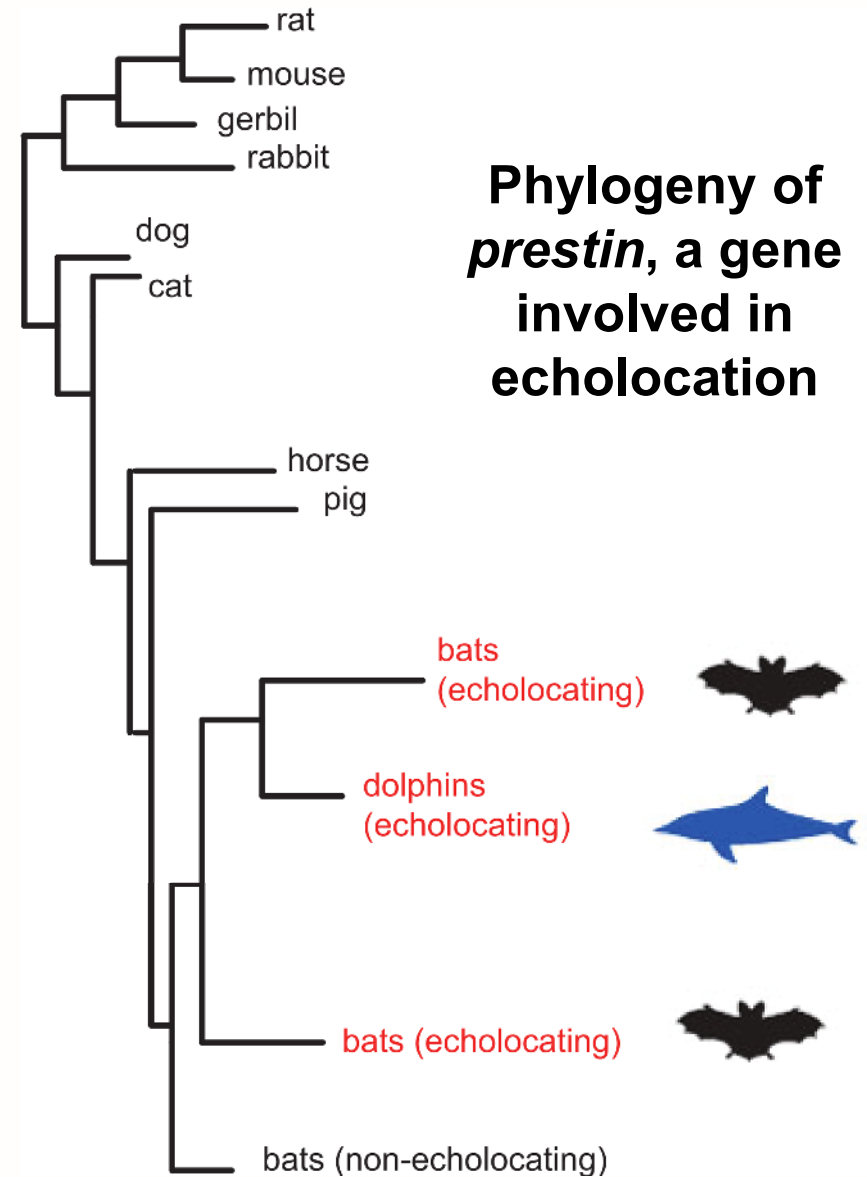
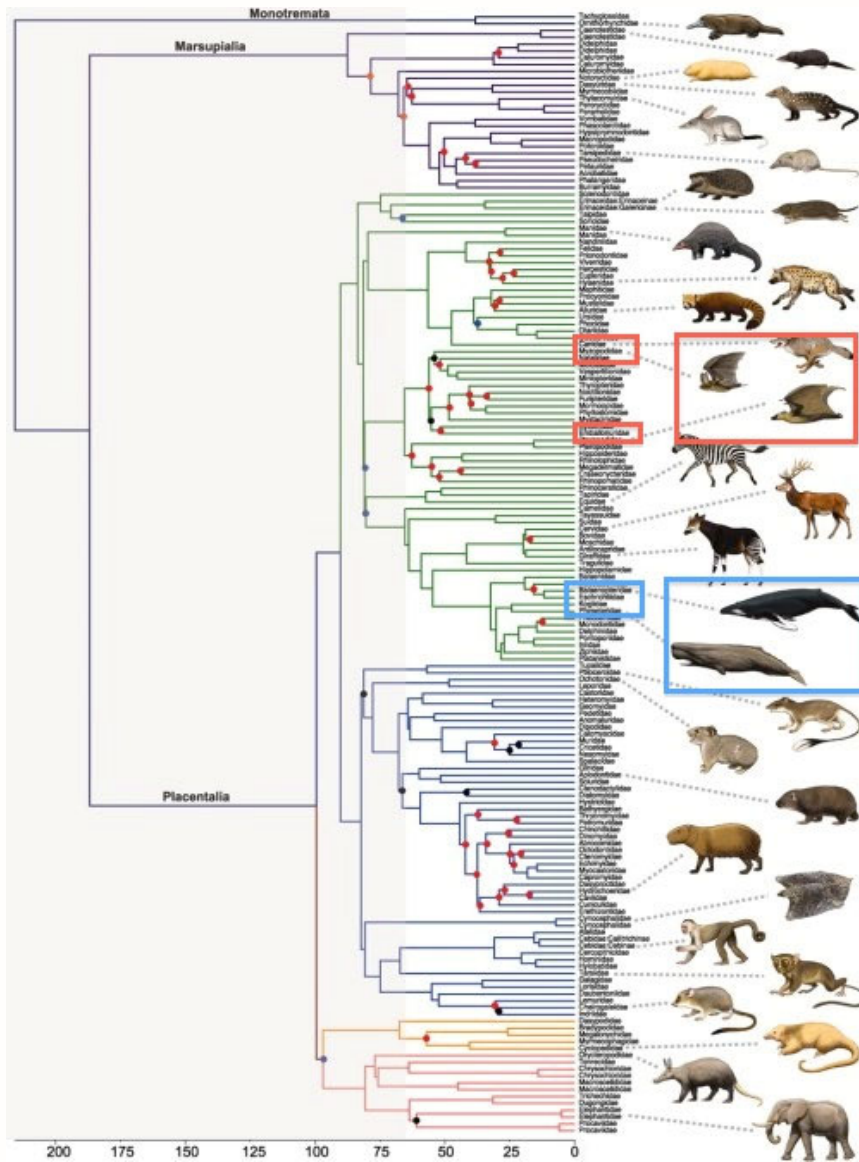
Balancing selection can maintain “trans-species polymorphisms”, in which the alleles are more ancient than the species

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!!!)

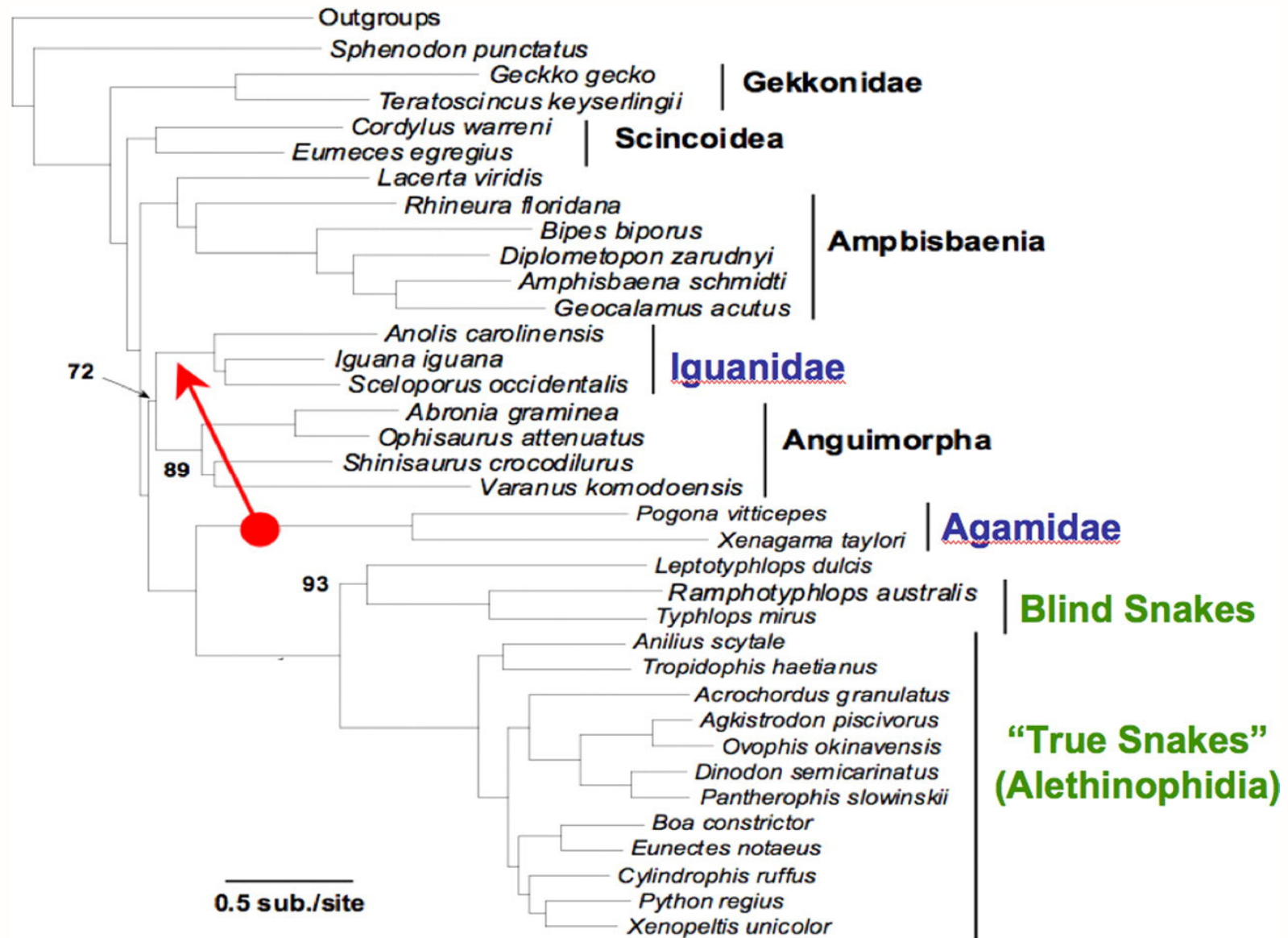


Positive Selection



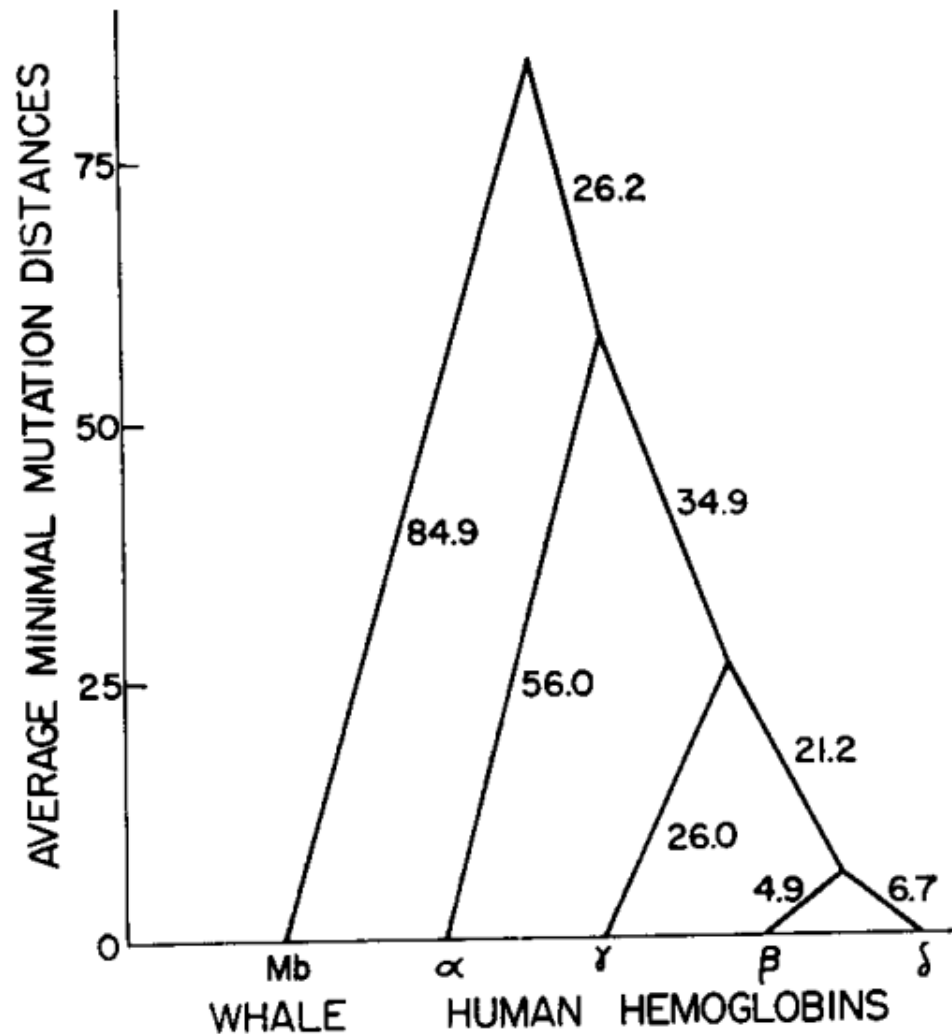
Li et al. (2010) Curr. Biol.

Positive Selection



Castoe et al. (2009) PNAS

Gene Duplication Can Confound Phylogeny



(15). A cautionary note may be derived from this. A wildly incorrect result could easily be obtained if the presence of multiple, homologous genes were not recognized and a phylogeny were constructed from sequences which were coded for, say, half by genes for alpha hemoglobin chains and half by genes for beta hemoglobin chains. This results from the speciation having occurred more recently than the gene duplication which permitted the separate evolution of the alpha and beta genes.

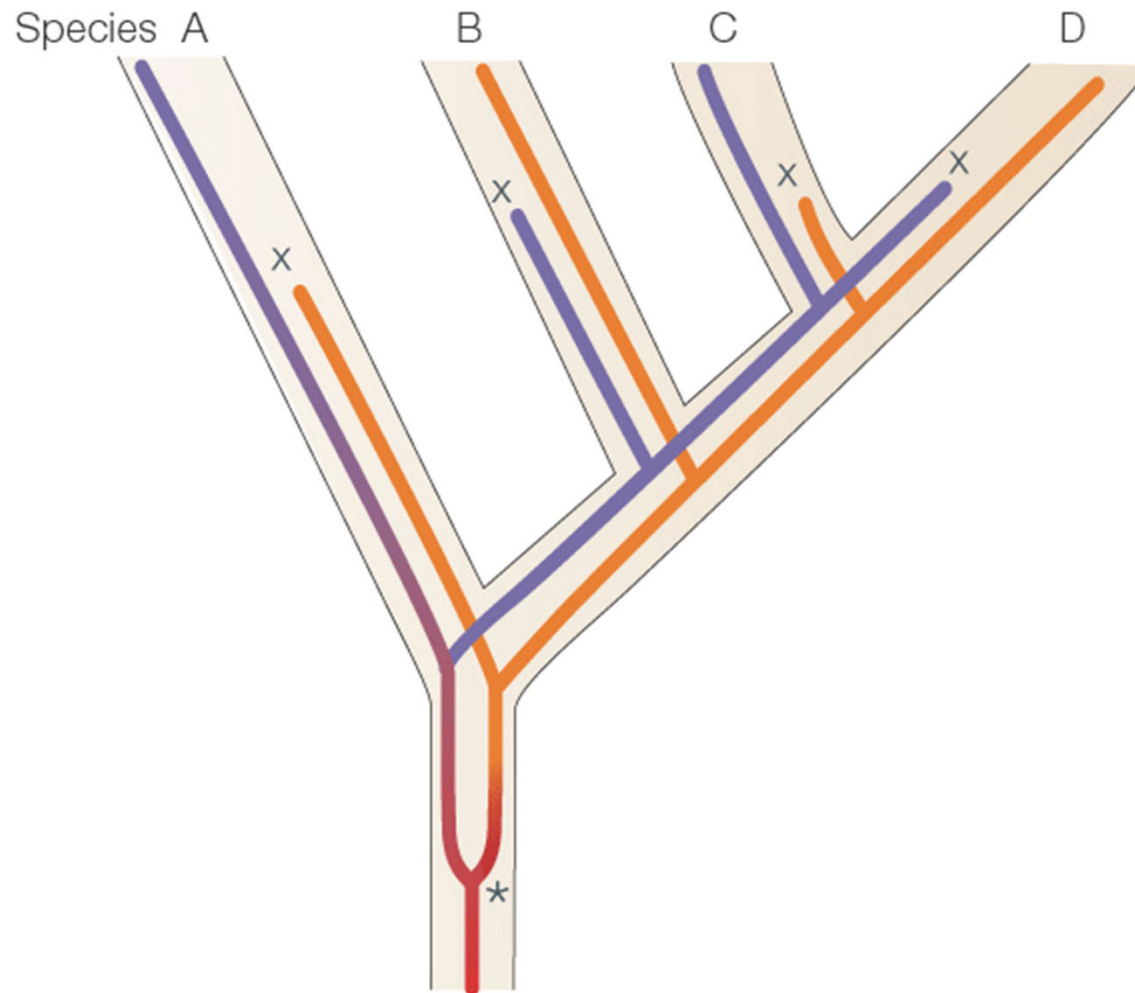


Gene Duplication Can Confound Phylogeny

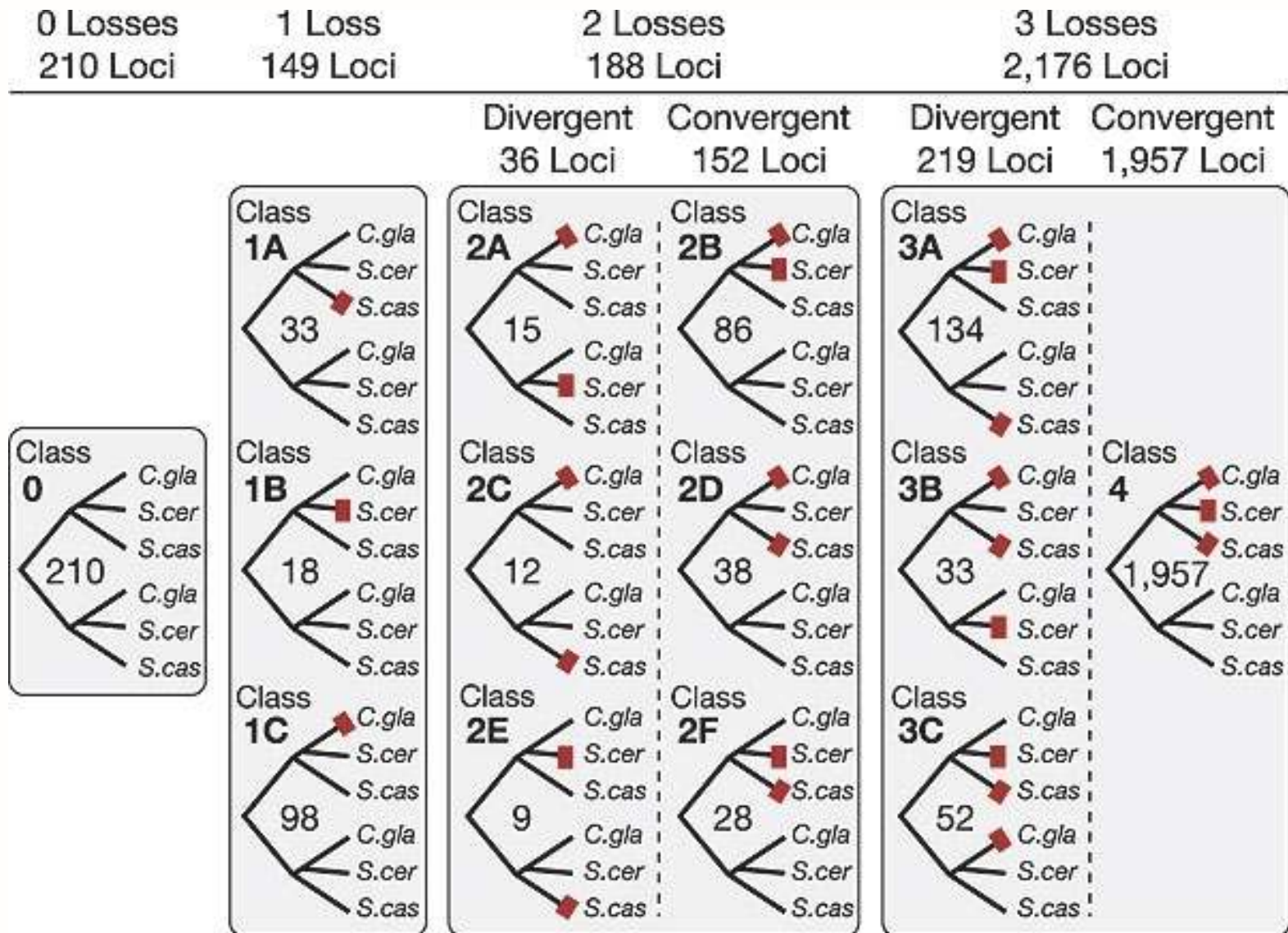
... amino acid sequences that the proteins be homologous. It has been pointed out before that a phylogeny of birds and mammals based upon a haphazard mixture of α and β hemoglobins would be biological nonsense since the initial dichotomy would be on the distinction between the α and β genes rather than between the birds and the mammals (Fitch and Margoliash, 1967). Therefore, there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism, (for example, α and β hemoglobin) the genes should be called *paralogous* (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called *orthologous* (ortho = exact). Phylogenies require orthologous, not paralogous, genes. Note



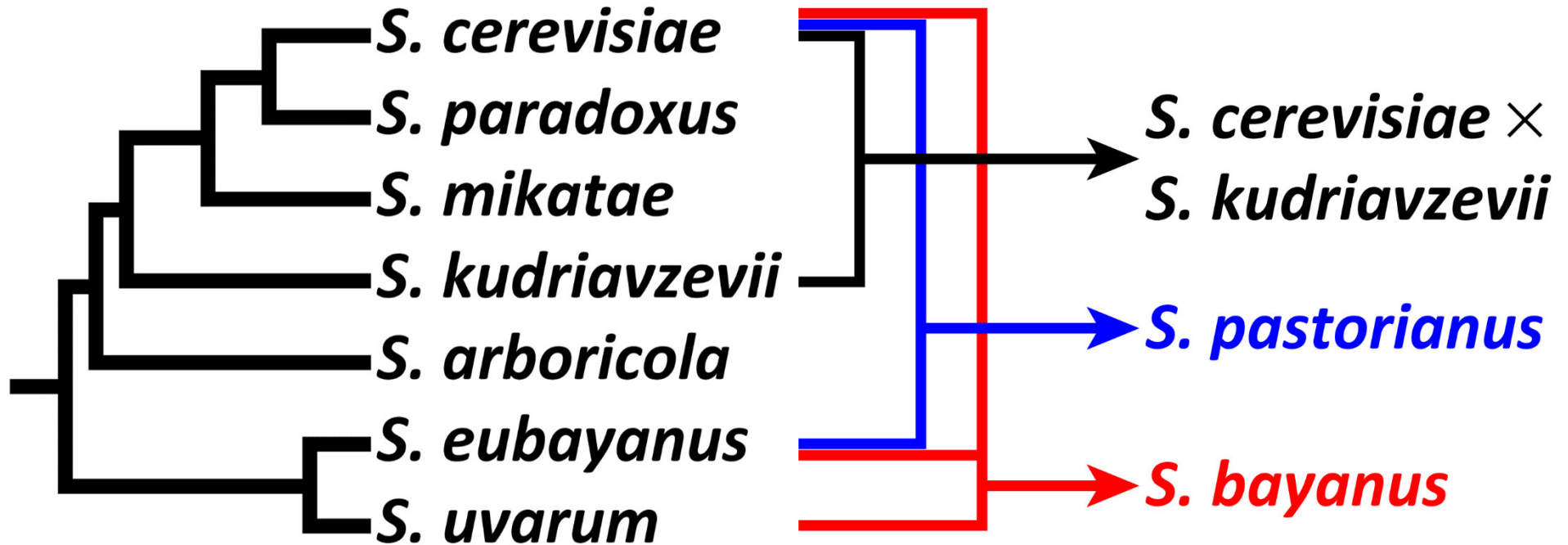
Gene Duplication and Loss



Gene Duplication and Loss



Hybridization / Introgression



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

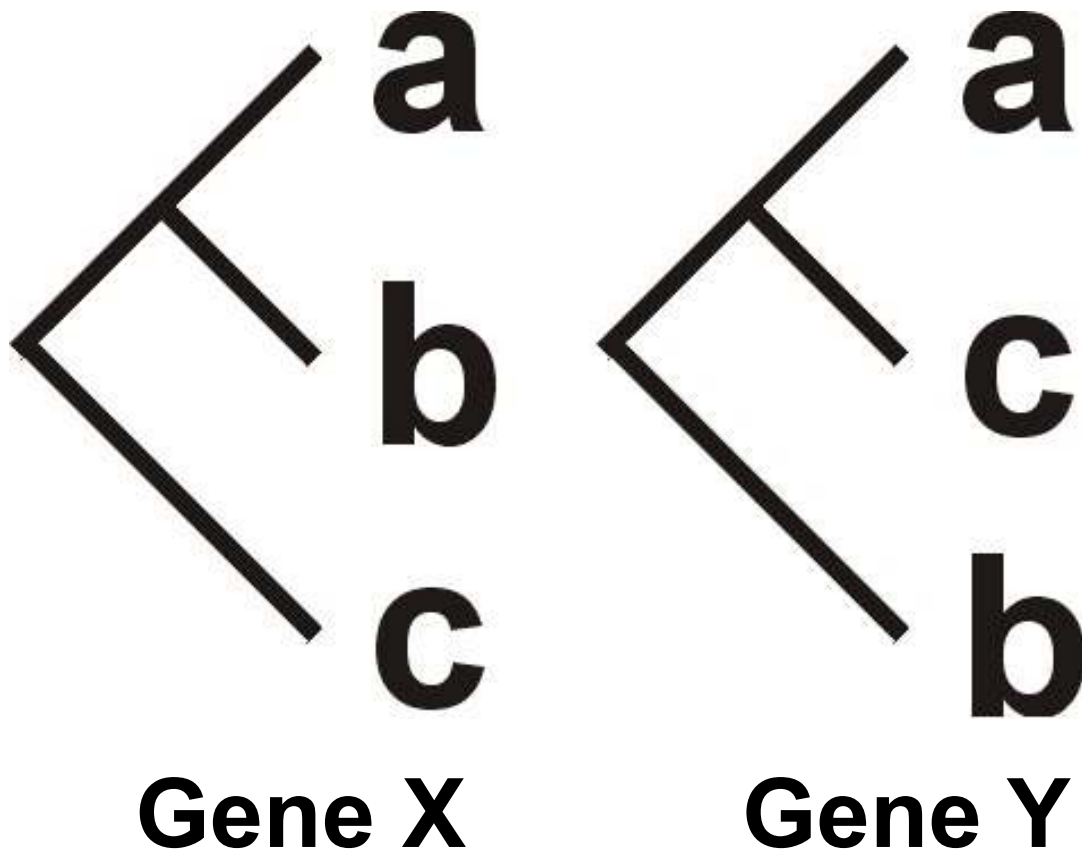
S. cerevisiae – *S. paradoxus* divergence $\approx$ human – mouse divergence
S. cerevisiae – *S. uvarum* divergence $\approx$ human – chicken divergence



**OK, I now get why
gene trees $\neq$ species trees**

**What does this have to do with
phylogenomics?**

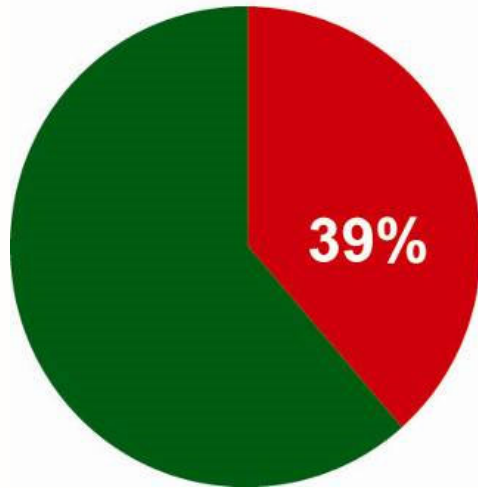
All this Manifests Itself as Incongruence



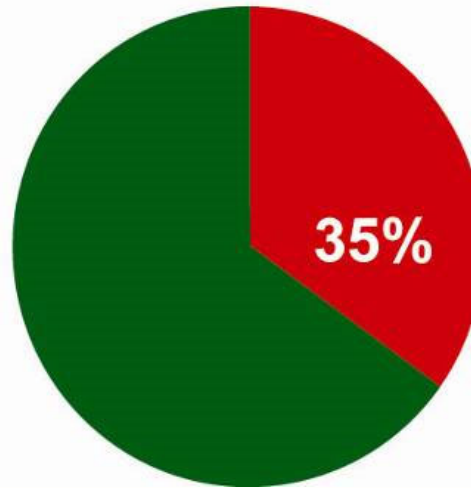
**Species
phylogeny?**

Incongruence is Pervasive in the Phylogenetics Literature

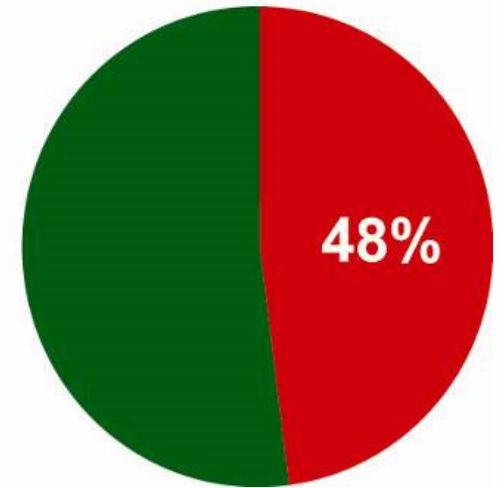
A: All organisms



B: Mammals



C: Insects



A Systematic Evaluation of Single Gene Phylogenies



S. cerevisiae

S. paradoxus

S. mikatae

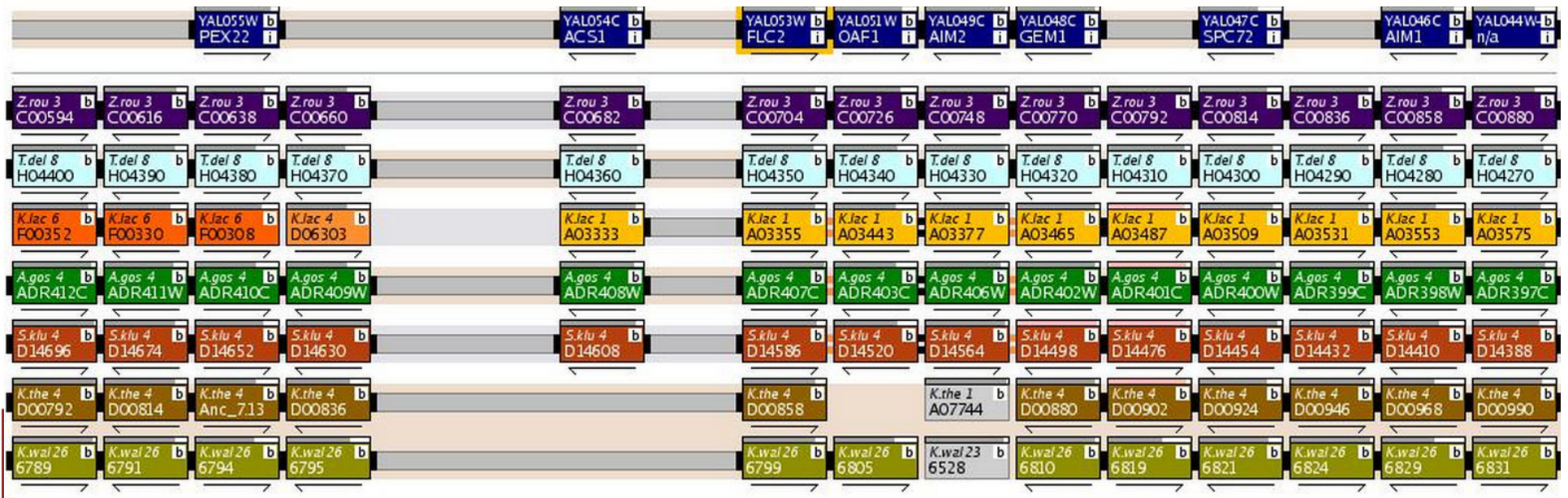
S. kudriavzevii

S. bayanus

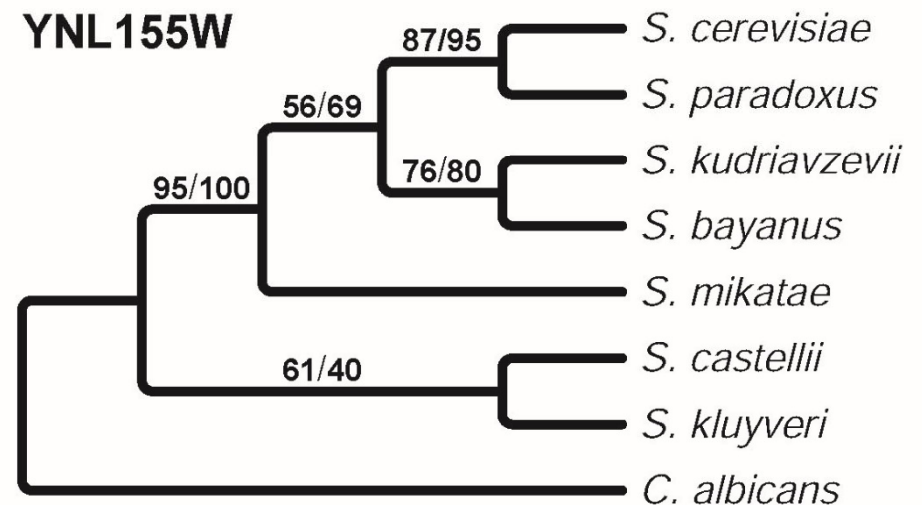
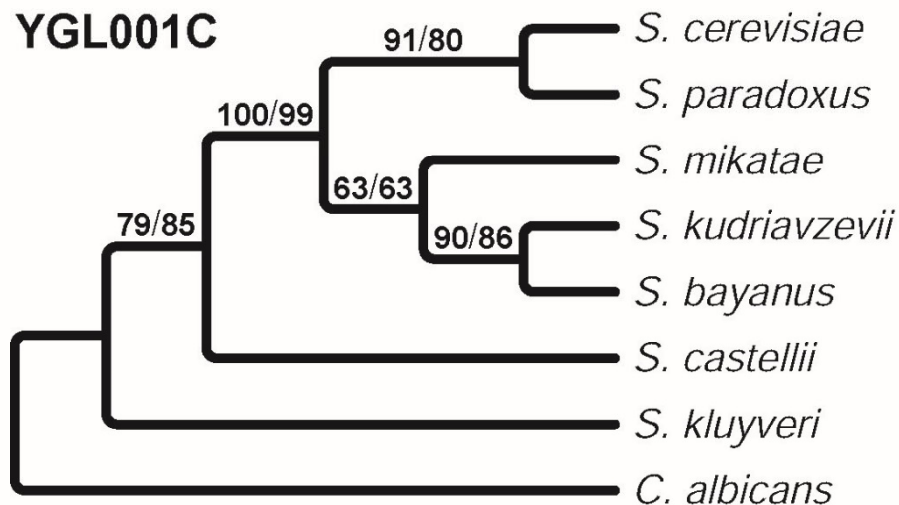
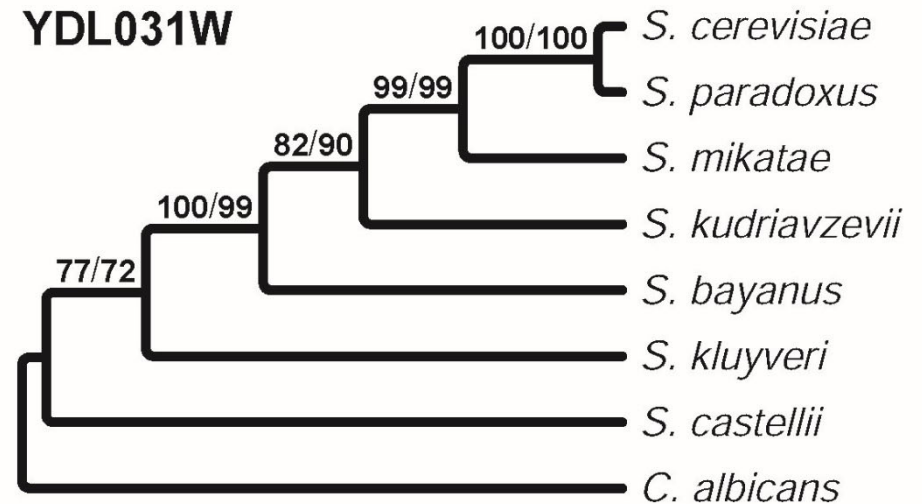
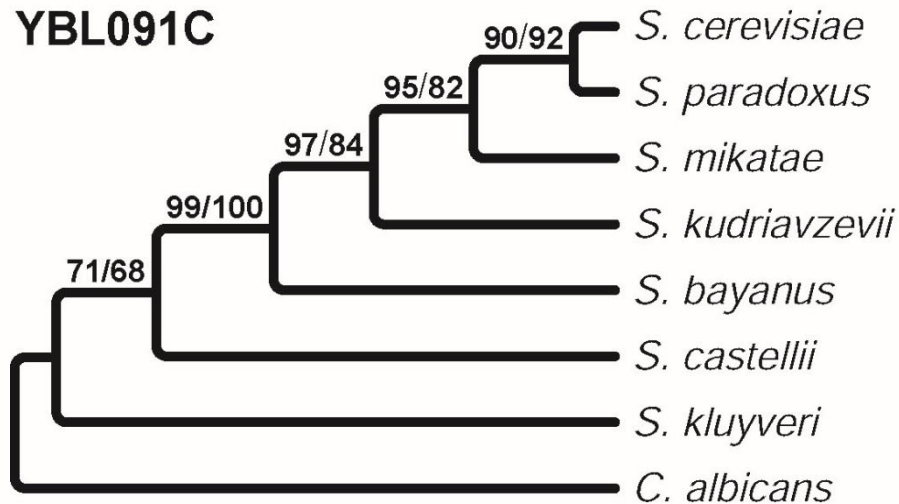
S. castellii

S. kluyveri

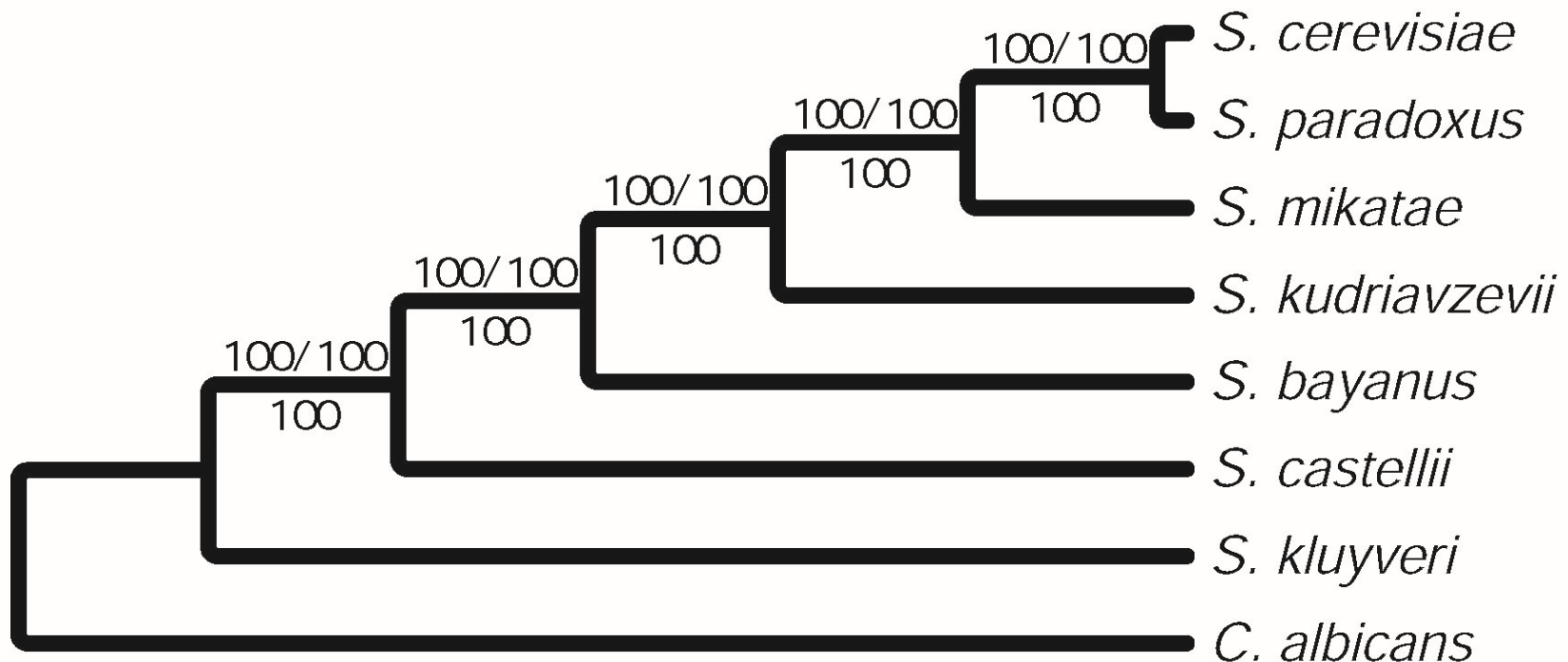
Candida glabrata



Incongruence at the Single Gene Level



Concatenation of 106 Genes Yields a Single Yeast Phylogeny

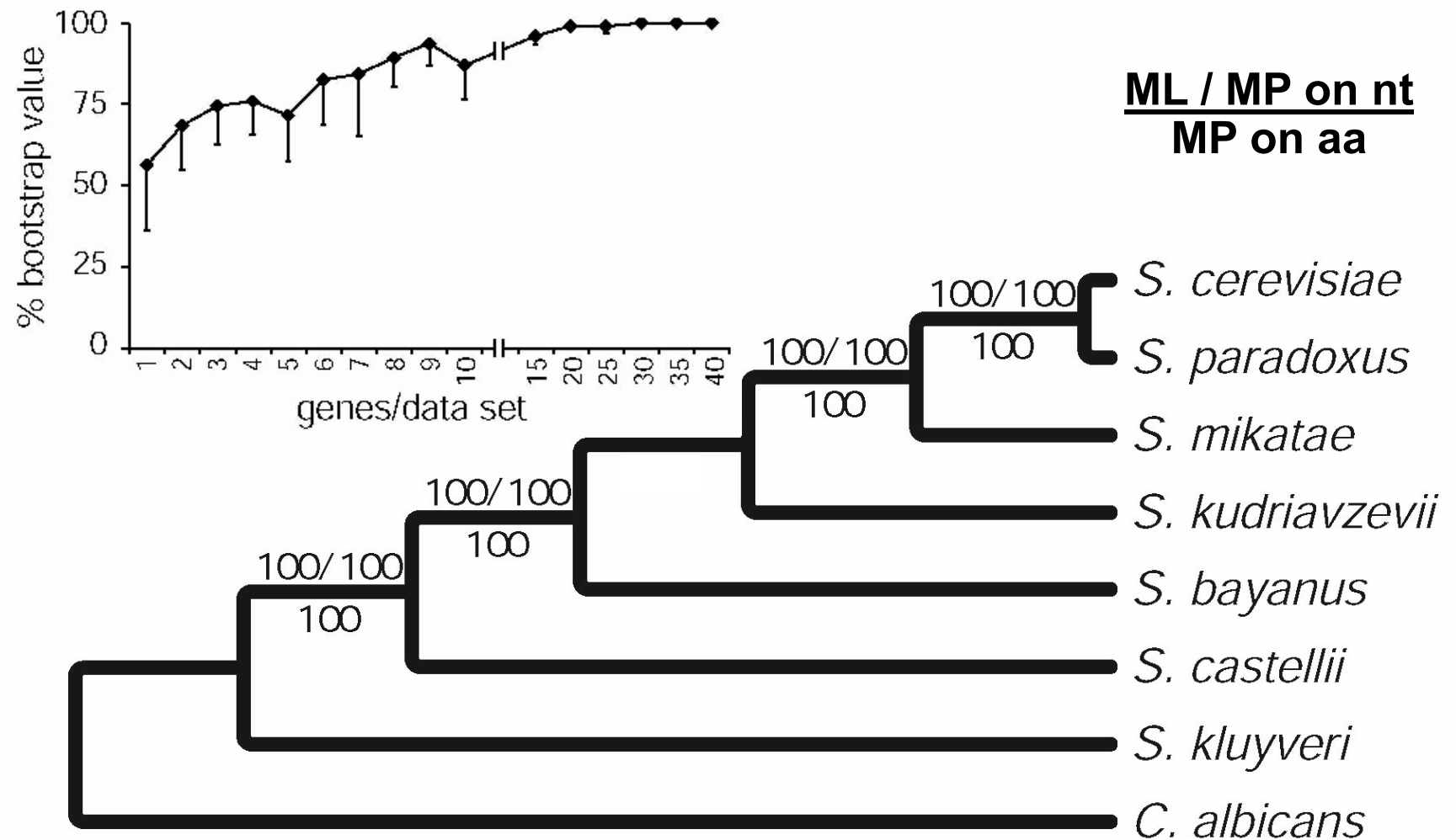


ML / MP on nt
MP on aa



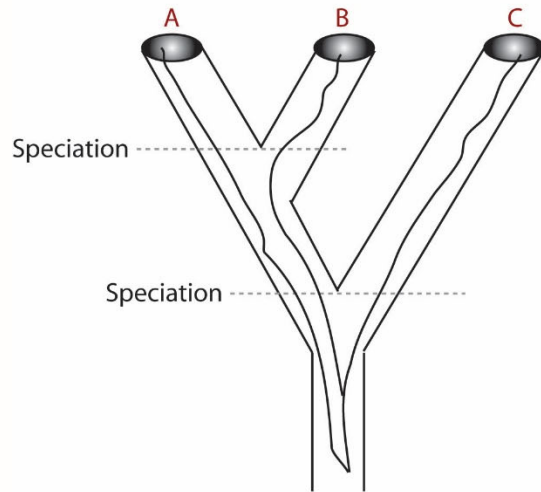
Rokas et al. (2003) Nature

The Use of Many Genes Eliminates Stochastic Error

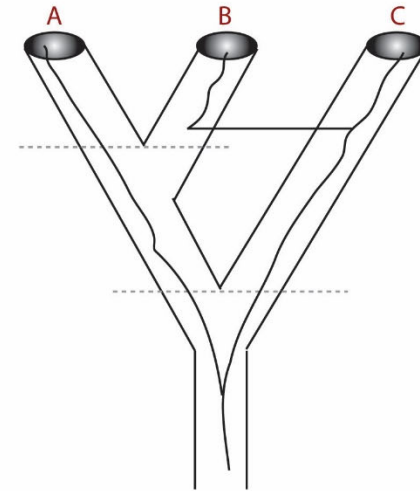


Gene Trees Can Differ from Species Trees

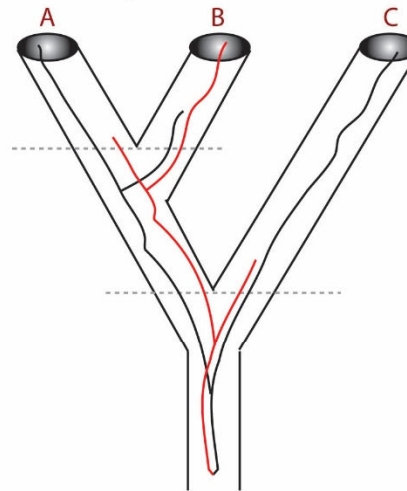
Lineage Sorting



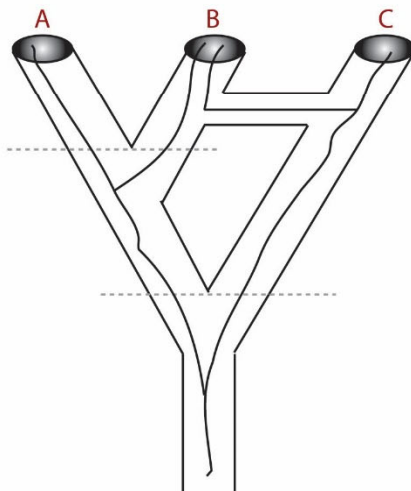
Horizontal Gene Transfer



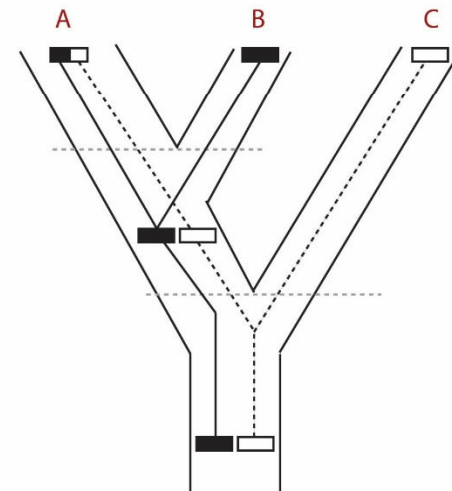
Gene Duplication and Loss



Hybridization

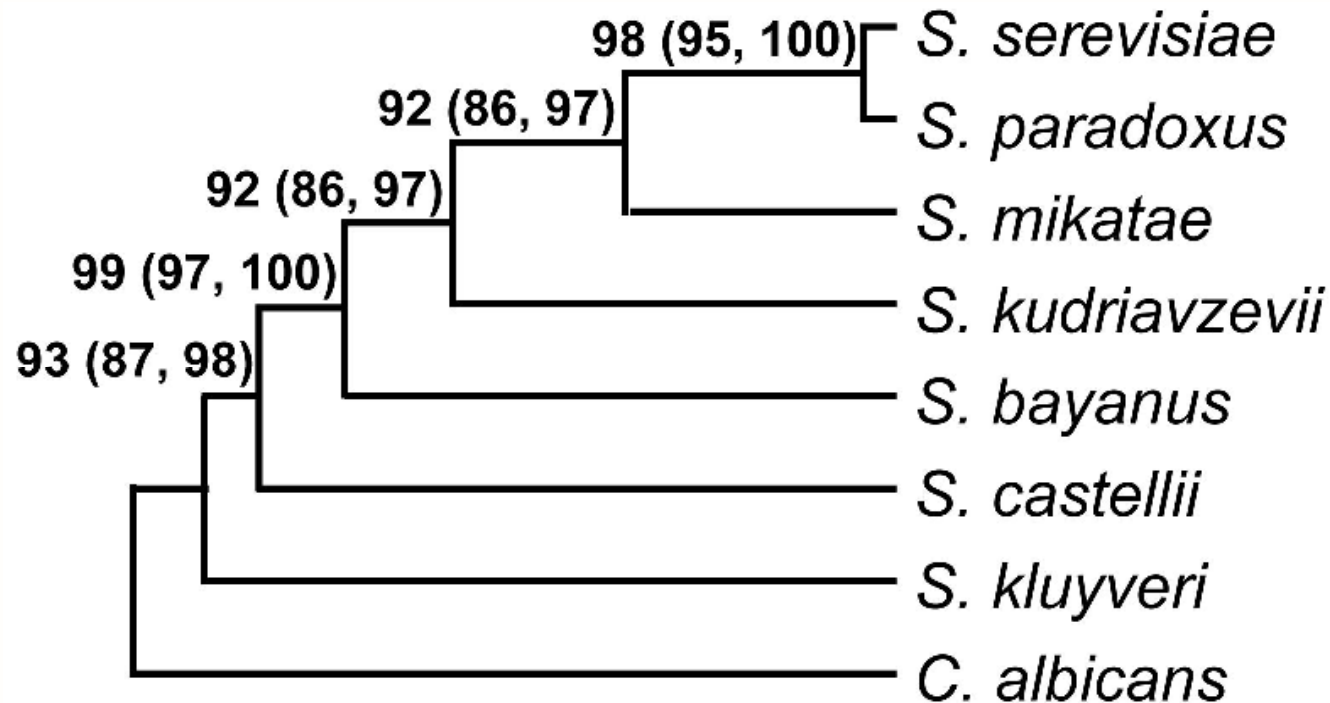


Recombination



Inferring the Species Tree from Individual Gene Histories

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



New methods to calculate concordance factors for phylogenomic datasets

Bui Quang Minh, Matthew Hahn, Robert Lanfear

doi: <https://doi.org/10.1101/487801>

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

This article is a preprint and has not been peer-reviewed [what does this mean?].



Ané et al. (2007) Mol. Biol. Evol.

The Phylogenomics Era – “Resolving” the Tree of Life

Syst. Biol. 61(1):150–164, 2012

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

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

DOI:10.1093/sysbio/syr089

Advance Access publication on September 7, 2011

LETT
LETT

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

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

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

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

Resolving the evolutionary relationships of molluscs with phylogenomic tools

nature

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

Syst. Biol. 57(6):920–938, 2008

Copyright © Society of Systematic Biologists

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

DOI: 10.1080/10635150802570791

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

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

Toward Resolving the Tree: The Phylogeny of Jakobids and Cercozoans

Yeast

An

Toward Resolving Priors

Prion-Like Proteins in the Fungal Kingdom

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

OPEN ACCESS Free

Towards

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

Samuli Lehtonen

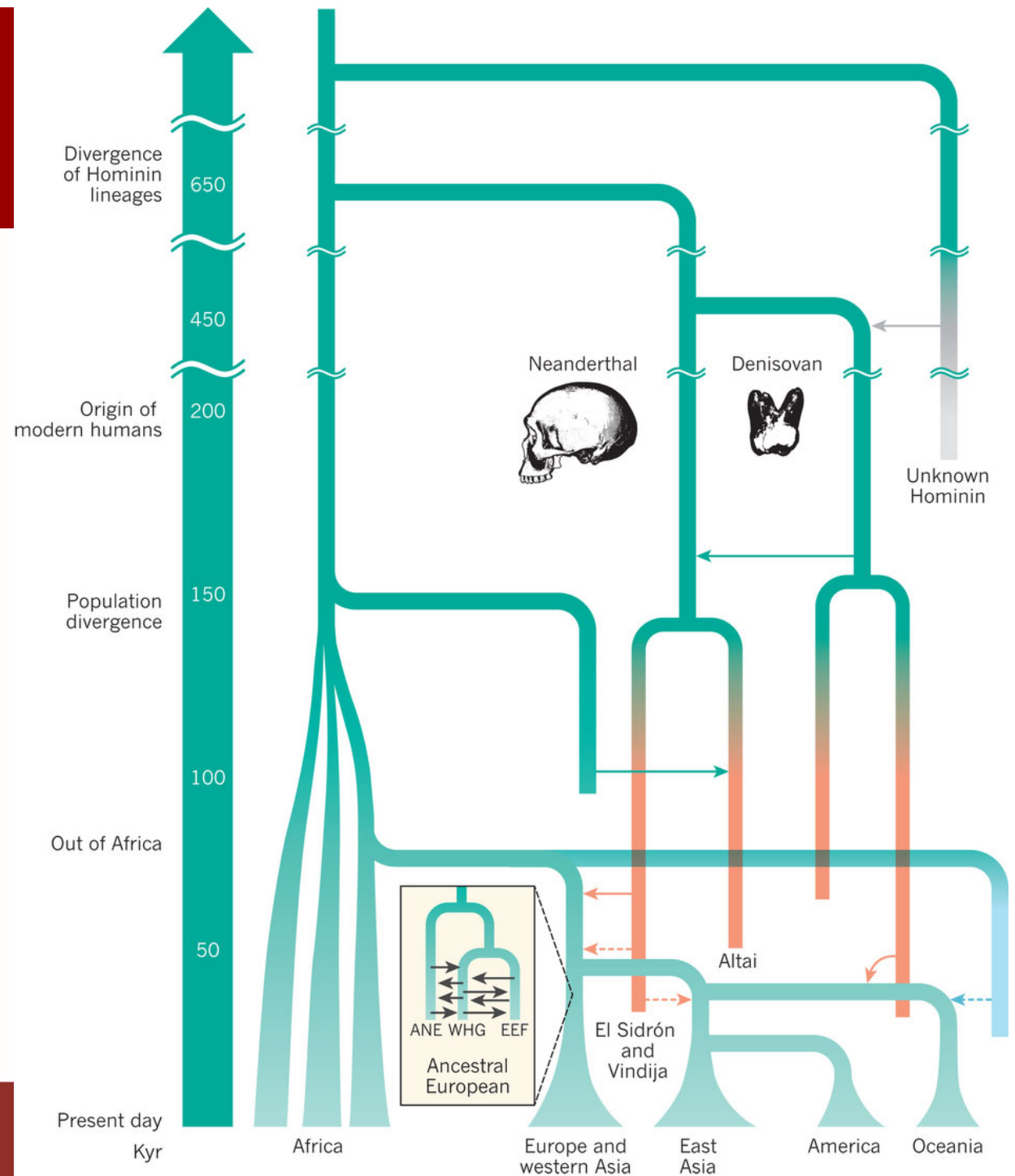
Department of Biology, U

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

**Have we eliminated
incongruence?**

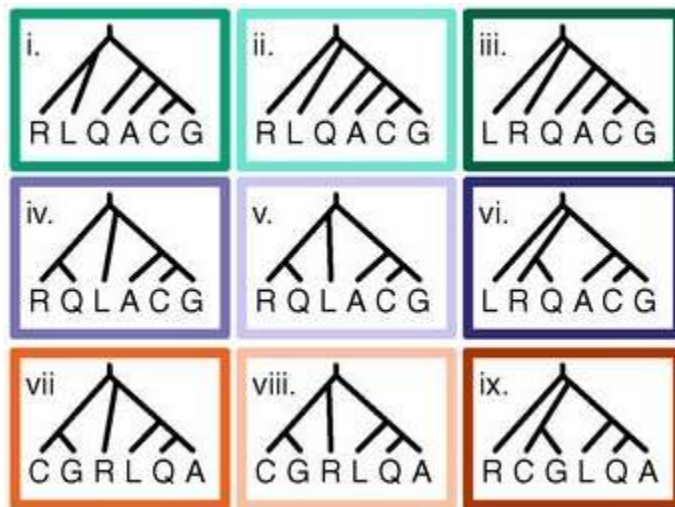
**Figuring out what's going on is
easier at shallow depths, b/c
systematic error is usually
absent**

The Evolution of Human Populations



Nielsen et al. (2017)
Nature

“Easier” Doesn’t Mean “Easy”!



Fontaine *et al.* adhere to a classical view that there is a “true species tree” [...]. But given that the bulk of the genome has a network of relationships that is different from this true species tree, **perhaps we should dispense with the tree and acknowledge that these genomes are best described by a network, and that they undergo rampant reticulate evolution**



The Phylogeny of Primate Genera

*Nomascus
leucogenys*



NLE

*Hoolock
leuconedys*



HLE

*Symphalangus
syndactylus*



SSY

*Hylobates
pileatus*



HPI

*Hylobates
moloch*

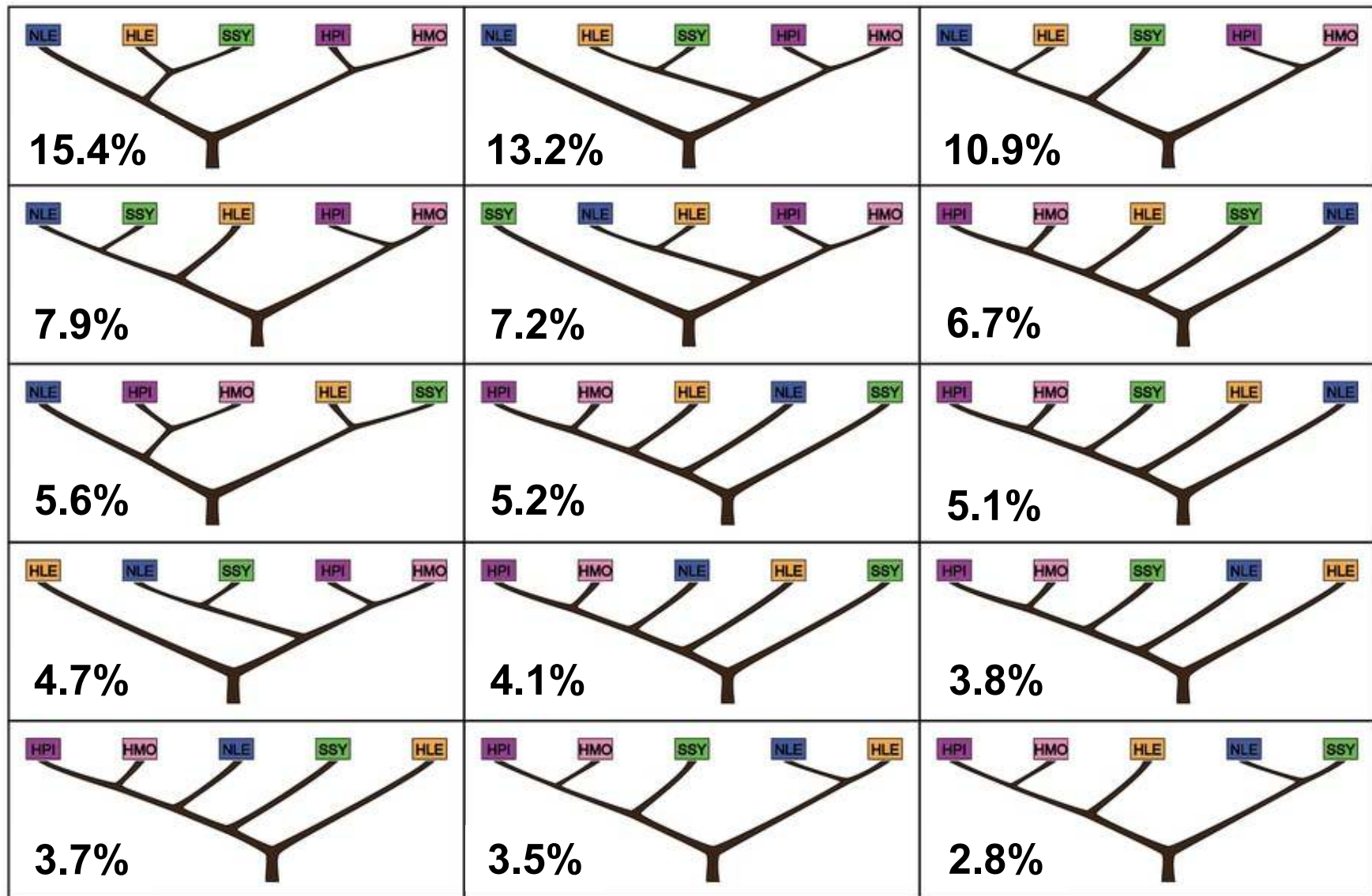


HMO

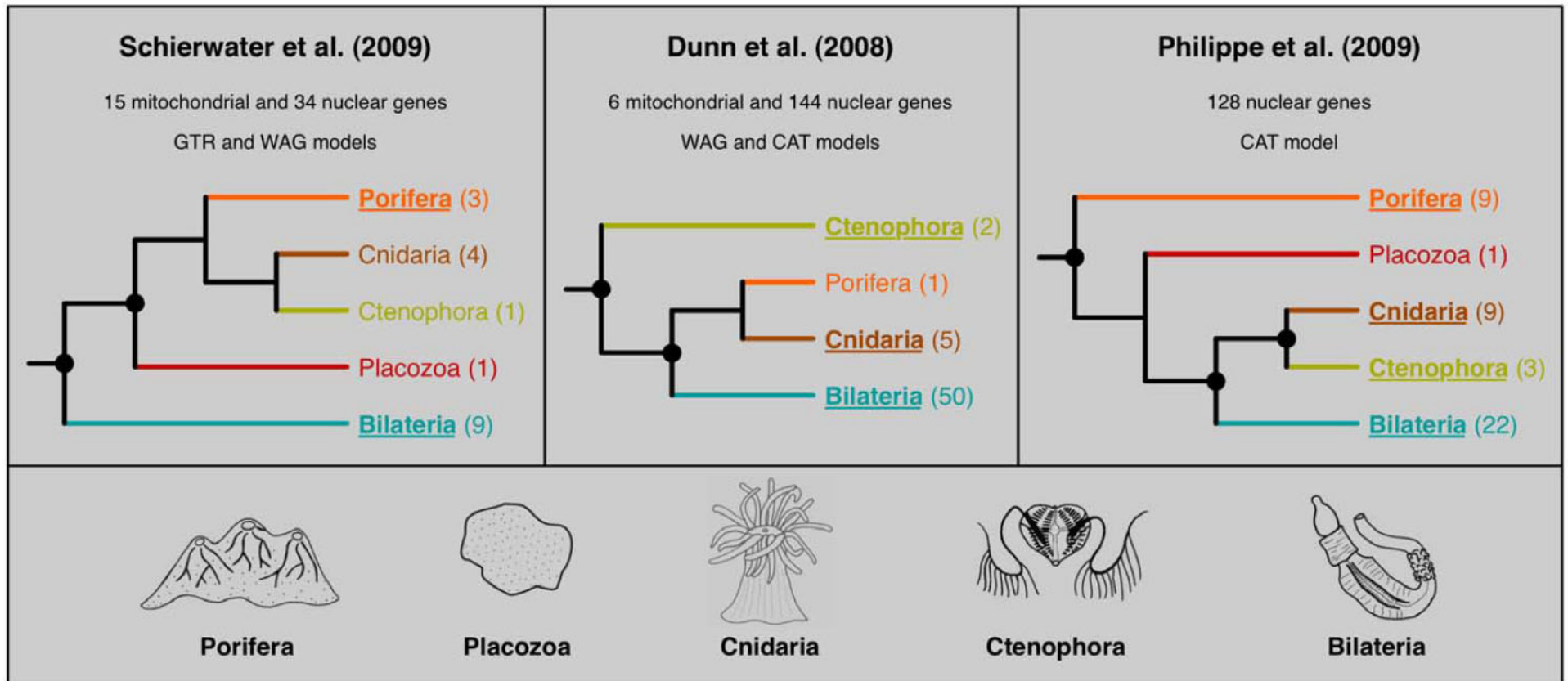


Carbone et al. (2014) Nature

“Easier” Doesn’t Mean “Easy”!



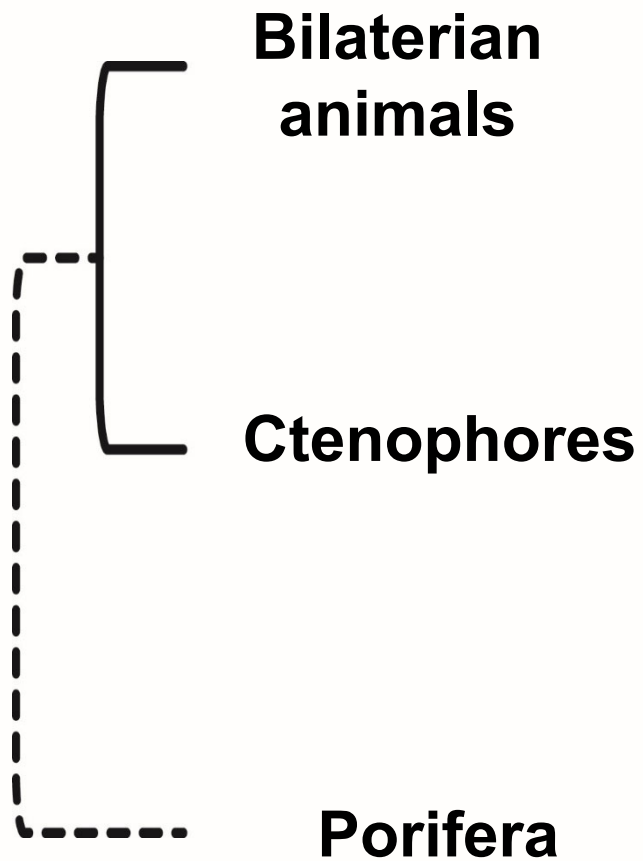
Incongruence in Deep Time is More Challenging



Incongruence in Deep Time is More Challenging

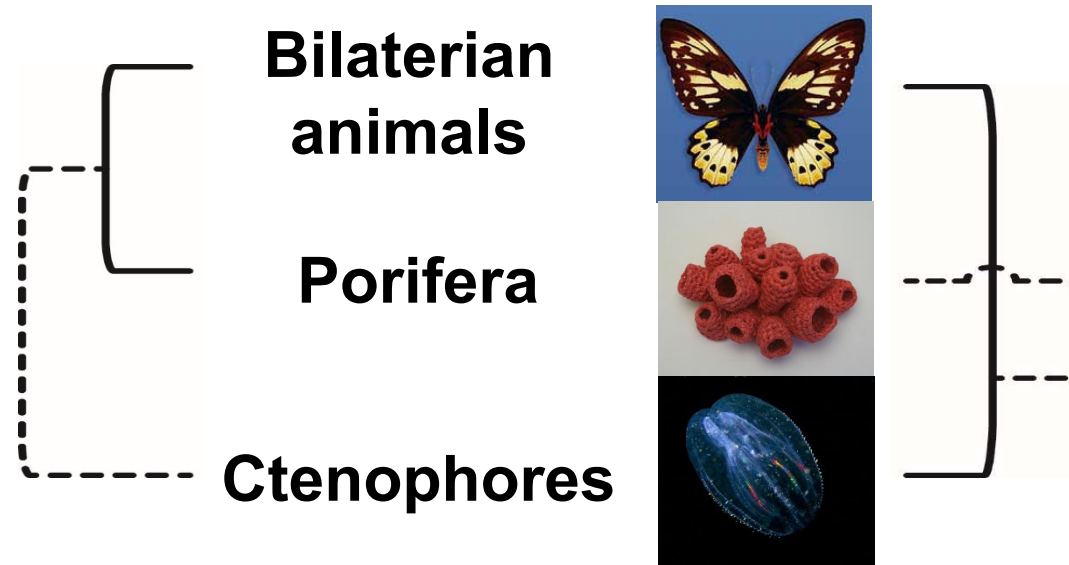


Incongruence in Deep Time is More Challenging



Why the disconnect?

Coffee Break



**Why are deep branches incongruent?
(How) can we resolve them?**



An Expanded Yeast Data Matrix

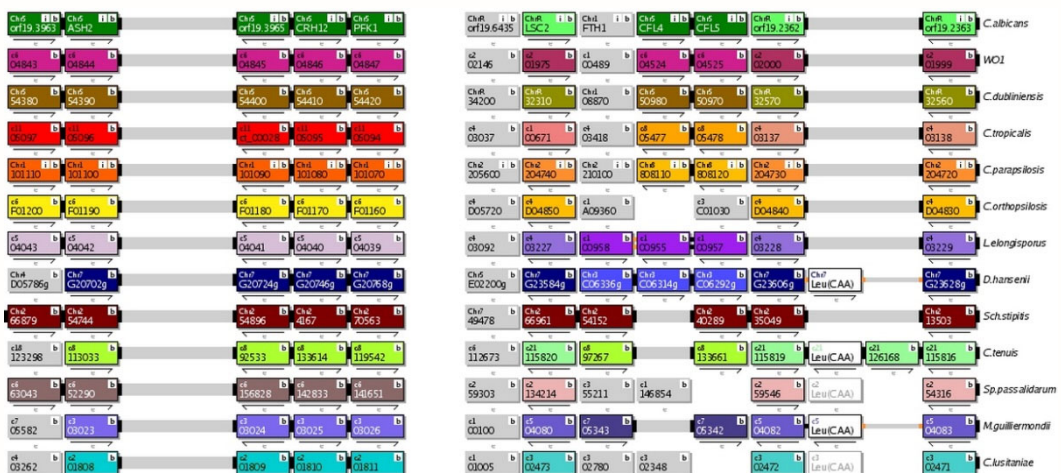
Yeast Gene Order Browser (YGOB)



**Saccharomyces
lineage**

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

Candida Gene Order Browser (CJOB)



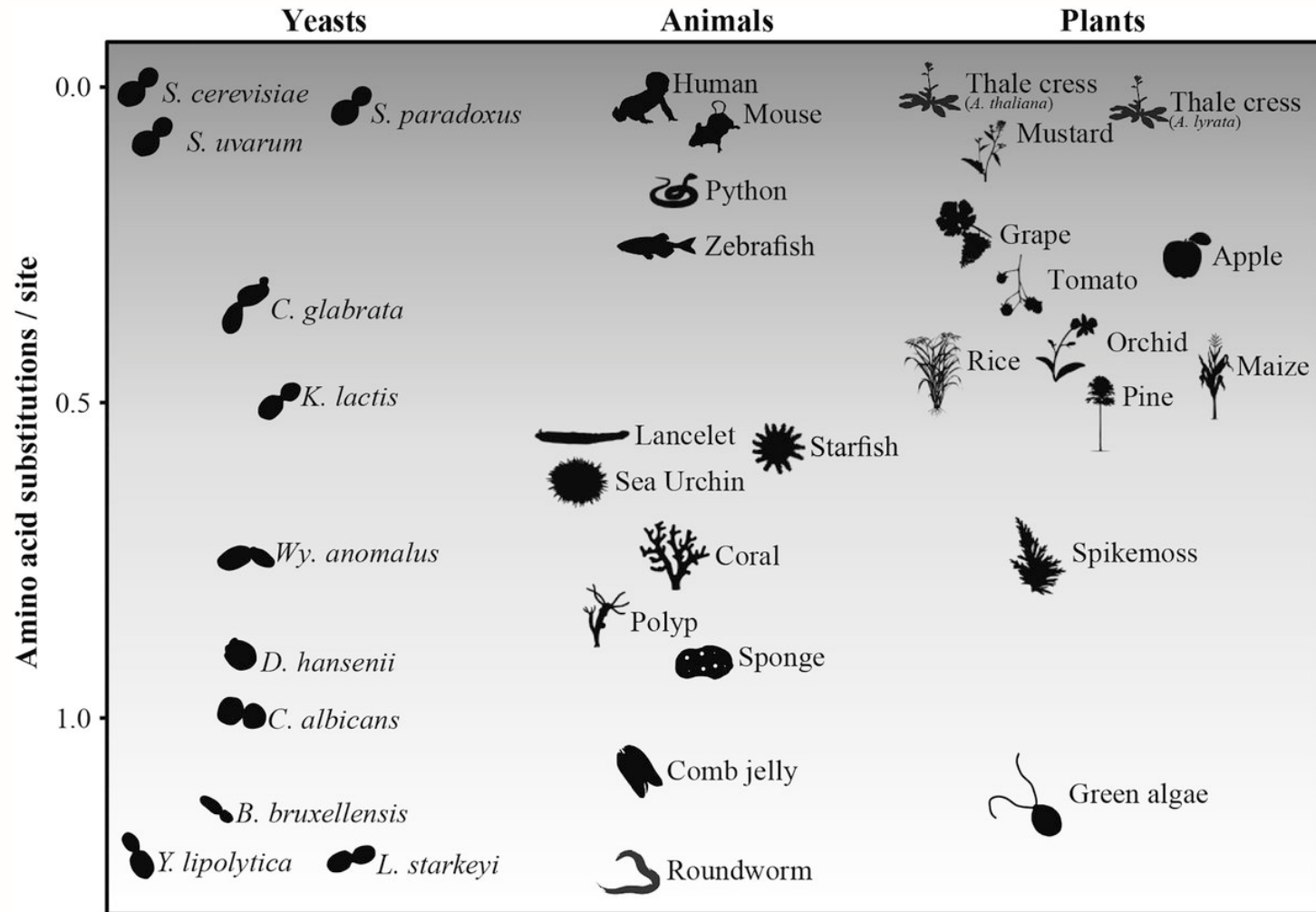
**Candida
lineage**



Byrne & Wolfe (2005) Genome Res.

Fitzpatrick et al. (2010) BMC Genom.

Budding Yeasts Exhibit Striking Genomic Diversity

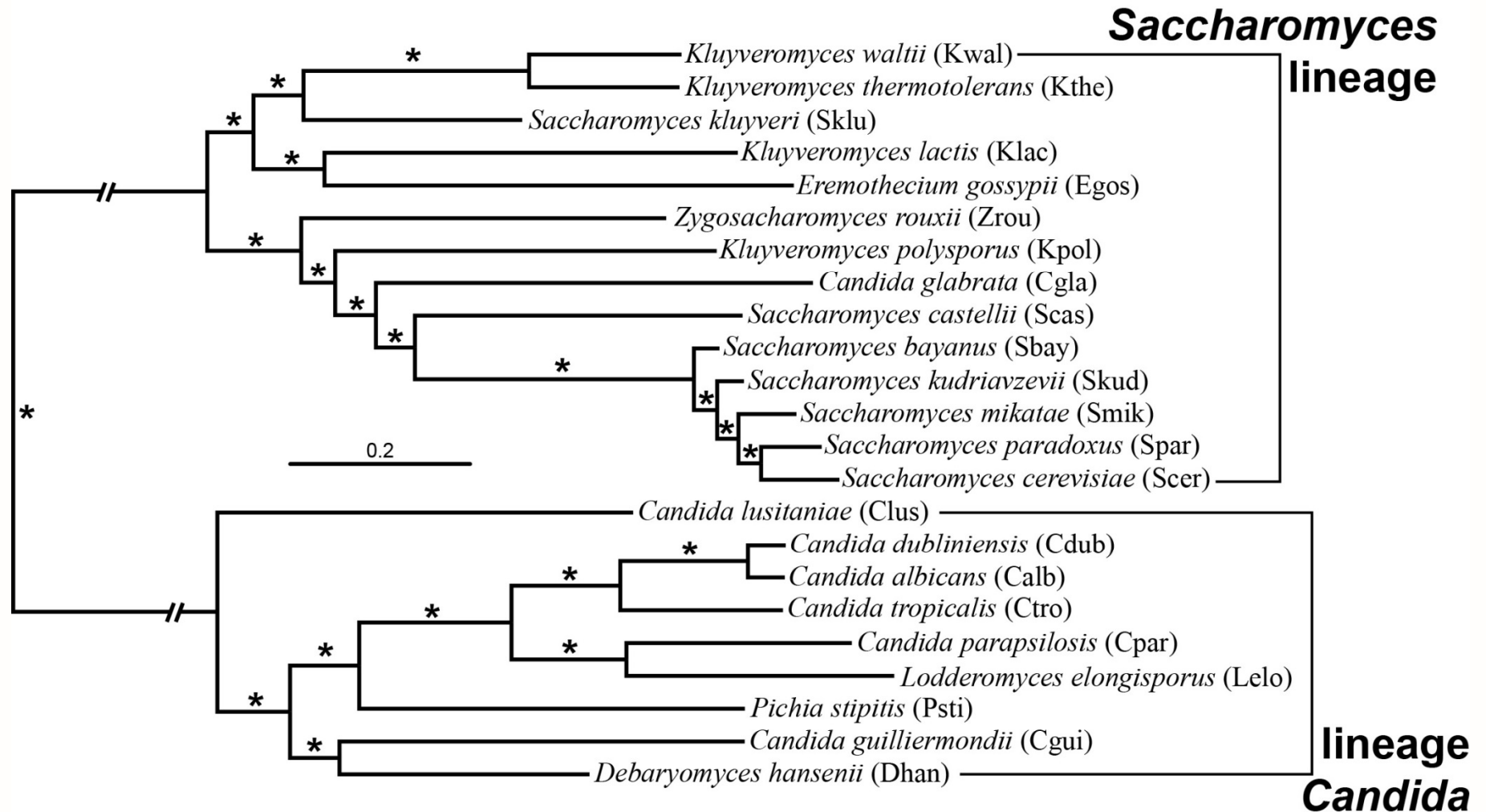


Saccharomyces, Candida, Kluyveromyces, etc. are all polyphyletic genera

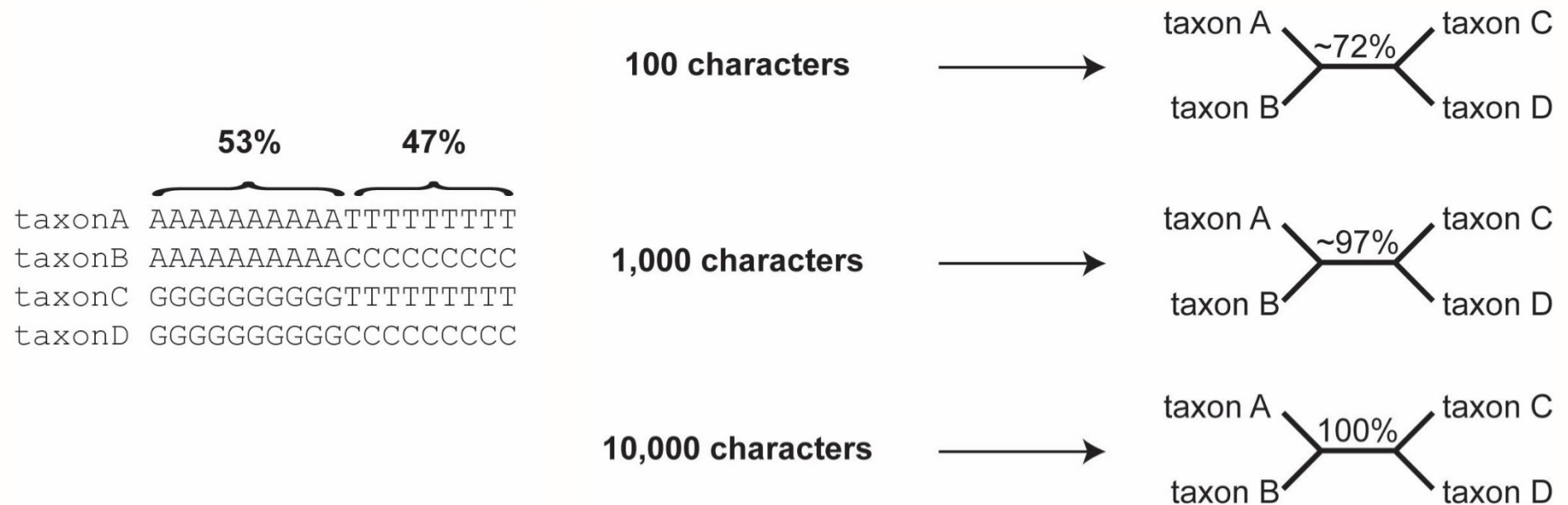


Shen, Opulente, Kominek, Zhou et al. (2018) Cell

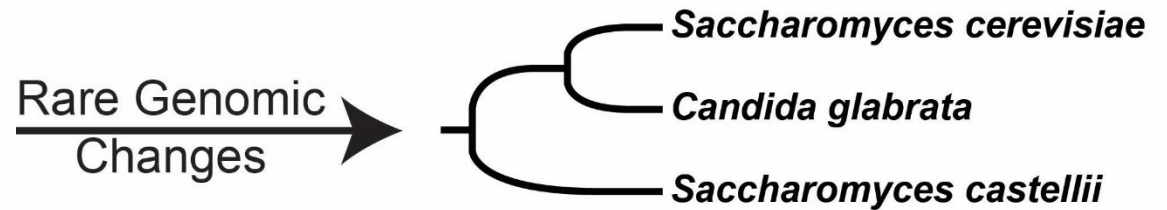
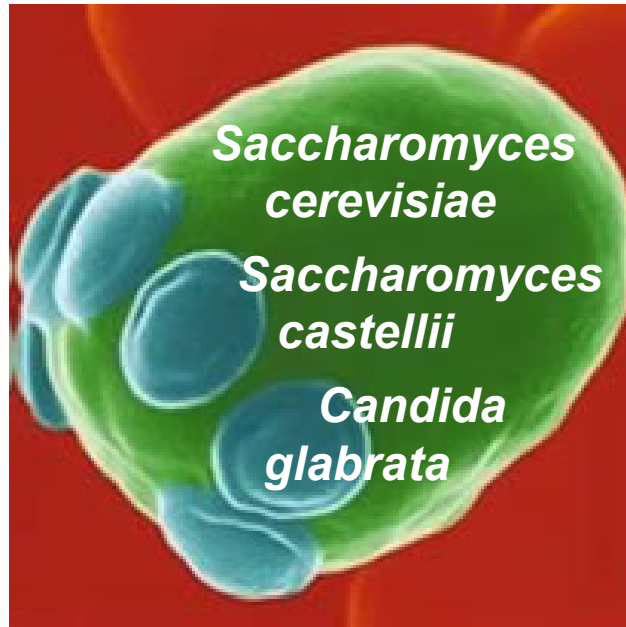
Concatenation Yields an Absolutely Supported Phylogeny



Bootstrap Support is Misleading When Used in Large Datasets

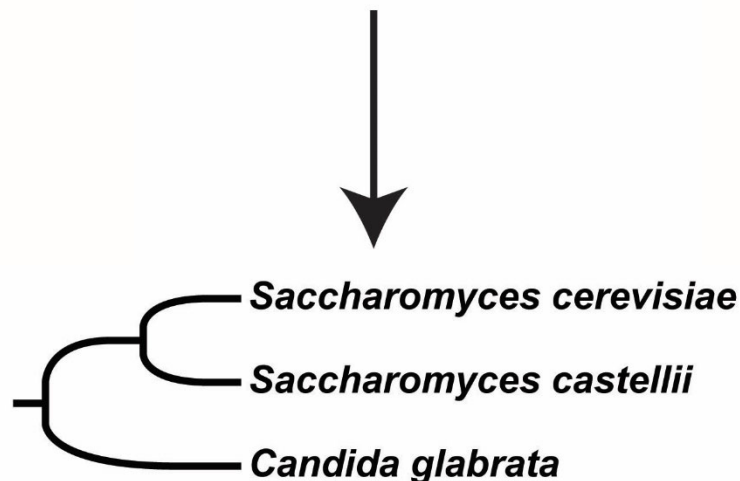


The Concatenation Phylogeny is at Least Partly Wrong

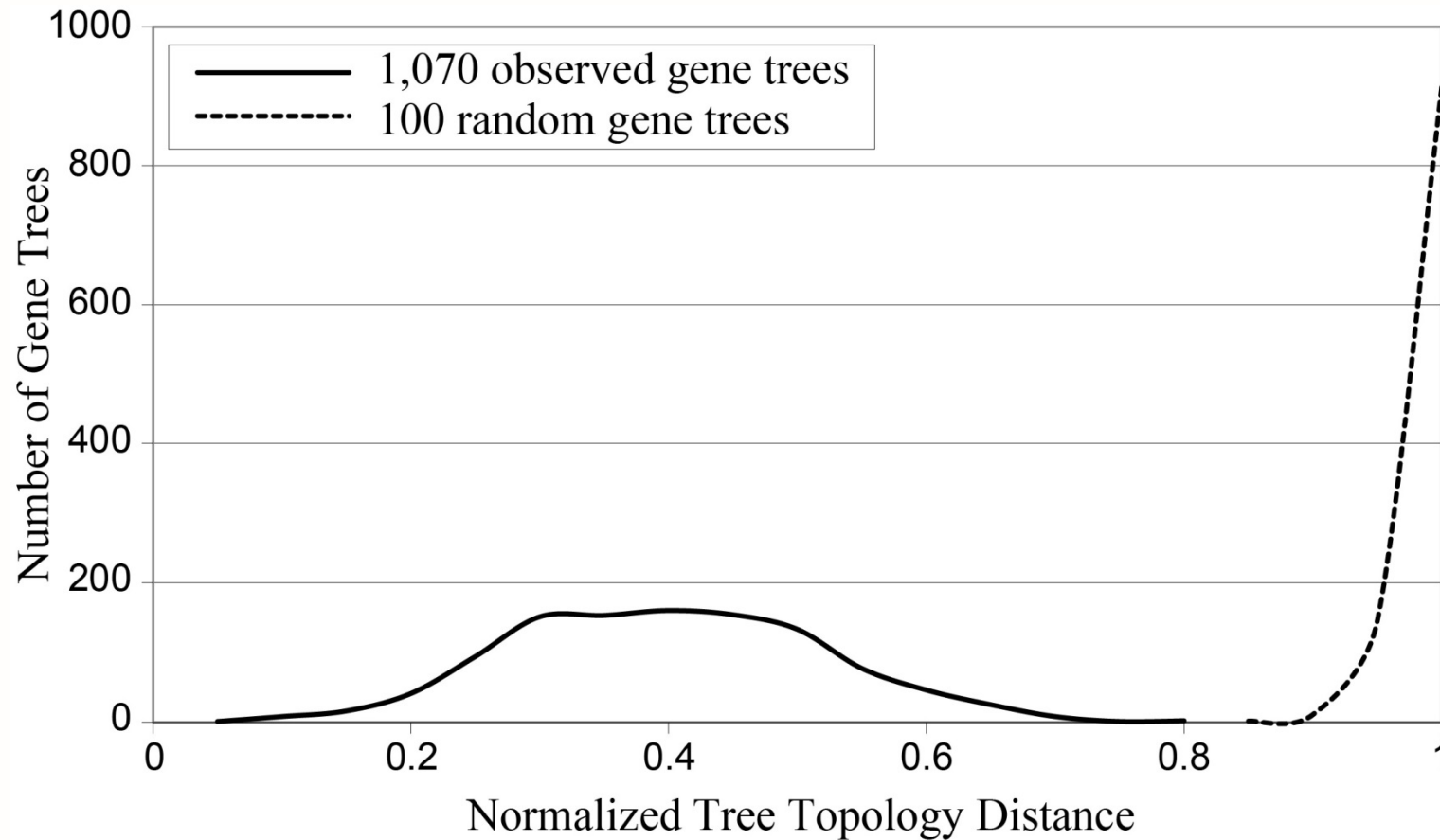


- ❖ 5 genomic rearrangements that are uniquely shared by *S. cerevisiae* and *C. glabrata*
- ❖ Much higher proportion of shared gene losses in *S. cerevisiae* and *C. glabrata*
- ❖ Bias in the placement of *C. glabrata* as an outgroup of *S. cerevisiae* and *S. castellii*

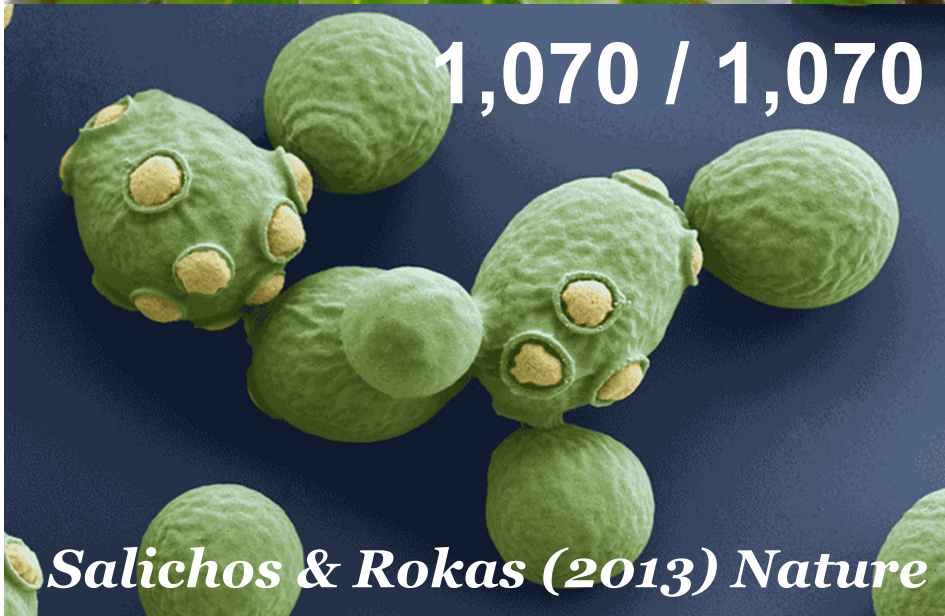
Linear Sequence Data



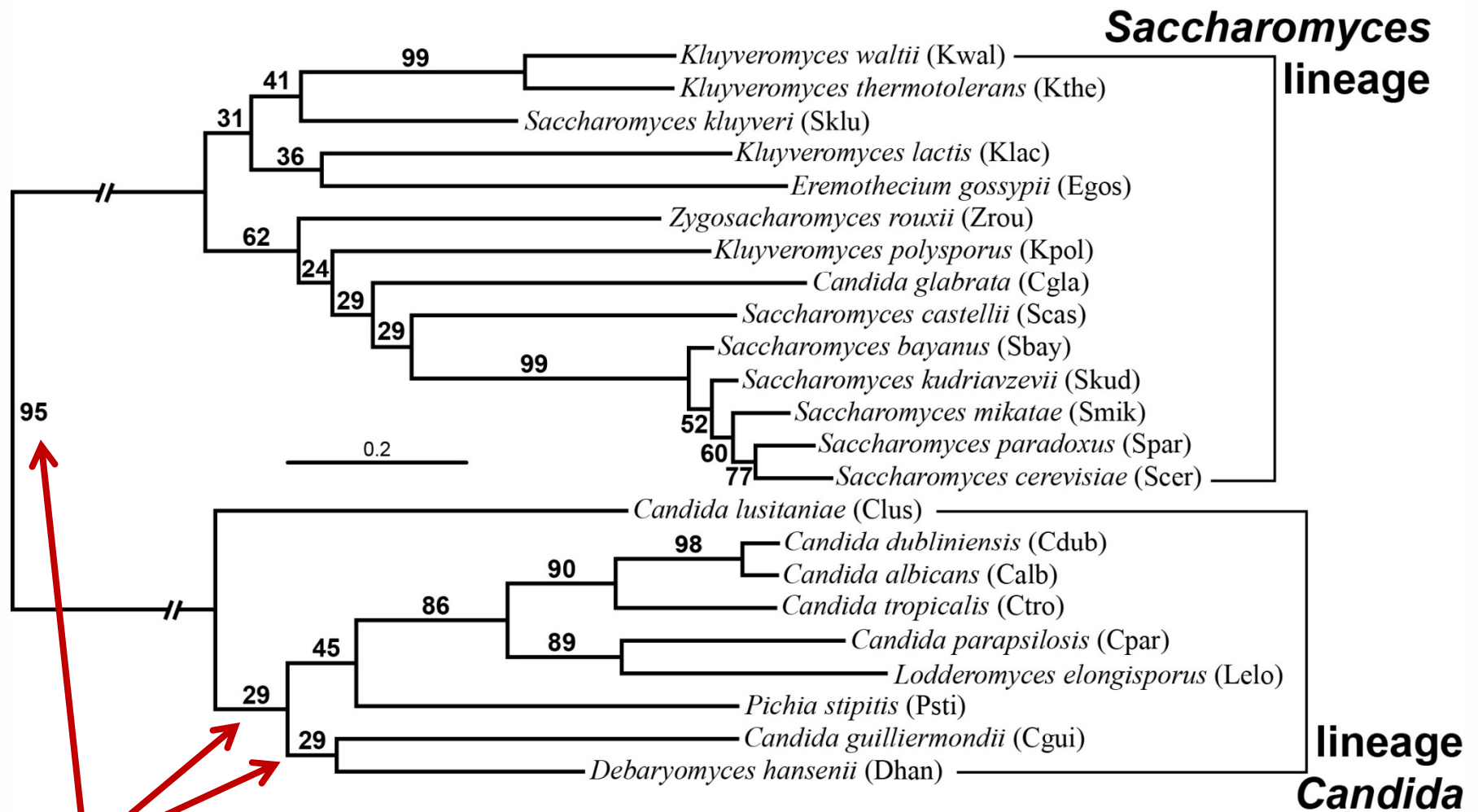
All Gene Trees Differ from the Concatenation Phylogeny



Gene Trees are Incongruent in Most Datasets



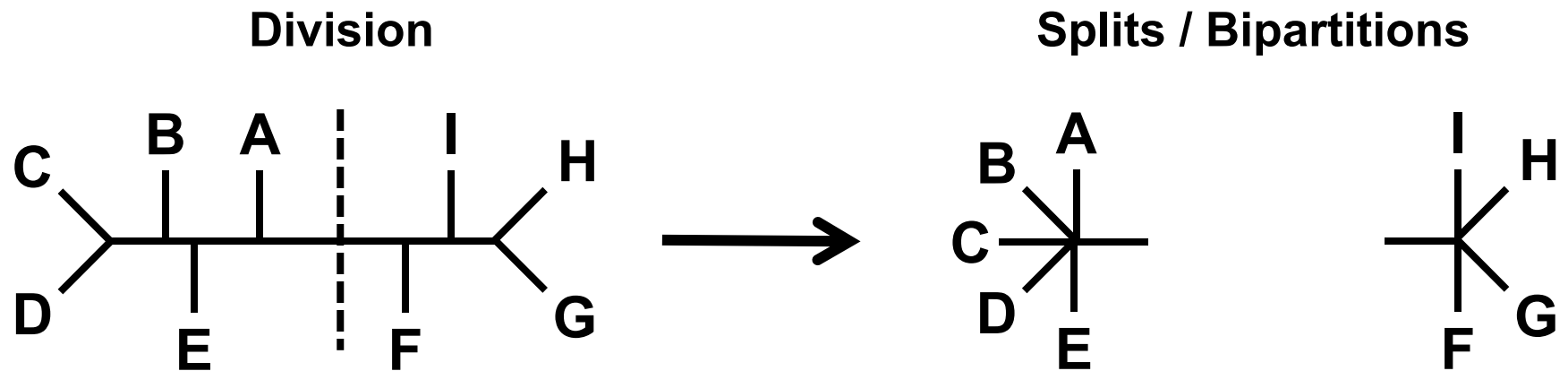
The Yeast Phylogeny Inferred by Majority-Rule Consensus



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

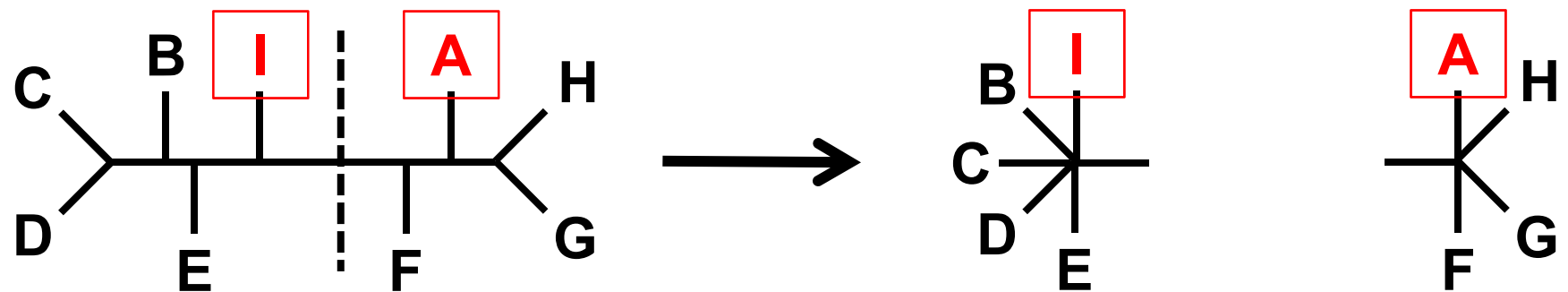


Phylogenetic Trees are Sets of Splits / Bipartitions



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

{F, G, H, I}



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

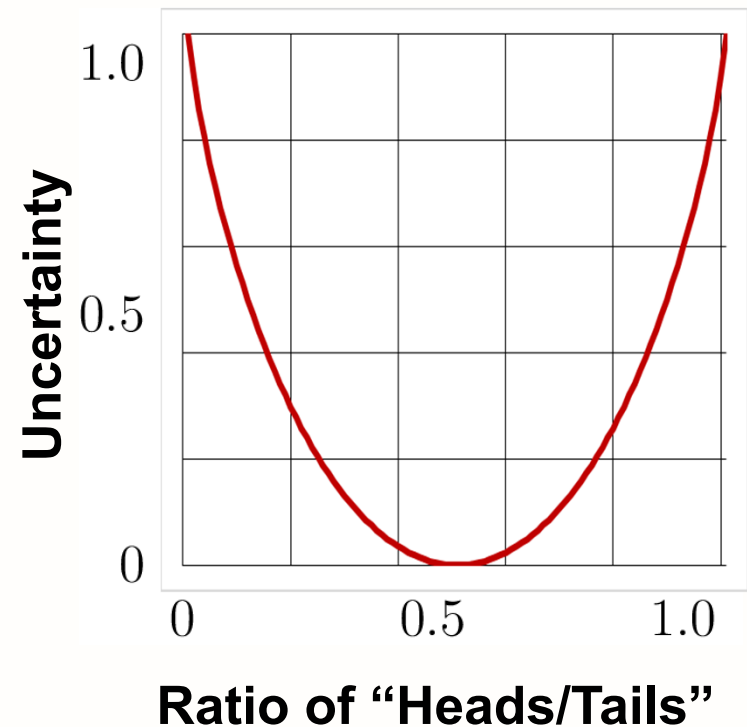
{F, G, H, A}

Quantifying Incongruence

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

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

IC and TC are implemented in the latest versions of RAxML

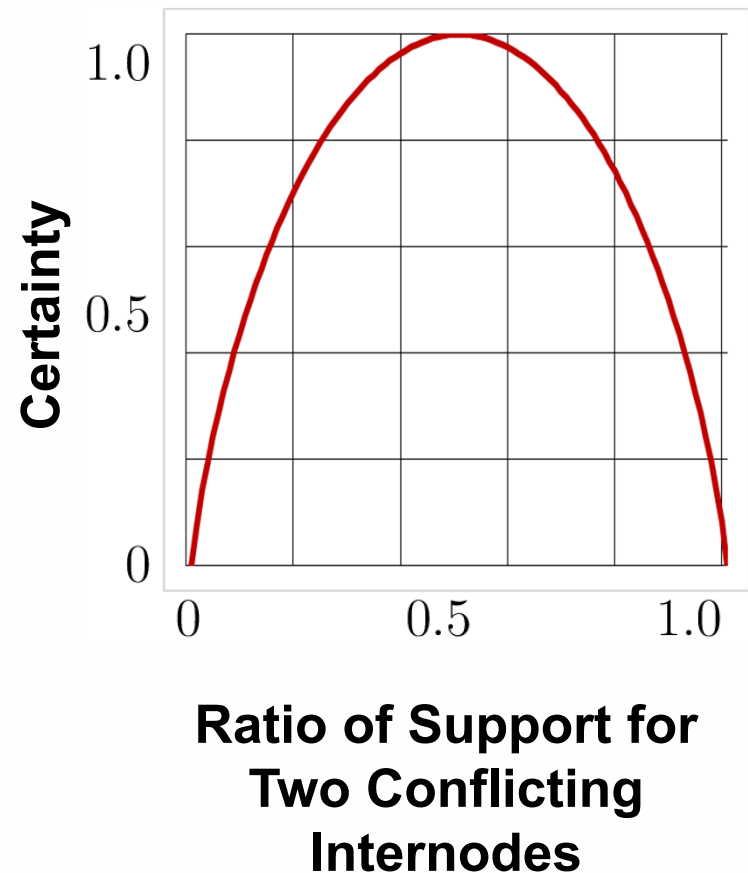


Quantifying Incongruence

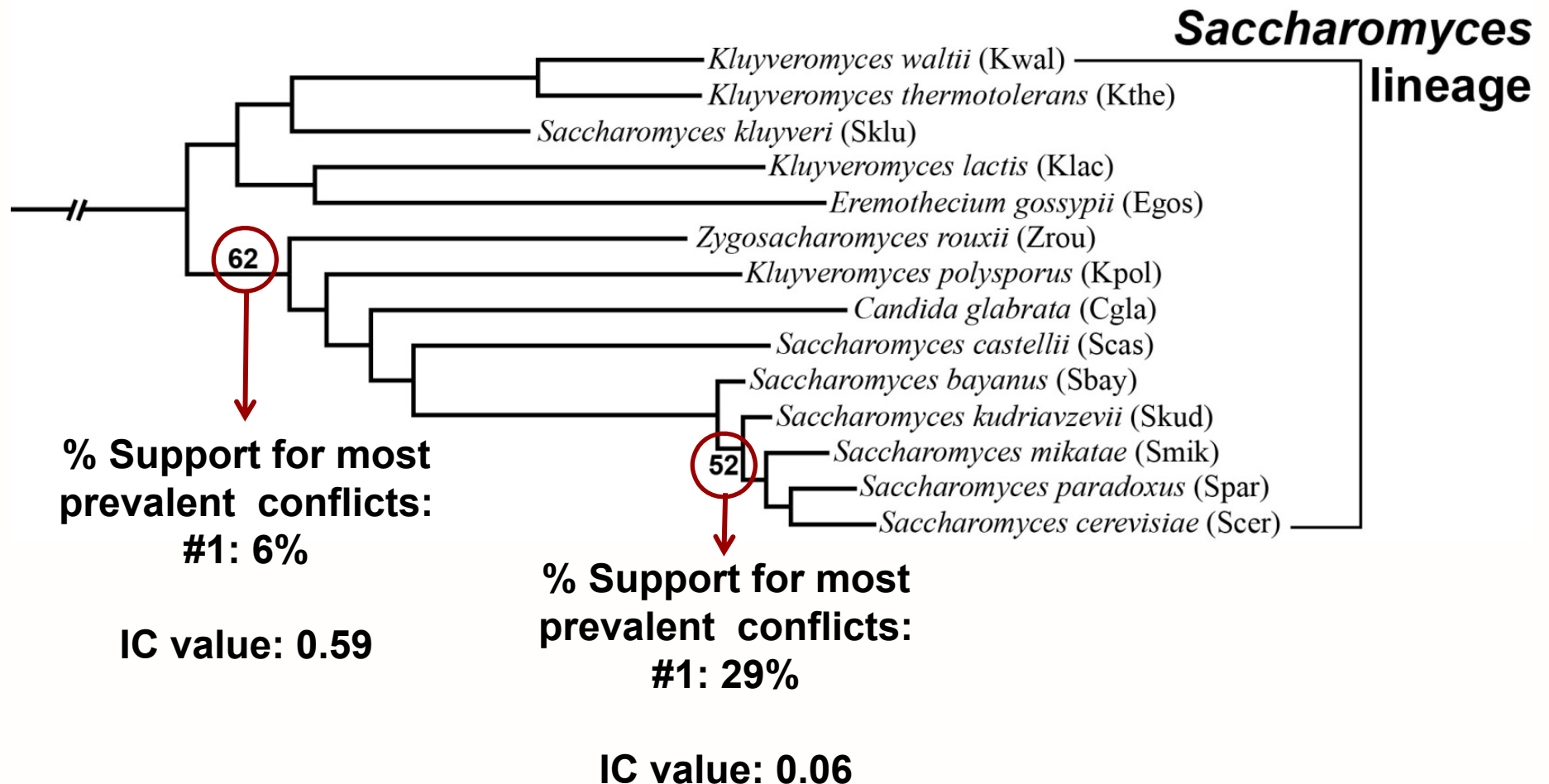
Internode Certainty (IC): a measure of the support for a given internode by considering its frequency in a given set of trees jointly with that of the most prevalent conflicting internode in the same set of trees

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

IC and TC are implemented in the latest versions of RAxML

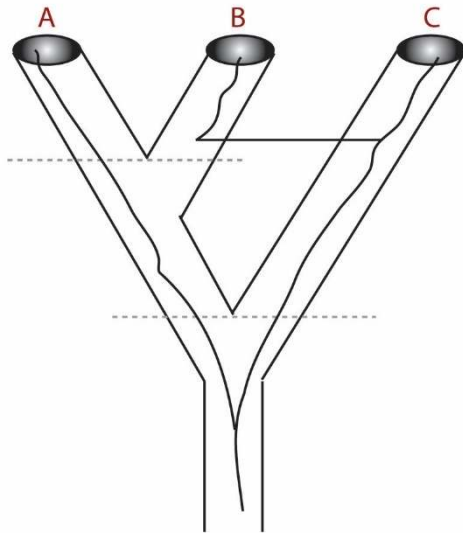


IC Can Be More Informative Measure of Internode Support

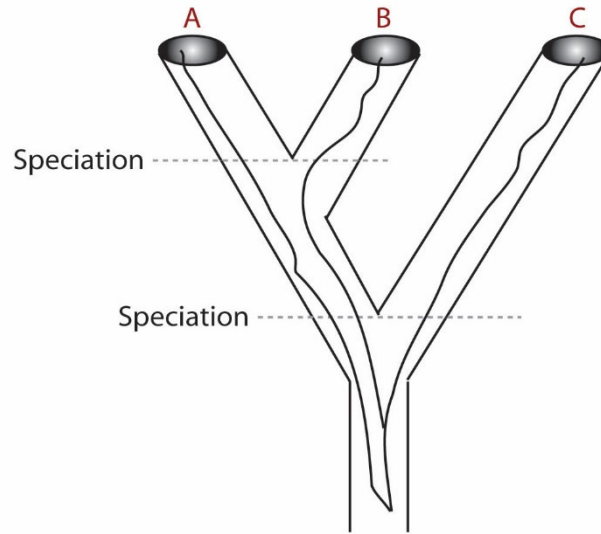


Why So Much Incongruence? Biological Factors

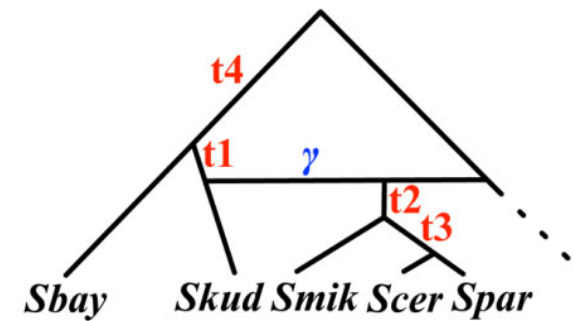
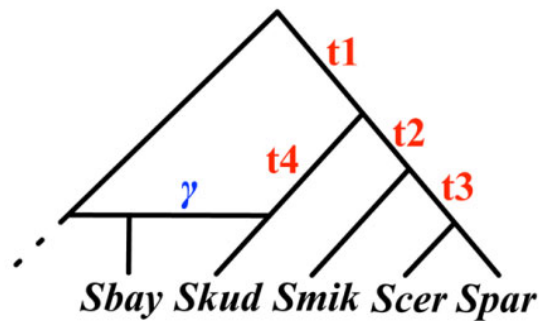
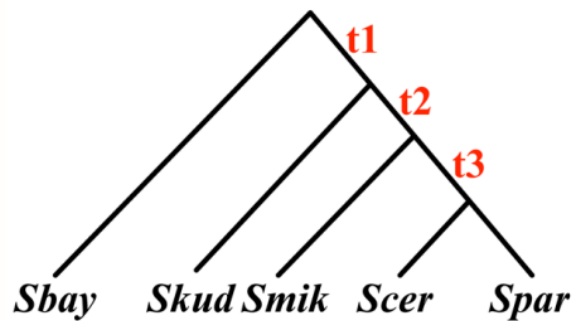
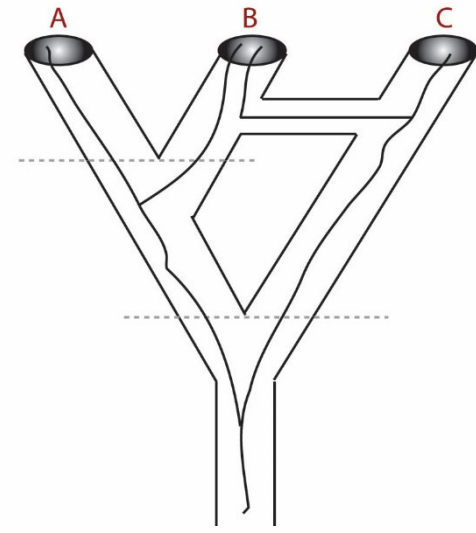
Horizontal Gene Transfer



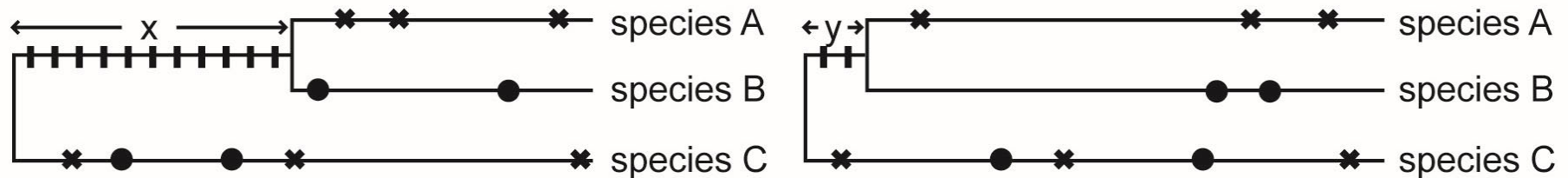
Lineage Sorting



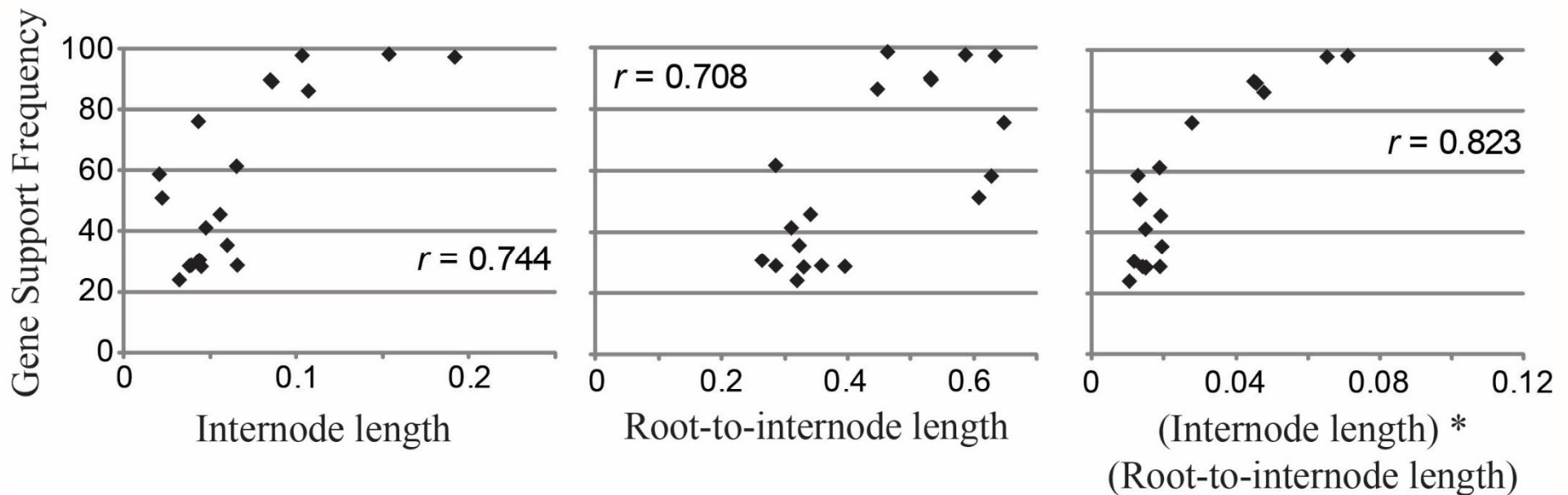
Hybridization



Why So Much Incongruence? Analytical Factors



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





Standard Recipes for Handling Incongruence Didn't Help

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

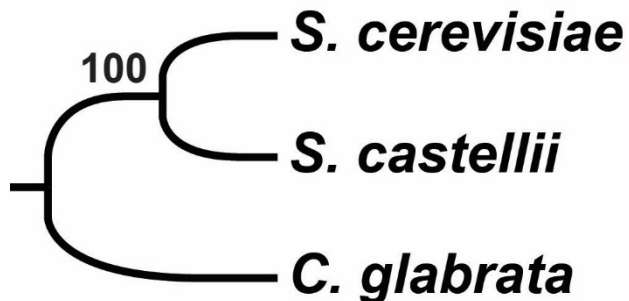




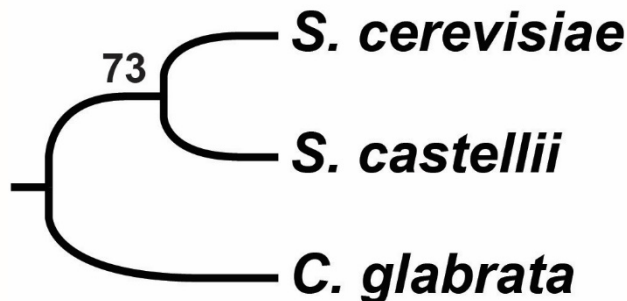
What Do We Do Then?

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

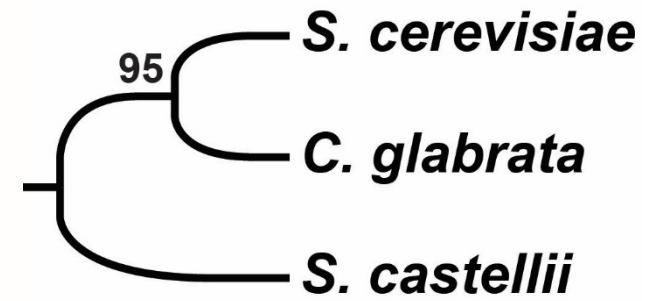
average BS $\geq 60\%$



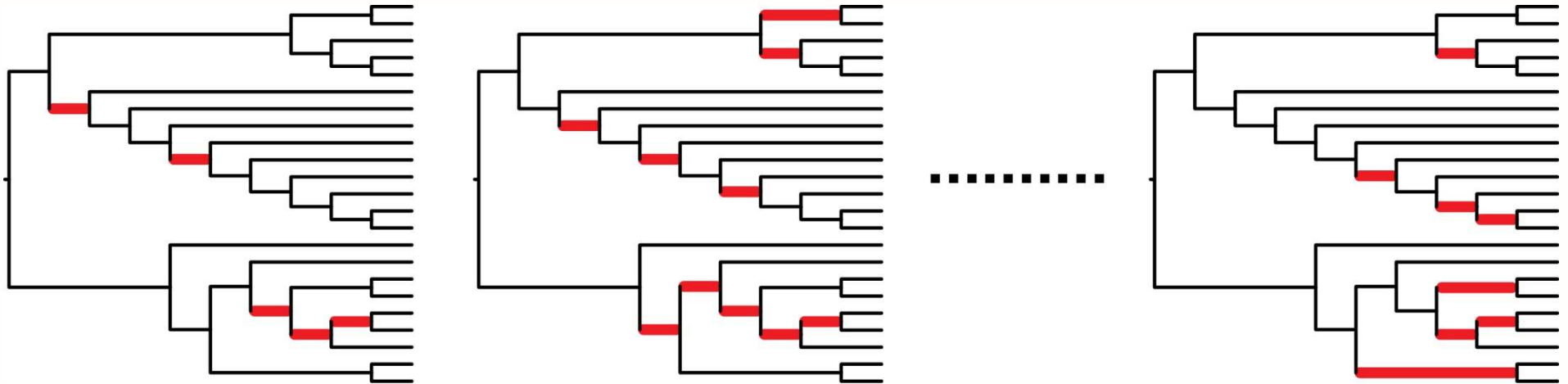
average BS $\geq 70\%$



average BS $\geq 80\%$



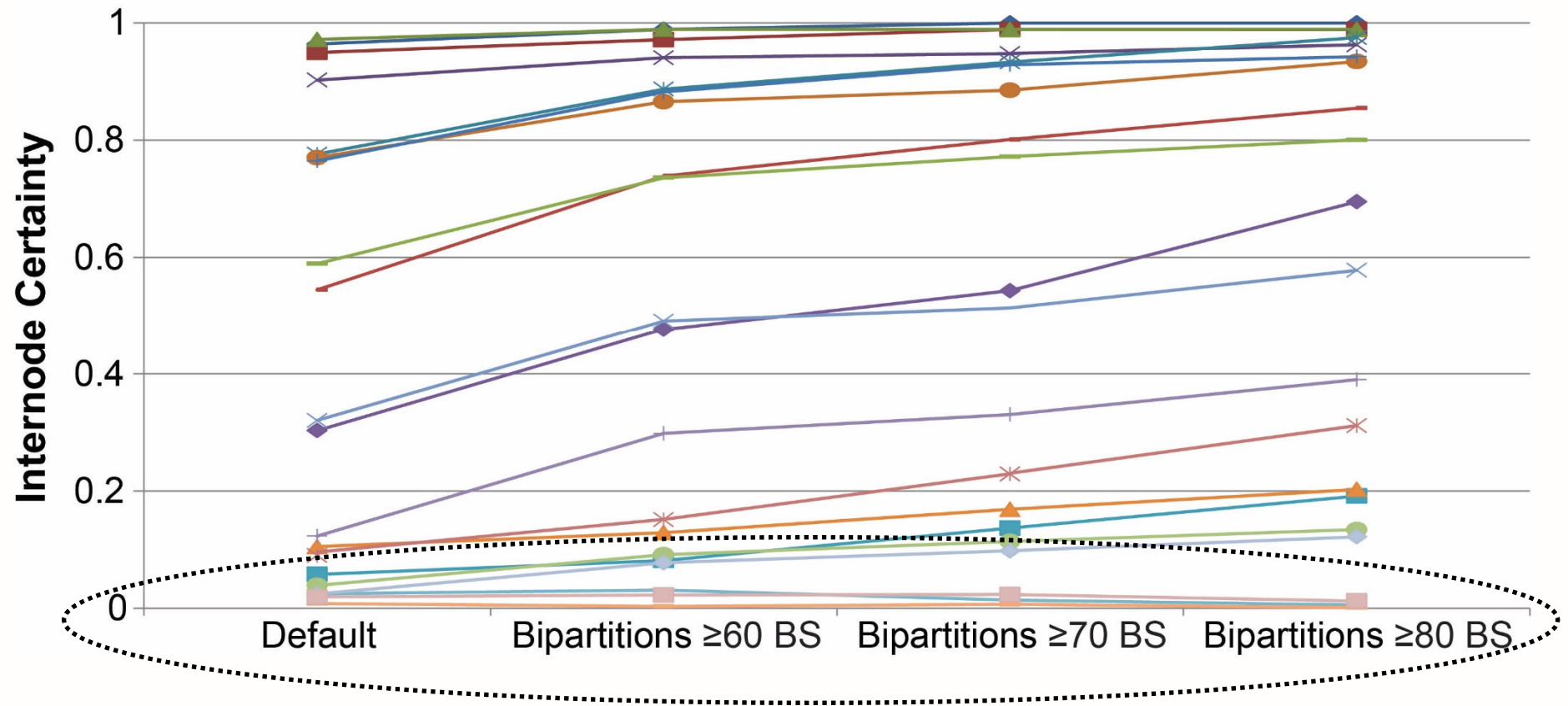
Selecting Specific Bipartitions Dramatically Improves Phylogeny



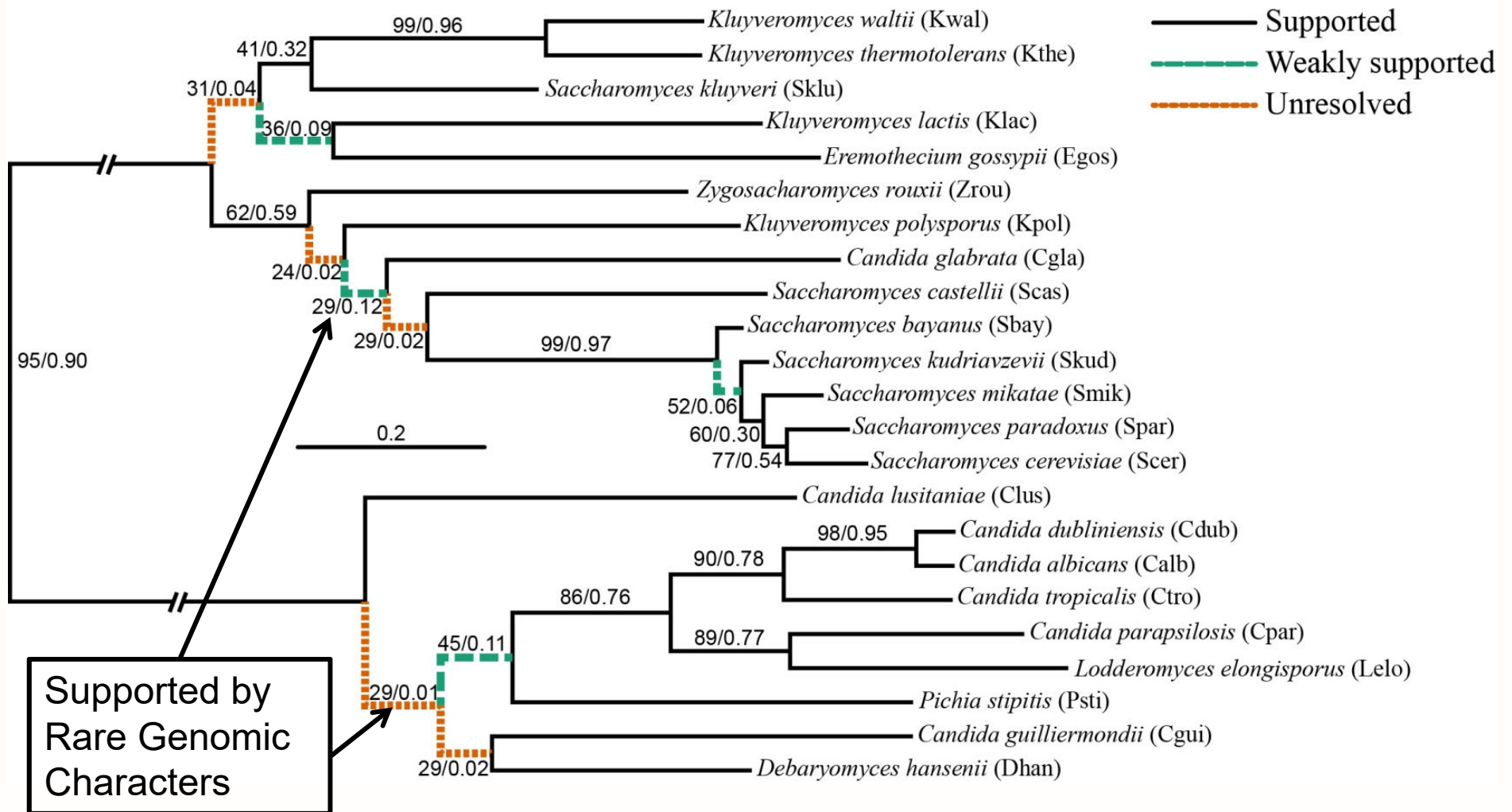
Treatment	Tree Certainty	# of Internodes where IC increased decreased
Default analysis	8.35	n/a
<i>Selecting genes whose bootstrap consensus trees have high average support</i>		
All bipartitions with BS $\geq$ 60%	10.11	14 0
All bipartitions with BS $\geq$ 70%	10.70	16 0
All bipartitions with BS $\geq$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



Least Supported Internodes Harbor the Most Conflict



The Status of the Yeast Phylogeny



Similar Results in Other Lineages

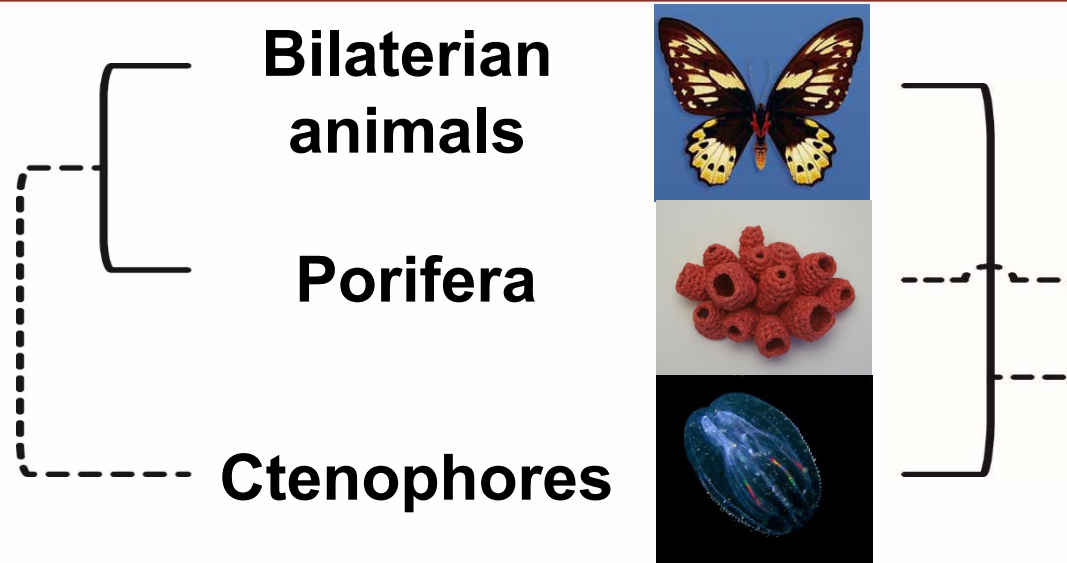
Vertebrates
(1,086 genes, 18 taxa)

Animals
(225 genes, 21 taxa)

Mosquitoes
(2,007 genes, 20 taxa)



Incongruence in Phylogenomic Datasets



These debates concern internodes that are poorly supported by individual gene trees



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

Definitions of Phylogenetic Signal

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

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

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

Sanderson (2008) Science

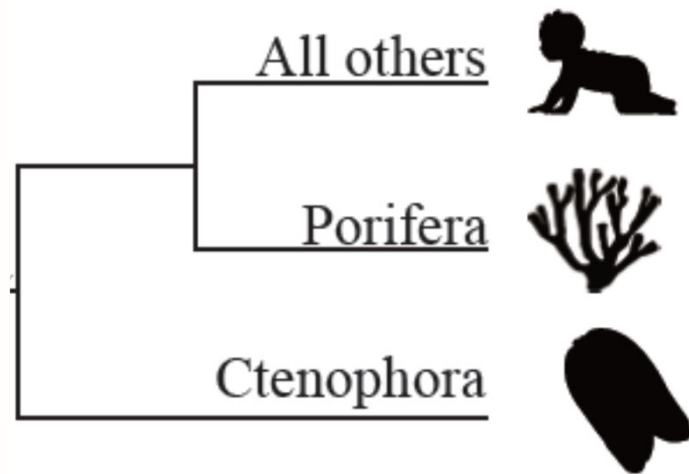
A measure of the substitutions occurring along a given branch of the evolutionary tree. In parsimony methods, the signal is encoded in shared derived characters. In probabilistic methods, the amount of phylogenetic signal actually extracted from a given dataset depends on the model and is expected to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

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

Our Definition

Maximum Likelihood tree

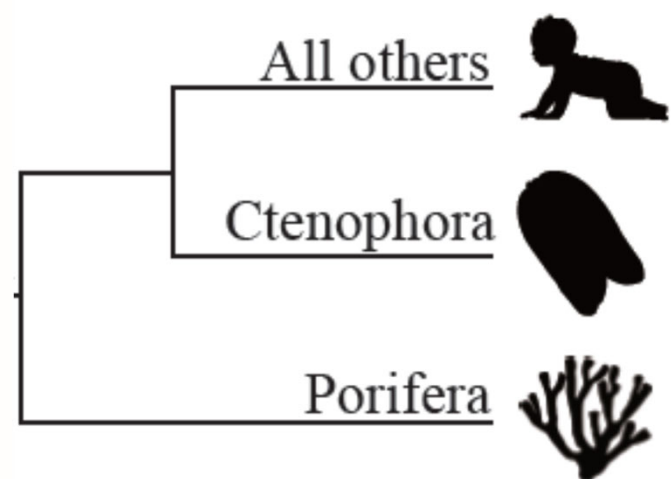
(T1)



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

Conflicting tree

(T2)



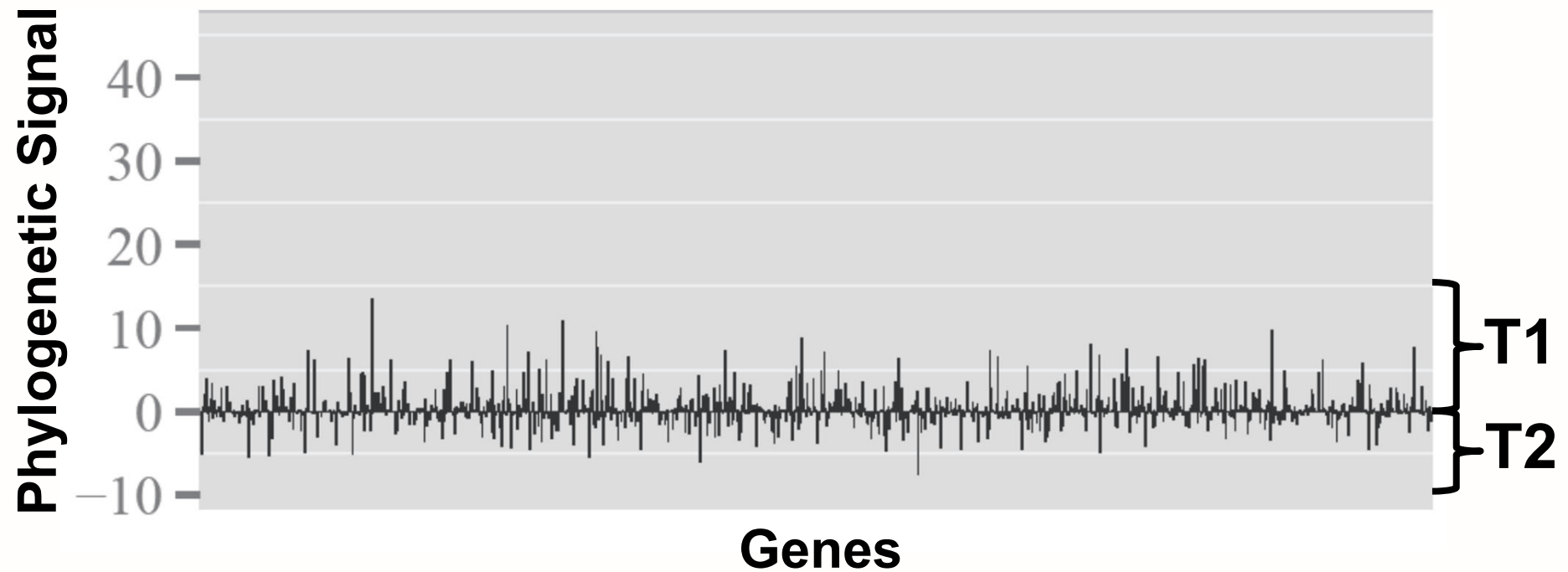
$$\ln(T_2|X_i) = -150$$

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



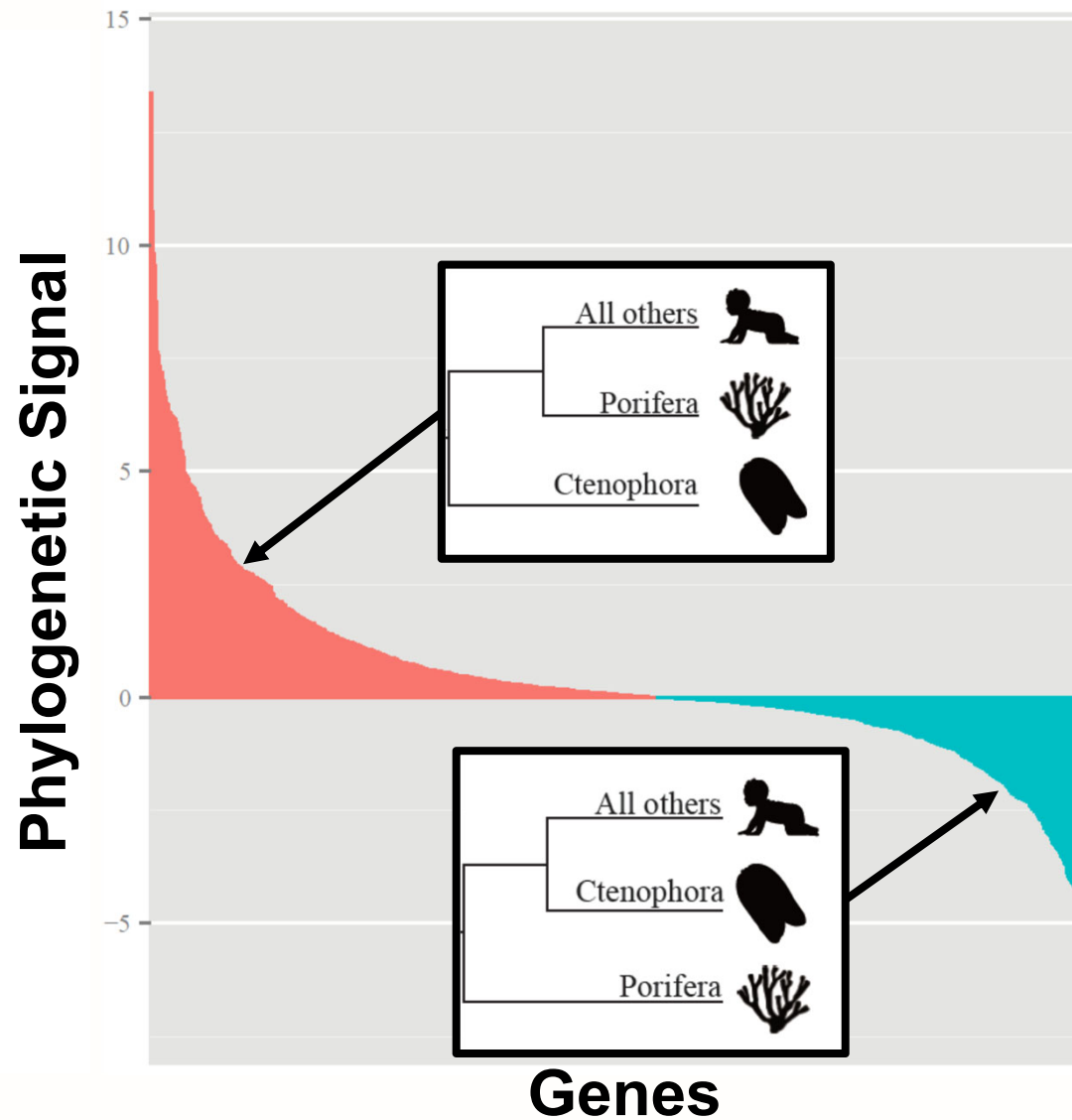
Signal of the Genes in a Phylogenomic Data Matrix

1,080 genes from 36 animal taxa

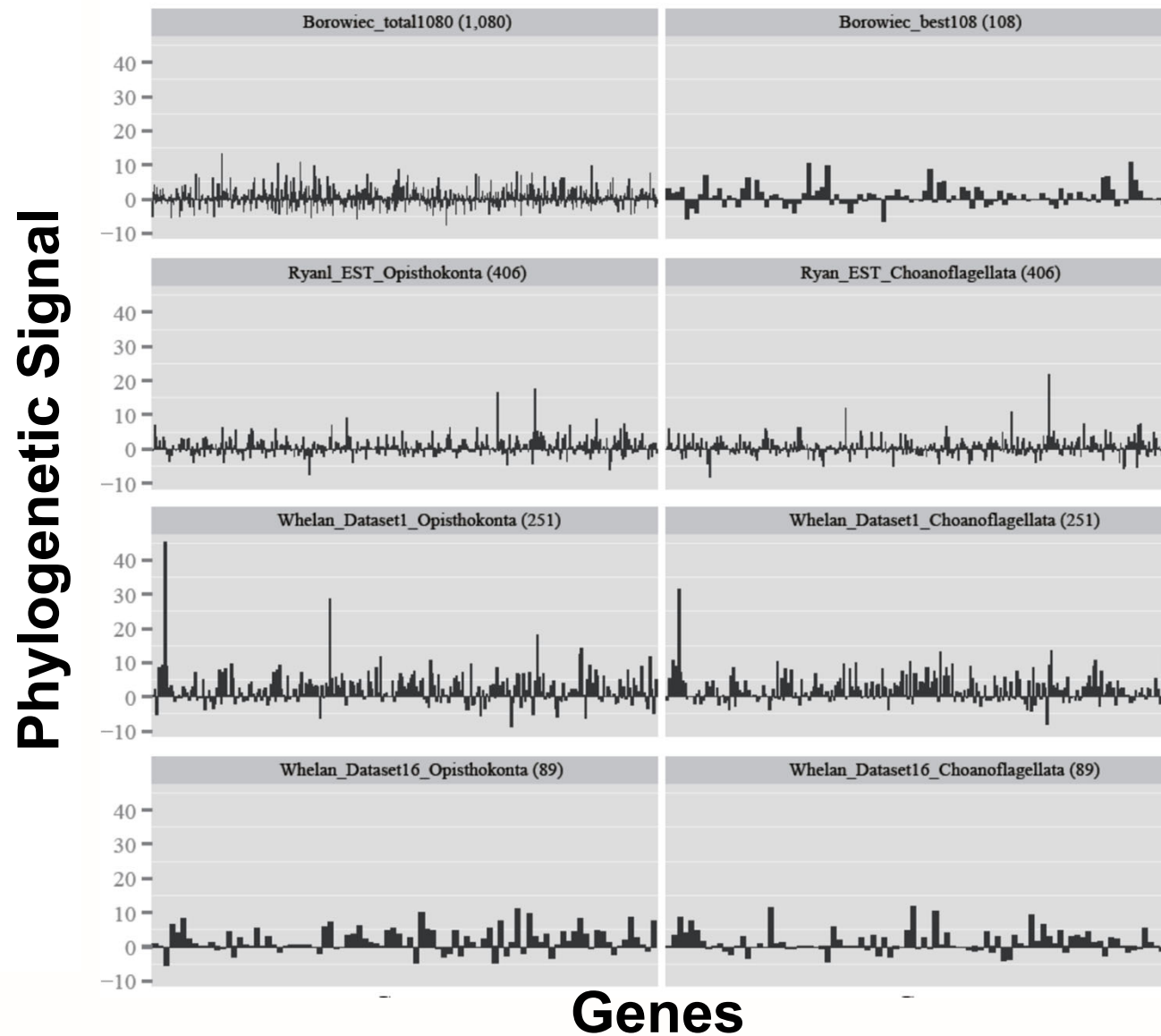


Shen et al. (2017) Nature Ecol. Evol.; data from Borowiec et al. (2015) BMC Genomics

Signal of the Genes in a Phylogenomic Data Matrix

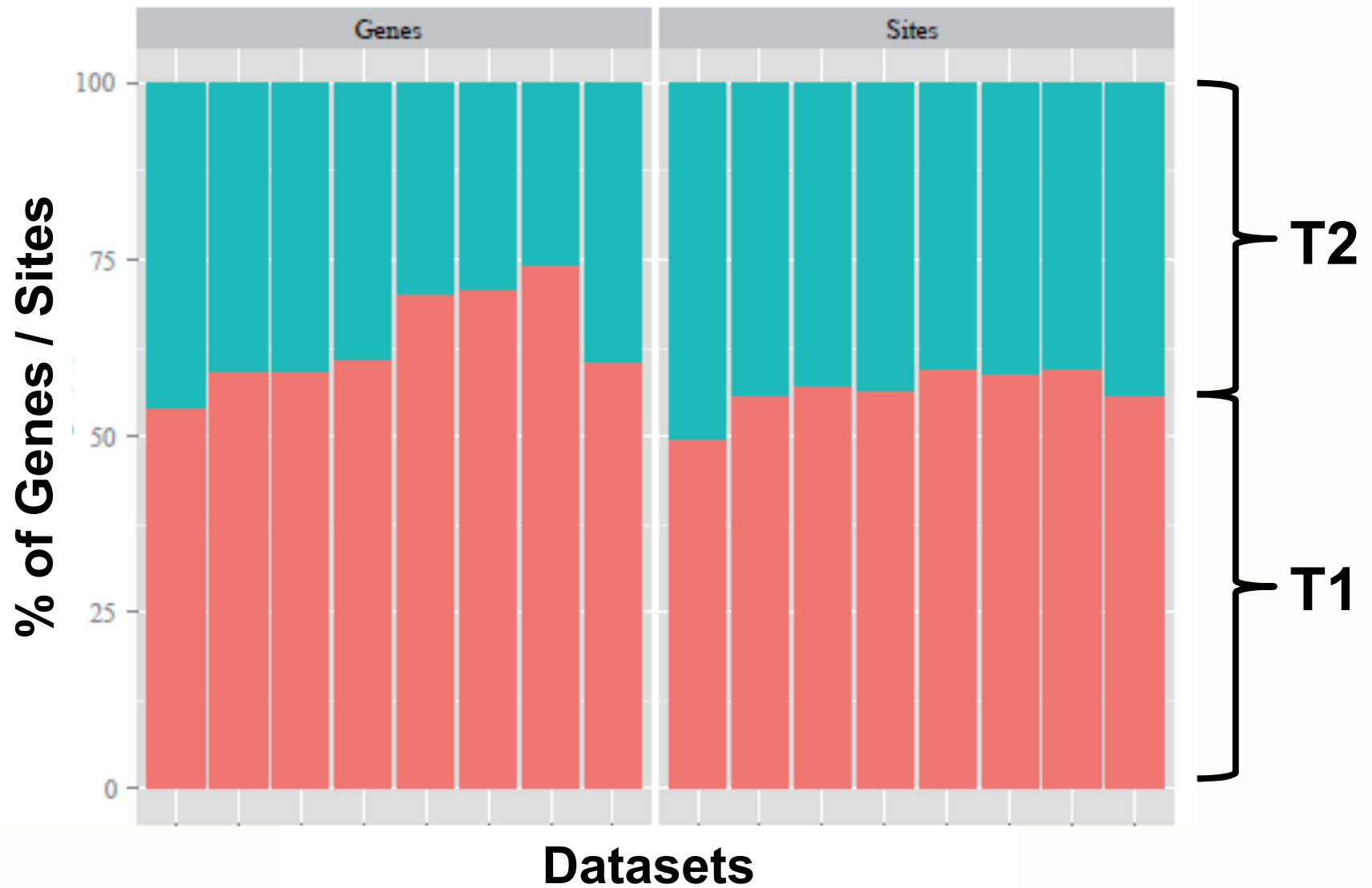


Signal of the Genes in Multiple Phylogenomic Data Matrices

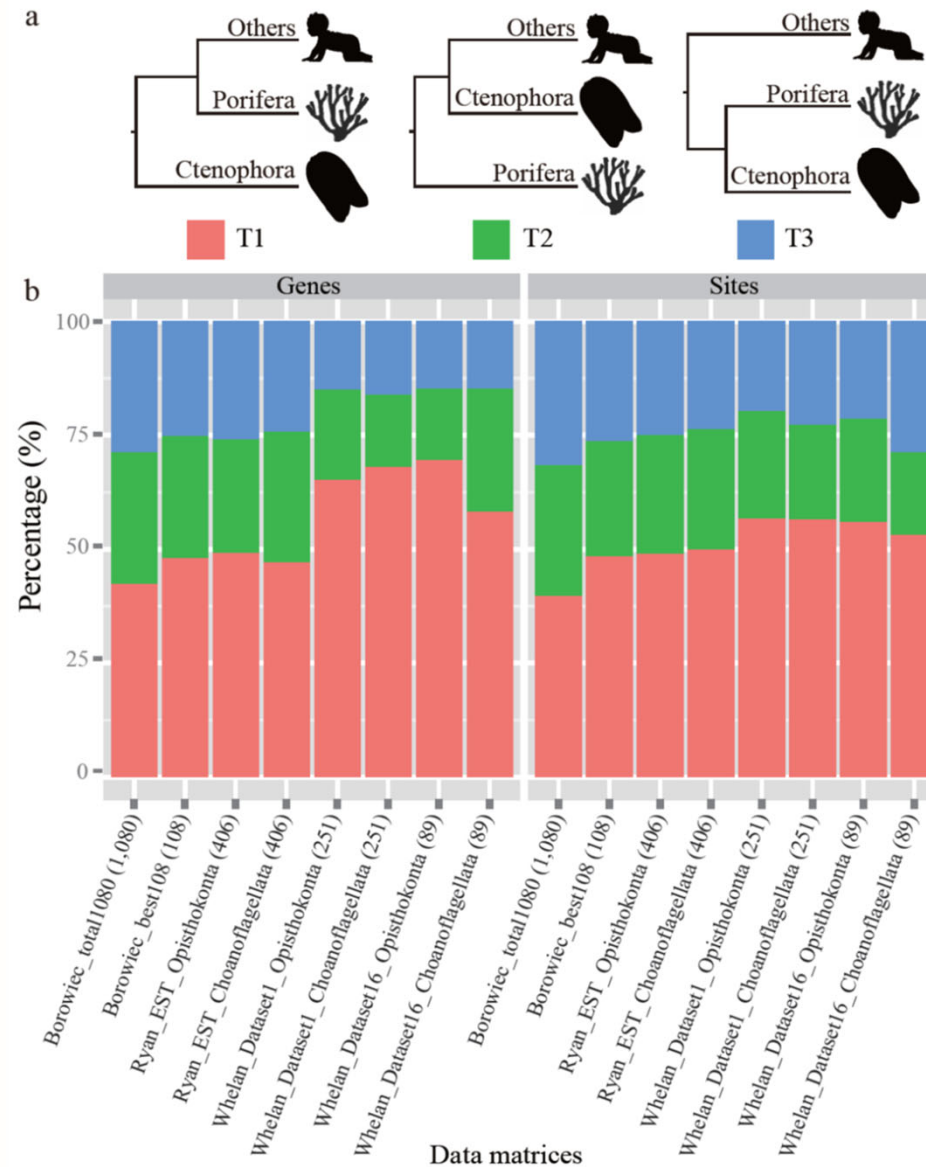


Shen et al. (2017) Nature Ecol. Evol.

Summarizing Phylogenetic Signal Across Genes and Sites



Summarizing the Signal Across All 3 Possible Topologies

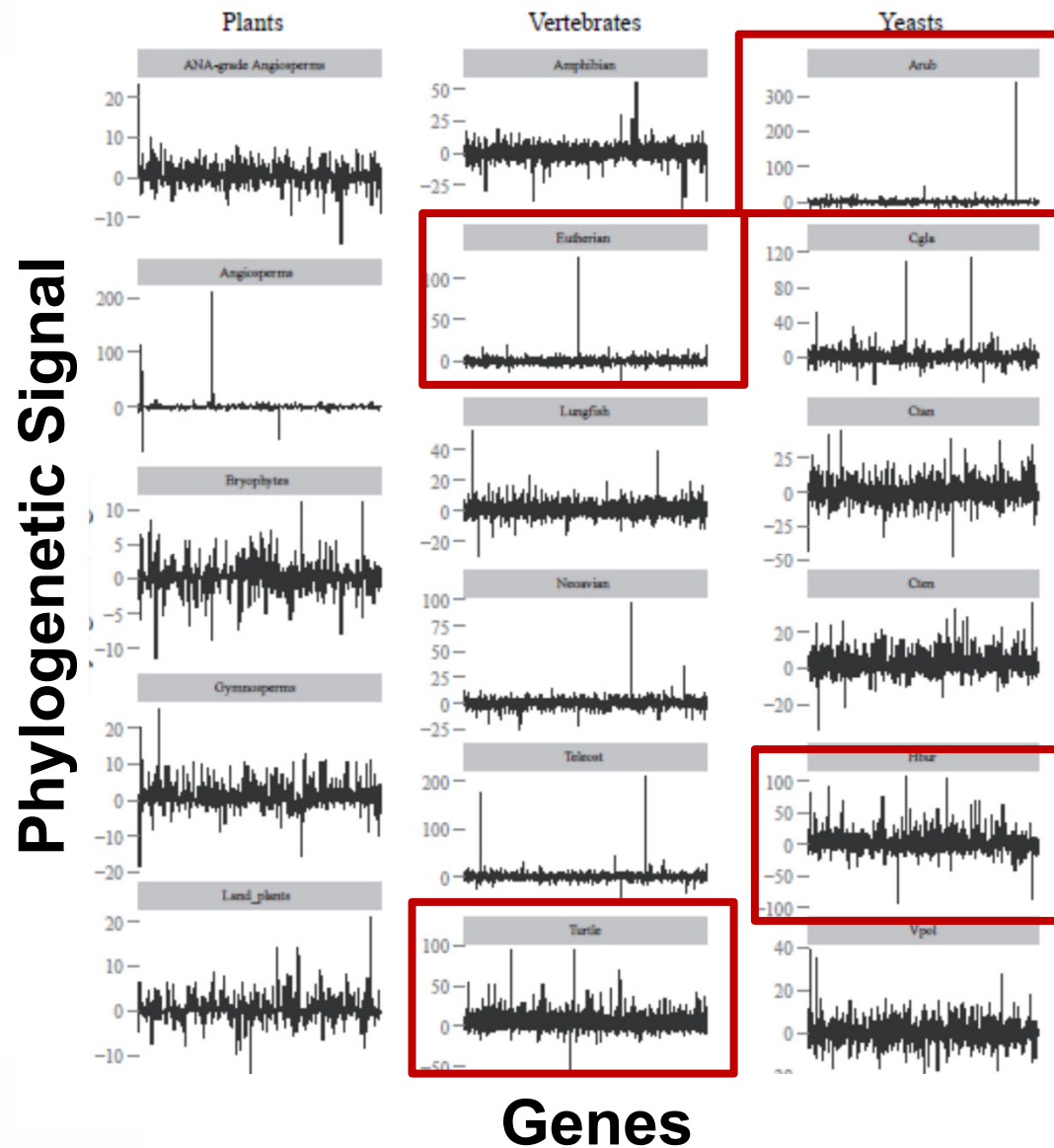


Testing Several Contentious Branches of the Tree of Life

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
	Zygnematomyceae 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 Saccharomyces sensu stricto yeasts	<i>Naumovozya castellii</i> rather than <i>Candida glabrata</i> sister to Saccharomyces sensu stricto 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>

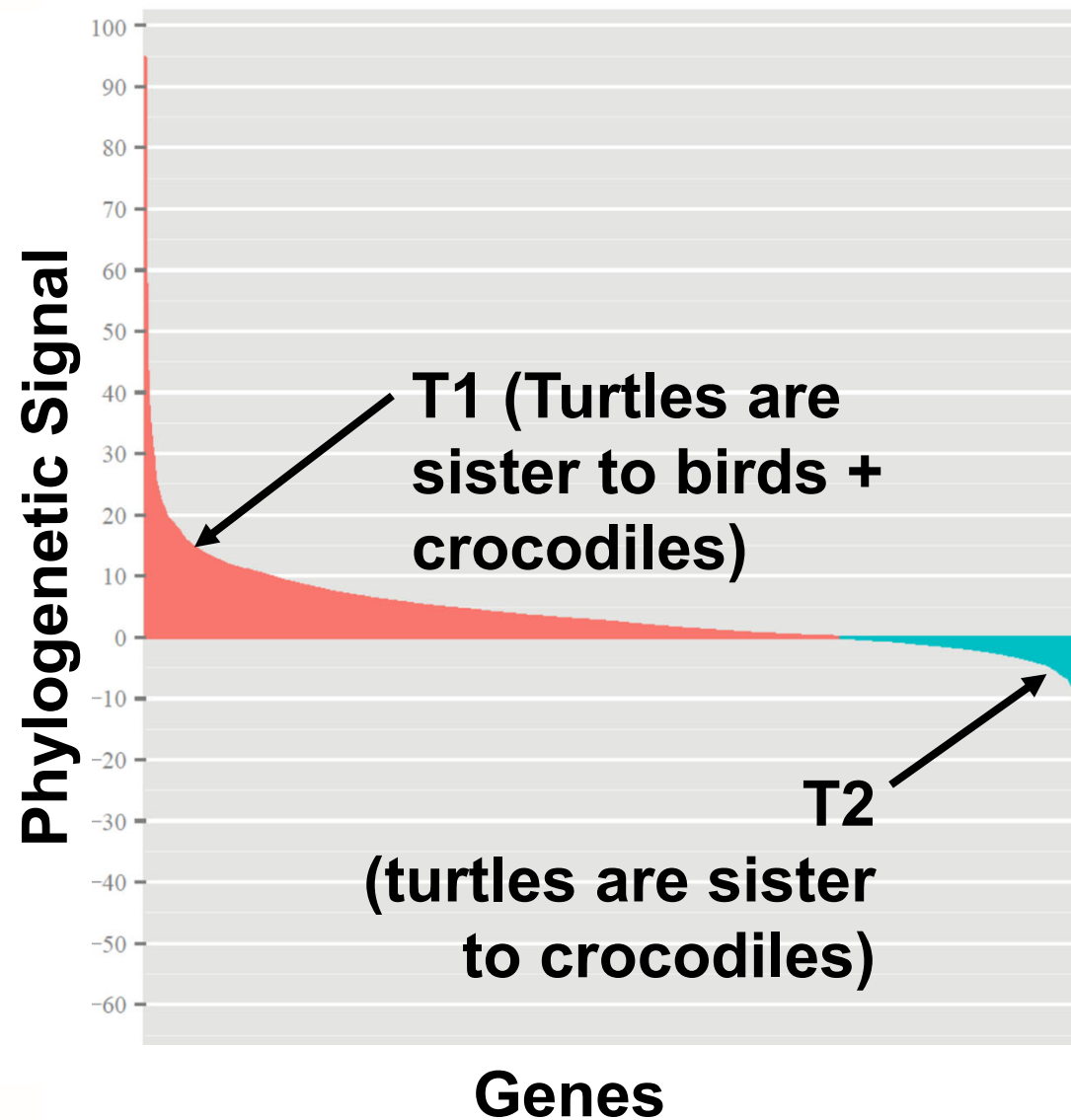


Phylogenetic Signal in Contentious Branches of the ToL

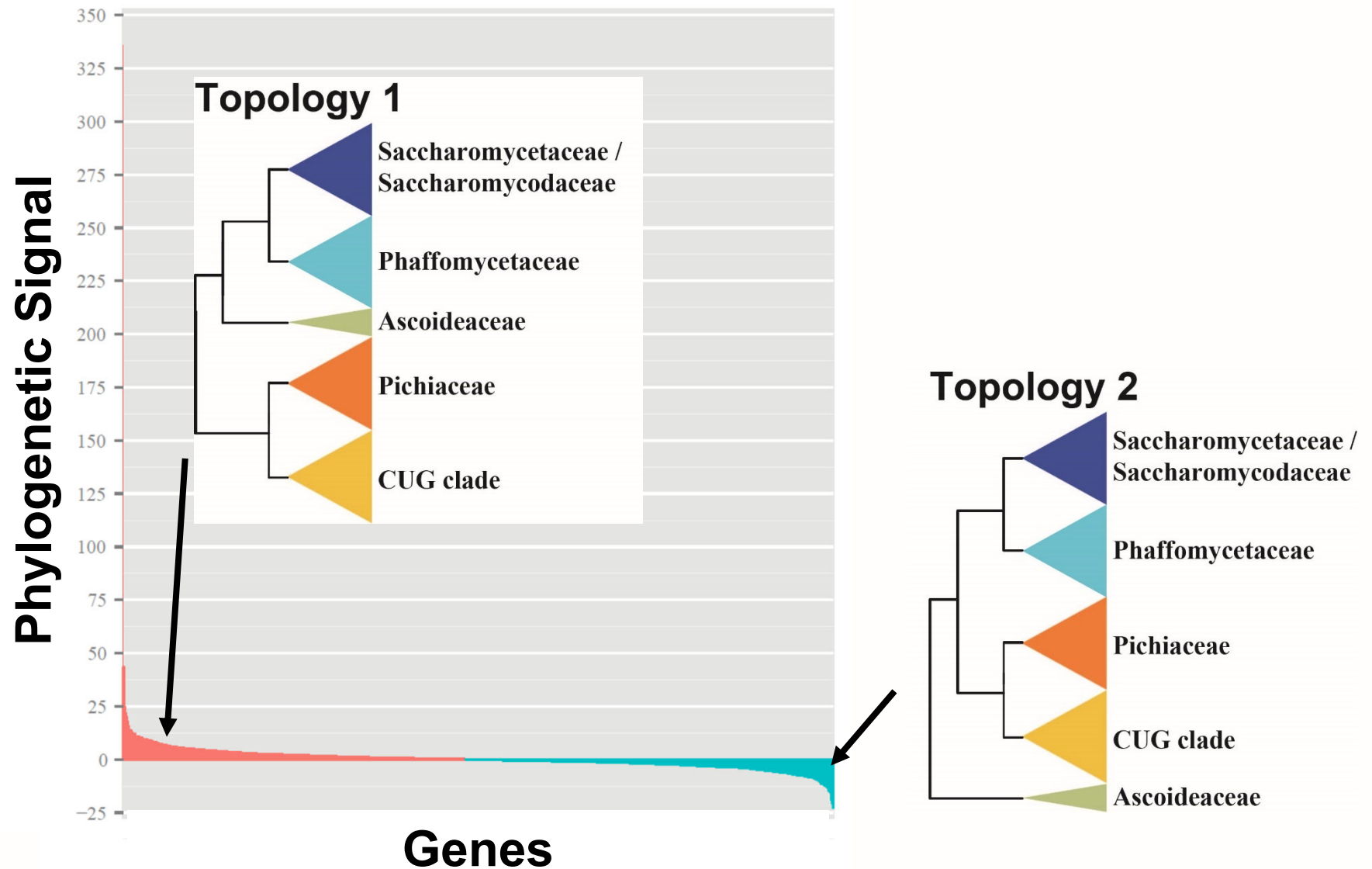


Shen et al. (2017) Nature Ecol. Evol.

The Signal in Some Branches is Very Strong...

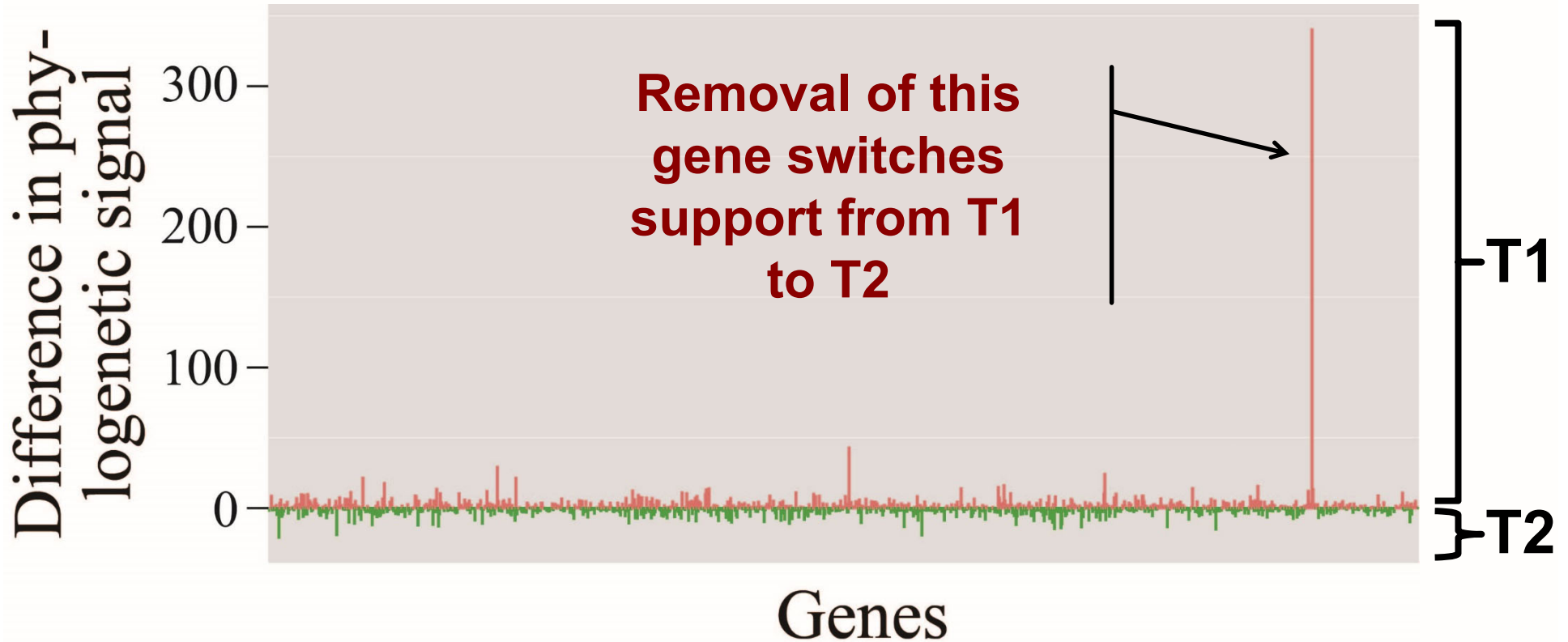


...But in Others It Stems from One or Two Genes

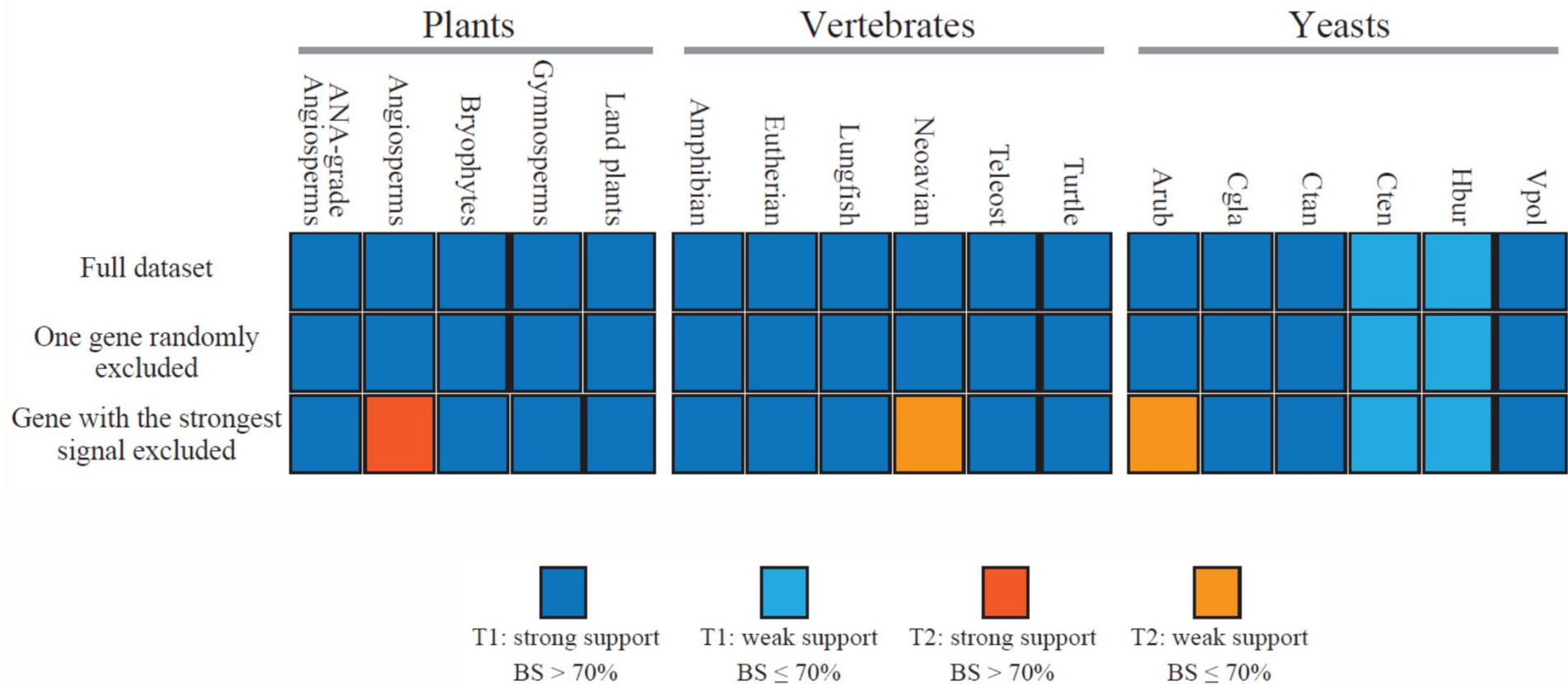


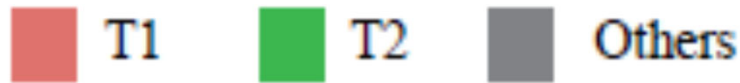
Phylogenetic Signal per Gene for the Two Hypotheses

1233 genes, 86 yeast taxa



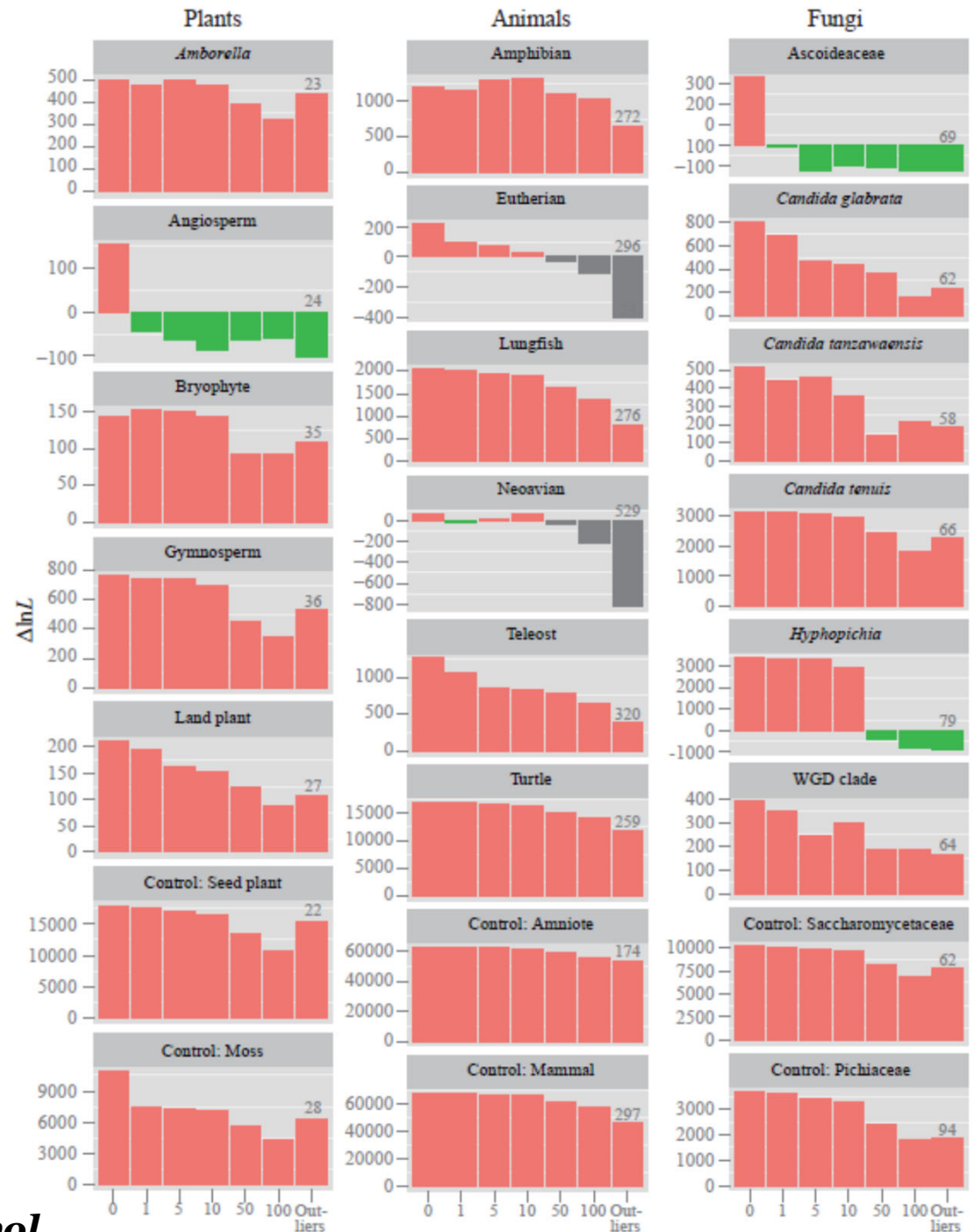
What Happens if we Remove That One Gene?



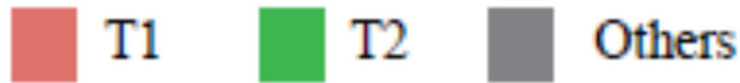


of genes excluded

Quantifying the Impact of Removing Opinionated Genes

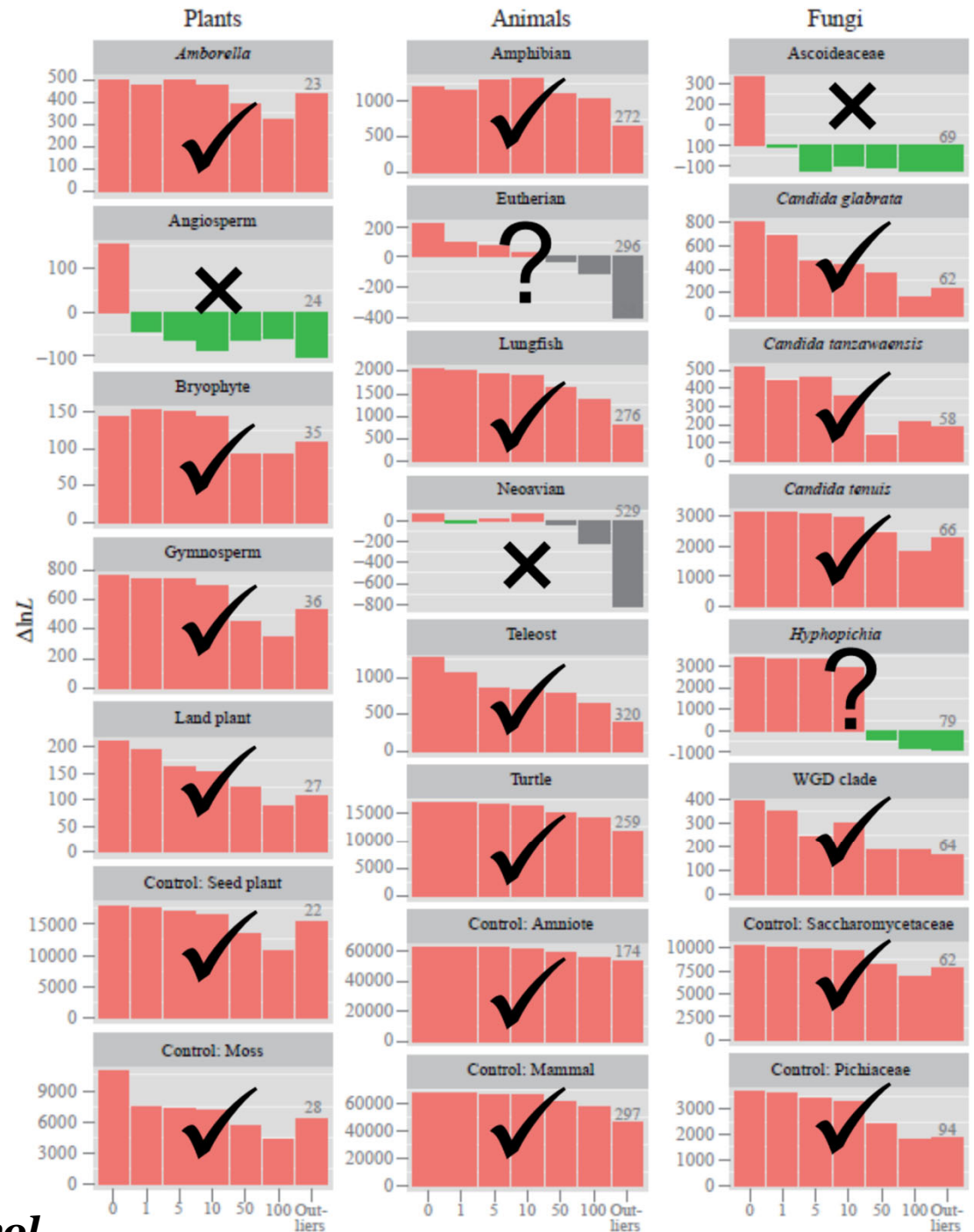


Shen et al. (2017) *Nature Ecol. Evol.*



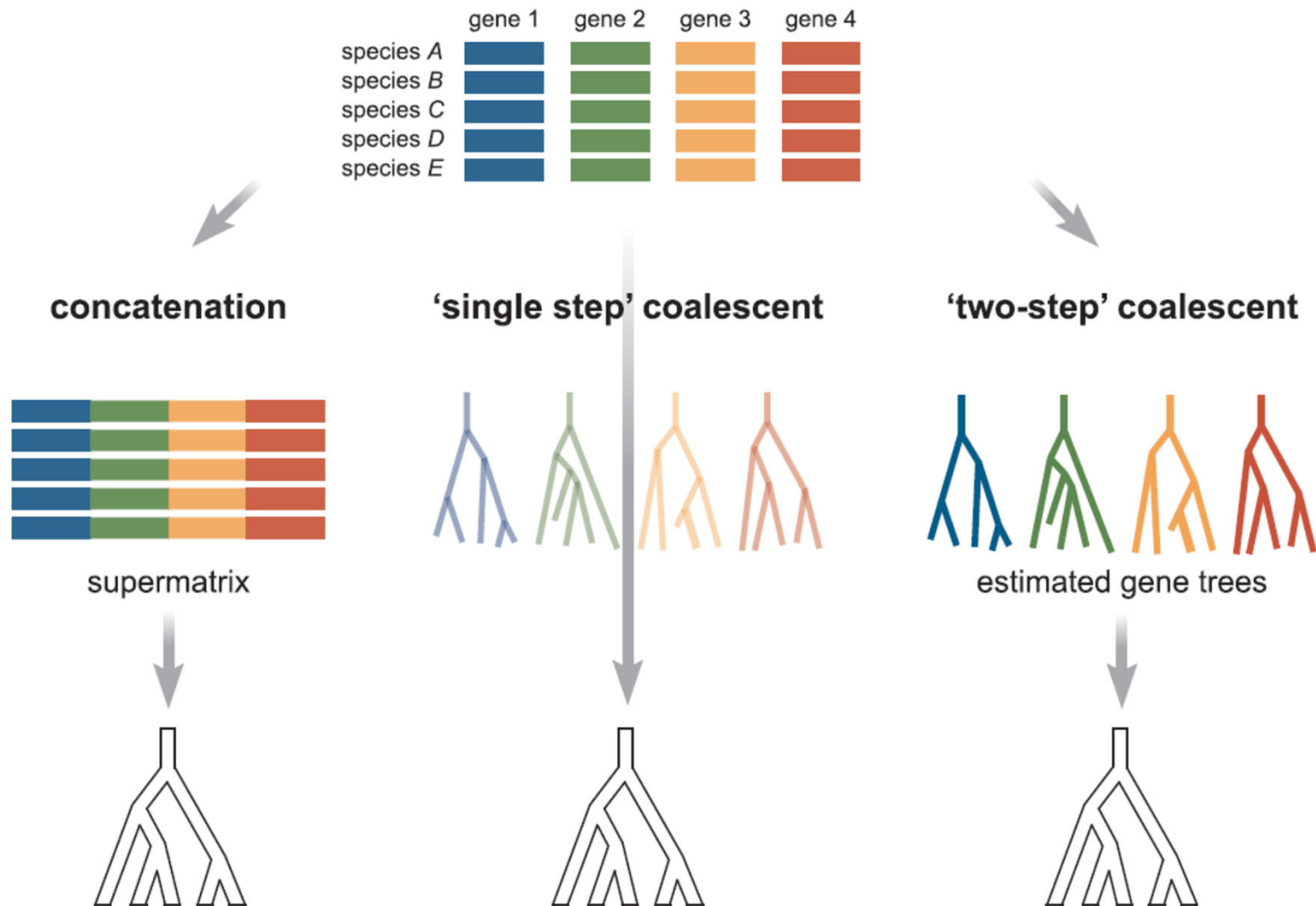
of genes excluded

*Which Branches
are Resolved and
Which are
Unresolved?*

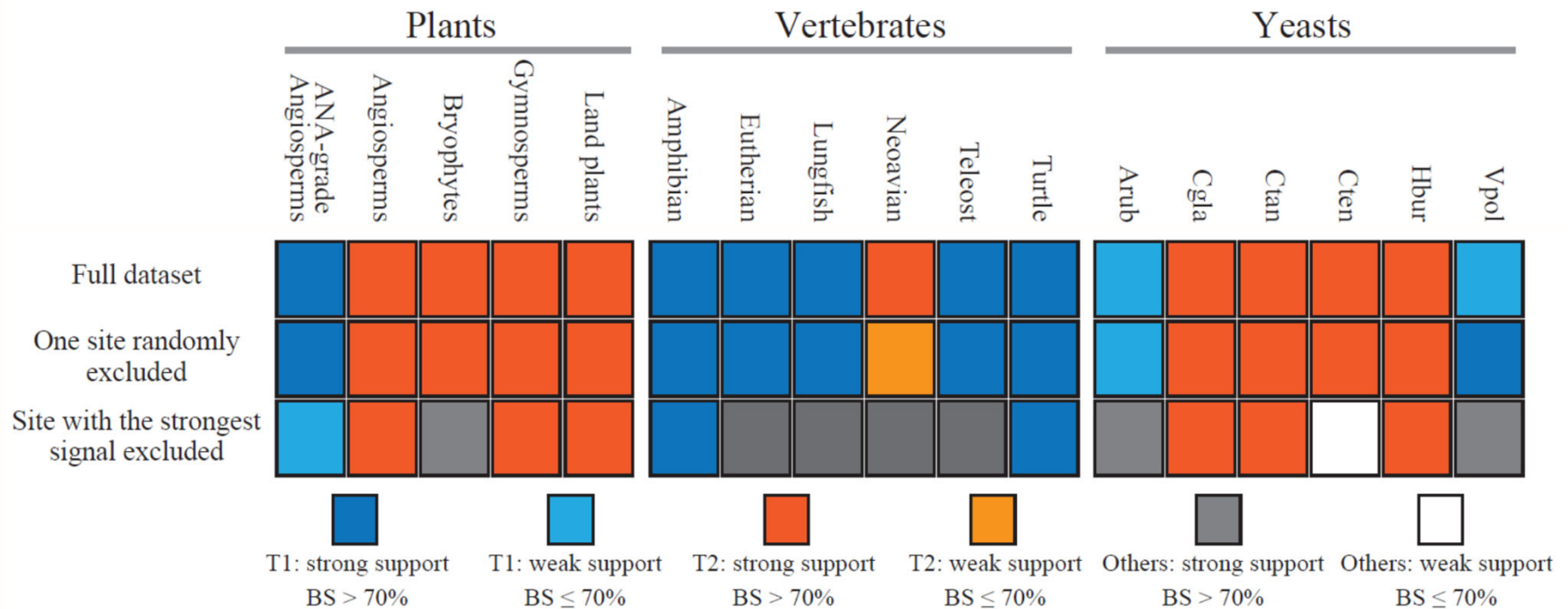


Shen et al. (2017) Nature Ecol. Evol.

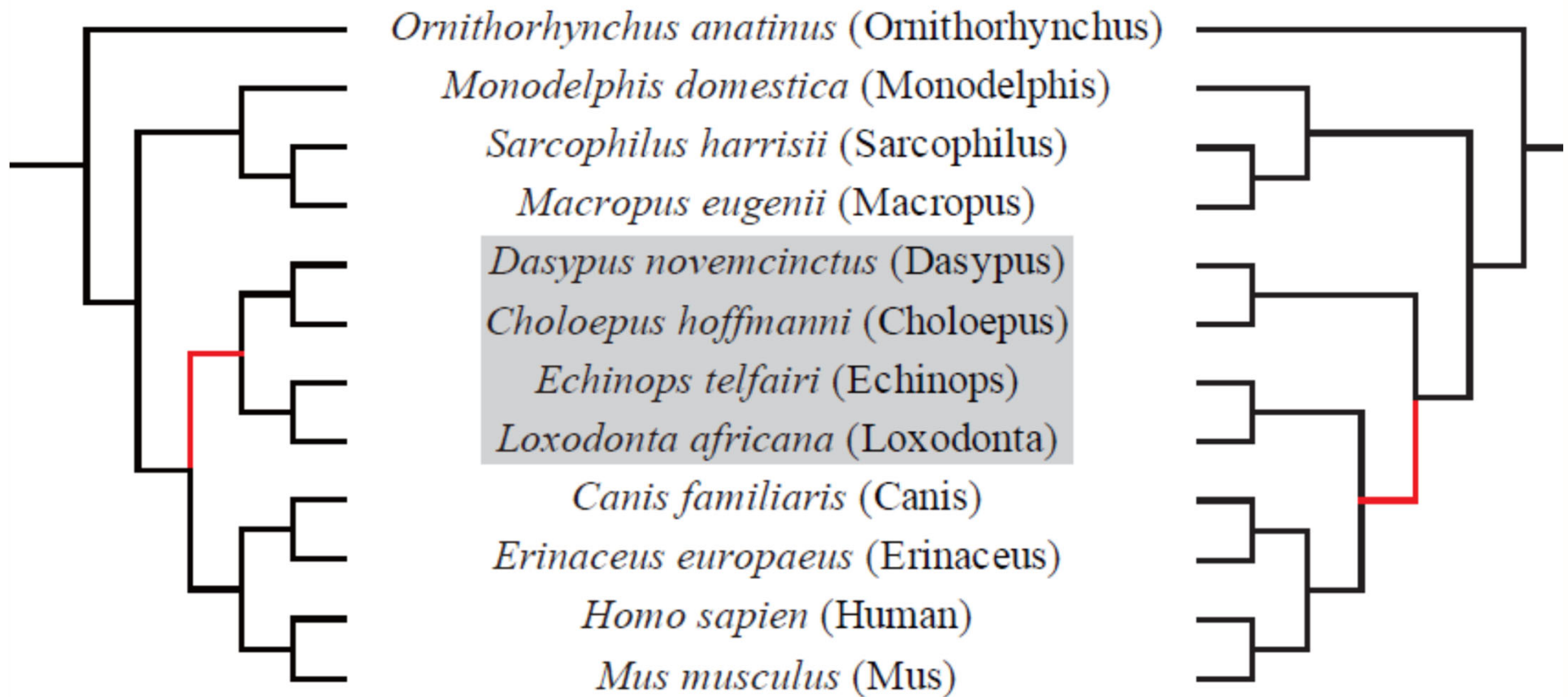
Methods for Phylogenomic Inference



What Happens if we Remove One Site from Every Gene?



Removing 1 Site Alters the Topology



What's Going On?

Explanation #1: 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)

The Making of Biodiversity across the Yeast Subphylum



Hittinger lab



Kurtzman lab



Rokas lab

The Making of Biodiversity across the Yeast Subphylum

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



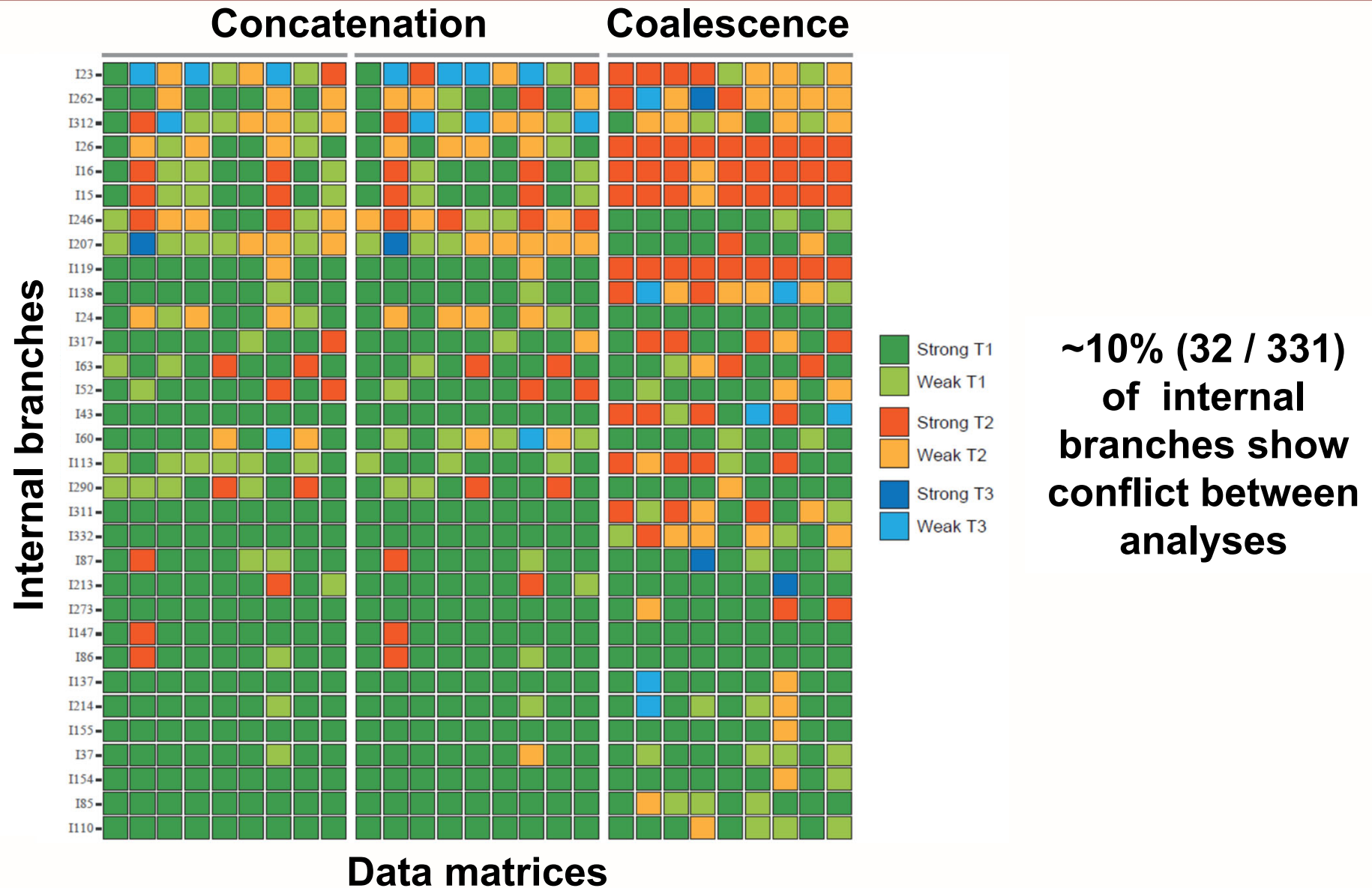
- ❖ Sequenced the genomes of 220 species (196 Y1000+ species + 24 RIKEN genomes); most of them are from type strains
- ❖ + 112 publicly available genomes -> 332 genomes
- ❖ Sampled taxa from 79 / 92 genera (~85%)

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

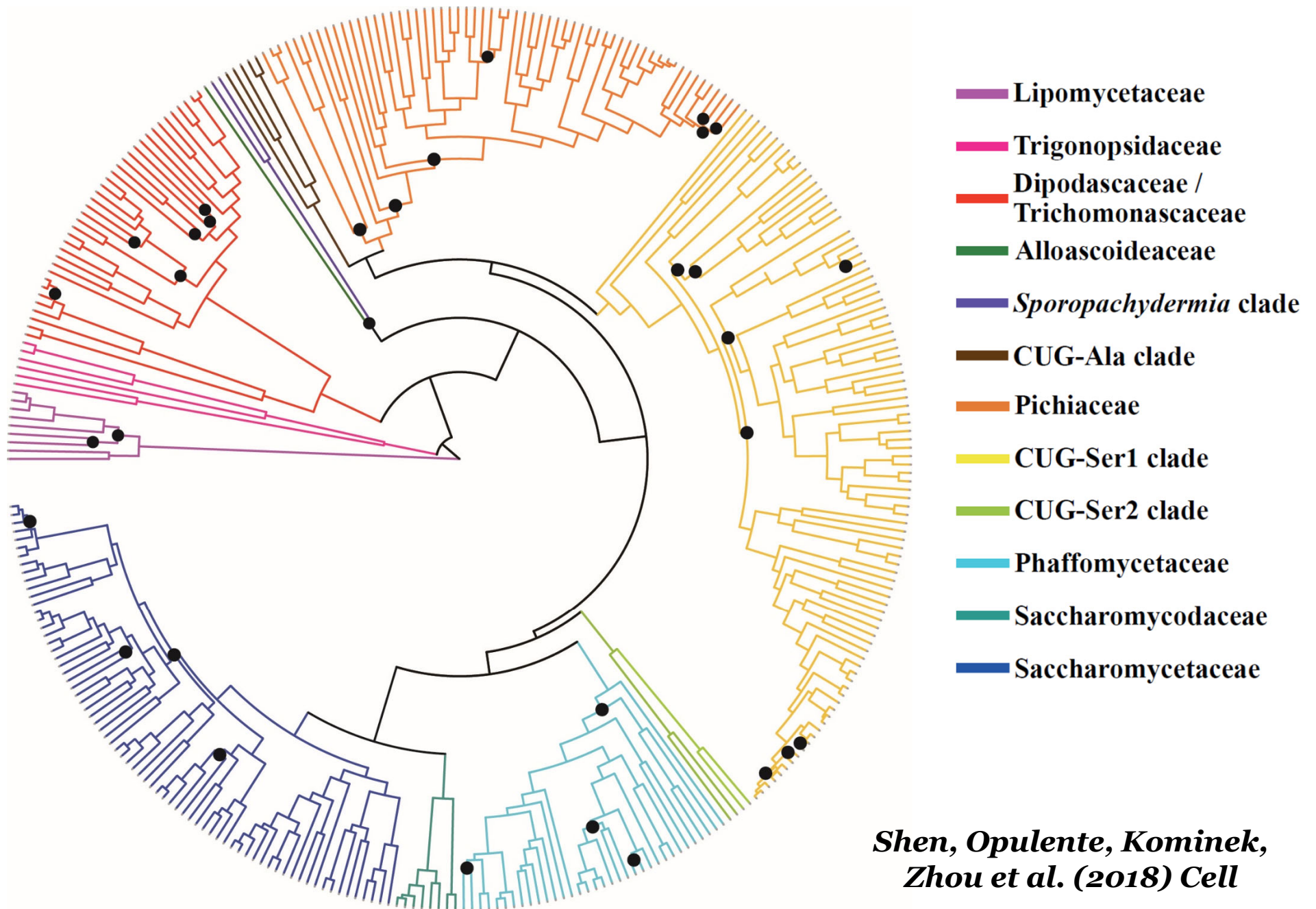
A circular phylogenetic tree of Ascomycota species. The tree is rooted at the center and branches outwards. Major clades are labeled: *Brettanomyces* (orange), *Yarrowia* (red), *Saccharomyces* (blue), *Candida albicans* (yellow), and *Shen* (green). A specific clade within the *Saccharomyces* group is highlighted with a blue circle and labeled "WGD clade". The tree is surrounded by a circular color gradient bar.

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

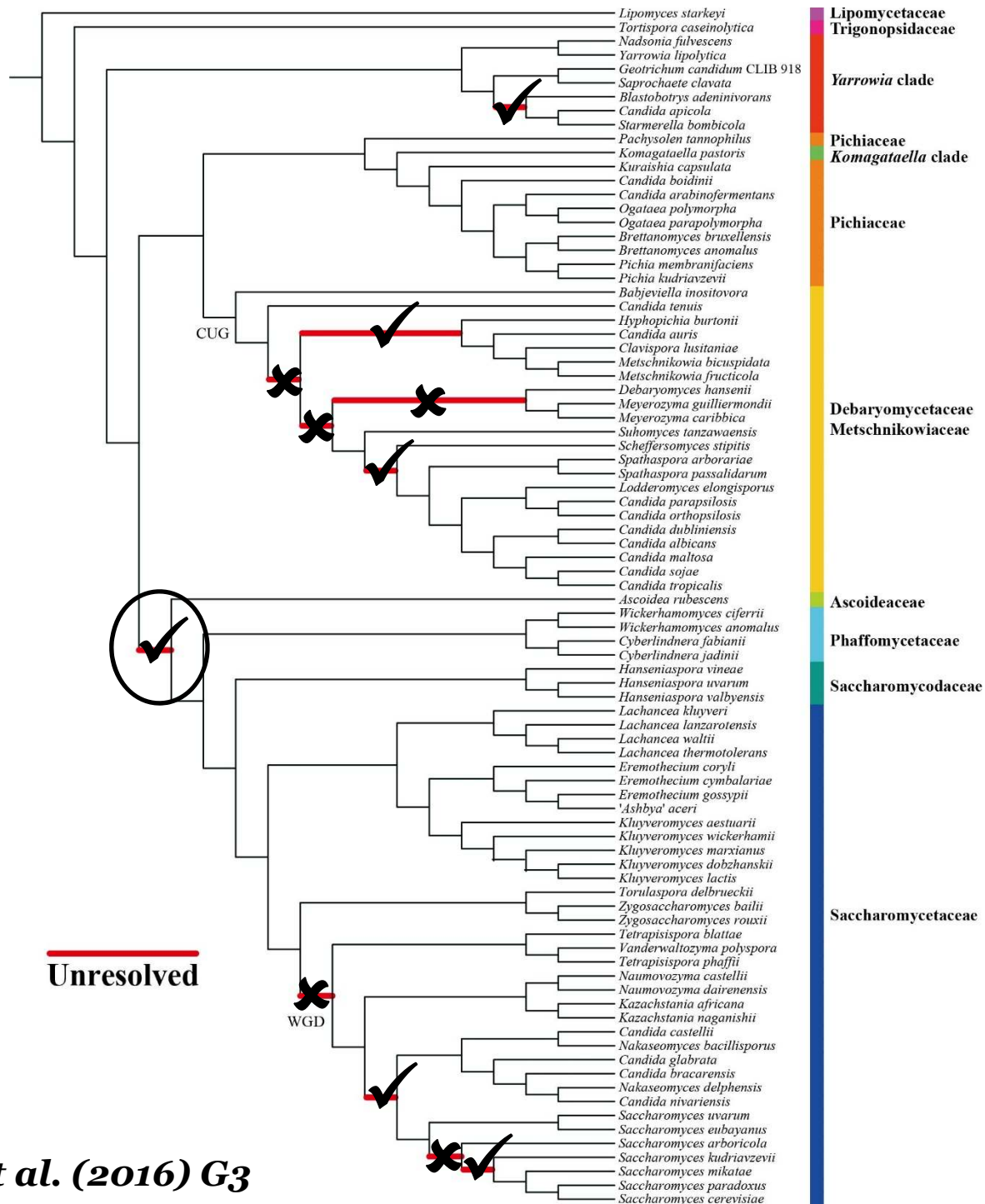
The 32 Conflicting Branches in the Yeast Phylogeny



Distribution of Conflict on the Yeast Phylogeny



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

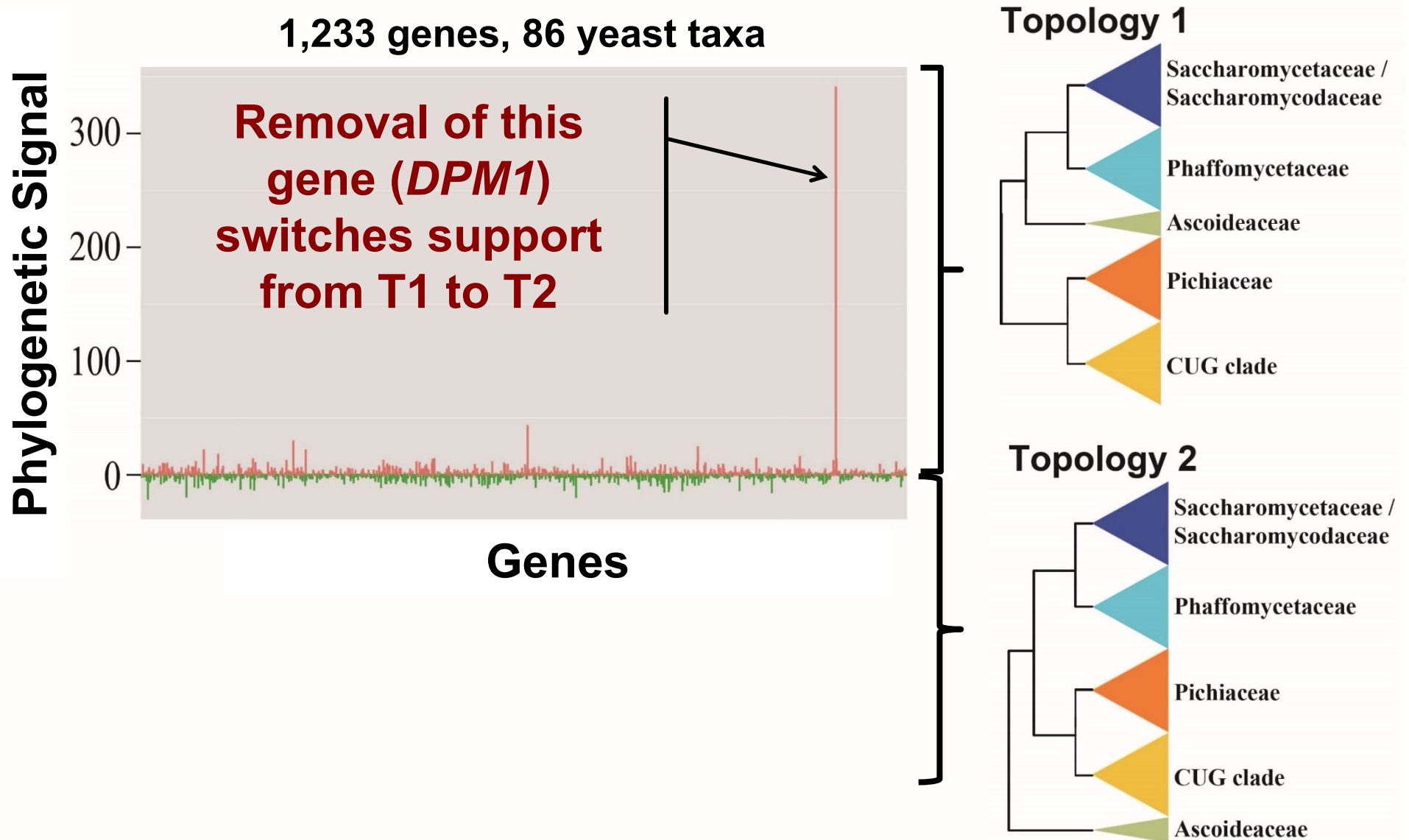


1,233-gene, 86-taxon data matrix

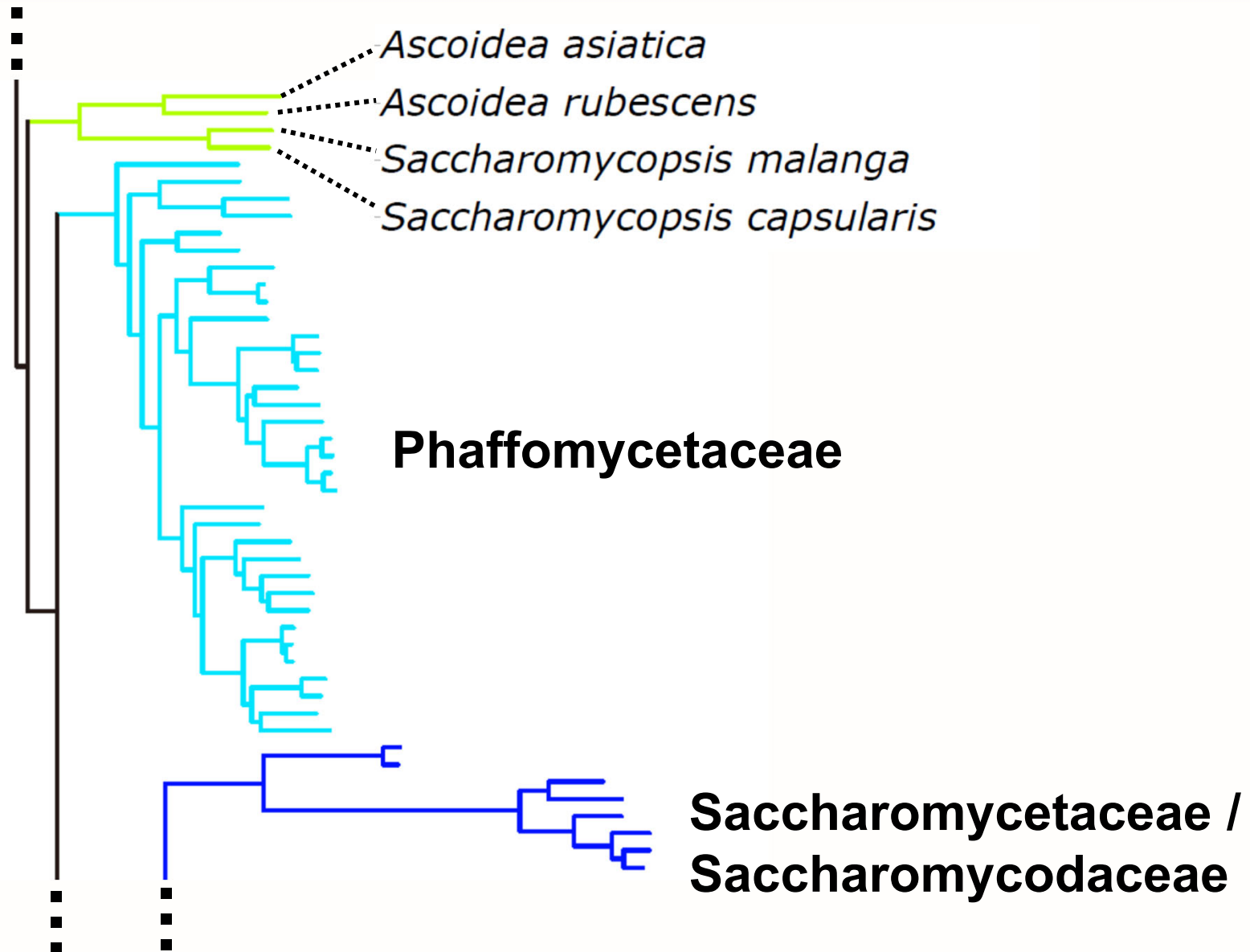
~13% (11 / 85) of internal branches conflict between analyses

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

A Single Gene Governs the Placement of Ascoideaceae

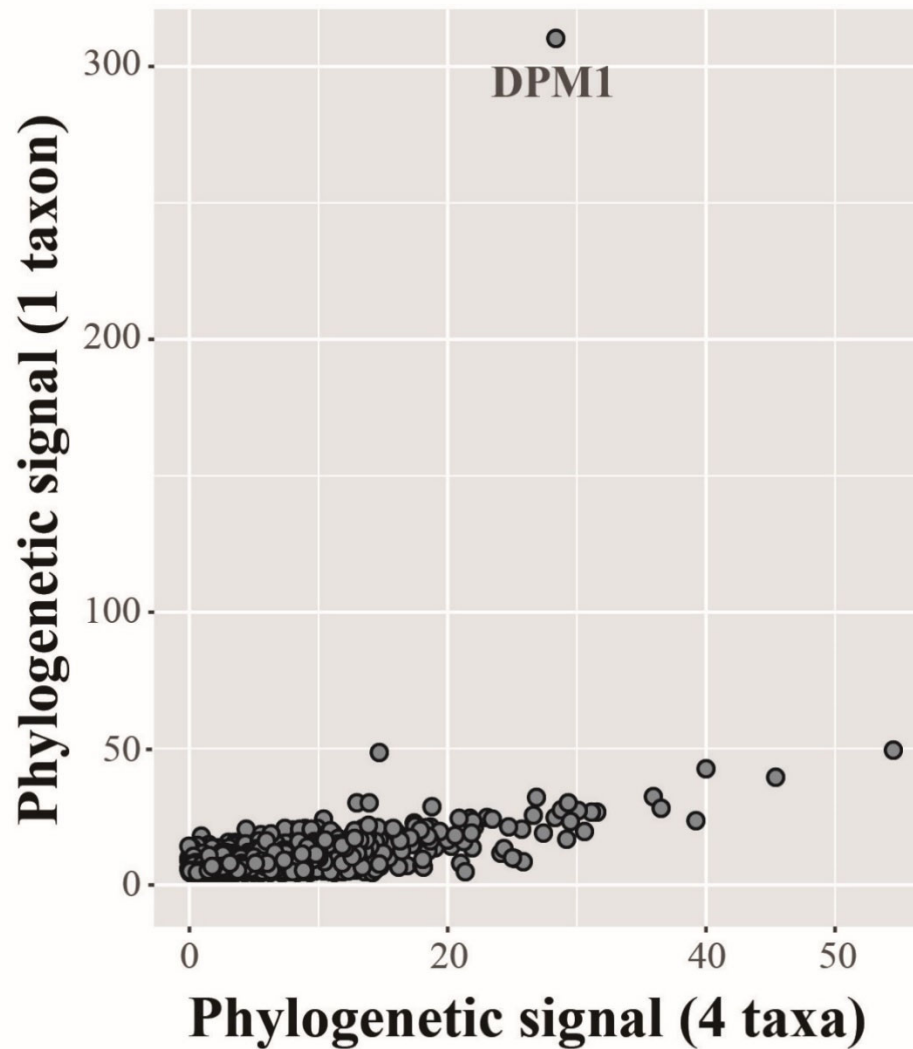


Sampling of 3 Additional Taxa “Breaks” the Long Branch



Sampling of 3 Additional Taxa Decreases Gene's Signal

2,408 genes, 329 – 332 yeast taxa



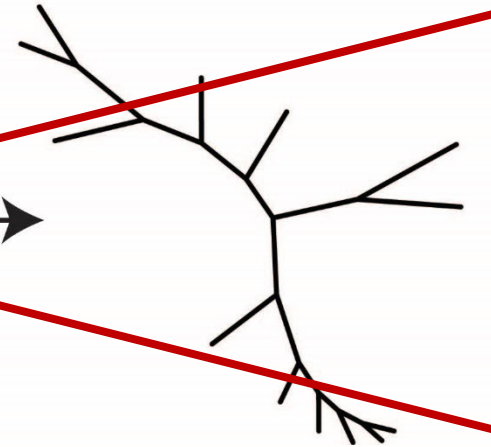
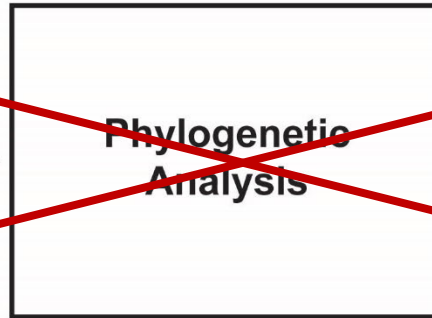
Shen, Opulente, Kominek, Zhou et al. (2018) Cell

Genomfart?

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

The Way Forward

~~taxon_1~~ ~~ACCCGATAGACAA~~
~~taxon_2~~ ~~.C.G.....~~
~~taxon_3~~ ~~.....CT..~~
~~taxon_4~~ ~~....A.....C~~
~~taxon_5~~ ~~T.A.....~~
~~taxon_7~~ ~~.....TT....~~
~~taxon_8~~ ~~..G....TT....~~
~~taxon_9~~ ~~.....G.....~~
~~taxon_10~~ ~~T.....~~
~~taxon_11~~ ~~T.....~~
~~taxon_12~~ ~~..GG.....T..~~
~~taxon_13~~ ~~..GG...C..T..~~



**Multiple sequence
alignment / data
matrix
reconstruction**



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



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



**Investigate alternative
hypotheses for branches
showing conflict / assess
sensitivity of results**



**Only report resolution of
branches that you have
support for**

Take Home Messages



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

Rodrigo et al. (1994)

**Chapter in: Sponge in Time and Space;
Biology, Chemistry, Paleontology**

Rokas Lab



polytomies happen...

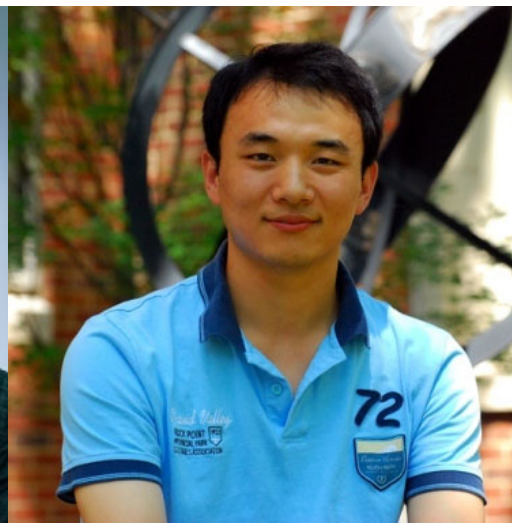
Acknowledgements



**Leonidas
Salichos**



**Xing-Xing
Shen**



**Xiaofan
Zhou**



**Alexis
Stamatakis**



National Science Foundation
WHERE DISCOVERIES BEGIN



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

@RokasLab