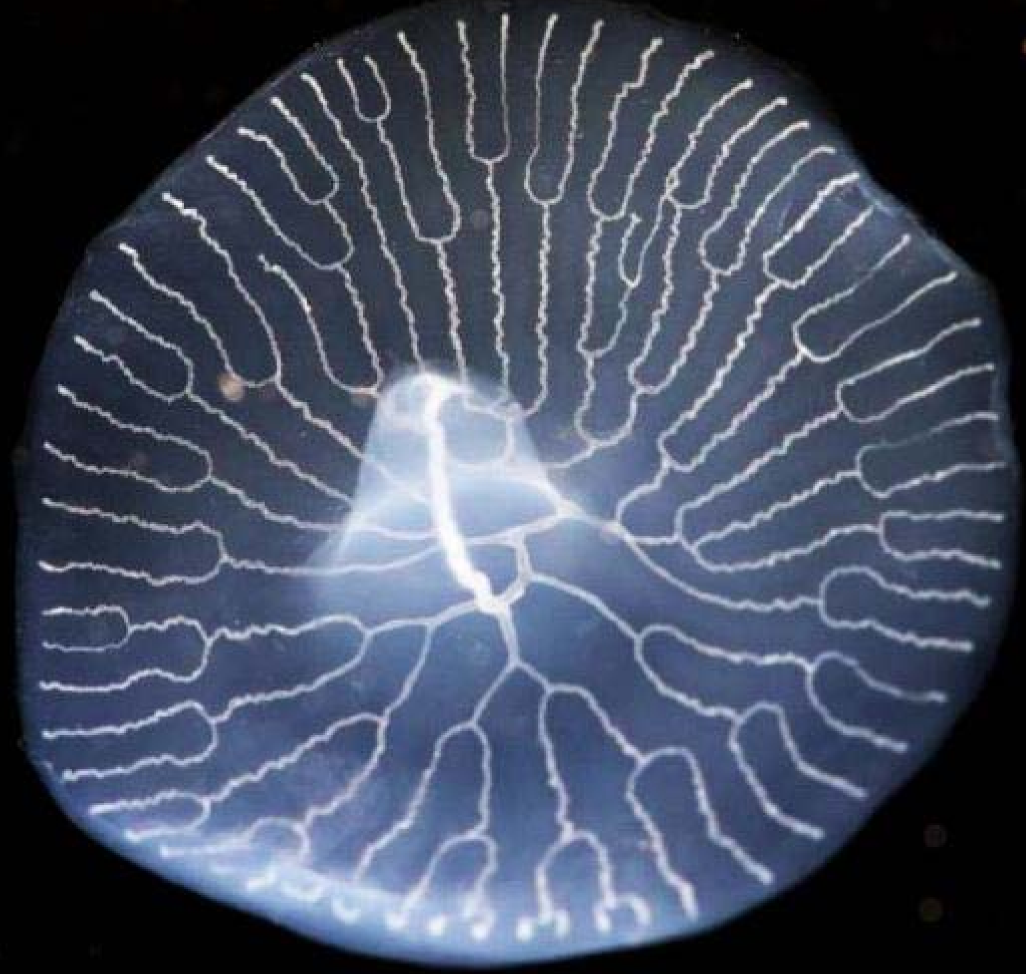
Introduction to Phylogenomics



Antonis Rokas

Department of Biological Sciences, Vanderbilt University

http://as.vanderbilt.edu/rokaslab

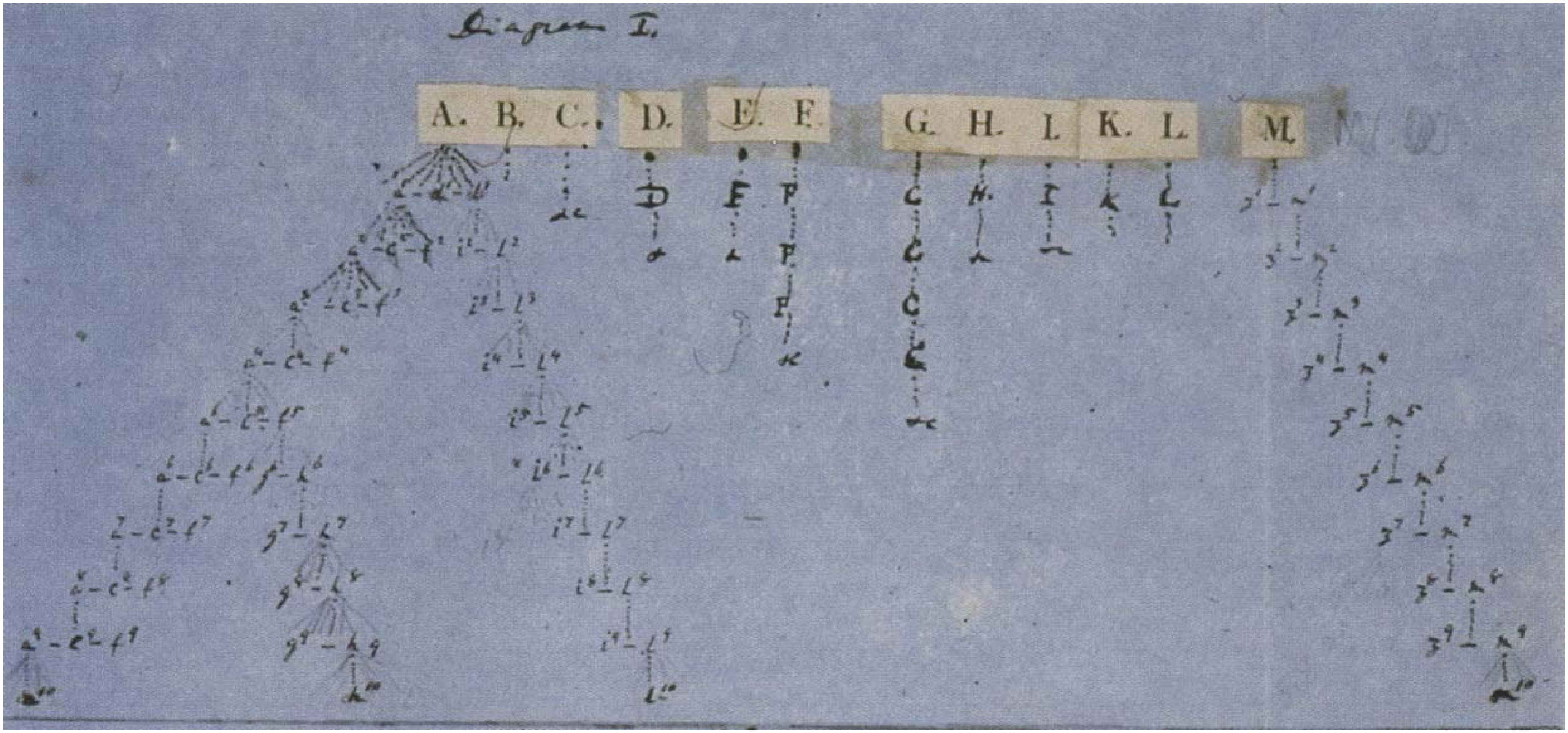
Lecture Outline

- **From Darwin to Phylogenomics**

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

- **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"

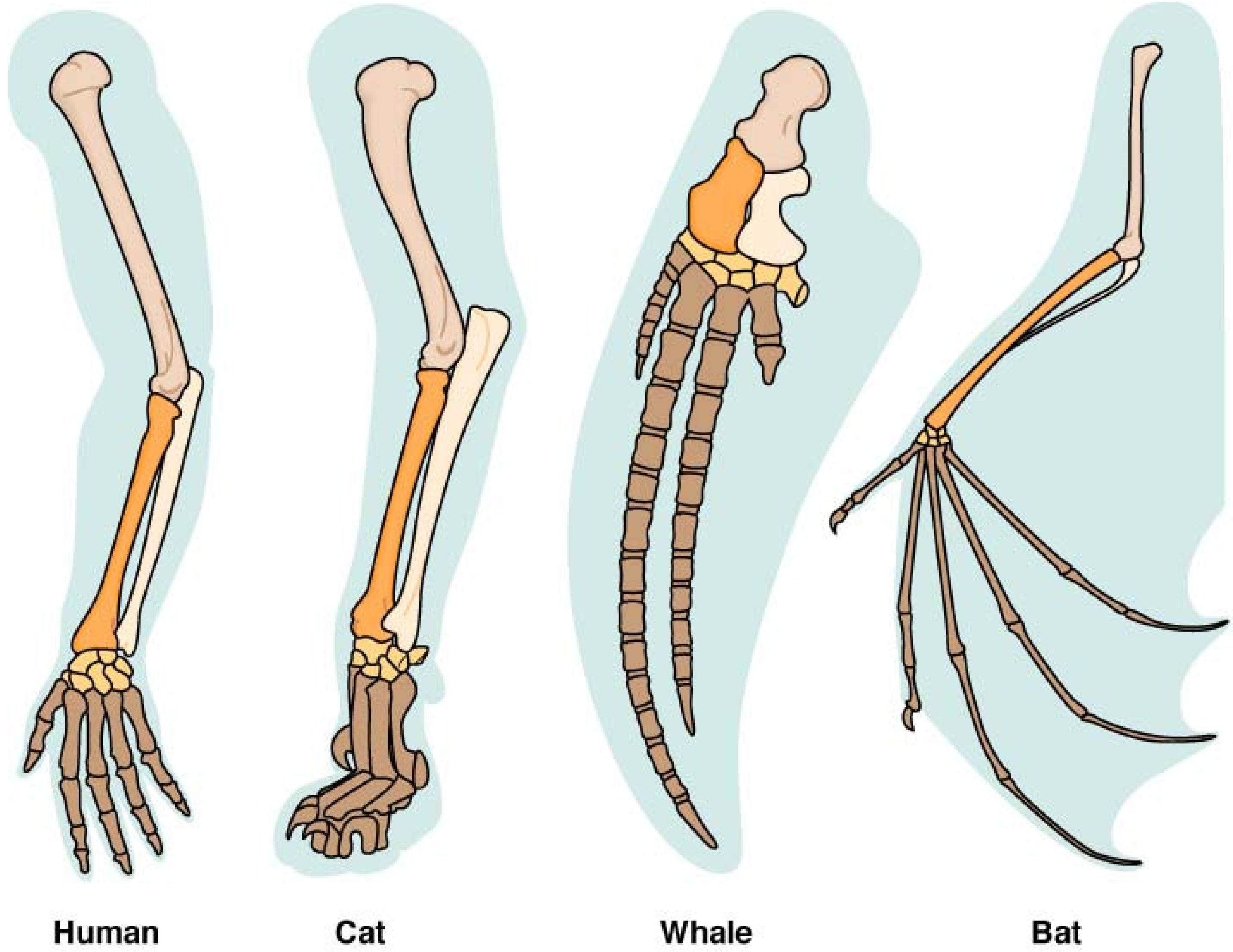
Darwin's hand-made proof of the famous diagram in his Origin of Species; Maderspacher (2006) Curr. Biol.

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.

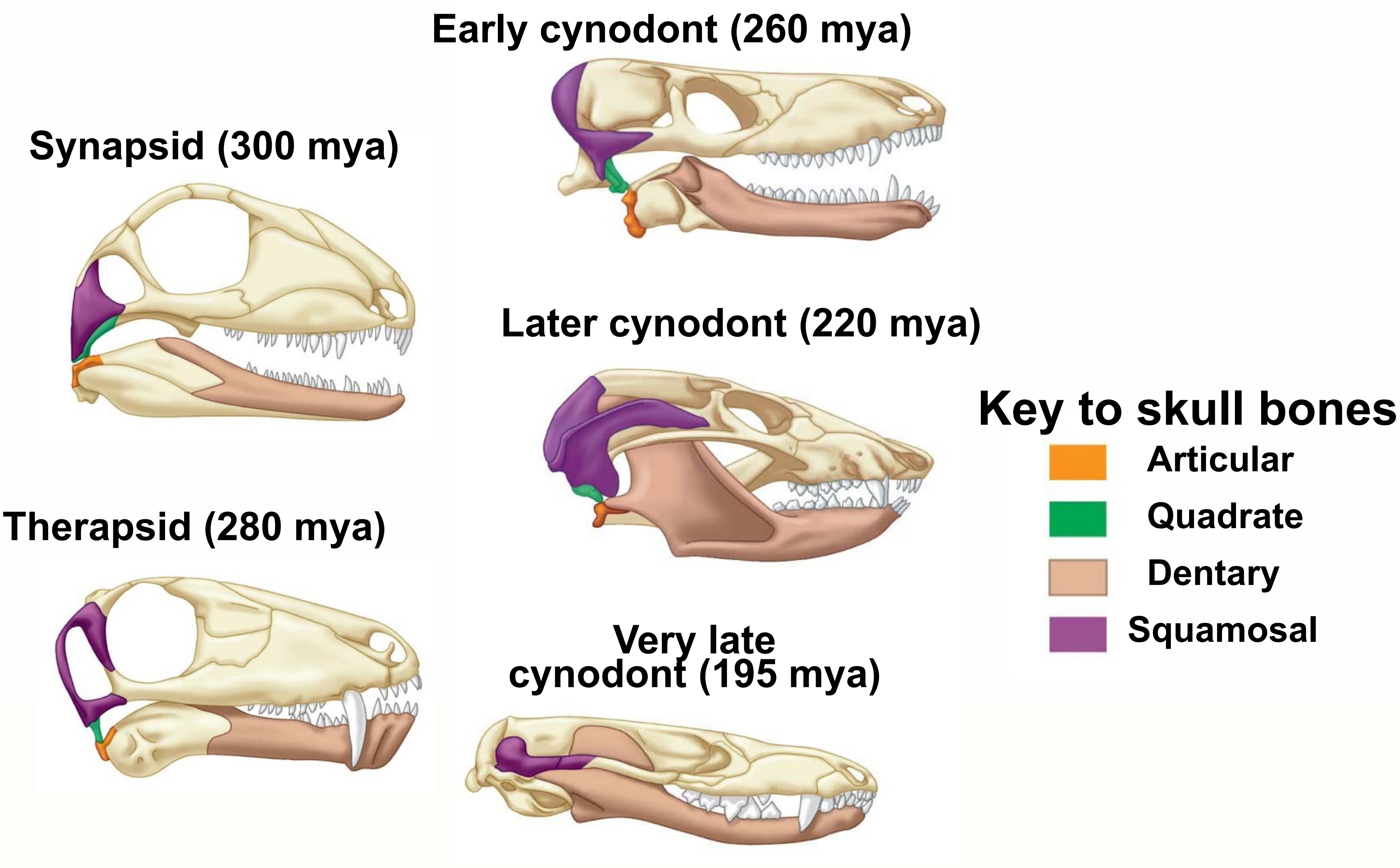
When we can feel assured that all the individuals of the same species, and all the closely allied species of most genera, have within a not very remote period de-

Comparative Morphology of Extant Organisms



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

Comparative Anatomy of Fossils



Synapsid (300 mya)

Therapsid (280 mya)

Early cynodont (260 mya)

Later cynodont (220 mya)

Very late cynodont (195 mya)

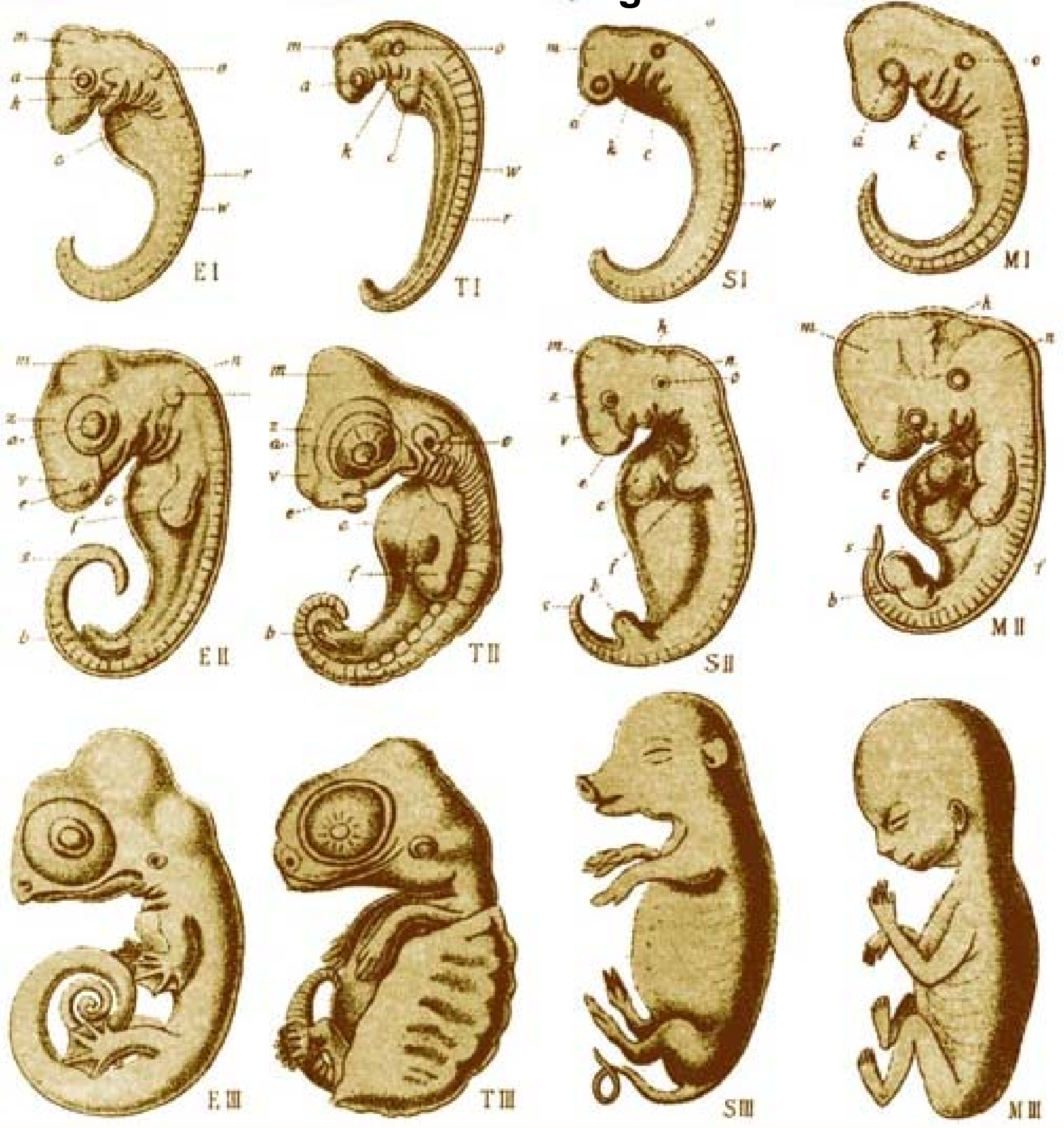
Key to skull bones

- 🟧 **Articular**
- 🟩 **Quadrate**
- 🟫 **Dentary**
- 🟪 **Squamosal**

Campbell (2016) Biology, 10th Ed., Fig. 25.7

Comparative Embryology

Lizard **Tortoise** **Pig** **Human**

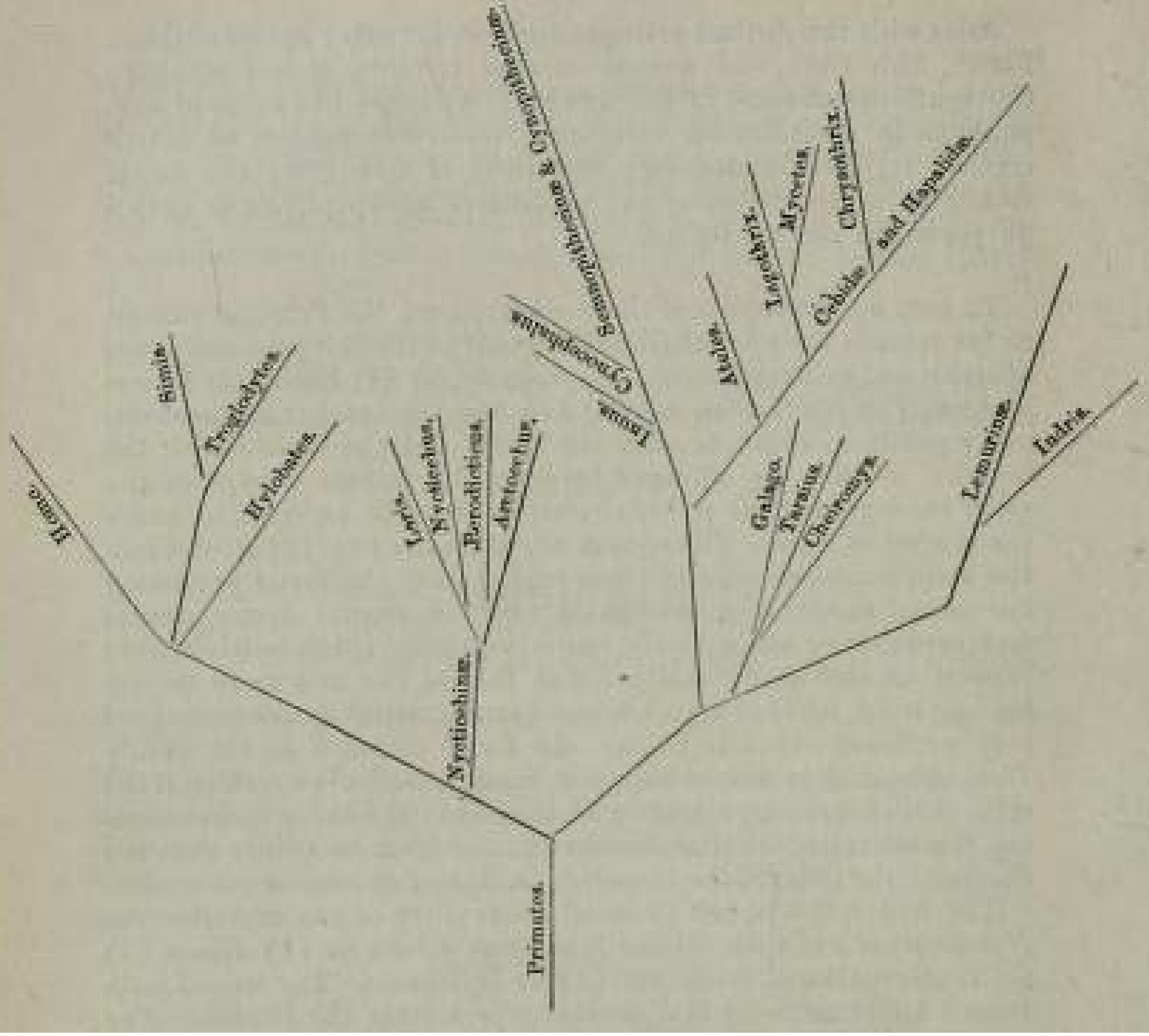


http://www.nature.com/nrg/journal/v7/n11/images/nrg1918-f2.jpg

The First Published Phylogeny

St. George Jackson Mivart

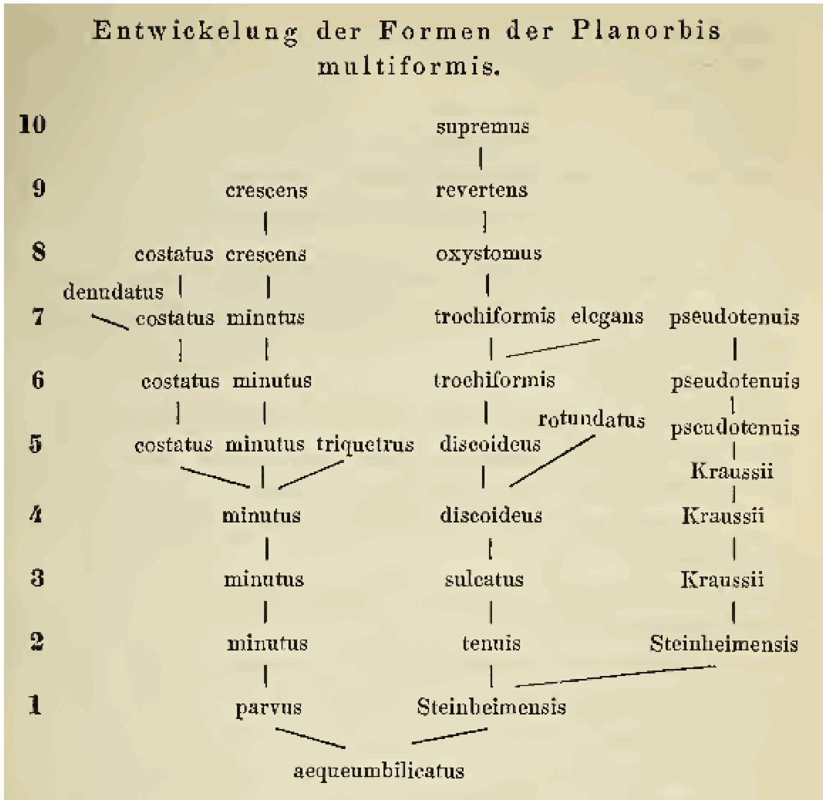
Fig. 13.



Mivart (1865) Proc. Zool. Soc. London

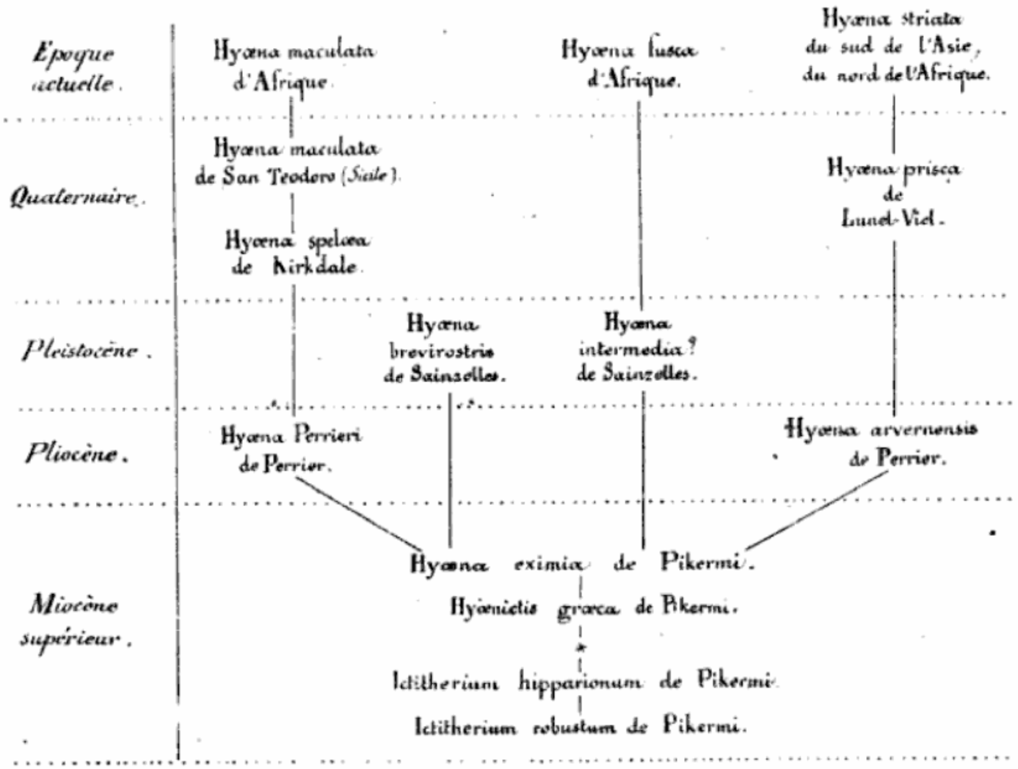
Inferring Phylogenies Becomes a Cottage Industry

Fossil gastropods



Hilgendorf, 1867

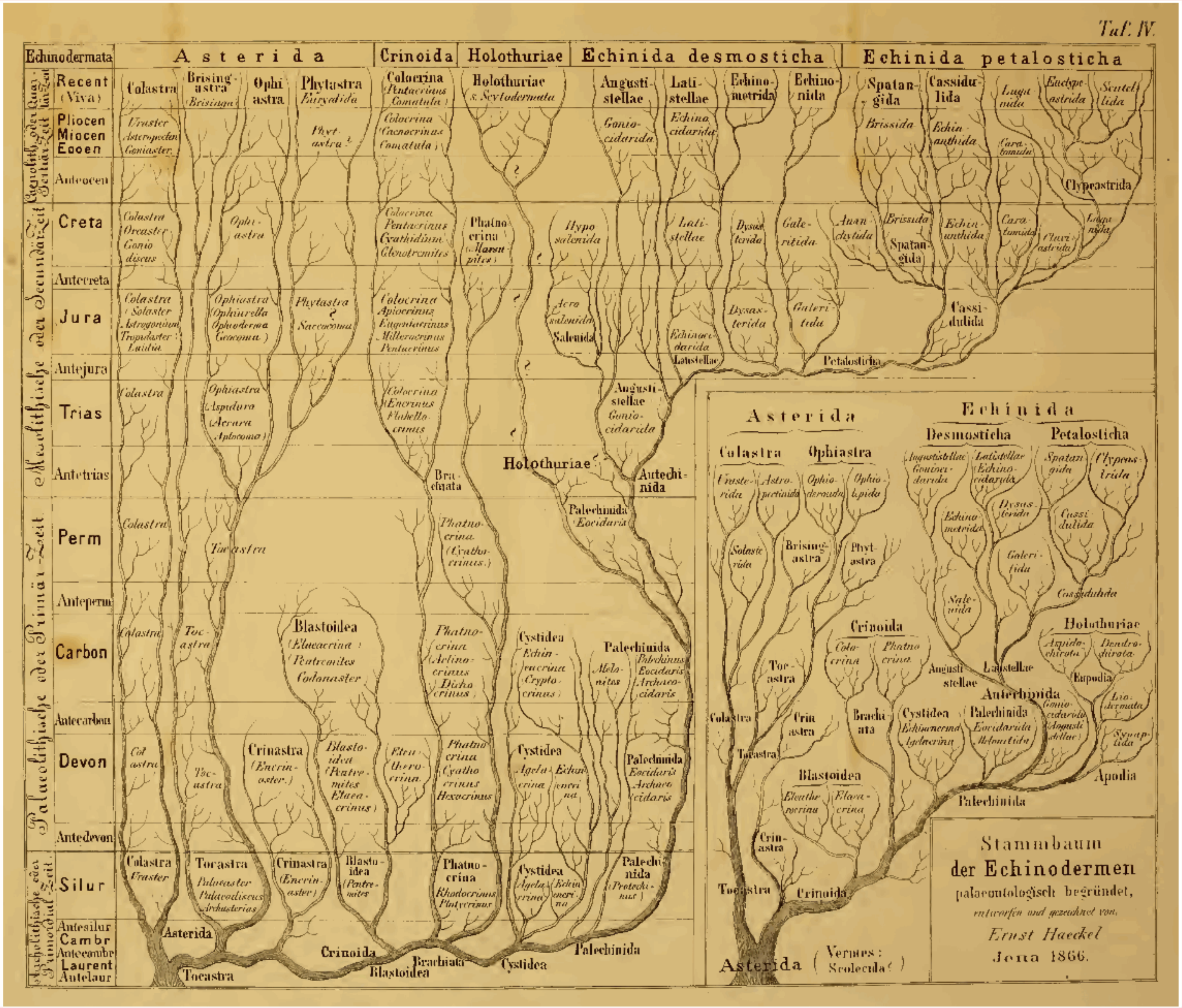
Extant and extinct mammals



Gaudry, 1866

http://phylonetworks.blogspot.com/2012/08/who-published-first-phylogenetic-tree.html

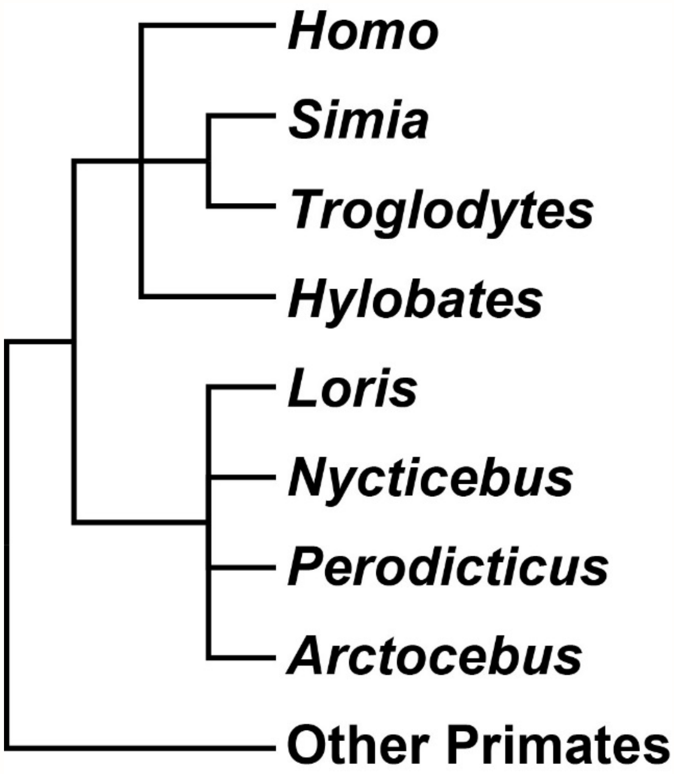
Haeckel's Phylogenies



Haeckel (1866)

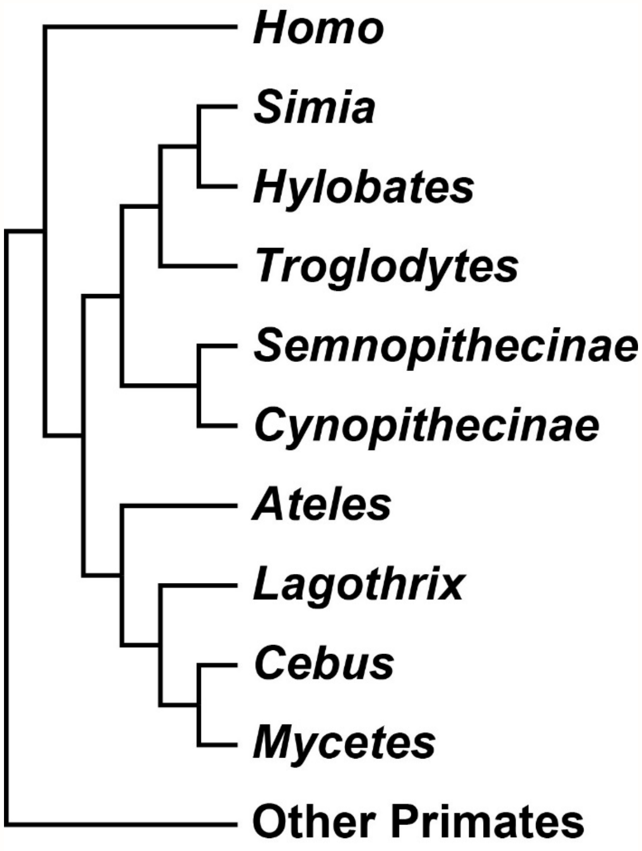
Disagreement Between Phylogenies

1865: SPINAL COLUMN

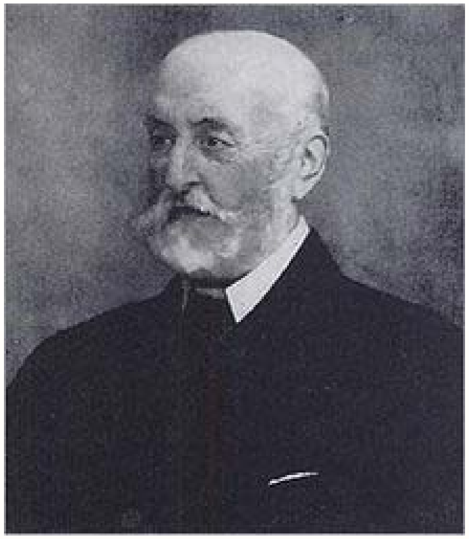


Homo
Simia
Troglodytes
Hylobates
Loris
Nycticebus
Perodicticus
Arctocebus
Other Primates

1867: LIMBS



Homo
Simia
Hylobates
Troglodytes
Semnopithecinae
Cynopithecinae
Ateles
Lagothrix
Cebus
Mycetes
Other Primates



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

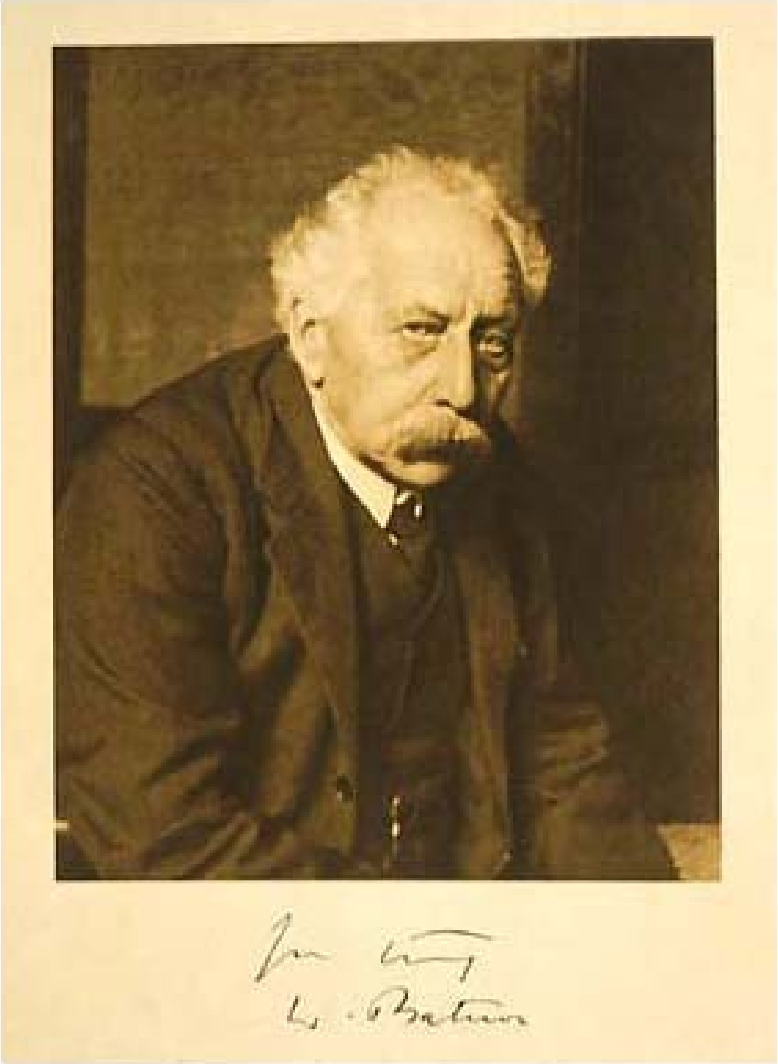
Comparative Morphology & Embryology in Trouble

By the turn of the century, a sense of despair about the conflicting phylogenies produced by comparative morphology and embryology has well settled 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



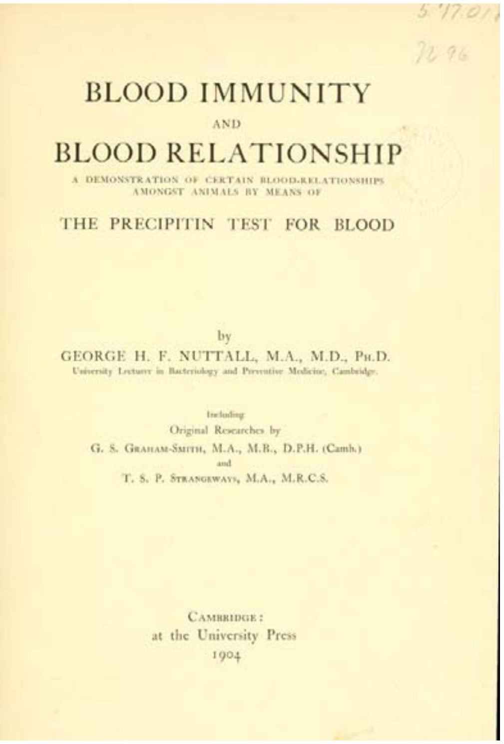
Courtesy of American Philosophical Society, Curt Stern Papers.
Noncommercial, educational use only.

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



BLOOD IMMUNITY
AND
BLOOD RELATIONSHIP
A DEMONSTRATION OF CERTAIN BLOOD-RELATIONSHIPS
AMONGST ANIMALS BY MEANS OF
THE PRECIPITIN TEST FOR BLOOD

by
GEORGE H. F. NUTTALL, M.A., M.D., PH.D.
University Lecturer in Bacteriology and Preventive Medicine, Cambridge.

Including
Original Researches by
G. S. GRAHAM-SMITH, M.A., M.B., D.P.H. (Camb.)
and
T. S. P. STRANGEWAYS, M.A., M.R.C.S.

CAMBRIDGE:
at the University Press
1904

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

“Molecular” Phylogenetics of Drosophila

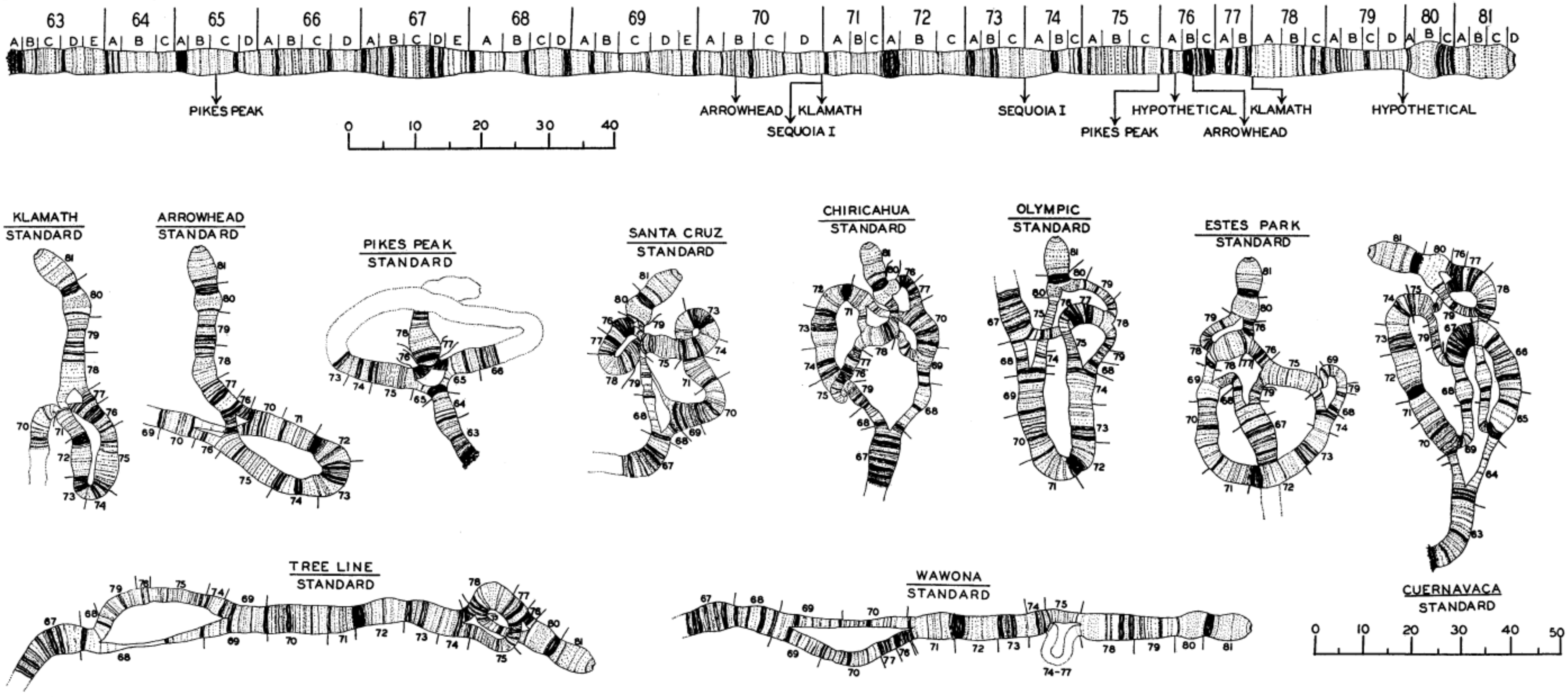
Theodosius Dobzhansky

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

Using Chromosomal Rearrangements as Markers…

Dobzhansky and Sturtevant, Chromosomes of Drosophila Pseudoobscura



Genetics 23: 28 Jan. 1938

Dobzhansky & Sturtevant (1938) Genetics

...To Infer the History of Species

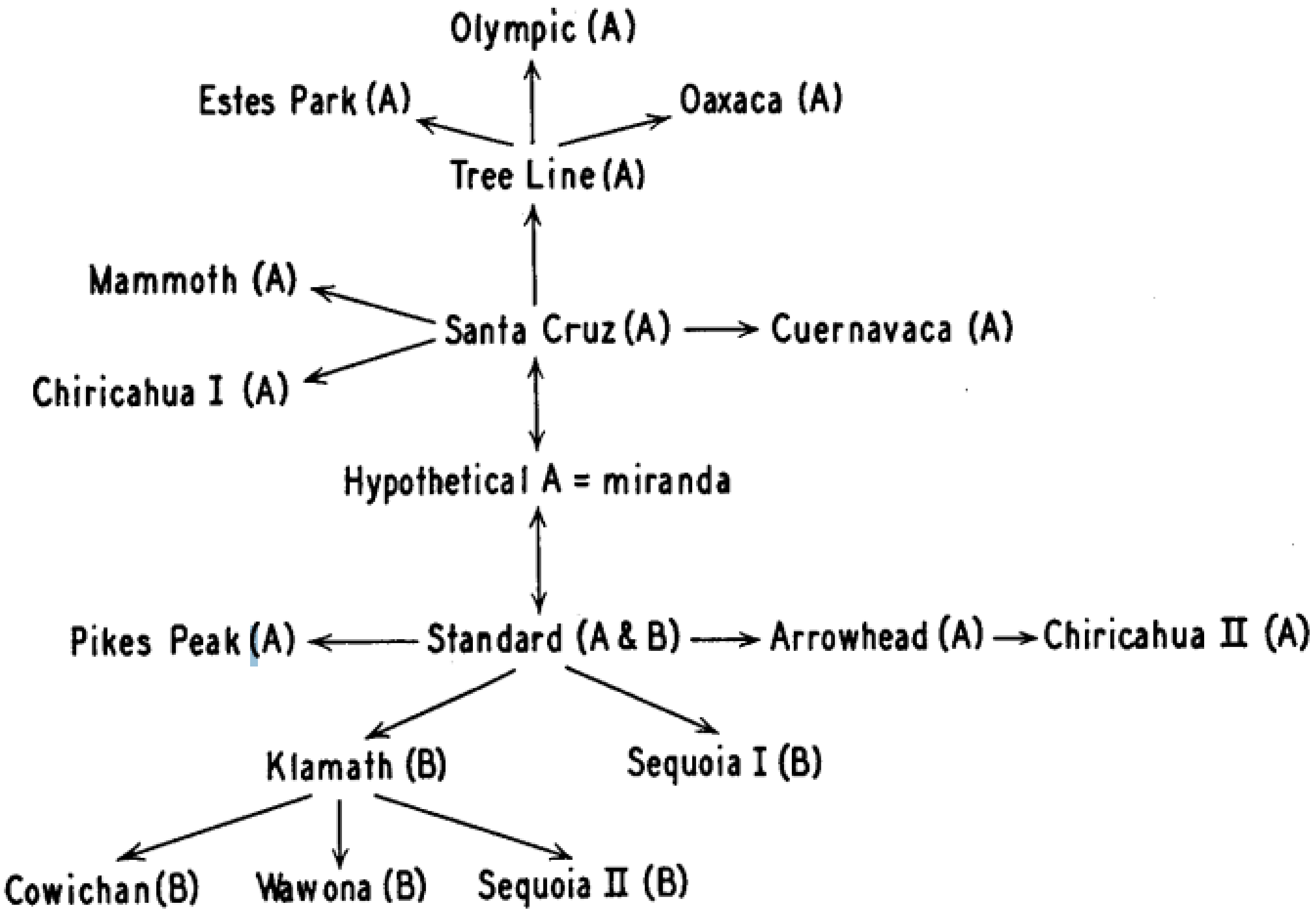
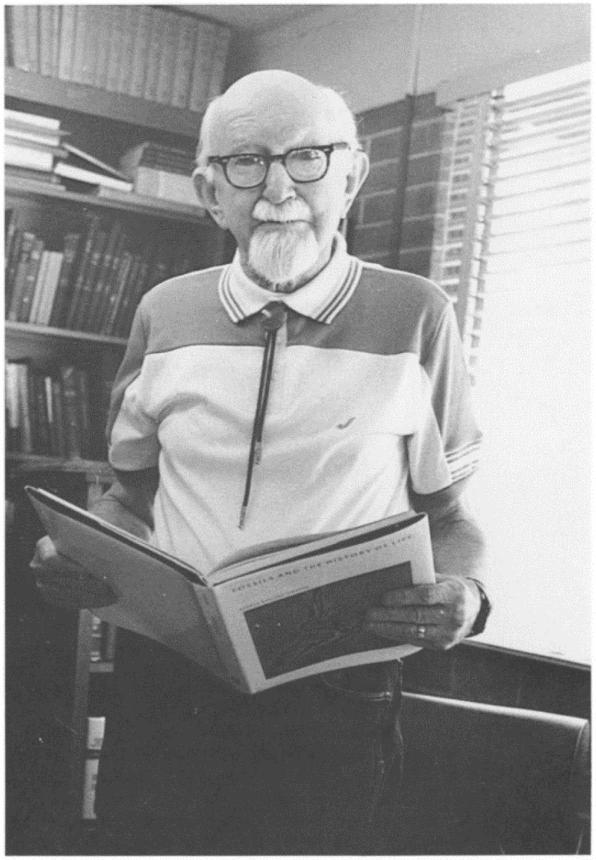
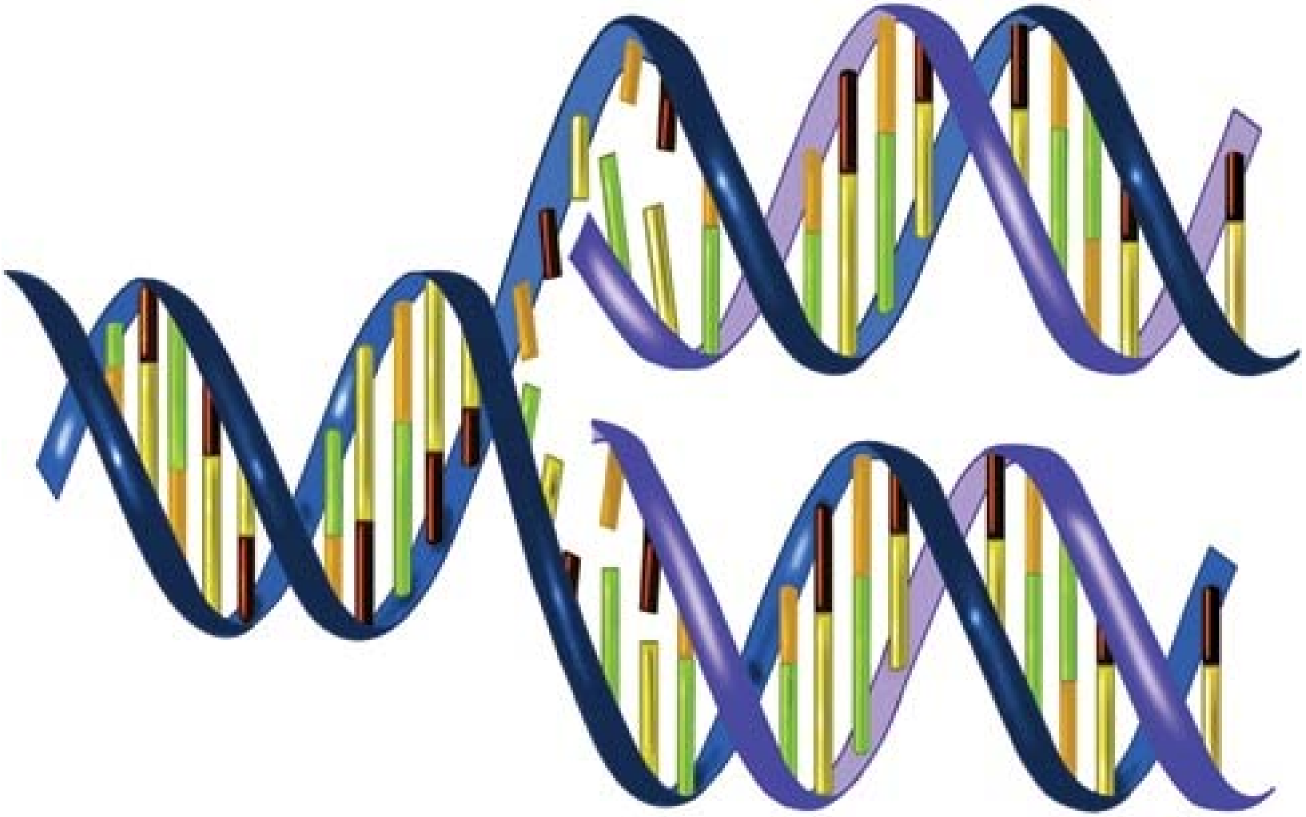


FIGURE 3.—Phylogeny of the gene arrangements in the third chromosome of *Drosophila pseudoobscura*. Any two arrangements connected by an arrow in the diagram differ by a single inversion. Further explanation in text.

Dobzhansky & Sturtevant (1938) Genetics



George Gaylord Simpson

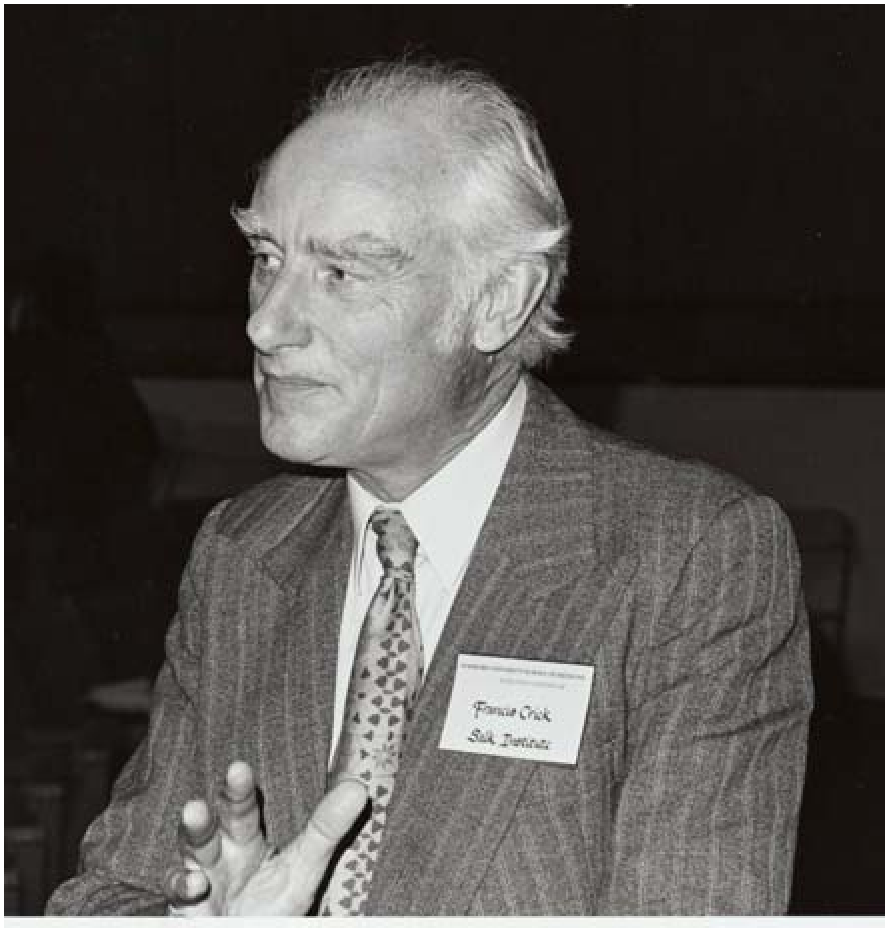


“The stream of heredity makes phylogeny; in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream”

G. G. Simpson, 1945

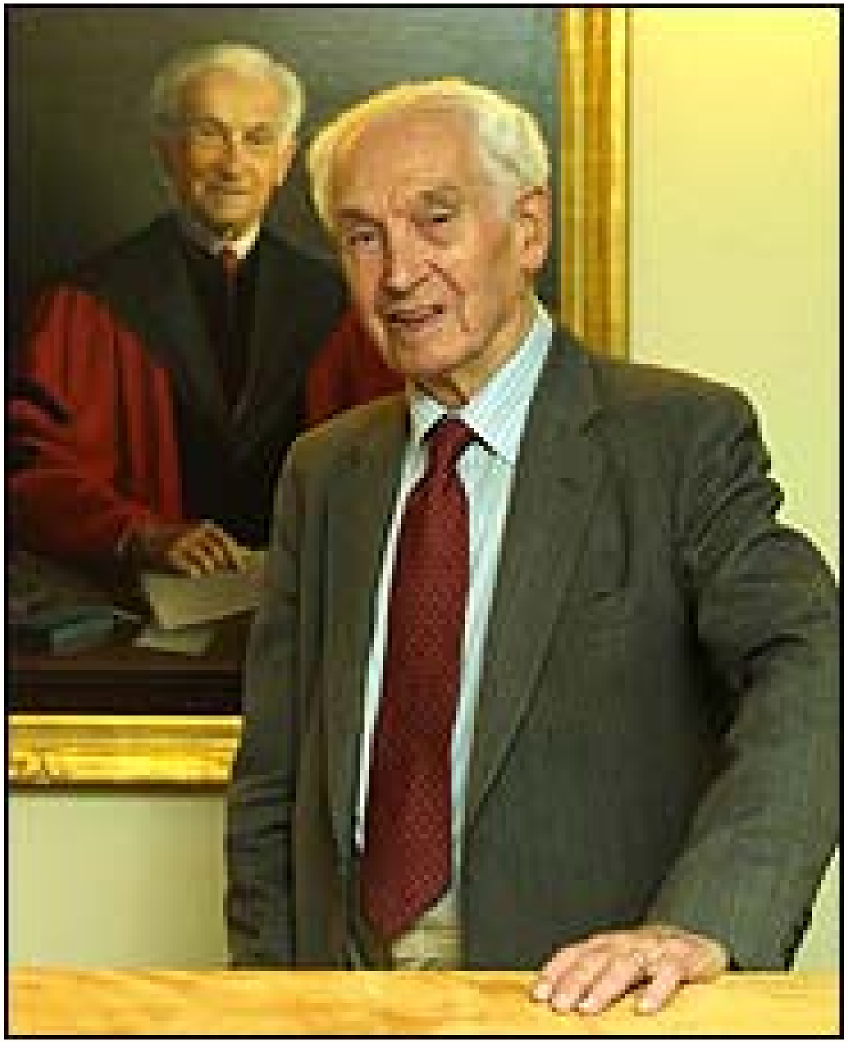
“...before long we shall have a subject which might be called “protein taxonomy”; the study of amino acid sequences of the proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism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

Francis Crick, 1957

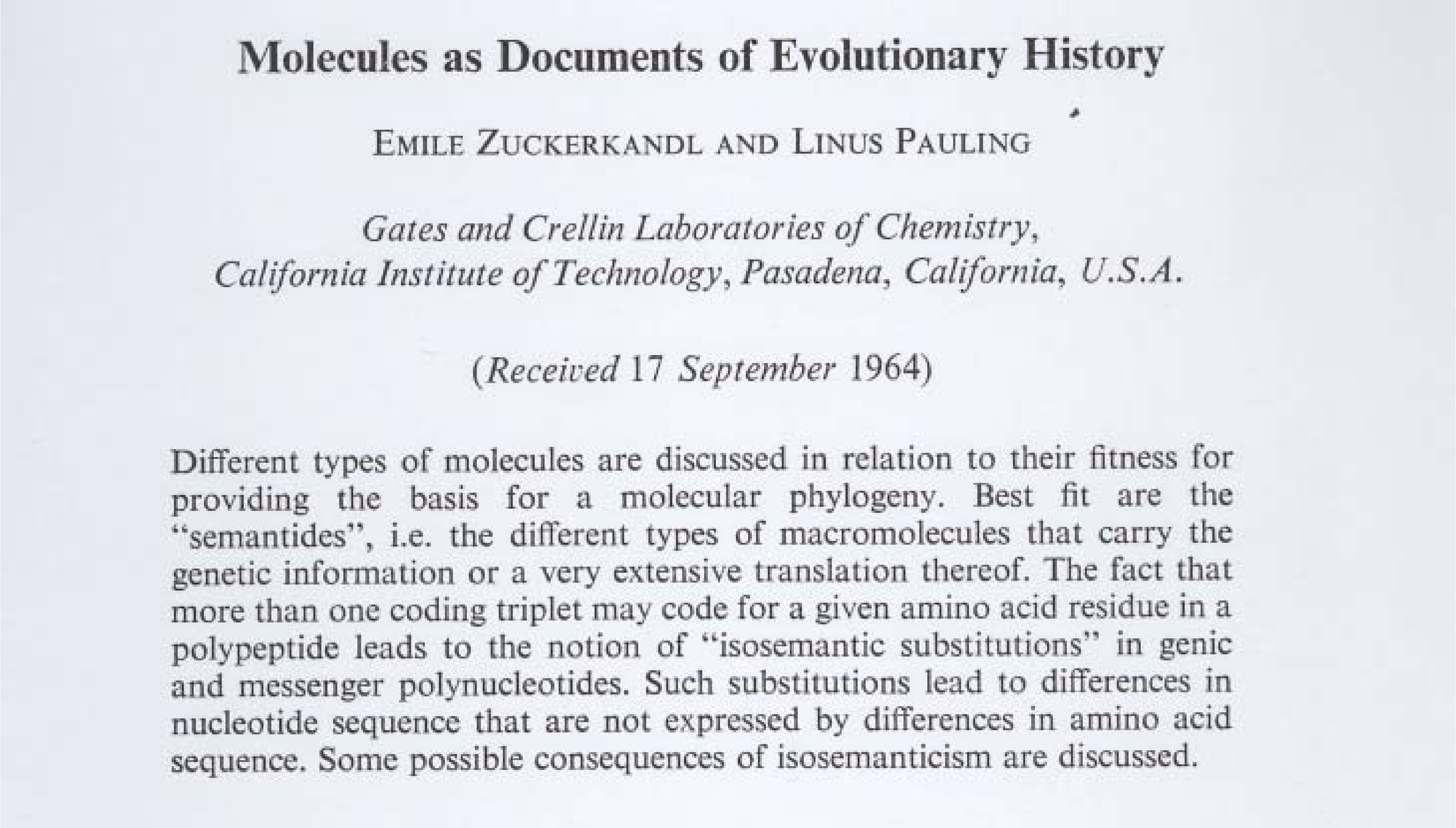


“…the search for homologous genes is quite futile except in very close relatives”

Ernst Mayr, 1963



DNA & Protein Sequences Record Evolutionary History



Molecules as Documents of Evolutionary History

EMILE ZUCKERKANDL AND LINUS PAULING

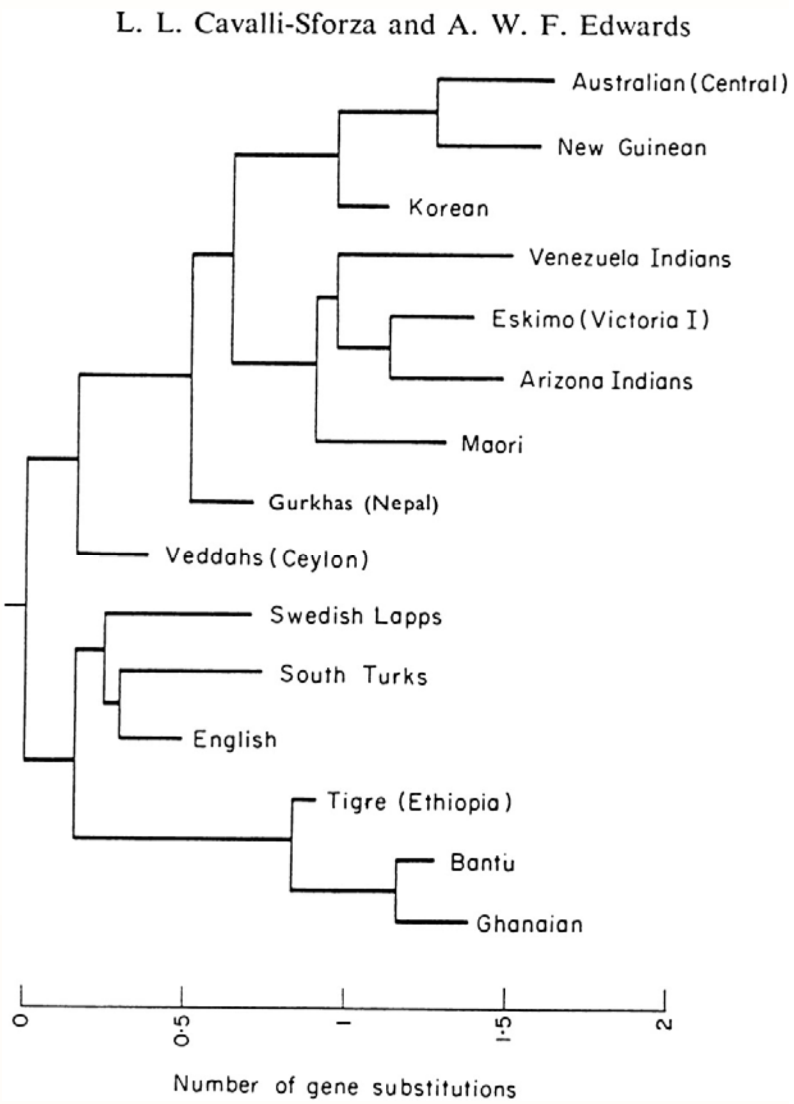
Gates and Crellin Laboratories of Chemistry,
California Institute of Technology, Pasadena, California, U.S.A.

(*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.

The Phylogeny of Human Populations

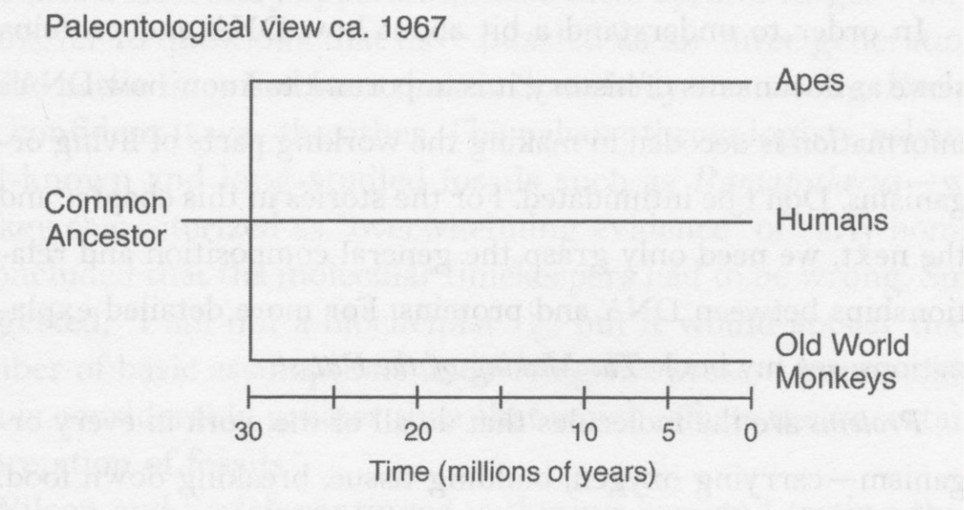


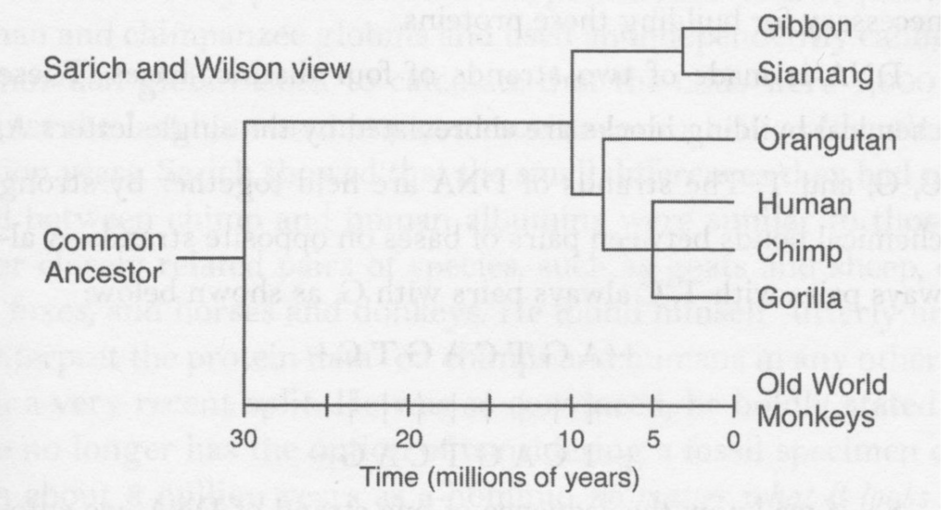
Phylogeny inferred from blood group allele frequencies from 15 populations

Cavalli-Sforza & Edwards (1965) in Genetics Today

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


Paleontological view ca. 1967


Sarich and Wilson view

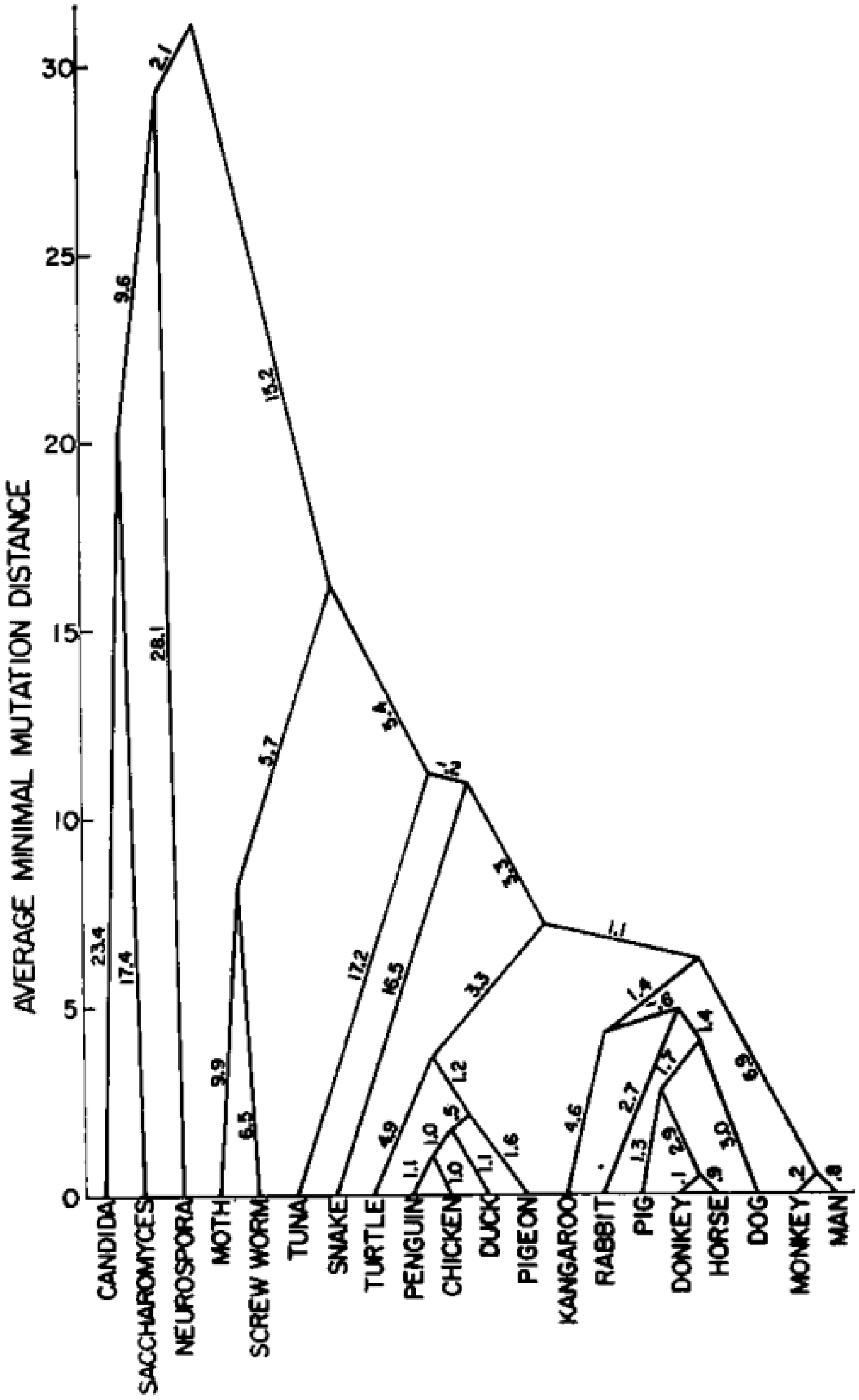
“no fuss, no muss, no dishpan hands. Just throw some proteins into a laboratory apparatus, shake them up, and bingo! – we have an answer to questions that have puzzled us for three generations.”

Sarich & Wilson (1967) Science

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

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

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

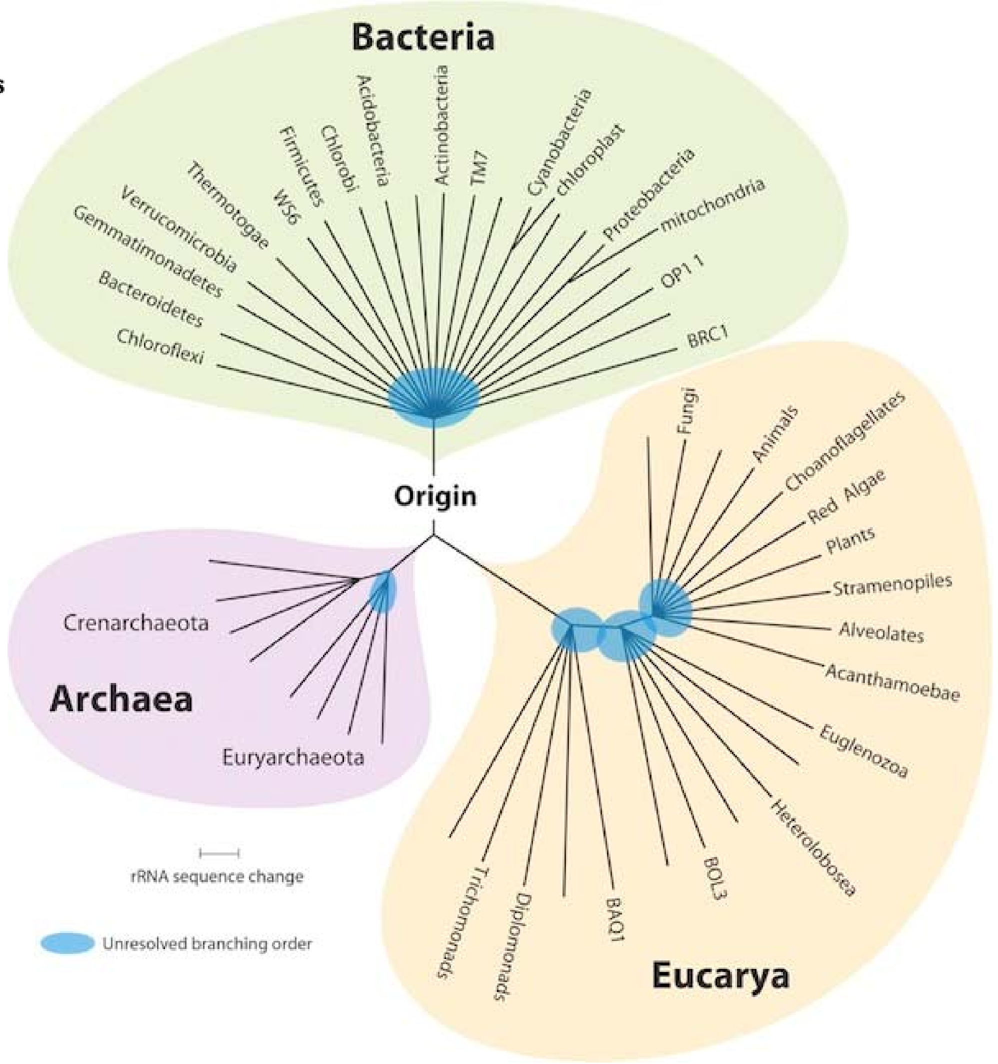
(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

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.



~~Gene tree ≈ Species phylogeny~~

Gene tree ≠ Species phylogeny

Two Types of Factors Influence the Relationship

Analytical factors

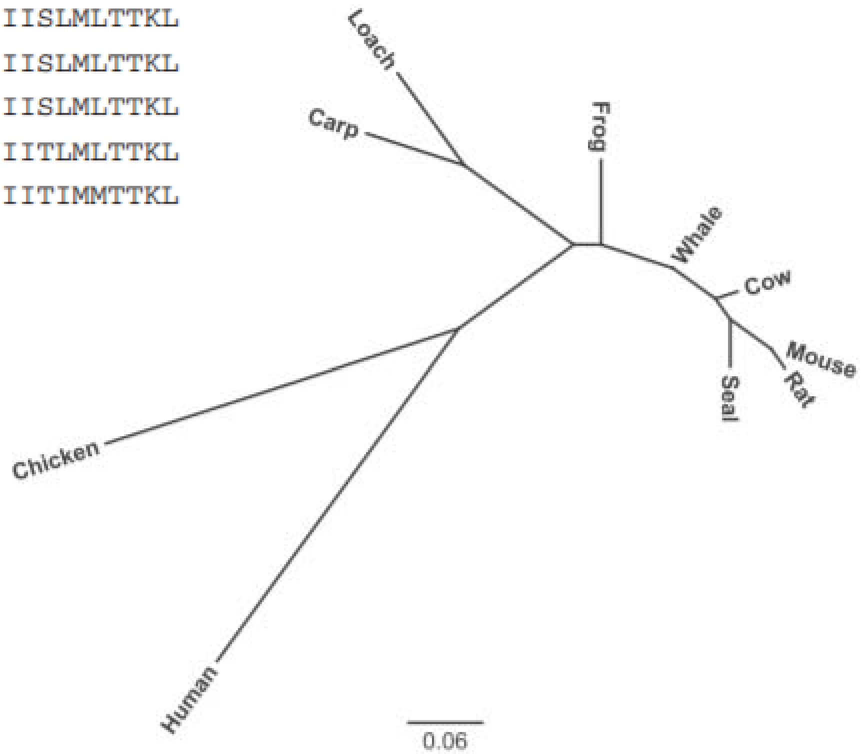
They lead to failure in accurately inferring a gene tree; these can be either due to **stochastic error** (e.g., insufficient sequence length or taxon samples) or due to **systematic error** (e.g., observed data far depart from model assumptions)

Biological factors

They lead to gene trees that are topologically distinct from each other and from the species tree. Known factors include **stochastic lineage sorting**, **hidden paralogy**, **horizontal gene transfer**, **recombination** and **natural selection**

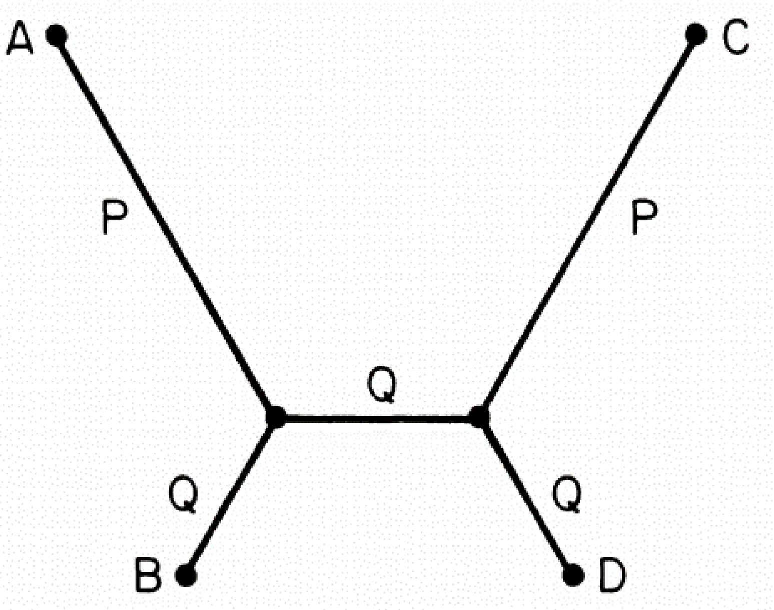
Sampling Error

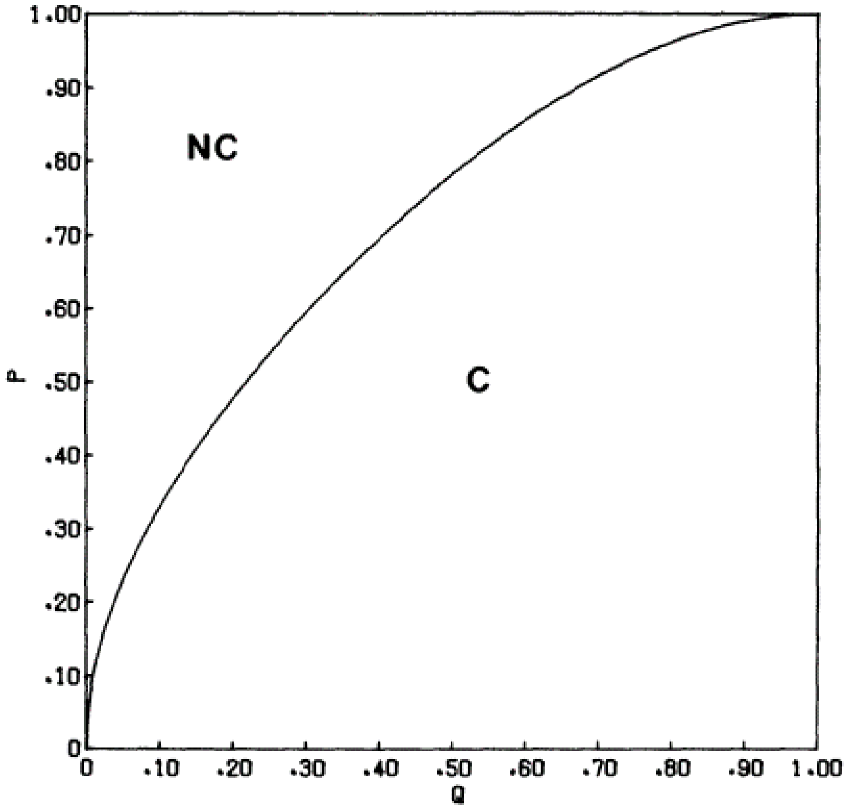
```
10 50
Cow         MAYPMQLGFQ DATSPIMEEL LHFHDHTLMI VFLISSLVLY IISLMLTTKL
Carp        MAHPTQLGFK DAAMPVMEEL LHFHDHALMI VLLISTLVLY IITAMVSTKL
Chicken     MANHSQLGFQ DASSPIMEEL VEFHDHALMV ALAICSLVLY LLTLMLMEKL
Human       MAHAAQVGLQ DATSPIMEEL ITFHDHALMI IFLICFLVLY ALFLTLTTKL
Loach       MAHPTQLGFQ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

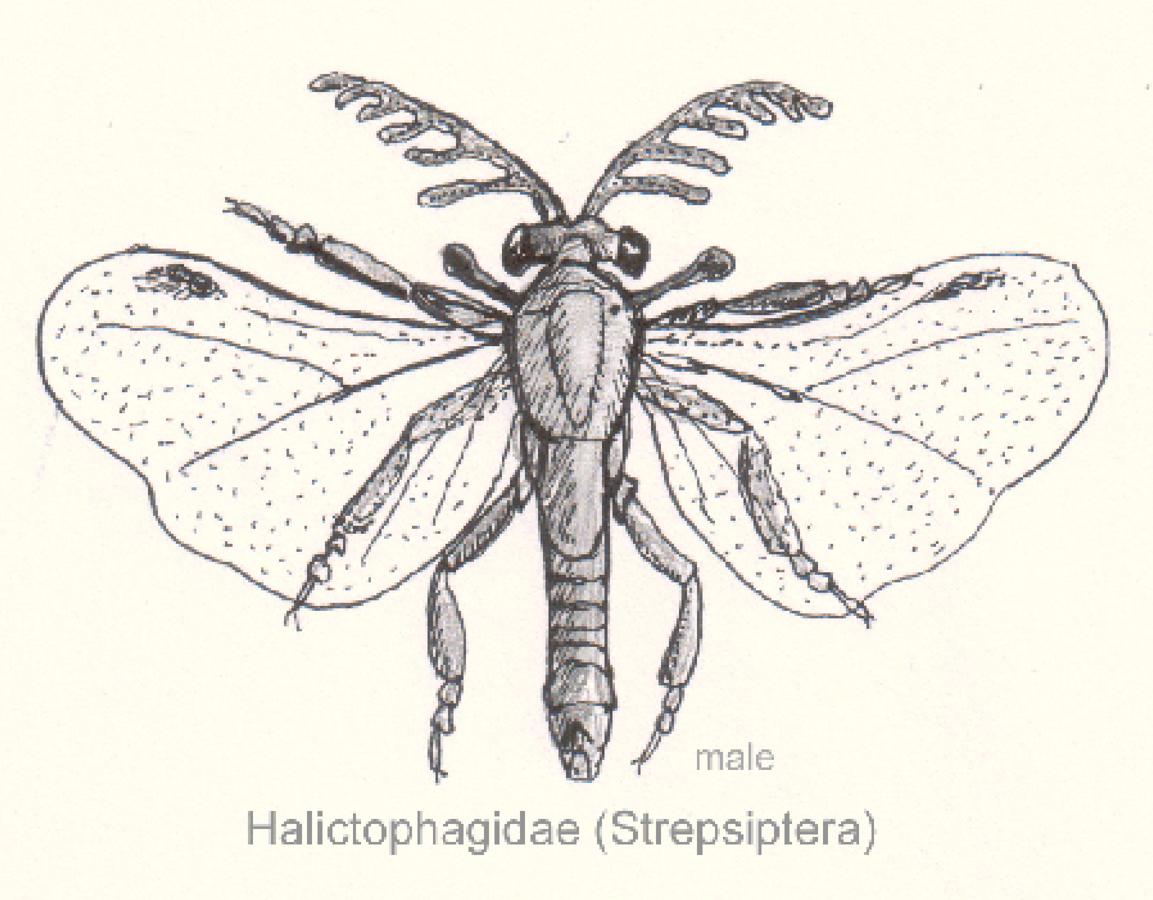




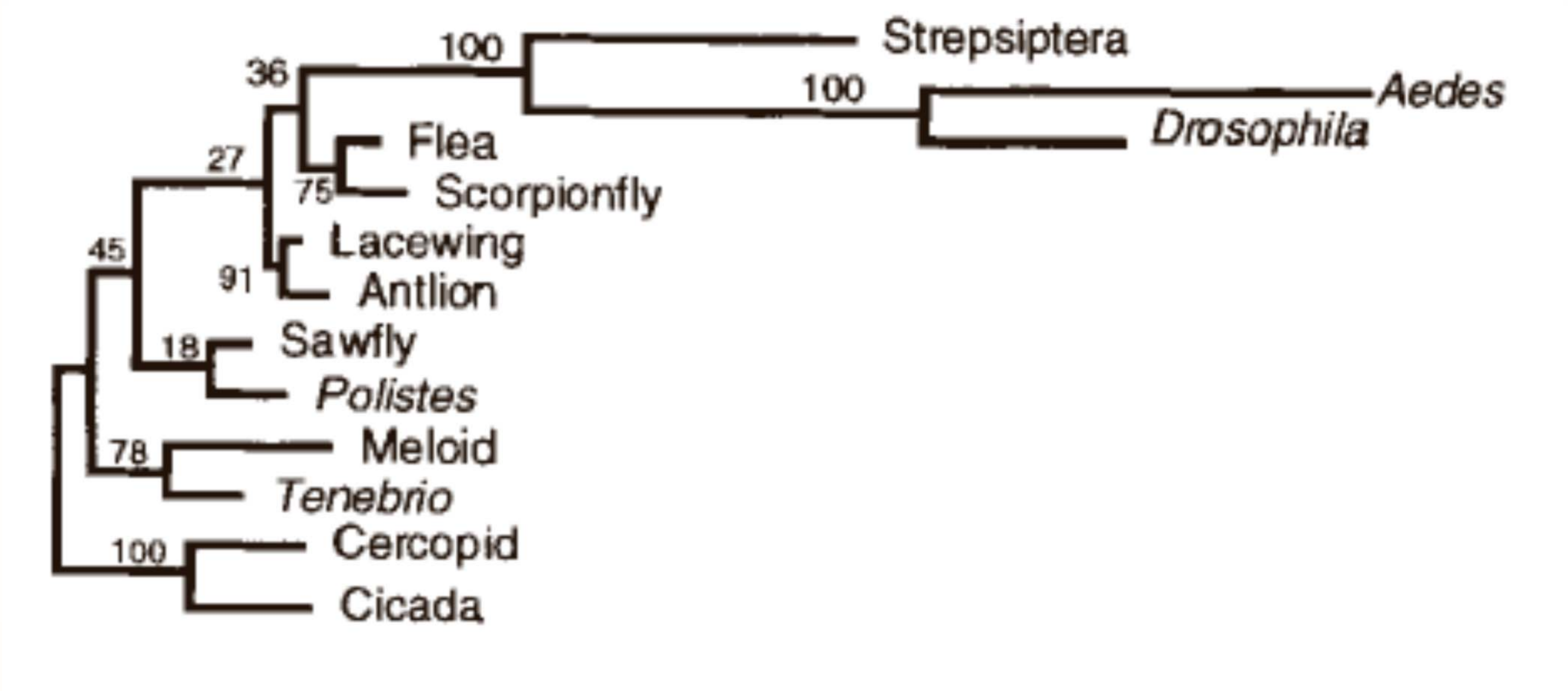
Long branch attraction

Felsenstein (1978) Syst. Zool.

“The Strepsiptera Problem” is the Classic Example of LBA

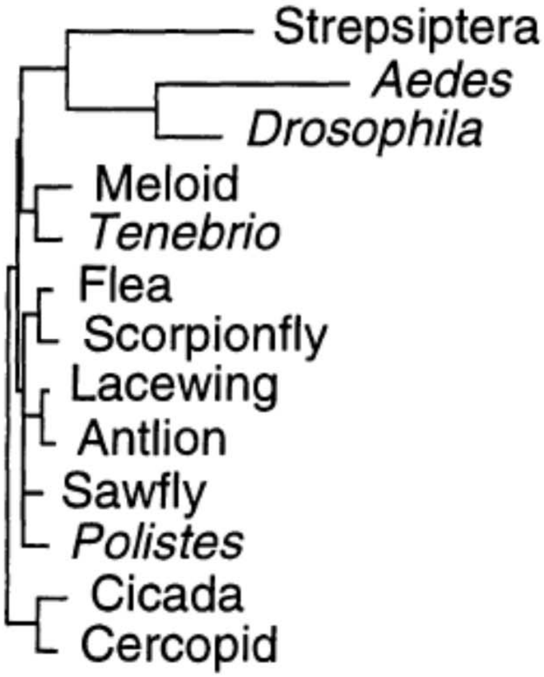


The Strepsiptera Problem

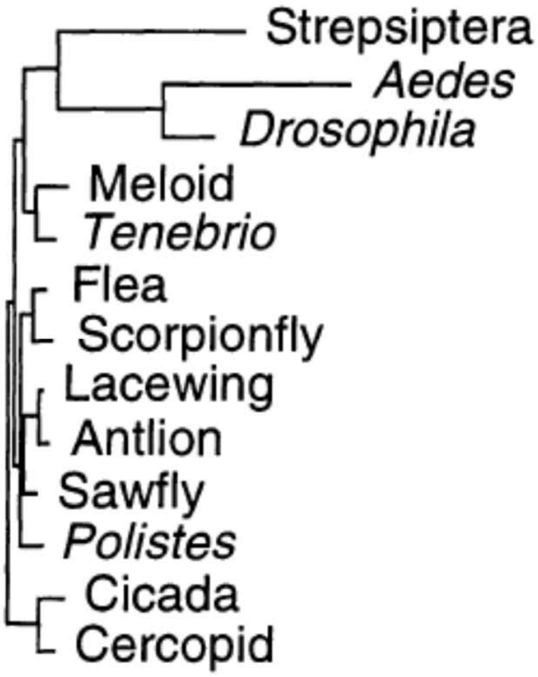


Carmean & Crespi (1995) Nature

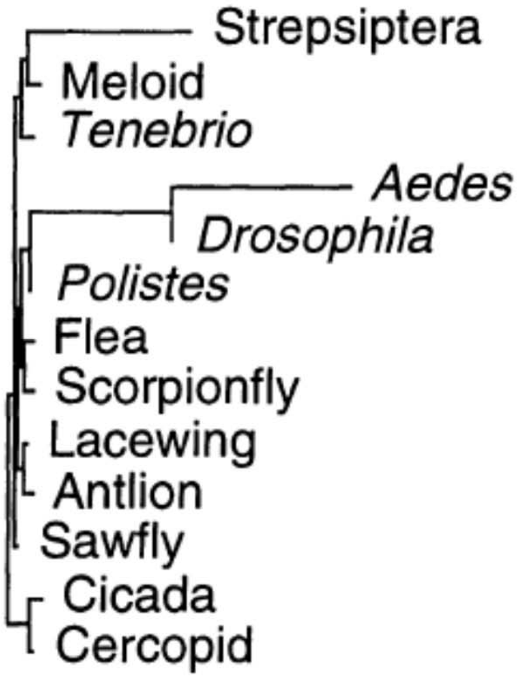
The Strepsiptera Problem



p distance



HKY85

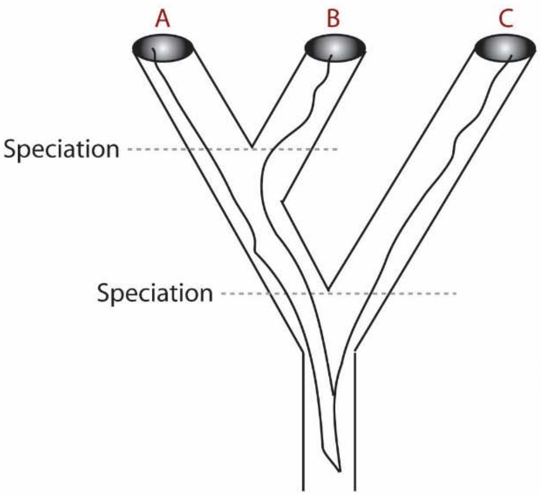


HKY85+GAMMA

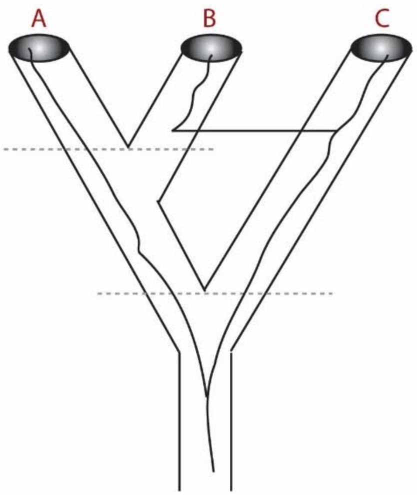
Huelsenbeck (1997) Syst. Biol.

Biological Factors

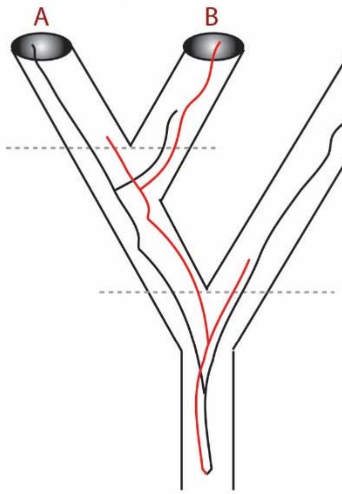
Lineage Sorting



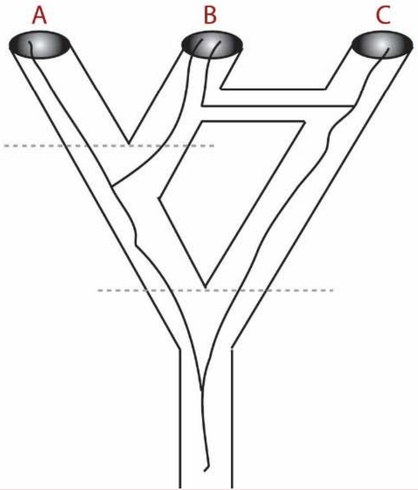
Horizontal Gene Transfer



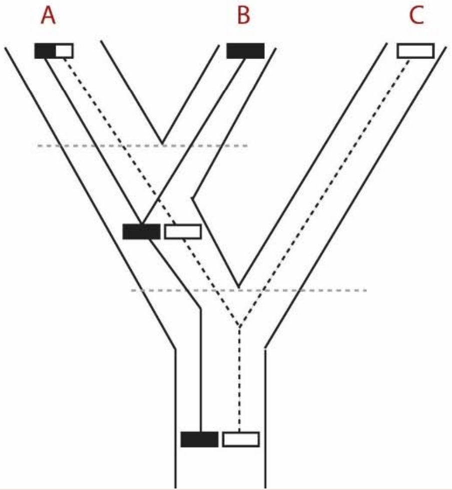
Gene Duplication and Loss



Hybridization



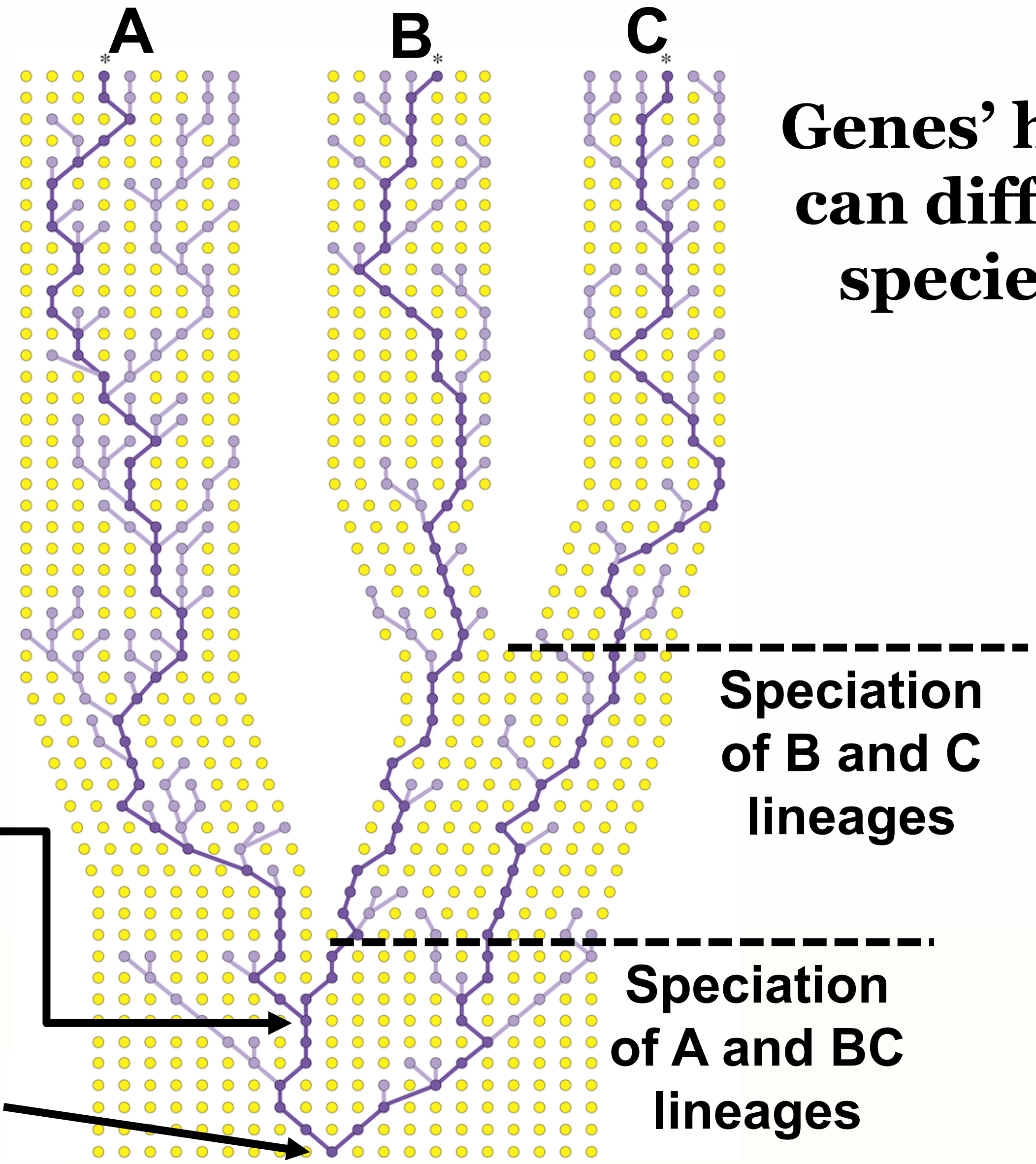
Recombination



Degnan & Rosenberg (2009) Trends Ecol. Evol.

Stochastic Lineage Sorting of Ancestral Polymorphisms

Genes' histories can differ from species ones



Speciation of B and C lineages

Coalescence of A and B alleles

Speciation of A and BC lineages

Coalescence 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%)

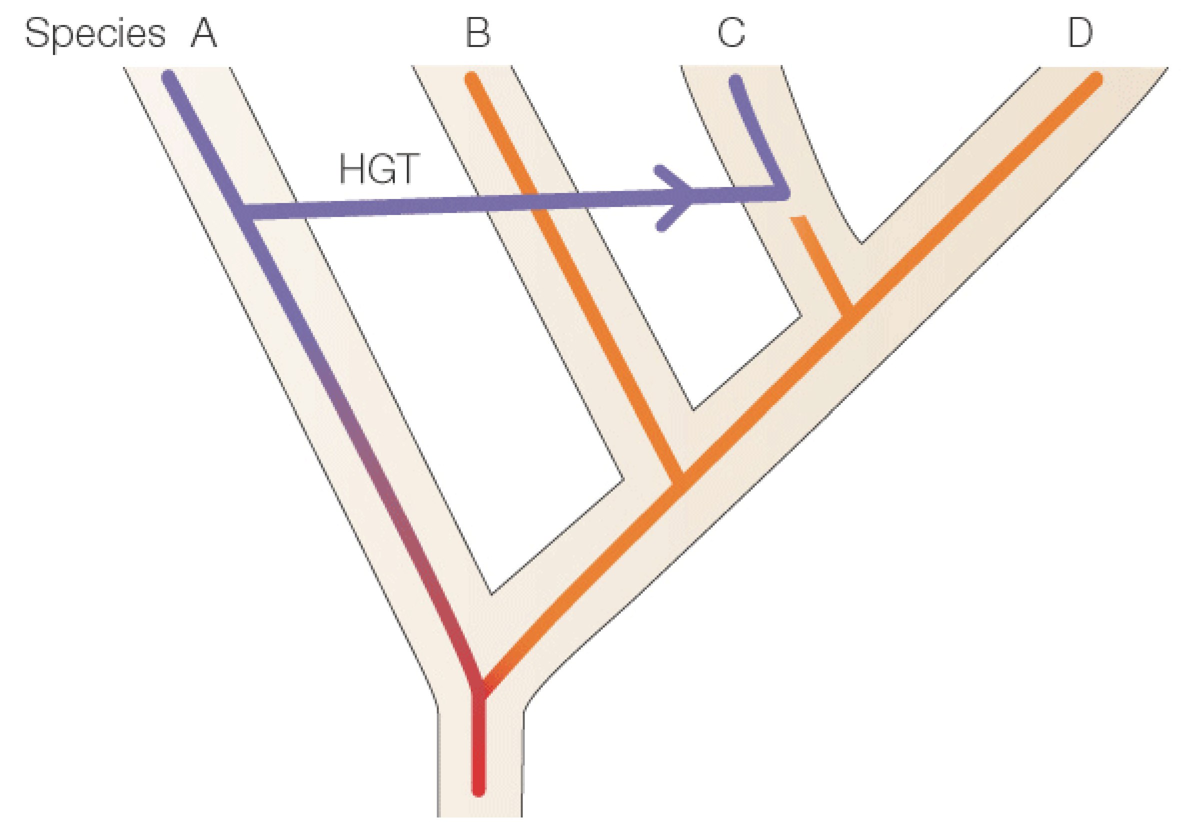
1,430 / 11,293

(~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 Gene Transfer in Fungi

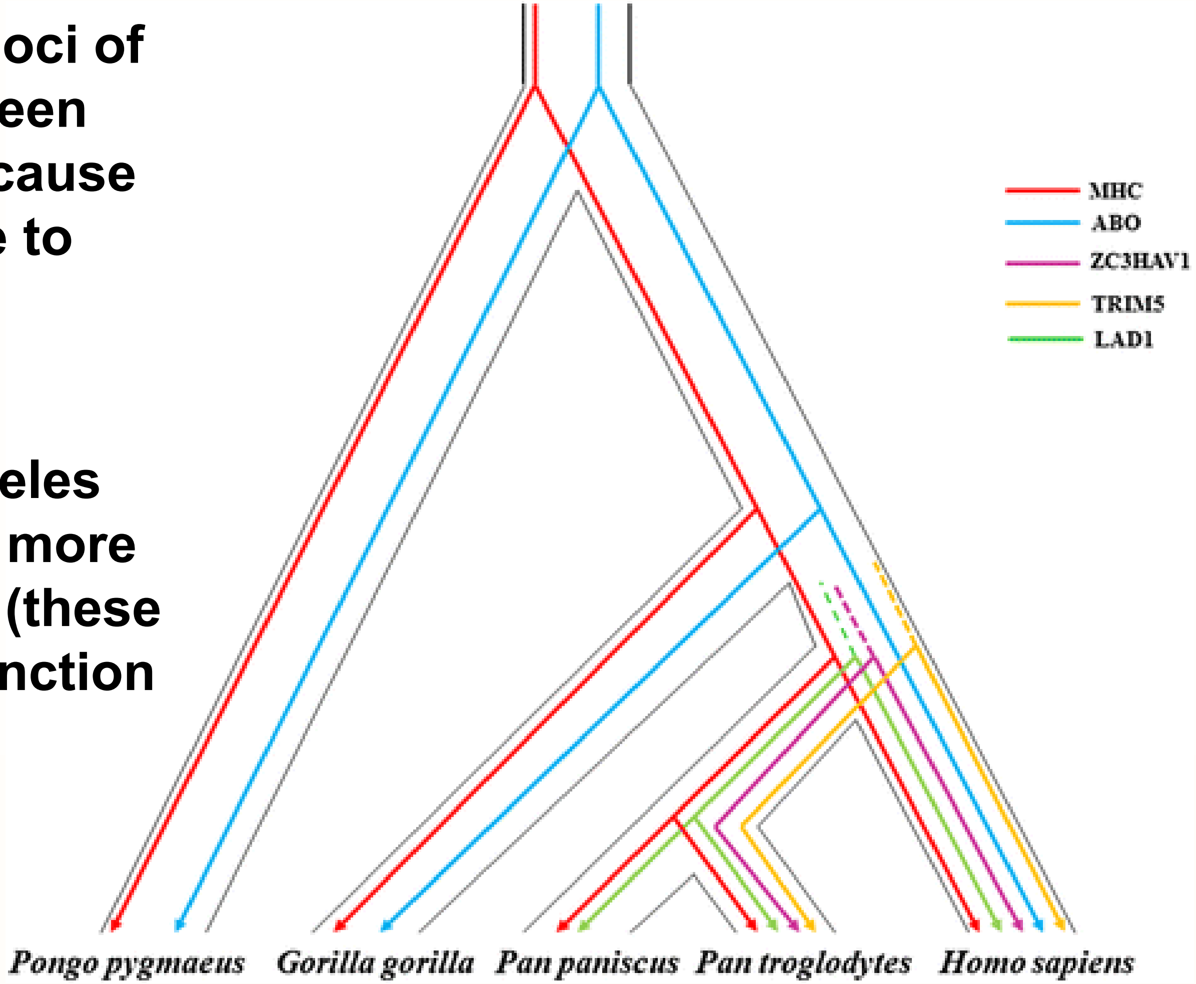
Slot & Rokas (2010) PNAS

Balancing Selection

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

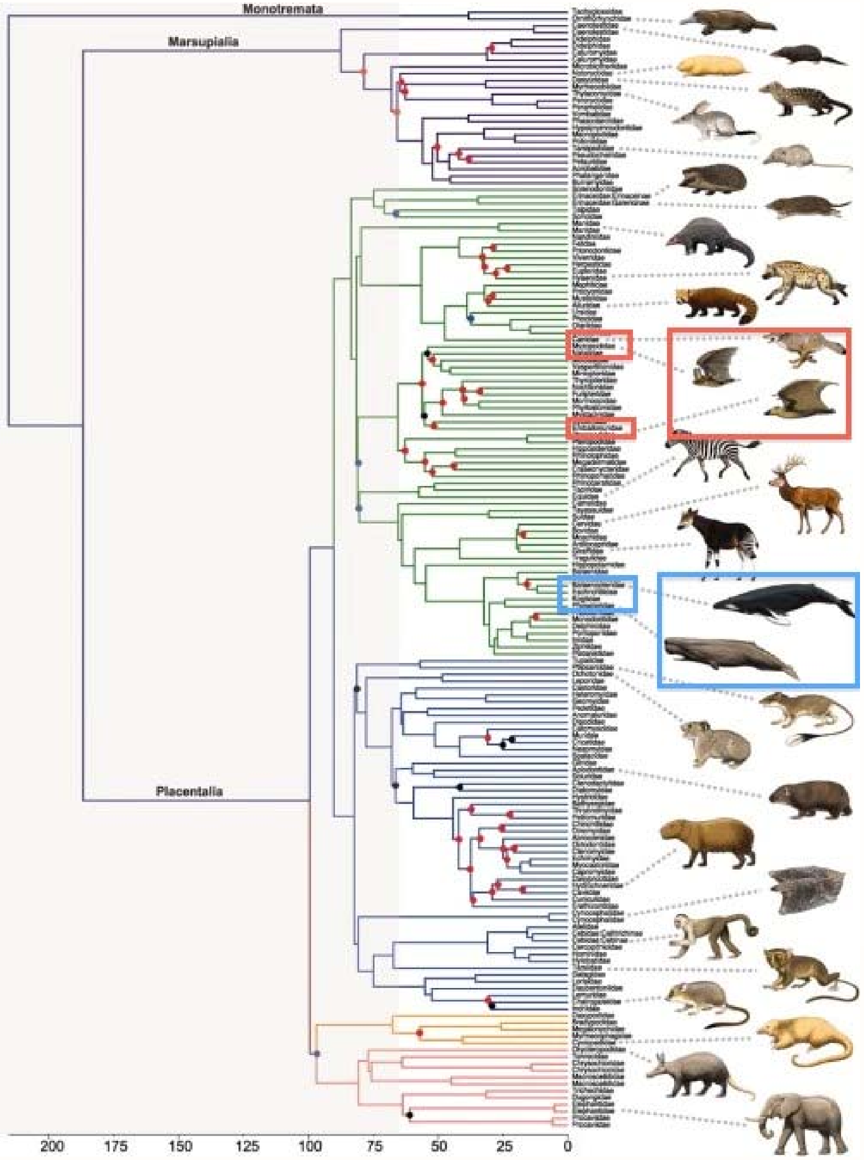
Best example: alleles at loci of the MHC – they have been retained by selection because they confer resistance to infection

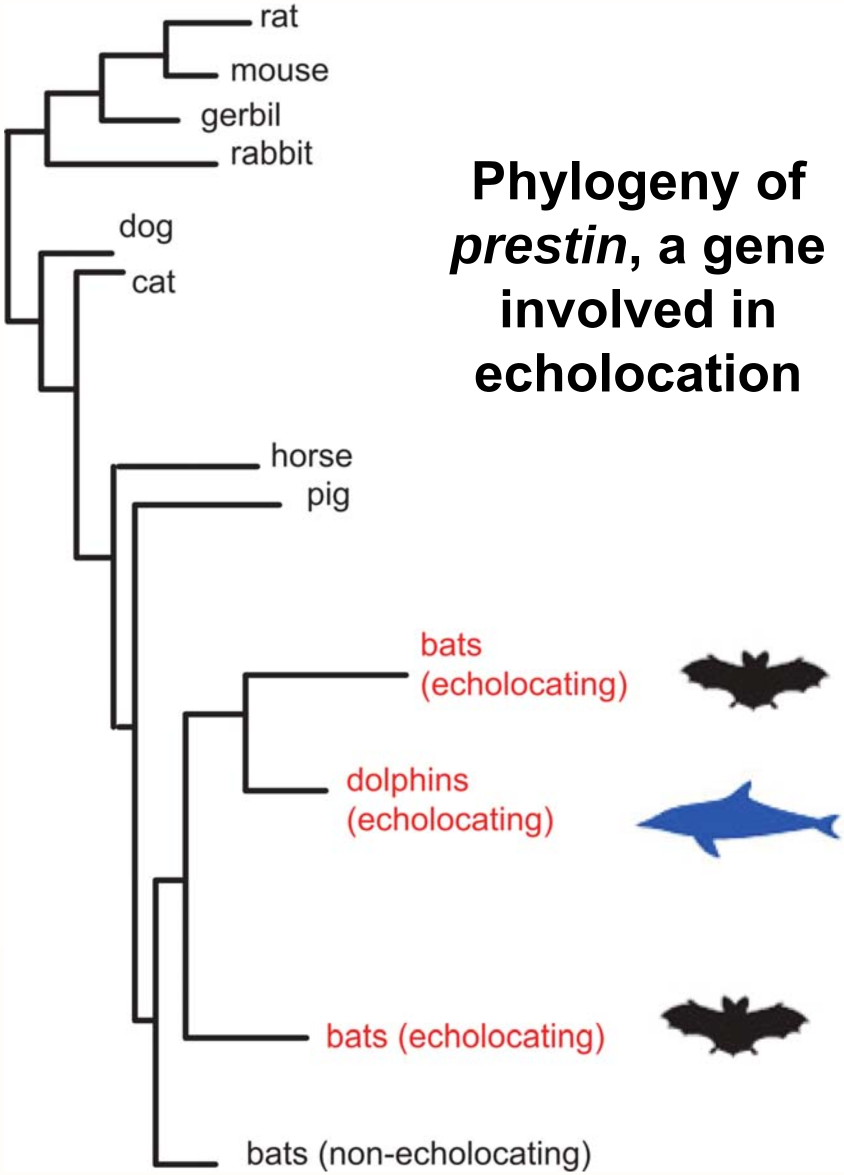
Certain human MHC alleles appear to have diverged more than 65 million years ago (these alleles witnessed the extinction of dinosaurs!!!)



Azevedo et al. (2015) Human Genomics

Positive Selection

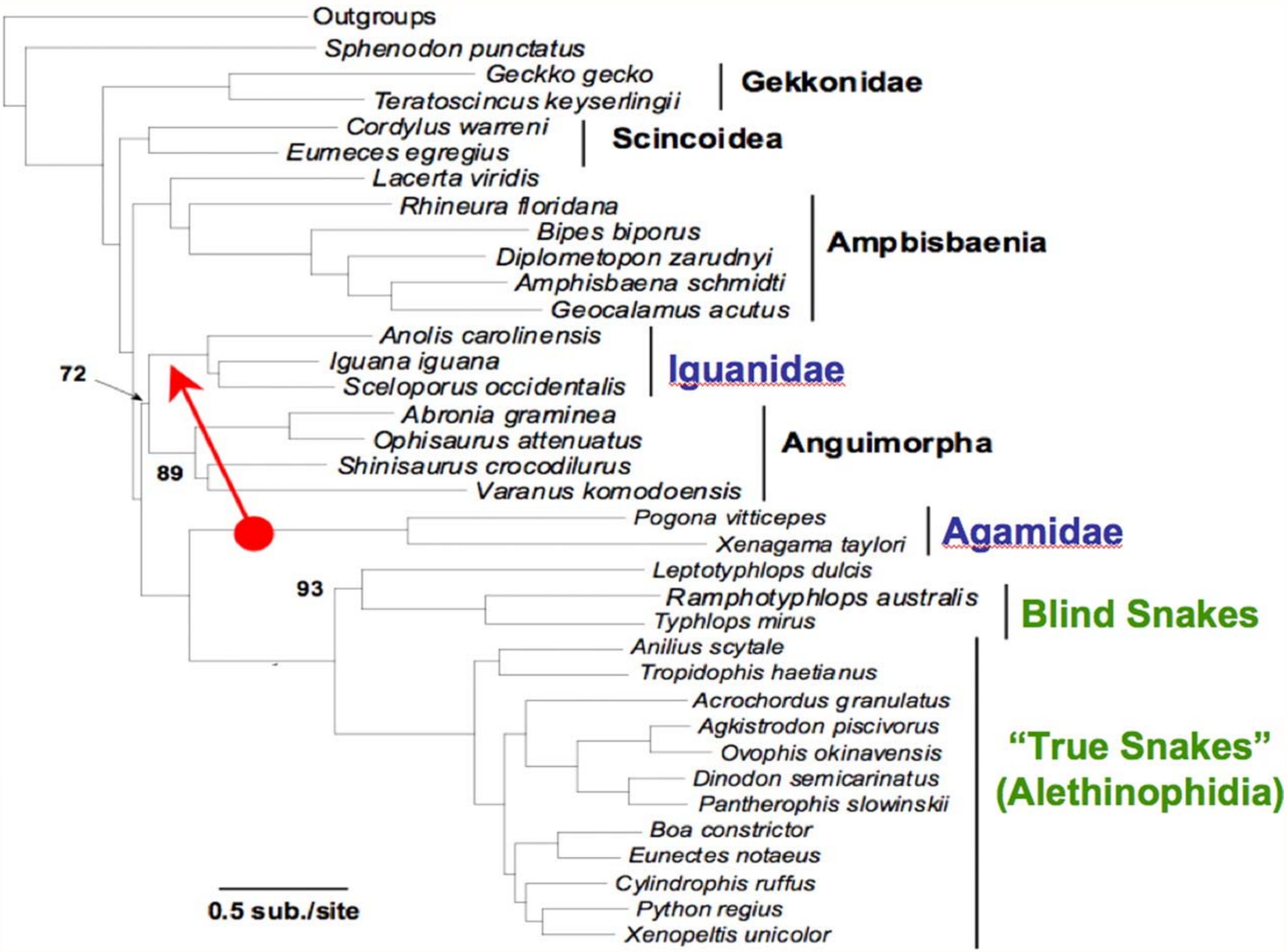




Phylogeny of *prestin*, a gene involved in echolocation

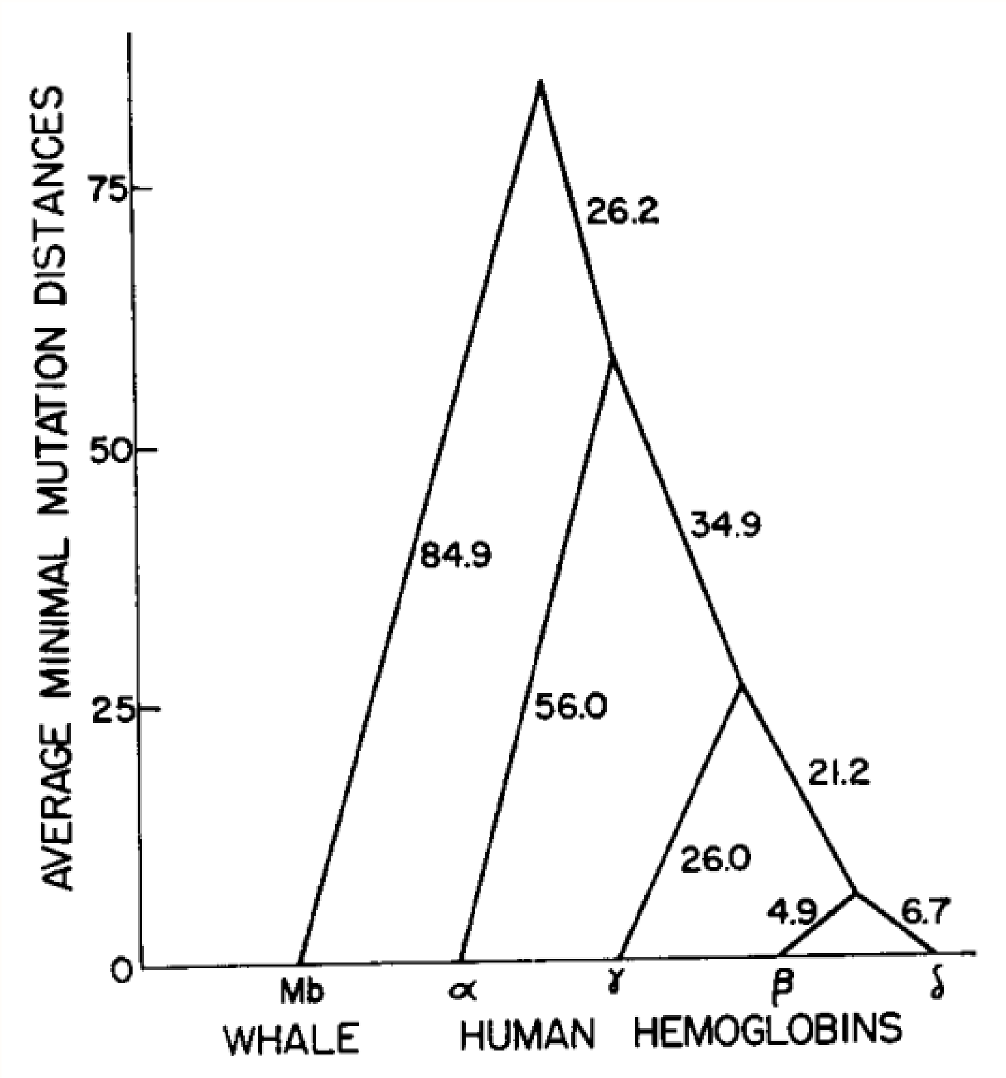
Li et al. (2010) Curr. Biol.

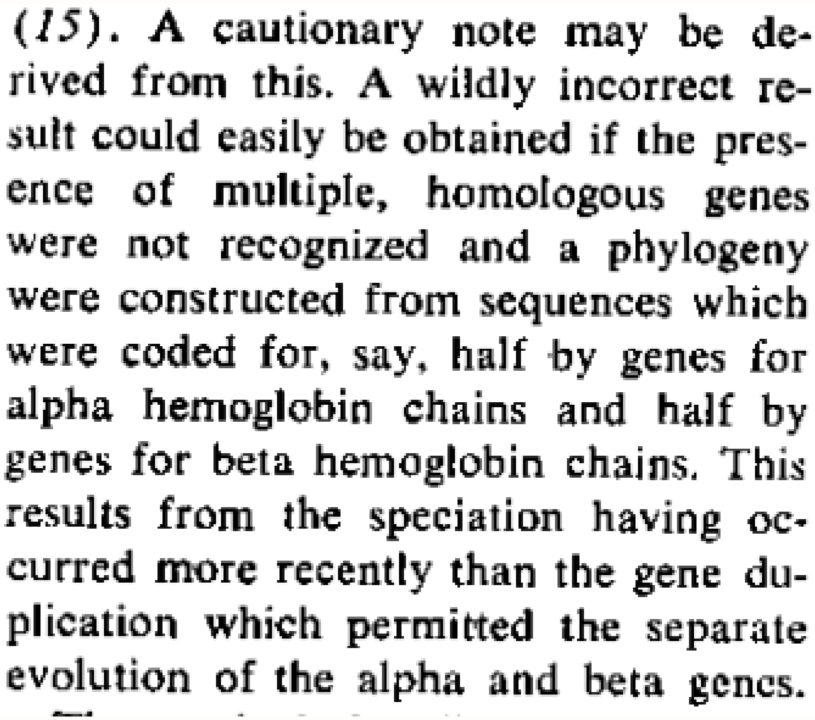
Positive Selection



Castoe et al. (2009) PNAS

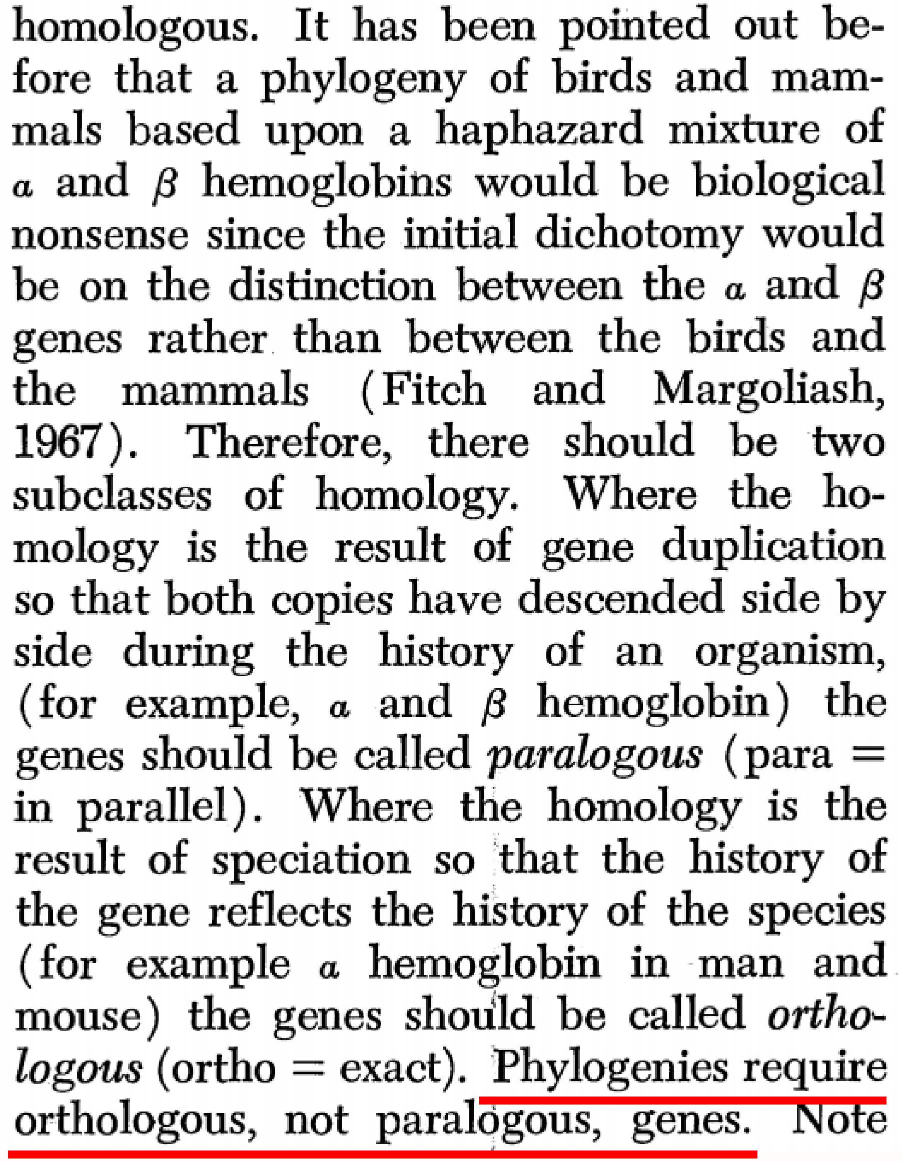
Gene Duplication Can Confound Phylogeny




(*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.

Fitch & Margoliash (1967) Science

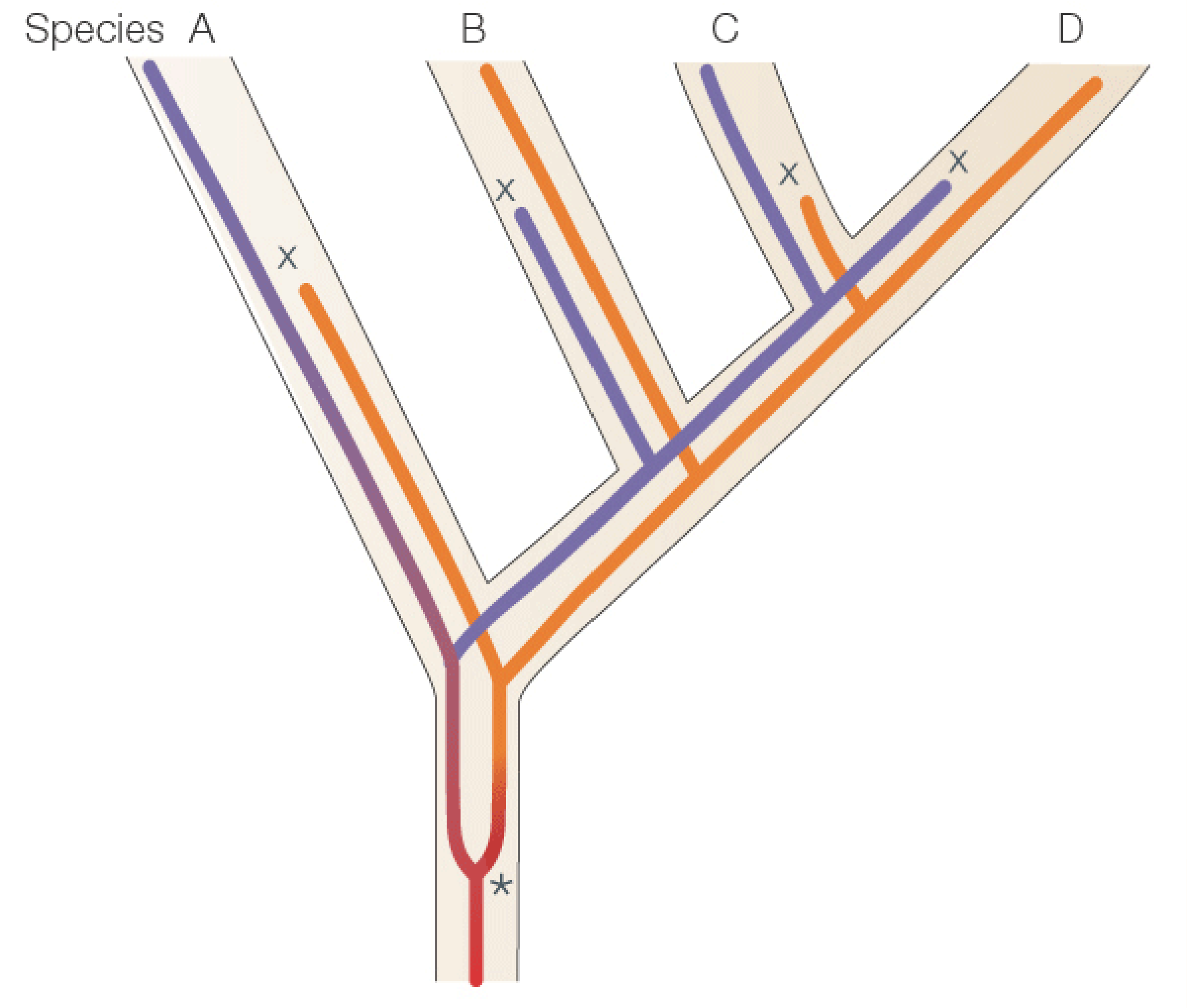
Gene Duplication Can Confound Phylogeny



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

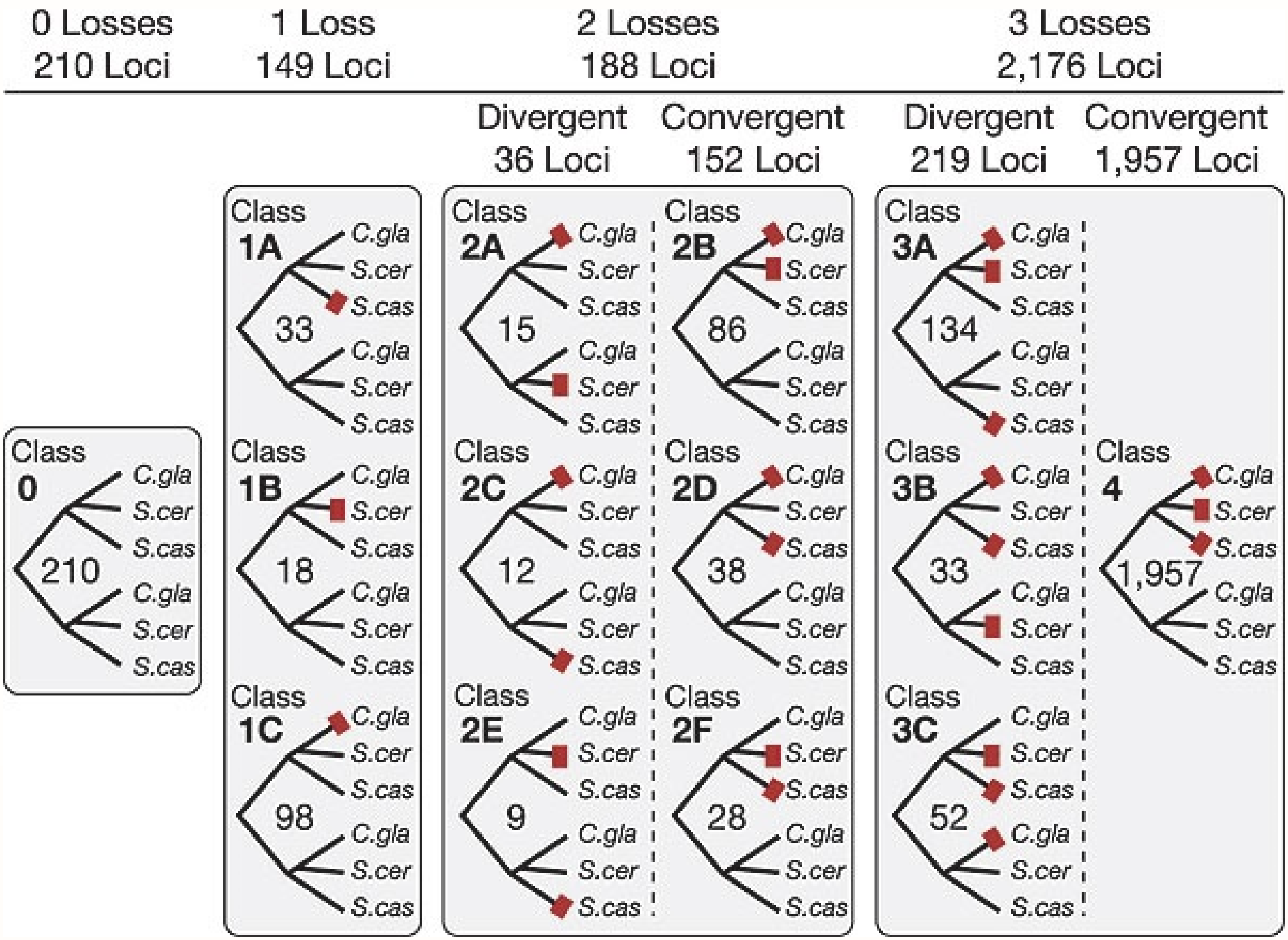
Fitch (1970) Syst. Zool.

Gene Duplication and Loss



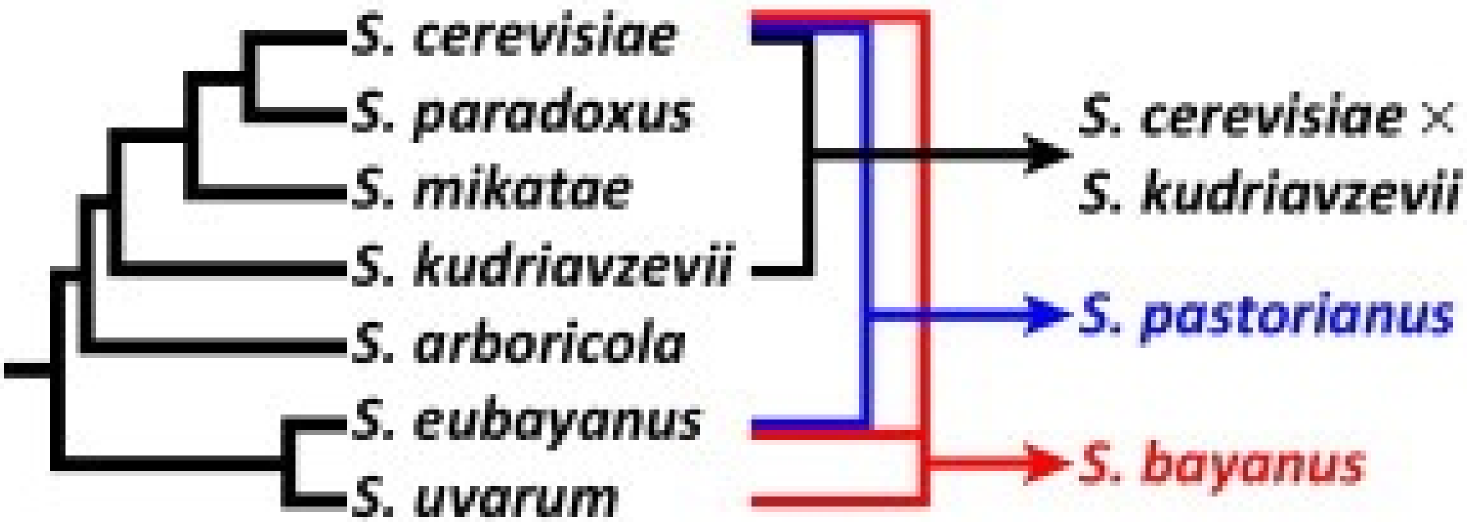
Gogarten & Townsend (2005) Nature Rev. Genet.

Gene Duplication and Loss



Scannell et al. (2006) Nature

Hybridization / Introgression



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

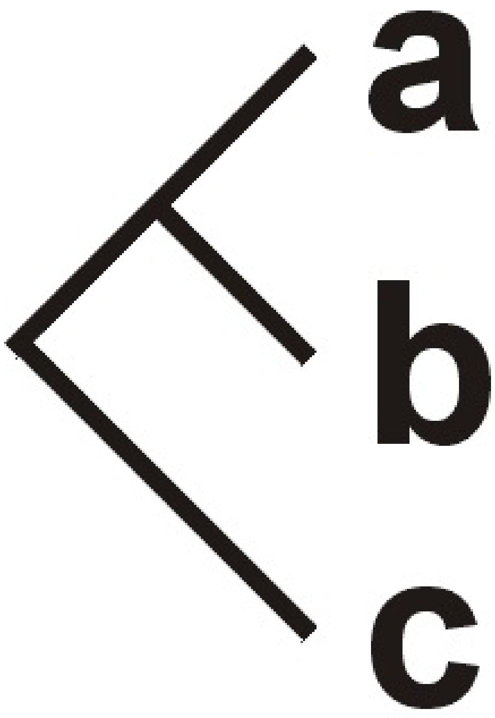
S. cerevisiae – S. paradoxus **divergence ≈ human – mouse divergence**
S. cerevisiae – S. uvarum **divergence ≈ human – chicken divergence**

Hittinger (2013) Trends Genet.

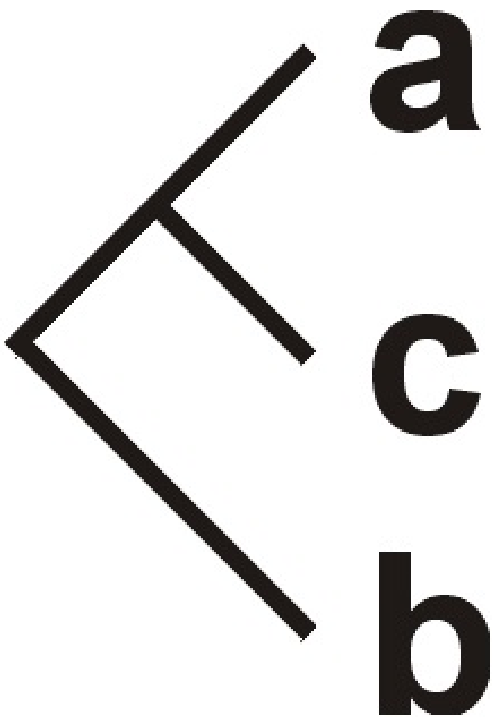
OK, I now get why gene trees ≠ species phylogenies

What does this have to do with phylogenomics?

All this Manifests Itself as Incongruence



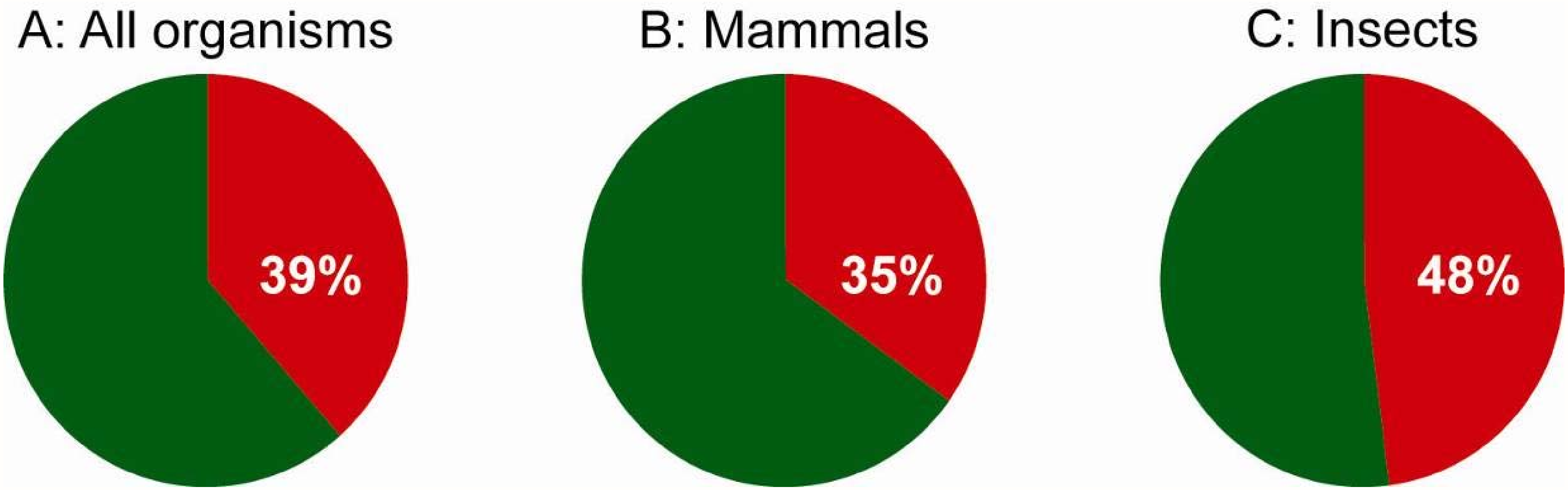
Gene X



Gene Y

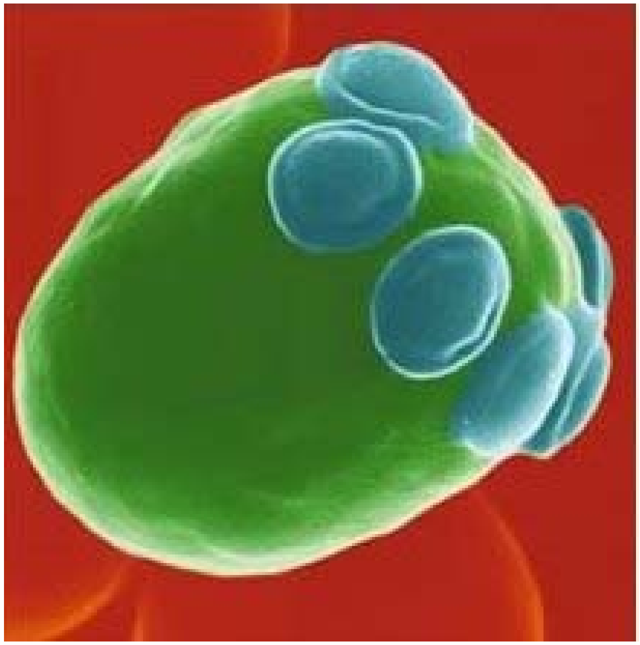
Species phylogeny?

Incongruence is Pervasive in the Phylogenetics Literature



Rokas & Chatzimanolis (2008) in Phylogenomics (W. J. Murphy, Ed.)

A Systematic Evaluation of Single Gene Phylogenies



S. cerevisiae

S. paradoxus

S. mikatae

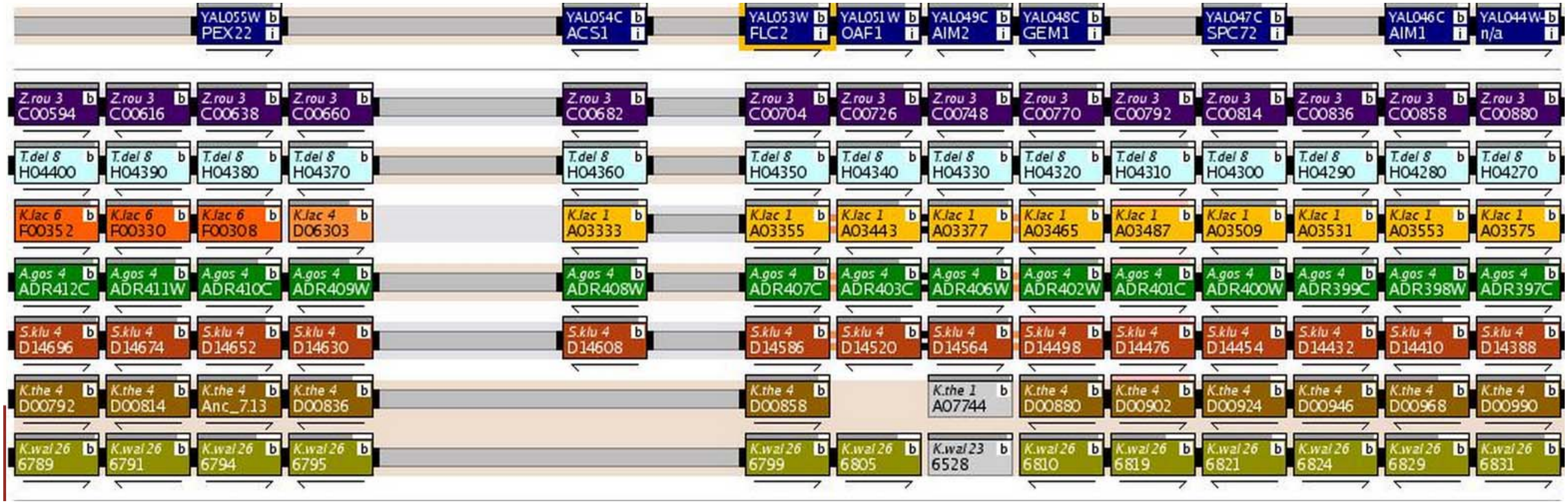
S. kudriavzevii

S. bayanus

S. castellii

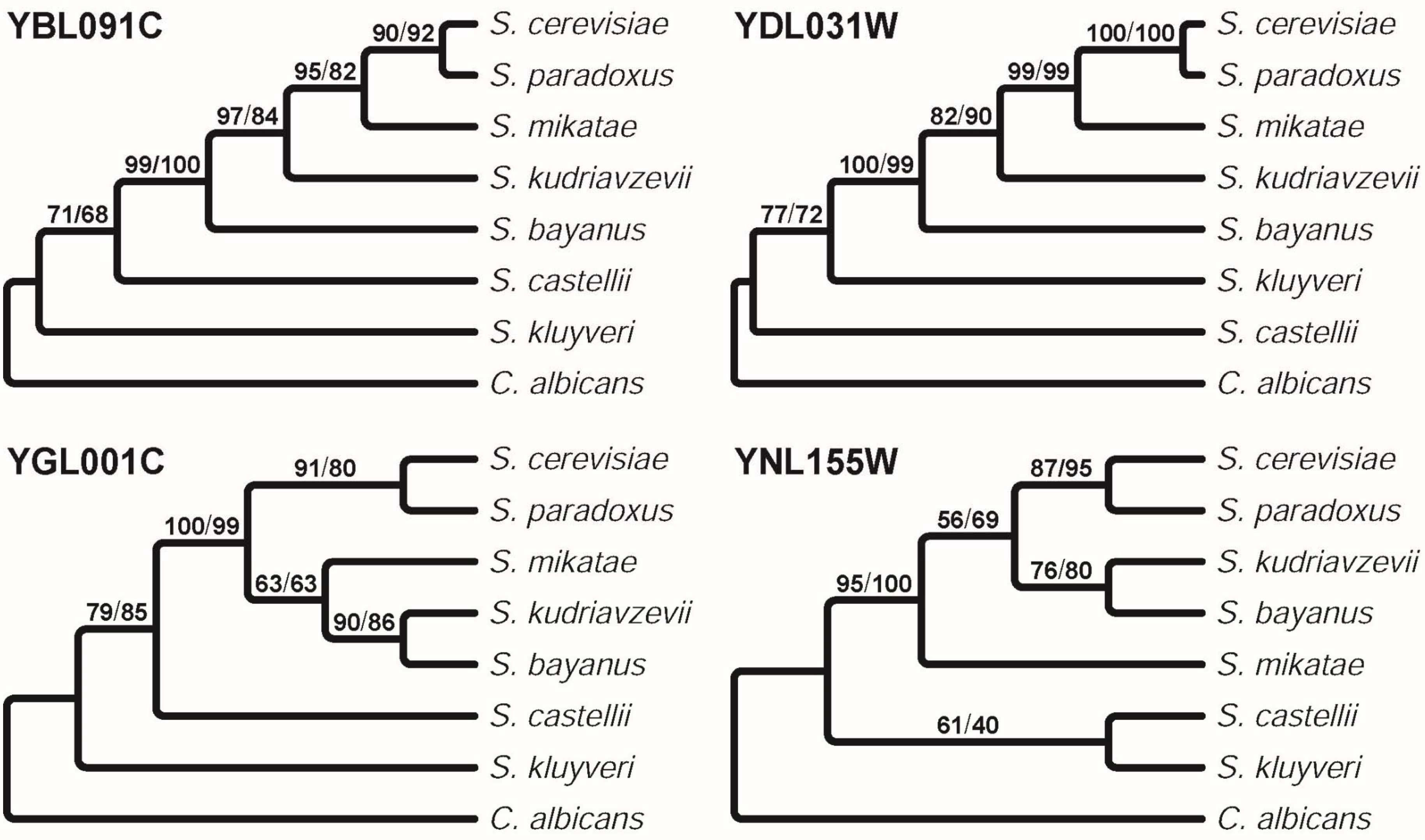
S. kluyveri

Candida glabrata



		YAL055W PEX22		YAL054C ACS1	YAL053W FLC2	YAL051W OAF1	YAL049C AIM2	YAL048C GEM1		YAL047C SPC72		YAL046C AIM1	YAL044W-B 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

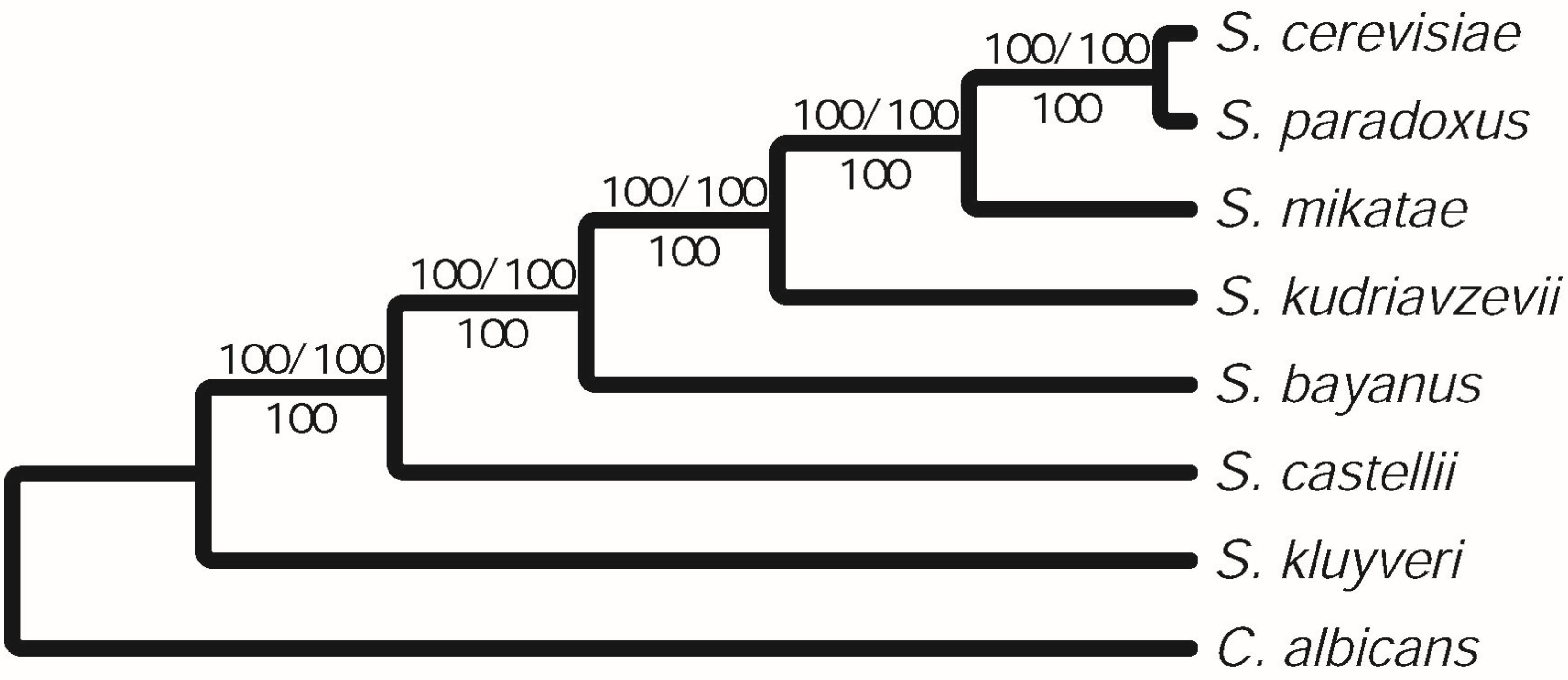


ML / MP

Anonymous Reviewer for Nature (2003)

Rokas et al. (2003) Nature

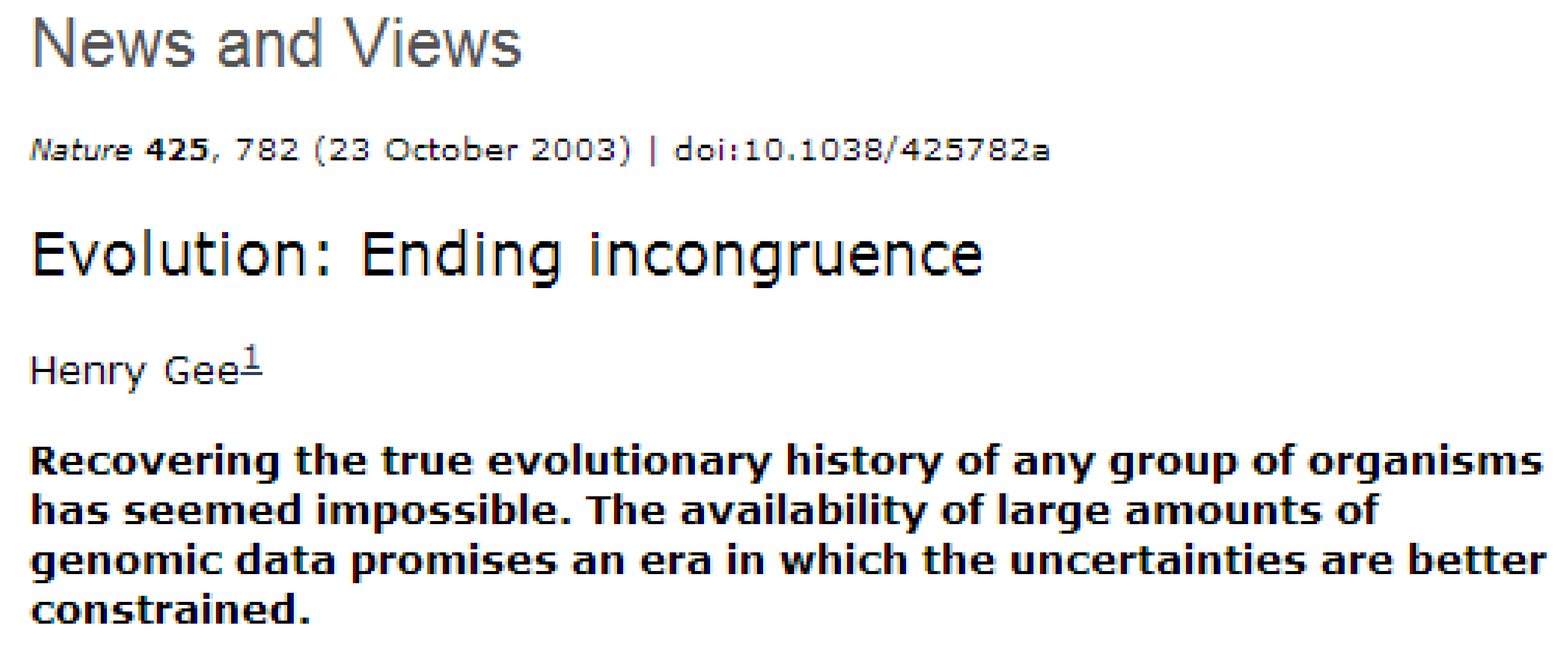
Concatenation of 106 Genes Yields a Single Yeast Phylogeny



ML / MP on nt
MP on aa

Rokas et al. (2003) Nature

The Phylogenomics Era – “Resolving” the Tree of Life



News and Views

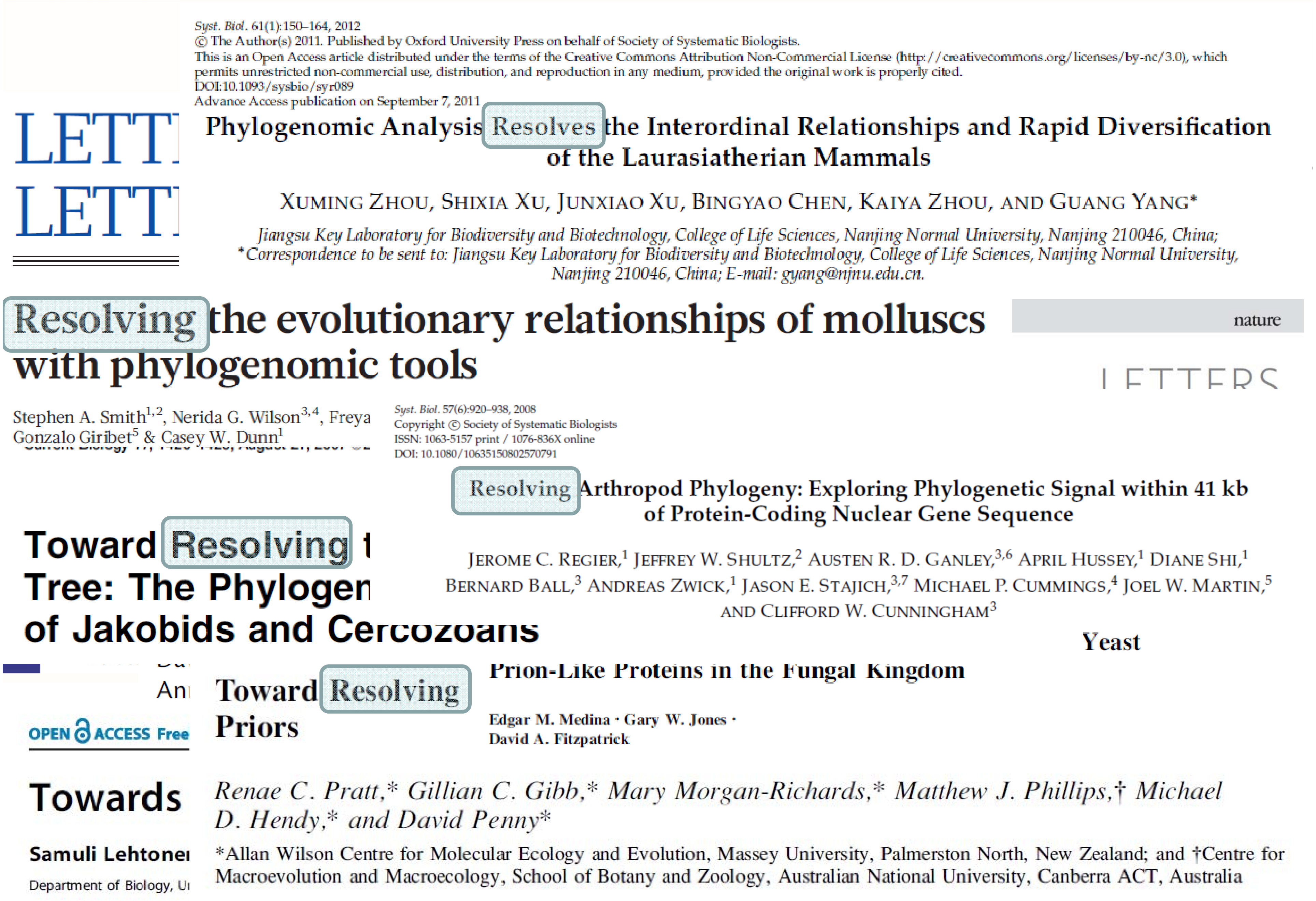
Nature **425**, 782 (23 October 2003) | doi:10.1038/425782a

Evolution: Ending incongruence

Henry Gee[1]

Recovering the true evolutionary history of any group of organisms has seemed impossible. The availability of large amounts of genomic data promises an era in which the uncertainties are better constrained.

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

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.*

LETT
LETT

Resolving the evolutionary relationships of molluscs with phylogenomic tools

nature

LETTERS

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

Syst. Biol. 57(6):920–938, 2008
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

Toward Resolving
Priors

OPEN ACCESS Free

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

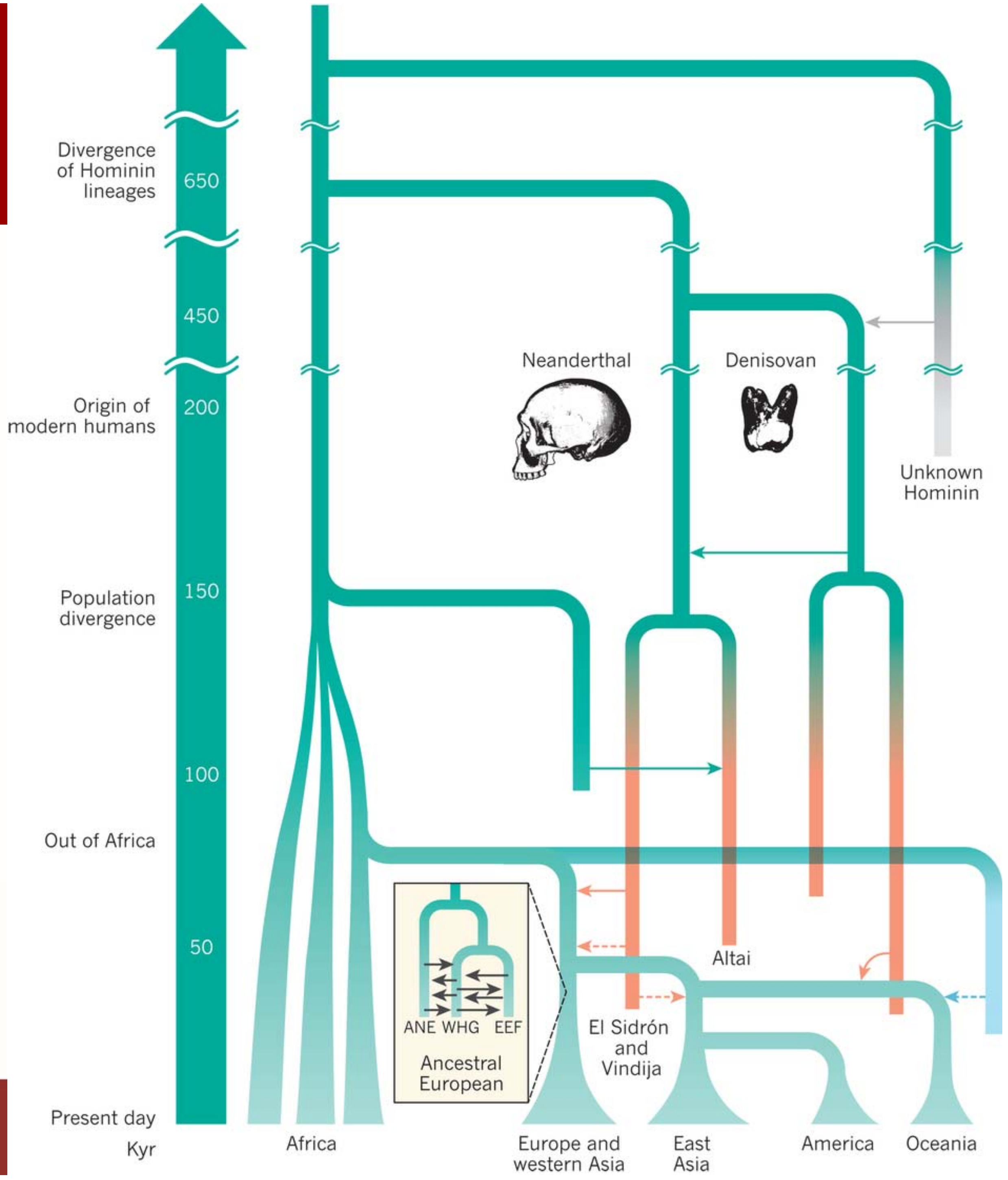
Samuli Lehtone

Department of Biology, U

Have we eliminated incongruence?

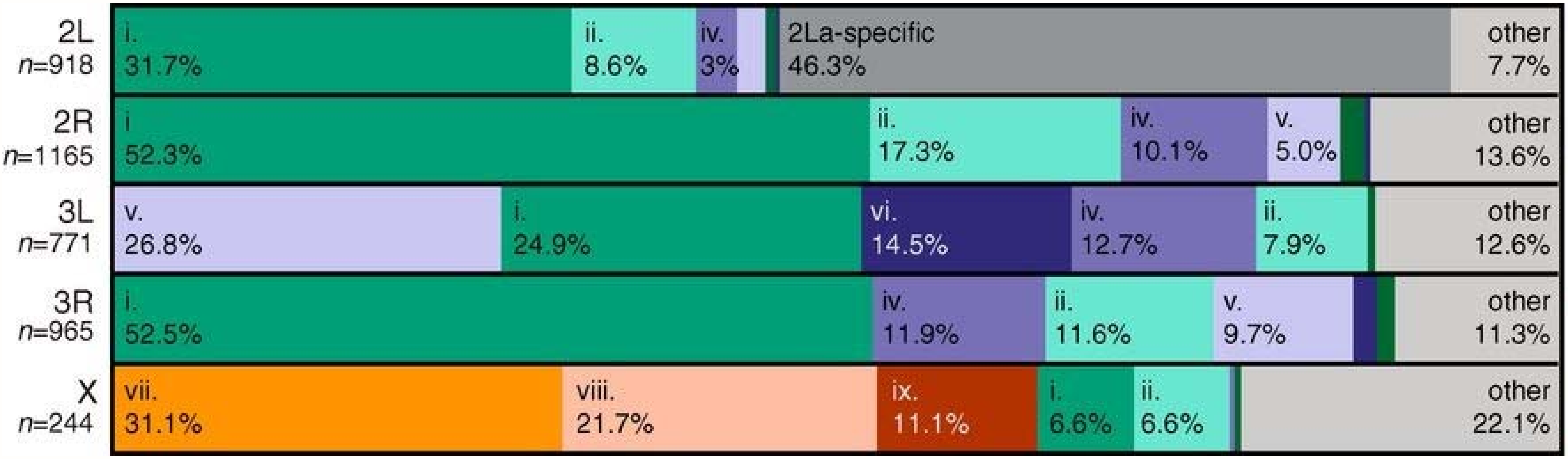
Figuring out what’s going on is easier at shallow depths, b/c systematic error is usually absent

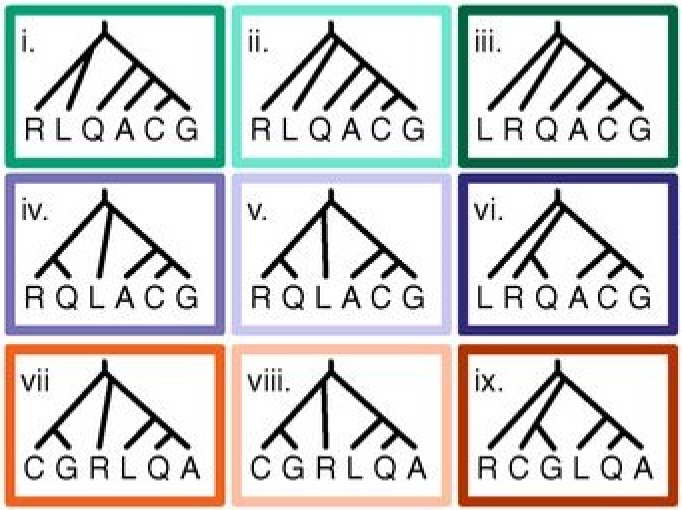
The Evolution of Human Populations



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

Fontaine et al. (2015) Science; Clark & Messer (2015) Science

The Phylogeny of Primate Genera

Nomascus leucogenys

NLE

Hoolock leuconedys

HLE

Symphalangus syndactylus

SSY

Hylobates pileatus

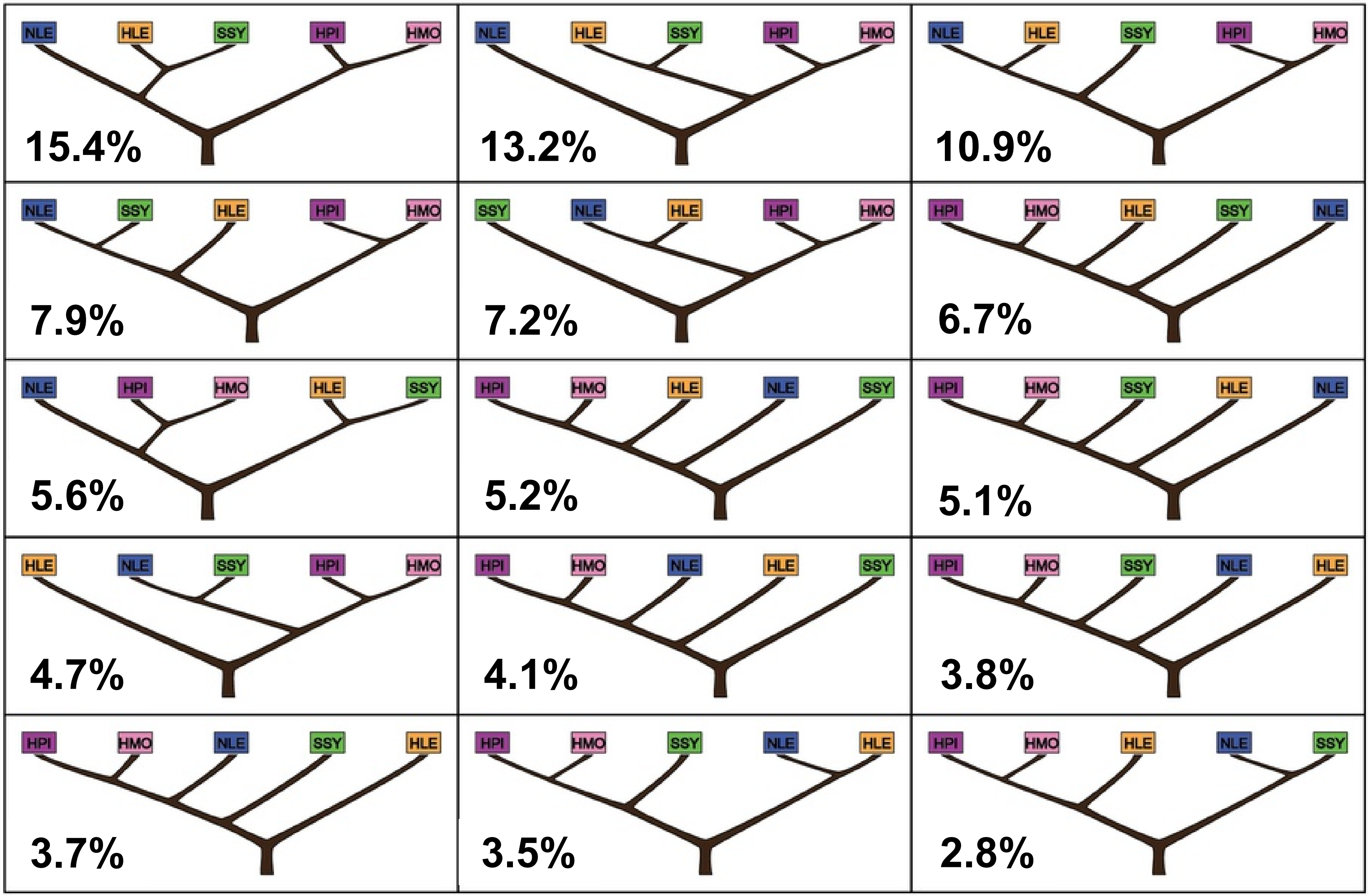
HPI

Hylobates moloch

HMO

Carbone et al. (2014) Nature

“Easier” Doesn’t Mean “Easy”!

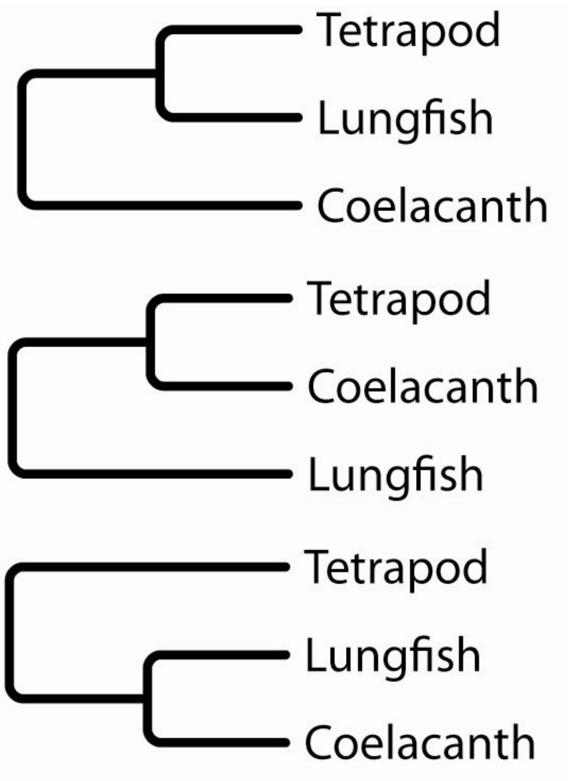


Carbone et al. (2014) Nature

Disentanglement Becomes More Challenging Deeper...

The three major lineages first appeared within 20 – 30 million years ago, approximately 390 million years ago

Inform. Sites



Tetrapod
Lungfish
Coelacanth

96 / 294

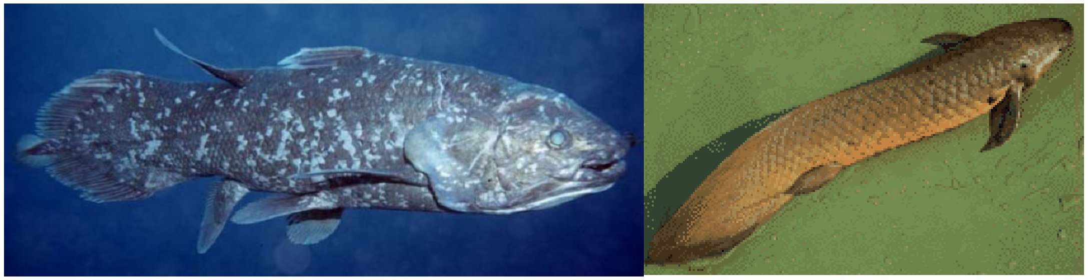
Tetrapod
Coelacanth
Lungfish

92 / 294

Tetrapod
Lungfish
Coelacanth

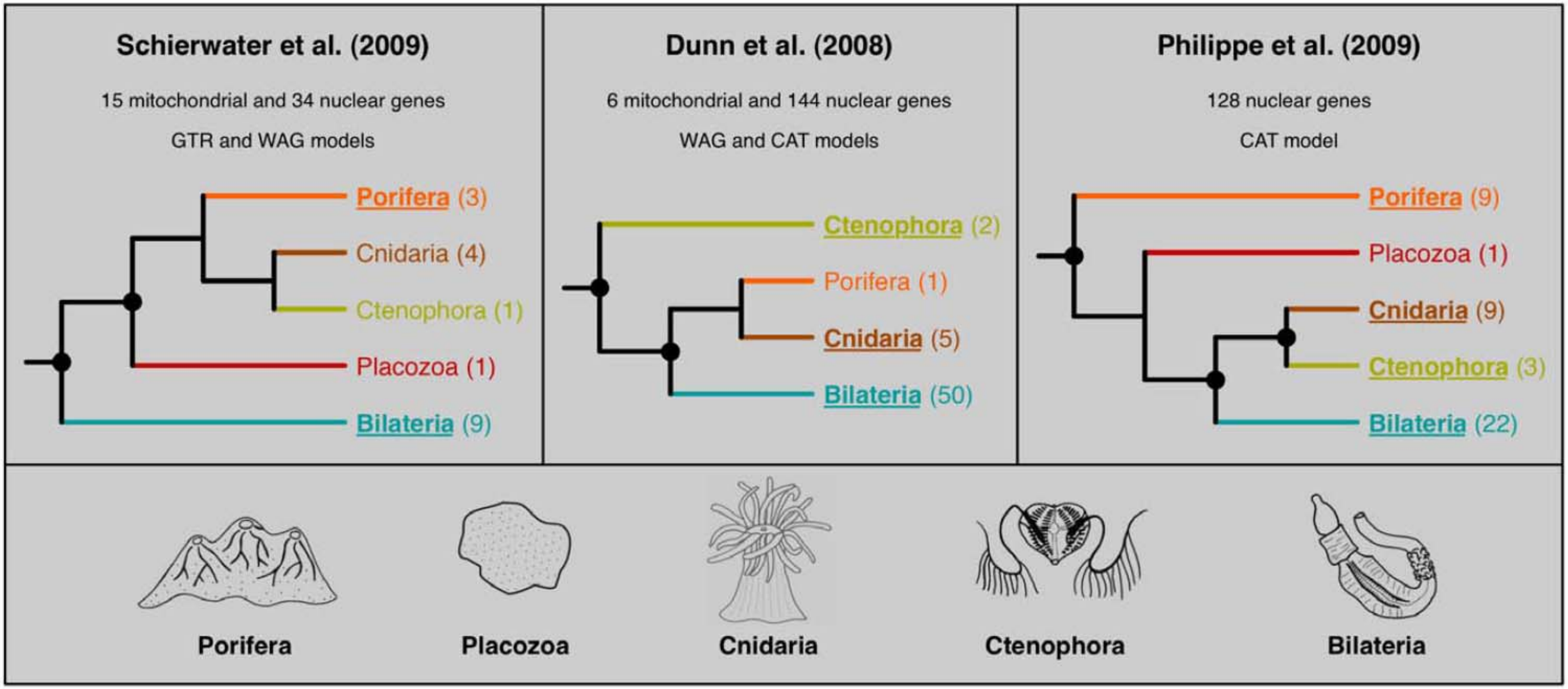
106 / 294

44 genes, ML/MP/NJ



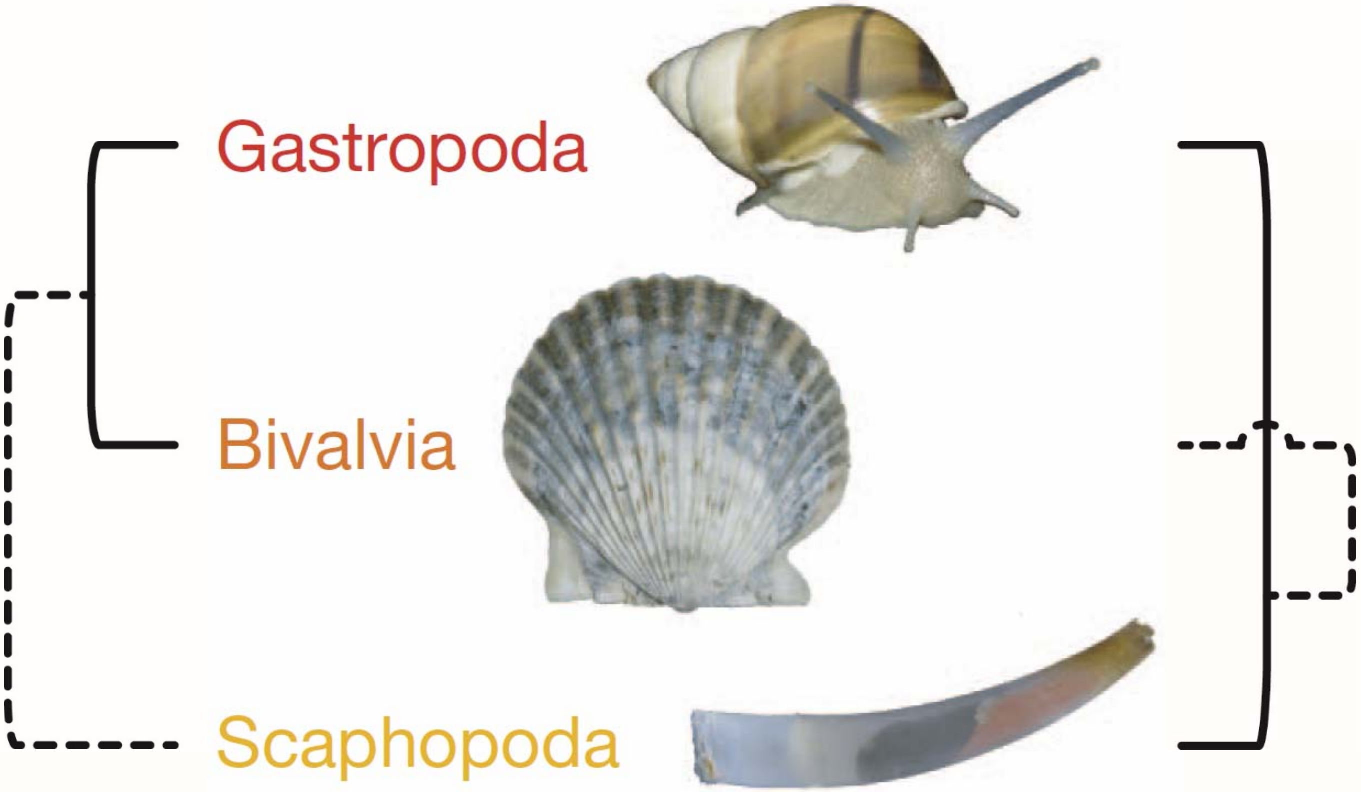
Takezaki et al. (2004) Mol. Biol. Evol.

Incongruence in Deep Time is More Challenging



Philippe et al. (2011) PLoS Biol.

Incongruence in Deep Time is More Challenging



Gastropoda

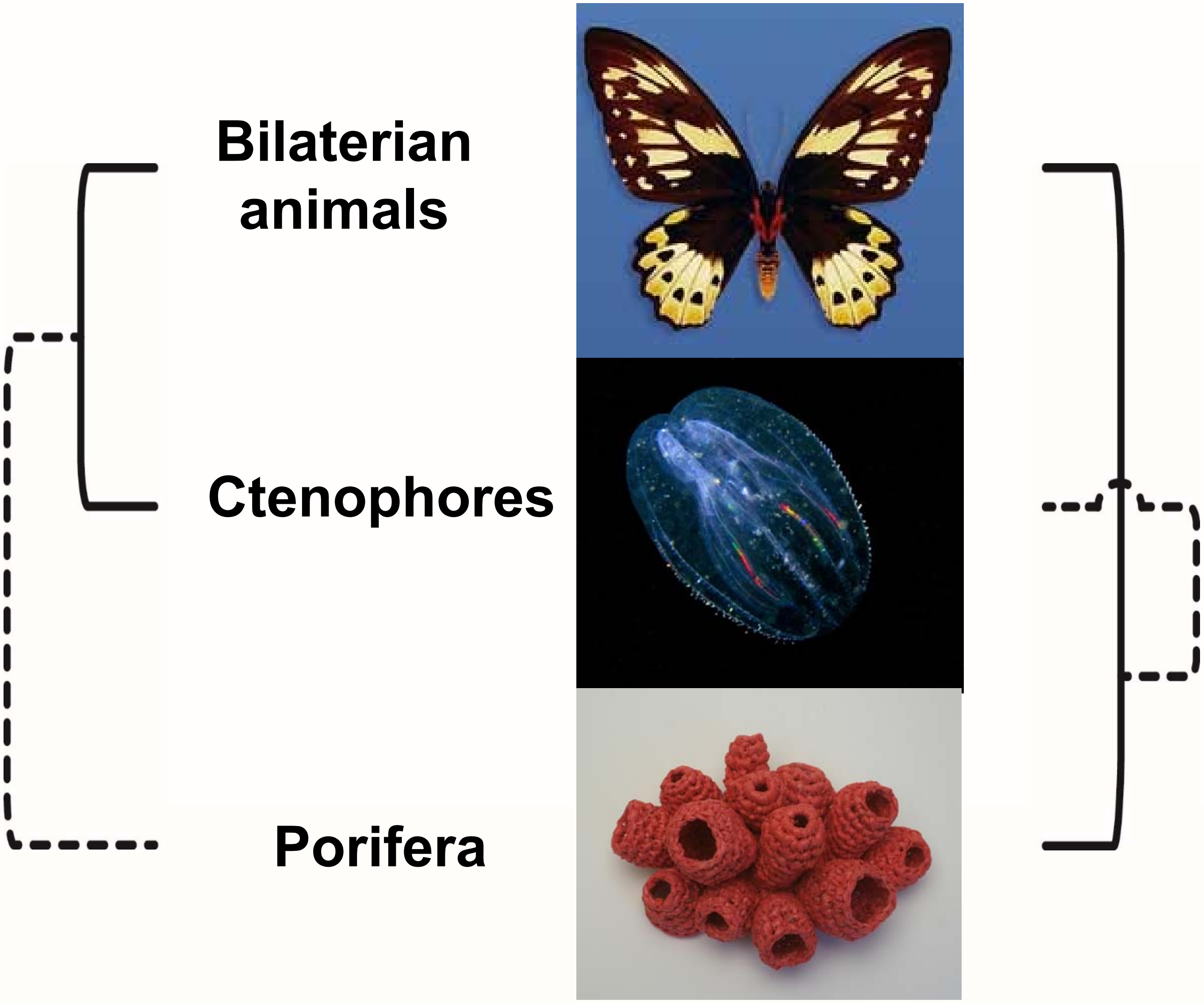
Bivalvia

Scaphopoda

Kocot et al. (2011) Nature

Smith et al. (2011) Nature

Incongruence in Deep Time is More Challenging



Bilaterian animals

Ctenophores

Porifera

Pisani et al. (2015) PNAS

Chang et al. (2015) PNAS

Why the disconnect?

Lecture Outline

- **From Darwin to Phylogenomics**

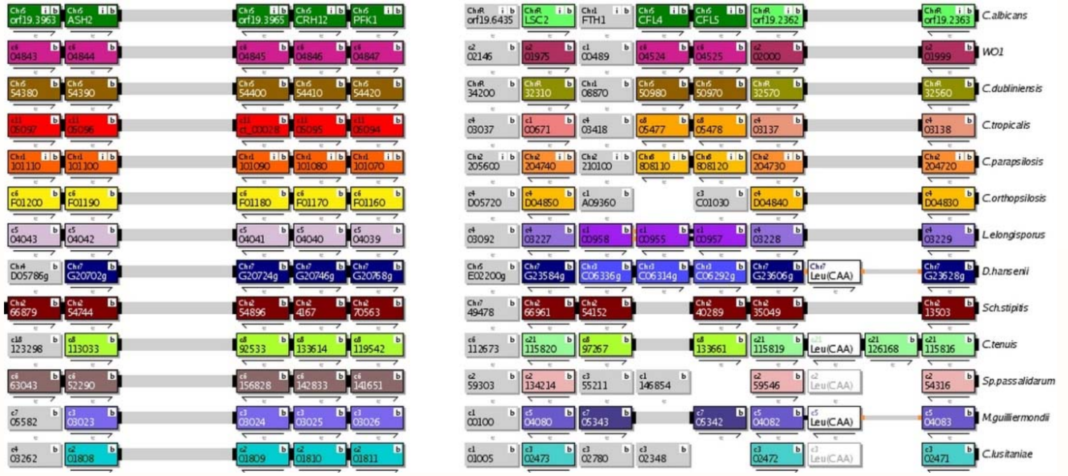
-------------------- Coffee Break -----------------------

- **Phylogenomics**

An Expanded Yeast Data Matrix

Yeast Gene Order Browser (YGOB)

Candida Gene Order Browser (CGOB)



***Saccharomyces* lineage**

***Candida* lineage**

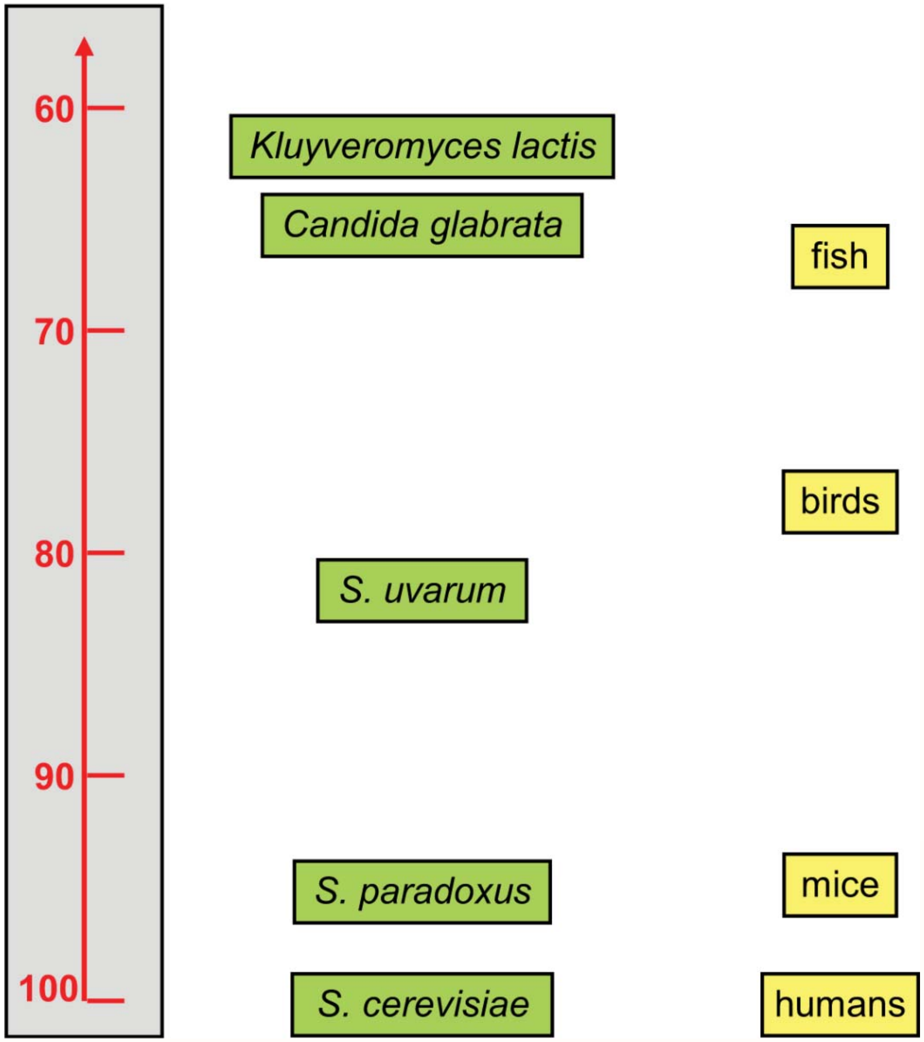
1,070 genes
23 taxa
no missing data

Byrne & Wolfe (2005) Genome Res. *Fitzpatrick et al. (2010) BMC Genom.*

Fungal Genomes are Similar in Divergence to Animals

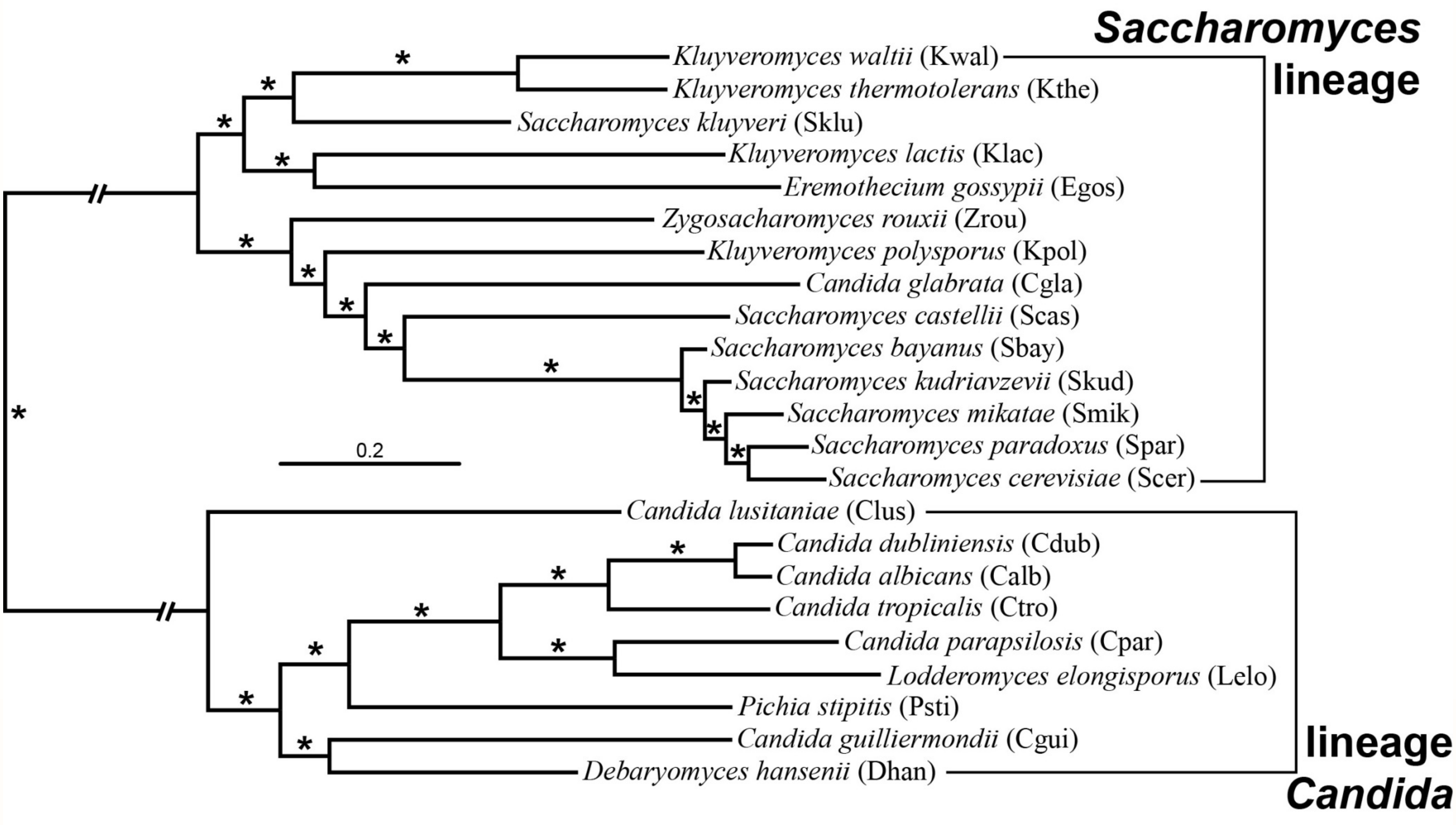
Proteome-wide average pairwise amino acid sequence similarity

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



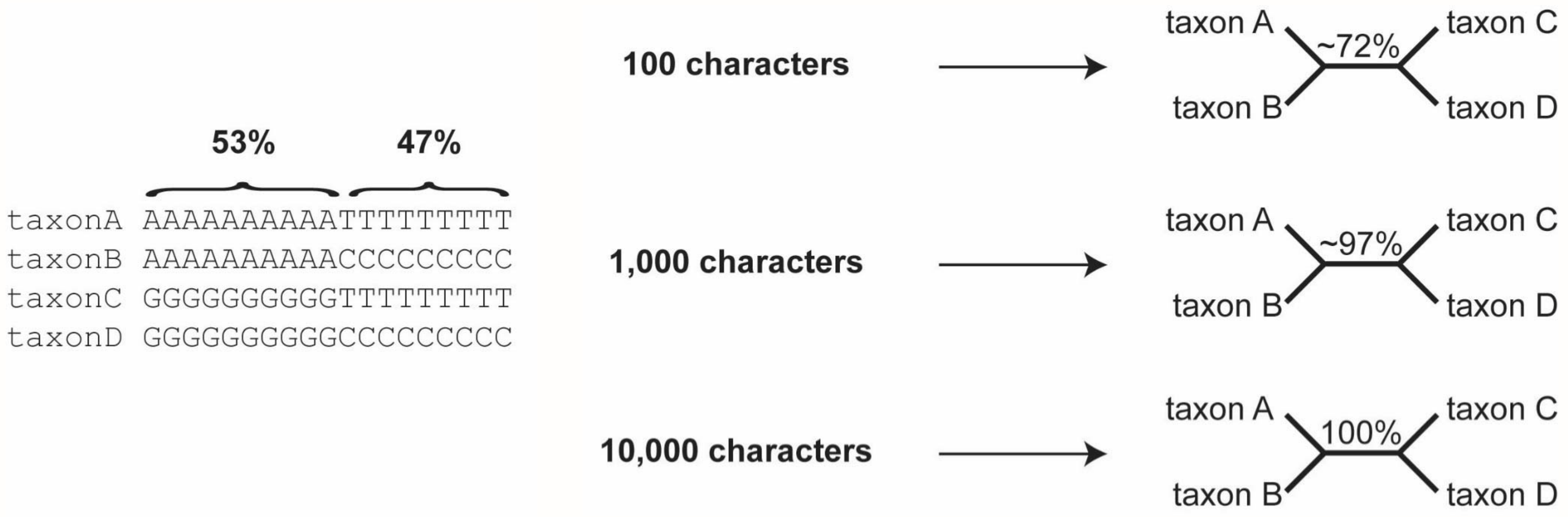
Fedorova et al. (2008) PLoS Genet.

Concatenation Yields an Absolutely Supported Phylogeny



Salichos & Rokas (2013) Nature

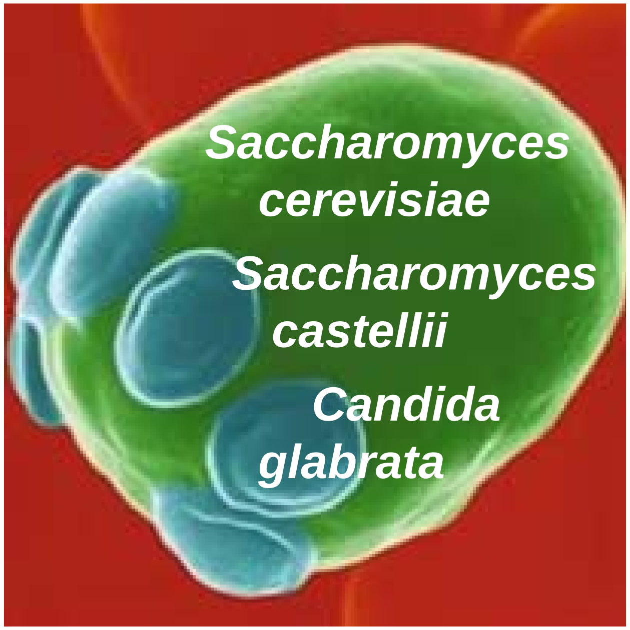
Bootstrap Support is Misleading When Used in Large Datasets

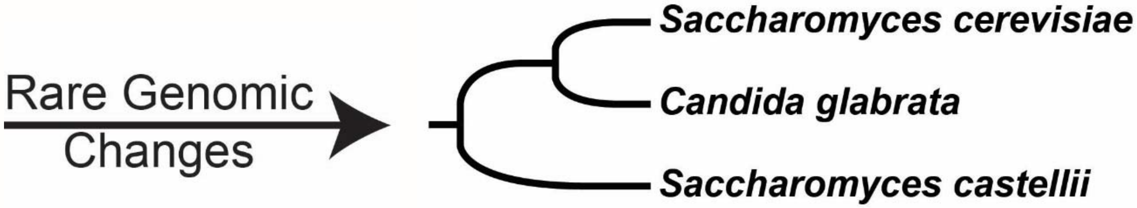


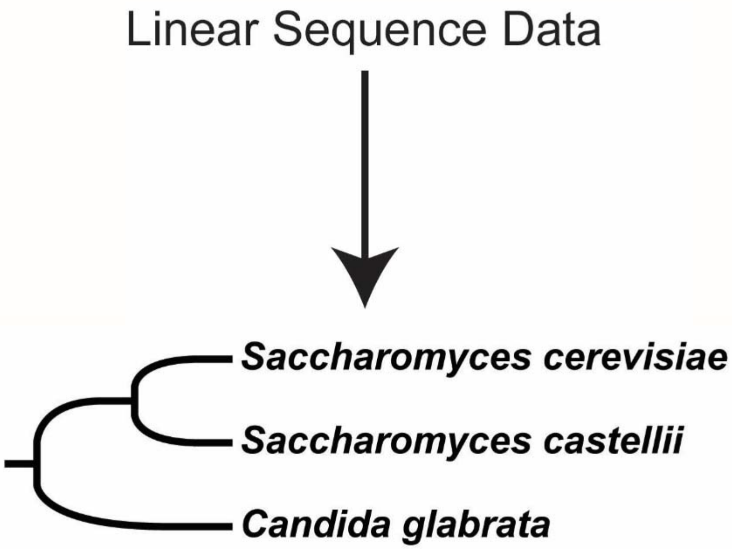
```
taxonA AAAAAAAAAATTTTTTTTT
taxonB AAAAAAAAAACCCCCCCCC
taxonC GGGGGGGGGGTTTTTTTTT
taxonD GGGGGGGGGGCCCCCCCCC
```

Rokas & Carroll (2006) PLoS Biol.

The Concatenation Phylogeny is at Least Partly Wrong



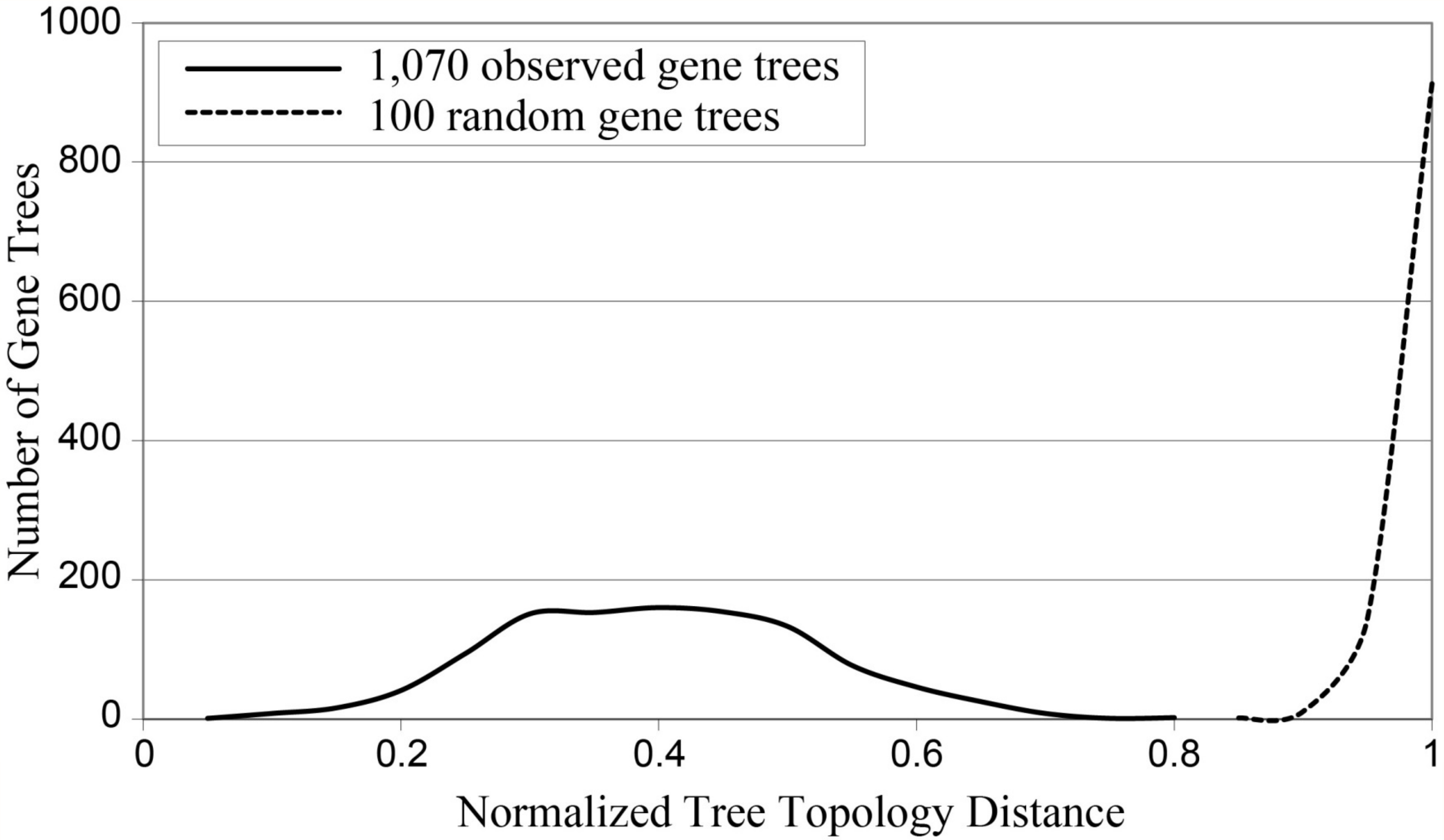




- **5 genomic rearrangements that are uniquely shared by *S. cerevisiae* and *C. glabrata***
- **Much higher proportion of shared gene losses in *S. cerevisiae* and *C. glabrata***
- **Bias in the placement of *C. glabrata* as an outgroup of *S. cerevisiae* and *S. castellii***

Scannell et al. (2006) Nature

All Gene Trees Differ from the Concatenation Phylogeny



Salichos & Rokas (2013) Nature

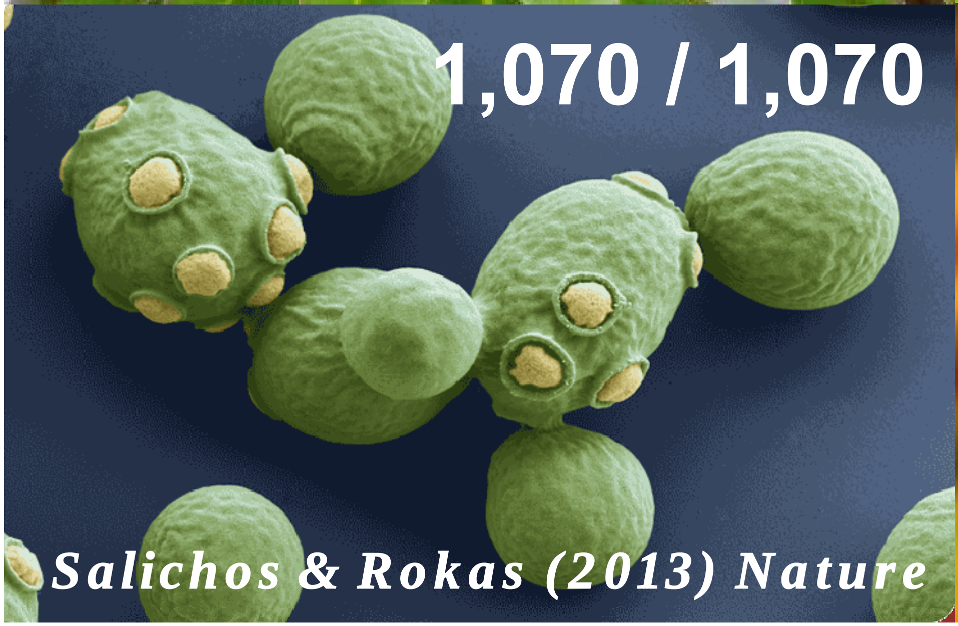
Gene Trees are Incongruent in Most Datasets

182 / 184

Zhong et al. (2013) Trends Plant Sci.

440 / 447

Song et al. (2012) PNAS



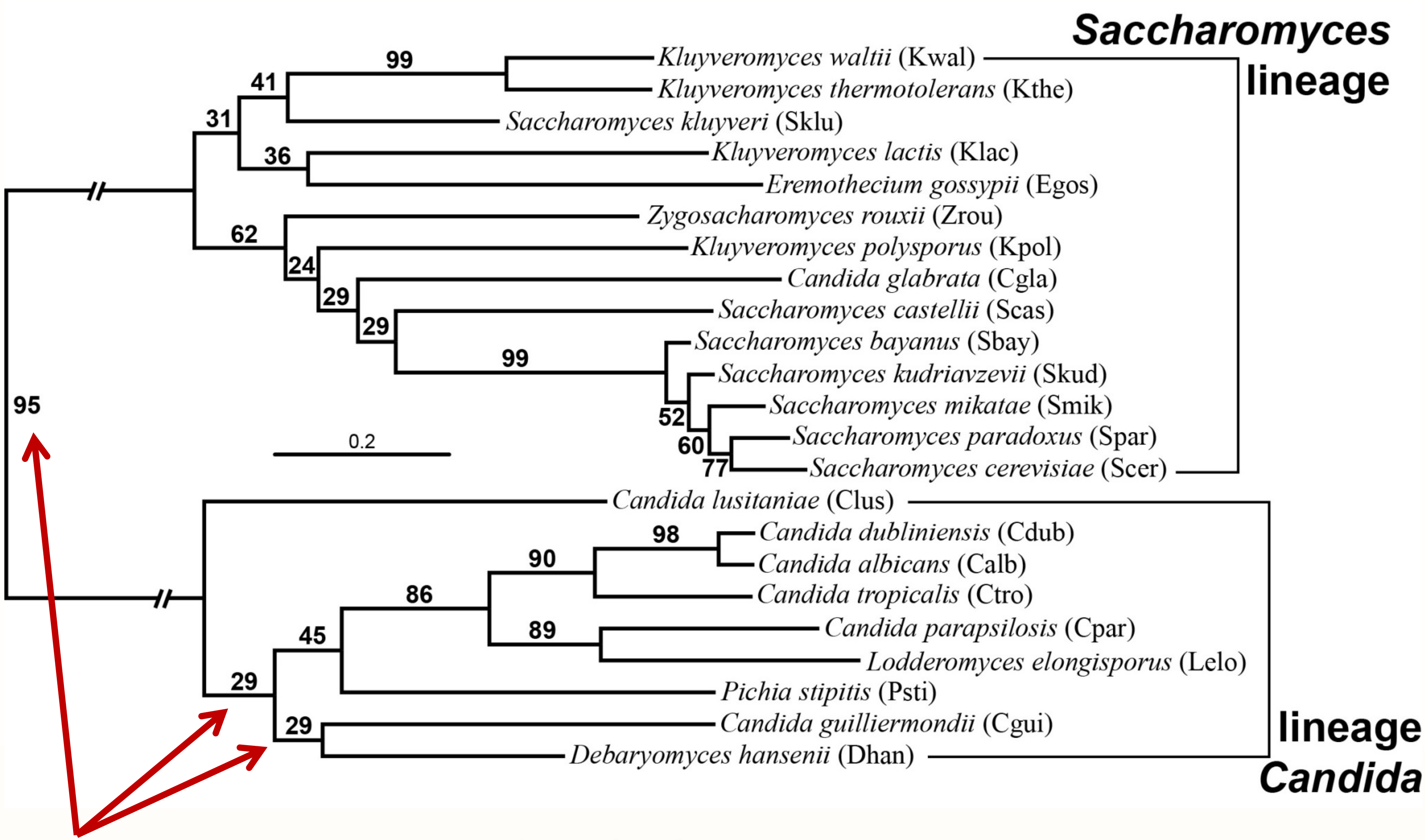
1,070 / 1,070

Salichos & Rokas (2013) Nature

14,536 / 14,536

Jarvis et al. (2014) Science

The Yeast Phylogeny Inferred by Majority-Rule Consensus

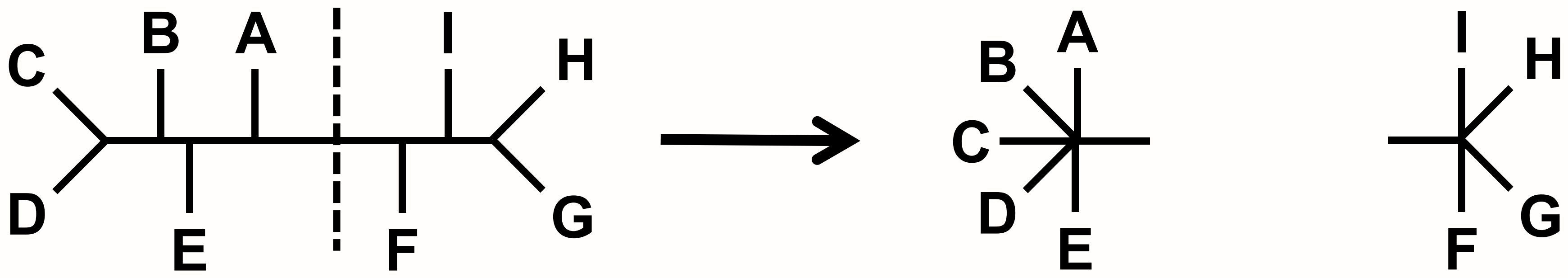


Gene Support Frequency (GSF): % of single gene trees supporting a given internode

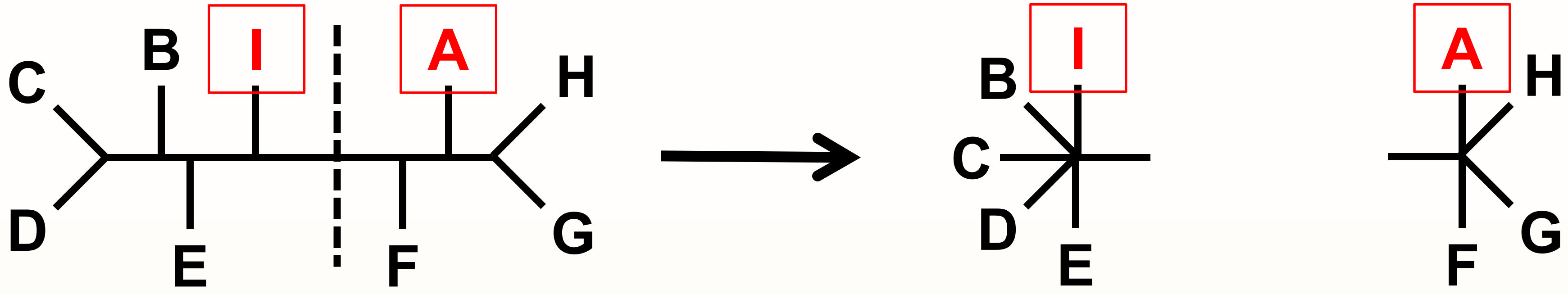
Salichos & Rokas (2013) Nature

Phylogenetic Trees are Sets of Splits

Division **Splits**



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



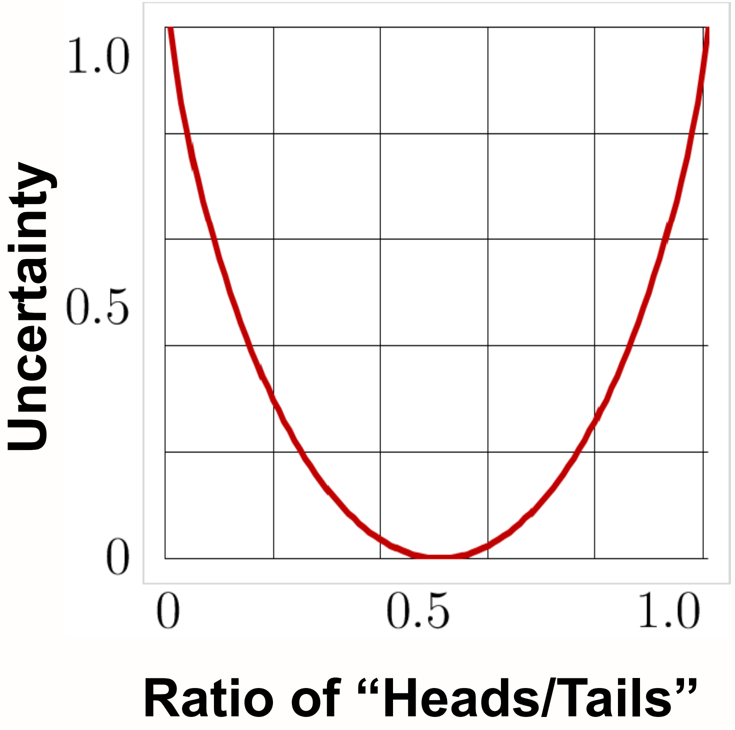
Conflicting Splits: {I, B, C, D, E} {F, G, H, A}

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



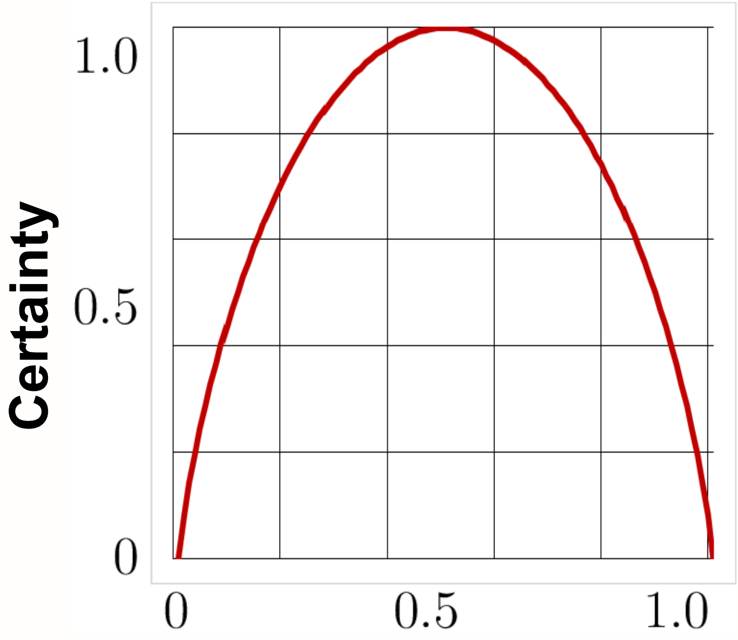
Salichos et al. (2014) Mol. Biol. Evol.; Kobert et al. (2016) Mol. Biol. Evol.

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

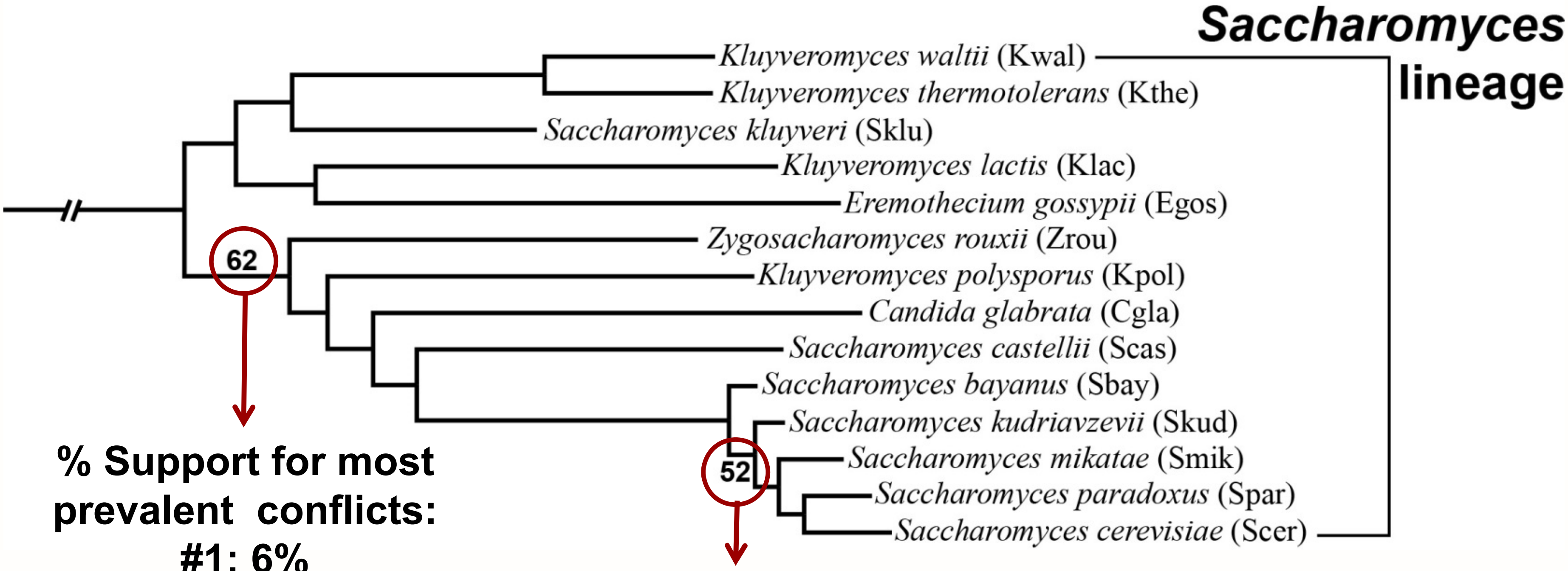
IC and TC are implemented in the latest versions of RAxML



Ratio of Support for Two Conflicting Internodes

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

IC Can Be More Informative Measure of Internode Support



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

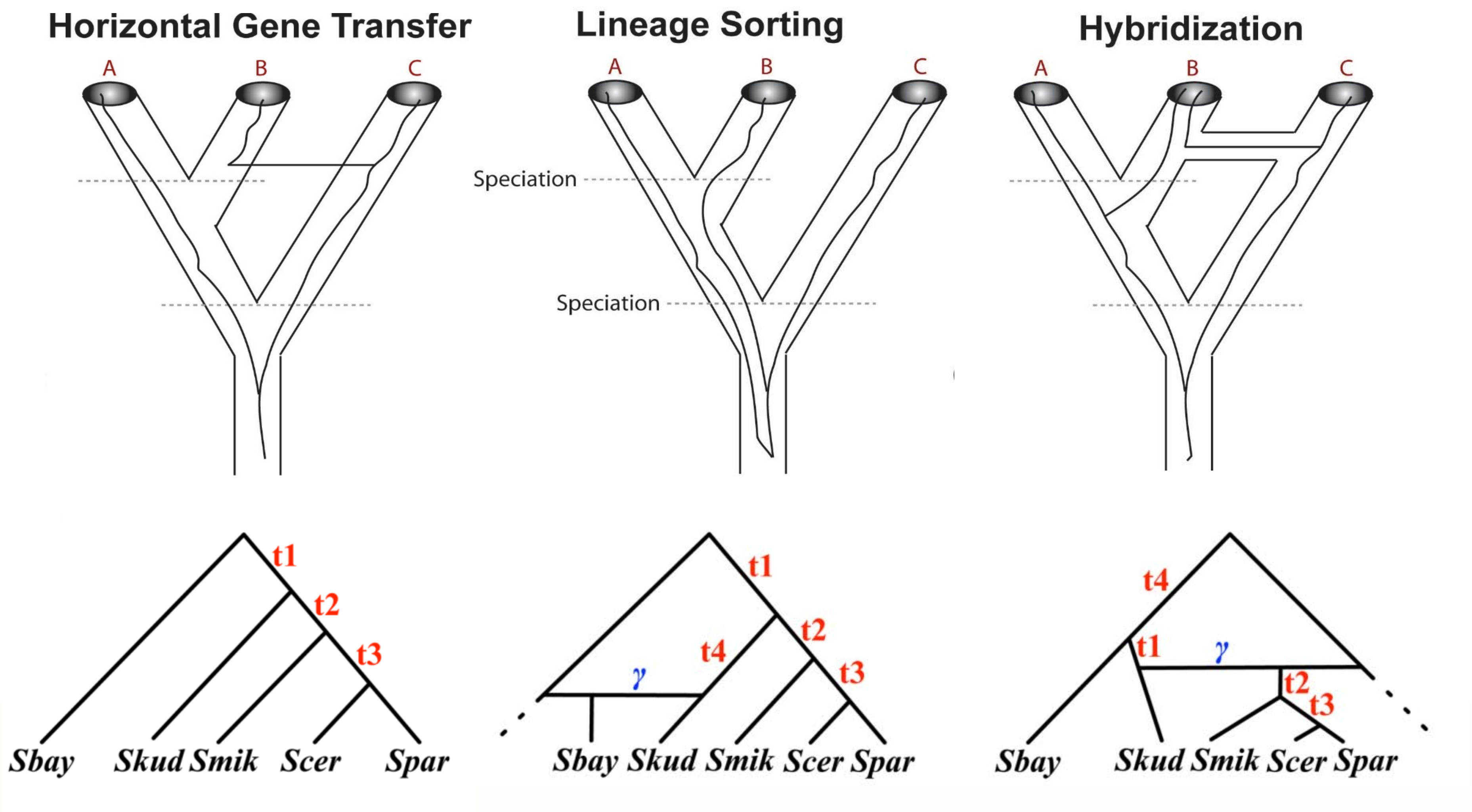
IC value: 0.59

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

IC value: 0.06

Salichos & Rokas (2013) Nature

Why So Much Incongruence? Biological Factors



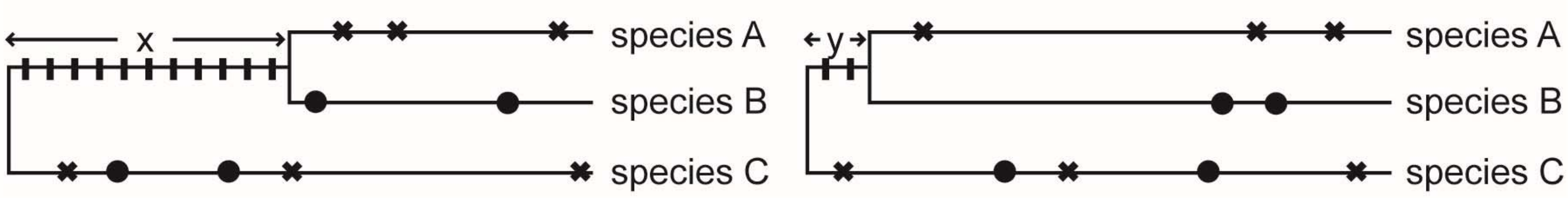
Horizontal Gene Transfer

Lineage Sorting

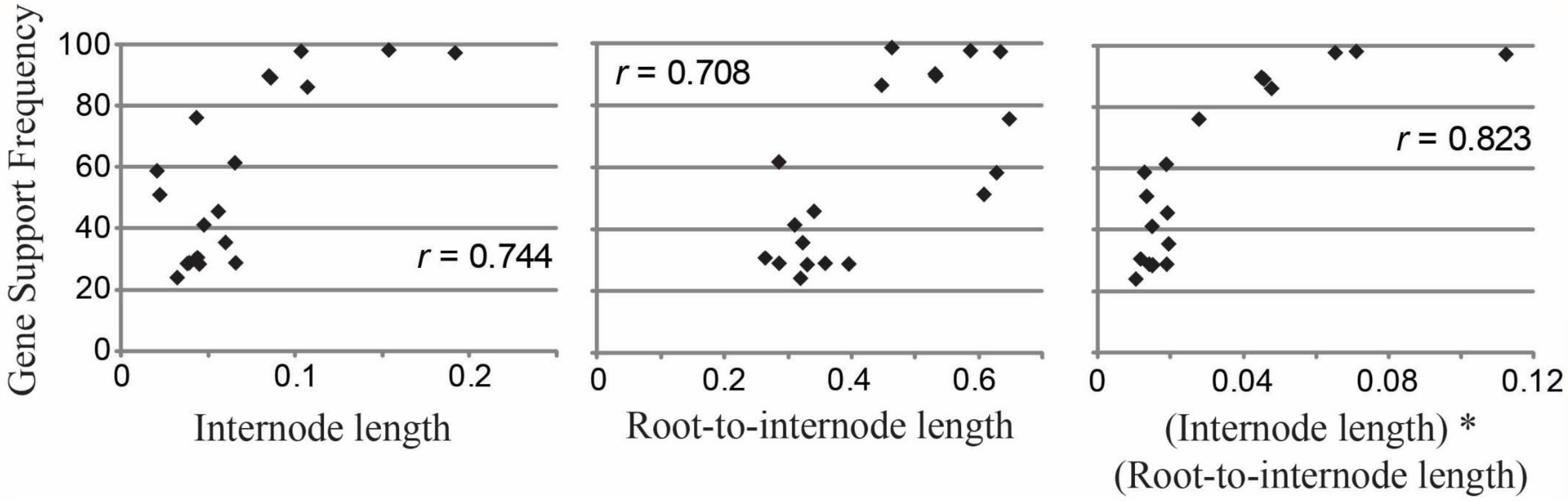
Hybridization

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

Standard Recipes for Handling Incongruence Didn't Help

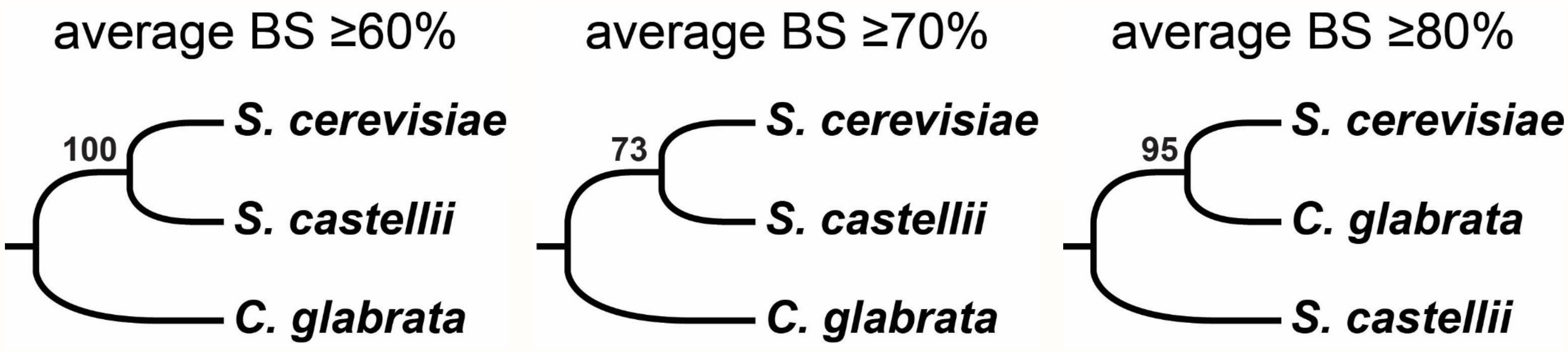
Treatment	Tree Certainty	# of Internodes where IC increased | decreased
Default analysis	**8.35**	**n/a**
Removing sites containing gaps		
All sites with gaps excluded	**7.91**	**0 | 7**
Removing fast-evolving or unstable species		
C. lusitaniae	**8.15**	**1 | 2**
C. glabrata	**8.30**	**2 | 2**
E. gossypii,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

Salichos & Rokas (2013) Nature


WHAT CAN I DO?!!

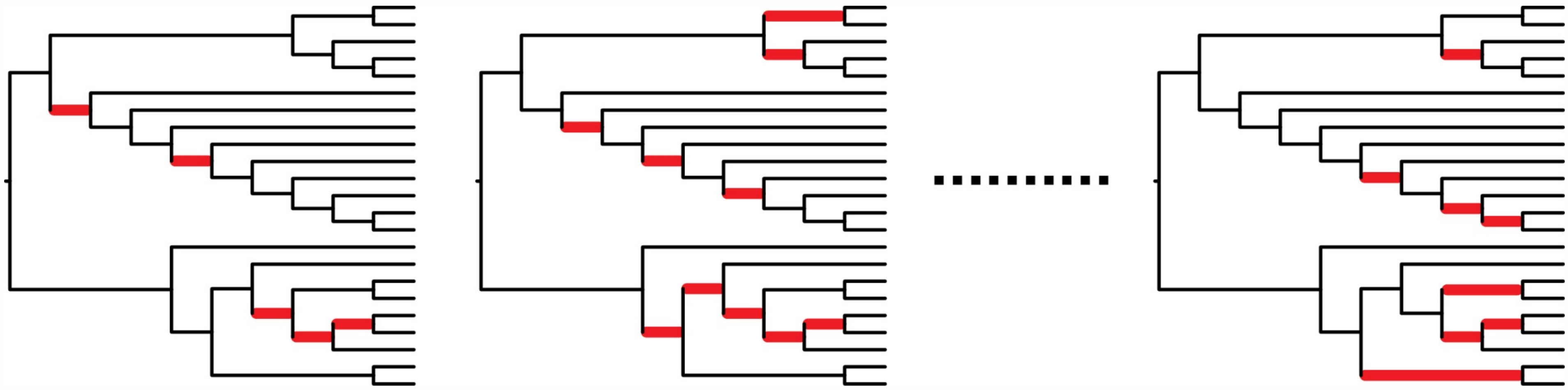
What Do We Do Then?

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



Salichos & Rokas (2013) Nature

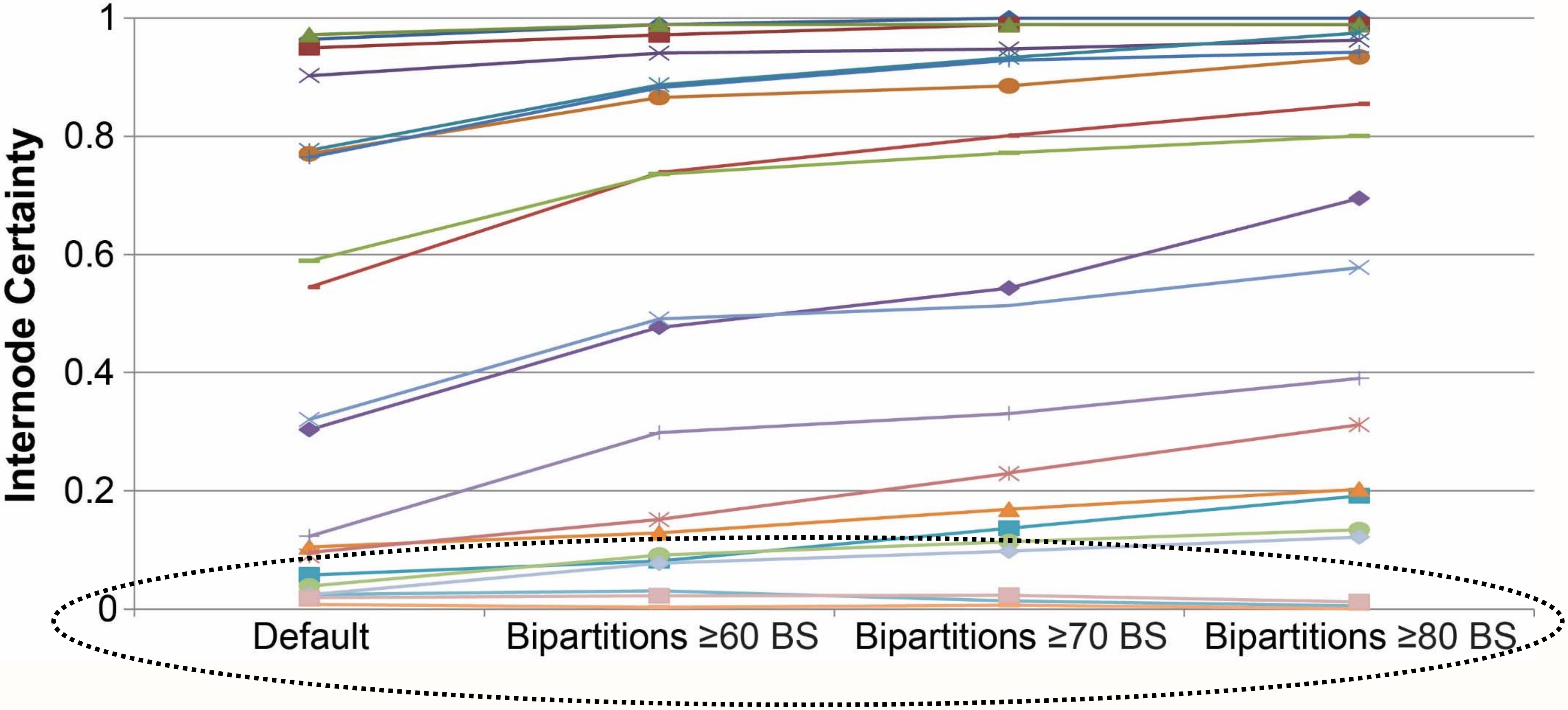
Selecting Specific Bipartitions Dramatically Improves Phylogeny



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**

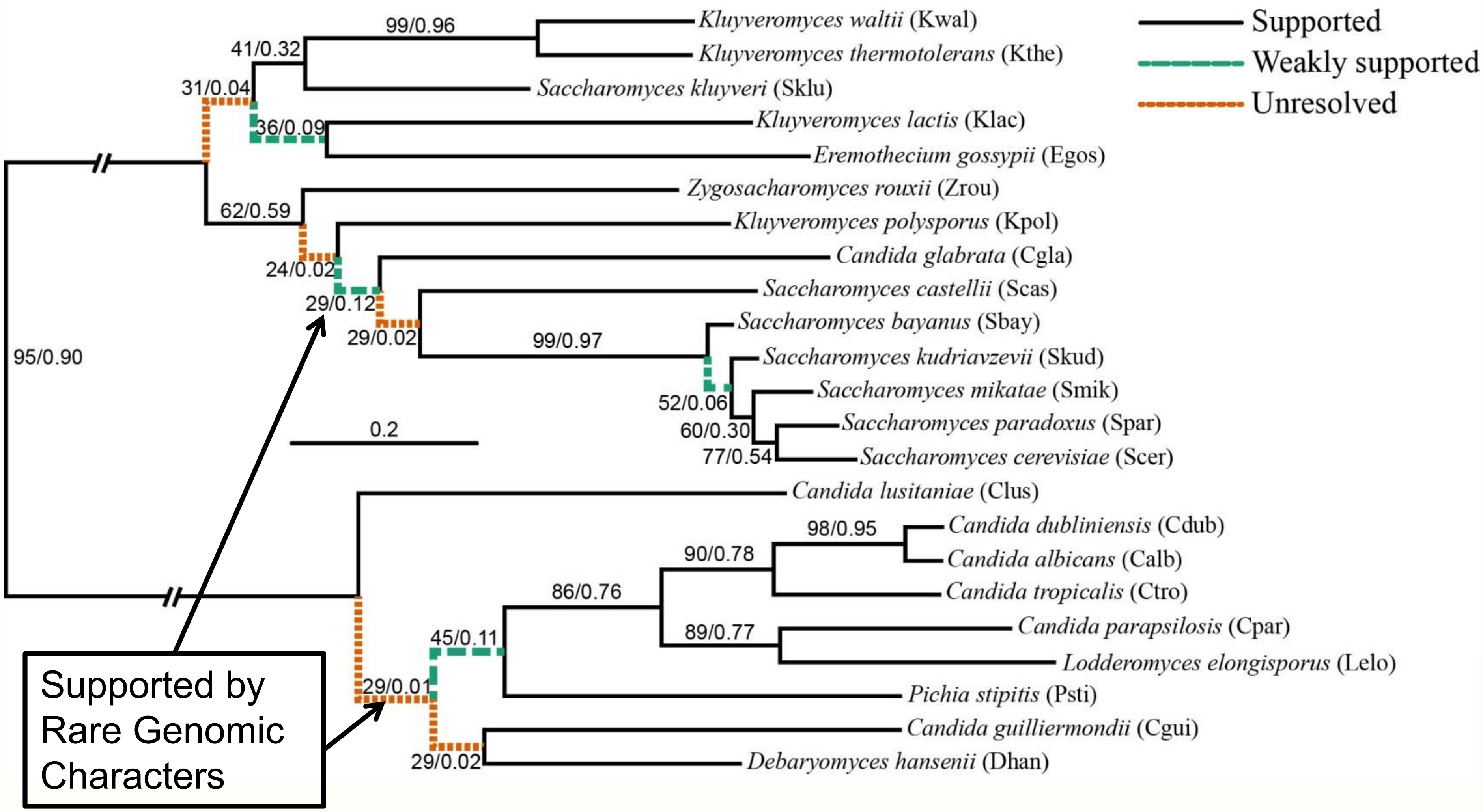
Salichos & Rokas (2013) Nature

Least Supported Internodes Harbor the Most Conflict



Salichos & Rokas (2013) Nature

The Status of the Yeast Phylogeny



Supported by Rare Genomic Characters

Gene Support Frequency / Internode Certainty

Similar Results in Other Lineages

Vertebrates

(1,086 genes, 18 taxa)

Animals

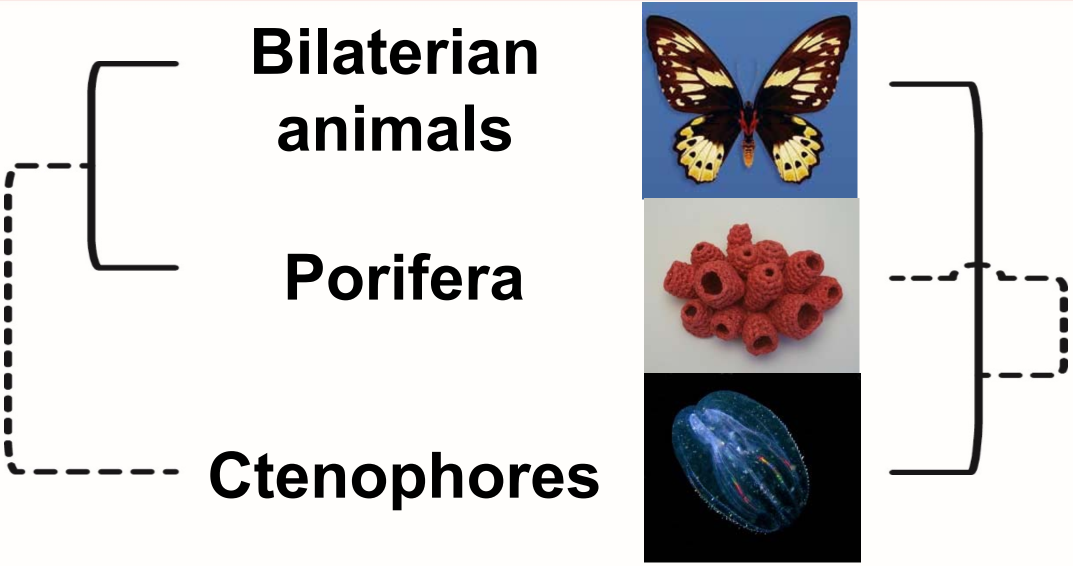
(225 genes, 21 taxa)

Mosquitoes

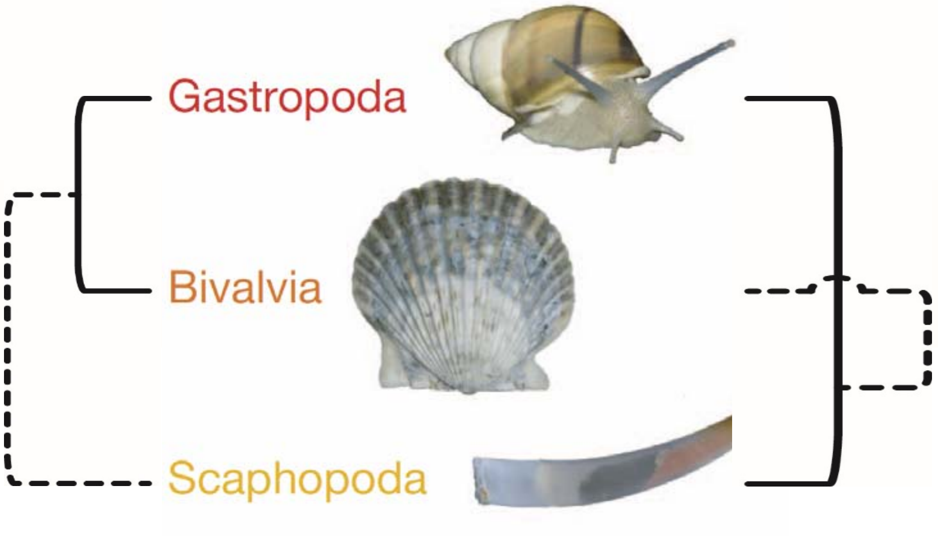
(2,007 genes, 20 taxa)

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

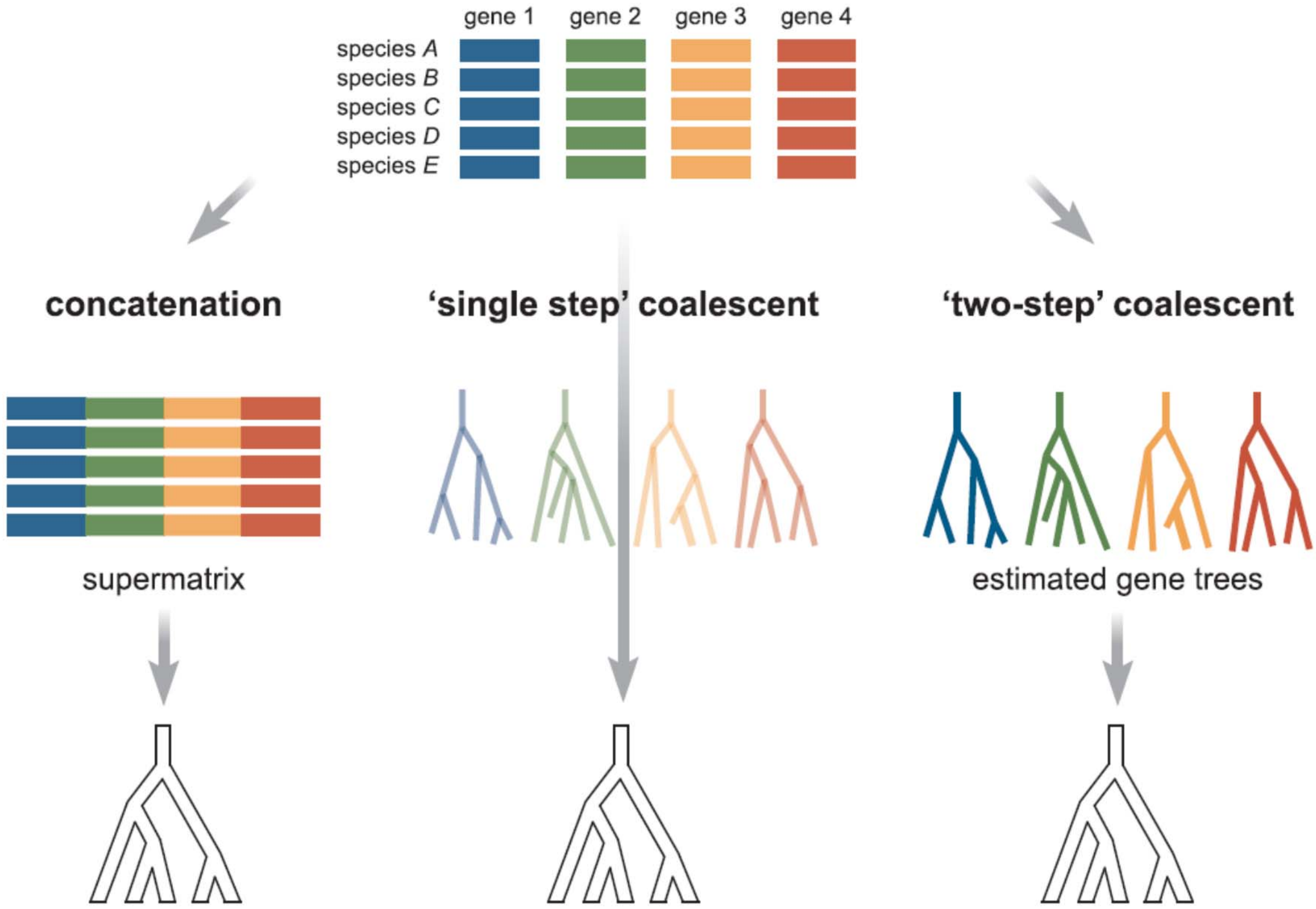
Incongruence in Phylogenomic Datasets



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

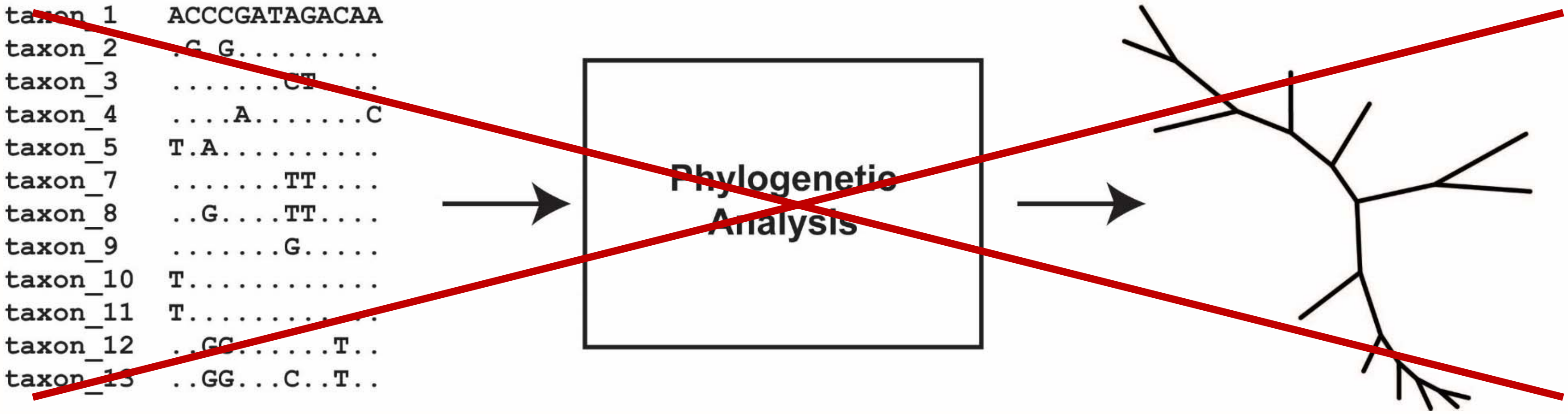


Methods for Phylogenomic Inference



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

The Way Forward



Multiple sequence alignment / data matrix reconstruction →

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

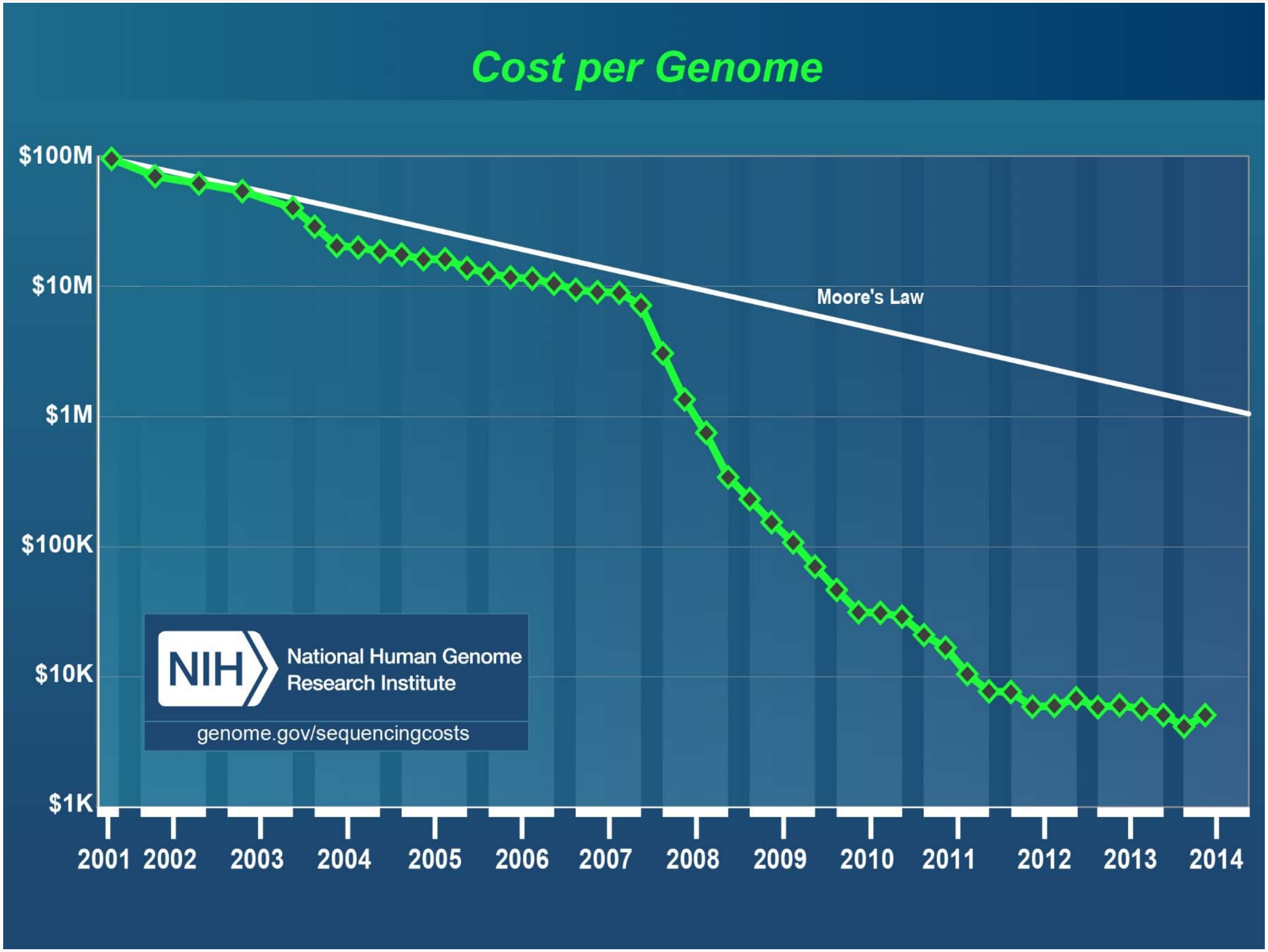
Assess conflict (e.g., use internode certainty) →

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

Only report resolution of branches that you have support for

What lies ahead

Cost of DNA Sequencing in this Century



http://www.genome.gov/sequencingcosts/

The Genomes of Non-Model Organisms are the New Frontiers

Rokas & Abbot (2009) Trends Ecol. Evol.

The Making of Biodiversity across the Yeast Subphylum

THE UNIVERSITY of WISCONSIN
MADISON

USDA

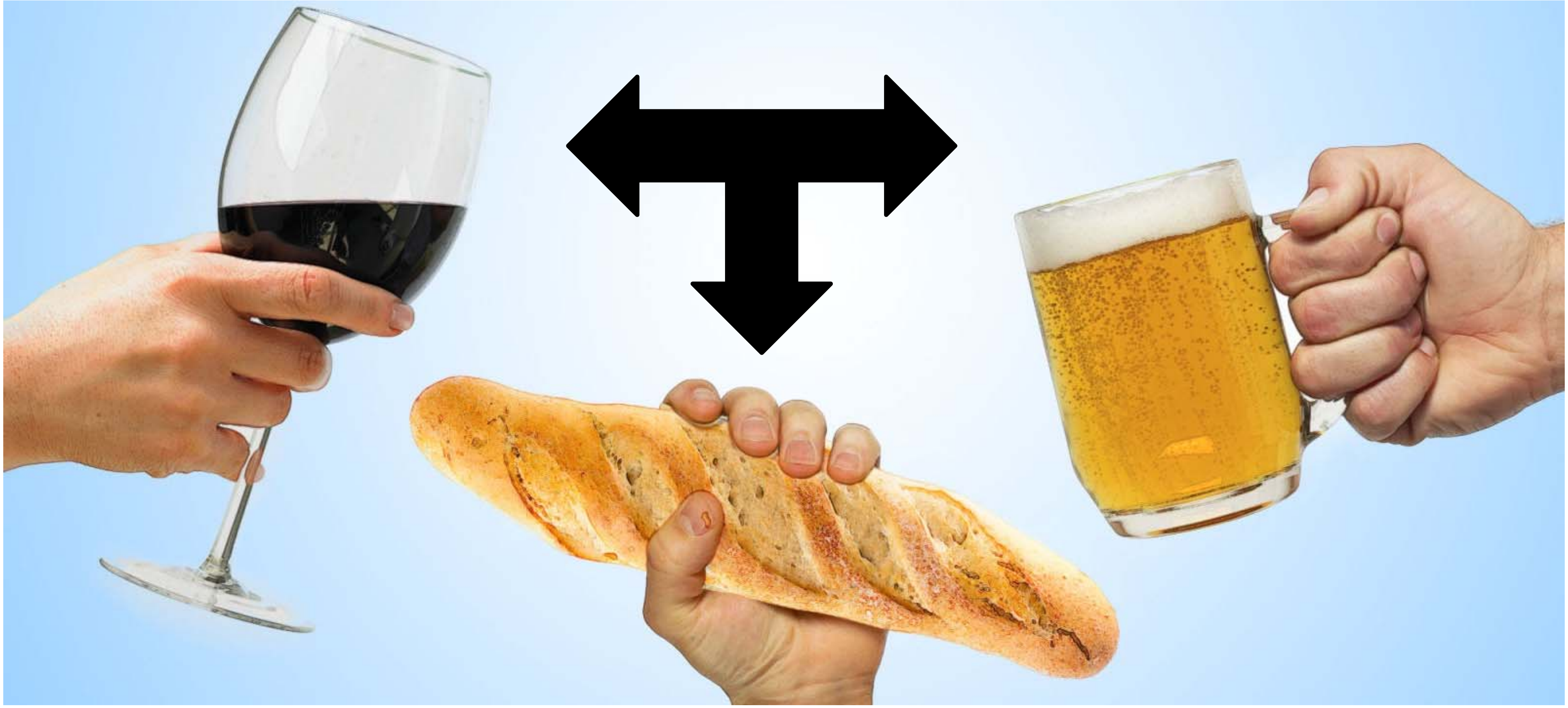
VANDERBILT
UNIVERSITY

Hittinger lab

Kurtzman lab

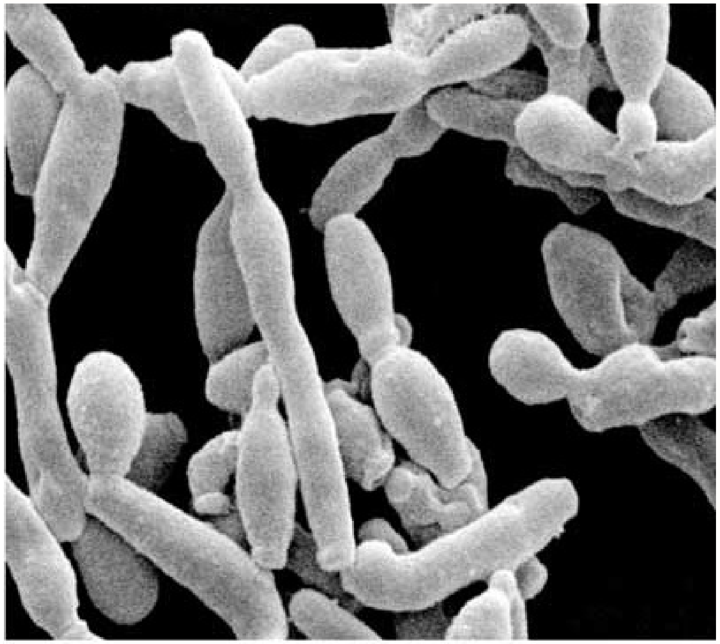
Rokas lab

Saccharomyces cerevisiae, *Molecular Biology's Workhorse*



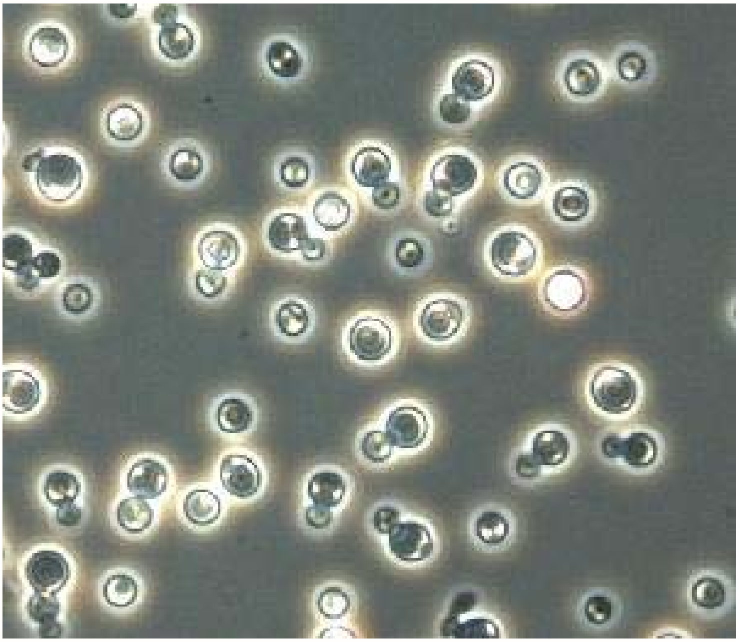
The Metabolisms of the 1,000+ Species Vary Widely

Xylose fermenters
(*Scheffersomyces (Pichia) stipitis*)

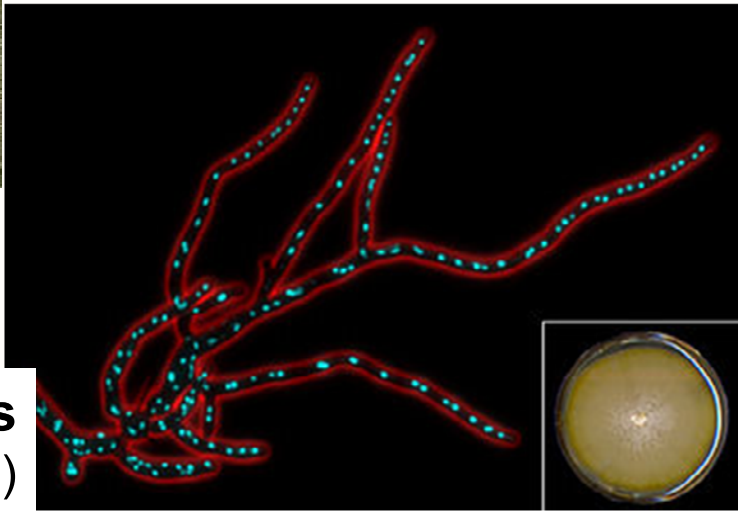


Cactophilic yeasts

Oil producers
(*Lipomyces*)



Animal pathogens
(*Candida*)



Plant pathogens
(*Eremothecium* sp.)

The Making of Biodiversity Across the Yeast Subphylum

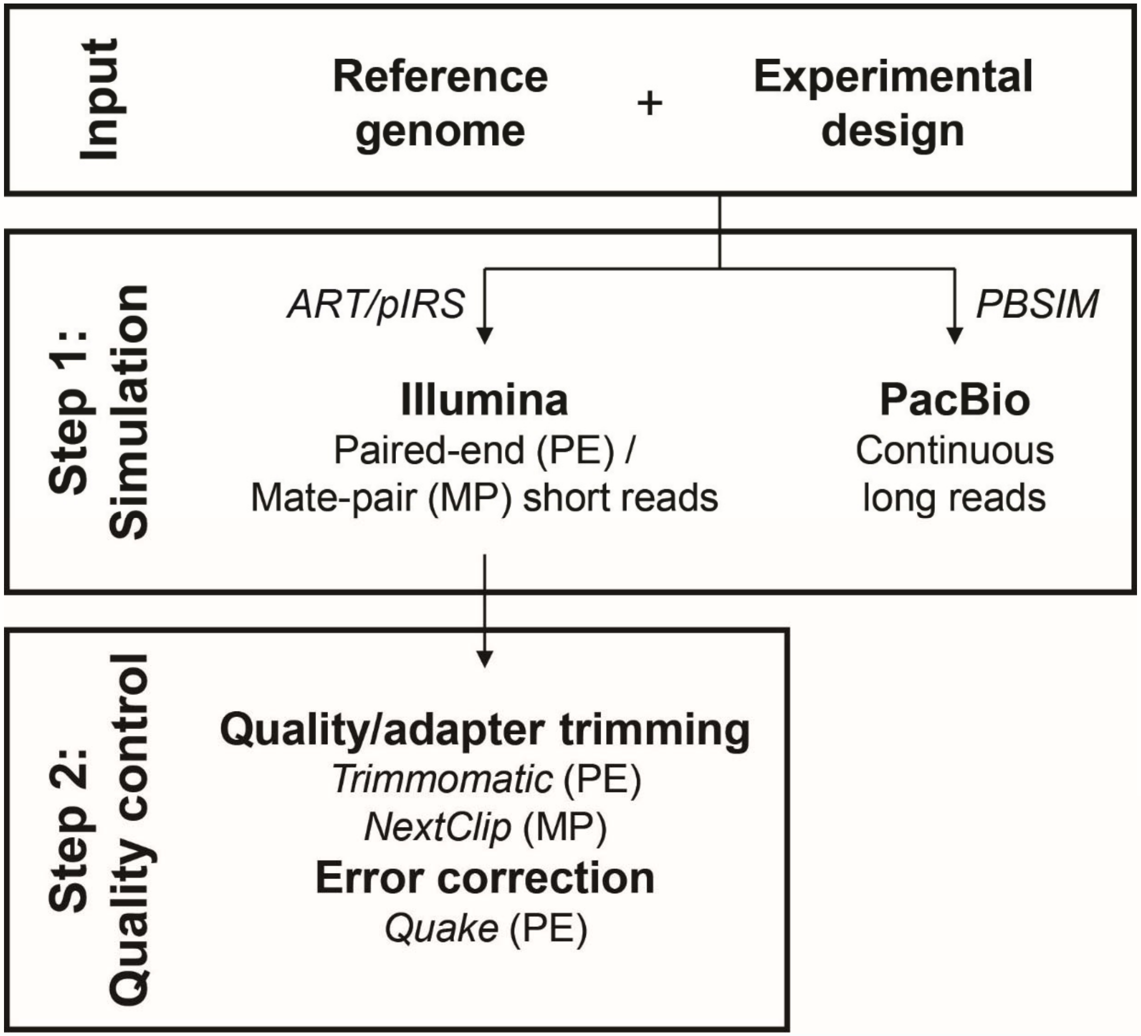
- ❖ Sequence the genome of every known of the ~1,000+ yeast species
- ❖ Construct their definitive phylogeny
- ❖ Revise their taxonomy
- ❖ Examine the impact of metabolism on yeast diversification

http://y1000plus.org

developing pipelines for handling genomic and phylogenomic data

In silico *Whole Genome Sequencer and Analyzer (iWGS)*

iWGS seamlessly integrates the four key steps of a *de novo* genome sequencing project:



Input

Reference genome + **Experimental design**

Step 1: Simulation

ART/pIRS

PBSIM

Illumina
Paired-end (PE) /
Mate-pair (MP) short reads

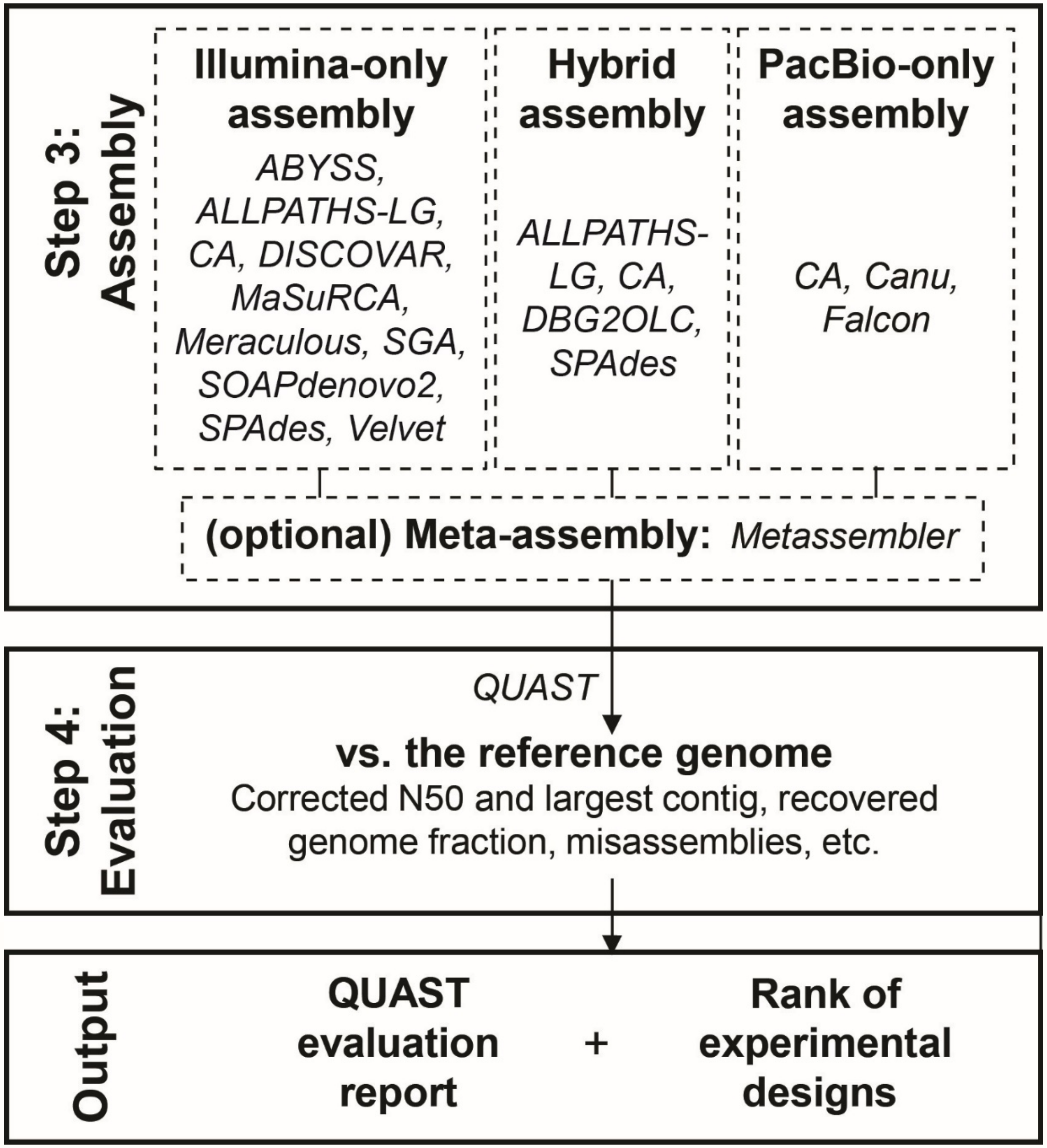
PacBio
Continuous
long reads

Step 2: Quality control

Quality/adapter trimming
Trimmomatic (PE)
NextClip (MP)
Error correction
Quake (PE)

Zhou et al. (2016) G3

In silico *Whole Genome Sequencer and Analyzer (iWGS)*



Step 3: Assembly

Illumina-only assembly
ABYSS, ALLPATHS-LG, CA, DISCOVAR, MaSuRCA, Meraculous, SGA, SOAPdenovo2, SPAdes, Velvet

Hybrid assembly
ALLPATHS-LG, CA, DBG2OLC, SPAdes

PacBio-only assembly
CA, Canu, Falcon

(optional) Meta-assembly: *Metassembler*

Step 4: Evaluation

QUAST

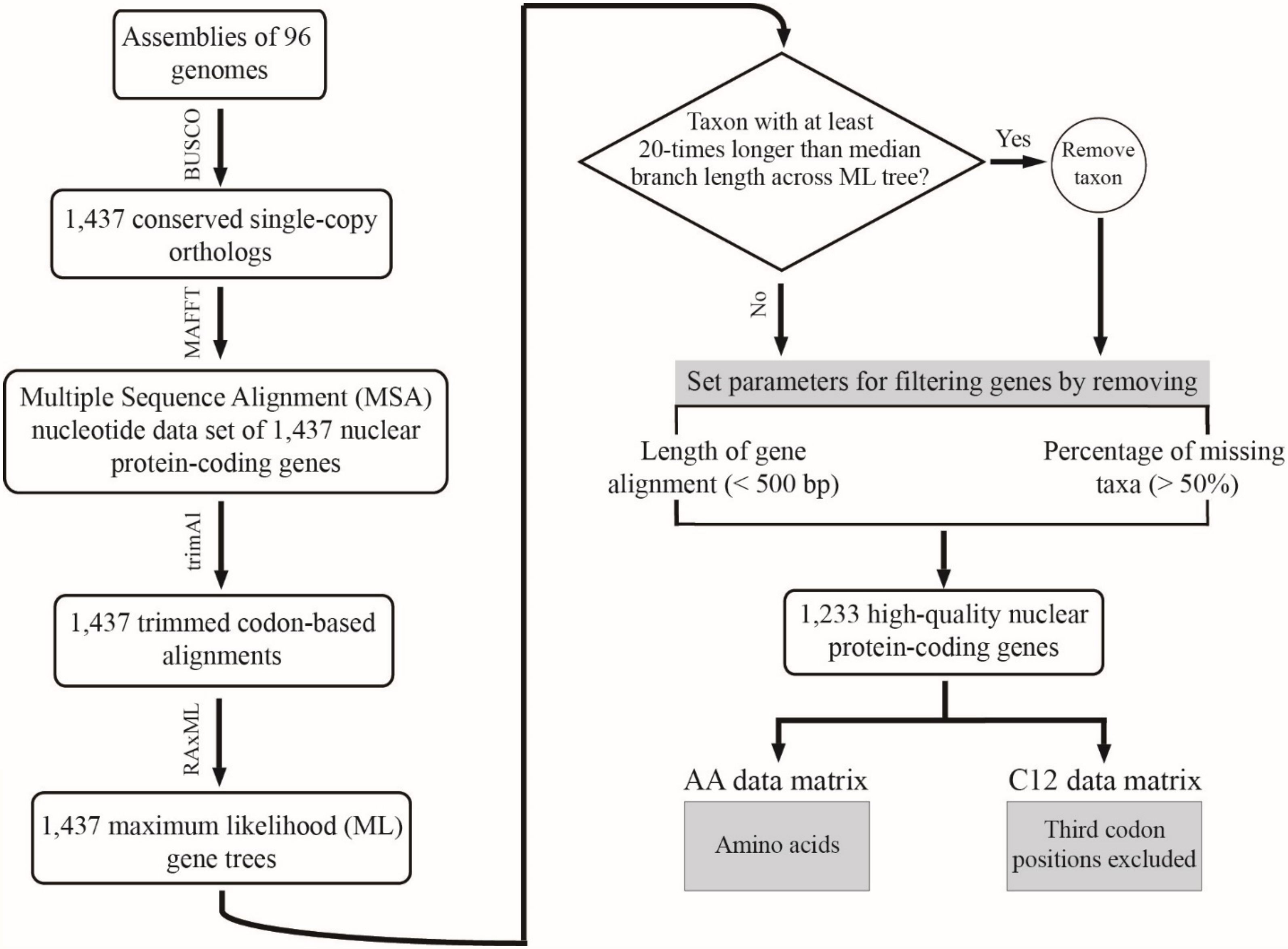
vs. the reference genome
Corrected N50 and largest contig, recovered genome fraction, misassemblies, etc.

Output

QUAST evaluation report + **Rank of experimental designs**

Zhou et al. (2016) G3

Developing Pipelines for Phylogenomic Studies



Assemblies of 96 genomes

BUSCO

1,437 conserved single-copy orthologs

MAFFT

Multiple Sequence Alignment (MSA) nucleotide data set of 1,437 nuclear protein-coding genes

trimAl

1,437 trimmed codon-based alignments

RAxML

1,437 maximum likelihood (ML) gene trees

Taxon with at least 20-times longer than median branch length across ML tree?

Yes

Remove taxon

No

Set parameters for filtering genes by removing

Length of gene alignment (< 500 bp)

Percentage of missing taxa (> 50%)

1,233 high-quality nuclear protein-coding genes

AA data matrix

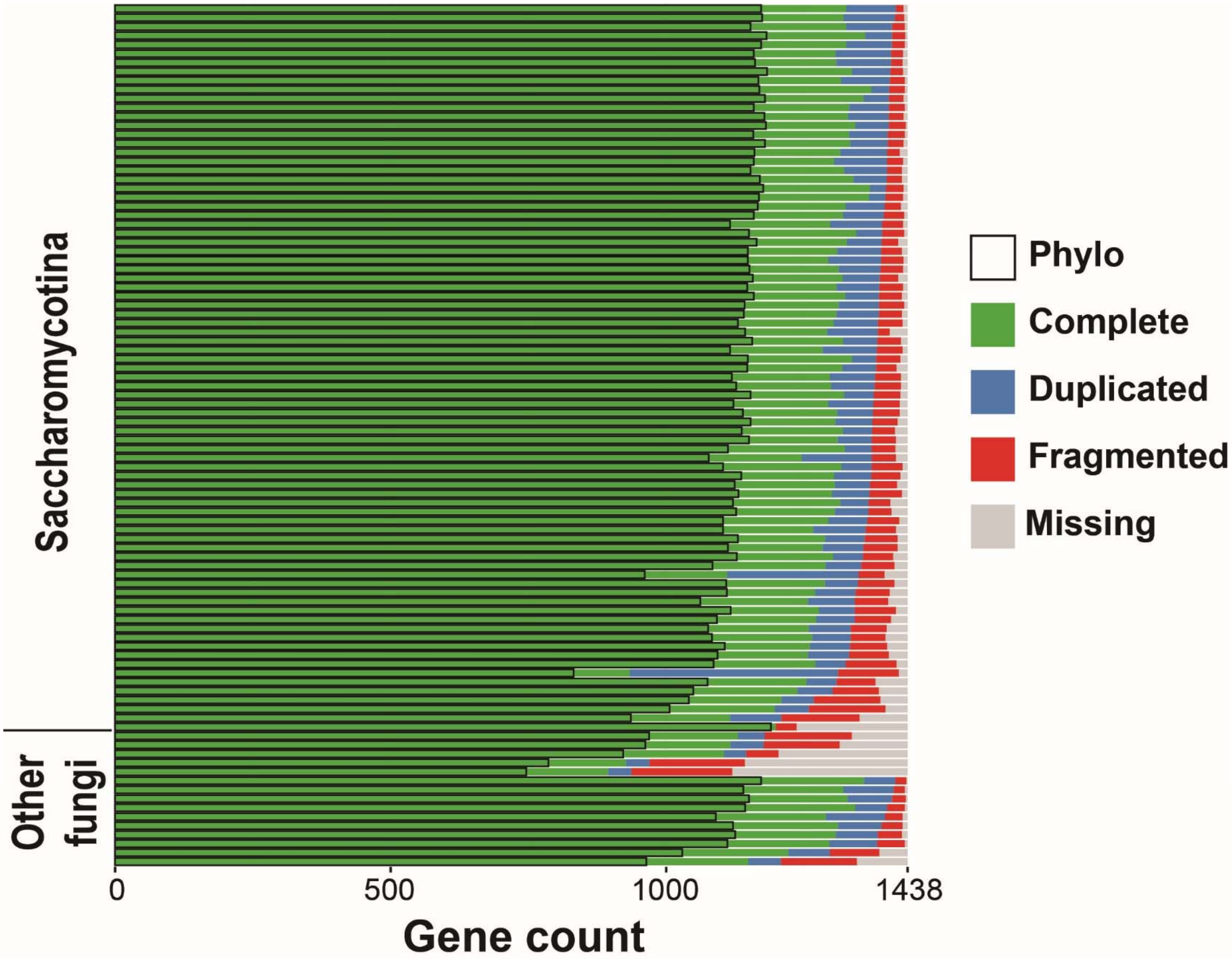
Amino acids

C12 data matrix

Third codon positions excluded

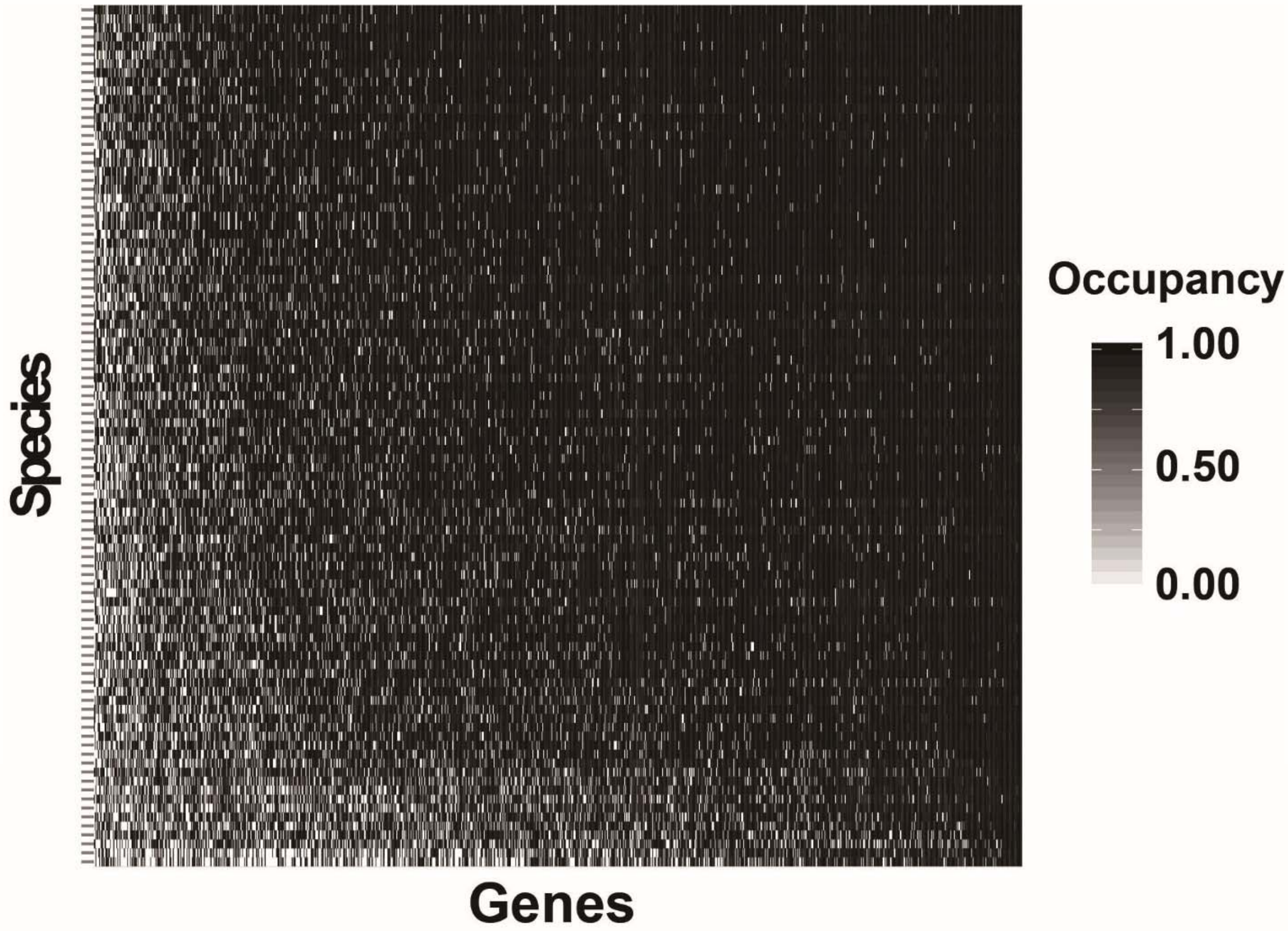
Shen et al. (2016) G3

Identifying Molecular Markers for Phylogenomics



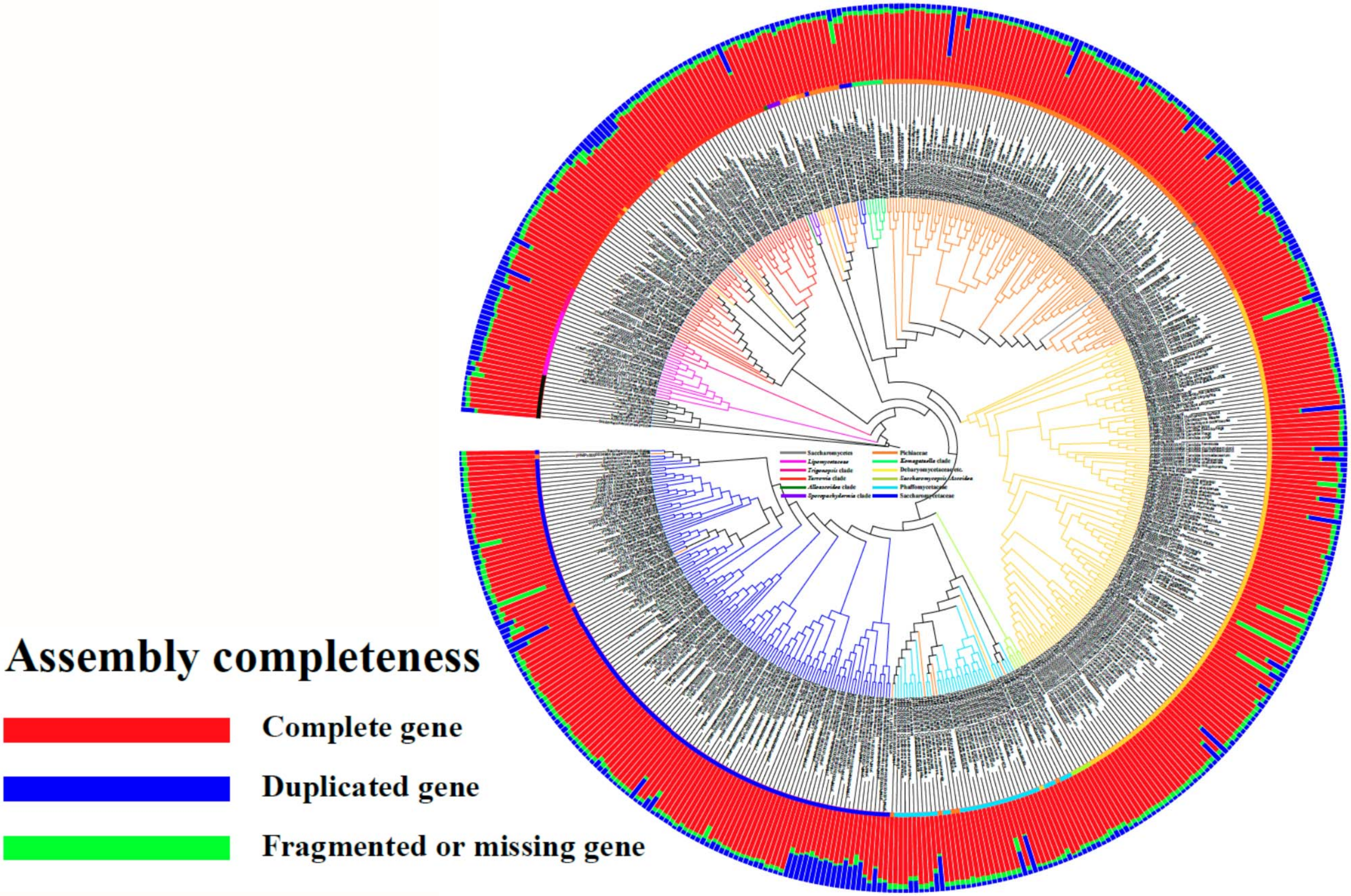
Shen et al. (2016) G3

Assessing Occupancy of the Phylogenomic Data Matrix



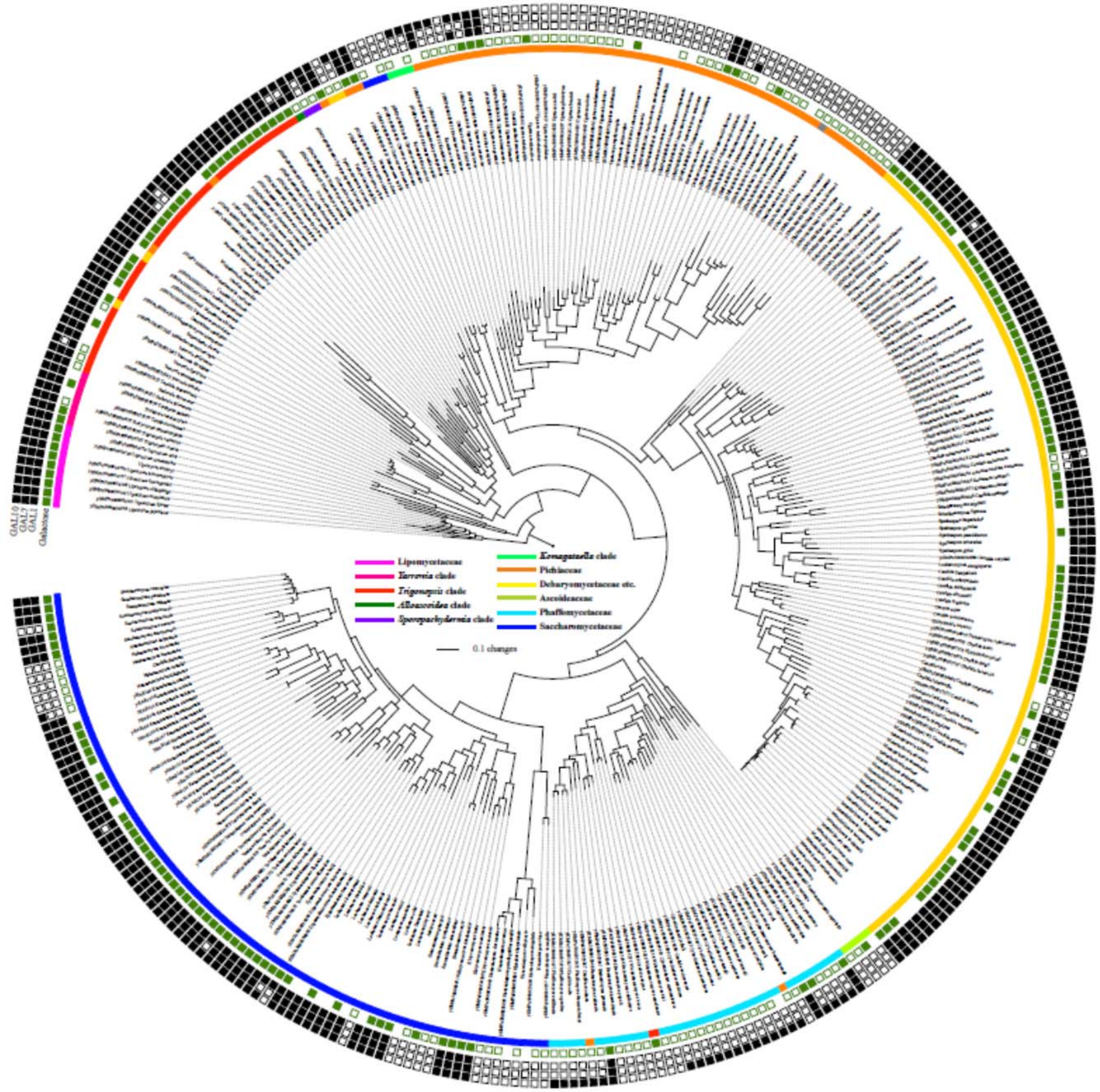
Shen et al. (2016) G3

So Far, We Have Sequenced ~530 Genomes



Xing-Xing Shen, Xiaofan Zhou, Dana Opulente, Jacek Kominek and co-workers

Inferring Evolution of Traits on the Yeast Phylogeny



Xing-Xing Shen, Xiaofan Zhou, Dana Opulente, Jacek Kominek and co-workers

Mind the Gap Between Real Data and Models

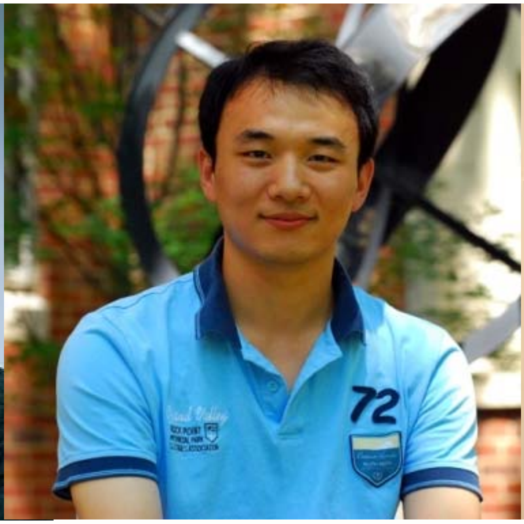
“One can use the most sophisticated audio equipment to listen, for an eternity, to a recording of white noise and still not glean a useful scrap of information”

Rodrigo et al. (1994)
Chapter in: Sponge in Time and Space; Biology, Chemistry, Paleontology

Acknowledgements

Leonidas Salichos

Xing-Xing Shen



Xiaofan Zhou

Alexis Stamatakis

National Science Foundation
WHERE DISCOVERIES BEGIN

http://as.vanderbilt.edu/rokaslab

Rokas Lab

polytomies happen...