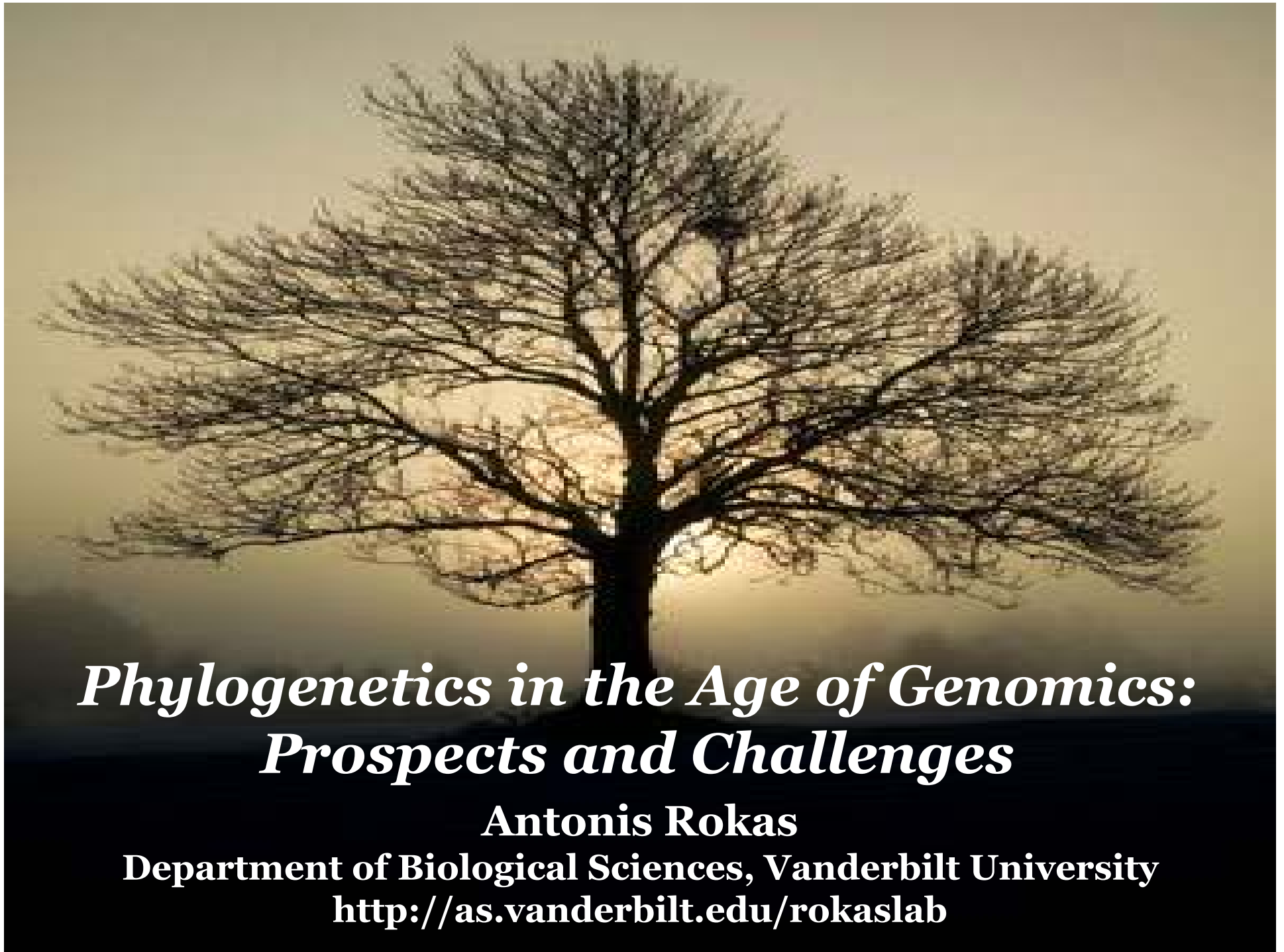


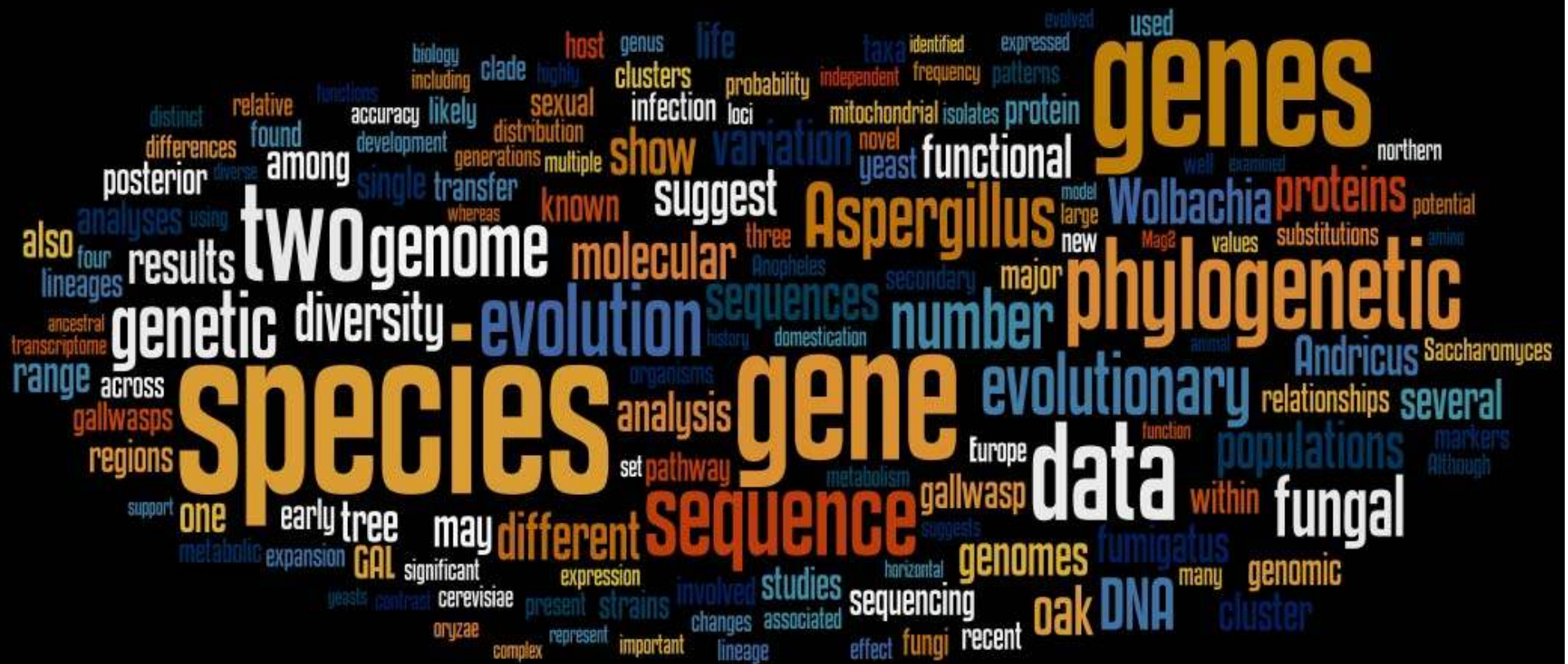


Phylogenetics in the Age of Genomics: Prospects and Challenges

Antonis Rokas

Department of Biological Sciences, Vanderbilt University

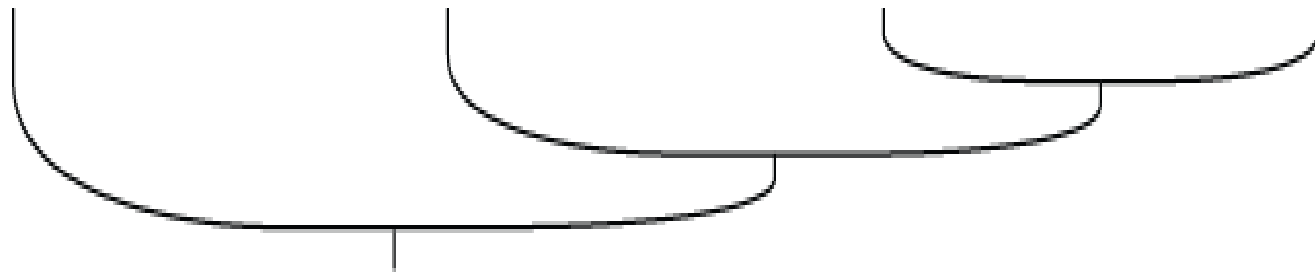
<http://as.vanderbilt.edu/rokaslab>



<http://pubmed2wordle.appspot.com/>

What is a Phylogenetic Tree?

A phylogenetic tree is the mathematical structure used to depict the evolutionary history of a group of organisms or genes



**Phylogenetic trees show historical relationships,
not similarities**



Why are Phylogenetic Trees Useful?

Benefits to Science:

- ❖ **The origin and history of life**
- ❖ **Evolution of molecules, phenotypes and developmental mechanisms**

Benefits to Society:

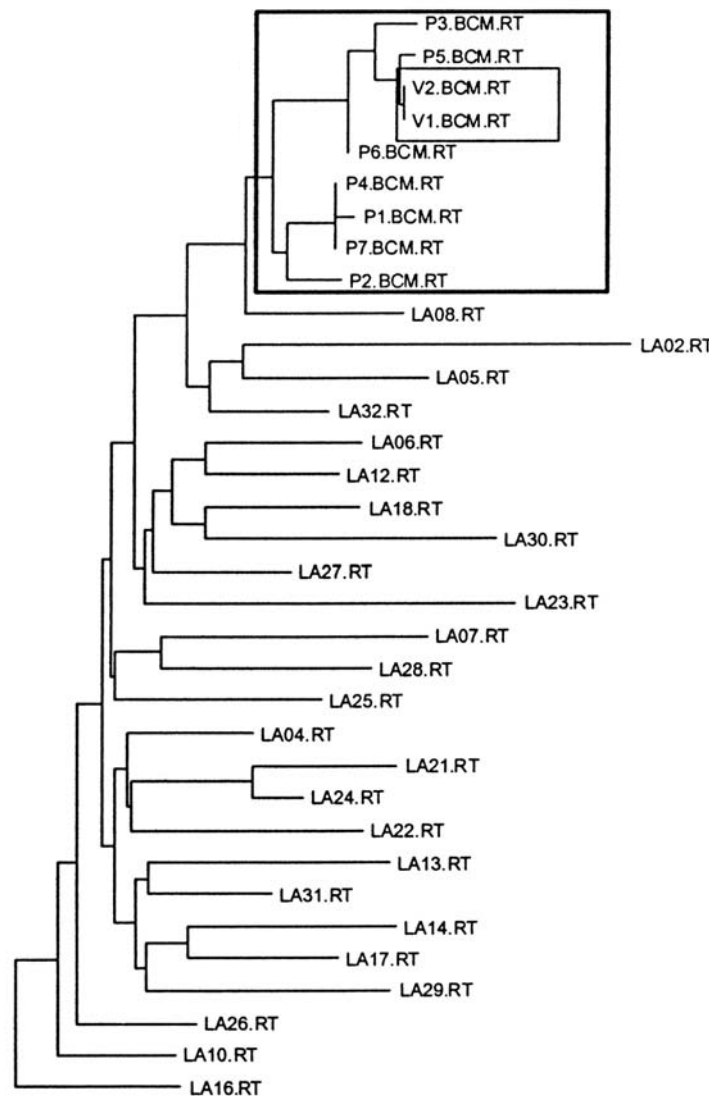
- ❖ **Human health (identification of disease agents & reservoirs)**
- ❖ **Agriculture (crops' wild relatives)**
- ❖ **Biodiversity (conservation strategies)**



CSI: Phylogeny

❖ A gastroenterologist was accused of second-degree murder for injecting his former girlfriend with blood obtained from an HIV type 1 (HIV-1)-infected patient under his care

❖ Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in the court

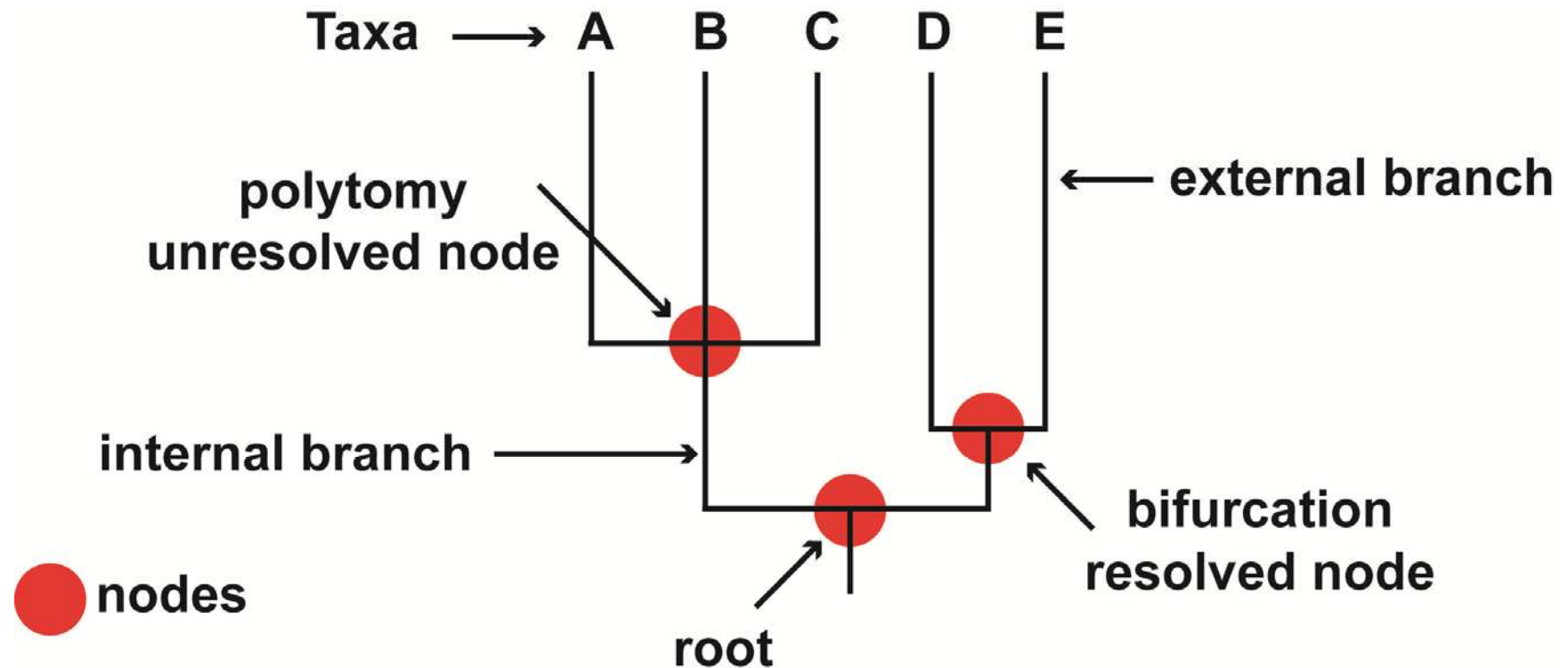


**Patient and
Victim HIV-1
sequences**

**Sequences
from local
population
sample of
HIV-1
infected
individuals**



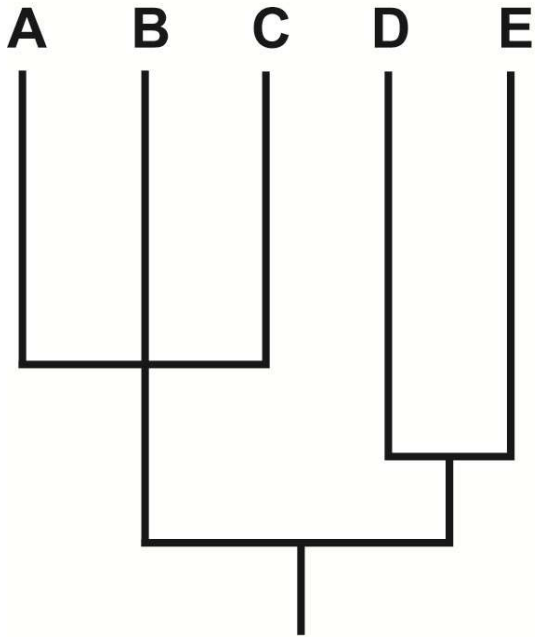
Phylogenetic Concepts and Terminology



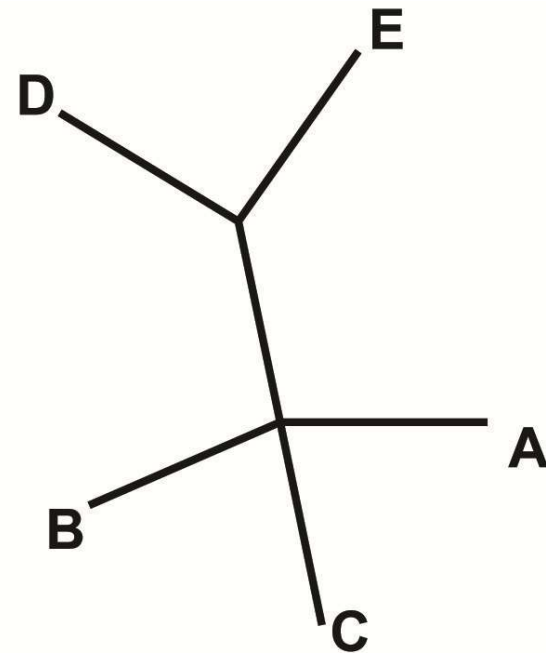
tree / topology | taxon - taxa | branches, nodes & internodes | polytomy & resolution



Rooted and Unrooted Trees



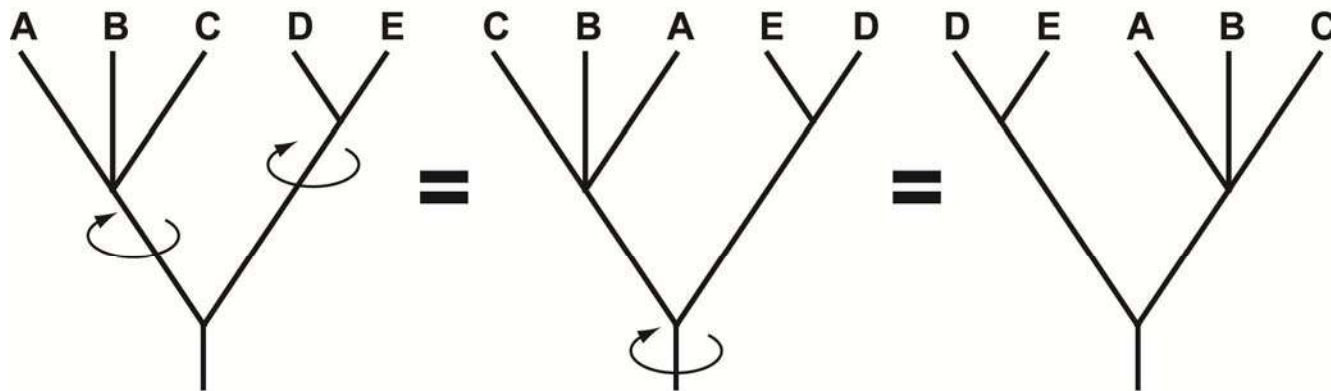
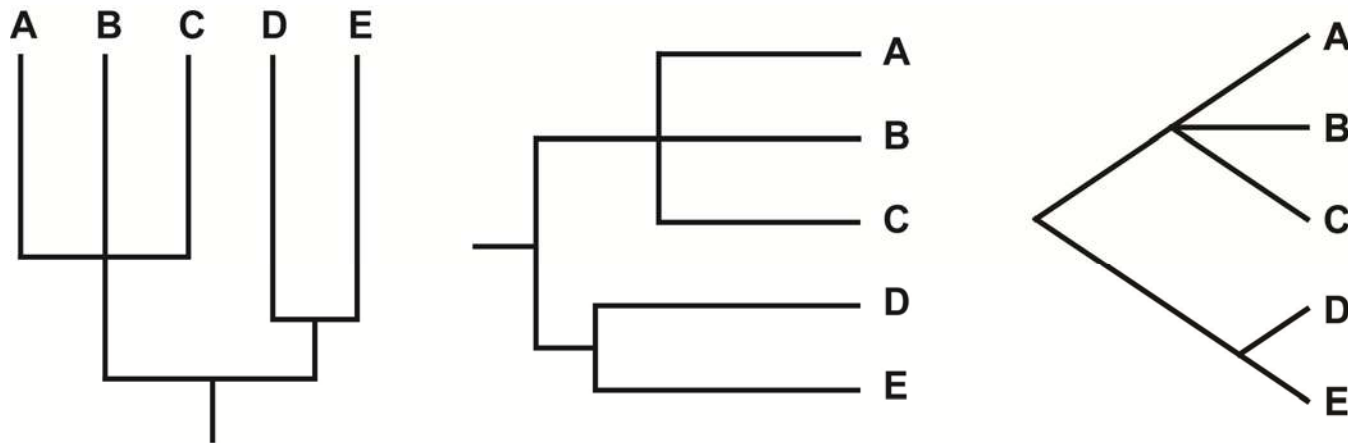
Rooted topology



Unrooted topology



Phylogenetic Concepts and Terminology

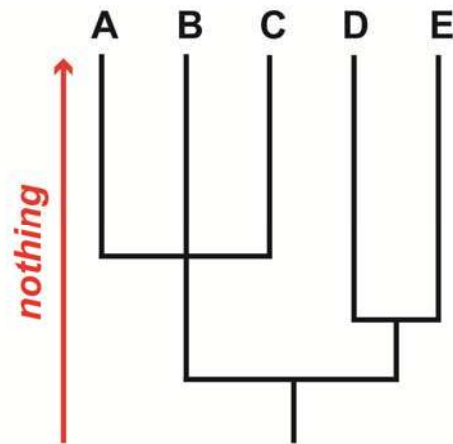


Newick standard: ((A,B,C),(D,E));

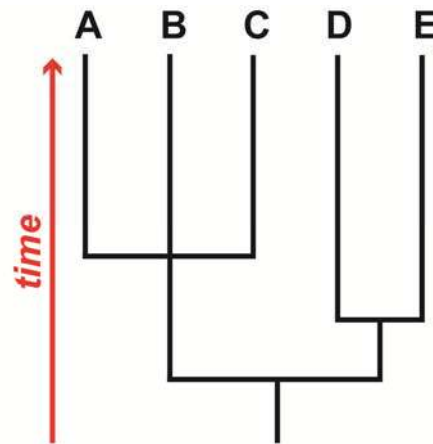


The Meaning of Vertical and Horizontal Axes in Trees

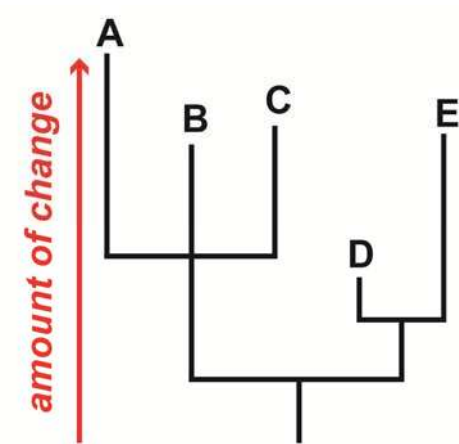
Vertical Axis:



Cladograms

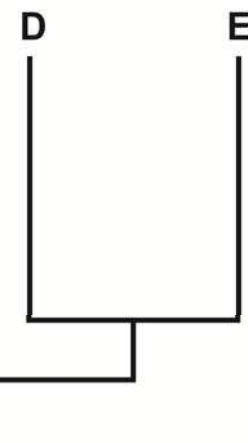
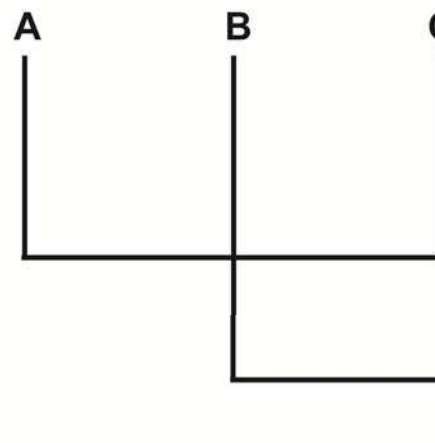
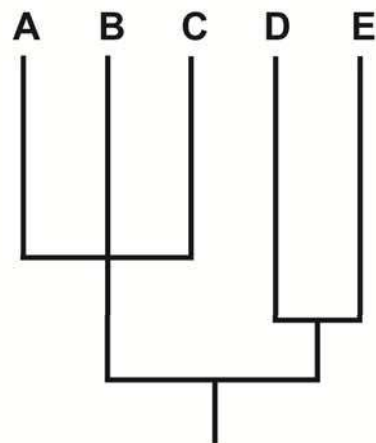


Chronograms



Phylograms

Horizontal Axis:



nothing



Estimating Phylogenies

1. Select an *optimality criterion*
2. Select a *search strategy*
3. Use selected search strategy to generate a series of trees, and apply selected optimality criterion to each, always keeping track of the 'best' tree(s) examined thus far



Optimality Criteria

Criterion	Data
Minimum Evolution	pairwise distances
Least Squares	pairwise distances
Parsimony	discrete characters
Maximum Likelihood	discrete characters
Bayesian Inference	discrete characters

Distance methods first convert an alignment into a matrix of pairwise distances

Discrete character methods use the alignment as is and consider directly each site/character



Search Strategies

There is a need for a search strategy because the number of possible trees goes up very fast with increasing number of species

$$B(T) = \prod_{n=3}^T (2n - 5)$$

Number of taxa T	Number of unrooted bifurcating trees $B(T)$
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
22	3×10^{23}
50	3×10^{74}

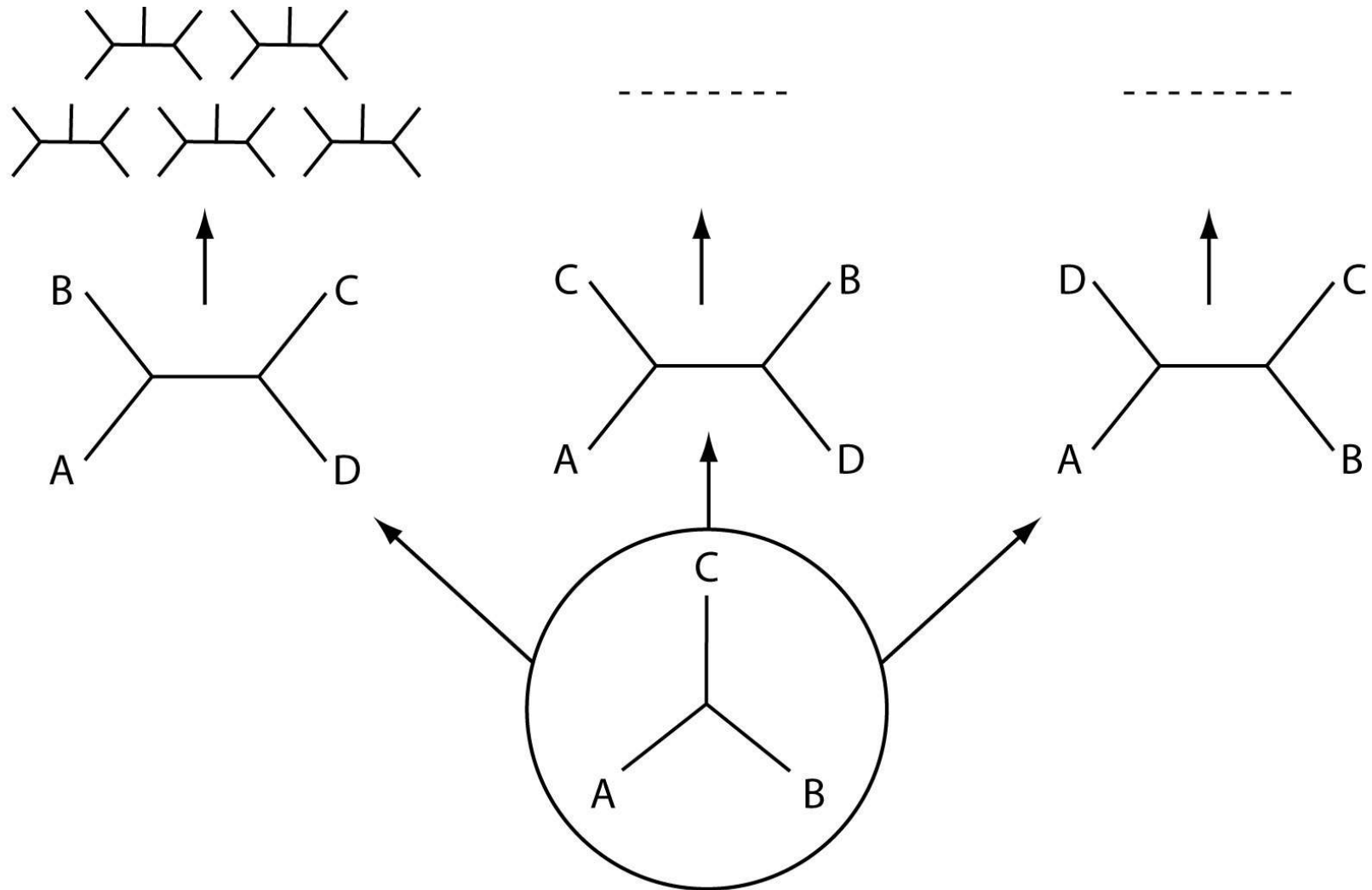
Search Strategy

Exhaustive

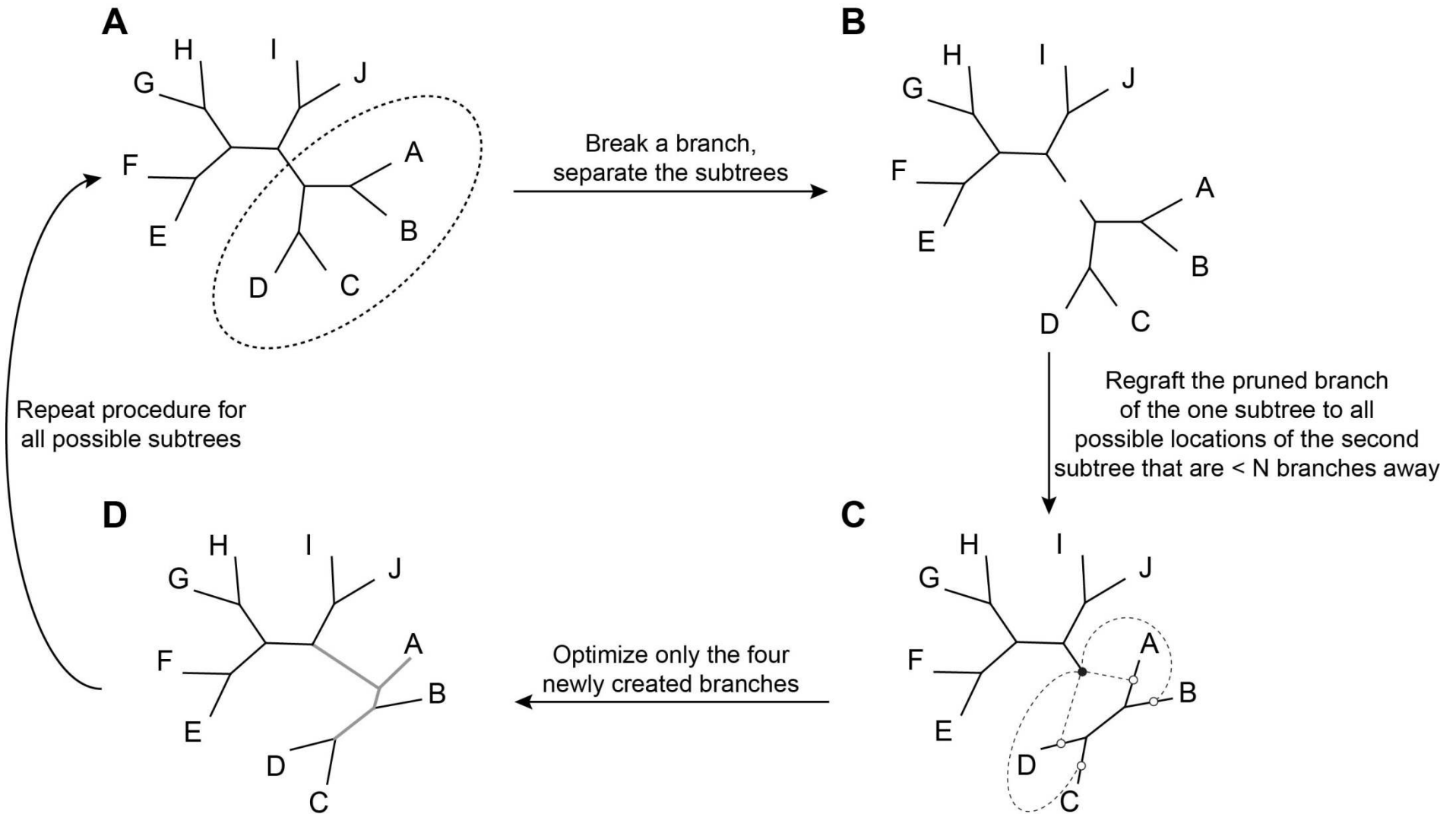
Heuristic



Exhaustive Search



Heuristic Search



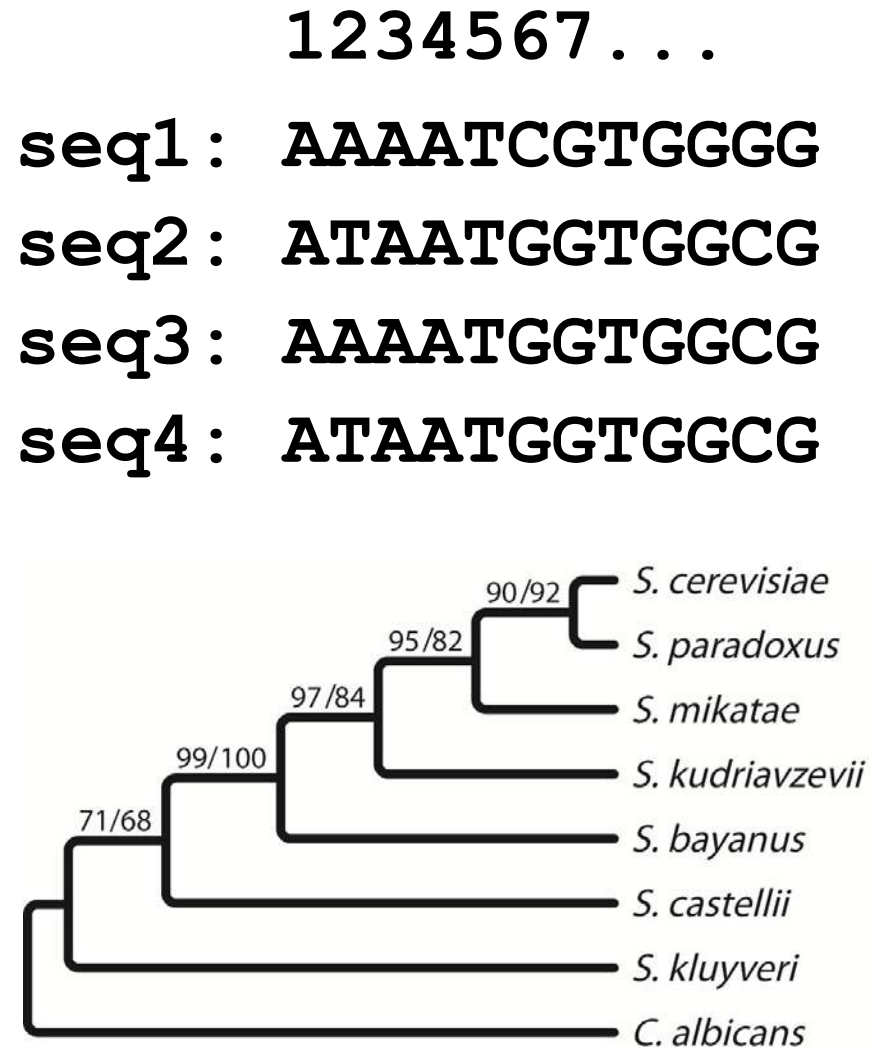
Heuristic searches don't guarantee discovery of optimal topology



The lazy subtree rearrangement (LSR) tree search strategy used by the RAxML program

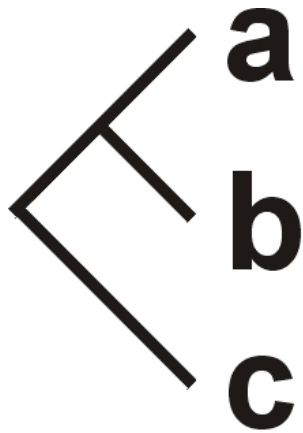
Assessing Robustness in Inference: Bootstrap

- ❖ Characters are sampled with replacement to create many bootstrap replicate data sets equal in size to the original one
- ❖ Each bootstrap replicate data set is analysed (e.g., with parsimony, likelihood)
- ❖ Agreement among the resulting trees is summarized with a majority-rule consensus tree
- ❖ The frequency of occurrence of clades, also known as bootstrap values, is a measure of support for those groups

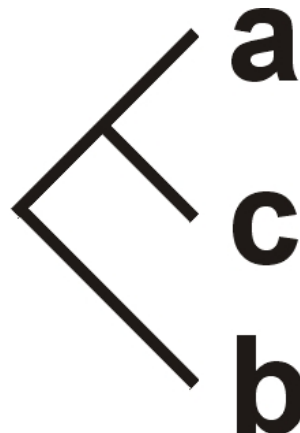


The Problem of Incongruence

Gene X

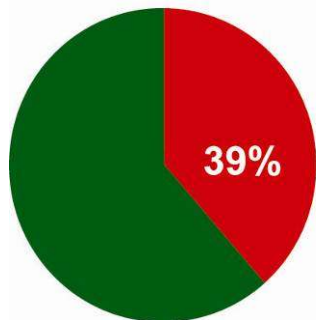


Gene Y

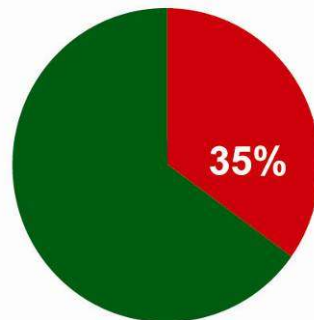


Species tree?

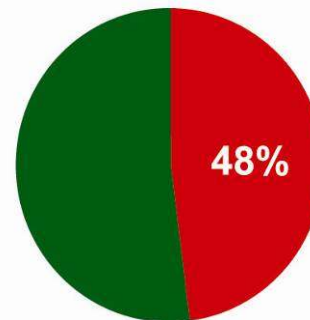
A: All organisms



B: Mammals



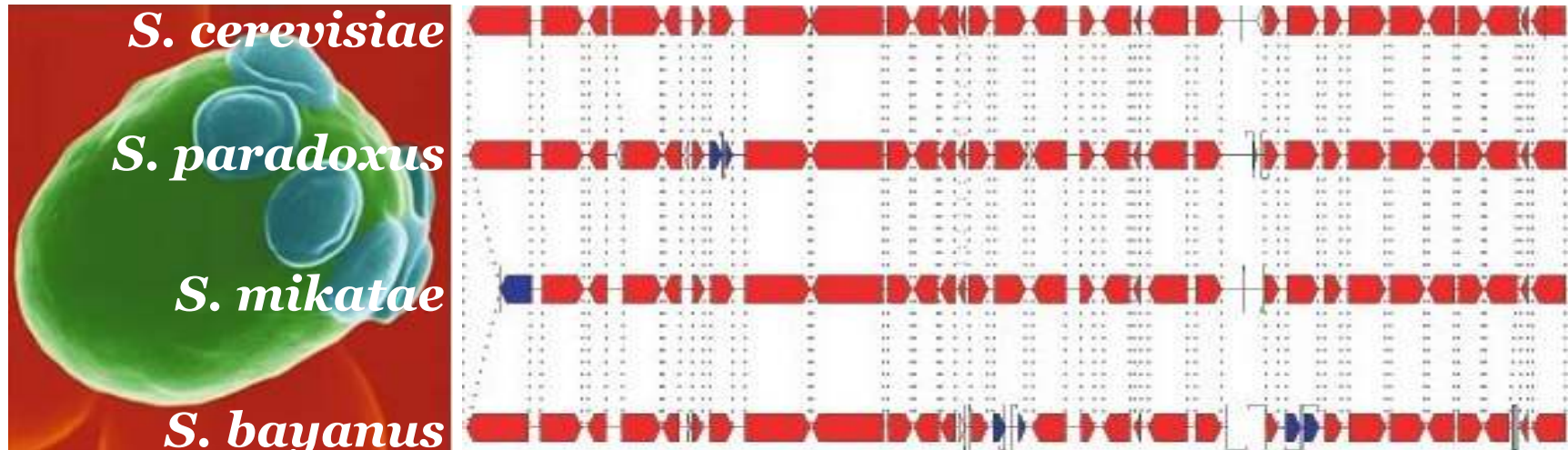
C: Insects



Incongruence is
pervasive in the
phylogenetics
literature



A Systematic Evaluation of Single Gene Phylogenies



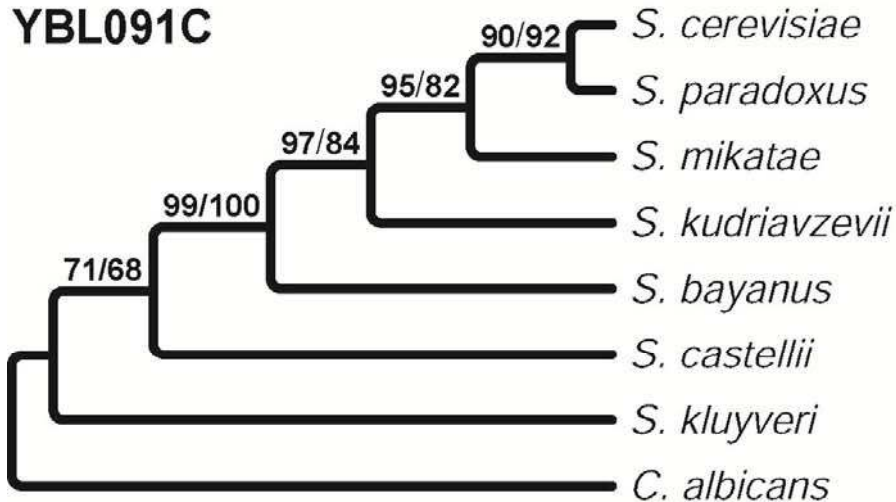
Dataset: 106 genes on all 16 chromosomes totaling 127kb corresponding roughly to 1% of the genomic sequence, 2% of genes

Analyses: Maximum Likelihood (ML) & Maximum Parsimony (MP) on nt data sets and MP on amino acid data sets

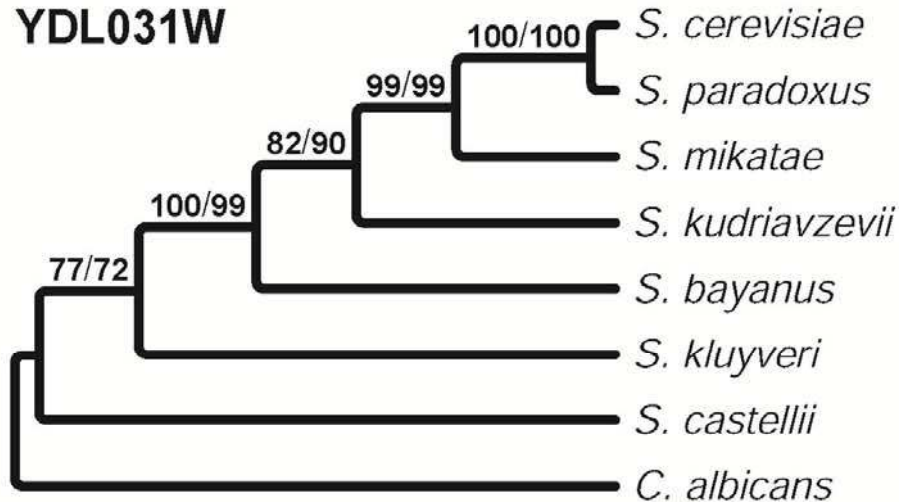


Incongruence in Shallow Time

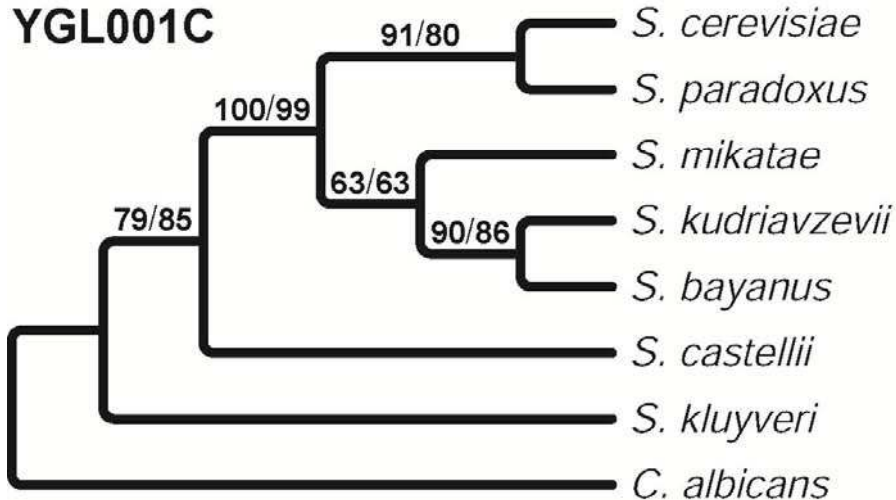
YBL091C



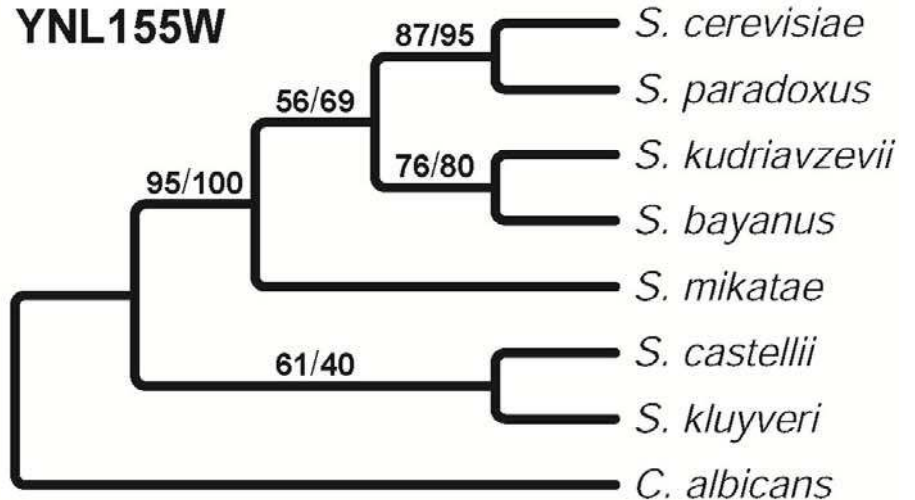
YDL031W



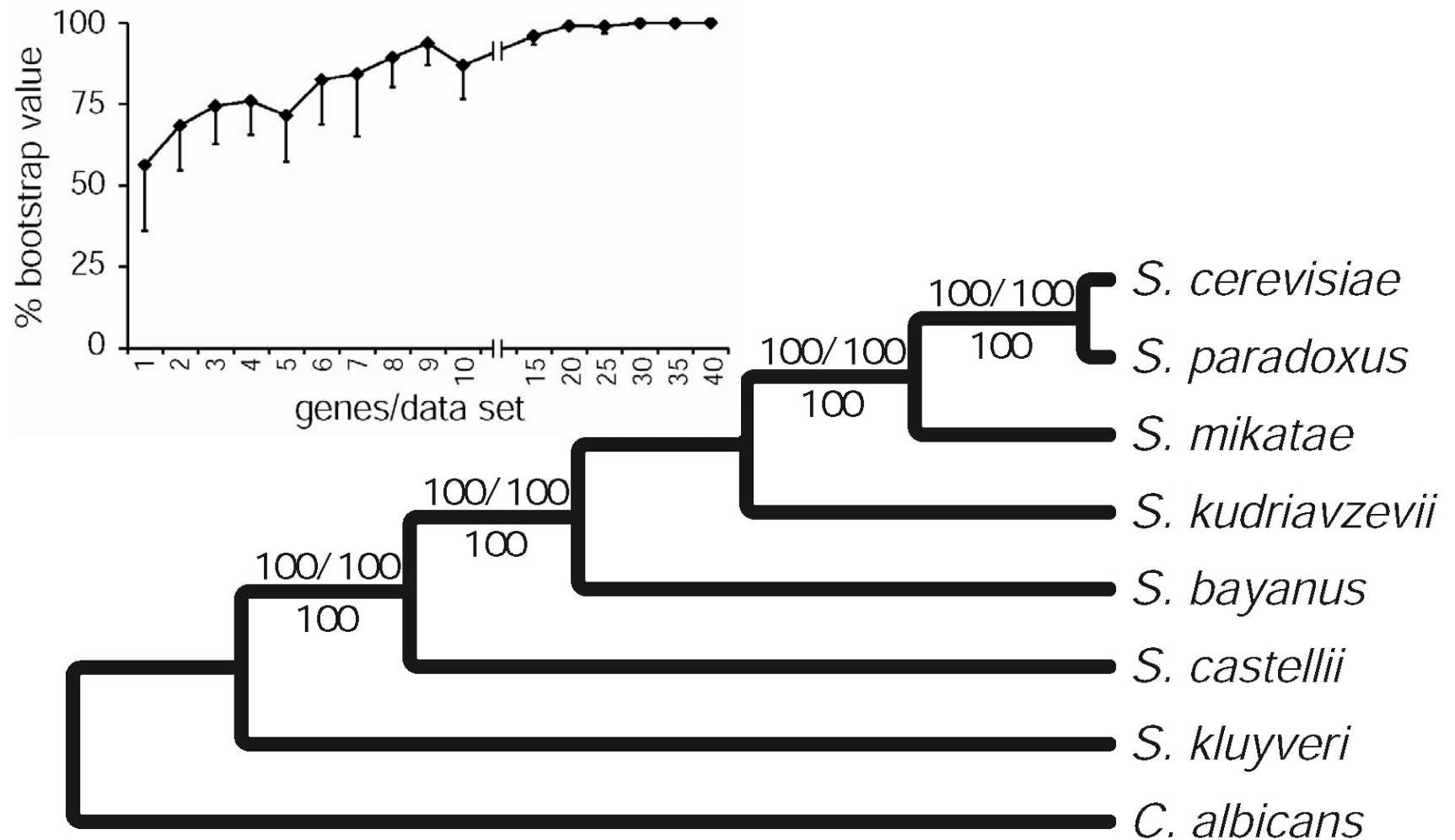
YGL001C



YNL155W



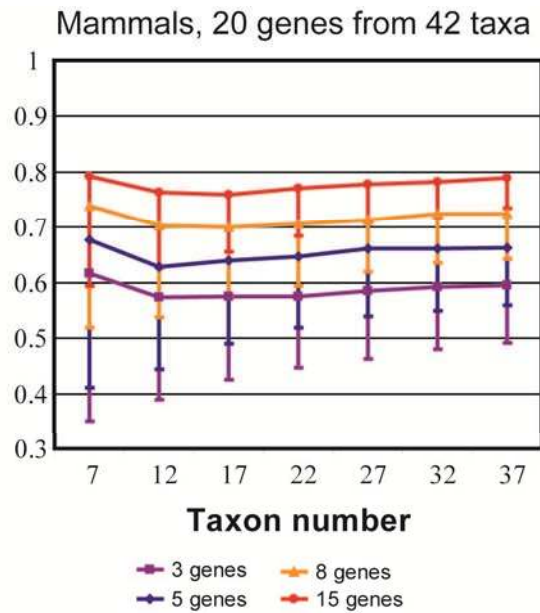
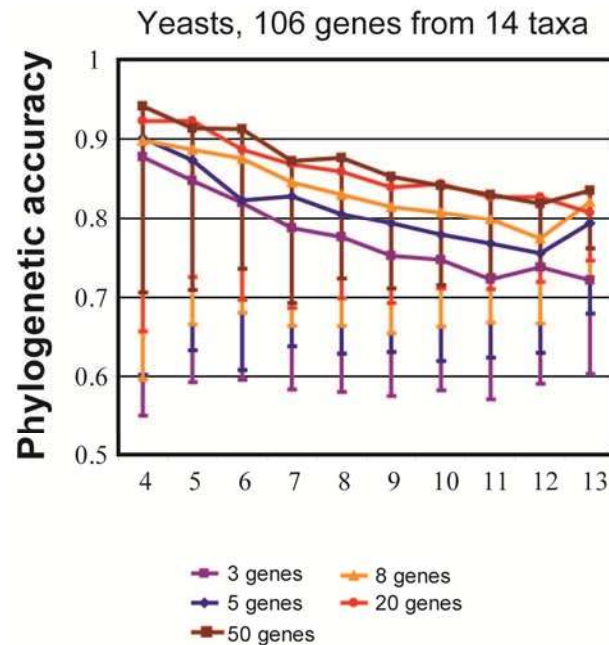
Concatenation of 106 Genes Yields a Single Yeast Phylogeny



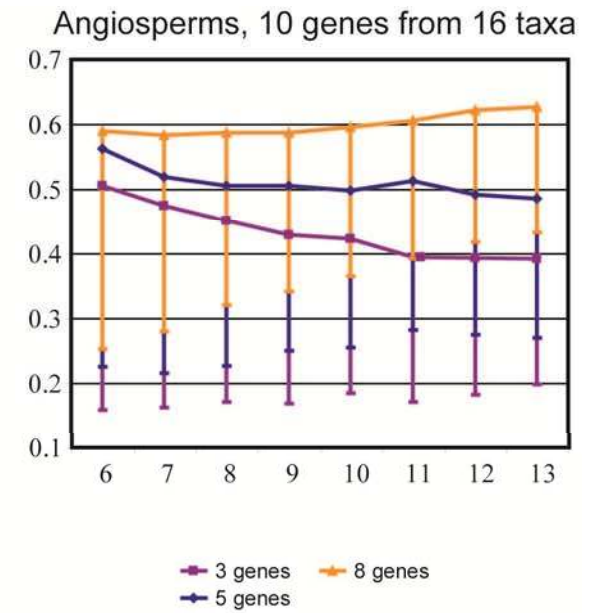
ML / MP on nt
MP on aa

Rokas et al. (2003) Nature

Phylogenetic Accuracy is Positively Correlated with Gene Number



Murphy et al. (2001) Science

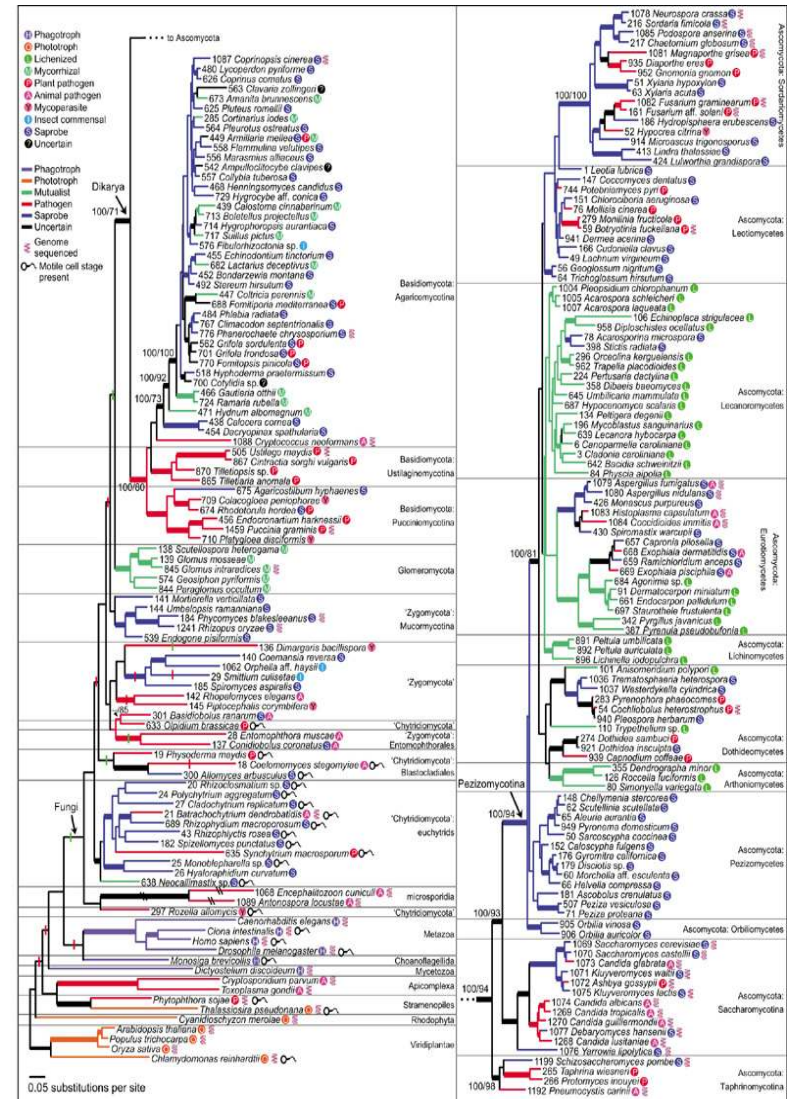
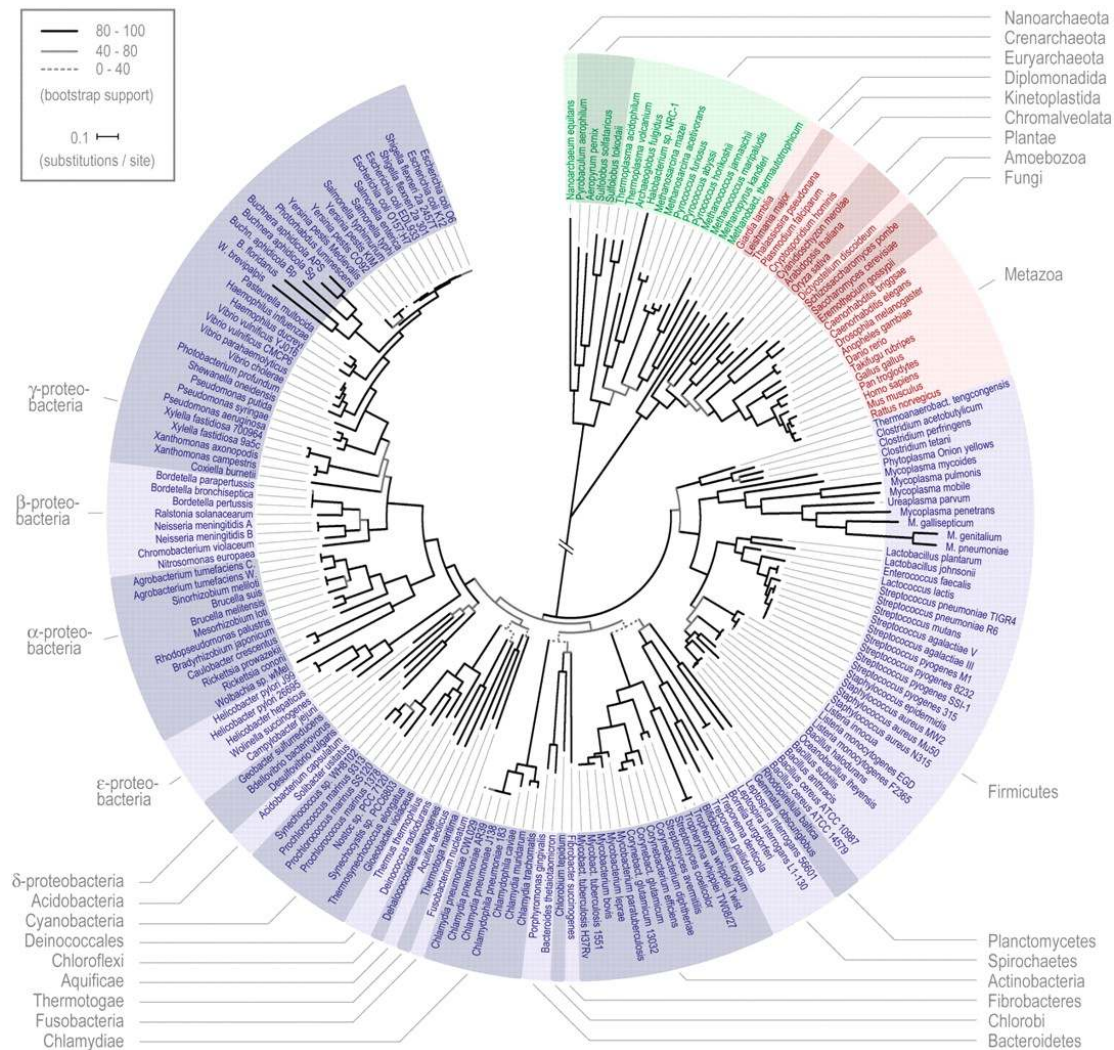


Zanis et al. (2002) PNAS

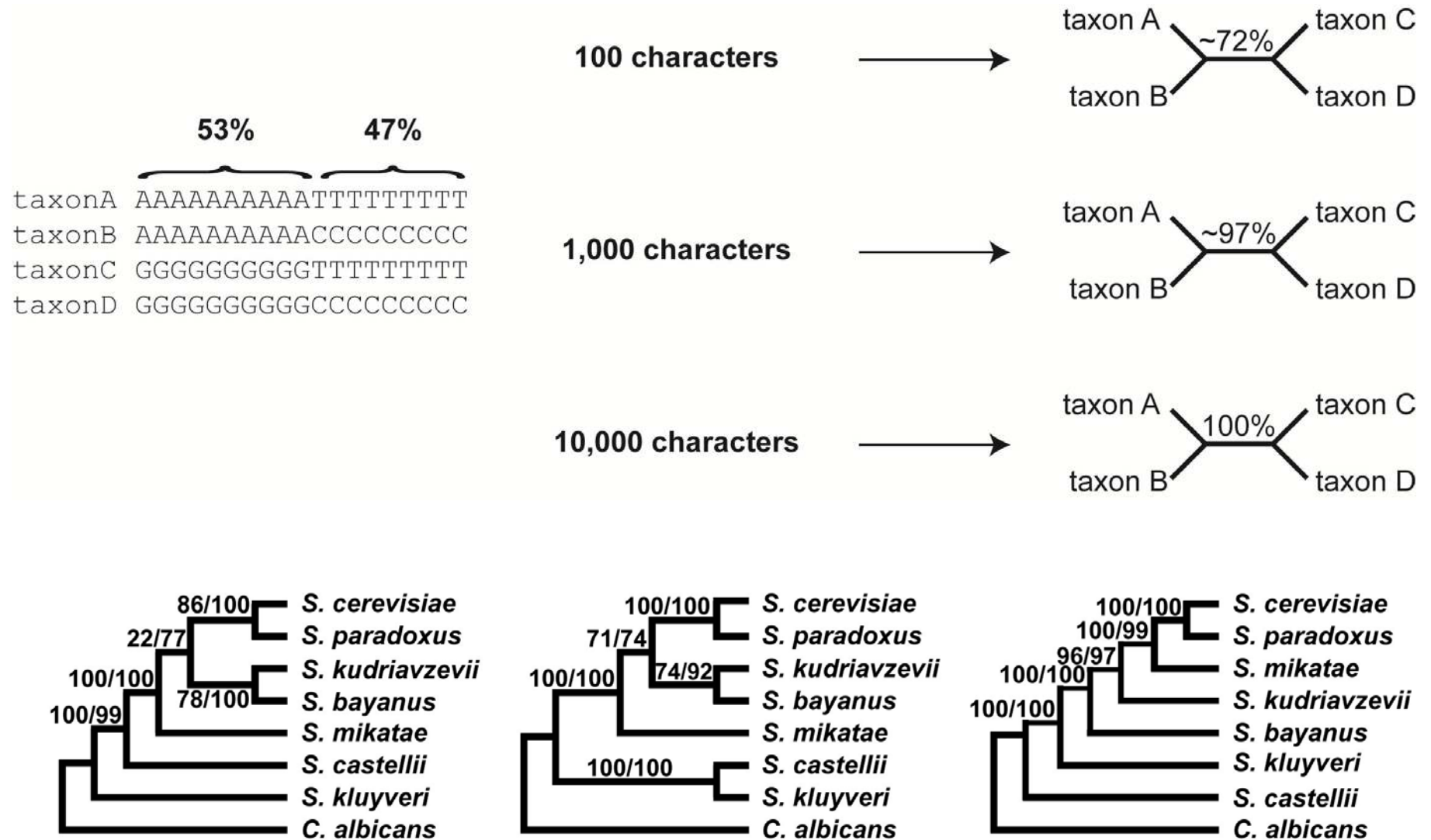


Rokas & Carroll (2005) Mol. Biol. Evol.

Reconstructing the Tree of Life

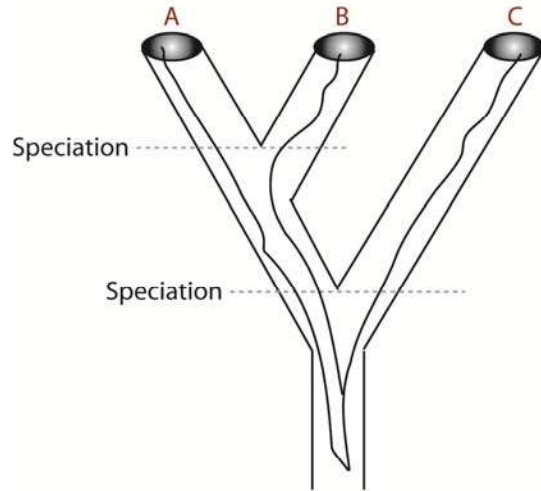


Methods of Assessing Robustness are Subject to Systematic Error

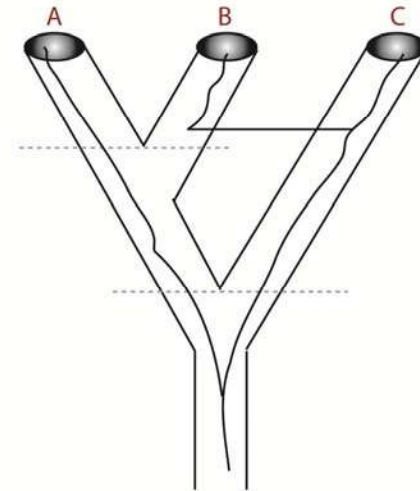


Gene Trees Can Differ from Species Trees

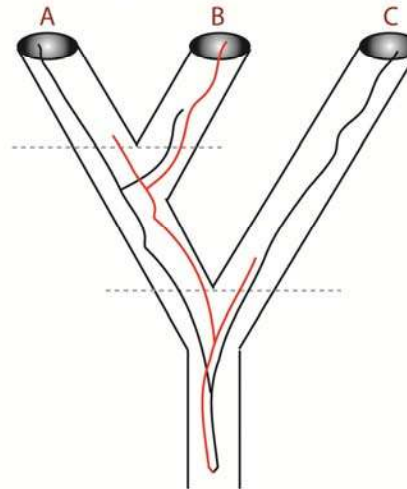
Lineage Sorting



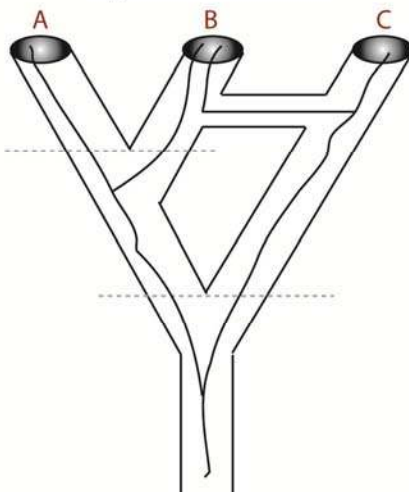
Horizontal Gene Transfer



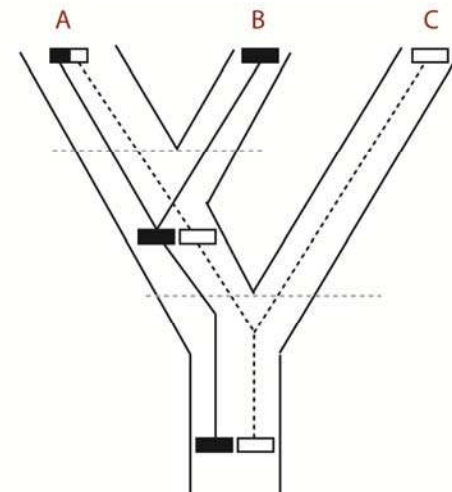
Gene Duplication and Loss



Hybridization

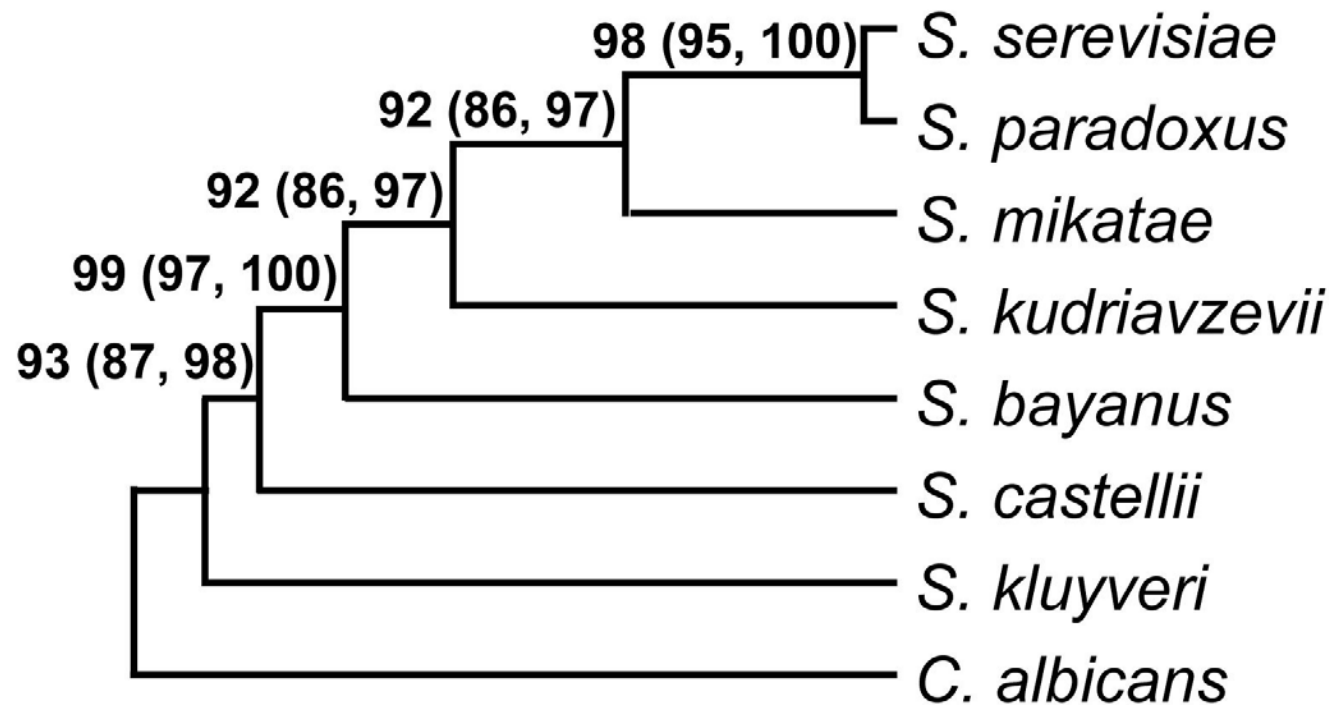


Recombination



Inferring the Species Tree from Individual Gene Histories

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



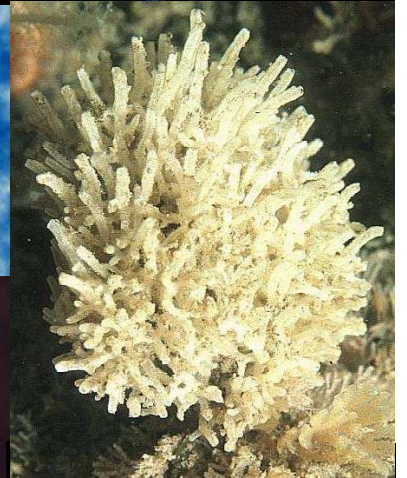
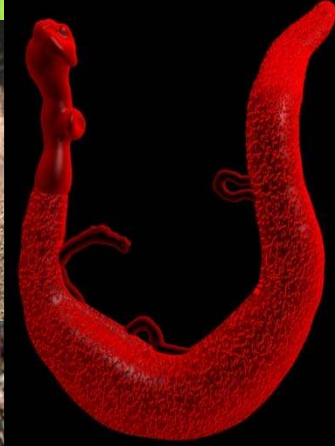
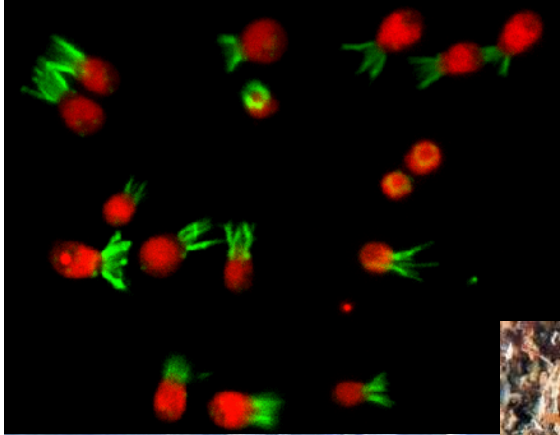
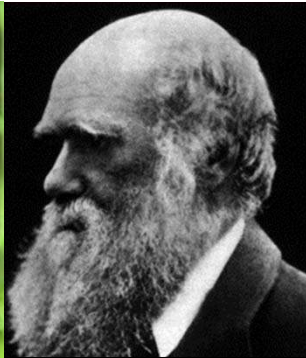
STEM: Maximum likelihood estimation of species trees

Kubatko et al.(2009) Bioinformatics

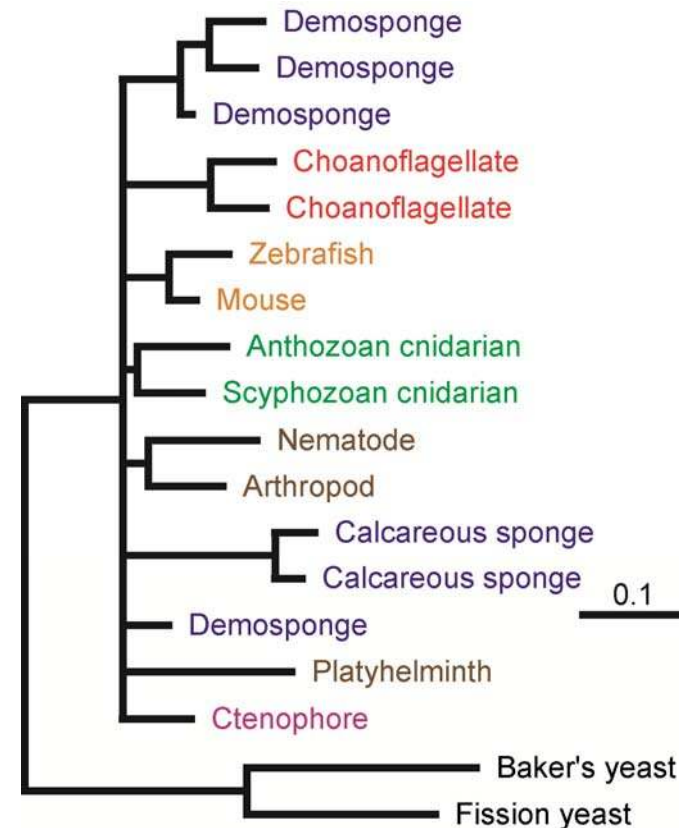
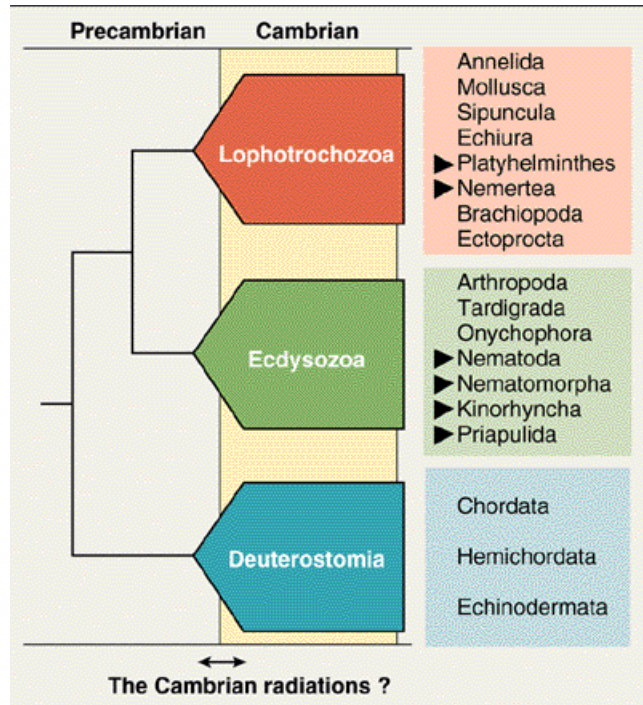
BEST: Bayesian estimation of species trees

Liu (2008) Bioinformatics

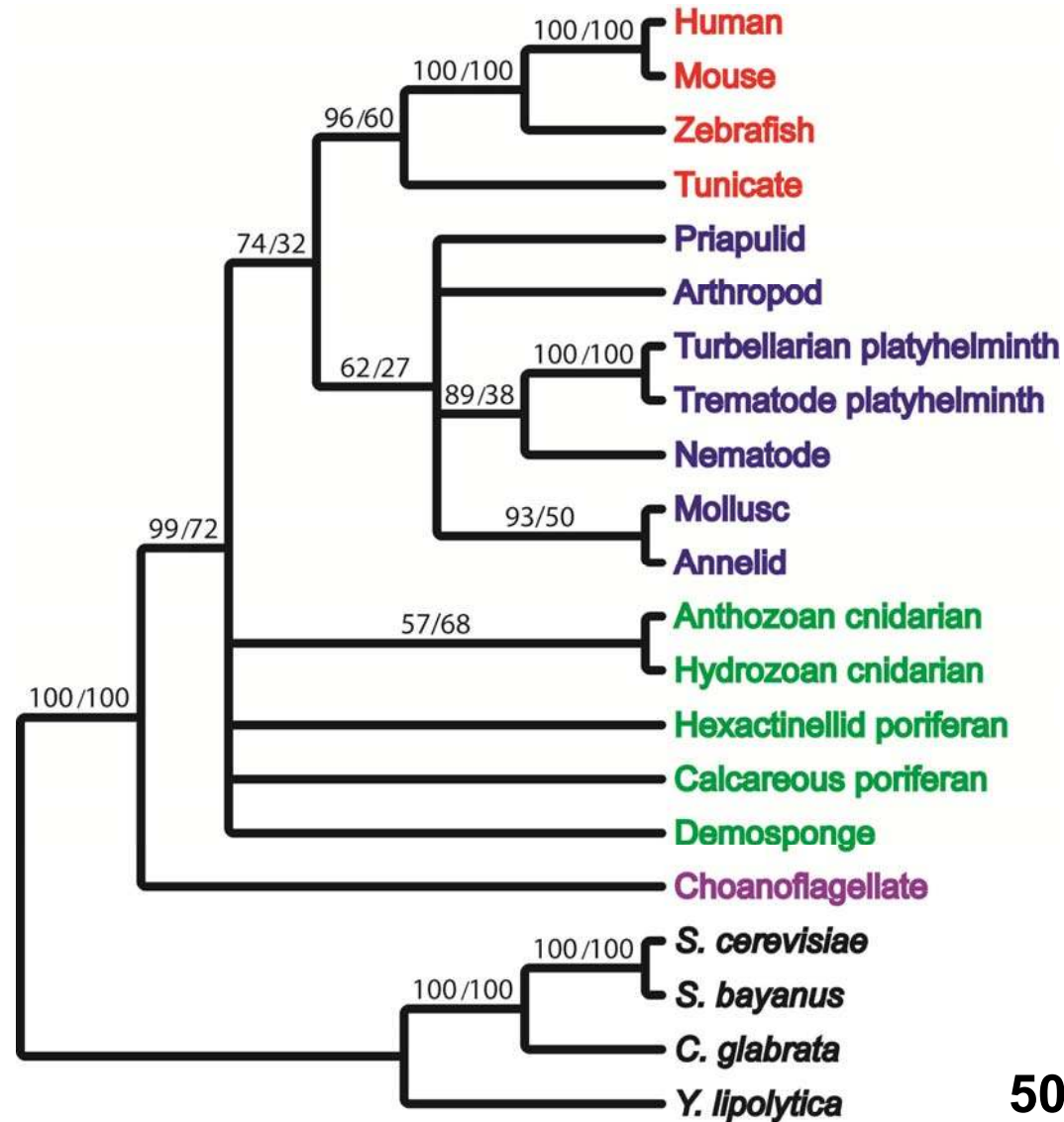




Lack of Resolution Due to Lack of Data or to Radiation?



Lack of Resolution Among Most Metazoan Phyla



50 genes, ML/MP



Biological Explanations for the Lack of Resolution

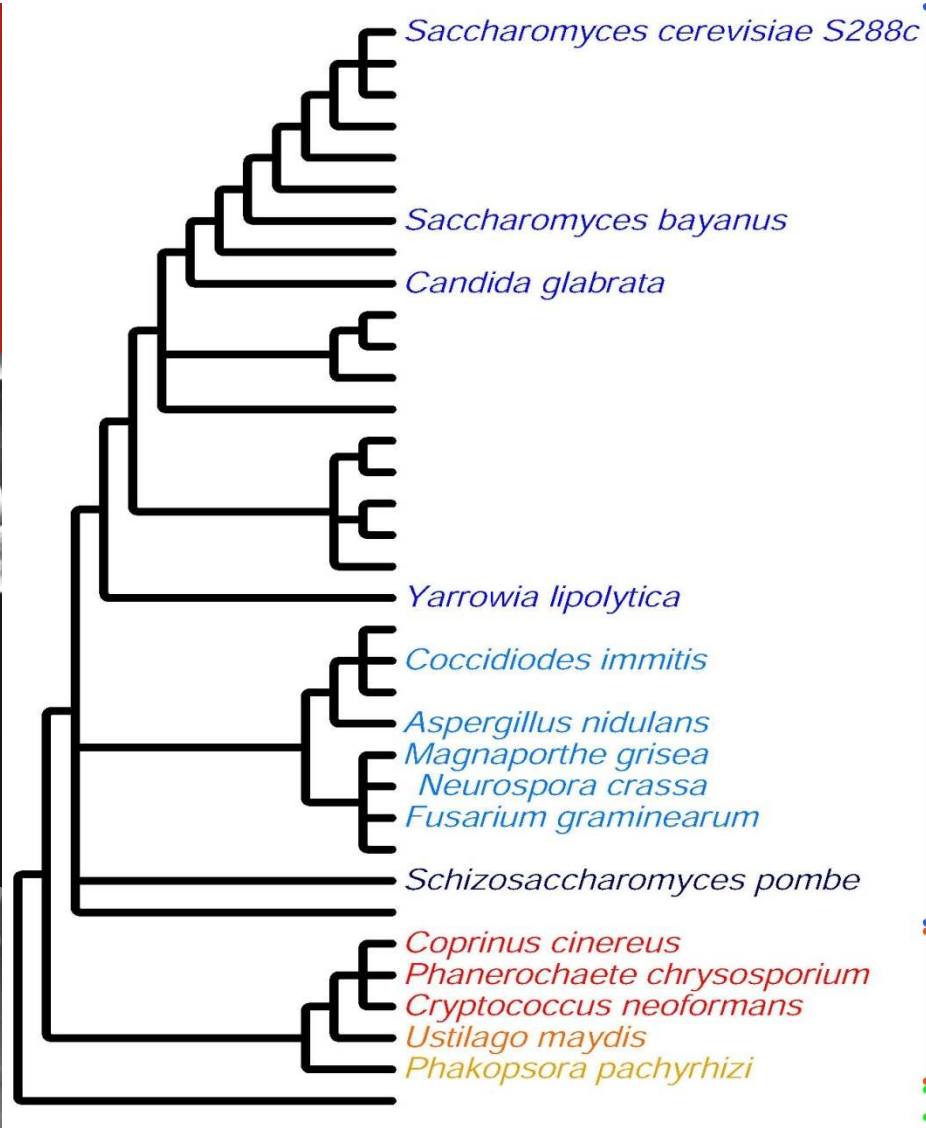
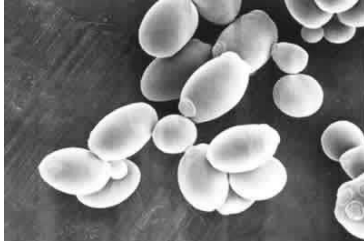
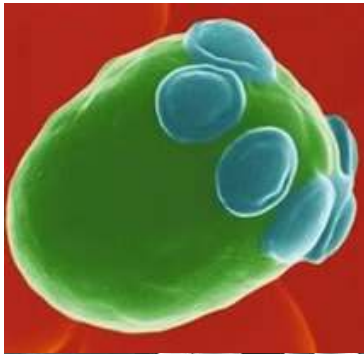
Hypothesis I: Mutational Saturation

the phylogenetic signal originally contained in these protein sequences has been erased by multiple substitutions

Hypothesis II: Evolutionary Radiation

the close spacing of cladogenetic events early in the history of metazoans limits their resolution

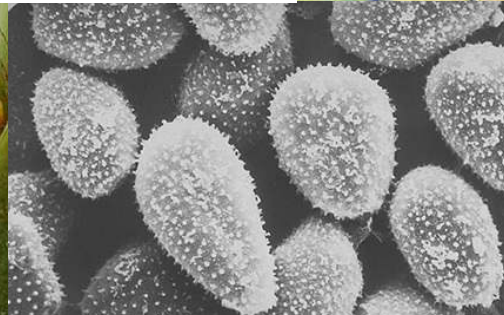
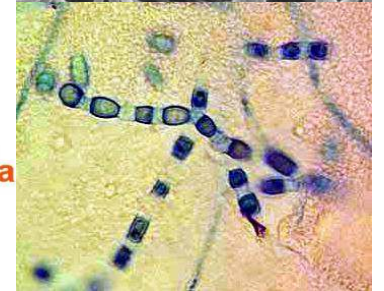
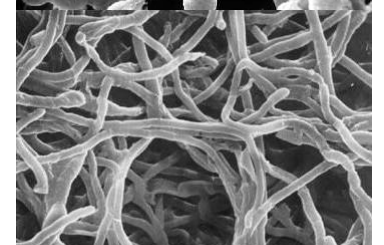
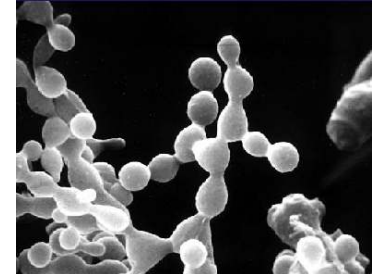
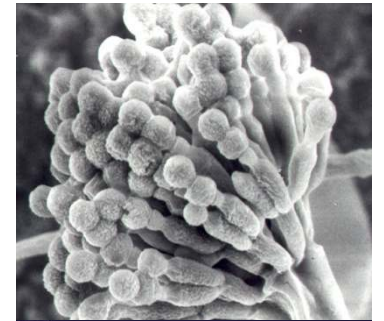




Ascomycota

Basidiomycota

Zygomycota

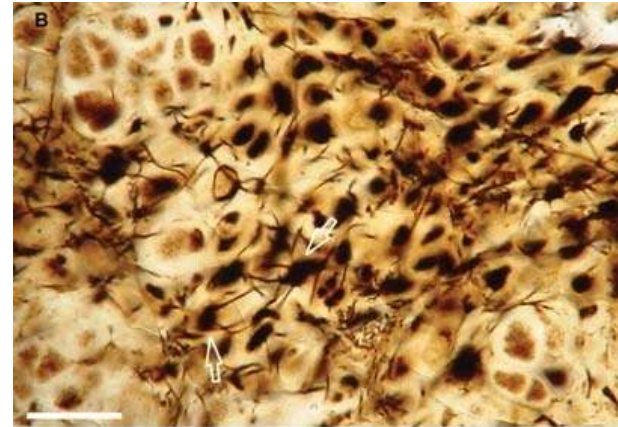


Time of Origin of Metazoa & Fungi

1) Fossils



Metazoa: *Xiao et al. (1998) Nature*



Fungi: *Yuan et al. (2005) Science*

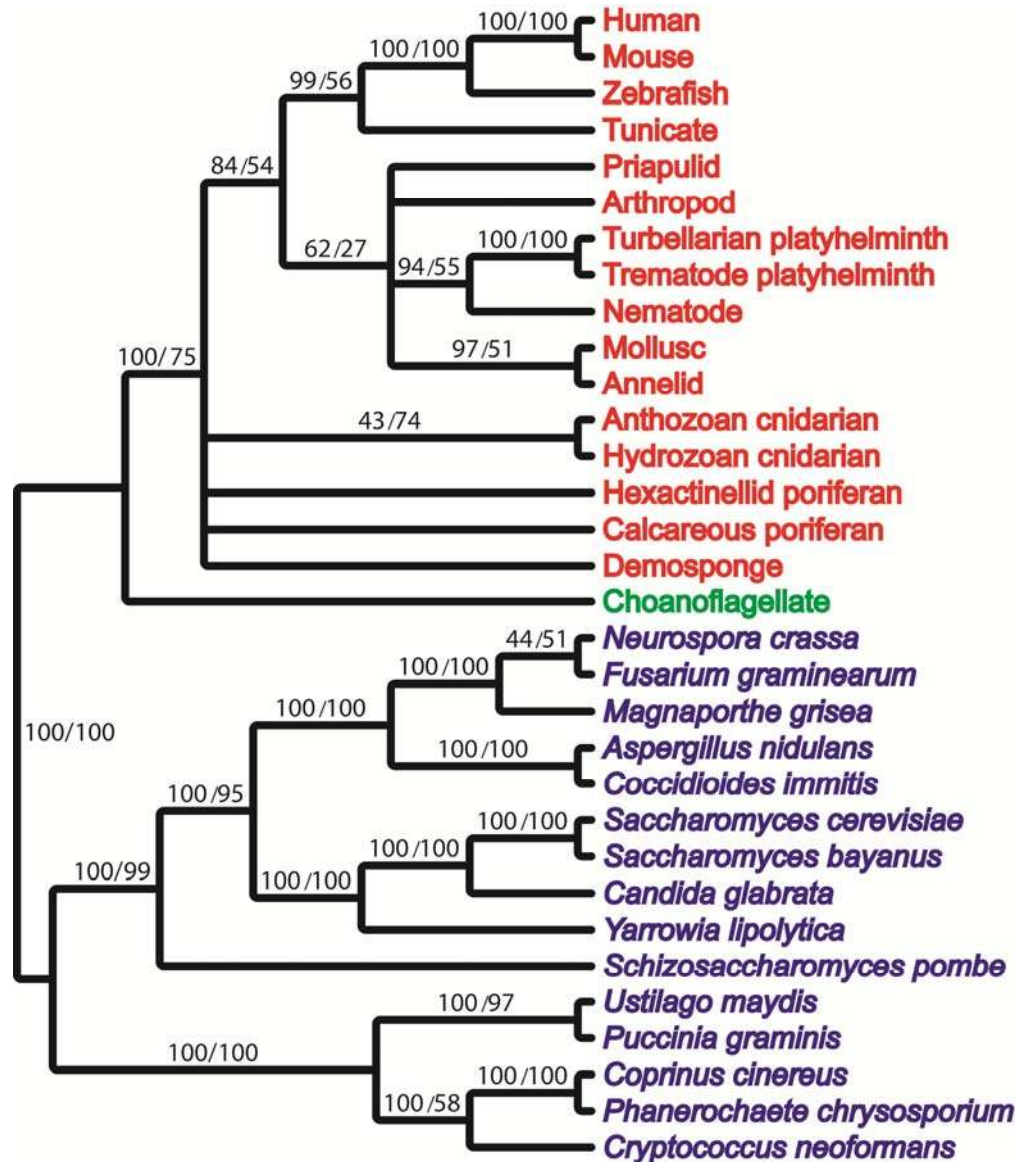
2) Relative molecular clocks

Study	Metazoa	Fungi
<i>Douzery et al. (2004) PNAS</i>	849 Myr	727 Myr
<i>Hedges et al. (2004) BMC Evol. Biol.</i>	976 Myr	968 Myr
<i>Heckman et al. (2001) Science</i>	1177 Myr	1208 Myr

3) The way this set of 50 genes has evolved across the two Kingdoms is remarkably similar



A Remarkable Contrast in Phylogenetic Resolution



Given:

- ❖ lack of resolution in Metazoa (using lots of genes)
- ❖ contrast in resolution between Metazoa and Fungi (using identical genes)

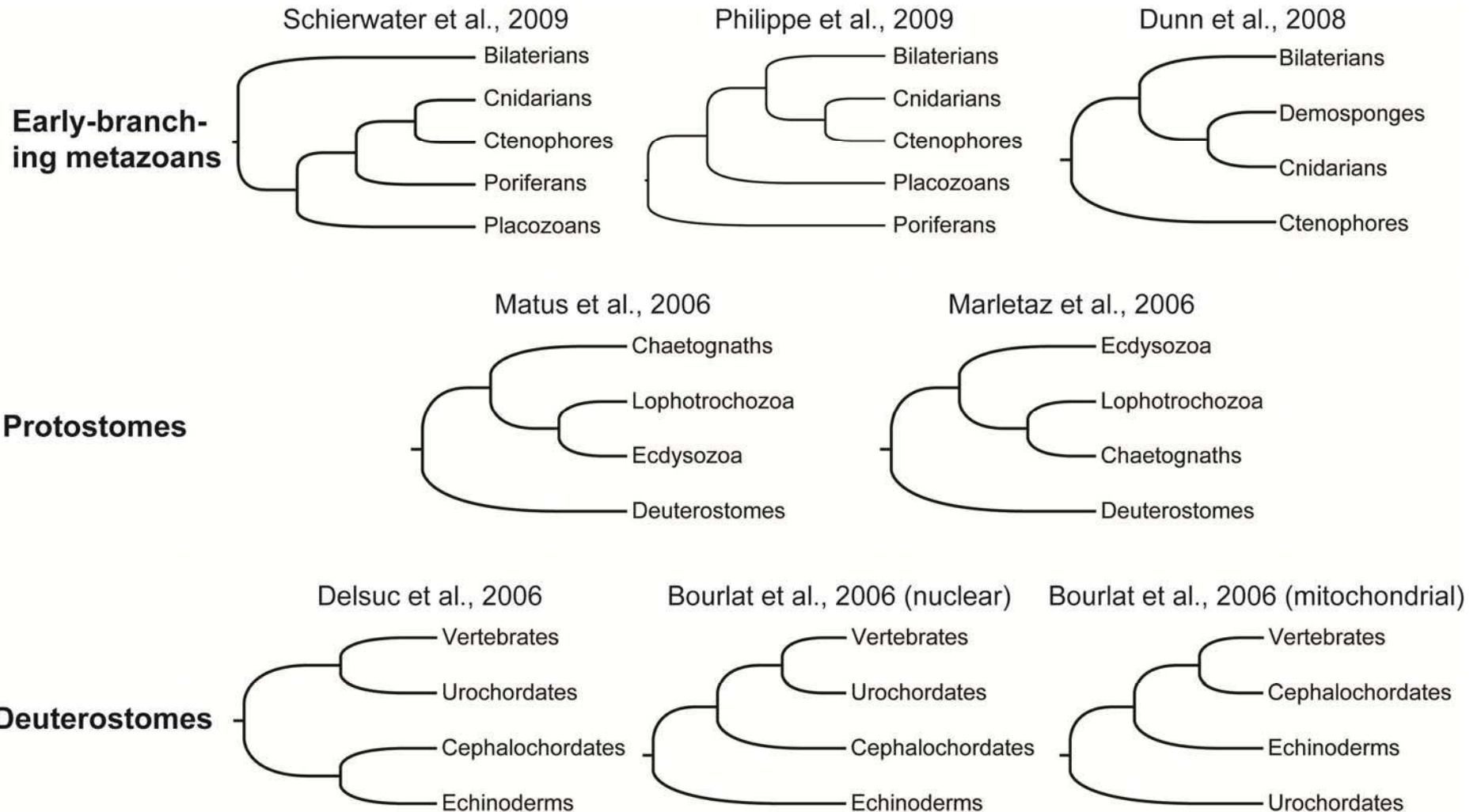
The early history of animals is best viewed as an evolutionary radiation



Rapid Tempo of Cladogenesis Early in Metazoan History

[illegible]

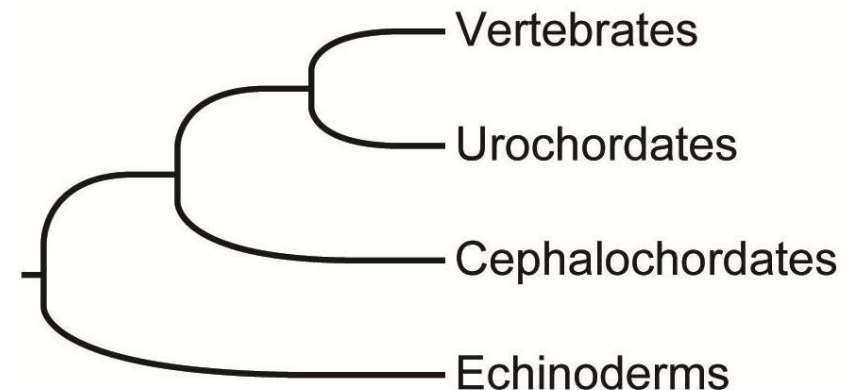
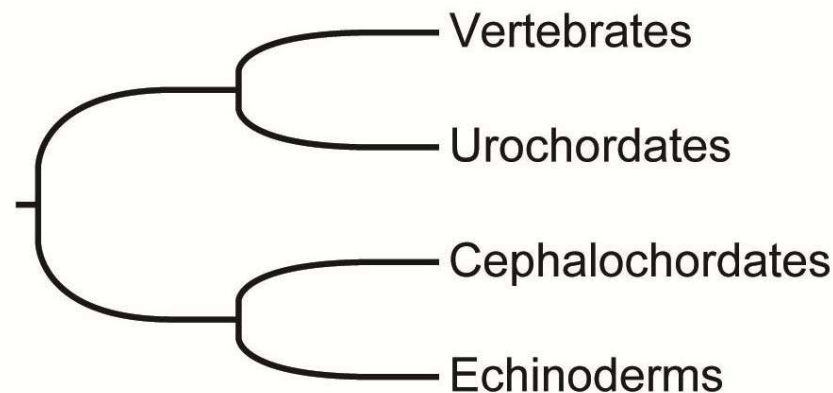
Incongruence in Deep Time



LETTERS

Tunicates and not cephalochordates are the closest living relatives of vertebrates

Frédéric Delsuc^{1†}, Henner Brinkmann¹, Daniel Chourrout² & Hervé Philippe¹

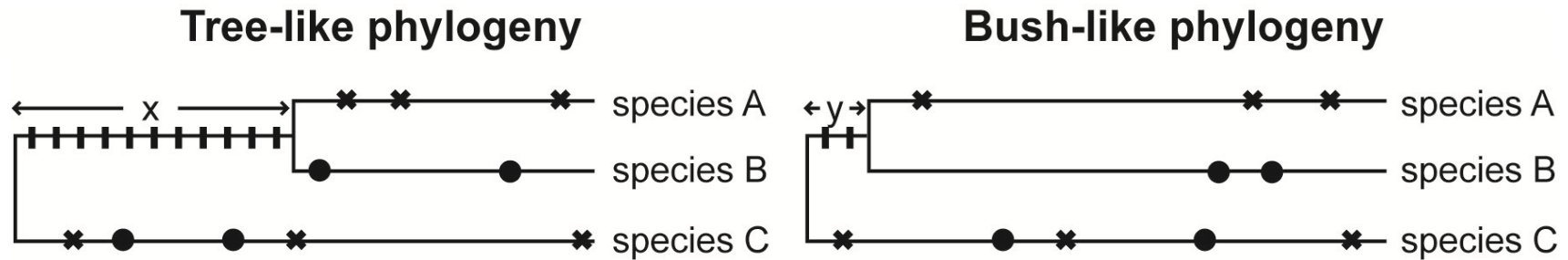


LETTERS

Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida

Sarah J. Boulrat¹, Thorhildur Juliusdottir², Christopher J. Lowe³, Robert Freeman⁴, Jochanan Aronowicz³, Mark Kirschner⁵, Eric S. Lander^{4,6}, Michael Thorndyke⁷, Hiroaki Nakano⁷, Andrea B. Kohn⁸, Andreas Heyland⁸, Leonid L. Moroz⁸, Richard R. Copley² & Maximilian J. Telford¹

What Makes Animals Hard to Resolve But Yeasts Easy?



Internode length: influences amount of phylogenetic signal (I)

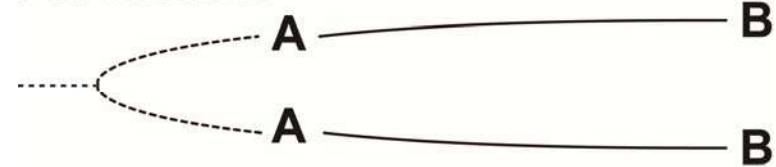
Homoplasy: independent evolution of identical characters (*, •)



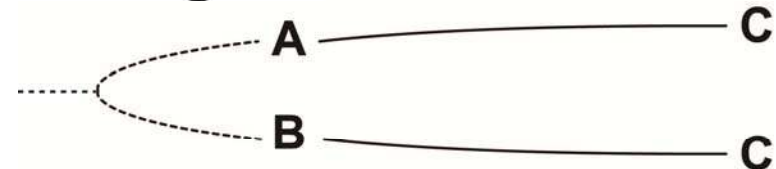
Homoplasy

the independent evolution of identical character states (e.g., amino acid residues) in different branches of a phylogenetic tree that are not directly inherited from a common ancestor

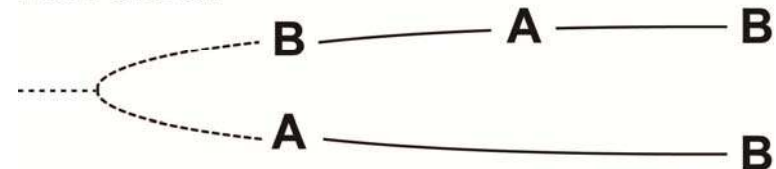
Parallelism



Convergence

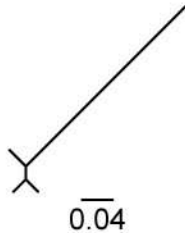


Reversal



Quantifying Homoplasy Across the Tree of Life

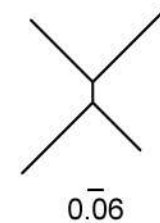
***Saccharomyces* yeasts
(106)**



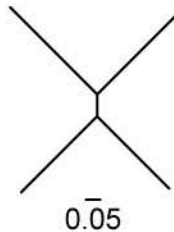
***Aspergillus* ascomycetes
(200)**



Fungal phyla (200)



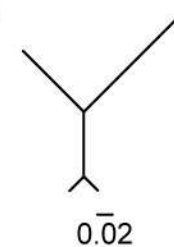
Eukaryotic phyla (200)



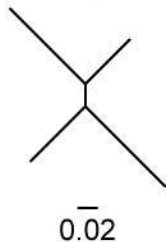
Land plants (39)



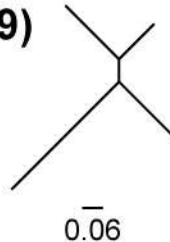
***Drosophila* fruit-flies
(200)**



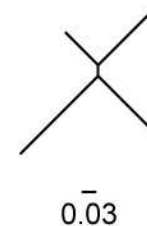
**Cetartiodactyl mammals
(12)**



**Metazoan phyla
(239)**



**Paenungulata mammals
(12)**



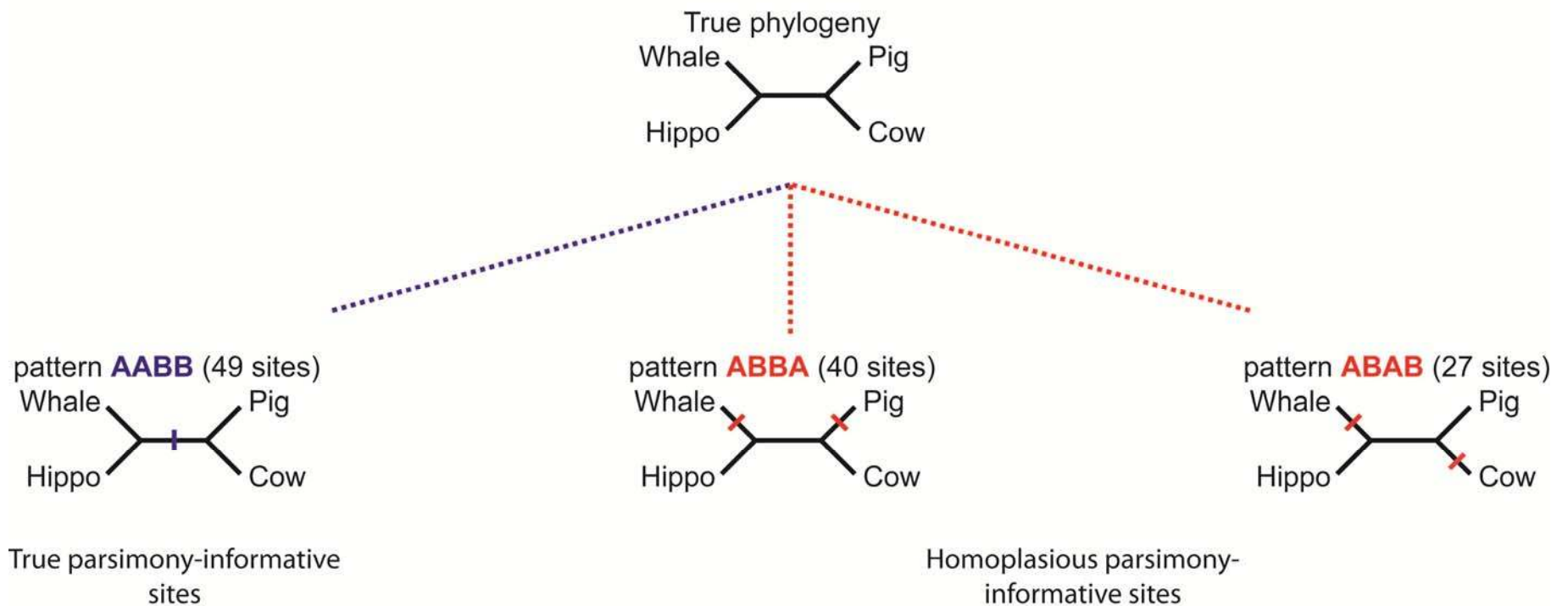
Vertebrates (44)



Measuring Homoplasy Across the Tree of Life

	1	2	3	4	5	6	7	8
Whale	1	0	0	0	0	0	0	0
Hippo	1	0	0	0	0	0	0	0
Cow	1	0	0	0	0	0	0	0
Pig	1	0	0	0	0	0	0	0

Whale	TGFRNKT	KMSLAHLLPQGTPTFLIPMLVIIETISLFIQ	V	ALAVRLTANITAGHLL	M	HLIGETTLVLM	S	TSLFTAIITFT...
Hippo	MGFRNKT	KISLAHFLPQGTPTPLIPMLVIIETISLFIQ	M	ALAVRLTANITAGHLL	M	HLIGGATLALM	N	ISMTTALITFI...
Cow	TGFRNKT	KASLAHFLPQGTPTPLIPMLVIIETISLFIQ	M	ALAVRLTANITAGHLL	I	HLIGGATLALM	S	ISTTTALITFT...
Pig	TGFRYKT	KTSLAHFLPQGTALLIPMLVIIETISLFIQ	V	ALAVRLTANITAGHLL	I	HLIGGATLALL	N	INTMTAFITFT...



Measuring Homoplasy Across the Tree of Life

Pattern	Observed sites	Observed substitutions (observed sites x substitutions)	Expected sites	Expected substitutions (expected sites x substitutions)
ABBB	460	460 x 1 = 460	456	456 x 1 = 456
BABB				
BBAB				
BBBA				
AABB	49	49 x 1 = 49	37	37 x 1 = 37
ABAB	67	67 x 2 = 134	25.5	25.5 x 2 = 51
ABBA				
AABC	171	171 x 2 = 342	185.5	185.5 x 2 = 371
ABAC				
ABCA				
BAAC				
BACA				
BCAA	34	34 x 3 = 102	40	40 x 3 = 120
ABCD				

$$\text{Observed Homoplasy} = \frac{\text{ABAB+ ABBA observed substitutions}}{\text{Total number of observed substitutions}} = \frac{134}{1087} = 12.3\%$$

$$\text{Expected Homoplasy} = \frac{\text{ABAB+ABBA expected substitutions}}{\text{Total number of expected substitutions}} = \frac{51}{1035} = 5.0\%$$



An Overabundance of Homoplastic Substitutions

Dataset	Obs(H)	Exp(H)	Obs(H)/Exp(H)
<i>Saccharomyces</i> yeasts	4.7%	2.0%	2.4
<i>Aspergillus</i> filamentous fungi	6.9%	2.5%	2.8
Fungal phyla	6.7%	3.5%	1.9
Eukaryote phyla	7.5%	3.5%	2.1
Plants	2.3%	0.9%	2.7
<i>Drosophila</i> fruit-flies	2.6%	1.3%	2.0
Cetartiodactyls mammals	12.3%	4.9%	2.5
Paenungulata mammals	10.4%	5.0%	2.1
Metazoan phyla	7.4%	3.3%	2.3
Vertebrates	10.2%	3.2%	3.2
Average:	7.1%	2.3%	2.4



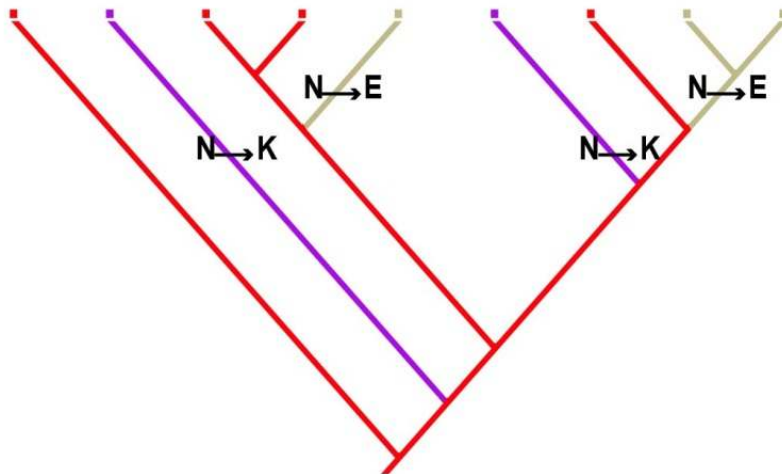
Excess Homoplasy is Specific to Homoplastic Substitutions

Clade	Parsimony-informative sites		
	Obs(H)	Exp(H)	Obs(H)/Exp(H)
<i>Saccharomyces</i> yeasts	3.3%	3.3%	0.0
<i>Aspergillus</i> filamentous ascomycetes	10.3%	9.8%	1.1
Fungal phyla	2.5%	2.2%	1.1
Eukaryotic phyla	2.3%	2.1%	1.1
Land plants	31.6%	31.7%	1.0
<i>Drosophila</i> fruit-flies	10.2%	10.1%	1.0
Cetartiodactyl mammals	4.5%	3.6%	1.3
Metazoan phyla	3.1%	2.8%	1.1
*Paenungulata mammals	5.4%	2.4%	2.3
*Vertebrates	2.9%	2.4%	1.2
Average:	7.6%	7.0%	1.1



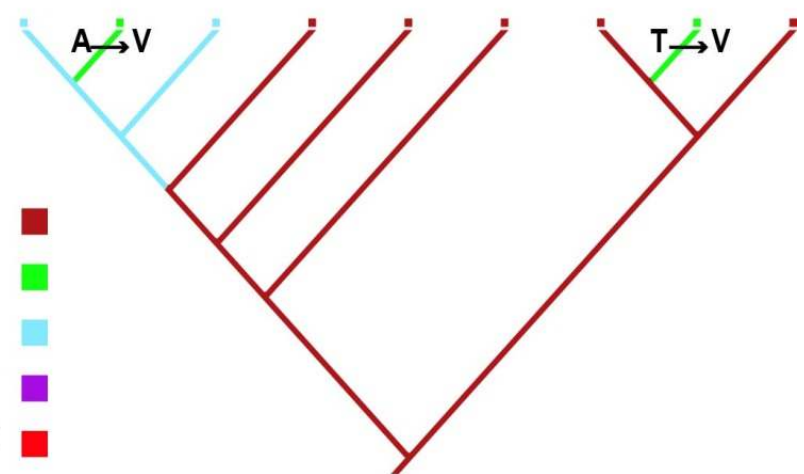
Homoplasy Stems From Frequently Exchanged Amino Acids

Aspergillus filamentous ascomycetes



T (Tyr) ■
V (Val) ■
A (Ala) ■
K (Lys) ■
N (Asn) ■
E (Glu) ■

Cetartiodactyl mammals



190 possible interchanges among 20 amino acids

- 75 can be achieved via a single nucleotide substitution
- The other 115 require two or three substitutions

65% of observed interchanges is between the top12 most frequently observed amino acid interchanges a single mutational step away

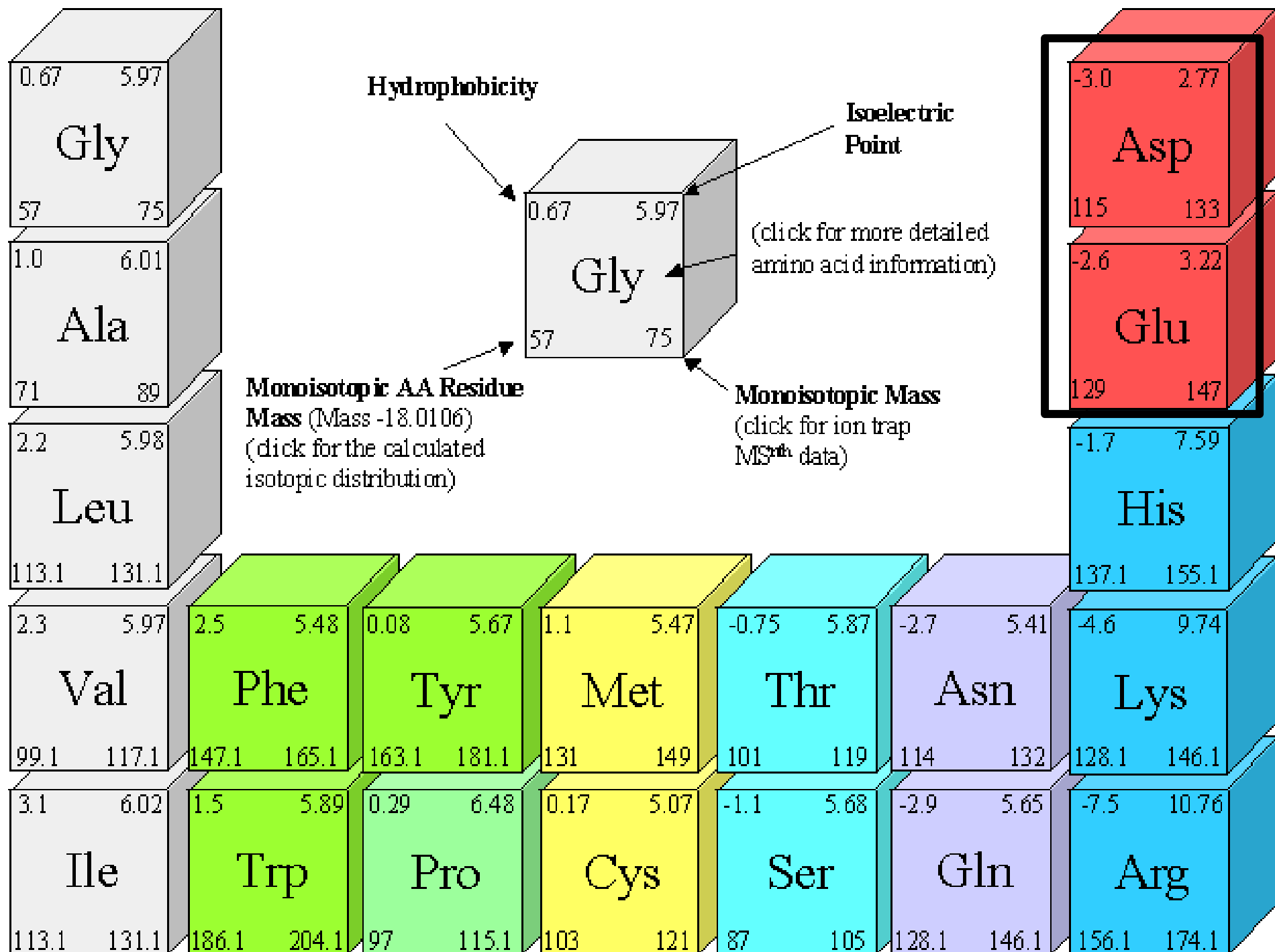


Conservative AA Substitutions Are Very Common in Alignments

>Cimi	EGAAGPIRFNMRLPESRYHCAGGITTIVVY
>Acla	AATTSPEDVETRKDDERVEHGGGITTVLVY
>Afis	AATASAEFEELRKEDERVEHGDGITTILY
>Afla	SPAAAADFEFDLQRDEDQNDYADSVSSVFIF
>Afum	AATTSAEEFEELRKEDERVEHGDGITTILY
>Anid	APAATADVSDTRLKQEQAEQAGGITAILVY
>Anig	APAAAPDDLETRLEEEQADYAGSVSSILY
>Aory	SPAAAADFEFDLQRDEDQNDYADSVSSVFIF
>Ater	SPAAAPDDVDVTQREEDQNDYAGSVSSIIFV

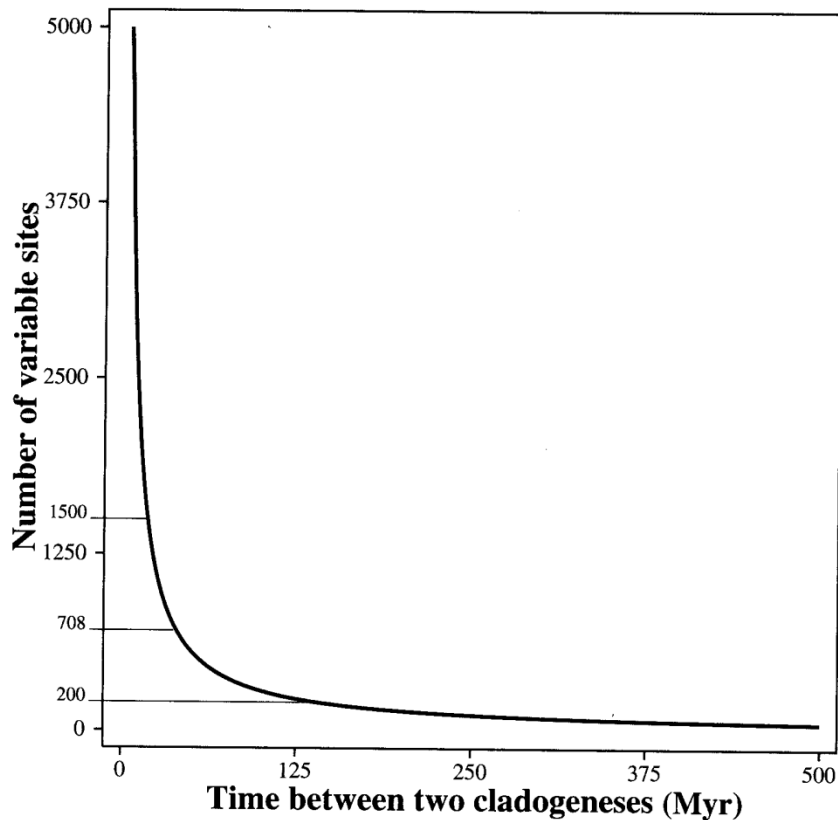
GAC/T (D)
Aspartic Acid
↕
GAA/G (E)
Glutamic Acid





What Are the Limits to Resolution?

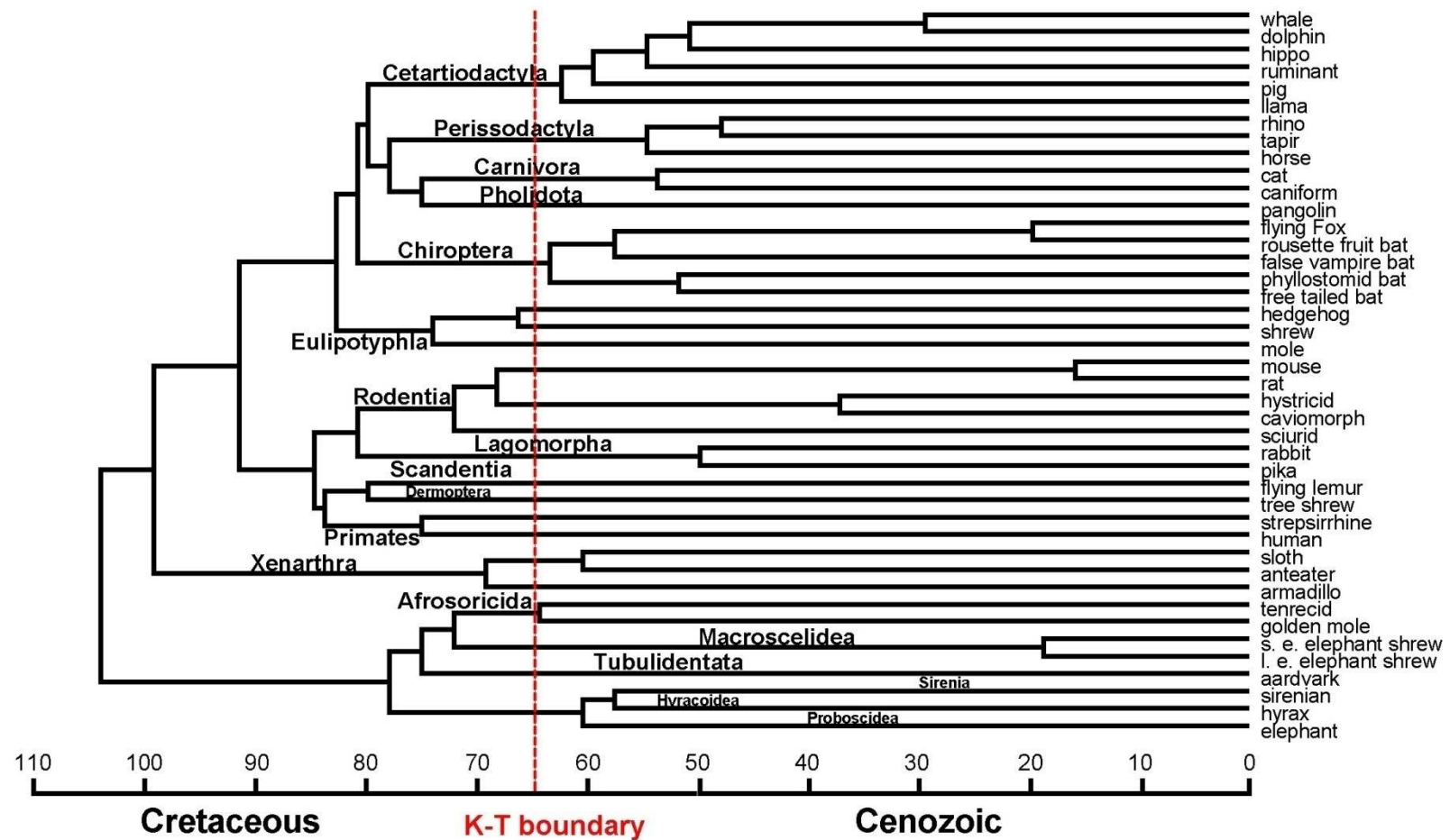
For internodes in ~ 600 my animal phylogeny



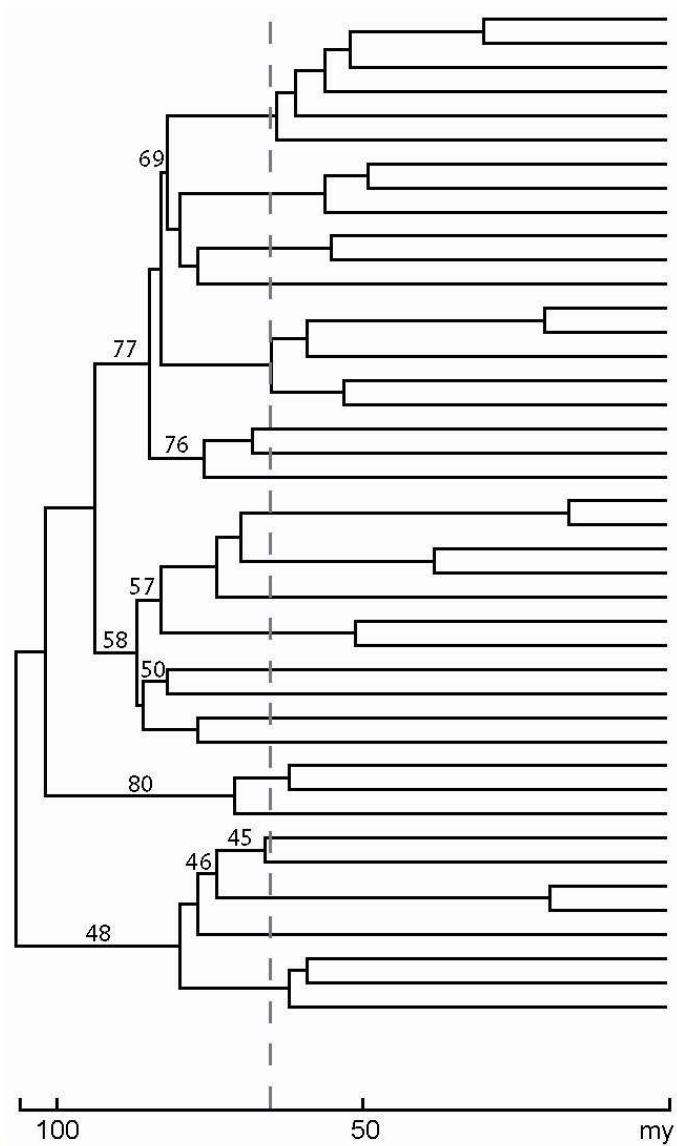
Time	Amount of Data
40 my ->	700 variable sites
10 my ->	2,800 variable sites
1 my ->	28,000 variable sites
0.1 my ->	?????? variable sites
1 y ->	?????? variable sites



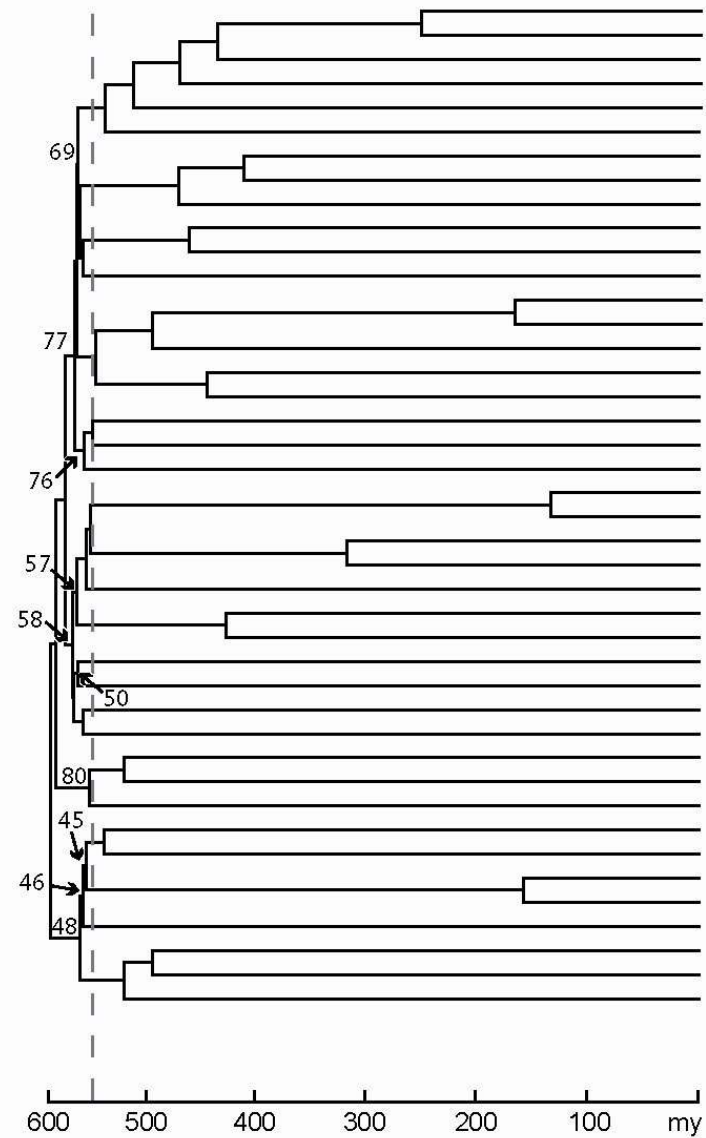
What If Mammals Had Diversified in the Cambrian?



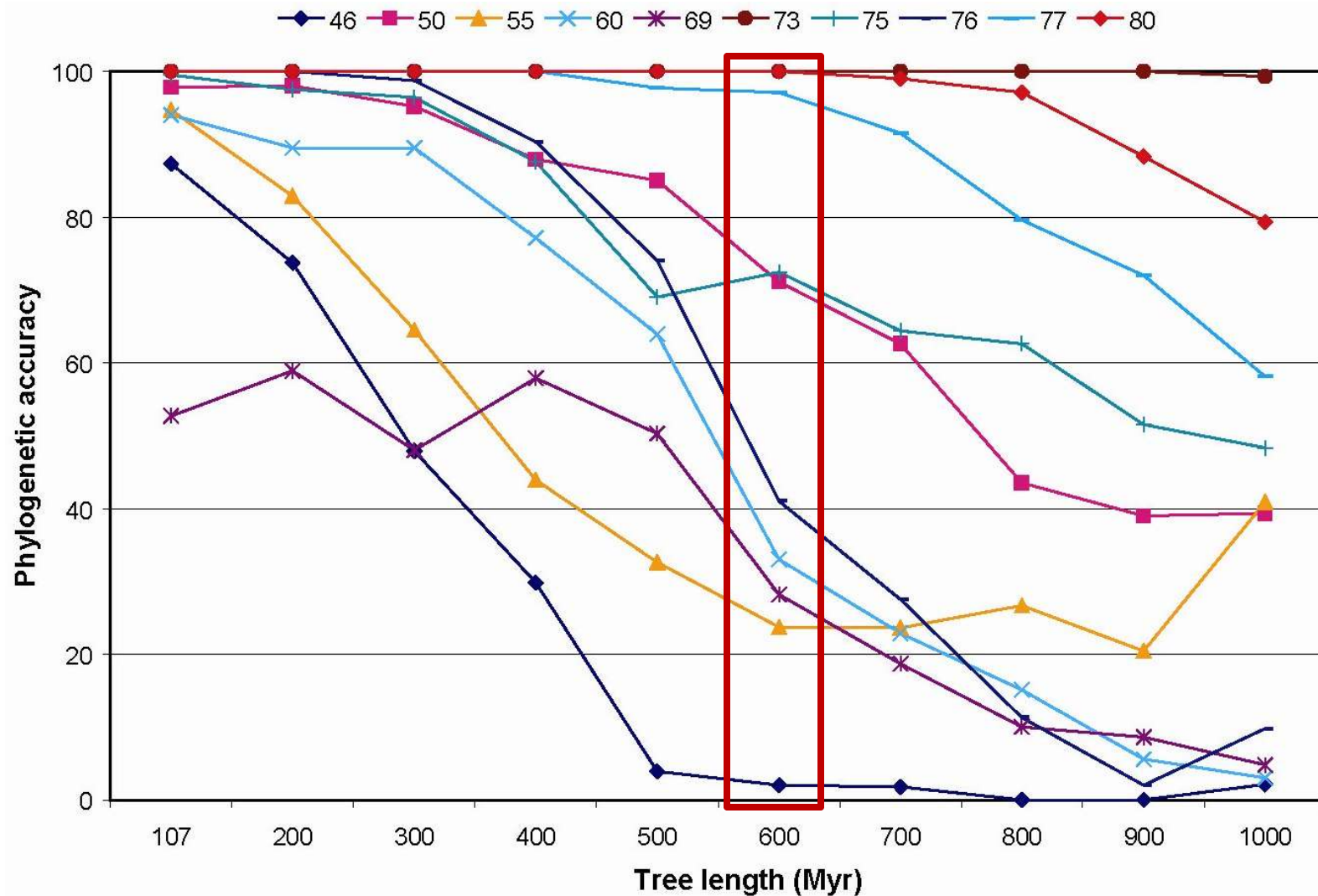
107 Myr tree



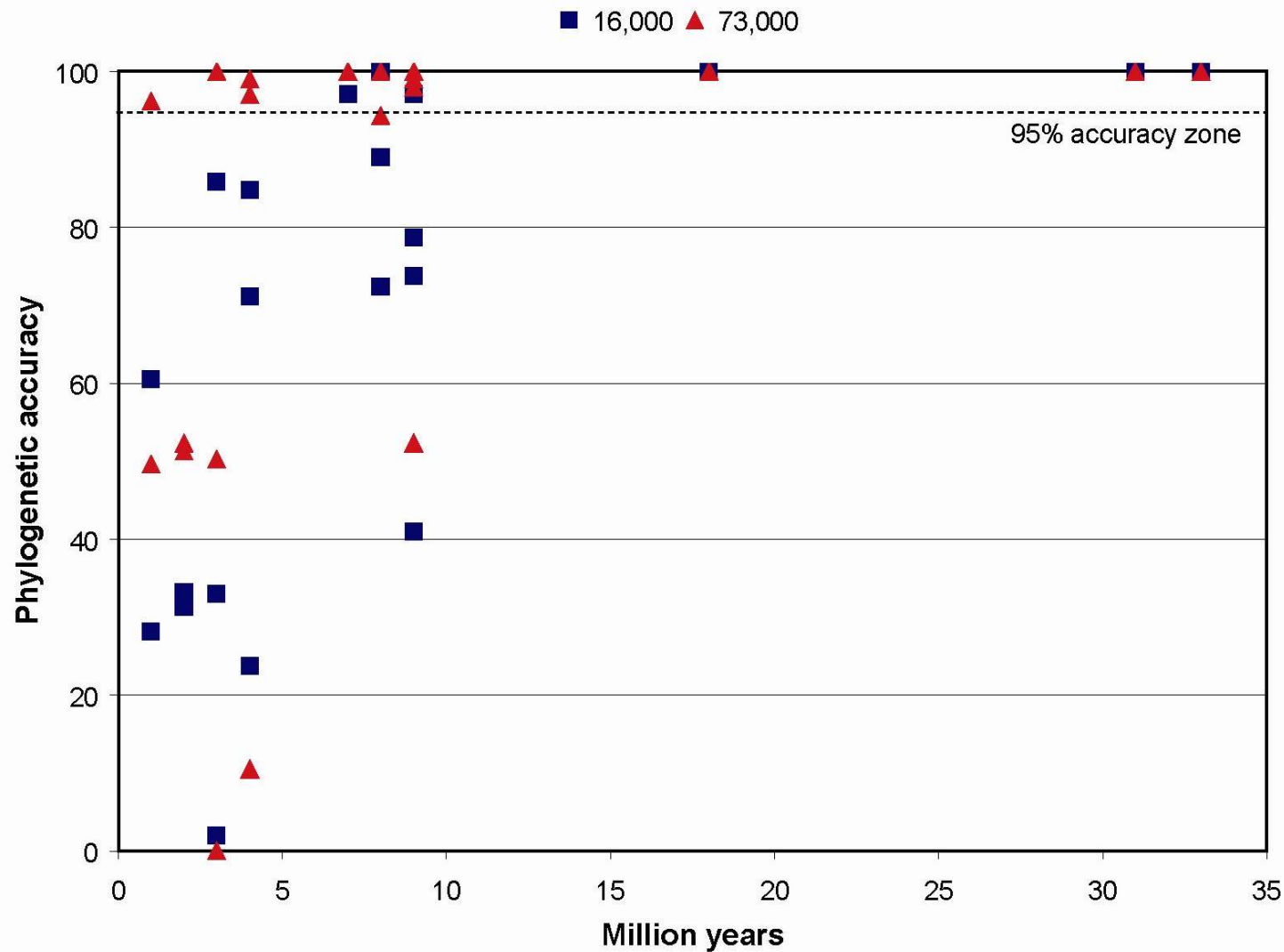
600 Myr tree



Phylogenetic Accuracy is Inversely Correlated with Elapsed Time

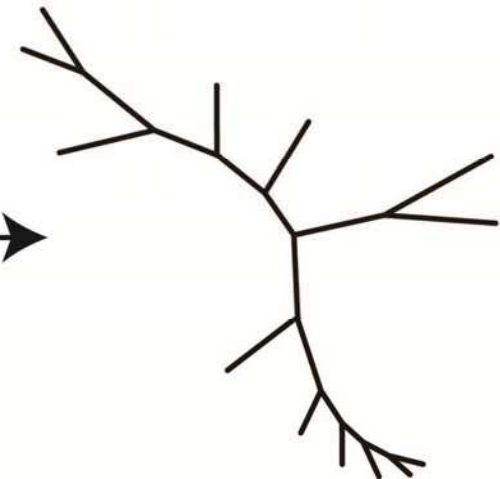
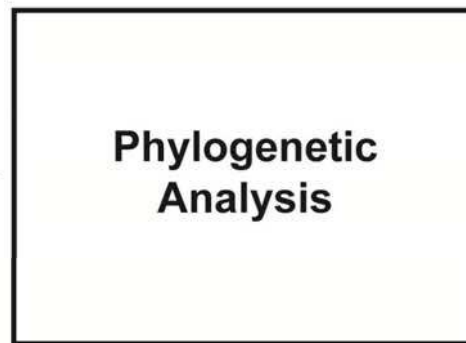


Failure to Resolve Internodes < 10 Million Years



Why Are There So Few Bushes?

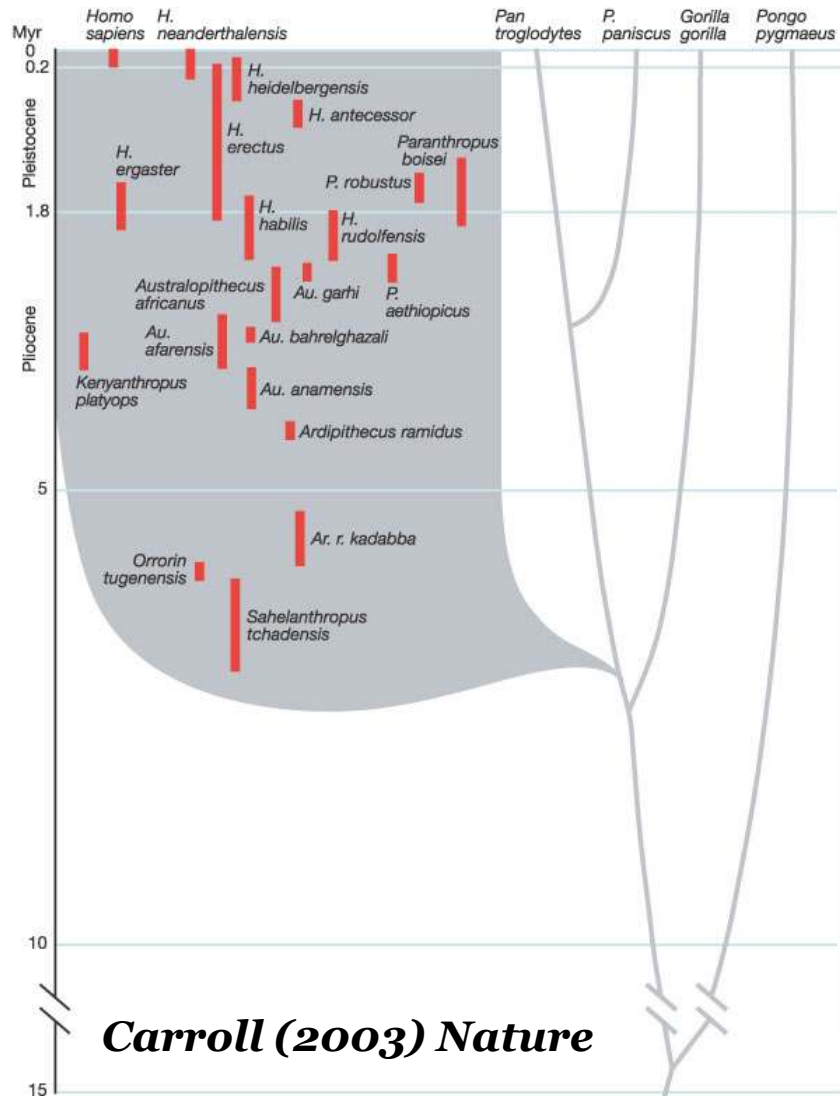
taxon_1	ACCCGATAGACAA
taxon_2	.G.G.....
taxon_3	CT....
taxon_4	A.....C
taxon_5	T.A.....
taxon_7	TT....
taxon_8	..G....TT....
taxon_9	G.....
taxon_10	T.....
taxon_11	T.....
taxon_12	..GG.....T..
taxon_13	..GG...C..T..



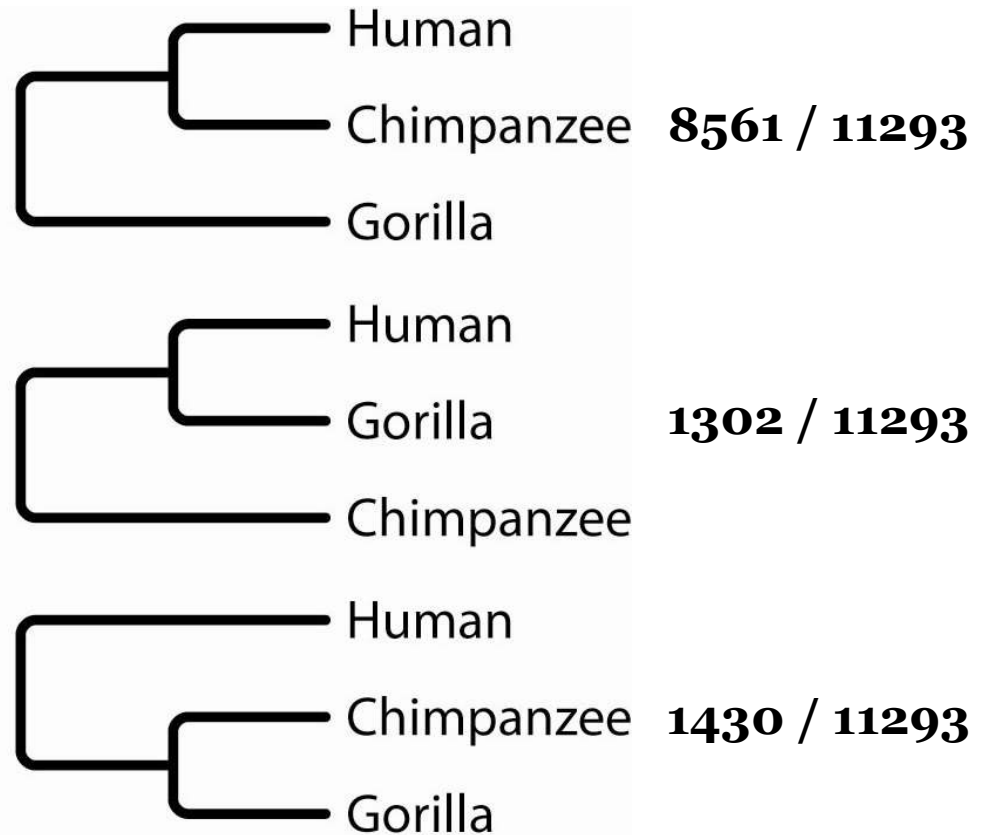
Data in, fully-resolved phylogenetic tree out



The Human / Chimp / Gorilla Tree

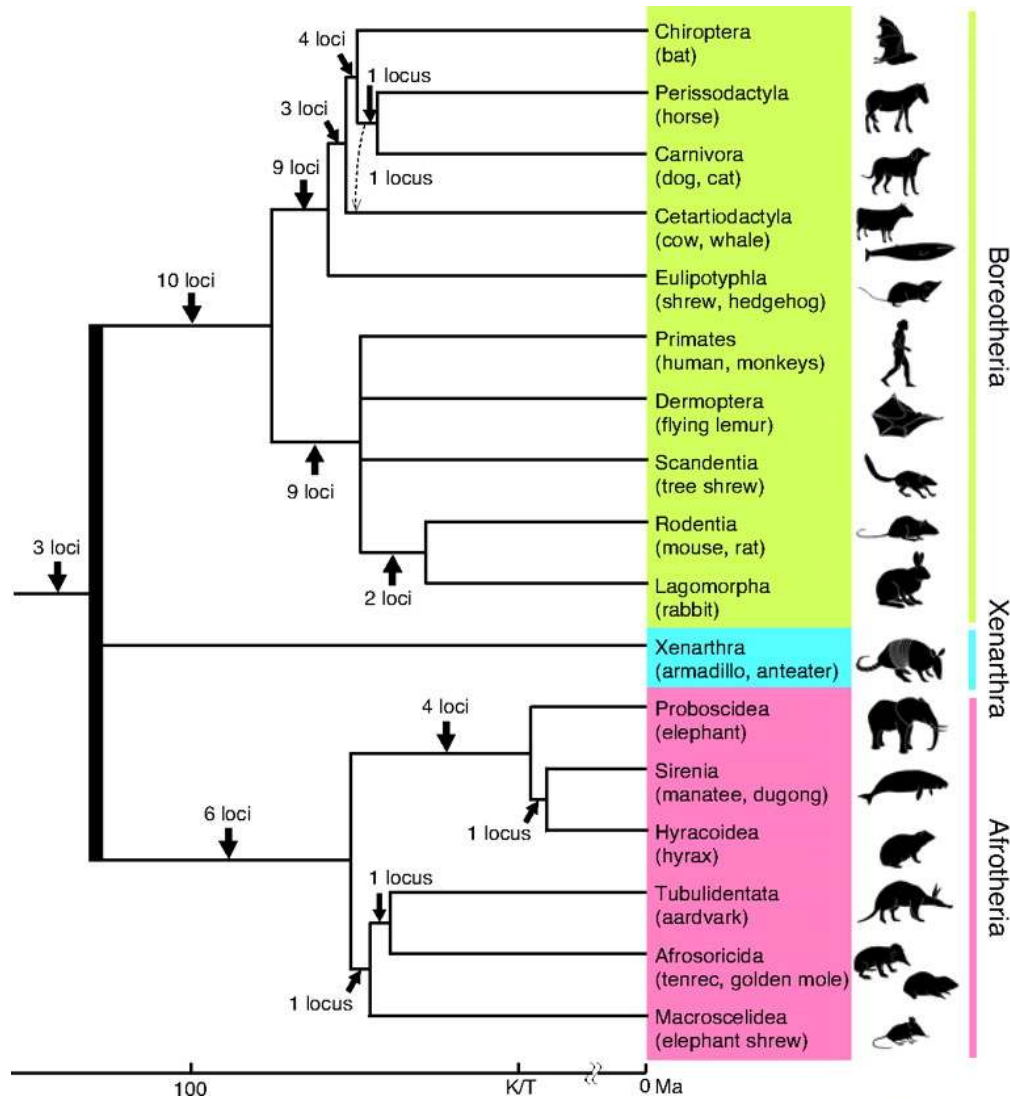


Inform. Sites

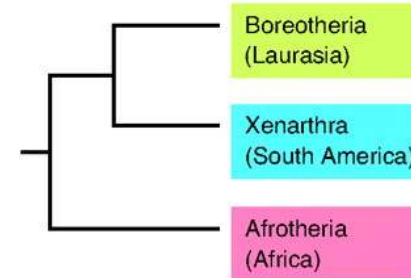


Patterson et al. (2006) Nature

Bushes in the Mammalian Tree



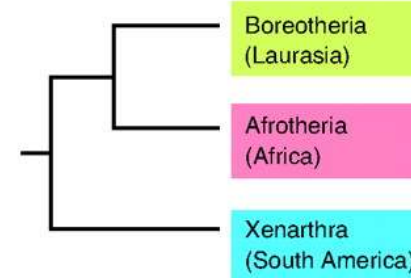
Tree 1 (basal-Afrotheria)



Inform. Sites

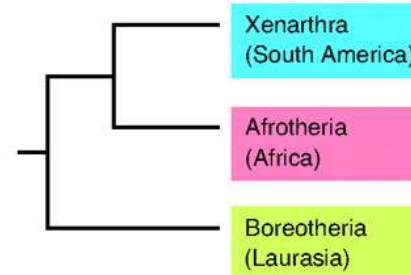
22

Tree 2 (basal-Xenarthra)



25

Tree 3 (basal-Boreotheria)



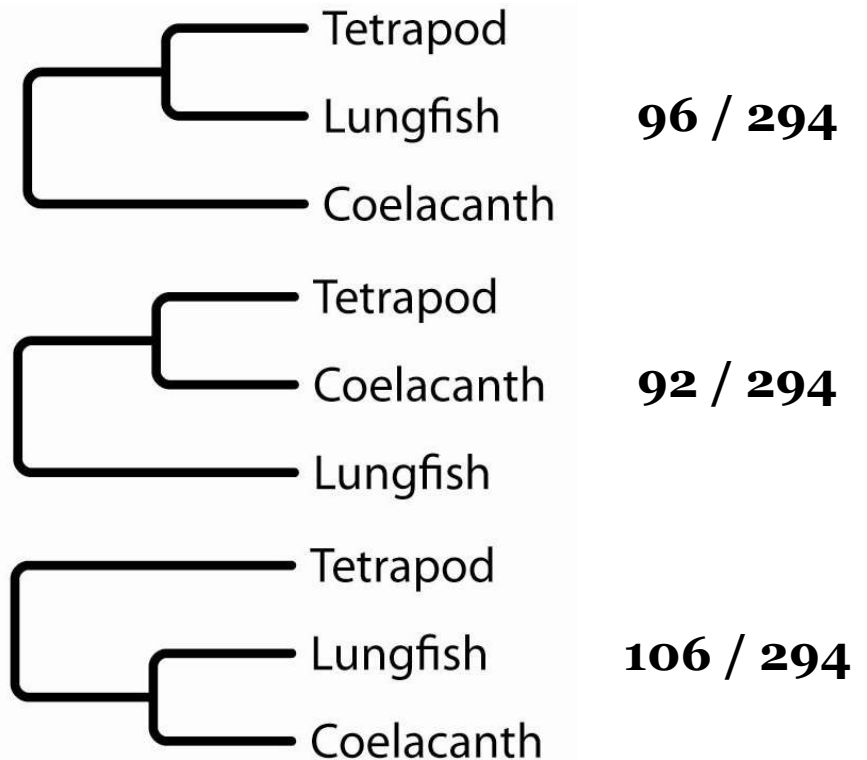
21



The Tetrapod / Lungfish / Coelacanth Bush

Inform. Sites

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



44 genes, ML/MP/NJ



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

Mind the Gap Between Real Data and Models



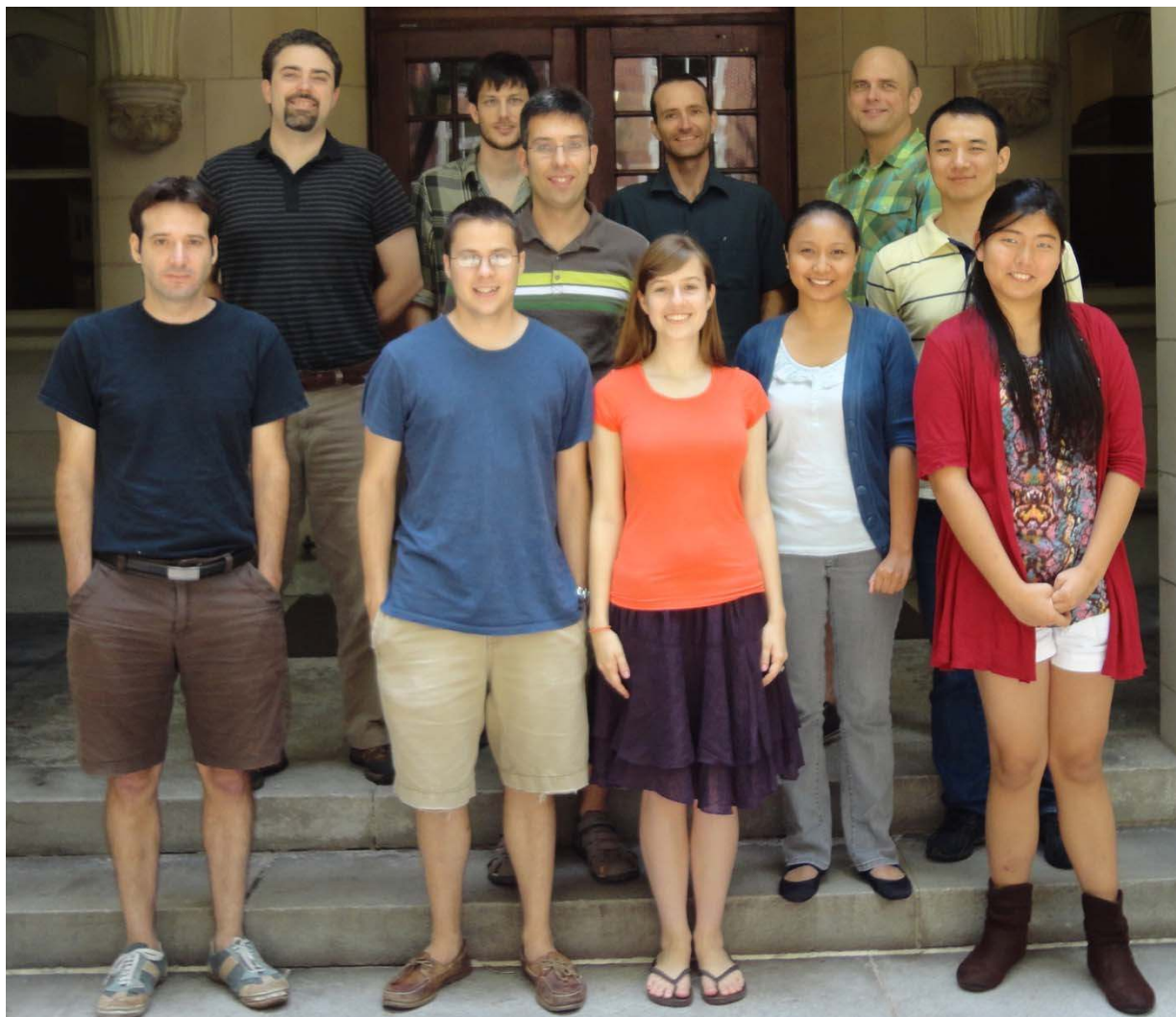
MIND THE GAP

“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



 **SEARLE SCHOLARS PROGRAM**



National Science Foundation
WHERE DISCOVERIES BEGIN



<http://as.vanderbilt.edu/rokaslab>