

# Phylogenetics

# What is phylogenetics?

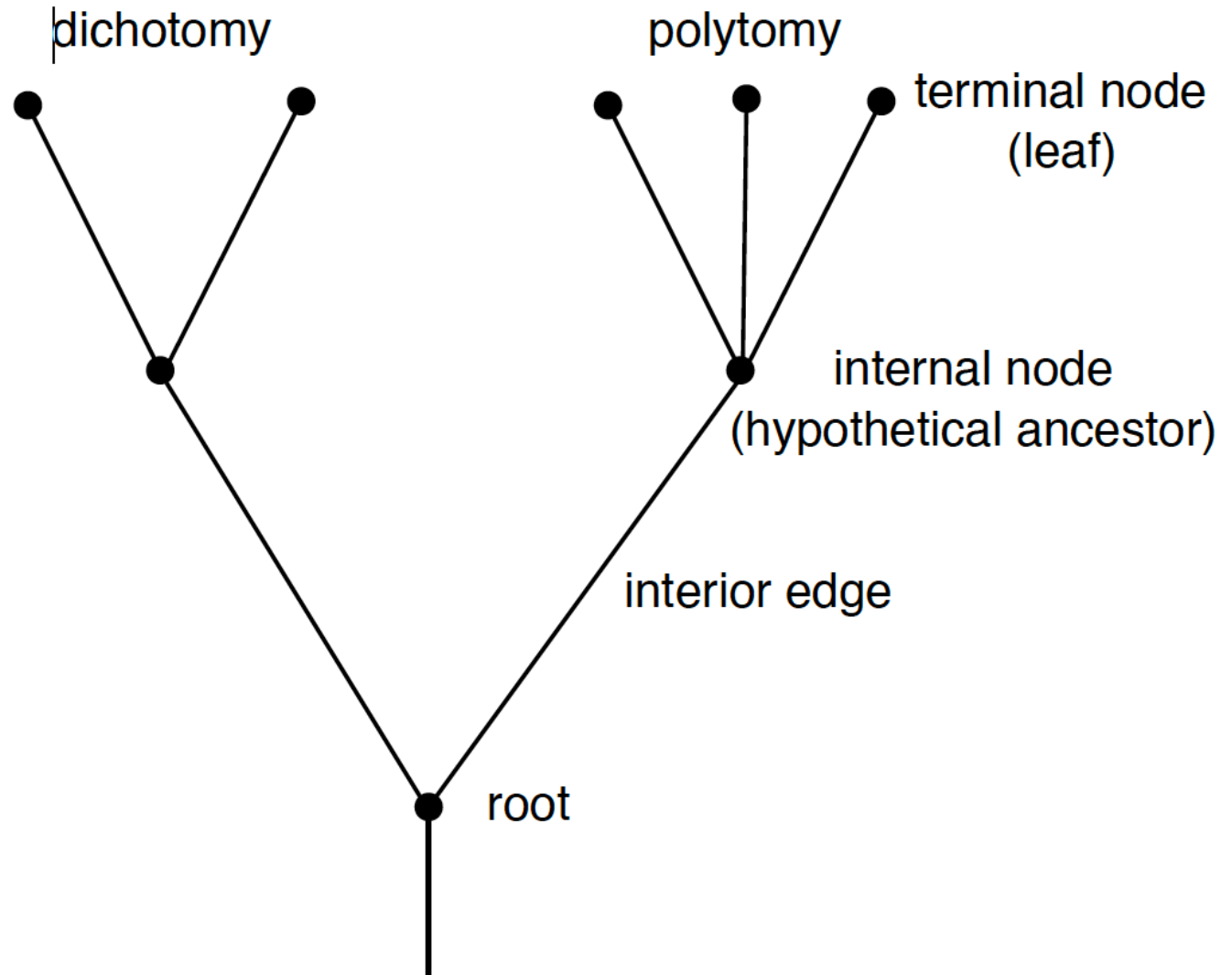
- Study of branching patterns of descent among lineages
- Lineages
  - Populations
  - Species
  - Molecules
- Shift between population genetics and phylogenetics is often the species boundary
  - Distantly related populations also show patterning
  - Patterning across geography

# What is phylogenetics?

- Goal: Determine and describe the evolutionary relationships among lineages
  - Order of events
  - Timing of events
- Visualization: Phylogenetic trees
  - Graph
  - No cycles

# Phylogenetic trees

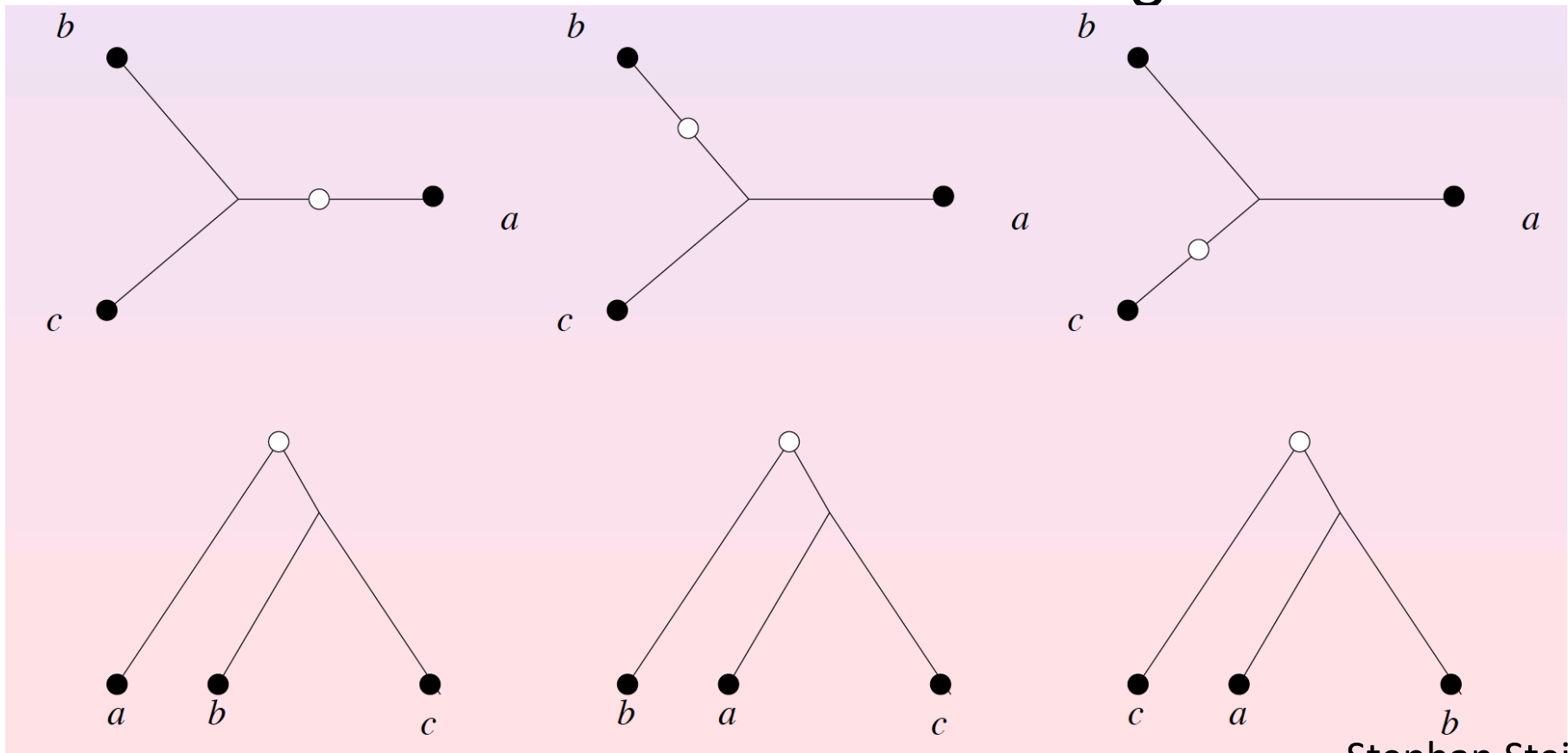
- Nodes
  - Terminal
  - Internal
  - Degree
- Branches
- Topology





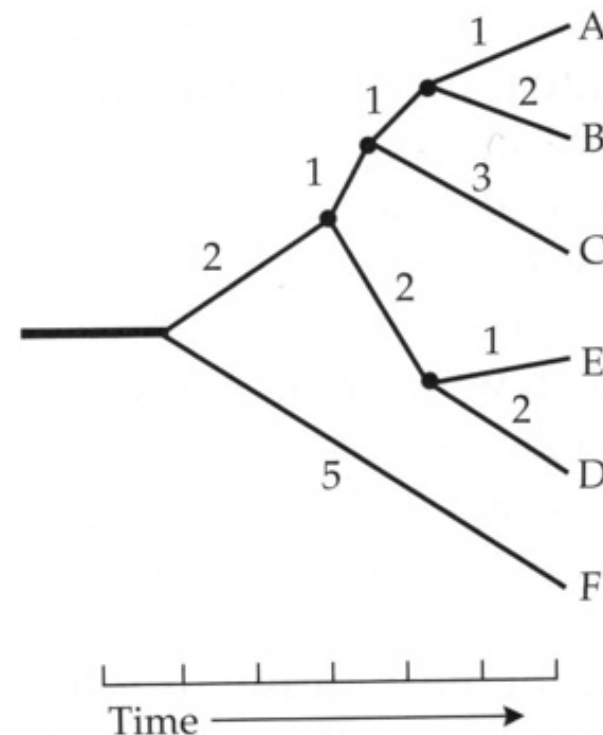
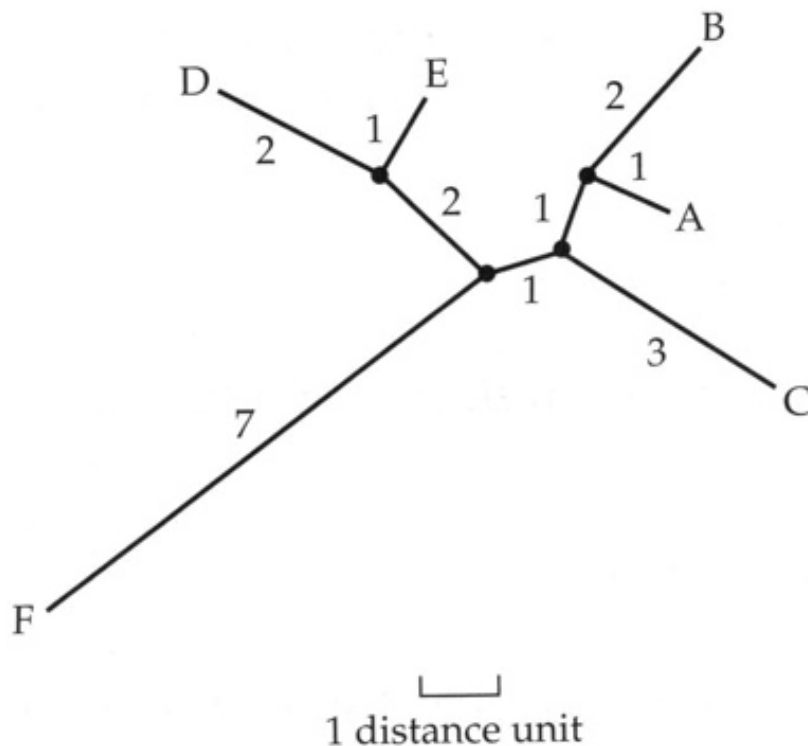
# Phylogenetic trees

- Rooted or unrooted
  - Rooted: Precisely 1 internal node of degree 2
    - Node that represents the common ancestor of all taxa
  - Unrooted: All internal nodes with degree 3+



# Phylogenetic trees

- Rooted or unrooted
  - Rooted: Precisely 1 internal node of degree 2
    - Node that represents the common ancestor of all taxa
  - Unrooted: All internal nodes with degree 3+



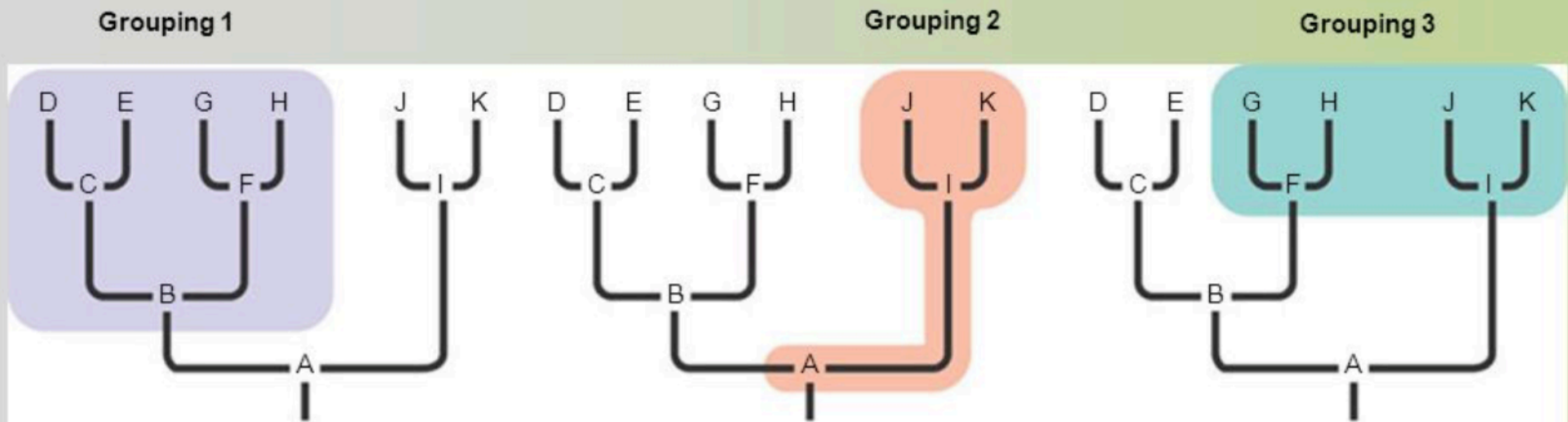
# Phylogenetic trees

- Rooted or unrooted
  - Rooted: Precisely 1 internal node of degree 2
    - Node that represents the common ancestor of all taxa
  - Unrooted: All internal nodes with degree 3+
- Binary: all speciation events produce two lineages from one
- Cladogram: Topology only
- Phylogram: Topology with edge lengths representing time or distance
- Ultrametric: Rooted tree with time-based edge lengths (all leaves equidistant from root)

# Phylogenetic trees

- Clade: Group of ancestral and descendant lineages
- Monophyly: All of the descendants of a unique common ancestor
- Polyphyly: Descendants include lineages from multiple ancestors
- Paraphyly: One or more monophyletic subgroups are left apart from all other descendants of a unique common ancestor

# Phylogenetic trees



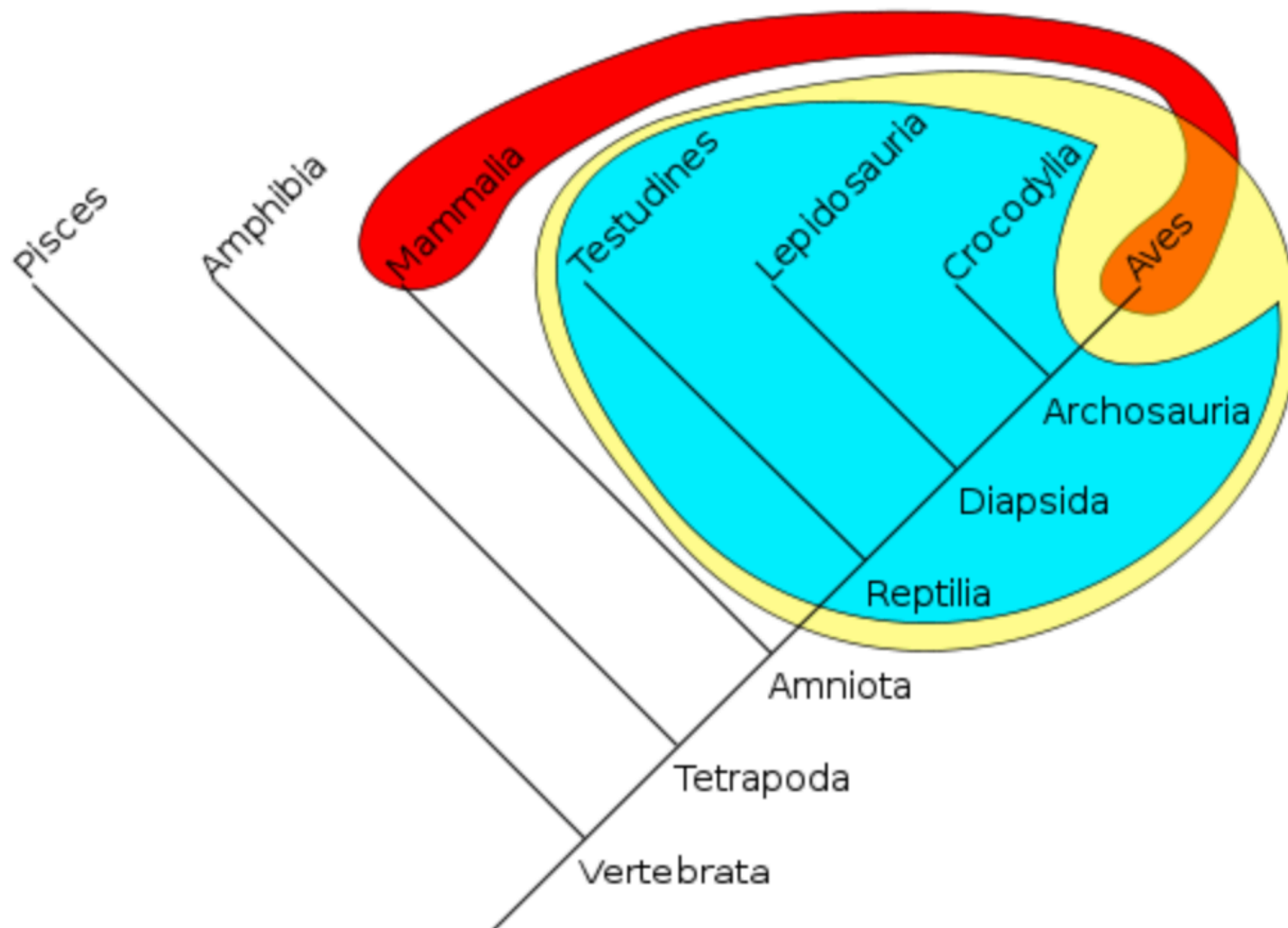
**(a) Monophyletic.** In this tree, grouping 1, consisting of the seven species B–H, is a monophyletic group, or clade. A monophyletic group is made up of an ancestral species (species B in this case) and *all* of its descendant species. Only monophyletic groups qualify as legitimate taxa derived from cladistics.

**(b) Paraphyletic.** Grouping 2 does not meet the cladistic criterion: It is paraphyletic, which means that it consists of an ancestor (A in this case) and *some*, but not all, of that ancestor's descendants. (Grouping 2 includes the descendants I, J, and K, but excludes B–H, which also descended from A.)

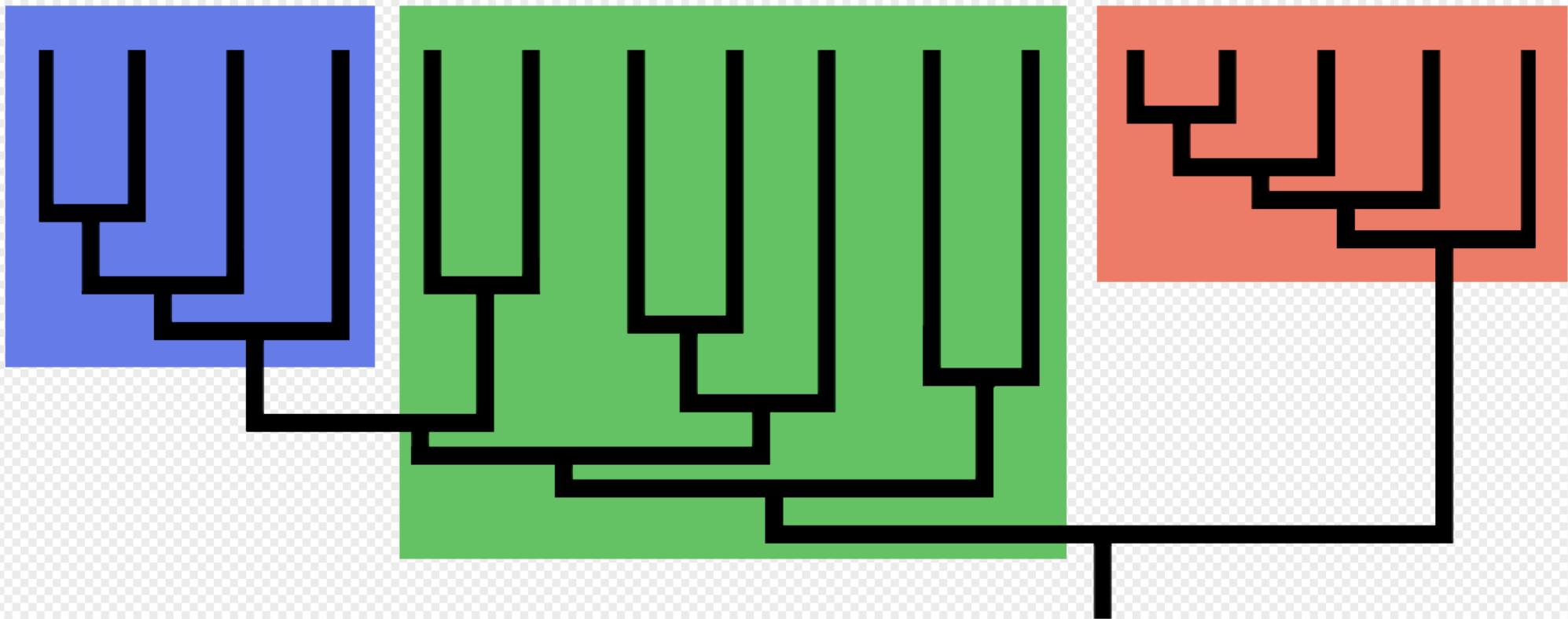
**(c) Polyphyletic.** Grouping 3 also fails the cladistic test. It is polyphyletic, which means that it lacks the common ancestor of (A) the species in the group. Furthermore, a valid taxon that includes the extant species G, H, J, and K would necessarily also contain D and E, which are also descended from A.

# Phylogenetic trees

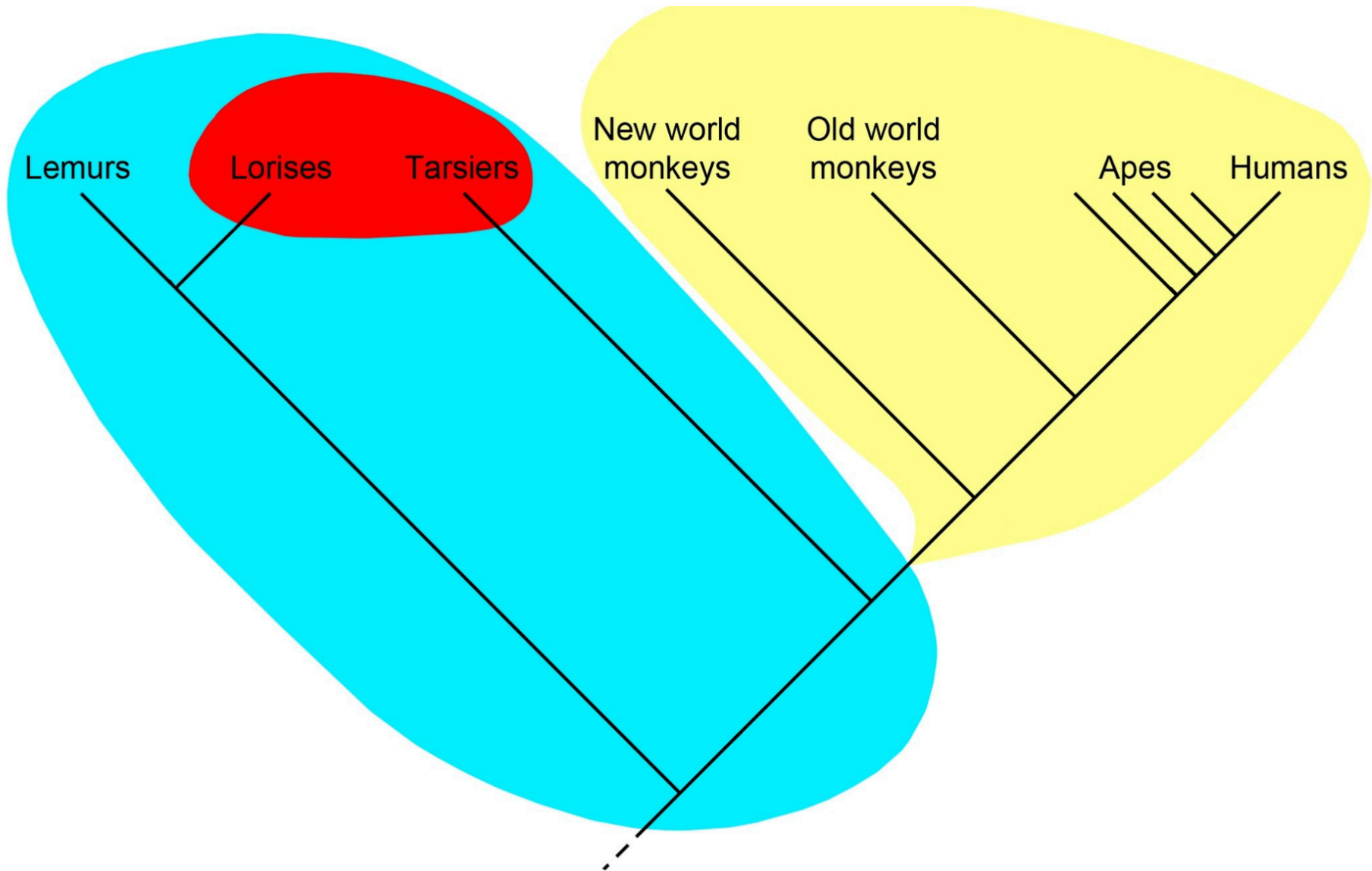
- Monophyly
- Paraphyly
- Polyphyly



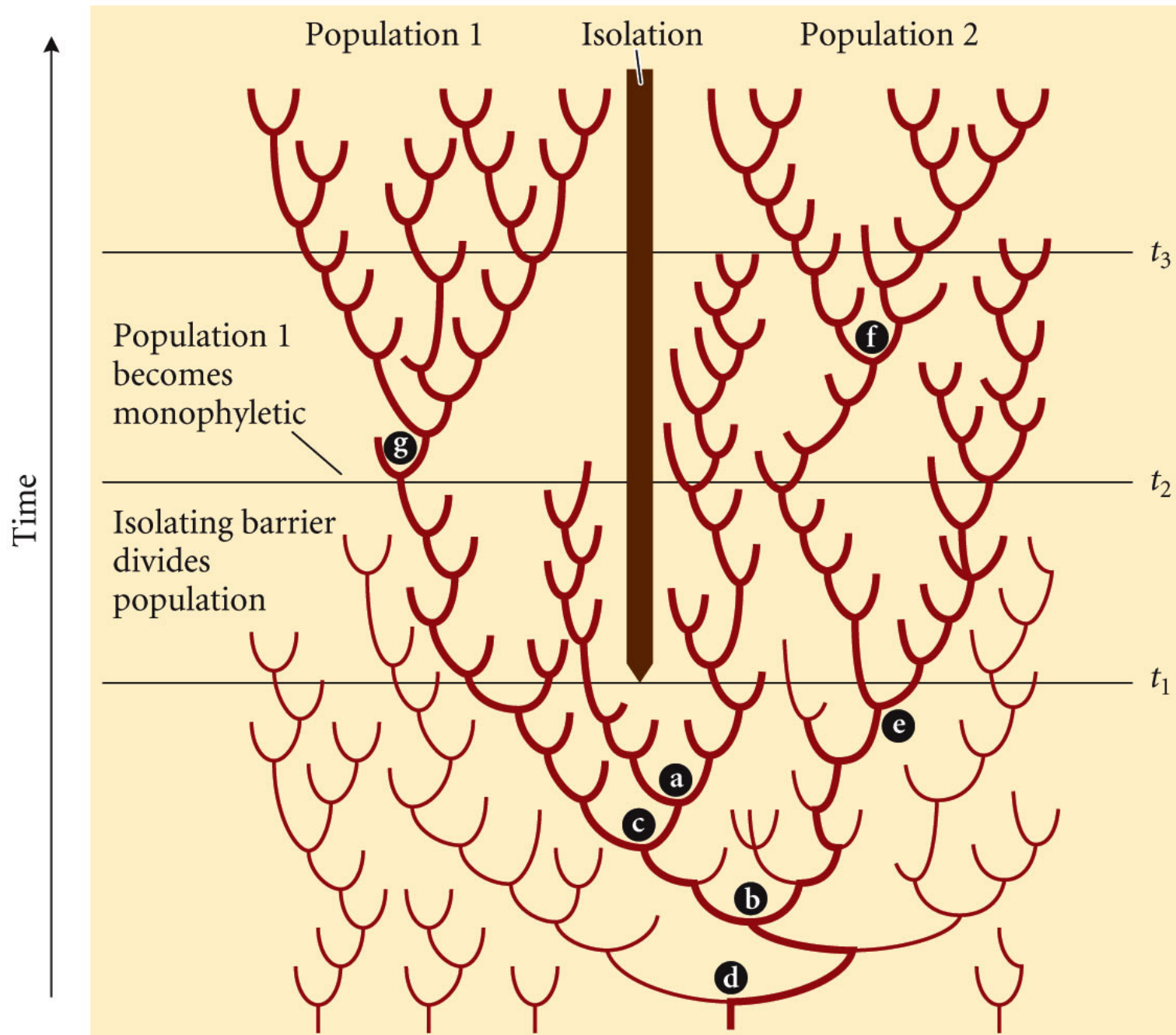
# Phylogenetic trees



# Phylogenetic trees

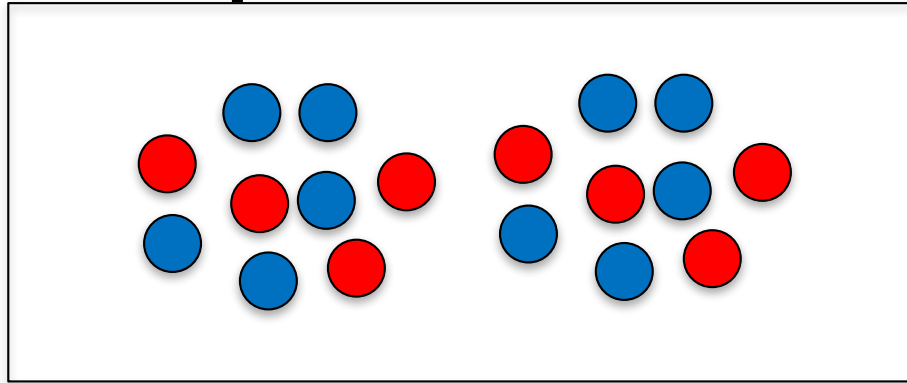




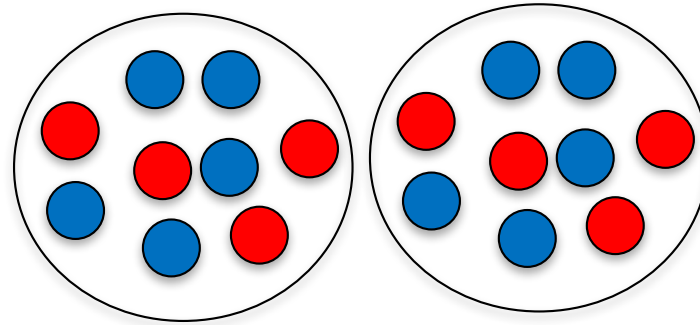


**EVOLUTION 2e, Figure 17.16**

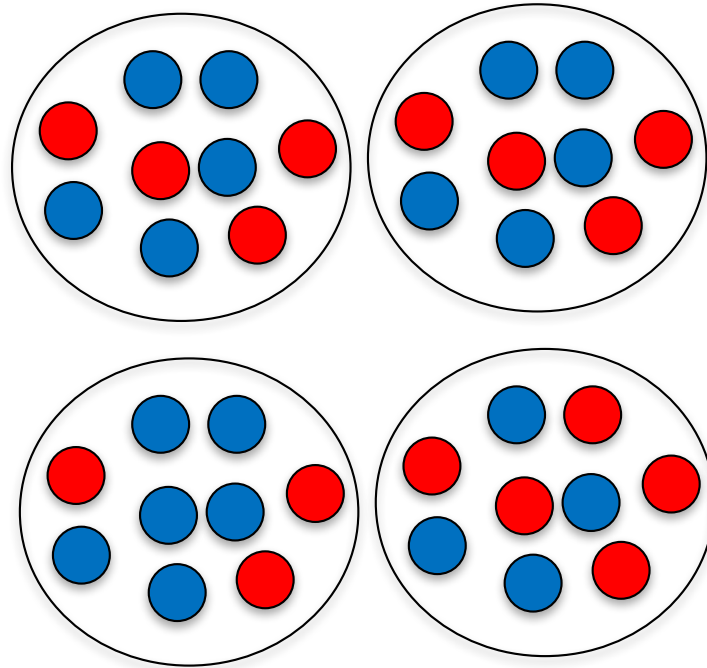
# Conceptual framework



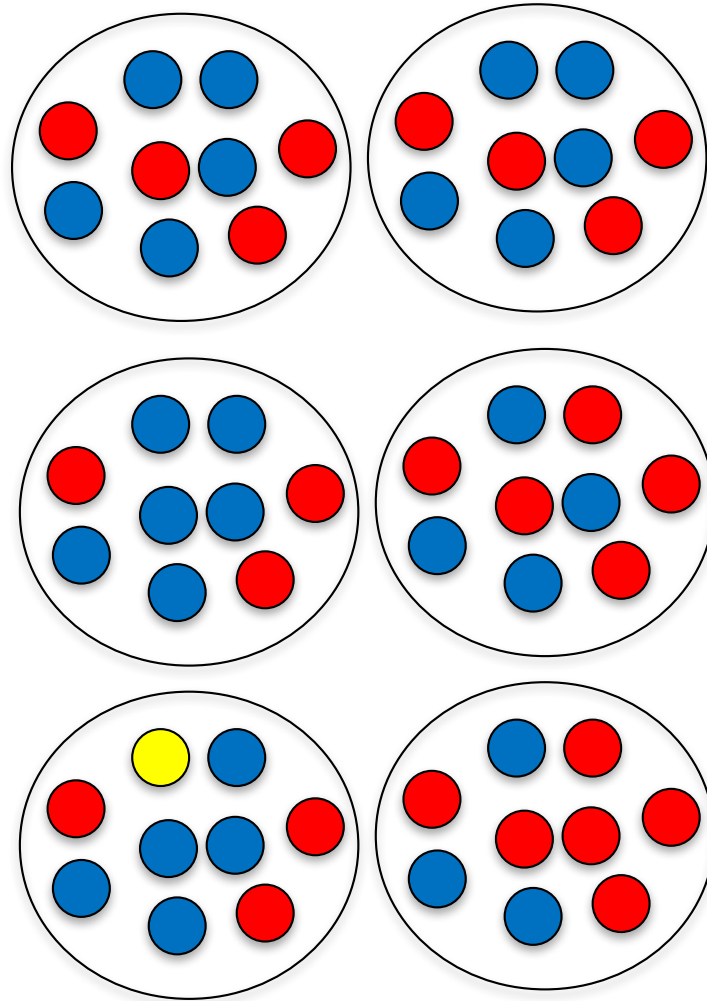
# Conceptual framework



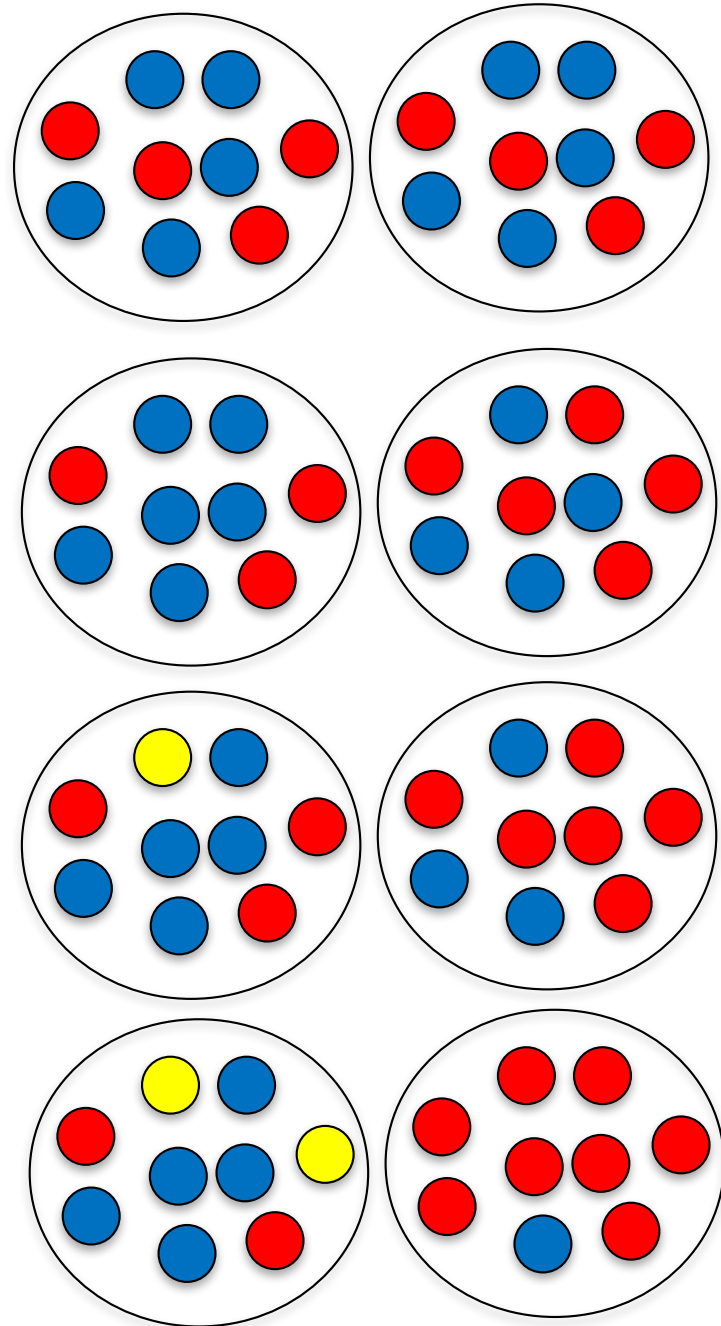
# Conceptual framework



# Conceptual framework



# Conceptual framework



# Conceptual framework

- In each lineage, new mutations are fixed independently
- Each subsequent mutation is placed on a the previously fixed sequence
- Given enough time, each lineage goes through a unique sequence of (nested) fixation events
- Sequential fixation generates a readable history of sequence similarity and differences
- How to read this sequence?
  - Given extant taxa, how to reconstruct history?
  - Use nested shared similarity to infer history

# Building phylogenetic trees

- Distance: Estimate distance matrix given data, generate tree that represents these distances
- Parsimony: Attempts to find a tree that minimizes the number of changes given data
- Maximum likelihood & Bayesian: Model-based approach to find the most-likely tree given the data



# Distance

- Compute a distance matrix given a set of biological data
  - UPGMA
  - Neighbor joining
- Compute a tree that most resembles this distance matrix

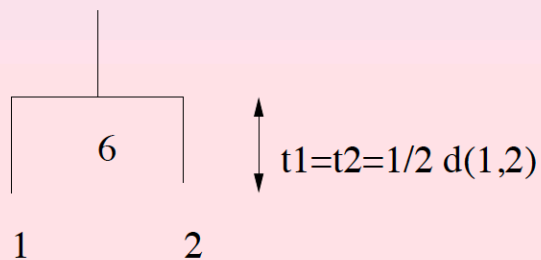
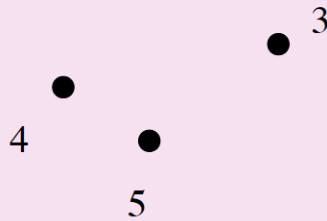
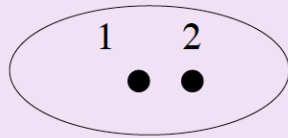
# Distance

- Unweighted pair group method using arithmetic averages (UPGMA)
- Given a set of taxa and a distance matrix, UPGMA produces a rooted tree with edge lengths
- Clusters taxa, then merging clusters
- Assembled outside in

# Distance

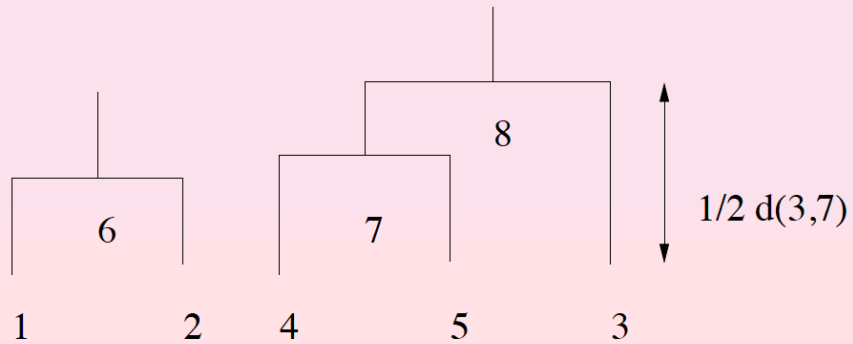
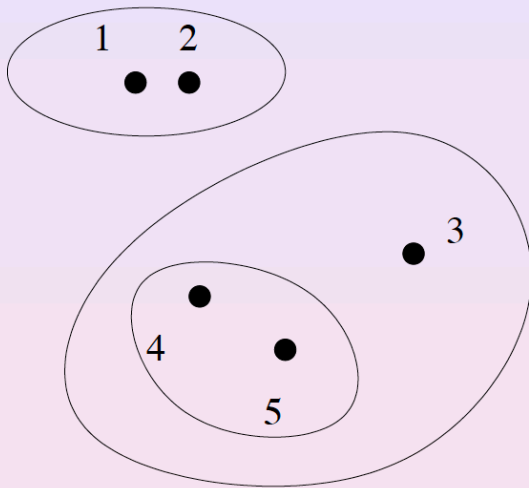
Example  $X = \{1, 2, 3, 4, 5\}$ , distances given by distance in the plane:

cluster 1 and 2:



# Distance

cluster 7 and 3:





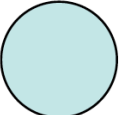
# Distance

- Neighbor-Joining (NJ)
- Saitou and Nei 1987
- Given distance matrix, produces an unrooted phylogenetic tree with edge lengths
- Repeatedly pairing neighboring taxa
- Determine which nodes are neighbors based only on distance matrix

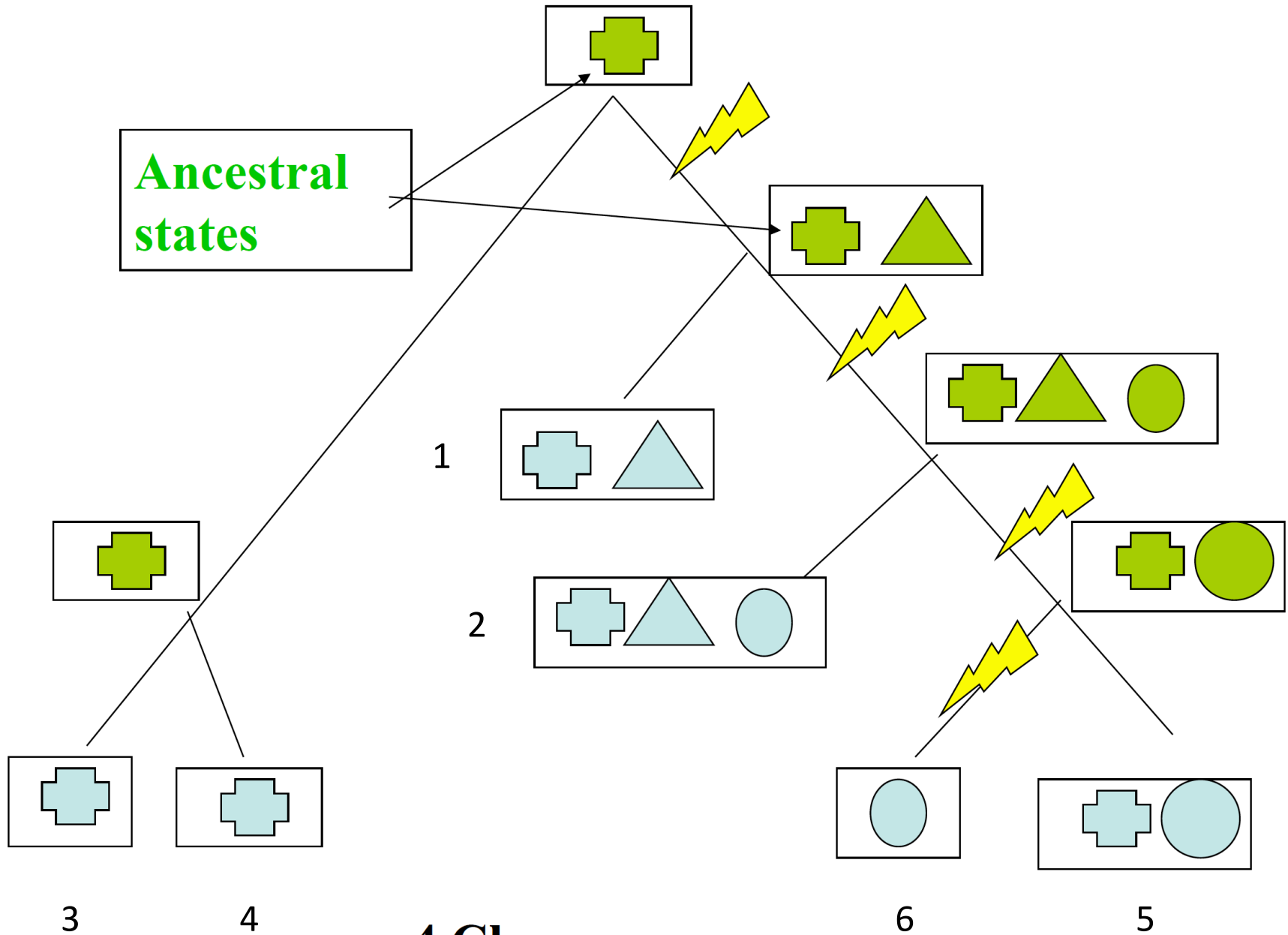
# Parsimony

- The preferred evolutionary tree is the one that requires “the minimum net amount of evolution” (Edwards and Cavalli-Sforza, 1963)
- Each taxon described by a set of characters
- Each character can be in one of a finite number of states
- Steps = changes in character states
- Goal: Find the tree that explains the distribution of character sets across taxa with the fewest number of steps

# Parsimony

			
Taxon 1	Yes	Yes	No
Taxon 2	YES	Yes	Yes
Taxon 3	Yes	No	No
Taxon 4	Yes	No	No
Taxon 5	Yes	No	Yes
Taxon 6	No	No	Yes

# Parsimony



**4 Changes**



# Parsimony

- Binary characters
- Multistate characters
- Ordered changes
- Reversible changes

# Parsimony

- Fitch
- Wagner
- Dollo
- Carmin-Sokal
- prefect

# Maximum likelihood

- Given a set of biological data, and a probabilistic model of evolution, find the tree that has the highest probability of generating the data
  - Multiple sequence alignment
  - Nucleotide substitution model

# Tree Search methods

- Exhaustive search (exact)
- Branch and bound (exact)
- Heuristic (approximate)

# Exhaustive

- Evaluate lengths of every possible tree

Number of Taxa	Number of trees
3	1
5	15
10	2,027,025
20	$10^{20}$
50	$10^{74}$

# Branch and Bound

- Hendy and Penny 1982
- Much faster, still guaranteed to find the best tree
- Determine upper bound of length of shortest tree  
Follow predictable search path through tree  
possible tree topology
- Abandon any fork of search tree where the upper bound is exceeded before the last taxon is added

# Heuristic

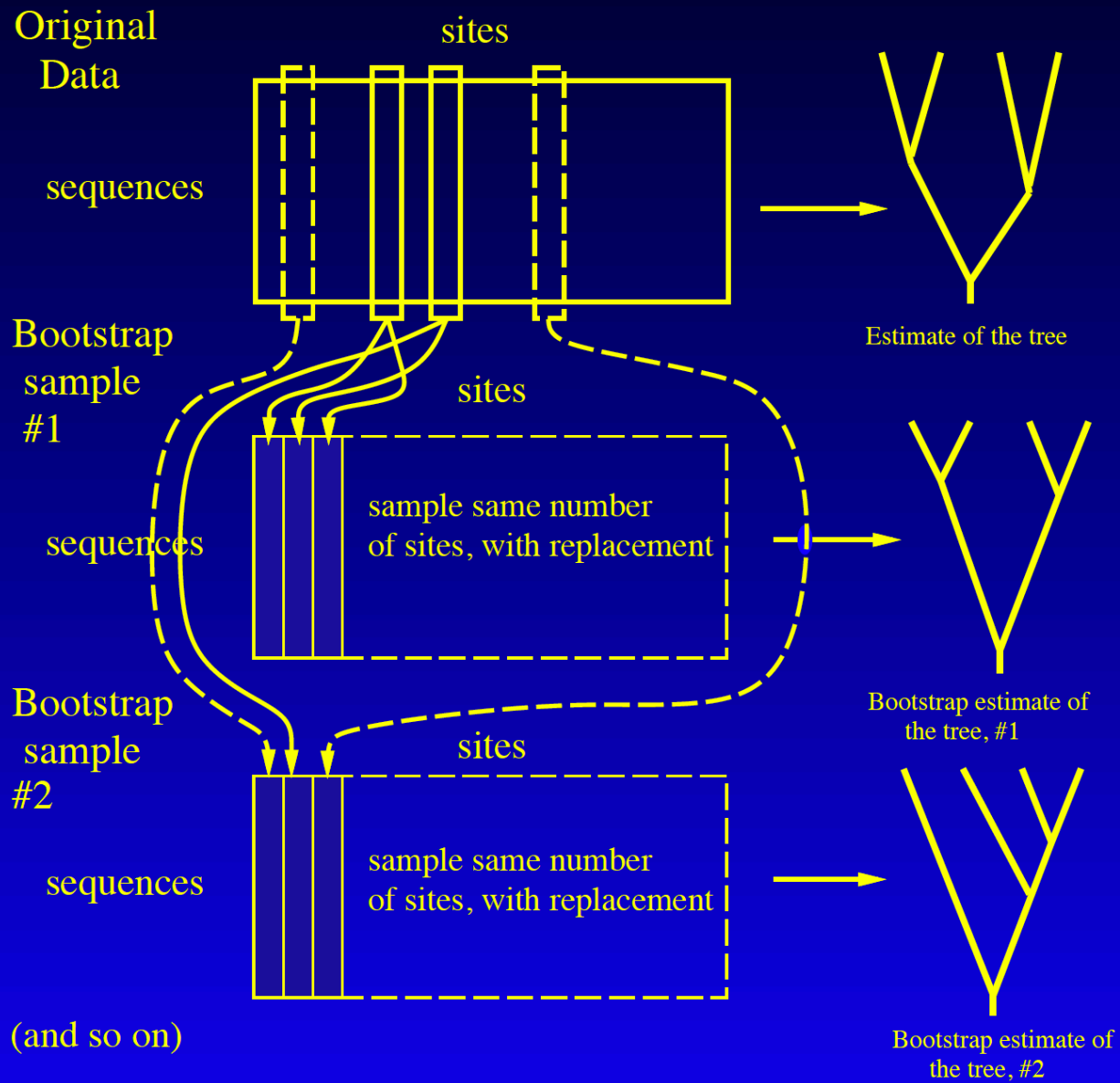
- Search trees by swapping
- Very fast
- Find a starting tree
  - Stepwise addition
  - Star decomposition
- Rearrange tree to find better trees
  - Nearest neighbor Interchange (NNI)
  - Subtree pruning and regrafting (SPR)
  - Tree bisection and Reconnection (TBR)

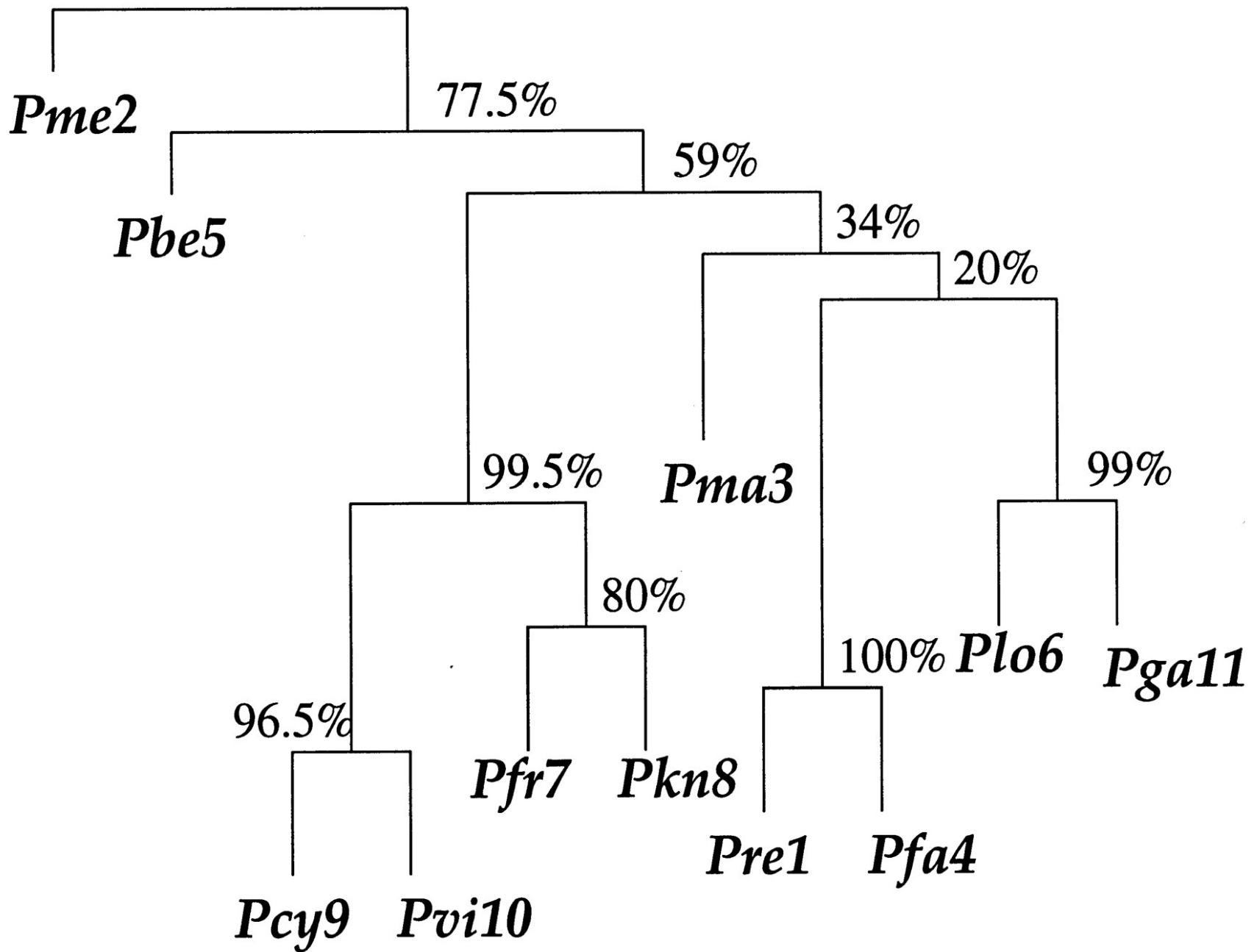
# Heuristic

- Bootstrap
  - Randomly resample the data with replacement
  - Rebuild tree
  - What fraction of the bootstrap samples show support for a particular node?



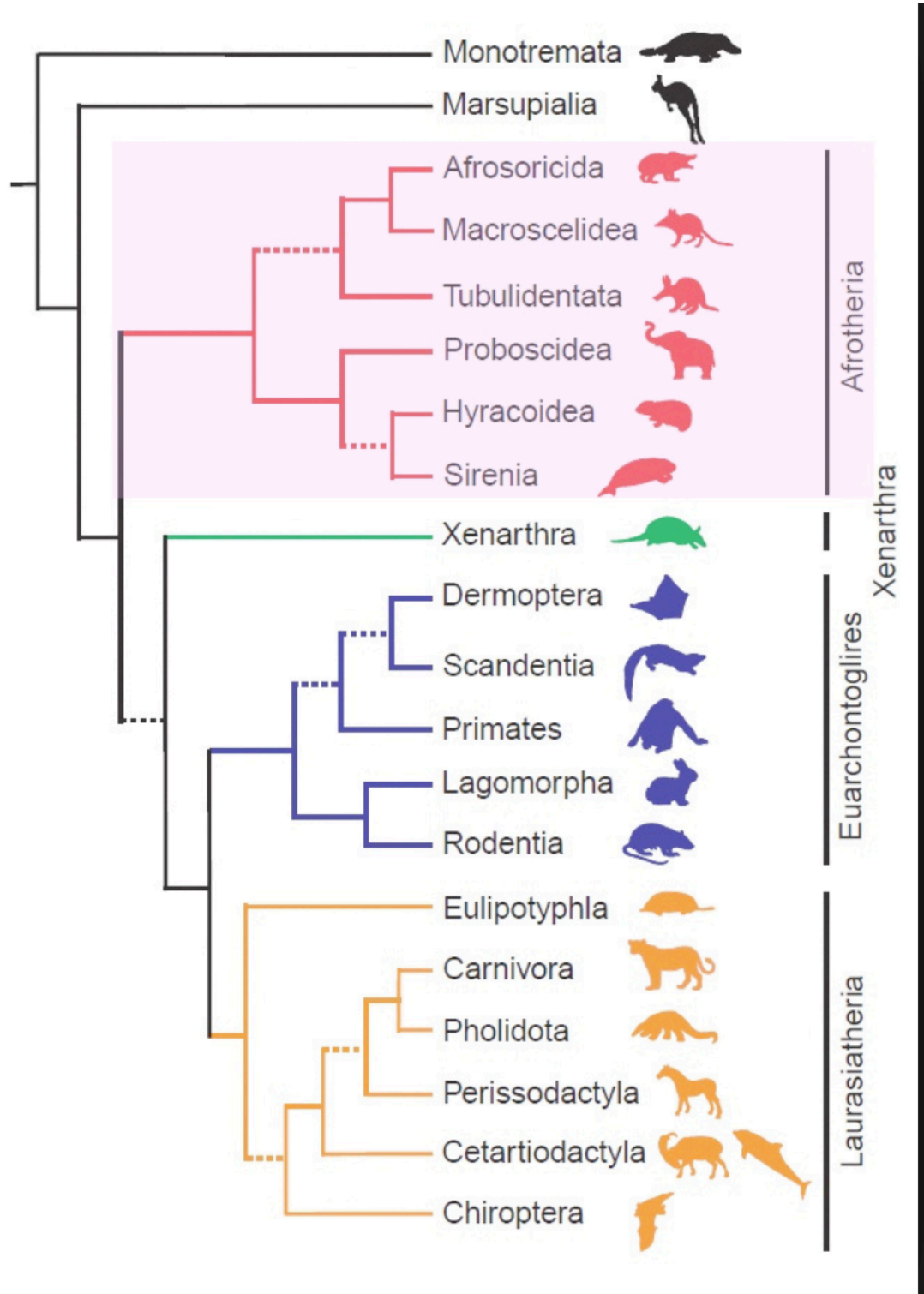
# Heuristic

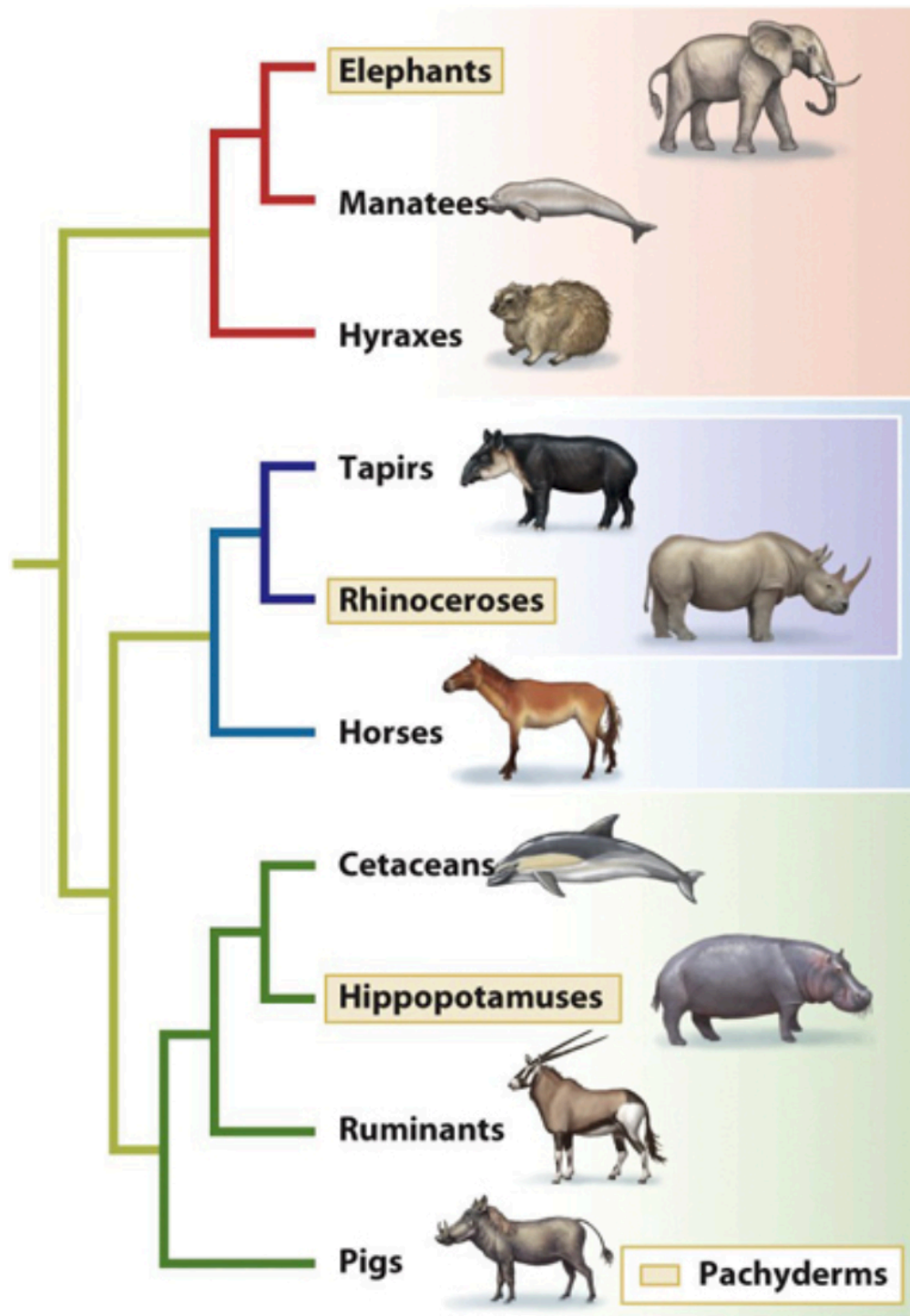




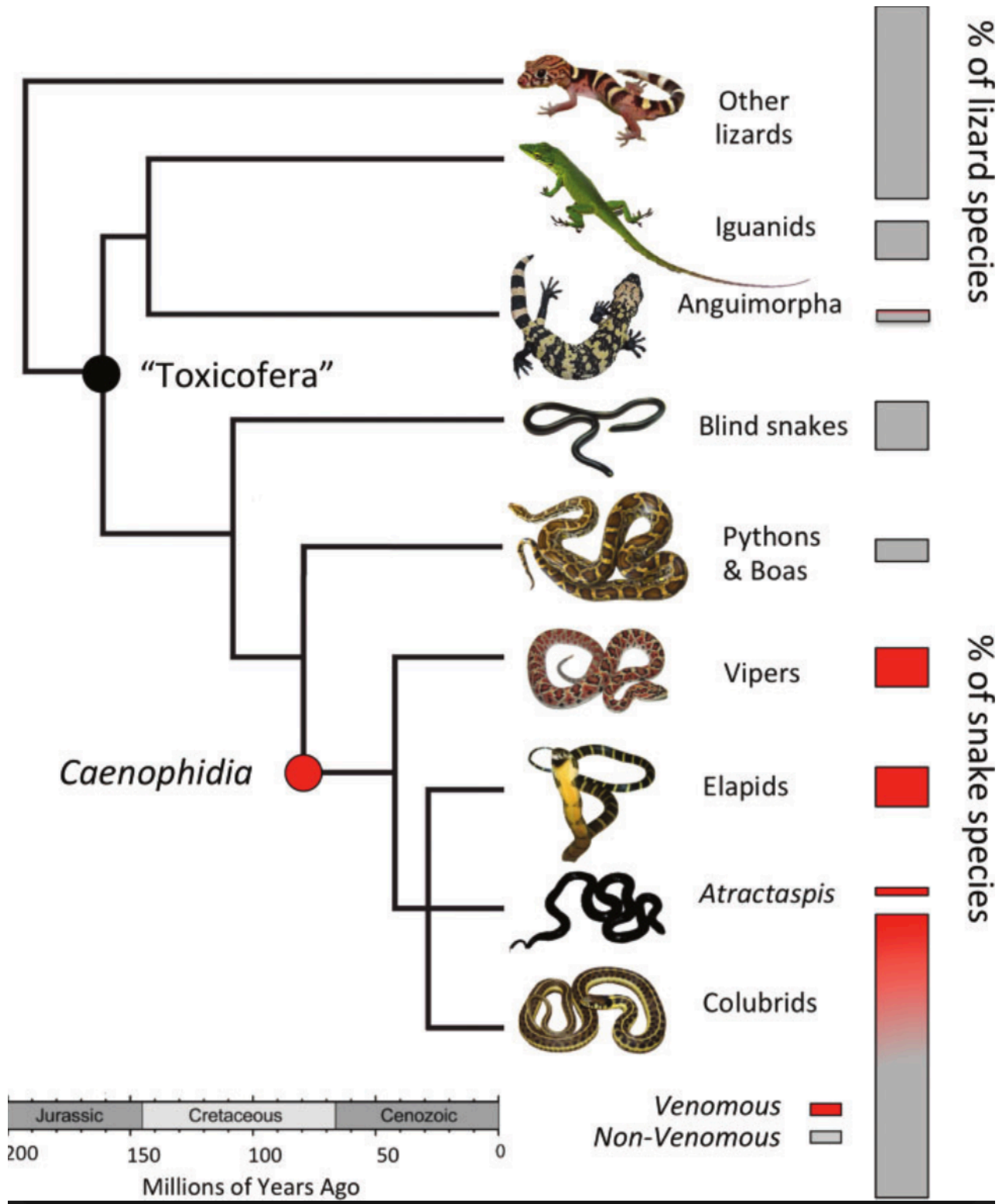
# Heuristic

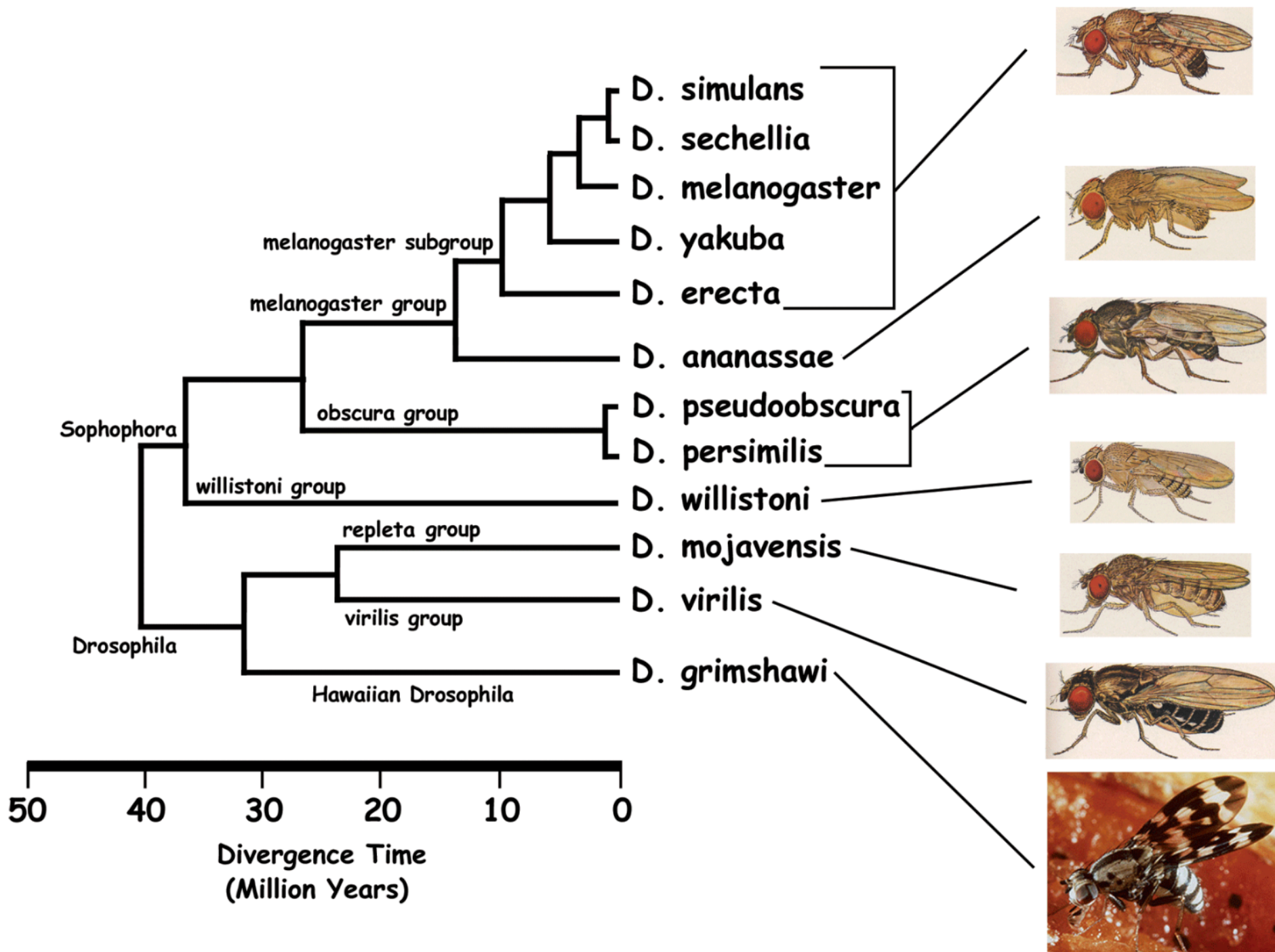
- Bootstrap
  - Randomly resample the data with replacement
  - Rebuild tree
  - What fraction of the bootstrap samples show support for a particular node?
- Jackknife
  - Randomly subset data
  - Rebuild tree
  - What fraction of jackknife samples show support for particular node
  - Whether excluding certain characters has major effect on tree





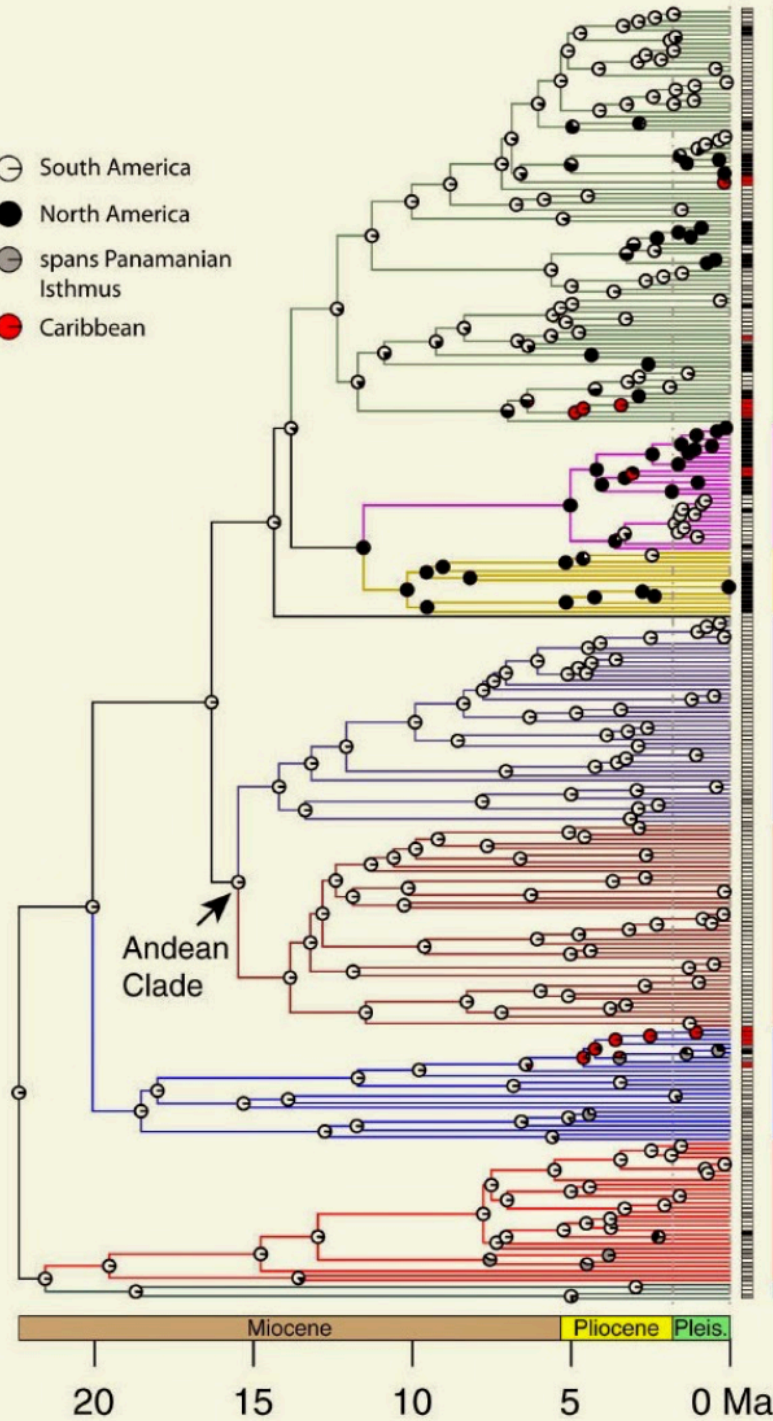
Evolution, 1/e Figure 4.11  
 © 2012 W. W. Norton & Company, Inc.







- South America
- North America
- ◐ spans Panamanian Isthmus
- Caribbean



Emeralds



Bees



Mtn. Gems  
Patagona

Coquettes



Brilliants



Mangoes



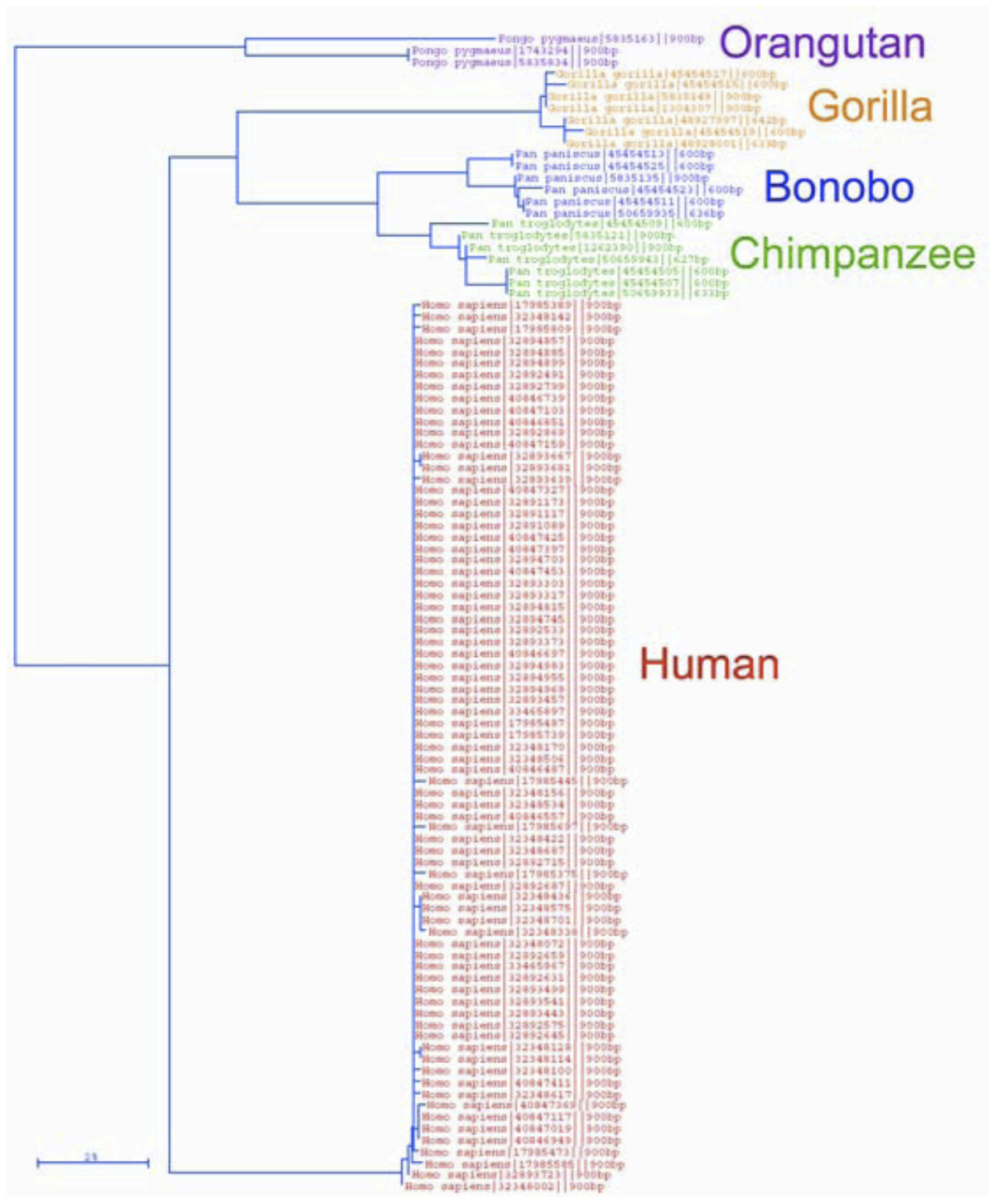
Hermits



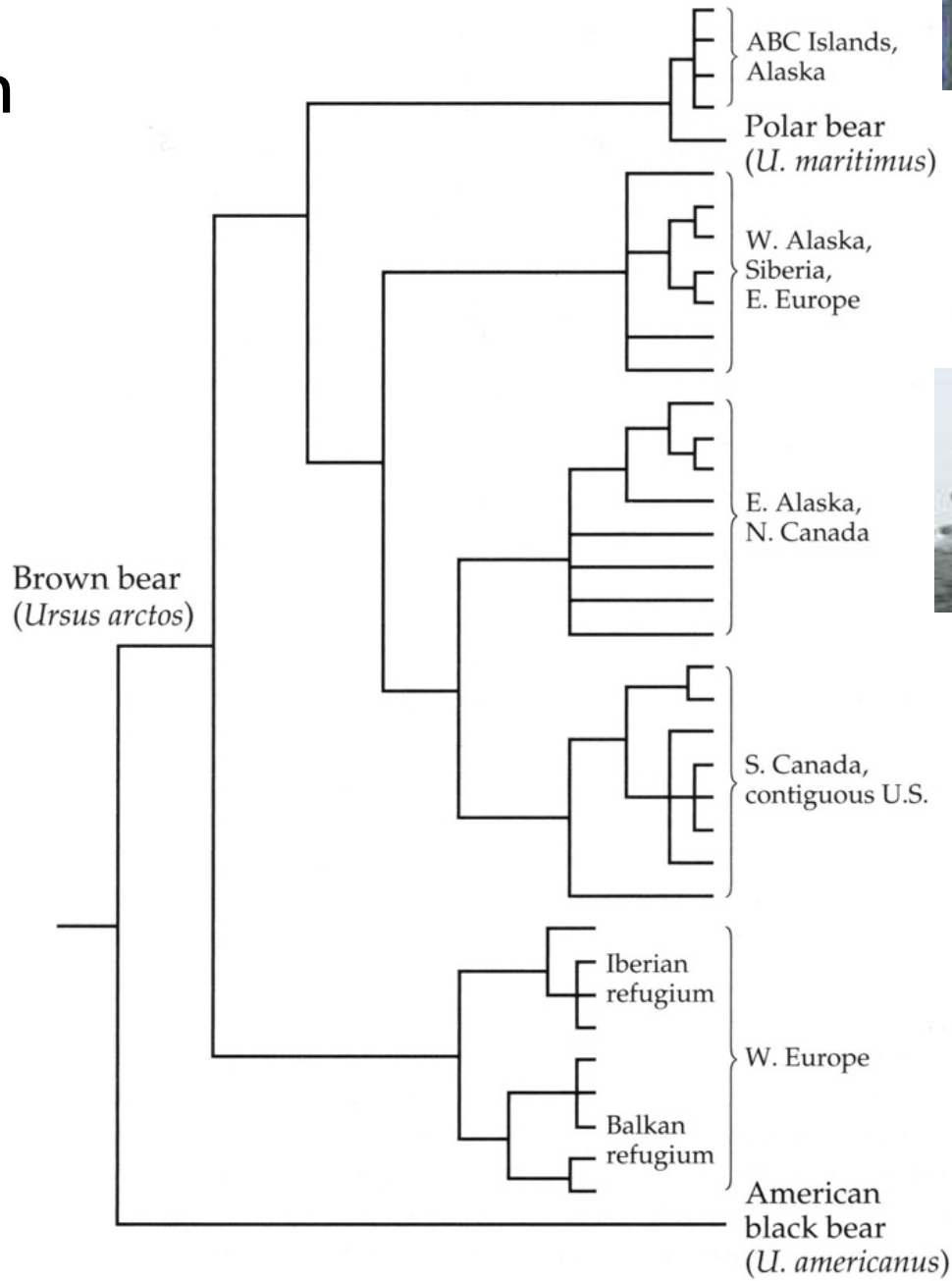
Topazes







# Populations of North American Bears



# Software: PHYLIP

- Joe Felsenstein (1980)
- Over 29,000 registered users
- Parsimony, distance matrix, ML
- DNA, RNA, protein, restriction sites, discrete characters, continuous characters, allele frequencies, distance matrices
- Freely available
- <http://evolution.gs.washington.edu/phylip.html>
- Webservers available as well

# Software: MEGA

- Molecular Evolutionary Genetic Analysis
- Sudhir Kumar
- Parsimony, distance matrix, ML
  - NJ
  - UPGMA
  - Minimum evolution
- Molecular data (nucleic acid, protein sequences)
- Bootstrapping, consensus trees
- Data editing
- Sequence alignment (with ClustalW)
- <http://www.megasoftware.net>

# Software: EMBOSS

- European Molecular Biology Open Software Suite
- Peter Rice, Alan Bleasby, Jon Ison
- General sequence analysis
- Phylogeny (PHYLIP)
- Alignment (CLUSTAL)
- All sequence formats
- Many alignment formats
- <http://emboss.sourceforge.net/what/>

# Software: Mesquite

- Wayne and David Maddison, Peter Midford, Danny Mandel, Jeff Oliver
- Set of Java modules for comparative biology
  - Over 500 functions available for data editing, management, and processing
  - Sequence alignment, visualization
  - Coalescent simulations
  - Inferences of fit of gene tree to species tree
  - Reconstruction of ancestral state (ML, parsimony)
  - Tree visualization, manipulation
- <http://mesquiteproject.org>



# Phylogeny Programs

- 392 phylogeny packages
- 54 free webservers
- Most comprehensive list

Joe Felsenstein  
<http://evolution.gs.washington.edu>



# Table of contents by methods available

- [General-purpose packages](#)
- [Parsimony programs](#)
- [Distance matrix methods](#)
- [Computation of distances](#)
- [Maximum likelihood methods](#)
- [Bayesian inference methods](#)
- [Quartets methods](#)
- [Artificial-intelligence and genetic algorithms methods](#)
- [Invariants \(or Evolutionary Parsimony\) methods](#)
- [Interactive tree manipulation](#)
- [Looking for hybridization or recombination events](#)
- [Bootstrapping and other measures of support](#)
- [Compatibility analysis](#)
- [Consensus trees, subtrees, supertrees, distances between trees](#)
- [Tree-based alignment](#)
- [Gene duplication and genomic analysis](#)
- [Biogeographic analysis and host-parasite comparison](#)
- [Comparative method analysis](#)
- [Simulation of trees or data](#)
- [Examination of shapes of trees](#)
- [Clocks, dating and stratigraphy](#)
- [Model Selection](#)
- [Description or prediction of data from trees](#)
- [Tree plotting/drawing](#)
- [Sequence management/job submission](#)
- [Teaching about phylogenies](#)
- [Web or e-mail servers that can analyze data for you](#)

Joe Felsenstein

<http://evolution.gs.washington.edu>



## General-purpose packages

- [PHYLIP](#)
- [PAUP\\*](#)
- [MEGA](#)
- [Phylo win](#)
- [ARB](#)
- [DAMBE](#)
- [PAL](#)
- [Bionumerics](#)
- [Mesquite](#)
- [PaupUp](#)
- [BIRCH](#)
- [Bosque](#)
- [EMBOSS](#)
- [phangorn](#)
- [Bio++](#)
- [ETE](#)
- [DendroPy](#)
- [SeaView](#)
- [Crux](#)

## Parsimony programs

- [PHYLIP](#)
- [PAUP\\*](#)
- [Hennig86](#)
- [MEGA](#)
- [RA](#)
- [NONA](#)
- [CAFCA](#)
- [PHYLIP](#)
- [Phylo win](#)
- [sog](#)
- [gmaes](#)
- [LVB](#)
- [GeneTree](#)
- [ARB](#)
- [DAMBE](#)
- [MALIGN](#)
- [POY](#)
- [Gambit](#)
- [TNT](#)
- [GelCompar II](#)
- [Bionumerics](#)
- [Network](#)
- [TCS](#)
- [GAPars](#)
- [CRANN](#)
- [Mesquite](#)
- [PAST](#)
- [FootPrinter](#)
- [BPAnalysis](#)
- [Simplot](#)
- [Parsimov](#)
- [NimbleTree](#)
- [PaupUp](#)
- [Notung](#)
- [BIRCH](#)
- [IDEA](#)
- [PSODA](#)
- [PRAP](#)
- [SeqState](#)
- [Bosque](#)
- [PhyloNet](#)
- [EMBOSS](#)
- [phangorn](#)
- [Murka](#)
- [Freqpars](#)
- [SeaView](#)
- [PAUPRat](#)

Joe Felsenstein

<http://evolution.gs.washington.edu>

## Distance matrix methods

- [PHYLIP](#)
- [PAUP\\*](#)
- [MEGA](#)
- [MacT](#)
- [ODEN](#)
- [TREECON](#)
- [DISPAN](#)
- [RESTSITE](#)
- [NTSYSpc](#)
- [METREE](#)
- [GDA](#)
- [SeqPup](#)
- [PHYLTEST](#)
- [Lintre](#)
- [Phylo\\_win](#)
- [POPTREE2](#)
- [Gambit](#)
- [gmaes](#)
- [DENDRON](#)
- [BIONJ](#)
- [TFPGA](#)
- [MVSP](#)
- [ARB](#)
- [Darwin](#)
- [T-REX](#)
- [sendbs](#)
- [nneighbor](#)
- [DAMBE](#)
- [weighbor](#)
- [DNASIS](#)
- [MINSPNET](#)
- [PAL](#)
- [Arlequin](#)
- [PEBBLE](#)
- [HY-PHY](#)
- [Vanilla](#)
- [GelCompar II](#)
- [Bionumerics](#)
- [qclust](#)
- [TCS](#)
- [Populations](#)
- [Winboot](#)
- [SYN-TAX](#)
- [PTP](#)
- [SplitsTree](#)
- [FastME](#)
- [APE](#)
- [MacVector](#)
- [QuickTree](#)
- [Simplot](#)
- [ProfDist](#)
- [START2](#)
- [STC](#)
- [NimbleTree](#)
- [CBCAnalyzer](#)
- [PaupUp](#)
- [Geneious](#)
- [BIRCH](#)
- [SEMPHY](#)
- [FASTML](#)
- [Rate4Site](#)
- [SWORDS](#)
- [IDEA](#)
- [FAMD](#)
- [Bosque](#)
- [GAME](#)
- [Bioinformatics Toolbox](#)
- [TreeFit](#)
- [EMBOSS](#)
- [phangorn](#)
- [PC-ORD](#)
- [Bio++](#)
- [UGENE](#)
- [NINJA](#)
- [SeaView](#)
- [Statio](#)
- [TIMER](#)
- [Crux](#)
- [Ancestor](#)
- [ANC-GENE](#)
- [Bn-Bs](#)

Joe Felsenstein

<http://evolution.gs.washington.edu>

## Computation of distances

- [PHYLIP](#)
- [PAUP\\*](#)
- [RAPDistance](#)
- [MULTICOMP](#)
- [Microsat](#)
- [DIPLOMO](#)
- [OSA](#)
- [DISPAN](#)
- [RESTSITE](#)
- [NTSYSpc](#)
- [TREE-PUZZLE](#)
- [GCUA](#)
- [DERANGE2](#)
- [POPGENE](#)
- [TFPGA](#)
- [REAP](#)
- [MVSP](#)
- [RSTCALC](#)
- [Genetix](#)
- [DISTANCE](#)
- [Darwin](#)
- [sendbs](#)
- [Arlequin](#)
- [DAMBE](#)
- [DnaSP](#)
- [PAML](#)
- [puzzleboot](#)
- [PAL](#)
- [Vanilla](#)
- [GelCompar II](#)
- [Bionumerics](#)
- [qclust](#)
- [Populations](#)
- [Winboot](#)
- [FSTAT](#)
- [SYN-TAX](#)
- [Phylo win](#)
- [Phyltools](#)
- [MSA](#)
- [APE](#)
- [YCDMA](#)
- [NSA](#)
- [T-REX](#)
- [LDDist](#)
- [DIVAGE](#)
- [Genepop](#)
- [START2](#)
- [Swaap](#)
- [Swaap PH](#)
- [SPAGeDi](#)
- [CBCAnalyzer](#)
- [PaupUp](#)
- [SEMPHY](#)
- [SWORDS](#)
- [rRNA phylogeny](#)
- [FAMD](#)
- [GAME](#)
- [Bioinformatics Toolbox](#)
- [GenoDive](#)
- [analysis](#)
- [TreeFit](#)
- [EMBOSS](#)
- [Murka](#)
- [Bio++](#)
- [UGENE](#)
- [POPTREE2](#)
- [DISTREE](#)
- [SeaView](#)
- [Crux](#)
- [Bn-Bs](#)
- [HON-new](#)

Joe Felsenstein

<http://evolution.gs.washington.edu>

## Maximum likelihood methods

- [PHYLIP](#)
- [PAUP\\*](#)
- [fastDNAm1](#)
- [MOLPHY](#)
- [PAML](#)
- [Spectrum](#)
- [SplitsTree](#)
- [TREE-PUZZLE](#)
- [SeqPup](#)
- [Phylo win](#)
- [PASSML](#)
- [ARB](#)
- [Darwin](#)
- [Modeltest](#)
- [DAMBE](#)
- [PAL](#)
- [dnarates](#)
- [HY-PHY](#)
- [Vanilla](#)
- [DT-ModSel](#)
- [Bionumerics](#)
- [fastDNAm1Rev](#)
- [RevDNARates](#)
- [rate-evolution](#)
- [CONSEL](#)
- [EDIBLE](#)
- [PLATO](#)
- [Mesquite](#)
- [PTP](#)
- [Treefinder](#)
- [MetaPIGA](#)
- [RAxML](#)
- [PHYML](#)
- [r8s-bootstrap](#)
- [MrMTgui](#)
- [MrModeltest](#)
- [BootPHYML](#)
- [PARBOOT](#)
- [p4](#)
- [Porn\\*](#)
- [SIMMAP](#)
- [Spectronet](#)
- [Rhino](#)
- [TipDate](#)
- [ProtTest](#)
- [ModelGenerator](#)
- [Simplot](#)
- [MrAIC](#)
- [Modelfit](#)
- [IQPNNI](#)
- [PARAT](#)
- [ALIFRITZ](#)
- [PhyNav](#)
- [DPRML](#)
- [MultiPhyl](#)
- [NimbleTree](#)
- [PaupUp](#)
- [SSA](#)
- [CoMET](#)
- [BIRCH](#)
- [Mac5](#)
- [Kakusan4](#)
- [GARLI](#)
- [PHYSIG](#)
- [SEMPHY](#)
- [FASTML](#)
- [Rate4Site](#)
- [aLRT](#)
- [McRate](#)
- [EREM](#)
- [PROCOV](#)
- [DART](#)
- [PhyloCoCo](#)
- [PRAP](#)
- [SeqState](#)
- [Leaphy](#)
- [NHML](#)
- [SLR](#)
- [rRNA phylogeny](#)
- [Bosque](#)
- [Concaterpillar](#)
- [PHYLLAB](#)
- [NEPAL](#)
- [EMBOSS](#)
- [CodeAxe](#)
- [phangorn](#)
- [Bio++](#)
- [FastTree](#)
- [nhPhyML](#)
- [PhyML-Multi](#)
- [Segminator](#)
- [raxmlGUI](#)
- [MixtureTree](#)
- [SeaView](#)
- [GZ-Gamma](#)
- [PAUPRat](#)
- [Crux](#)

Joe Felsenstein

<http://evolution.gs.washington.edu>

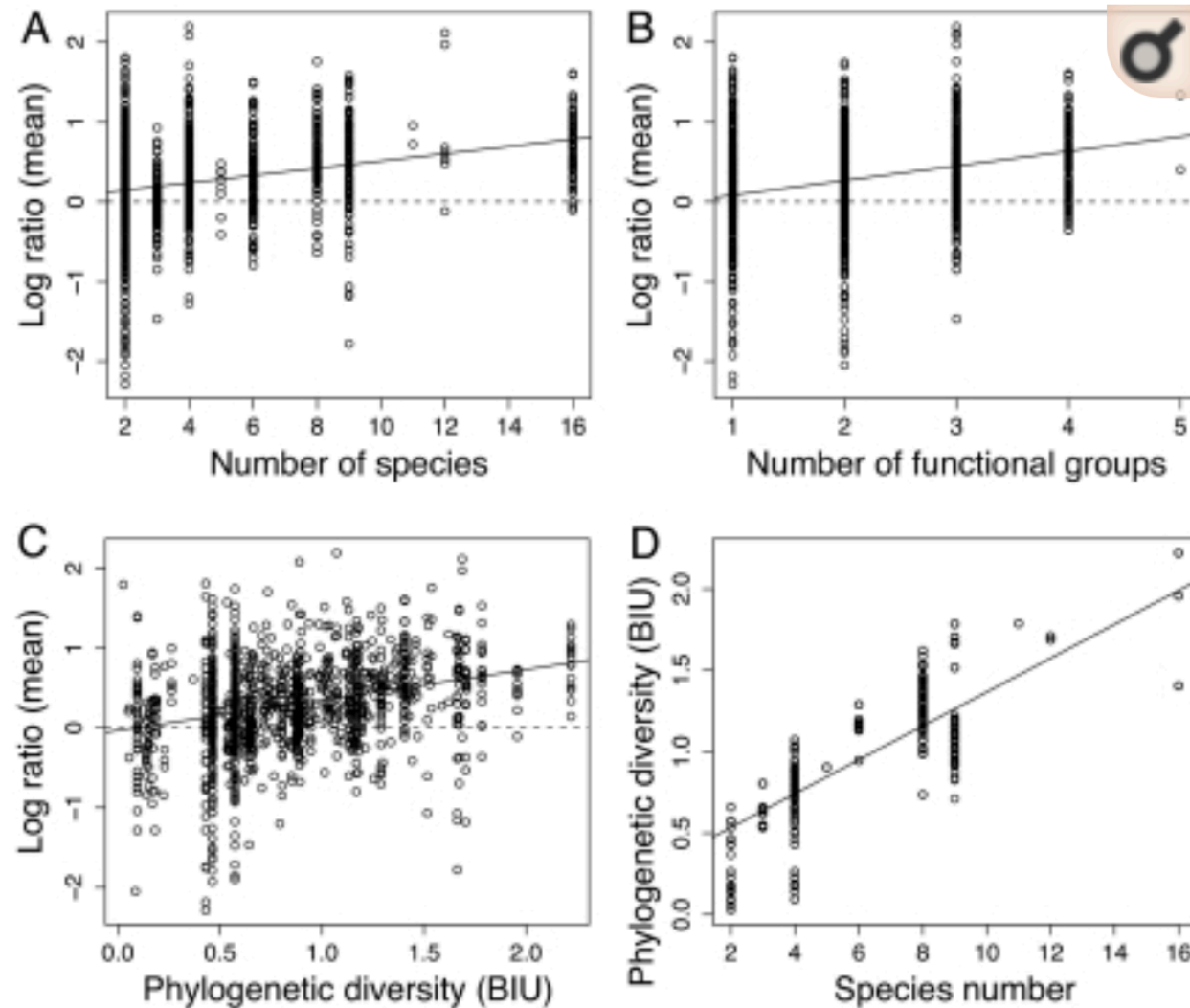
## Bayesian inference methods

- [PAML](#)
- [BAMBE](#)
- [PAL](#)
- [Vanilla](#)
- [MrBayes](#)
- [Mesquite](#)
- [PHASE](#)
- [BEAST](#)
- [MrBayes tree scanners](#)
- [p4](#)
- [SIMMAP](#)
- [IMa2](#)
- [BAli-Phy](#)
- [BayesPhylogenies](#)
- [MrBayesPlugin](#)
- [PhyloBayes](#)
- [PHASE](#)
- [Cadence](#)
- [Multidivtime](#)
- [BEST](#)
- [AMBIORE](#)
- [PHYLLAB](#)
- [bms\\_runner](#)
- [tracer](#)
- [burntrees](#)
- [Bio++](#)
- [Crux](#)
- [ANC-GENE](#)

Joe Felsenstein  
<http://evolution.gs.washington.edu>

# Phylogenetics in conservation

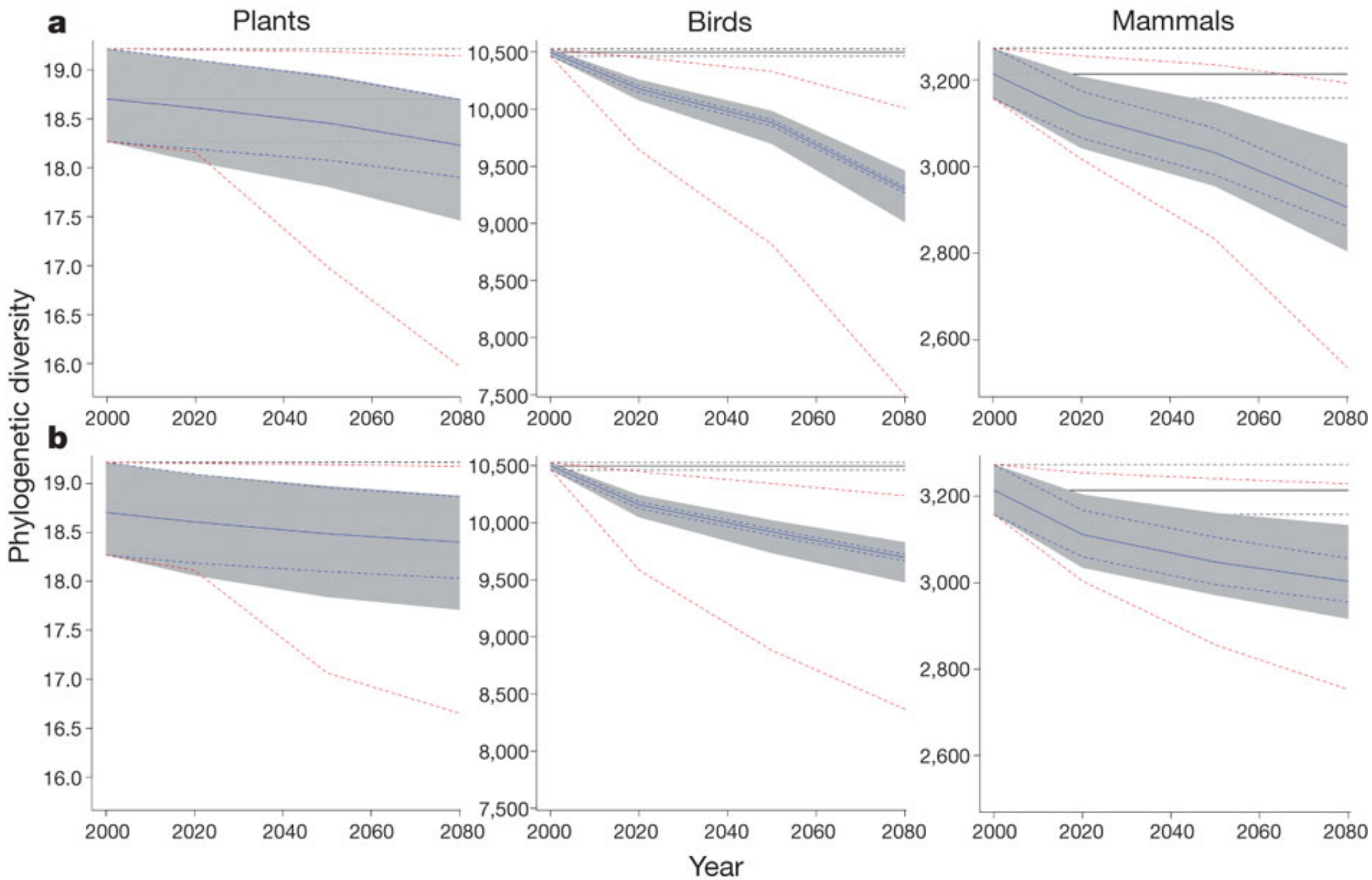
- Phylogenetic diversity and ecosystem function



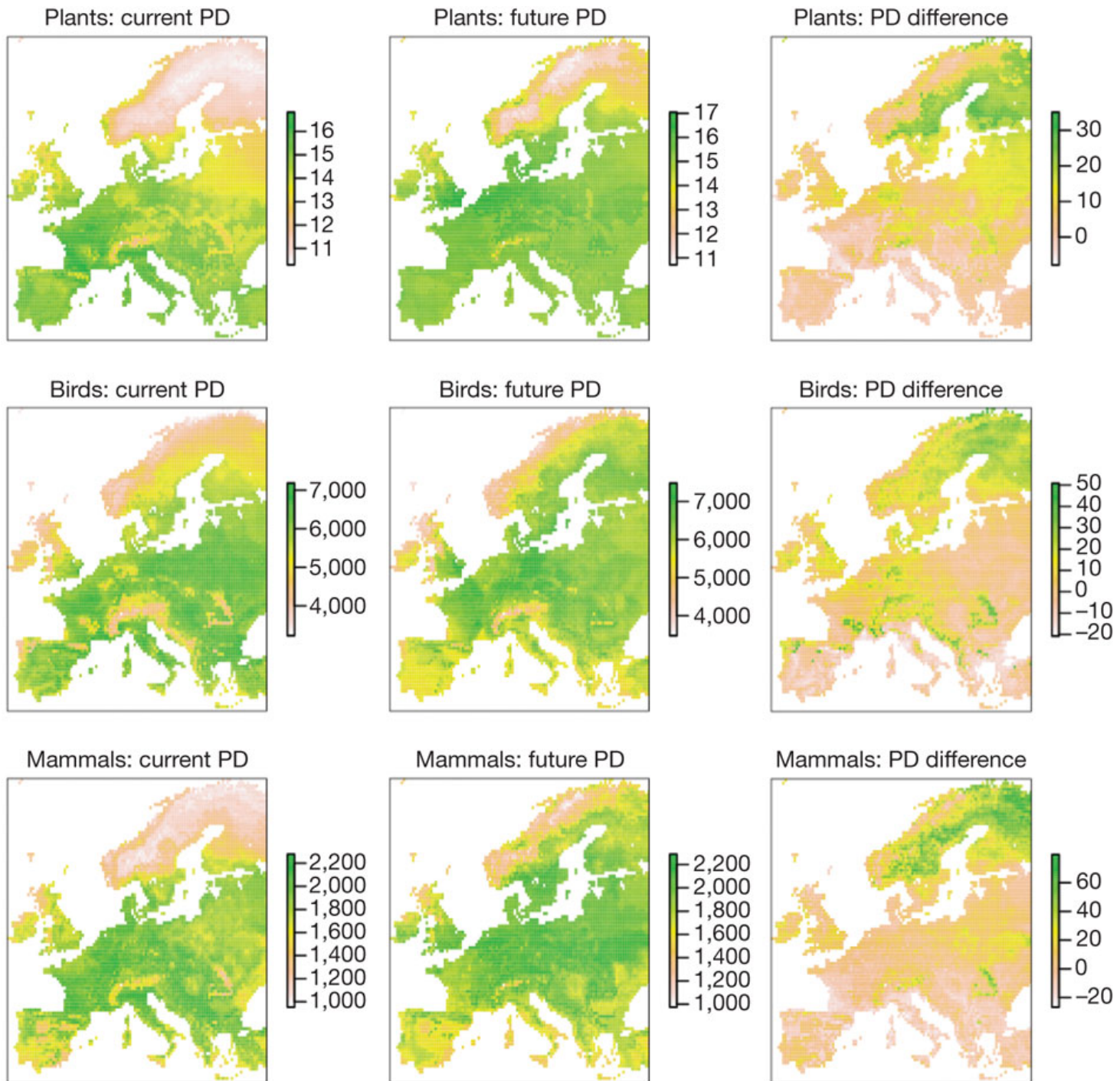
# Phylogenetics in conservation

- Phylogenetic diversity and ecosystem function
- Global change and extinction





Thuiller et al. 2011.



Thuiller et al. 2011.

# Phylogenetics in conservation

- Phylogenetic diversity and ecosystem function
- Global change and extinction
- Conservation planning

