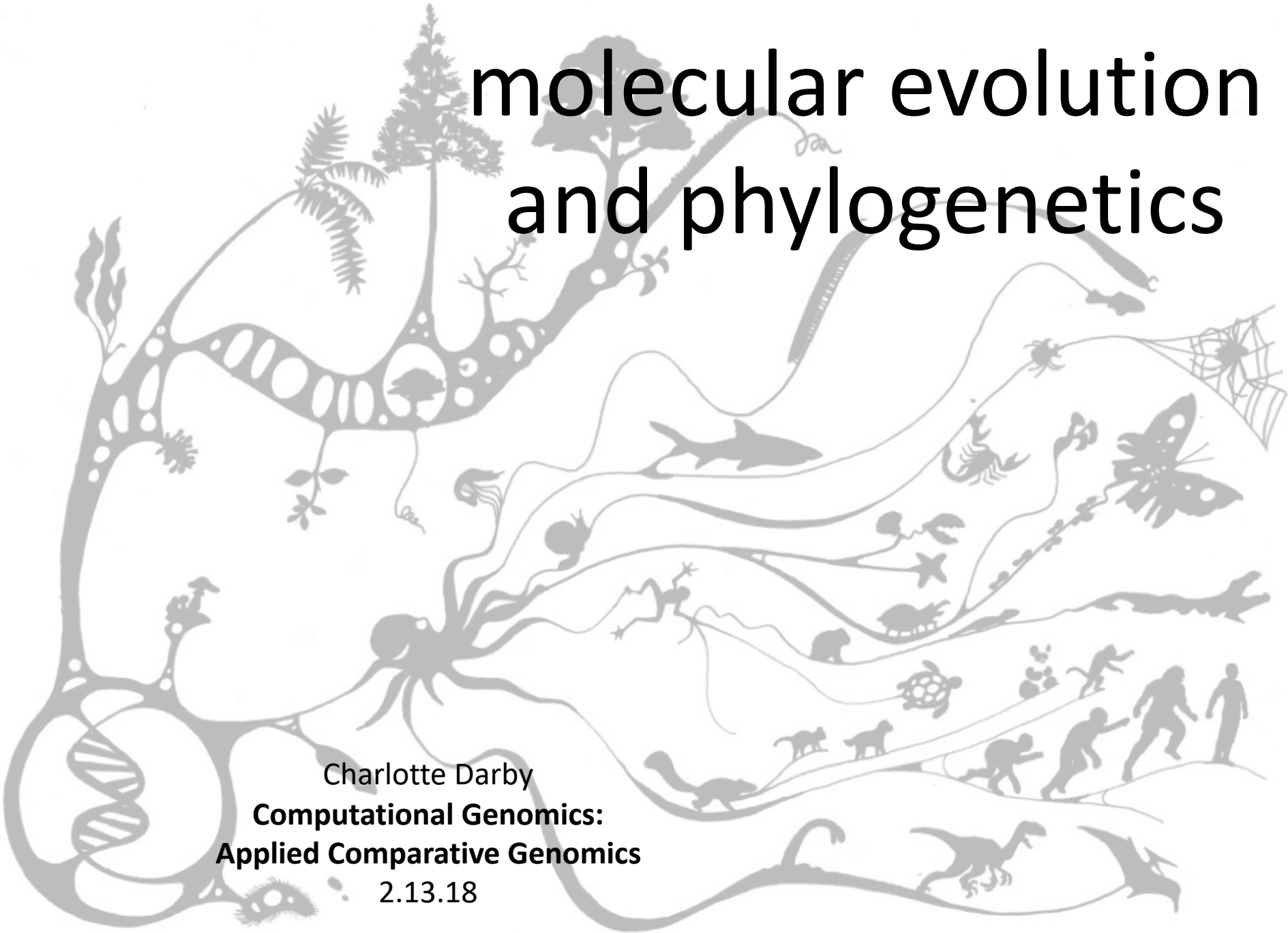
molecular evolution and phylogenetics

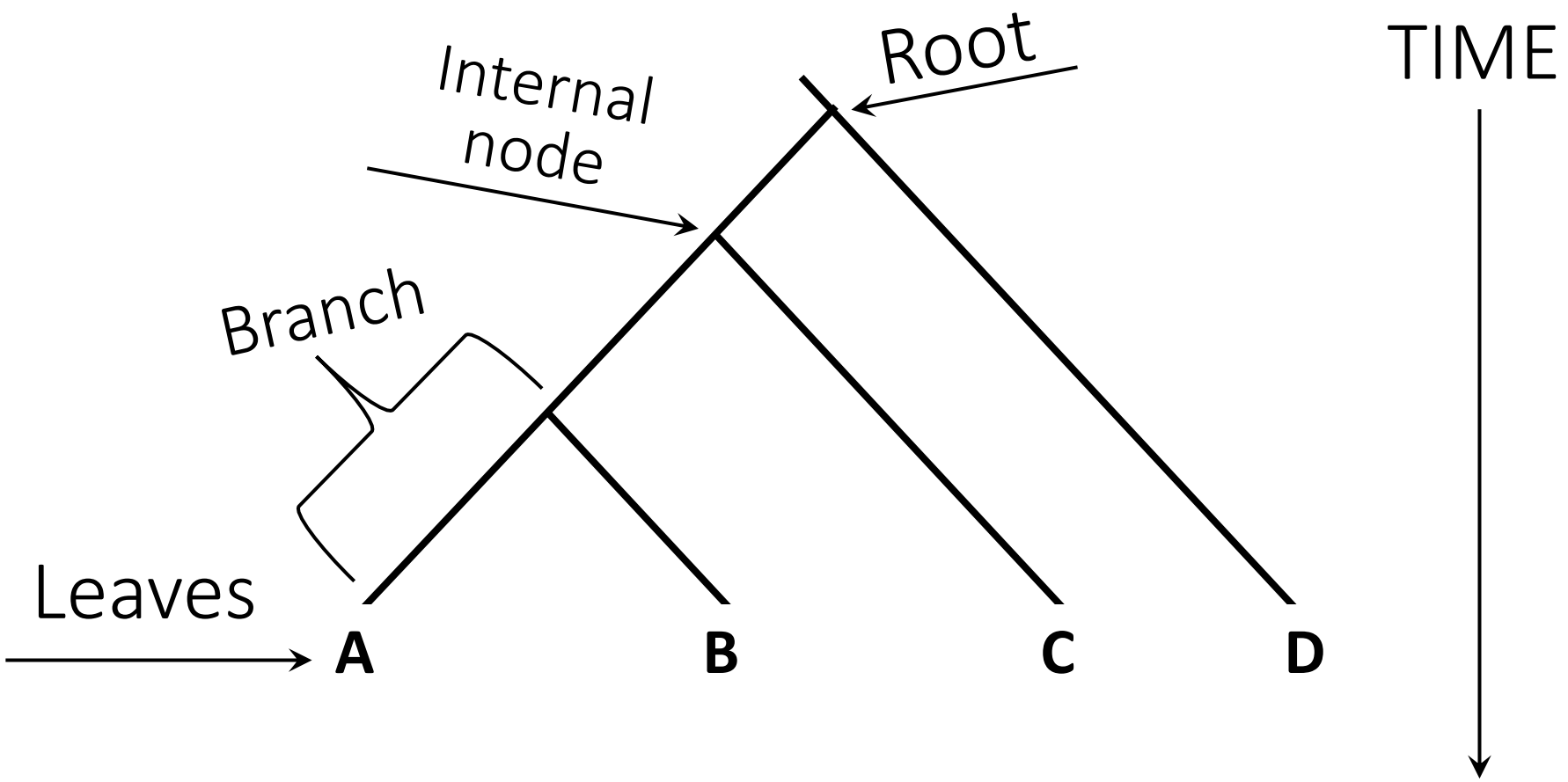


Charlotte Darby

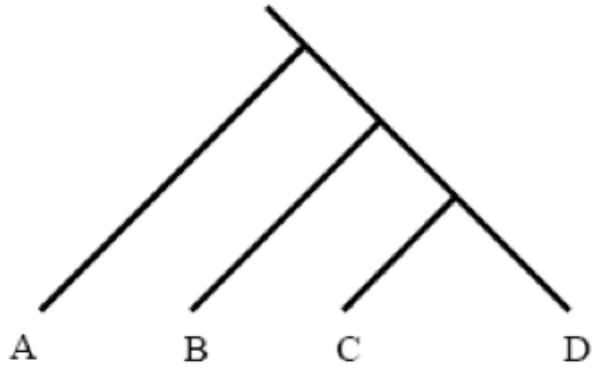
Computational Genomics:
Applied Comparative Genomics

2.13.18

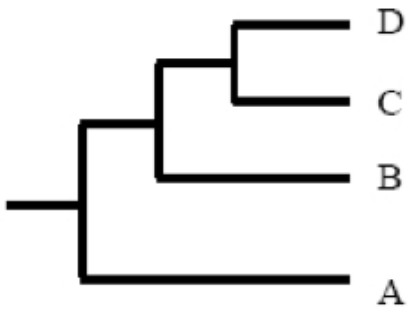
https://www.thinglink.com/scene/762084640000311296


Root
Internal node
Branch
Leaves
A
B
C
D
TIME

computer scientist

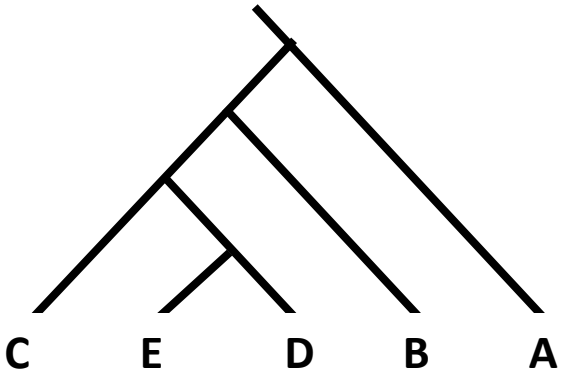


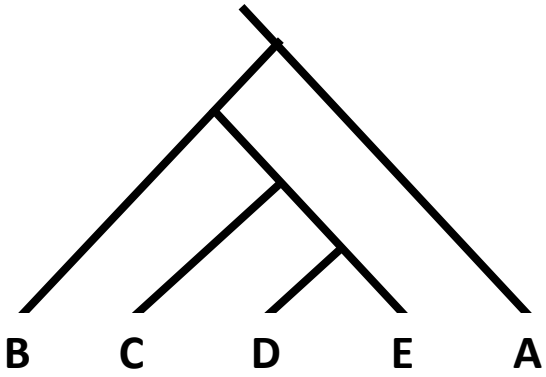
biologist

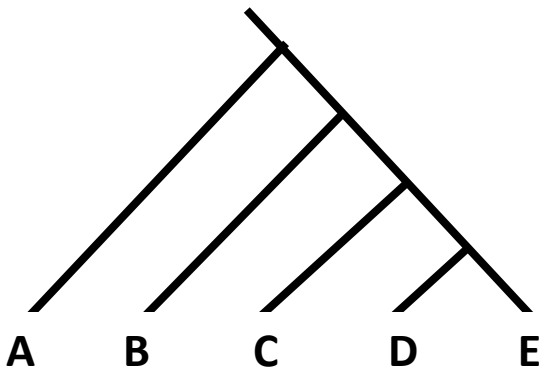


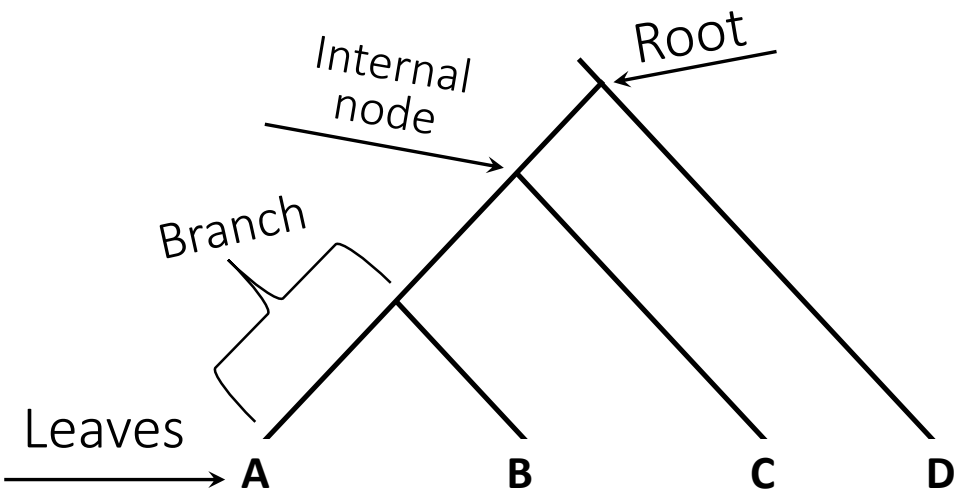
???

Each set of three figures shows the same topological relationship among the leaves





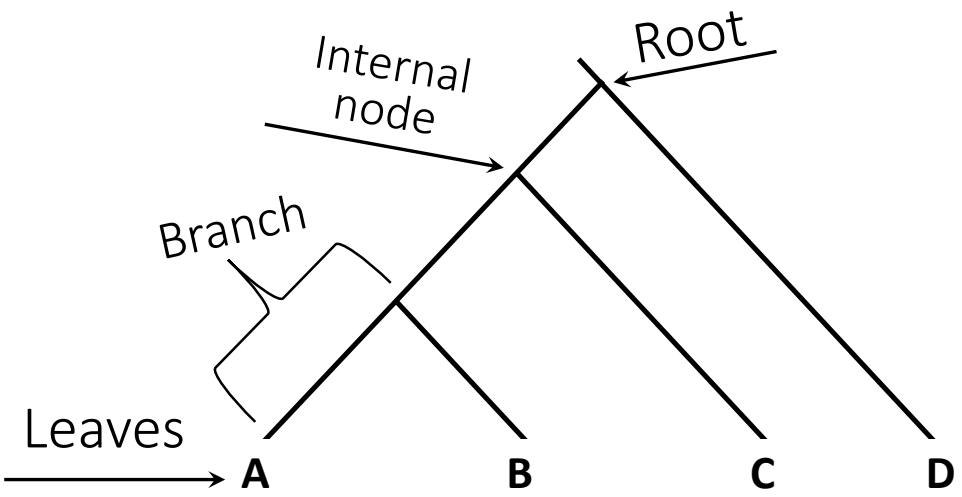




Leaves

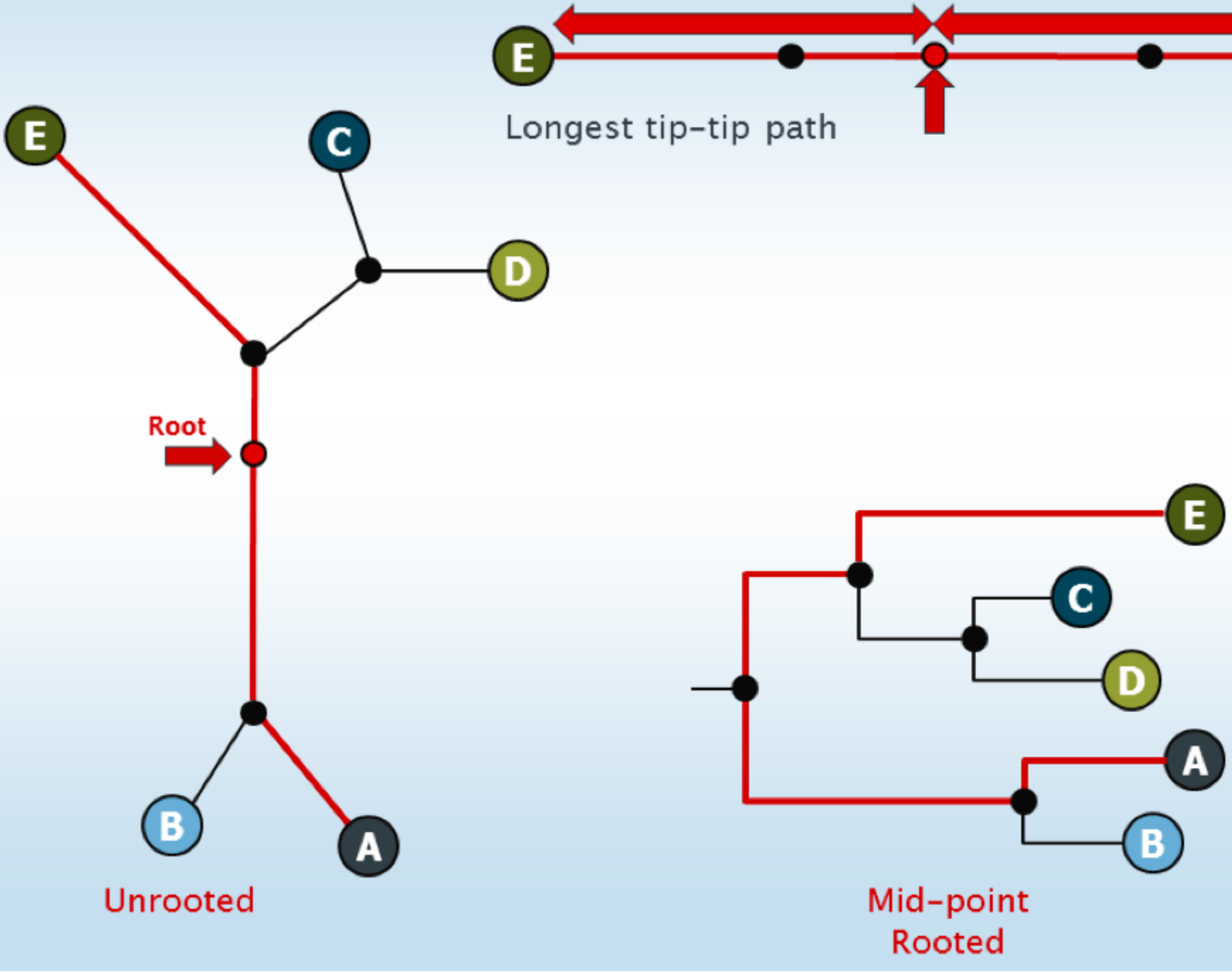
- Strains or isolates of the same species
- OR different species
- OR DNA or protein sequences
  - From the same genome or different genomes

Leaves are implied to be extant
(present at the current time)

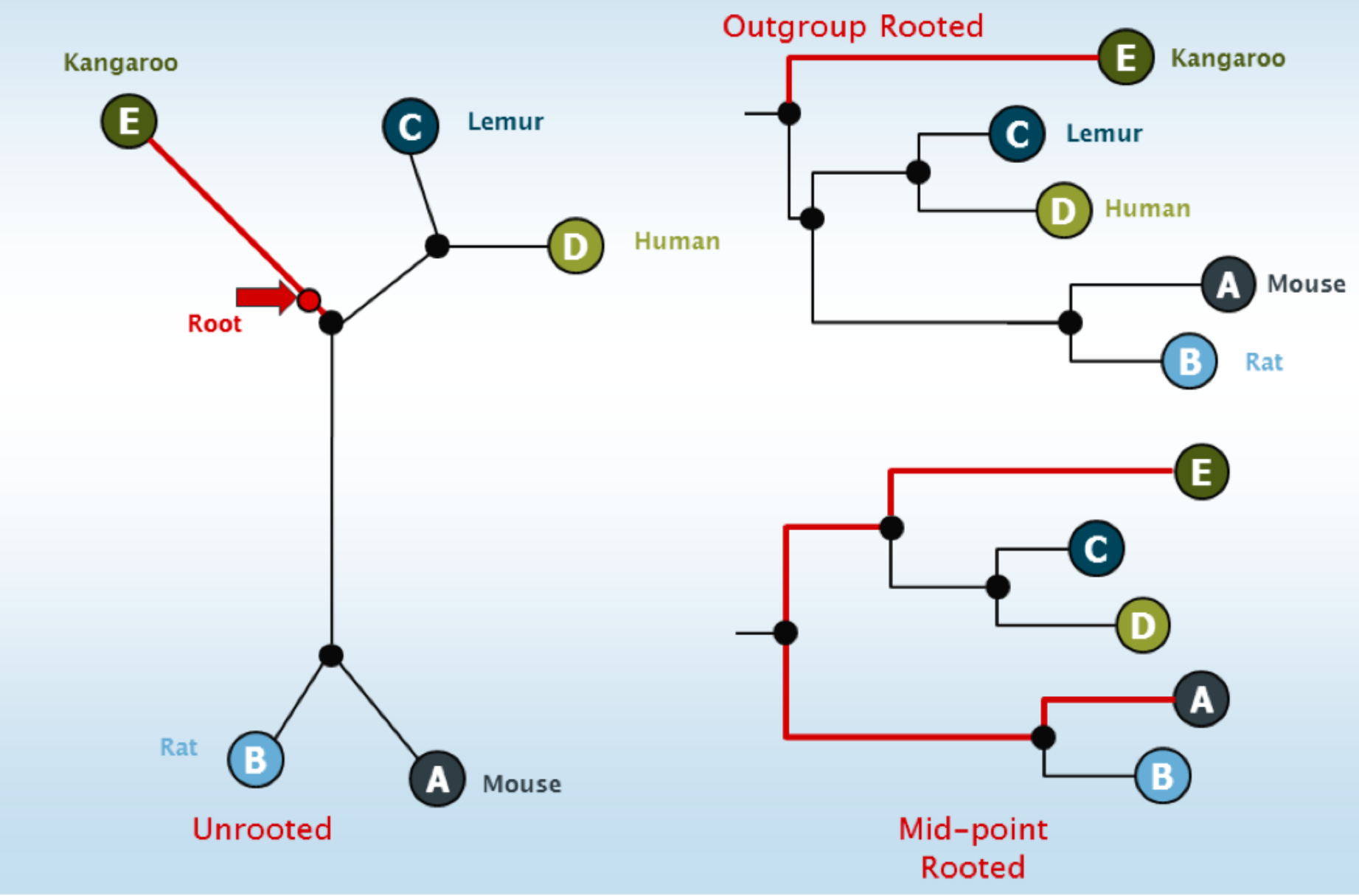


Root

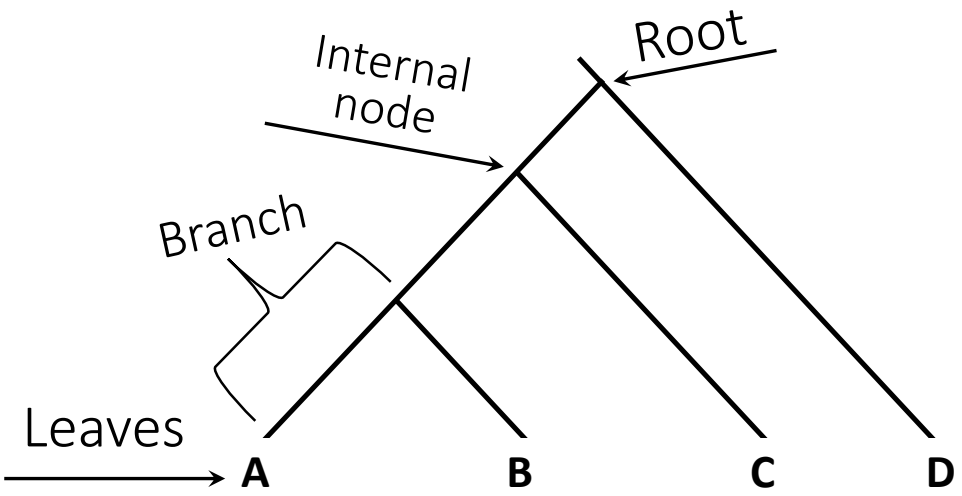
- Implies what happened first in chronological time
- Tree-building algorithms may not infer the root of a tree
  - Represented as a node with 3 outgoing branches where the root ought to be
- Could root halfway along the longest path
  - midpoint rooting
- Could root by prior knowledge of an outgroup
  - e.g. bacteria in the tree above



http://cabbagesofdoom.blogspot.com/2012/06/how-to-root-phylogenetic-tree.html



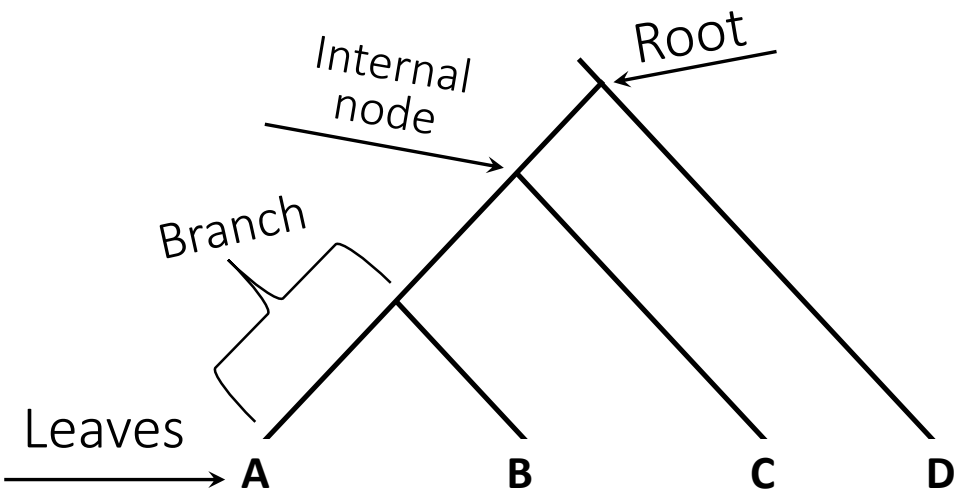
http://cabbagesofdoom.blogspot.com/2012/06/how-to-root-phylogenetic-tree.html



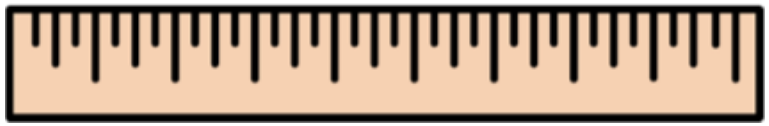
Internal node

- Represents some ancestral state
- Most recent common ancestor (MRCA) of the leaves following it

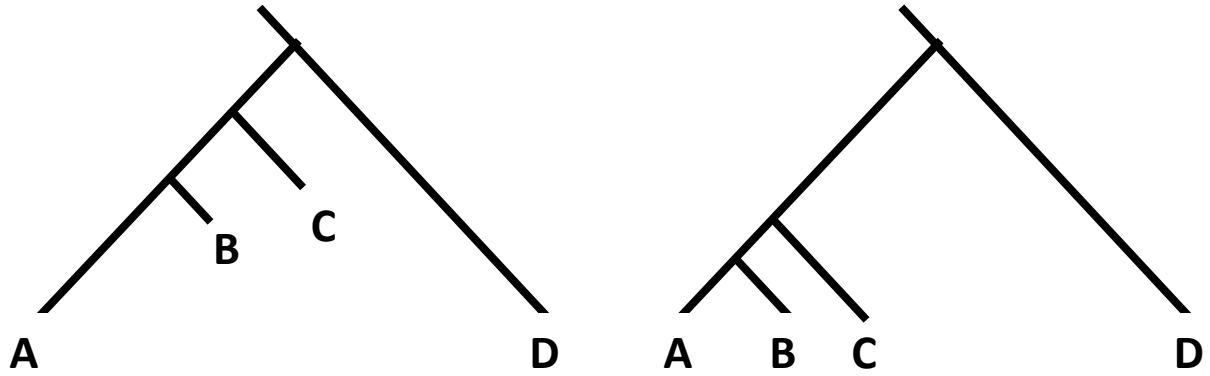
Internal nodes are implied to NOT be extant



Branch length



- Chronological time
- OR number of changes to the DNA / protein sequence
  - Depending on assumptions you make about rate of evolution when building tree, may not be ultrametric (every root-leaf path has same length)
- OR no distance implied – just defines a branching order



Why visualize biological information as trees?

(besides having a colorful Figure 1 for your Nature paper!)

Trees show fundamental structure in your data.

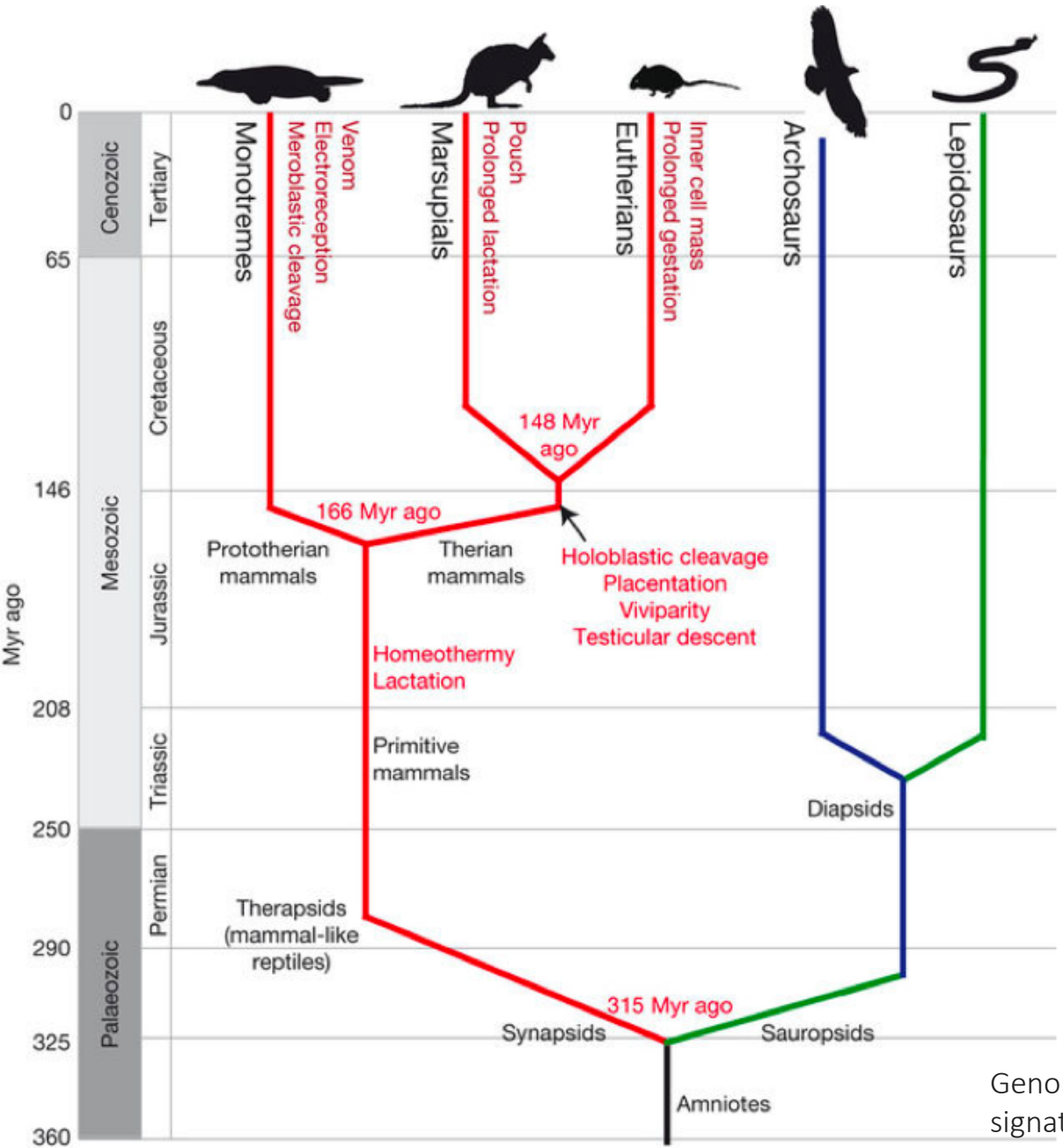
A picture's worth a thousand words!

The last amniote common ancestor of synapsids and sauropsids is placed 315 MYA. The sauropsid lineage evolved into archosaurs (such as birds, crocodilians, and dinosaurs) and lepidosaurs (such as snakes and lizards), diverging in the Triassic period. In the synapsid lineage, these primitive mammals acquired homothermy and lactation before the divergence of protetherian and therian mammals 166 MYA in the Jurassic period.

While descendant species of the prototherian mammals - extant monotremes such as the platypus *Ornithorhynchus anatinus* - have characteristics such as venom, electroreception, and meroblastic cleavage, therian mammals evolved holoblastic cleavage, placentation, viviparity, and testicular descent before their divergence 148 MYA into marsupials and eutherians. Further diversification leads us to observe a pouch and prolonged lactation in extant marsupials, and inner cell mass and prolonged gestation in eutherians.

Text composed by me for illustration only

species phylogeny

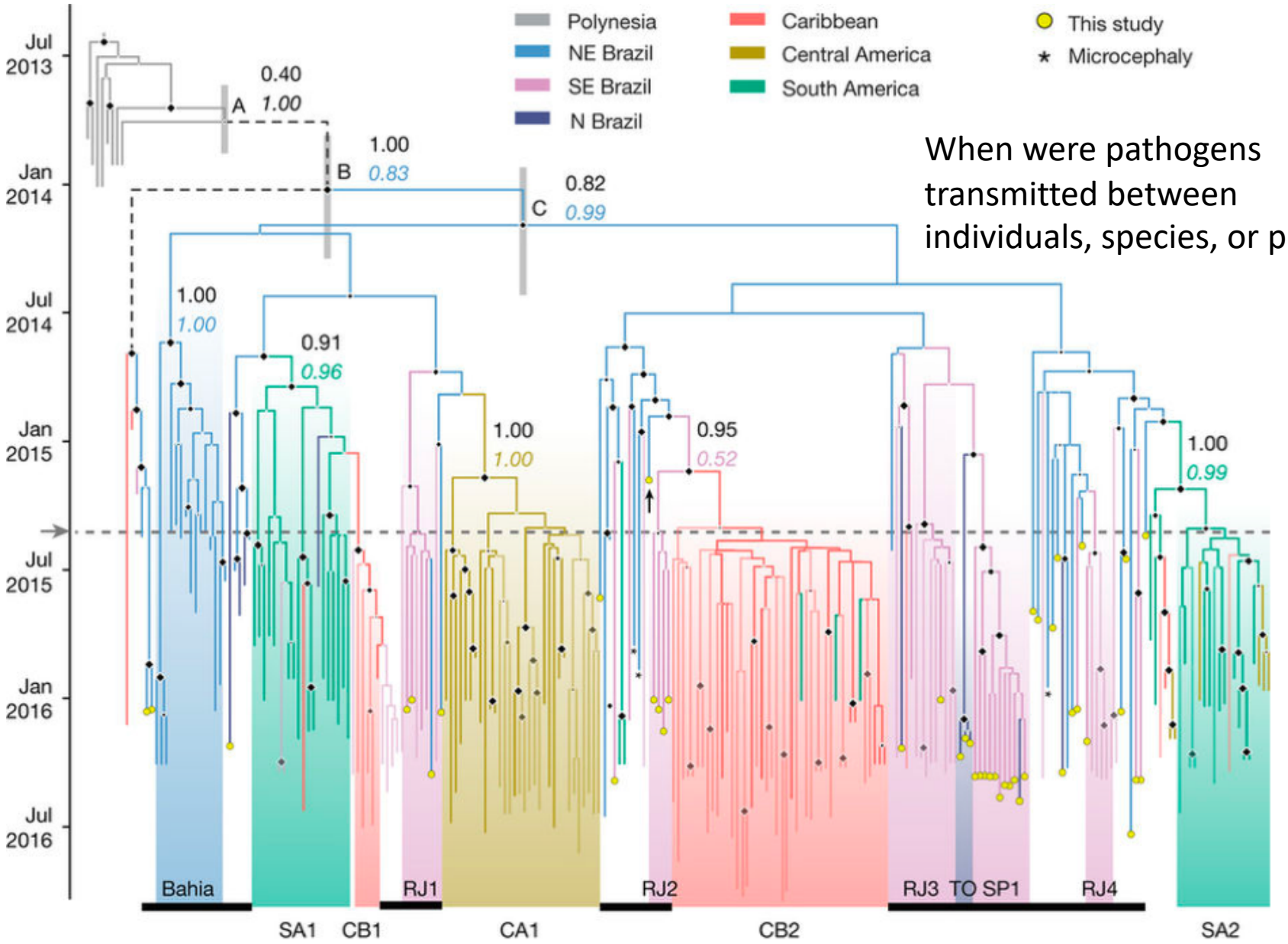


When did species diverge?

What characteristics define extant (surviving) species, and when in evolution did they arise?

Genome analysis of the platypus reveals unique signatures of evolution. Warren et al., Nature 2008

strain phylogeny

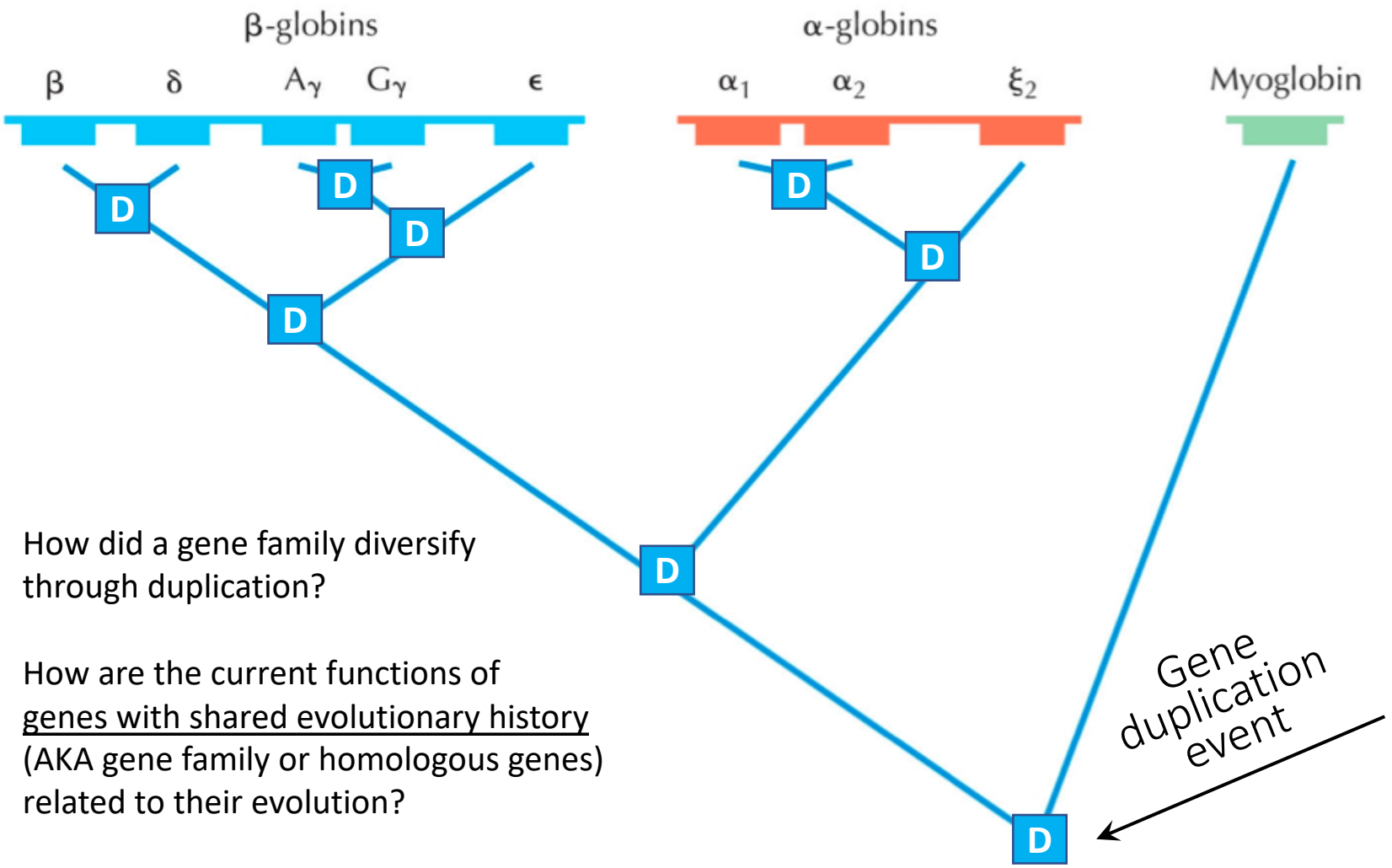


When were pathogens transmitted between individuals, species, or places?

Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Faria et al., Nature 2017

Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Faria et al., Nature 2017

gene phylogeny



How did a gene family diversify through duplication?

How are the current functions of <u>genes with shared evolutionary history</u> (AKA gene family or homologous genes) related to their evolution?

http://evolution-textbook.org/content/free/figures/27_EVOW_Art/29_EVOW_CH27.jpg

(((A,B),C),(D,E))

"The Newick Standard was adopted 26 June 1986 by an informal committee meeting convened by me [Joe Felsenstein] during the Society for the Study of Evolution meetings in Durham, New Hampshire and consisting of James Archie, William H.E. Day, Wayne Maddison, Christopher Meacham, F. James Rohlf, David Swofford, and myself. (The committee was not an activity of the SSE nor endorsed by it). The reason for the name is that the second and final session of the committee met at Newick's restaurant in Dover, New Hampshire, and we enjoyed the meal of lobsters."

```
((raccoon:19.19959,bear:6.80041):0.84600[50],((sea_lion:11.99700,
seal:12.00300):7.52973[100],((monkey:100.85930,cat:47.14069):20.59
201[80], weasel:18.87953):2.09460[75]):3.87382[50],dog:25.46154);

(((A:5,B:5)f:3,C:9)g:1,(D:2,E:6)h:2);
```

http://evolution.genetics.washington.edu/phylip/newicktree.html

Demo: http://tree.bio.ed.ac.uk/software/figtree/

Building Trees

You need a principled way of choosing the tree that best represents the relationship among leaves

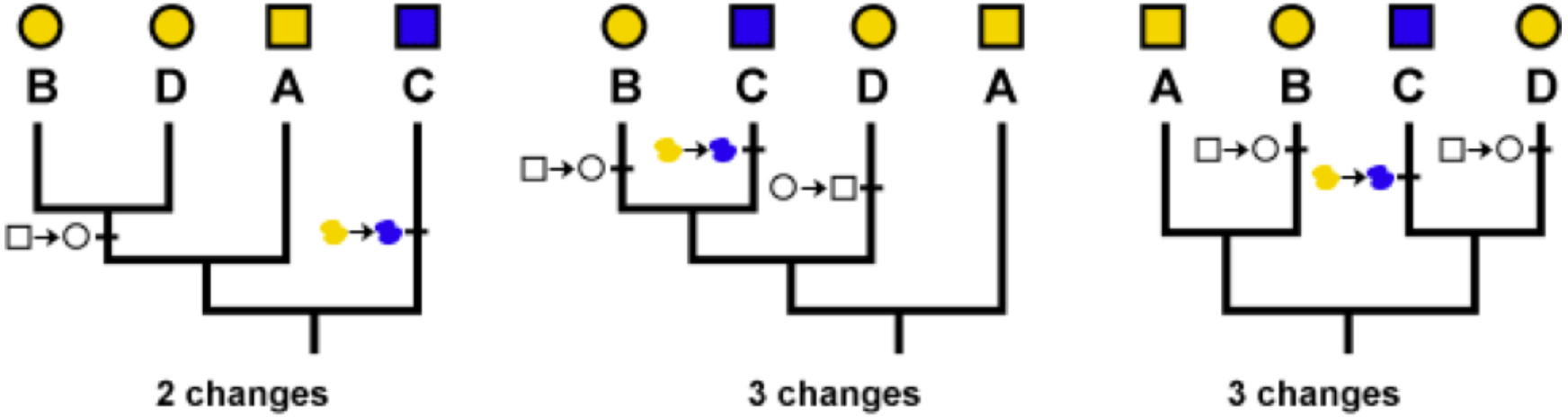
There are (2n-3)!! rooted trees for n taxa

- 3 taxa: 3 trees
- 5 taxa: 105 trees
- 10 taxa: 34,459,425 trees
- 20 taxa: 8.2×10^{21} trees
- 50 taxa: 2.8×10^{76} trees
- 100 taxa: 3.3×10^{184} trees

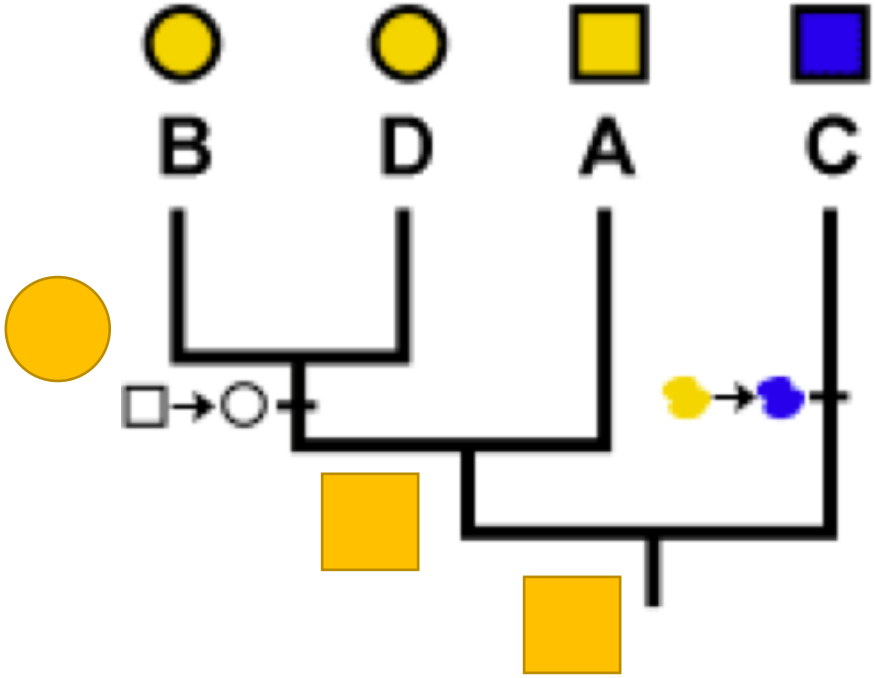
The **double factorial**, symbolized by two exclamation marks (!!), is a quantity defined for all integers greater than or equal to -1. For an even integer n , the **double factorial** is the product of all even integers less than or equal to n but greater than or equal to 2.

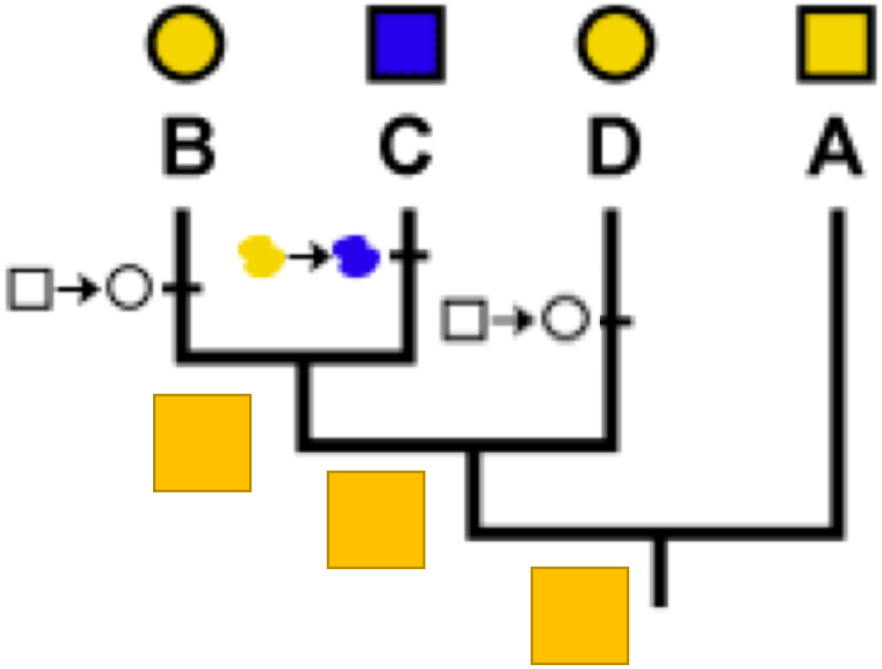
Parsimony: scenario that requires the fewest changes is the best

taxon	characters	
	shape	color
A	□	
B	○	
C	□	
D	○	



https://evolution.berkeley.edu/evolibrary/article/0_0_0/evotrees_build_07


B
D
A
C


B
C
D
A

Clustering a distance matrix

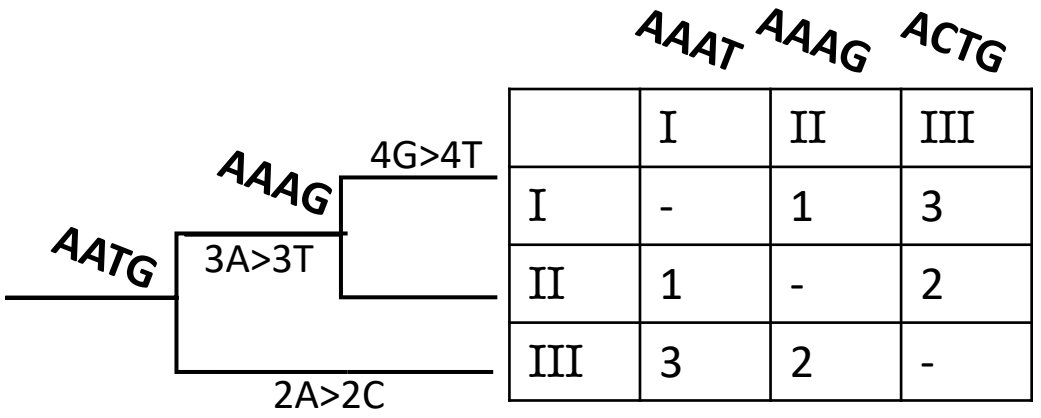
- Calculate all pairwise distances
  - Use a mathematical model of protein / nucleotide evolution to calculate the distance between two sequences

Here: Hamming distance

Example algorithms:

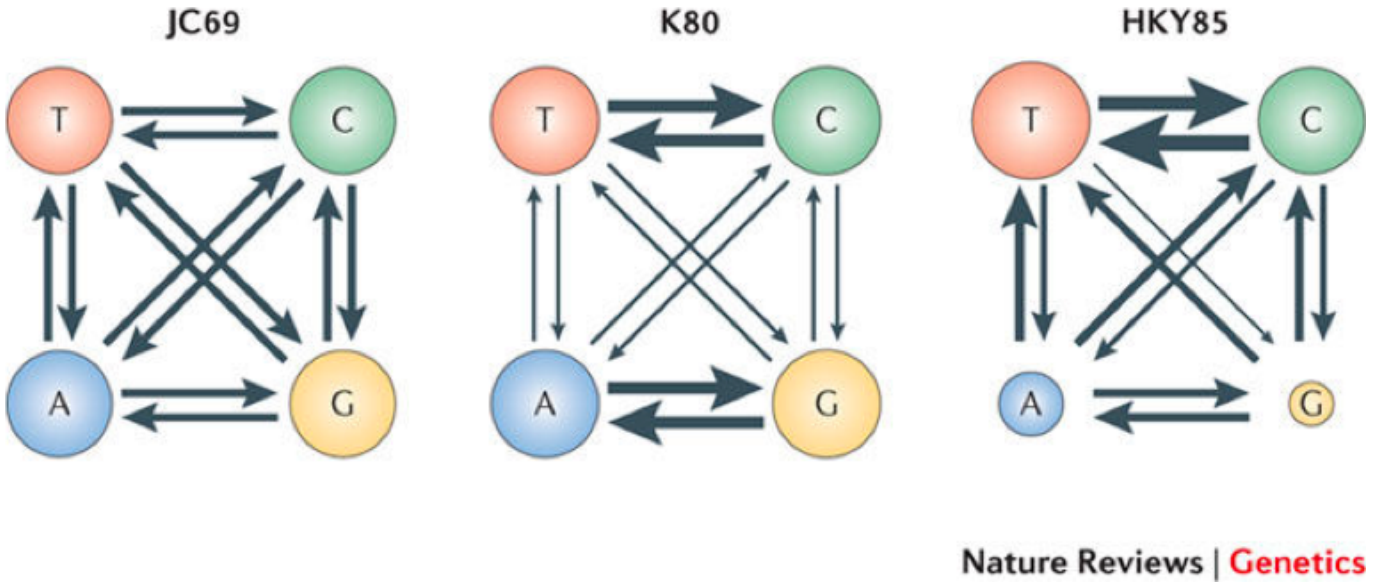
- Least squares
- Neighbor joining
- Minimum evolution
- UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

Software: PHYLIP



	I	II	III
I	-	1	3
II	1	-	2
III	3	2	-

Estimating distance between sequences: nucleotide evolution models



"The thickness of the arrows indicates the substitution rates of the four nucleotides (T, C, A and G), and the sizes of the circles represent the nucleotide frequencies when the substitution process is in equilibrium. Note that both JC69 and K80 predict equal proportions of the four nucleotides."

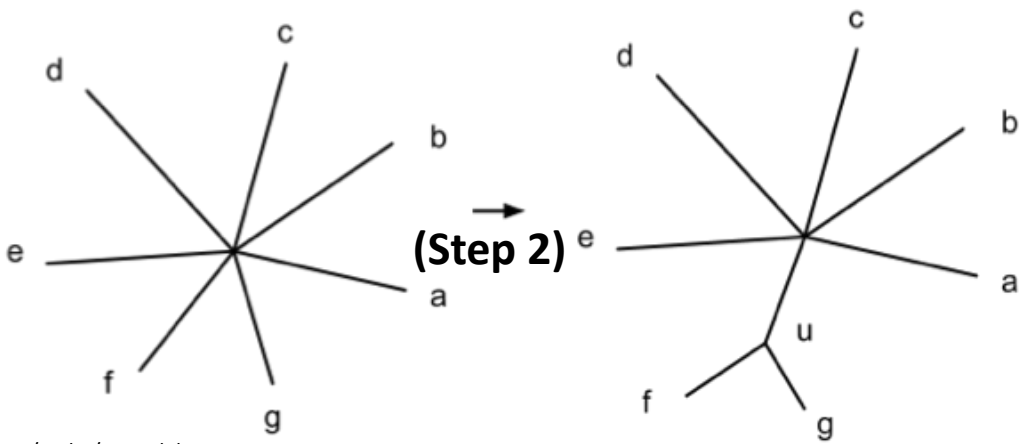
Molecular phylogenetics: principles and practice.
Yang and Rannala, Nature Reviews Genetics 2012

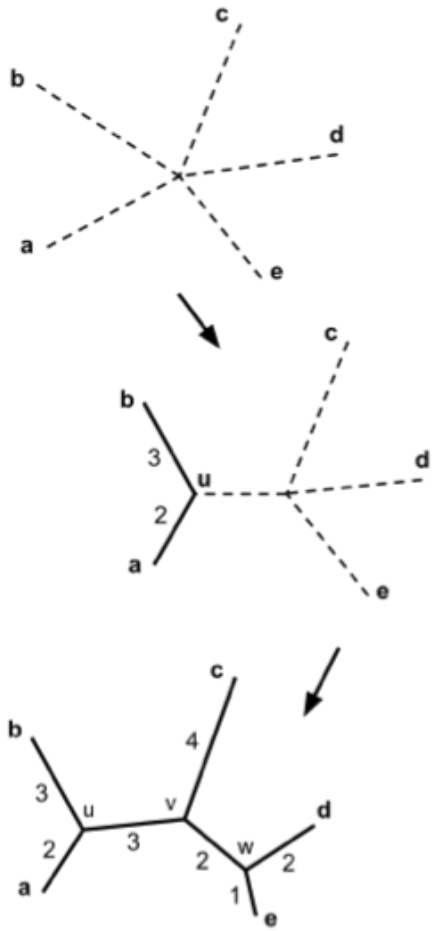
Clustering: Neighbor joining

0. Start with a star graph
1. Calculate distance matrix
2. Find the minimum entry (closest leaves) and join those leaves with a new internal node, with branch lengths based on the values in the matrix
3. Replace the rows/columns of the leaves with a single row/column of the new internal node
4. Recalculate distance matrix

Repeat 2,3,4 until you have an unrooted binary tree

*Resulting tree is unrooted and not guaranteed to be ultrametric





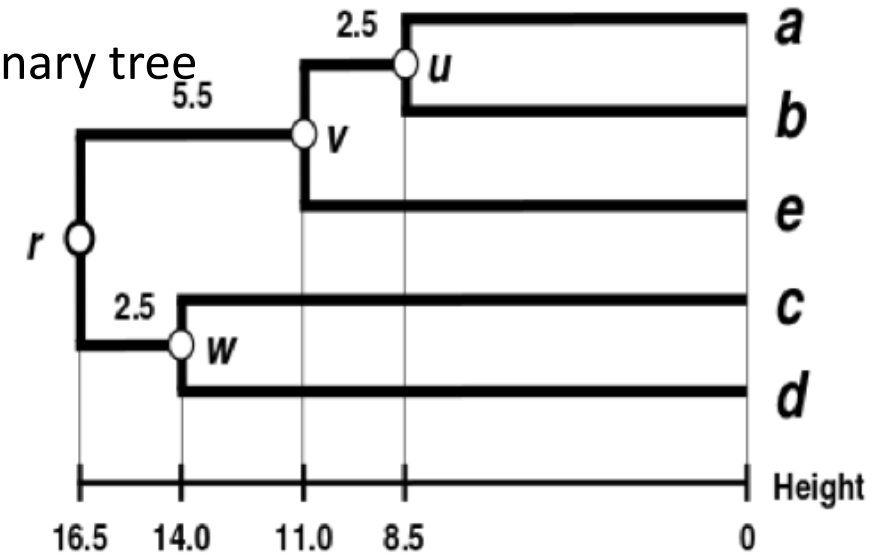
https://en.wikipedia.org/wiki/Neighbor_joining

Clustering: UPGMA

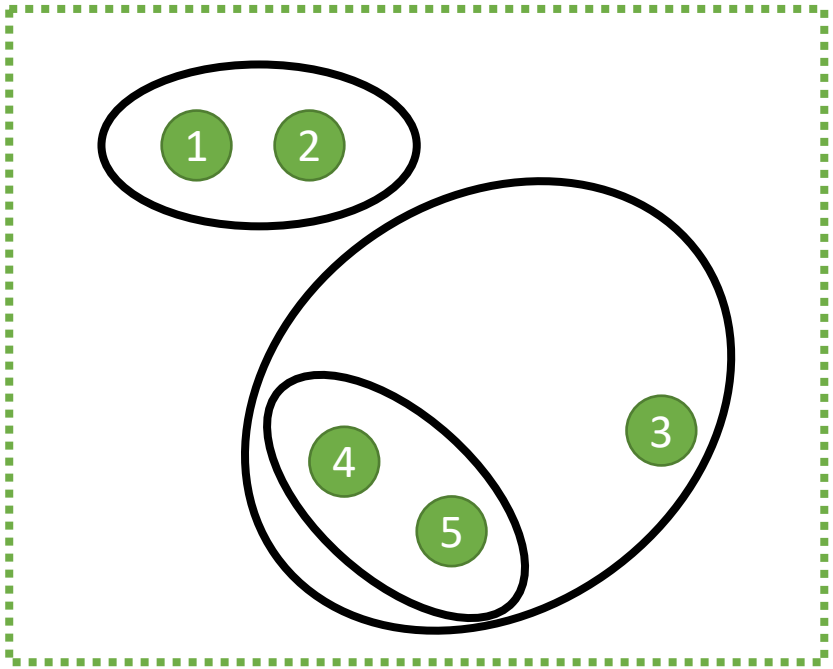
0. Start with a star graph
1. Calculate distance matrix
2. Find the minimum entry (closest leaves) and join those leaves at a distance of the mean distance between their elements
3. Replace the rows/columns of the leaves with a single row/column of their union (cluster)
4. Recalculate distance matrix

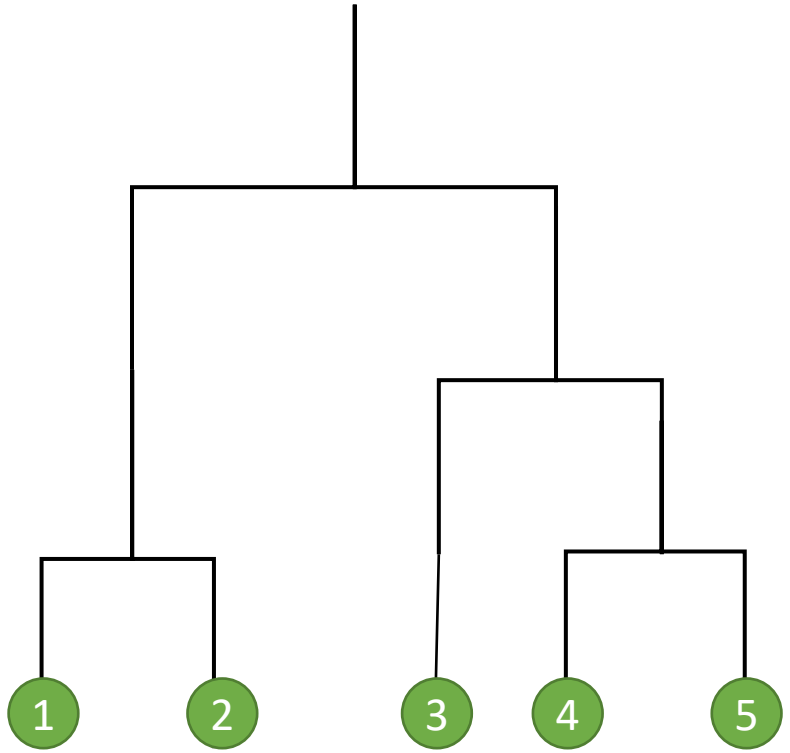
Repeat 2,3,4 until you have an rooted binary tree

*Resulting tree *is* rooted and ultrametric: assumes rate of evolution is constant on all branches (molecular clock)

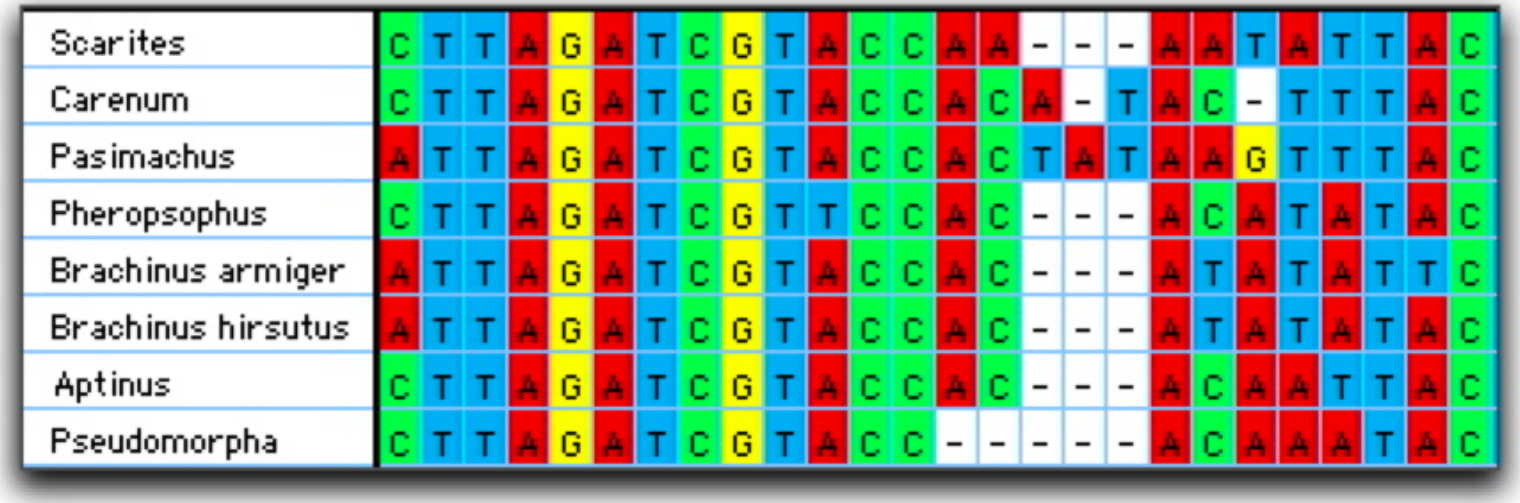


https://en.wikipedia.org/wiki/UPGMA


1
2
3
4
5


1
2
3
4
5

Multiple sequence alignment



http://www.sequence-alignment.com/sequence-alignment.jpg

Maximum likelihood estimation from an MSA

0. Start with a multiple sequence alignment
1. Build a starting tree
2. Calculate the likelihood of the tree based on an evolutionary model
3. Change the tree slightly to increase the likelihood

Repeat 2,3 until some convergence criterion met

ML is a heuristic search of “tree space”

Bootstrapping: resampling from “tree space” to gauge the quality of your solution

- Proportion of resamples that support a certain branching pattern

Software:

PhyML/RAxML

Genetic algorithm - GARLI

Bayesian estimation from an MSA

0. Start with a multiple sequence alignment
1. Build a starting tree and assume some prior probability on the distribution of trees
2. Calculate the posterior probability of the tree based on an evolutionary model
3. Sample a ”nearby” tree, usually using MCMC (Markov chain Monte Carlo) algorithms
4. Accept the new tree if it is “better”, otherwise keep with the old tree

Repeat 2,3,4 until some convergence criterion met

Bayesian methods are also a heuristic search of “tree space”

The posterior probability (probability correct *given model*) gauges the quality of your solution

Software: BEAST, MrBayes

Table 2 | **A summary of strengths and weaknesses of different tree reconstruction methods**

Strengths	Weaknesses
Parsimony methods	
• Simplicity and intuitive appeal • The only framework appropriate for some data (such as SINES and LINES)	• Assumptions are implicit and poorly understood • Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution • Branch lengths are substantially underestimated when substitution rates are high • Maximum parsimony may suffer from long-branch attraction
Distance methods	
• Fast computational speed • Can be applied to any type of data as long as a genetic distance can be defined • Models for distance calculation can be chosen to fit data	• Most distance methods, such as neighbour joining, do not consider variances of distance estimates • Distance calculation is problematic when sequences are divergent and involve many alignment gaps • Negative branch lengths are not meaningful
Likelihood methods	
• Can use complex substitution models to approach biological reality • Powerful framework for estimating parameters and testing hypotheses	• Maximum likelihood iteration involves heavy computation • The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret
Bayesian methods	
• Can use realistic substitution models, as in maximum likelihood • Prior probability allows the incorporation of information or expert knowledge • Posterior probabilities for trees and clades have easy interpretations	• Markov chain Monte Carlo (MCMC) involves heavy computation • In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify • Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge • Posterior probabilities often appear too high • Model selection involves challenging computation[138,139]

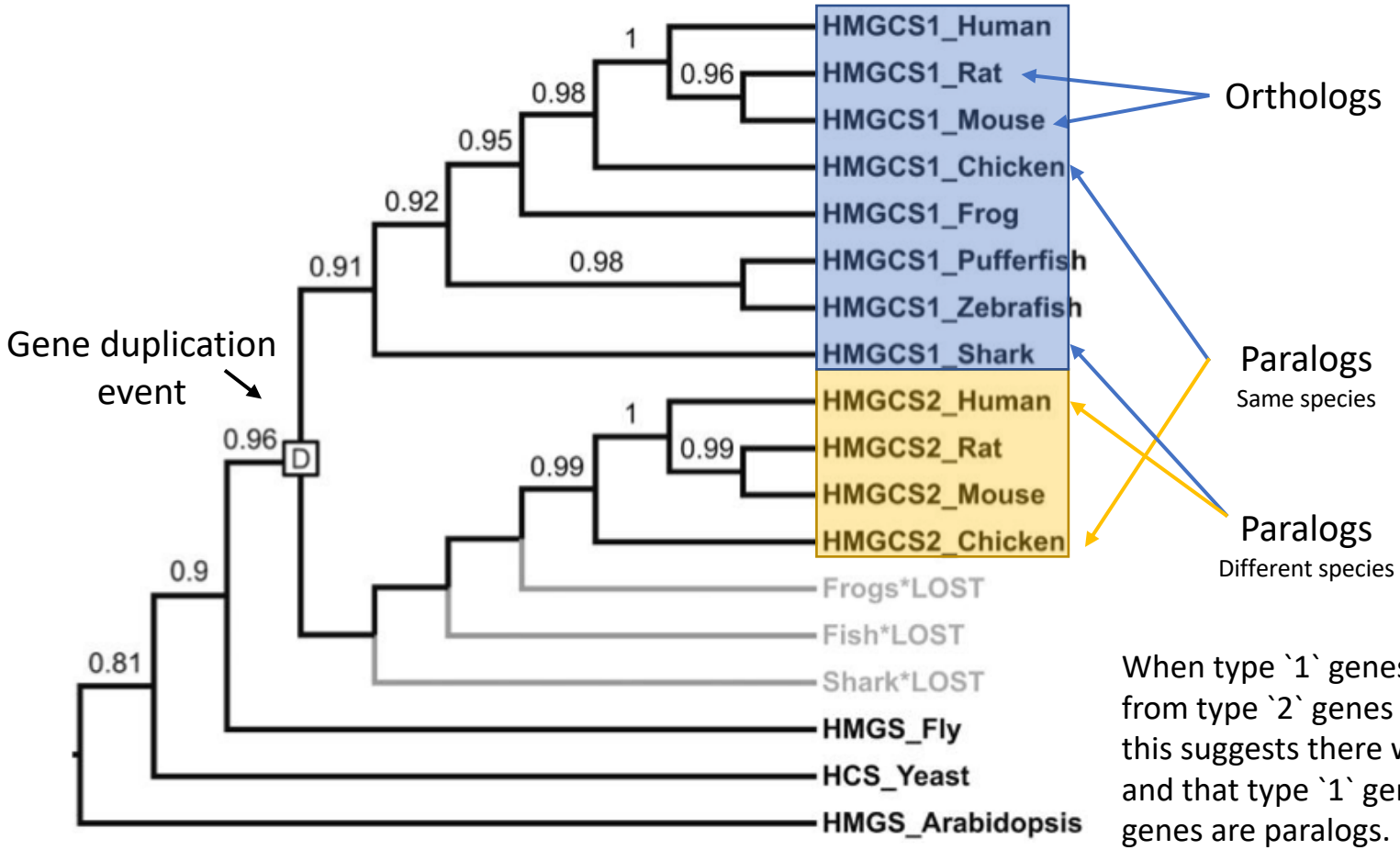
Molecular phylogenetics: principles and practice.
Yang and Rannala, Nature Reviews Genetics 2012

3-hydroxy-3-methylglutaryl coenzyme A synthase (HMGCS) enzyme family

Homologs: Genes with shared evolutionary history

Orthologs: Genes that diverged through speciation (i.e. not duplication)

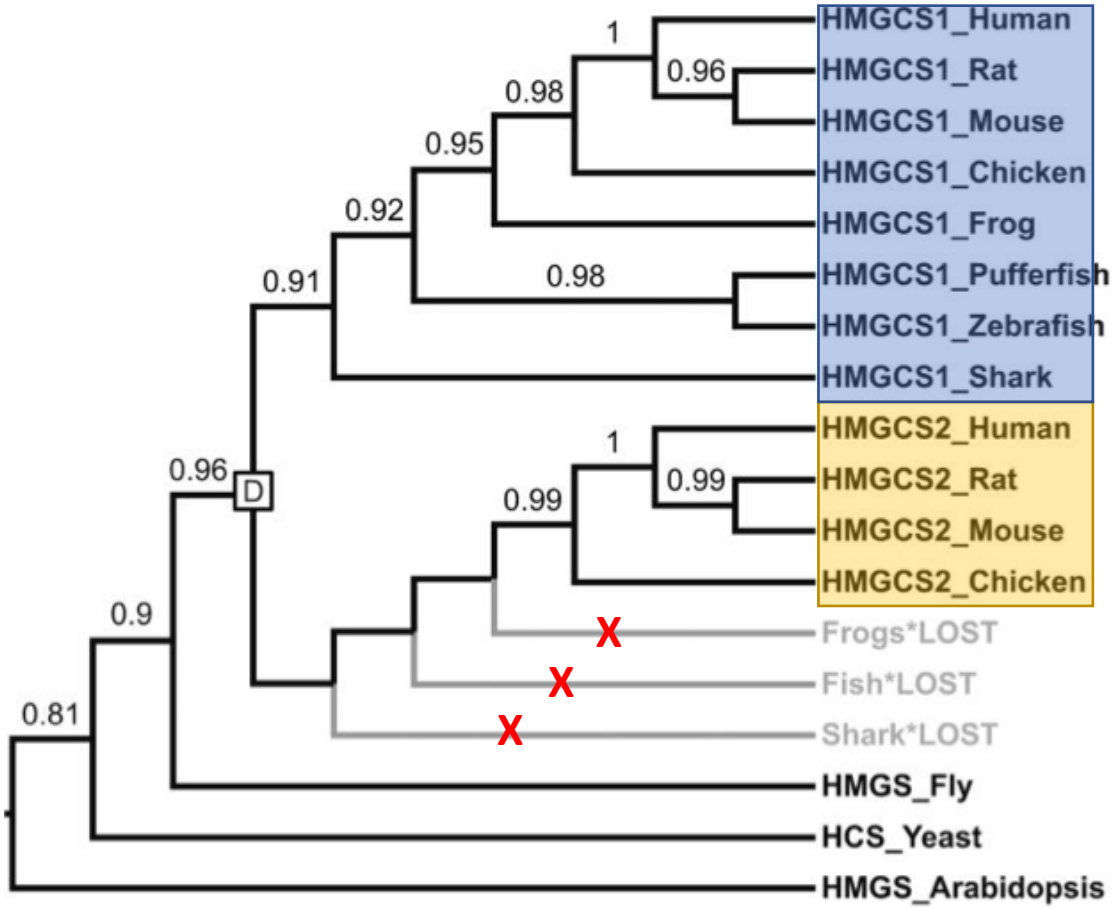
Paralogs: Genes that diverged through duplication



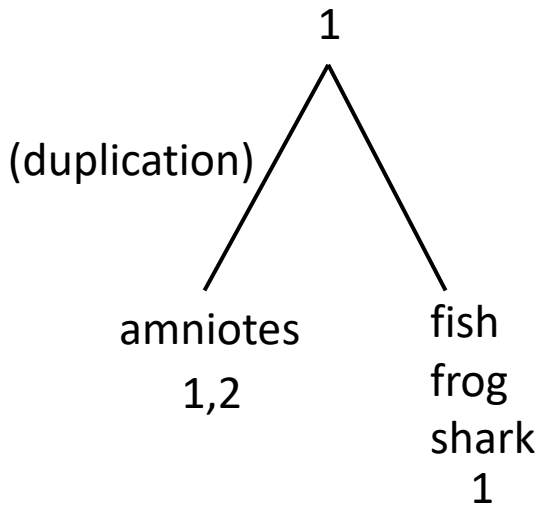
When type `1` genes cluster separately from type `2` genes in the gene tree, this suggests there was a duplication and that type `1` genes and type `2` genes are paralogs.

How old is my gene? Capra et al., Trends in Genetics 2013

3-hydroxy-3-methylglutaryl coenzyme A synthase (HMGCS) enzyme family

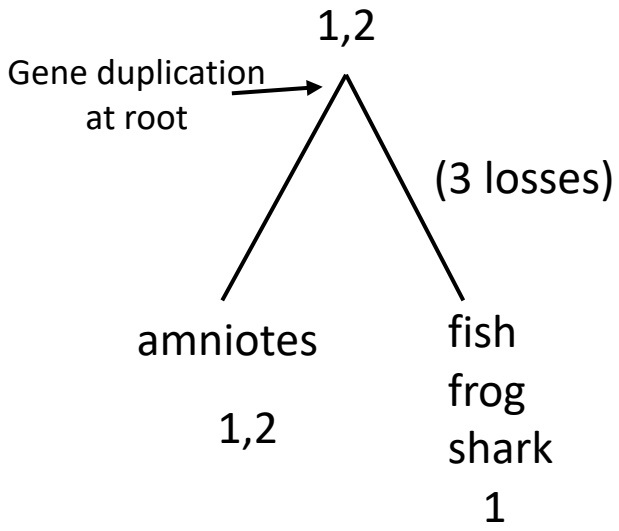


"Gene tree inferred using PhyML from sequences aligned with MAFFT and rooted using three invertebrate outgroup sequences. Branch support was assessed using aLRT scores."



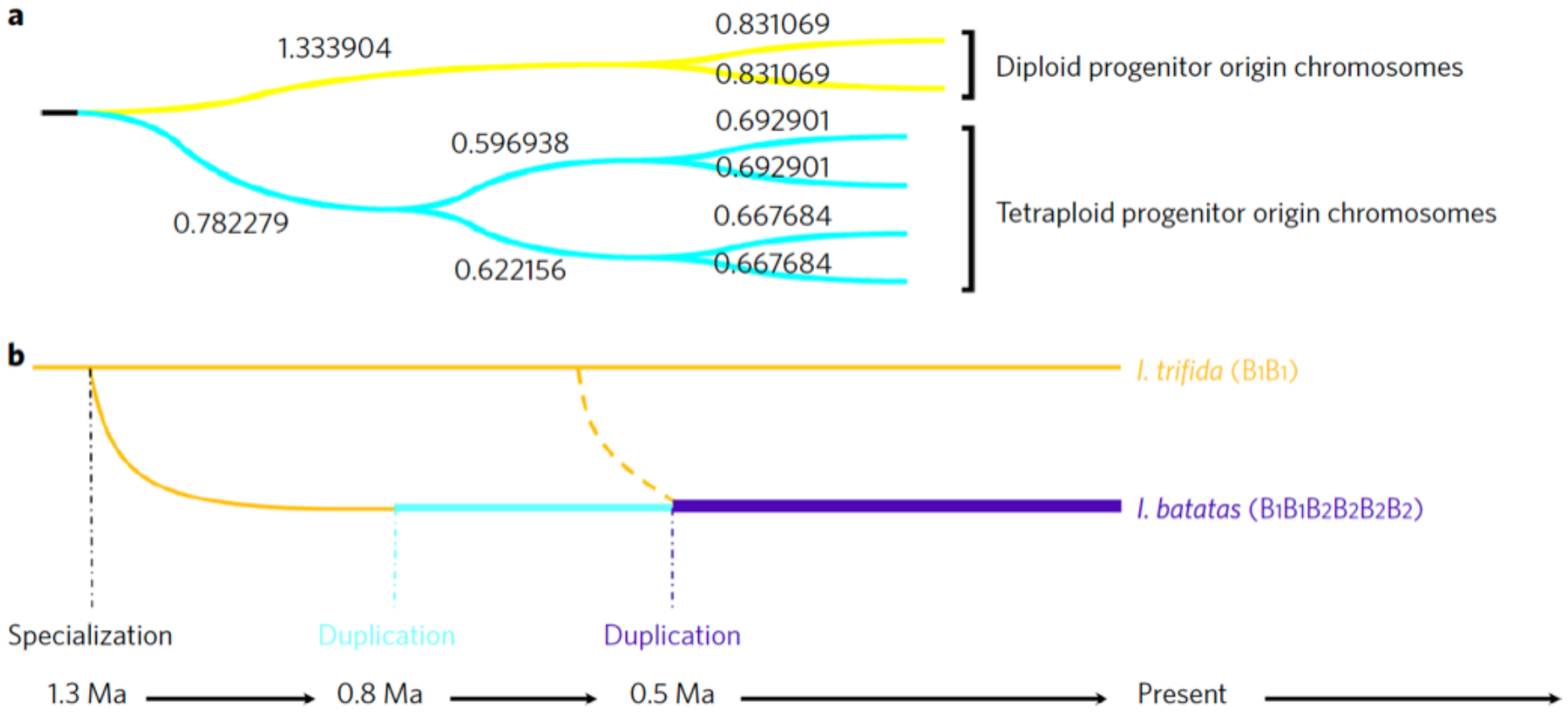
More parsimonious

Supported by gene tree evidence



How old is my gene? Capra et al., Trends in Genetics 2013

Evolution of chromosomes in hexaploid sweet potato

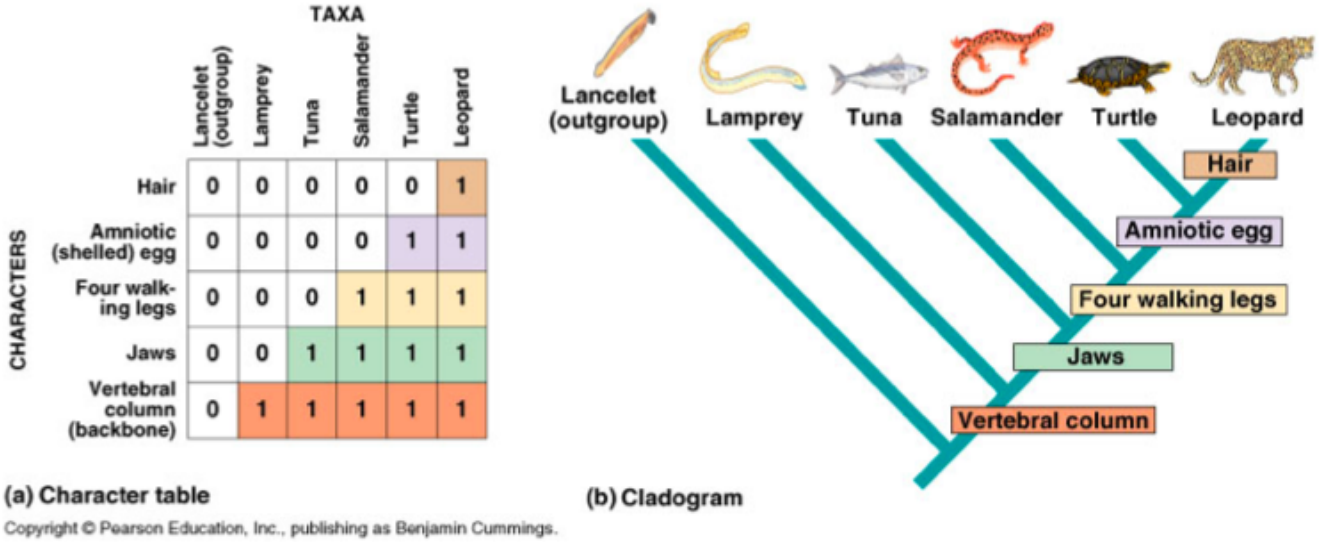


a) Evolutionary history of cultivated *I. batatas* revealed by phylogenetic analysis of homologous chromosome regions. a, The dominant topology structure of all phylogenetic trees. Numbers indicate the average branch length of trees in this structure.
b) The time points of B2 subgenome specialization and two whole-genome duplication events were estimated as 1.3, 0.8 and 0.5 million years ago (Ma). The estimation is based on 0.7% mutation rate per million years. The dashed curve indicates the crossing between diploid and tetraploid progenitors.

Haplotype-resolved sweet potato genome traces back its hexaploidization history.
Yang et. al, Nature Plants 2017

Building a species phylogeny

- Macroscopic characteristics

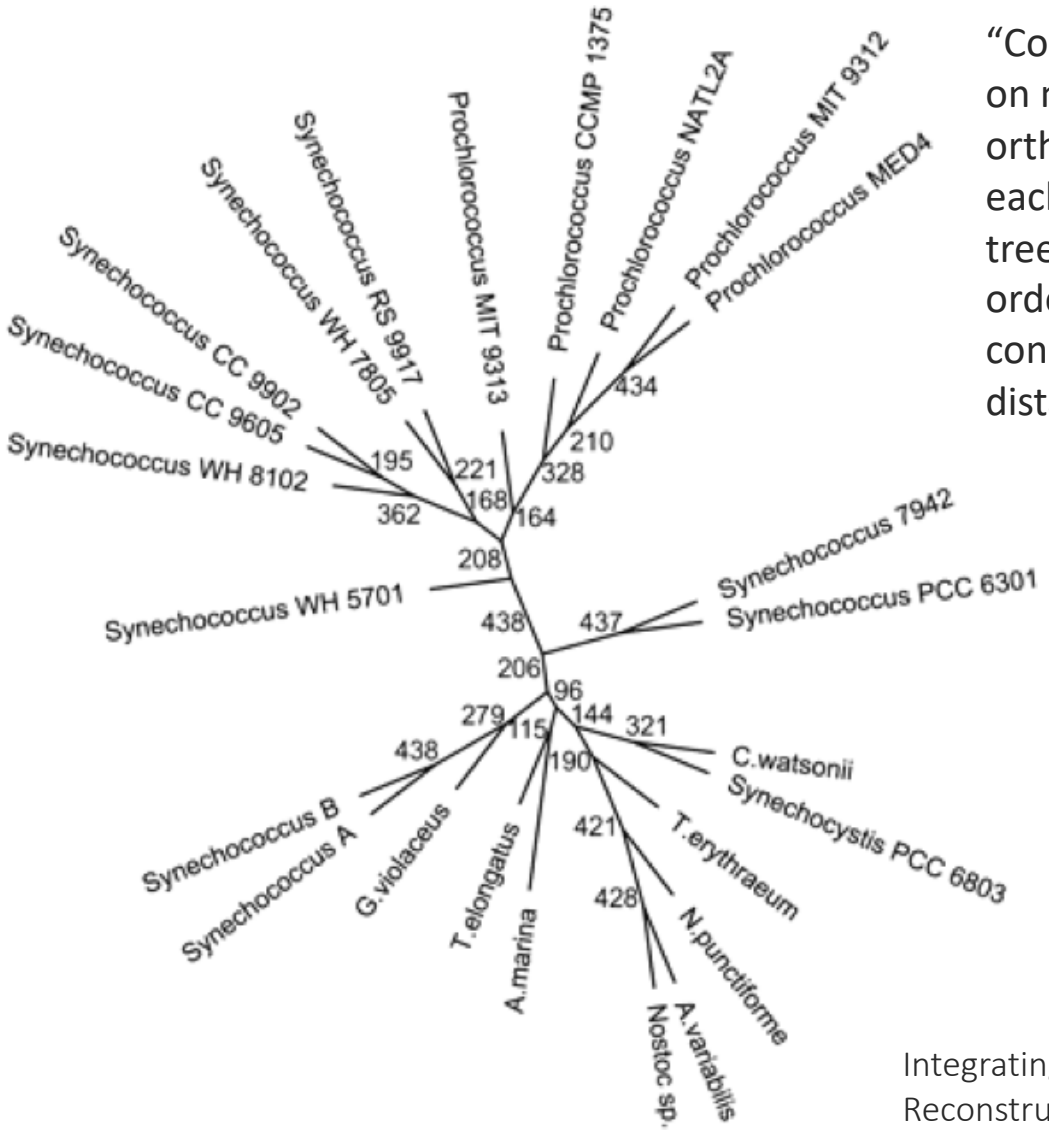


Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

- Molecular characteristics
  - Conserved sequence, e.g. ribosome
  - Concatenation of conserved genes
  - Consensus of many individual-gene trees (supertree)

http://www.bio.utexas.edu/faculty/sjasper/Bio213/phylogeny.html

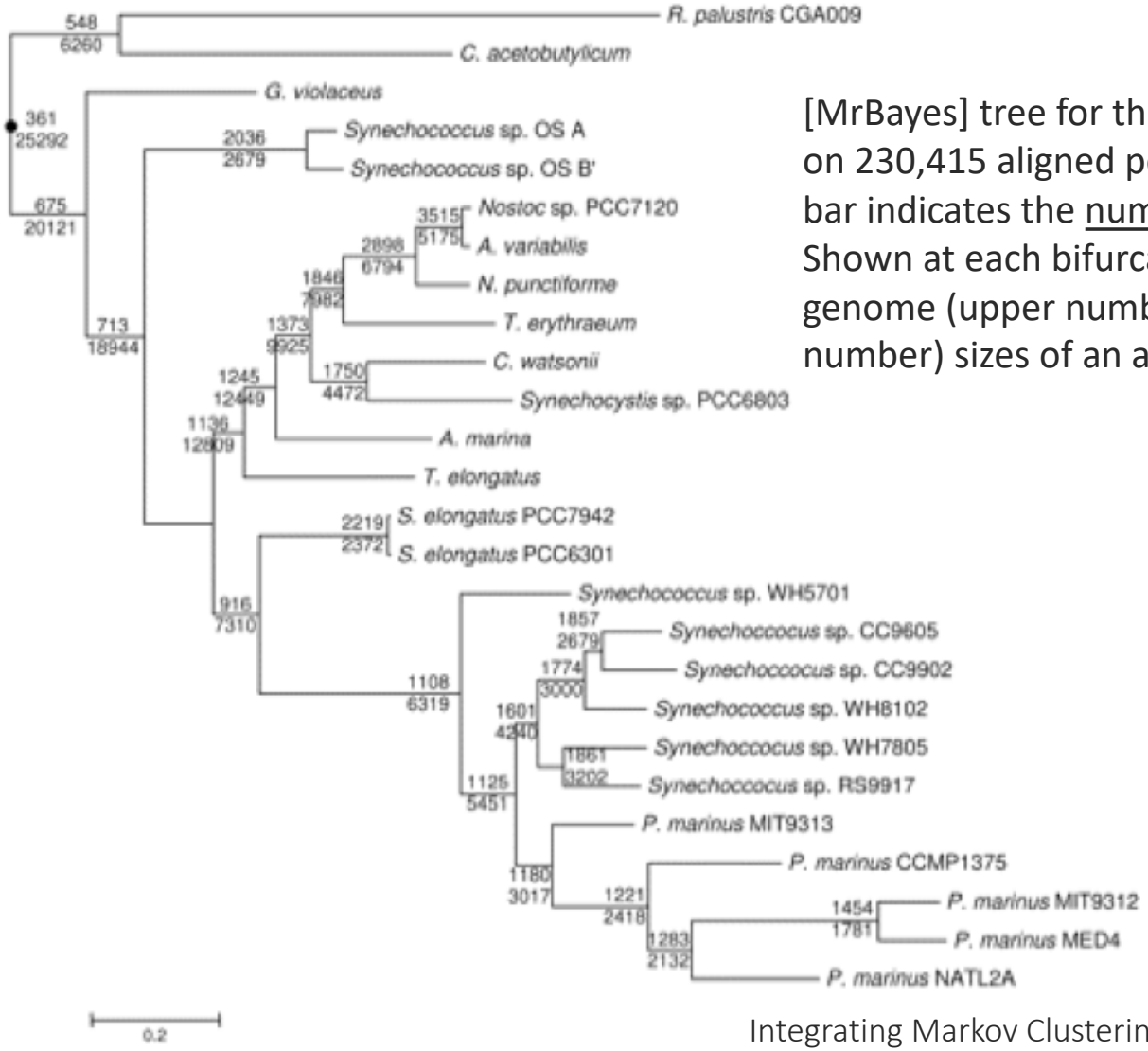
Species phylogeny of Cyanobacteria from consensus of many protein families



"Consensus cyanobacterial phylogeny based on maximum likelihood trees for each of 438 orthologous protein families. The numbers at each bifurcation indicate the total number of trees where that exact bifurcation/branching order is observed … Note that as this is a consensus tree, topology is meaningful but distances are not."

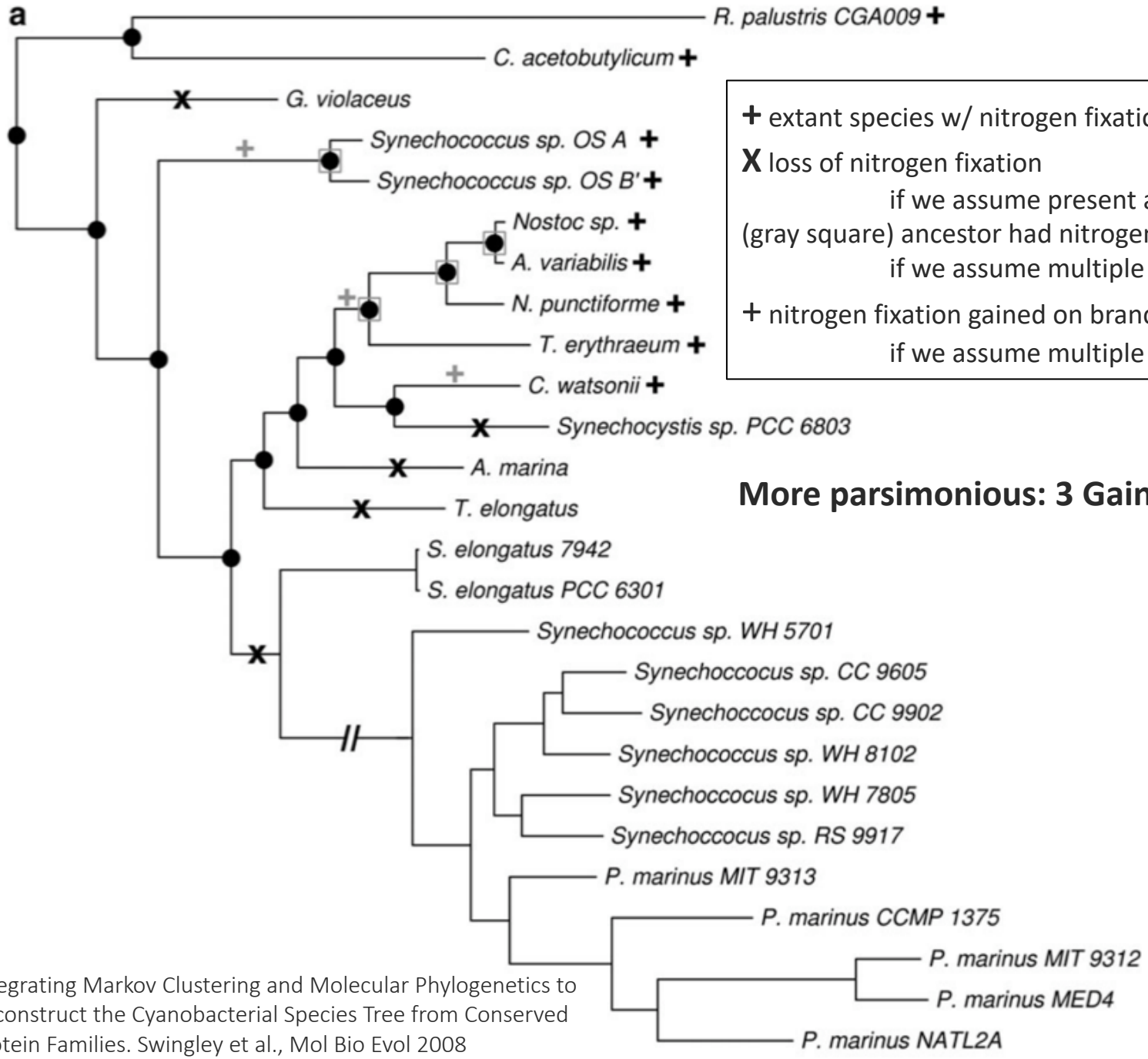
Integrating Markov Clustering and Molecular Phylogenetics to Reconstruct the Cyanobacterial Species Tree from Conserved Protein Families. Swingley et al., Mol Bio Evol 2008

Species phylogeny of Cyanobacteria from concatenation of many multiple sequence alignments



[MrBayes] tree for the full-concatenated data set, based on 230,415 aligned positions in 26 genomes. The scale bar indicates the number of substitutions per site. Shown at each bifurcation are the predicted core-genome (upper number) and pan-genome (lower number) sizes of an ancestor at that point.

Integrating Markov Clustering and Molecular Phylogenetics to Reconstruct the Cyanobacterial Species Tree from Conserved Protein Families. Swingley et al., Mol Bio Evol 2008



+ extant species w/ nitrogen fixation

X loss of nitrogen fixation
if we assume present at root and parallel loss (gray square) ancestor had nitrogen fixation
if we assume multiple gains

+ nitrogen fixation gained on branch
if we assume multiple gains

More parsimonious: 3 Gains or 5 Losses?

Integrating Markov Clustering and Molecular Phylogenetics to Reconstruct the Cyanobacterial Species Tree from Conserved Protein Families. Swingley et al., Mol Bio Evol 2008

Resources

- Review article
  - Molecular phylogenetics: principles and practice. Yang and Rannala, Nature Reviews Genetics 2012
- Opinion piece
  - Homology: a personal view on some of the problems. Fitch, Trends In Genetics 2000
- Textbooks
  - *Molecular Evolution: A Phylogenetic Approach*. Page and Holmes
  - *Inferring Phylogenies.* Felsenstein

My work in evolutionary biology (horizontal gene transfer)

Xenolog classification. Darby, Stolzer et. al., Bioinformatics 2017