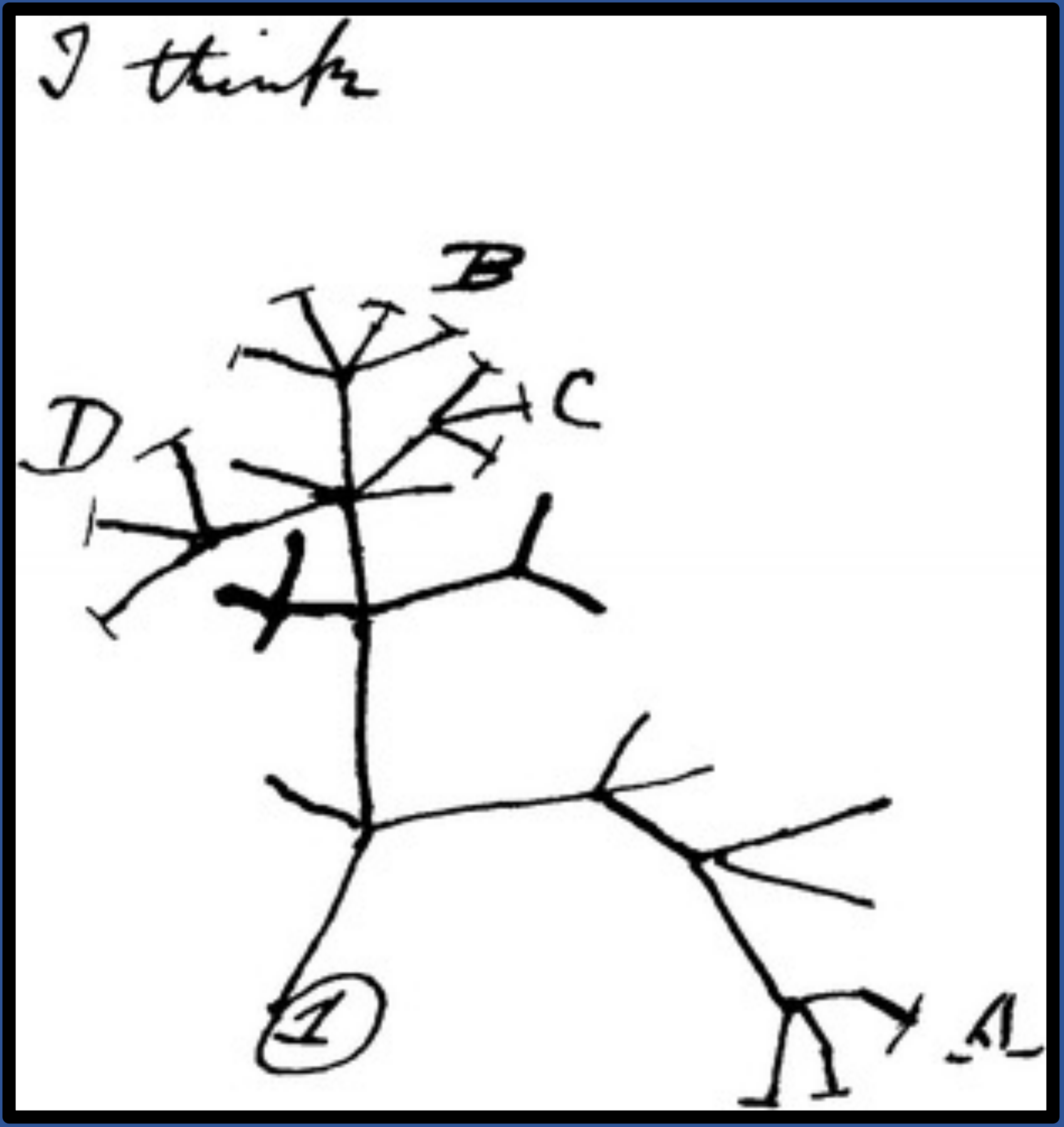
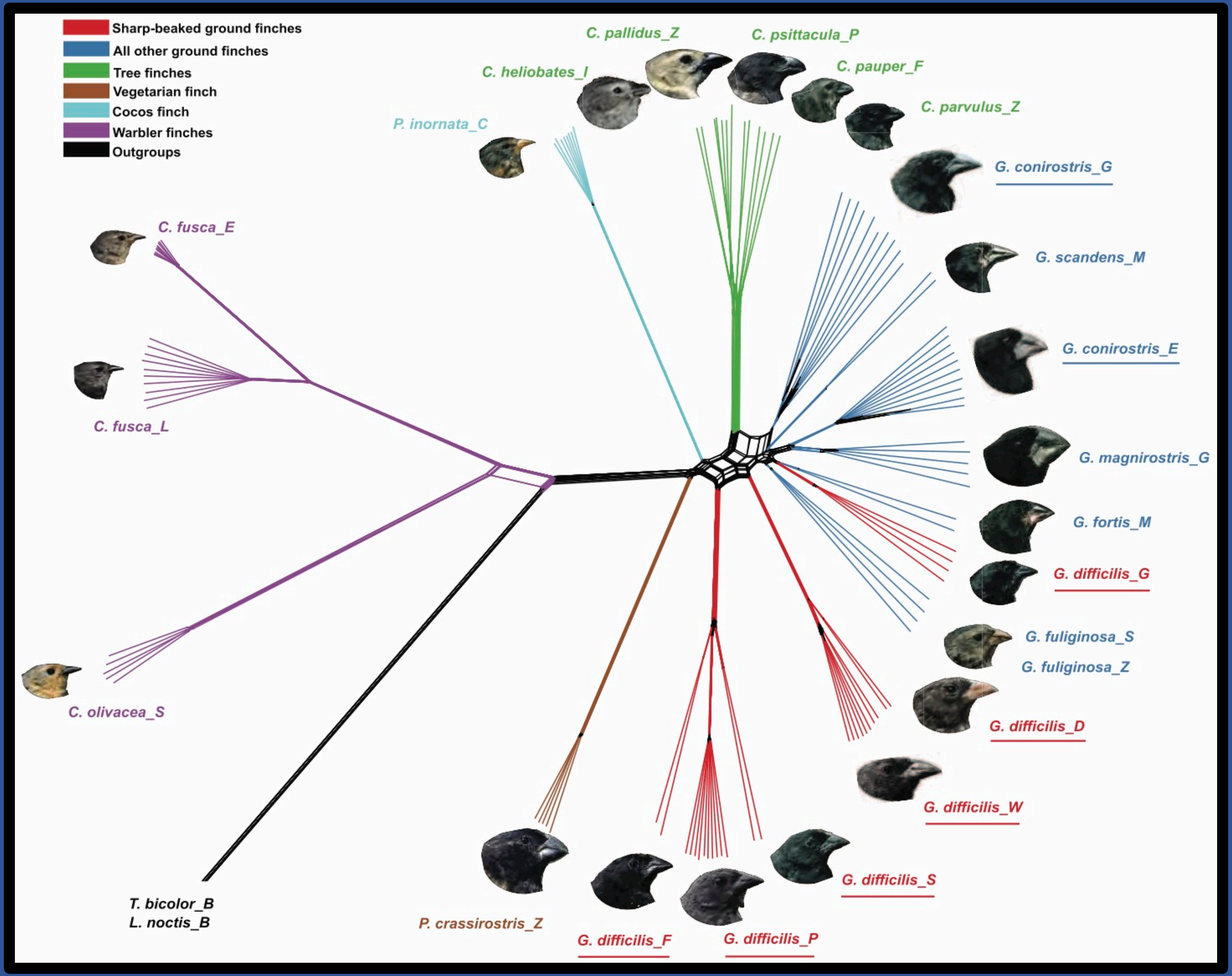
Phylogenetics



Erick Gagne

Phylogenetics

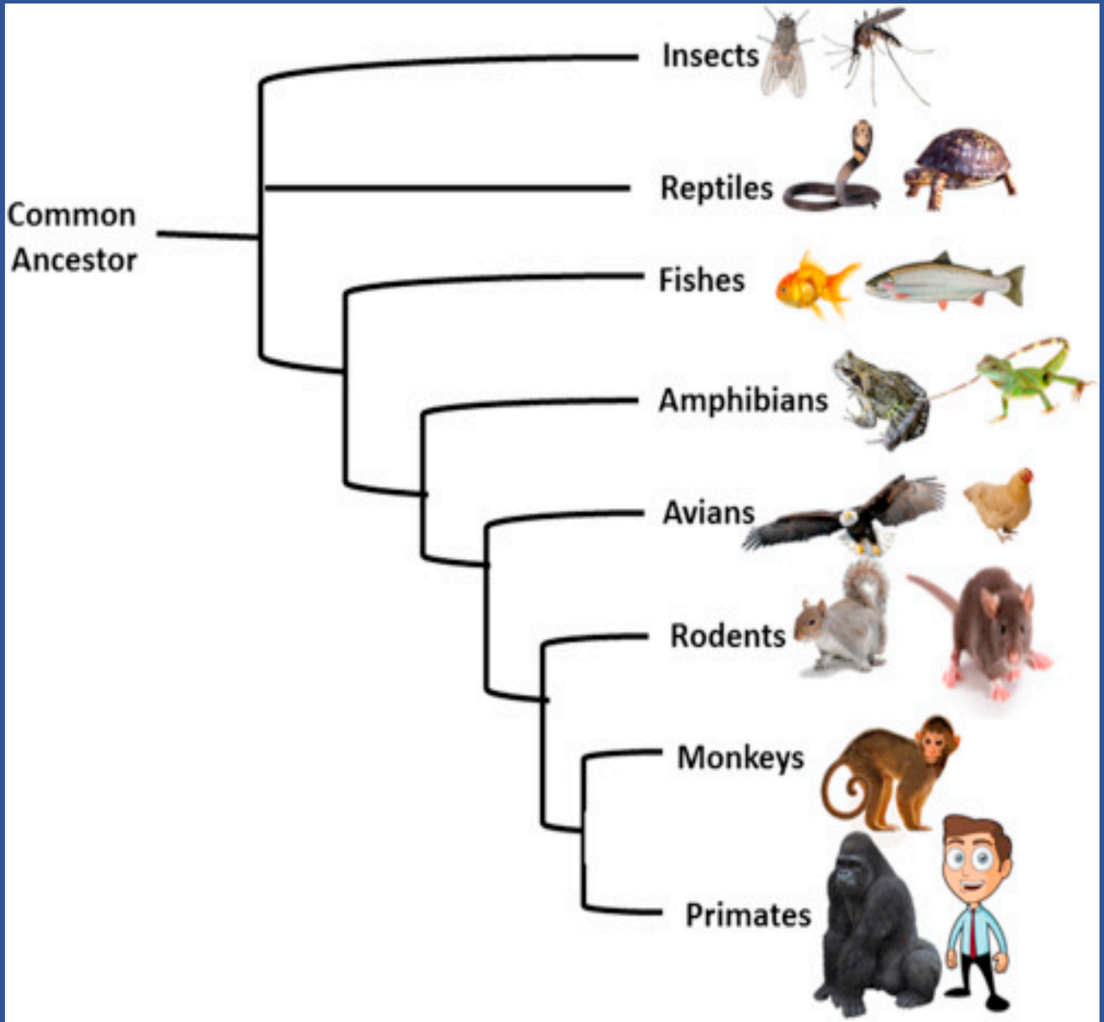


Sallman et al. 2015, BioEssays

Phylogenetics – the basics

- Generate trees using genetic sequence data
    - Reconstructing the ancestral relationships among taxa

- A tree is only an estimate the “truth” usually unknown



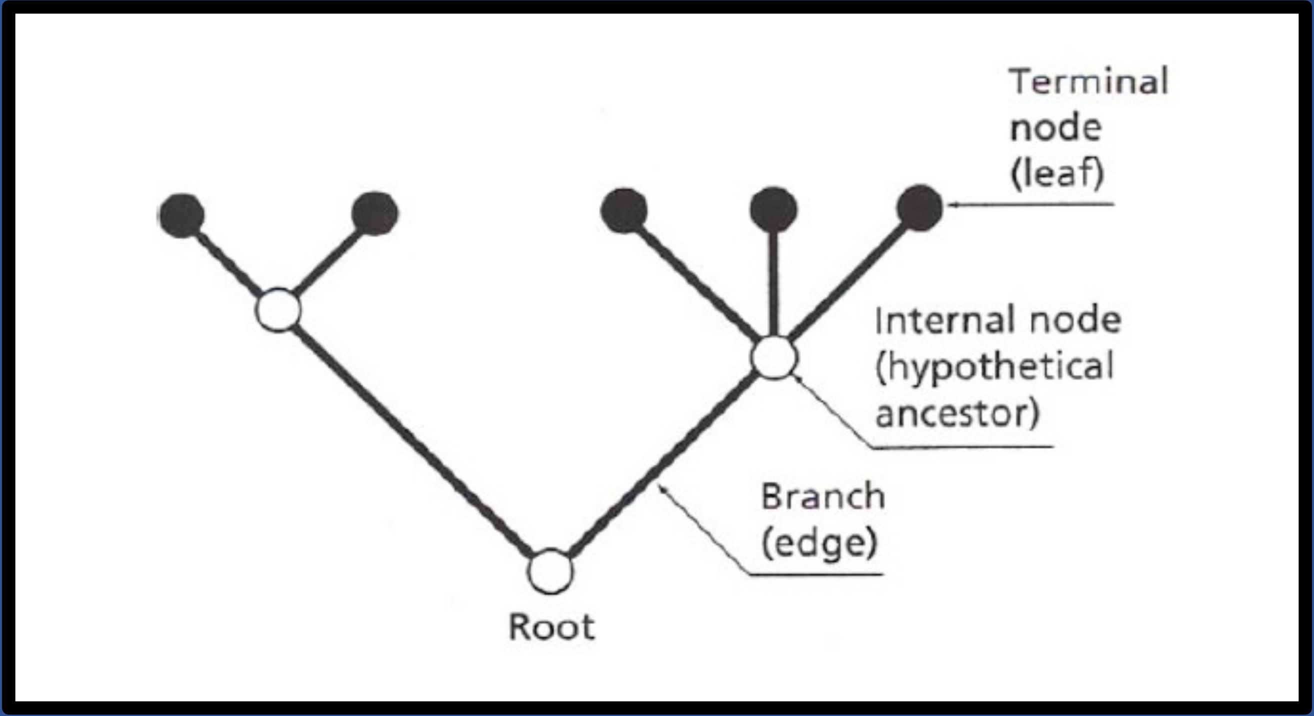
Phylogenetics – Terminology

Tree: A mathematical structure used to model evolutionary history of a group of sequences or organisms

Node: Taxonomic unit (e.g. species, population, individual, gene), can be tip or tree or internal

Branch: Evolutionary pathways between nodes

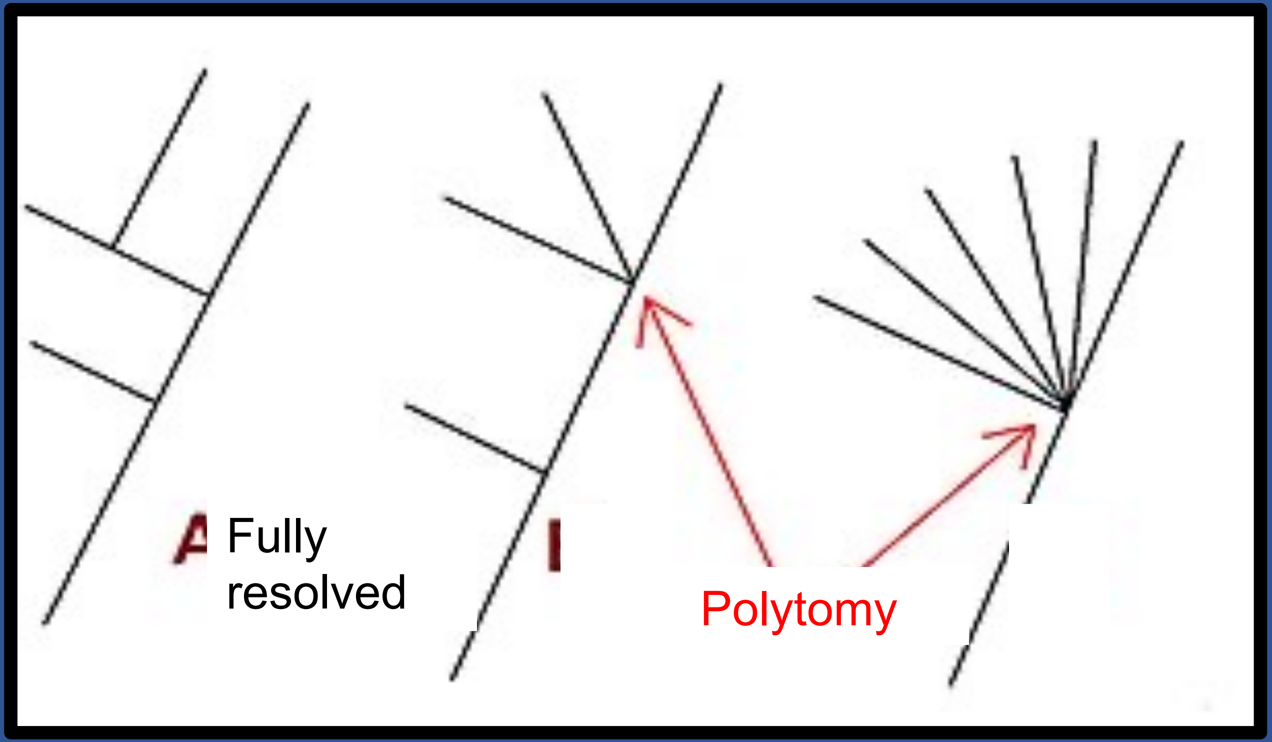
Root: Most recent common ancestor of all terminal in tree



Phylogenetics – Terminology

Bifurcating node: when internal node gives rise to only two immediate descendant lineages

Multifurcating node: when internal node gives rise to 3+ immediate descendant lineages; sometimes referred to as a polytomy

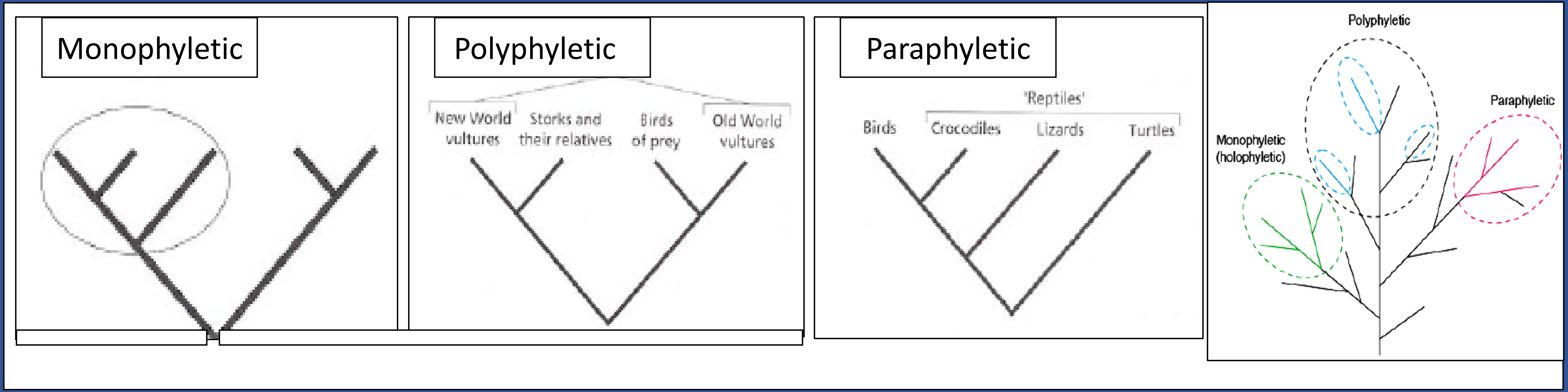


Phylogenetics – Terminology

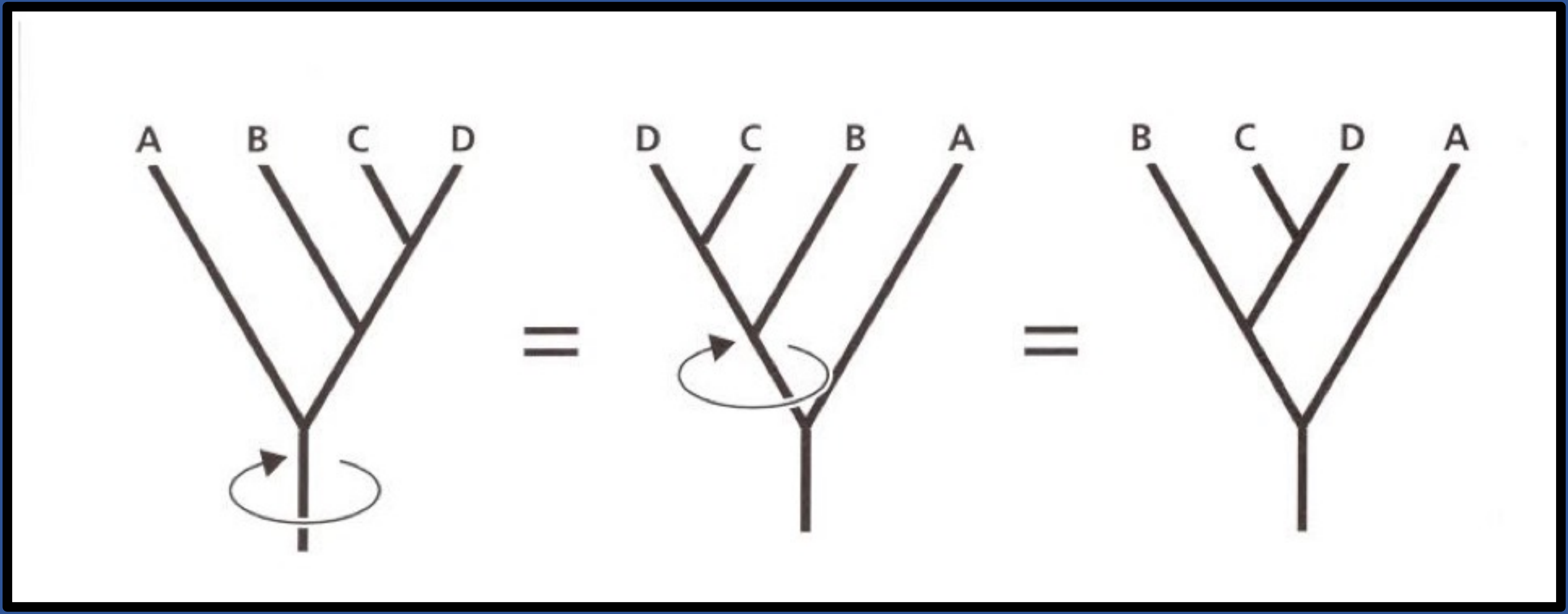
Monophyletic clade: when all taxa within a group are derived from a common ancestor (that is contained within the group)

Polyphyletic clade: a taxonomic group with at least one member who’s last common ancestor is not a member of the group.

Paraphyletic clade: a group that contains the most recent common ancestor but does not contain all the descendants of that ancestor



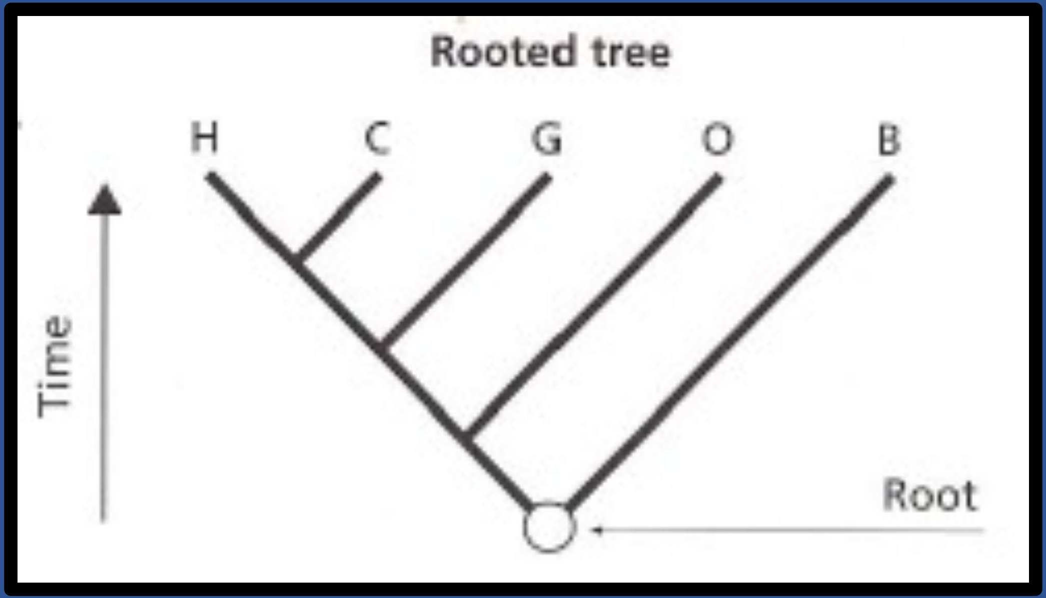
Phylogenetics – Trees can rotate!

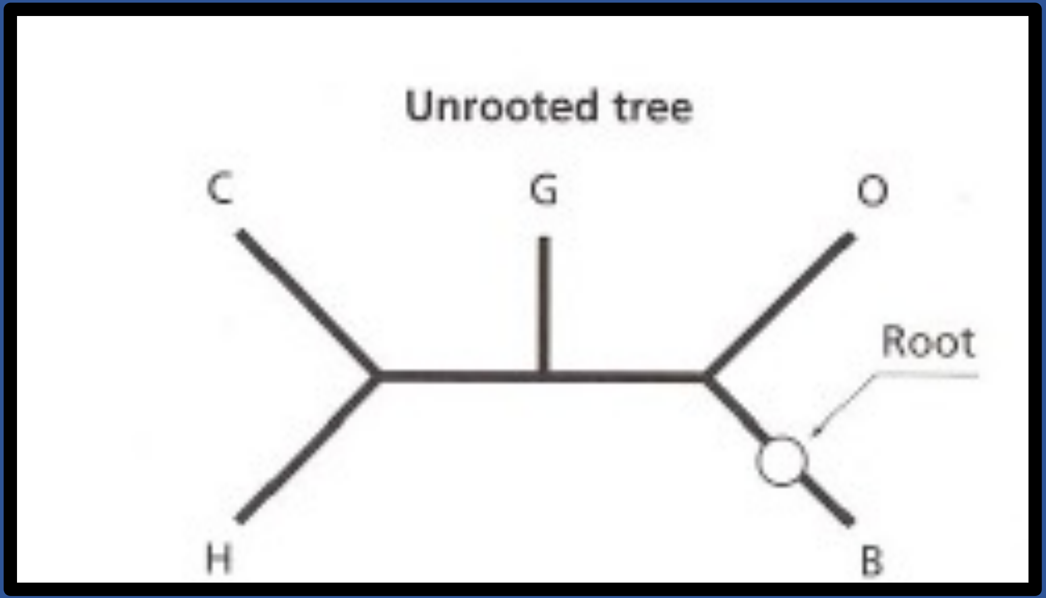


Phylogenetics – Rooted vs Unrooted

Rooted tree: when the direction of each path reflects evolutionary time; a tree that reflects kinship and evolutionary pathways

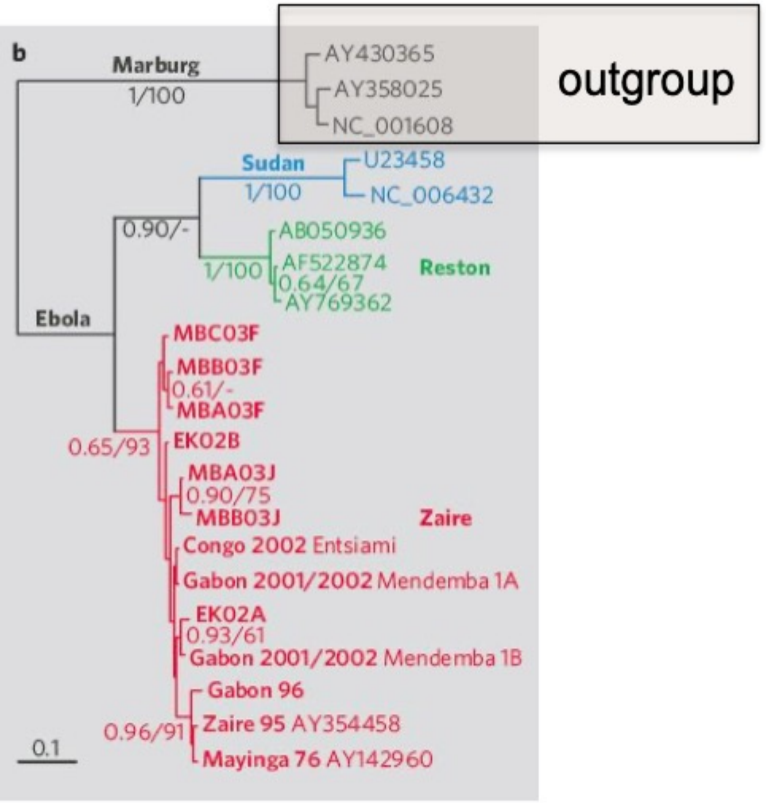
Unrooted tree: a tree that lacks a root that only reflects kinship



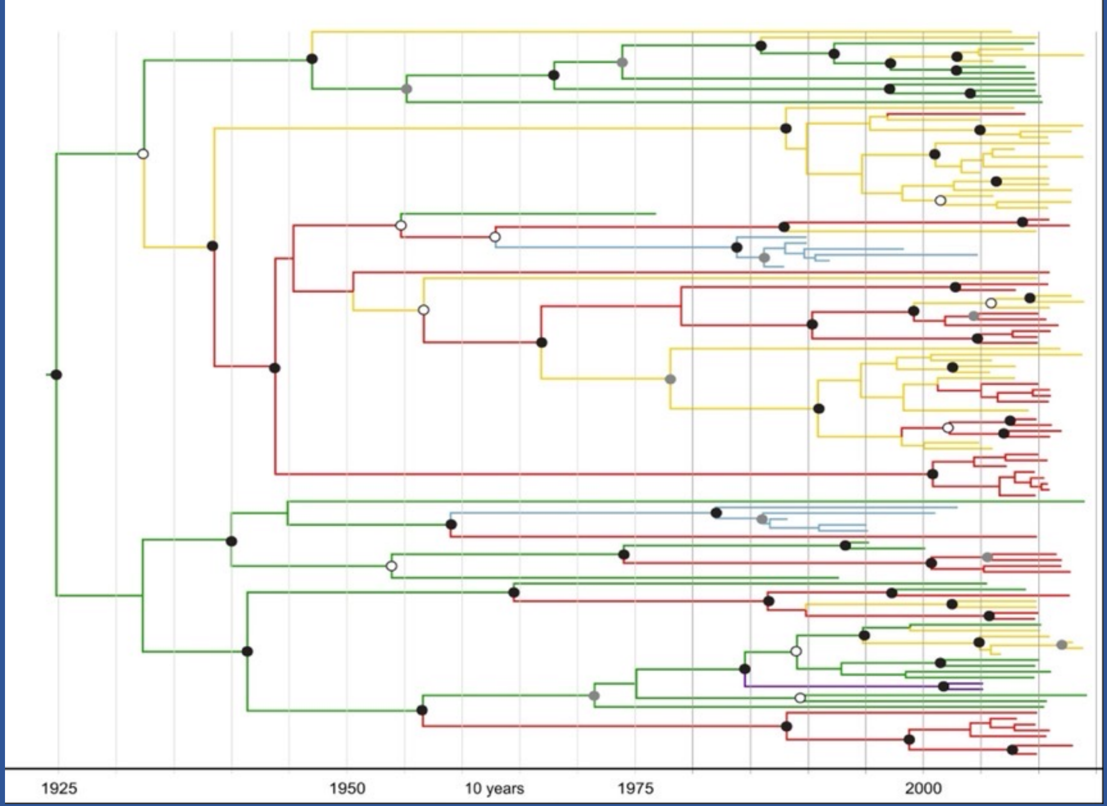


How to root a tree?

1. Root using an **outgroup** – one or more taxa outside the group of interest



2. Using a **molecular clock** – orients tree with a time axis



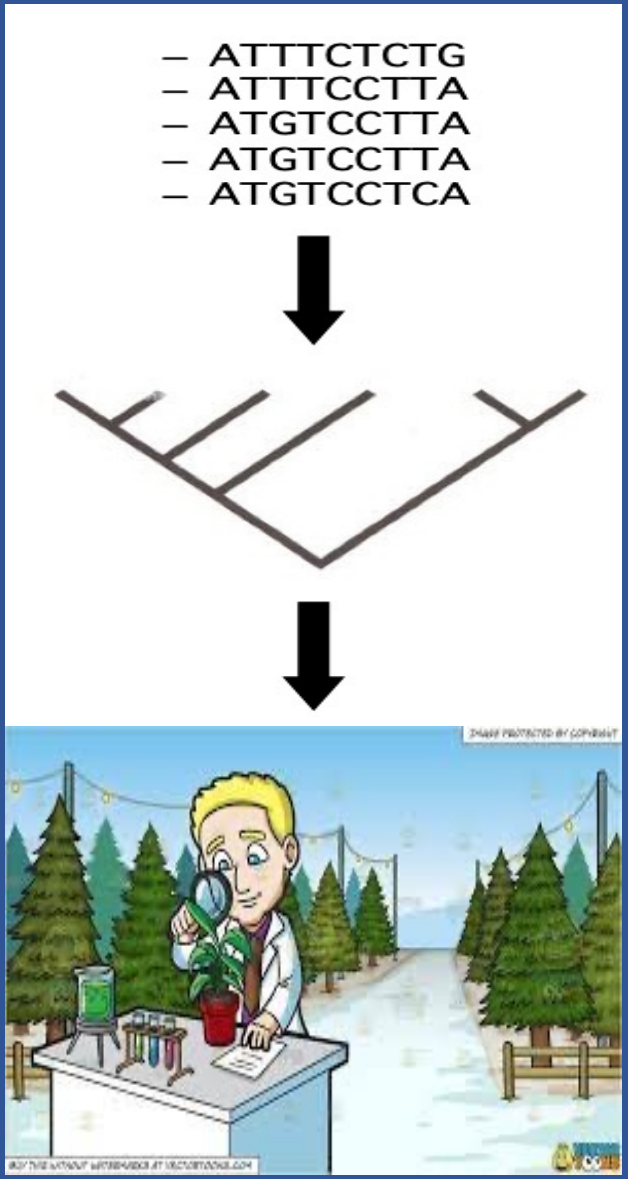
FFV in puma

Generating trees – overall aim

1. Measure genetic variation

2. Develop models that fit the observed patterns

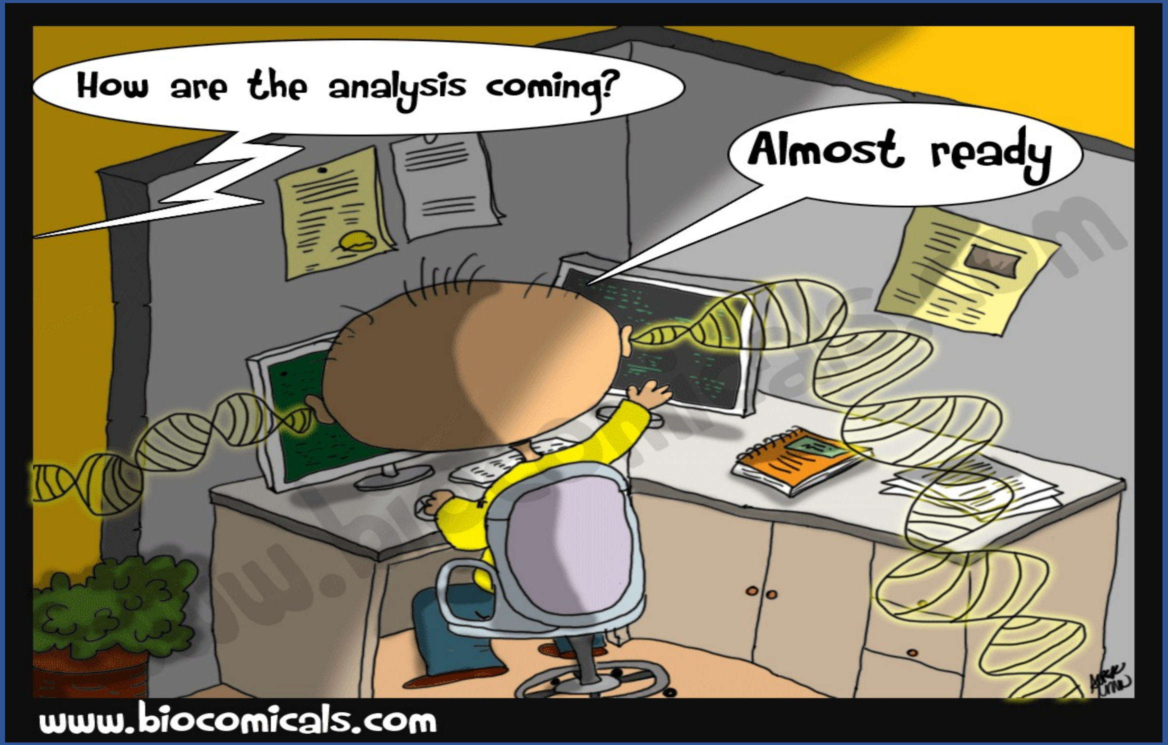
3. Infer process from patterns



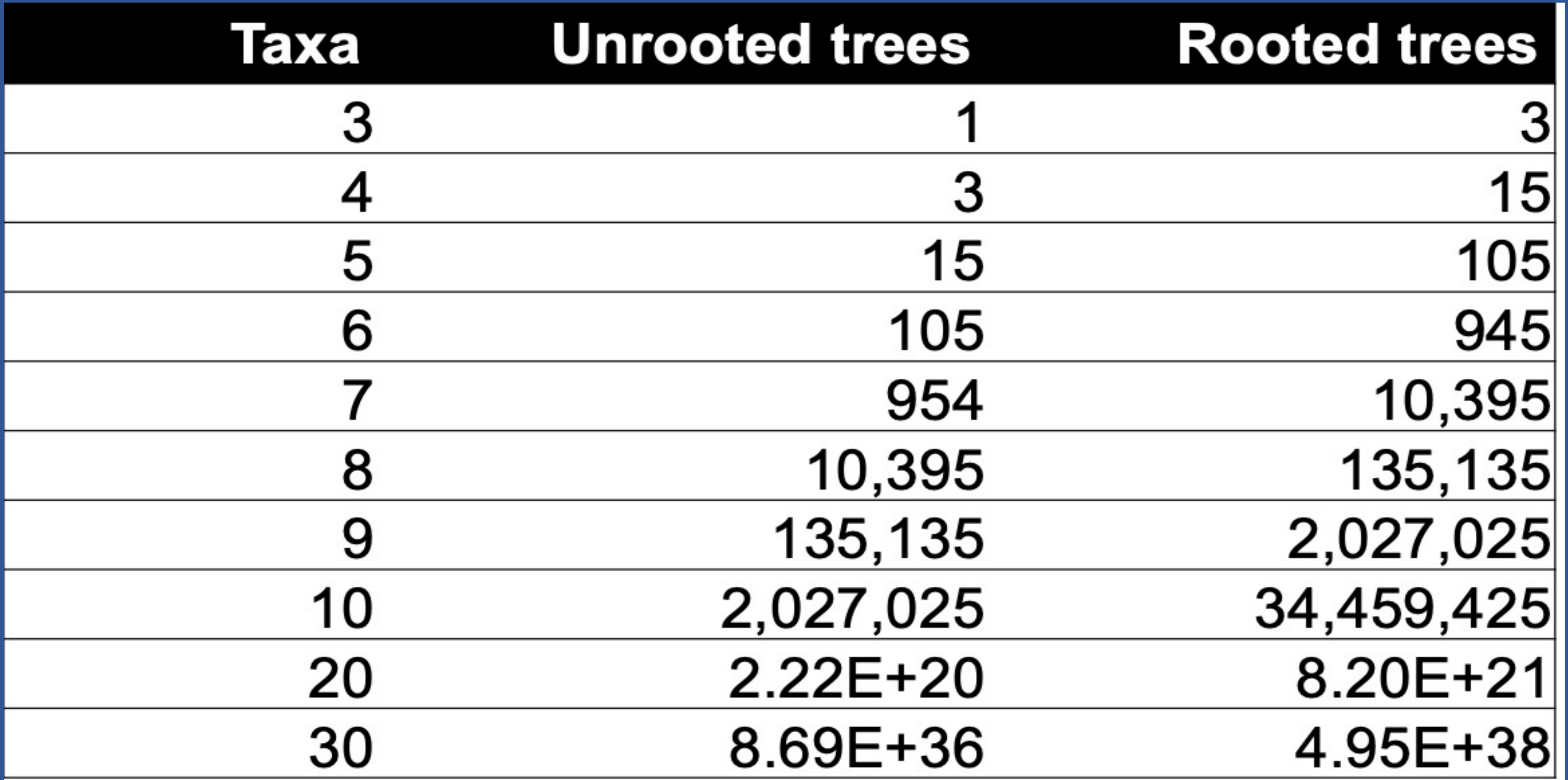
How do we build a tree?

Four basic categories:

- Distance
- Maximum parsimony
- Maximum likelihood
- Bayesian methods



Number of overall trees rises quickly!

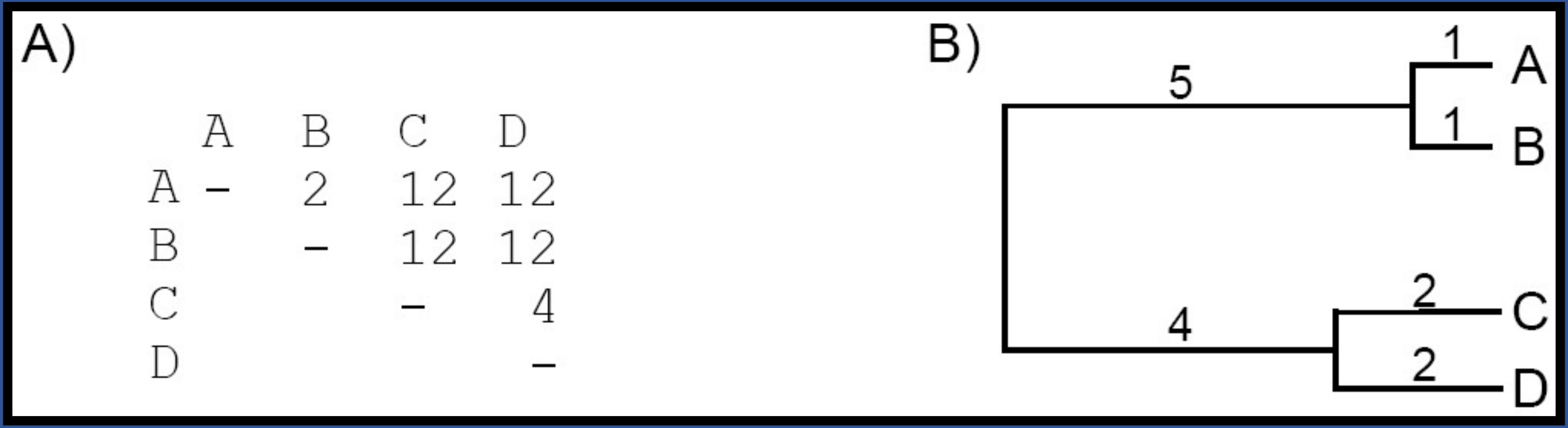


Taxa	Unrooted trees	Rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	954	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
20	2.22E+20	8.20E+21
30	8.69E+36	4.95E+38

Distance based

A. The pairwise genetic distances between species are provided in a matrix – number represents the percent different.

B. The genetic distances are used to generate the tree



A)

	A	B	C	D
A	–	2	12	12
B		–	12	12
C			–	4
D				–

Distance based

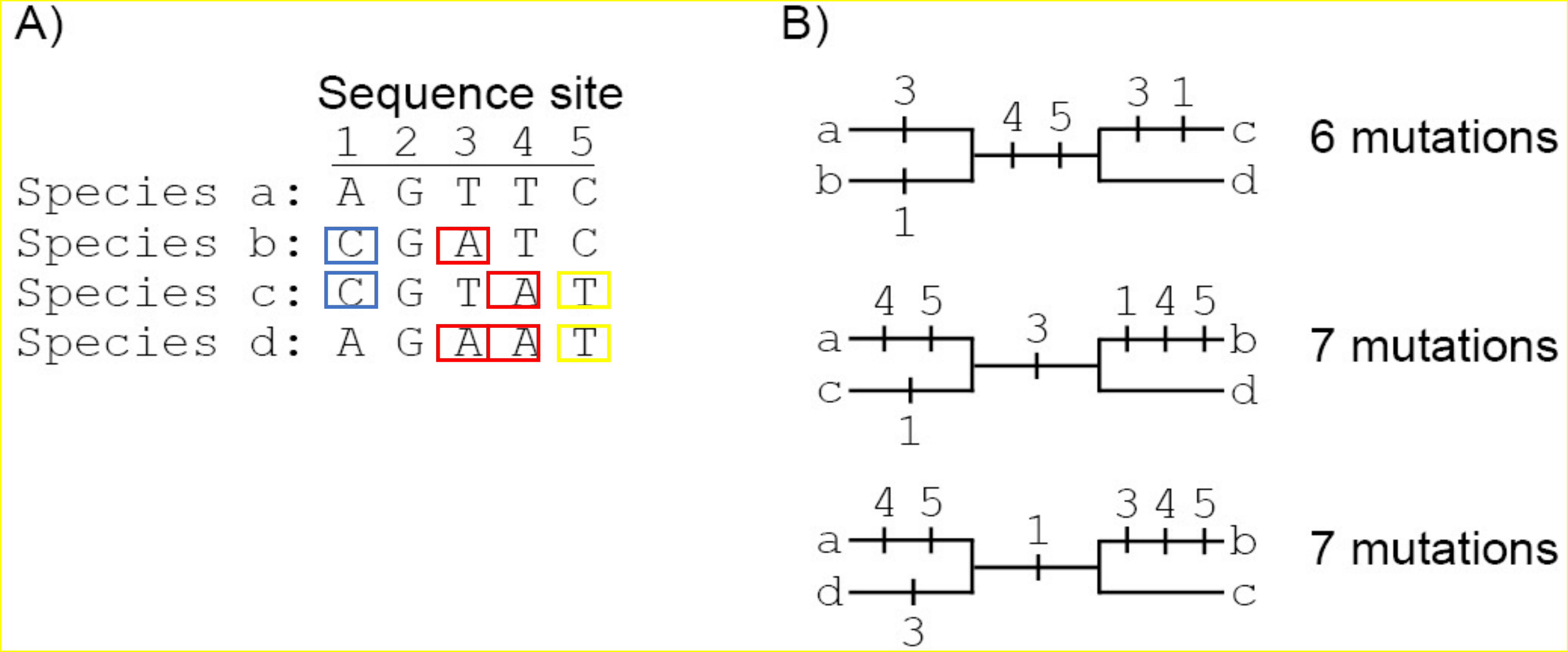
A. Calculate pairwise distances among all sequences (according to some substitution model)

B. Use distances to build tree (according to some rule e.g. “neighbor joining” method)

Highlights

- Fast to compute, even for large data sets
- Information about character state change is lost

Maximum Parsimony



A)

Sequence site

	1	2	3	4	5
Species a:	A	G	T	T	C
Species b:	C	G	A	T	C
Species c:	C	G	T	A	T
Species d:	A	G	A	A	T

Maximum Parsimony

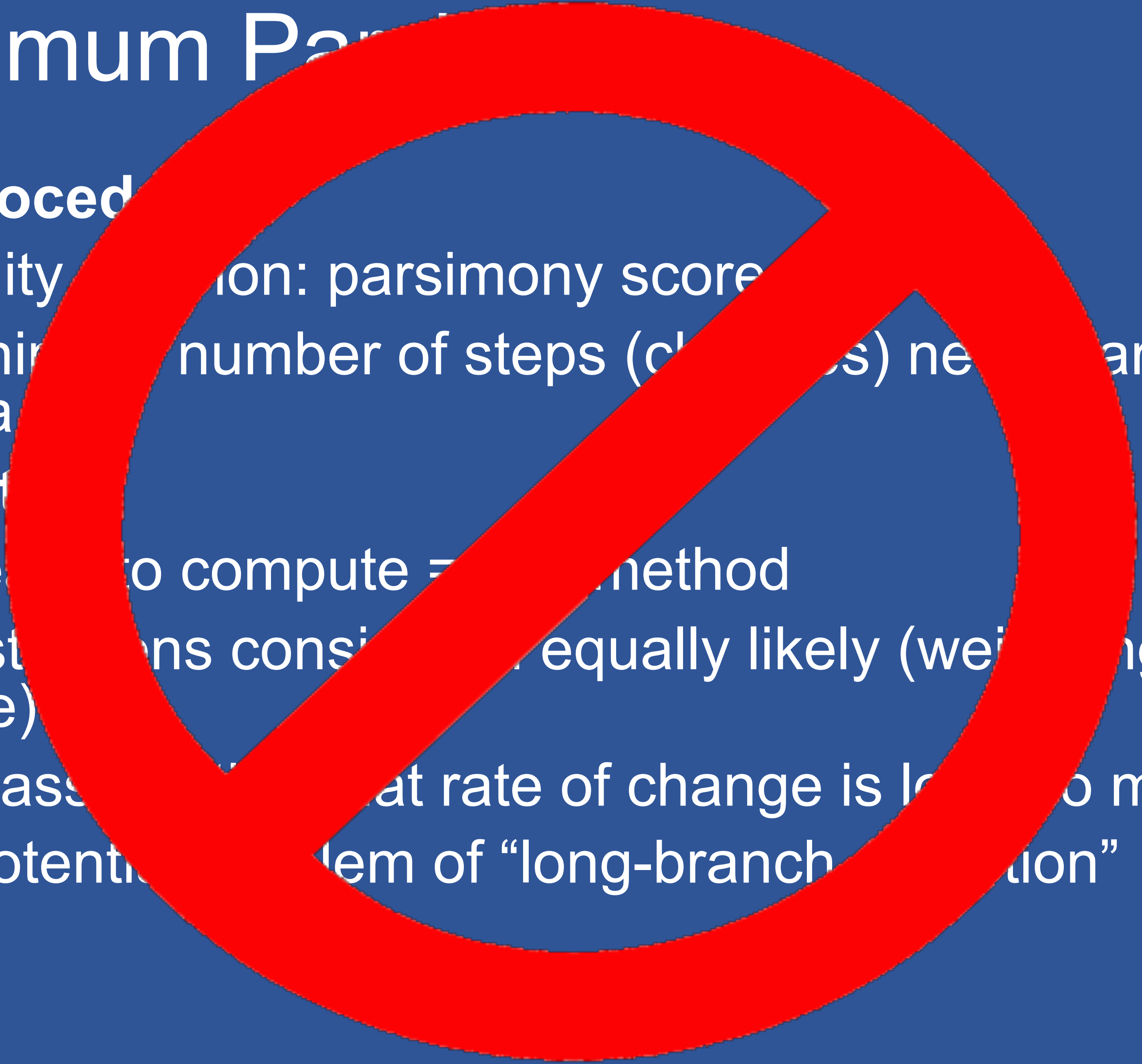
Basic procedure

- Optimality criterion: parsimony score
- The minimum number of steps (changes) necessary to explain the data

Highlights

- Score easy to compute = fast method
- All substitutions considered equally likely (weighting schemes possible)
- Implicit assumption that rate of change is low (no multiple hits)
  - Potential problem of “long-branch attraction”

Maximum Pa



Basic proced

- Optimality ion: parsimony score
- The mini number of steps (c s) ne ary to explain the data

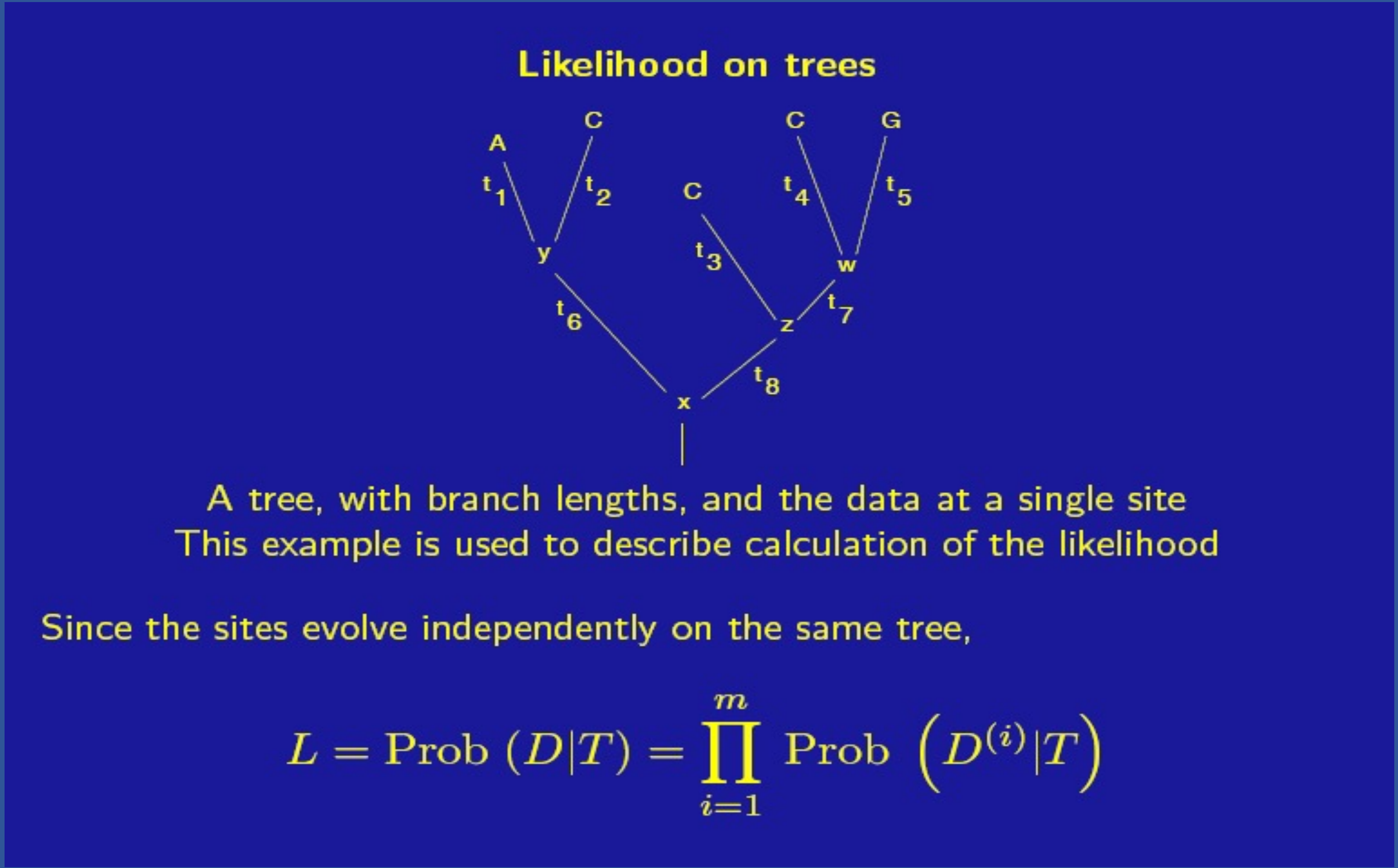
Highlight

- Score e o compute = method
- All subst ns consi equally likely (wei ng schemes possible)
- Implicit ass at rate of change is lo o multiple hits)
    - Potenti lem of “long-branch tion”

Maximum Likelihood

Given a model of sequence evolution, the ML tree is the combination of topology and branch lengths that maximizes the likelihood (probability) of the observed data (i.e., character state patterns among taxa in the data set)

Maximum Likelihood



Likelihood on trees

A tree, with branch lengths, and the data at a single site
This example is used to describe calculation of the likelihood

Since the sites evolve independently on the same tree,

$$L = \text{Prob}\ (D|T) = \prod_{i=1}^{m} \text{Prob}\ \left(D^{(i)}|T\right)$$

Maximum Likelihood

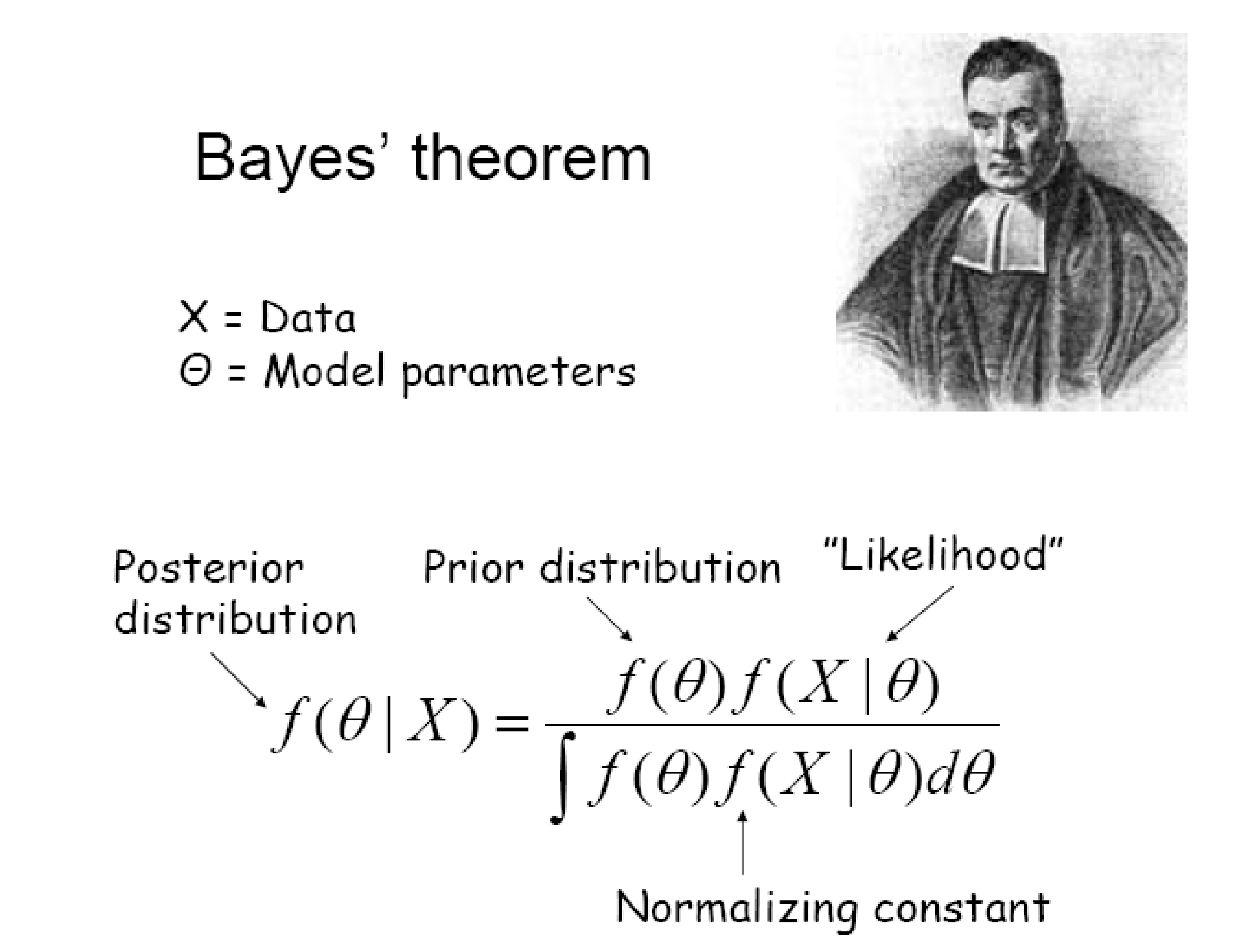
Basic procedure

- Optimality criterion: likelihood score
- Maximize the probability of the sequences, given a tree and its branch lengths plus an evolutionary model and its parameters

Highlights

- Allows full use of evolutionary models
- Relies heavily on model chosen = can be misleading if there is much variation in the substitution process among lineages
- Computationally much more demanding

Bayesian inference: the most probable outcome according to prior knowledge



Bayes' theorem

X = Data
Θ = Model parameters

Posterior distribution

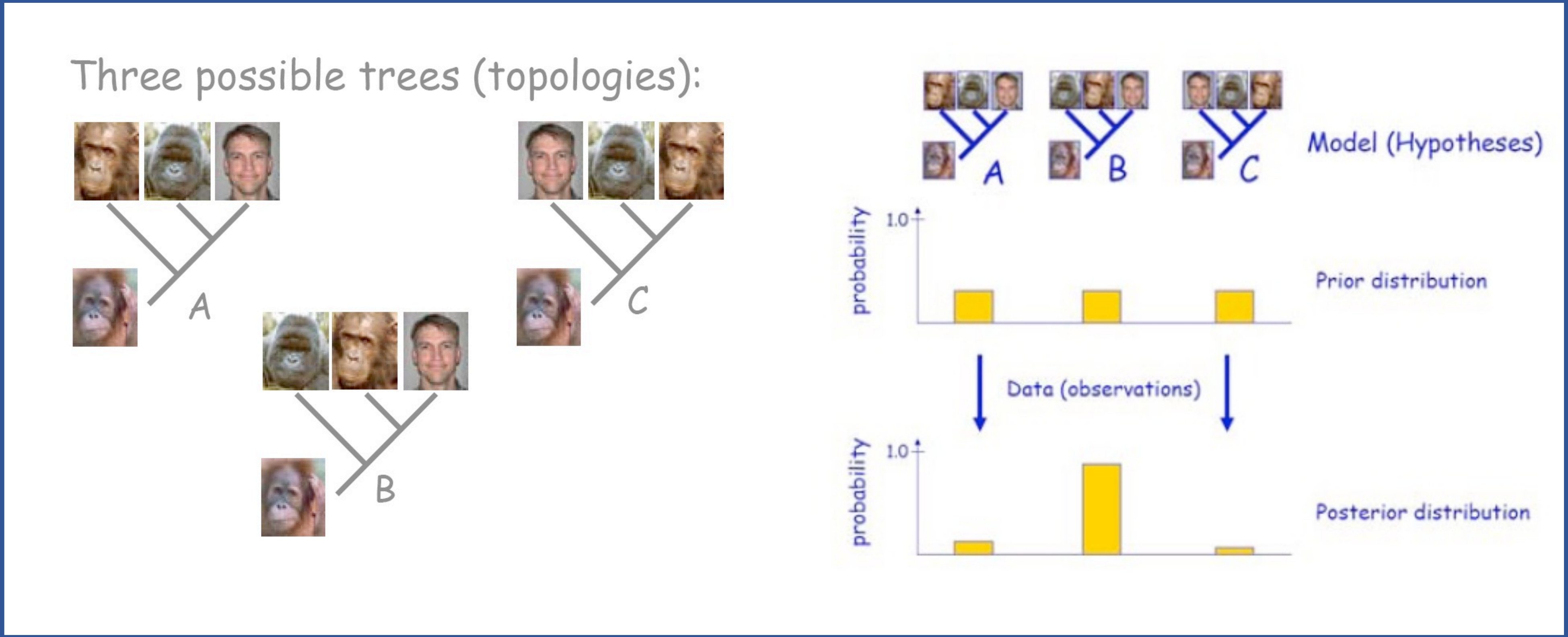
Prior distribution

"Likelihood"

$$f(\theta \mid X) = \frac{f(\theta) f(X \mid \theta)}{\int f(\theta) f(X \mid \theta) d\theta}$$

Normalizing constant

Bayesian inference of phylogeny



Bayesian Phylogenetics

Basic procedure

- Objective: determine the posterior distribution of trees given the sequence data
- Based on this distribution, ‘best’ tree can be identified

Highlights

- Allows full use of evolutionary models
- Need to include priors BUT this can also expand inferences
- Posterior probabilities are approximated through Markov Chain Monte Carlo (MCMC) methods that sample from the posterior
- Clade probabilities provide measure of uncertainty

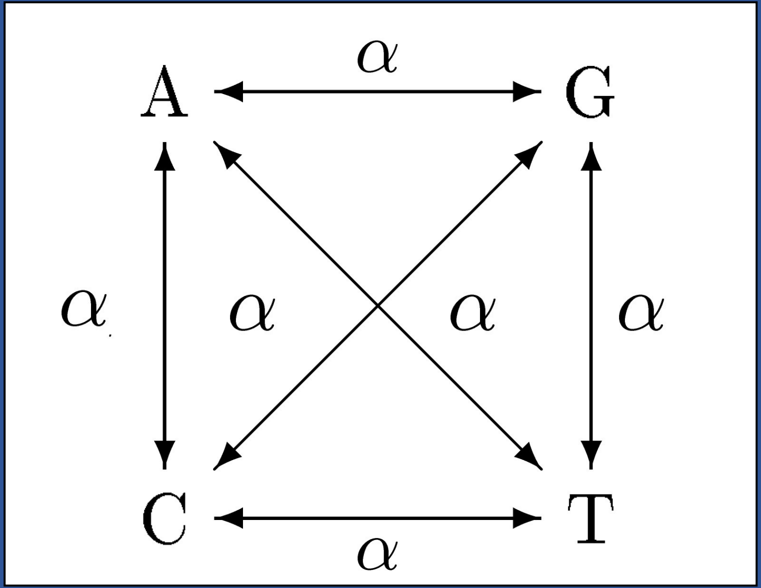
Stretch-think-discuss-share

What approach would you use for your datasets? When might you use a different approach?

Models of sequence evolution

Models that assume all nucleotides occur at equal frequencies (25%)

1. The Jukes-Cantor (JC) model
   a. All substitutions are equally likely.
   b. All nucleotides occur at the same frequency (25%).
   c. One parameter: the rate of substitution (alpha).



2. Kimura two parameter (K2P) model
   a. Transitions (α) (purine to purine or pyrimidine to pyrimidine substitutions) are more common than transversions (β)
   b. All nucleotides occur at the same frequency.
   c. Two parameters: transition rate (alpha) and transversion rate (beta).

More complicated but more biologically realistic

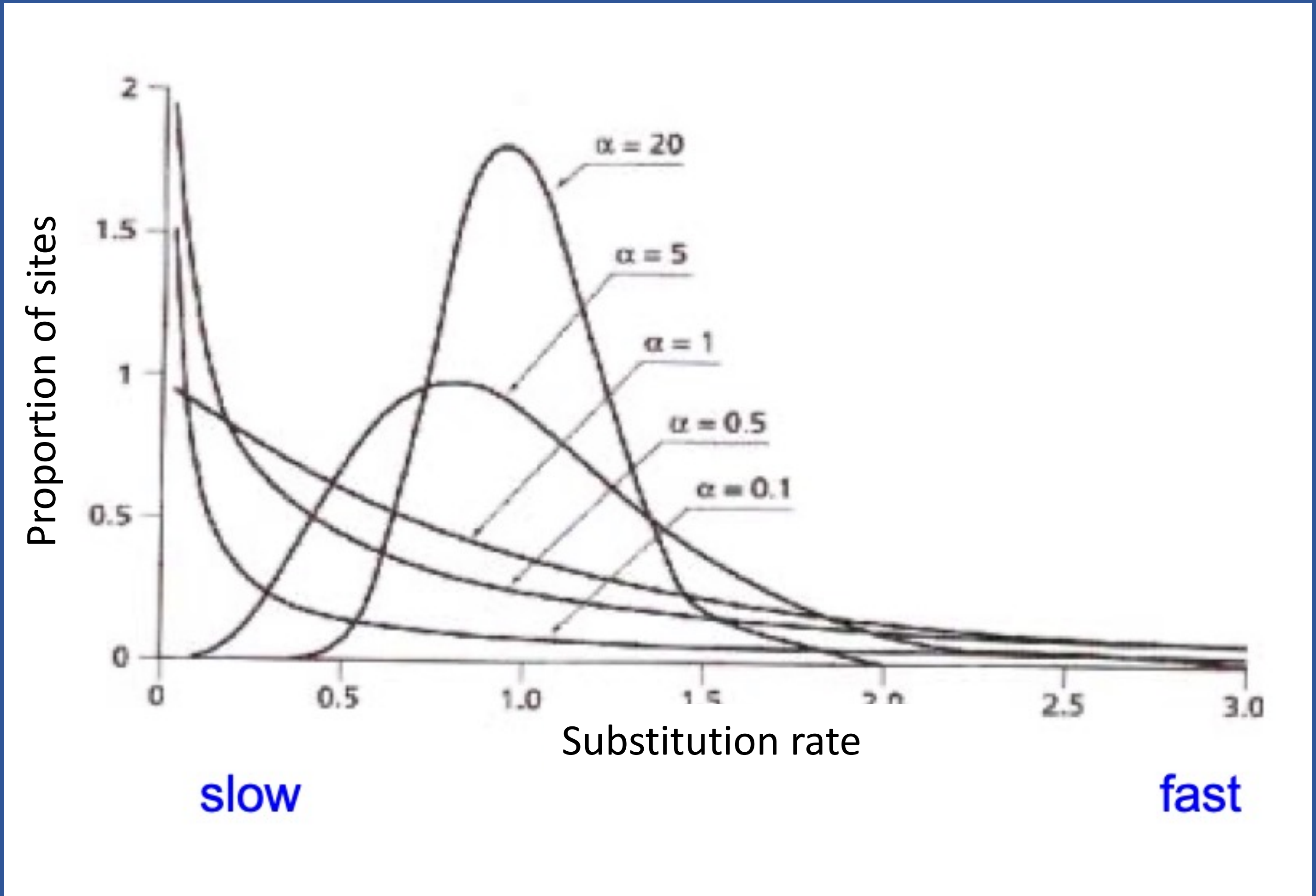
Models that allow the four nucleotides to be present in different frequencies

3. Felsenstein (F84) & Hasegawa-Kishono-Yano (HKY85) models
    a. Two closely related models -- they use different calculations to model essentially the same thing
    b. Transitions and transversions occur at different rates
    c. Nucleotides occur at different frequencies

4. General time reversible (GTR) model
    a. Assumes a symmetric substitution matrix (and thus is time reversible)
    b. In other words, A changes into T with the same rate that T changes into A.
    c. Each pair of nucleotide substitutions has a different rate
    d. Nucleotides can occur at different frequencies

Variation among sites

Some sites undergo changes more frequently than others - can be expressed using a gamma distribution



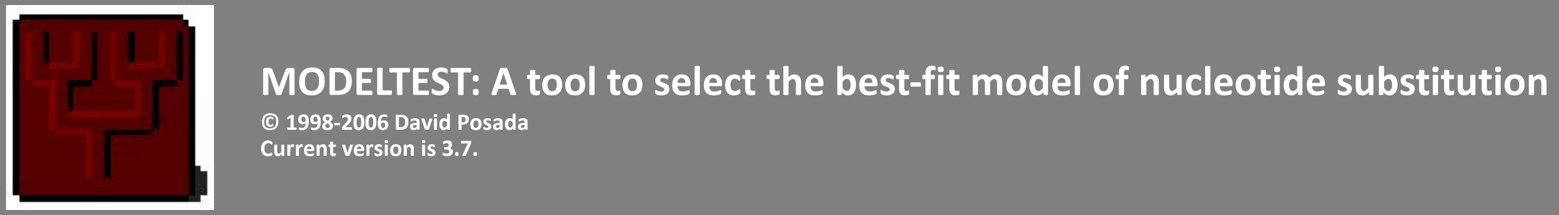
Think-pair-share

If more complex models are more biologically realistic, then are more complex models always better to use than simpler models?

Choosing a model of nucleotide substitution

More complex does not mean better, due to sampling error
Select model according to sequence:

1. **Length**
2. **Composition (base pair frequencies, protein-coding?)**
3. **Polymorphism (syn. Vs. nonsyn. Substitutions?)**



MODELTEST: A tool to select the best-fit model of nucleotide substitution
© 1998-2006 David Posada
Current version is 3.7.

jModeltest is program for the selecting the model of nucleotide substitution that best fits the data

Available from:

http://darwin.uvigo.es/software/jmodeltest.html

Fits up to 88 candidate models fit to your sequence data

In practice…..

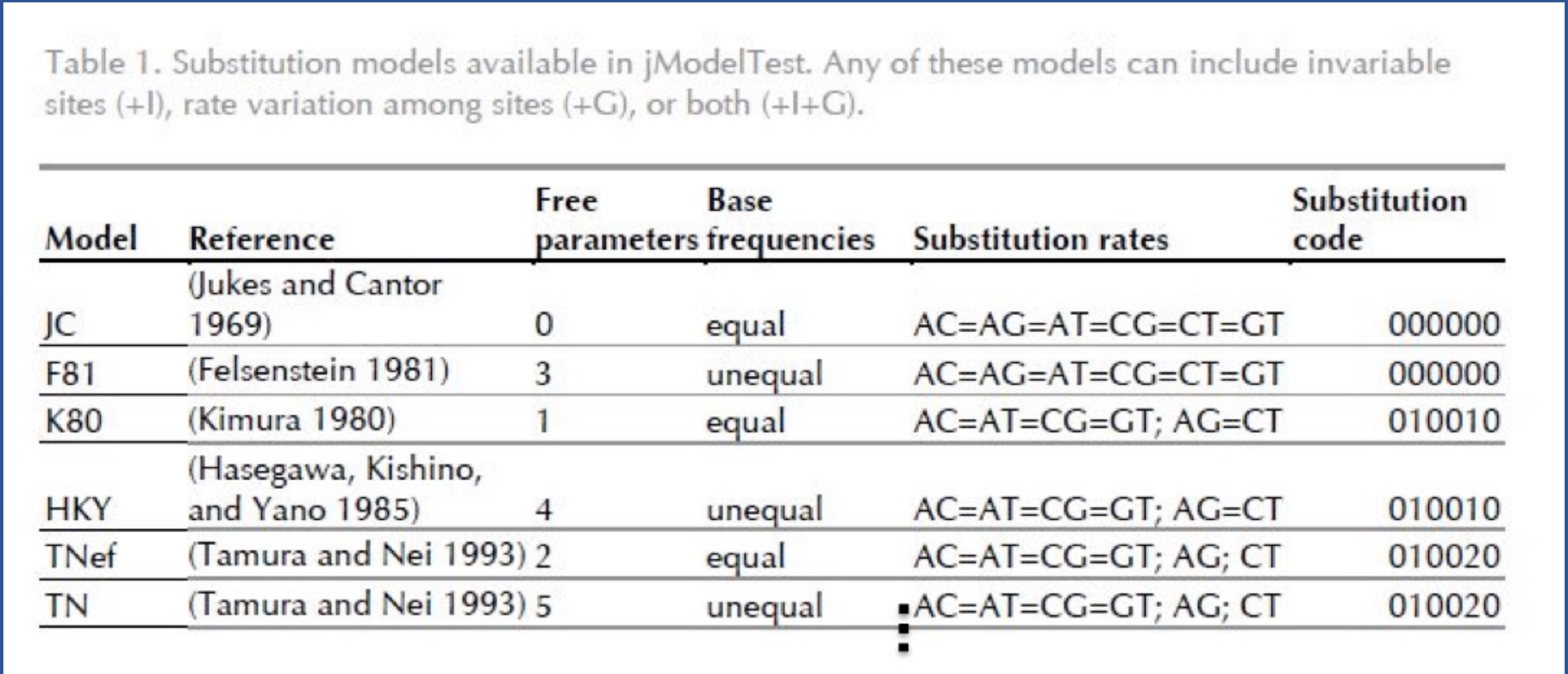


Table 1. Substitution models available in jModelTest. Any of these models can include invariable sites (+I), rate variation among sites (+G), or both (+I+G).

Model	Reference	Free parameters	Base frequencies	Substitution rates	Substitution code
JC	(Jukes and Cantor 1969)	0	equal	AC=AG=AT=CG=CT=GT	000000
F81	(Felsenstein 1981)	3	unequal	AC=AG=AT=CG=CT=GT	000000
K80	(Kimura 1980)	1	equal	AC=AT=CG=GT; AG=CT	010010
HKY	(Hasegawa, Kishino, and Yano 1985)	4	unequal	AC=AT=CG=GT; AG=CT	010010
TNef	(Tamura and Nei 1993)	2	equal	AC=AT=CG=GT; AG; CT	010020
TN	(Tamura and Nei 1993)	5	unequal	AC=AT=CG=GT; AG; CT	010020

⋮

So now you have a tree

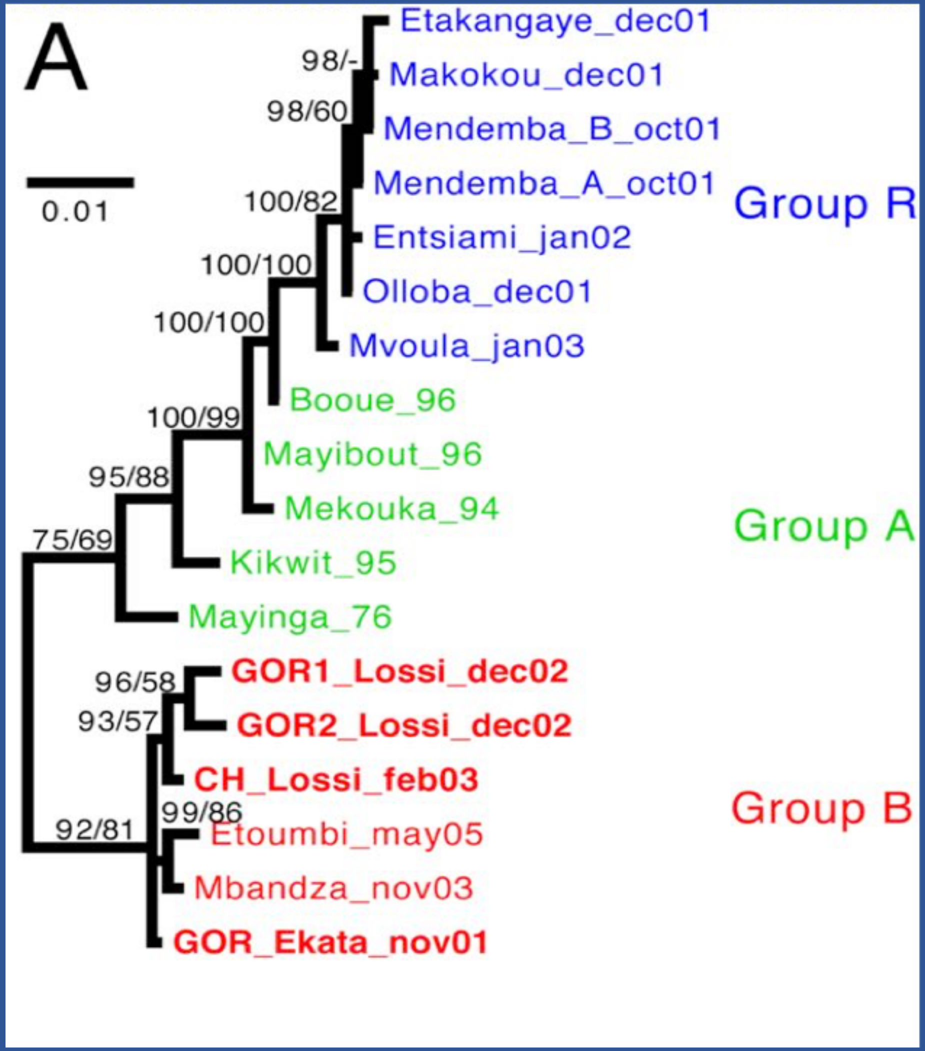
Can it be trusted?

Can it be trusted?

Non-parametric bootstrap

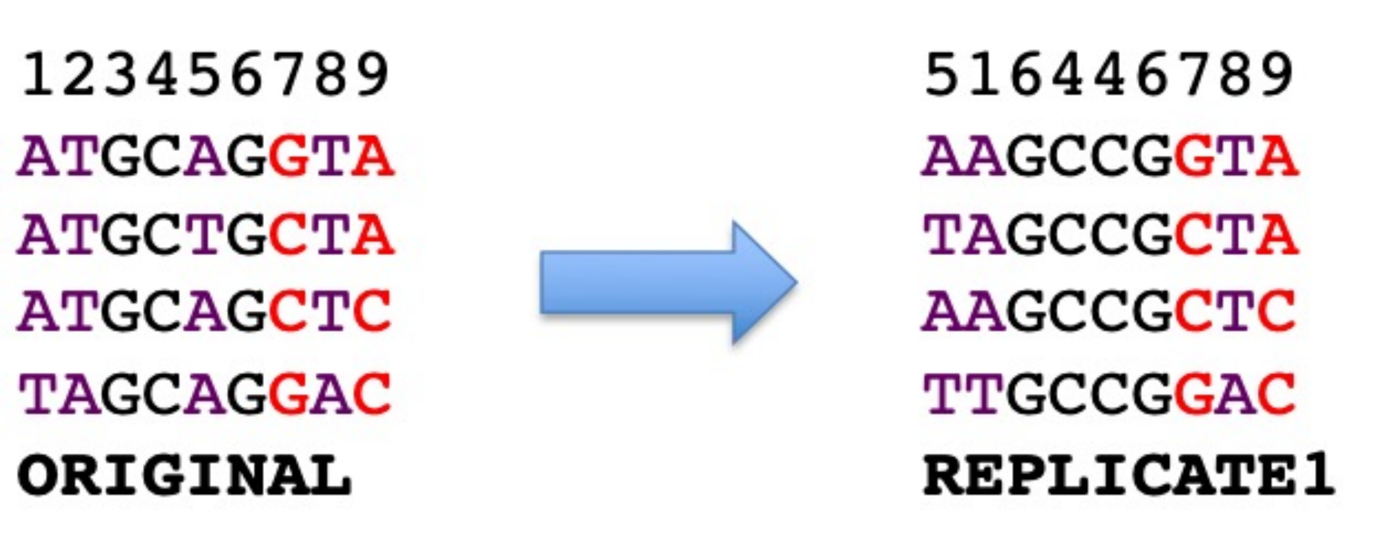
Sample from the original data to create ‘new’ data sets

Count how often a particular clade appears in the resampled data



Bootstrapping

Generate “new” datasets of the same size from the original data by sampling columns with replacement



```
123456789                516446789
ATGCAGGTA                AAGCCGGTA
ATGCTGCTA                TAGCCGCTA
ATGCAGCTC                AAGCCGCTC
TAGCAGGAC                TTGCCGGAC
ORIGINAL                 REPLICATE1
```

Trees build from these new data sets

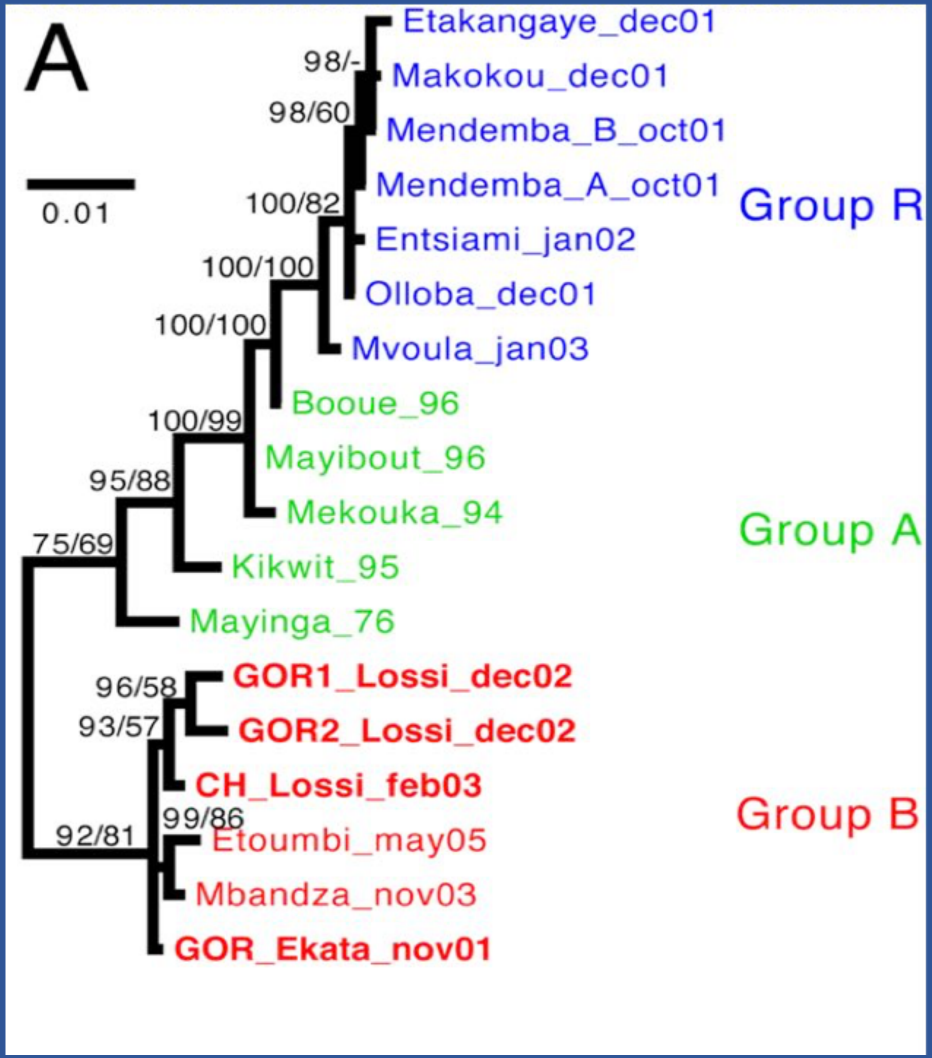
The frequency with which a node appears across replicate trees is taken as a measure of confidence for that node

How well supported is a grouping?

Posterior probabilities

Count the frequency of a clade within the posterior distribution of trees
Less conservative: tend to be much higher than bootstrap values

<u>Strong support:</u>
Bootstrap >0.7
Posterior probabilities >0.95



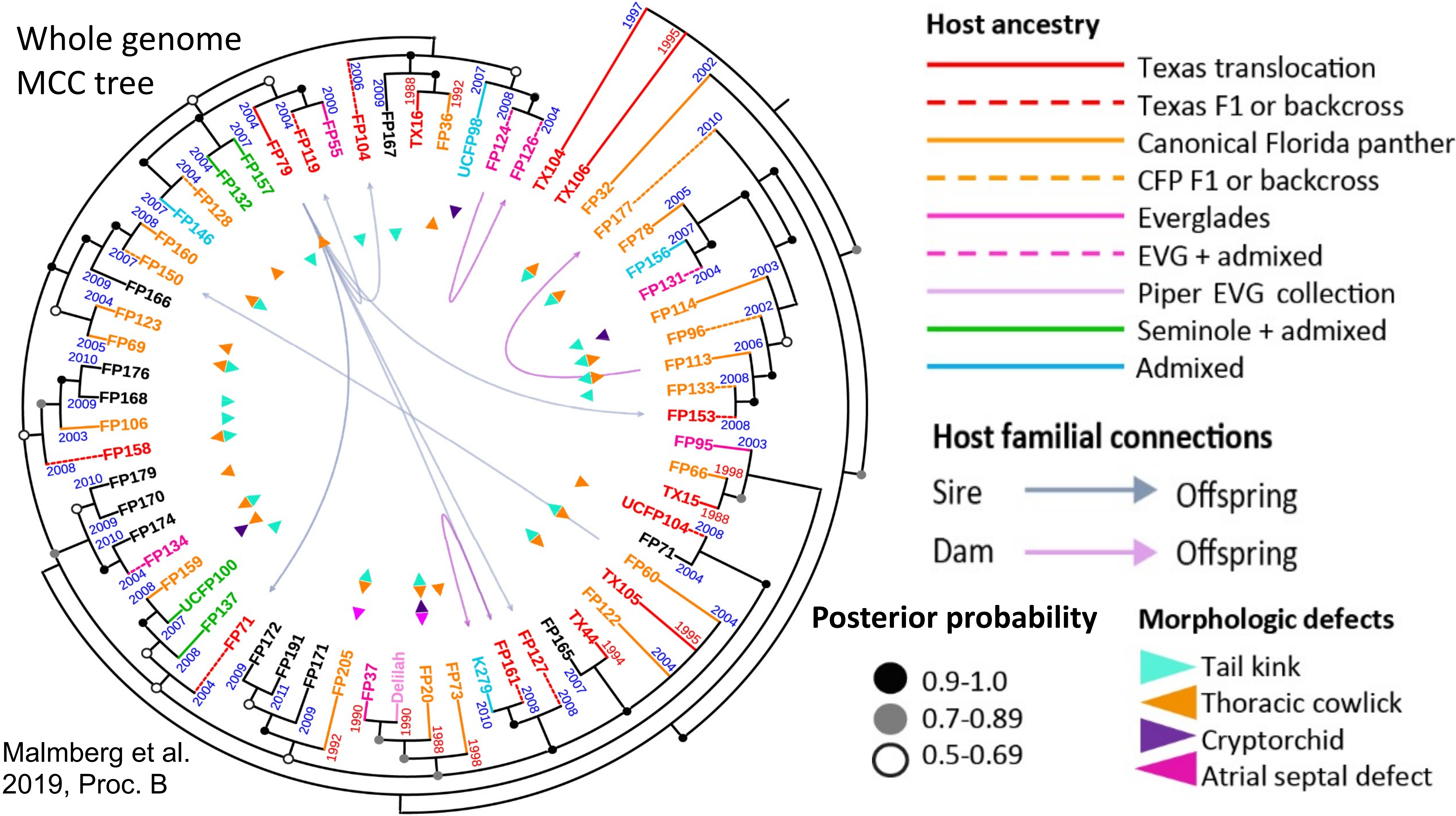
Phylogenetic analysis in practice

1) Collect homologous sequences
2) Conduct multiple alignment
3) Fit an appropriate substitution model
4) Estimate tree(s) under that model
5) Test the reliability of the estimated tree(s)
6) Interpret and apply the phylogenetic tree
7) Potentially repeat steps 4-6 using different tree building methods and/or additional data

What can we learn from our trees?

- Host pathogen co-evolution
- Cross species transmissions
- Geographic structuring
- Temporal structuring
- Transmission events
- Use in more complex downstream analyses…stay tuned

Whole genome MCC tree



Malmberg et al. 2019, Proc. B