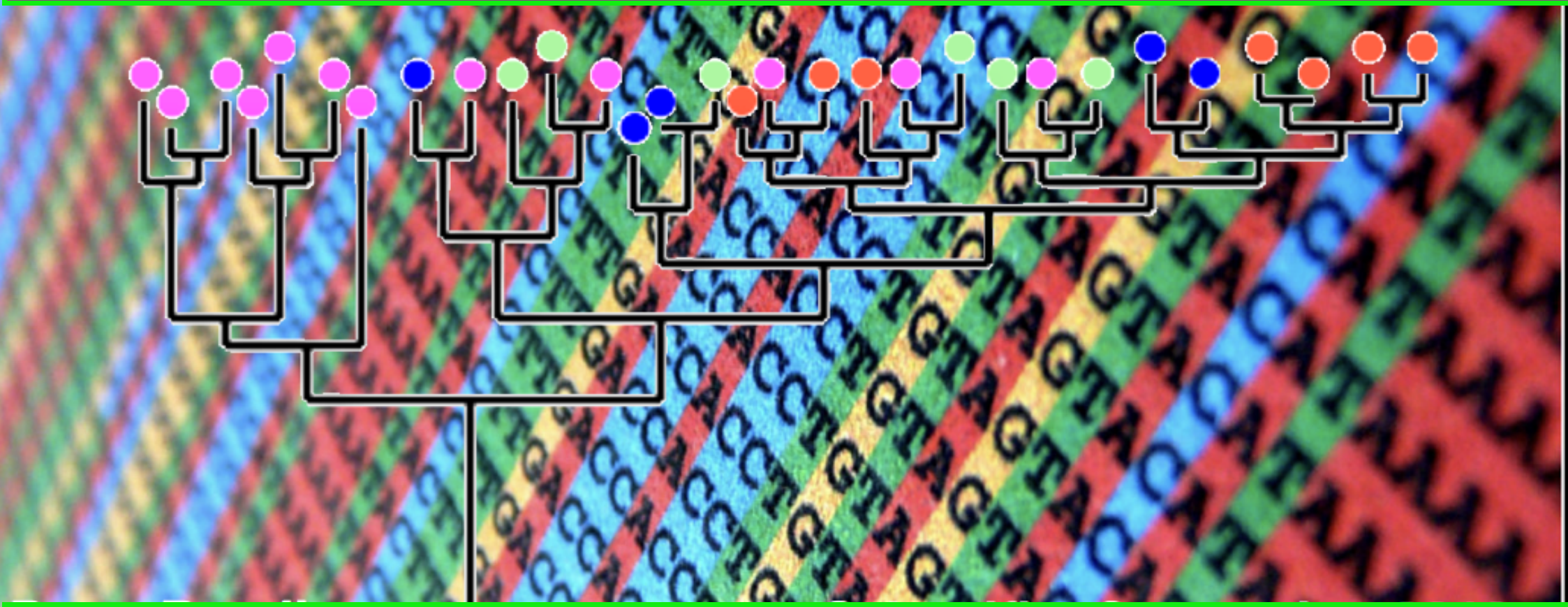
Phylogenetics and some of its applications

an introduction for Scientific Computing folks

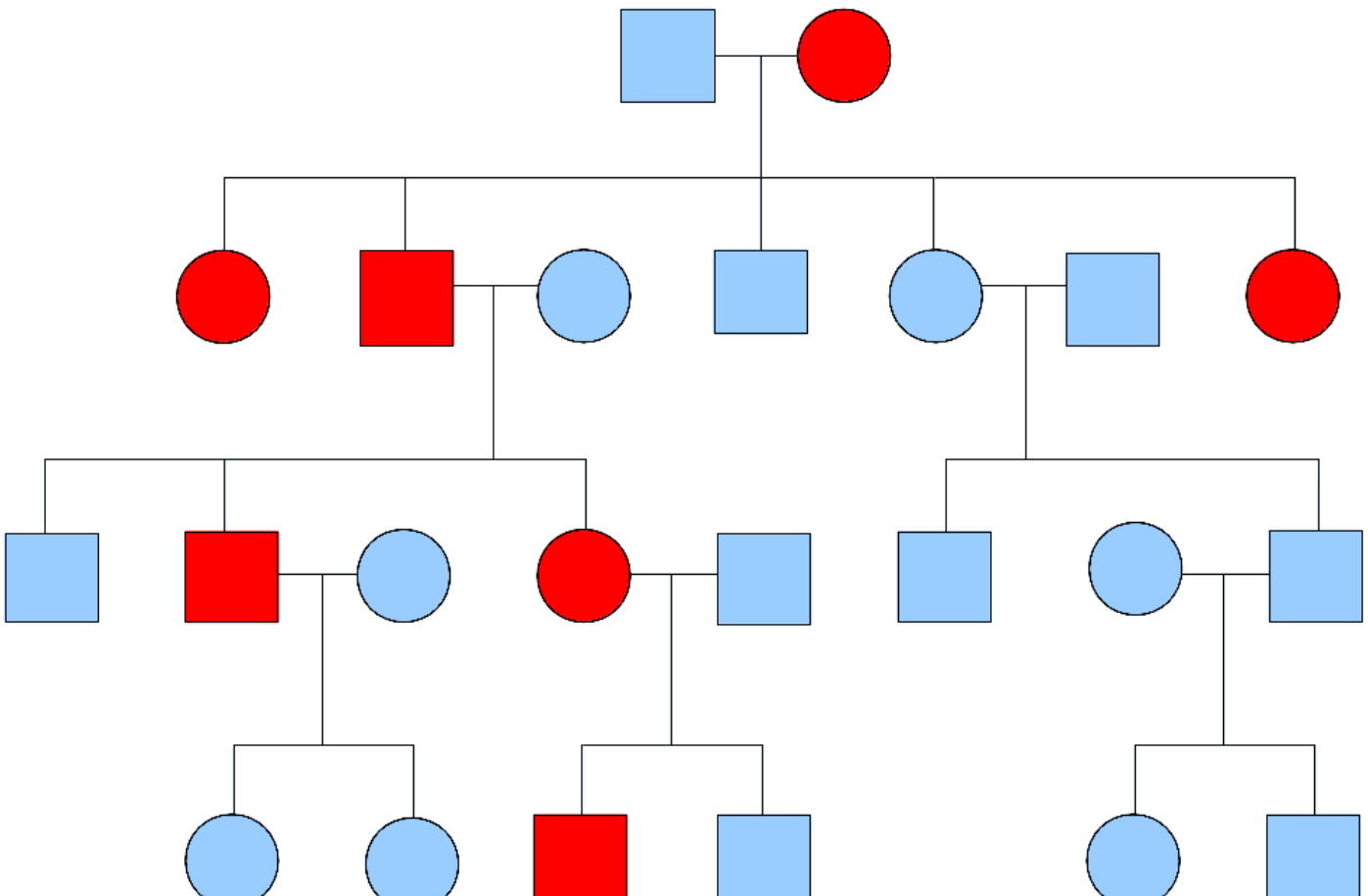


Peter Beerli Scientific Computing, Florida State University Twitter:@peterbeerli

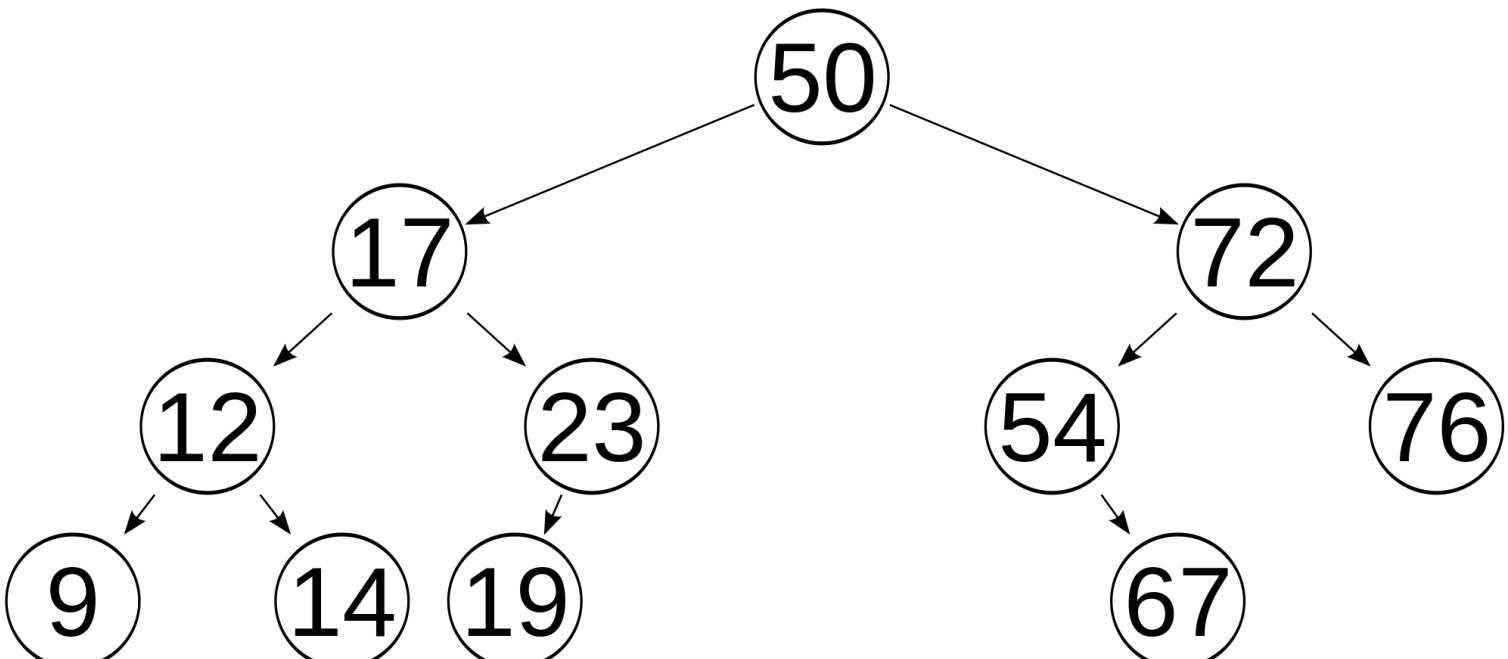
What is a phylogenetic tree

Long leafed Pine

What is a phylogenetic tree

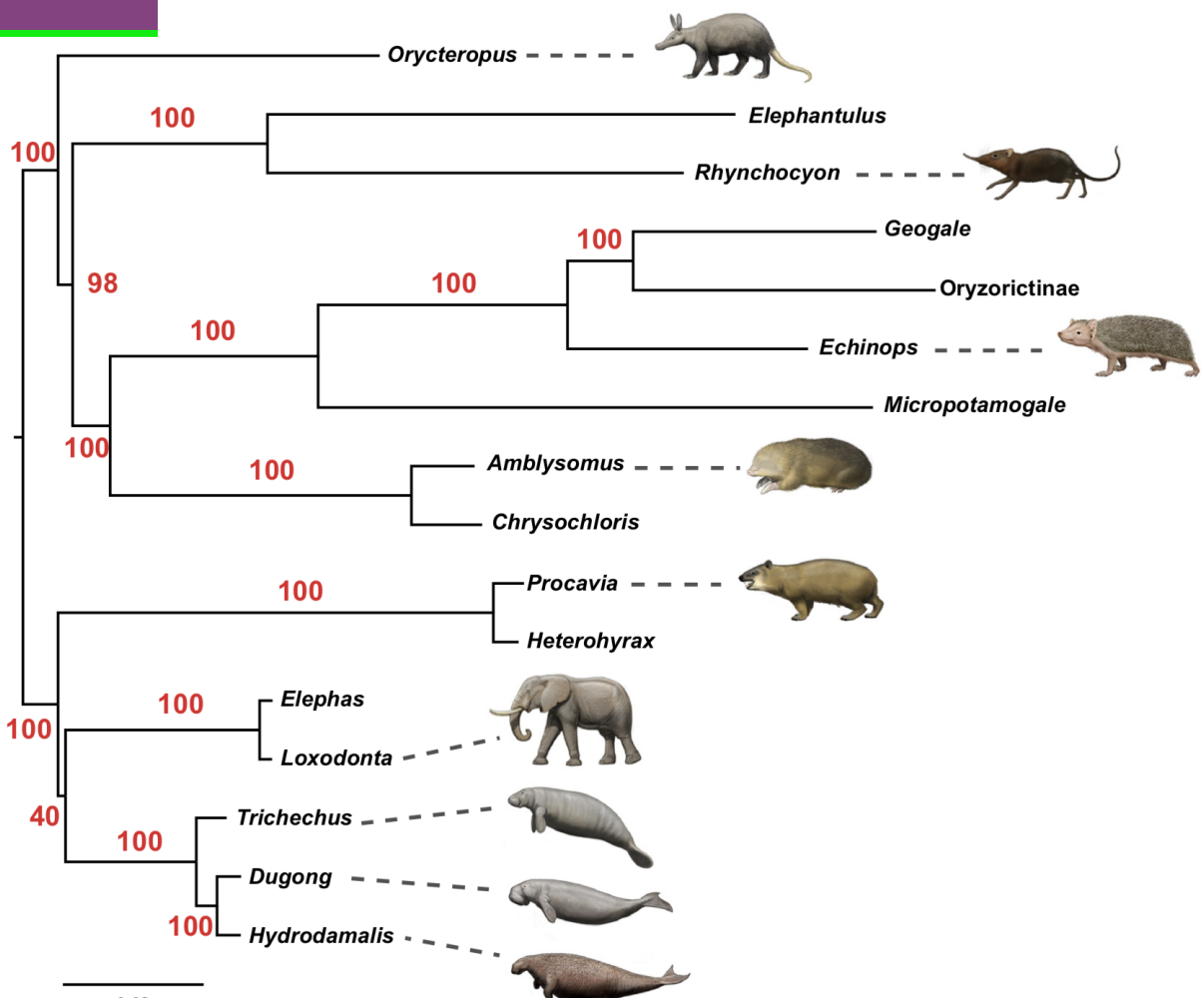


What is a phylogenetic tree



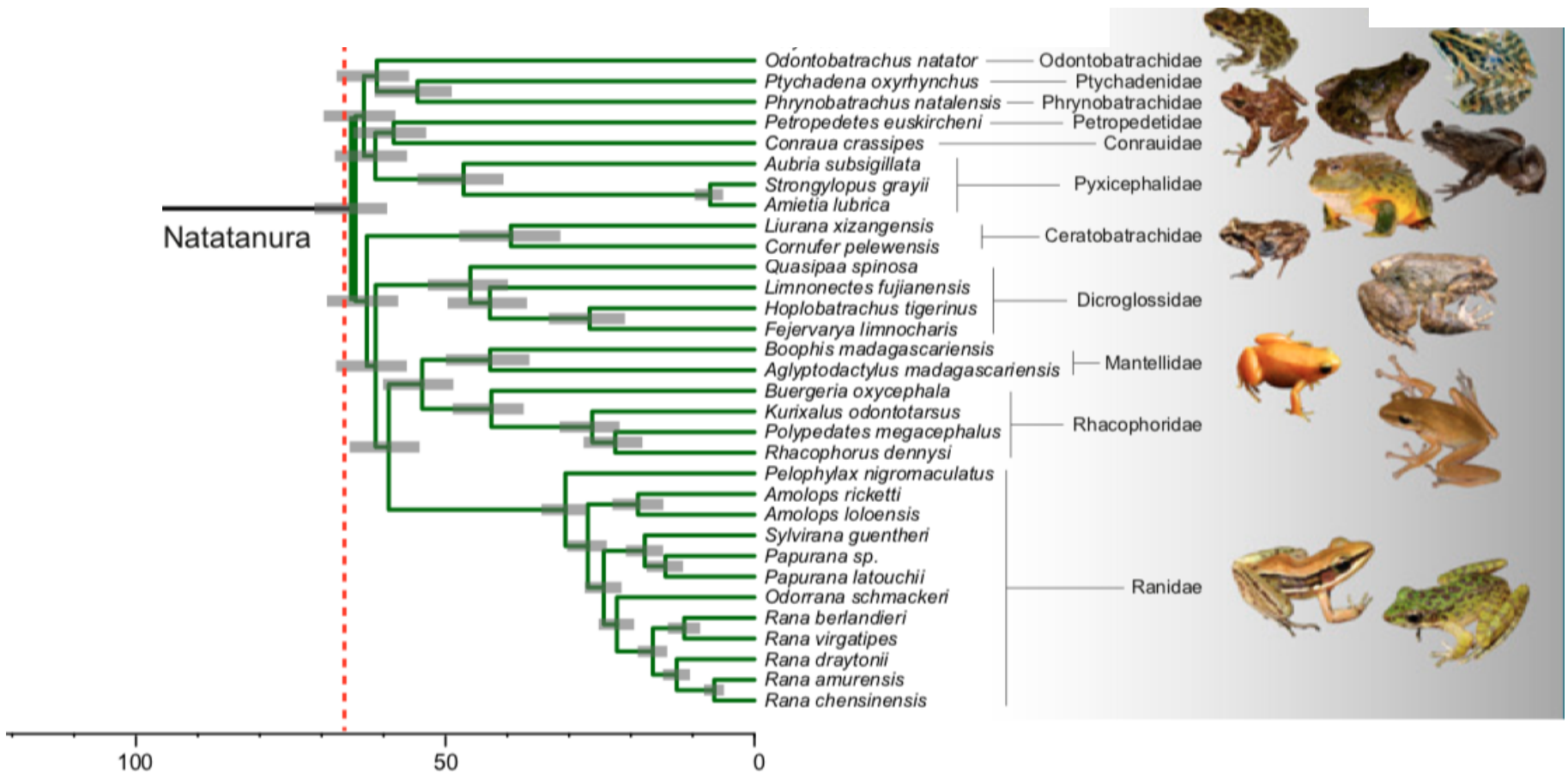
What is a phylogenetic tree

Google says: *A phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities – their phylogeny – based upon similarities and differences in their physical or genetic characteristics.*



A phylogenetic tree [Springer et al 2015, MPE]

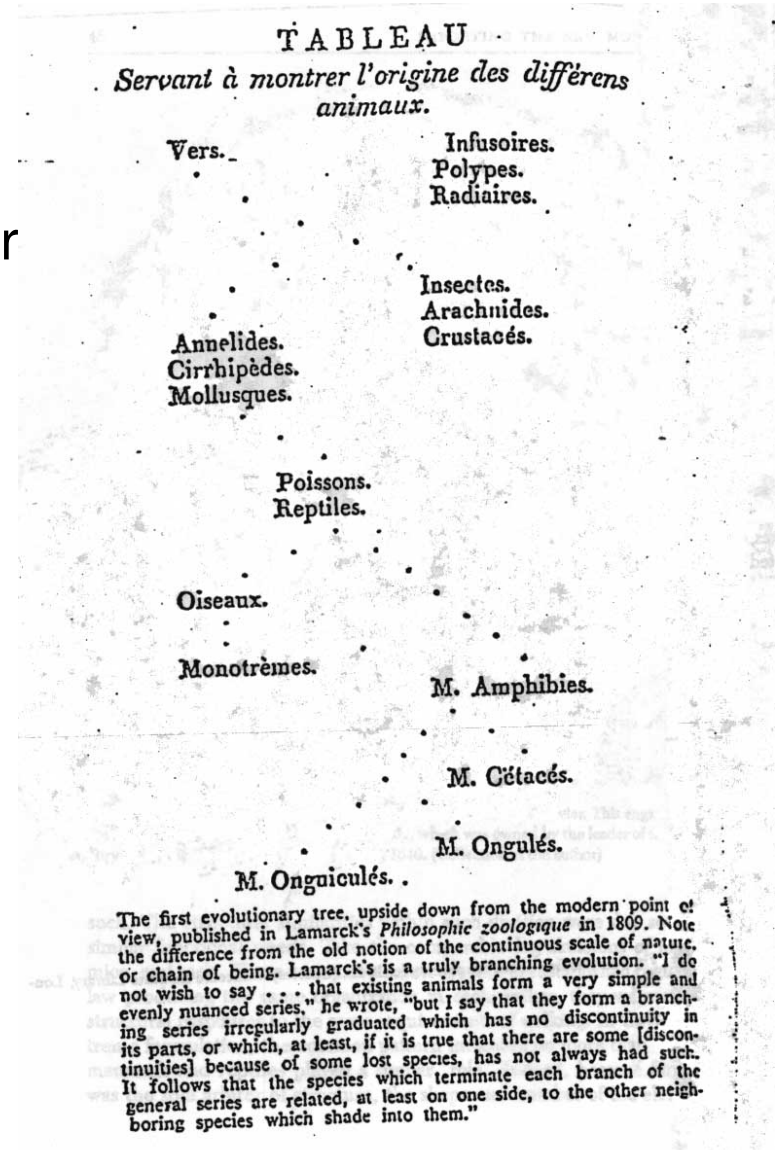
What is a phylogenetic tree



A phylogenetic tree [Feng et al 2017, PNAS]

History of phylogenetic trees

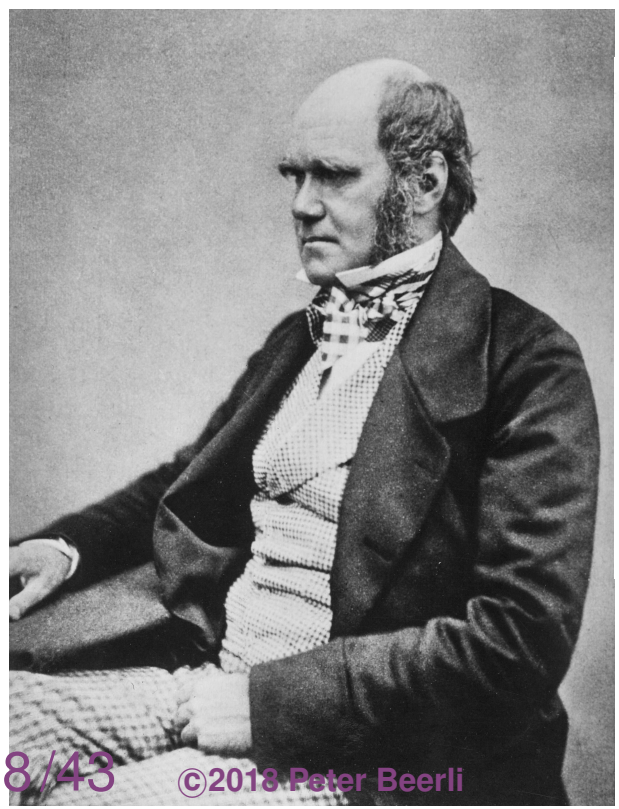
Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck (1 August 1744-18 December 1829)

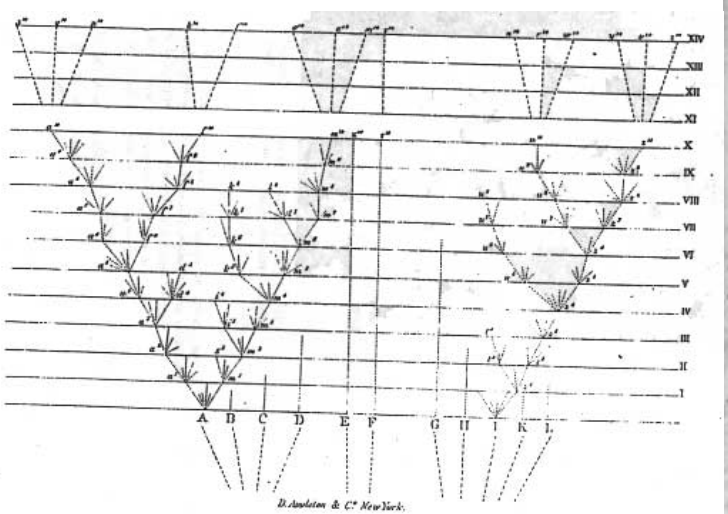


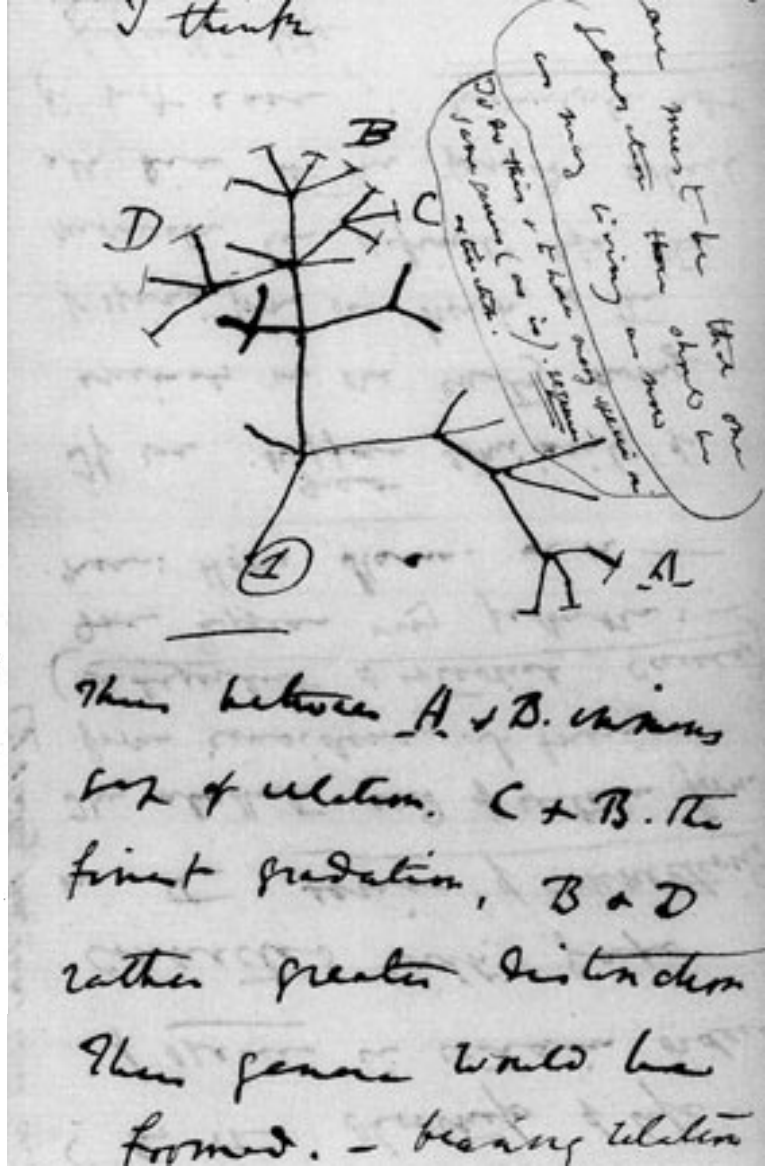
The first evolutionary tree, upside down from the modern point of view, published in Lamarck's *Philosophic zoologique* in 1809. Note the difference from the old notion of the continuous scale of nature, or chain of being. Lamarck's is a truly branching evolution. "I do not wish to say . . . that existing animals form a very simple and evenly nuanced series," he wrote, "but I say that they form a branching series irregularly graduated which has no discontinuity in its parts, or which, at least, if it is true that there are some [discontinuities] because of some lost species, has not always had such. It follows that the species which terminate each branch of the general series are related, at least on one side, to the other neighboring species which shade into them."

History of phylogenetic trees

Charles Robert Darwin
(12 February 1809 - 19 April 1882)

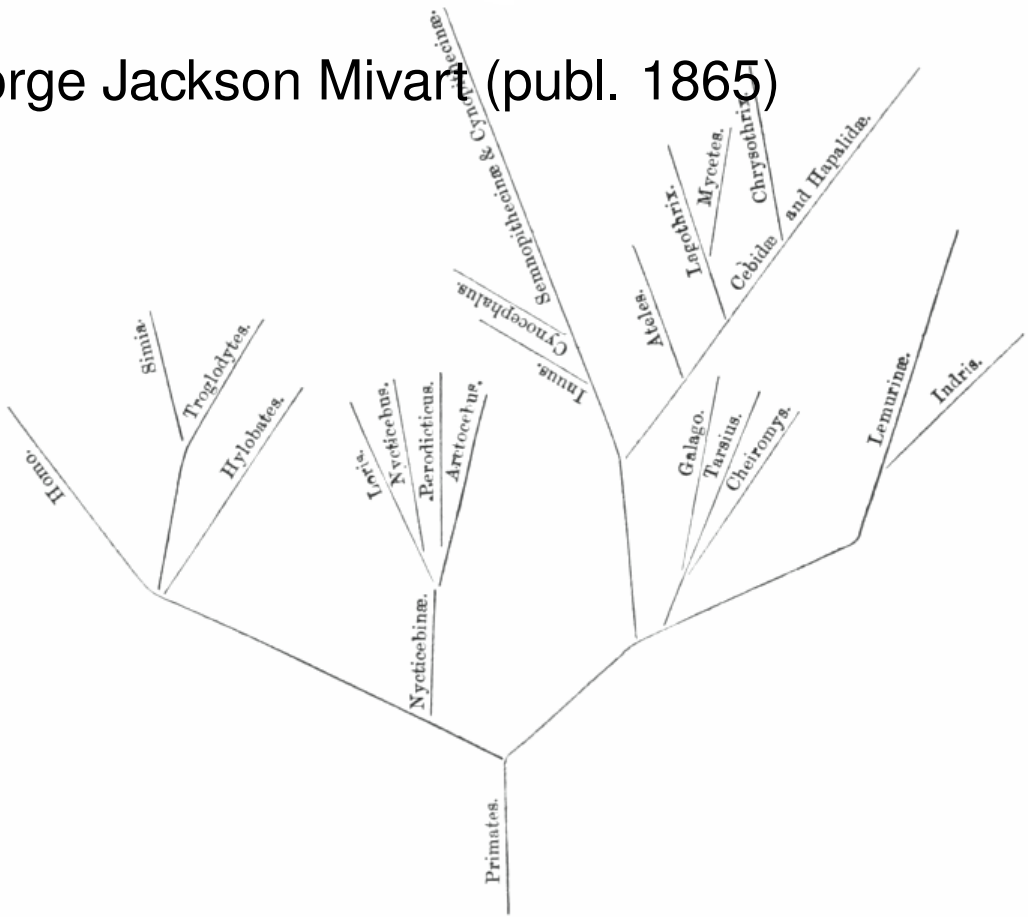




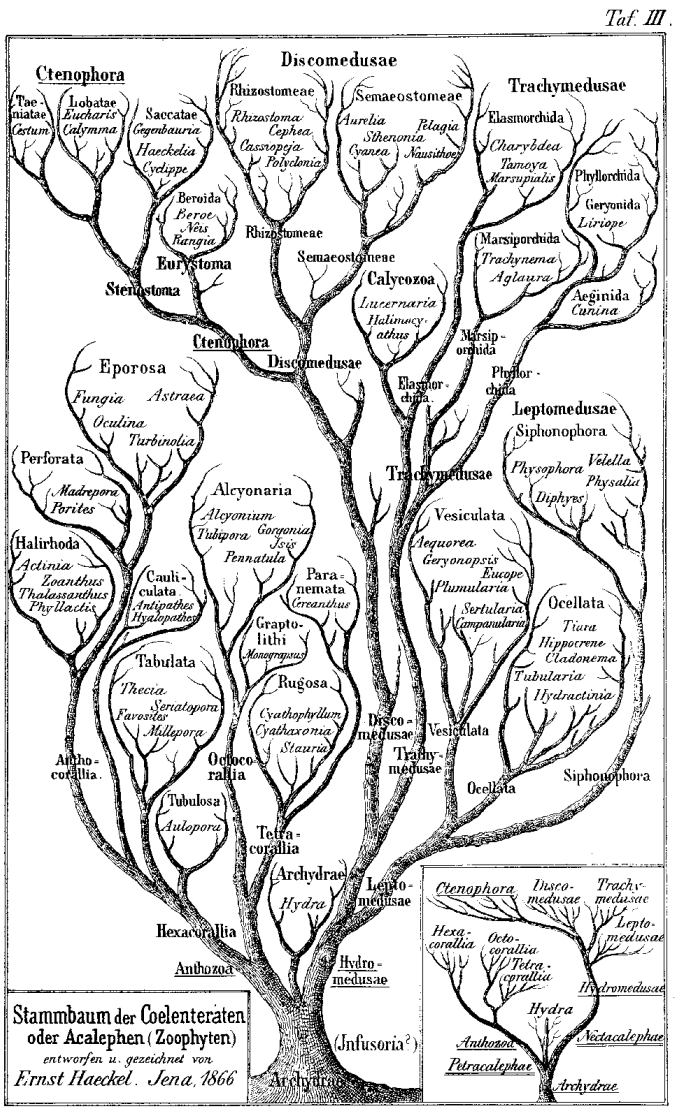


History of phylogenetic trees

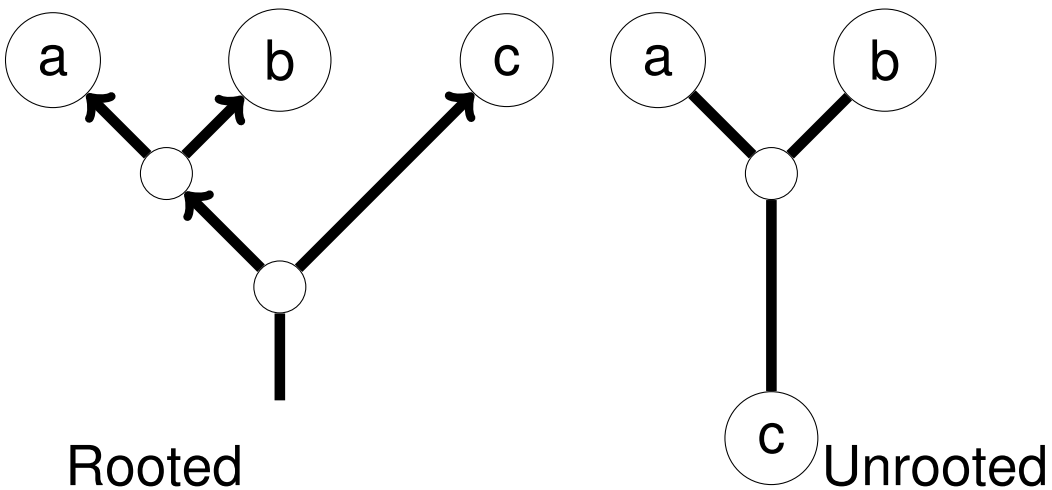
St George Jackson Mivart (publ. 1865)



Ernst Haeckel (publ. 1866)

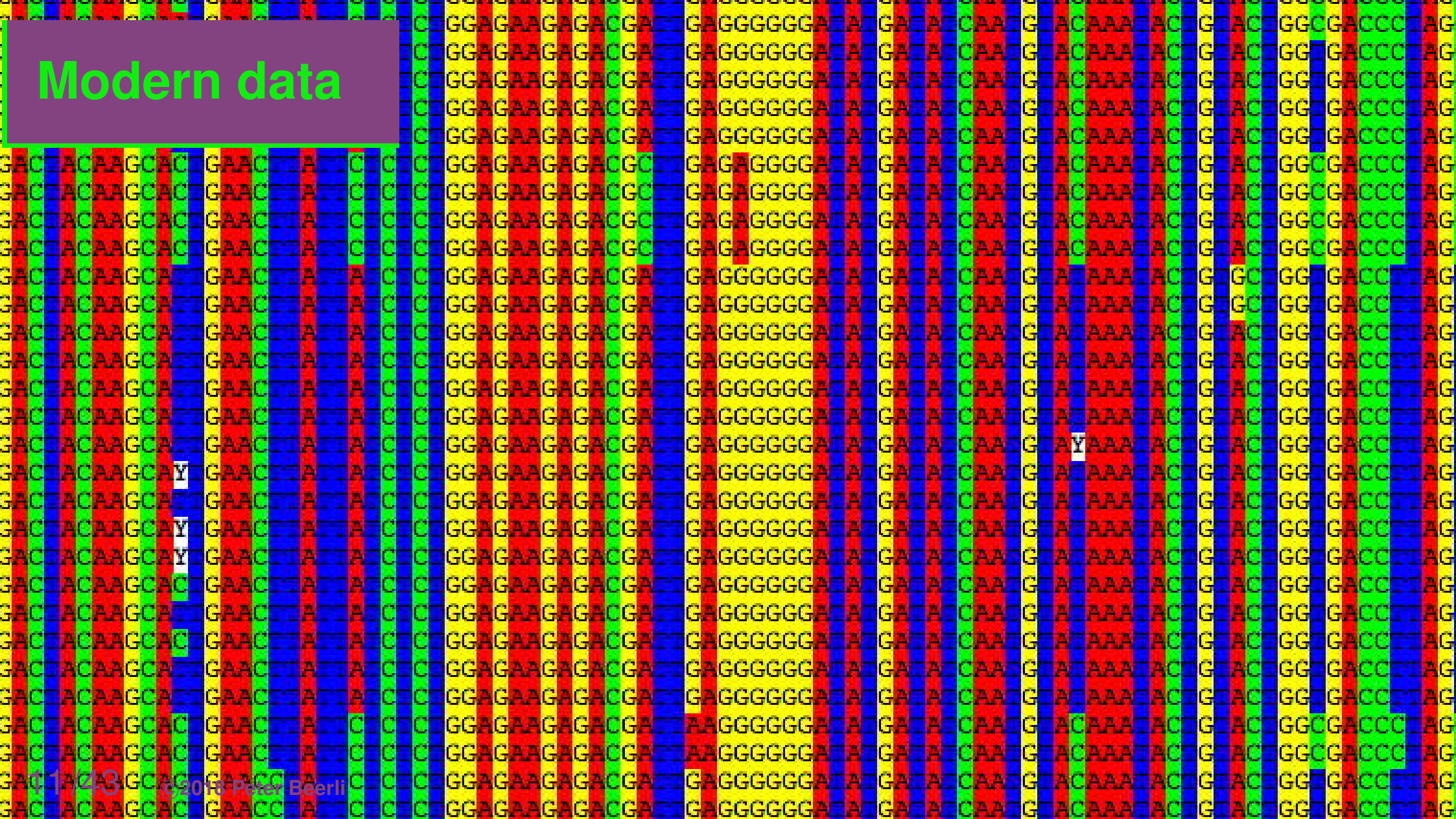


Modern phylogenetic tree



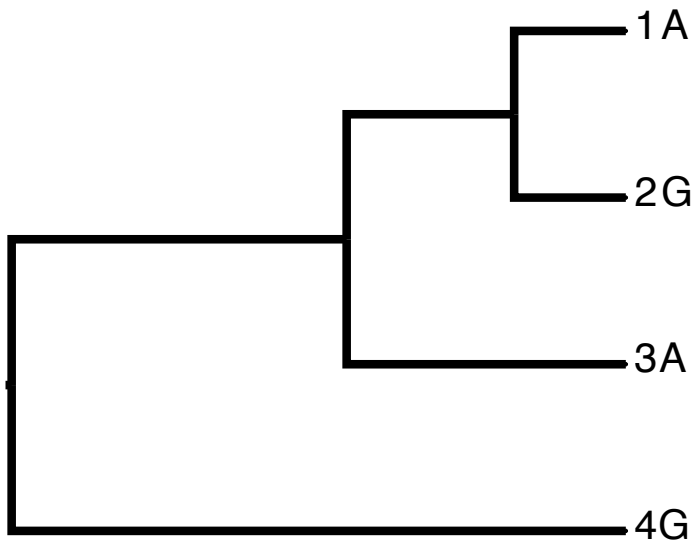
- rooted: directed, acyclic graph
- unrooted: acyclic graph
- leaves are labeled [tips], internal nodes are usually unlabelled [interior], edges [branches]

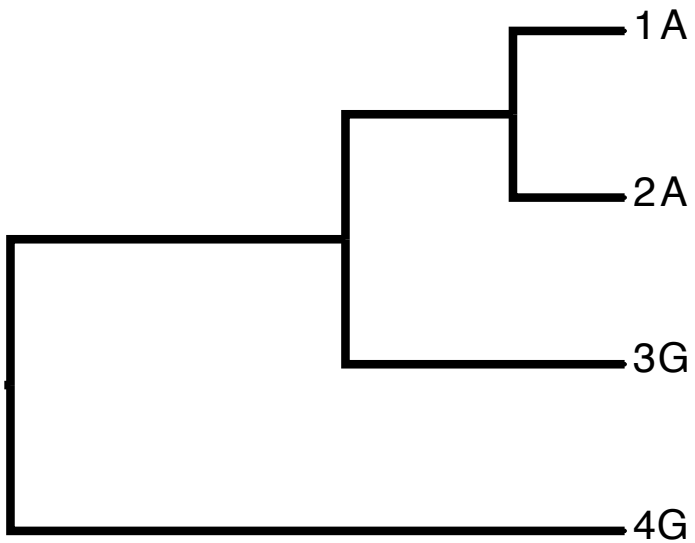
Modern data



Parsimony

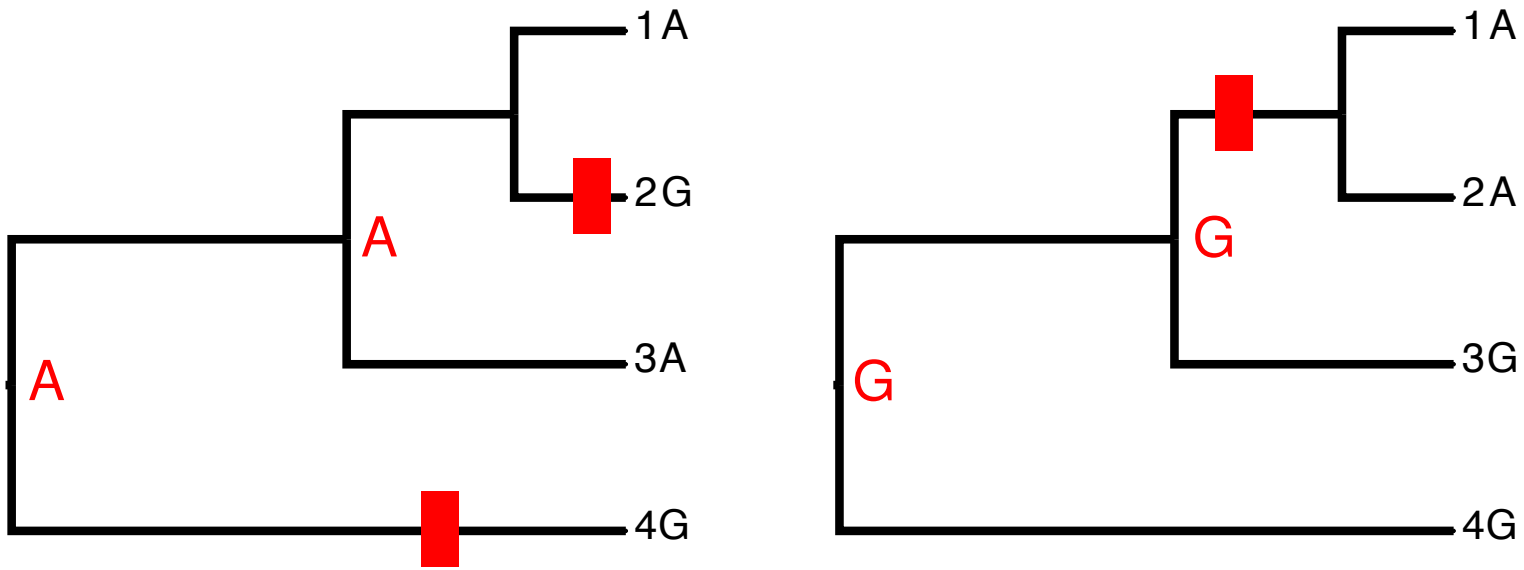
Maximum parsimony is an optimality criterion under which the phylogenetic tree that minimizes the total number of character-state changes is to be preferred.





Occam's razor (also Ockham's razor; Latin: *lex parsimoniae* "law of parsimony") is a problem-solving principle attributed to William of Ockham (c. 1287-1347), who was an English Franciscan friar, scholastic philosopher, and theologian. His principle states that among competing hypotheses, the one with the fewest assumptions should be selected or when you have two competing theories that make exactly the same predictions, the simpler one is the better.

Parsimony



Using the Fitch-algorithm, we calculate 2 necessary changes on the left and 1 change on the right tree (the right tree is better)

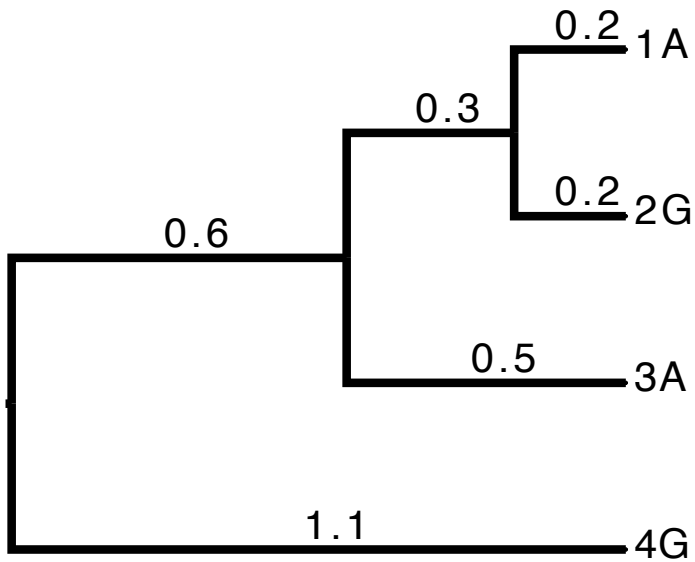
Maximum Likelihood

Maximum likelihood is an optimality criterion under which the phylogenetic tree that maximizes the probability of the data fitting the tree is to be preferred.

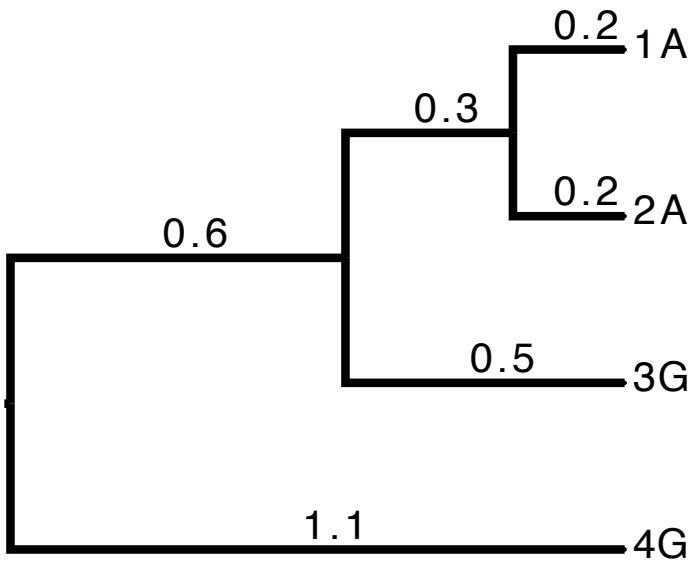
- Model how a nucleotide changes into others given time t (or branch length), Usually we assume that mutation are rare and use a Poisson process, we will only need to know the last state. Information on the DNA has no memory.
- Frequency of nucleotides at the root
- The matrix Q describes all rates of change among nucleotides (many specific models are available: JC, K81, F81, F84, HKY, TN94, GTR, ...)

Maximum Likelihood

$$Q_{JC} = \begin{pmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{pmatrix}$$



$\ln L = -6.21$



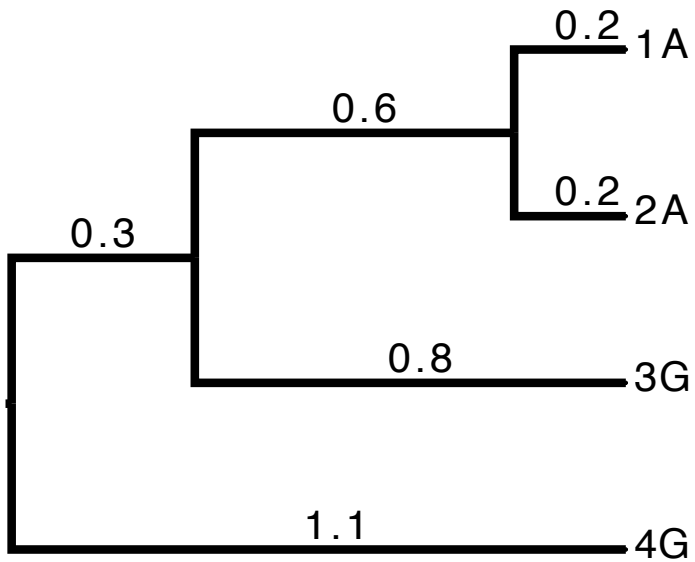
$\ln L = -4.91$

$$L = \pi \cdot (((c_1 \cdot e^{Qb_1} \odot c_2 \cdot e^{Qb_2}) \odot c_3 \cdot e^{Qb_3}) \odot c_4 \cdot e^{Qb_4})$$

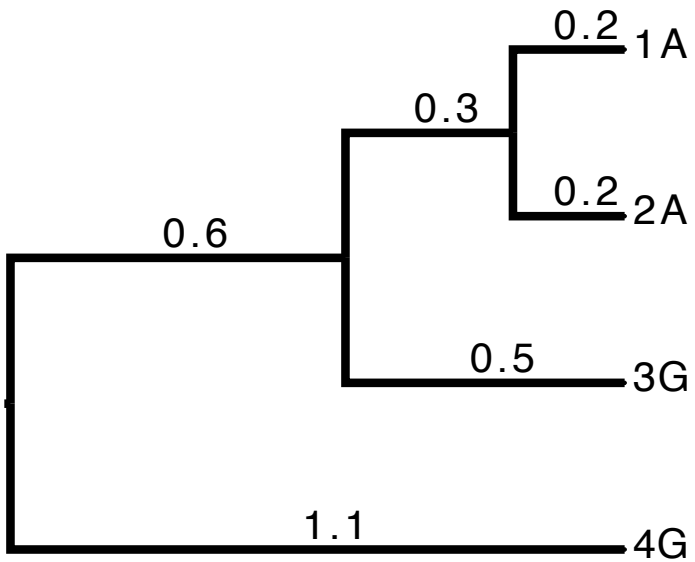
$$c_i = [\ell_A, \ell_C, \ell_G, \ell_T]$$

Maximum Likelihood

$$Q = \begin{pmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{pmatrix}$$



$\ln L = -4.64$

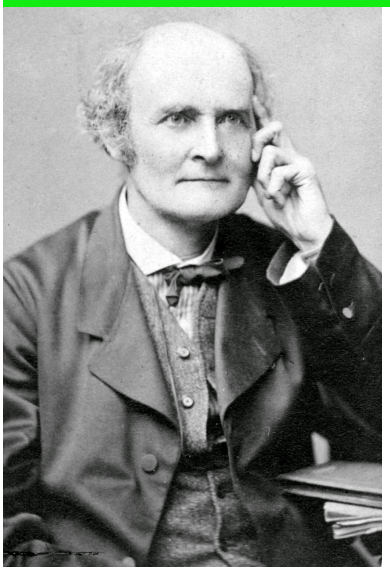


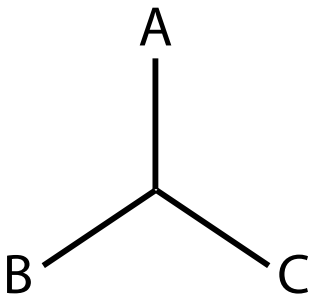
$\ln L = -4.91$

$$L = \pi \cdot (((c_1 \cdot e^{Qb_1} \odot c_2 \cdot e^{Qb_2}) \odot c_3 \cdot e^{Qb_3}) \odot c_4 \cdot e^{Qb_4})$$

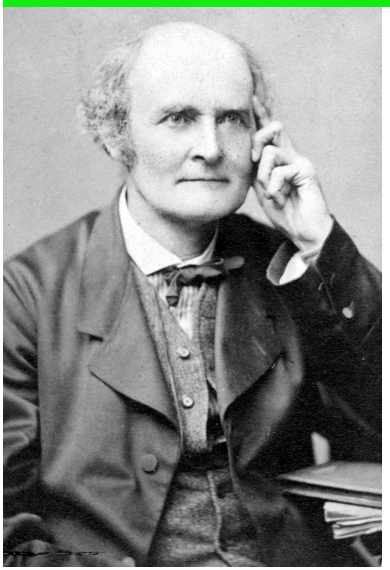
$$c_i = [\ell_A, \ell_C, \ell_G, \ell_T]$$

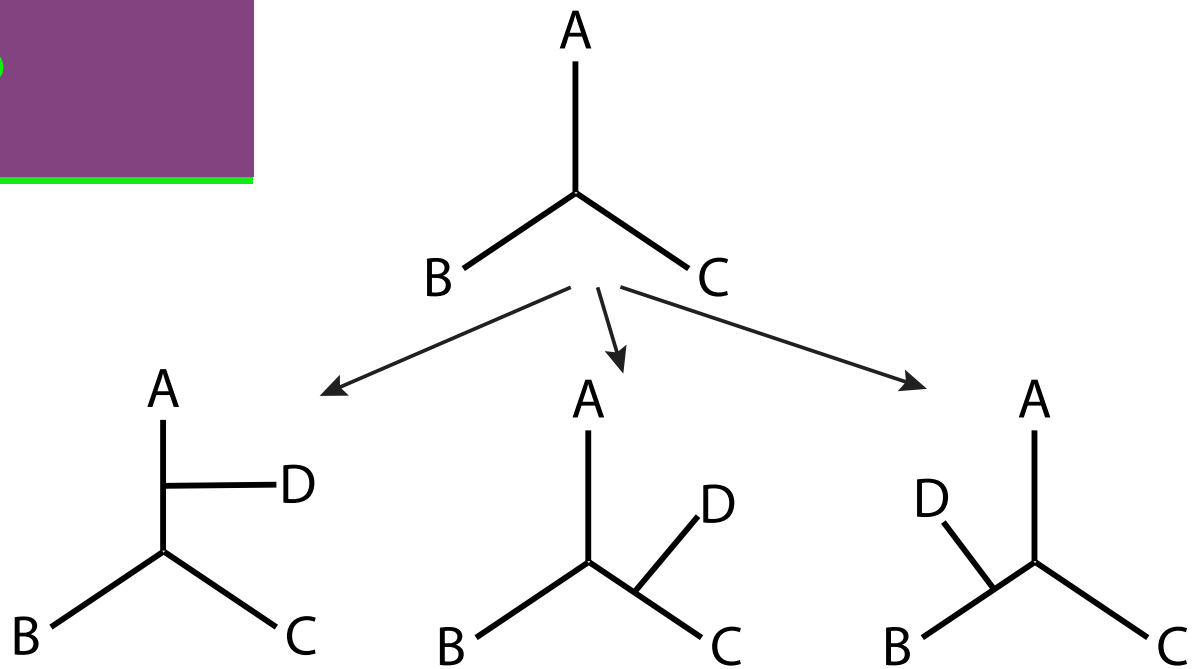
How many trees?



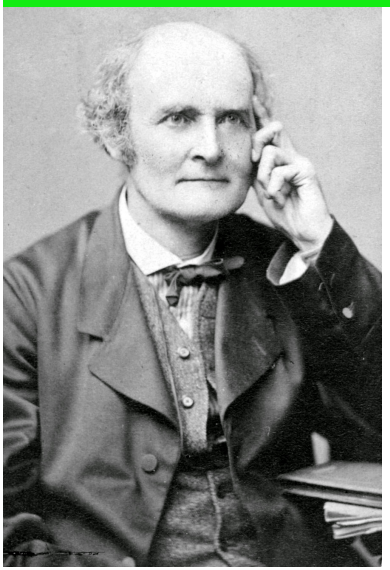


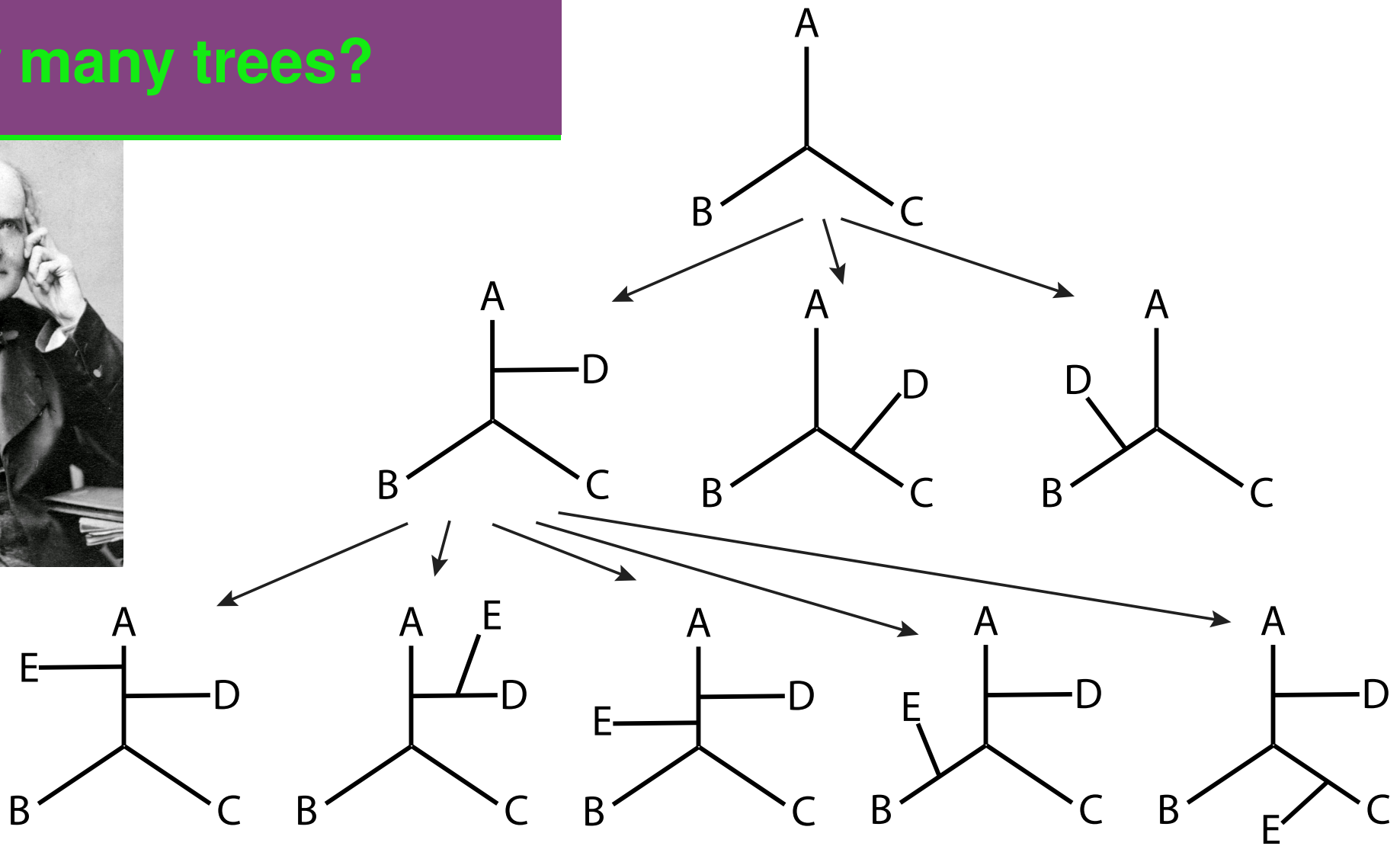
How many trees?



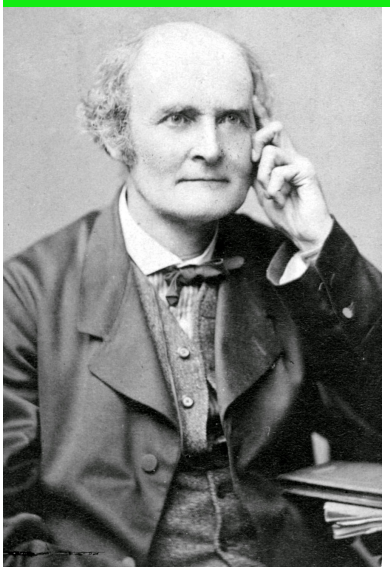


How many trees?

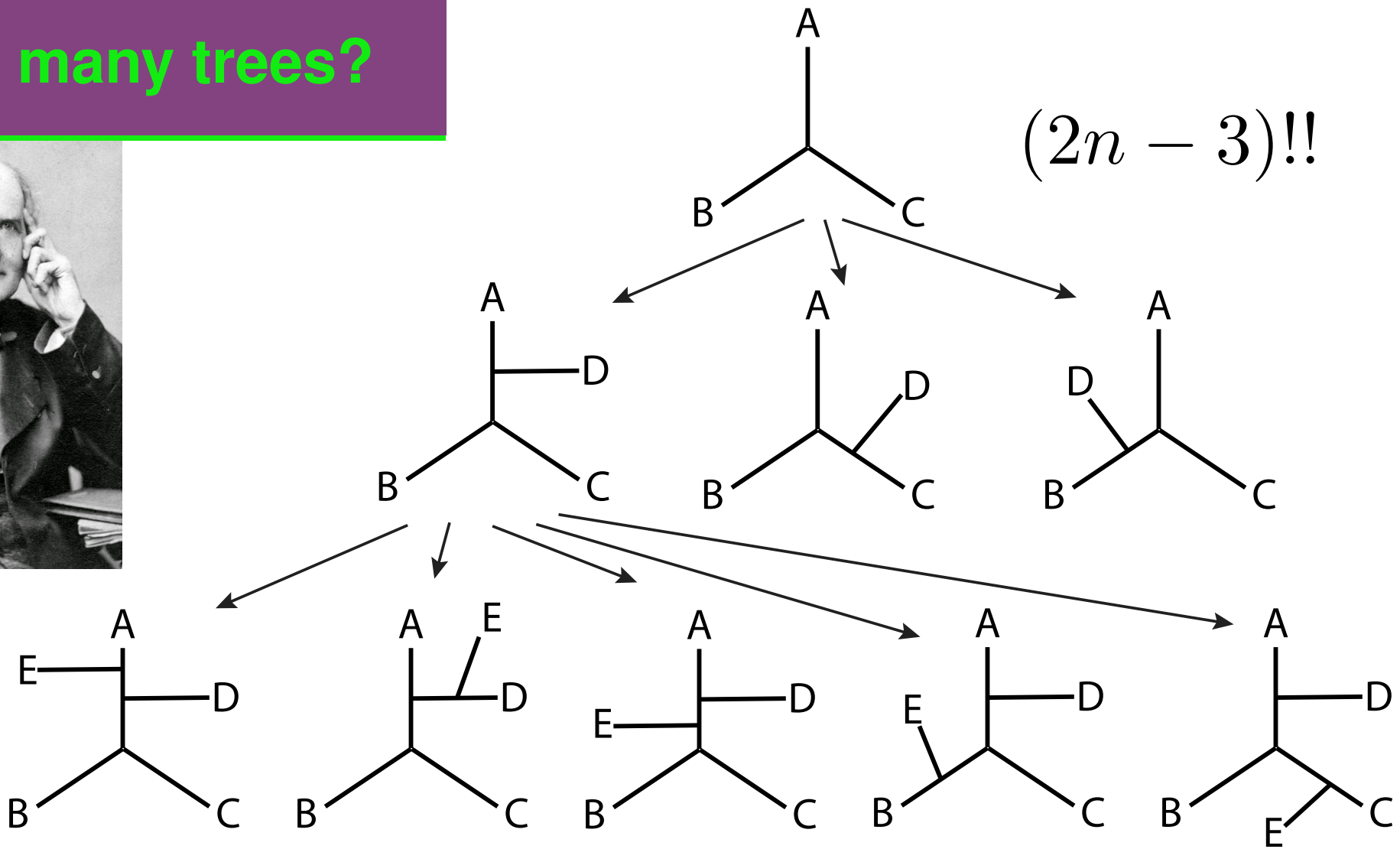




How many trees?



$$(2n-3)!!$$



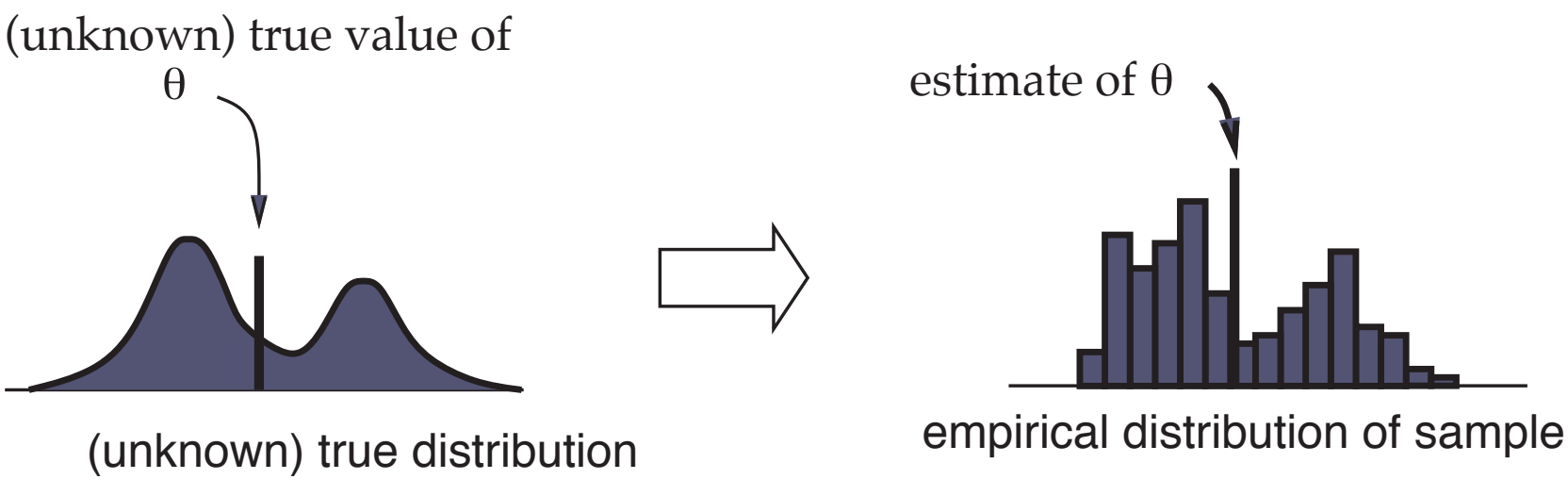
How many?

Tips	Number of labeled histories	Number of rooted trees
2	1	1
3	3	3
4	18	15
5	180	105
6	2700	945
7	56700	10395
8	1587600	135135
9	57153600	2027025
10	2571912000	34459425
11	141455160000	654729075
12	9336040560000	13749310575
13	728211163680000	316234143225
14	66267215894880000	7905853580625
15	6958057668962400000	213458046676875
16	834966920275488000000	6190283353629375
17	113555501157466368000000	191898783962510625
18	17373991677092354304000000	6332659870762850625
19	2970952576782792585984000000	221643095476699771875
20	564480989588730591336960000000	82007945326378915593 75

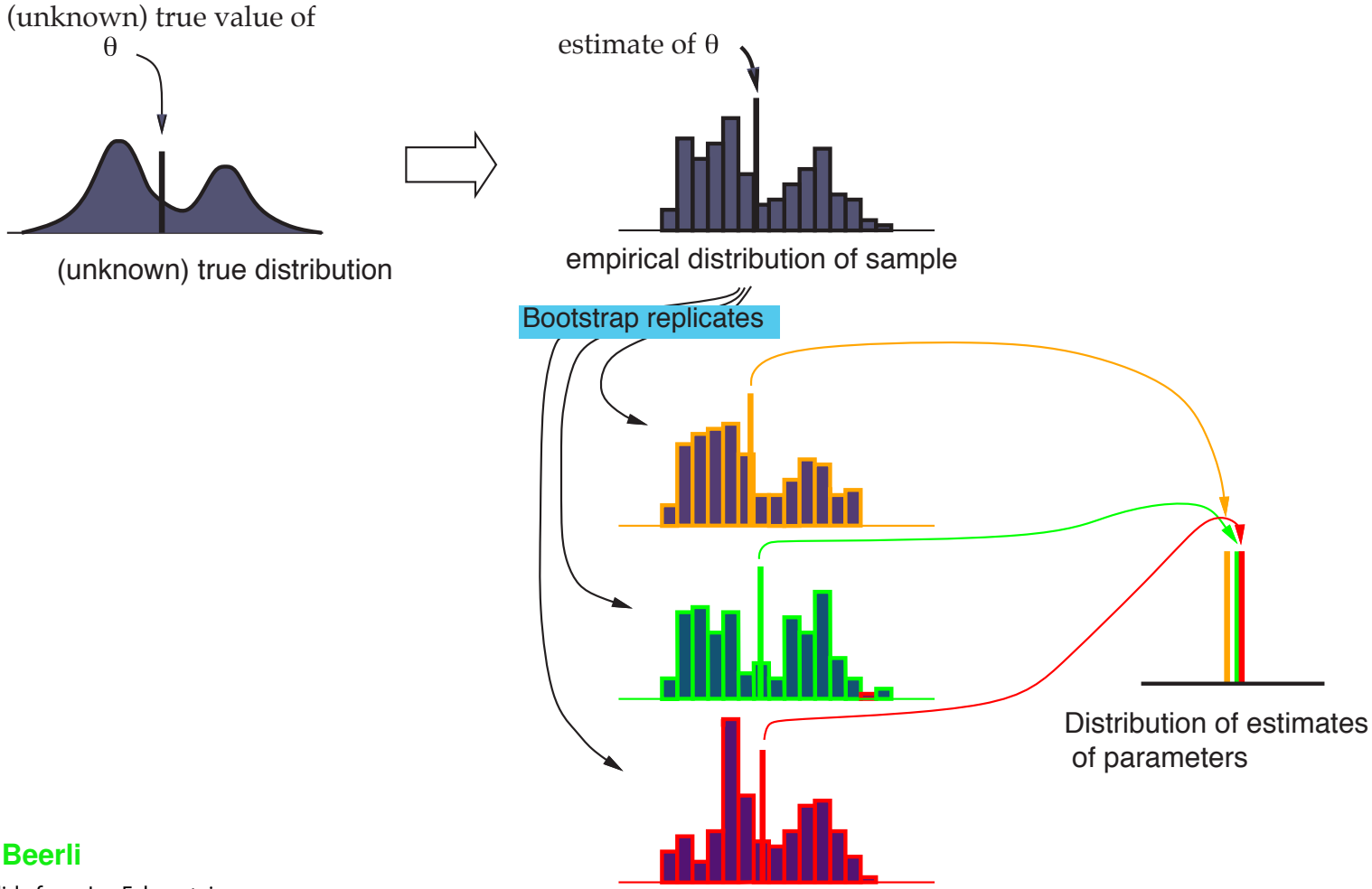
Searching tree space

Bootstrapping phylogenies

The bootstrap

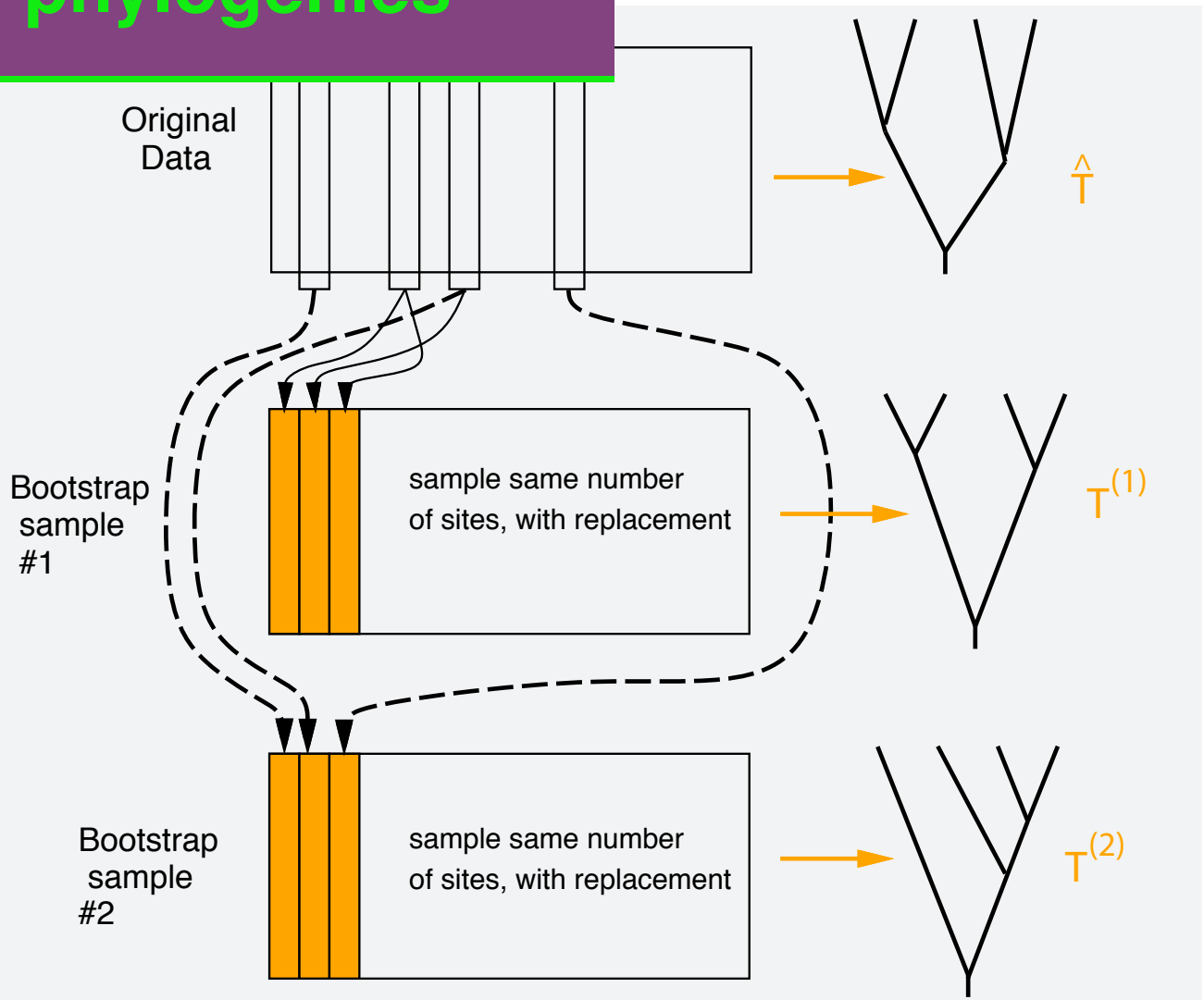


Bootstrapping phylogenies



Slide from Joe Felsenstein

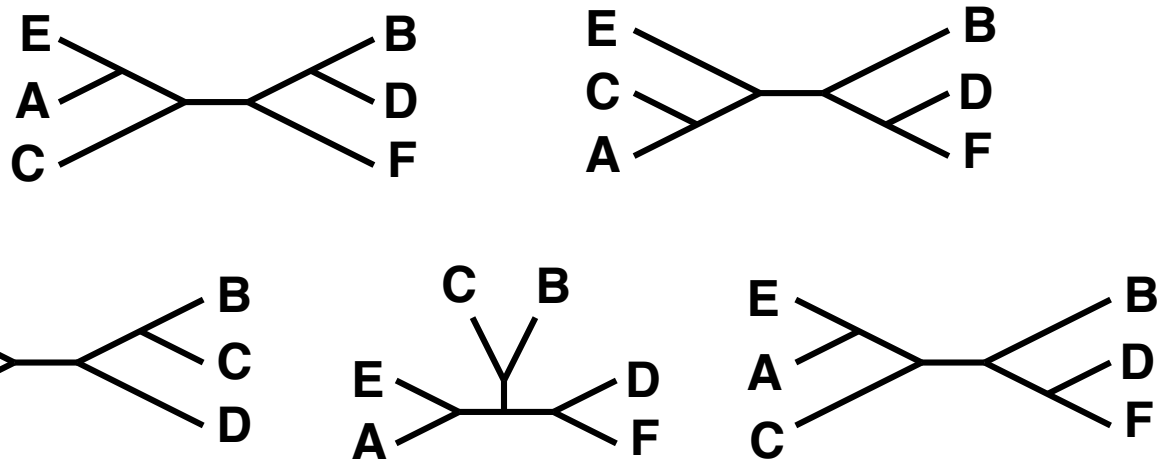
Bootstrapping phylogenies



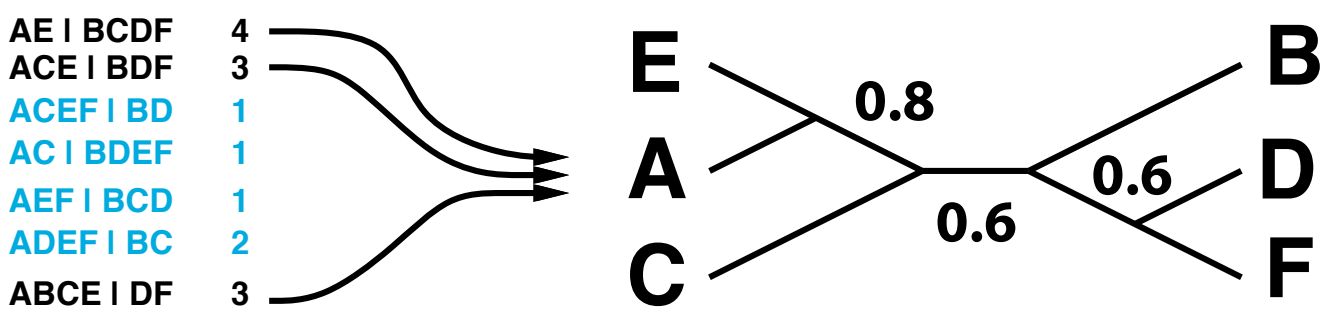
Bootstrapping

The majority-rule consensus tree

Trees:



How many times each partition of species is found:



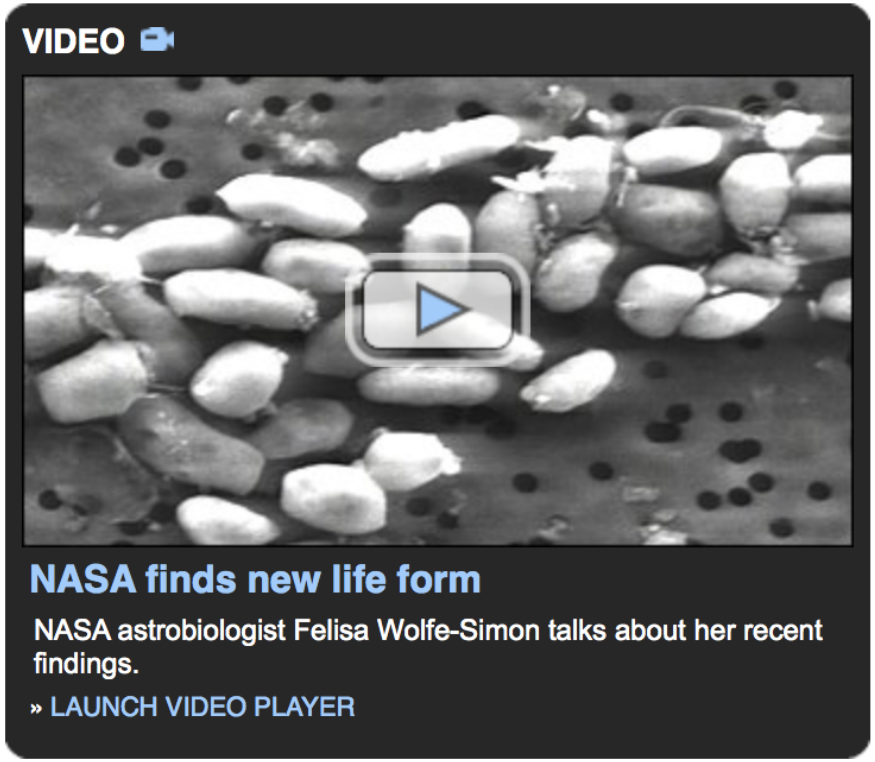
Tree of Life

NASA announcement: Arsenic-based life form discovered on Earth

Washington Post Staff
washingtonpost.com
Thursday, December 2, 2010; 4:00 PM

NASA held a press conference Thursday afternoon in which they revealed the discovery of arsenic-based life forms on Earth. As Marc Kaufman explained:

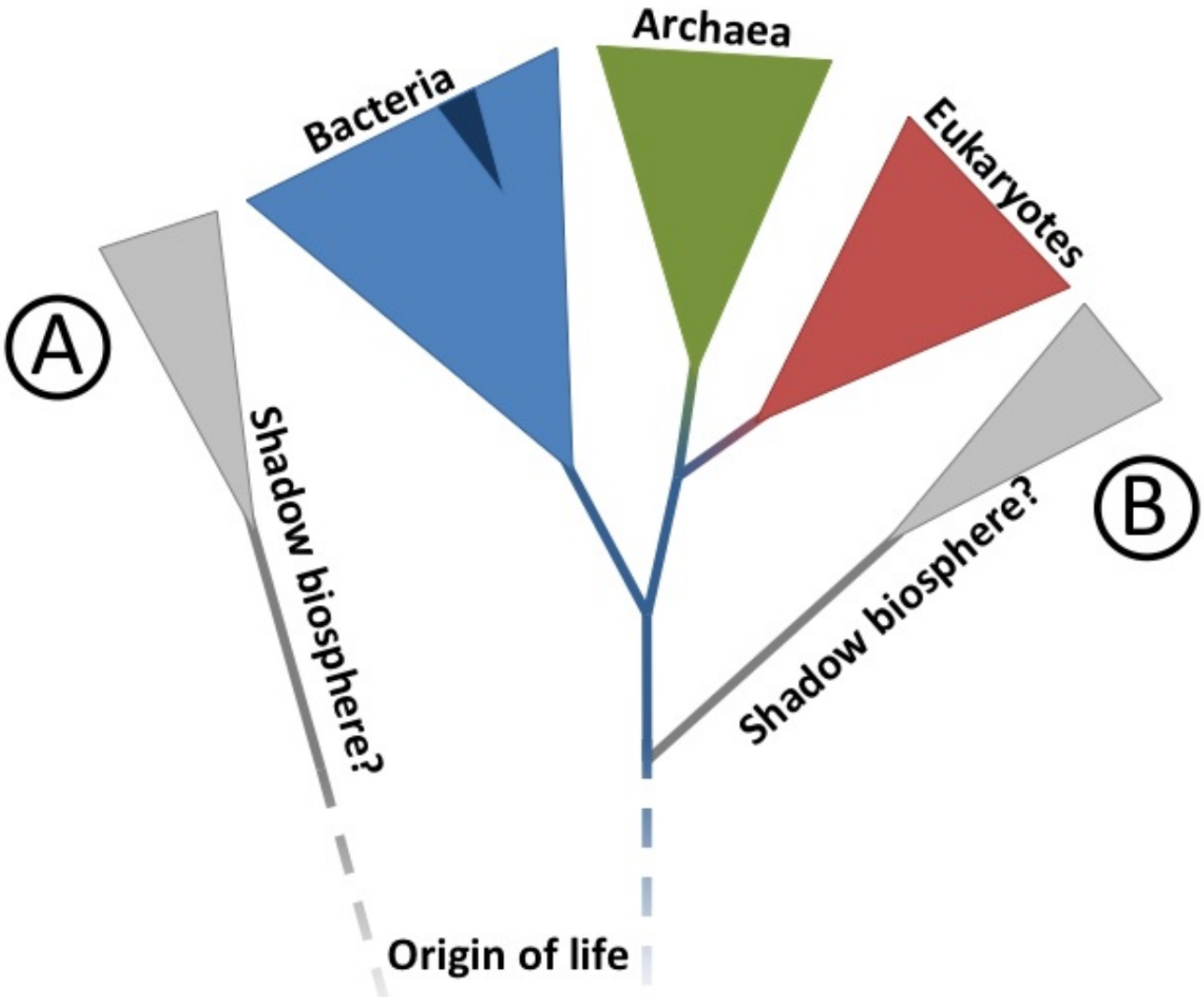
> All life on Earth - from microbes to elephants and us - is based on a single genetic model that requires the element phosphorus as one of its six essential components.



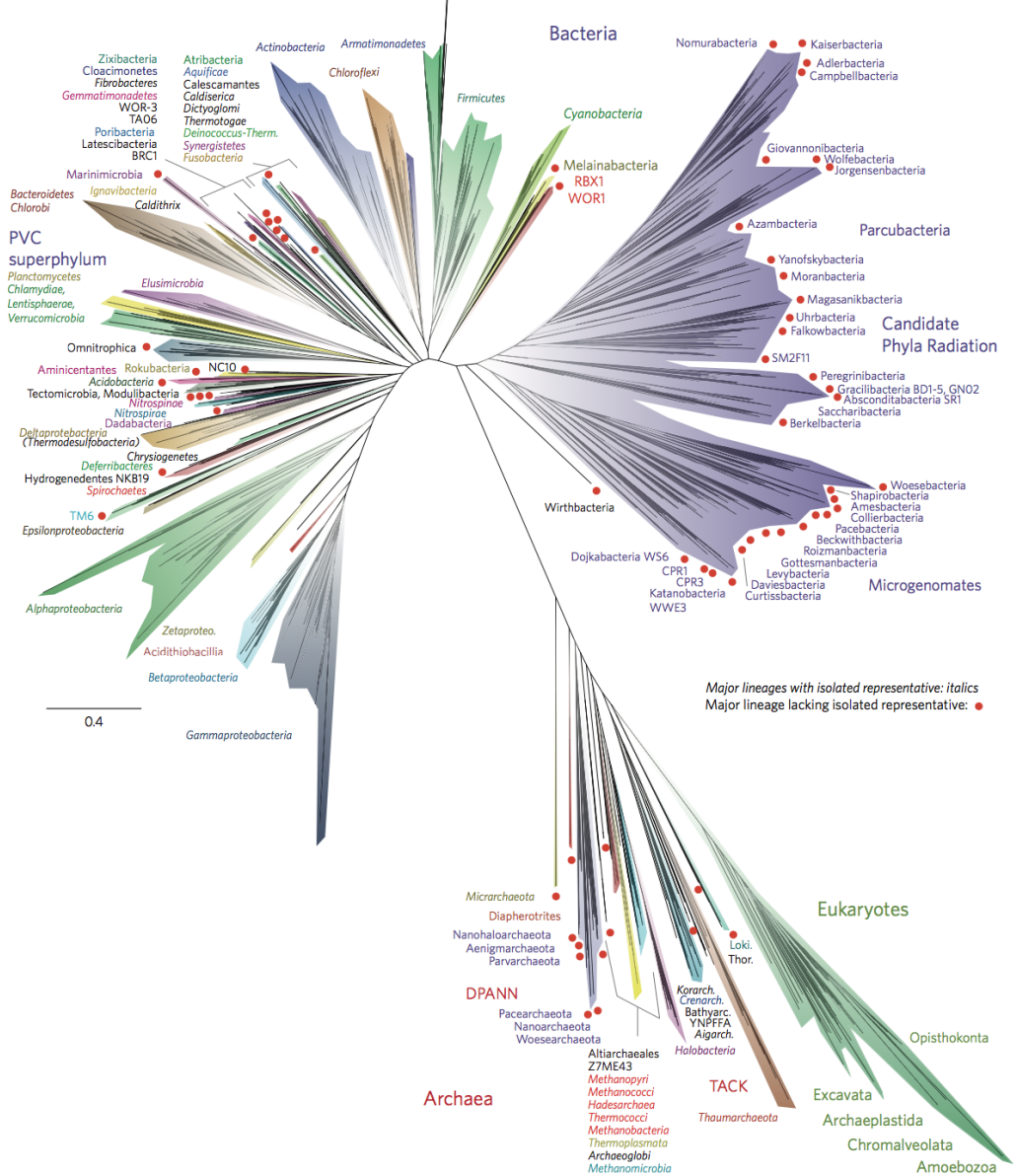
But now researchers have uncovered a bacterium that has five of those essential elements but has, in effect, replaced phosphorus with its look-alike but toxic cousin arsenic.

News of the discovery caused a scientific commotion, including calls to NASA from the White House and Congress asking whether a second line of

Tree of Life



Tree of Life



Tree of Life

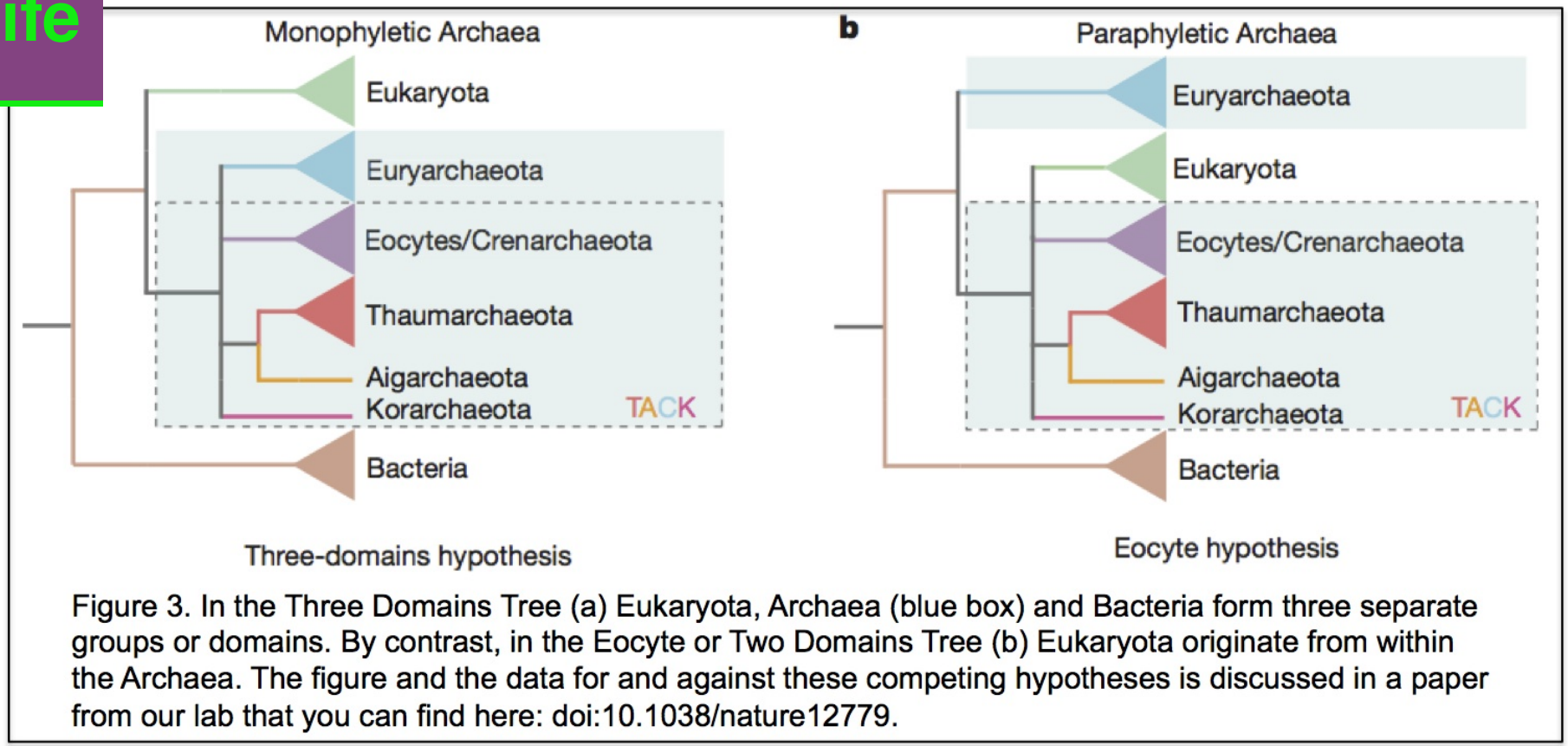


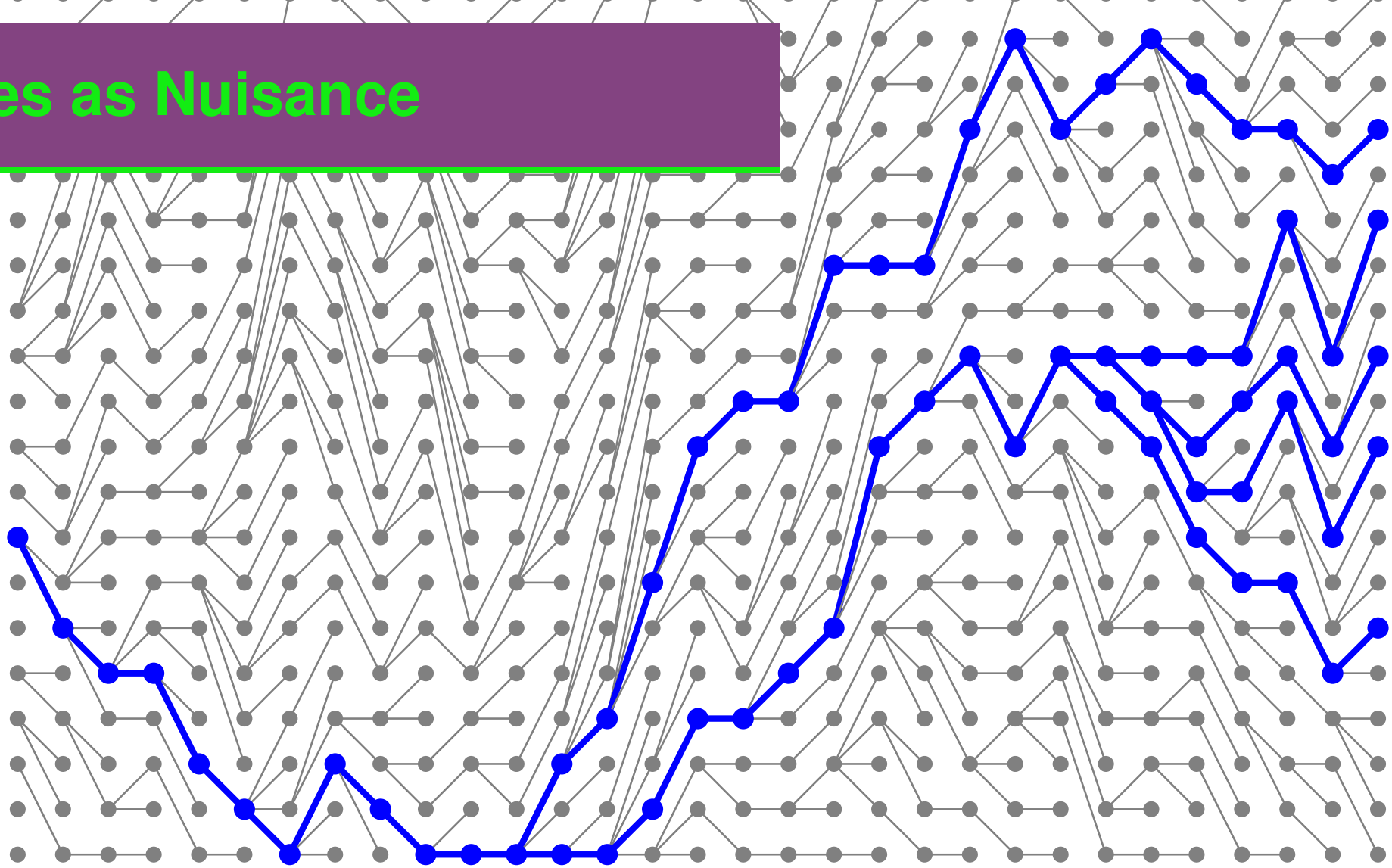
Figure 3. In the Three Domains Tree (a) Eukaryota, Archaea (blue box) and Bacteria form three separate groups or domains. By contrast, in the Eocyte or Two Domains Tree (b) Eukaryota originate from within the Archaea. The figure and the data for and against these competing hypotheses is discussed in a paper from our lab that you can find here: doi:10.1038/nature12779.

Embley 2015: https://blogs.ncl.ac.uk/icamblog/tag/tree-of-life/

Tree of Life: just in case you wondered:

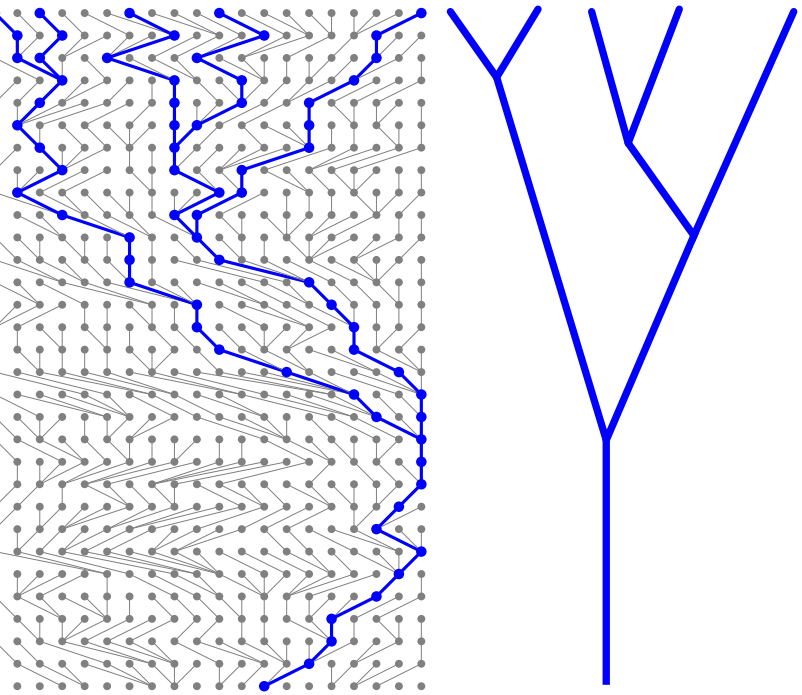
We and Bananas are about 50% similar

Trees as Nuisance



sample genealogy in a Wright-Fisher population

Kingman's n-coalescent



$$\mathrm{f}(G|\Theta) = \prod_{j=0}^{T} e^{-u_j \frac{k_j(k_j-1)}{\Theta}} \frac{2}{\Theta}$$

The parameter Θ is the mutation-scaled population size and is of interest. Using Bayesian inference and MCMC, we calculate its posterior density from sequence data D with

$$g(\Theta|D) = \frac{p(\Theta) \int_G f(G|\Theta) P(D|G) dG}{\int_\theta p(\Theta) \int_G f(G|\Theta) P(D|G) dG d\Theta}$$

Estimation of population size of vulnerable species

Fractional calculus meets the Coalescent, soon

The Fractional Coalescent (V1.6)

Somayeh Mashayekhi[*,1] **and Peter Beerli**[†]
[*†]Department of Scientific Computing, Florida State University, Tallahassee, FL 32306

ABSTRACT A new approach to the coalescent, the *fractional* coalescent (f-coalescent), is introduced. Two derivations are presented: first, the f-coalescent is based on an extension of the Canning population model that the variance of the number of offspring is assumed as a random variable. Second, the f-coalescent emerges as a continuous-time semi-Markov process. The f-coalescent extends Kingman's n-coalescent by introducing an additional parameter α that affects the variability in the patterns of the waiting times; values of $\alpha < 1$ lead to an increase of short lineages, but allows occasionally for very long lineages. When $\alpha = 1$, the f-coalescent and the Kingman's n-coalescent are equivalent. The f-coalescent has been implemented in the population genetic model inference software MIGRATE. Simulation studies suggest that it is possible to infer the correct α values from data that was generated with known α values. When data is simulated using models with $\alpha < 1$ or for three example datasets (HIV-1 influenza, Malaria parasites, Humpback whales), Bayes factor comparisons show an improved model

Phylogeny in Court

Court of Appeal of Louisiana,Third Circuit.

STATE of Louisiana v. Richard J. SCHMIDT.

No. 99-1412.

Decided: July 26, 2000

(Court composed of Judge HENRY L. YELVERTON, Judge BILLIE COLOMBARO WOODARD, Judge GLENN B. GREMILLION). Michael Harson, District Attorney, Keith A. Stutes, Assistant District Attorney, Lafayette, LA, Counsel for Plaintiff/Appellee. Michael S. Fawer, Covington, LA, Gerald J. Block, Lafayette, LA, William R. Campbell, Herbert V. Larson, Jr., New Orleans, LA, Thomas E. Guilbeau, Lafayette, LA, Counsel for Defendant/Appellant.

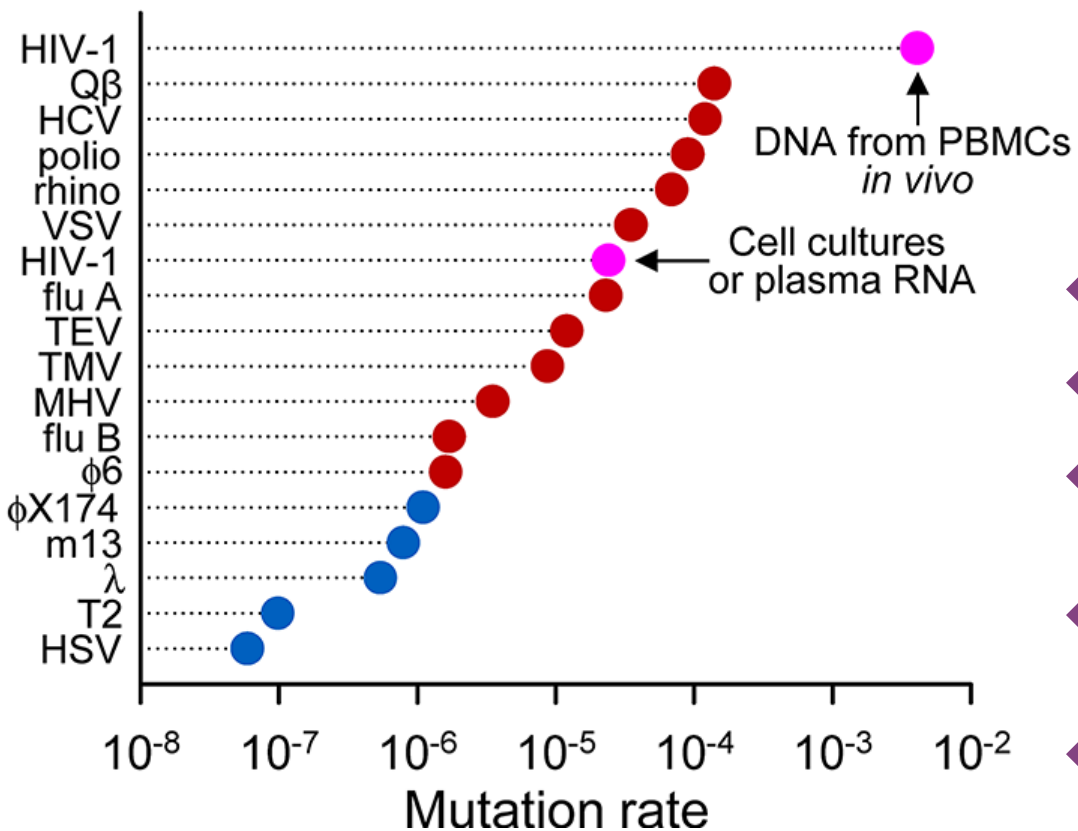
In this case, the defendant, Richard J. Schmidt, appeals his conviction and sentence of fifty years at hard labor for attempted second degree murder. For the following reasons, we affirm.

Murder with HIV?

FACTS

The State charged Defendant, a medical doctor specializing in gastroenterology, with the attempted second degree murder of Janice Trahan, a nurse, alleging that he committed the crime on August 4, 1994, by injecting the human immunodeficiency virus (HIV) into Trahan under the guise of giving her a Vitamin B-12 shot. HIV is the virus which causes acquired immune deficiency syndrome (AIDS), a disease for which there is no cure and which is ultimately fatal.[1] Trahan testified that she is now HIV-positive and suffers from Hepatitis C, and that she became infected by those diseases when Defendant injected those viruses into her on August 4, 1994. She initially sought medical attention for symptoms of a viral infection on August 16, 1994, and was informed she was HIV positive on January 3, 1995.

Mutation rate in HIV

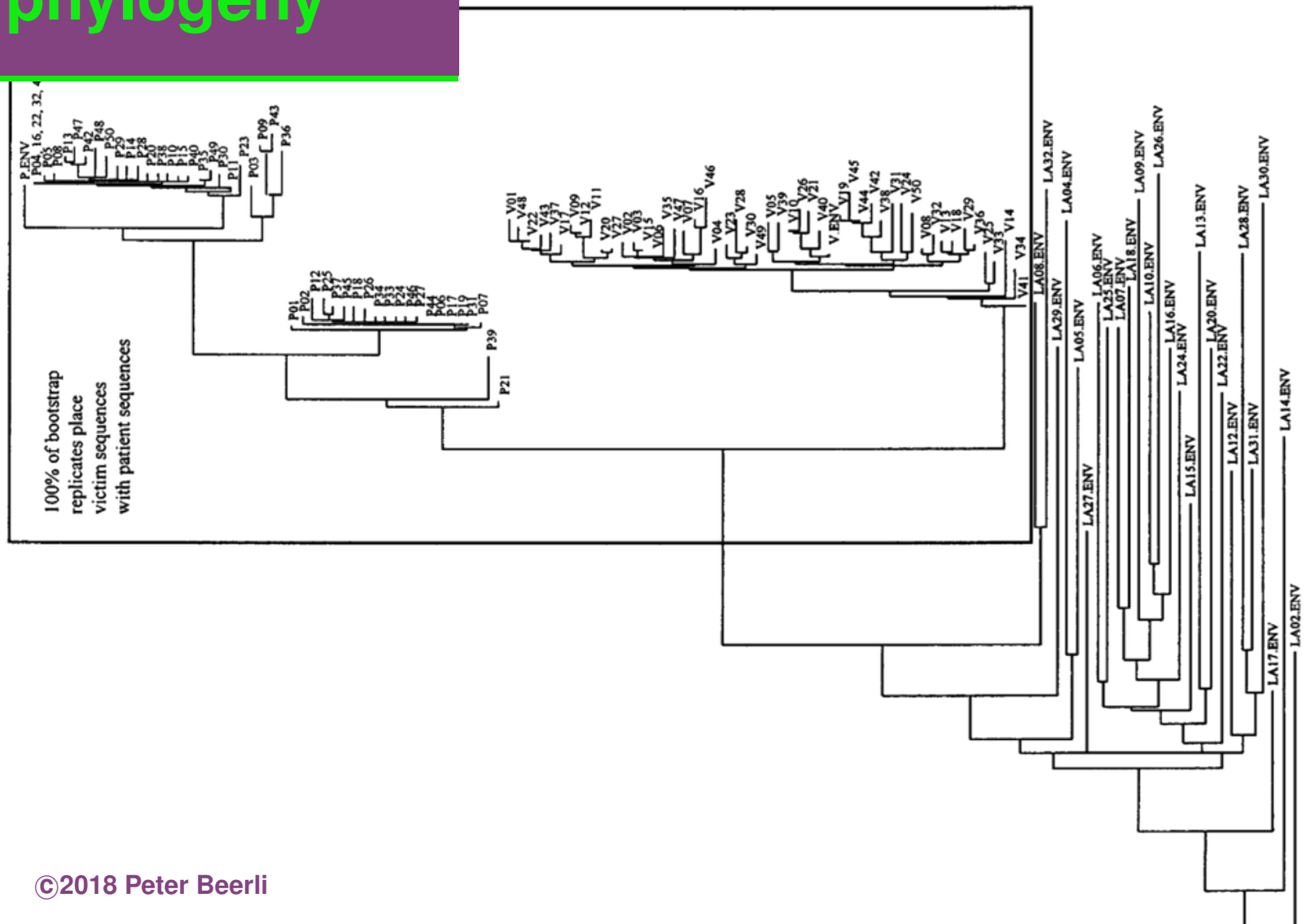


Some viruses mutate very fast (HIV, Hepatitis, Influenza This allows to construct relationship trees of virus samples Such phylogenetic trees are statistical constructs works best with clearly defined hypotheses, backed up by other evidence

- Mutation ~ 0.01 per nucleotide and generation
- Total of ~ 9800 nucleotides
- About 10 Billion new HIV particles per day in untreated patients
- $0.01 \times 9800 \times 10^9 = 98 \times 10^9$ mutations per day
- Some of these escape the immune system

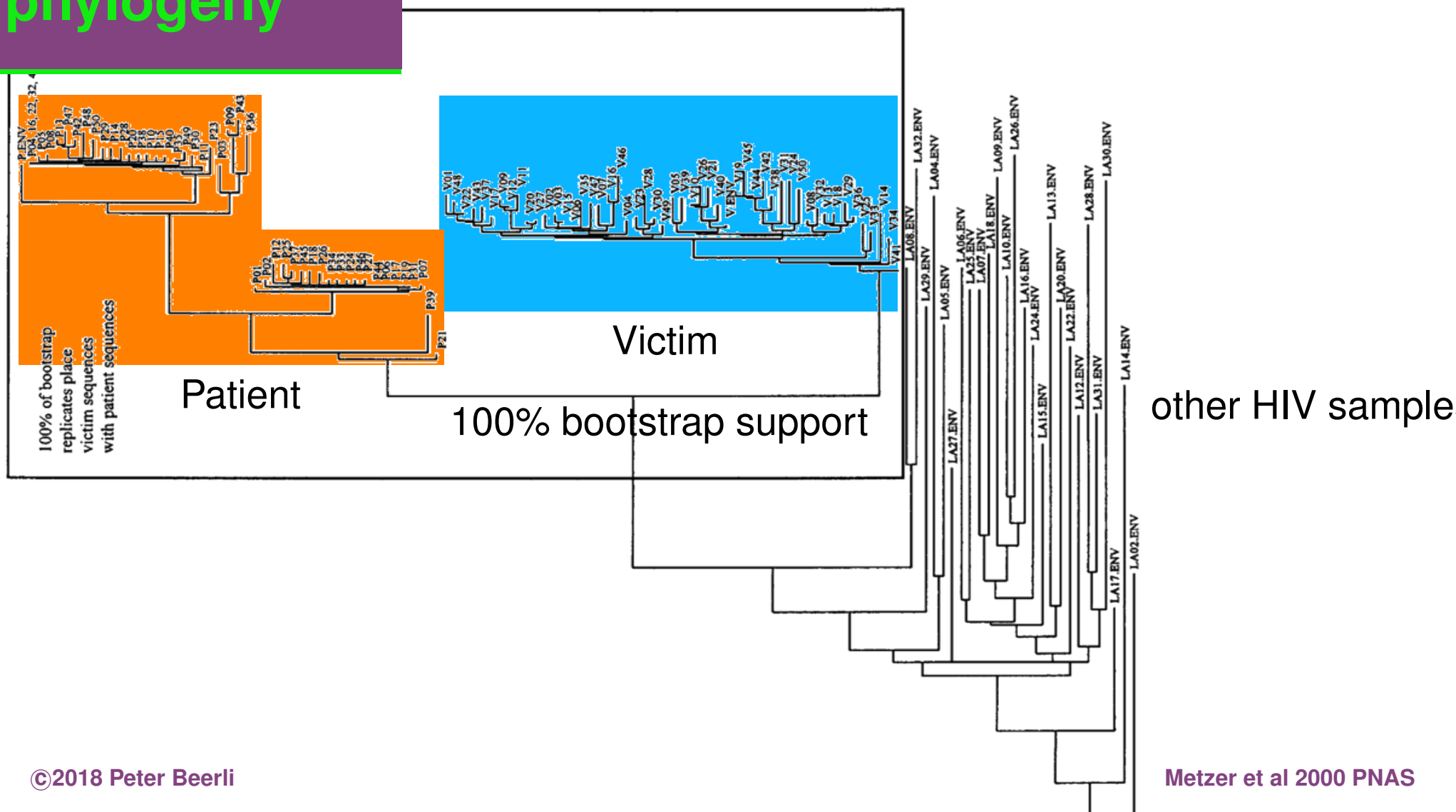
Cuevas et al. 2015 PLoS Bio

HIV phylogeny



Metzer et al 2000 PNAS

HIV phylogeny



Metzer et al 2000 PNAS

Epilog

David Hillis (he was a scientific witness in that case):

".. it is important to have a clear a priori hypothesis to test, and it is important to blind the identities of samples during the analysis. Clearly, a case cannot rely merely on phylogenetic analysis"

"There has to be clear epidemiological evidence and a criminal investigation and forensic standards of investigation must be maintained."

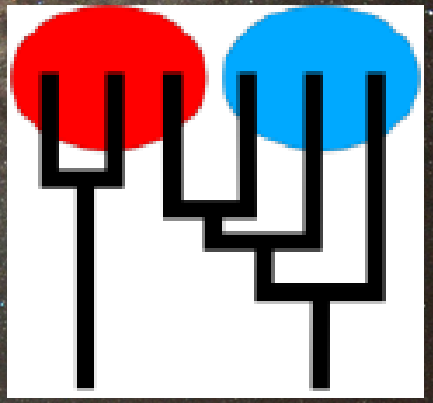
"All scientific inferences from data, of any kind, represent circumstantial evidence, by definition. That's the way the analyses are presented in court"

Thank you

Lucrezia Bieler

National Science Foundation

Somayeh Mashayekhi
Kyle Shaw
Marjan Sadeghi
Tara Khodaie



http://popgen.sc.fsu.edu