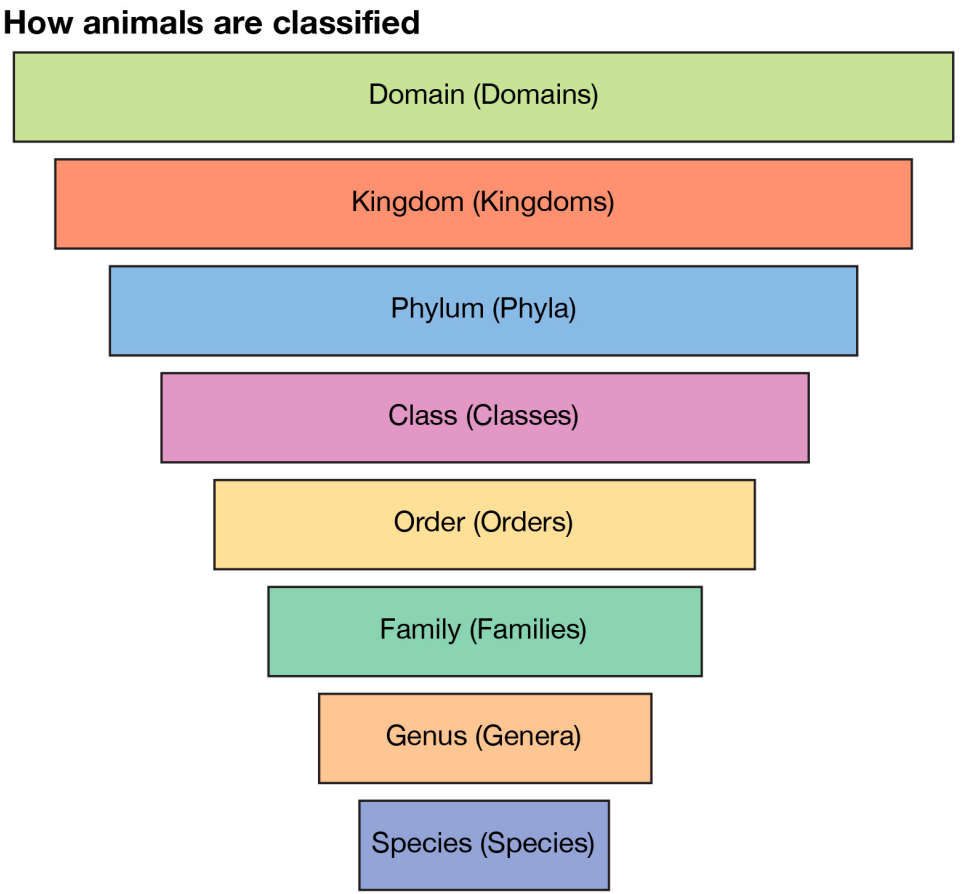
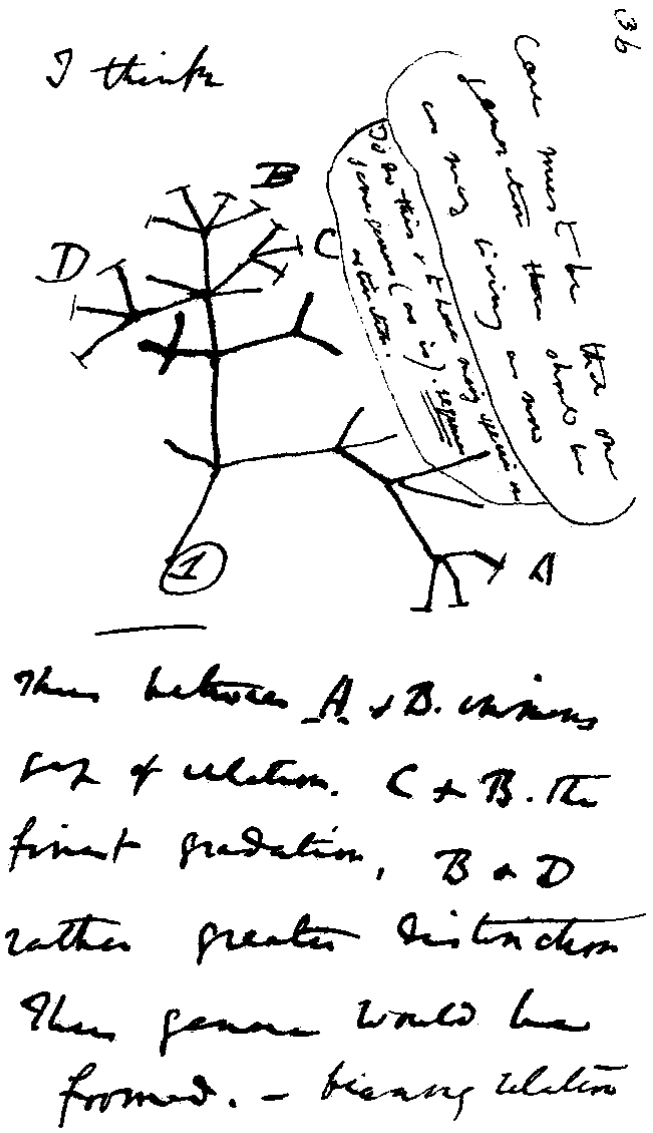
Phylogenetics

wikipedia

"Great Chain of Being"
scala naturae



Linnaean Classification

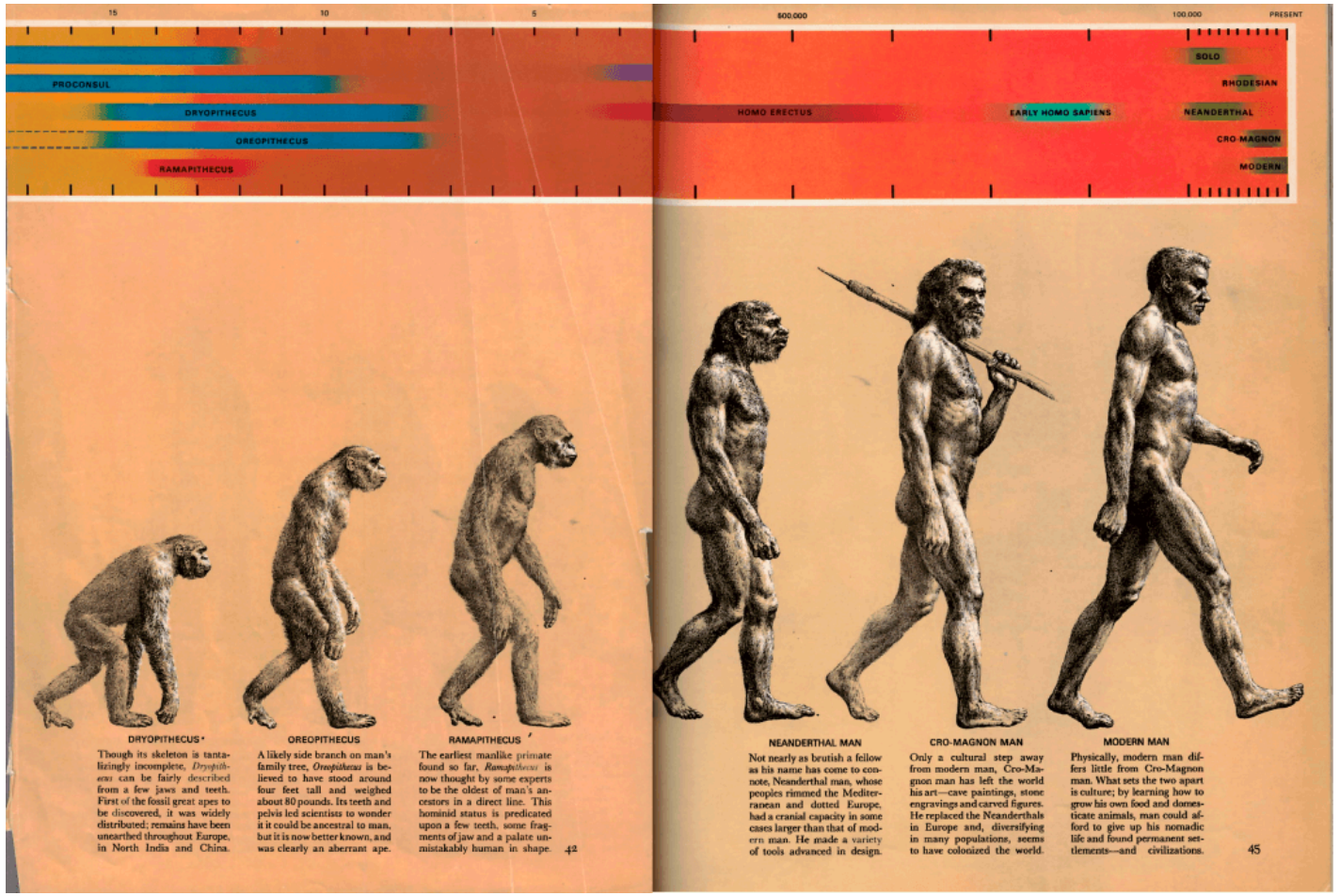


wikipedia

Charles Darwin

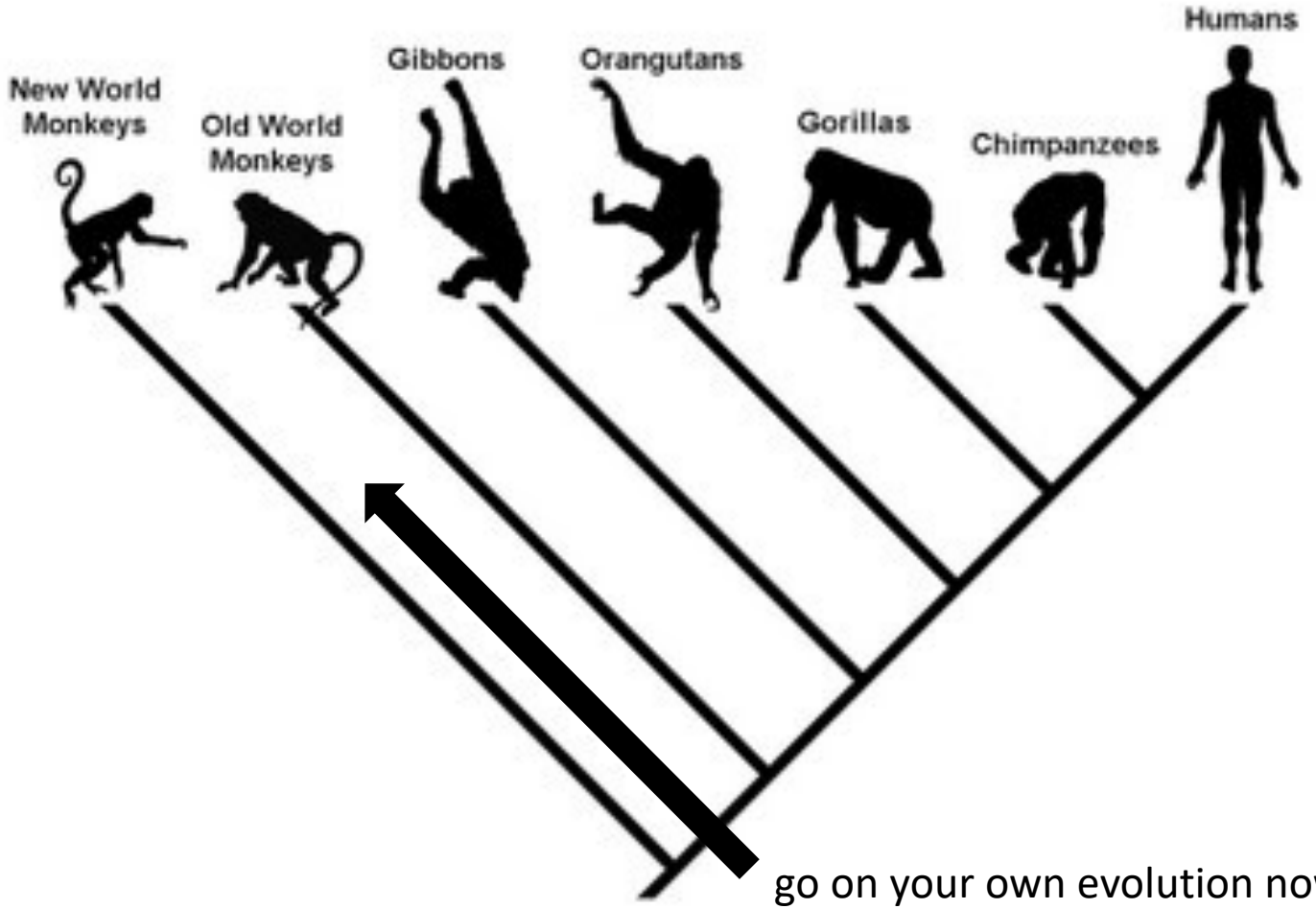
Blackwell Dictionary of Western Philosophy, p. 289 (2004)

What is **NOT** a phylogenetic tree?



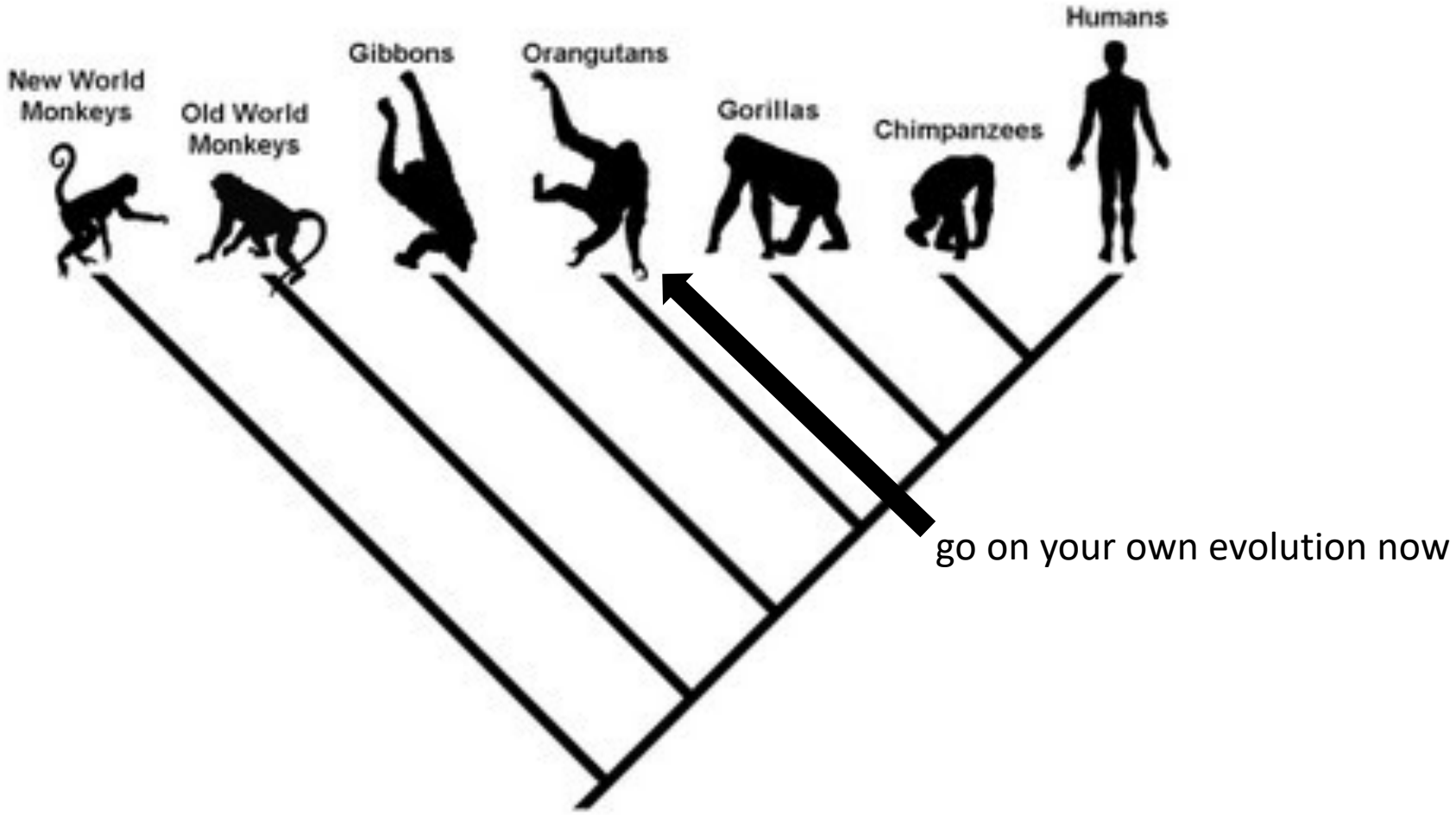
The abbreviated version of "The Road to Homo Sapiens" from Early Man (1965)
https://sites.wustl.edu/prosper/on-the-origins-of-the-march-of-progress/

Tips do not evolve from other tips

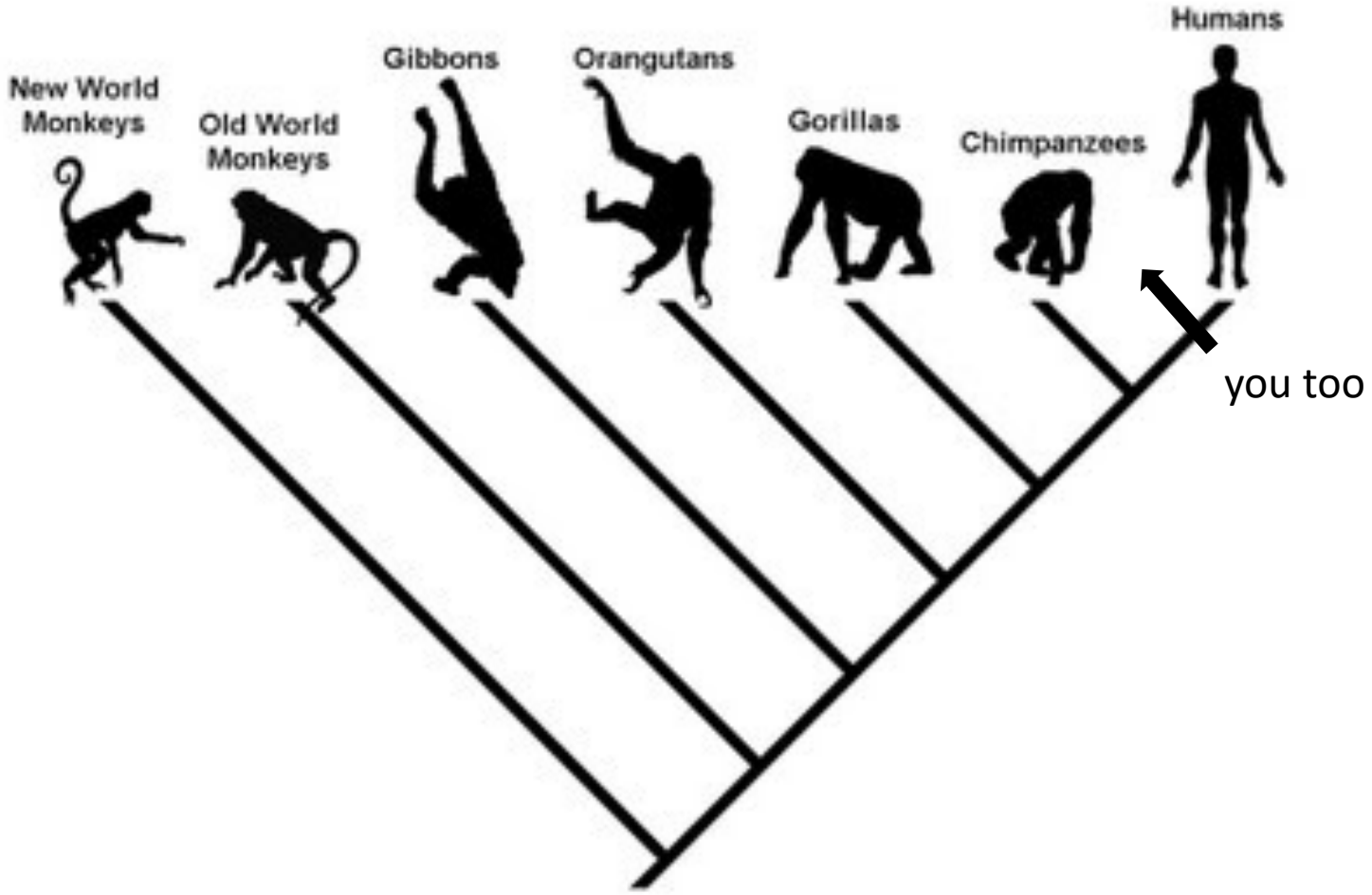


go on your own evolution now

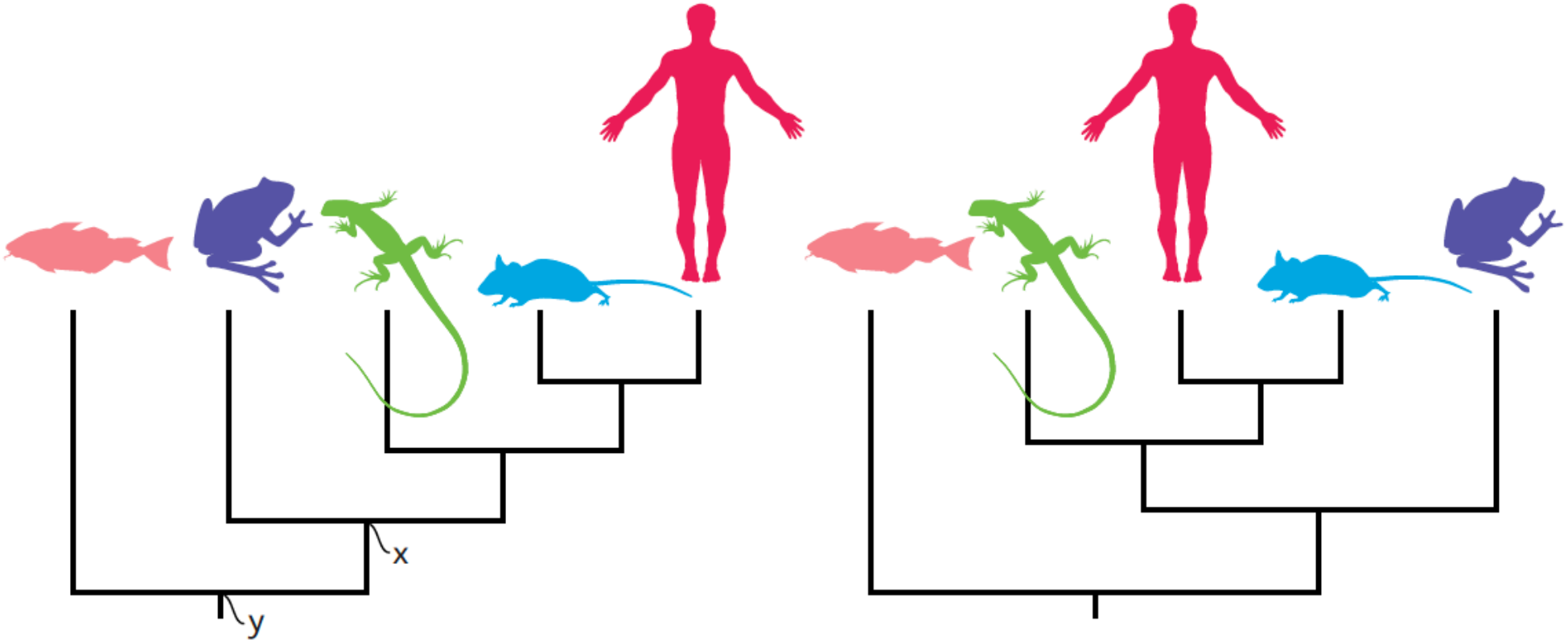
Tips do not evolve from other tips



Who is the “most evolved”?

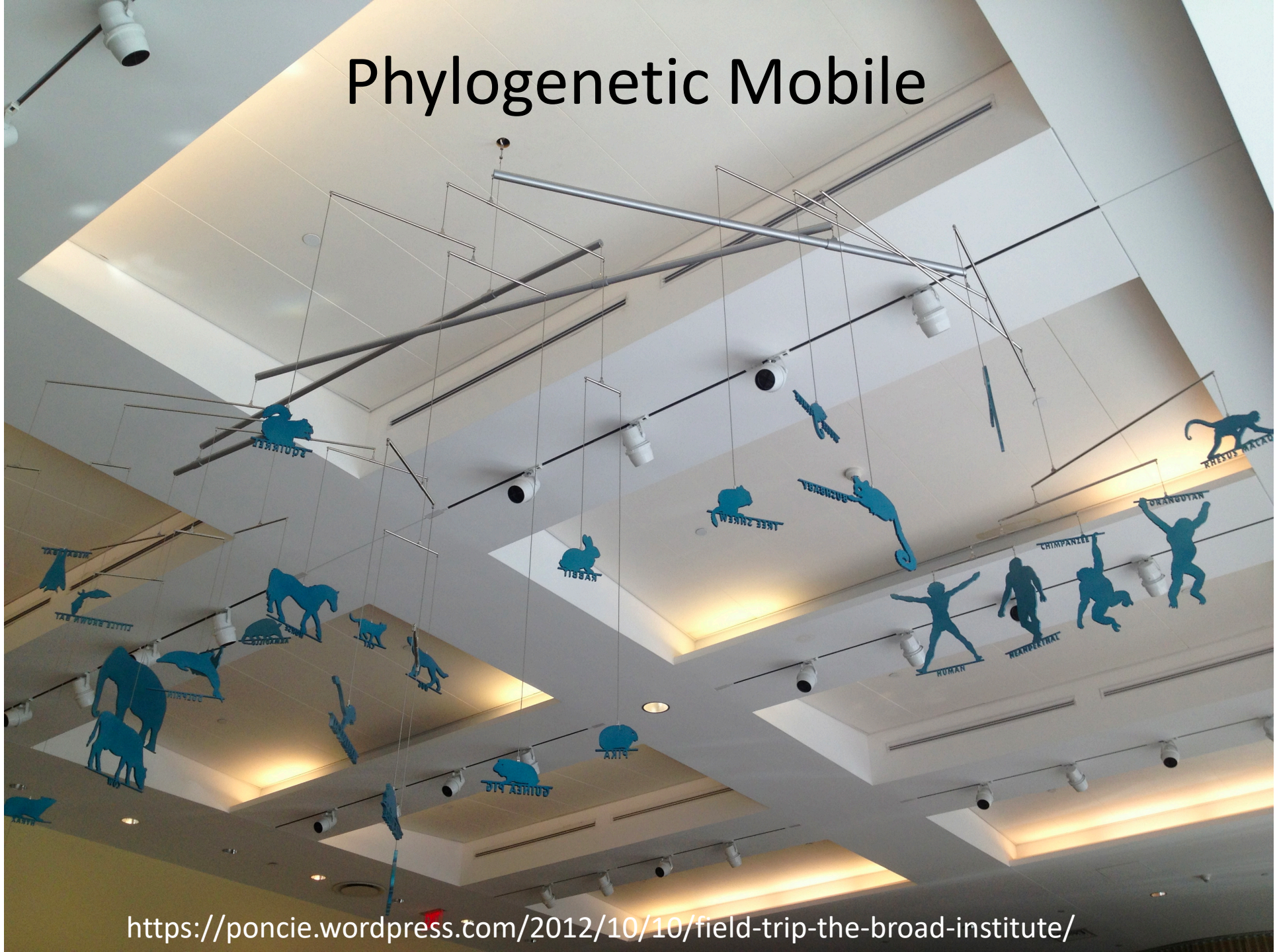


How to interpret a phylogenetic tree: common ancestry



Baum et al. 2005. The Tree-Thinking Challenge. *Science*

Phylogenetic Mobile

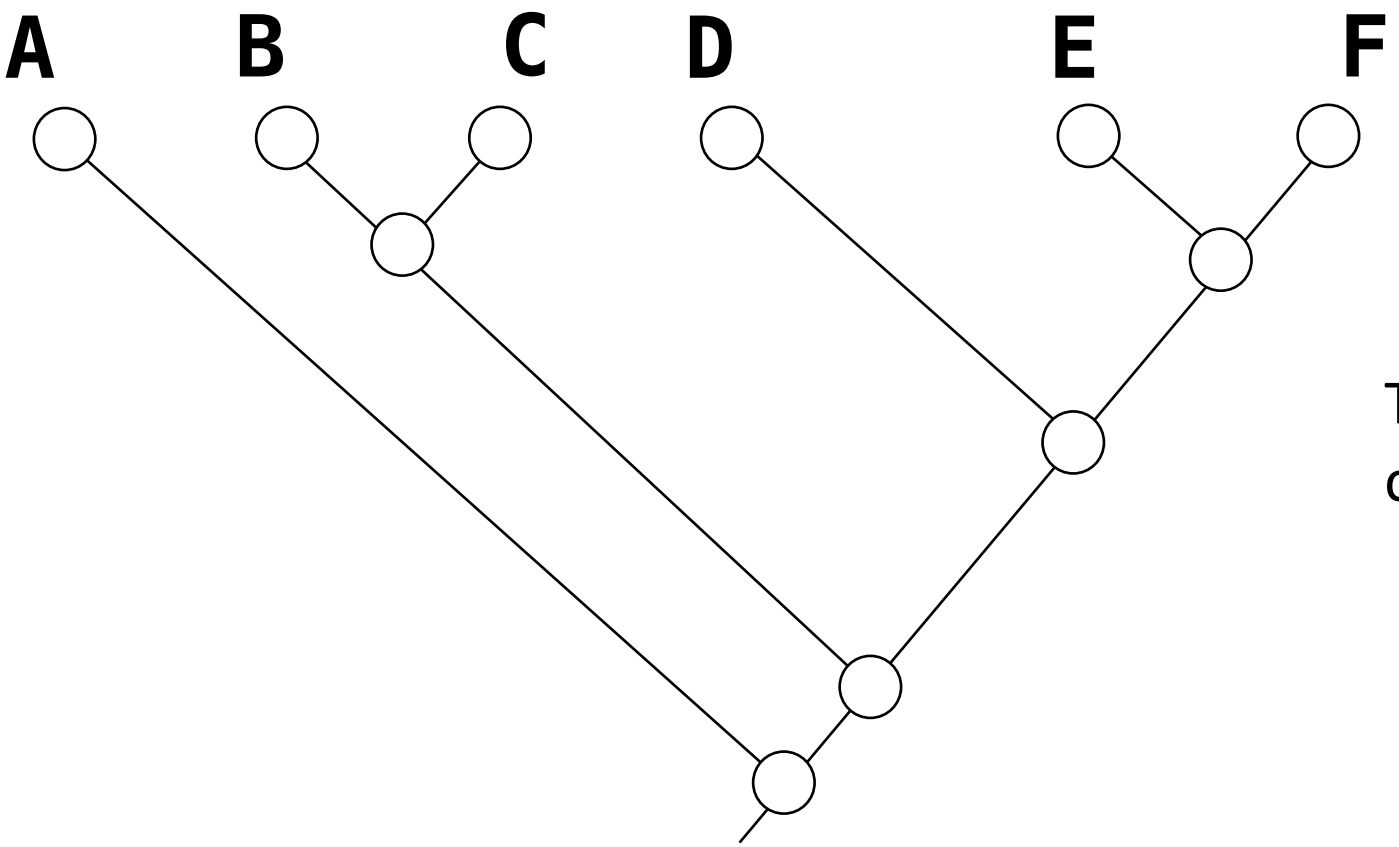


https://poncie.wordpress.com/2012/10/10/field-trip-the-broad-institute/

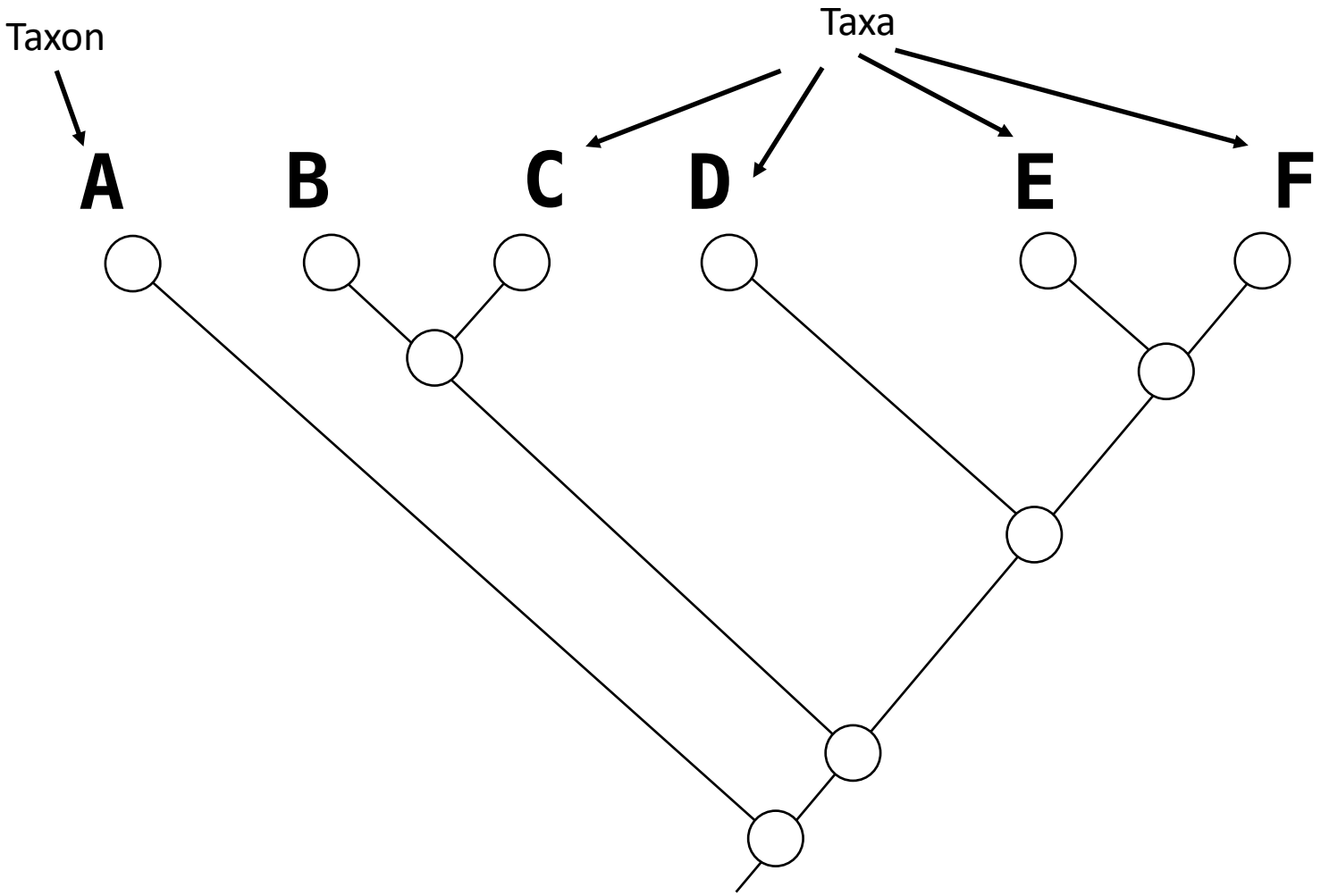
Tree of Life

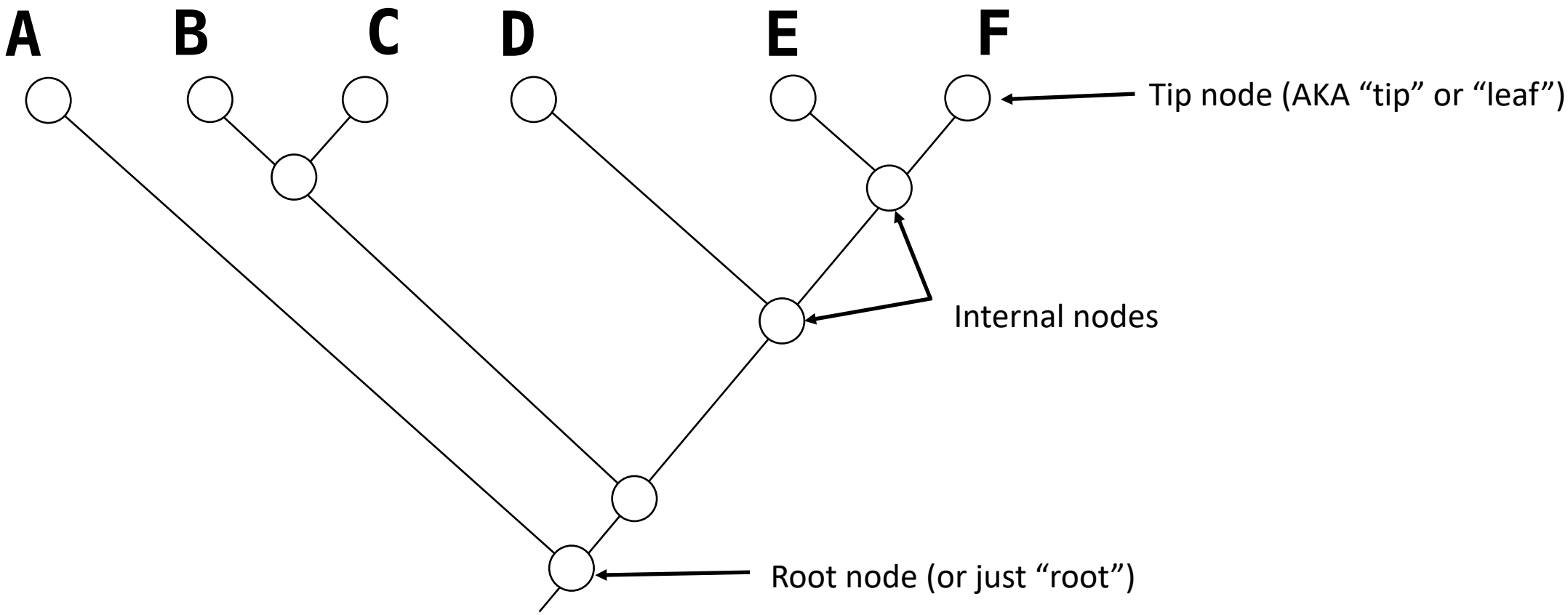
Phylogeny

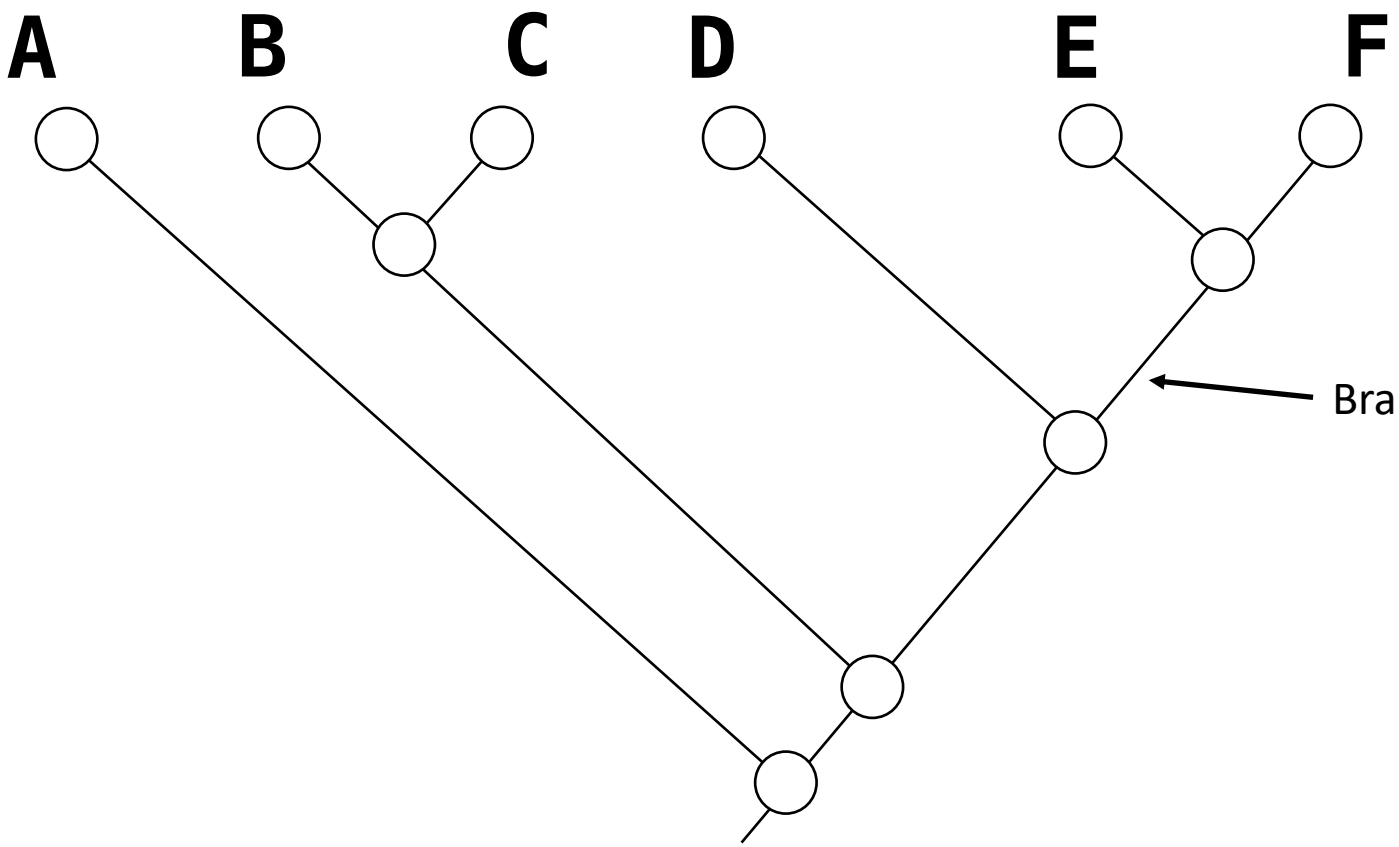
A tree depicting the branching order of evolutionary relationships



The structure of the tree is called the ***topology.***


Taxon
Taxa
A
B
C
D
E
F

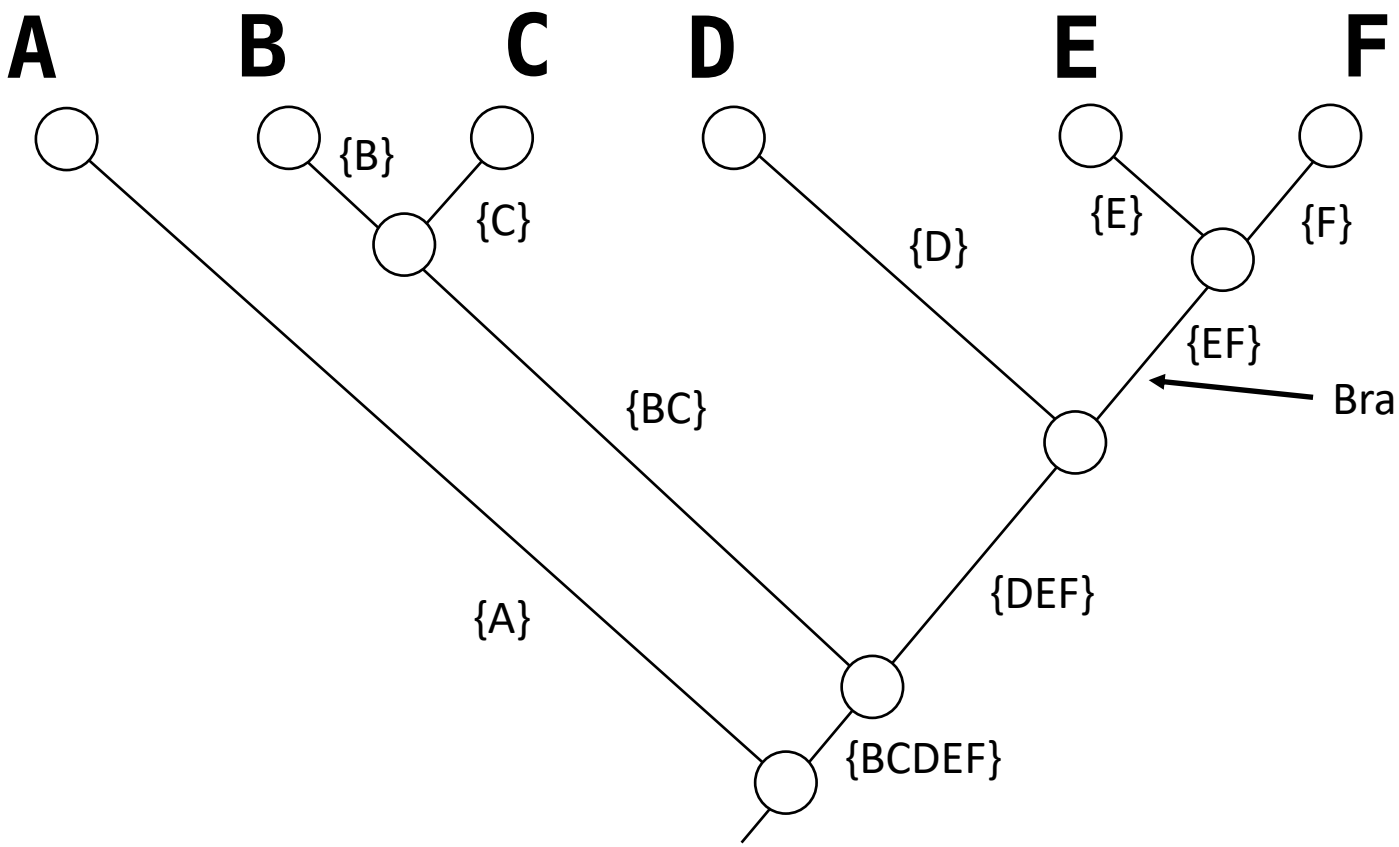

A
B
C
D
E
F
Tip node (AKA “tip” or “leaf”)
Internal nodes
Root node (or just “root”)



Branch (edge, bipartition)

- branches divide the tree into two sets of tips
- you only need to know one set to know what the other is

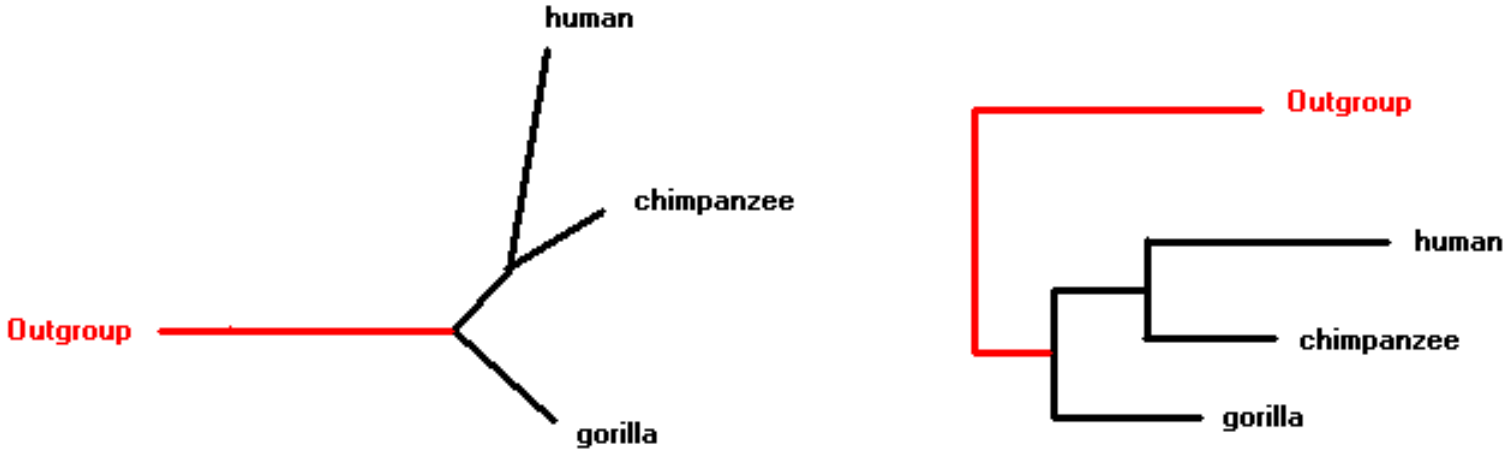
Branches as Bipartitions



Branch (edge, bipartition)

- branches divide the tree into two sets of tips
- you only need to know one set to know what the other is

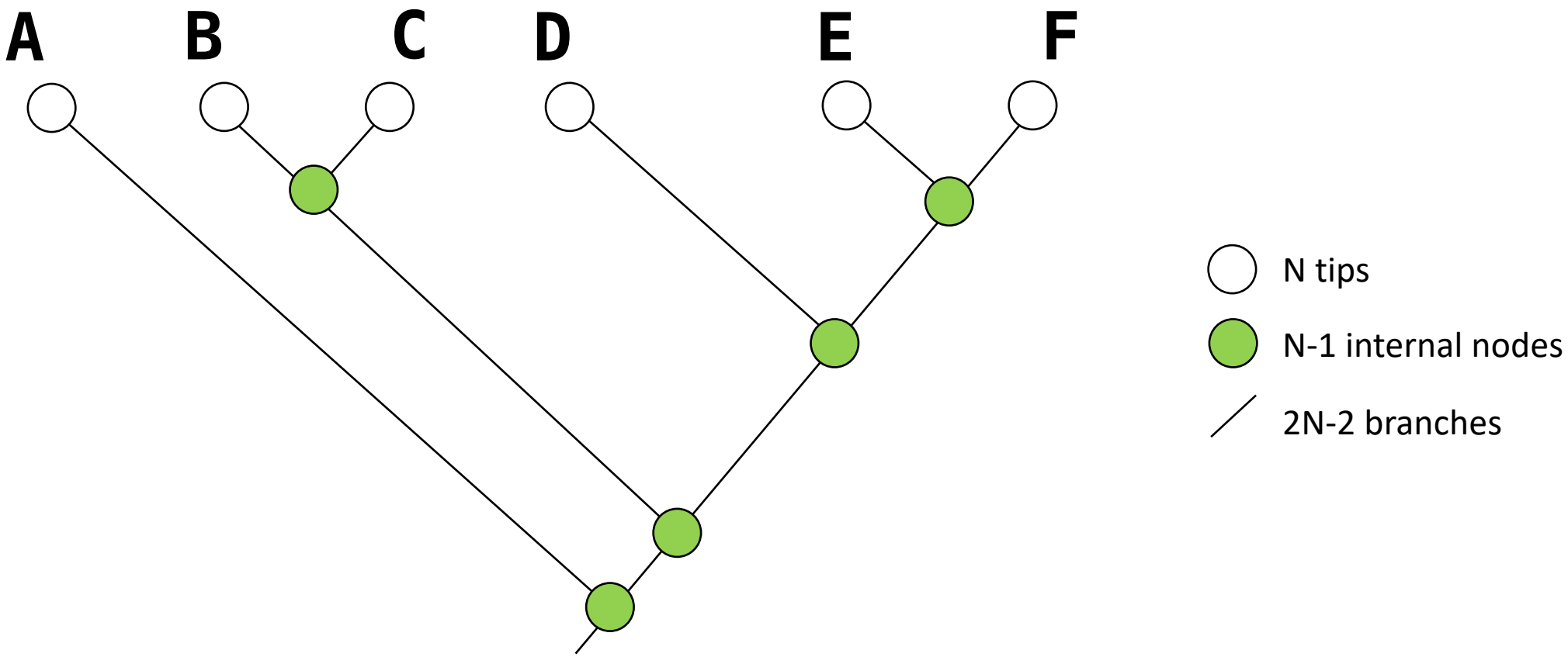
Unrooted and rooted trees



Two equivalent representations of the same unrooted tree

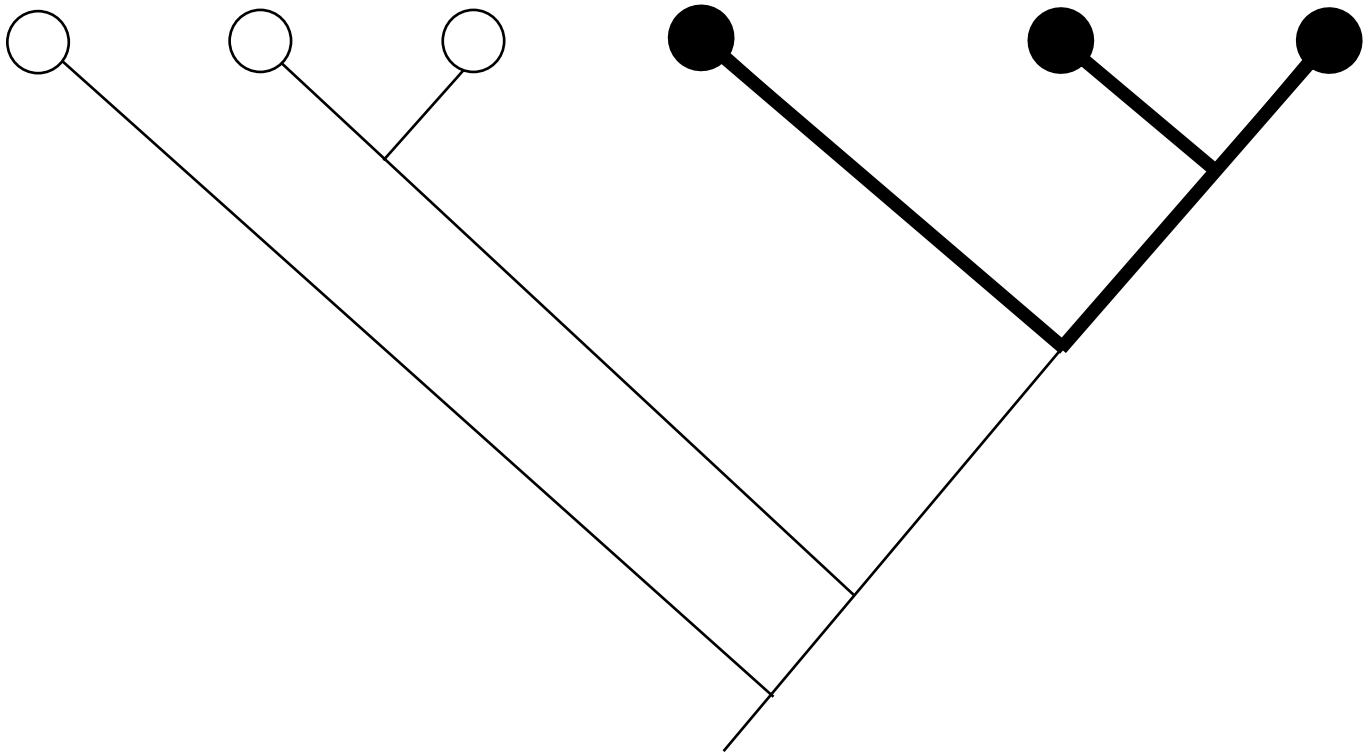
http://www.bioinf.man.ac.uk/resources/phase/manual/node59.html

Properties of Rooted Trees



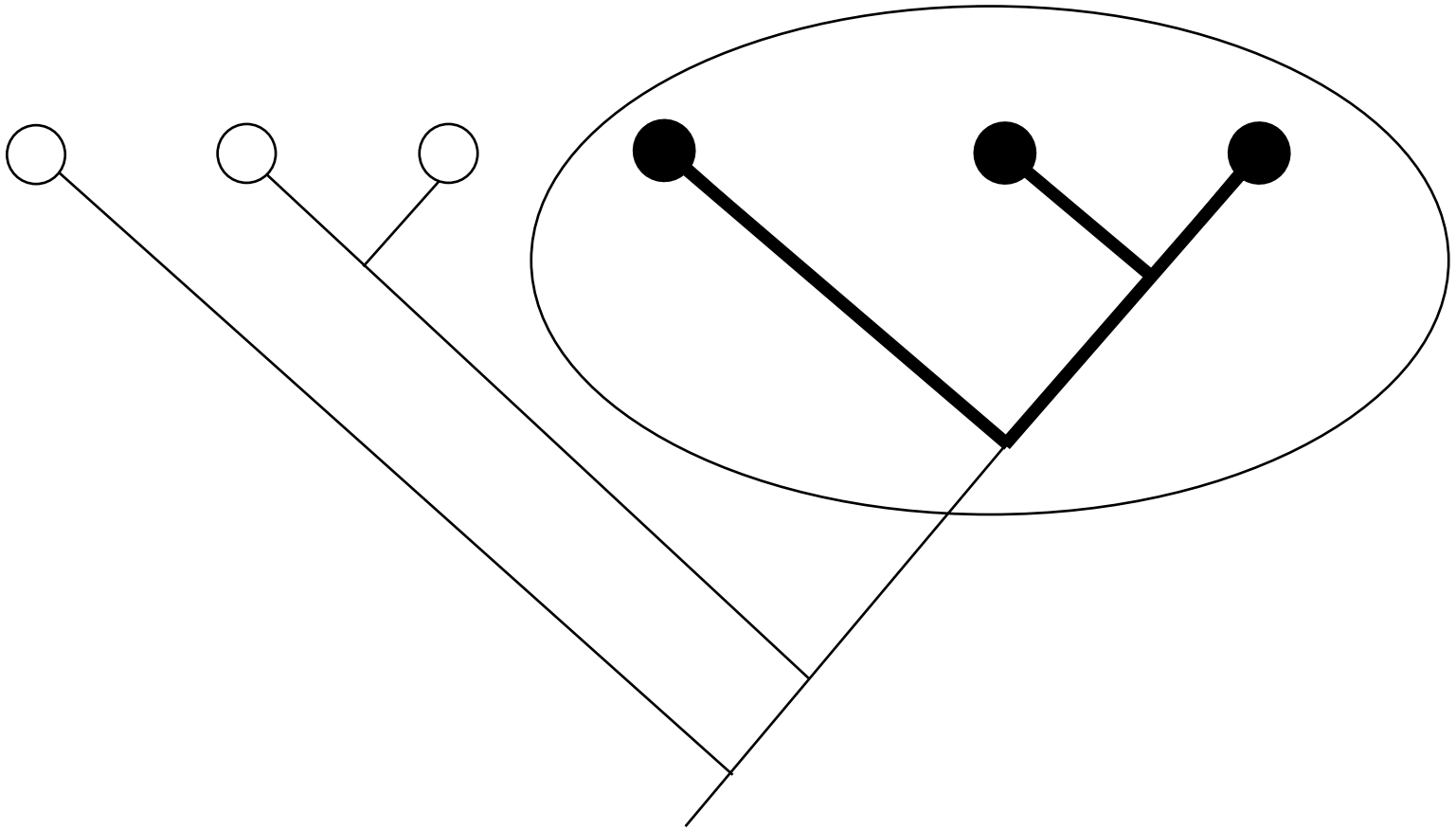
Monophyletic

Group includes an ancestor and all of its descendants



Monophyletic

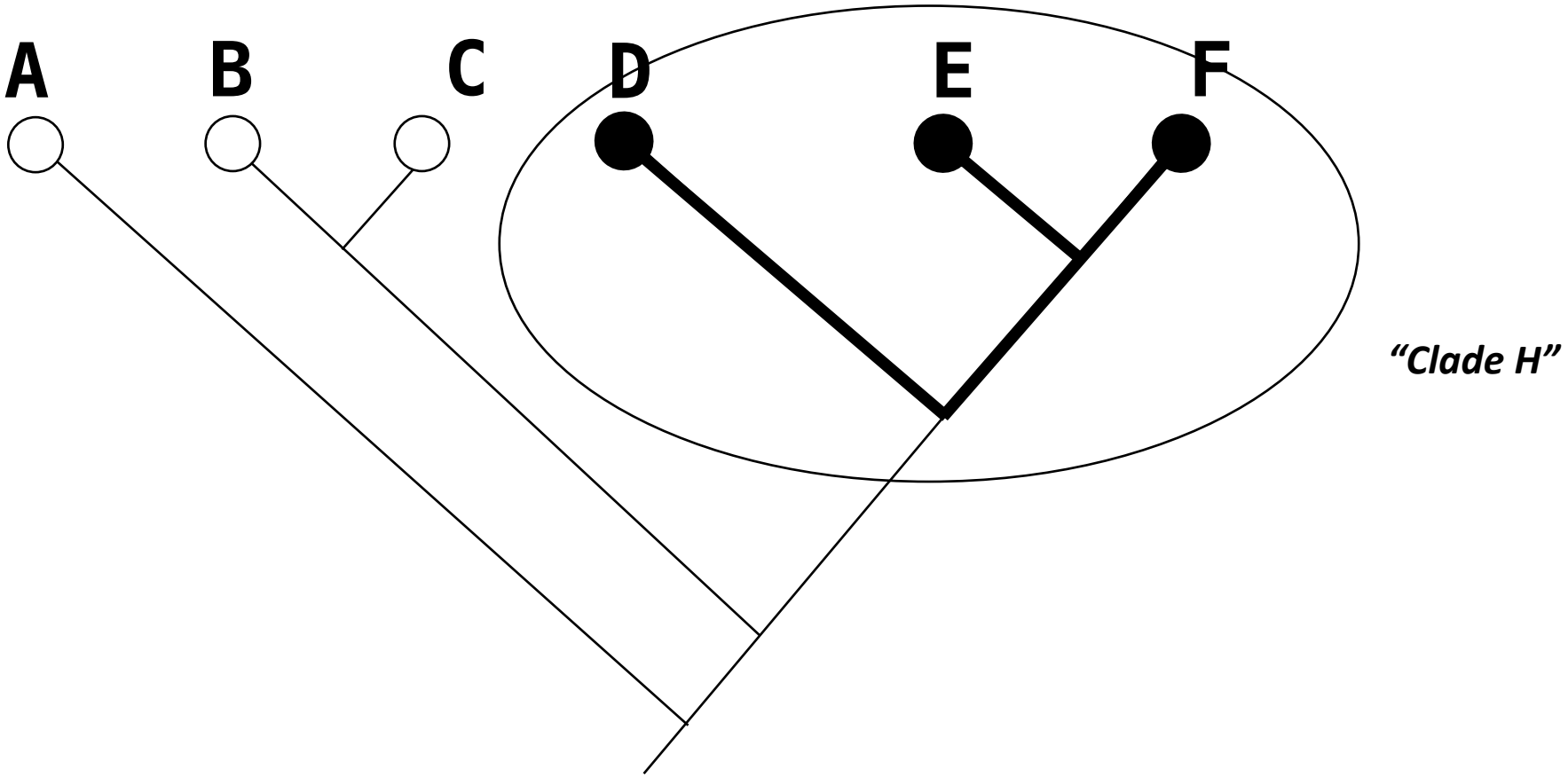
Group includes an ancestor and all of its descendants



Clade *– a monophyletic group*

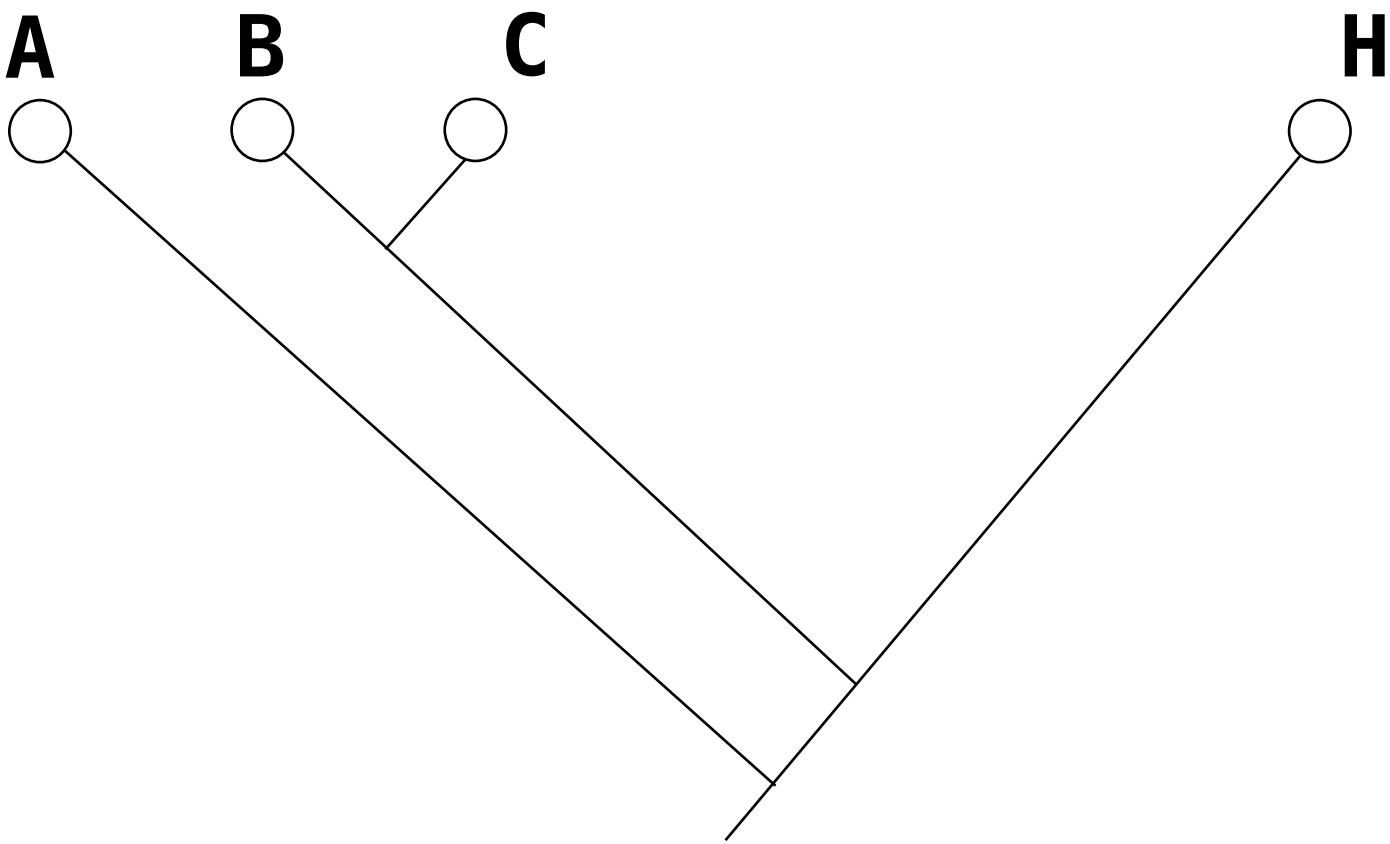
Monophyletic

Group includes an ancestor and all of its descendants



Monophyletic

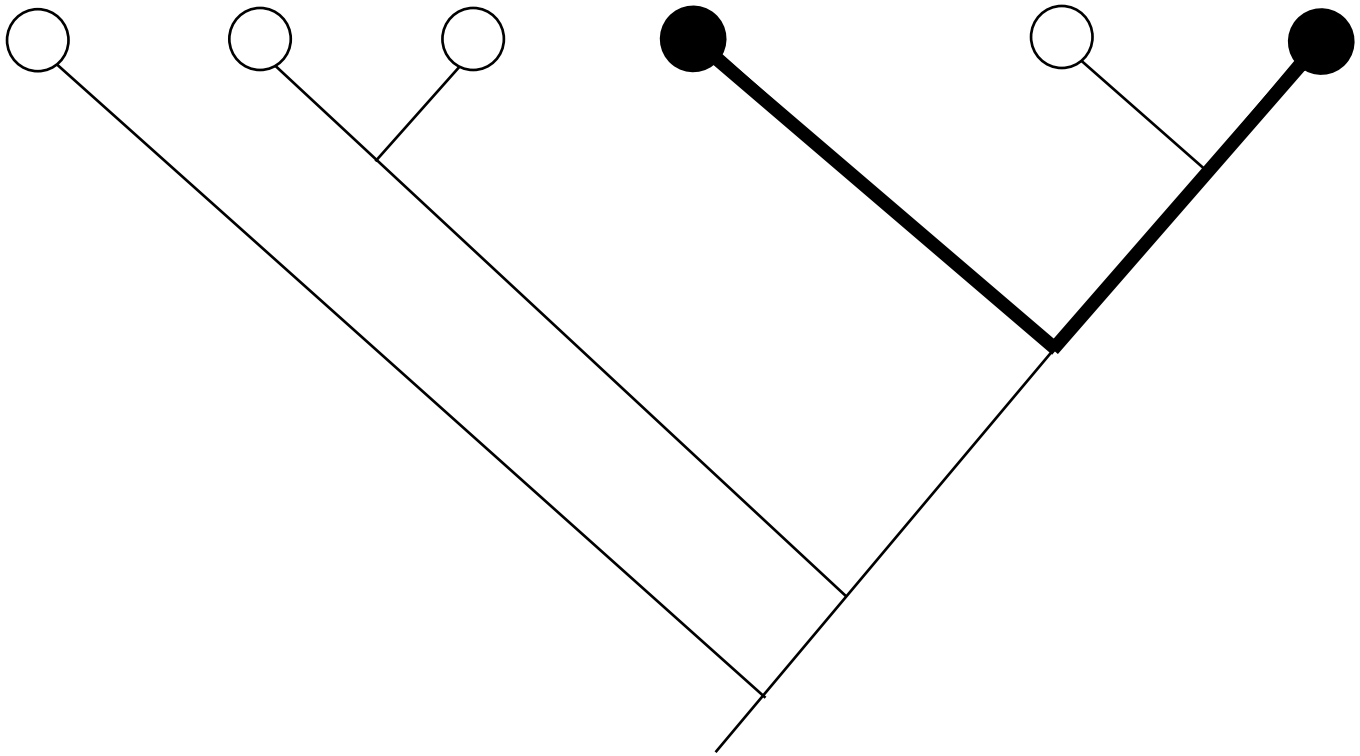
Group includes an ancestor and all of its descendants



H is a valid taxon

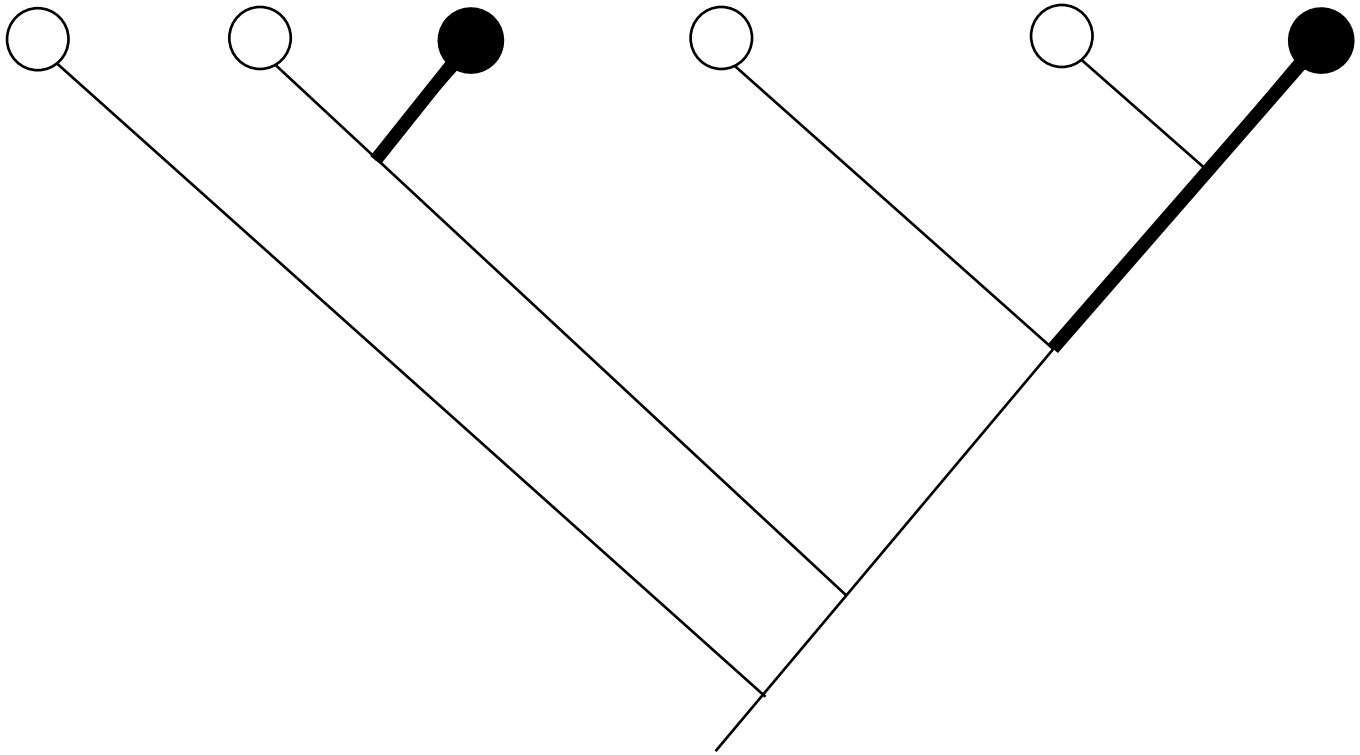
Paraphyletic

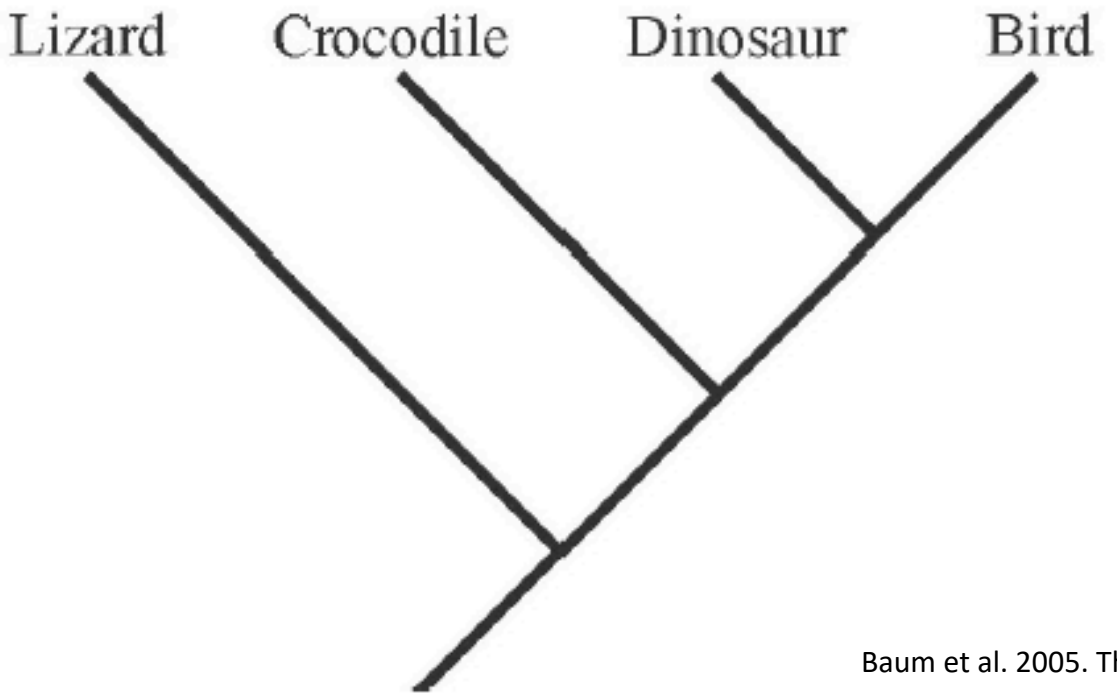
Leaves out some descendants of an ancestor



Polyphyletic

Leaves out an ancestor





Baum et al. 2005. The Tree-Thinking Challenge. *Science*

2) By reference to the tree above, which of the following is an accurate statement of relationships?

a) A crocodile is more closely related to a lizard than to a bird
b) A crocodile is more closely related to a bird than to a lizard
c) A crocodile is equally related to a lizard and a bird
d) A crocodile is related to a lizard, but is not related to a bird

Characters

Any heritable attribute of a taxon

- Could be genetic, morphological

Homologous characters

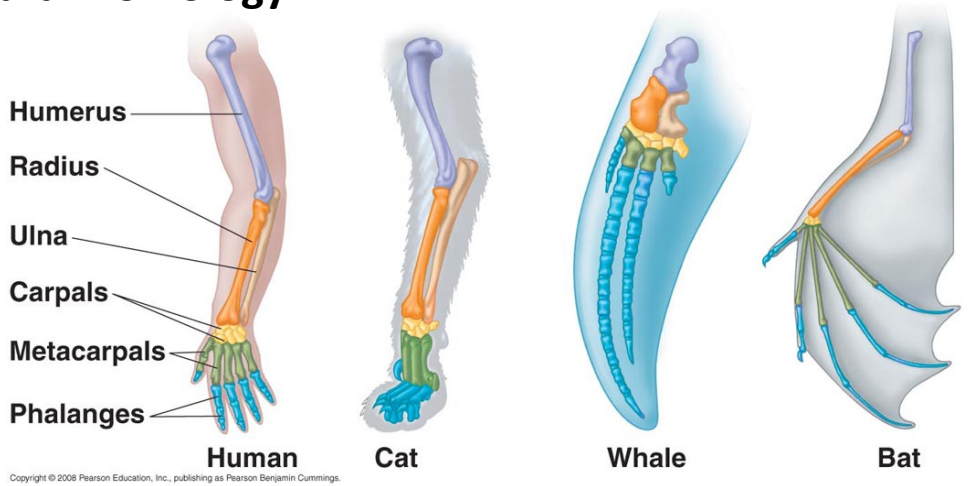
- Any trait present in taxa because they were also present in the most recent common ancestor of those taxa

Character state – value that a character can have

- A,T,C,G for DNA
- 0,1 for binary traits

Homology and Bioinformatics

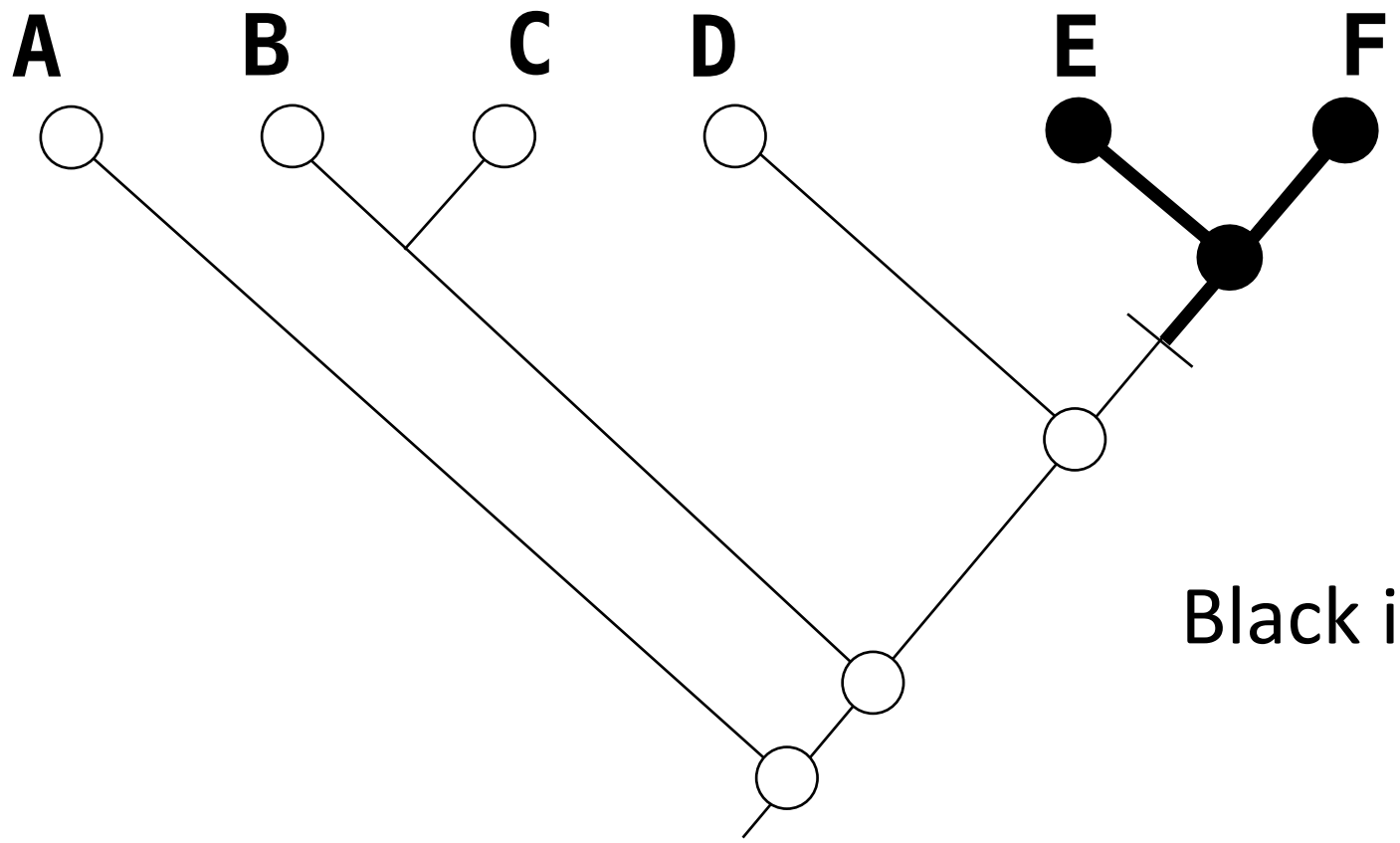
Structural Homology:



Sequence Homology (Alignment):

```
Human KKASKPKKAASKAPTKKPKATPKATPVKKAKKKLAAT
Mouse KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKPAAT
  Rat KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKPAAT
  Cow KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKPAAT
Chimp KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKLAAT
      ***:**********:******************** ***
```

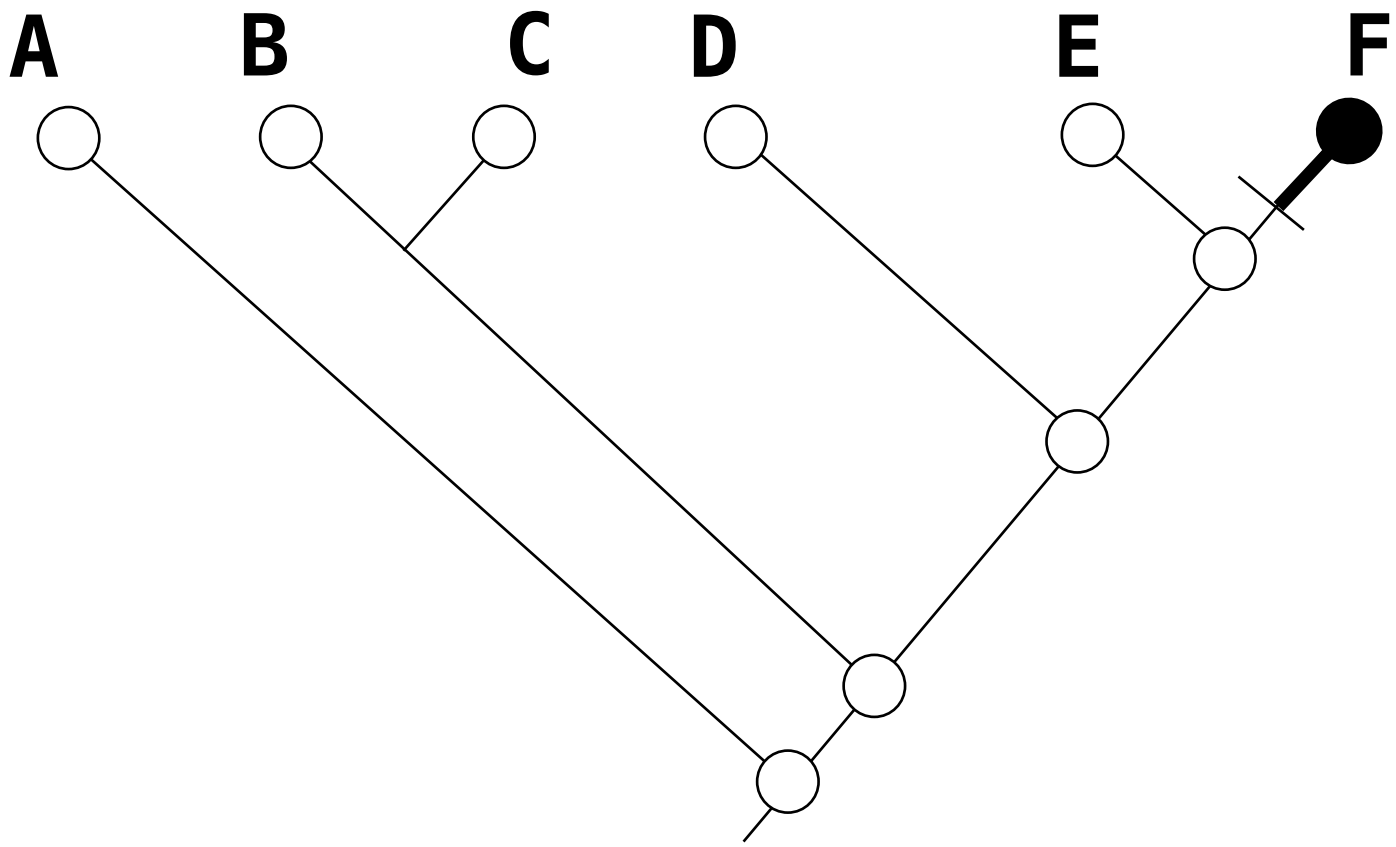
Apomorphy
A derived character state
A character state that is different from that of an ancestor



Black is a derived character state.

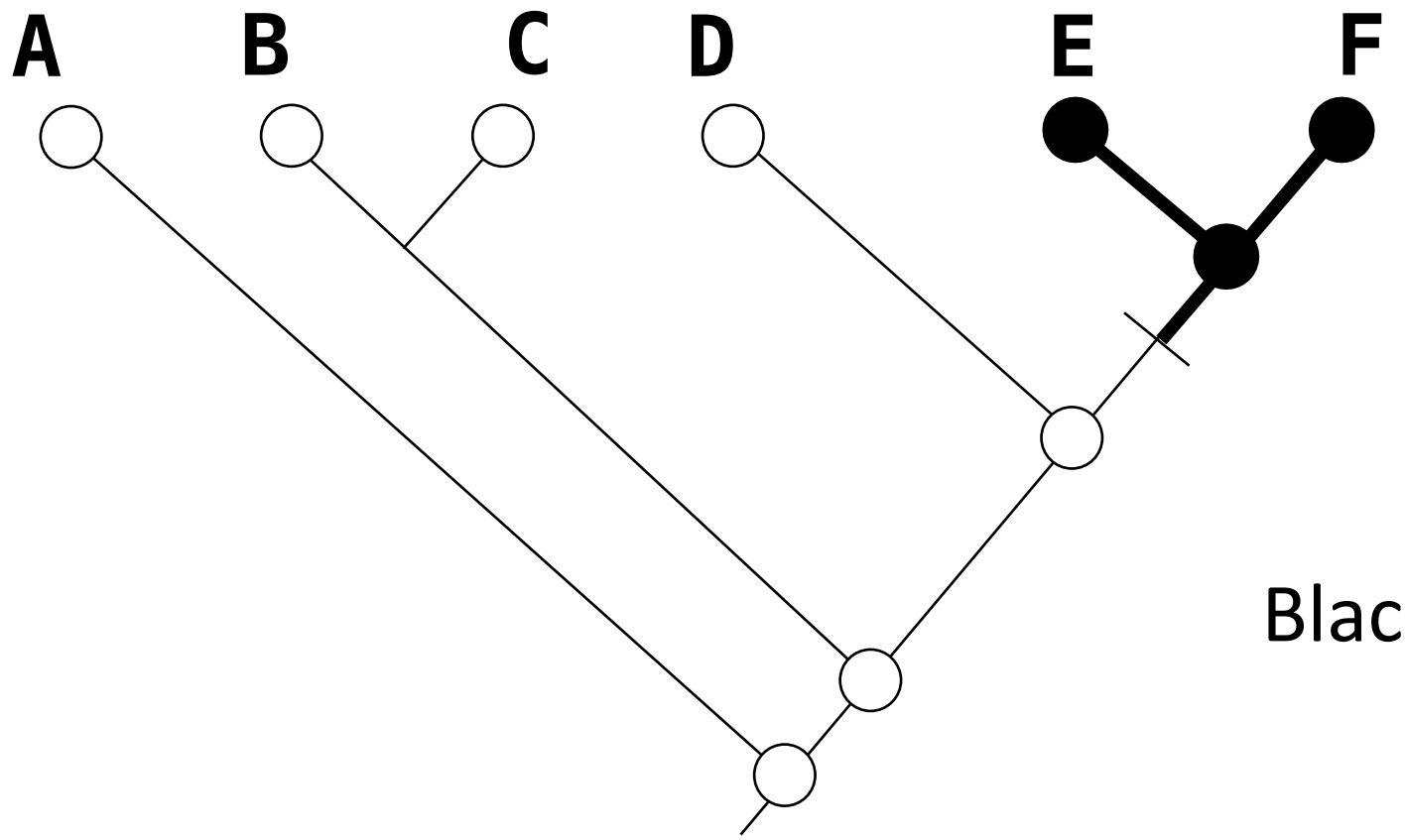
Autapomorphy

An apomorphy that is unique to a single taxon



Synapomorphy

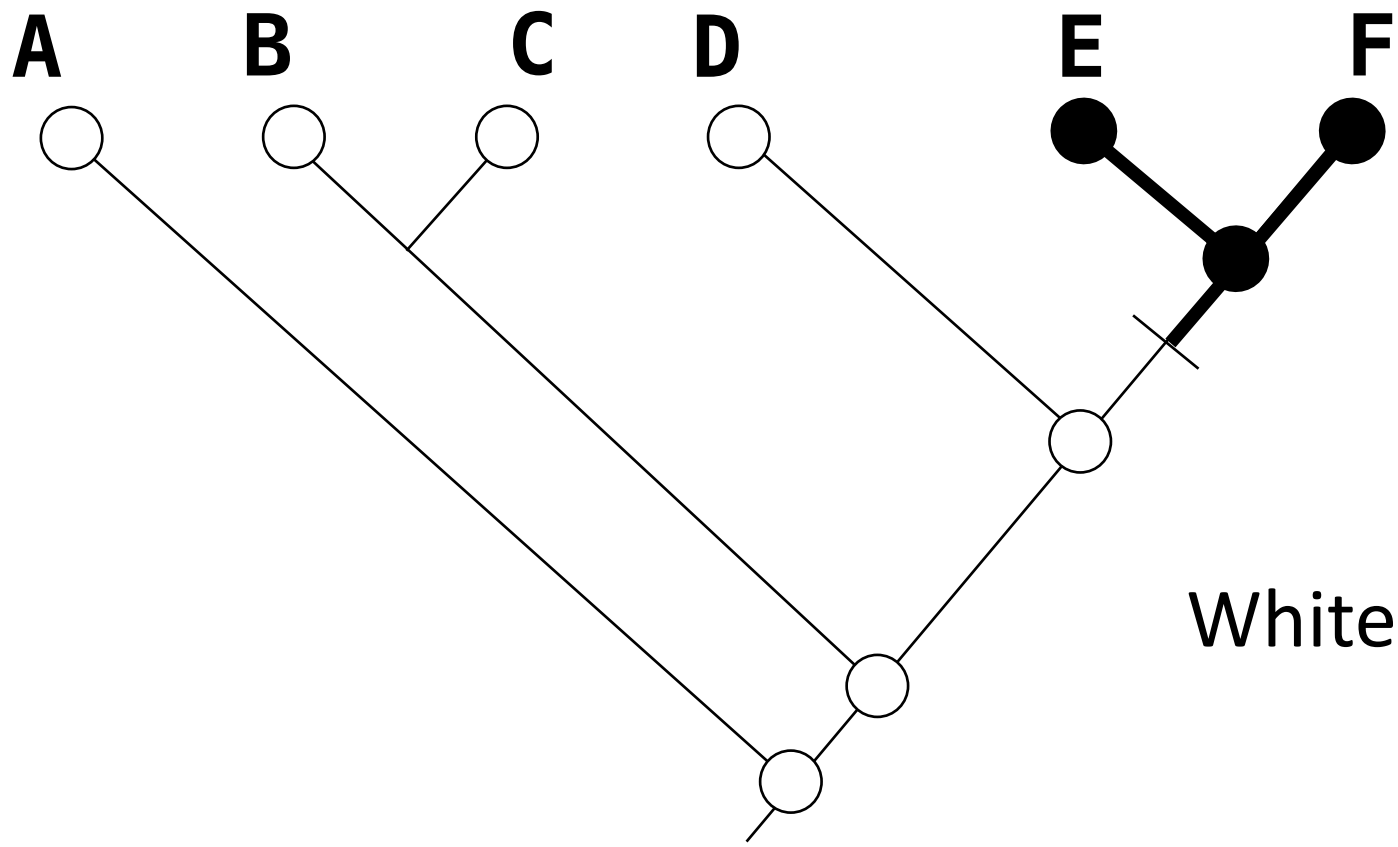
A shared apomorphy



Black is a a synapomorphy for the clade **EF**.

Pleisiomorphy

A character that is the same as that of the ancestor



White is a pleisiomorphy for **B**.

Character matrix

A table of character states. Each row corresponds to a taxon, each column is a homologous character.

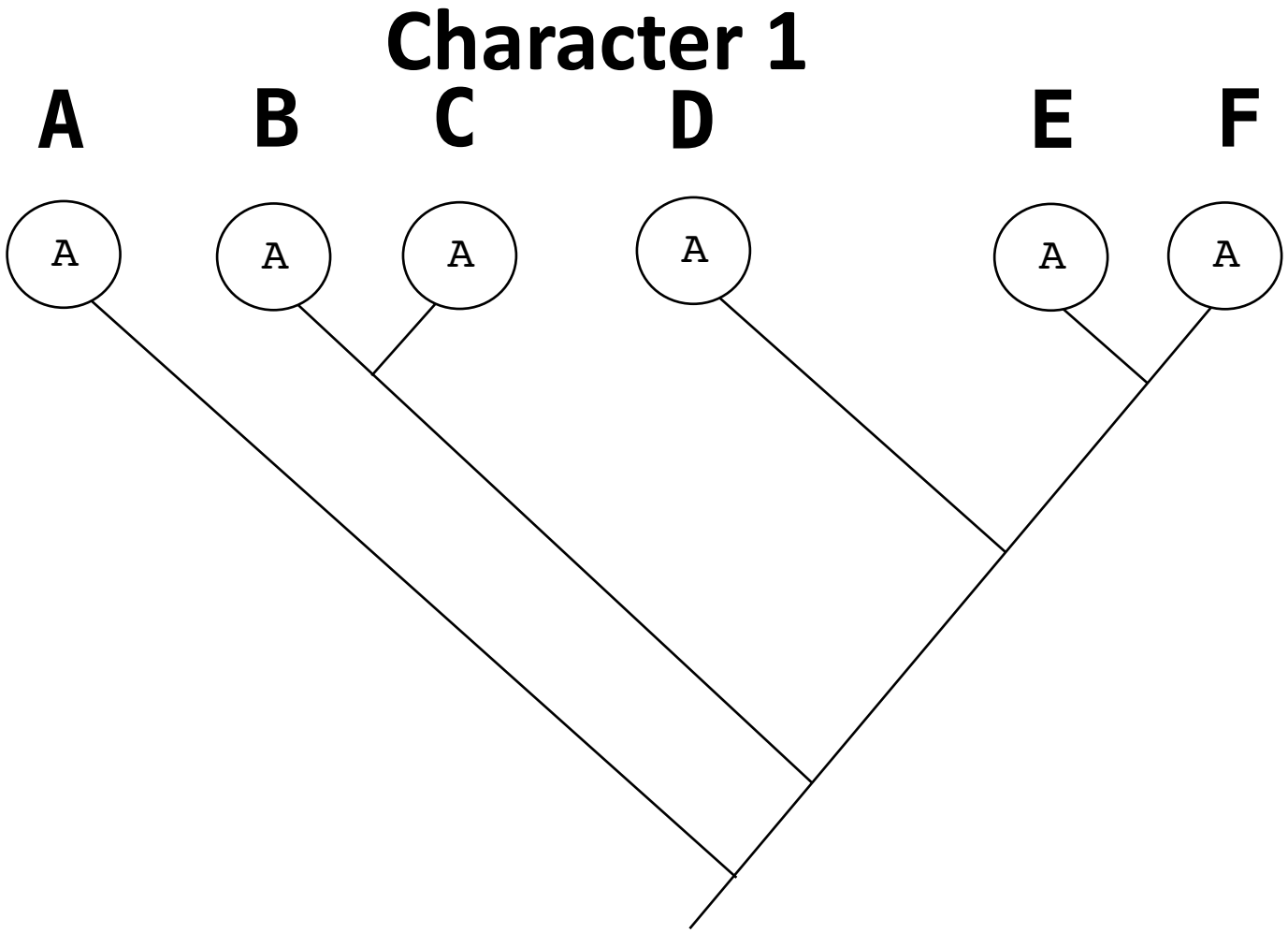
Characters

Taxa	1	2	3
A	A	C	A
B	A	G	A
C	A	G	A
D	A	C	A
E	A	C	T
F	A	C	T

Character matrix

A table of character states. Each row corresponds to a taxon, each column is a homologous character.

Taxa	Characters		
	1	**2**	**3**
A	A	C	A
B	A	G	A
C	A	G	A
D	A	C	A
E	A	C	T
F	A	C	T

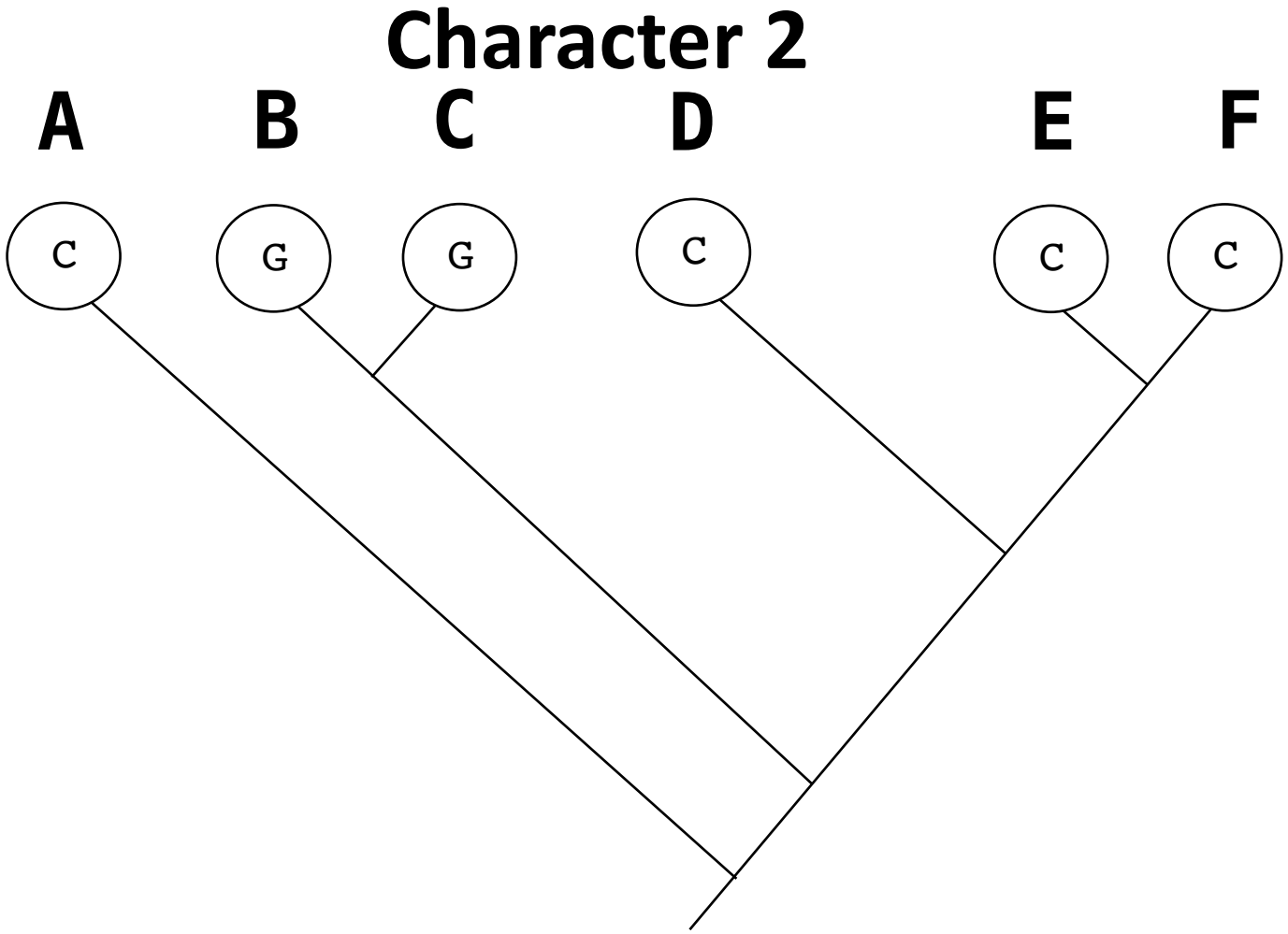


Character matrix

A table of character states. Each row corresponds to a taxon, each column is a homologous character.

Characters

Taxa	1	2	3
A	A	C	A
B	A	G	A
C	A	G	A
D	A	C	A
E	A	C	T
F	A	C	T

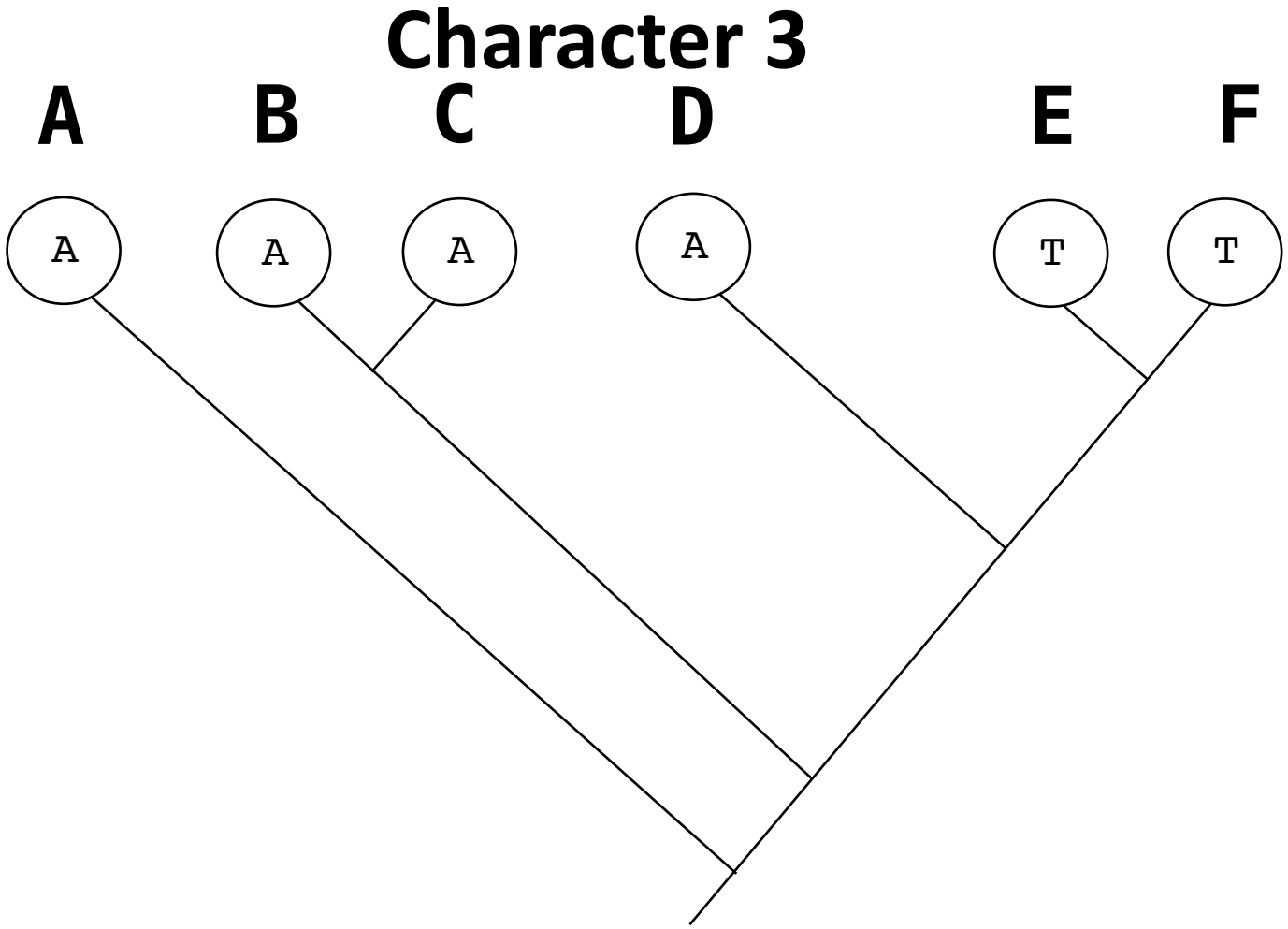


Character matrix

A table of character states. Each row corresponds to a taxon, each column is a homologous character.

Characters

Taxa	1	2	3
A	A	C	A
B	A	G	A
C	A	G	A
D	A	C	A
E	A	C	T
F	A	C	T

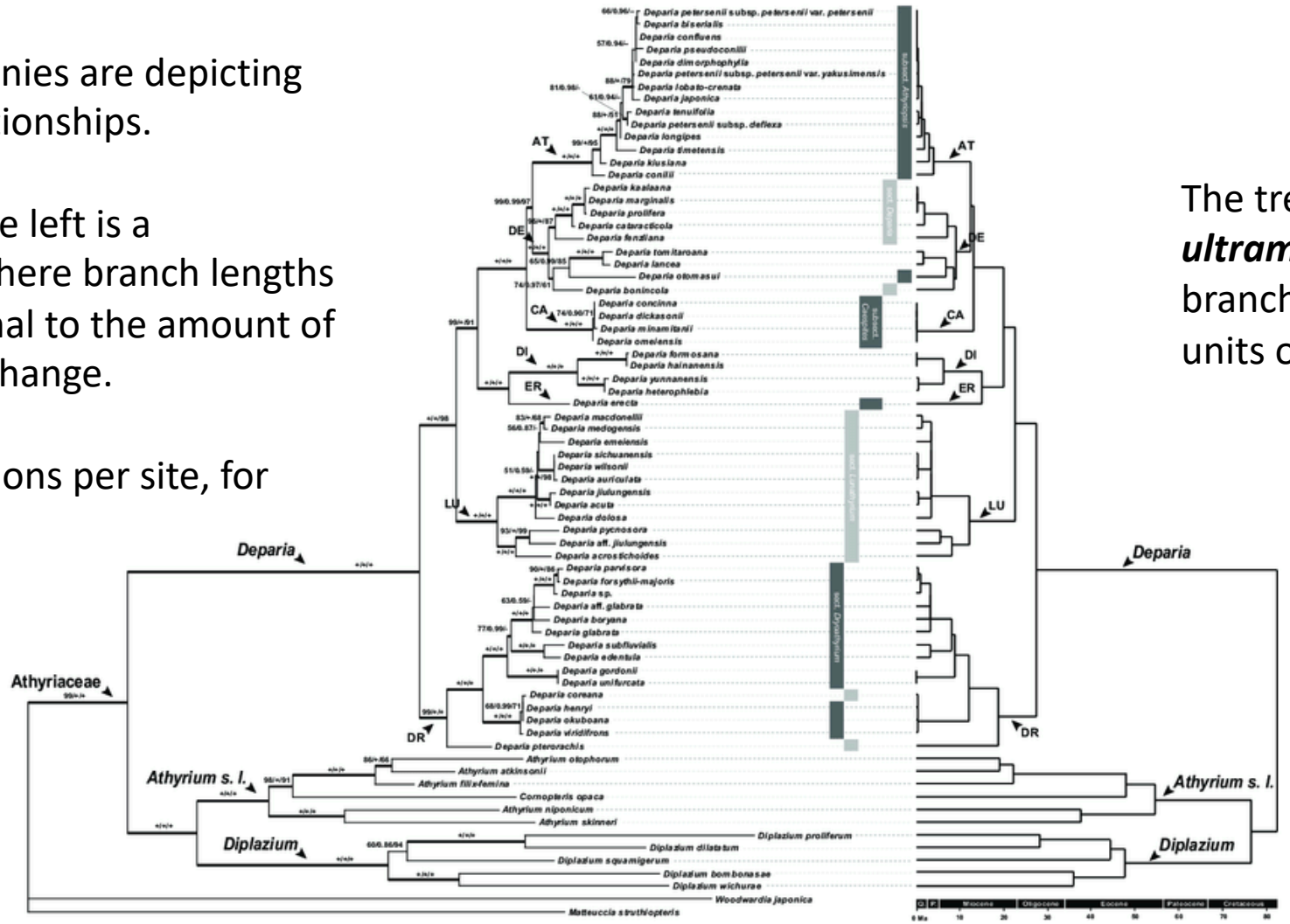


Branch lengths matter

These phylogenies are depicting the same relationships.

The tree on the left is a ***phylogram***, where branch lengths are proportional to the amount of evolutionary change.

DNA substitutions per site, for instance.



The tree on the right is an ***ultrametric chronogram***, where branch lengths are depicted in units of time.

Kuo et al. 2017. Morphological characterization of infra-generic lineages in *Deparia* (Athyriaceae: Polypodiales). *Cladistics*

Phylogenetic Inference

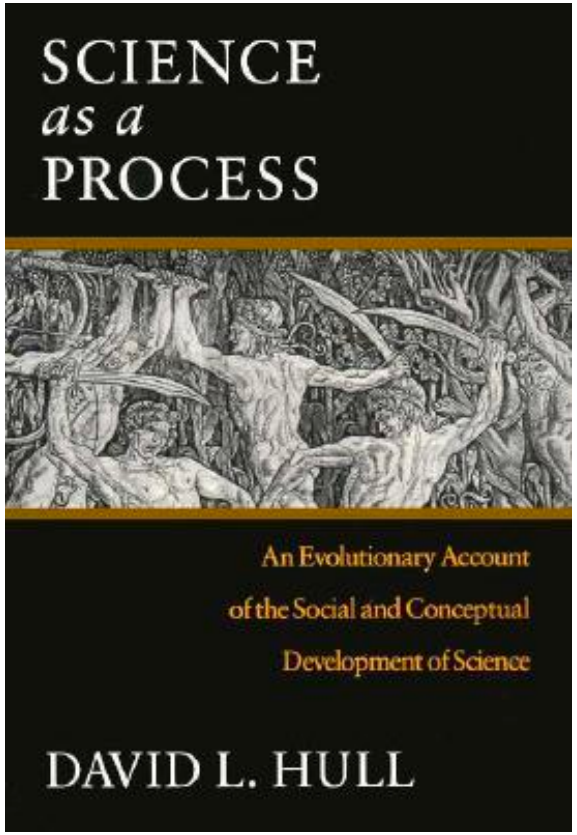
The estimation of a phylogeny based on character data and a model of character evolution

Usually tries to find the “best” tree

Phylogenetic Inference

The estimation of a phylogeny based on character data and a model of character evolution

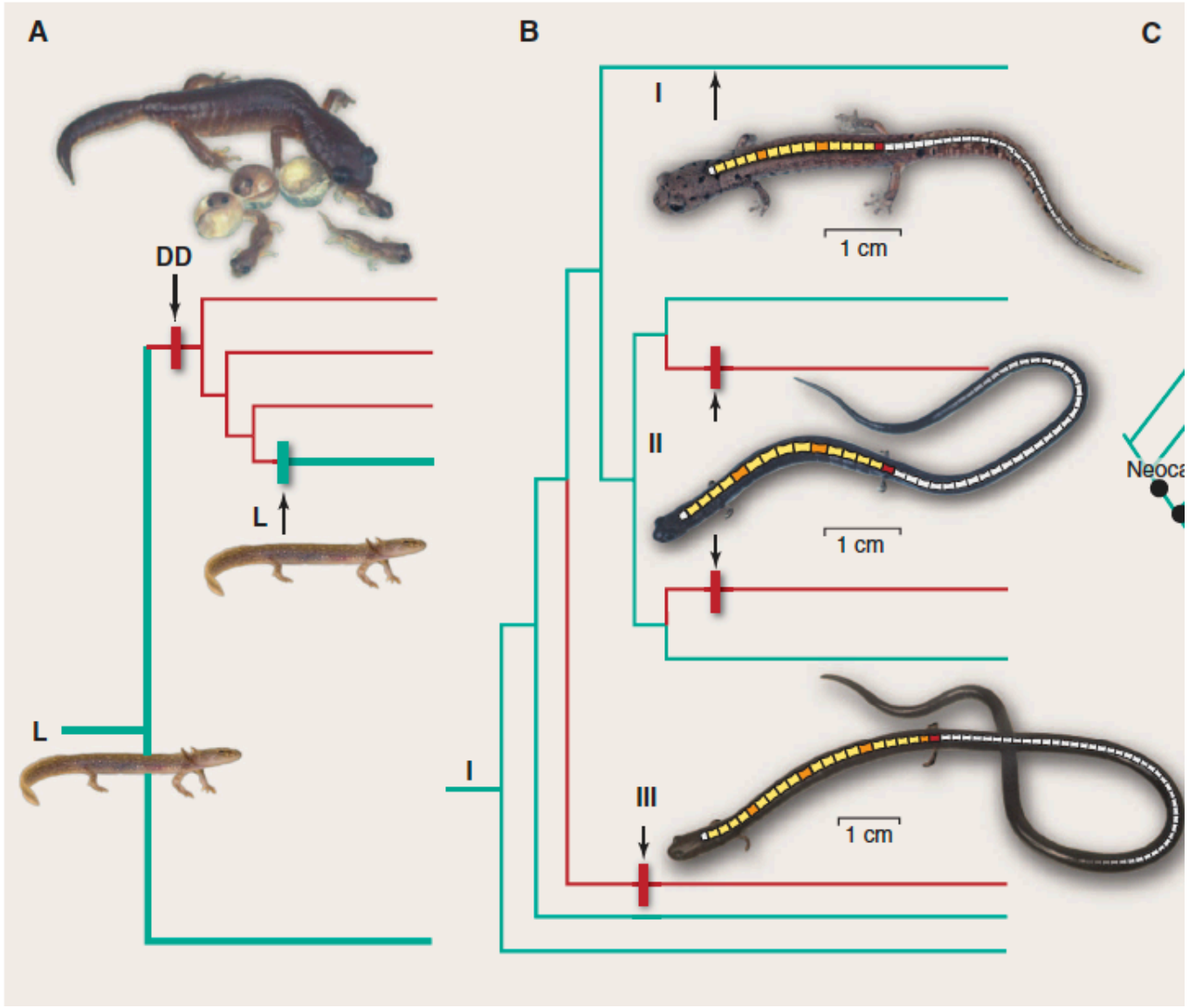
Usually tries to find the “best” tree



Homoplasy

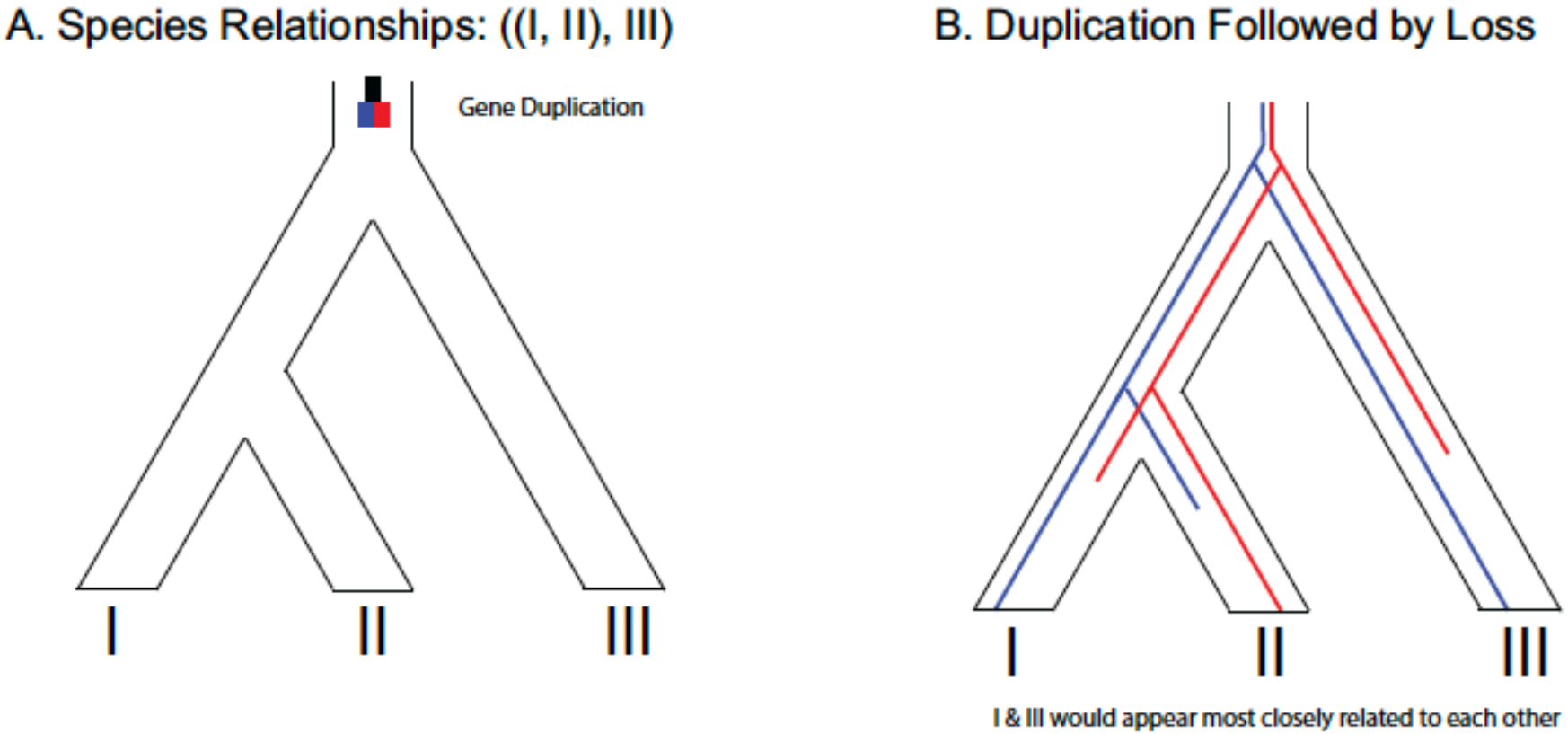
Independent origins of the same character state.

Can be due to convergence or reversal.



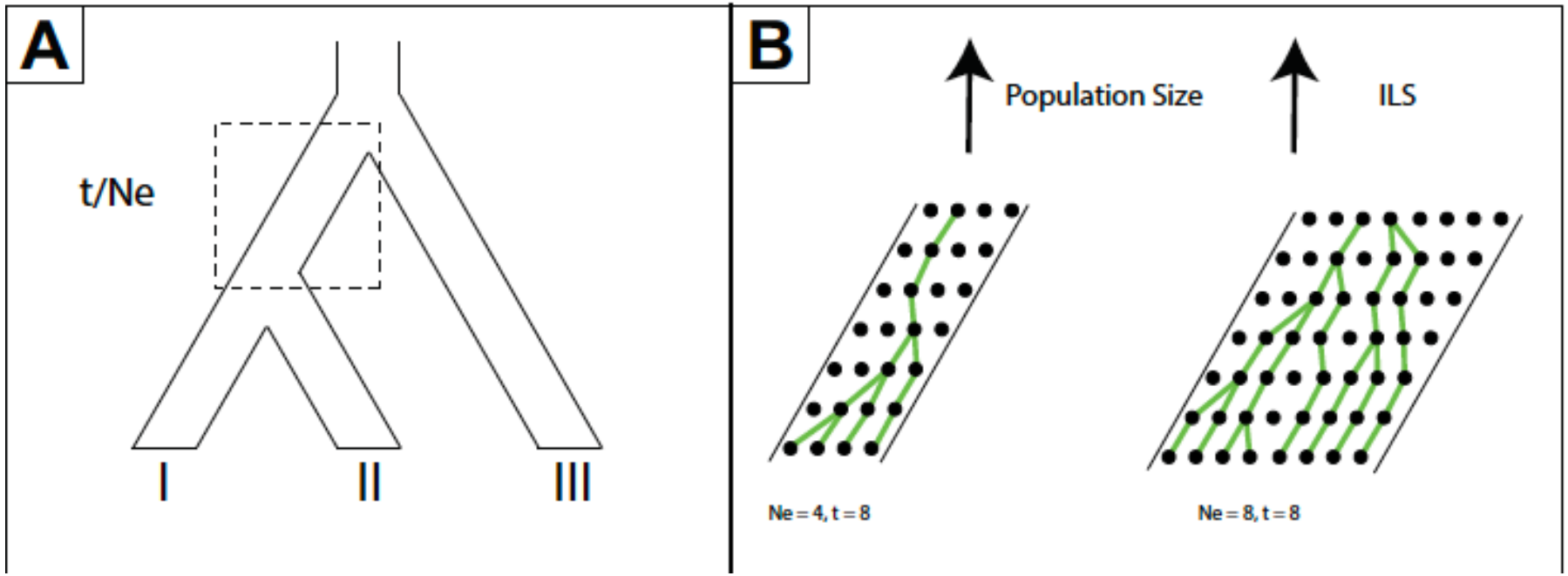
Wake et al. 2011. *Science*

Gene duplication



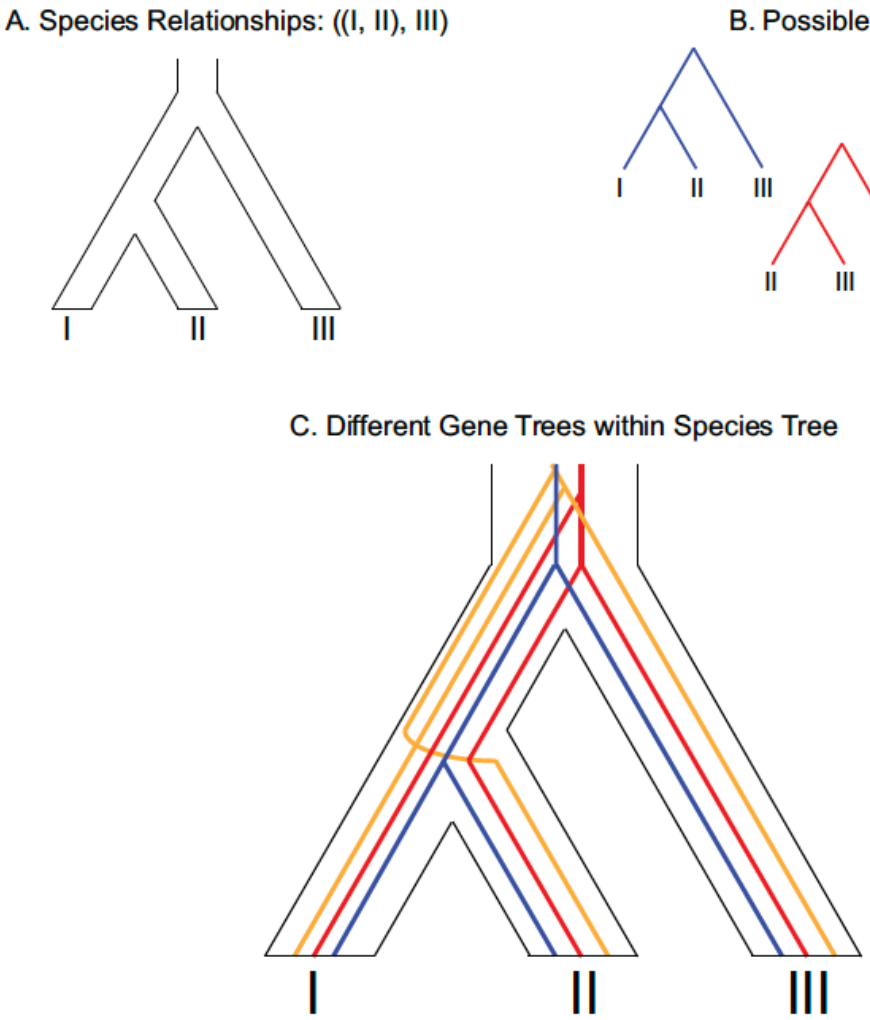
Campbell et al. 2020. Addressing incomplete lineage sorting and paralogy in the inference of uncertain salmonid phylogenetic relationships. *PeerJ*

Incomplete Lineage Sorting



Campbell et al. 2020. Addressing incomplete lineage sorting and paralogy in the inference of uncertain salmonid phylogenetic relationships. *PeerJ*

Gene-Tree/Species-Tree Discordance



Campbell et al. 2020. Addressing incomplete lineage sorting and paralogy in the inference of uncertain salmonid phylogenetic relationships. *PeerJ*

Approaches

Distance methods

- Construct a tree from pairwise distance matrix
- Use ***similarity*** as a proxy for relatedness
- ***UPGMA*** – heuristic, assumes equal rates
- ***Neighbor-joining*** – works efficiently, except with homoplasy and unobserved changes

Character Matrix

TABLE 8.3 Ten characters from the carnivoran morphology data set

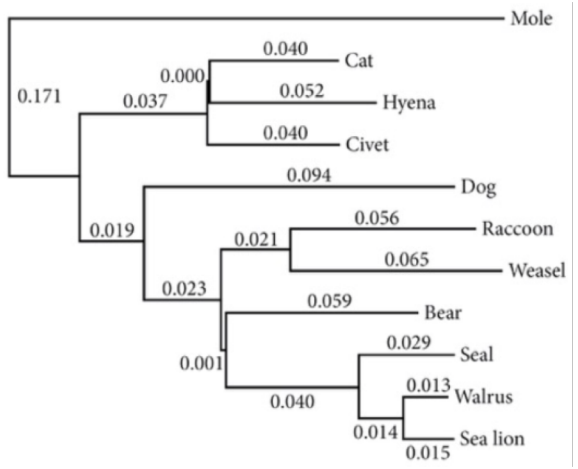
	Characters									
	1	2	3	4	5	6	7	8	9	10
Outgroup	0	0	0	0	0	0	0	0	0	0
Cat	0	1	0	1	0	0	1	1	1	0
Hyena	0	1	0	1	0	0	1	0	1	0
Civet	0	1	0	0	0	0	0	0	1	0
Dog	1	0	0	0	1	0	0	0	0	0
Raccoon	1	0	0	0	1	0	0	0	0	0
Bear	1	0	0	0	1	1	0	0	0	1
Otter	1	0	0	0	1	0	0	0	0	1
Seal	1	0	1	0	1	1	0	0	0	1
Walrus	1	0	1	0	1	1	0	0	0	1
Sea lion	1	0	1	0	1	1	0	0	0	1

Computed Distances

TABLE 8.4 Pairwise distances for the morphological data

	Outgroup	Cat	Hyena	Civet	Dog	Raccoon	Bear	Otter	Seal	Walrus
Cat	0.5									
Hyena	0.4	0.1								
Civet	0.2	0.3	0.2							
Dog	0.2	0.7	0.6	0.4						
Raccoon	0.2	0.7	0.6	0.4	0					
Bear	0.4	0.9	0.8	0.6	0.2	0.2				
Otter	0.3	0.8	0.7	0.5	0.1	0.1	0.1			
Seal	0.5	1	1	0.7	0.3	0.3	0.1	0.2		
Walrus	0.5	1	1	0.7	0.3	0.3	0.1	0.2	0	
Sea lion	0.5	1	1	0.7	0.3	0.3	0.1	0.2	0	0

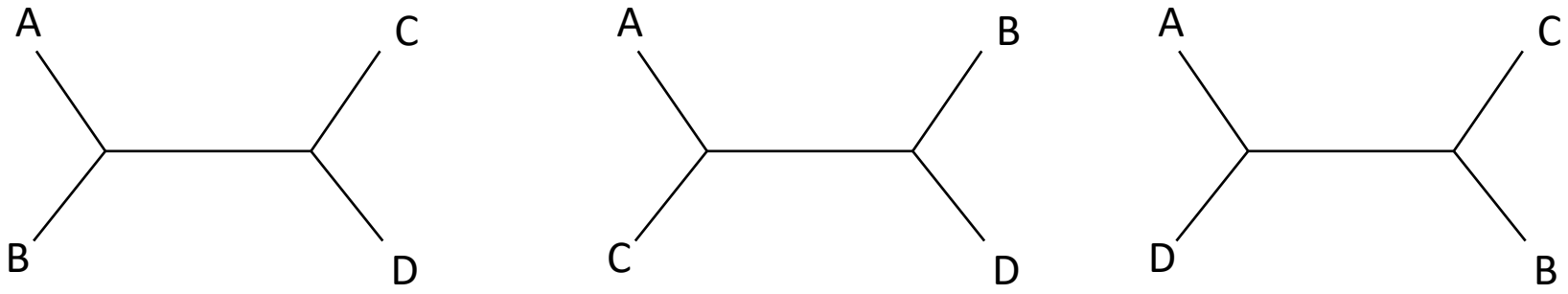
Tree



Baum and Smith. Tree Thinking: An Introduction to Phylogenetic Biology. 2012. Chapter 8

Finding the “best tree”: enumerating trees

Four-taxon unrooted tree:



Labeled leaves	Unrooted Trees	Rooted Trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425

Total number of unrooted trees:

$$(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Total number of rooted trees:

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Approaches

Distance methods

- Construct trees from pairwise distance matrix
- Uses ***similarity*** as a proxy for relatedness
- ***UPGMA*** – heuristic, assumes equal rates
- ***Neighbor-joining*** – works efficiently, except with homoplasy and unobserved changes

Optimization methods

- Use evolutionary models to define ***optimality criteria***
  - *"How well does this tree account for the observed character?"*
- Measures the optimality criteria for *many* trees, then picks the best value
- Methods include:
  - **Maximum parsimony** – the best tree is the one with the least number of changes
    - based on "Occam's Razor"
    - seeks to minimize homoplasy
  - **Maximum likelihood** – the best tree has the highest likelihood of observing the data

Why parsimony is wrong

Long branch attraction

- Parsimony assumes all lineages evolved at the same rate (Felsenstein 1981)
- Rapidly evolving taxa are grouped together
- There are ways to account for LBA, but it's a feature and not a bug

Bawono and Heringa. 2014. *Comprehensive Biomedical Physics.*

Maximum Likelihood Estimation

Asks the question: ***How well does the model fit the data?***

The ***likelihood*** is the probability of the data given the model

- P(D|H)
- Felsenstein (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution.*

Process:

1. Generate a large number of trees using a model of evolution
2. Calculate the likelihood for each tree
3. Choose the tree with the highest likelihood

Phylip, RaxML, PhyML, MEGA

Maximum Likelihood Estimation

To Compute Likelihood, You Need:

1. Data
2. A model of evolution
3. Hypothesis (trees)
4. Framework for calculating the likelihoods

Data:

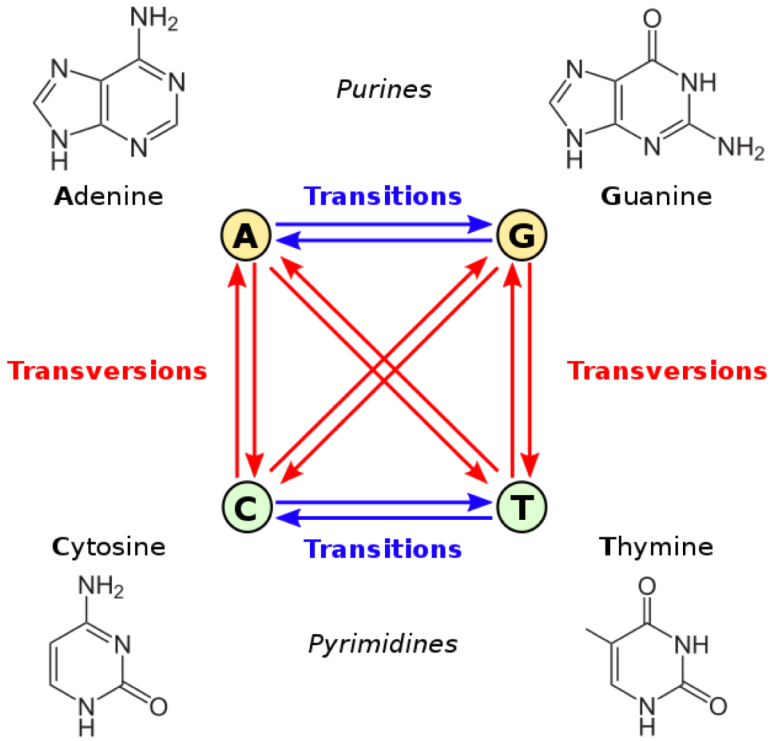
A matrix with:

- rows for taxa
- columns for characters
- cells are character states for each character for each taxon
- protein and DNA alignments are perfect

Model of DNA Sequence Evolution:

Based on the ***rate matrix (Q)***

This is the rate of change from one character state to another



https://upload.wikimedia.org/wikipedia/commons/8/8a/All_transitions_and_transversions.svg

DNA Substitution Models

Many different models to choose from

Jukes-Cantor: simplest model, assumes equal substitution rates

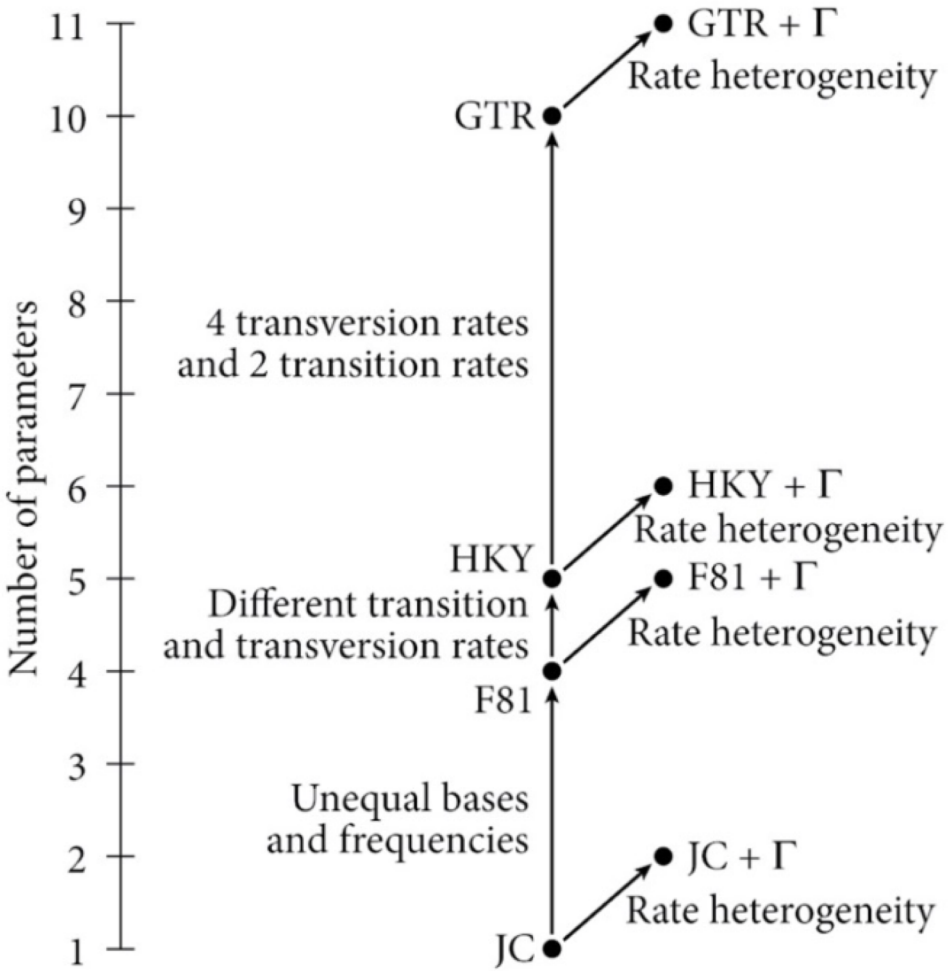
HKY: allows different Ts and Tv rates

GTR: most parameter-rich

Additional parameters such as rate heterogeneity (lambda) and rate of invariant sites (gamma)

Model testing

- AIC, Bayes Factors
- Software: JModelTest, Partitionfinder, MEGA

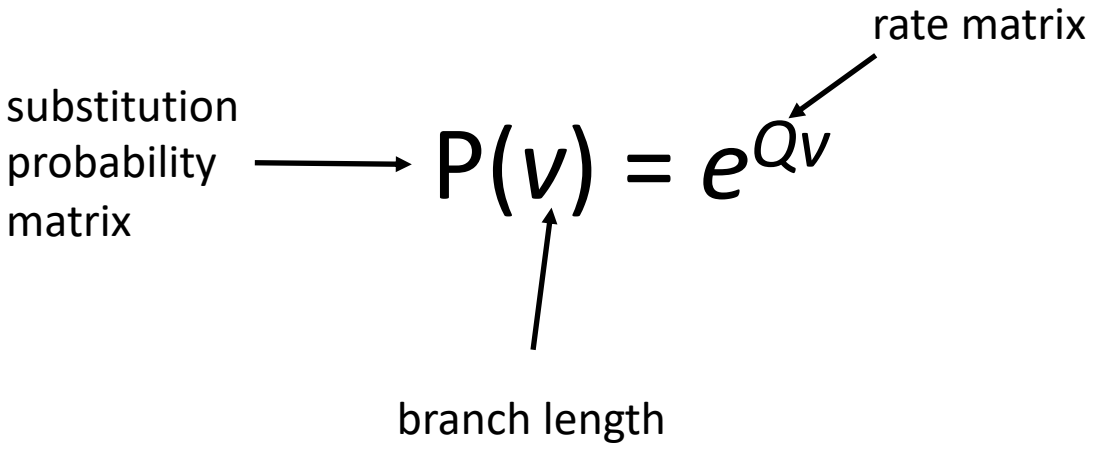


Baum and Smith. Tree Thinking: An Introduction to Phylogenetic Biology. 2012. Chapter 8

Calculating Likelihood

Substitution Probability Matrix

- The probability that a given substitution occurs in a given interval
- Intervals are in branch lengths.



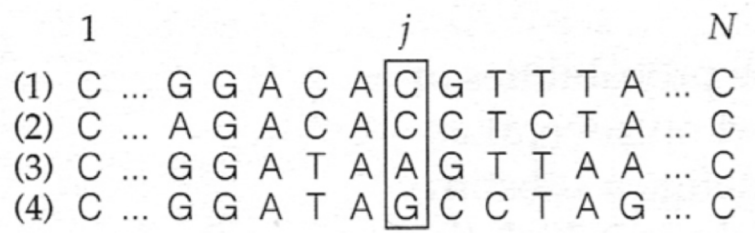
In an F81 model, when branch length is 0, **P** is a diagonal matrix

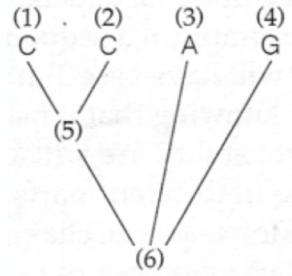
	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Calculating Likelihood

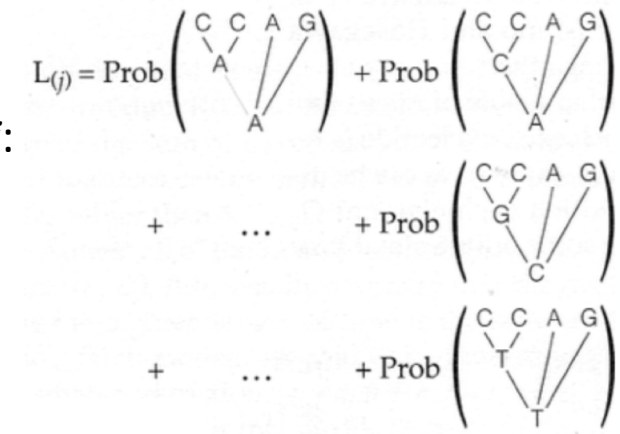
Calculate:

- Probability of observing the data in each column given the tree and the model
- Sum over alternative trees
- Probability of observing the entire matrix by multiplying probabilities across columns





Likelihood for site *j*:



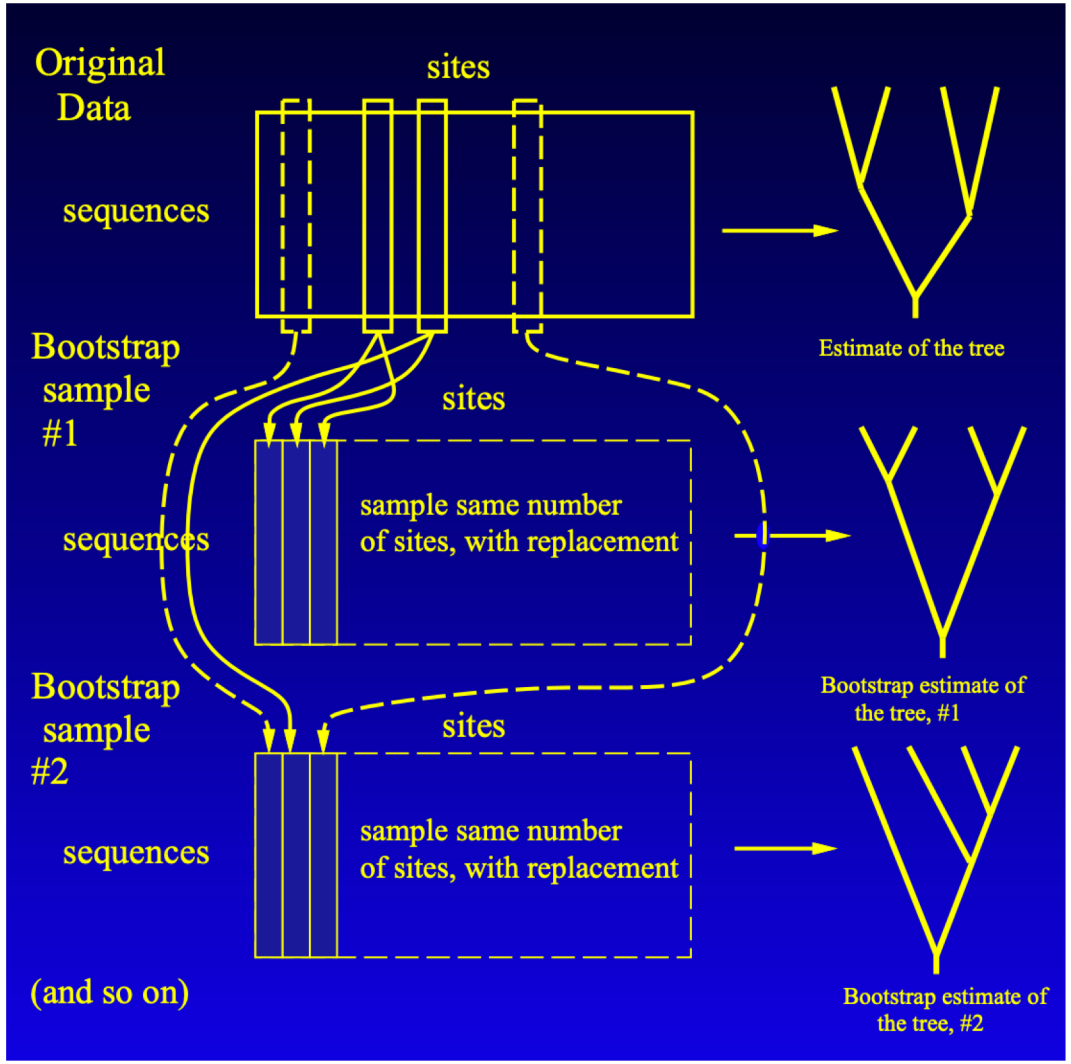
Swofford, 1996. Phylogenetic Inference. In: Molecular Systematics. Sinauer

Bootstrapping

Sampling with Replacement

- repeats the analysis iteratively
- removes sites, then replaces
- generates trees from replicates
- calculates the frequency of bipartitions across all bootstrapped trees

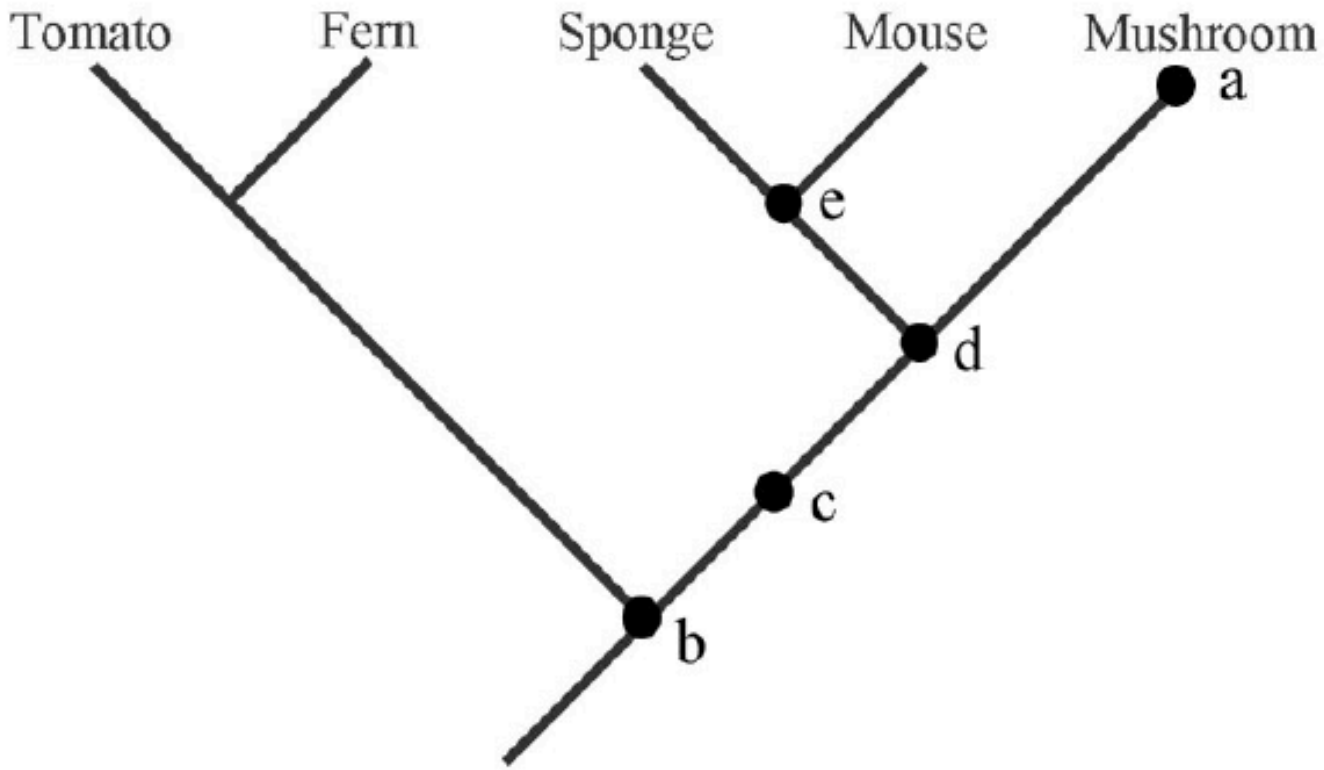
Allows you to report the frequency of each bipartition in your most likely tree ***as a measure of statistical support*** for your clades.



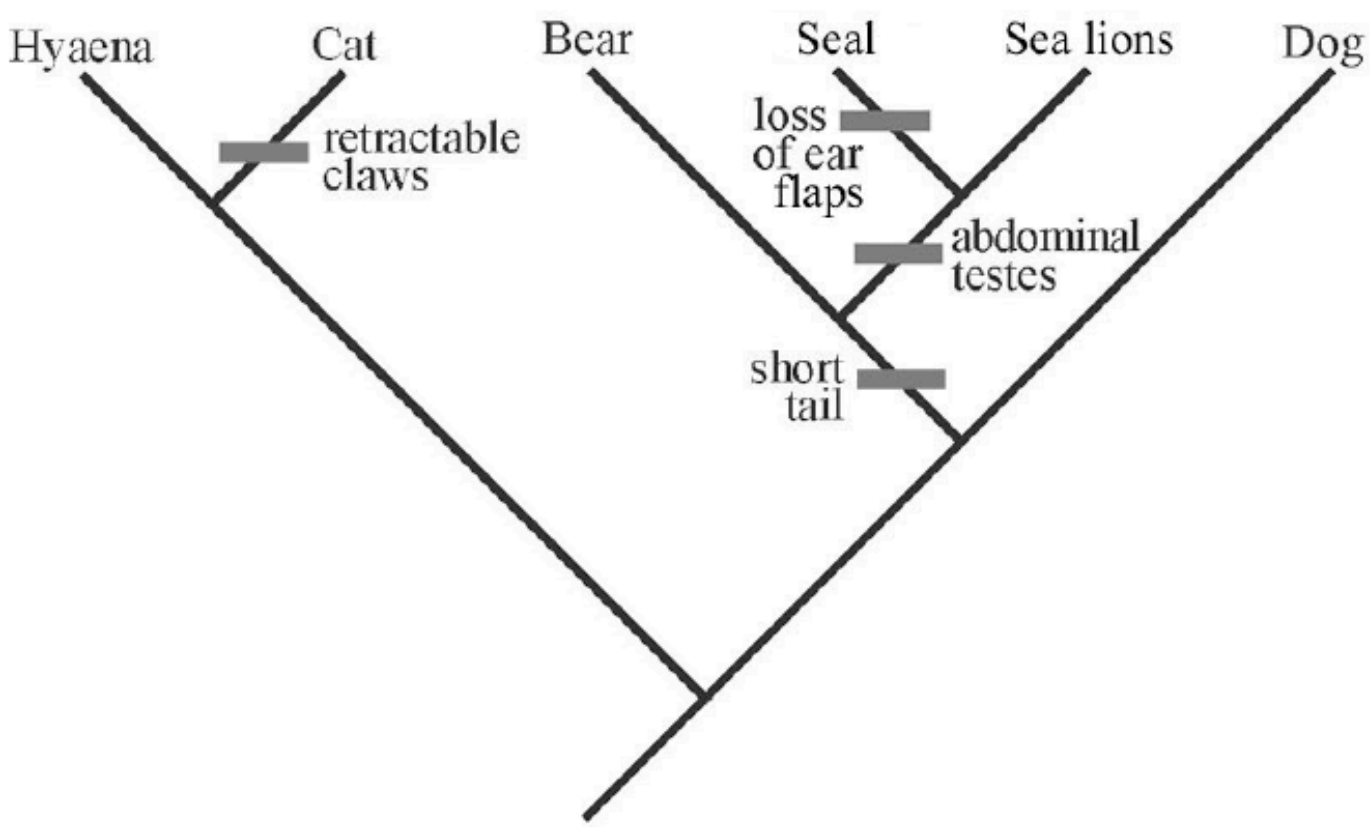
Joe Felsenstein Lecture, Univ. of Washington

Next

1. Phylogenetics lab Th: build trees from data
2. Phylogenomics 2 3/3: we will cover Bayesian methods and divergence time estimation

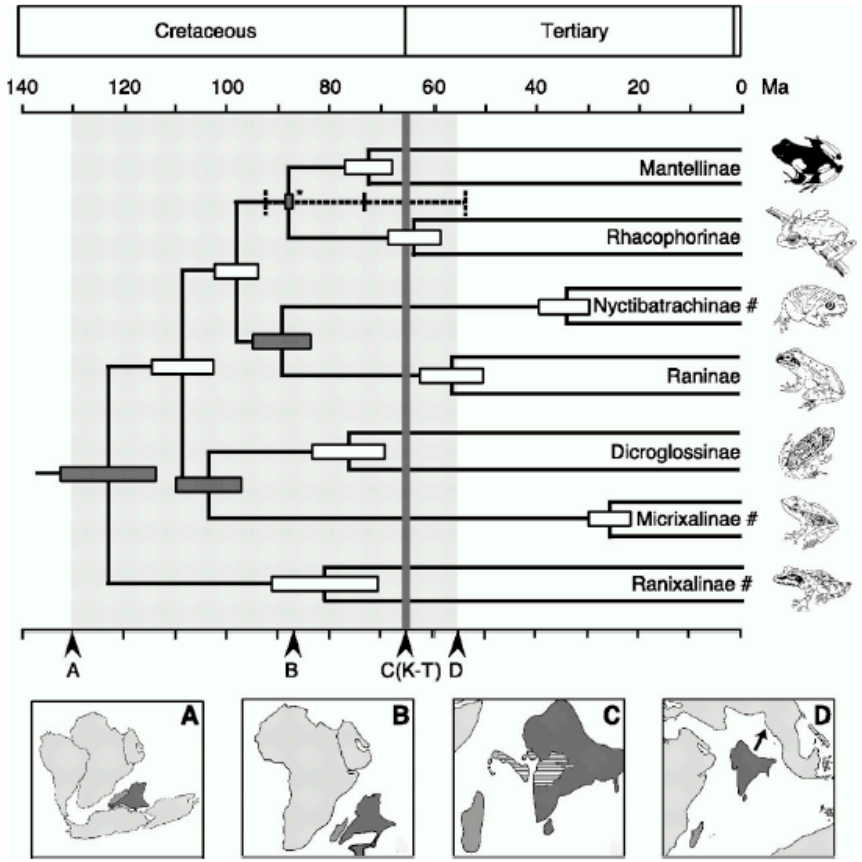


4) Which of the five marks in the tree above corresponds to the most recent common ancestor of a mushroom and a sponge?



9) In the above tree, assume that the ancestor had a long tail, ear flaps, external testes, and fixed claws. Based on the tree and assuming that all evolutionary changes in these traits are shown, what traits does a sea lion have?

a) long tail, ear flaps, external testes, and fixed claws
b) short tail, no ear flaps, external testes, and fixed claws
c) short tail, no ear flaps, abdominal testes, and fixed claws
d) short tail, ear flaps, abdominal testes, and fixed claws
e) long tail, ear flaps, abdominal testes, and retractable claws



F. Bossuyt, M. C. Milinkovitch. Amphibians as indicators of early tertiary "out-of-India" dispersal of vertebrates. *Science* **292**, 93 (2001).

3) This tree depicts inferred relationships among some major frog groups with branches drawn proportional to absolute time. Error bars on internal nodes depict confidence intervals on the dates of estimated nodes. Assuming this tree and the associated ages are correct which of the following statements is true?

a) No individual living before 70 million years ago is an ancestor of Raninae
b) Raninae and Dicroglossinae shared a common ancestor about 75 million years ago
c) The divergence of Raninae and Nyctibatrachinae occurred more recently than the 85 million year old separation of India from Madagascar
d) The last common ancestor of Micrixalinae and Dicroglossinae lived before India and Madagascar separated (85 million years ago)