Tree-thinking and basic approaches to building phylogenies

Nothing in biology makes sense except in the light of evolution.

— Theodosius Dobzhansky, 1973

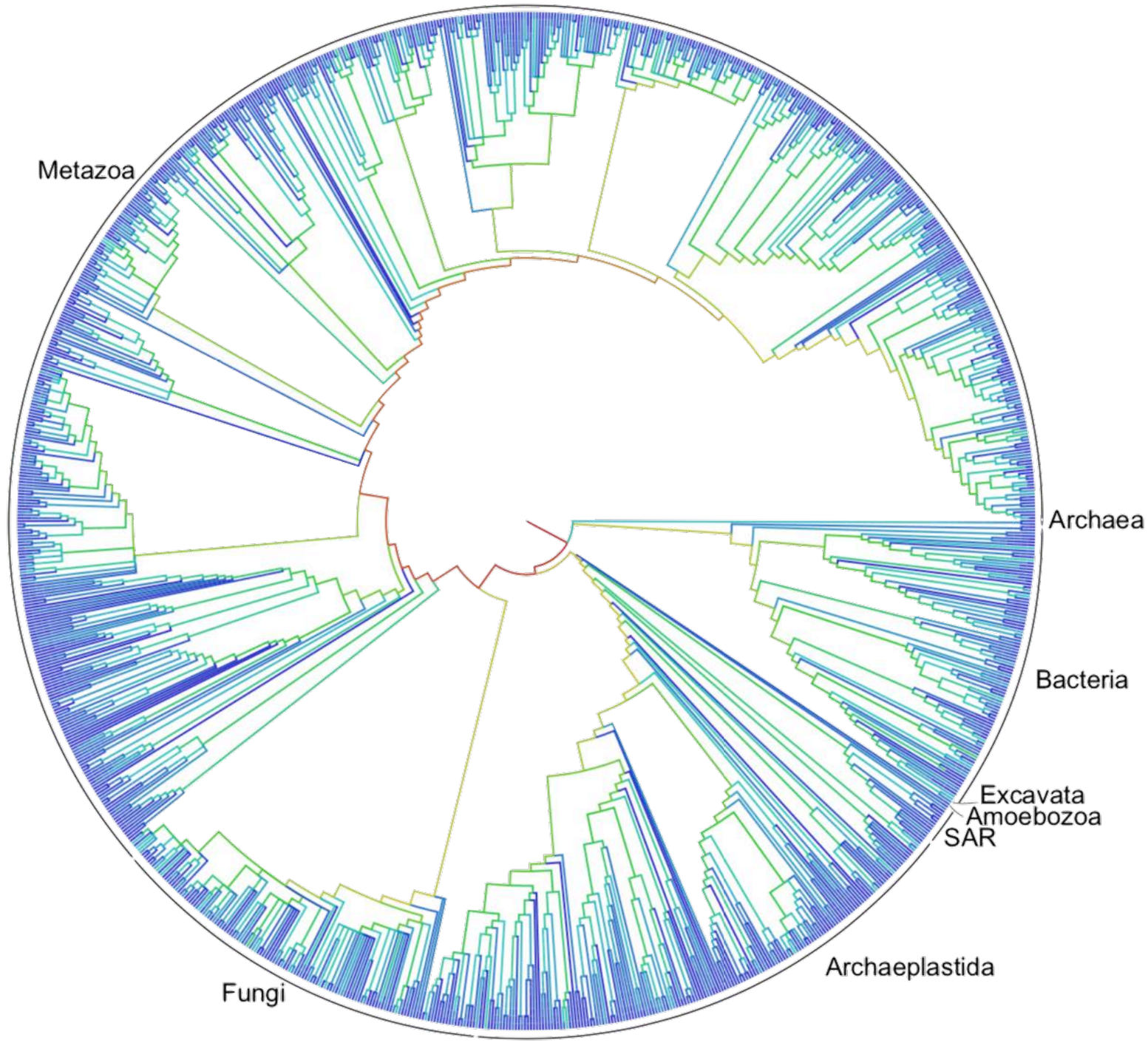
Nothing in evolution makes sense except when seen in the light of phylogeny.

— Jay Savage, 1997

The Tree of Life

Three billion years the Tree has grown
From replicators' first seed sown
To branches rich with progeny:
The wonder of phylogeny.

excerpt from the poem "The Tree of Life"
by David Maddison

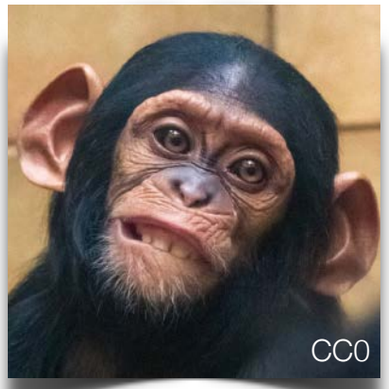


Open Tree of Life Synthetic Tree from Hinchliff et al. (PNAS 2015)

Phylogeny Terminology & Concepts

a taxon

taxa

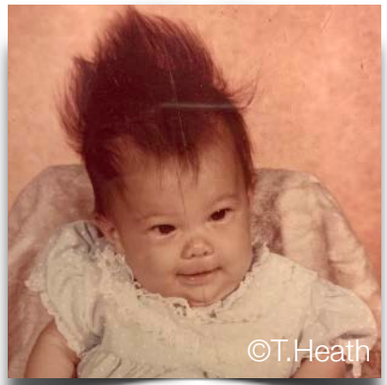


tip

node

branches

Phylogeny Terminology & Concepts

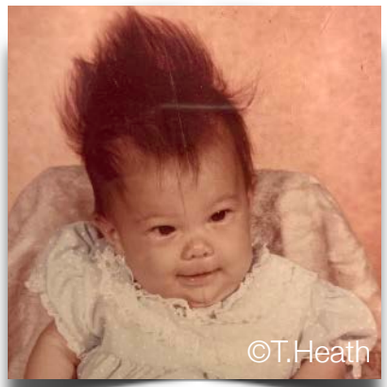


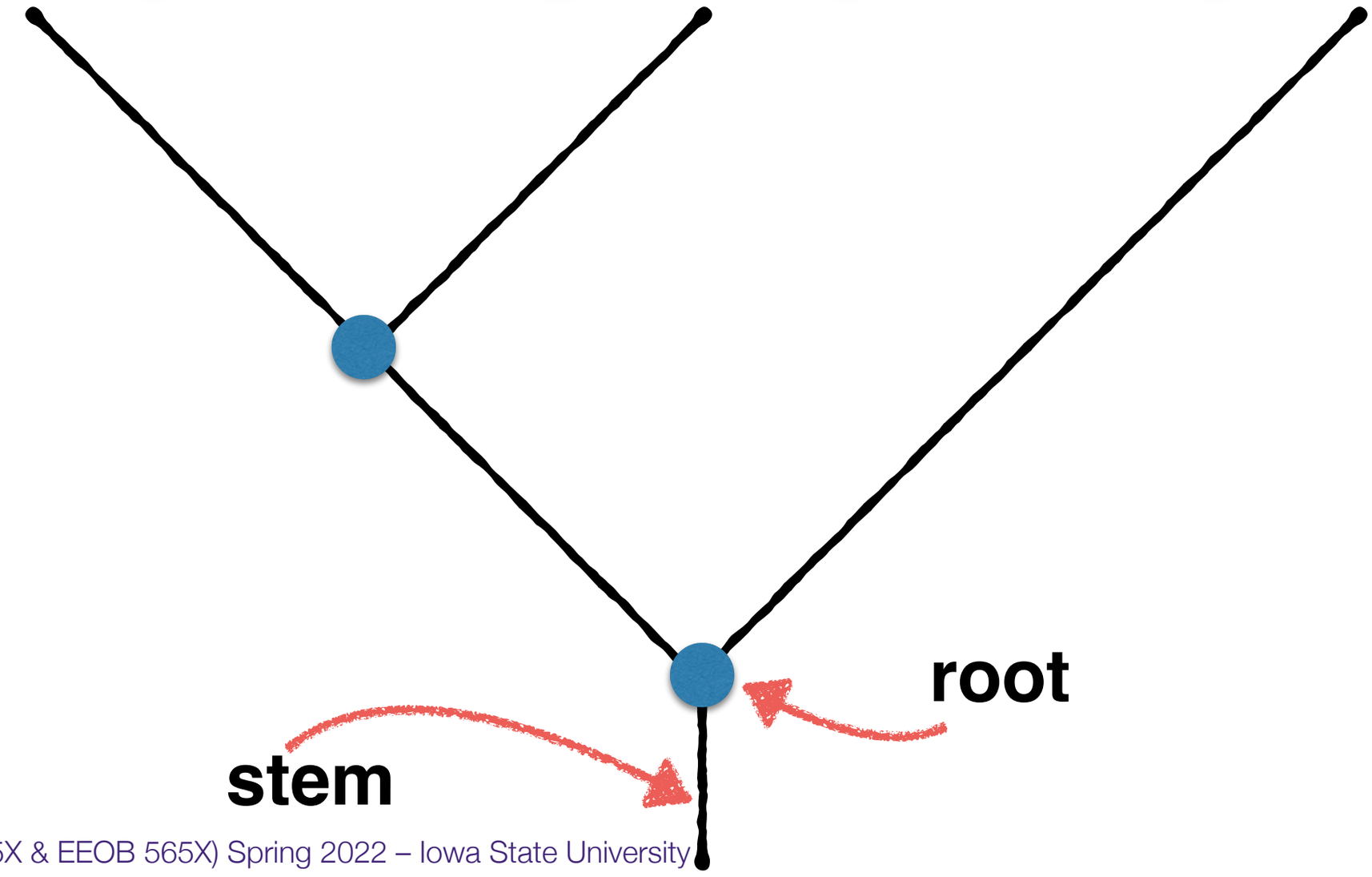
leaf

vertex

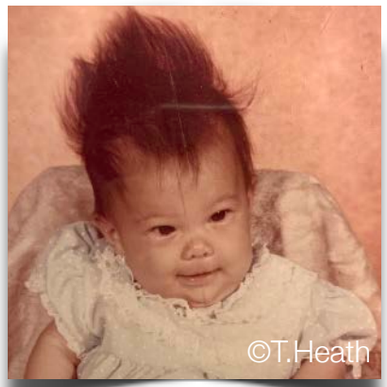
edges

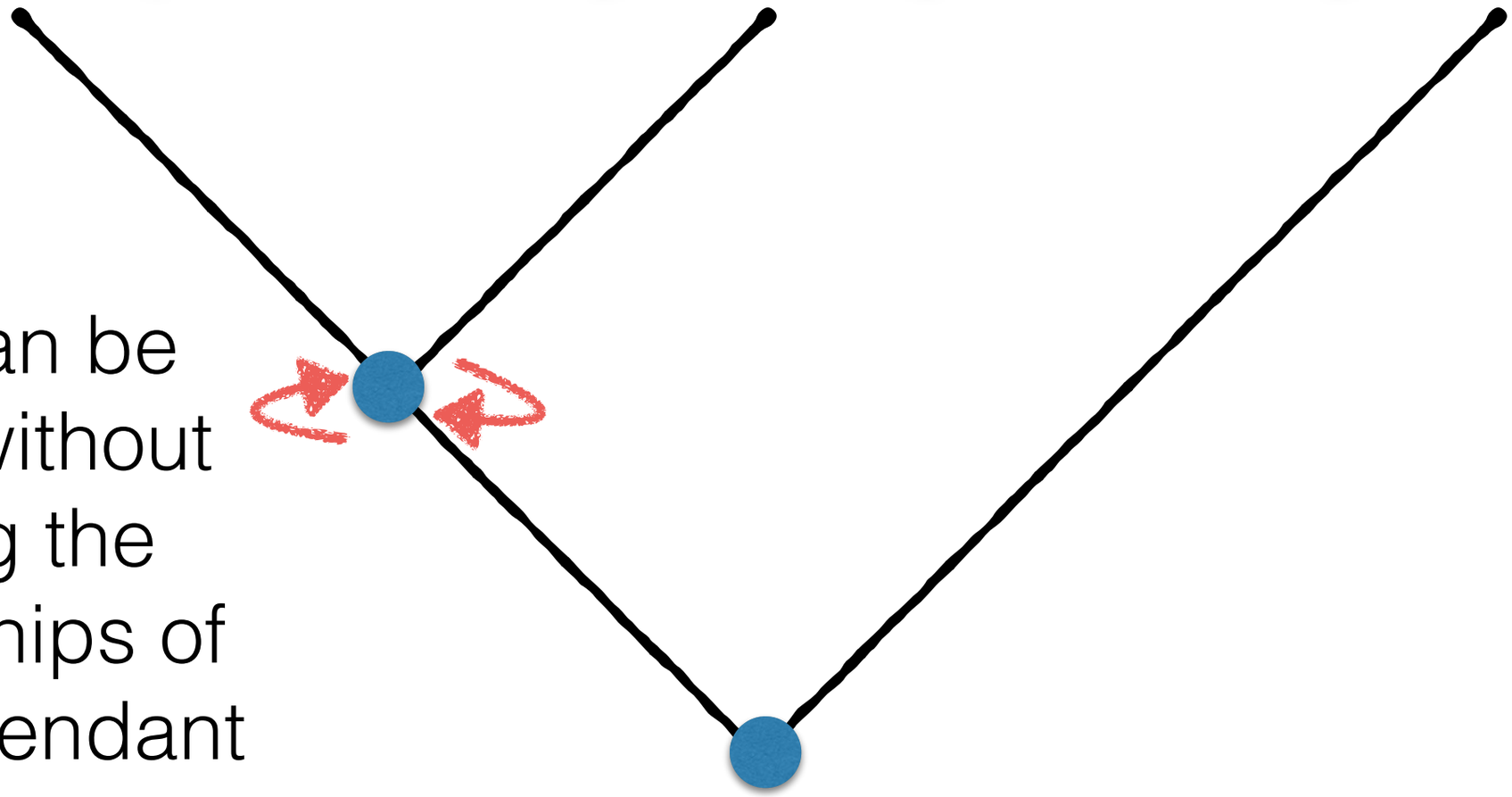
Phylogeny Terminology & Concepts





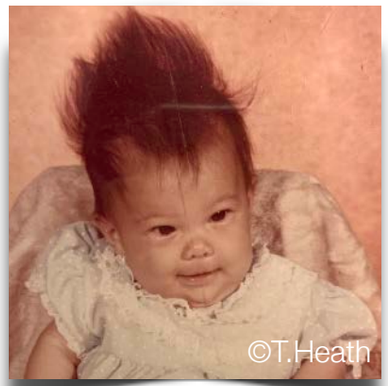
Phylogeny Terminology & Concepts

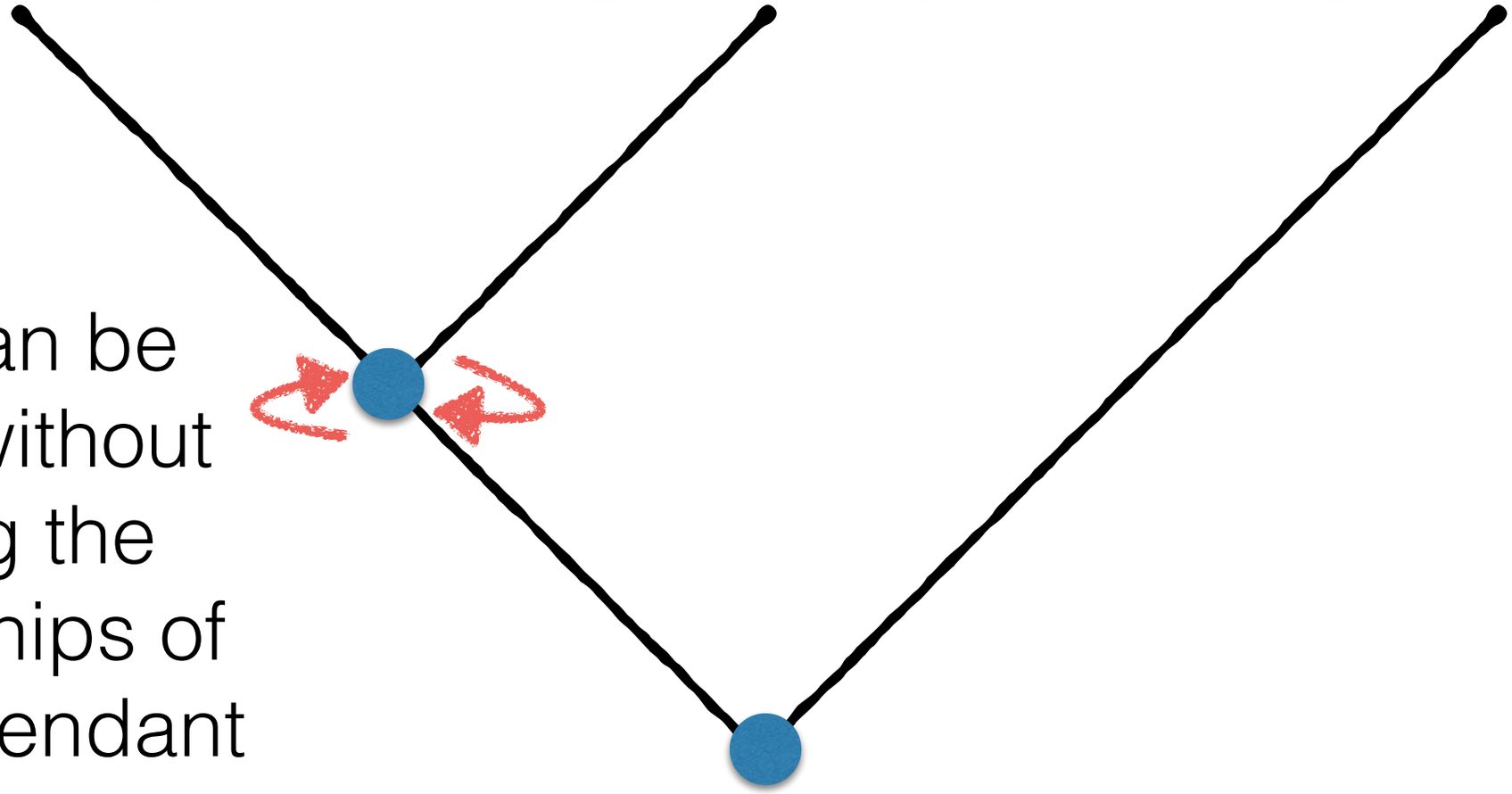




nodes can be rotated without changing the relationships of the descendant branches

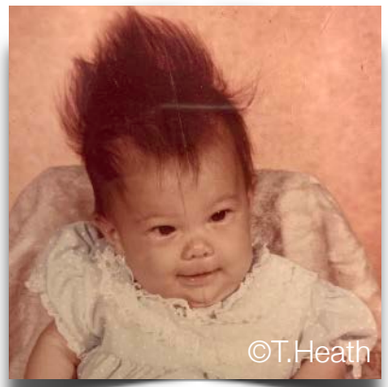
Phylogeny Terminology & Concepts

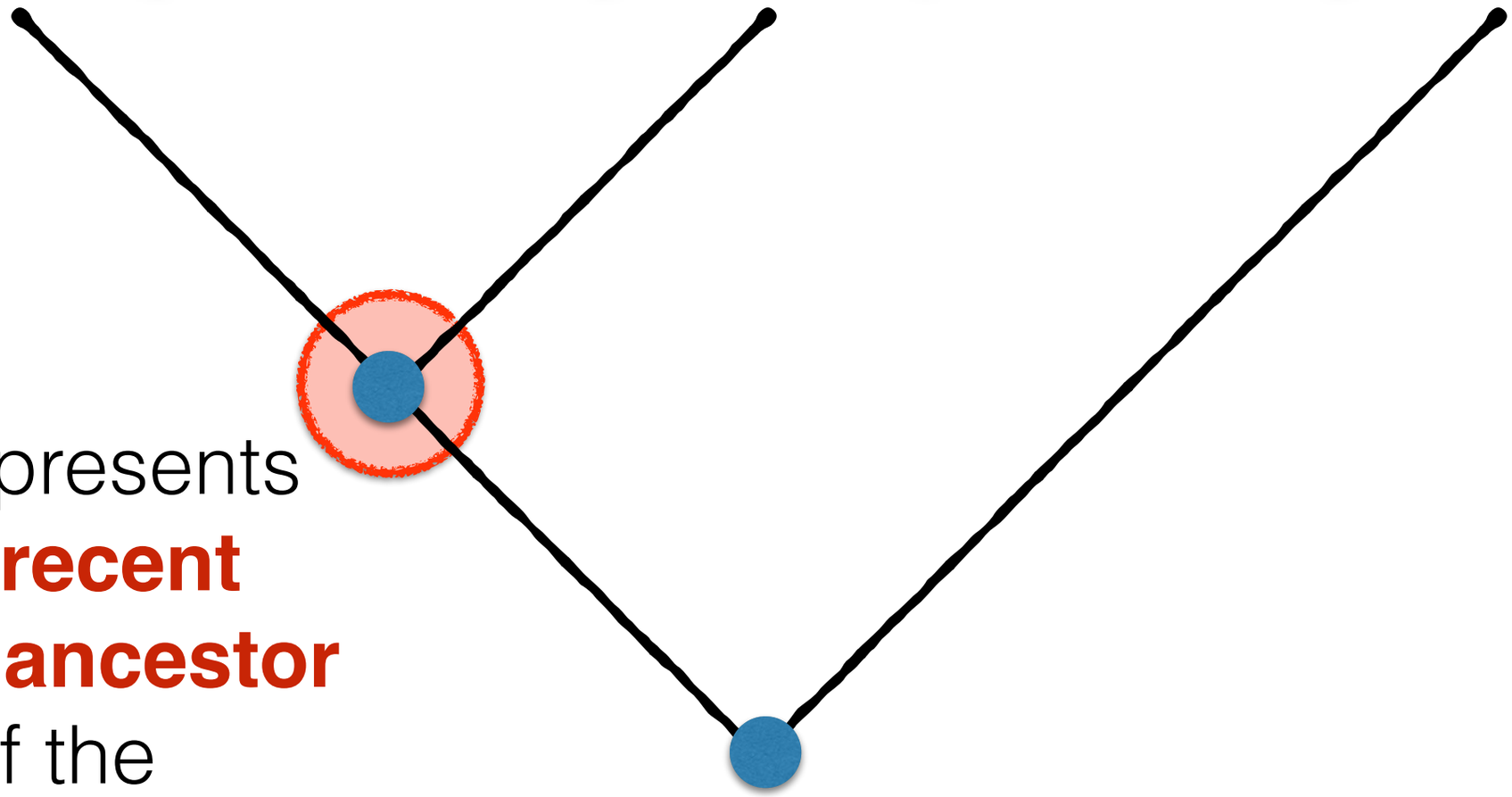




nodes can be rotated without changing the relationships of the descendant branches

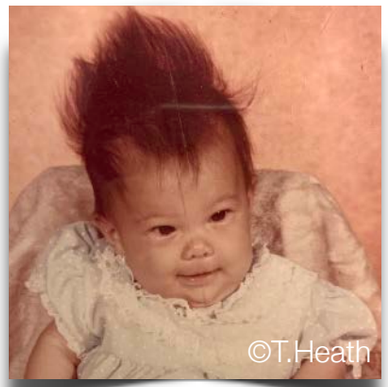
Phylogeny Terminology & Concepts

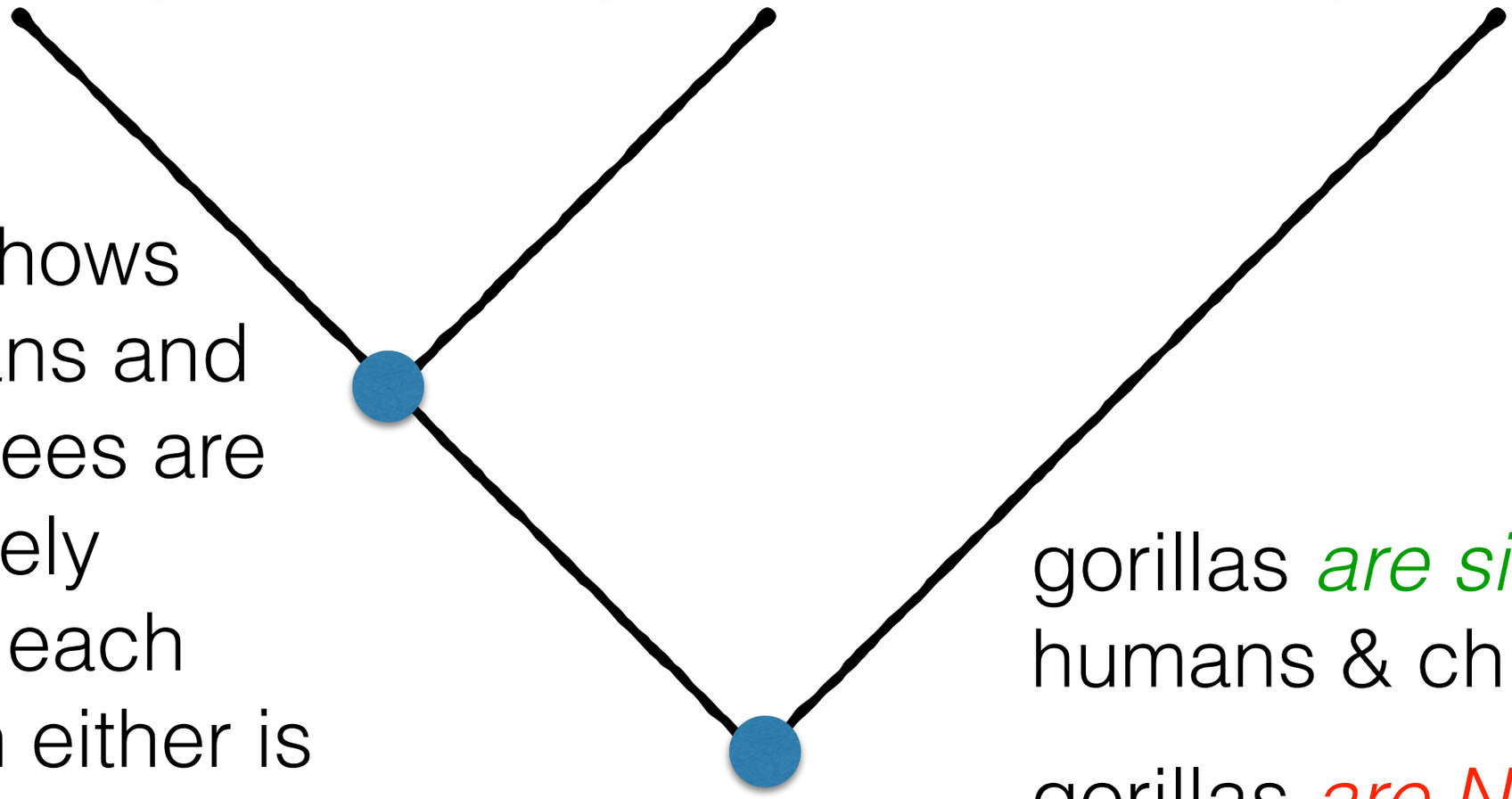




a node represents the **most-recent common ancestor** (MRCA) of the descendant taxa

Phylogeny Terminology & Concepts



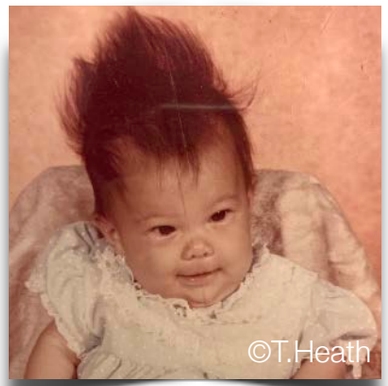


this tree shows that humans and chimpanzees are more closely related to each other than either is to gorillas

gorillas *are sister to* humans & chimps

gorillas *are NOT basal* to chimps & humans

Phylogeny Terminology & Concepts

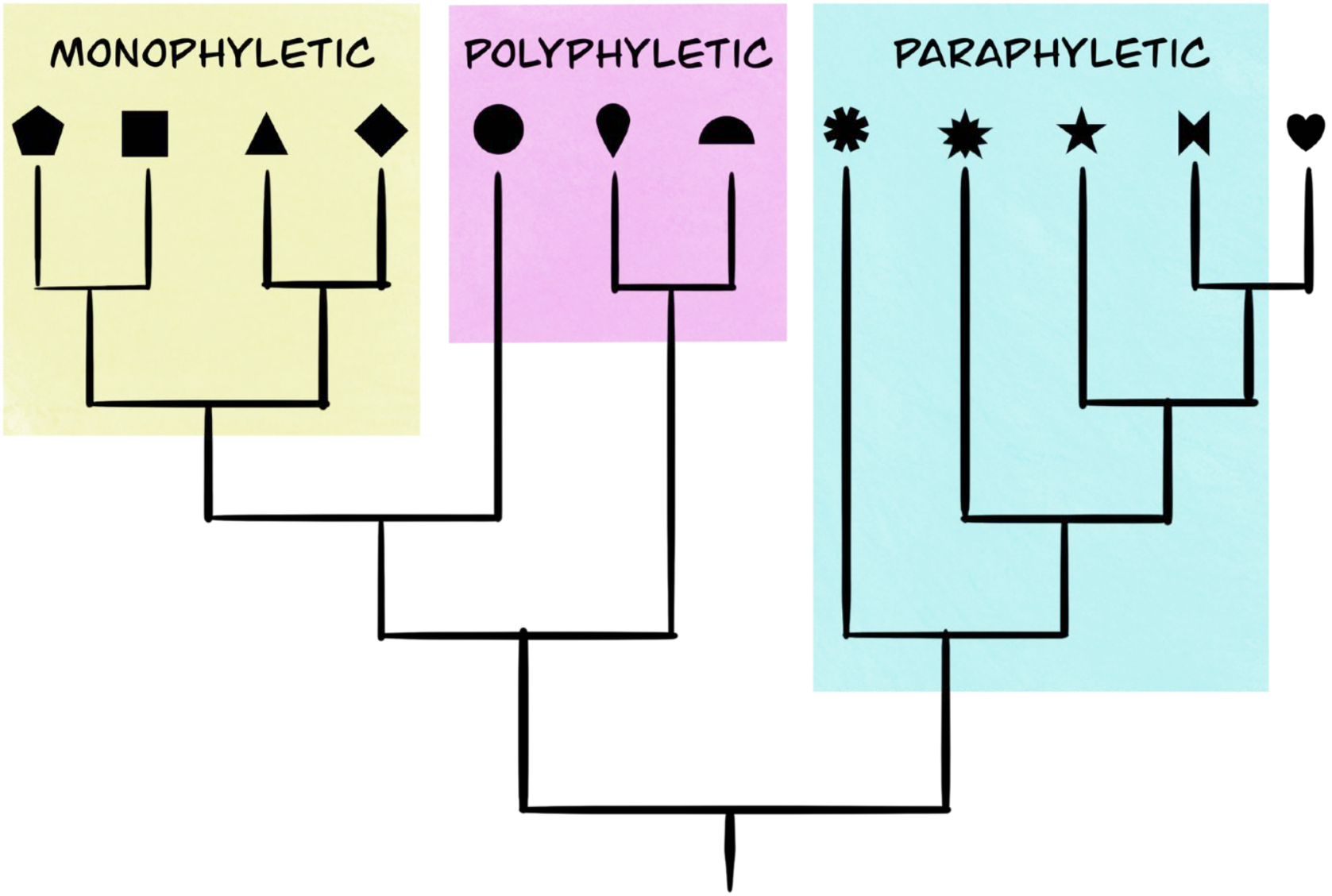


the term basal refers to something that is "closer to the base", so please never use this for extant taxa

the 10 Mya fossil *Chororapithecus abyssinicus* can be called a basal ape (Sewa et al 2007)

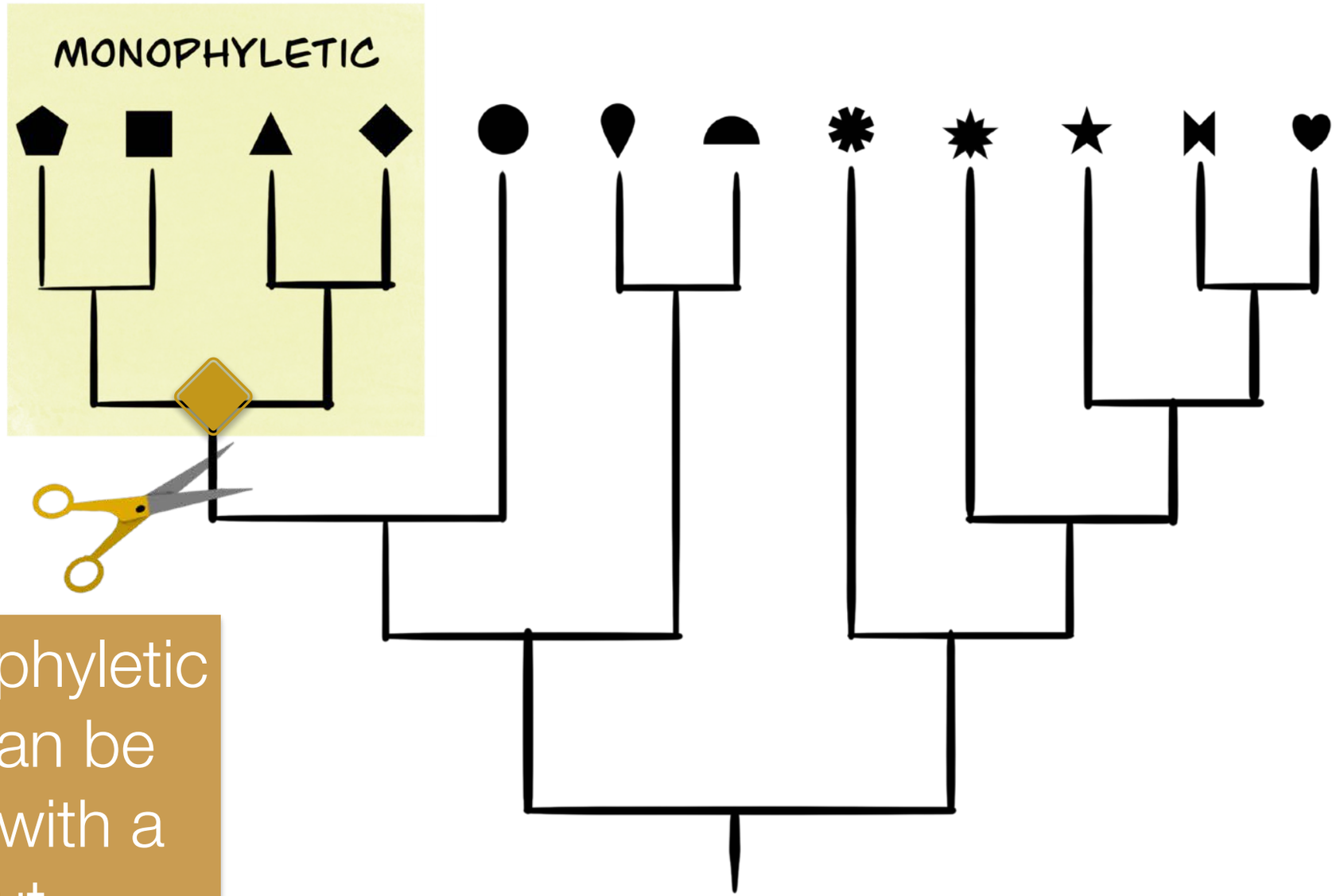
see: http://for-the-love-of-trees.blogspot.com/2016/09/the-ancestors-are-not-among-us.html

Groups in a Phylogeny



Monophyly

a monophyletic group includes an ancestor and all of its descendants

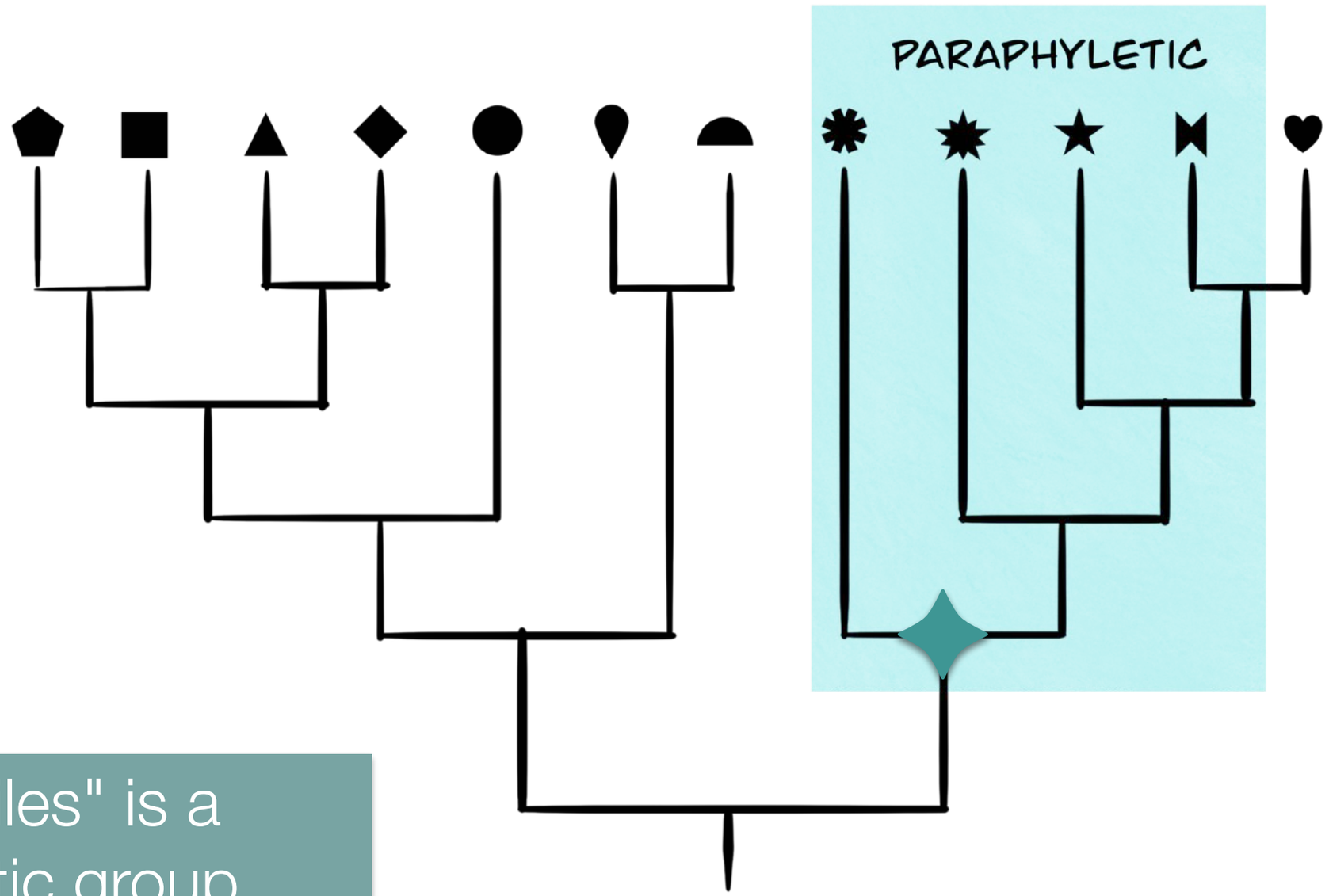


a monophyletic group can be pruned with a single cut

monophyletic group = clade

Paraphyly

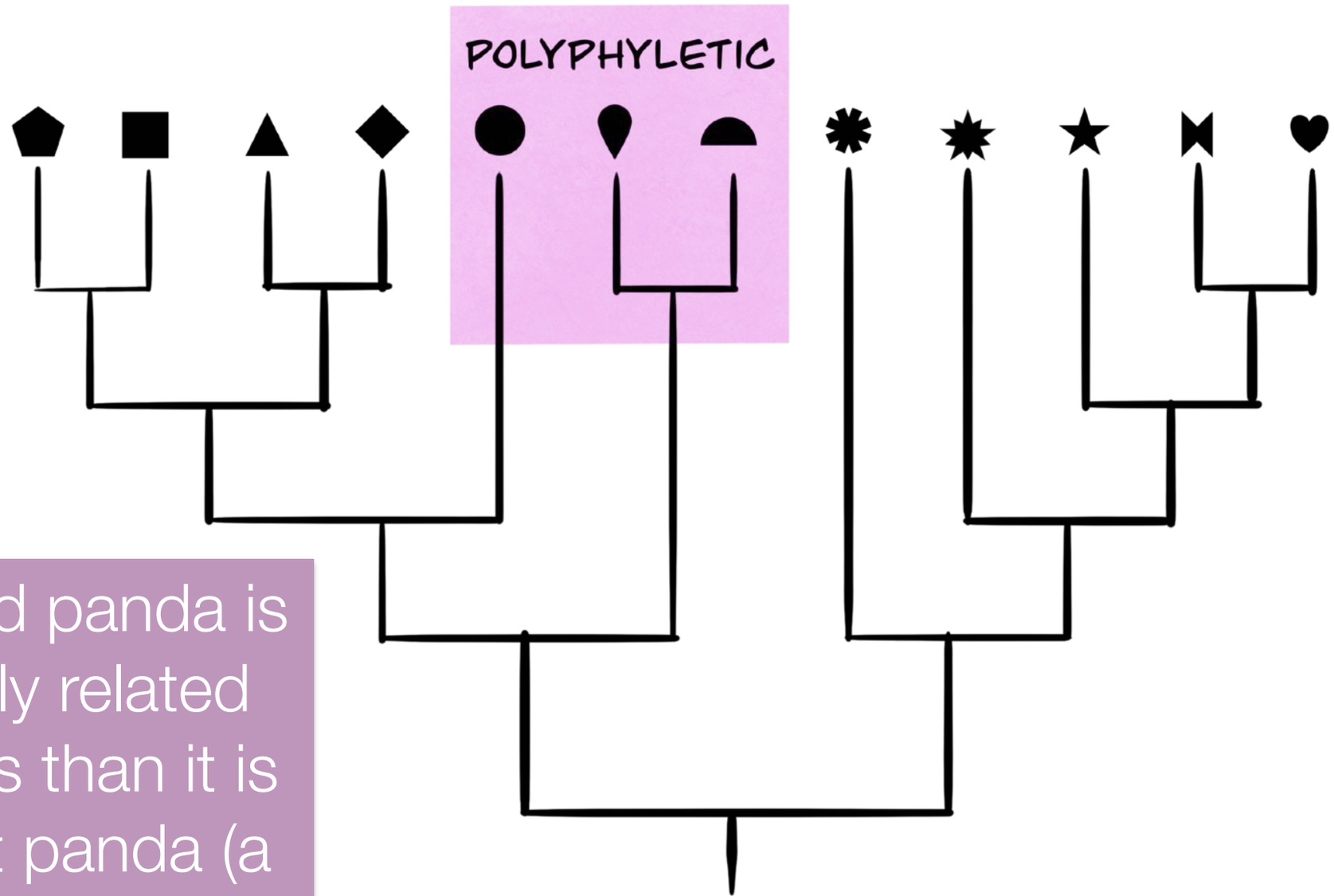
a paraphyletic group includes an ancestor and a subset of its descendants



e.g., "reptiles" is a paraphyletic group unless it includes birds

Polyphyly

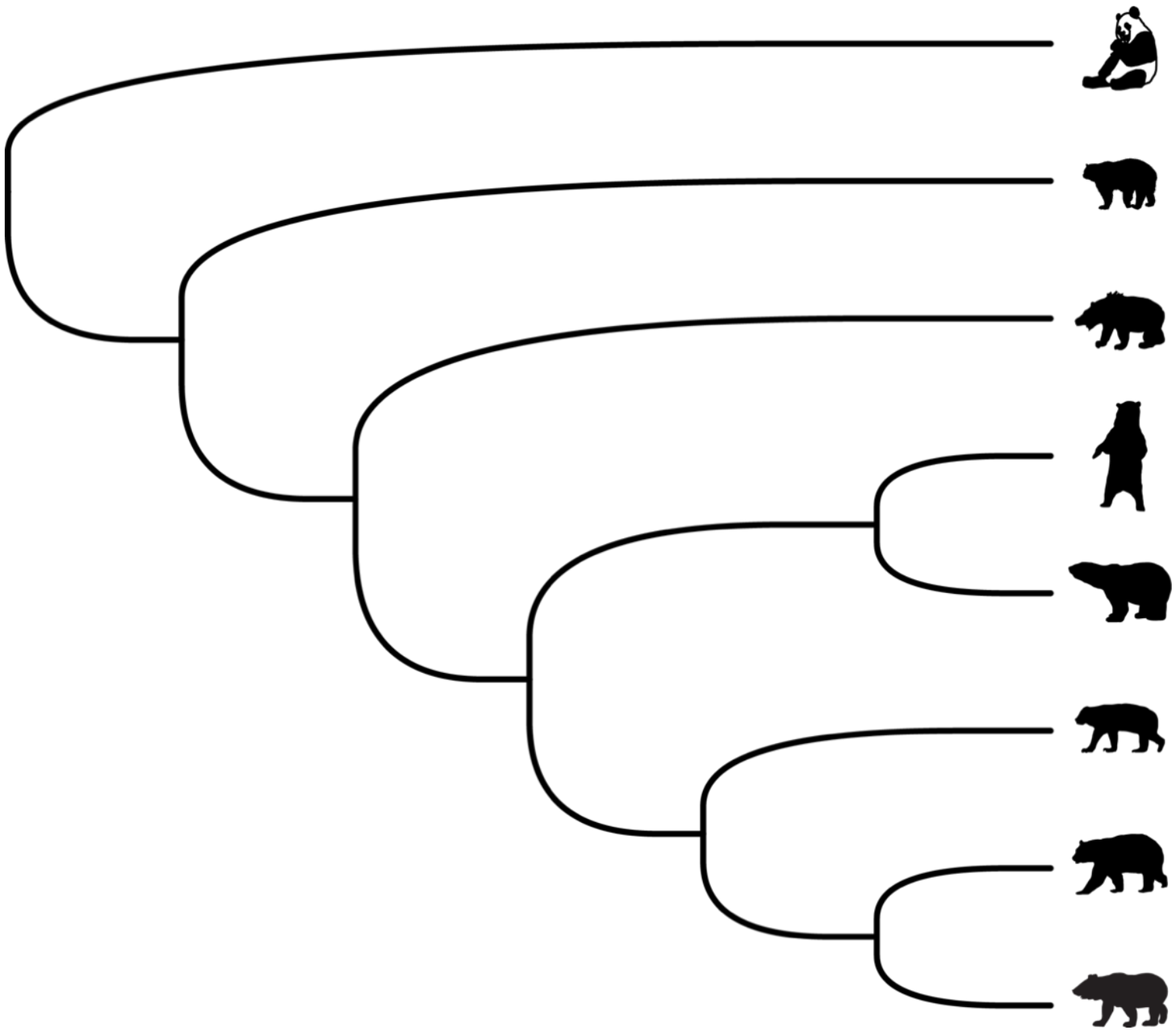
a polyphyletic group includes a set of taxa, but not their common ancestor



e.g., the red panda is more closely related to raccoons than it is to the giant panda (a bear), so "pandas" is a polyphyletic group

Phylogeny: Branch Lengths

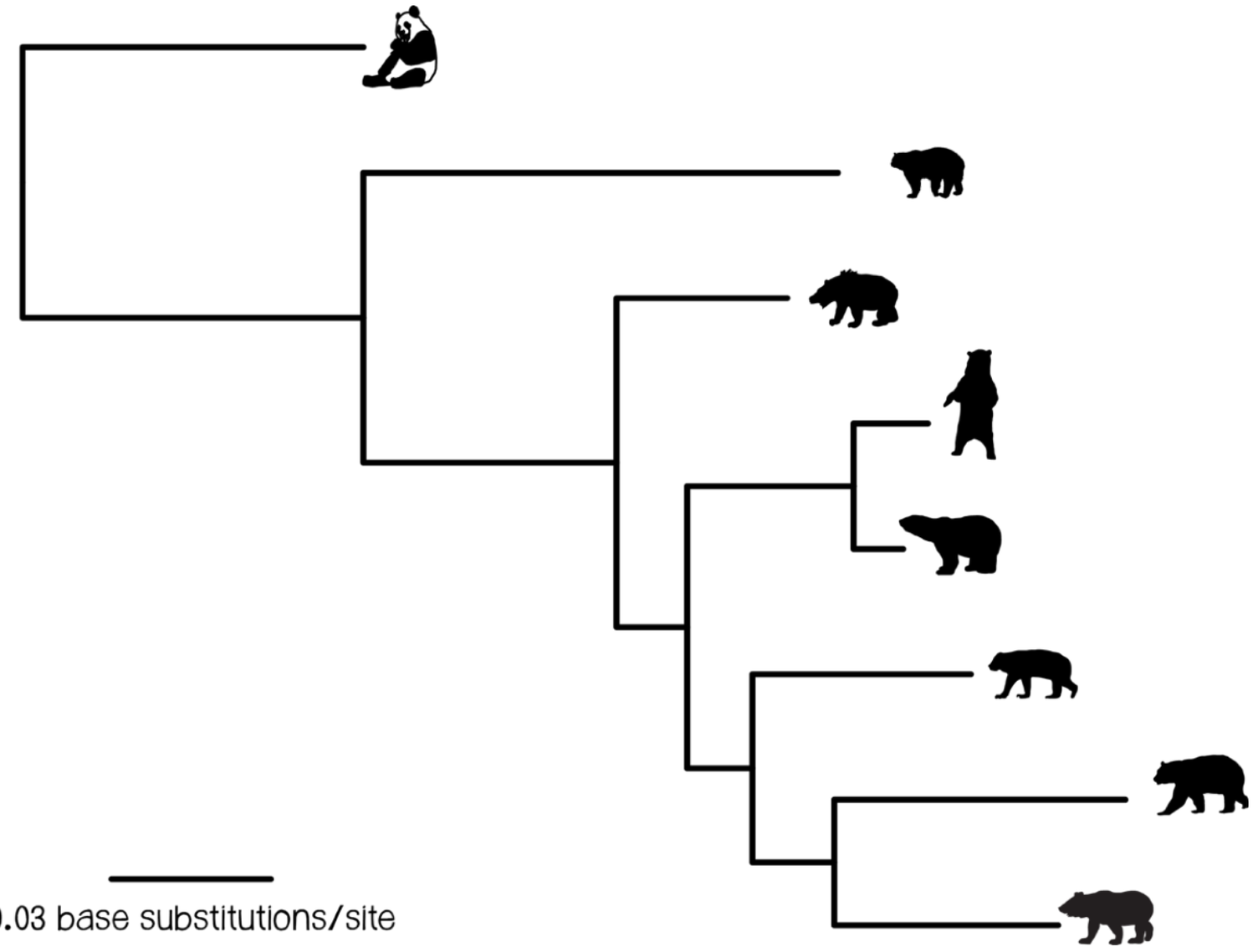
can have no meaning & just show the pattern of relationships



(figure adapted from Heath et al. 2014)

Phylogeny: Branch Lengths

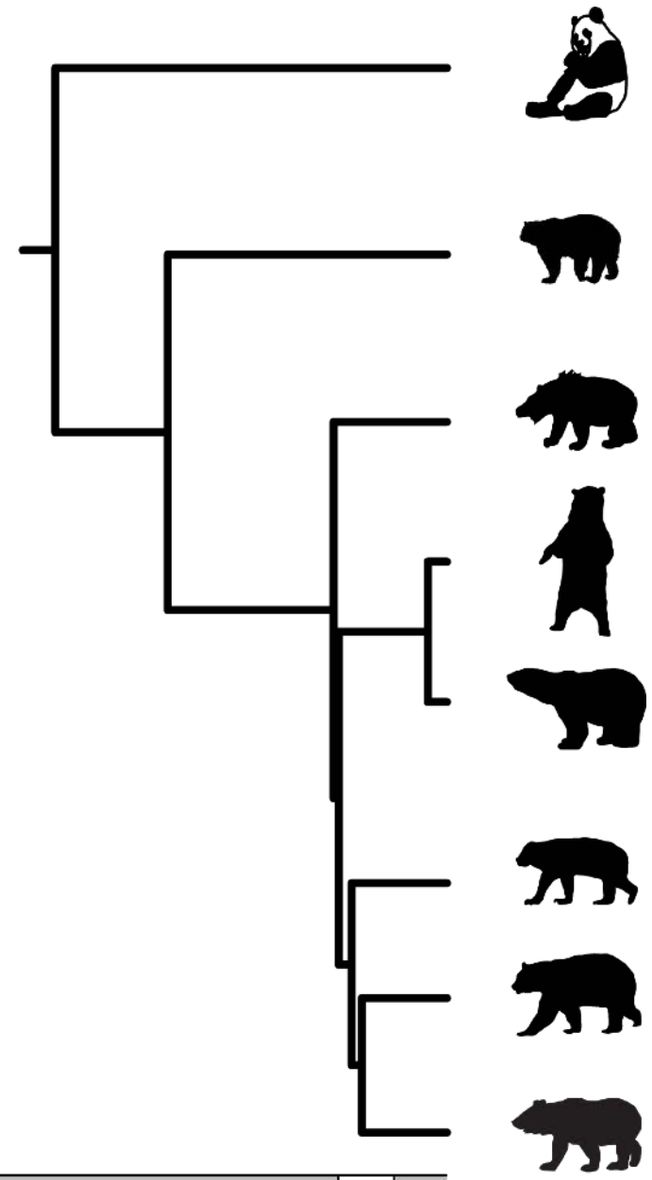
can represent the amount of **genetic difference**



(figure adapted from Heath et al. 2014)

Phylogeny: Branch Lengths

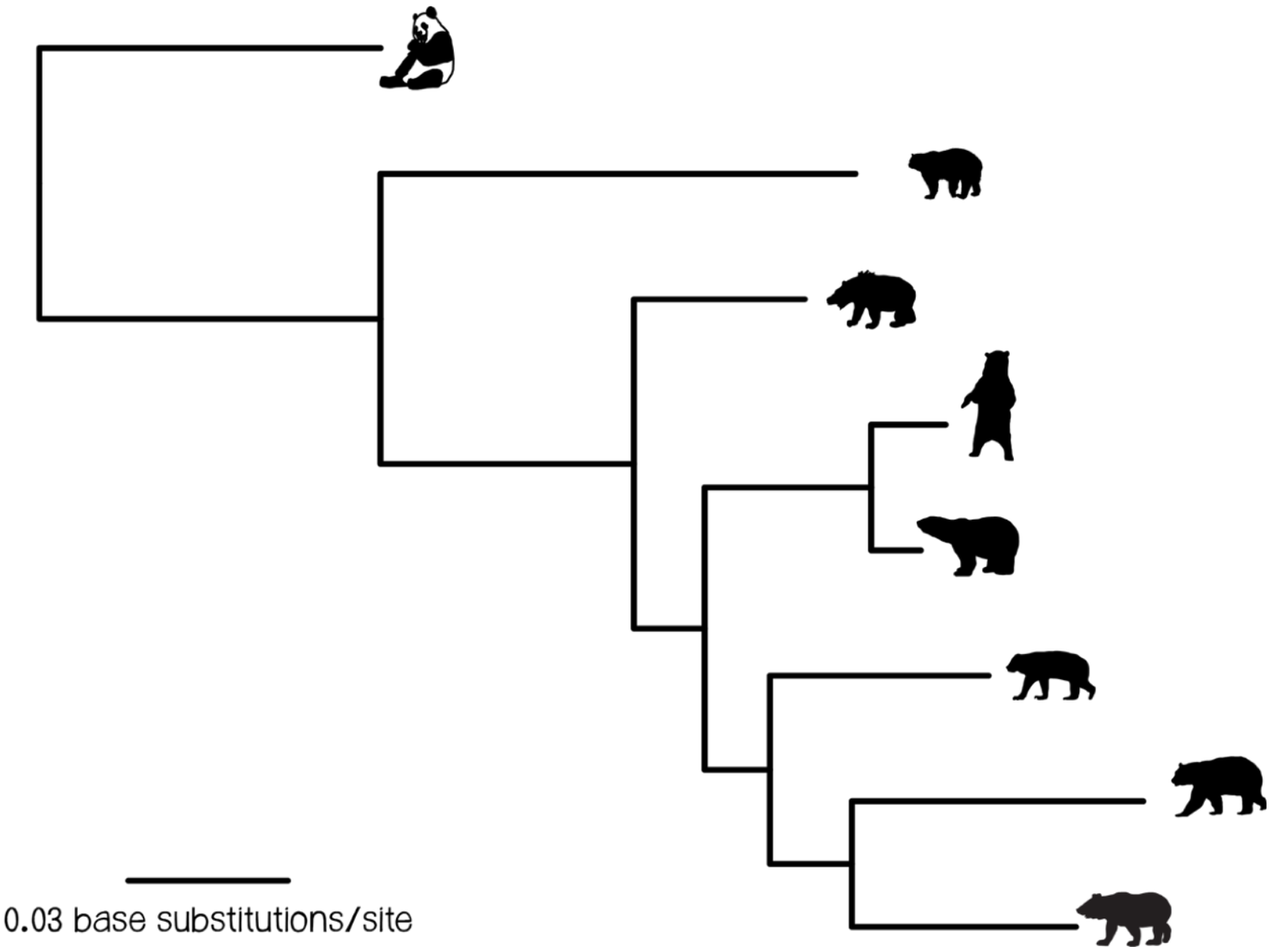
can represent the **duration of time** between nodes



(figure adapted from Heath et al. 2014)

Phylogeny: Rooting

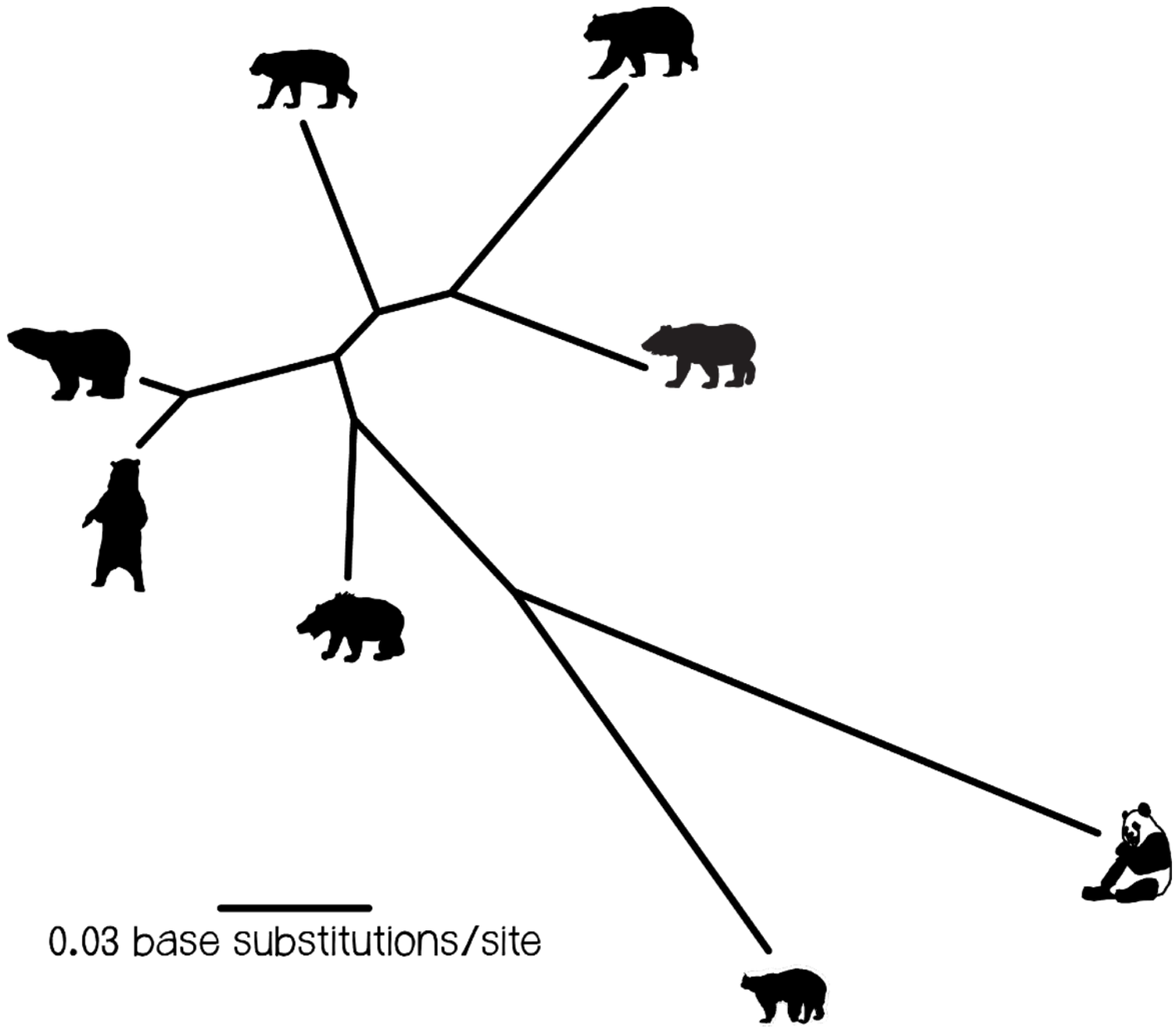
a tree can be **rooted** to show the direction or relative timing of divergence



(figure adapted from Heath et al. 2014)

Phylogeny: Rooting

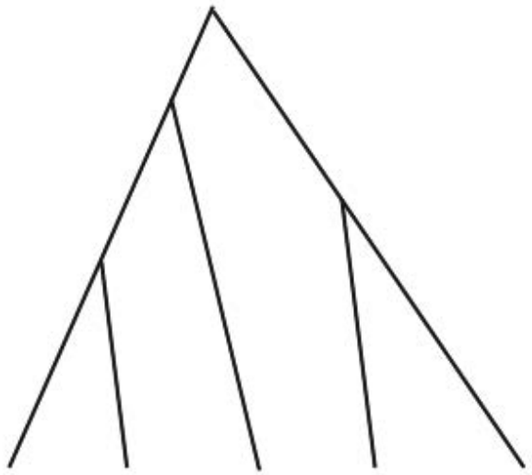
a tree can be **unrooted** showing only relationships among lineages



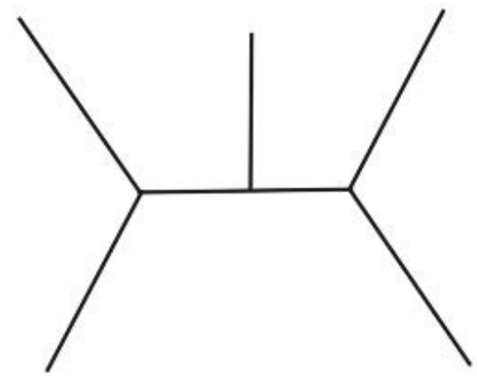
(figure adapted from Heath et al. 2014)

Resolved Phylogeny

a tree can be **bifurcating** or **binary** when all nodes split into only 2 descendants (such a tree is also called "resolved")



fully-resolved tree

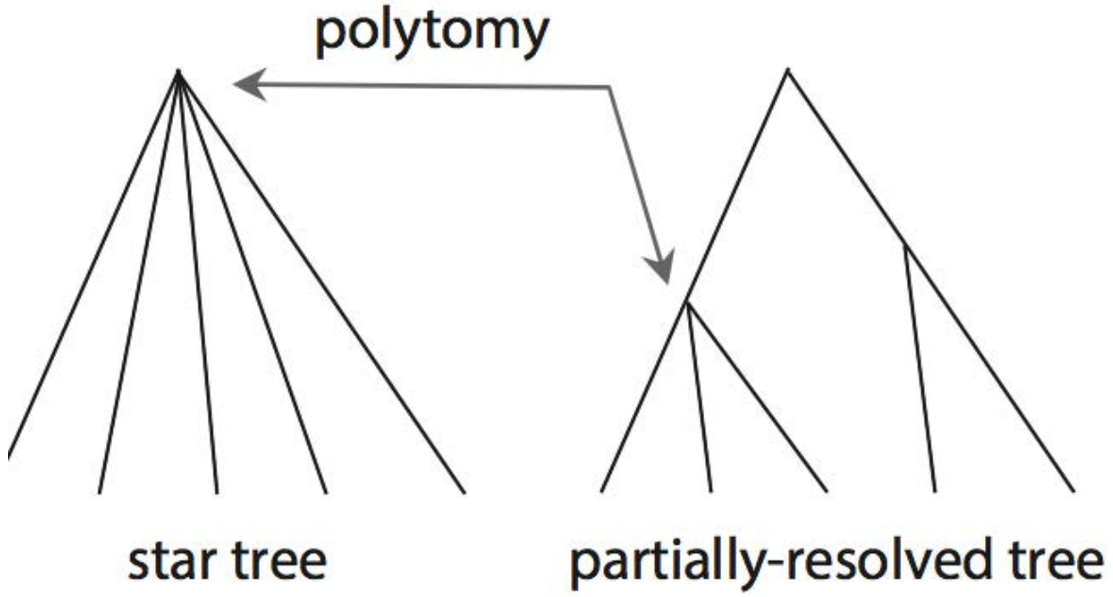


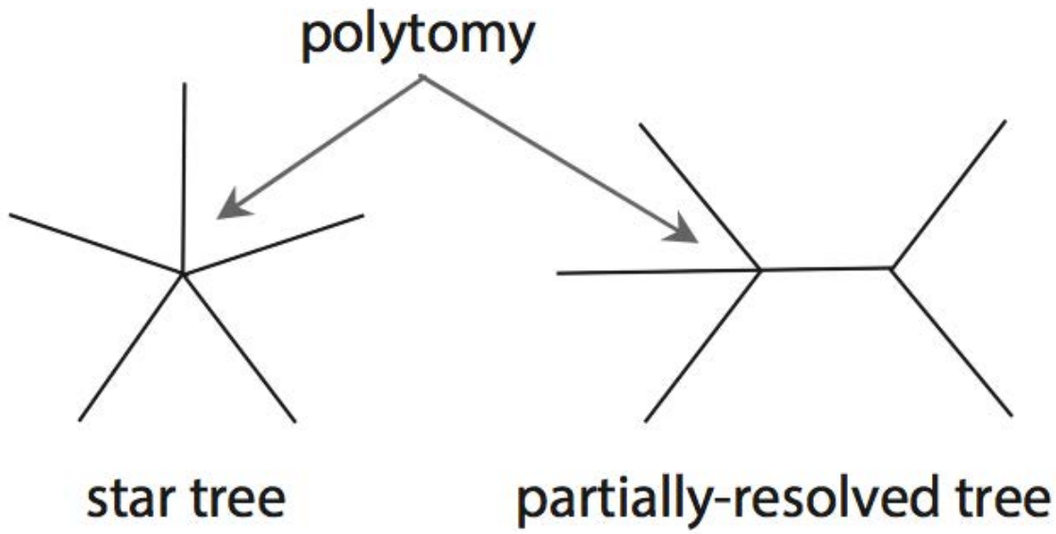
fully-resolved tree

(Yang 2014)

Unresolved Phylogeny

polytomies allow us to represent unresolved nodes



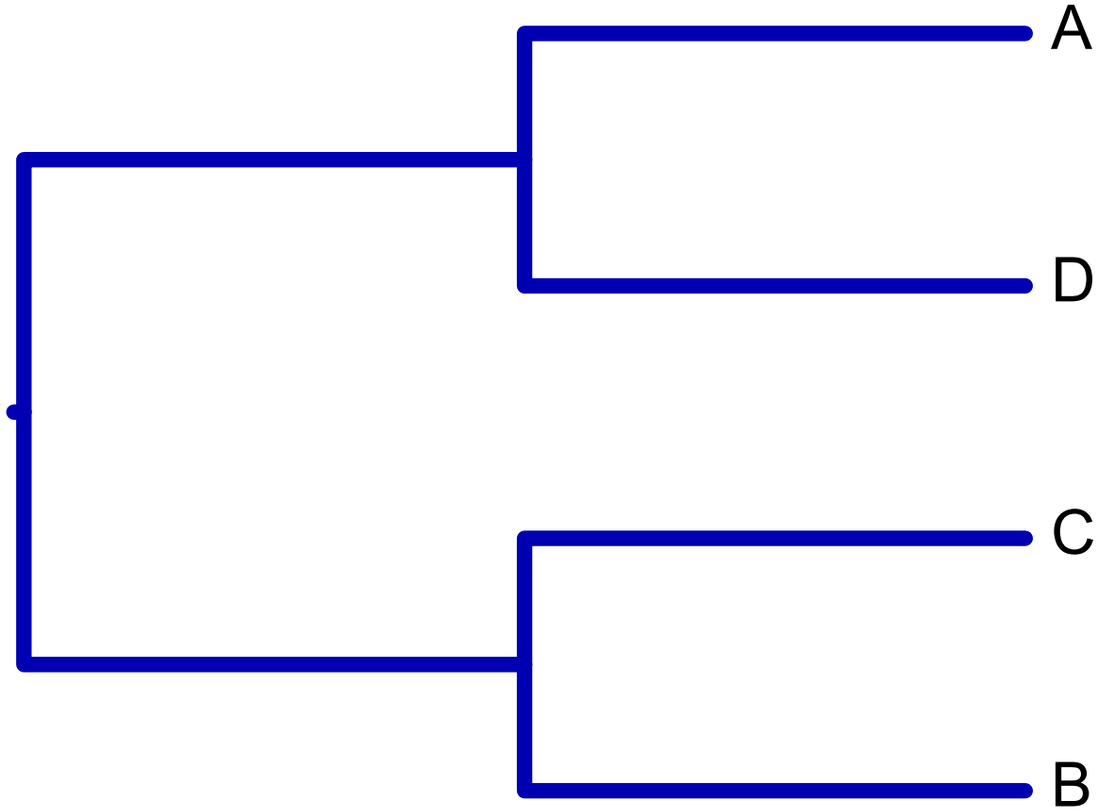


(Yang 2014)

Representing Trees

```
((A,D),(C,B));
```

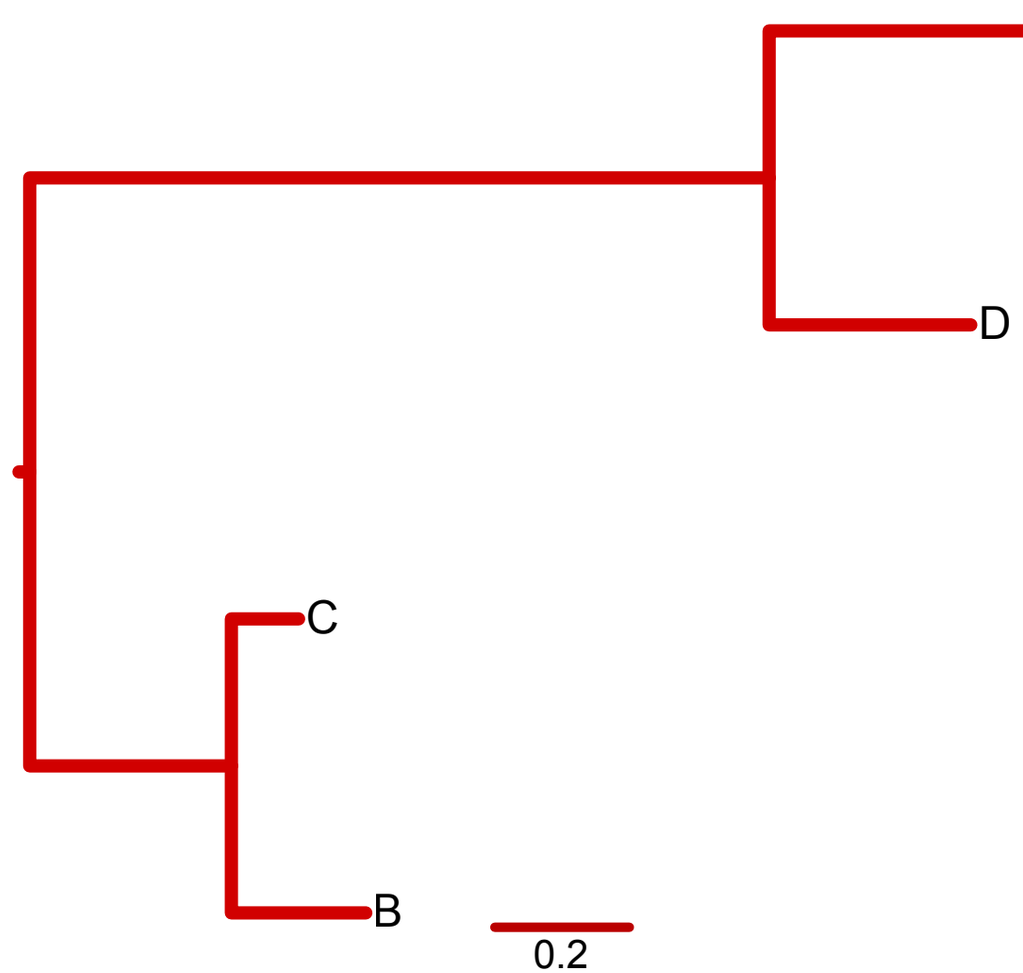
we can represent trees using Newick format, which uses sets of nested parentheses



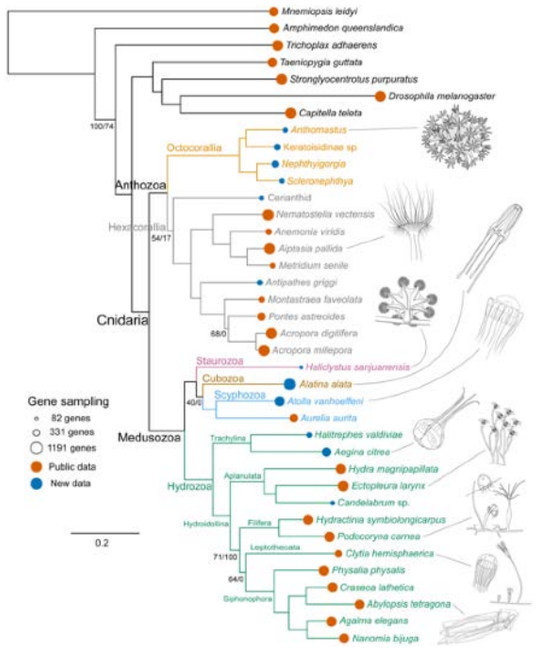
Representing Trees

```
((A:0.5,D:0.3):1.1,(C:0.1,B:0.2):0.3);
```

this format also accommodates branch information

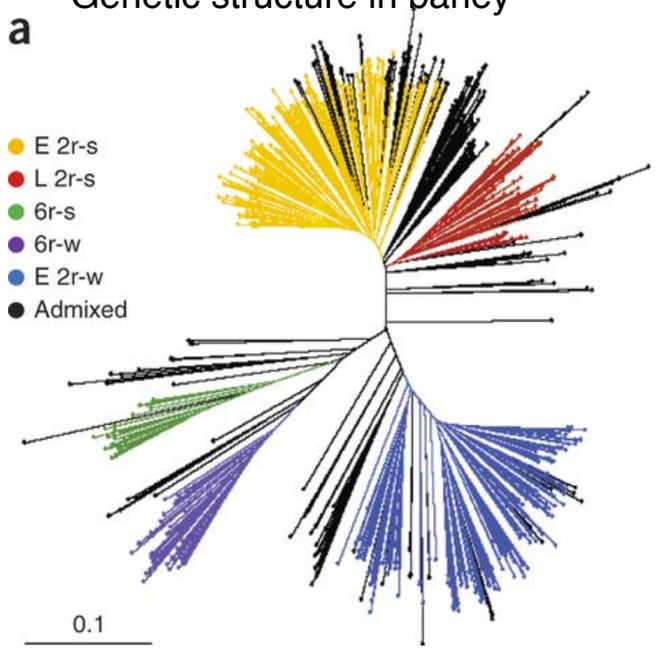


Phylogenies

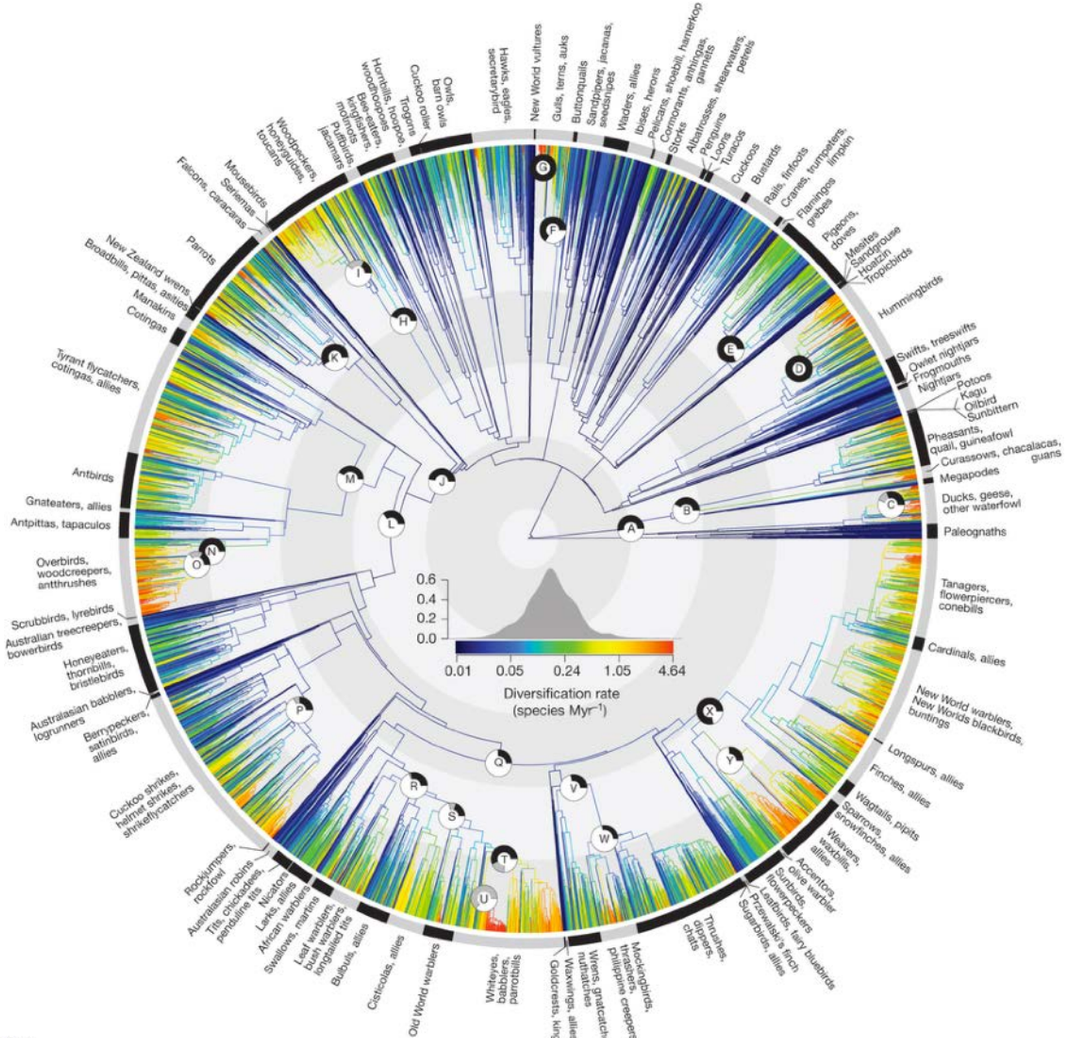


doi.org/10.1371/journal.pone.0139068

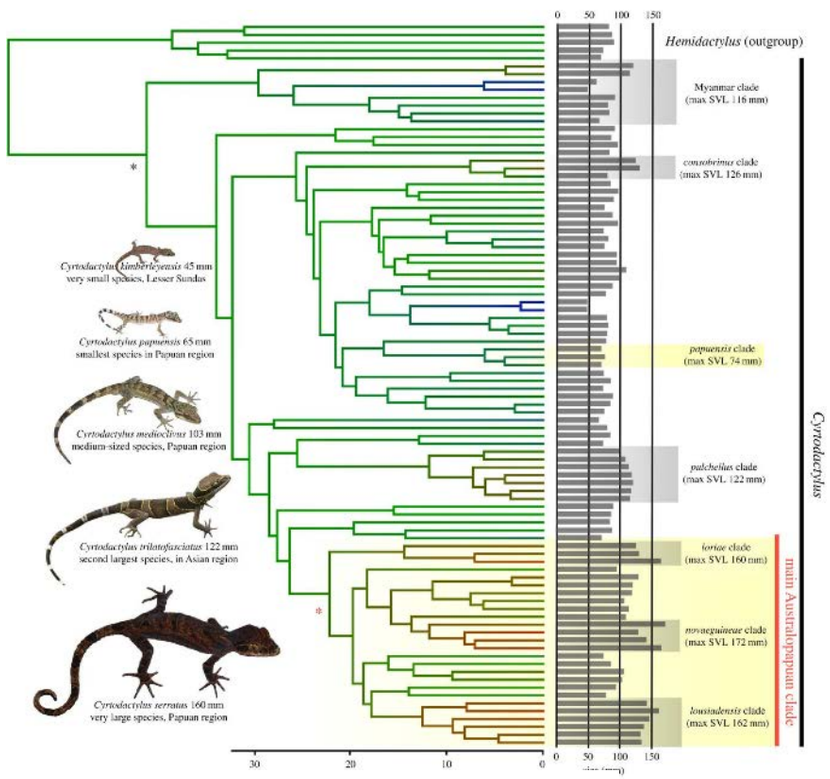
Genetic structure in barley



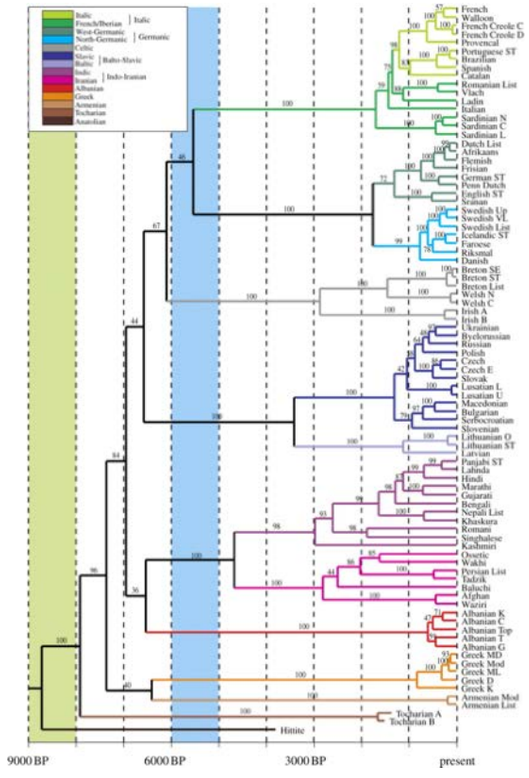
doi.org/10.1038/ng.2447



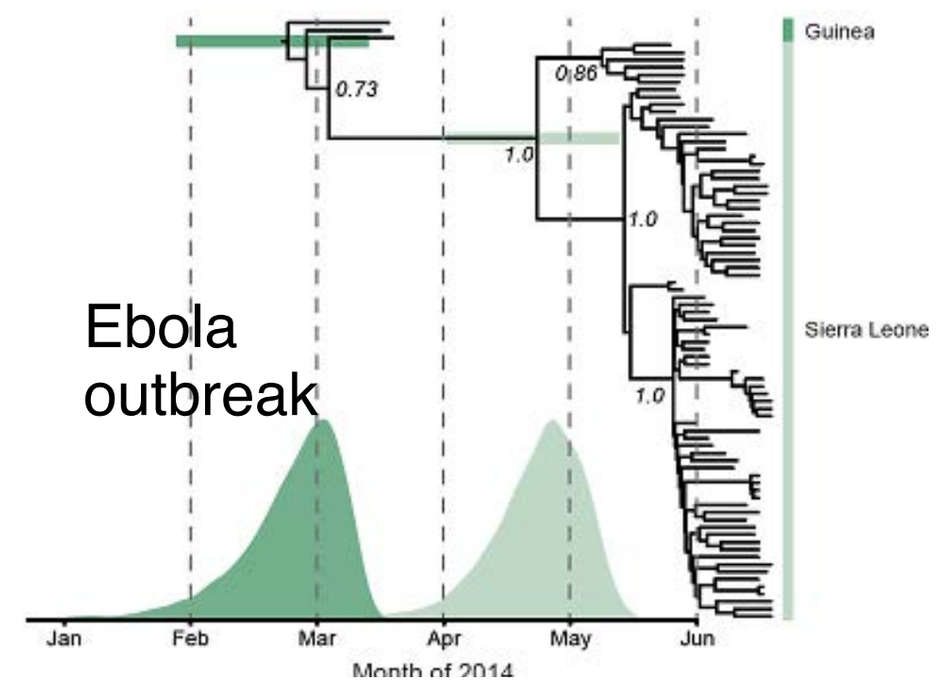
doi.org/10.1038/nature11631



doi.org/10.1098/rsbl.2014.0479

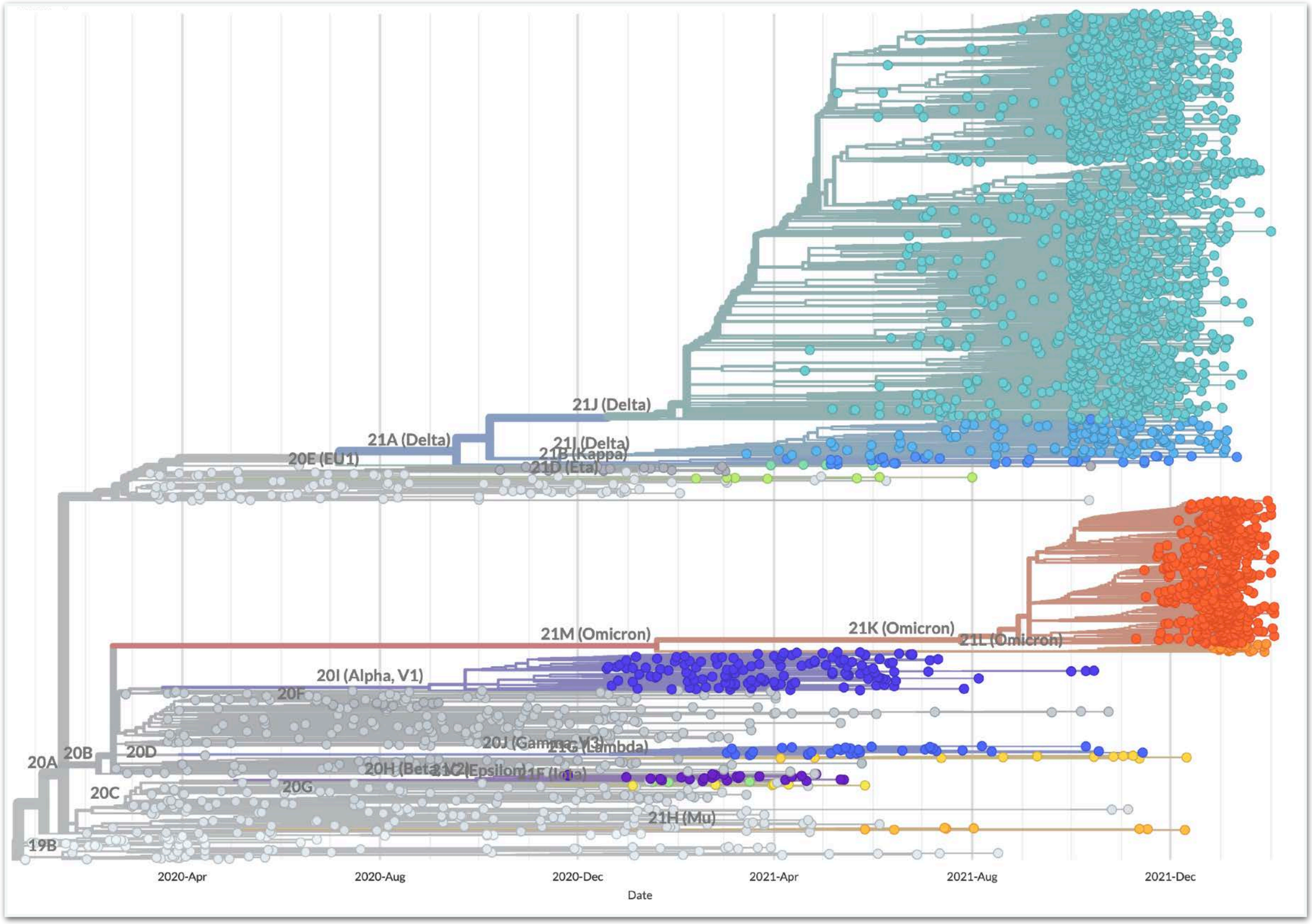


doi.org/10.1098/rstb.2010.0378



doi.org/10.1126/science.1259657

Phylogenetics & Epidemiology



https://nextstrain.org/ncov/gisaid/global

Building Phylogenies

to build a phylogeny, we need to start with data

the data are observations of character states for a set of taxa

taxa	character 1	character 2	character 3
T1	pointed	blue	present
T2	pointed	blue	present
T3	round	blue	absent
T4	round	black	absent

a column in the matrix is a **character**

the form that character takes is its **state**

Building Phylogenies

to build a phylogeny, we need to start with data

taxa	character 1	character 2	character 3
T1	A	T	C
T2	A	T	T
T3	G	T	G
T4	G	T	G

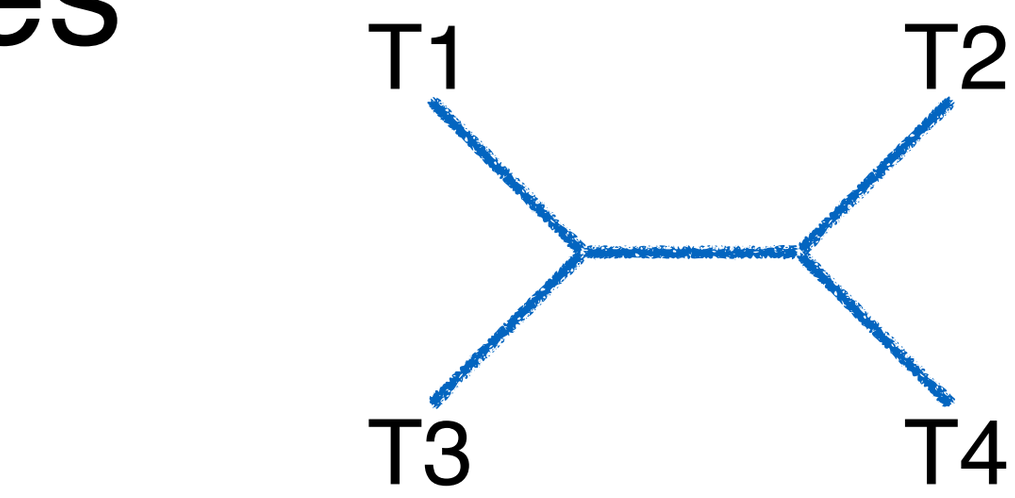
the data are observations of character states for a set of taxa

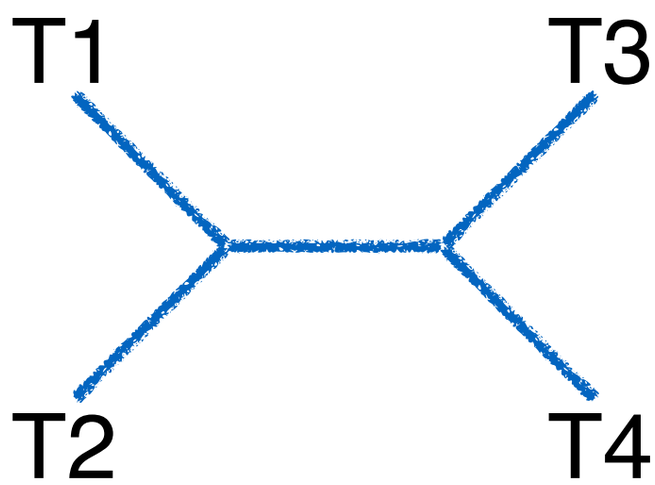
discrete characters can be molecular or morphological

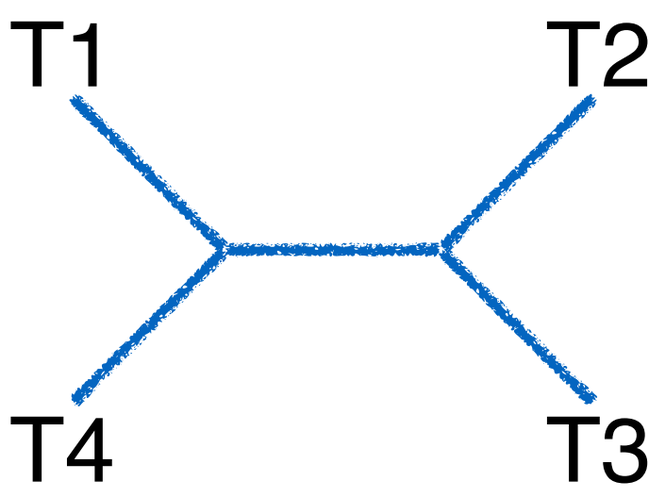
Building Phylogenies

we build phylogenies by evaluating tree topologies

for 4 taxa we can evaluate all possible unrooted topologies (there are only 3)







How many trees?

n	Unrooted trees (U_n)
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
20	~2.22×10^{20}
50	~2.84×10^{74}

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

(Yang 2014)

How many trees?

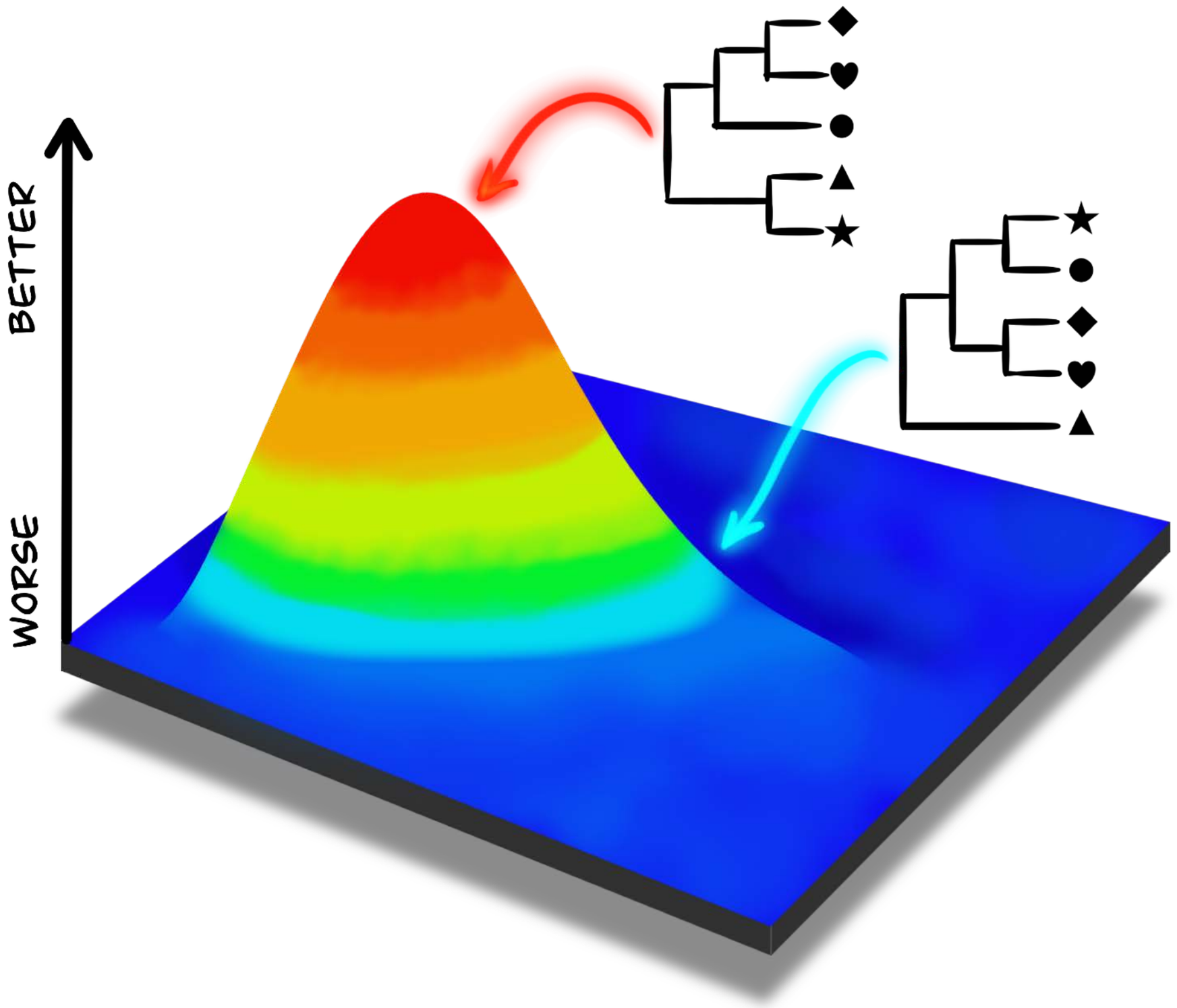
n	Unrooted trees (U_n)	Rooted trees (R_n)
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
20	~2.22×10^{20}	~8.20×10^{21}
50	~2.84×10^{74}	~2.75×10^{76}

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

at 51 taxa, the number of trees exceeds the number of particles in the observable universe

(Yang 2014)

How to find the "best" tree?



It depends on how you measure "best"

Table 3.2 Optimality criteria used for phylogeny reconstruction

Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes, minimized over ancestral states
Maximum likelihood	Log likelihood score, optimized over branch lengths and model parameters
Minimum evolution	Tree length (sum of branch lengths, often estimated by least squares)
Bayesian	Posterior probability, calculated by integrating over branch lengths and substitution parameters

(Yang 2014)

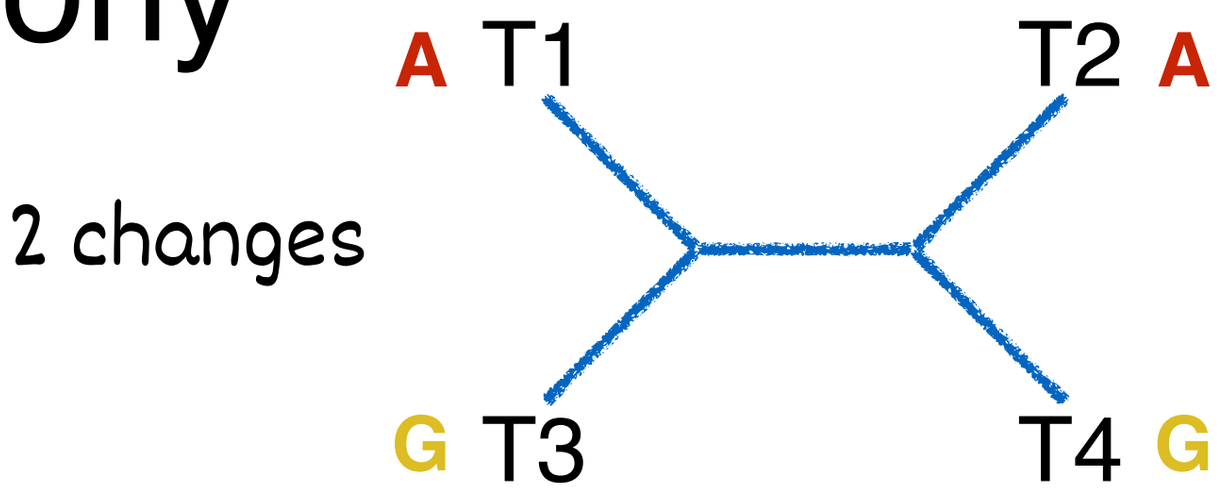
Maximum Parsimony

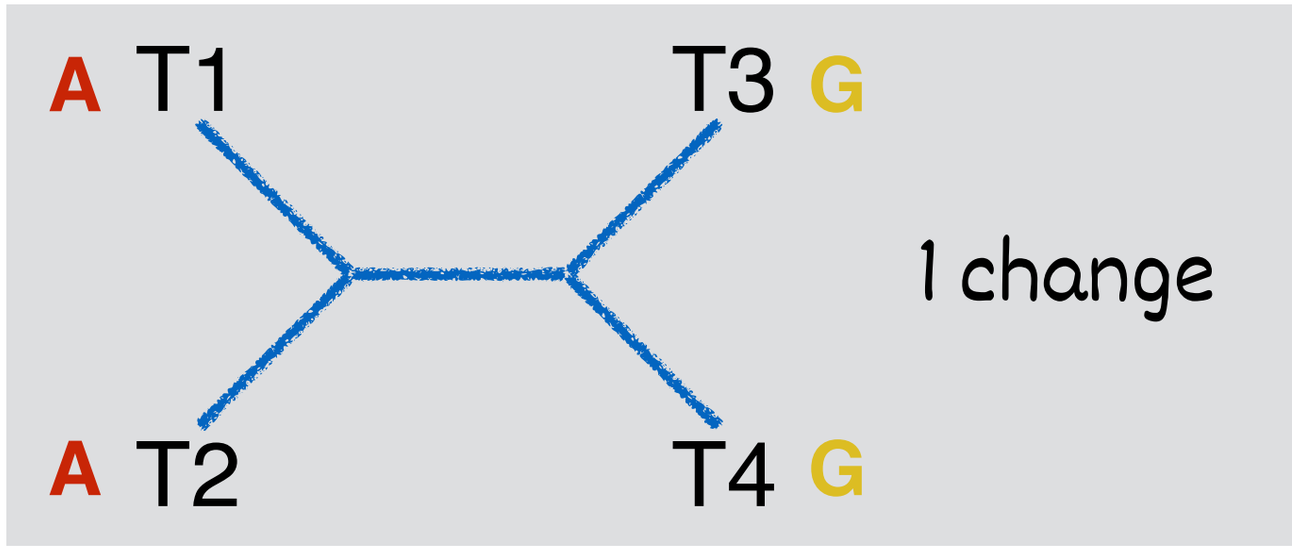
the optimal tree is the one that has the fewest number of changes, given an observed set of discrete characters

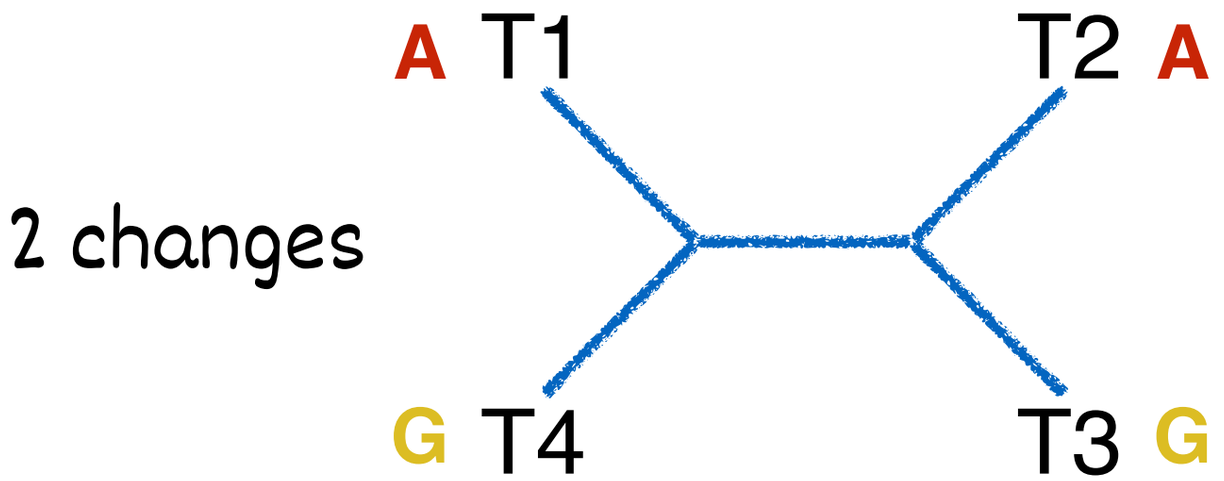
based on the ***parsimony principle***: assumes simpler explanations are better than complex ones

Maximum Parsimony

taxa	character 1
T1	A
T2	A
T3	G
T4	G



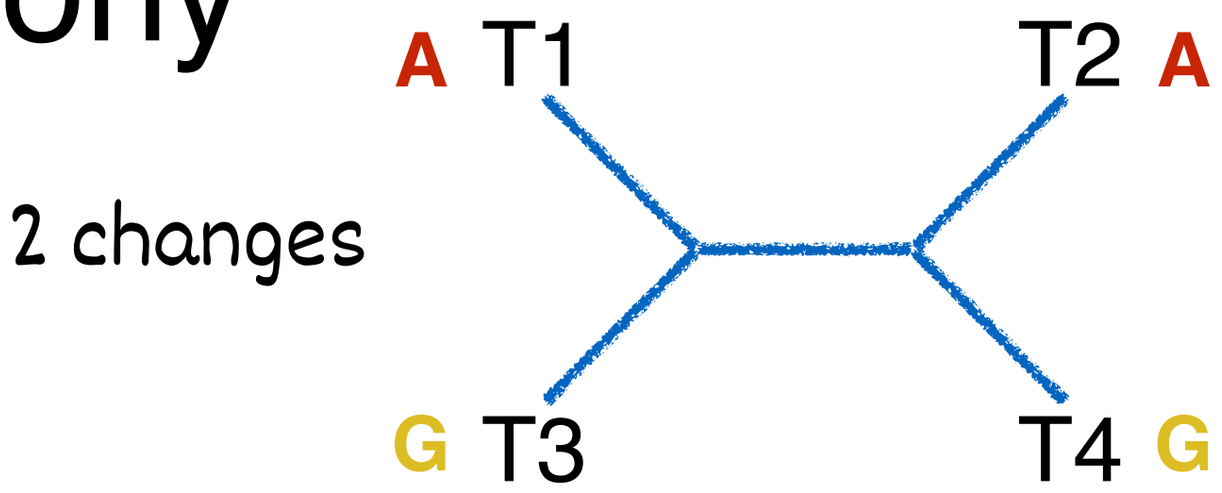


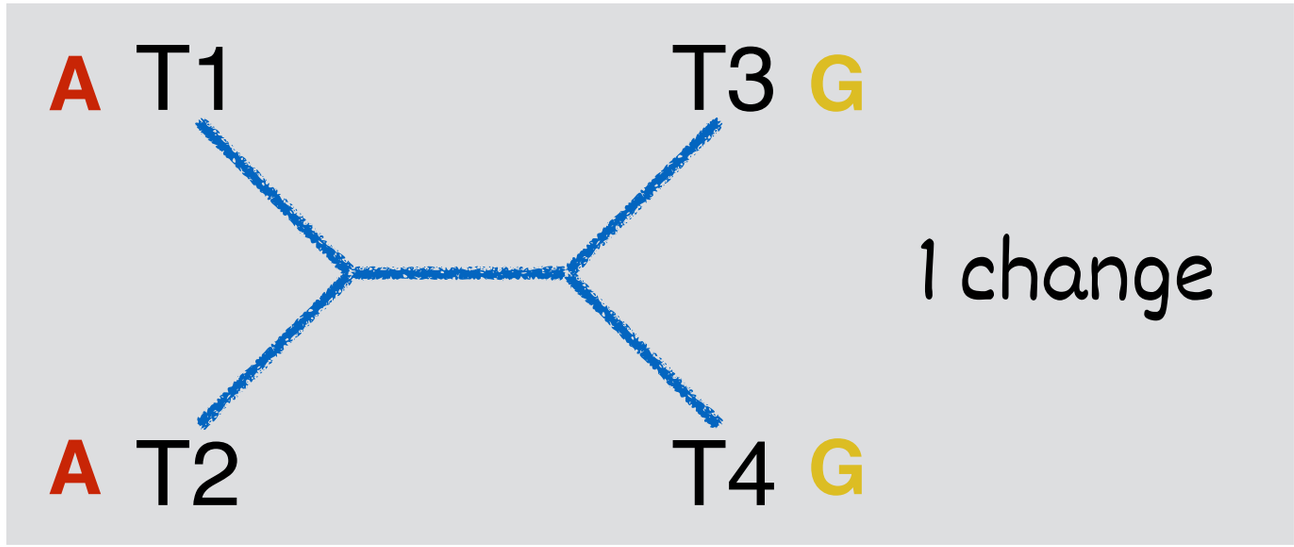


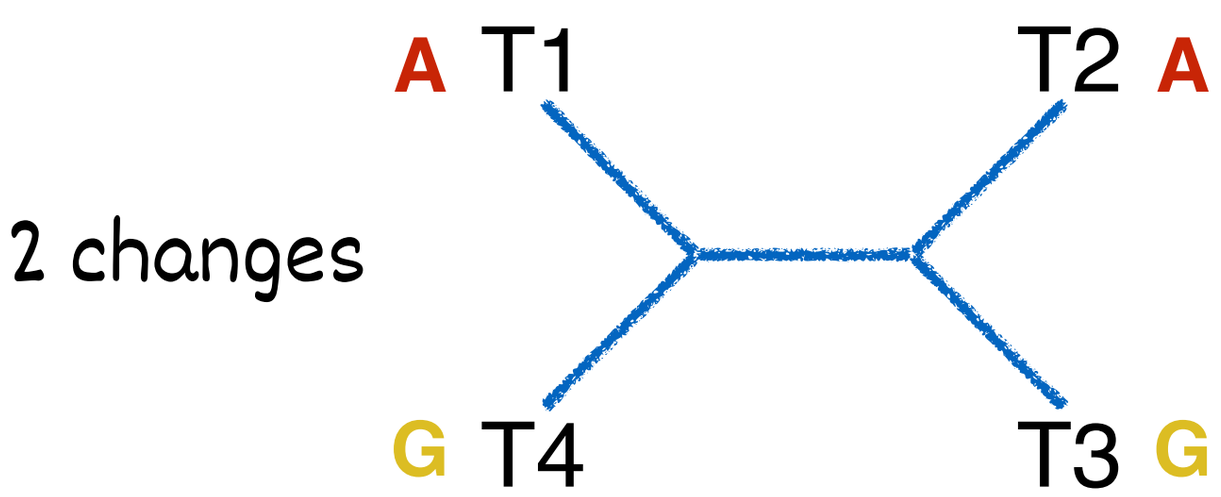
Maximum Parsimony

to find the tree with the fewest changes:

1. construct all possible trees
2. count the minimum number of changes for every character in the matrix
3. sum the counts across all characters to obtain the "total tree length"
4. choose the tree with the lowest total tree length

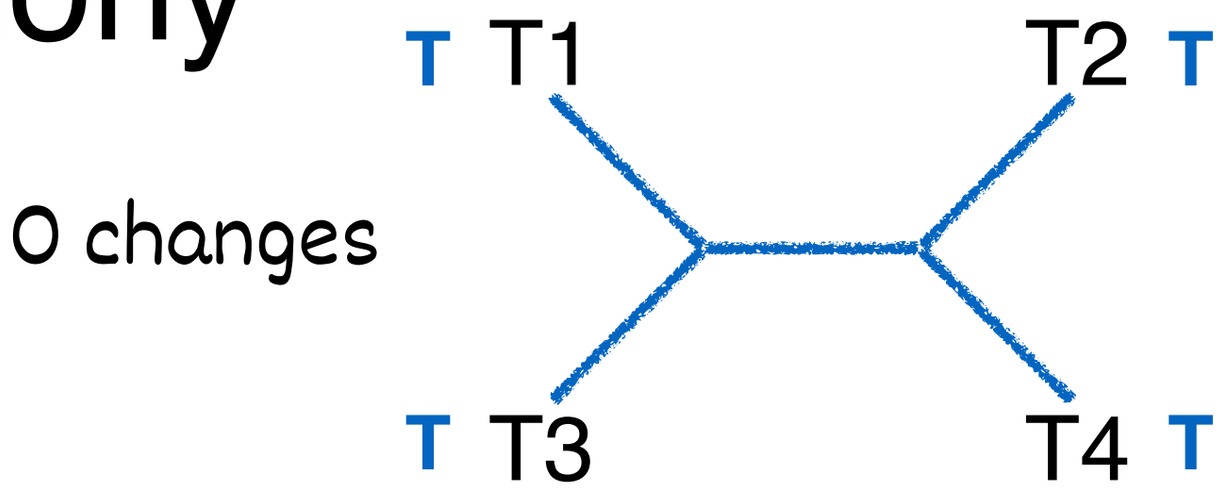


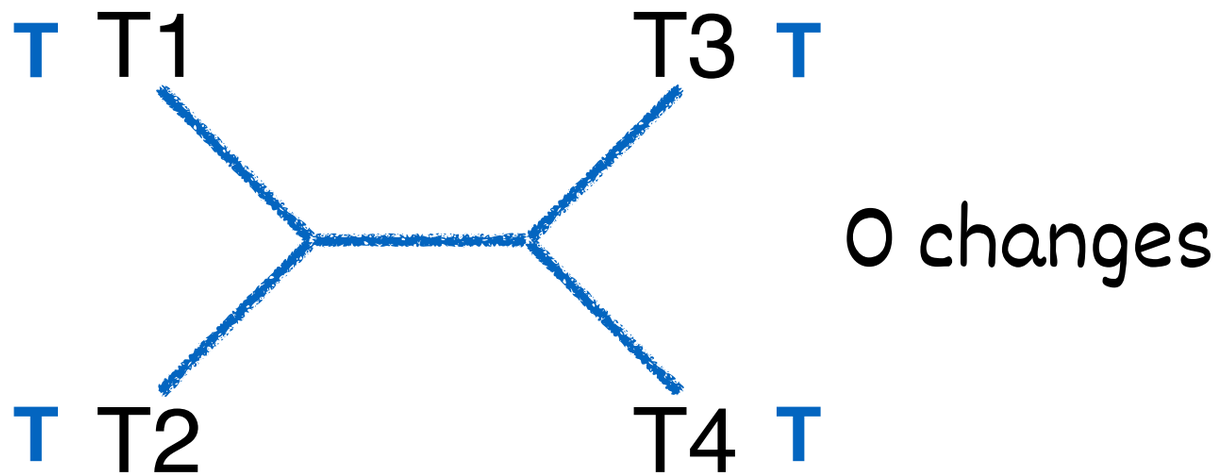




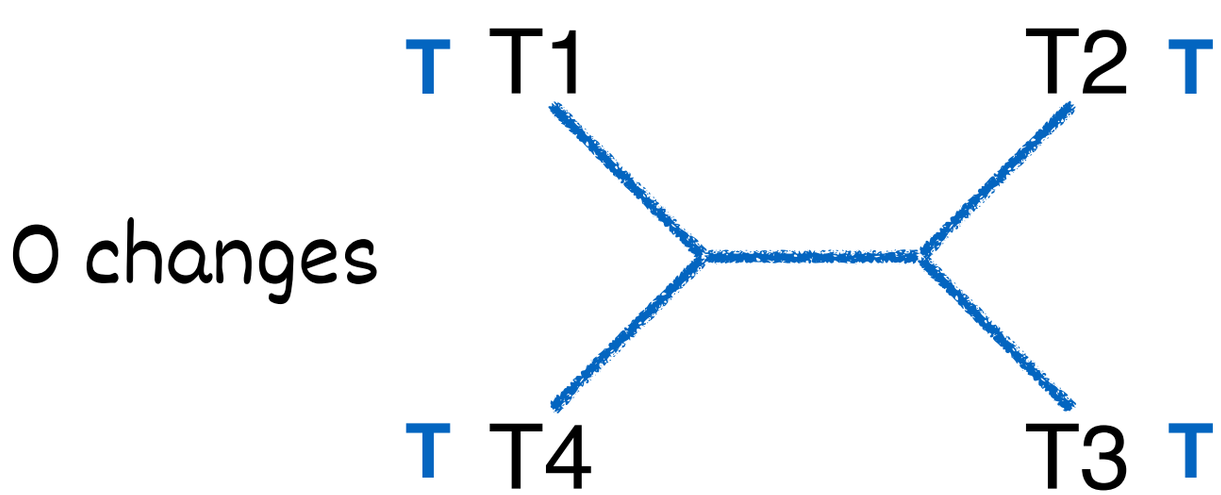
Maximum Parsimony

taxa	character 2
T1	T
T2	T
T3	T
T4	T





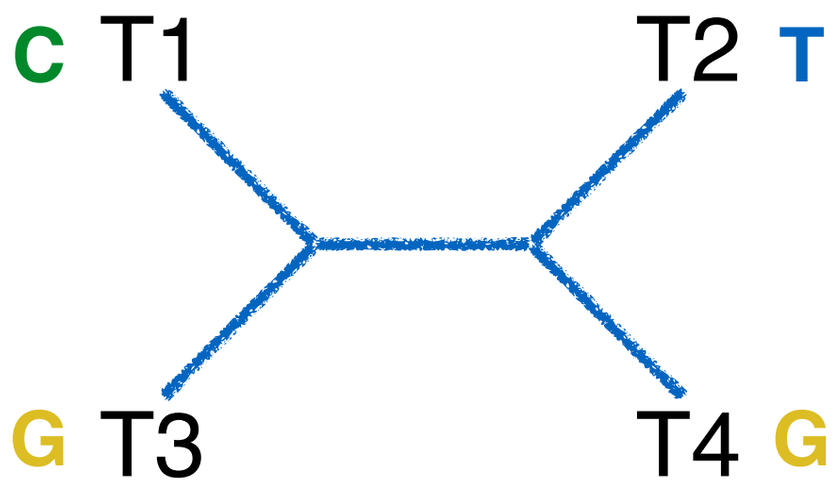
not all patterns of characters are parsimony informative because all topologies have the same length

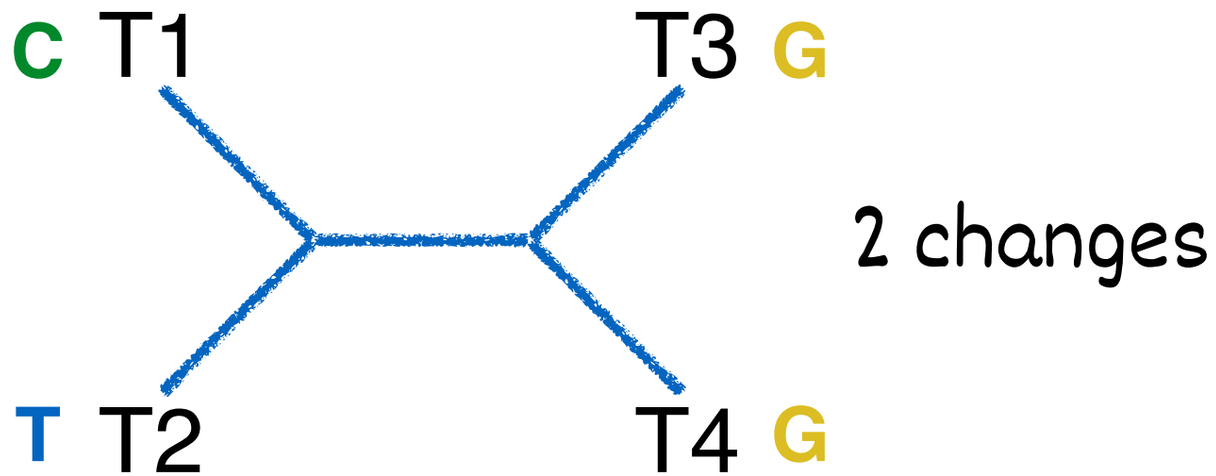


Maximum Parsimony

taxa	character 3
T1	C
T2	T
T3	G
T4	G

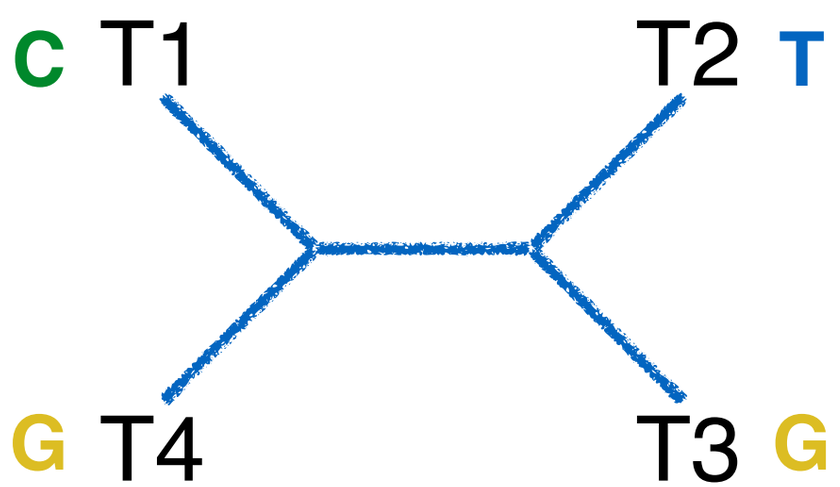
2 changes





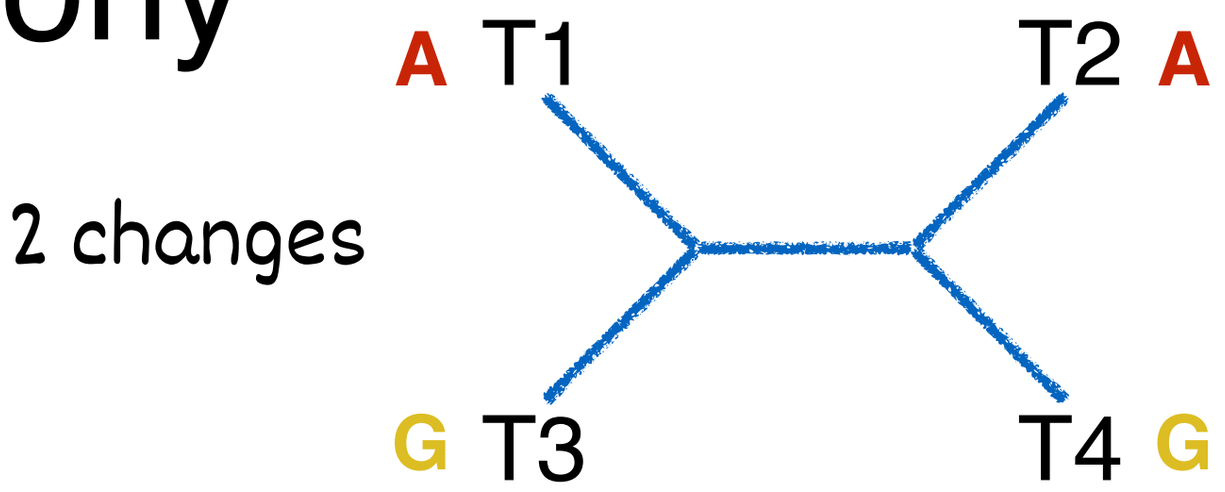
not all patterns of characters are parsimony informative because all topologies have the same length

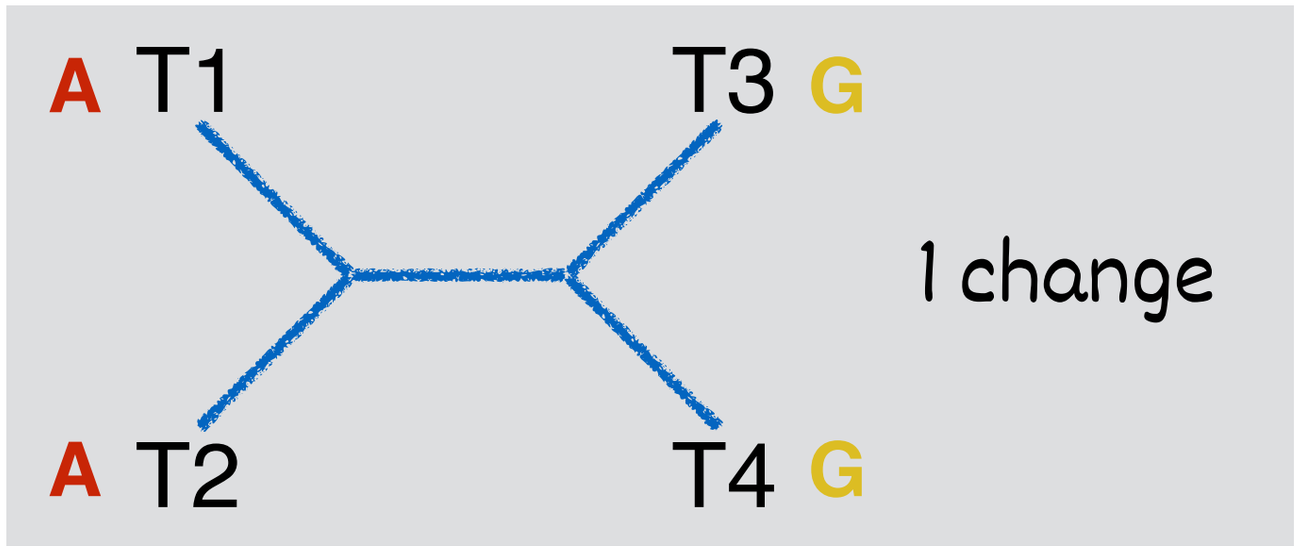
2 changes



Maximum Parsimony

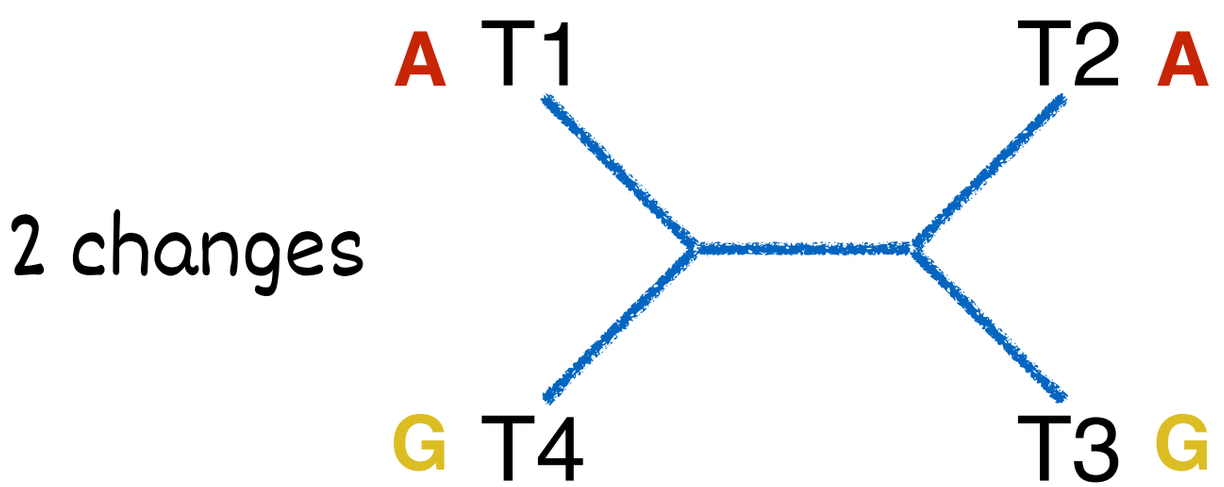
taxa	character 1
T1	A
T2	A
T3	G
T4	G





for a 4-taxon tree, the only parsimony informative patterns are:

xxyy, xyxy, xyyx

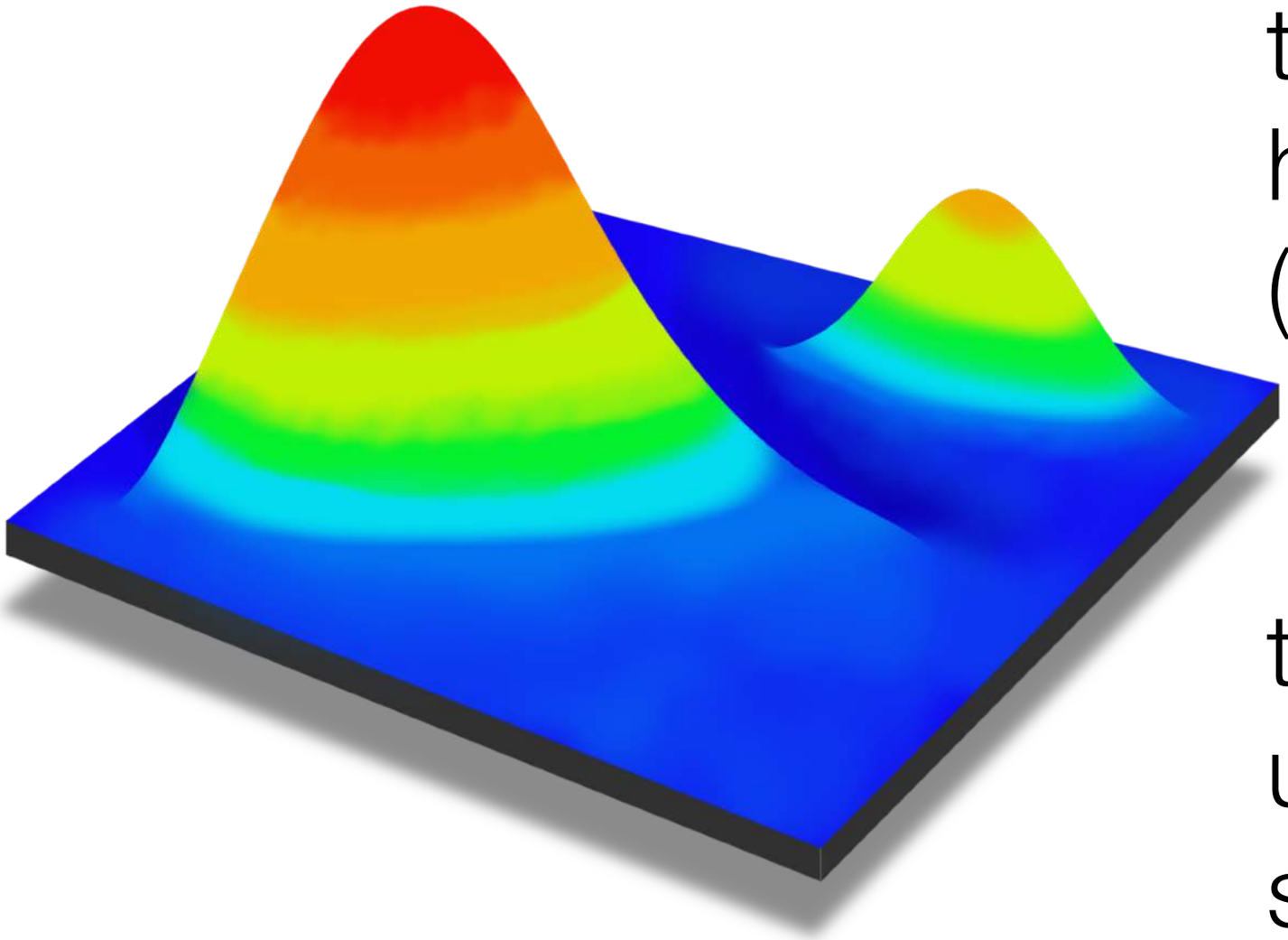


Maximum Parsimony

computing the total tree length (parsimony score) is easy to do given a tree topology and observed character states (see Yang, 2014, ch. 3)

finding the tree topology that has the optimal parsimony score is hard (it is actually NP-hard)

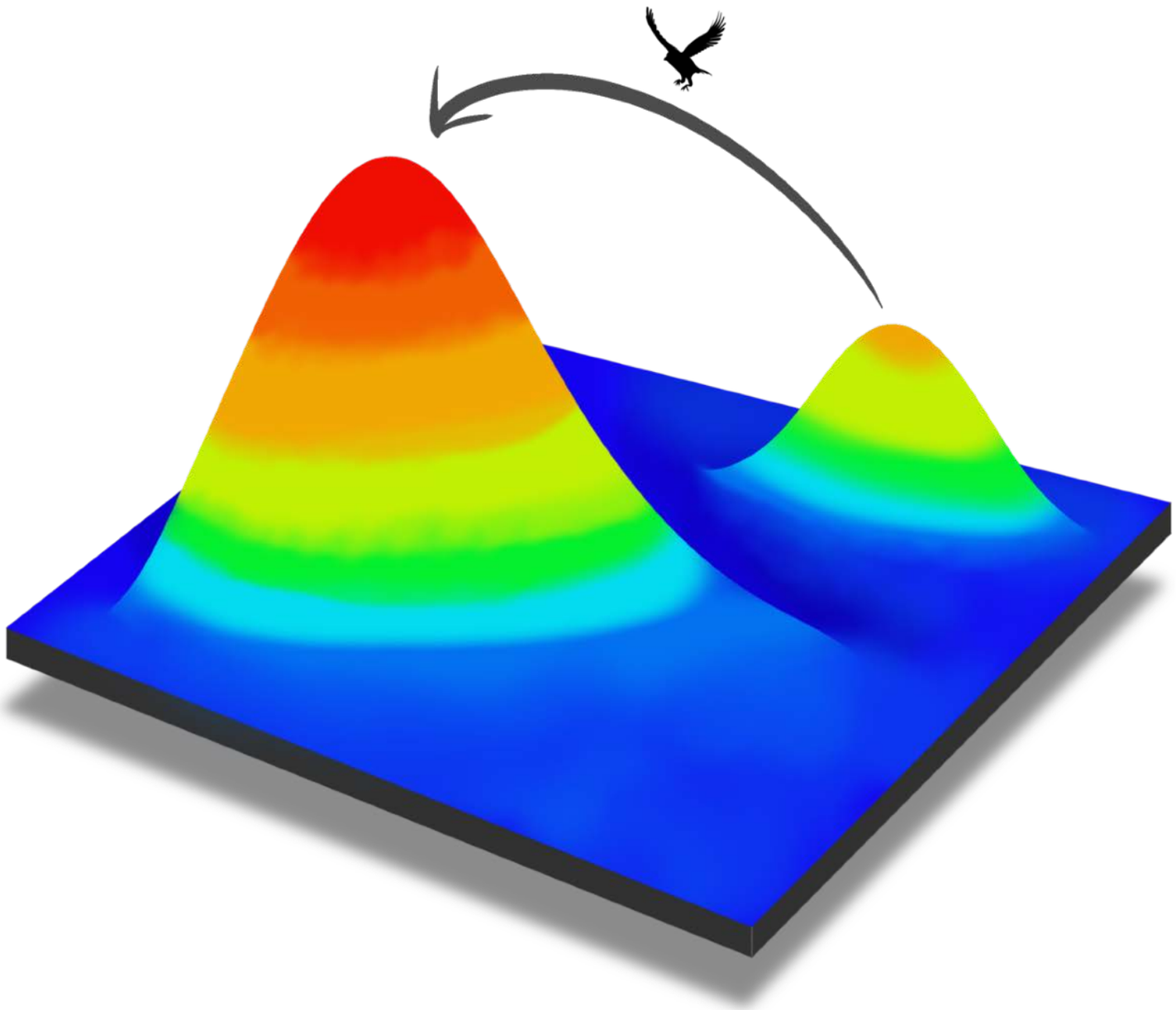
Searching Tree Topologies



a number of heuristic tree-search algorithms have been developed (Yang 2014, Ch. 3)

these methods allow us to evaluate just a subset of the possible trees

Searching Tree Topologies



importantly, we need tree-search methods that can find the global optimum

these approaches are useful for maximum parsimony, maximum likelihood, and Bayesian methods

Parsimony and Assumptions

the parsimony principle is based on Occam's Razor: the simplest explanation that fits the data is preferred and *ad hoc* explanations should be avoided

parsimony does not make *explicit* assumptions about the evolutionary process that generated observed character states

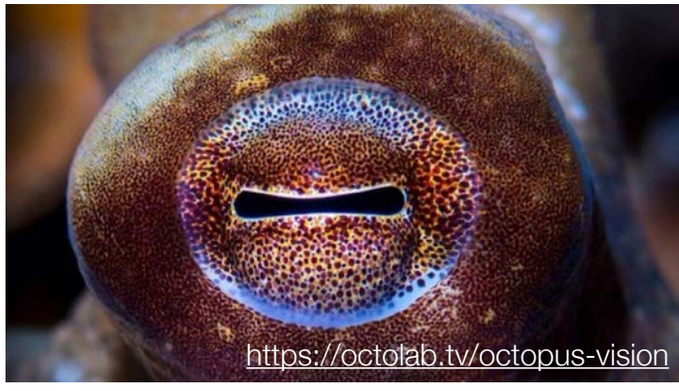
Parsimony and Assumptions

with datasets comprising multiple characters there is typically no single topology that is the most parsimonious for every observation

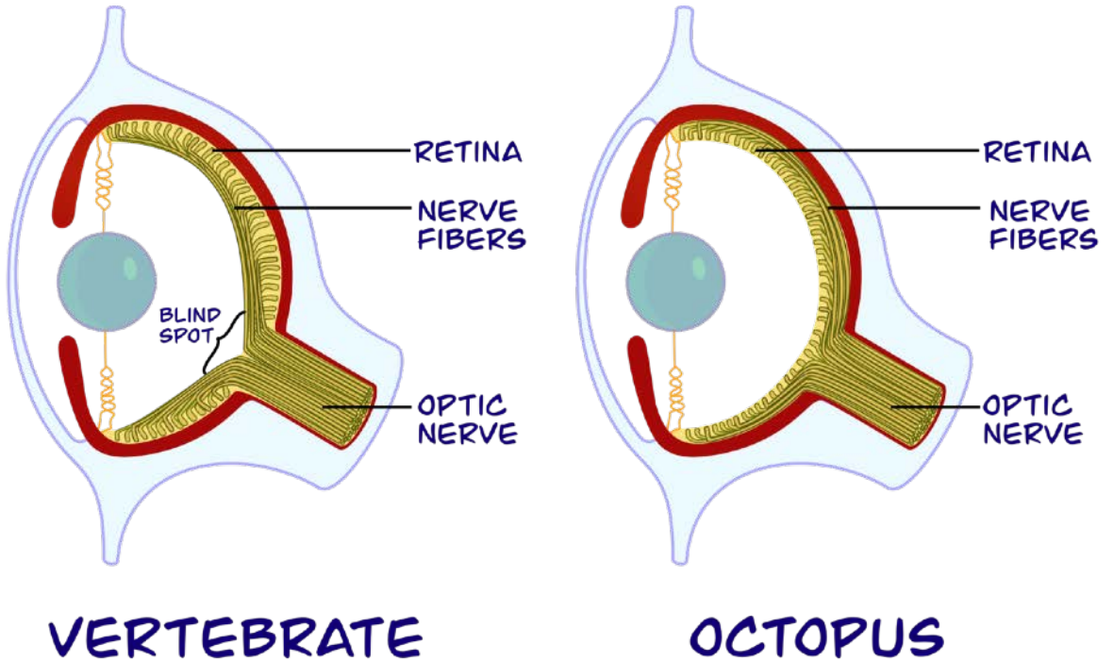
thus, *ad hoc* explanations (convergence, reversals) must be invoked to explain these patterns

Homoplasy

A trait that is found in two species, but not in their common ancestor is an example of **homoplasy**

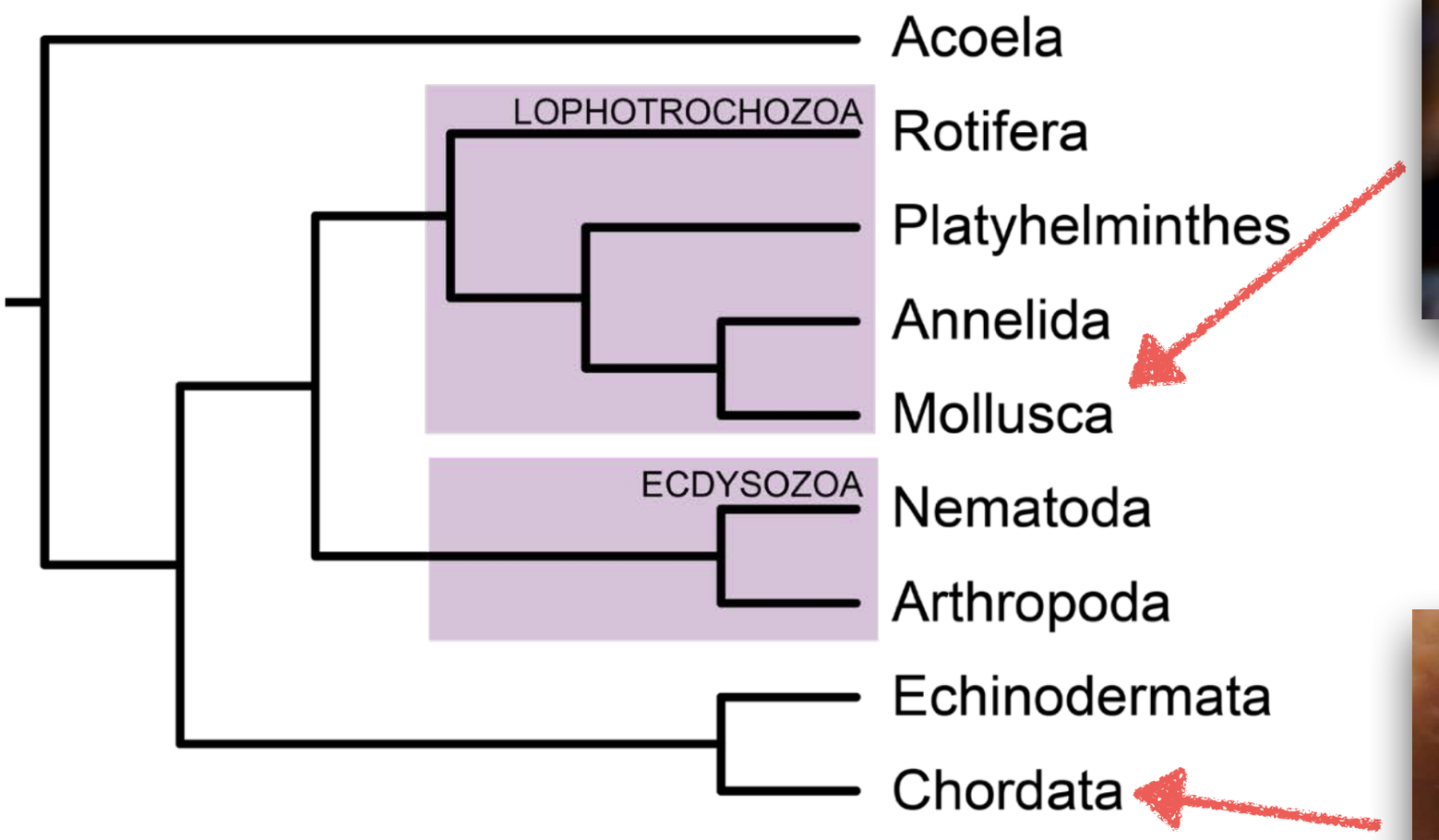


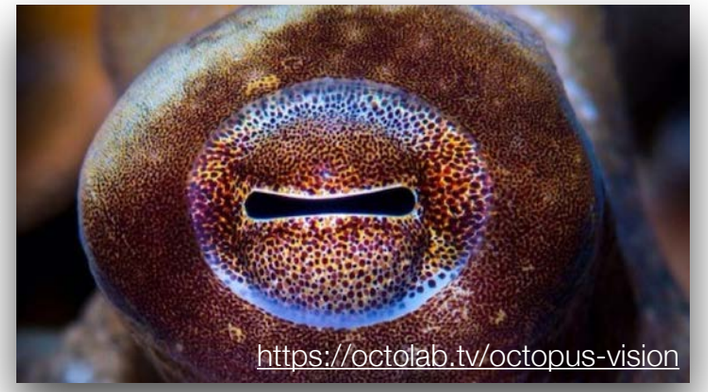
The eye structures of a human & a giant Pacific octopus are similar but evolved independently

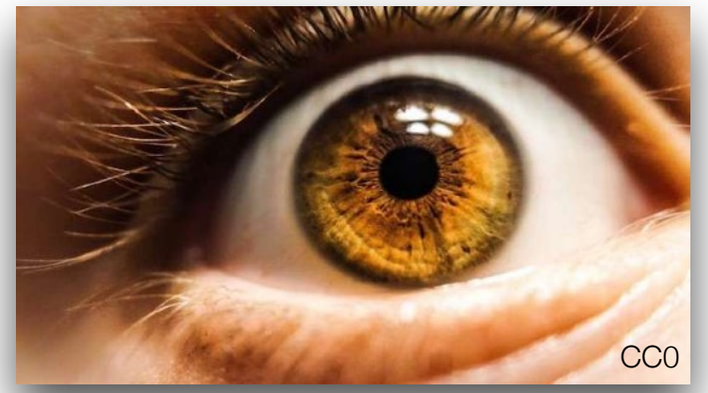


(figure credit: Jerry Crimson Mann)

Homoplasy and Parsimony







when characters conflict, ad hoc explanations (e.g., convergence) cannot be avoided

Arguments Against Parsimony

parsimony does make *implicit* assumptions about evolutionary processes, though is is difficult to identify exactly what these are

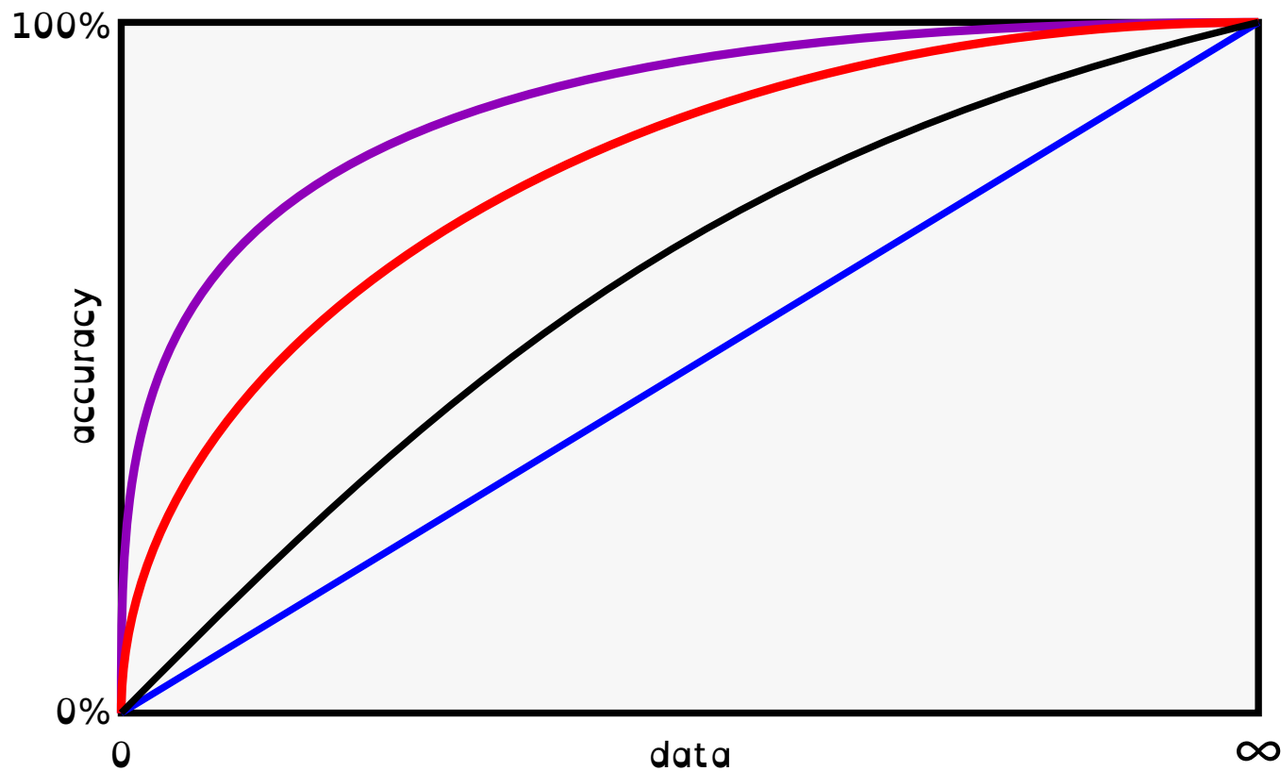
parsimony has been demonstrated to be statistically inconsistent

for more on this, see Yang (2014), Ch. 5

Statistical Consistency

an estimator is consistent if it is guaranteed to get the correct answer with an infinite amount of data

we would prefer our estimators to be consistent



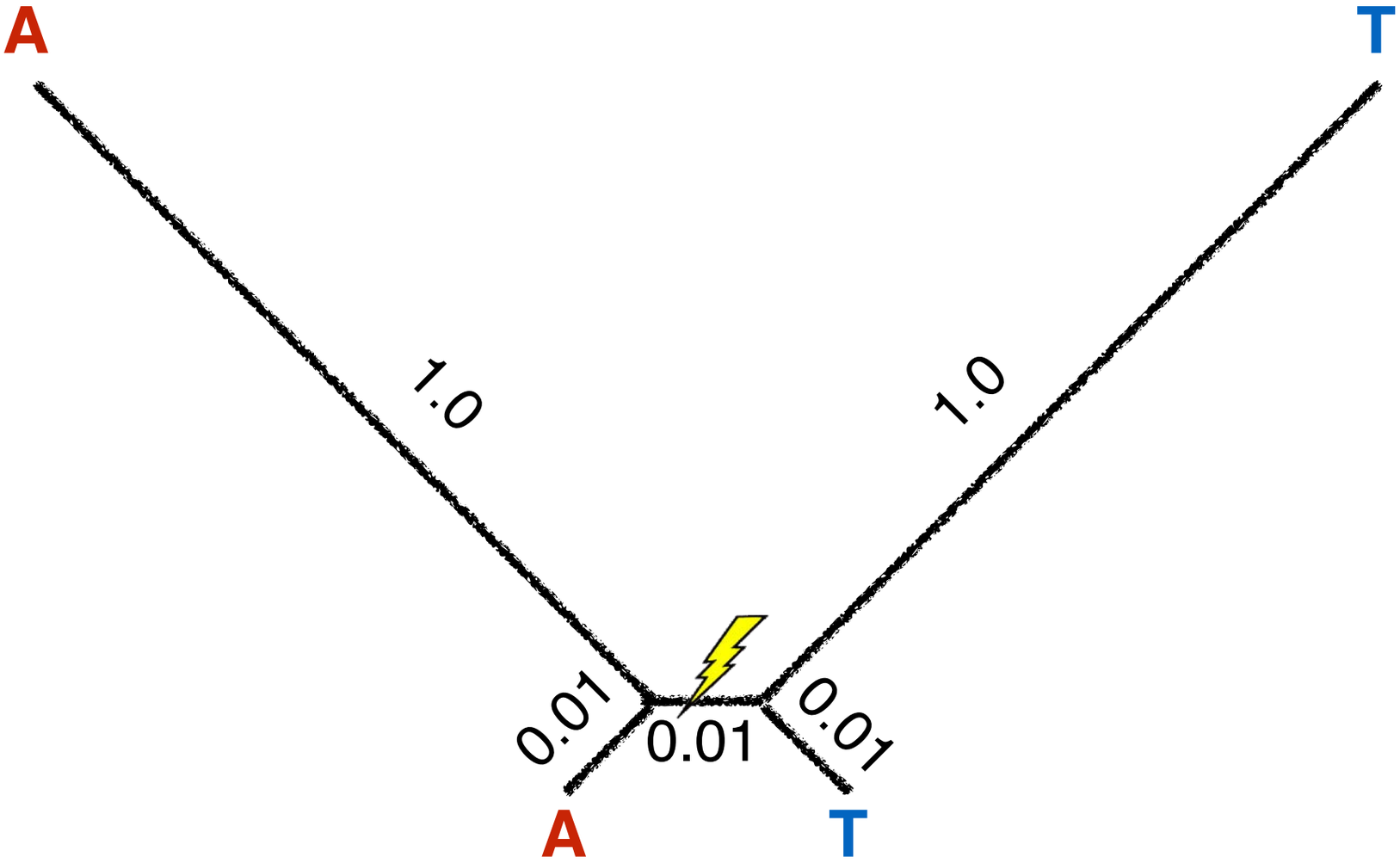
Parsimony can be Inconsistent

Felsenstein (1978) demonstrated that for some situations, parsimony is inconsistent and yields the wrong tree, even with an infinite amount of data

this issue is also known as **long-branch attraction** and is one of the strongest criticisms of the parsimony method

Long-Branch Attraction

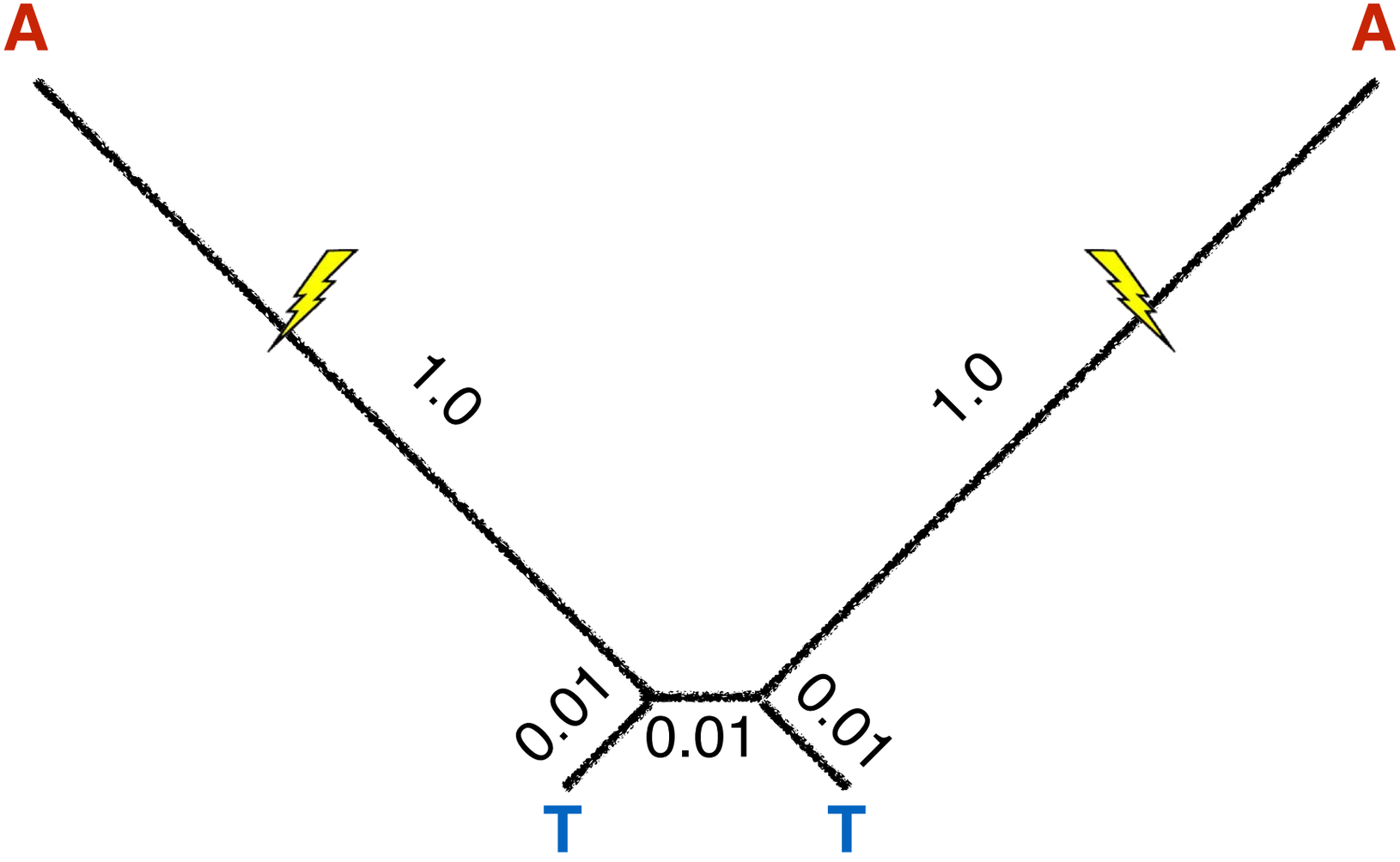
the probability of a parsimony informative site due to inheritance is very low



(based on slides by M.T. Holder)

Long-Branch Attraction

the probability of a misleading parsimony informative site due to parallelism is much higher

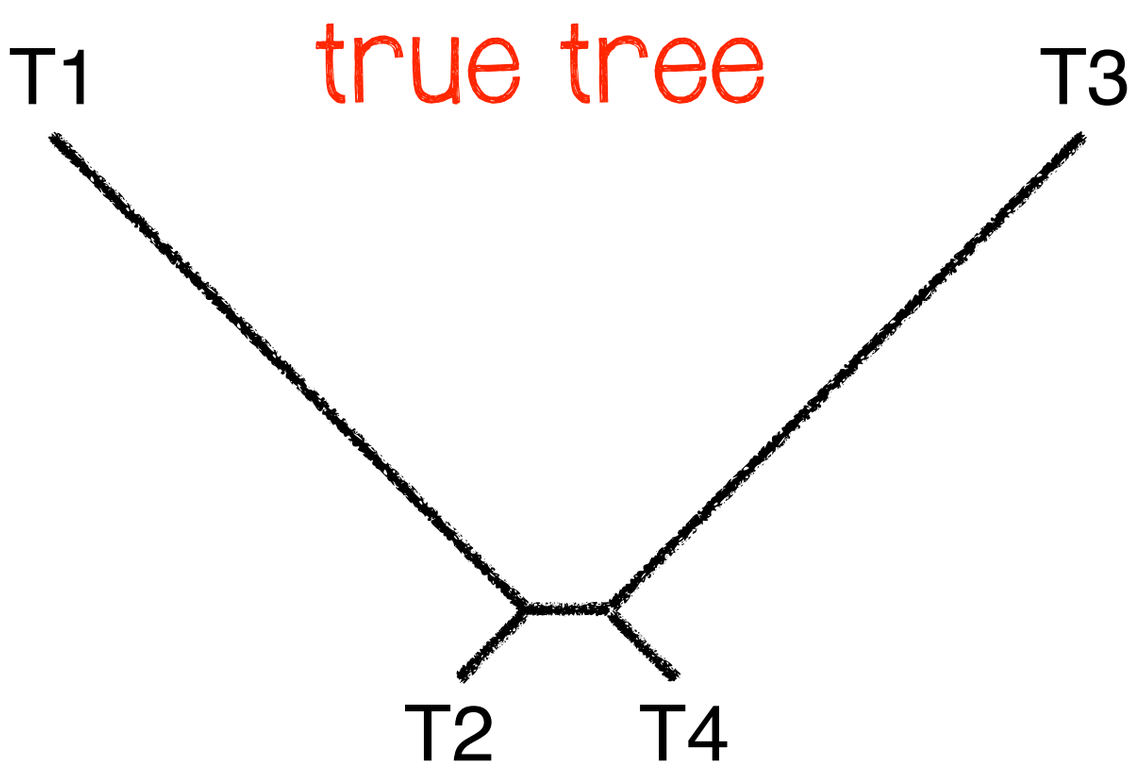


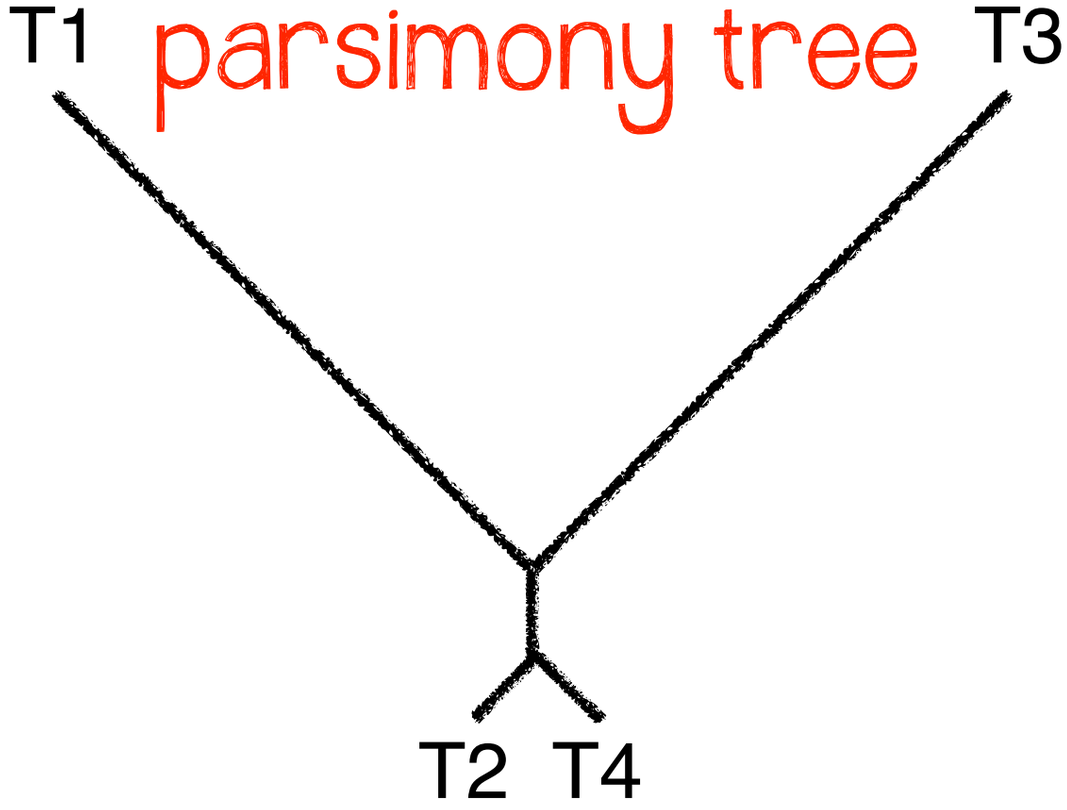
(based on slides by M.T. Holder)

Long-Branch Attraction

parsimony is almost guaranteed to get this tree wrong

parsimony will incorrectly place two long branches as sister lineages



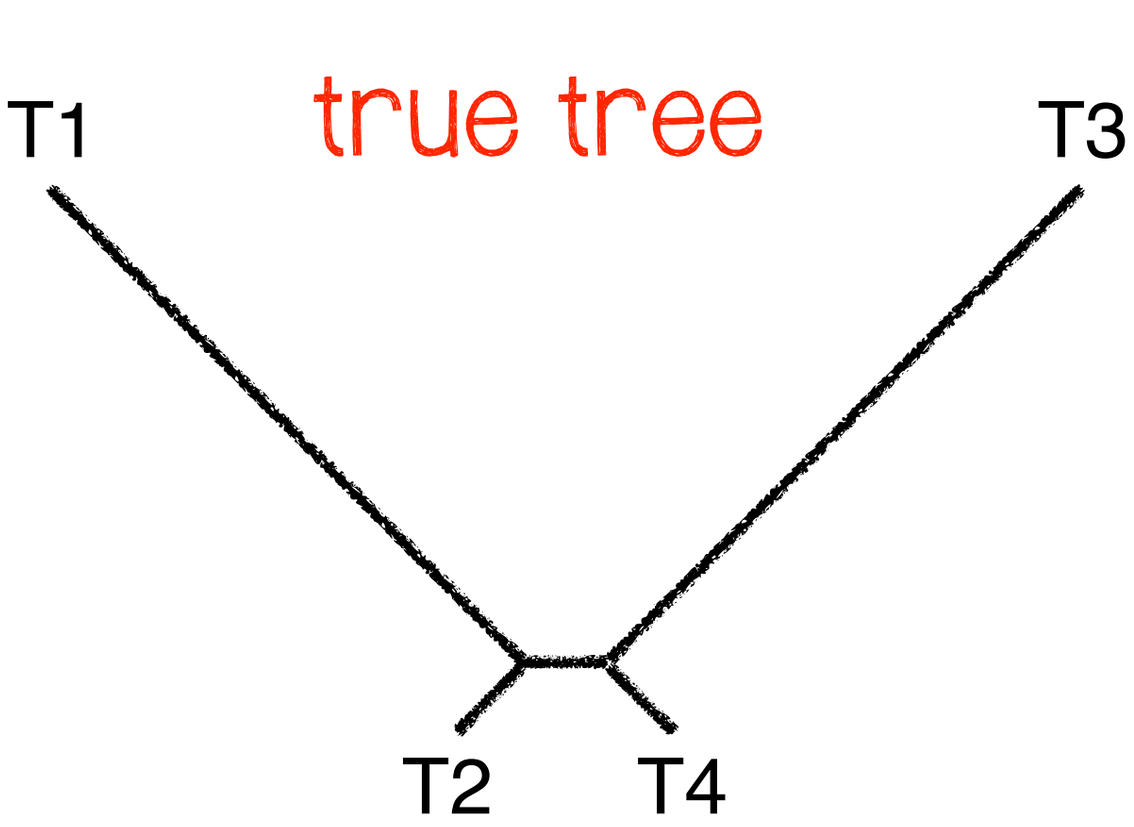


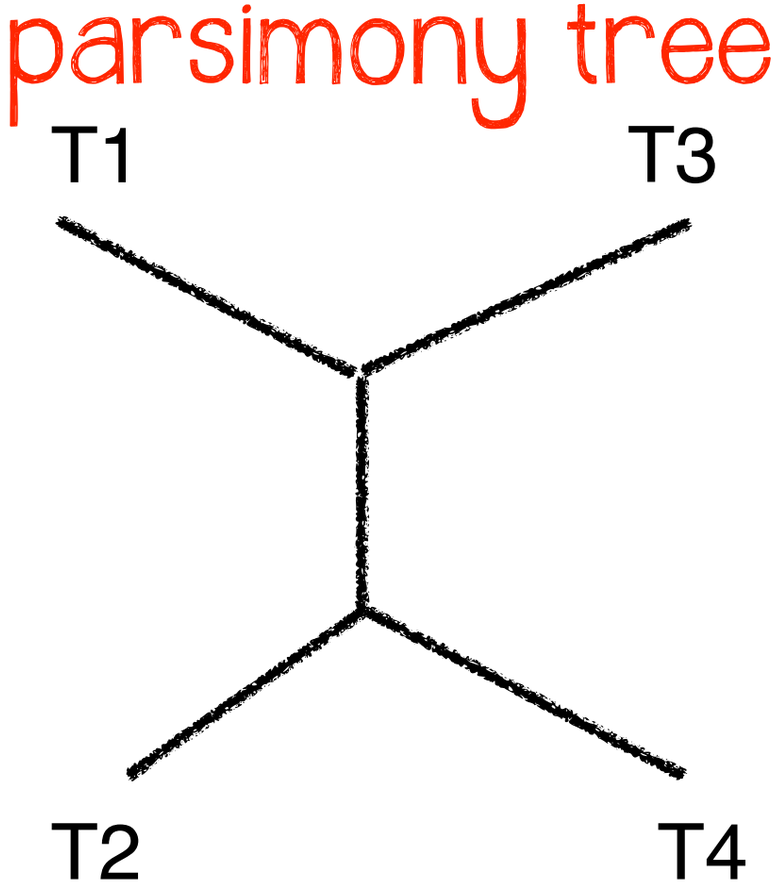
(based on slides by M.T. Holder)

Long-Branch Attraction

under parsimony more change will be attributed to the internal branch

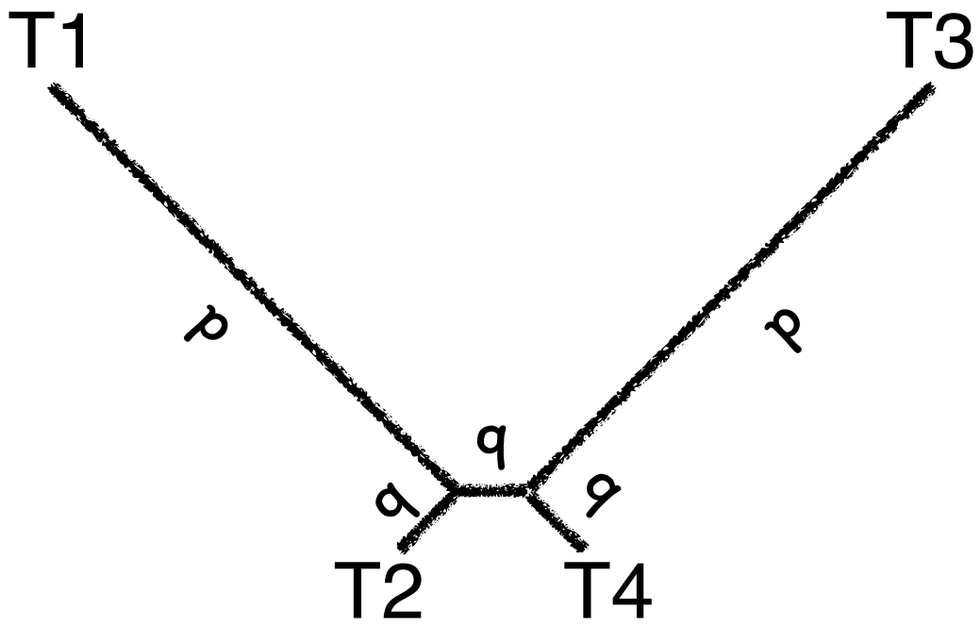
in the case of long branch attraction, parsimony is positively misleading

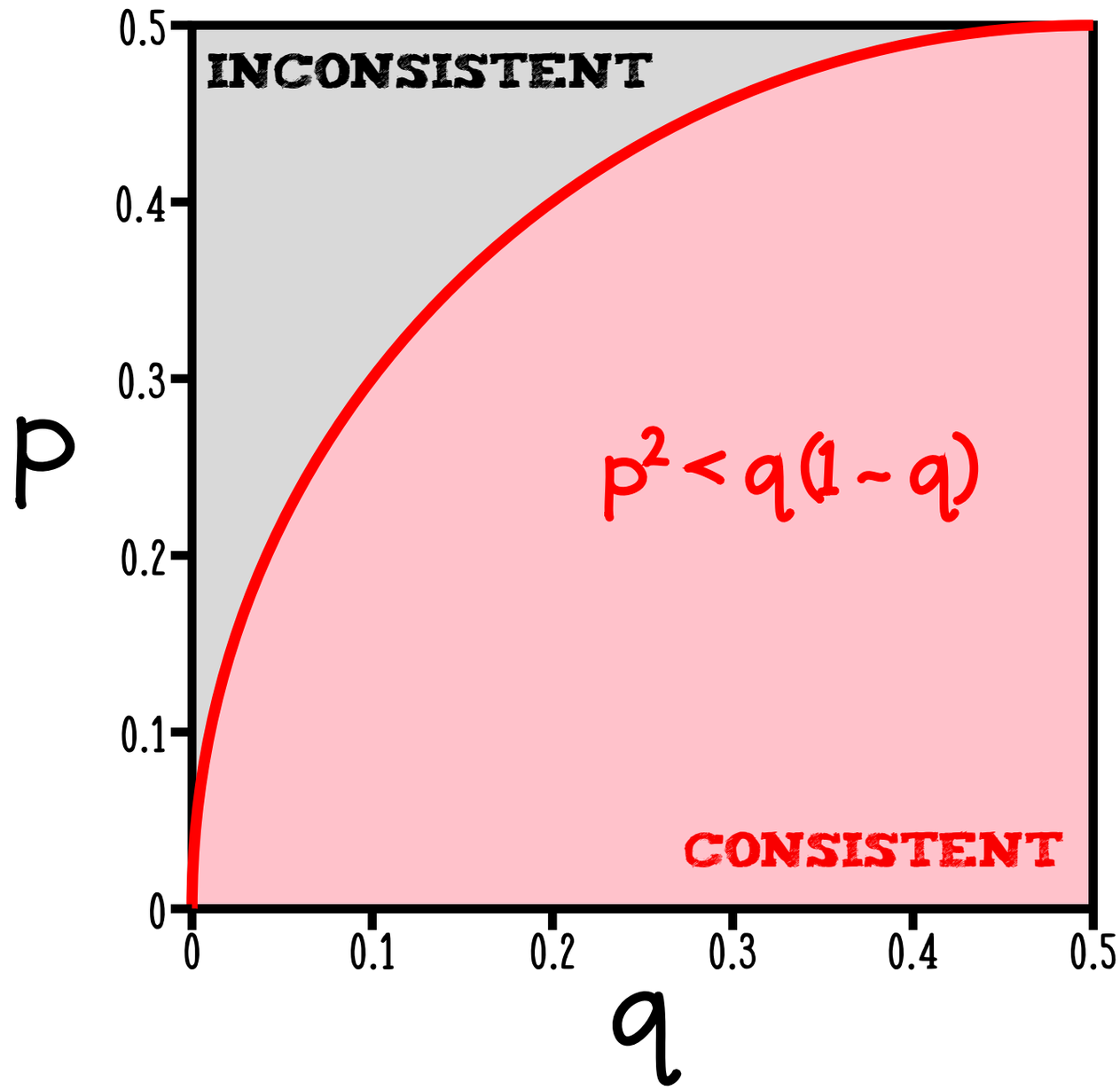




(based on slides by M.T. Holder)

Parsimony can be Inconsistent

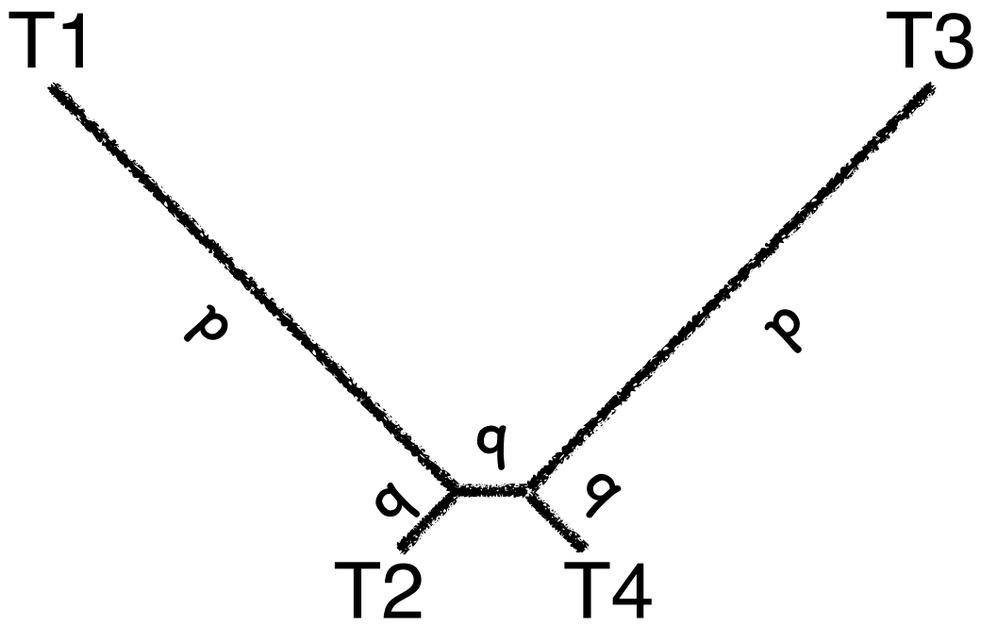




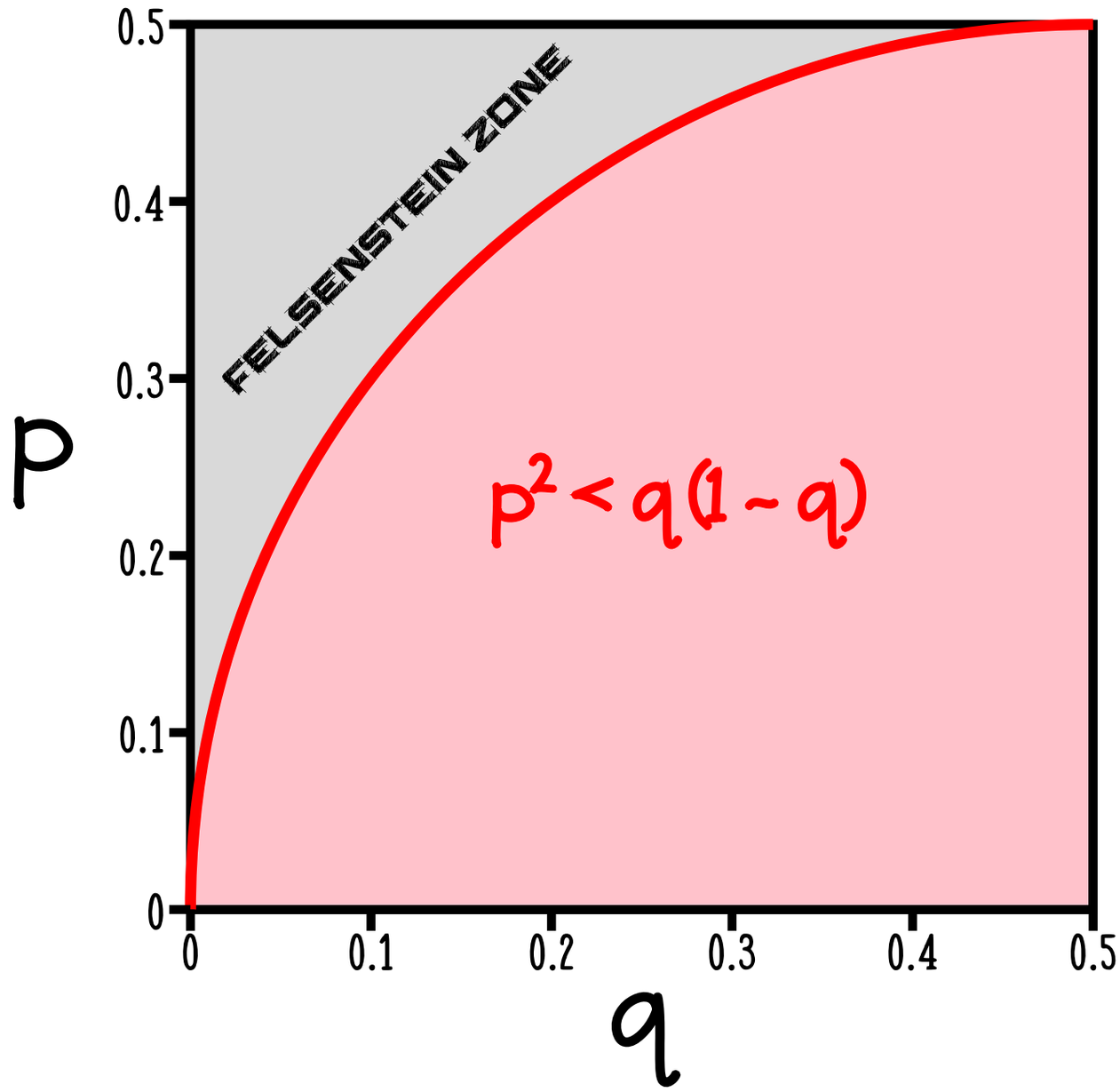
the branch lengths (p, q) represent the probability of change along a branch

(Felsenstein 2004)

The Felsenstein Zone

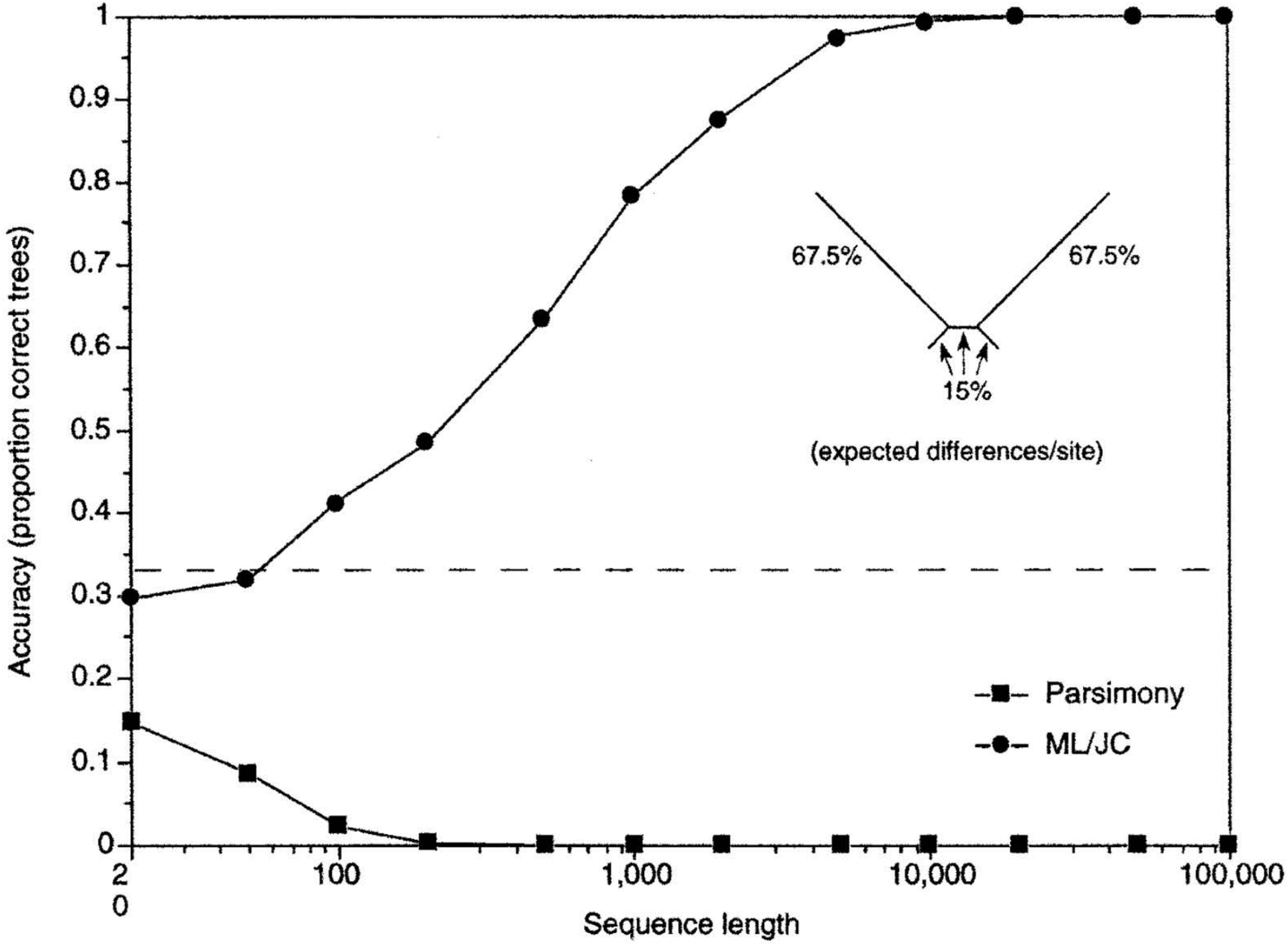


the branch lengths (p, q) represent the probability of change along a branch



(Felsenstein 2004)

Long-Branch Attraction



(Swofford et al. 2001)

Parsimony can be Inconsistent

if one feels that consistency is a desirable property for an estimator…

the inconsistency of parsimony is the strongest argument against its use