Phylogenetic inference
Phylogenetic Biology -Week 2

Biology 1425

Professor: Casey Dunn, dunnlab.org

Brown University

Front matter...

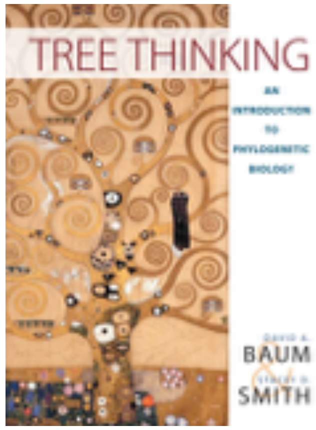
All original content in this document is distributed under the following license:

Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (http://creativecommons.org/licenses/by-nc-sa/3.0/deed.en_US)

See sources for copyright of non-original content

Sources

Some non-original content is drawn from:



Baum, D and S. Smith (2012) Tree Thinking: and Introduction to Phylogenetic Biology. Roberts and Company Publishers. ISBN 9781936221165

Other non-original content is referenced by url.

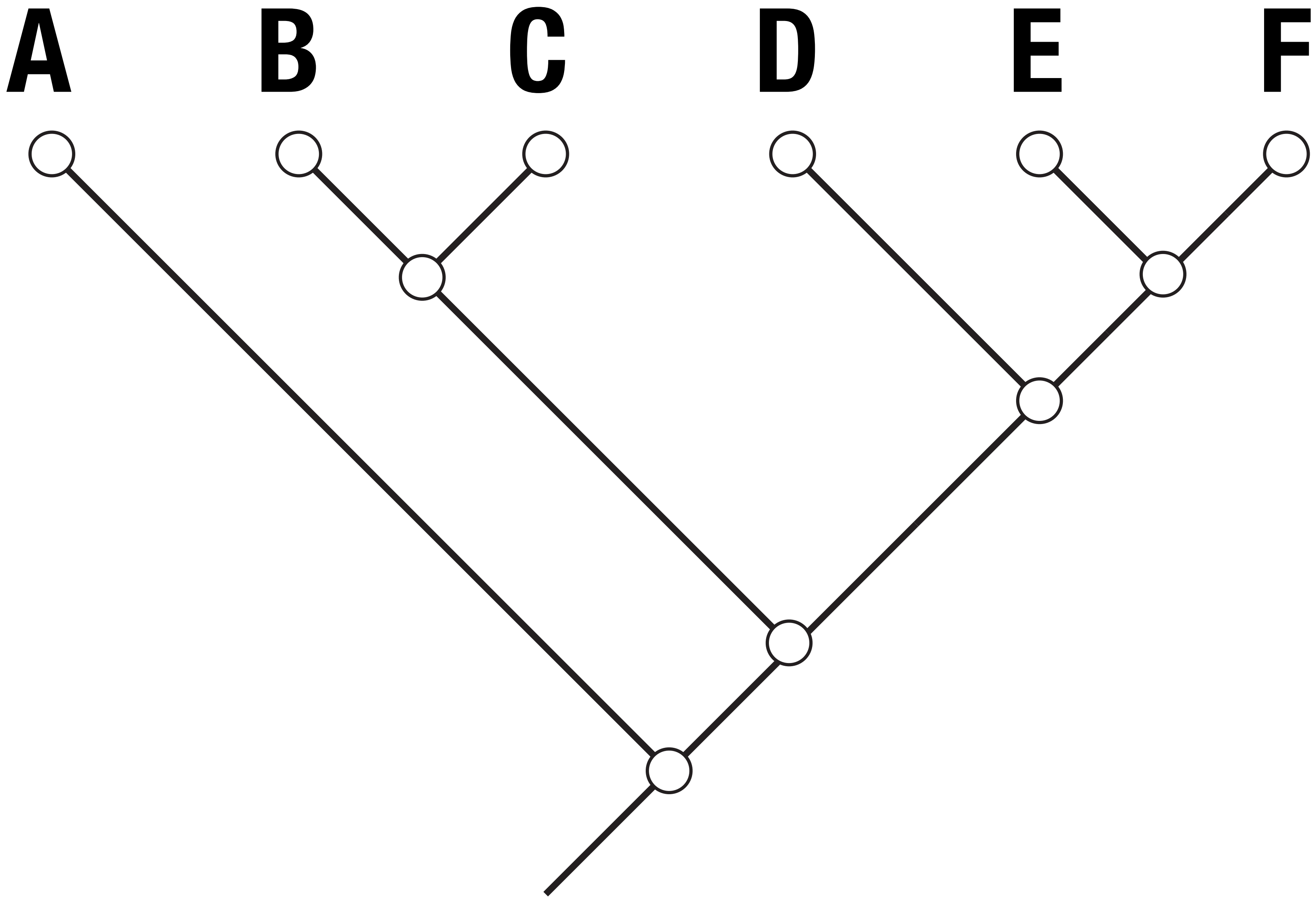
Some definitions...

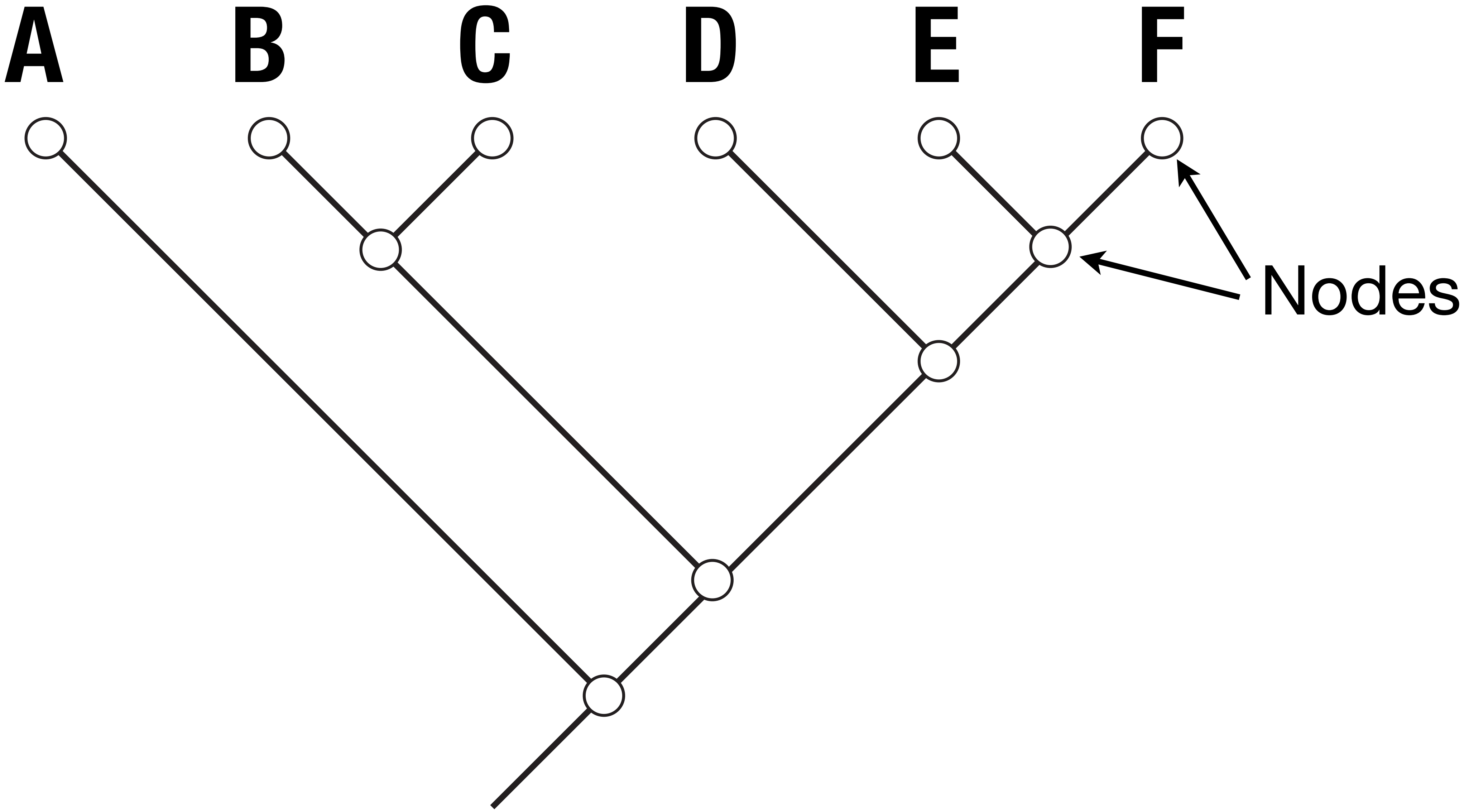
Taxon

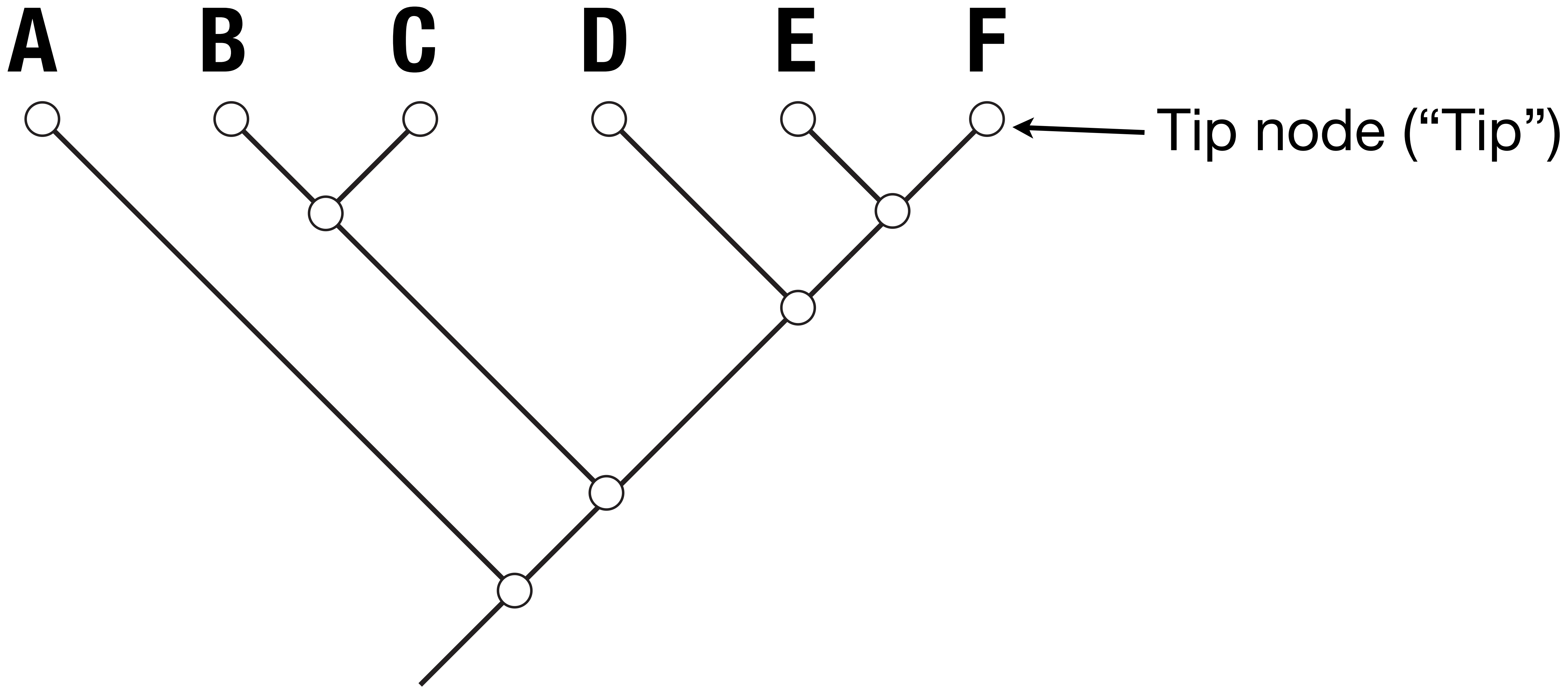
A named group of evolving entities (*e.g.* genes, species, languages)

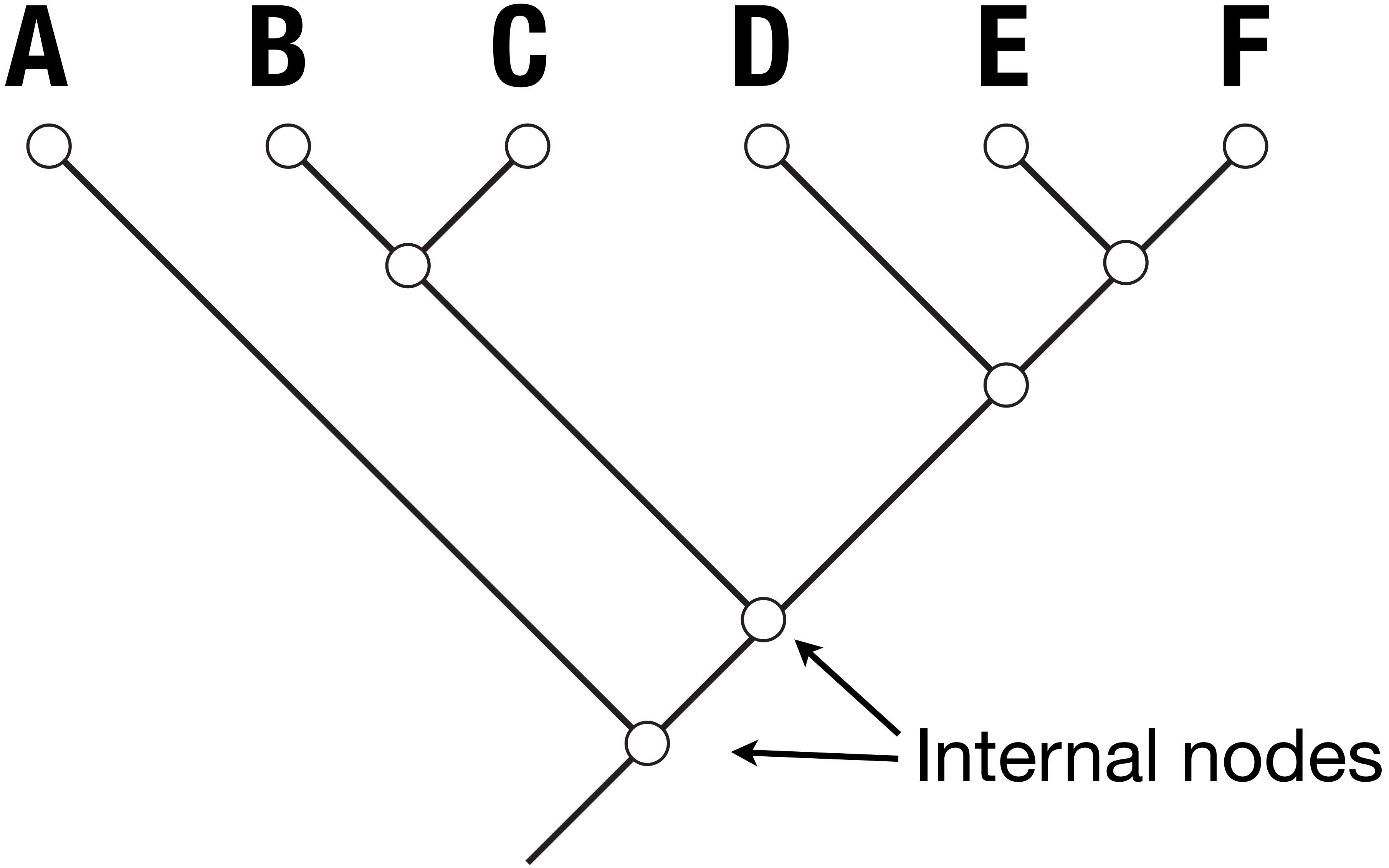
Phylogeny

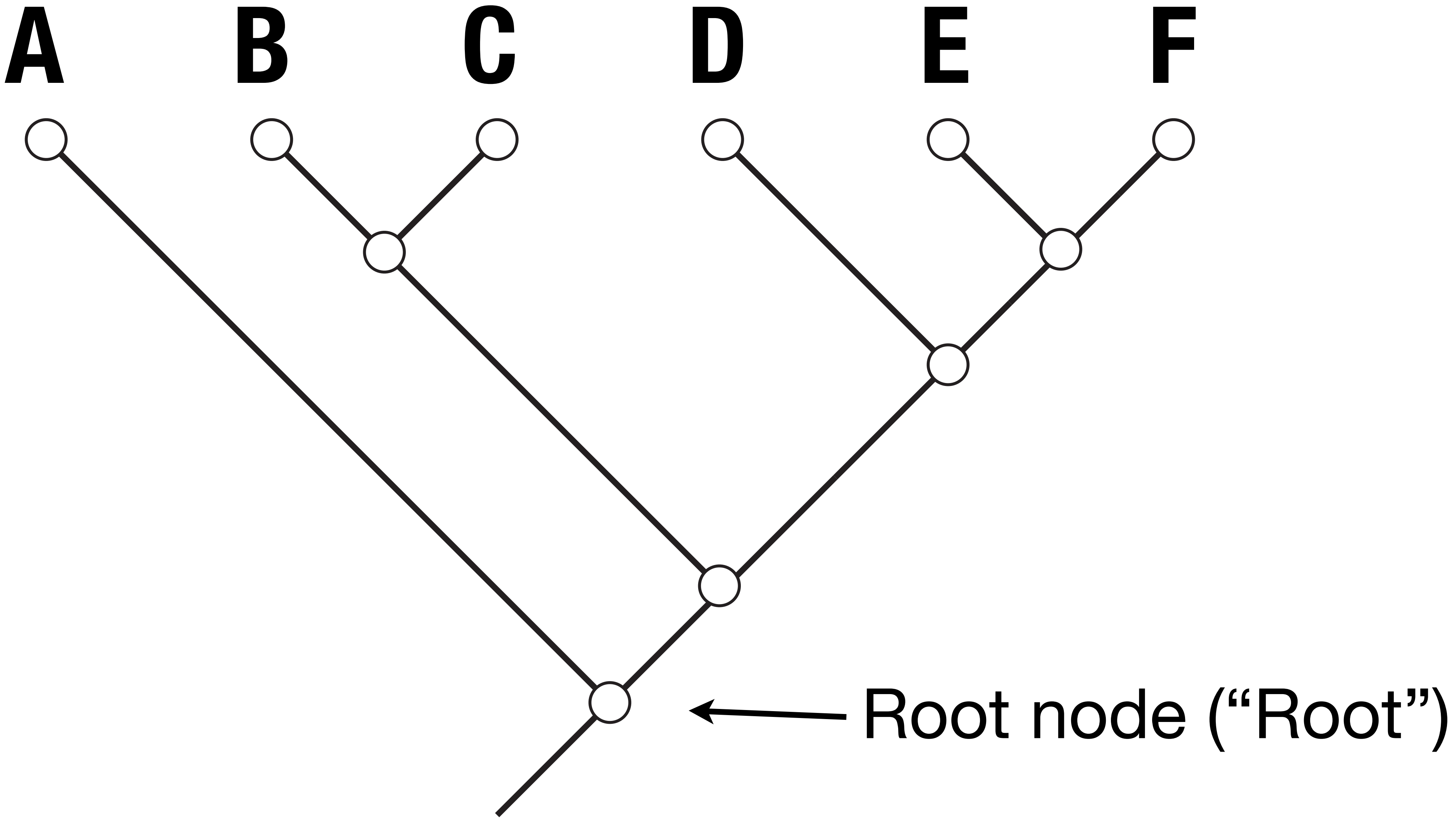
A tree depicting the evolutionary relationships between taxa

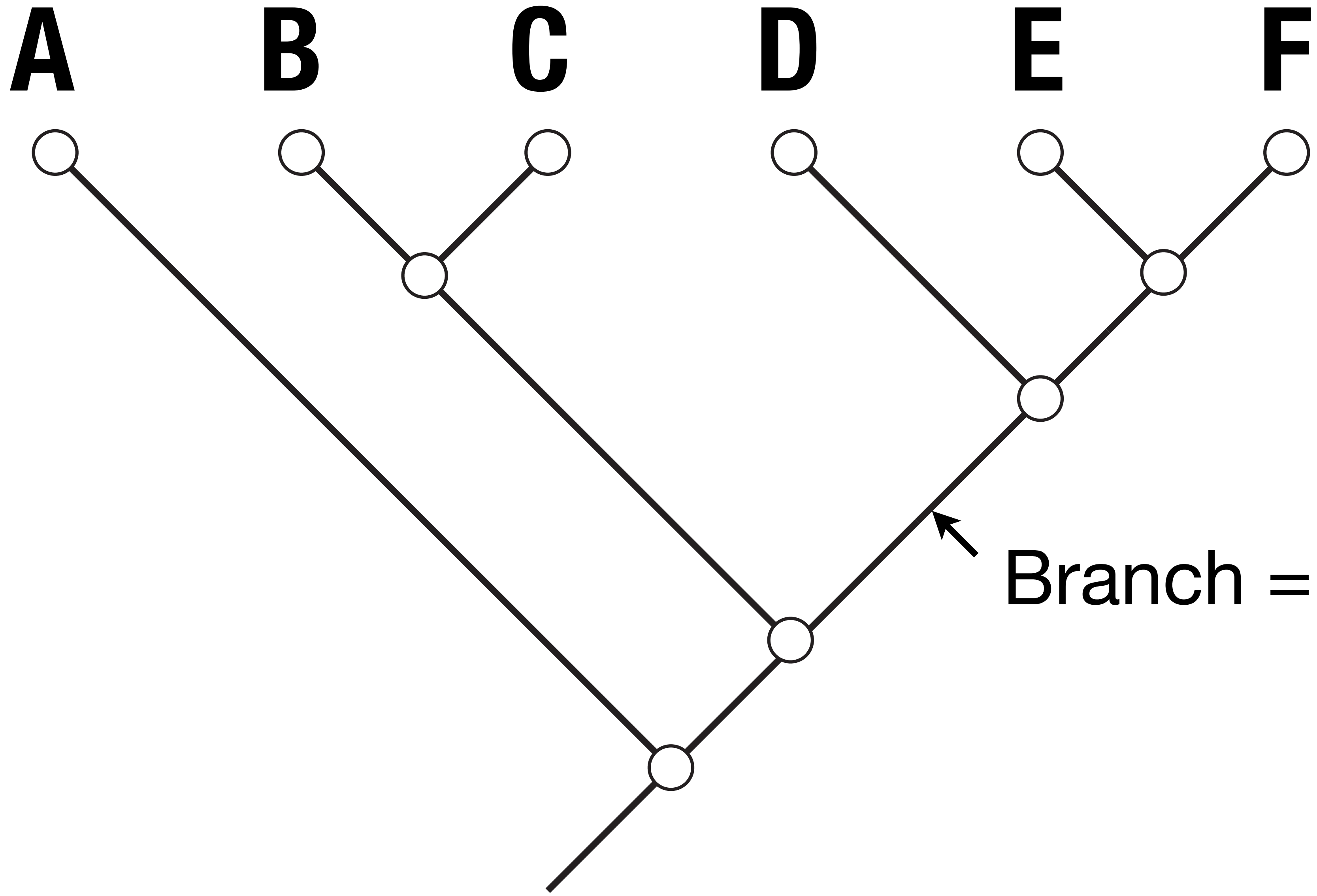



A
B
C
D
E
F
Nodes

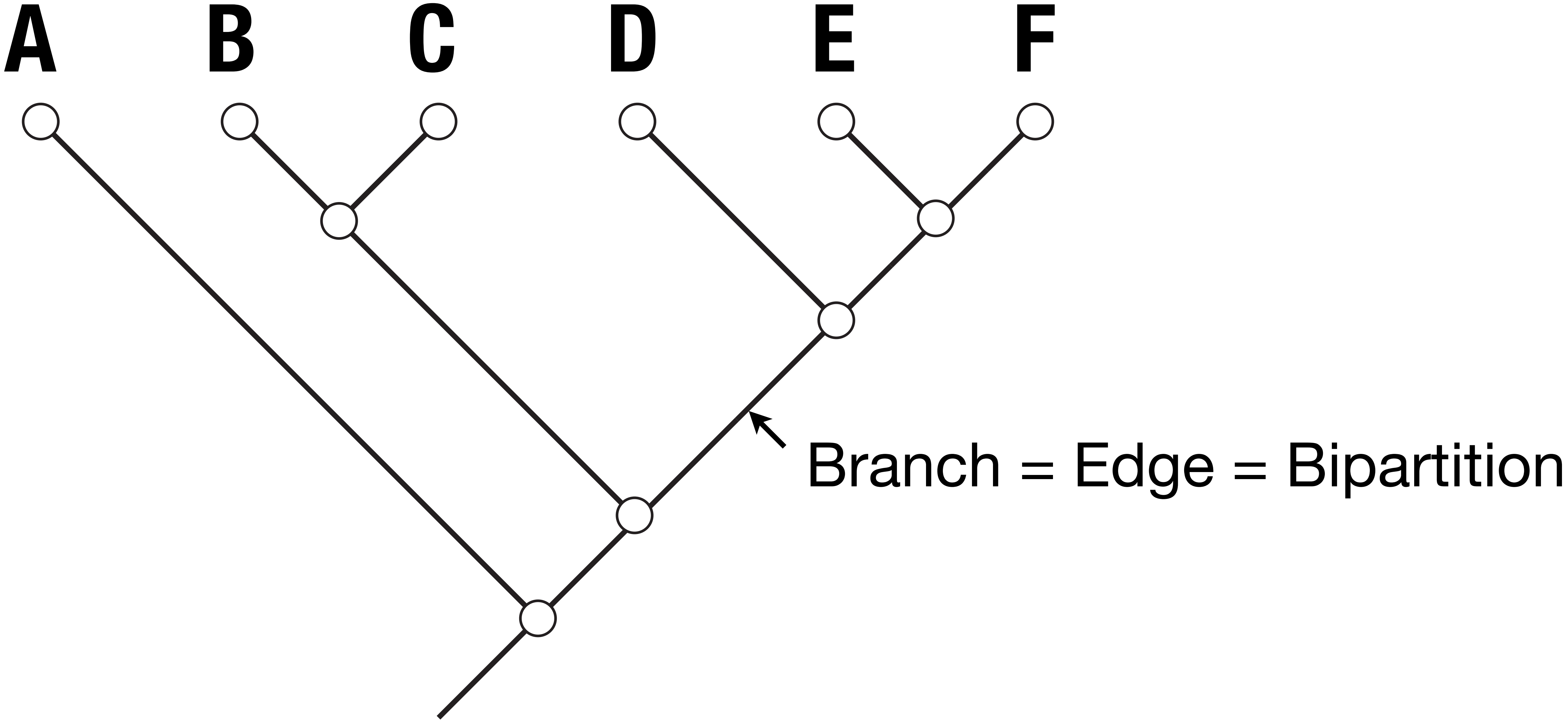

A
B
C
D
E
F
Tip node ("Tip")


A
B
C
D
E
F
Internal nodes


A
B
C
D
E
F
Root node ("Root")

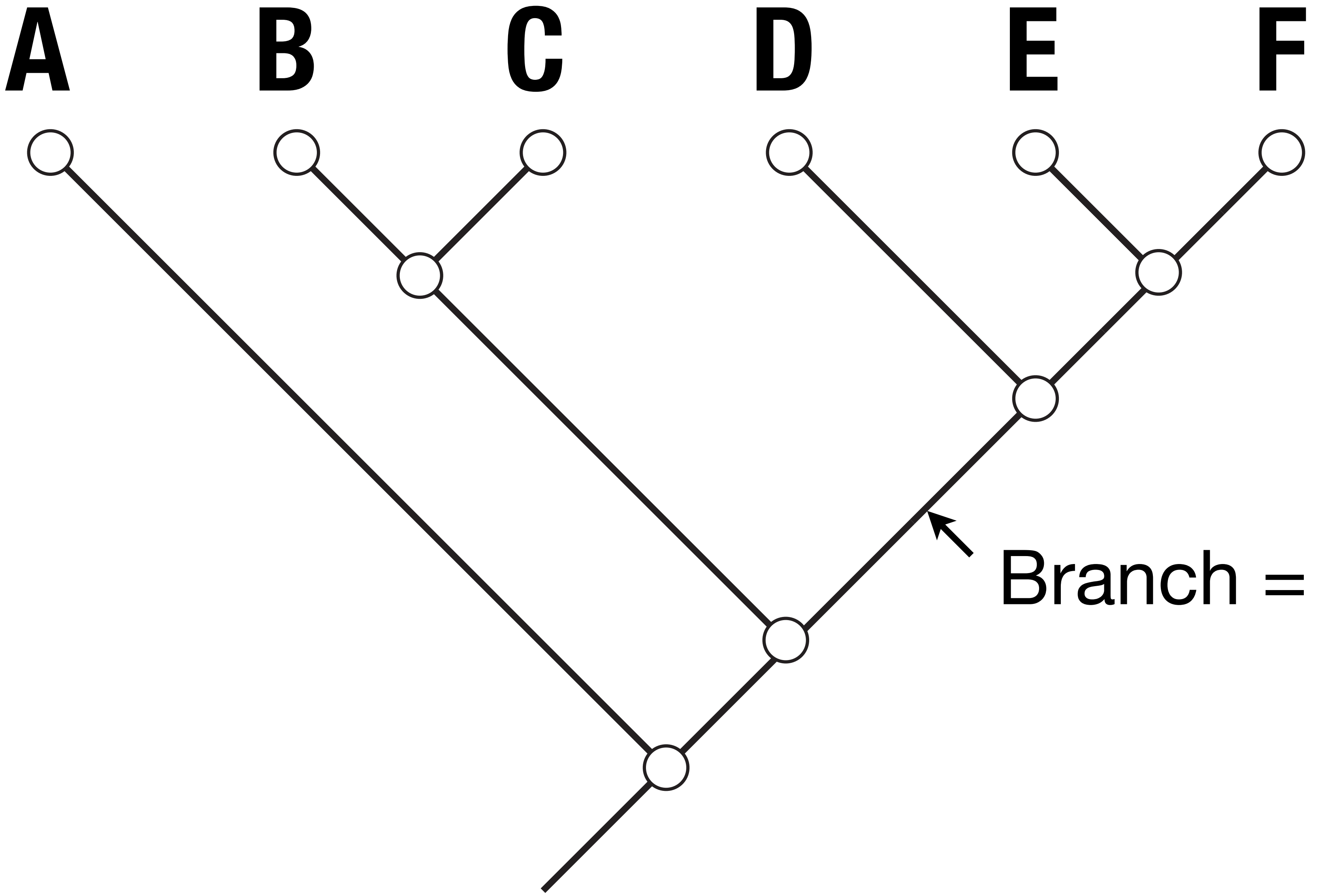


Branch = Edge = Bipartition



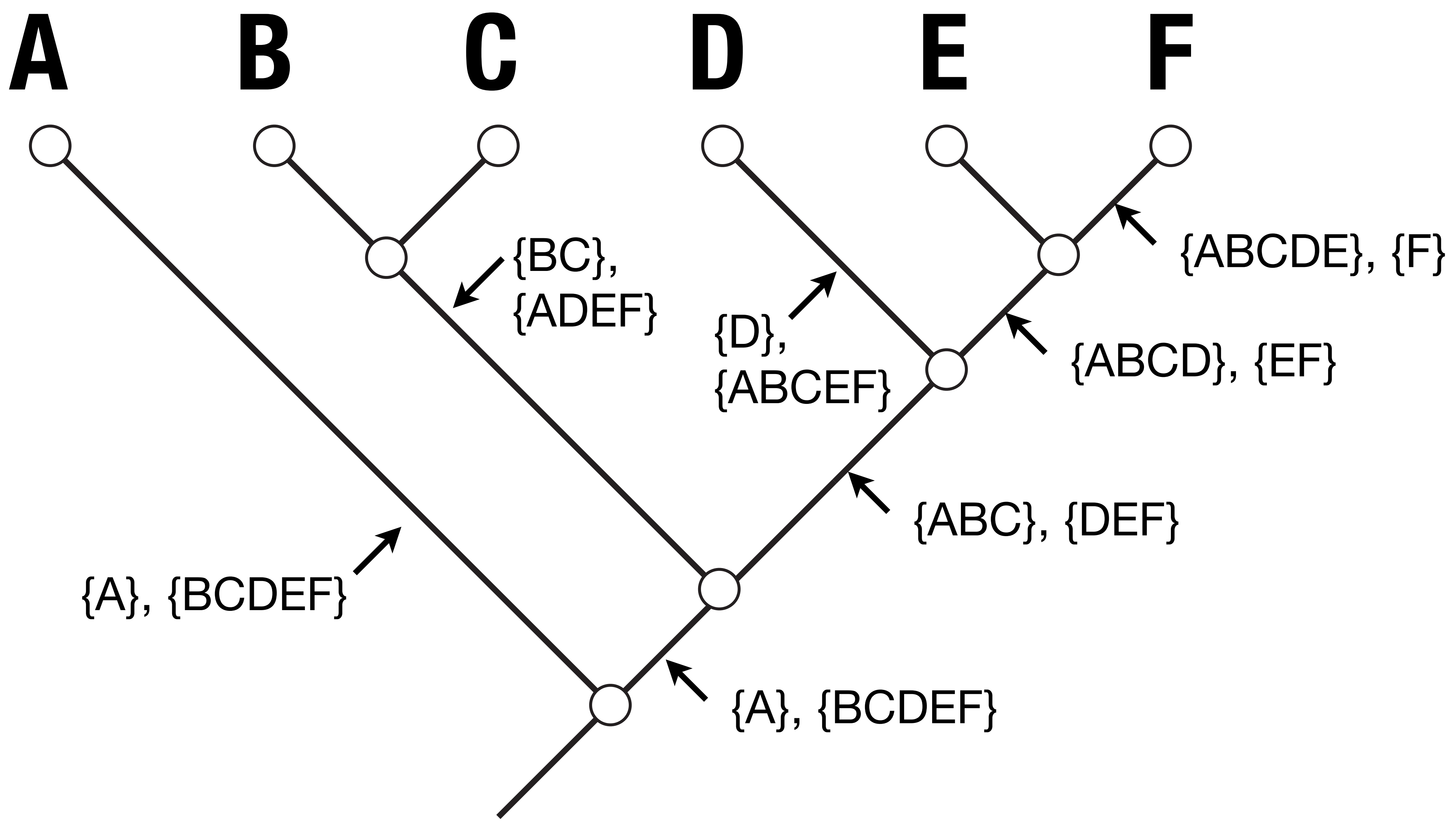
Branch = Edge = Bipartition

Why are branches sometimes called bipartitions?
Because each divides the tree into 2 sets of tips



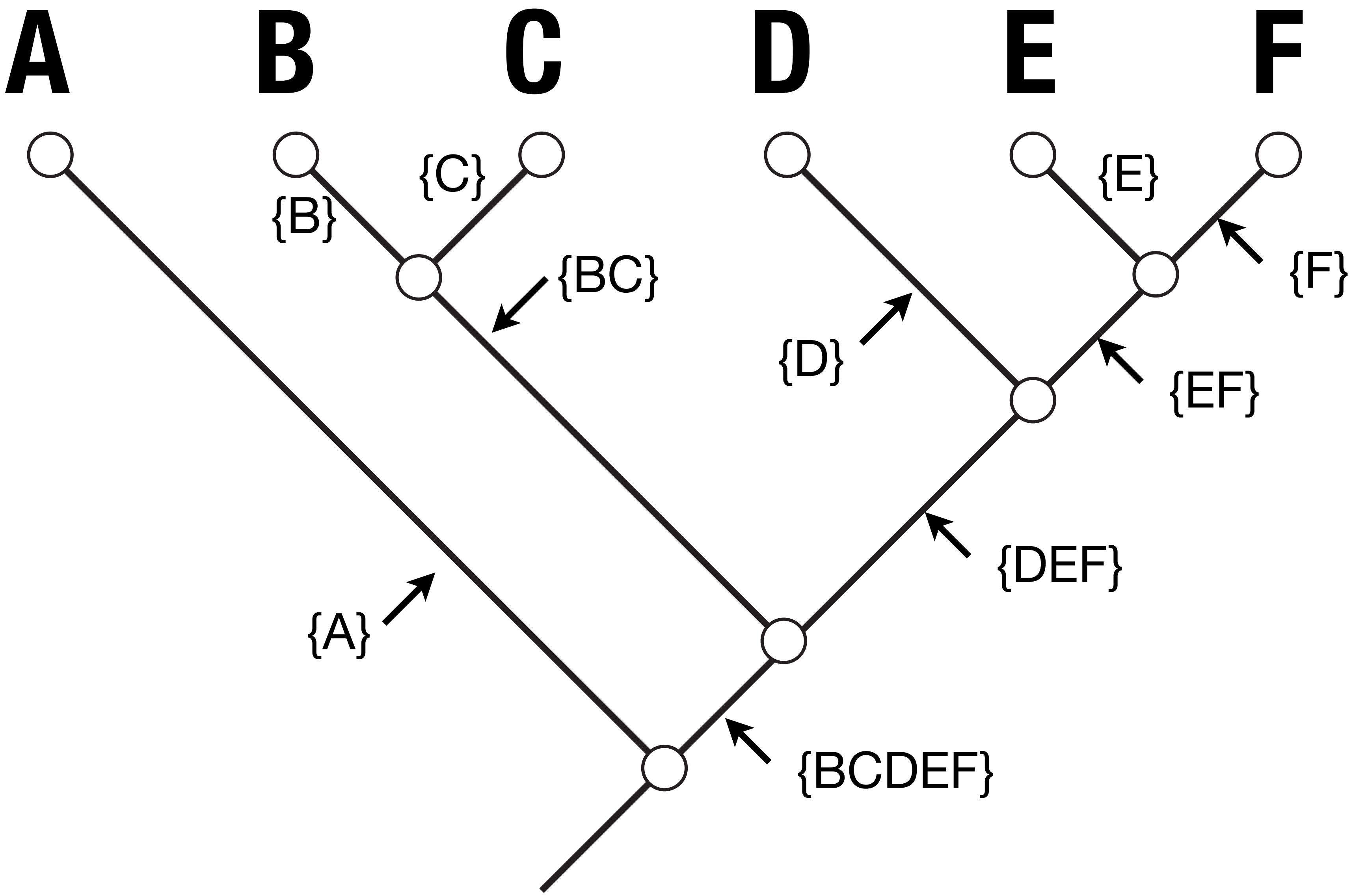
Branch = Edge = Bipartition

Why are branches sometimes called bipartitions?
Because each divides the tree into 2 sets of tips



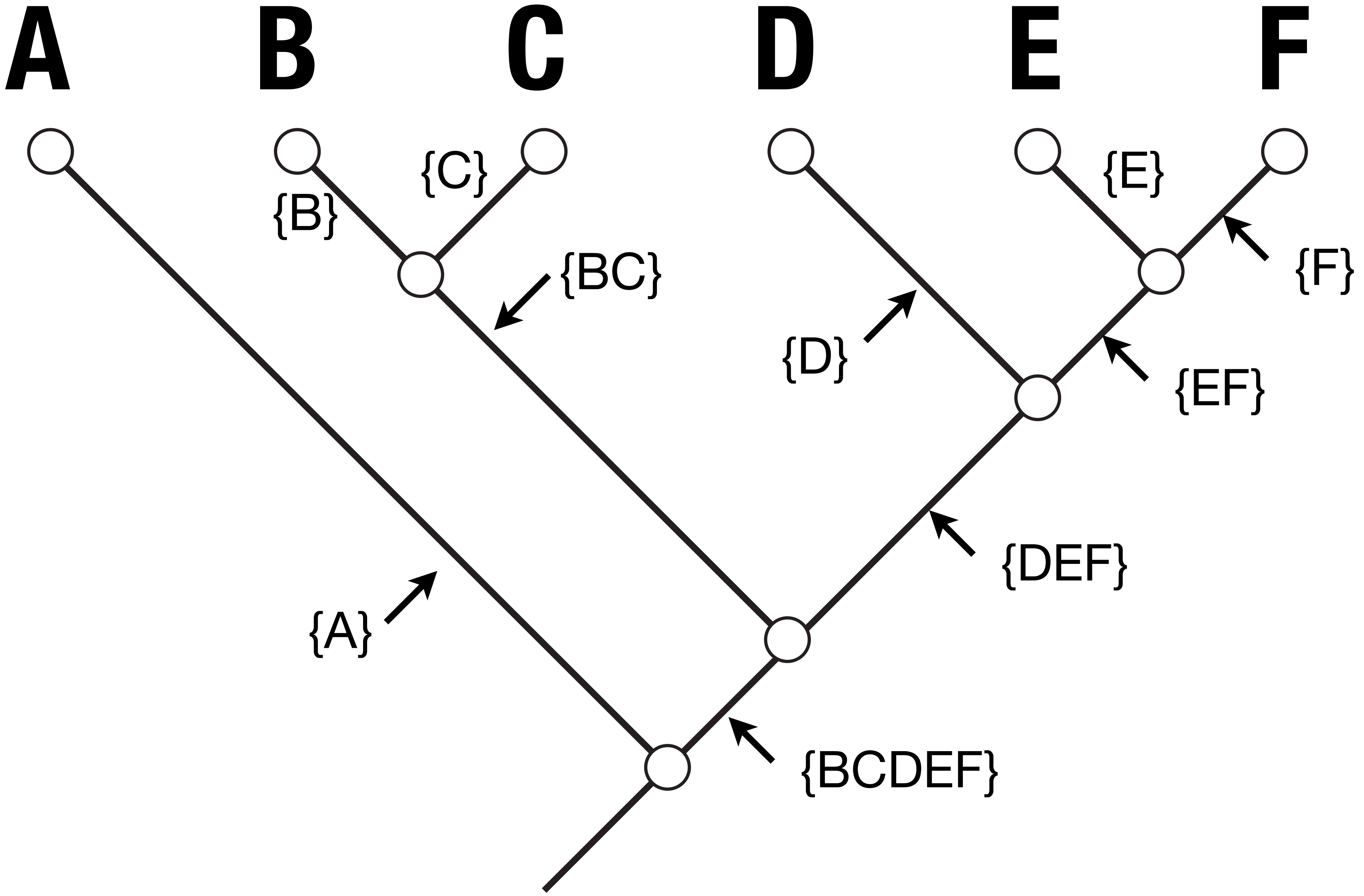
(not all shown)

Since the two sets include all tips, you only need to state one set to know what the other is.



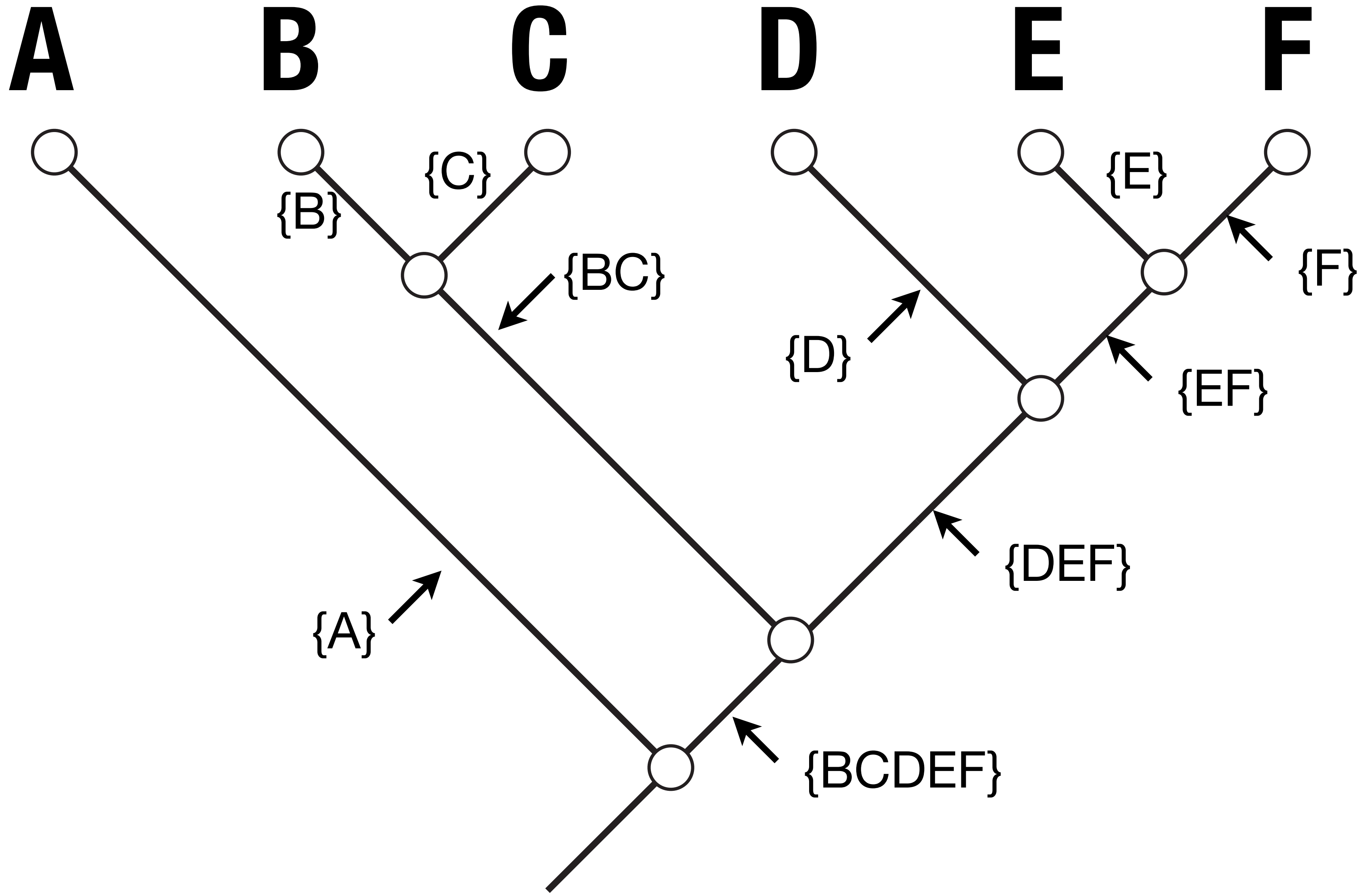
Any fully resolved tree with more than 3 taxa has only a subset of the possible bipartitions.

Can describe a tree by the set of its bipartitions:



{A}
{B}
{C}
{D}
{E}
{F}
{BC}
{EF}
{DEF}
{BCDEF}

Two bipartitions can be compatible, i.e. they can both exist on the same fully resolved tree, or incompatible.

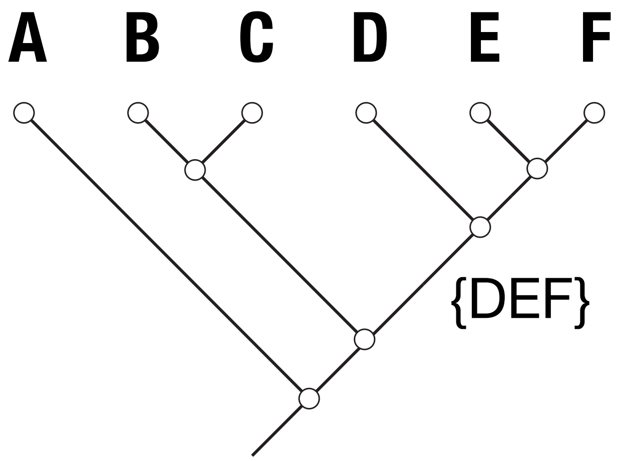


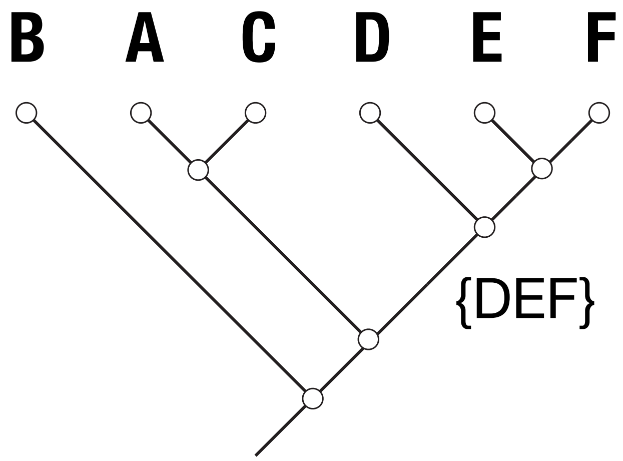
{DEF} and {EF} are compatible (both on this tree)

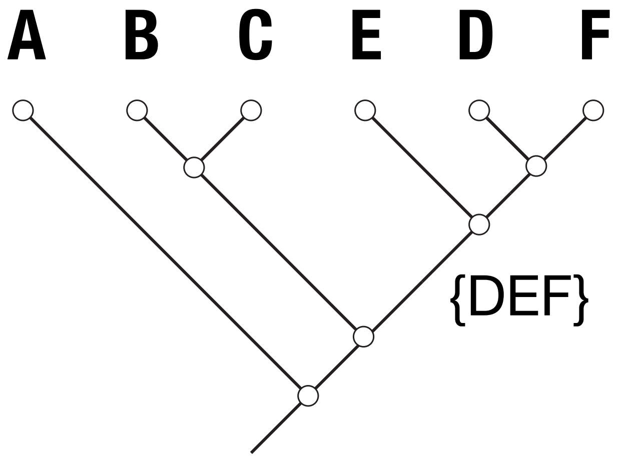
{DEF} and {BCF} are incompatible ({BCF} not on this tree)

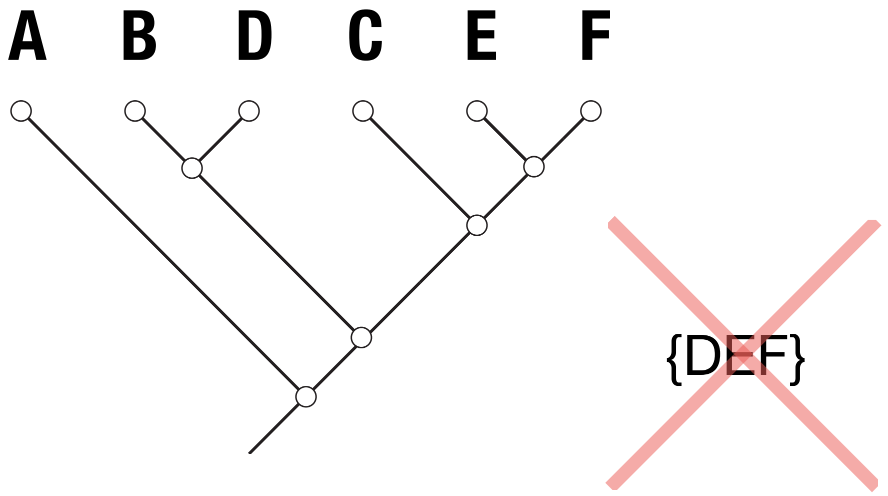
{CA} and {DF} are compatible (neither is on this tree)

The nice thing about thinking of trees as collections of bipartitions is that we can define and talk about one particular branch independent of the other branches



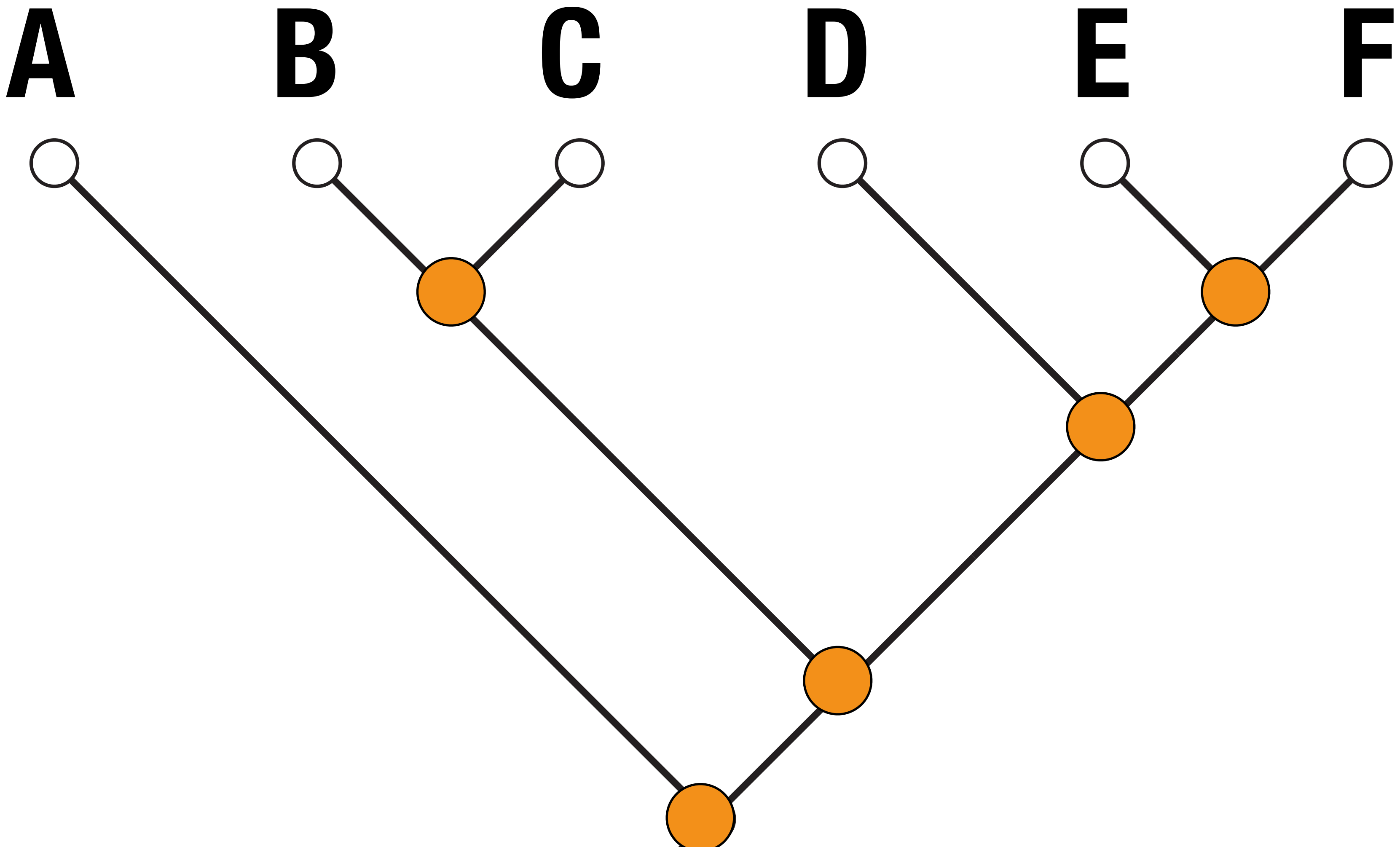






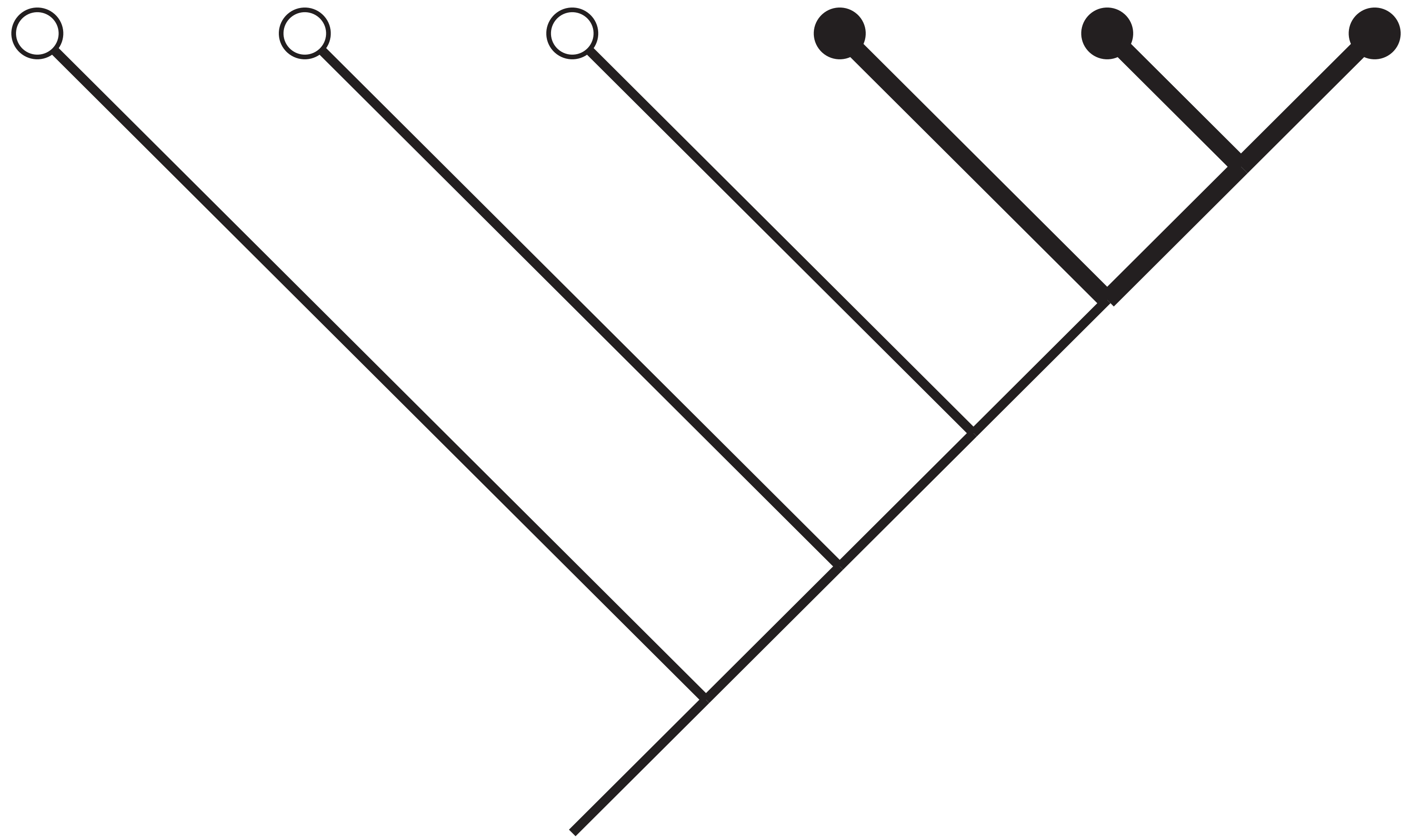
On a fully resolved rooted tree:

- ○ N tip nodes
- ● N-1 internal nodes
- / 2N-2 branches



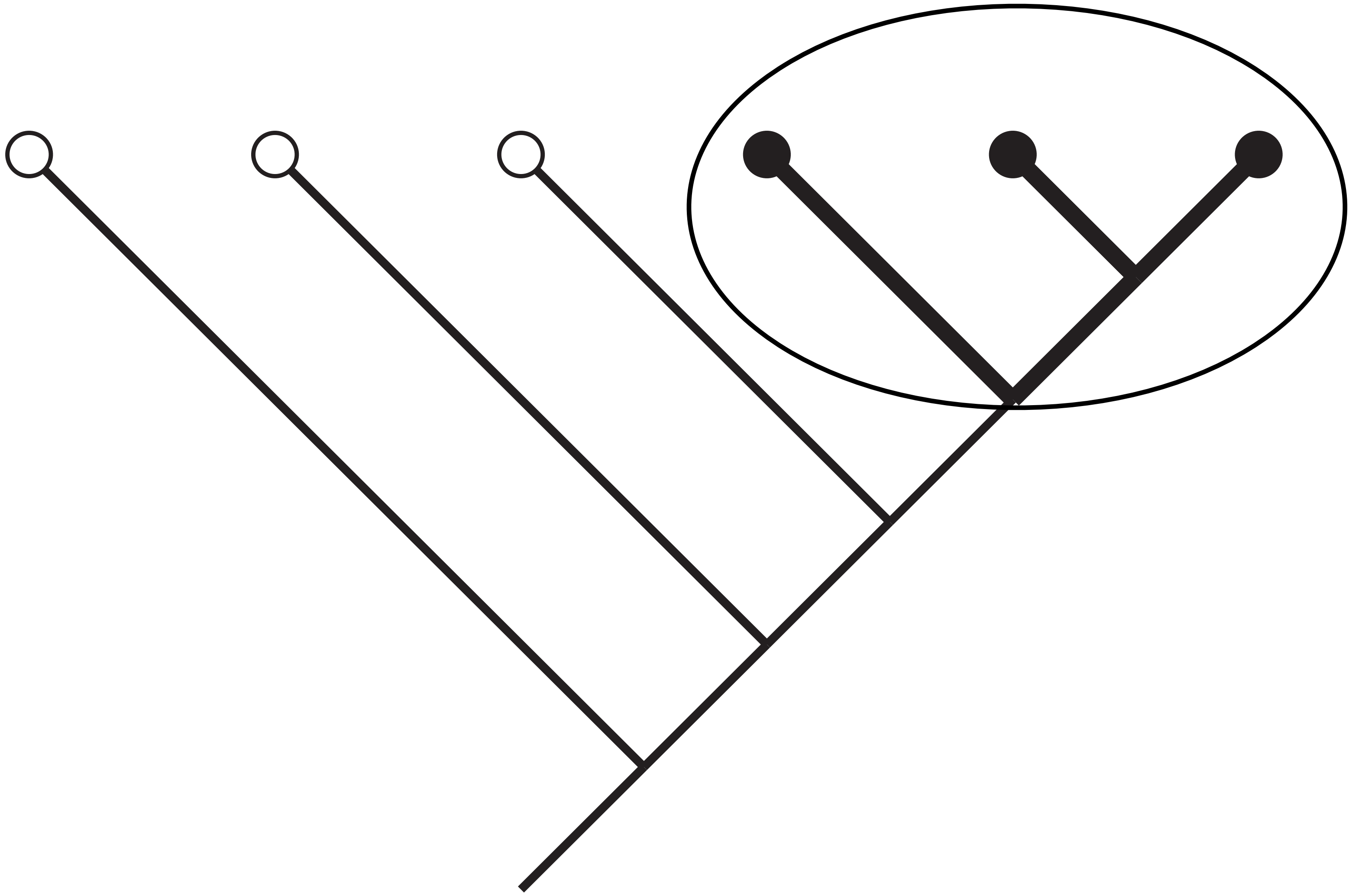
Monophyletic

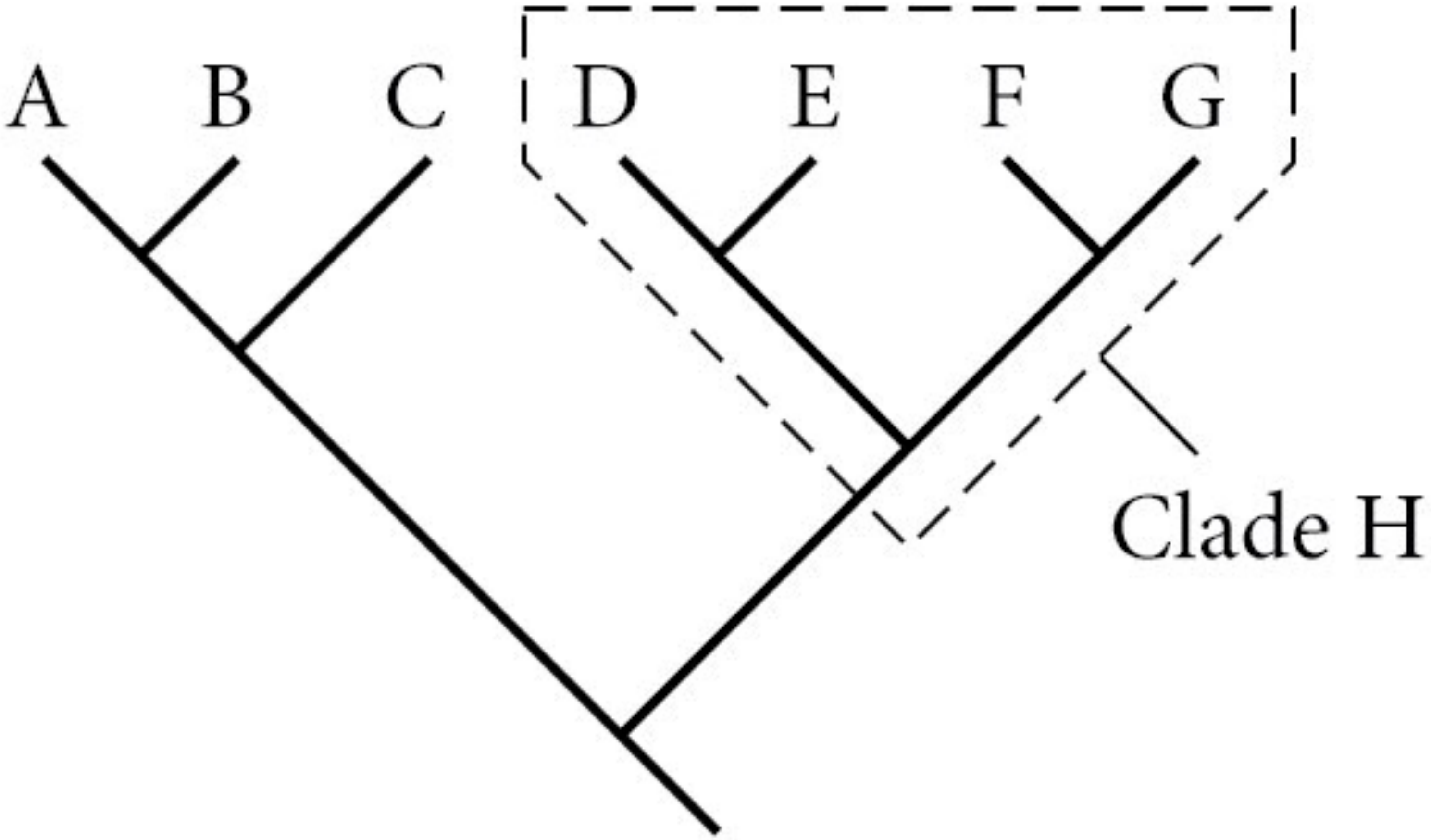
To include an ancestor, all of its descendants, and nothing that isn’t a descendant

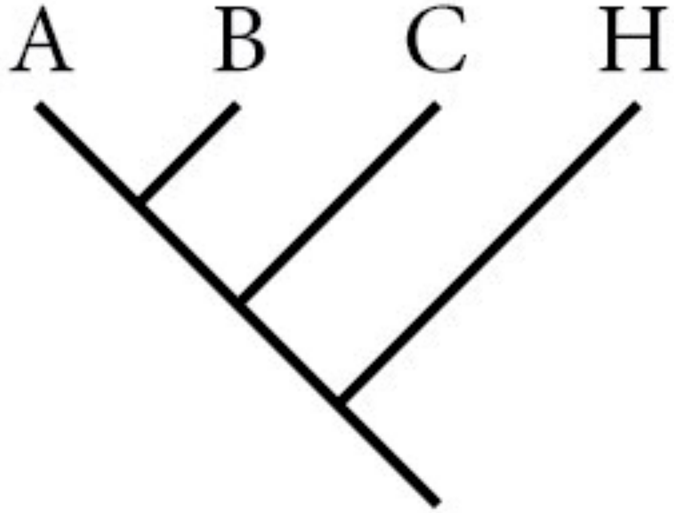


Clade

A monophyletic group

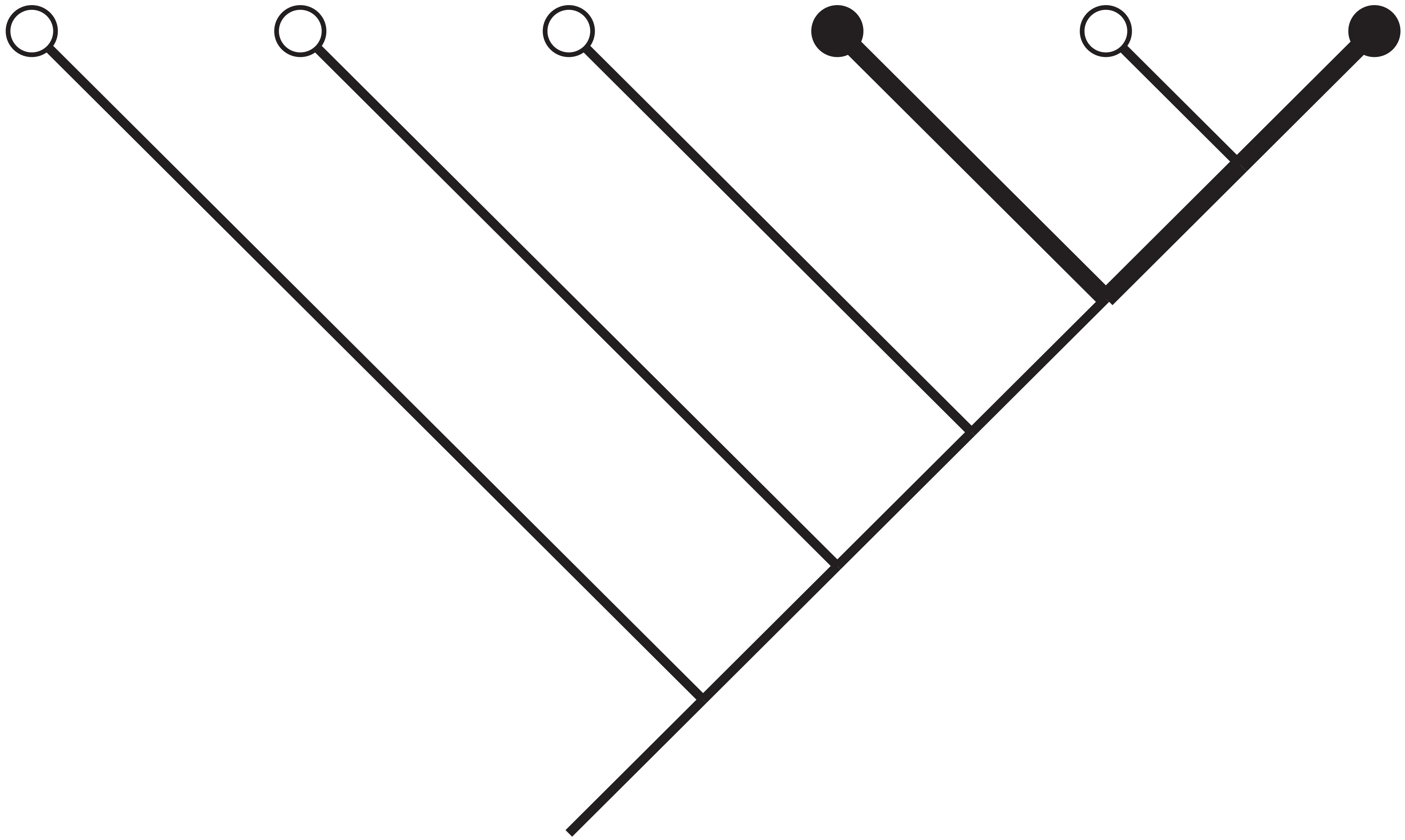



A
B
C
D
E
F
G
Clade H


A
B
C
H

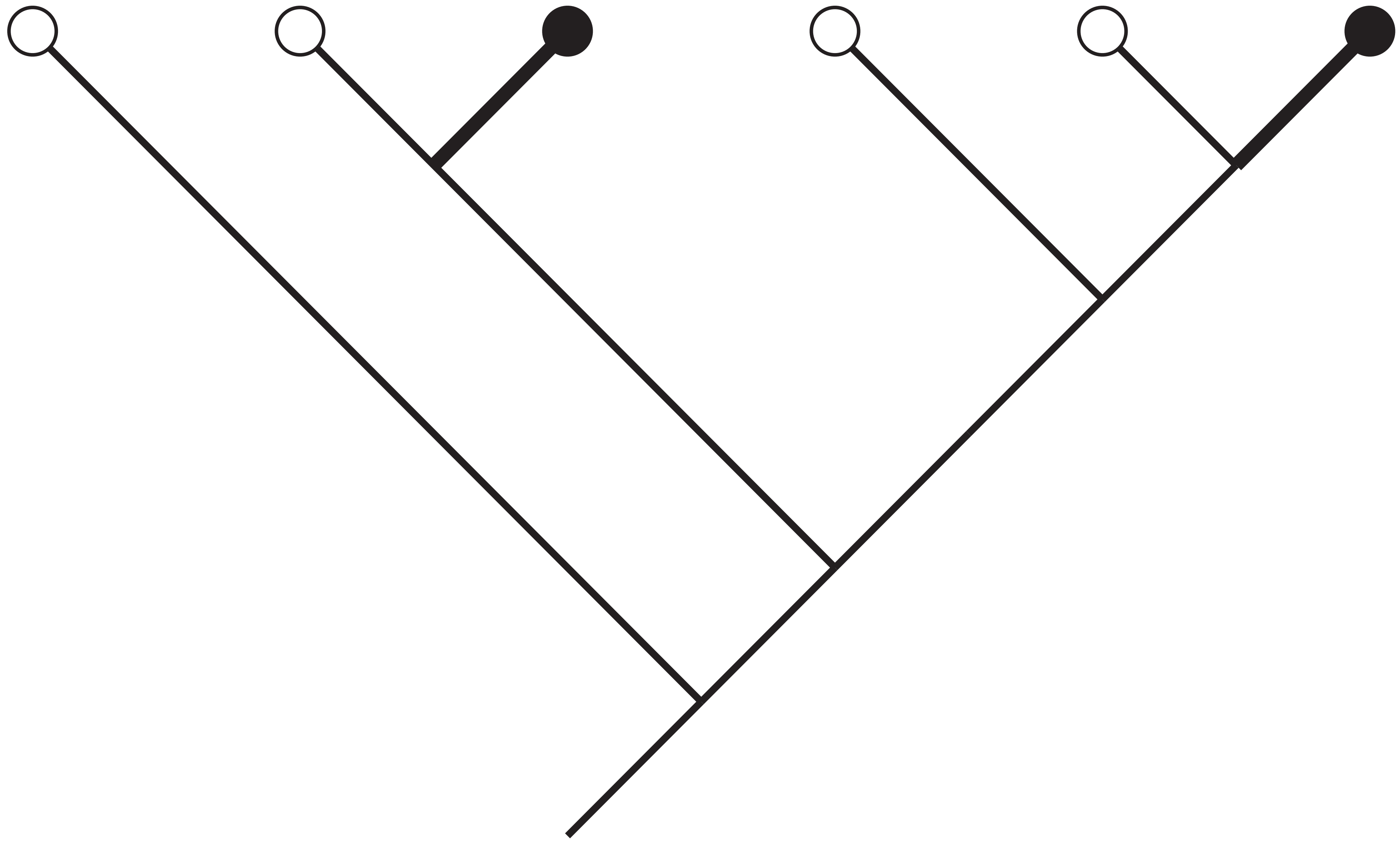
Paraphyletic

To include an ancestor and only some of its descendants



Polyphyletic

To not include an ancestor



Character

Any heritable attribute of a taxon

Homologous characters

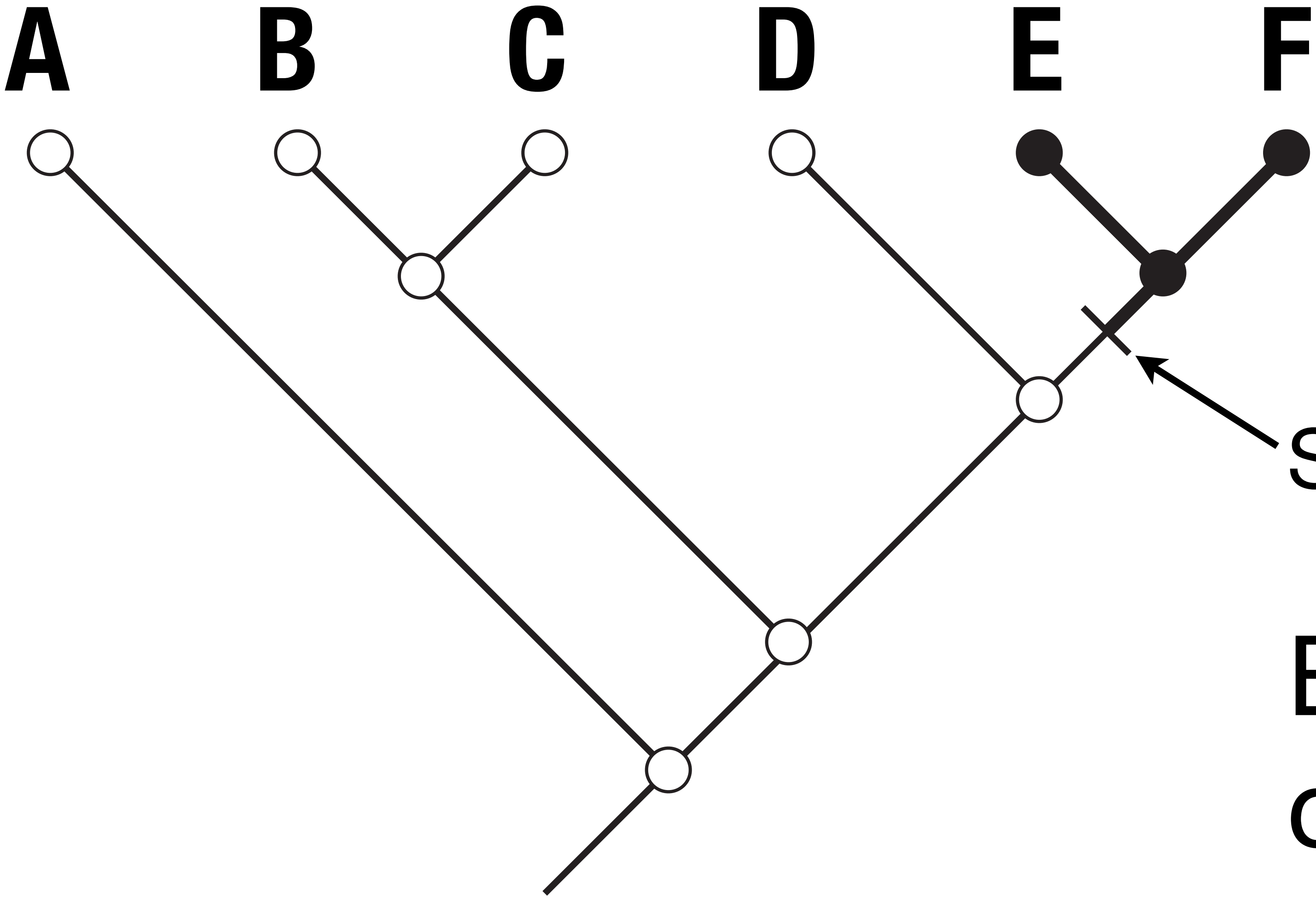
Characters that are present in taxa because they were also present in the most recent common ancestor of those taxa

Character state

A value that a character can have. (*e.g.*, A, C,G, or T for a DNA nucleotide)

Apomorphy (Derived character state)

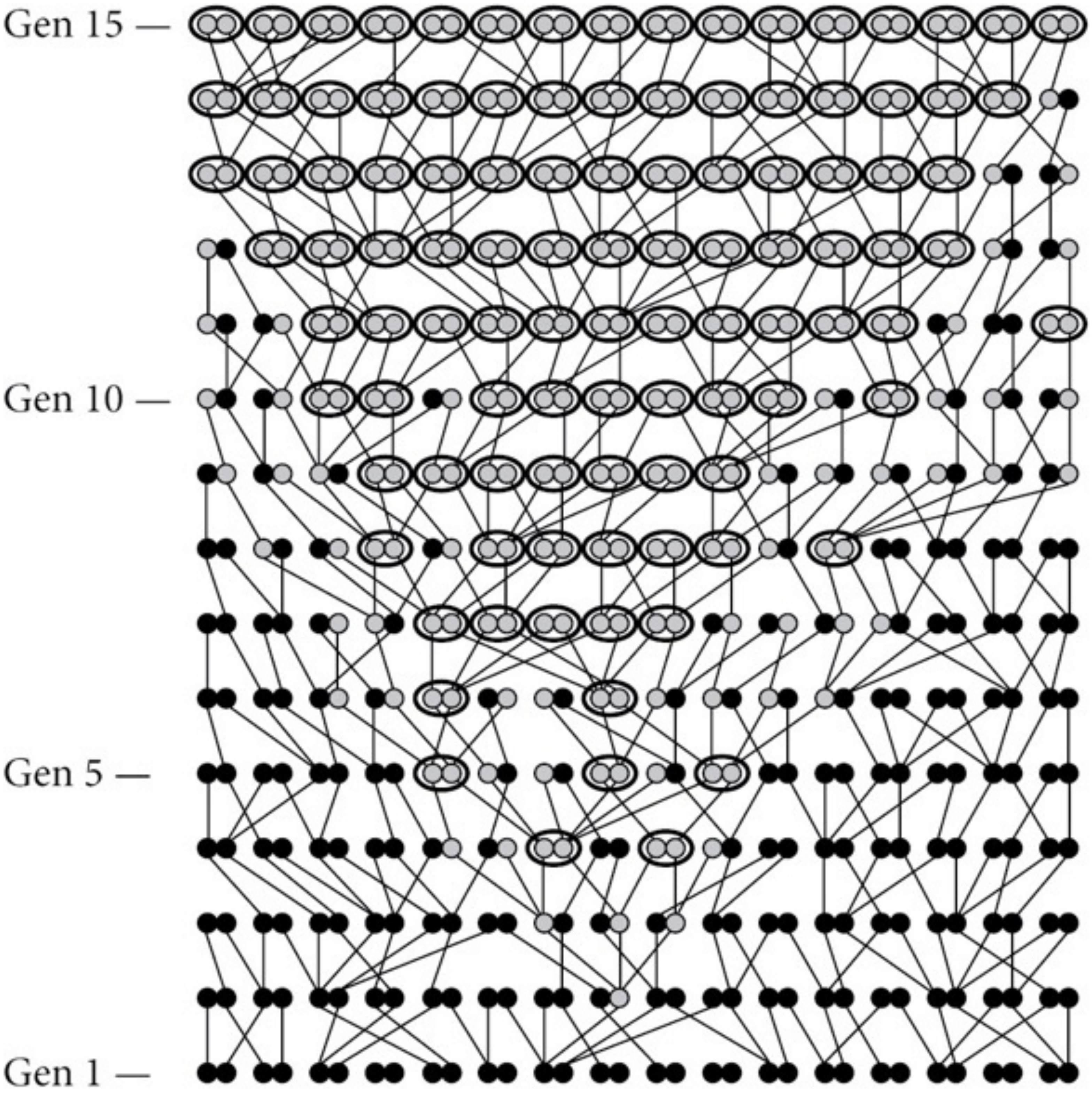
A character state that is different from the character state of an ancestor



Switch to black

Black is a derived character state (it replaced white)

Apomorphy (Derived character state)

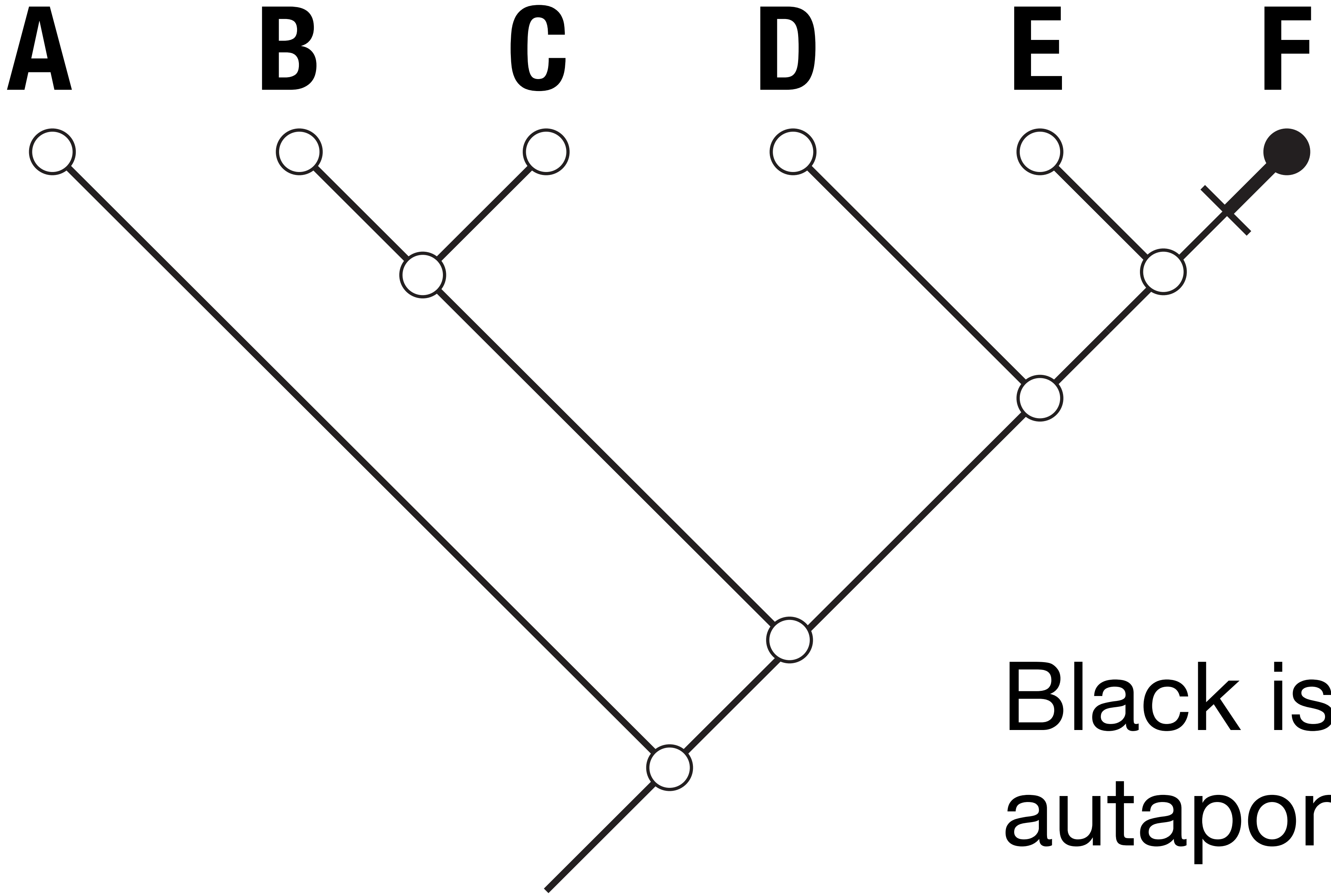


Fixation of derived character states

Baum and Smith 2012, Figure 4.1

Autapomorphy

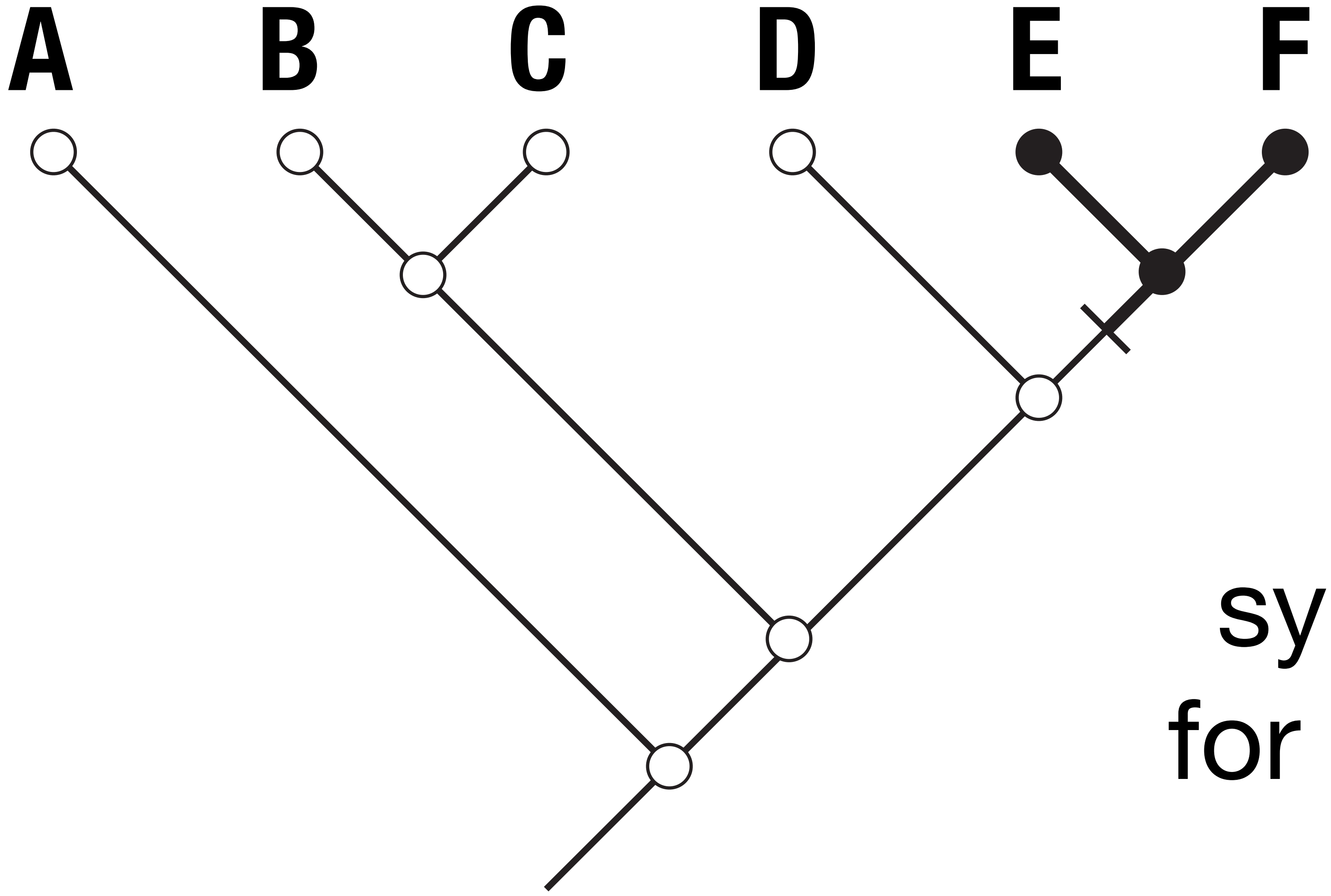
An apomorphy that is unique to a single taxon



Black is an autapomorphy of **F**

Synapomorphy

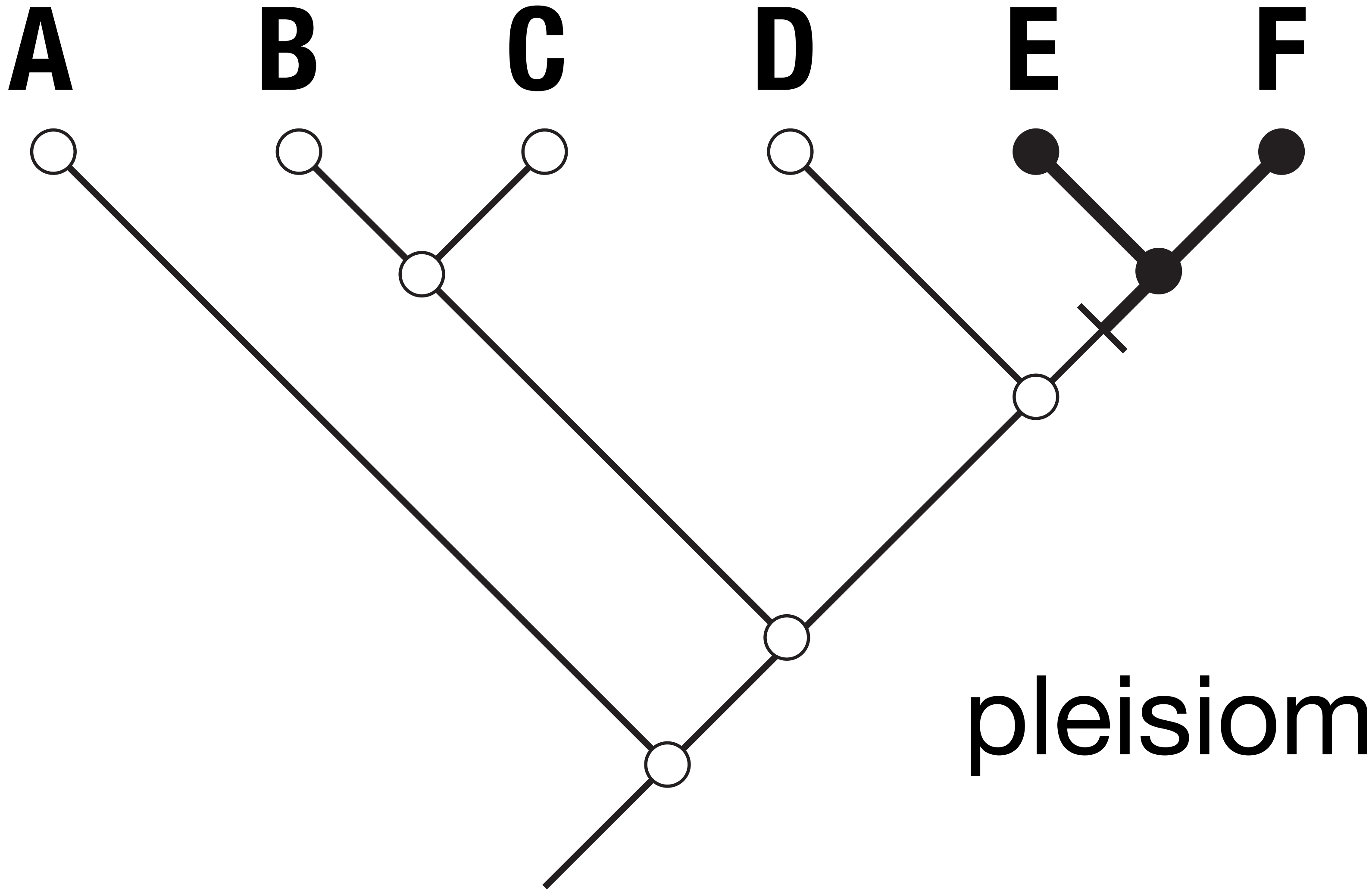
A shared apomorphy



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

Pleisiomorphy

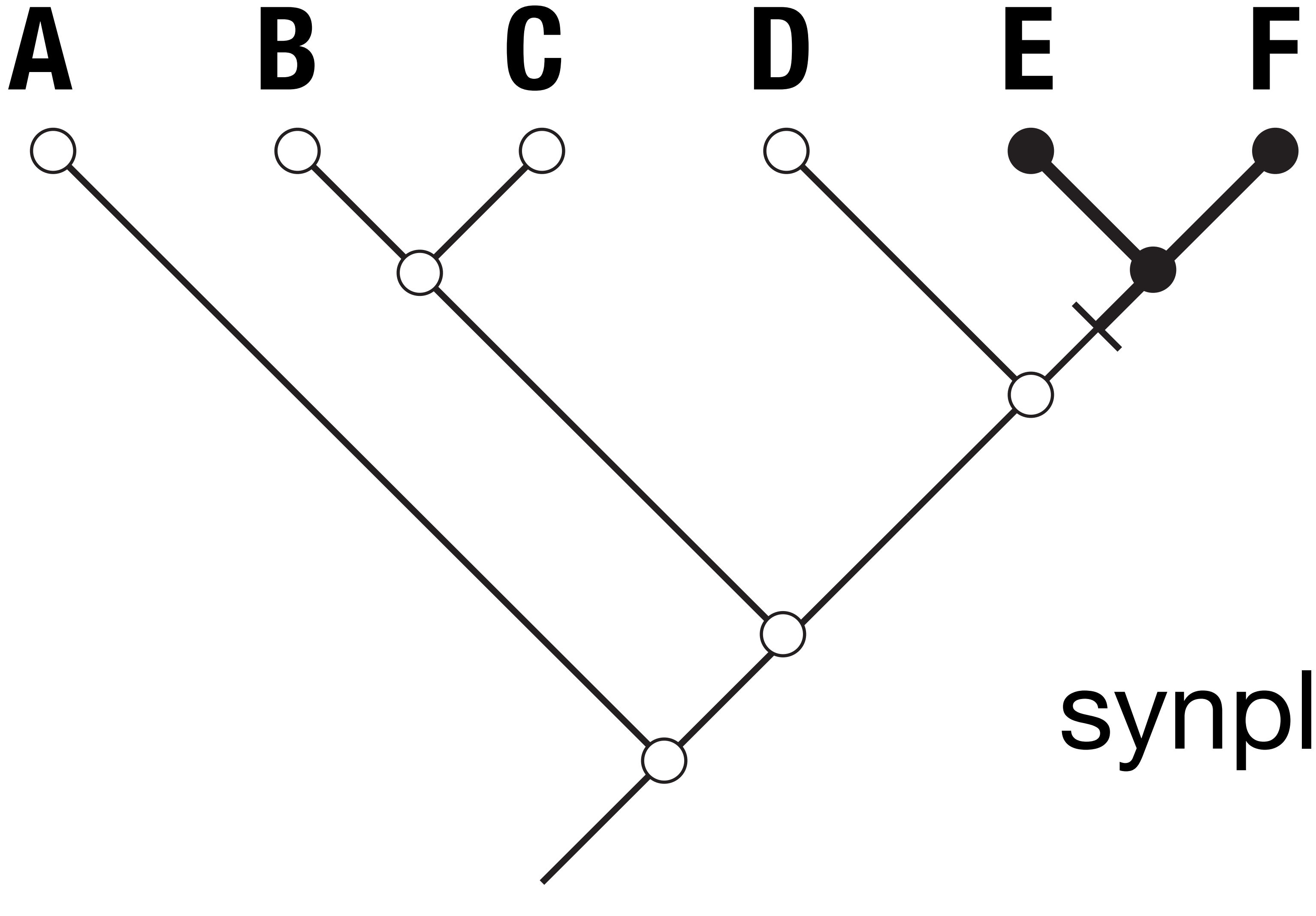
A character state that is the same as that of an ancestor



White is a pleisiomorphy for **B**

Synpleisiomorphy

A shared pleisiomorphy



White is a synpleisiomorphy for **BC**

Homoplasy

Independent origin of the same character state. Causes include convergence and reversals.

Character matrix

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

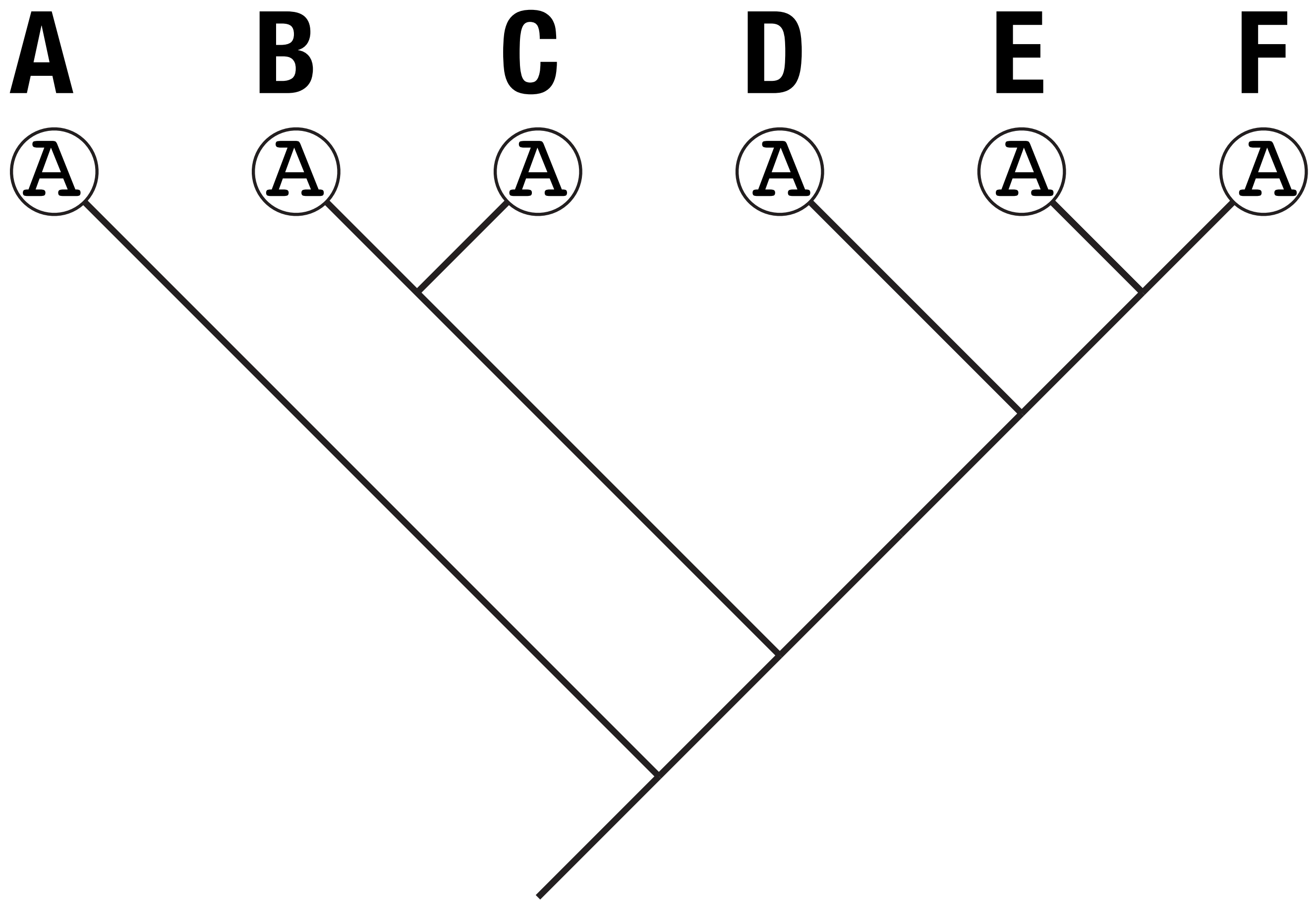
A table of character states. Each row corresponds to a taxon, and each column to a homologous character.

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 1



Character matrix

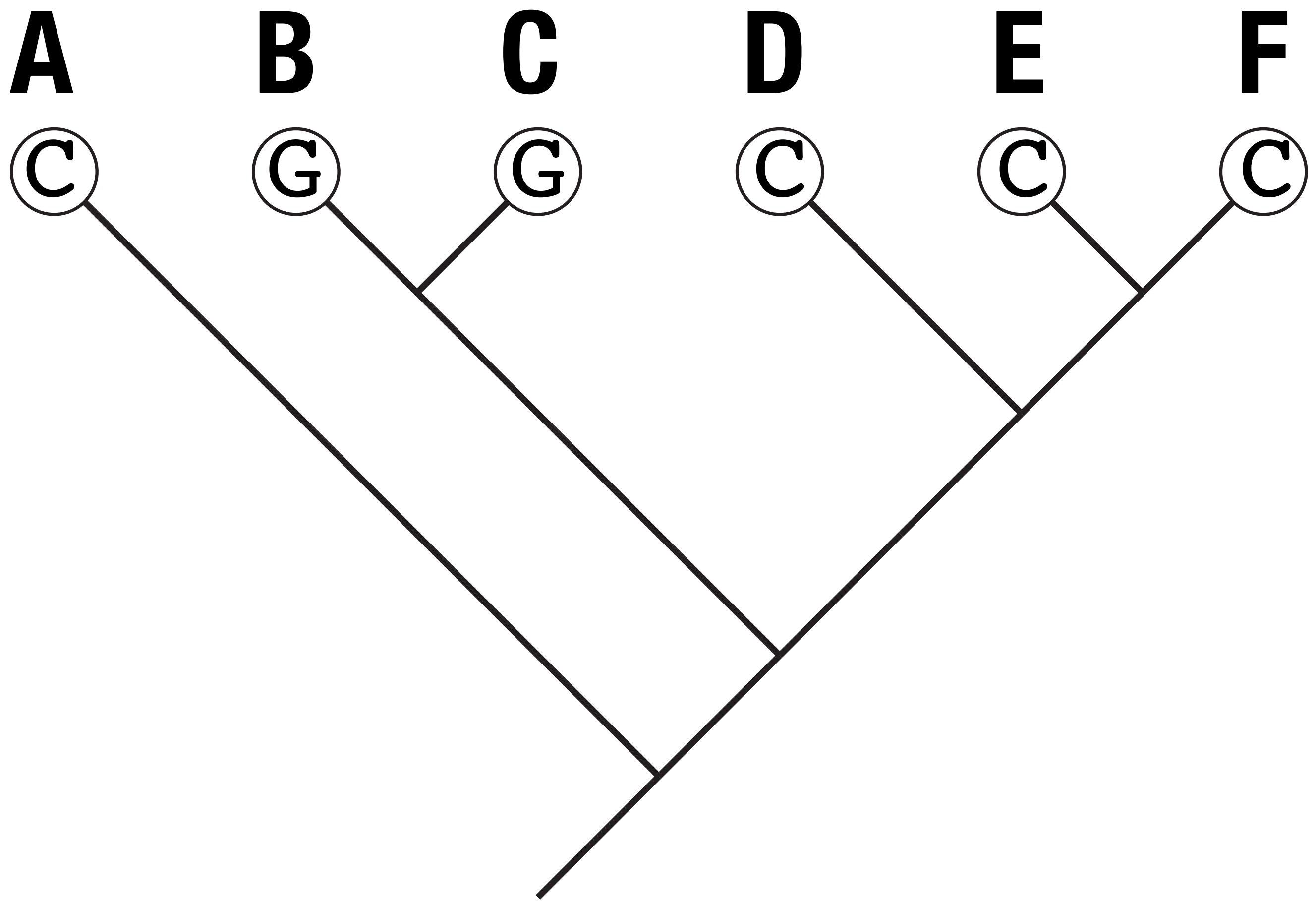
A table of character states. Each row corresponds to a taxon, and each column to a homologous character.

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 2



Character matrix

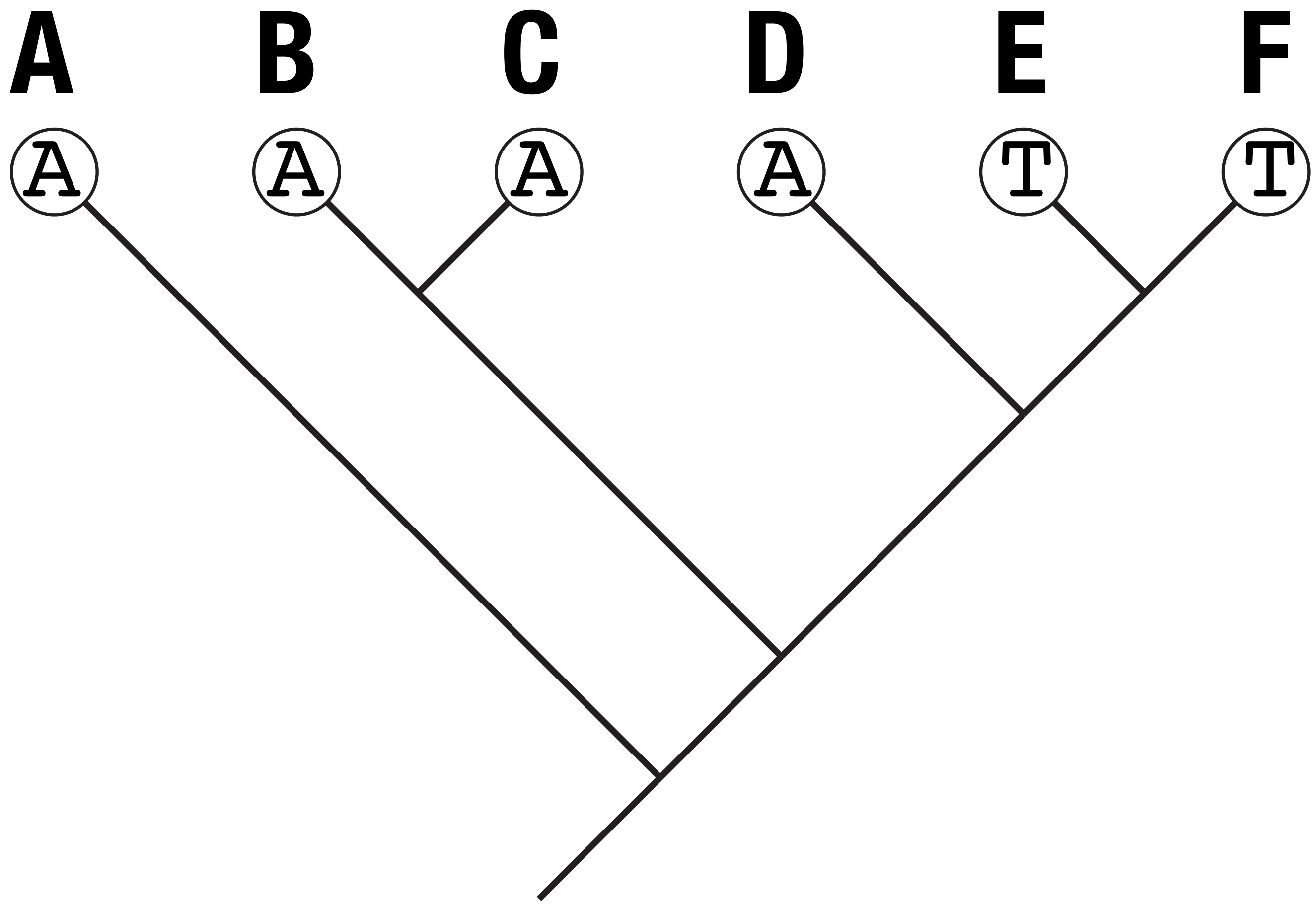
A table of character states. Each row corresponds to a taxon, and each column to a homologous character.

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 3



Phylogenetic Inference

In a typical phylogenetic study, we have a character matrix that describes the character states at the tips of the tree, and we want to know the phylogeny

Phylogenetic inference is the estimation of the phylogeny based on the character data and a model of character evolution

In the most basic phylogenetic inference projects, we are looking for the single "best" phylogeny (later we'll expand beyond this)

Analytic approaches

Neighbor joining - An algorithm that that constructs a tree from pairwise distances between species.

Advantages - Extremely fast

Limitations - While it works well on “clean” data, it performs poorly when there are unobserved changes and homoplasy

Optimization approaches

Evolutionary models allows us to define an optimality criterion that, for each tree, answers the question, “**How well does this tree account for the observed character?**”

We then measure the optimality criterion for many trees, and pick the tree with the best value

There are different optimality criteria. These include:

Maximum parsimony - The best tree is the tree that minimizes homoplasy, i.e. the simplest (most parsimonious) explanation

Maximum likelihood - The best tree is the tree that maximizes the likelihood of observing the character matrix

How many trees are there?

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

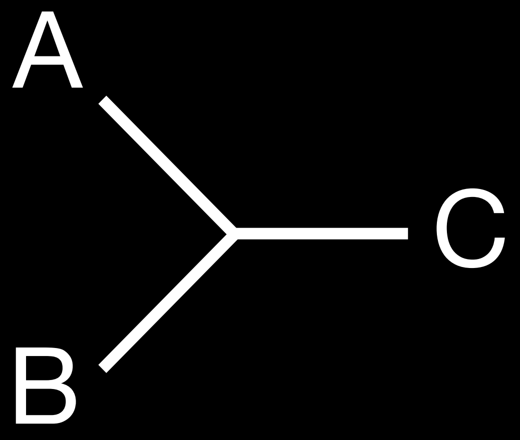
Where:

N is the number of trees

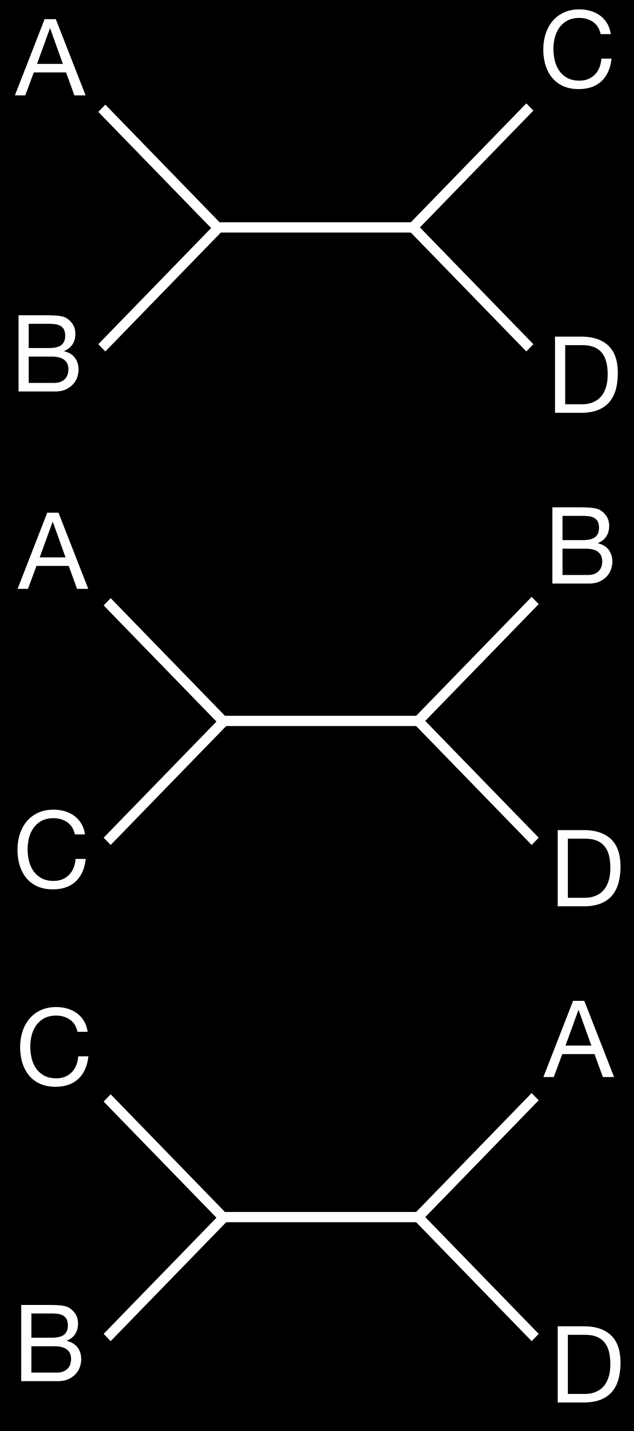
t is the number of taxa

How many trees are there?

$t = 3$, $N = 1$:



$t = 4$, $N = 3$:



How many trees are there?

t	*N*
3	1
4	3
5	15

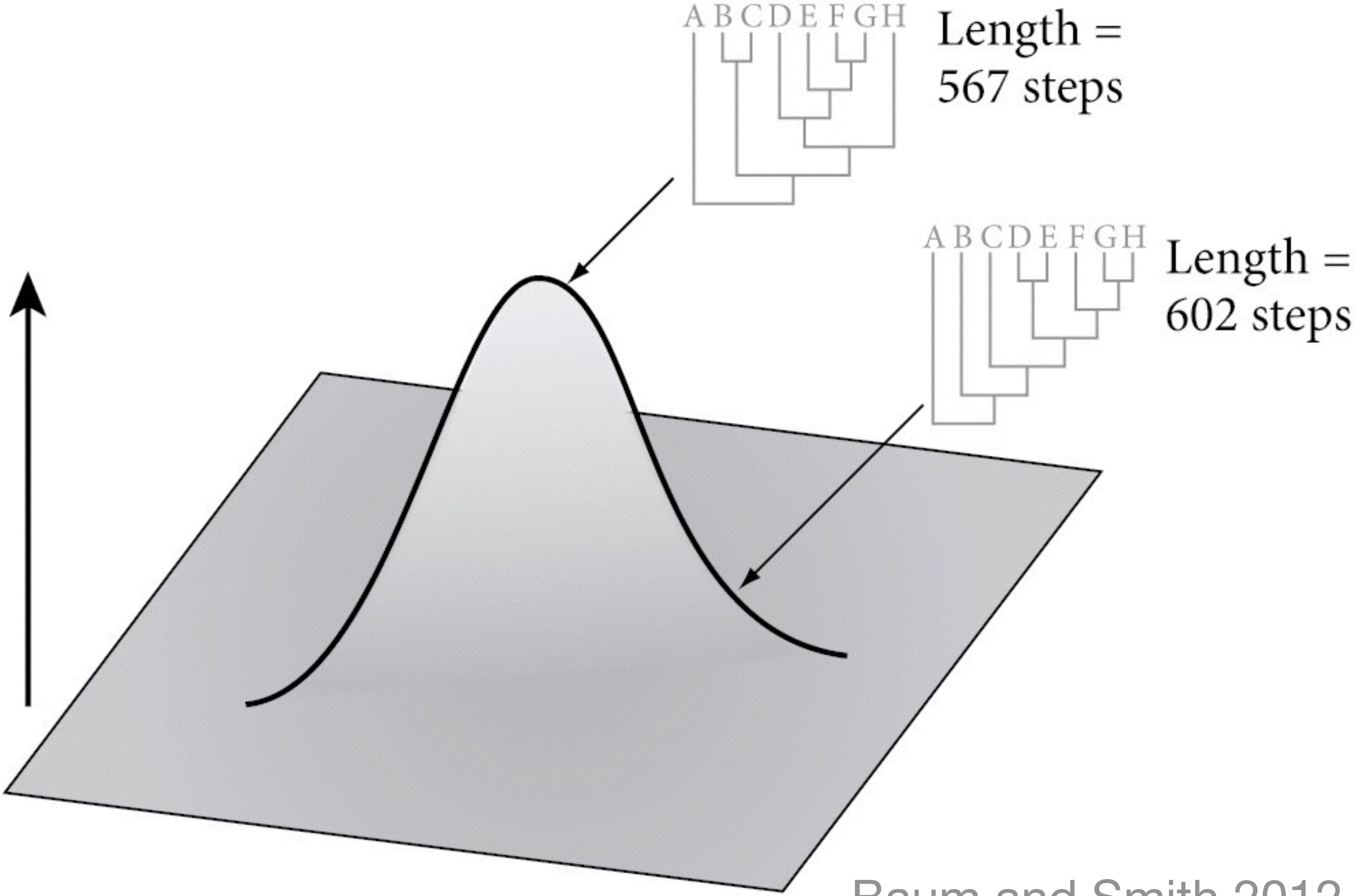
← 1,470,000- The number of pixels on this slide

← 6.02x10^{23}- Avagadro’s number (1 mole)

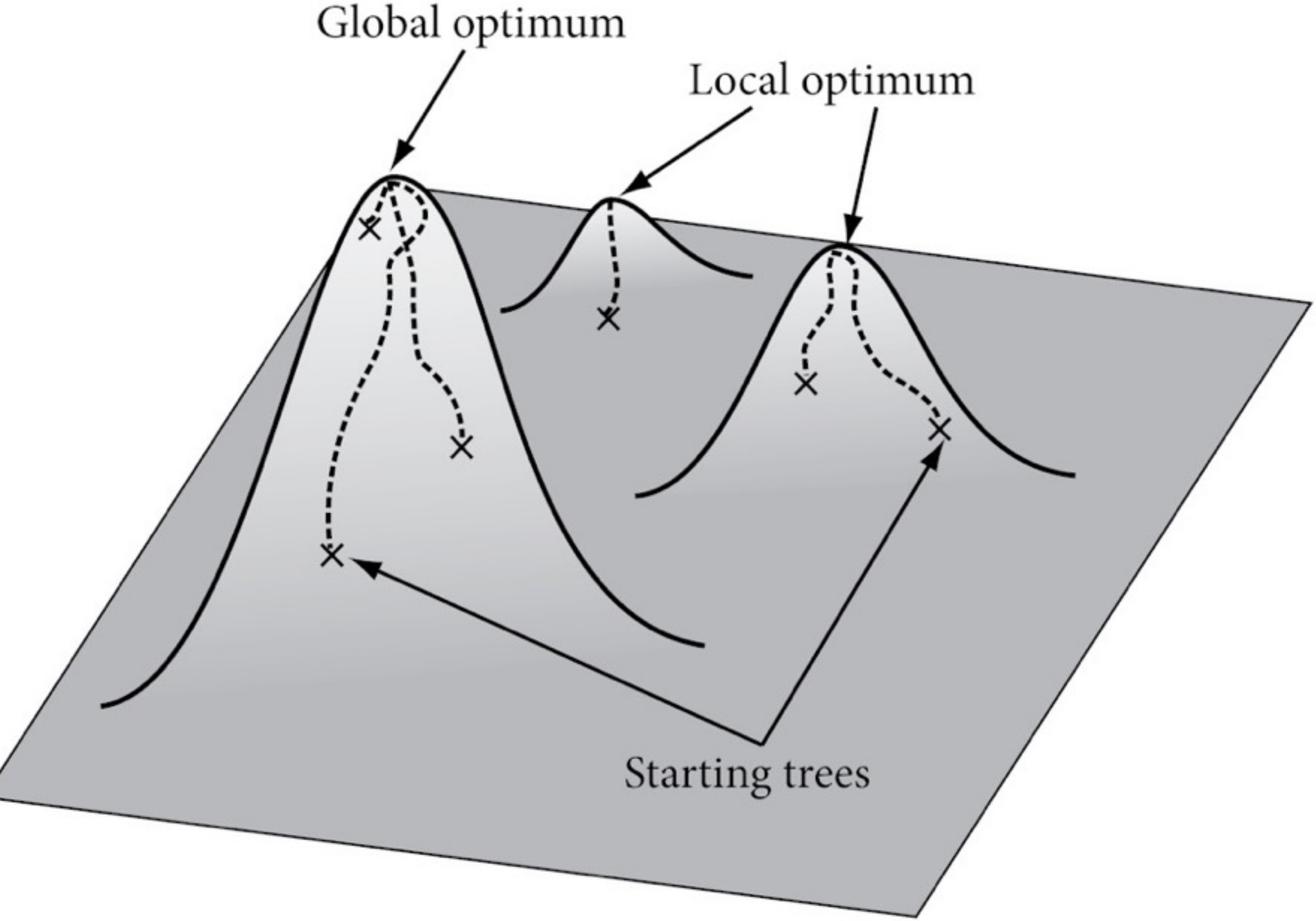
← 5x10^{27}- Molecules of gas in this room

← 7x10^{79}- Number of atoms in the observable universe

← 1x10^{100}- Googol



Baum and Smith 2012, Figure 7.8



Baum and Smith 2012, Figure 7.9

Maximum parsimony

Maximum parsimony

1. For each tree, find the minimum number of steps (character changes) needed to explain the character data.

2. Find the tree with the fewest steps

An example:

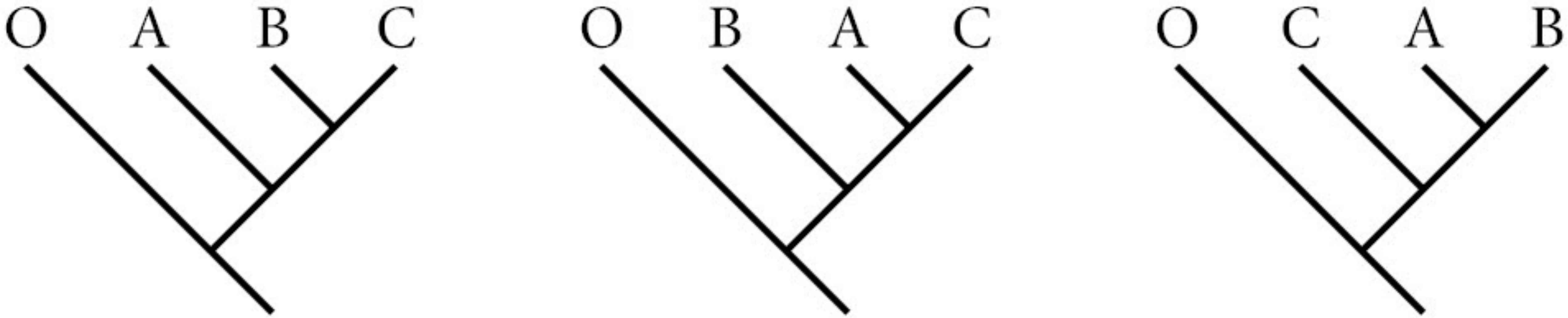
4 Taxa (O, A, B, C)

8 Characters

2 Character states for each character

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
O	0	0	0	0	0	0	0	0
A	0	1	0	0	0	1	1	0
B	1	1	0	1	1	1	1	1
C	0	0	1	1	0	0	0	0

Baum and Smith 2012, Table 7.5



Tree 1

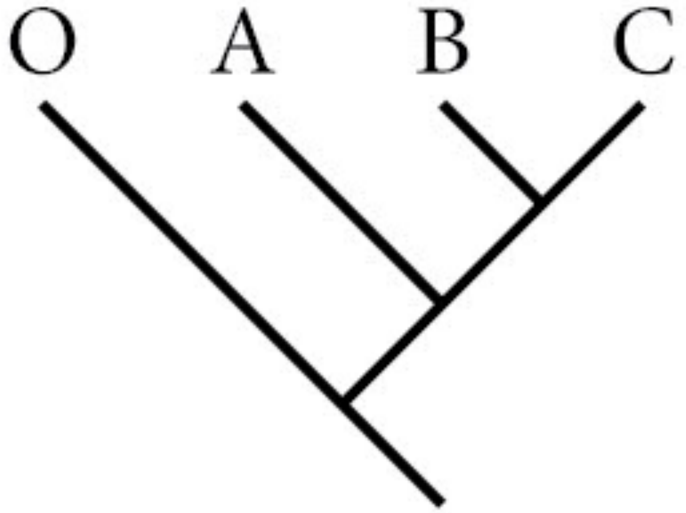
Tree 2

Tree 3

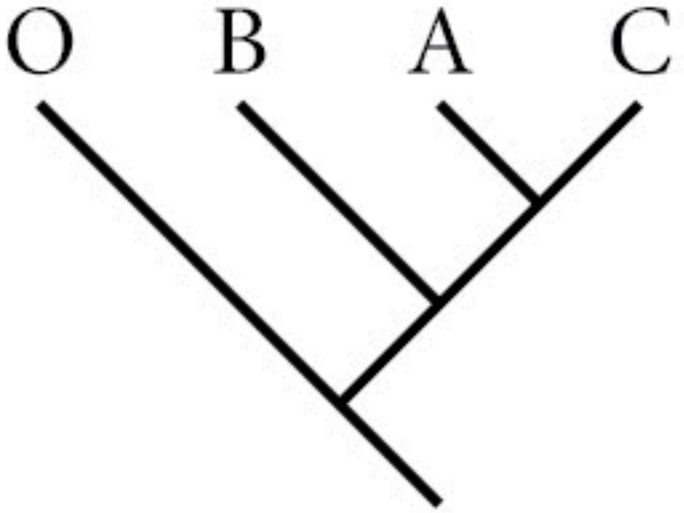
Baum and Smith 2012, Figure 7.3

	1	2	3	4	5	6	7	8
O	0	0	0	0	0	0	0	0
A	0	1	0	0	0	1	1	0
B	1	1	0	1	1	1	1	1
C	0	0	1	1	0	0	0	0

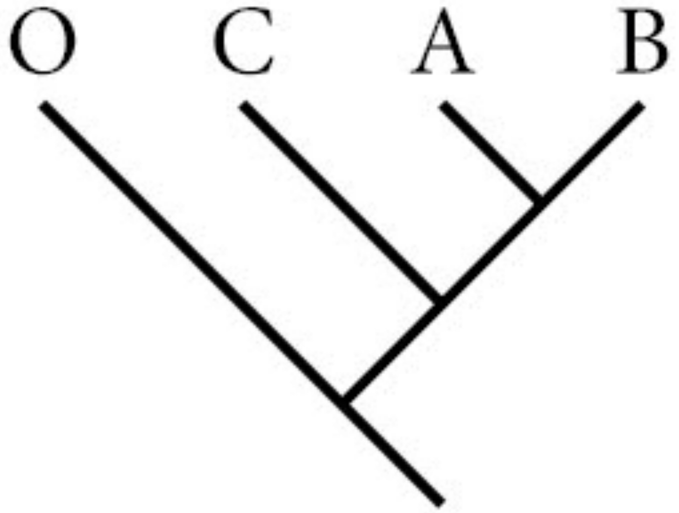
For each tree, what is the minimum number of changes needed to explain the character data?



Tree 1



Tree 2



Tree 3

Baum and Smith 2012,

Maximum parsimony

Advantages - Very fast

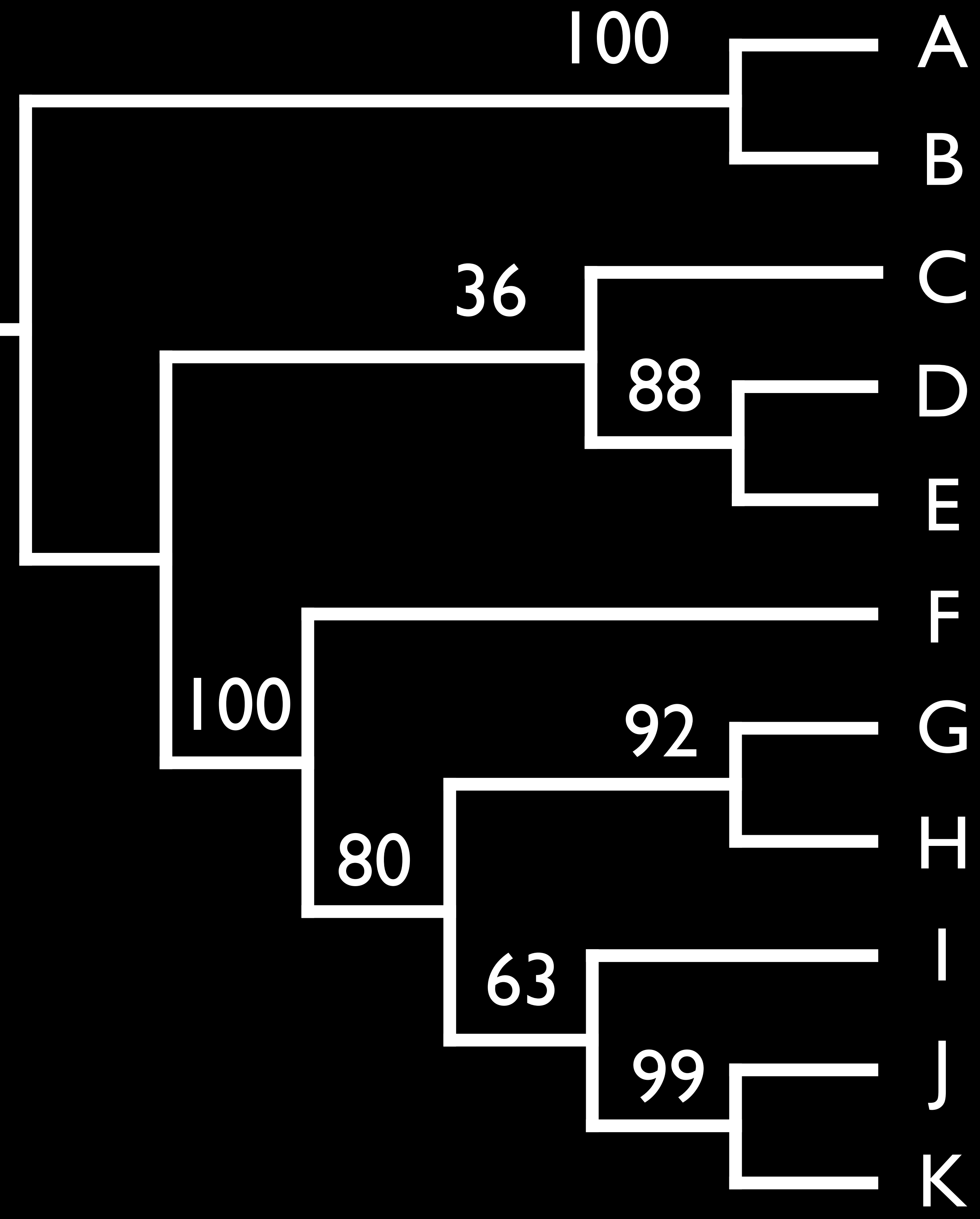
Limitations:

- Doesn’t account for expectation of more changes on longer branches
- Only considers scenario with the smallest number of changes, doesn’t take into account how many plausible historical scenarios could have produced tree.

(See Holder and Lewis 2003 for more)

Interpreting support

Confidence values



Presenting tree support:

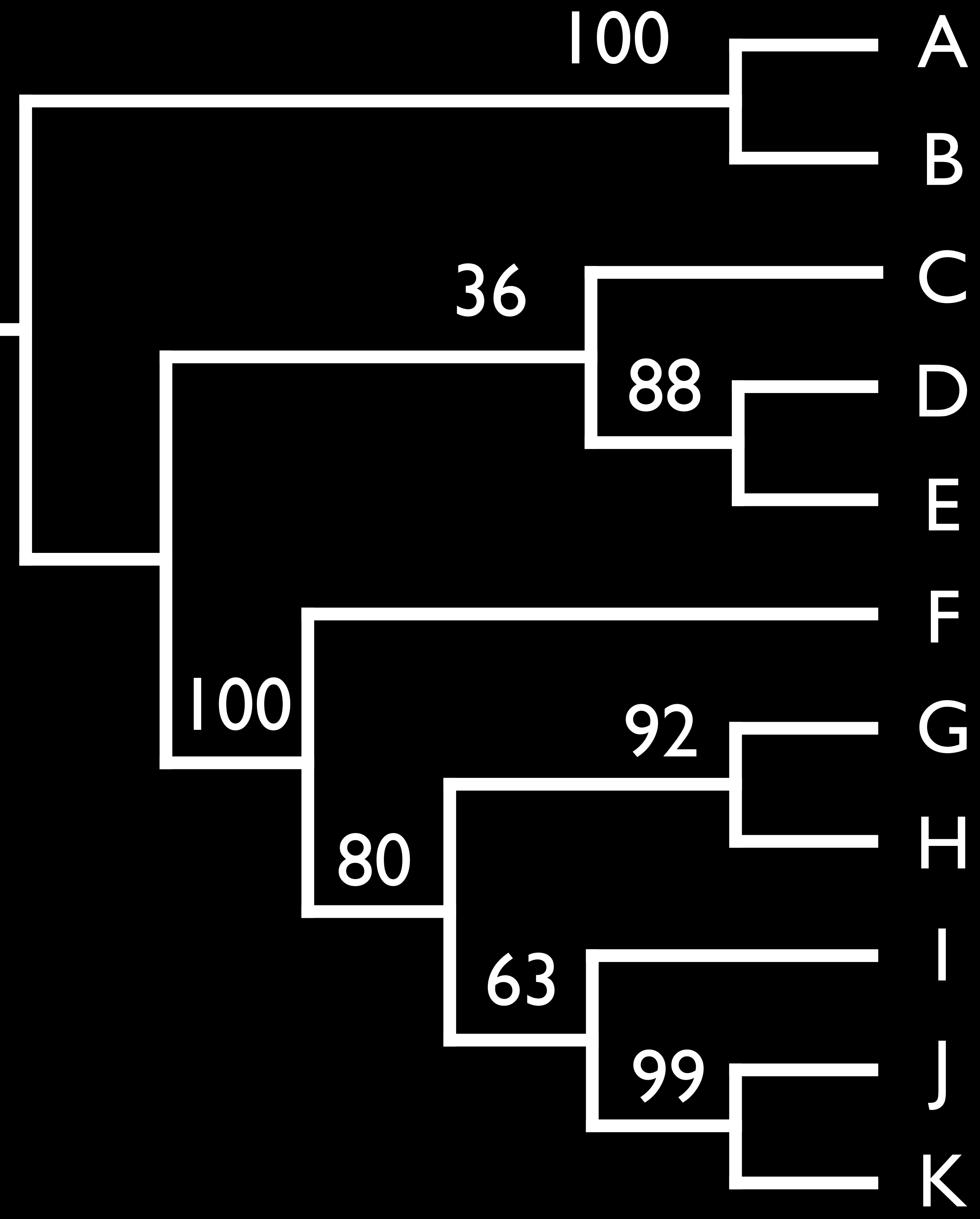
1. Get a set of trees with the same tips

2. Calculate the frequency of each bipartition in the tree set

3. Pick a display tree to show support on (could be best tree, consensus tree, etc)

4. For each bipartition in the display tree, write the observed frequency of the bipartition from the tree set

Confidence values - Bootstraps



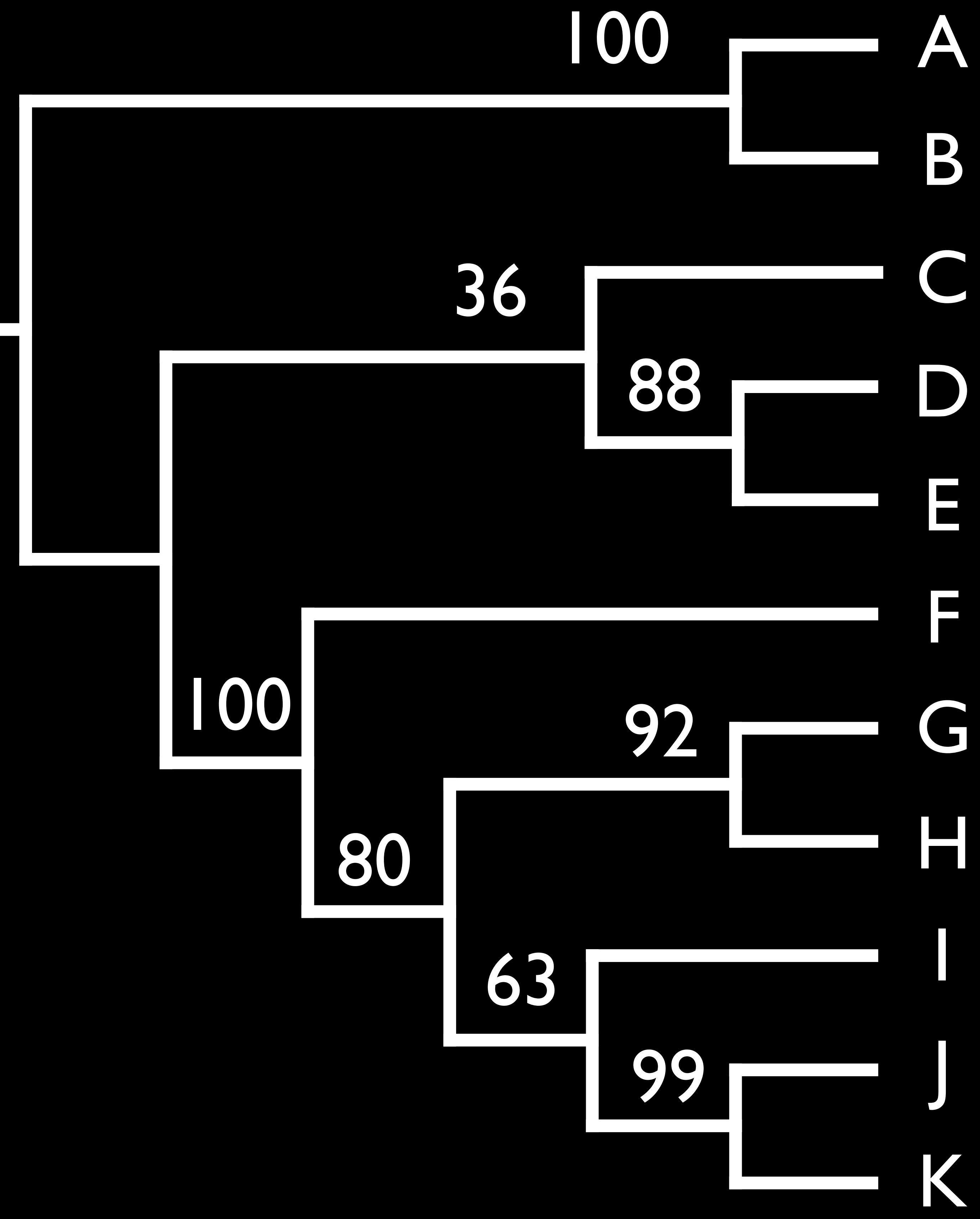
A relative index of how well supported clades are across characters

1. Resample from the real data to generate many pseudoreplicate datasets

2. Look for the best tree supported by each dataset

3. Measure how often each relationship is recovered when the resampled data are analyzed

Confidence values - Bayesian analyses



Bayesian support values indicate the probability of a relationship given the data

A few things to keep in mind about summarizing tree support:

1. Support refers to bipartitions, ie edges, not nodes. But support is often written on nodes and called “nodal support”. 😬

2. Much information can be lost when summarizing trees this way. Doesn’t show bipartition in tree set that conflict with display tree. These can have high support.

Other topics addressed

Rooted vs. unrooted trees

Outgroup/ ingroup