Many of the slides that I'll use have been borrowed from Dr. Paul Lewis, Dr. Joe Felsenstein. Thanks!

Paul has many great tools for teaching phylogenetics at his web site:

http://hydrodictyon.eeb.uconn.edu/people/plewis

The main subject of this course: estimating a tree from character data

Tree construction:

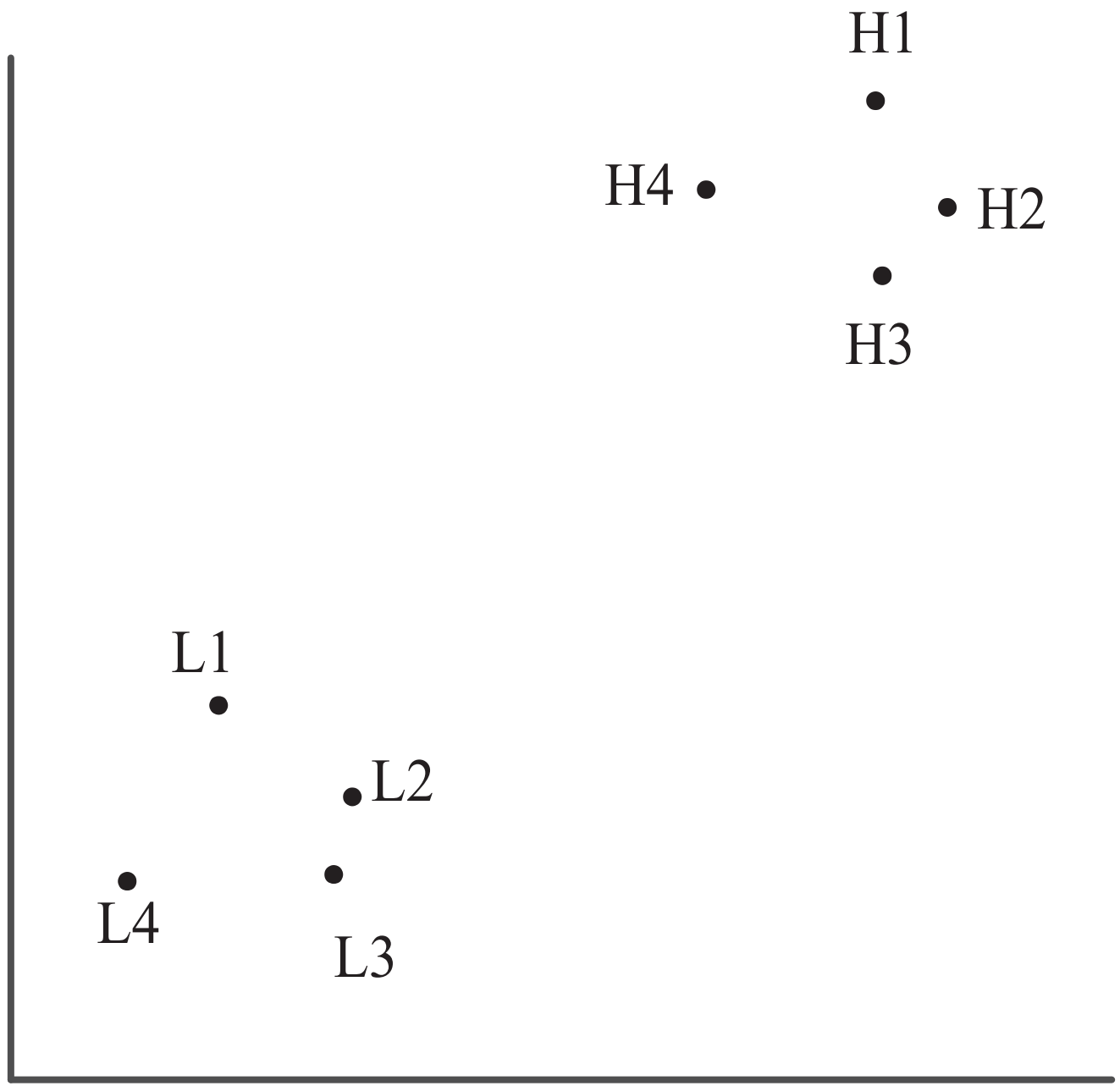
- strictly algorithmic approaches - use a "recipe" to construct a tree
- optimality based approaches - choose a way to "score" a trees and then search for the tree that has the best score.

Expressing support for aspects of the tree:

- bootstrapping,
- testing competing trees against each other,
- posterior probabilities (in Bayesian approaches).

Simple test of Bergmann's rule: comparing latitude and mass (I made these data up)
lat. offset = degrees north of the 49th parallel.

species	lat. offset	mass
L1	3.1	5.9
L2	5.4	4.3
L3	5.1	3.1
L4	1.8	3.6
H1	13.5	15.2
H2	14.6	13.5
H3	13.6	12.4
H4	10.8	13.7

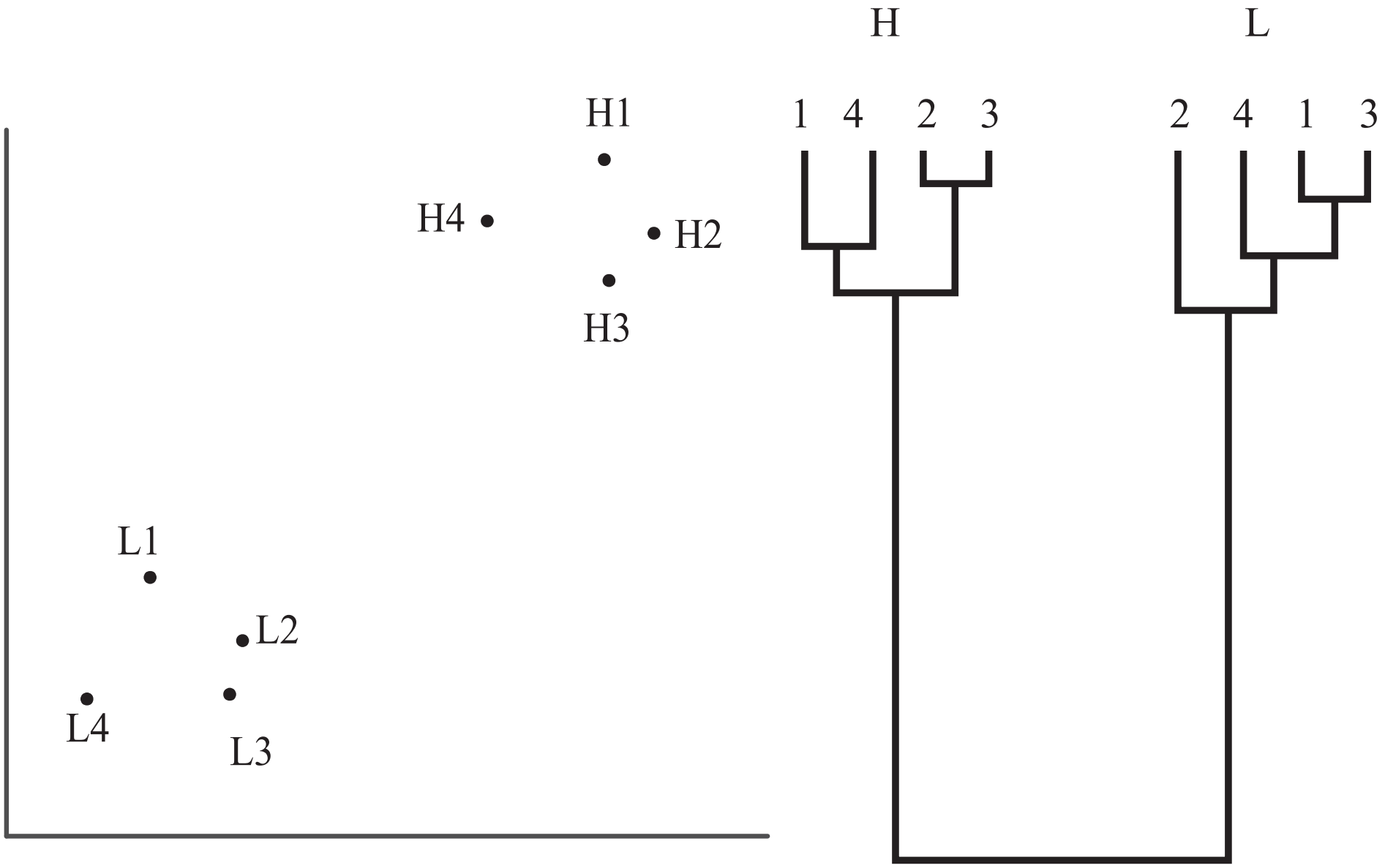

H1
H4
H2
H3
L1
L2
L4
L3

(cue cartoon videos)

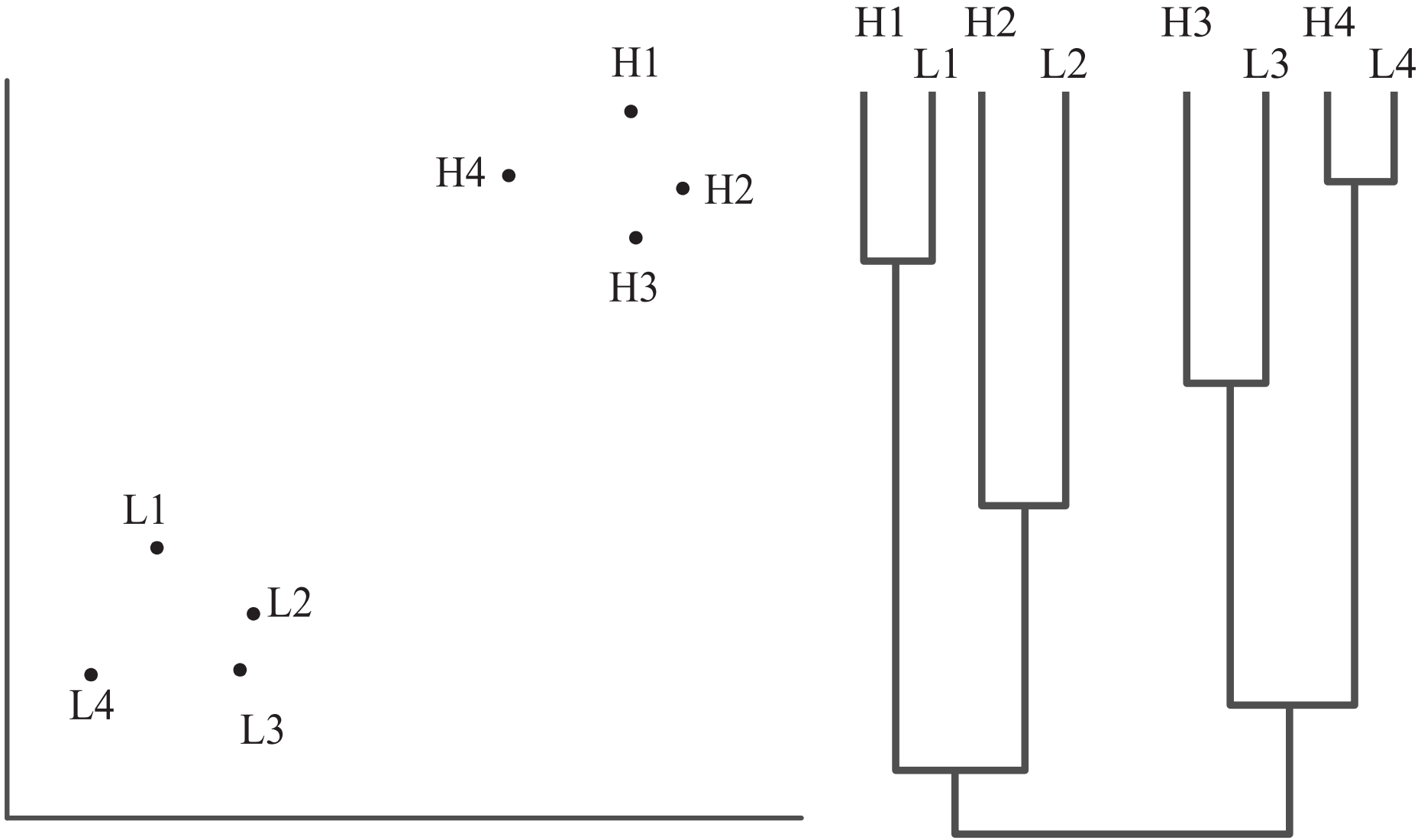
See http://phylo.bio.ku.edu/slides/no-correl-anim.mov

and http://phylo.bio.ku.edu/slides/correl-anim2.mov

No (or little) evidence for correlation



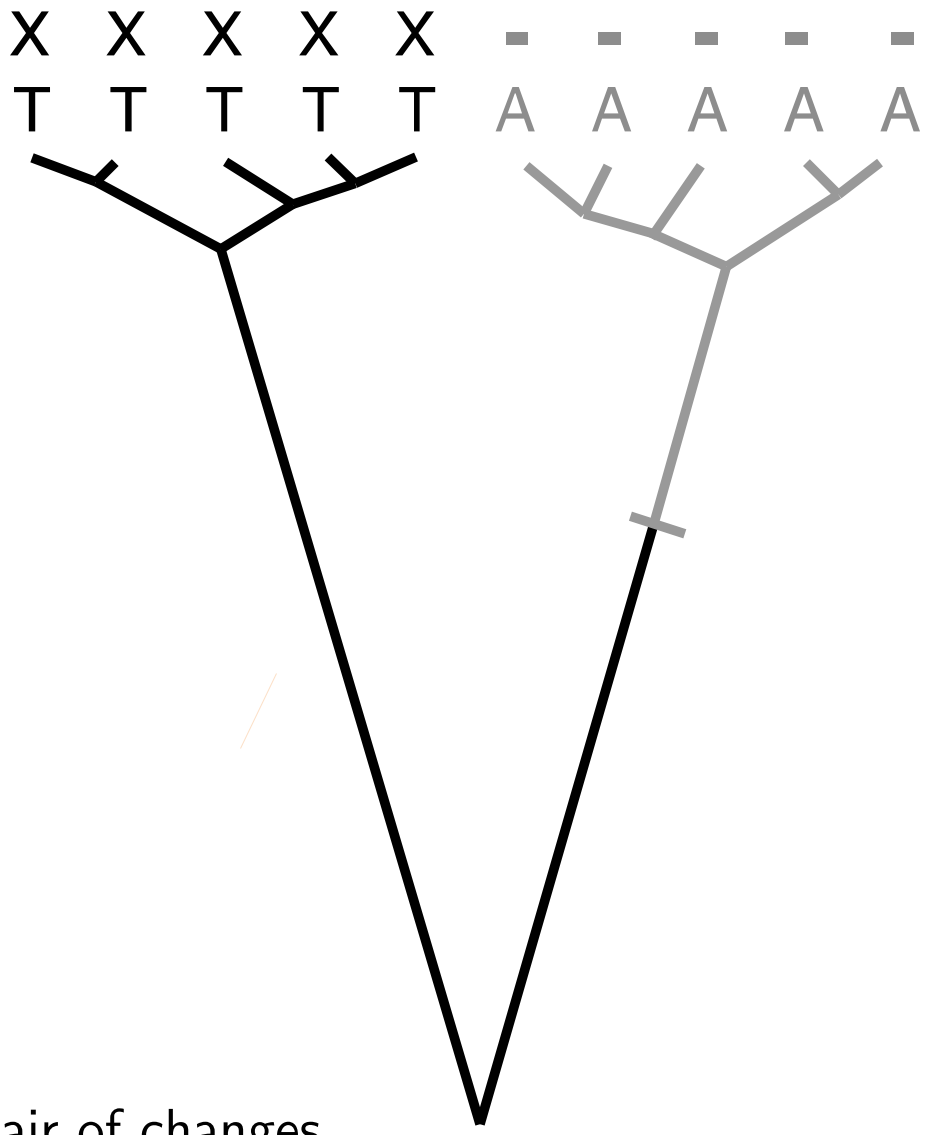
Evidence for correlation



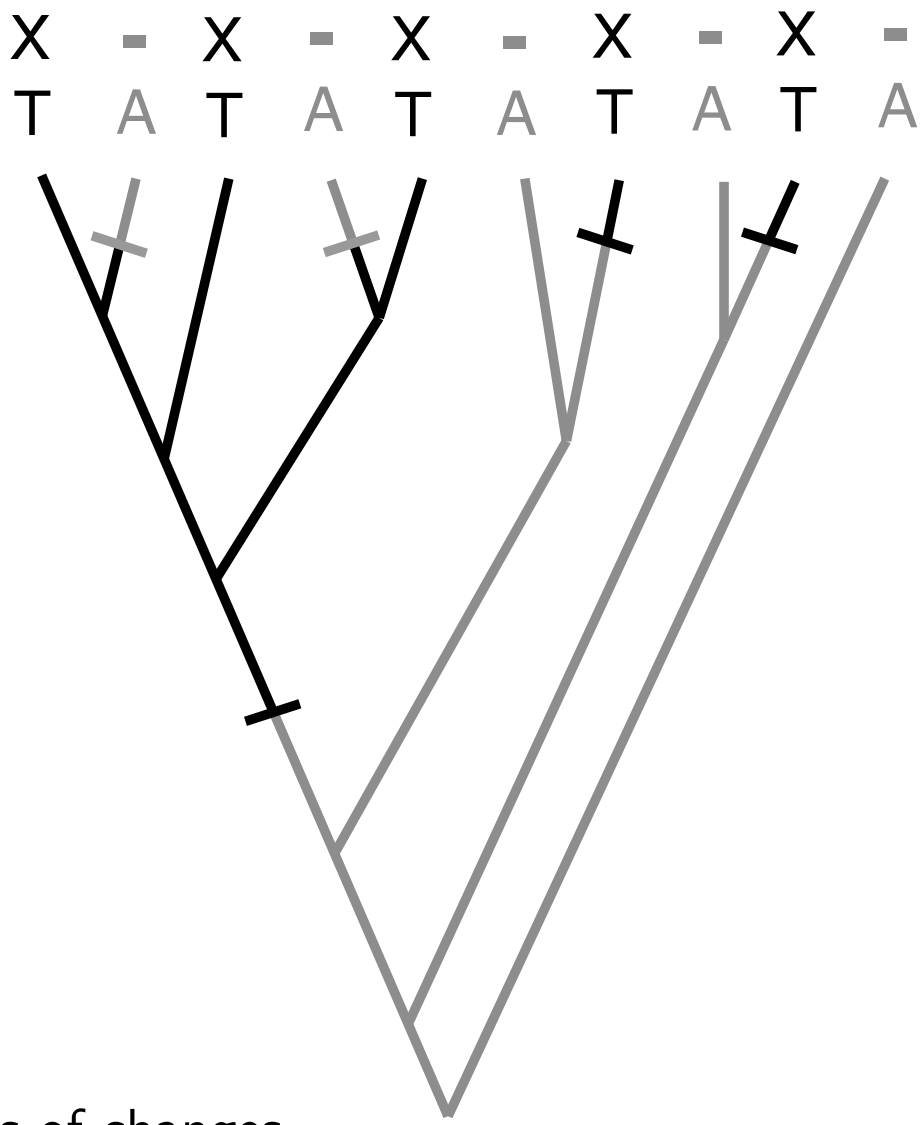
Do desert green algae use xanthophyll to protect against excessive light intensities?

Species	Habitat	Photoprotection
1	terrestrial	xanthophyll
2	terrestrial	xanthophyll
3	terrestrial	xanthophyll
4	terrestrial	xanthophyll
5	terrestrial	xanthophyll
6	aquatic	none
7	aquatic	none
8	aquatic	none
9	aquatic	none
10	aquatic	none

Phylogeny reveals the events that generate the pattern



1 pair of changes.
Coincidence?



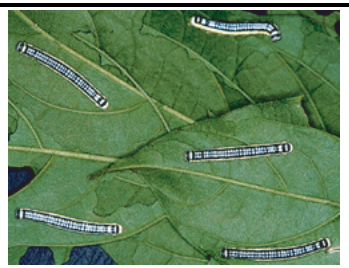
5 pairs of changes.
Much more convincing

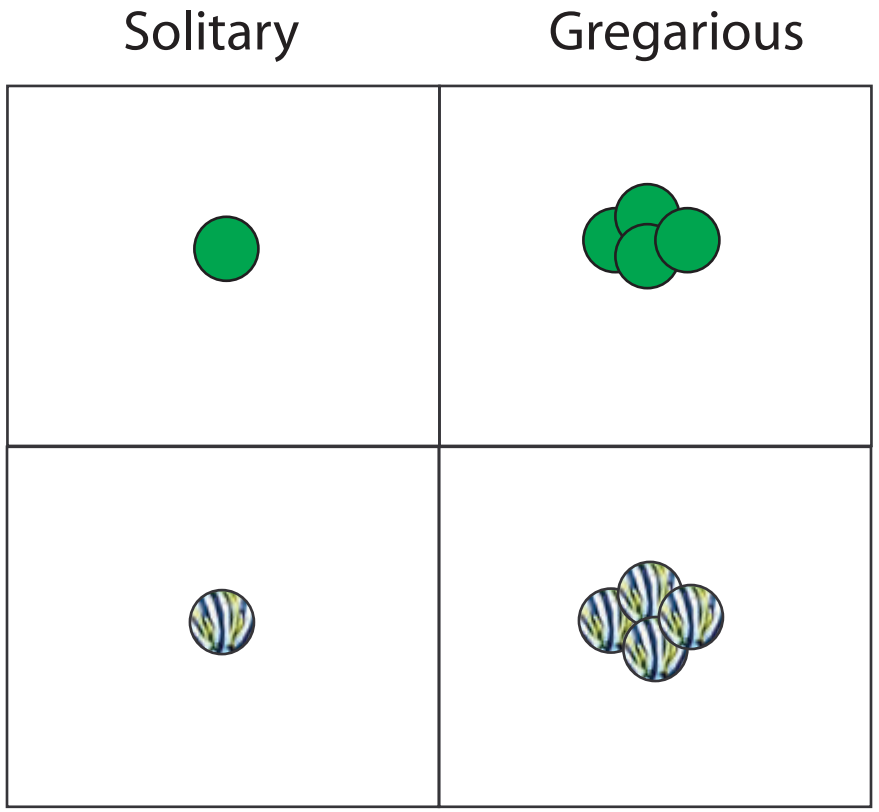
Inferring Process from Pattern

Hypothesis:

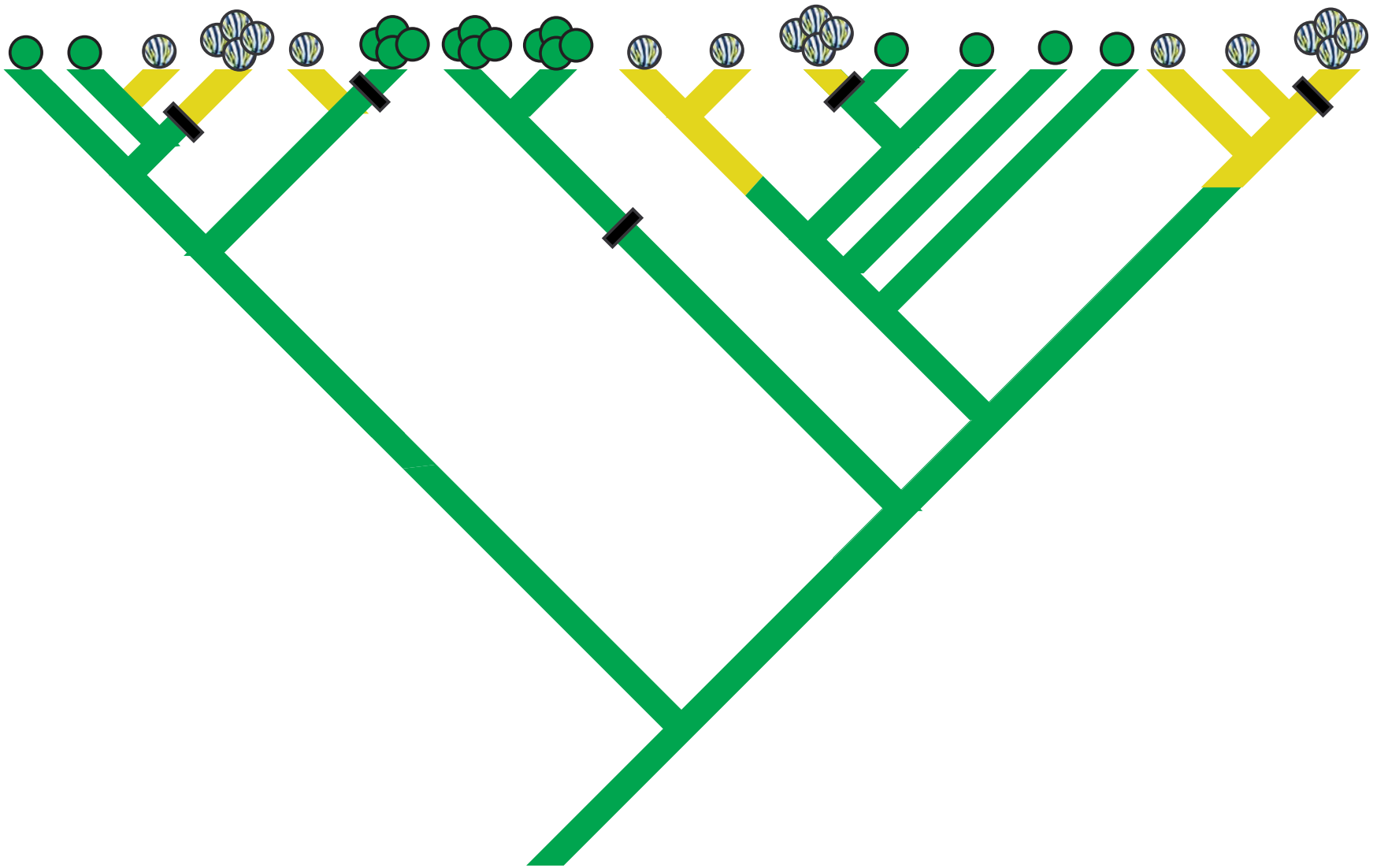
Gregariousness should arise more frequently in unpalatable organisms than in tasty ones (Sillén-Tullberg, 1988)

Inferring Process from Pattern



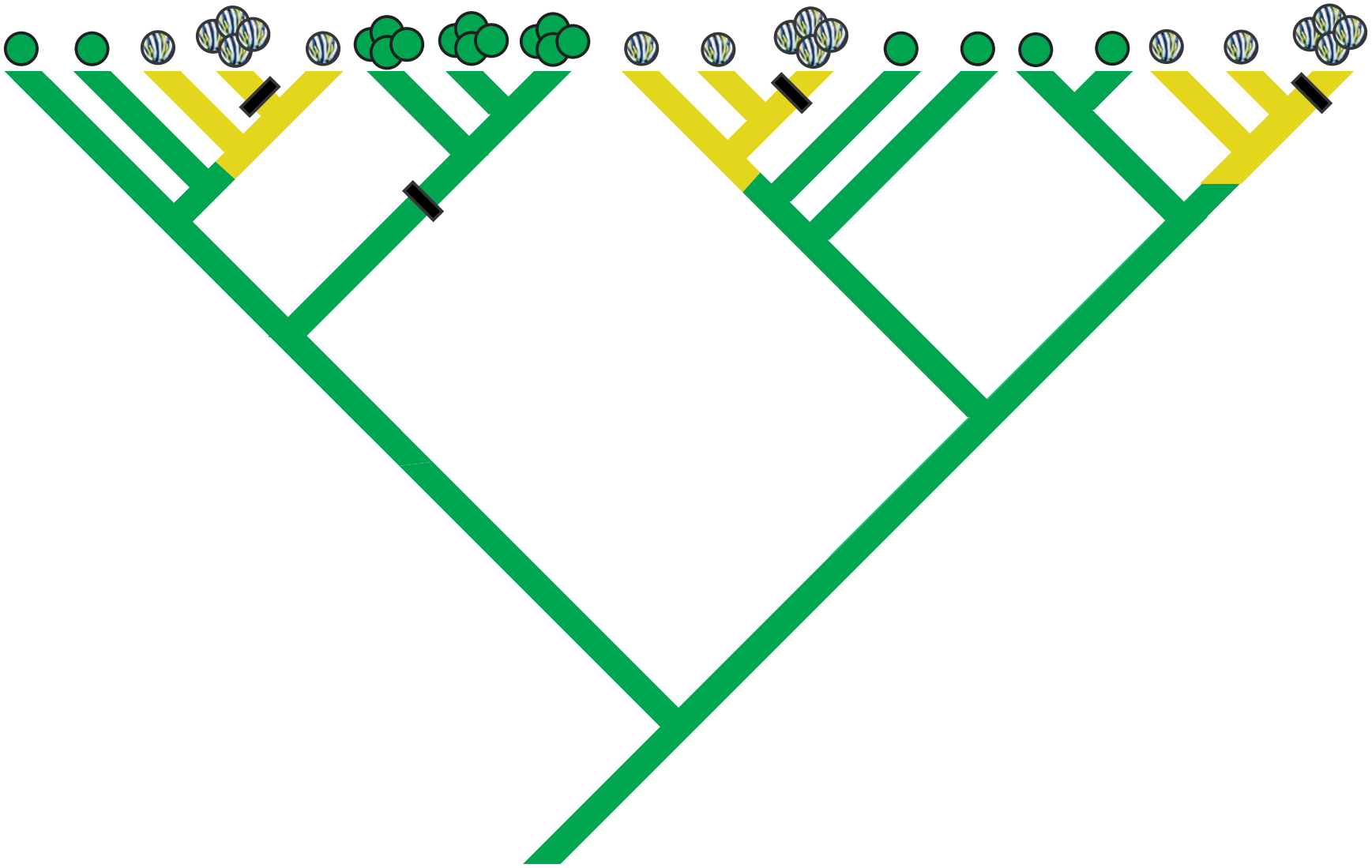


Sillén-Tullberg (1988), Dyer and Gentry (2002), Hill (2001)



One possible outcome:

No clear evidence of associations between traits



Cartoon of the real results (Sillén-Tullberg, 1988)

Aposematic species are more likely to evolve gregarious larvae

Importance of phylogeny

The previous slides had identical patterns of traits if the phylogeny is ignored.

Without knowledge of the tree, no conclusion would be reached.

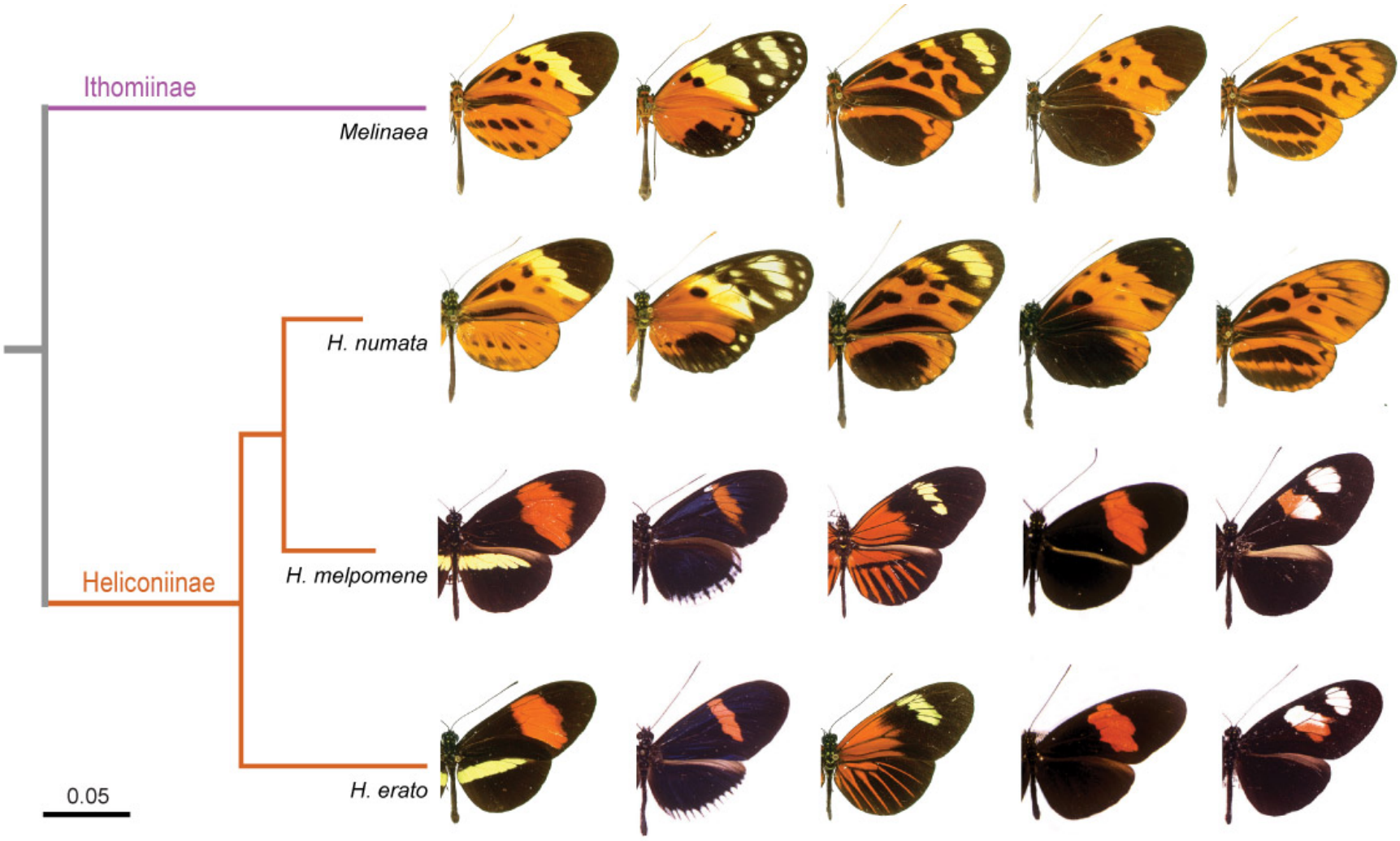


Figure by Mathieu Joron: http://xyala.cap.ed.ac.uk/joron/

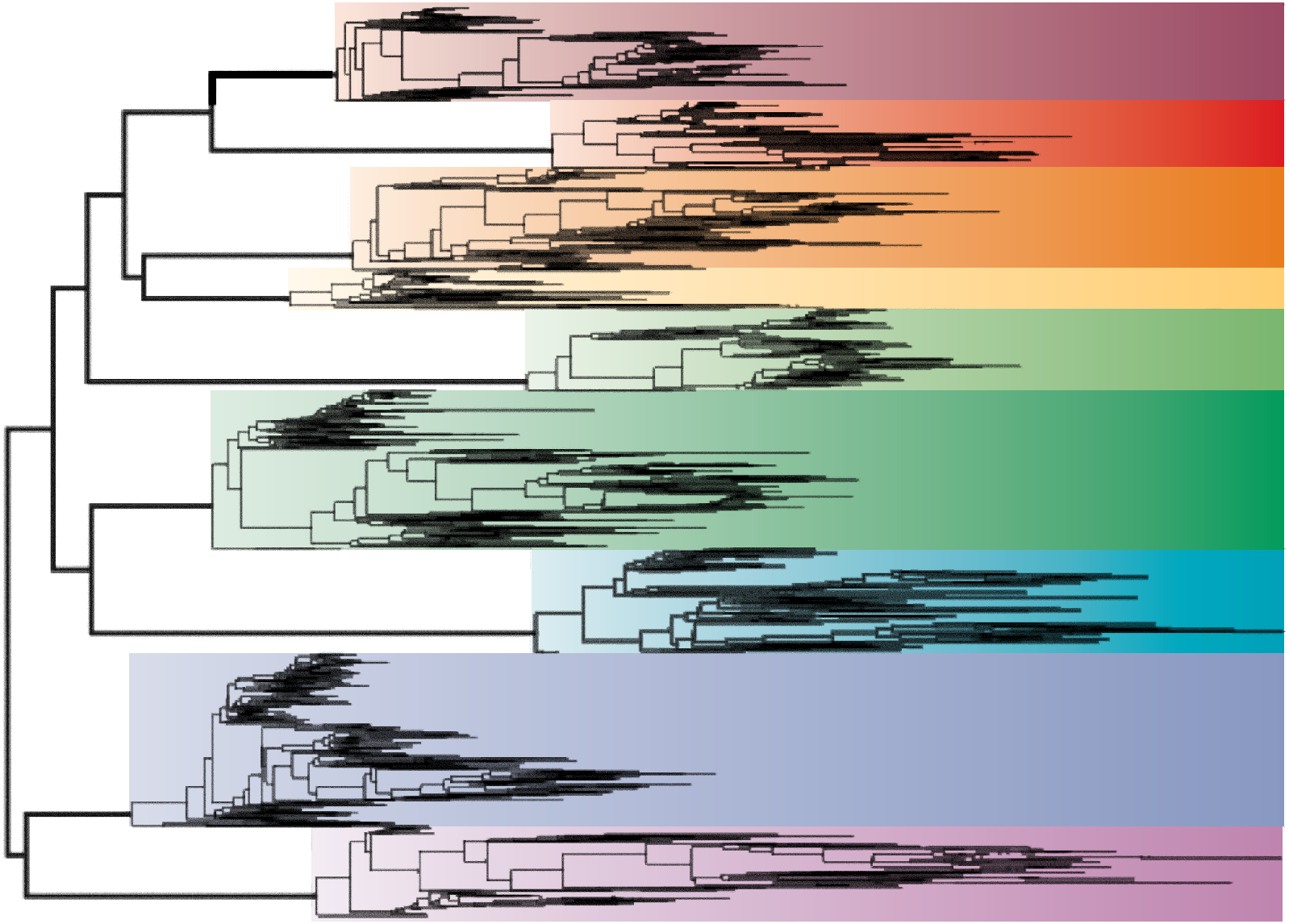


Figure from Rambaut, Posada, Crandall, and Holmes
Nature Reviews Genetics, 2004

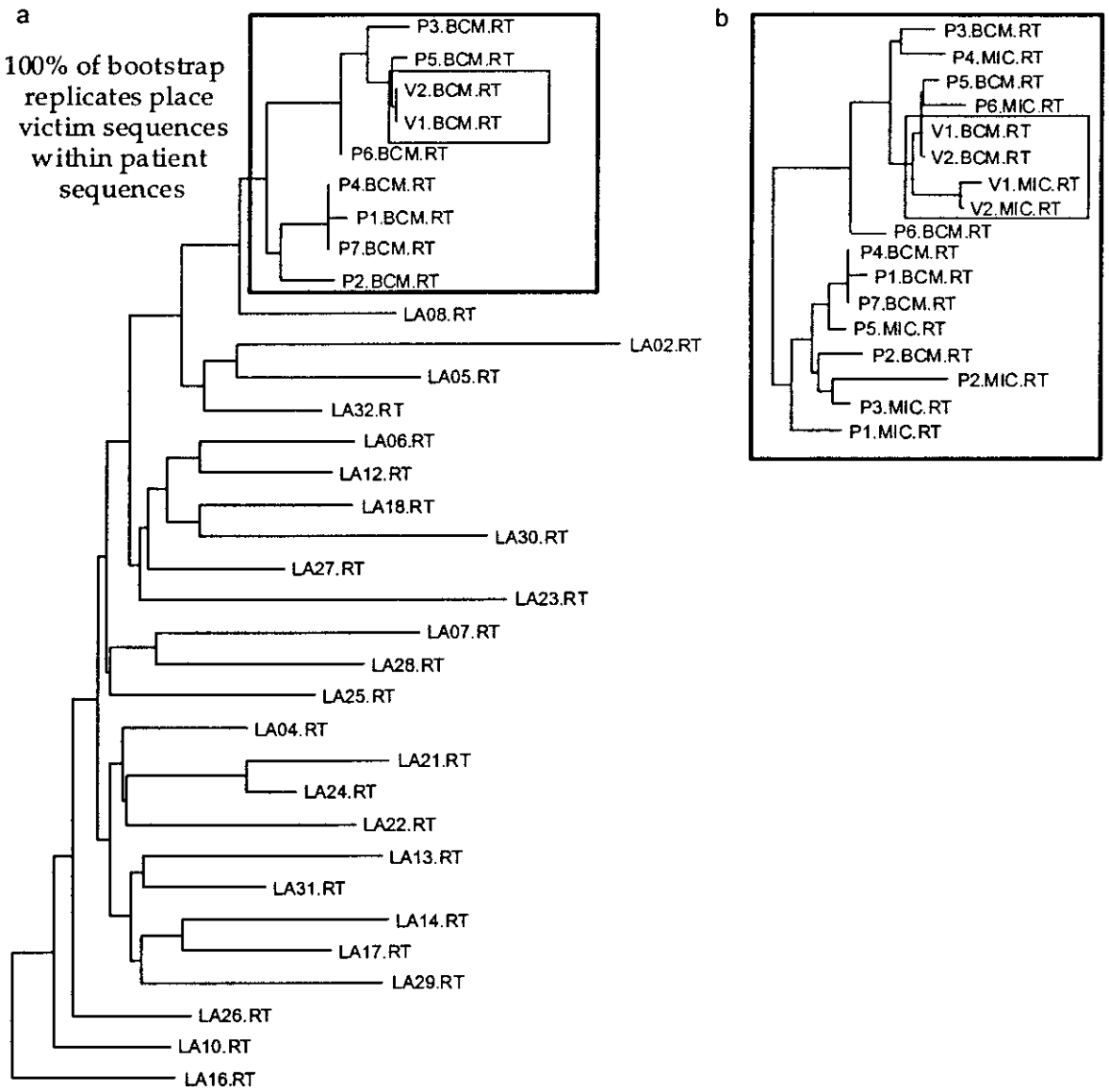
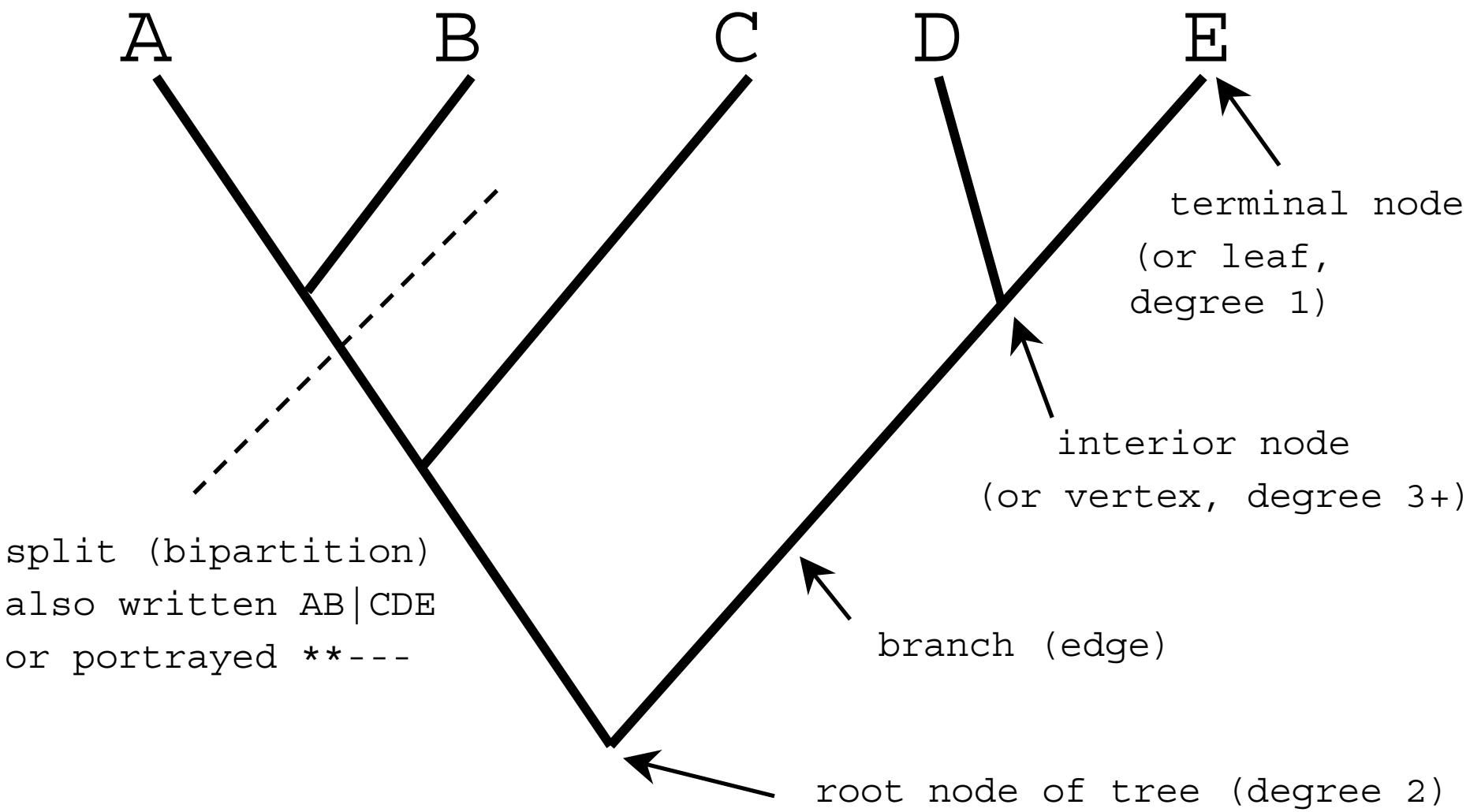
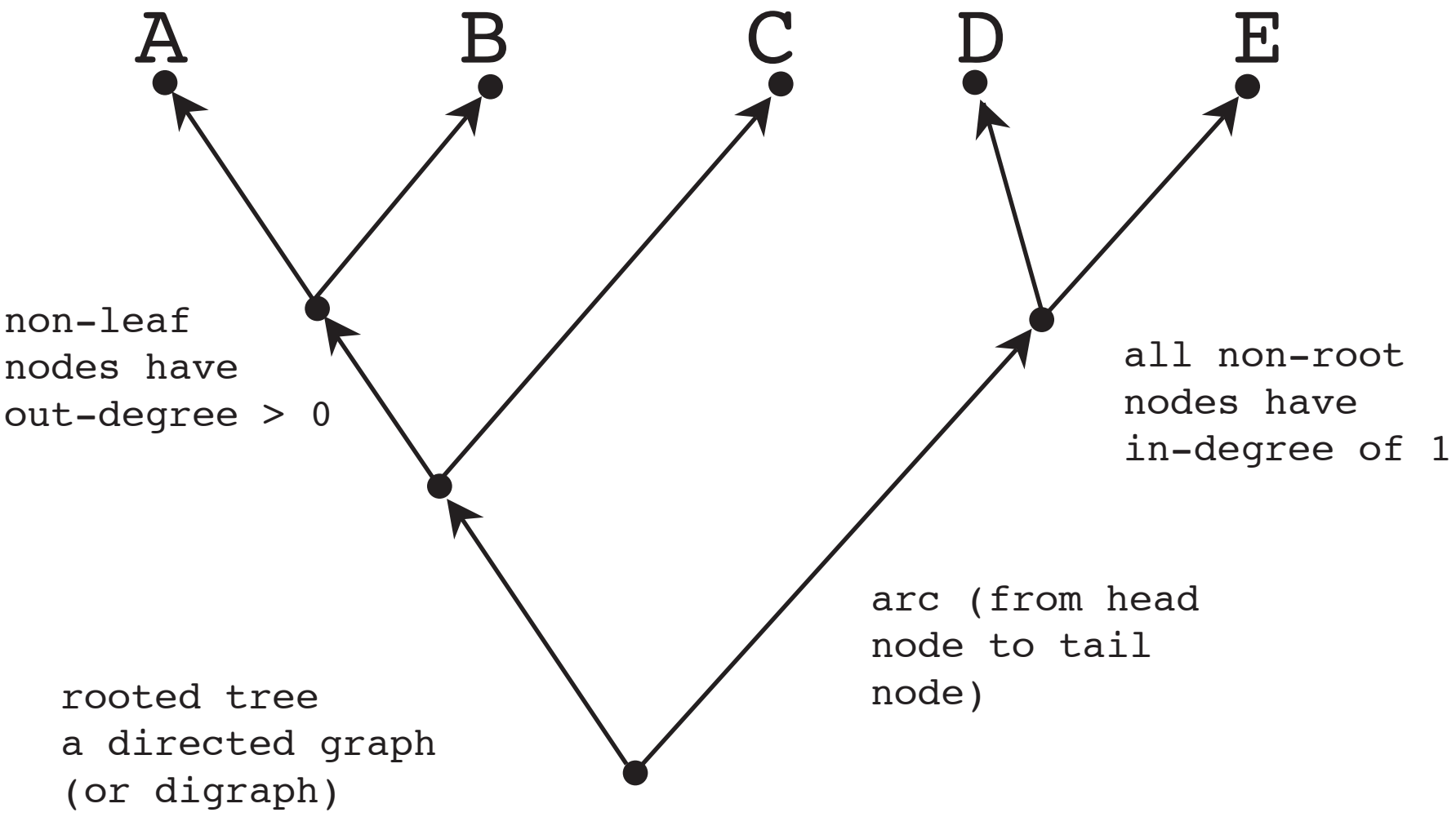


Figure from Metzker et al. (2002), 2004

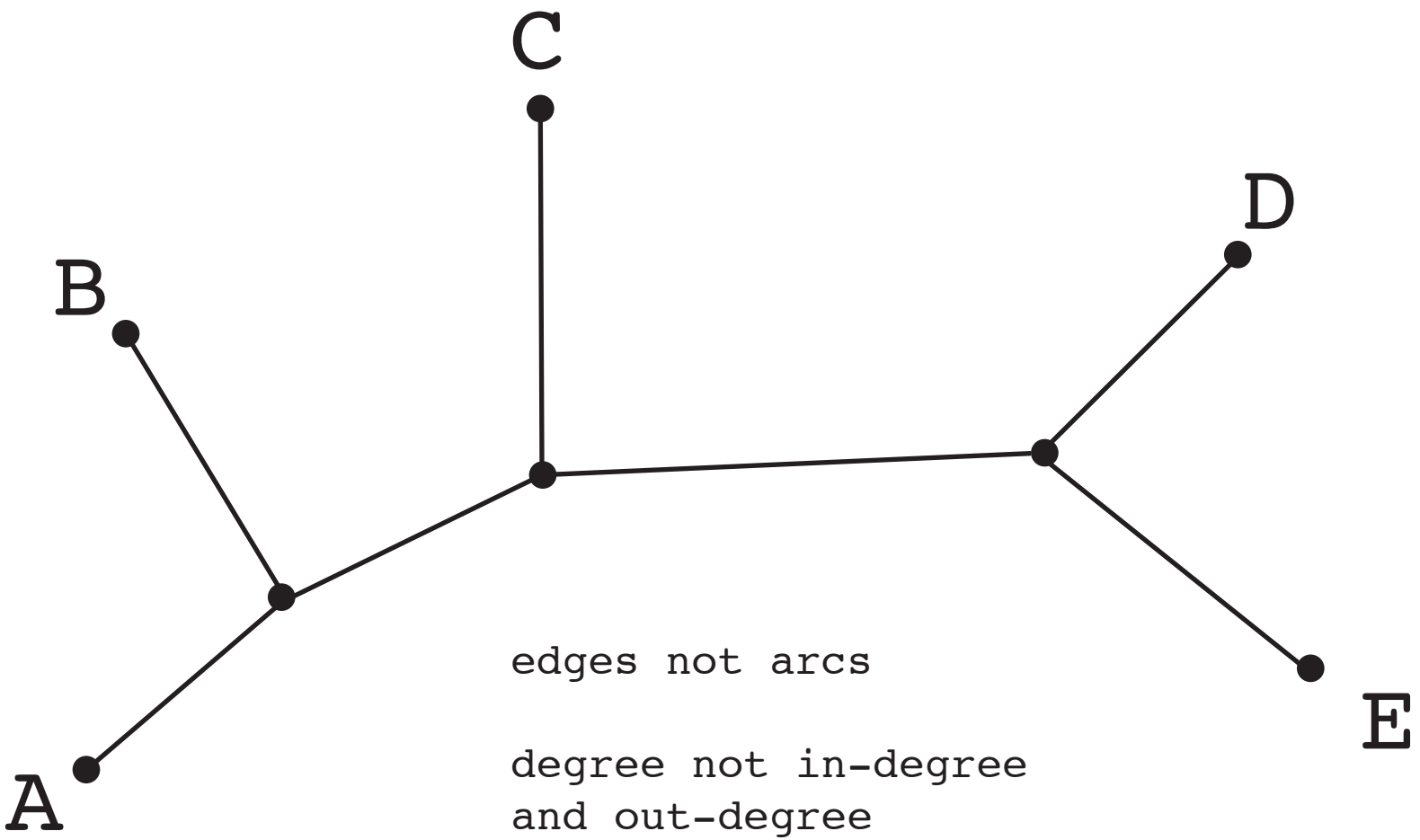
Tree terminology



Rooted tree terminology



Rooted tree terminology



Tree terms

A tree is a connected, acyclic graph.

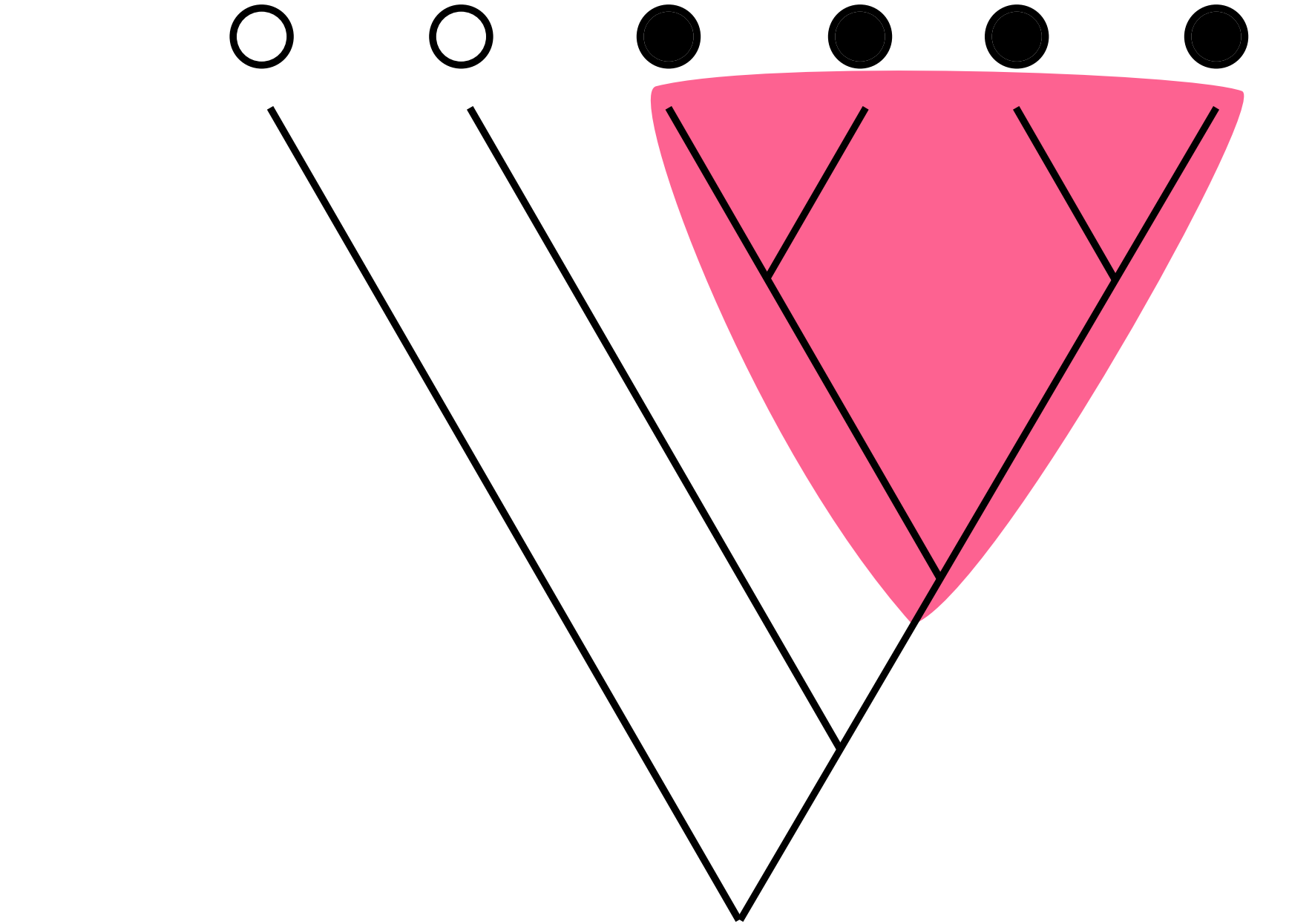
A rooted tree is a connected, acyclic directed graph.

A polytomy or multifurcation is a node with a degree > 3 (in an unrooted tree), or a node with an out-degree > 2 (in a rooted tree).

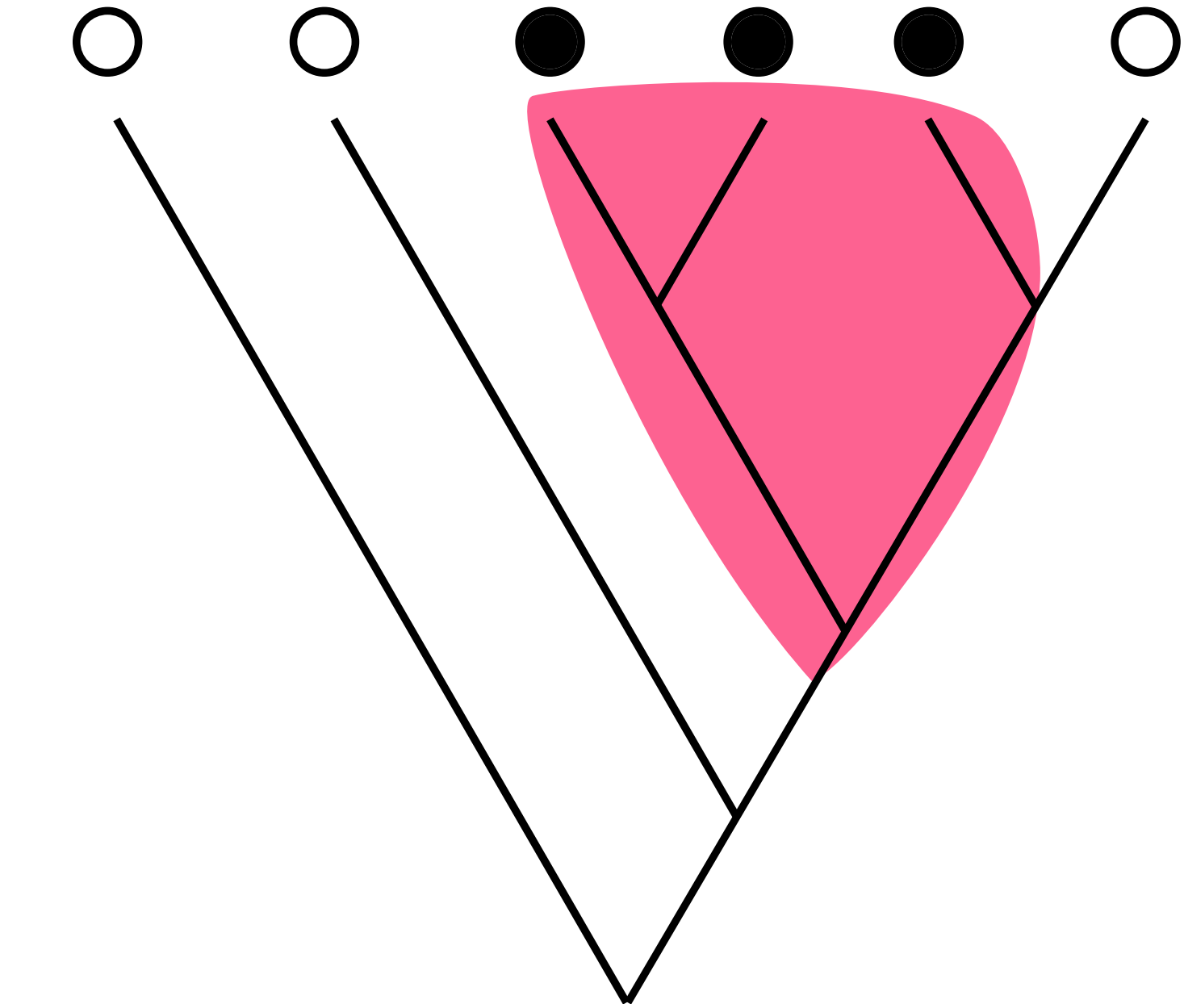
Collapsing an edge means to merge the nodes at the end of the branch (resulting in a polytomy in most cases).

Refining a polytomy means to "break" the node into two nodes that are connected by an edge.

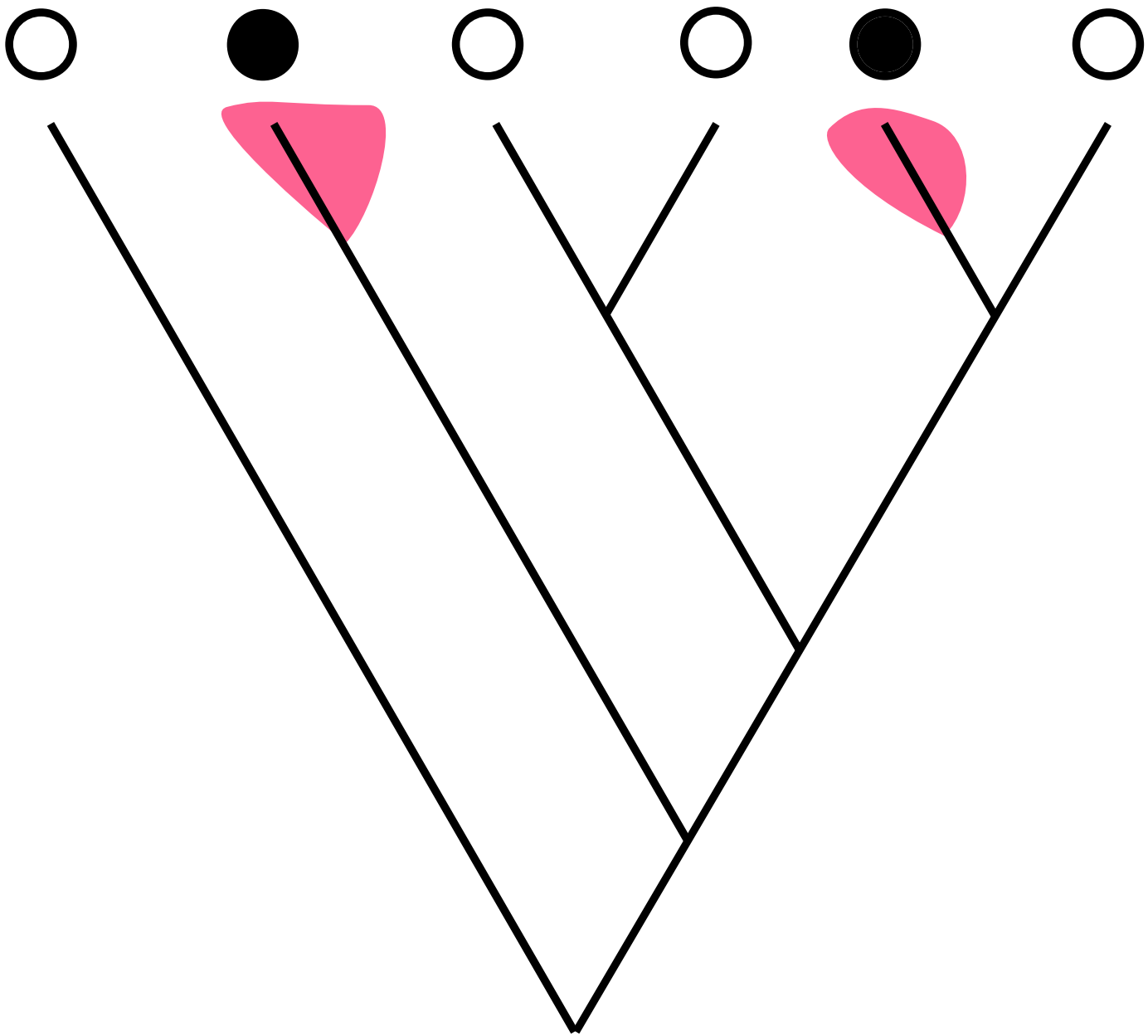
Monophyletic groups ("clades"): the basis of phylogenetic classification



Paraphyletic groups: error of omitting some species



Polyphyletic groups: error of grouping "unrelated" species



Homework #1 – (due Monday, Aug 28th)

Draw an unrooted tree from the table of splits shown on the next page. The frequencies shown in the table represent bootstrap proportions. We'll cover bootstrapping later in the course – for now you can treat the "Freq" column as label for the branches.

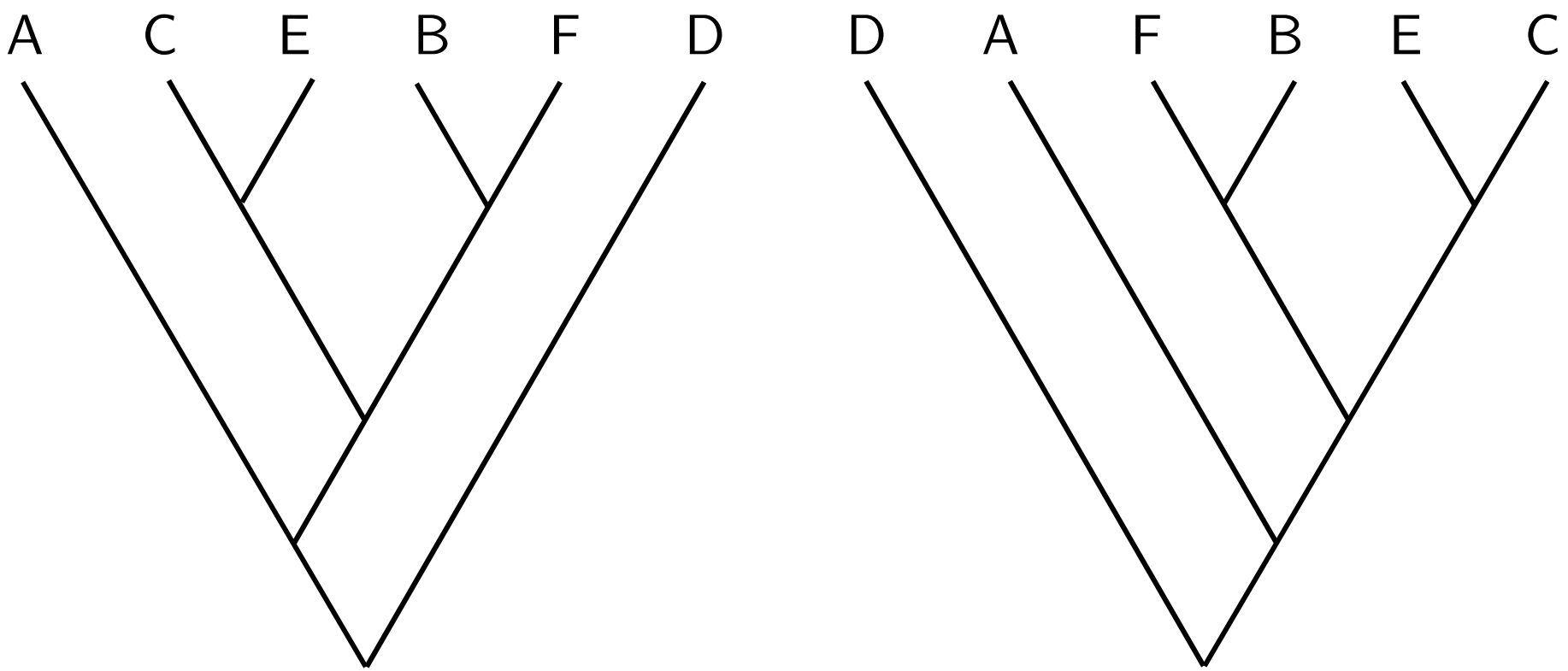
Start at the first row and add splits until you cannot add any more splits to the tree.

Make sure to label the leaves of the tree with the taxon number and the edges with the value found in the "Freq" column.

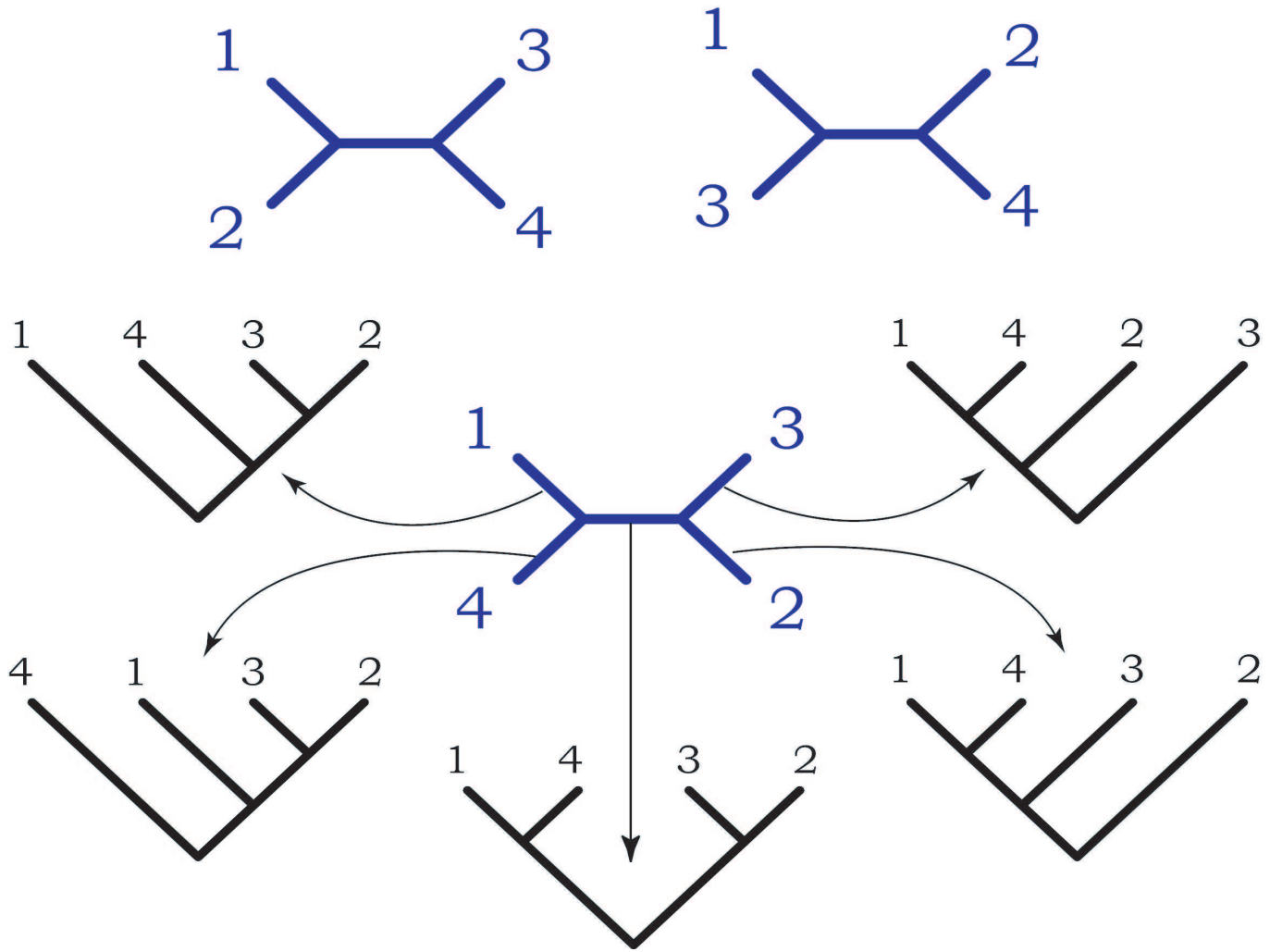
000000000111111	
123456789012345	Freq
..........*.*.*	100
.........**.....	99
.**..........*.	97
.........***.*.*	94
......*....*...	78
...**********.*	67
.**............	61
......*.*****.*	60
..........*...*	56
...*.*.........	41
..........*.*..	39
..*..........*.	37
.....********.*	33

/end-of-homework

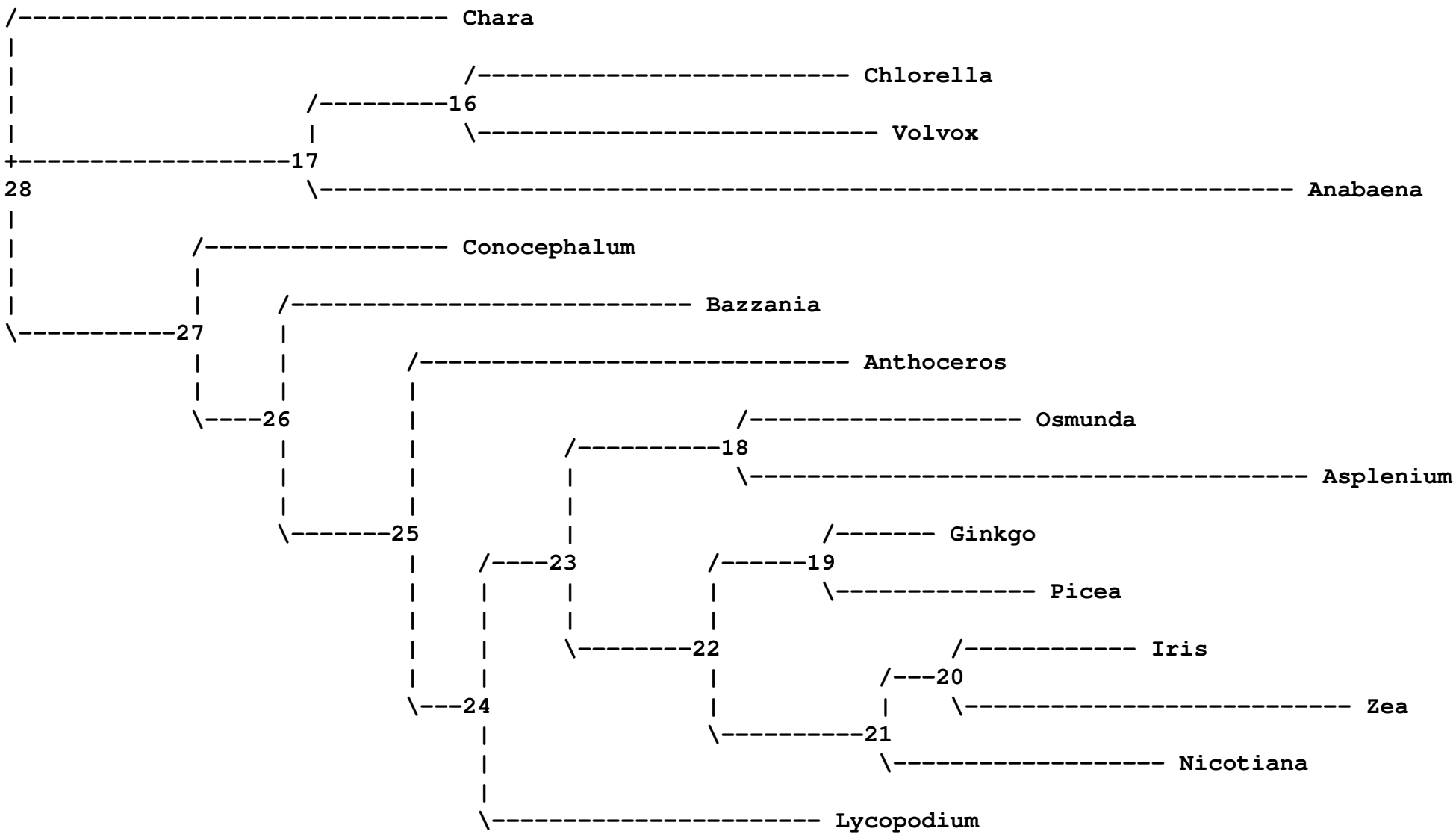
Branch rotation does not matter



Rooted vs unrooted trees



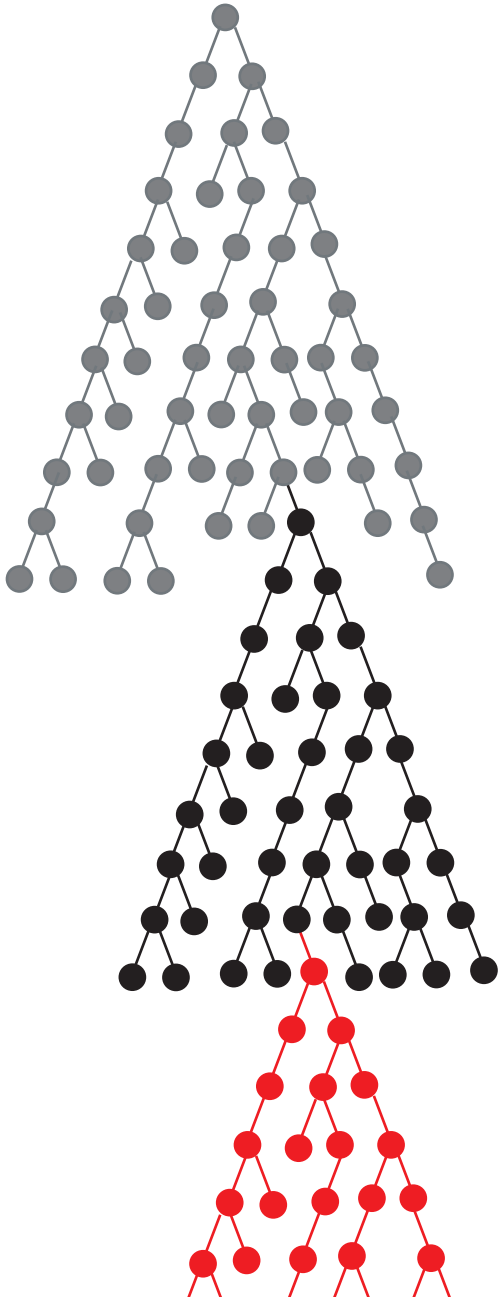
Warning: software often displays unrooted trees like this:



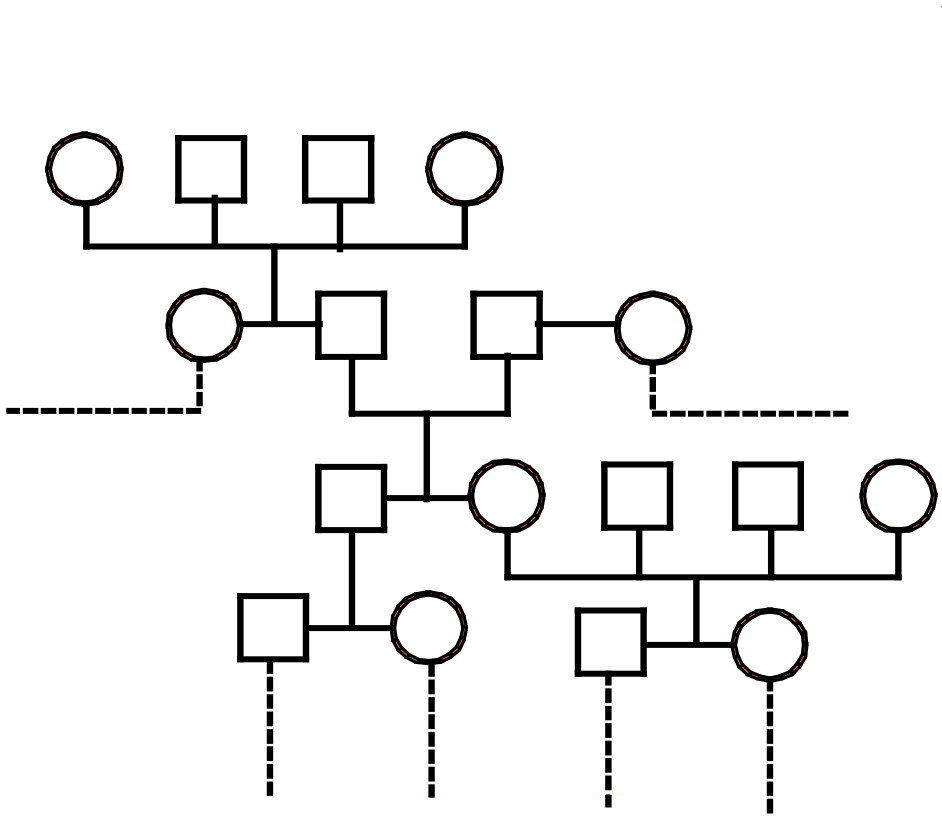
We use trees to represent genealogical relationships in several contexts.

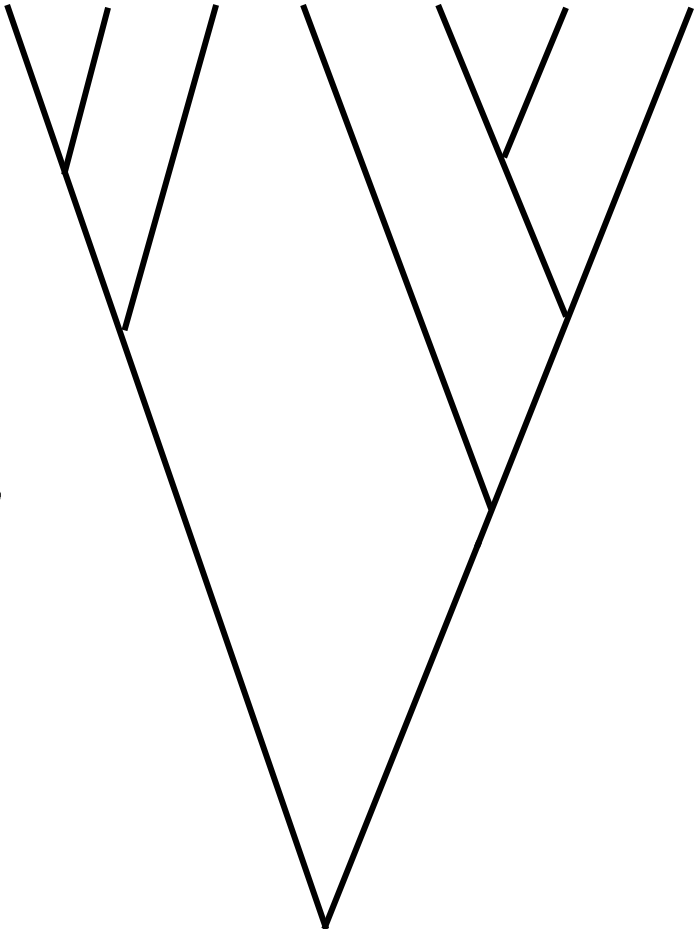
Domain	Sampling	tree	The cause of splitting
Pop. Gen.	> 1 indiv/sp. Few species	Gene tree	> 1 descendants of a single gene copy
Phylogenetics	Few indiv/sp. Many species	Phylogeny	speciation
Mol. Gen.	> 1 locus/sp. > 1 species	Gene tree. Gene family tree	speciation or duplication

Phylogenies are an inevitable result of molecular genetics

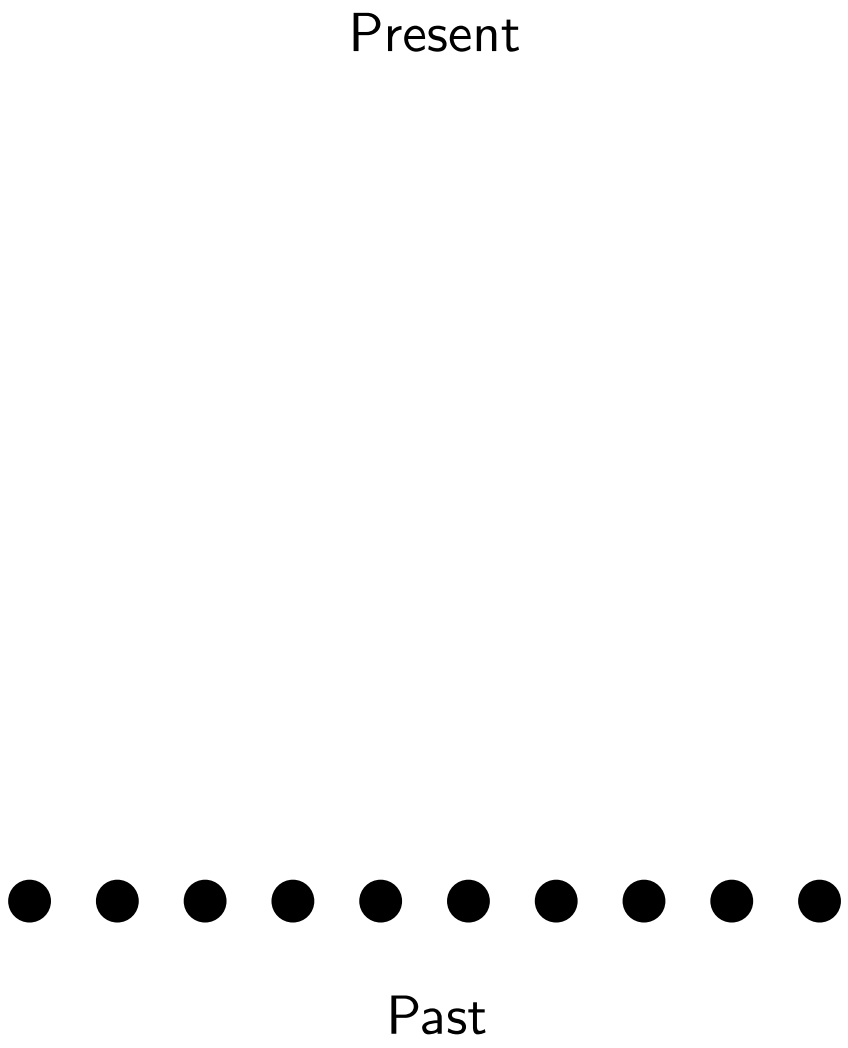


Two types of genealogies



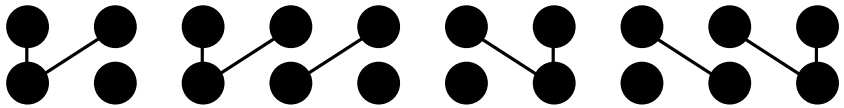


Genealogies within a population



Genealogies within a population

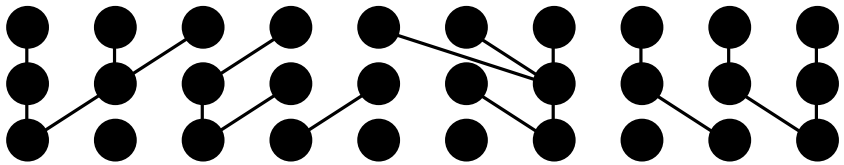
Present



Past

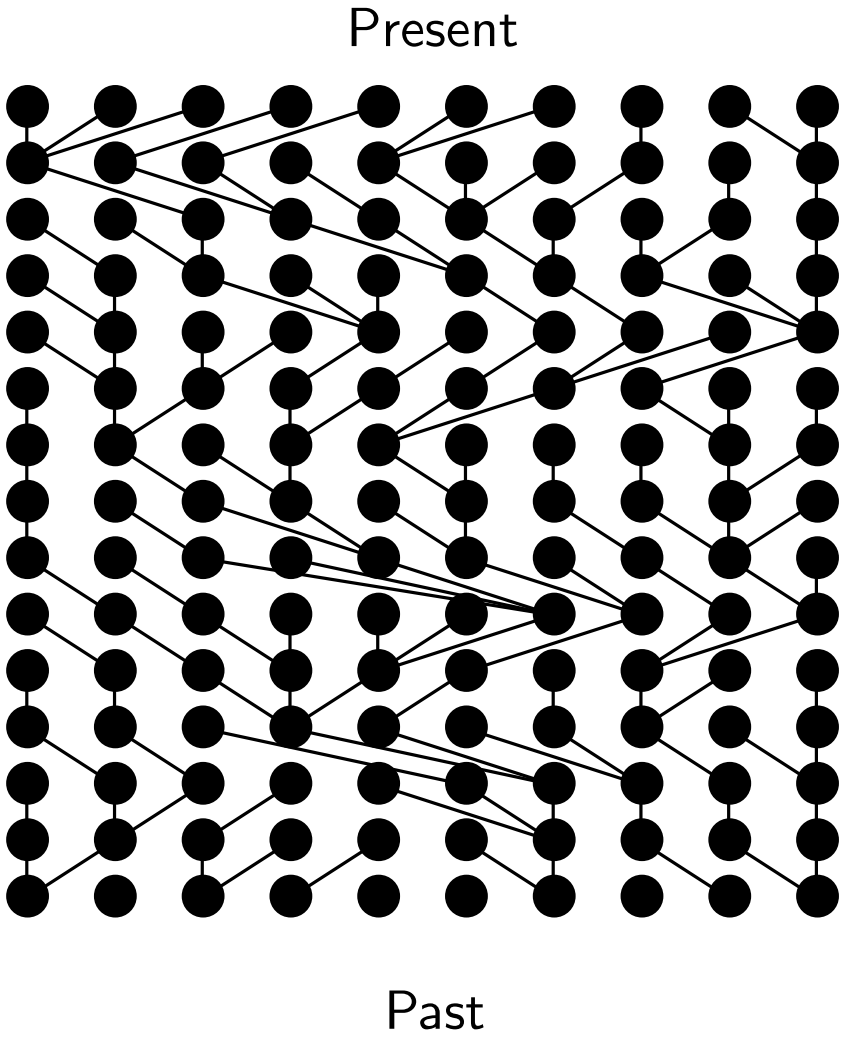
Genealogies within a population

Present

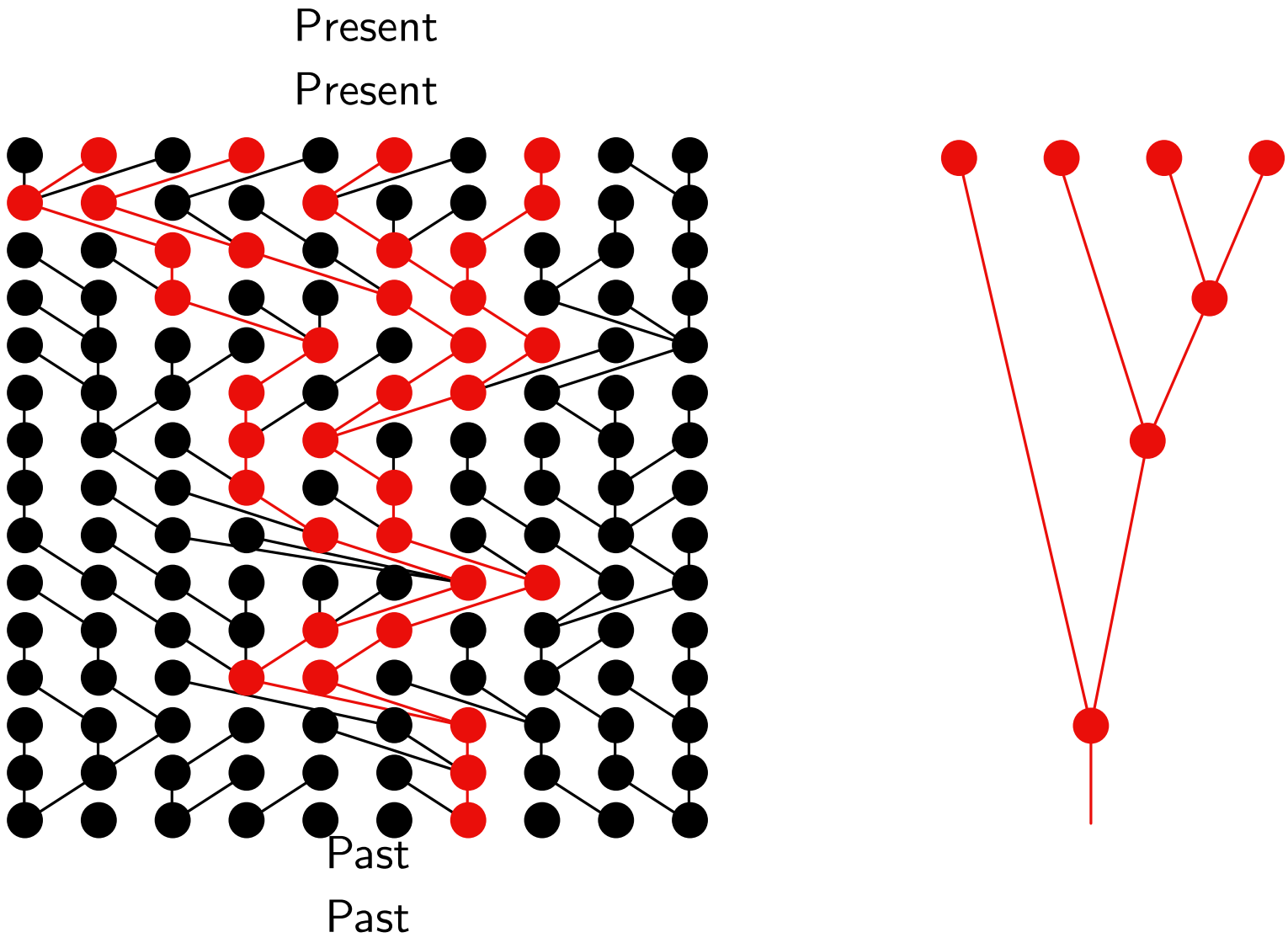


Past

Genealogies within a population



Genealogies within a population



Biparental inheritance would make the picture messier, but the genealogy of the gene copies would still form a tree (if there is no recombination).

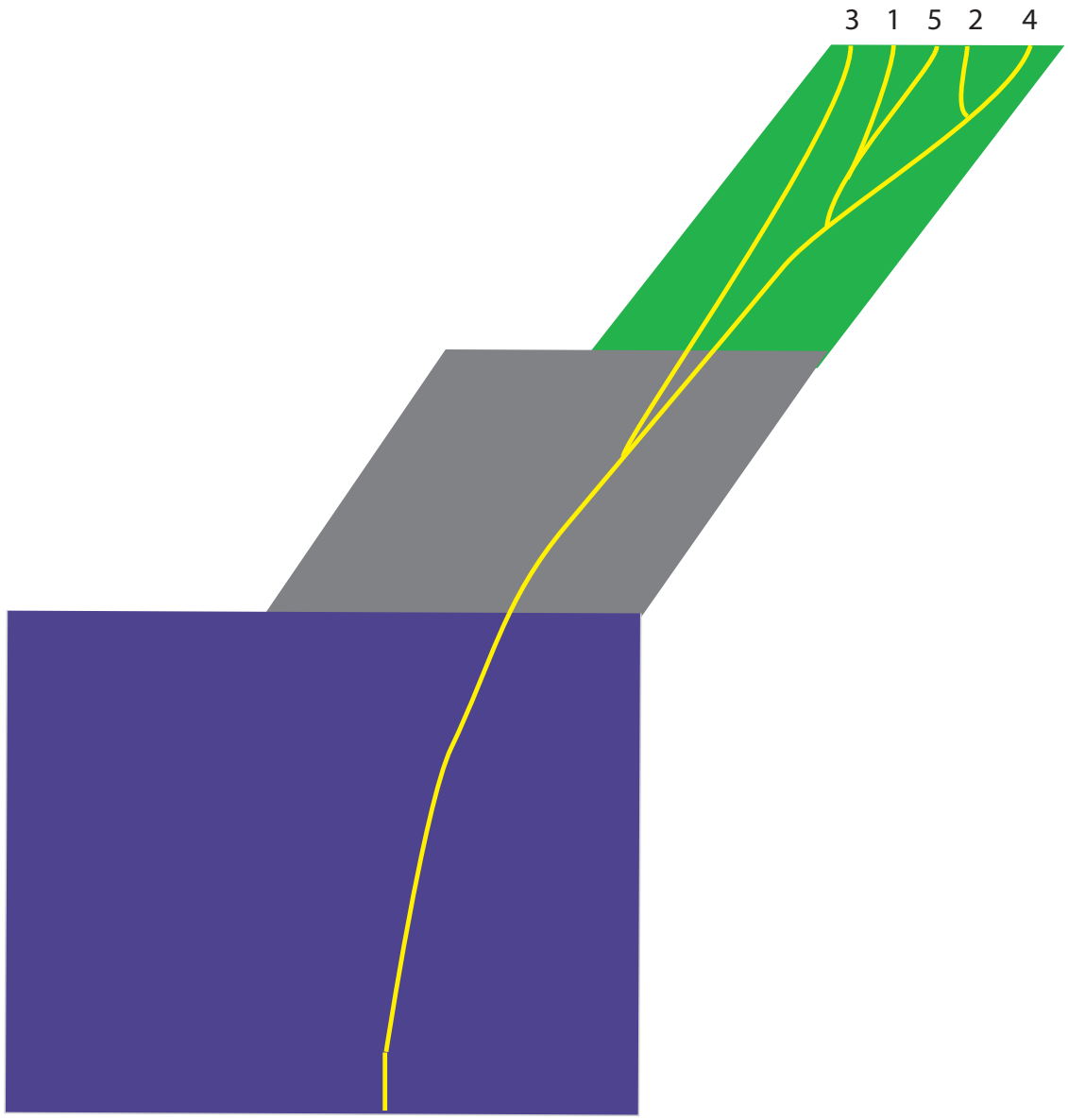
terminology: genealogical trees within population or species trees

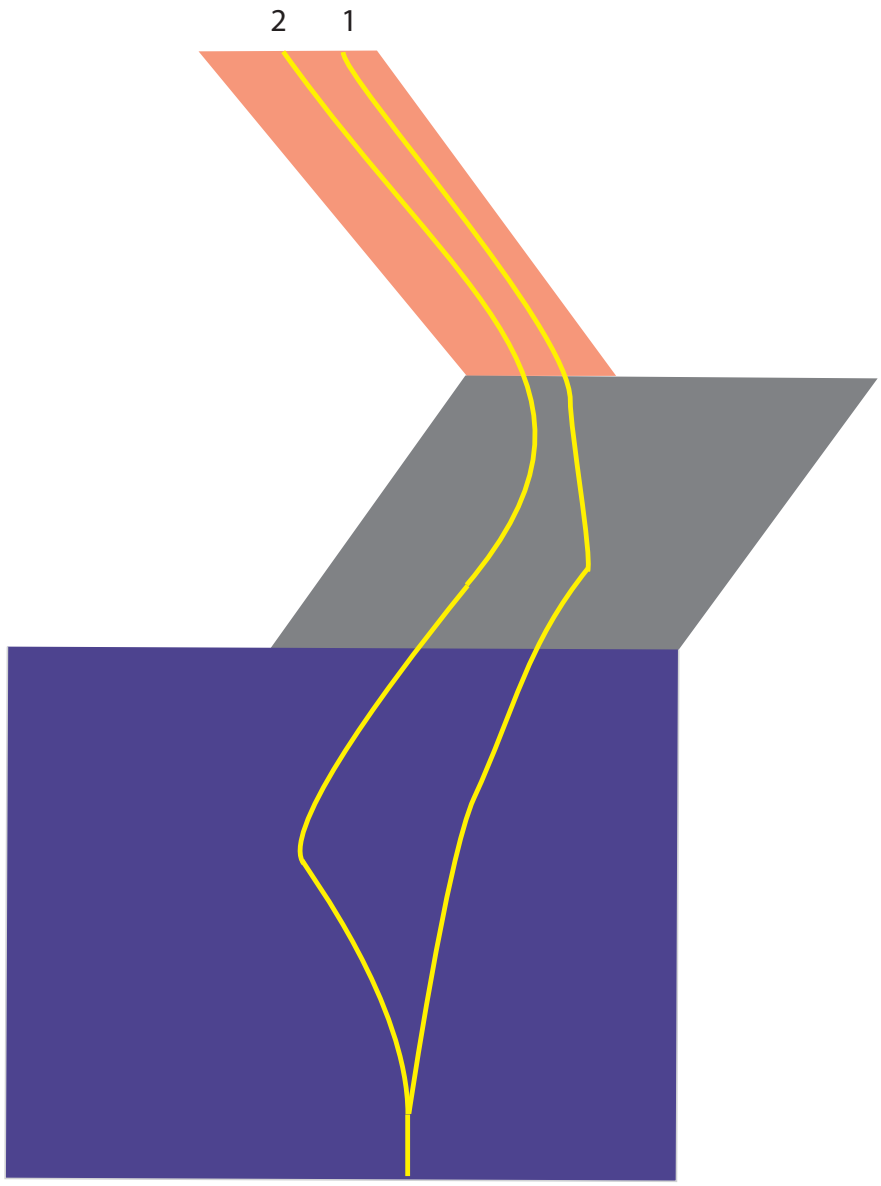
It is tempting to refer to the tips of these gene trees as alleles or haplotypes.

- allele – an alternative form a gene.
- haplotype – a linked set of alleles

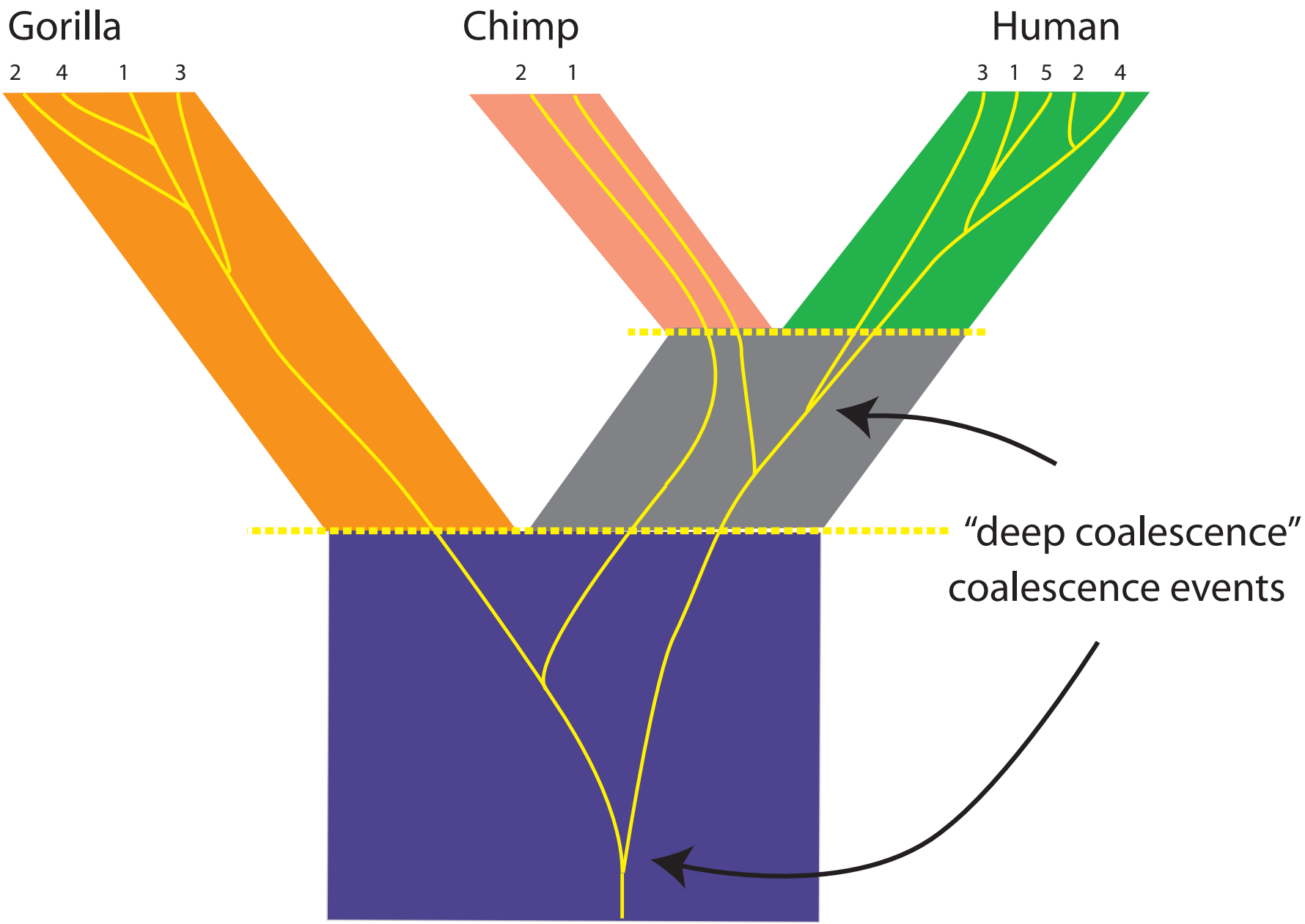
But both of these terms require a differences in sequence.

The gene trees that we draw depict genealogical relationships – regardless of whether or not nucleotide differences distinguish the "gene copies" at the tips of the tree.


3 1 5 2 4


2
1

A “gene tree” within a species tree



terminology: genealogical trees within population or species trees

- coalescence – merging of the genealogy of multiple gene copies into their common ancestor. "Merging" only makes sense when viewed *backwards in time*.

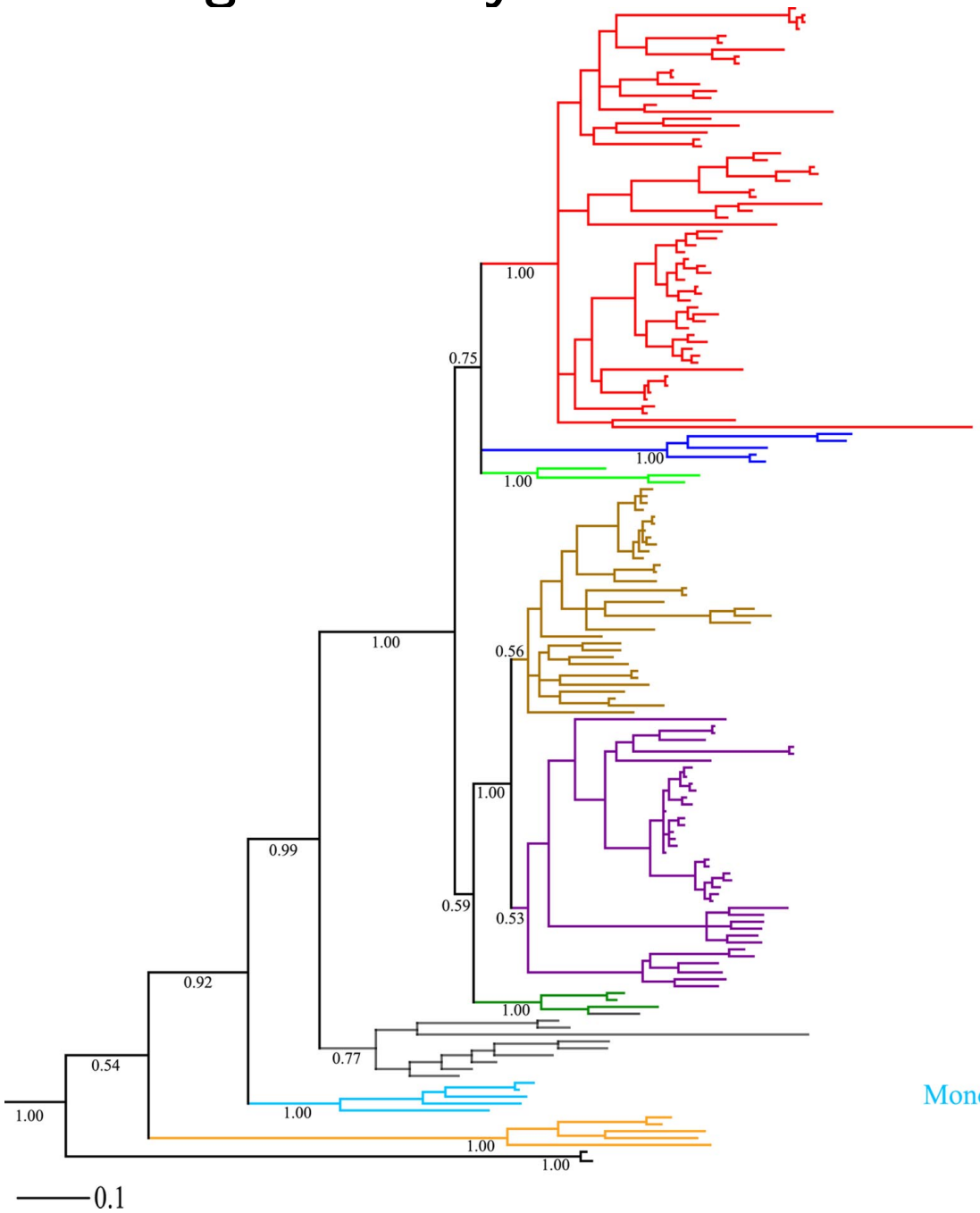
- "deep coalescence" or "incomplete lineage sorting" refer to the *failure* of gene copies to coalesce within the duration of the species – the lineages coalesce in an ancestral species

terminology: genealogical trees within population or species trees

- coalescence – merging of the genealogy of multiple gene copies into their common ancestor. "Merging" only makes sense when viewed *backwards in time*.

- "deep coalescence" or "incomplete lineage sorting" refer to the *failure* of gene copies to coalesce within the duration of the species – the lineages coalesce in an ancestral species

A "gene family tree"



Eutherian β-δ globin

Marsupial β-globin

Monotreme ε^P-β^Pglobin

Eutherian ε-globin

Eutherian γ-globin

Marsupial ε-globin

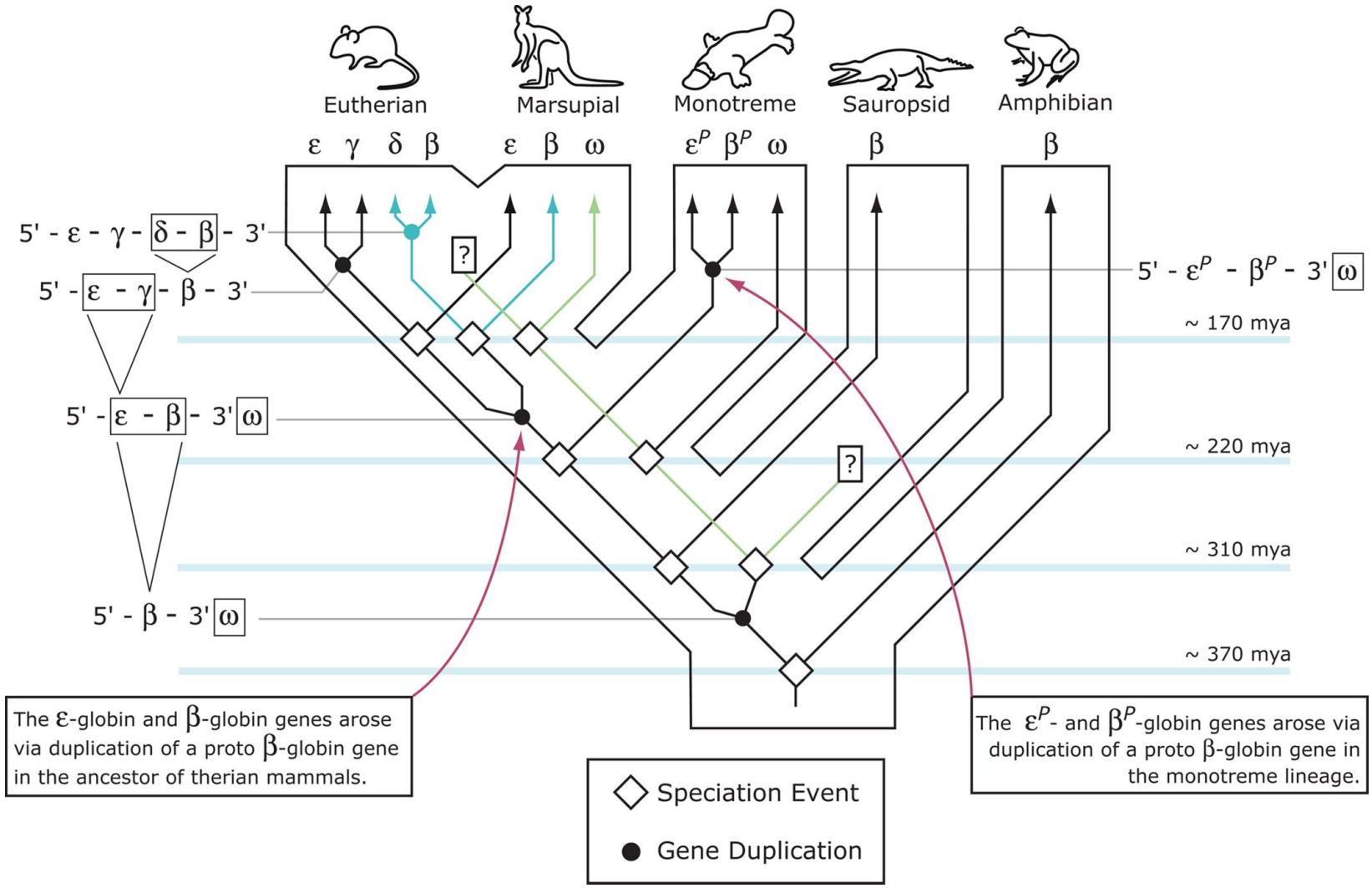
Sauropsida β-like globins

Monotreme and marsupial ω-globin

Amphibian β-like globins

Fish β-like globins

Opazo, Hoffmann and Storz "Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals" PNAS **105(5)** 2008



Opazo, Hoffmann and Storz "Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals" PNAS **105(5)** 2008

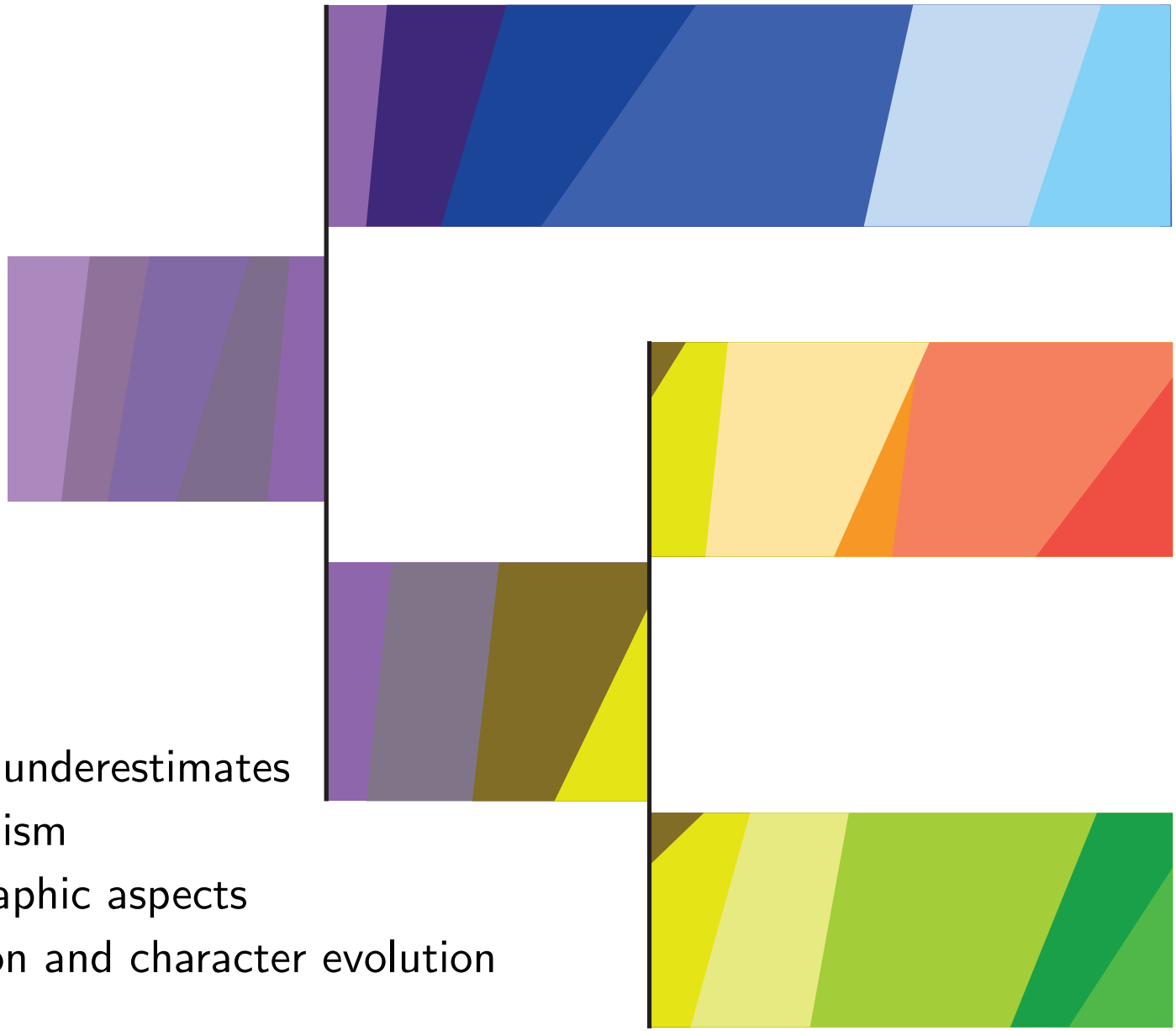
terminology: trees of gene families

- duplication – the creation of a new copy of a gene within the same genome.
- homologous – descended from a common ancestor.
- paralogous – homologous, but resulting from a gene duplication in the common ancestor.
- orthologous – homologous, and resulting from a speciation event at the common ancestor.

Multiple contexts for tree estimation (again):

	The cause of splitting	**Important caveats**
"Gene tree"	DNA replication	recombination is usually ignored
Species tree Phylogeny	speciation	recombination, hybridization, and deep coalescence cause conflict in the data we use to estimate phylogenies
Gene family tree	speciation or duplication	recombination (eg. domain swapping) is not tree-like

Phylogeny with complete genome + "phenome" as colors:

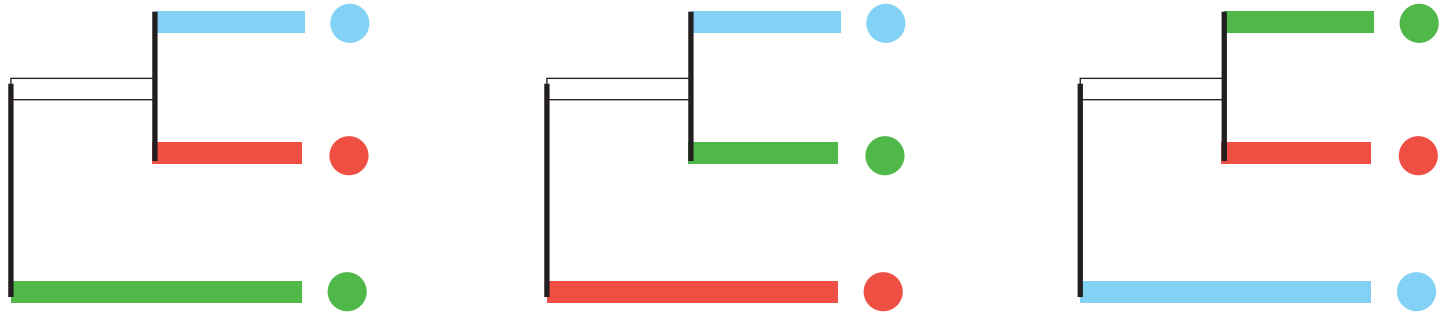


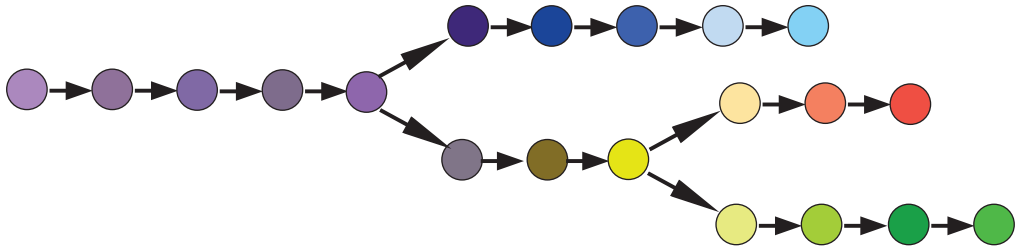
This figure:
dramatically underestimates
polymorphism
ignore geographic aspects
of speciation and character evolution

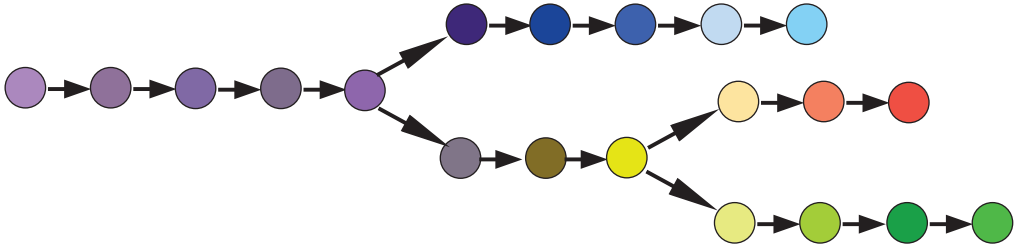
Extant species are just a thin slice of the phylogeny:

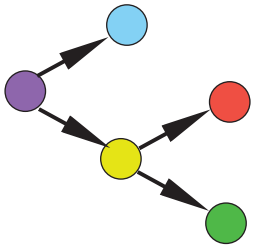
Our exemplar specimens are a subset of the current diversity:

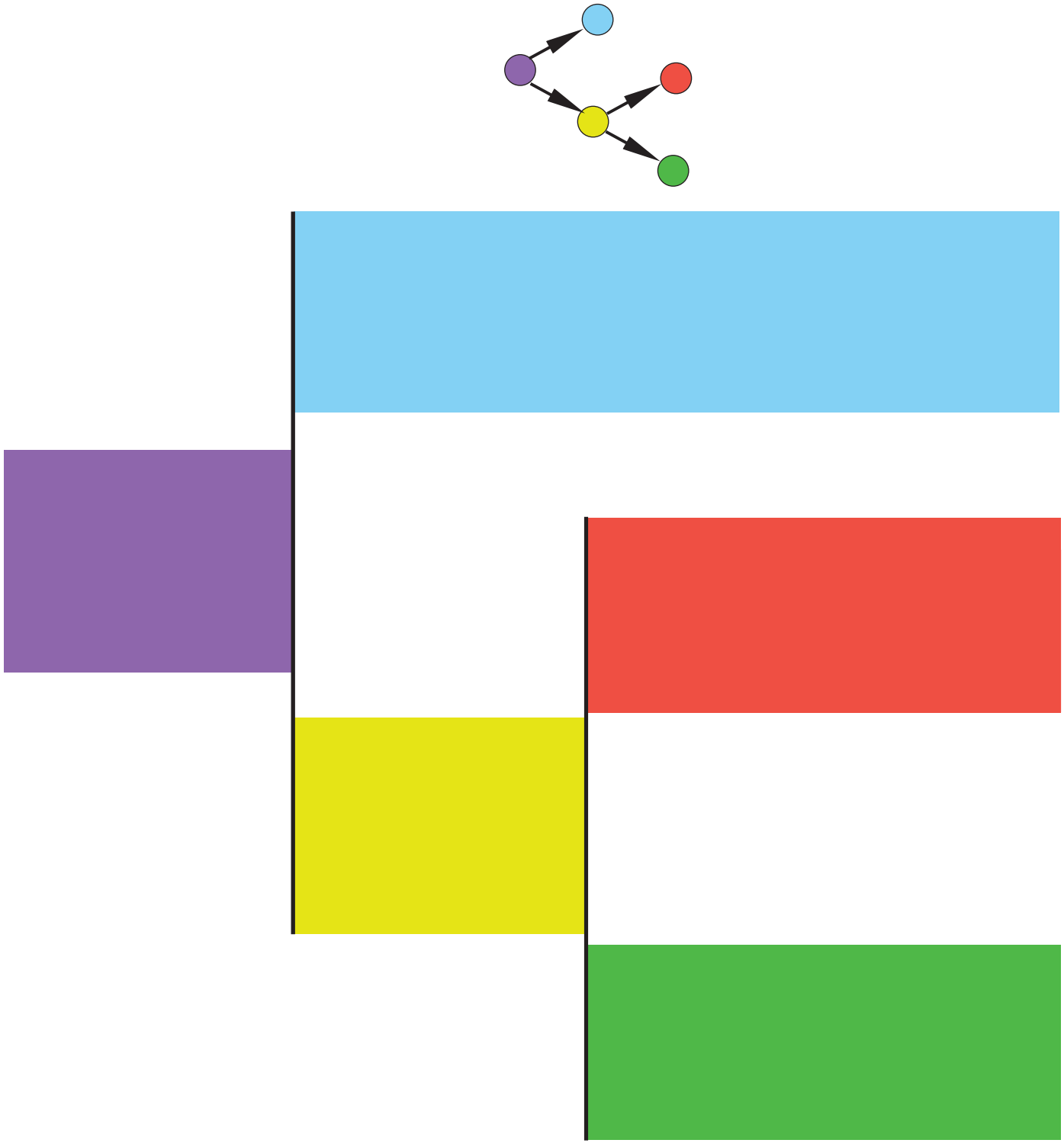
The phylogenetic inference problem:

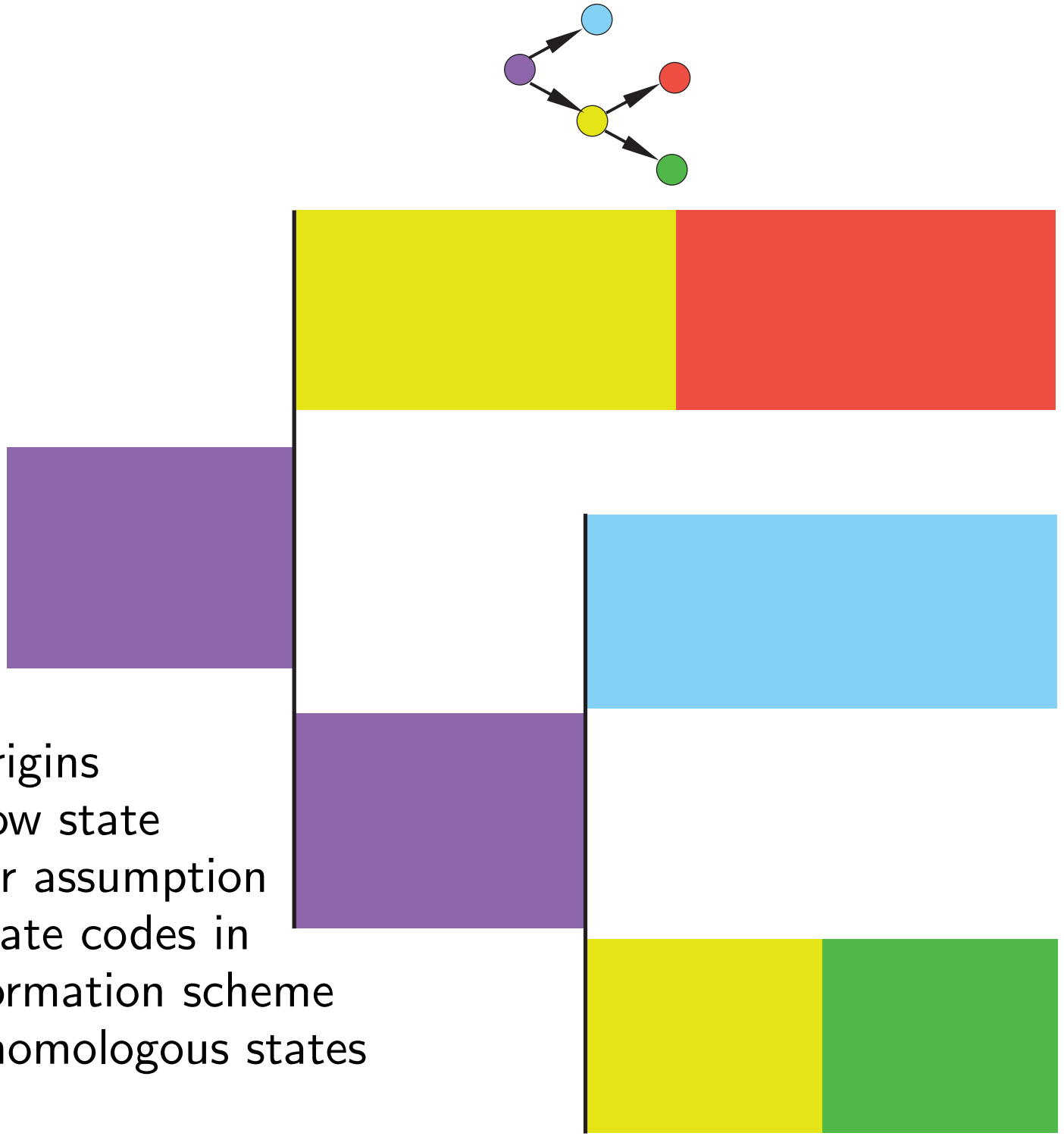




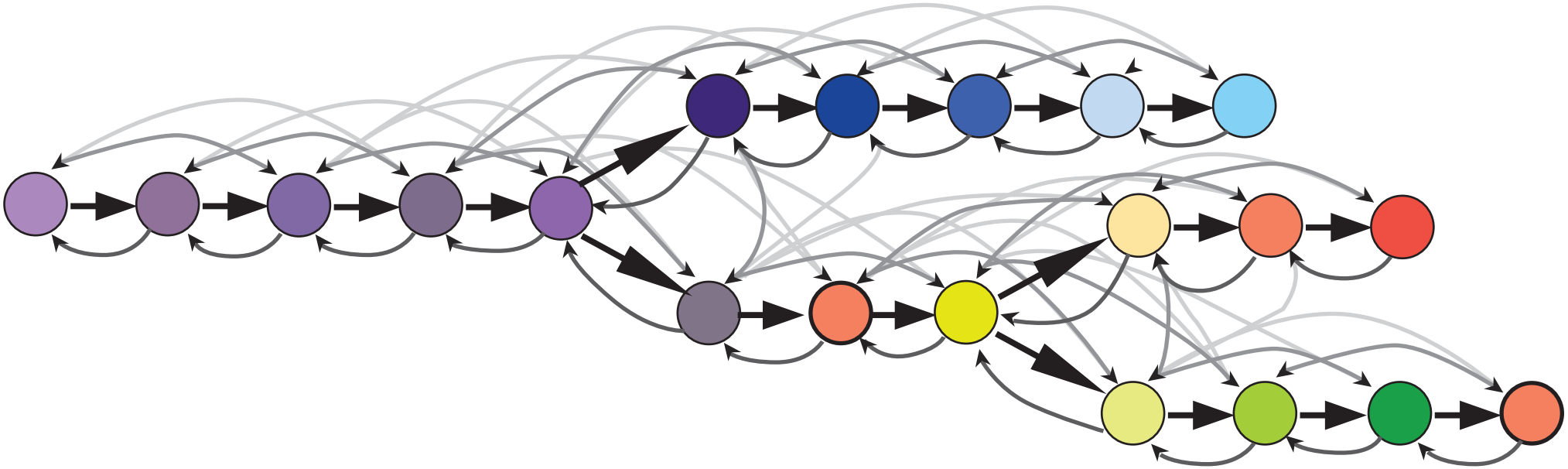








Multiple origins
of the yellow state
violates our assumption
that the state codes in
our transformation scheme
represent homologous states



Character matrices:

		Characters					
		1	2	3	4	5	6
Taxa	*Homo sapiens*	0.13	A	A	rounded	1	1610 - 1755
	Pan paniscus	0.34	A	G	flat	2	0621 - 0843
	Gorilla gorilla	0.46	C	G	pointed	1	795 - 1362

Characters (aka "transformation series") are the columns.
The values in the cells are character states (aka "characters").

		Characters					
		1	2	3	4	5	6
Taxa	*Homo sapiens*	0.13	A	A	rounded	1	1610 - 1755
	Pan paniscus	0.34	A	G	flat	2	0621 - 0843
	Gorilla gorilla	0.46	C	G	pointed	1	795 - 1362

Character coding:

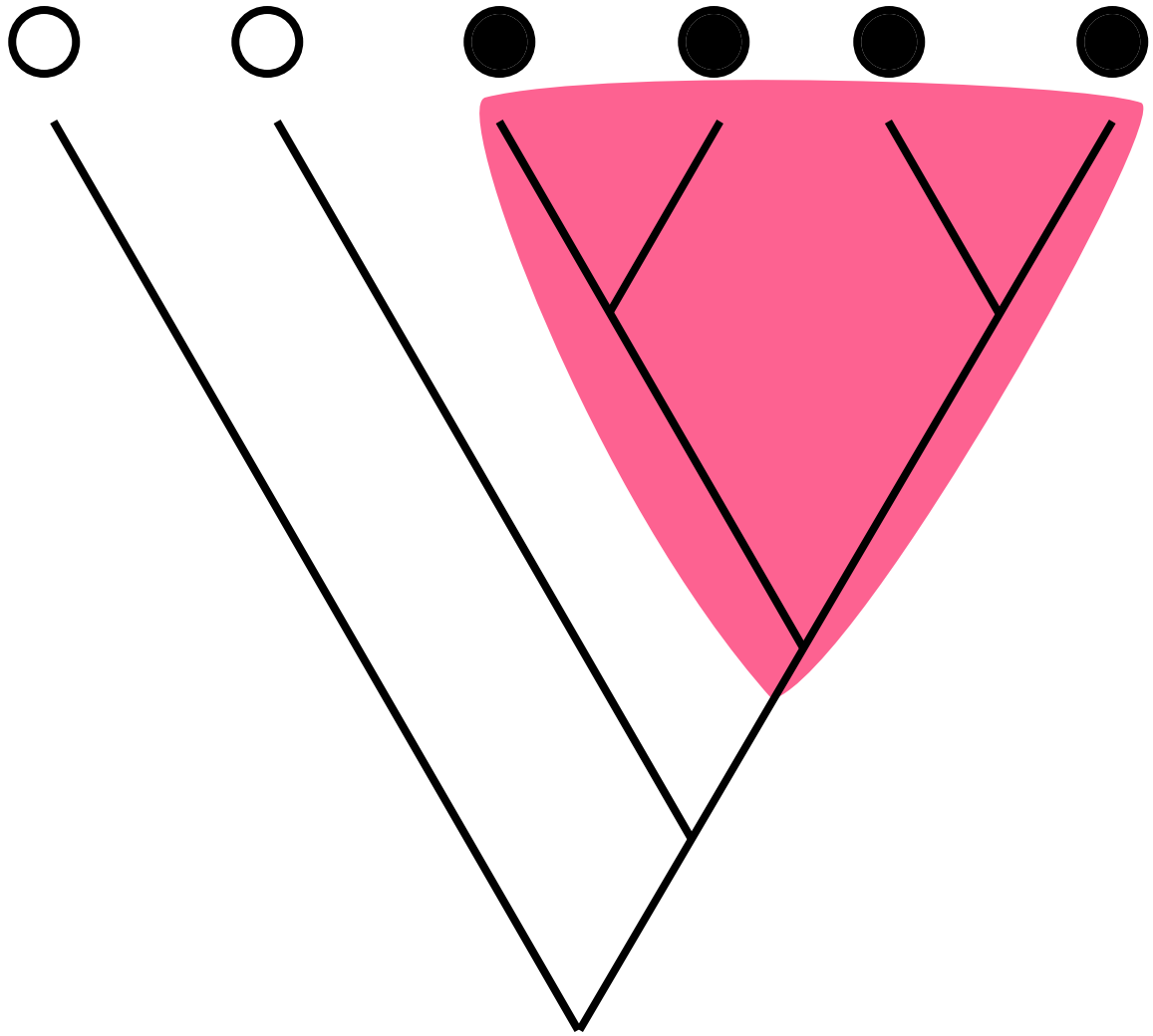
		Characters					
		1	2	3	4	5	6
Taxa	*Homo sapiens*	0	A	A	0	1	4
	Pan paniscus	2	A	G	1	2	0,1
	Gorilla gorilla	3	C	G	2	1	1,2

The meaning of homology (**very roughly**):

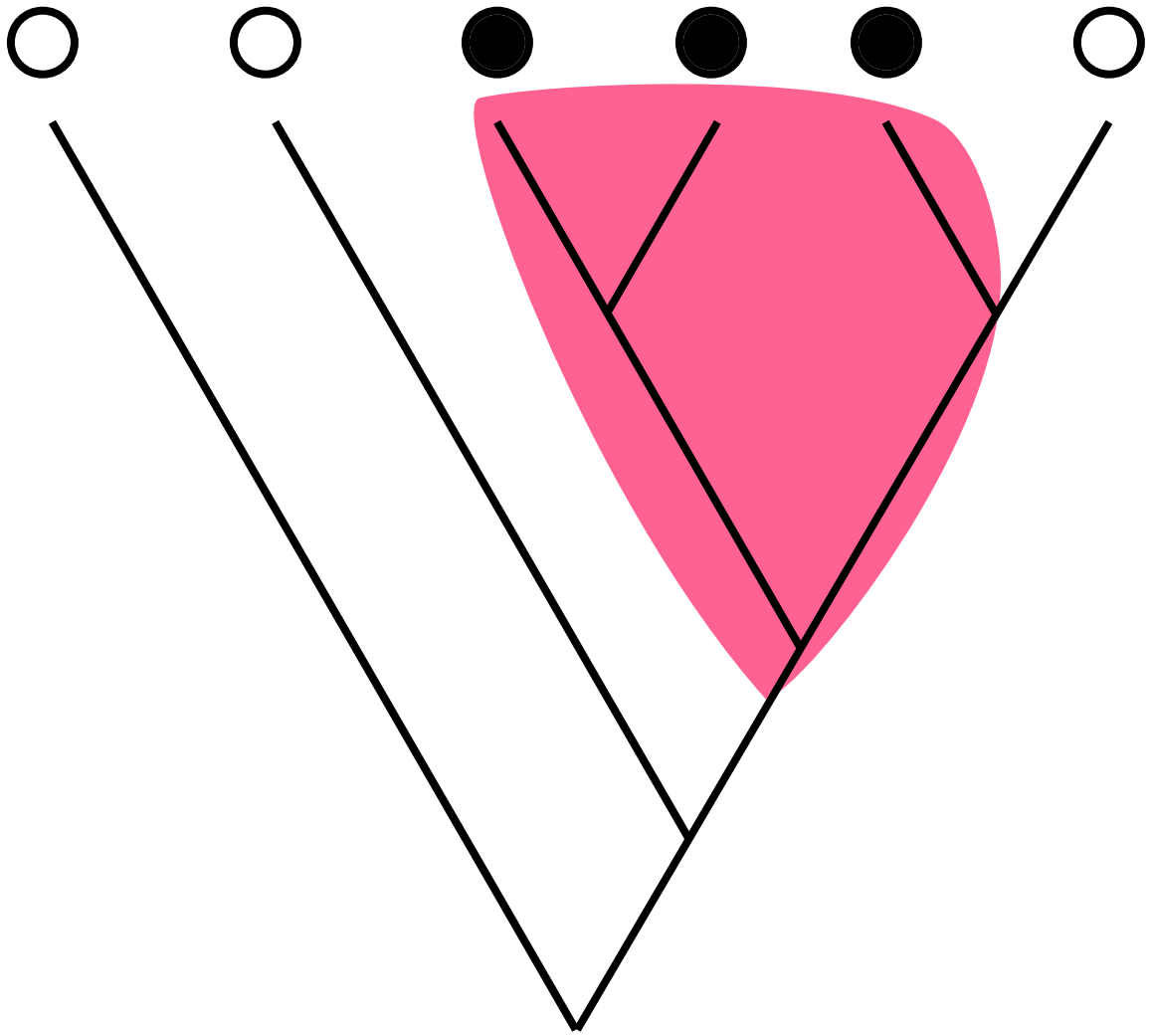
1. comparable (when applied to characters)
2. identical by descent (when applied to character states)

Ideally, each possible character state would arise once in the entire history of life on earth.

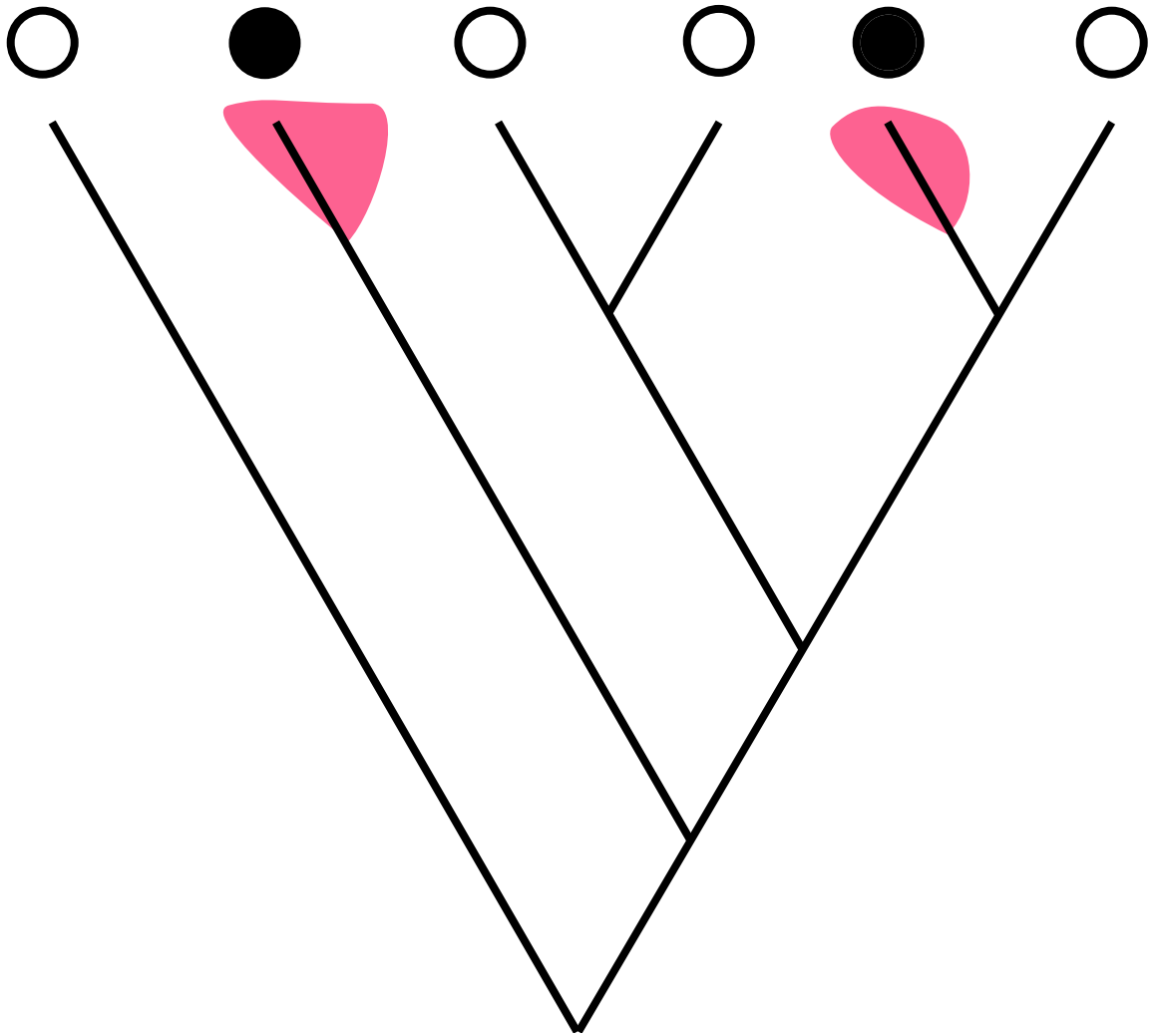
Instances of the filled character state are homologous
Instances of the hollow character state are homologous



Instances of the filled character state are homologous
Instances of the hollow character state are NOT homologous



Instances of the filled character state are NOT homologous
Instances of the hollow character state are homologous

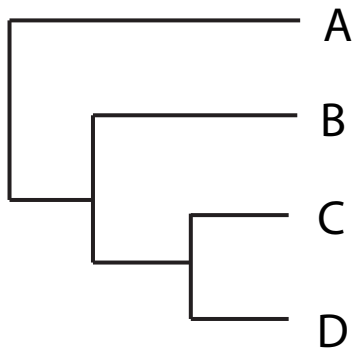


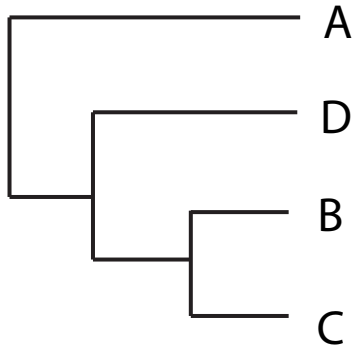
Inference

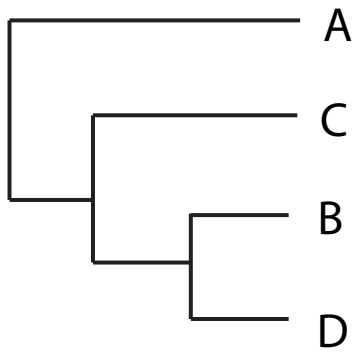
"deriving a conclusion based solely on what one already knows" [1]

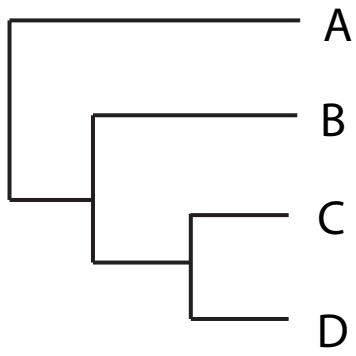
- logical
- statistical

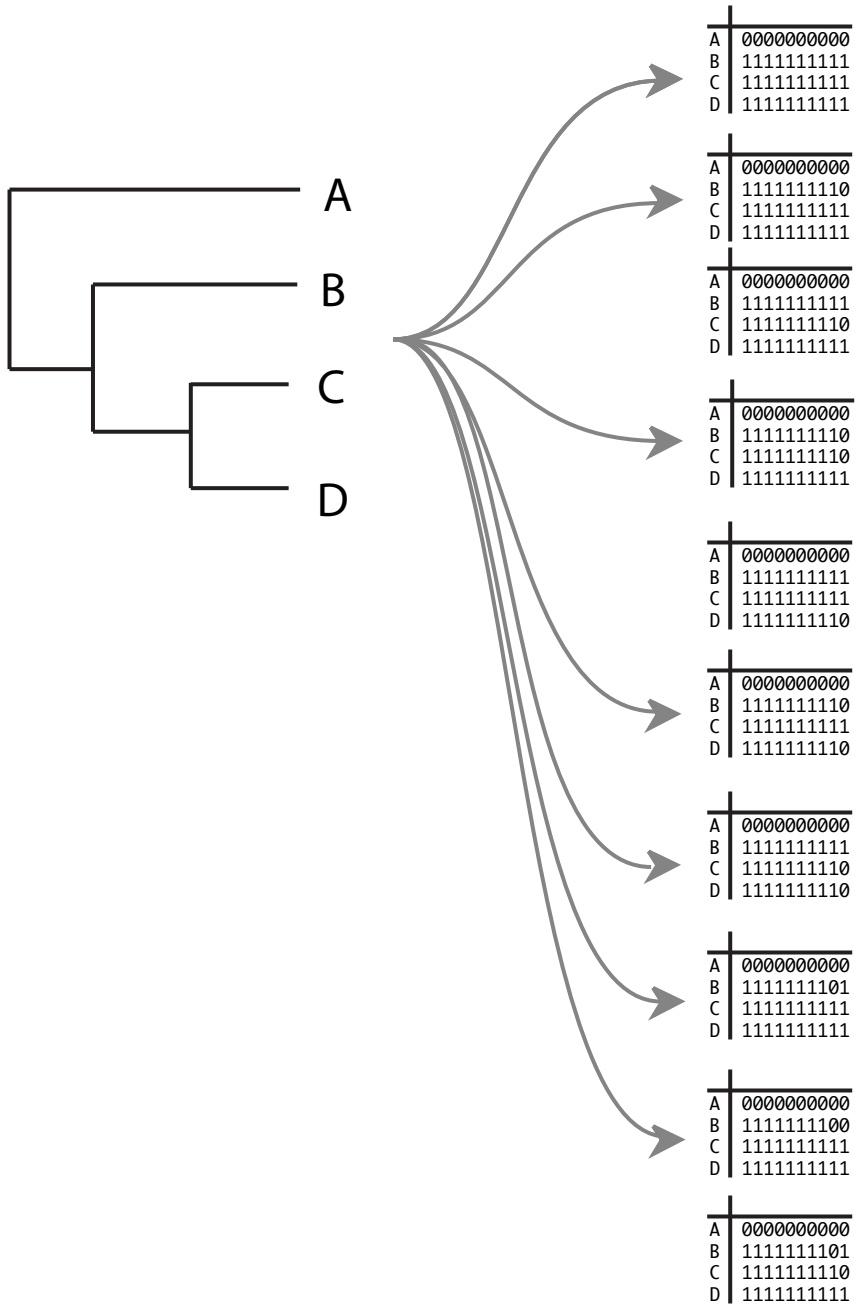
[1] definition from Wikipedia, so it must be correct!

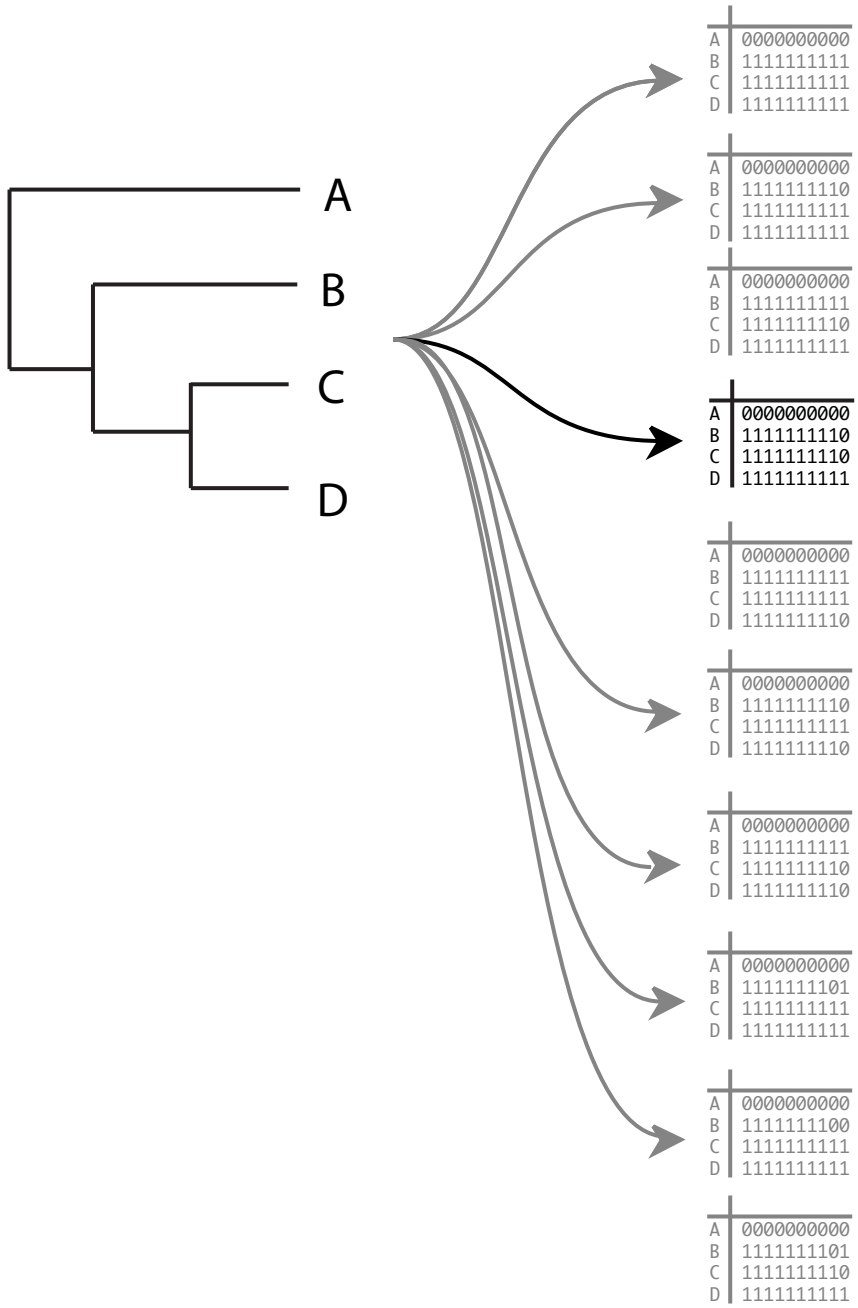

A
B
C
D

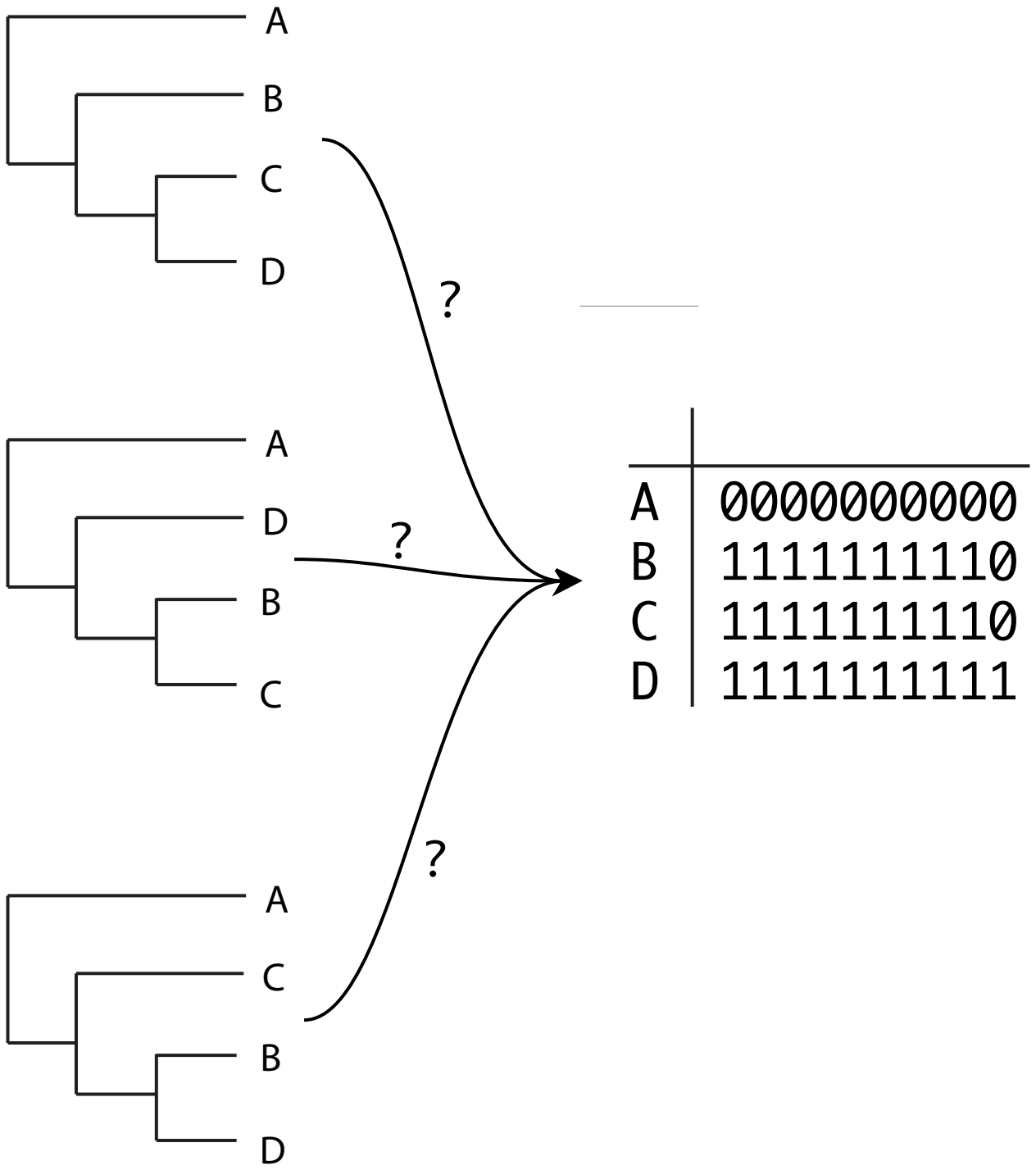

A
D
B
C

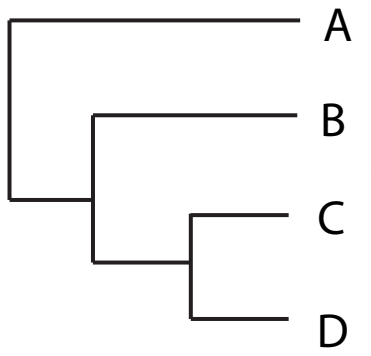

A
C
B
D

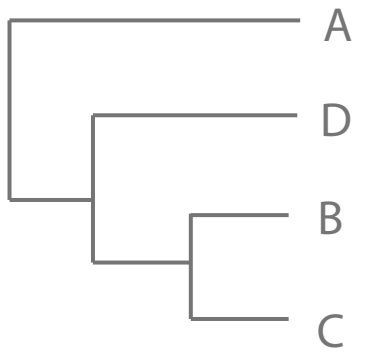

A
B
C
D

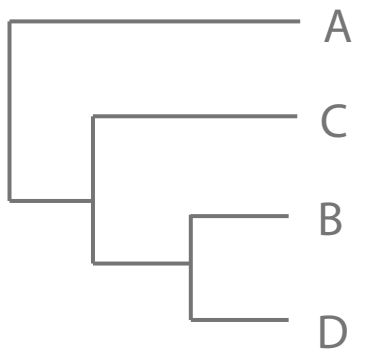

A
B
C
D
A 00000000000
B 1111111111
C 1111111111
D 1111111111
A 00000000000
B 1111111110
C 1111111111
D 1111111111
A 00000000000
B 1111111111
C 1111111110
D 1111111111
A 00000000000
B 1111111110
C 1111111110
D 1111111111
A 00000000000
B 1111111111
C 1111111111
D 1111111110
A 00000000000
B 1111111110
C 1111111111
D 1111111110
A 00000000000
B 1111111111
C 1111111110
D 1111111110
A 00000000000
B 1111111101
C 1111111111
D 1111111111
A 00000000000
B 1111111100
C 1111111111
D 1111111111
A 00000000000
B 1111111101
C 1111111110
D 1111111111


A
B
C
D
A 0000000000
B 1111111111
C 1111111111
D 1111111111
A 0000000000
B 1111111110
C 1111111111
D 1111111111
A 0000000000
B 1111111111
C 1111111110
D 1111111111
A 0000000000
B 1111111110
C 1111111110
D 1111111111
A 0000000000
B 1111111111
C 1111111111
D 1111111110
A 0000000000
B 1111111110
C 1111111111
D 1111111110
A 0000000000
B 1111111111
C 1111111110
D 1111111110
A 0000000000
B 1111111101
C 1111111111
D 1111111111
A 0000000000
B 1111111100
C 1111111111
D 1111111111
A 0000000000
B 1111111101
C 1111111110
D 1111111111


A
B
C
D
?
A
D
B
C
?
A
C
B
D
?
A 0000000000
B 1111111110
C 1111111110
D 1111111111







A	0000000000
B	1111111110
C	1111111110
D	1111111111

Logical Inference

Deductive reasoning:

1. start from premises
2. apply proper rules
3. arrive at statements that were not obviously contained in the premises.

If the rules are valid (logically sound) and the premises are true, then the conclusions are *guaranteed* to be true.

Deductive reasoning

```
All men are mortal.
Socrates is a man.
--------------------
Therefore Socrates is mortal.
```

Can we infer phylogenies from character data using deductive reasoning?

Logical approach to phylogenetics

Premise: The following character matrix is correctly coded (character states are homologous in the strict sense):

	1
taxon A	Z
taxon B	Y
taxon C	Y

Is there a valid set of rules that will generate the tree as a conclusion?

Logical approach to phylogenetics (cont)

Rule: Two taxa that share a character state must be more closely related to each other than either is to a taxon that displays a different state.

Is this a valid rule?

Invalid rule

Here is an example in which we are confident that the homology statements are correct, but our rule implies two conflicting trees:

	placenta	vertebra
Homo sapiens	Z	A
Rana catesbiana	Y	A
Drosophila melanogaster	Y	B

Hennigian logical analysis

The German entomologist Willi Hennig (in addition to providing strong arguments for phylogenetic classifications) clarified the logic of phylogenetic inference.

Hennig's correction to our rule: Two taxa that share a **derived** character state must be more closely related to each other than either is to a taxon that displays the **primitive** state.

Hennig's logic is valid

Here we will use 0 for the primitive state, and 1 for the derived state.

	placenta	vertebra
Homo sapiens	1	1
Rana catesbiana	0	1
Drosophila melanogaster	0	0

Now the character "placenta" does not provide a grouping, but "vertebra" groups human and frog as sister taxa.

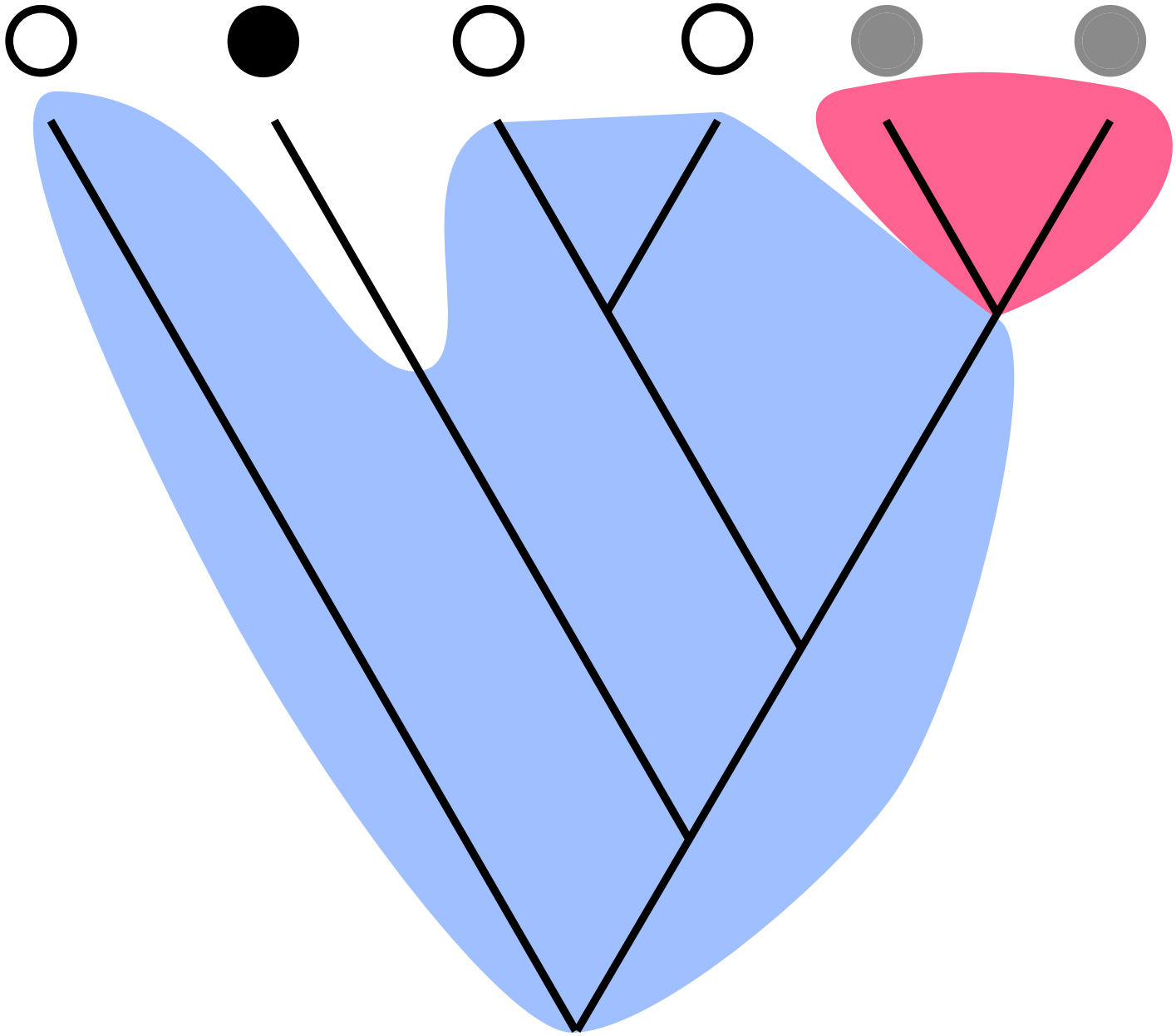
Hennigian terminology

prefixes:

- “apo” - refers to the new or derived state
- “plesio” - refers to the primitive state
- “syn” or “sym” - used to indicate shared between taxa
- “aut” - used to indicate a state being unique to one taxon

Hennigian rules

- synapomorphy - shared, derived states. Used to diagnose monophyletic groups.
- symplesiomorphy - shared, primitive states. Diagnose icky, unwanted paraphyletic groups.
- autapomorphy – a unique derived state. **No** evidence of phylogenetic relationships.
- constant characters – columns in a matrix with no variability between taxa. **No** evidence of phylogenetic relationships.



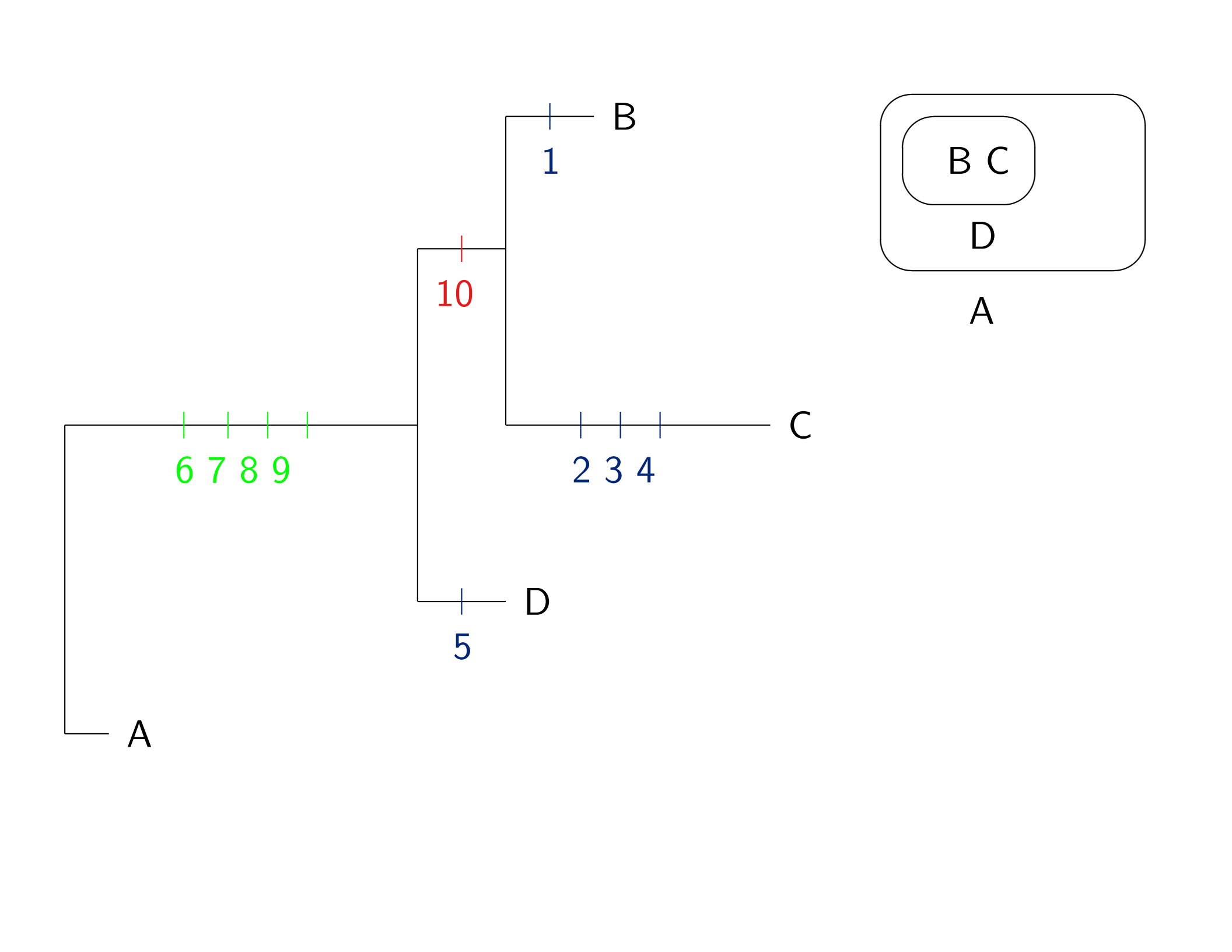
Hennigian inference

When we create a character matrix for Hennig's system, it is crucial that:

- traits assigned the same state represent homologous states (trace back to the MRCA)
- we correctly identify the directionality of the transformations (which state is plesiomorphic and which is apomorphic). The process of identifying the direction of change is called polarization.

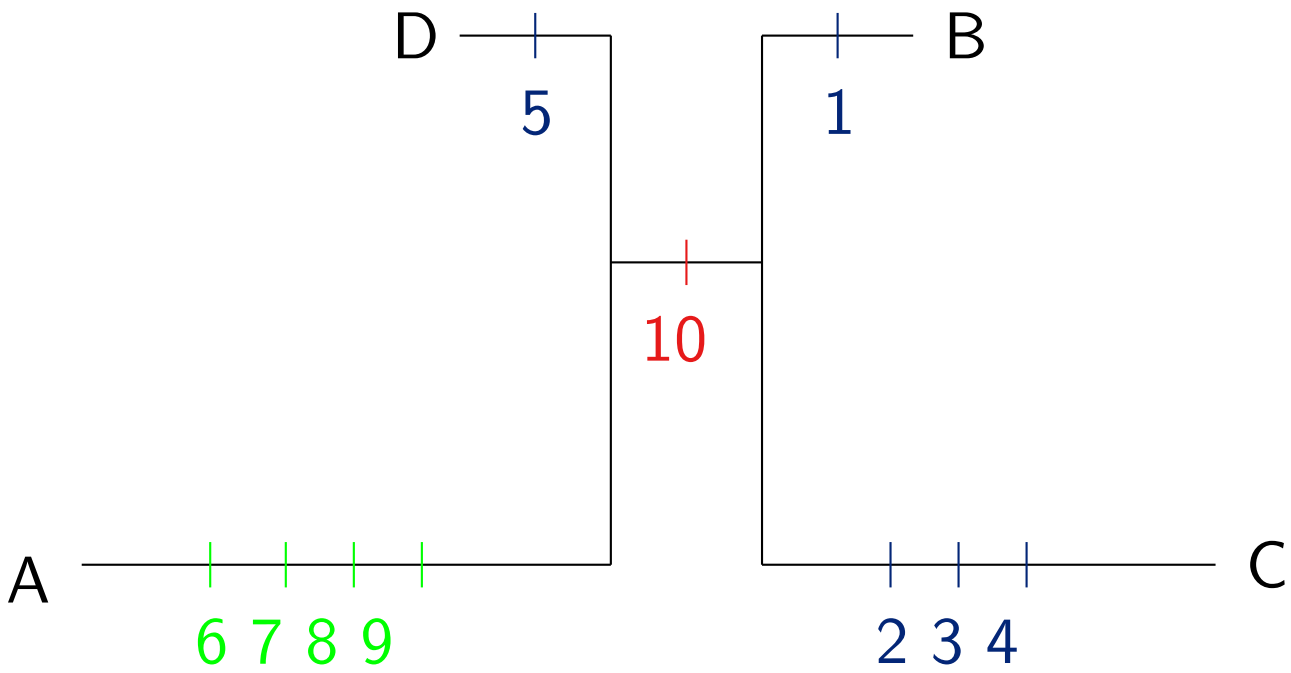
Polarization could be done based on developmental considerations, paleontological evidence, or biogeographic considerations, but the most common technique is outgroup polarization.

Taxon	Character #									
	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1
D	0	0	0	0	1	1	1	1	1	0


B
1
10
6 7 8 9
C
2 3 4
D
5
A
B C
D
A

Interestingly, without polarization Hennig's method can infer unrooted trees. We can get the tree topology, but be unable to tell paraphyletic from monophyletic groups.

The outgroup method amounts to inferring an unrooted tree and then rooting the tree on the branch that leads to an outgroup.



B C

A D

Inadequacy of logic

Unfortunately, though Hennigian logic is valid we quickly find that we do not have a reliable method of generating accurate homology statements.

The logic is valid, but we don't know that the premises are true.

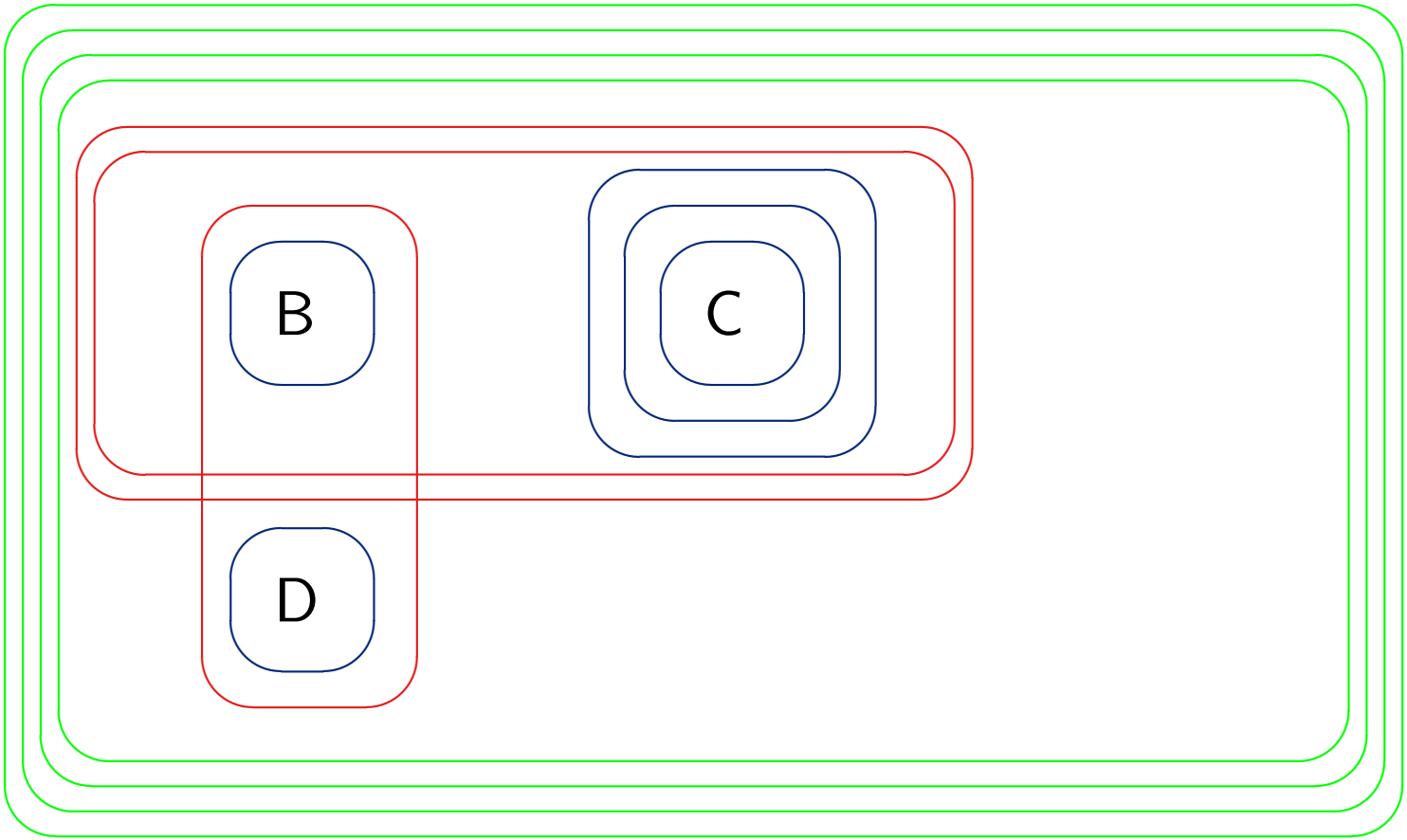
In fact, we almost always find that it is impossible for all of our premises to be true.

Character conflict

Homo sapiens	AGTTCAAGT
Rana catesbiana	AATTCAAGT
Drosophila melanogaster	AGTTCAAGC
C. elegans	AATTCAAGC

The red character implies that either (*Homo* + *Drosophila*) is a group (if G is derived) and/or (*Rana* + *C. elegans*) is a group.
The green character implies that either (*Homo* + *Rana*) is a group (if T is derived) and/or (*Drosophila* + *C. elegans*) is a group.
The green and red character cannot both be correct.

Taxon	Character #											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1	1	0
D	0	0	0	0	1	1	1	1	1	0	0	1



A

References

Dyer, L. A. and Gentry, G. L. (2002). Caterpillars and parasitoids of a tropical lowland wet forest. *http://www.caterpillars.org*, Accessed: 2006.

Hill, J. (2001). Monarch caterpillar image. *University of Minnesota / National Science Foundation Image Library*.

Metzker, M. L., Mindell, D. P., Liu, X.-M., Ptax, R. G., Gibbs, R. A., and Hillis, D. M. D. M. (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Science USA*, 99(22):14292–14297.

Sillén-Tullberg, B. (1988). Evolution of gregariousness

in aposematic butterfly larvae: a phylogenetic analysis. *Evolution*, 42(2):293–305.