Phylogenetic inference from data

Emily Jane McTavish

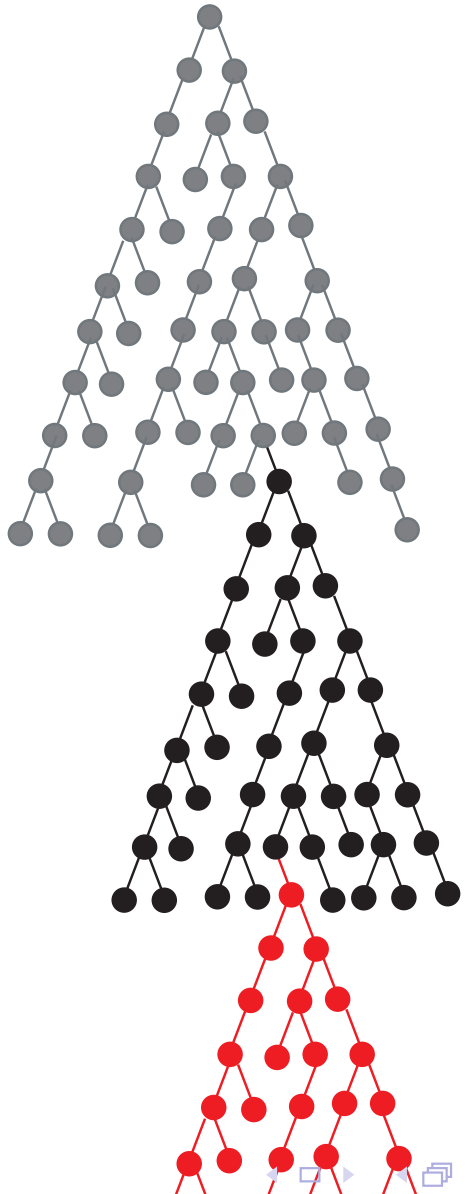
Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Mark Holder, Paul Lewis, Joe Felsenstein, and David Hillis for slides)

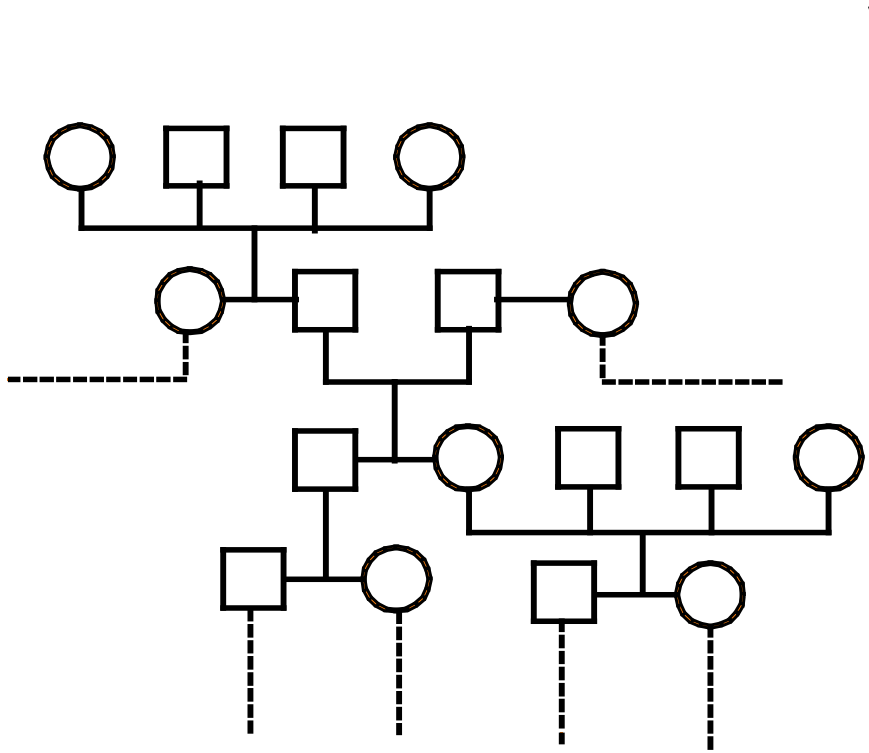
We use trees to represent genealogical relationships in several contexts.

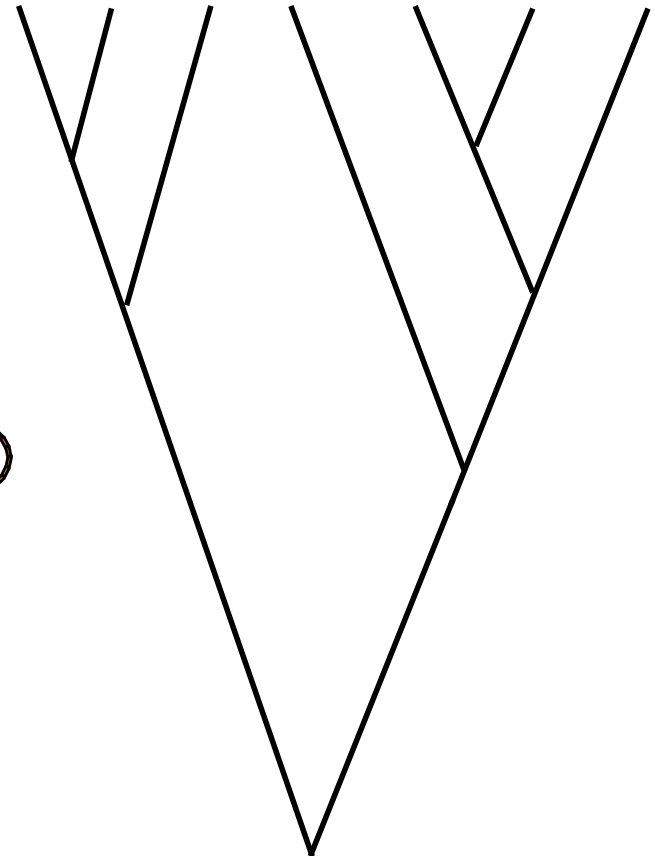
- **Population genetics** estimate gene trees, using several individuals per species, but few species.
  - Splits represent descendents of a single gene copy
- **Phylogenetics** estimate phylogenies, using 1 or few individuals per species, across several species
  - Splits represent speciation events.
- Across molecular evolution, gene duplication + speciation creates more complex patterns.

Phylogenies are an inevitable result of molecular genetics



Two types of genealogies





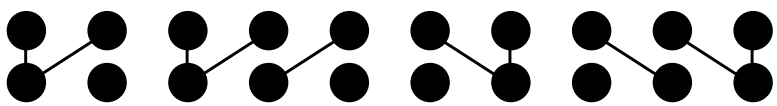
Genealogies within a population

Present

Past

Genealogies within a population

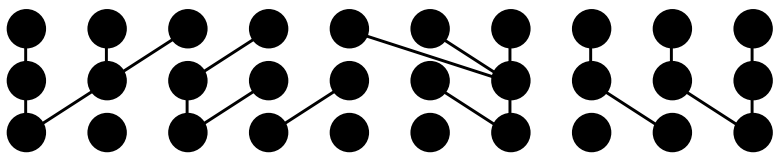
Present



Past

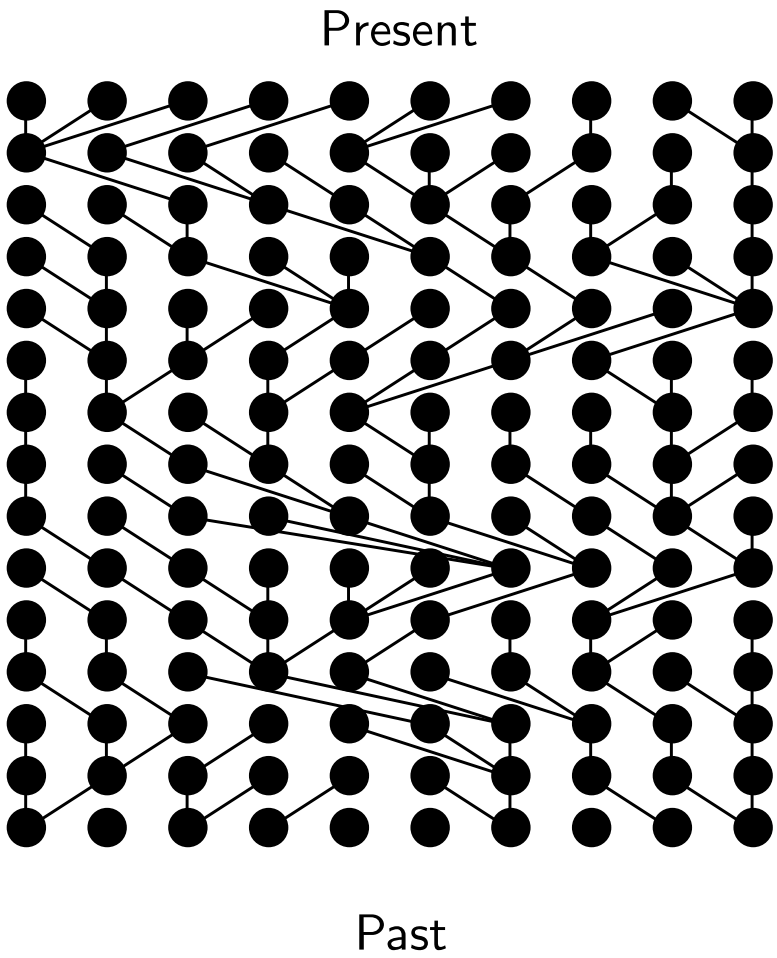
Genealogies within a population

Present

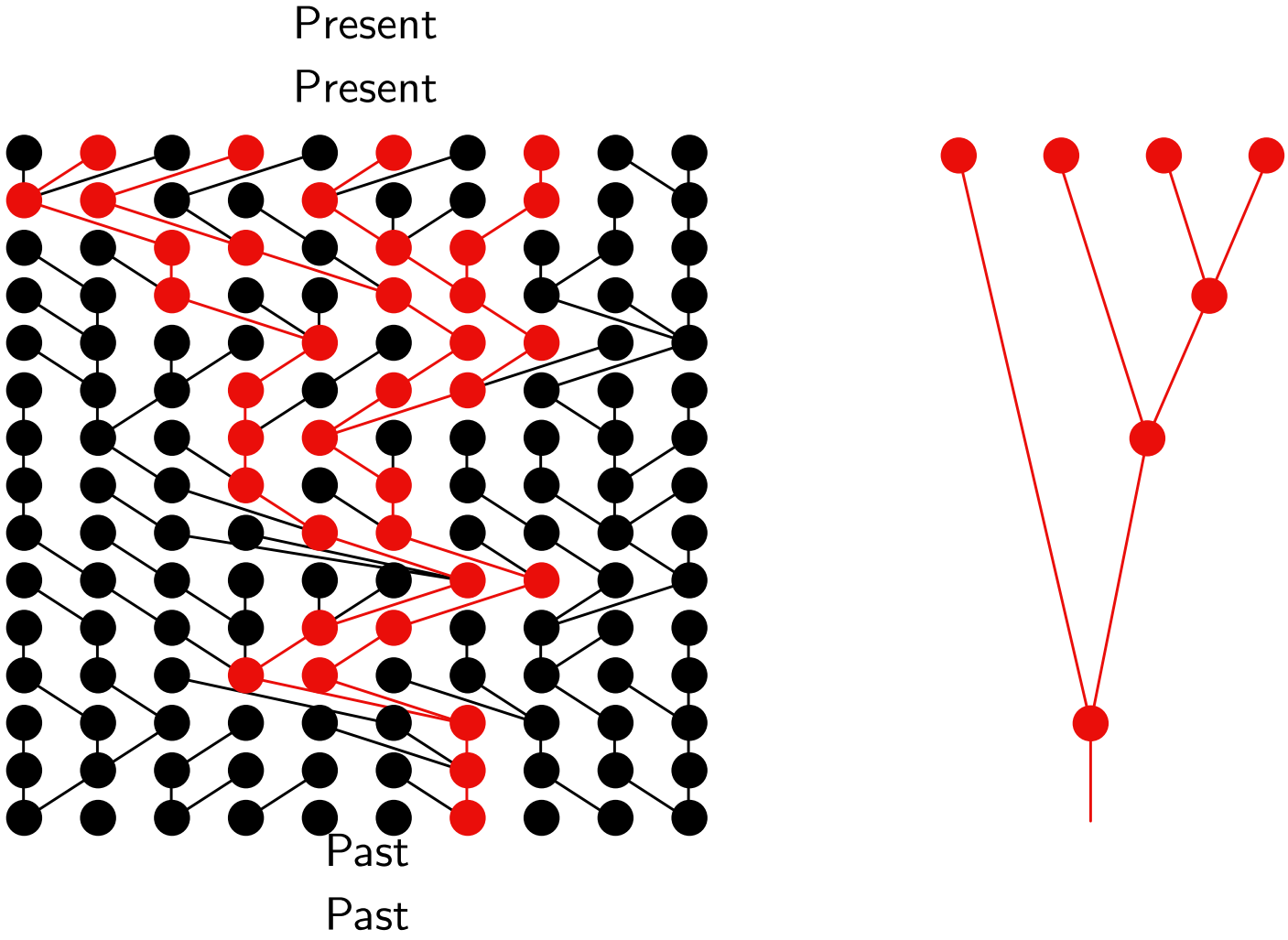


Past

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).

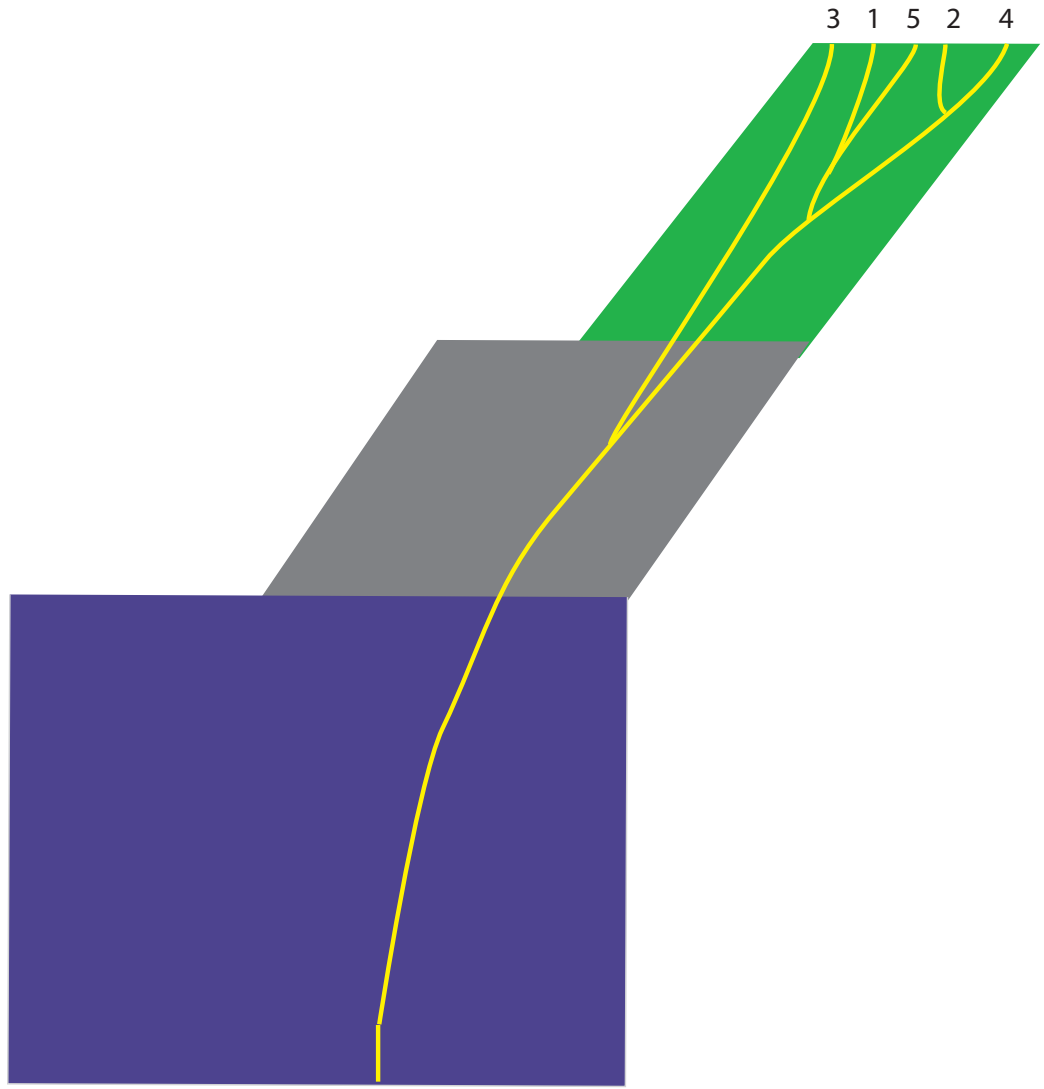
terminology: genealogical trees within population or species trees

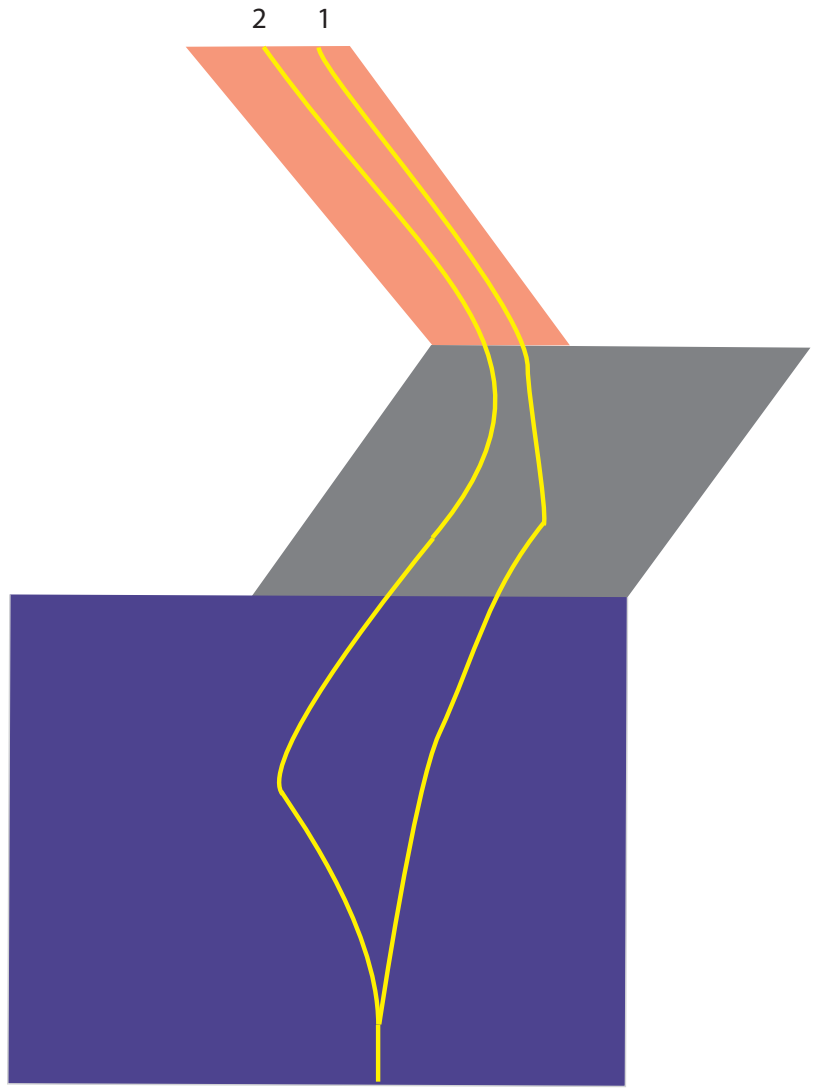
It is tempting to refer to the tips of these gene trees as alleles or haplotypes.

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

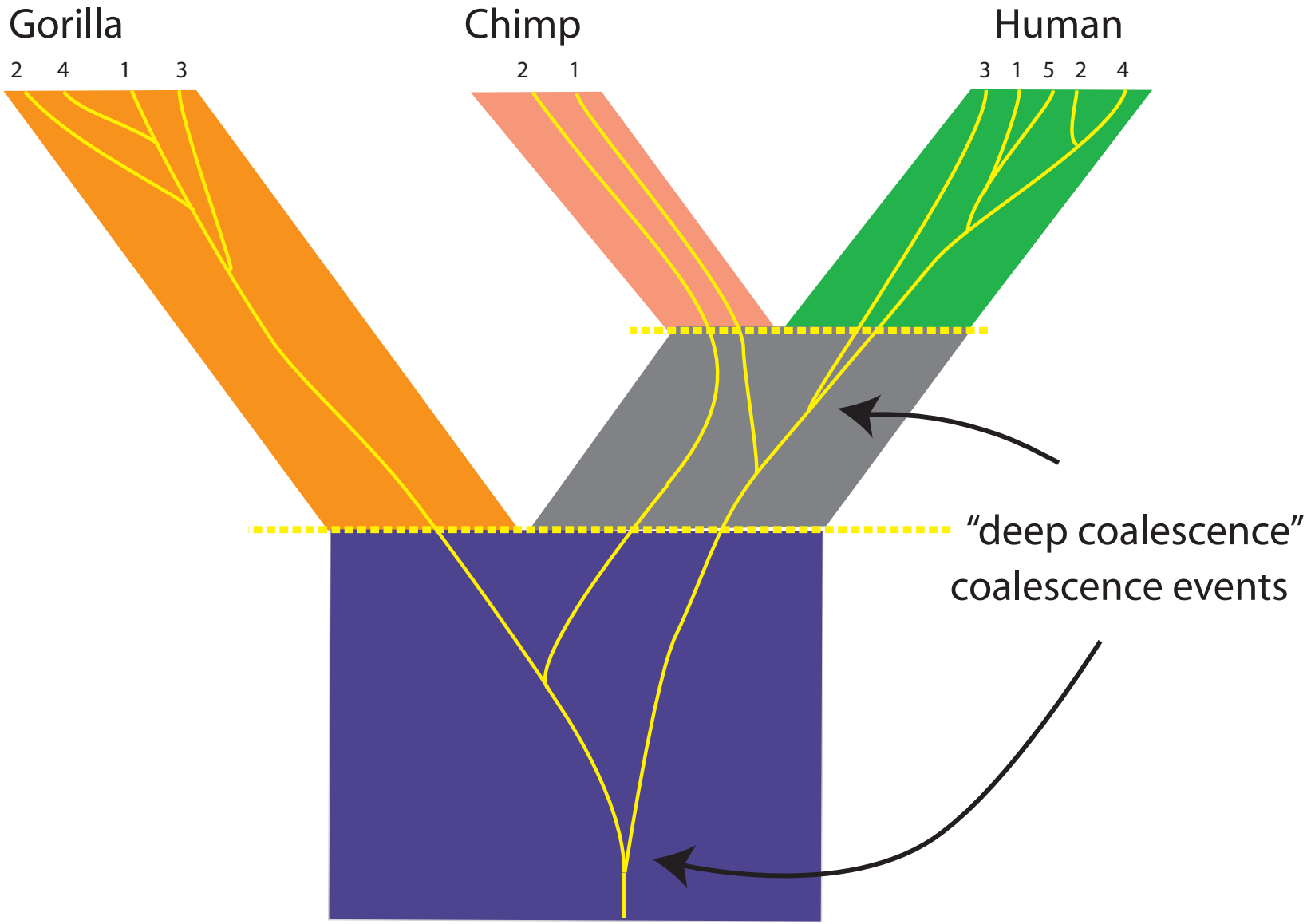
But both of these terms require a differences in sequence.

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*.

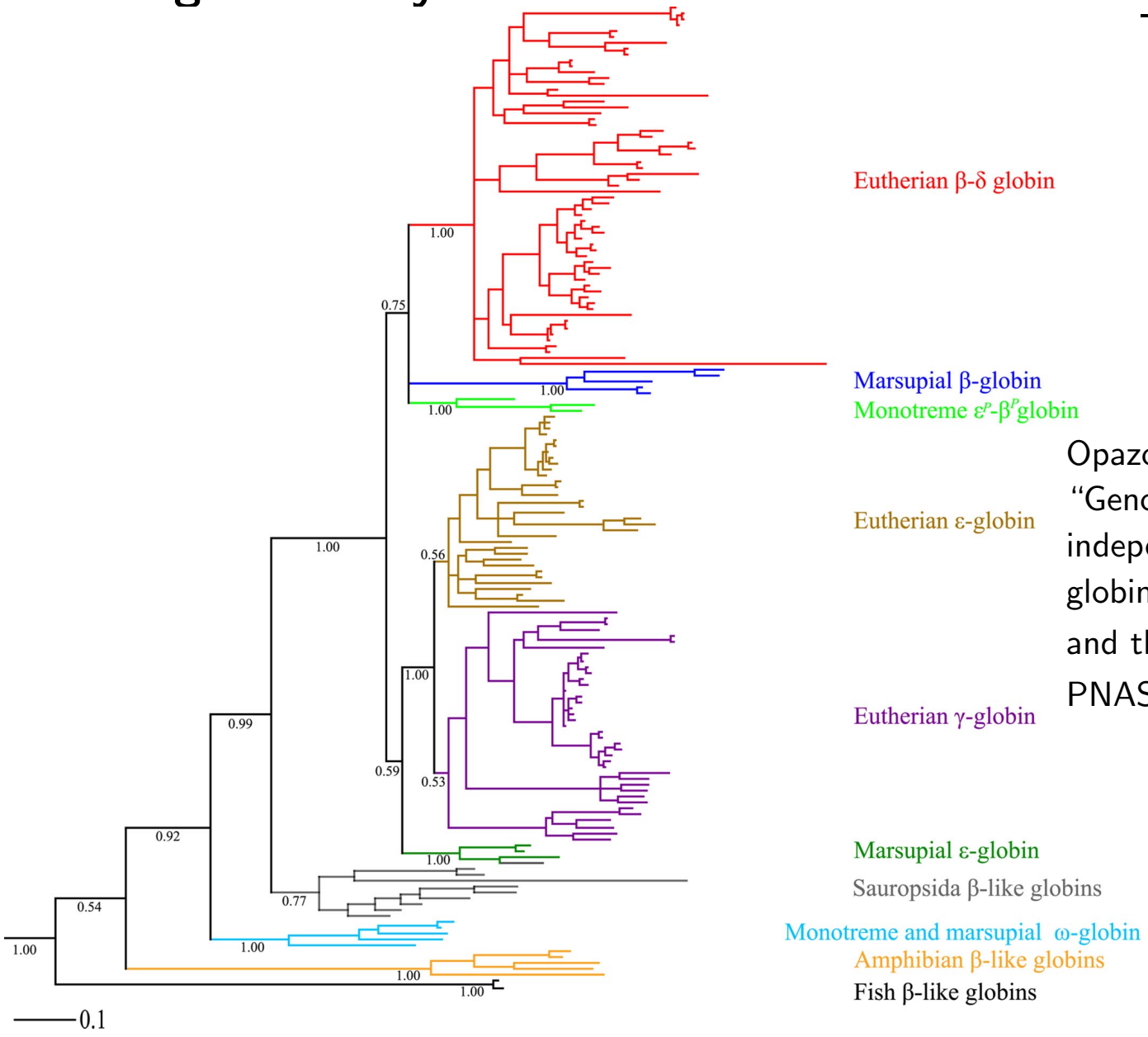
- "deep coalescence" or "incomplete lineage sorting" refer to the *failure* of gene copies to coalesce within the duration of the species – the lineages coalesce in an ancestral species

terminology: genealogical trees within population or species trees

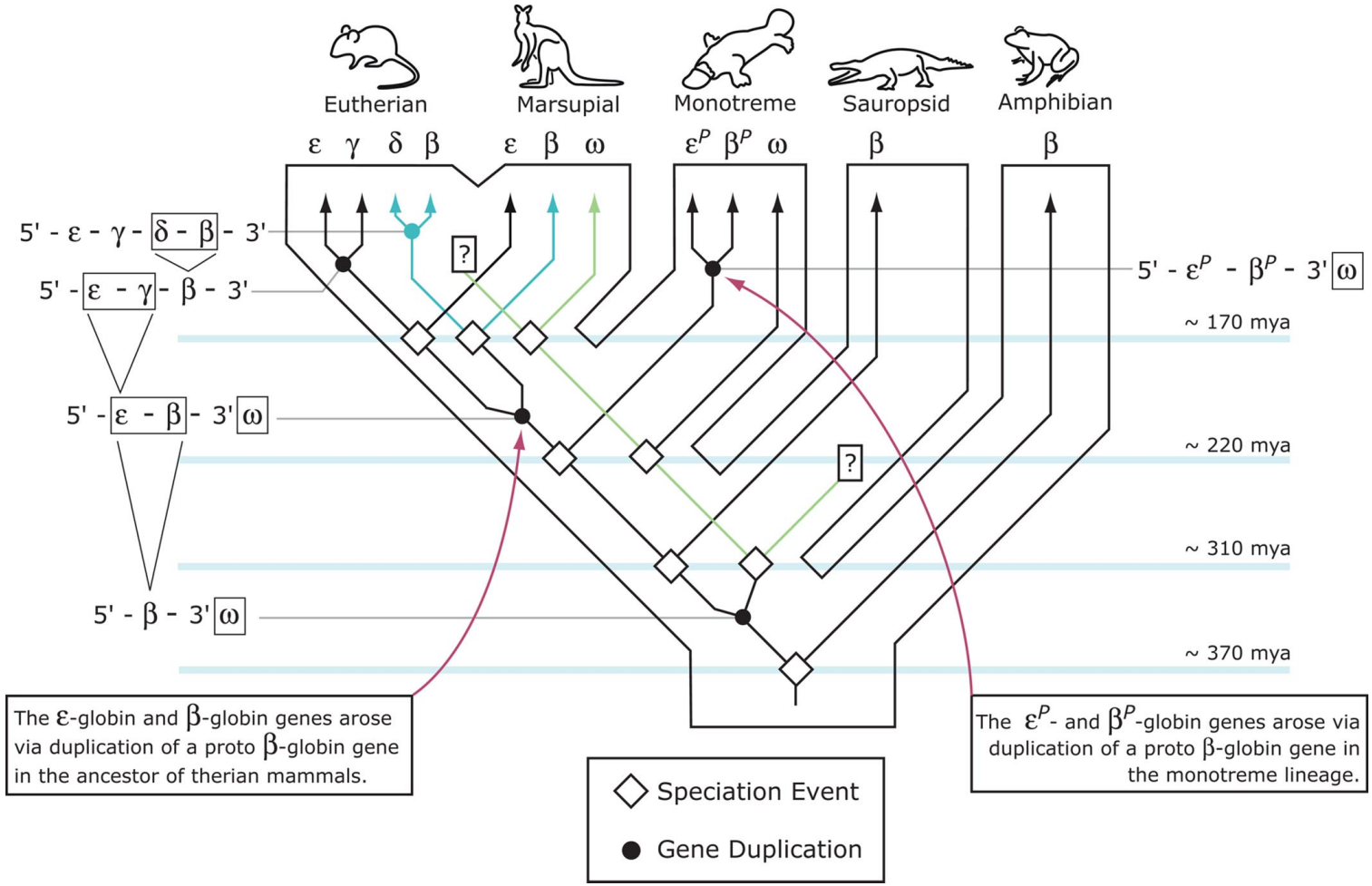
- coalescence – merging of the genealogy of multiple gene copies into their common ancestor. "Merging" only makes sense when viewed *backwards in time*.

- "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"



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

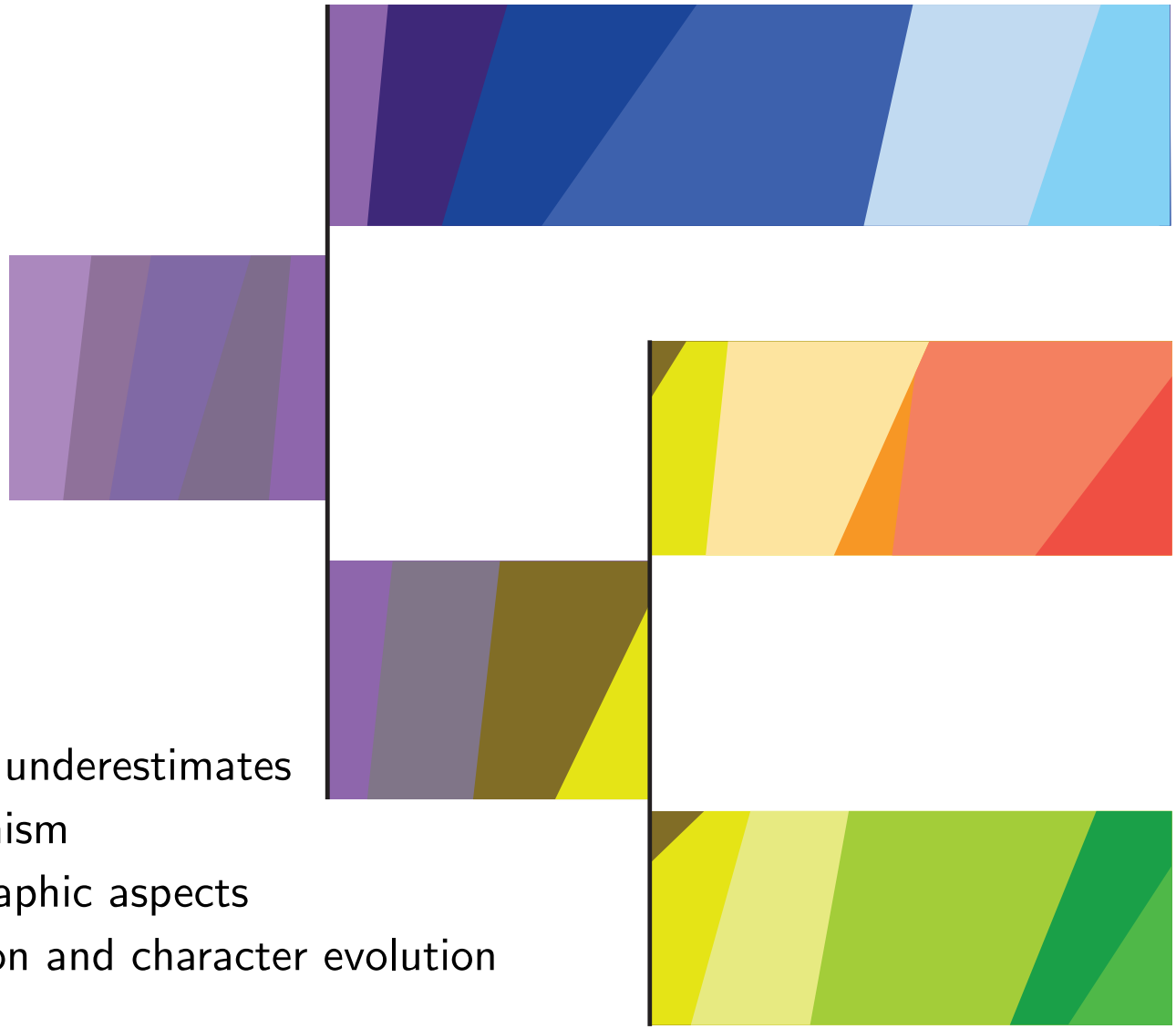
- duplication – the creation of a new copy of a gene within the same genome.
- homologous – descended from a common ancestor.
- paralogous – homologous, but resulting from a gene duplication in the common ancestor.
- orthologous – homologous, and resulting from a speciation event at the common ancestor.

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.

Phylogeny with complete genome + "phenome" as colors:

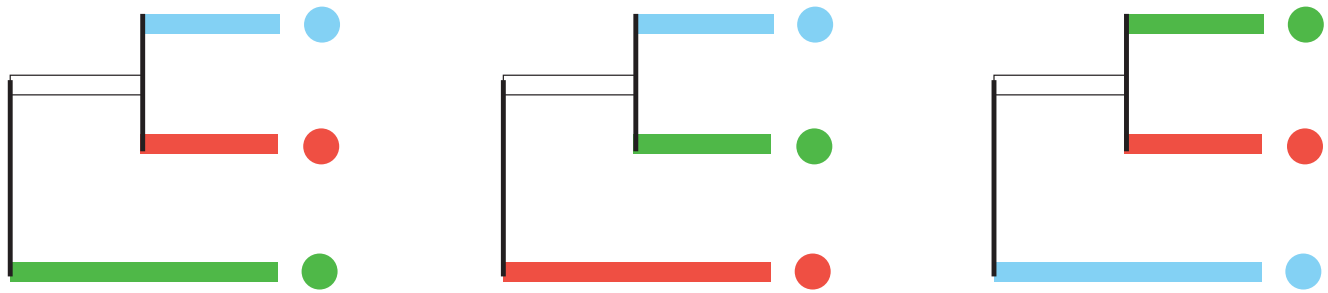


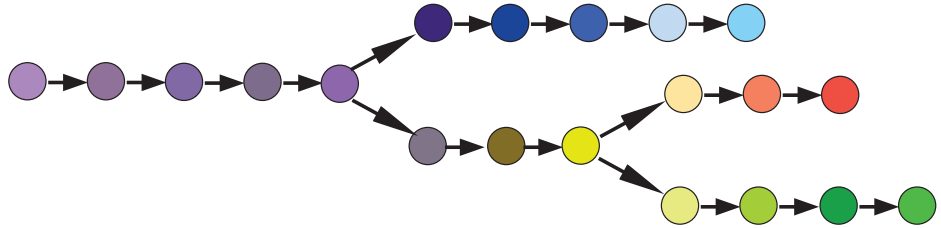
This figure:
dramatically underestimates
polymorphism
ignore geographic aspects
of speciation and character evolution

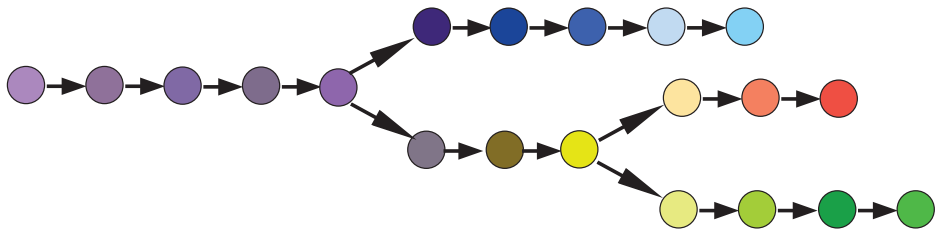
Extant species are just a thin slice of the phylogeny:

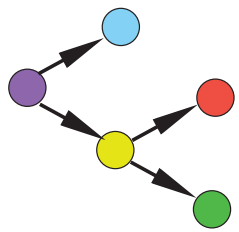
Our exemplar specimens are a subset of the current diversity:

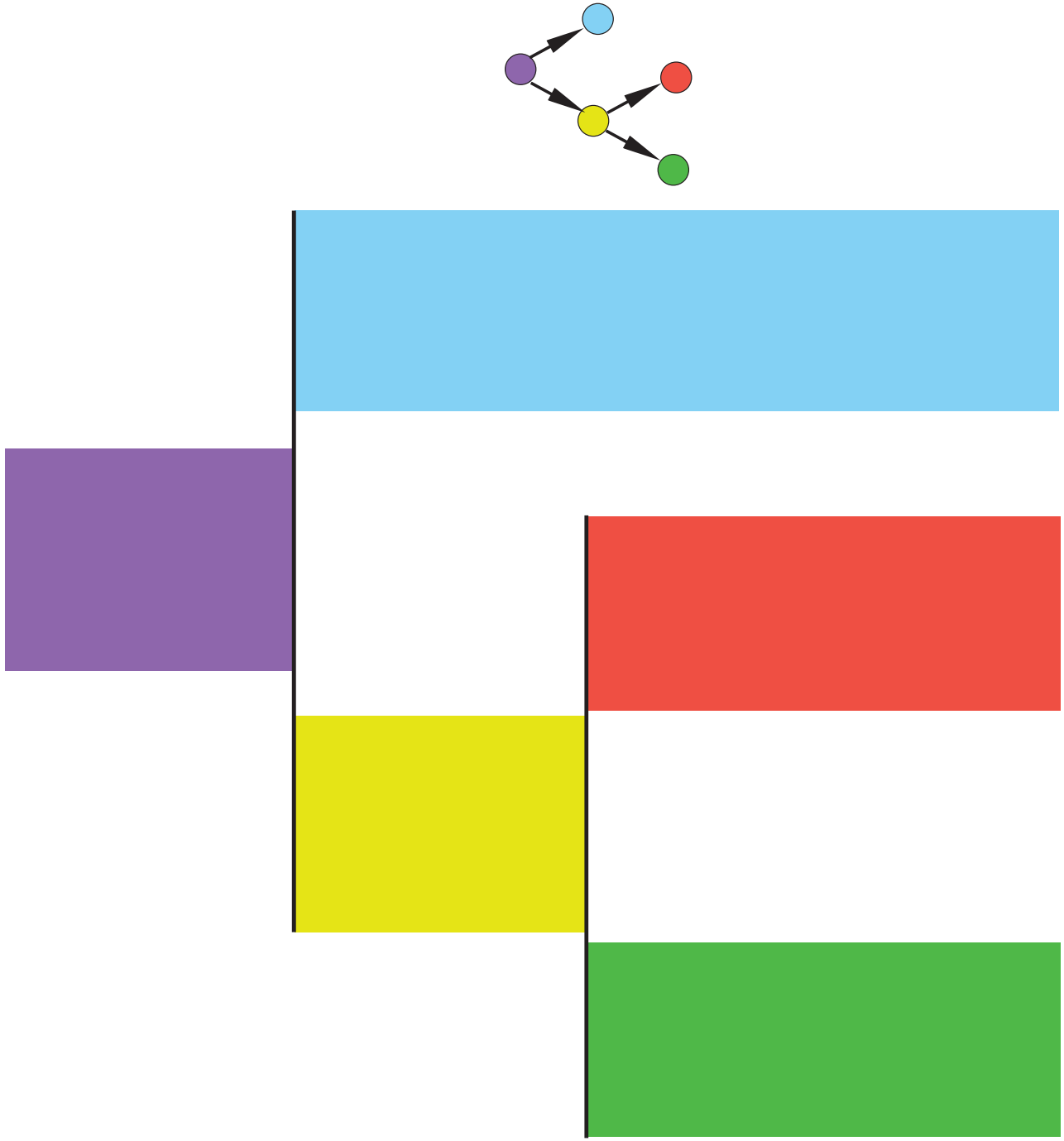
The phylogenetic inference problem:

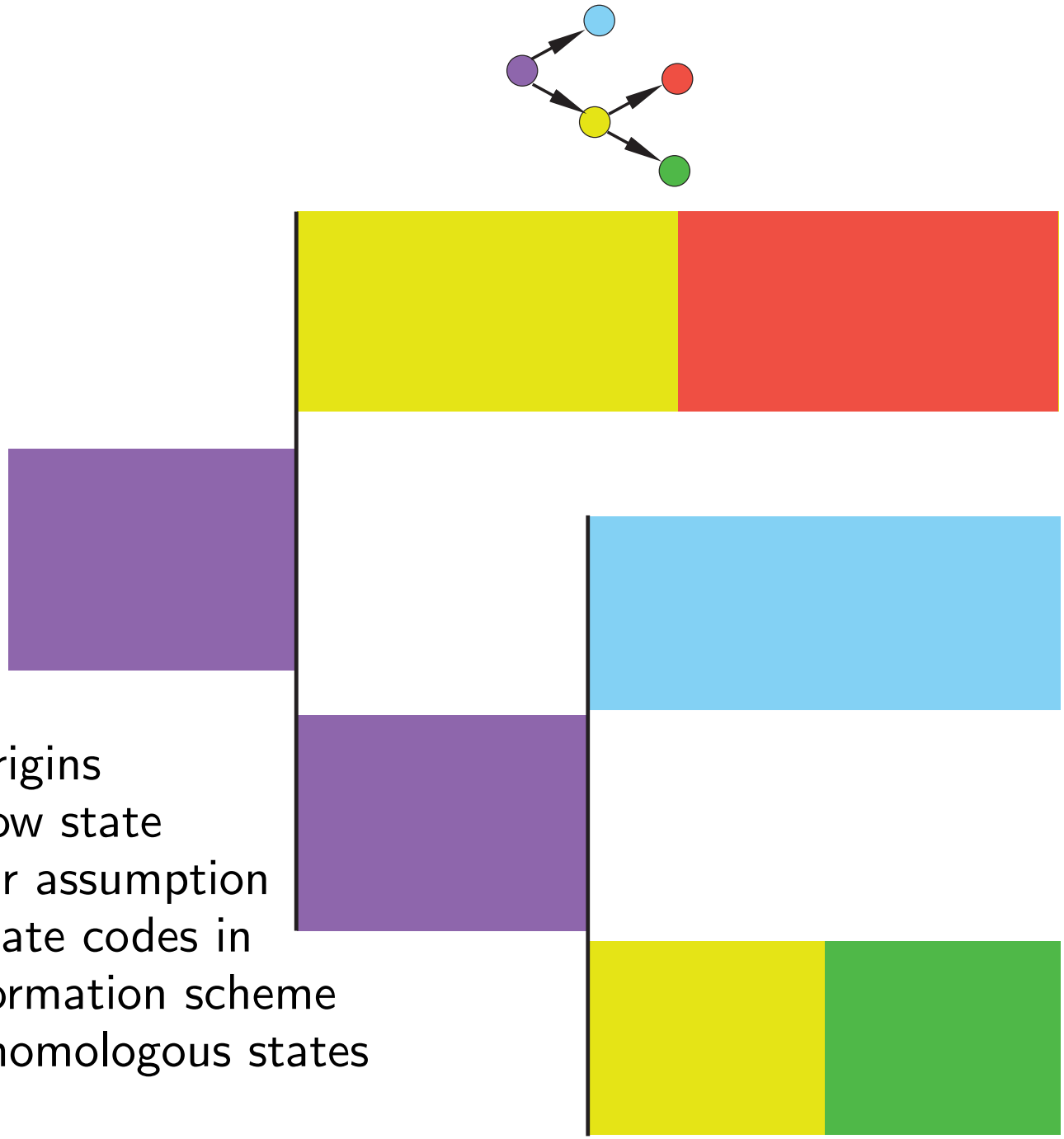




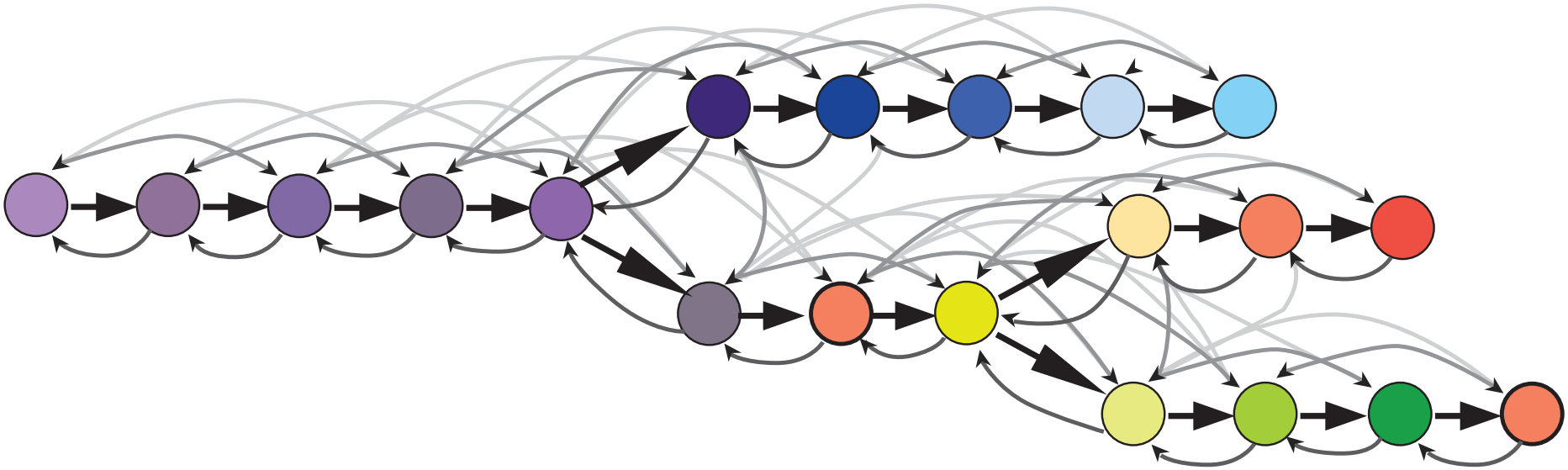








Multiple origins
of the yellow state
violates our assumption
that the state codes in
our transformation scheme
represent homologous states

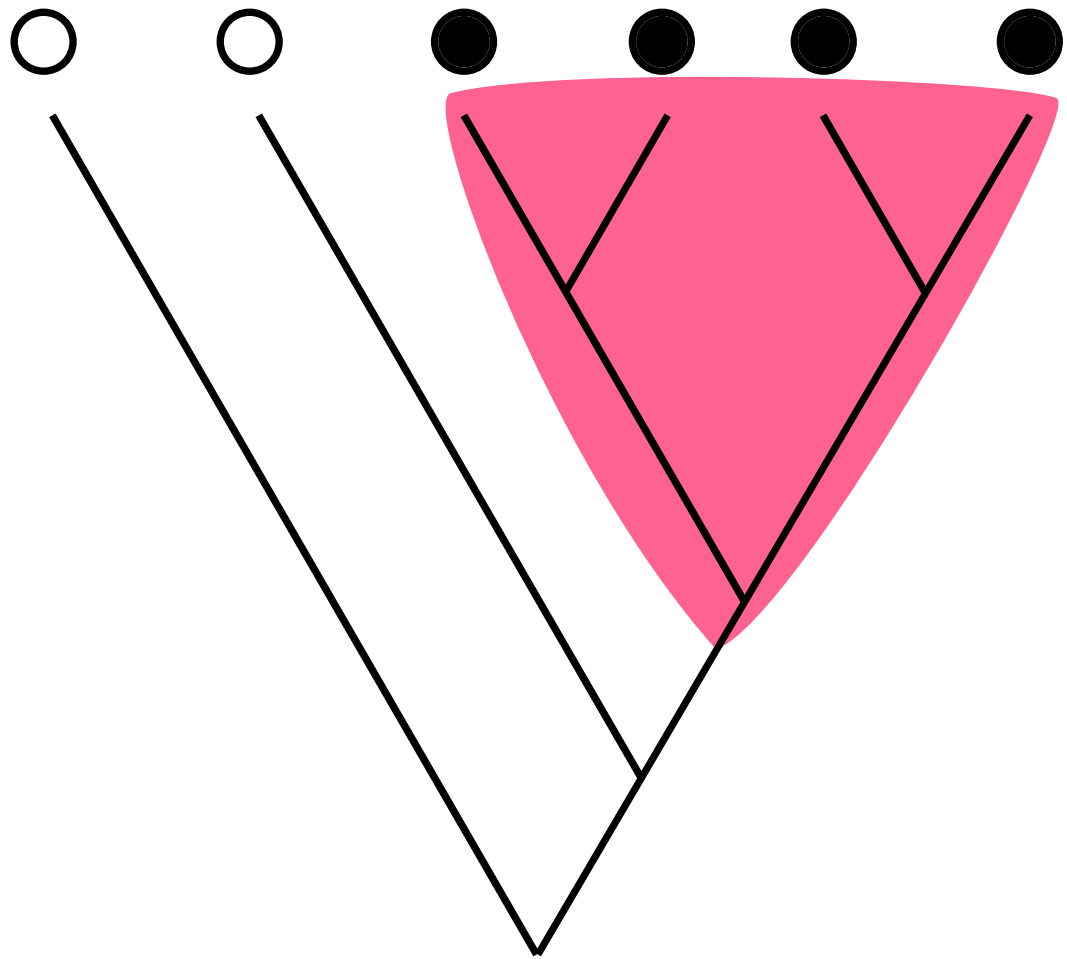


The meaning of homology (**very roughly**):

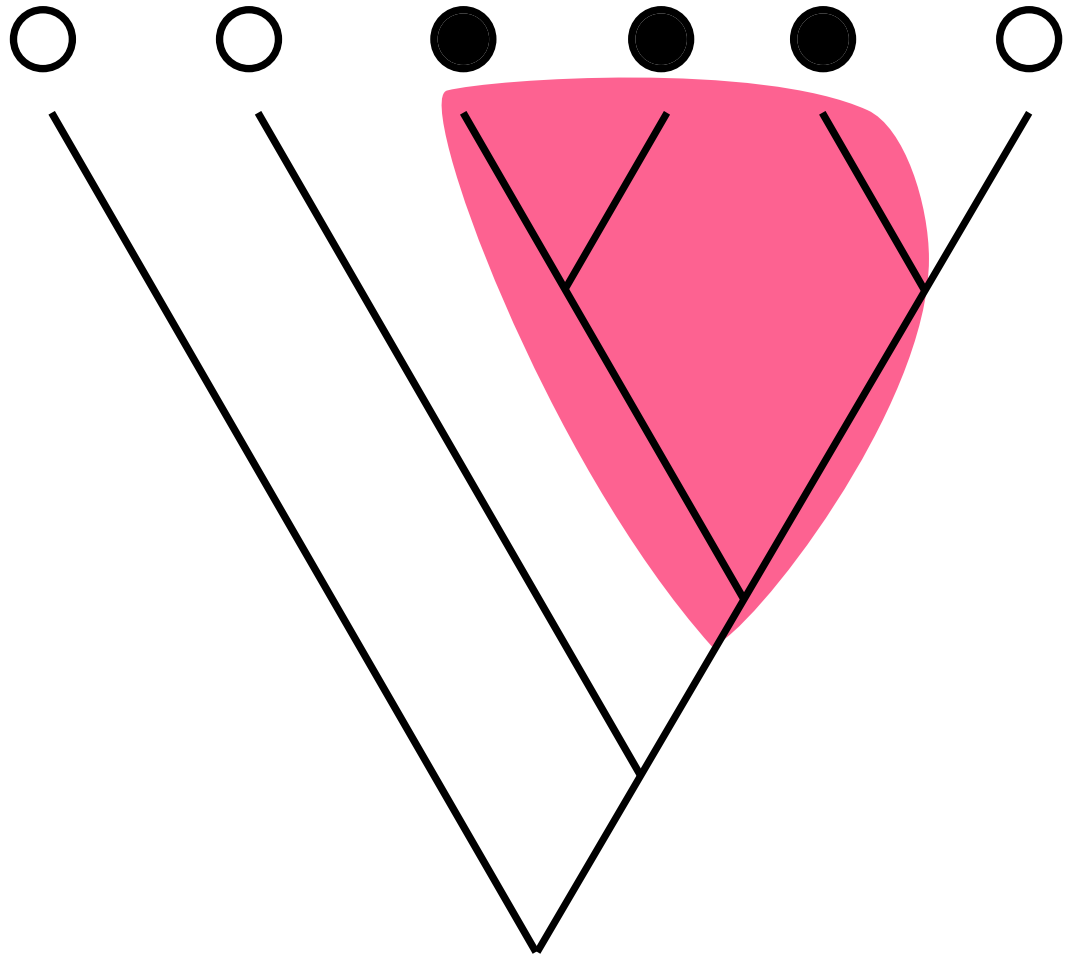
1. comparable (when applied to characters)
2. identical by descent (when applied to character states)

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

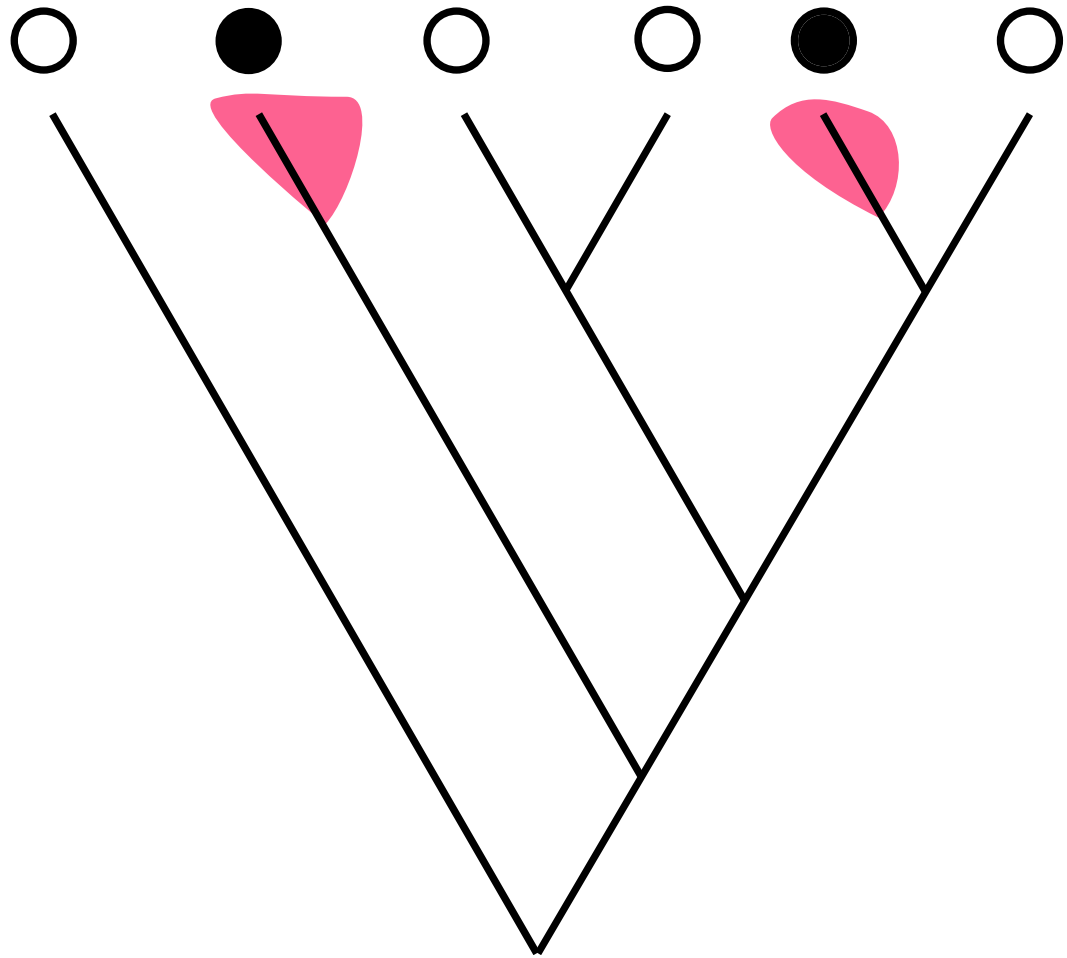
Instances of the filled character state are homologous
Instances of the hollow character state are homologous



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



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.(method suggested by Hennig)
Is this a valid rule?

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.

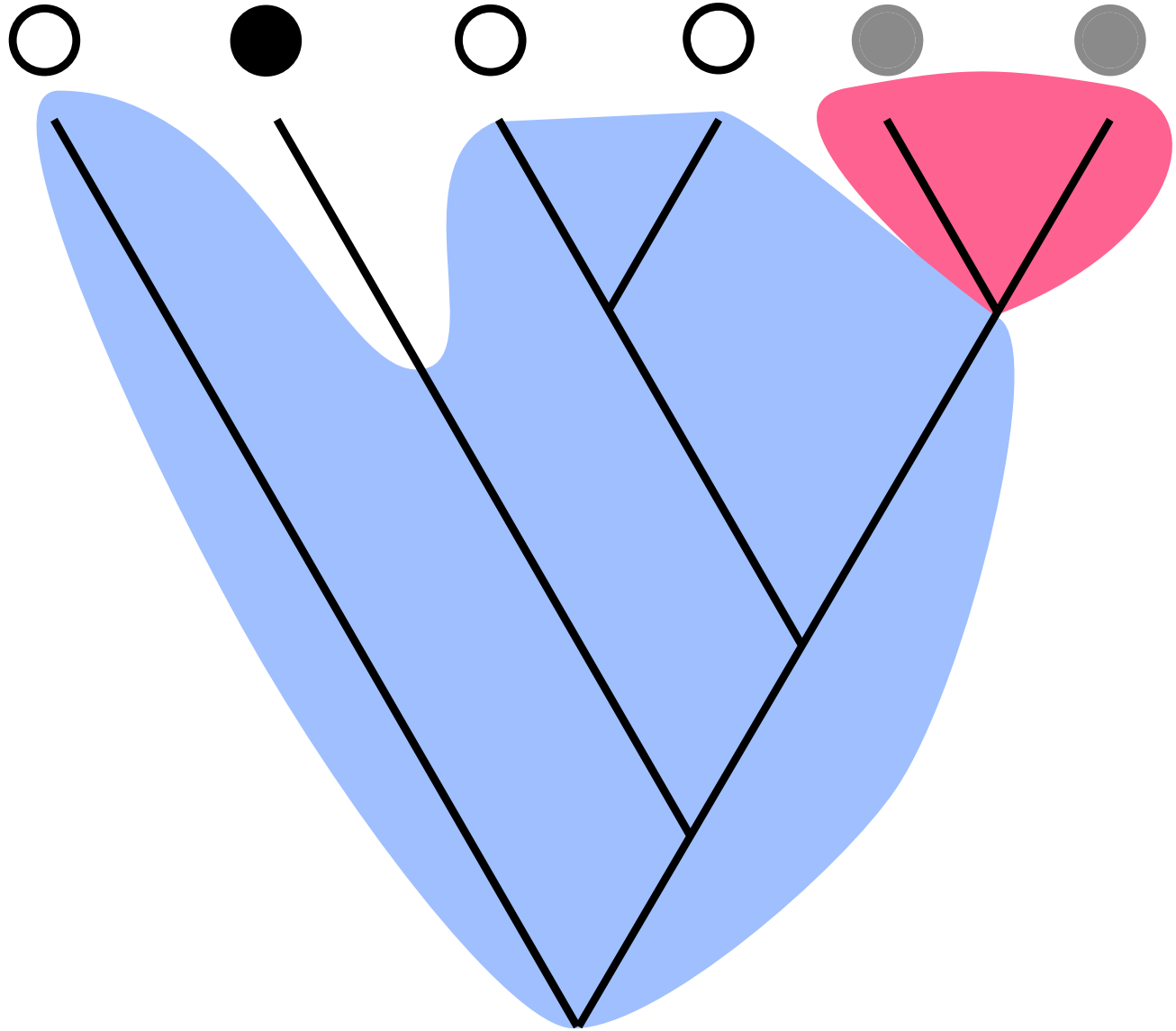
Hennigian terminology

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.



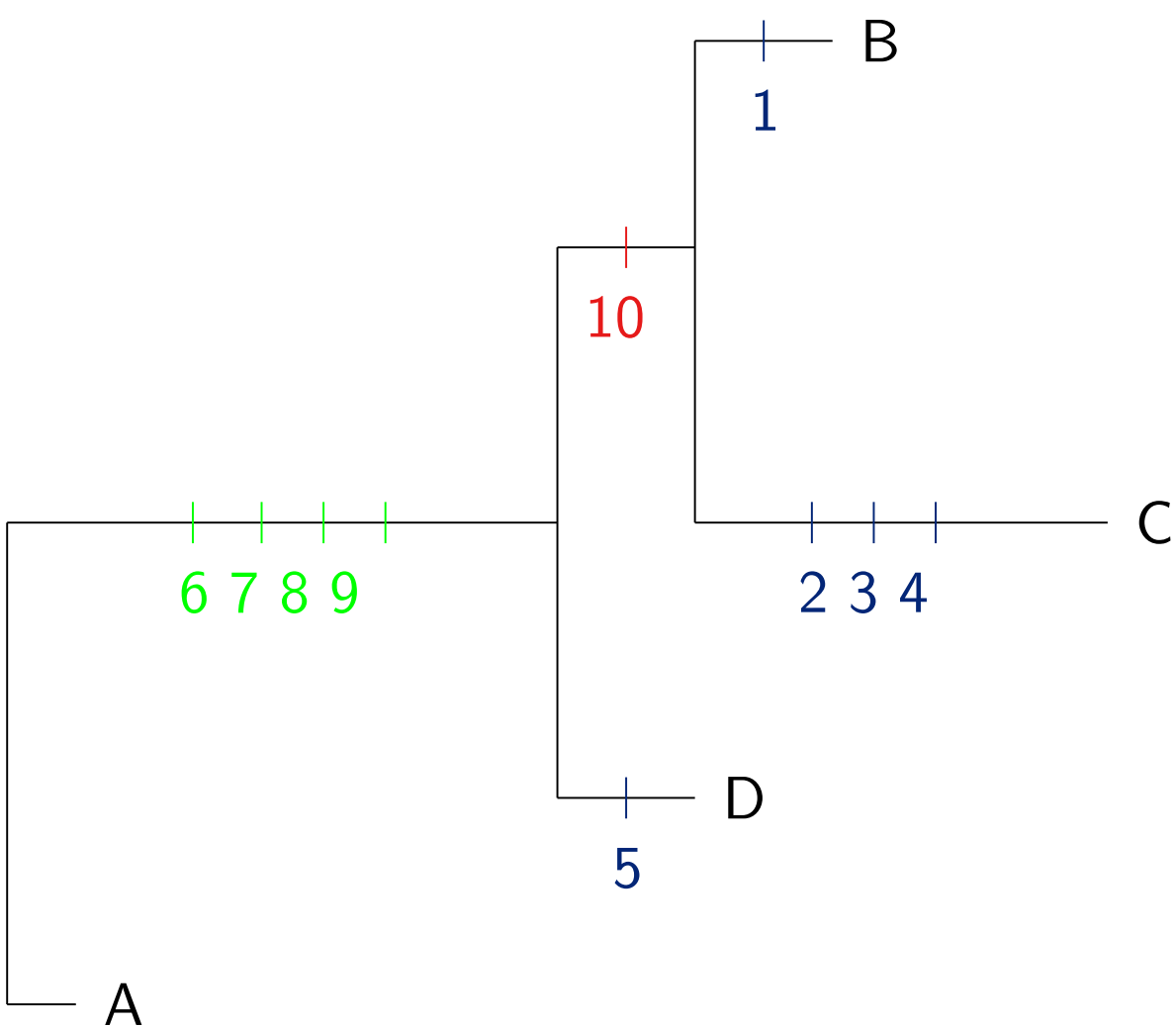
Hennigian inference

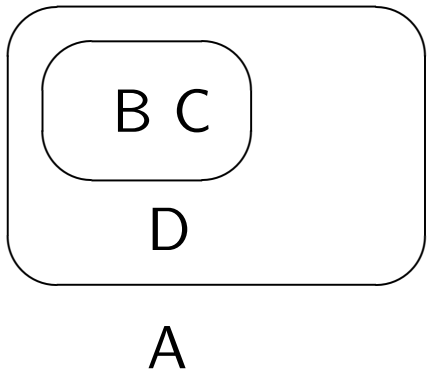
When we create a character matrix for Hennig's system, it is crucial that:

- 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

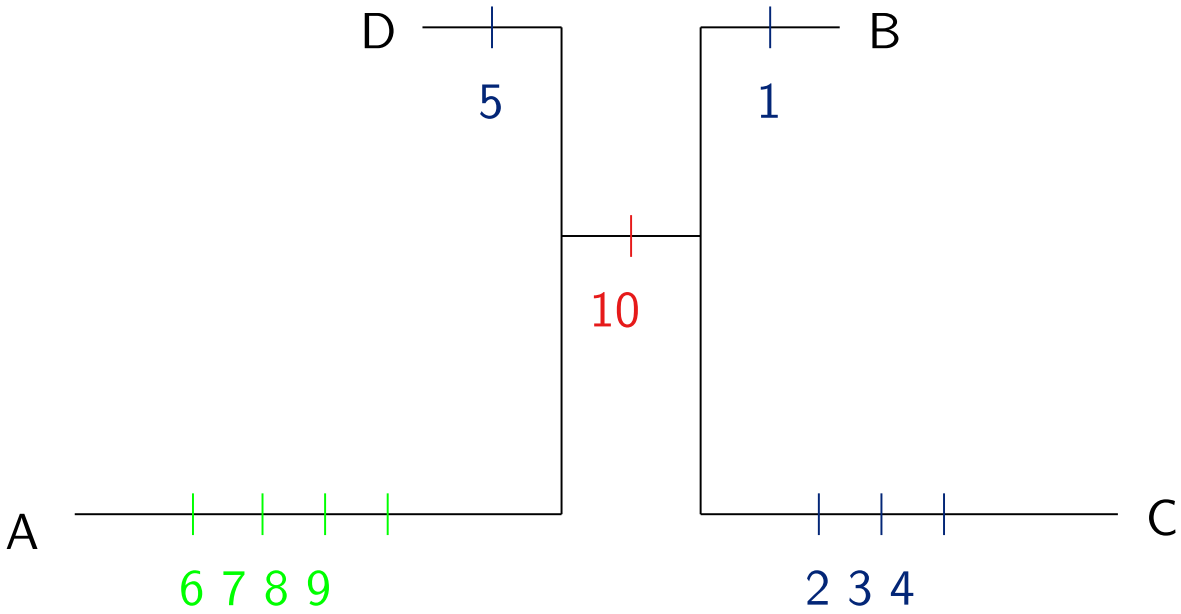

B
1
10
C
6 7 8 9
2 3 4
D
5
A


B C
D
A

If characters are not polarized (ancestral and descendent states known) method can infer unrooted trees.
Infer tree topology, but be unable to tell paraphyletic from monophyletic groups.

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.

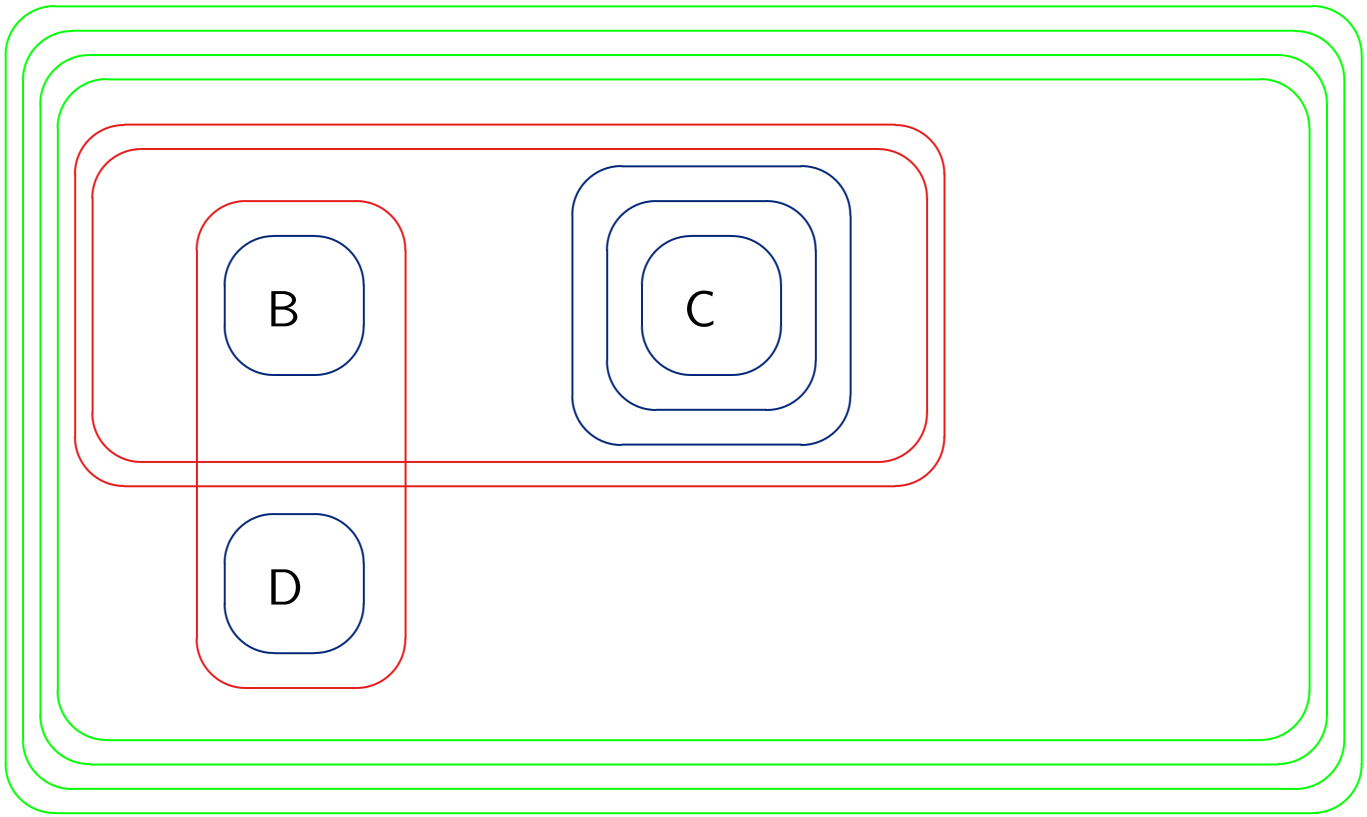
In fact, we almost always find that it is impossible for all of our premises to be true.

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

Character conflict

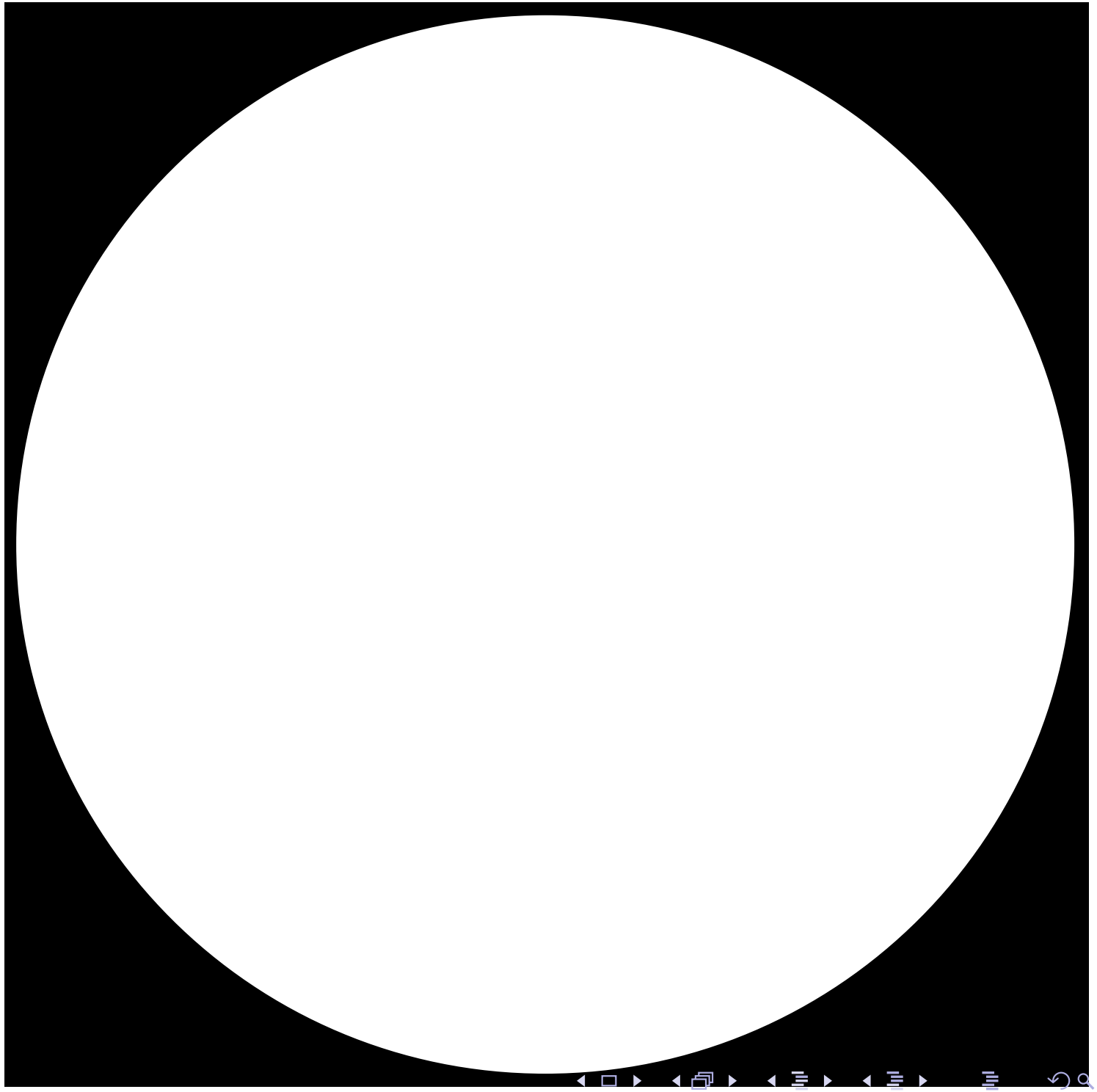
Two characters are compatible if they can both be mapped on the same tree so that all of the character states displayed could be homologous.

Incompatible characters are evidence of *homoplasy* in the data

Homoplasy literally means the "same change" has occurred more than once in the evolutionary history of the group.

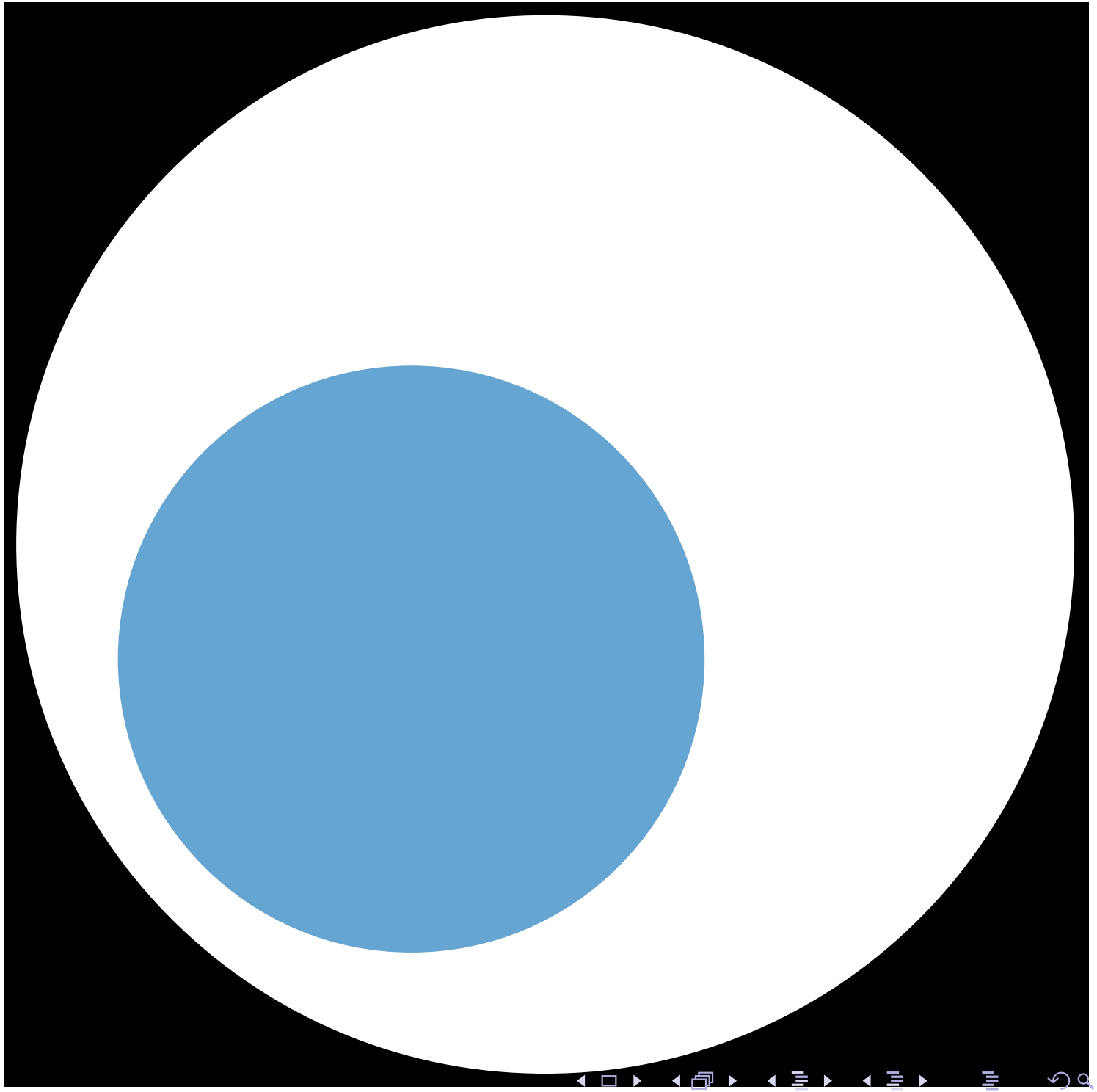
The presence of homoplasy undermines Hennigian analyses.

white = space of all possible matrices



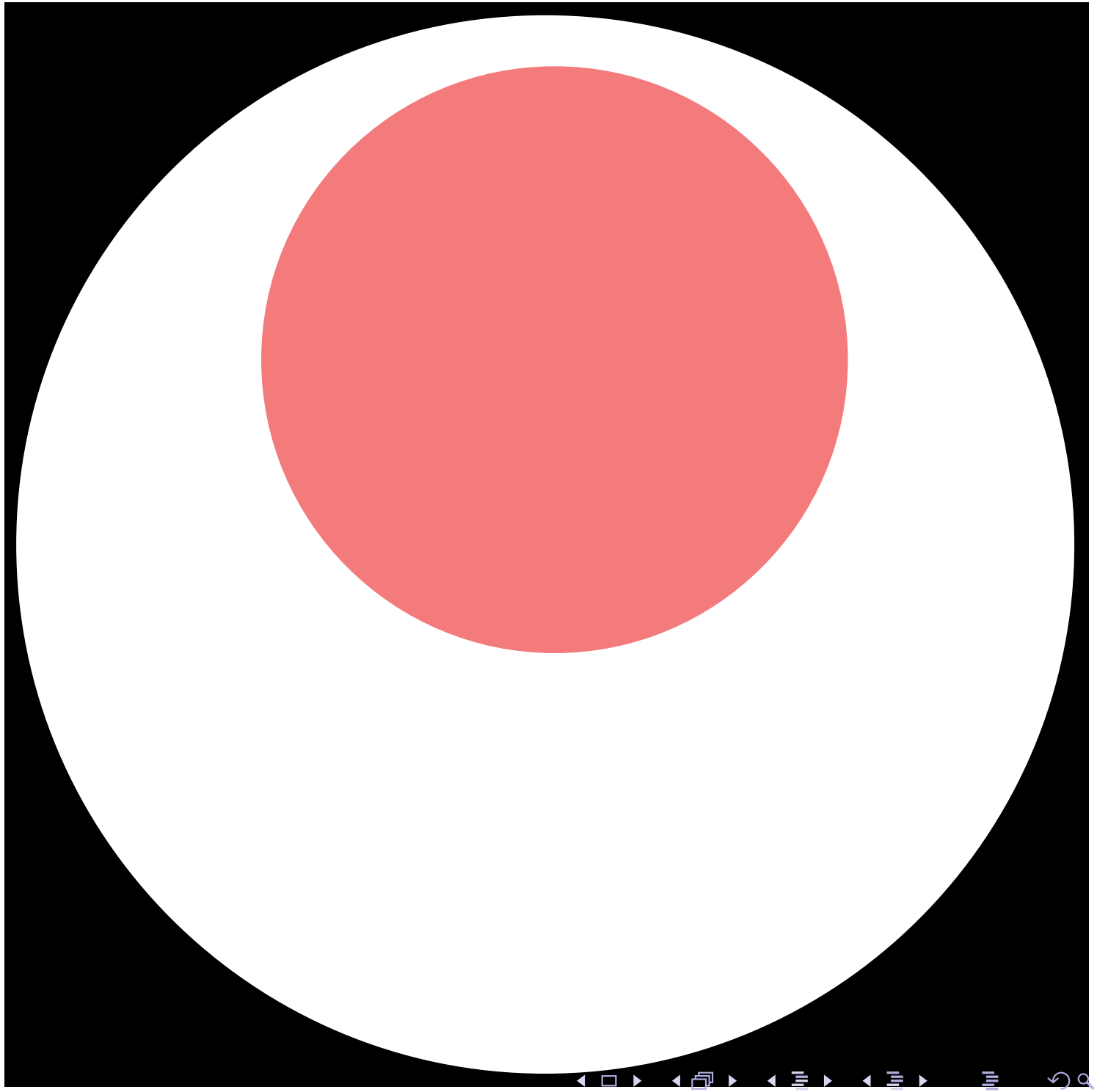
blue = space of matrices with the pattern:

```
A B C D
- * * -
```



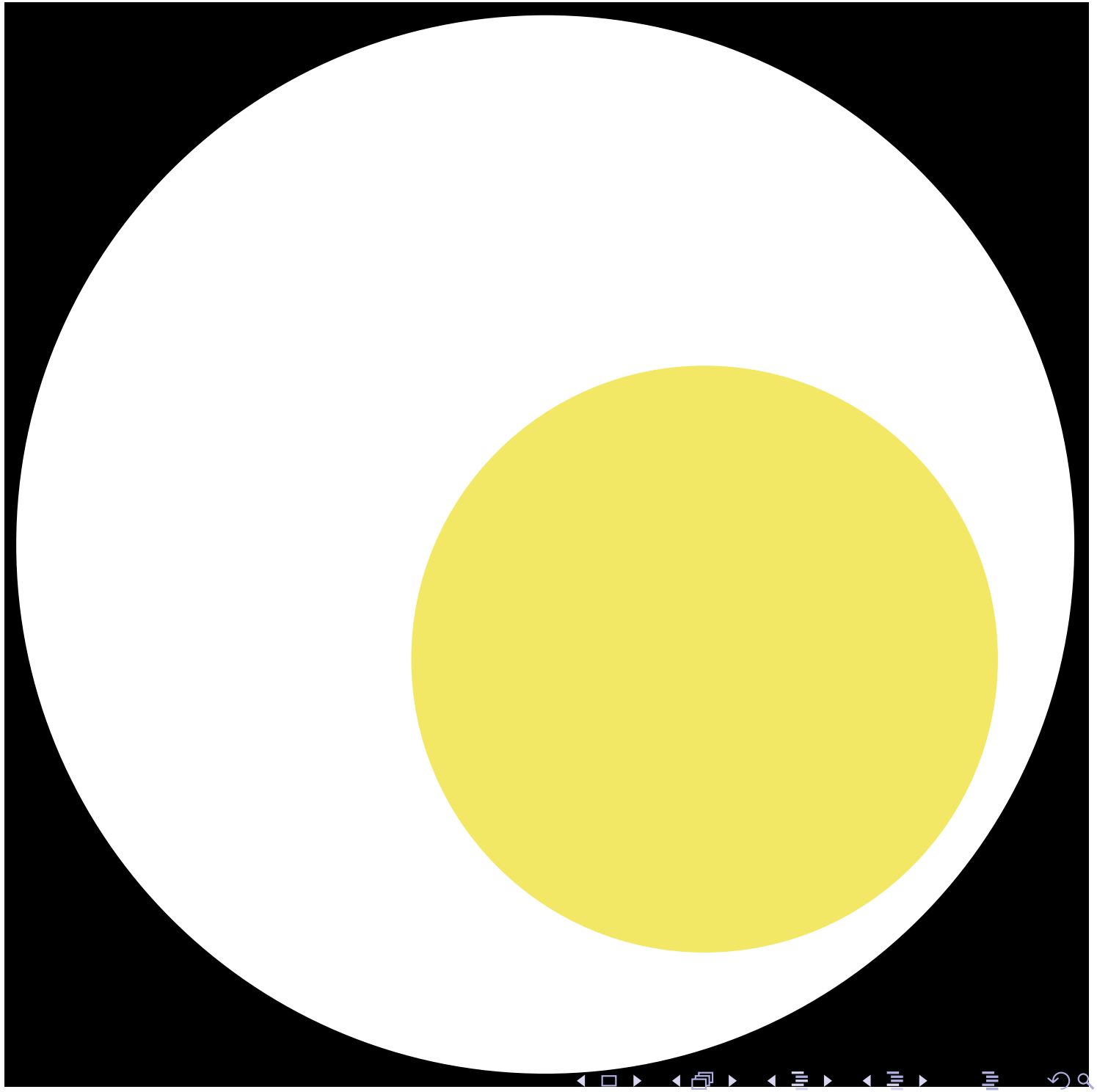
red = space
of matrices with
the pattern:

```
A B C D
- * - *
```



yellow = space of matrices with the pattern:

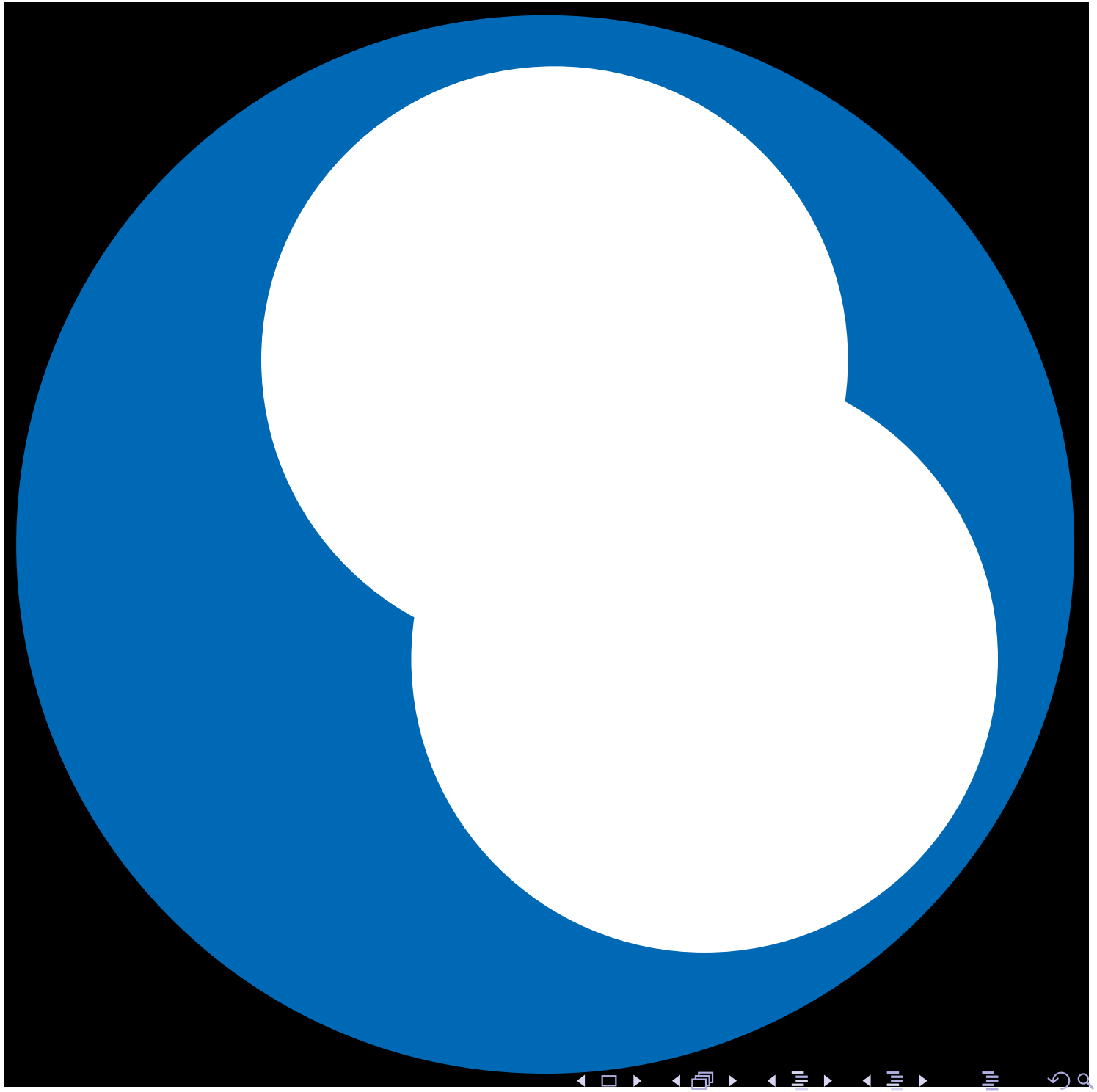
```
A B C D
- - * *
```



all eight categories of matrices

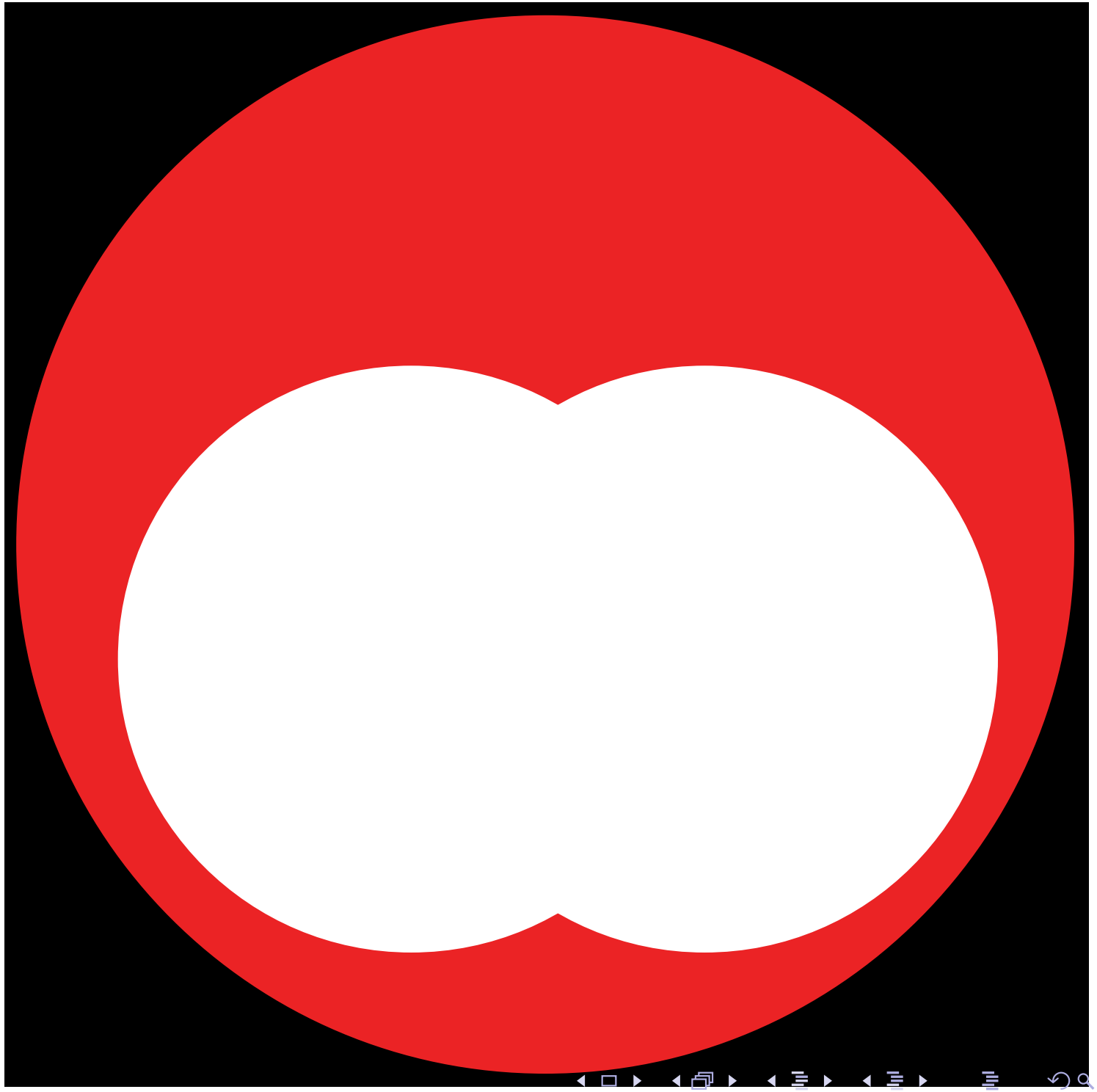
blue = space of matrices compatible with tree:

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



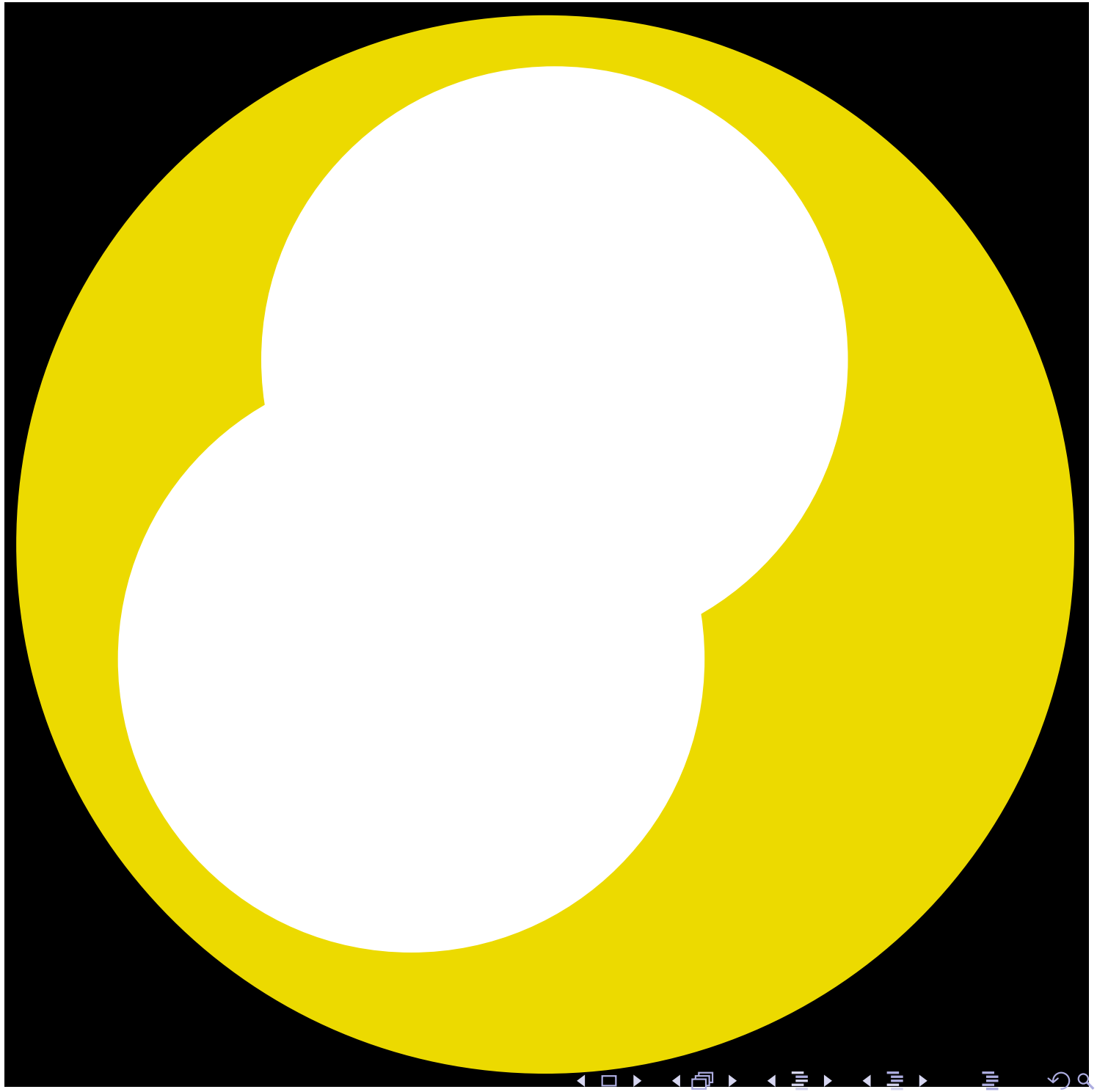
blue = space of matrices compatible with tree:

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



blue = space of matrices compatible with tree:

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



Hennigian:

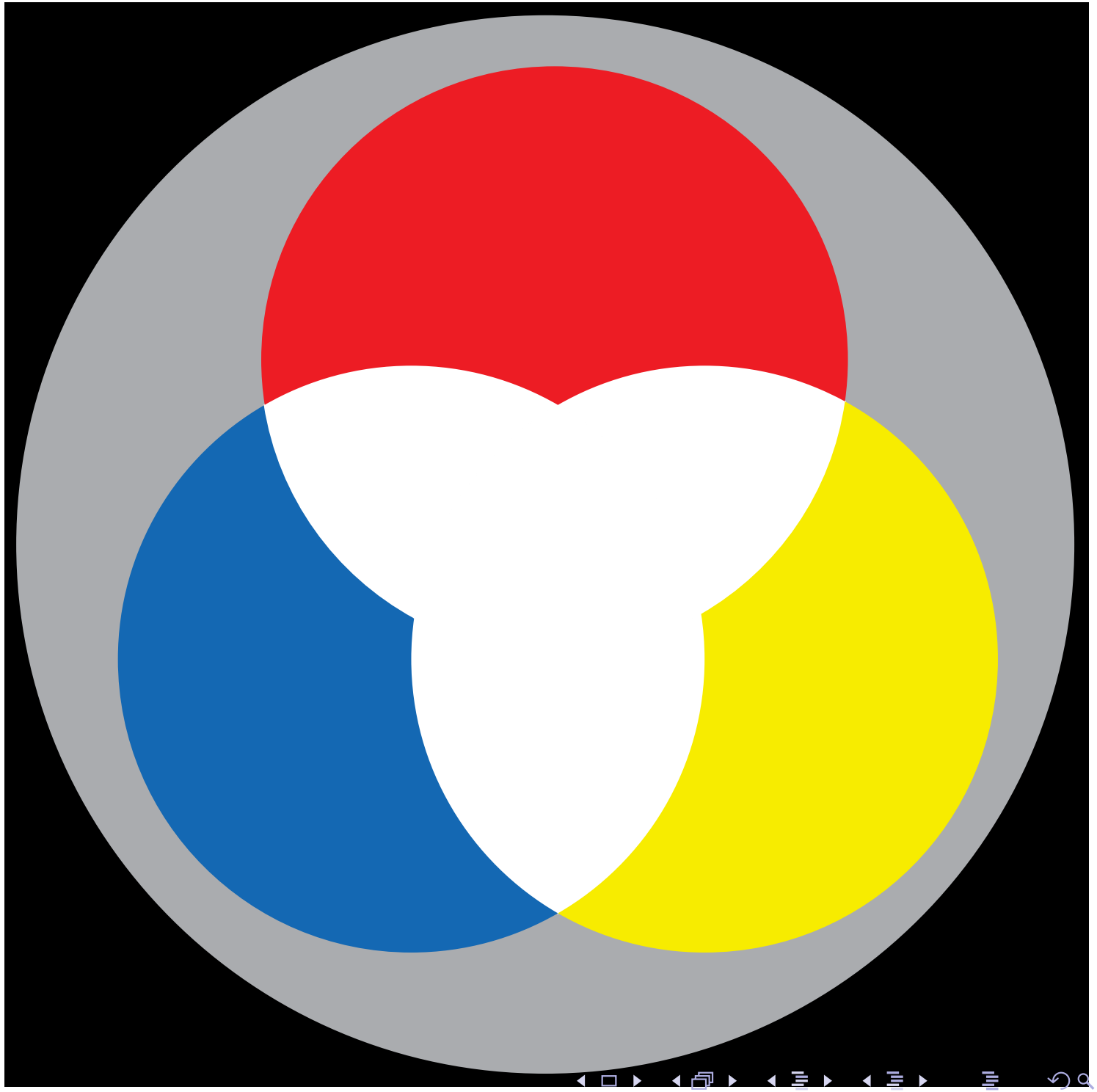
grey = any tree

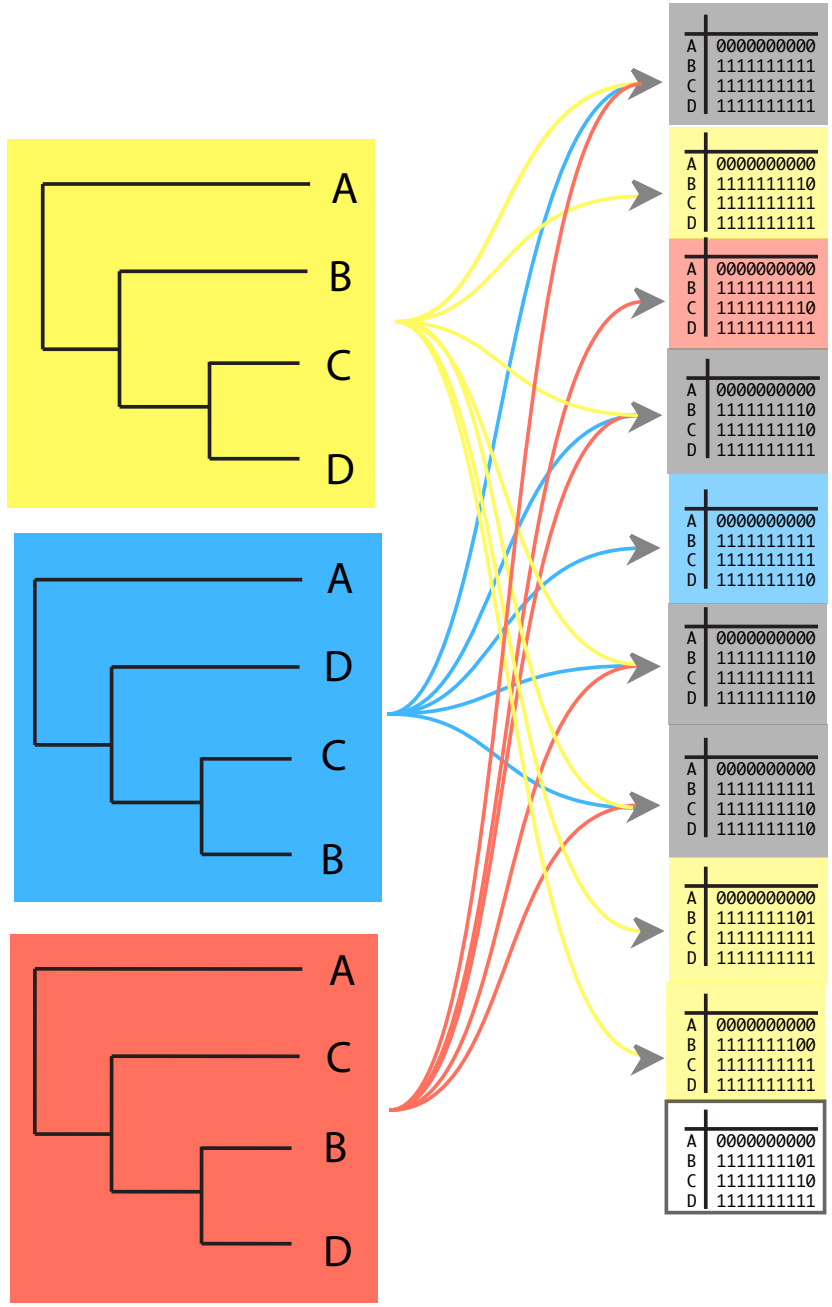
blue = B+C

red = B+D

yellow = C+D

white = no tree (conflicting characters)




A
B
C
D
A
D
C
B
A
C
B
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

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- can we detect character conflict?
- is there a logic-based solution to the problem of character conflict?

Detecting character conflict in binary characters

Consider the four possible combinations of states in a two-character matrix.

The characters are incompatible *iff* (when you look across all taxa) you see all four state combinations.

		Char 1	
		0	1
Char 2	0	×	×
	1	×	×

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- Can we detect character conflict? Yes
- Is there a logic-based solution to the problem of character conflict?
  - recoding characters?
  - "reciprocal illumination"?

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- Can we detect character conflict? Yes
- Is there a logic-based solution to the problem of character conflict? No, nothing purely based on logic (and the suggestions for culling data to make matrices suitable for logical inference can lead to unsatisfyingly subjecive analyses).
- What can we do?

We must have an "error model"

Next class - statistical estimators for phylogenetics.