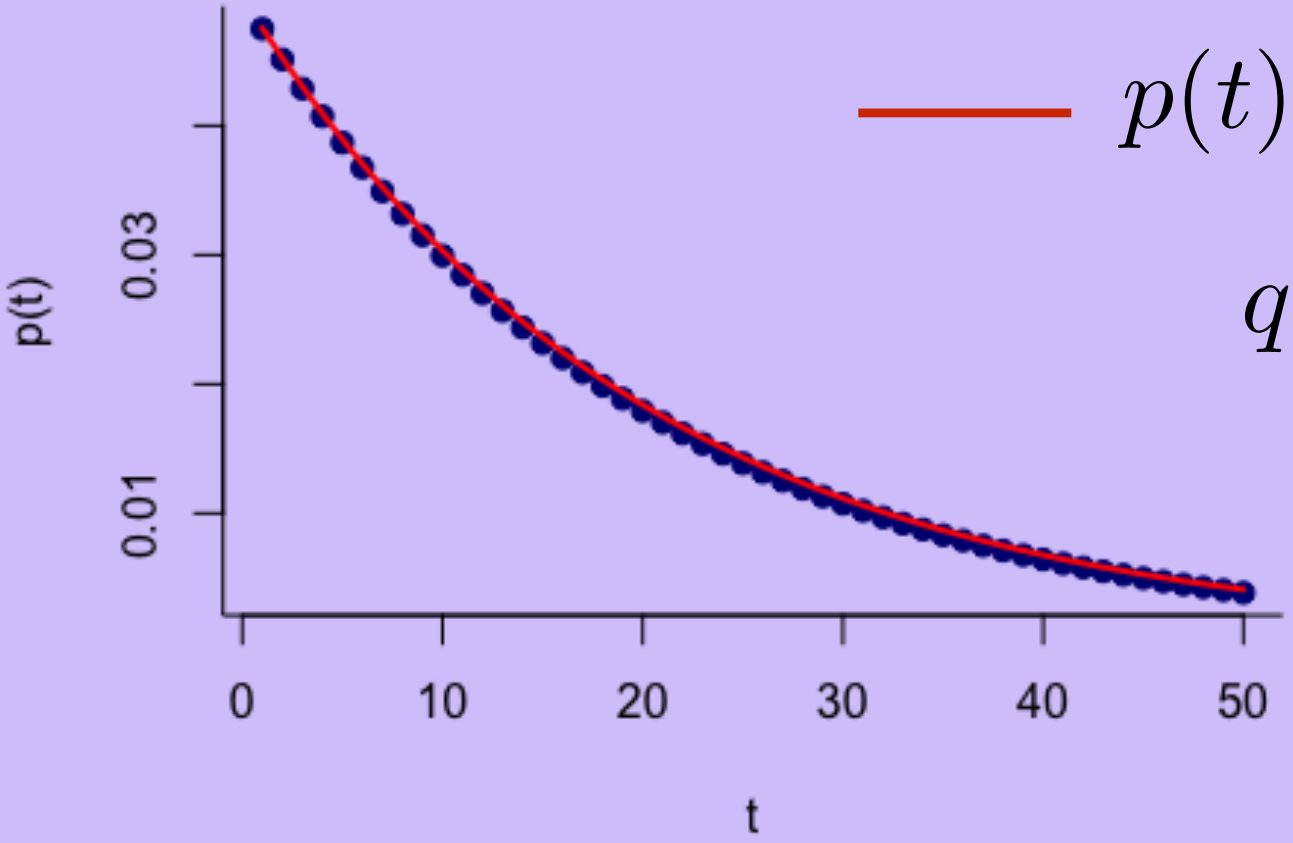
Discrete vs Continuous Coalescence Probabilities

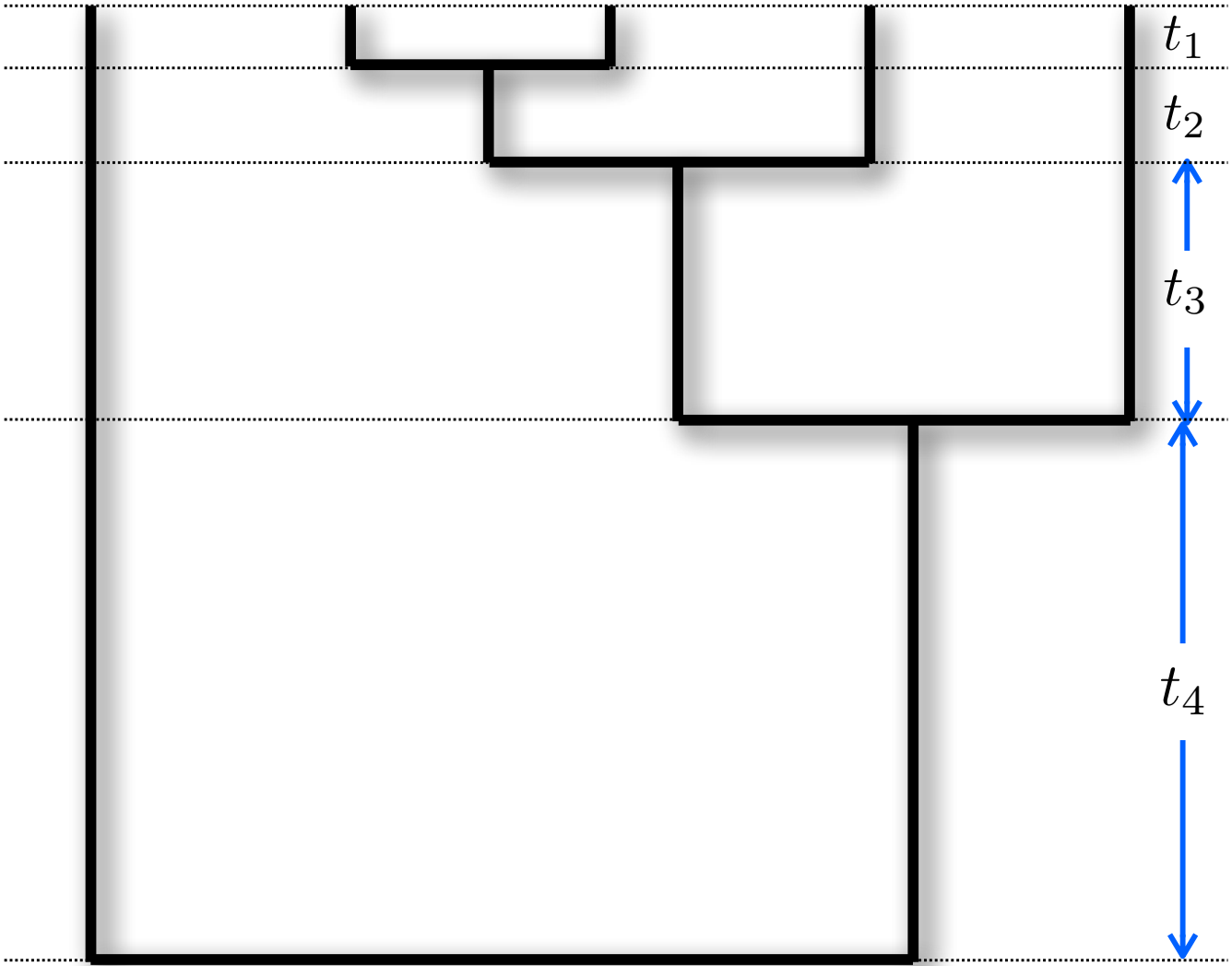


- $p(t) = q(1-q)^{t-1}$

—— $p(t) = \lambda e^{-\lambda t}$

$$q = \lambda = \frac{\binom{n}{2}}{2N}$$

Simulating coalescence

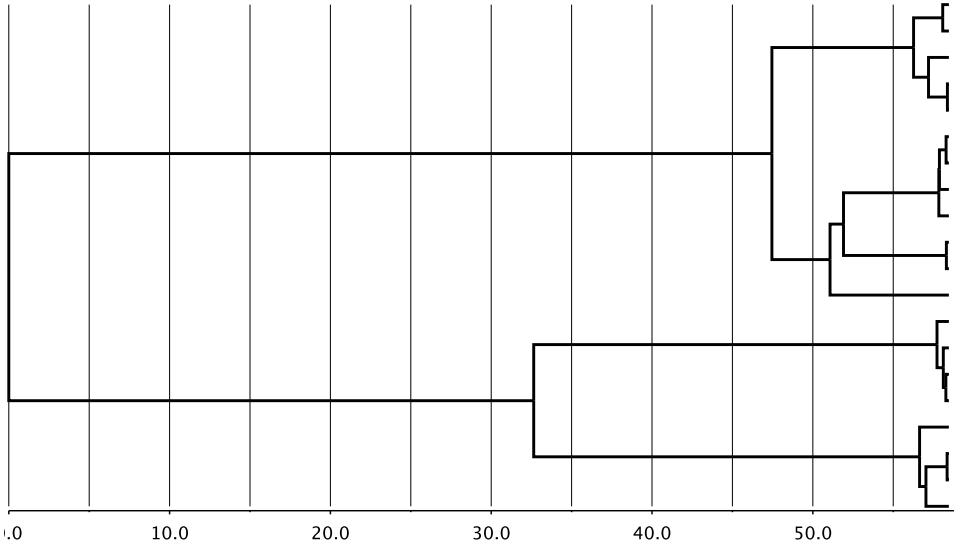


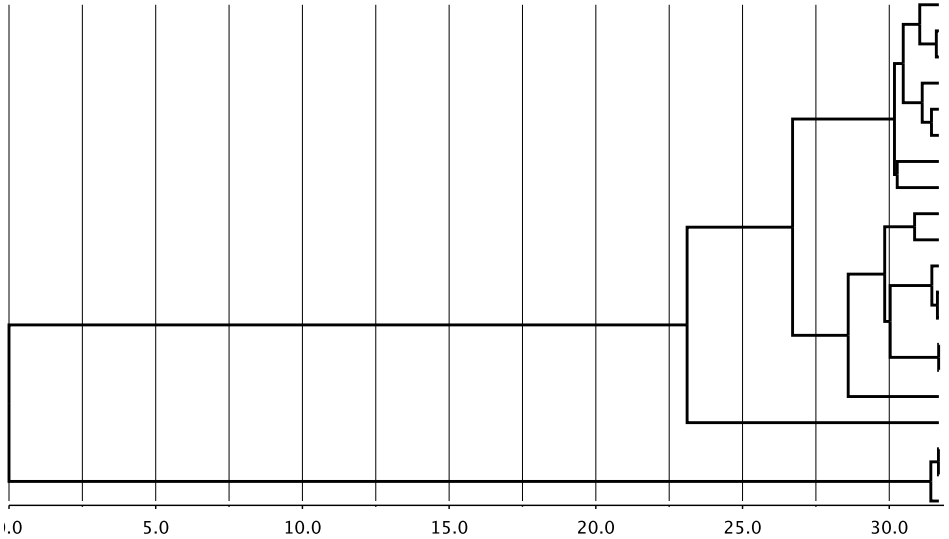
$$E[t_1] = 2N/10$$
$$E[t_2] = 2N/6$$
$$E[t_3] = 2N/3$$
$$E[t_4] = 2N/1$$

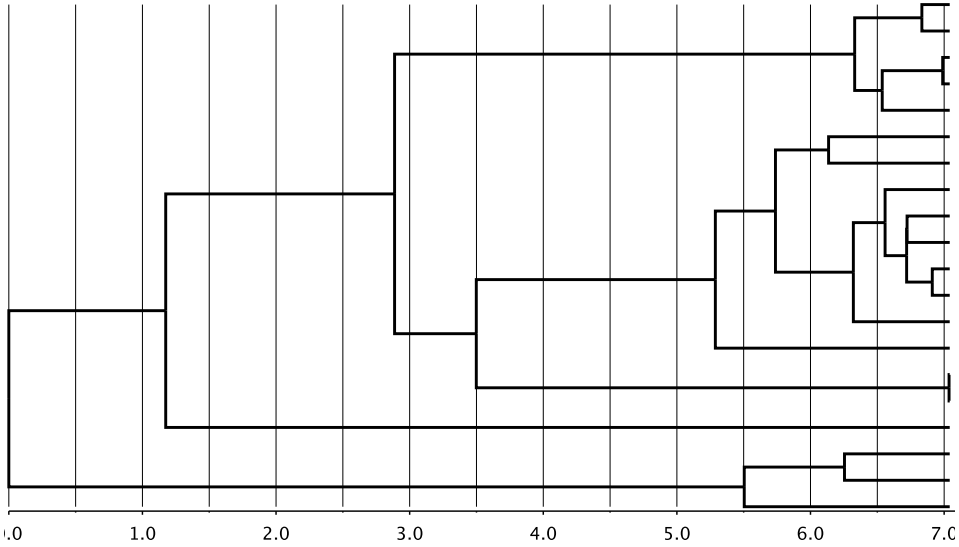
Uniform(0,1)

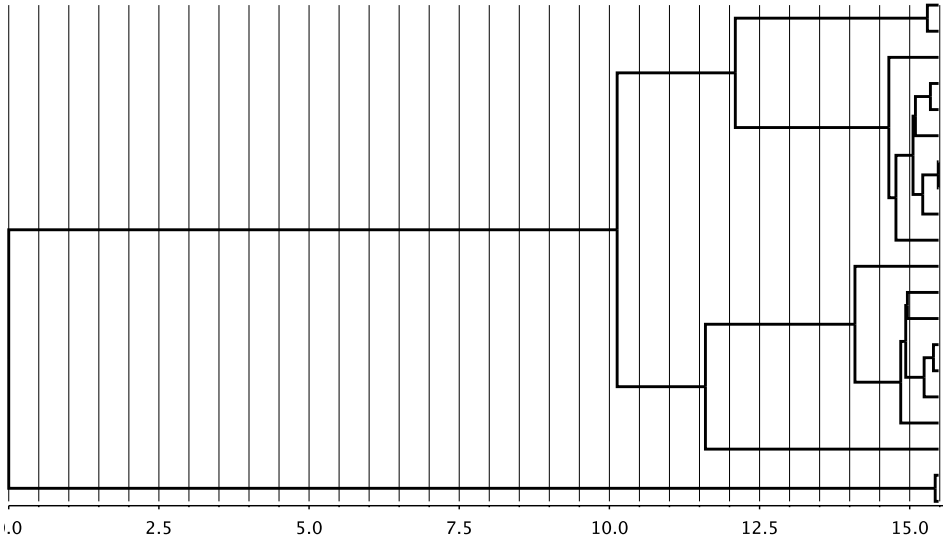
Exponential(λ) $\longrightarrow t = -\log(u)/\lambda$

Examples of coalescent trees









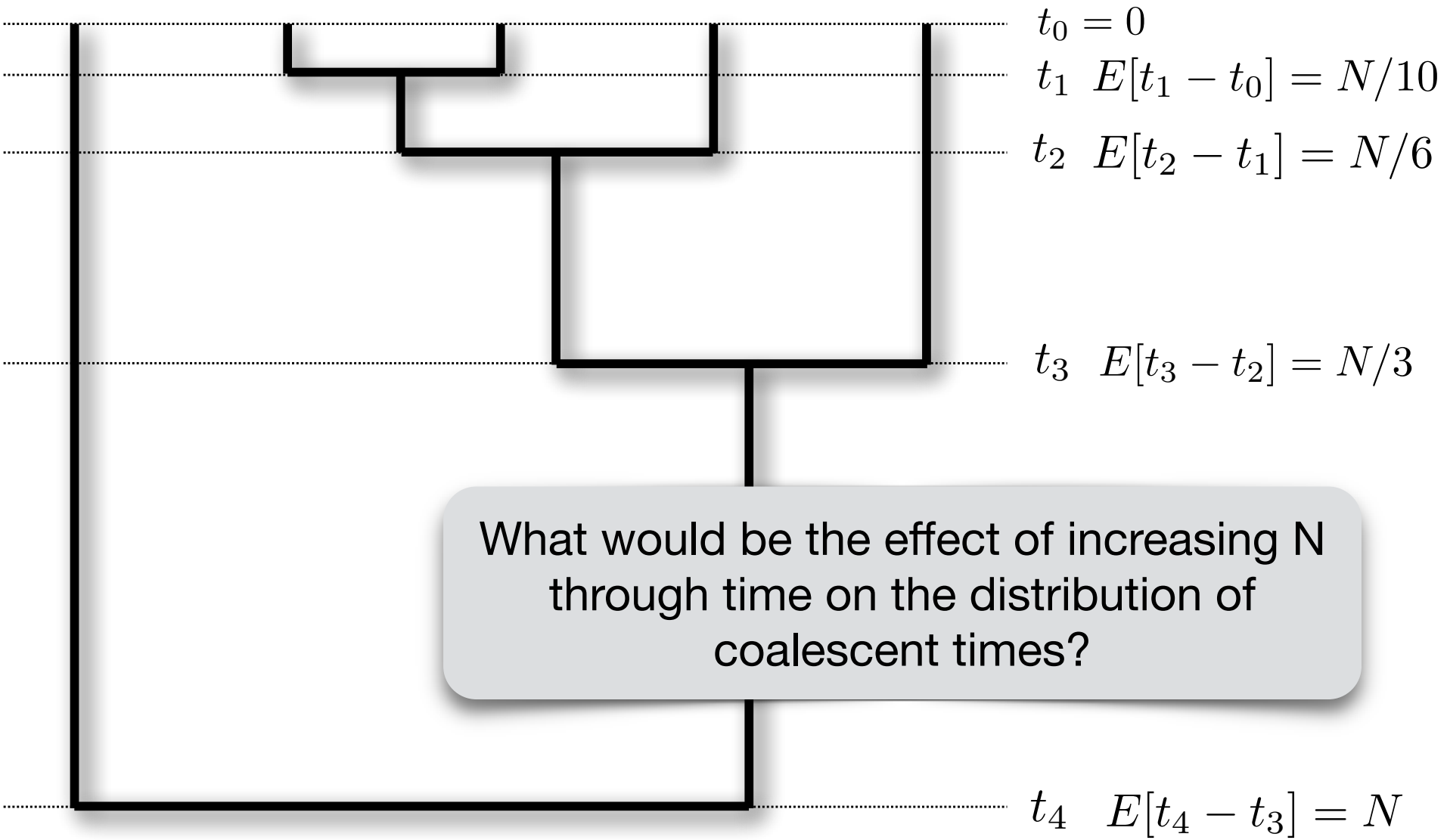
depth of trees simulated with same N quite variable

Bugs in a box

written by Peter Beerli

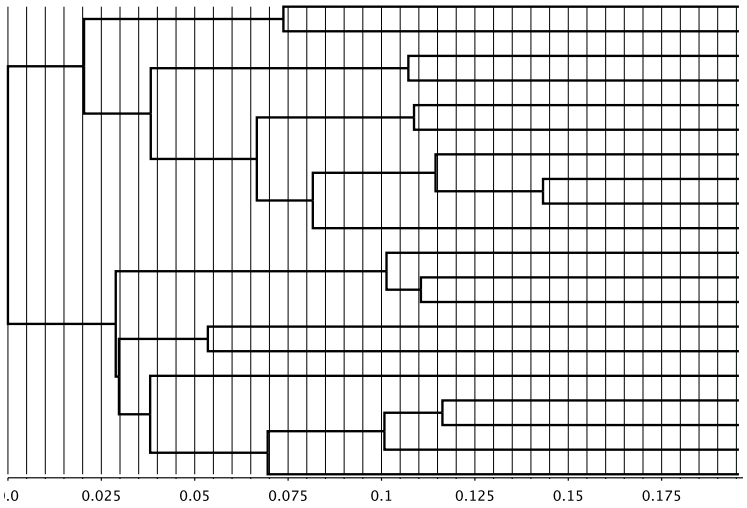
https://github.com/pbeerli/bugsinbox

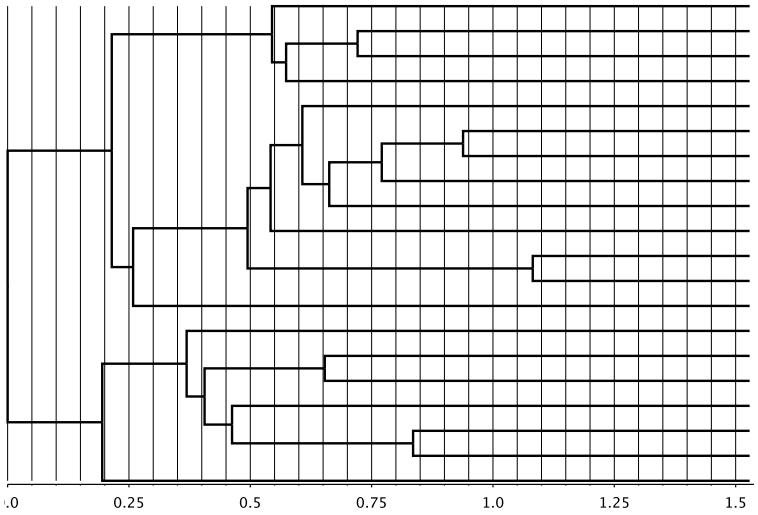
Effect of population growth



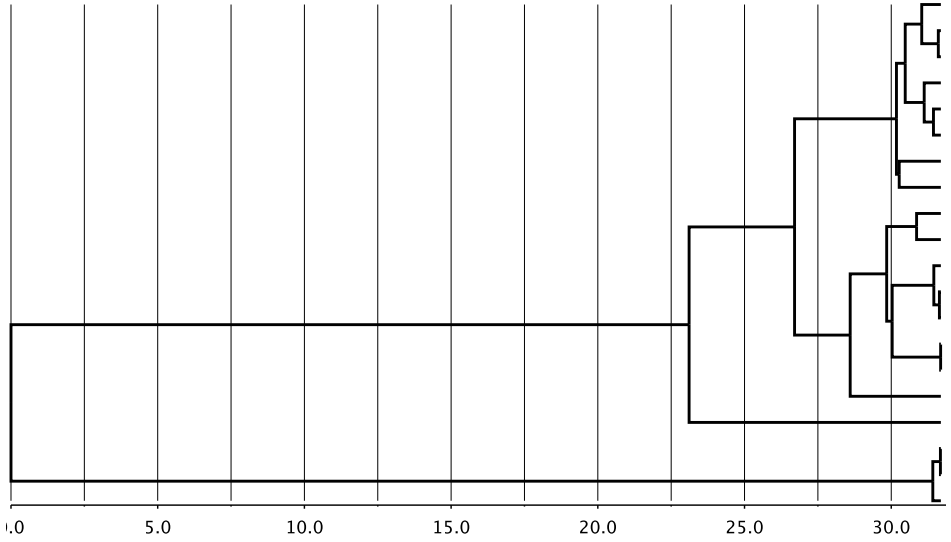
Exponential growth reduces compression near present

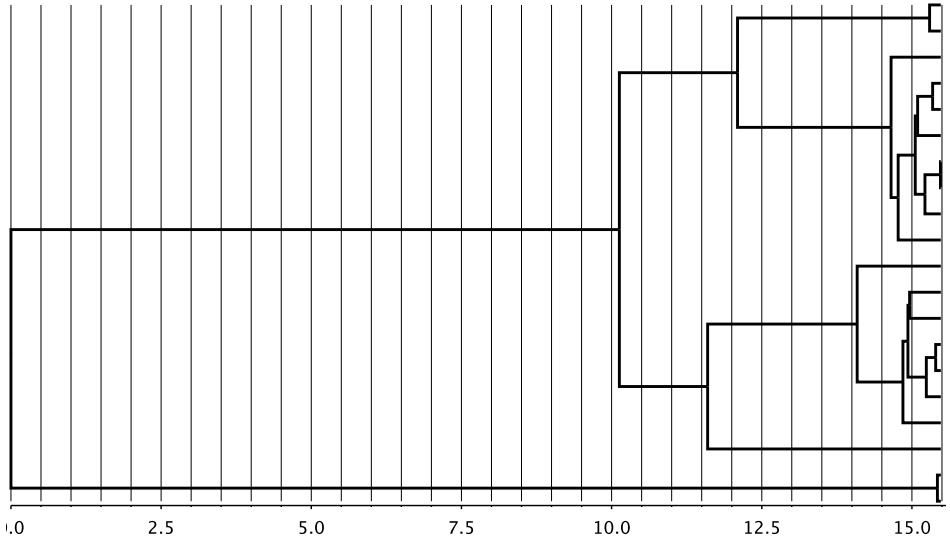
Exponential growth





Constant population size

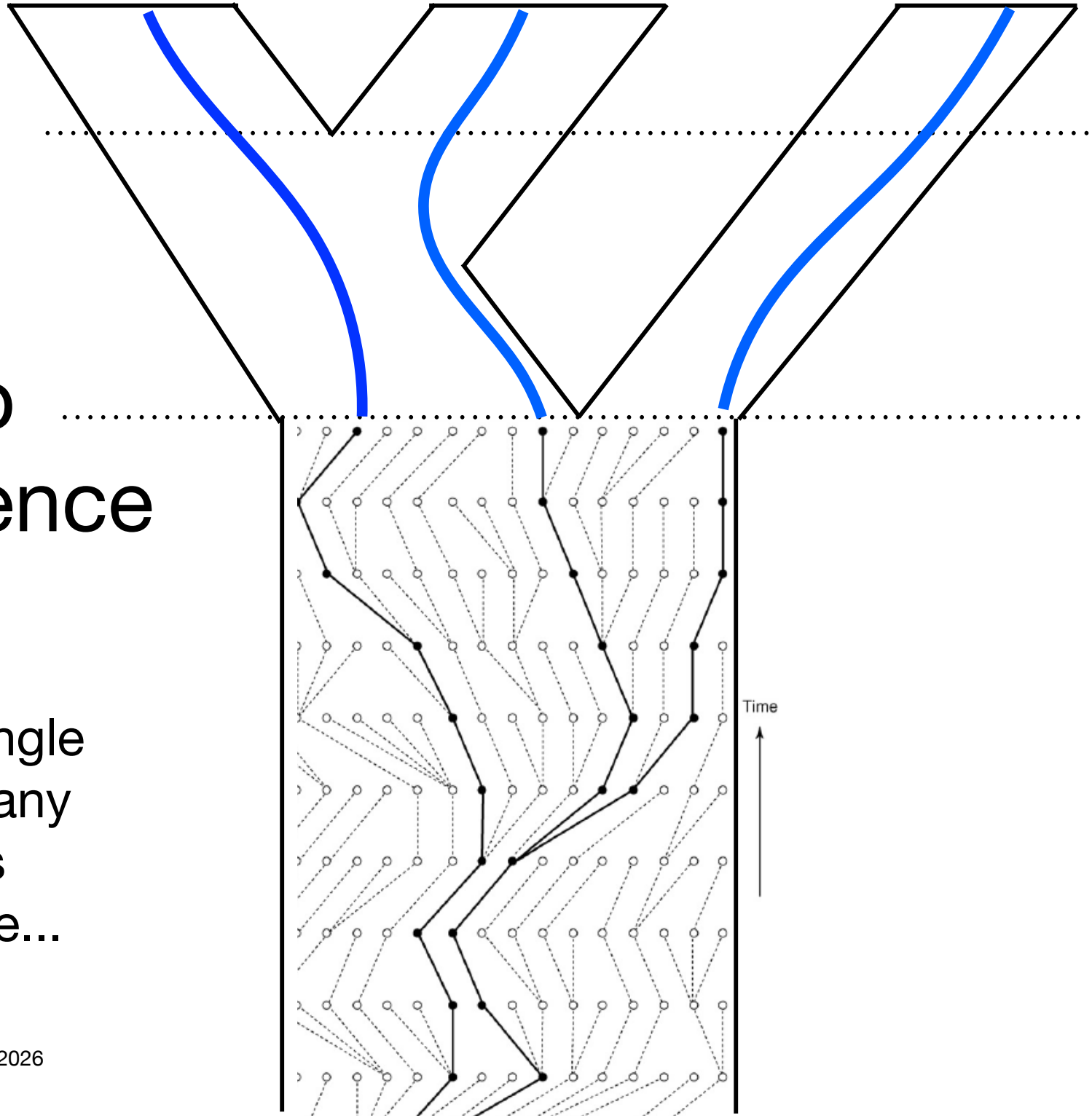




exponential growth makes edge lengths more even

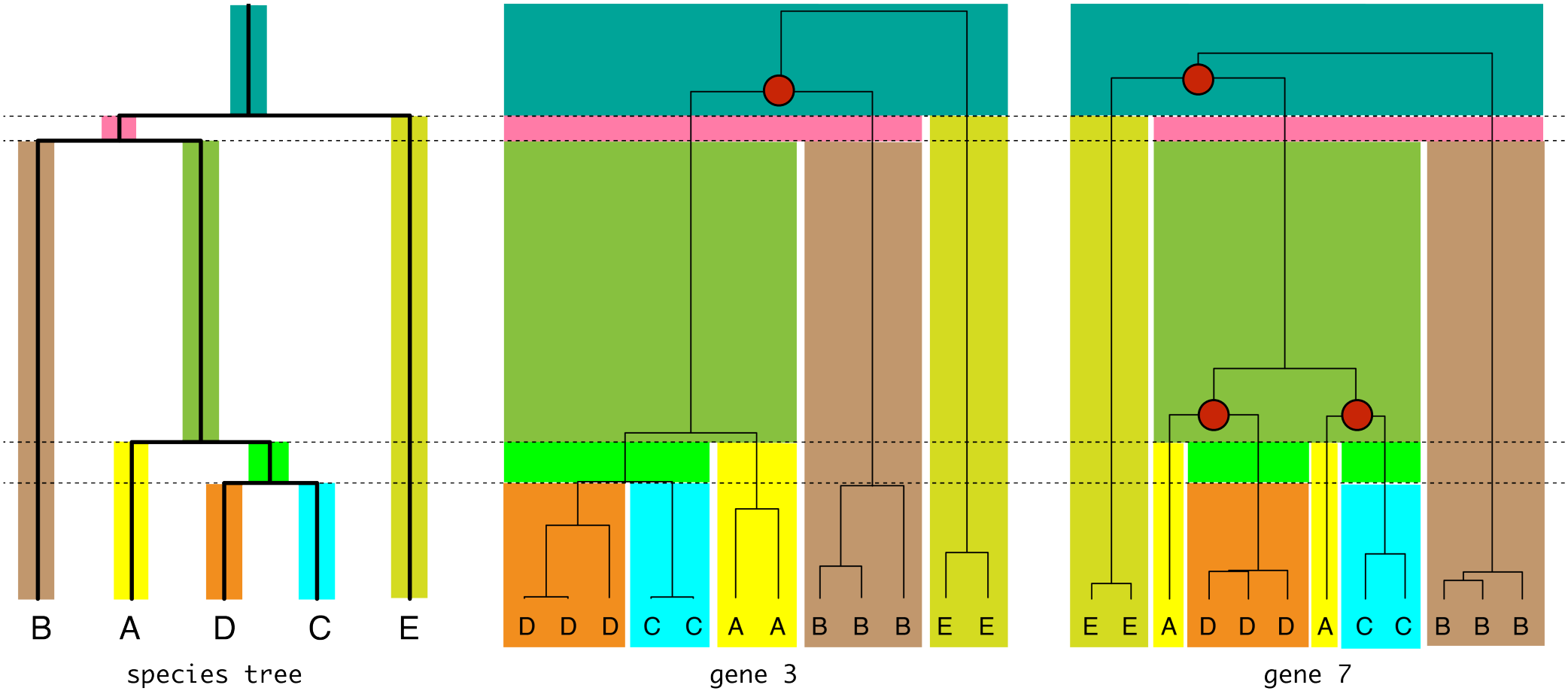
Deep coalescence can cause conflict among gene trees

Deep coalescence



Once in a single population, any two lineages can coalesce...

Gene tree conflict



species tree

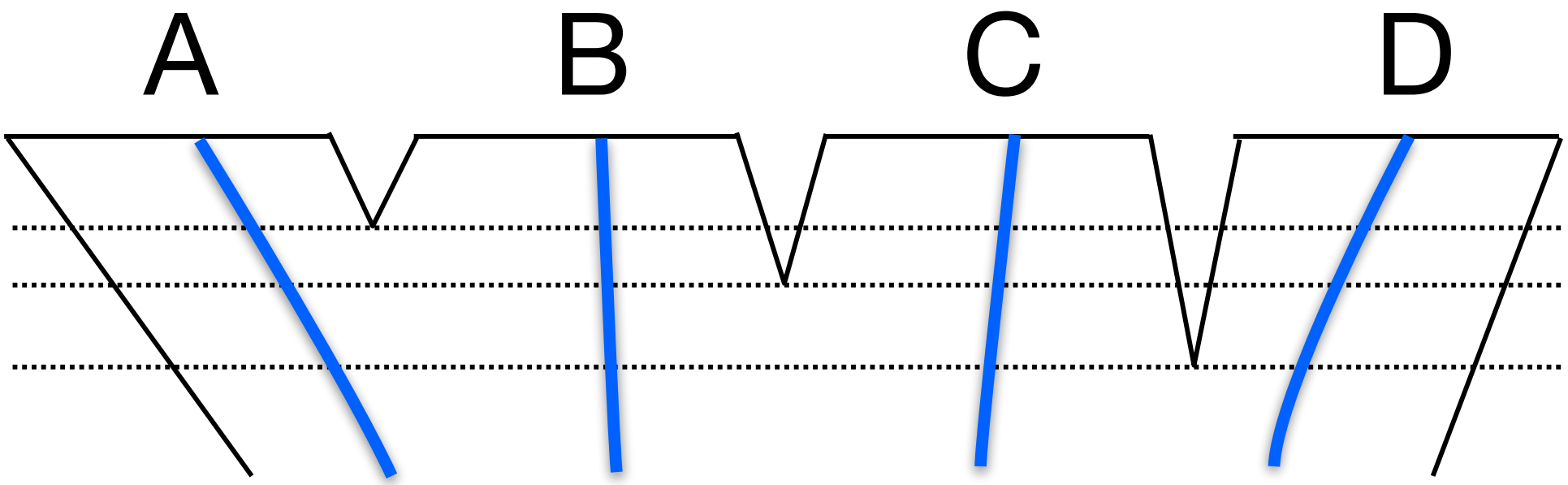
gene 3

gene 7

Gene 3 agrees with the species tree (even though there is one deep coalescence)

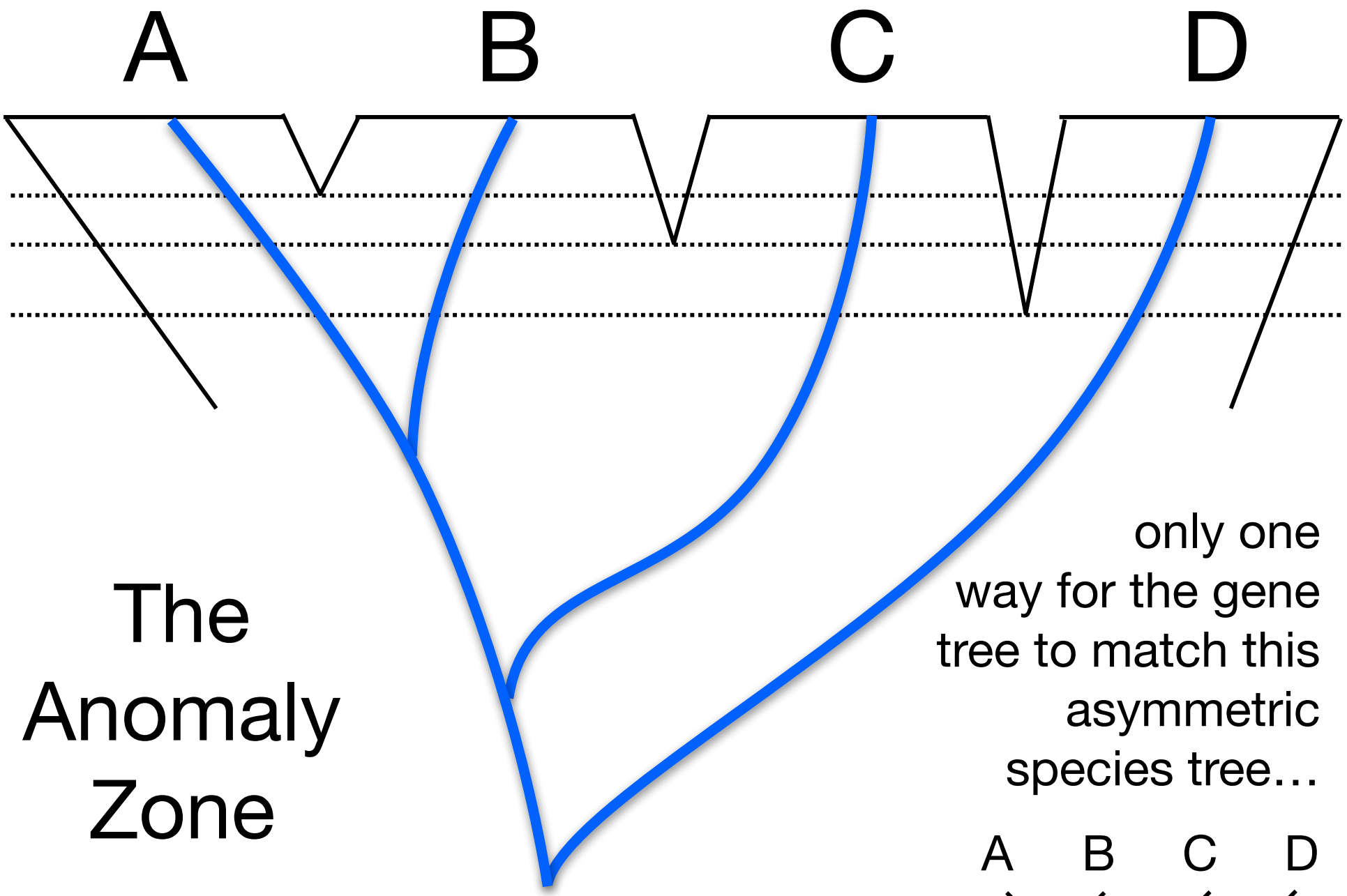
Gene 7 conflicts with the species tree (and thus also with gene 3)

The anomaly zone



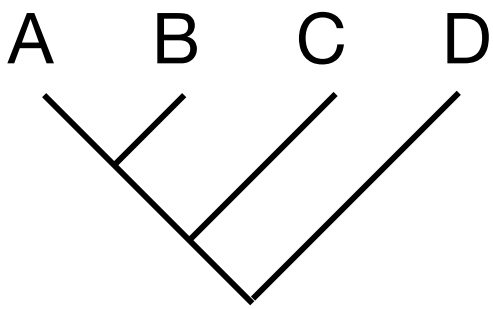
Anything can happen now!

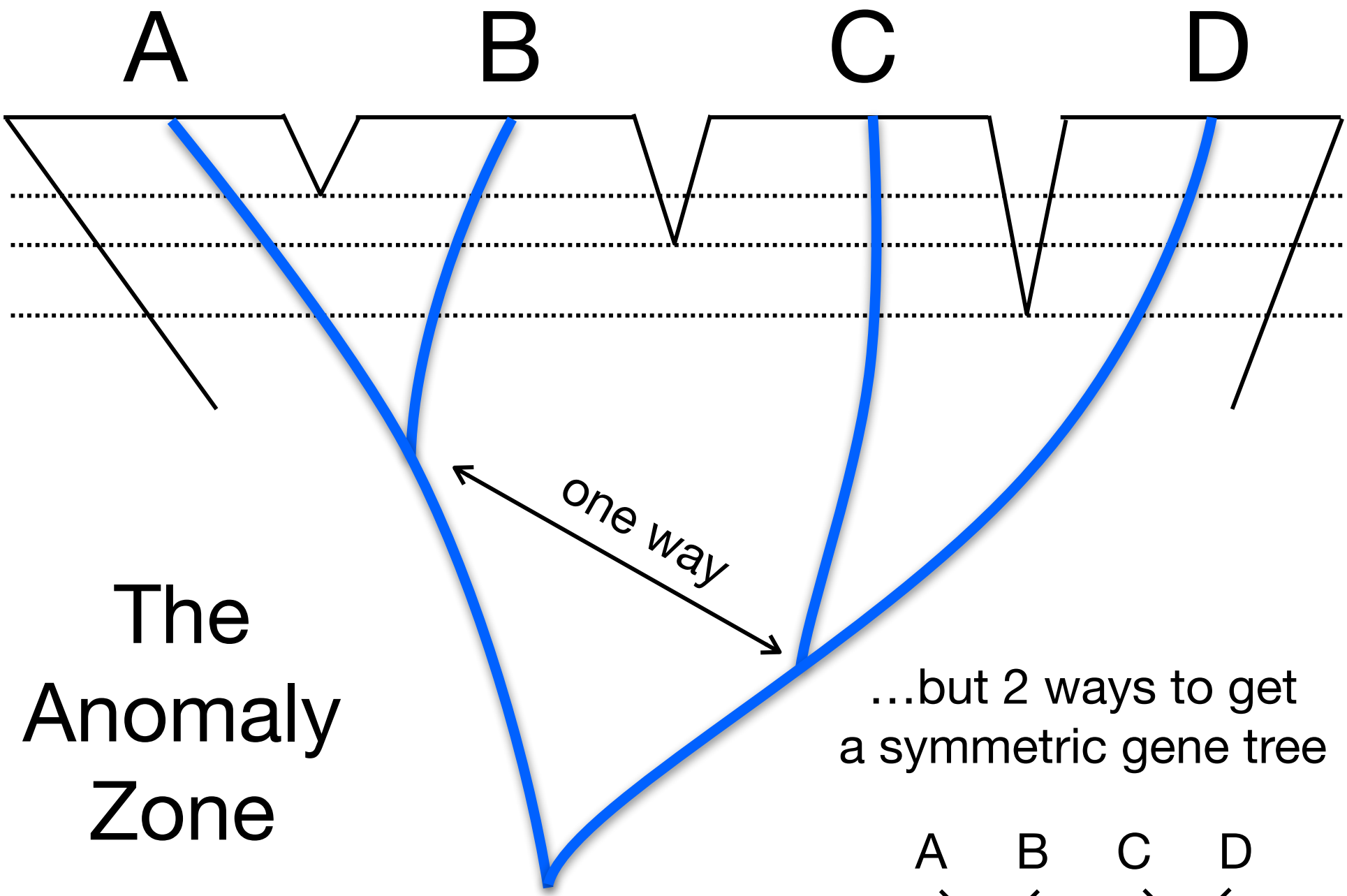
The Anomaly Zone



The Anomaly Zone

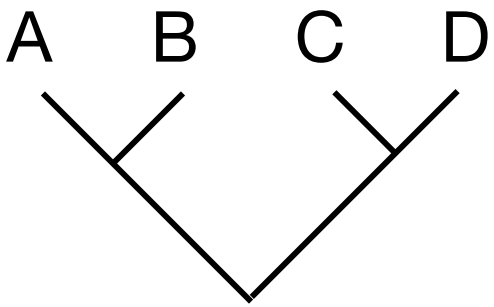
only one way for the gene tree to match this asymmetric species tree…

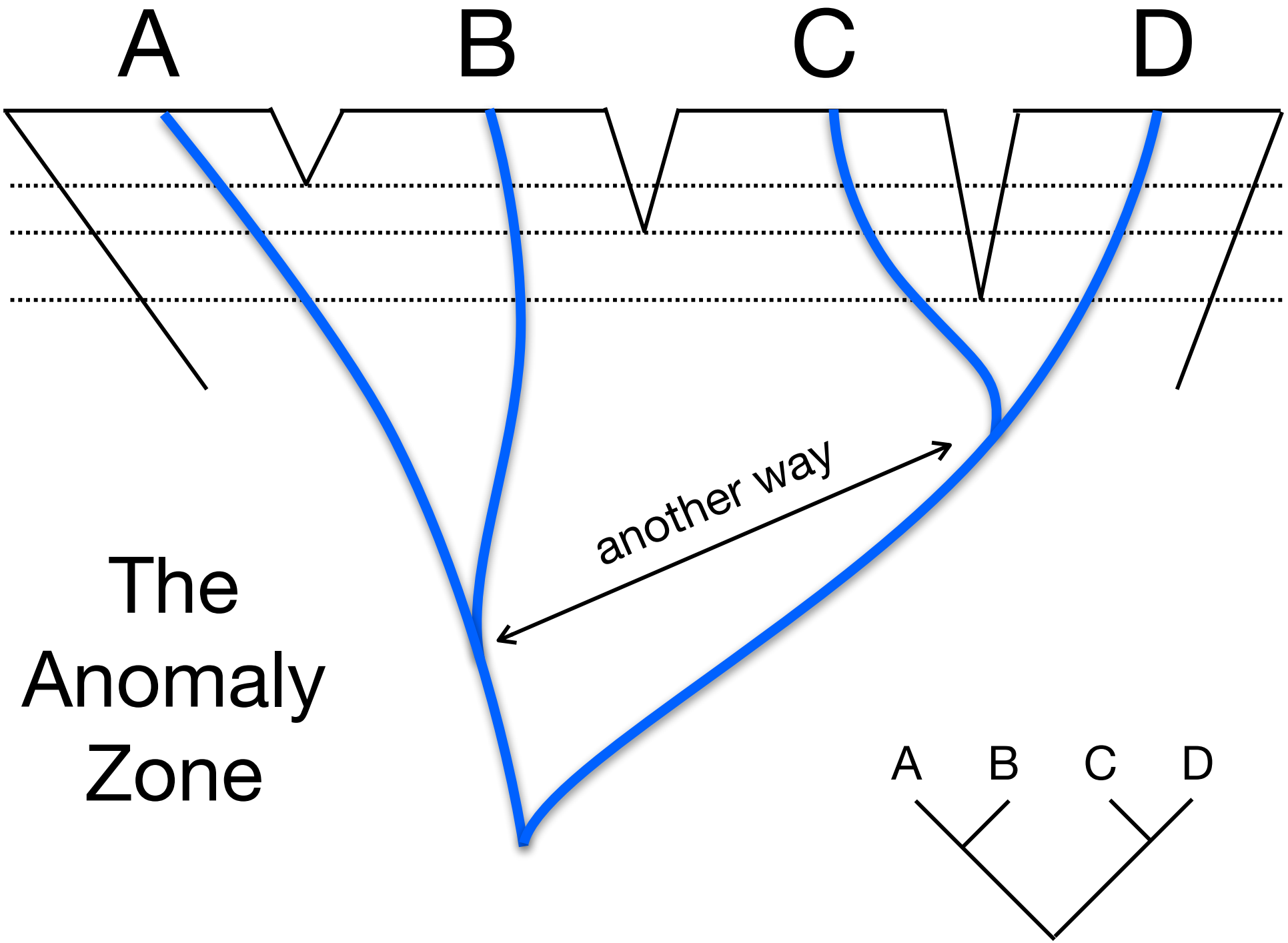




The Anomaly Zone

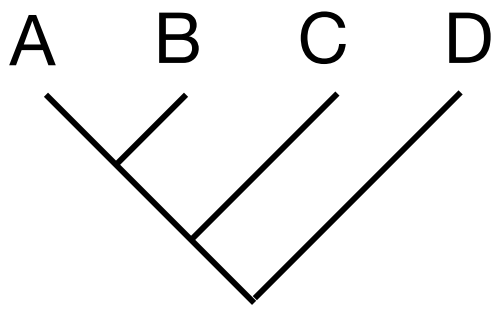
…but 2 ways to get a symmetric gene tree



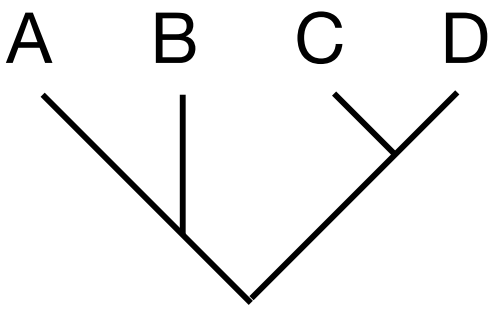


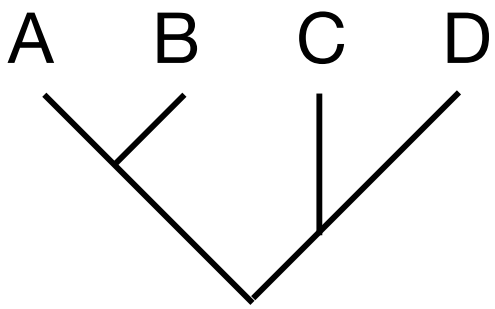
The Anomaly Zone

The Anomaly Zone



1 way to get gene tree that matches species tree





2 ways to get symmetric gene tree that does not match species tree

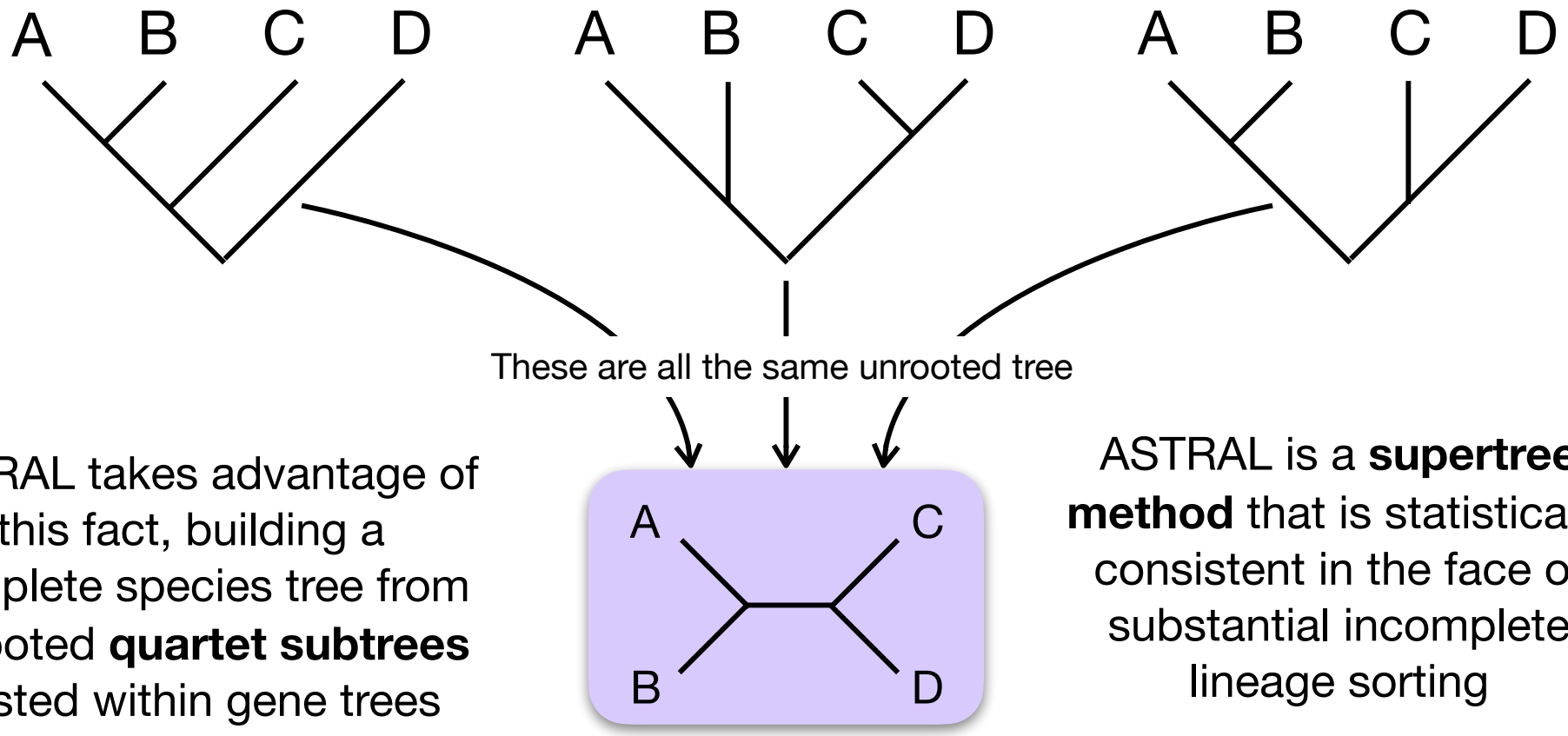
Anomalous gene trees (AGTs) are gene trees that are more probable than the species tree

Degnan and Rosenberg (2006)

Quartet-based species tree methods

ASTRAL

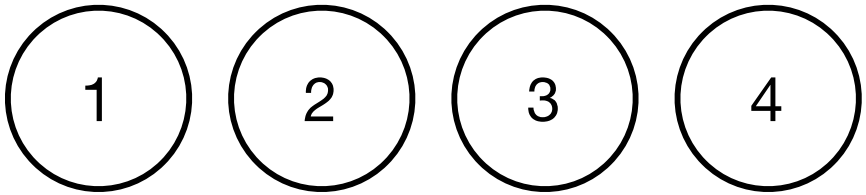
For 4 taxa, anomalous gene trees only exist if *rooted* trees are considered…



ASTRAL takes advantage of this fact, building a complete species tree from unrooted **quartet subtrees** nested within gene trees

ASTRAL is a **supertree method** that is statistically consistent in the face of substantial incomplete lineage sorting

Mirarab and Warnow (2015)



Combinations

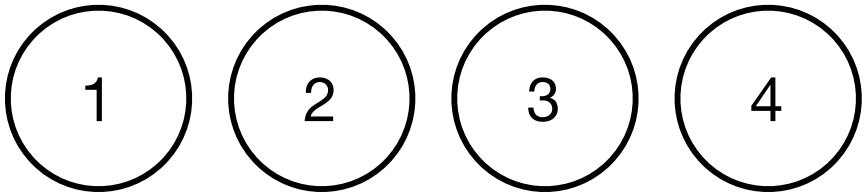
1. 1234	13. 3124
2. 1243	14. 3142
3. 1324	15. 3214
4. 1342	16. 3241
5. 1423	17. 3412
6. 1432	18. 3421
7. 2134	19. 4123
8. 2143	20. 4132
9. 2314	21. 4213
10. 2341	22. 4231
11. 2413	23. 4312
12. 2431	24. 4321

$n! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ total orderings

1. 12|34	4. 23|14
2. 13|24	5. 24|13
3. 14|23	6. 34|12

$$\binom{4}{2} = \frac{4!}{2!\ 2!} = \frac{24}{2 \cdot 2} = 6$$

6 ways of choosing 2 out of the 4 if order doesn't matter



Combinations

1. 1234
2. ~~1243~~
3. 1324
4. ~~1342~~
5. 1423
6. ~~1432~~
7. ~~2134~~
8. ~~2143~~
9. 2314
10. ~~2341~~
11. 2413
12. ~~2431~~
13. ~~3124~~
14. ~~3142~~
15. ~~3214~~
16. ~~3241~~
17. 3412
18. ~~3421~~
19. ~~4123~~
20. ~~4132~~
21. ~~4213~~
22. ~~4231~~
23. ~~4312~~
24. ~~4321~~

1. 12|34
2. 13|24
3. 14|23
4. 23|14
5. 24|13
6. 34|12

$$\binom{4}{2} = \frac{4!}{2!\,2!} = \frac{24}{2 \cdot 2} = 6$$

6 ways of choosing 2 out of the 4 if order doesn't matter

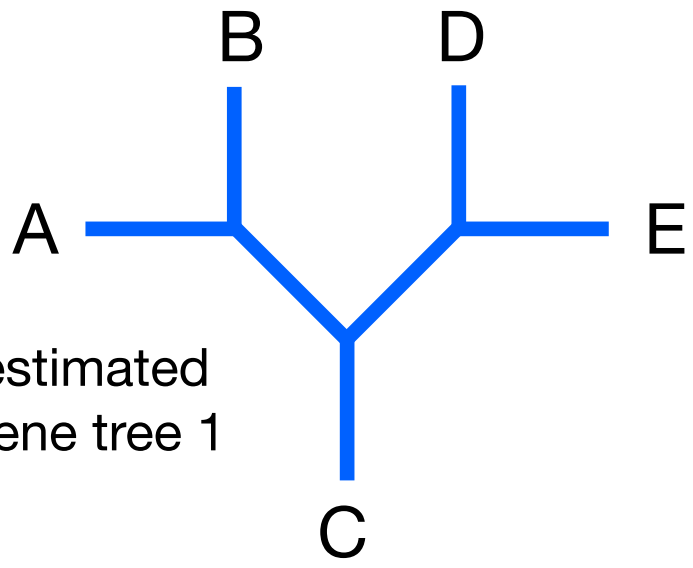
right side alt. order

left side alt. order

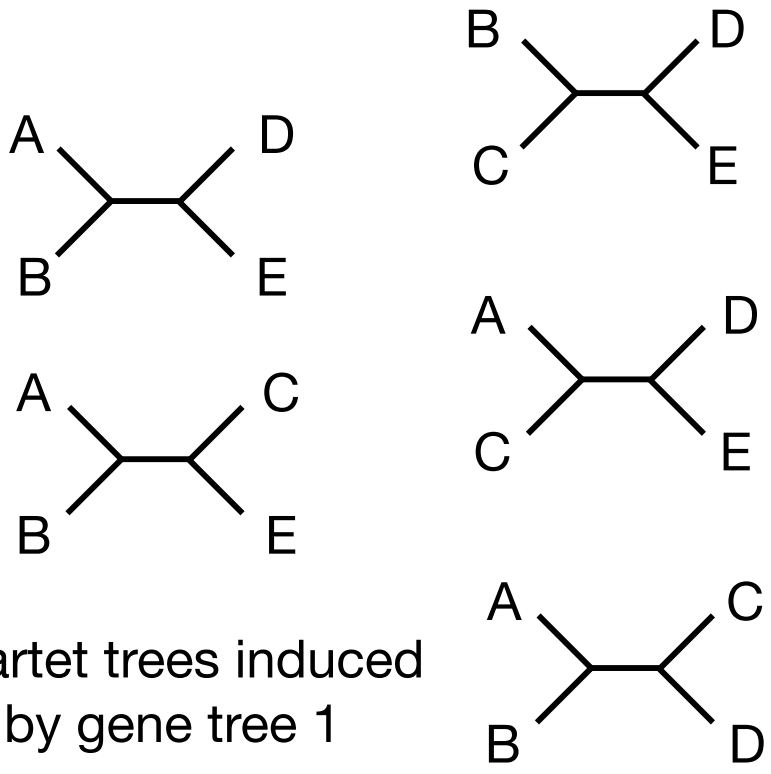
$n! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ total orderings

How many ways are there of choosing
4 taxa (a quartet) out of 5?

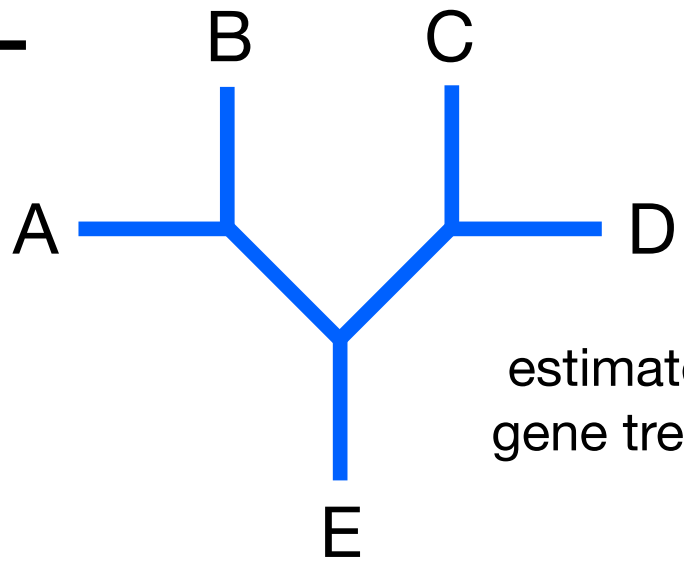
ASTRAL



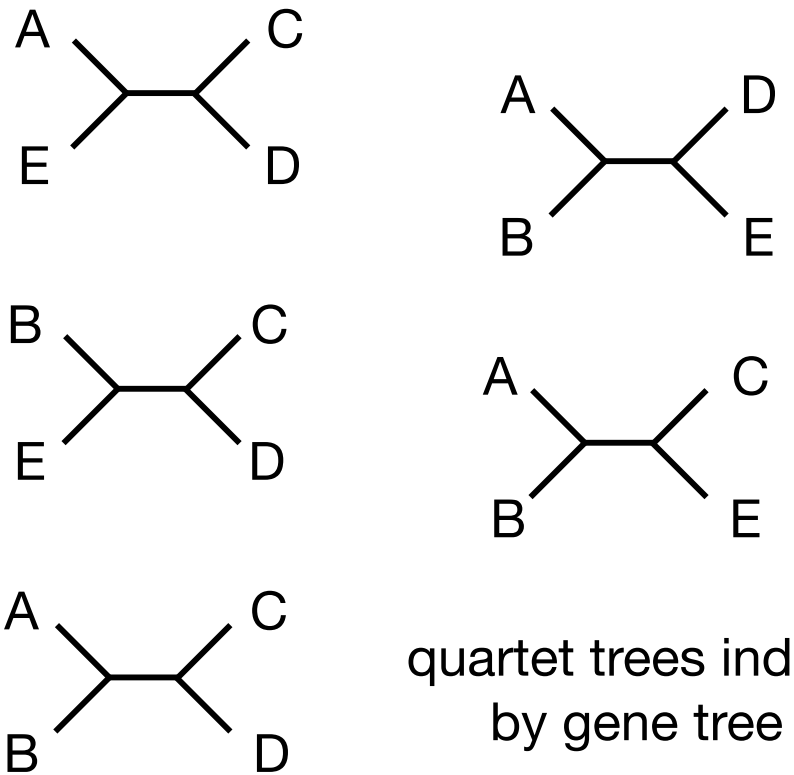
estimated gene tree 1



quartet trees induced by gene tree 1

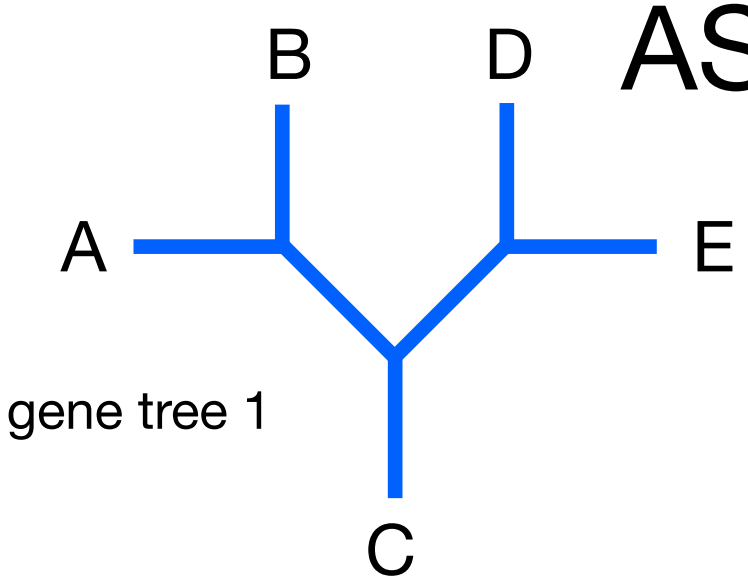


estimated gene tree 2

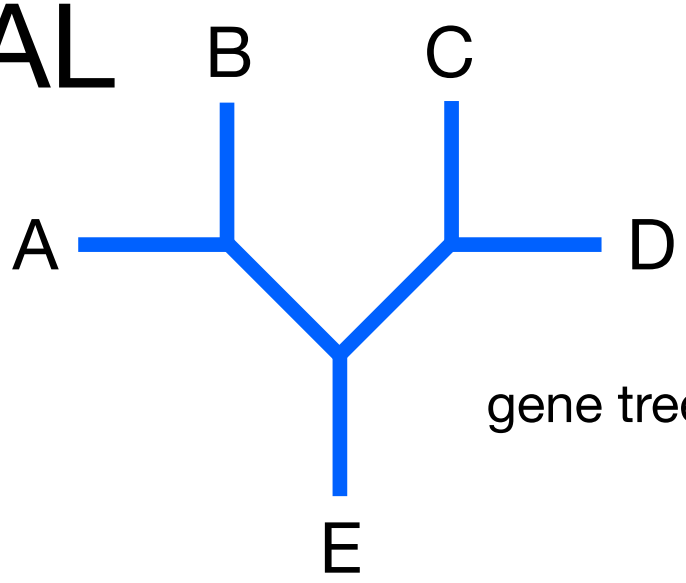


quartet trees induced by gene tree 2

ASTRAL

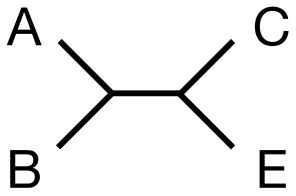


gene tree 1

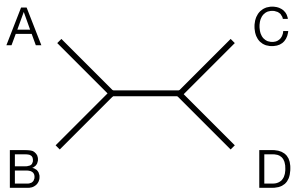


gene tree 2

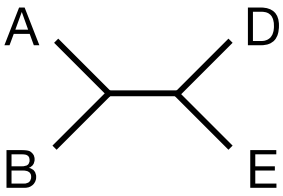
weight (w) = number of gene trees in which quartet is found



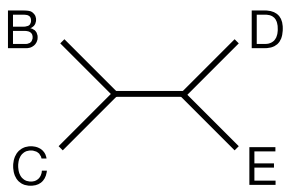
w=2



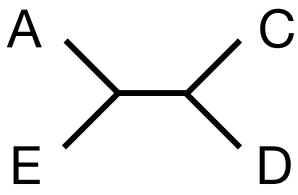
w=2



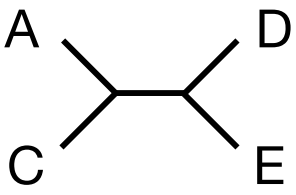
w=2



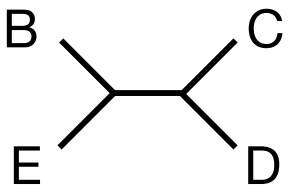
w=1



w=1

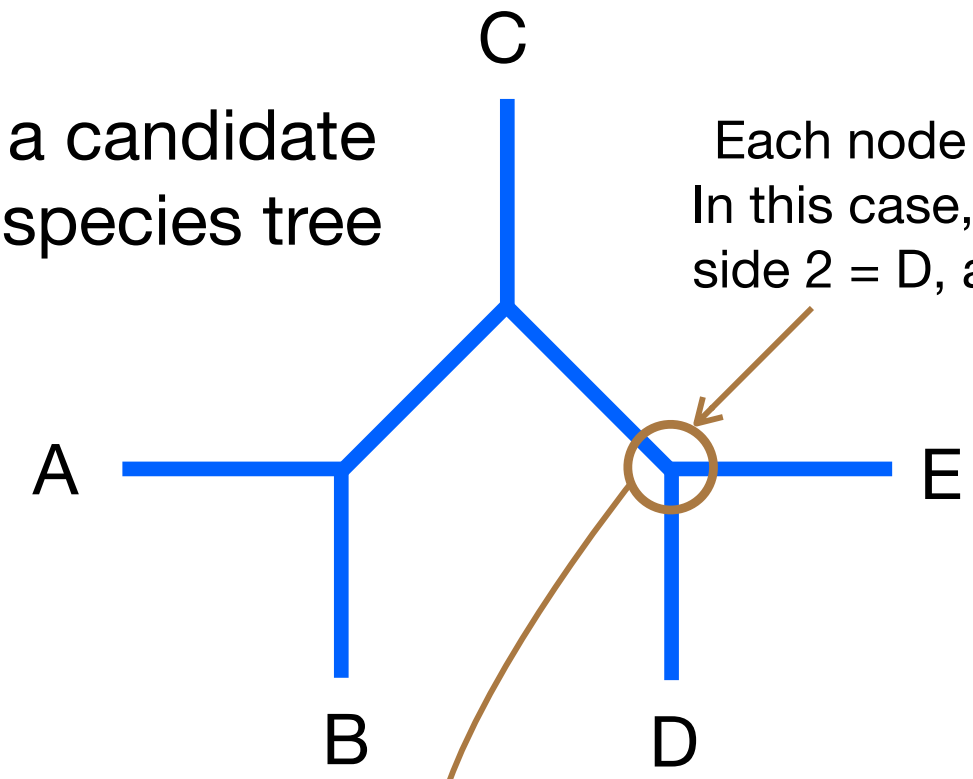


w=1



w=1

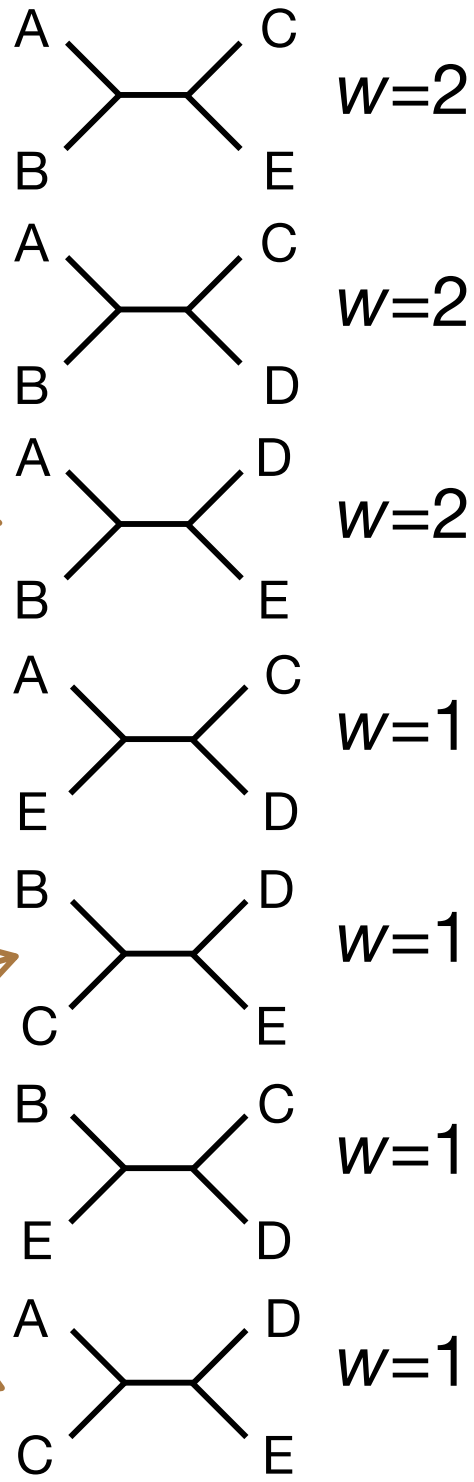
a candidate species tree



Each node has 3 “sides”
In this case, side 1 = ABC, side 2 = D, and side 3 = E)

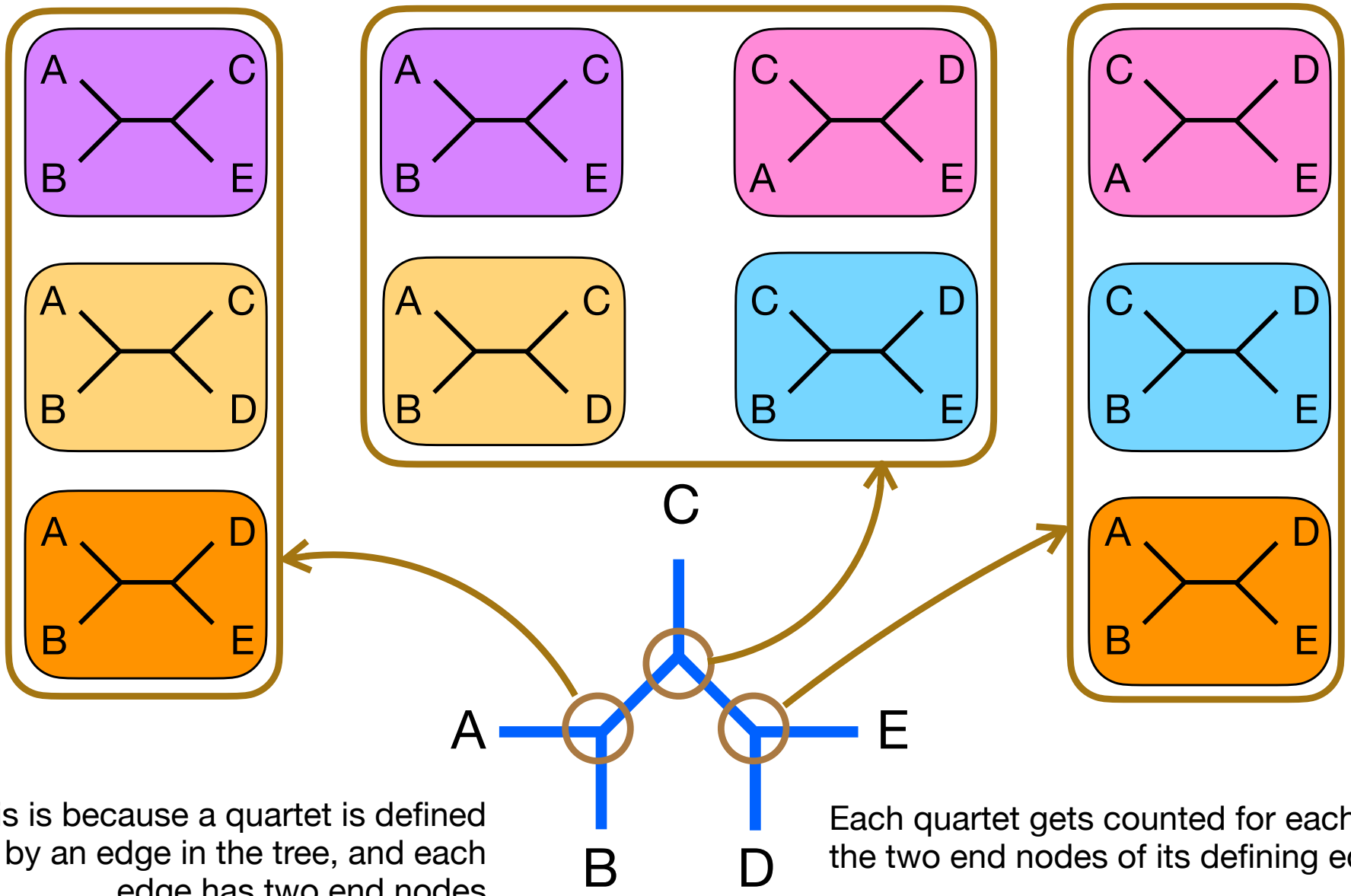
weight for each node in candidate tree = sum of weights for all quartets that map to the node

w(ABC|D|E) = 2+1+1



ASTRAL

Each quartet gets counted twice

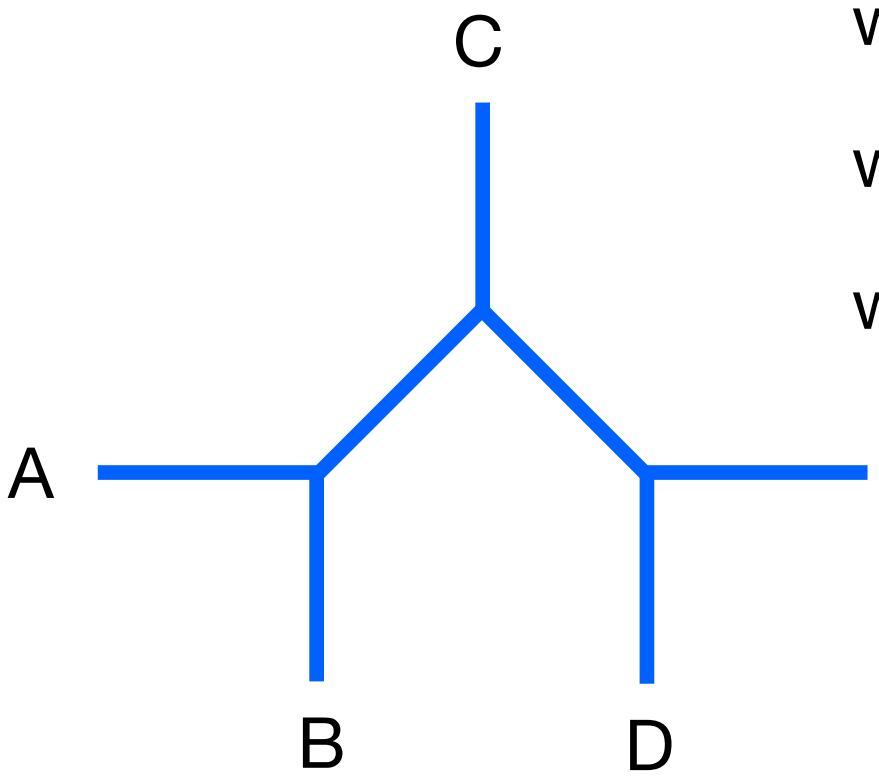


This is because a quartet is defined by an edge in the tree, and each edge has two end nodes

Each quartet gets counted for each of the two end nodes of its defining edge

ASTRAL

Total support for candidate species tree
= (4+6+6)/2 = **8**



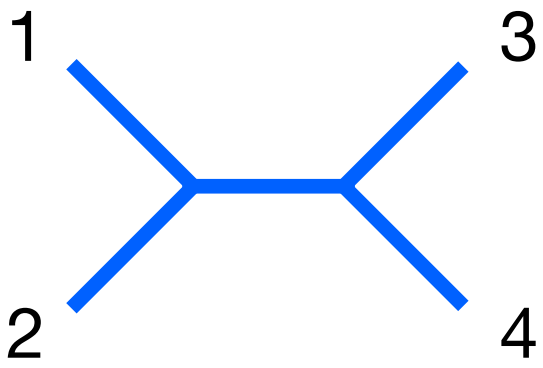
w(ABC|D|E) = 2+1+1 = **4**

w(AB|C|DE) = 2+2+1+1 = **6**

w(A|B|CDE) = 2+2+2 = **6**

total weight is half the sum of the node weights because each quartet gets counted twice

https://github.com/smirarab/ASTRAL



SVDQuartets

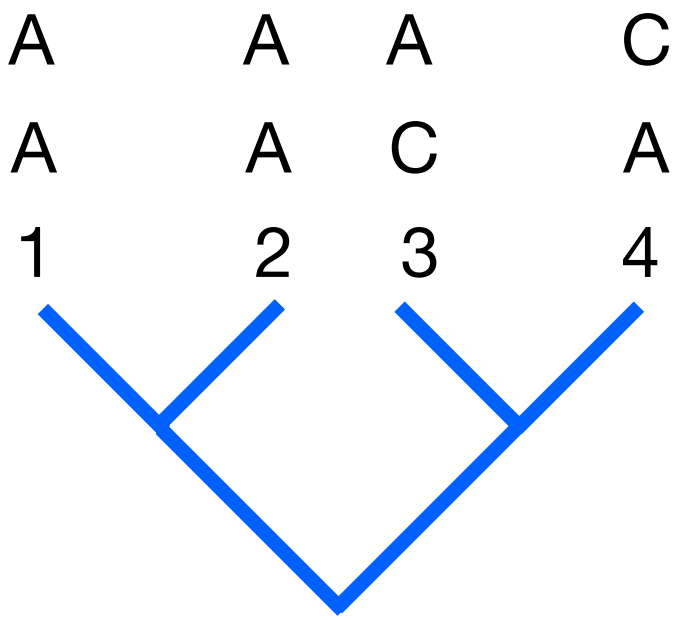
Quartet method that consults the original sequences rather than depending on gene trees being correctly estimated

```
taxon1   AAAGATTACAGGTTGACTTATTACACCCCGGAG...
taxon2   AAAGATTATCGACTGACTTATTACACCCCCGAA...
taxon3   AAAGATTACAGATTAACTTATTATACTCCTGAA...
taxon4   AAAGATTATAAATTGACTTACTACACCCCGGAG...
```

$$\text{flat}_{12|34} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} AA & AC & AG & AT & CA & \end{array} \\ \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

The flat matrix for any quartet tree stores counts of sites with patterns that match the quartet tree

Chifman and Kubatko (2014)



SVDQuartets

Symmetries among columns in the flat matrix are expected, even if there is ILS

For example p_{AAAC} should approximately equal p_{AACA} (and the same is true for every pair in these two columns)

$$\text{flat}_{12|34} = \begin{array}{c} \\ AA \\ AC \\ AG \\ AT \\ CA \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} AA & AC & AG & AT & CA \end{array} \\ \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

Such symmetries reduce the **rank** of the flat matrix, which is a measure of how many columns (rows) are independent

SVDQuartets

- Flat matrix should have rank
  - 16 for quartets **not** in the species tree
  - ≤10 for quartets that **are** in the species tree under coalescent model
  - ≤4 for quartets that are in the true tree if true tree is common to all sites
- Evaluate all quartets (or a sample)
- Construct tree from quartets that have rank ≤10
- To use version in PAUP*:
  - load concatenated data matrix
  - use svdquartets command