From Trees

https://en.wikipedia.org/wiki/Strangler_fig

To Networks

Two questions

- How can we assess whether the signal in our data is tree-like?

- How do we combine and display information from different data sets?

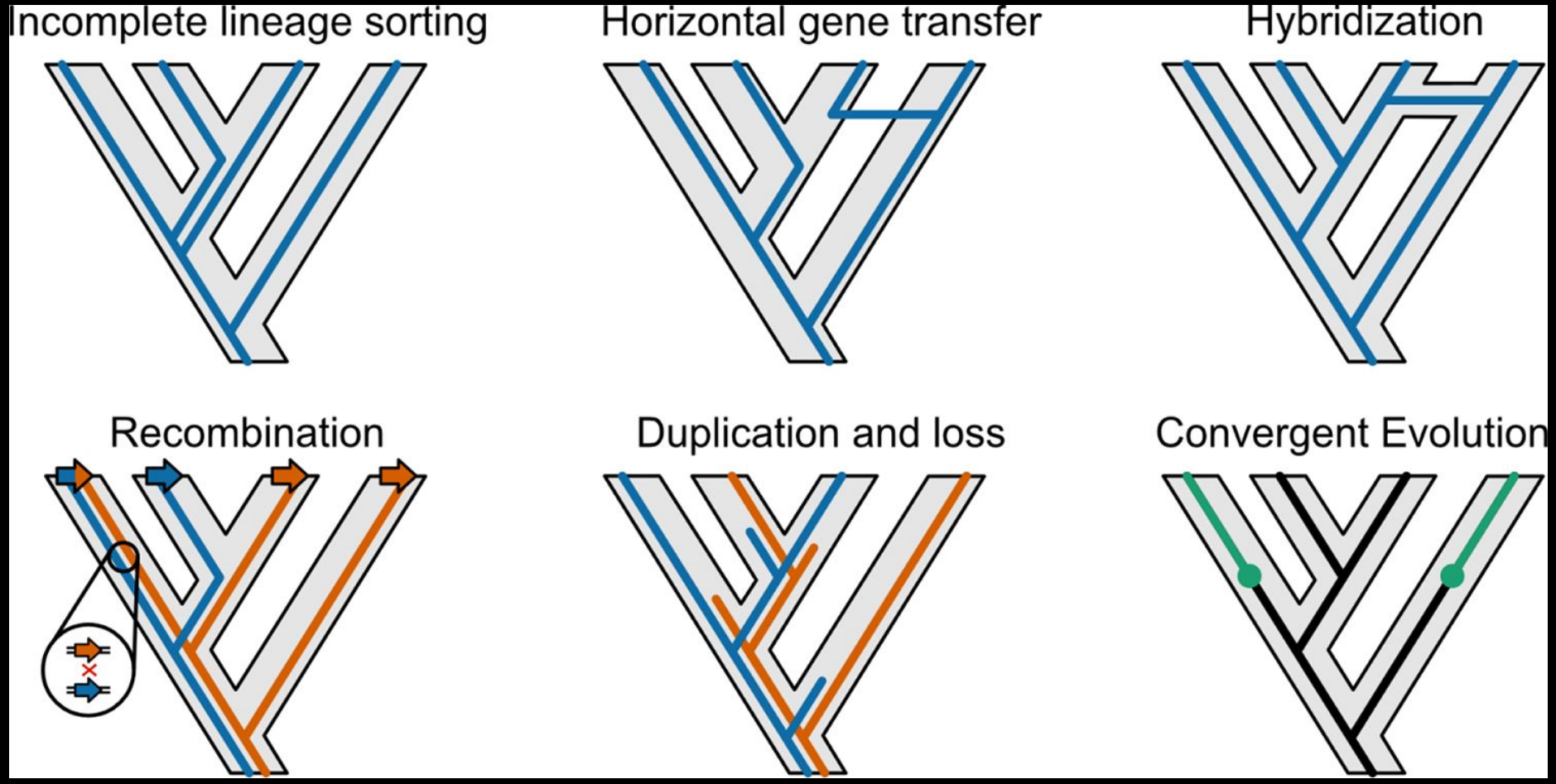
Violations of tree-like signal

- Does my entire data set (single alignment, set of alignments) present the same phylogenetic signal?

- i.e., do all alignment columns in a gene or genes in a set of genomes support the same tree?

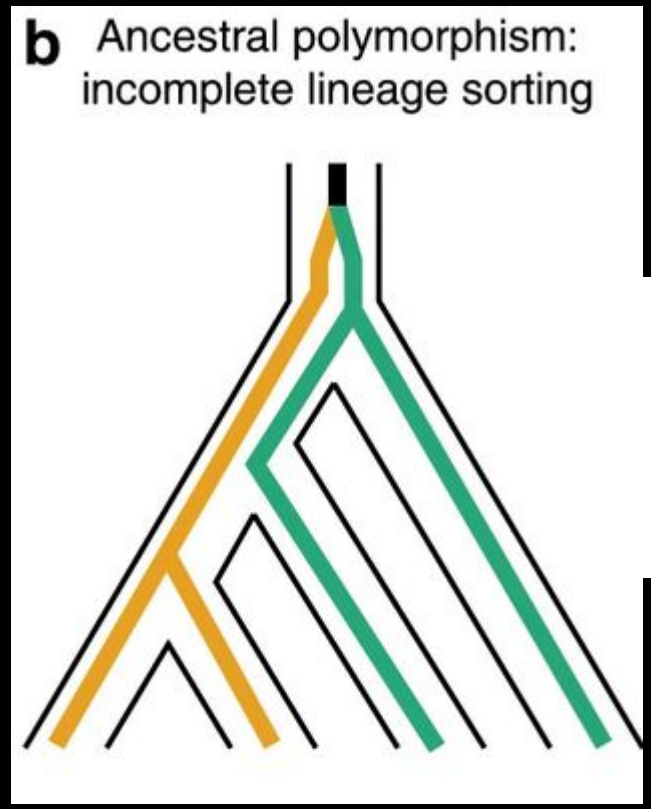
- Why wouldn't they?

Evolutionary noise in individual genes



Steenwyk, Jacob L., et al. "Incongruence in the phylogenomics era."
Nature Reviews Genetics 24.12 (2023): 834-850.

Incomplete Lineage Sorting

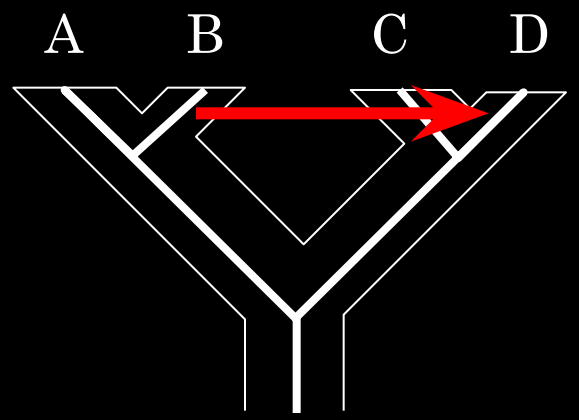


Palmer, D.H., Kronforst, M.R. A shared genetic basis of mimicry across swallowtail butterflies points to ancestral co-option of doublesex. Nat Commun 11, 6 (2020). https://doi.org/10.1038/s41467-019-13859-y

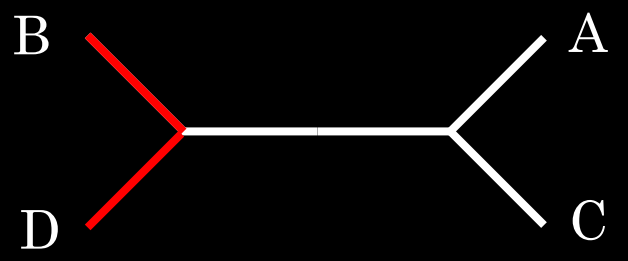
Lateral gene transfer

- The transmission of genetic material between genomes in a manner other than parent to offspring

- A gene from organism ‘B’ is acquired by organism ‘D’ - thus B and D appear to be very closely related

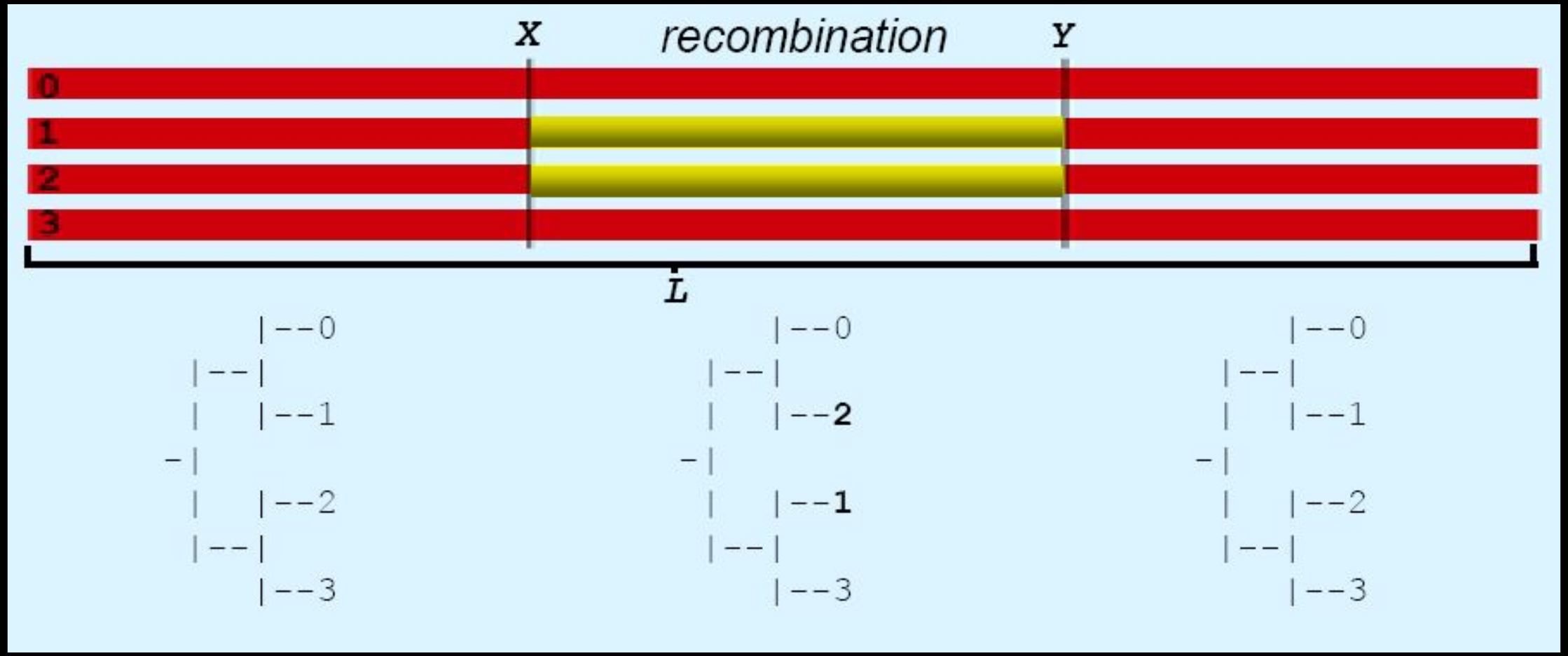


Organism phylogeny with LGT event



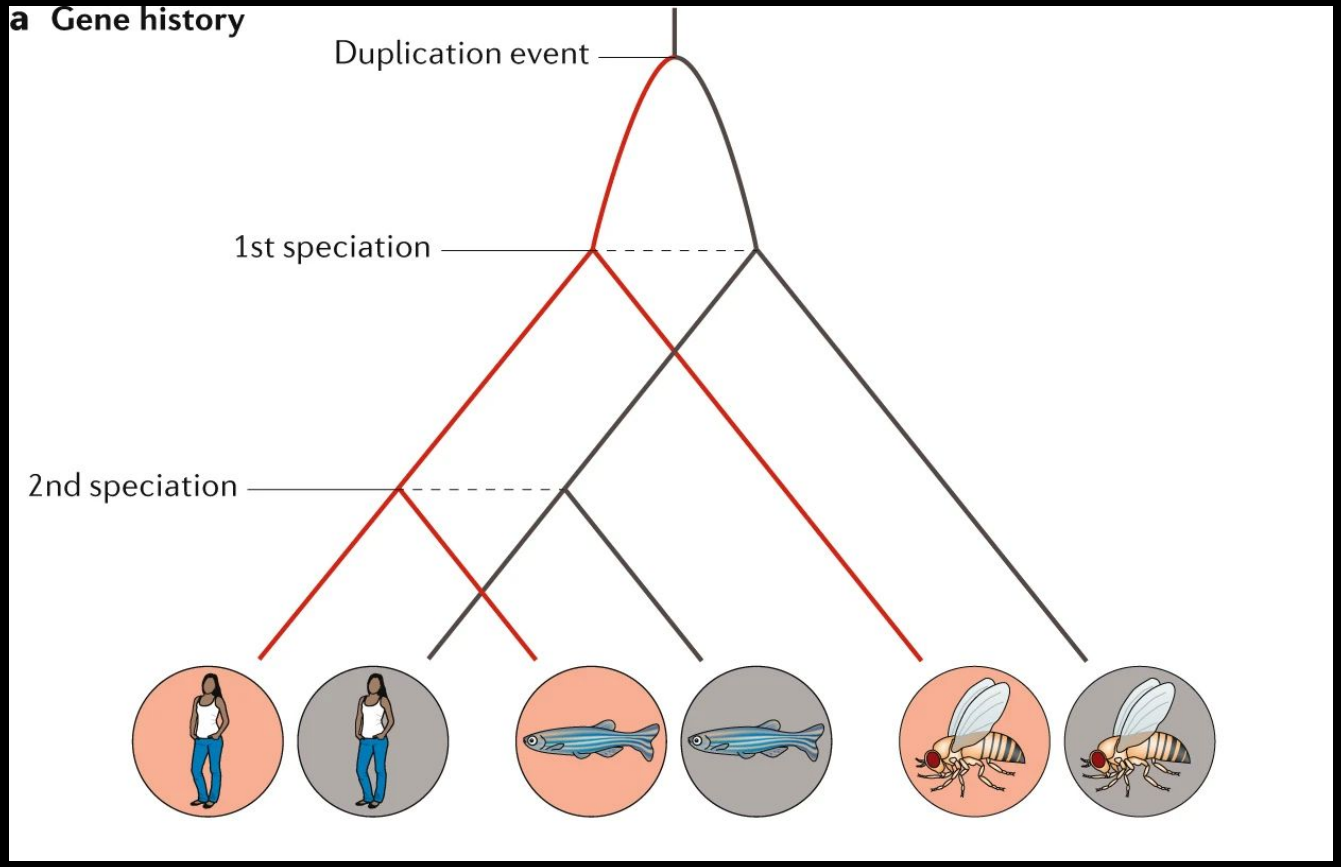
Inferred gene tree

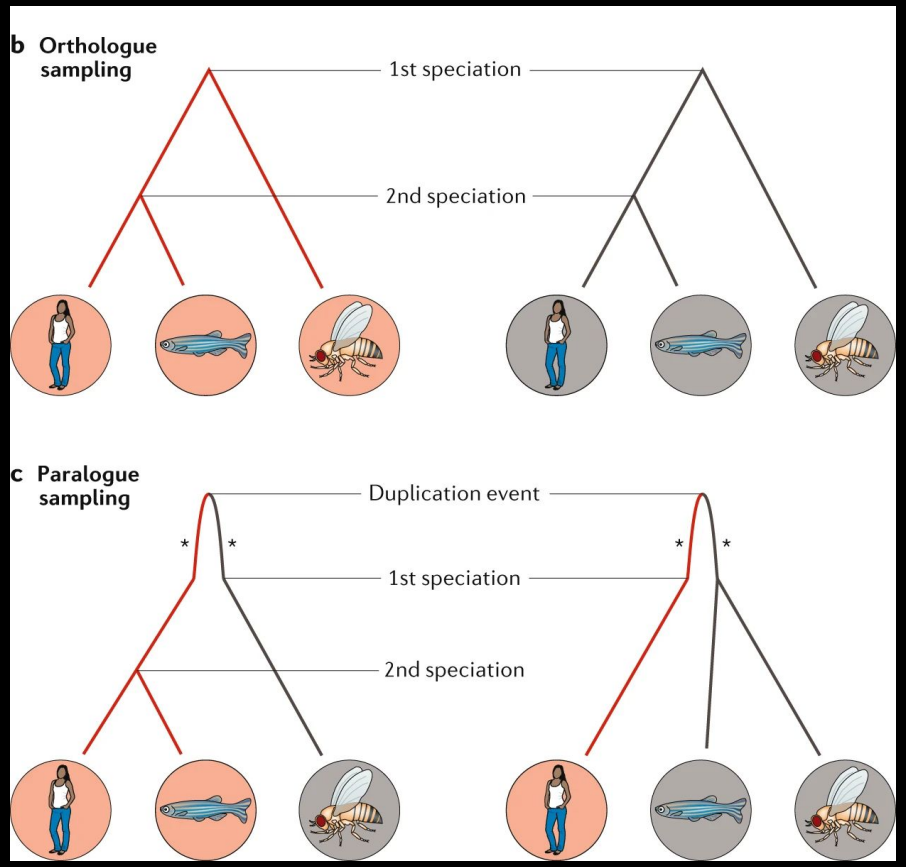
Reticulate evolution



Chan et al. (2005) *BMC Bioinformatics*

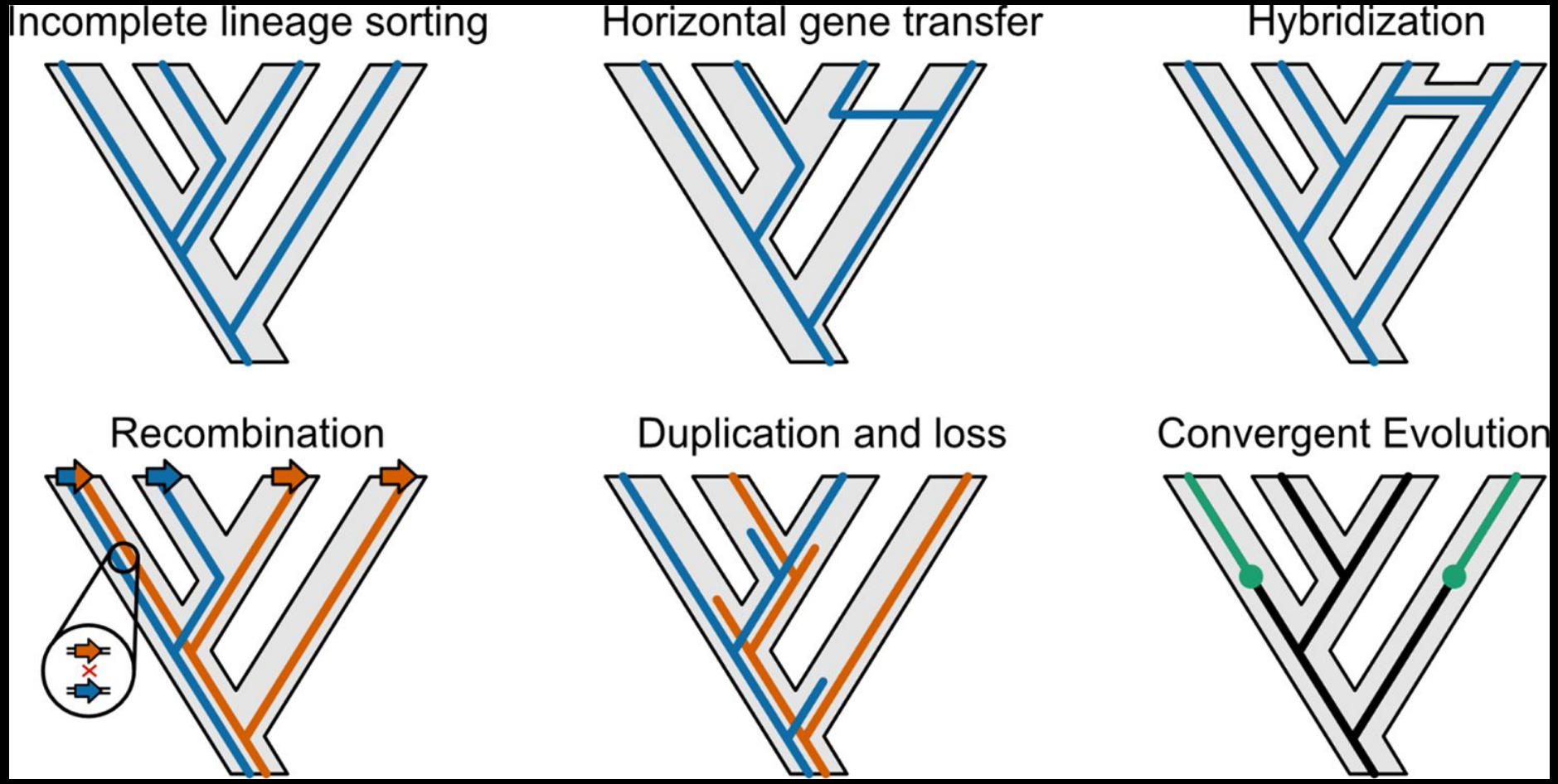
Paralogy (hidden or otherwise)





Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. "Phylogenetic tree building in the genomic age." *Nature Reviews Genetics* 21.7 (2020): 428-444.

Evolutionary noise in individual genes

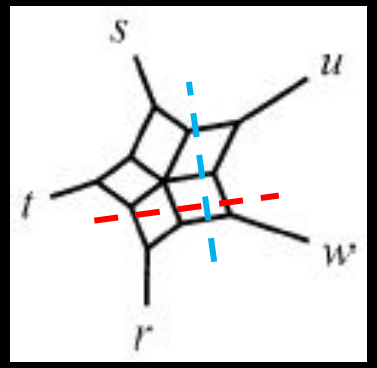


Steenwyk, Jacob L., et al. "Incongruence in the phylogenomics era."
Nature Reviews Genetics 24.12 (2023): 834-850.

Phylogenetic Networks

- Networks can represent incompatible splits
- Generalize the notion of a tree
- A given group of taxa can have multiple affinities or connections

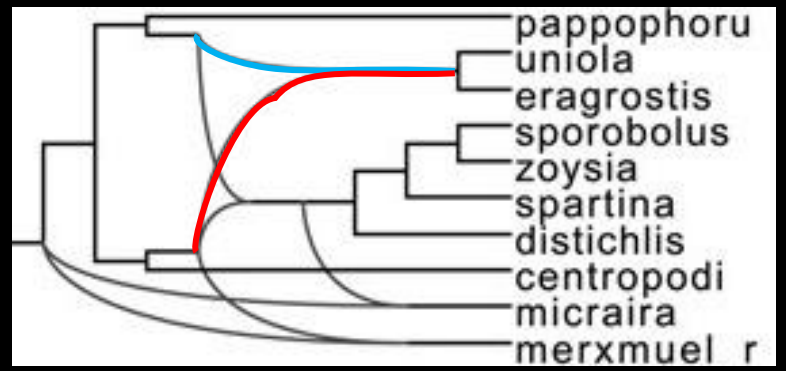
Implicit network



(uw | str)

(rw | stu)

Explicit network

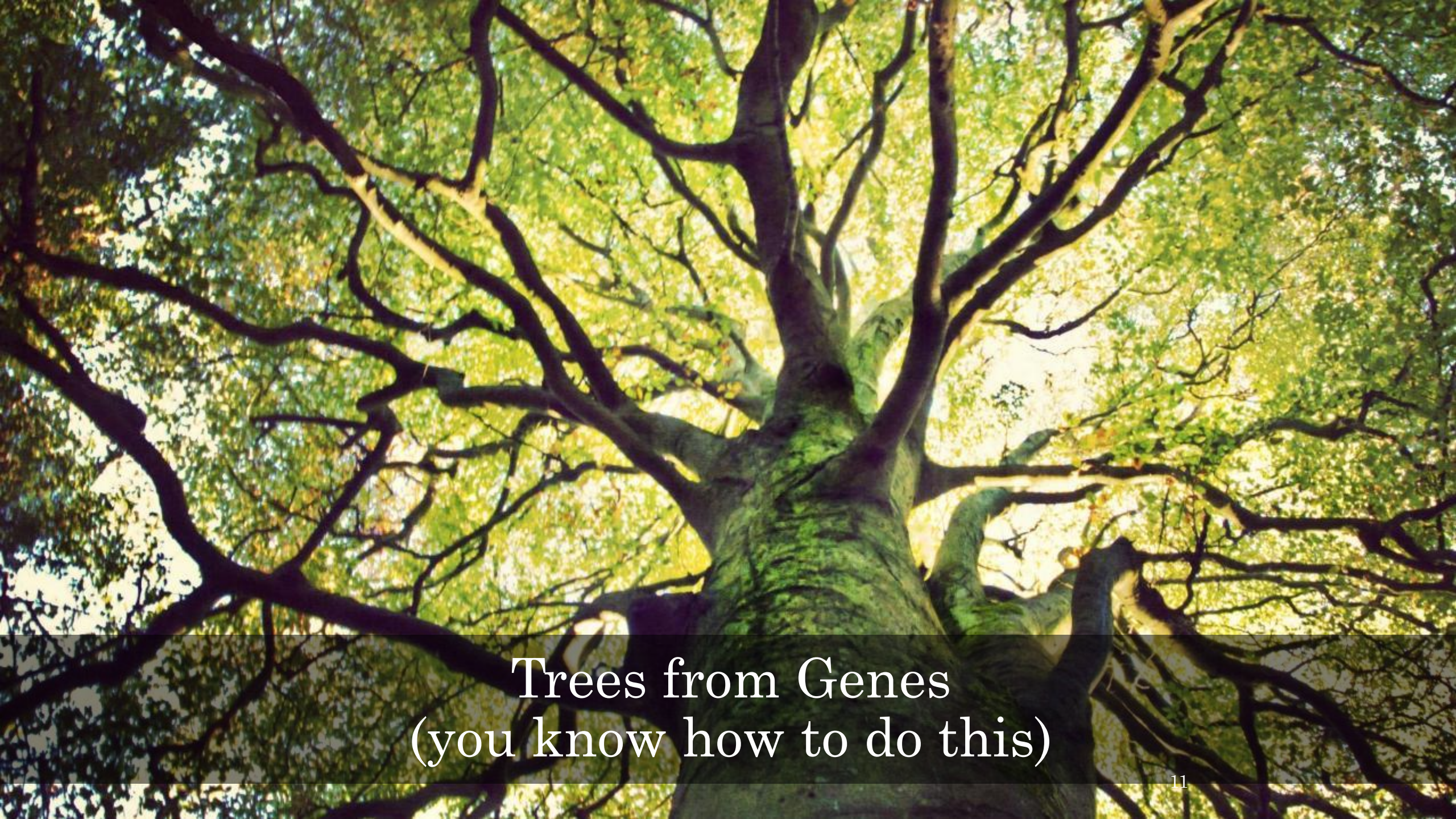


(pappophoru, uniola, eragrostis)

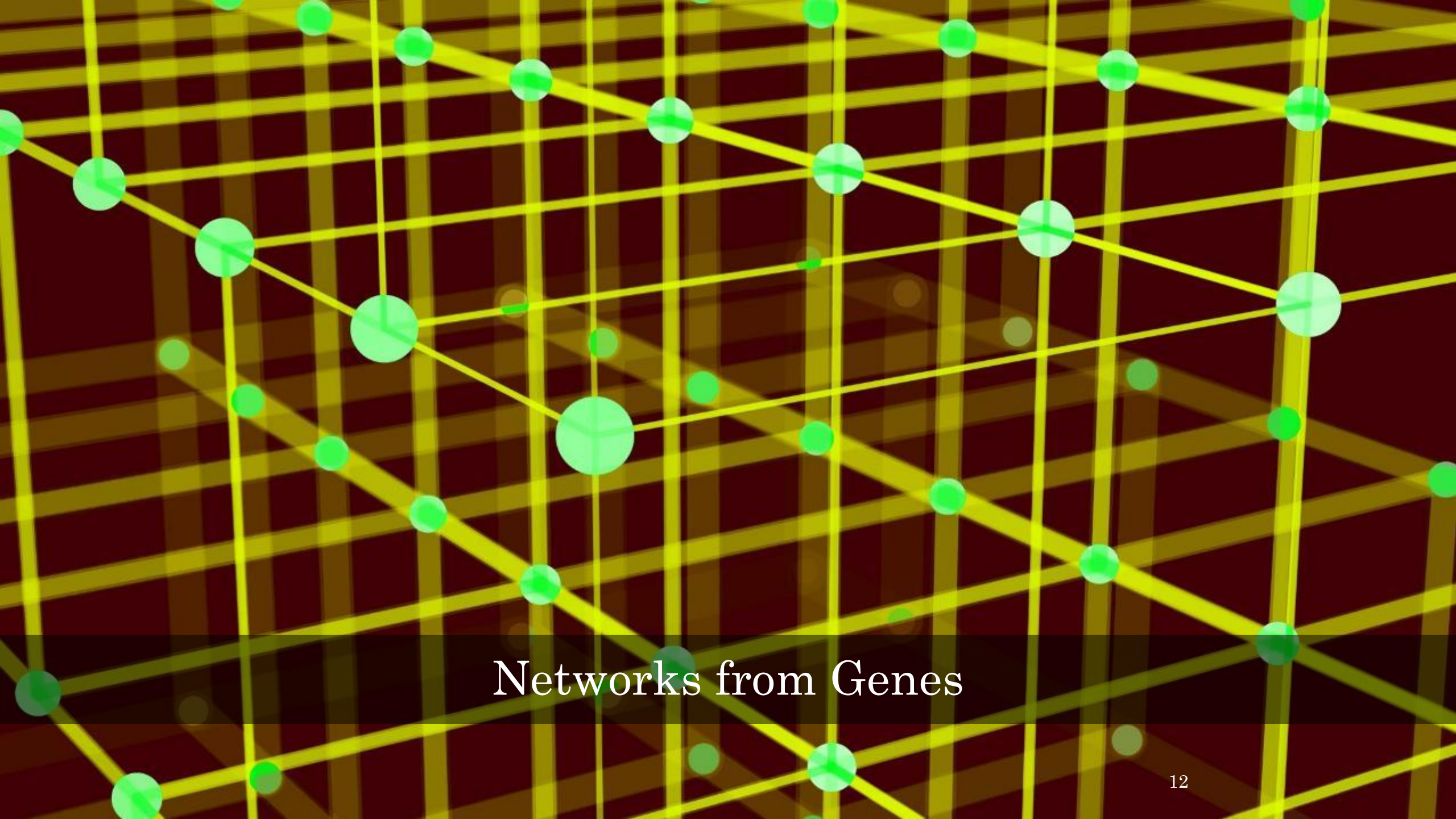
(centropodi, uniola, eragrostis)

grass hybridisation

Bryant and Moulton (2004) *Molecular Biology and Evolution*
van Iersel et al. (2010) *Bioinformatics*



Trees from Genes
(you know how to do this)



Networks from Genes

NeighborNet

- Based on neighbor joining, but allows multiple connections among taxa

- At each iteration, choose the pairing of taxa that minimizes the total tree length (that's the Q-criterion again)

Bryant and Moulton (2004) *Molecular Biology and Evolution*

Like neighbor-joining…

At every step, choose the pair of remaining clusters of taxa C_i, C_j that minimize the same Q criterion as before with NJ:

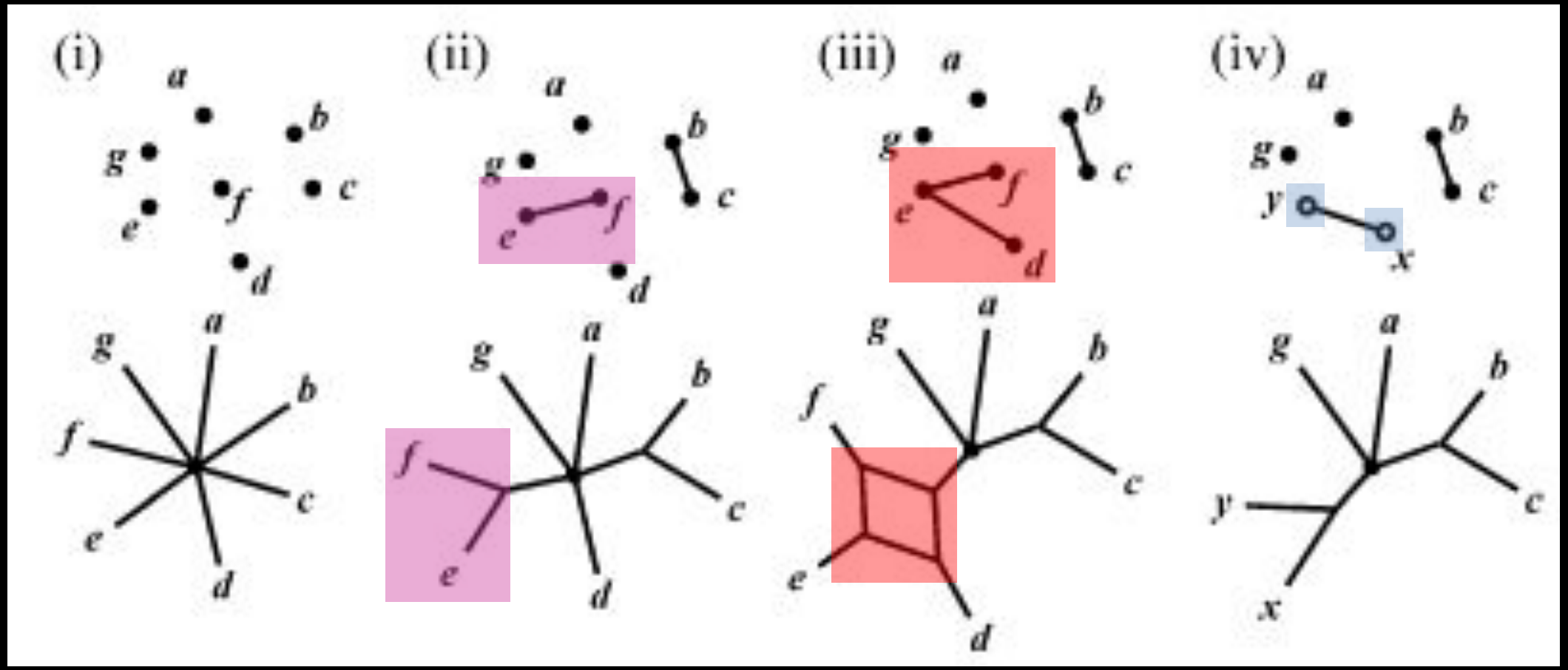
Distance matrix entry for (x,y)

$$\delta(x,y) - \frac{1}{(n-2)}\sum_z \delta(x,z) - \frac{1}{(n-2)}\sum_z \delta(y,z)$$

Weighted distance from x and y to each other leaf z

but x and y remain accessible for pairing off with other taxa

Building the splits graph

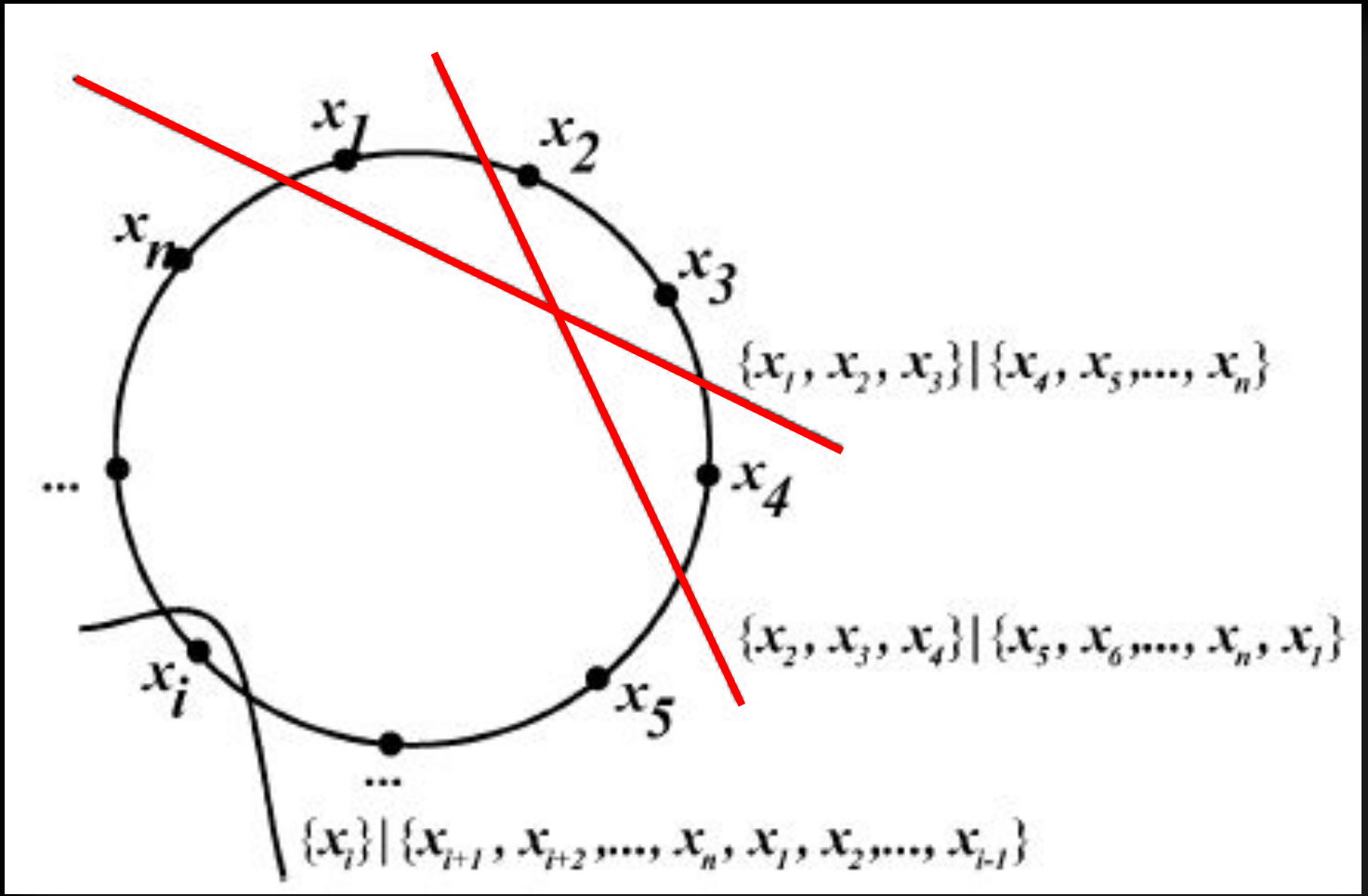


e and *f* optimize the Q criterion – join them but keep them in the pool

Now, *e* and *d* optimize the Q criterion – join them

Replace *d*, *e*, *f* with *x*, *y*

Constraining the splits graph

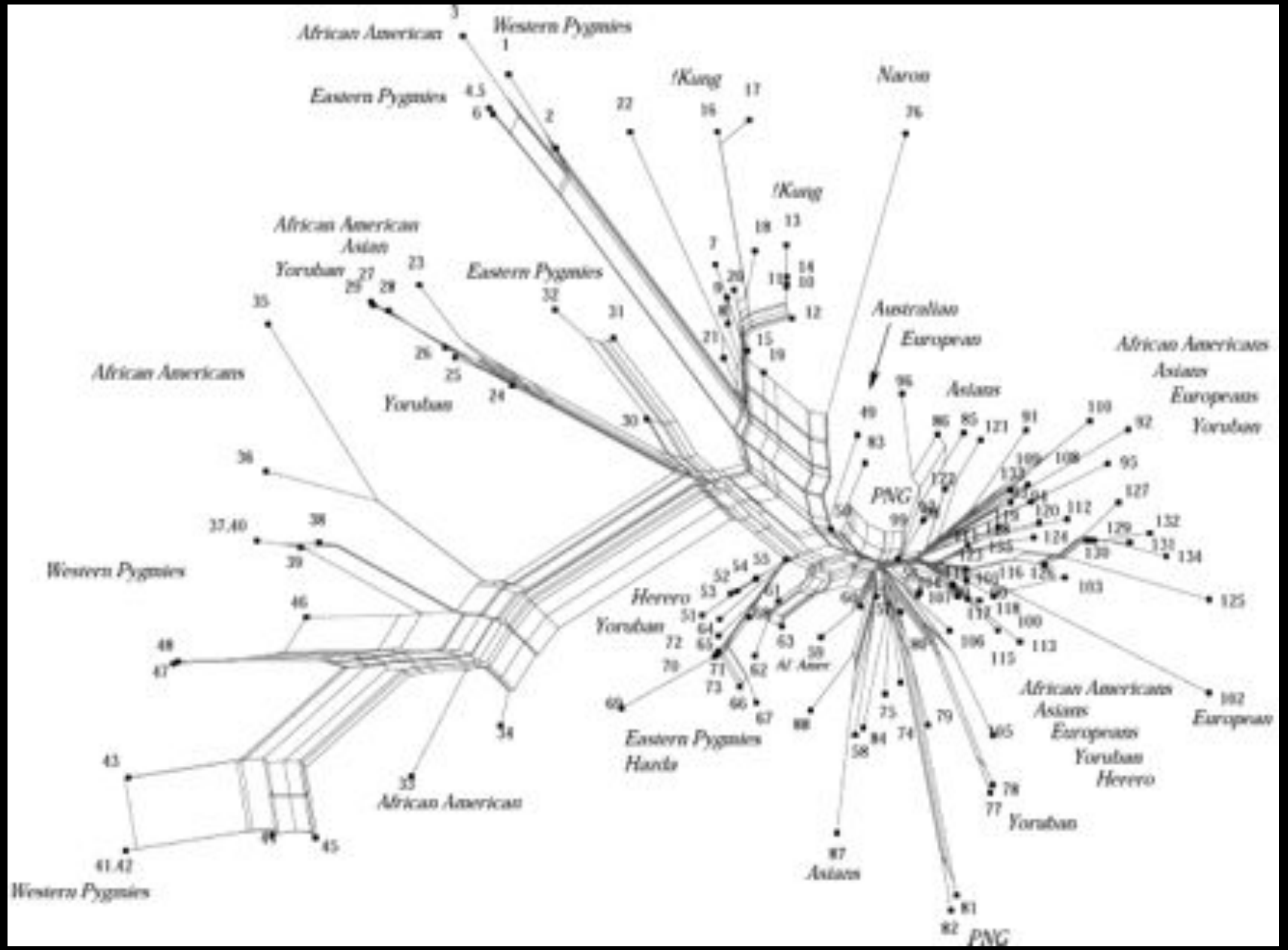


Neighbor-net is guaranteed to produce a circularly compatible set of splits

Fix (x_1, x_2, x_3) and (x_2, x_3, x_4):

- (x_3, x_4, x_5) is legal
- (x_1, x_2, x_4) is not

This leads to a planar graph that can be drawn on a two-dimensional surface without crossings

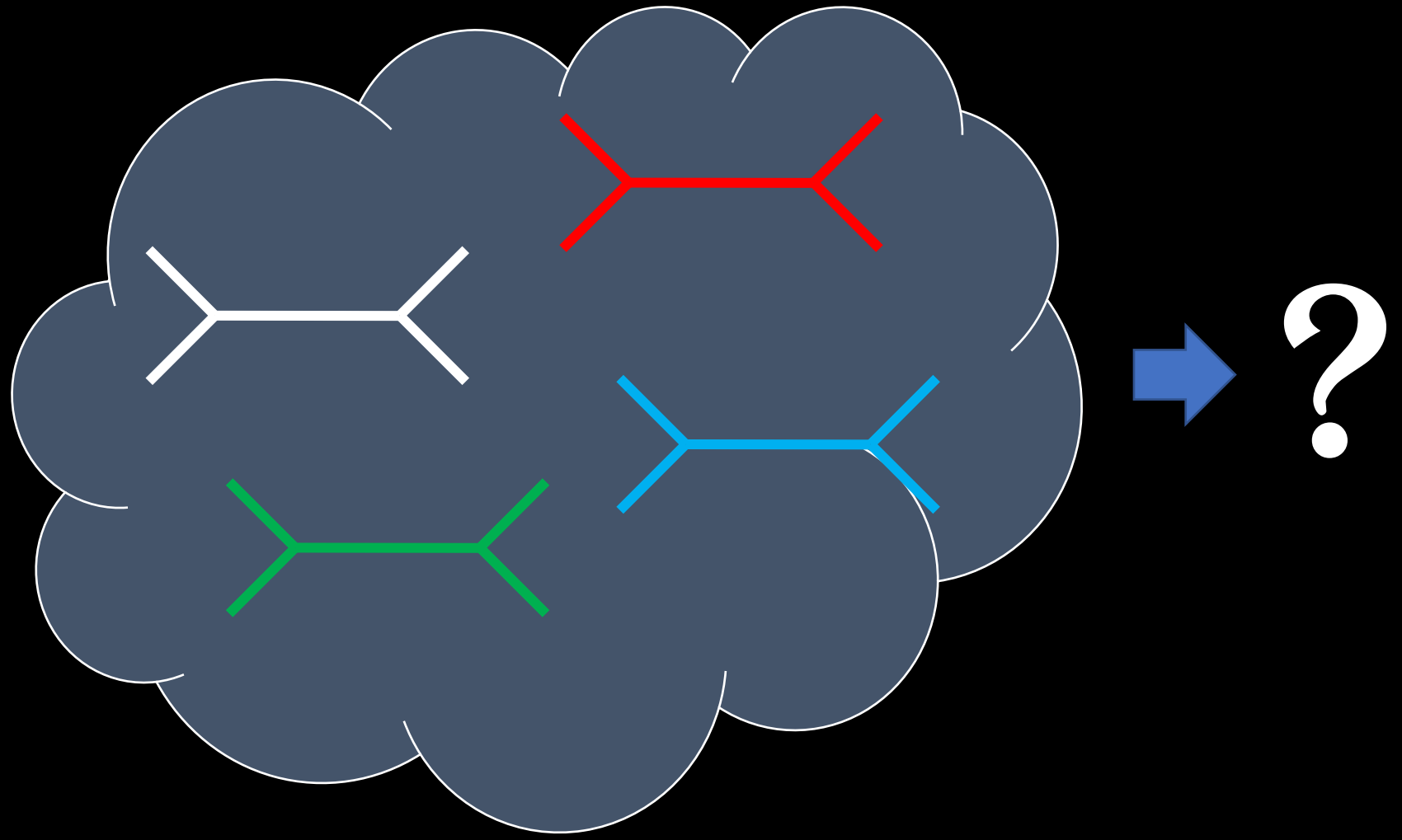


Mitochondrial genomes from 135 individuals

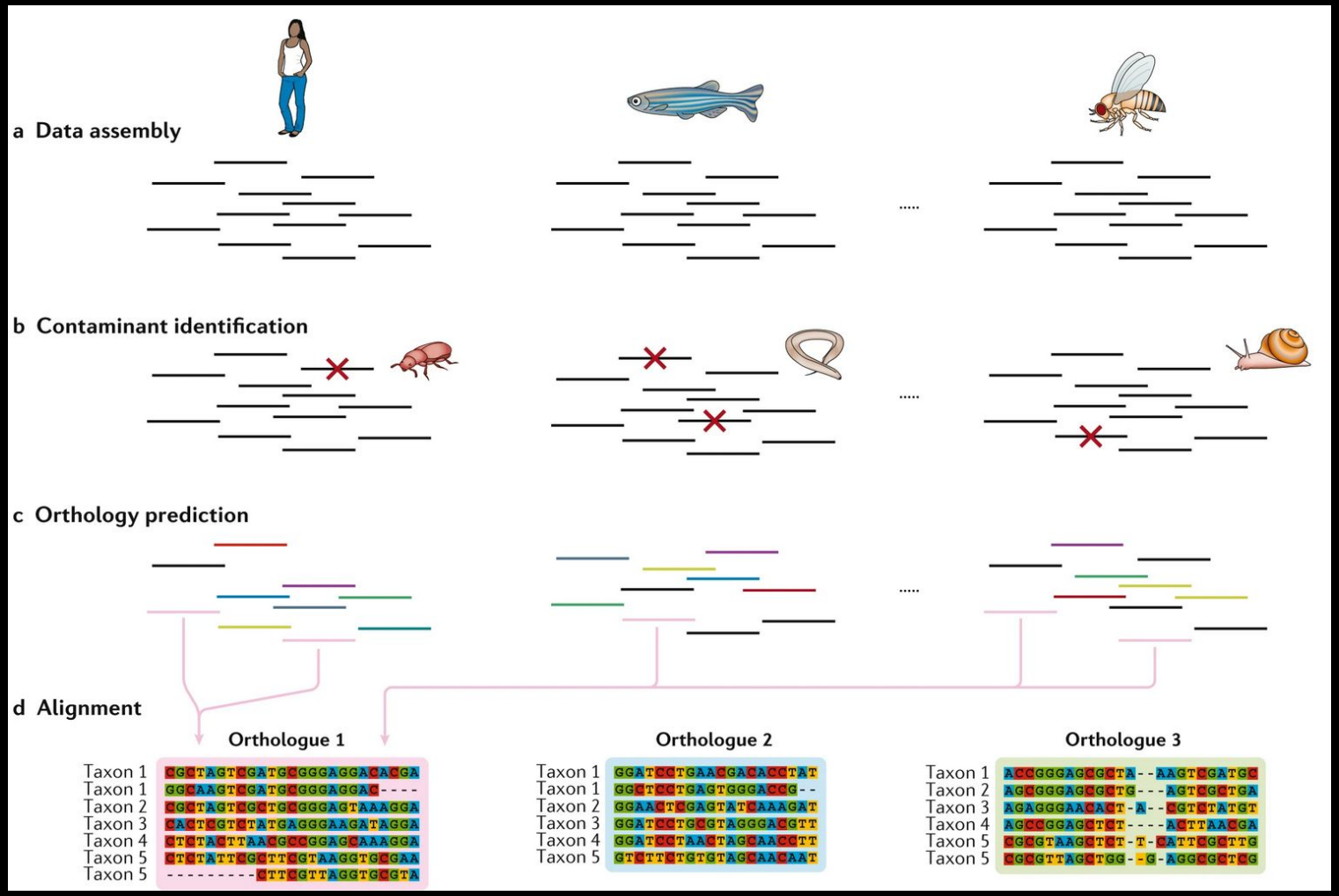
Messy but still planar!

Trees from Many Genes

Reconciling Data Sets

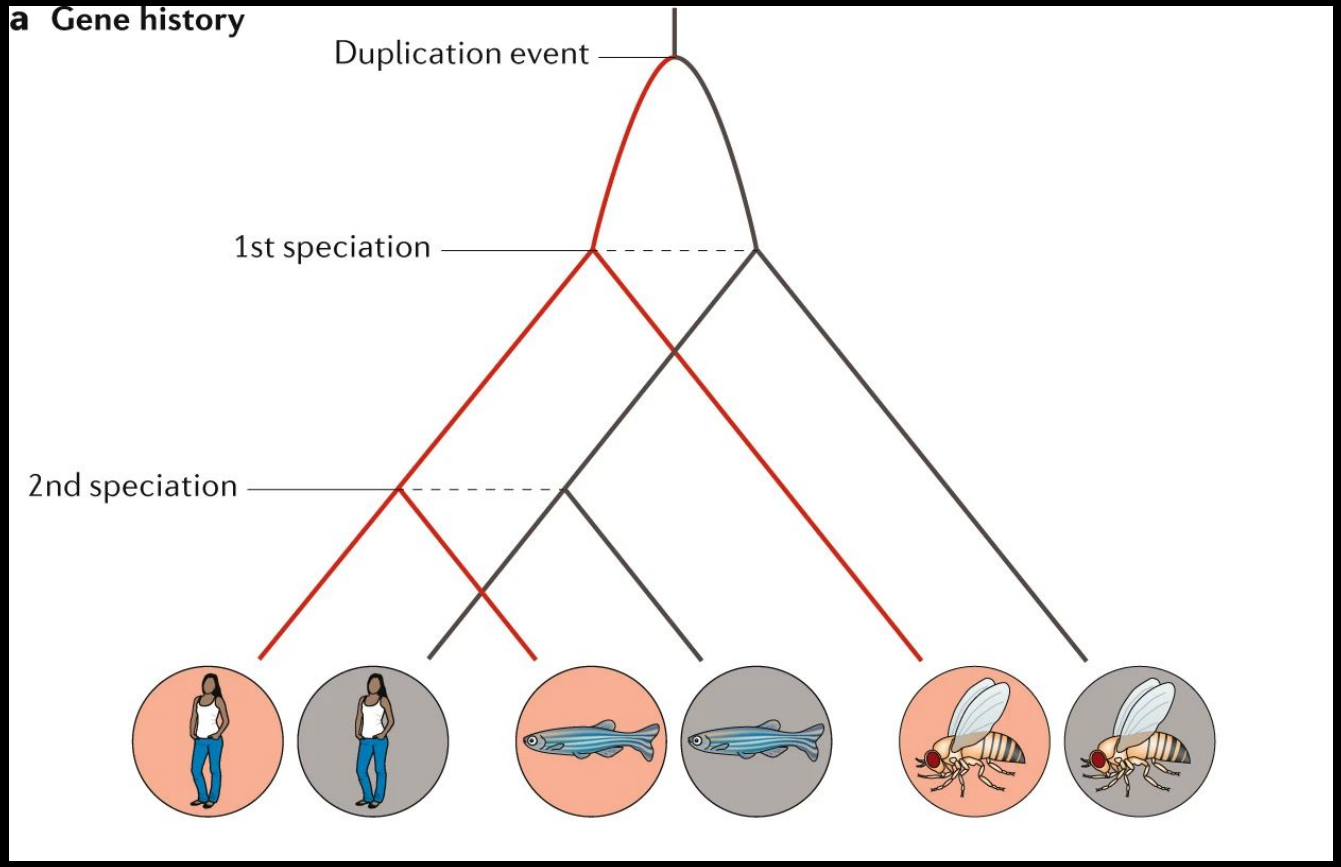


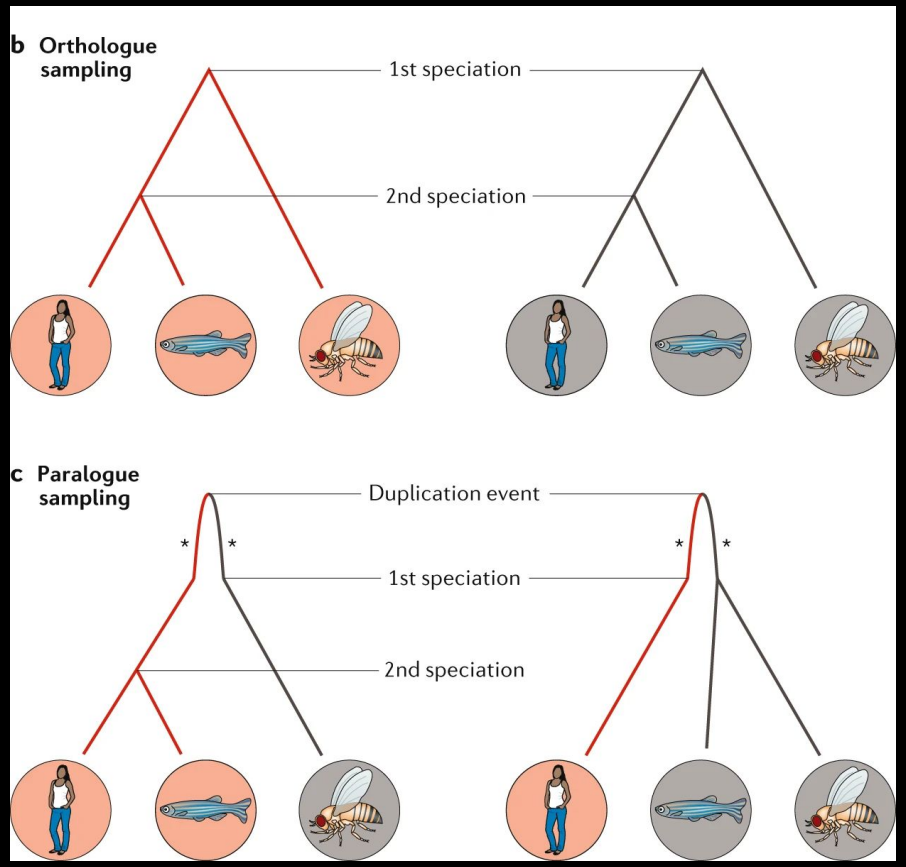
Central to Phylogenomics



Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. "Phylogenetic tree building in the genomic age." *Nature Reviews Genetics* 21.7 (2020): 428-444.

Identifying Orthologs Important



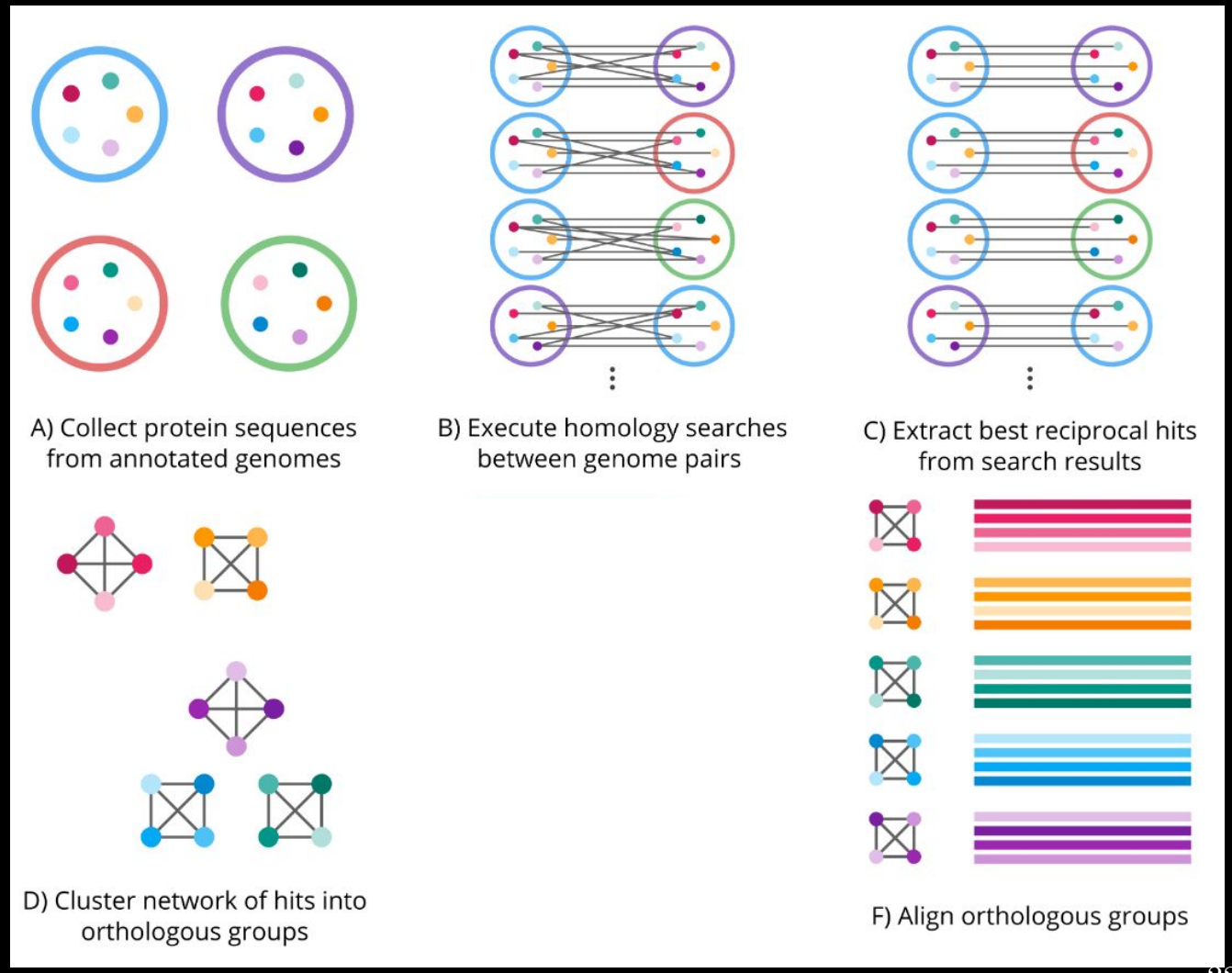


Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. "Phylogenetic tree building in the genomic age." *Nature Reviews Genetics* 21.7 (2020): 428-444.

Graph-Based Orthology Inference

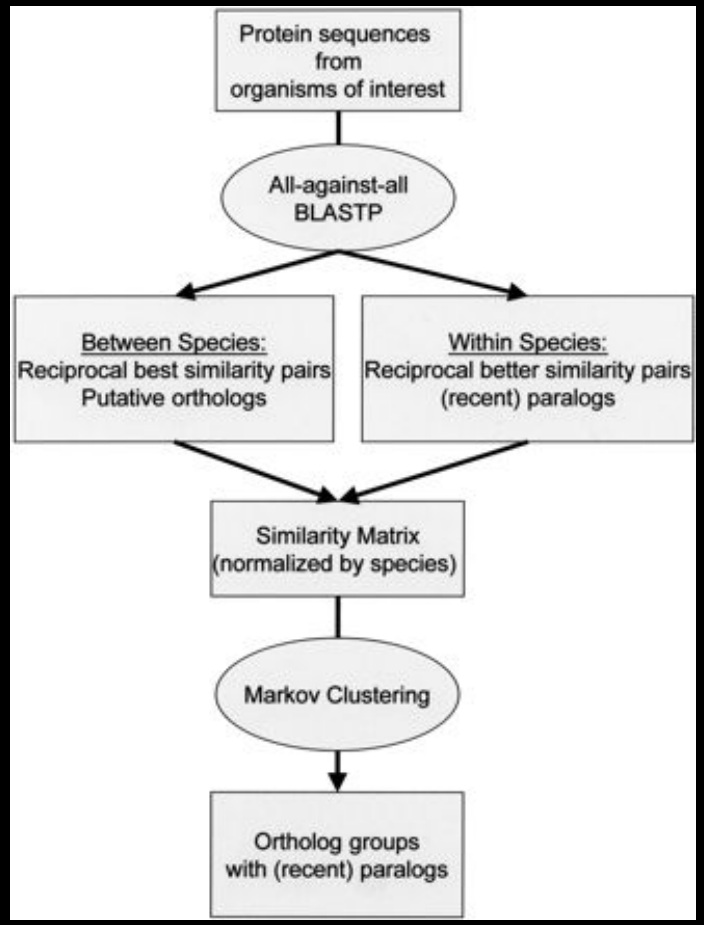
Key assumption:

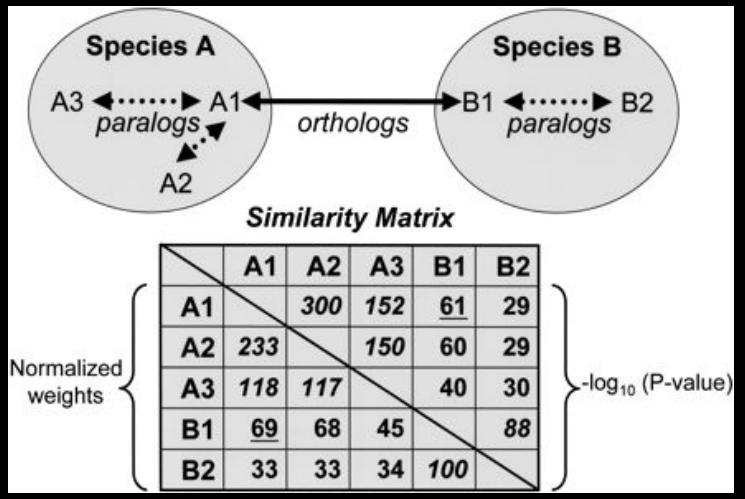
Gene A in Species 1
most reciprocally similar to
Gene A in Species 2



https://doi.org/10.1101/2023.01.24.525427

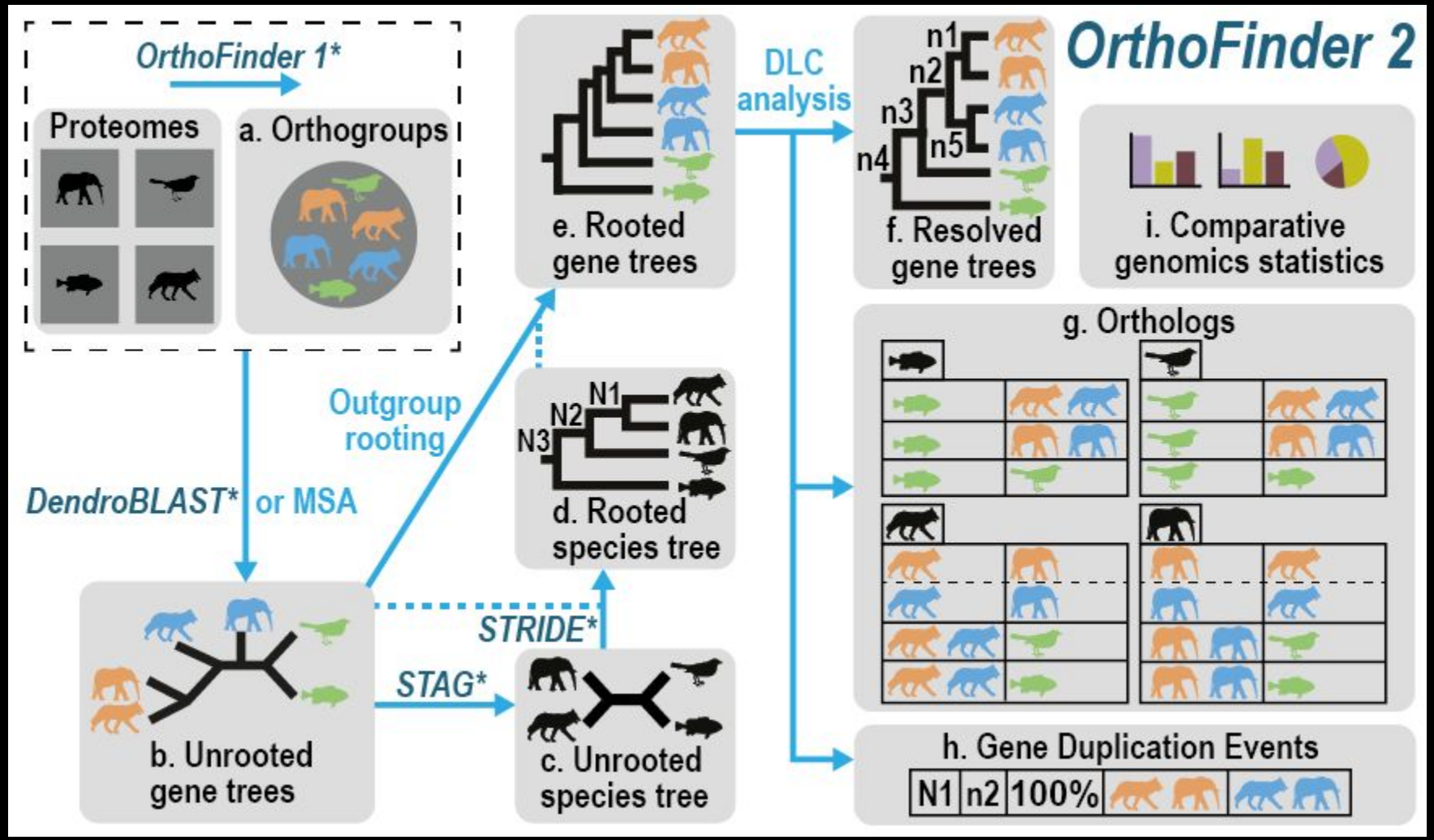
Orthology Inference - OrthoMCL



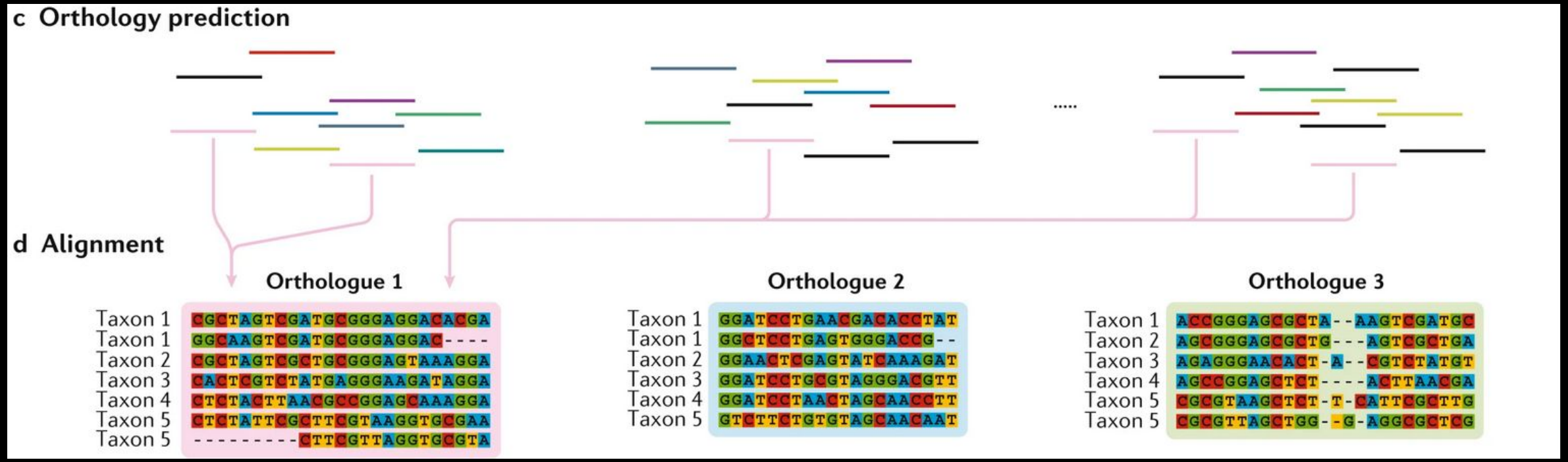


	A1	A2	A3	B1	B2
A1		*300*	*152*	61	29
A2	*233*		*150*	60	29
A3	*118*	*117*		40	30
B1	69	68	45		*88*
B2	33	33	34	*100*	

Orthology Inference - OrthoFinder



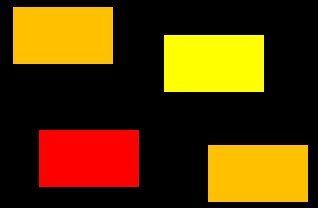
Many “ortholog” alignments: what now?



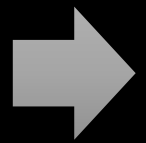
Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. "Phylogenetic tree building in the genomic age." *Nature Reviews Genetics* 21.7 (2020): 428-444.

The first way: supermatrices

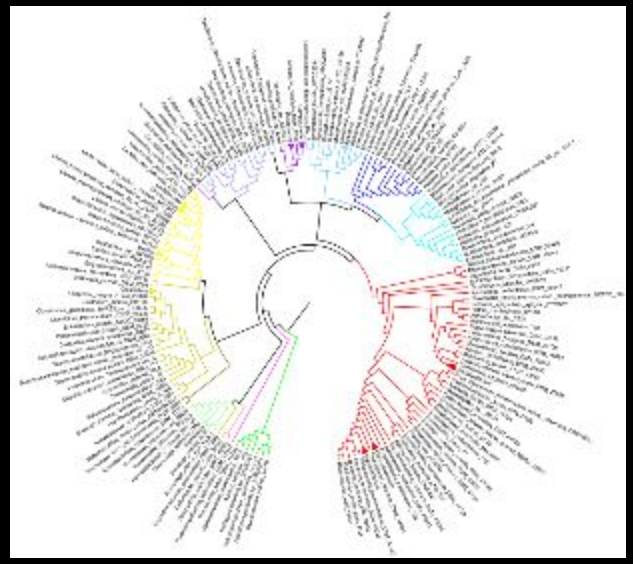
Many alignments



One alignment



One tree or network

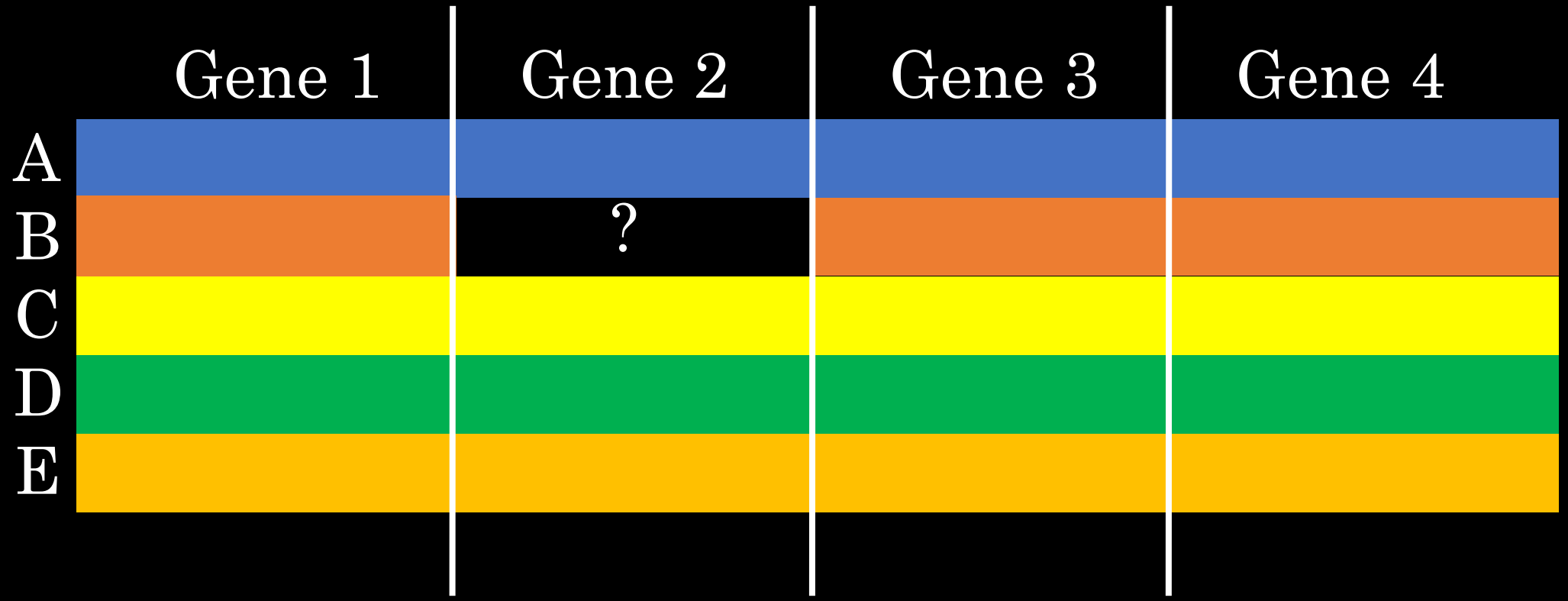


Supermatrices

Very simple: since many alignment columns are (potentially) needed for ML inference, chain many single-gene alignments together and build a tree

Inaccuracy due to small sample size is sometimes called ‘stochastic error’

Inaccuracy due to violation of model assumptions are called “systematic error”

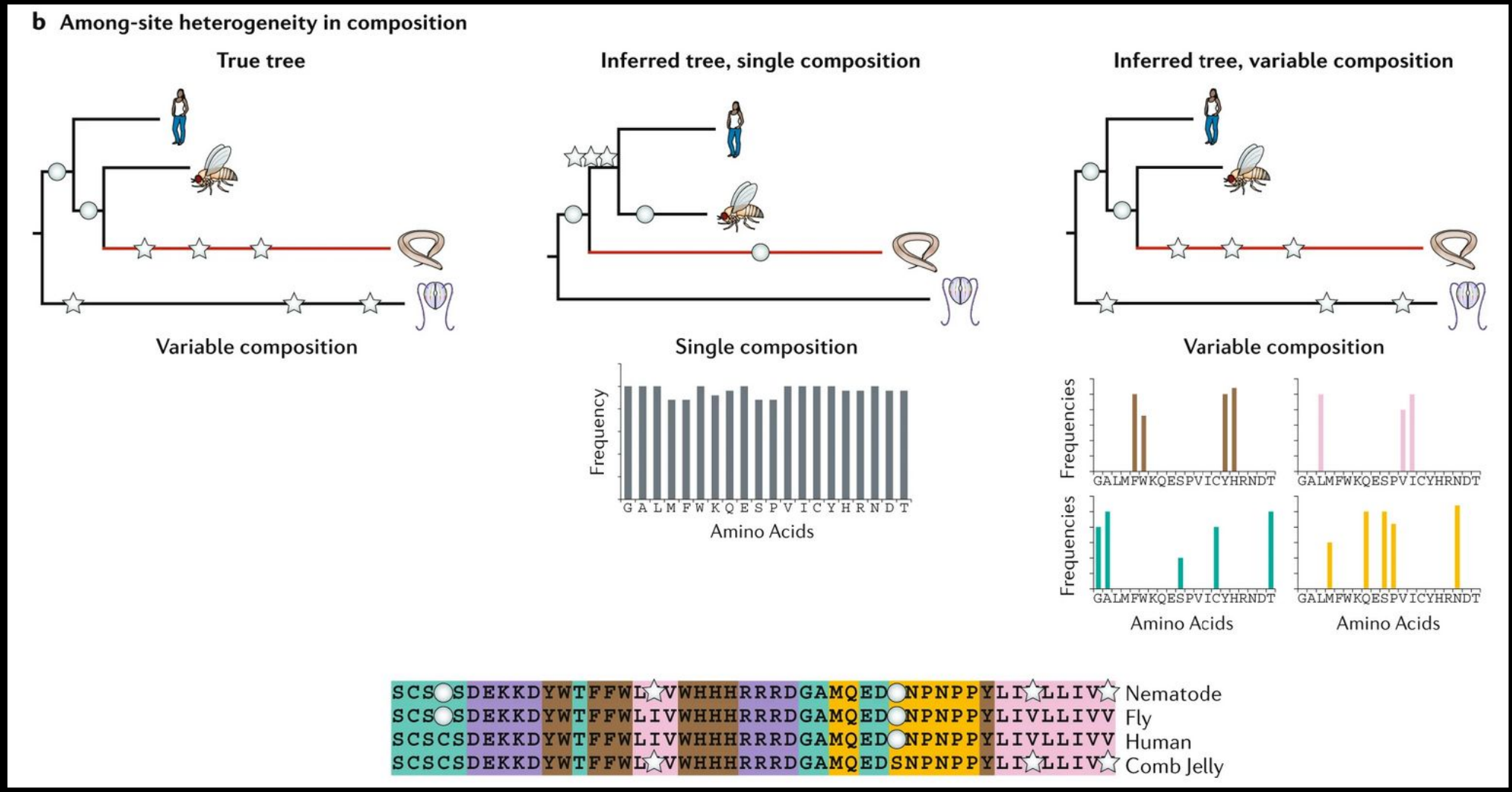


What do we do about genes that are missing from one organism?

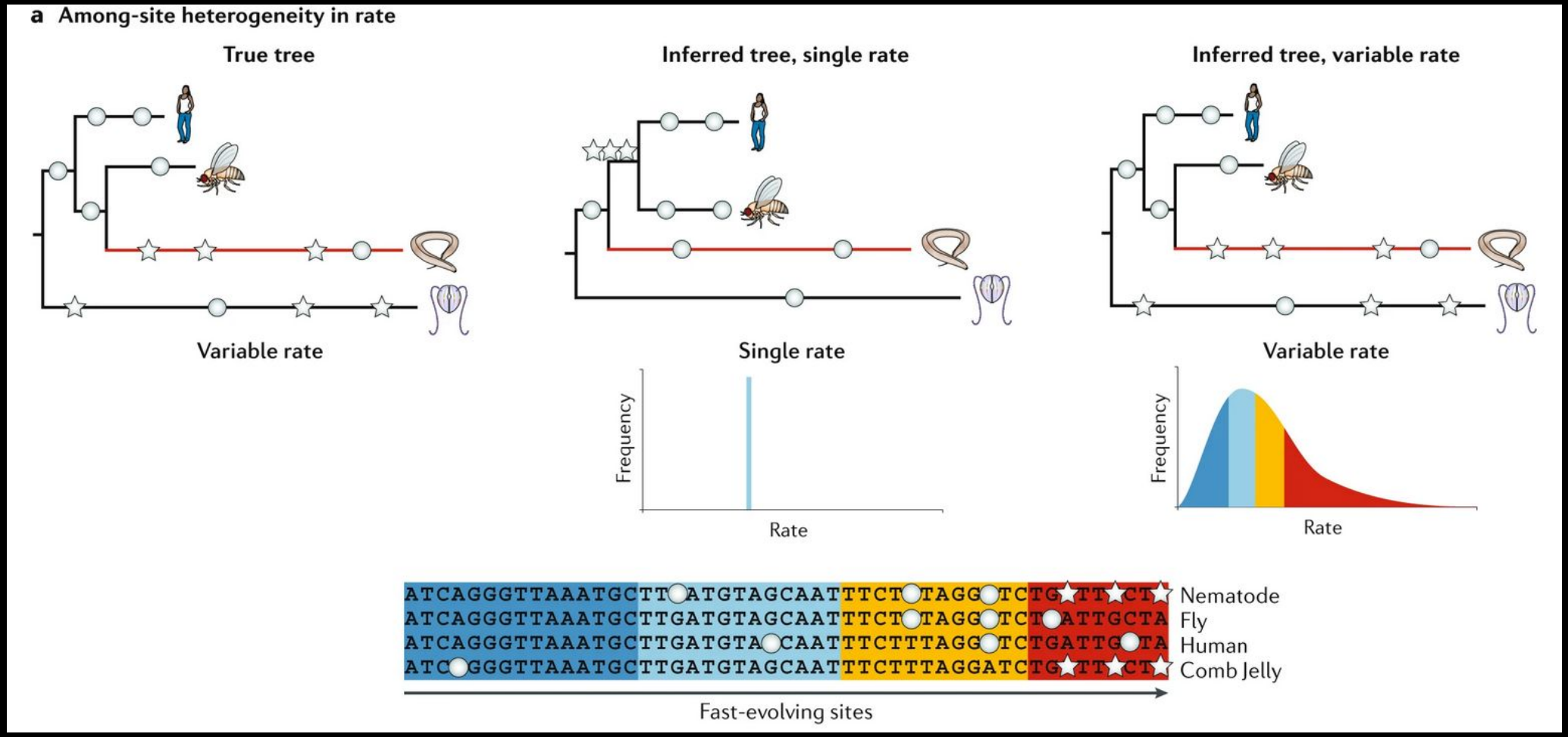
Supermatrices

- Also known as ‘concatenated phylogeny’
- Since we’re building a tree, assumes that each gene has the same evolutionary history (hmmm…)
- Set a minimum threshold (presence in % of organisms) for inclusion of each gene
- Typically the model will allow different genes to evolve at different rates

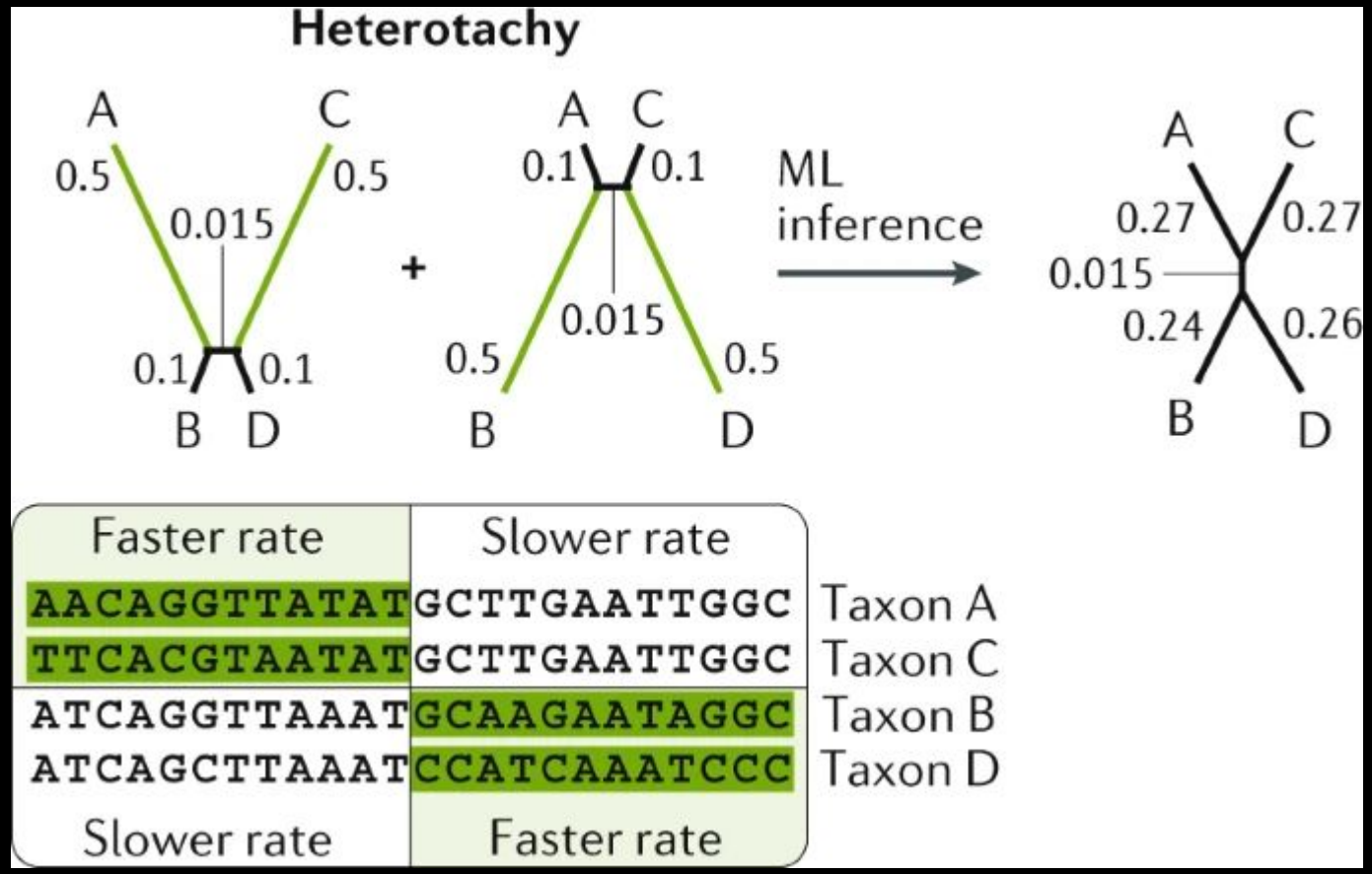
Compositional Biases



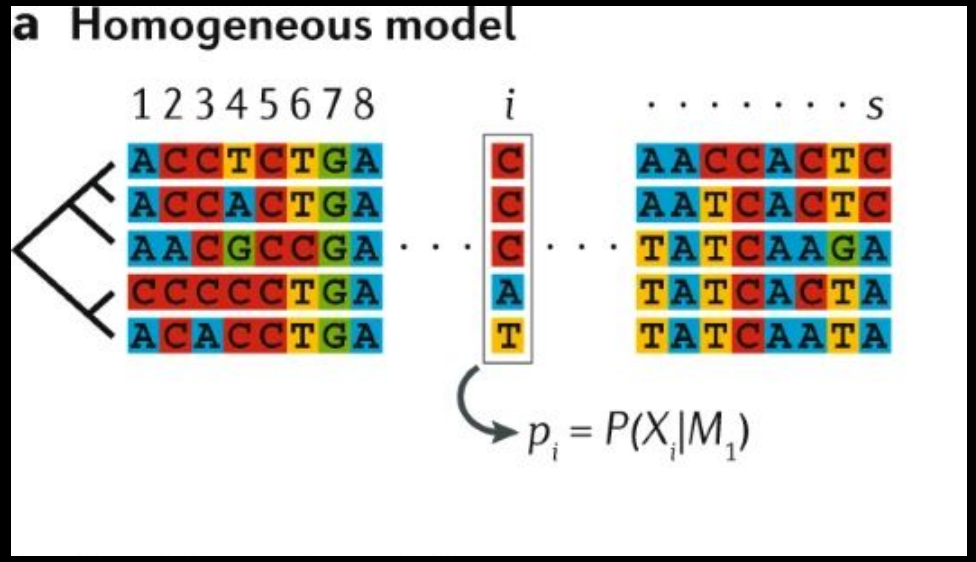
Among-Site Rate Heterogeneity

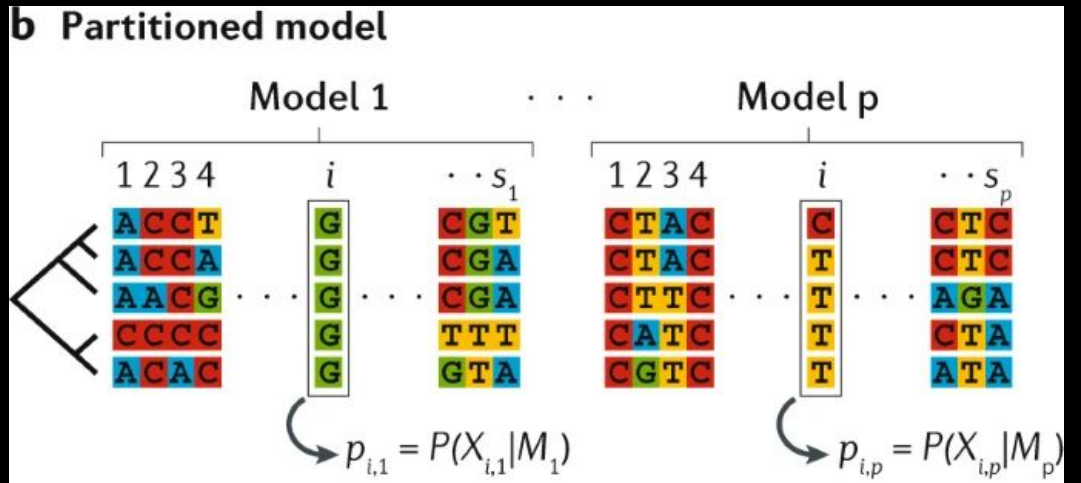


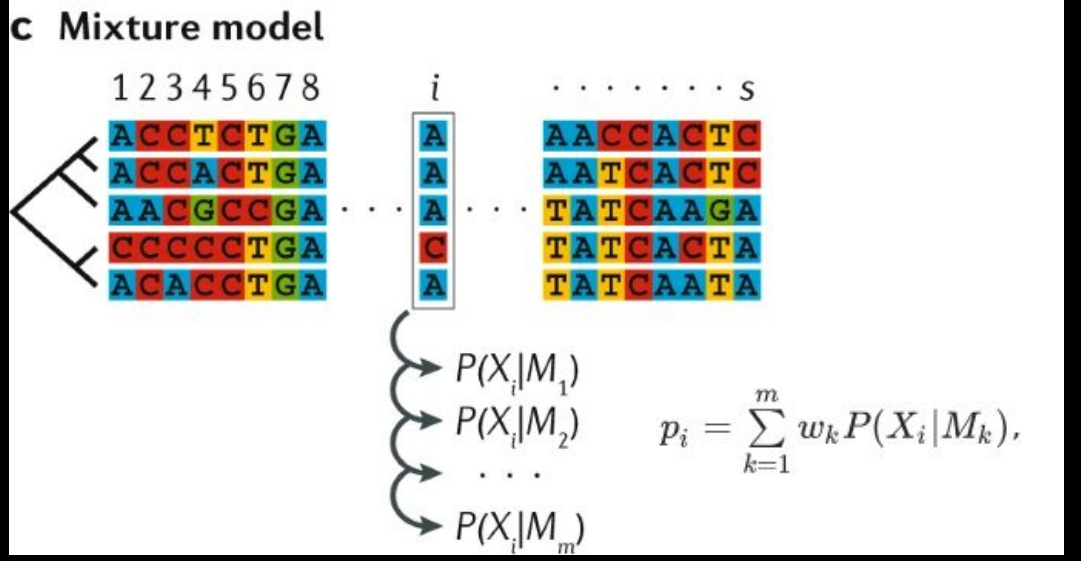
Heterogeneity across time/lineages



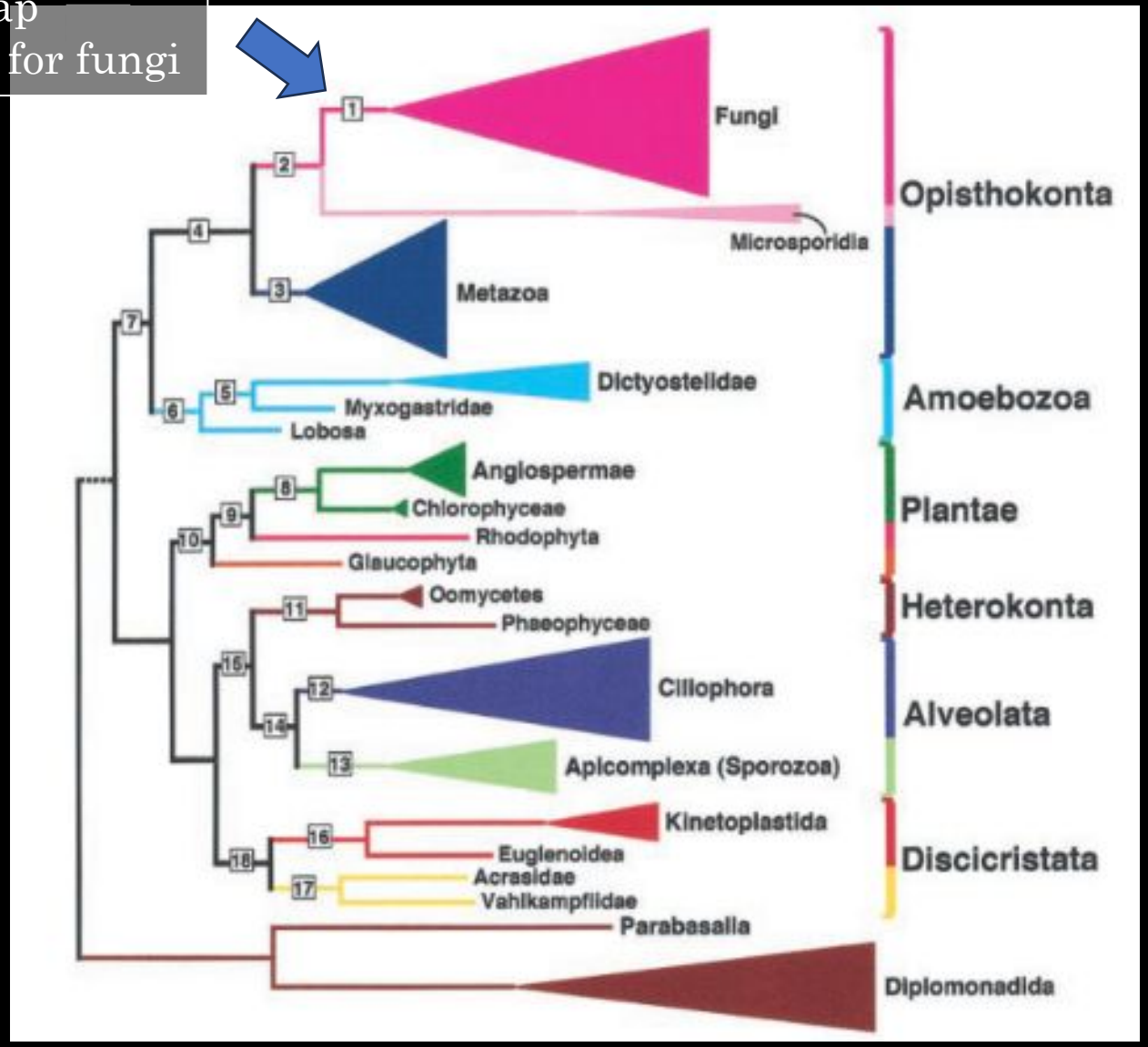
Partition and Mixtures

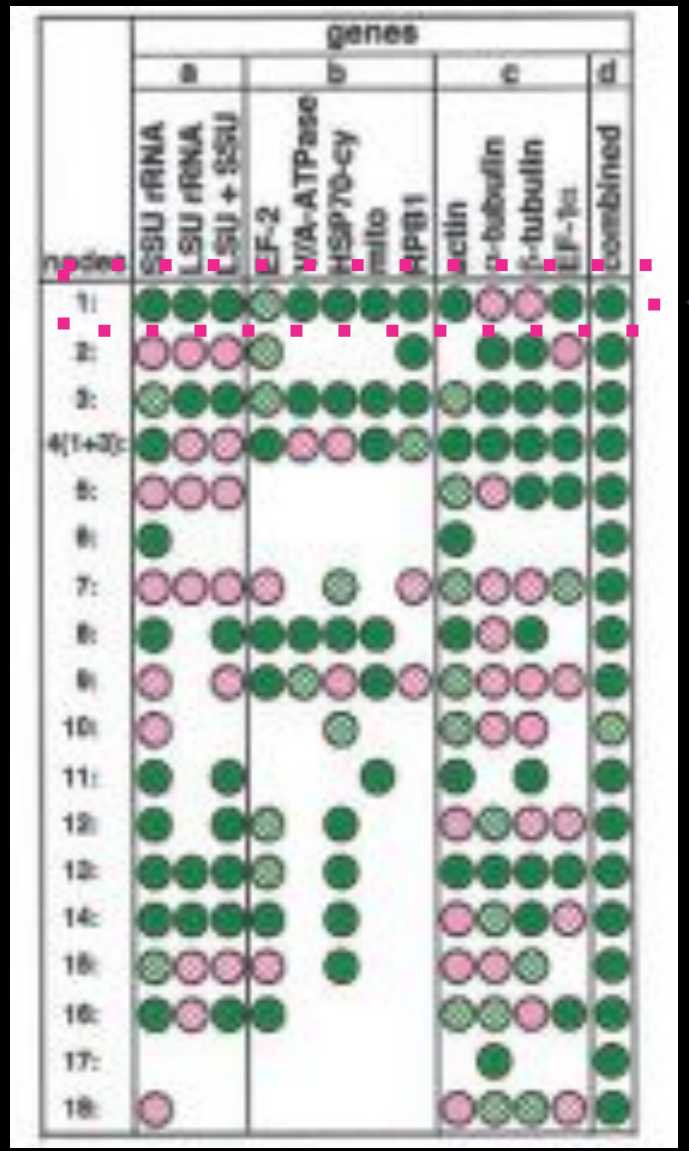






Bootstrap support for fungi





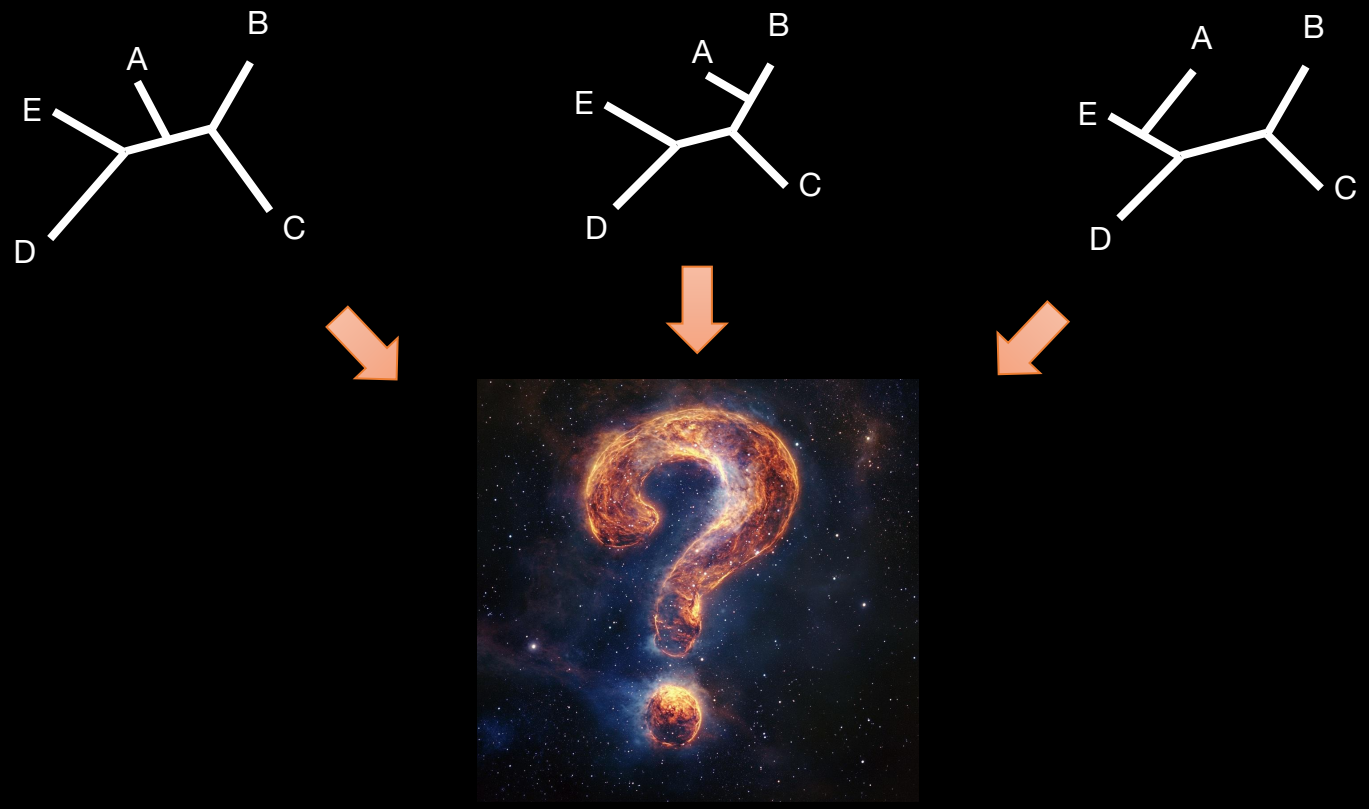
Tree based on actin, alpha/beta tubulin, EF-1alpha
Green circles = strong support, red circles = poor support

Baldauf et al. (2000)


Trees from many trees`

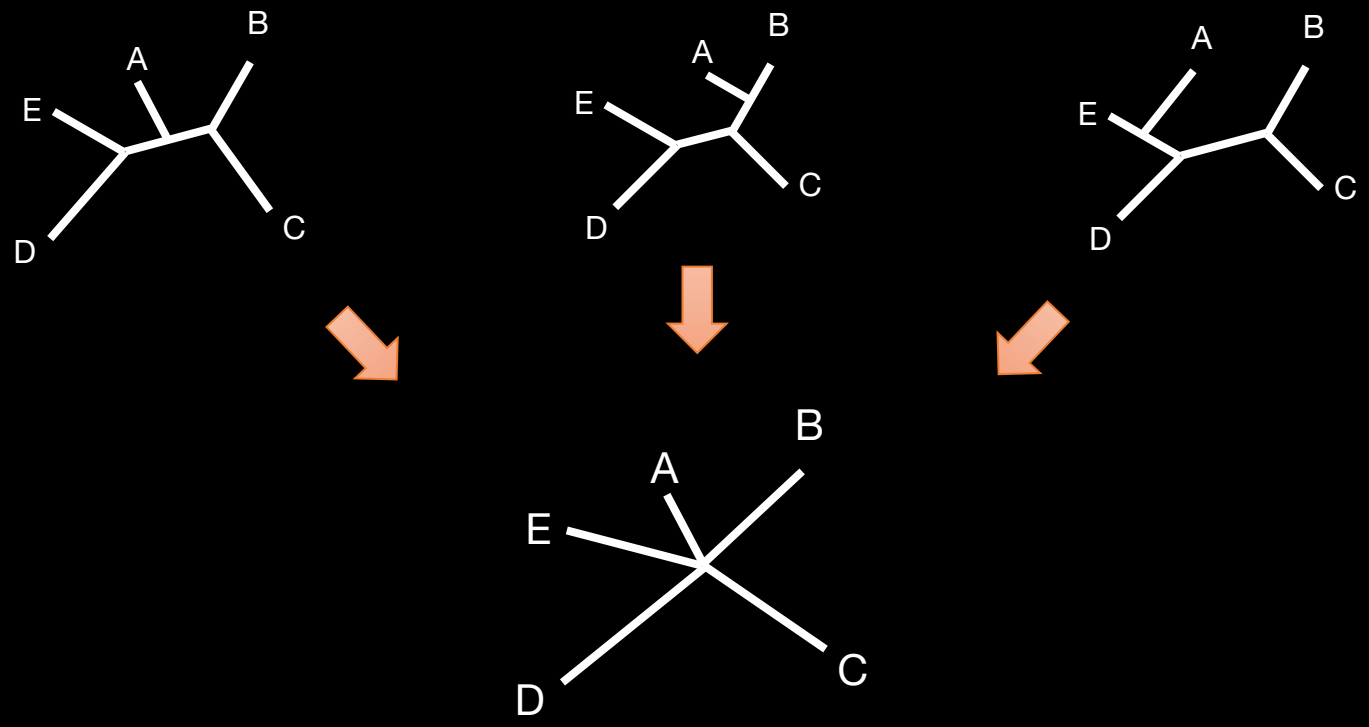
Consensus trees

If each member of a set of trees covers the exact same set of taxa, then they can be combined into a single tree representation



Strict consensus

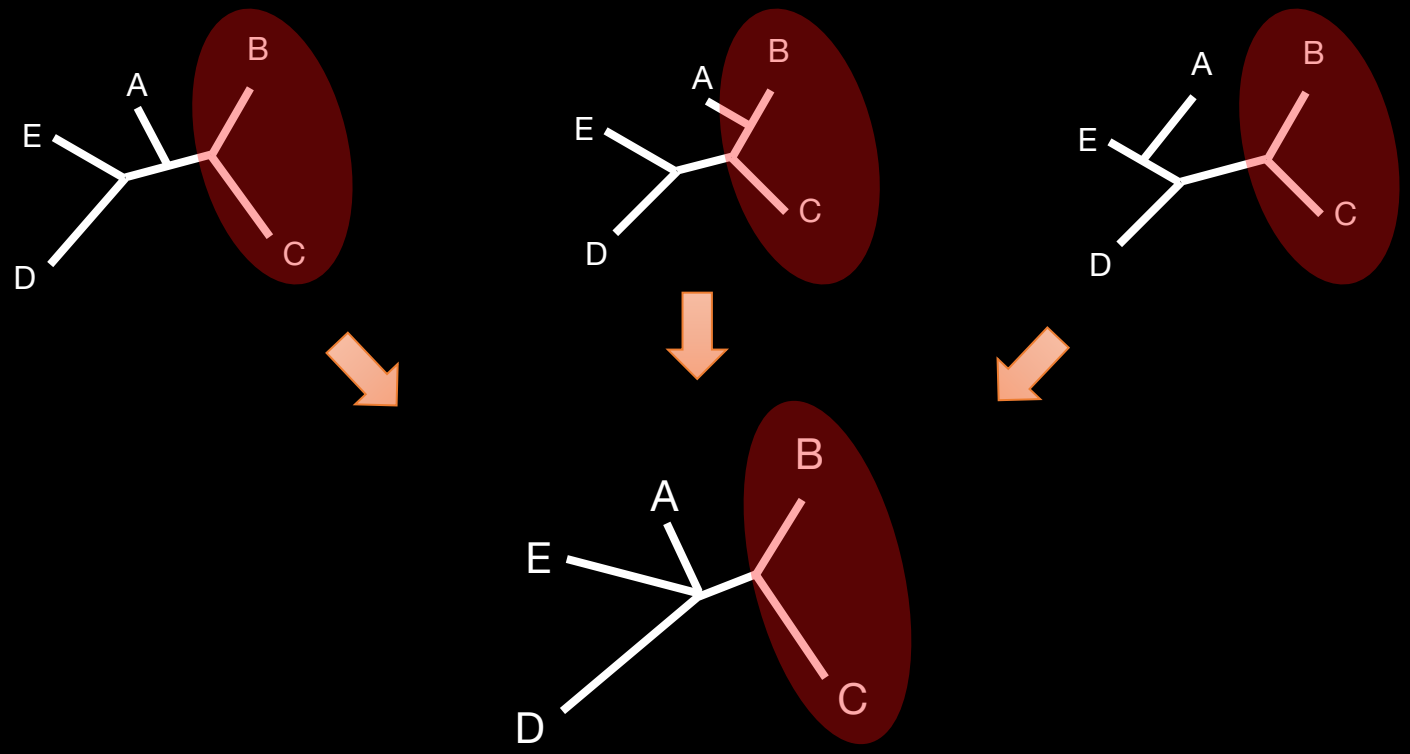
The consensus tree contains only those bipartitions that are found in all trees



No bipartitions / splits are found in all three trees, so consensus is a star

Majority-rule consensus

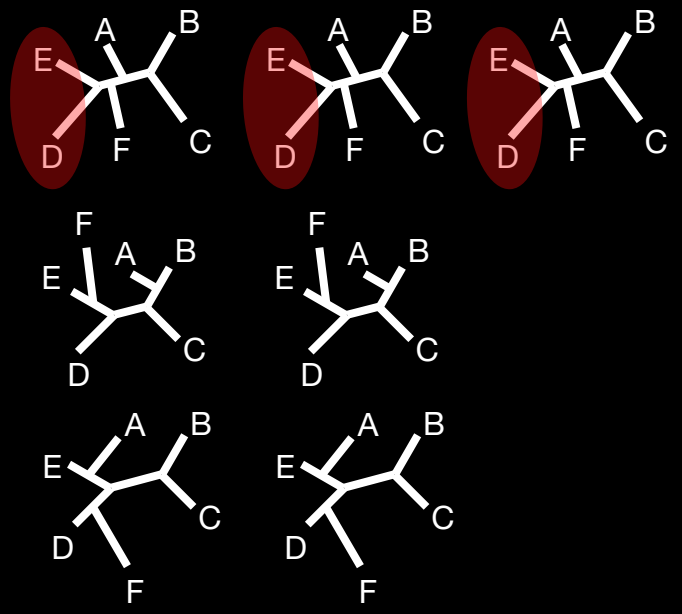
The consensus tree contains only those bipartitions that are found in the majority of trees

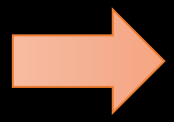


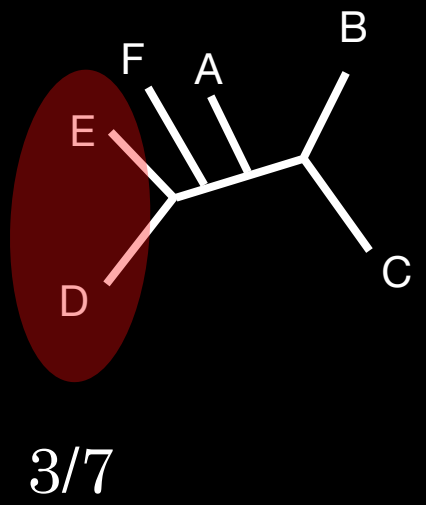
(ADE|BC) found in 2/3 trees

Extended majority-rule consensus

- Greedily add features from the set of trees until the consensus tree is completely resolved







3/7

(ABCF|DE) not found in a majority of trees, but is more common than any other resolution of the two

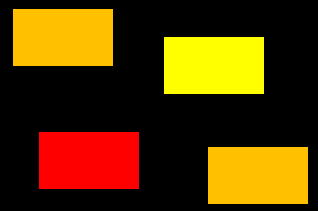
Supertrees

- What if the set of trees cover overlapping but non-identical sets of taxa?

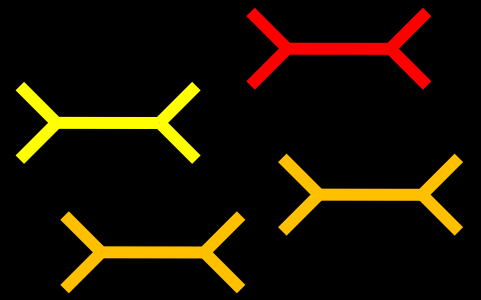
- We can build "all the trees of all the genes" and then combine the trees

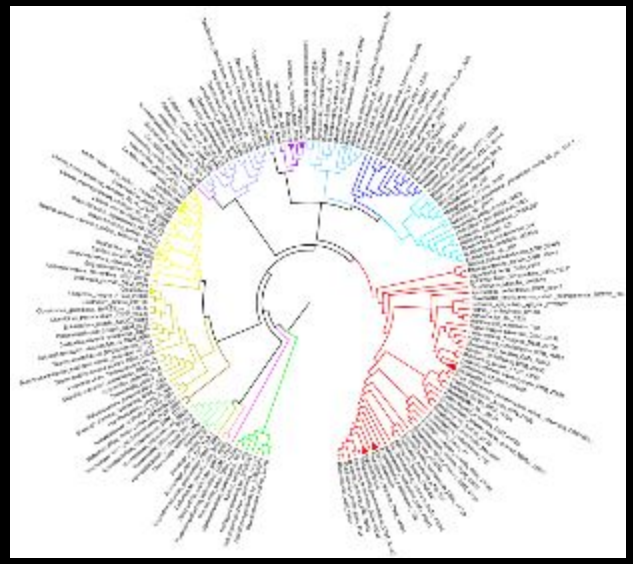
Supertrees

Many alignments



Many trees



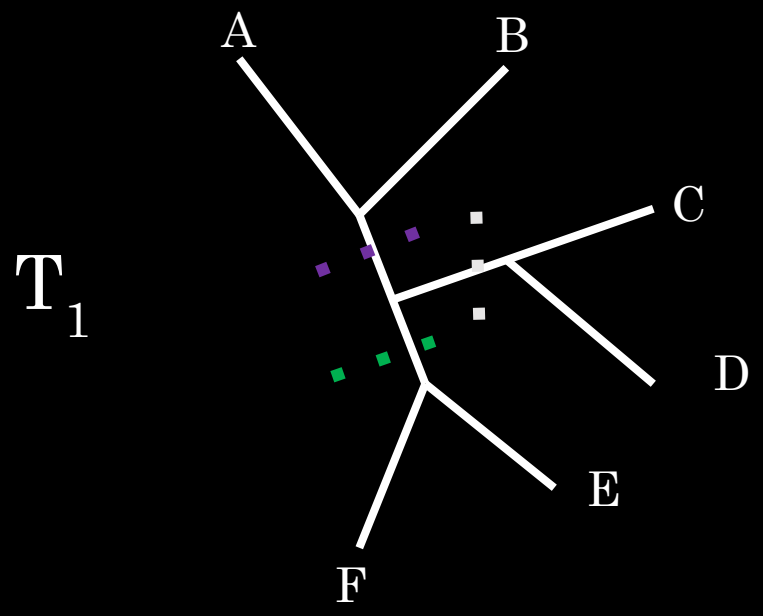


One tree or network

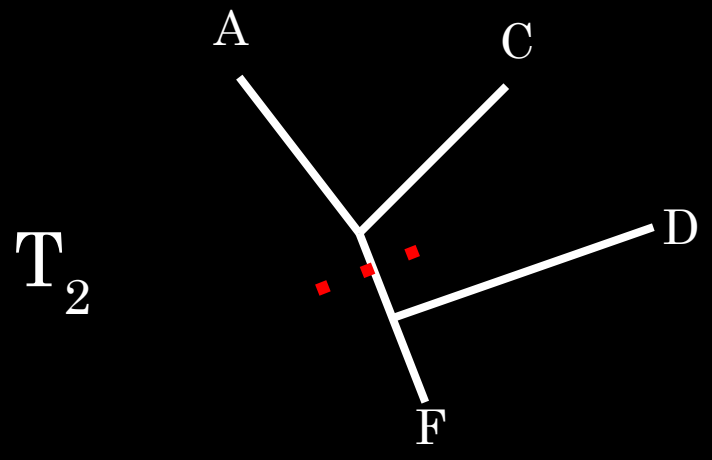
Matrix Representation with Parsimony

- Build an alignment from trees (!?)
- Each bipartition is represented as a character (column) in the alignment

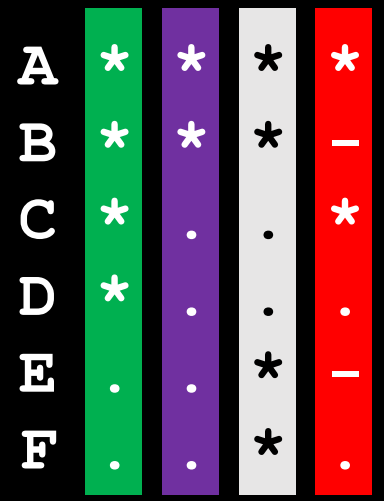
Baum (1992) *Taxon*
Ragan (1992) *Molecular Phylogenetics and Evolution*



3 informative bipartitions



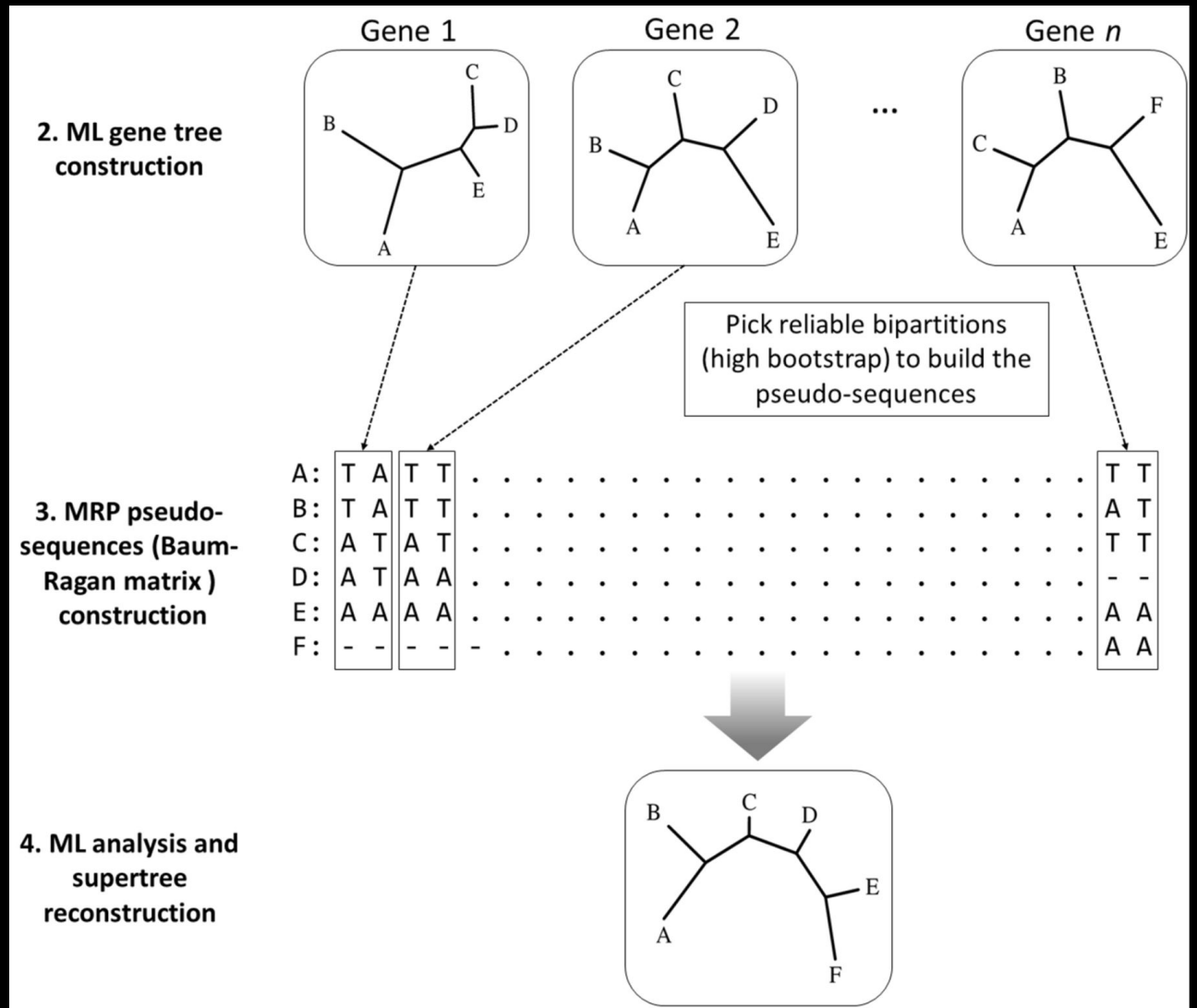
1 informative bipartition



T_1 does not have B and E, so these are gaps

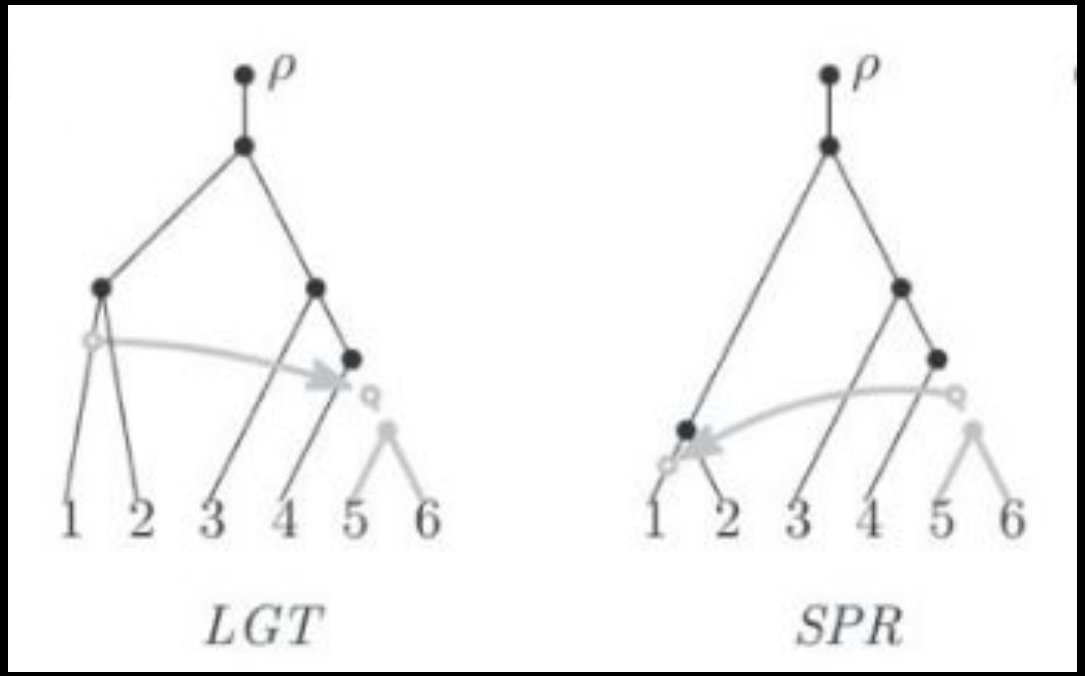
Run parsimony!

Pseudosequence Matrix



SPR supertrees

Lateral gene transfer events are equivalent to subtree prune-and-regraft moves

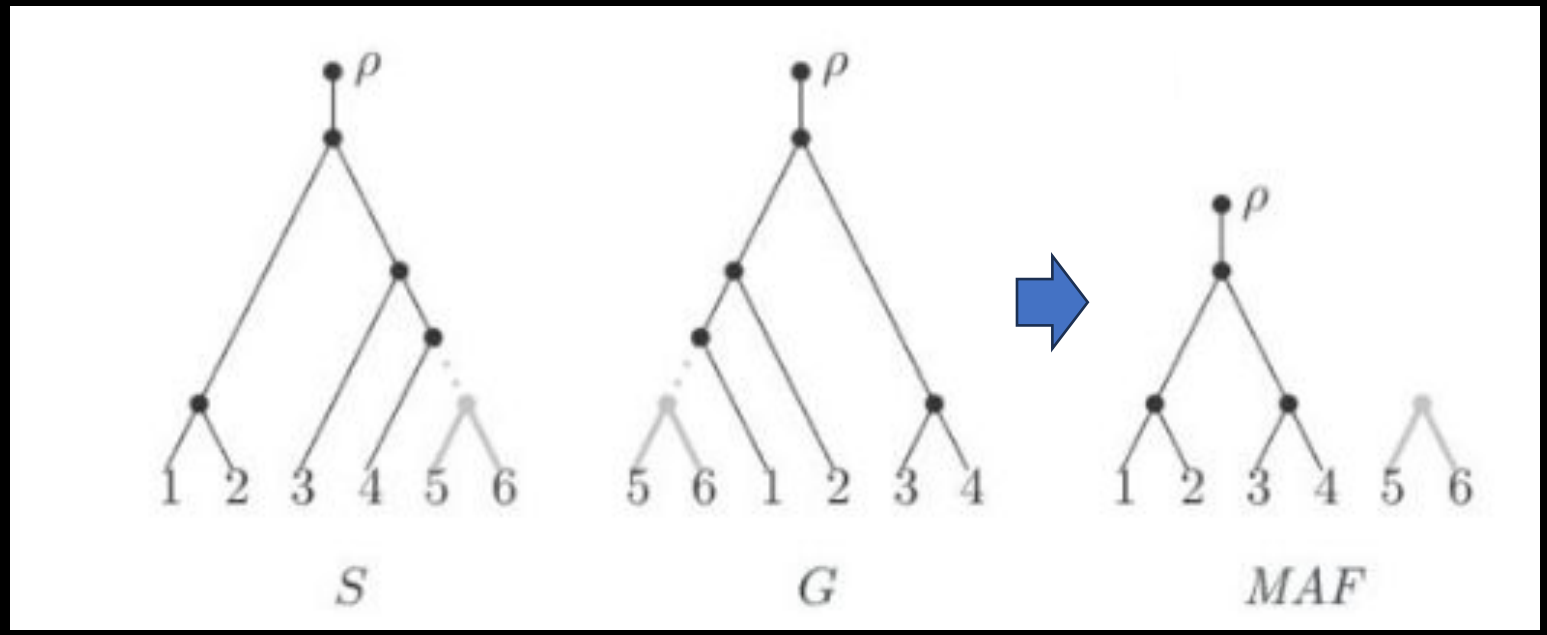


Naïve approach: apply all possible SPR moves in a breadth-first search to reconcile reference ("species") tree with gene tree

Beiko and Hamilton (2005) *BMC Evolutionary Biology*

SPR supertrees

Smarter approach: cut edges in both trees to generate maximum agreement forests
Previous algorithms were exponential in the size of the tree
rSPR algorithm is exponential in the SPR distance (much smaller!)



Chris Whidden's algorithm: $O(2.42^k n)$, where
n = number of leaves
k = SPR distance between two trees

Whidden et al. (2014) *Syst Biol*

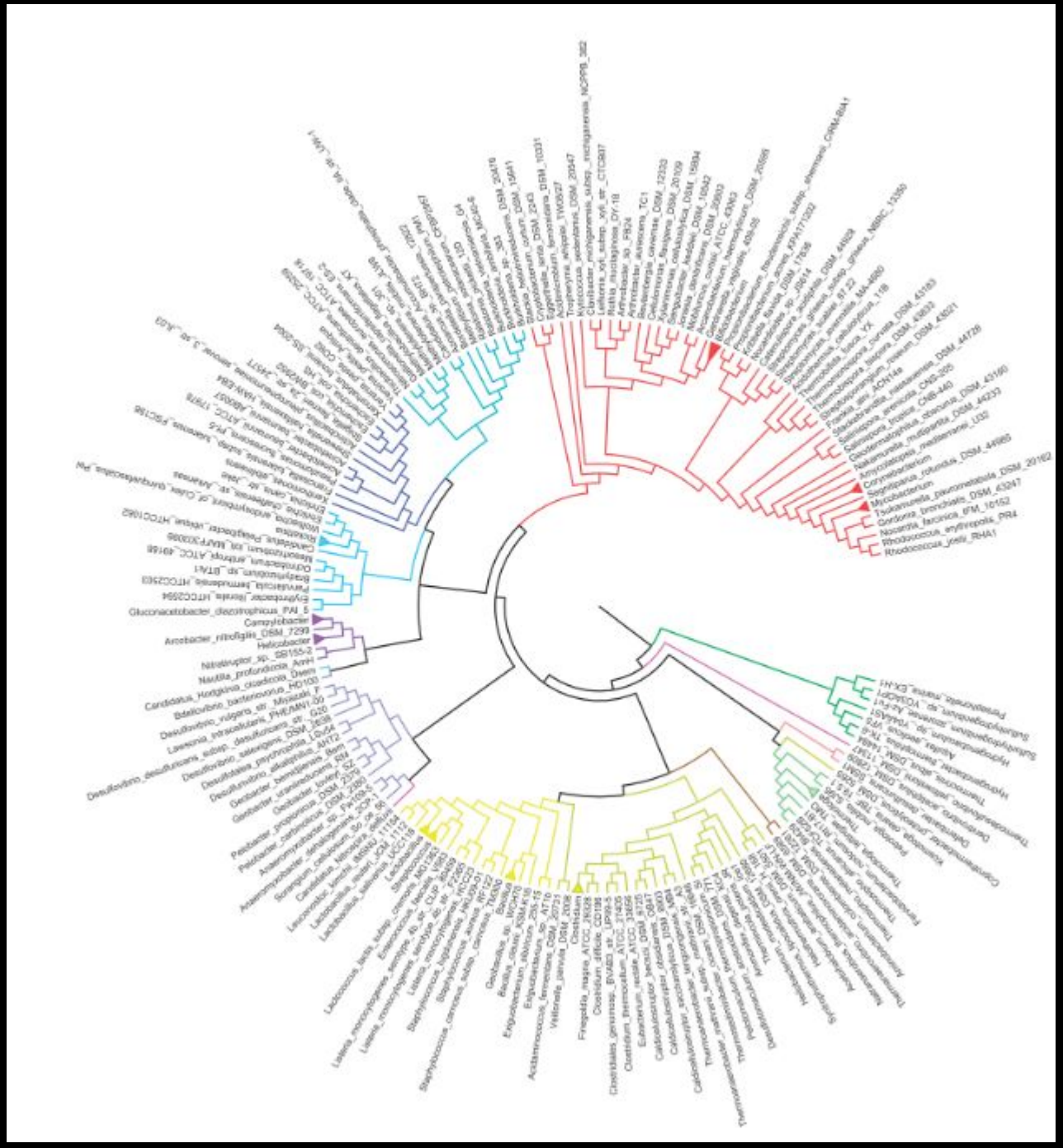
SPR supertrees

Given a set of input trees $t_1, t_2, \ldots t_p$,
find the supertree S that minimizes the total SPR distance to all t_i

Step 1:
Stepwise addition (start with 4 most-common taxa) to build the initial tree

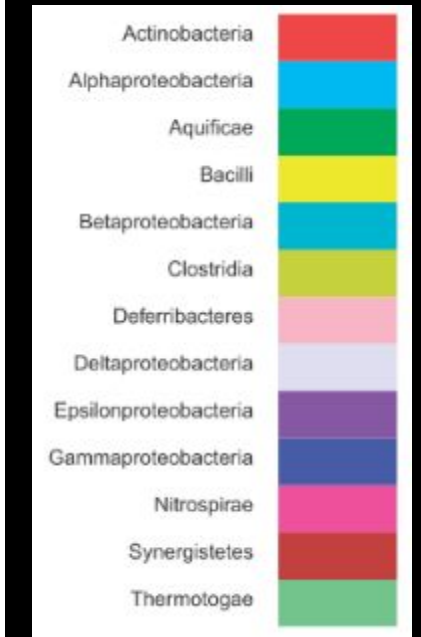
Step 2:
Propose new trees via SPR rearrangements
Various speed-ups including freezing bipartitions with strong support.

Whidden et al. (2014) *Syst Biol*

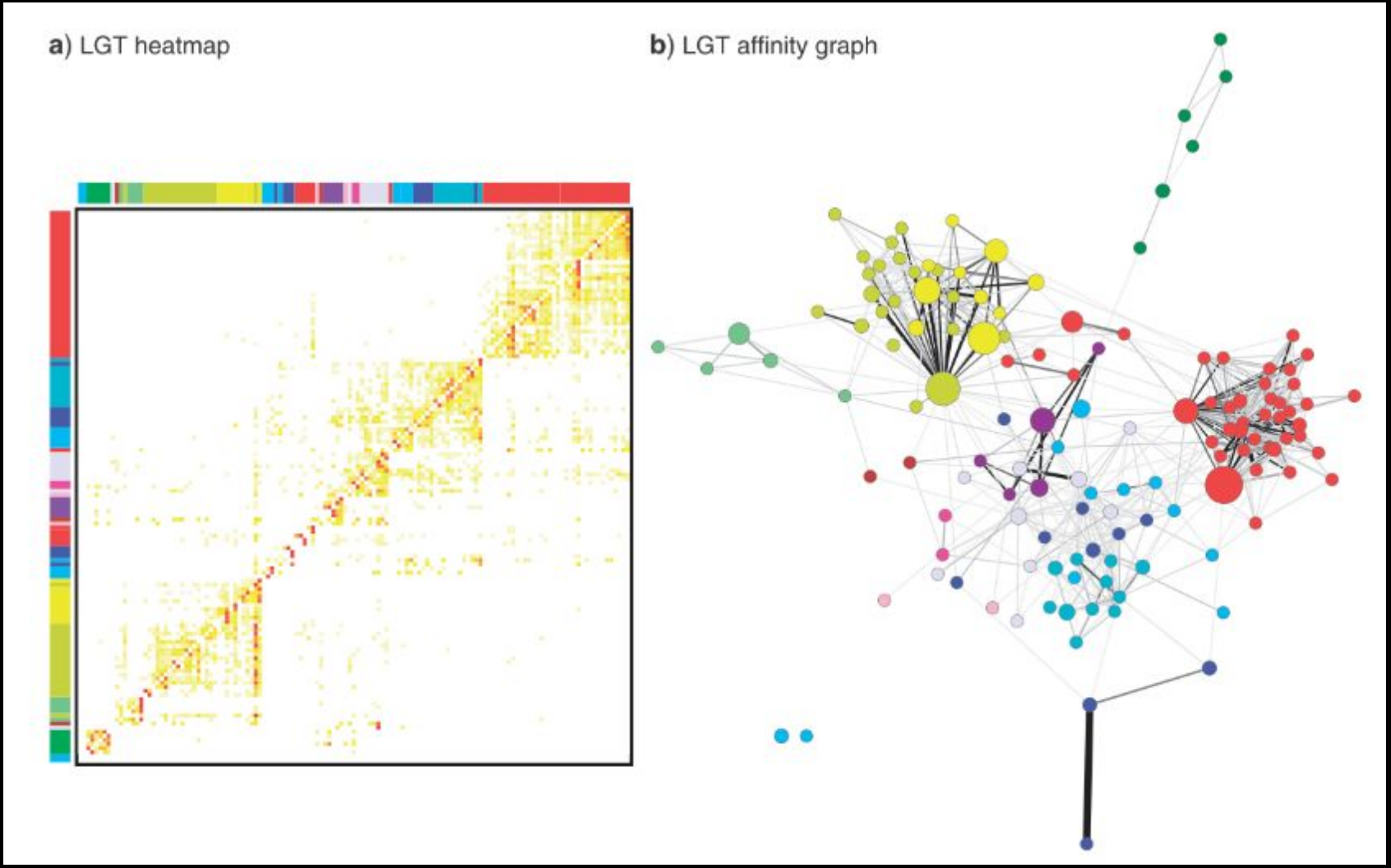


Example

244 microbial taxa
40,631 trees
393,876 leaves (= genes)



Mapping LGT events between lineages



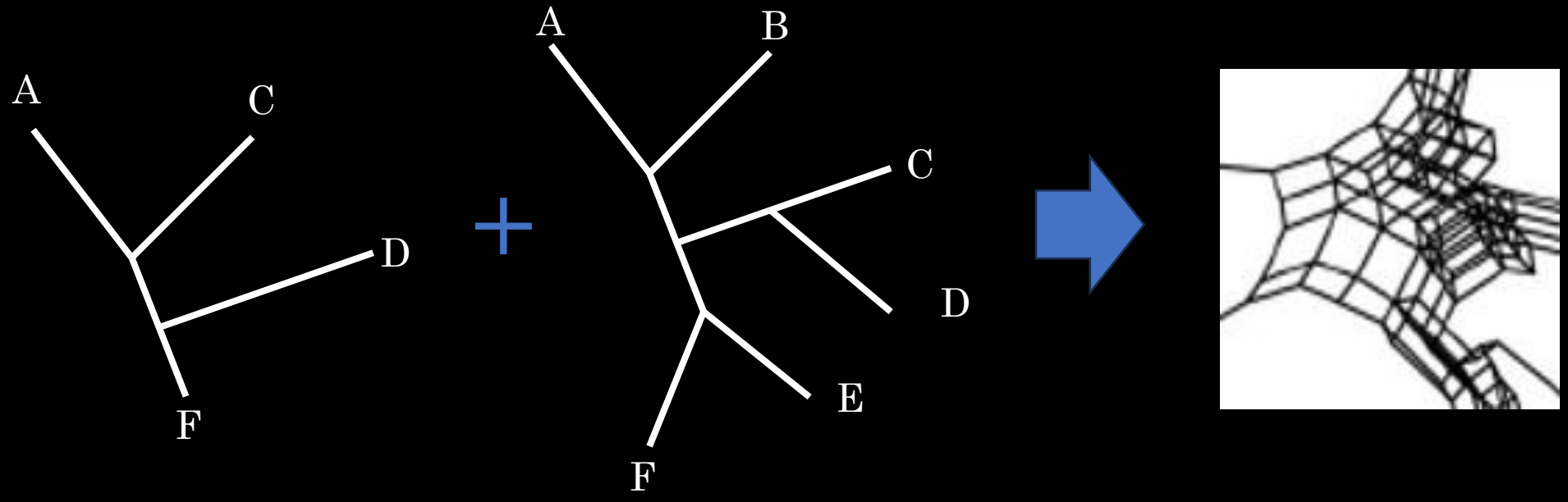
This comparison step took < 1 minute for all 40,000+ trees

Whidden et al. (2014) *Syst Biol*

Networks from Many Trees

Consensus networks and supernetworks

Like consensus tree methods, but do not constrain the results to a tree



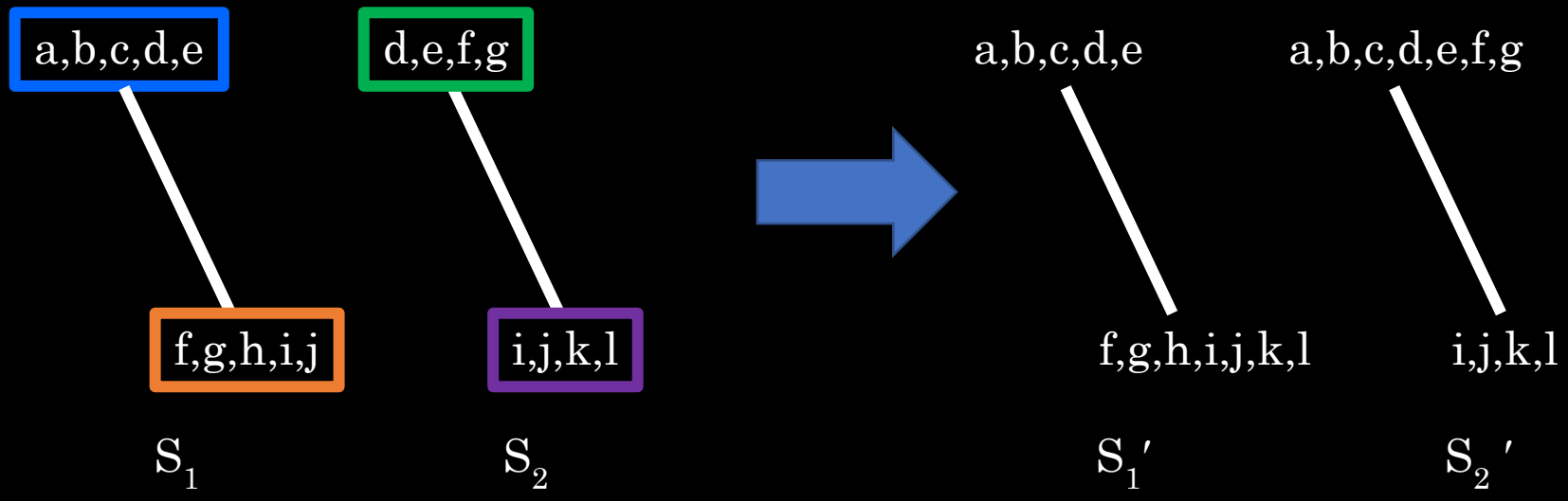
Z-closure supernetworks

- We have a set of trees T that define a set of splits Σ
- These splits are not all necessarily *compatible*, nor do they necessarily all cover the same set of taxa
- How can we reconcile them?

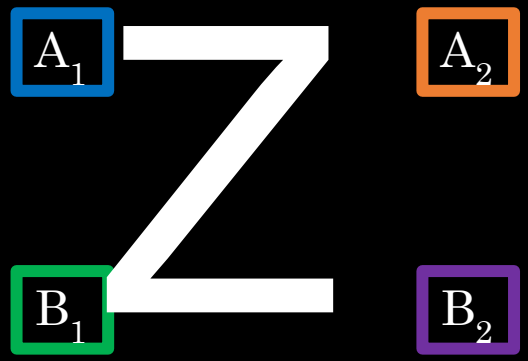
Huson et al. (2004) *IEEE/ACM Trans Comput Biol Bioinform*

The Z-rule

For any two splits $S_1 = \frac{A_1}{B_1} \in \Sigma$ and $S_2 = \frac{A_2}{B_2} \in \Sigma$:
if $A_1 \cap A_2 \neq \emptyset$, $A_2 \cap B_1 \neq \emptyset$, $B_1 \cap B_2 \neq \emptyset$, and $A_1 \cap B_2 = \emptyset$,
then
replace S_1 and S_2 by $S_1' = \frac{A_1}{B_1 \cup B_2}$, and $S_2' = \frac{A_1 \cup A_2}{B_2}$.



Z?



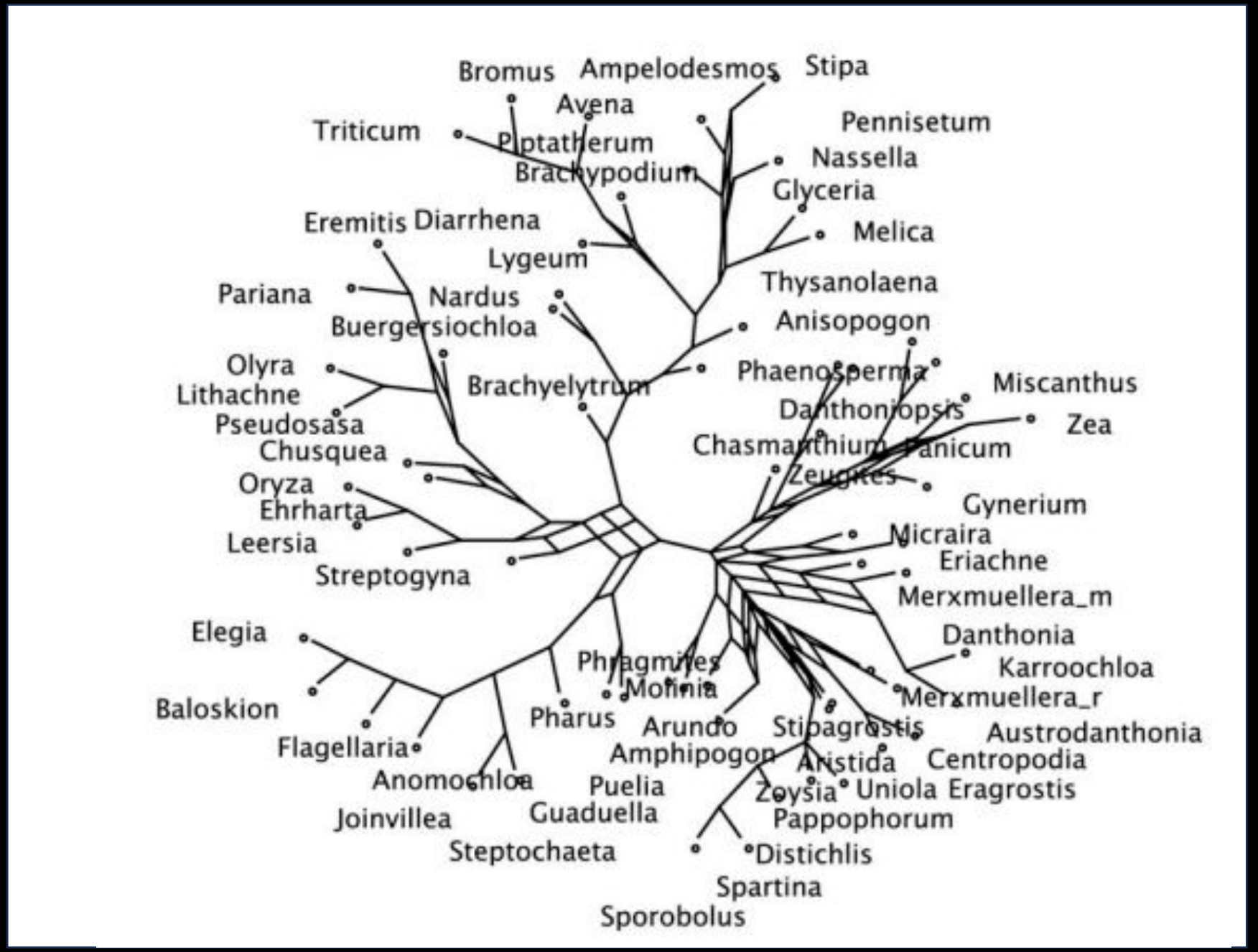
Intersecting sets

Obtaining the Z-closure

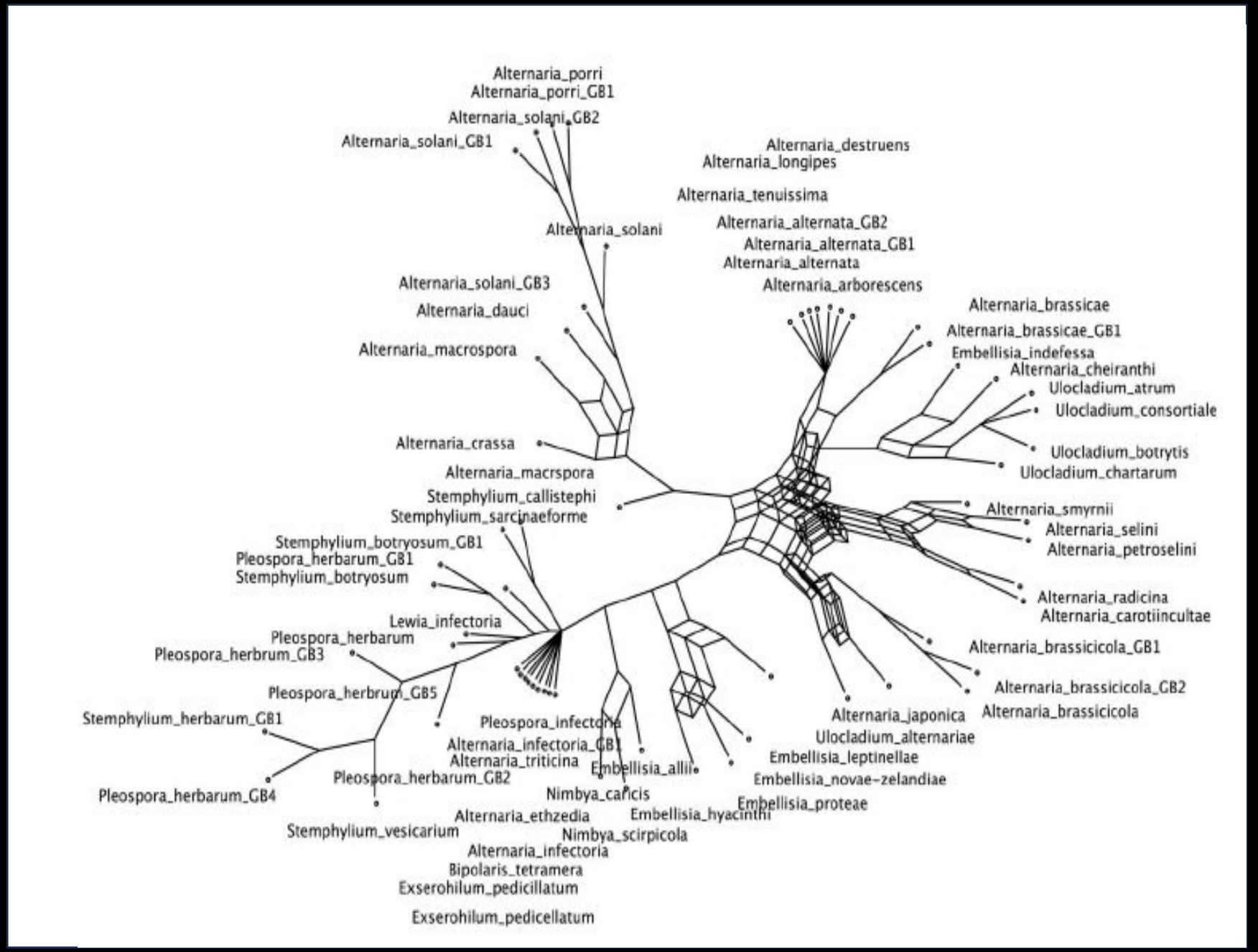
- Build up sets of splits using the Z-rule until all splits are complete

- The weighting of each complete split is proportional to the corrected edge weights in all splits that support it

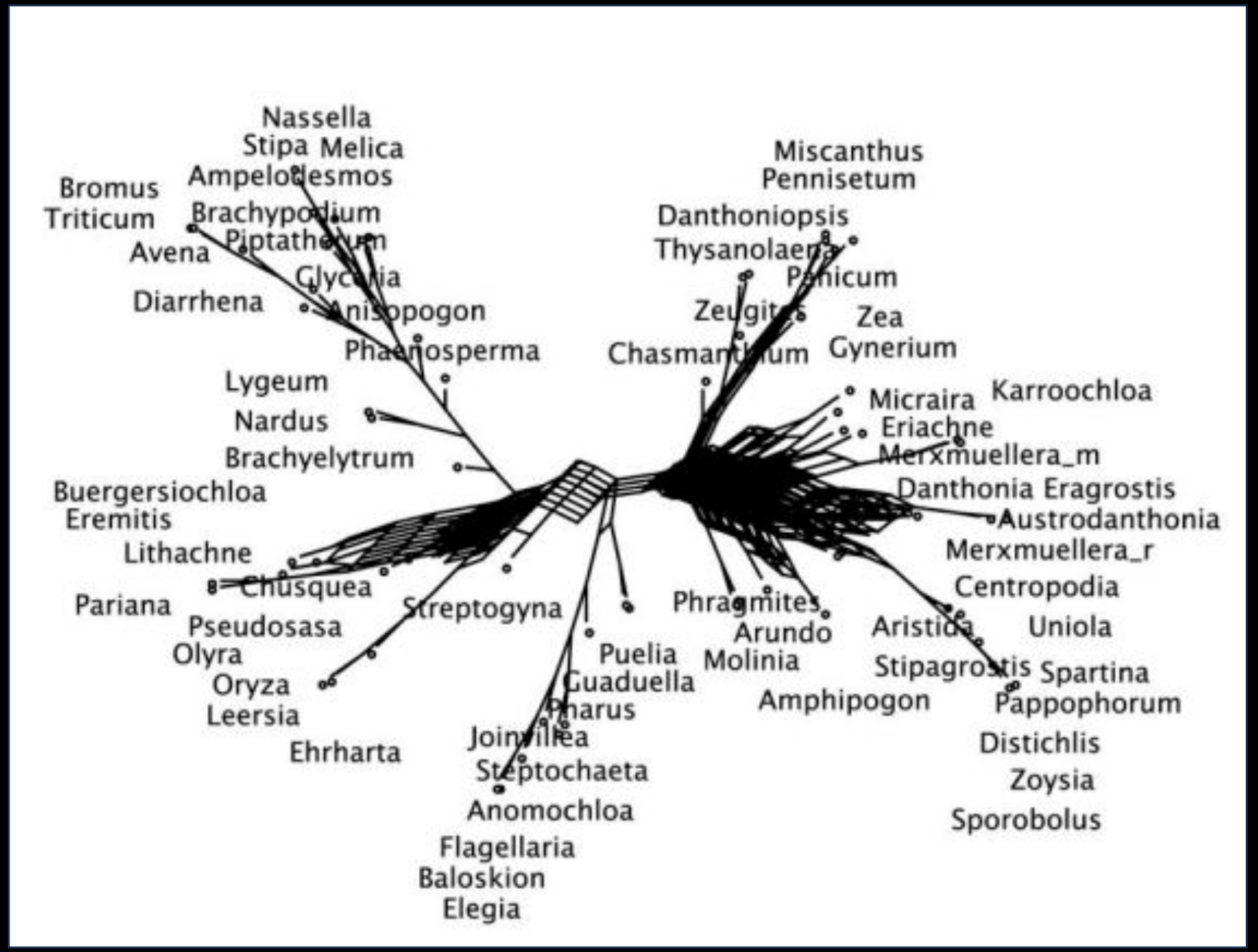
Z-closure supernetwork from five fungal gene trees
(max two dimensions)

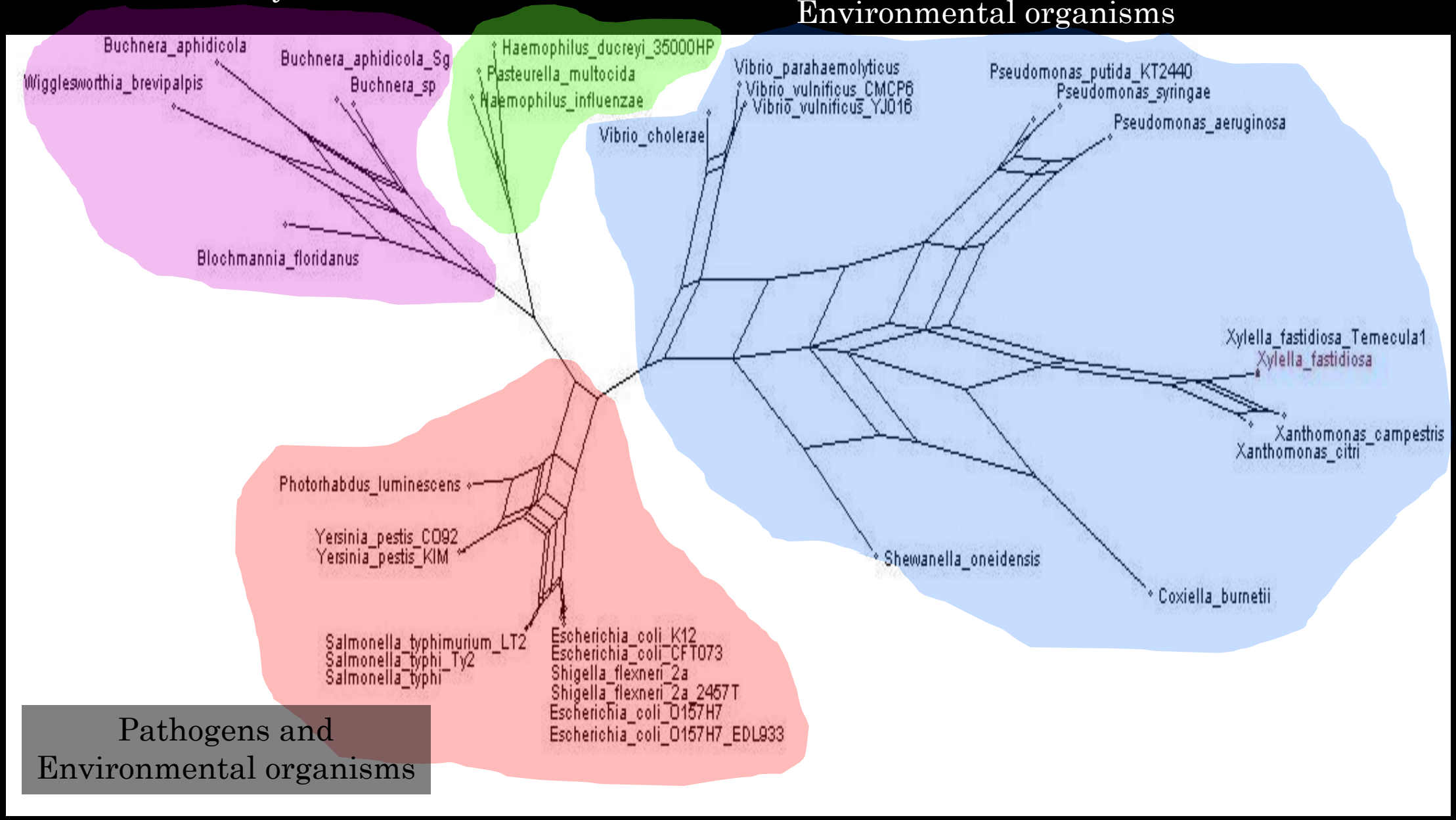


Z-closure supernetwork from five fungal gene trees
(max three dimensions)



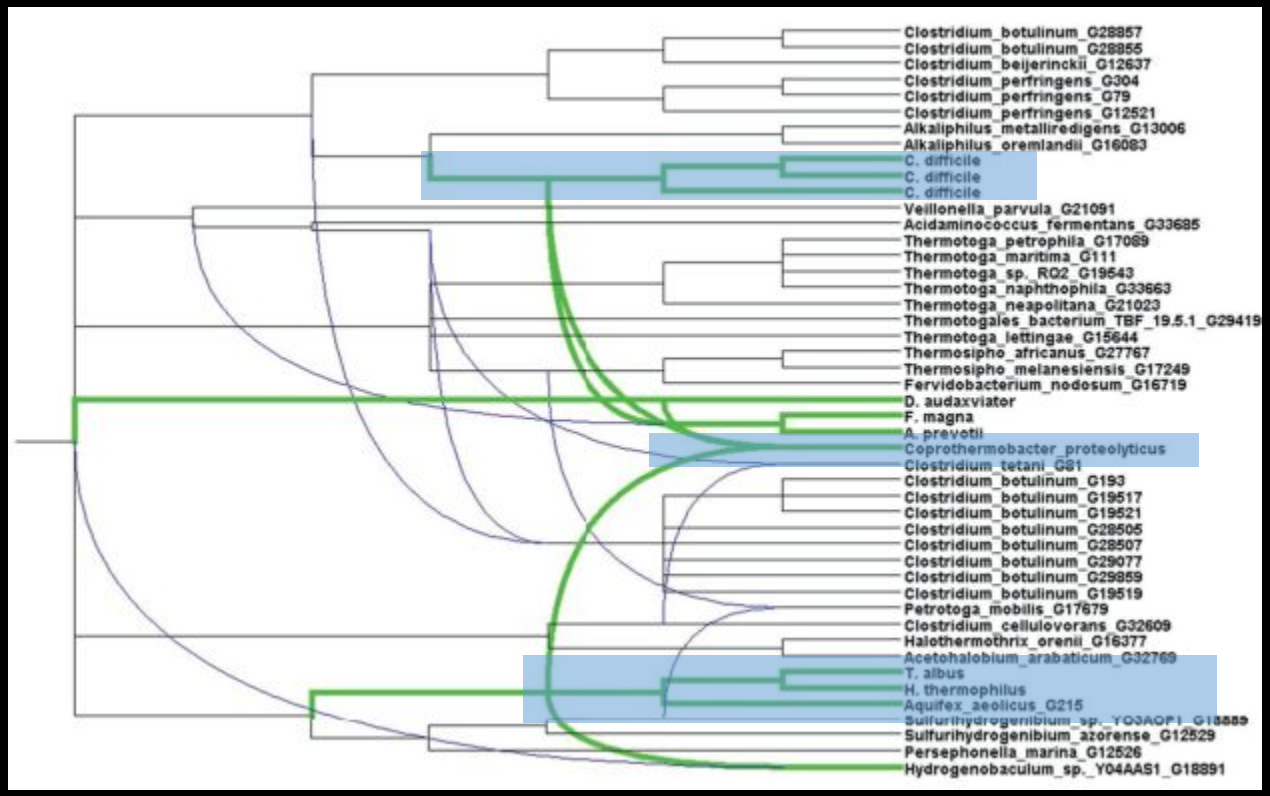
Z-closure supernetwork
from five fungal gene trees
(no constraint)





Z-closure supernetwork from 1721 trees (me, 2005)

Galled Networks (in very brief)



Galled network
Method: Huson et al. (2009) *Bioinformatics*
This network: Me (2011) *Biol Direct*

Summary

- There are several ways to represent incompatible signals within a single alignment
  - Are these signals clustered together?

- There are also several ways to generate a single tree or network topology from many data sets
  - How constrained should the answer be?