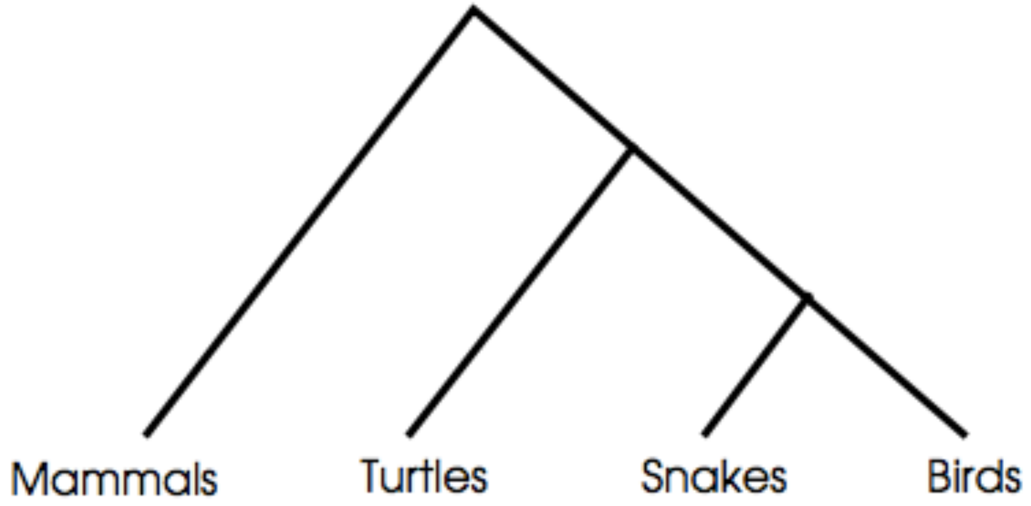
Topic 10: Phylogenomics with SNPs

Biol 525D - Bioinformatics for Evolutionary Biology
2018



Overview

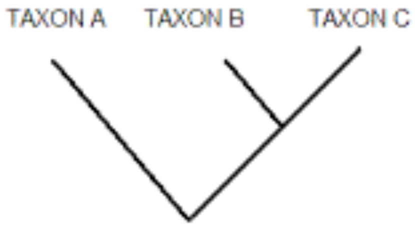
- What is phylogenetic
- Terms and outline
- Methods of trees
    - UPGMA
    - Neighbour joining
    - Maximum parsimony
    - Maximum likelihood
- Distance calculations
- Considerations for SNPs
- Reticulate networks

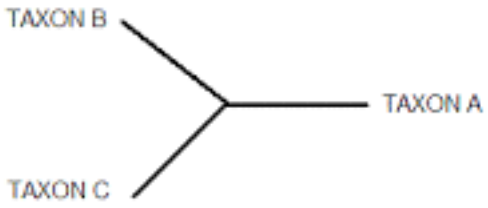
Learning Goals

- What are the different methods of building a phylogenetic tree?
- What are the different methods of calculating phylogenetic distance?
- How can confidence be measured in phylogenetic trees?
- How can you use SNPs for phylogenetic analysis?

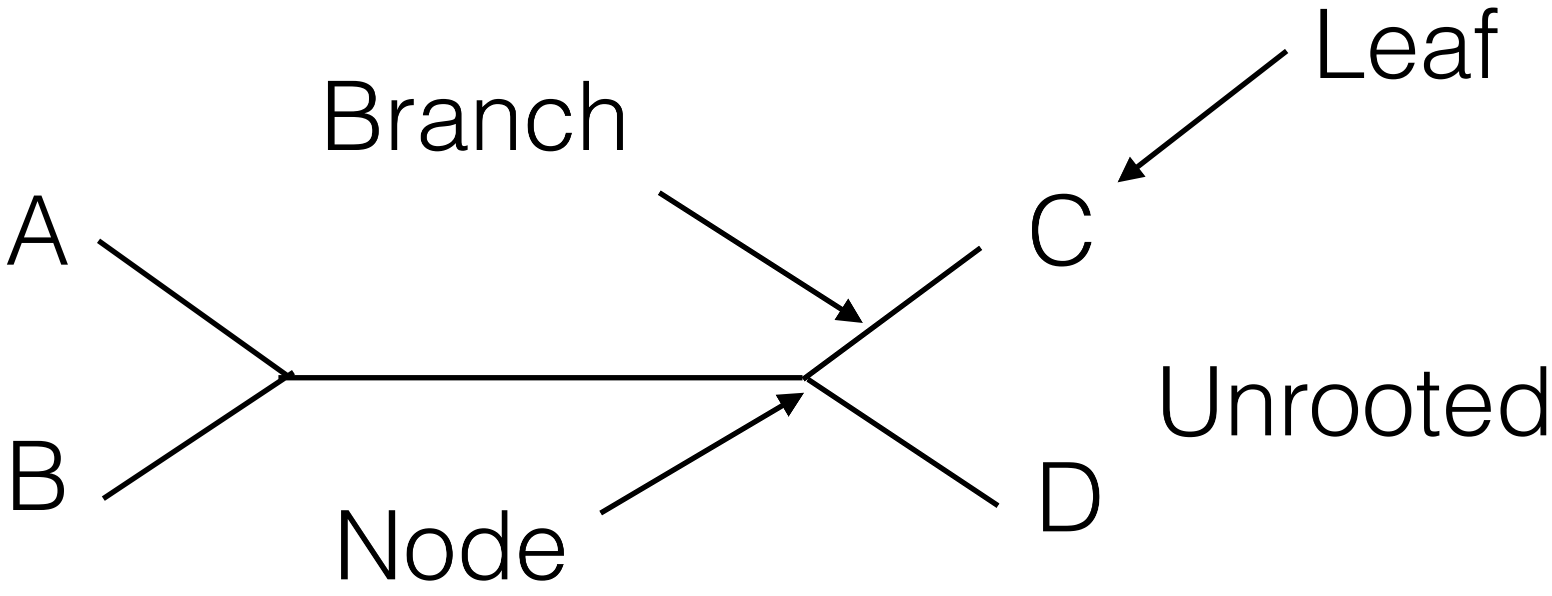
Phylogenetics

- Reconstructs evolutionary ties between organisms.
- Estimates divergence times between organisms.
- Can use morphological or genetic data.





Terms


Leaf

Branch

A

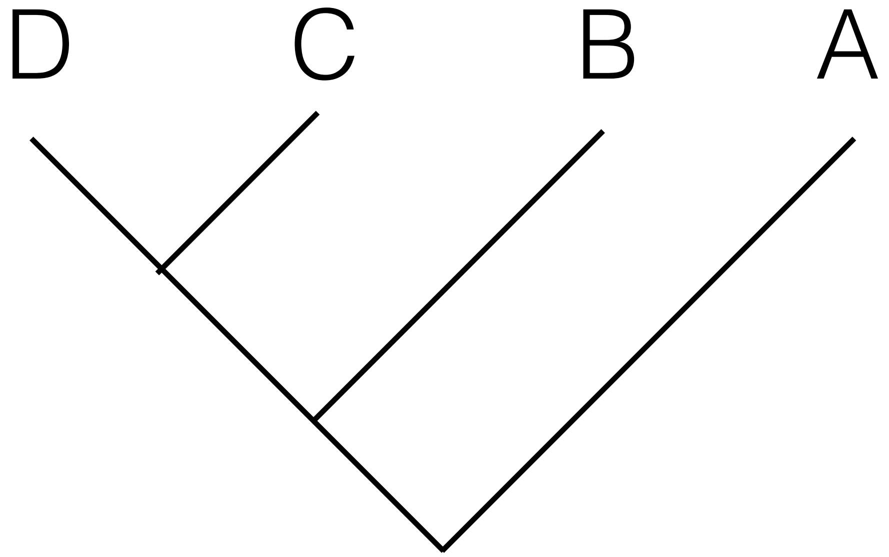
C

B

D

Node

Unrooted tree


D

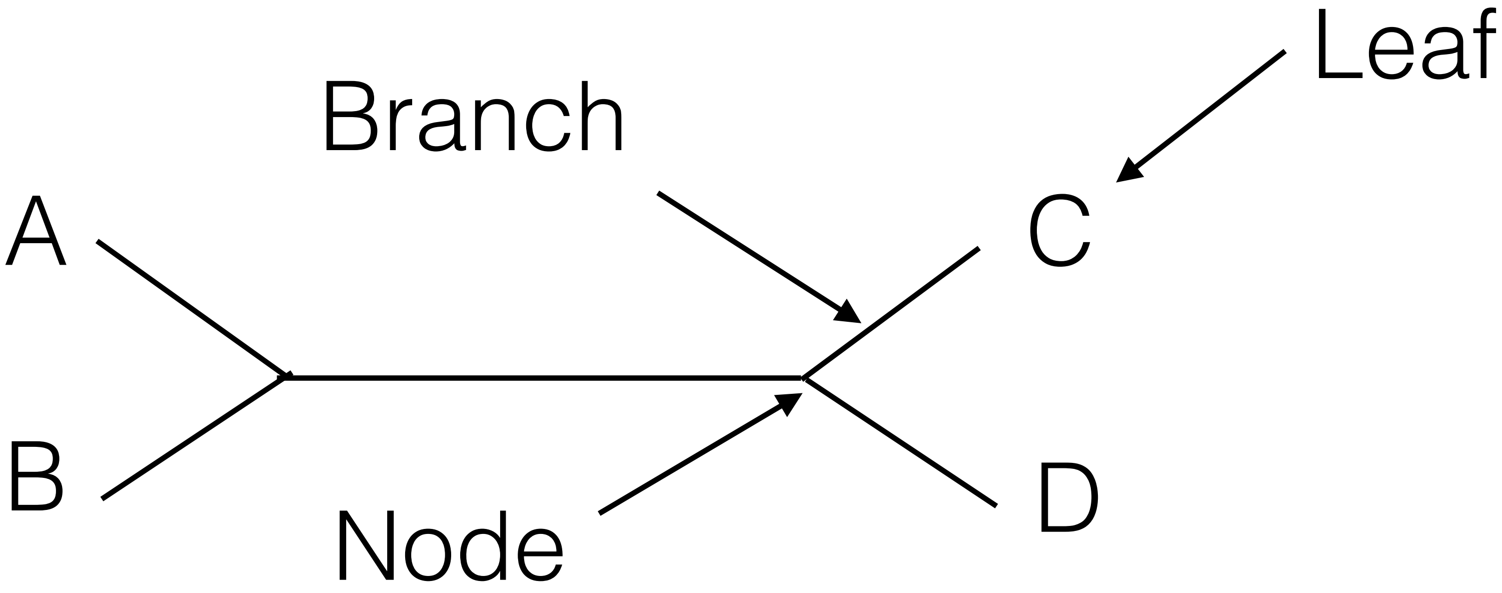
C

B

A

Rooted tree

Terms



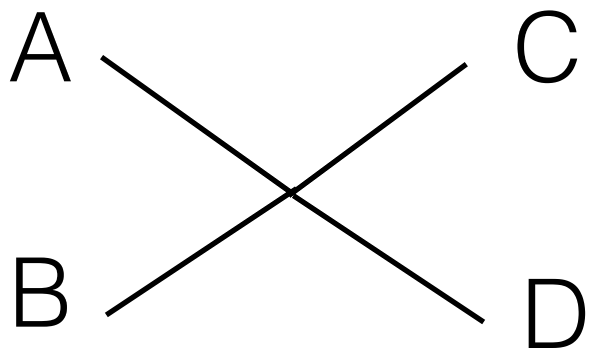
Operational Taxonomic Unit (OTU): An external node representing a monophyletic group

Distance methods

- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects

Distance methods

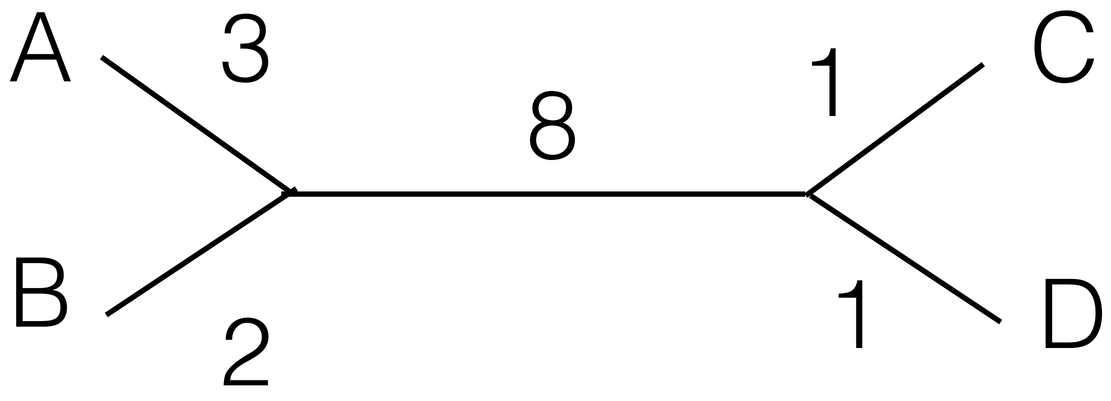
- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects



	A	B	C	D
A	0			
B	5	0		
C	12	11	0	
D	12	11	2	0

Distance methods

- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects



	A	B	C	D
A	0			
B	5	0		
C	12	11	0	
D	12	11	2	0

Distance methods

- Tries to build a tree where the distances measured between leaves on the tree correspond to the actual distance between objects

- Easy to calculate when distance matrix is additive, but often is not.

- Need to use heuristics to pick best fitting tree because there are too many to try all.

Neighbour-joining

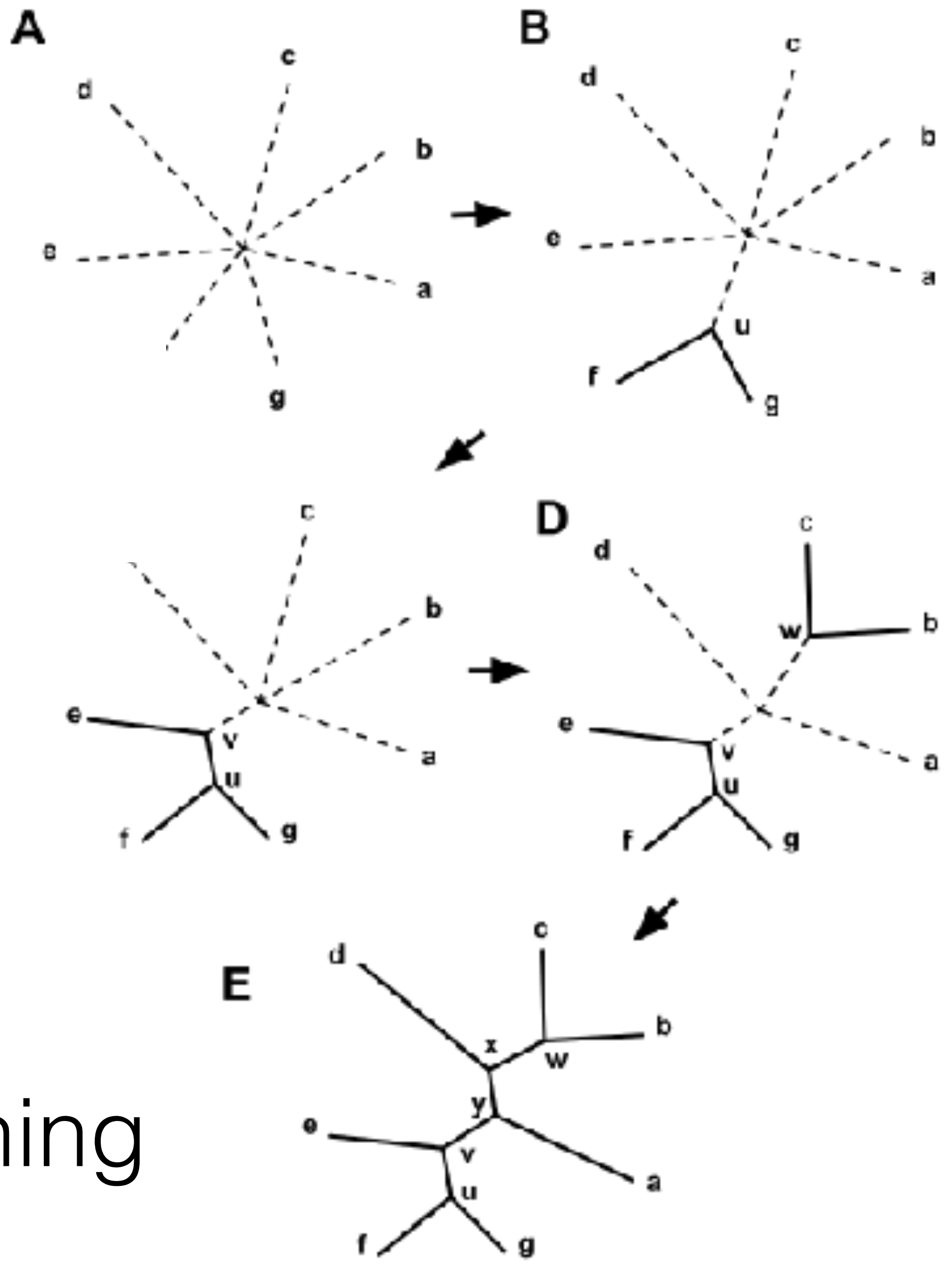
- Similar to UPGMA, but uses a Q-matrix.

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

Distance i to j

Distance i to everything

Distance j to everything



Maximum Likelihood

- Calculates the likelihood of trees based on substitution models and picks the model with the highest likelihood.
- Uses heuristics to search through the possible tree space.

Bayesian Trees

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

- Finds the tree with the highest posterior probability based on a model of evolution and prior probabilities.

- Can use MCMC to search the tree space.

Substitution models

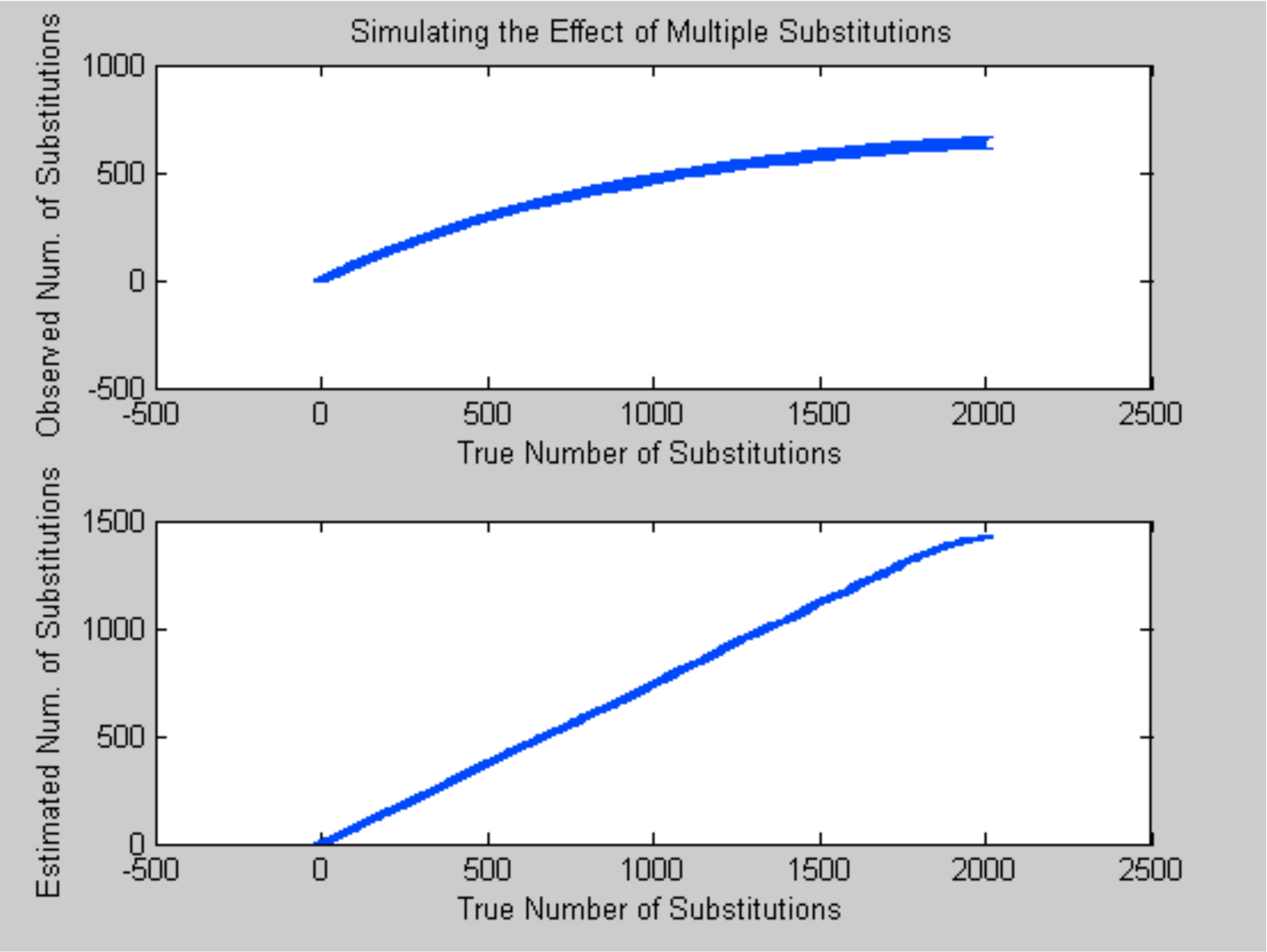
- Many different models of varying complexity.
  - Equal or unequal mutation rates
  - Equal or unequal base frequencies
  - GC bias or not
- More parameters not always better, can overfit to your data, so you should use a program to pick the best model.

Substitution models

Hidden Changes

Seq1 A -> T -> C
Seq2 A -> C

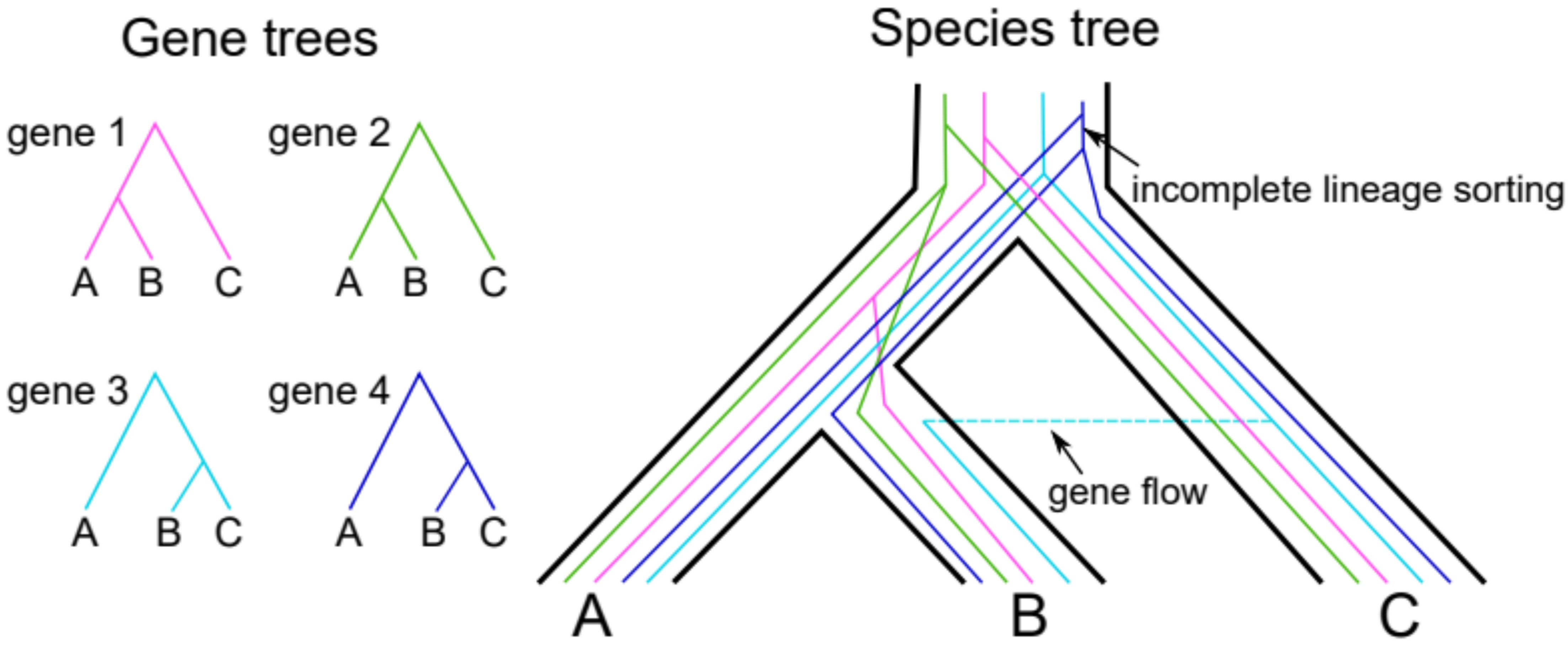
- Also helps control for saturation of mutations and hidden changes



Bootstrapping

- Repeat your analysis with a bootstrapped version of your dataset X1000 times.
    - Bootstrapping involves building a dataset of the same number of characters by sampling with replacement from your original dataset.
- The percent of bootstrap datasets that produce the same tree is your confidence value.

Gene trees



Marin et al., 2018

Gene tree assumptions

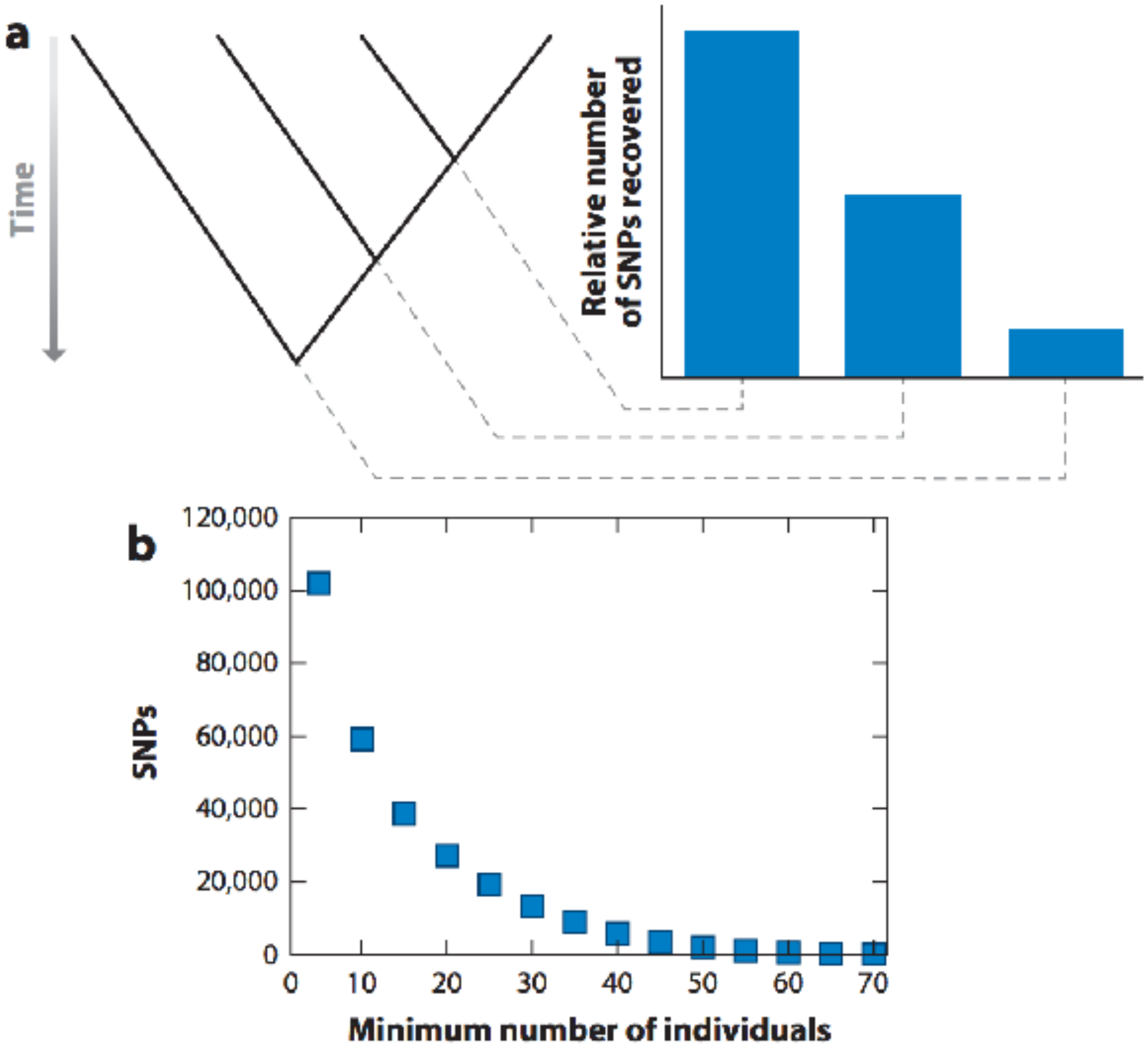
- Haplotypes are phased within diploid individuals
- No recombination within a gene
- No linkage between genes

Considerations for SNPs

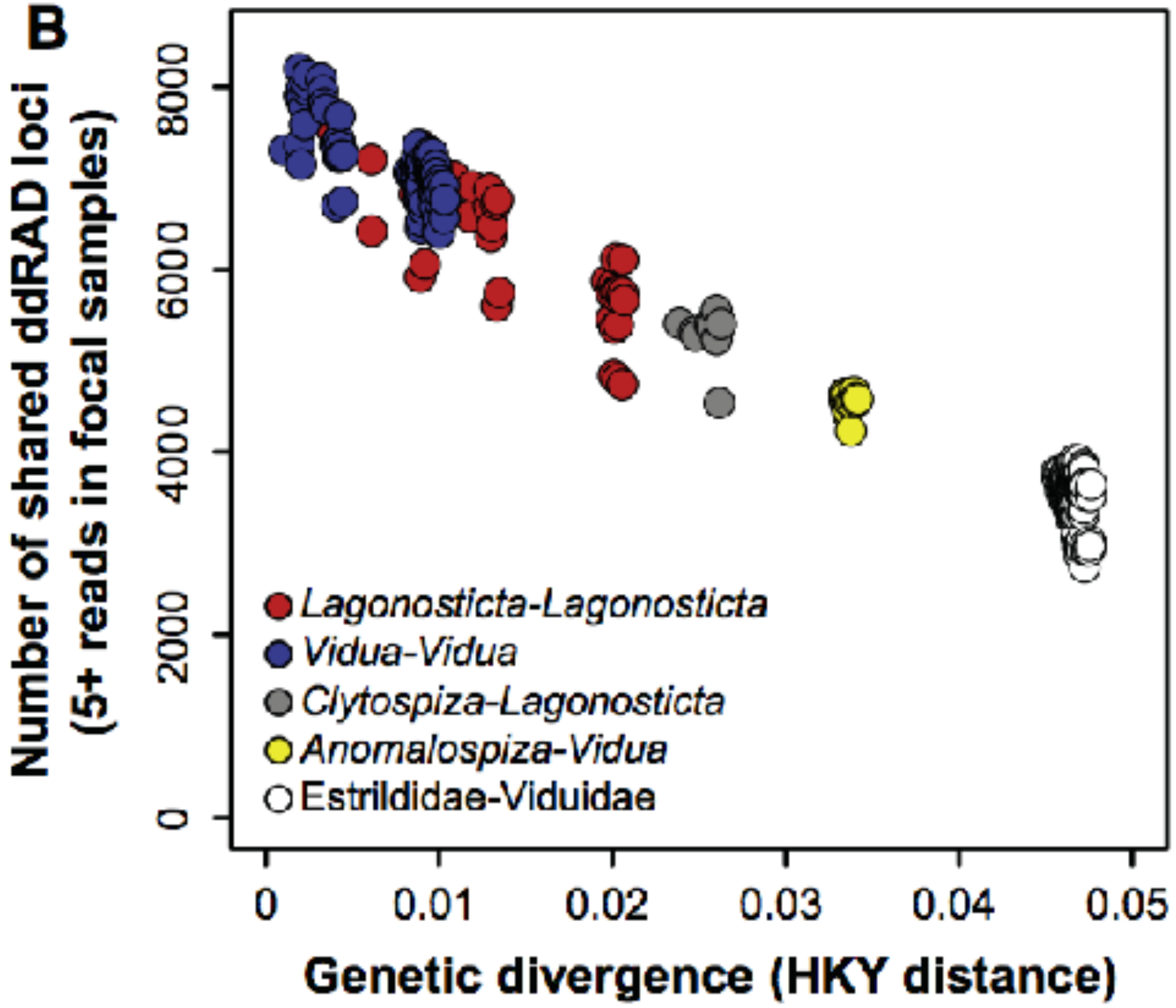
- Generally you only keep variable sites, while many phylogenetic algorithms assume invariant sites are included.

- Need to use models that explicitly control for the ascertainment bias, or include invariant sites.

Allelic dropout



Allelic dropout



DaCosta & Sorenson 2016

Allelic dropout

- Overestimates genetic variation
- Causes drop out of high-frequency alleles.

Gautier et al. 2013

Ascertainment Bias

- Problem when using probe based SNP detection
- Biases SNP set to intermediate frequency alleles.

Gautier et al. 2013

Ways to use SNPs

- Concatenation
- SNAPP (+PoMo)
- Quartet methods

Ways to use SNPs

- Concatenation
- SNAPP (+PoMo)
- Quartet methods

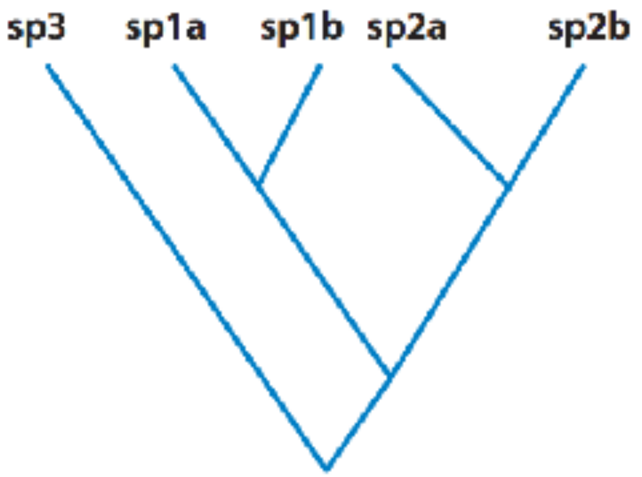
Concatenation

a SNP data matrix

	locus1	locus2	locus3	locus4	locus5
species1_a	A	T	C	G	A
species1_b	A	T	G	G	A
species2_a	T	T	G	G	G
species2_b	A	C	C	T	G
species3	A	C	C	T	G

Super locus

species1_a	ATCGA...
species1_b	ATGGA...
species2_a	TTGGG...
species2_b	ACCTG...
species3	ACCTG...



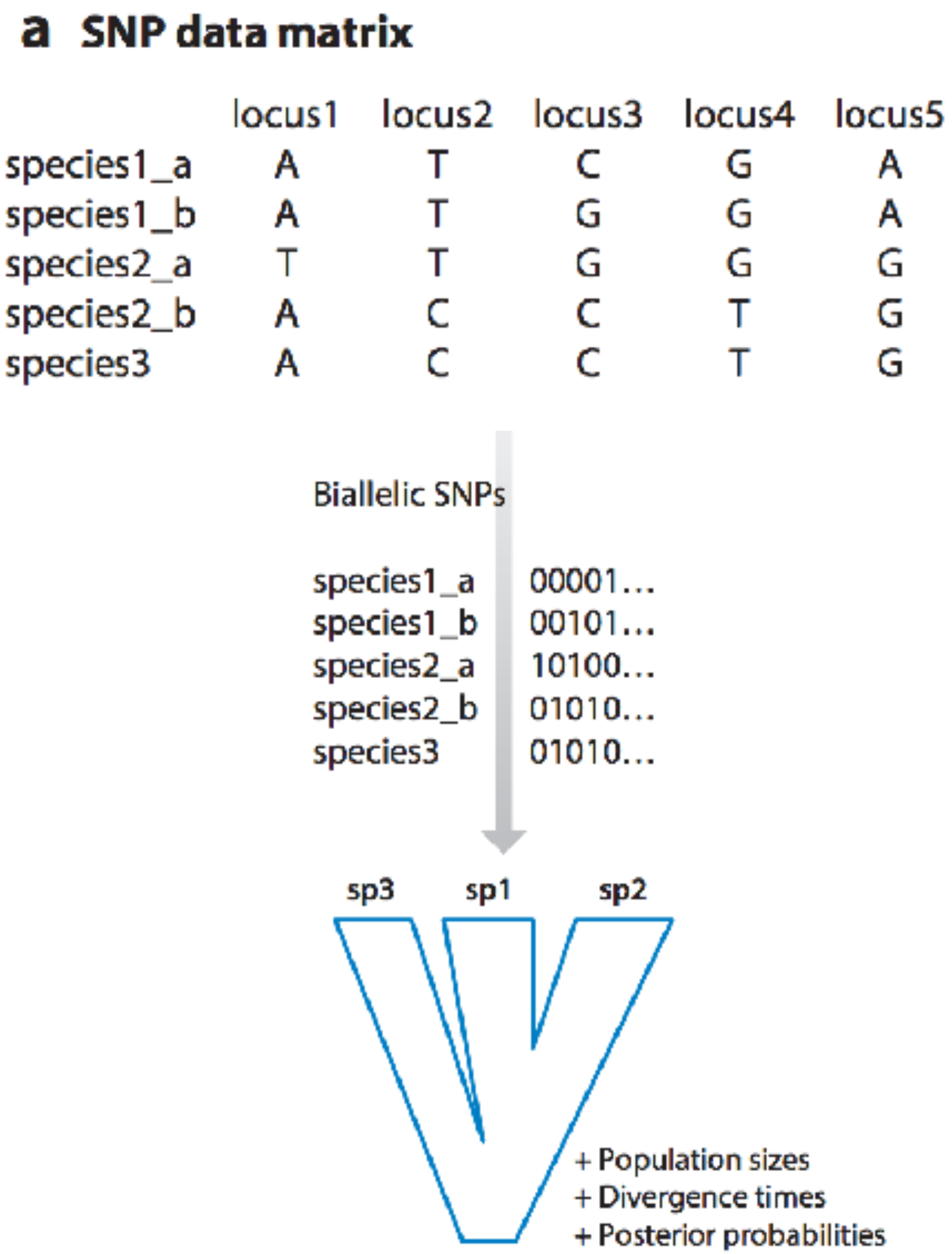
Concatenation

- Phylogeny estimated using other methods
- Problem: Ignores incomplete lineage sorting and assumes a single coalescent history.
- Problem: Overestimates support and can bias toward incorrect trees
- Solution: Include invariant sites from original alignments

Ways to use SNPs

- Concatenation
- SNAPP (+PoMo)
- Quartet methods

SNAPP (+PoMo)



a SNP data matrix

	locus1	locus2	locus3	locus4	locus5
species1_a	A	T	C	G	A
species1_b	A	T	G	G	A
species2_a	T	T	G	G	G
species2_b	A	C	C	T	G
species3	A	C	C	T	G

Biallelic SNPs

species1_a	00001...
species1_b	00101...
species2_a	10100...
species2_b	01010...
species3	01010...

+ Population sizes
+ Divergence times
+ Posterior probabilities

SNAPP (+PoMo)

- Uses allele frequencies to model demographic and phylogenetic history that matches data.
- Doesn’t use gene trees, so its faster than gene tree methods.
- Also estimates divergence times and population sizes.

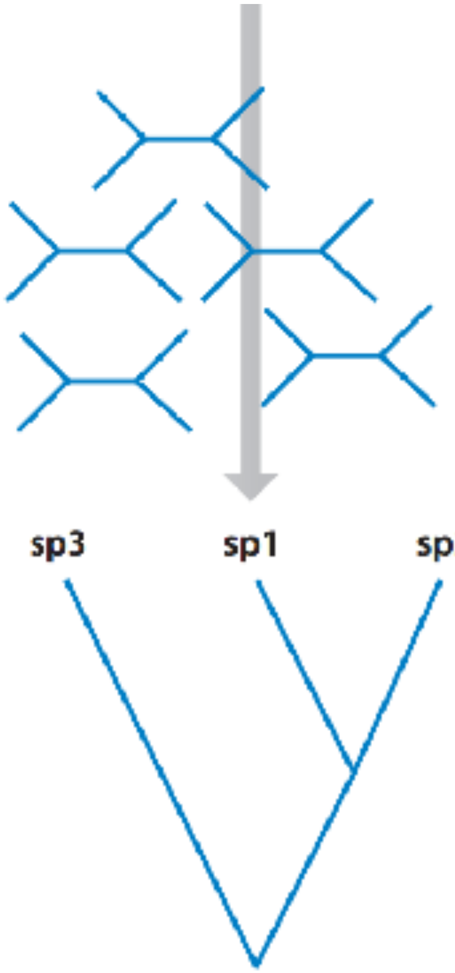
Ways to use SNPs

- Concatenation
- SNAPP (+PoMo)
- Quartet methods

Quartet methods

a SNP data matrix

	locus1	locus2	locus3	locus4	locus5
species1_a	A	T	C	G	A
species1_b	A	T	G	G	A
species2_a	T	T	G	G	G
species2_b	A	C	C	T	G
species3	A	C	C	T	G



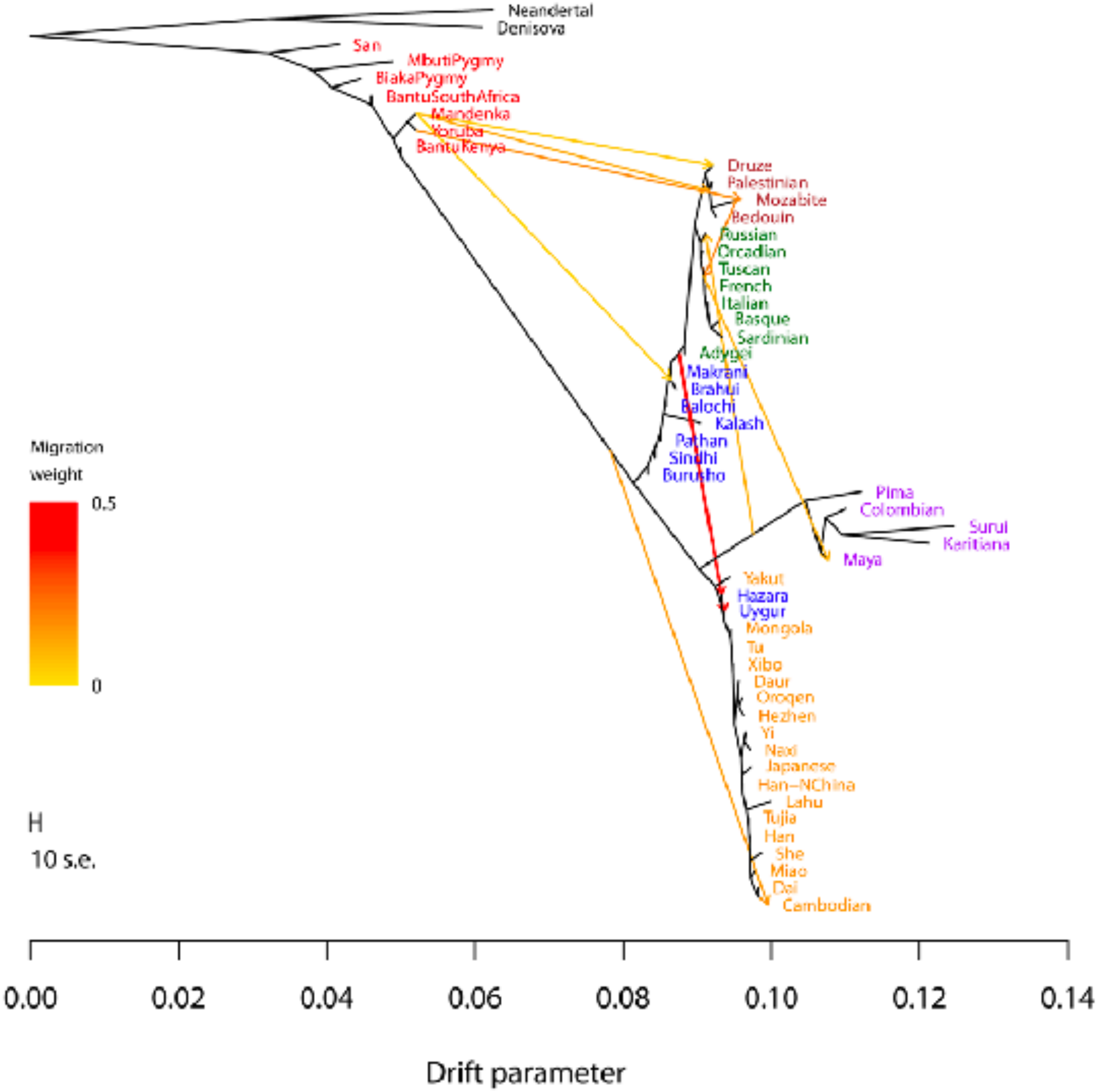
Quartet methods

- Estimates quartet trees and then combines them into a larger species tree.
- Doesn’t need gene trees

Reticulate networks

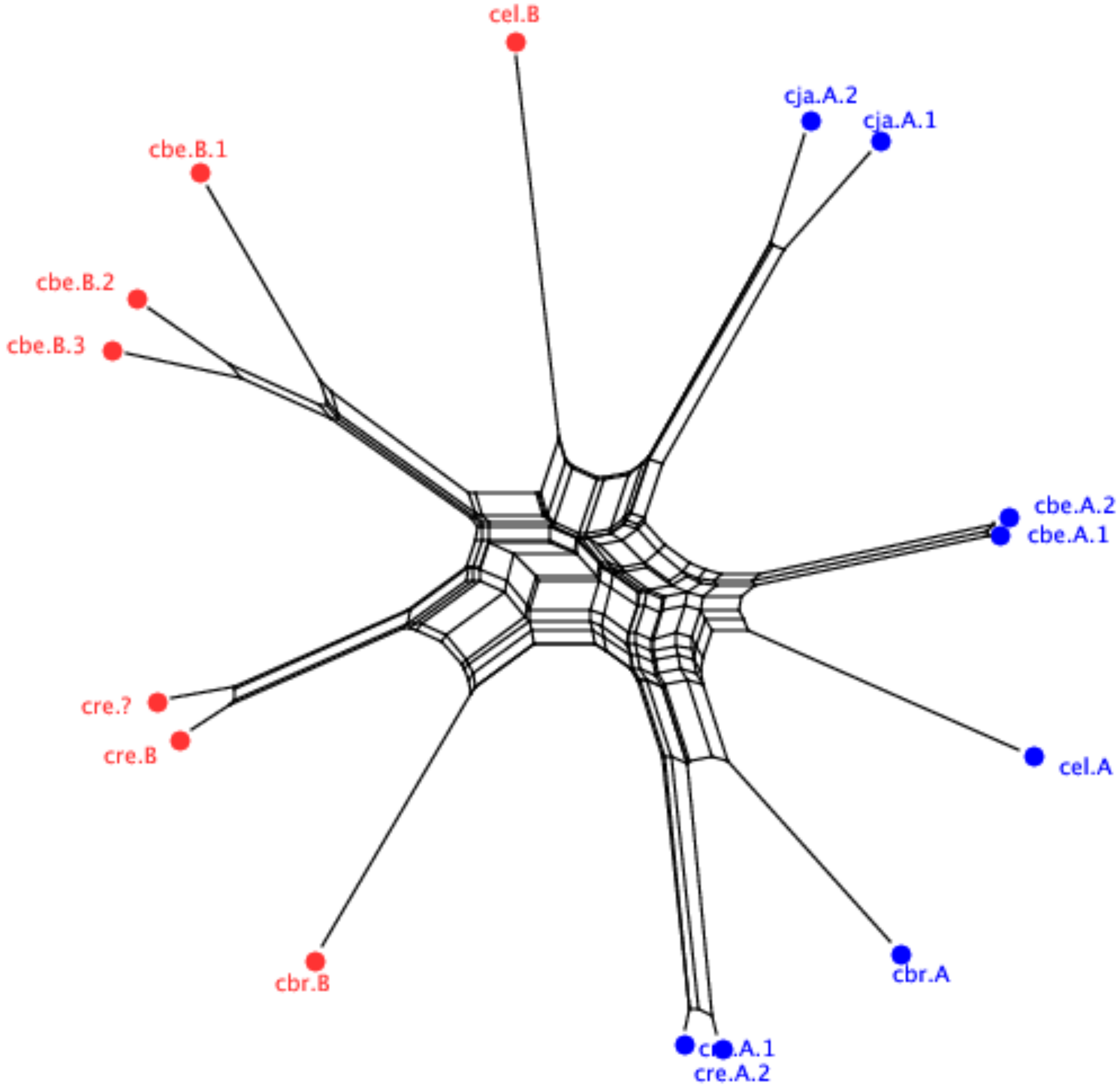
- Phylogenetic history is not always perfectly bifurcating. Gene flow can occur between species and different loci can have different phylogenetic histories.

Reticulate networks



Pickrell & Pritchard 2012

Splitstree


0.01
cel.B
cja.A.2
cja.A.1
cbe.B.1
cbe.B.2
cbe.B.3
cbe.A.2
cbe.A.1
cel.A
cre.?
cre.B
cbr.B
cbr.A
cre.A.1
cre.A.2