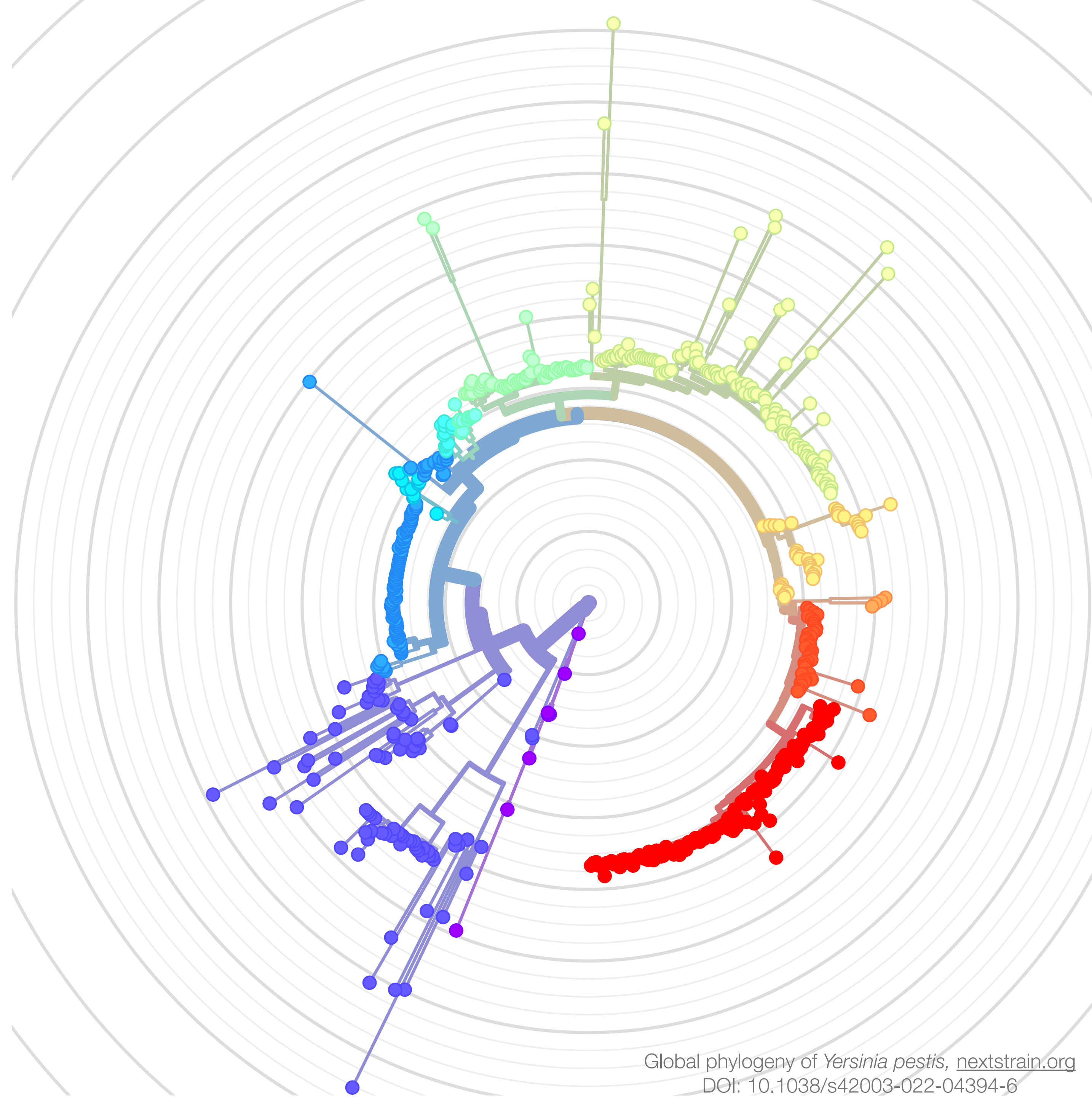


# Introduction to Phylogenetics

BIO298 Microbial bioinformatics block course  
Fanny Wegner  
2023-03-29

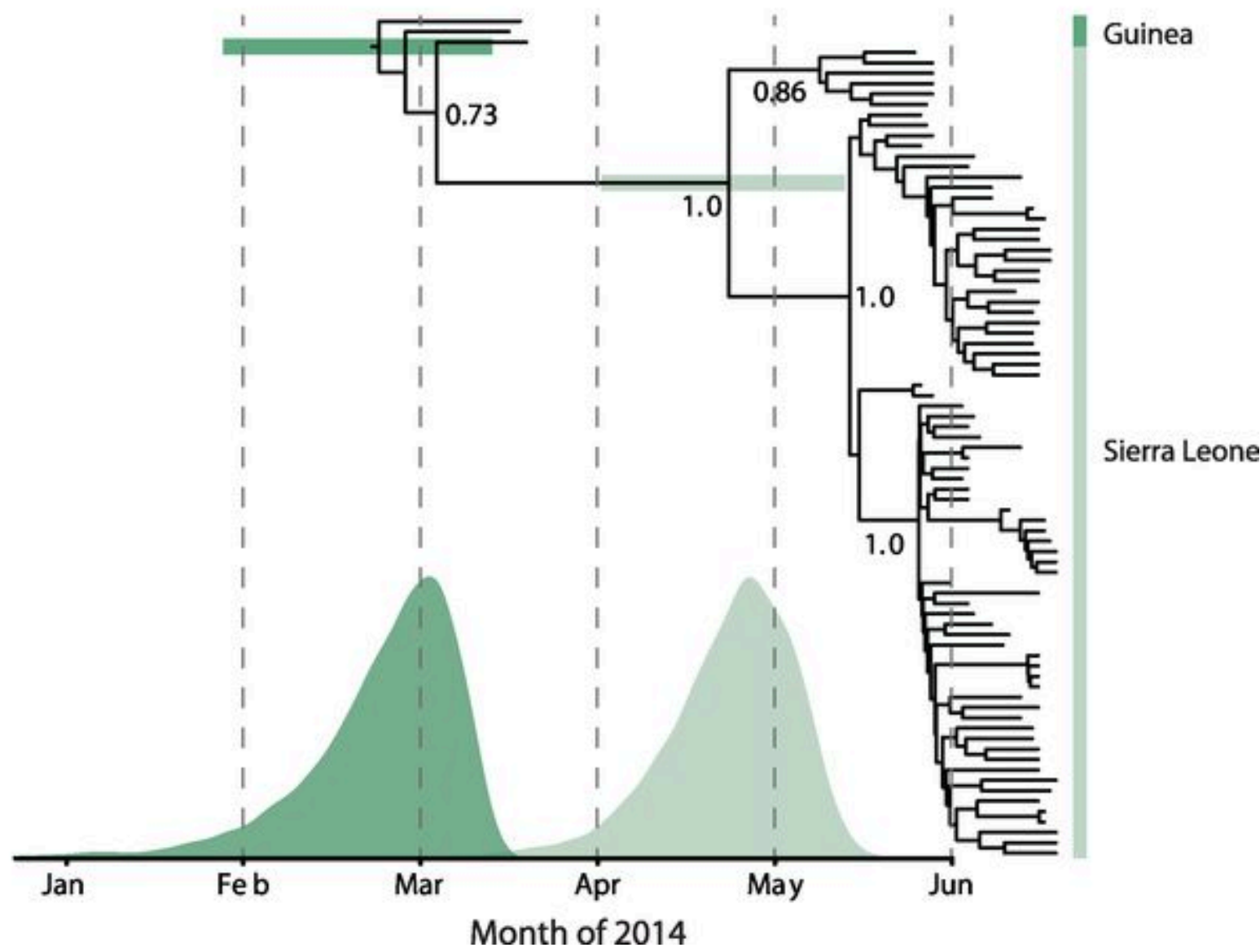




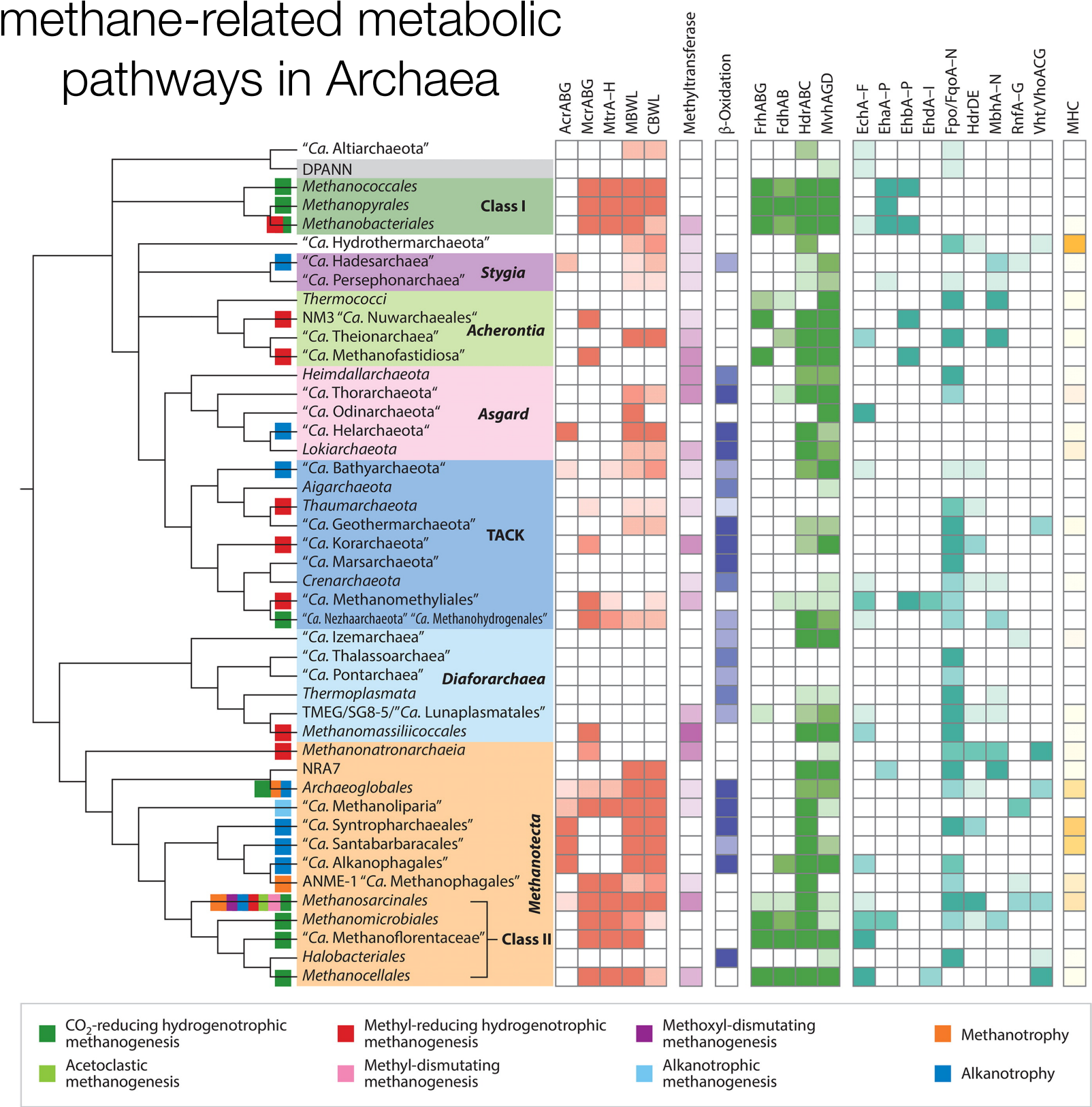
# Why phylogenetics?

- A graphical representation to visualise the evolutionary relationships between genes or organisms

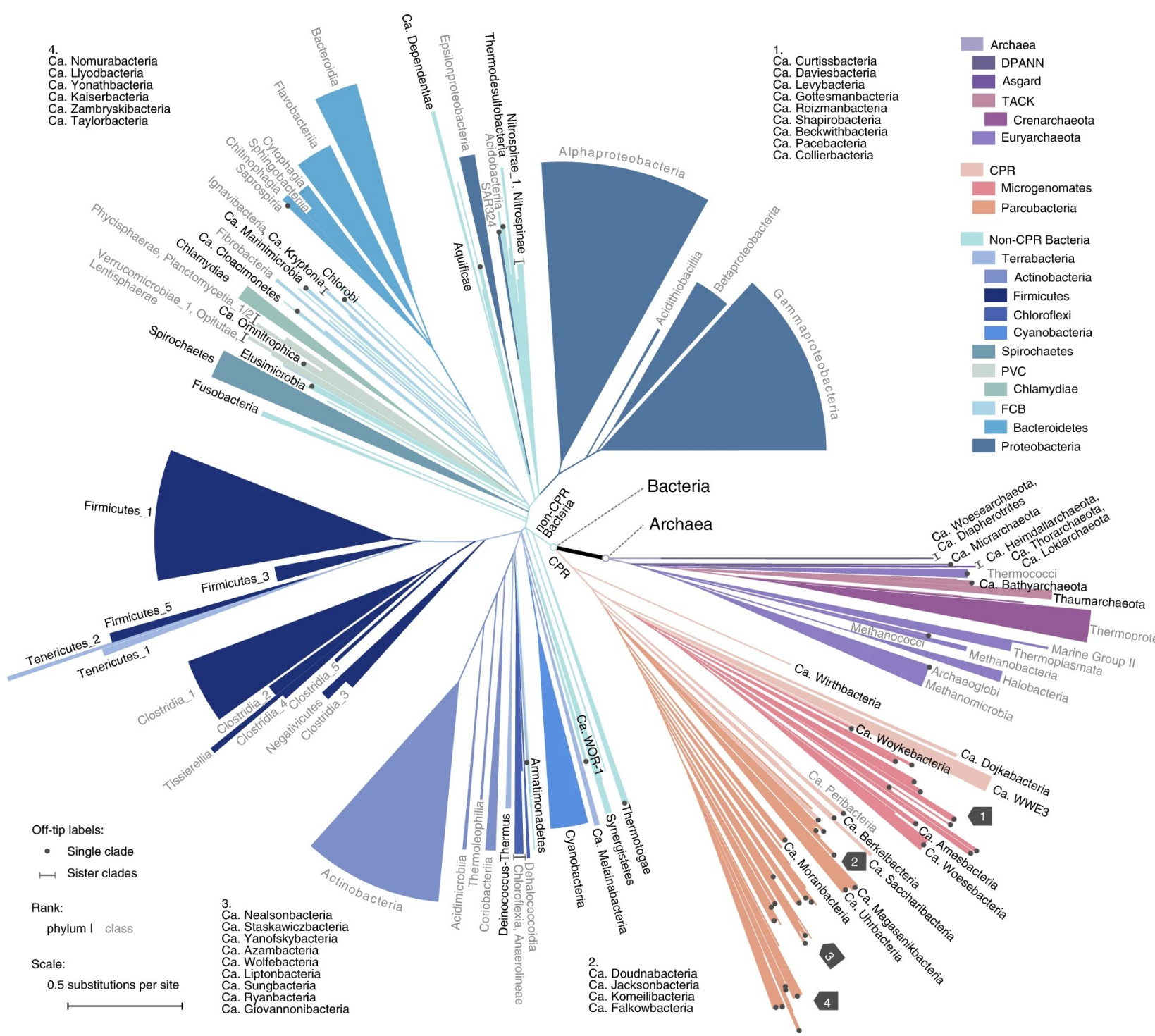
2014 West African Ebola outbreak



Taxonomic distribution of methane-related metabolic pathways in Archaea



Bacterial and archaeal tree of life



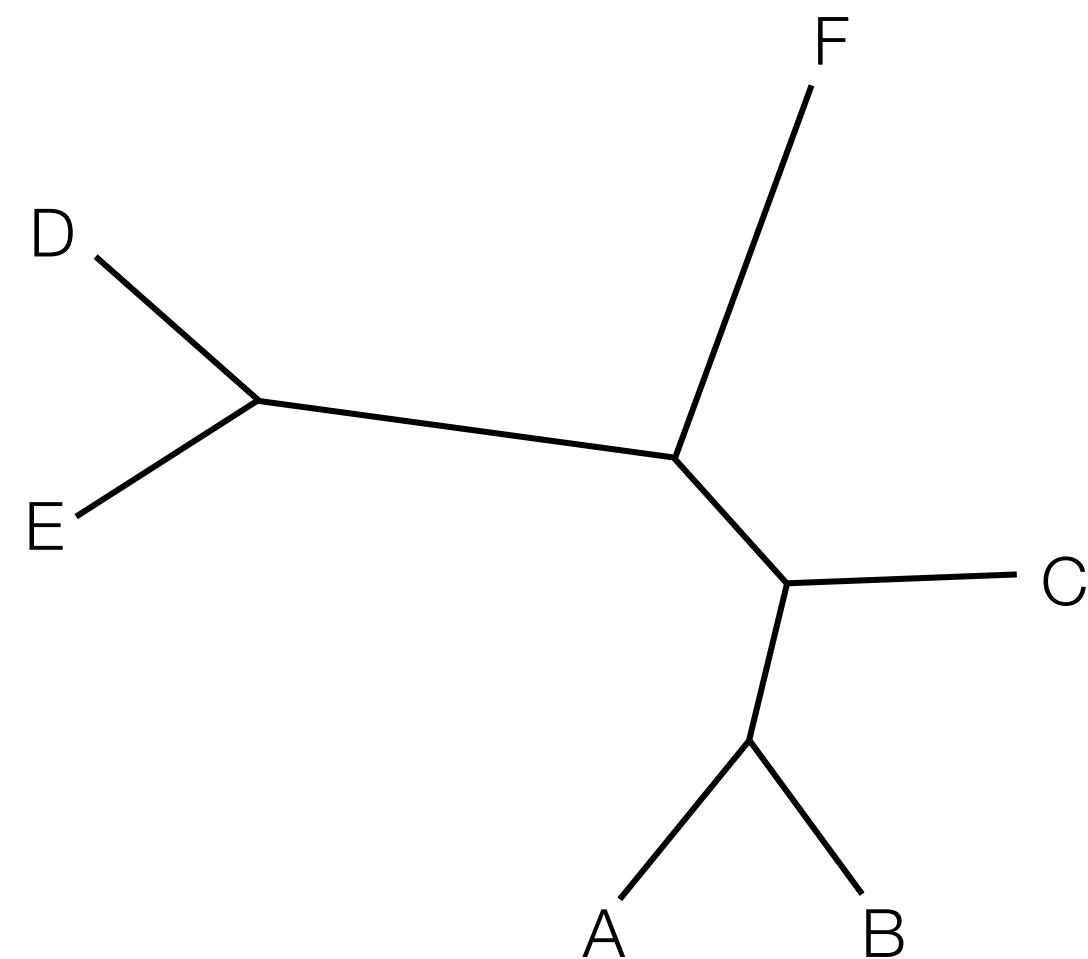
# Outline

---

- Basics of phylogeny, definitions
- Alignments
- Tree construction methods

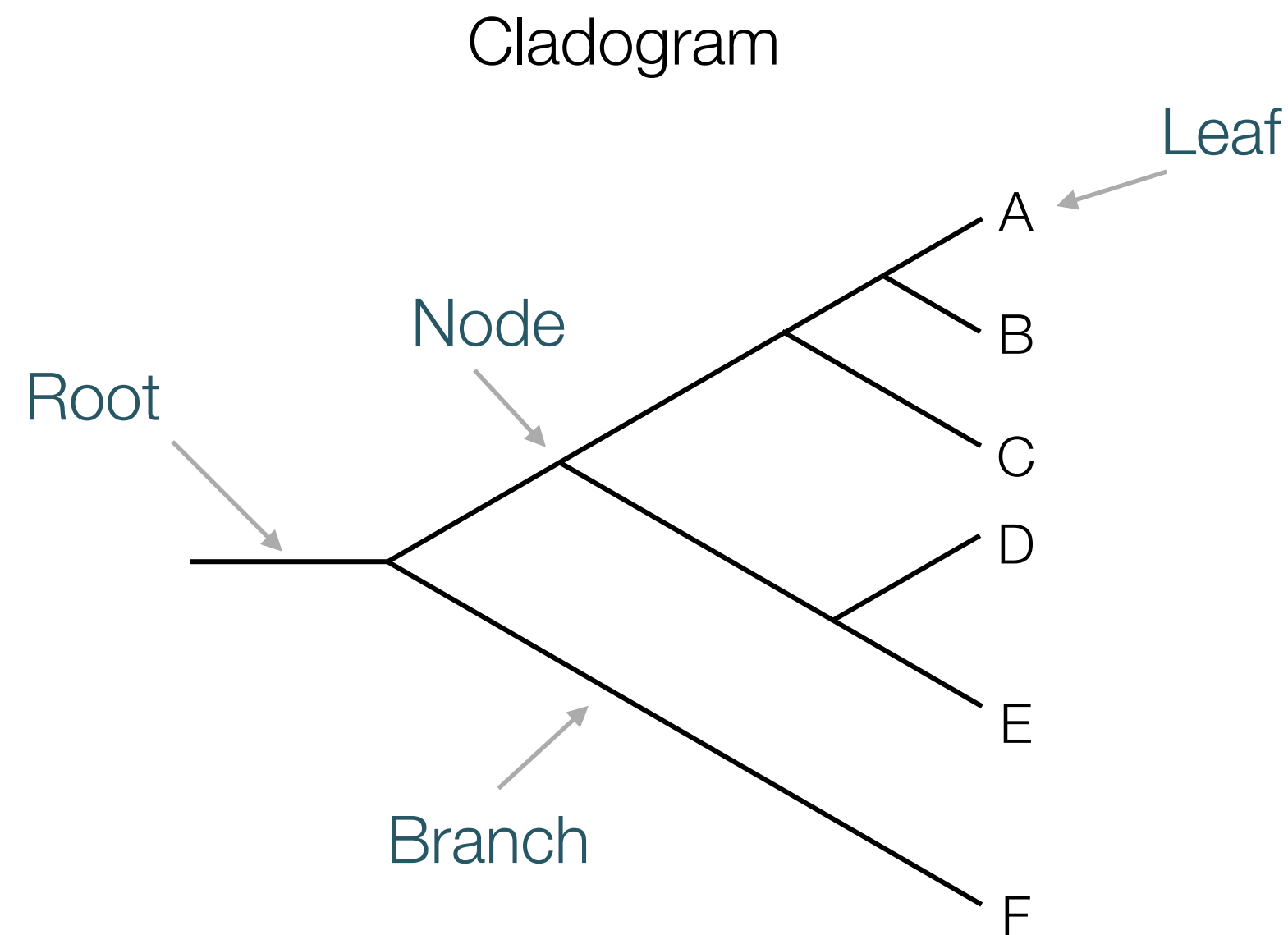
# What are phylogenetic trees?

- A graphical representation to visualise the evolutionary relationships between genes or organisms



(b) unrooted tree

- Describes relatedness between taxa
- Undirected



(a) rooted tree

- Describes ancestry of taxa
- Directed
- Root = most recent common ancestor

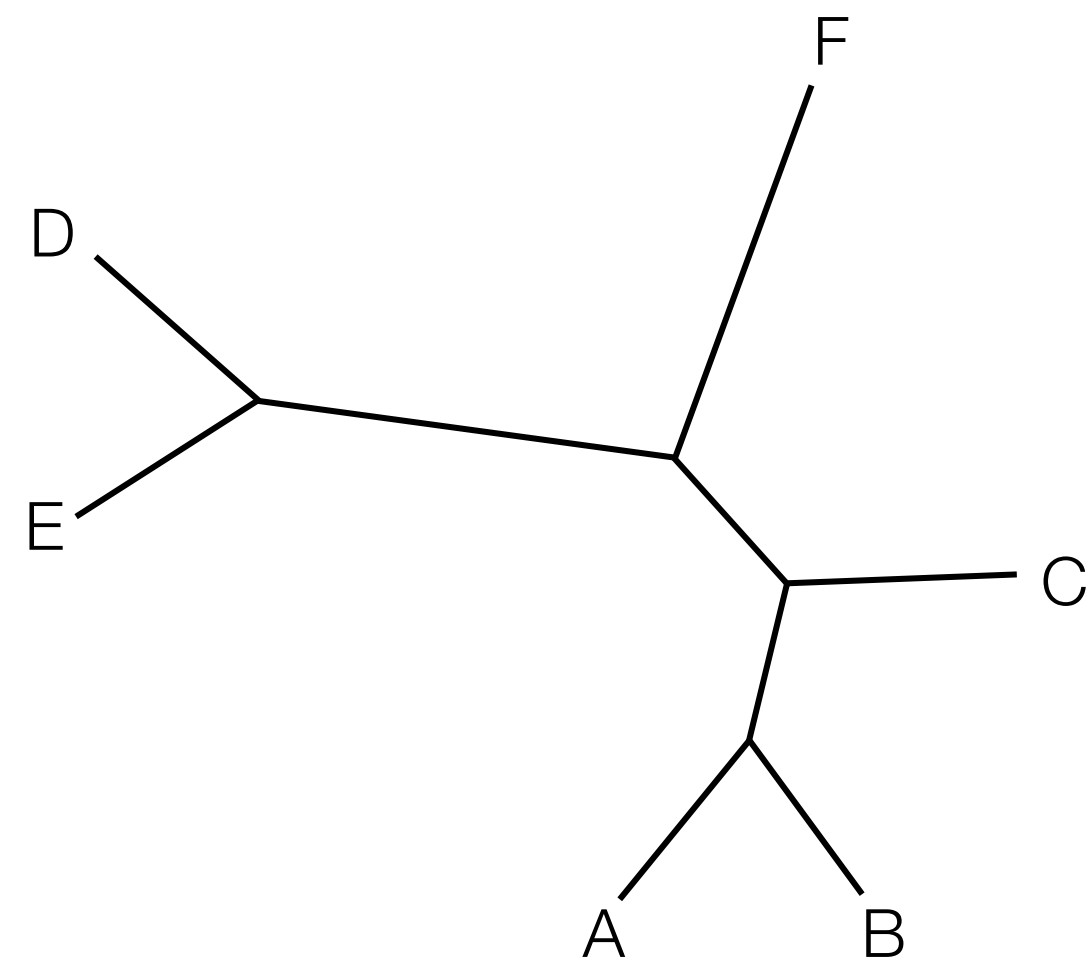
*Rooting:*

- Outgroup
- Mid-point rooting



# What are phylogenetic trees?

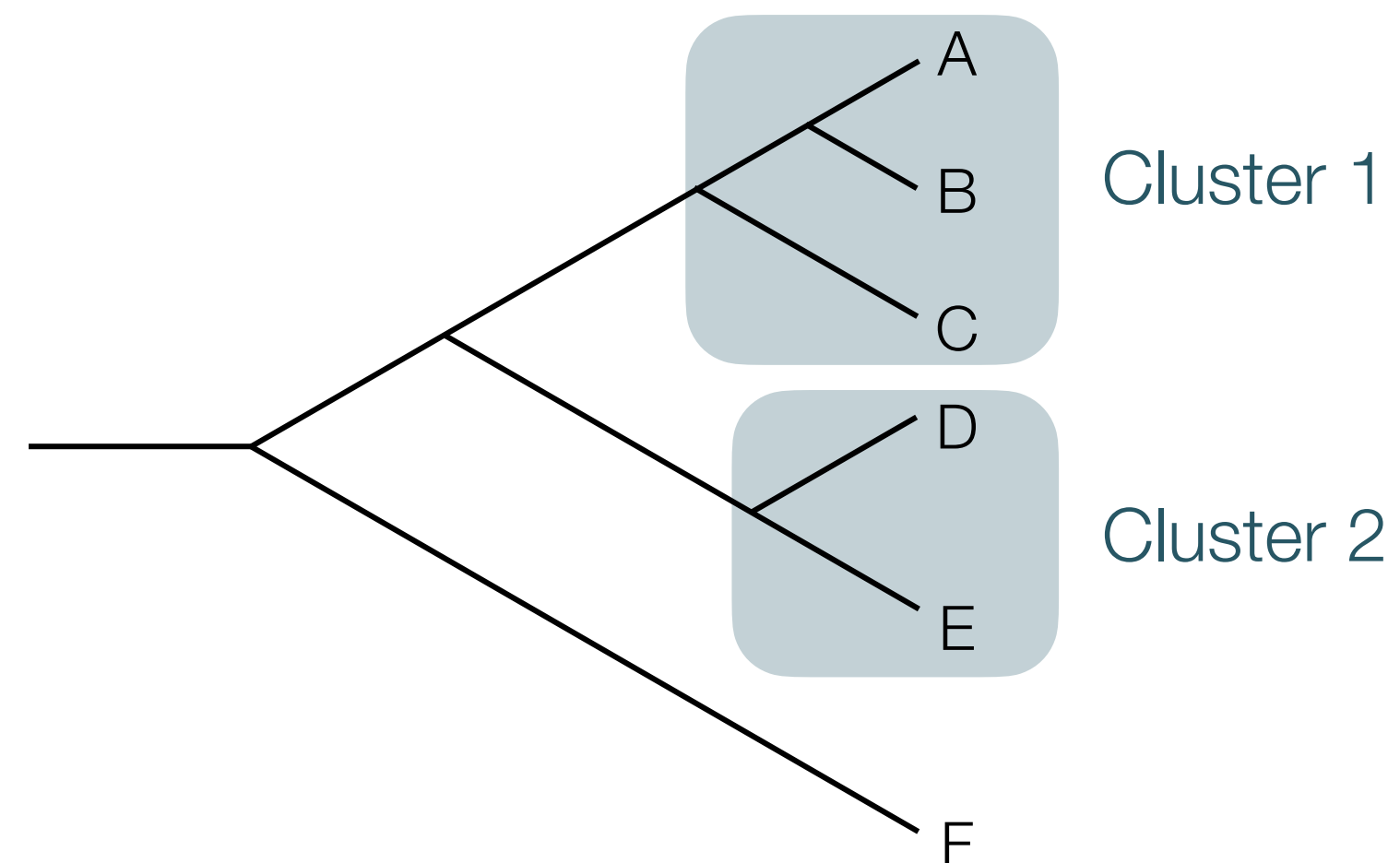
- A graphical representation to visualise the evolutionary relationships between genes or organisms



(b) unrooted tree

- Describes relatedness between taxa
- Undirected

Cladogram



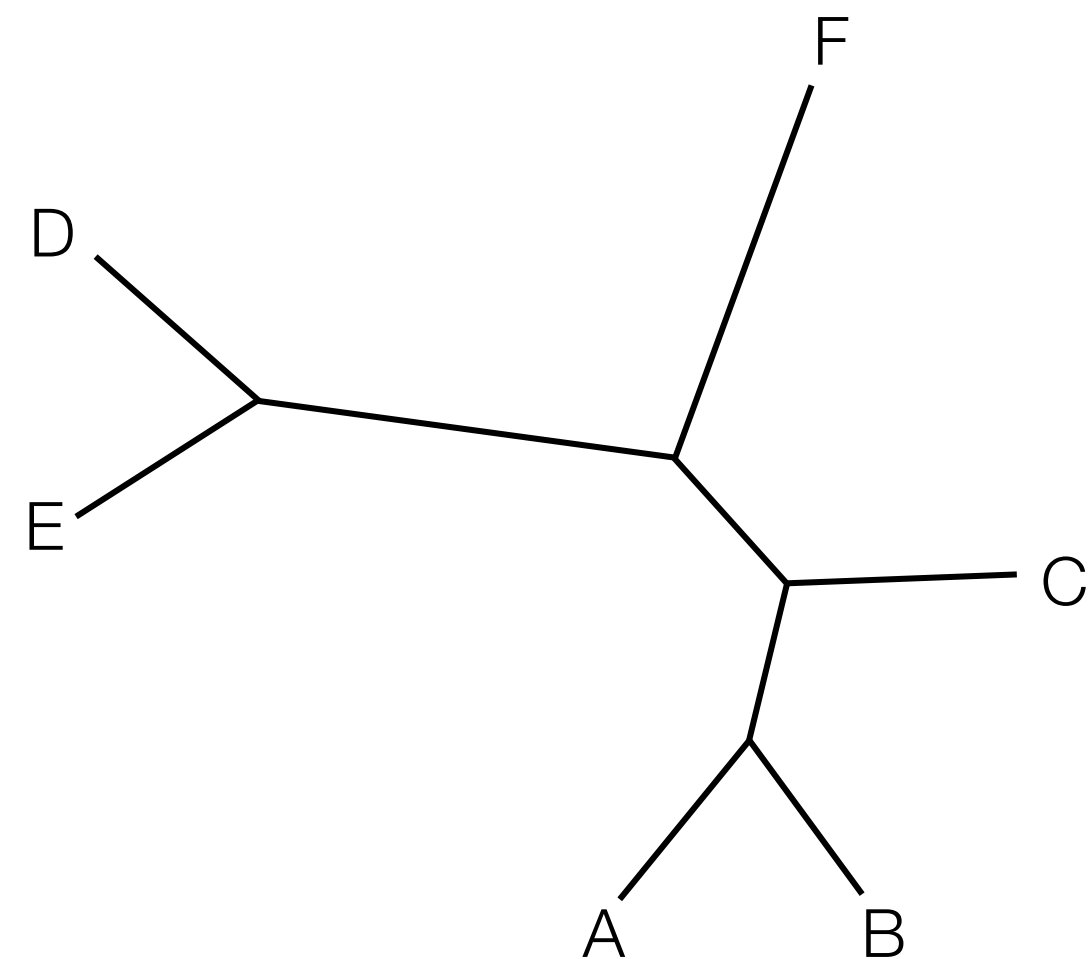
**Monophyletic groups = clade  
= shared common ancestor**

(a) rooted tree

- Describes ancestry of taxa
- Directed
- Root = most recent common ancestor

# What are phylogenetic trees?

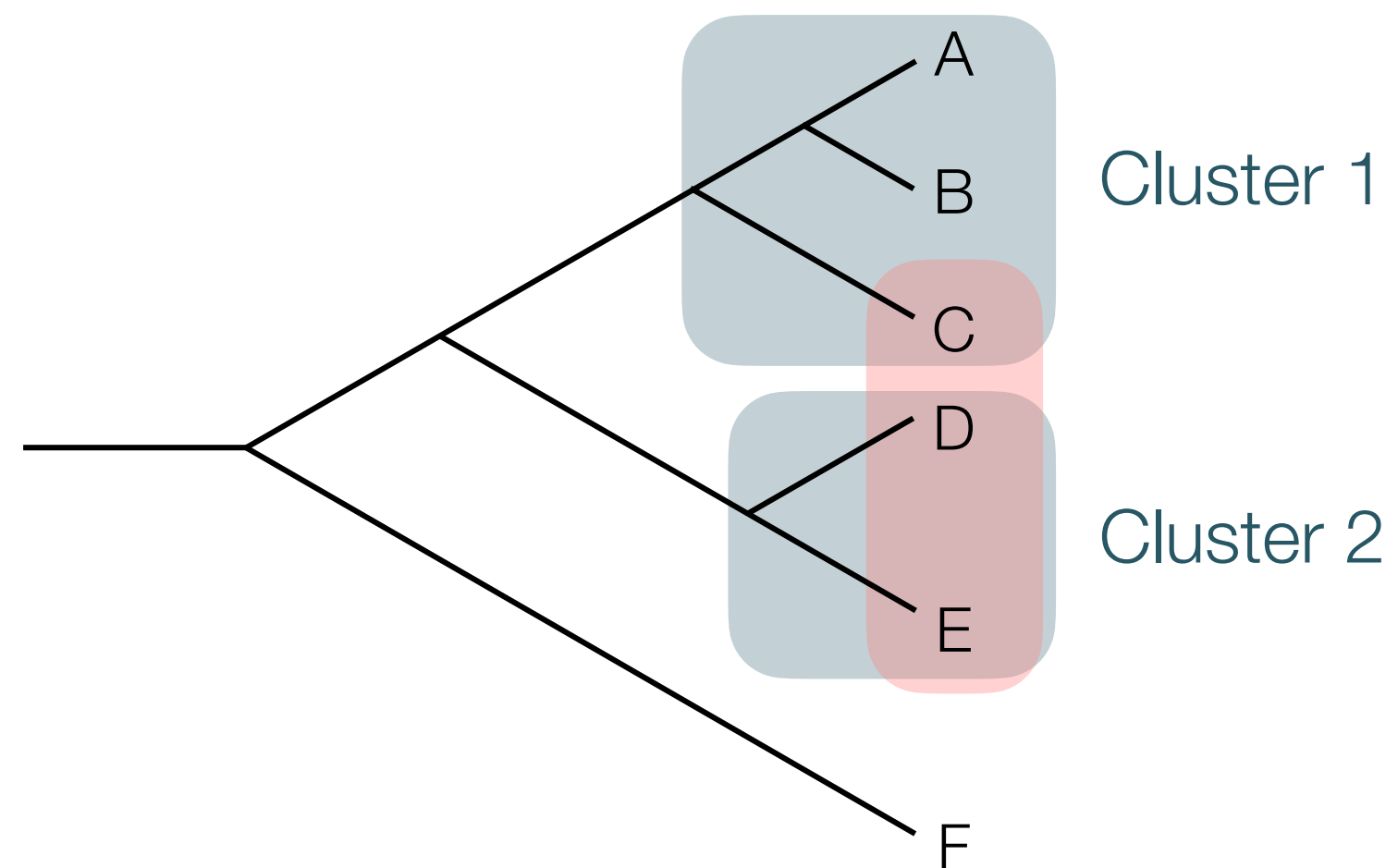
- A graphical representation to visualise the evolutionary relationships between genes or organisms



(b) unrooted tree

- Describes relatedness between taxa
- Undirected

Cladogram



(a) rooted tree

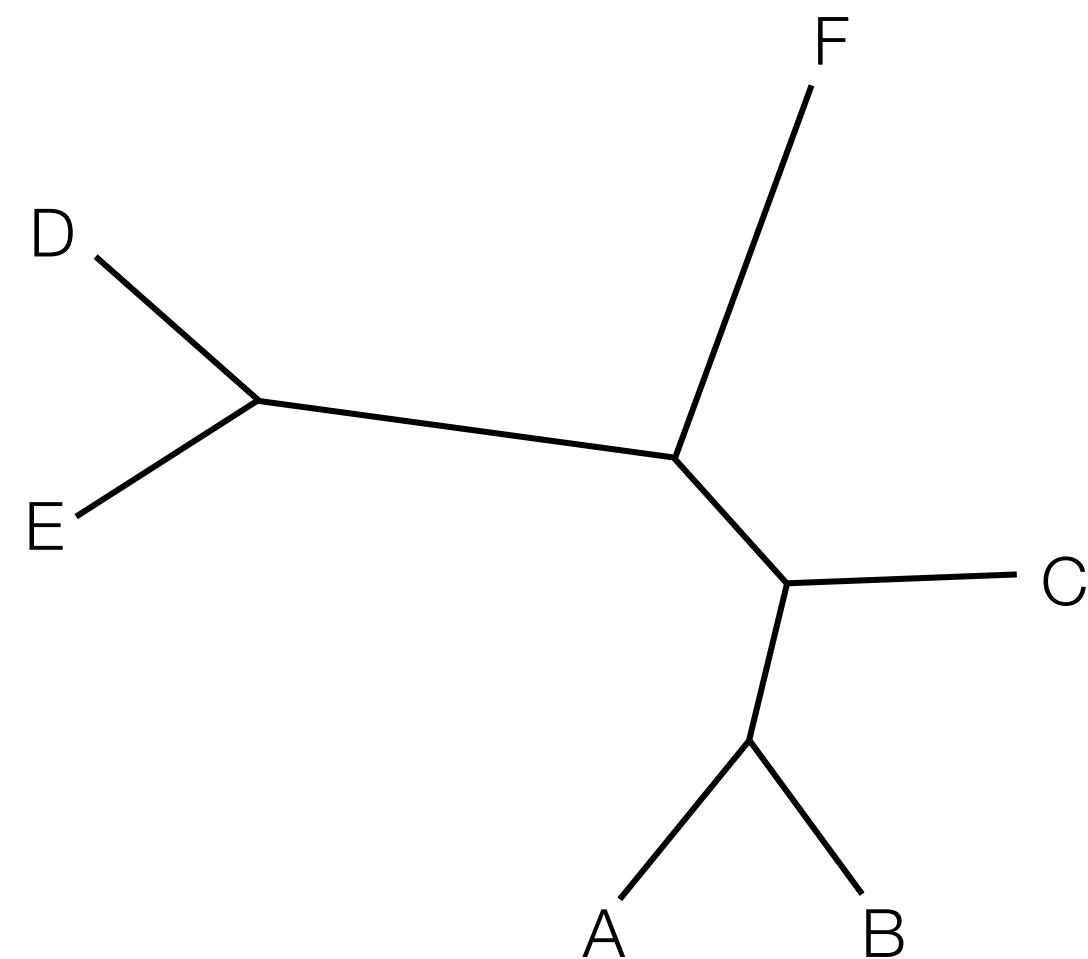
- Describes ancestry of taxa
- Directed
- Root = most recent common ancestor

**Monophyletic groups = clade  
= shared common ancestor**

**Paraphyletic group  
= does not include all descendants  
from inferred common ancestor**

# What are phylogenetic trees?

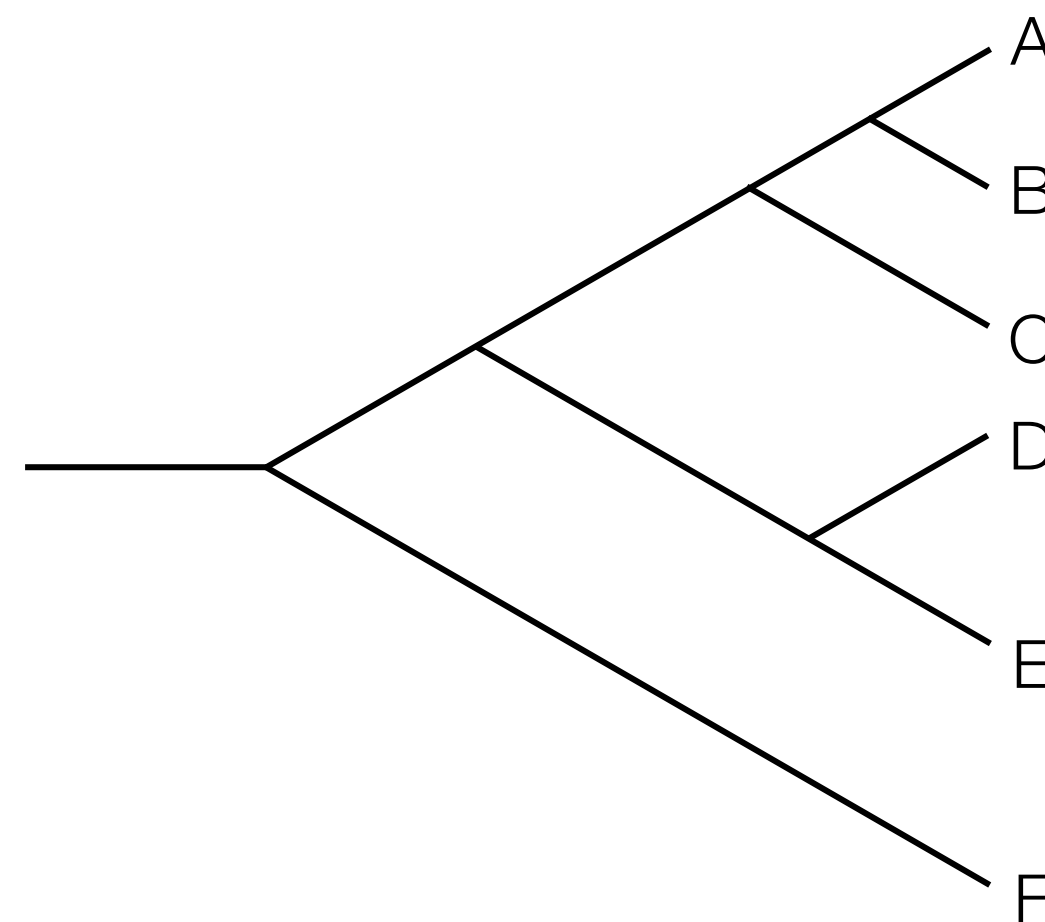
- A graphical representation to visualise the evolutionary relationships between genes or organisms



(b) unrooted tree

- Describes relatedness between taxa
- Undirected

Cladogram

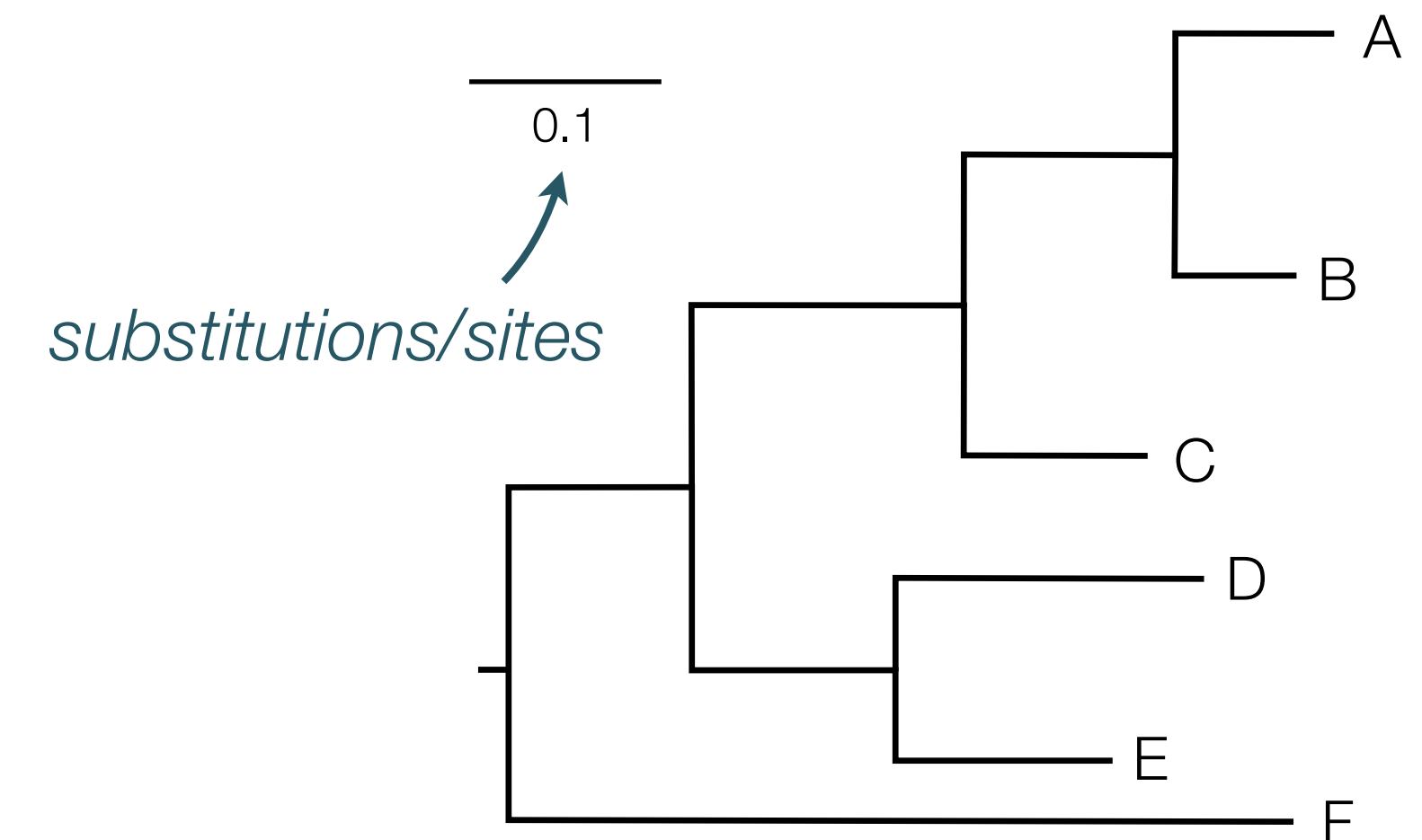


(a) rooted tree

- Describes ancestry of taxa
- Directed
- Root = most recent common ancestor

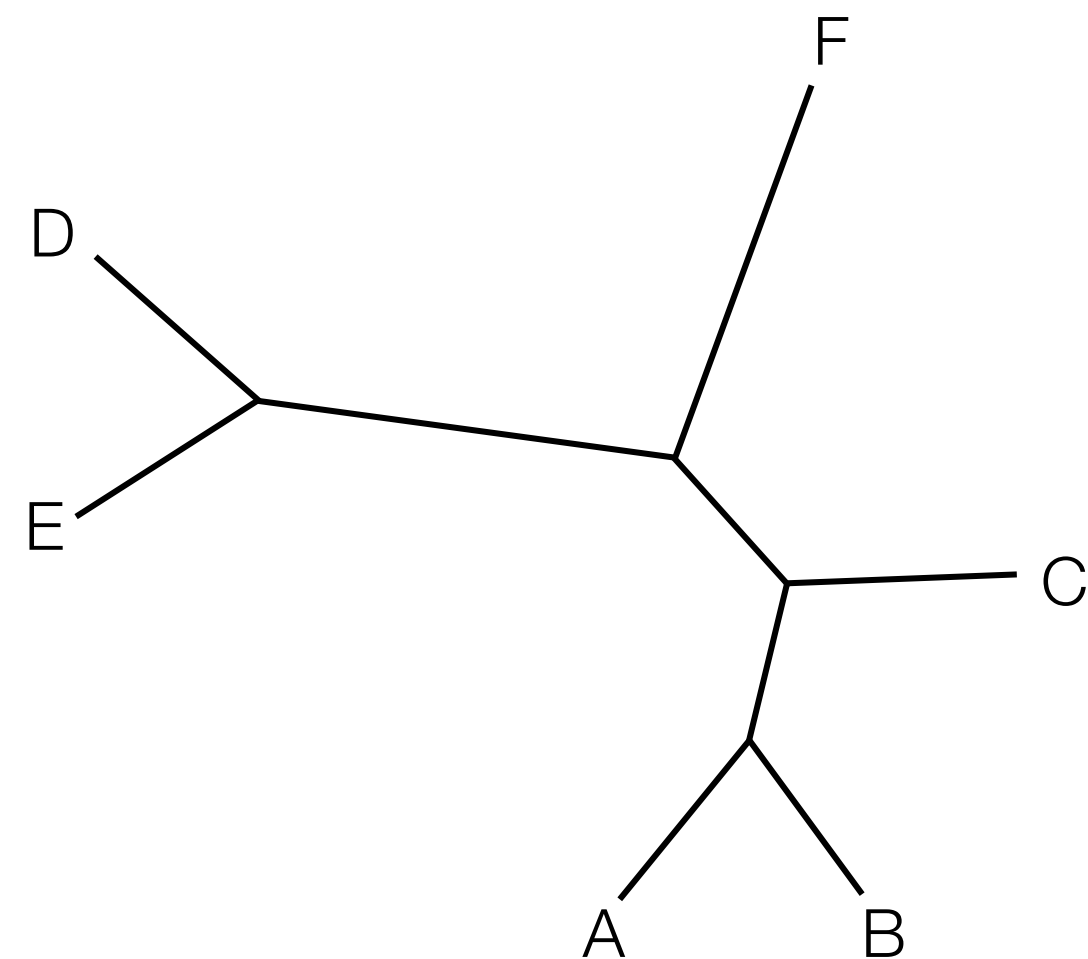
Phylogram

*Branch lengths = evolutionary distances*



# What are phylogenetic trees?

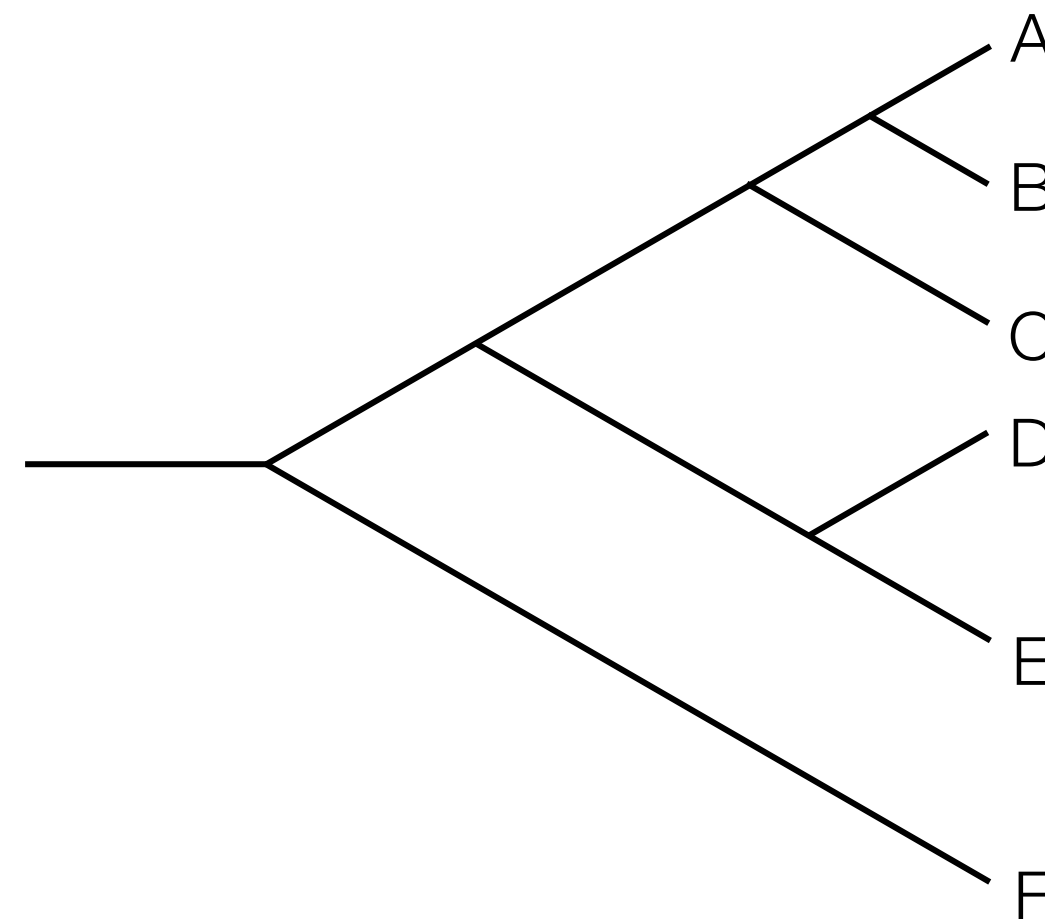
- A graphical representation to visualise the evolutionary relationships between genes or organisms



(b) unrooted tree

- Describes relatedness between taxa
- Undirected

Cladogram

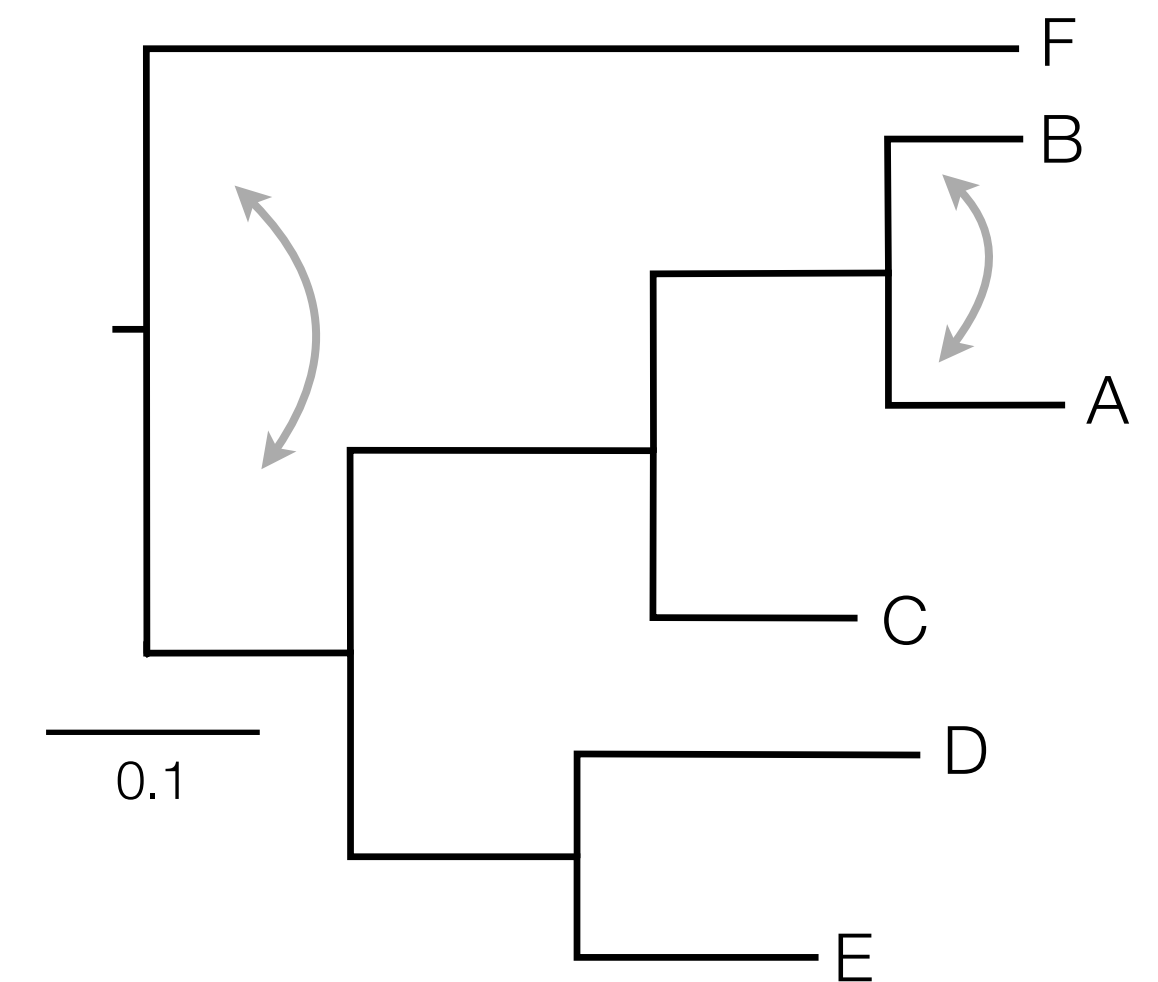


(a) rooted tree

- Describes ancestry of taxa
- Directed
- Root = most recent common ancestor

Phylogram

*Branch lengths = evolutionary distances*

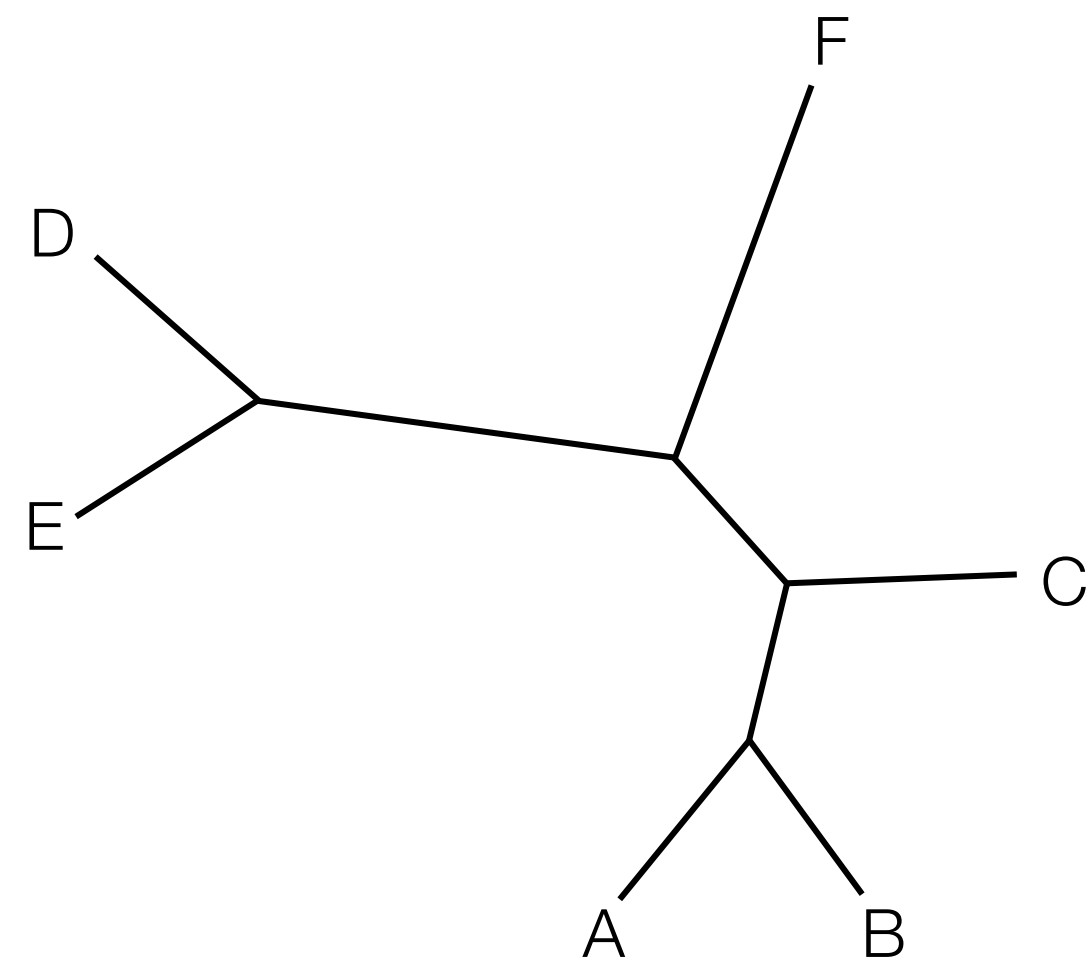


*Nodes can be rotated!  
Same tree topology*



# What are phylogenetic trees?

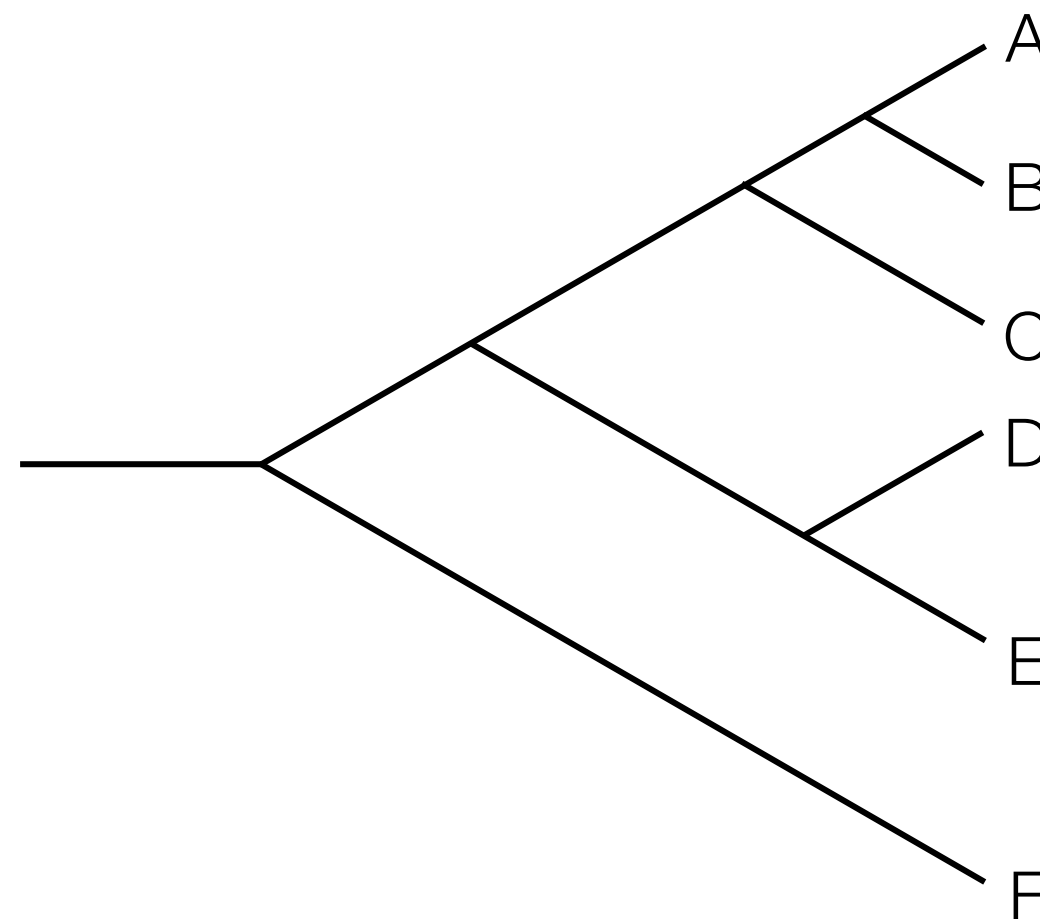
- A graphical representation to visualise the evolutionary relationships between genes or organisms



(b) unrooted tree

- Describes relatedness between taxa
- Undirected

Cladogram

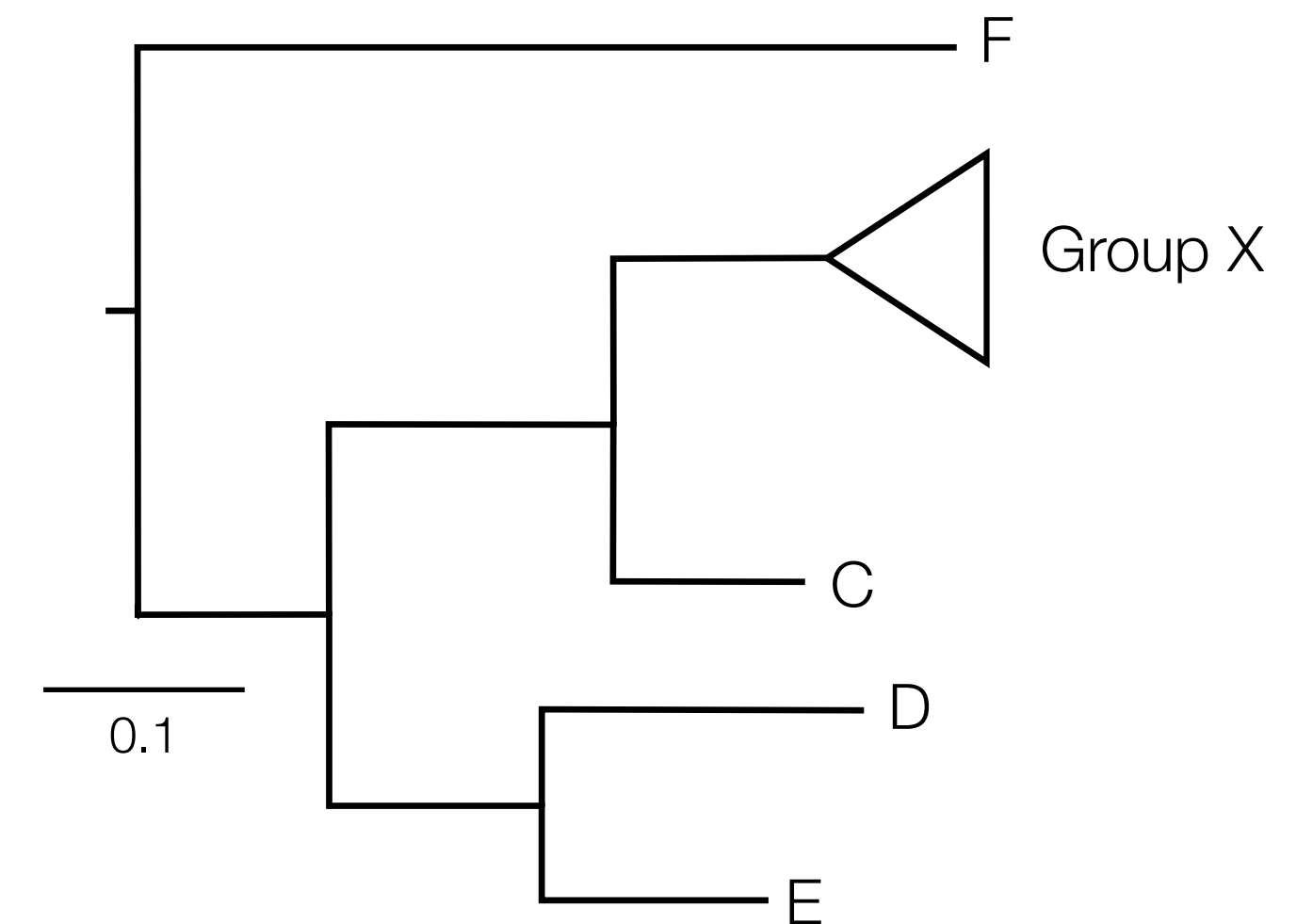


(a) rooted tree

- Describes ancestry of taxa
- Directed
- Root = most recent common ancestor

Phylogram

*Branch lengths = evolutionary distances*

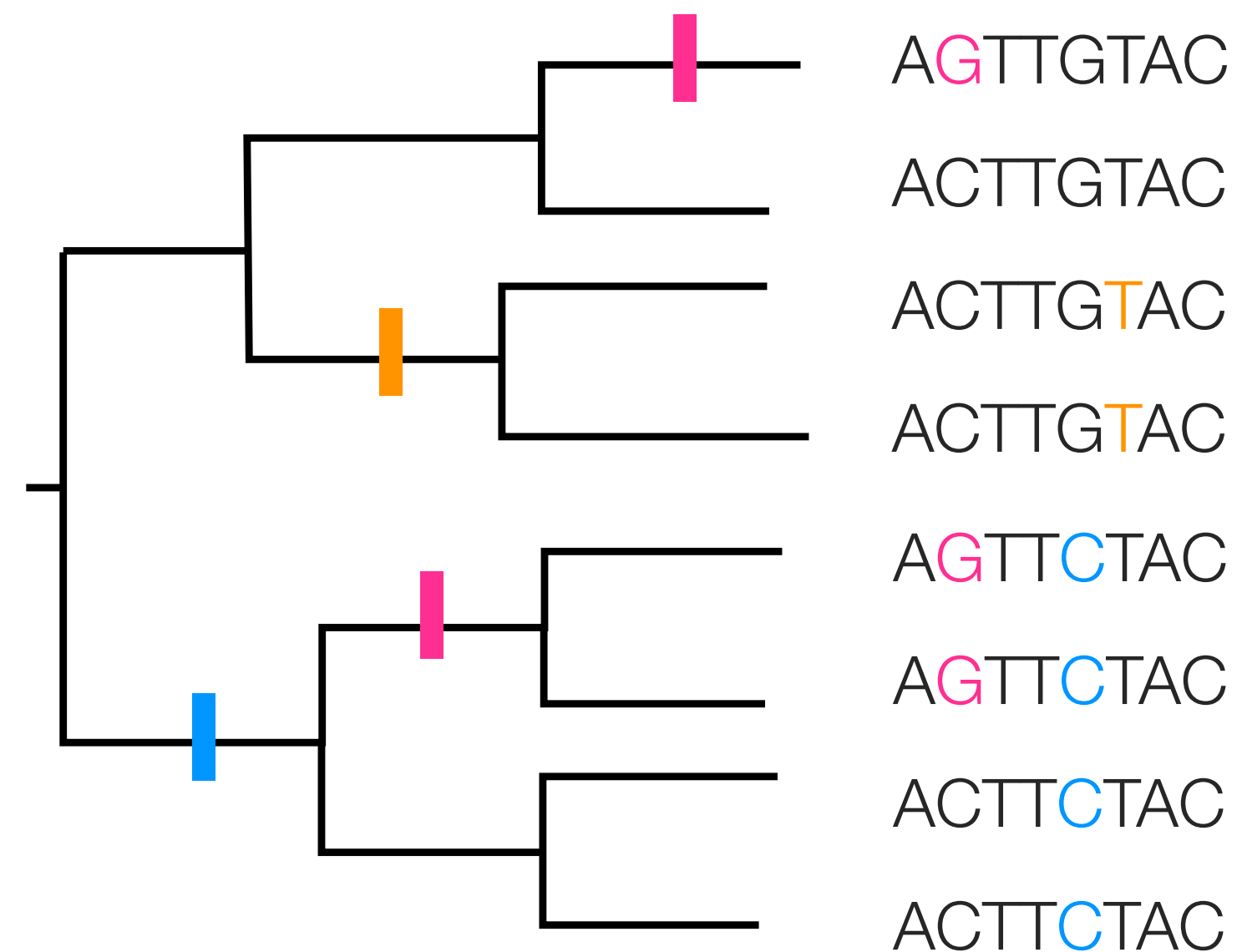


*Nodes can be rotated!  
Same tree topology*

*Branches can be  
collapsed*

# What are phylogenetic trees?

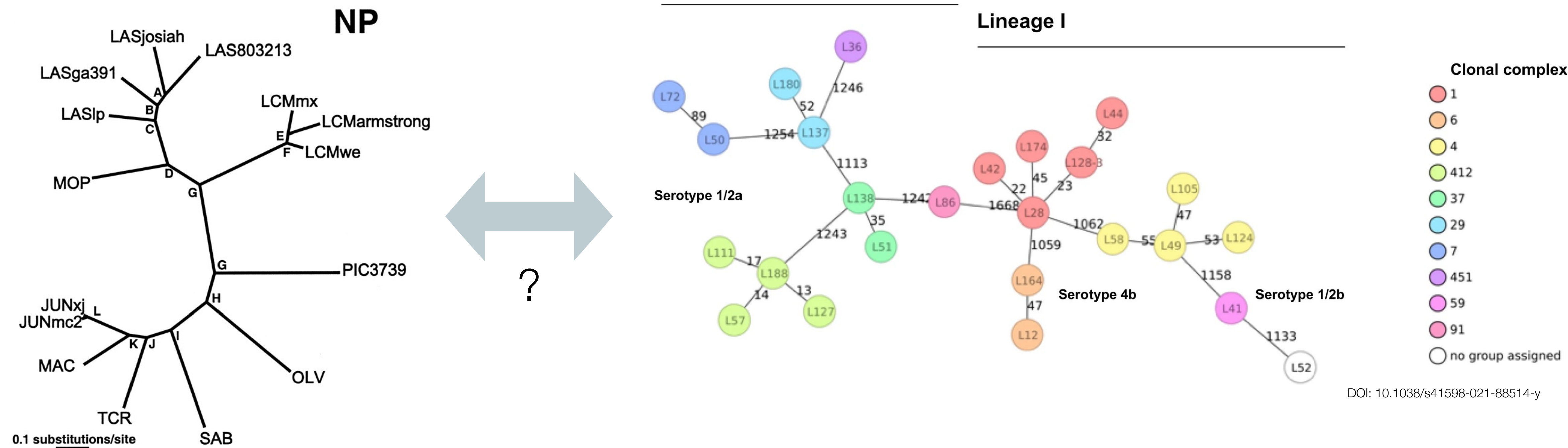
- Phylogenetics is based on the idea of **homology**, i.e. similarity due to shared ancestry
- Can be based e.g., on morphological traits, nucleotide polymorphisms (SNPs) or amino acid polymorphisms



- **Homoplasy:** Similarity that cannot be explained by shared ancestry, e.g., loss of feature, features that have arisen multiple times independently, or that have recombined

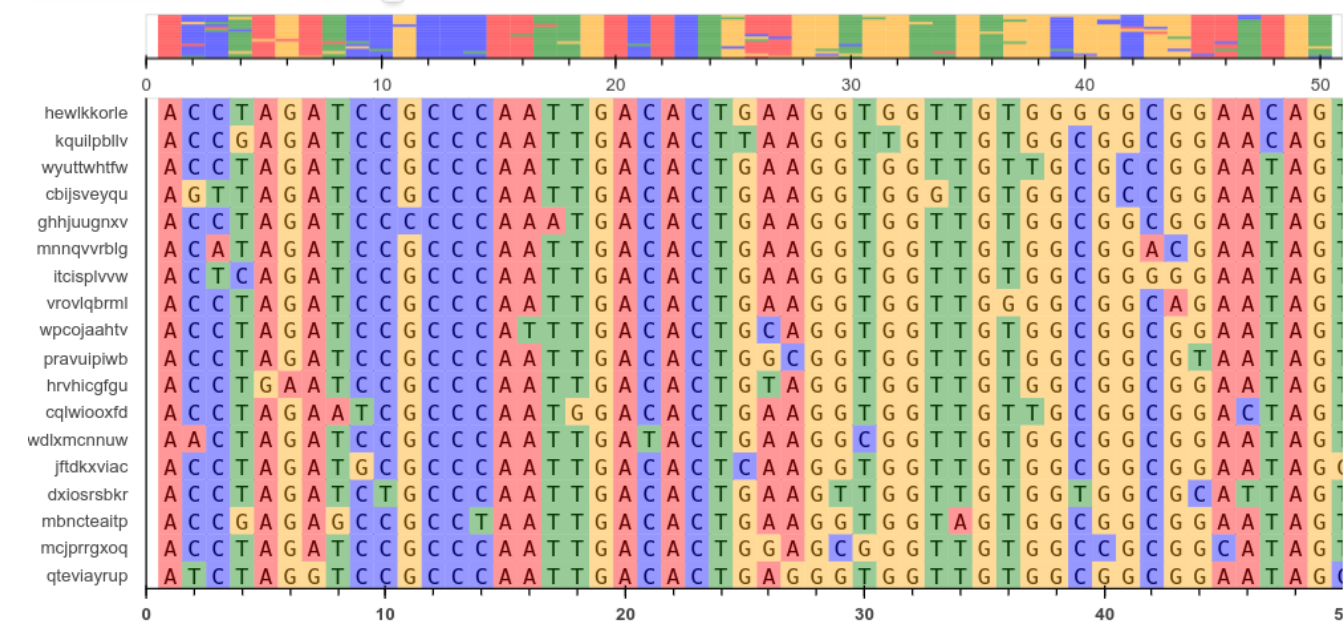
# What are NOT phylogenetic trees?

- Minimum spanning trees (MST)

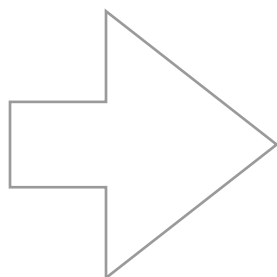


# Approaches for phylogenetic inference

- Distance-based

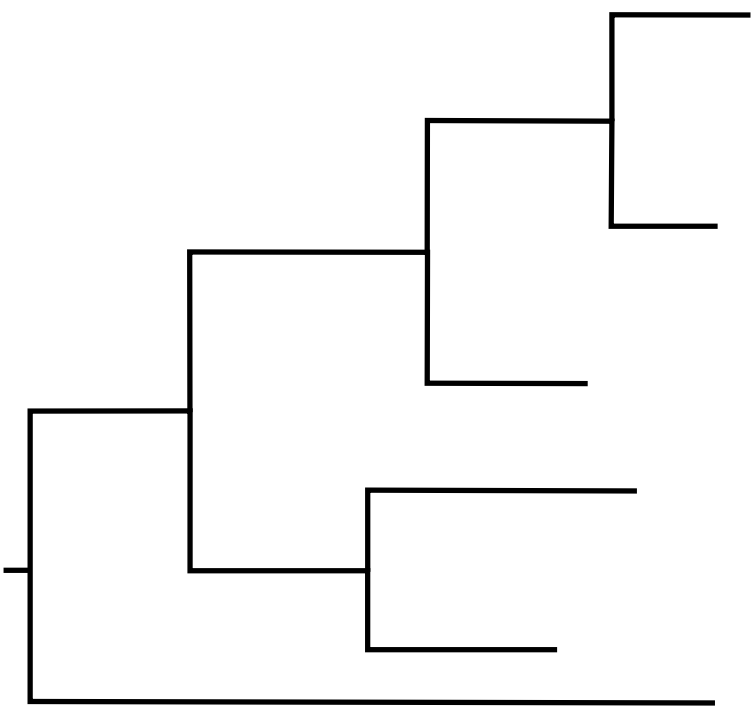
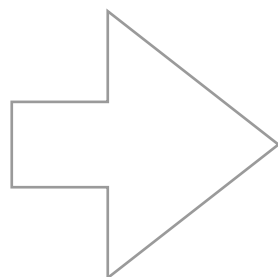


Sequence alignment



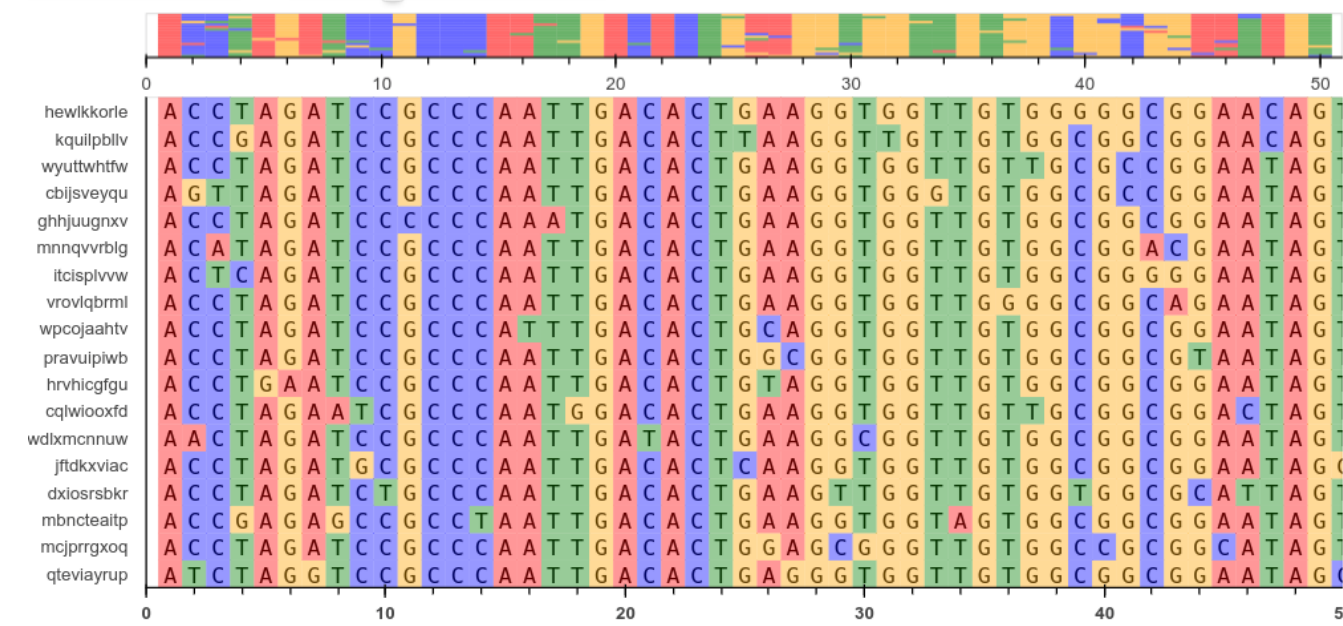
	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

Pairwise distance matrix

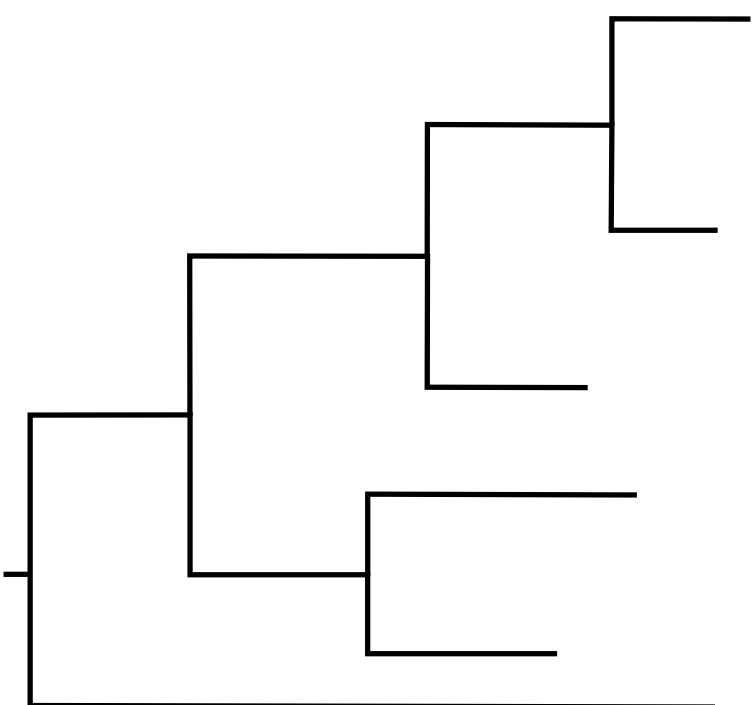
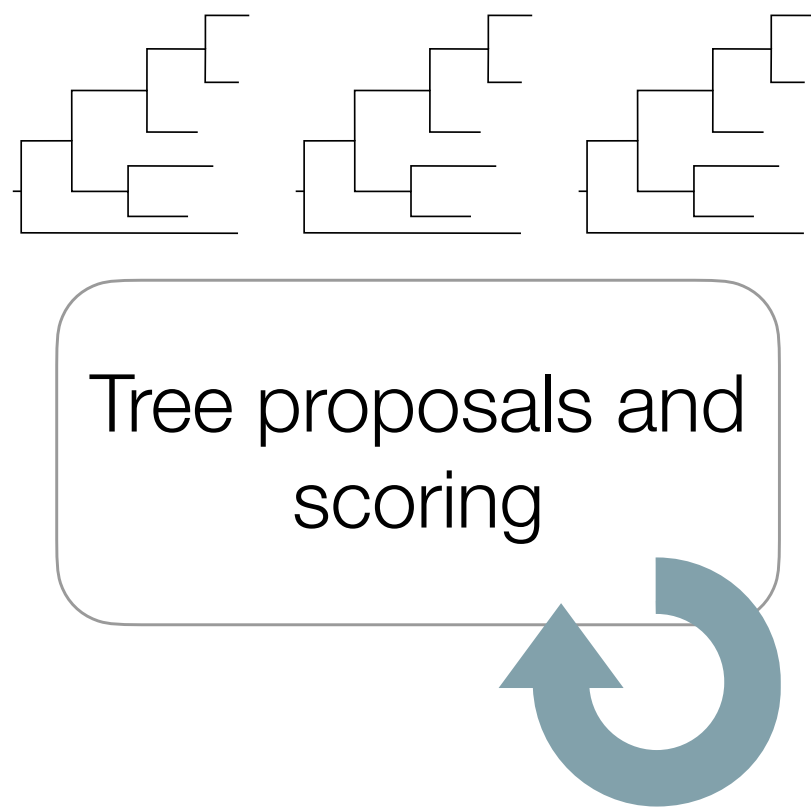
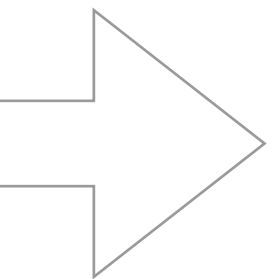


Output tree

- Character-based



Sequence alignment

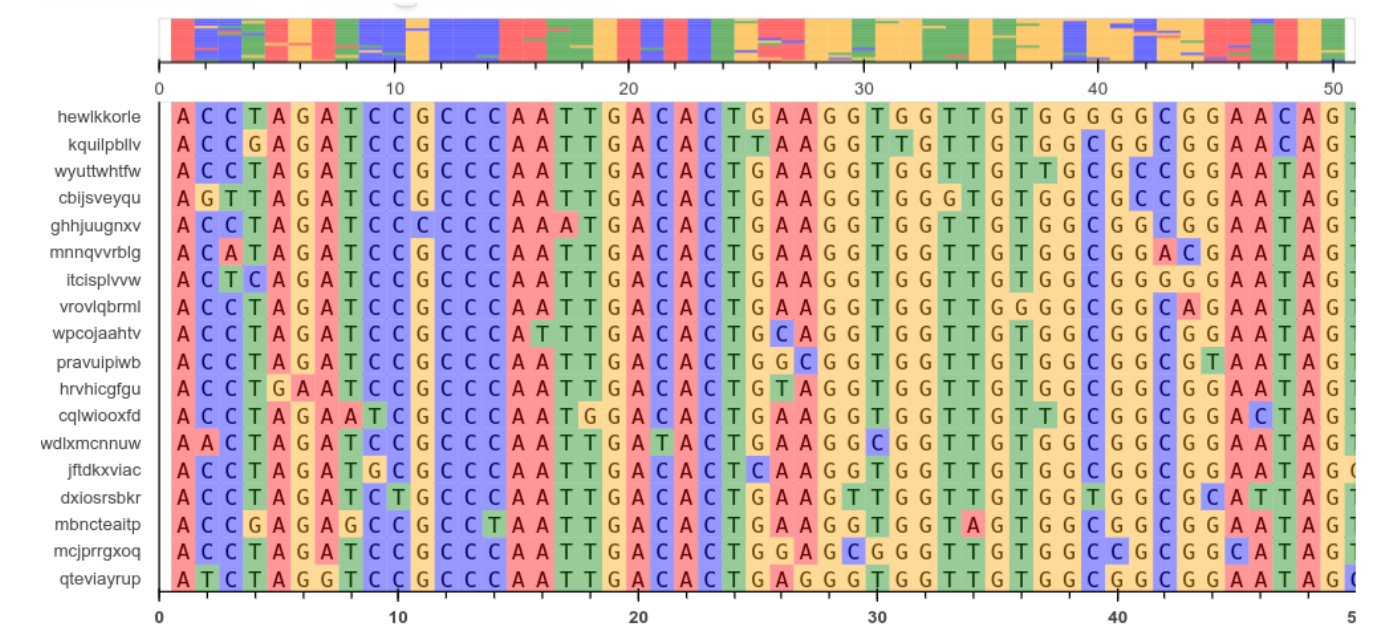


Output tree



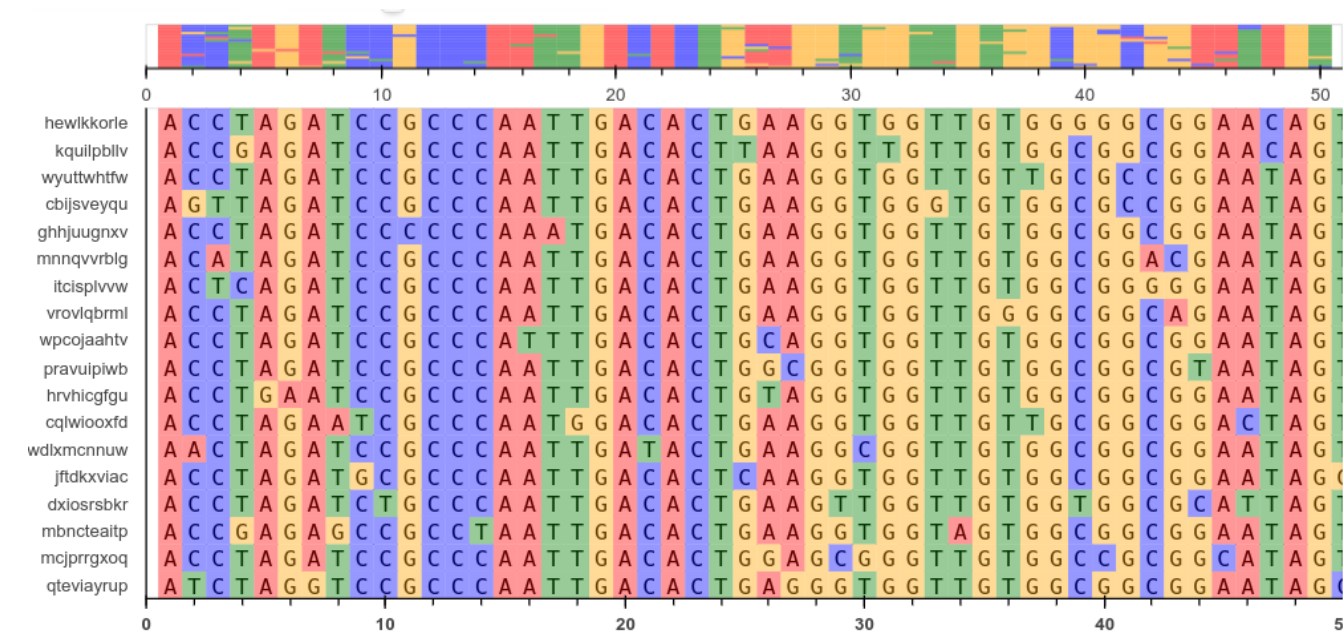
# Alignments/SNPs

- Tool to identify evolutionary relationships, usually in biological sequences, by aligning characters based on homology
- For phylogenetics:
  - Align whole genes, concatenated genes (e.g., marker genes) or whole genomes (e.g., viruses) to each other (e.g., mafft, Clustal, blast, ...)
  - Call SNPs against a reference directly from reads (e.g., snippy) or from our alignment made during assembly (e.g., samtools, bcftools, ...)
  - Kmer approaches, e.g.,
    - split kmer analysis (SKA) for small, closely related haploid genomes
    - Mash for distance calculation more distant genomes

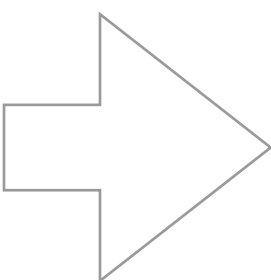


# Approaches for phylogenetic inference

- Distance-based

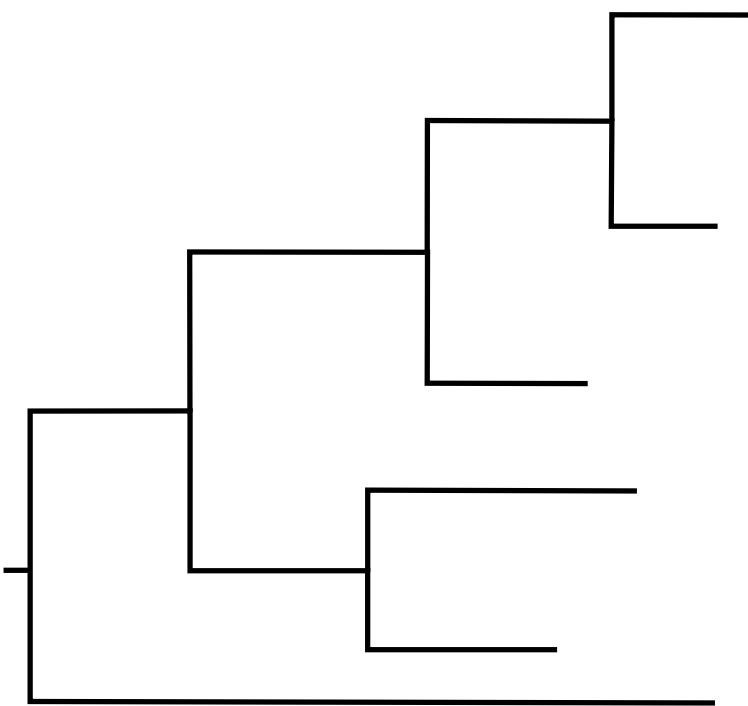
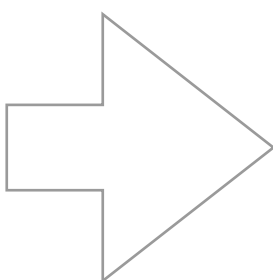


Sequence alignment



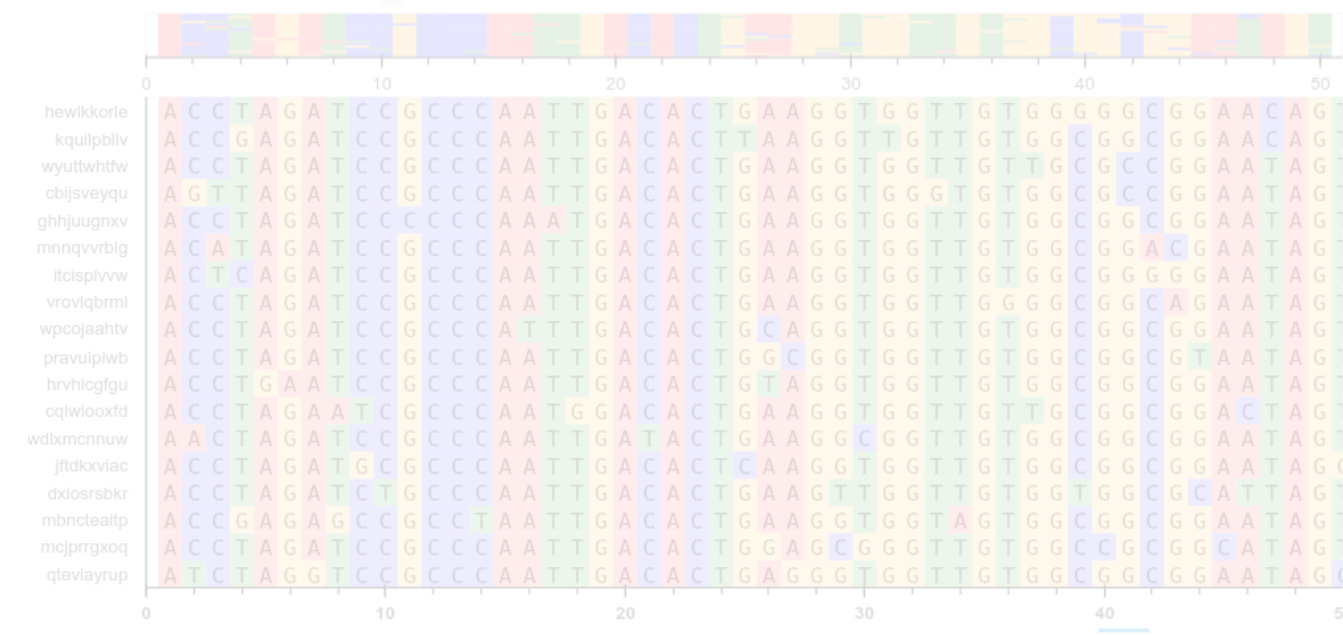
	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

Pairwise distance matrix

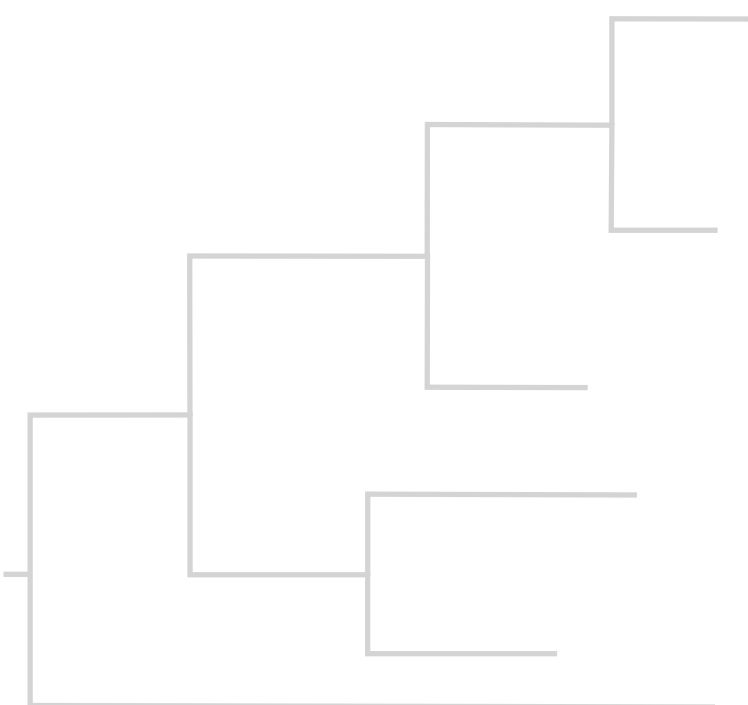
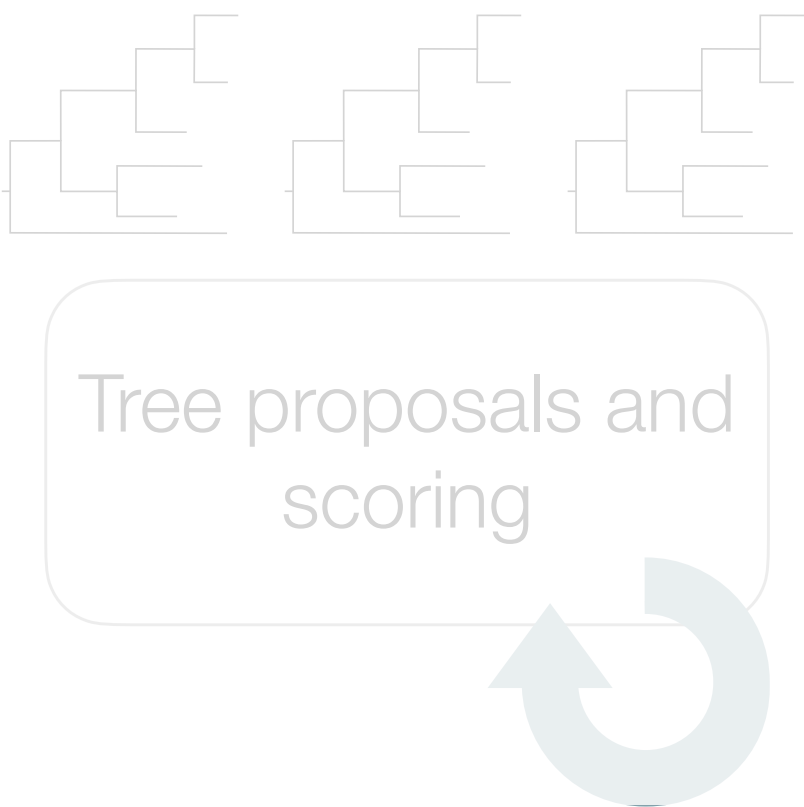


Output tree

- Character-based



Sequence alignment

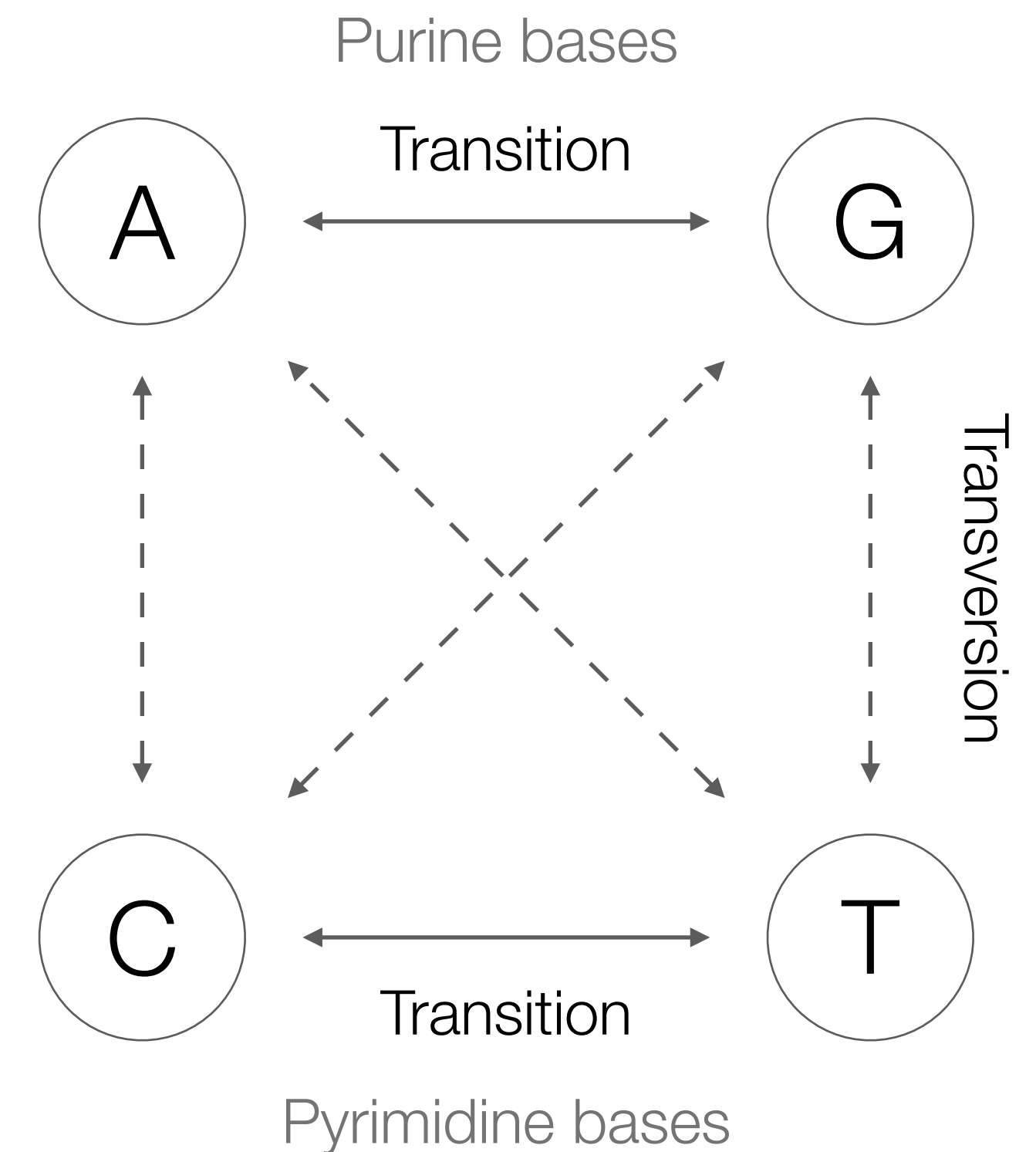


Output tree

# Distance-based tree inference

---

- (Uncorrected) p distance: proportion of sites at which two sequences are different
- Evolutionary nucleotide models:
  - Transitions/transversions
  - $N_{\text{actual}} > N_{text{observed}}$  : some fraction of “conserved” positions mutated twice
  - Examples:
    - Jukes Cantor: all equal
    - Kimura 80: transitions more likely than transversions
- Models for amino acid evolution or codon evolution exist as well



# Distance-based tree inference

---

- Example p distance vs evolutionary model

```
ATGTCACCACAAACAGAACTAAAGCAAGTGTTGGATT
|||||X|||||||X|||||X|||||||
ATGTCTCCACAAACAGAAGCTAAAGGAAGTGTTGGATT
```

- p distance

000001000000000001000000100000000000 = 3

$$p = \frac{D}{L} = \frac{3}{38} = 0.079$$

- Evolutionary model K80: transitions = 1; transversions: 2

$$d = 0.083$$



# Distance-based tree inference

NUCLEOTIDE SUBSTITUTION MODELS	Unequal base frequency		Equal base frequency	
	1 substitution type	Felsenstein (F81)	Jukes Cantor (JC, JC69)	
	2 substitution types	Hasegawa-Kishino-Yano (HKY85) Felsenstein (F84)  <i>1 transition</i> <i>1 transversion</i>	Kimura-2-parameter (K2P, K80)  <i>1 transition</i> <i>1 transversion</i>	
	3 substitution types	Tamura-Nei (TN, TN93)  <i>2 transitions</i> <i>1 transversion</i>	Kimura-3-parameter (K3P, K81, K3ST)  <i>1 transition</i> <i>2 transversions</i>	
	4 substitution types	Generalised time-reversible model (GTR)	Symmetric model (SYM)	
RHAS*	Gamma distribution (G) Proportion of invariable sites (I)		<i>Gamma distributed rate variation among sites</i> <i>Extend of static, unchanging sites in a dataset</i>	

\*Rate heterogeneity among sites

# Distance-based tree inference

---

- Pairwise distances are calculated between all sequences
- Tree topology and branch length is calculated, e.g. through clustering **UPGMA** (Unweighted Pair Group Method with Arithmetic mean = hierarchical clustering): implies constant molecular clock for all species/taxa/sequences  
(→ *pretty poor trees*)

## **Neighbour-Joining:**

Minimises the sum of all branch lengths in the tree

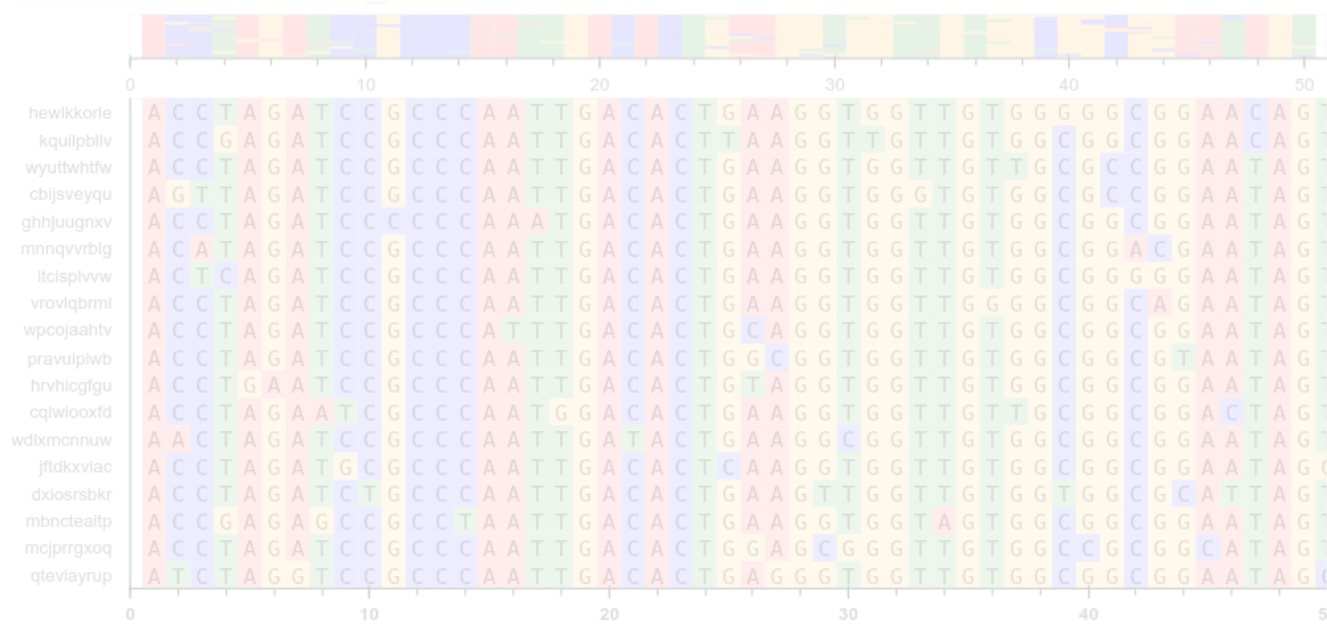
If input matrix is correct, the output tree will be correct

even if matrix is not correct, tree topology is *usually pretty good*

- Main advantages: gives only one tree as output, very fast computation

# Approaches for phylogenetic inference

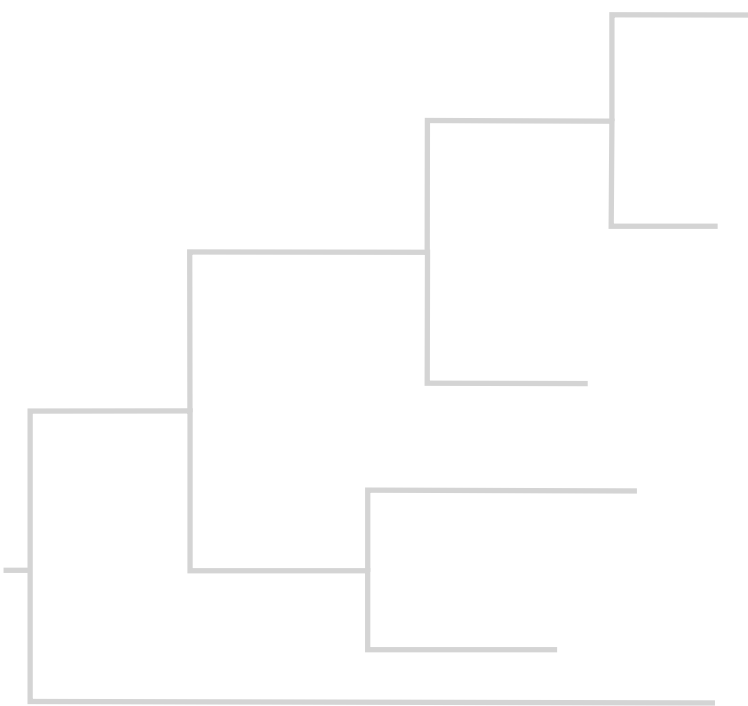
- Distance-based



Sequence alignment

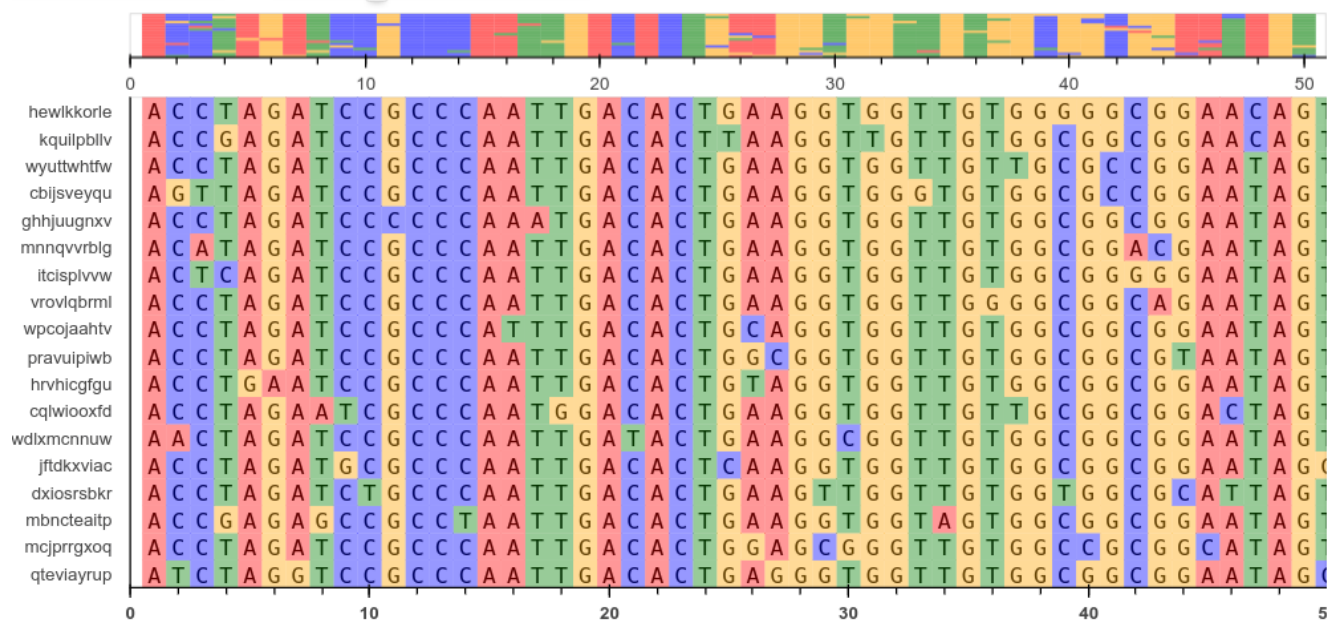
	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

Pairwise distance matrix

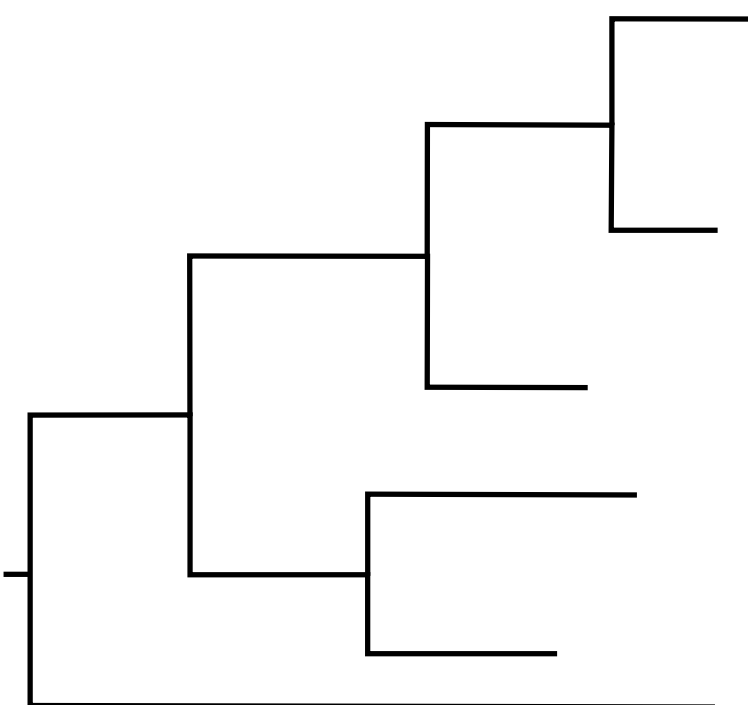
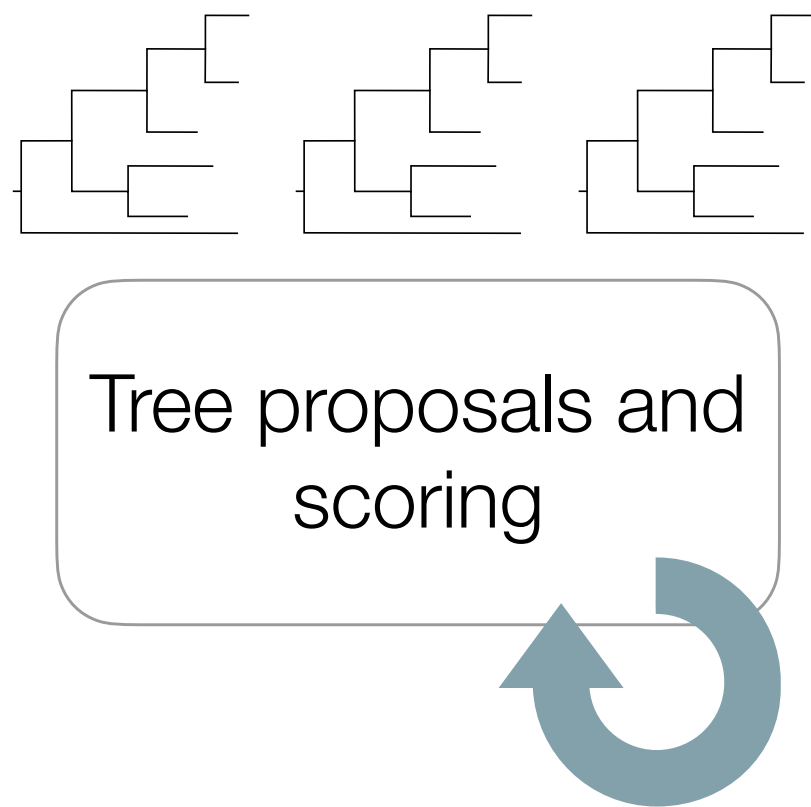


Output tree

- Character-based



Sequence alignment



Output tree

# Character-based phylogenetic inference - Maximum Likelihood

---

- Likelihood refers to the probability of the data given a model (e.g., of nucleotide sequence evolution)
  - We test different hypotheses = tree topologies with branch lengths, parameters of sequence model
  - We aim to find the tree topology, branch length and parameters of evo model that maximises the probability of observed data (= sequences)
- Advantages:
  - ✓ Inherently based on statistical and evolutionary models
  - ✓ Can be used for character and rate analyses
  - ✓ Can be used to infer sequences of extinct ancestors
- Disadvantages:
  - Computationally expensive
  - Violations of model assumptions can lead to incorrect trees



# Tree reliability - Bootstrapping

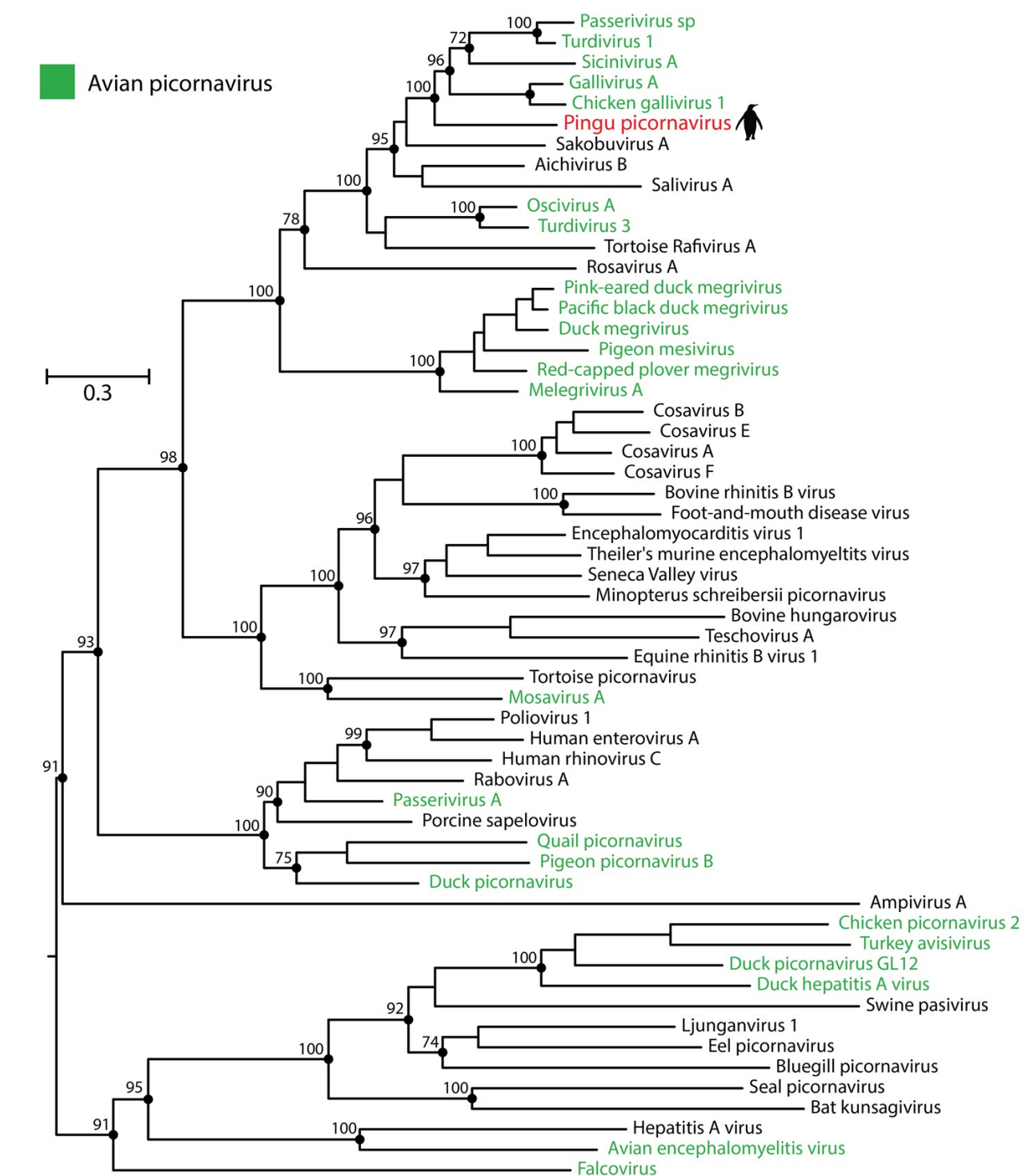
- Re-sample alignment: Randomly sample alignment columns with replacement to create many replicate data sets of equal size
- Build phylogenetic tree for each replicate data set
- Most frequently constructed tree is considered the one with most support
- Bootstrap values are the observation frequency of each branch

Original sequence

1	2	3	4	5	6
A	T	G	A	C	C
A	T	A	A	C	T
A	T	A	A	C	T
A	T	G	A	C	T

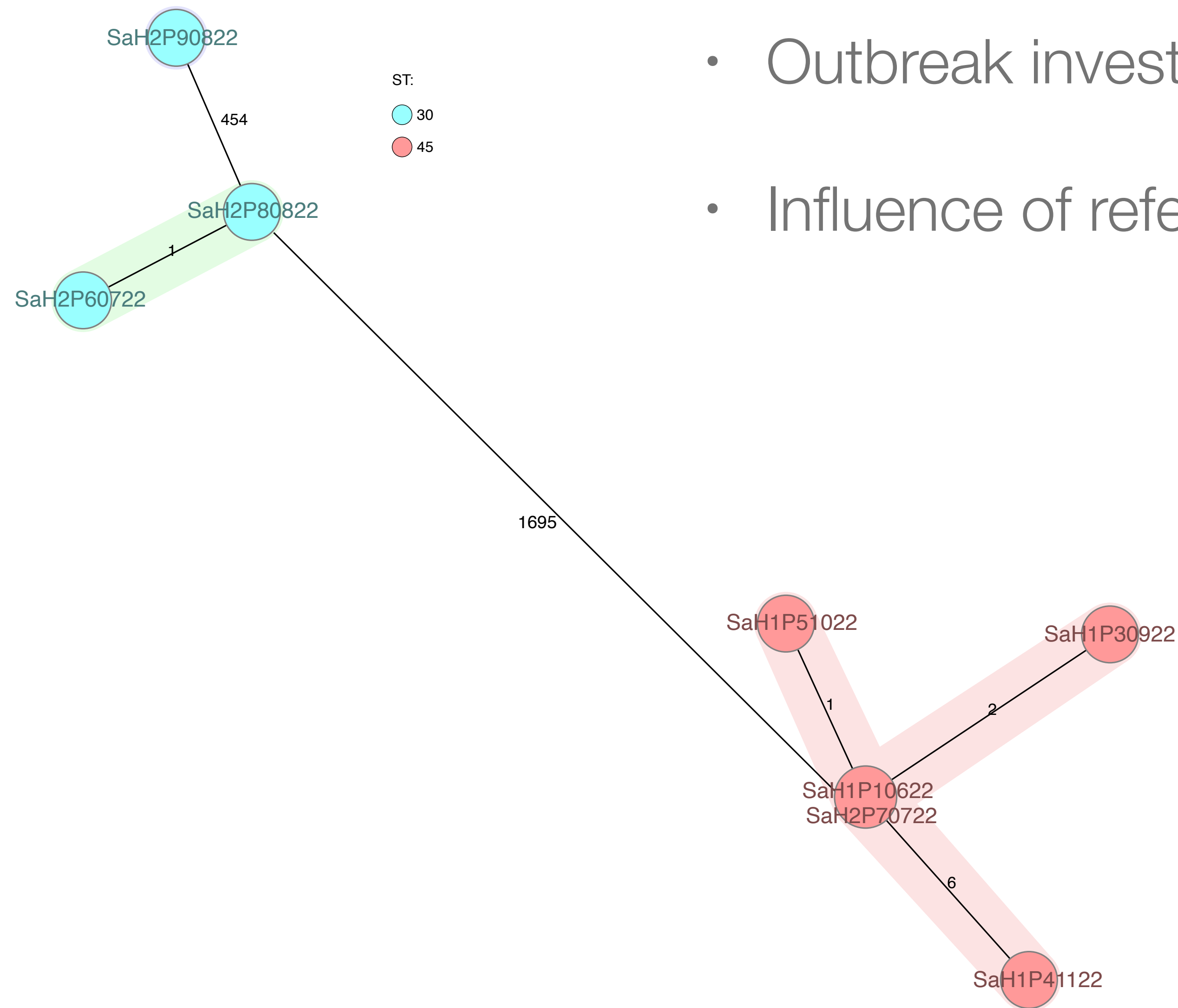
Bootstrap sequence

3	2	1	1	5	4
G	T	A	A	C	A
A	T	A	A	C	A
A	T	A	A	C	A
G	T	A	A	C	A



# Exercise tomorrow

---



- Outbreak investigation for our MRSA in hospital 1
- Influence of reference selection

Question: How much do you want to code in the next exercise?