Phylogenetics

Dr Johanna Rhodes

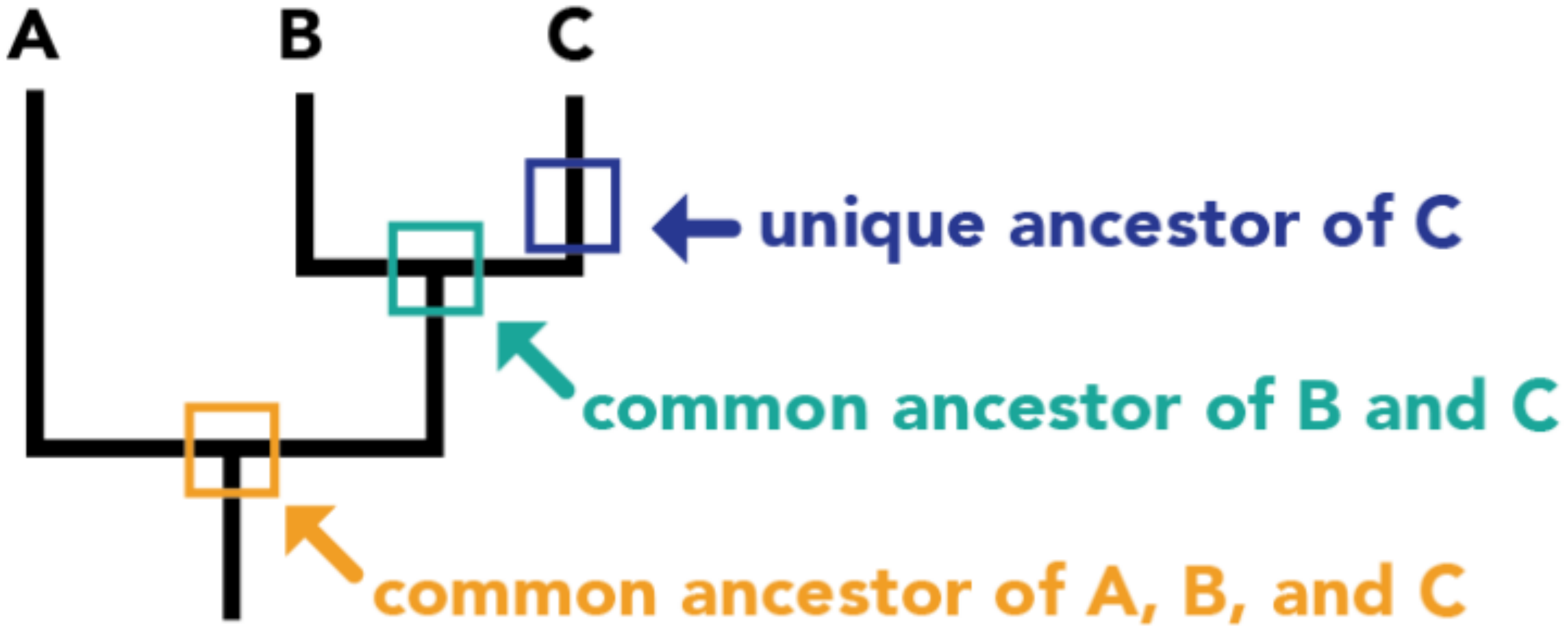
Learning outcomes

- **Basic phylogenetic theory**
- **How to construct a phylogeny**
  - different methods/input data
  - visualisation

Phylogenetics

Shared ancestry

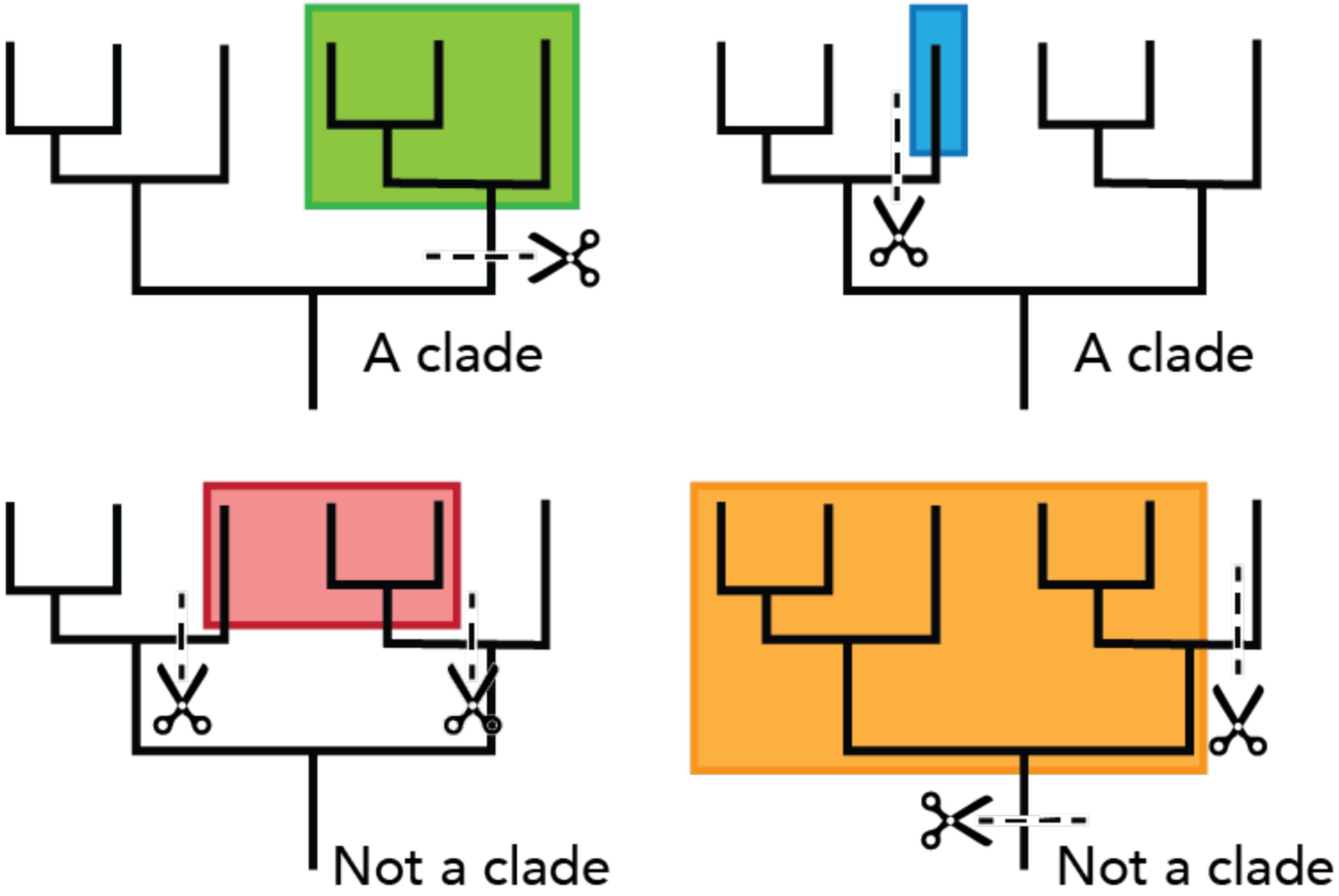
- Phylogenies trace patterns of shared ancestry between lineages.
- Each lineage also has ancestors that are unique to that lineage and ancestors that are shared between lineages - common ancestors



Phylogenetics

Clades

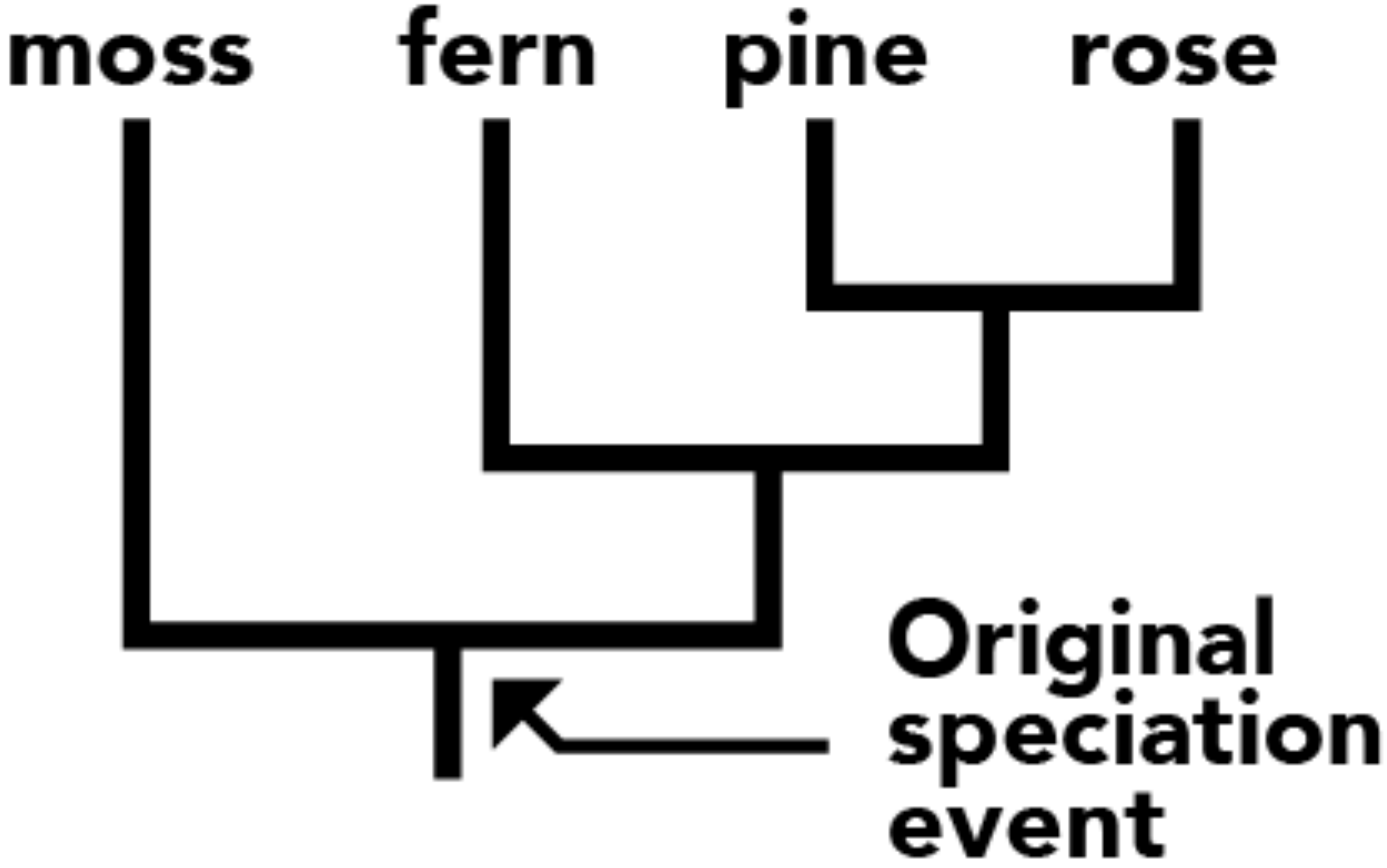
- A clase is a grouping that includes a common ancestor and all descendants (living or extinct) of that ancestor
- Phylogenies make it easy to tell if a group of lineages forms a clade.
- Clades may include a few isolates/species or many thousands, and can be nested within larger clades



Phylogenetics

Interpretation

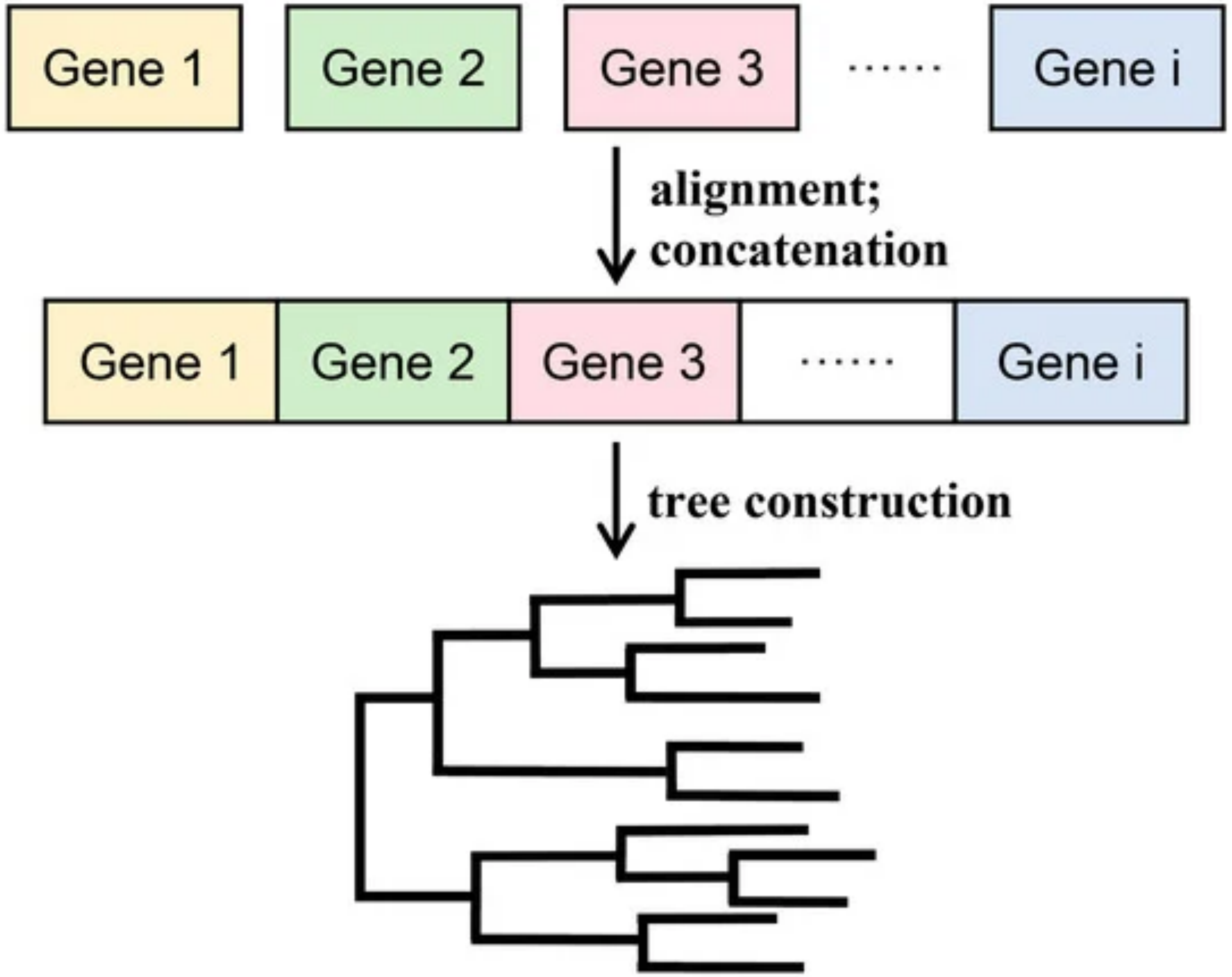
- They do not imply some species/ isolates are more ‘advanced’ than others.
- A speciation event resulted in two lineages - moss, and fern/pine/ rose. Both lineages have had equal time to evolve.
  - Mosses are not more primitive, nor ancestral to other plants
  - They share a common ancestor



Phylogenetics

Building a tree

- Reconstruct a phylogeny to form a hypothesis about how isolates are related
  - Data can be physical characteristics (morphology), behavioural traits, or genetic data
    - DNA, protein sequences
      - Whole genome?
      - SNPs?
      - Sets of genes?

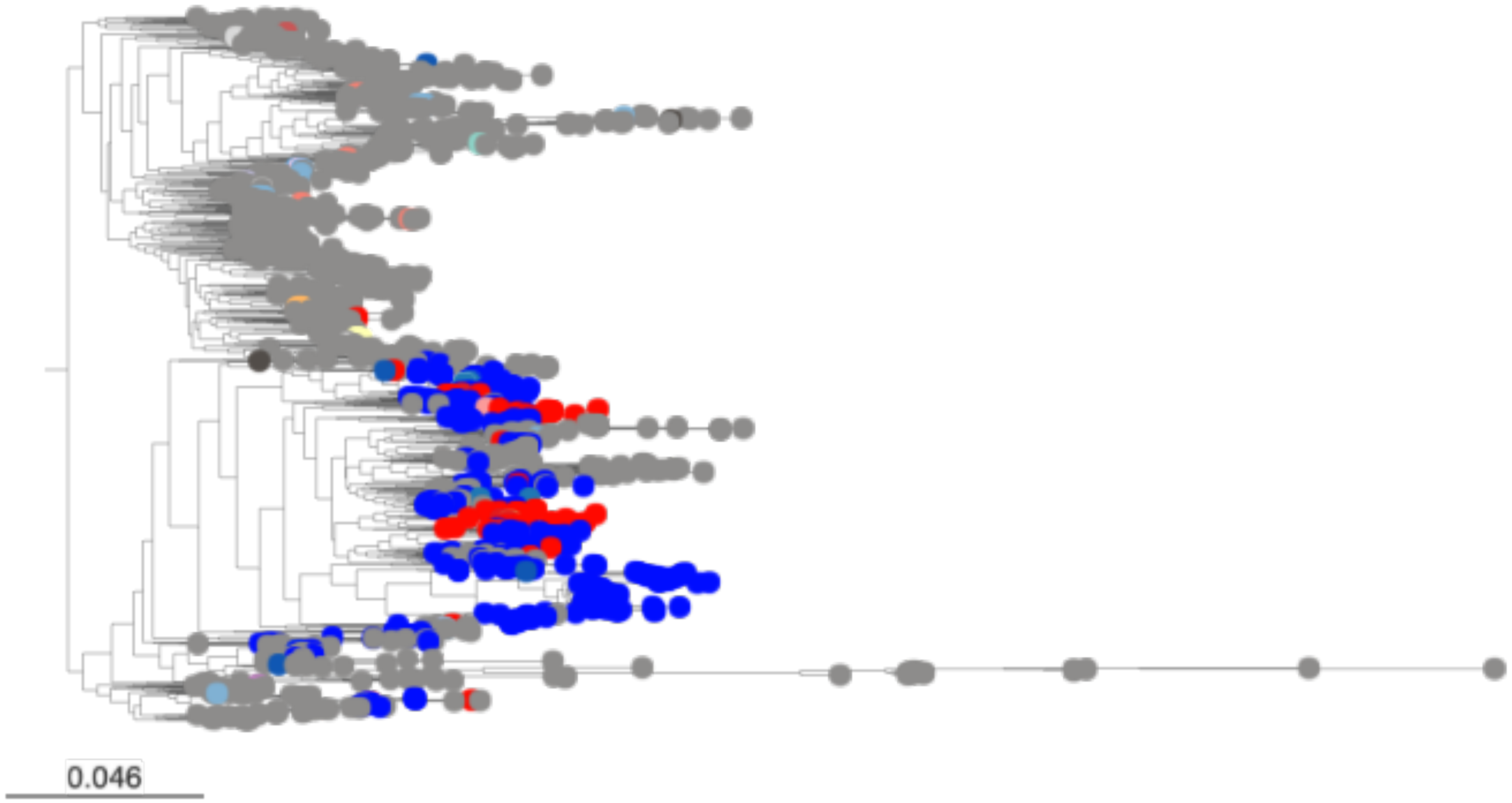


Zou, Y., *et al.* *Bioengineering* **11(5)**; 480 (2024)

Phylogenetics

Considerations for building a tree for fungi

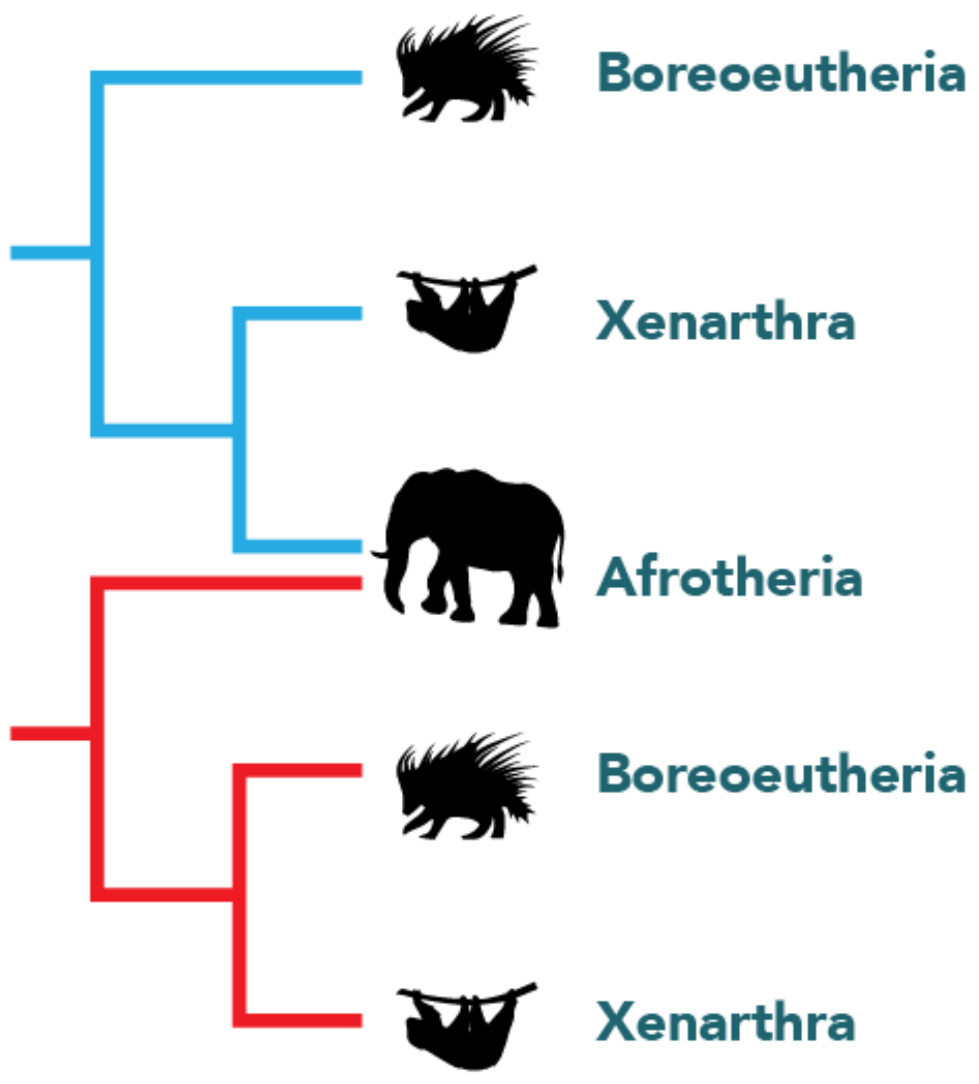
- Size of the genome
- Amount of recombination/ clonality
- Choice:
  - whole genome sequence
  - wgSNPs
  - all genes/subset of genes

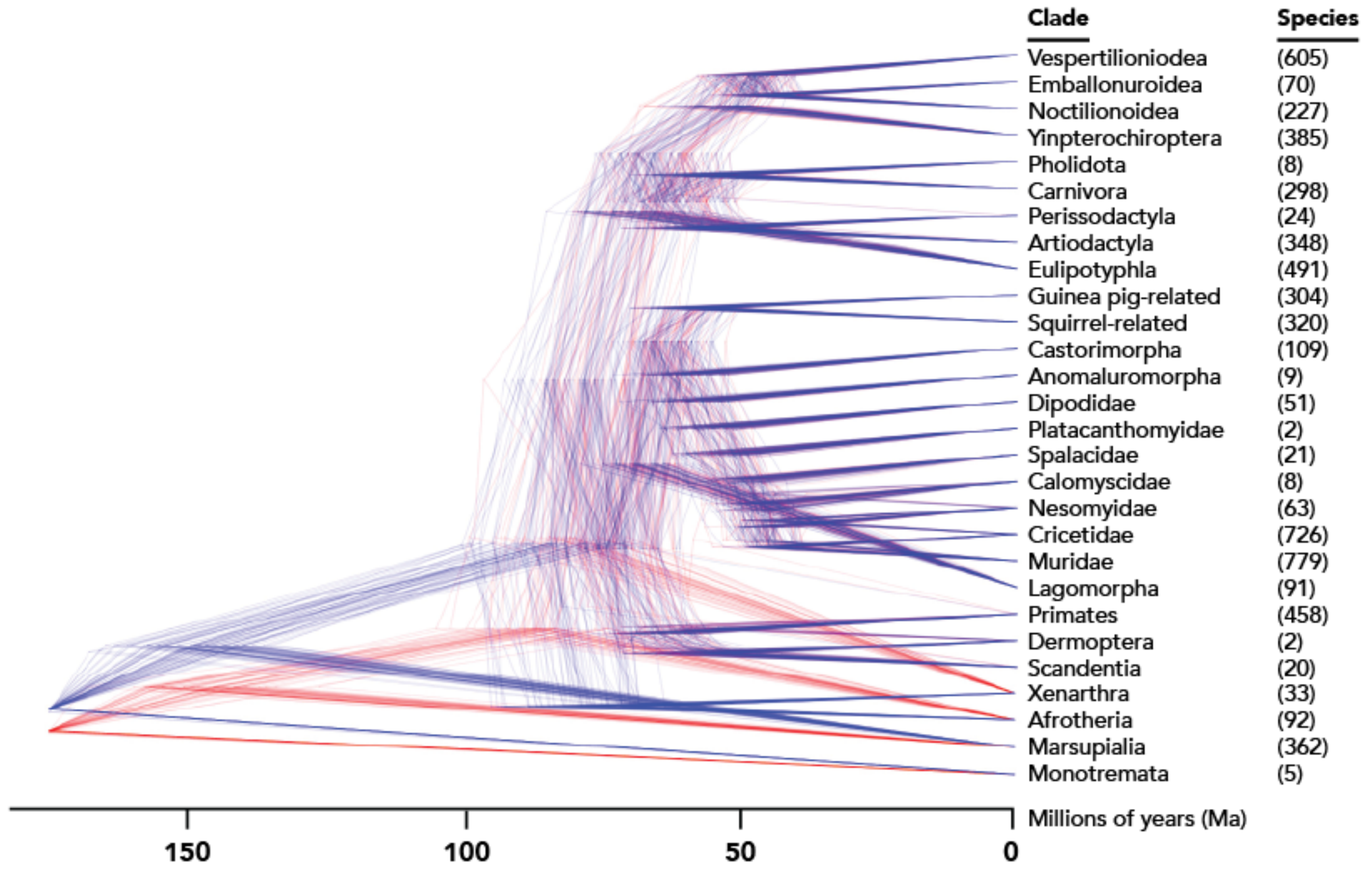


Building trees

Methods

- UPGMA
  - assumes equal rates of evolution throughout
- Neighbor-Joining (NJ)
  - generates sub-trees, and closest sub-trees are joined together in a step-wise manner
- Parsimony
  - grouping isolates in ways that minimise number of evolutionary changes
    - simple answer is often true (Occam's razor)
- Bayesian/likelihood based
  - produce lots of trees covering various hypotheses to produce a 'best' supported phylogeny





Building phylogenies with wgSNPs

RAxML

- Maximum Likelihood (Bayesian) approach
  - enables bootstrapping to get support for branches
- Requires a PHYLIP or FASTA input
- We can combine multiple vcf files (filtered snps/‘LowConf’ labelled) in to a single FASTA file and construct a phylogeny on multiple isolates
  - First, all vcf files need to be in the same directory (using the command ‘cp’ or ‘mv’)

Building phylogenies with wgSNPs

How to handle low confidence SNPs?

- Do not remove them
    - Consider them as ‘missing’ instead
    - If that position is low confidence because it has low mapping quality or coverage, it may be we don’t have enough evidence for or against it being a real SNP
    - Change low confidence positions to ’N’, which will be read by phylogeny software as ‘missing’
- Index all vcf files and merge into one file using vcftools
- Convert into a multiFASTA file using bcftools

Building phylogenies with wgSNPs

RAxML

- Two models of approximation: CAT and GAMMA
  - Do not use CAT if you have less than 50-100 taxa in your input file - use GAMMA instead
  - CAT is better than GAMMA if using protein sequence as it accommodates rate heterogeneity
- Bootstrap over 100 replicates
- Can have BIN or GTR
  - BIN is for binary (presence/absence)
  - GTR = ‘General Time Reversible’ model of nucleotide substitution under Gamma model of rate heterogeneity, which reduced computational burden
  - -p random seed
  - -x rapid bootstrap random seed
  - -f a = bootstrap
  - -N = number of bootstrap iterations
- `raxml -s file.fa -m GTRGAMMA -p 12345 -f a -x 12345 -N 100 -n trichophyton -w /path/to/output/directory`

Building phylogenies with wgSNPs

RAxML output files

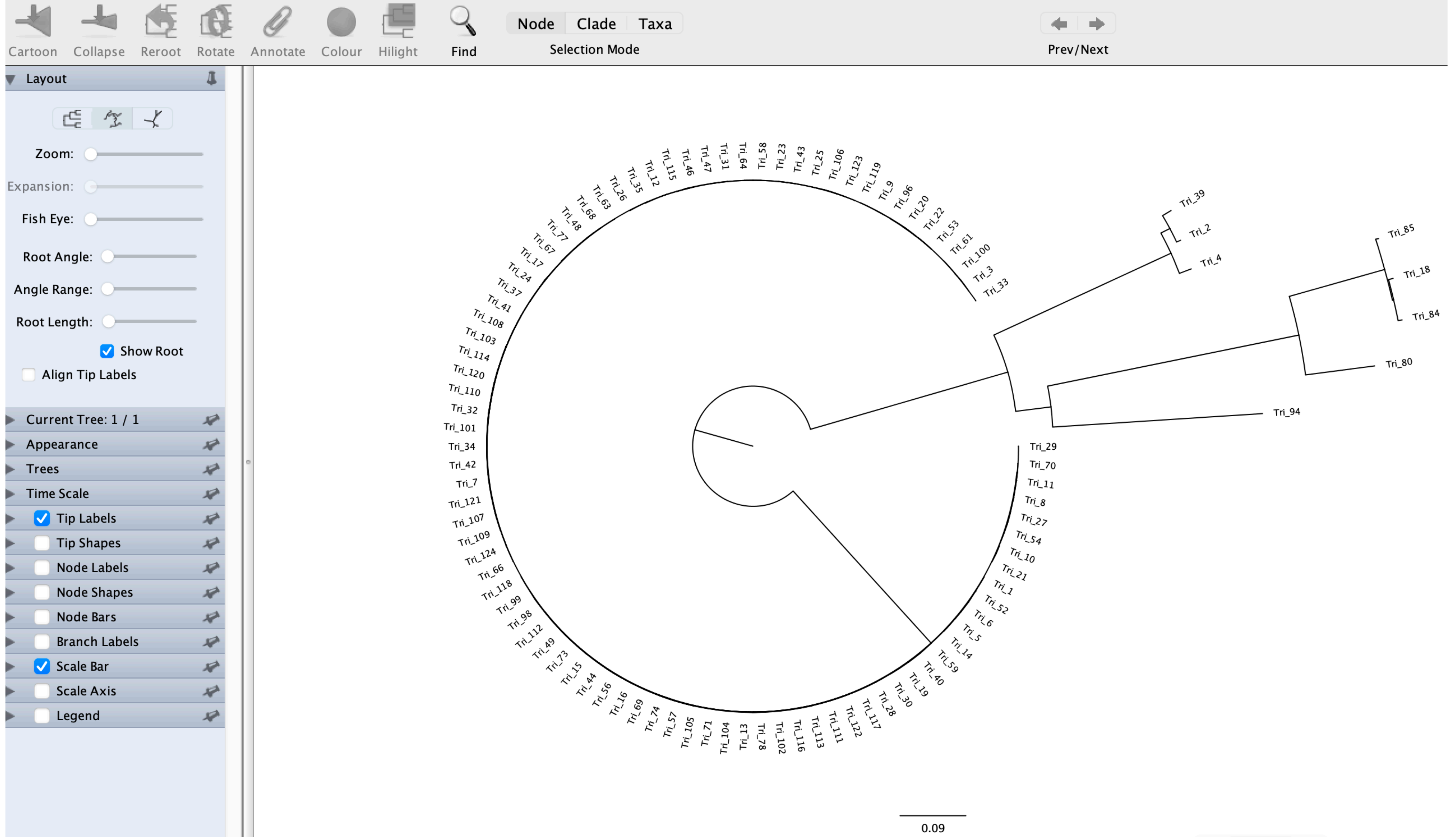
1. Best-scoring ML tree
2. Best-scoring ML tree with support values (bipartitions)
3. Best-scoring ML tree with support values as branch labels (bipartitionsBranchLabels)

- Depends on your analysis requirements which one you use, and which software you use to visualise
    - e.g. FigTree does not understand the bipartitionsBranchLabels format but will understand the bipartitions file

Visualising phylogenies

FigTree

- Free to download (https://github.com/rambaut/figtree/) and use
- Available for all OS
- Quite basic



☑ Show Root

□ Align Tip Labels

☑ Tip Labels

□ Tip Shapes

□ Node Labels

□ Node Shapes

□ Node Bars

□ Branch Labels

☑ Scale Bar

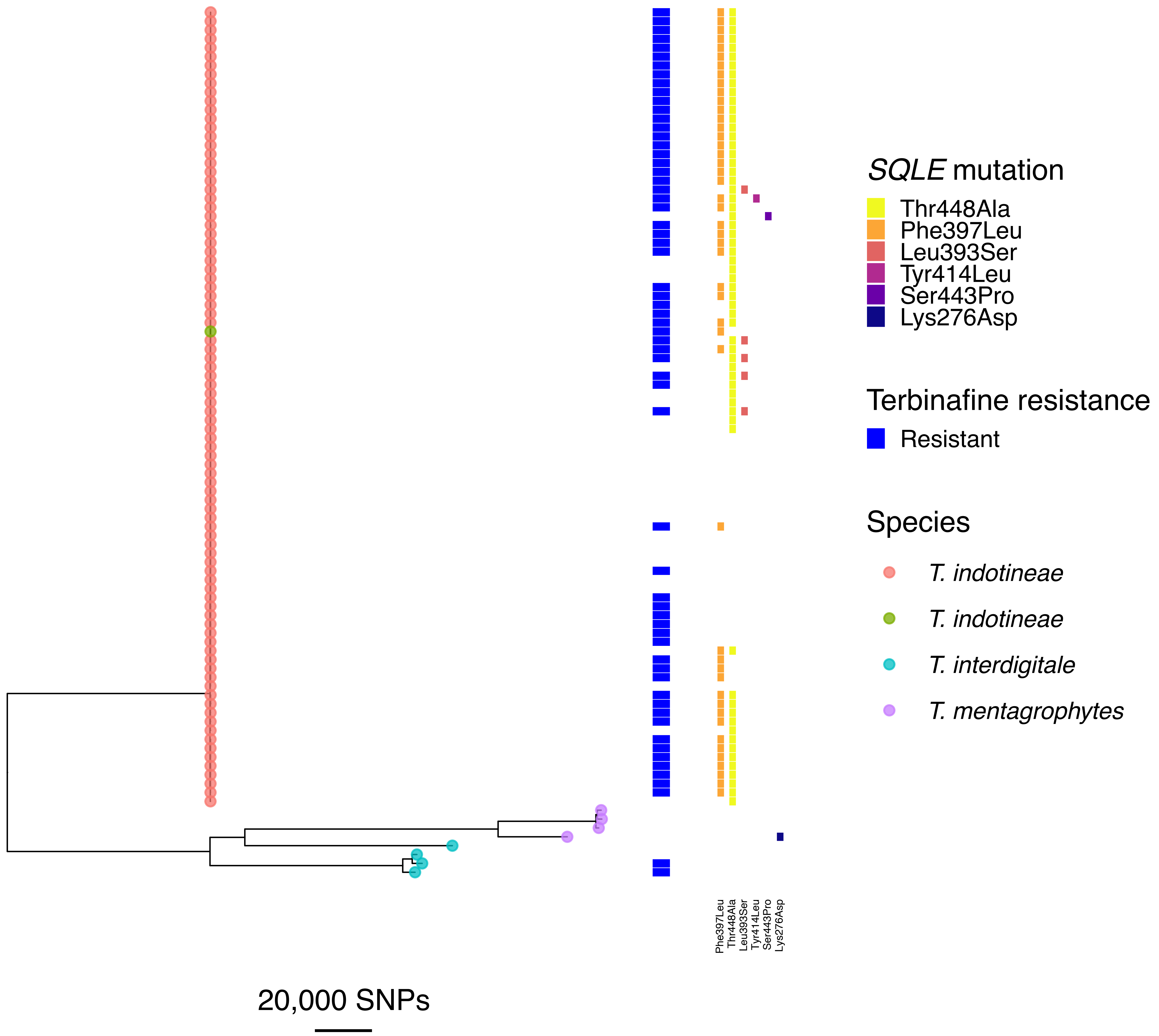
□ Scale Axis

□ Legend

Visualising phylogenies

ggTree

- Derived from ggplot2 as an R graphics package
- https://yulab-smu.top/treedata-book/


SQLE mutation
Thr448Ala
Phe397Leu
Leu393Ser
Tyr414Leu
Ser443Pro
Lys276Asp
Terbinafine resistance
Resistant
Species
T. indotineae
T. indotineae
T. interdigitale
T. mentagrophytes
Phe397Leu
Thr448Ala
Leu393Ser
Tyr414Leu
Ser443Pro
Lys276Asp
20,000 SNPs