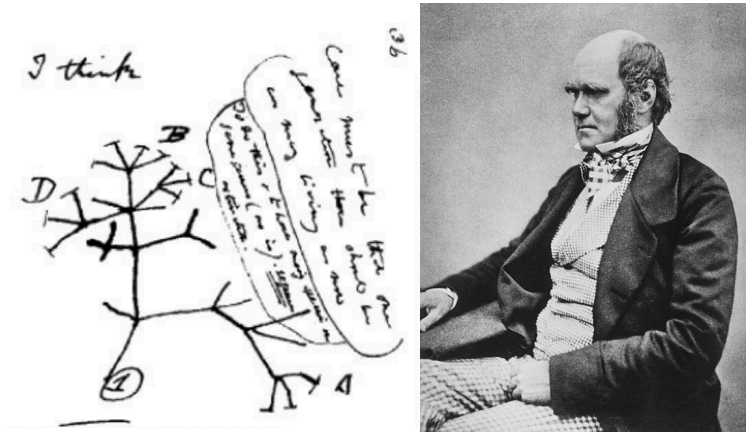
Module Phylogenetics and Phylogenomic

Helminth Bioinformatics - Latin America & the Caribbean

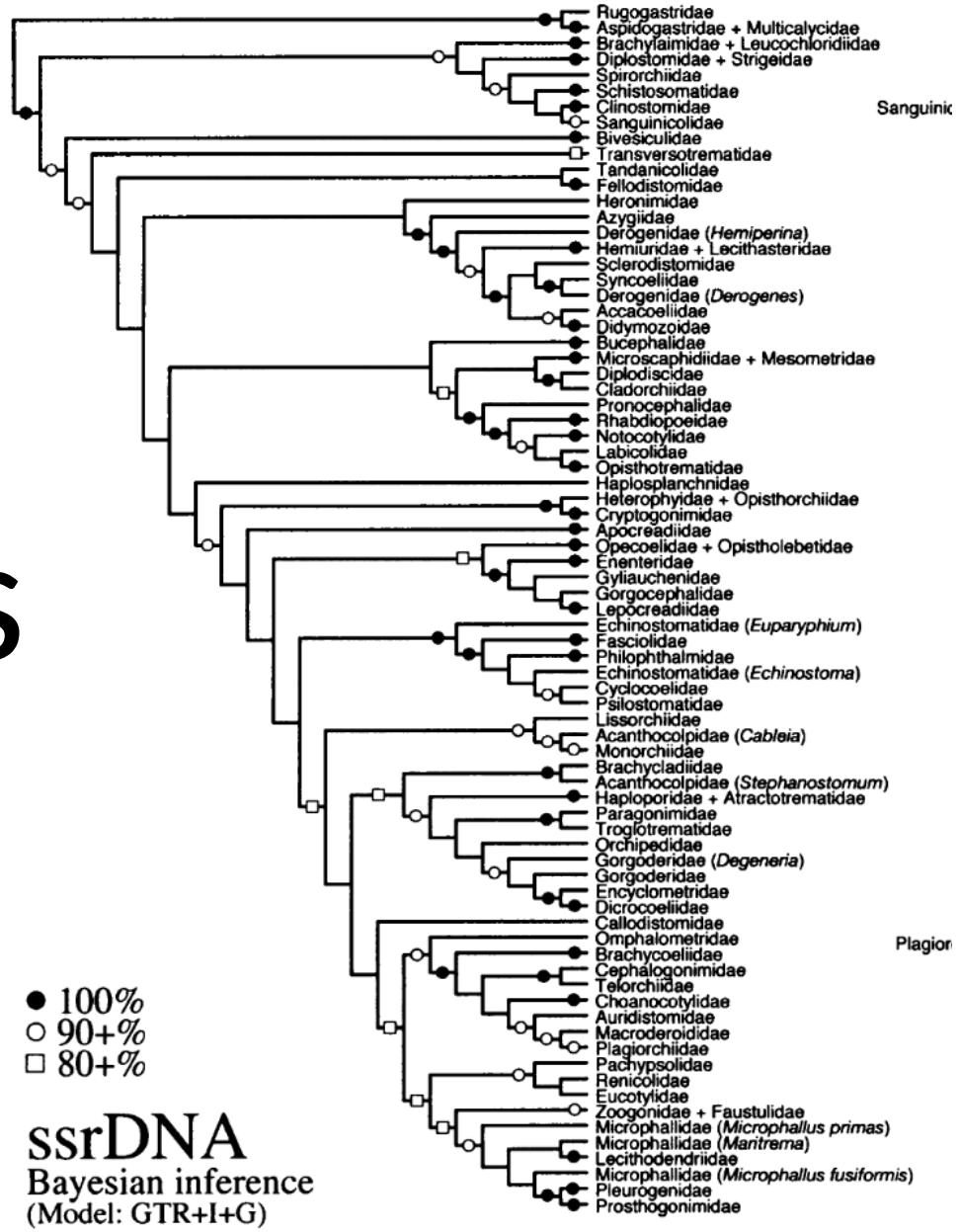
18th – 24th May 2025

Andrés Iriarte

Adapted from a presentation of: Daryl Domman, Marcela Suarez, and Sushmita Sridhar



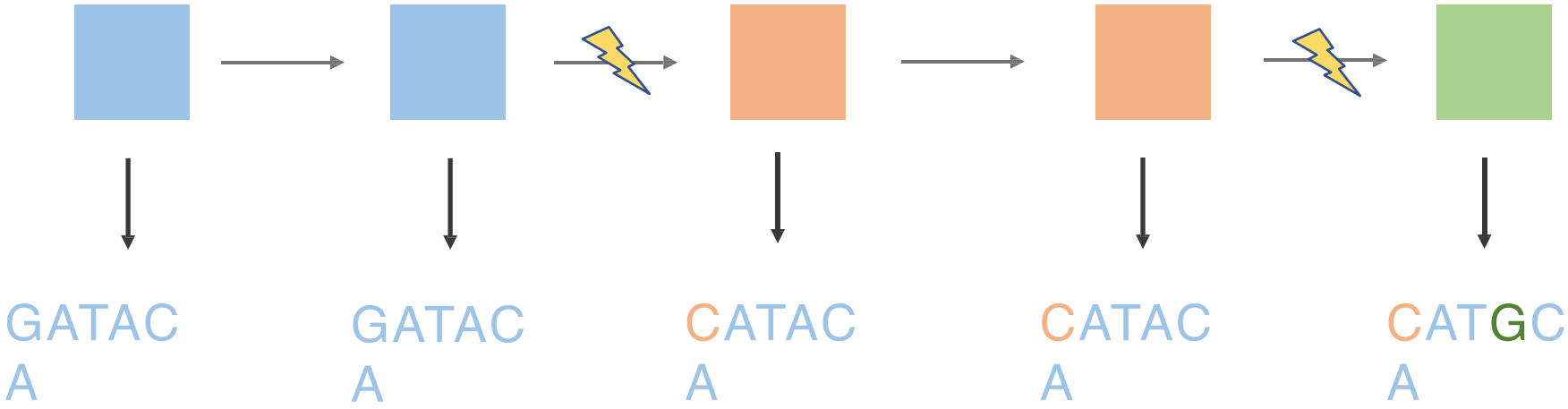
Phylogenetics



Modified from: Fig. 1. Olson et al. 2003
Int. Journal of Parasitology.

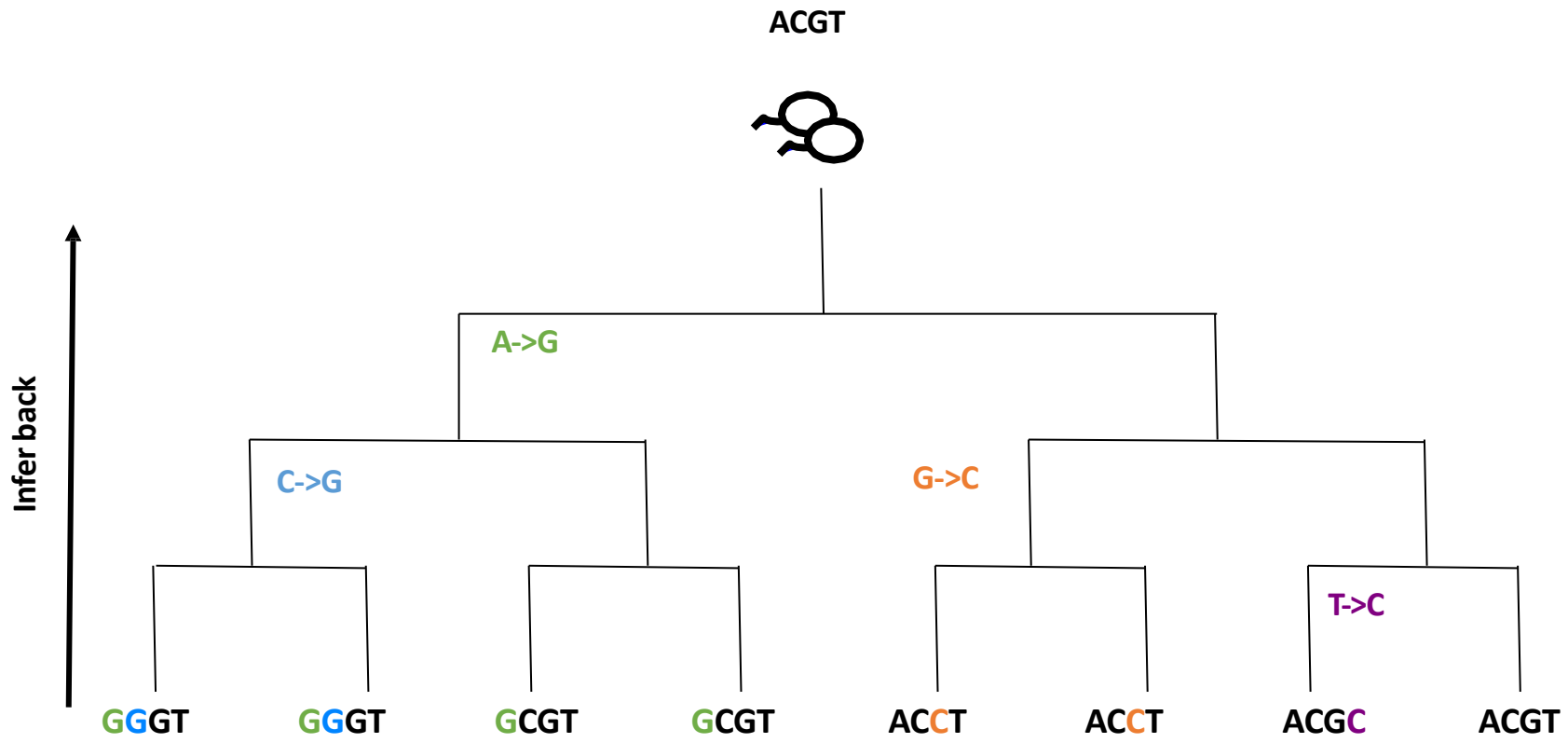
wellcome
connecting
science

Organisms acquire mutations



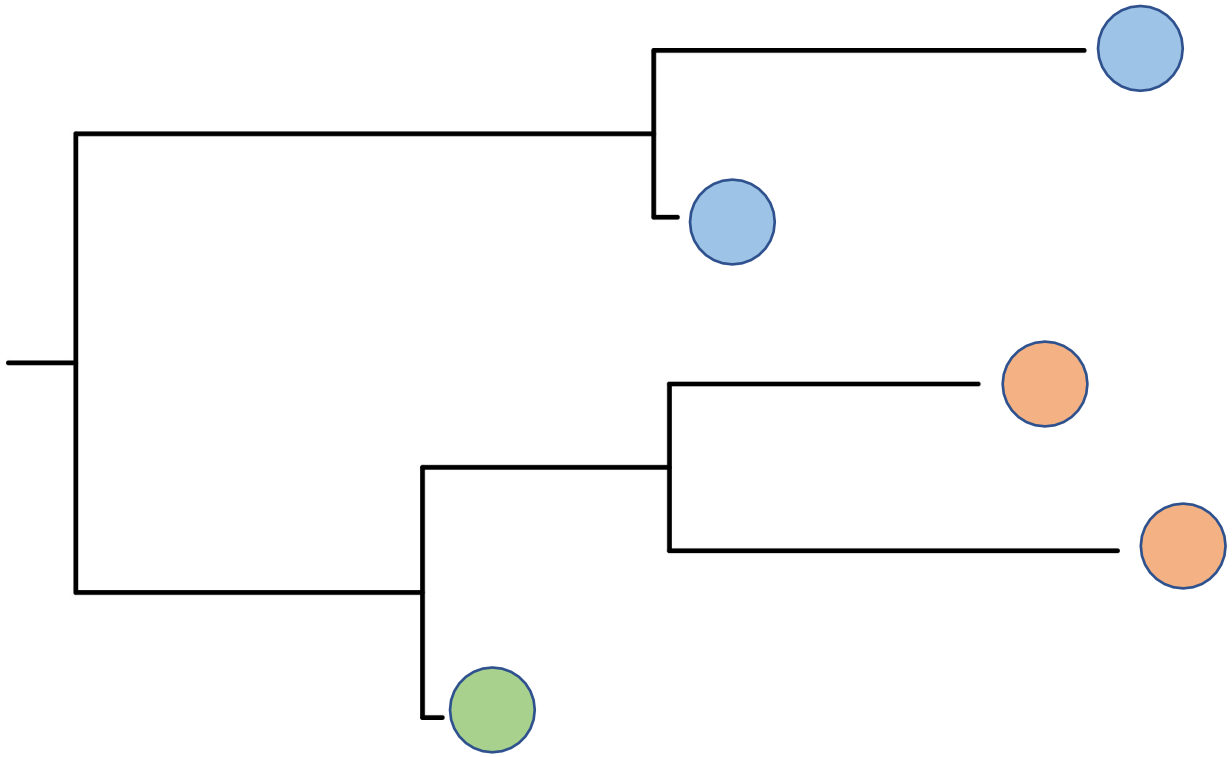
wellcome connecting science

Mutations tell us about relationships



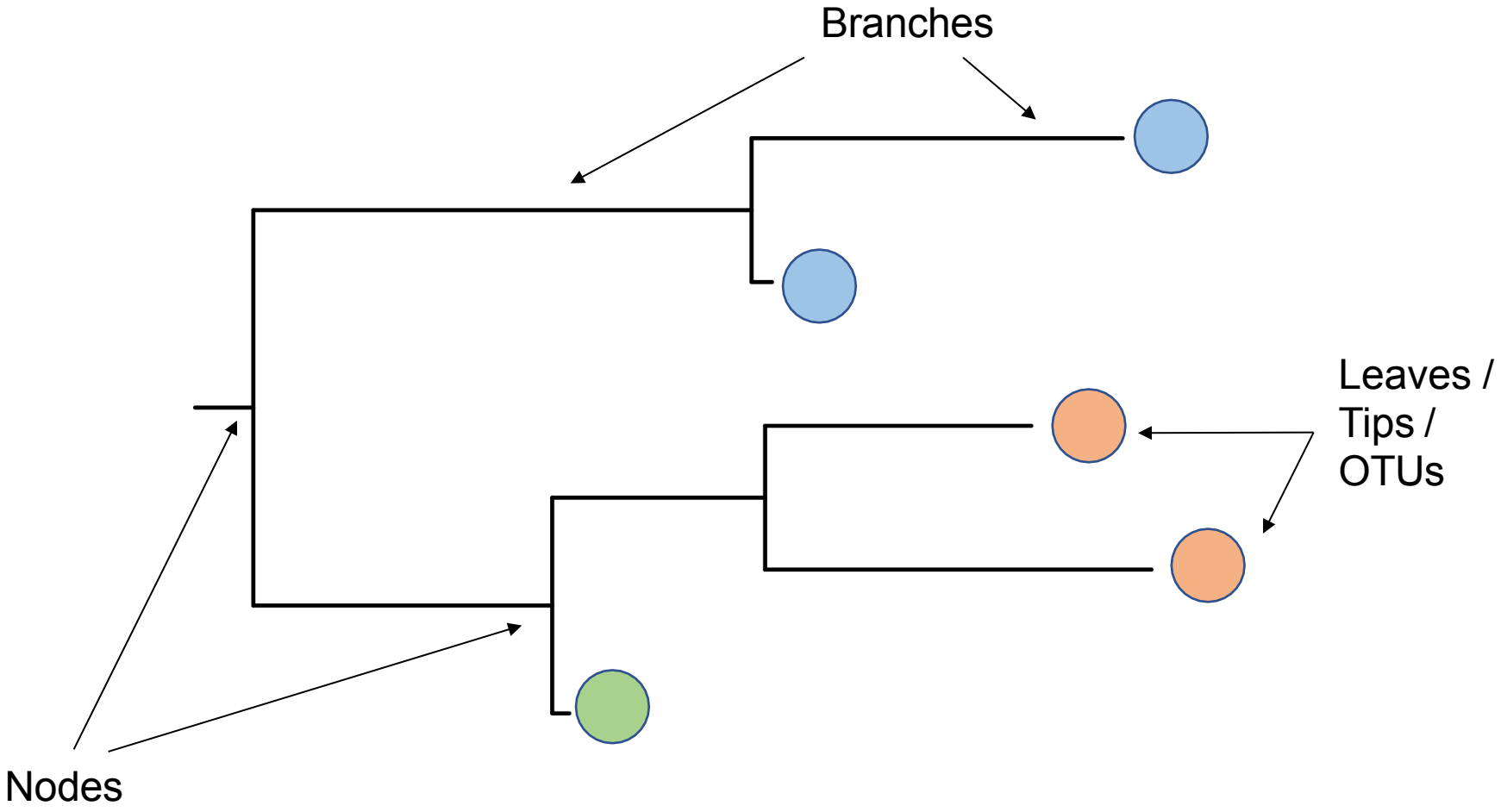
wellcome connecting science

Phylogenetic trees reveal relationships



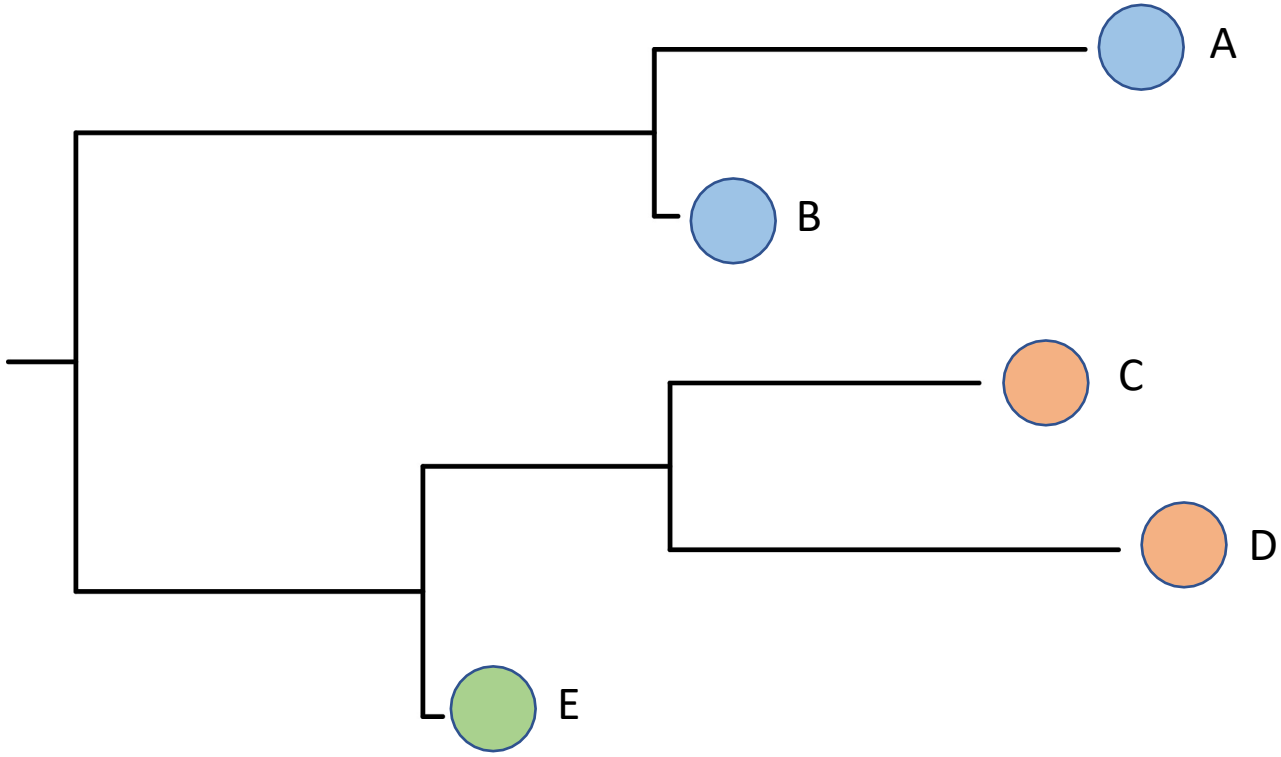
Genetic similarity

Phylogenetic trees



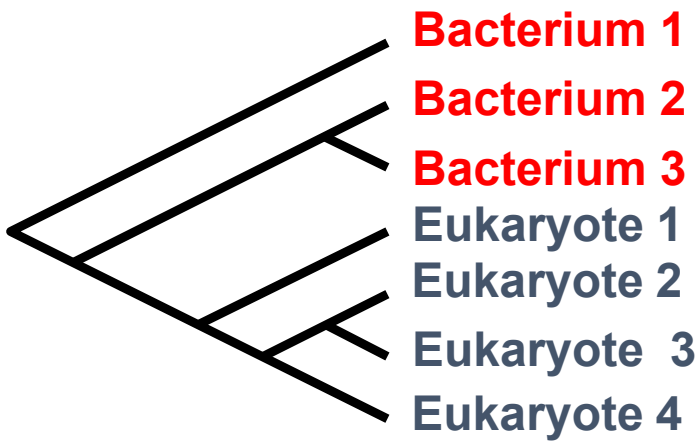
wellcome
connecting
science

Which taxa are the most distantly related?

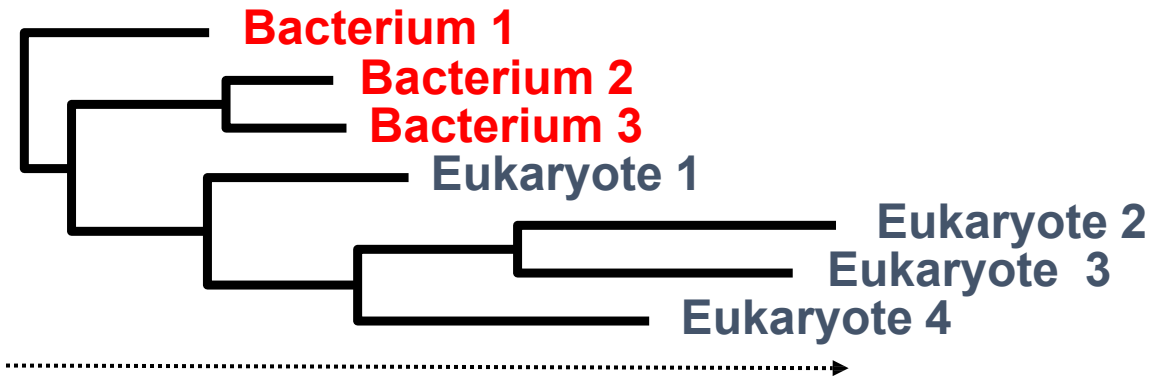


wellcome connecting science

Cladograms vs Phylograms



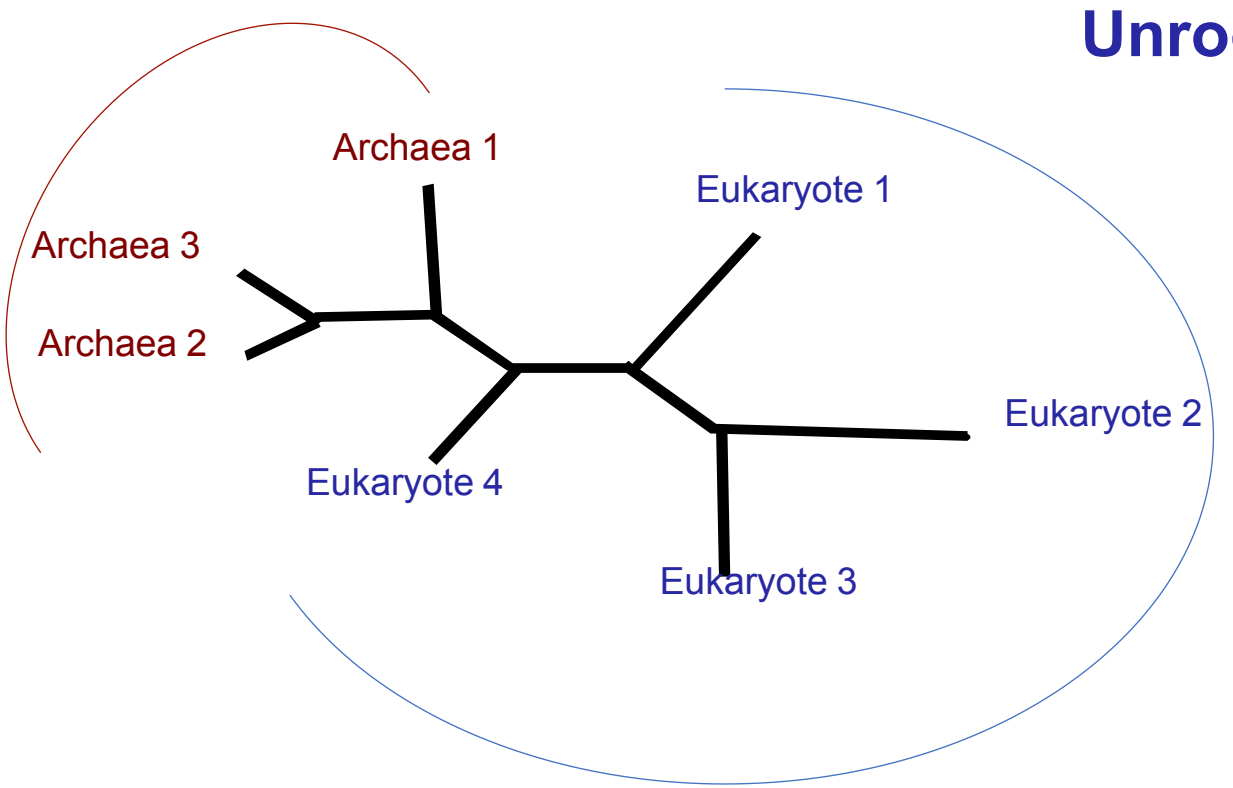
Cladograms show branch order (topology) only - branch lengths are meaningless



Phylograms show branch order and branch lengths with scale

wellcome connecting science

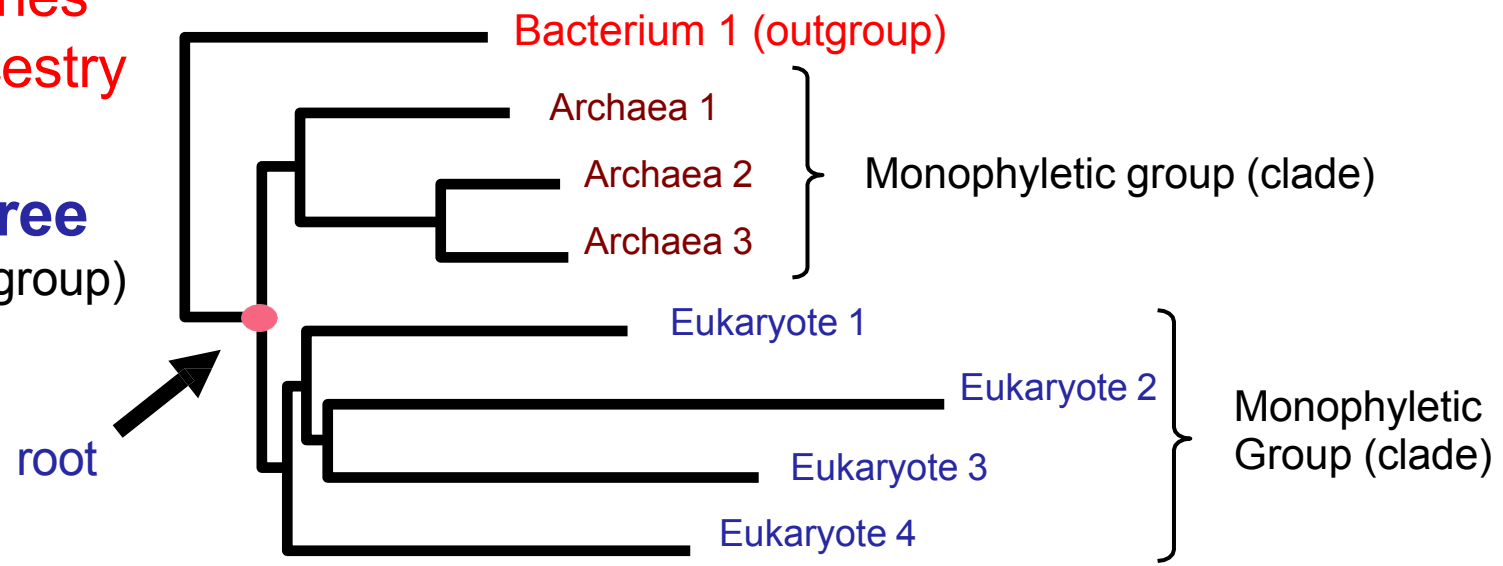
Rooted and Unrooted trees



Rooted and Unrooted trees

The root defines common ancestry

Rooted tree
(by using outgroup)



Where to root a tree?

Midpoint or Outgroup

Best to check what other people in the field are doing and define outgroup

Include published references in phylogeny, choose midpoint root and check to see where the published sequences cluster

If in doubt start with midpoint root and work from there

Building a Phylogenetic Tree

1. Identify protein, DNA or RNA sequences of interest
   (Format file, fasta, nexus, phylip…)

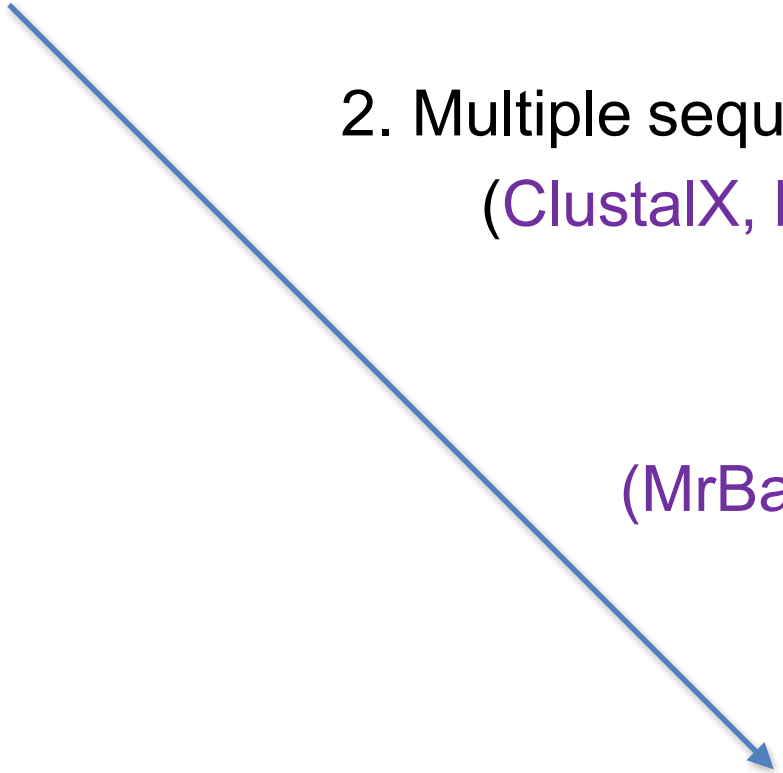
2. Multiple sequence alignment
   (ClustalX, Muscle, Mafft…)

3. Construct phylogeny
   (MrBayes, RAxML, IQ-Tree2, FastTree…)

4. View and edit tree
   (FigTree, MEGAX, tree view…)

Remember:

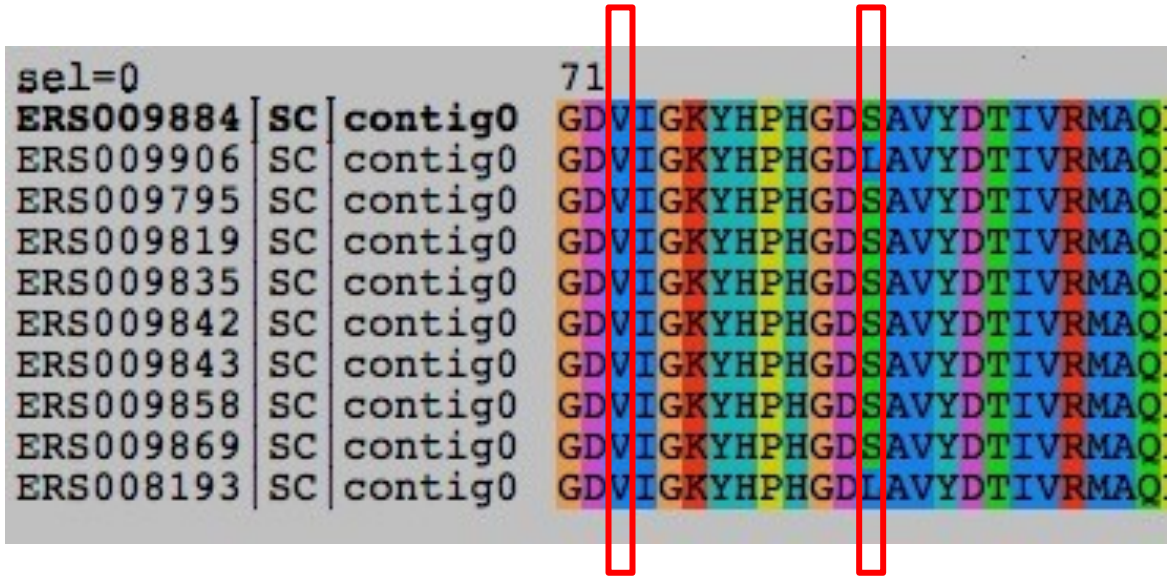
Crap in == Crap out!



wellcome connecting science

Multiple sequence alignment (MSA)

MSA is best hypothesis of **positional homology**
between bases/amino acids of different sequences



wellcome connecting science

MSA - can be easy but also tricky

```
GCGGCCCA  TCAGGTAGTT  GGTGG
GCGGCCCA  TCAGGTAGTT  GGTGG
GCGTTCCA  TCAGCTGGTT  GGTGG
GCGTCCCA  TCAGCTAGTT  GGTGG
GCGGCGCA  TTAGCTAGTT  GGTGA
********  **********  *****
```

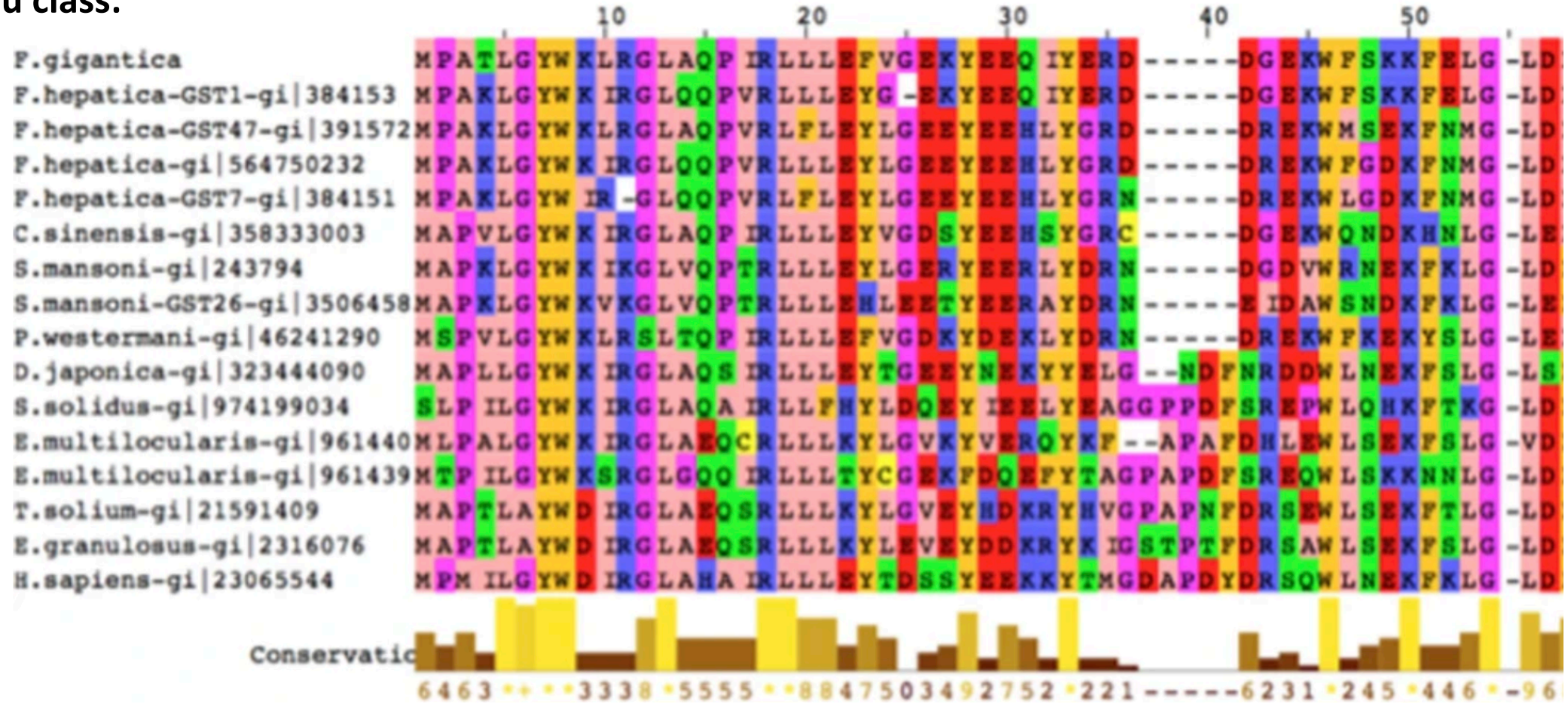
Easy

```
TTGACATG  CCGGGG---A  AACCG
TTGACATG  CCGGTG--GT  AAGCC

TTGACATG  -CTAGG---A  ACGCG
TTGACATG  -CTAGGGAAC  ACGCG

TTGACATC  -CTCTG---A  ACGCG
********  ??????????  *****
```

Tricky

MSA - can be easy but also **tricky**

GSTs mu class:



Modified from Kalita et al. 2017. Scientific Reports volume 7, Article number: 17547 (2017)

wellcome connecting science

Important note on MSAs

COMPUTERS DO NOT NECESSARILY KNOW BETTER

Check your alignments by eye

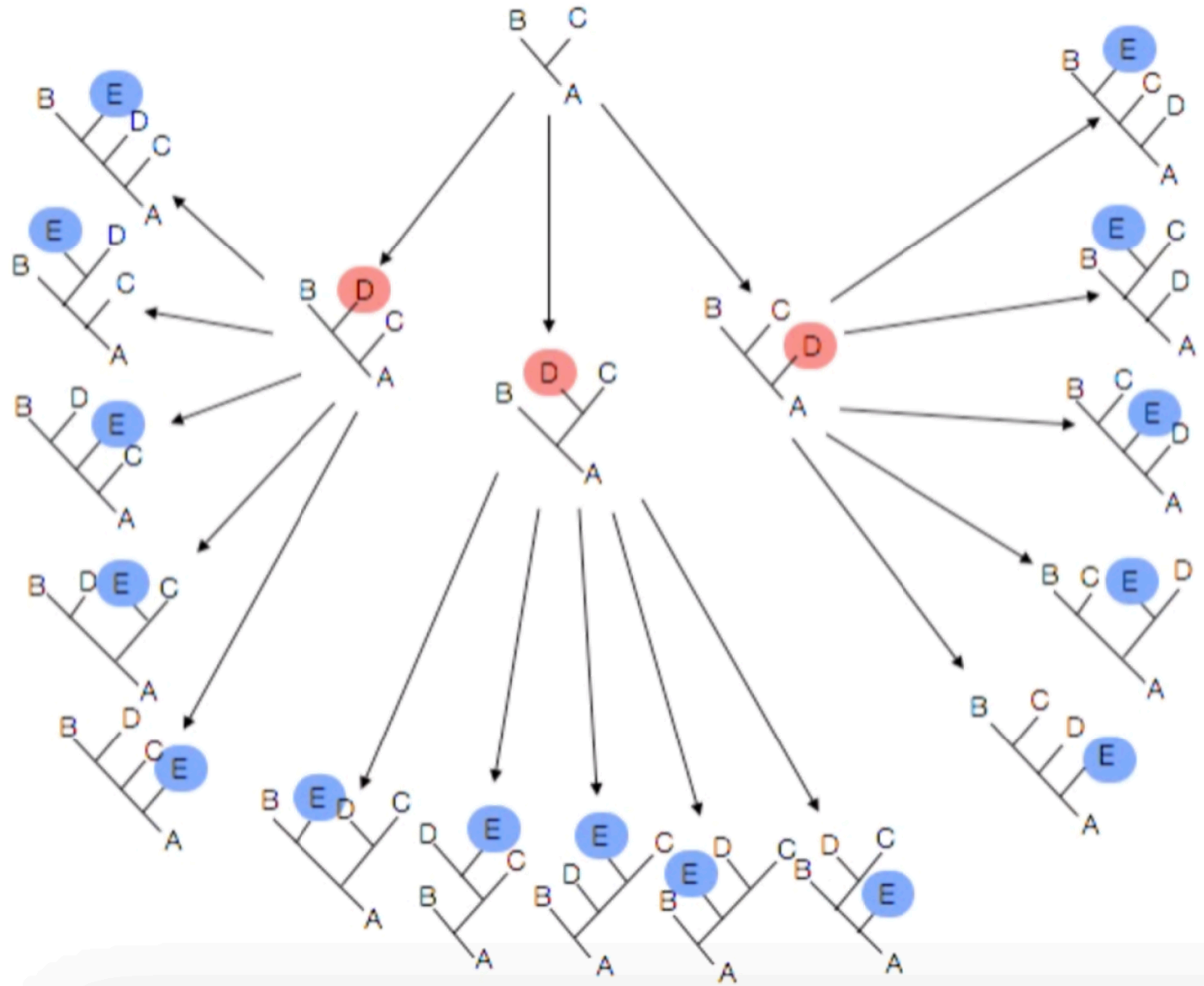
Can I remove erroneous sections?

Constructing a phylogenetic tree

Method	Data used	Tree search	Evolutionary Model
Distance	Pairwise distance	Simple algorithm	Can be complex
Parsimony	Informative sites	Mainly hill climbing	Simple
Maximum likelihood	All sites	Hill climbing	Can be complex
Bayesian Methods	All sites (+ other info)	MCMC	Can be very complex

Tree searching algorithms

Taxones: A B C D E



Possible number of trees for *n* taxa
$(2n - 3)!!$

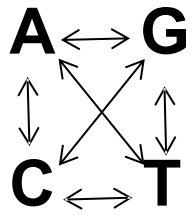
No. taxa	No. unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10395
80	2.18×10^{137}

Phylogenetic models:

Simple

JC69:	all substitutions equally likely, all bases equally frequent.
JC69+I+Γ:	as for JC69, but with additional parameters for invariant sites and gamma distribution.
K2P:	specific probabilities for transitions and transversions, all bases equally frequent.
HKY85:	specific probabilities for transitions and transversions, specific base frequencies.
GTR:	each substitution has a specific probability, moderated by specific base frequencies.
GTR+I+Γ:	as for GTR, but with additional parameters for invariant sites and gamma distribution.

Complex



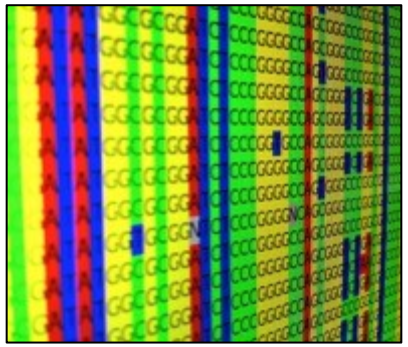
4 equilibrium base frequency parameters and 6 substitution rate parameters and

wellcome connecting science

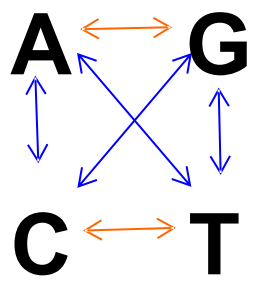
Putting it together

Maximum likelihood phylogenetic models maximize the probability of achieving …

these data…

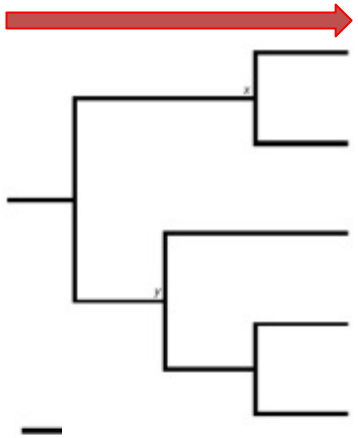


… if this happens…



+ composition

… over this tree



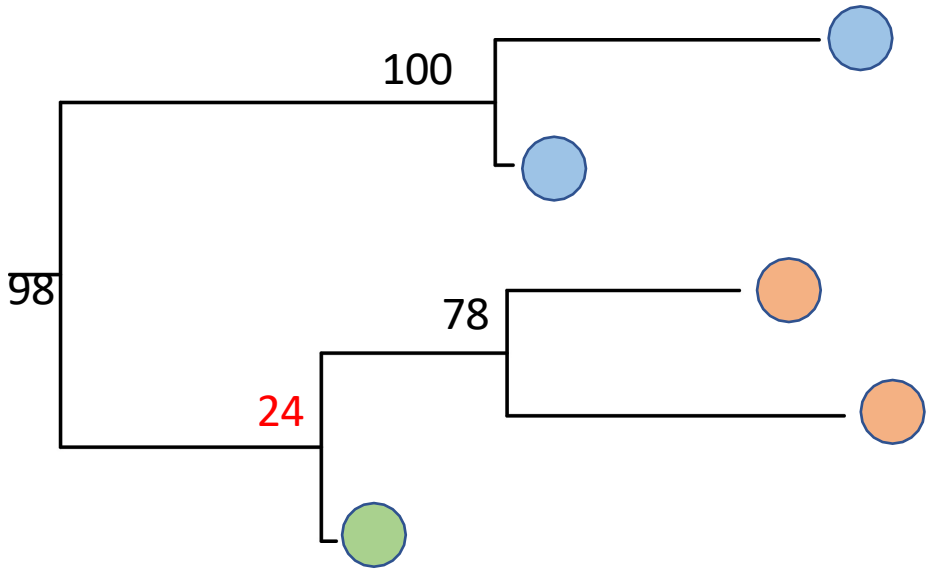
Gaining confidence: Bootstrapping

Bootstrapping is a way to produce a confidence measure in the topology relationships found in a phylogenetic analysis

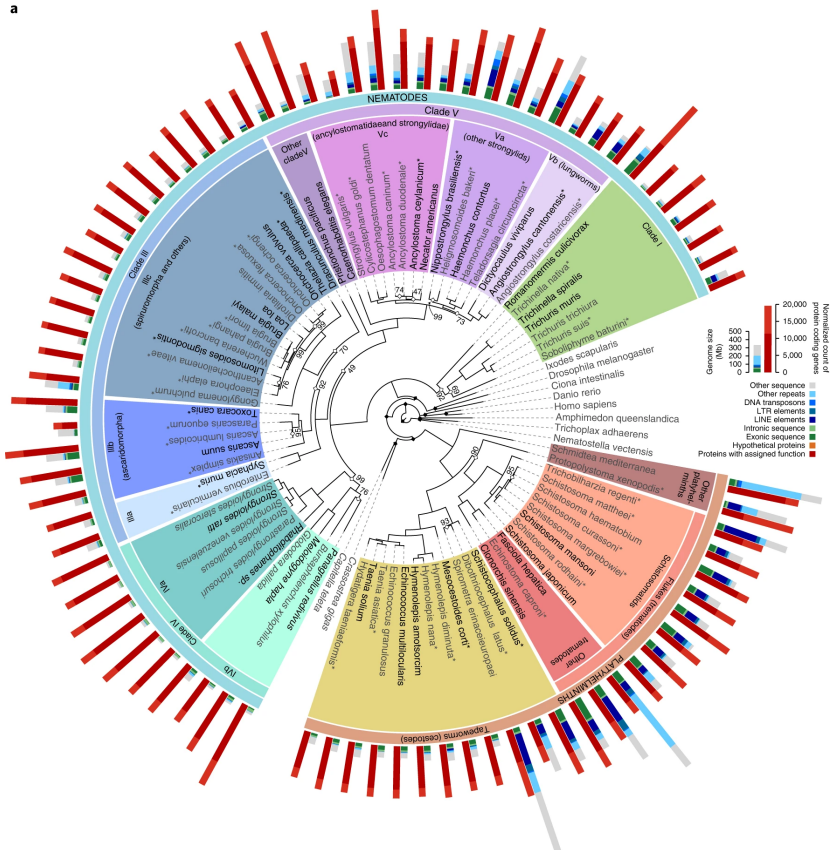
X number of **bootstraps** (resampled replicates) are created of your input data (MSA)

Typically run 100 – 1,000 bootstraps for ML analysis

These are commonly used as a measure of support for these branches and are represented as a number on each tree branch



Phylogenomics



Modified from: International Helminth Genomes Consortium, Nature Genetics volume 51, pages163–174 (2019).

wellcome connecting science

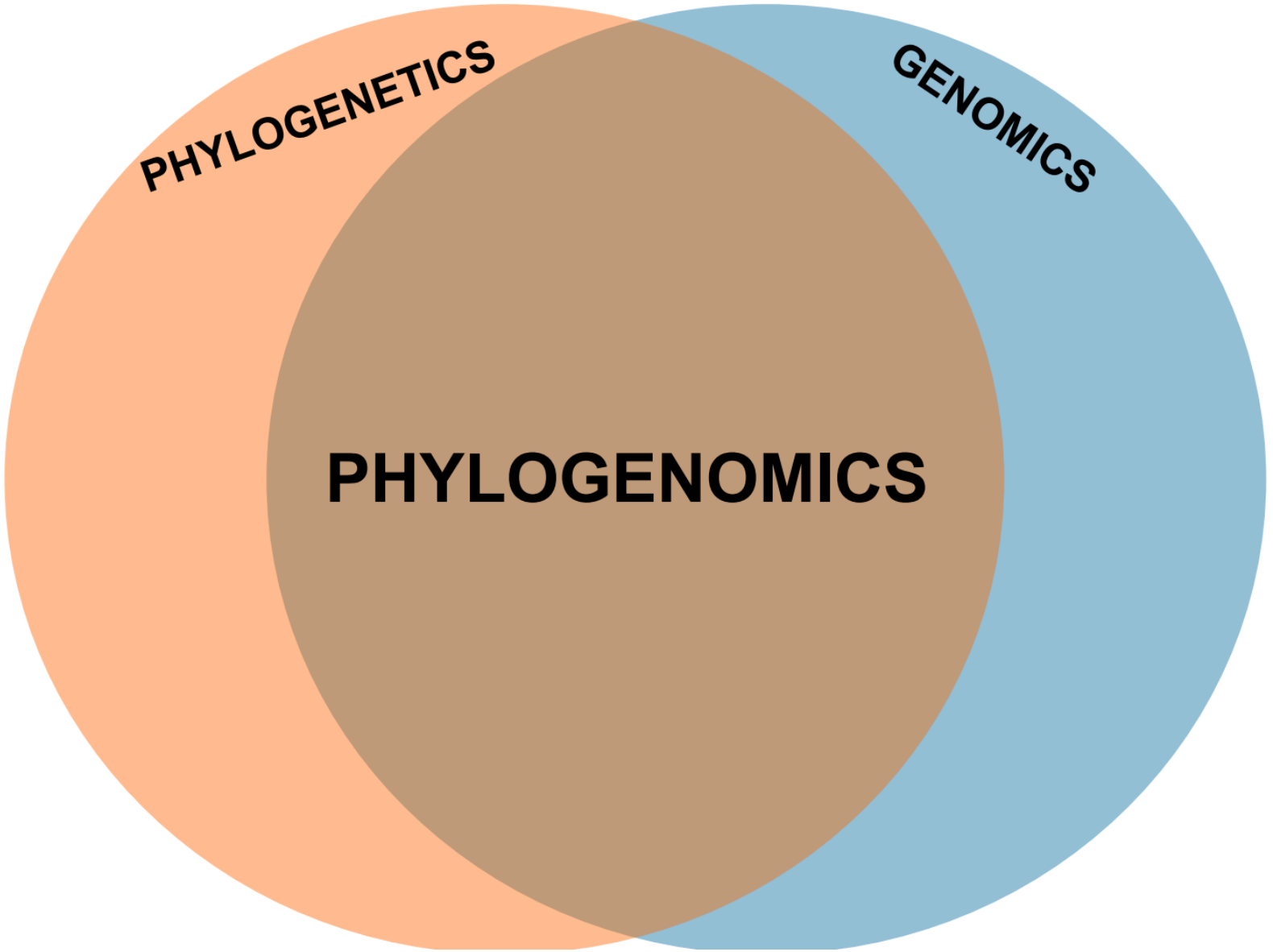
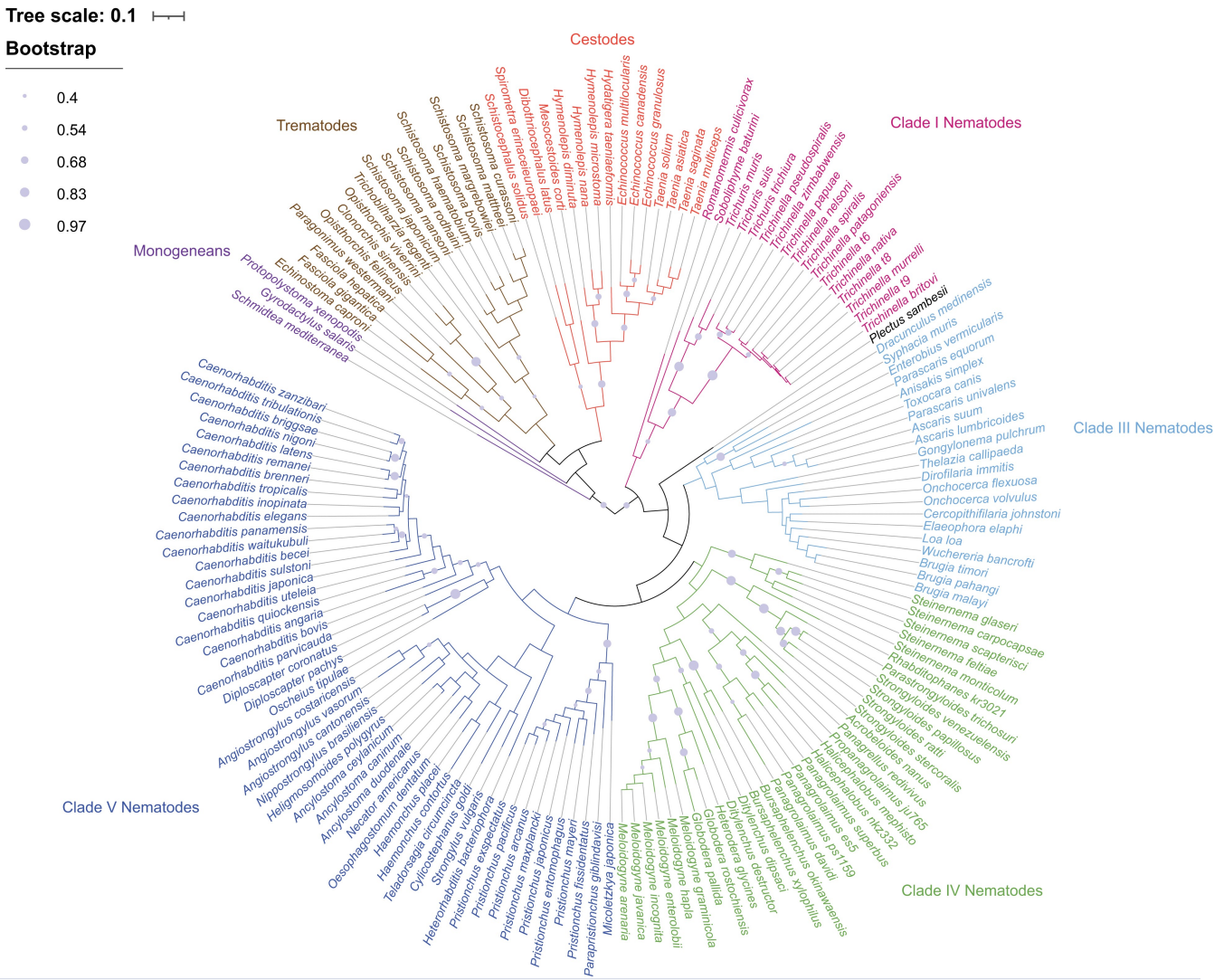
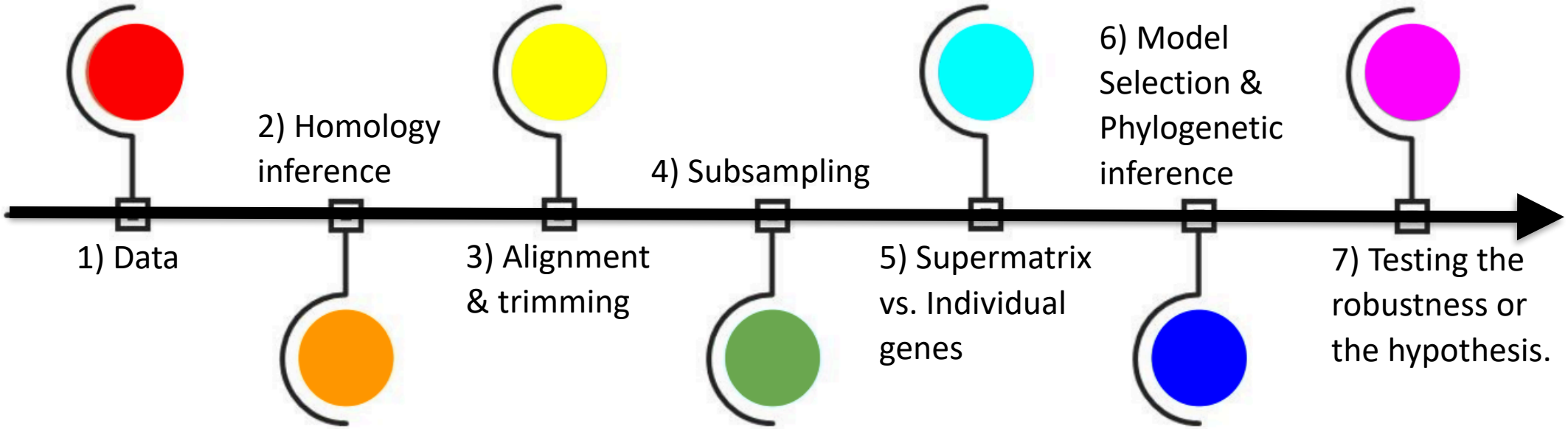

PHYLOGENETICS
GENOMICS
PHYLOGENOMICS

Fig. 1. Phylogenetic tree of 155 species of helminth. This tree is based on 339 orthogroups identified with ...

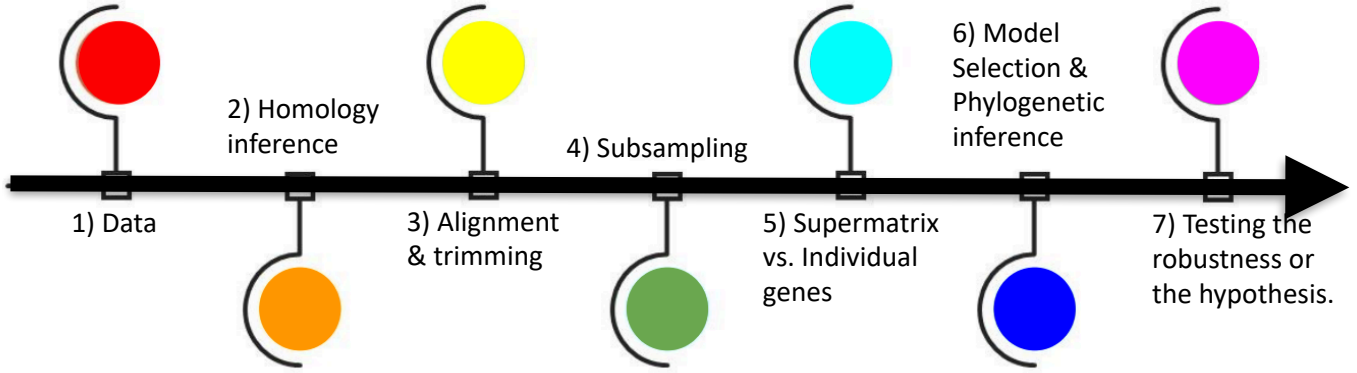
Reconstructing the evolutionary histories of organisms taking into account whole genomes or large fractions of genomes.



Taken from Collington et al. 2023. Genome Biol Evol, Volume 15, Issue 8, August 2023, evad135.


1) Data
2) Homology inference
3) Alignment & trimming
4) Subsampling
5) Supermatrix vs. Individual genes
6) Model Selection & Phylogenetic inference
7) Testing the robustness or the hypothesis.

Key Considerations in Phylogenomics:



Sampling Design

- What taxa are included? Outgroup vs. ingroup?
- Is the sampling aligned with the biological question?

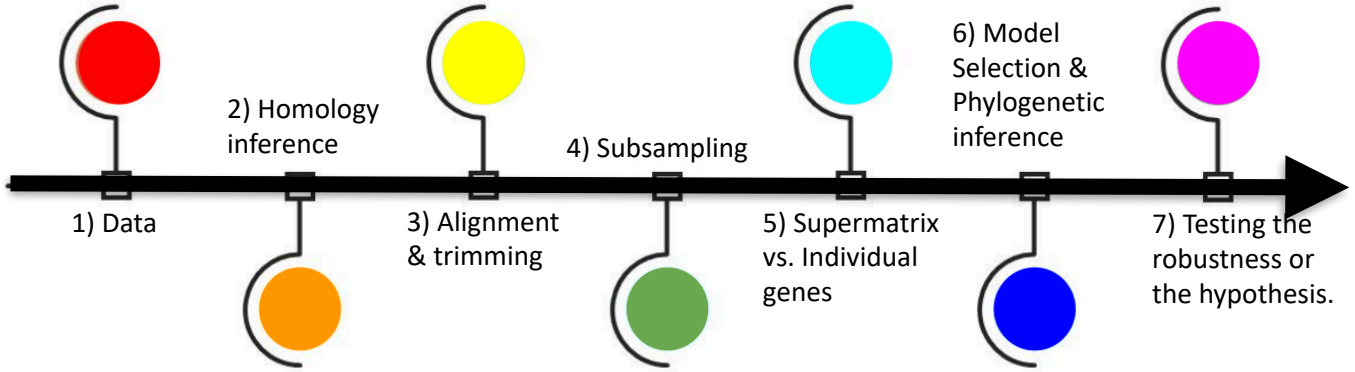
Homology Inference

- What thresholds are used for detecting homology?
- Strategy for distinguishing orthologs (1:1, many:1), paralogs, and general homologs.

Alignment & Trimming

- Should we evaluate the quality of alignments, incorporate structural or functional information?
- Comparison of different alignment methods/software.
- Manual curation vs. automated, objective trimming.

Key Considerations in Phylogenomics:



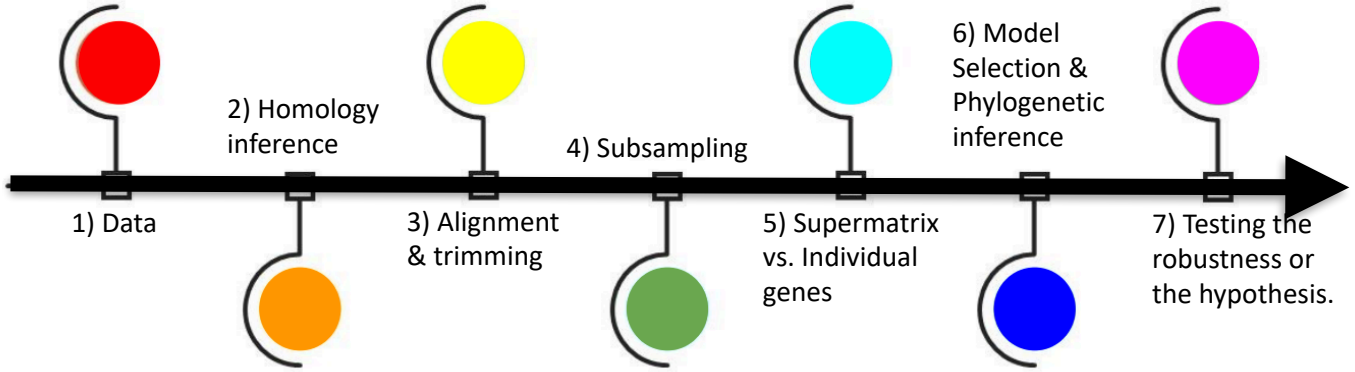
Subsampling

- Should we reduce the dataset (loci or taxa) to lower computational cost?
- Can we reduce noise while still answering the main question?

Supermatrix vs. Consensus Approaches

- Concatenate all loci (supermatrix) or analyze gene trees separately?
- Use of reconciliation or consensus methods.
- Do different approaches yield consistent phylogenetic signals?

Key Considerations in Phylogenomics:



Model Selection & Phylogenetic Inference

- Why is model selection important?
- Should multiple inference methods be compared?

Testing & Validation

- Why is it necessary to test phylogenetic hypotheses?
- What kinds of support values or statistical tests are available?


Questions?