



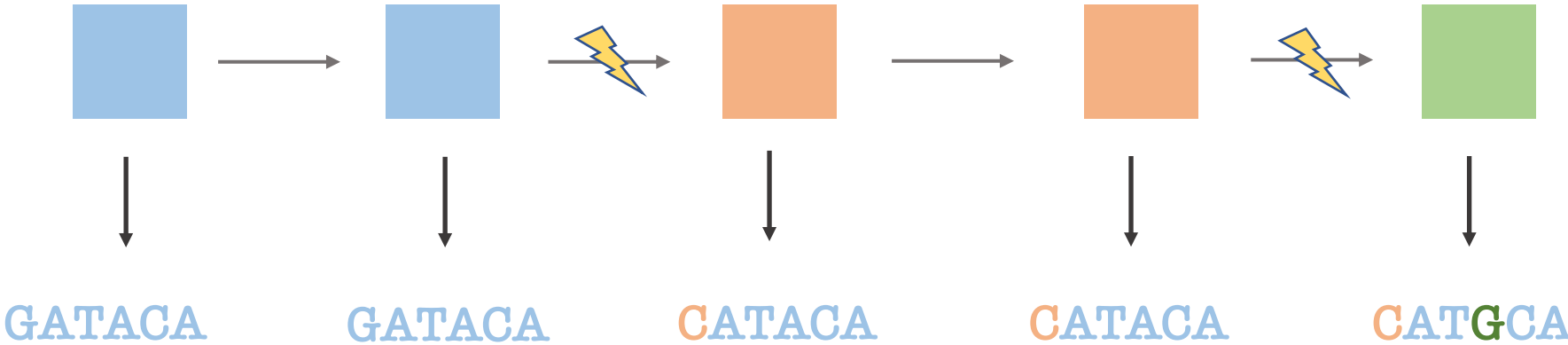
# Module 4 - Phylogenetics

Working with Pathogen Genomes

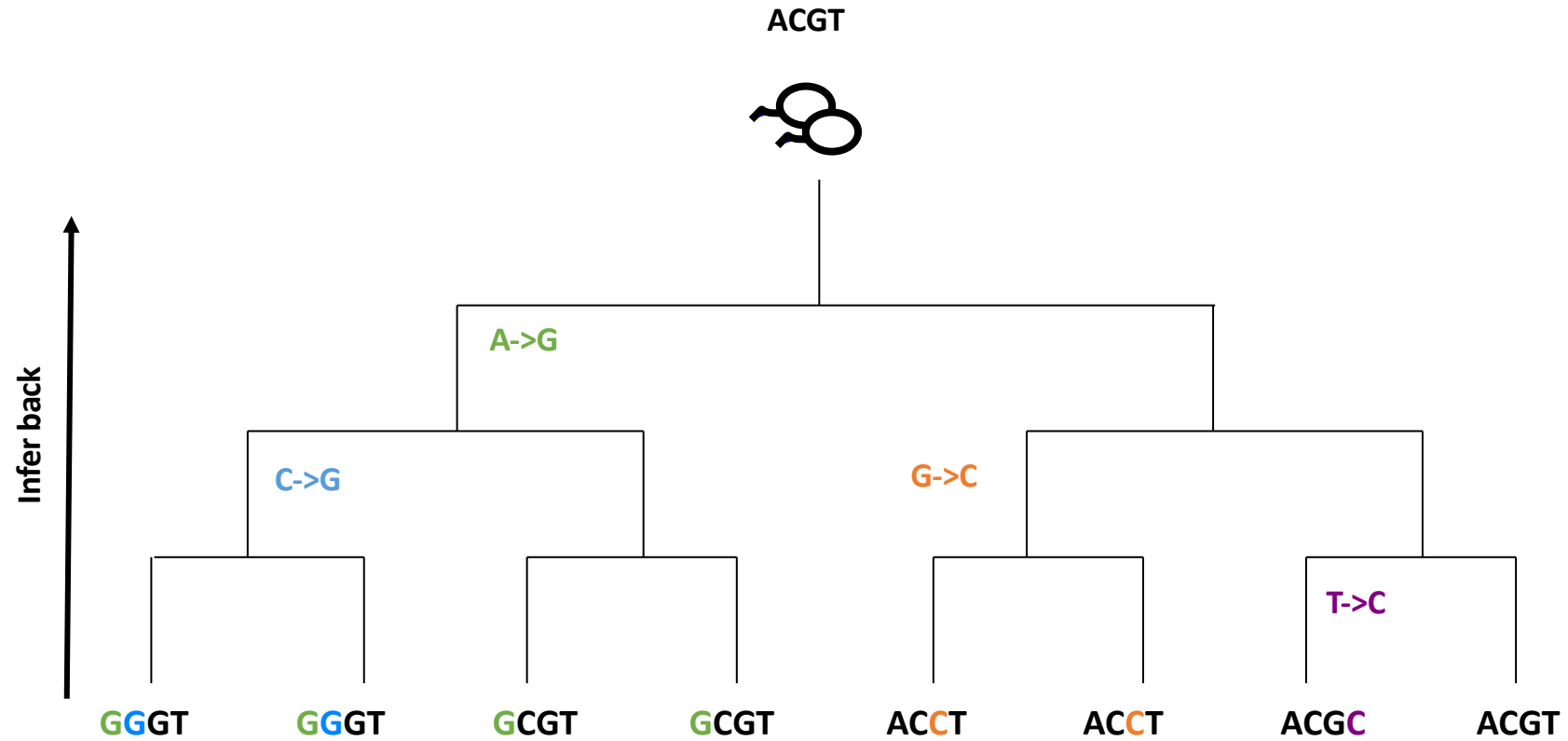
7<sup>th</sup> - 11<sup>th</sup> February 2022

Daryl Domman  
Marcela Suarez  
Sushmita Sridhar

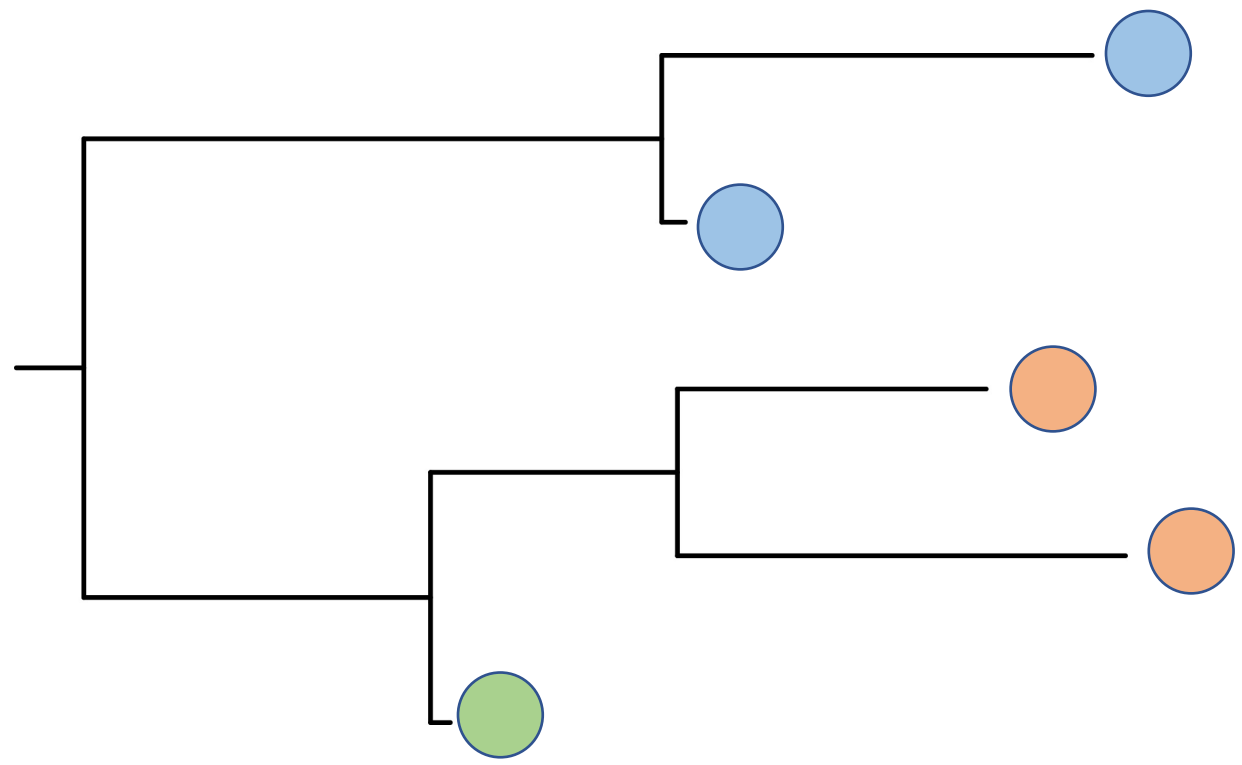
# Organisms acquire mutations



# Mutations tell us about relationships

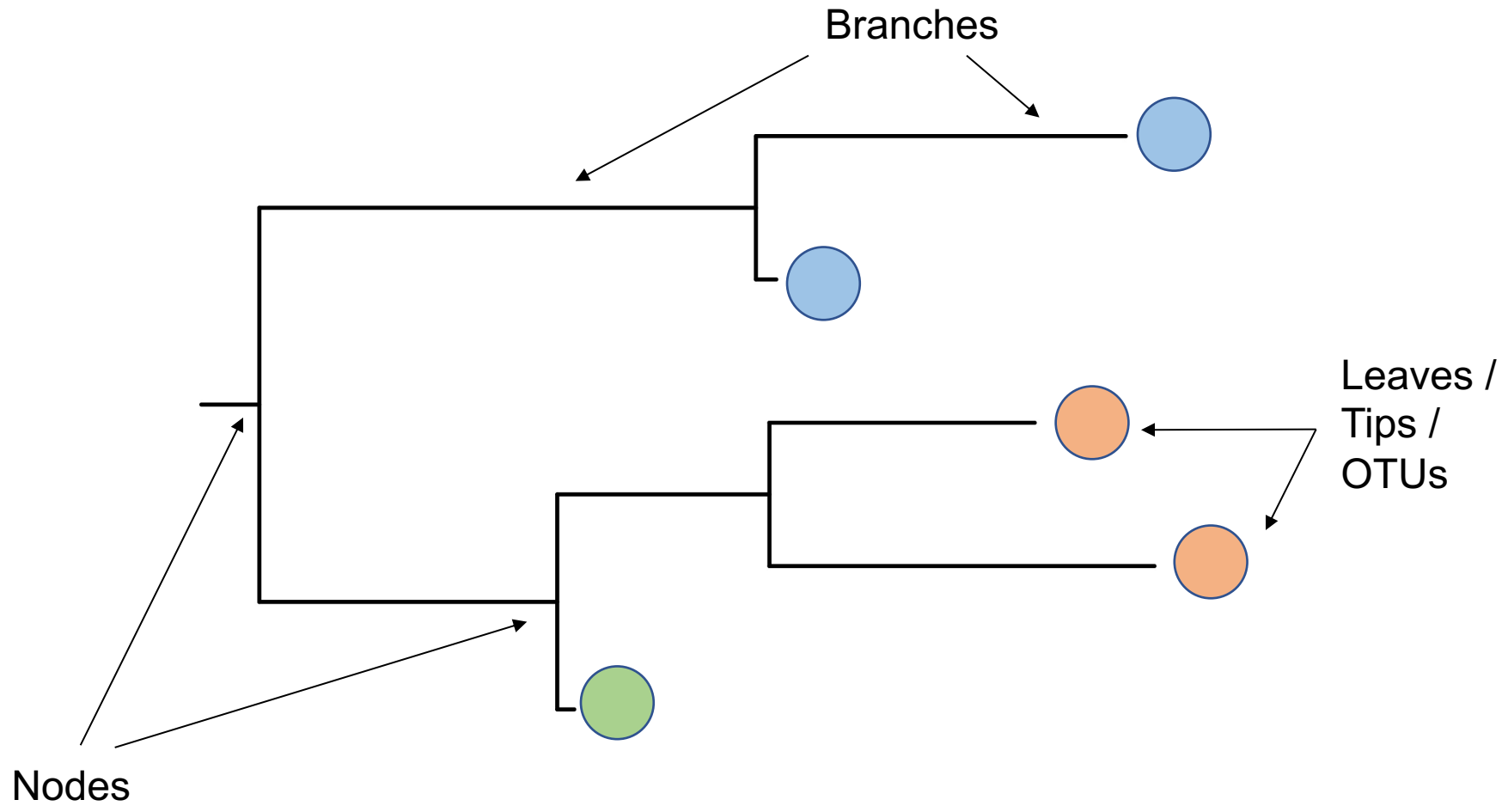


# Phylogenetic trees reveal relationships

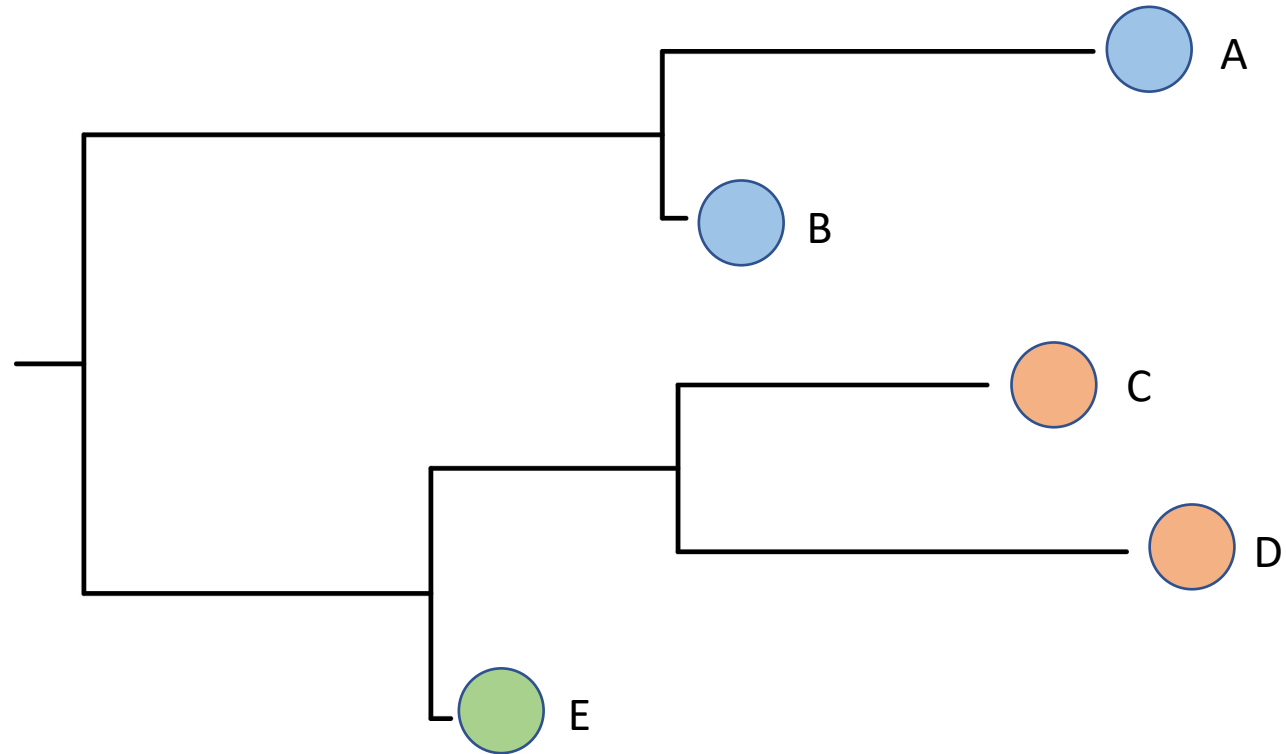


Genetic similarity

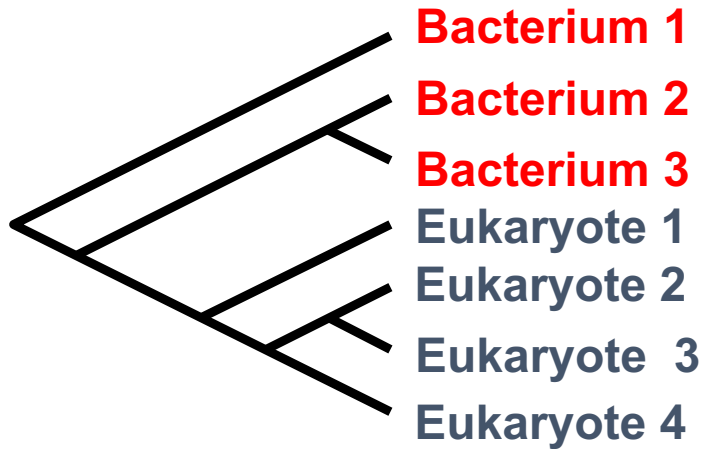
# Phylogenetic trees



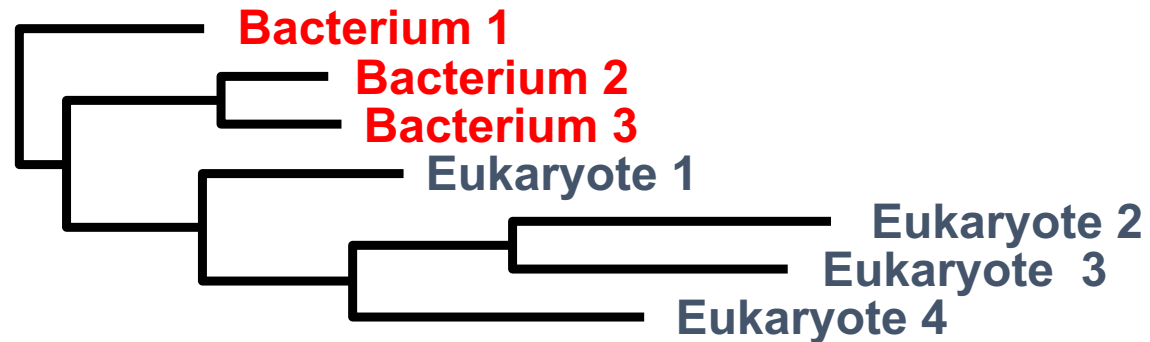
Which taxa are the most distantly related?



# Cladograms vs Phylograms



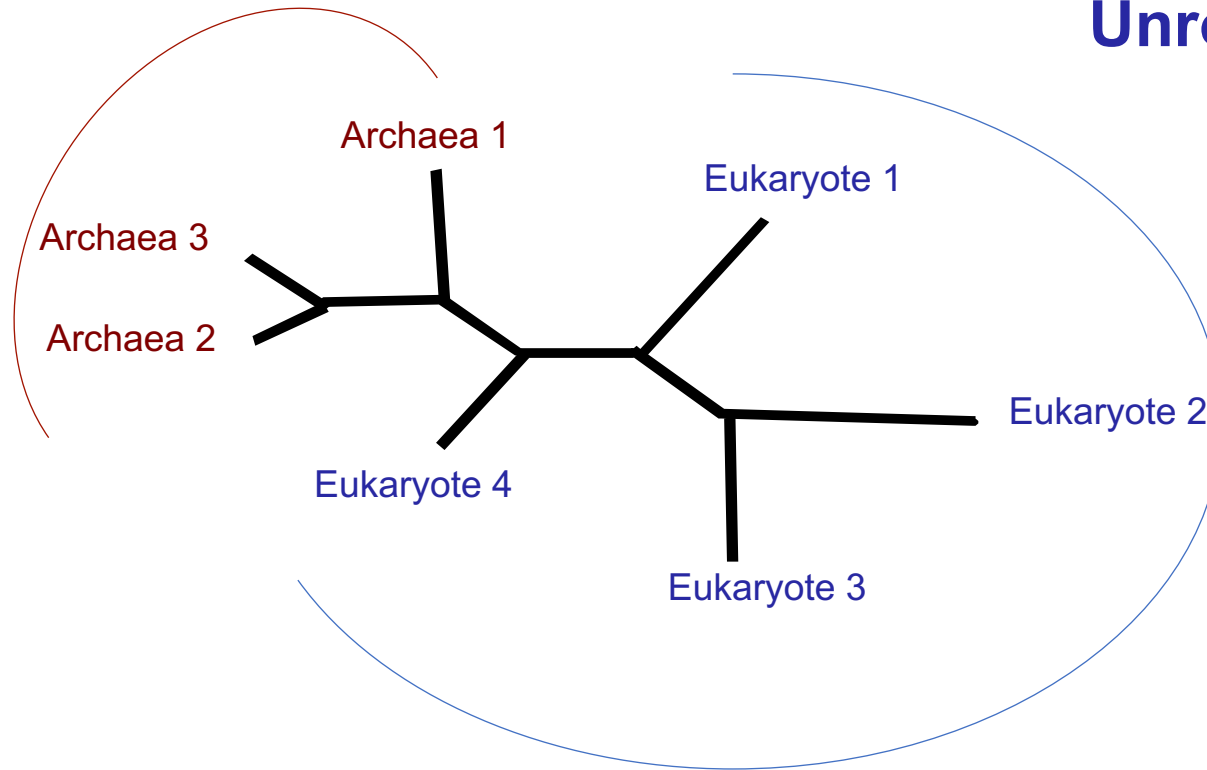
**Cladograms** show branch order (topology) only - branch lengths are meaningless



**Phylograms** show branch order and branch lengths with scale

# Rooted and Unrooted trees

Unrooted tree

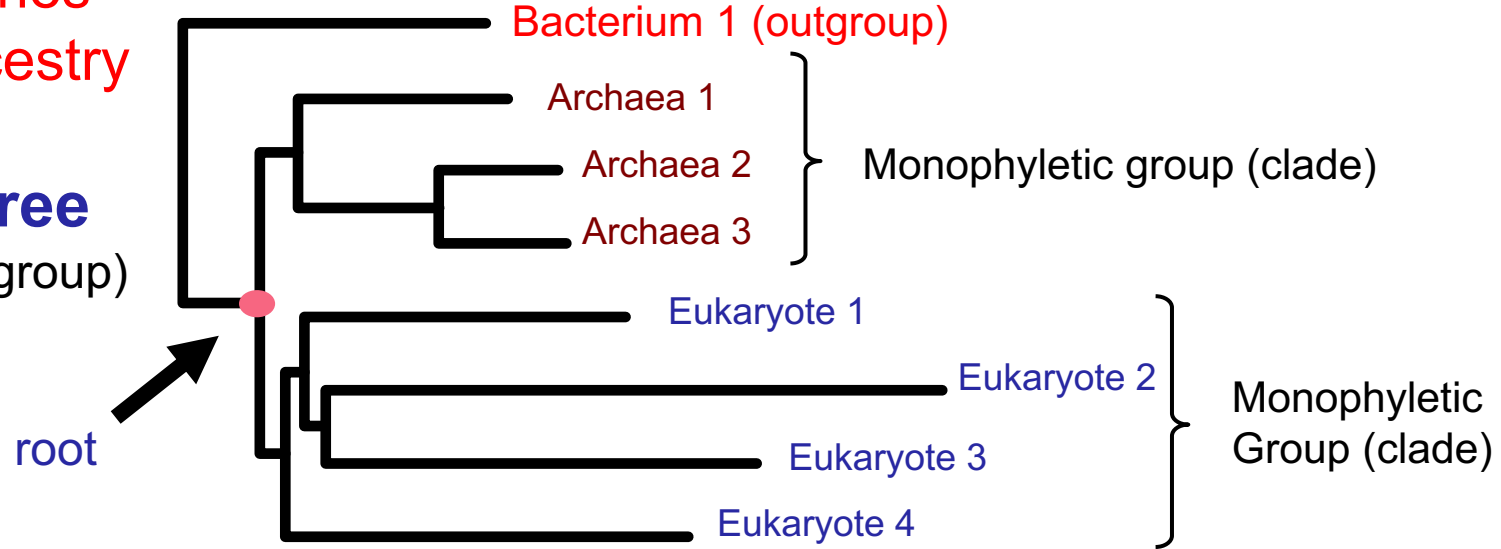




# Rooted and Unrooted trees

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

---

Identify protein, DNA or RNA sequences of interest

Fasta format file of concatenated sequences

Multiple sequence alignment

ClustalX, Muscle, Mafft

Construct phylogeny

PHYML, RAxML, IQ-Tree, FastTree

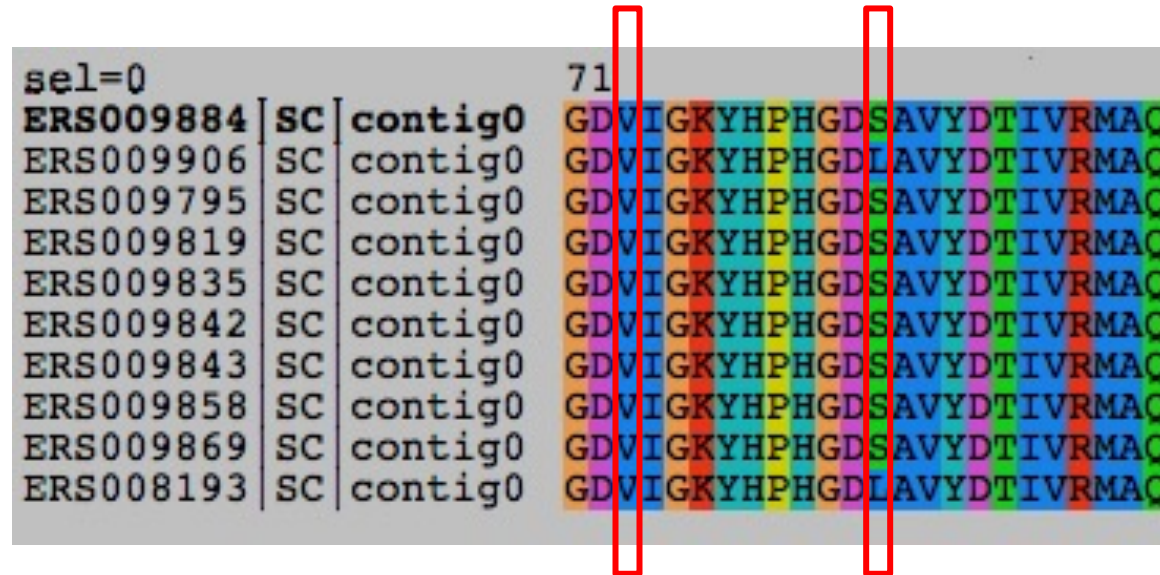
View and edit tree

FigTree

# Multiple sequence alignment (MSA)

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

```
sel=0 71
ERS009884 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS009906 | SC | contig0 | GDVIGKYHPHGDLAVYDTIVRMAQ
ERS009795 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS009819 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS009835 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS009842 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS009843 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS009858 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS009869 | SC | contig0 | GDVIGKYHPHGDSAVYDTIVRMAQ
ERS008193 | SC | contig0 | GDVIGKYHPHGDLAVYDTIVRMAQ
```



This is perhaps most important step!!

Crap in == Crap out!

# MSA - can be easy but also tricky

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

Easy

**Alignments can be difficult to get right!**

# Important note on MSAs

---

COMPUTERS DO NOT NECESSARILY KNOW BETTER

check your alignments by eye if possible and  
remove erroneous sections

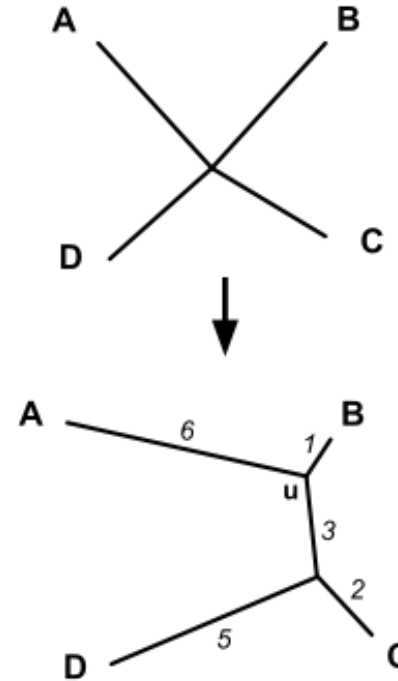
# Constructing a phylogenetic tree

Method	Data used	Tree search	Evolutionary Model
Distance	Pairwise distance	Simple algorithm	Can be complex
Parsimony	All 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

Method	Data used	Tree search	Evolutionary Model
Distance	Pairwise distance	Simple algorithm	Can be complex

	A	B	C	D
A	0	7	11	14
B	7	0	6	9
C	11	6	0	7
D	14	9	7	0

Distance matrix



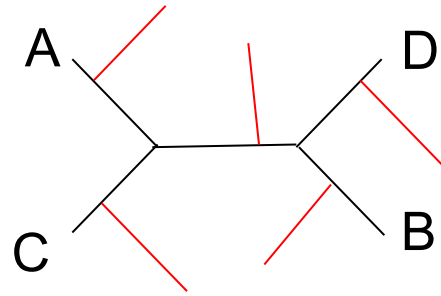
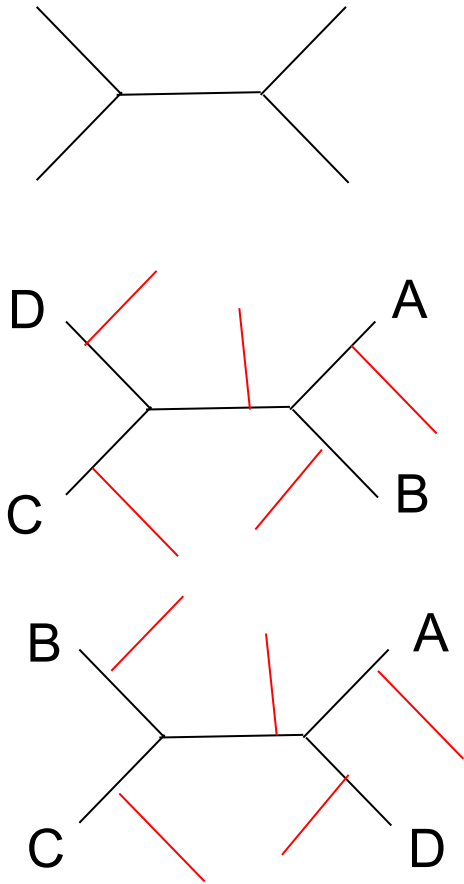
e.g. UPGMA, Neighbour joining, minimum evolution, BIONJ



# Methods that attempt to find the BEST tree

Method	Data used	Tree search	Evolutionary Model
Distance	Pairwise distance	Simple algorithm	Can be complex
Parsimony	All sites	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

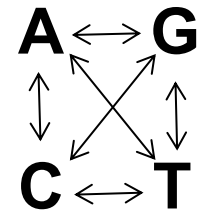
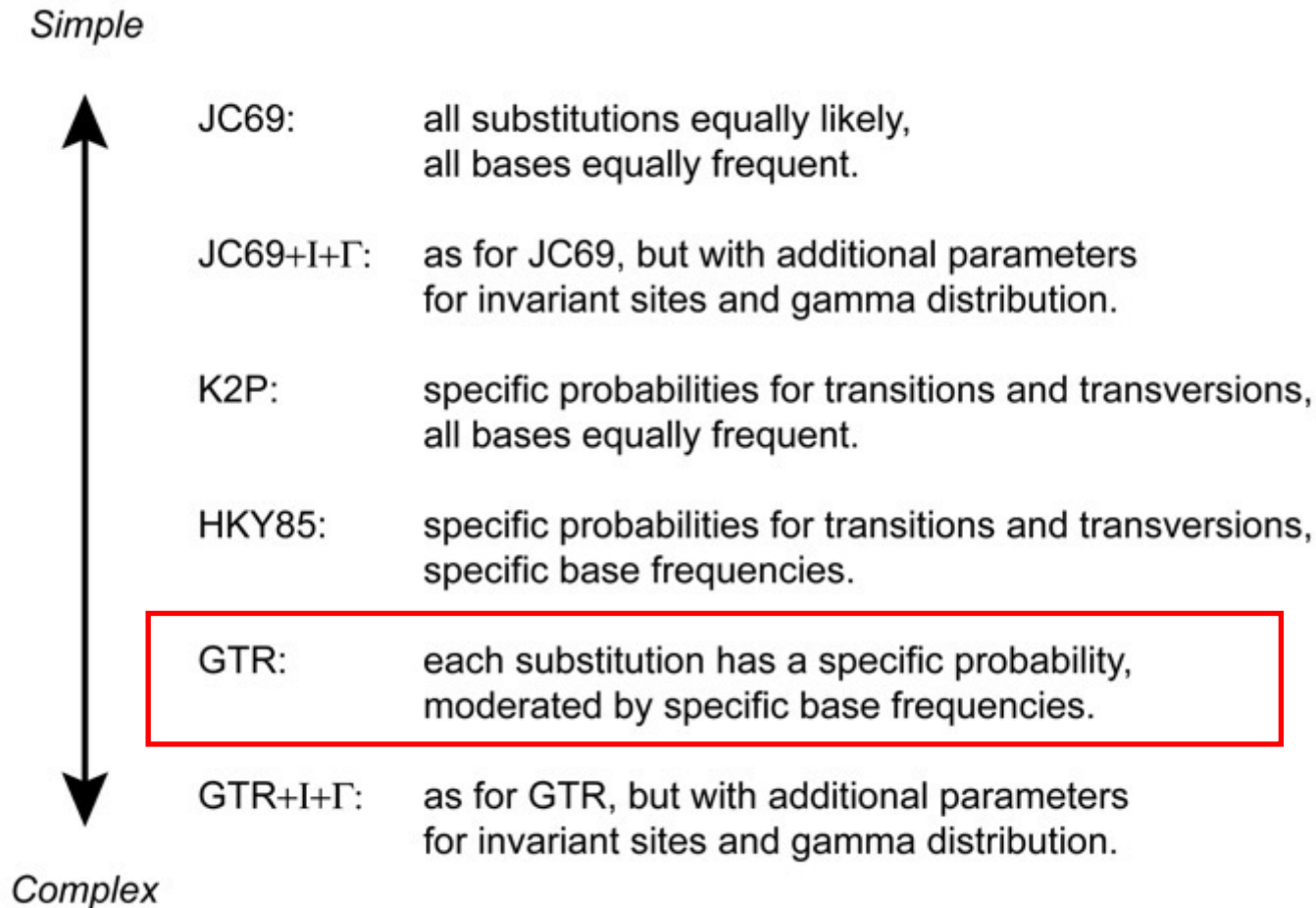


**E?**

**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	$2.18 \times 10^{137}$

# Maximum likelihood phylogenetic models

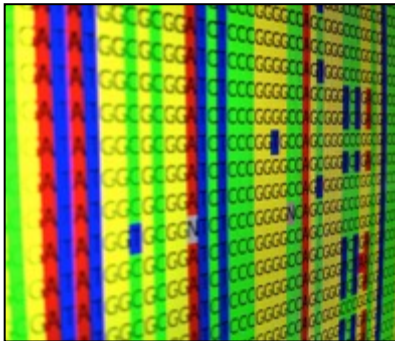


4 equilibrium base frequency parameters and 6 substitution rate parameters and

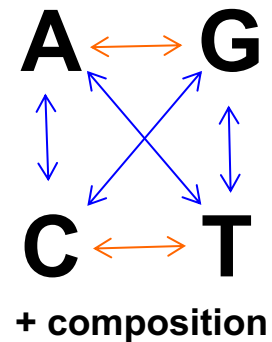
# Putting it together

Maximum likelihood phylogenetic models maximize the probability of achieving ...

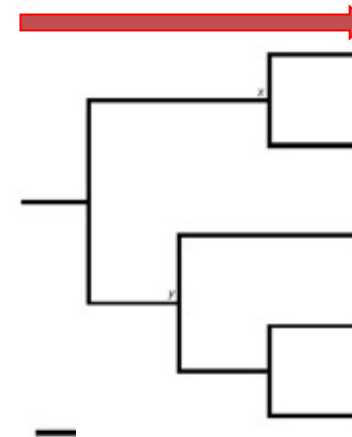
these data...



... if this happens...



... over this tree



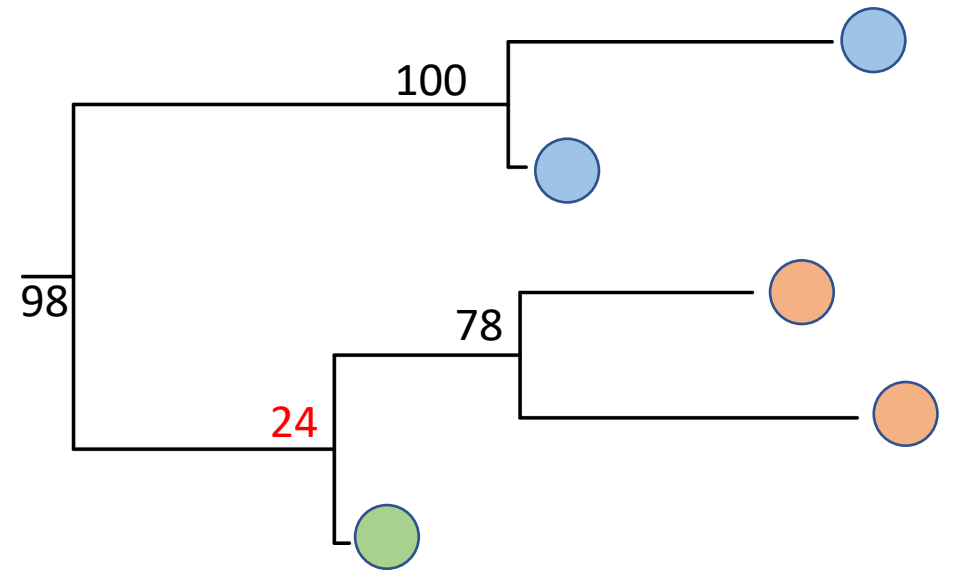
# Gaining confidence : Bootstrapping

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

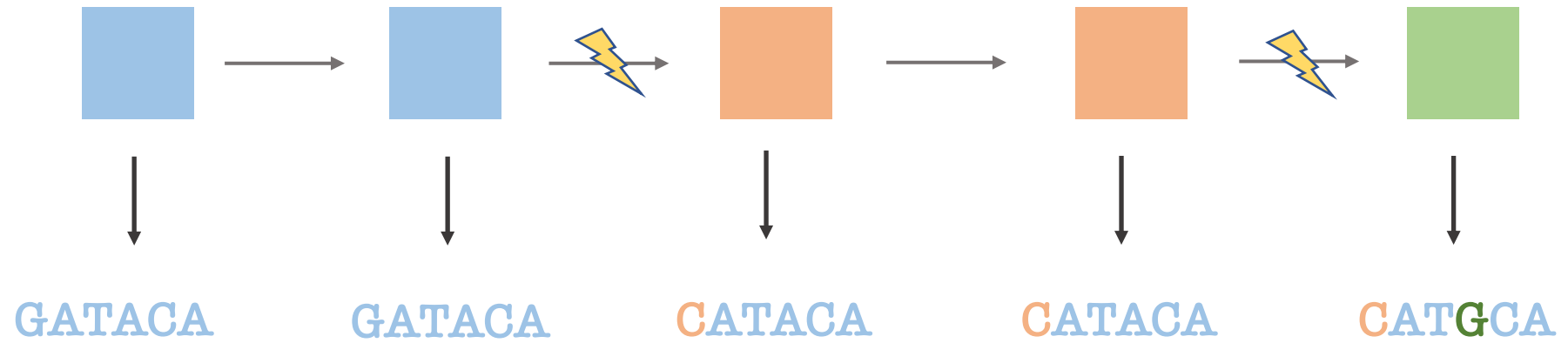
**X** number of **bootstraps** (resampled replicates) are created of your input data (MSA)

Typically run 100 – 1,000 bootstraps for ML analysis

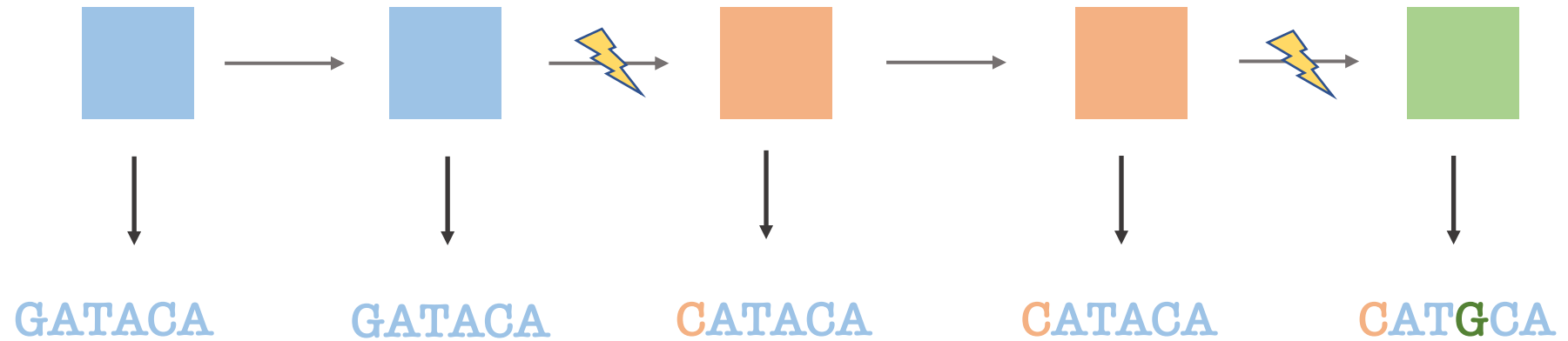
These are commonly used as a measure of support for these branches and are represented as a number on each tree branch



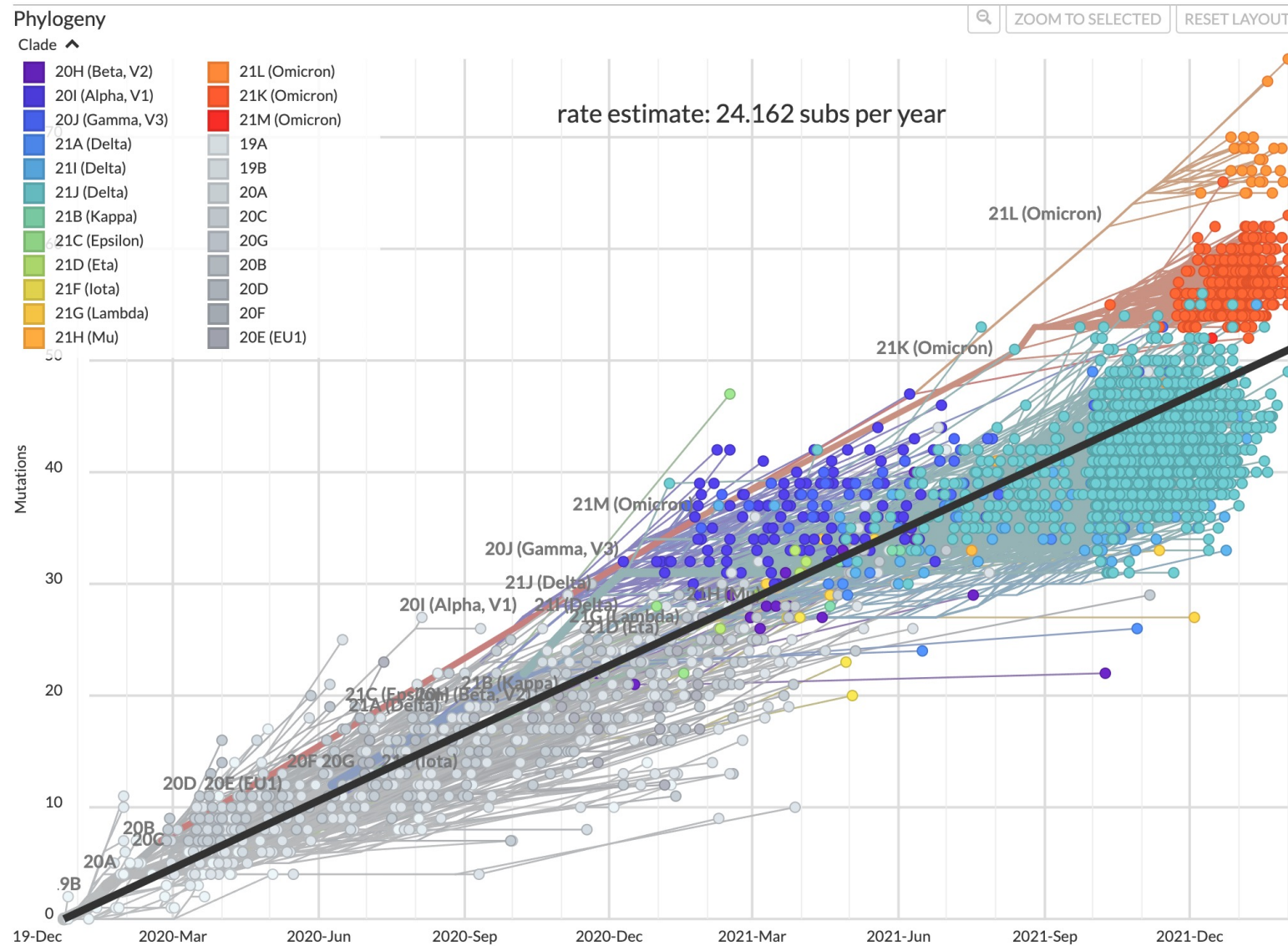
# Pathogens mutate as they transmit



# SARS-CoV-2 mutates once every two weeks

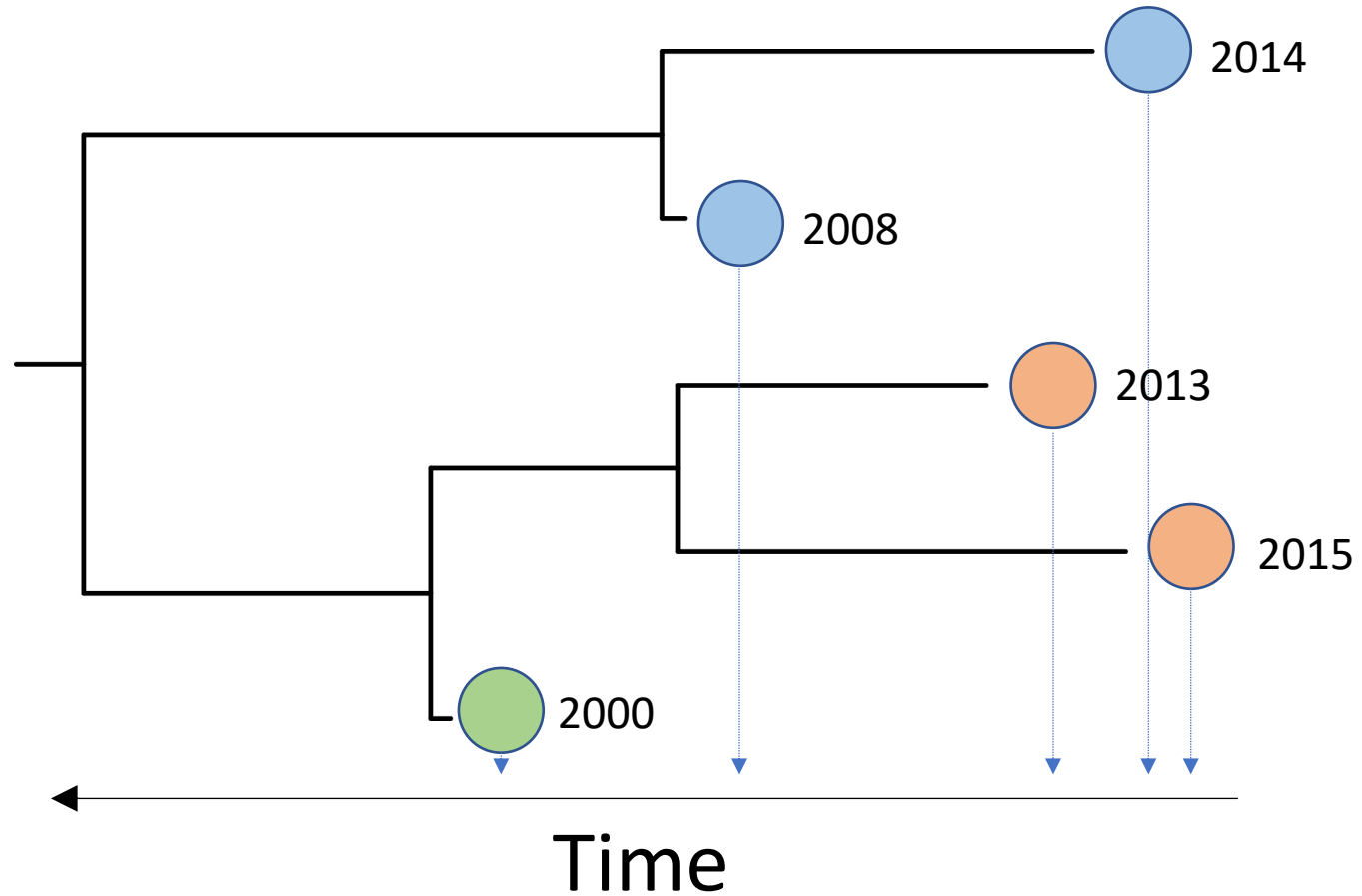


# SARS-CoV-2 molecular clock





# Trees reveal timing



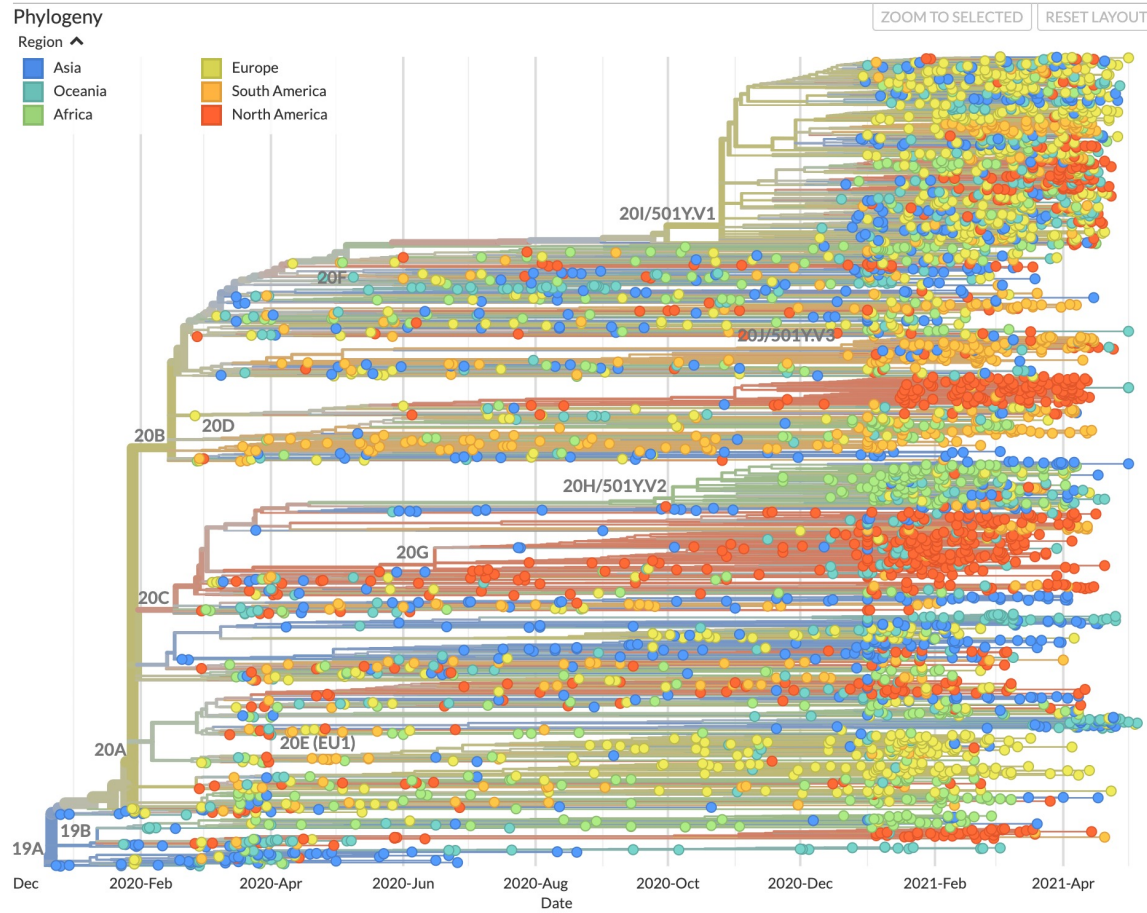
Typically use BEAST, BactDater, or TreeTime to generate

# SARS-CoV-2 phylodynamics

## Genomic epidemiology of novel coronavirus - Global subsampling

Built with [nextstrain/ncov](#). Maintained by the [Nextstrain team](#). Enabled by data from [GISAID](#).

Showing 3825 of 3825 genomes sampled between Dec 2019 and May 2021.



The background is a complex geometric pattern composed of numerous triangles of varying sizes and orientations. The color palette is primarily purple and blue, with a vibrant red section on the left and a teal/green section on the right. The triangles are arranged in a way that creates a sense of depth and movement.

Questions?