

# SARS-COV-2 ARTIC PIPELINE



DARYL DOMMAN, PHD   DARRELL DINWIDDIE, PHD

[DDOMMAN@GMAIL.COM](mailto:DDOMMAN@GMAIL.COM)



# Today's Agenda



Lecture



SARS-CoV-2 lineage calling with Pango and NextClade

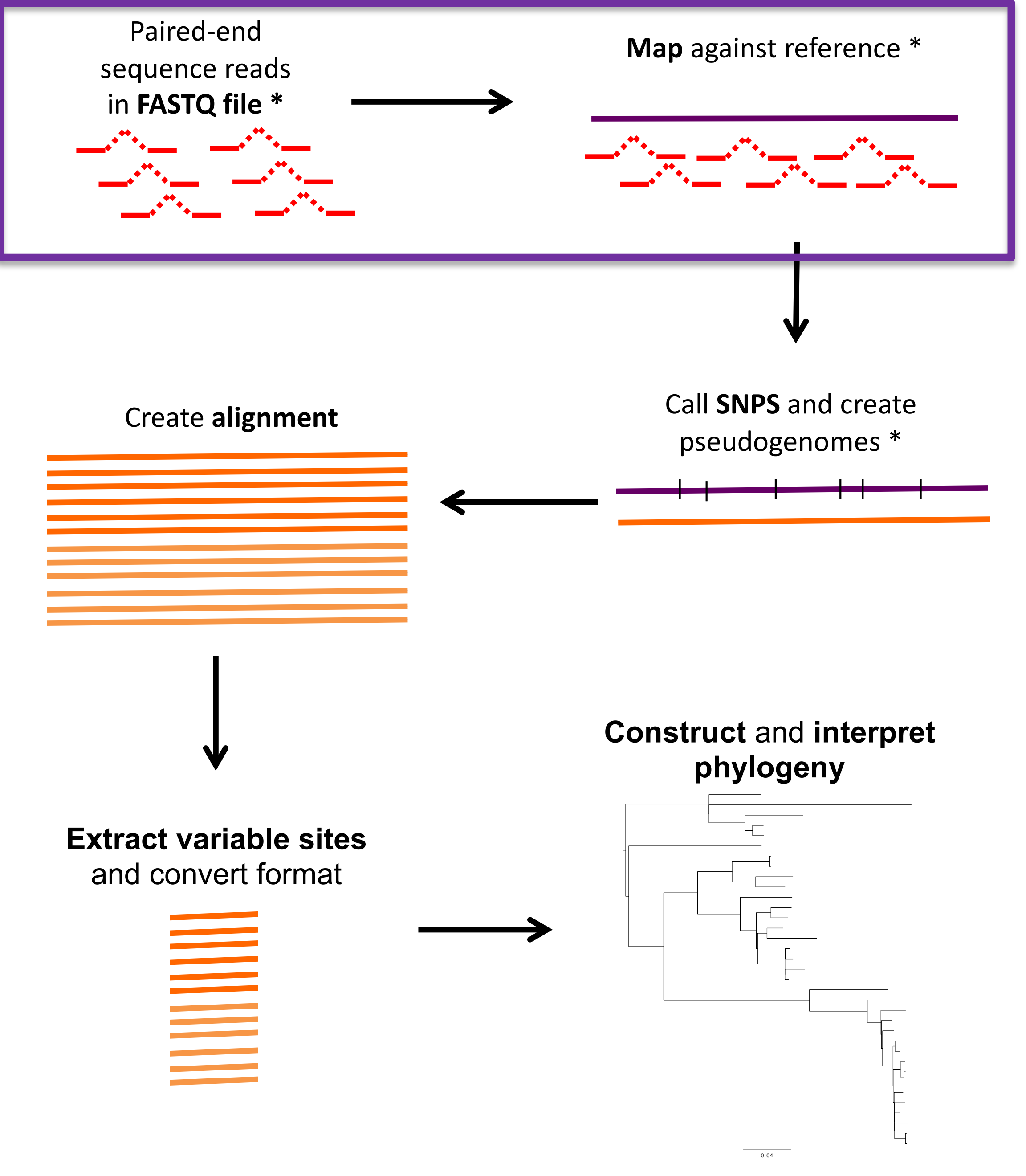


Align sequences with Mafft

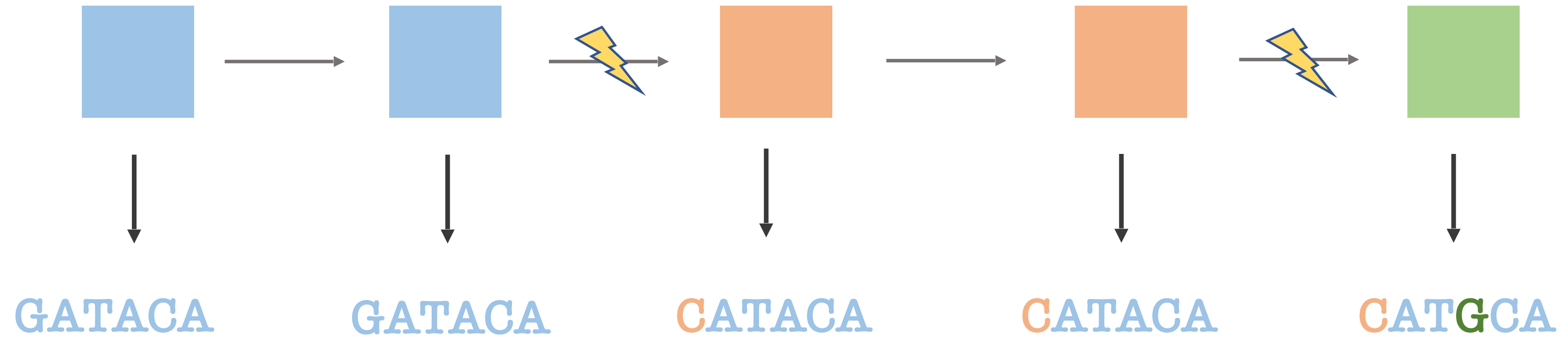


Phylogeny

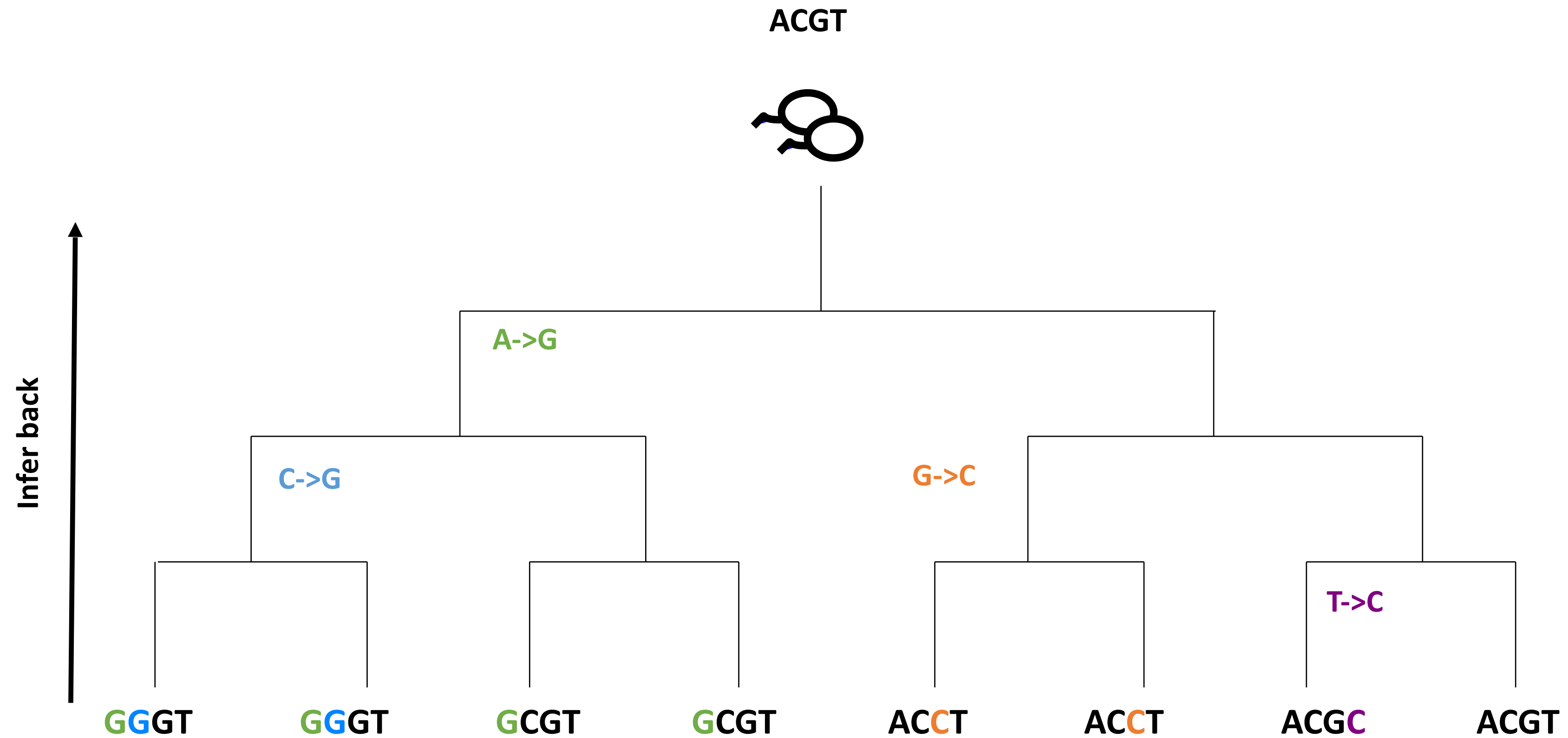
---



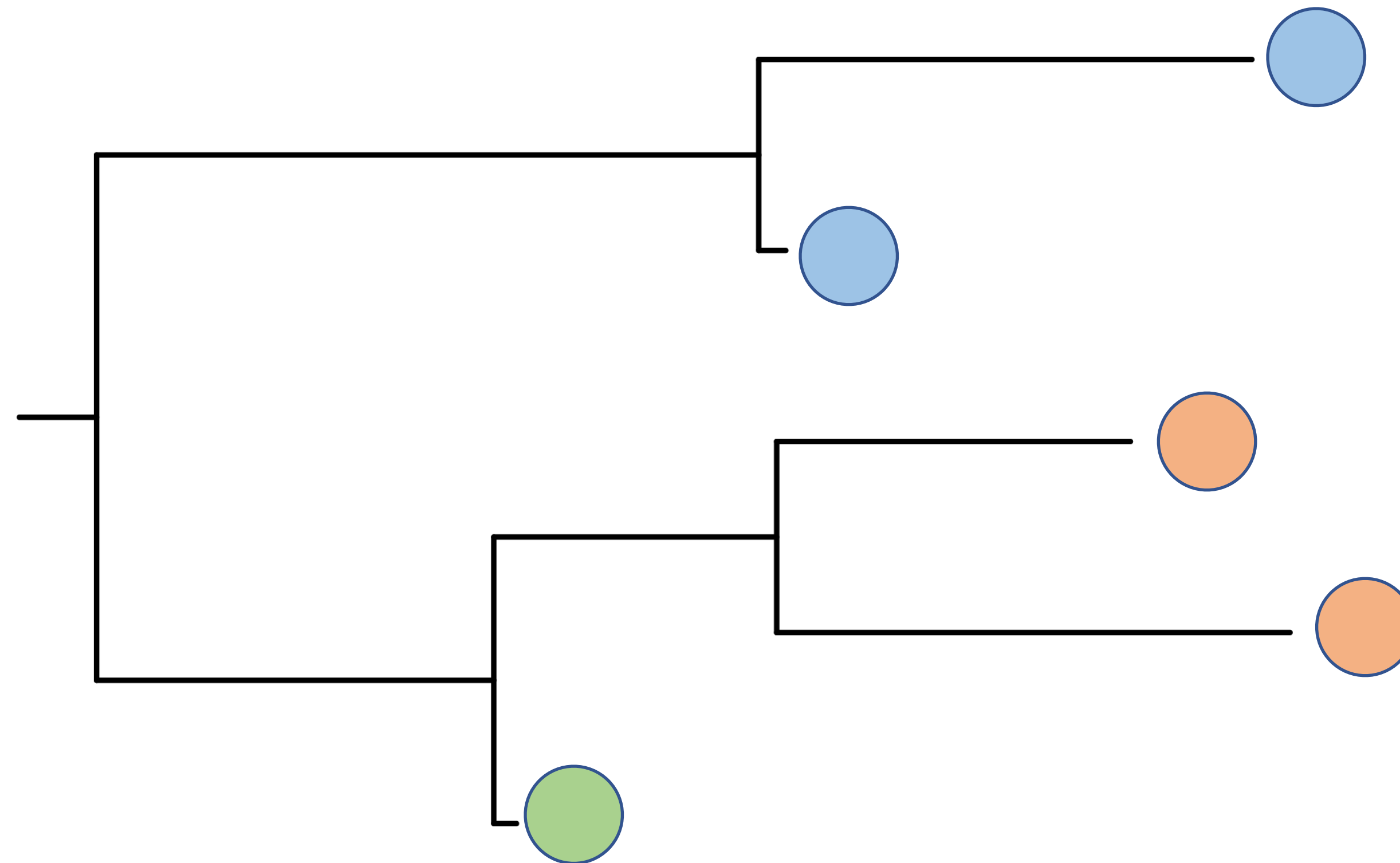
# Organisms acquire mutations



# Mutations tell us about relationships

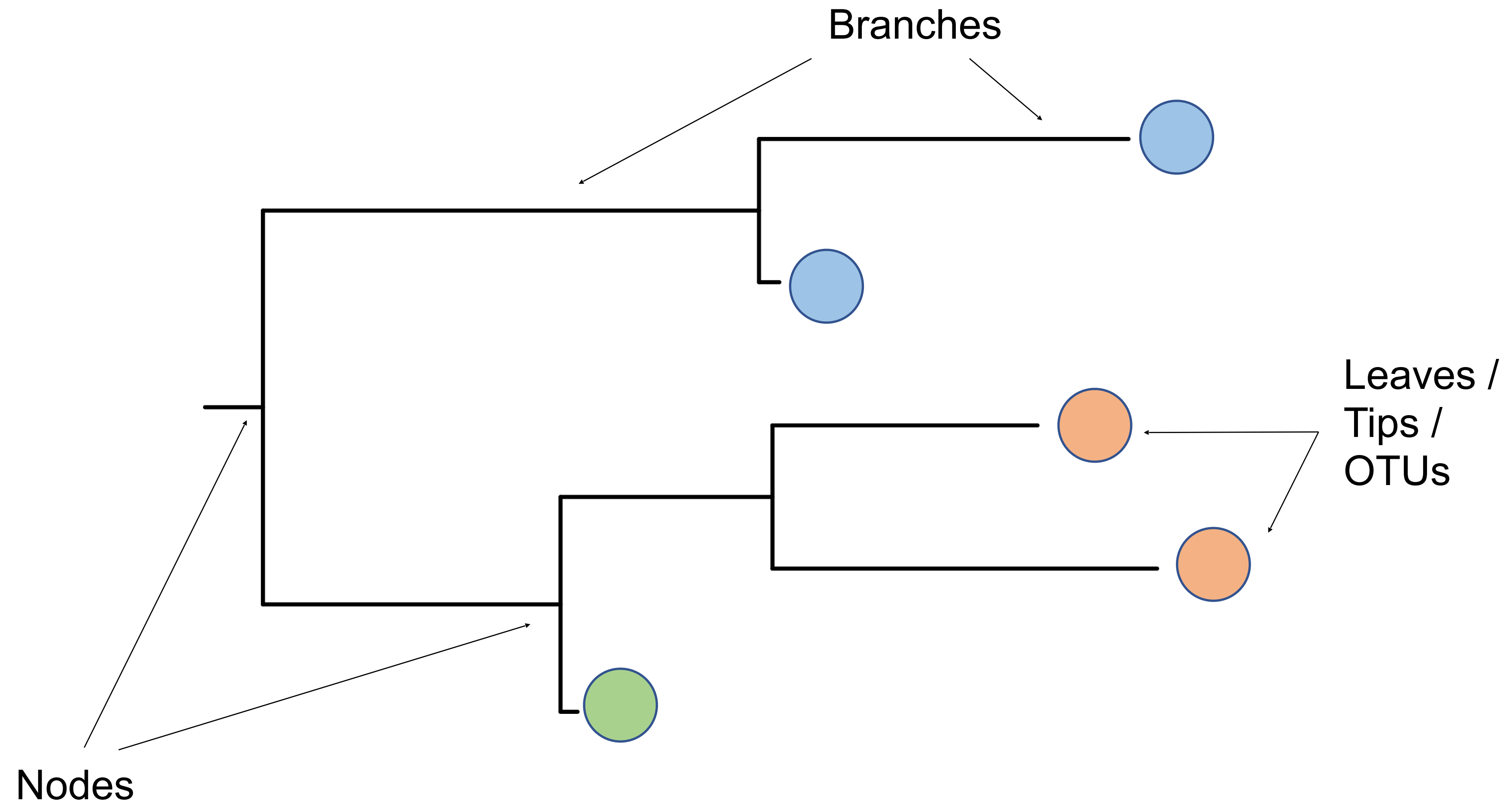


# Phylogenetic trees reveal relationships



Genetic similarity

# Phylogenetic trees



# 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

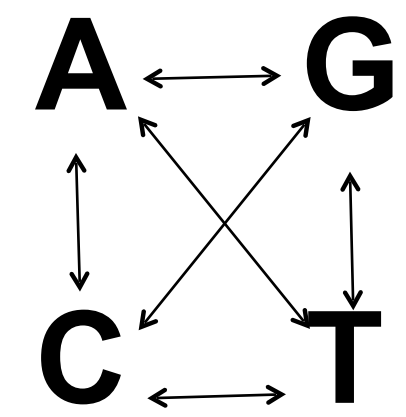
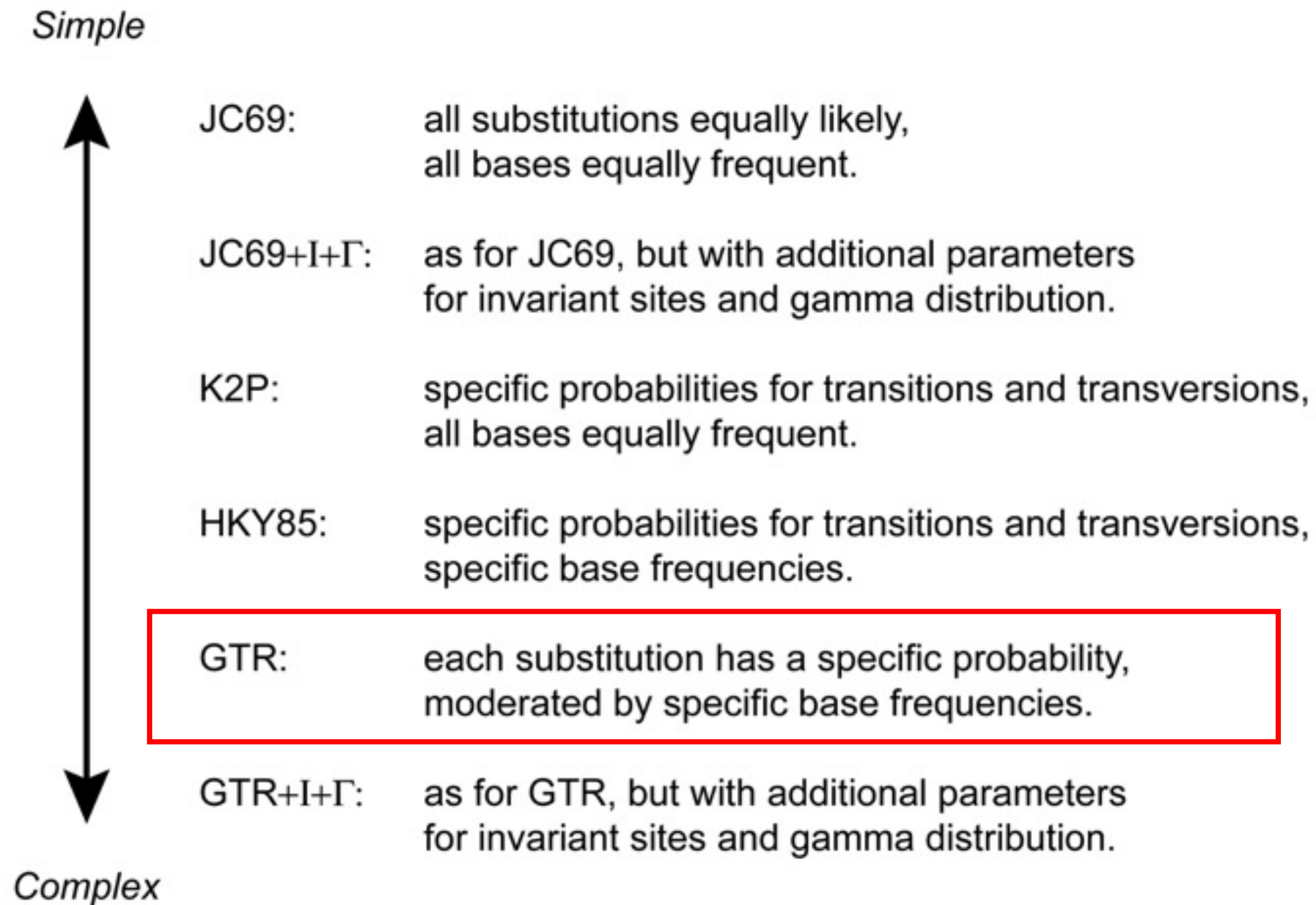
This is perhaps most important step!!

Crap in == Crap out!

# 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

# Maximum likelihood phylogenetic models



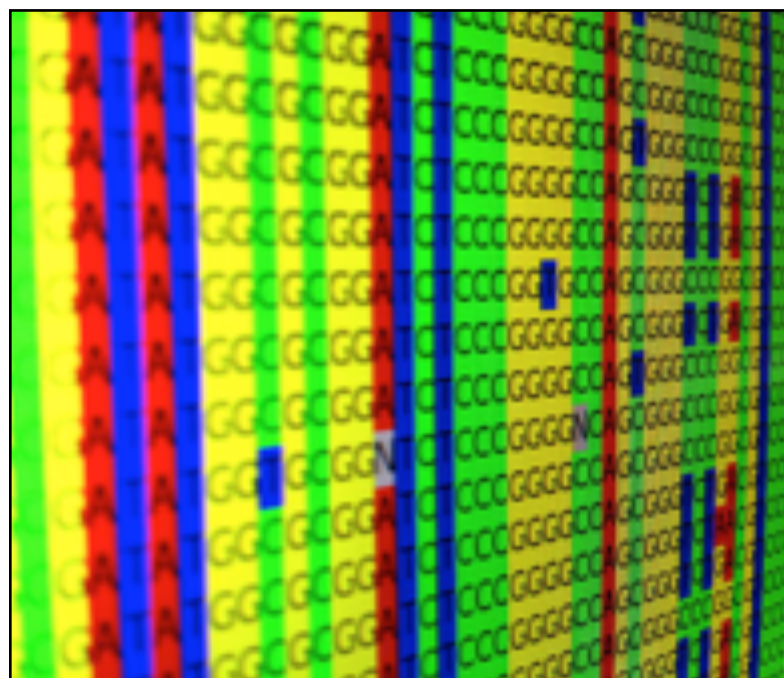
4 equilibrium base  
frequency  
parameters and 6  
substitution rate  
parameters and



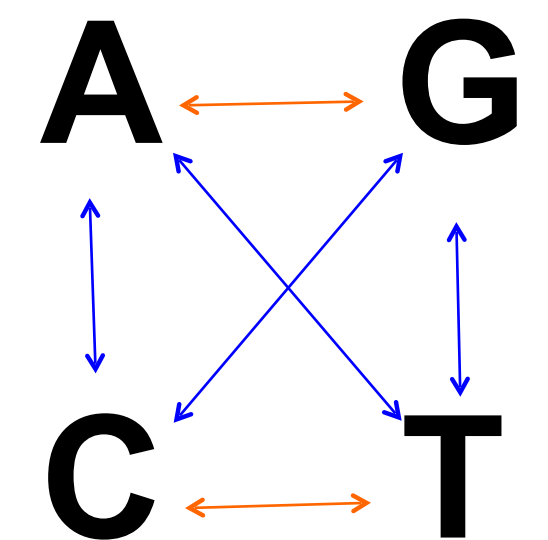
# Putting it together

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

these data...

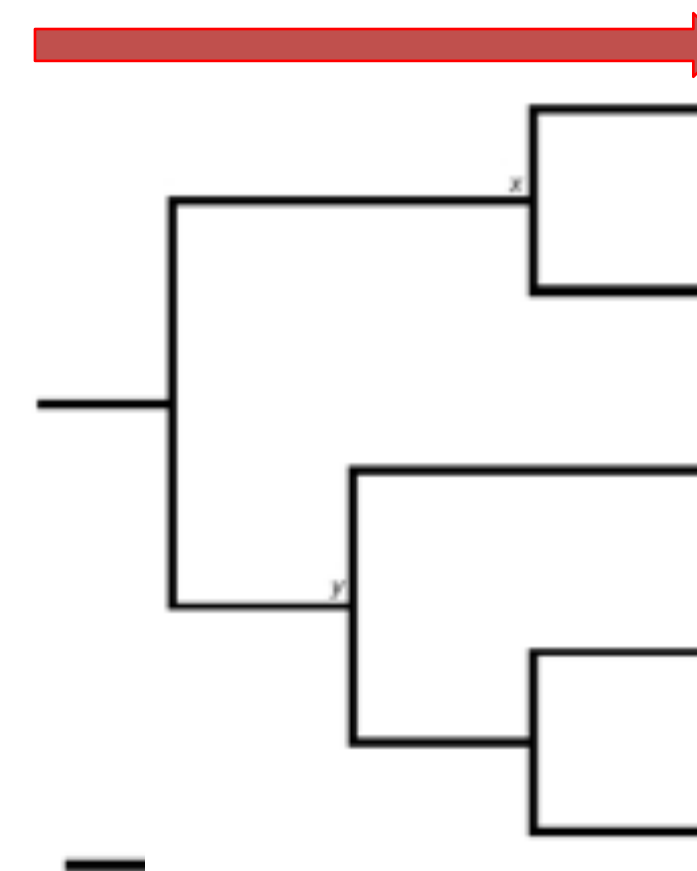


... if this happens...



+ composition

... over this tree



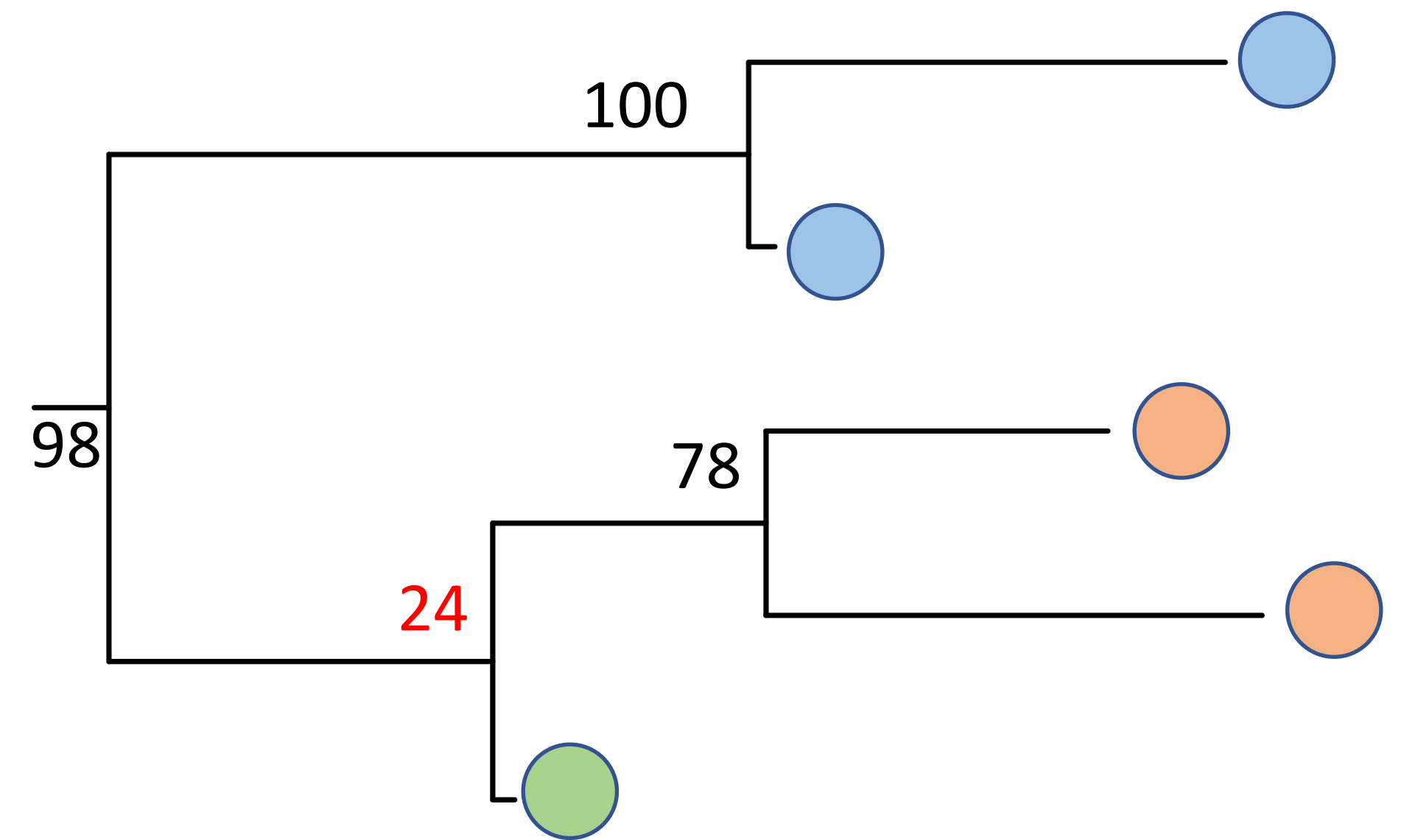
# Gaining confidence : Bootstrapping

**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



**Questions?**