NIH Fogarty International Center
Advancing Science for Global Health

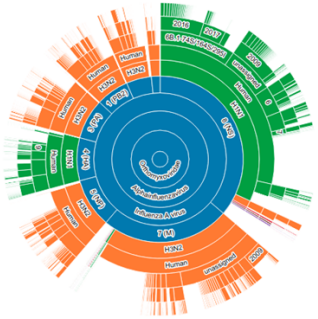
COV-IRT
COVID-19 International Research Team

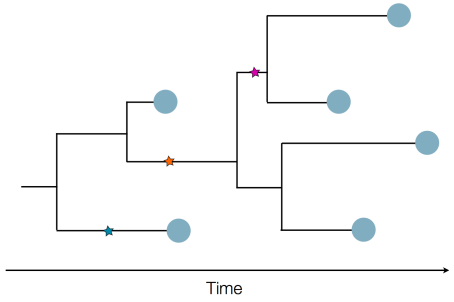
Phylogenetic Trees

Methods and Interpretation

James R. Otieno (PhD), Nídia Trovão (PhD) and Martha Nelson (PhD)
Division of International Epidemiology and Population Studies
Fogarty International Center
National Institutes of Health

Genomic epidemiology workflow





Obtain Sample

DNA/RNA for sequencing

Sequence DNA

Initial Sample Processing

In-depth Analysis

Background Dataset

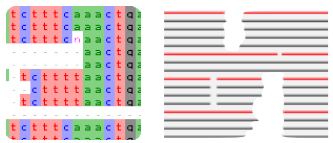
Multiple Sequence Alignment

Building a Phylogeny

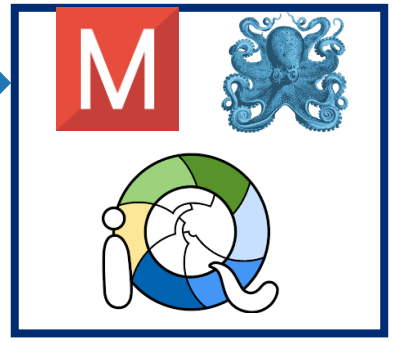
Interpreting a Phylogenetic Tree

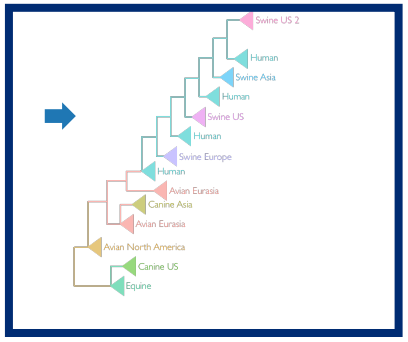
GISAID

- Download from public database



Sequence alignment
Clean and manually edition

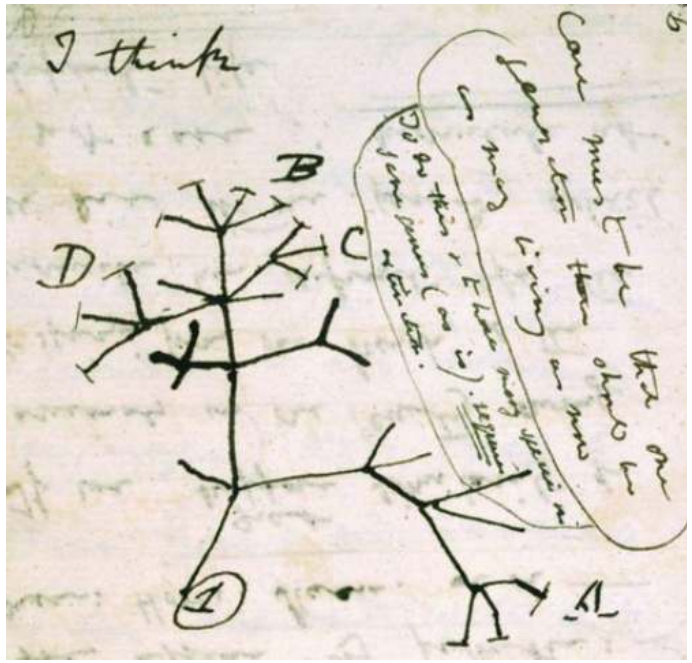




Phylogenetic Trees

What is a phylogenetic Tree?

- A diagram used for depicting evolutionary relationships among genes and organisms.
- Shows which genes/organisms are more closely related and which are more distant.
- If rooted, also shows the direction of the evolutionary process across time.

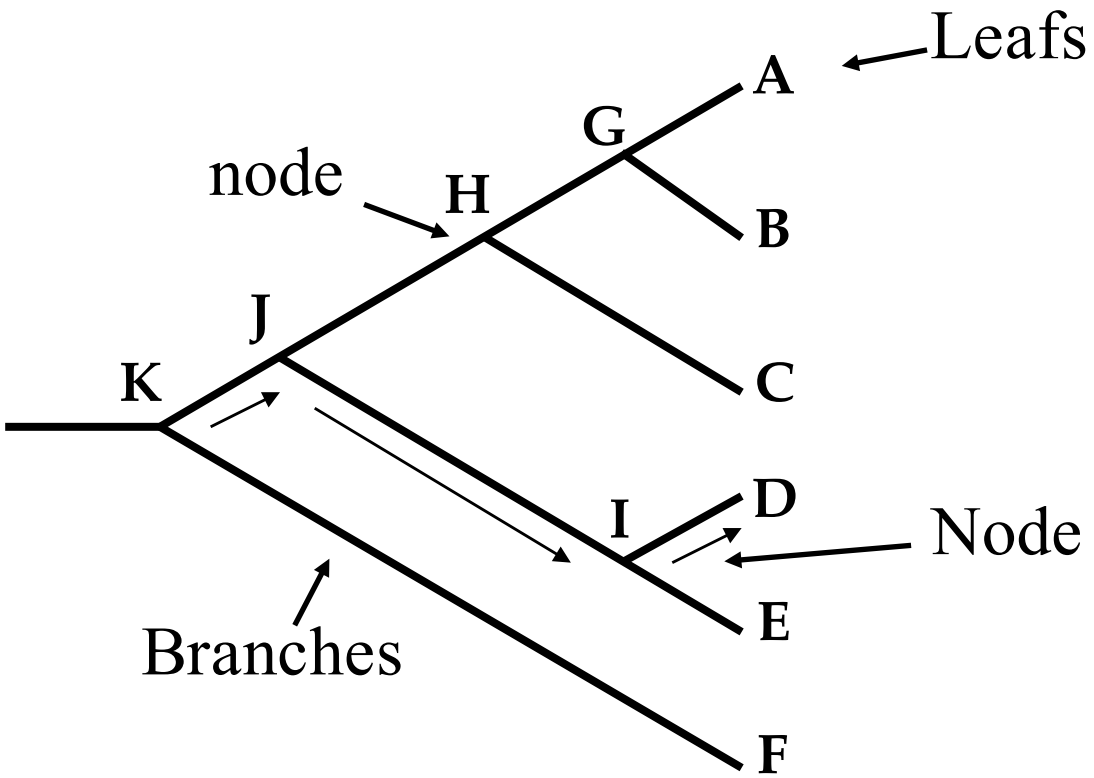


Components of a phylogenetic Tree?

Leafs/Taxa/Tips:
A, B, C, D, E, F

Nodes (External/Internal)
G, H, I, J, K

Branches



Ancestors → Descendants

Important Problems in Molecular Phylogenetic Analysis

- Is there a tree at all (e.g. recombination)?

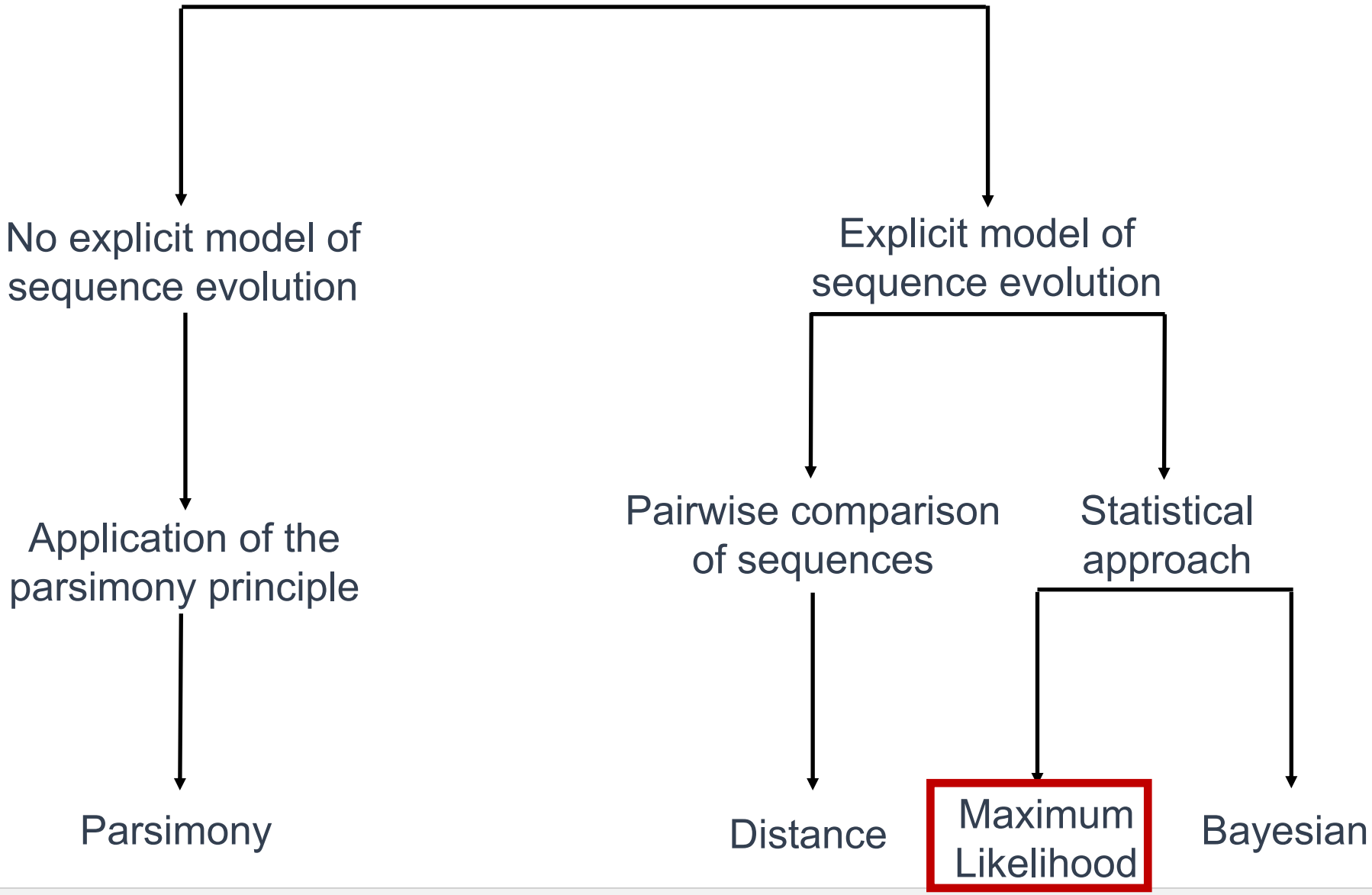
- Many possible trees:
  - For 10 taxa there are 2×10^6 unrooted trees
  - For 50 taxa there are 3×10^{74} unrooted trees

- Efficient and powerful search algorithms

- Choosing the right model of nucleotide substitution

- Rate variation among lineages (causes “long branch attraction”)
  → Need a representative sample of taxa.

Phylogenetic Tree Building Methods



Software for Inferring Phylogenetic Trees

- *Parsimony (PAUP*)*

Find tree with the minimum number of mutations between sequences (i.e. choose tree with the least convergent evolution)

- *Neighbor-Joining (PAUP*, MEGA)*

Estimate genetic distances between sequences and cluster these distances into a tree that minimizes genetic distance over the whole tree

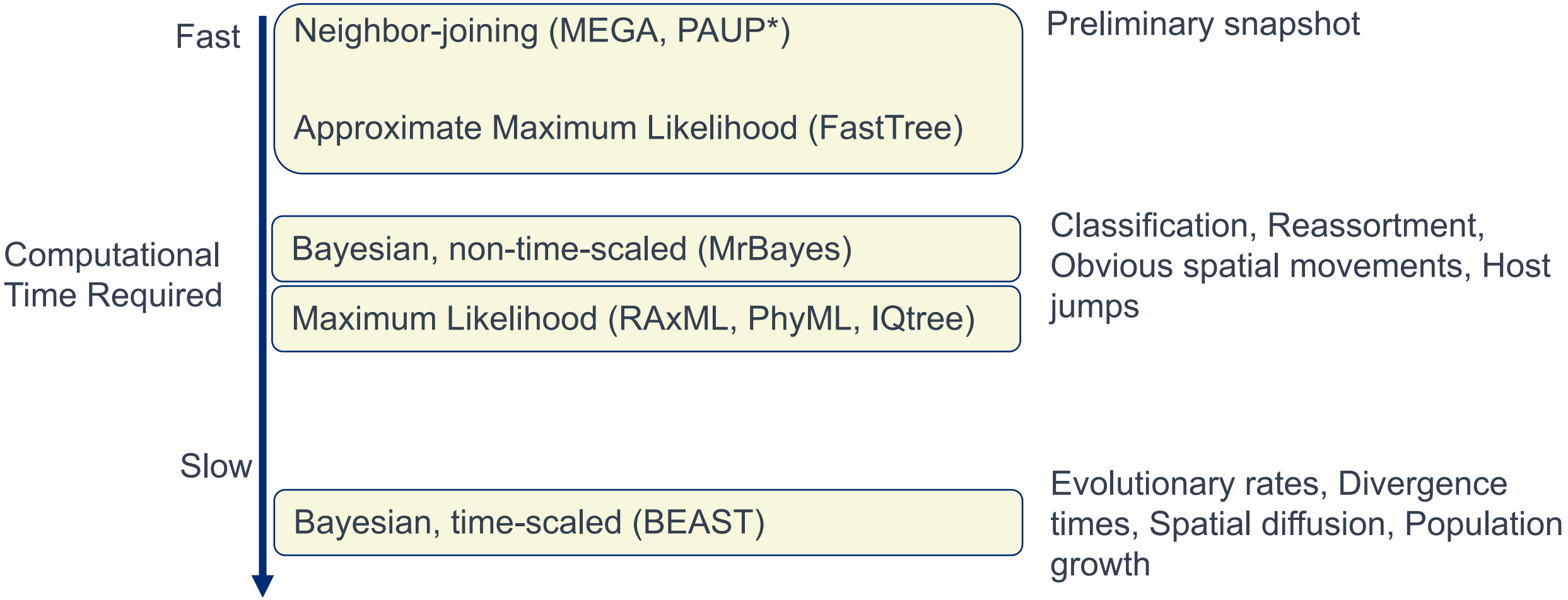
- *Maximum Likelihood (PAUP*, GARLi, PhyML, RaxML, FastTree, IQ-Tree)*

Determine the probability of a tree (and branch lengths) given a particular *model* of molecular evolution and the observed sequence data

- *Bayesian (BEAST, Mr.Bayes)*

Similar to likelihood but where there is information about the prior distribution of parameters. Also returns a (posterior) distribution of trees

Advantages of Tree-Building Methods



Likelihood

- Maximum likelihood methods allow us to incorporate extremely detailed models of molecular evolution
- Likelihood is a quantity *proportional* to the probability of observing an outcome/data/event X given a hypothesis H

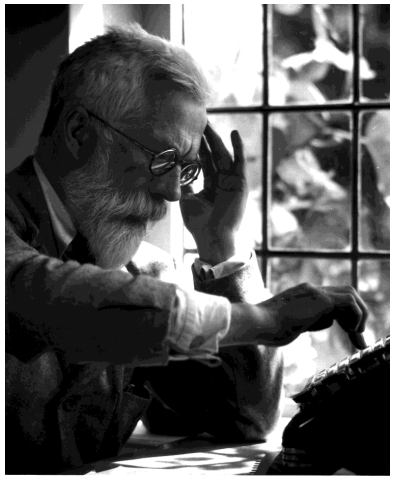
$$P(X \mid H) \text{ or } P(X \mid p)$$

- then we would talk about the likelihood

$$L(p \mid X)$$

that is, the likelihood of the parameters given the data.

- In this case the hypothesis is a tree + branch lengths and the data are the sequences

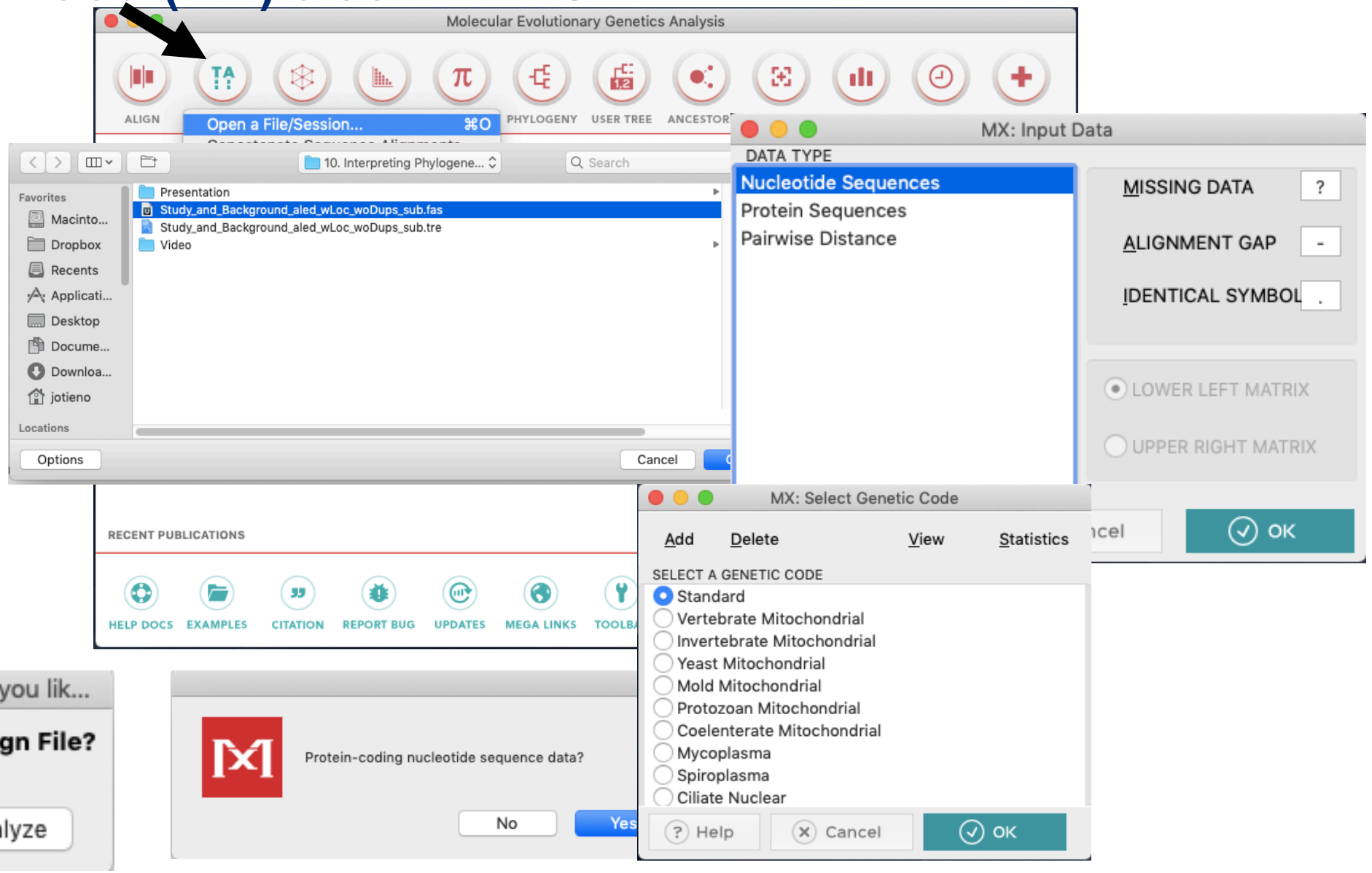


R.A. Fisher

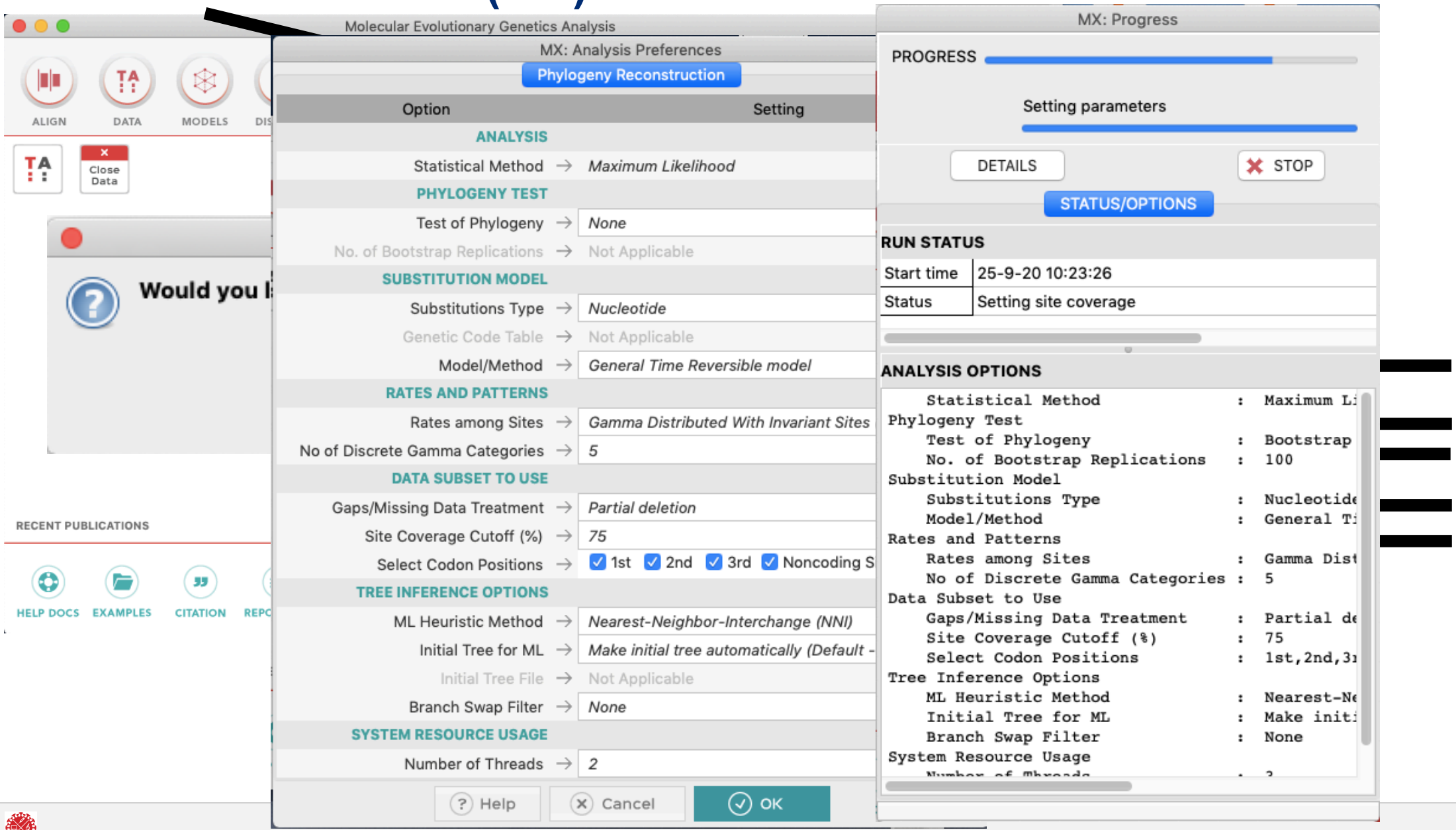
Practical: Building A Phylogenetic Tree

Maximum likelihood (ML) tree in MEGA

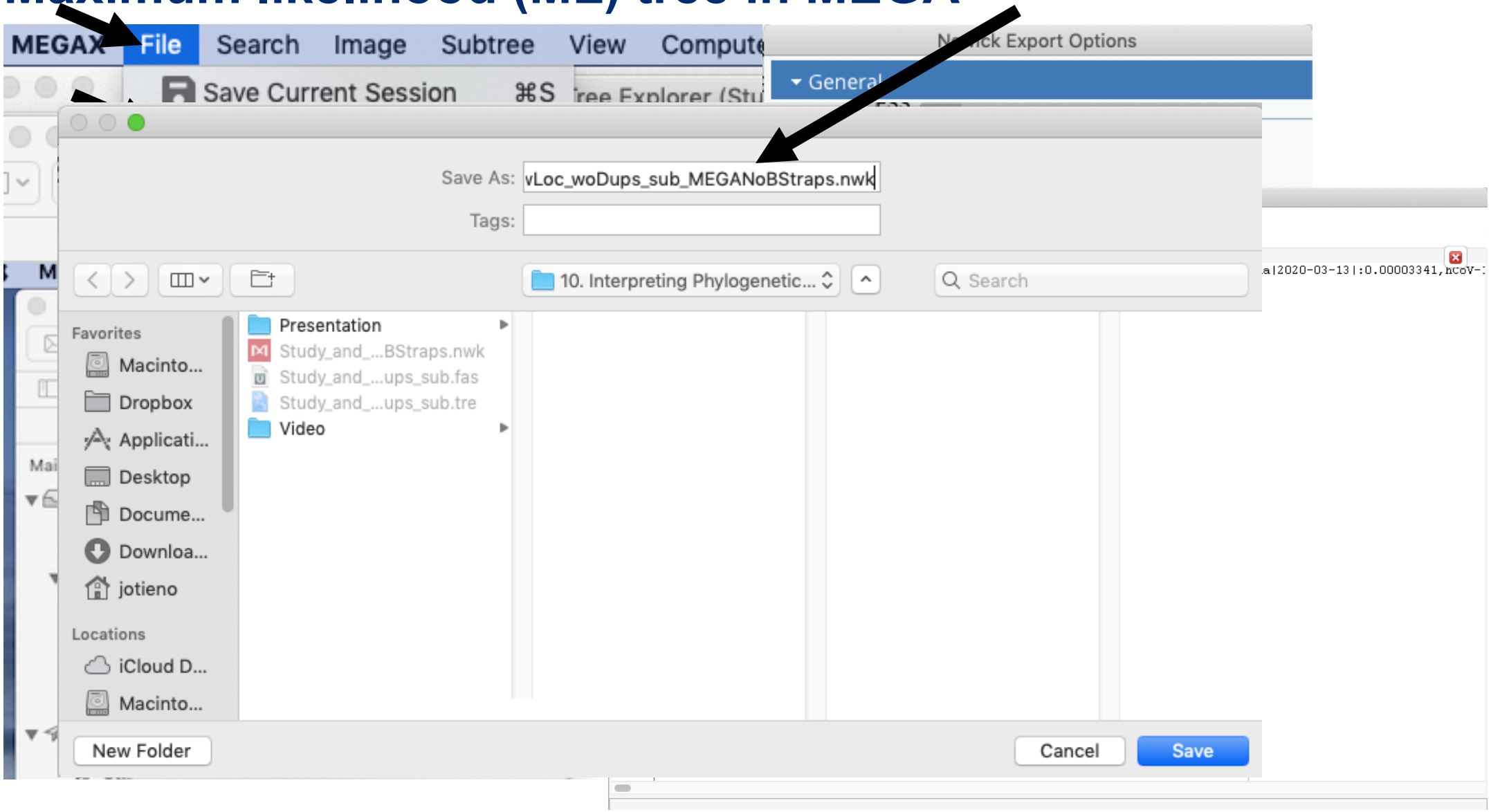
MEGA



Maximum likelihood (ML) tree in MEGA

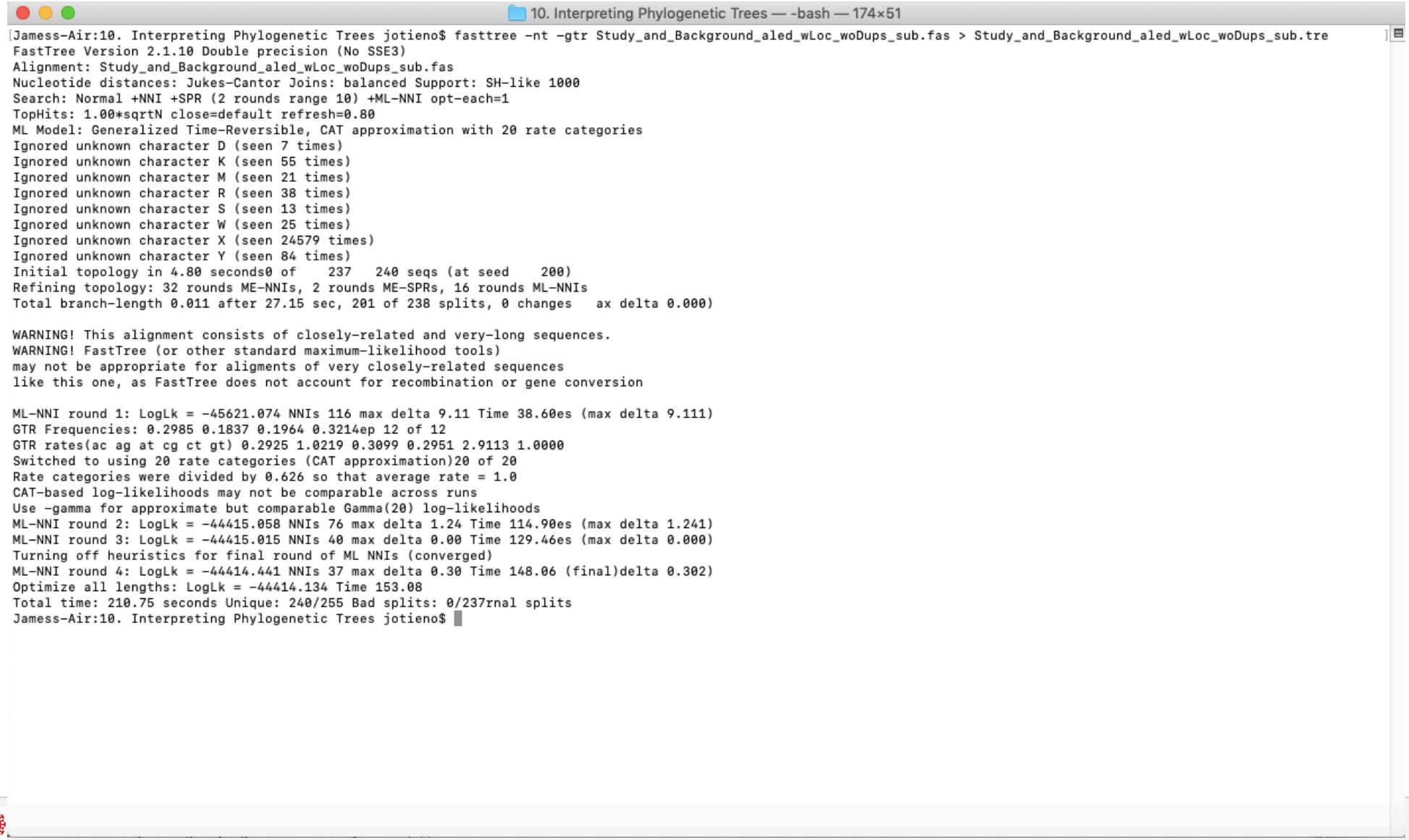


Maximum likelihood (ML) tree in MEGA



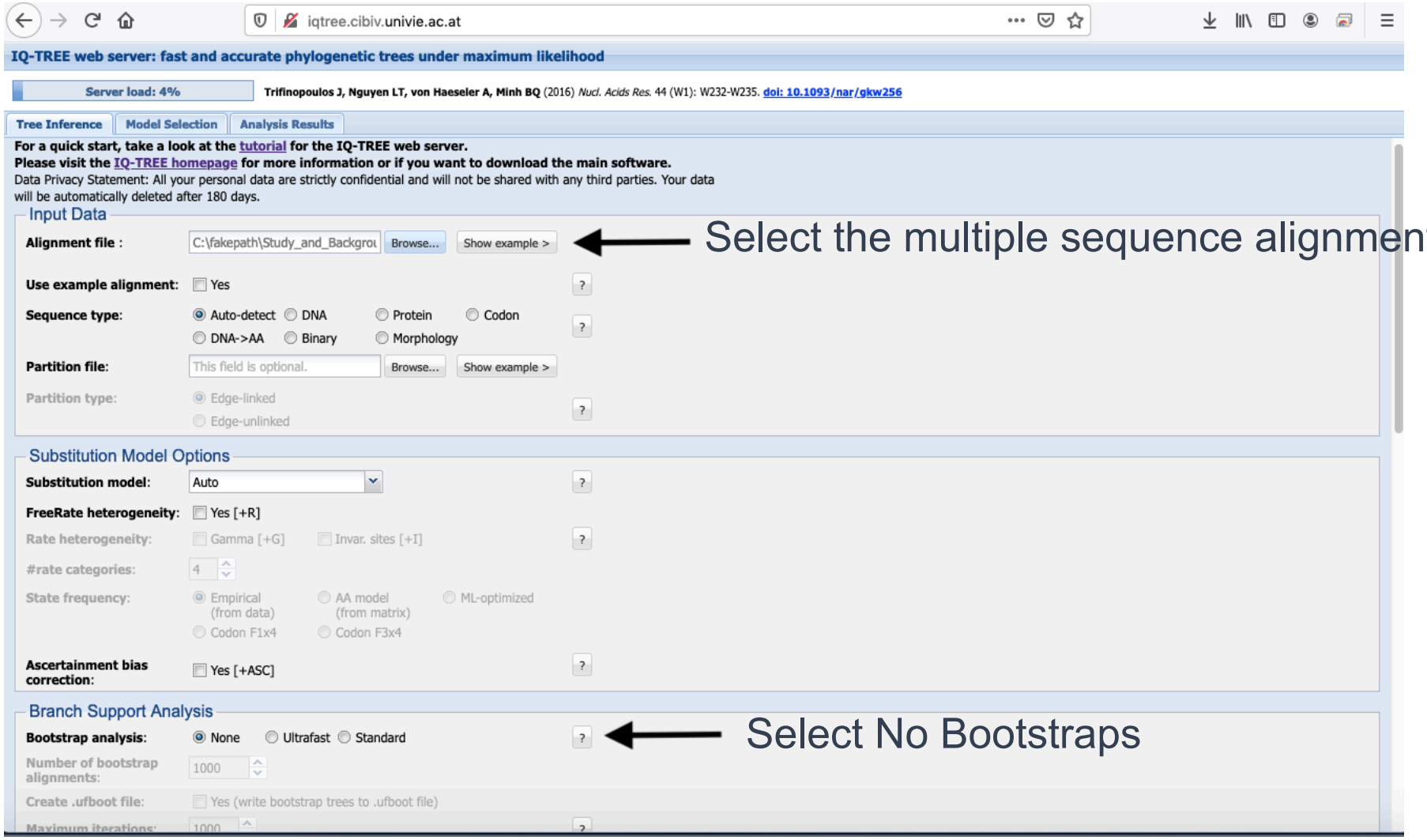
Maximum likelihood using FastTree

Fasttree –nt –gtr Study_and_Background_aled_wLoc_woDups_sub.fas > Study_and_Background_aled_wLoc_woDups_sub.tre



Maximum likelihood using IQ-Tree Webserver

http://iqtree.cibiv.univie.ac.at/



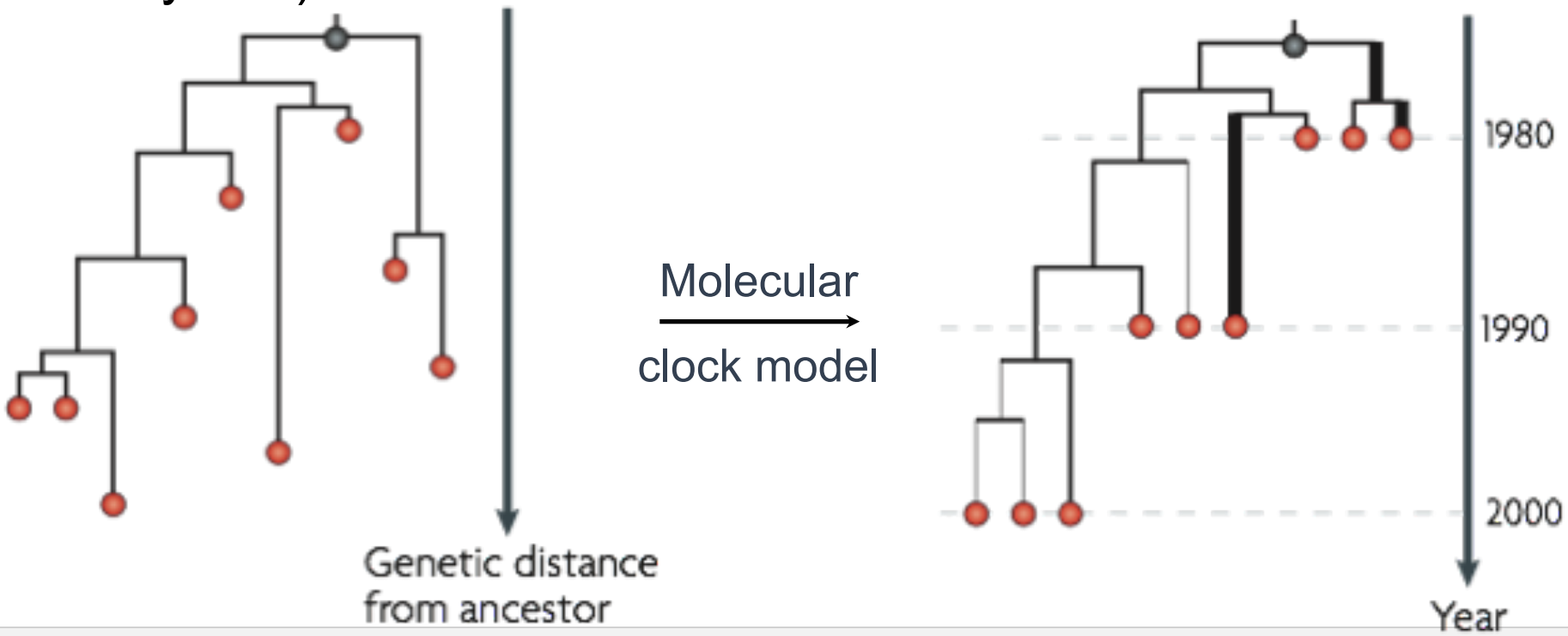
Select the multiple sequence alignment file

Select No Bootstraps

Phylogenetic Signal Considerations (TempEst)

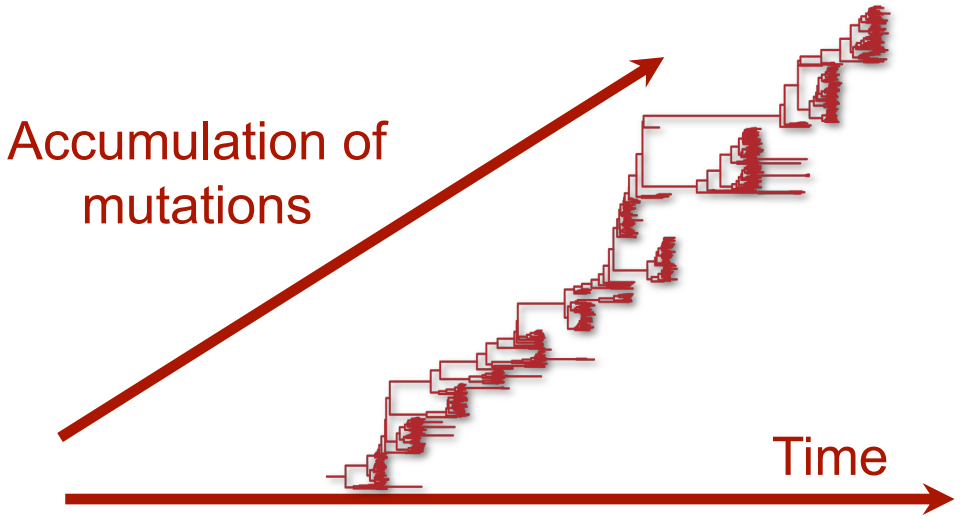
Temporal Signal

- Pathogen genomes are usually sampled at different points in time (heterochronous sequences)
- Transmission history is estimated on a real time-scale (e.g. days, months or years)



Temporal Signal

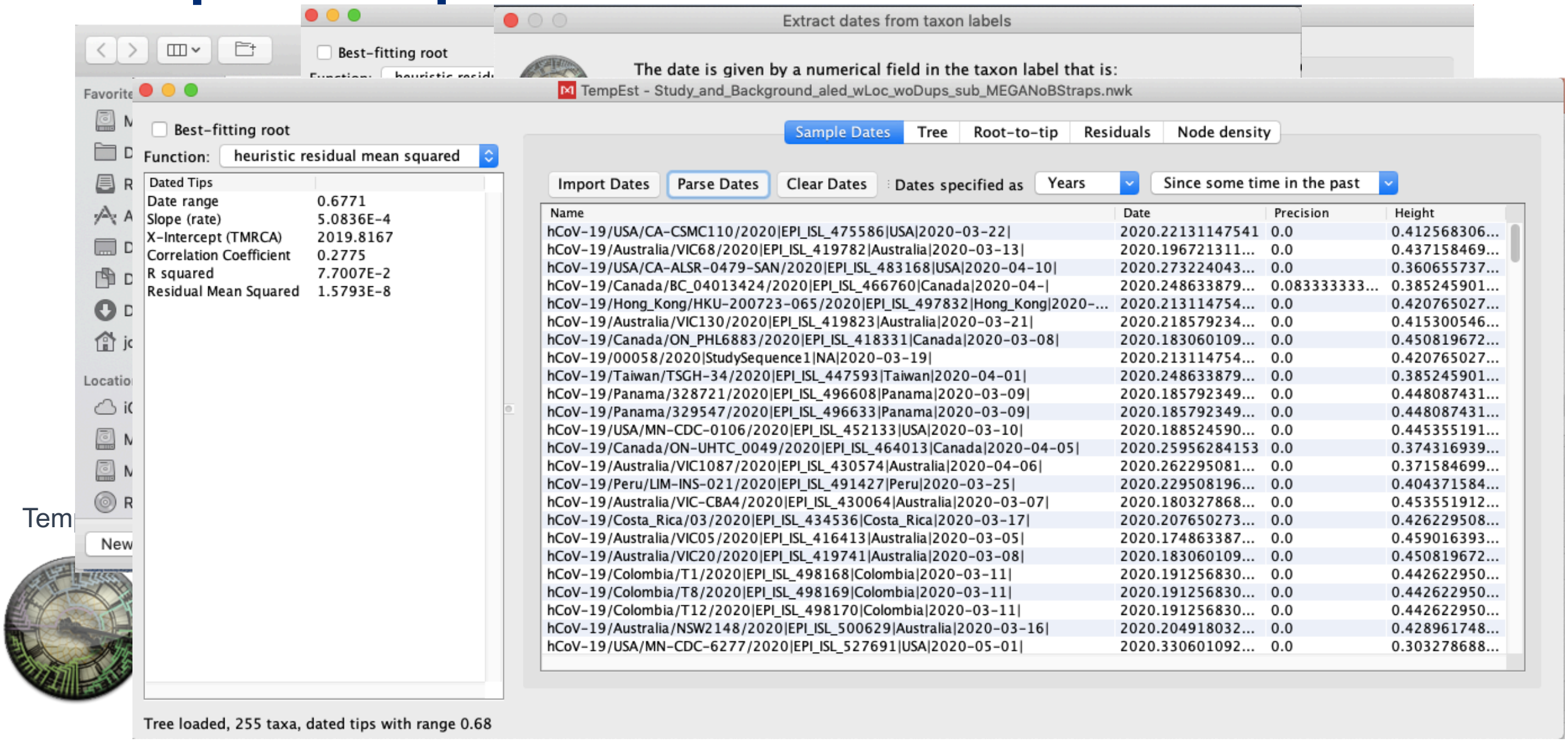
- Before building a time-scaled phylogenetic tree from heterochronous sequences, it is advisable to confirm that the sequences under investigation contain sufficient ‘temporal signal’ or ‘clockiness’ for reliable estimation.
- In other words, there must be sufficient genetic change between sampling times to reconstruct a statistical relationship between genetic divergence and time.



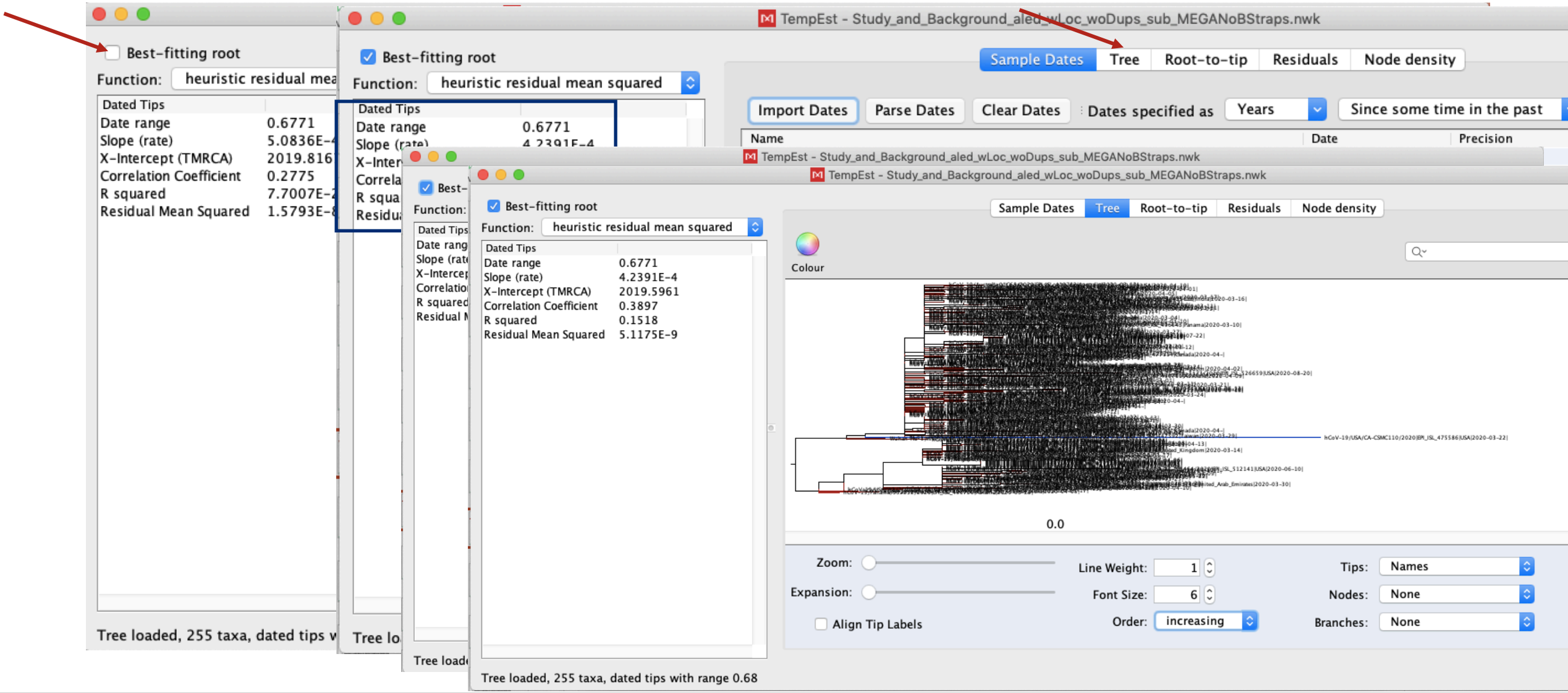
Temporal Signal

- The ability to genetically distinguish sequences sampled at different times depends on:
  - the rate of evolution of the gene / genome that is obtained
  - the length of time between samples
  - the sequence length of the gene/ genome that is obtained

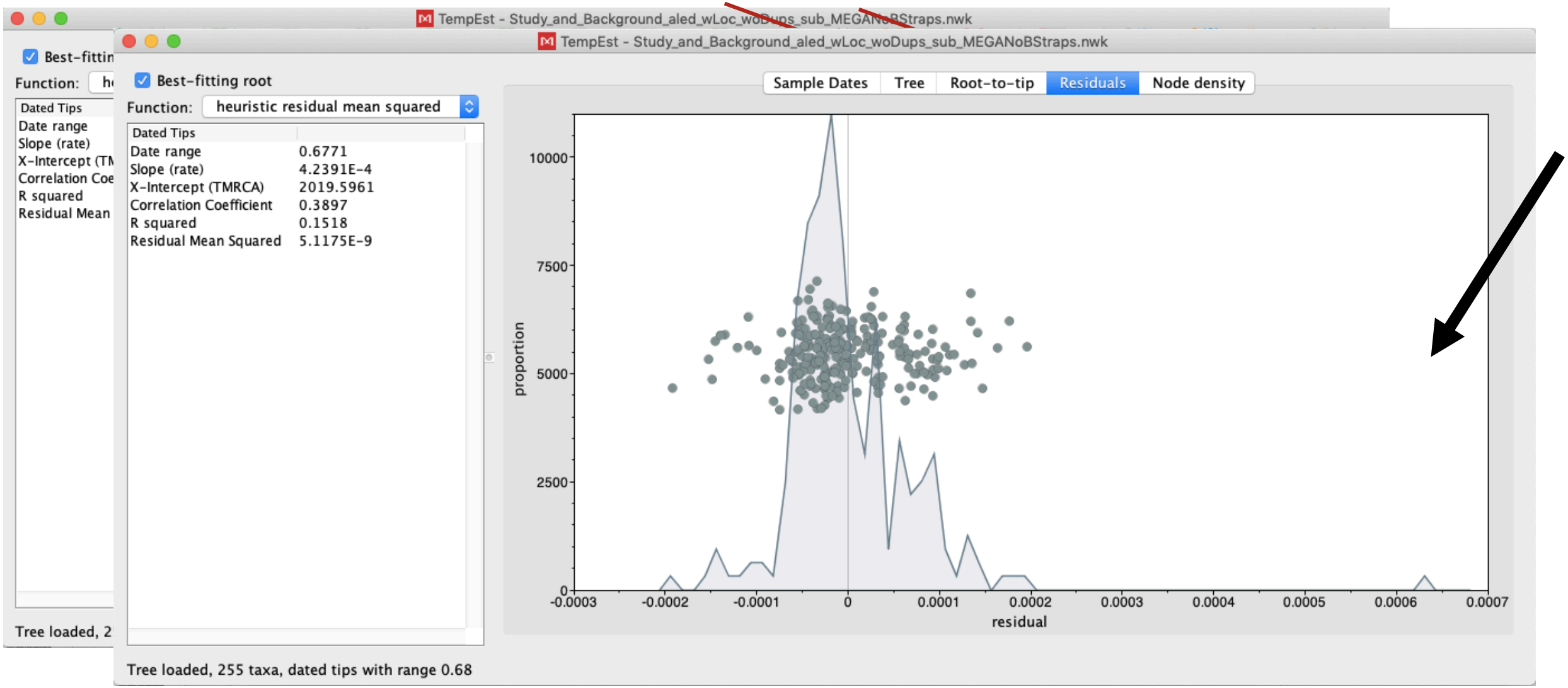
Open TempEst → Load Data → Parse Dates



Compute best-fitting root



Investigating root-to-tip divergence



Temporal Signal

- Regression of root-to-tip genetic distance against sampling time can be used as a simple diagnostic tool for molecular clock models.
- A linear trend with small residuals indicates that evolution will be adequately represented by a *strict* molecular clock. The same trend with greater scatter from the regression line suggests a *relaxed* molecular clock model may be most appropriate.
- A strong non-linear trend suggests that evolutionary rate has systematically changed through time.
- No trend at all indicates that the data contain little temporal signal and is unsuitable for inference using phylogenetic molecular clock models.

Temporal Signal

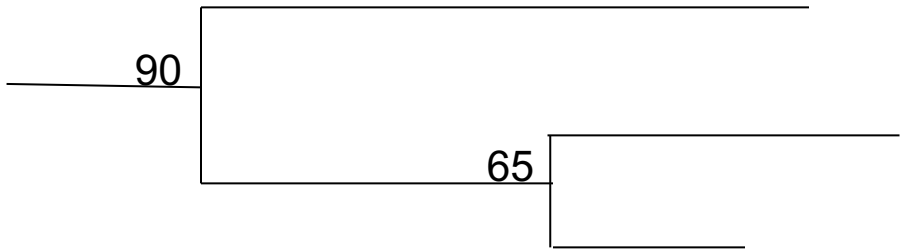
Note:

- Regressions of sampling time versus root-to-tip genetic distance can only be used to in an exploratory manner to investigate temporal signal and data quality in heterochronous alignments, and not for statistical hypothesis testing!

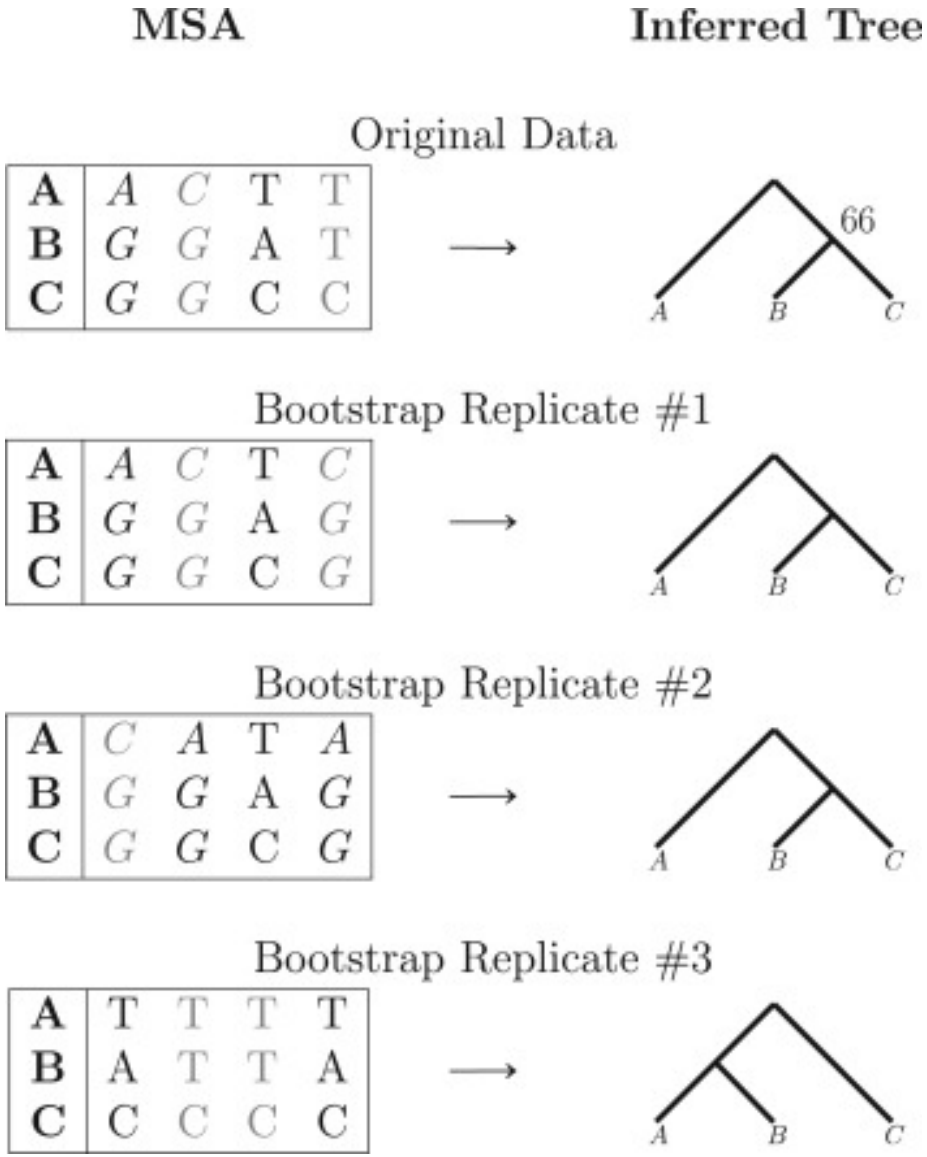
Understanding the Bootstrap

Bootstrapping (How Robust is the Tree?)

- Statistical technique that uses random resampling of data to determine sampling error.
- Gives an idea about the ‘reliability’ of branches and clusters.
- Characters are resampled with replacement to create many replicate data sets. A tree is then inferred from each replicate.
- Agreement among the resulting trees is summarized with a consensus tree. The frequencies of occurrence of groups, bootstrap proportions (BPs), are a measure of support for those groups
- Usually considered significant is higher than 70% (or 0.7 or 70/100)

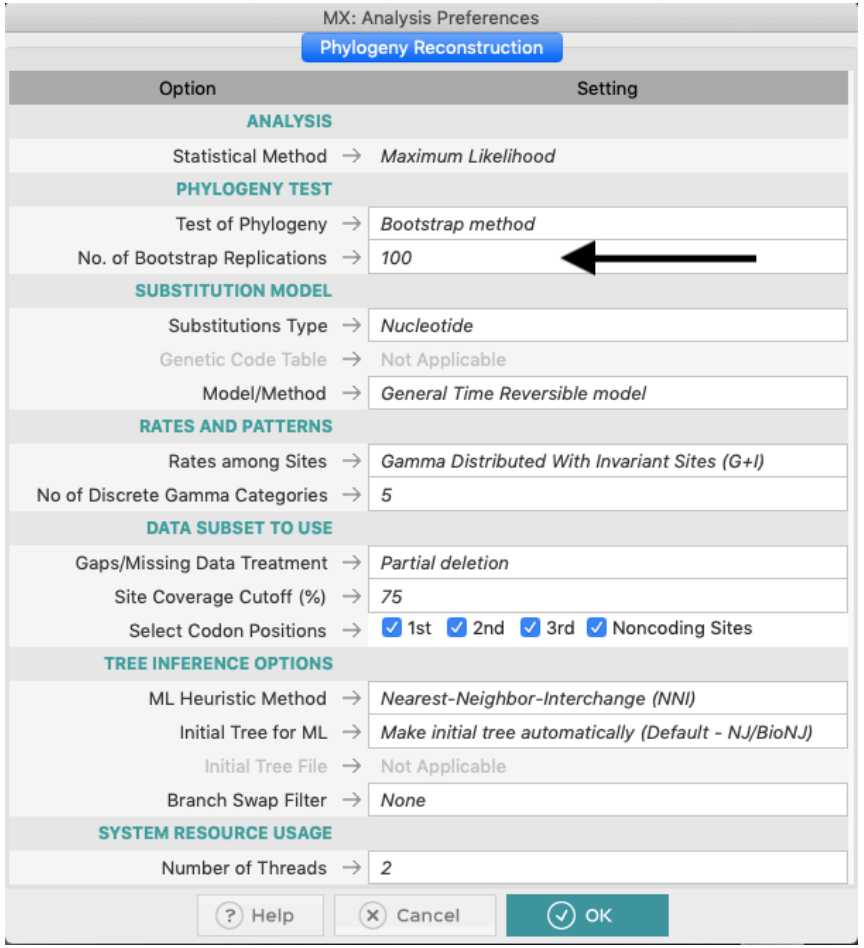


Bootstrapping (How Robust is the Tree?)



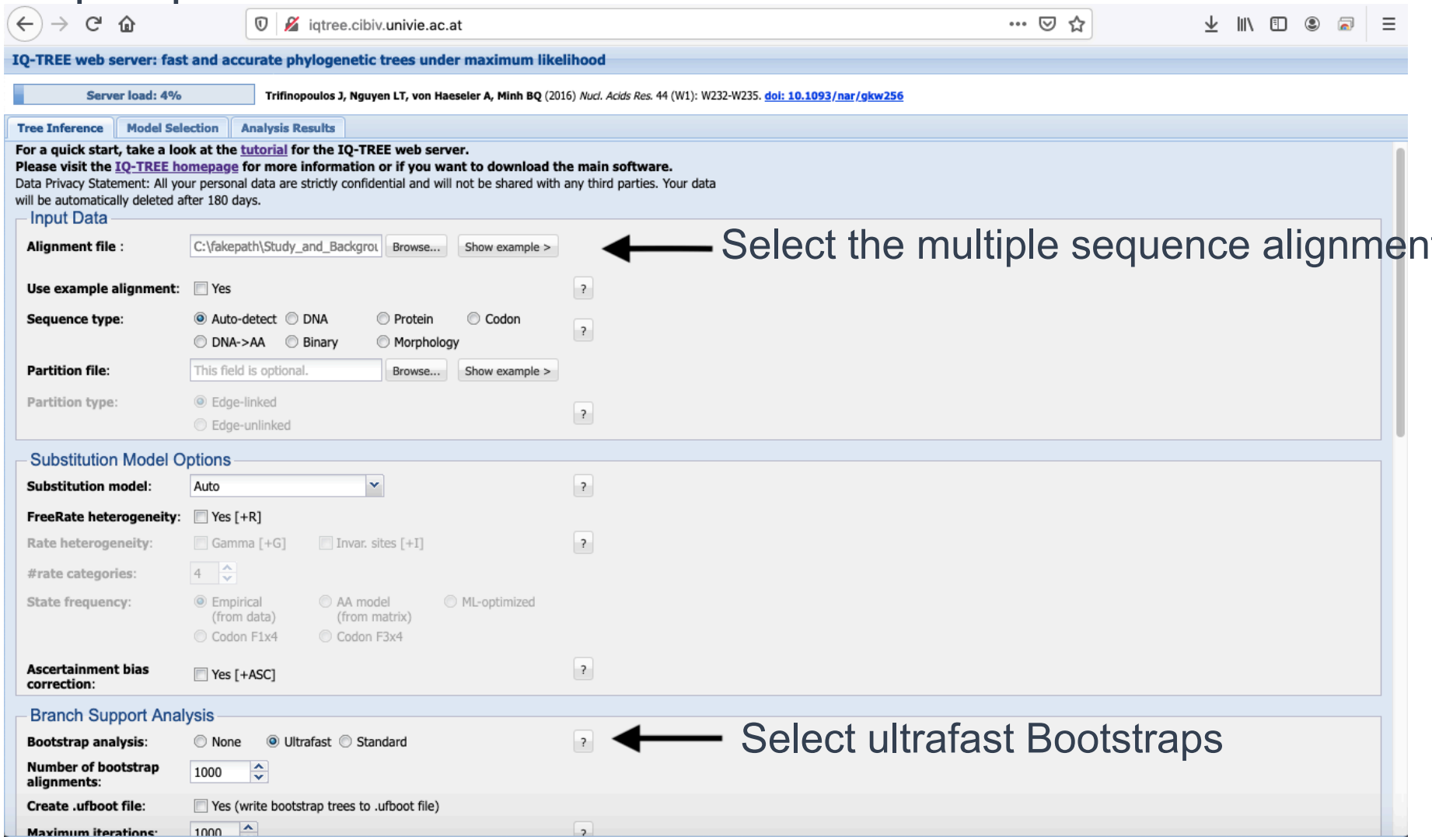
Non-Parametric Bootstrapping

- Repeat the previous tree building exercise with MEGA, but this time add bootstraps replications at 100



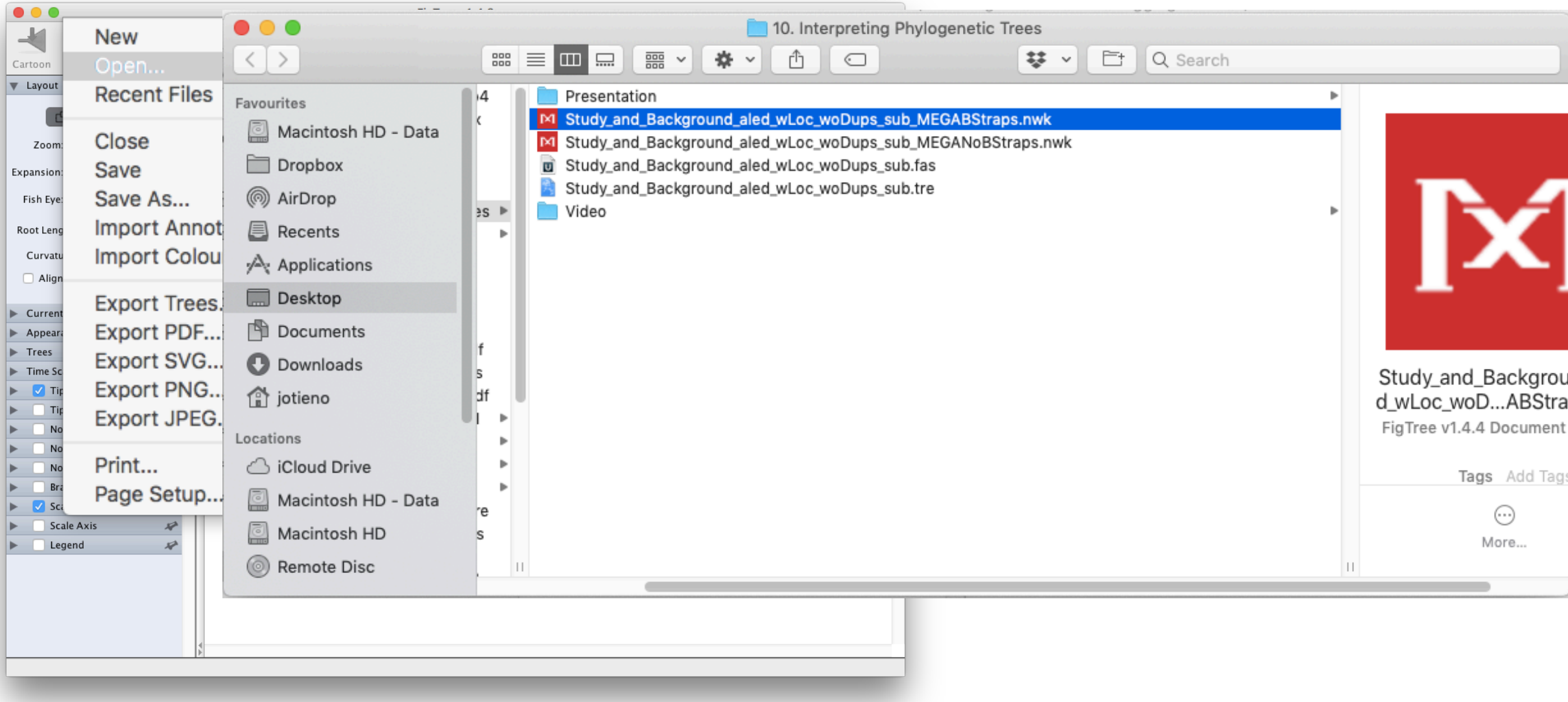
Maximum likelihood using IQ-Tree Webserver

http://iqtree.cibiv.univie.ac.at/

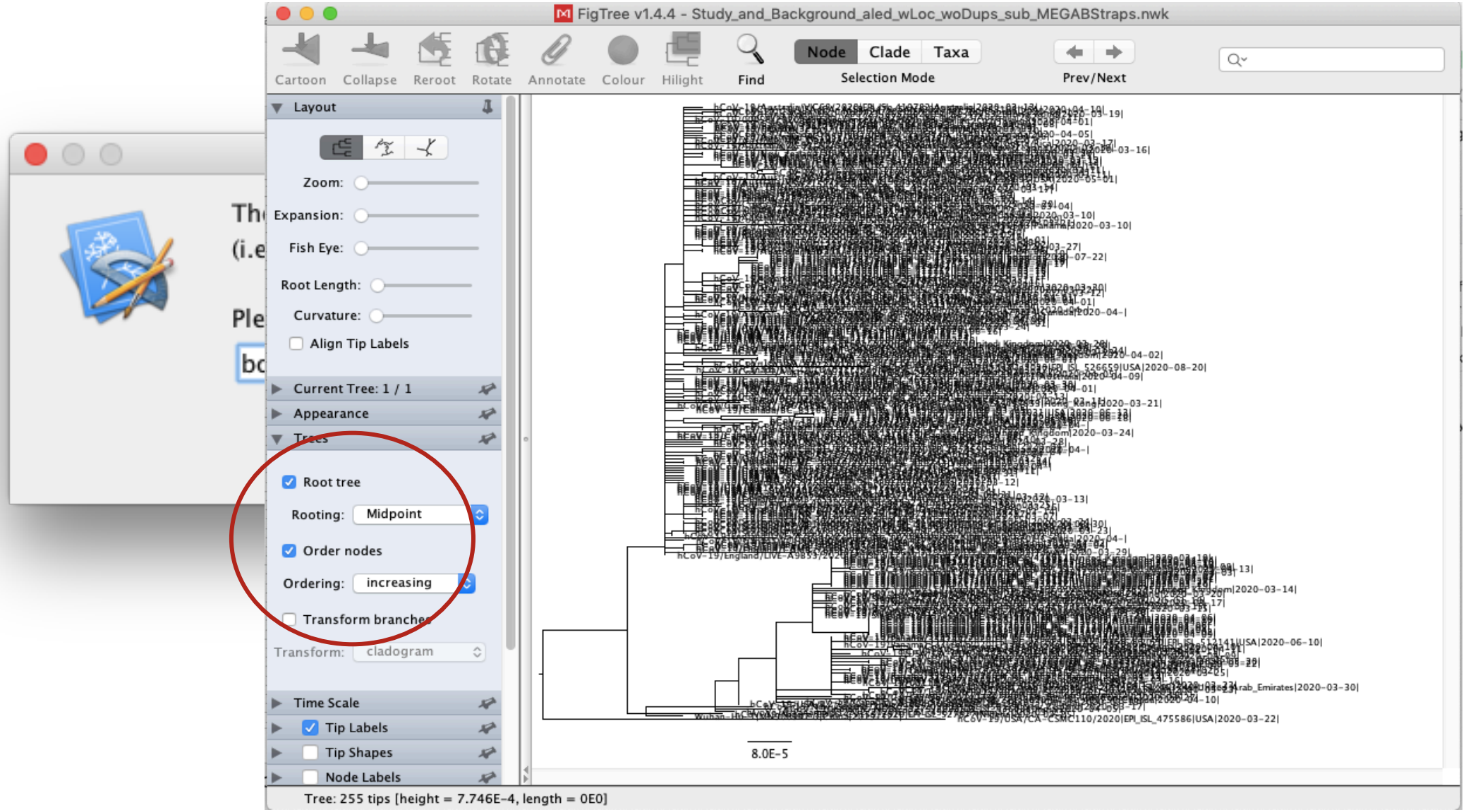


Visualizing Phylogenetic Trees in FigTree

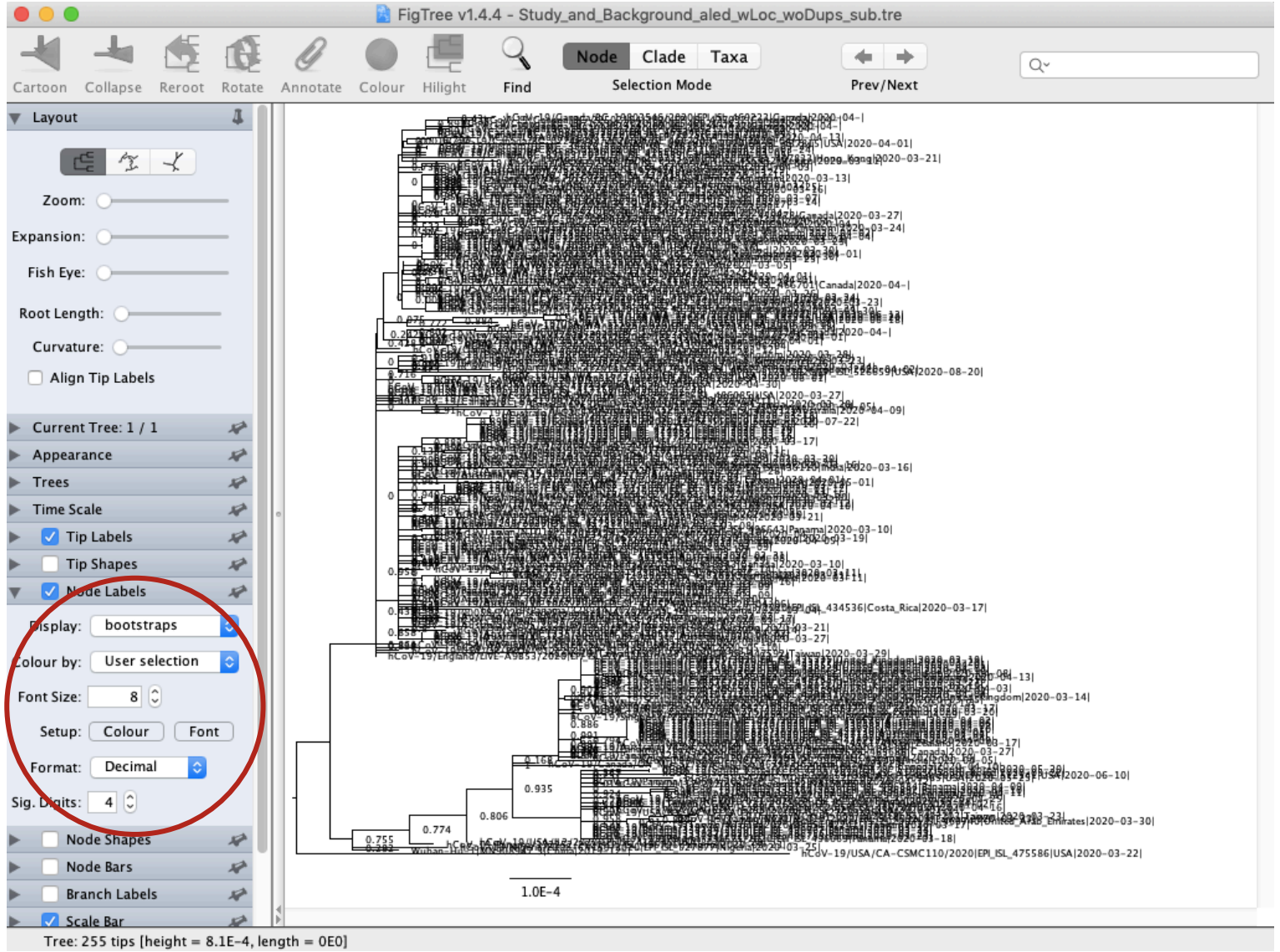
Open FigTree → Load Boostraped Tree



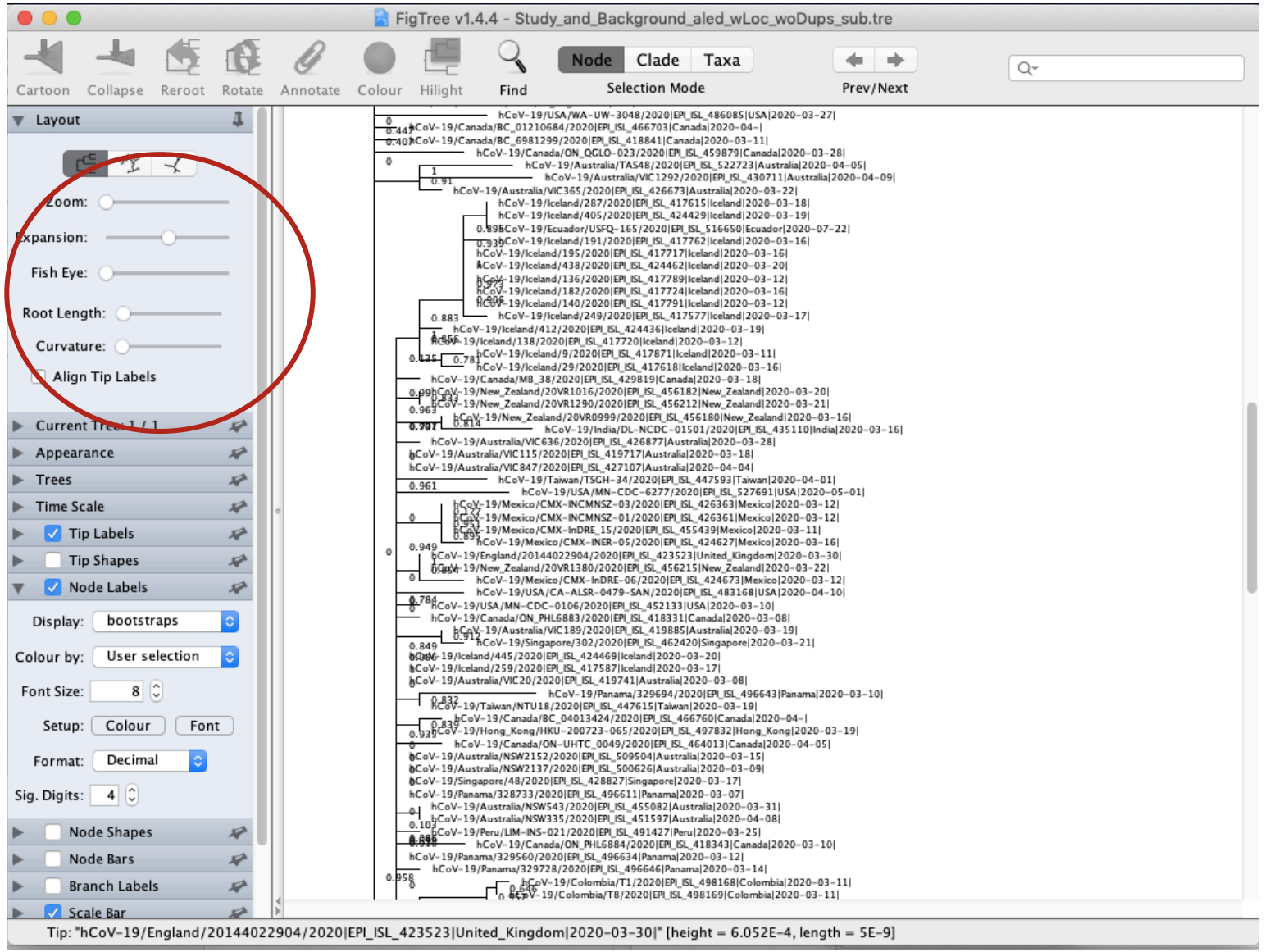
Trees: 1) Midpoint Root; 2) Increase nodes order



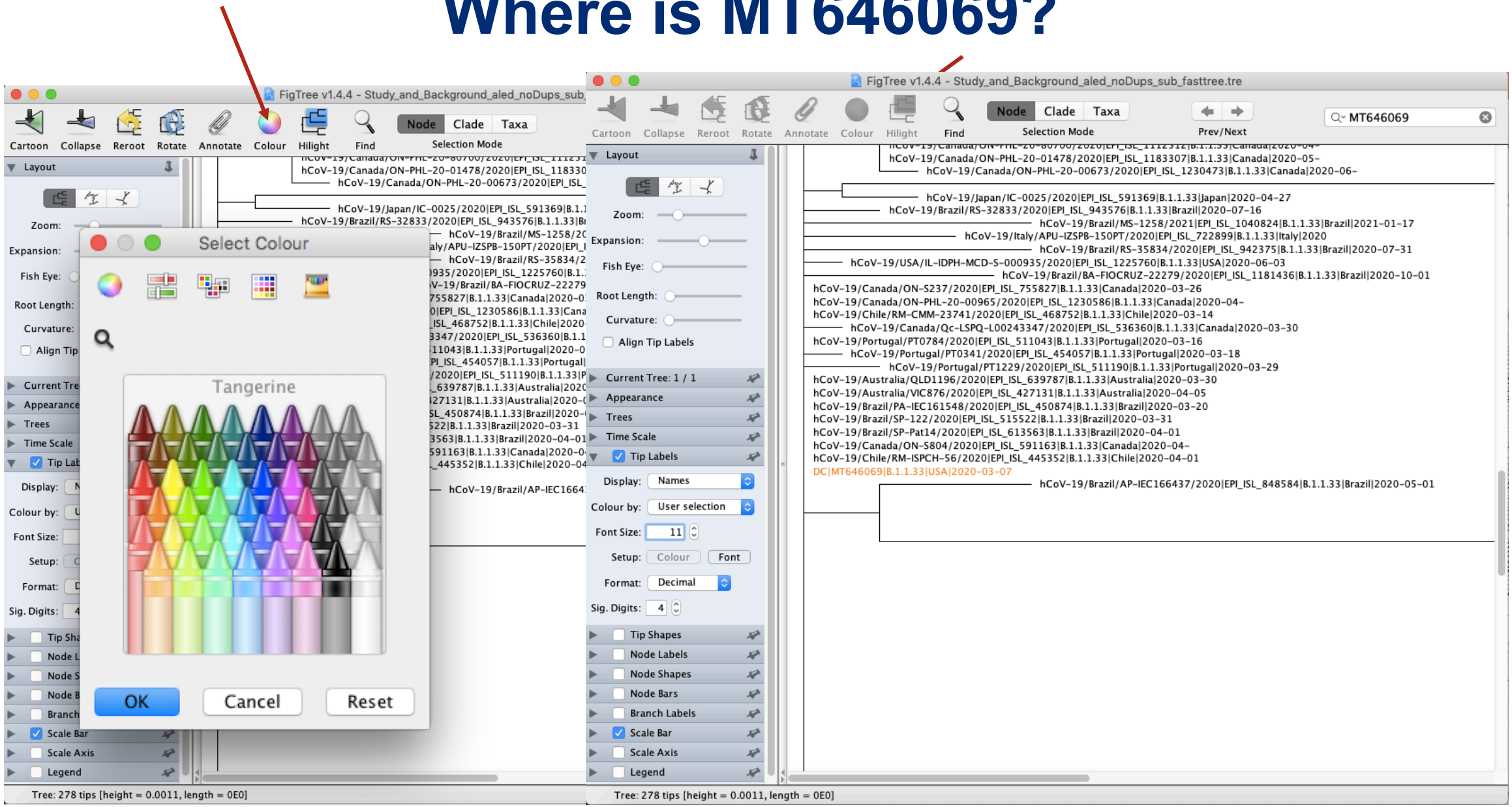
Add bootstrap support labels to the nodes



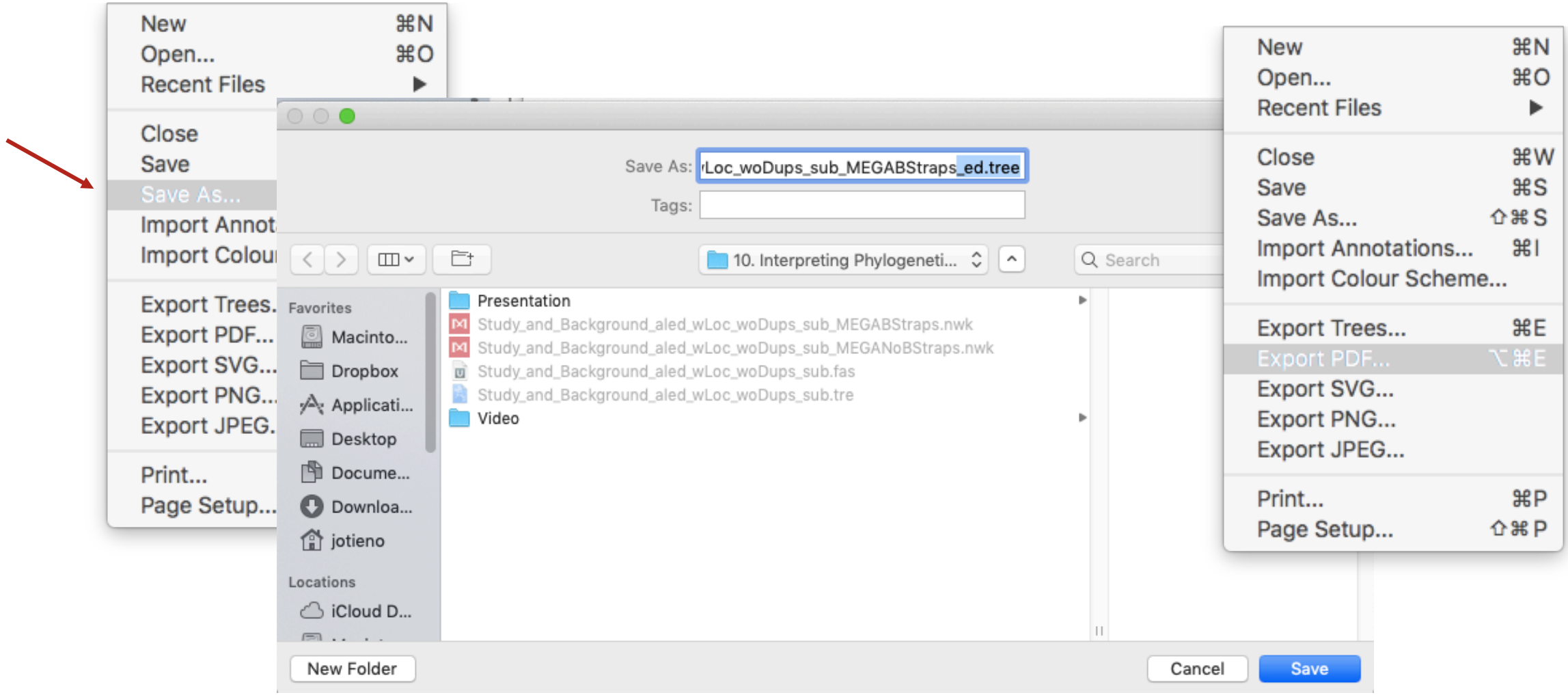
Expand tree to make it readable



Where is MT646069?



Save Tree and Export as a PDF

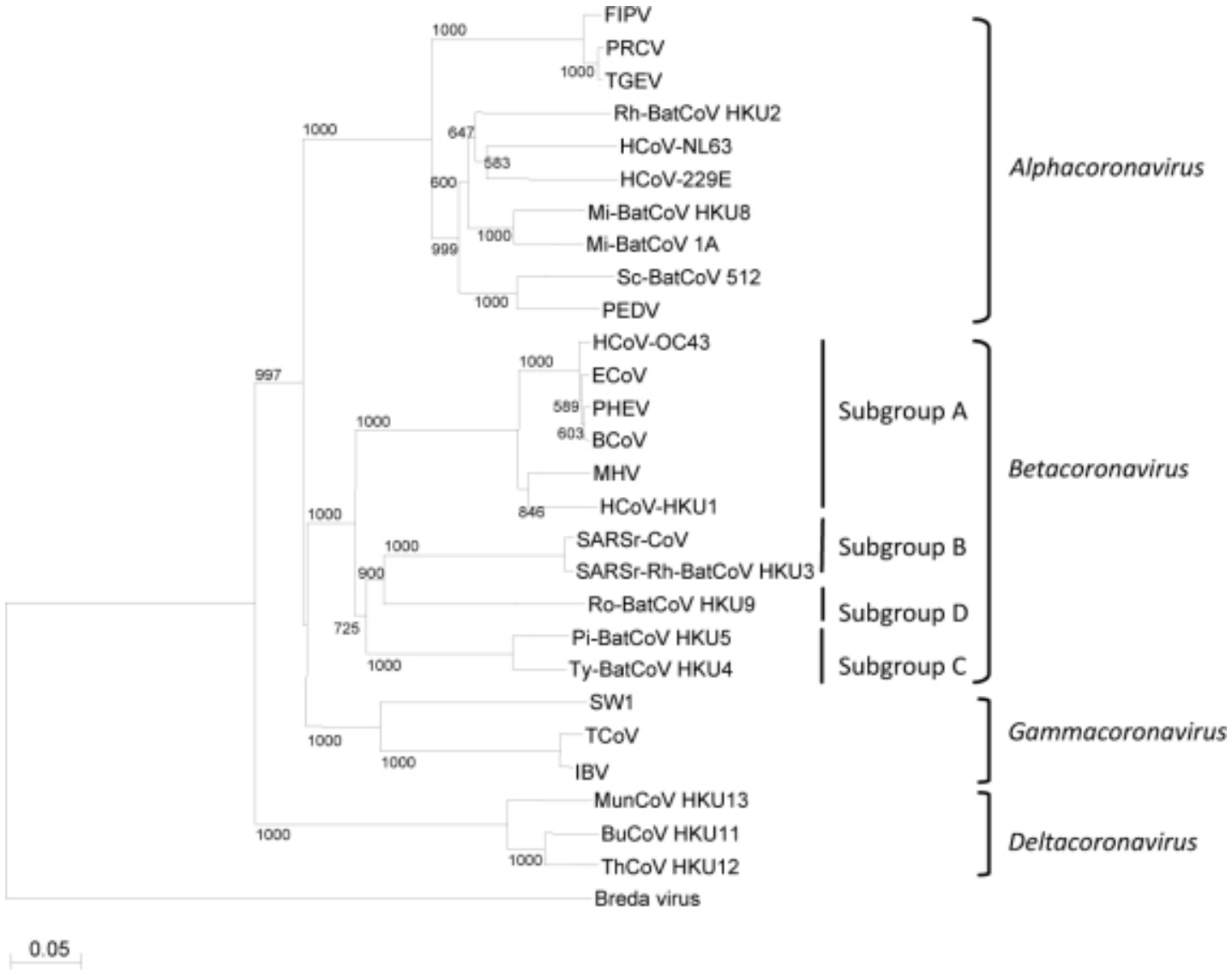


Basic Interpretation of Phylogenetic Trees

Building a tree is only half the challenge

Inferring phylogenies enables you to reconstruct ancestral relationships and recover hidden information

1. Pathogen Origins



Trees inferred from different gene regions

SARS-CoV-2

What can we infer from the four tree topologies?

Current Biology
Report
CellPress

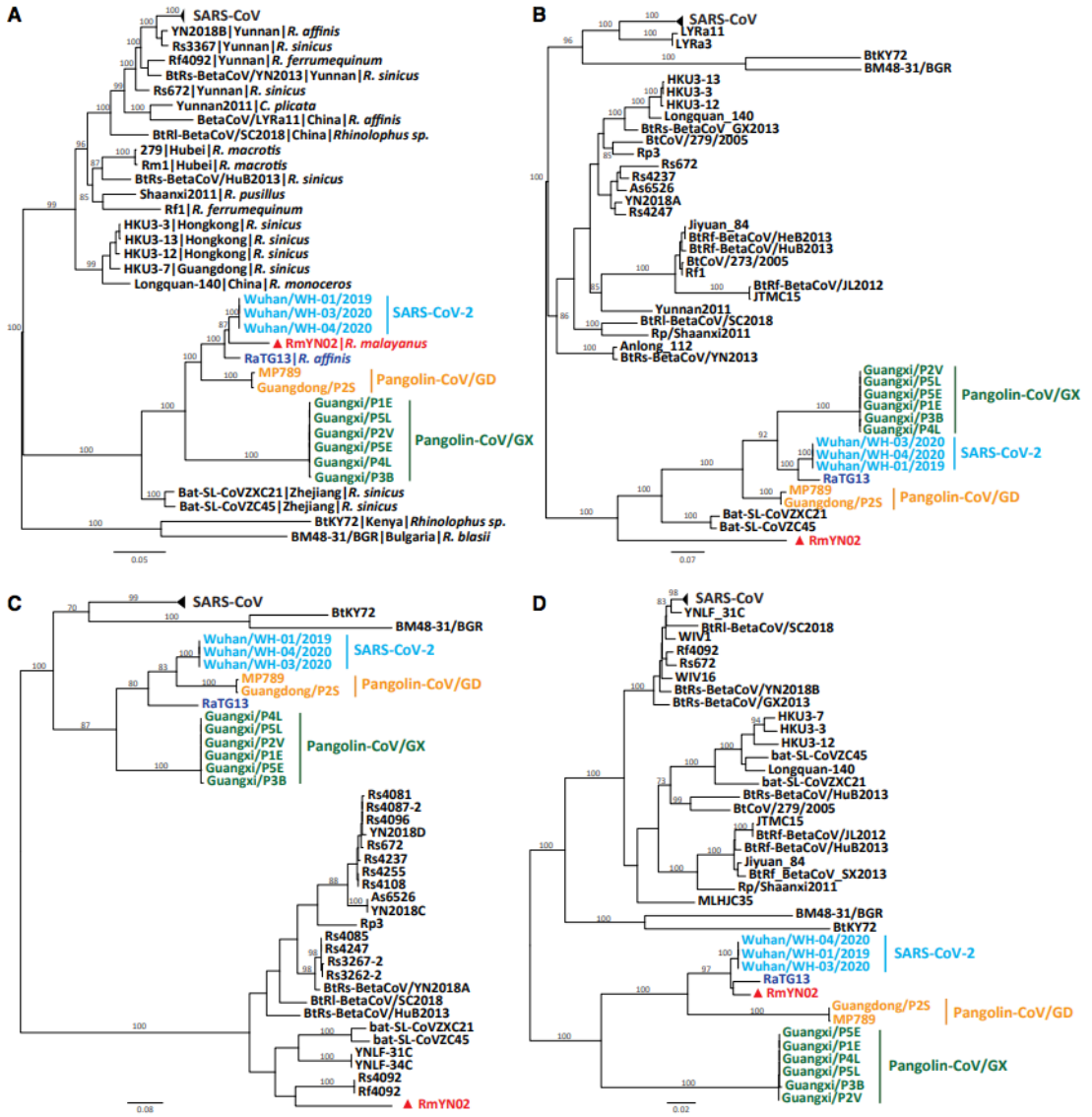
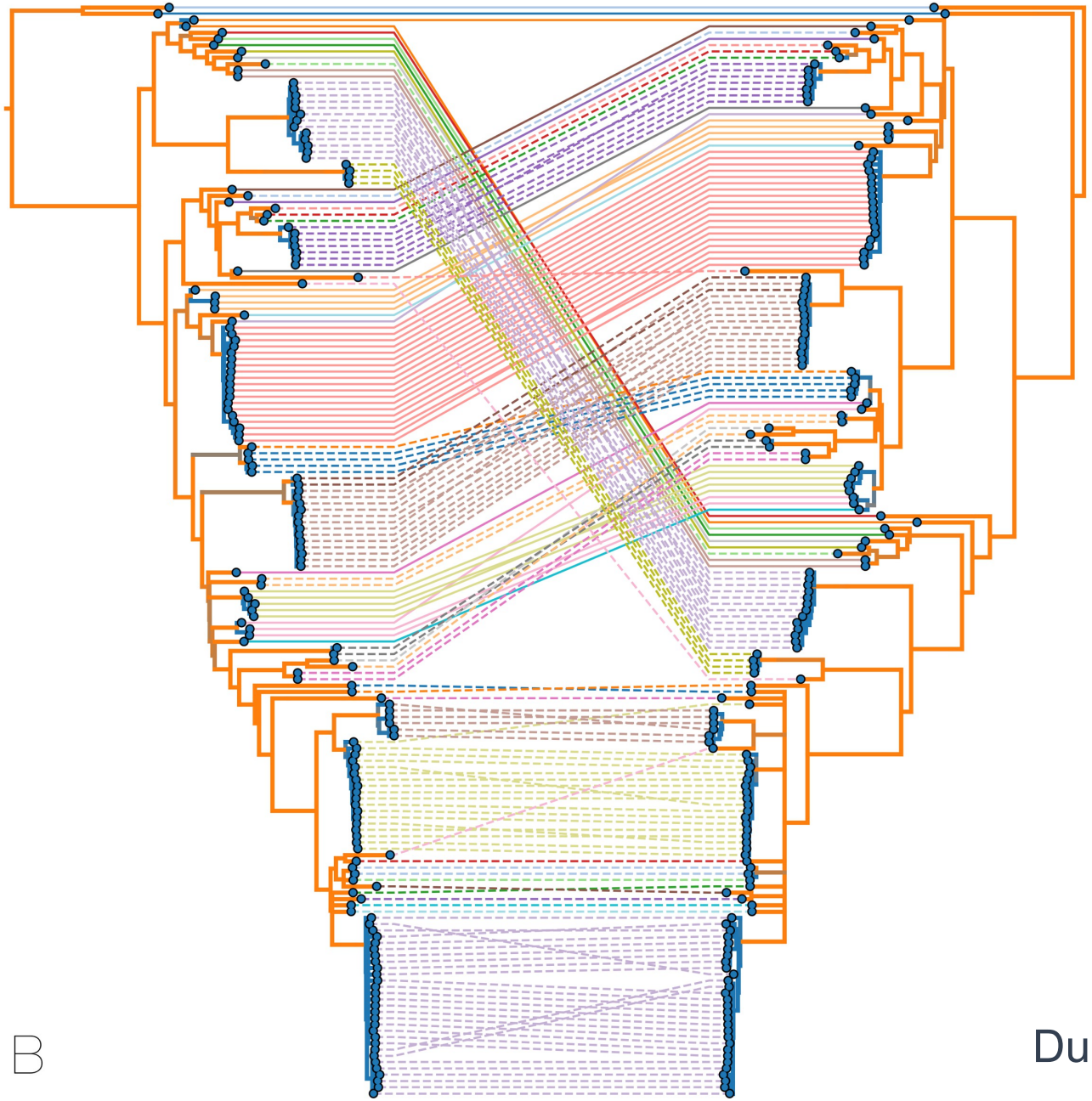


Figure 3. Phylogenetic Analysis of SARS-CoV-2 and Representative Viruses from the Subgenus *Sarbecoronavirus*

Trees inferred from different gene regions

MERS

Tanglegram

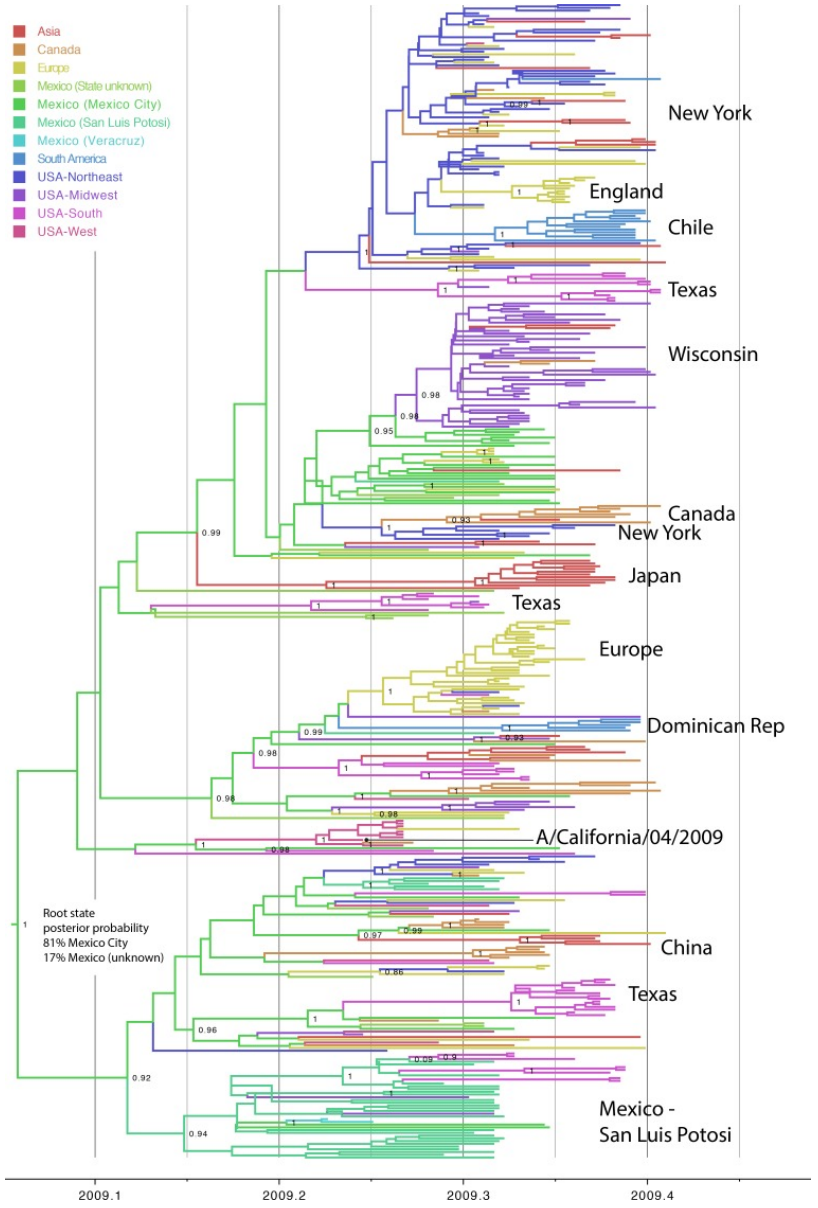


B

Dudas et al., 2018

2. Pathogen Transmission

Pandemic H1N1

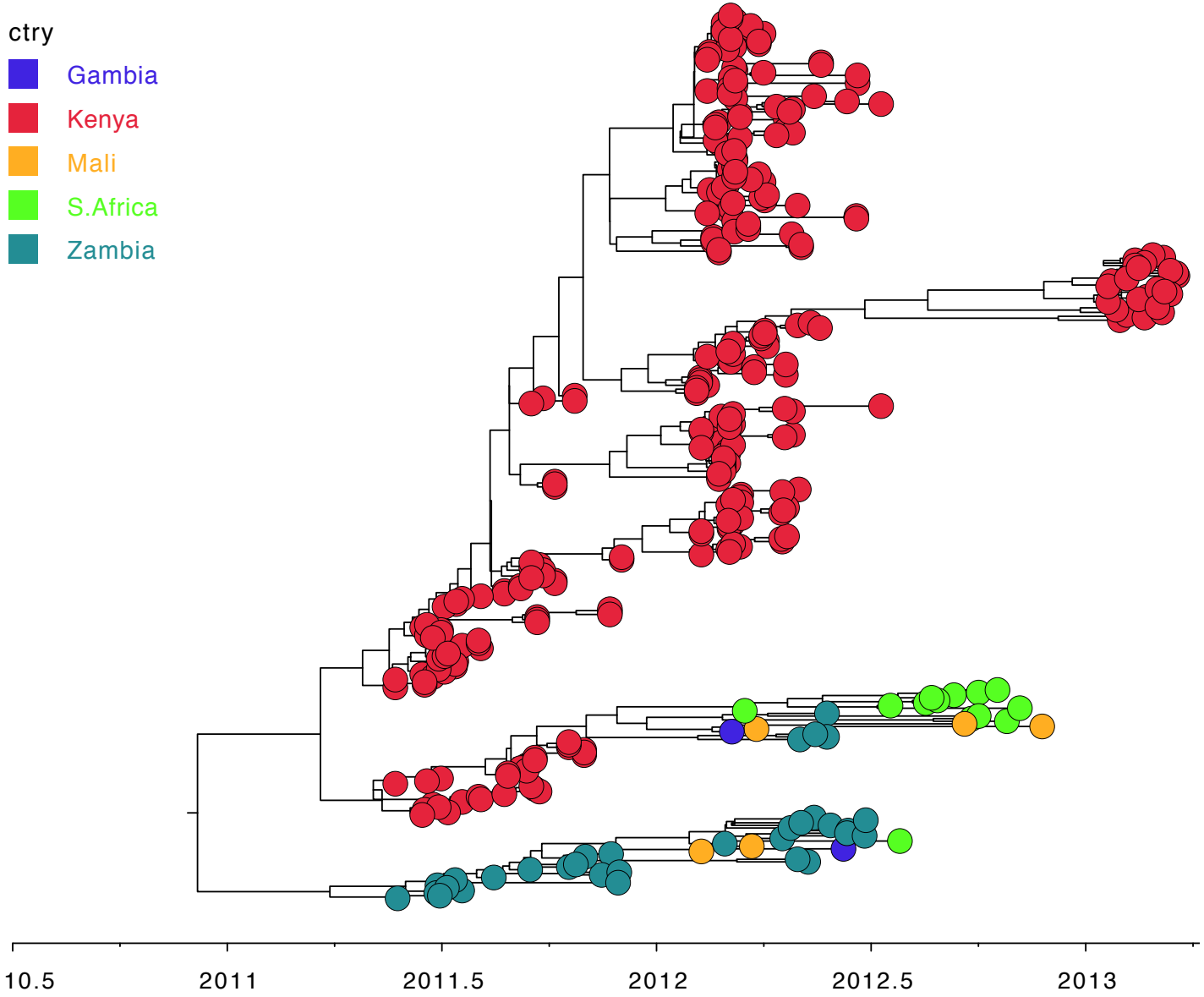


Ignacio Mena,
Martha I Nelson, et al., 2016

RSV genotype ON1

Africa

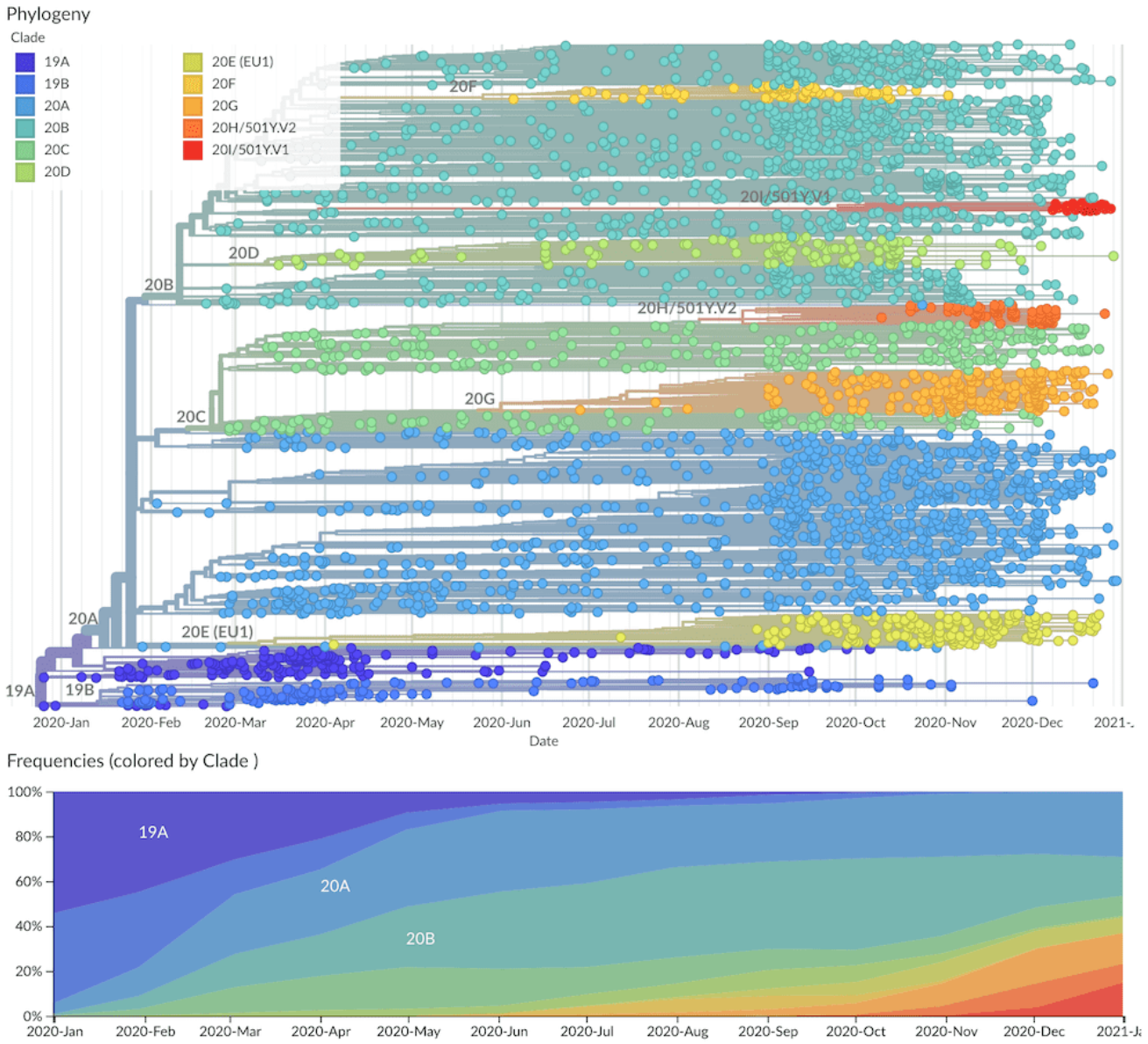
Where did the Mali RSV strains originate from?



Pathogen Evolution

SARS-CoV-2

Clades/Variants/Lineages



Branch lengths show you how many mutations occurred in the evolutionary time between lineages?

https://clades.nextstrain.org

Common Phylogenetic Tree Building Pitfalls

- Not using enough background data
- Not enough phylogenetic signal in dataset
- Annotation errors on sequence alignments and dates
- Not removing recombinants
- Not removing viruses with sequencing errors/contaminants
- Not questioning odd results
- Over-interpreting gaps in tree

Conclusions

- Inferring phylogenies enables you to reconstruct ancestral relationships and recover hidden information
- It is important to use a good tree-building method -- ML/Bayesian have many advantages

Building a tree is only half the challenge

- The real challenge is to interpret the tree
  - –When/where did a virus come from?
  - –How is it evolving/adapting/spreading spatially?
- New tools are allowing for more sophisticated interpretation of phylogenetic patterns

Resources

- https://evolution.berkeley.edu/evolibrary/article/0_0_0/evo_05
- https://artic.network/how-to-read-a-tree.html
- https://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956/


COV-IRT
COVID-19 International Research Team


NIH
Fogarty International Center
Advancing Science for Global Health