



Fogarty International Center
Advancing Science for Global Health



COVID-19 International Research Team

Phylogenetic Trees

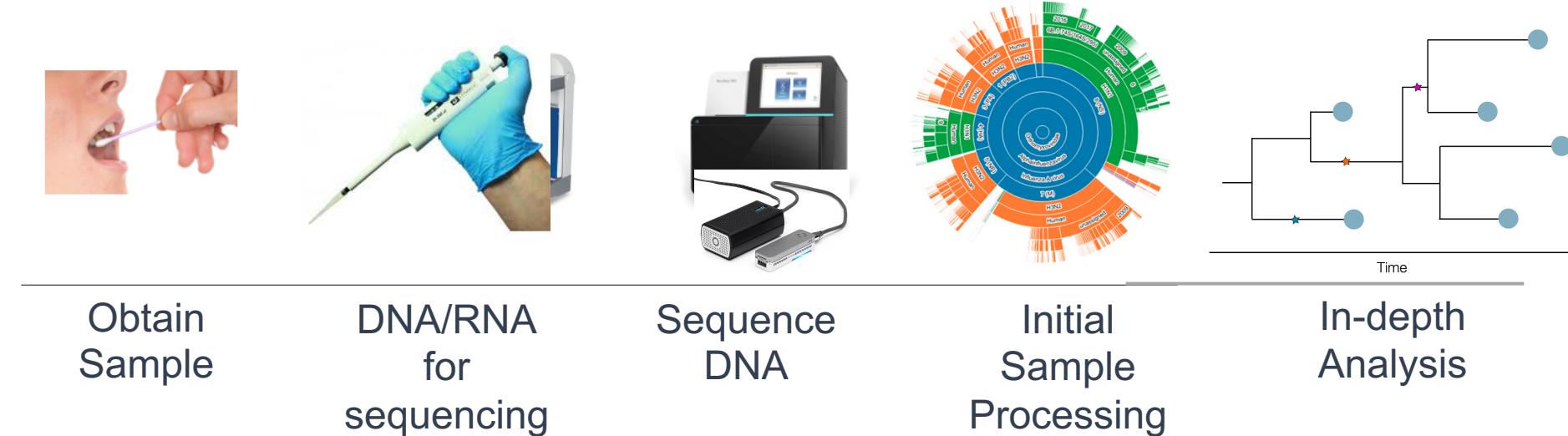
Methods and Interpretation

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

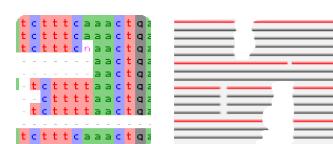


Background Dataset



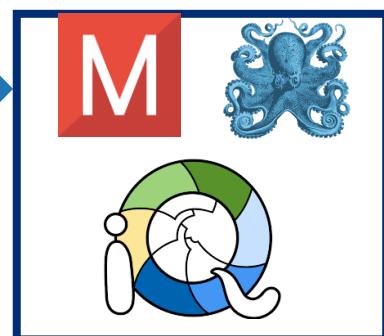
- Download from public database

Multiple Sequence Alignment

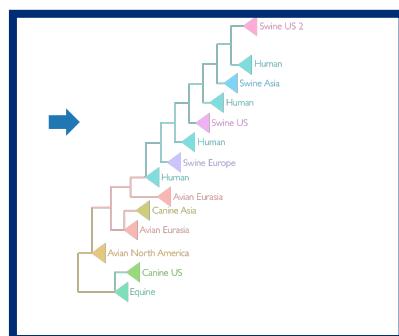


Sequence alignment
Clean and manually edition

Building a Phylogeny



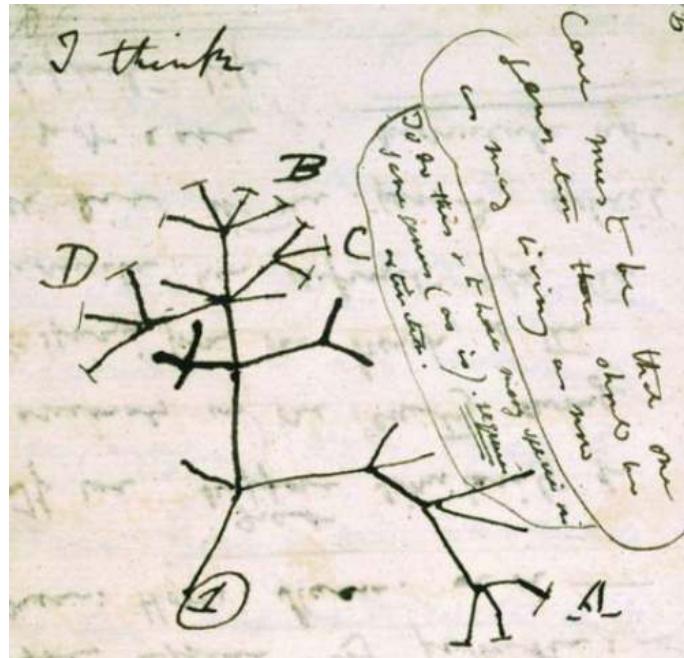
Interpreting a Phylogenetic Tree



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?

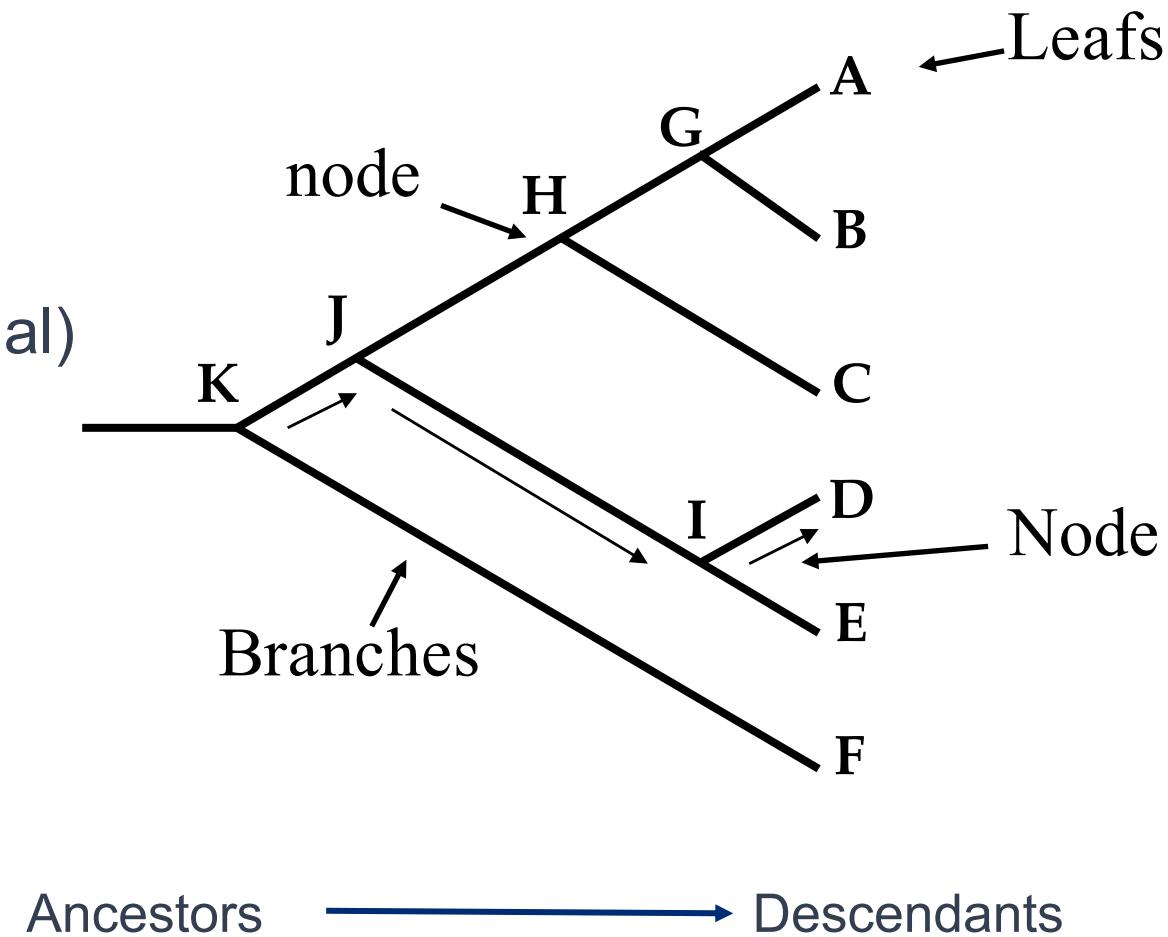
Leafs/Taxa/Tips:

A, B, C, D, E, F

Nodes (External/Internal)

G, H, I, J, K

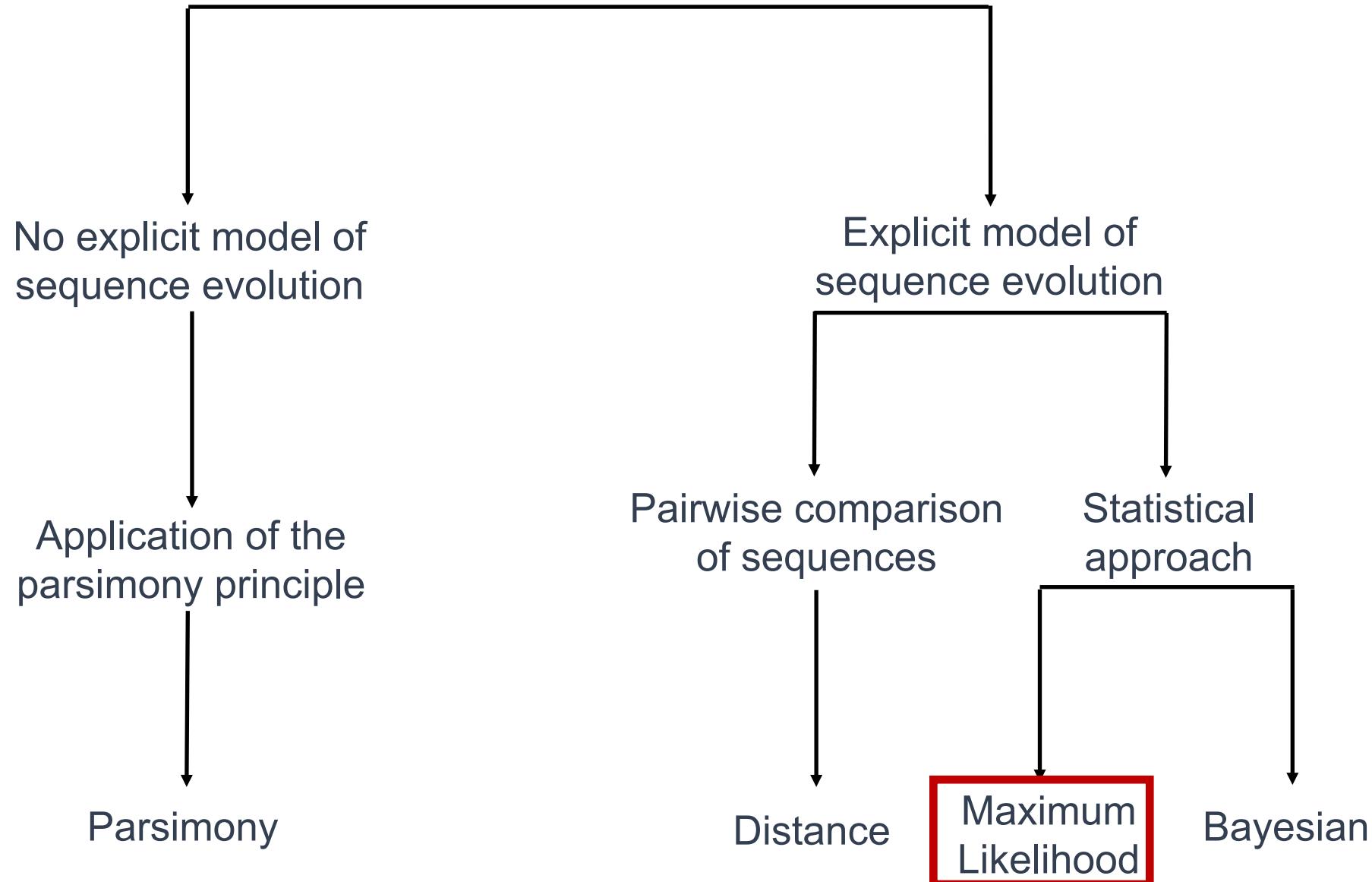
Branches



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

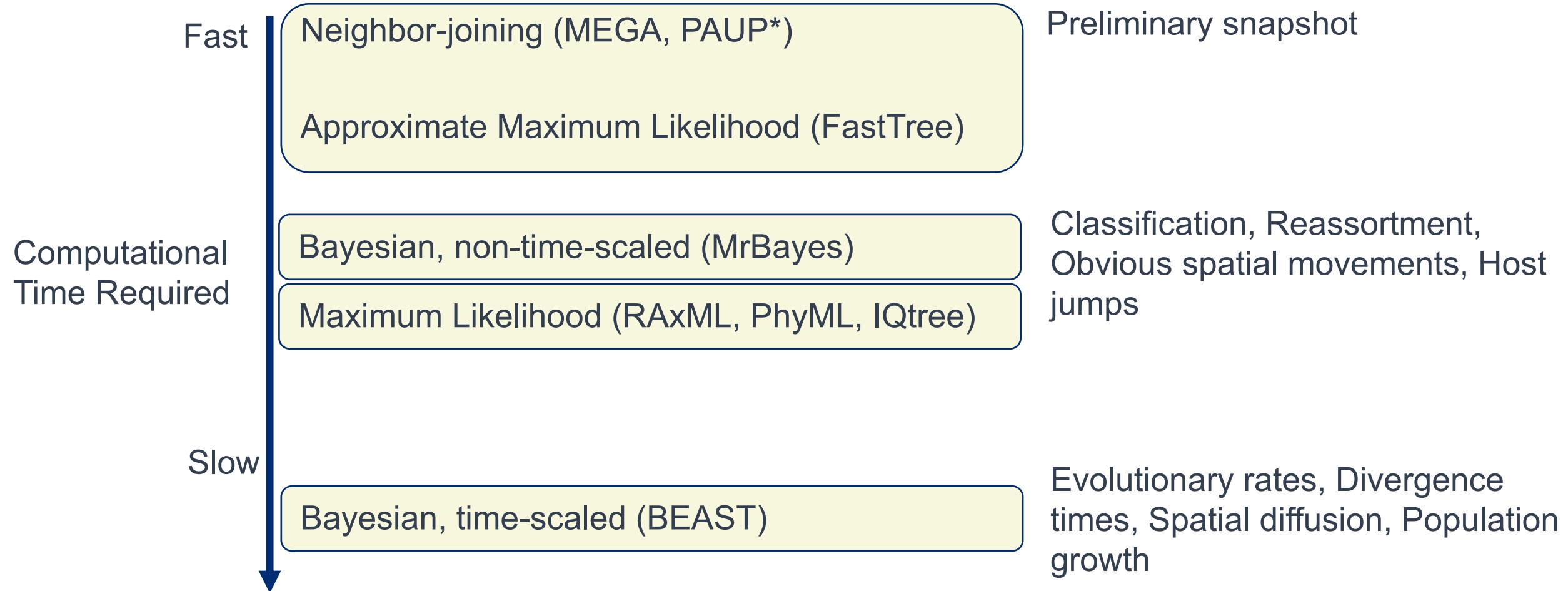
- *Maximum Likelihood (PAUP*, GARLi, PhyML, RaxML, FastTree, IQ-Tree)*

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

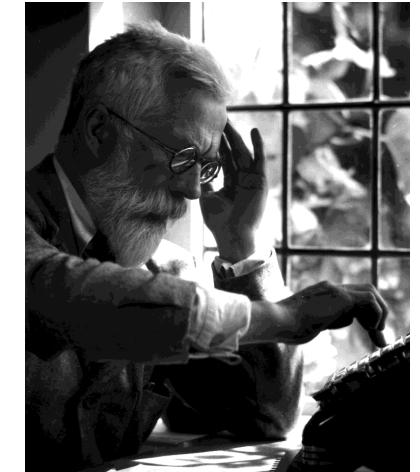
$$P(X | H) \text{ or } P(X | p)$$

- then we would talk about the likelihood

$$L(p | X)$$

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

The screenshot illustrates the workflow for creating a Maximum Likelihood (ML) tree in MEGA. It shows the main MEGA window, a file selection dialog, a data type dialog, a genetic code selection dialog, and a small pop-up window.

MEGA Logo: A large red square containing a white stylized 'M' logo.

Main MEGA Window: Shows the Molecular Evolutionary Genetics Analysis (MEGA) interface with various icons in the toolbar and a sidebar with 'RECENT PUBLICATIONS' and links like 'HELP DOCS', 'EXAMPLES', 'CITATION', 'REPORT BUG', 'UPDATES', 'MEGA LINKS', and 'TOOLBAR'.

File Selection Dialog: A modal window titled 'Open a File/Session...' with a blue background. It lists 'Favorites' (Presentation, Macintosh..., Dropbox, Recents, Applications, Desktop, Documents, Downloads, jotieno) and 'Locations' (Presentation). The file 'Study_and_Background_aled_wLoc_woDups_sub.fas' is selected. An arrow points to the 'Open' button in the bottom right corner.

Data Type Dialog: A modal window titled 'MX: Input Data'. It shows 'DATA TYPE' options: 'Nucleotide Sequences' (selected), 'Protein Sequences', and 'Pairwise Distance'. Other tabs include 'MISSING DATA', 'ALIGNMENT GAP', 'IDENTICAL SYMBOL', 'LOWER LEFT MATRIX', and 'UPPER RIGHT MATRIX'. An arrow points to the 'OK' button in the bottom right corner.

Genetic Code Selection Dialog: A modal window titled 'MX: Select Genetic Code'. It has tabs for 'Add', 'Delete', 'View', and 'Statistics'. Under 'SELECT A GENETIC CODE', the 'Standard' option is selected (radio button is checked). Other options listed are: Vertebrate Mitochondrial, Invertebrate Mitochondrial, Yeast Mitochondrial, Mold Mitochondrial, Protozoan Mitochondrial, Coelenterate Mitochondrial, Mycoplasma, Spiroplasma, and Ciliate Nuclear. An arrow points to the 'OK' button in the bottom right corner.

Small Pop-up Window: A small window titled 'How would you like to analyze your sequence?' with a red 'X' icon. It asks 'Analyze or Align File?' and has two buttons: 'Align' (blue border) and 'Analyze' (white). An arrow points to the 'Analyze' button.

Bottom Left: NIH Fogarty logo.

Bottom Right: Page number '12'.

Maximum likelihood (ML) tree in MEGA

Molecular Evolutionary Genetics Analysis

MX: Analysis Preferences

Phylogeny Reconstruction

Option	Setting
ANALYSIS	
Statistical Method	→ Maximum Likelihood
PHYLOGENY TEST	
Test of Phylogeny	→ None
No. of Bootstrap Replications	→ Not Applicable
SUBSTITUTION MODEL	
Substitutions Type	→ Nucleotide
Genetic Code Table	→ Not Applicable
Model/Method	→ General Time Reversible model
RATES AND PATTERNS	
Rates among Sites	→ Gamma Distributed With Invariant Sites
No of Discrete Gamma Categories	→ 5
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Partial deletion
Site Coverage Cutoff (%)	→ 75
Select Codon Positions	→ <input checked="" type="checkbox"/> 1st <input checked="" type="checkbox"/> 2nd <input checked="" type="checkbox"/> 3rd <input checked="" type="checkbox"/> Noncoding S
TREE INFERENCE OPTIONS	
ML Heuristic Method	→ Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	→ Make initial tree automatically (Default -)
Initial Tree File	→ Not Applicable
Branch Swap Filter	→ None
SYSTEM RESOURCE USAGE	
Number of Threads	→ 2

? Help ✖ Cancel ✓ OK

MX: Progress

PROGRESS

Setting parameters

DETAILS ✖ STOP

STATUS/OPTIONS

RUN STATUS

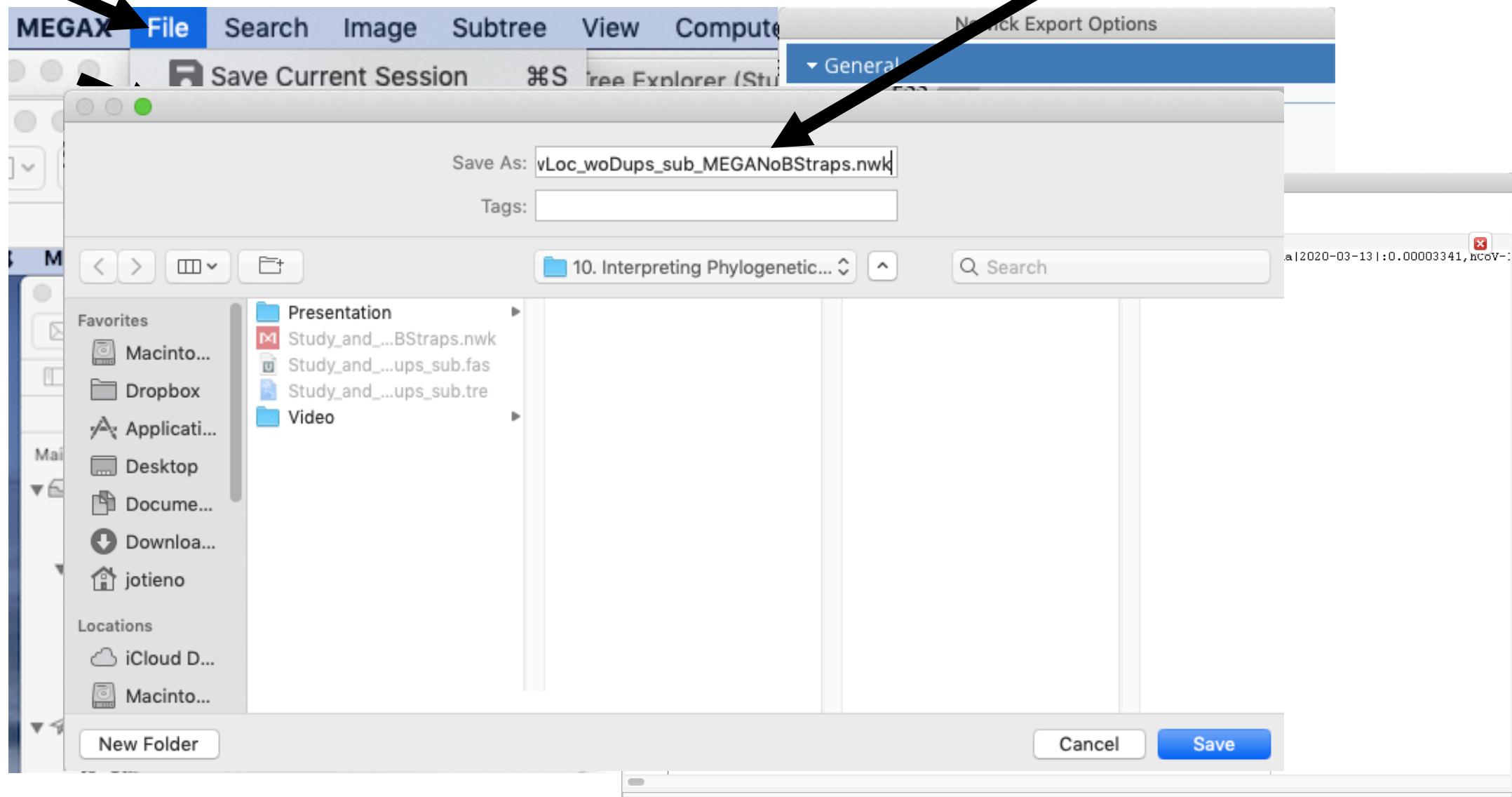
Start time | 25-9-20 10:23:26

Status | Setting site coverage

ANALYSIS OPTIONS

Statistical Method	:	Maximum Li
Phylogeny Test	:	
Test of Phylogeny	:	Bootstrap
No. of Bootstrap Replications	:	100
Substitution Model	:	
Substitutions Type	:	Nucleotide
Model/Method	:	General Ti
Rates and Patterns	:	
Rates among Sites	:	Gamma Dist
No of Discrete Gamma Categories	:	5
Data Subset to Use	:	
Gaps/Missing Data Treatment	:	Partial de
Site Coverage Cutoff (%)	:	75
Select Codon Positions	:	1st, 2nd, 3
Tree Inference Options	:	
ML Heuristic Method	:	Nearest-Ne
Initial Tree for ML	:	Make initi
Branch Swap Filter	:	None
System Resource Usage	:	
Number of Threads	:	2

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

```
[Jamess-Air:10. Interpreting Phylogenetic Trees — bash — 174x51
Jamess-Air:10. Interpreting Phylogenetic Trees jotieno$ fasttree -nt -gtr Study_and_Background_aled_wLoc_woDups_sub.fas > Study_and_Background_aled_wLoc_woDups_sub.tre
FastTree Version 2.1.10 Double precision (No SSE3)
Alignment: Study_and_Background_aled_wLoc_woDups_sub.fas
Nucleotide distances: Jukes-Cantor Joins: balanced Support: SH-like 1000
Search: Normal +NNI +SPR (2 rounds range 10) +ML-NNI opt-each=1
TopHits: 1.00*sqrtN close=default refresh=0.80
ML Model: Generalized Time-Reversible, CAT approximation with 20 rate categories
Ignored unknown character D (seen 7 times)
Ignored unknown character K (seen 55 times)
Ignored unknown character M (seen 21 times)
Ignored unknown character R (seen 38 times)
Ignored unknown character S (seen 13 times)
Ignored unknown character W (seen 25 times)
Ignored unknown character X (seen 24579 times)
Ignored unknown character Y (seen 84 times)
Initial topology in 4.80 seconds0 of 237 240 seqs (at seed 200)
Refining topology: 32 rounds ME-NNIs, 2 rounds ME-SPRs, 16 rounds ML-NNIs
Total branch-length 0.011 after 27.15 sec, 201 of 238 splits, 0 changes ax delta 0.000)

WARNING! This alignment consists of closely-related and very-long sequences.
WARNING! FastTree (or other standard maximum-likelihood tools)
may not be appropriate for alignments of very closely-related sequences
like this one, as FastTree does not account for recombination or gene conversion

ML-NNI round 1: LogLk = -45621.074 NNIs 116 max delta 9.11 Time 38.60es (max delta 9.111)
GTR Frequencies: 0.2985 0.1837 0.1964 0.3214ep 12 of 12
GTR rates(ac ag at cg ct gt) 0.2925 1.0219 0.3099 0.2951 2.9113 1.0000
Switched to using 20 rate categories (CAT approximation)20 of 20
Rate categories were divided by 0.626 so that average rate = 1.0
CAT-based log-likelihoods may not be comparable across runs
Use -gamma for approximate but comparable Gamma(20) log-likelihoods
ML-NNI round 2: LogLk = -44415.058 NNIs 76 max delta 1.24 Time 114.90es (max delta 1.241)
ML-NNI round 3: LogLk = -44415.015 NNIs 40 max delta 0.00 Time 129.46es (max delta 0.000)
Turning off heuristics for final round of ML NNIs (converged)
ML-NNI round 4: LogLk = -44414.441 NNIs 37 max delta 0.30 Time 148.06 (final)delta 0.302
Optimize all lengths: LogLk = -44414.134 Time 153.08
Total time: 210.75 seconds Unique: 240/255 Bad splits: 0/237 rnal splits
Jamess-Air:10. Interpreting Phylogenetic Trees jotieno$ ]
```

Maximum likelihood using IQ-Tree Webserver

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

The screenshot shows the IQ-TREE web server interface. At the top, the URL is iqtree.cibiv.univie.ac.at. Below the header, there are tabs for Tree Inference, Model Selection, and Analysis Results. A message encourages users to look at the tutorial and visit the homepage for more information. It also includes a Data Privacy Statement.

Input Data:

- Alignment file: C:\fakepath\Study_and_Backgrou (with a 'Browse...' button and a 'Show example >' link). An arrow points to this field with the text "Select the multiple sequence alignment file".
- Use example alignment: Yes
- Sequence type:
 - Auto-detect
 - DNA
 - Protein
 - Codon
 - DNA->AA
 - Binary
 - Morphology
- Partition file: This field is optional. (with a 'Browse...' button and a 'Show example >' link)
- Partition type:
 - Edge-linked
 - Edge-unlinked

Substitution Model Options:

- Substitution model: Auto
- FreeRate heterogeneity: Yes [+R]
- Rate heterogeneity:
 - Gamma [+G]
 - Invar. sites [+I]
- #rate categories: 4
- State frequency:
 - Empirical (from data)
 - AA model (from matrix)
 - ML-optimized
 - Codon F1x4
 - Codon F3x4
- Ascertainment bias correction: Yes [+ASC]

Branch Support Analysis:

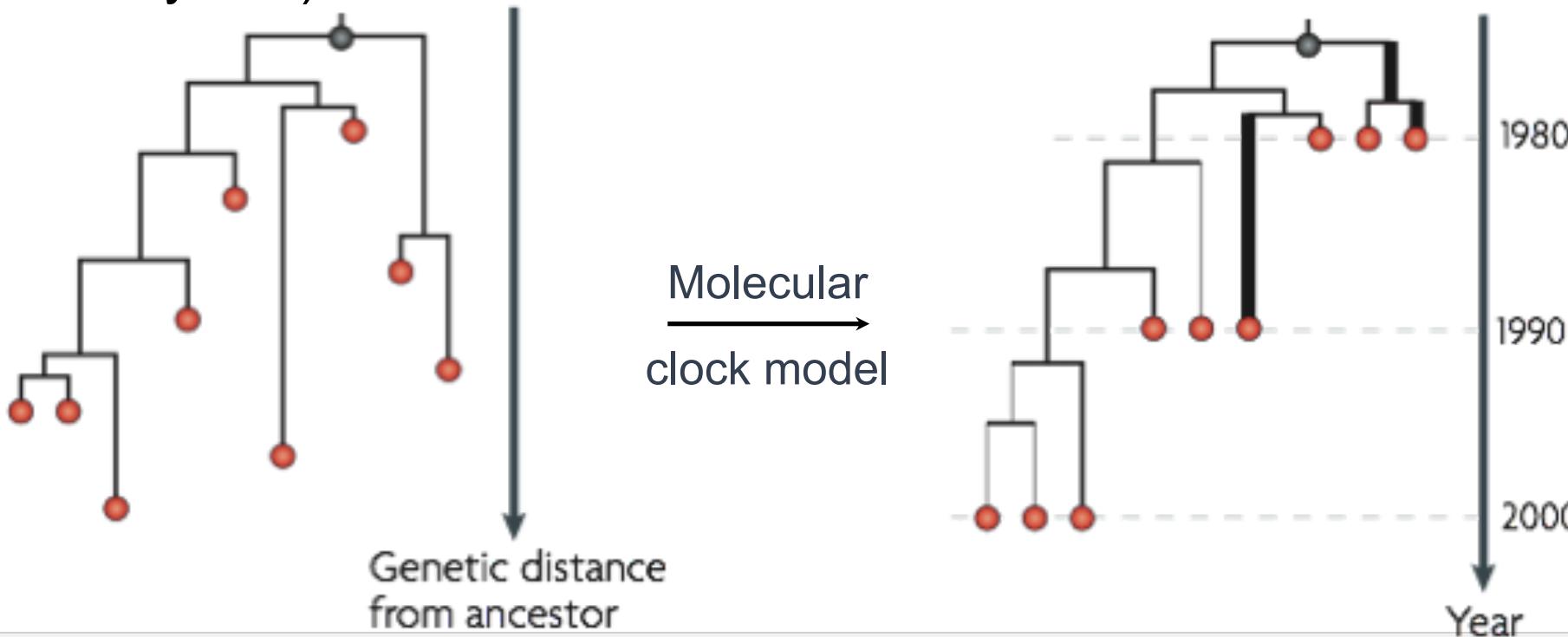
- Bootstrap analysis:
 - None
 - Ultrafast
 - Standard
- Number of bootstrap alignments: 1000
- Create .ufboot file: Yes (write bootstrap trees to .ufboot file)
- Maximum iterations: 1000

An arrow points to the "None" radio button in the Bootstrap analysis section with the text "Select No Bootstraps".

Phylogenetic Signal Considerations (TempEst)

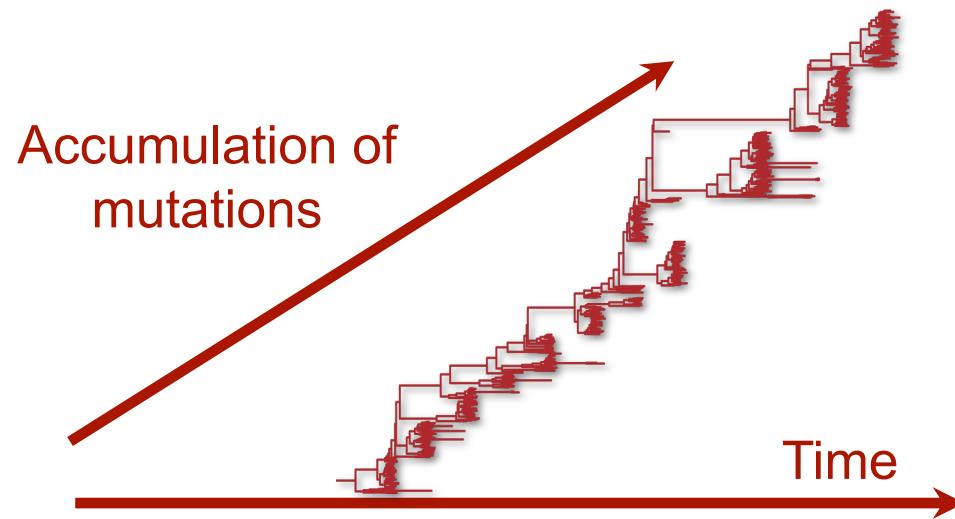
Temporal Signal

- 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

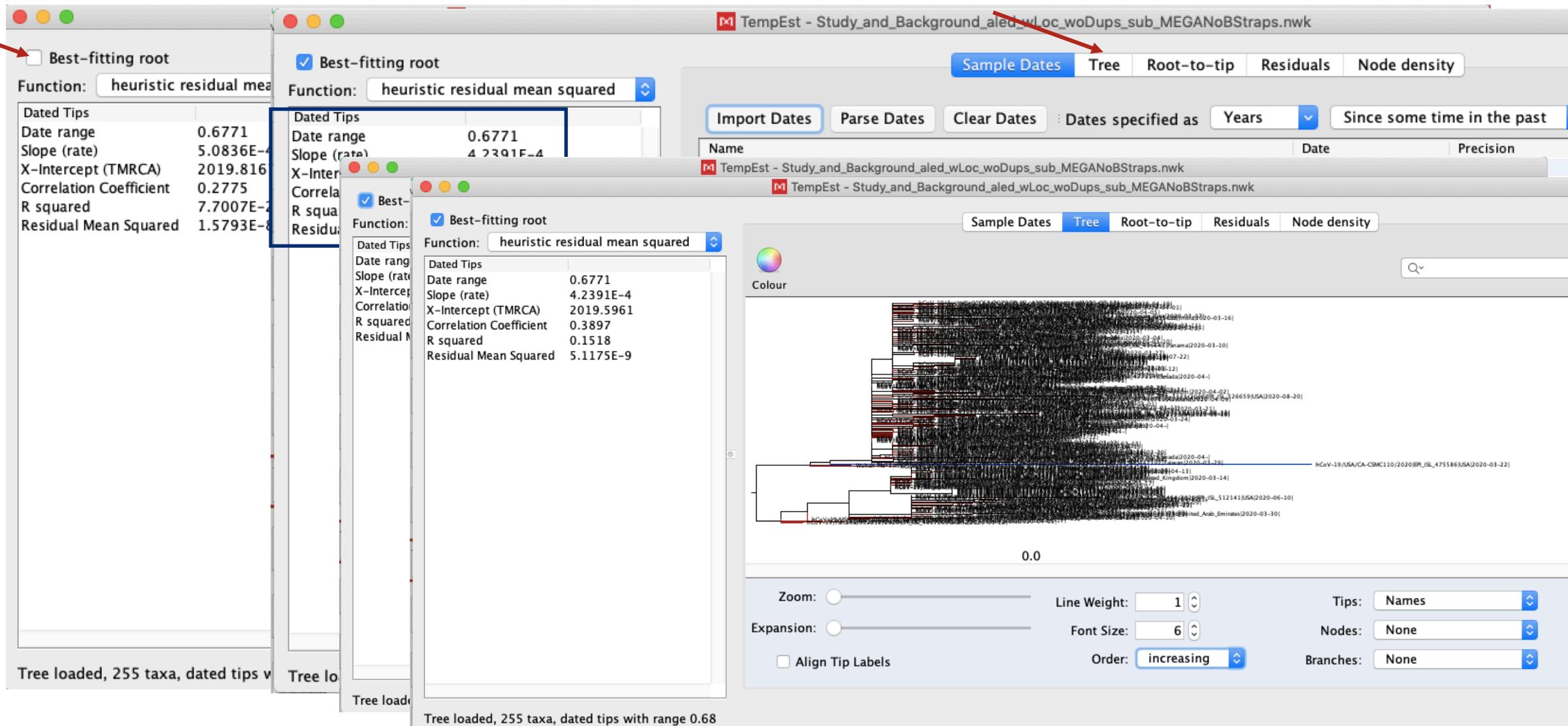
The date is given by a numerical field in the taxon label that is:

TempEst - Study_and_Background_aled_wLoc_woDups_sub_MEANoBStraps.nwk

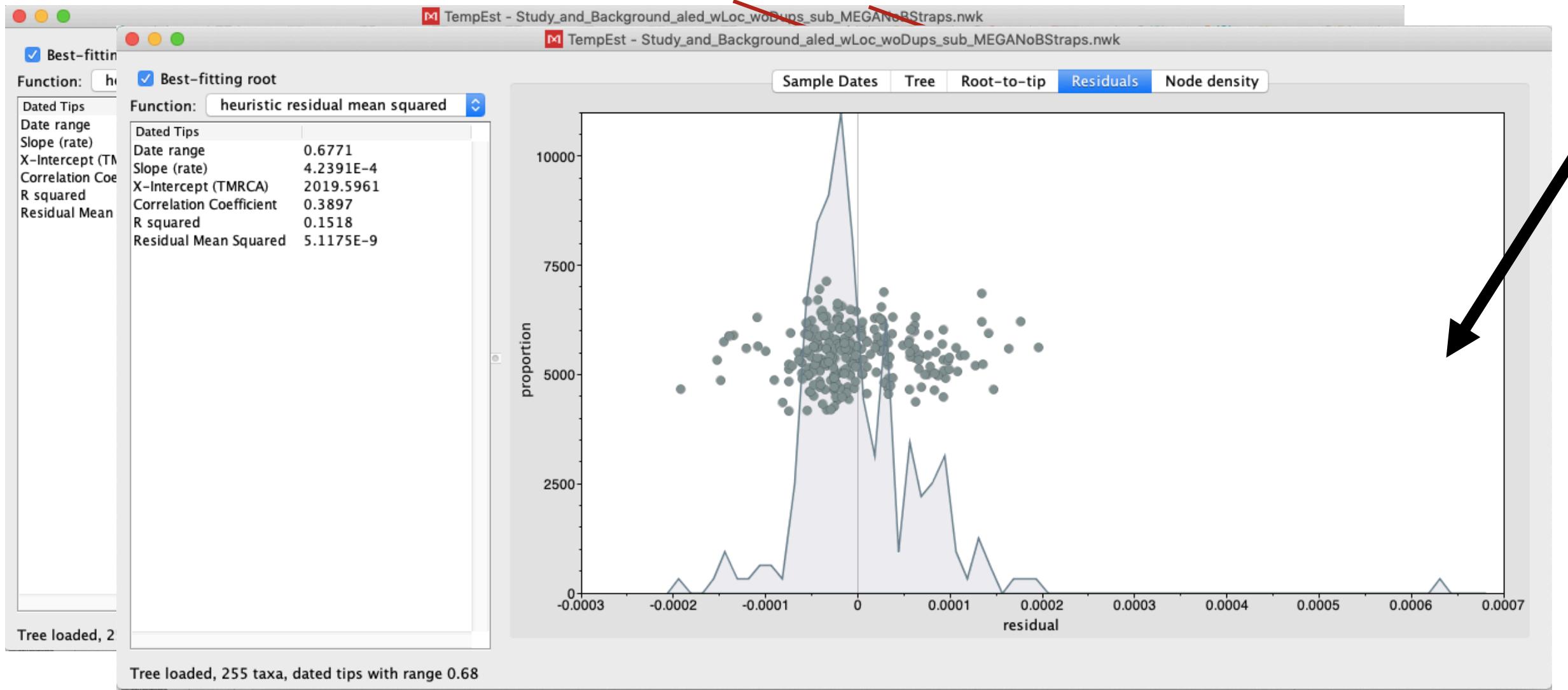
Name	Date	Precision	Height
hCoV-19/USA/CA-CSMC110/2020 EPI_ISL_475586 USA 2020-03-22	2020.22131147541	0.0	0.412568306...
hCoV-19/Australia/VIC68/2020 EPI_ISL_419782 Australia 2020-03-13	2020.196721311...	0.0	0.437158469...
hCoV-19/USA/CA-ALSR-0479-SAN/2020 EPI_ISL_483168 USA 2020-04-10	2020.273224043...	0.0	0.360655737...
hCoV-19/Canada/BC_04013424/2020 EPI_ISL_466760 Canada 2020-04-	2020.248633879...	0.0833333333...	0.385245901...
hCoV-19/Hong_Kong/HKU-200723-065/2020 EPI_ISL_497832 Hong_Kong 2020-...	2020.213114754...	0.0	0.420765027...
hCoV-19/Australia/VIC130/2020 EPI_ISL_419823 Australia 2020-03-21	2020.218579234...	0.0	0.415300546...
hCoV-19/Canada/ON_PHL6883/2020 EPI_ISL_418331 Canada 2020-03-08	2020.183060109...	0.0	0.450819672...
hCoV-19/00058/2020 StudySequence1 NA 2020-03-19	2020.213114754...	0.0	0.420765027...
hCoV-19/Taiwan/TSGH-34/2020 EPI_ISL_447593 Taiwan 2020-04-01	2020.248633879...	0.0	0.385245901...
hCoV-19/Panama/328721/2020 EPI_ISL_496608 Panama 2020-03-09	2020.185792349...	0.0	0.448087431...
hCoV-19/Panama/329547/2020 EPI_ISL_496633 Panama 2020-03-09	2020.185792349...	0.0	0.448087431...
hCoV-19/USA/MN-CDC-0106/2020 EPI_ISL_452133 USA 2020-03-10	2020.188524590...	0.0	0.445355191...
hCoV-19/Canada/ON-UHTC_0049/2020 EPI_ISL_464013 Canada 2020-04-05	2020.25956284153	0.0	0.374316939...
hCoV-19/Australia/VIC1087/2020 EPI_ISL_430574 Australia 2020-04-06	2020.262295081...	0.0	0.371584699...
hCoV-19/Peru/LIM-INS-021/2020 EPI_ISL_491427 Peru 2020-03-25	2020.229508196...	0.0	0.404371584...
hCoV-19/Australia/VIC-CBA4/2020 EPI_ISL_430064 Australia 2020-03-07	2020.180327868...	0.0	0.453551912...
hCoV-19/Costa_Rica/03/2020 EPI_ISL_434536 Costa_Rica 2020-03-17	2020.207650273...	0.0	0.426229508...
hCoV-19/Australia/VIC05/2020 EPI_ISL_416413 Australia 2020-03-05	2020.174863387...	0.0	0.459016393...
hCoV-19/Australia/VIC20/2020 EPI_ISL_419741 Australia 2020-03-08	2020.183060109...	0.0	0.450819672...
hCoV-19/Colombia/T1/2020 EPI_ISL_498168 Colombia 2020-03-11	2020.191256830...	0.0	0.442622950...
hCoV-19/Colombia/T8/2020 EPI_ISL_498169 Colombia 2020-03-11	2020.191256830...	0.0	0.442622950...
hCoV-19/Colombia/T12/2020 EPI_ISL_498170 Colombia 2020-03-11	2020.191256830...	0.0	0.442622950...
hCoV-19/Australia/NSW2148/2020 EPI_ISL_500629 Australia 2020-03-16	2020.204918032...	0.0	0.428961748...
hCoV-19/USA/MN-CDC-6277/2020 EPI_ISL_527691 USA 2020-05-01	2020.330601092...	0.0	0.303278688...

Tree loaded, 255 taxa, dated tips with range 0.68

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

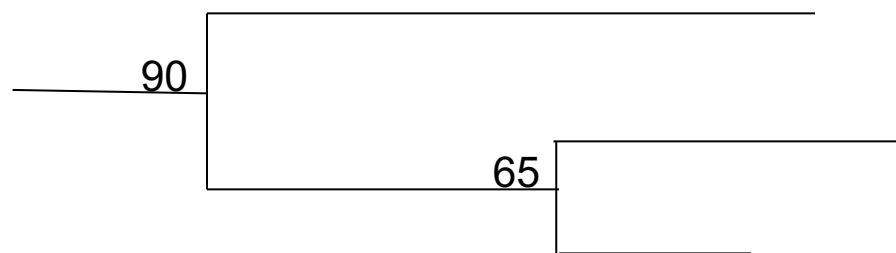
Note:

- 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)



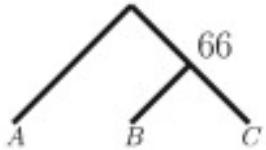
Bootstrapping (How Robust is the Tree?)

MSA

Inferred Tree

Original Data

A	A	C	T	T
B	G	G	A	T
C	G	G	C	C



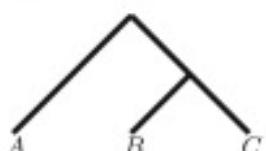
Bootstrap Replicate #1

A	A	C	T	C
B	G	G	A	G
C	G	G	C	G



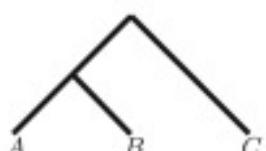
Bootstrap Replicate #2

A	C	A	T	A
B	G	G	A	G
C	G	G	C	G



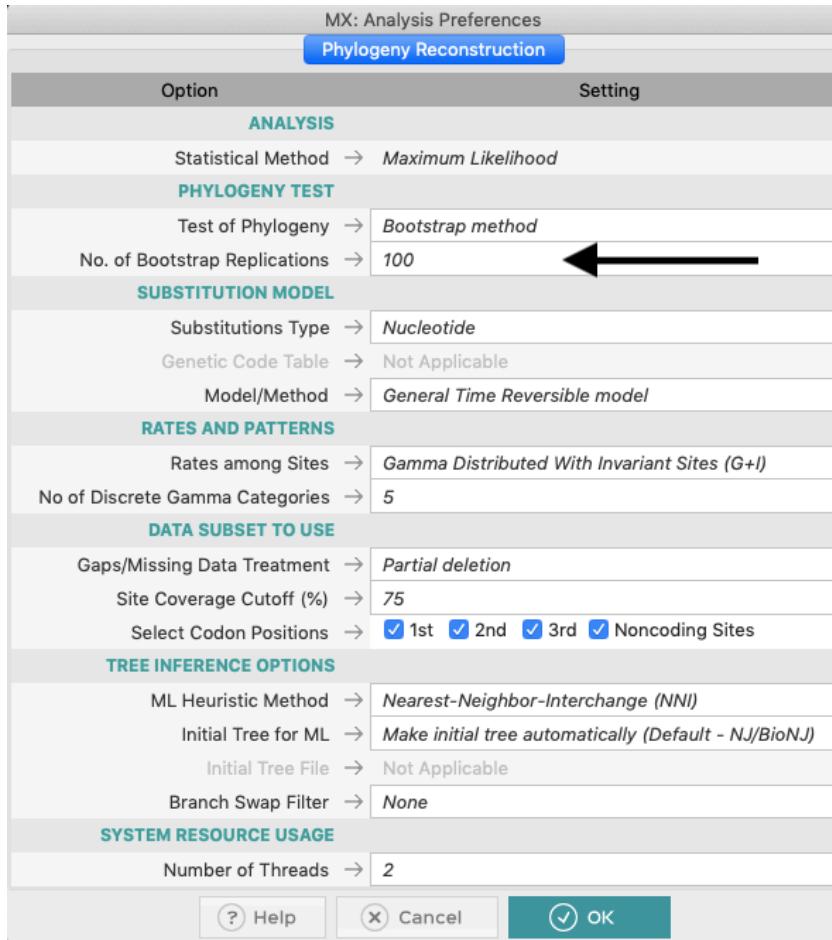
Bootstrap Replicate #3

A	T	T	T	T
B	A	T	T	A
C	C	C	C	C



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/>

The screenshot shows the IQ-TREE web server interface. At the top, it displays the URL <http://iqtree.cibiv.univie.ac.at/>. Below the URL, the page title is "IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood". It includes a note about server load (4%), the citation information (Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. doi: 10.1093/nar/gkw256), and links for Tree Inference, Model Selection, and Analysis Results.

Input Data:

- Alignment file: C:\fakepath\Study_and_Background (with a "Browse..." button and a "Show example >" link). An arrow points to this field with the text "Select the multiple sequence alignment file".
- Use example alignment: Yes
- Sequence type:
 - Auto-detect
 - DNA
 - Protein
 - Codon
 - DNA->AA
 - Binary
 - Morphology
- Partition file: This field is optional. (with a "Browse..." button and a "Show example >" link)
- Partition type:
 - Edge-linked
 - Edge-unlinked

Substitution Model Options:

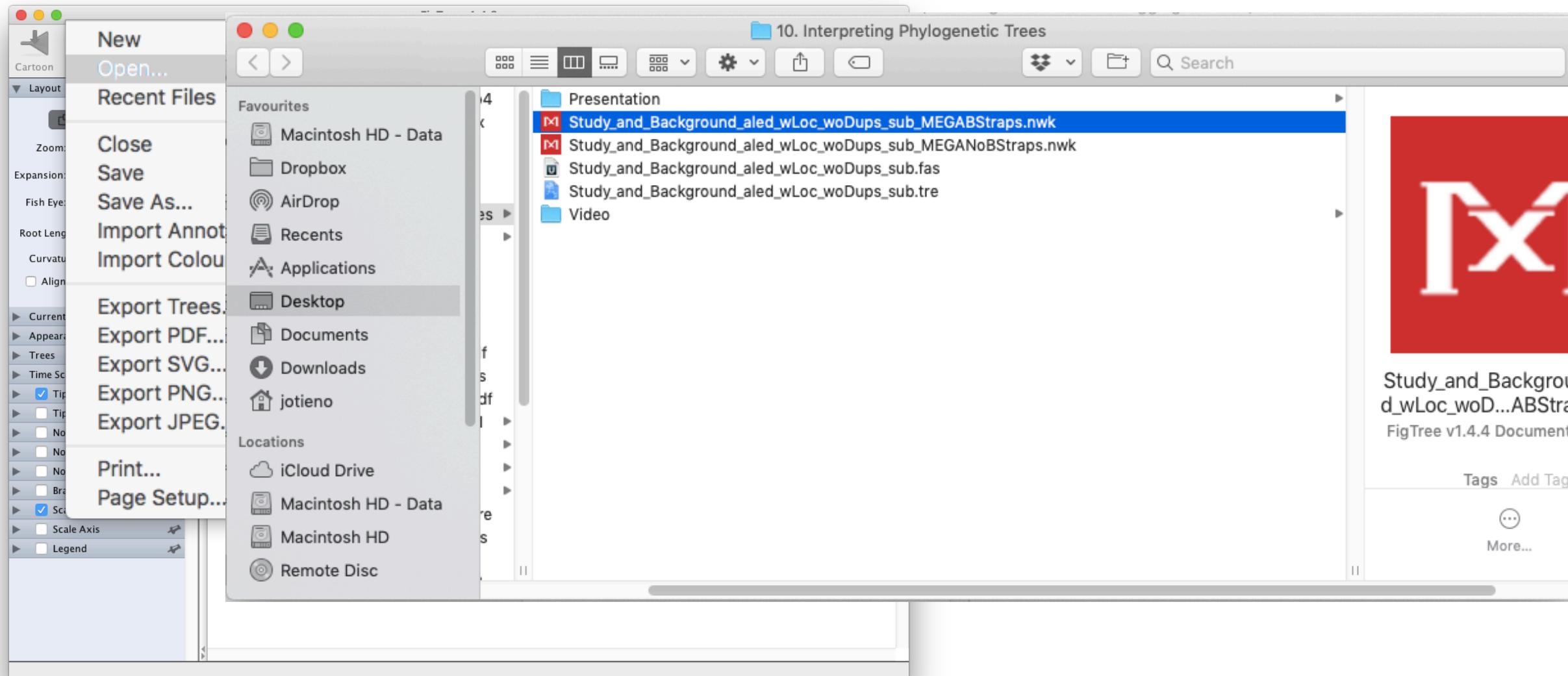
- Substitution model: Auto
- FreeRate heterogeneity: Yes [+R]
- Rate heterogeneity:
 - Gamma [+G]
 - Invar. sites [+I]
- #rate categories: 4 (with up and down arrows)
- State frequency:
 - Empirical (from data)
 - AA model (from matrix)
 - ML-optimized
 - Codon F1x4
 - Codon F3x4
- Ascertainment bias correction: Yes [+ASC]

Branch Support Analysis:

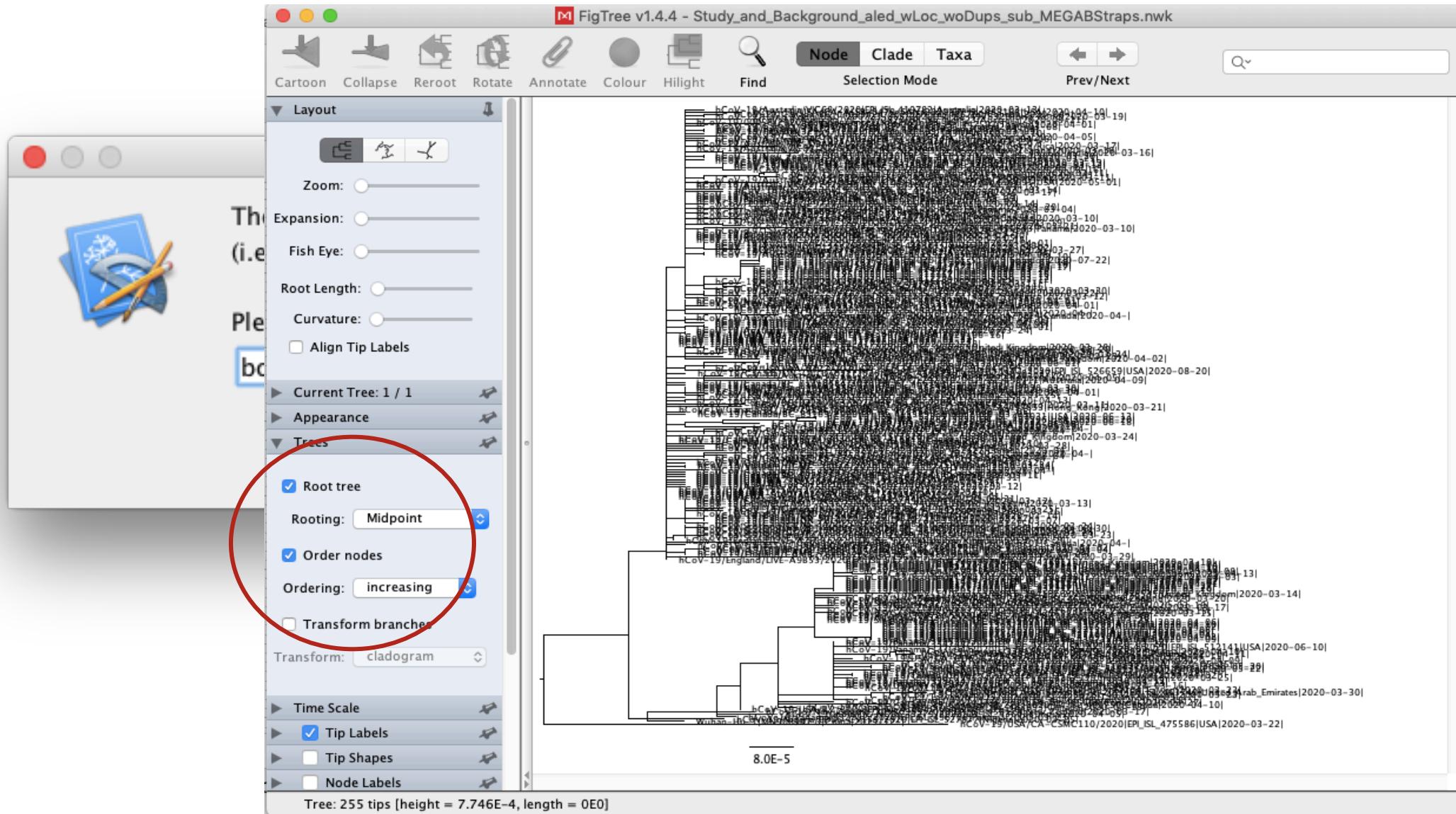
- Bootstrap analysis: None Ultrafast Standard
- Number of bootstrap alignments: 1000 (with up and down arrows). An arrow points to this field with the text "Select ultrafast Bootstraps".
- Create .ufboot file: Yes (write bootstrap trees to .ufboot file)
- Maximum iterations: 1000 (with up and down arrows)

Visualizing Phylogenetic Trees in FigTree

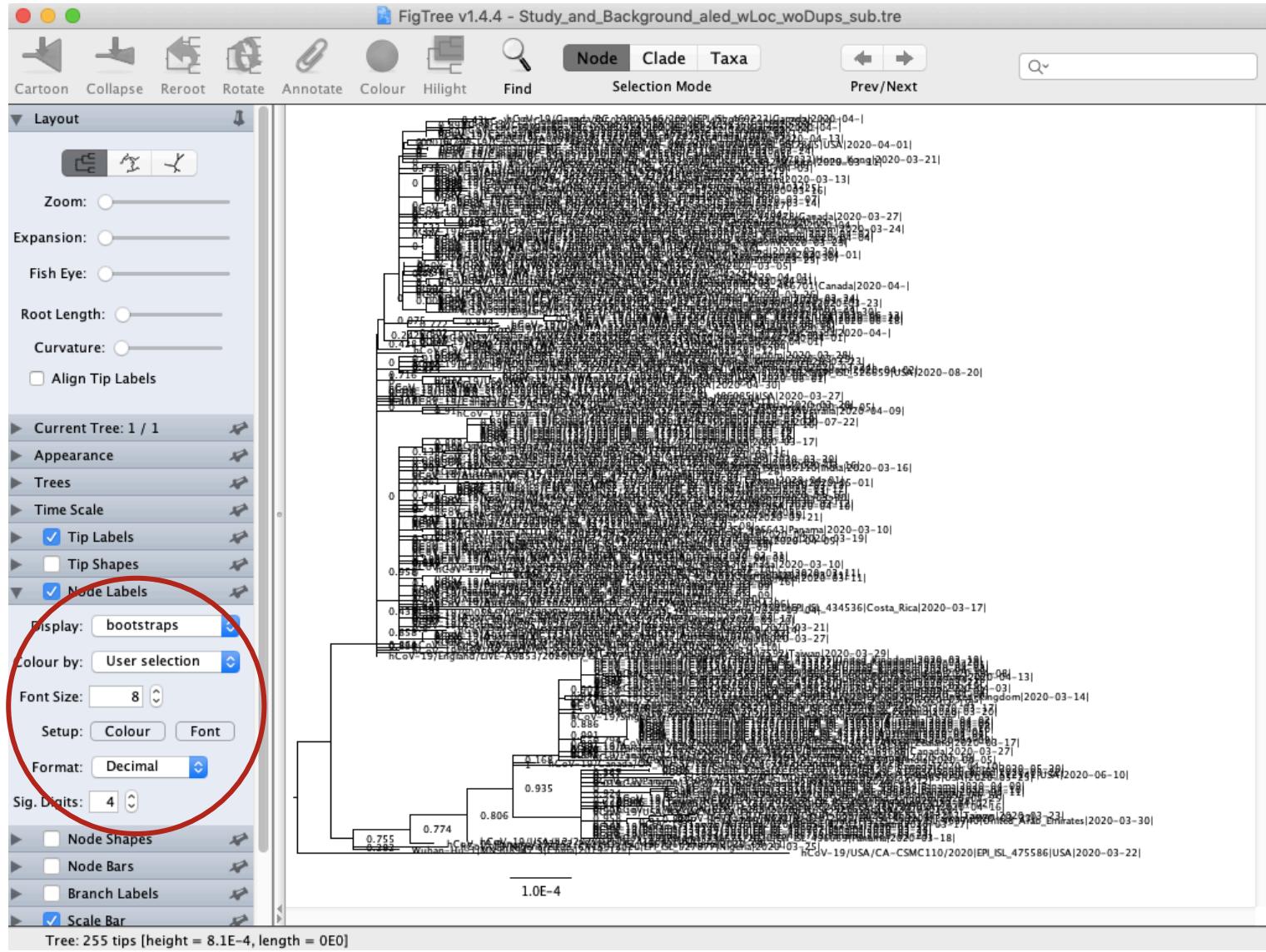
Open FigTree → Load Bootstrapped Tree



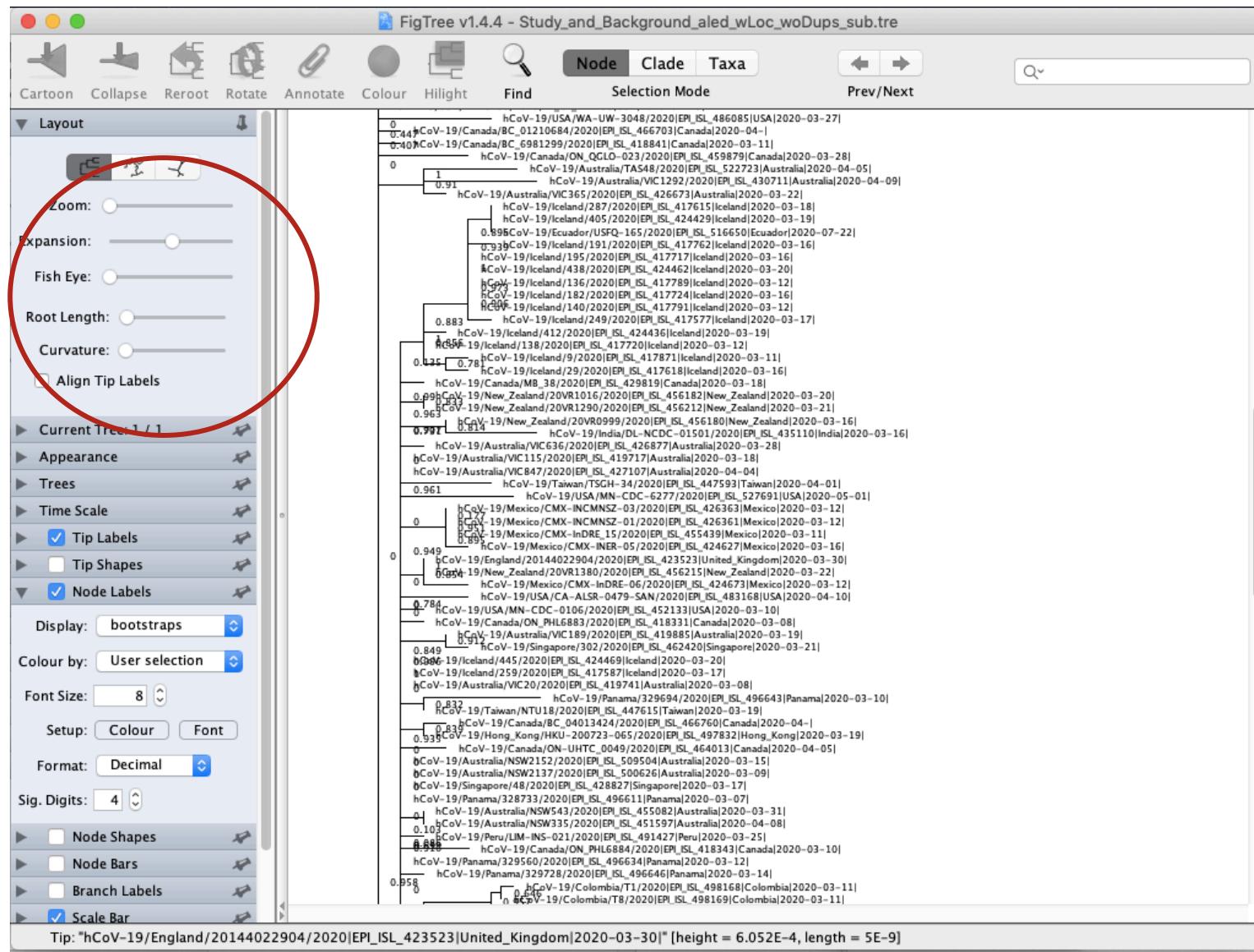
Trees: 1) Midpoint Root; 2) Increase nodes order



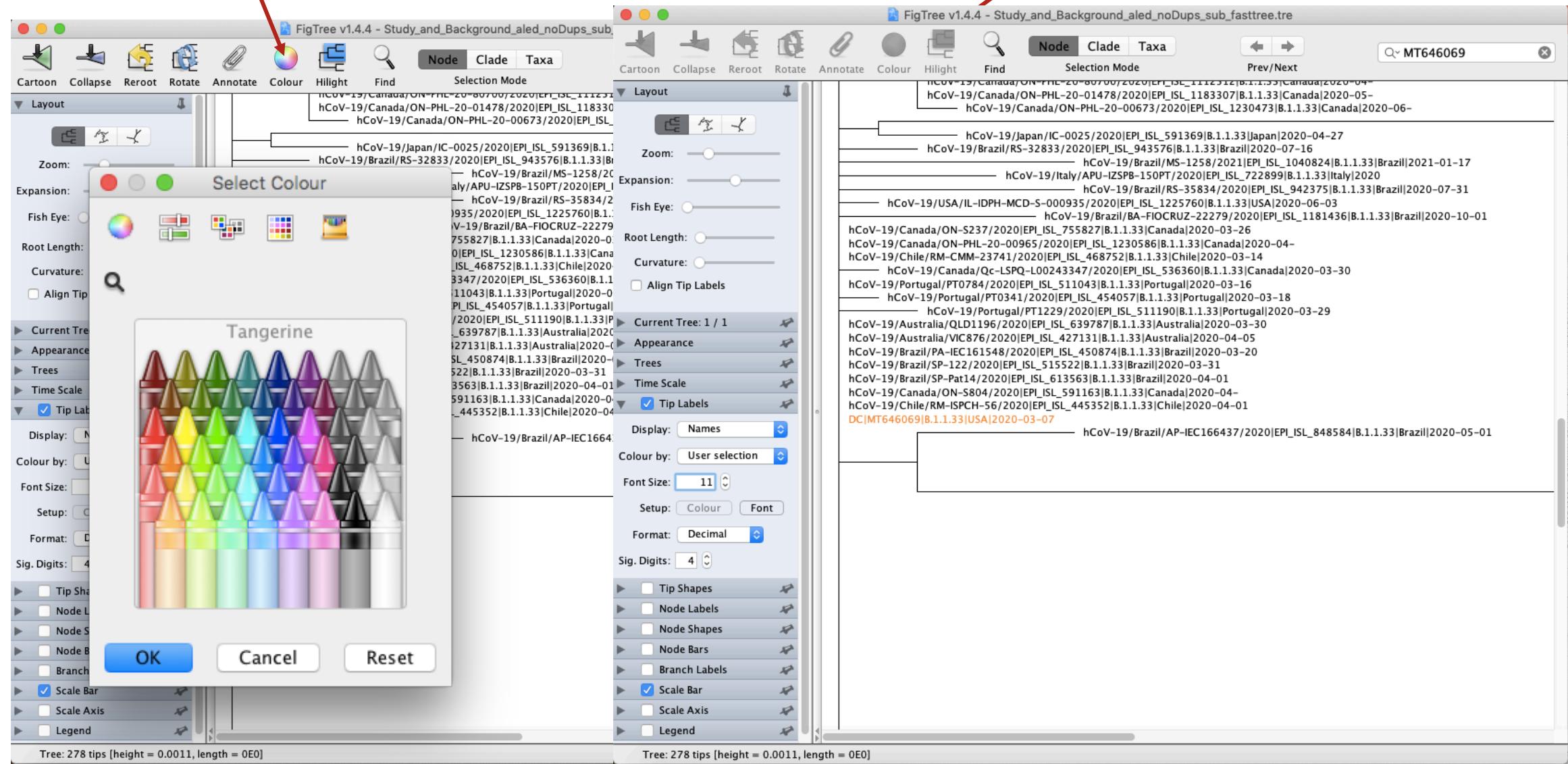
Add bootstrap support labels to the nodes



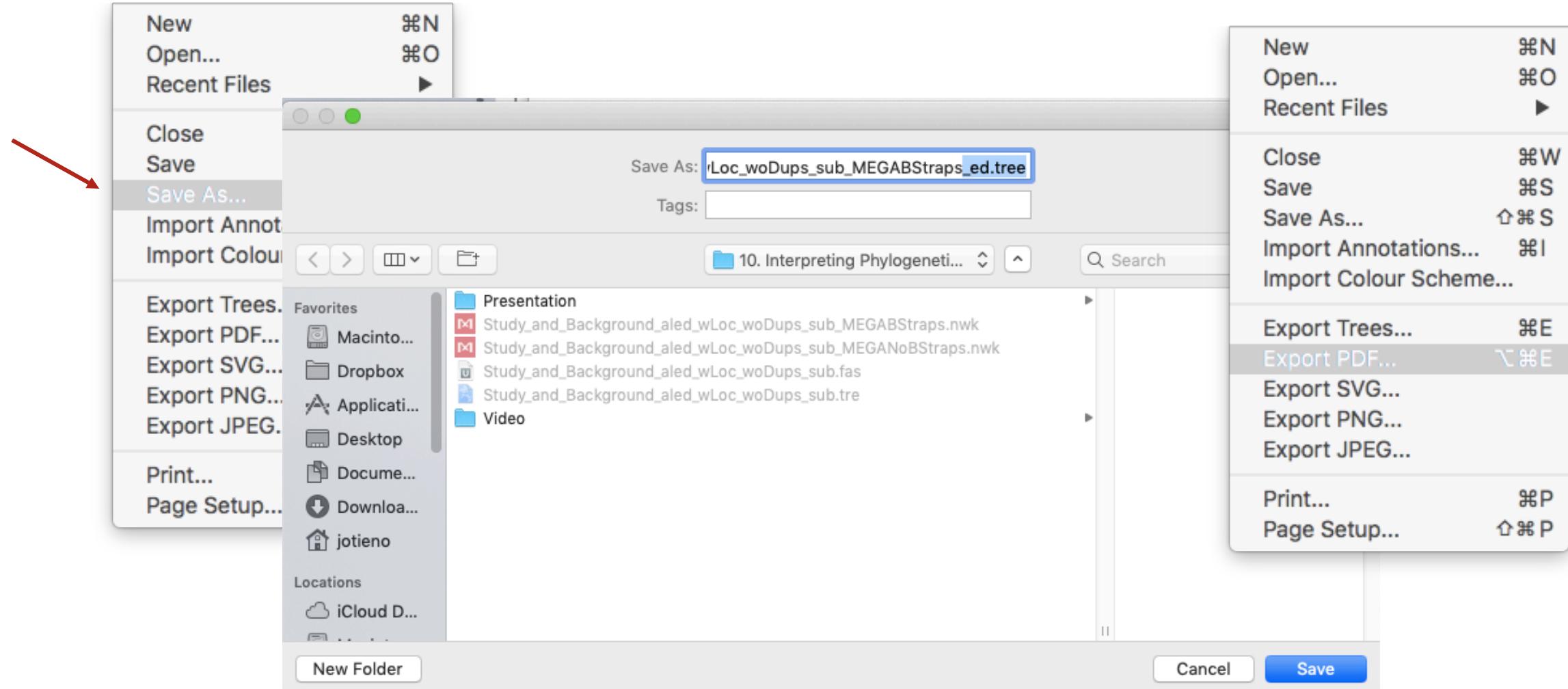
Expand tree to make it readable



Where is MT646069?



Save Tree and Export as a PDF

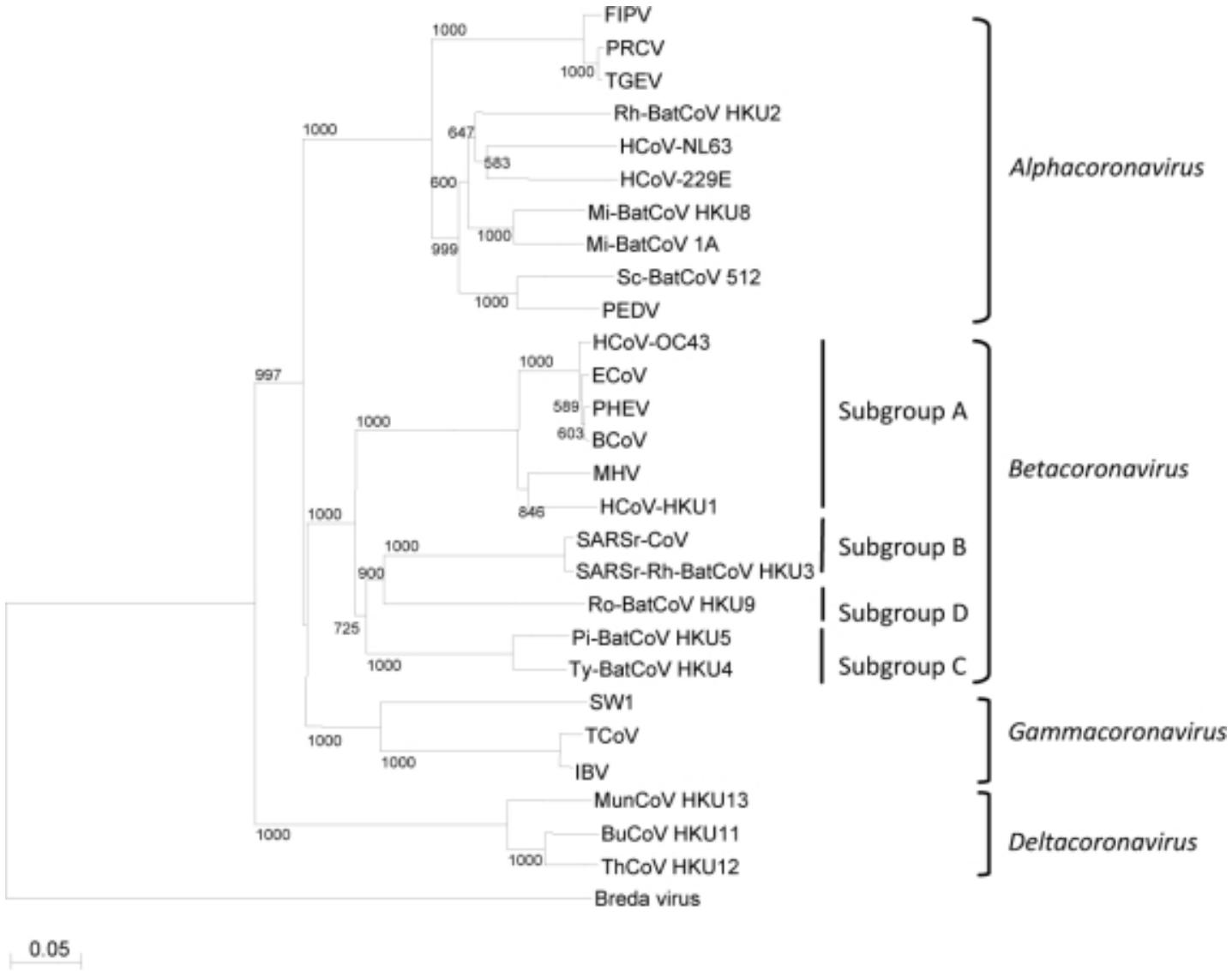


Basic Interpretation of Phylogenetic Trees

**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?

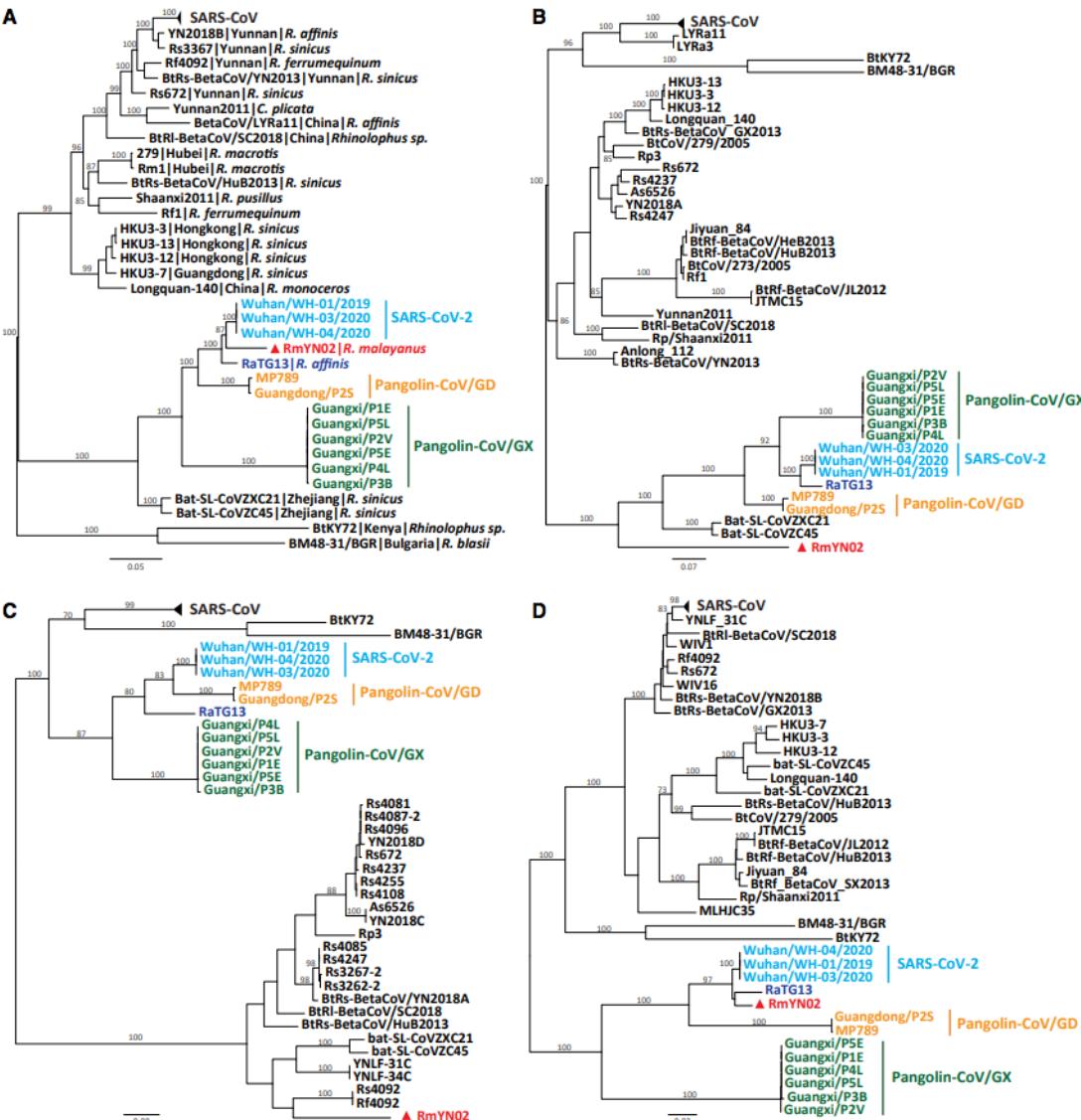
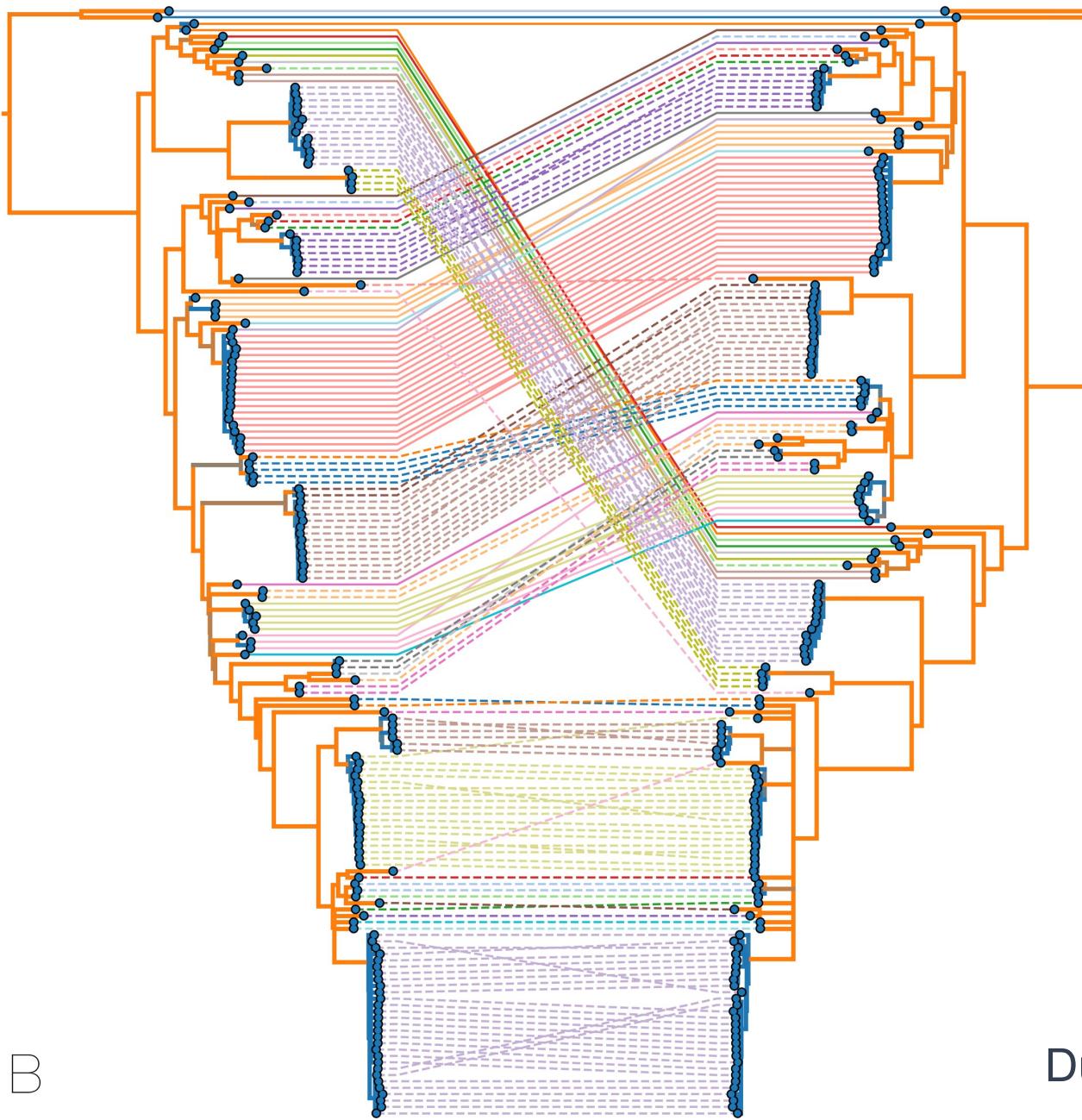


Figure 3. Phylogenetic Analysis of SARS-CoV-2 and Representative Viruses from the Subgenus Sarbecoronavirus

Trees inferred
from different
gene regions

MERS

Tanglegram

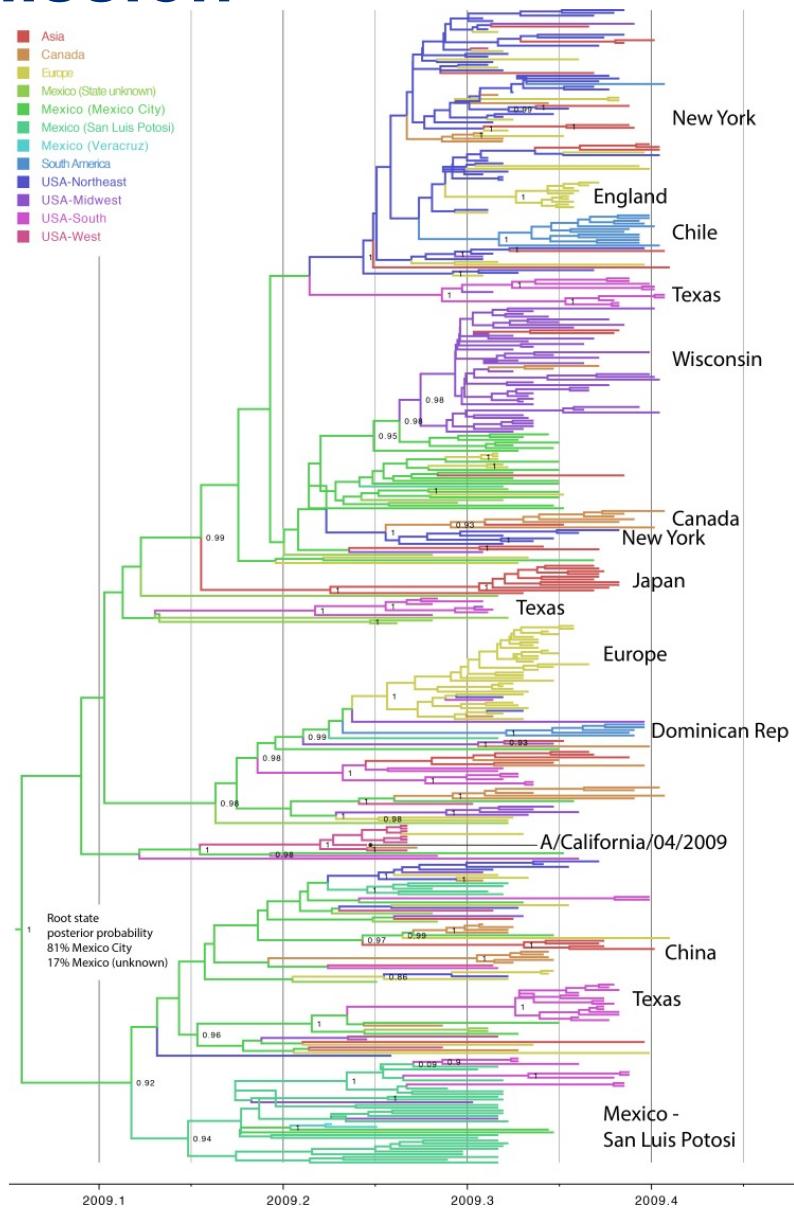


B

Dudas et al., 2018

2. Pathogen Transmission

Pandemic H1N1

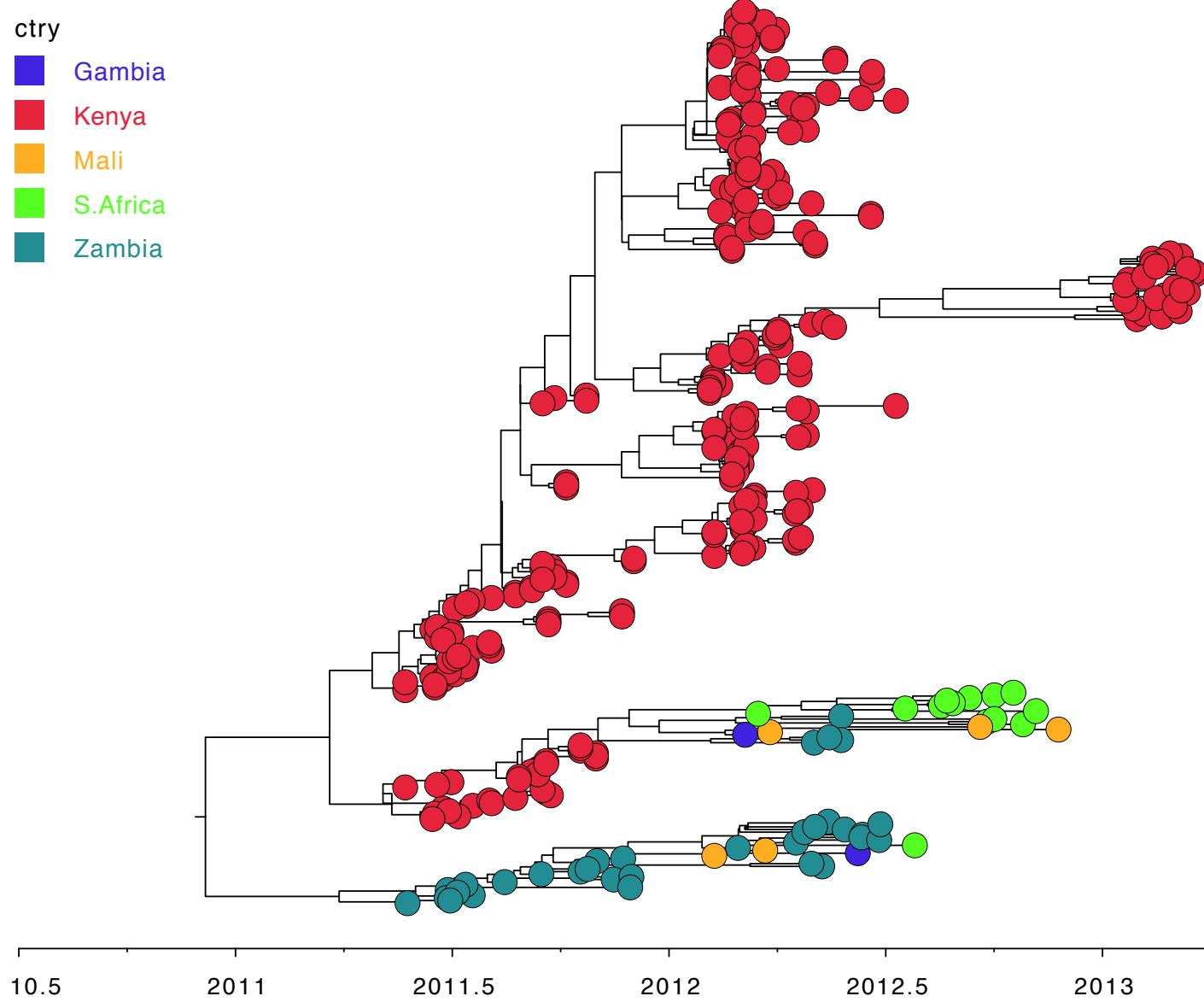


Ignacio Mena,
Martha I Nelson, et al., 2016

RSV genotype ON1

Africa

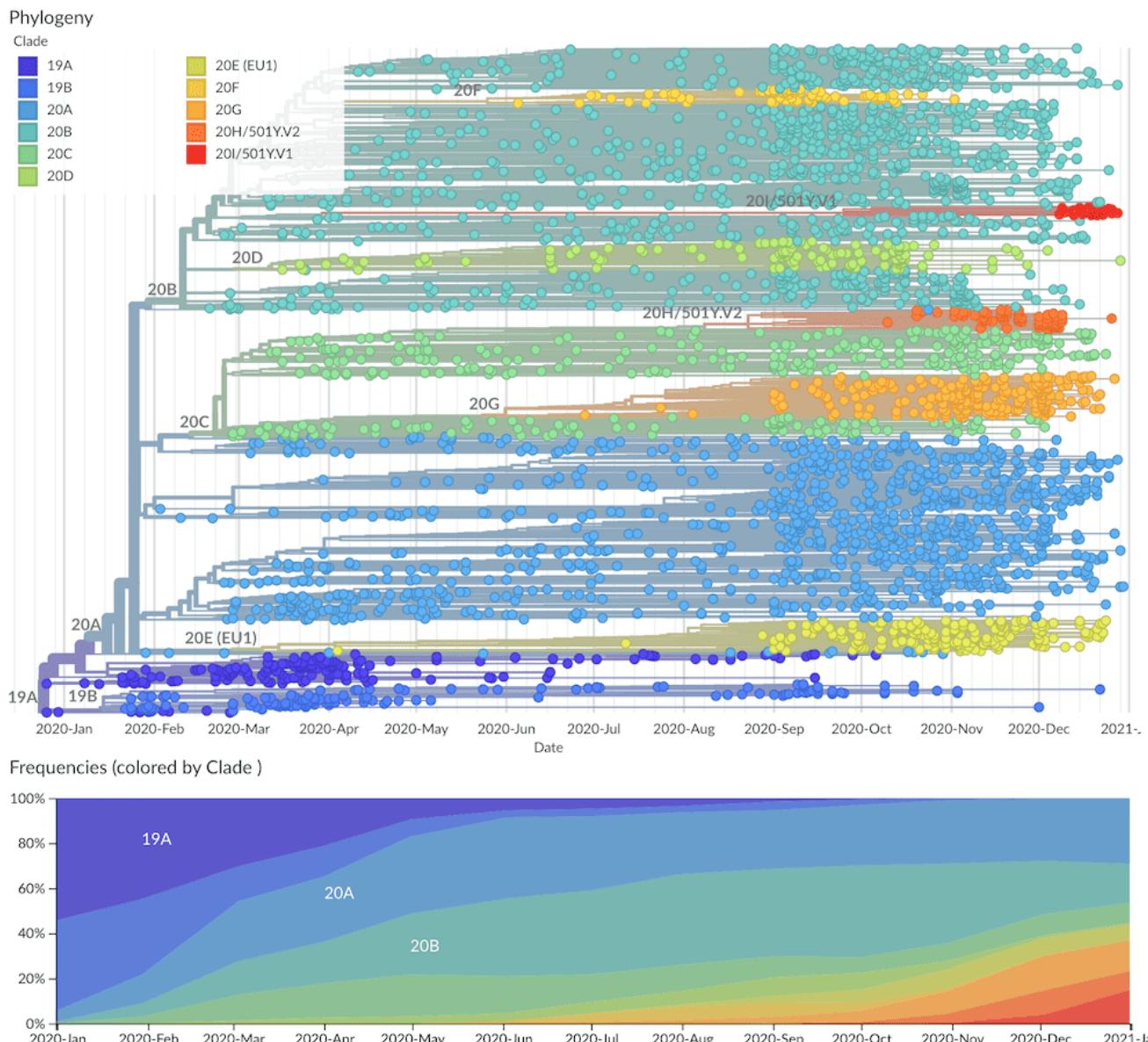
Where did the
Mali RSV
strains
originate
from?



Pathogen Evolution

SARS-CoV-2

Clades/Variant s/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

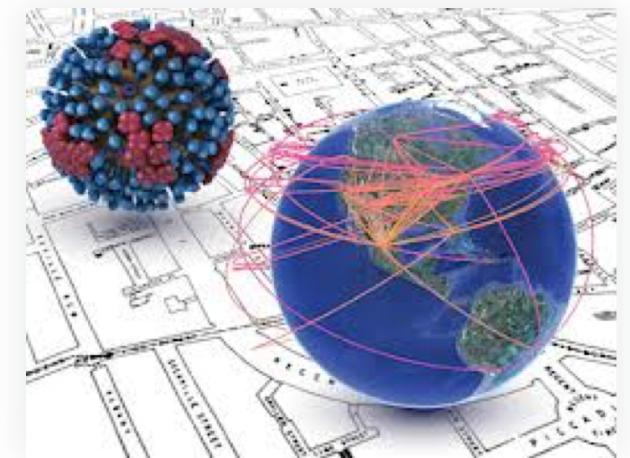


Conclusions

- 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/>



COVID-19 International Research Team



Fogarty International Center
Advancing Science for Global Health