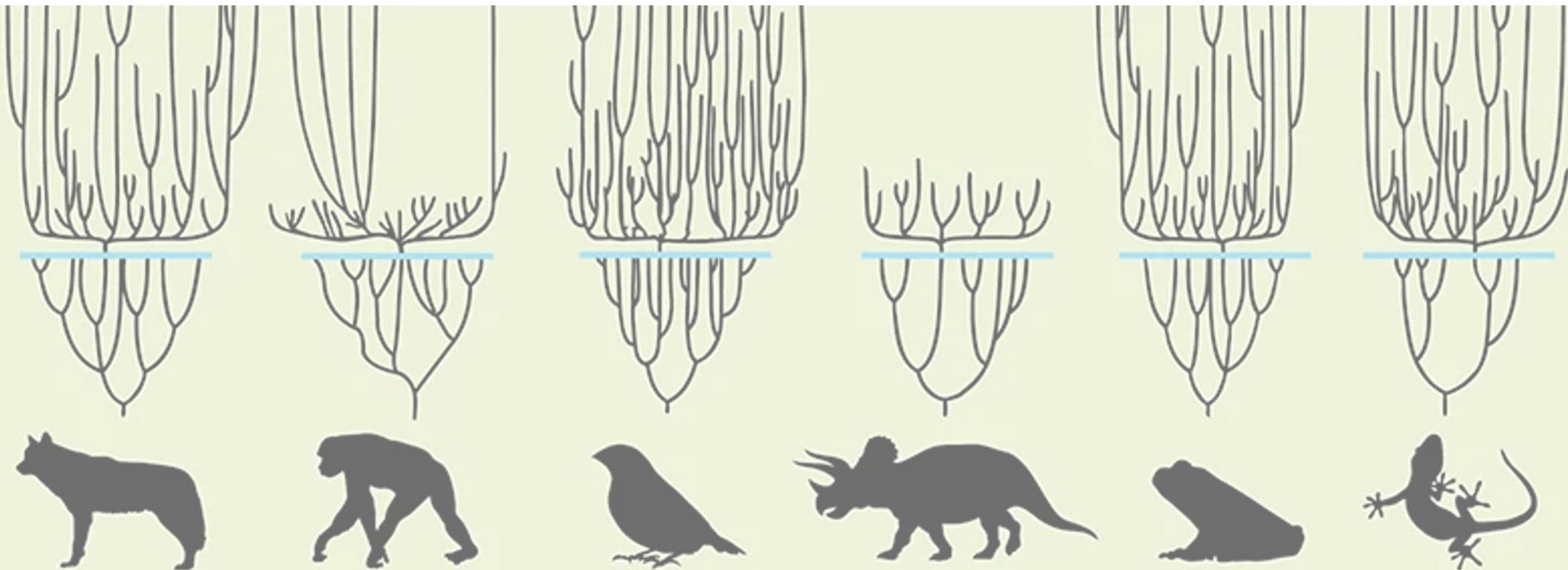


Introduction to phylogenetics

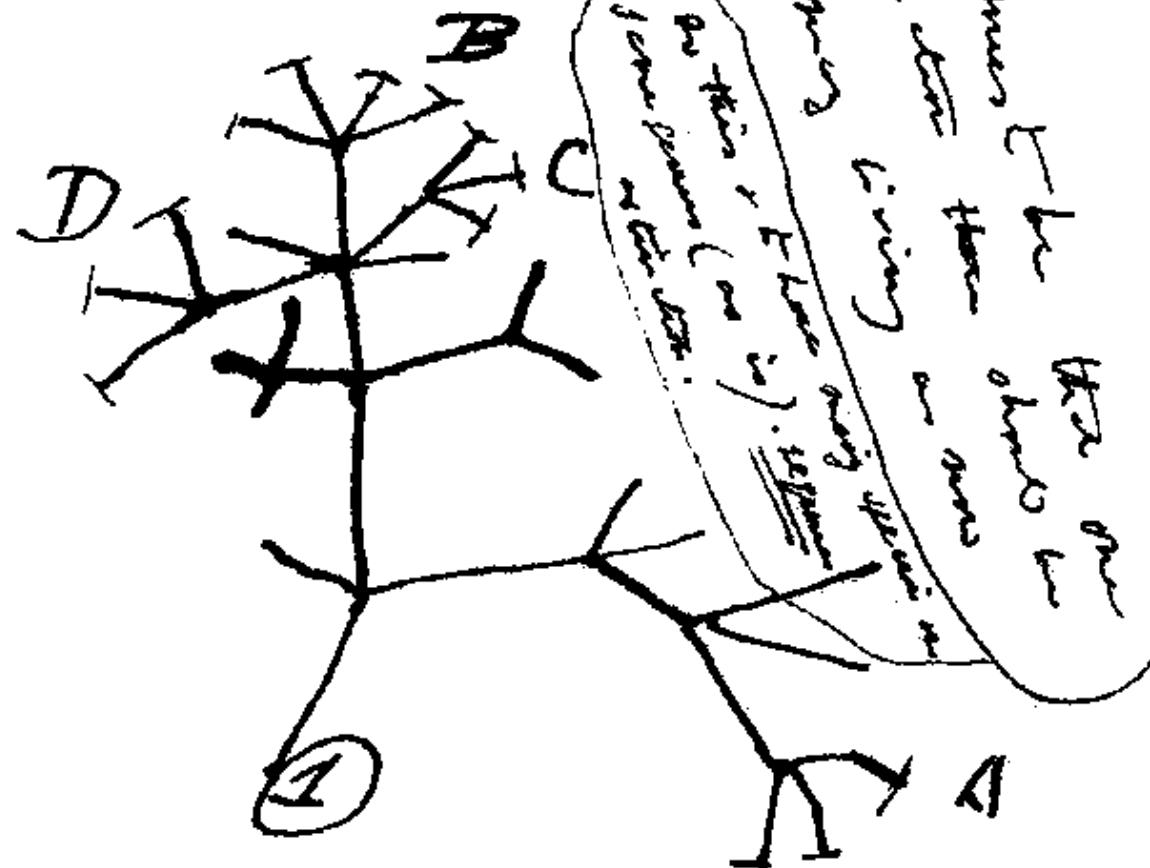
Gwen Kettenburg

Slides adapted from Richard Ree and Andrew Hipp, University of Chicago

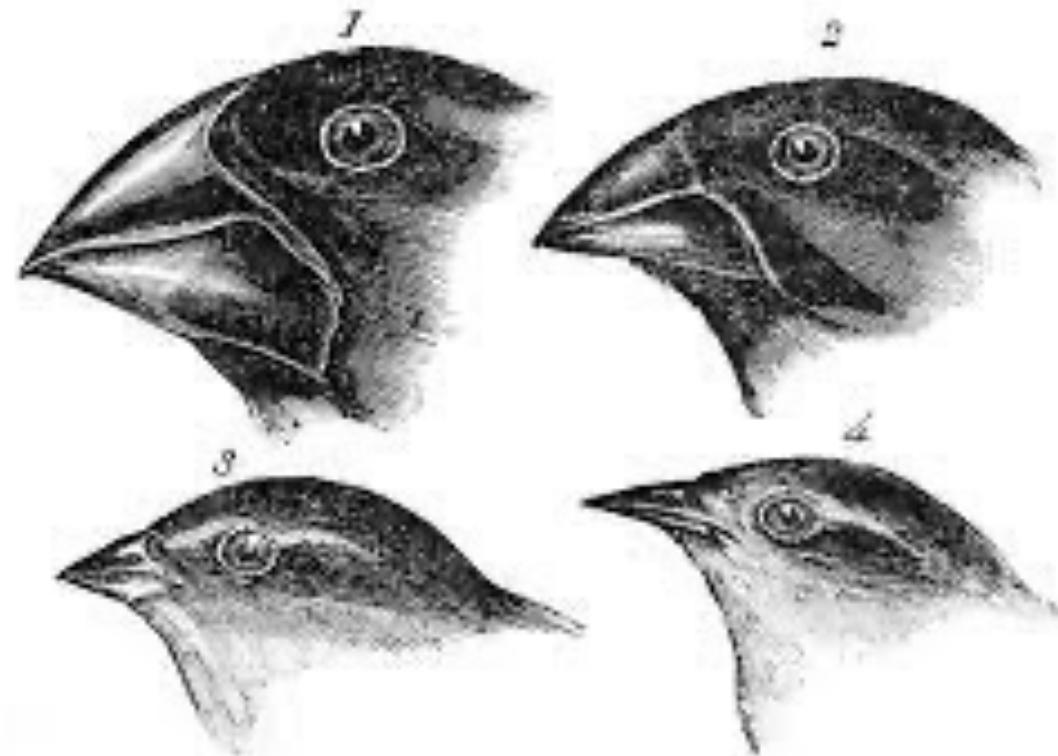


What is a phylogeny?

I think



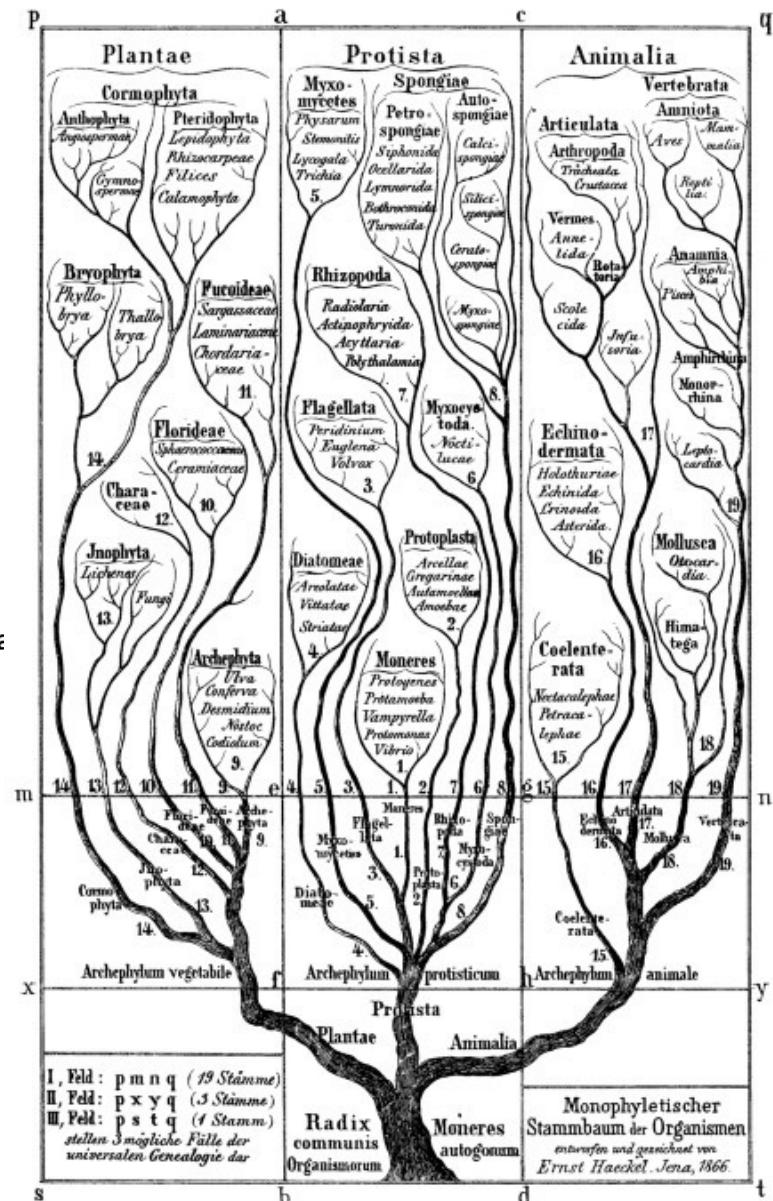
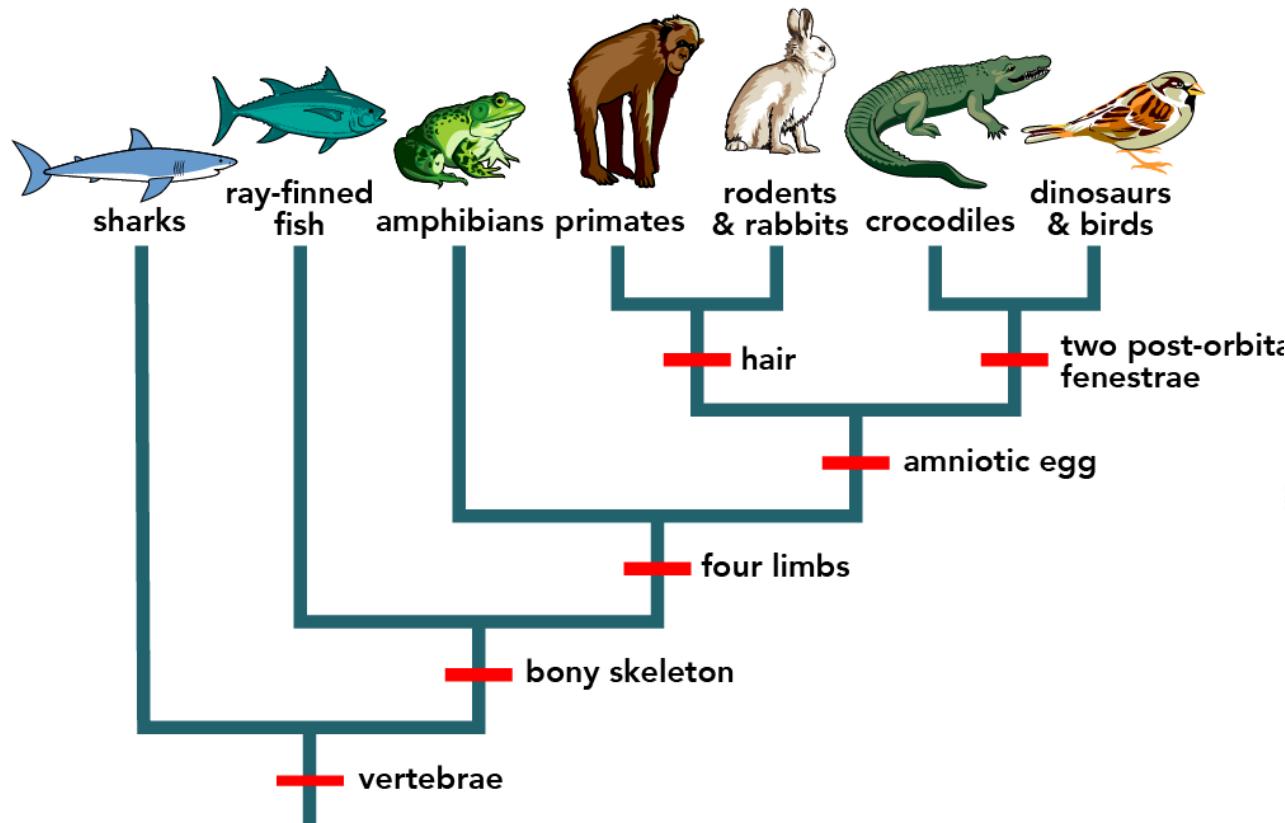
(36)



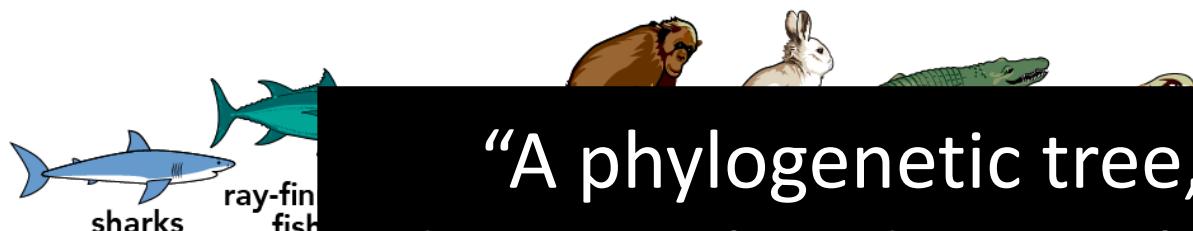
1. *Geopelia magnirostris*,
2. *Geopelia parvula*,

3. *Geopelia fuscata*,
4. *Coturnis olivaceus*.

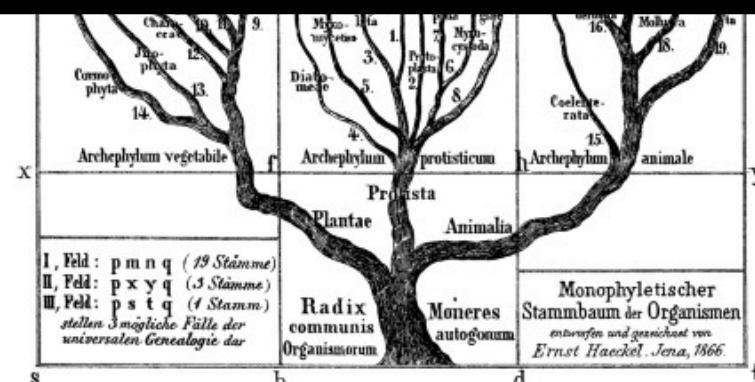
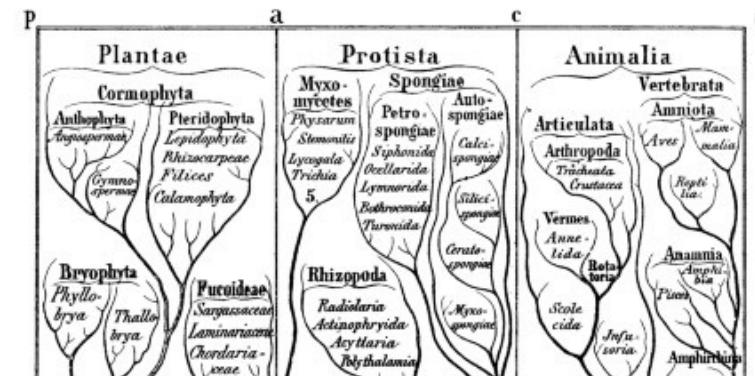
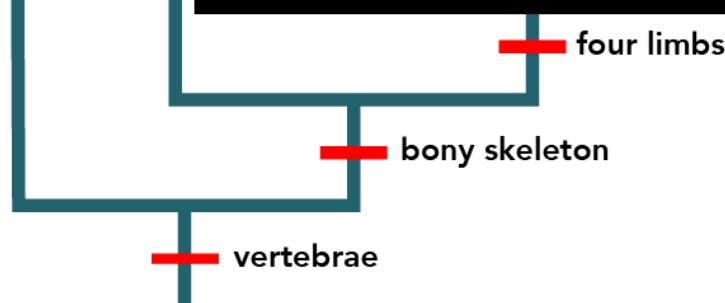
What is a phylogeny?



What is a phylogeny?

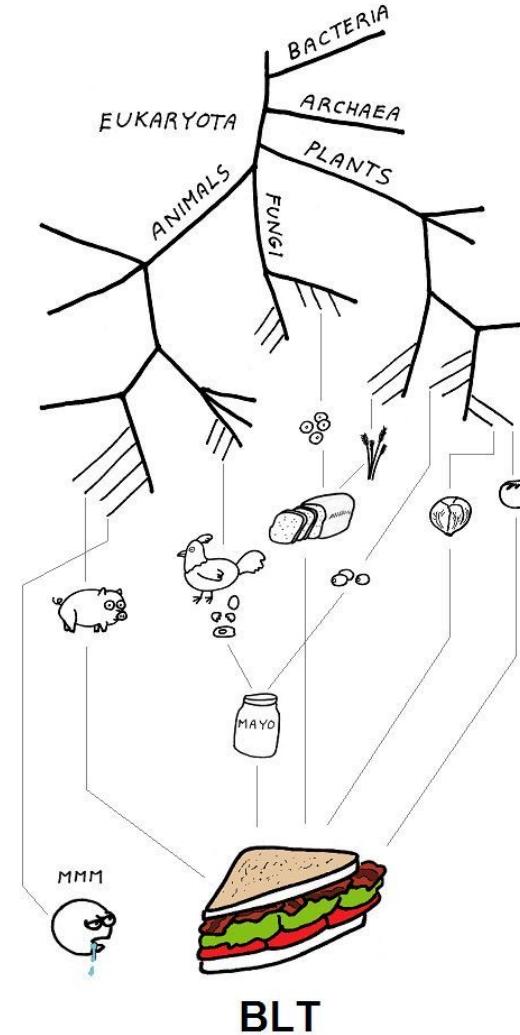
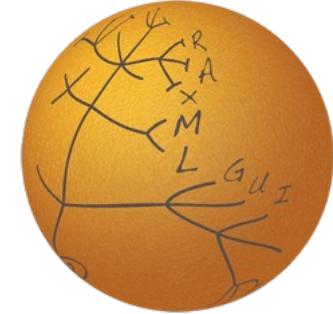


"A phylogenetic tree, or a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor."



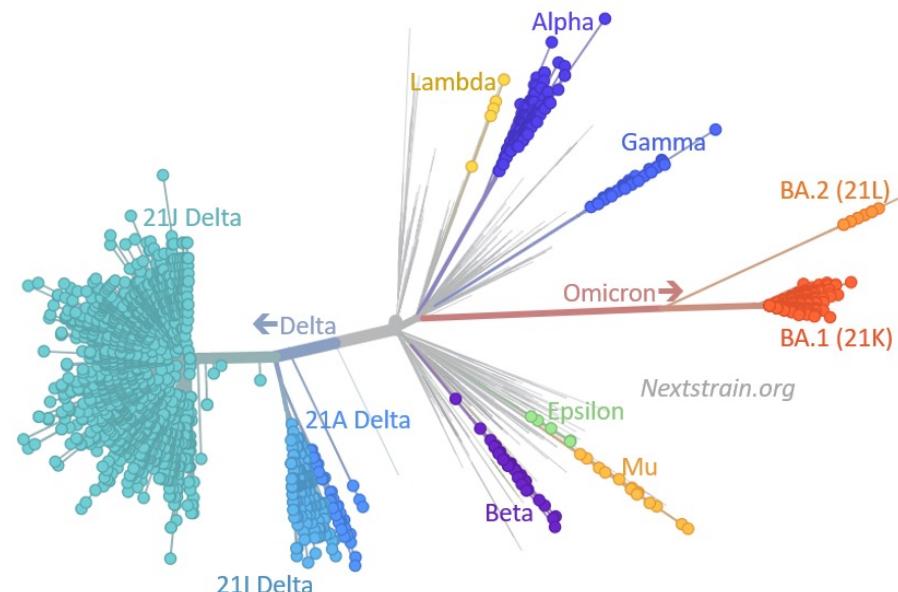
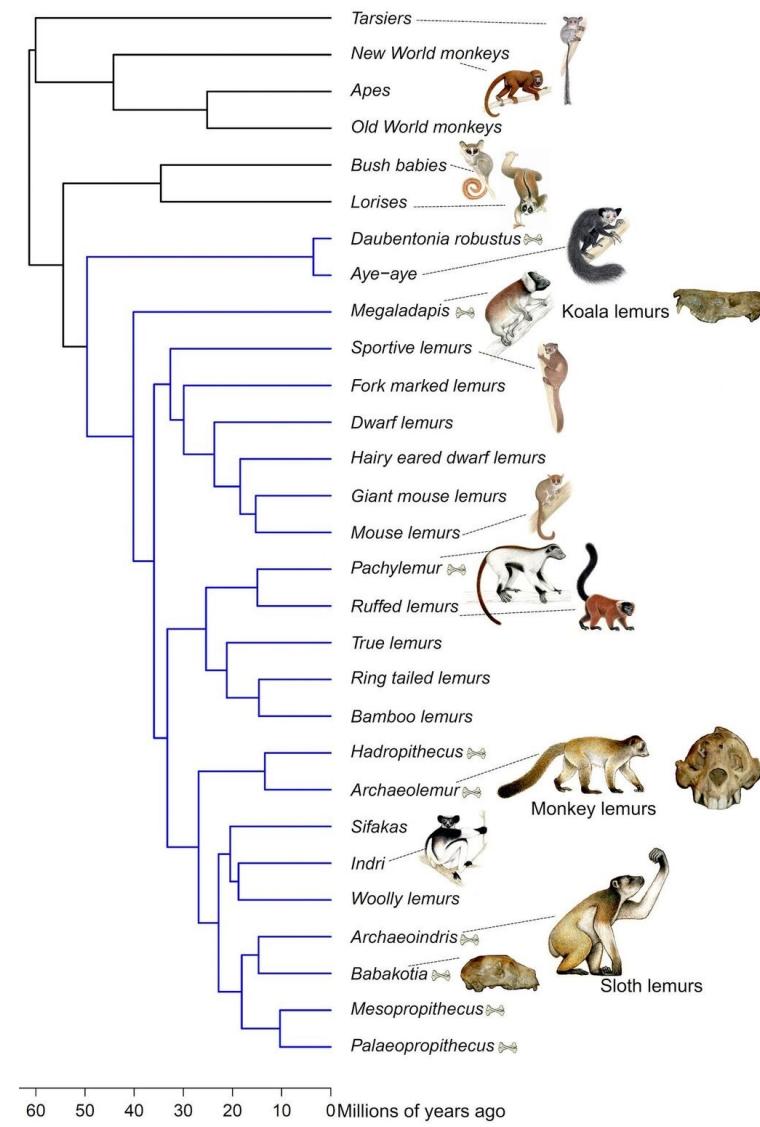
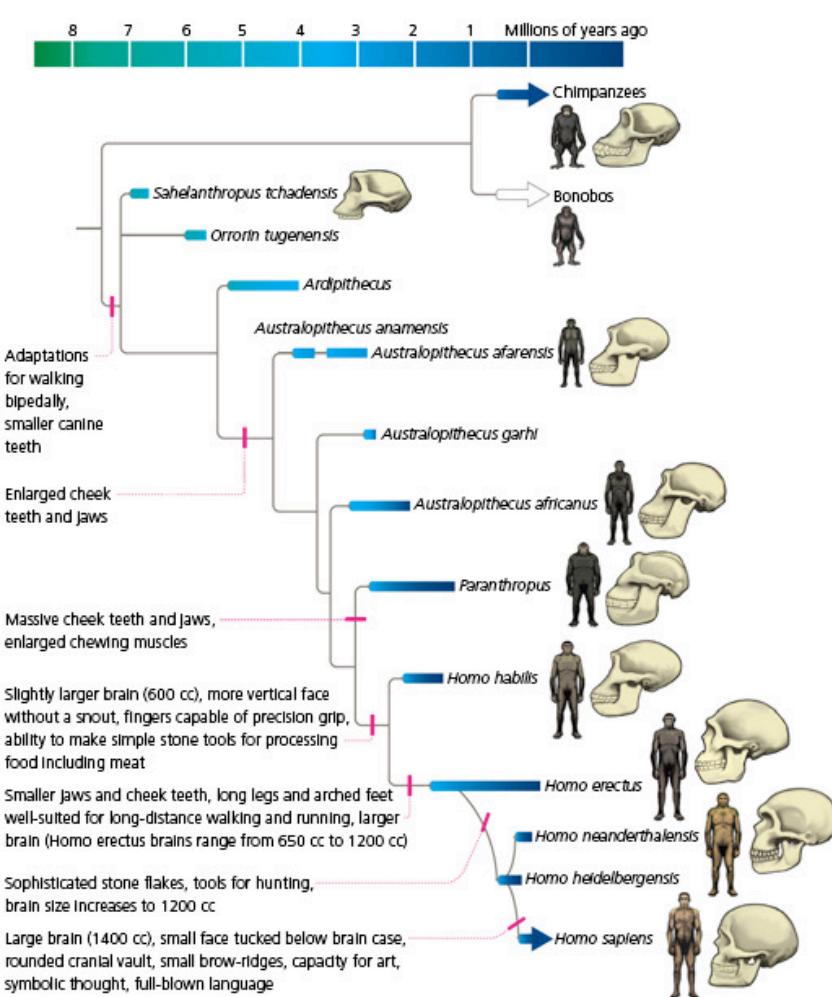
Goals:

- Lecture component
 - Learn basics of what a phylogeny is
 - Learn how to read phylogenies
 - Basics of phylogenetic modeling
- Tutorial component
 - Learn how to make a phylogenetic tree from sequencing data
 - Using influenza protein sequences in MEGA software
 - Edit and visualize tree in R and FigTree

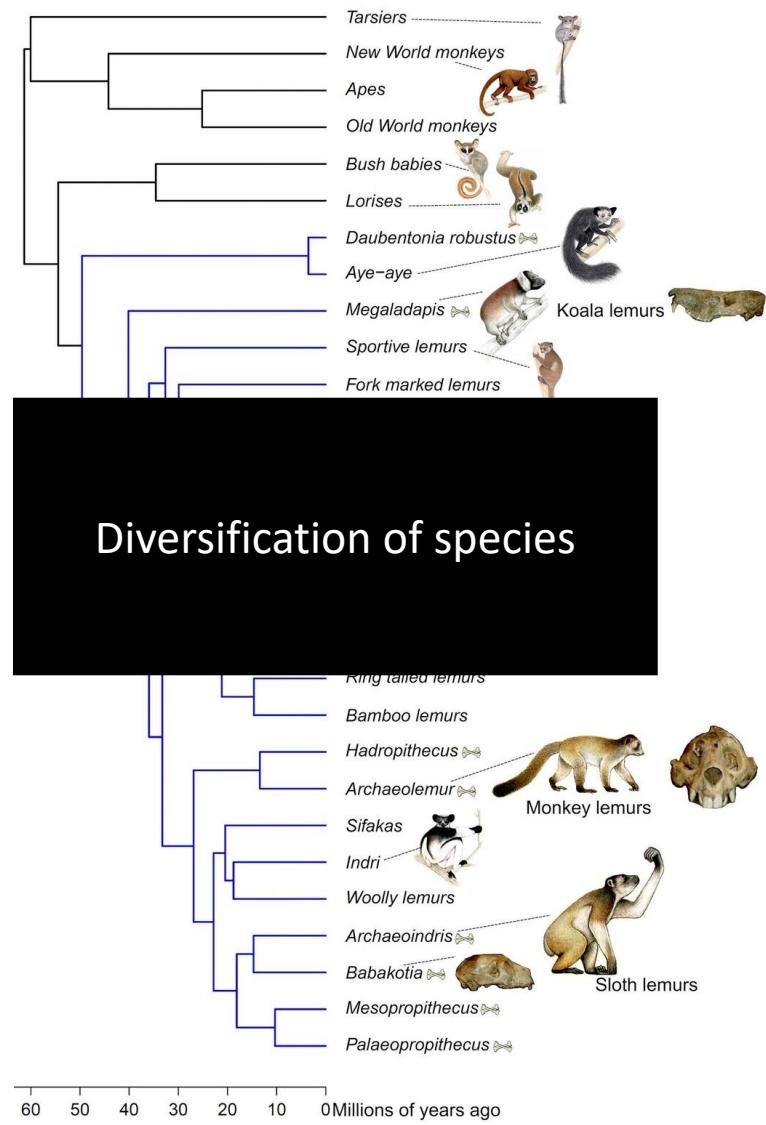
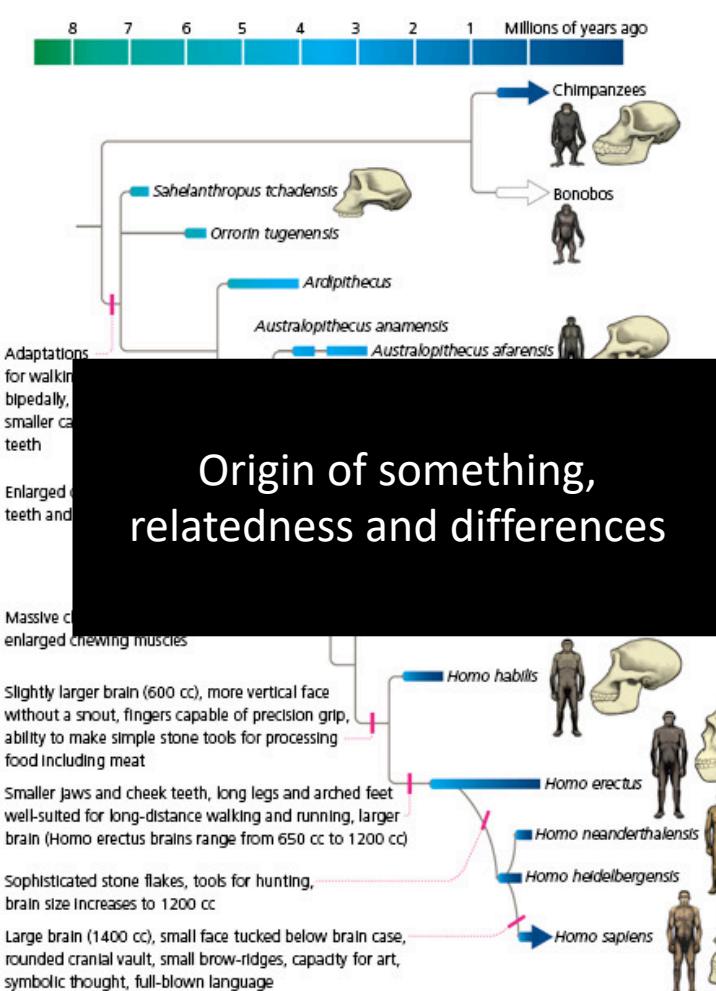


Molecular Evolutionary
Genetics Analysis

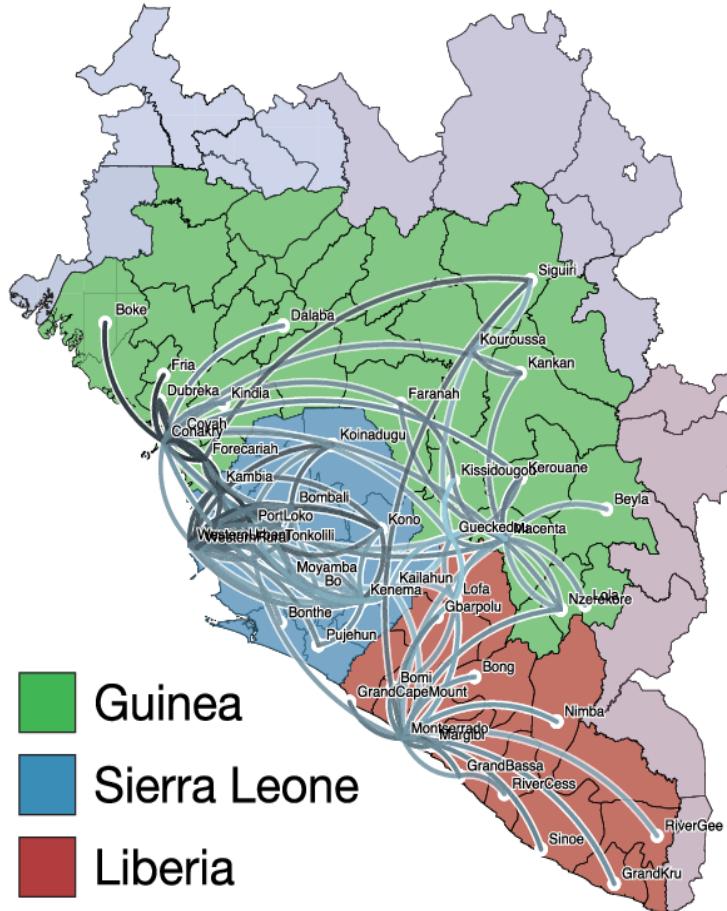
What can you do with phylogenies?



What can you do with phylogenies?

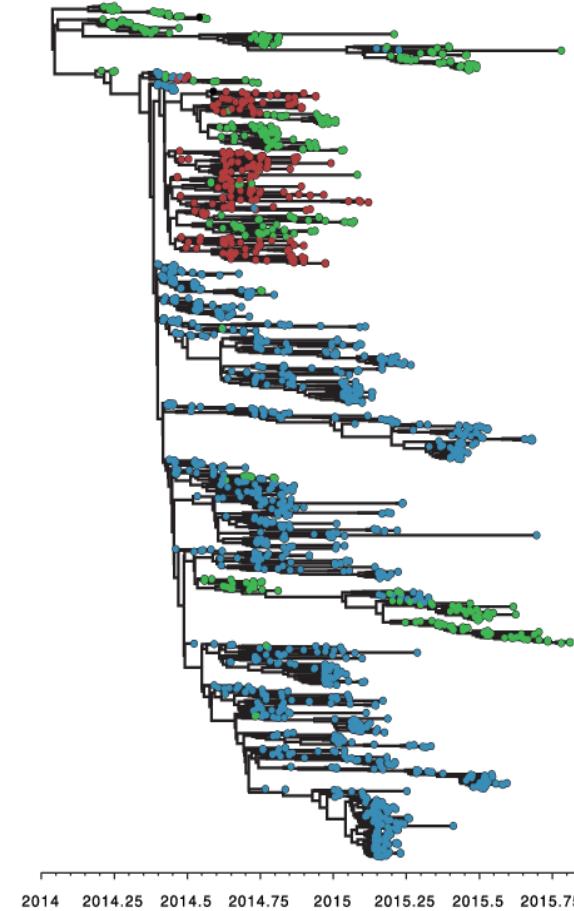


Phyldynamics



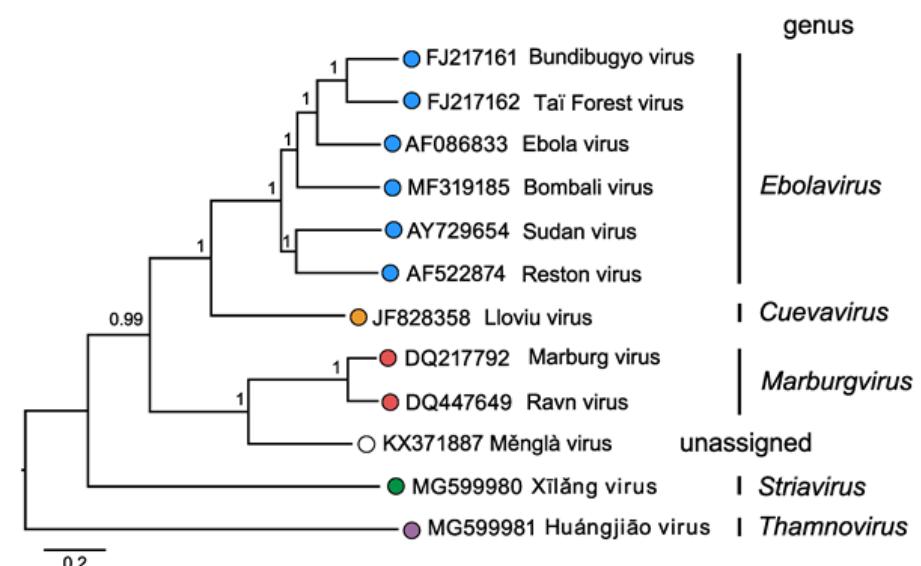
WHERE does it go?

Bayesian trees



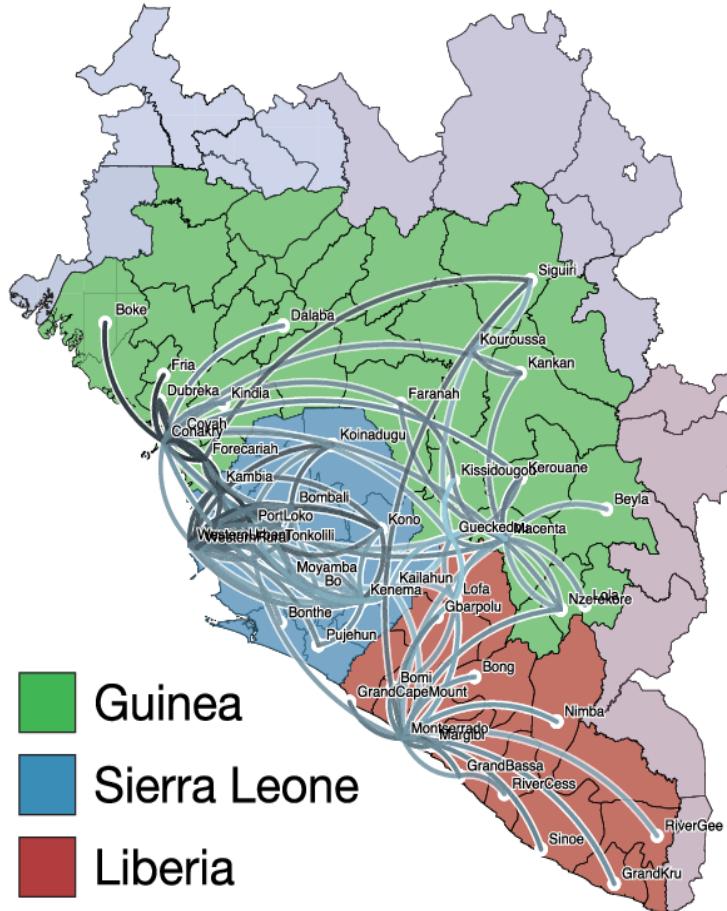
WHEN is the most recent common ancestor?

Maximum likelihood



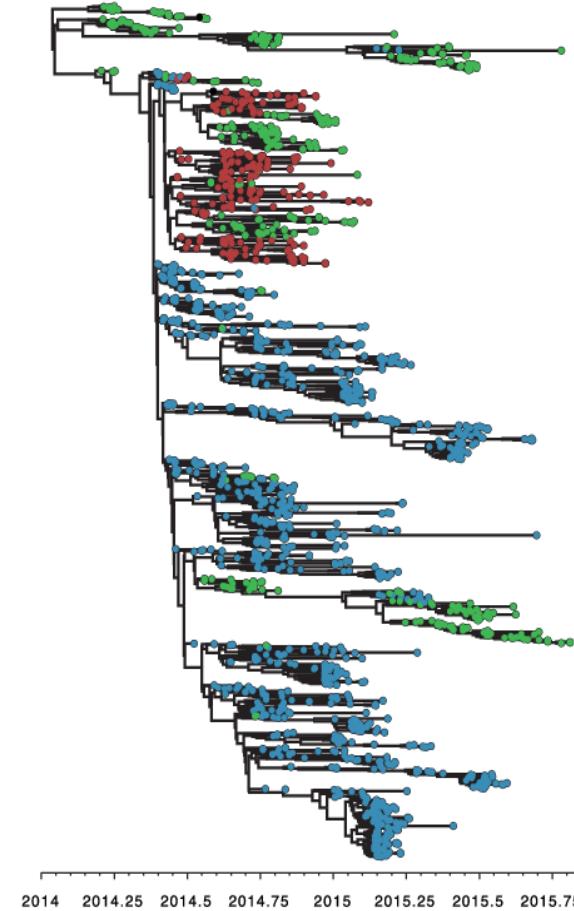
HOW different is it to what's known?

Phyldynamics



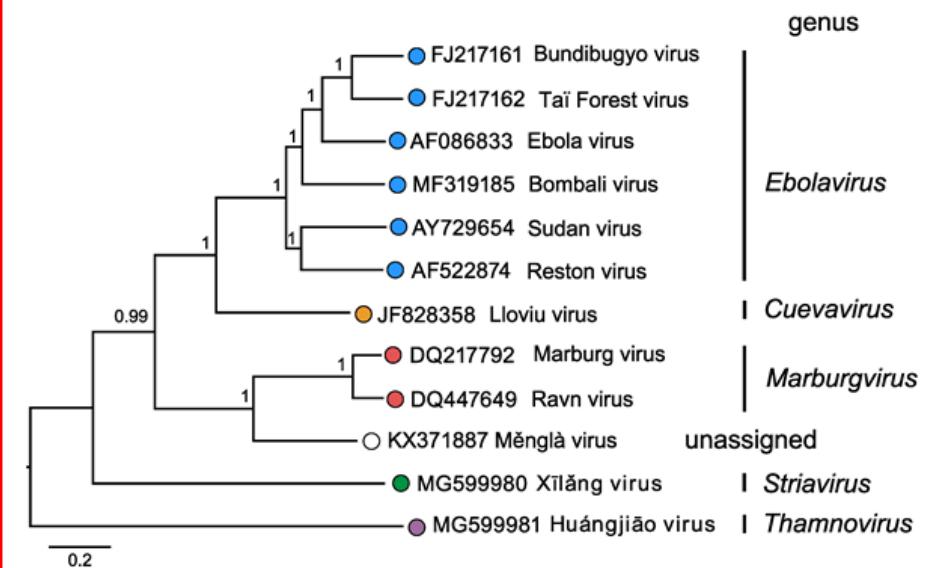
WHERE does it go?

Bayesian trees



WHEN is the most recent common ancestor?

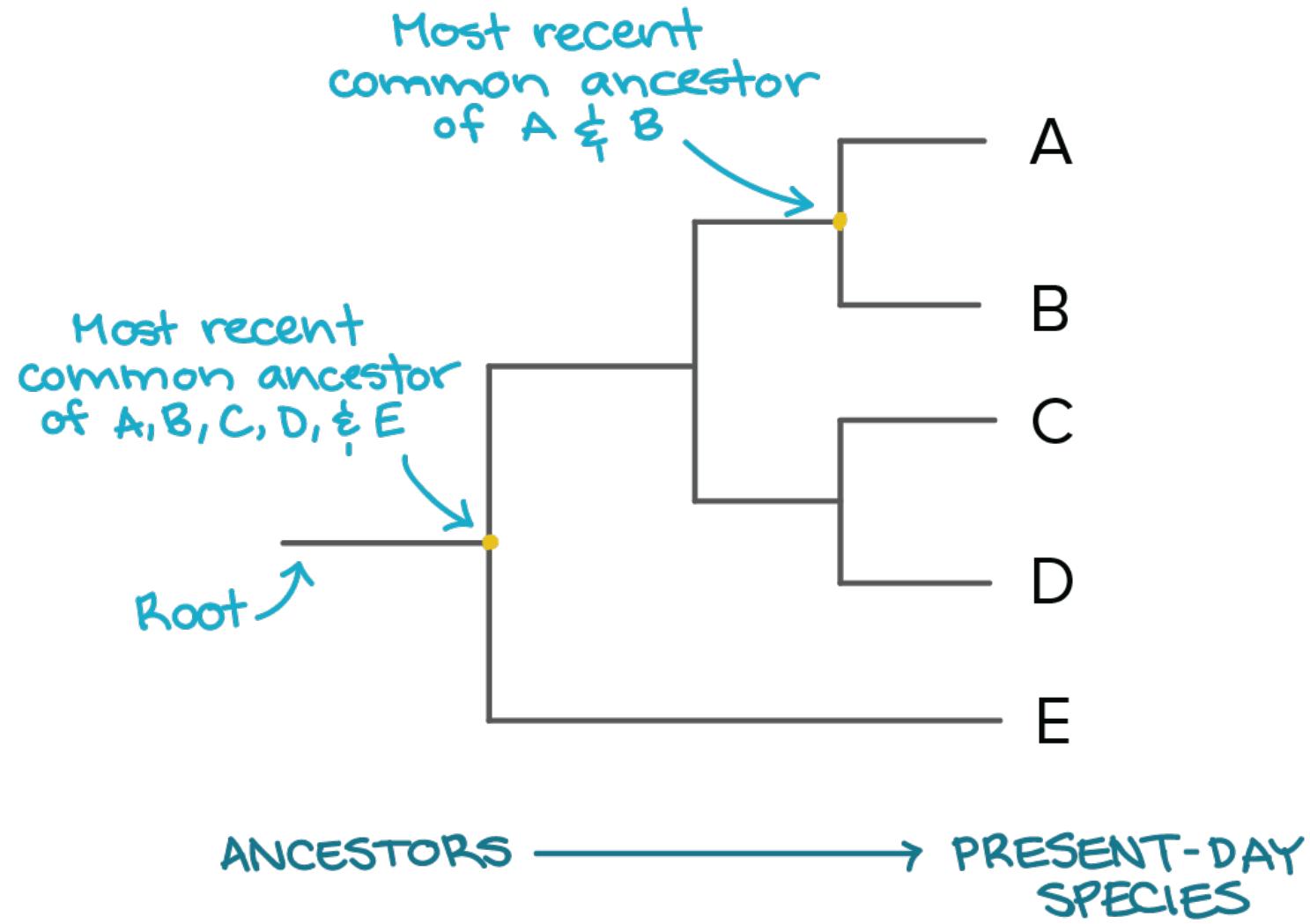
Maximum likelihood



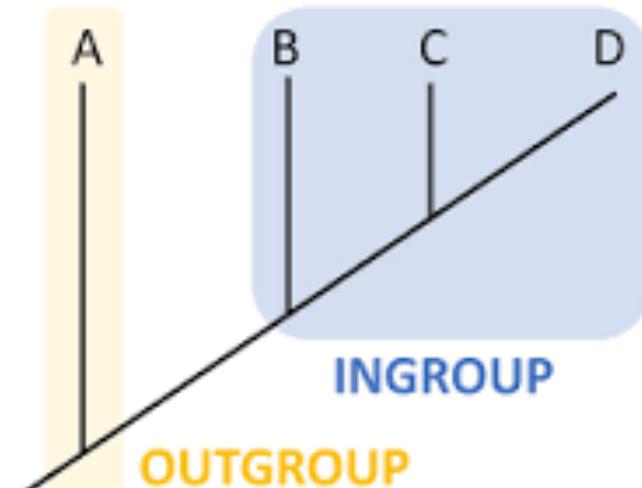
HOW different is it to what's known?

Anatomy of a phylogeny

CONFIDENCE

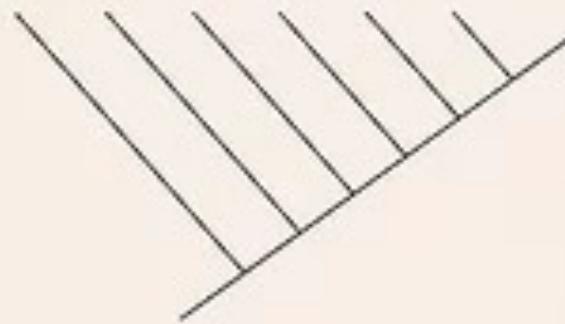


BOOTSTRAP VALUE	
STRONGLY SUPPORTED	>90%
WELL SUPPORTED	70%-90%
WEAKLY SUPPORTED	50%-70%
NOT SUPPORTED	<50%

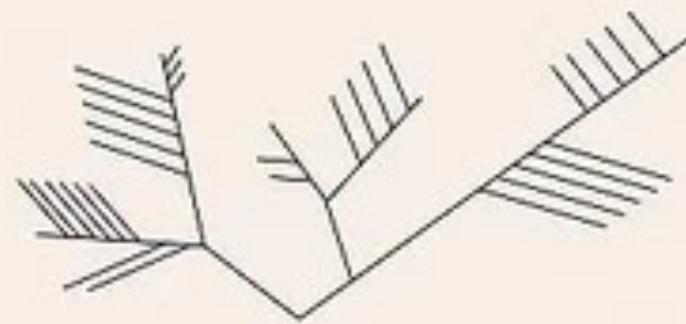


Cladogram versus phylogenetic tree

CLADOGRAM



PHYLOGENETIC TREE

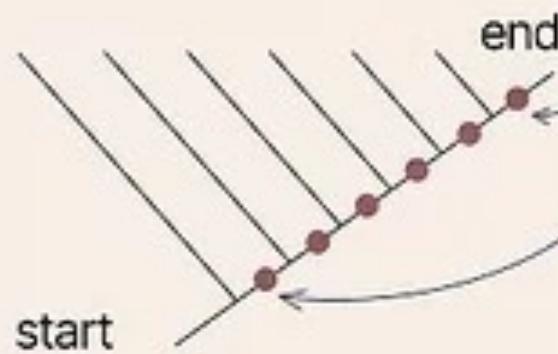


- the relationships are *hypothetical*
- you can easily make on your own

- the relationships are *backed by molecular evidence*
- should have access to DNA or other molecular data

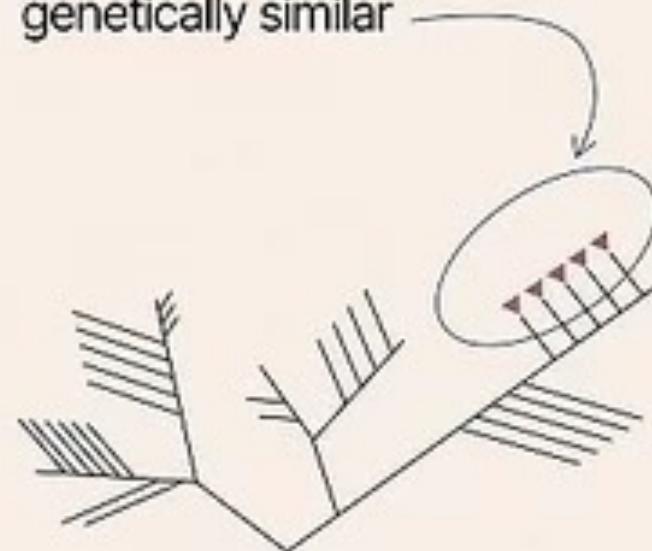
Cladogram versus phylogenetic tree

Nodes closer to the start of the main line happened longer ago than nodes closer to the end



CLADOGRAM

Animals that are closer together are also more genetically similar



PHYLOGENETIC TREE

Why aren't we using cladograms for epidemiology instead of phylogenetic trees?

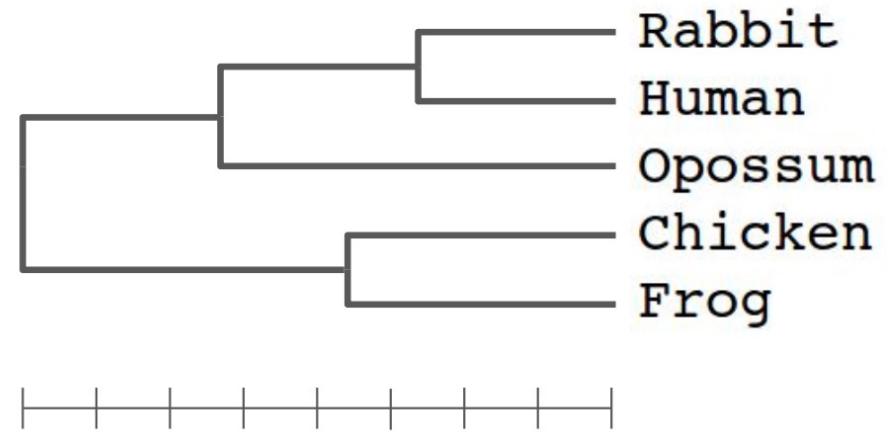
Parsimony versus likelihood

- Parsimony: minimum number of changes
- Likelihood: maximum probability of the data having evolved on the tree



branch length can mean different things:

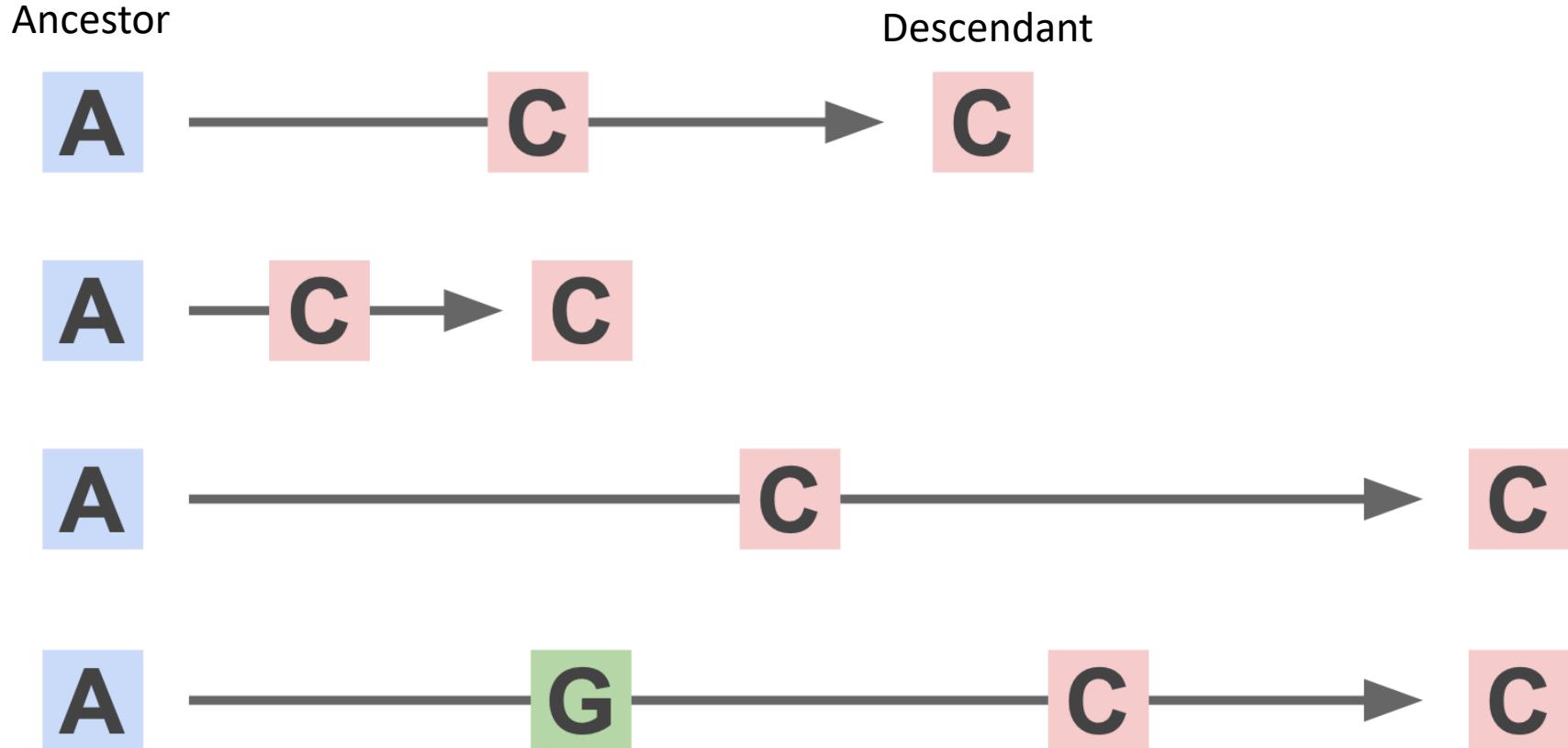
- minimum number of changes (parsimony)
- time; opportunity for change
- expected number of changes, given a model of evolution



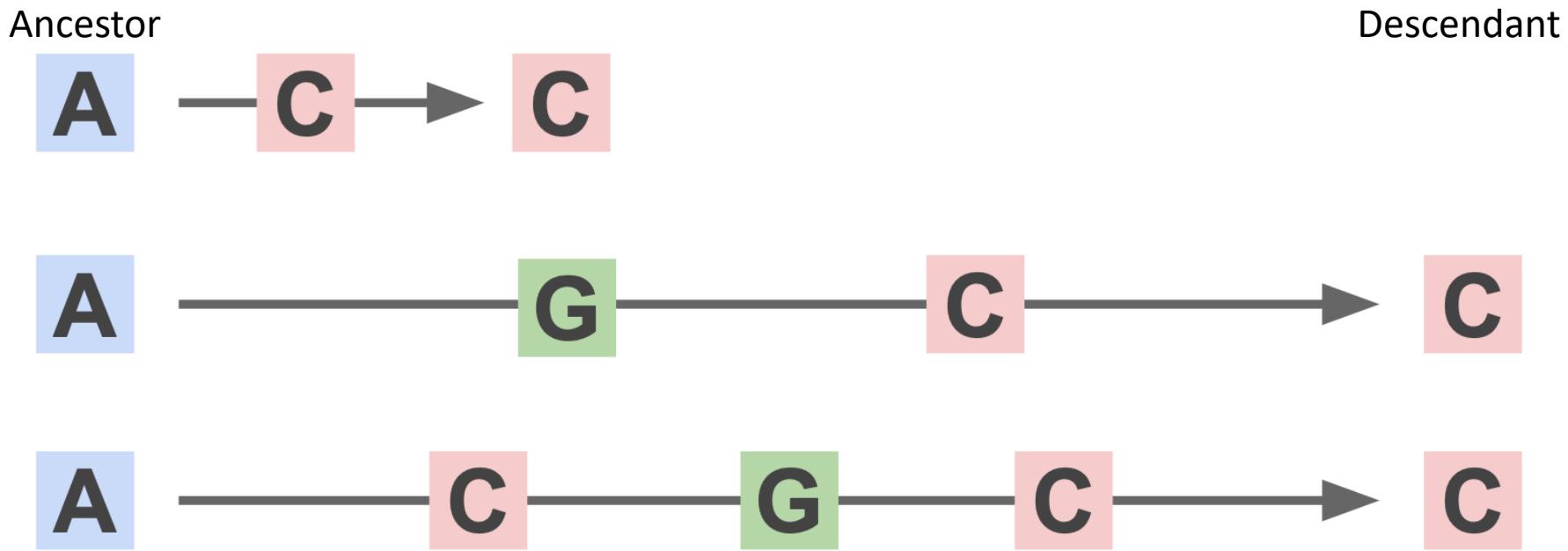
proposed tree has **branch lengths** in units of expected number of changes per site

Parsimony: minimum number of changes
regardless of time/opportunity

**Which is the
least
parsimonious?**



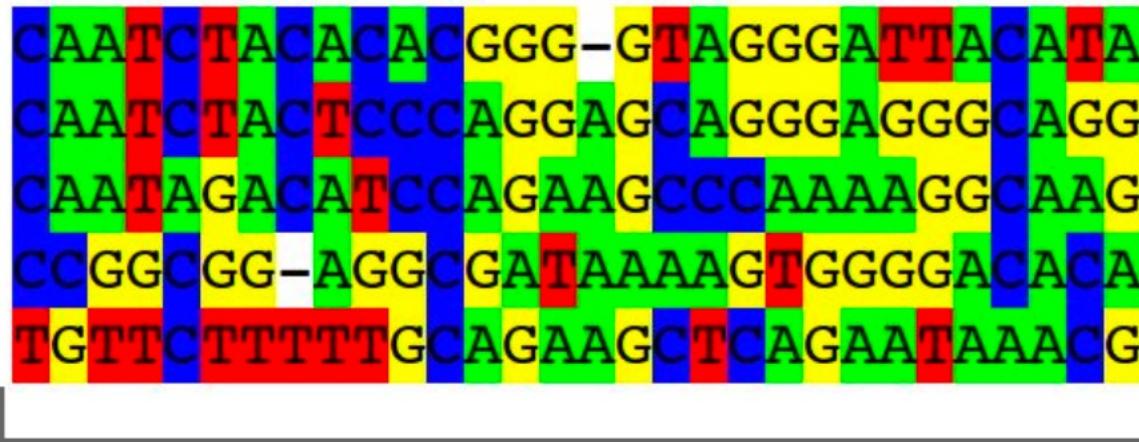
Likelihood: probability of ancestral and descendant status is a function of time (branch length)



We don't know what the actual history of the change is, so use a model of evolution to consider all possible histories (**maximum likelihood**)

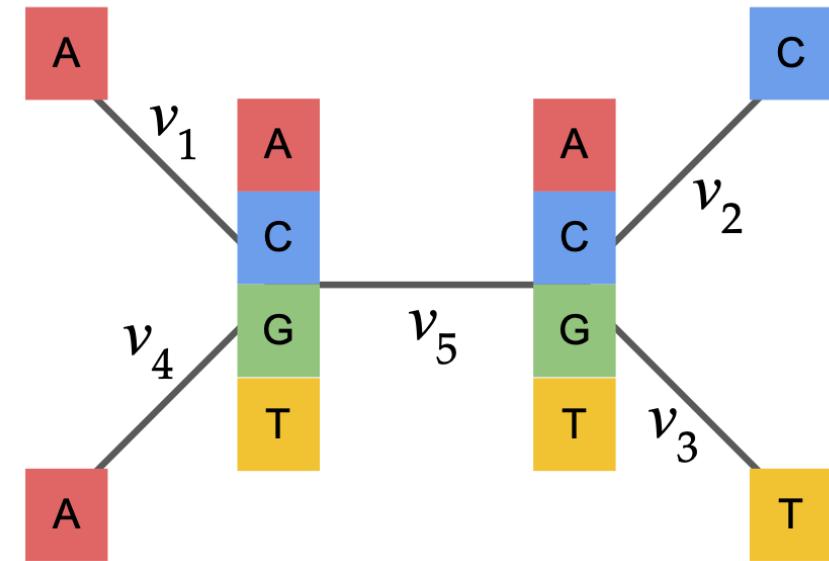
Likelihood cont'd.

Rabbit
Human
Opossum
Chicken
Frog



overall likelihood is the product of likelihoods across characters (sites)

Parameters: tree topology, branch lengths, substitution rates estimated to maximize likelihood of data



Consider *all possible ancestral states* at internal nodes, and calculate their contribution to the overall likelihood.*

Models of DNA evolution

- Markov models that describe relative rates of different changes
 - JC69 (Jukes and Cantor 1969)
 - K80 model (Kimura 1980)
 - K81 model (Kimura 1981)
 - F81 (Felsenstein 1981)
 - HKY85 model (Hasegawa, Kishino and Yano 1985)
 - T92 model (Tamura 1992)
 - TN93 model (Tamura and Nei 1993)
 - GTR model (Tavaré 1986)
 - Yep there's a lot of them! How do I know what's best for my data?

Good news, most people don't need to know the mathematical specifics of these models

JC69 model (Jukes and Cantor 1969) [\[edit\]](#)

JC69, the [Jukes and Cantor 1969](#) model,^[2] is the simplest substitution model. There are several assumptions. It assumes equal base frequencies

$(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4})$ and equal [mutation rates](#). The only parameter of this model is therefore μ , the overall substitution rate. As previously

mentioned, this variable becomes a constant when we normalize the mean-rate to 1.

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

When branch length, ν , is measured in the expected number of changes per site then:

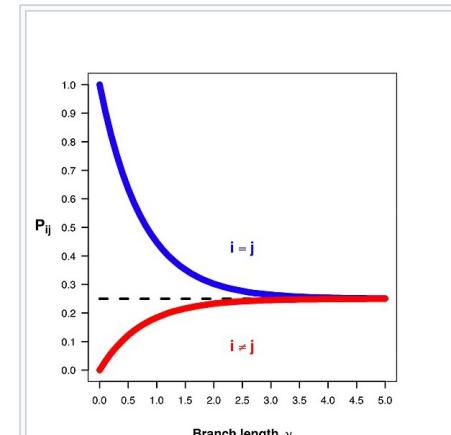
$$P_{ij}(\nu) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

It is worth noticing that $\nu = \frac{3}{4}t\mu = (\frac{\mu}{4} + \frac{\mu}{4} + \frac{\mu}{4})t$ what stands for sum of any column (or row) of matrix

Q multiplied by time and thus means expected number of substitutions in time t (branch duration) for each particular site (per site) when the rate of substitution equals μ .

Given the proportion p of sites that differ between the two sequences the Jukes-Cantor estimate of the evolutionary distance (in terms of the expected number of changes) between two sequences is given by

$$\hat{d} = -\frac{3}{4} \ln(1 - \frac{4}{3}p) = \hat{\nu}$$



Probability P_{ij} of changing from initial state i to final state j as a function of the branch length (ν) for JC69. Red curve: nucleotide states i and j are different. Blue curve: initial and final states are the same. After a long time, probabilities tend to the nucleotide equilibrium frequencies (0.25: dashed line).

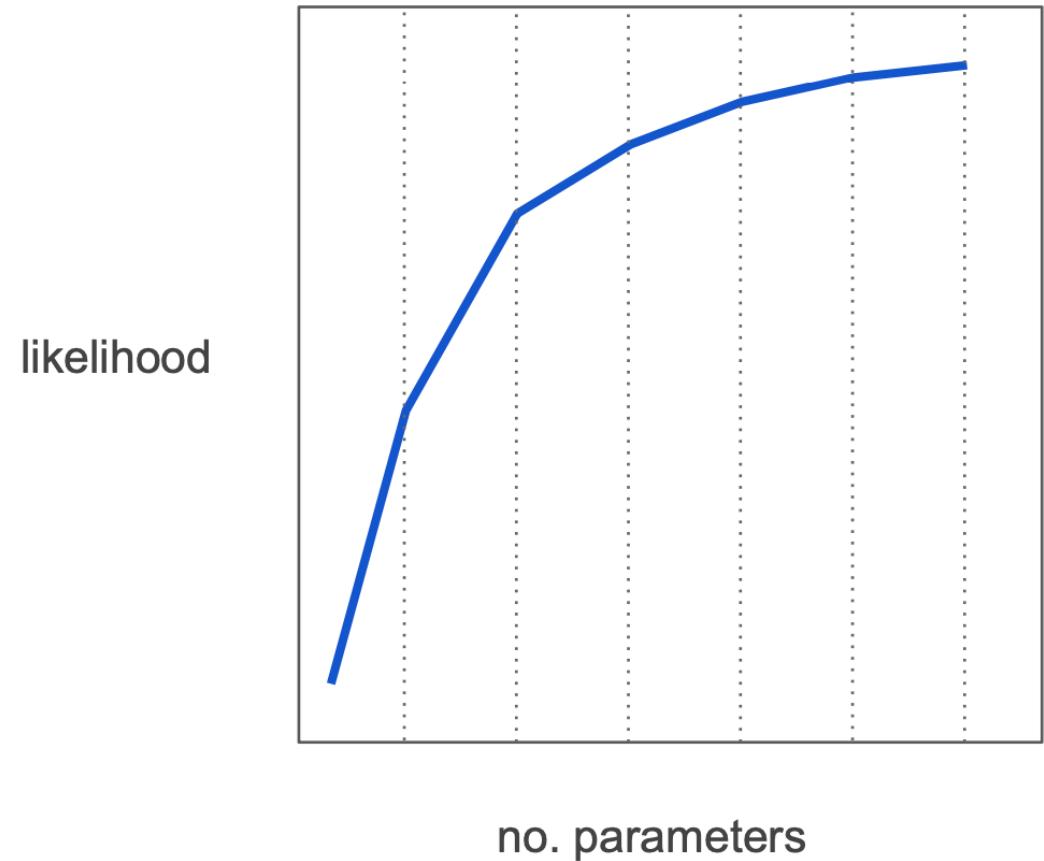
Model selection

More parameters means higher likelihood, but is the increase in likelihood necessary? Adds much more complexity

- Programs will use statistical methods to answer this question using Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC), likelihood ratio tests, etc.

Model testing will give you BIC and AIC score

- AIC score: tries to select the model that most adequately describes an unknown, high dimensional reality
- BIC score: tries to find the TRUE model among the set of candidates



DNA models

Base substitution rates

IQ-TREE includes all common DNA models (ordered by complexity):

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like HKY but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like TN but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like K81 but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like TPM2 but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like TPM3 but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIMe	3	Like TIM but equal base freq.	012230

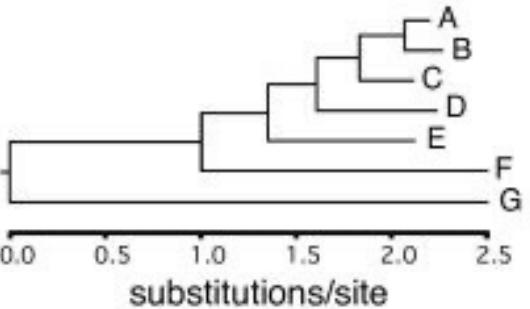
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like TIM2 but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like TIM3 but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like TVM but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

Rate heterogeneity across sites

- Do we expect all sites in an alignment to evolve at the same rate?
What kind events would affect this?

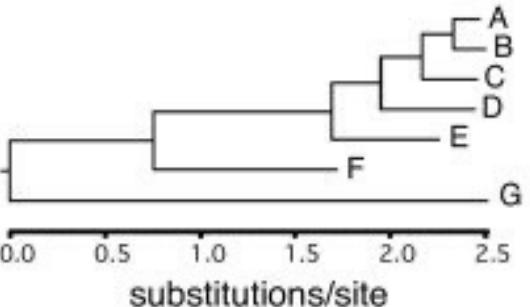
A Alignment simulated with rate heterogeneity

A TTEQGIKSGTSIPAPQLPNWSGQYHEWVLKS---FQNEVK----KTLHCALSQGTITATOSVLDELHADVWALLASSEVCYAKPCGQVKPELAFLRKRA
B TAGAGEXTGTSLPAPNLPNWSGQYHEWVLKS---ERADVI----KTMHCRALSDGTTATOSVLDELHPNVWALLASSEVCYARPCGQVKPELIYMKVKQA
C VAGGCCEKAGTSIPAPYLPN-SGQYGEWVLKS---LSTHVI----KHMHCGLSDSDTITOSVLDELHGERWALLOSSEVCYAKPCGQEKRVLEHECYKRA
D CACQAEKXTGTSLPALHLPNWS-QYGEWVLKS---FPSQPV----MPIQCVPLSDARTAQSVLDELHVESDALDSSEVCYAAPCGA-RHDLXFCVCYSKA
E DEGLTQKXTGTSLPALALPNWSGQYTFEWVLKS---YG---FGQGGGAAICCKPLSGDKTSIHSVLDELHAVLAALLMSGEVCYALPCGAYKKALEFKCYLKKA
F GEQFIKKXTGTSAAPAVLPDFAEQYDEWPALKSTLAYGRVNFF---AAVPGAYLSDFTGTGSHSVLDELHDNHAALLSSEVCFAAPCGESRGALVVVCYSHA
G NDGPHIKKXGTSGPAAELPNQDPIOYQEWVLKS---CEAKSI----NGSNWKPLSGKYTGLOJVLDELHAMKDALLHATBVCLAFPCGY-TADLXAALYGP



B Alignment simulated without rate heterogeneity

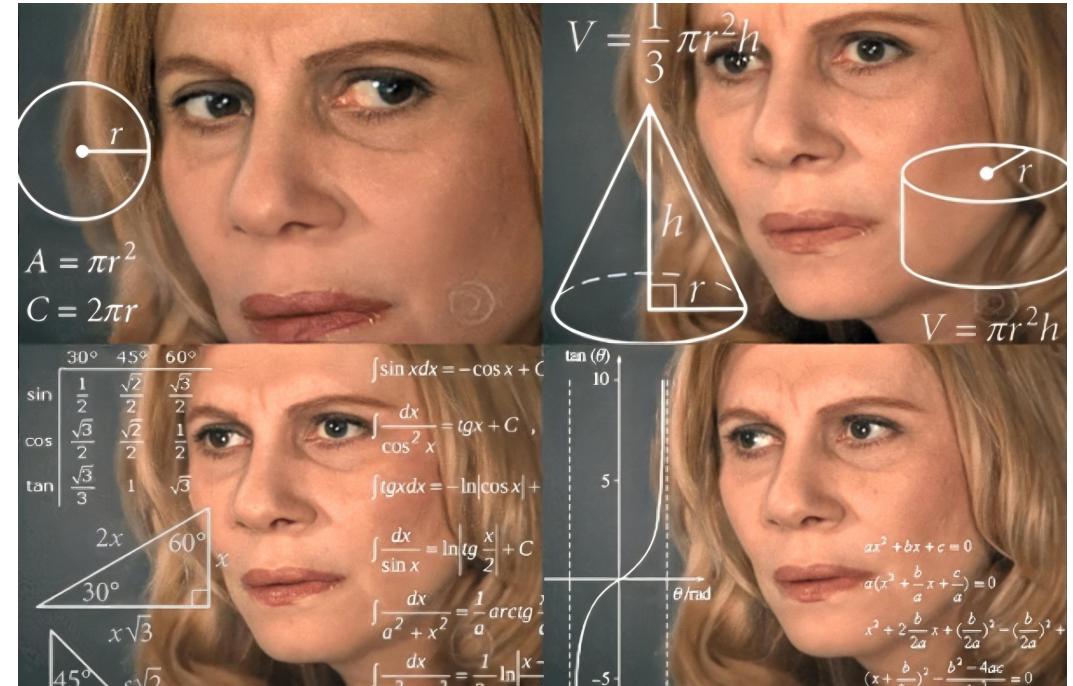
A H-----ENYPC--SQVAKYLAF-YSHNYL--EALLRHATLEIQR--RKSNAEHGTGLEGPESA-FDPR--VPAGNEKLLGKVWNNEFSAPNL---IKRP
B Q-----GENYPC--GGVAKYLAW-VGHNYL--EALLRHATHEINR--KKSKEEQGNGLDGPESA-FEPR--IPAGEEKLLGNYHWNMFCAASL---VKKP
C S-----ENYPC--PQVAKYLAW-MSNNYL--HAFLTQAKLEIER--KRNOAEHGCGLQDGNGO-FDPR--IKNGGOKLLGGY--KFLKNFSL---FVKP
D Q-----AHQYPC--SHIGKYFW-VANAYM--HVLLRYAKLEVER--KTRRADHSTDLVAPNGA-RKSV--LLPGPDKLL-RYHNKFISTPLA---FIKT
E E-----GDQYPC--KEVQKYLAW-VGHGYLRAHALSKHAKLATEK--KMTEDDHNTKLETAEGP-LVPC--IPPLPDTRVAIYANTPFSAQNL---FIKT
F Q-----GKKDLC--ENLN---TN-MQNRLW--QALHK-TITVVQHDGKSSMGDRGCKAIDSKAS-LSPC--VSSCEGYLQKSNQIDFPEVSNTV---YLKS
G NNDPSKPFLPCWYTGIL---ILQCAG-----YLDGETMIGRFQ--STQVGLYSTREFDFRYKCMGPTHKATNNTDTFGDRKAFKKRVSVKAFKQQTAPQ



Rate heterogeneity across sites

- Changes in rate heterogeneity:
 - Codon positions
 - Exons (coding regions) versus introns (non-coding regions)
 - Housekeeping genes versus non-functional genes
 - Structure in RNA (stems vs. loops)

We can make inference about selection from these values, but makes things much more complicated



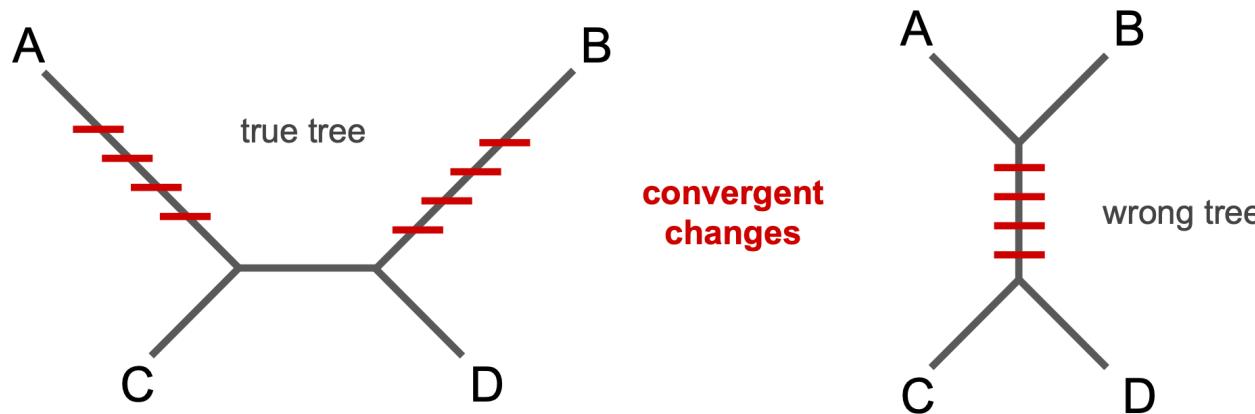
Rate heterogeneity across sites

IQ-TREE supports all common rate heterogeneity across sites models:

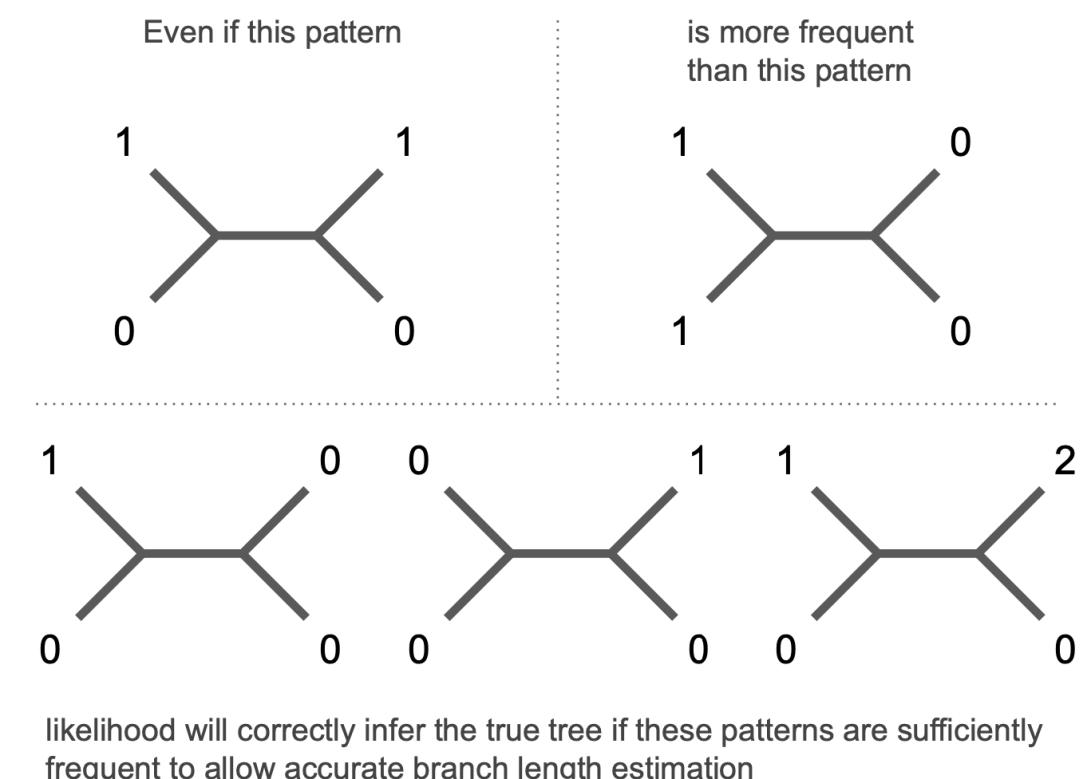
RateType	Explanation
+I	allowing for a proportion of invariable sites.
+G	discrete Gamma model (Yang, 1994) with default 4 rate categories. The number of categories can be changed with e.g. <code>+G8</code> .
+GC	continuous Gamma model (Yang, 1994) (for AliSim only).
+I+G	invariable site plus discrete Gamma model (Gu et al., 1995).
+R	FreeRate model (Yang, 1995 ; Soubrier et al., 2012) that generalizes the <code>+G</code> model by relaxing the assumption of Gamma-distributed rates. The number of categories can be specified with e.g. <code>+R6</code> (default 4 categories if not specified). The FreeRate model typically fits data better than the <code>+G</code> model and is recommended for analysis of large data sets.
+I+R	invariable site plus FreeRate model.

Felsenstein zone

- Branch lengths for which parsimony confidently infers the wrong topology, these can affect bootstrap values

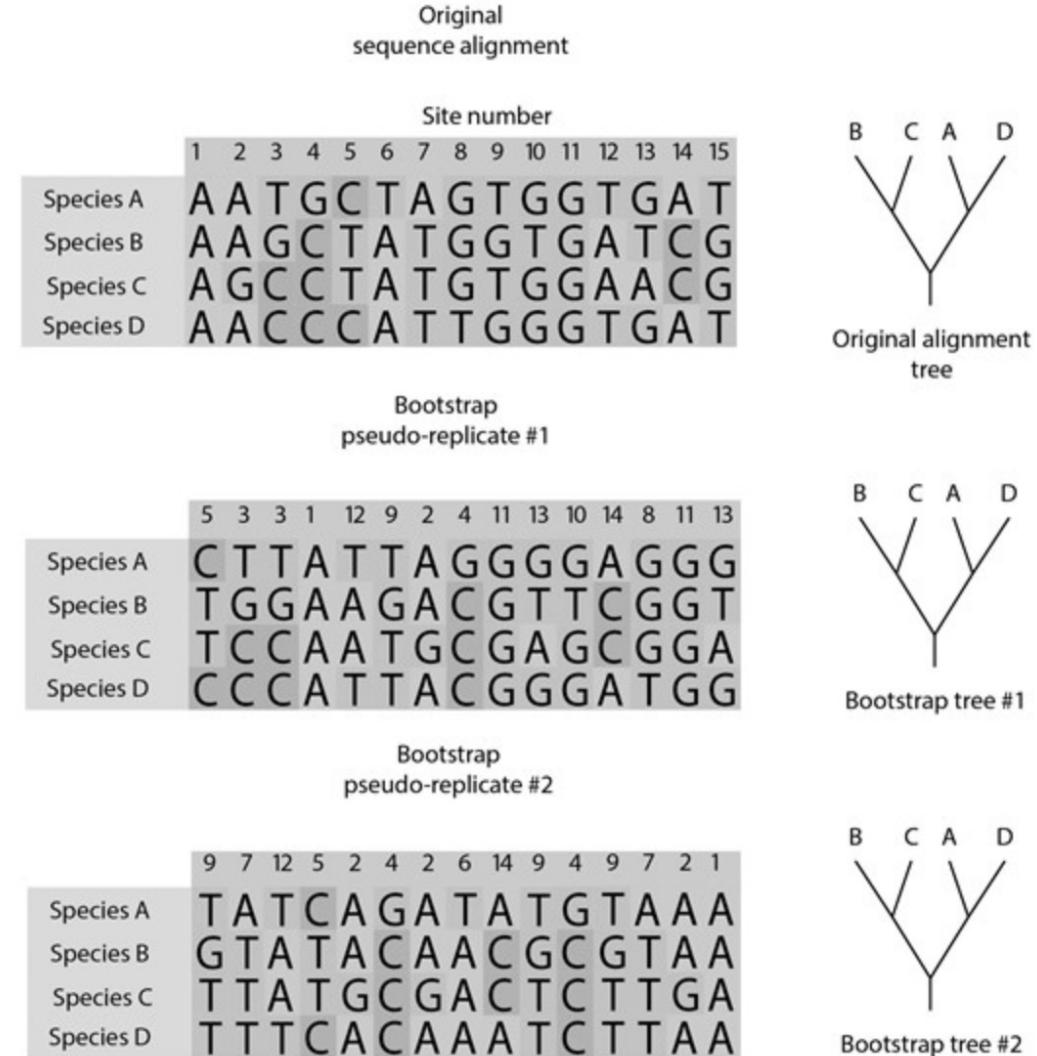


likelihood is a **consistent estimator** of tree topology because it converges on the correct value with increasing data



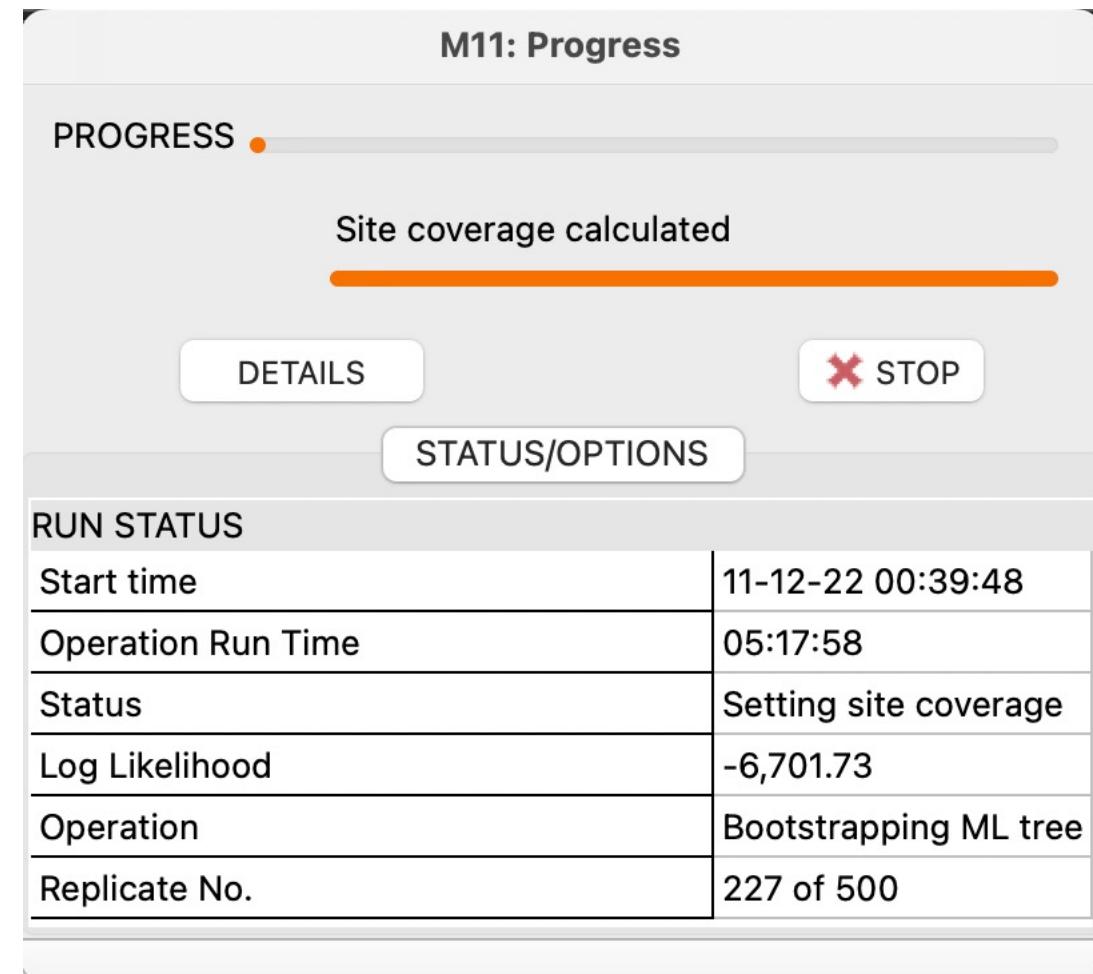
Bootstrapping

- Specify number of replicates: how many times does the test replicate the original sequence alignment?
- Standard in MEGA is 500 replicates, 1000 is better but takes longer



Warnings and limitations

- Building phylogenies takes a LONG time, larger ones can take up to a week to run and Bayesian phylogenies can run for months, so a computing cluster is almost necessary for this
- Without a proper outgroup or root, a phylogeny doesn't tell you much about order of descent
- **Why was it so hard to make phylogenies for SARS-CoV-2?**



Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny

Lenore Pipes, Hongru Wang, John P Huelsenbeck , Rasmus Nielsen 

Molecular Biology and Evolution, Volume 38, Issue 4, April 2021, Pages 1537–1543,

<https://doi.org/10.1093/molbev/msaa316>

Published: 09 December 2020

 PDF  Split View  Cite  Permissions  Share ▾

Abstract

The rooting of the SARS-CoV-2 phylogeny is important for understanding the origin and early spread of the virus. Previously published phylogenies have used different rootings that do not always provide consistent results. We investigate several different strategies for rooting the SARS-CoV-2 tree and provide measures of statistical uncertainty for all methods. We show that methods based on the molecular clock tend to place the root in the B clade, whereas methods based on outgroup rooting tend to place the root in the A clade. The results from the two approaches are statistically incompatible, possibly as a consequence of deviations from a molecular clock or excess back-mutations. We also show that none of the methods provide strong statistical support for the placement of the root in any particular edge of the tree. These results suggest that phylogenetic evidence alone is unlikely to identify the origin of the SARS-CoV-2 virus and we caution against strong inferences regarding the early spread of the virus based solely on such evidence.

Putting your tree in context is important, without a control, can you really infer anything?

How are phylogenies used in global
epidemiology/health?

Case study: Marburg virus – where did it come from?

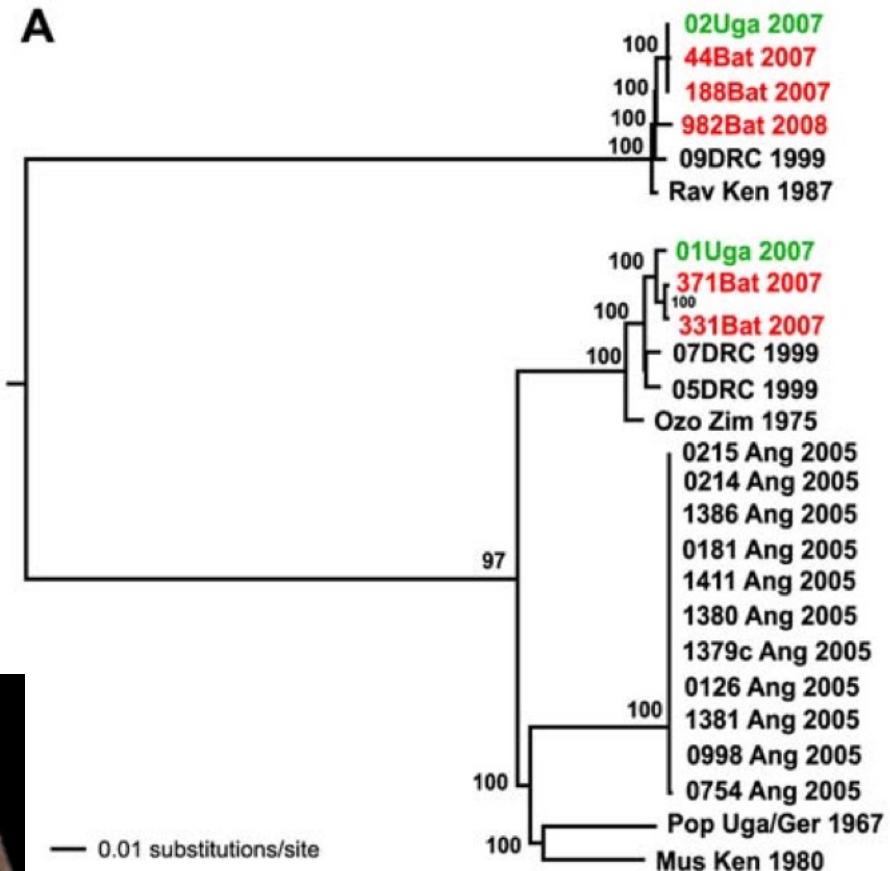
- Marburg filovirus causes severe hemorrhagic fever in humans through sporadic spillovers
- Like Ebola virus, it was suspected to have been hosted by bats, but we did not know for sure

What do we need to see if the cases in humans were spillovers from bats?



Case study: Marburg virus – where did it come from?

- Marburg filovirus causes severe hemorrhagic fever in humans through sporadic spillovers
- Like Ebola virus, it was suspected to have been hosted by bats, but we did not know for sure

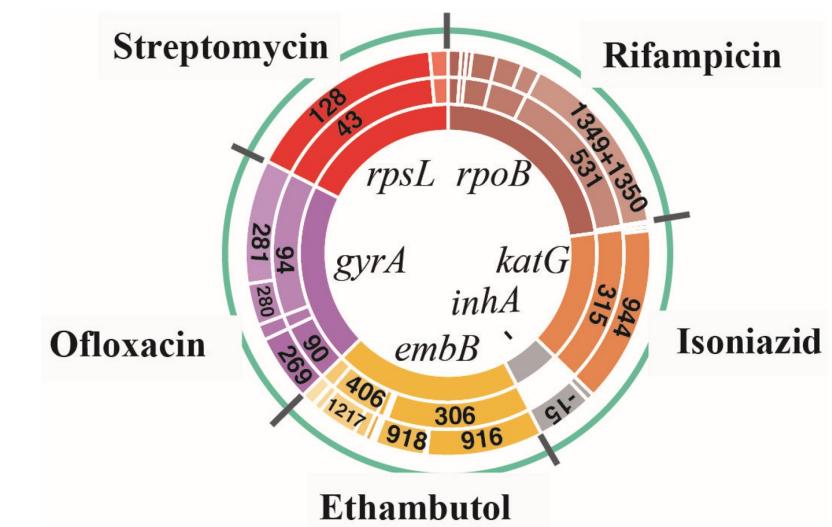
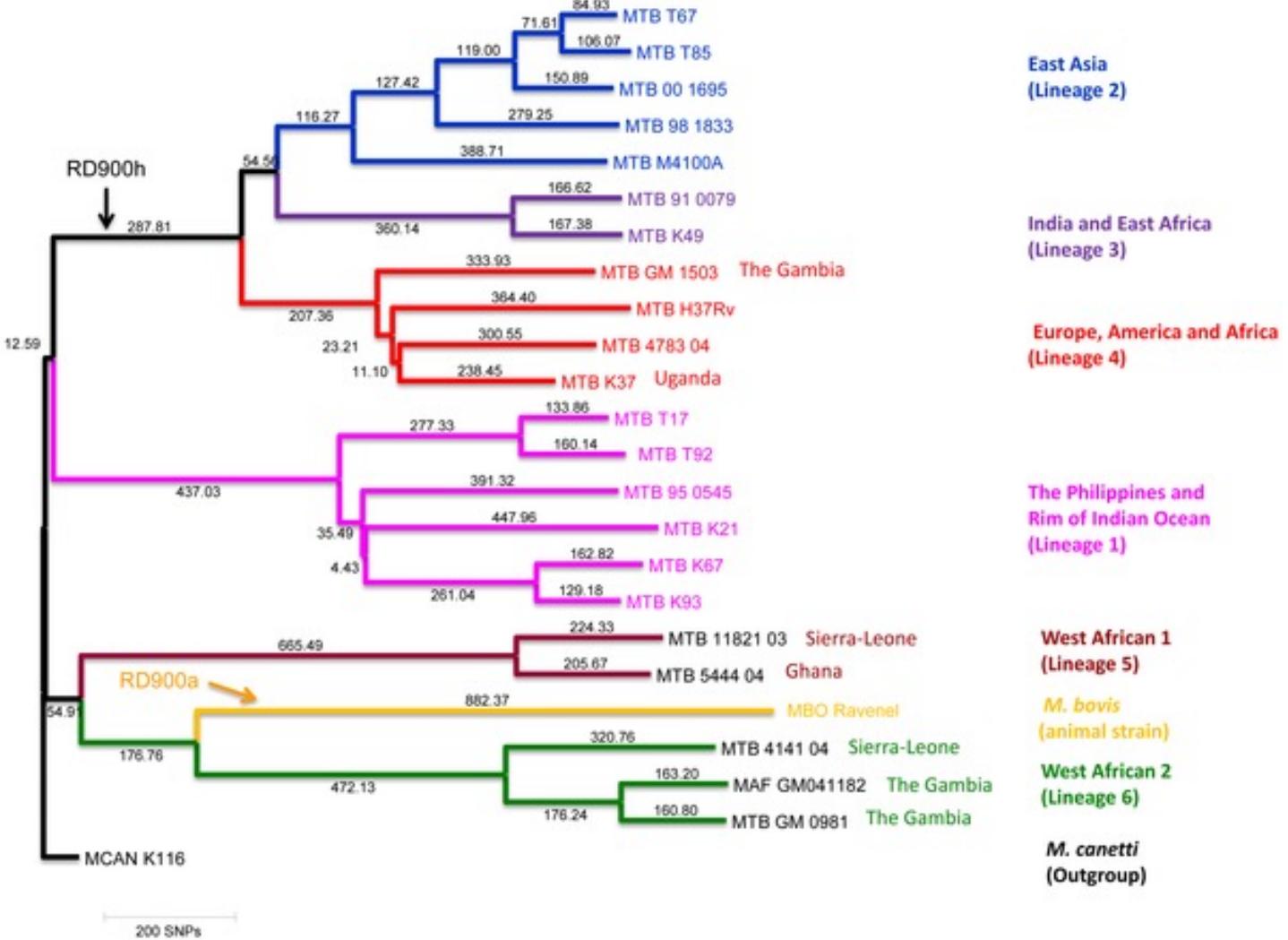


Green: human sequences

Red: bat sequences

Black: reference sequences

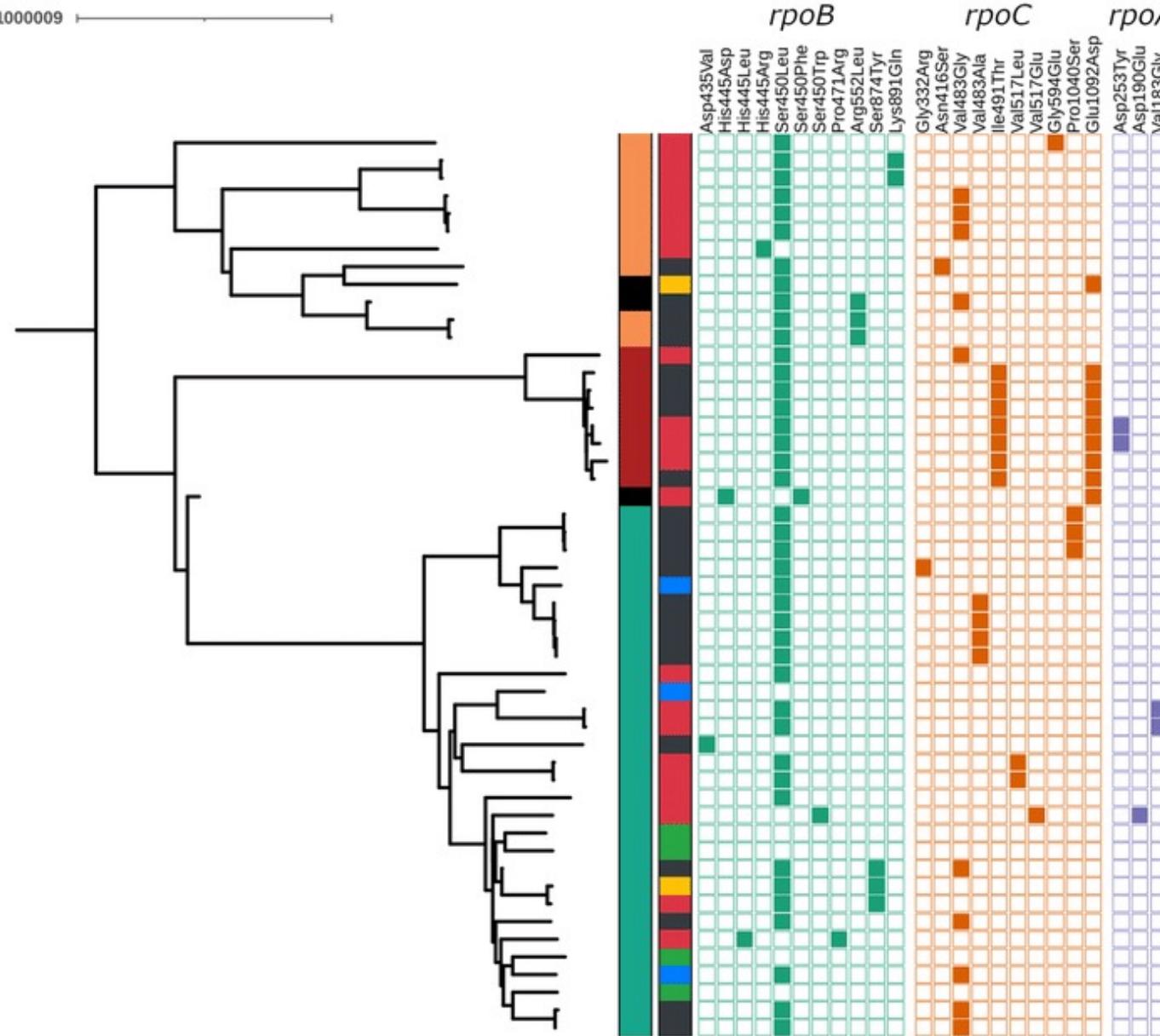
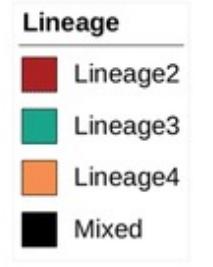
Topic: tuberculosis



What do you think the above plot is telling you?
It has to do with drug resistance...the phylogeny can help us track something...

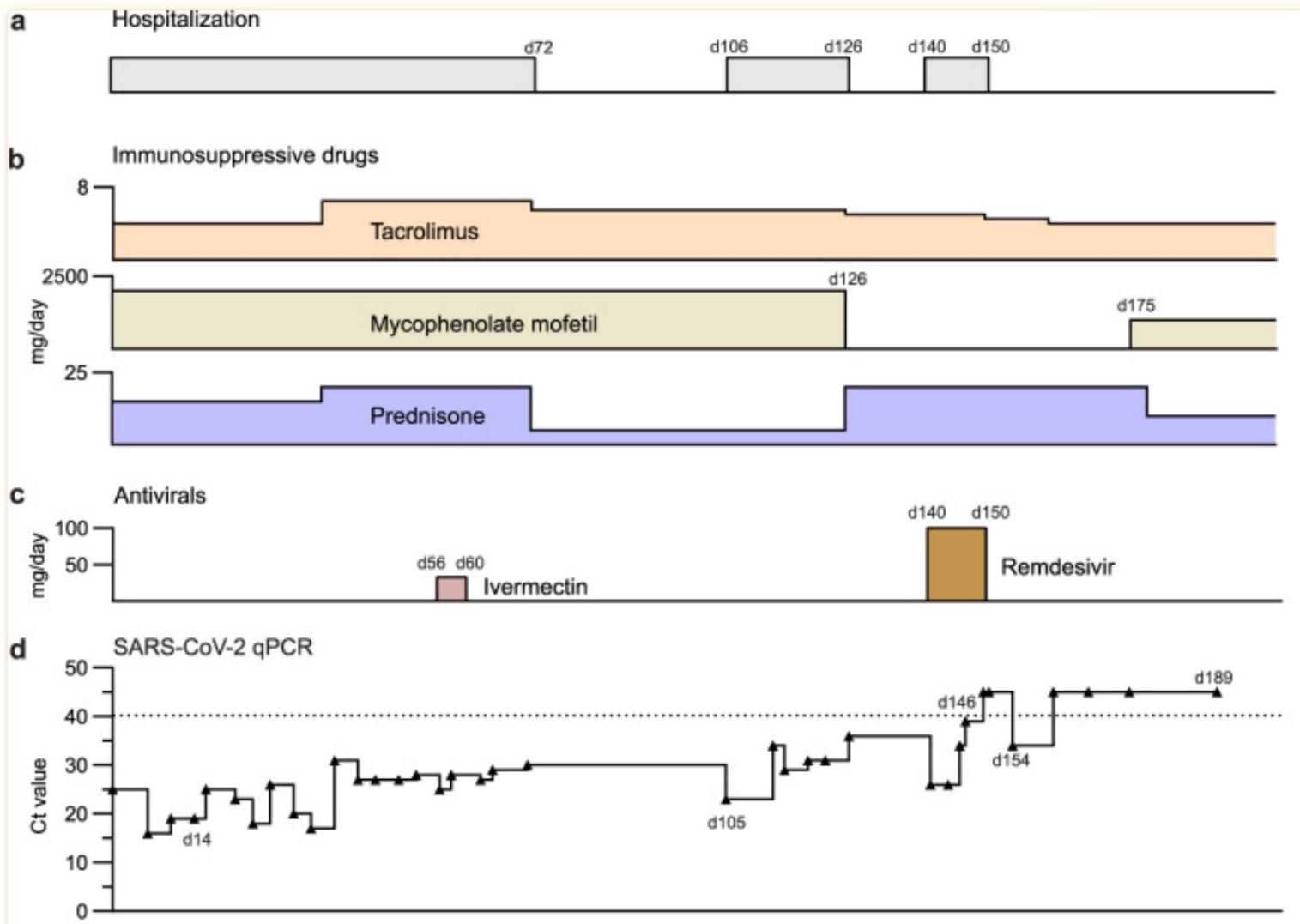
Topic: tuberculosis

Tree scale: 0.00010000100001000009



We can track mutations associated with drug resistance and visualize it next to a tree

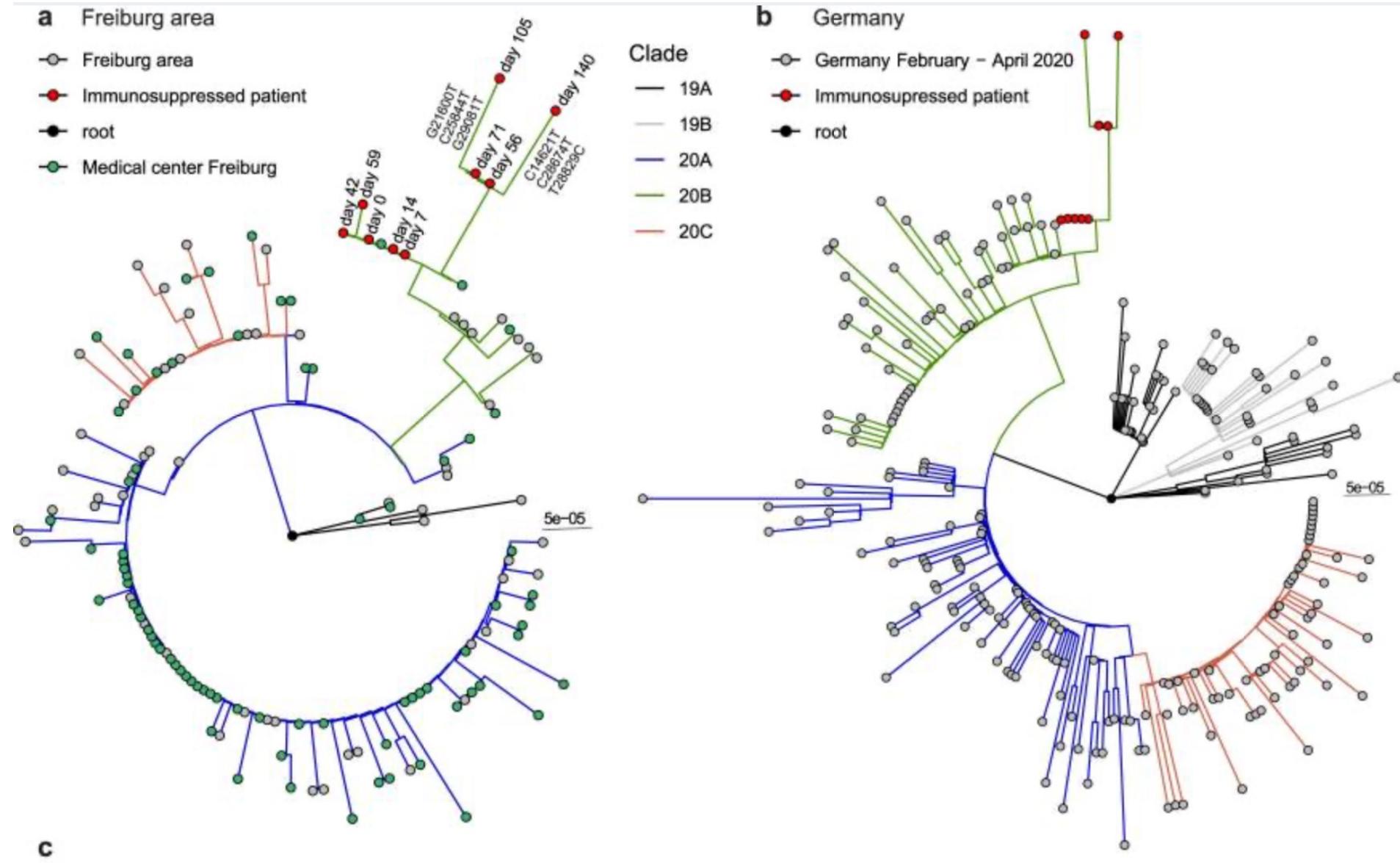
Case study: within-host evolution of SARS-CoV-2



"A 58-year-old male with a history of autosomal dominant polycystic kidney disease was admitted to our university hospital, for renal transplantation performed on March 2020. The patient was treated with a cocktail of tacrolimus, mycophenolate, and prednisone from March until the end of September"

What do you expect to happen to the virus?

Case study: within-host evolution of SARS-CoV-2



Case study: a literal court case

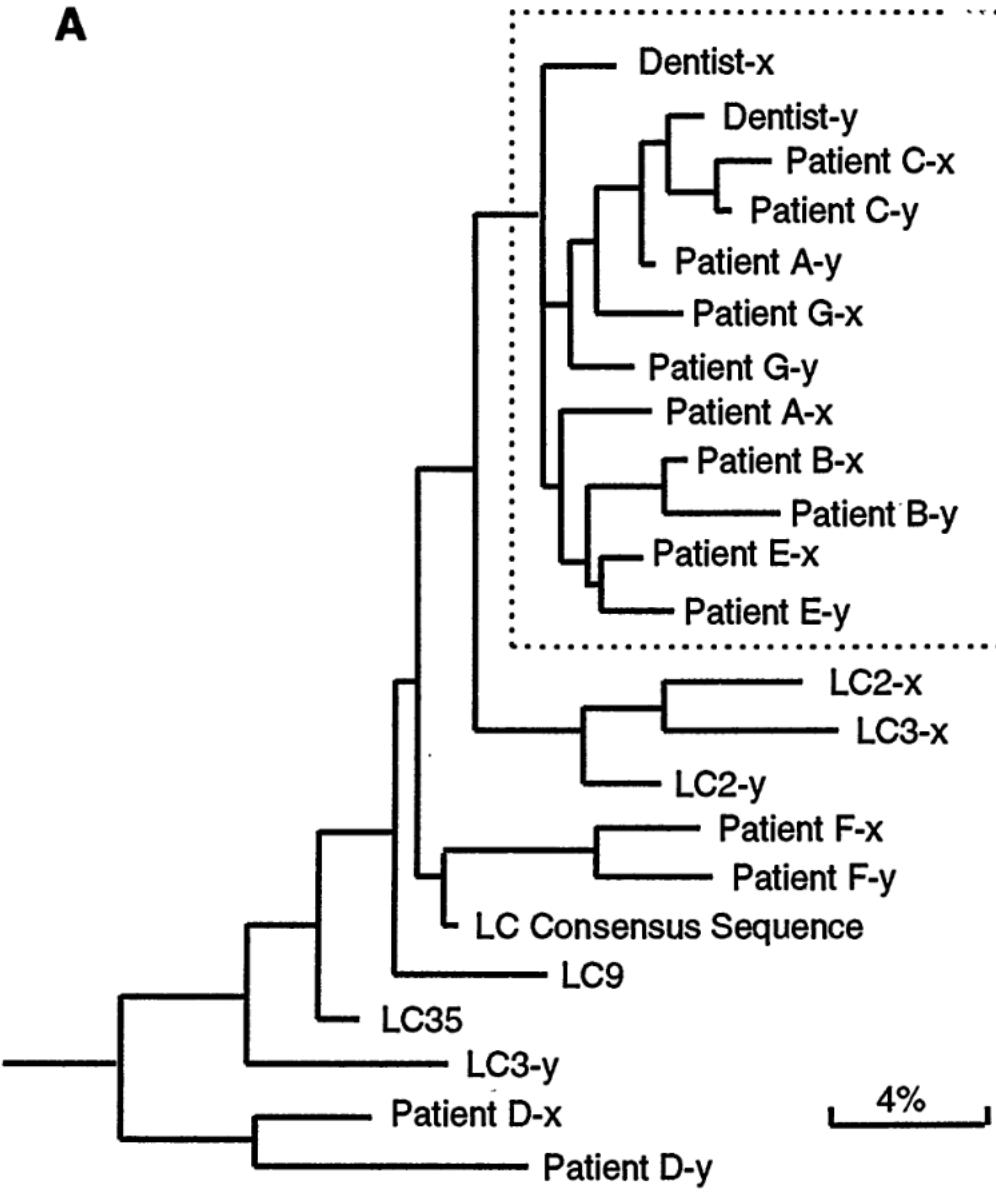
In 1990:

There was a case of a dentist who possibly infected 6 of his patients with a virus, a phylogenetic analysis was used to support this statement

Considered first case of a health worker giving a disease to a patient

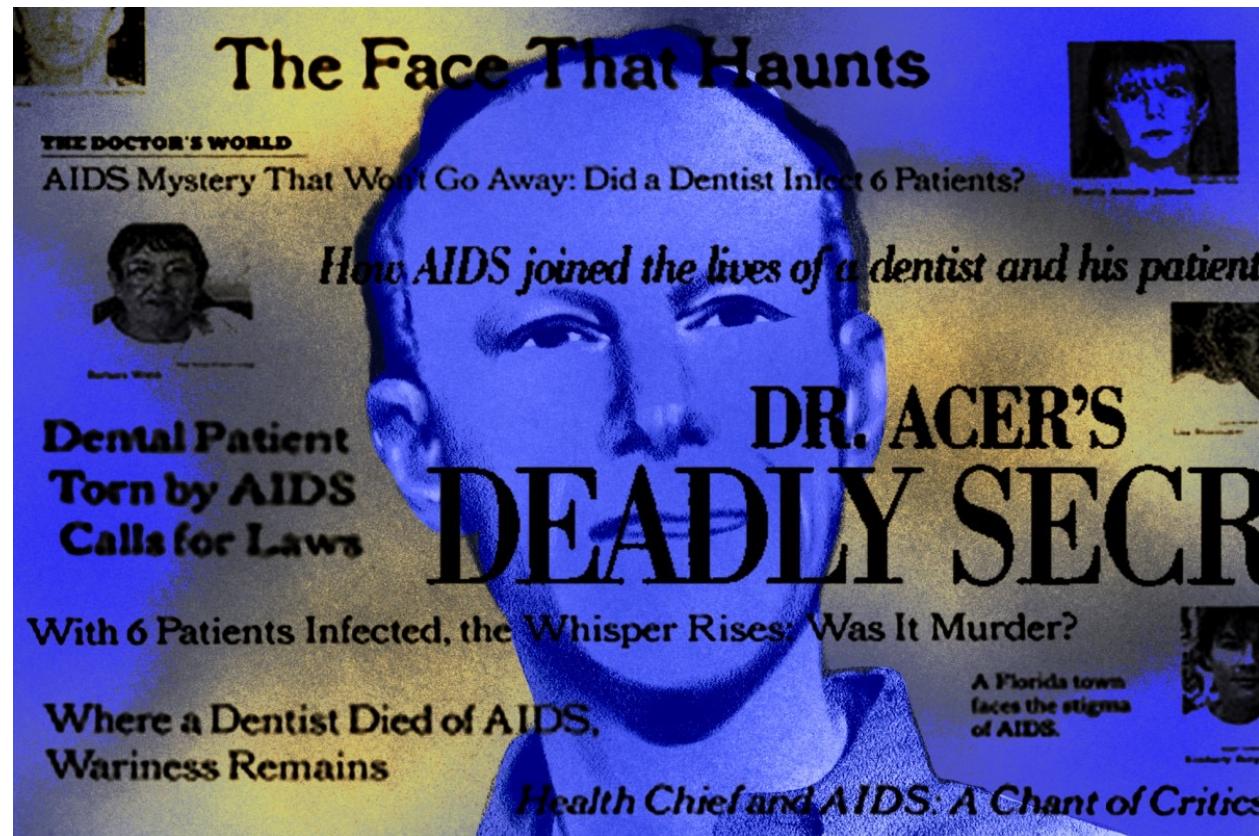


A

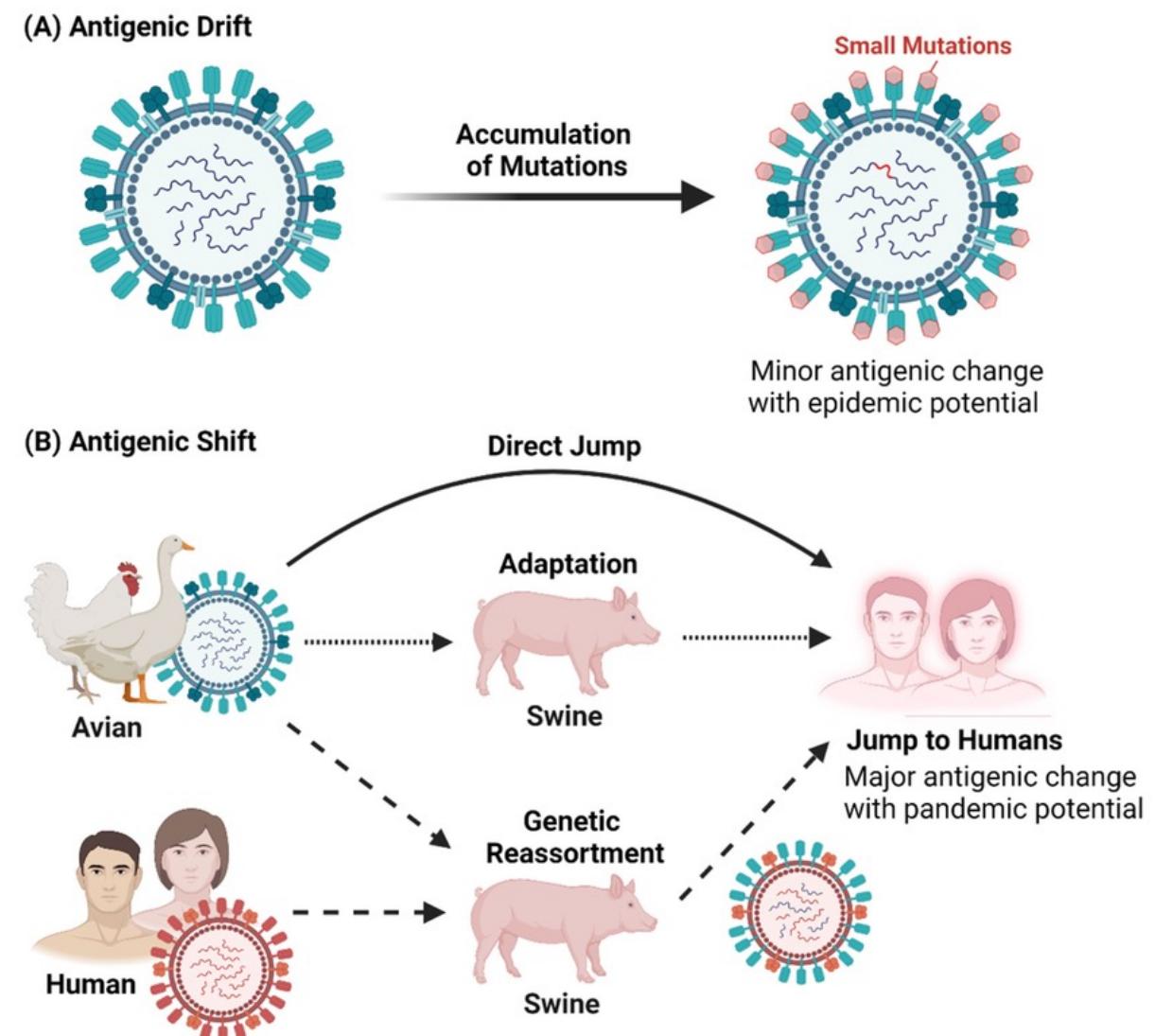
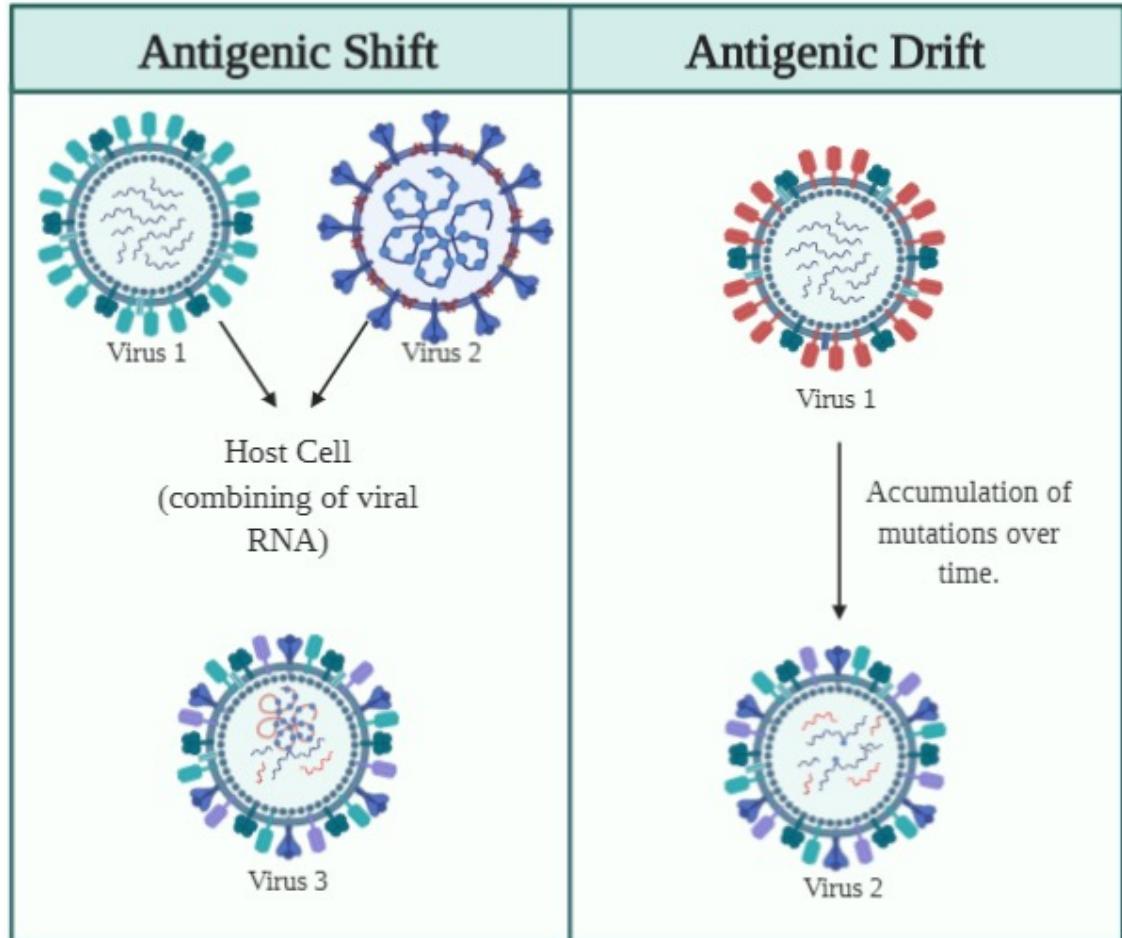


Case study: a literal court case

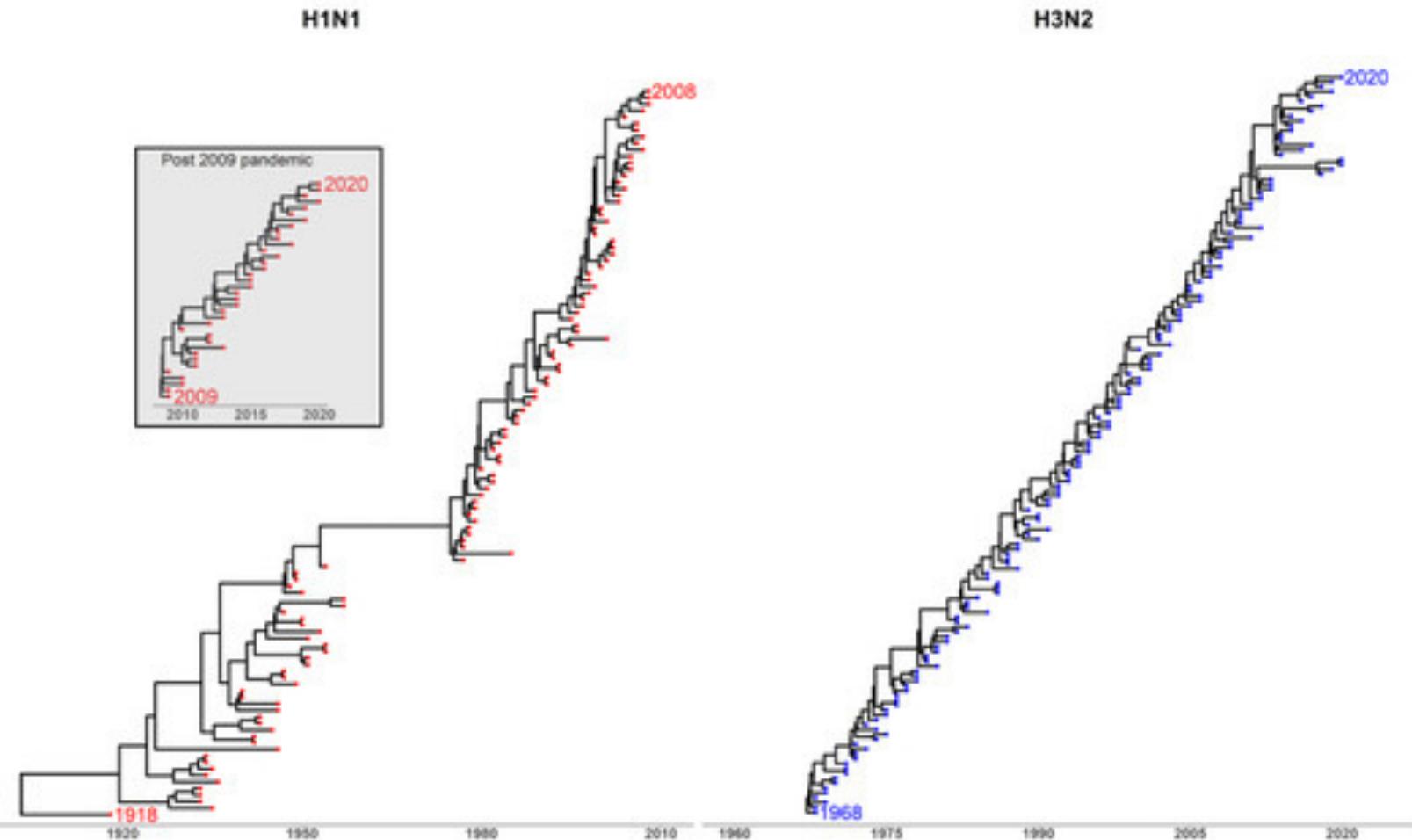
- David Acer was a dentist who practiced in Florida throughout the 1980s, by 1986 he knew he had AIDS
- As he got sicker, he continued to see patients. He treated Kimberly Bergalis in 1987
- In 1990, Bergalis was diagnosed with AIDS and Acer died from Kaposi's sarcoma
- Bergalis sued for medical malpractice immediately after his death
- Ethical issues arose with disclosure of medical status
- Investigations found that many patients lied, HIV likely acquired from the same community as Acer
- Still unknown for sure whether Acer infected his patients



Case study: influenza – antigenic shift and drift



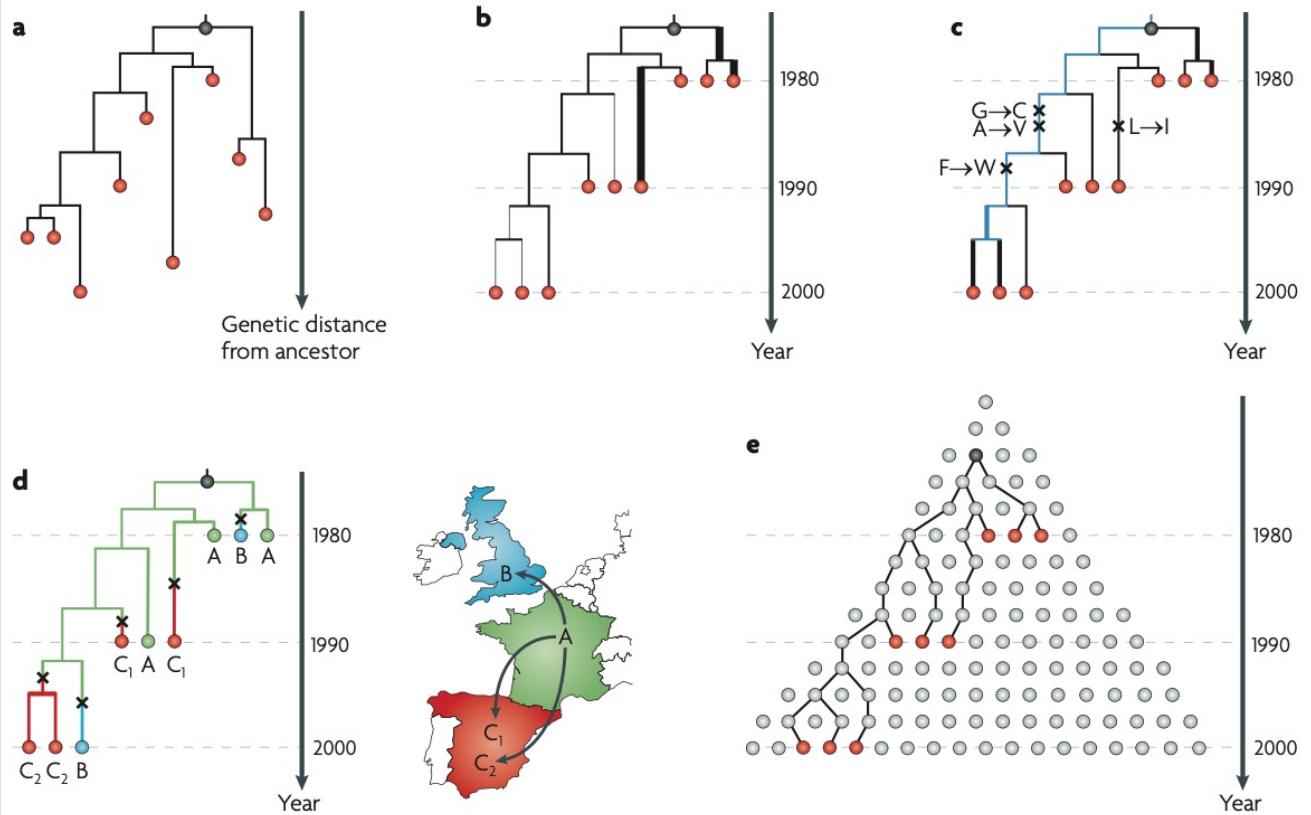
Case study: influenza – antigenic shift and drift



Is drift or shift
happening here?

How does this impact
vaccine design?

Box 1 | Phylodynamic techniques



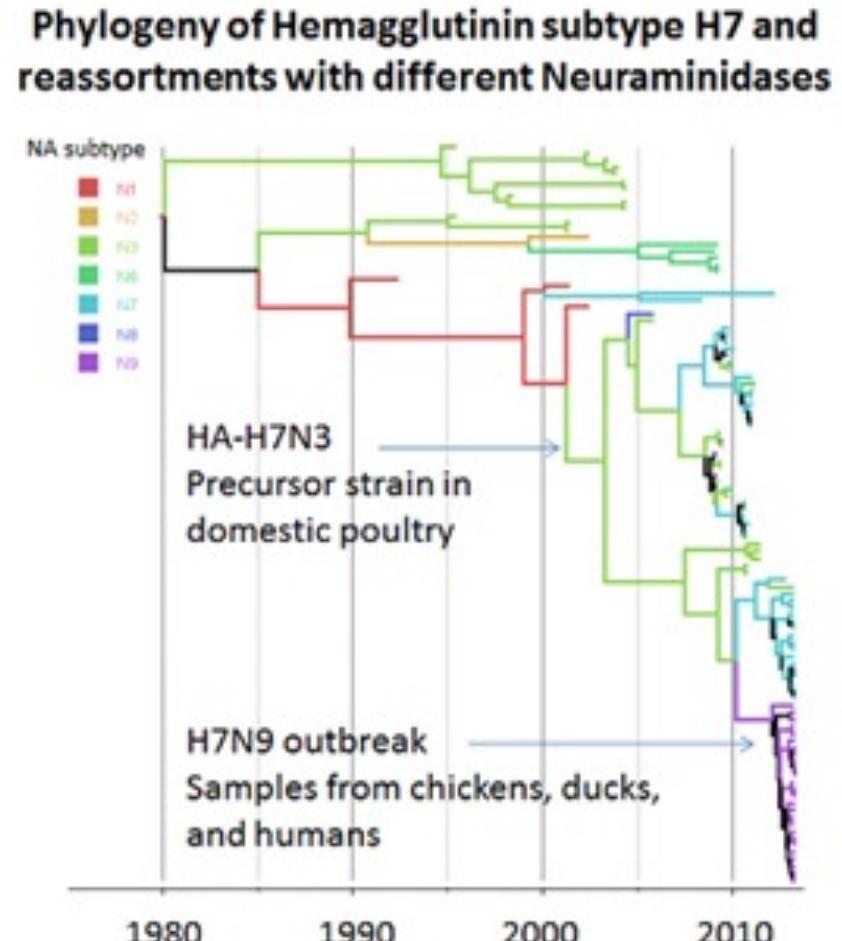
Units matter

Shapes matter

Continual Immune Selection	Weak or Absent Immune Selection	
	Tree shape controlled by non-selective population dynamic processes	
Idealized Phylogeny Shapes	Population size dynamics Exponential growth	Spatial dynamics Strong spatial structure
	Constant size	Weak spatial structure
Examples	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV
Tree Inferences	Detection of antigenic escape mutations	Estimation of population growth rates
		Measles, rabies inter-host HIV
		Estimation of population migration rates

You can do a lot with phylogenies...

- Phylodynamics (tomorrow's lecture)
- Can look at phylogeny in context of other factors
 - Time (how long ago did this virus diverge)
 - Location (how did a virus change as it spread?)
 - Host (how did a virus change in different hosts?)
- Maximum likelihood phylogenies good for:
 - How different/similar is one thing in comparison to known things



BEAST

Bayesian Evolutionary Analysis Sampling Trees

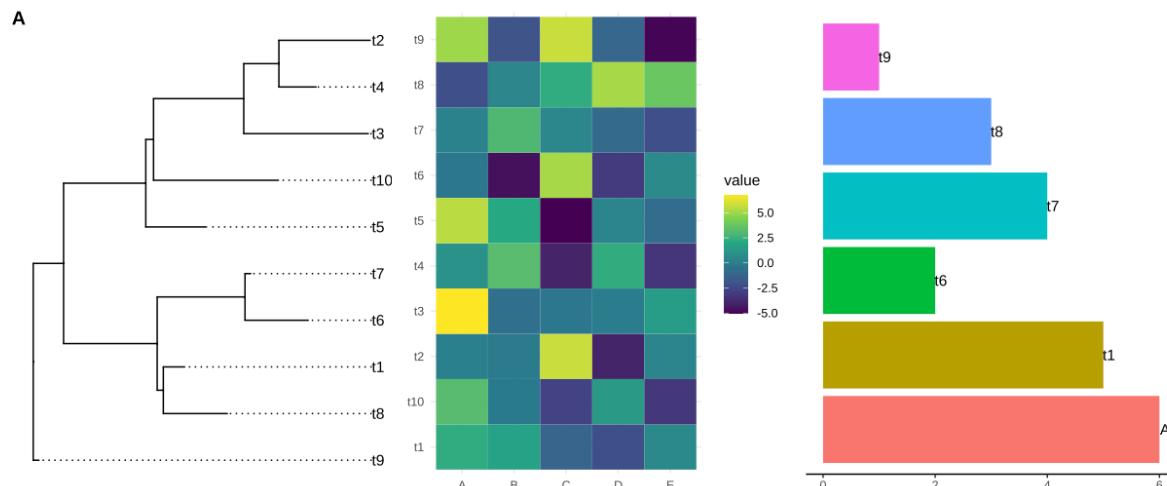
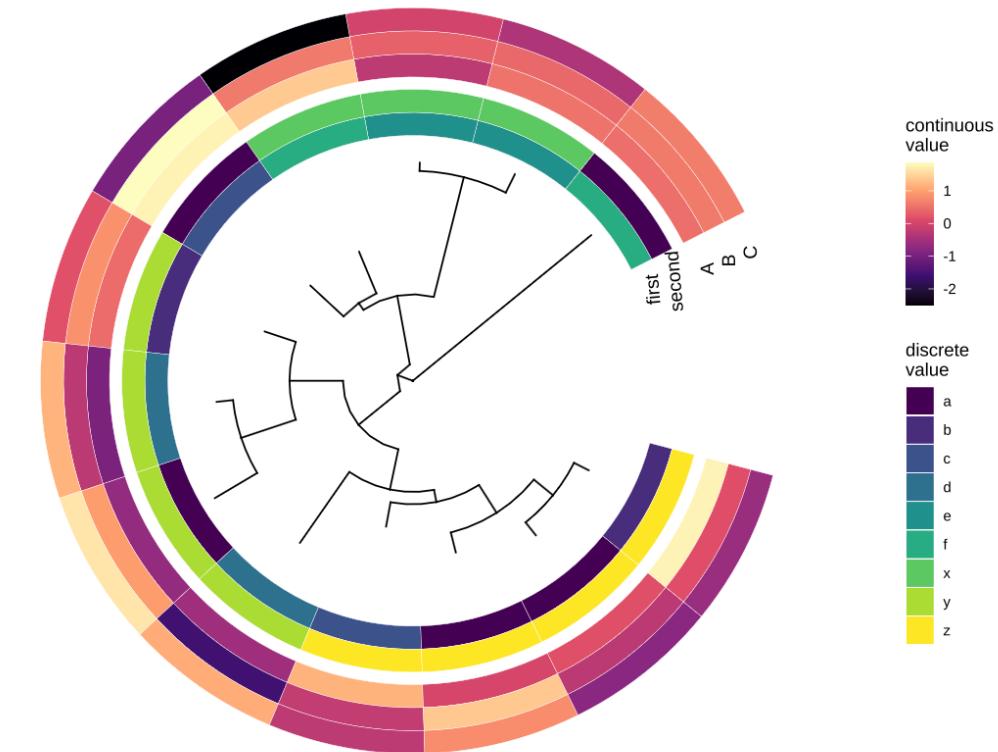
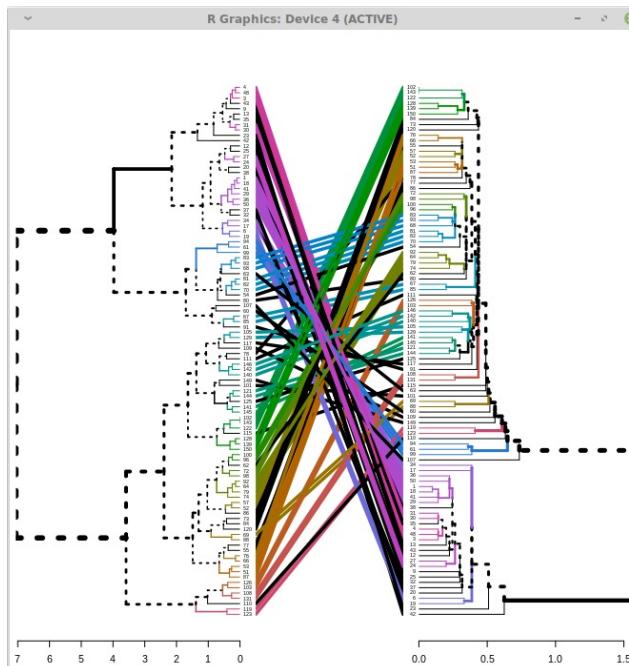
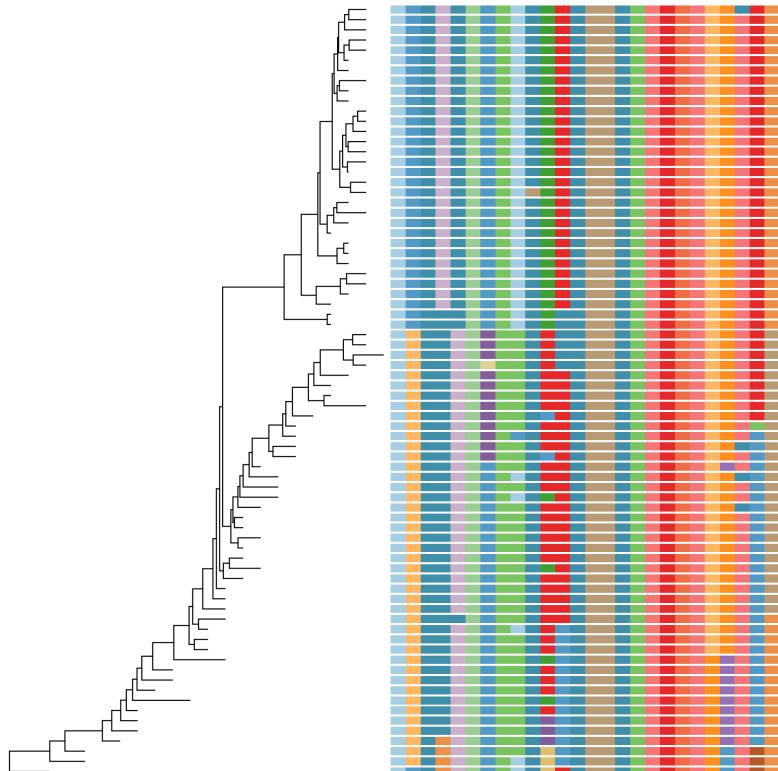
So you have your sequences, now what?

- Get some reference sequences from NCBI
- Get an outgroup from NCBI
- Align them (use a software like MEGA or online like MAFFT)
- Pick the best model (use a software like MEGA or ModelTest-NG)
- Run the phylogeny using your aligned sequences and chosen model (use a software like MEGA or RAxML)
- Visualize/edit tree in either R or FigTree

All of this listed is free to use ☺

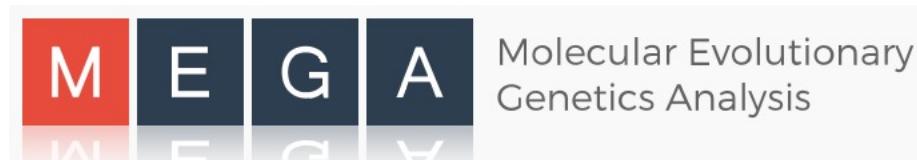
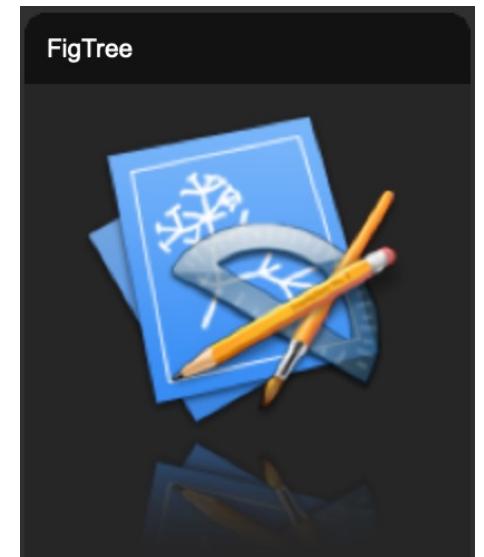
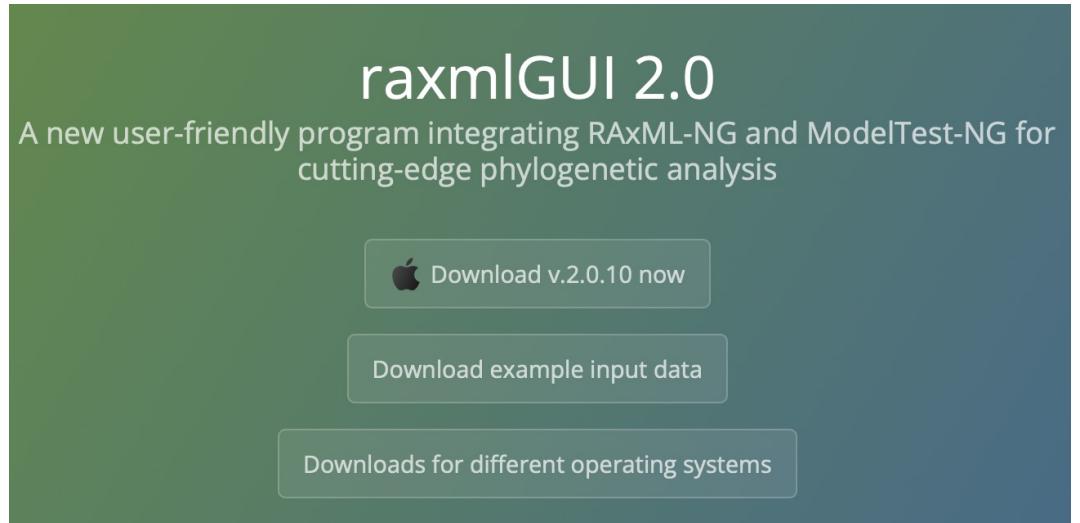
Potential for pretty figures

ggtree: an R package for visualization of tree and annotation data



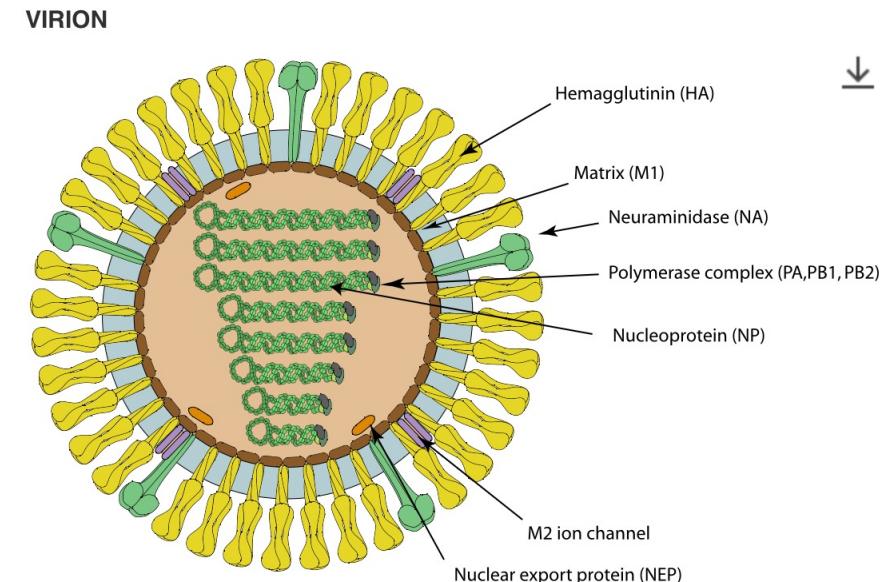
TUTORIAL

- Please make sure RAxML-GUI and MEGA opens on your computer and you have the sequences downloaded from GitHub
- Also have FigTree downloaded, or just follow along by watching
- Additionally, open internet and go to NCBI virus...if this does not open just follow along by watching



Differences between human-hosted and swine-hosted H1N1 influenza

- Hemagglutinin (HA)
 - Cell entry protein for influenza, the “H” in H1N1, 18 different types
- PB2
 - Polymerase protein for influenza, has to do with viral replication, host immune evasion
- Prompt: Let's look at some influenza sequences from France, from both humans and pigs. We'll look at two trees (HA and PB2) for H1N1 influenza
- **What kind of patterns do you think we'll see?**



Steps to revisit later

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

BLAST+ 2.13.0 is here!

Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.

Thu, 17 March 2022 [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

If you have a new sequence and need to figure out what it is...go to BLAST



Log in

BLAST® > blastp suite

Home Recent Results Saved Strategies Help

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

[Reset page](#)

[Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

DTLCIGYHANNSTDVTDTVLEKNVTVTHSVNLLEDKHNGKLCKLRGVAPLHLGK
CNIAGWILGNPECESL
STASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSFERFEIFPKTSSWP
NHDSNKGVTAACPHAG

Or, upload file [Choose File](#) no file selected [?](#)

Job Title

Choose File no file selected [?](#)

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Standard databases (nr etc.): [New](#) Experimental databases

[Try experimental clustered nr database](#)

For more info see [What is clustered nr?](#)

Databases

[Feedback](#)

Compare

Standard

Database

Non-redundant protein sequences (nr) [?](#)

BLASTn: nucleotide sequences

BLASTx: translates your nucleotide sequences to amino acid sequences

BLASTp: protein sequences

BLAST® > blastp suite > results for RID-UD8GFKBS013

Home Recent Results Saved Strategies Help

[Edit Search](#)

[Save Search](#)

[Search Summary](#) [?](#)

[How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title

Protein Sequence

RID

[UD8GFKBS013](#) Search expires on 01-17 22:34 pm [Download All](#) [?](#)

Program

BLASTP [?](#) [Citation](#) [?](#)

Database

nr

[See details](#) [?](#)

Query ID

lcl|Query_52943

Description

unnamed protein product

Molecule type

amino acid

Query Length

530

Other reports

[Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

E value

Query Coverage

[Filter](#) [Reset](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

[Download](#) [Select columns](#) [Show](#) 100 [?](#)

Sequences producing significant alignments

select all 100 sequences selected

[GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	hemagglutinin [Influenza A virus (A/Vienna/INS294/2009(H1N1))]	Influenza A virus (A/Vienna/INS294/2009(H1N1))	1105	1105	100%	0.0	100.00%	566	ADK33981.1
<input checked="" type="checkbox"/>	hemagglutinin [Influenza A virus (A/Canada-ON/RV2965/2009(H1N1))]	Influenza A virus (A/Canada-ON/RV2965/2009(H1N1))	1105	1105	100%	0.0	99.81%	557	ADN24149.1
<input checked="" type="checkbox"/>	hemagglutinin [Influenza A virus (A/Thailand/CU-MV14/2009(H1N1))]	Influenza A virus (A/Thailand/CU-MV14/2009(H1N1))	1105	1105	100%	0.0	99.62%	575	ADK26552.1
<input checked="" type="checkbox"/>	hemagglutinin [Influenza A virus (A/Reunion/2923-3-M1E/2009(H1N1))]	Influenza A virus (A/Reunion/2923-3-M1E/2009(H1N1))	1104	1104	100%	0.0	99.81%	566	AFC18606.1

Okay! Our sequence is influenza A hemagglutinin...so we'll want to build a phylogeny using amino acid influenza A hemagglutinin sequences for reference

Feedback

 Quick Access to SARS-CoV-2 Data!

- Novel Severe acute respiratory syndrome coronavirus 2 RefSeq genomes, nucleotide, and protein sequences.
- View our new [SARS-CoV-2 interactive dashboard](#).
- How to [submit SARS-CoV-2 sequences](#).
- Visit our new SARS-CoV-2 [Variants Overview](#) New!.

NCBI Virus is a community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories. To find, retrieve and analyze data, please select an option below.



Search by sequence

Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.



Search by virus

Use virus name or taxid to find viral nucleotide and protein sequences.



NCBI Visual Data Dashboard

Refine Results Reset

- Virus +
 - Alphainfluenzavirus, taxid:197911 ×
- Accession +
- Sequence Length +
- Ambiguous Characters +
- Sequence Type +
- RefSeq Genome Completeness +
- Nucleotide Completeness +
- Isolate +
- Proteins +
- Provirus +
- Geographic Region +
- Host +

Influenza Virus Data Hub Download

Select genus: Alphainfluenzavirus (A) Gammainfluenzavirus (C)
 Betainfluenzavirus (B) Deltainfluenzavirus (D)

 Advanced Filters for GenBank Sequences

 Visual Filters for GenBank Sequences i

Selected Results: 0 Align Build Phylogenetic Tree

 New! Randomized subsets in Downloads

You now have the option of downloading a smaller, randomized subset of the data shown in the Results table. Begin by using filters to refine your dataset, select the Nucleotide, Protein, or RefSeq Genome tab above the table for the datatype you would like to download, then follow the prompts in the Download menu. Our [Help documentation](#) has more information.

Refine Results Reset

Virus +

Alphainfluenzavirus, taxid:197911 ×

Accession +

Nucleotide (986,562) Protein (1,318,961) RefSeq Genome (7) Select Columns

Expand Table

<input type="checkbox"/> Accession <small>◆</small>	Organism Name <small>New! ◆</small>	Submitters <small>◆</small>	Organization <small>◆</small>
<input type="checkbox"/> NC_026431 <small>RefSeq</small>	Influenza A virus (A/Califo... Garten,R.J., ...	National Center for Biotech...	
<input type="checkbox"/> NC_026432 <small>RefSeq</small>	Influenza A virus (A/Califo... Garten,R.J., ...	National Center for Biotech...	

Feedback

On NCBI virus, we can download some published sequences to make our tree:

1. Click search by virus
2. Type in whatever you're looking for, influenza virus A or alphainfluenzavirus is fine
3. You can select the protein sequences or the nucleotide sequences, refseq refers to "good" genomes that have been determined to be a good representative
4. You can filter by protein, geographic region, host species, and more on the side
5. You can then download all selected sequences, a random subset of sequences, or a csv file with all of the information if you want to manually sort through the results and use R to download sequences using that csv file

Influenza Virus BLAST Influenza Virus Annotation Tool (FLAN) Influenza Virus Articles in PubMed Legacy Influenza Resource	Submit assembled sequences to GenBank Submit sequence reads to SRA
---	---

Download options on NCBI virus

NIH National Library of Medicine
National Center for Biotechnology Information

NCBI Virus
Sequences for discovery

Influenza Virus Data Hub

Download

Advanced Filters for GenBank Sequences

New! Randomized sub

You now have the option of dataset, select the Nucleot prompts in the Download m

Download Results

Step 1 of 3: Select Data Type

Sequence Data (FASTA format)	Accession List	Results Table
<input checked="" type="radio"/> Nucleotide	<input type="radio"/> Nucleotide	<input type="radio"/> CSV format
<input type="radio"/> Coding Region	<input type="radio"/> Protein	<input type="radio"/> Legacy Influenza Resource
<input type="radio"/> Protein	<input type="radio"/> Assembly	

Next

Refine Results Reset

Nucleotide (986,562)	Protein (1,318,961)	RefSeq Genome (7)	Select Columns	Feedback
<input type="checkbox"/> Accession <input type="checkbox"/> Organism Name <input type="checkbox"/> New! <input type="checkbox"/> Submitters <input type="checkbox"/> Organization				

Sample of csv file that you can download instead of the sequences

pb2_seq.fasta

>AHC68917.1 polymerase PB2 [Influenza A virus (A/StEtienne/1691/2009(H1N1))]
MERIKELRDLMSQSRTREILTTVDHMAIIKKYTSQRQEKNPALRMKWMAMRYPITADKRIMDMIPER
NEQGQTLWSKTNAGSDRVLVSPLAFTWNRNGPTTSTVHYPKVYKTYFEKVERLKHGTGPFVFRNQVK
IRRVDTNPNGHADLSAKEAQDVIMEVVFPEVGRILTSESQALITKEKKEELQDCKIAPLMVAYMLERE
LVRKTRFLPVAGGTGSVYIEVHLHTQGTCWEQMYTPGGEVNRDDIDQSLIIARNIVRRAAVSADPLASL
LEMCNSTQIGGVRMVDILRQNPTEEQAVDICKAAIGLRISSSFSFGFTFKRTSGSSVKKEEEVLTGNLQ
TLKIRVHEGYEEFTMVRRAITALRKATRRLIQLIVSGRDEQSAEAIIVAMVFSQEDCMIKAVRGDLNF
VNANQRLNPMHQLLRFQDAKVLFQNWGIESIDNVGMGILPDMPTSTEMLRGIRVSKMGVDEYSS
TERVVVSIDRFLRVRDQRGNILLSPPEEVSETQGTEKLITYSSSMWEINGPESVLVNTYQWIIRNWEIV
KIQWSQDPTMLYNKMEFEPFQSLVPKAIRSRYSGFVRTLFQQMRDVLGTFDTVQIICKLLPFAAAPPEQSR
MQFSSLTVNRGSGRLILVRGNSPVFNYNKATKRLTVLGKDAGALTEDPDEGTSGVESAVLRGFLILGKE
DKRYGPALSINELSNLAKGEKANVLIGQGDVVLVMKRKDSSI LTDQSTATKIRMAIN

>AHC68936.1 polymerase PB2 [Influenza A virus (A/Lyon/52.16/2010(H1N1))]
MERIKELRDLMSQSRTREILTTVDHMAIIKKYTSQRQEKNPALRMKWMAMRYPITADKRIMDMIPER
NEQGQTLWSKTNAGSDRVLVSPLAFTWNRNGPTTSTVHYPKVYKTYFEKVERLKHGTGPFVFRNQVK
IRRVDTNPNGHADLSAKEAQDVIMEVVFPEVGRILTSESQALITKEKKEELQDCKIAPLMVAYMLERE

>AHC68969.1 polymerase PB2 [Influenza A virus (A/Lyon/1.12/2011(H1N1))]
MERIKELRDLMSQSRTREILTTVDHMAIIKKYTSQRQEKNPALRMKWMAMRYPITADKRIMDMIPER
NEQGQTLWSKTNAGSDRVMVSPLAFTWNRNGPTTSTVHYPKVYKTYFEKVERLKHGTGPFVFRNQVK
IRRVDTNPNGHADLSAKEAQDVIMEVVFPEVGRILTSESQALITKEKKEELQDCKIAPLMVAYMLERE

FASTA file of all your selected sequences

	A	B	C	D	E	F	G
1	Accession	Organism_N	Genus	Family	Country	Host	Collection_Date
2	AHC68917.1	Influenza A v Alphainfluen	Orthomyxovi	France	Homo sapier	2009	
3	AHC68936.1	Influenza A v Alphainfluen	Orthomyxovi	France	Homo sapier	2010	
4	AHC68969.1	Influenza A v Alphainfluen	Orthomyxovi	France	Homo sapier	2011	
5	AKJ80495.1	Influenza A v Alphainfluen	Orthomyxovi	France	Sus scrofa	2010	
6	AKJ80506.1	Influenza A v Alphainfluen	Orthomyxovi	France	Sus scrofa	2011	
7	AKJ83080.1	Influenza A v Alphainfluen	Orthomyxovi	France	Sus scrofa	2009	
8	AID48619.1	Influenza A v Alphainfluen	Orthomyxovi	France	Sus scrofa	1980	
9	AFG99544.1	Influenza A v Alphainfluen	Orthomyxovi	France	Homo sapier	1991	
10	AFG99964.1	Influenza A v Alphainfluen	Orthomyxovi	France	Homo sapier	1992	
11	APT36494.1	Influenza A v Alphainfluen	Orthomyxovi	France	Sus scrofa	2015	
12	AGC13478.1	Influenza A v Alphainfluen	Orthomyxovi	France	Sus scrofa	2012	
13	NC_006505	Isavirus salar	Isavirus	Orthomyxovi	Norway	Salmo salar	2012
14	WRK84861	Influenza B v Betainfluen	Orthomyxovi	USA	Homo sapier	2023	

We need to pick an outgroup to root the tree...otherwise we can not determine how many changes in the genome led to the separation in the tree, an unrooted tree can not tell us which sequences descended from what

1. You can go to Wikipedia for Influenza A virus and look at virus classification, you can also go to other sites like ICTV which are dedicated to taxonomy of viruses
2. In a general rule...you want something different enough from your virus of interest but not too different...so go one level "up" to Orthomyxoviridae and pick another non-flu genus
3. I picked Isavirus, pick a sequence from there that is from the same region of your gene of interest
 1. If you are making a PB2 tree...download a OB2 sequence from an Isavirus
 2. Download the sequence and add to master FASTA file of your reference sequences

Family: ***Orthomyxoviridae***

Genera

- *Alphainfluenzavirus*
- *Betainfluenzavirus*
- *Gammainfluenzavirus*
- *Deltainfluenzavirus*
- *Isavirus*
- *Quaranjavirus*
- *Thogotovirus*

Virus classification



(unranked): **Virus**

Realm: ***Riboviria***

Kingdom: ***Orthornavirae***

Phylum: ***Negarnaviricota***

Class: ***Insthoviricetes***

Order: ***Articulavirales***

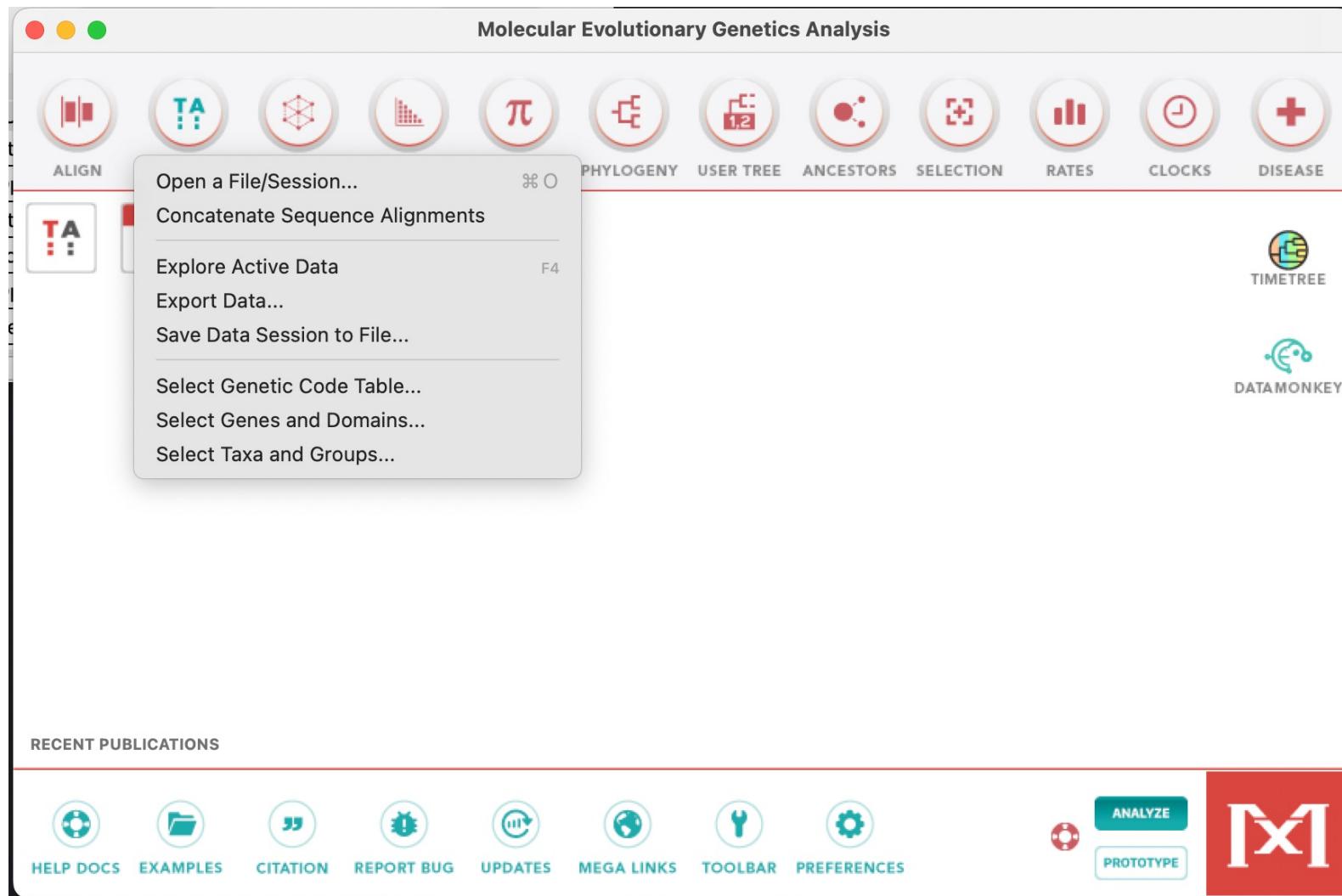
Family: ***Orthomyxoviridae***

Genus: ***Alphainfluenzavirus***

Species: ***Influenza A virus***

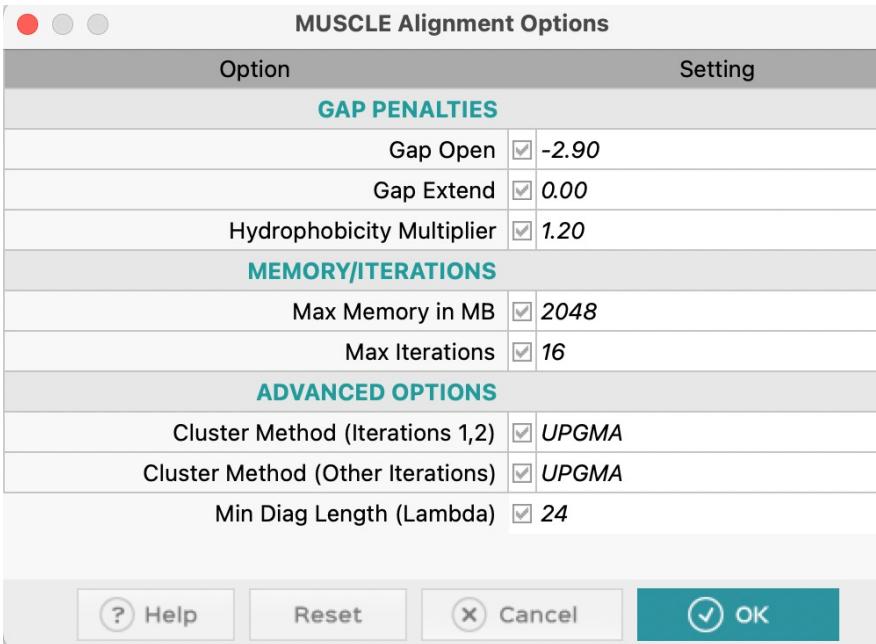
Subtypes

See text



We need to align our sequences, so the analysis is comparing sequences in the same region of the genome

1. Open MEGA, and open a file/session, select your concatenated fasta file
2. MEGA will ask if you want to align or analyze, click on align



So the sequences are loaded into MEGA like this (below):

1. Click on the muscle arm to align with MUSCLE program
2. Leave all the preset options the same, click okay and click yes if it asks to select all sequences

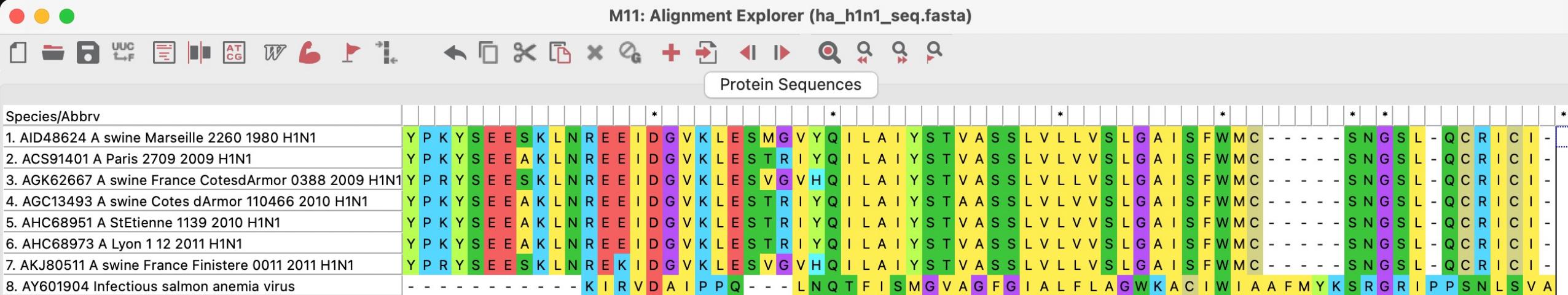
M11: Alignment Explorer (ha_h1n1_seq.fasta)

Align Protein Protein Sequences

Species/Abbrv

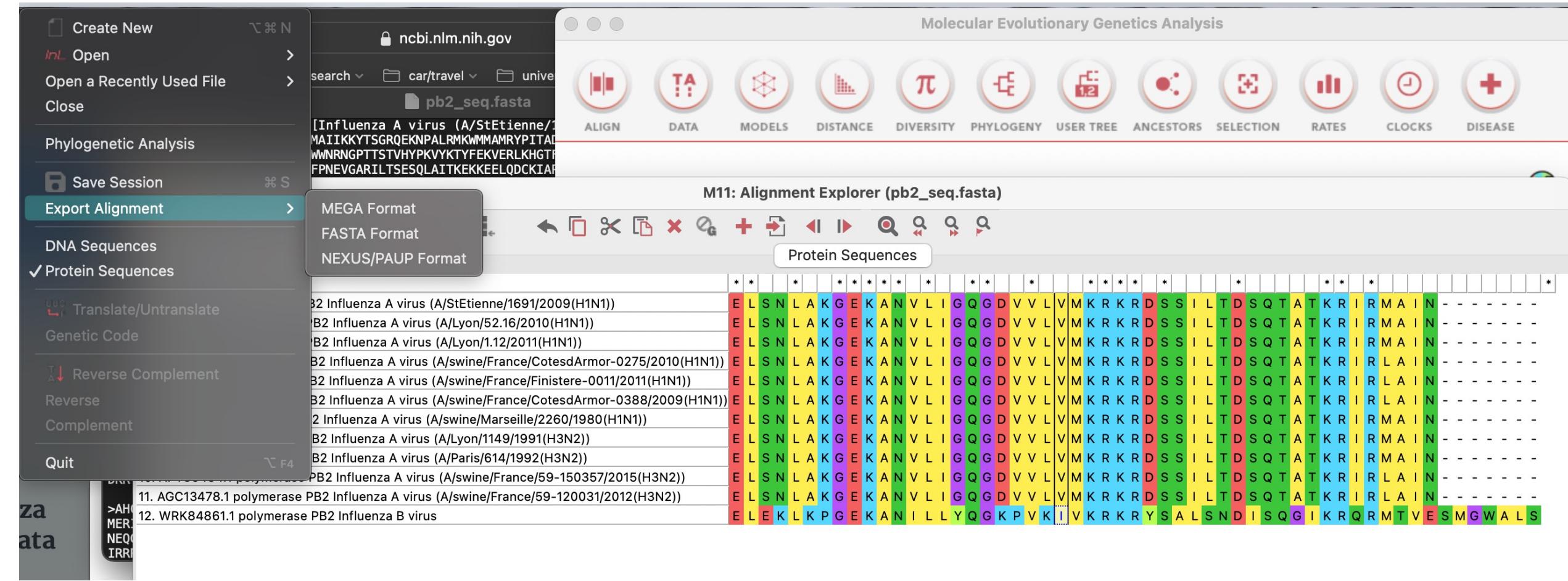
- 1. AID48624 A swine Marseille 2260 1980 H1N1
- 2. ACS91401 A Paris 2709 2009 H1N1
- 3. AGK62667 A swine France CotesdArmor 0388 2009 H1N1
- 4. AGC13493 A swine Cotes dArmor 110466 2010 H1N1
- 5. AHC68951 A StEtienne 1139 2010 H1N1
- 6. AHC68973 A Lyon 1 12 2011 H1N1
- 7. AKJ80511 A swine France Finistere 0011 2011 H1N1
- 8. AY601904 Infectious salmon anemia virus

The alignment grid shows the amino acid sequence alignment of the eight selected H1N1 influenza A virus hemagglutinin (HA) genes. The sequences are color-coded by residue type: M (blue), E (red), K (yellow), A (cyan), L (light blue), F (green), V (light green), I (purple), L (light purple), Y (pink), T (light pink), C (light cyan), P (light blue), S (light green), R (light pink), N (light purple), D (light blue), G (light green), H (light pink), N (light purple), S (light green), T (light blue), D (light blue), and S (light green). The alignment shows high conservation of the HA protein across the different strains.



So this is the aligned data file, at this point you can trim ends if necessary to prepare for making a tree. You would want to do that if you have one sequence that "hangs" off past the others

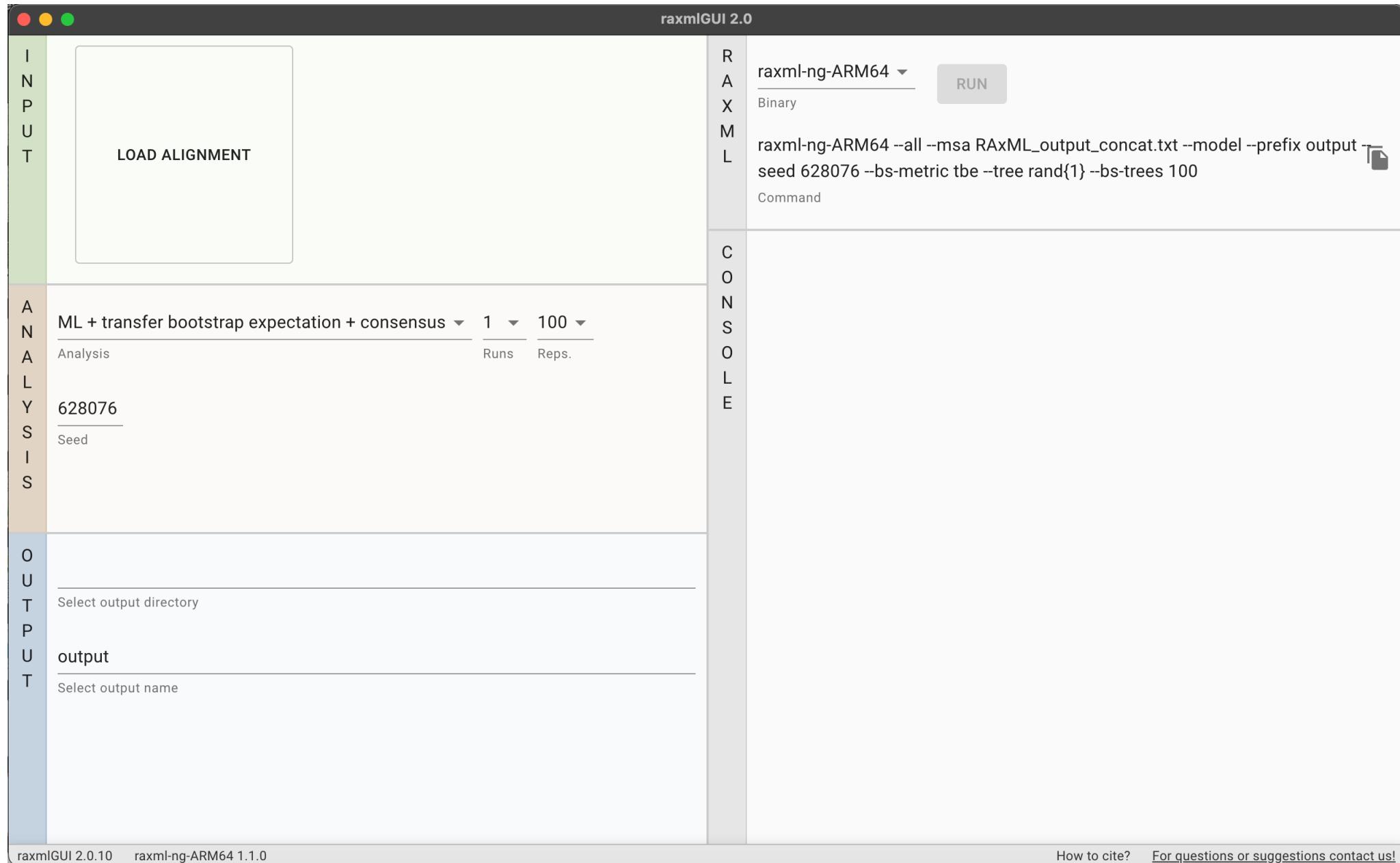
However you can also NOT do the trimming, depending on how much overhang there is it isn't necessary...here there is not much overhang so I would not trim



Save the aligned file in FASTA format, then we will proceed to model selection

You can save things in MEGA format, but they tend to only run well in MEGA and are not good using across different softwares

Open up RAxML GUI



INPUT

protein pb2_h1n1_seq_mega.fas
8 sequences of length 807

Blosum62 ▾ none ▾ none ▾

Substitution model Stationary frequencies Proportion of invariant sites

none ▾

Rate heterogeneity **RUN MODELTEST**

Partition 1/1:

	Model	Score	Weight
BIC	LG+I	9153.0821	0.5393
AIC	LG+I+F	9048.0610	0.4389
AICc	LG+I+F	9050.0610	0.4836

INPUT

protein ha_h1n1_seq_align_mega.fas
8 sequences of length 574

Blosum62 ▾ none ▾ none ▾

Substitution model Stationary frequencies Proportion of invariant sites

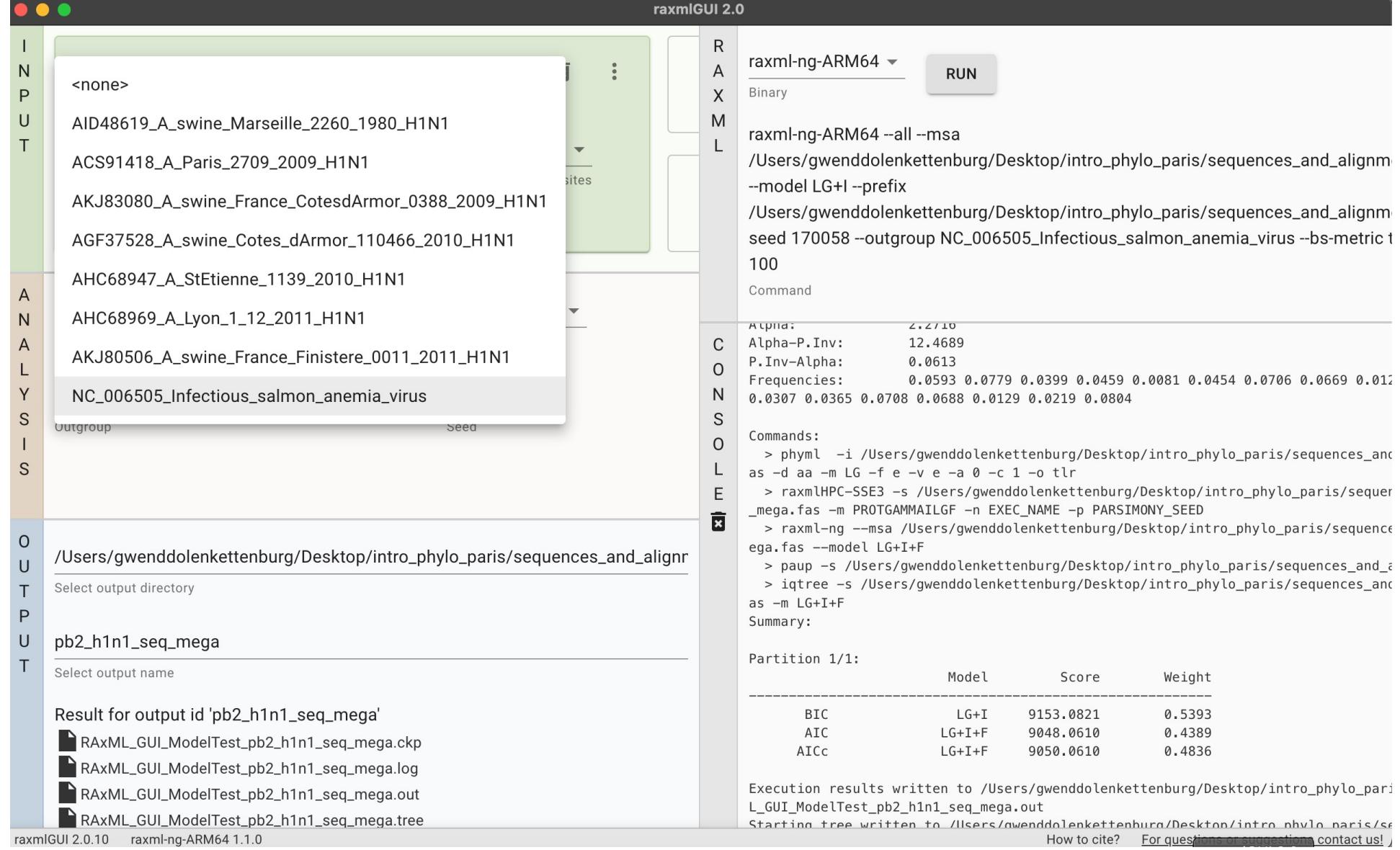
none ▾

Rate heterogeneity **RUN MODELTEST**

Partition 1/1:

	Model	Score	Weight
BIC	FLU+G4	7259.4956	0.6526
AIC	FLU+G4	7198.5588	0.8274
AICc	FLU+G4	7198.5588	0.8274

Load the aligned file and perform modeltest, for model selection go with the BIC score, it will spit out a report that saves to your files



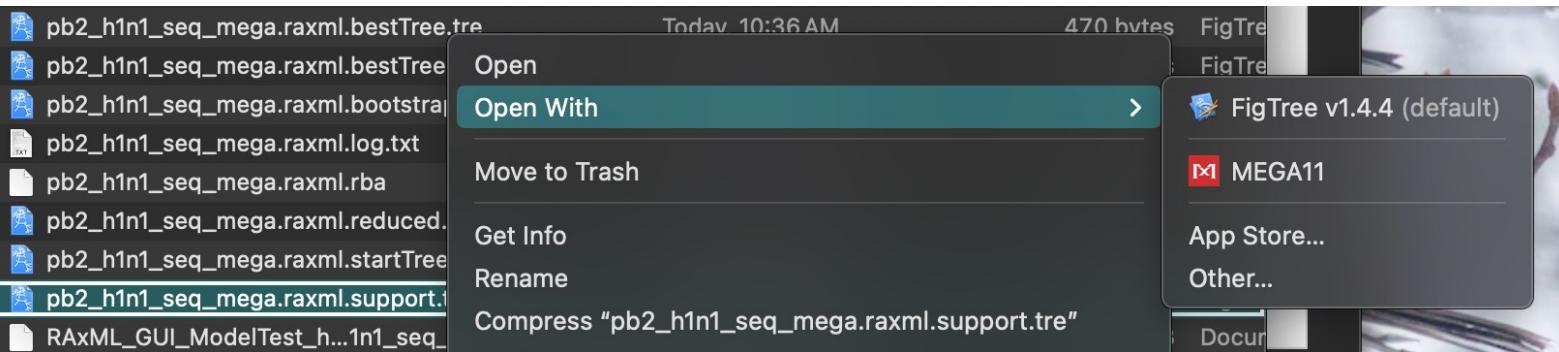
Using the model from
modeltest, change as
needed in the input box,
then in analysis set the
outgroup and then hit
run in the RAxML section

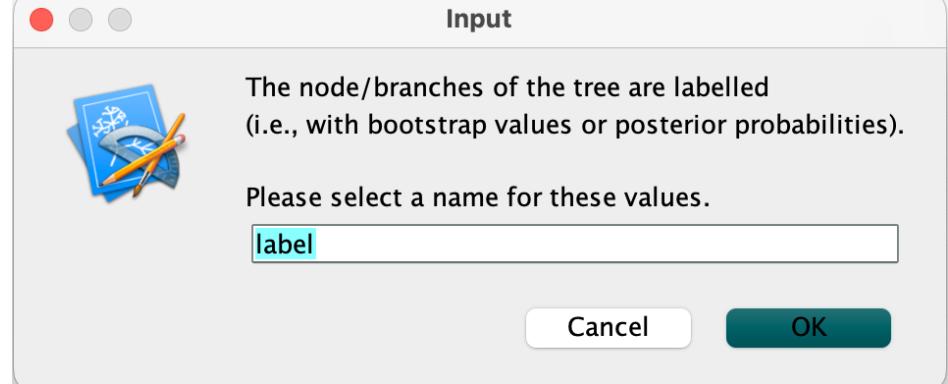
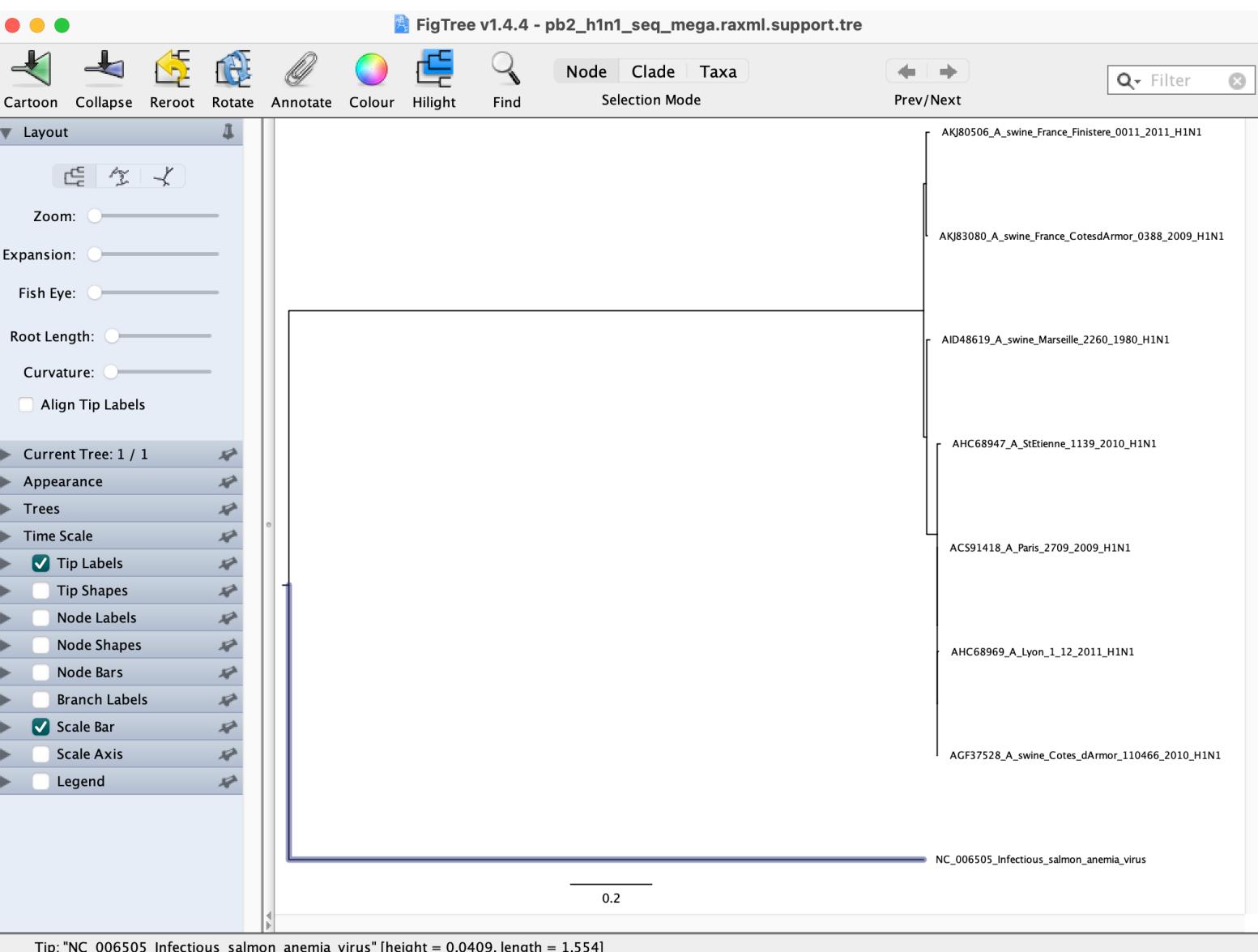
It will also spit out some files...

RAxML_output		Today, 10:40 AM	--	Folder
ha_h1n1_france_human_swine.csv		Today, 10:22 AM	1 KB	comma-separated values
ha_h1n1_seq_align_mega.raxml.bestModel.txt		Today, 10:40 AM	34 bytes	Plain Text
ha_h1n1_seq_align_mega.raxml.bestTree.tre		Today, 10:40 AM	469 bytes	FigTree
ha_h1n1_seq_align_mega.raxml.bootstraps.tre		Today, 10:40 AM	47 KB	FigTree
ha_h1n1_seq_align_mega.raxml.log.txt		Today, 10:40 AM	13 KB	Plain Text
ha_h1n1_seq_align_mega.raxml.rba		Today, 10:38 AM	5 KB	Document
ha_h1n1_seq_align_mega.raxml.startTree.tre		Today, 10:38 AM	458 bytes	FigTree
ha_h1n1_seq_align_mega.raxml.support.tre		Today, 10:40 AM	509 bytes	FigTree

We're interested in the .support.tre file

This will include the tree topology and bootstrap support values, open in Figtree





When you load..it will ask you select name for values...just leave as is and click okay

We want a file of the tree saved that can be read in R, so newick. Click export trees...choose format newick, and customize in R

You can customize in FigTree too...it's just more limited

Then make pretty in R!

- Follow instructions in `influenza_tree_editing.R` file