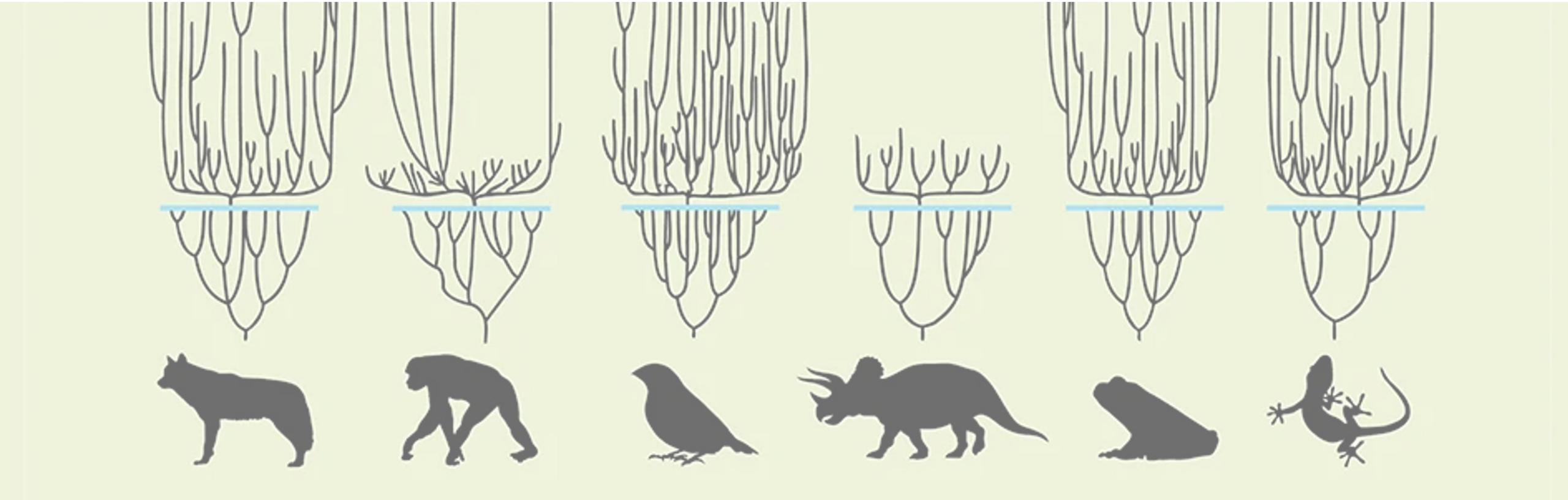
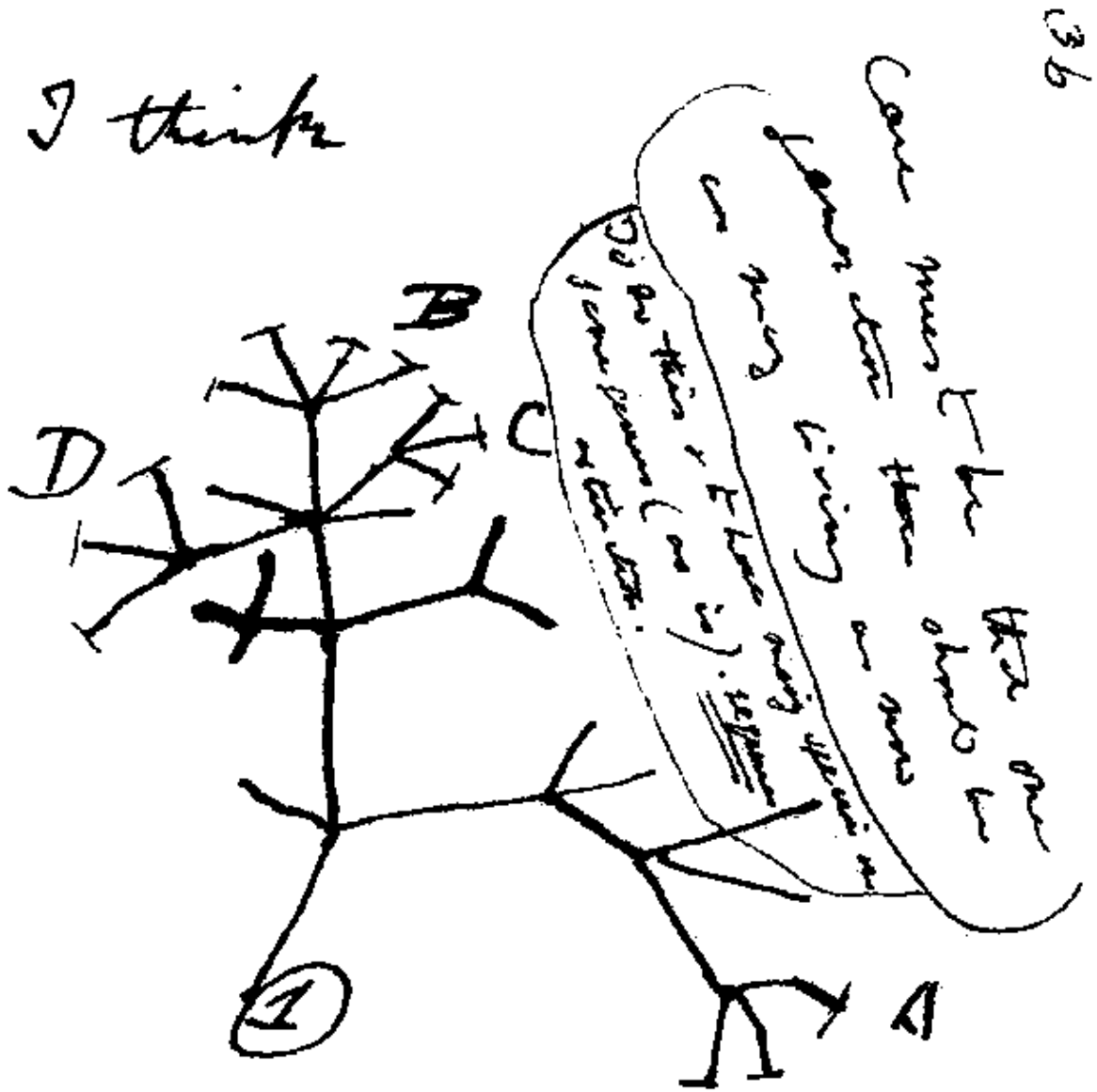
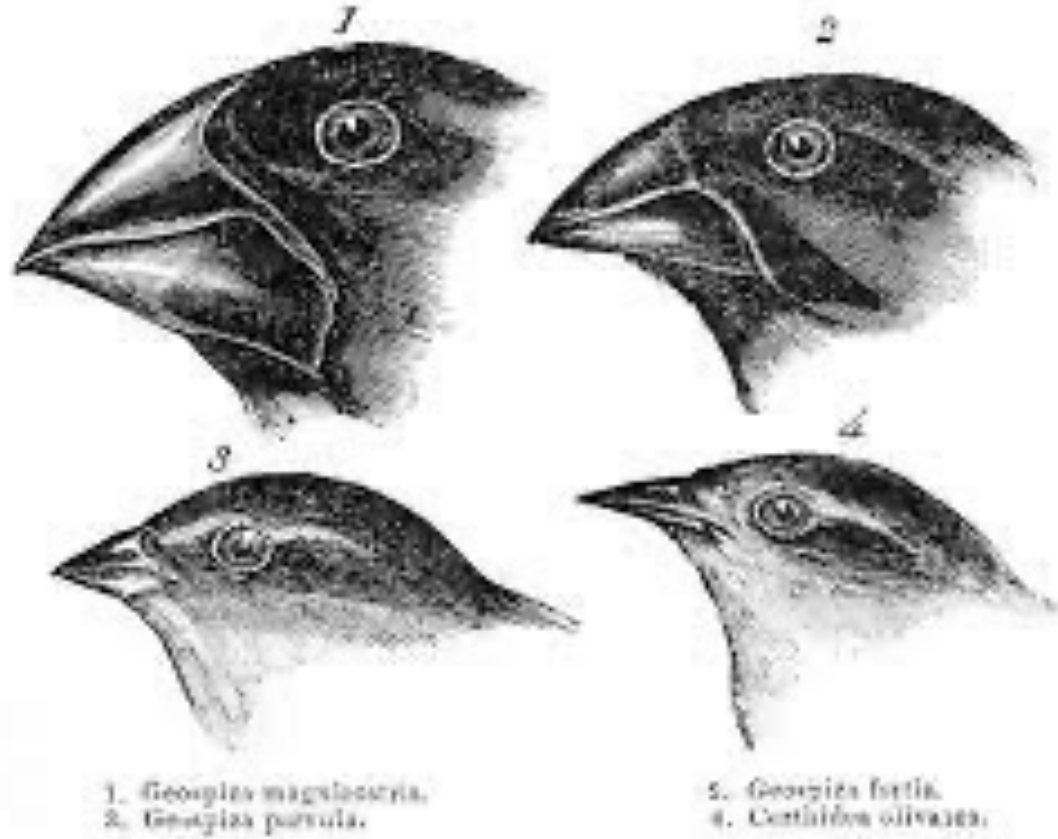
# Introduction to phylogenetics

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

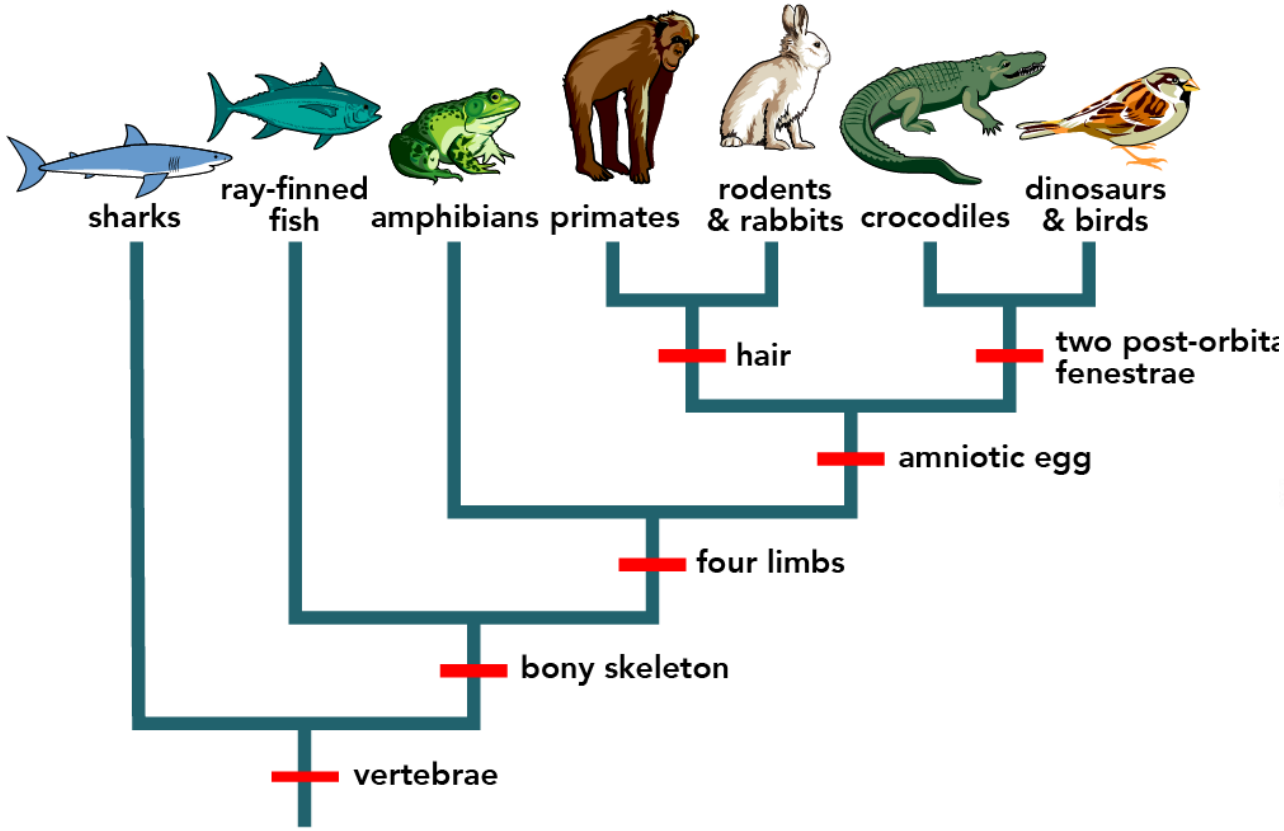


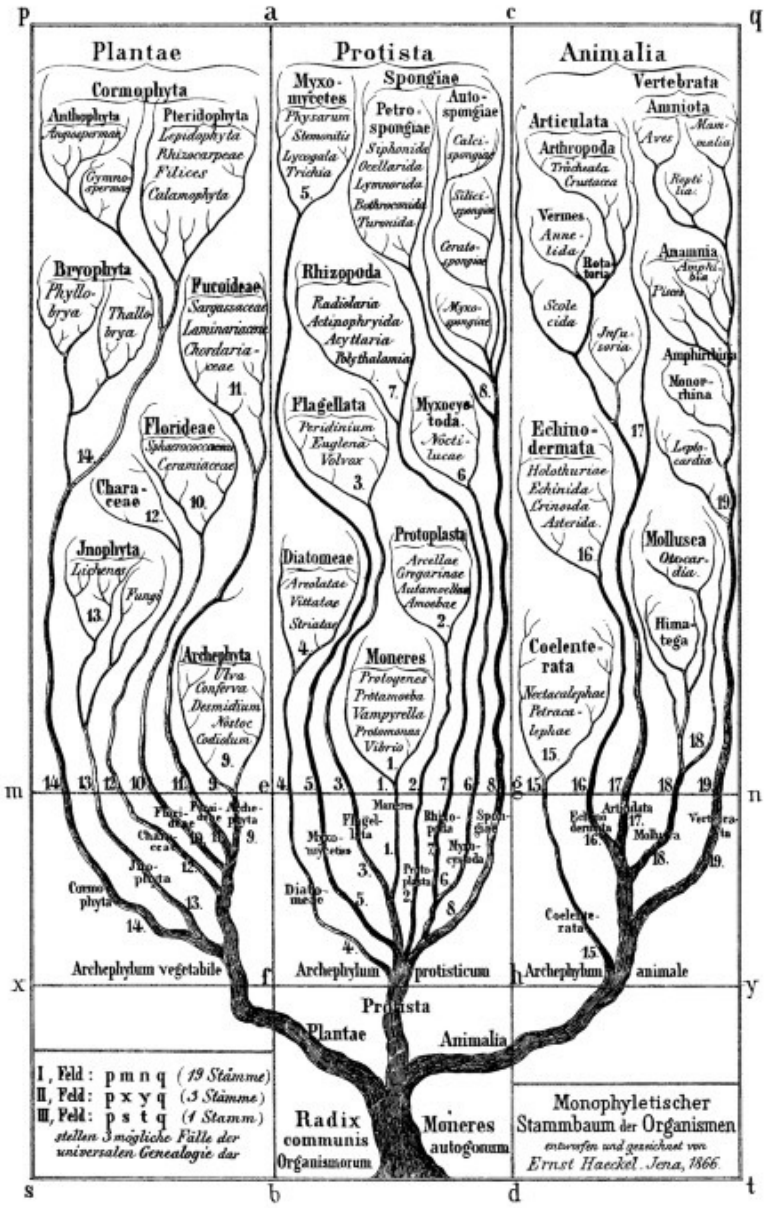
# What is a phylogeny?



# What is a phylogeny?

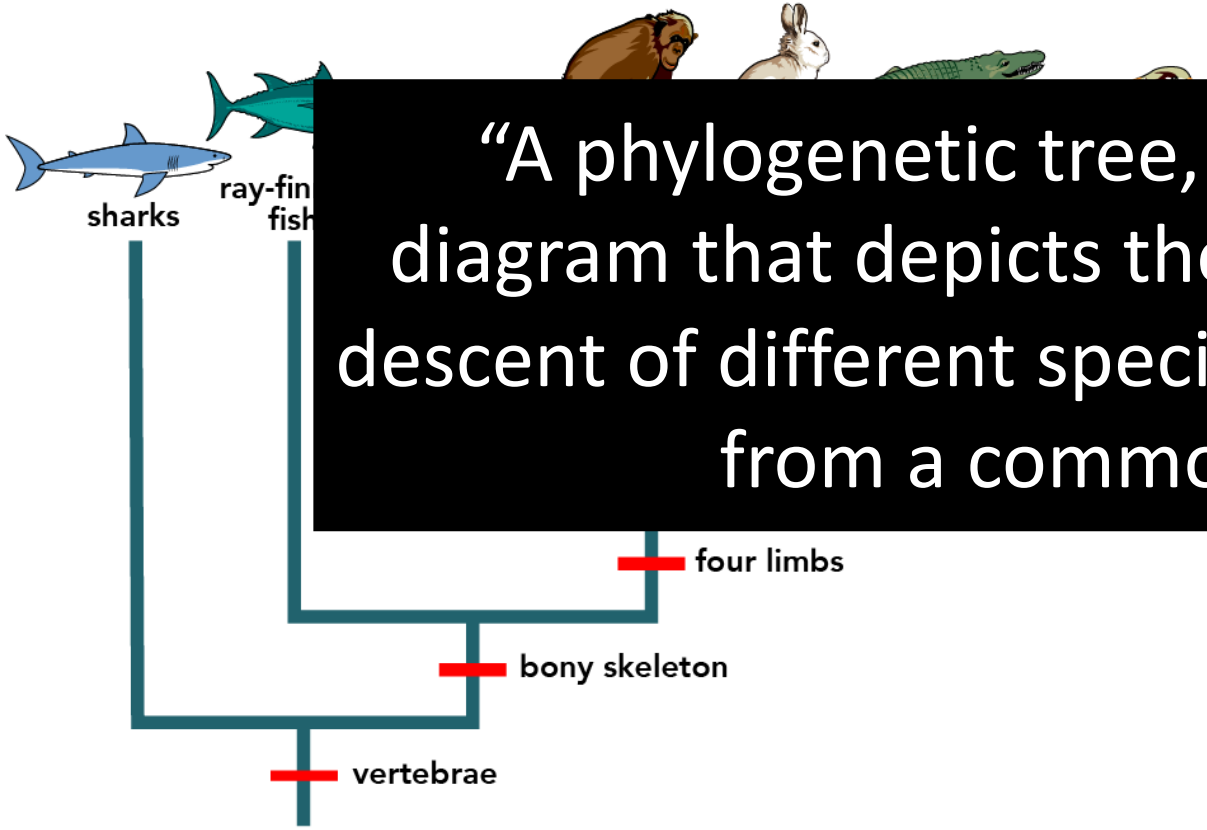


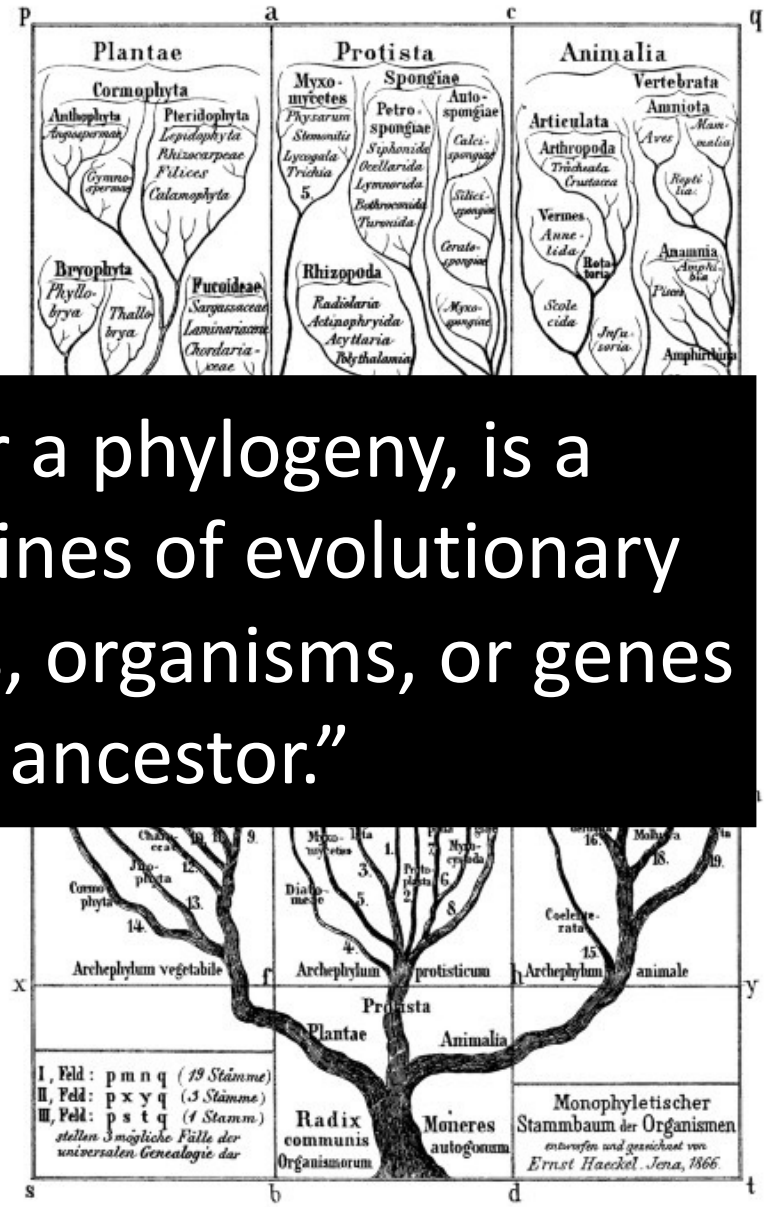

Baum et. al, Nature, 2008

Hossfeld and Levit, Nature, 2016

# What is a phylogeny?



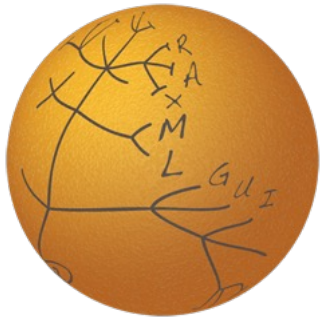

"A phylogenetic tree, or a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor."




Baum et. al, Nature, 2008

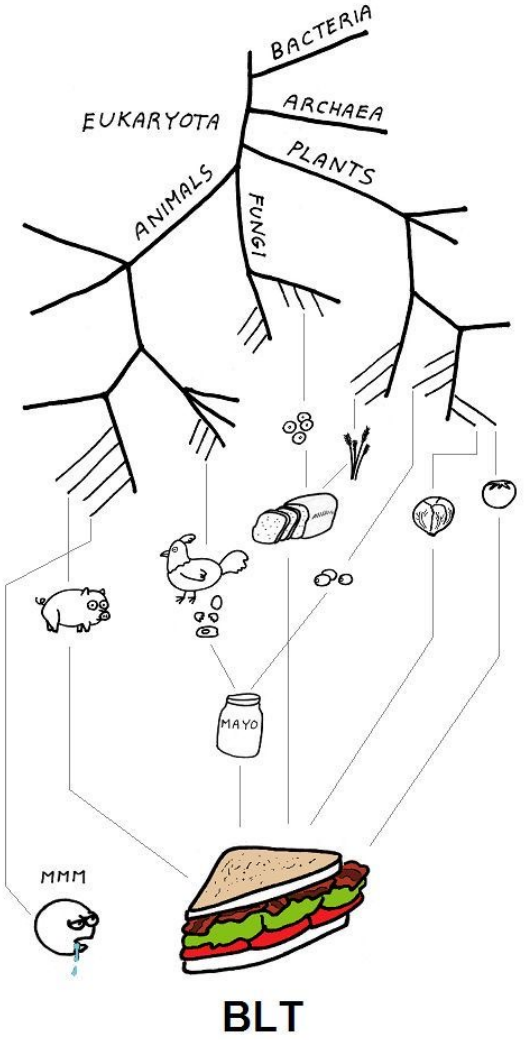
Hossfeld and Levit, Nature, 2016

# 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

RStudio®

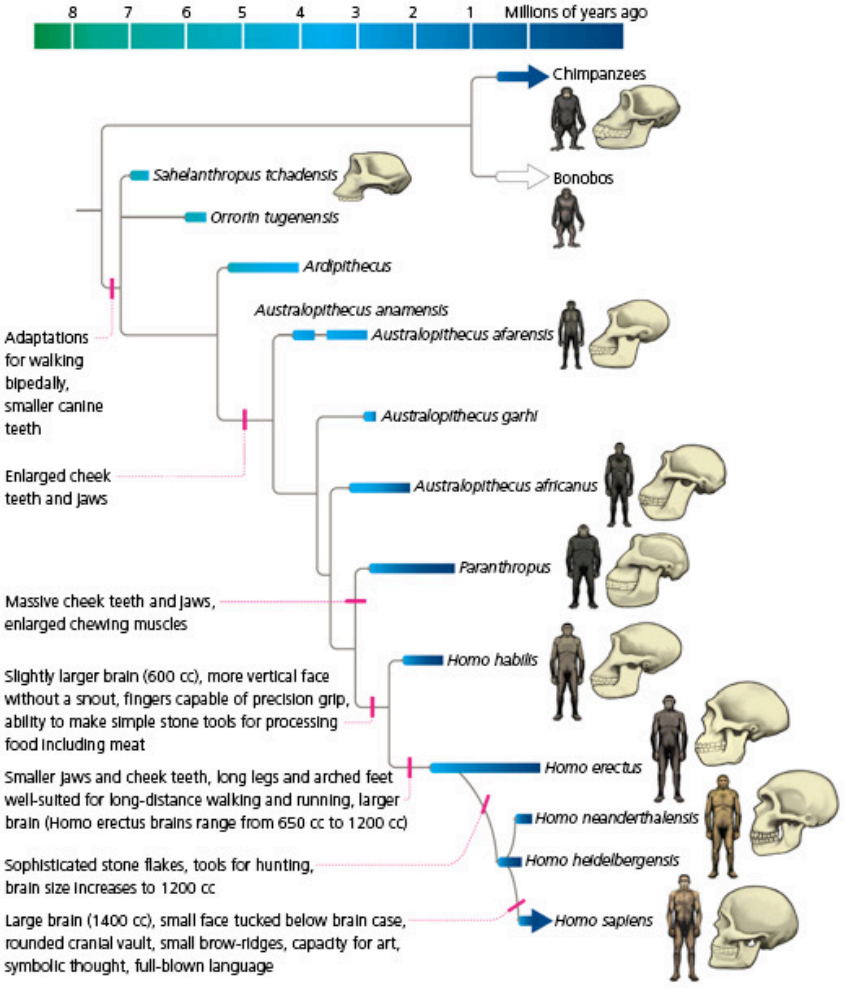



BLT

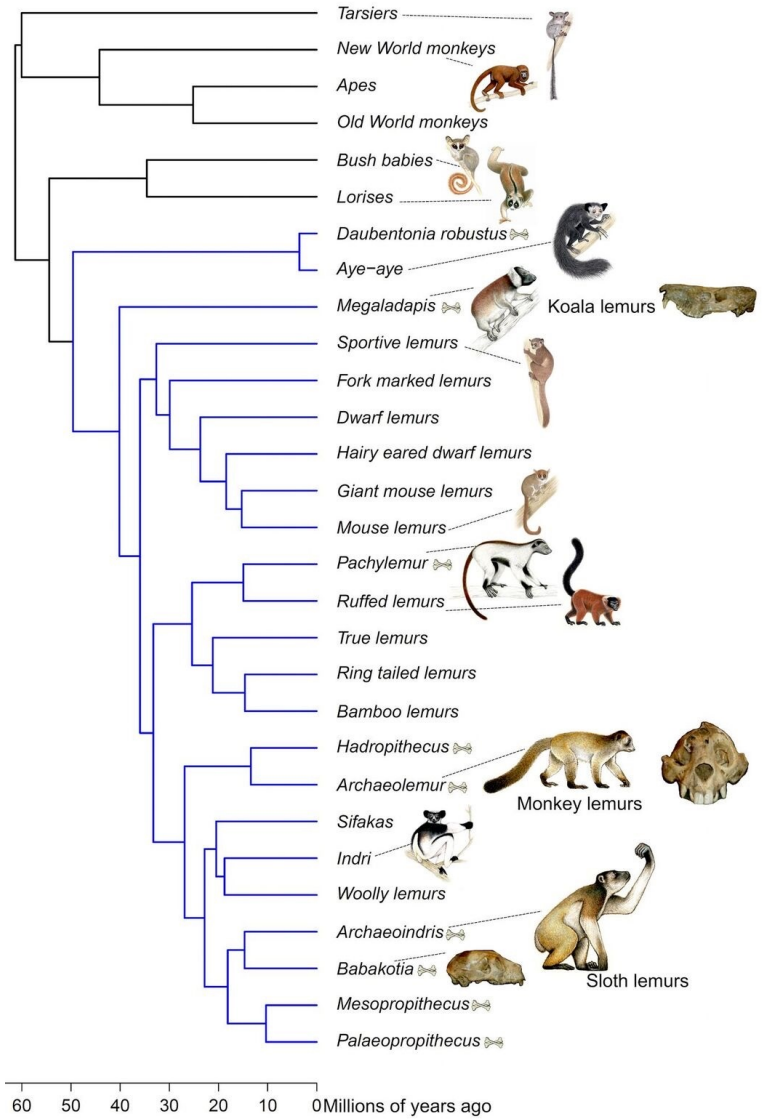
Nucleotide BLAST
nucleotide ▶ nucleotide

MEGA Molecular Evolutionary Genetics Analysis

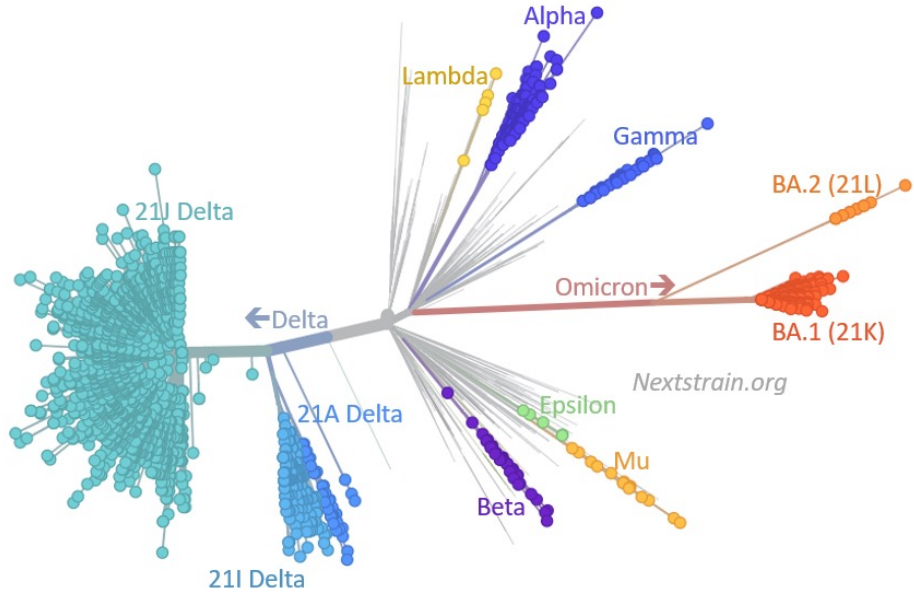
# What can you do with phylogenies?




UC Museum of Paleontology

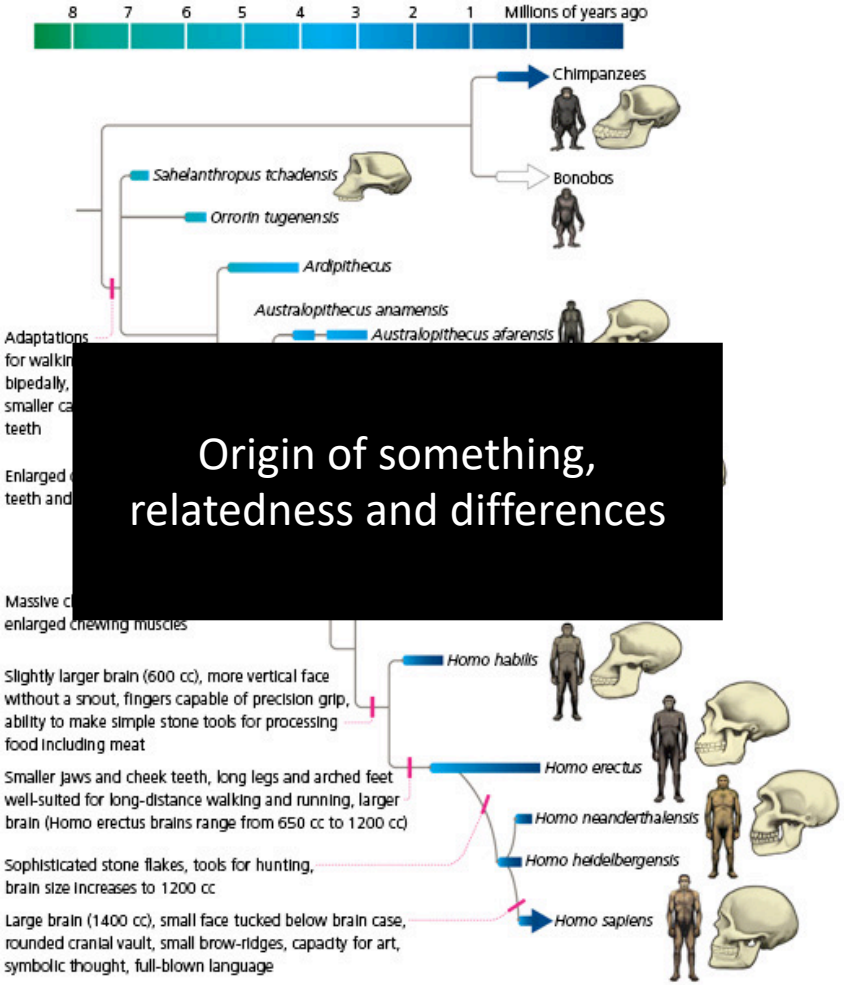



Herrera et al 2016

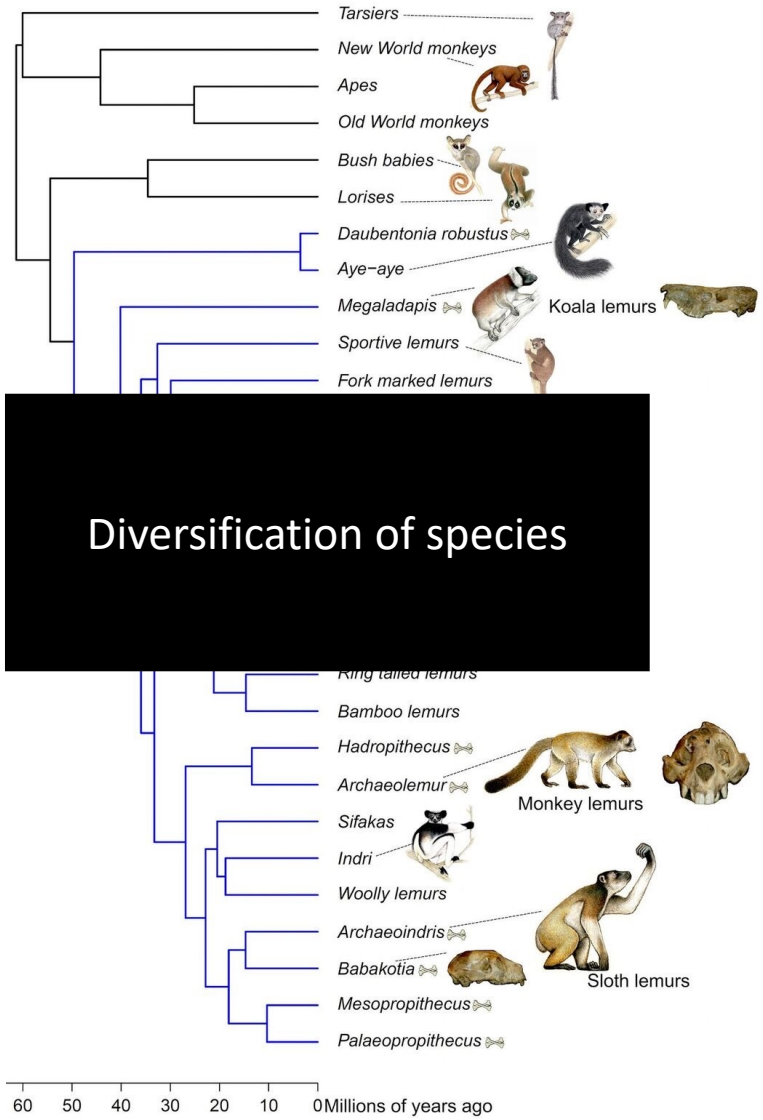



Lowe 2022

# What can you do with phylogenies?




Origin of something, relatedness and differences




Diversification of species

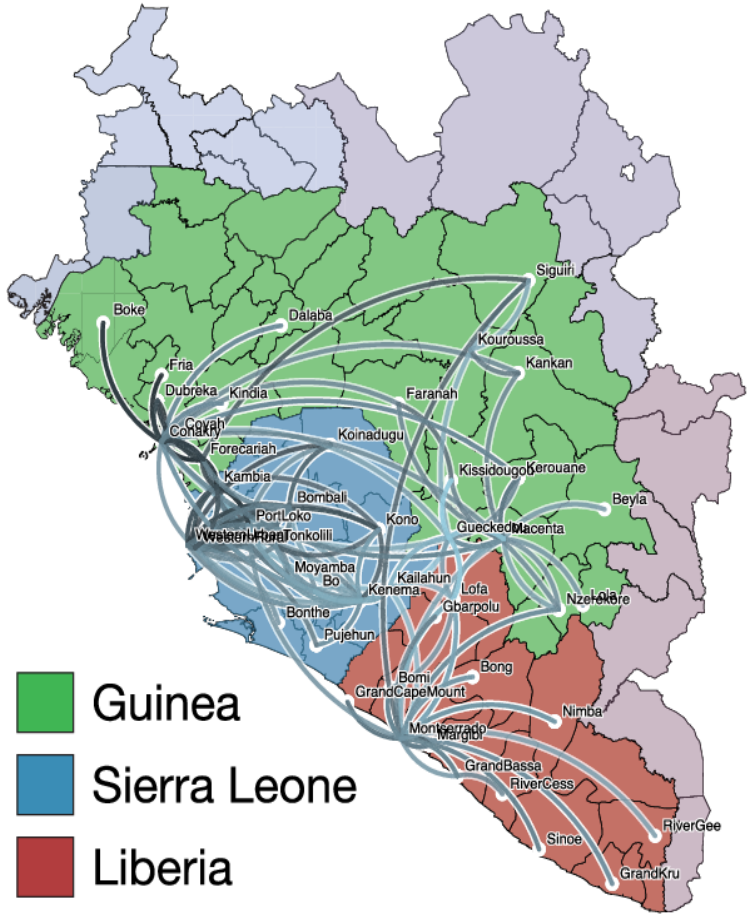

Changes in real time

UC Museum of Paleontology

Herrera et al 2016

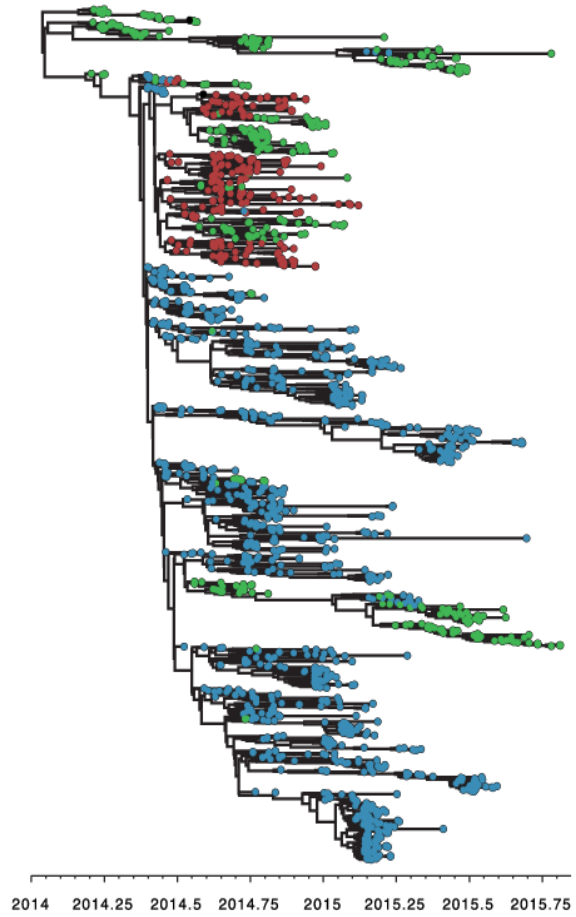
Lowe 2022

## Phylodynamics



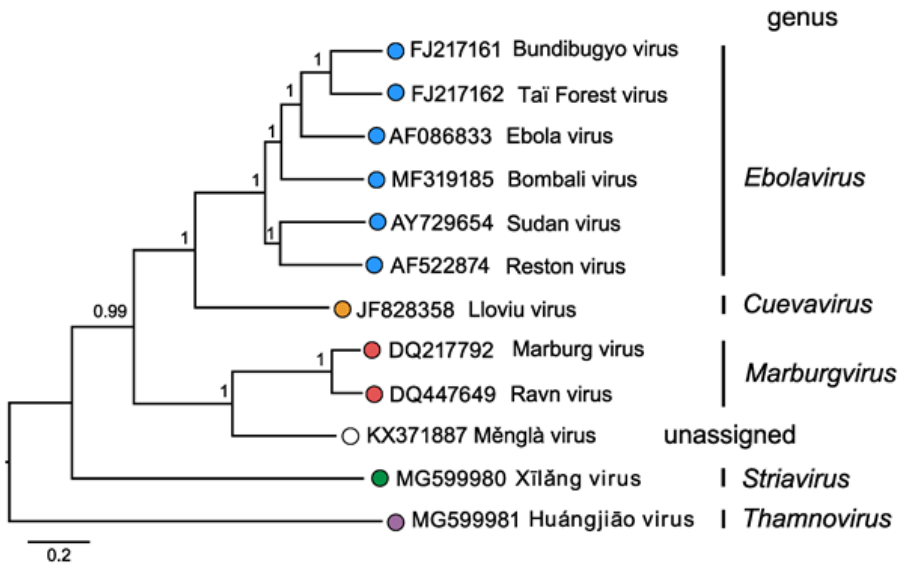

WHERE does it go?

## Bayesian trees




WHEN is the most recent common ancestor?

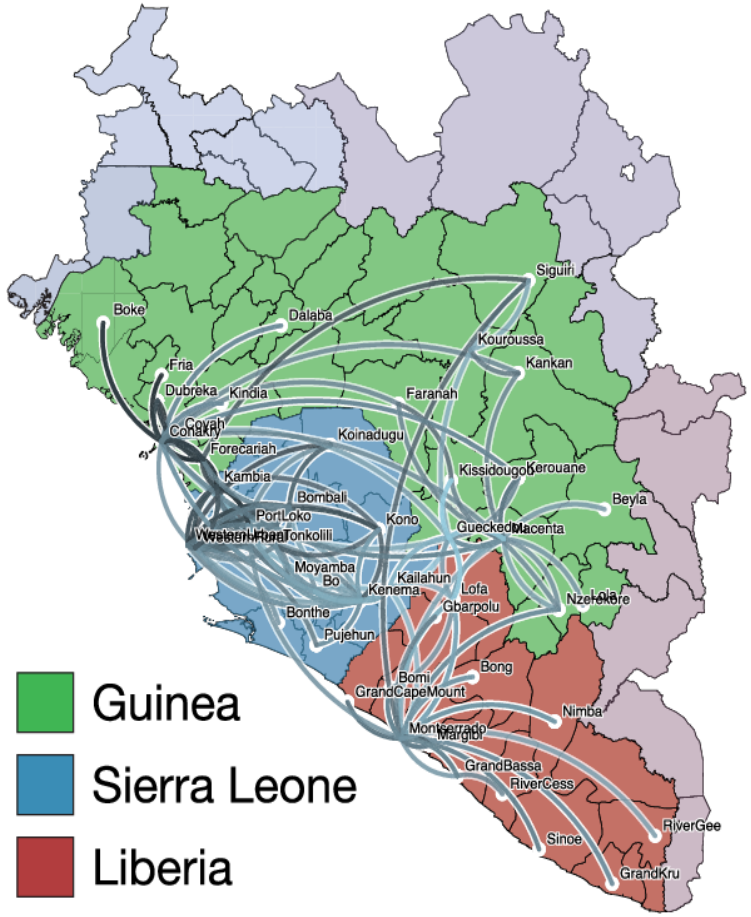
## Maximum likelihood




HOW different is it to what's known?

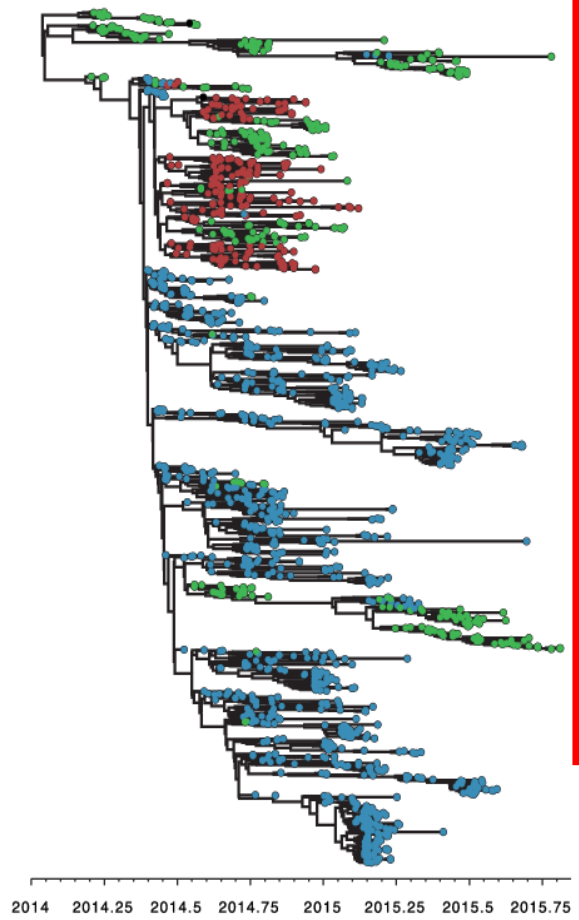
Suchard et al. 2018
Kuhn et al. 2019

## Phylodynamics



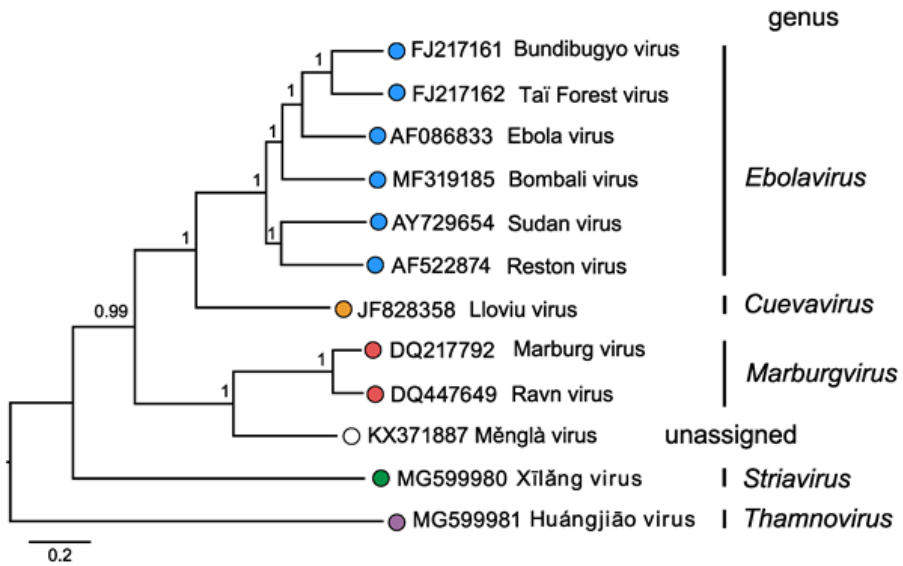

WHERE does it go?

## Bayesian trees




WHEN is the most recent common ancestor?

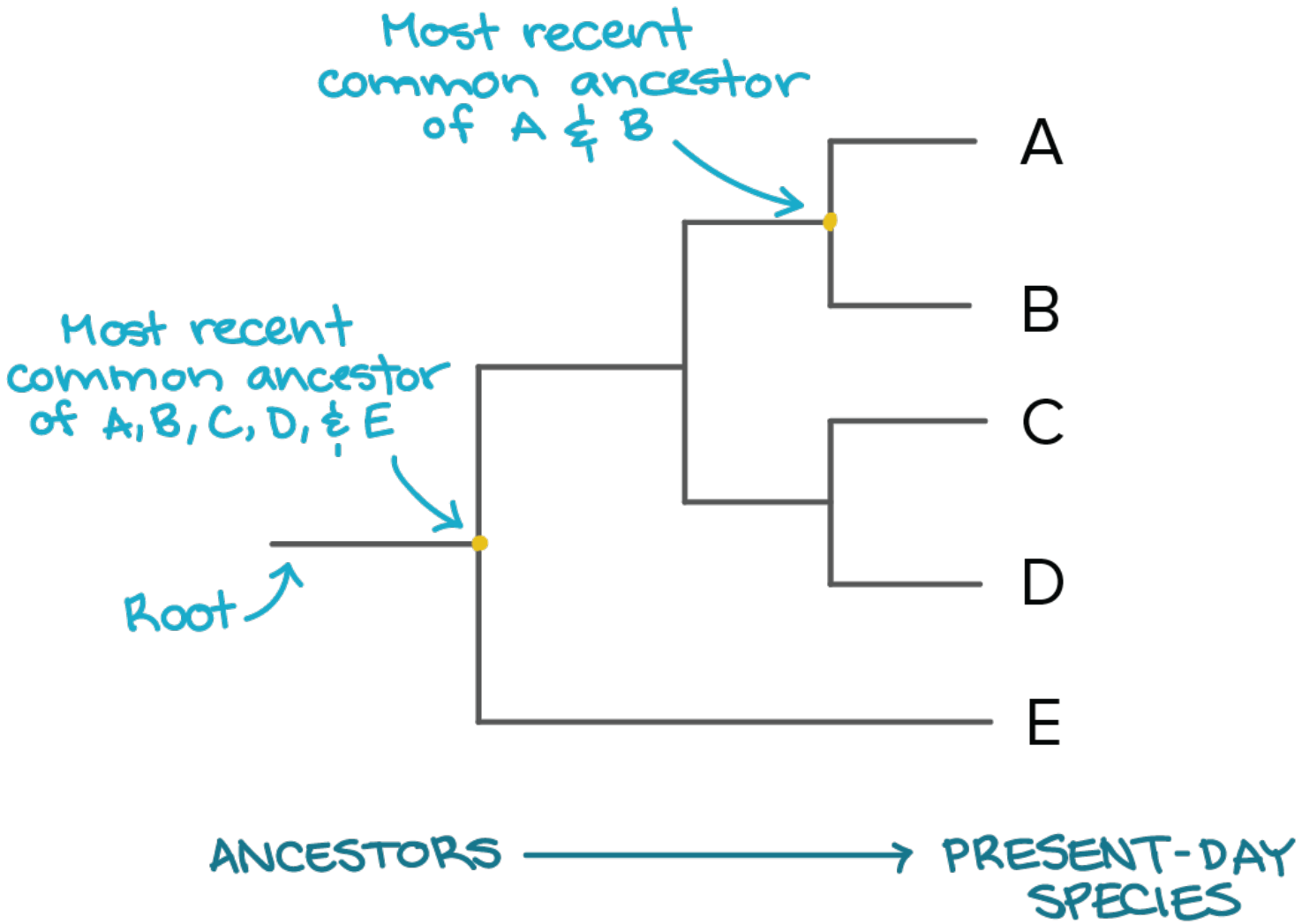
## Maximum likelihood




HOW different is it to what's known?

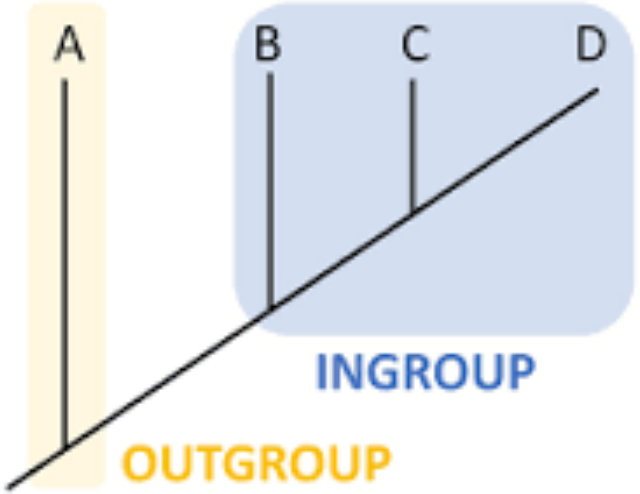
Suchard et al. 2018
Kuhn et al. 2019

# Anatomy of a phylogeny




**CONFIDENCE**

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



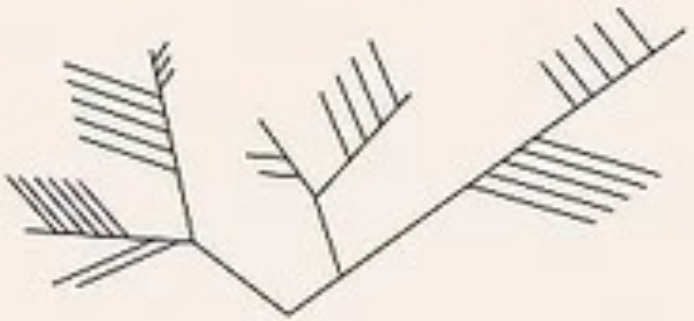
# Cladogram versus phylogenetic tree

## CLADOGRAM

– the relationships are ***hypothetical***

– you can easily make on your own

## PHYLOGENETIC TREE

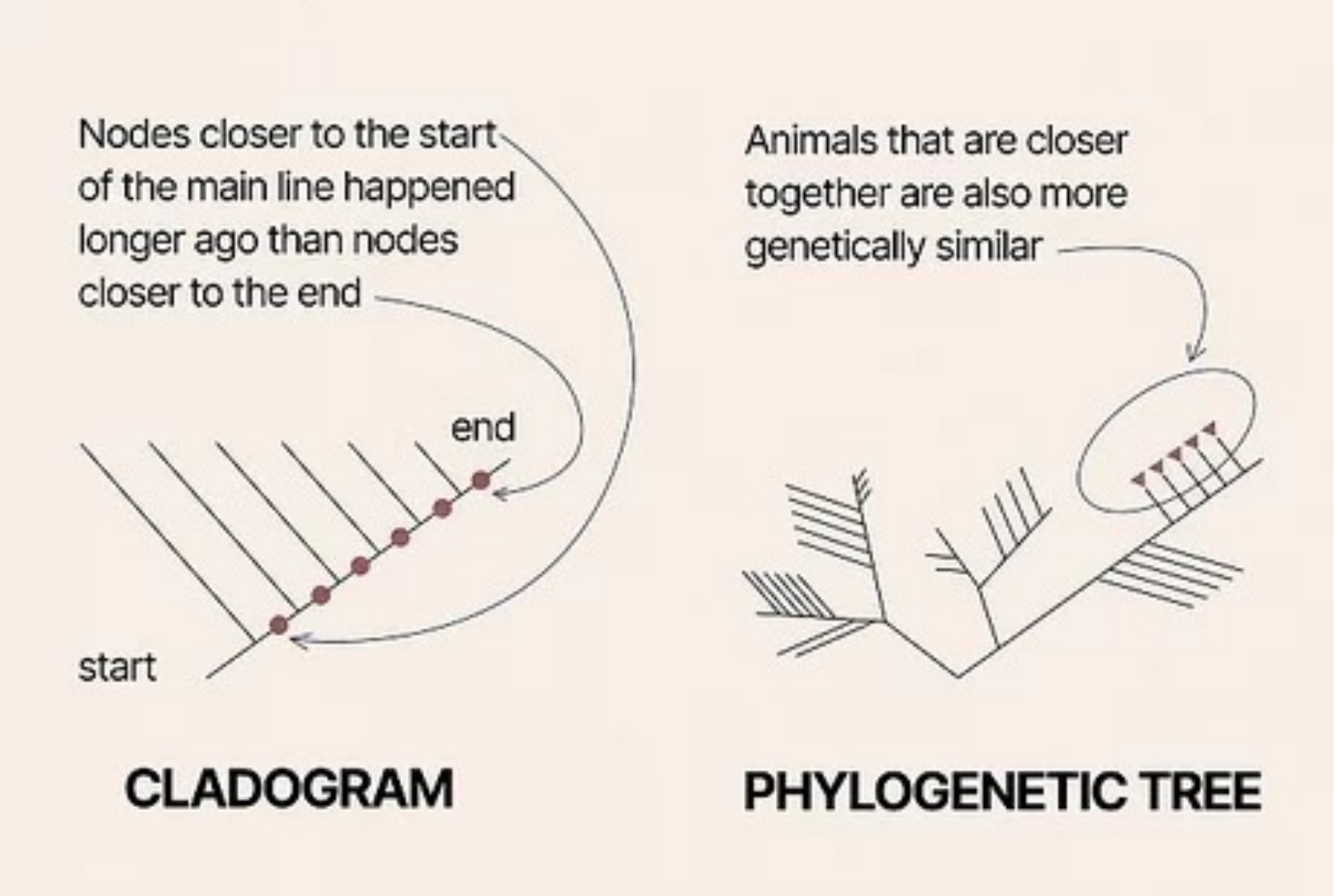


– the relationships are ***backed by molecular evidence***

– should have access to DNA or other molecular data

Wikipedia, 2022

# Cladogram versus phylogenetic tree



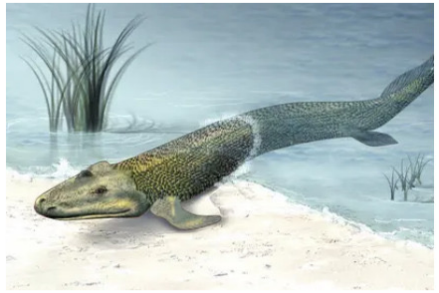

**Why aren't we using cladograms for epidemiology instead of phylogenetic trees?**

Wikipedia, 2022

# Parsimony versus likelihood

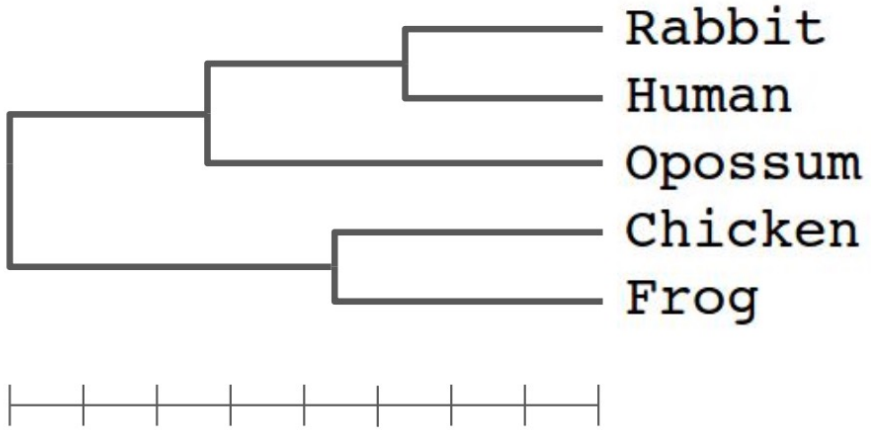
- Parsimony: minimum number of changes
- Likelihood: maximum probability of the data having evolved on the tree

Ancestor



Started Out as a Fish. How Did It End Up Like This?

Descendant


**branch length** can mean different things:

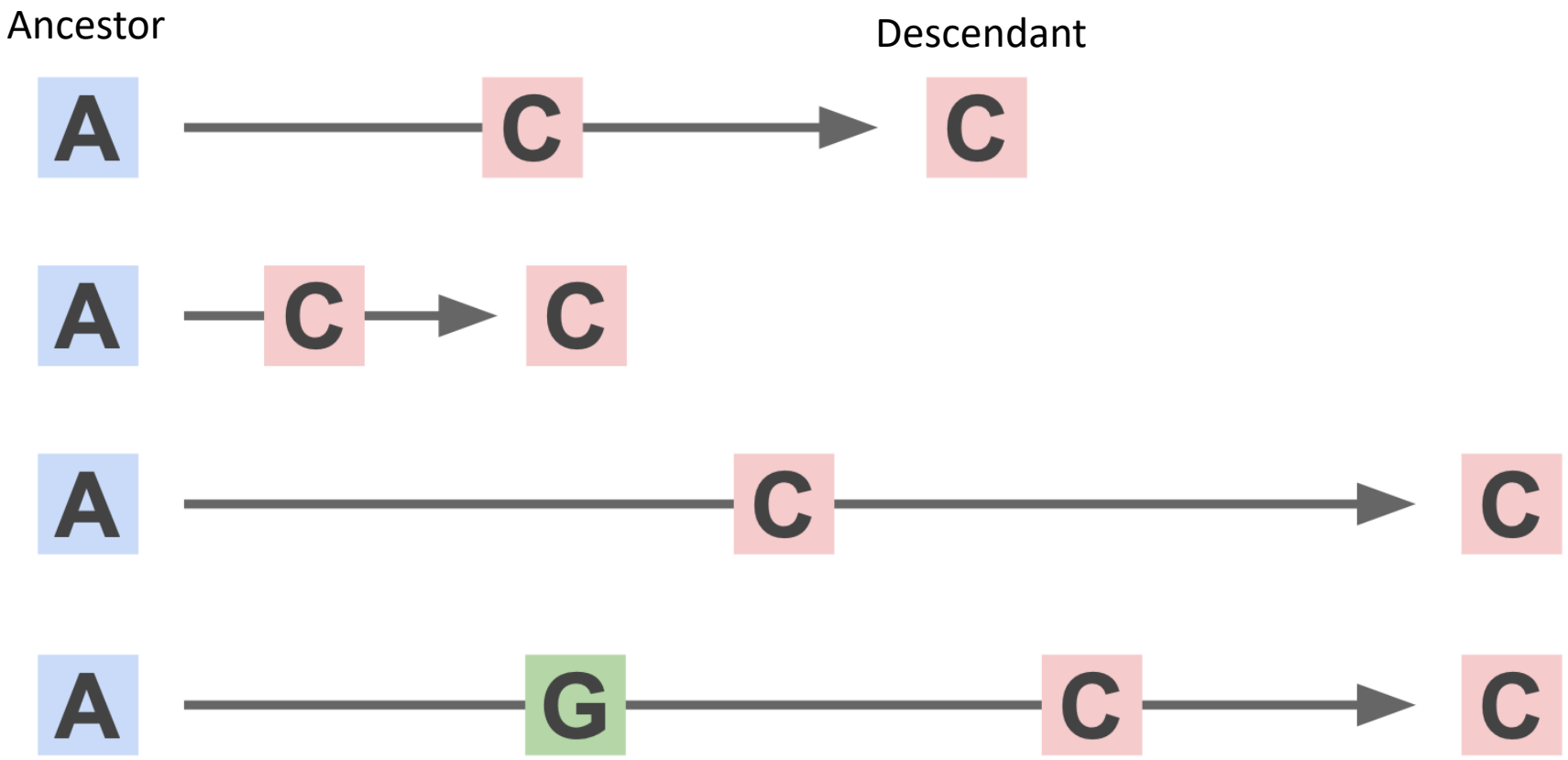
- minimum number of changes (parsimony)
- time; opportunity for change
- expected number of changes, given a model of evolution

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

Adapted from Ree and Hipp, UChicago, 2021

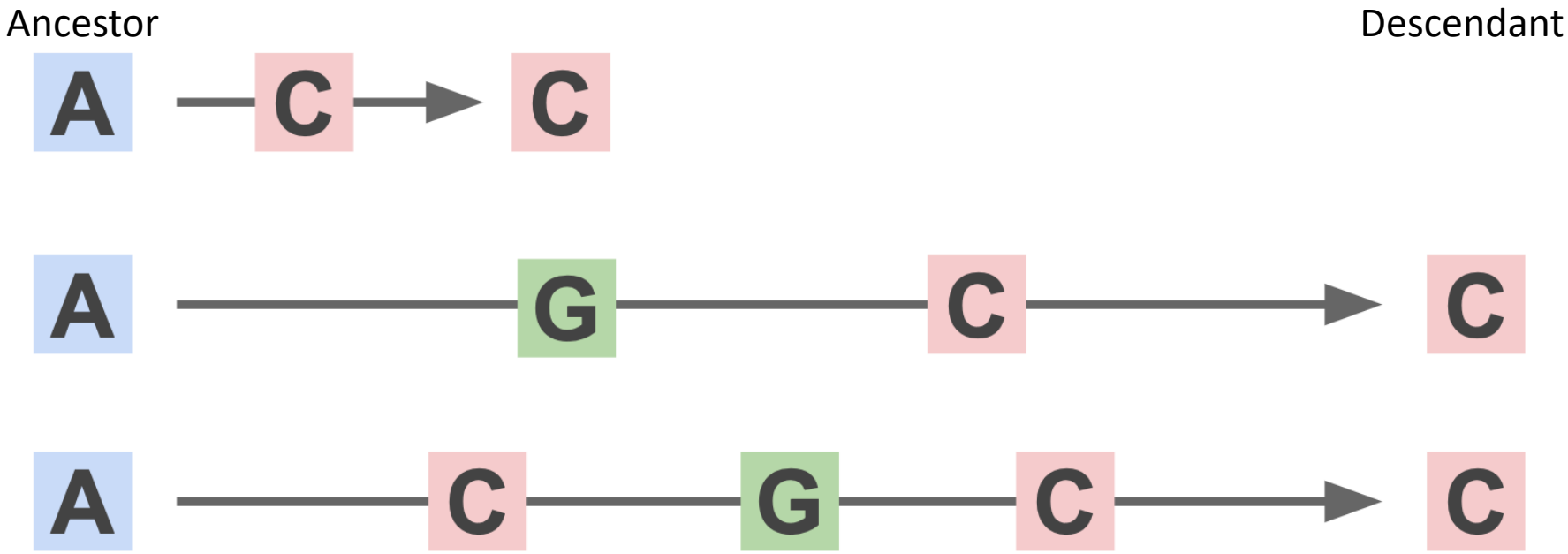
# Parsimony: minimum number of changes regardless of time/opportunity

**Which is the least parsimonious?**




Adapted from Ree and Hipp, UChicago, 2021

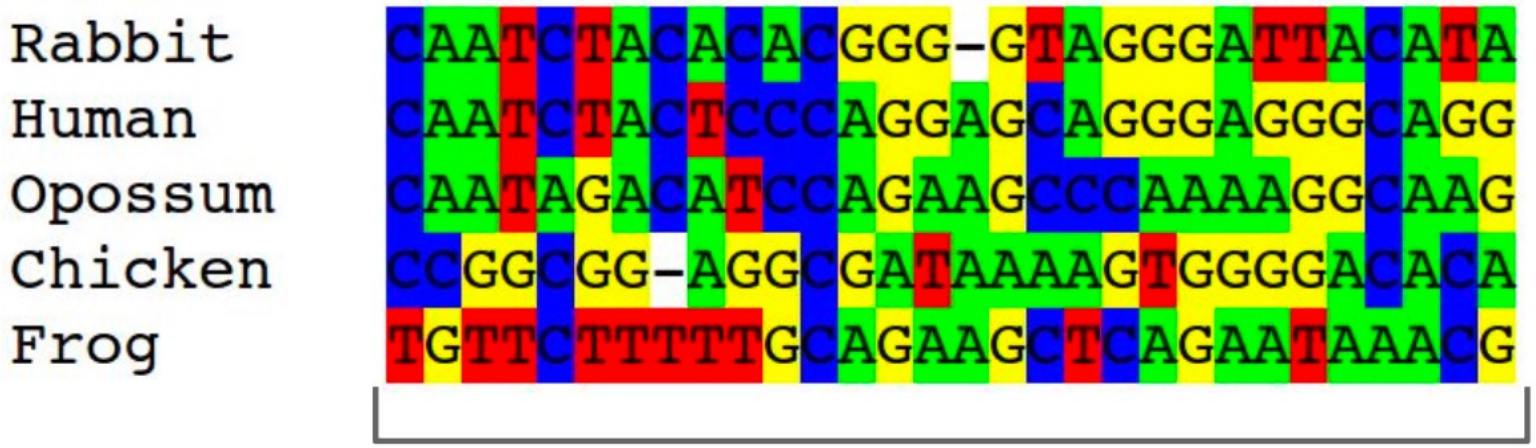
# Likelihood: probability of ancestral and descendant status is a function of time (branch length)



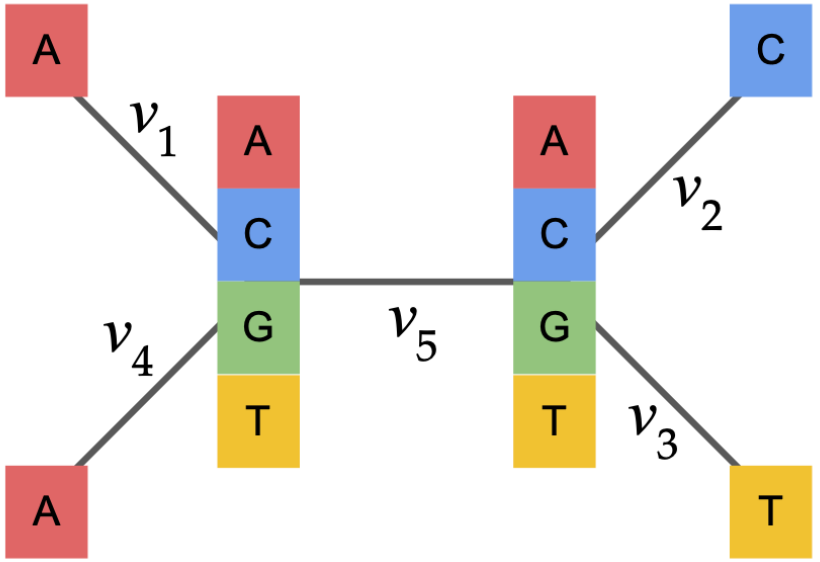

We don’t know what the actual history of the change is, so use a model of evolution to consider all possible histories (maximum likelihood)

Adapted from Ree and Hipp, UChicago, 2021

# Likelihood cont'd.




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




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

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

Adapted from Ree and Hipp, UChicago, 2021

# 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 $\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}\right)$ and equal mutation rates. The only parameter of this model is therefore $\mu$, the overall substitution rate. As previously mentioned, this variable becomes a constant when we normalize the mean-rate to 1.

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

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

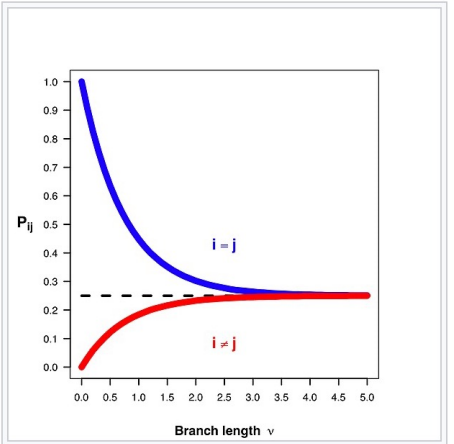
When branch length, $\nu$, is measured in the expected number of changes per site then:

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

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 $\mu$.

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 ($\nu$) 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).

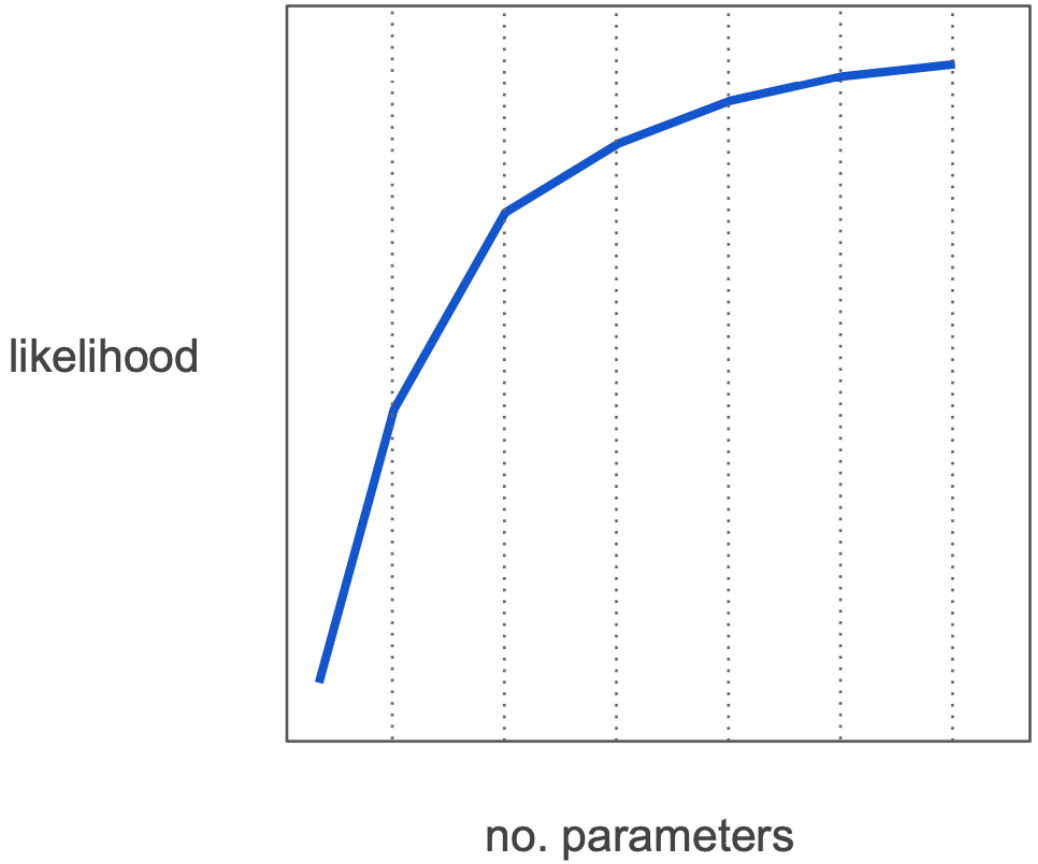
# Model selection

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




Adapted from Ree and Hipp, UChicago, 2021

# DNA models

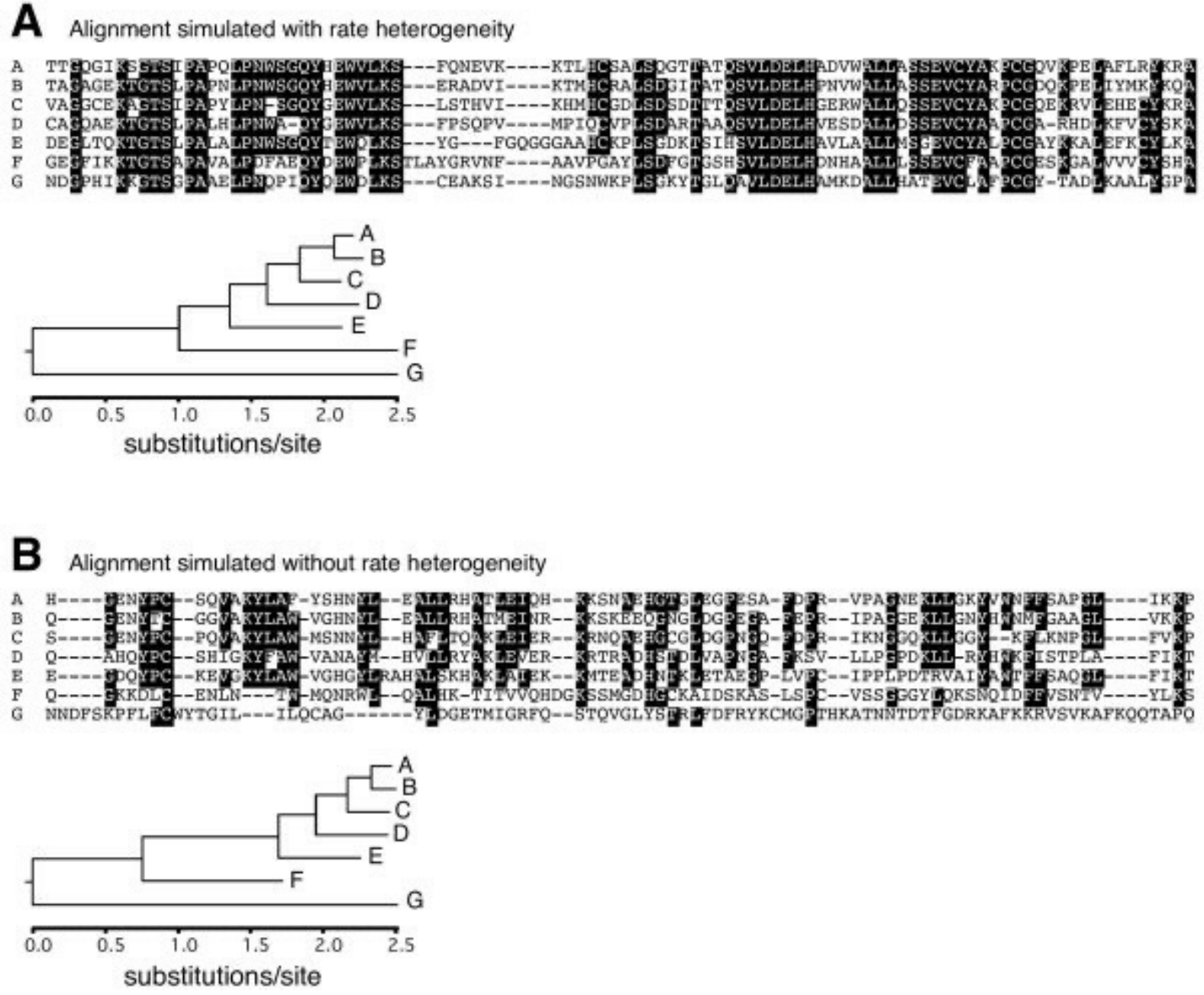
## Base substitution rates

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

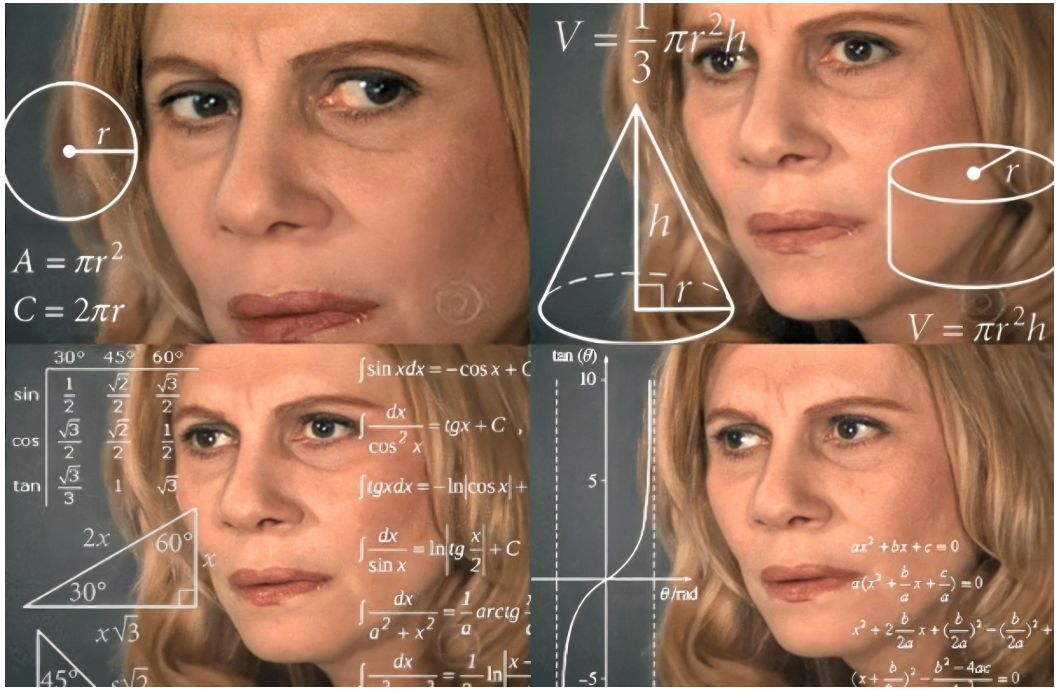
- Do we expect all sites in an alignment to evolve at the same rate? What kind events would affect this?



# 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




Adapted from Ree and Hipp, UChicago, 2021

# Rate heterogeneity across sites

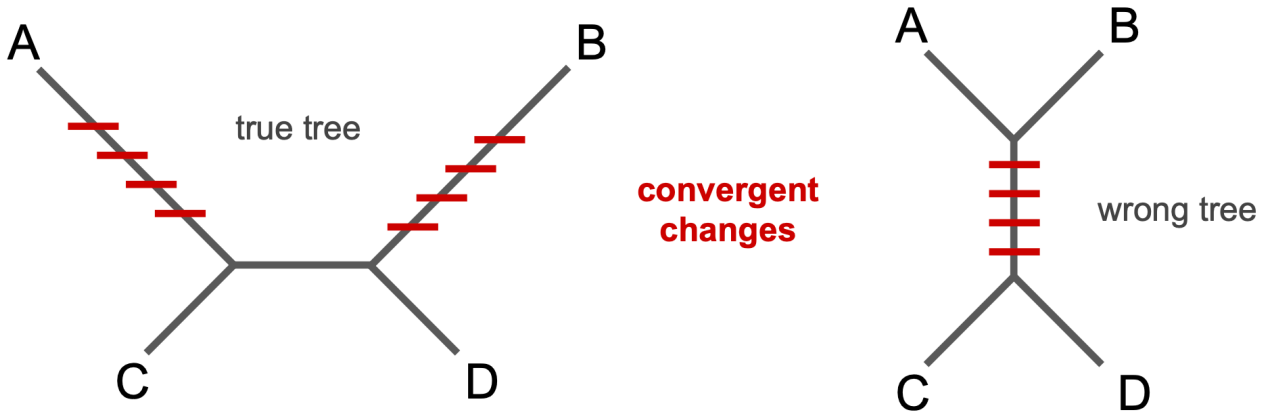
IQ-TREE supports all common rate heterogeneity across sites models:

| 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. `+G8` . |
| +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 `+G` model by relaxing the assumption of Gamma-distributed rates. The number of categories can be specified with e.g. `+R6` (default 4 categories if not specified). The FreeRate model typically fits data better than the `+G` model and is recommended for analysis of large data sets. |
| +I+R | invariable site plus FreeRate model. |

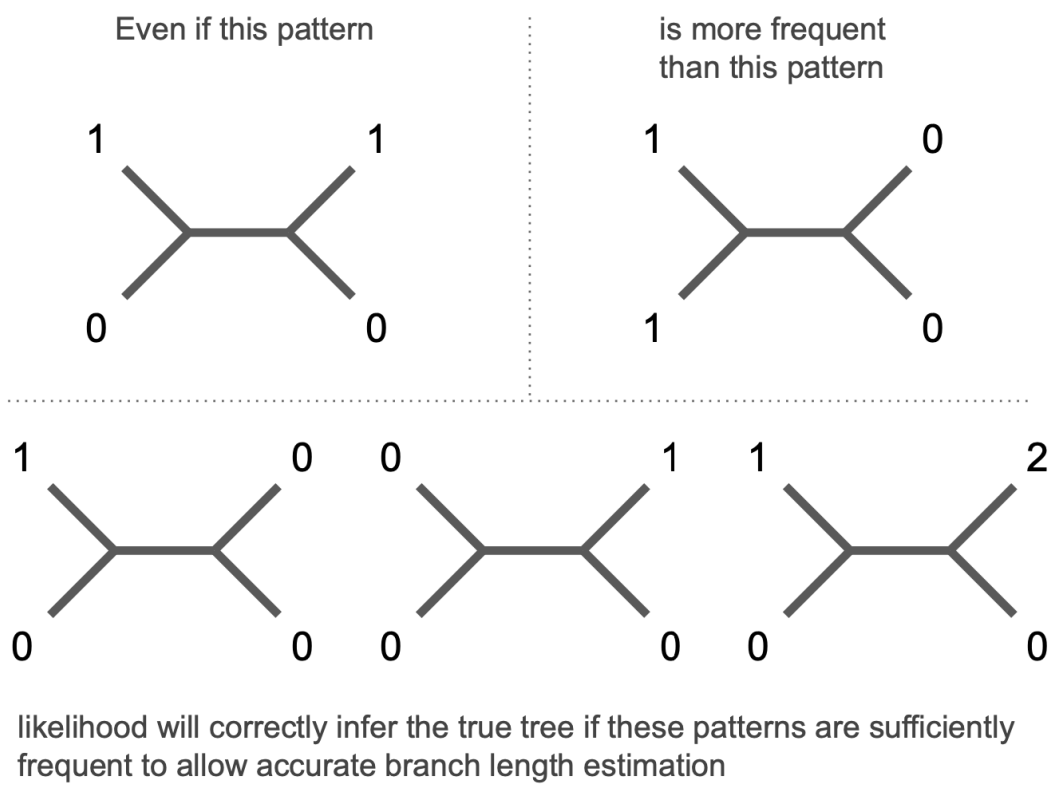
IQ-TREE, 2022

# 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




likelihood will correctly infer the true tree if these patterns are sufficiently frequent to allow accurate branch length estimation

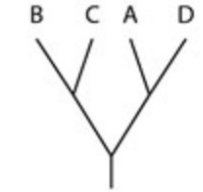
Adapted from Ree and Hipp, UChicago, 2021

# 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


Original
sequence alignment

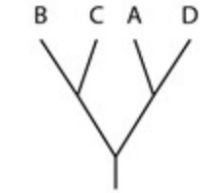
| | Site number | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Species A | A | A | T | G | C | T | A | G | T | G | G | T | G | A | T |
| Species B | A | A | G | C | T | A | T | G | G | T | G | A | T | C | G |
| Species C | A | G | C | C | T | A | T | G | T | G | G | A | A | C | G |
| Species D | A | A | C | C | C | A | T | T | G | G | G | T | G | A | T |




Original alignment
tree

Bootstrap
pseudo-replicate #1

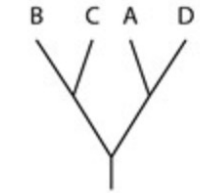
| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | 3 | 3 | 1 | 12 | 9 | 2 | 4 | 11 | 13 | 10 | 14 | 8 | 11 | 13 |
| Species A | C | T | T | A | T | T | A | G | G | G | G | A | G | G | G |
| Species B | T | G | G | A | A | G | A | C | G | T | T | C | G | G | T |
| Species C | T | C | C | A | A | T | G | C | G | A | G | C | G | G | A |
| Species D | C | C | C | A | T | T | A | C | G | G | G | A | T | G | G |




Bootstrap tree #1

Bootstrap
pseudo-replicate #2

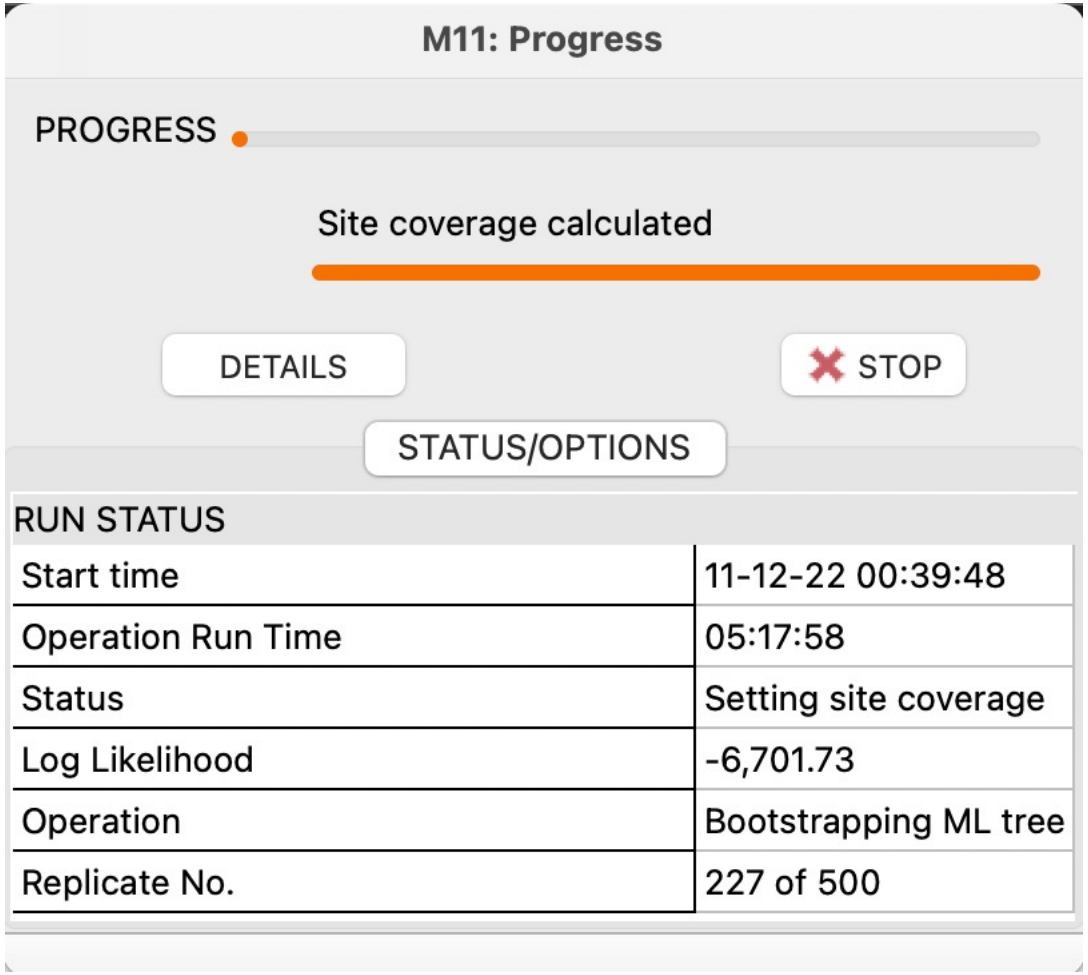
| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9 | 7 | 12 | 5 | 2 | 4 | 2 | 6 | 14 | 9 | 4 | 9 | 7 | 2 | 1 |
| Species A | T | A | T | C | A | G | A | T | A | T | G | T | A | A | A |
| Species B | G | T | A | T | A | C | A | A | C | G | C | G | T | A | A |
| Species C | T | T | A | T | G | C | G | A | C | T | C | T | T | G | A |
| Species D | T | T | T | C | A | C | A | A | A | T | C | T | T | A | A |




Bootstrap tree #2

# 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?**



JOURNAL ARTICLE

# 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?

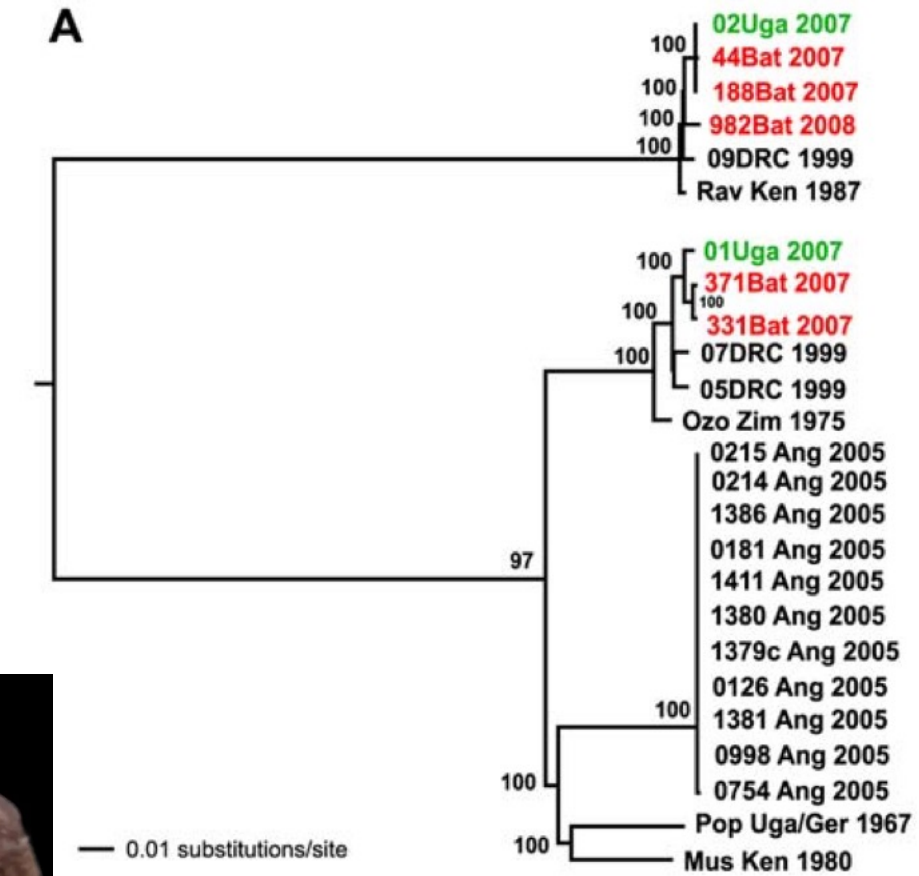
# Case study: Marburg virus – where did it come from?

- 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?

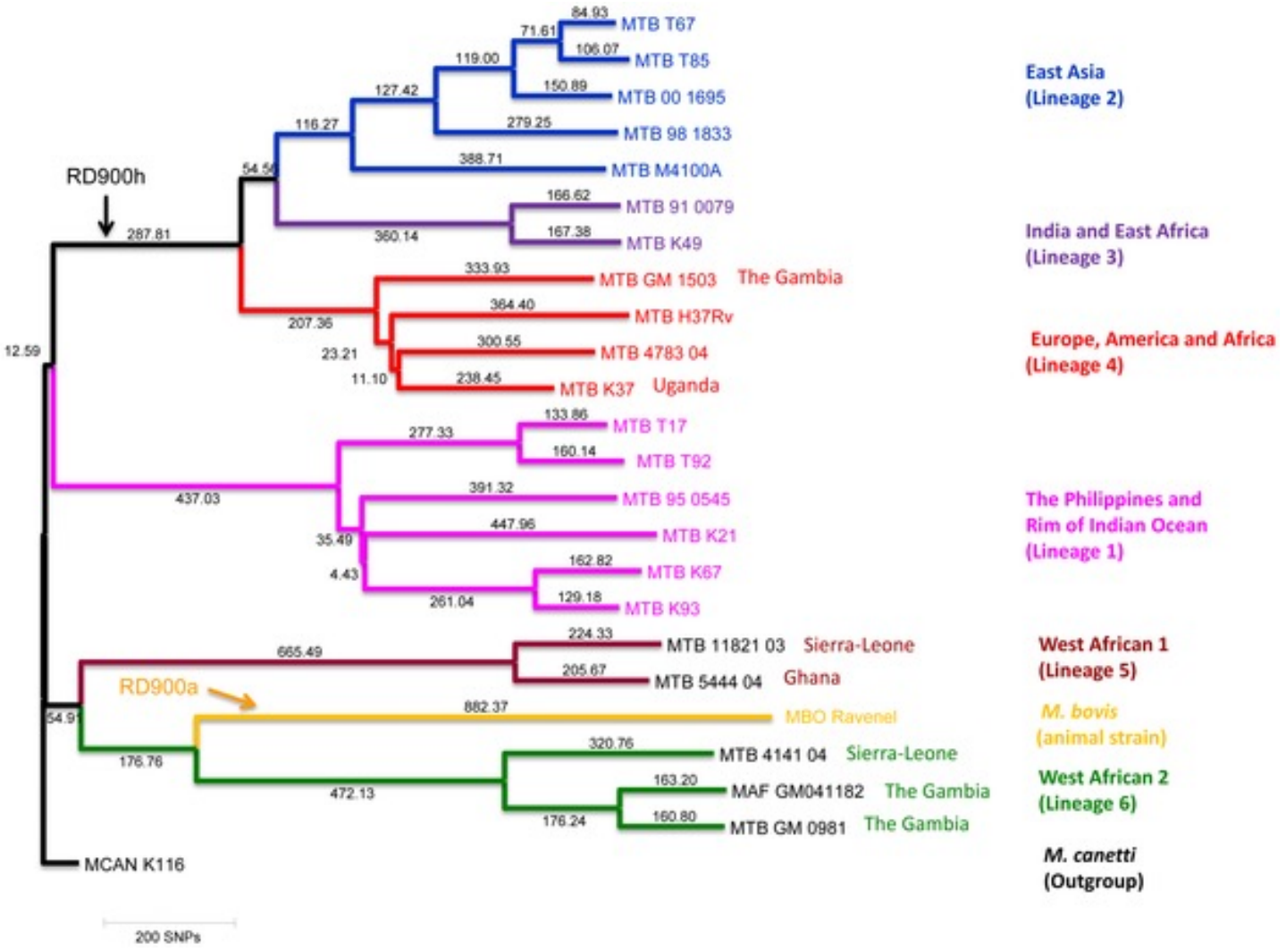
- Marburg filovirus causes severe hemorrhagic fever in humans through sporadic spillovers
- Like Ebola virus, it was suspected to have been hosted by bats, but we did not know for sure

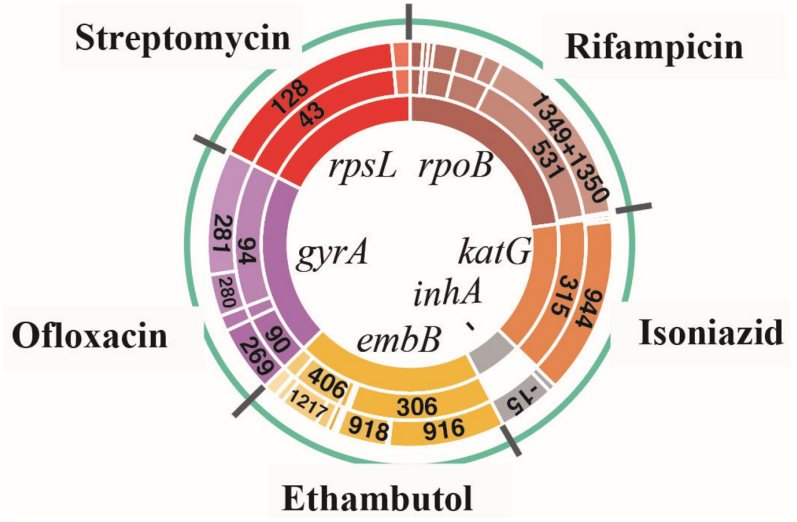



Green: human sequences
Red: bat sequences
Black: reference sequences

Towner et al. 2009

# Topic: tuberculosis

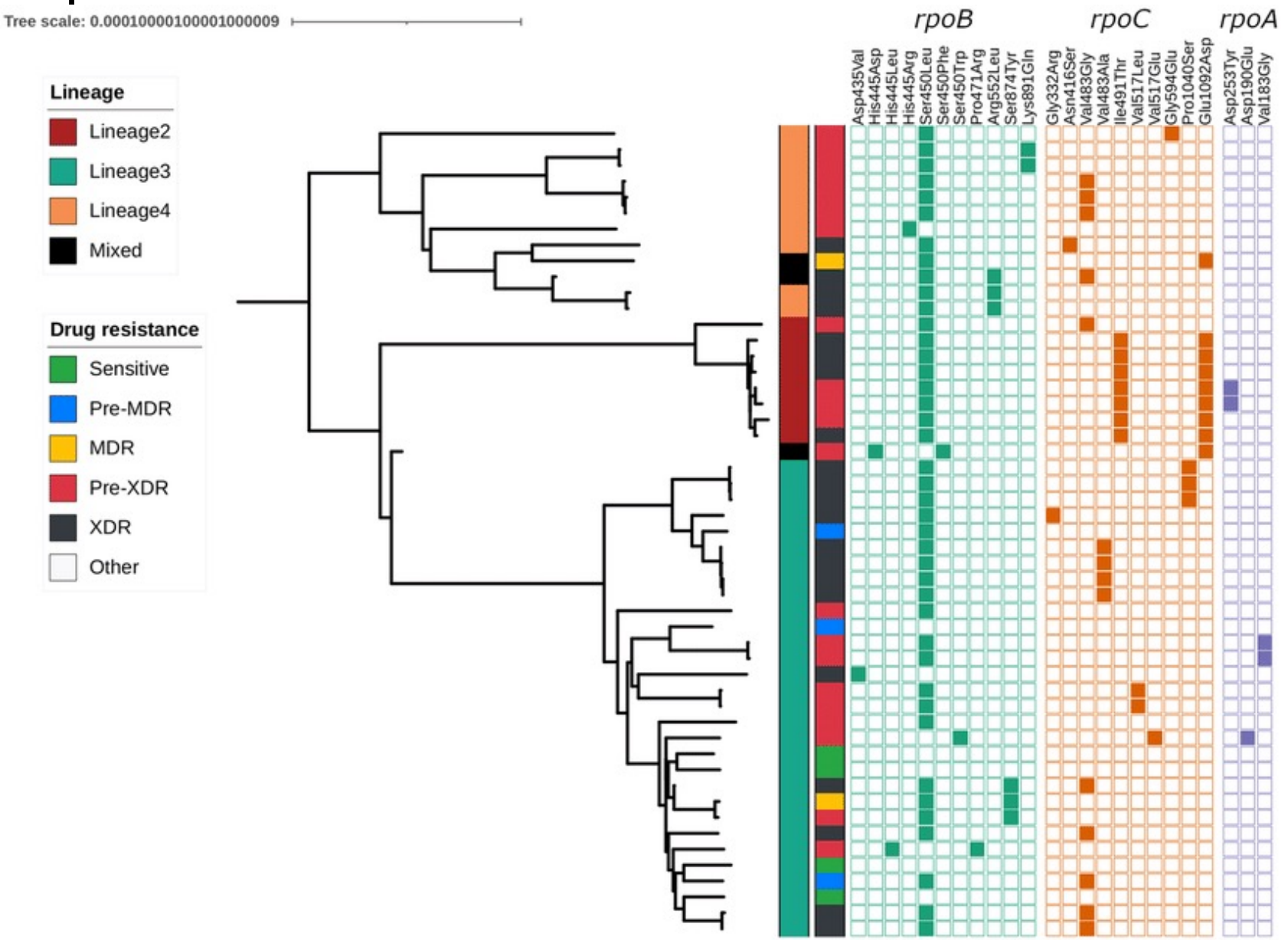



**What do you think the above plot is telling you? It has to do with drug resistance…the phylogeny can help us track something…**

Gagneux et al. 2013
Jia et al. 2021

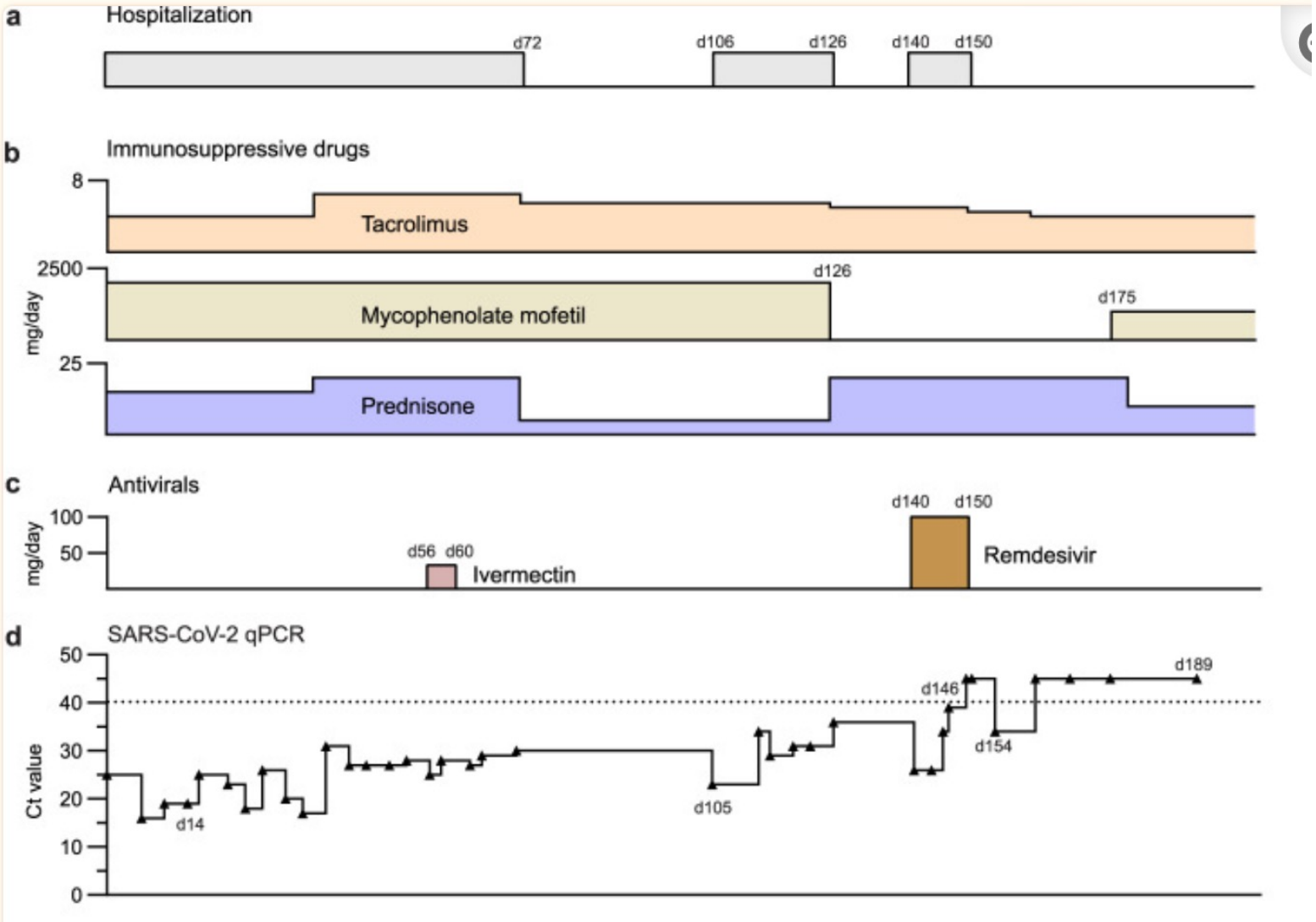
# Topic: tuberculosis




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

Khan et al. 2021

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

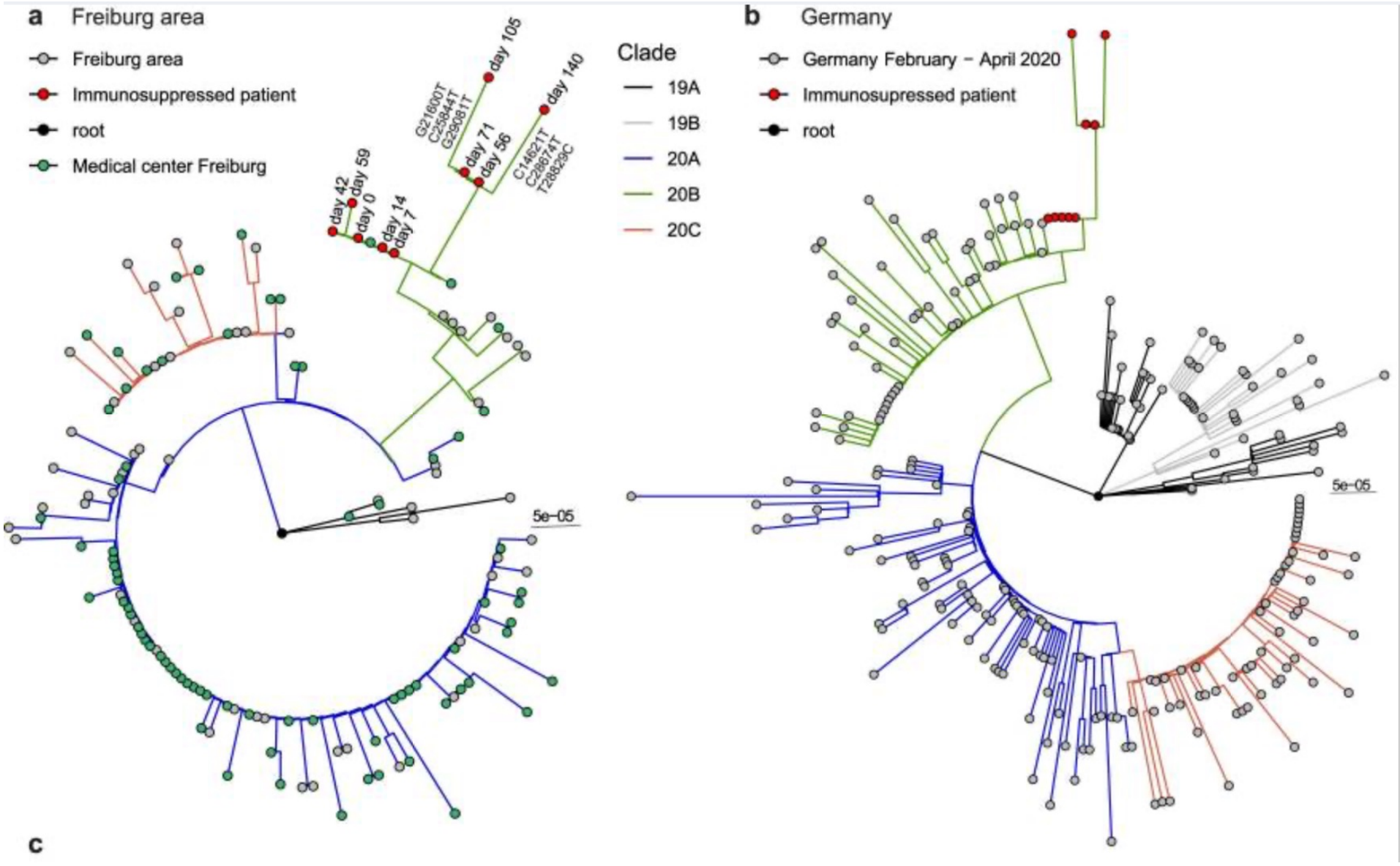



"A 58-year-old male with a history of autosomal dominant polycystic kidney disease was admitted to our university hospital, for renal transplantation performed on March 2020. The patient was treated with a cocktail of tacrolimus, mycophenolate, and prednisone from March until the end of September"

**What do you expect to happen to the virus?**

Weigang et al. 2021

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




Weigang et al. 2021

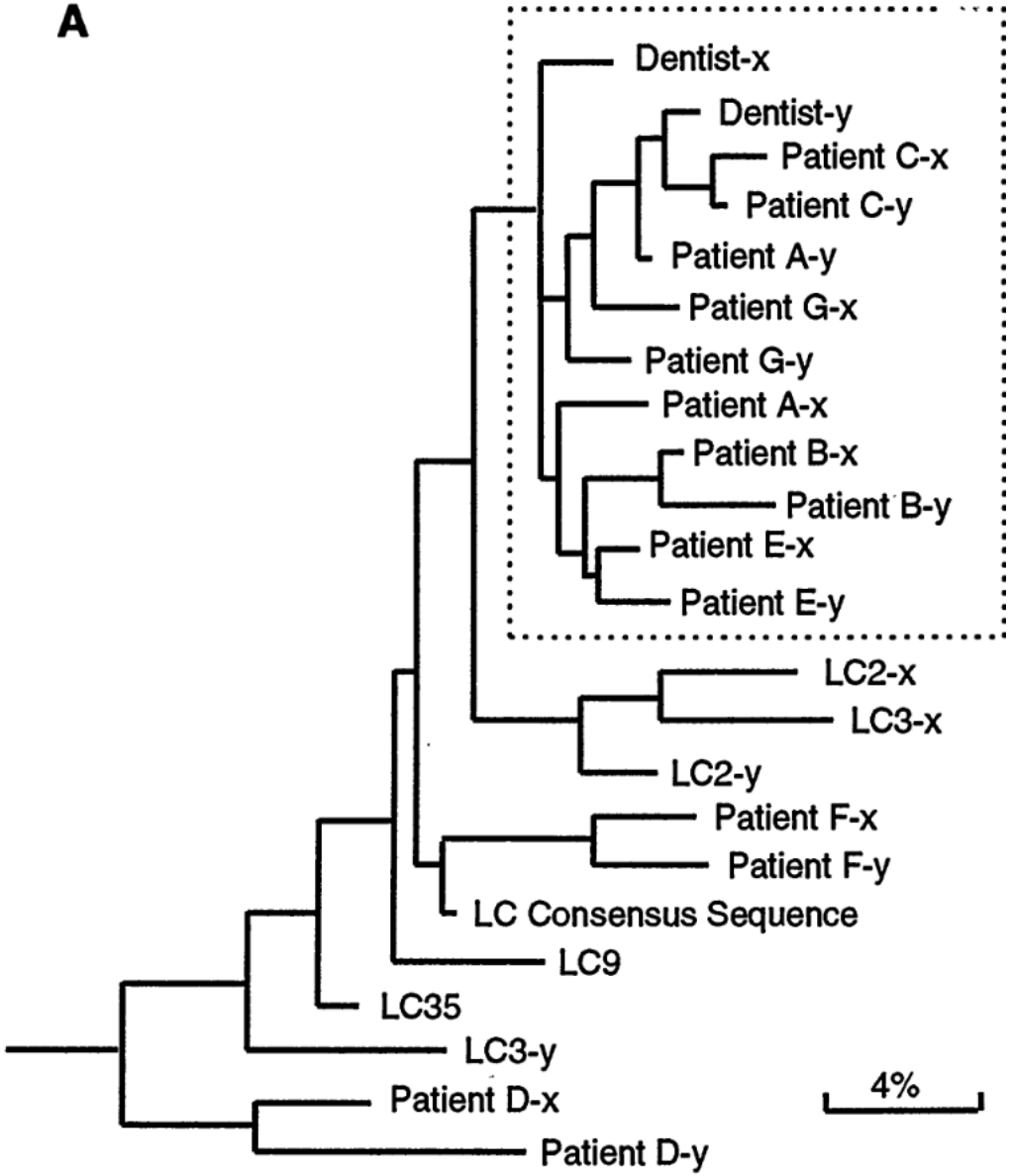
# Case study: a literal court case

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

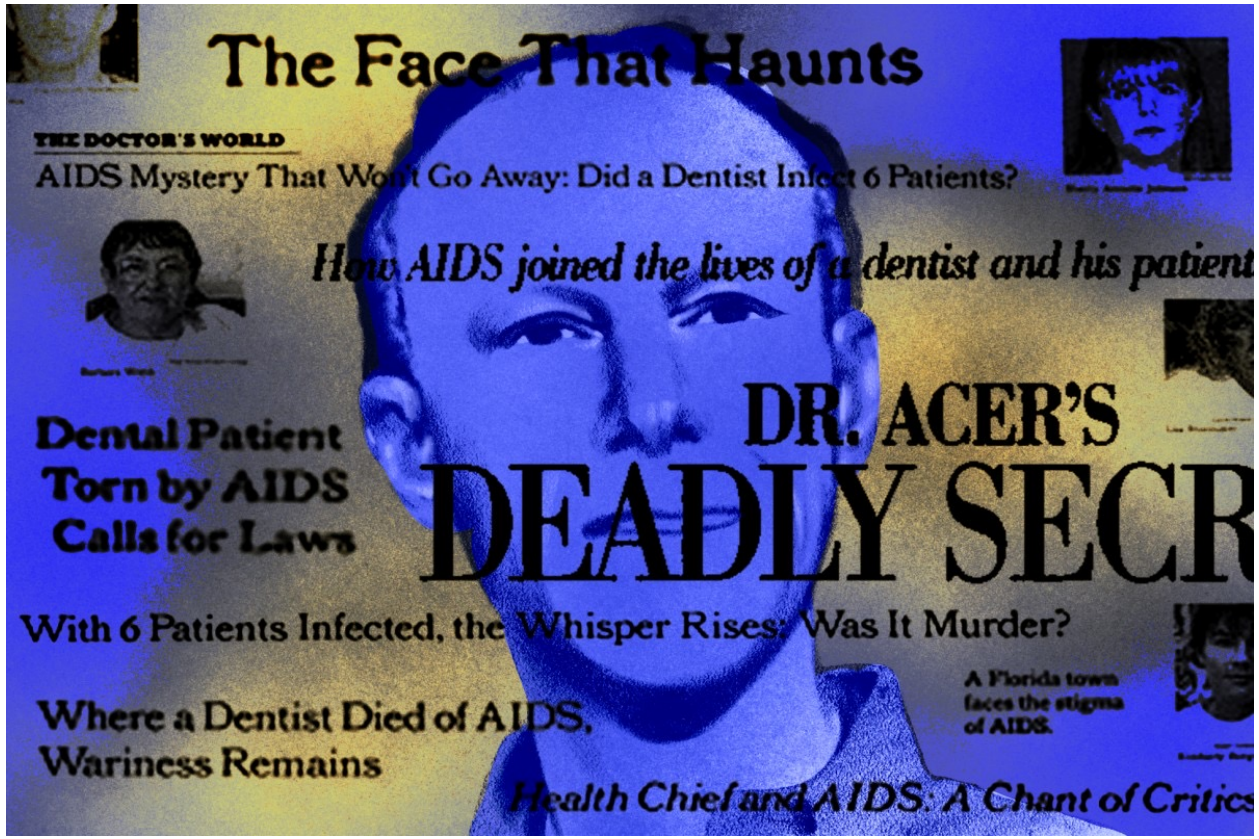

With 6 Patients Infected, the Whisper Rises: Was It Murder?



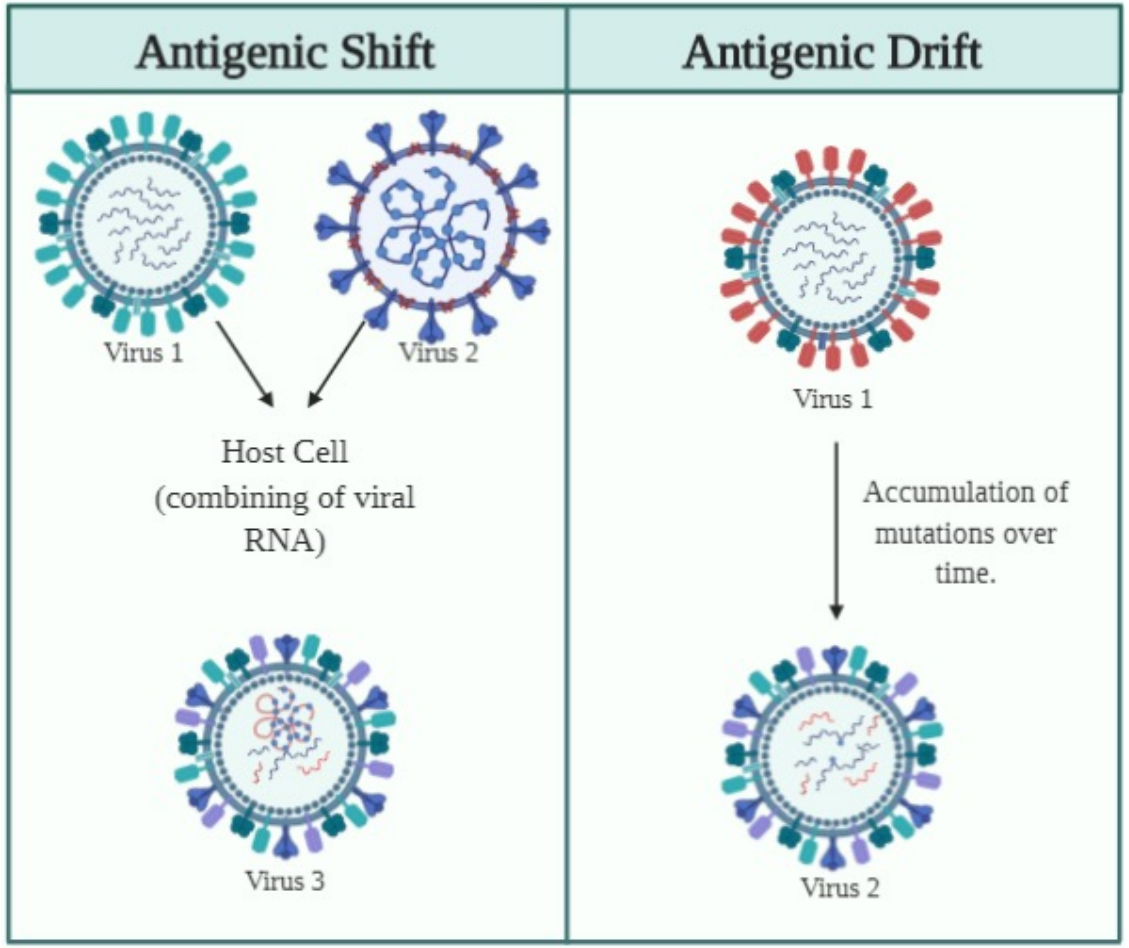

NYTimes archives
Ou et al. 1992

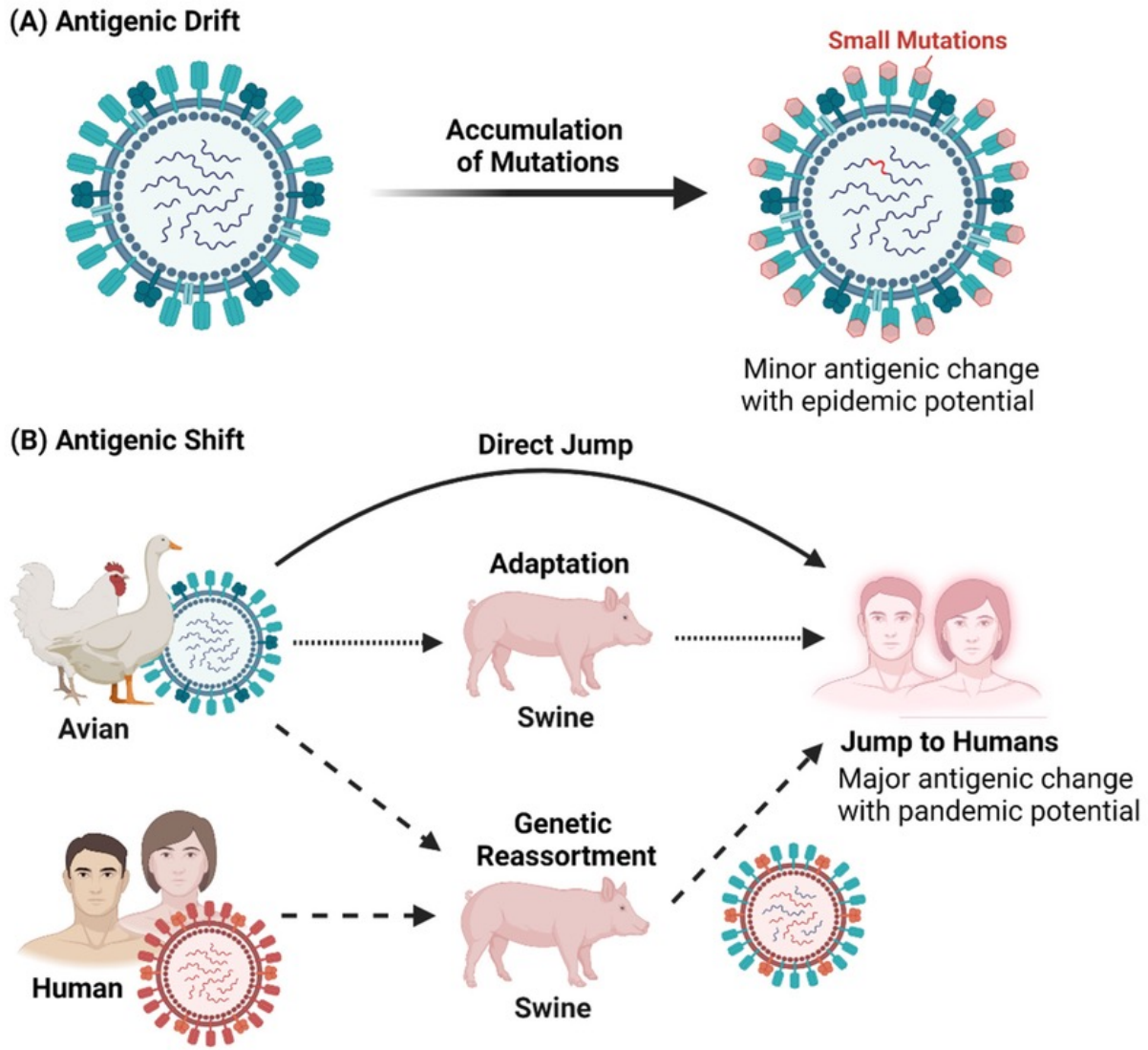
# Case study: a literal court case

- 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



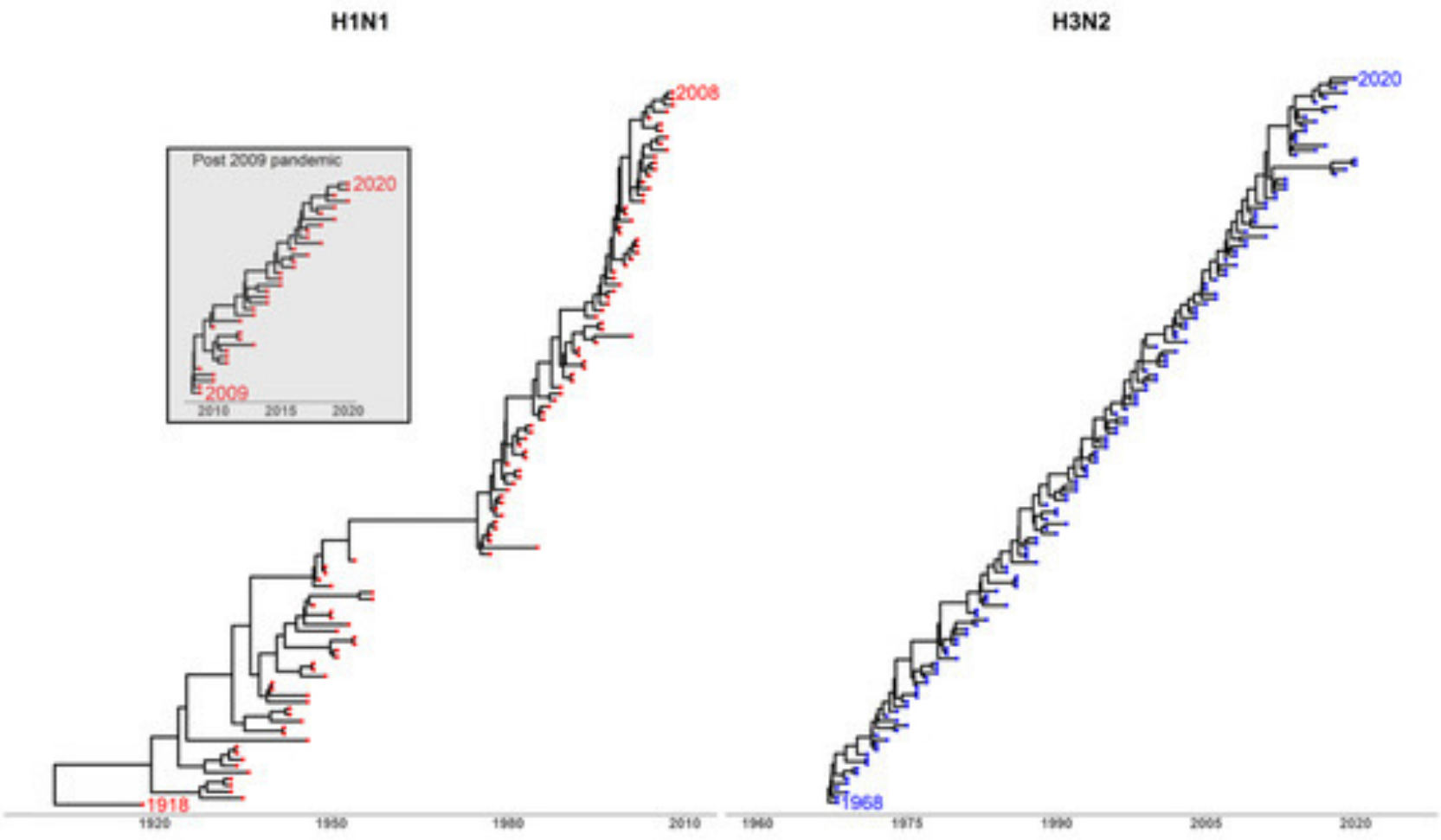
# Case study: influenza – antigenic shift and drift




Yeo et al. 2021

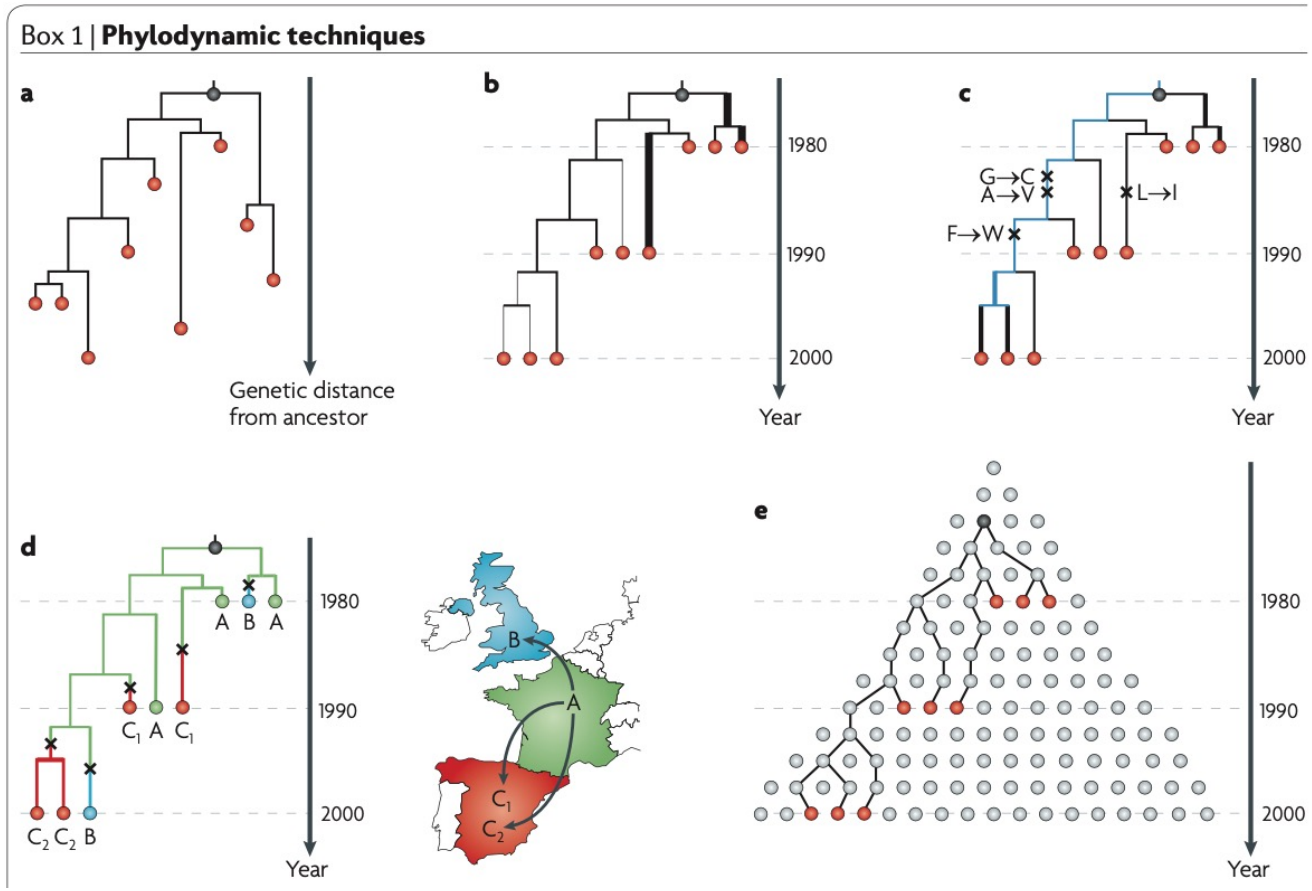
# Case study: influenza – antigenic shift and drift




Bolton et al. 2021

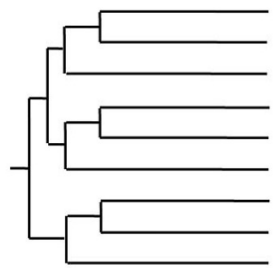
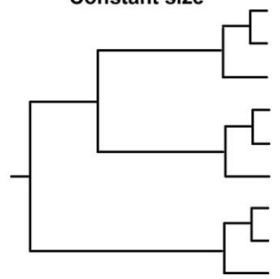
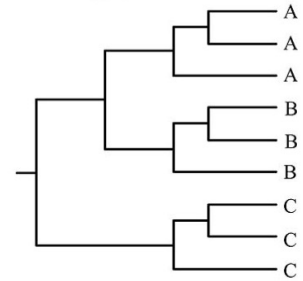
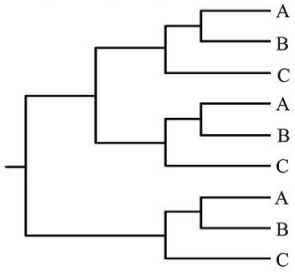
**Is drift or shift happening here?**

**How does this impact vaccine design?**




# Units matter

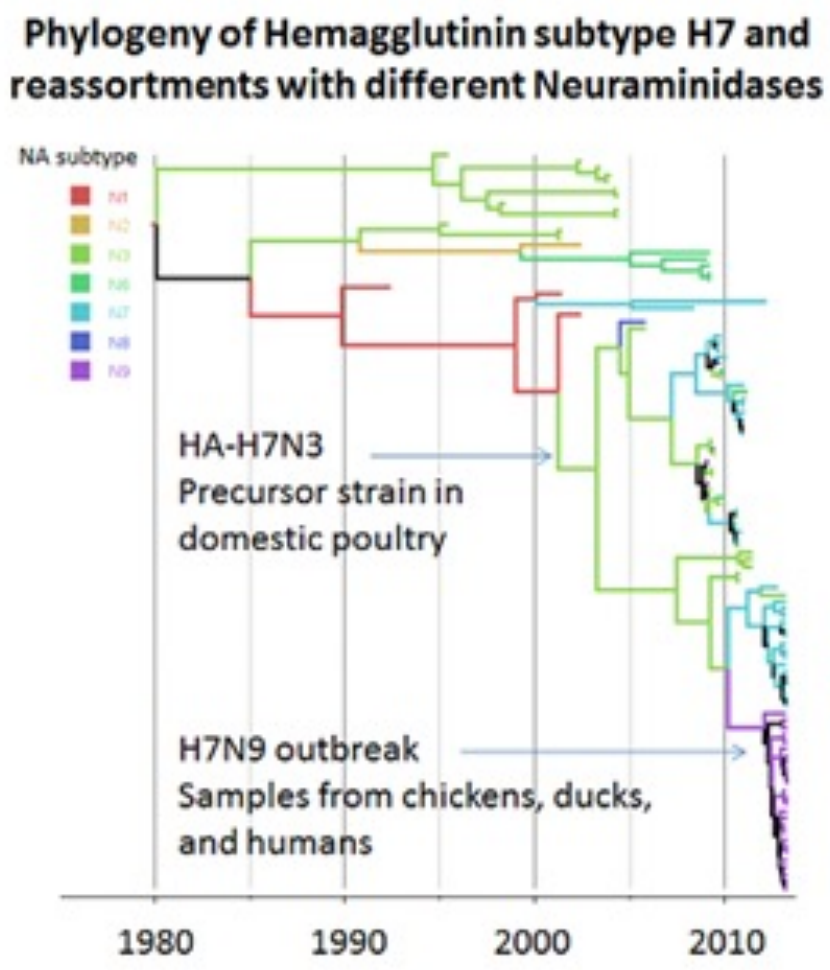
# Shapes matter

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

Grenfell et al. 2004
Pybus and Rambaut 2009

# 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

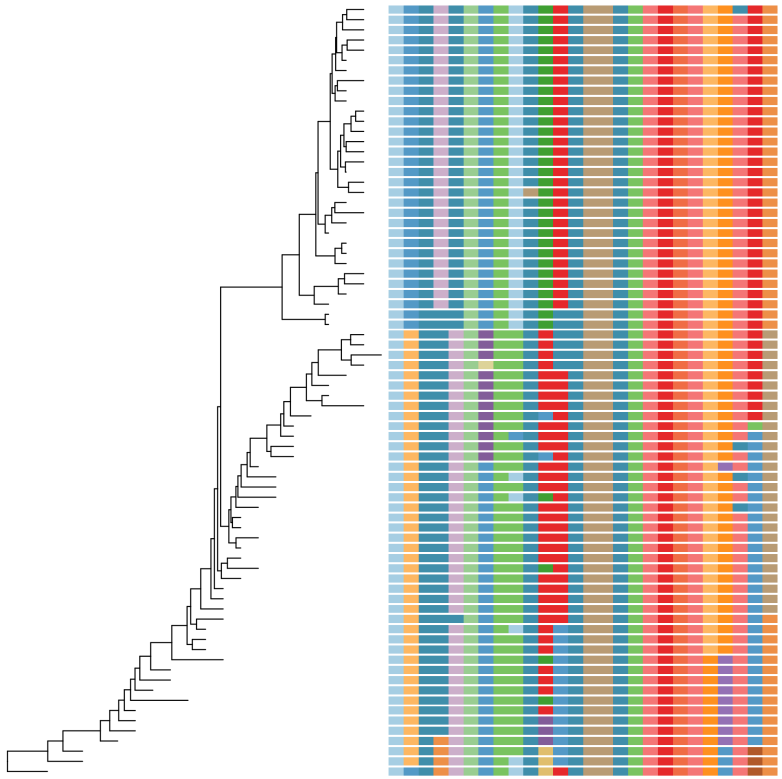
# So you have your sequences, now what?

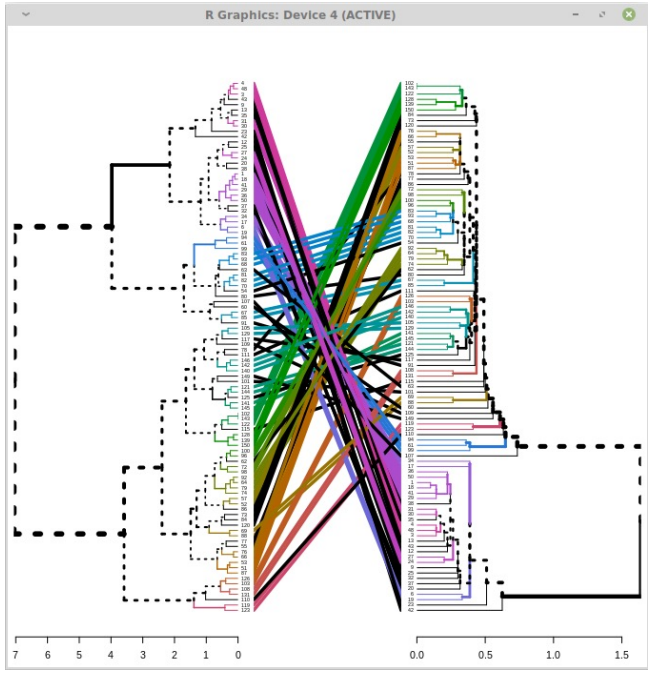
- Get some reference sequences from NCBI
- Get an outgroup from NCBI
- Align them (use a software like MEGA or online like MAFFT)
- Pick the best model (use a software like MEGA or ModelTest-NG)
- Run the phylogeny using your aligned sequences and chosen model (use a software like MEGA or RAxML)
- Visualize/edit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

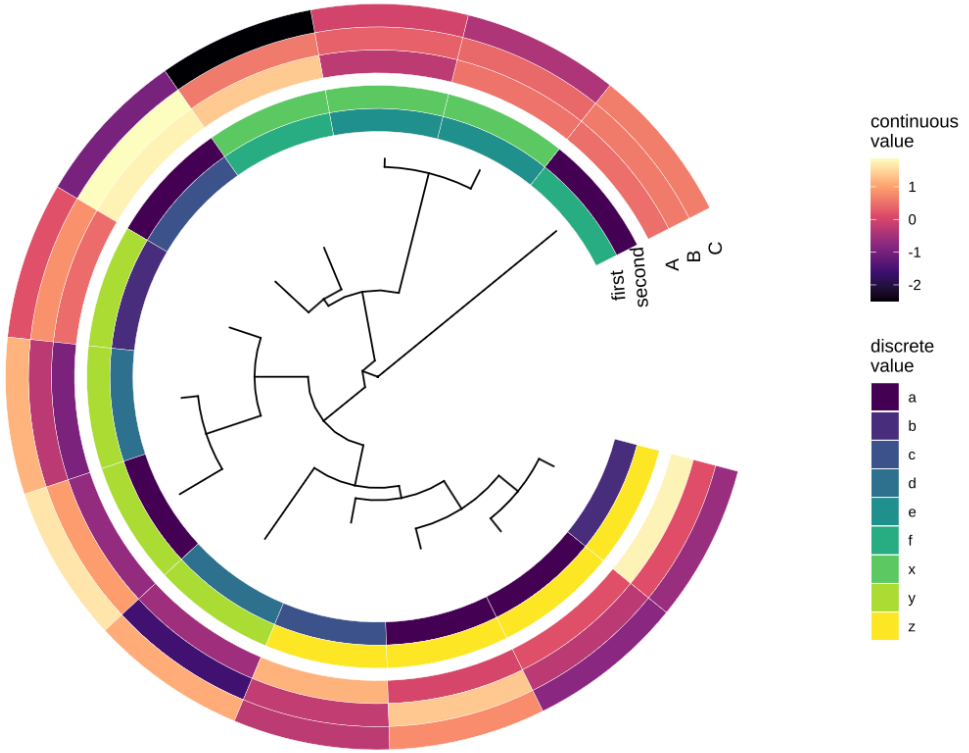
All of this listed is free to use ☺

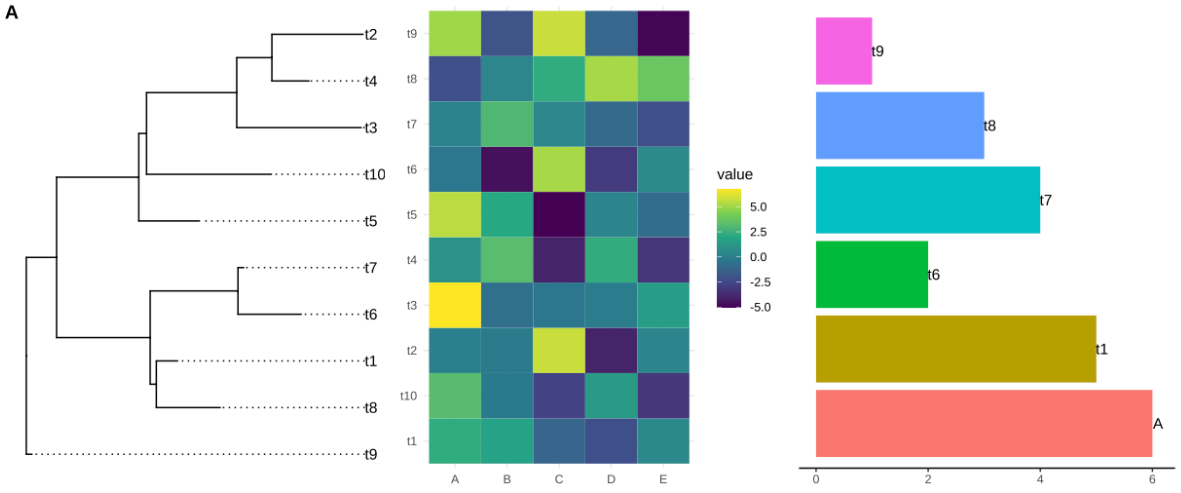
# Potential for pretty figures

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



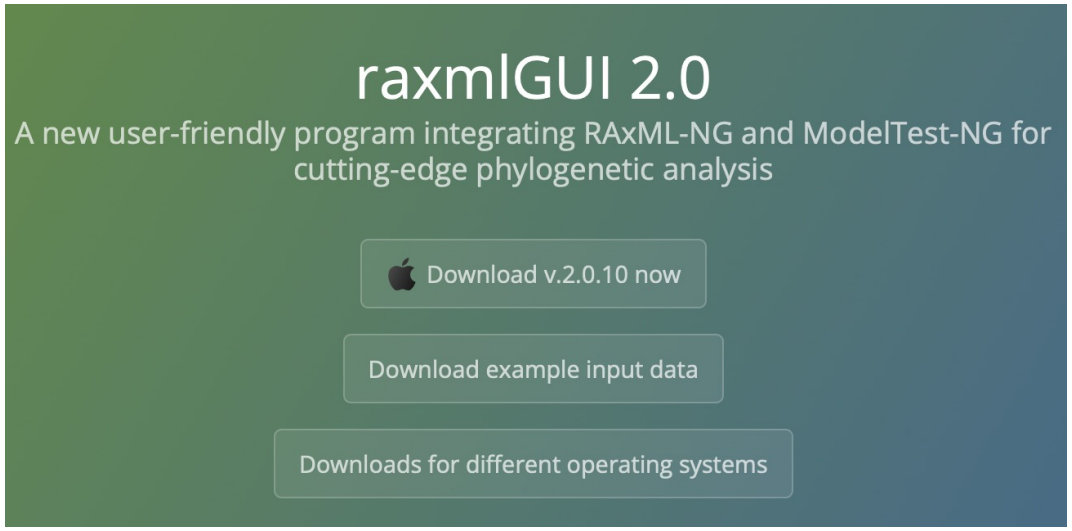


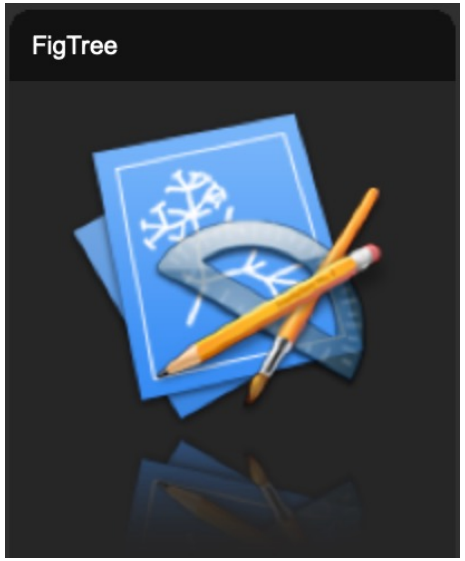



# TUTORIAL

- 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

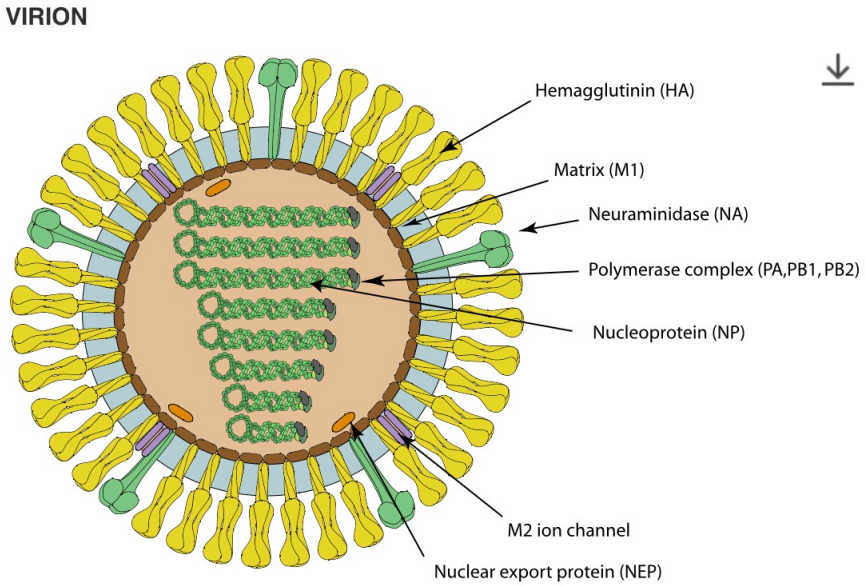


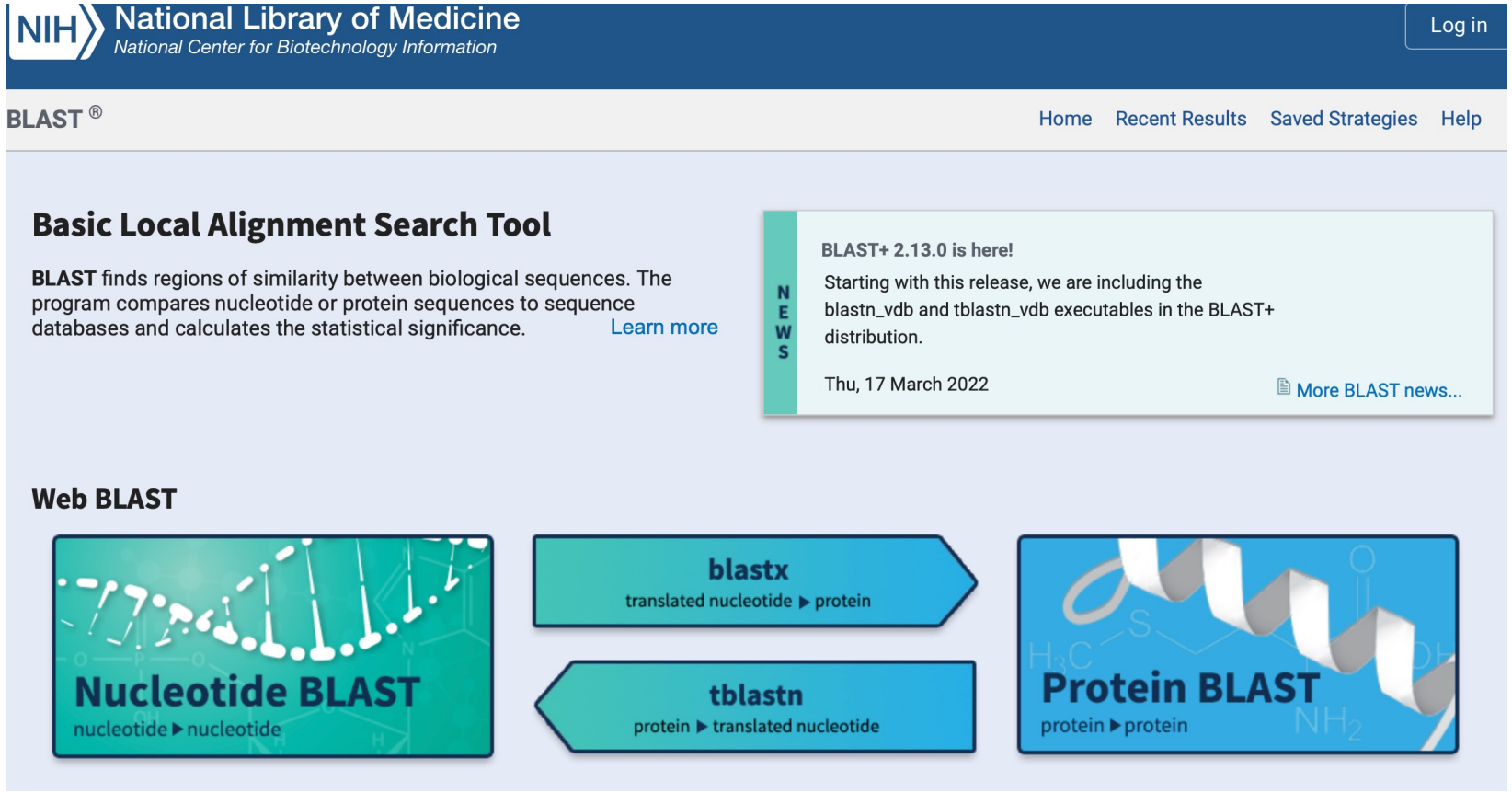

MEGA Molecular Evolutionary Genetics Analysis

# 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?**

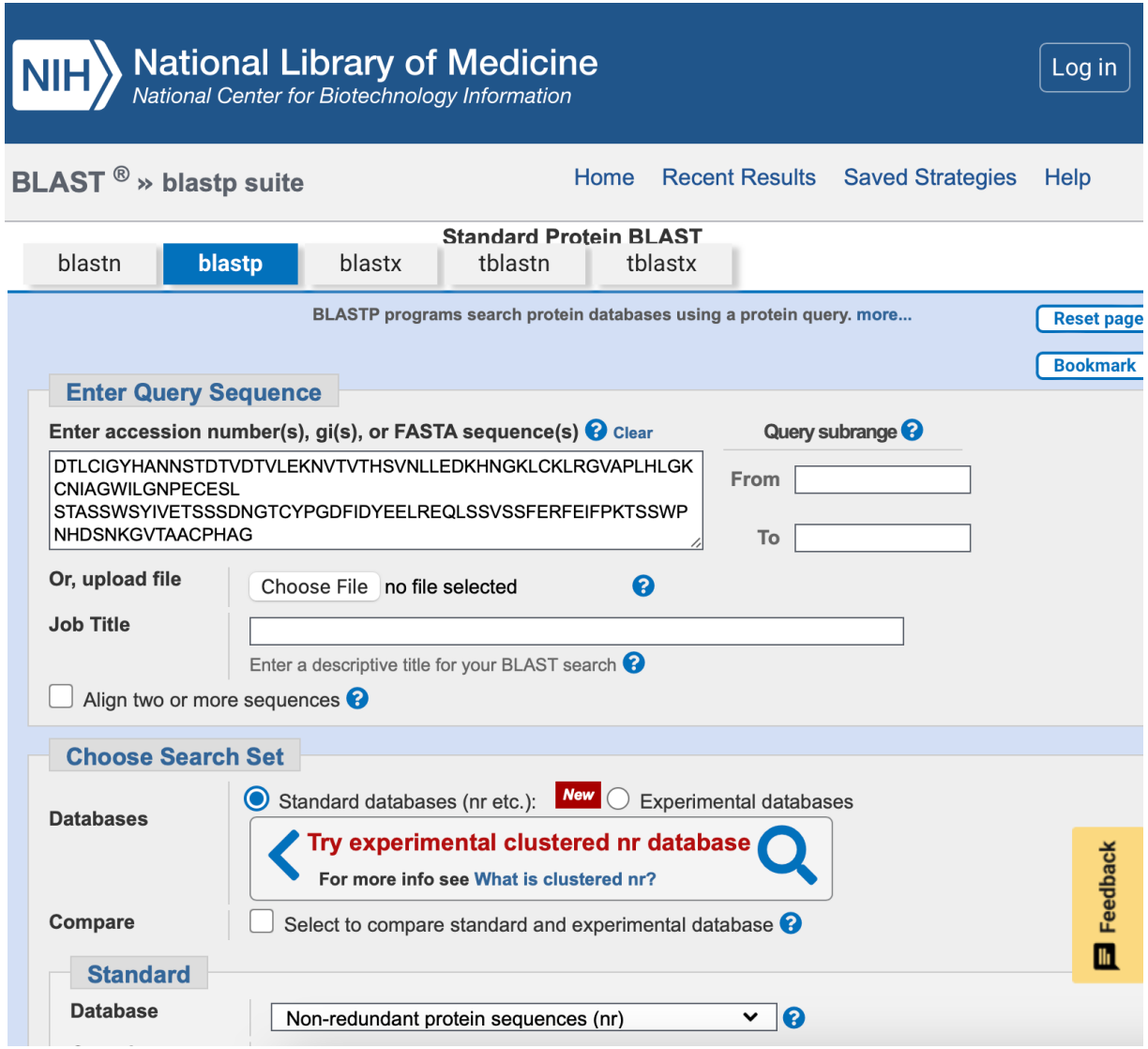


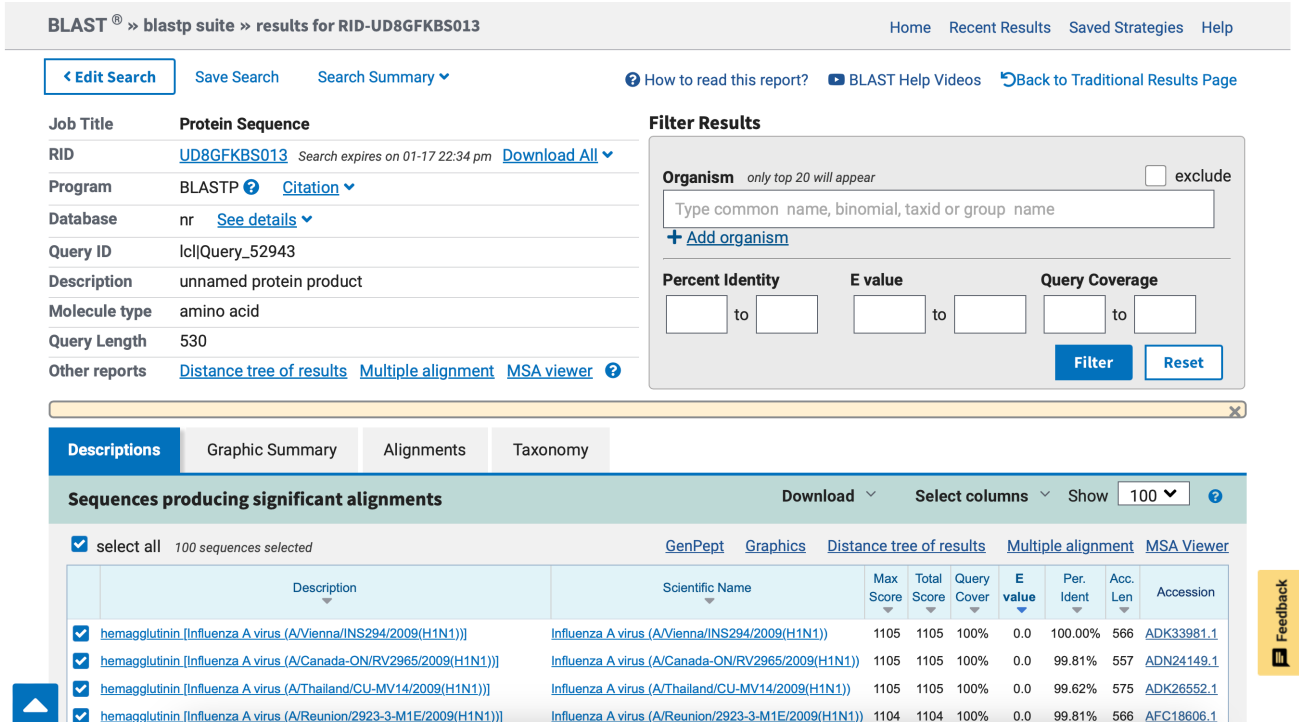
# Steps to revisit later



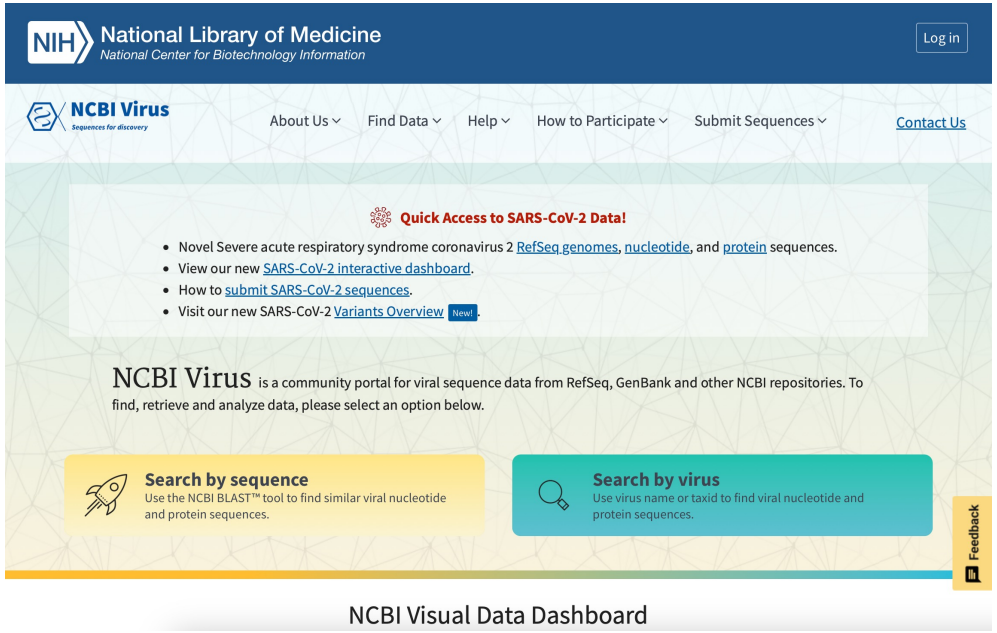

If you have a new sequence and need to figure out what it is…go to BLAST

BLASTn: nucleotide sequences
BLASTx: translates your nucleotide sequences to amino acid sequences
BLASTp: protein sequences

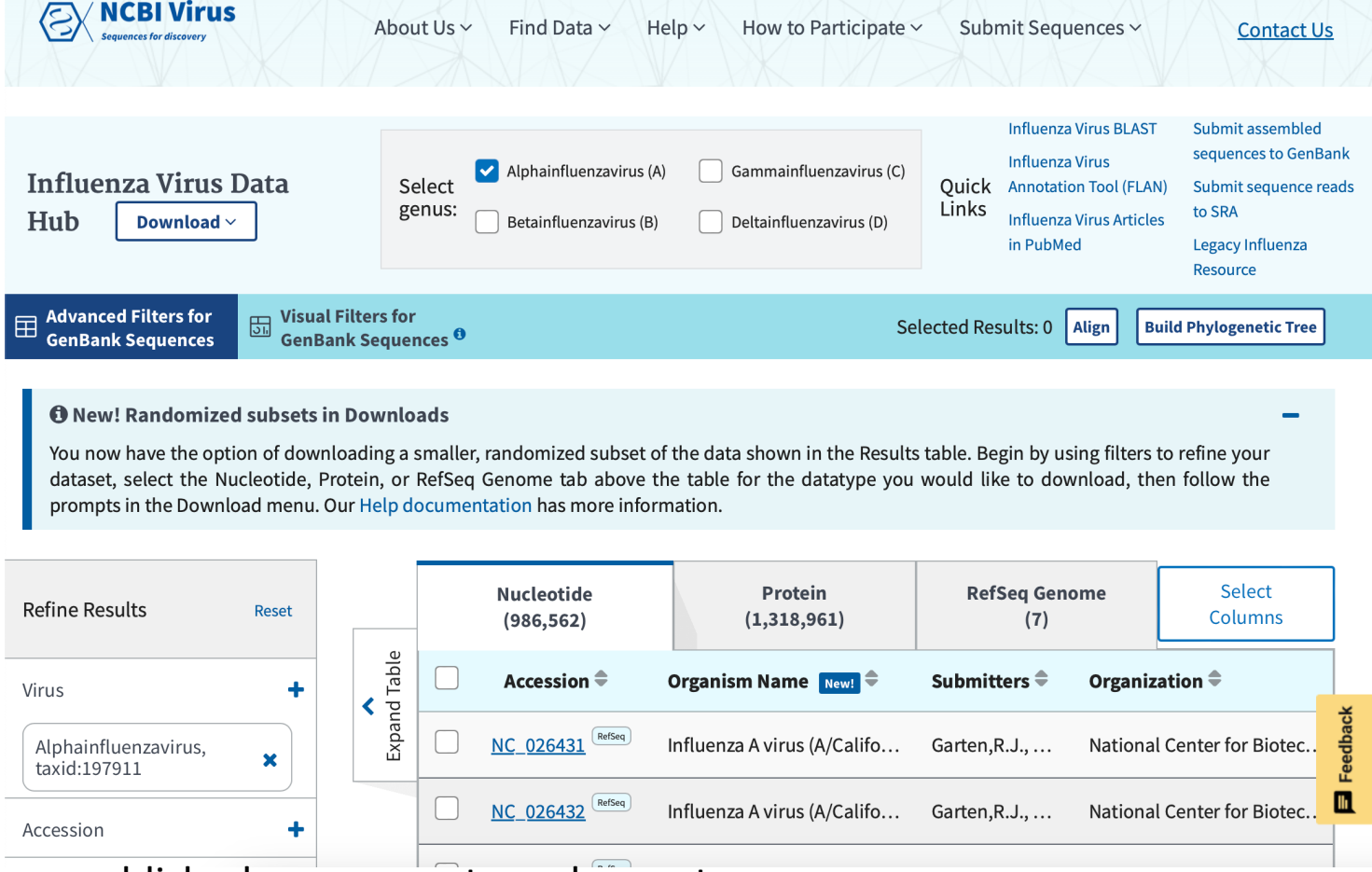


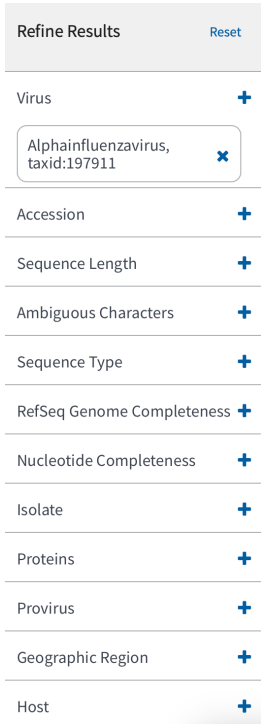




Okay! Our sequence is influenza A hemagglutinin…so we'll want to build a phylogeny using amino acid influenza A hemagglutinin sequences for reference




NCBI Visual Data Dashboard

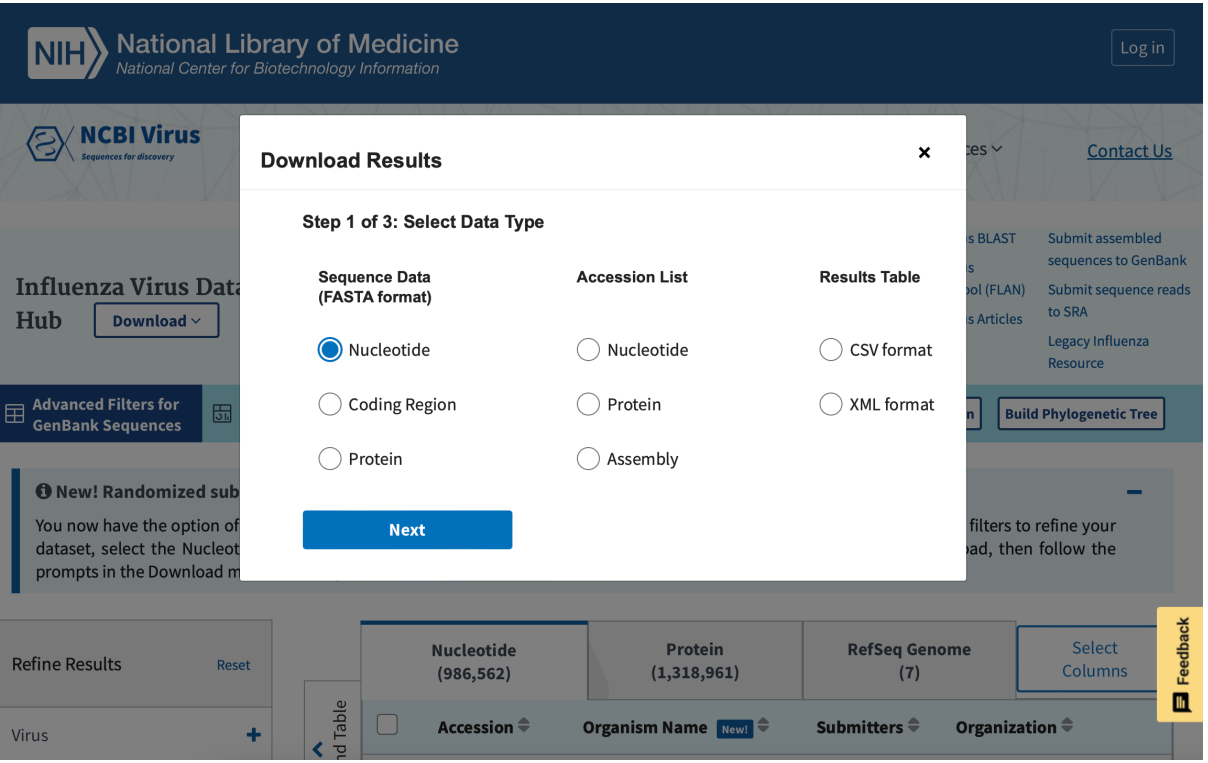



On NCBI virus, we can download some published sequences to make our tree:

1. Click search by virus
2. Type in whatever you're looking for, influenzavirus 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

# Download options on NCBI virus





FASTA file of all your selected sequences

Sample of csv file that you can download instead of the 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 | Betainfluenz | Orthomyxovi | USA | Homo sapier | 2023 |

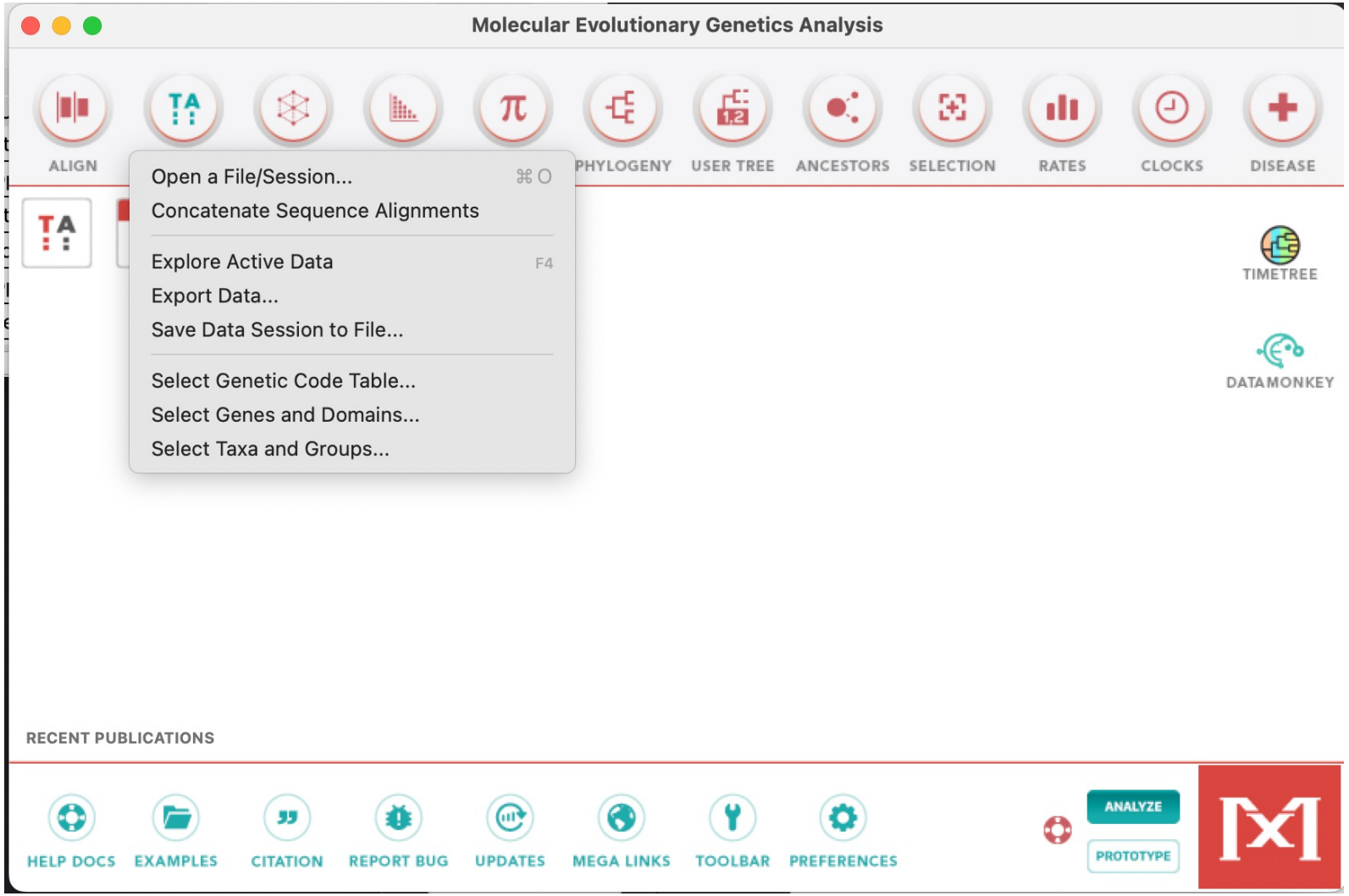
We need to pick an outgroup to root the tree...otherwise we can not determine how many changes in the genome led to the separation in the tree, an unrooted tree can not tell us which sequences descended from what

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

| 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 | |

| Family: | ***Orthomyxoviridae*** |
|---|---|
| **Genera** | |

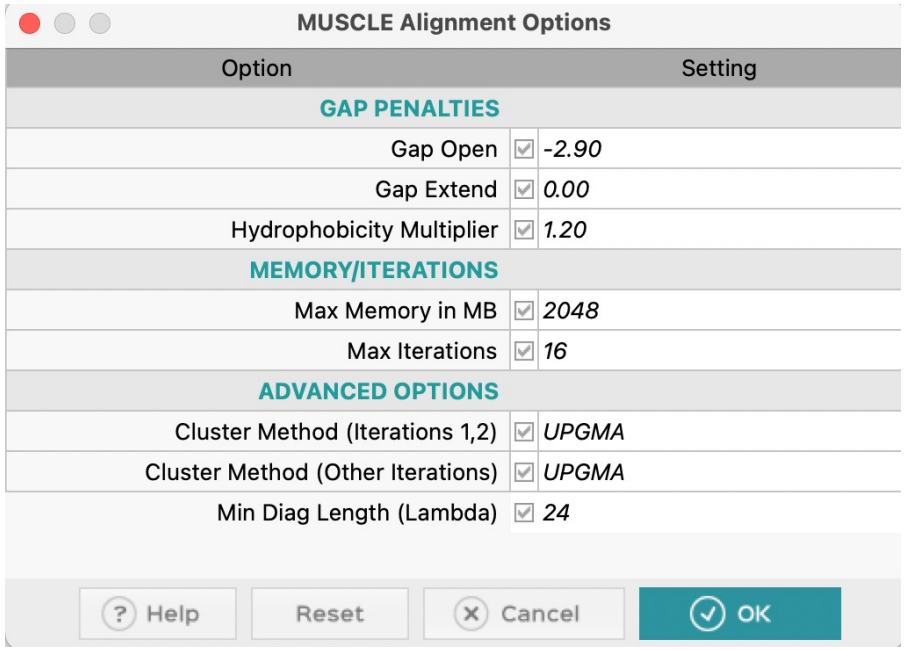
- *Alphainfluenzavirus*
- *Betainfluenzavirus*
- *Gammainfluenzavirus*
- *Deltainfluenzavirus*
- *Isavirus*
- *Quaranjavirus*
- *Thogotovirus*




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

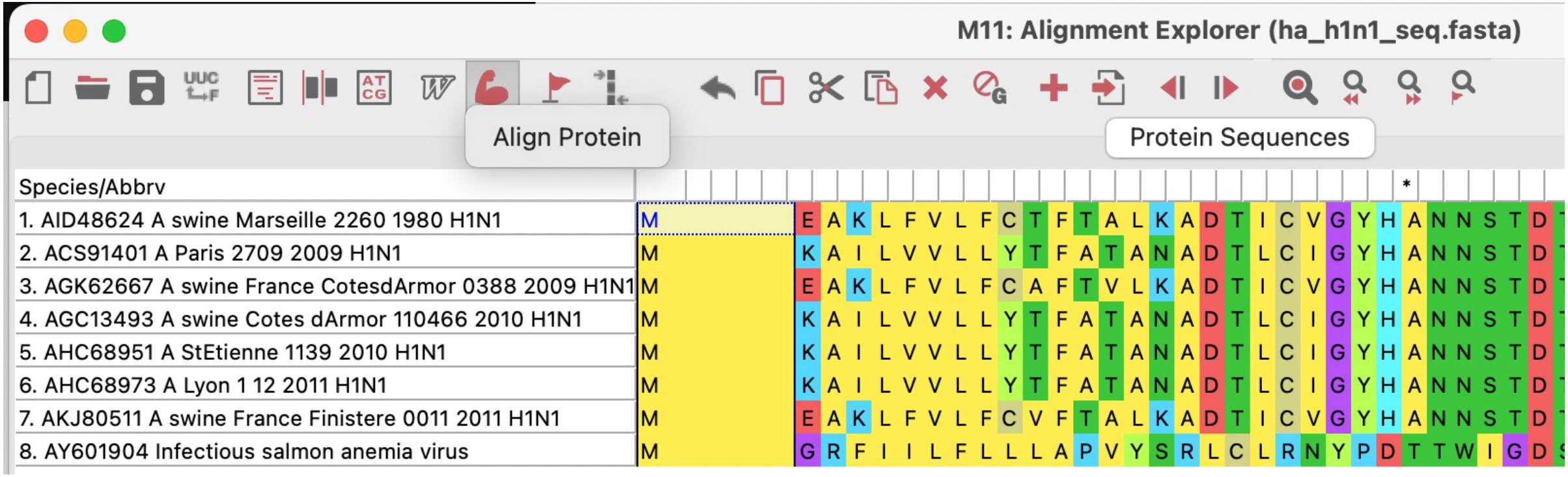
1. Open MEGA, and open a file/session, select your concatenated fasta file

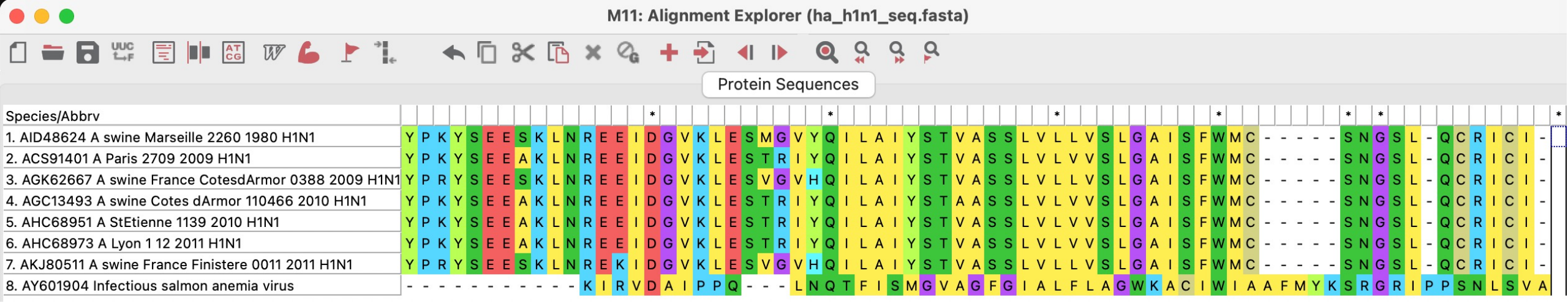
2. MEGA will ask if you want to align or analyze, click on align




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

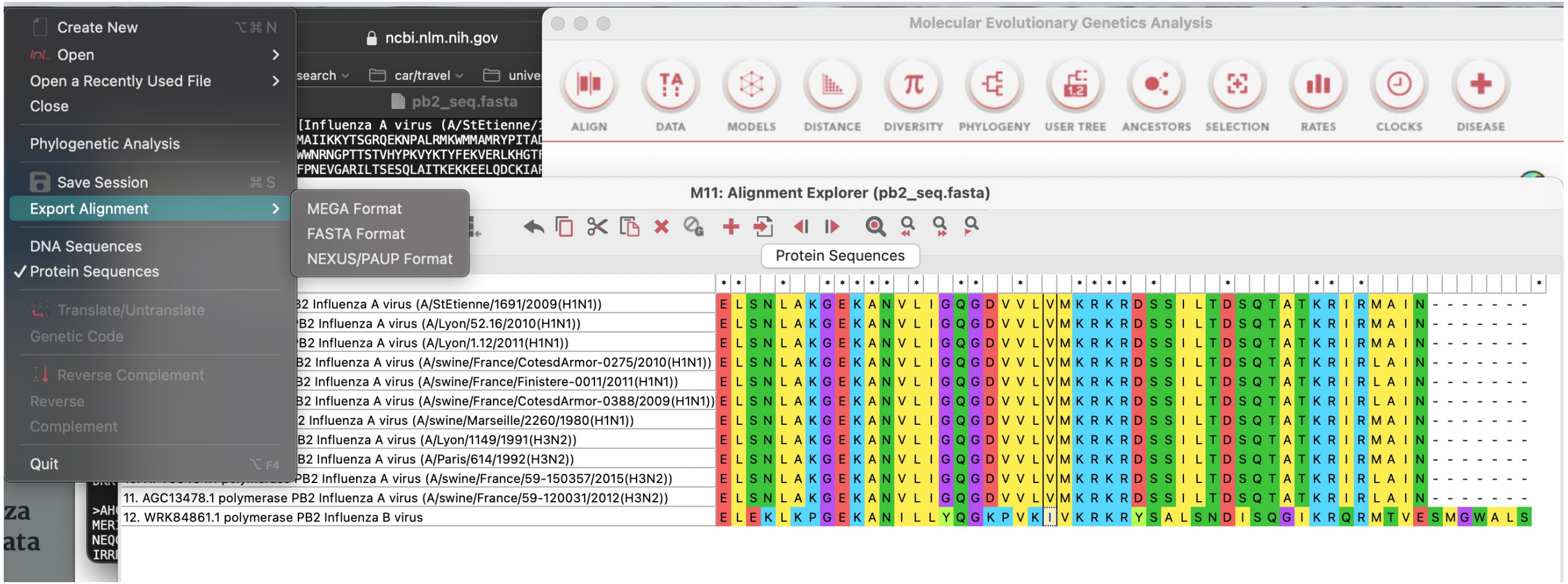
1. Click on the muscle arm to align with MUSCLE program
2. Leave all the preset options the same, click okay and click yes if it asks to select all sequences






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

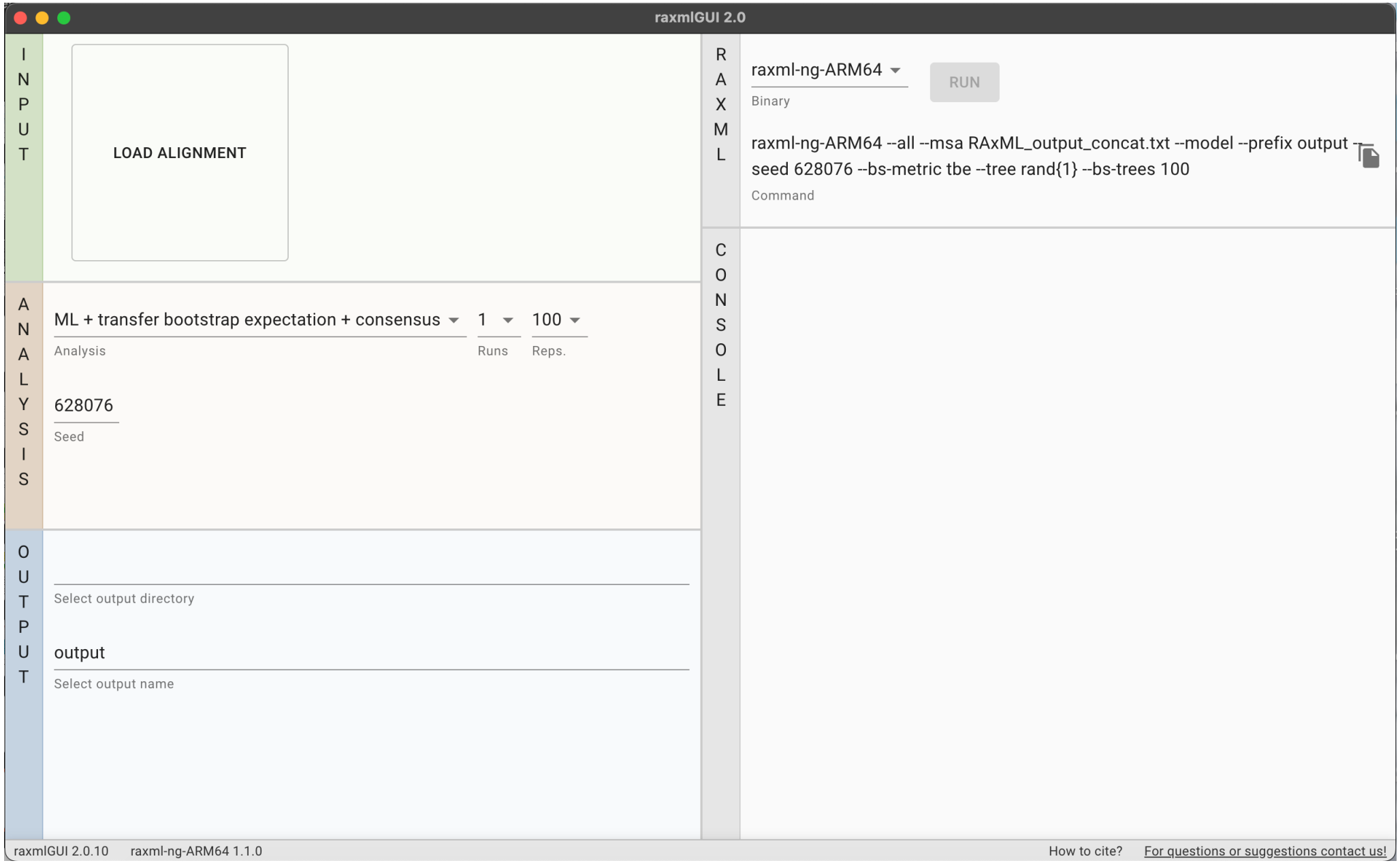
However you can also NOT do the trimming, depending on how much overhang there is it isn’t necessary…here there is not much overhang so I would not trim

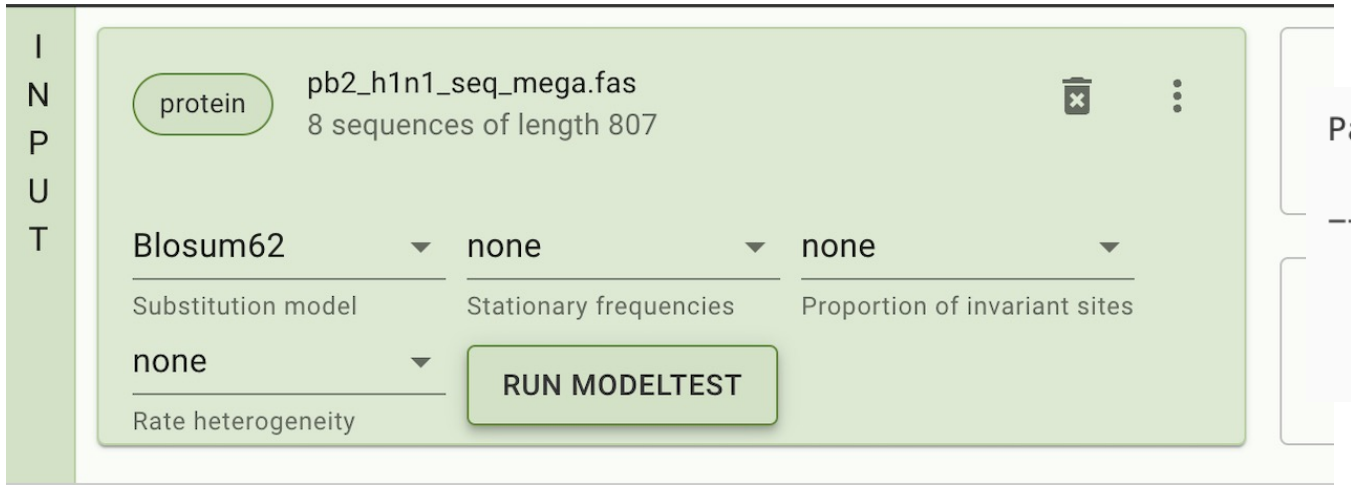



Save the aligned file in FASTA format, then we will proceed to model selection

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

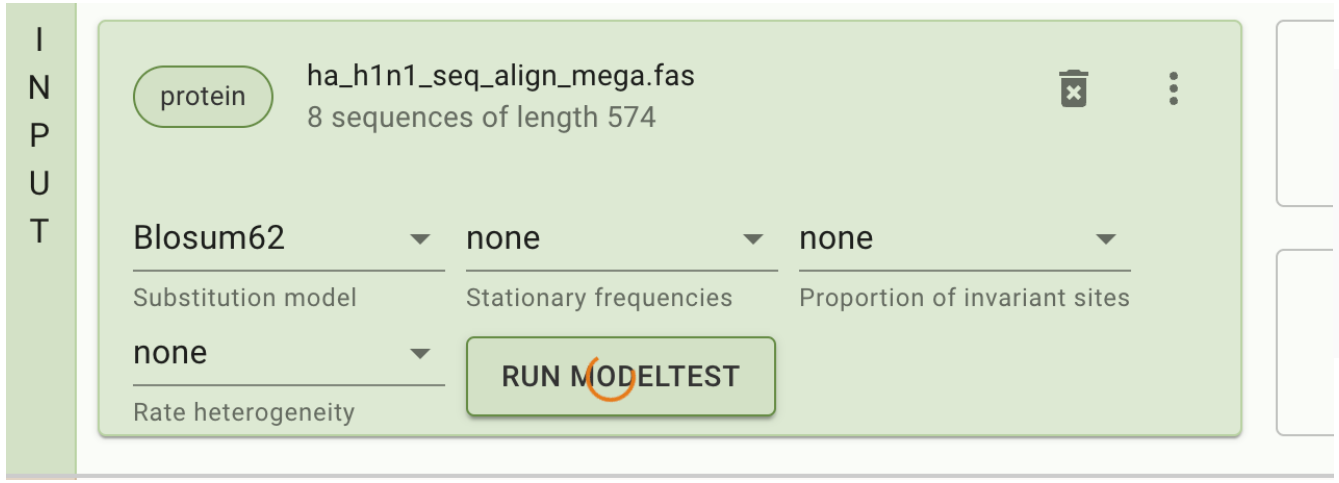
# Open up RAxML GUI






Partition 1/1:

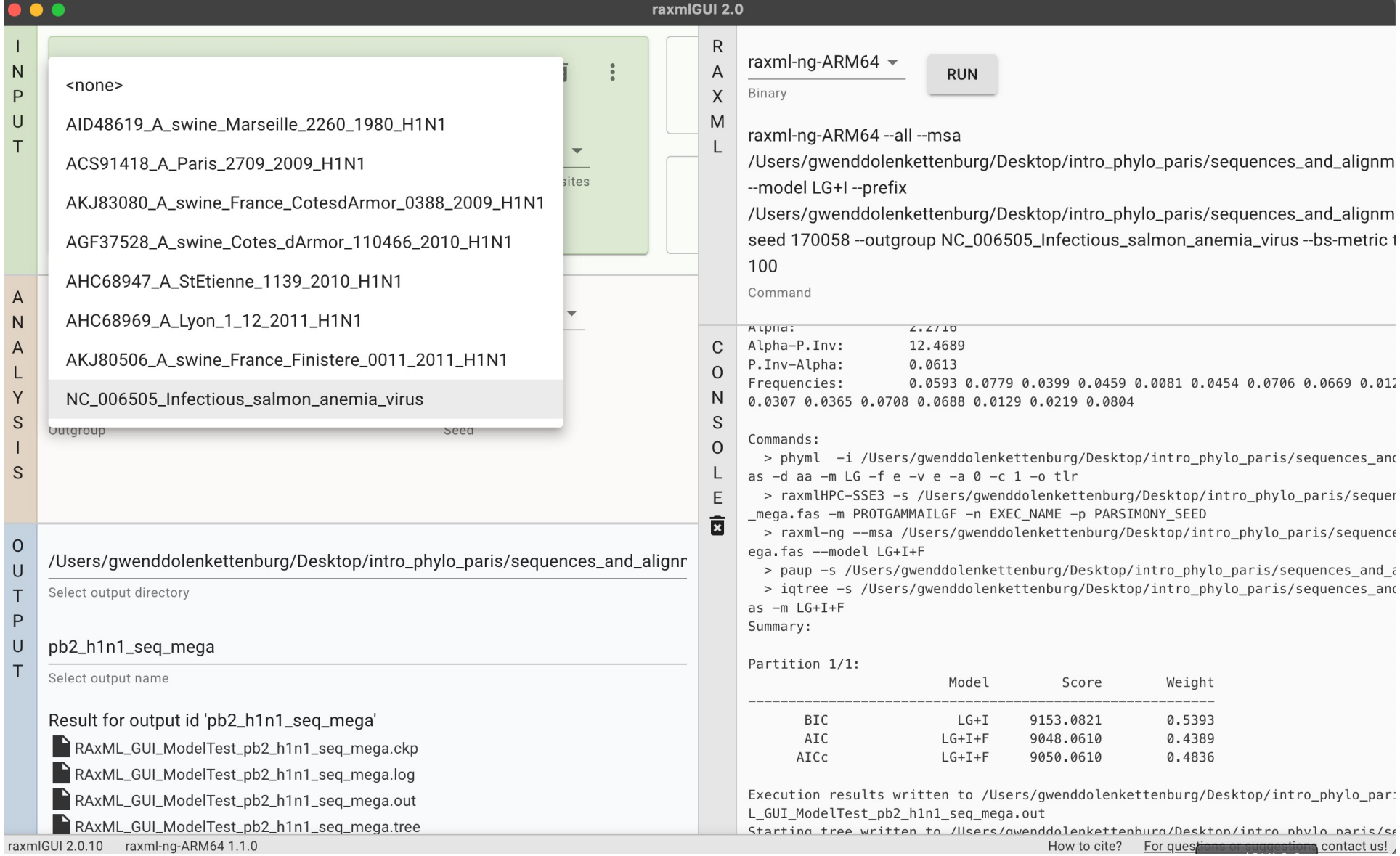
| | Model | Score | Weight |
|---|---|---|---|
| BIC | LG+I | 9153.0821 | 0.5393 |
| AIC | LG+I+F | 9048.0610 | 0.4389 |
| AICc | LG+I+F | 9050.0610 | 0.4836 |




Partition 1/1:

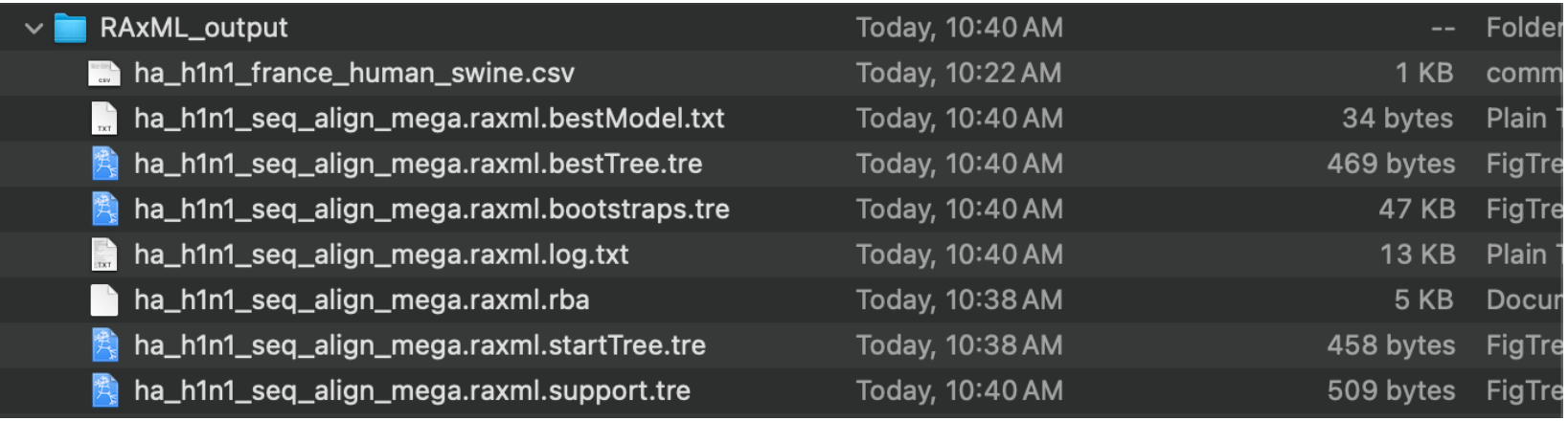
| | Model | Score | Weight |
|---|---|---|---|
| BIC | FLU+G4 | 7259.4956 | 0.6526 |
| AIC | FLU+G4 | 7198.5588 | 0.8274 |
| AICc | FLU+G4 | 7198.5588 | 0.8274 |

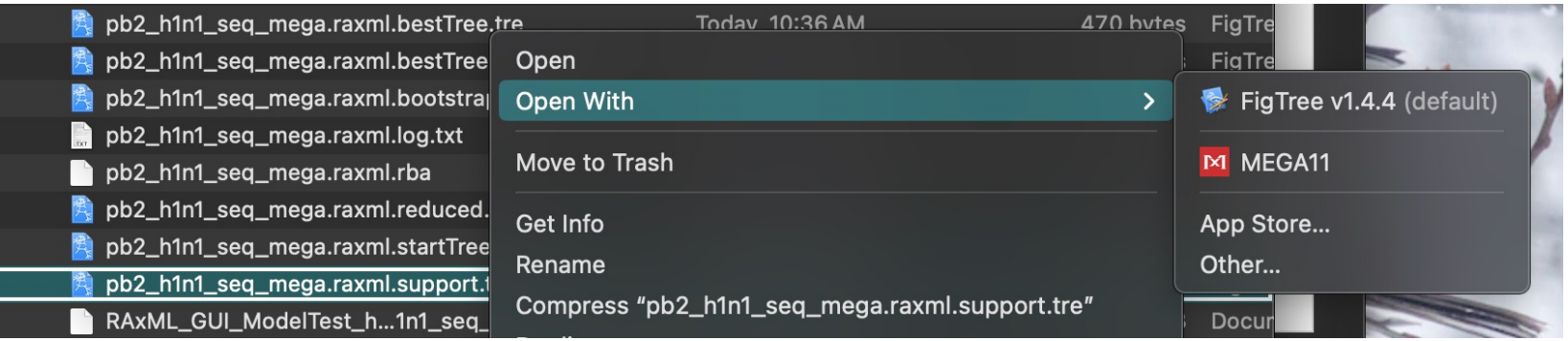
Load the aligned file and perform modeltest, for model selection go with the BIC score, it will spit out a report that saves to your files




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

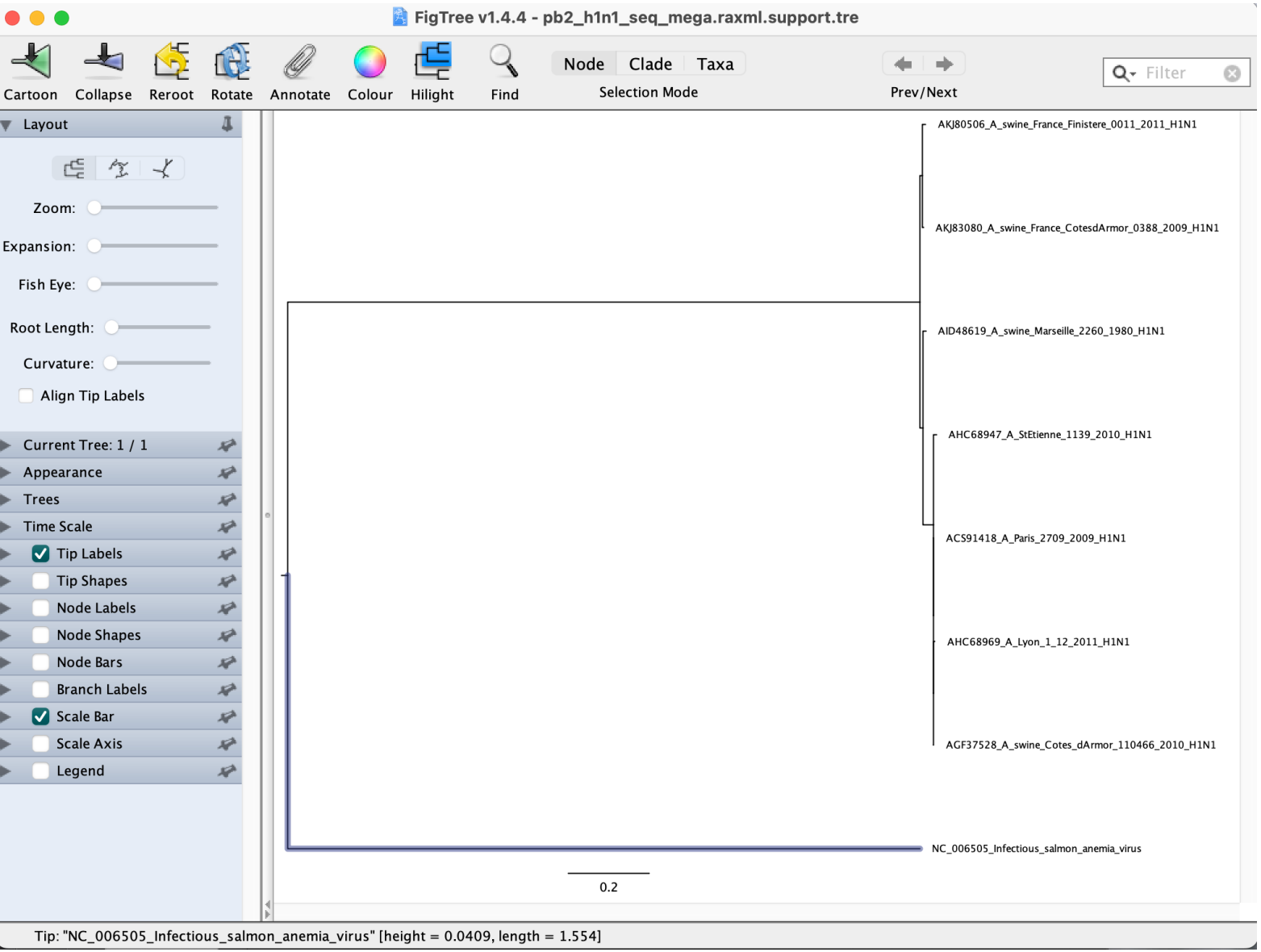
It will also spit out some files…

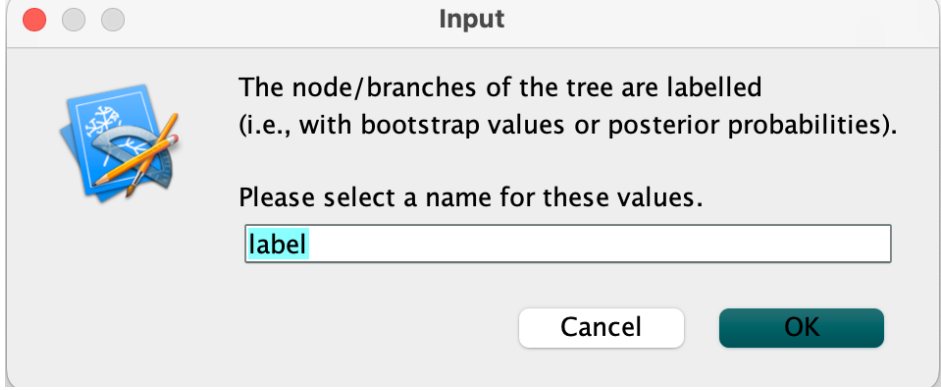



We're interested in the .support.tre file

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