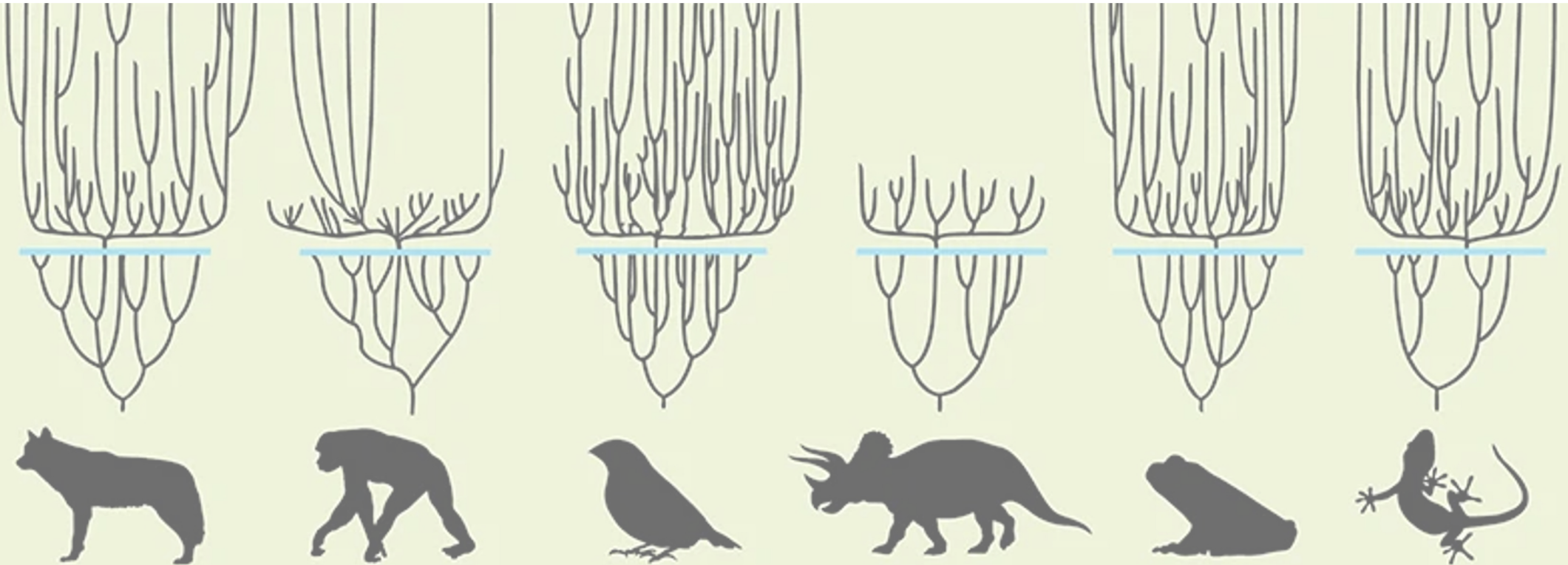


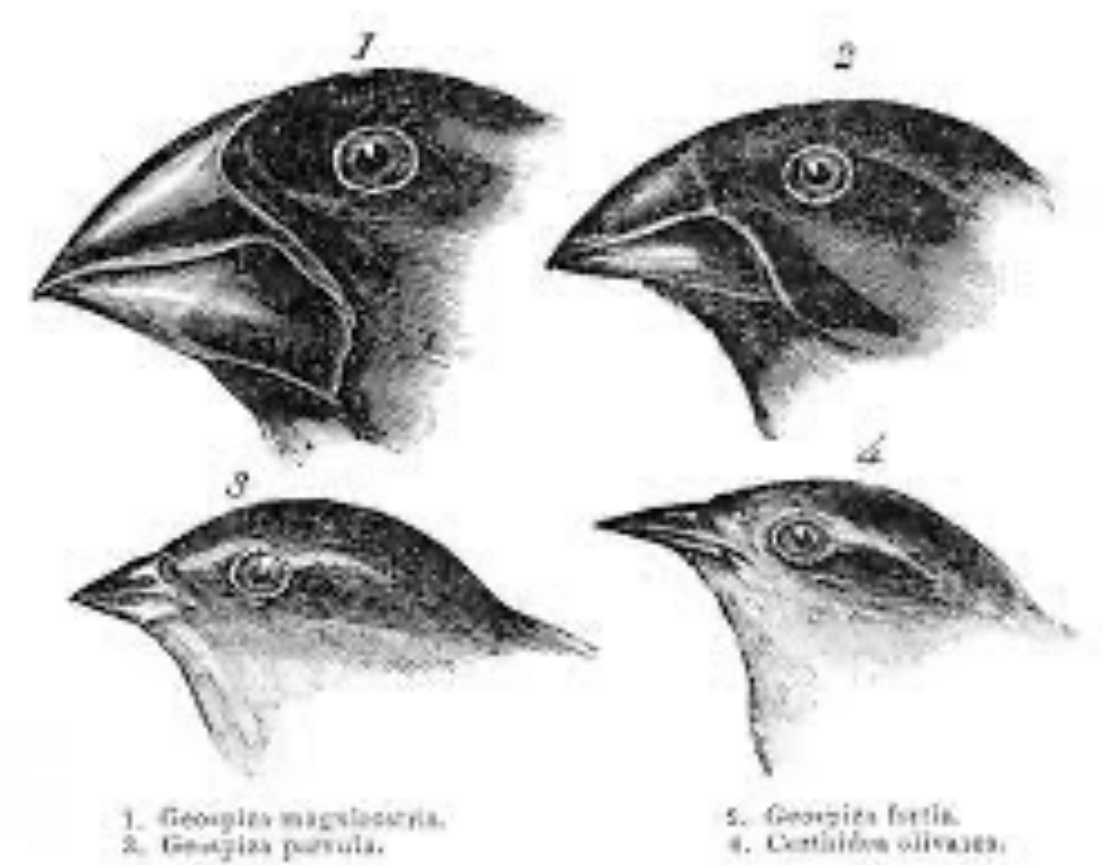
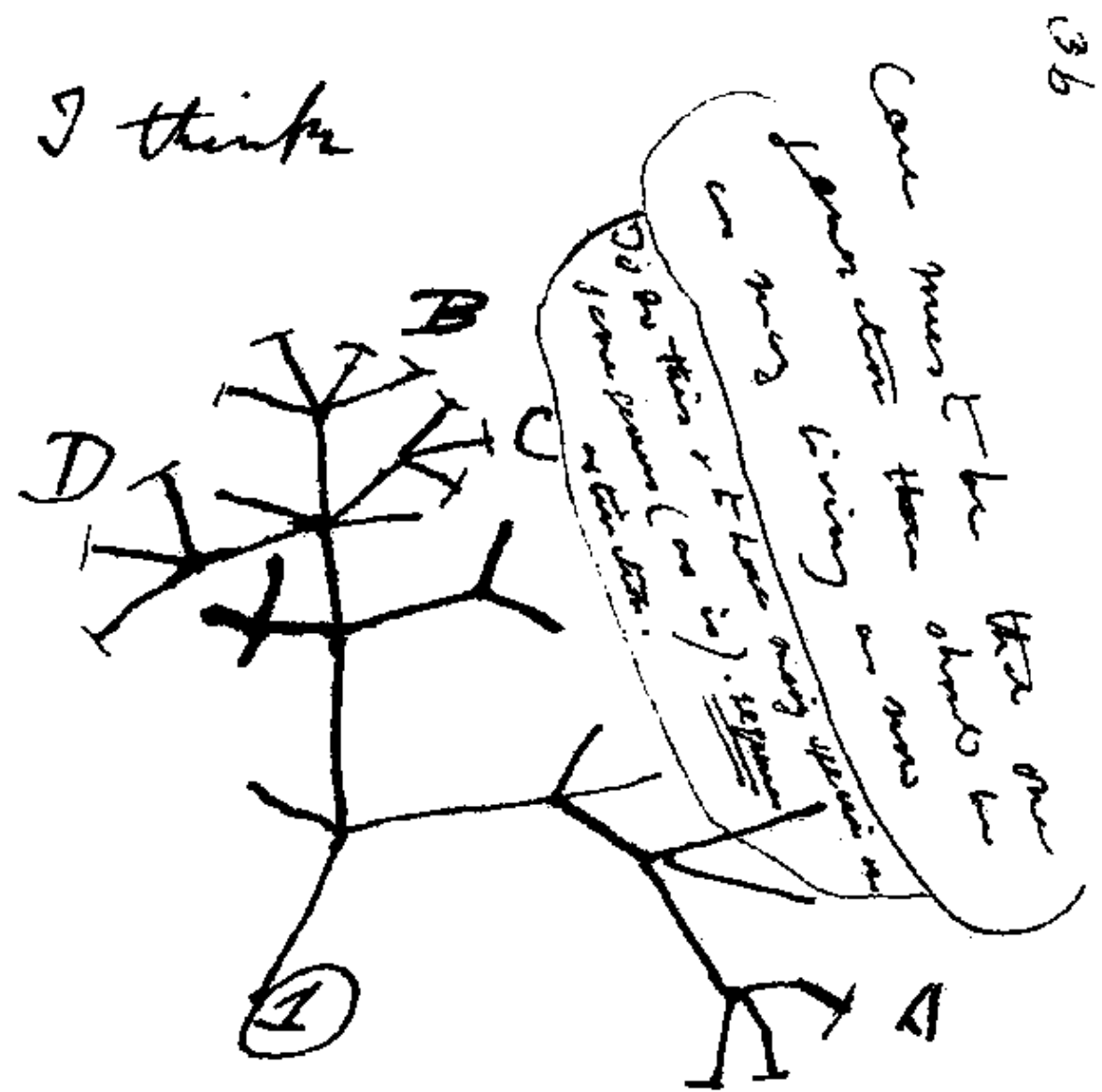
Introduction to phylogenetics

Gwen Kettenburg

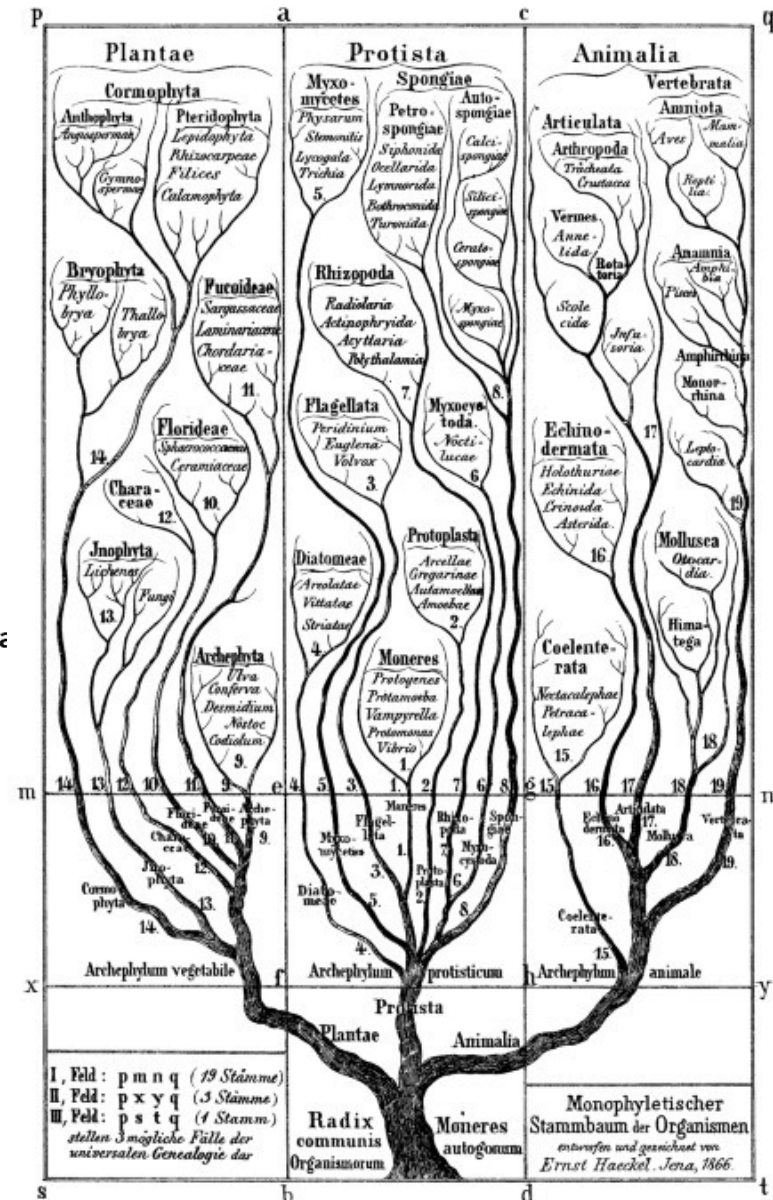
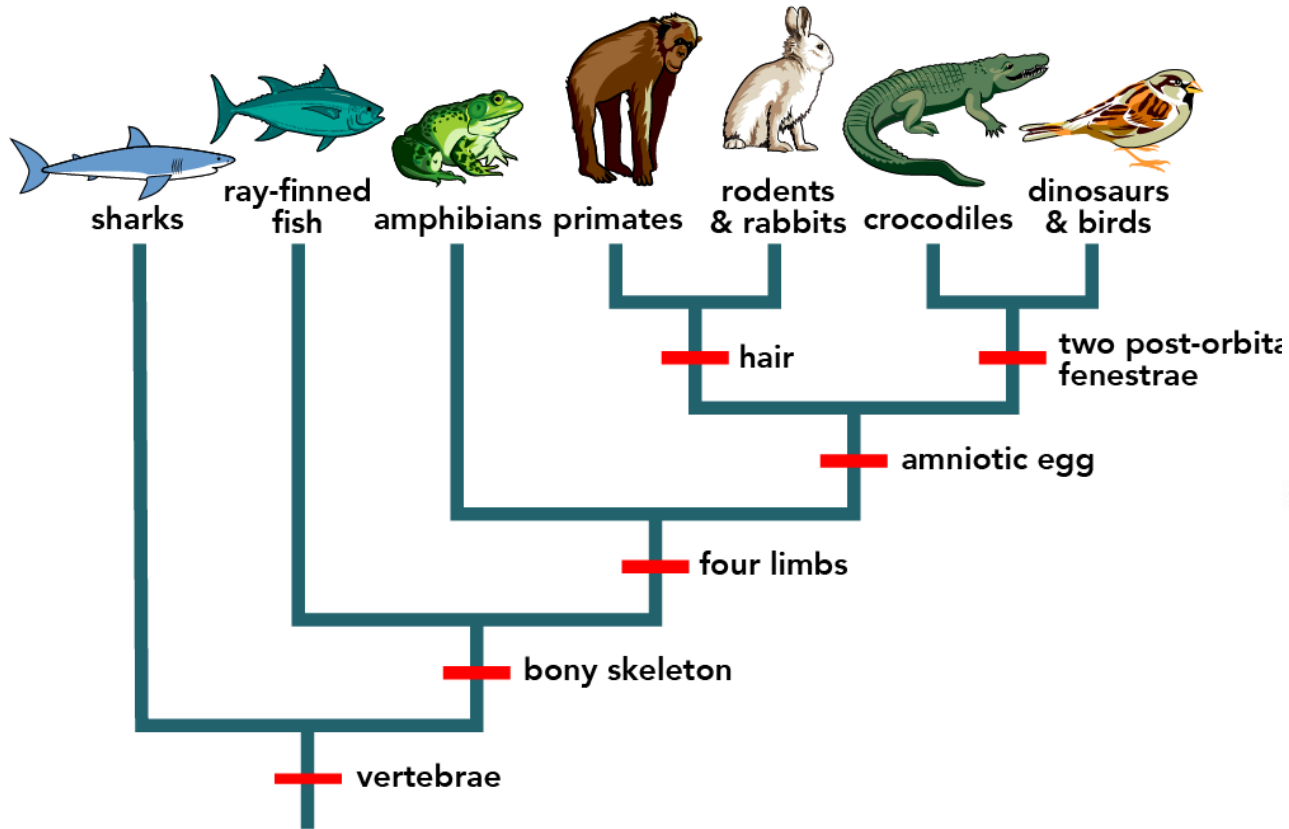
Slides adapted from Richard Ree and Andrew Hipp, University of Chicago



What is a phylogeny?

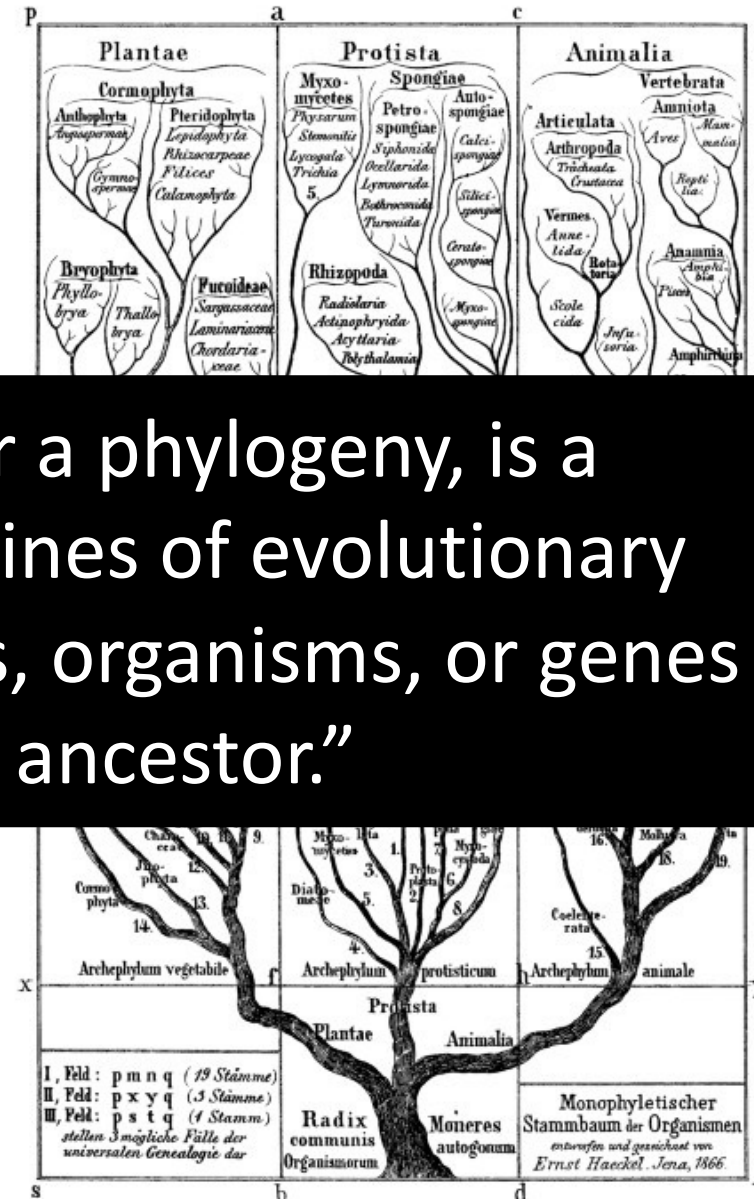
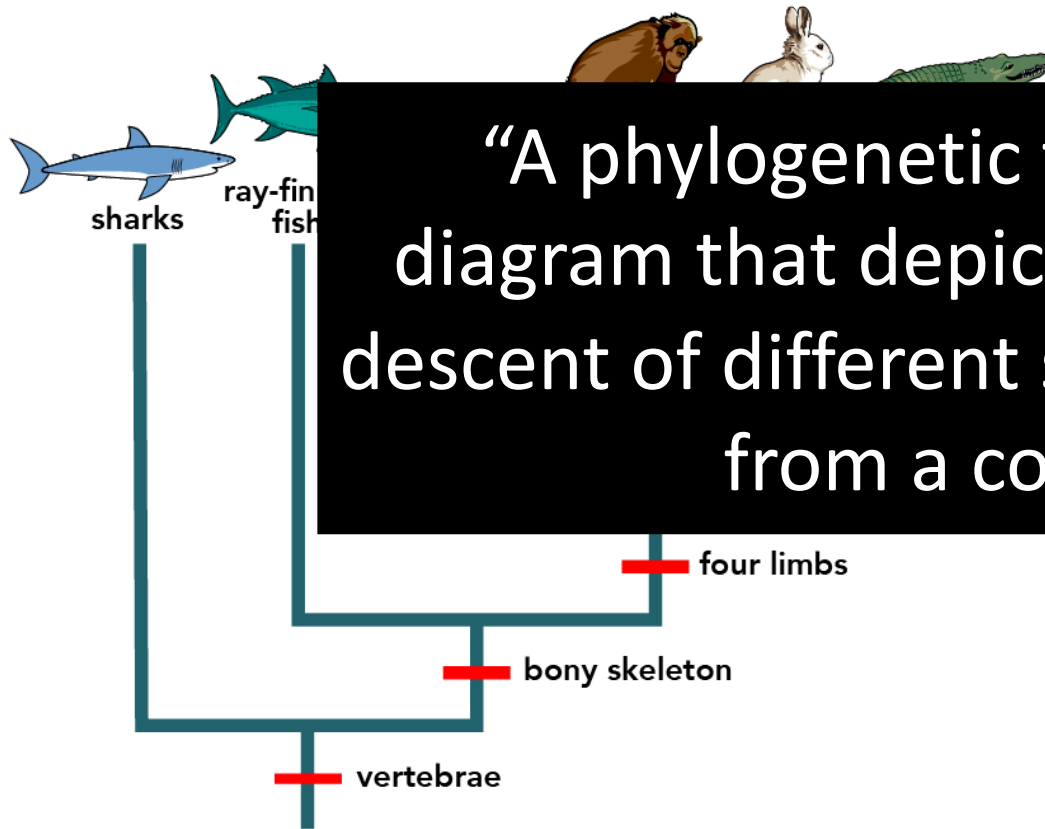


What is a phylogeny?



What is a phylogeny?

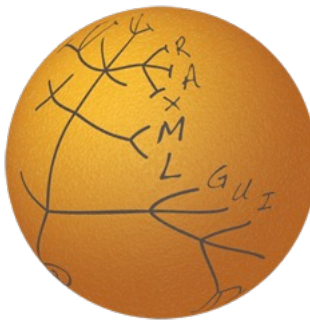
“A phylogenetic tree, or a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor.”



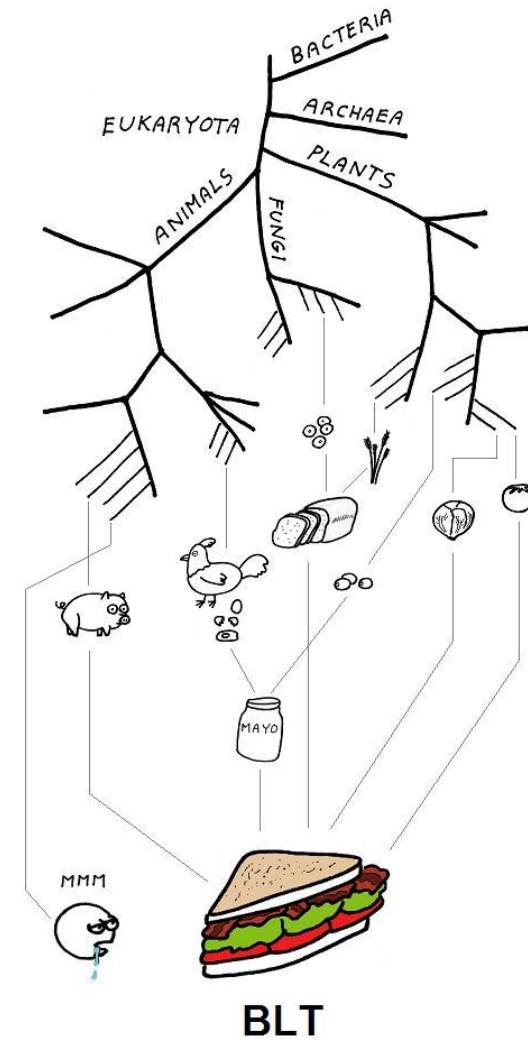
Baum et. al, Nature, 2008

Hossfeld and Levit, Nature, 2016

Goals:

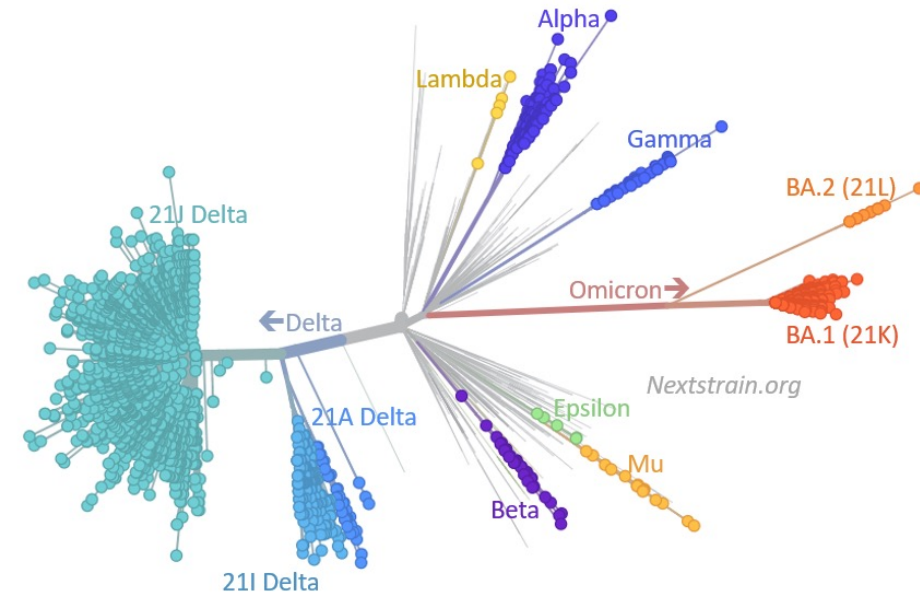
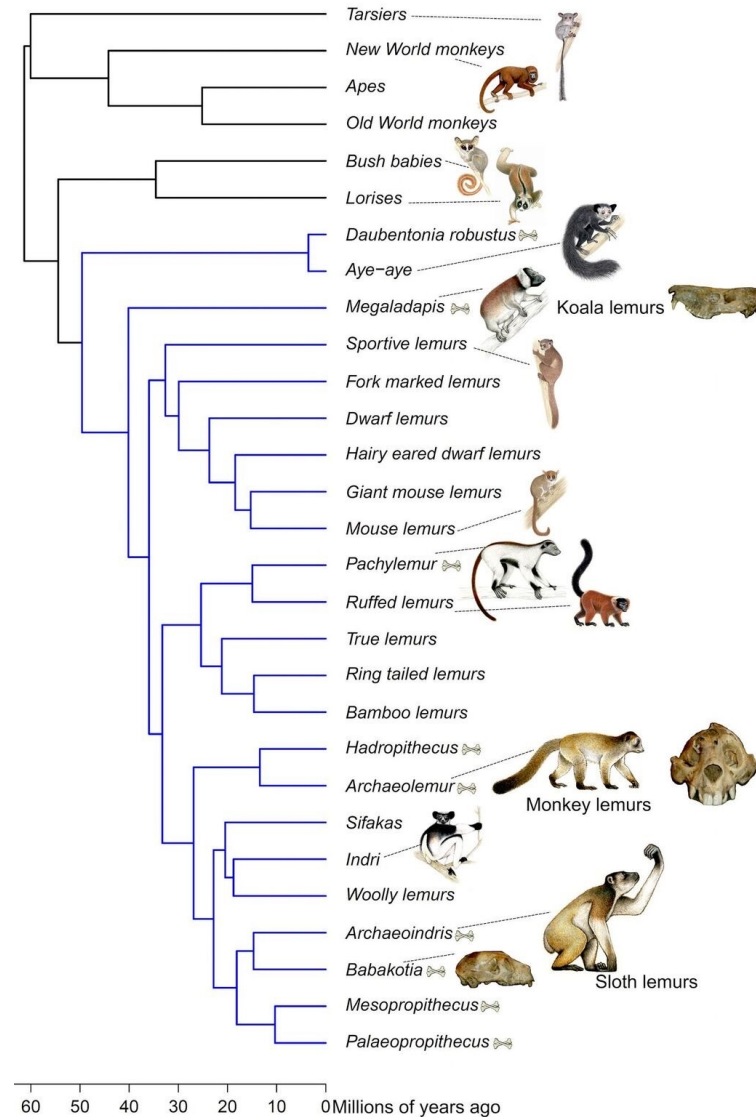
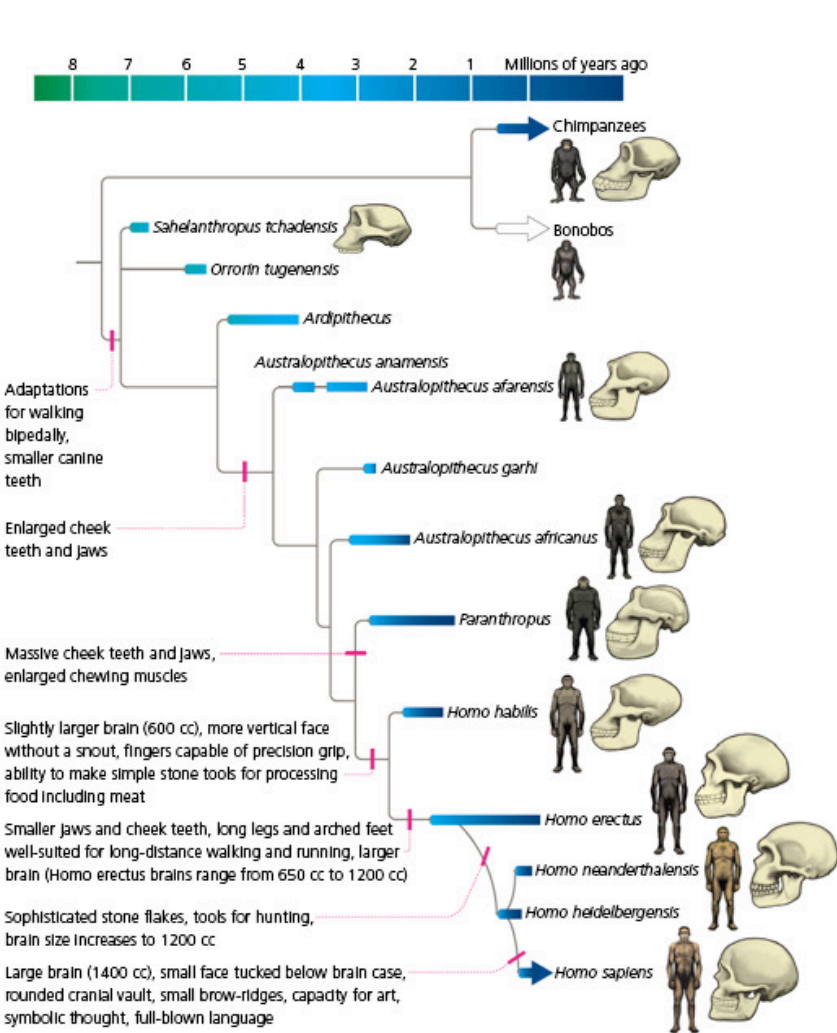


- Lecture component
 - Learn basics of what a phylogeny is
 - Learn how to read phylogenies
 - Basics of phylogenetic modeling
- Tutorial component
 - Learn how to make a phylogenetic tree from sequencing data
 - Using lemur cytochrome B protein sequences in RAxML software
 - Edit and visualize tree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

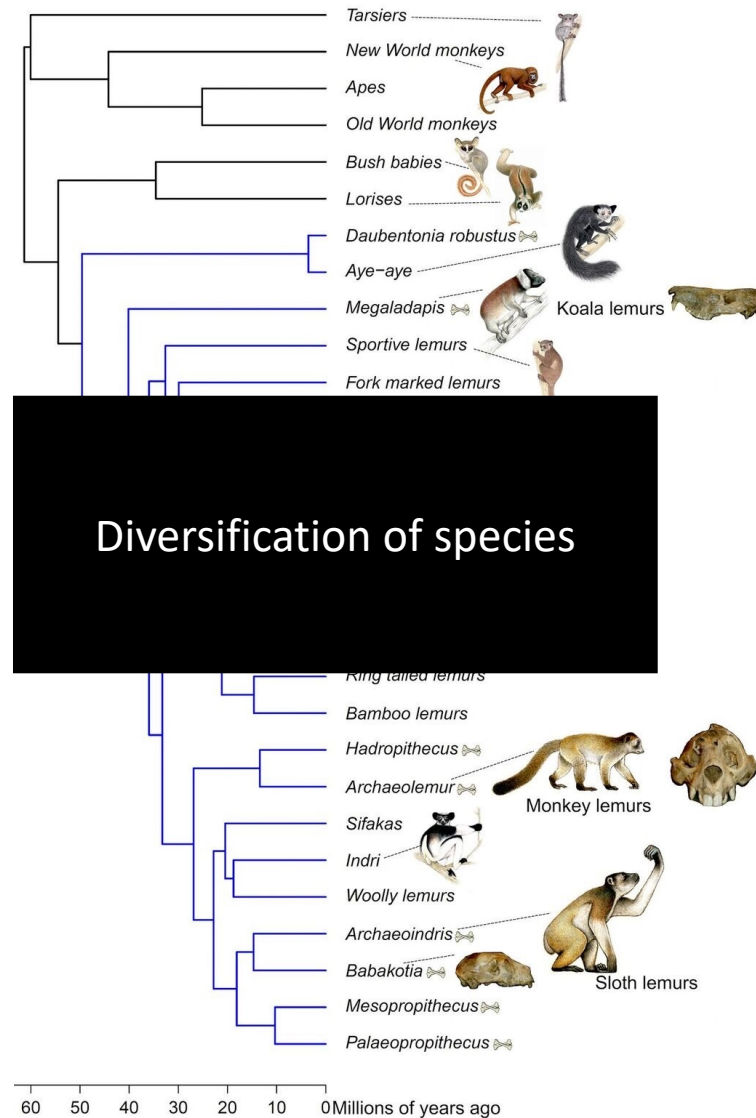
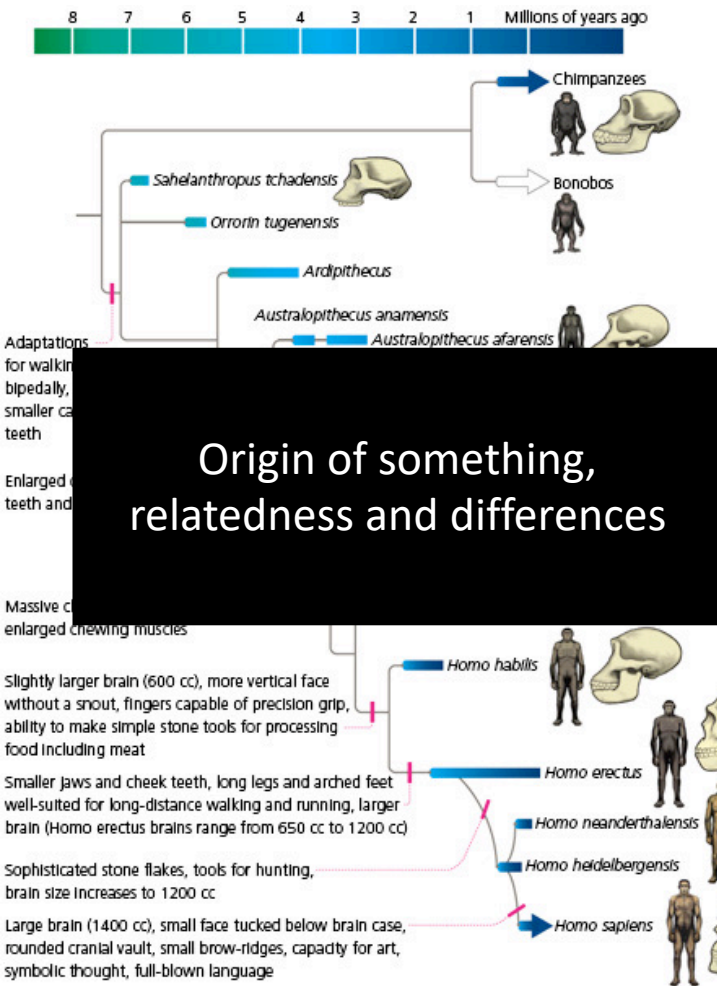


Molecular Evolutionary
Genetics Analysis

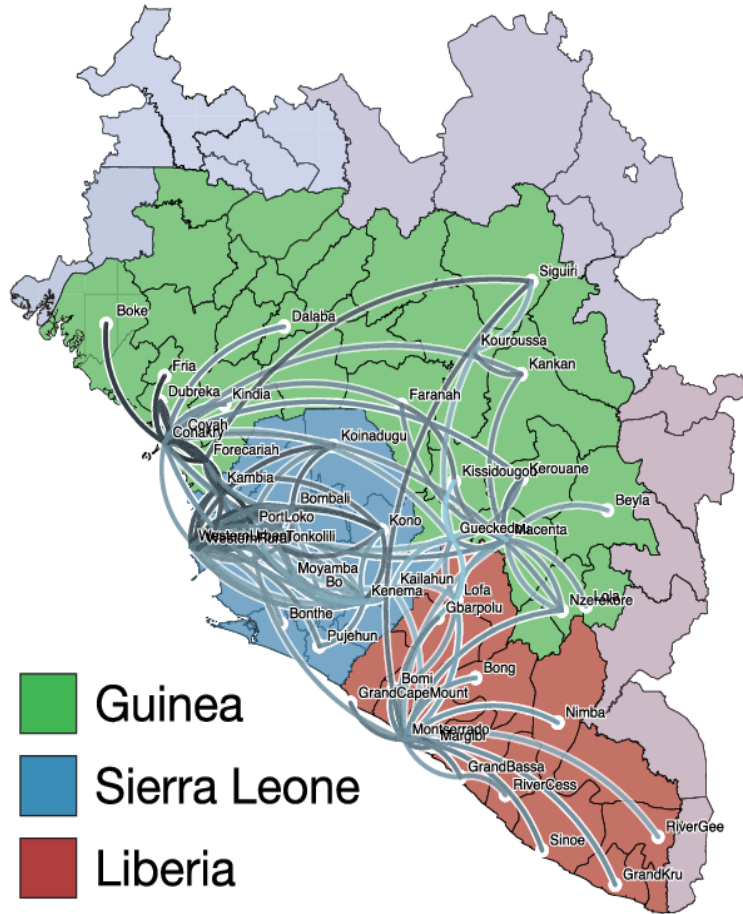
What can you do with phylogenies?



What can you do with phylogenies?

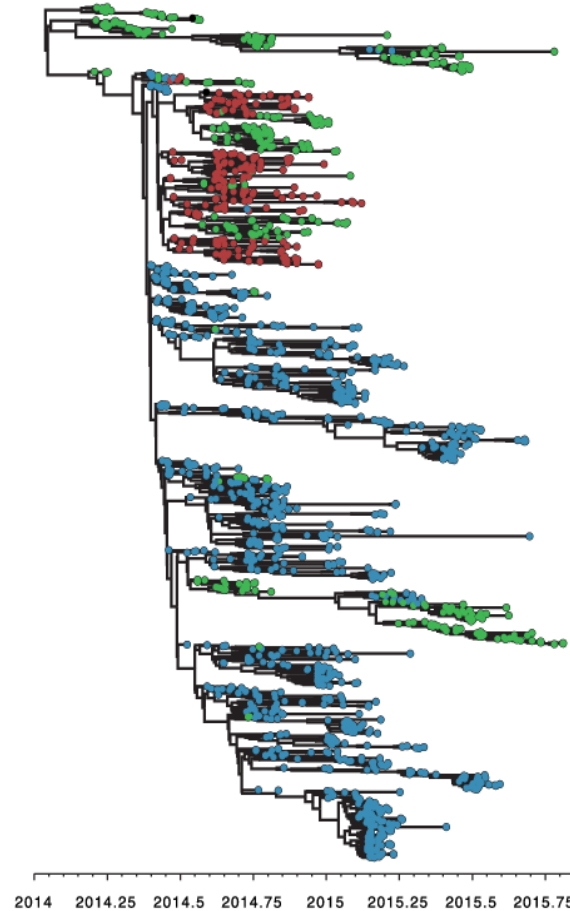


Phylodynamics



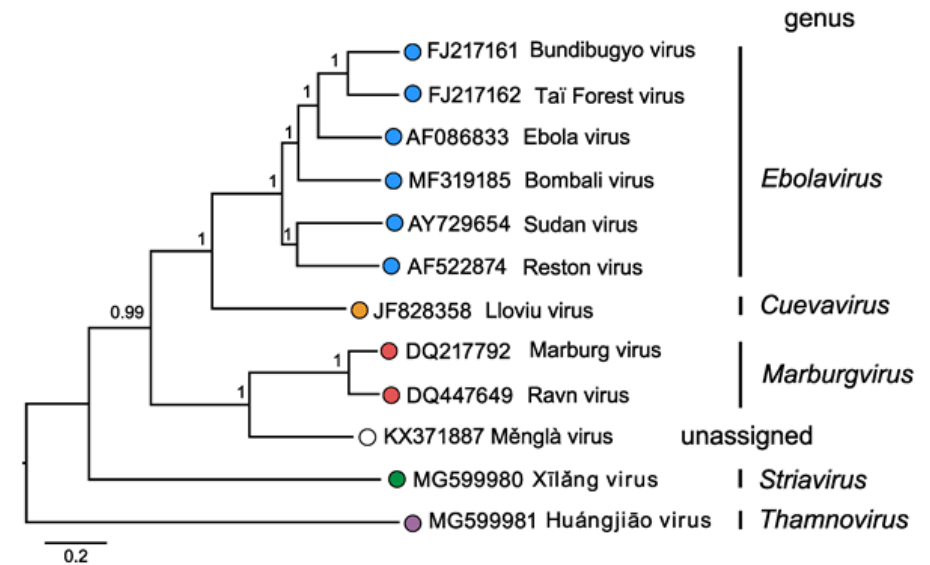
WHERE does it go?

Bayesian trees



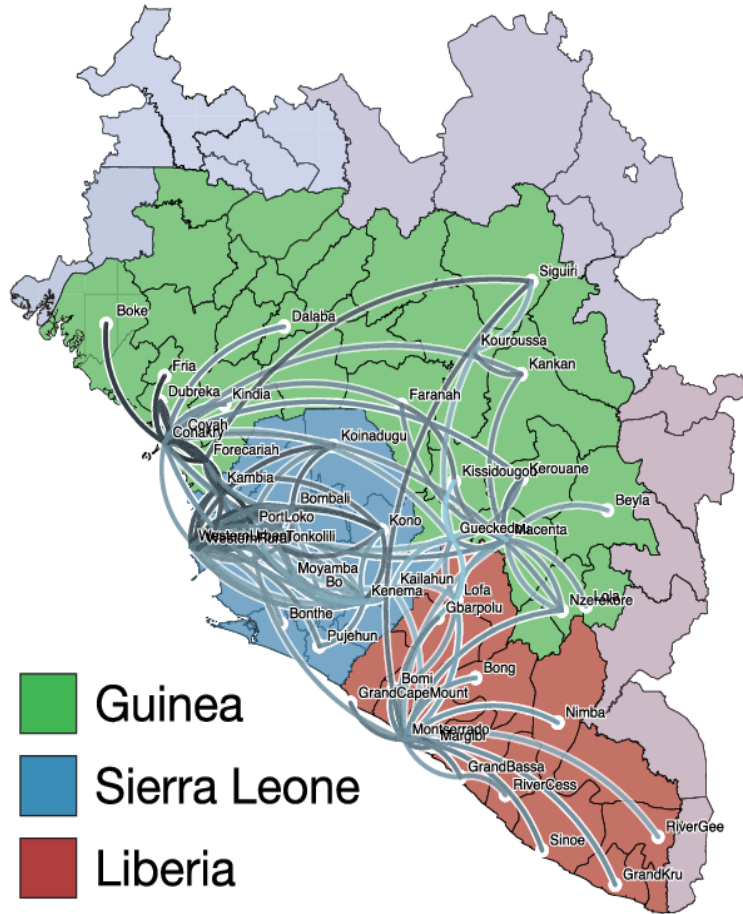
WHEN is the most recent common ancestor?

Maximum likelihood



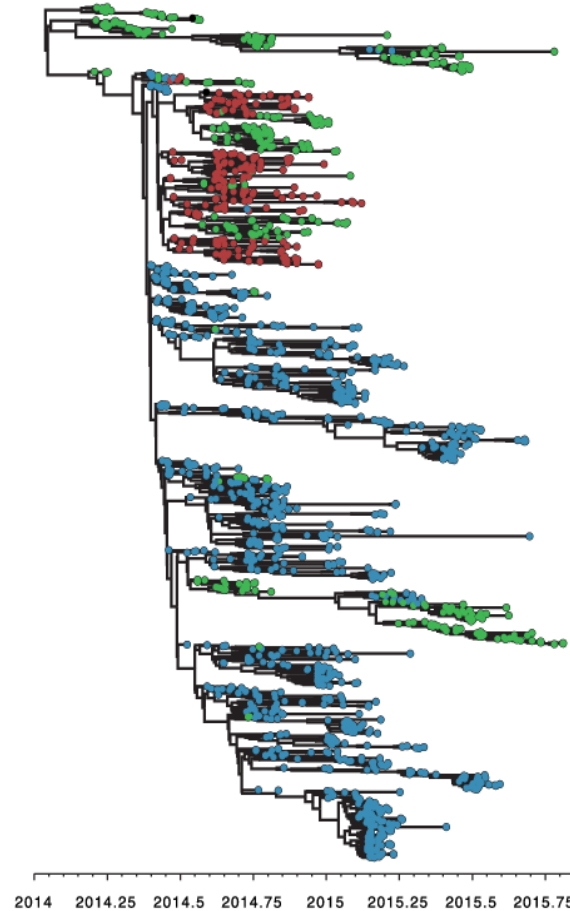
HOW different is it to what's known?

Phylodynamics



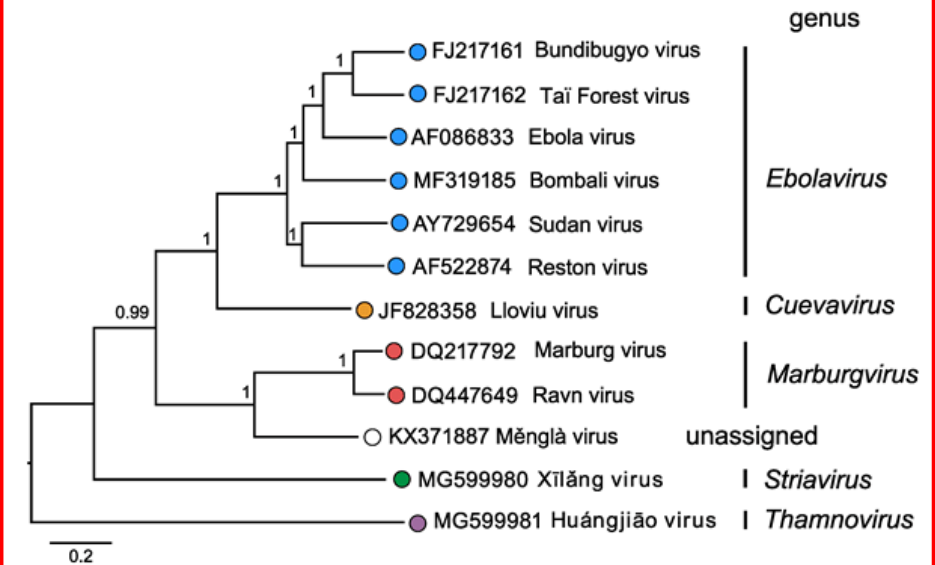
WHERE does it go?

Bayesian trees



WHEN is the most recent common ancestor?

Maximum likelihood

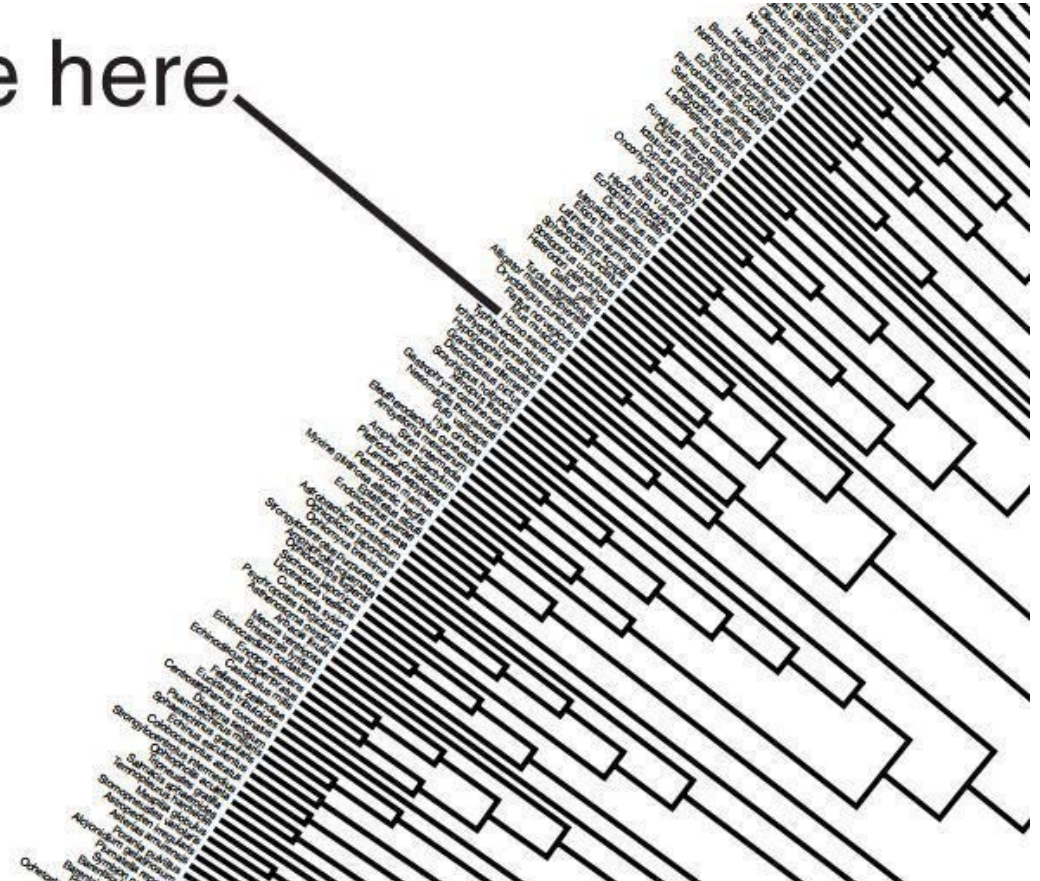


HOW different is it to what's known?

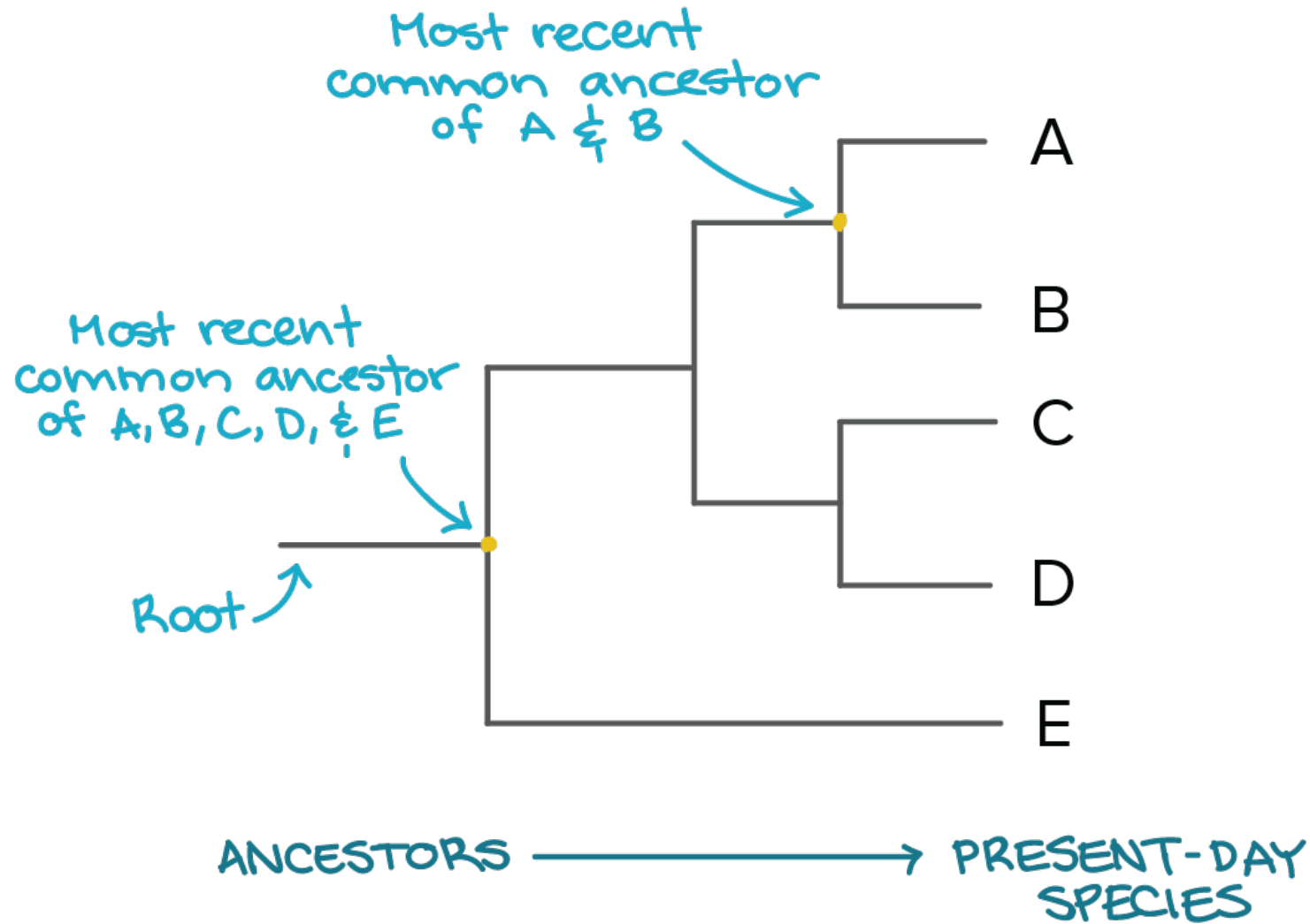
Checkpoint!

- What is a phylogeny?
- Can you use a phylogenetic analysis with time data?
- Can you use a phylogenetic analysis to see how similar something is to another?

You are here



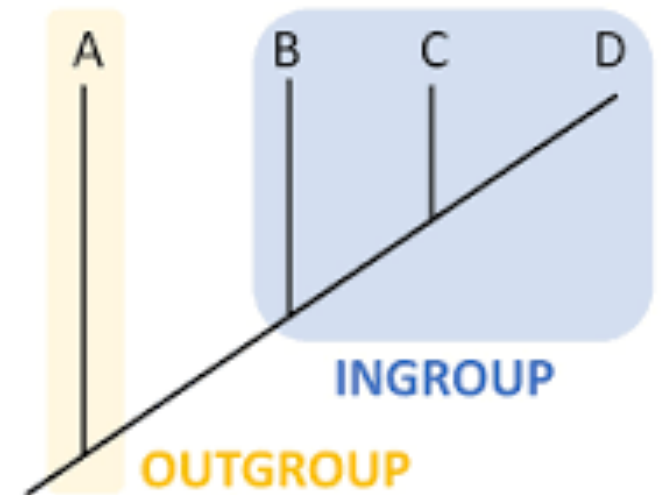
Anatomy of a phylogeny



CONFIDENCE

BOOTSTRAP VALUE

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



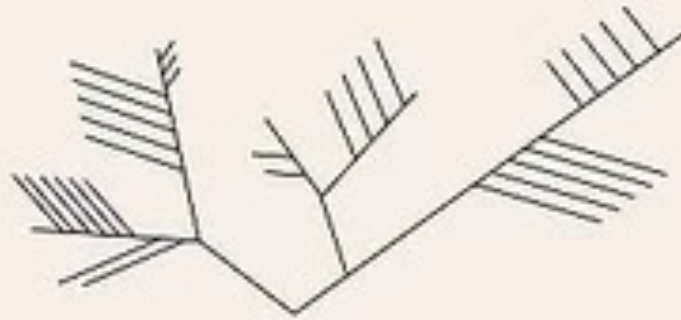
Cladogram versus phylogenetic tree

CLADOGRAM



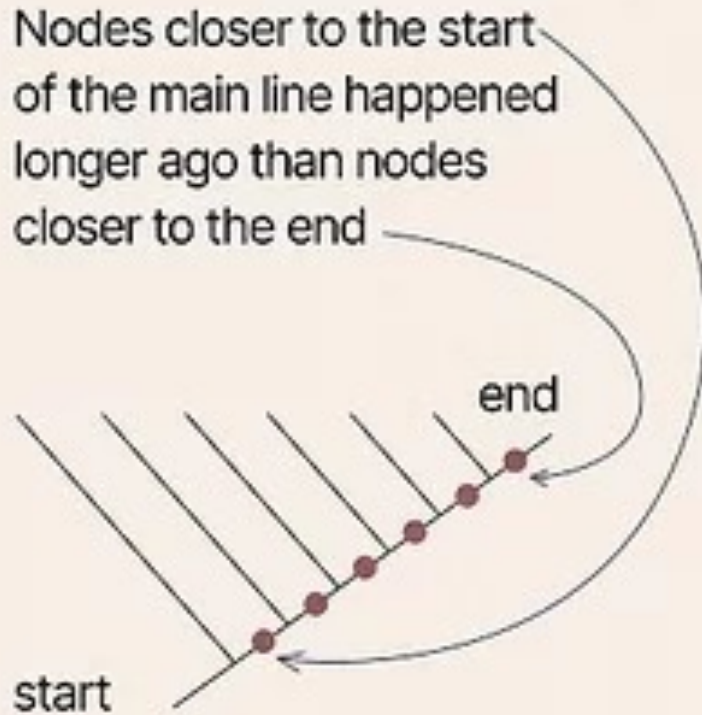
- the relationships are *hypothetical*
- you can easily make on your own

PHYLOGENETIC TREE



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

Cladogram versus phylogenetic tree



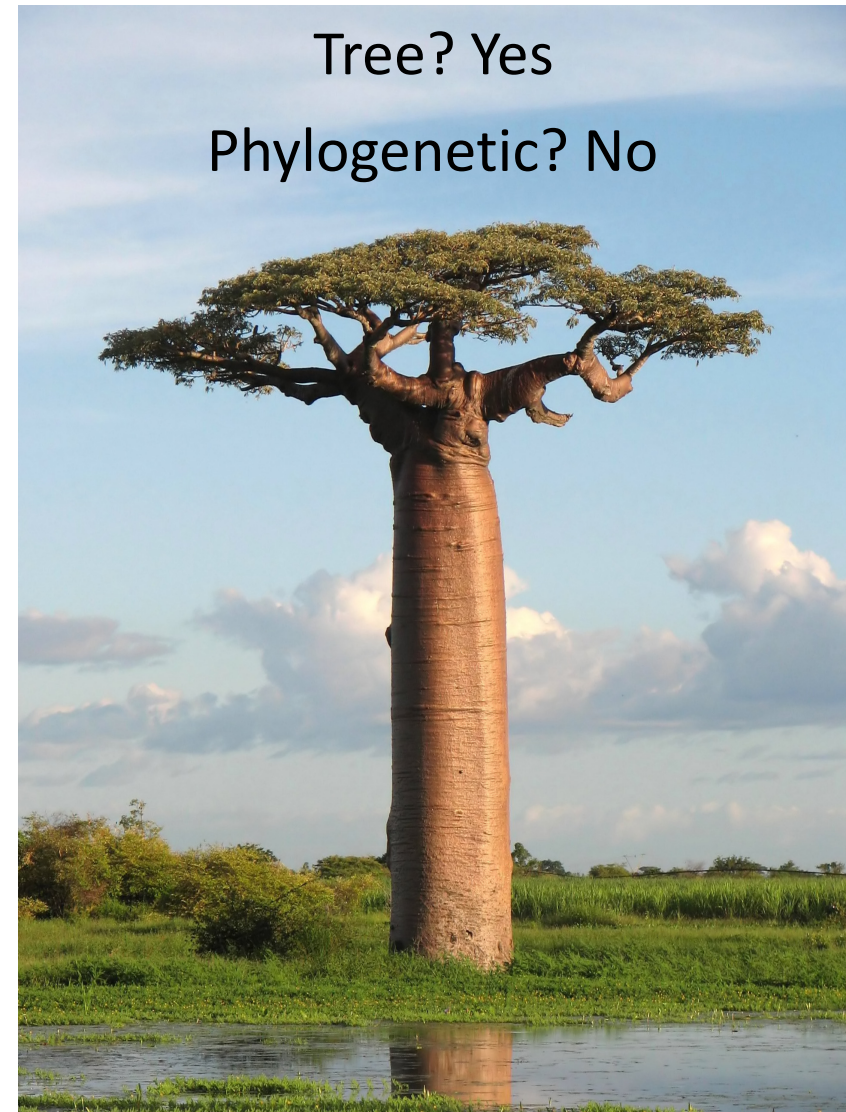
CLADOGRAM

Animals that are closer together are also more genetically similar

- Would you use a cladogram or phylogenetic tree for:**
- 1. Hypothesizing ancestors of humans**
 - 2. Tracking a new pathogen**

Checkpoint!

- Cladograms are good for generating hypotheses, phylogenies show genetic similarity
- What do bootstrap values show?
- What is a root?



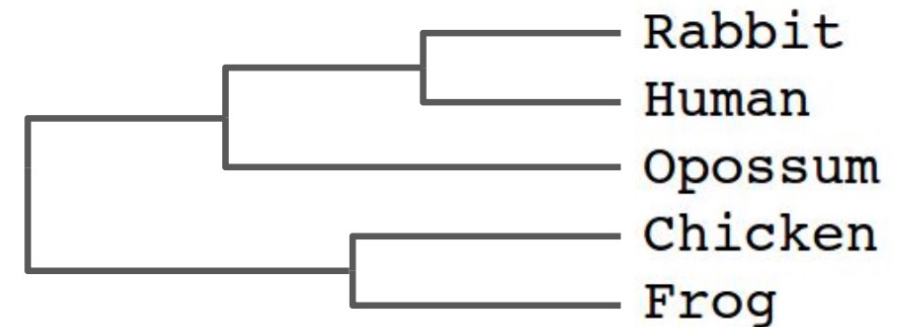
Parsimony versus likelihood

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



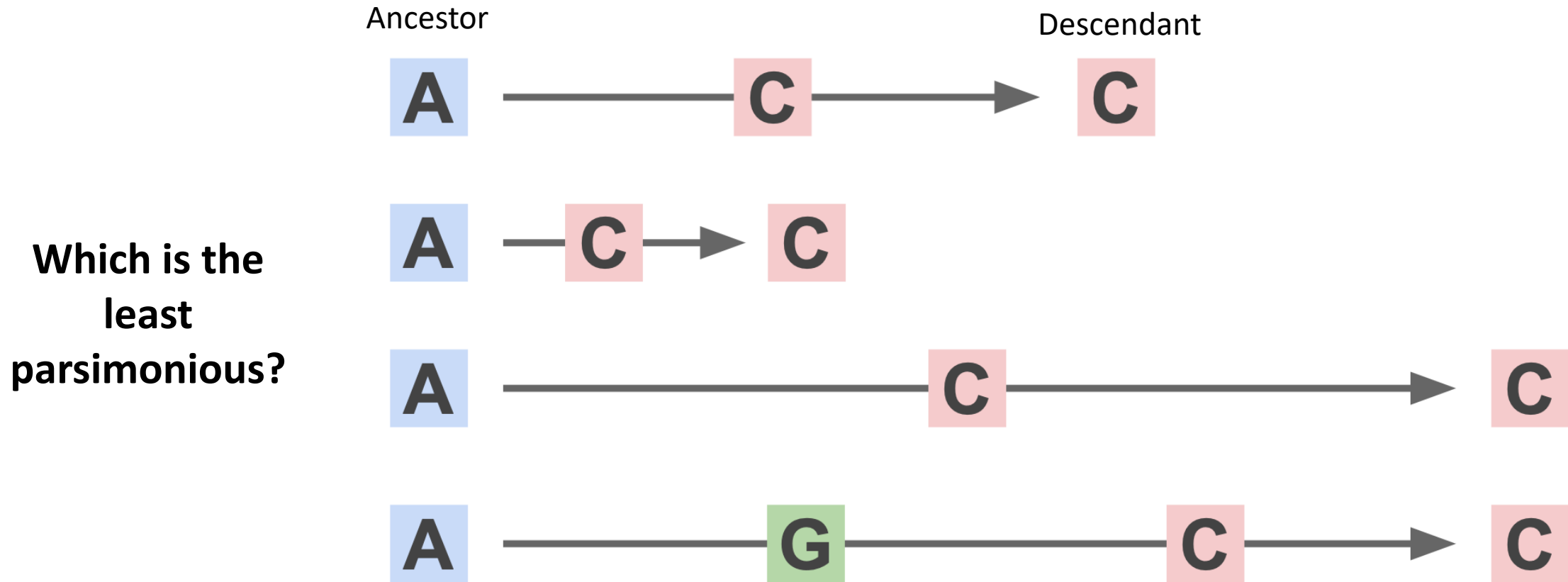
branch length can mean different things:

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

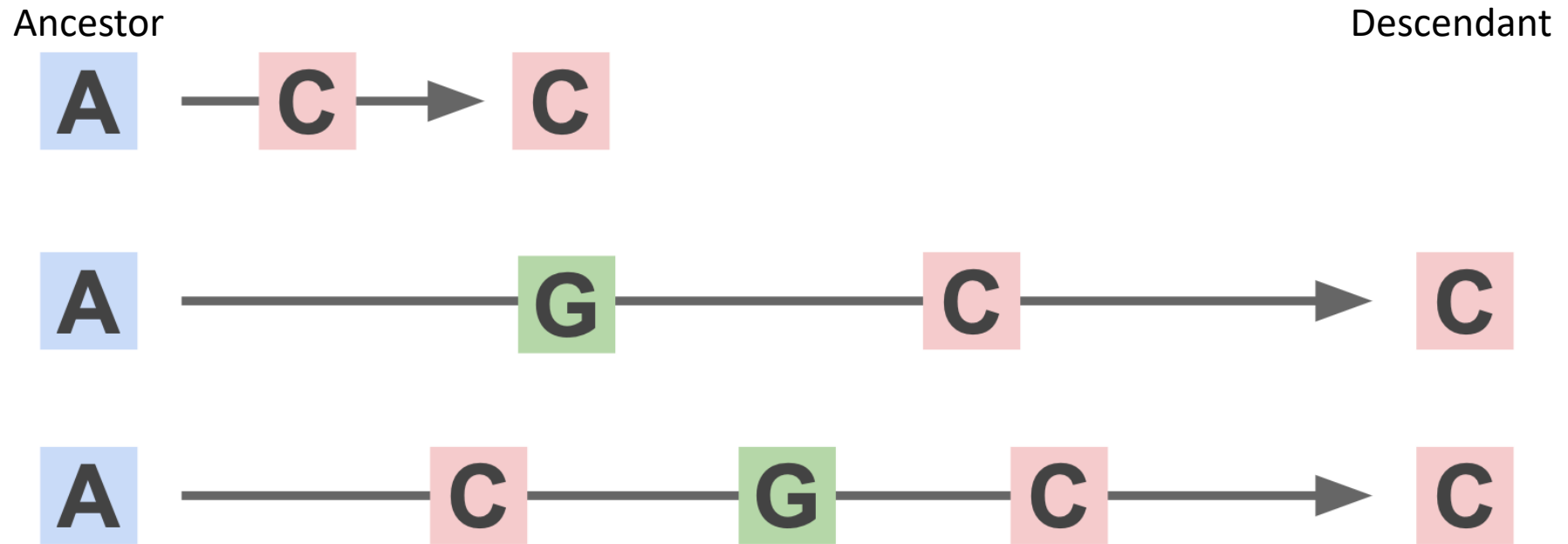


proposed tree has **branch lengths** in units of expected number of changes per site

Parsimony: minimum number of changes regardless of time/opportunity



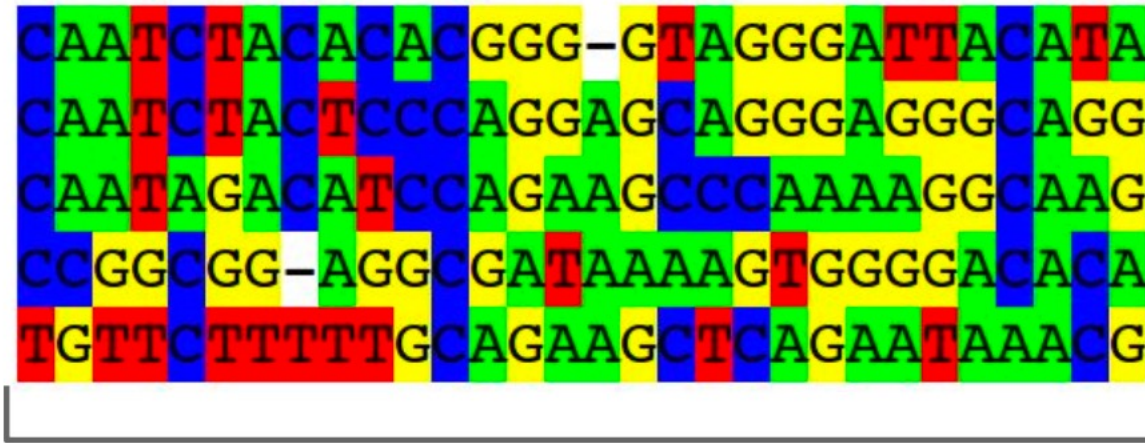
Likelihood: probability of ancestral and descendant status is a function of time (branch length)



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**)

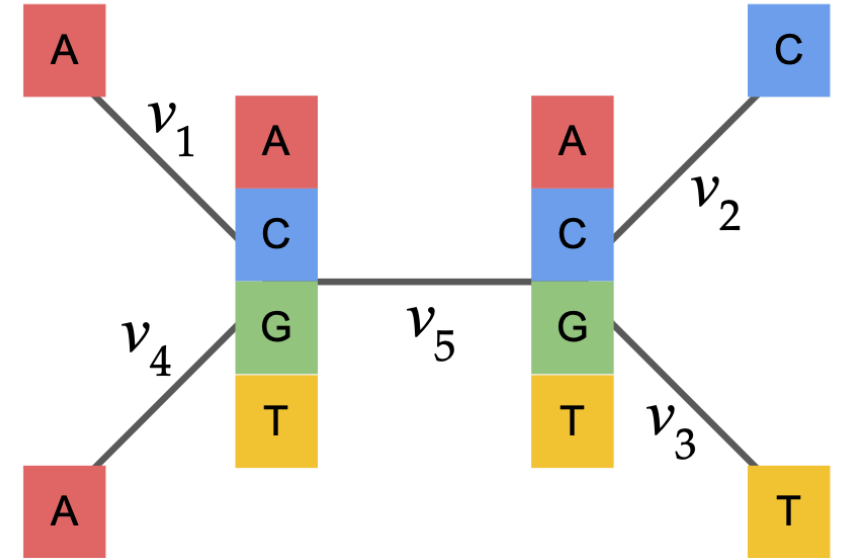
Likelihood cont'd.

Rabbit
Human
Opossum
Chicken
Frog



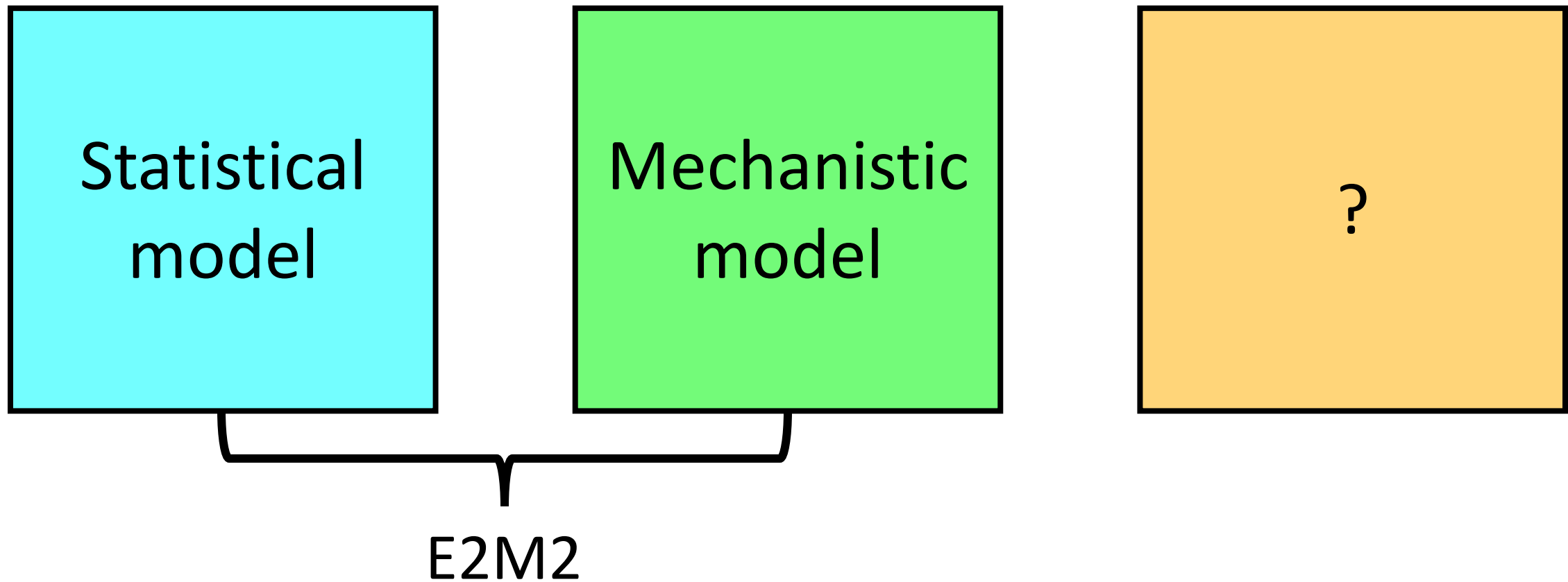
overall likelihood is the product of
likelihoods across characters (sites)

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

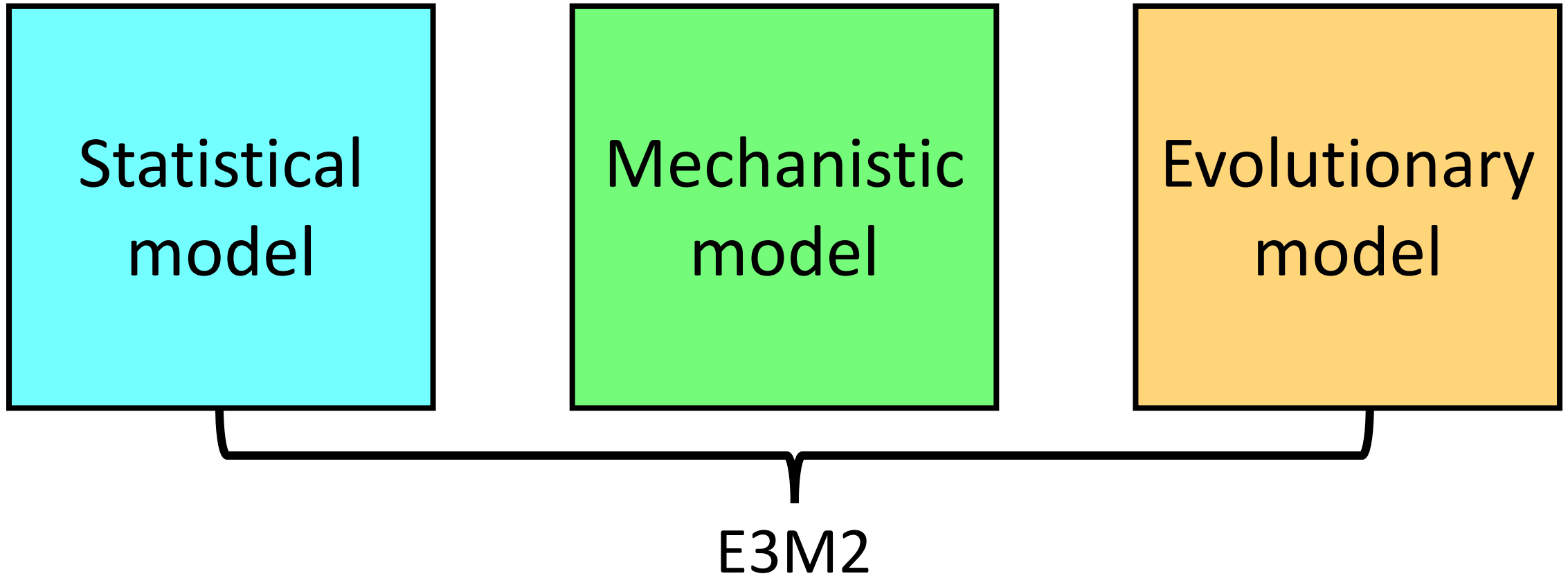


Consider *all possible ancestral states* at internal nodes, and calculate their contribution to the overall likelihood.*

So far...



Today...



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 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:

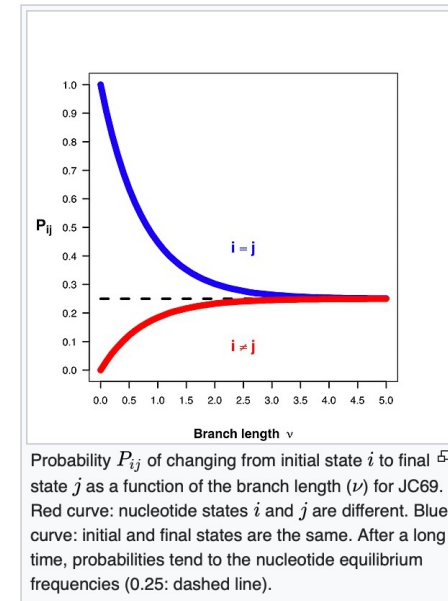
$$P_{ij}(\nu) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

It is worth noticing that $\nu = \frac{3}{4}t\mu = \left(\frac{\mu}{4} + \frac{\mu}{4} + \frac{\mu}{4} \right)t$ 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

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

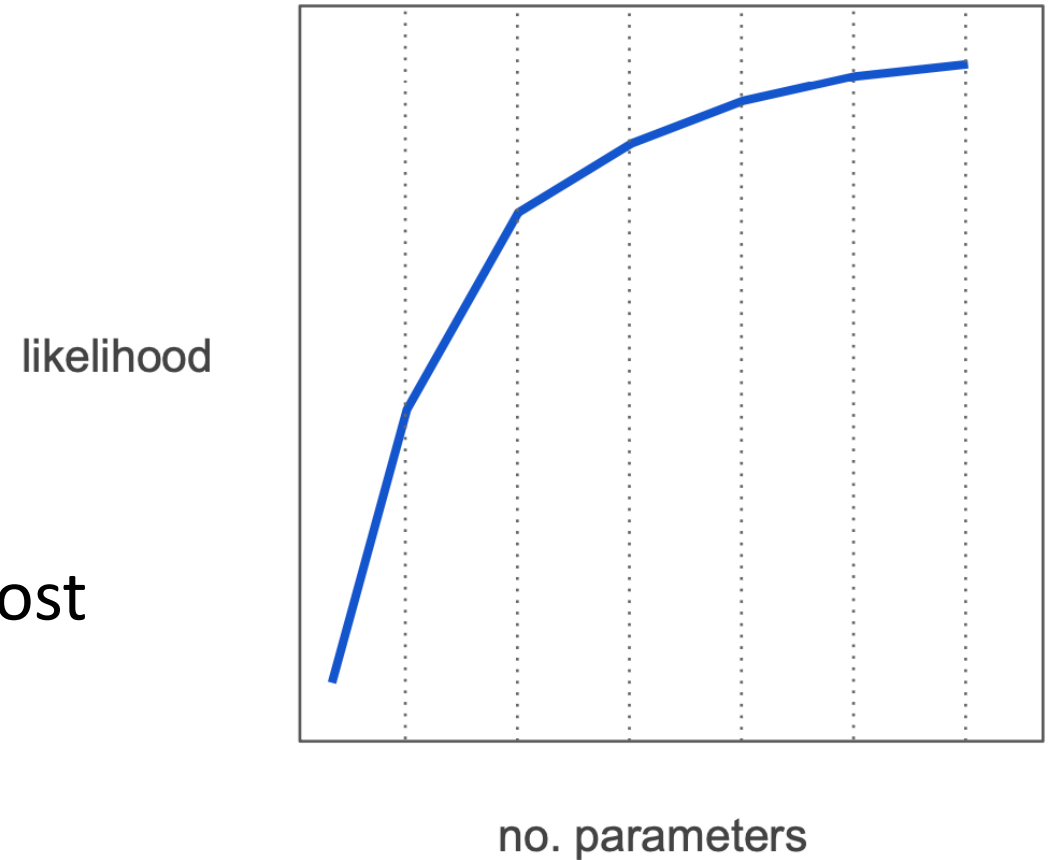


Model selection

More parameters means higher likelihood, but is the increase in likelihood necessary?

Model testing will give you two scores

- 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):

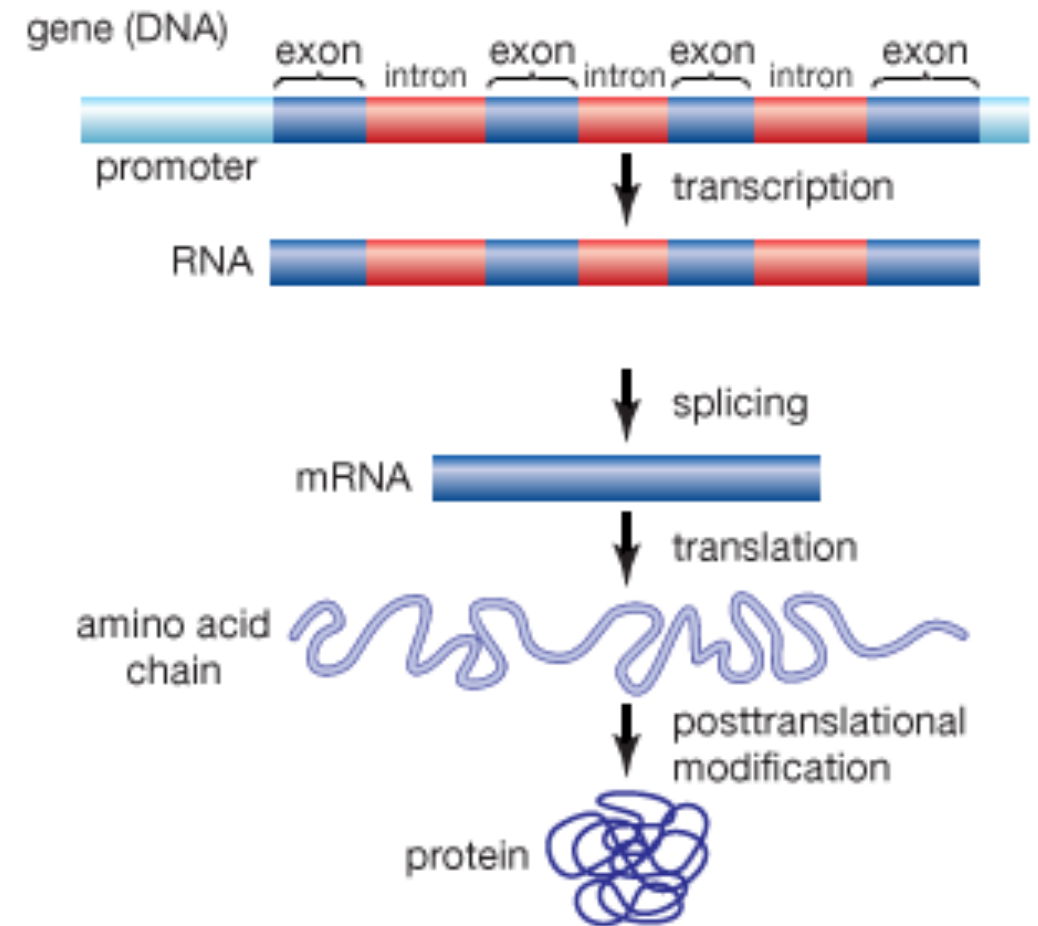
Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	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

- 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)

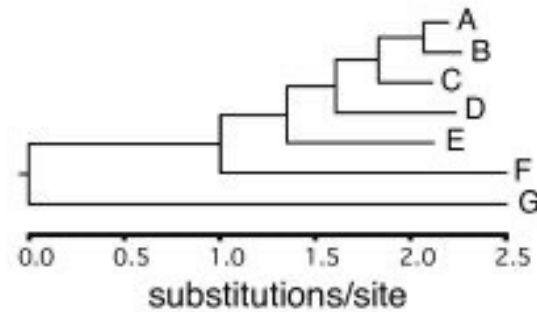
We can make inference about selection from these values, but makes things much more complicated



Rate heterogeneity across sites

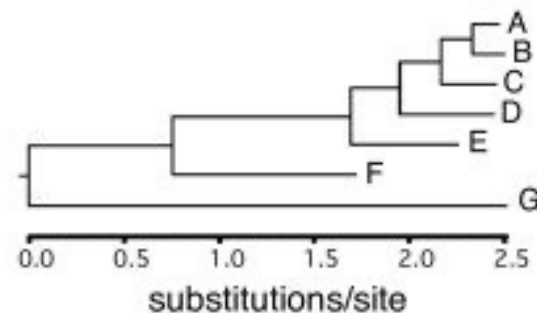
WITH rate heterogeneity

```
A TTEQGIKSSSTSLPAPQLPNWSGQYHEWVLKS---FQNEVK---KTLHCALSQGTATQSVLDELHADVWALLASSEVCYARPCGOVKPELAFLRKRA
B TACAGEKTGTSLPAPNLPNWSGQYHEWVLKS---ERADVI---KTMHCALSDGITATQSVLDELHFNVWALLASSEVCYARPCGOVKPELIYMKKQQA
C VAGGCEKAGTSLPAPYLPN--SGQYGEWVLKS---LSTHVI---KHMHCEDLSDSDTTTQSVLDELHGERWALLOSSEVCYARPCGOVKPELHEHECYKRA
D CAGQAEKTGTSLPALHLPNWA--QYGEWVLKS---FPSQPV---MPIQCVPLSDARTAAQSVLDELHVESDALLDSSEVCYARPCGA-RHDLKFVCYSKA
E DEGLTQKTGTSLPALALPNWSGQYFEWVLKS---YG---FGQGGAAGCKPLSGDKTSIHSLVDELHVAFLAALLMSGVCYALPCGAYKKALEFKCYLKA
F GEGFIKKTGTSPAPVALPDFAEQYDEWPLKSTLAYGRVNF---AAVPGAYLSDFGTGSHSVDELHONHAALLLSSEVCFAAPCGESKGALVVVCYSHA
G NDGPHIKKGTSGPAALPNQPIQYDEWVLKS---CEAKSI---NGSNWKPLSGKYTGLDYLDELHVMKDALLHATEVCLAPPCGY-TADLKAALGSPA
```



WITHOUT rate heterogeneity

```
A H----GENYFC--SQVAKYLAF-YSHNYL--EALLRHATLEIQH--KSNNEHGTGLEGPESA--EDPR--VPAGNEKLLGKVVNNEFSAPGL----IKKP
B Q----GENYFC--GGVAKYLAW-VGHNYL--EALLRHATMEINR--KSKKEEQNGLDGPESA--EDPR--IPAGGEKLLGMHNNMFGAAGL----VKRP
C S----GENYFC--PQVAKYLAW-MSNNYL--HAFETQAKLEIER--KRNQVEHGCGLDGPNGQ--EDPR--IKNSGQKLLGGY--KELKNPGL----FVRP
D Q----AHQYPC--SHIGKYFAW-VANAYH--HVLLRYAKLEVER--KRTADHSTDLVAPNGA--KFSV--LLPGPDALL-RHINKFISTPLA----FIKT
E E----EDQYPC--KENYKYLAW-VGHGELRAHSLSKHAKLATEK--KTEADHNTKLETAESP-LVEC--IPPLPDTRVAIVANTFFSAQSL----FIKT
F Q----GKKDLC--ENLN----TN--MQNRWL--QALHK-TITVVQHDGKSSMGDHCCKAIDSKAS-LSPC--VSSGGGYQKSNQIDFFVSNVTV----YLKS
G NNDFSKPFLFCNYTGIL---ILQCAG-----YLDGETMIGRFQ--STQVGLYSERLFDPRYKCMGETHKATNNTDTFGDRKAFKKRVSVKAFKQQTAPQ
```



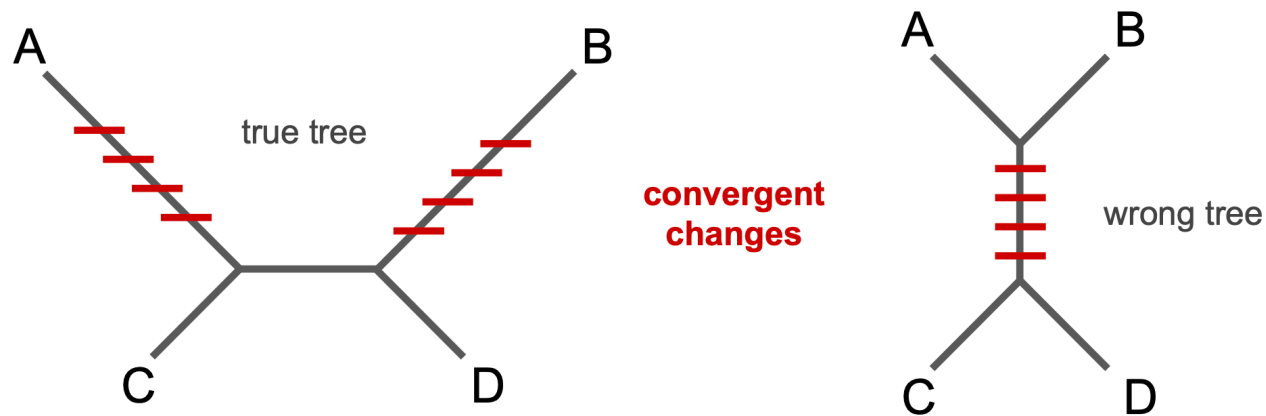
Rate heterogeneity across sites

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

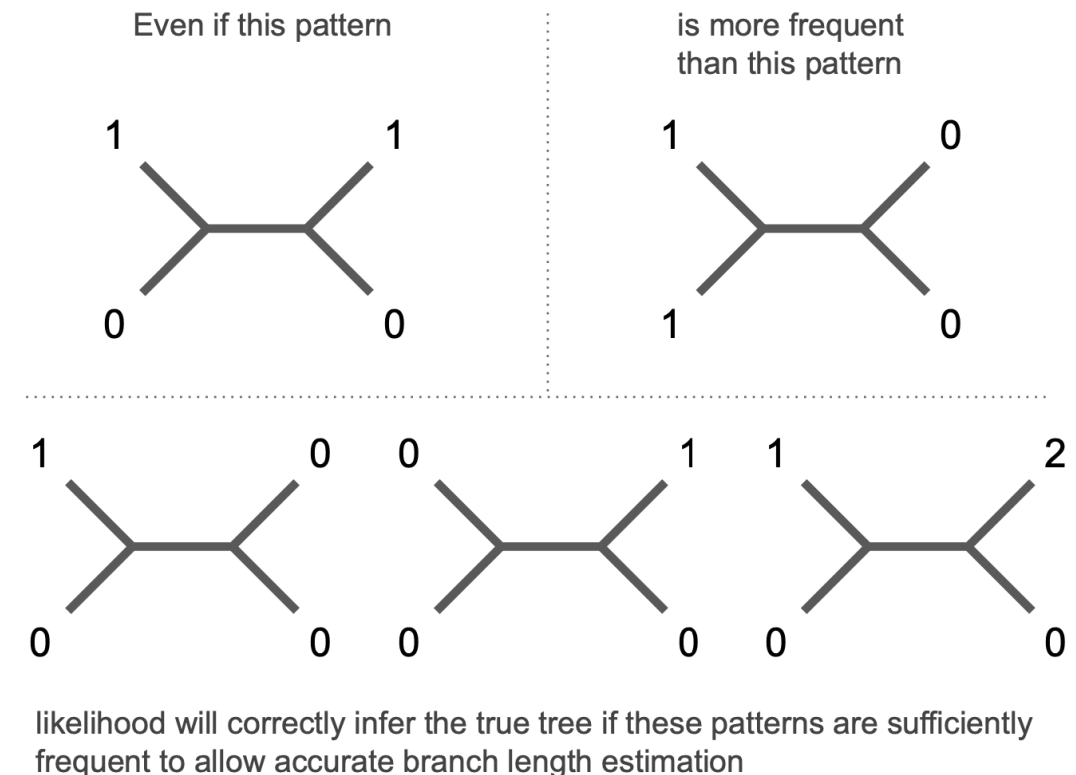
RateType	Explanation
+I	allowing for a proportion of invariable sites.
+G	discrete Gamma model (Yang, 1994) with default 4 rate categories. The number of categories can be changed with e.g. +G8 .
+GC	continuous Gamma model (Yang, 1994) (for AliSim only).
+I+G	invariable site plus discrete Gamma model (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.

Felsenstein zone

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

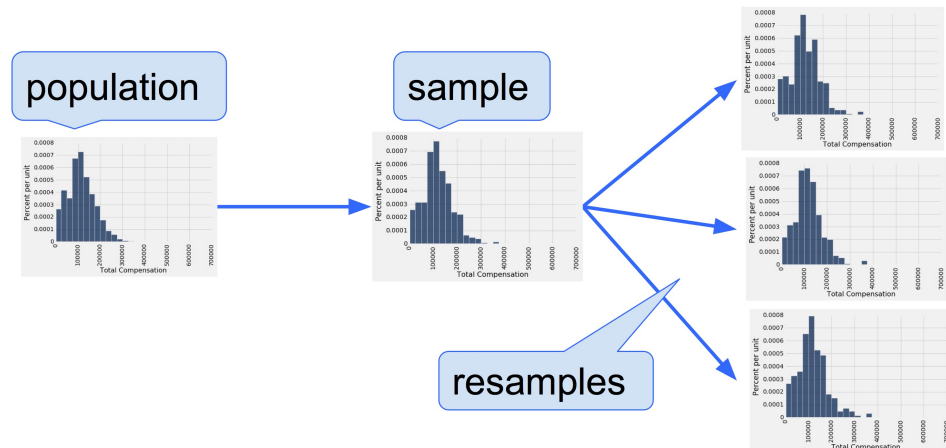


likelihood is a **consistent estimator** of tree topology because it converges on the correct value with increasing data



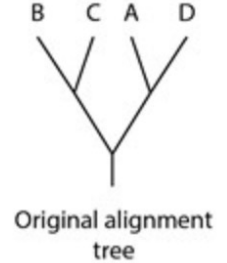
Bootstrapping

- Specify number of replicates: how many times does the test replicate the original sequence alignment?



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



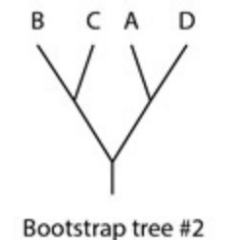
Bootstrap pseudo-replicate #1

	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



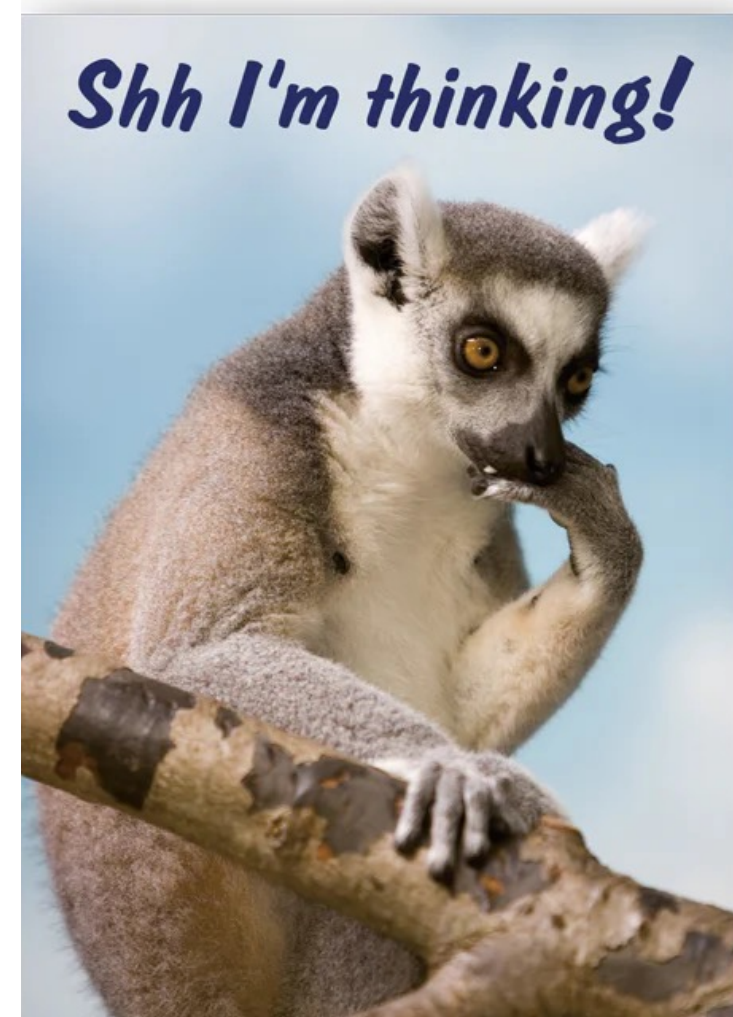
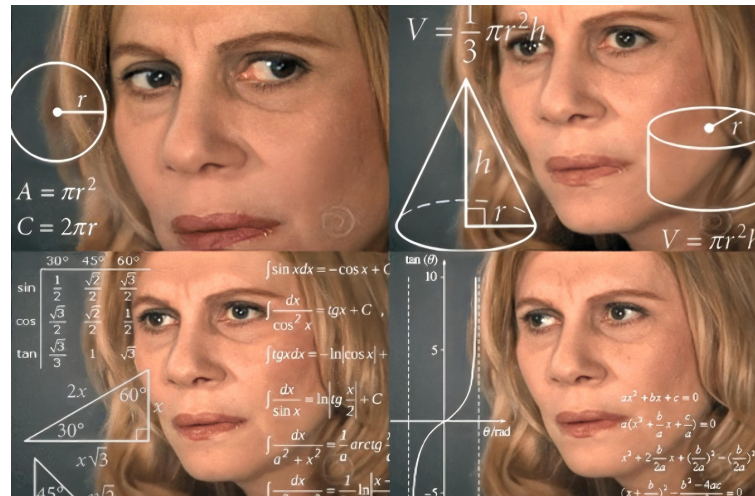
Bootstrap pseudo-replicate #2

	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



Checkpoint

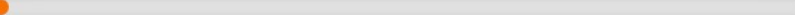
- More parameters = higher or lower likelihood?
- Low bootstrap values means low confidence in tree topology
- Not all sequences undergo evolution the same way, models account for this change




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

M11: Progress

PROGRESS 

Site coverage calculated 

[DETAILS](#) [✖ STOP](#)

[STATUS/OPTIONS](#)

RUN STATUS	
Start time	11-12-22 00:39:48
Operation Run Time	05:17:58
Status	Setting site coverage
Log Likelihood	-6,701.73
Operation	Bootstrapping ML tree
Replicate No.	227 of 500

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

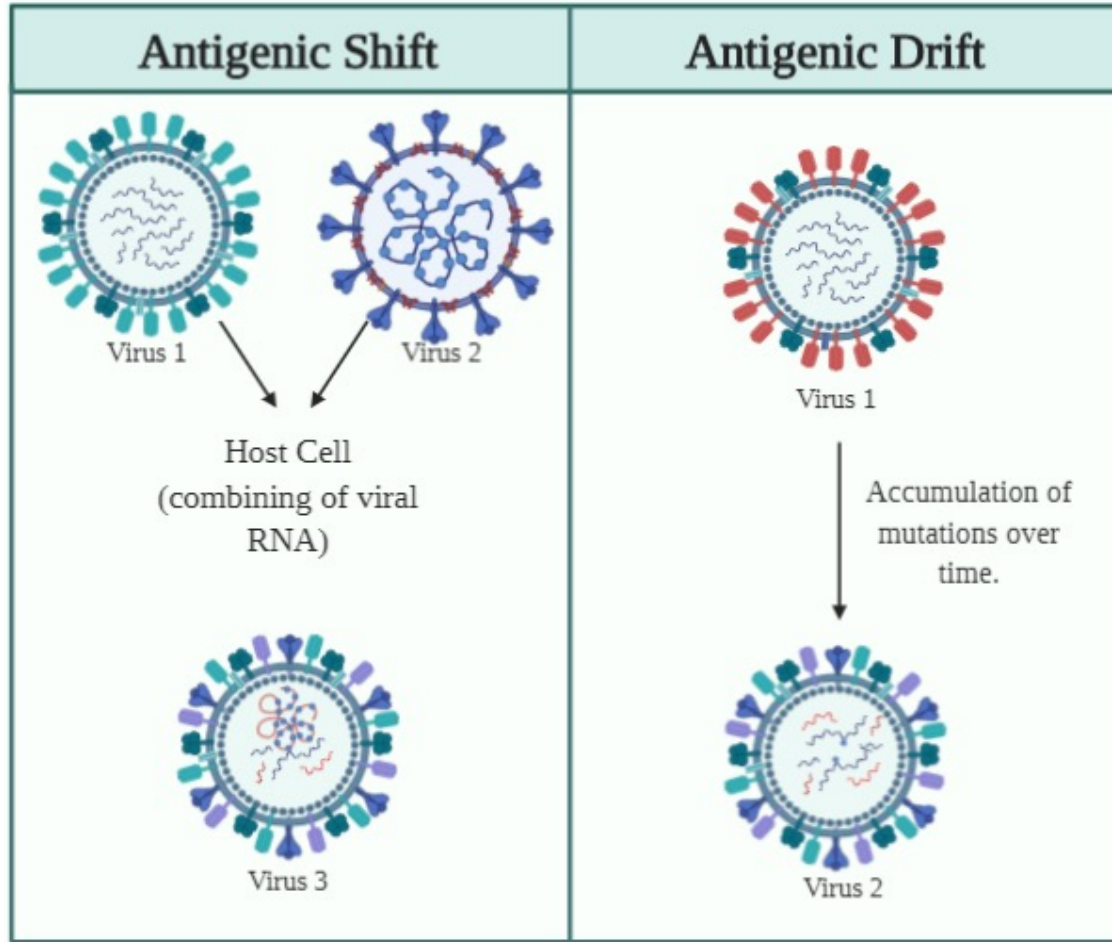
 PDF  Split View  Cite  Permissions  Share ▼

Abstract

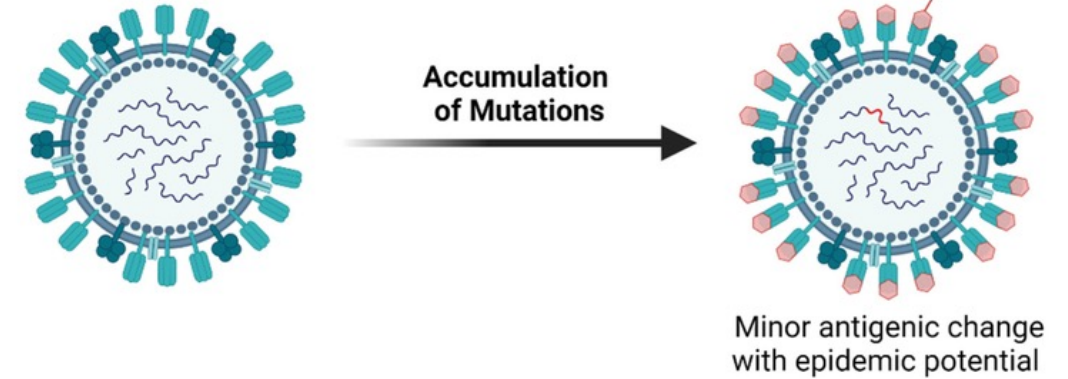
The rooting of the SARS-CoV-2 phylogeny is important for understanding the origin and early spread of the virus. Previously published phylogenies have used different rootings that do not always provide consistent results. We investigate several different strategies for rooting the SARS-CoV-2 tree and 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?**

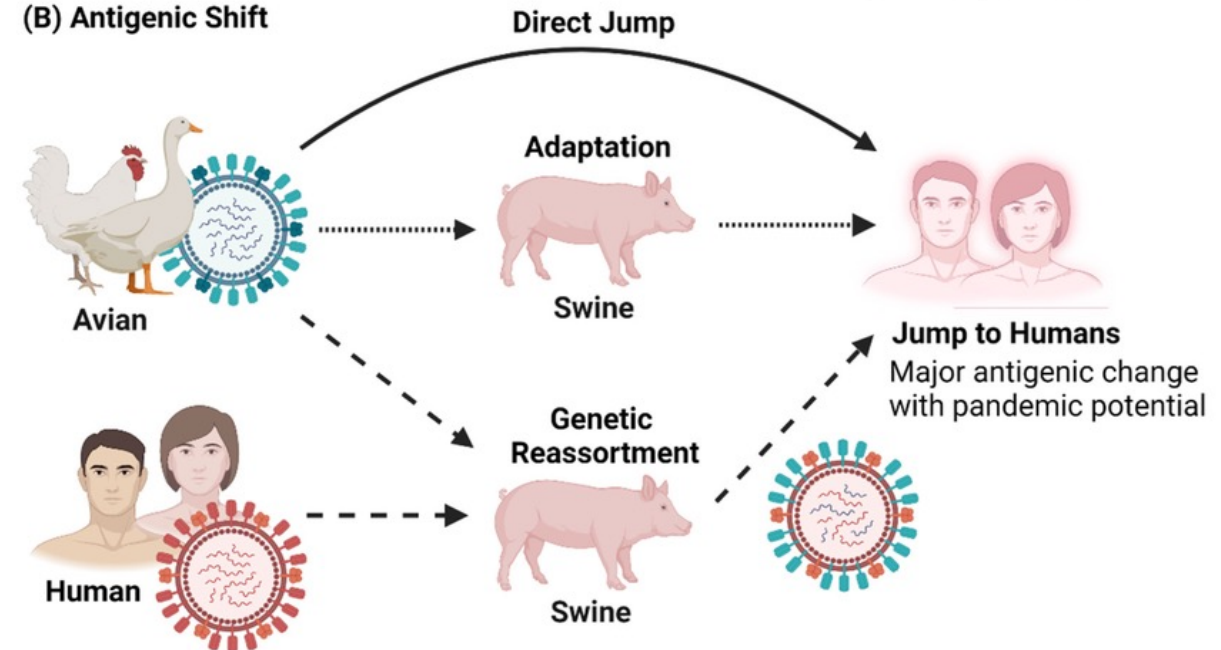
Case study: influenza – antigenic shift and drift



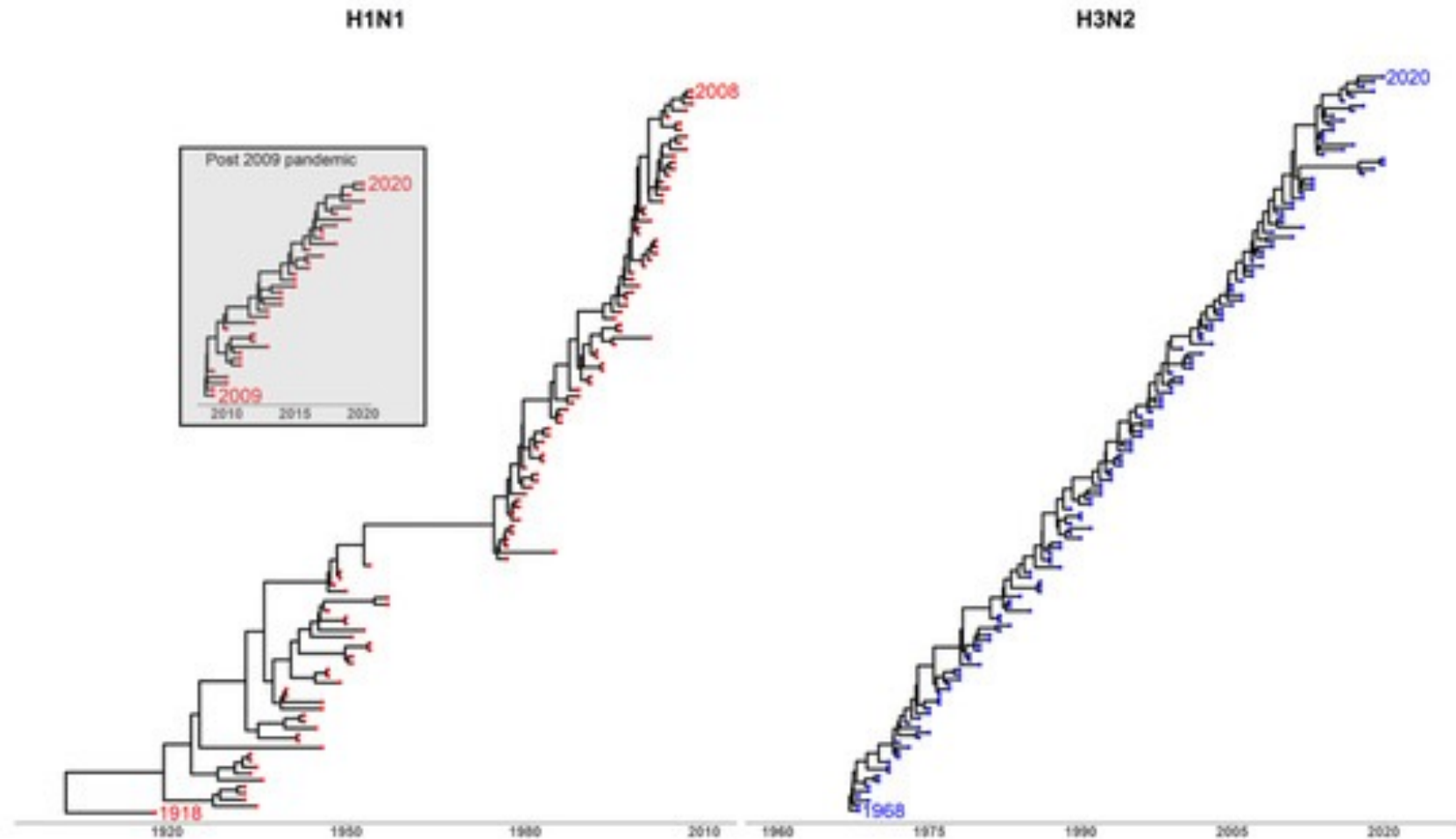
(A) Antigenic Drift



(B) Antigenic Shift



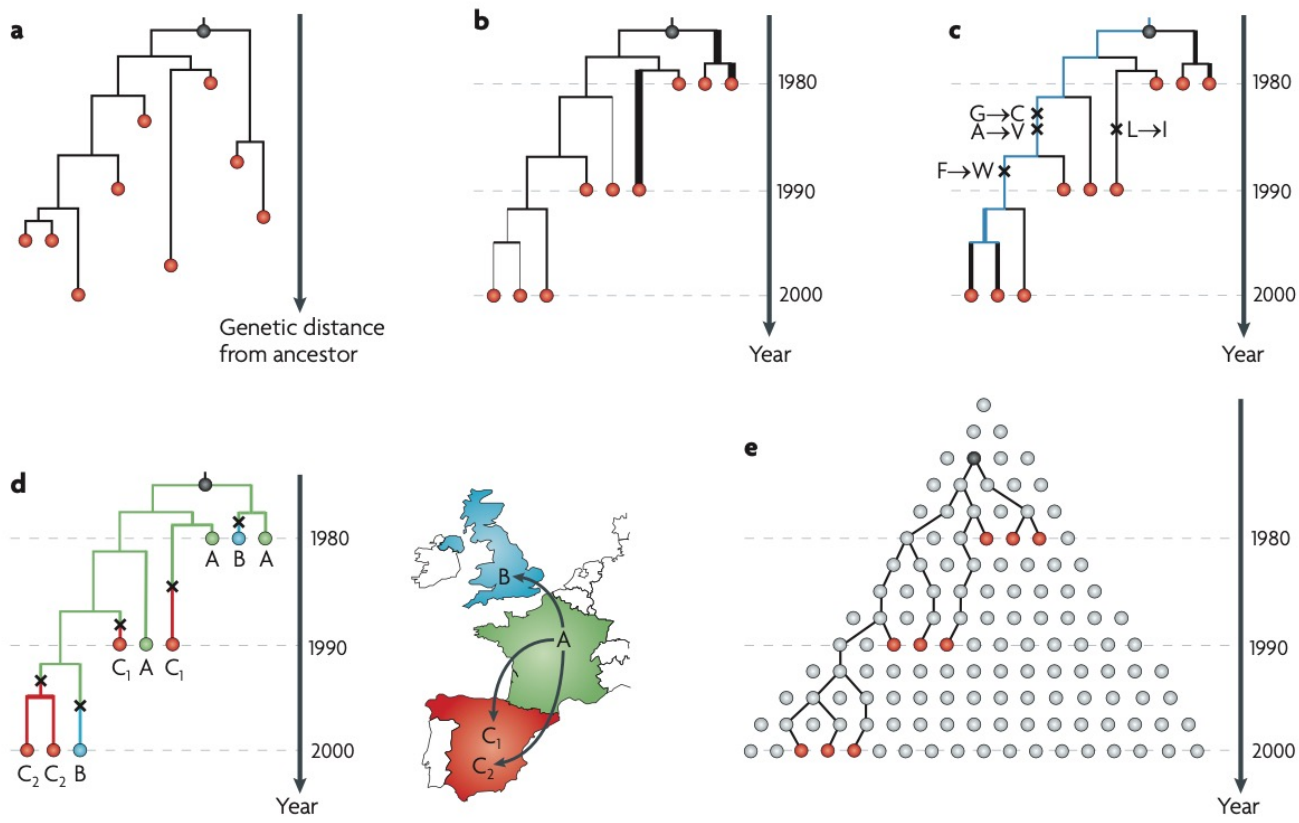
Case study: influenza – antigenic shift and drift



**Is drift or shift
happening here?**

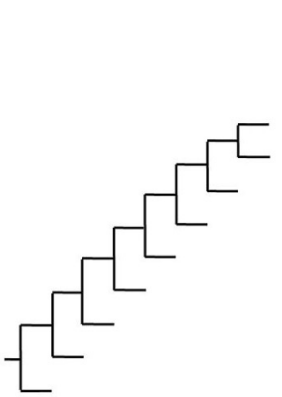
**How does this impact
vaccine design?**

Box 1 | **Phylogenetic techniques**



Units matter

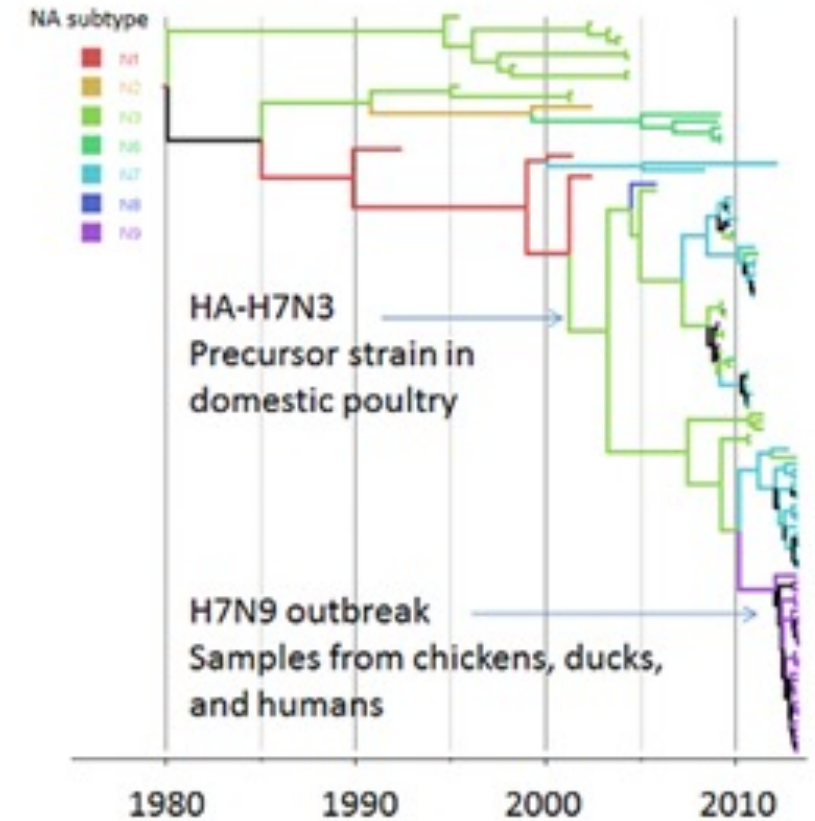
Shapes matter

Idealized Phylogeny Shapes	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective population dynamic processes	
		Population size dynamics	Spatial dynamics
 Time →		Exponential growth	Strong spatial structure
		Constant size	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

You can do a lot with phylogenies...

- Phylodynamics
- 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

Phylogeny of Hemagglutinin subtype H7 and reassortments with different Neuraminidases

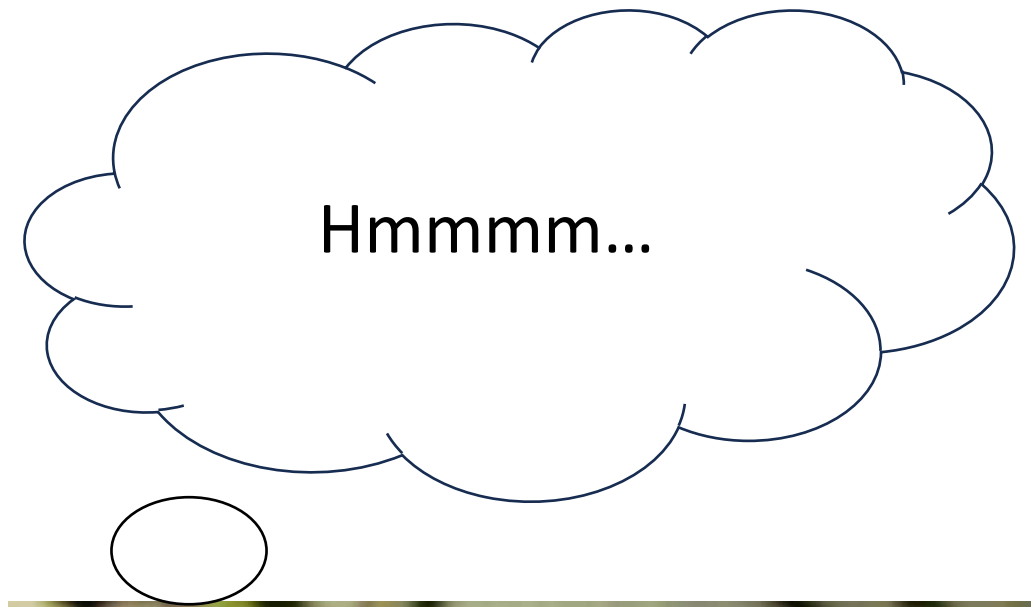


BEAST

Bayesian Evolutionary Analysis Sampling Trees

Checkpoint

- Phylogenetics is useful but computationally intensive
- Why do we need a root in phylogenetic trees?
- Can shape of a phylogeny tell us anything about the subject?



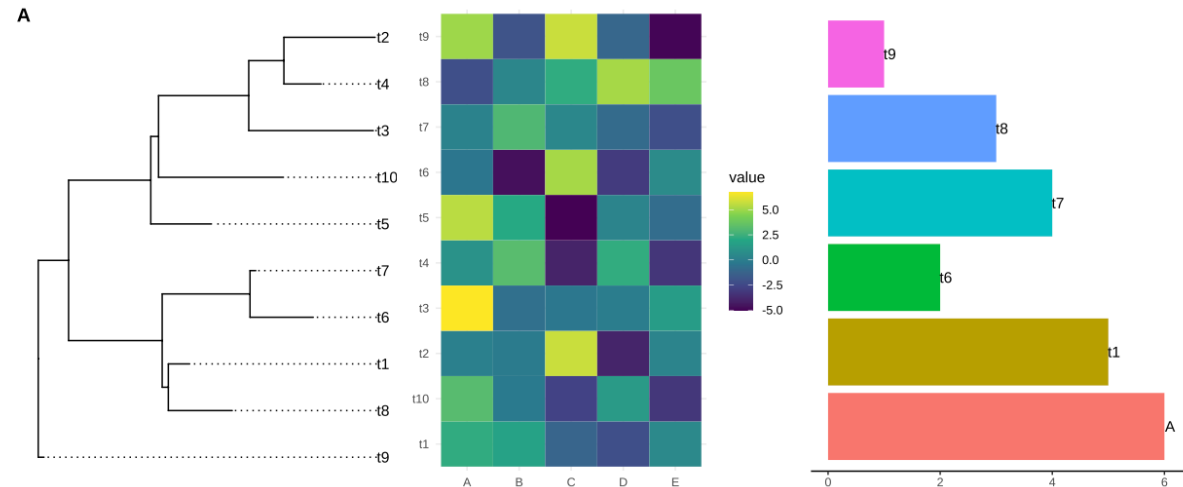
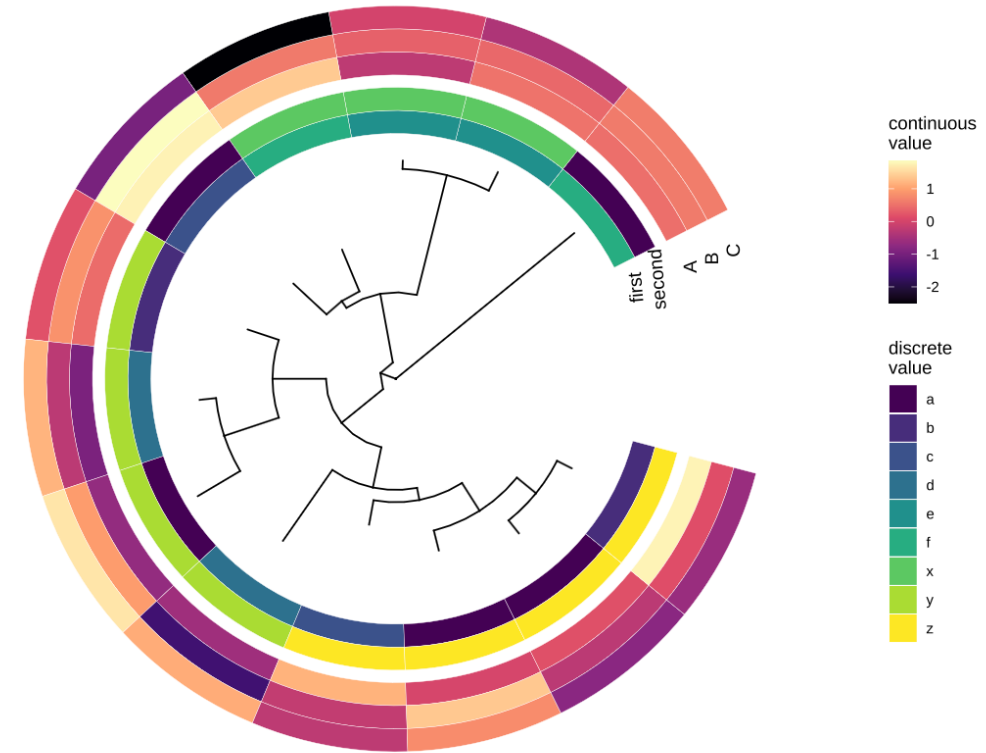
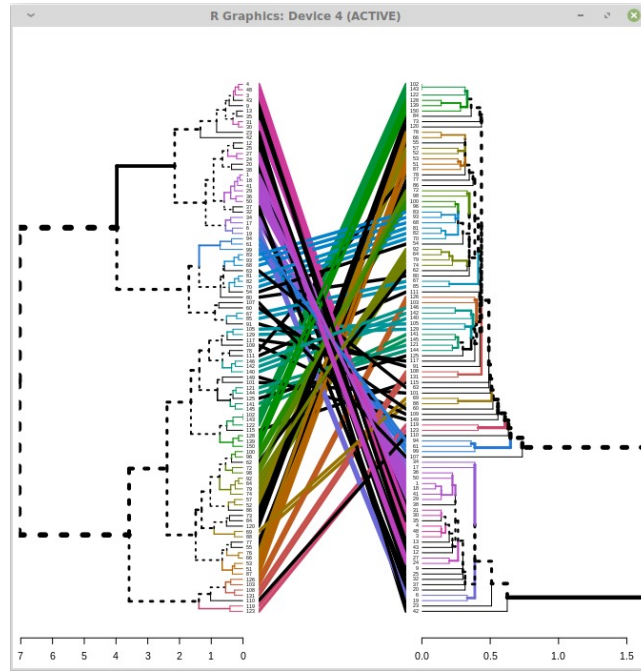
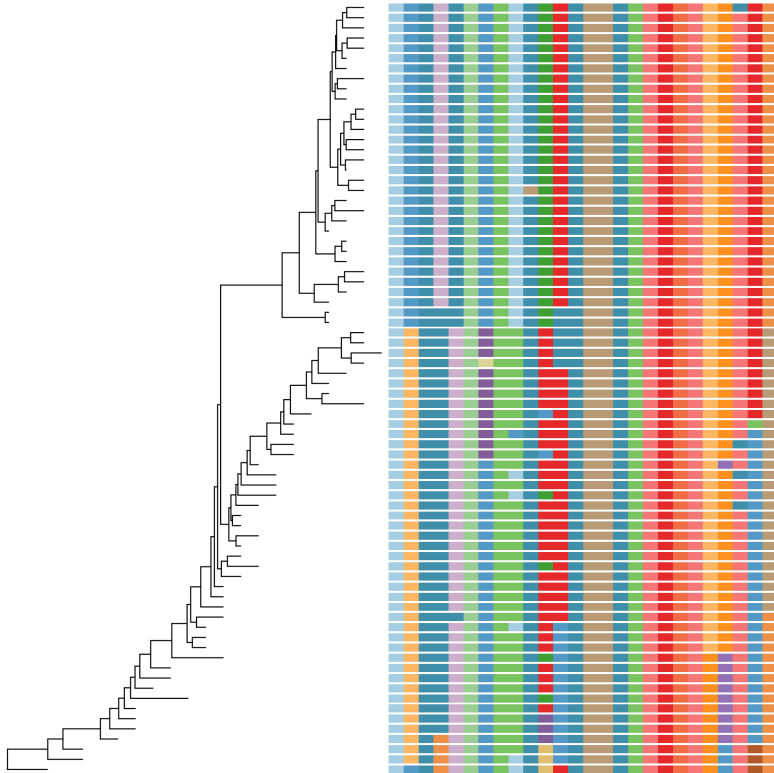
So you have your sequences, now what?

- 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 MEGA

All of this listed is free to use 😊

Potential for pretty figures

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



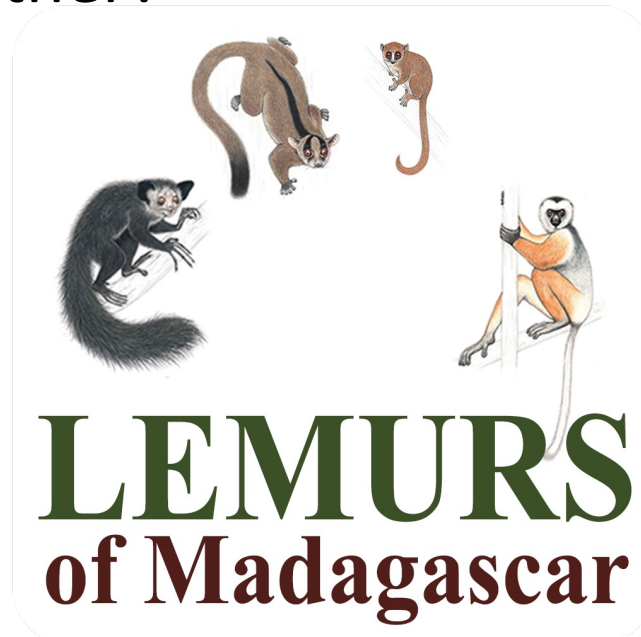
Final checkpoint

- Phylogenetic trees has a wide range of applications to many research topics!
- Uses modeling and stats “behind the scenes”
- **Can anyone give an example of how they could use a phylogeny for their own research? 😊**




Lemurs of Ranomafana national park - TUTORIAL

- Cytochrome B
 - Used a lot in species identification, limited variability within and much greater variation between species
- Prompt: how similar are the lemurs that can be found in Ranomafana National Park to each other?



SLIDES TO REVISIT ON YOUR OWN TIME

Steps to revisit later

 **National Library of Medicine**
National Center for Biotechnology Information

Log in

BLAST®

HomeRecent ResultsSaved StrategiesHelp

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...](#)

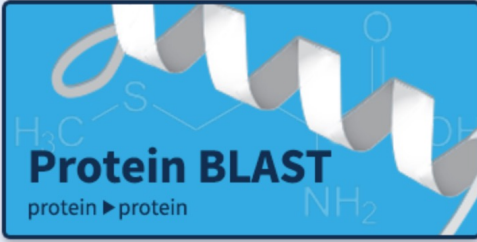
Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

Check what kind of sequence you are dealing with by doing a BLAST search

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

GGACAAGTAGCCTCCATTCTATACTTTTCTCTAATCCTTATTATTATACCAAC
TGTAAGCCTCATCGAAA
ACAAGATACTTAAATGAAGA

Or, upload file

Choose Fileno file selected ?

Job Title

Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?

Choose Search Set

Database

☒ Standard databases (nr etc.):☐ rRNA/ITS databases☐ Genomic + transcript databases☐ Betacoronaviru

Nucleotide collection (nr/nt)

?

Limit by

Organism

Optional

☐ exclude

Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ?

Exclude

Optional

☐ Models (XM/XP)☐ Uncultured/environmental sample sequences

Limit to

Optional

☐ Sequences from type material

Entrez Query

Optional

YouTube

Create custom database

Enter an Entrez query to limit search ?

Program Selection

Optimize for

☒ Highly similar sequences (megablast)☐ More dissimilar sequences (discontiguous megablast)☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

BLAST

Search using Megablast (Optimize for highly similar sequences)☐ Show results in a new window

NIH

National Library of Medicine

National Center for Biotechnology Information

Log in

BLAST® » blastn suite » results for RID-TBKASV0K016

HomeRecent ResultsSaved StrategiesHelp

< Edit Search

Save Search

Search Summary ▾

How to read this report?

BLAST Help Videos

Back to Traditional Results Page

Job TitleNC_035562.1:14221-15360 Microcebus rufus

RIDTBKASV0K016Search expires on 12-12 19:30 pmDownload All ▾

ProgramBLASTN ?Citation ▾

DatabasentSee details ▾

Query IDlcl|Query_55759

DescriptionNC_035562.1:14221-15360 Microcebus rufus isolate HAB...

Molecule typedna

Query Length1140

Other reportsDistance tree of resultsMSA viewer ?

Filter Results

Organismonly top 20 will appear☐ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity

E value

Query Coverage

to

to

to

Filter

Reset

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download ▾Select columns ▾Show100 ▾?

☒ select all100 sequences selectedGenBankGraphicsDistance tree of resultsMSA Viewer

	Description ▾	Scientific Name ▾	Max Score ▾	Total Score ▾	Query Cover ▾	E value ▾	Per. Ident ▾	Acc. Len ▾	Accession
<input checked="" type="checkbox"/>	Microcebus rufus isolate HAB06.12 mitochondrion, complete genome	Microcebus rufus	2106	2106	100%	0.0	100.00%	16819	KM112297.1
<input checked="" type="checkbox"/>	Microcebus rufus isolate VEV7.13 mitochondrion, complete genome	Microcebus rufus	1751	1751	100%	0.0	94.39%	16822	KM112317.1

An official website of the United States government [Here's how you know](#) ✓

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases Eulemur ruffrons cytochrome B **Search**

NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

COVID-19 Information

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Join NCBI at PAG 30
08 Dec 2022
San Diego, January 13-18, 2023 NCBI is looking forward to seeing you in person at the International Plant and Animal

Announcing the NCBI SARS-CoV-2 Variant Calling Pipeline and Related Data Products
01 Dec 2022
Still waiting for an analysis pipeline that

New Proximity Search Feature Available in PubMed
30 Nov 2022
PubMed, a free National Library of

1. Go to NCBI, and search for the thing you want to build a phylogeny for, in our case cytochrome B of lemurs in Ranomafana national park

Search NCBI

Eulemur rufifrons cytochrome B

×

Search

Results found in 4 databases

<div>Literature</div> <div>Bookshelf0</div> <div>MeSH0</div> <div>NLM Catalog0</div> <div>PubMed0</div> <div>PubMed Central4</div>	<div>Genes</div> <div>Gene0</div> <div>GEO DataSets0</div> <div>GEO Profiles0</div> <div>HomoloGene0</div> <div>PopSet0</div>	<div>Proteins</div> <div>Conserved Domains0</div> <div>Identical Protein Groups5</div> <div>Protein28</div> <div>Protein Family Models0</div> <div>Structure0</div>
<div>Genomes</div> <div>Assembly0</div> <div>BioCollections0</div> <div>BioProject0</div> <div>BioSample0</div> <div>Genome0</div> <div>Nucleotide28</div> <div>SRA0</div>	<div>Clinical</div> <div>ClinicalTrials.gov0</div> <div>ClinVar0</div> <div>dbGaP0</div> <div>dbSNP0</div> <div>dbVar0</div> <div>GTR0</div> <div>MedGen0</div>	<div>PubChem</div> <div>BioAssays0</div> <div>Compounds0</div> <div>Pathways0</div> <div>Substances0</div>

This is what it will look like, you can go to Nucleotide under the genome category and click on that

Nucleotide

Nucleotide

Eulemur rufifrons cytochrome b

Search

Create alertAdvancedHelp

Species

Animals (28)

Customize ...

Molecule types

genomic DNA/RNA (28)

Customize ...

Source databases

INSDC (GenBank) (28)

Customize ...

Sequence Type

Nucleotide (28)

Genetic compartments

Mitochondrion (28)

Sequence length

Custom range...

Release date

Custom range...

Revision date

Custom range...

Clear all

Show additional filters

Summary

20 per page

Sort by Default order

Send to:

Filters: [Manage Filters](#)

See Gene information for b cytochrome **cytochrome b**

b in [Drosophila melanogaster \(2\)](#) [Escherichia phage Lambda](#) [All 50 Gene records](#)

cytochrome in [Cricetulus griseus](#) [Tripterygium wilfordii \(2\)](#) [All 4 Gene records](#)

cytochrome b in [Pongo abelii](#) [1 Gene record](#)

Items: 1 to 20 of 28

<< First < Prev Page 1 of 2 Next > Last >>

☐ [Eulemur rufifrons clone Erufi-NHMB89006 cytochrome b gene, partial cds; mitochondrial](#)

1. 223 bp linear DNA

Accession: KF708347.1 GI: 556926369

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Eulemur rufifrons clone Erufi-NHM1882314 cytochrome b gene, partial cds; mitochondrial](#)

2. 223 bp linear DNA

Accession: KF708346.1 GI: 556926367

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Eulemur rufifrons clone Erufi-MCZ16357 cytochrome b gene, partial cds; mitochondrial](#)

3. 223 bp linear DNA

Accession: KF708345.1 GI: 556926365

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Eulemur rufifrons clone Erufi-MCZ16356 cytochrome b gene, partial cds; mitochondrial](#)

4. 223 bp linear DNA

Accession: KF708344.1 GI: 556926363

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

Find related data

Database: Select

Find items

Search details

("Eulemur rufifrons"[Organism] OR Eulemur rufifrons[All Fields]) AND cytochrome b[All Fields]

Search

See more...

Recent activity

Your browsing activity is temporarily unavailable.

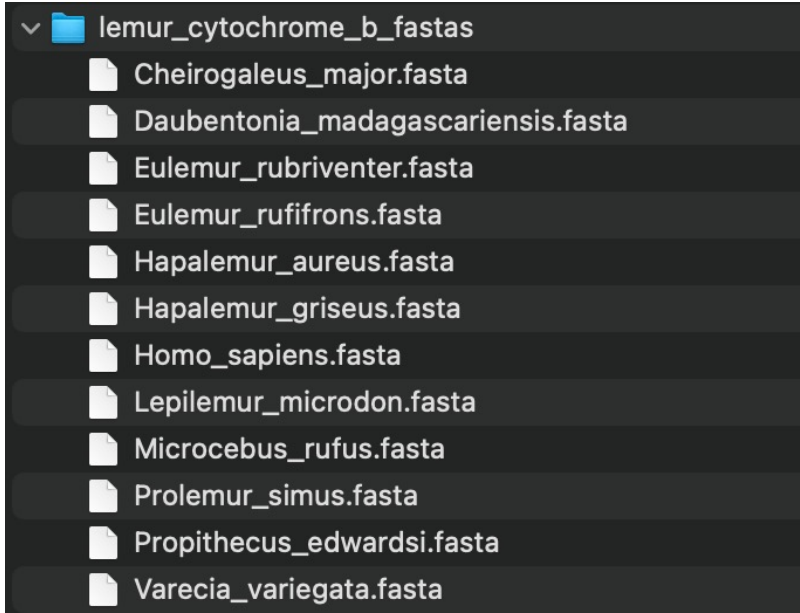
Pick the sequence of what you're interested in, in our case we want a complete cds

We might want partial cds if we have a partial sequence of interest, but right now we're just building a tree with known data, so complete cds is best

Cds: protein coding sequence

Then download the fastas

- ☐ [Eulemur rufifrons clone Erufi-MM-448 cytochrome b gene, complete cds; mitochondrial](#)
7. 1,140 bp linear DNA
- Accession: KF708293.1 GI: 556926260
- [Protein](#) [PubMed](#) [Taxonomy](#)
- [GenBank](#) [FASTA](#) [Graphics](#)



Step 2: when you have all your sequences of interest and your outgroup, you need to concatenate the sequences into one file, you can do this by making a text/edit file and pasting each sequence in, otherwise follow instructions on command line (mac) or powershell (windows) to do this

```
lemur_cytochrome_b_fastas — -bash — 121x27
Last login: Wed Nov 23 12:34:19 on ttys000

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Gwenddolens-MacBook-air:~ gwenddolenkettenburg$ cd Desktop
(base) Gwenddolens-MacBook-air:Desktop gwenddolenkettenburg$ cd Intro_phylogenetic_modeling_Kettenburg
(base) Gwenddolens-MacBook-air:Intro_phylogenetic_modeling_Kettenburg gwenddolenkettenburg$ cd lemur_cytochrome_b_fastas
(base) Gwenddolens-MacBook-air:lemur_cytochrome_b_fastas gwenddolenkettenburg$ cat *.fasta>lemur_cytB_concatenated
(base) Gwenddolens-MacBook-air:lemur_cytochrome_b_fastas gwenddolenkettenburg$
```

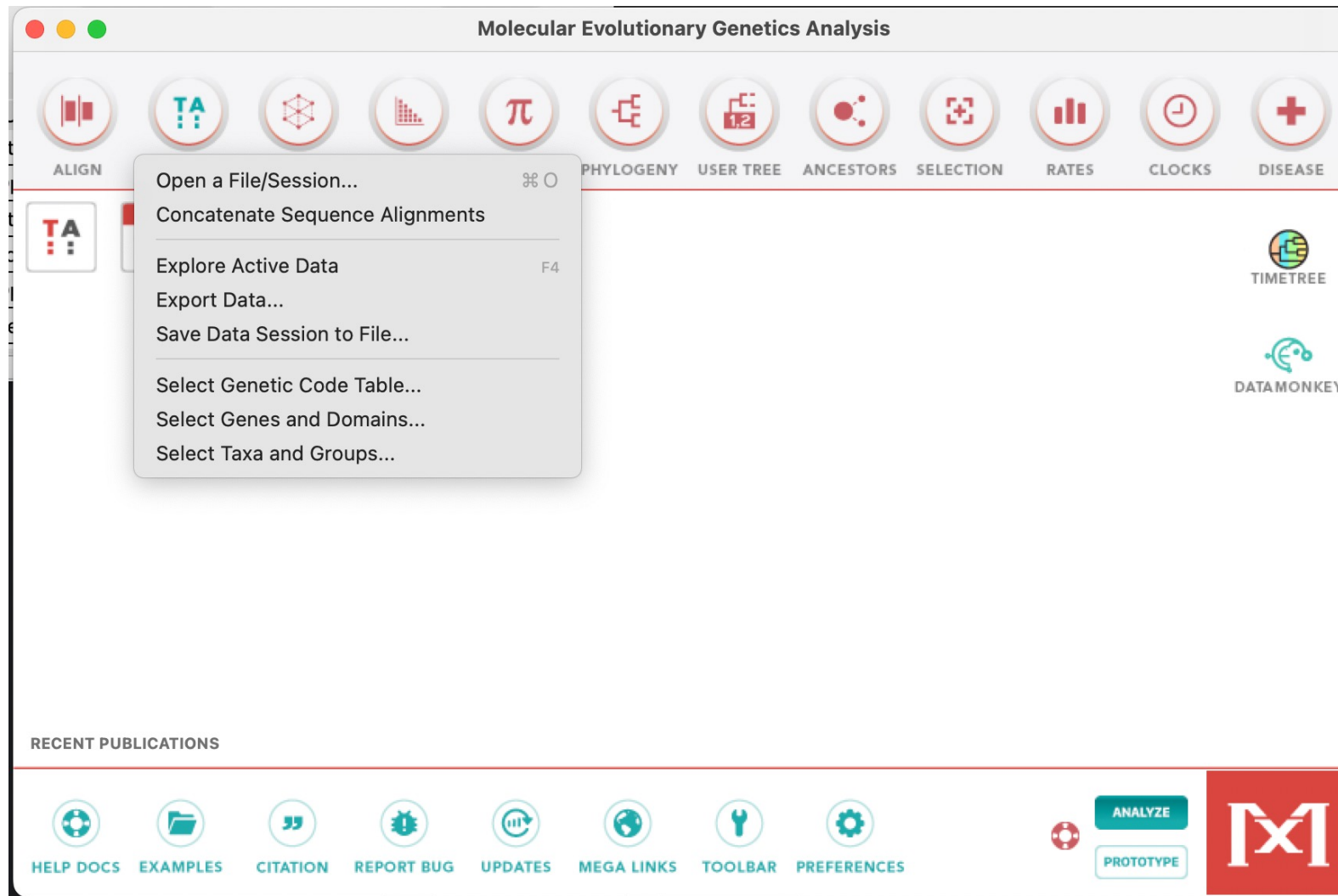
Example 1: Merge with file names (This will merge file1.csv & file2.csv to create concat.csv)

```
type file1.csv file2.csv > concat.csv
```

Example 2: Merge files with pattern (This will merge all files with csv extension and create concat.csv)

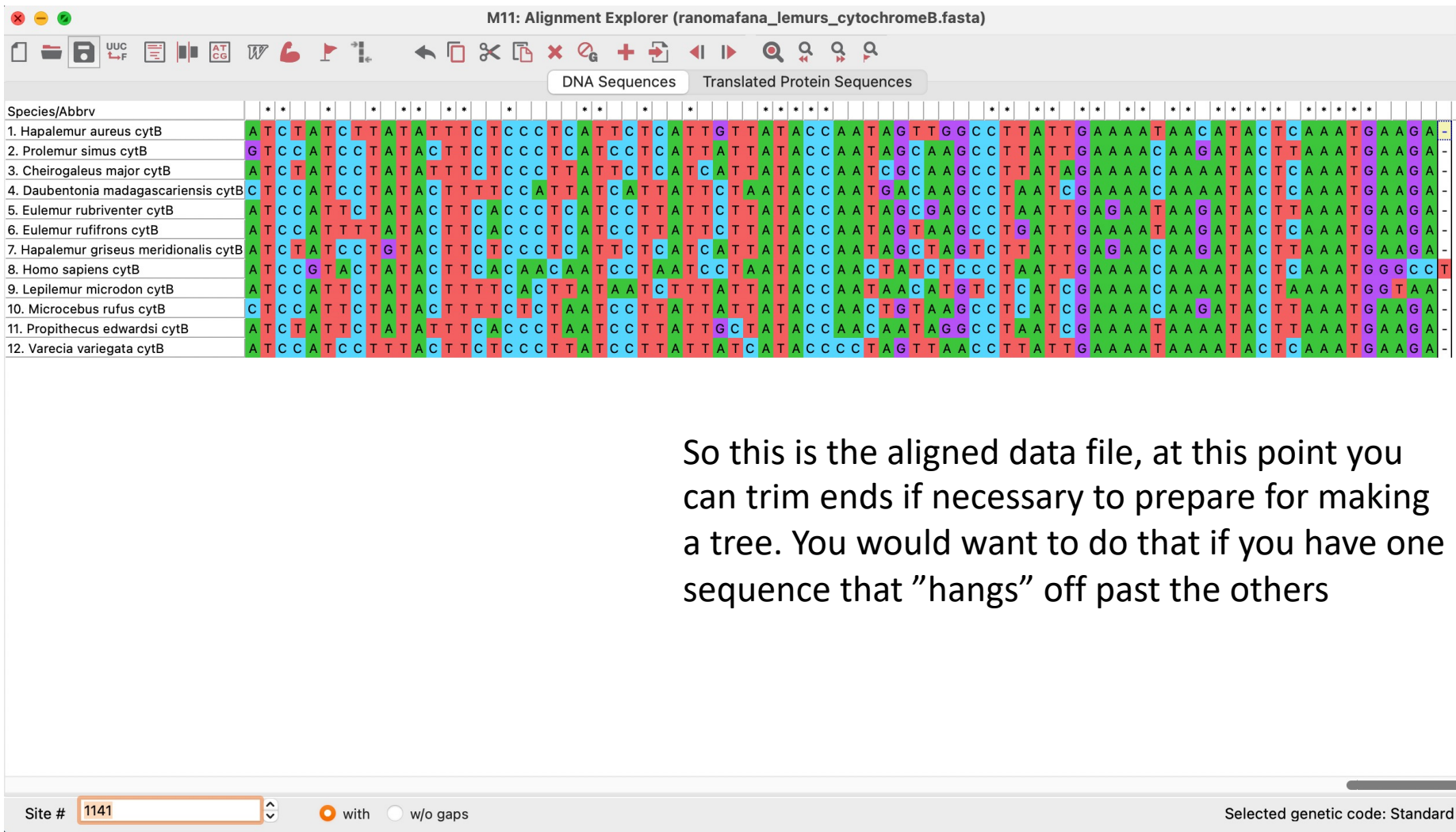
When using asterisk(*) to concatenate all files. Please DON'T use same extension for target file(Eg. .csv). There should be some difference in pattern else target file will also be considered in concatenation

```
type *.csv > concat_csv.txt
```



Step 3: open MEGA, and open a file/session, select your concatenated fasta file

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



MEGA11 Data Edit Search Alignment Web Sequencer Display Help

intro_phylogenetic_modeling_Kettenburg — Saved to my Mac

Transitions Animations Slide Show Review View Tell me Share Comments

Body) 28 A A A Paragraph Insert Drawing Sensitivity Designer

M11: Alignment Explorer (ranomafana_lemurs_cytochromeB_aligned.mas)

DNA Sequences Translated Protein Sequences

MEGA Format FASTA Format NEXUS/PAUP Format

Create New InL Open Open a Recently Used File Close Phylogenetic Analysis Save Session Export Alignment DNA Sequences Protein Sequences Translate/Untranslate Genetic Code Reverse Complement Reverse Complement Quit

11. Propithecus edwardsi cytB
12. Varecia variegata cytB

16

Sun Dec 11 12:31 AM

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

Open up RAxML GUI

INPUT

LOAD ALIGNMENT

ANALYSIS

ML + transfer bootstrap expectation + consensus ▾
Analysis

1 ▾
Runs

100 ▾
Reps.

628076
Seed

OUTPUT

Select output directory

output
Select output name

RAXML

raxml-ng-ARM64 ▾
Binary

RUN

raxml-ng-ARM64 --all --msa RAxML_output_concat.txt --model --prefix output
seed 628076 --bs-metric tbe --tree rand{1} --bs-trees 100
Command

CONSOLE



raxmlGUI 2.0.10 raxml-ng-ARM64 1.1.0

How to cite? [For questions or suggestions contact us!](#)

I
N
P
U
T

nucleotide

ranomafana_lemurs_cytochromeB_align.fas
12 sequences of length 1141



GTR

Substitution model

none

Stationary frequencies

none

Proportion of invariant sites

none

Rate heterogeneity

RUN MODELTEST

Partition 1/1:

	Model	Score	Weight
BIC	TIM2+I+G4	13908.7732	0.9990
AIC	TIM2+I+G4	13762.6230	0.8741
AICc	TIM2+I+G4	13763.6230	0.8741

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

nucleotide

ranomafana_lemurs_cytochromeB_align.fas

12 sequences of length 1141

TIM2

none

+I (ML estimate)

Substitution model

Stationary frequencies

Proportion of invariant sites

+GAMMA (mean)

RUN MODELTEST

Rate heterogeneity

<none>

Hapalemur_aureus_cytB

Prolemur_simus_cytB

Cheirogaleus_major_cytB

Daubentonia_madagascariensis_cytB

Eulemur_rubriventer_cytB

Eulemur_rufifrons_cytB

Hapalemur_griseus_meridionalis_cytB

Homo_sapiens_cytB

Lepilemur_microdon_cytB

Microcebus_rufus_cytB

Propithecus_edwardsi_cytB

Varecia_variegata_cytB

ML + transfer bootstrap expectation + consensus

1

100

Runs

Reps.

cytochromeB_align'

cytochromeB_align.ckp

cytochromeB_align.log

cytochromeB_align.out

cytochromeB_align.tree

RAXML

raxml-ng-ARM64

RUN

Binary

raxml-ng-ARM64 --all --msa /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cytochr --model TIM2+I+G --prefix /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cytochr seed 177748 --bs-metric tbe --tree rand{1} --bs-trees 100

Command

P.Inv: 0.4945
Alpha: 0.2560
Alpha-P.Inv: 0.7313
P.Inv-Alpha: 0.3876
Frequencies: 0.3075 0.3419 0.1142 0.2364

Commands:
> phylml -i /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cyto as -m 010232 -f m -v e -a e -c 4 -o tlr
> raxmlHPC-SSE3 -s /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lem align.fas -m GTRGAMMAIX -n EXEC_NAME -p PARSIMONY_SEED
> raxml-ng --msa /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs ign.fas --model TIM2+I+G4
> paup -s /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cytoch as -m TIM2+I+G4
> iqtree -s /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cyto as -m TIM2+I+G4

Summary:

Partition 1/1:

	Model	Score	Weight
BIC	TIM2+I+G4	13908.7732	0.9990
AIC	TIM2+I+G4	13762.6230	0.8741
AICc	TIM2+I+G4	13763.6230	0.8741

Execution results written to /Users/gwenddolenkettenburg/Desktop/RAXML/RAXML_C nomafana_lemurs_cytochromeB_align.out
Starting tree written to /Users/gwenddolenkettenburg/Desktop/RAXML/RAXML_GUI_M fana_lemurs_cytochromeB_align.tree

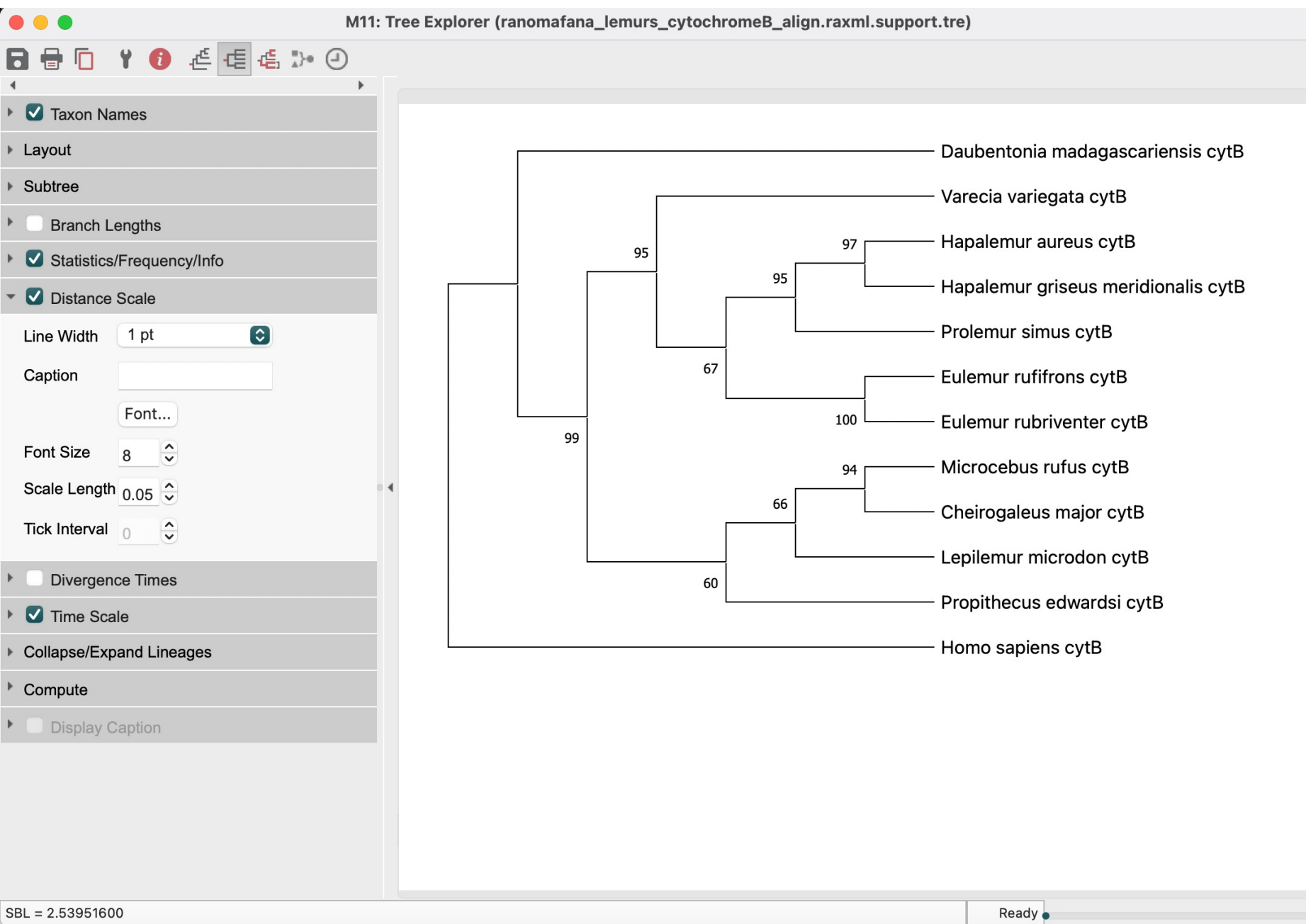
Using the model from modeltest, change as needed in the input box, then in analysis set the outgroup and then hit run in the RAXML section

It will also spit out some files...

>	modeltest	Today, 1:24 PM	--	Fo
	ranomafana_lemurs_cytochromeB_align.fas	Today, 1:22 PM	14 KB	De
▼	raxml	Today, 1:26 PM	--	Fo
	ranomafana_lemurs_cytochromeB_align.raxml.bestModel.txt	Today, 1:26 PM	90 bytes	Pl
	ranomafana_lemurs_cytochromeB_align.raxml.bestTree.tre	Today, 1:26 PM	518 bytes	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.bootstraps.tre	Today, 1:26 PM	52 KB	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.log.txt	Today, 1:26 PM	11 KB	Pl
	ranomafana_lemurs_cytochromeB_align.raxml.rba	Today, 1:26 PM	8 KB	De
	ranomafana_lemurs_cytochromeB_align.raxml.startTree.tre	Today, 1:26 PM	507 bytes	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.support.tre	Today, 1:26 PM	590 bytes	Fi
	RAxML_GUI_Settings_ranomafana_lemurs_cytochromeB_align.txt	Today, 1:26 PM	566 bytes	Pl

We're interested in the .support.tre file

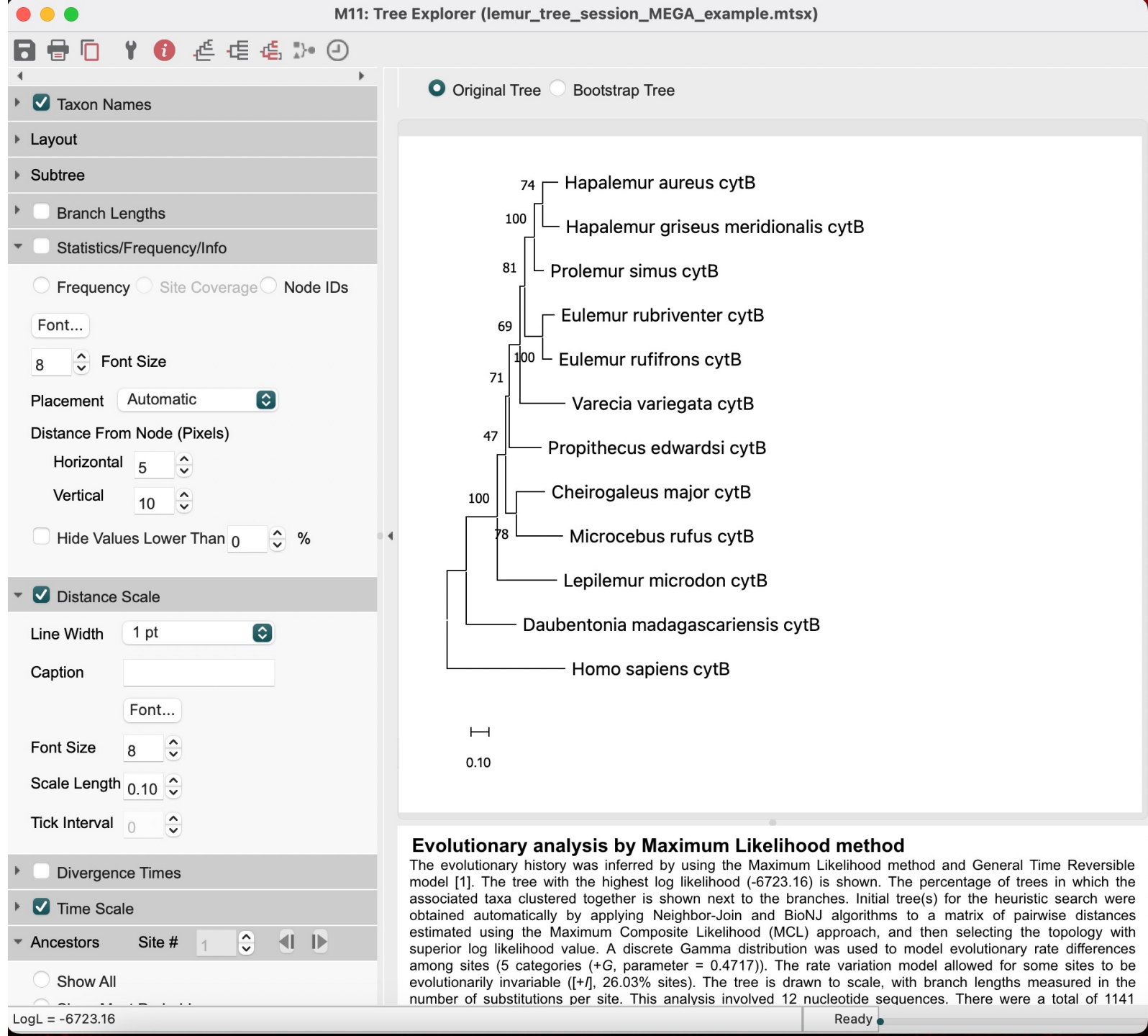
This will include the tree topology and bootstrap support values, open in MEGA



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 MEGA too...it's just more limited

We could have
done modeltest
and RAxML in
MEGA too...but
takes forever!



Then make pretty in R!

- Follow instructions in lemur_tree_editing.R file