

Conservation Genetics in the Tropics

Phylogenetics

Miguel Camacho Sánchez

October 25th, 2021

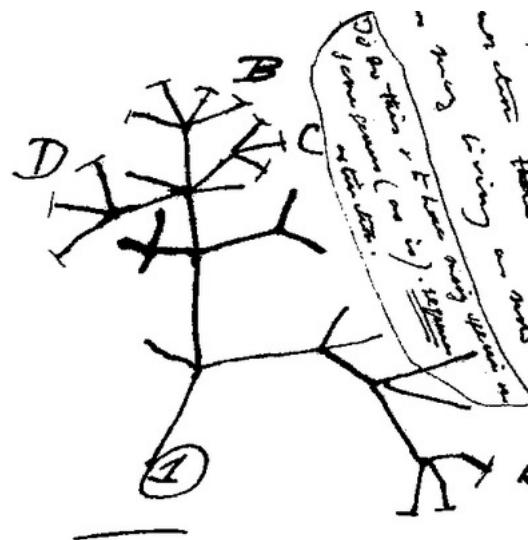
miguelcamachosanchez@gmail.com

miguelcamachosanchez.weebly.com

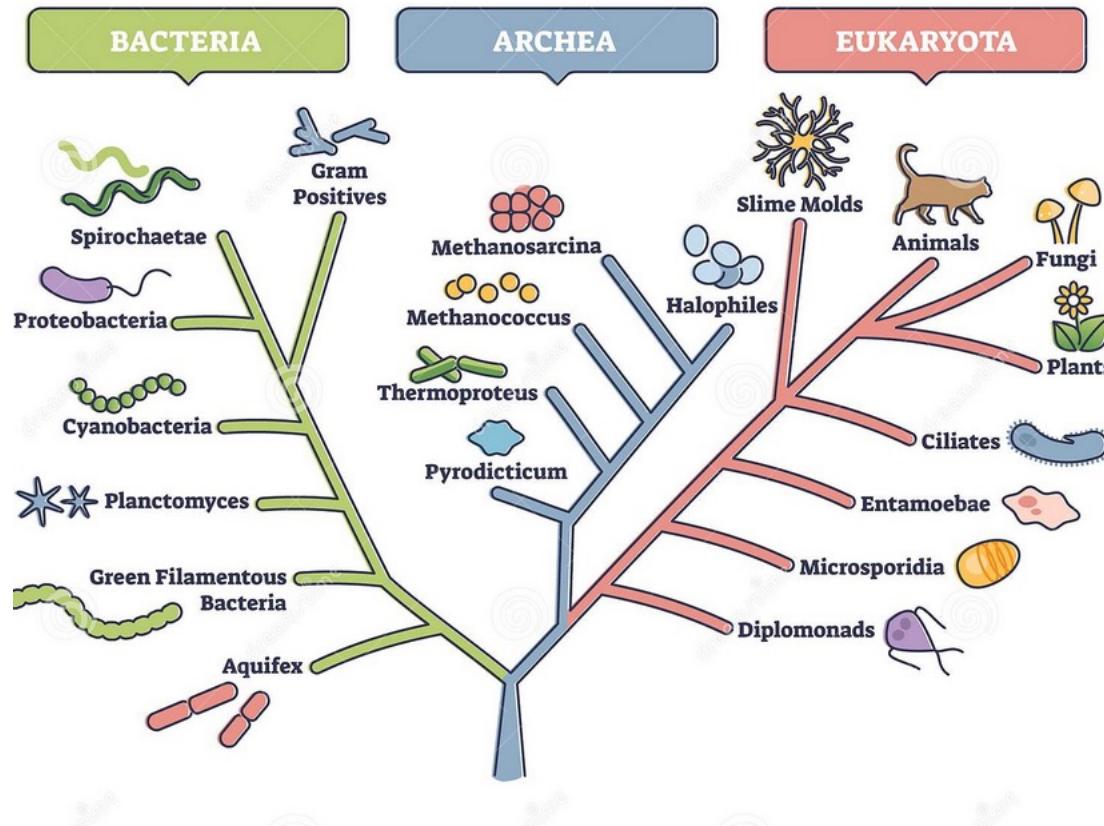


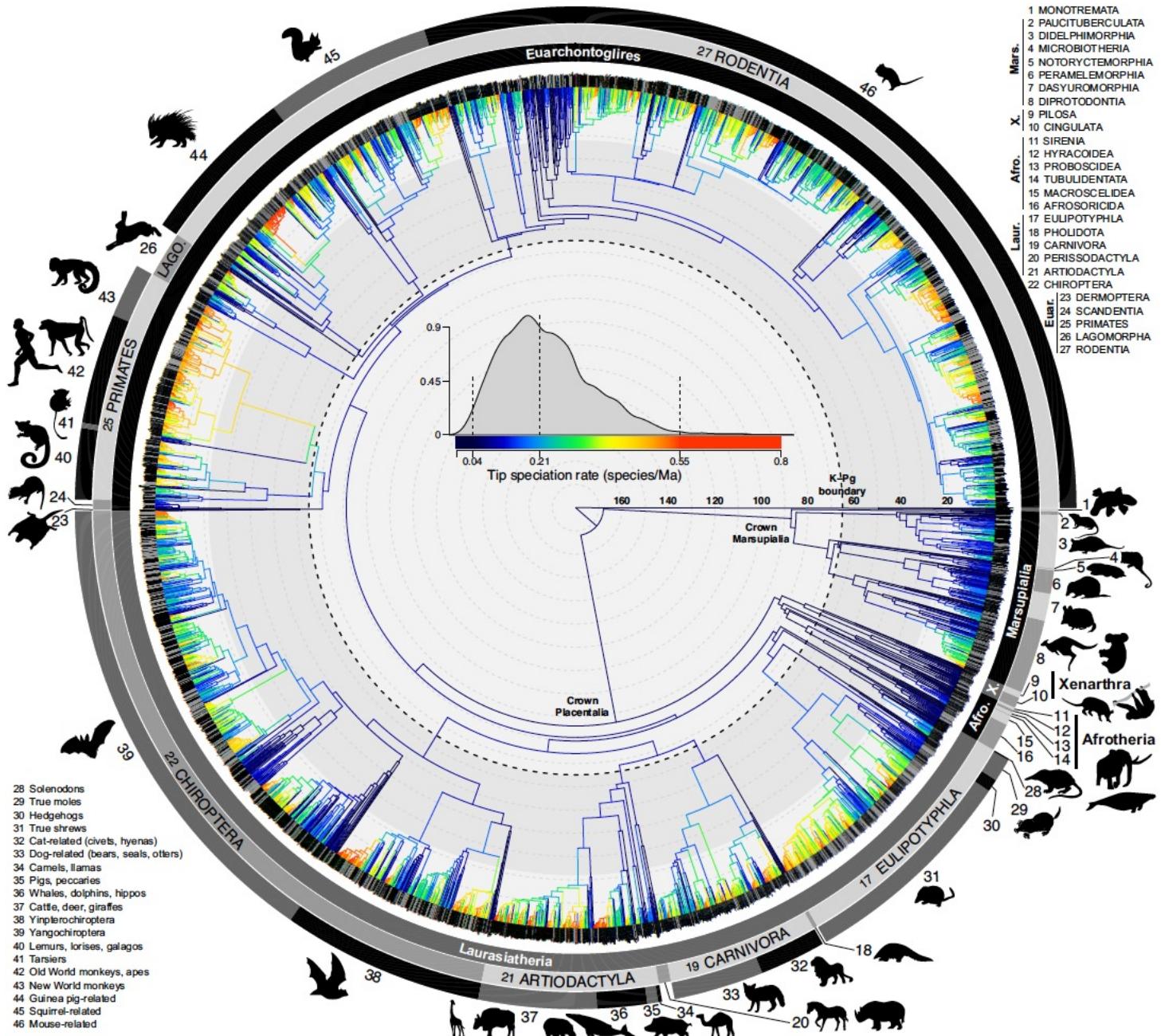
Phylogenetics

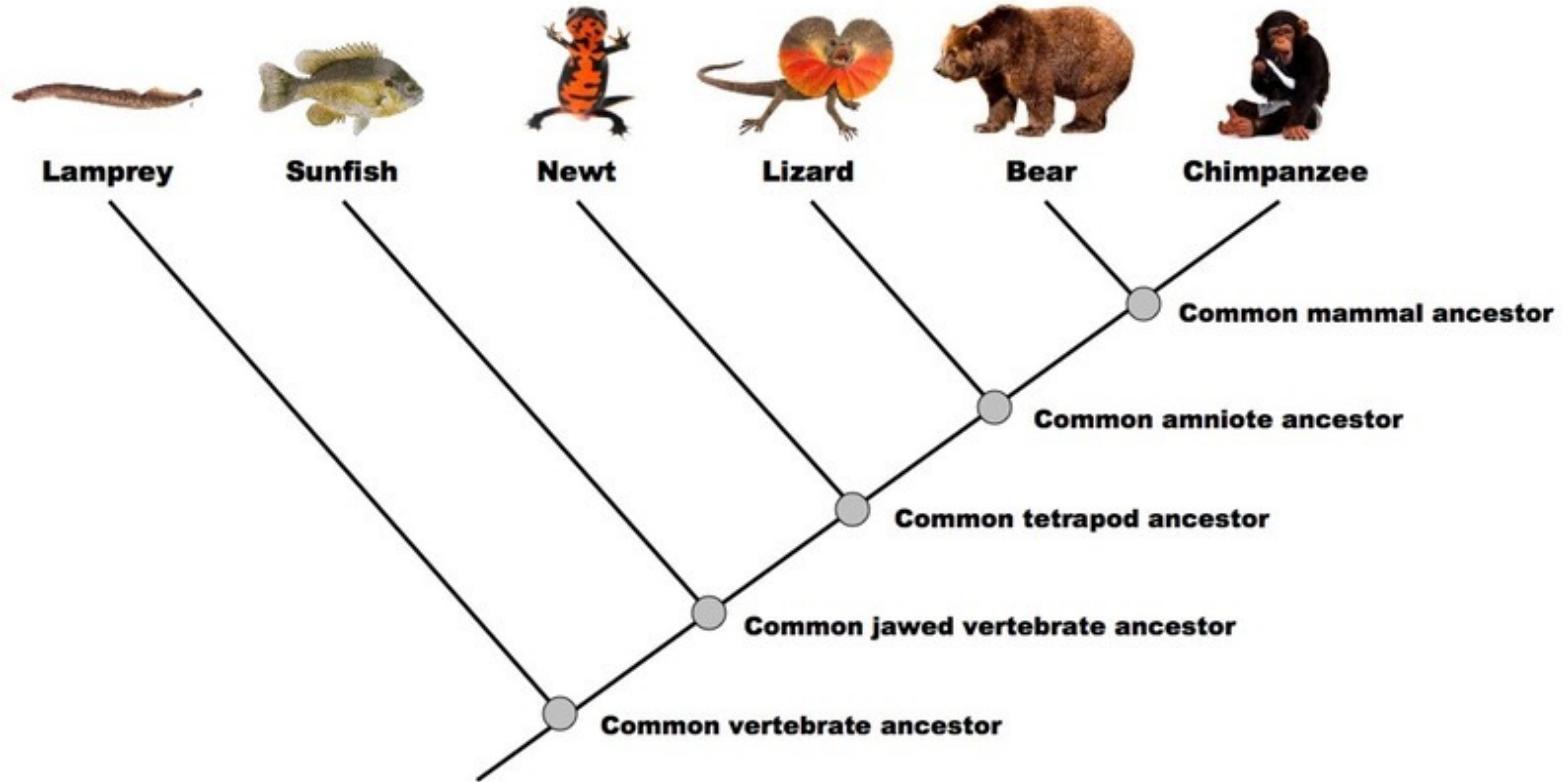
- study of evolutionary relationships.
- turns DNA/protein sequence data into a diagram which represents the evolutionary relationships between the species.



Examples

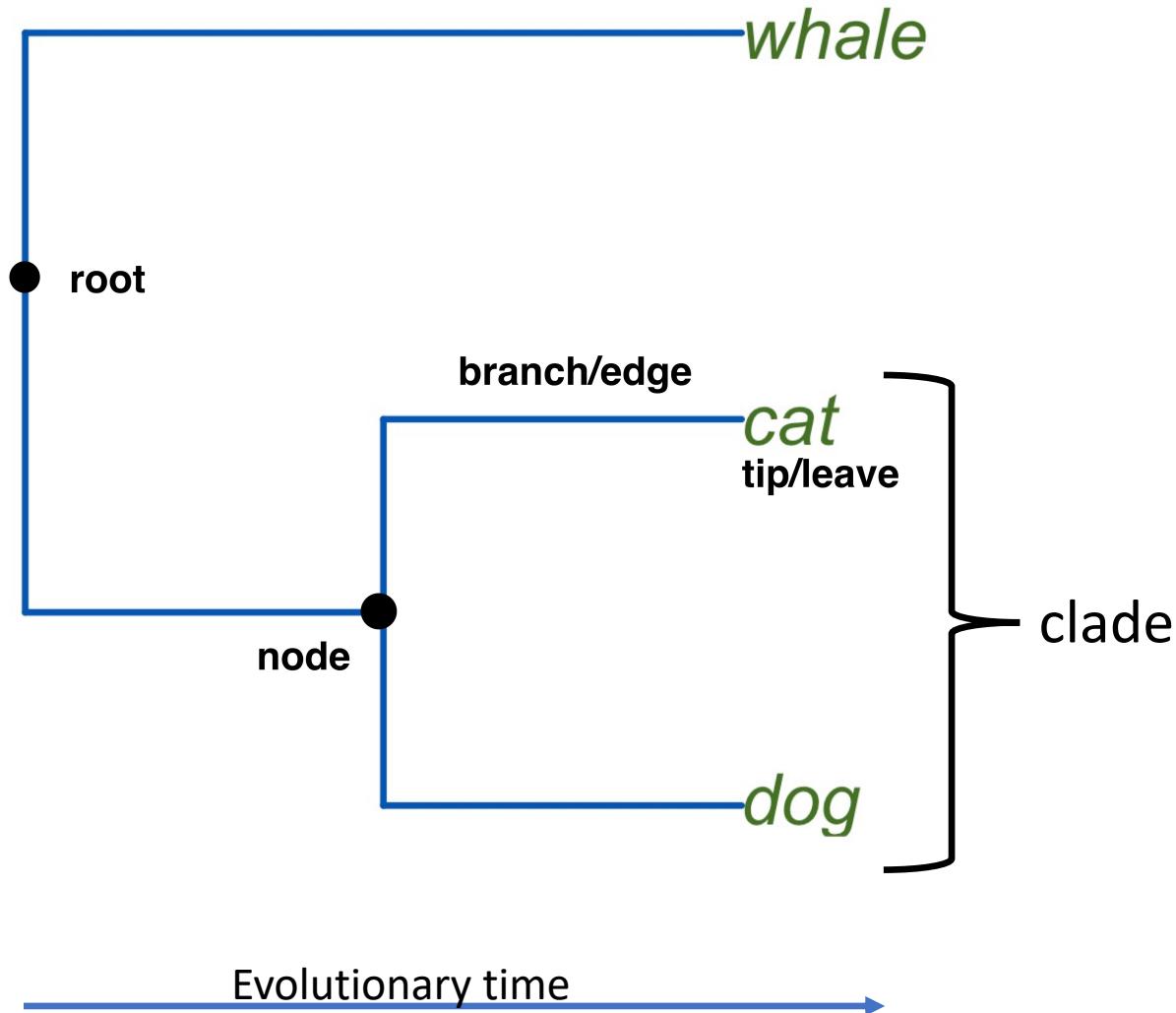




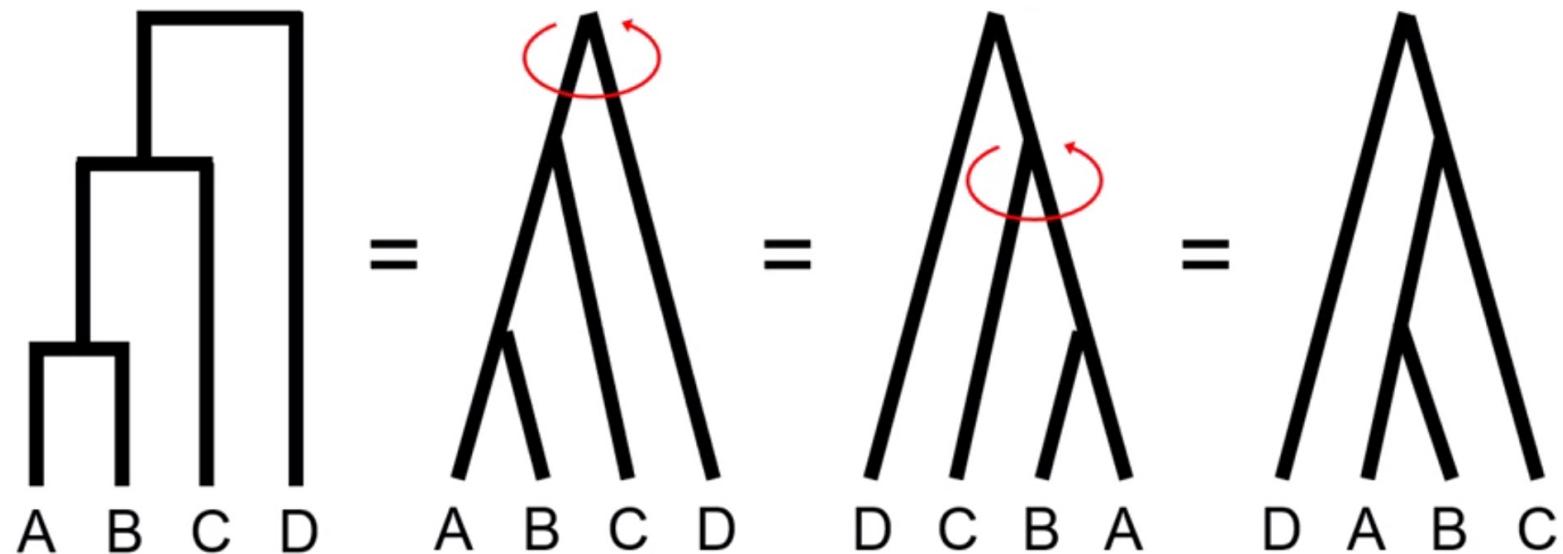


Parts of a tree

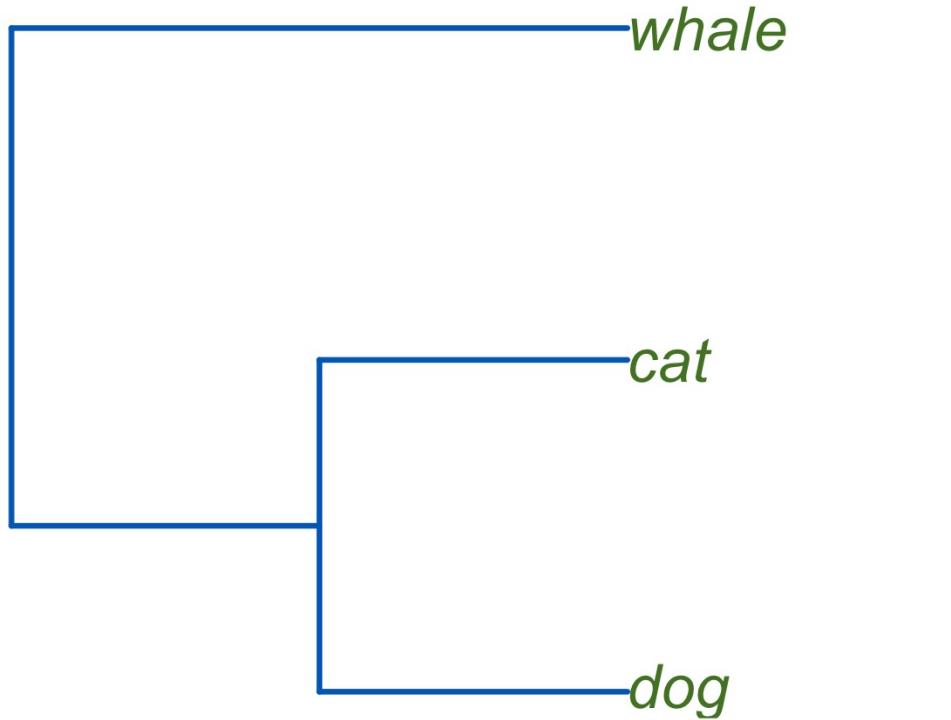
Topology = shape
Leave = terminal node



Ways of representing a tree



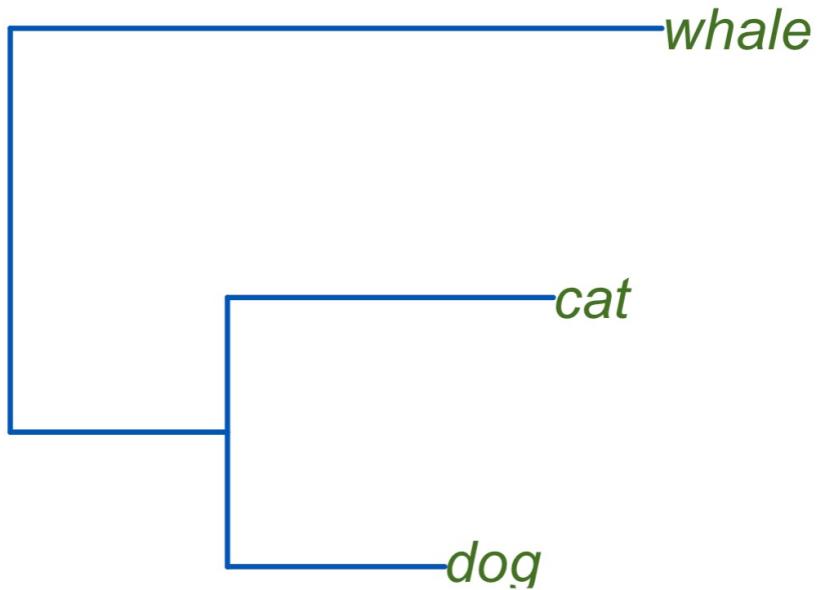
Newick format



- tips

((dog,cat),whale);

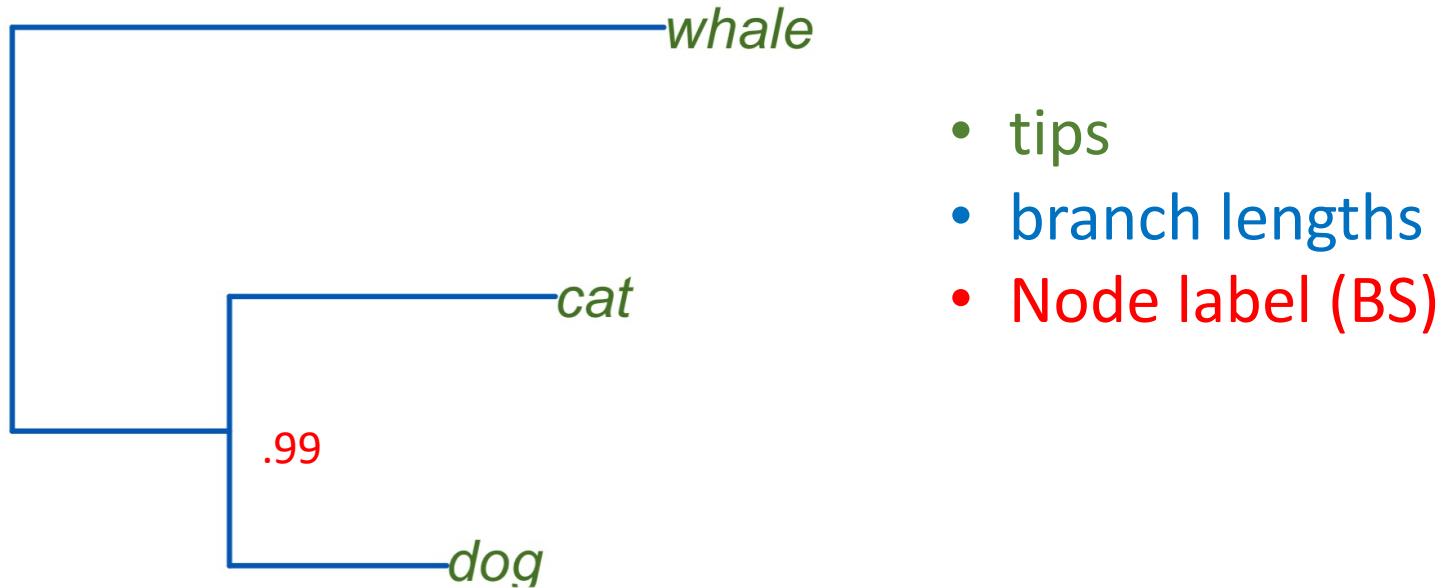
Newick format



- tips
- branch lengths

```
((dog:1,cat:1.5):1,whale:3);
```

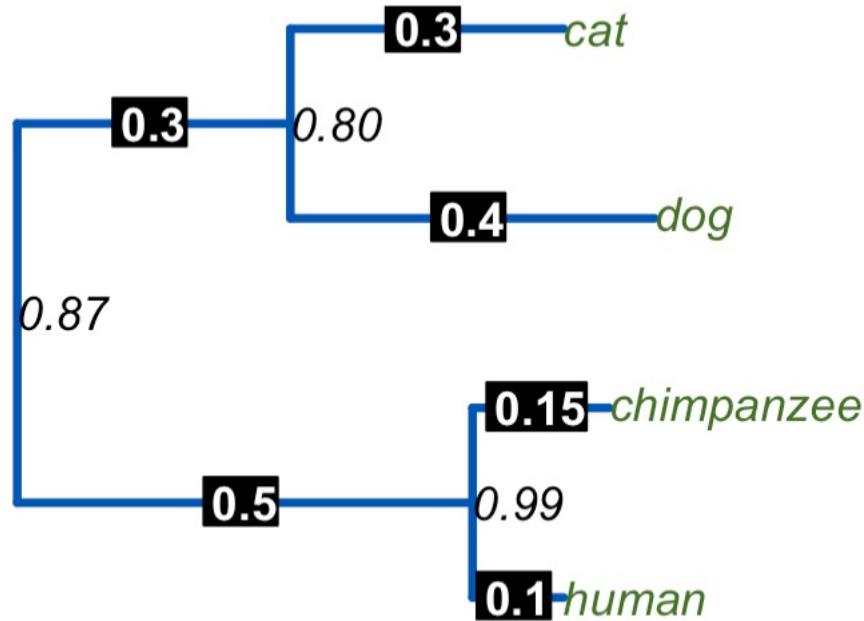
Newick format



```
((dog:1,cat:1.5).99:1,whale:3);
```

Exercise1: tree format

Goal: write in newick format



You can use:

<http://etetoolkit.org/treeview/>

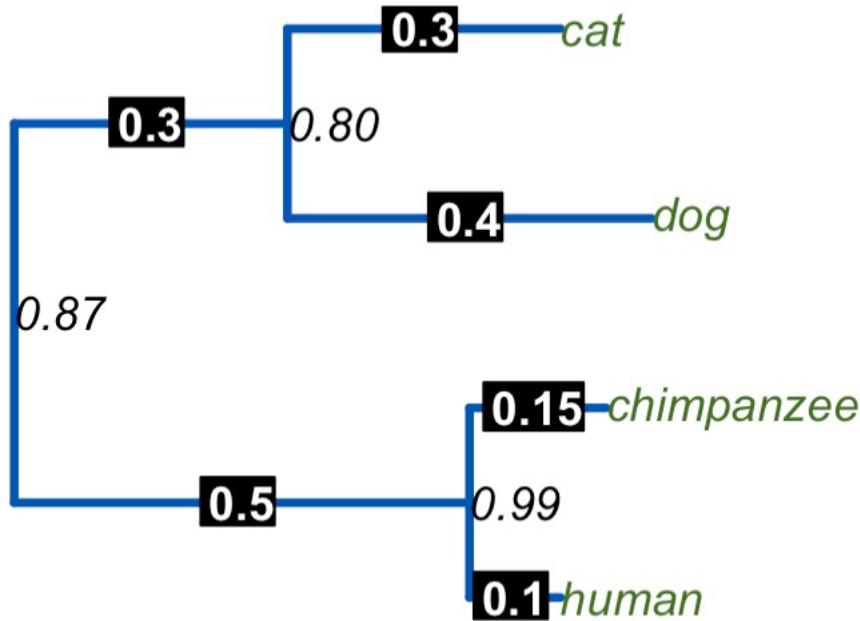
or

R

Tip:

((dog:1,cat:1.5).99:1,whale:3);

Exercise1: tree format

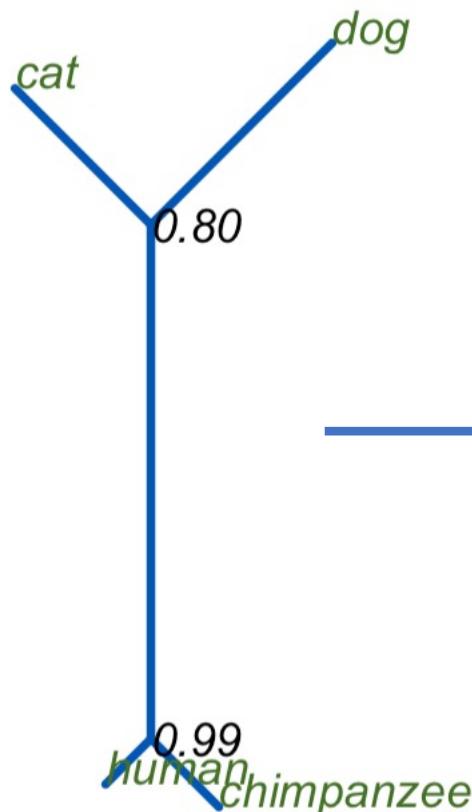


((human:0.1, chimpanzee:0.15)0.99:0.5,(dog:0.4, cat:0.3)0.80:0.3)0.87;

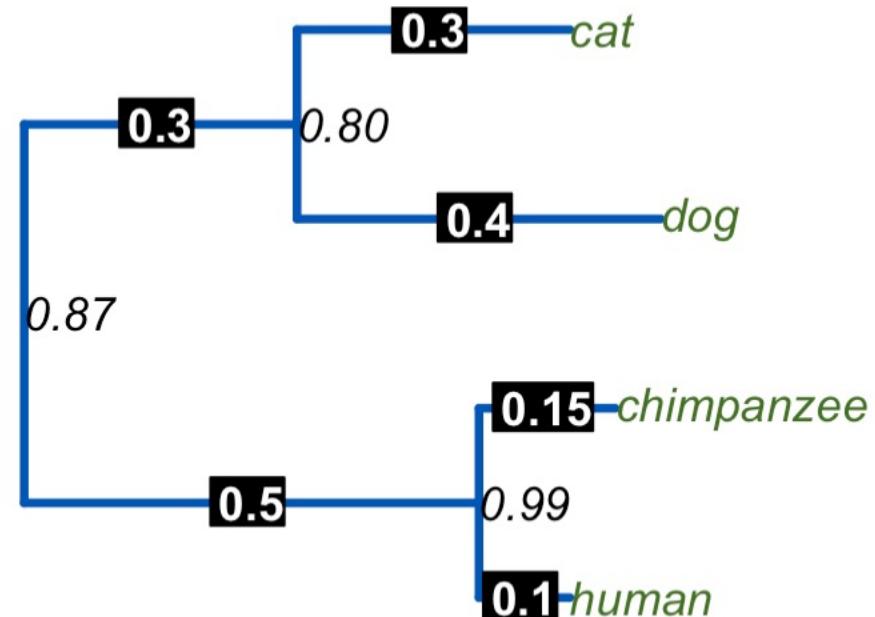
https://github.com/csmiguel/2021_ConGenTopics/tutorials/phylogenetics#Exercise1

Rooting

Implies giving a directionality to the changes

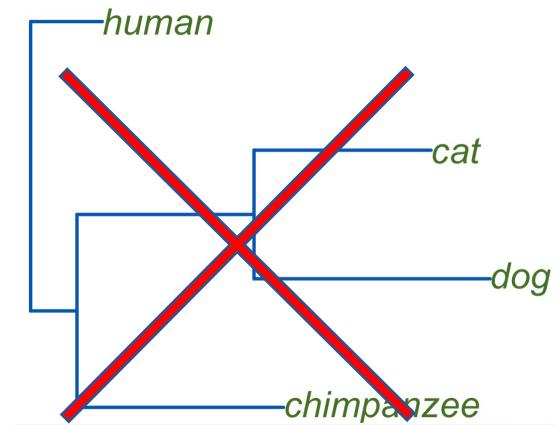
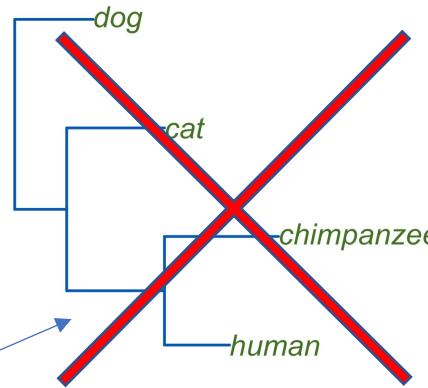
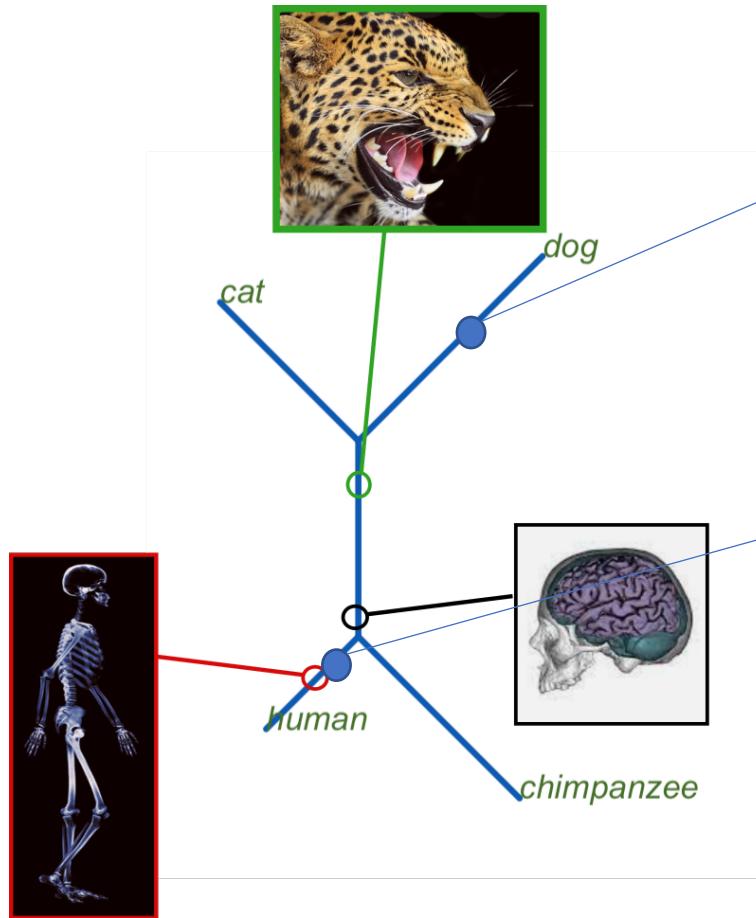


Unrooted tree



rooted tree

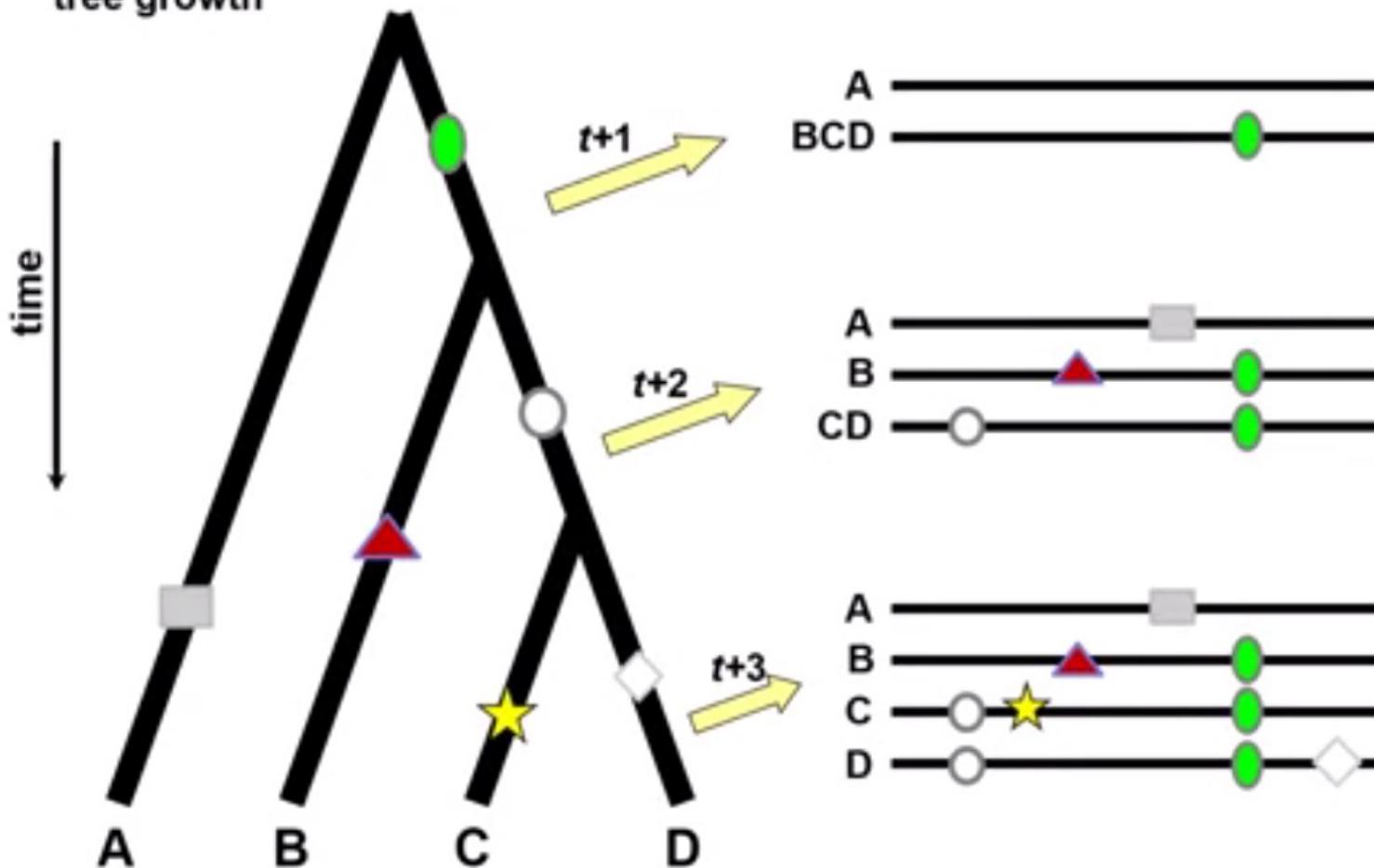
Why the root is important?



Directionality to the evolutionary changes

Phylogenetics

tree growth



Phylogenetic inference workflow

Multiple sequence alignment

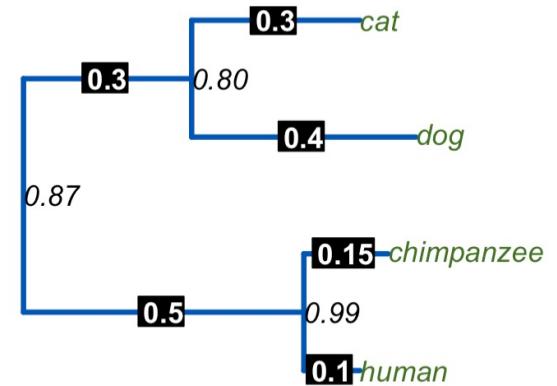
C	T	T	A	G	A	T	C	G	T	A	C
C	T	T	A	G	A	T	C	G	T	A	C
A	T	T	A	G	A	T	C	G	T	A	C
C	T	T	A	G	A	T	C	G	T	T	C

Data selection

Inference

Phylogenetics
methods

Tree

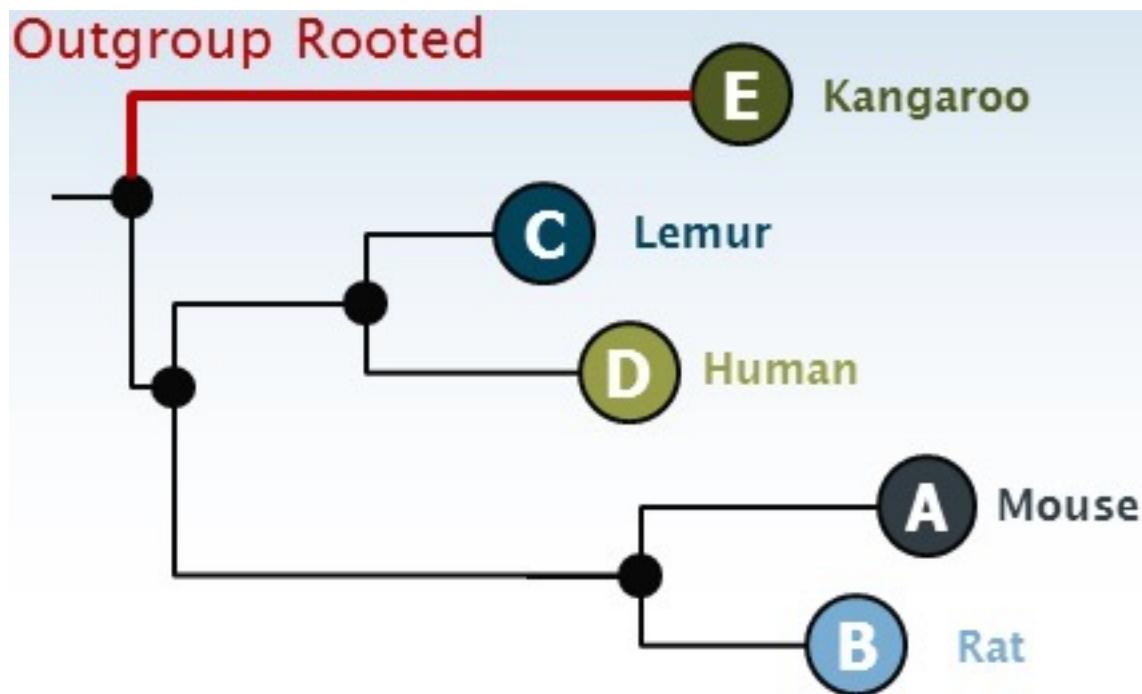


Data for phylogenetics

- Taxa taxonomic sampling
- Loci genetic sampling

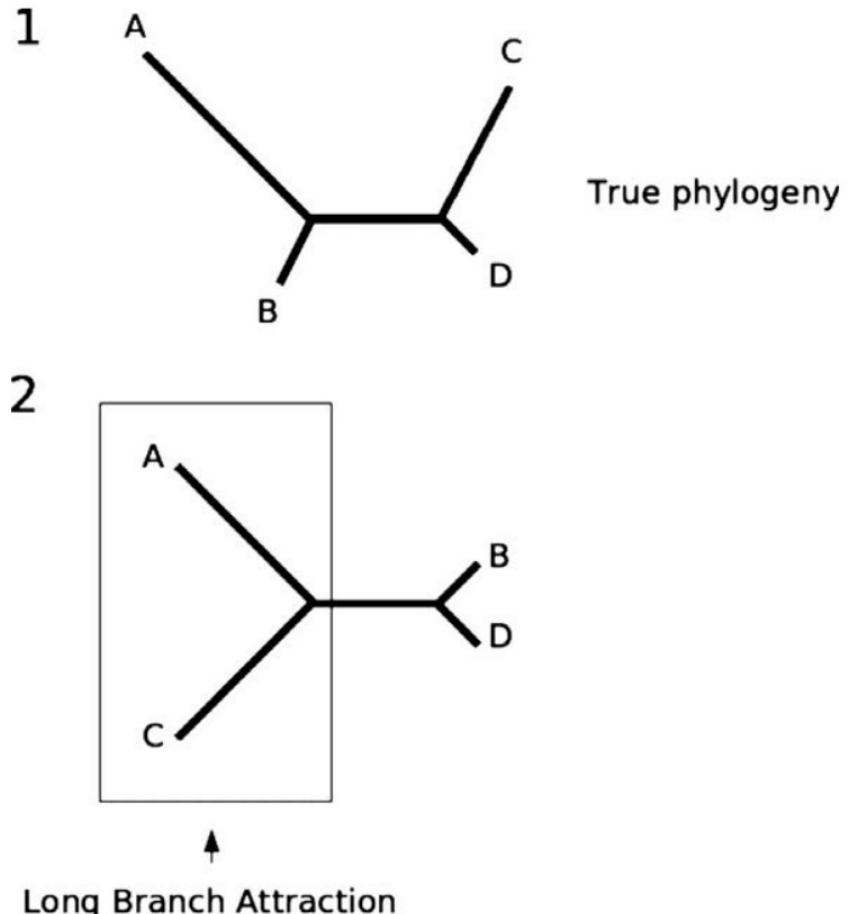
Taxa

- Ingroup: study species.
- outgroups: close but in another clade



Taxa

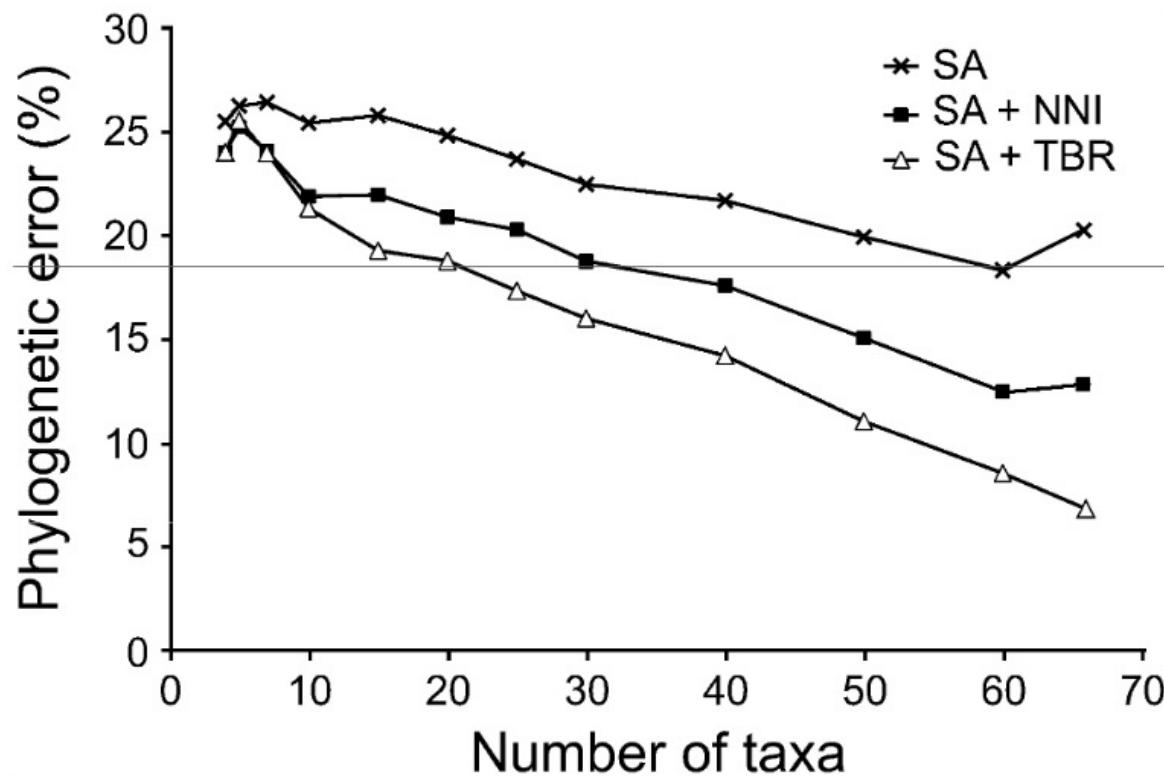
- Large taxonomic sampling prevents from long branch attraction
([10.1080/10635150500234583](https://doi.org/10.1080/10635150500234583))



10.1016/j.shpsc.2008.09.002

Taxa

Thorough taxon sampling is one of the most practical ways to improve the accuracy of phylogenetic estimates (10.3724/SP.J.1002.2008.08016)

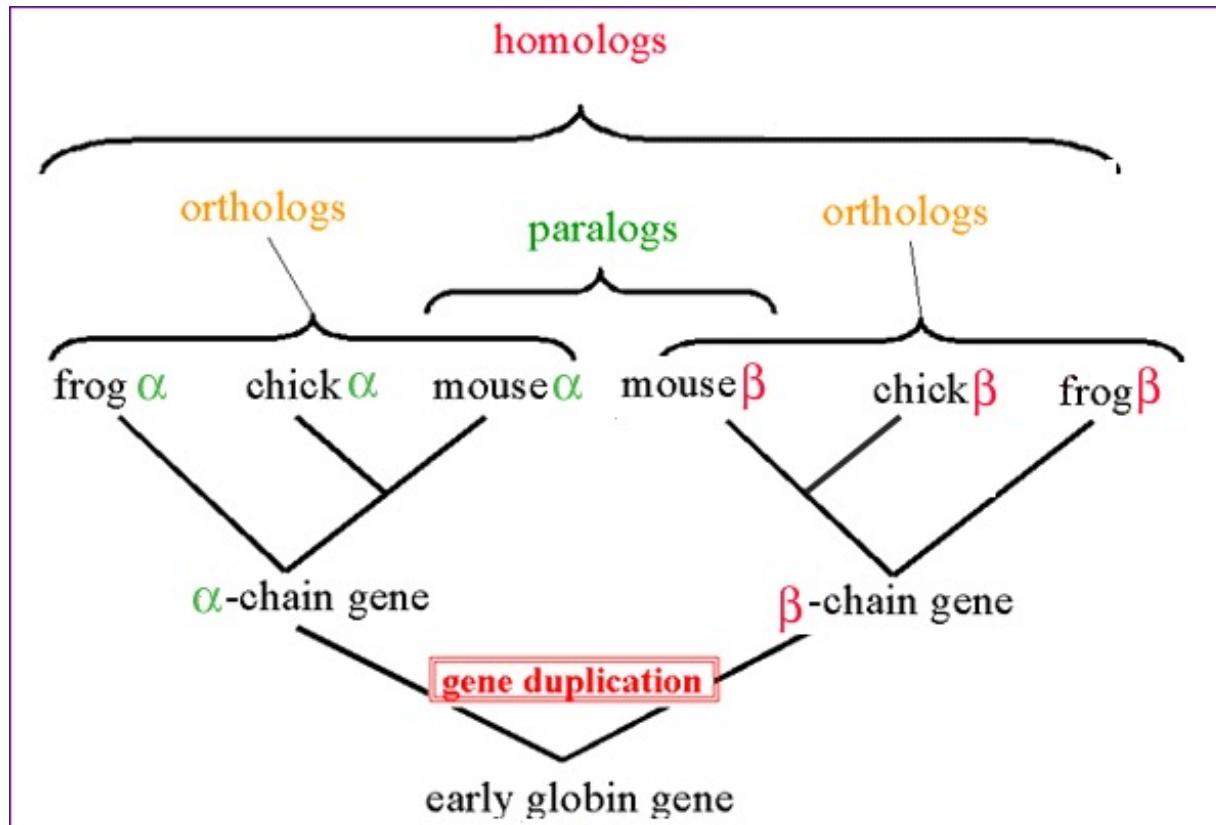


Loci: genetic markers

- Homologous
- Variable
- Independent

Loci: homology - common ancestor

Use single-copy genes



Loci: variable

Moderate variability for the phylogenetic scale of study

14. Rattu... AGTACGAGCATTACCCCTCGCTTCCG
15. Rattu... AGTACGAGCATTACCCCTCGCTTCCG
16. Rattu... AGTACGAGCATTACCCCTCGCTTCCG
17. Rattu... AGTACGAGCATTACCCCTCGCTTCCG

Too conserved: no variation

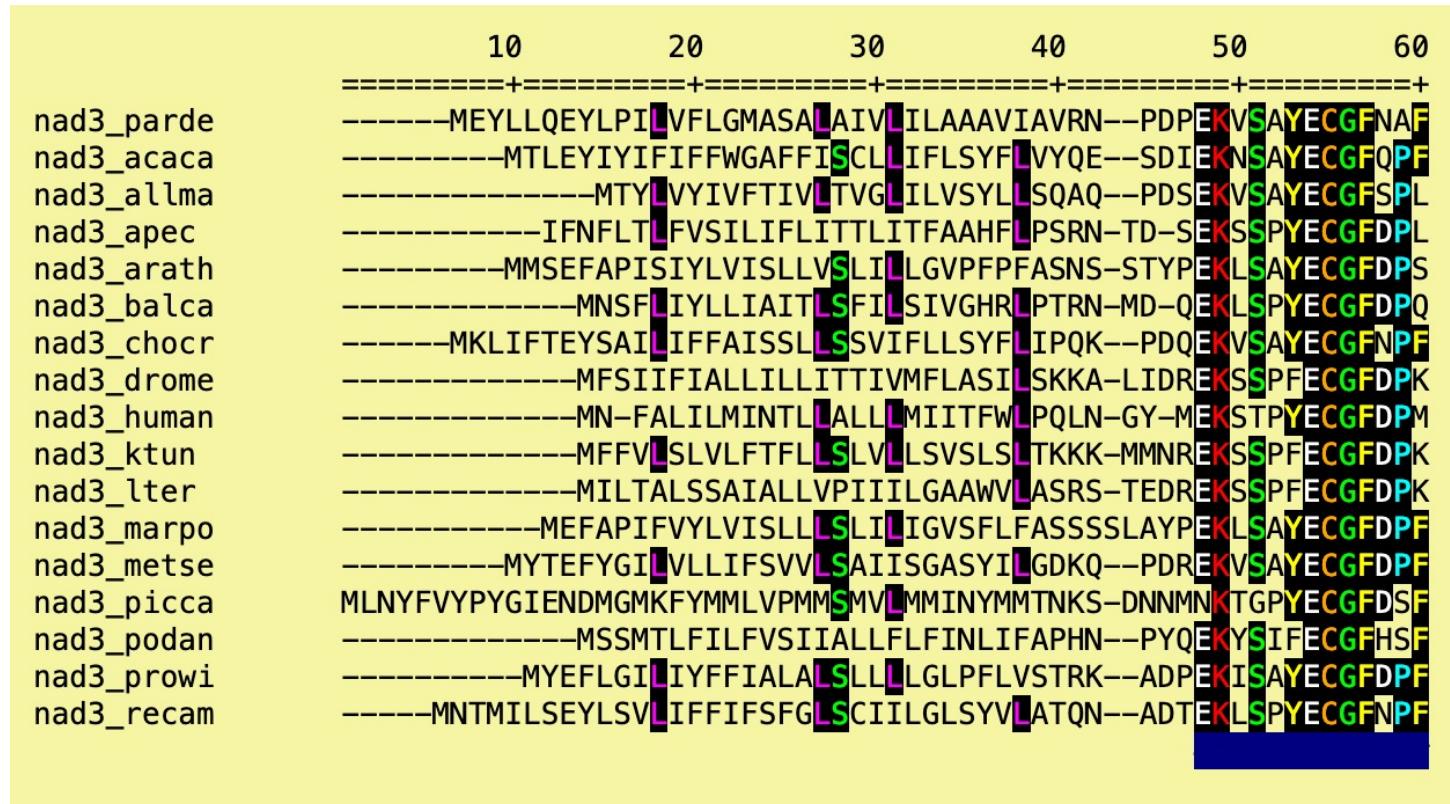
Too variable: difficult to find homology

1. DQ... TTTTCTAAGGG--TATTCAGGGAAAGAGGTC
2. mc... TCCCTAGGGCA---CTCAAAGGAAGAAAGAA
3. mc... TTCTTAGAGCA--AACTTCAGGGAAGAAAC
4. mc... TTCTTAGGGCA--AACTTCAGGGAAGAAAC
5. mc... TTCTTAGAGCA--AAC-TCAGGGAAGAAAC
6. mc... TTCTTAAGCA--AAACTTCAGGGAAGAAAT
7. mc... TTCTTAGAGCA--CCCCTCAGGGAAGAAAT
8. mc... TCCCTGAGACA-----TCAGAGAAGAGGC
9. mc... CTCTAGAACCA--AATCAAAGAACGAC

Right amount of variation

1. DQ316... GGACGCTTAGATCTATCGGCGACTGGCACCGGTCCTCT
2. mc105... GGACGCTTAGCTTATCGGCGACTGGCACCGGGCCTCT
3. mc213... GGACGCCCTAGCTTATCGGCGATTGGCACCGGACTTCT
4. mc458... GGACGCCCTAGCTTATCGGCGATTGGCACCGGACTTT
5. mc459... GGACGCCCTAGCTTATCGGCGACTGGCACCGGGCCTCT
6. mc453... GGACGCCCTAGCTTATCGGCGACTGGCACCGGACCTTT
7. mc708... GGACGCCCTAGTTTATCGGCGACTGGCACCGGGCCTCT
8. mc716... GGACGCCCTAGCTTATCGGCGATTGGCACCGGACCTCT
9. mc699... GGACGCTTAGCTTATCGGCGACTGGCACCGGGCCTTAT
10. mc49... GGACGCTTAGATTATCGGCGACTGGCACCGGGCCTCT
11. mc49... GGACGCTTAGATTATCGGCGACTGGCACCGGGCCTCT
12. mc74... GCAAGGGCTAGATTATCGGCGACTGGCACCGGGCCTCT

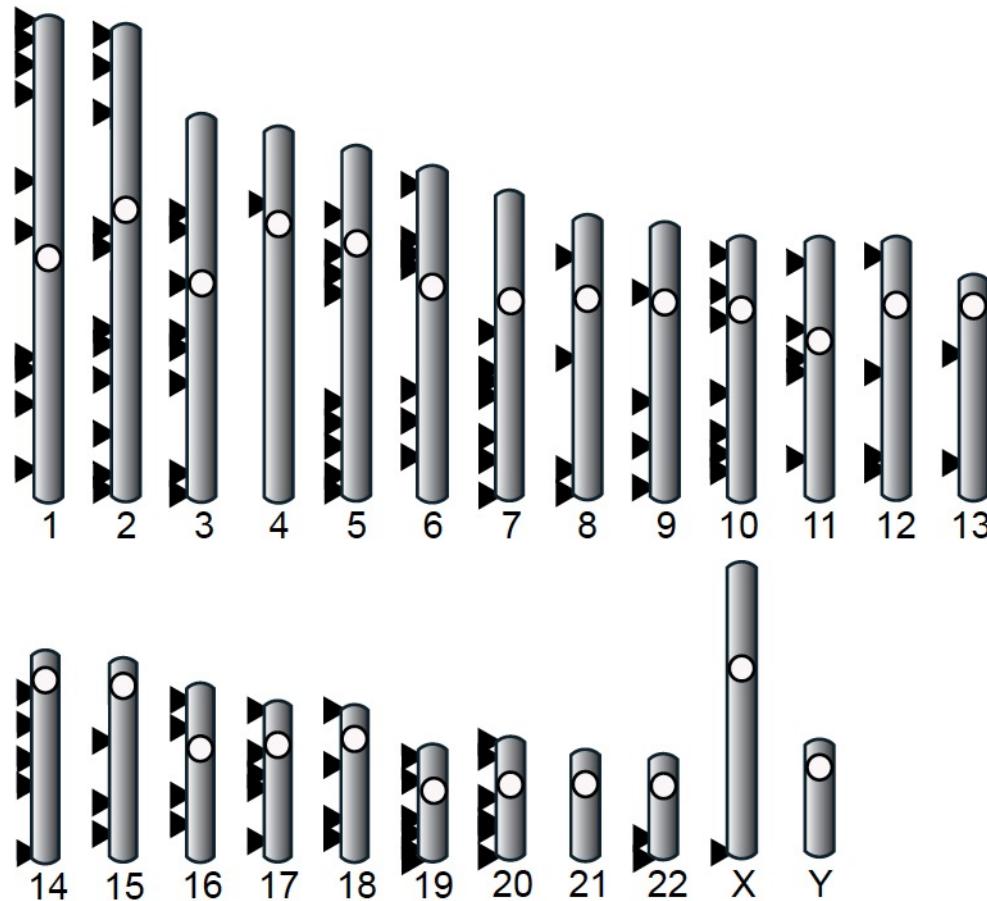
Loci: homology in regions



Software to clean non-homologous regions: gblocks, divvier, trimal

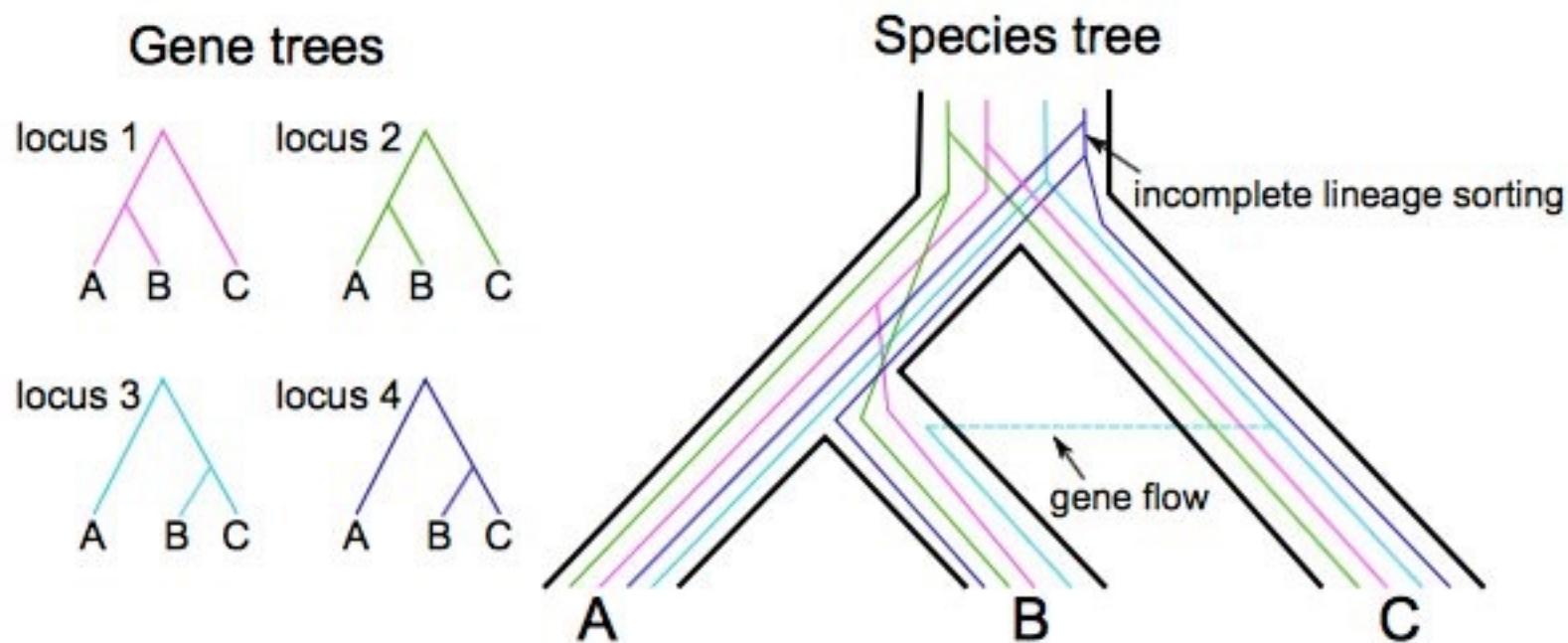
Loci: independent

- Independent => recombination
(eg, different chromosomes)



Loci: independent

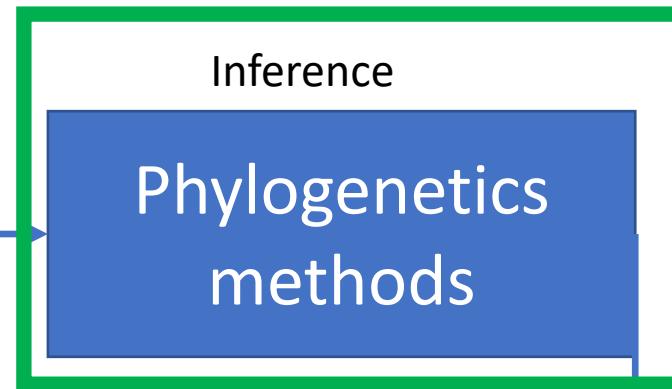
- Many **independent** loci to overcome bias by incomplete lineage sorting



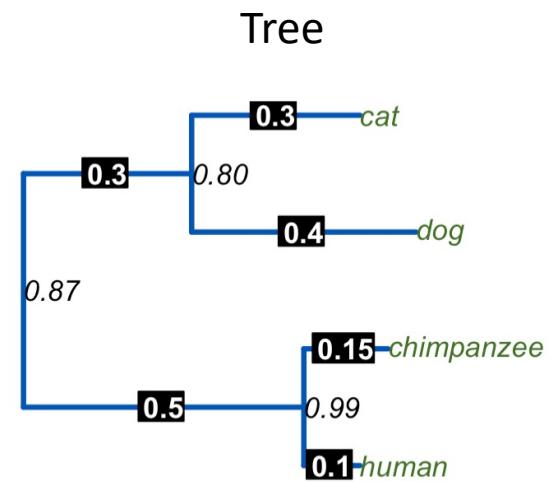
Phylogenetic inference workflow

Multiple sequence alignment

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



Choose method



Phylogenetics methods

- Distance-based (eg. neighbour joining)
- Maximum likelihood
- Bayesian

Advantages

Disadvantages

Distance methods

- Fast computational speed
- Can be applied to any type of data as long as a genetic distance can be defined
- Models for distance calculation can be chosen to fit data

- Most distance methods, such as neighbour joining, do not consider variances of distance estimates
- Distance calculation is problematic when sequences are divergent and involve many alignment gaps
- Negative branch lengths are not meaningful

Likelihood methods

- Can use complex substitution models to approach biological reality
- Powerful framework for estimating parameters and testing hypotheses

- Maximum likelihood iteration involves heavy computation
- The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret

Bayesian methods

- Can use realistic substitution models, as in maximum likelihood
- Prior probability allows the incorporation of information or expert knowledge
- Posterior probabilities for trees and clades have easy interpretations

- Markov chain Monte Carlo (MCMC) involves heavy computation
- In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify
- Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge
- Posterior probabilities often appear too high
- Model selection involves challenging computation^{138,139}

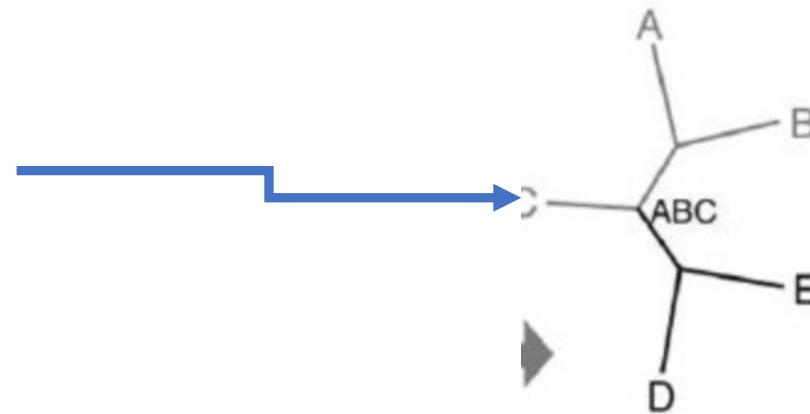
Distance based: Neighbour joining

1. DQ316...	GGACGCTT	TAGATC	TATCGGCGACT	TGGCACCGGT	CCTCTC
2. mc105...	GGACGCTT	TAGCTT	TATCGGCGACT	TGGCACCGGG	CCTCTC
3. mc213...	GGACGCC	TAGCTT	TATCGGCGATT	TGGCACCGGA	CTTCTC
4. mc458...	GGACGCC	TAGCTT	TATCGGCGATT	TGGCACCGGA	CTTCTC
5. mc459...	GGACGCC	TAGCTC	TATCGGCGACT	TGGCACCGGG	CTTCTC
5. mc453...	GGACGCC	TAGCTC	TATCGGCGACT	TGGCACCGGA	CCTTTT
7. mc708...	GGACGCC	TAGTTT	TATCGGCGACT	TGGCACCGGG	CTTCTC
3. mc716...	GGACGCC	TAGCTC	TATCGGCGATT	TGGCACCGGA	CCTCTC
9. mc699...	GGACGCT	TAGCTC	TATCGGCGACT	TGGCACCGGG	CTTATC
10. mc49...	GGACGCT	TAGATT	TATCGGCGACT	TGGCACCGGG	CTTCTC
11. mc49...	GGACGCT	TAGATT	TATCGGCGACT	TGGCACCGGG	CTTCTC
12. mc49...	GGACGCT	TAGATT	TATCGGCGACT	TGGCACCGGG	CTTCTC

Character matrix

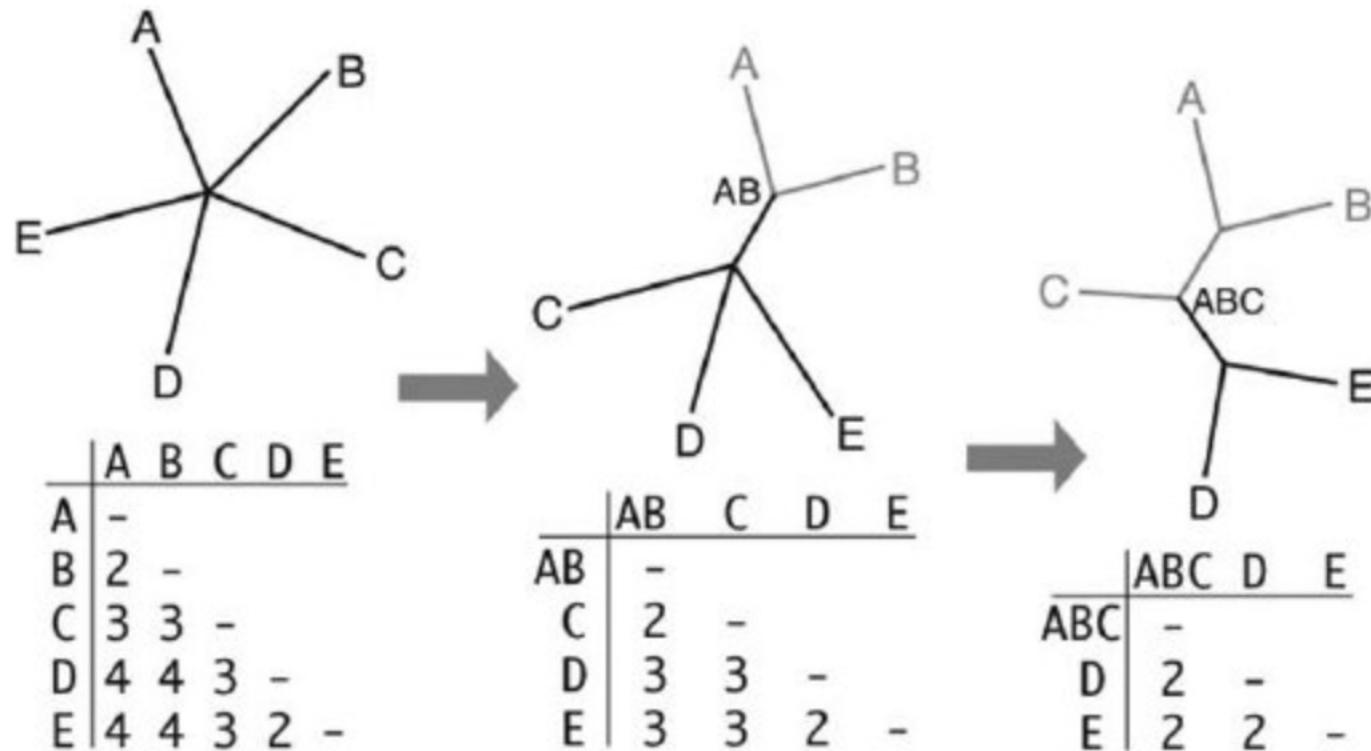
	A	B	C	D	E
A	-				
B	2	-			
C	3	3	-		
D	4	4	3	-	
E	4	4	3	2	-

Distance matrix



Tree

Neighbour joining



Software: R (ape), MEGA, Geneious

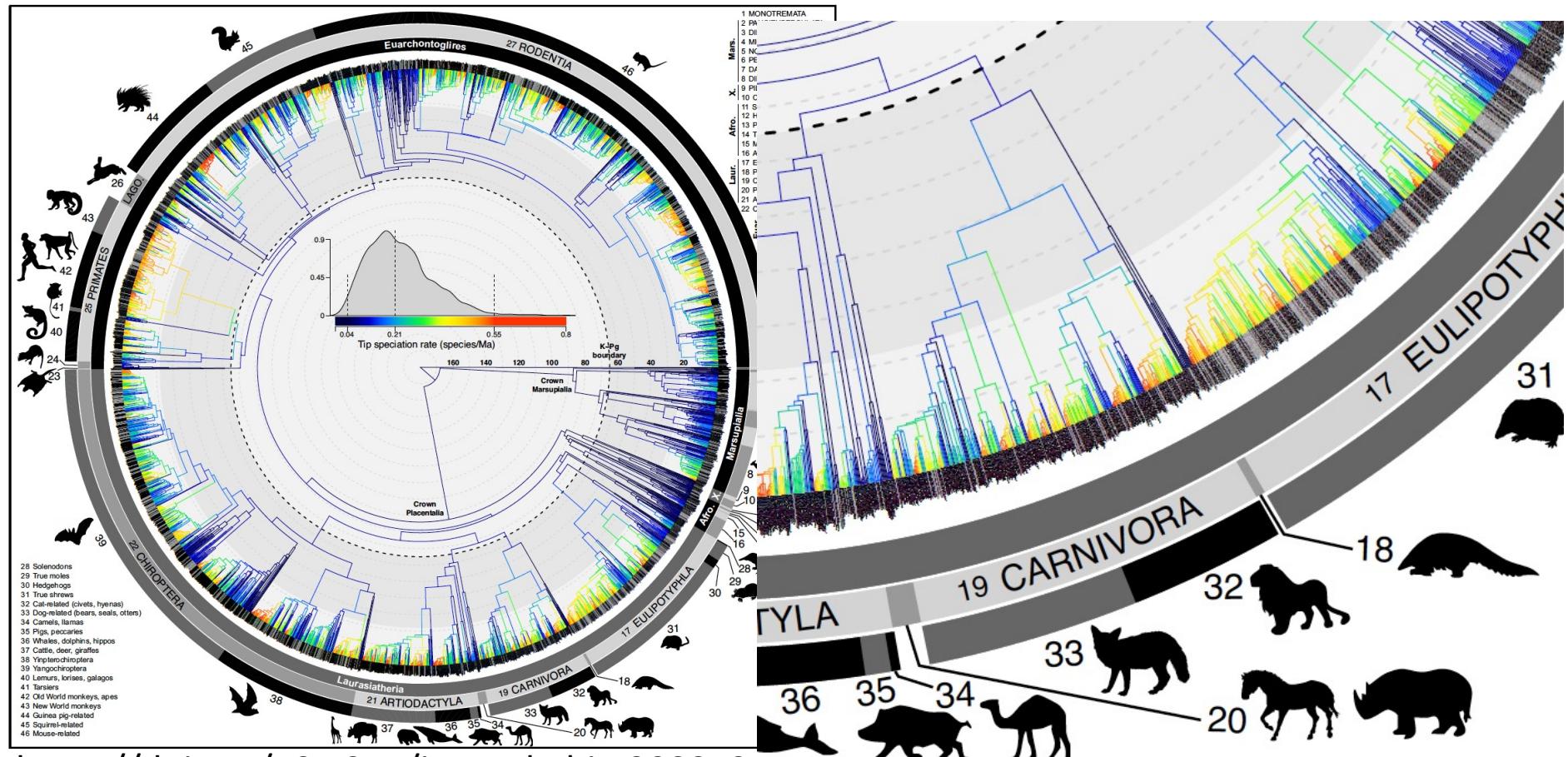
Exercise 2

Neighbour joining tree reconstruction

https://github.com/csmiguel/2021_ConGenTopics/blob/main/tutorials/phylogenetics

Exercise 2: dataset

Cytochrome b from carnivores + pangolins at outgroup



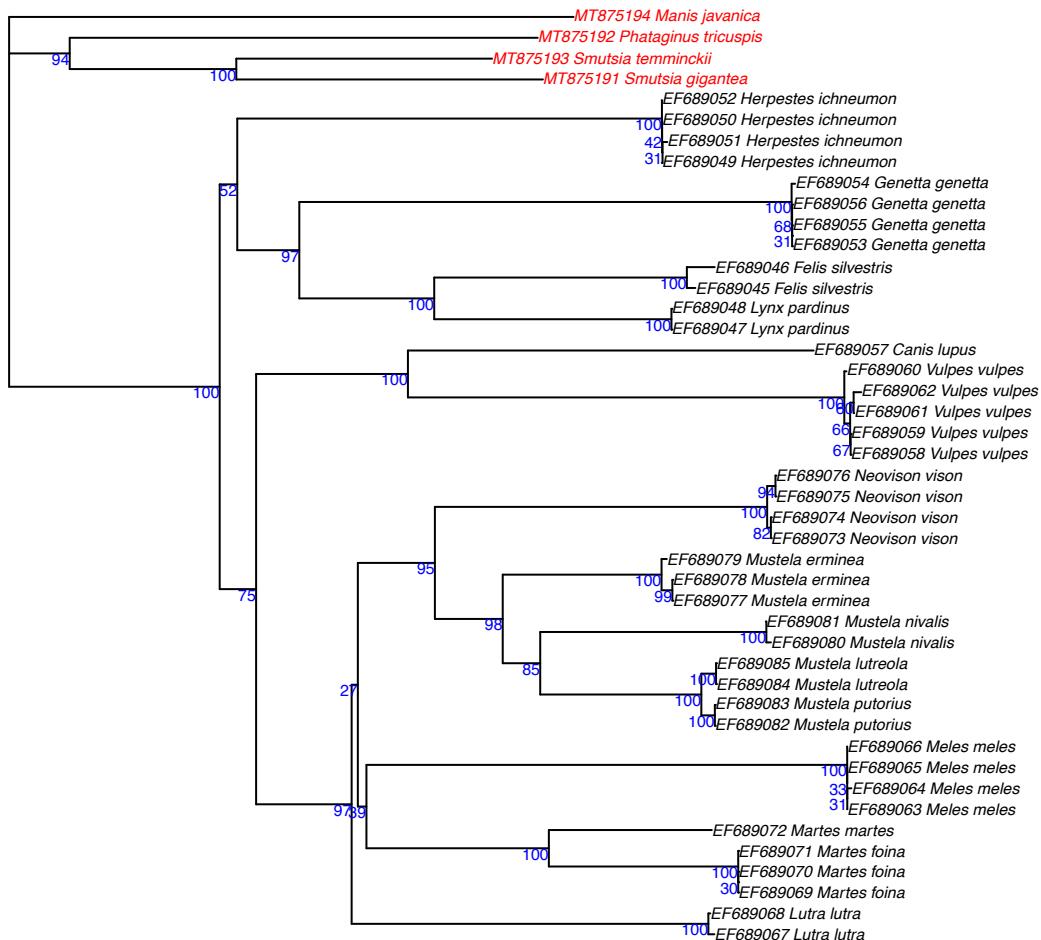
Exercise 2: dataset

Cytochrome b from carnivores + pangolins at outgroup



Exercise 2: result

Rooted NJ Tree



Phylogenetic inference workflow

Multiple sequence alignment

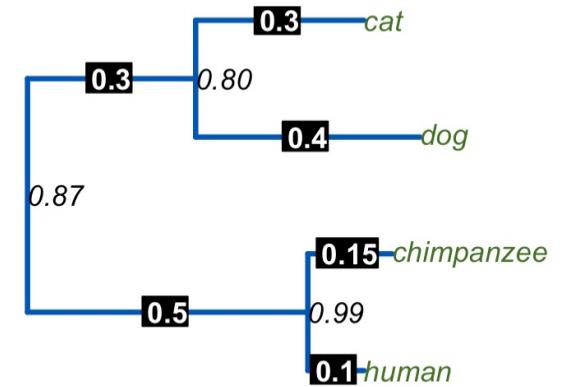
C	T	T	A	G	A	T	C	G	T	A	C
C	T	T	A	G	A	T	C	G	T	A	C
A	T	T	A	G	A	T	C	G	T	A	C
C	T	T	A	G	A	T	C	G	T	T	C

Inference

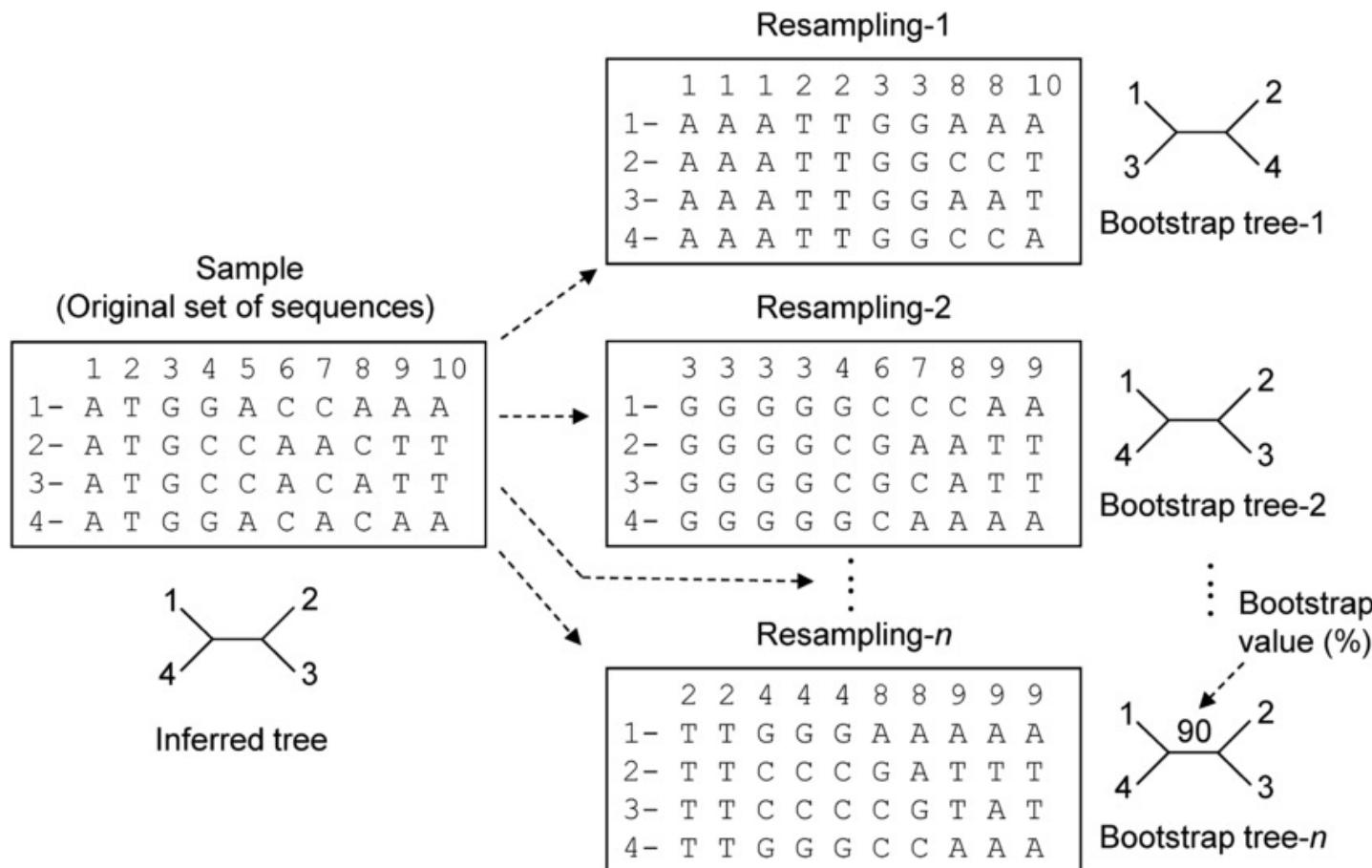
Phylogenetics
methods

- Bootstrapping
- Models of substitution

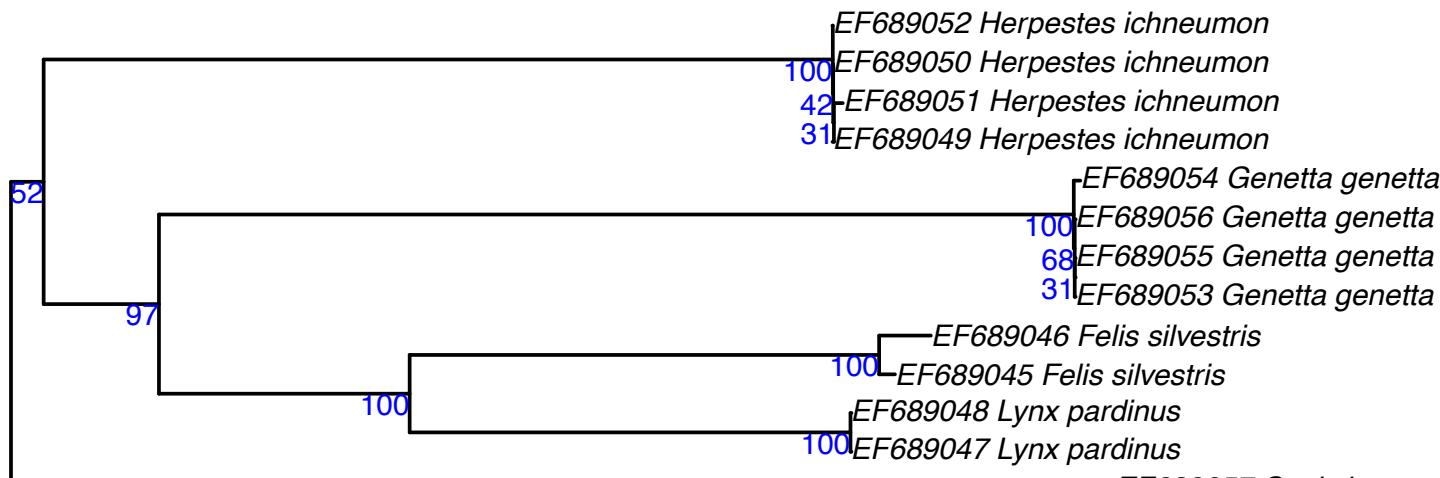
Tree



Bootstrapping = resampling with replacement



Bootstrapping



Bootstrap values

> 90% strongly supported
 70 > 90% well supported
 50 > 70% weakly supported
 < 50% not supported

(In Maximum likelihood)

Models of substitution

Model how DNA changes in evolutionary time

Elements:

- base frequencies
- *substitution probability matrix*
- rate heterogeneity among sites

Models of substitution.

the problem: changes not random

BEAR

BEER



	B	E	E	R
B	1	0	0	0
E	0	1	0.7	0
A	0	0	0.3	0
R	0	0	0	1

Rate heterogeneity

- Variation in mutation rate across sites.

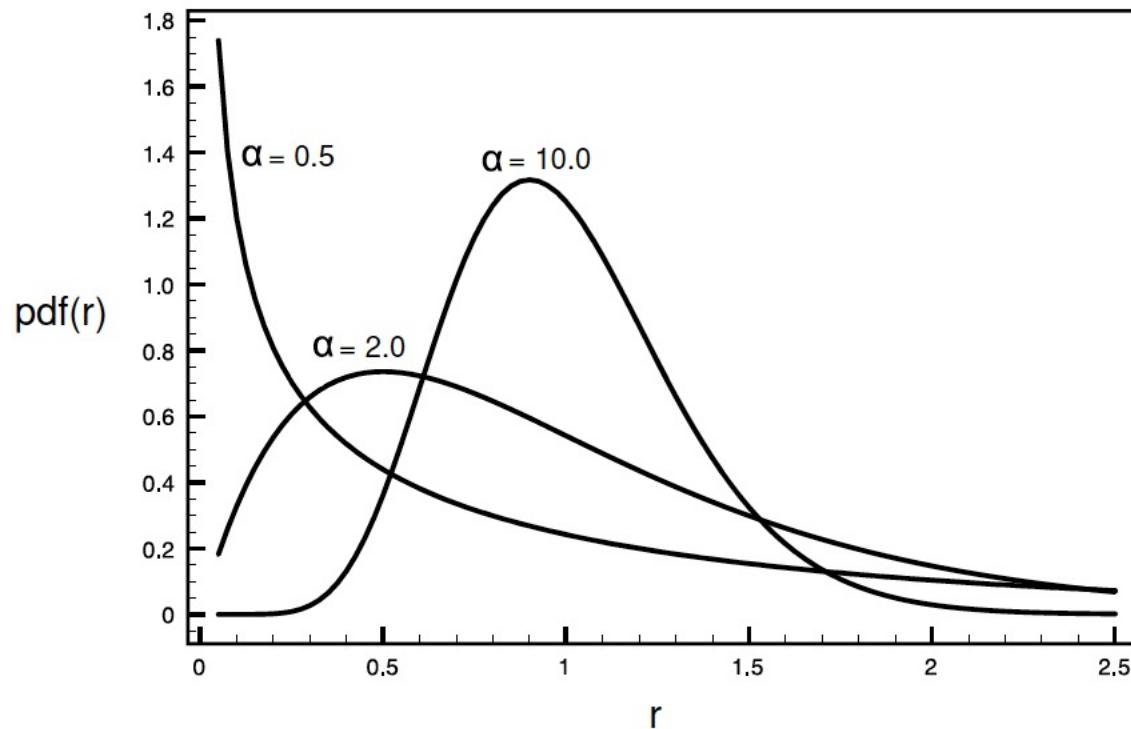
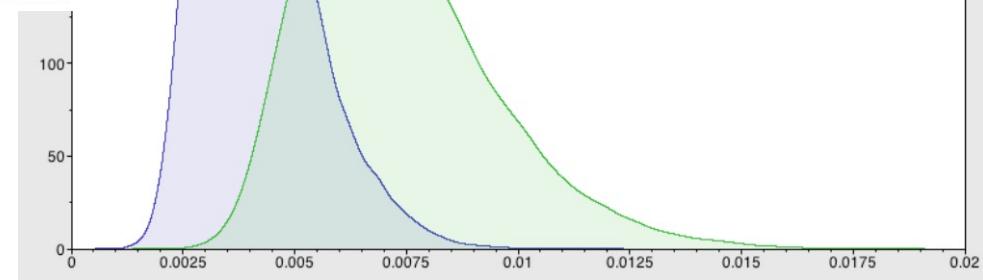


Fig. 4.8 Different shapes of the Γ -distribution depending on the α -shape parameter.

Rate heterogeneity

		Second letter							
		U	C	A	G	U	C	A	G
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Trp UGG Trp	U	C	A	G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U	C	A	G
	A	AUU } Ile AUC } AUA } Met AUG }	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Stop AGG } Stop	U	C	A	G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U	C	A	G

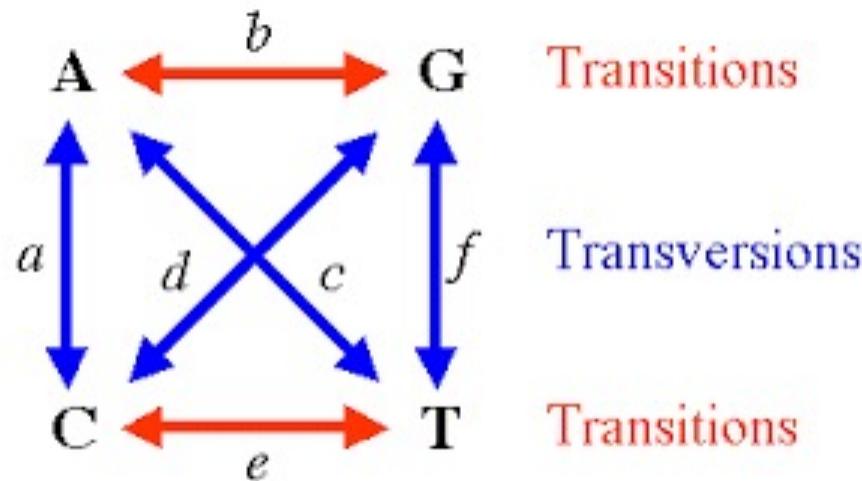
Third letter



Models of substitution

~base frequencies

~transition/transversion matrix



Trade off:

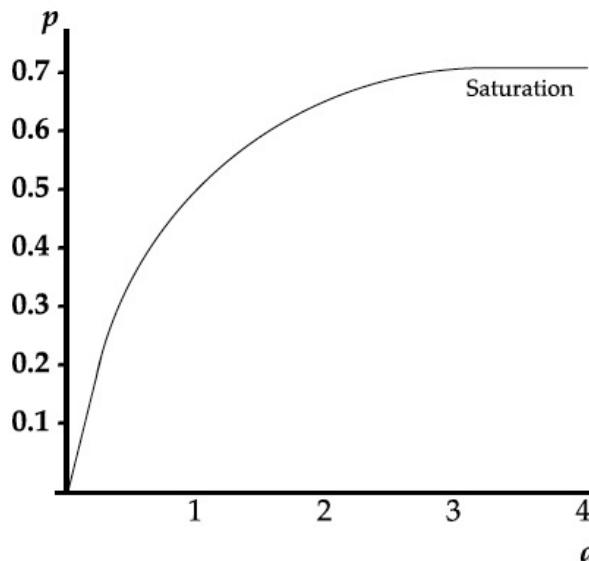
No of parameters vs simplification

Software to test models of substitution

- Modeltest: <https://github.com/ddarriba/modeltest>
- PartitionFinder: <https://www.robertlanfear.com/partitionfinder/>
- MEGA
- JModelTest: <https://github.com/ddarriba/jmodeltest2>

Exercise 3

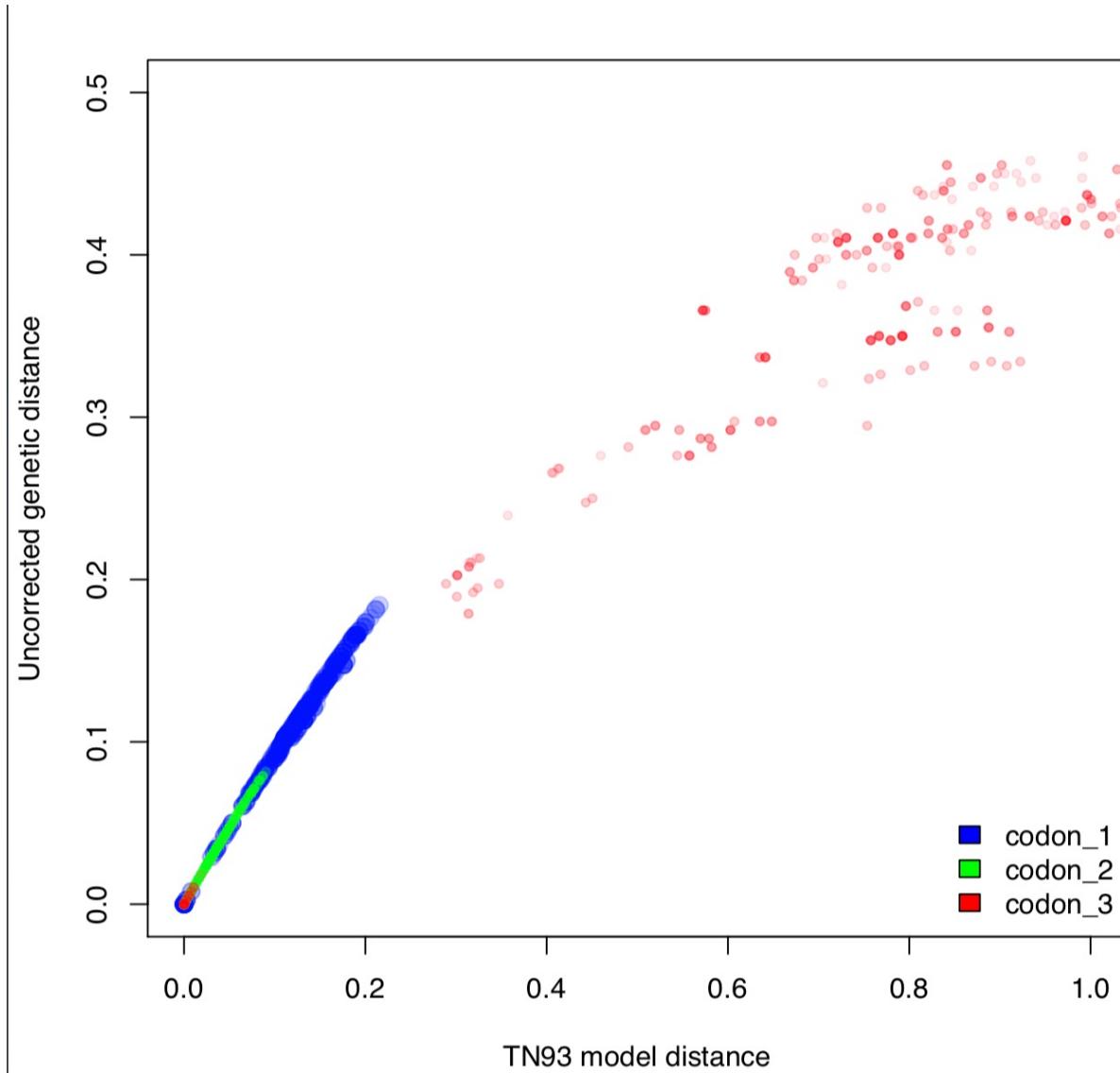
- Visualize saturation plots from DNA alignments for each codon position.



Relationships between expected ***genetic distance*** d and observed ***p-distance***.

		Second letter					
		U	C	A	G		
		UUU } Phe UUC } UUA } Leu UUG }	UCU } Ser UCC } UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Trp UGG Trp	U C A G	
		CUU } CUC } Leu CUA } CUG }	CCU } CCC } CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } CGA } Arg CGG }	U C A G	
		AUU } Ile AUC } AUA } Met AUG }	ACU } ACC } ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } AGC } Ser AGA Stop AGG Stop	U C A G	
		GUU } Val GUC } GUA } GUG }	GCU } GCC } GCA } Ala GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } GGA } Gly GGG }	U C A G	
		First letter				Third letter	

Exercise 3



Phylogenetics methods

- Distance-based (eg. neighbour joining)

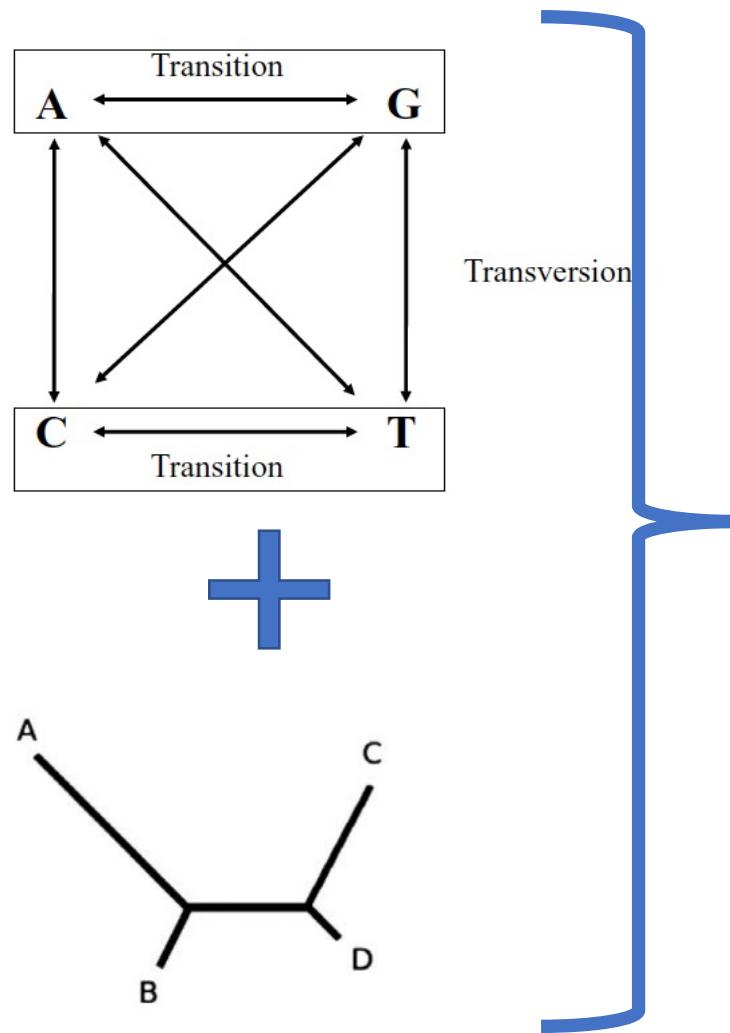
- Maximum likelihood

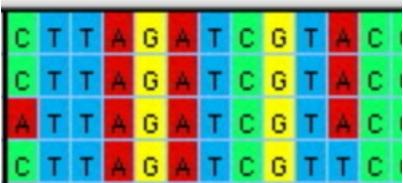
- Bayesian

Maximum likelihood

- probability of observing the data given a particular model of evolution and evolutionary history.
- Computationally intense
- Outperforms other methods
- Allows explicit evolutionary models
- Outputs likelihood of tree topologies

Maximum likelihood



$P ($  $)$

The sequence alignment matrix shows four rows of DNA sequence data. The columns are color-coded: green for C, blue for T, red for A, and yellow for G. The sequences are as follows:

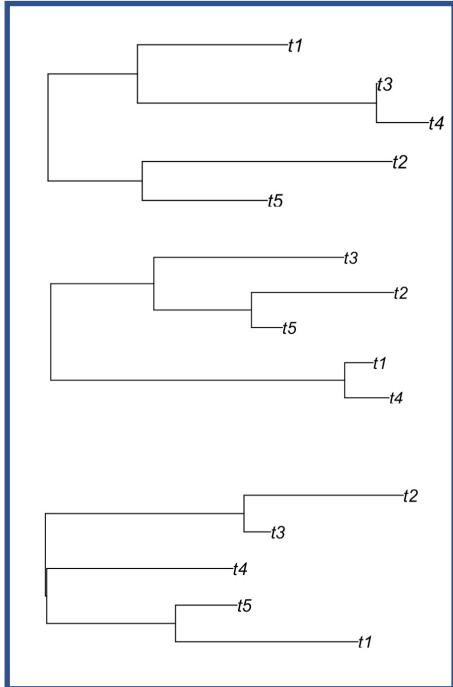
C	T	T	A	G	A	T	C	G	T	A	C
C	T	T	A	G	A	T	C	G	T	A	C
A	T	T	A	G	A	T	C	G	T	A	C
C	T	T	A	G	A	T	C	G	T	T	C

Example: RAxML Maximum likelihood

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

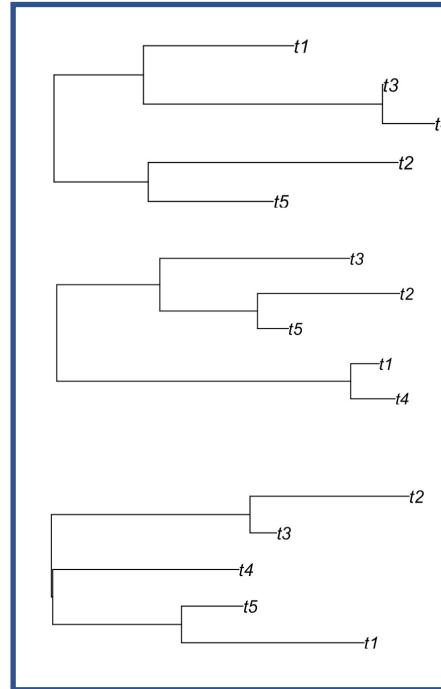
The RAxML: starting tree based on maximum parsimony, which is re-optimized using the current best tree.
For the 20 best trees, the final ML-value is re-optimized by adjusting all branch lengths. Tree search and re-optimization is repeated until no better tree is found.

Maximum parsimony



Tree search:
optimization

Best trees
Branch-length
adjustment



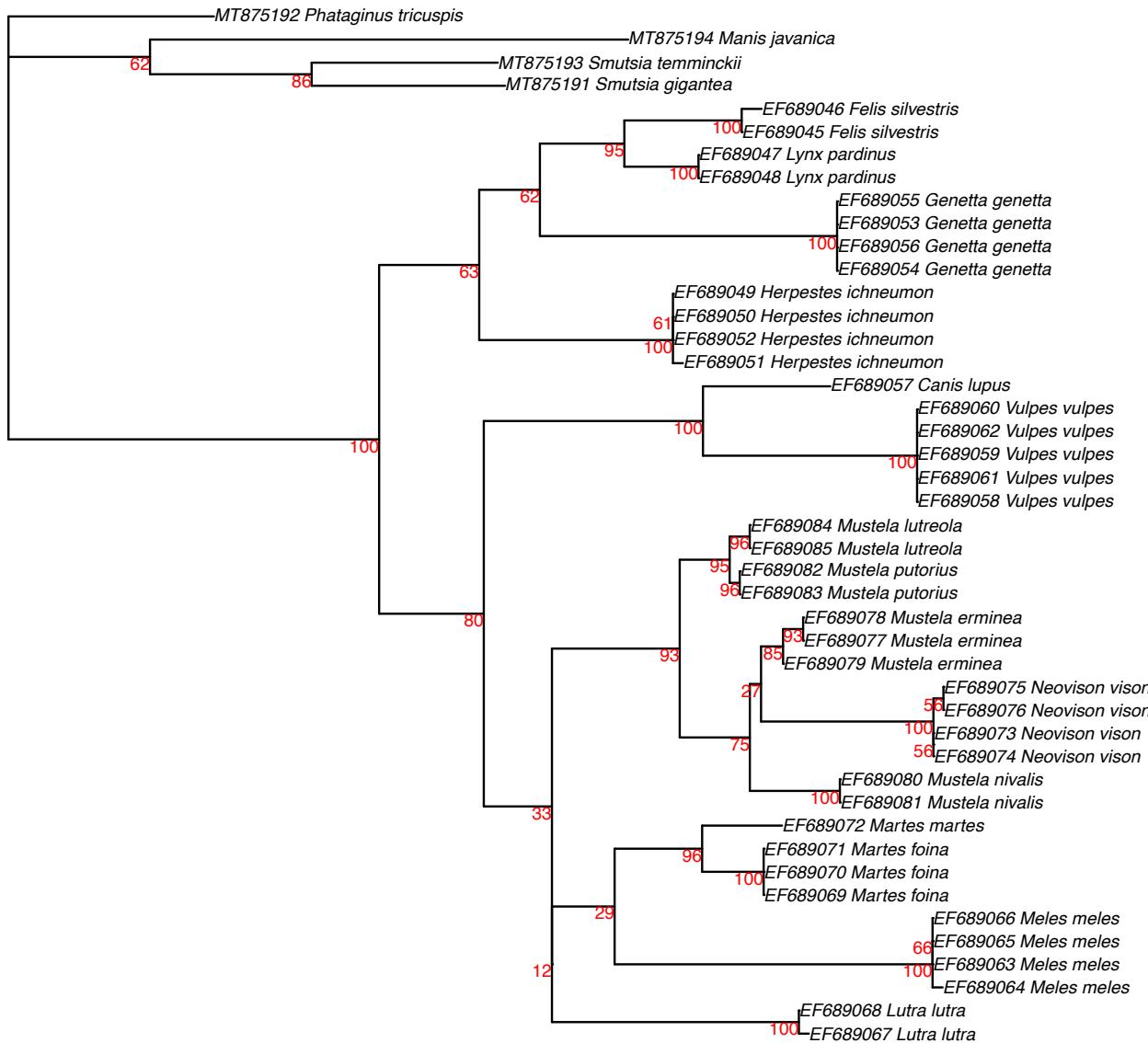
Best trees found

Exercise 4: Maximum likelihood

Reconstruct phylogenetic tree with ML in R

Go to the GitHub repository

Exercise 4: Maximum likelihood



Phylogenetics methods

- Distance-based (eg. neighbour joining)
- Maximum likelihood
- Bayesian

Bayesian inference

Bayes theorem applied to phylogenetics:

$$P(\text{tree}_j / D) = \frac{P(D / \text{tree}_j) P(\text{tree}_j)}{\sum_{i=1}^n P(D / \text{tree}_i) P(\text{tree}_i)}$$

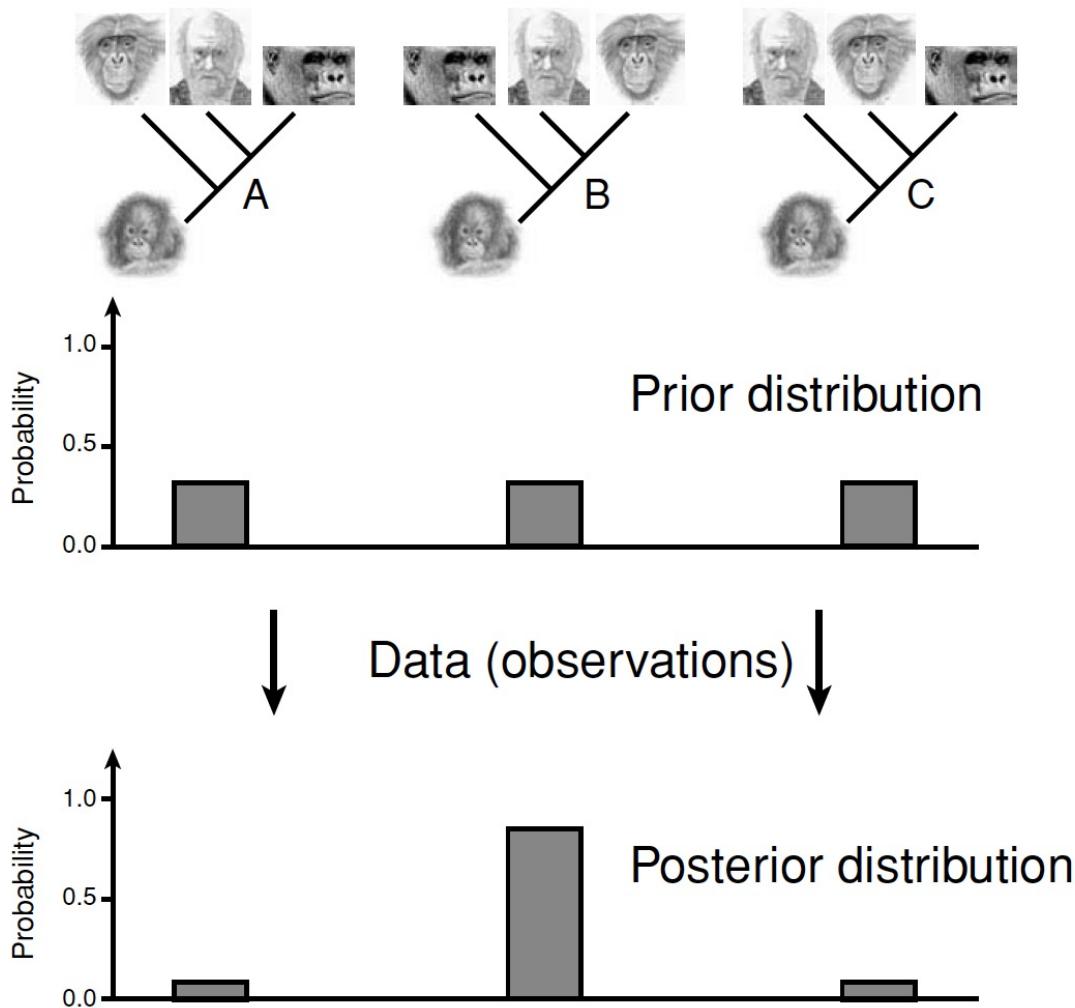
↓

$$\text{Posterior distribution} = \frac{\text{Prior distribution} * \text{likelihood of tree}}{\text{constant}}$$

The denominator is difficult to compute, but we know it is constant for a given problem

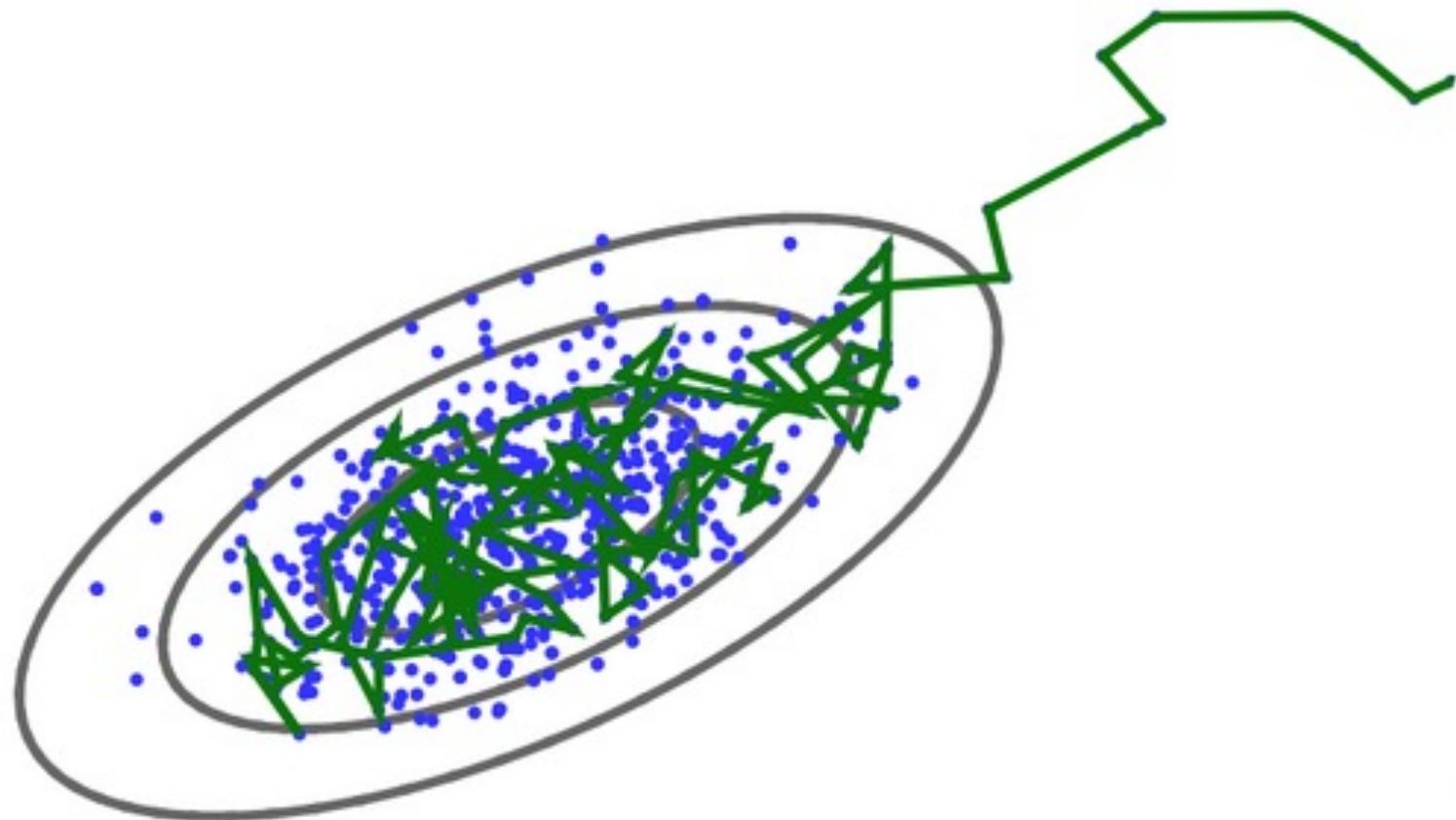
Bayesian inference

Allow the incorporation of expertise knowledge as prior info

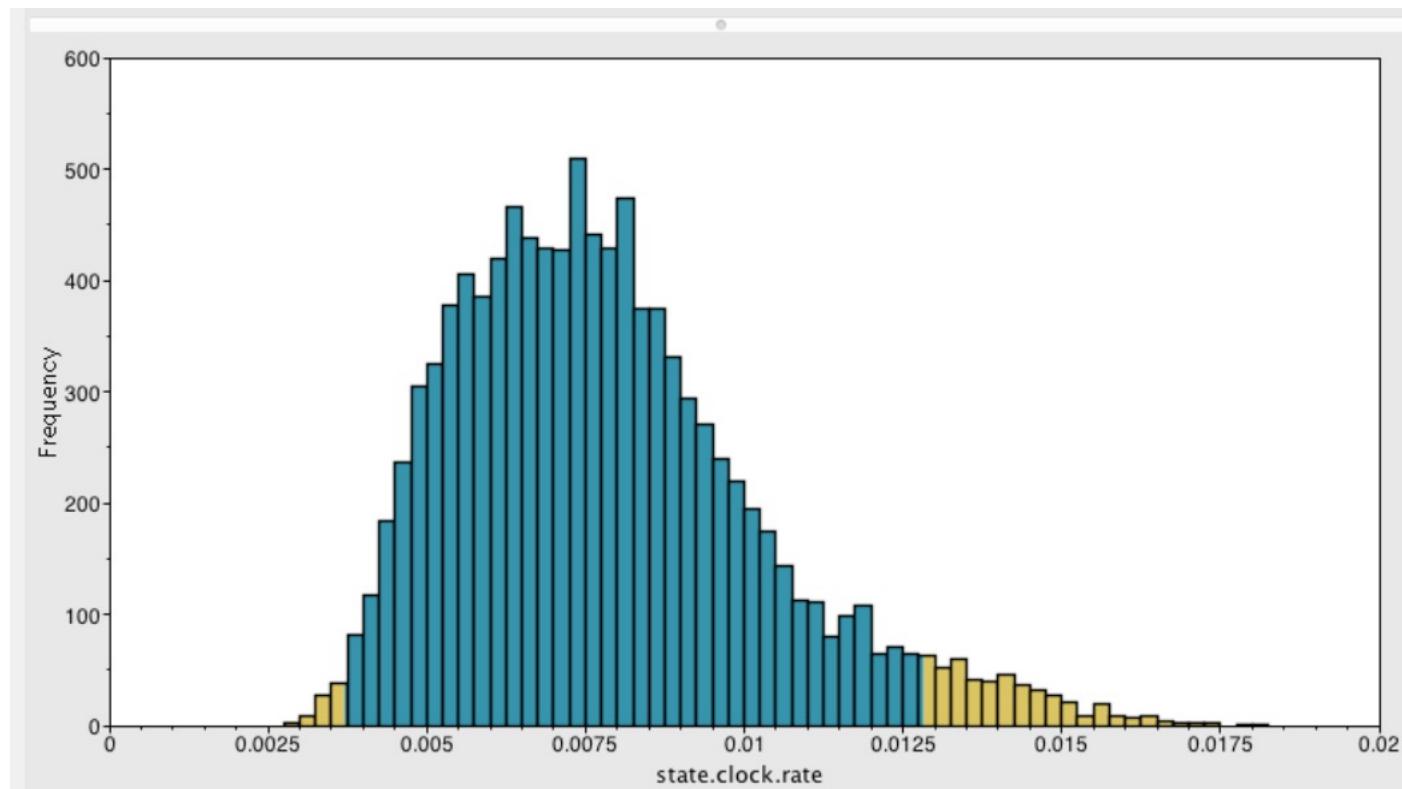


$$\textit{Posterior distribution} \propto \textit{Prior distribution} * \textit{likelihood of tree}$$

MCMC sampler



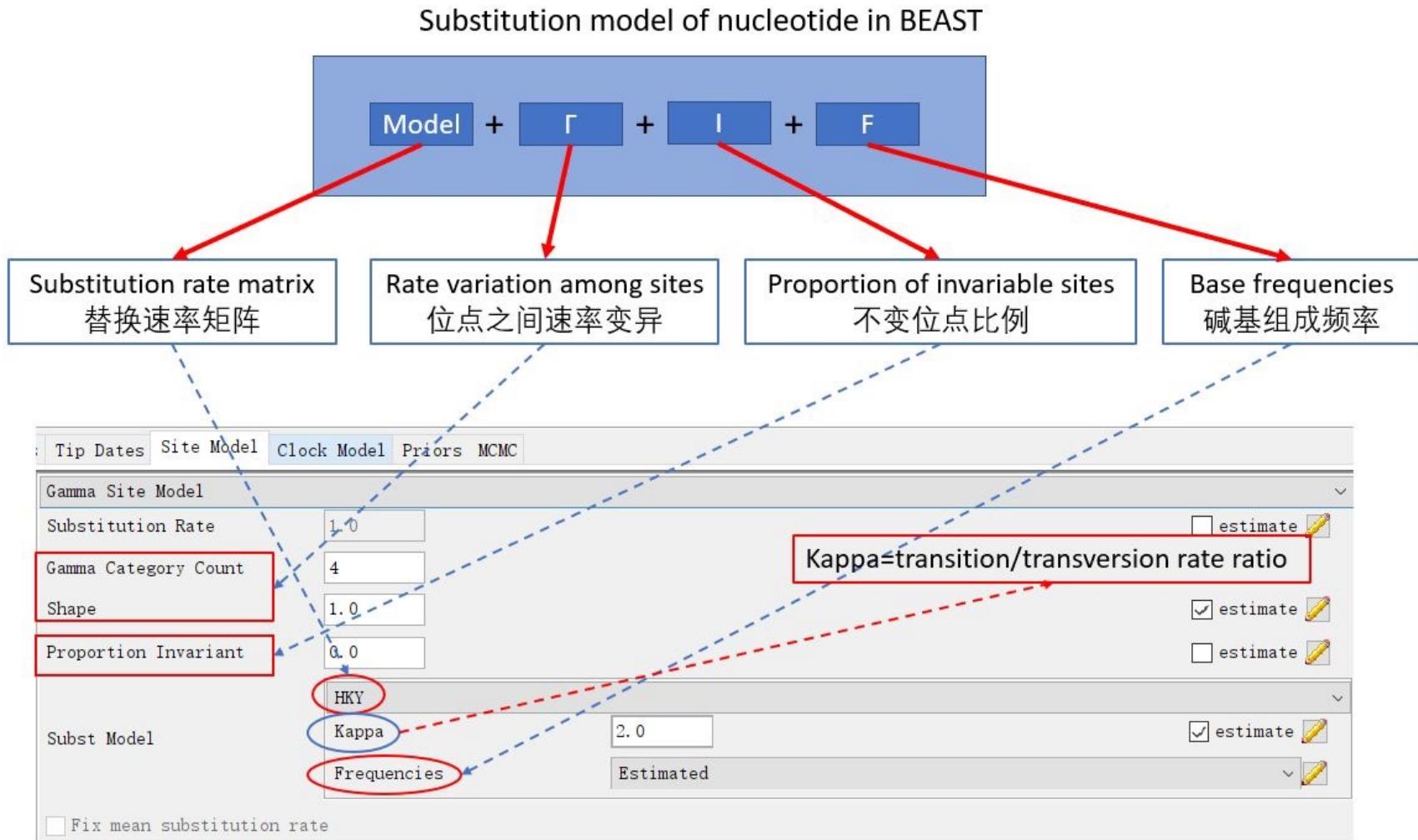
Interpreting output: HPD 95% (high posterior density)



Exercise 5: Bayesian inference

- BEAST:
 - Coalescent reconstructions.
 - Tutorial online:
 - https://github.com/csmiguel/2021_ConGenTopics/blob/main/tutorials/phylogenetics.md
- Install BEAST (instructions at <https://www.beast2.org/>):
 - Download for Windows [without java \(8MB\)](#) / [with java \(45MB\)](#)
 - Download for Mac OS X [without java \(8MB\)](#) / [with java \(46MB\)](#)
 - Download for Linux [without java \(8MB\)](#) / [with java \(47MB\)](#)

Exercise 5: Bayesian inference



Exercise 5: Bayesian inference

A screenshot of a software interface for a Bayesian phylogenetic analysis. The top navigation bar includes tabs for 'Partitions', 'Tip Dates', 'Site Model', 'Clock Model' (which is highlighted in blue), 'Priors', and 'MCMC'. Below the tabs, the 'Clock Model' section is titled 'Relaxed Clock Log Normal'. It contains the following settings:

- 'Number Of Discrete Rates': A dropdown menu set to '-1'.
- 'Normalize': An unchecked checkbox.
- 'Clock.rate': A text input field containing '1.0'.
- 'estimate': An unchecked checkbox followed by a pencil icon.

The uncorrelated relaxed molecular clock model is a **model of rate heterogeneity across** branches that assumes each branch has its own independent rate drawn from a single shared parametric rate distribution

Exercise 5: Bayesian inference

Partitions Tip Dates Site Model Clock Model **Priors** MCMC

▶ Tree.t:cytb_beast Yule Model

▶ birthRate.t:cytb_beast Uniform initial = [1.0] $[-\infty, \infty]$ Prior on Yule birth rate for partition s:cytb_beast

▶ gammaShape.s:cytb_beast Exponential initial = [1.0] $[-\infty, \infty]$ Prior on gamma shape for partition s:cytb_beast

▶ kappa.s:cytb_beast Log Normal initial = [2.0] $[0.0, \infty]$ HKY transition-transversion parameter of partition s:cytb_beast

▶ ucldMean.c:cytb_beast Uniform initial = [1.0] $[-\infty, \infty]$ uncorrelated lognormal relaxed clock mean of partition c:cytb_beast

▶ ucldStddev.c:cytb_beast Gamma initial = [0.1] $[0.0, \infty]$ uncorrelated lognormal relaxed clock stdev of partition c:cytb_beast

▼ cats.prior Log Normal monophyletic

M 12 estimate

S 0.2 estimate

Mean In Real Space

Offset 0.0

Tipsonly

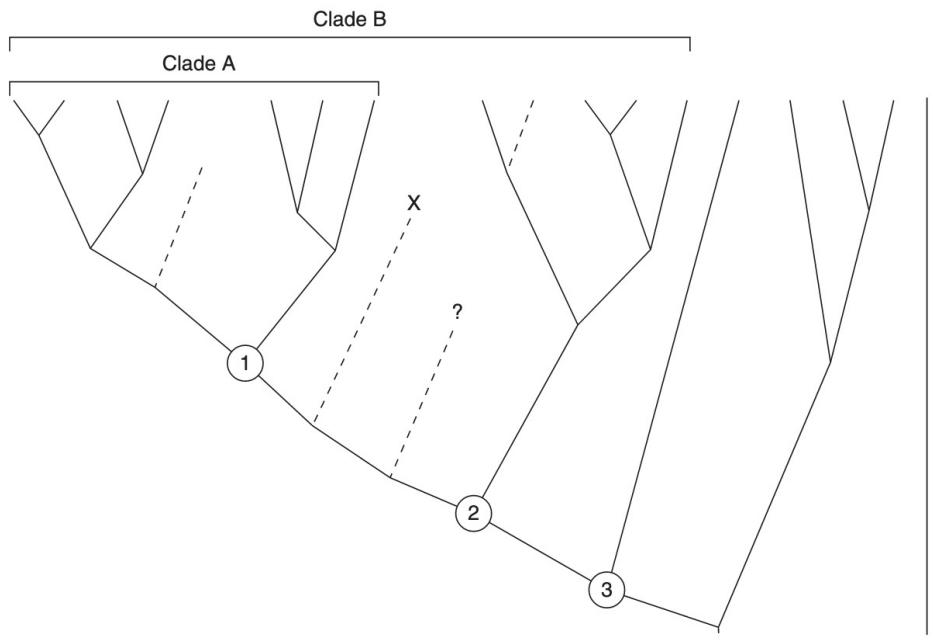
Use Originate

Yule tree prior that assumes a (unknown) constant lineage birth rate for each branch in the tree. This tree prior is most suitable for trees describing the relationships between individuals from different species.

Exercise 5:

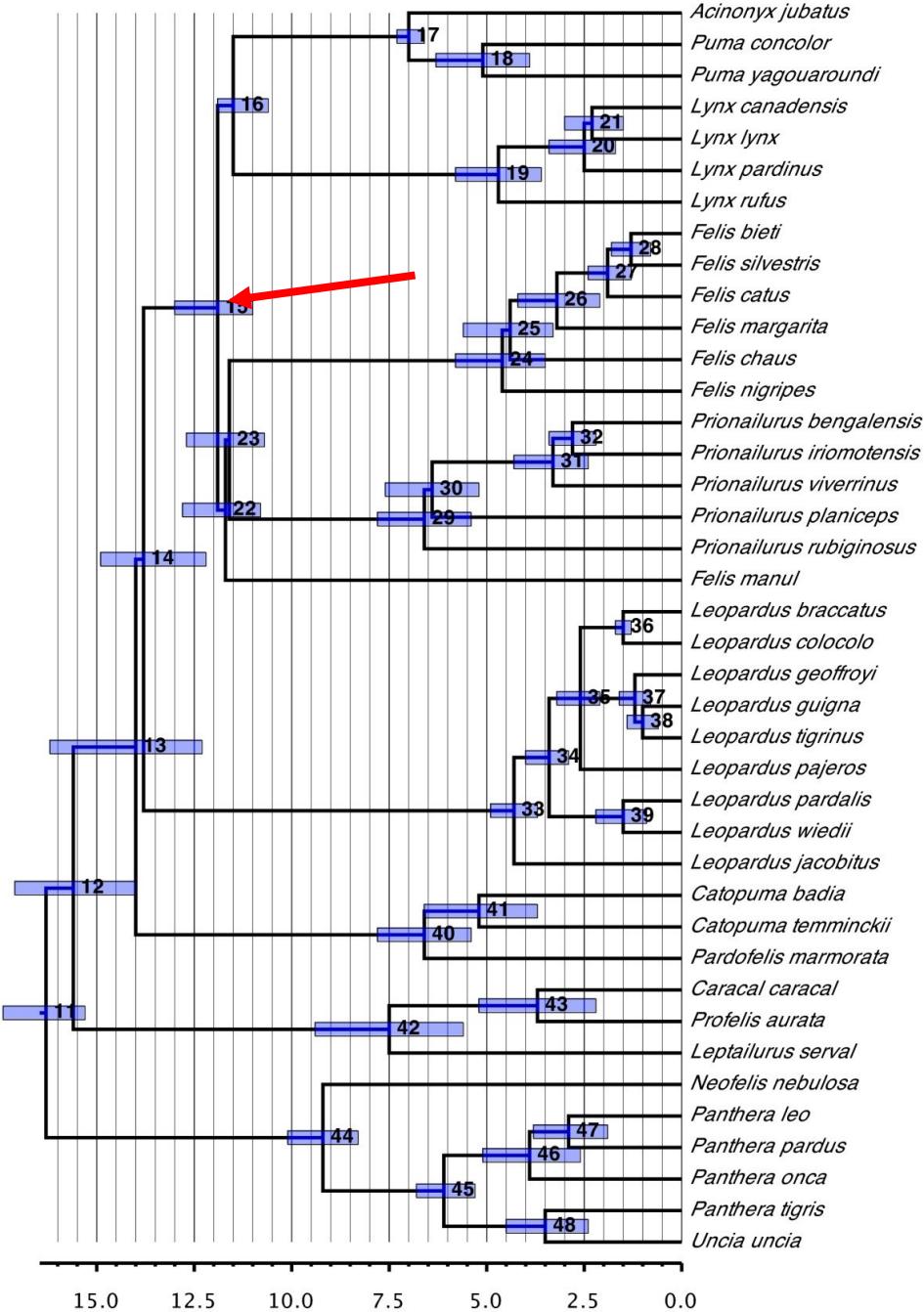
Bayesian inference-calibration

Goal. Put evolutionary time in time units



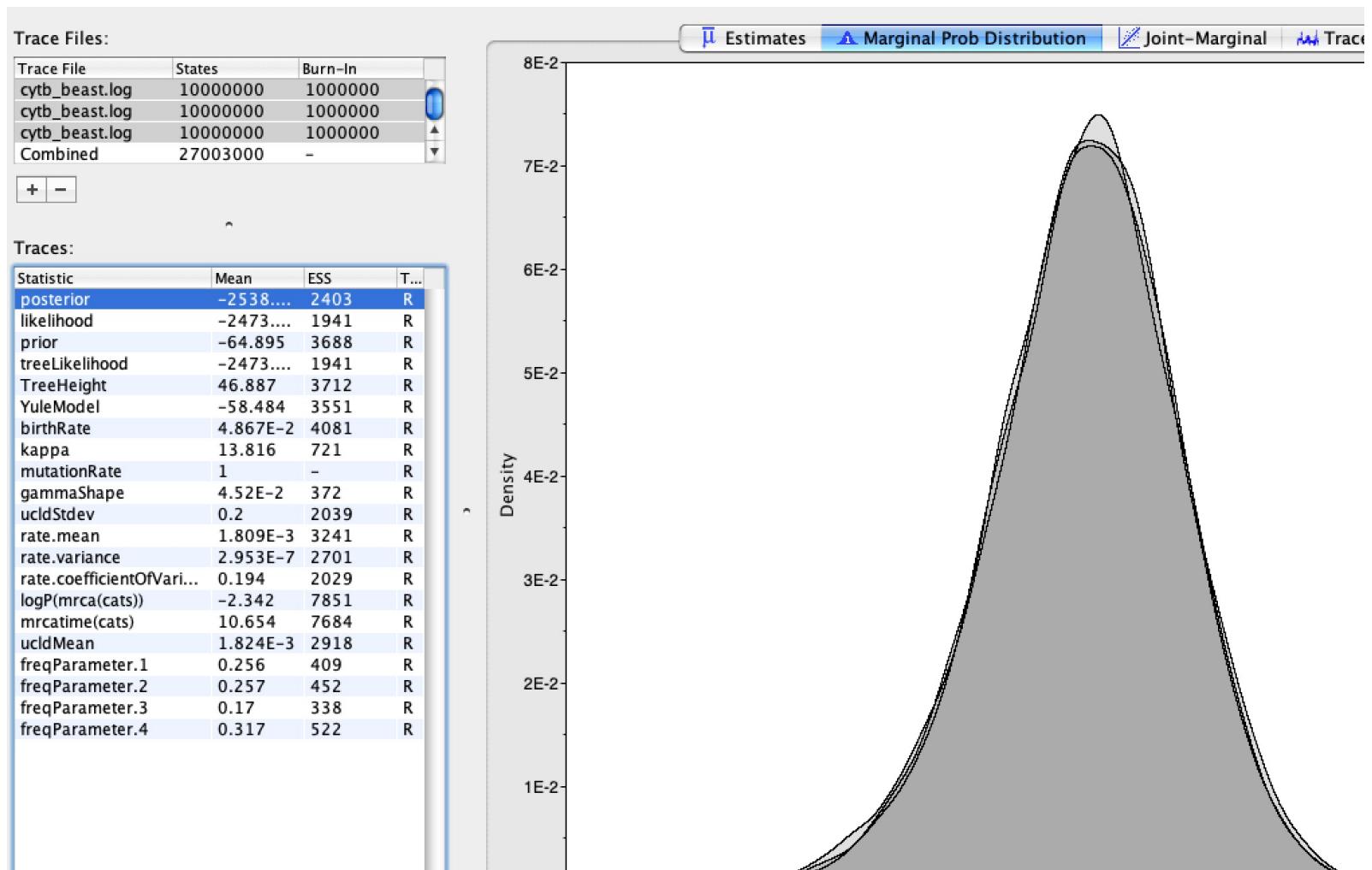
- Fossils
- Mutation rates
- Historical events
- (secondary calibration points) avoid!

Exercise 5: Bayesian inference- calibration



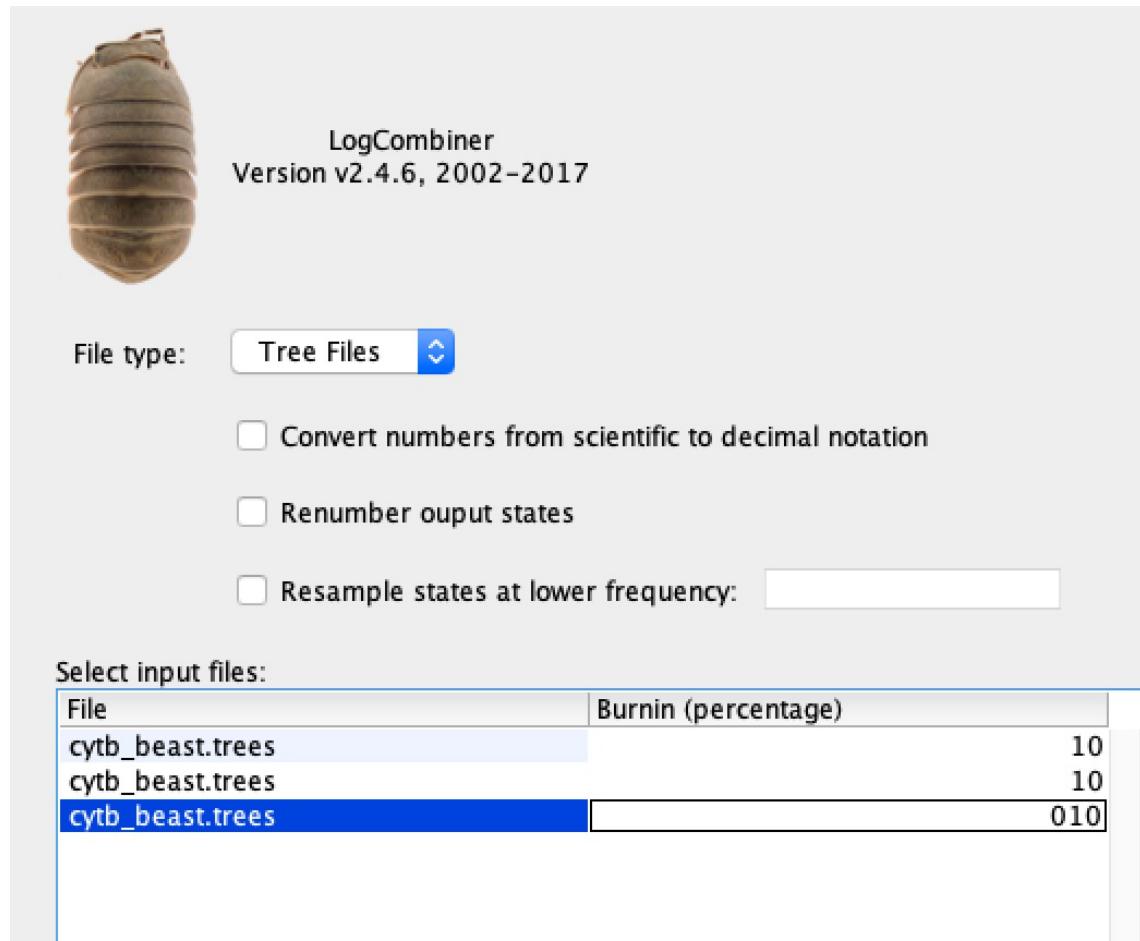
(10.1186/1741-7007-10-12)

Tracer: check convergence



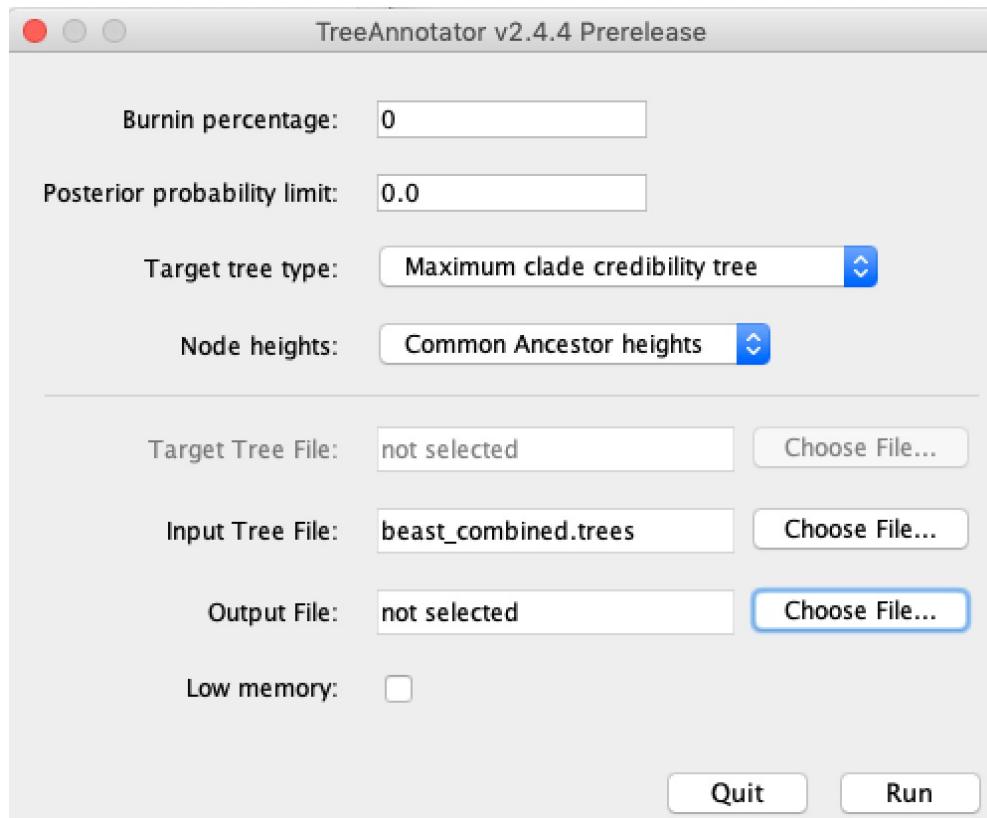
LogCombiner: combine trees from multiple chains.

- Discard 10% of trees

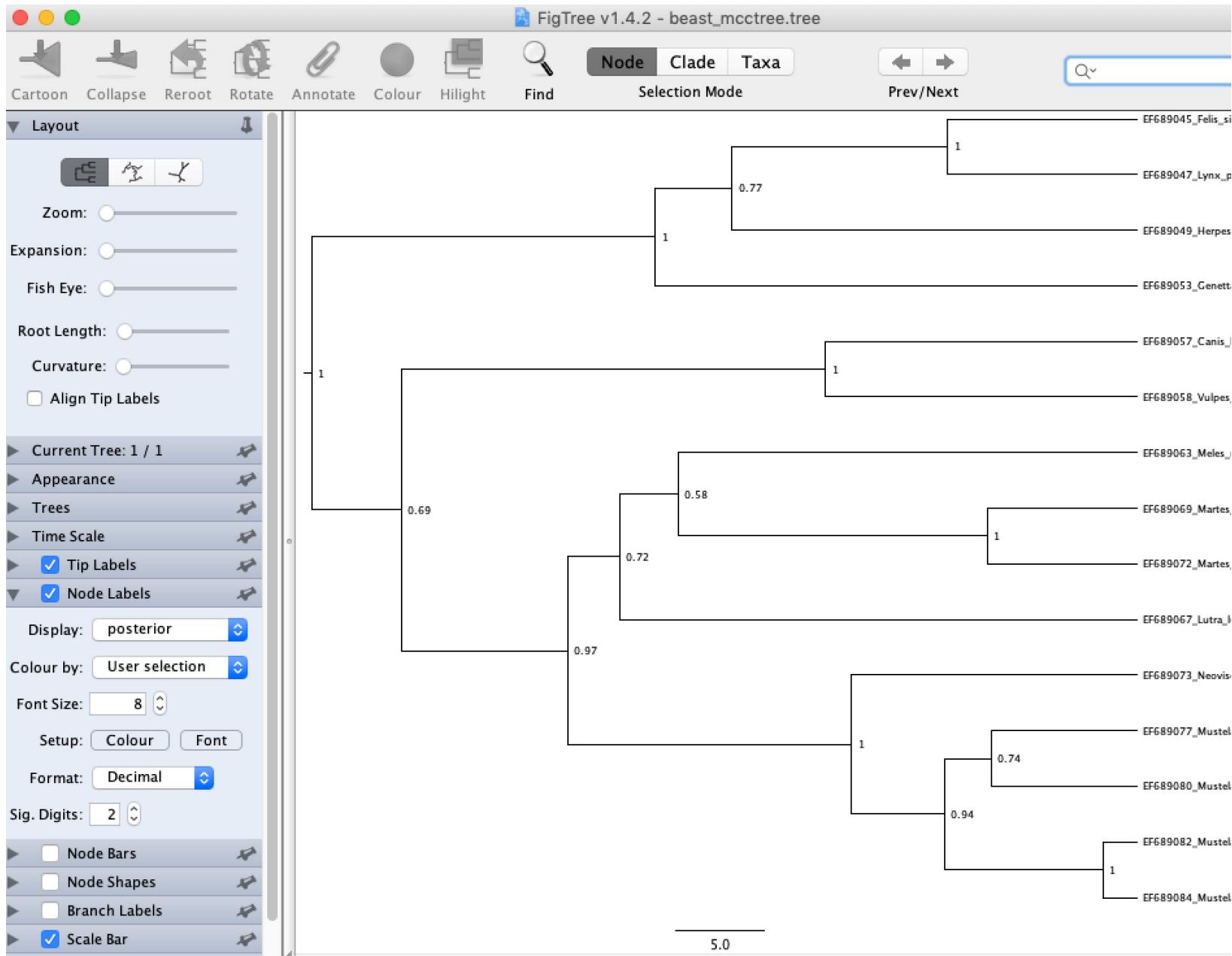


TreeAnnotator

- Discard 10% of trees
- Get Maximum clade credibility tree: **summarises the results of a Bayesian phylogenetic inference**



FigTree: visualize tree



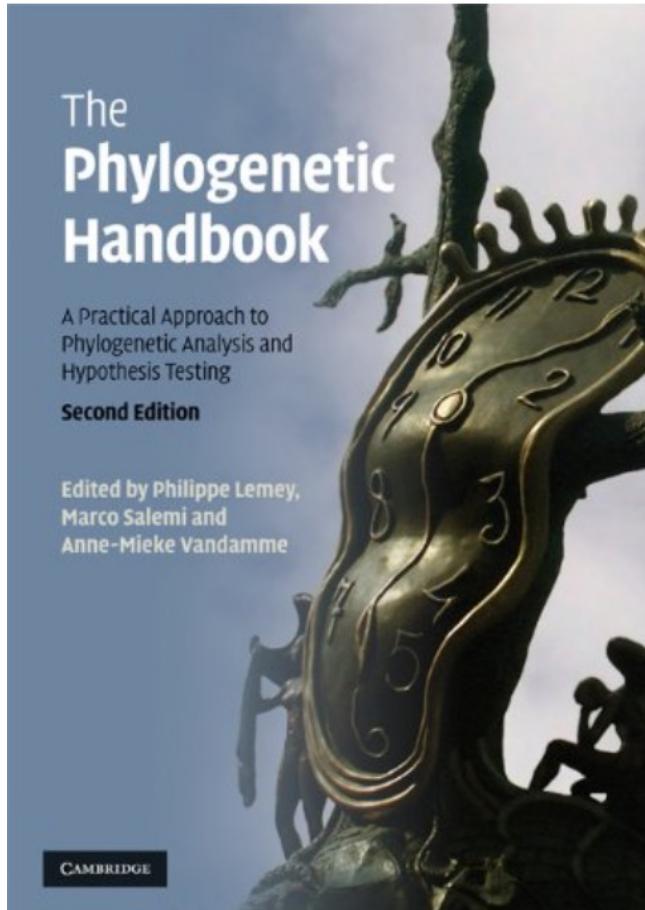
Software need

- In Galaxy
 - Raxml
 - Phyloviz
- Tracer <https://github.com/beast-dev/tracer/releases/tag/v1.7.2>
- Figtree <https://github.com/rambaut/figtree/releases>

More complex datasets

- Multiple partitions: PartitionFinder, RAxML, BEAST.
- Mutilocus: species trees vs gene trees.
- Avoid concatenation.

Bibliography



- Resources:
- <https://evolution.genetics.washington.edu/phylip/software.html>
- Manuals RAxML, BEAST