GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES CSIC

Estación Biológica Doñana

Conservation Genetics in the Tropics

Phylogenies

Miguel Camacho Sánchez

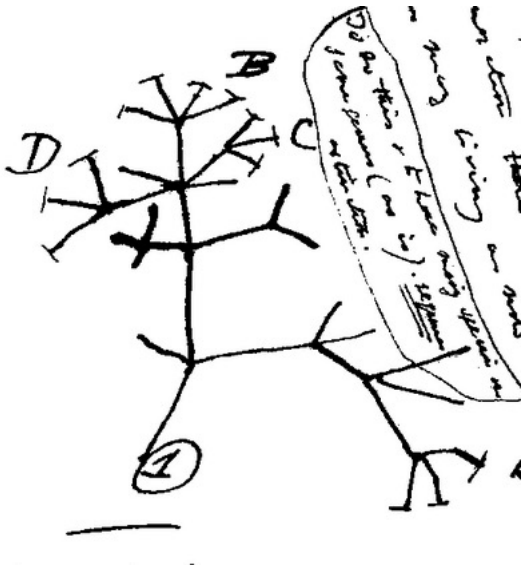
October 25th, 2021

miguelcamachosanchez@gmail.com

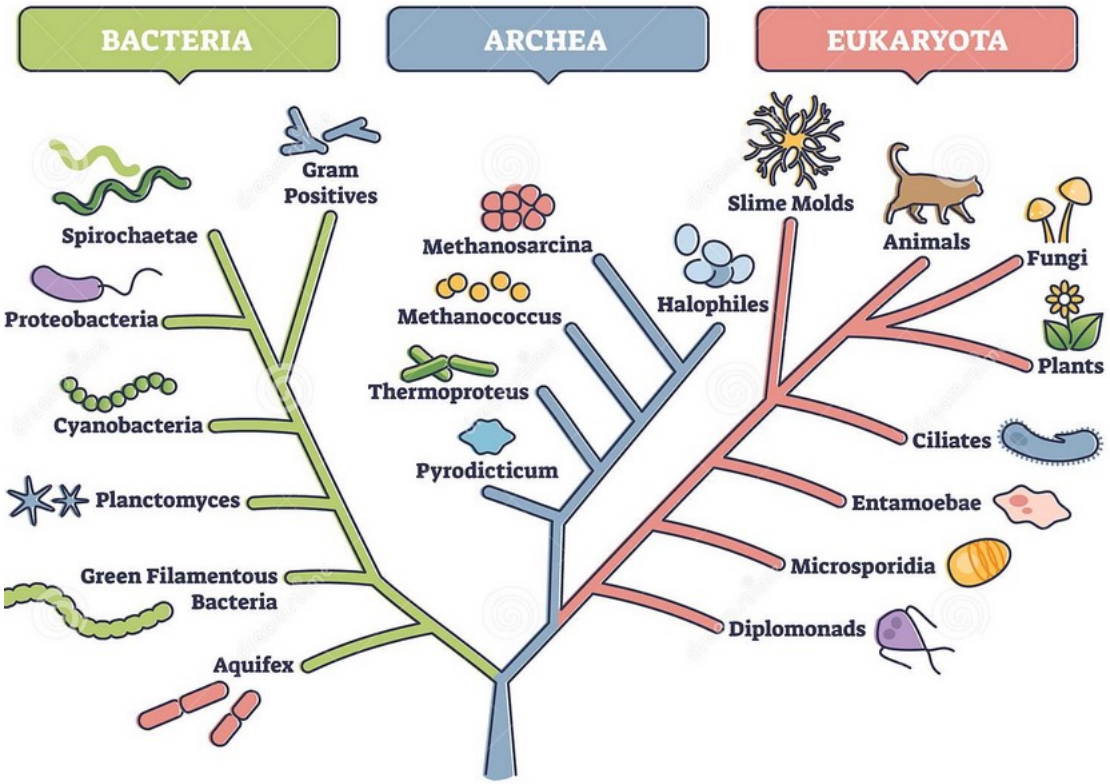
miguelcamachosanchez.weebly.com

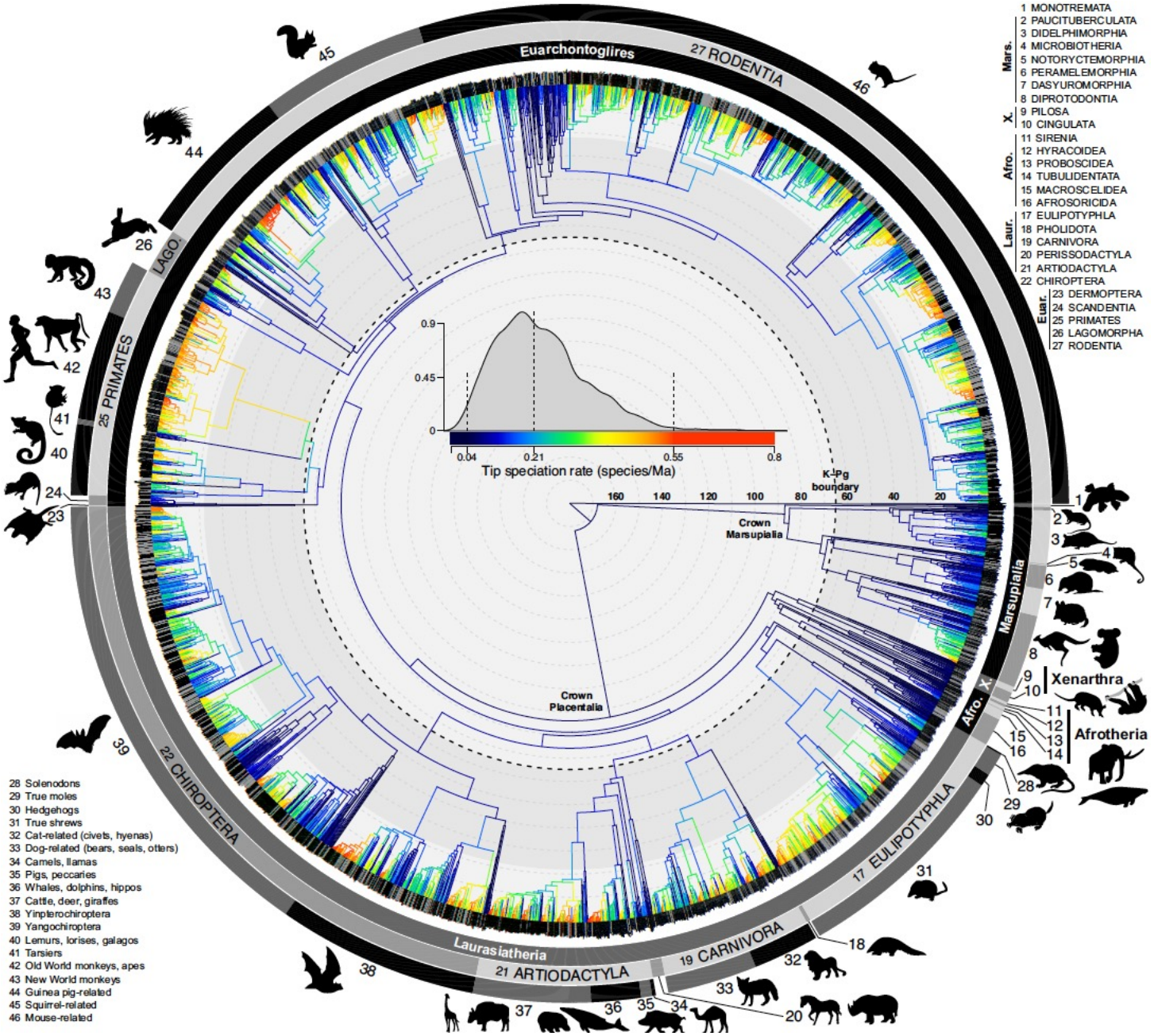
Phylogenetics

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



Examples

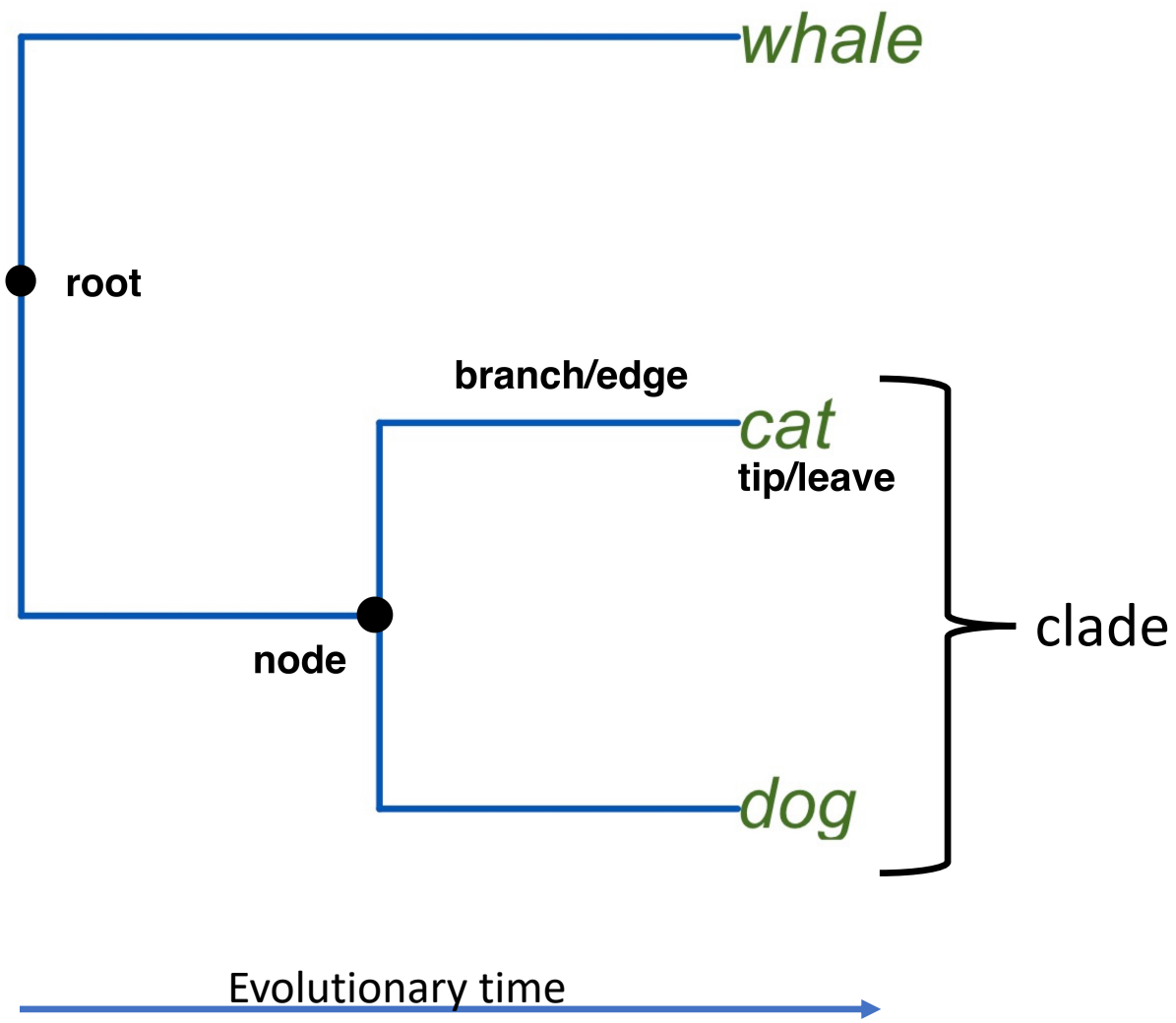

BACTERIA
ARCHEA
EUKARYOTA
Gram Positives
Spirochaetae
Proteobacteria
Cyanobacteria
Planctomyces
Green Filamentous Bacteria
Aquifex
Methanosarcina
Methanococcus
Thermoproteus
Pyrodicticum
Halophiles
Slime Molds
Animals
Fungi
Plants
Ciliates
Entamoebae
Microsporidia
Diplomonads



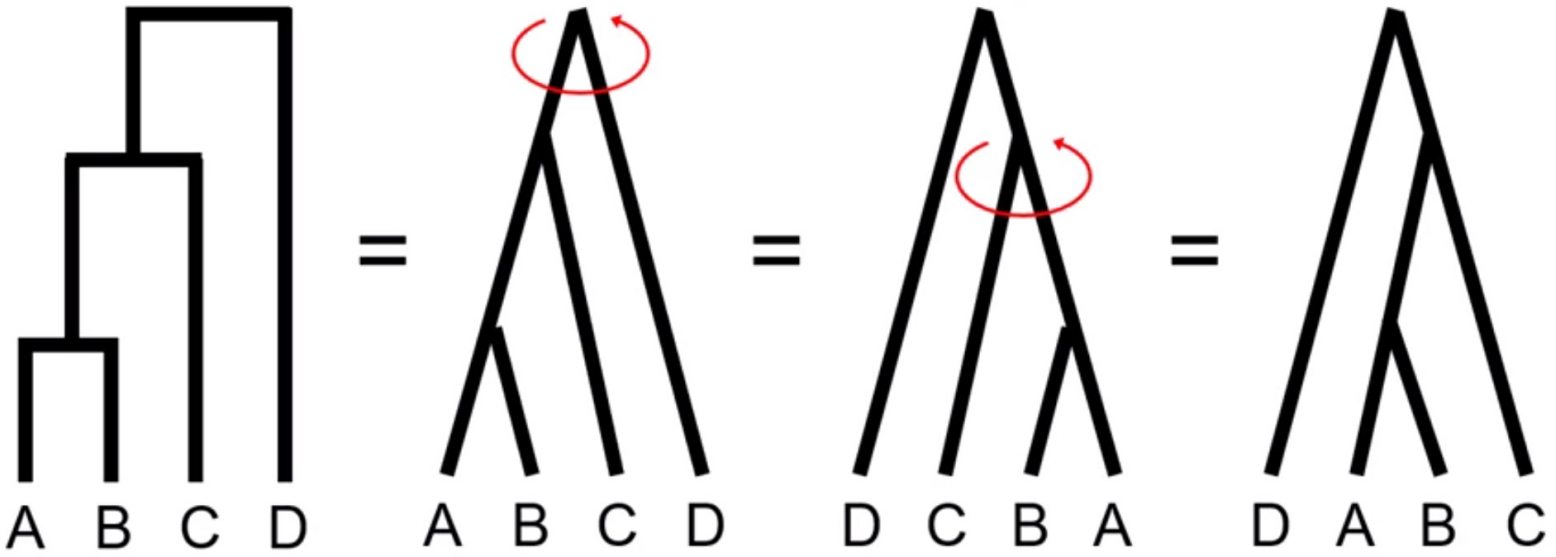
doi.org/10.1371/journal.pbio.3000494

Parts of a tree

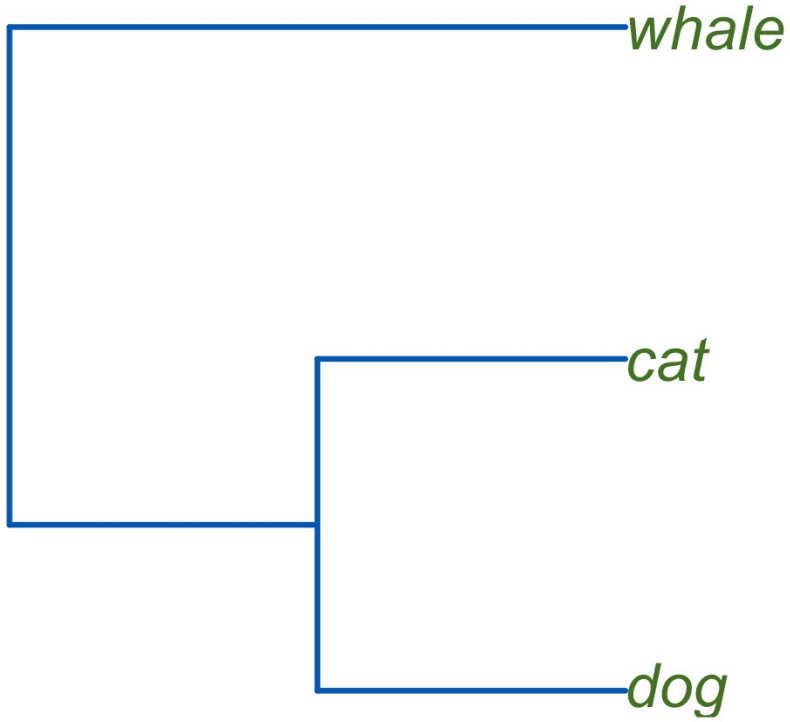
Topology = shape
Leave = terminal node



Ways of representing a tree



Newick format

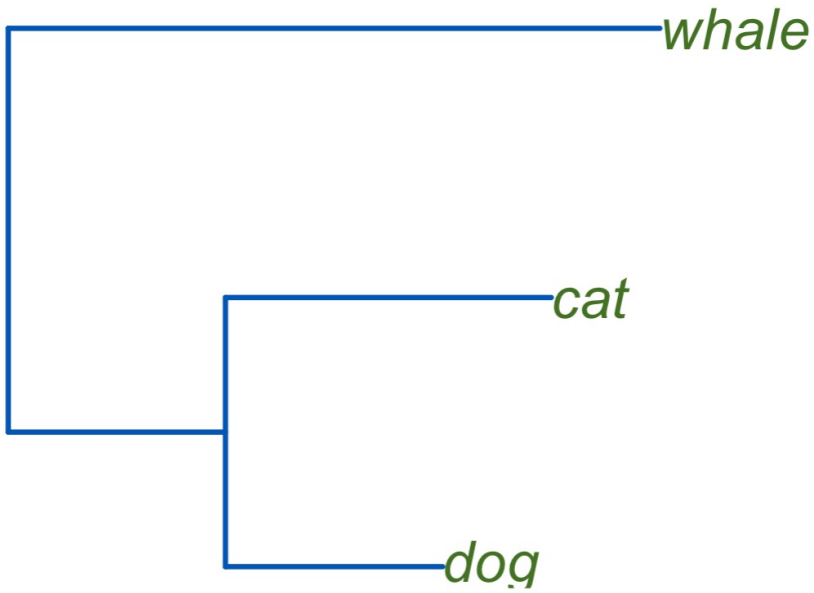


- tips

((dog,cat),whale);

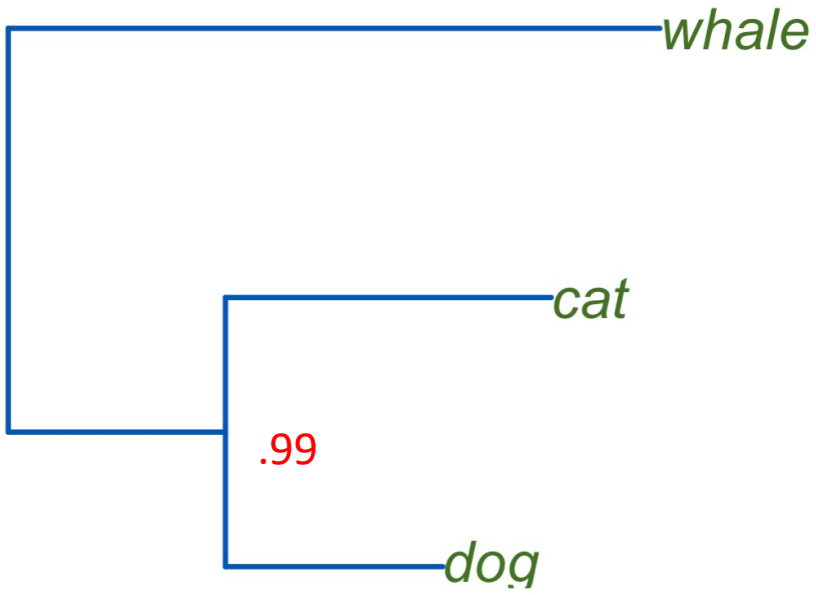
Newick format

- tips
- branch lengths



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

Newick format

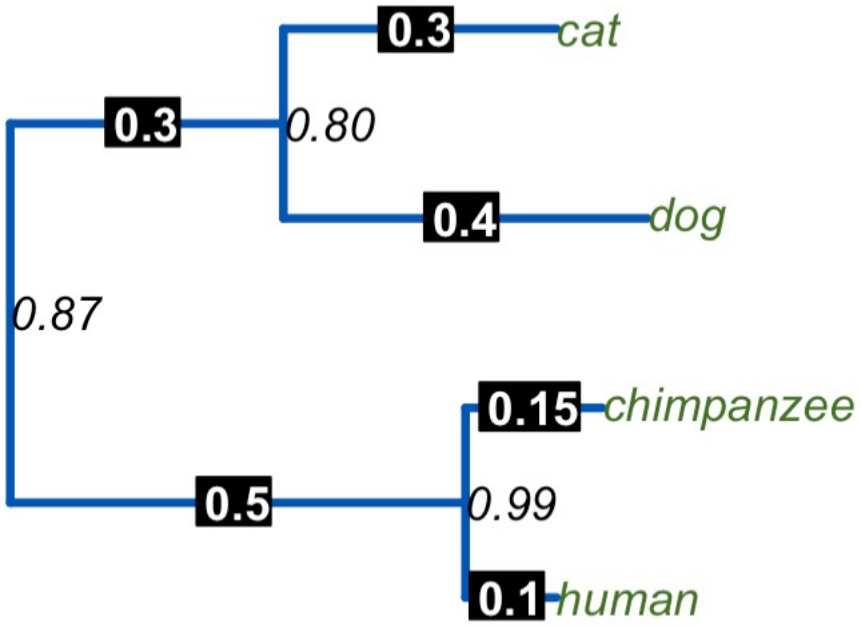


- tips
- branch lengths
- Node label (BS)

((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:

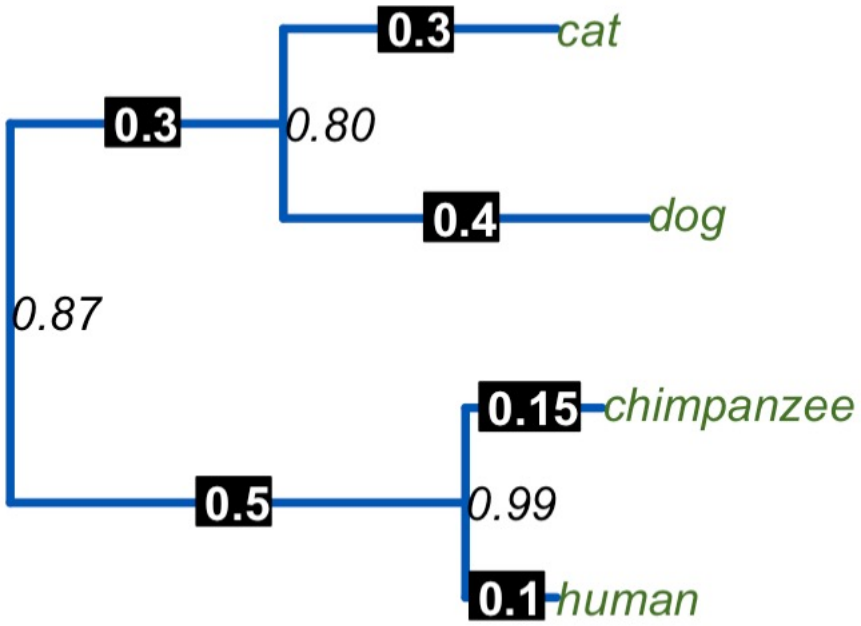
((dog:1,cat:1.5).99:1,whale:3);

Exercise1: set up environment in Rstudio

- Go to Rstudio in elabs Galaxy
- Run this:

```
download.file("https://raw.githubusercontent.com/csmiguel/2021_ConGenTropics/main/scripts/import_data.r", "import_data.r")
```

Exercise1: tree format

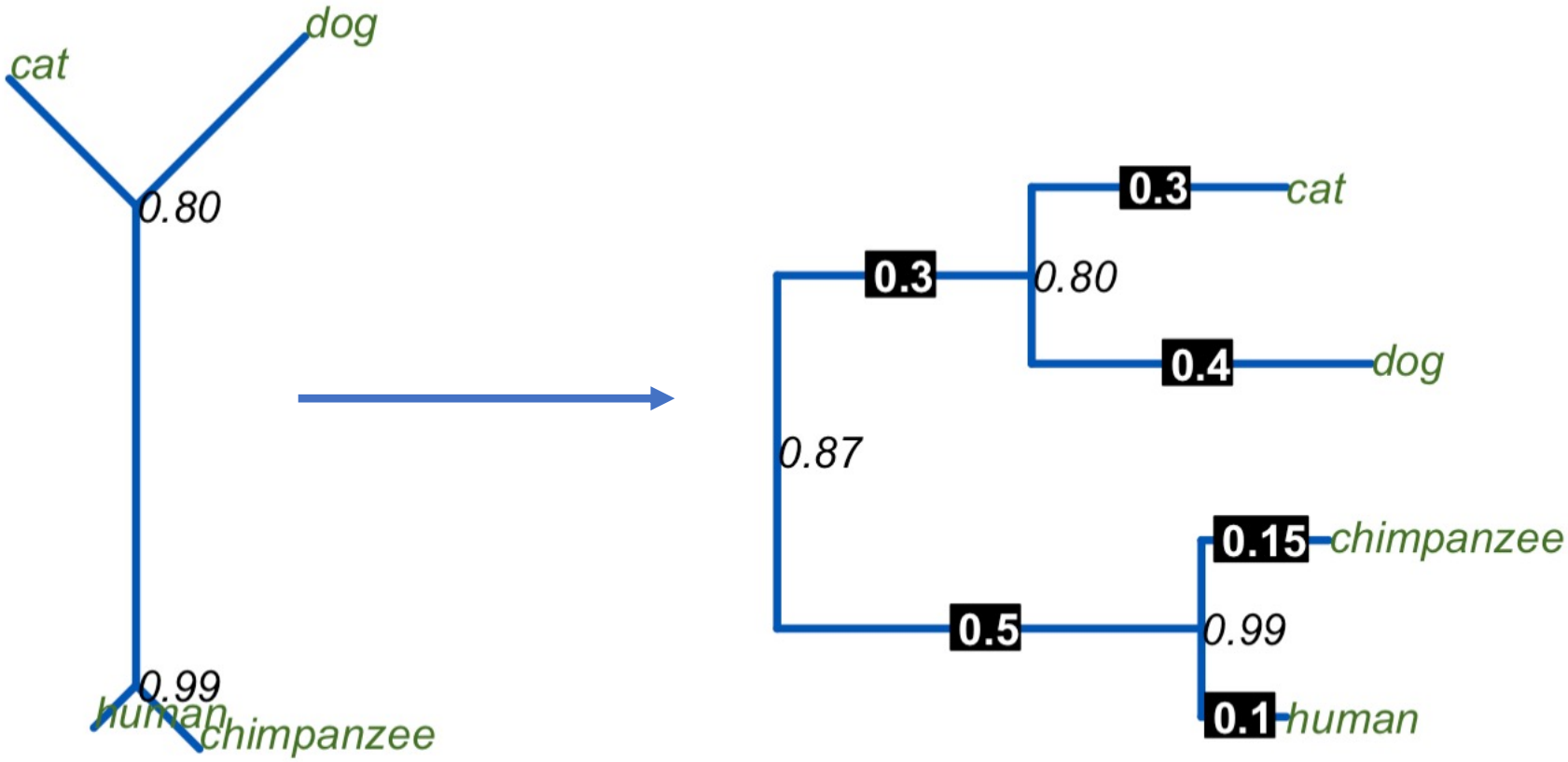


((human:0.1, chimpanzee:0.15)0.99:0.5,(dog:0.4, cat:0.3)0.80:0.3)0.87;

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

Rooting

Implies giving a directionality to the changes



Unrooted tree

rooted tree

Phylogenetic inference workflow

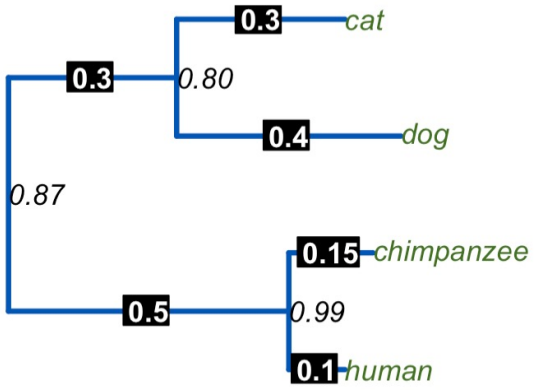
Multiple sequence alignment

Data selection

Inference

Phylogenetics methods

Tree

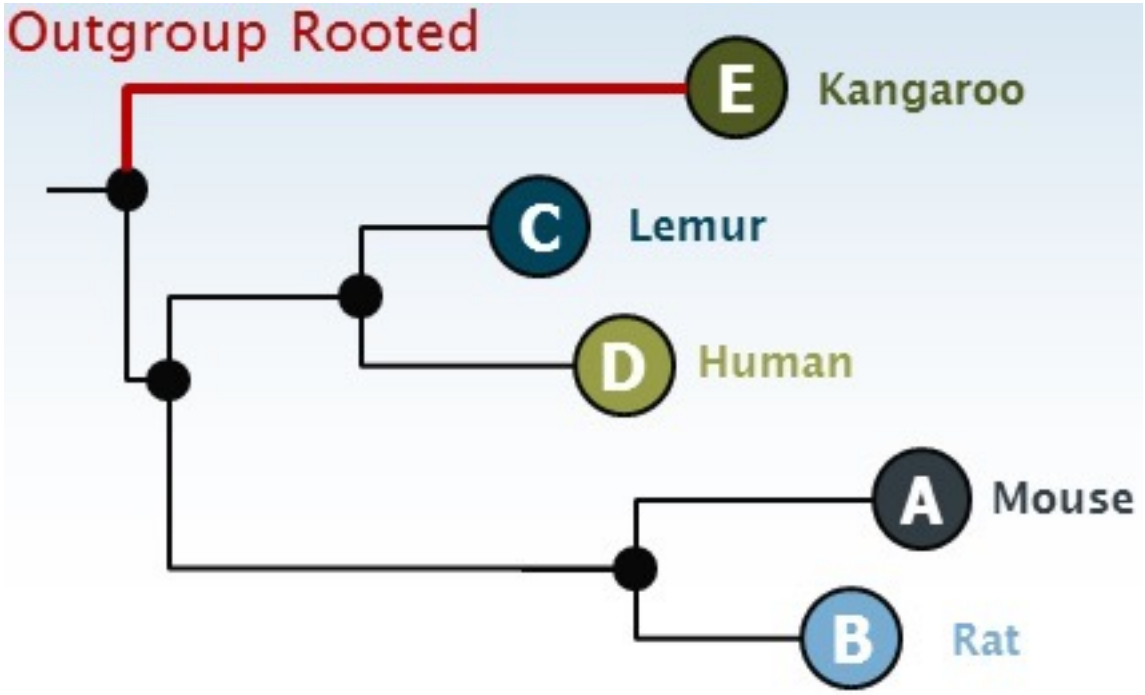


Data for phylogenetics

- Taxa taxonomic sampling
- Loci genetic sampling

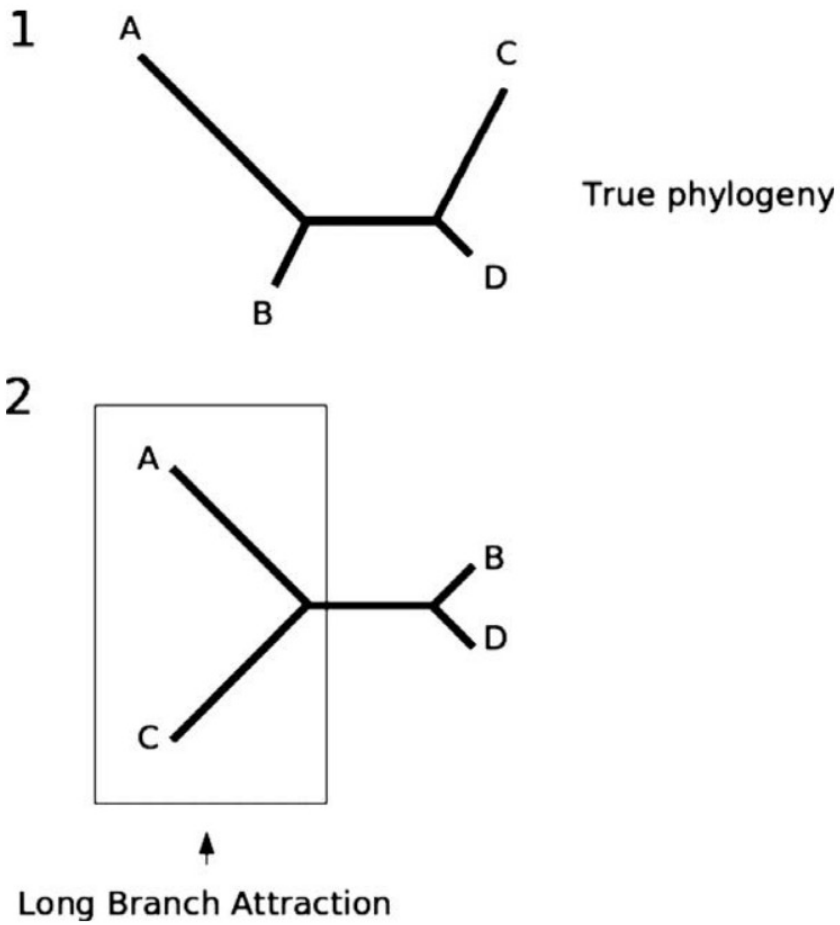
Taxa

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



Taxa

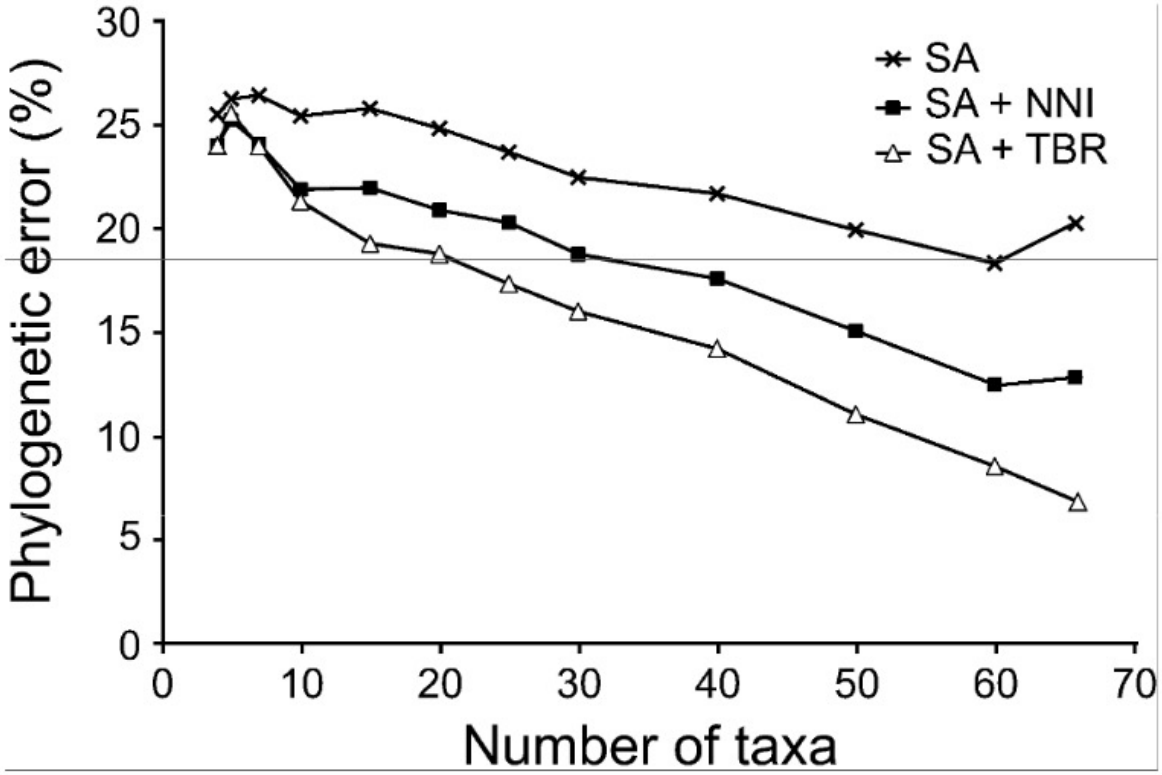
- Large taxonomic sampling prevents from long branch attraction (10.1080/10635150500234583)



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)

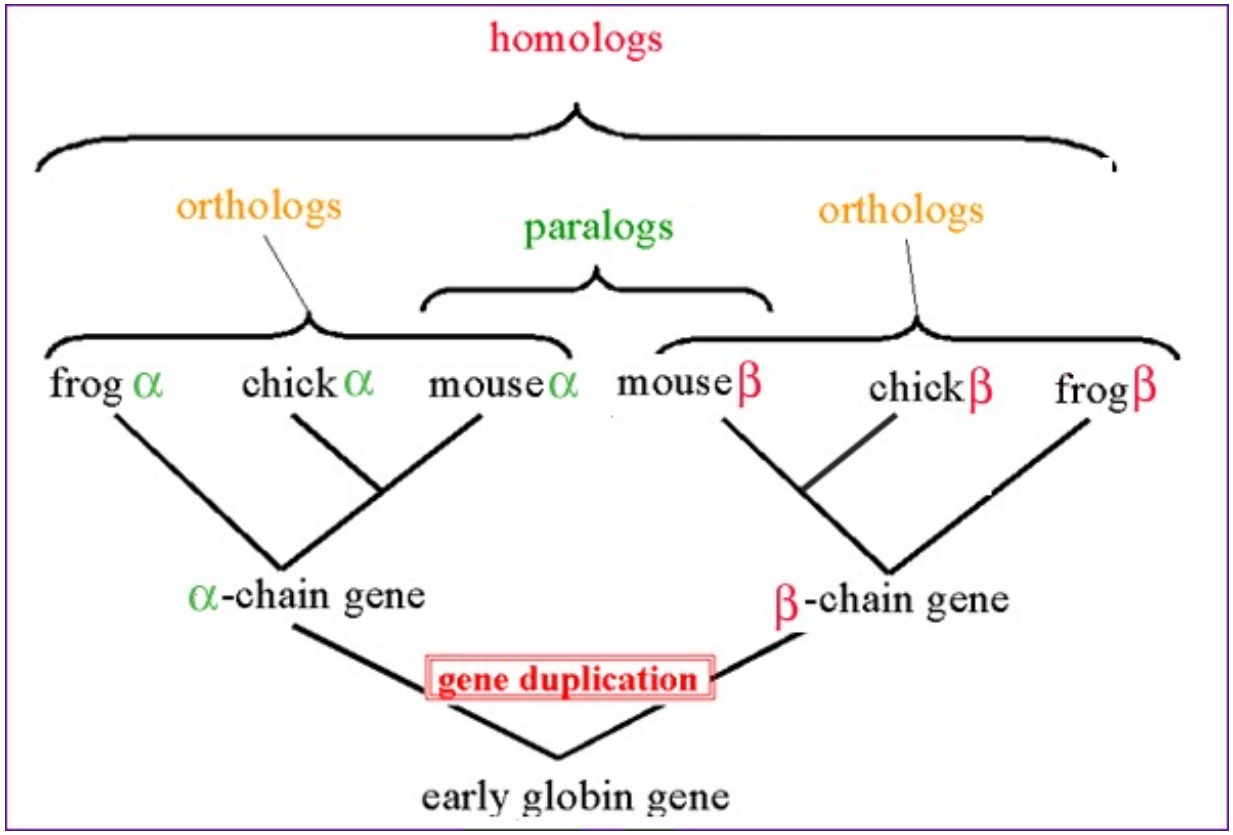


Loci: genetic markers

- Homologous
- Variable
- Independent

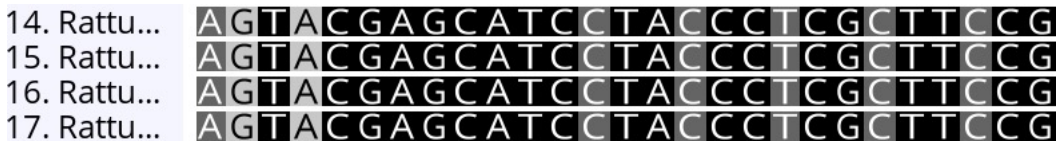
Loci: homology - common ancestor

Use single-copy genes



Loci: variable

Moderate variability for the phylogenetic scale of study

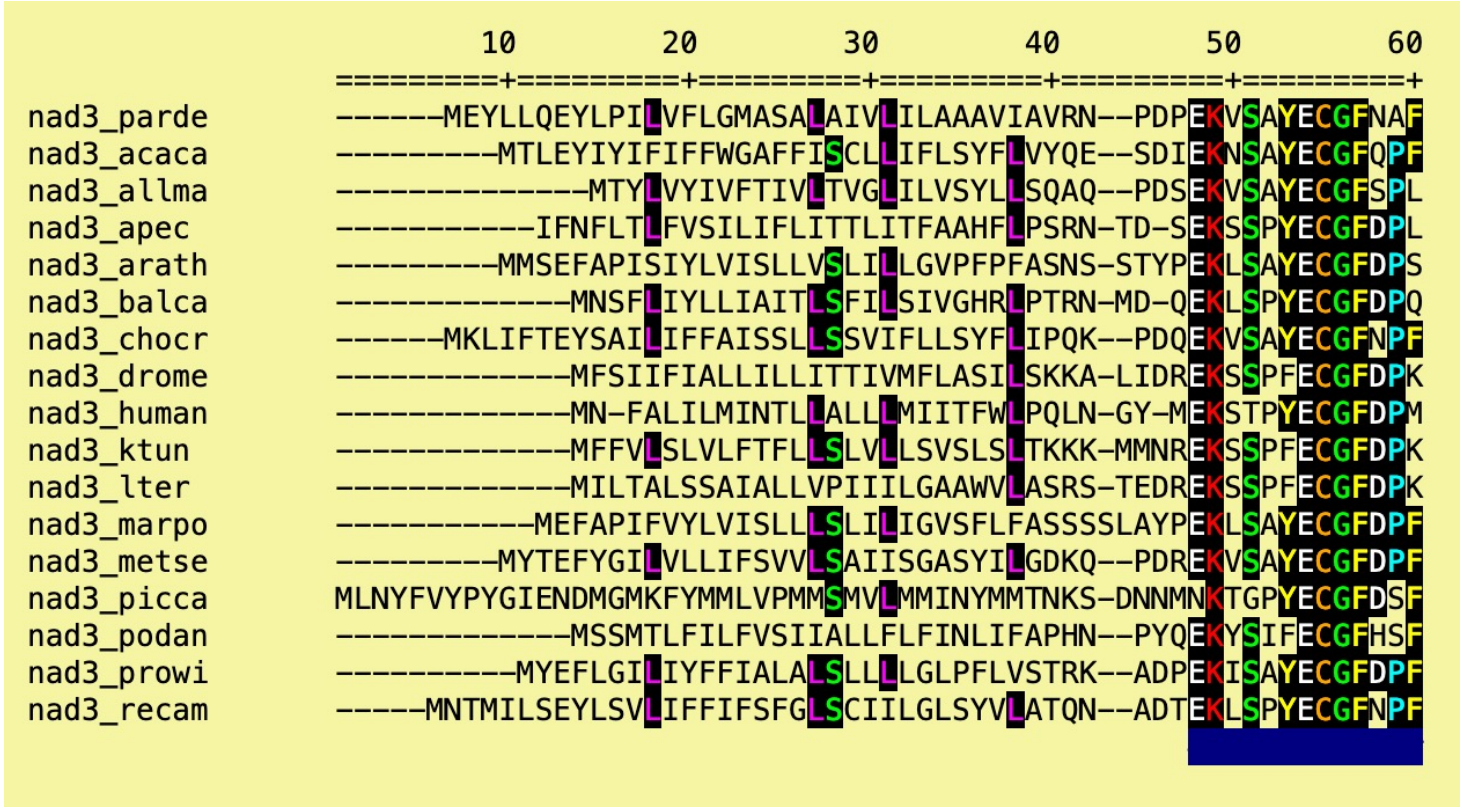


Too conserved: no variation

Too variable: difficult to find homology

Right amount of variation

Loci: homology in regions

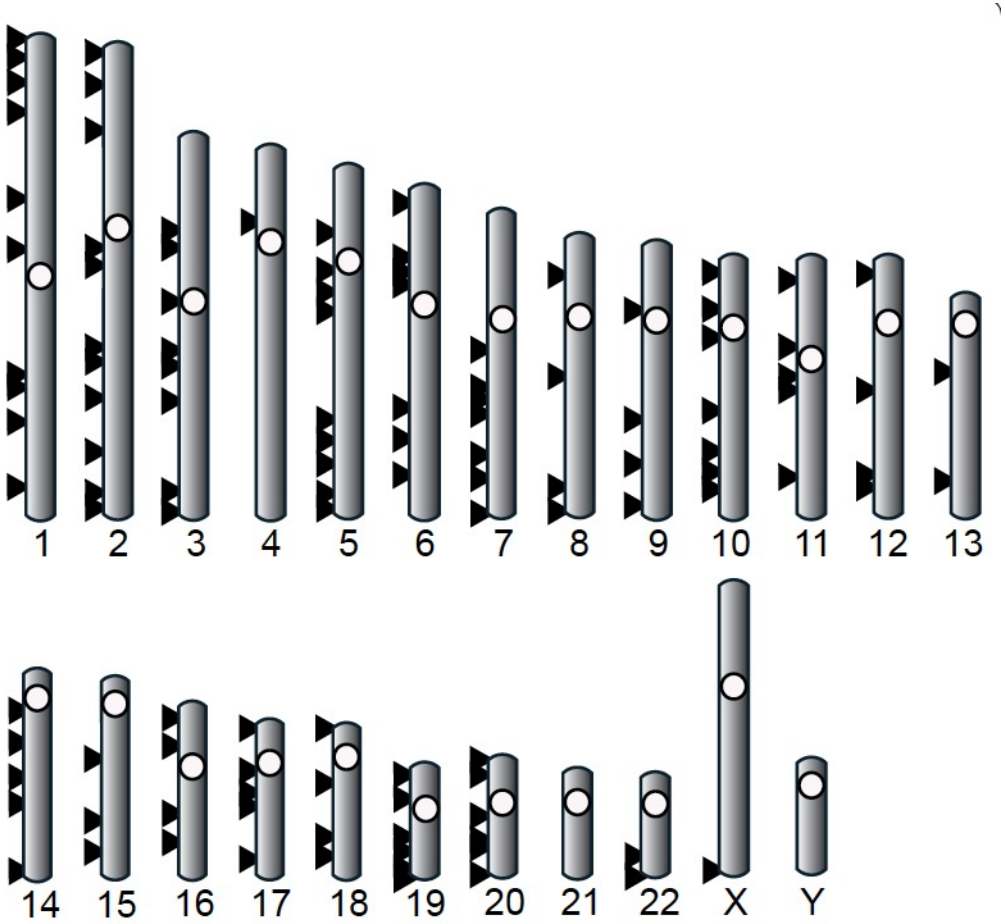


Software to clean non-homologous regions: gblocks, divvier, trimal

Loci: independent

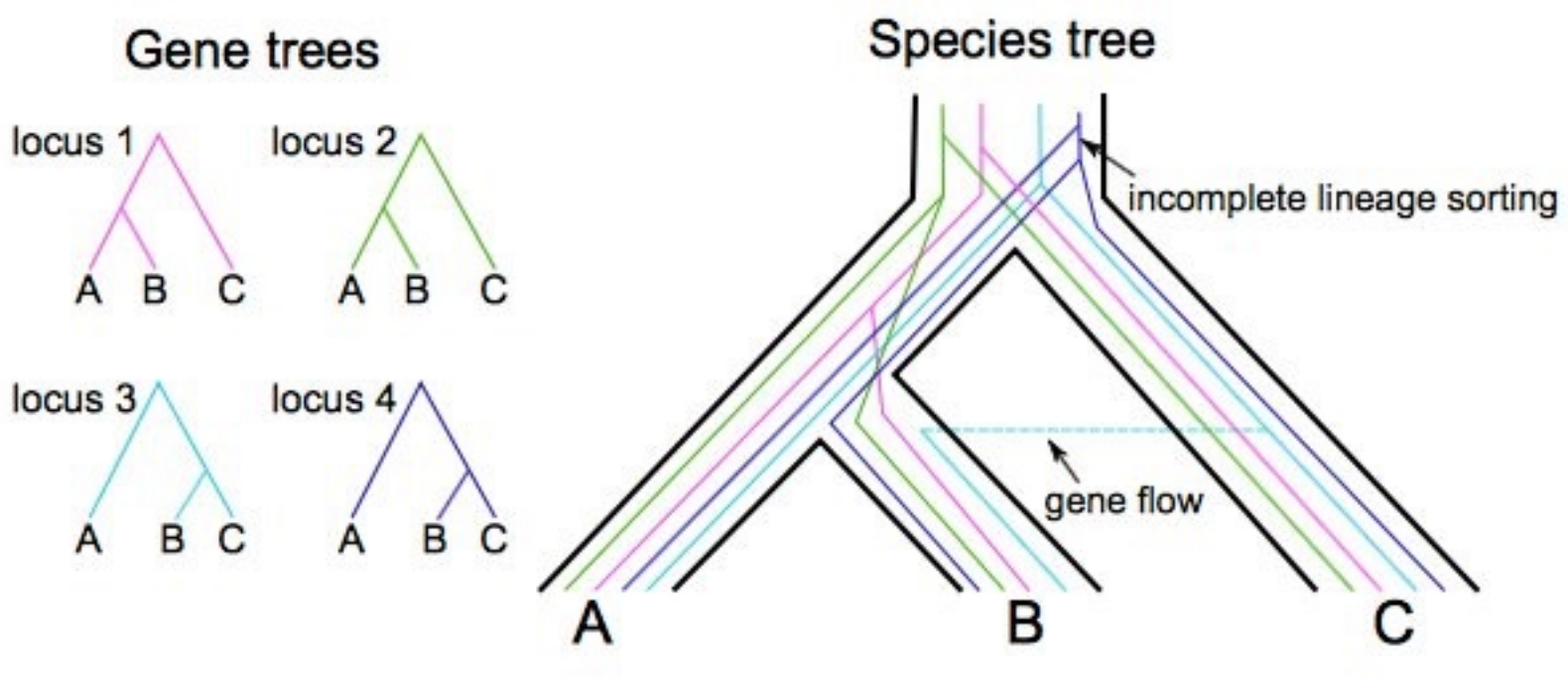
- Independent => recombination

(eg, different chromosomes)



Loci: independent

- Many independent loci to overcome bias by incomplete lineage sorting



Phylogenetic inference workflow

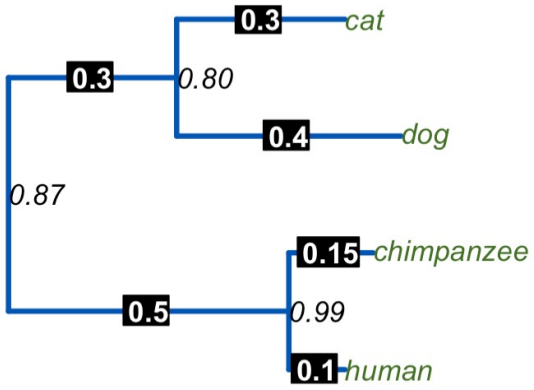
Multiple sequence alignment

Inference

Phylogenetics methods

Choose method

Tree



Phylogenetics methods

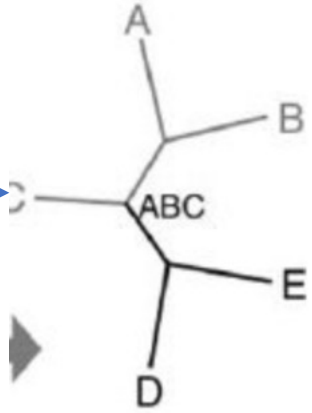
- Distance-based (eg. neighbour joining)
- Maximum likelihood
- Bayesian

Distance based: Neighbour joining

Character matrix

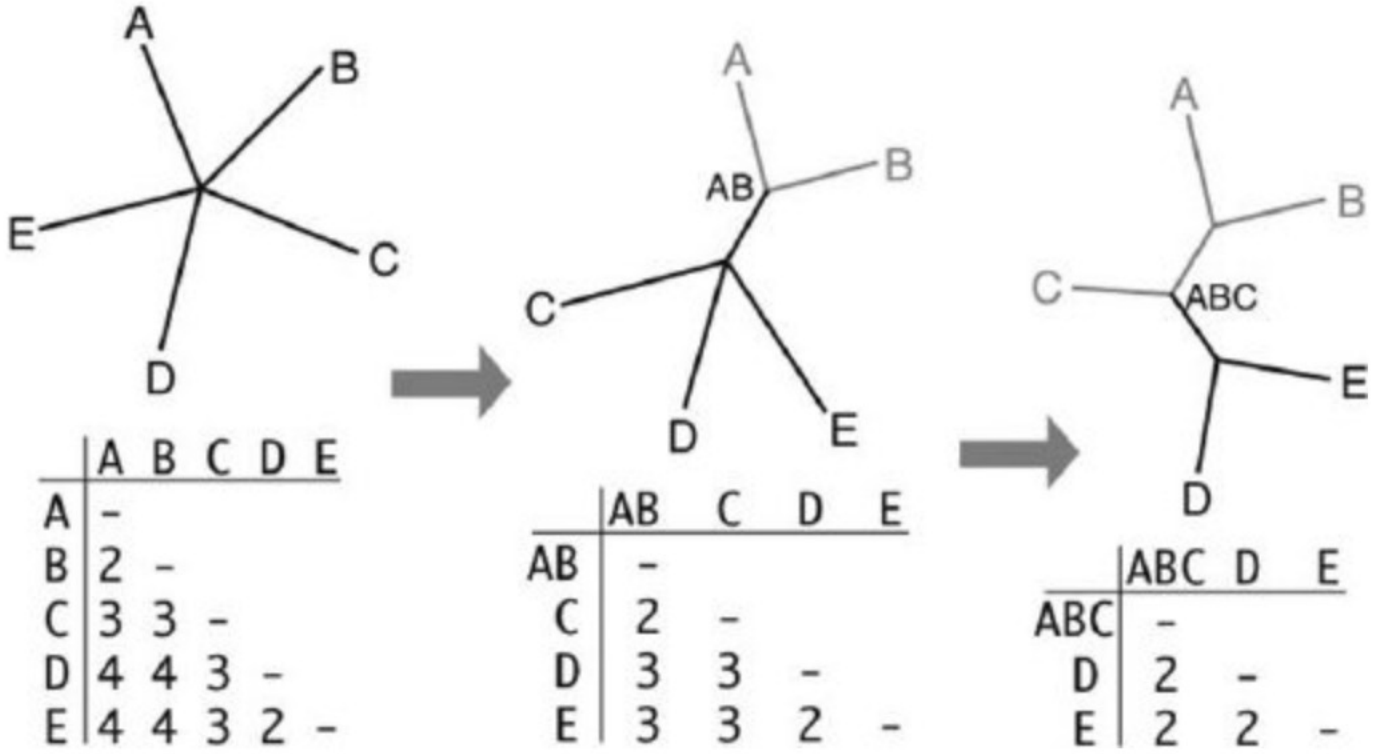
	A	B	C	D	E
A	-				
B	2	-			
C	3	3	-		
D	4	4	3	-	
E	4	4	3	2	-

Distance matrix



Tree

Neighbour joining



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

	AB	C	D	E
AB	–			
C	2	–		
D	3	3	–	
E	3	3	2	–

	ABC	D	E
ABC	–		
D	2	–	
E	2	2	–

Software: R (ape), MEGA, Geneious

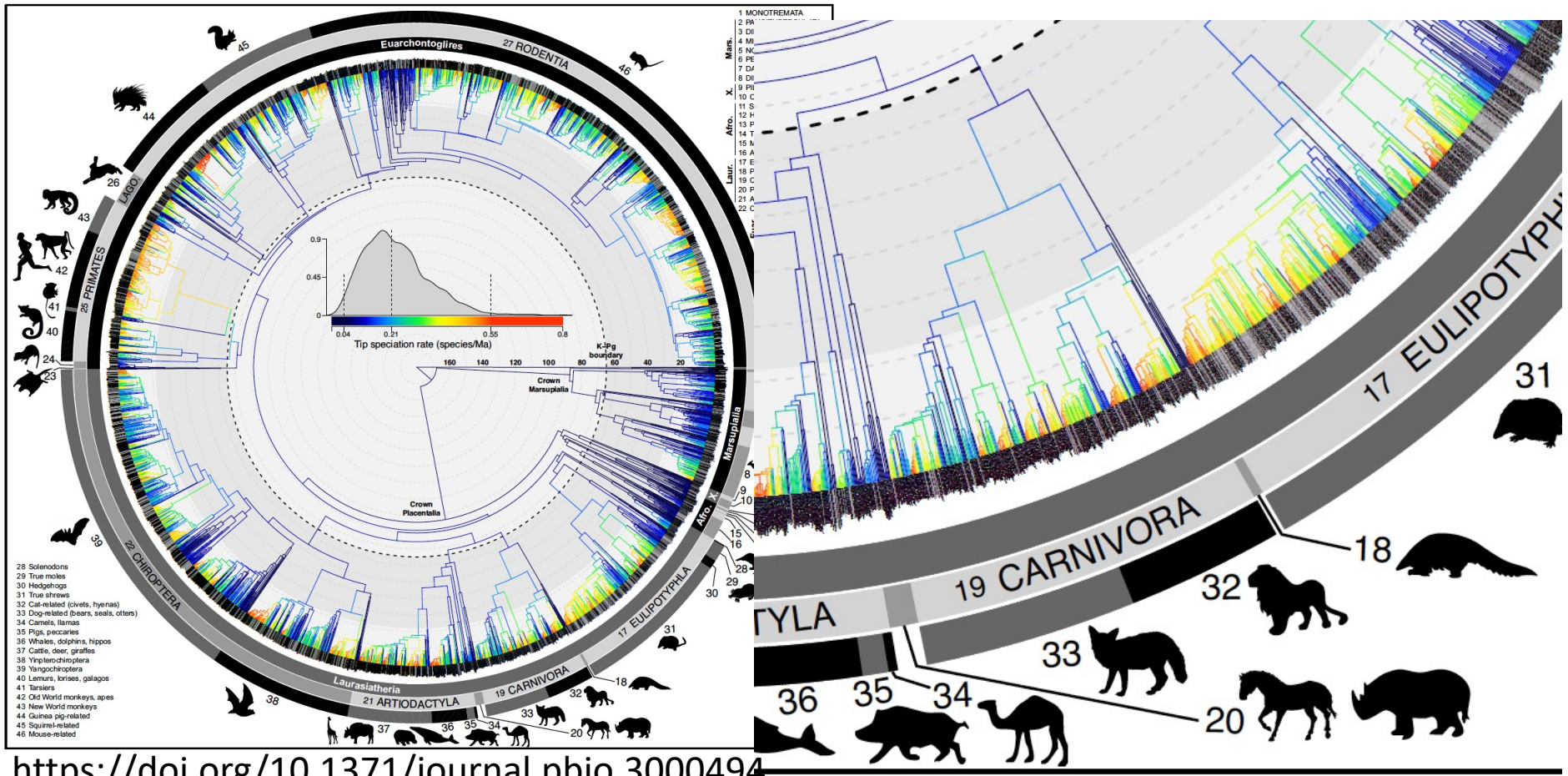
Exercise 2

Neighbour joining tree reconstruction

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

Exercise 2: dataset

Cytochrome b from carnivores + pangolins at outgroup

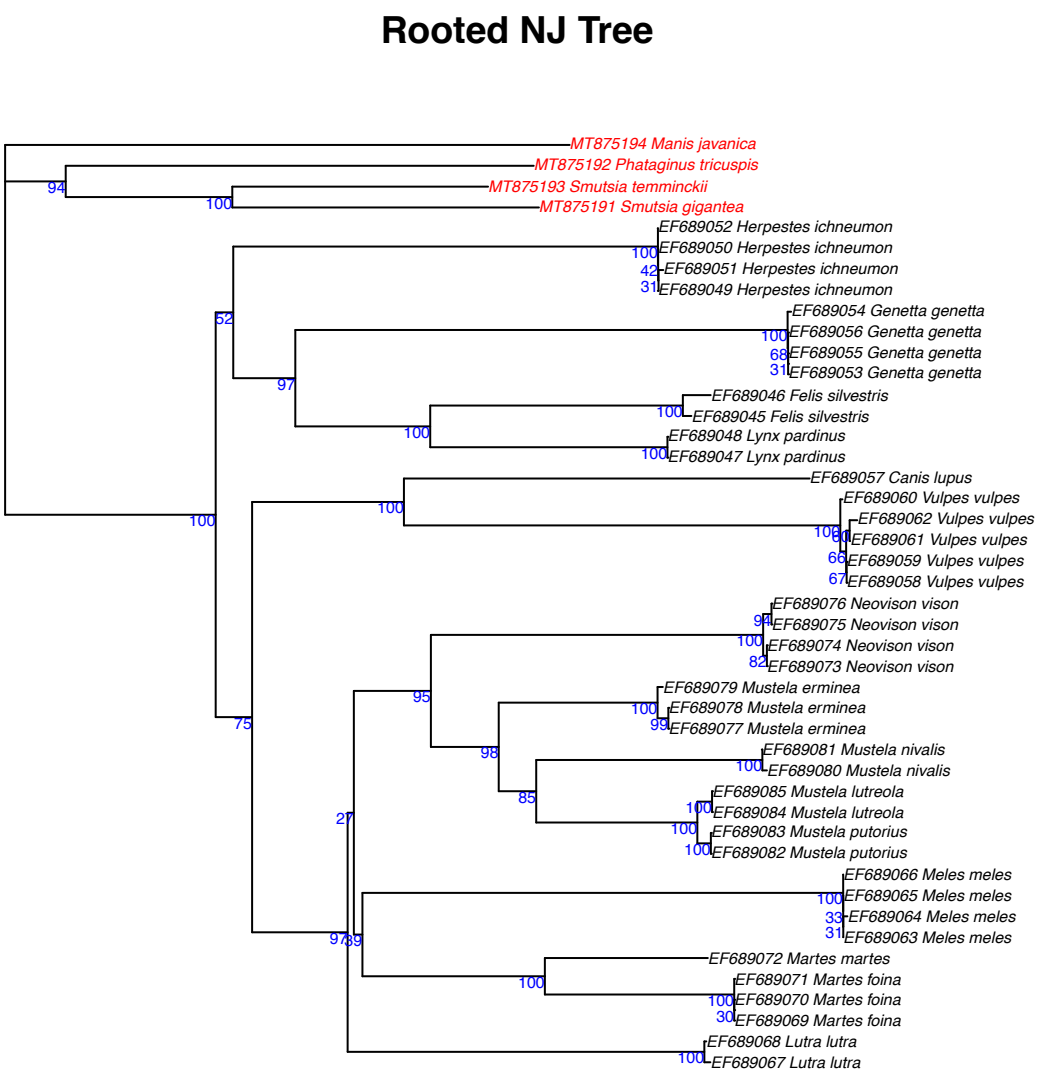


https://doi.org/10.1371/journal.pbio.3000494

Exercise 2: dataset

Cytochrome b from carnivores + pangolins at outgroup

Exercise 2: result



Phylogenetic inference workflow

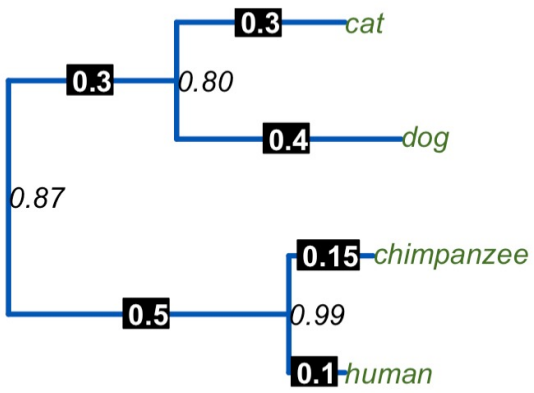
Multiple sequence alignment

Inference

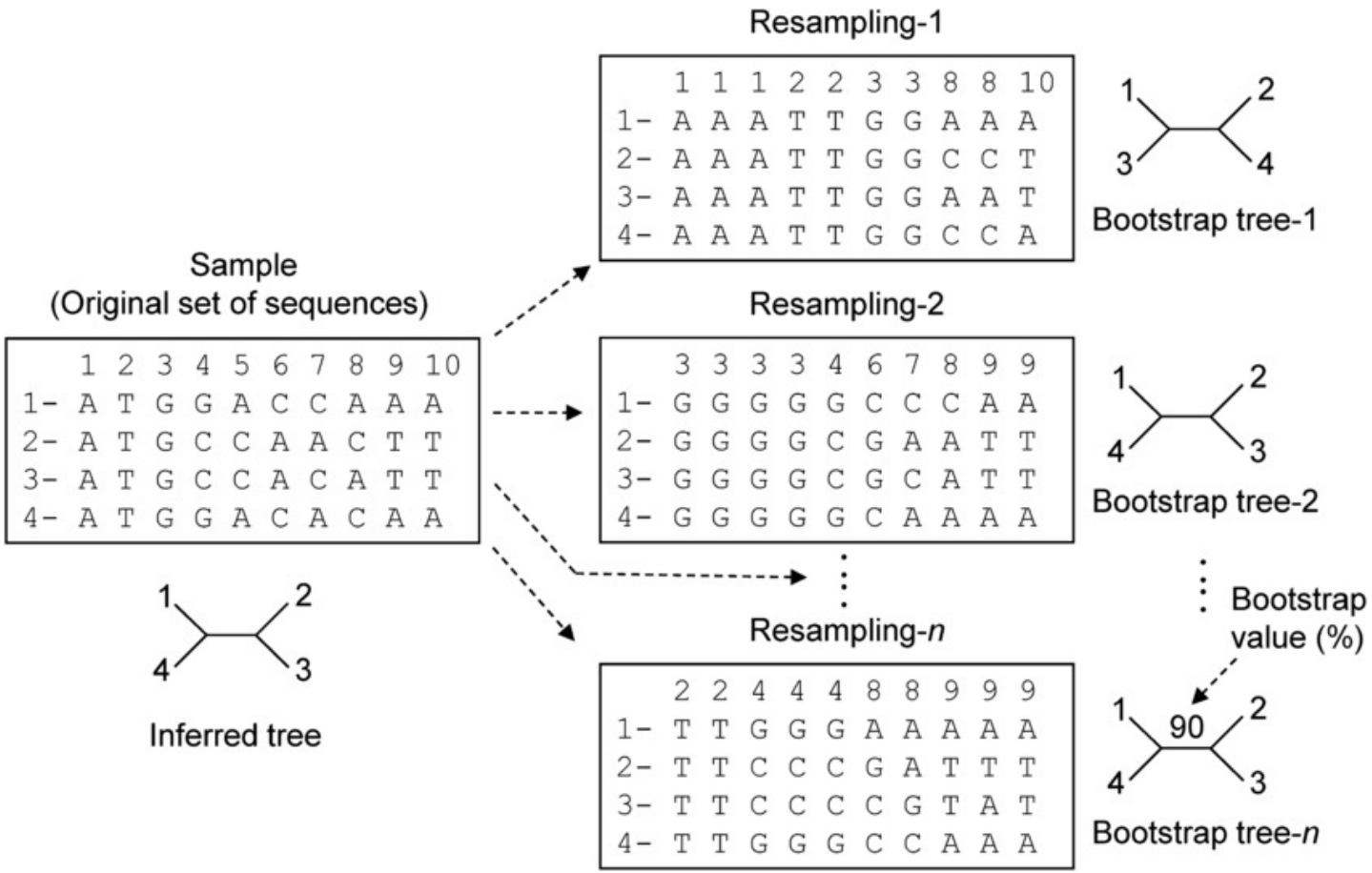
Phylogenetics methods

- 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

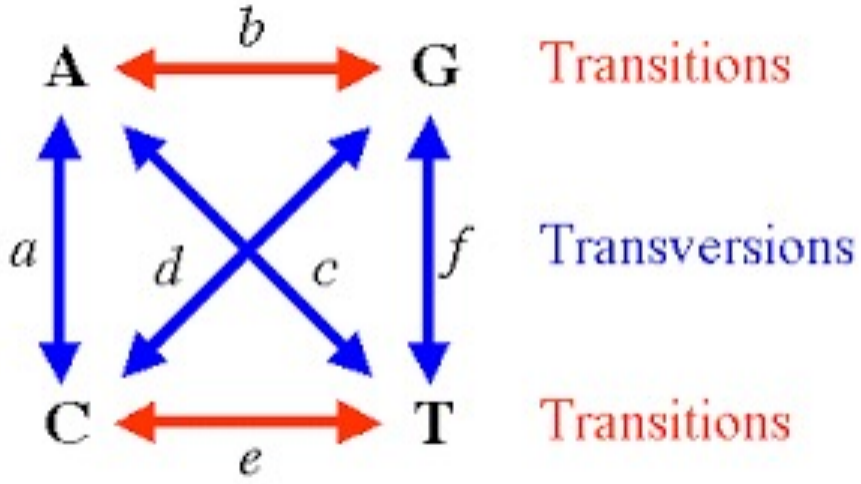
BEAR

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

Models of substitution

~base frequencies
~transition/transversion matrix



Trade off:
No of parameters vs simplification

Rate heterogeneity

- Variation in mutation rate across sites.

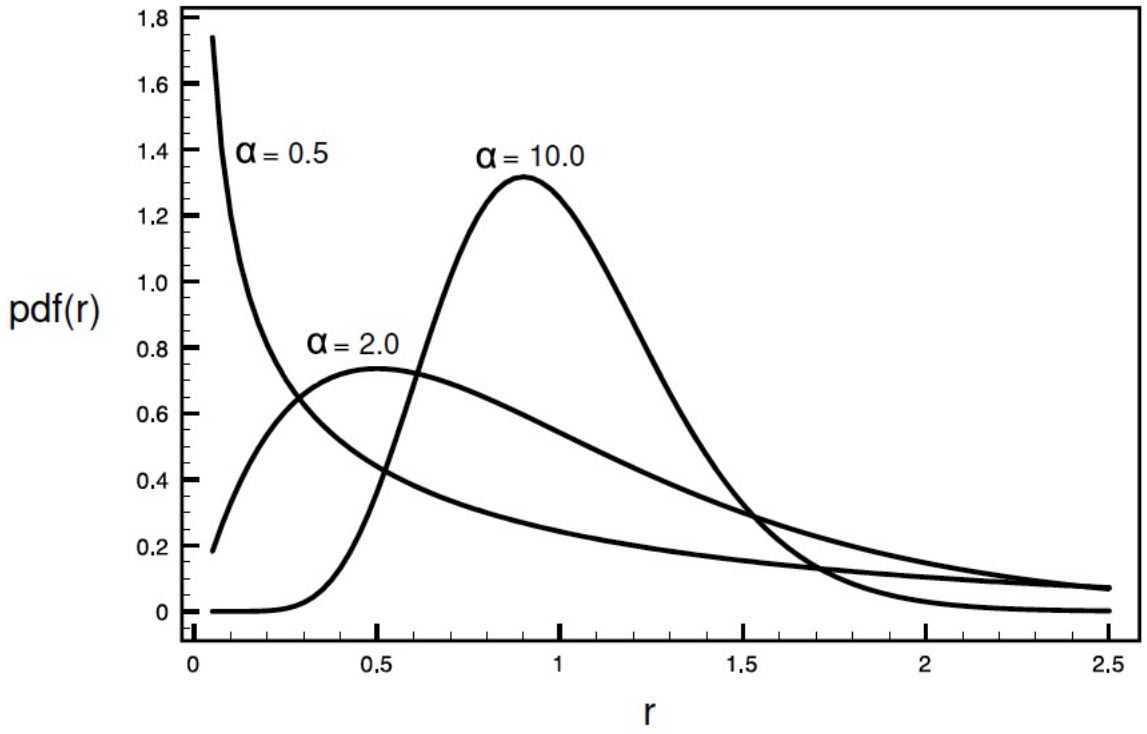
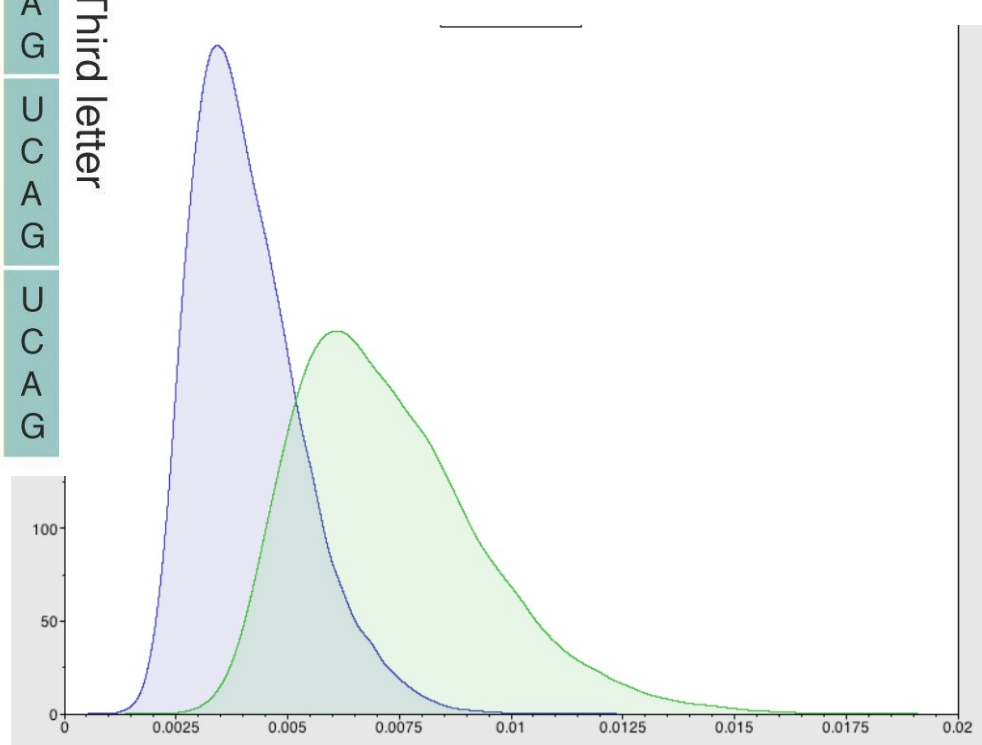


Fig. 4.8 Different shapes of the Γ-distribution depending on the α-shape parameter.

Rate heterogeneity

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

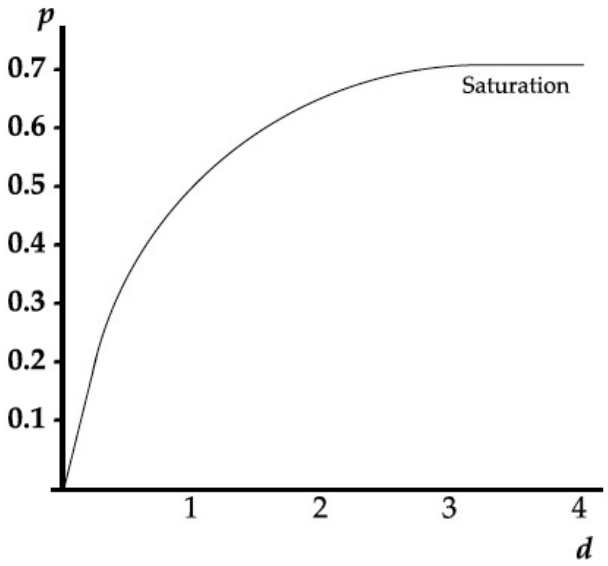


Software to test models of substitution

- 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.

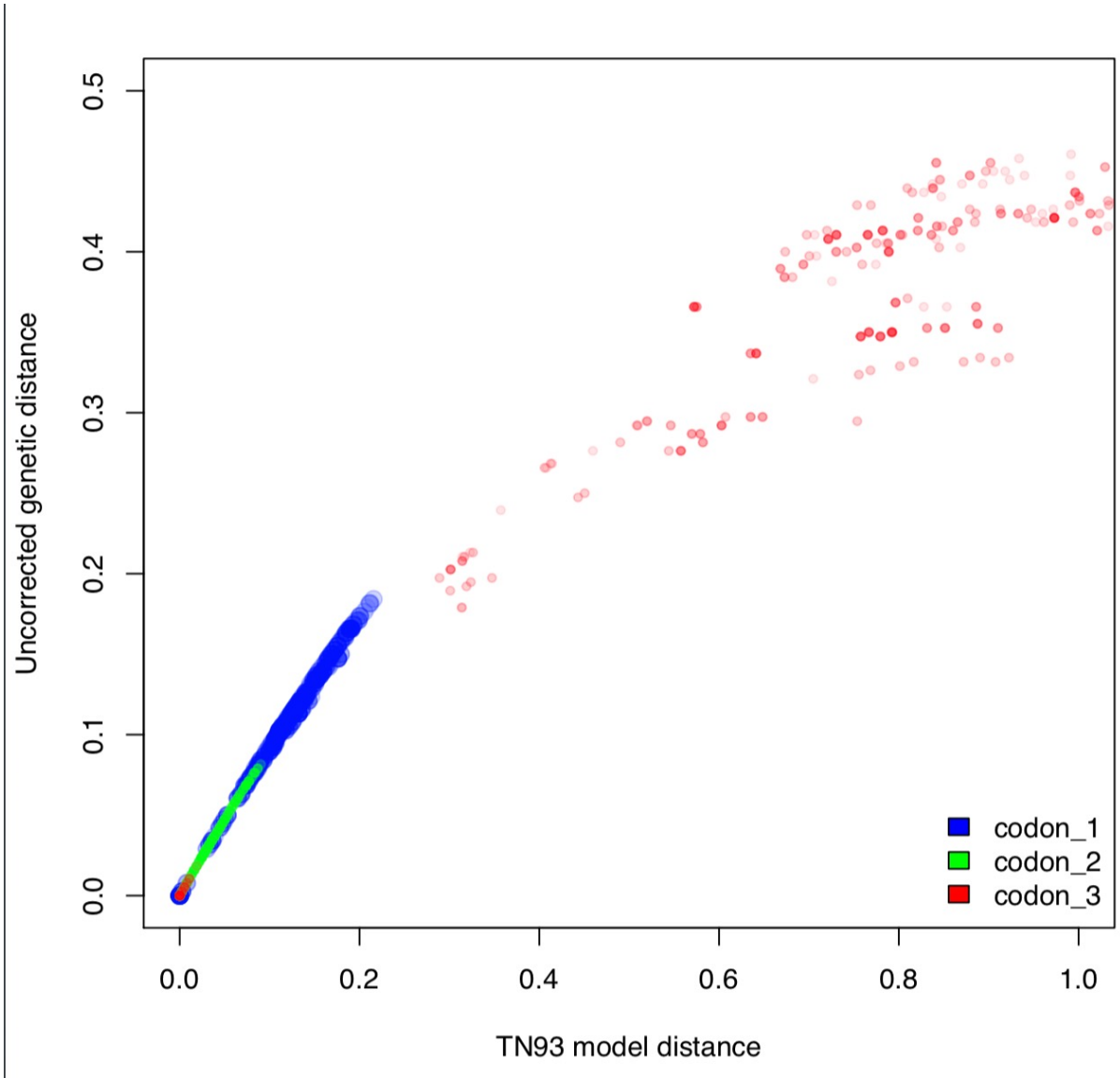


Relationships between expected ***genetic distance*** *d* and observed ***p-distance***.

Second letter

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

Exercise 3


Uncorrected genetic distance
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
TN93 model distance
codon_1
codon_2
codon_3

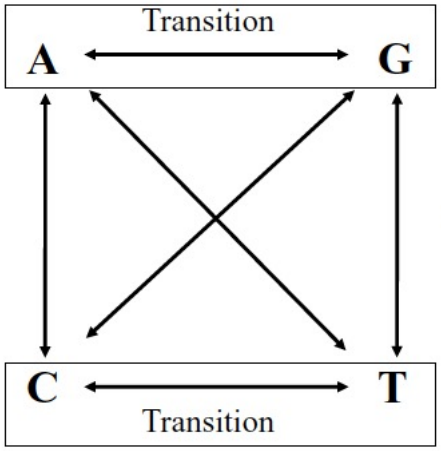
Phylogenetics methods

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

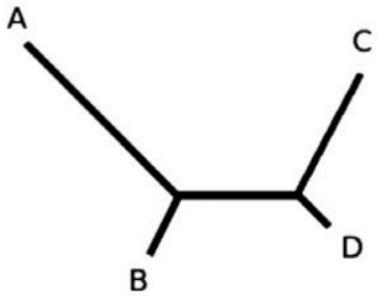
Maximum likelihood

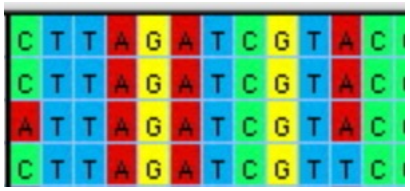
- probability of observing the data given and particular model of evolution and evolutionary history.
- Computationally intense
- Outperforms other methods
- Allows explicit evolutionary models
- Outputs likelihood of tree topologies

Maximum likelihood

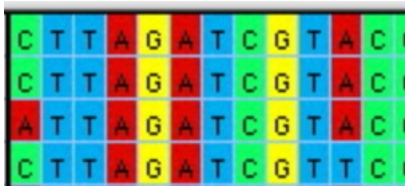


Transversion



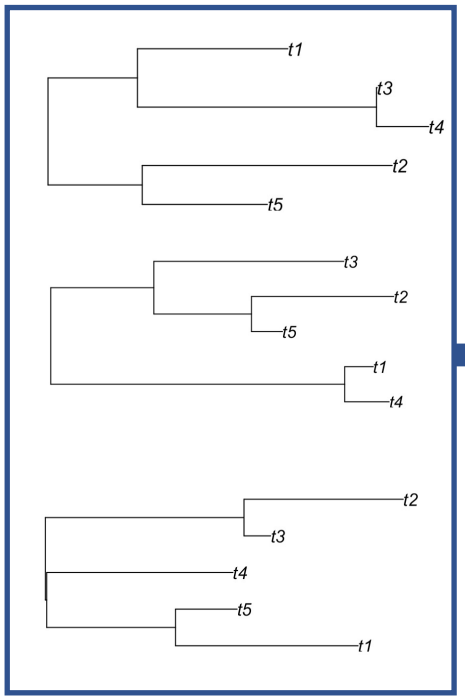
P ()

Example: RAxML Maximum likelihood



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.

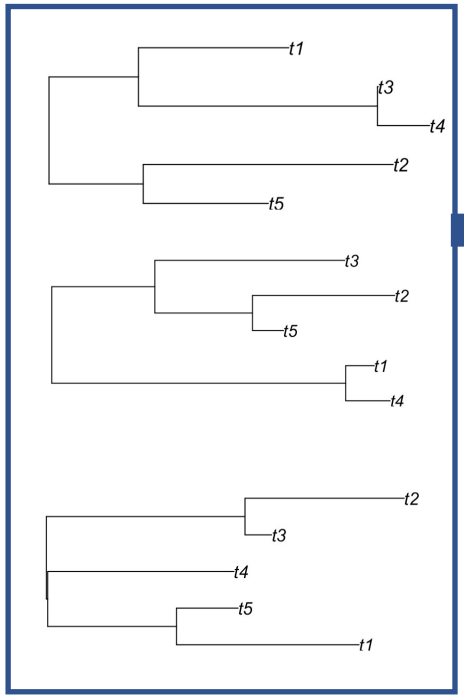
Maximum parsimony

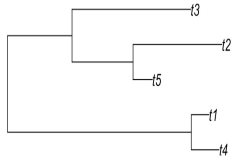


Tree search: optimization

Best trees

Branch-length adjustment



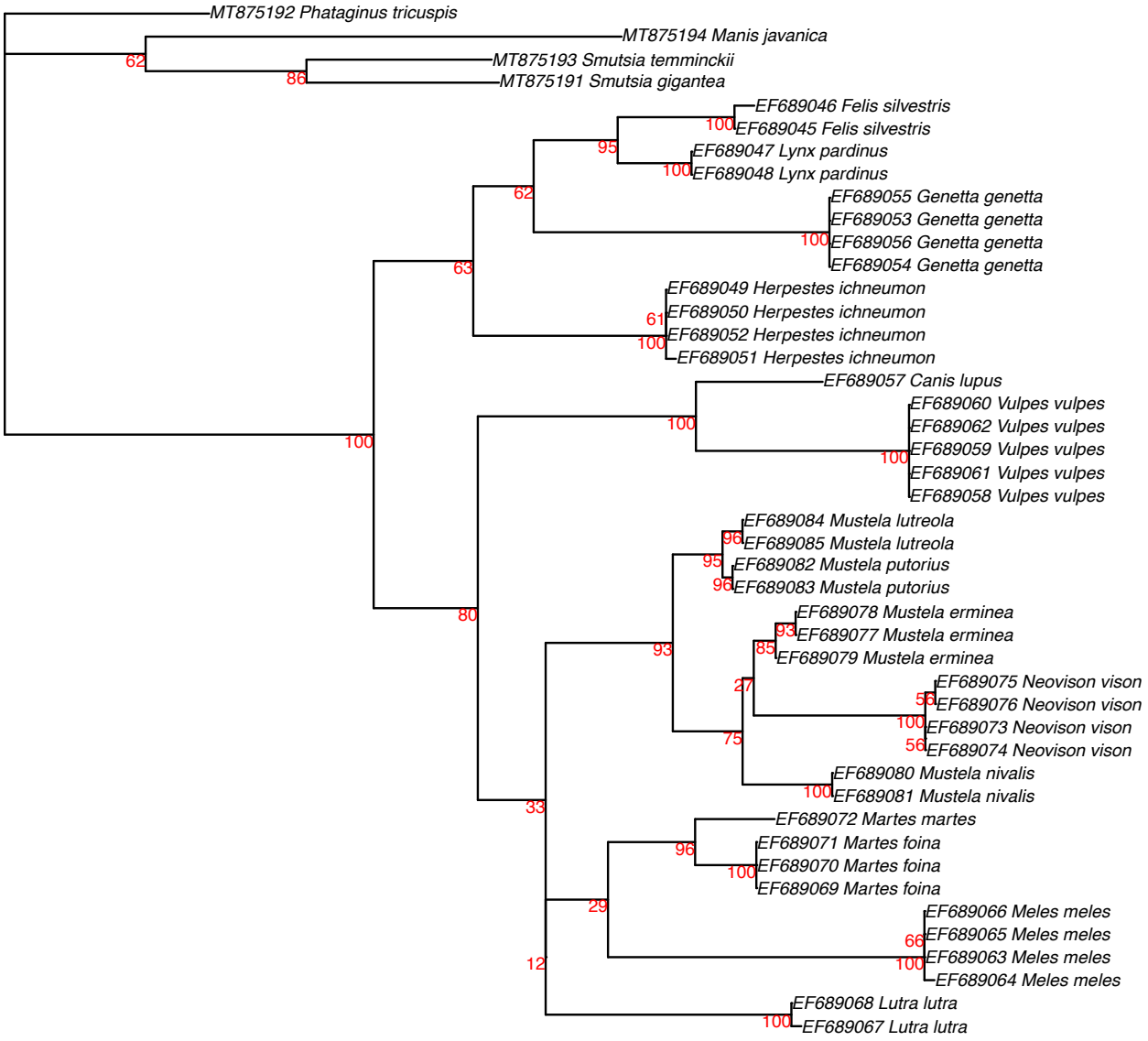


Best trees found

Exercise 4: Maximum likelihood

Reconstruct phylogenetic tree with ML in R

Exercise 4: Maximum likelihood

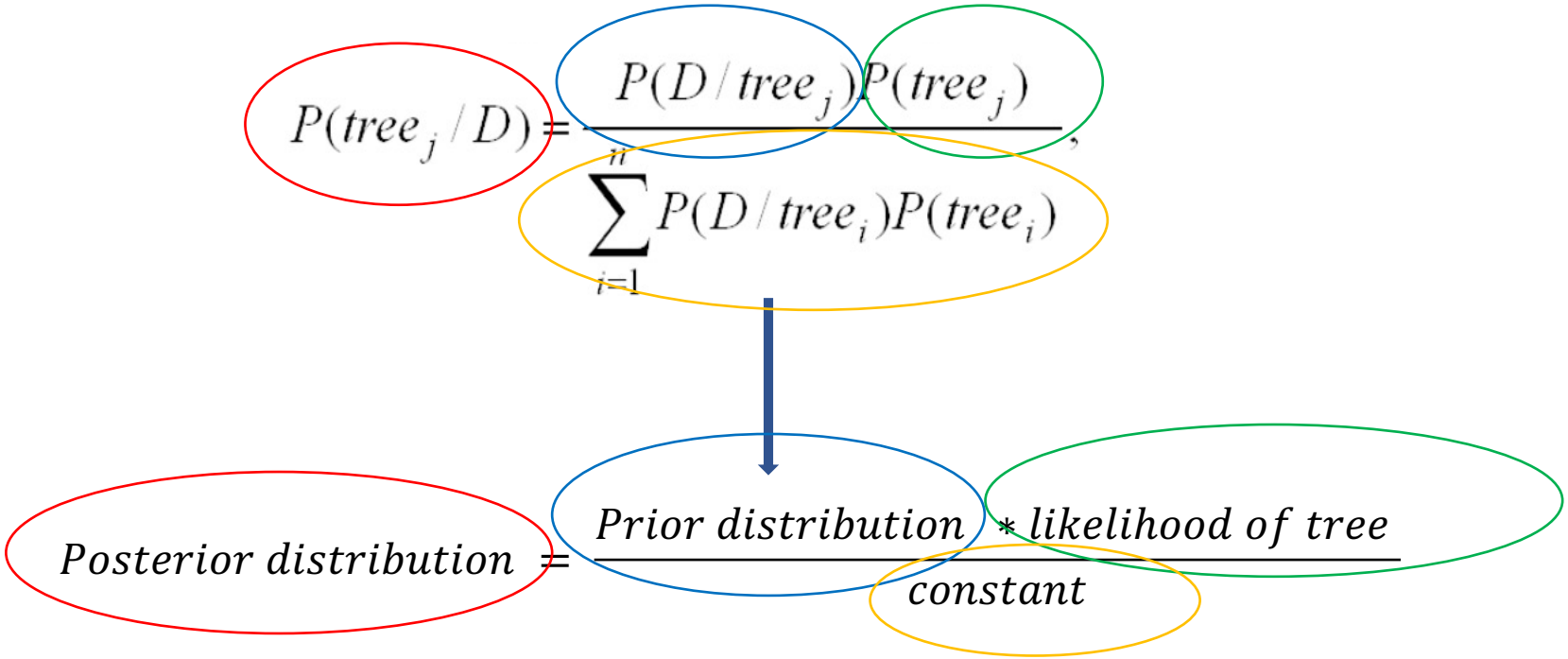


Phylogenetics methods

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

Bayesian inference

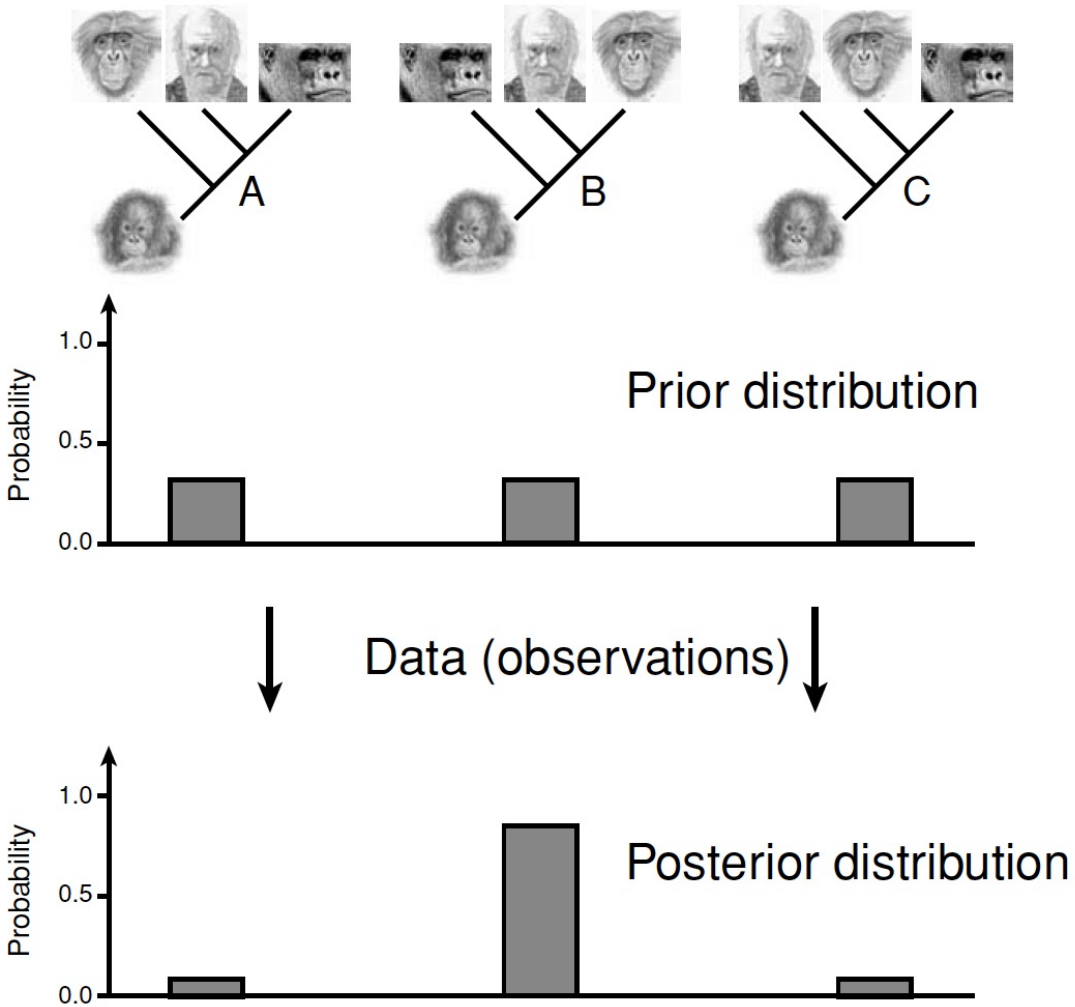
Bayes theorem applied to phylogenetics:



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

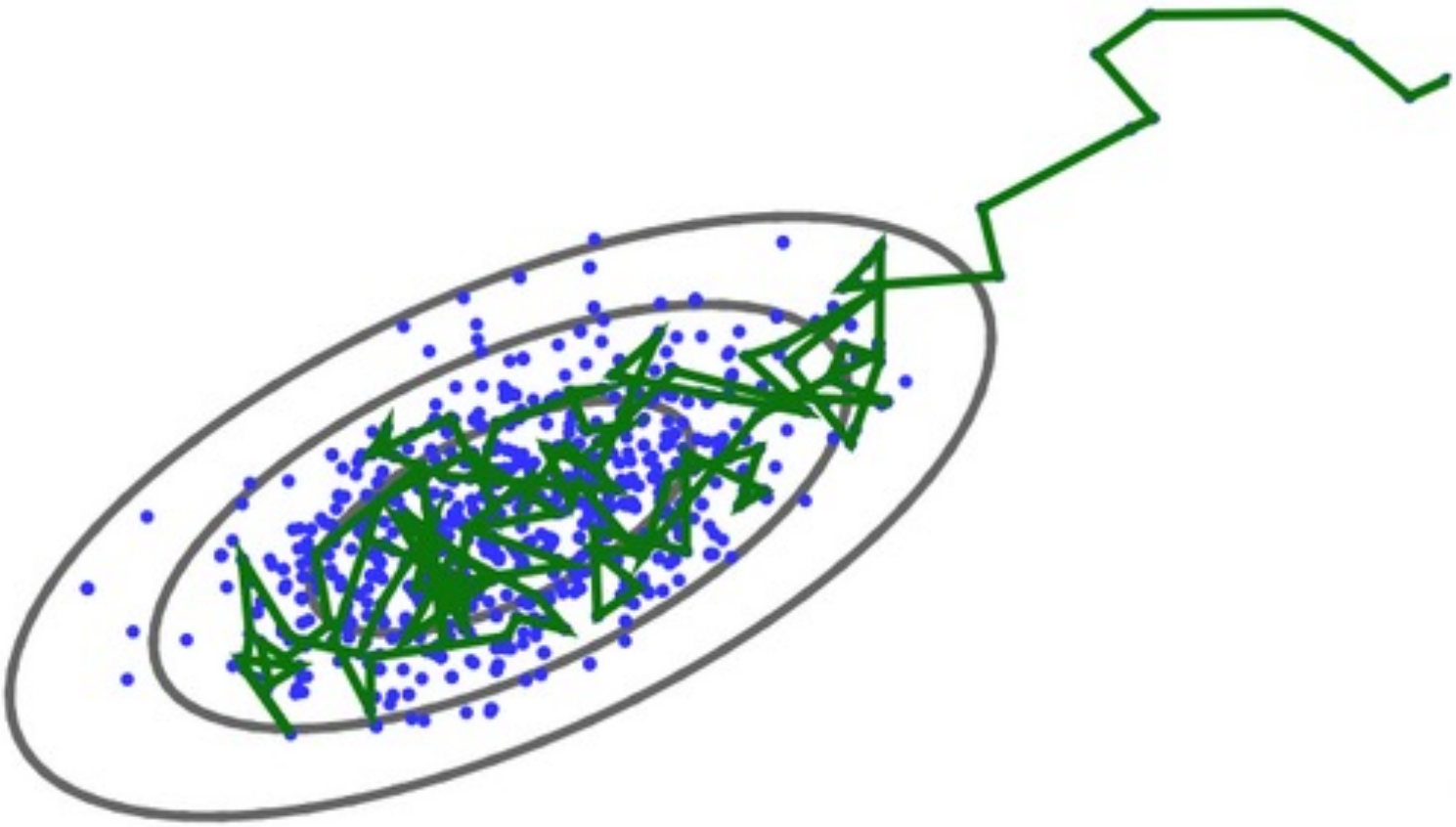
Bayesian inference

Allow the incorporation of expertise knowledge as prior info

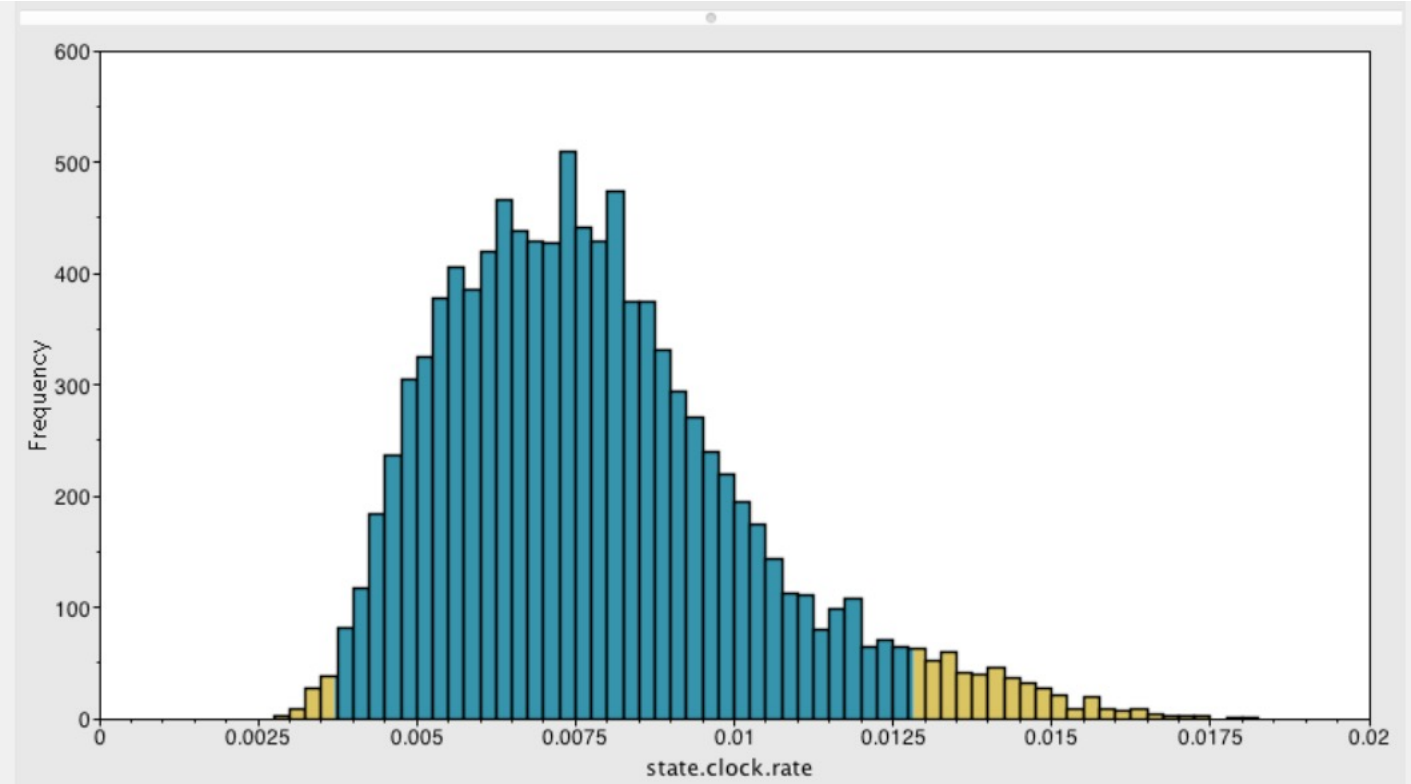


$$Posterior\ distribution \propto Prior\ distribution * likelihood\ of\ tree$$

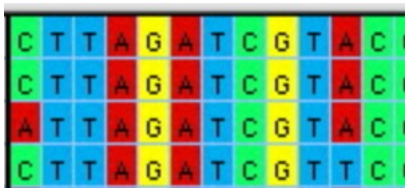
MCMC sampler



Interpreting output: HPD 95%
(high posterior density)



Beast workflow



Downsampled matrix:
- 1 sample/species
- no outgroups

BEAUTi

Prepare input for Beast
.xml file

Output.log: posterior distribution
Ouput.trees: trees

Beast2
Bayesian evolutionary analysis by sampling trees

Tracer

Check convergence
From logs

LogCombiner

Combine trees

TreeAnnotator

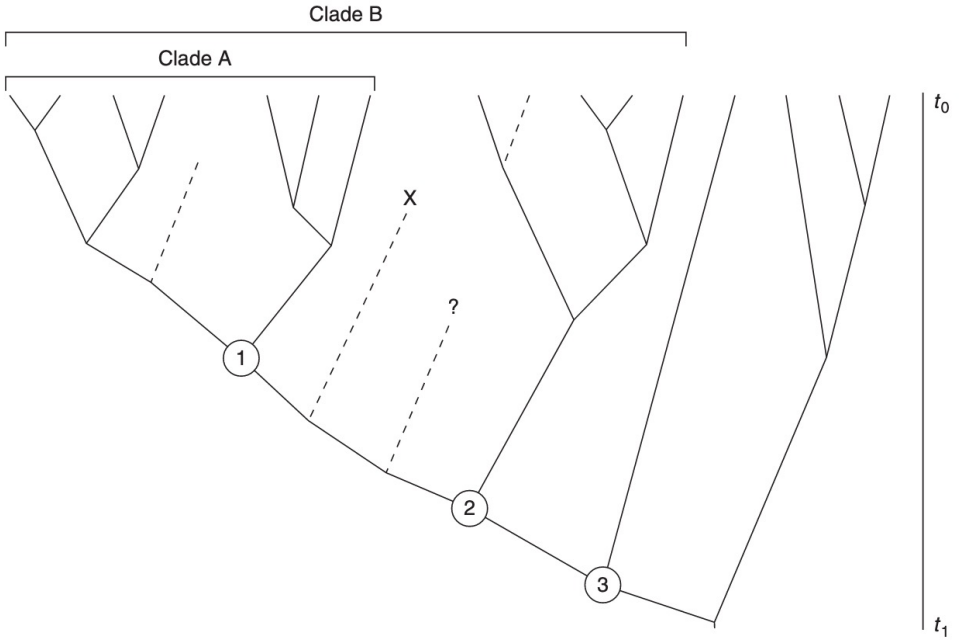
Visualize tree

FigTree

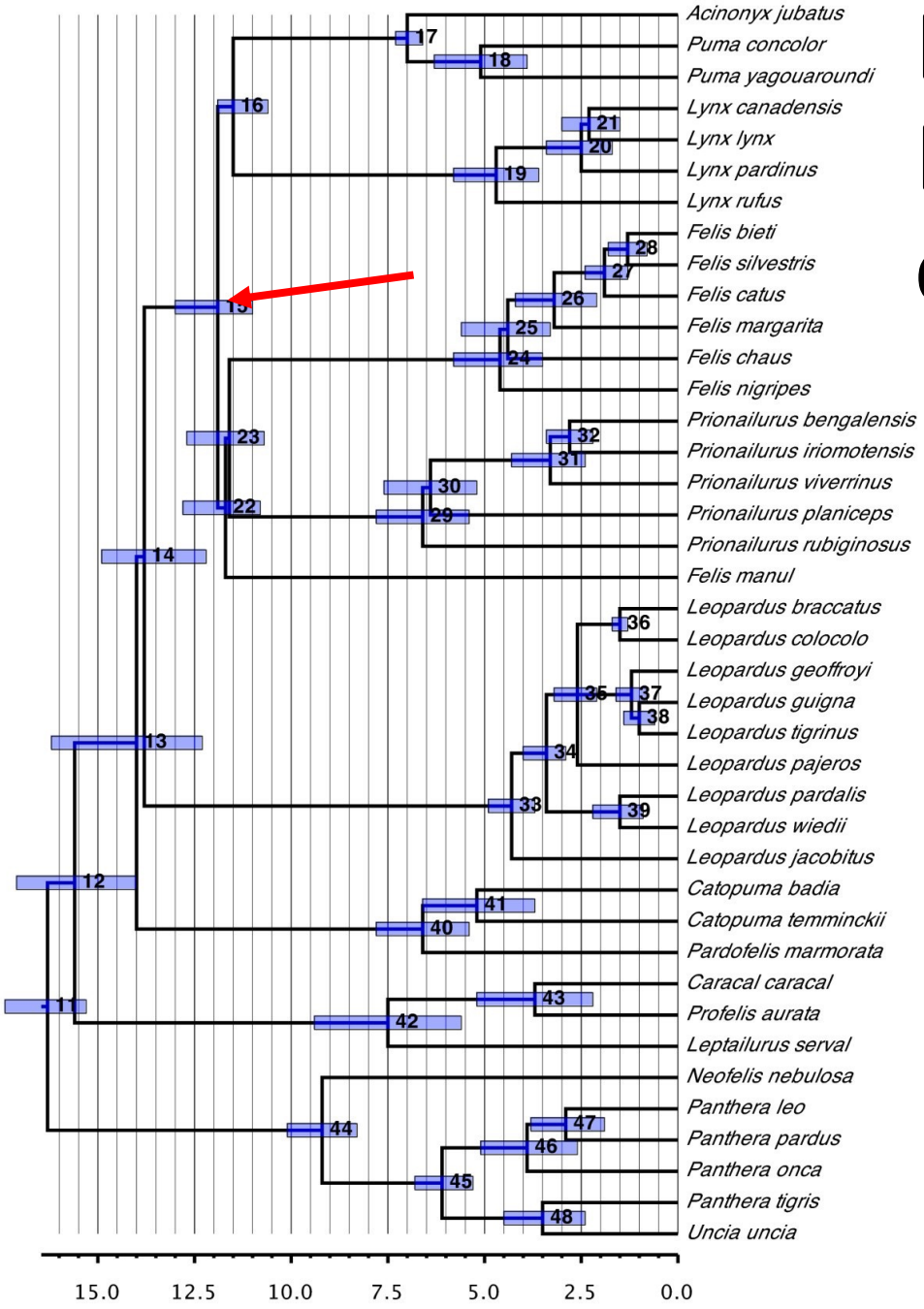
Visualize tree

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

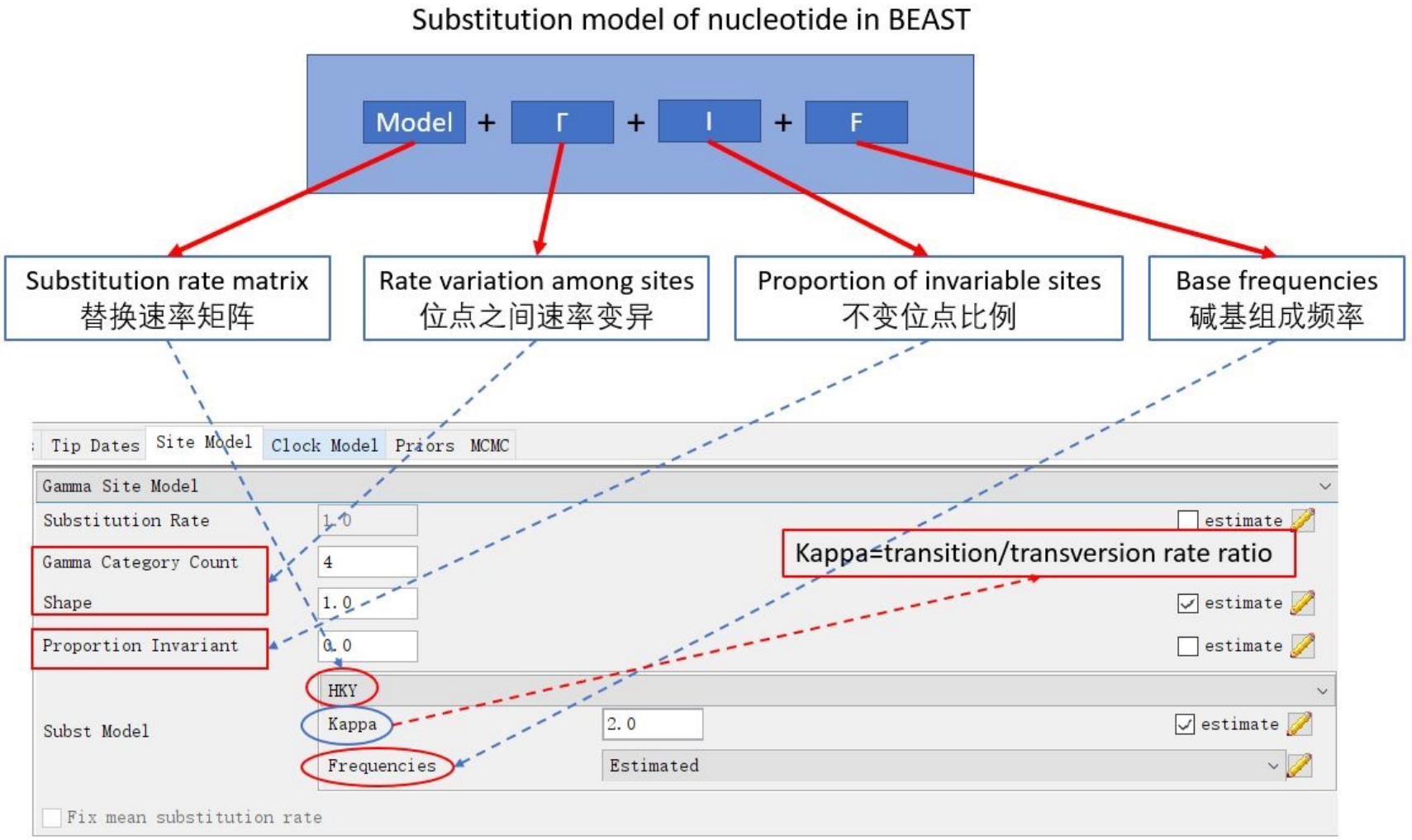
(10.1186/1741-7007-10-12)

Exercise 5: Bayesian inference

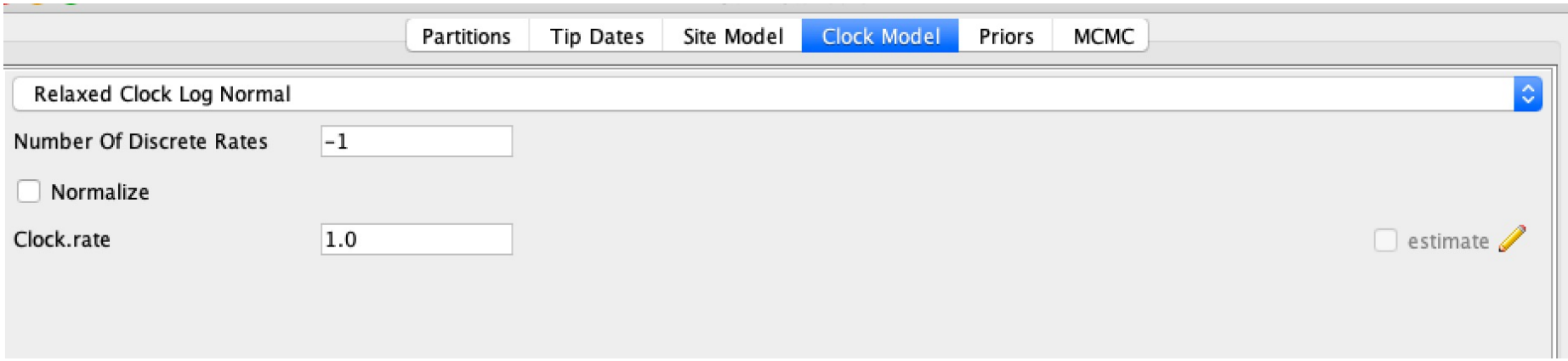
- BEAST:
  - Coalescent reconstructions.
  - Tutorial online:
    - https://github.com/csmiguel/2021_ConGenTropics/blob/main/tutorials/phylogenetics.md

- 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

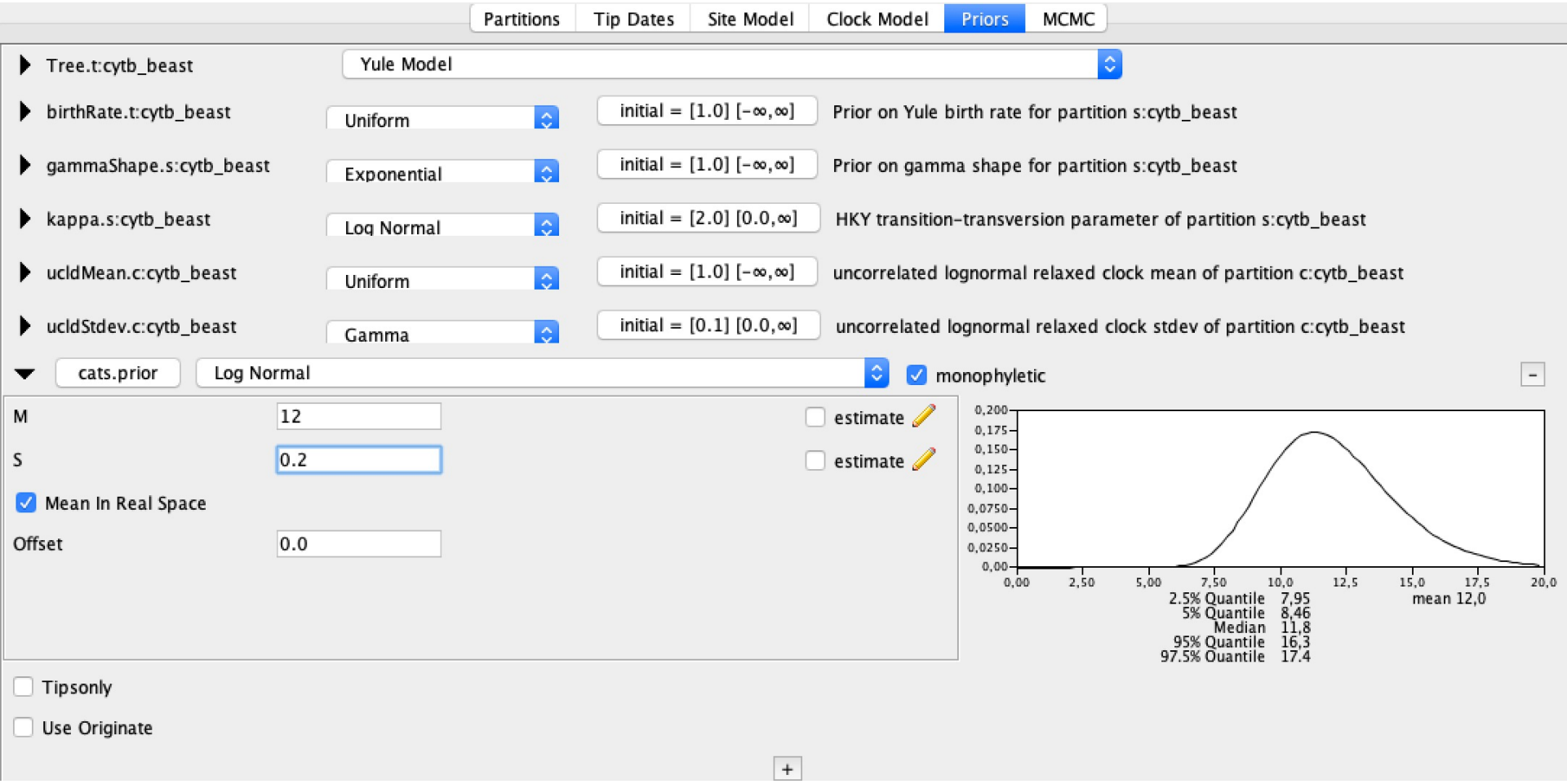


Exercise 5: Bayesian inference



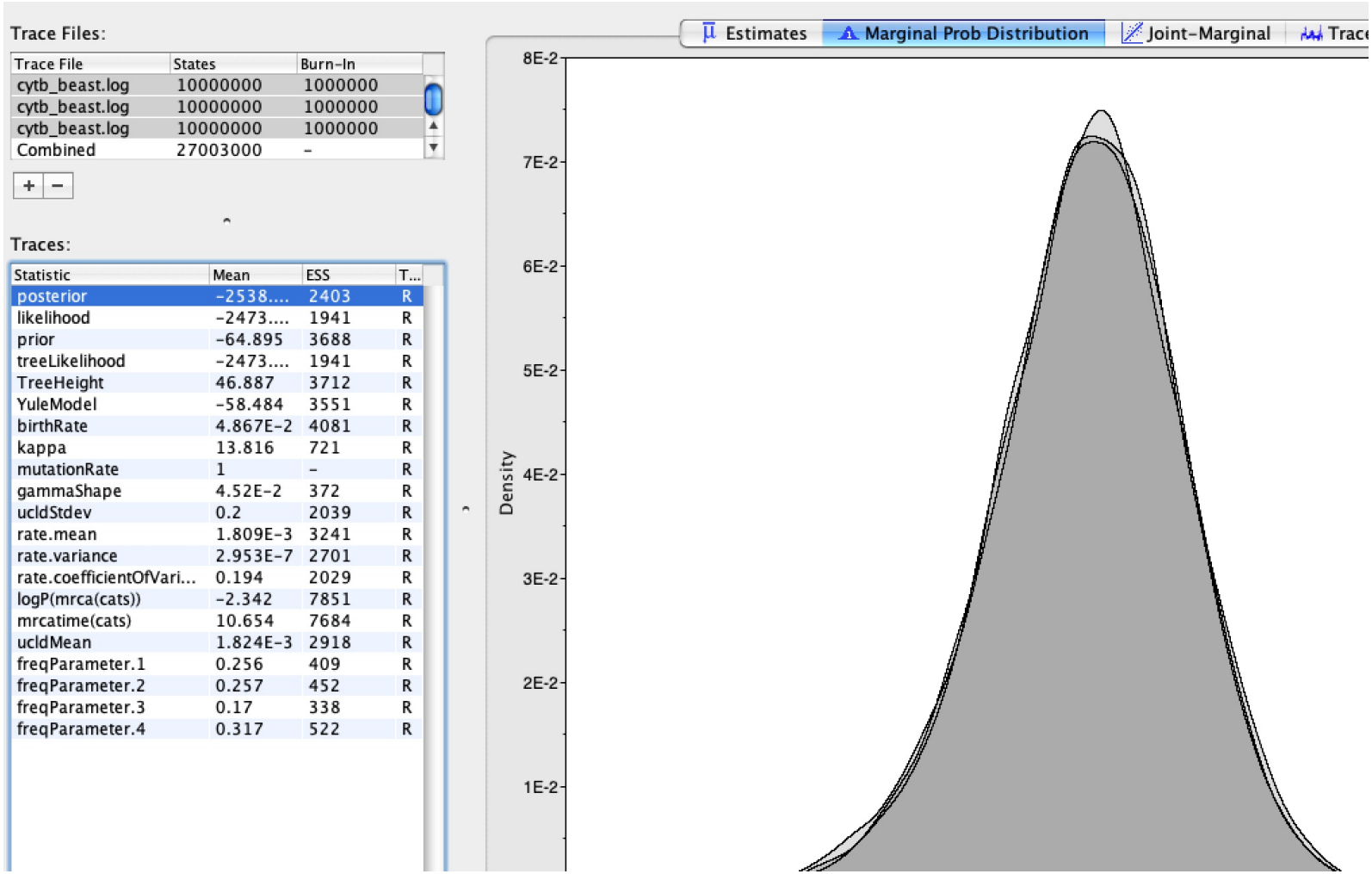
The uncorrelated relaxed molecular clock model is a **model of rate heterogeneity across** branches that assumes each branch has its own independent rate drawn from a single shared parametric rate distribution

Exercise 5: Bayesian inference



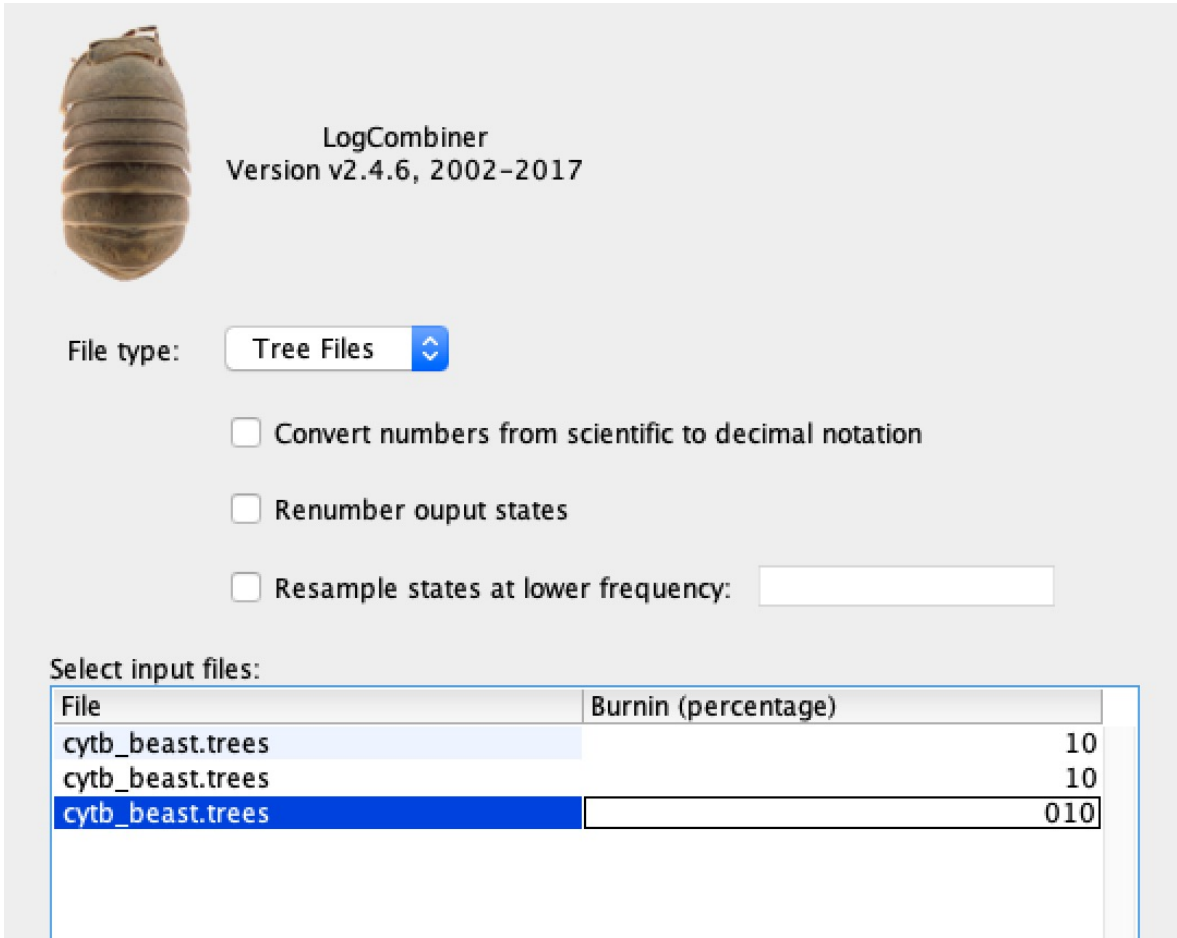
Yule tree prior that assumes a (unknown) constant lineage birth rate for each branch in the tree. This tree prior is most suitable for trees describing the relationships between individuals from different species.

Tracer: check convergence



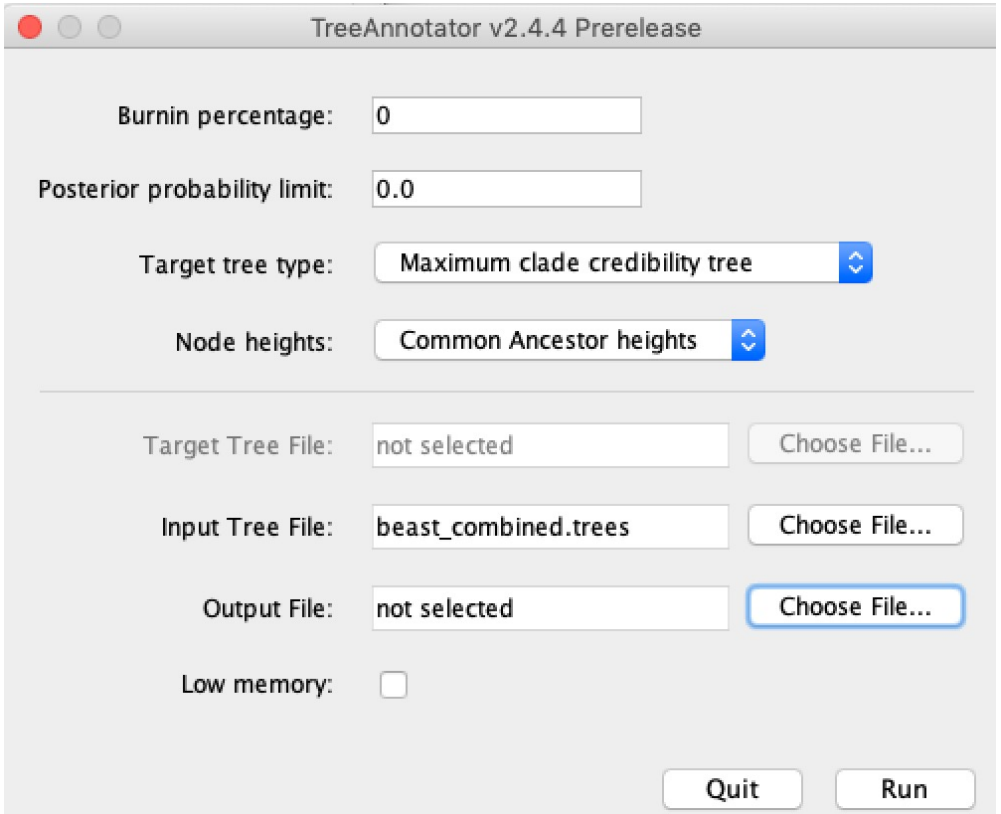
LogCombiner: combine trees from multiple chains.

- Discard 10% of trees

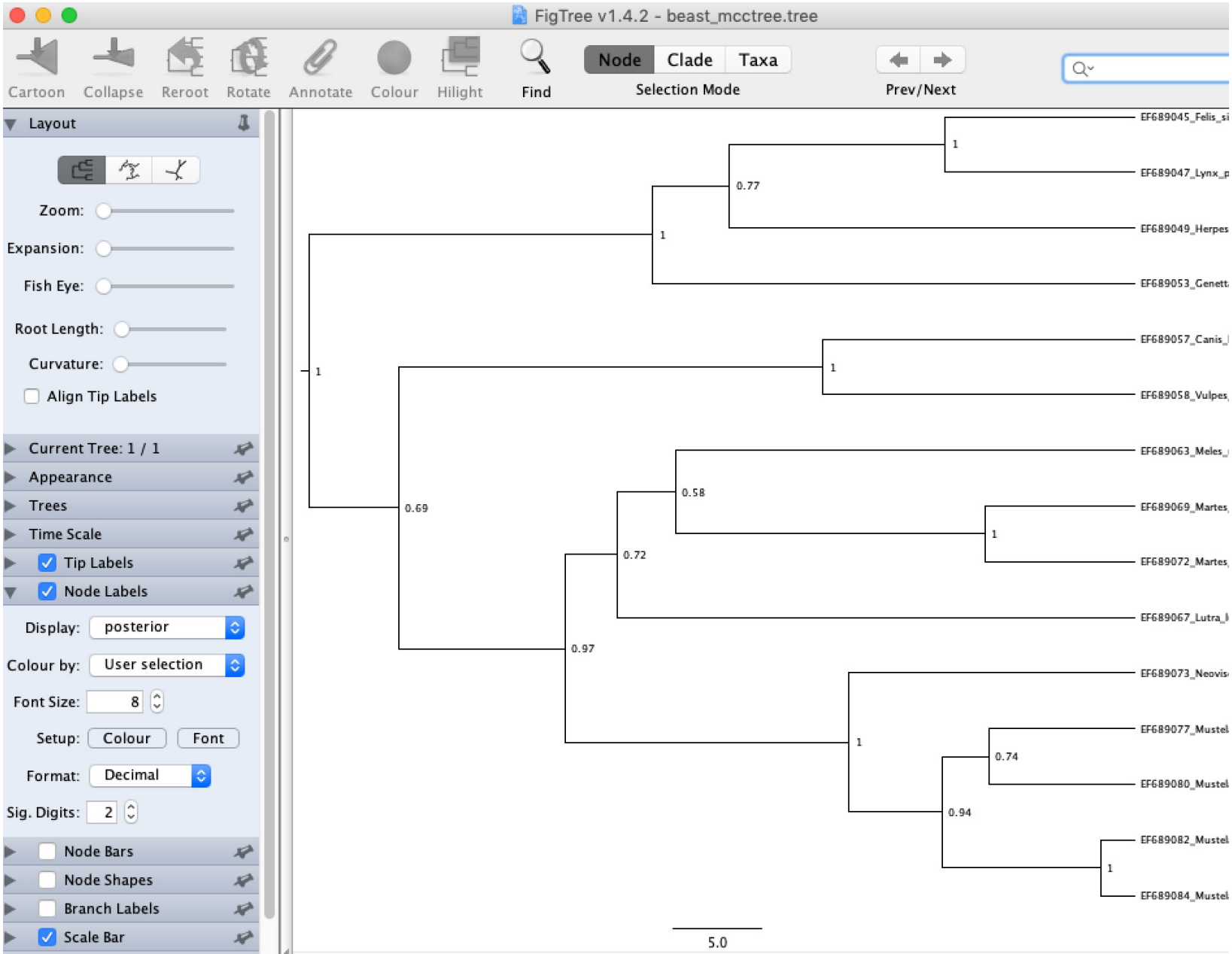


TreeAnnotator

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



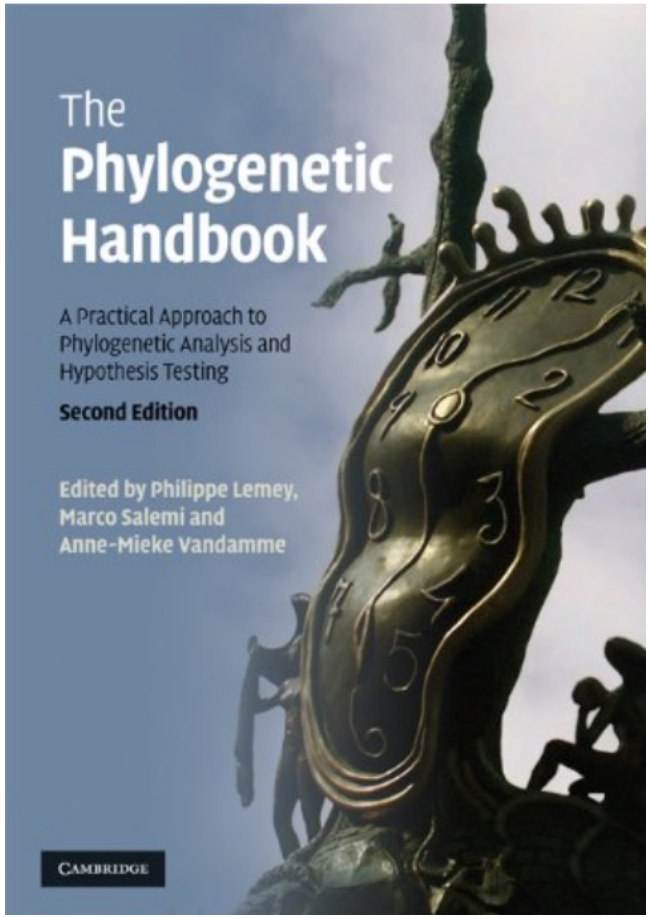
FigTree: visualize tree



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