Phylogenetics

Introduction to molecular dating

RL-V3 MPP

Rachel Warnock

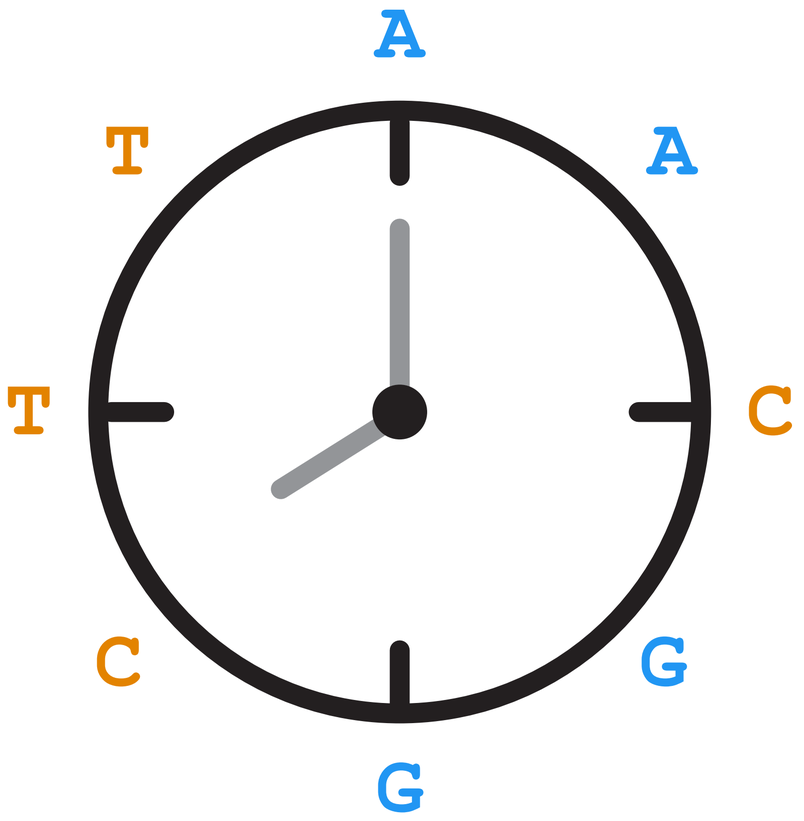
16.05.25

Can you try installing the following R packages?

- TreeSim
- FossilSim

Today’s objectives

- Homework
- **Recap**
  - Bayesian inference
  - MCMC
- Intro to molecular dating



Recap

Bayes' theorem

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model})\,\Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Bayes' theorem

The probability of the data given the model assumptions and parameter values

Likelihood

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model})\Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Bayes' theorem

This represents our prior knowledge of the model parameters

Priors

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model})\,\Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Bayes' theorem

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model})\,\Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Marginal probability

The probability of the data, given all possible parameter values. Can be thought of as a normalising constant

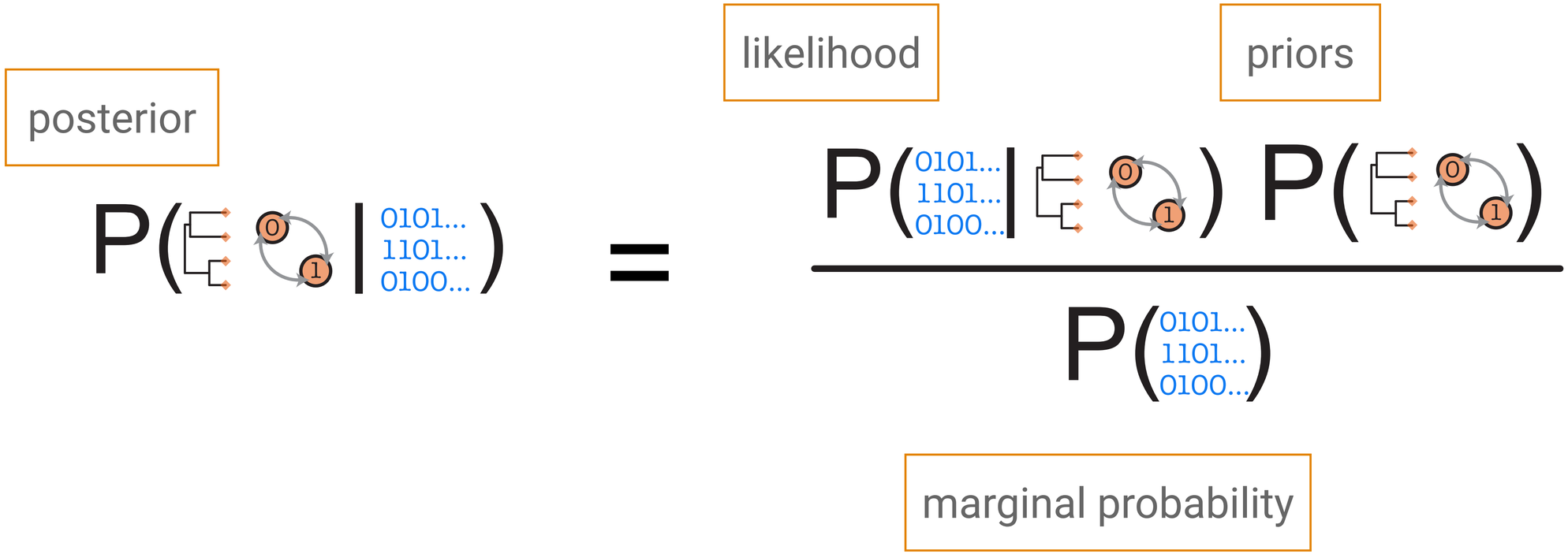
Bayes' theorem

Reflects our combined knowledge based on the likelihood and the priors

posterior

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model})\,\Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Bayesian tree inference



Recap

How do we find the ‘best’ tree?

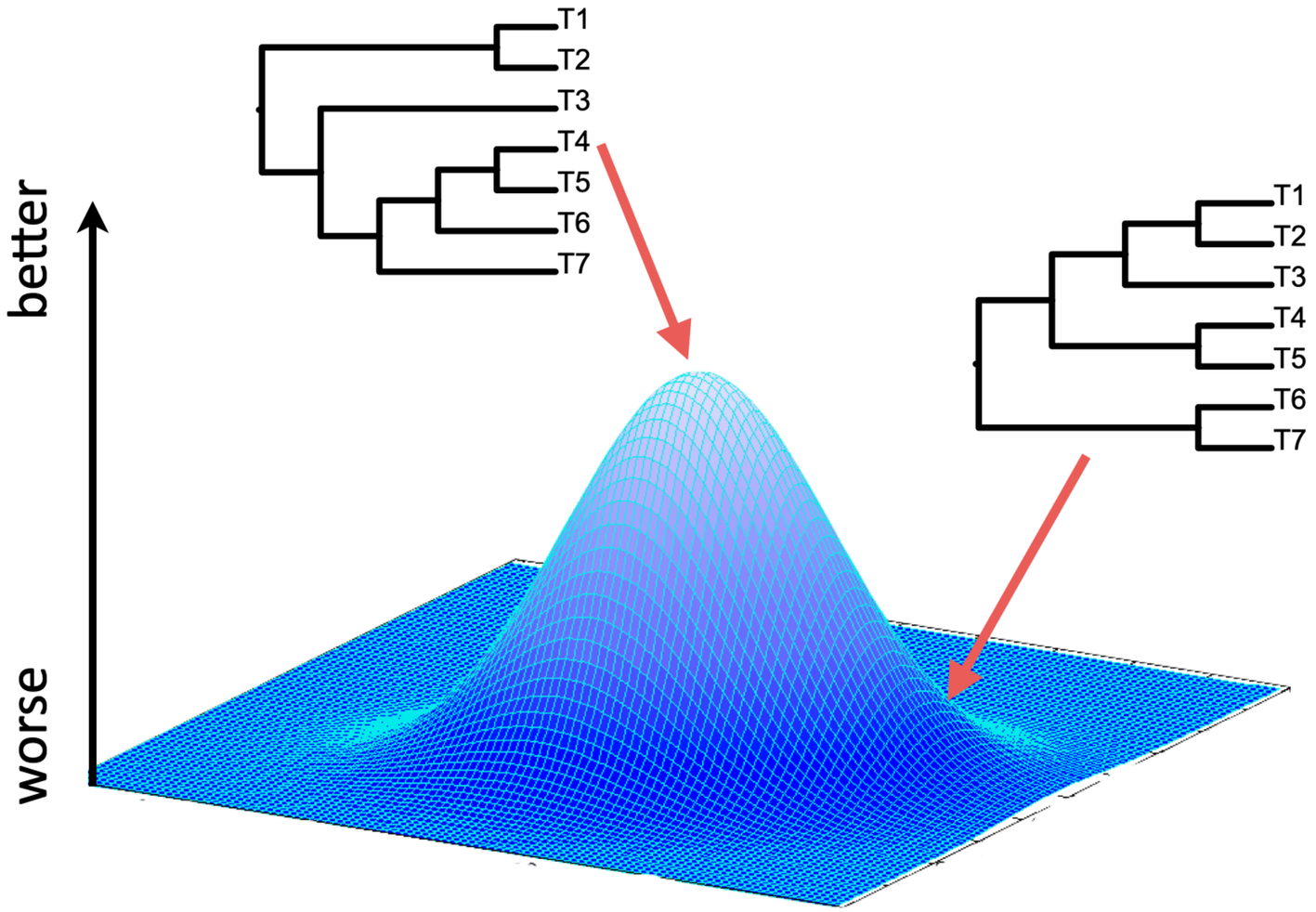


Image source Tracy Heath

It depends how you measure ‘best’

Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
Maximum likelihood	Likelihood score (probability), optimised over branch lengths and model parameters
Bayesian inference	Posterior probability, integrating over branch lengths and model parameters

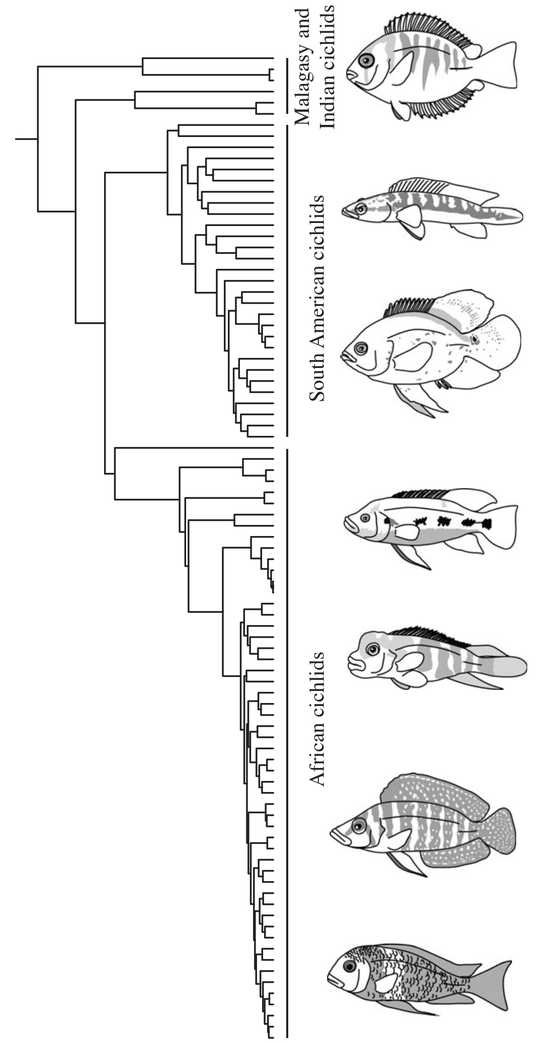
Both maximum likelihood and Bayesian inference are model-based approaches

Note these are not the only approaches to tree-building but they are the most widely used

Introduction to molecular dating

What can we learn from trees?

- Evolutionary relationships



What can we learn from trees?

- Evolutionary relationships
- Timing of diversification events
- Geological context
- Rates of phenotypic evolution
- Diversification rates

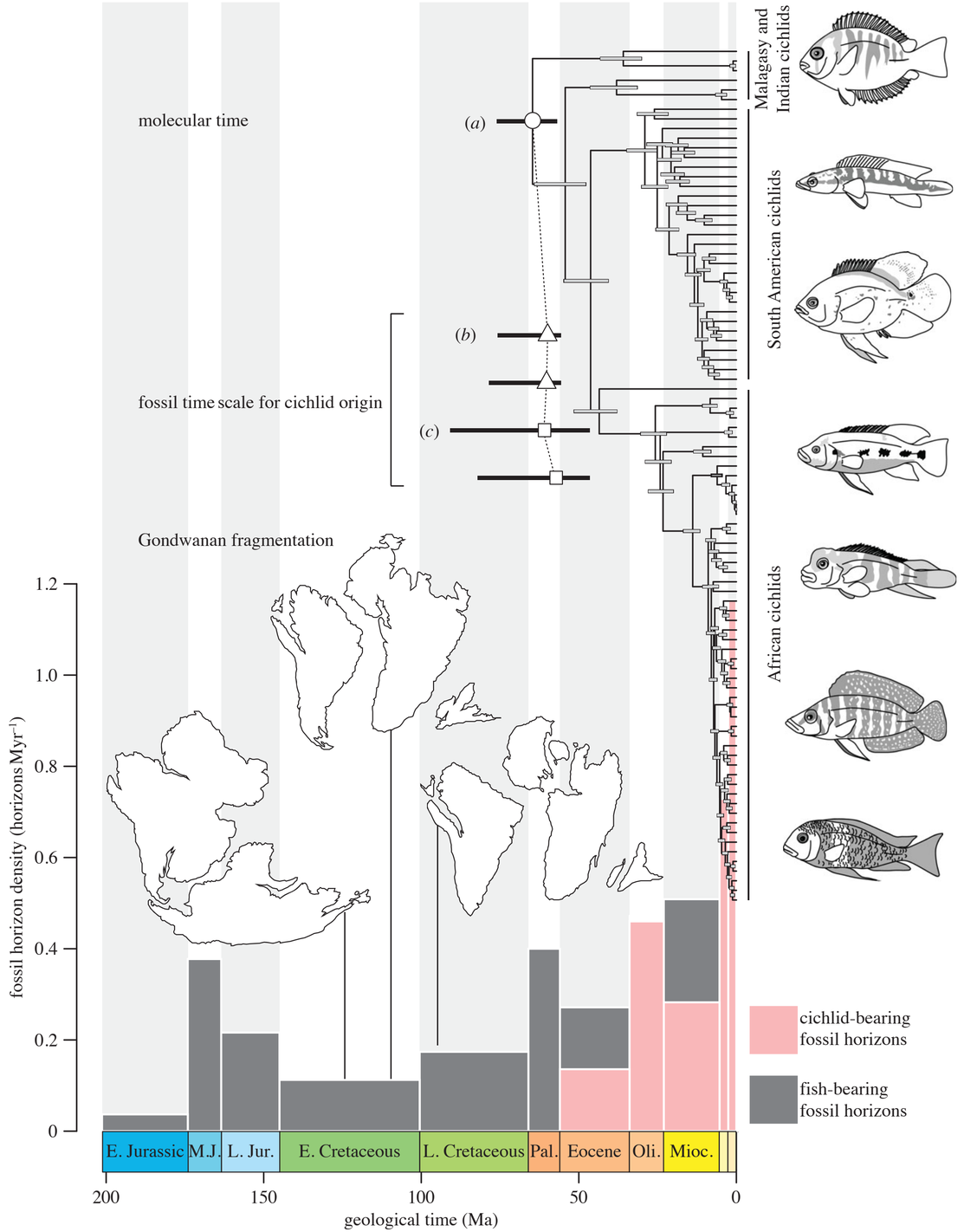
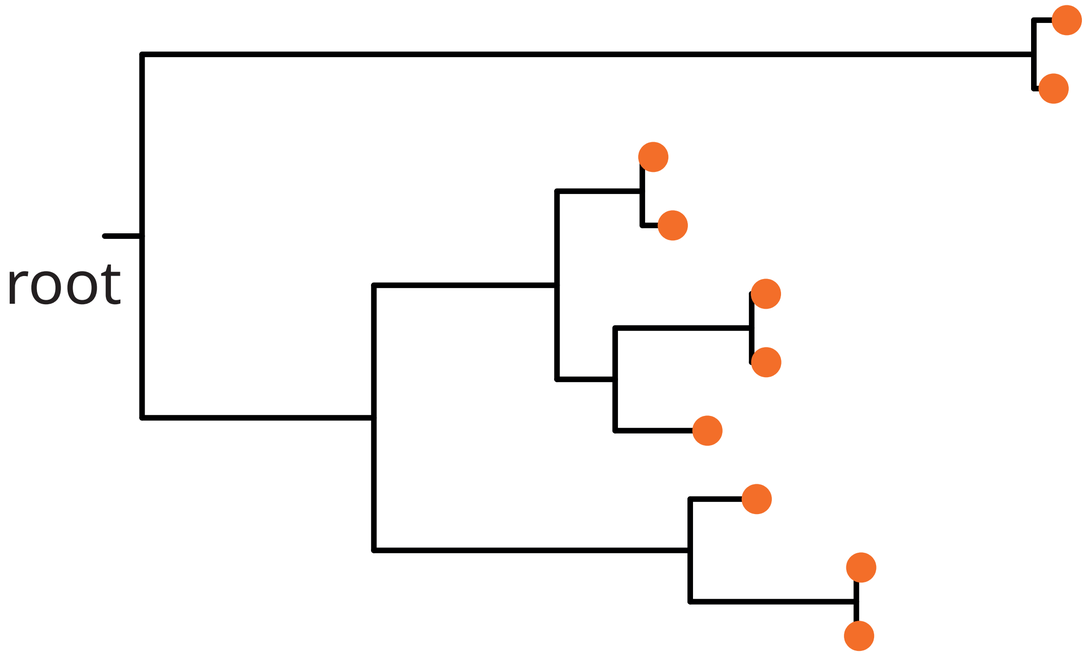


Image adapted from *Friedmann et al.* (2013)

Molecular (or morphological) characters are not independently informative about time



ATGCATGC
TTGCCTGC
TTGCATCG
ATGCATCG
ATGCATGG
TTGCCTGG
TAGCGTGC
TAGCGAGC





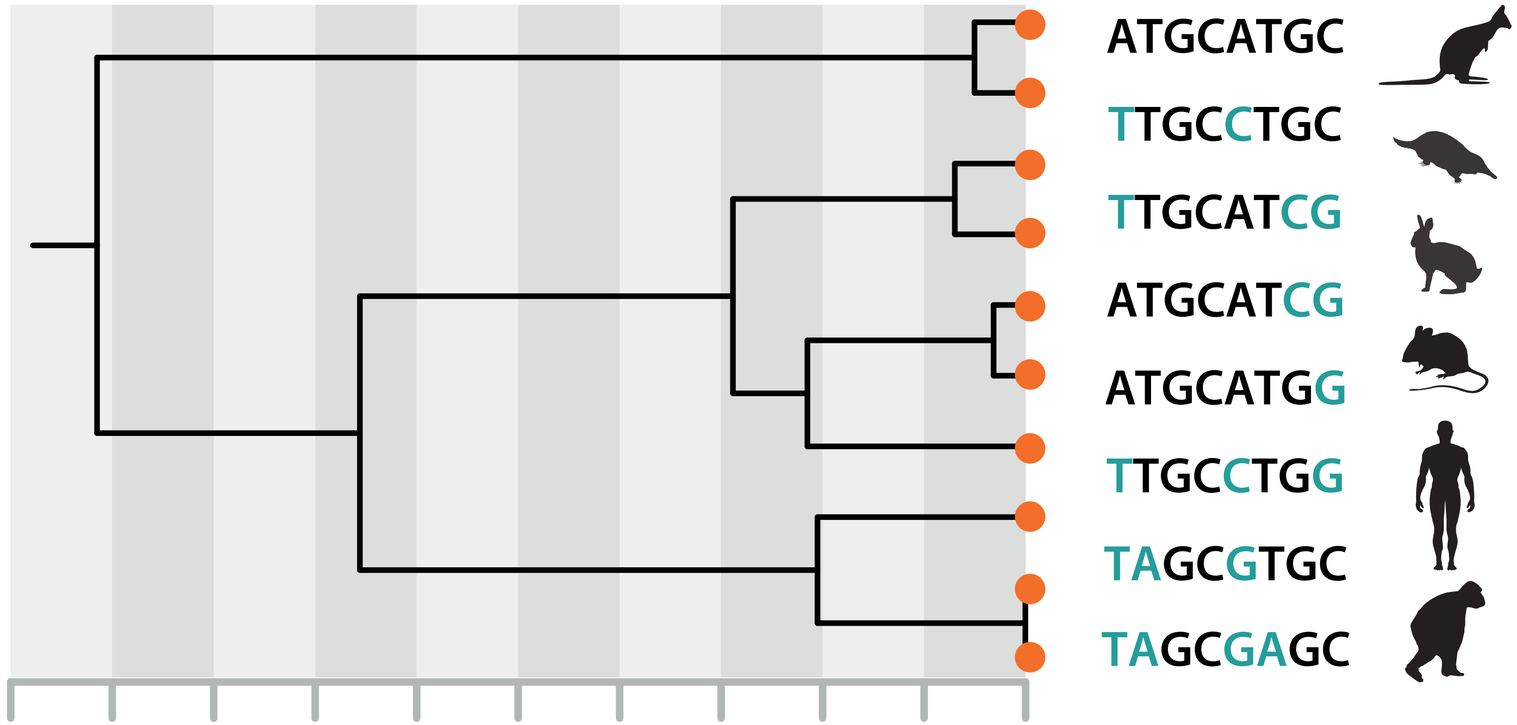

Slow rate, long interval **or** fast rate, short interval?

Goal: to disentangle evolutionary rate and time

branch lengths = genetic distance

$v = rt$

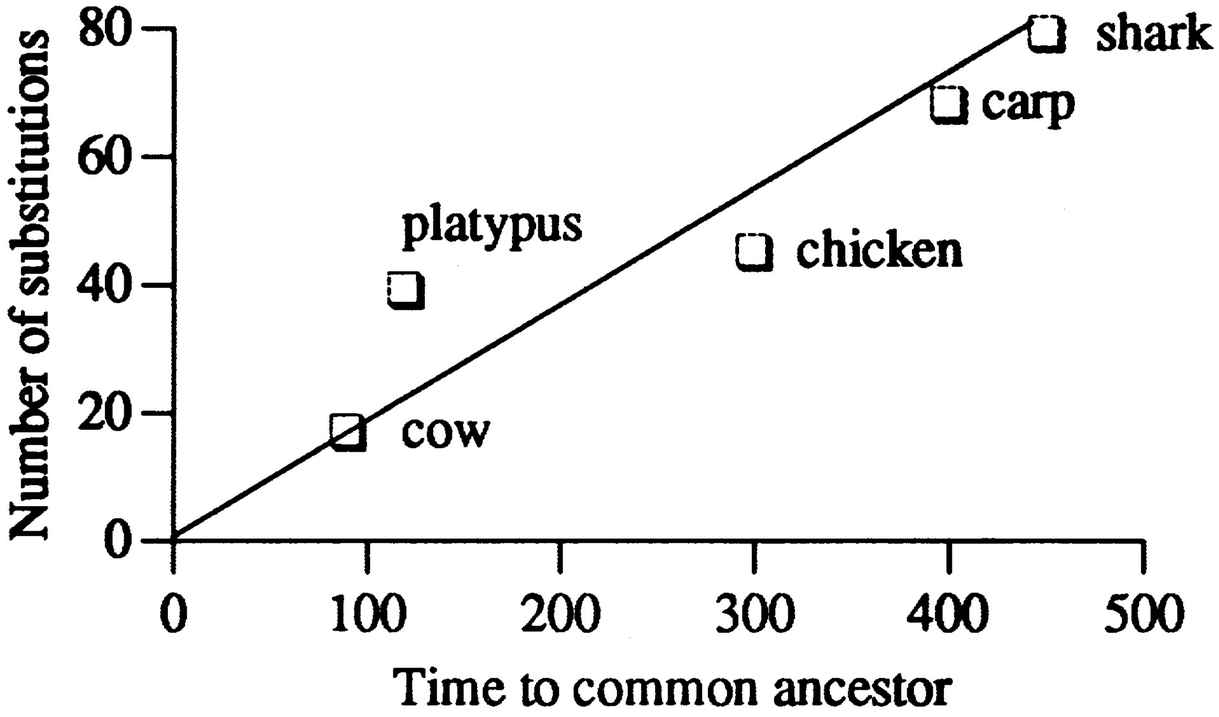
Molecular (or morphological) characters are not independently informative about time



branch lengths = time

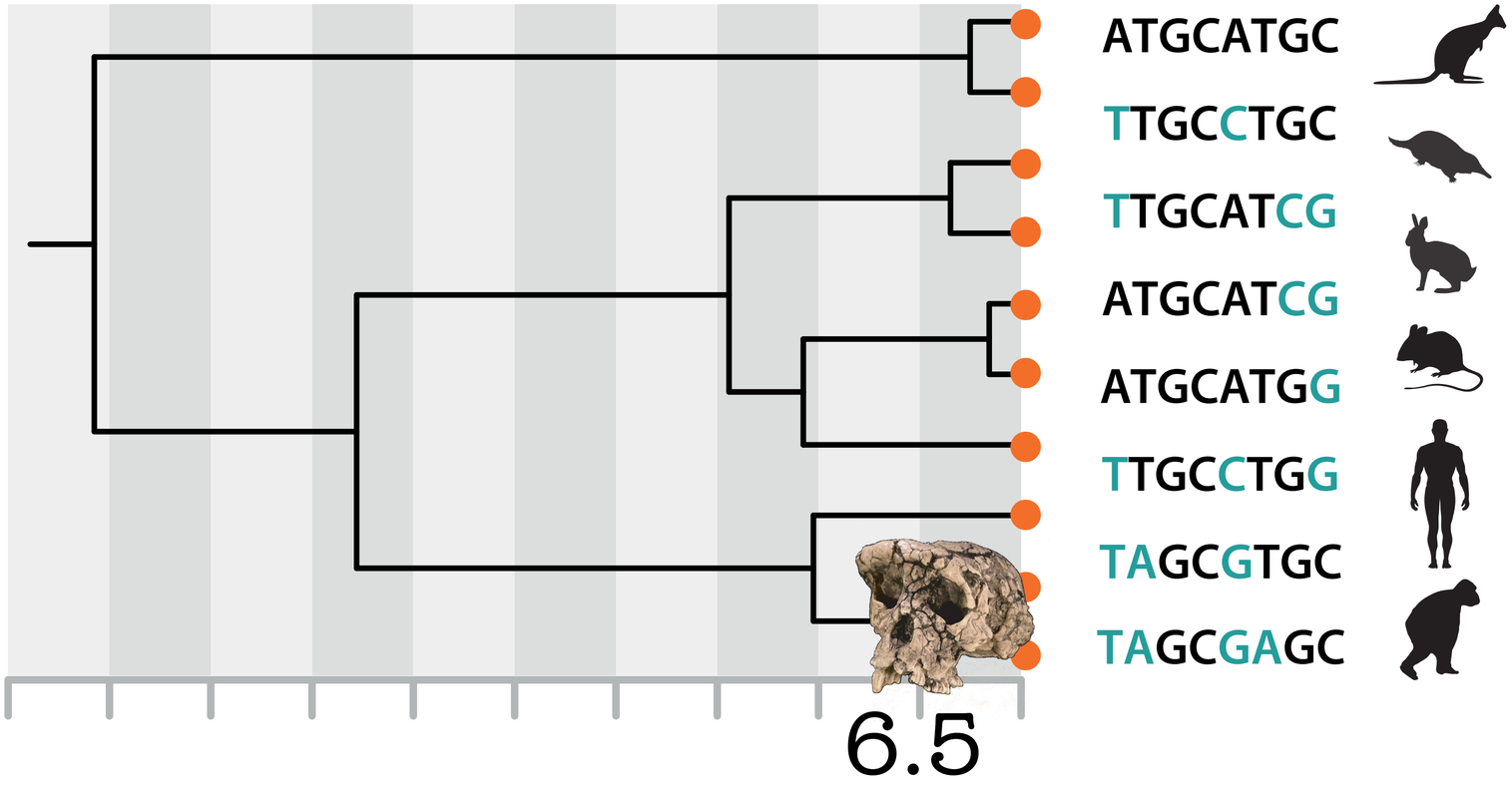
Goal: to disentangle evolutionary rate and time

The molecular clock hypothesis



Molecules as documents of evolutionary history *Zuckerkandl & Pauling* (1965)
A history of the molecular clock *Morgan* (1998)

Calibrating the substitution rate



branch lengths = time

ATGCATGC

TTGCCTGC

TTGCATCG

ATGCATCG

ATGCATGG

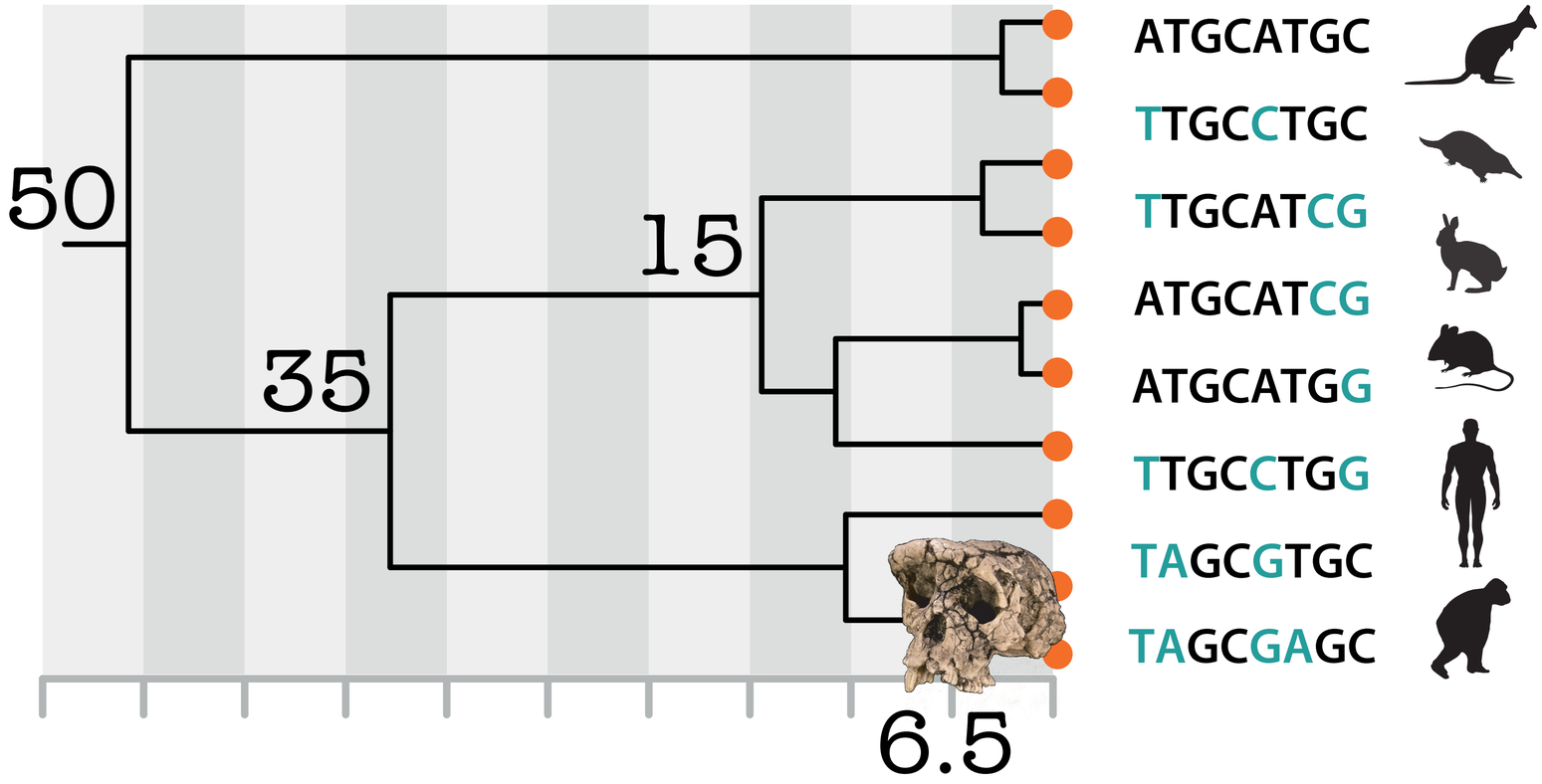
TTGCCTGG

TAGCGTGC

TAGCGAGC

Temporal evidence of divergence for one species pair let's us calibrate the average rate of molecular evolution

Calibrating the substitution rate

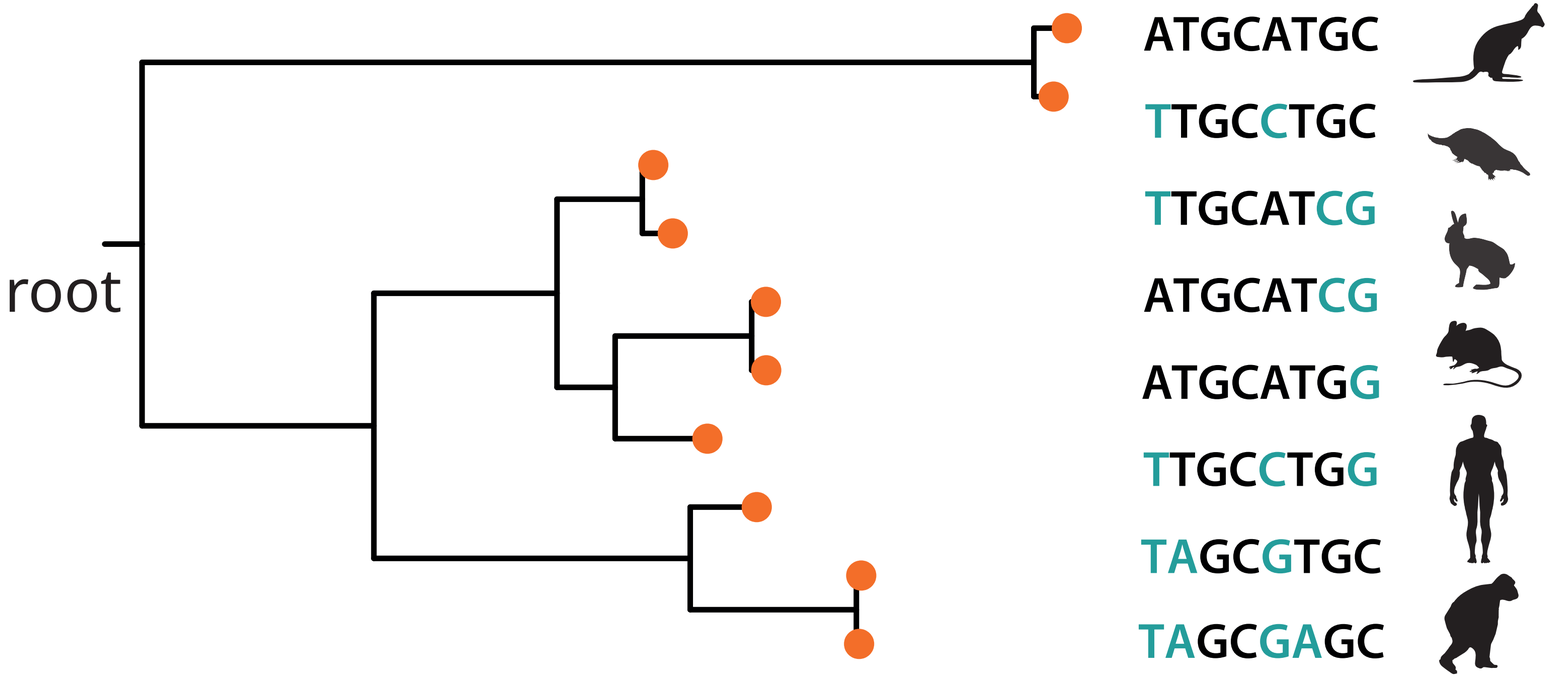


branch lengths = time

We can use this rate to extrapolate the divergence times for other species pairs

Molecular dating: challenges

Rate and time are not **fully identifiable**!

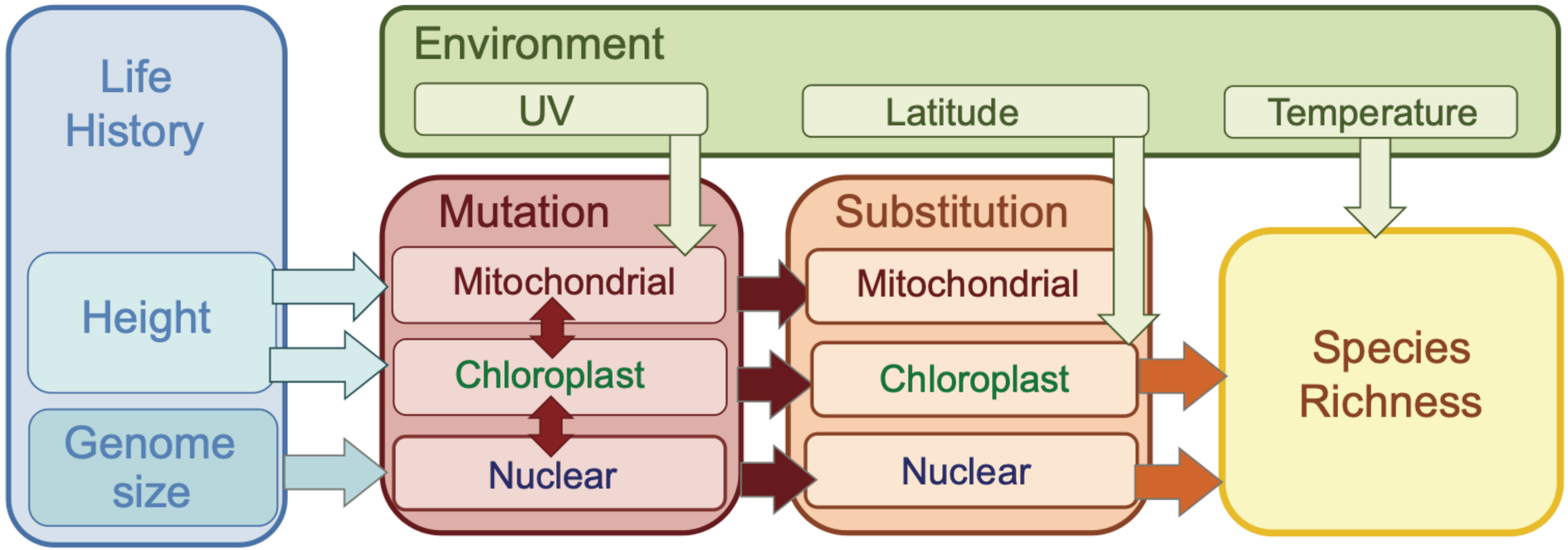


branch lengths = genetic distance

$v = rt$

Molecular dating: challenges

Many variables contribute to **variation in the substitution rate**



Bromham et al. (2015)

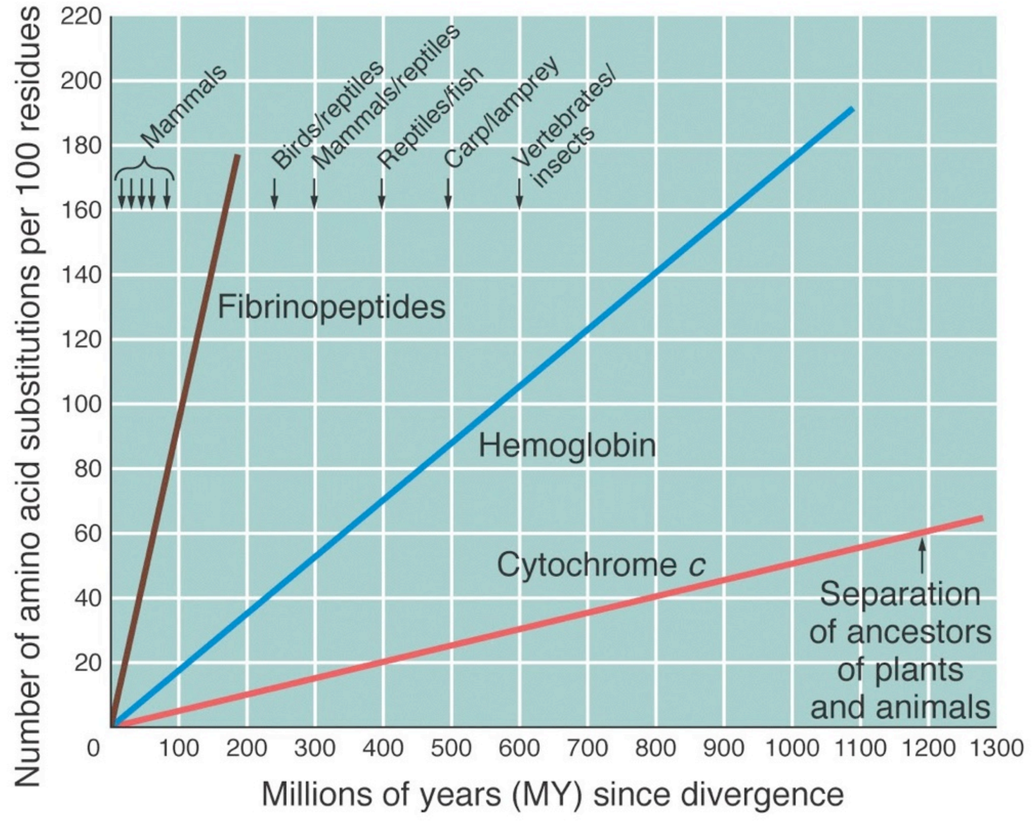
Molecular dating: challenges

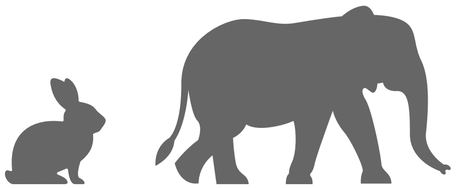
Many variables contribute to **variation in the substitution rate**

The molecular clock is not constant

Rates vary across:

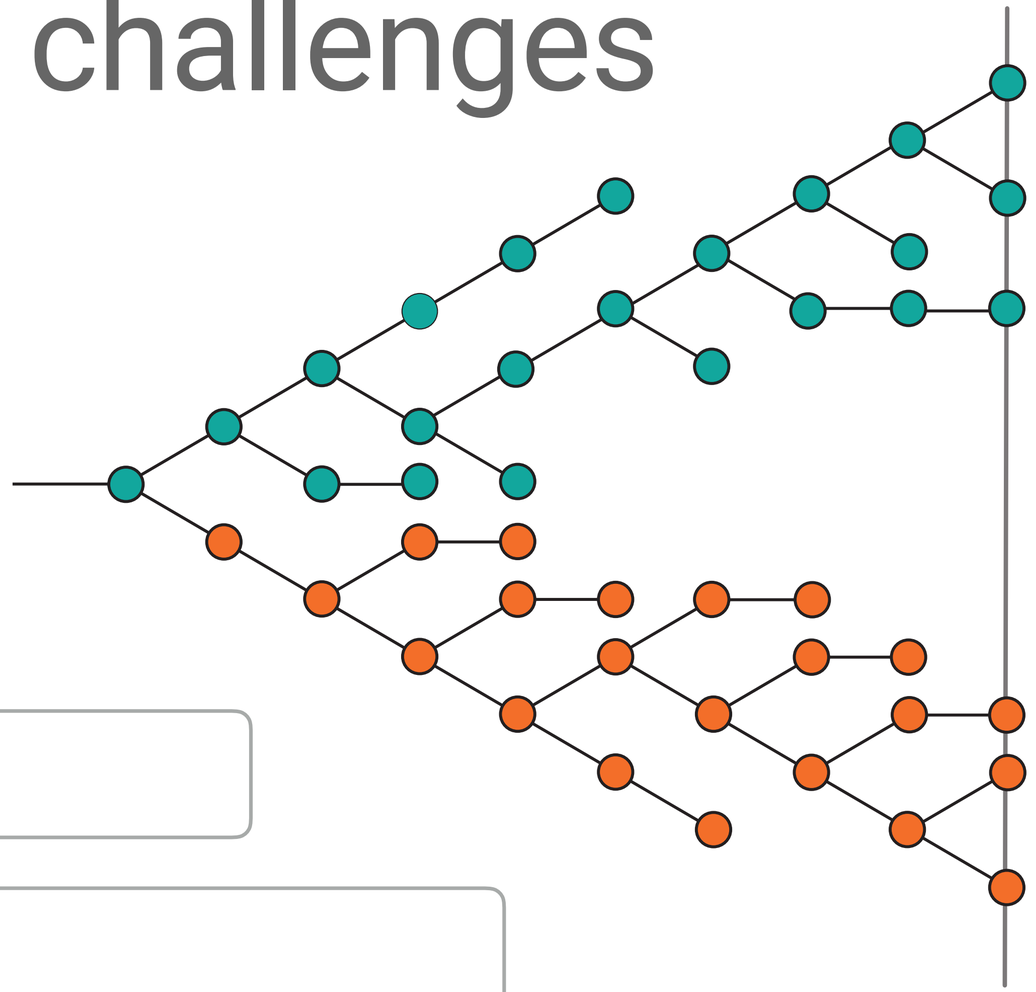
- taxa
- time
- genes
- sites within the same gene





Molecular dating: challenges

Calibrations are imprecise



Molecular evolution:

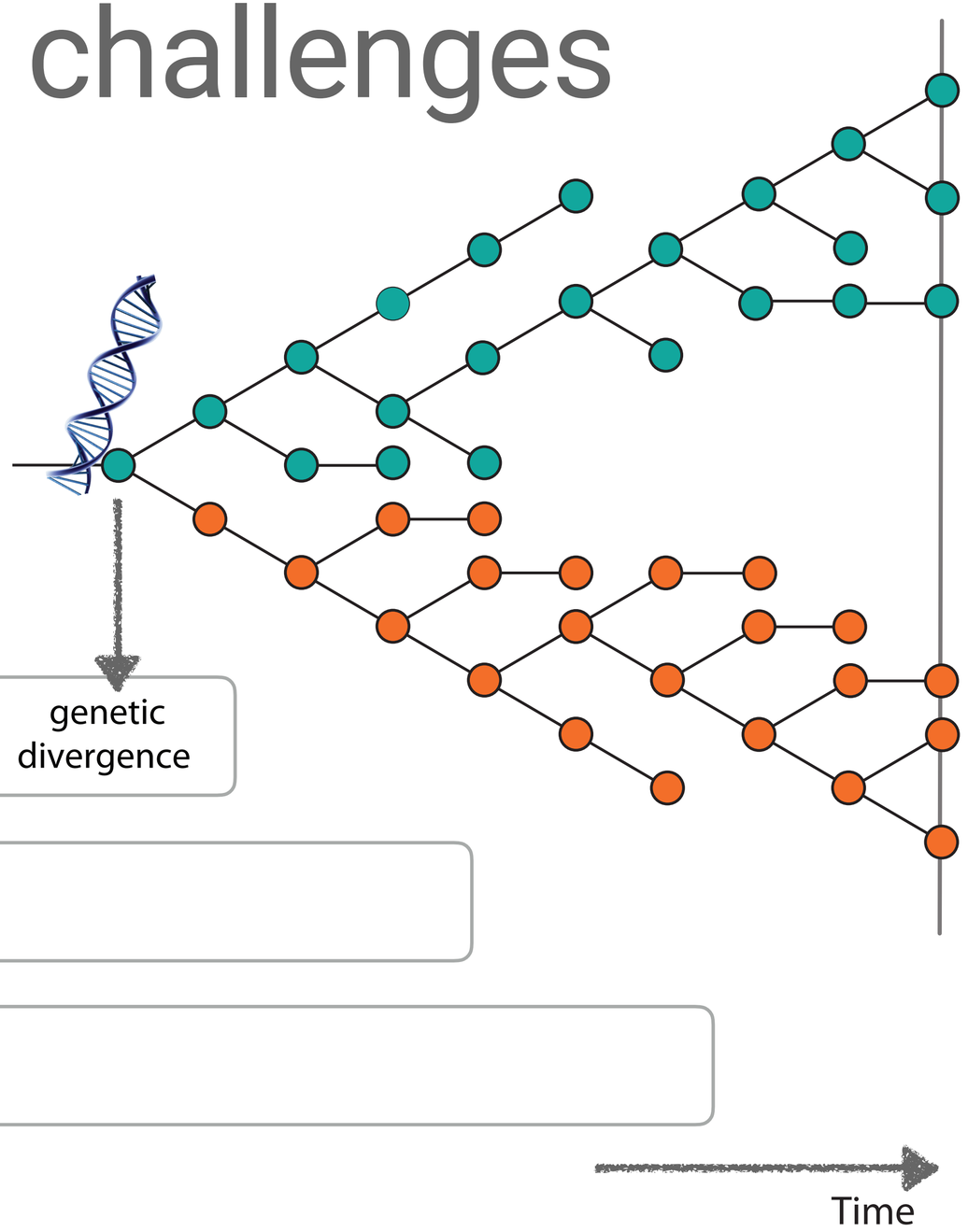
Morphological evolution:

Fossil preservation:

Time

Molecular dating: challenges

Calibrations are imprecise



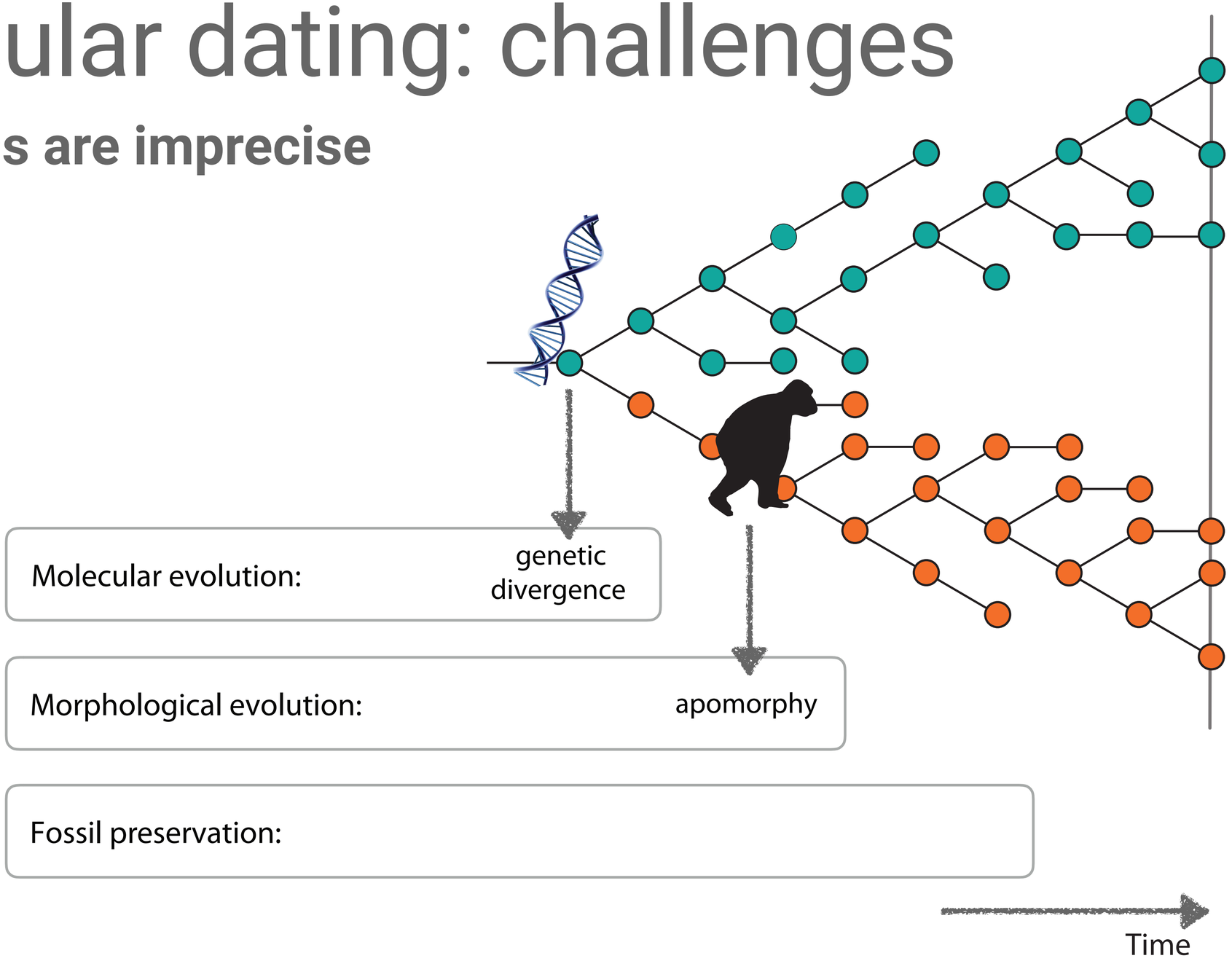
Molecular evolution:

Morphological evolution:

Fossil preservation:

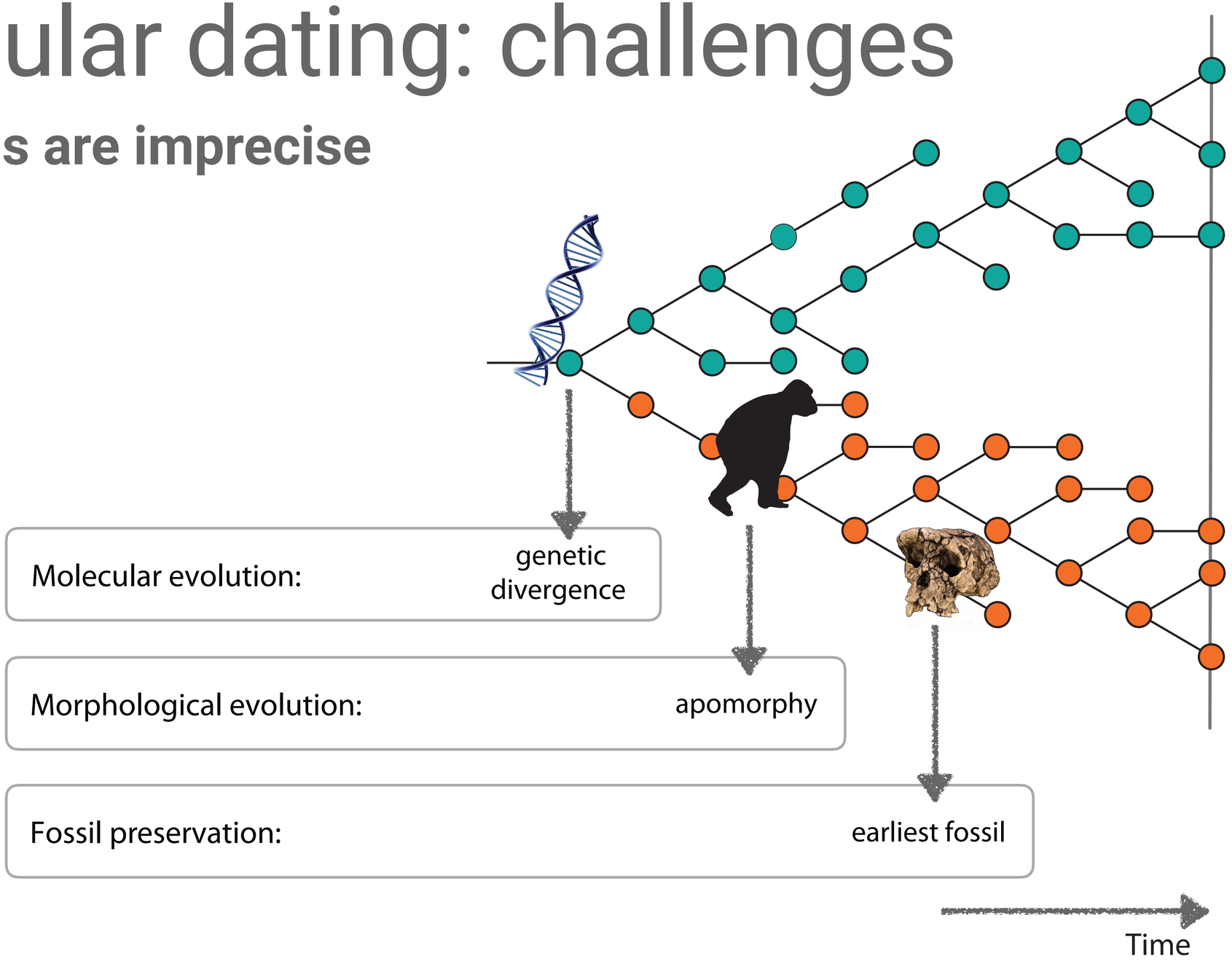
Molecular dating: challenges

Calibrations are imprecise



Molecular dating: challenges

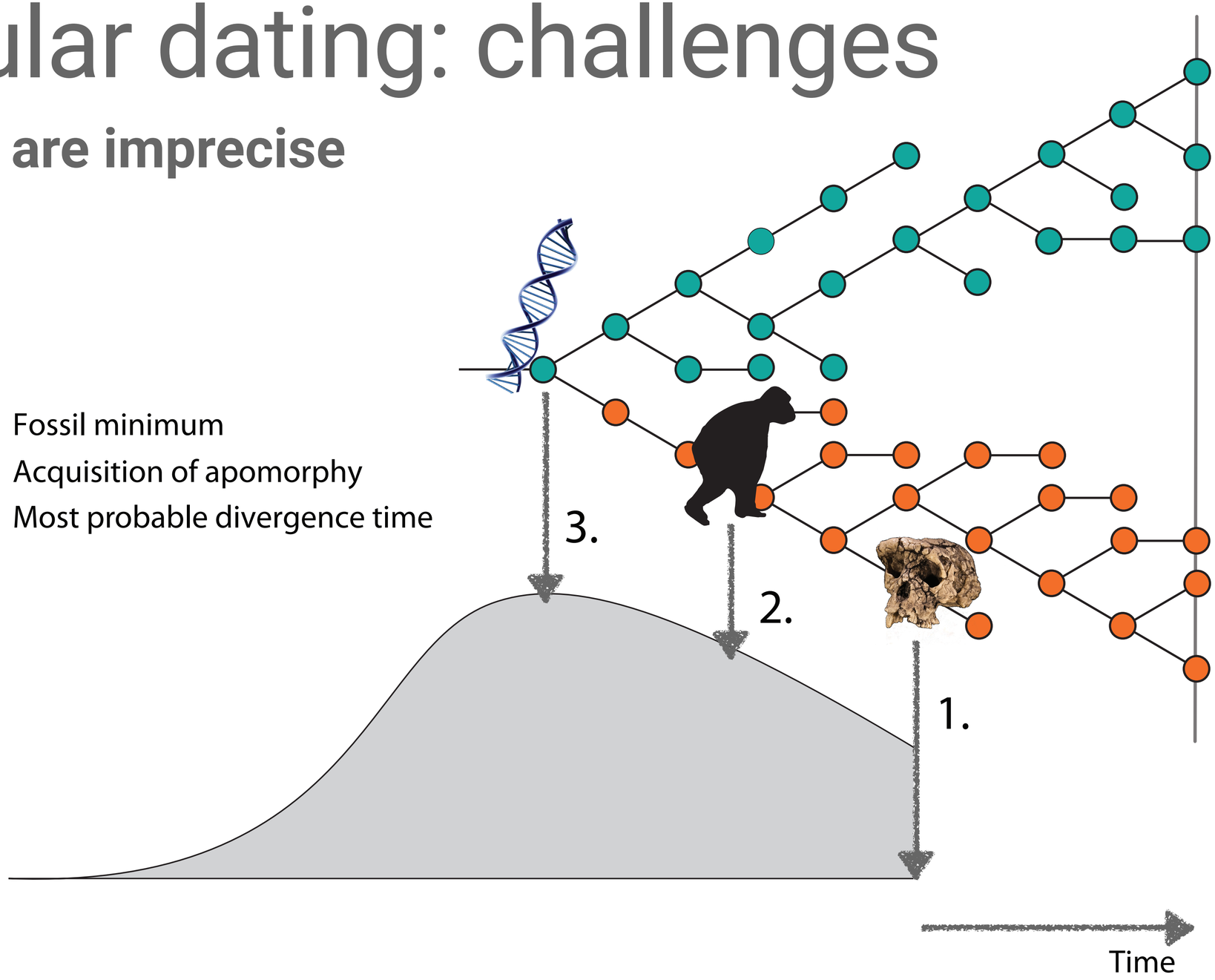
Calibrations are imprecise



Molecular dating: challenges

Calibrations are imprecise

1. Fossil minimum
2. Acquisition of apomorphy
3. Most probable divergence time



Molecular dating: challenges

Summary

1. Rate and time are not fully identifiable

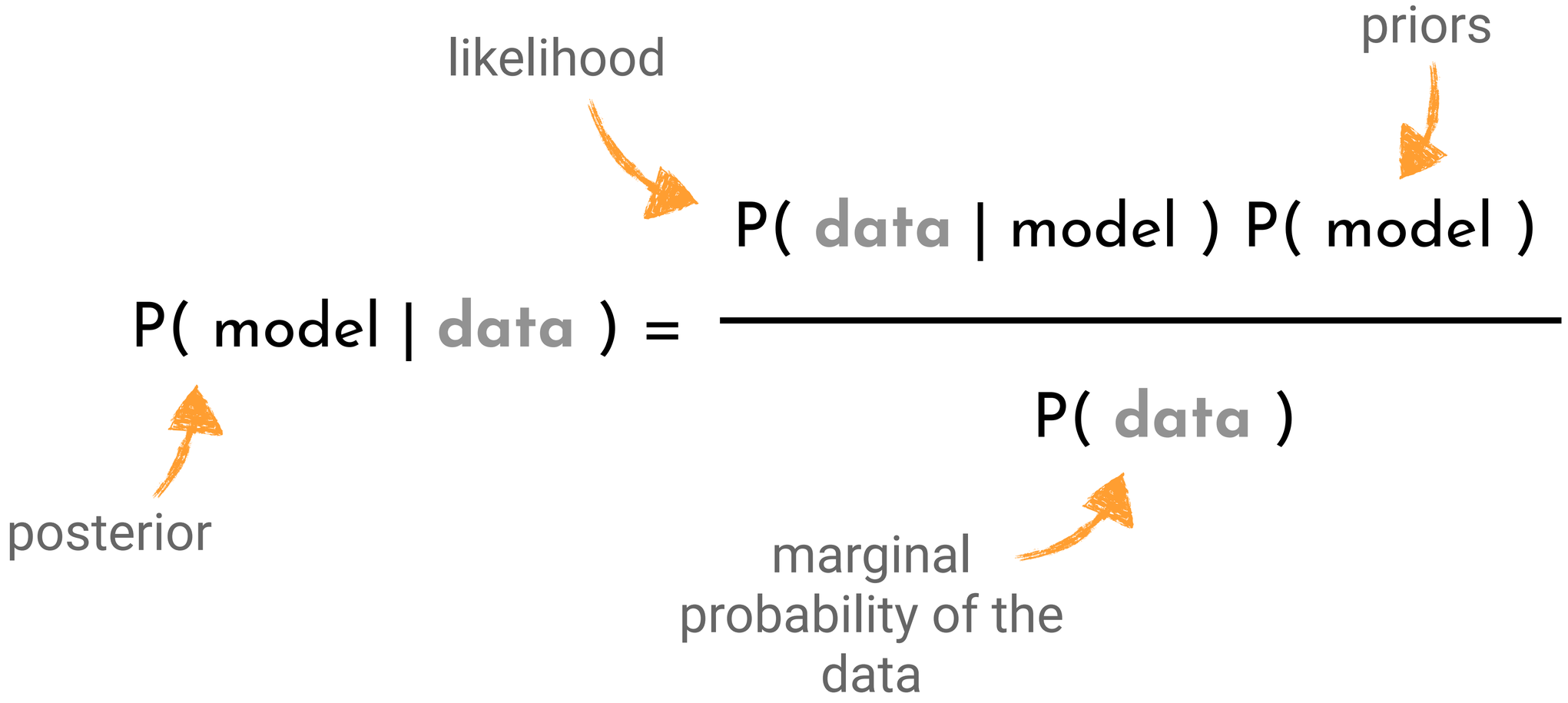
2. The substitution rate varies

3. Calibrations are imprecise

→ we need a flexible statistical framework that deals well with uncertainty!

Bayesian divergence time estimation

We use a Bayesian framework



Bayesian divergence time estimation

The data

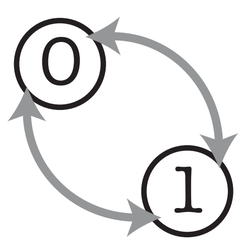
and / or

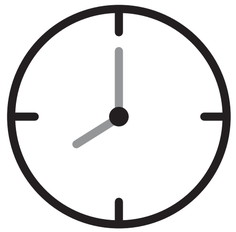
```
0101... ATTG...
1101... TTGC...
0100... ATTC...
```

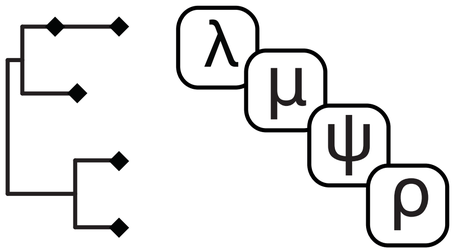
phylogenetics characters

sample ages

3 model components







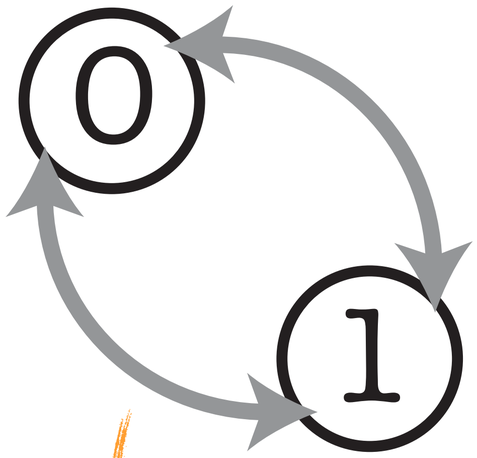
substitution model

clock model

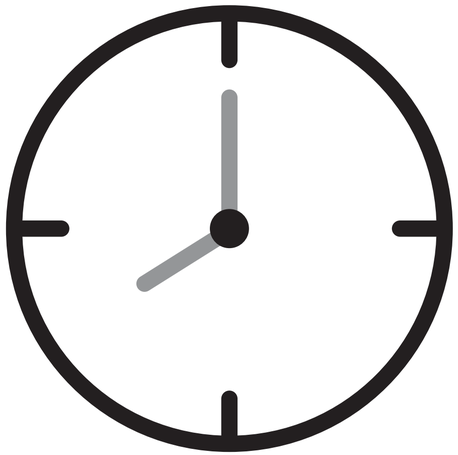
tree and tree model

Understanding the tripartite approach to Bayesian divergence time estimation
Warnock, Wright. (2020)

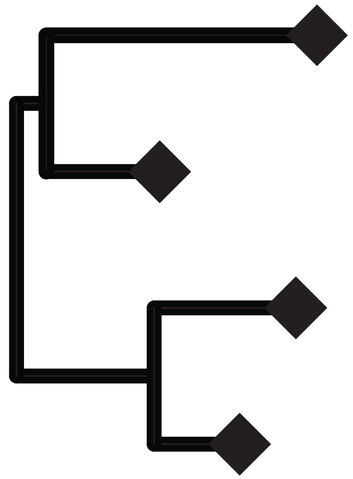
substitution model



clock model

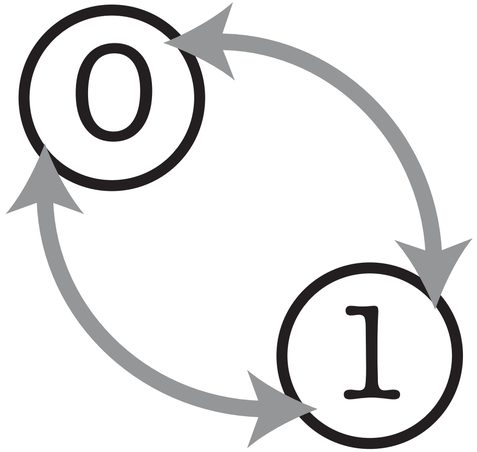


tree model

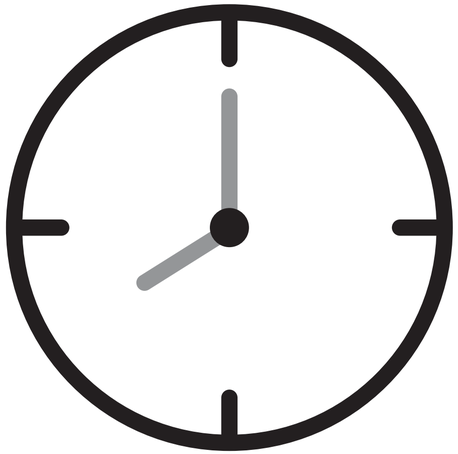


How likely are we to observe a change between character states? e.g., A → T

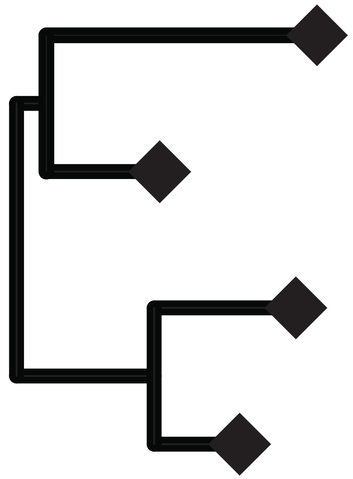
substitution model



clock model

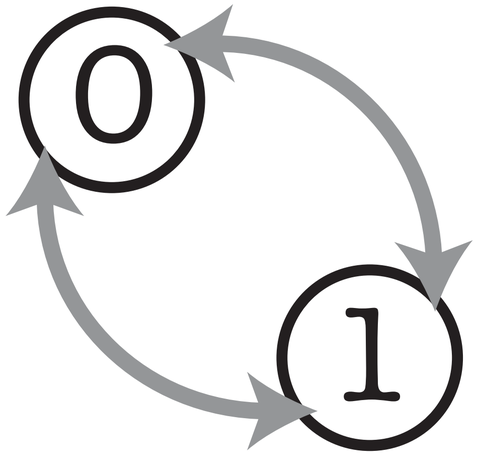


tree model

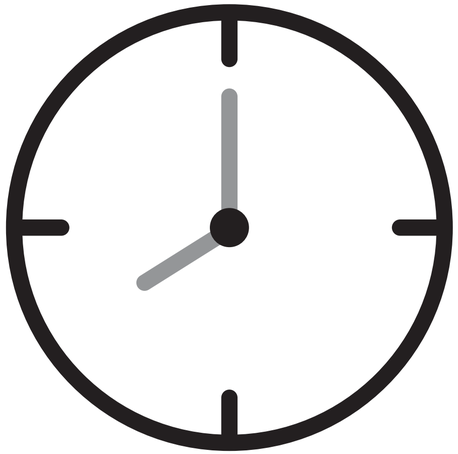


How have rates of evolution varied (or not) across the tree?

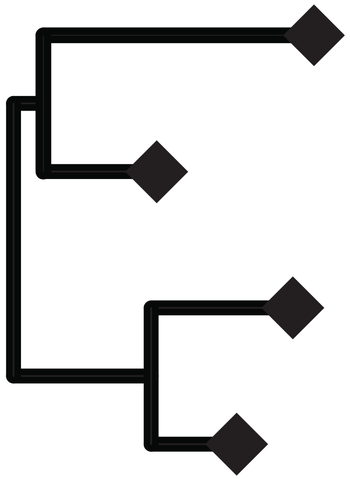
substitution model



clock model



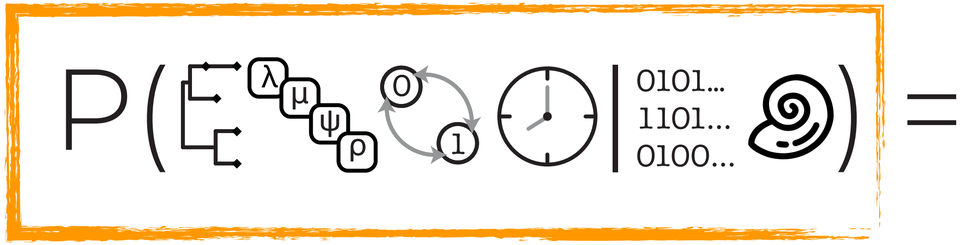
tree model



How have species originated, gone extinct and been sampled through time?

Bayesian divergence time estimation

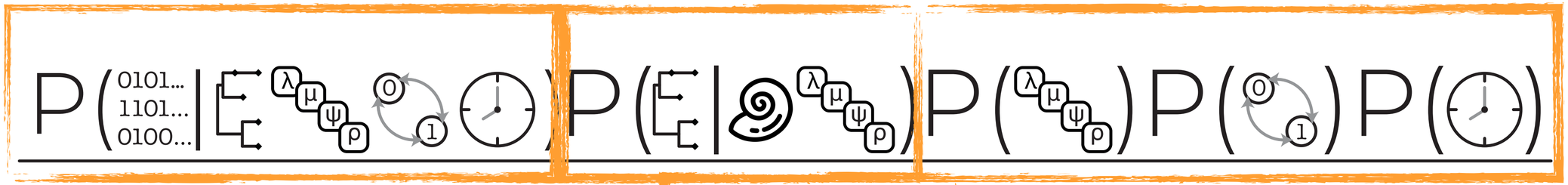
posterior

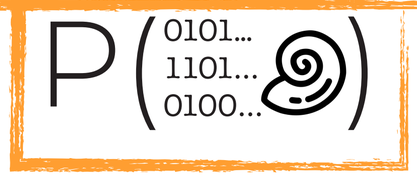


likelihood

probability of the time tree

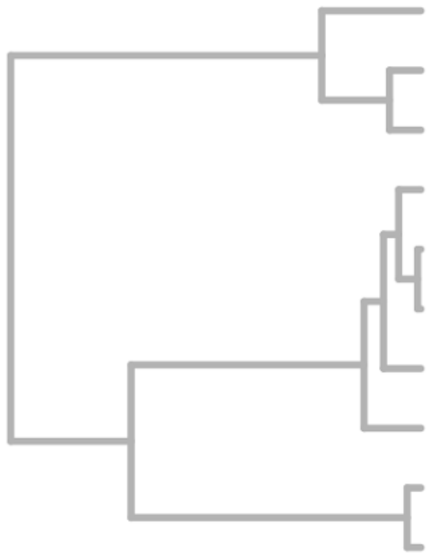
priors





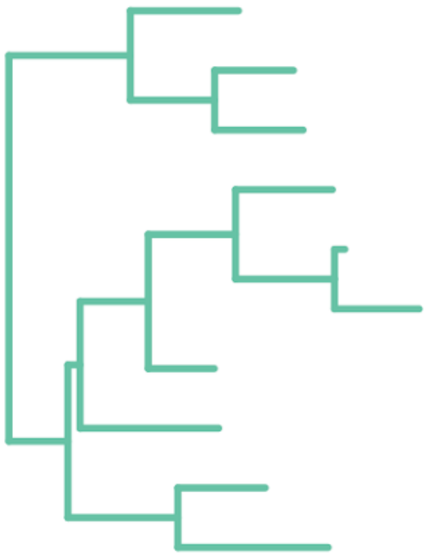
marginal pr of the data

Calculating the likelihood



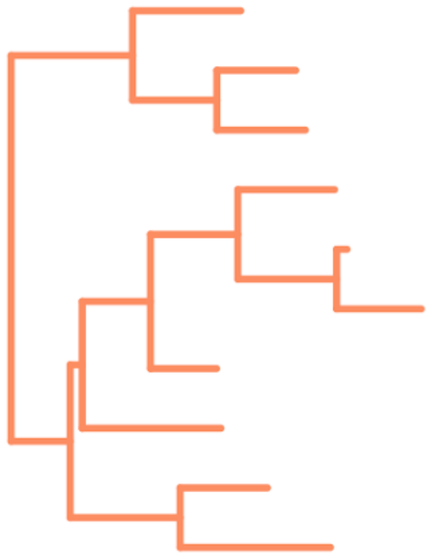
Time (year)

X



Clock rate (subs/site/year)

=



Genetic distance (subs/site)

Prior

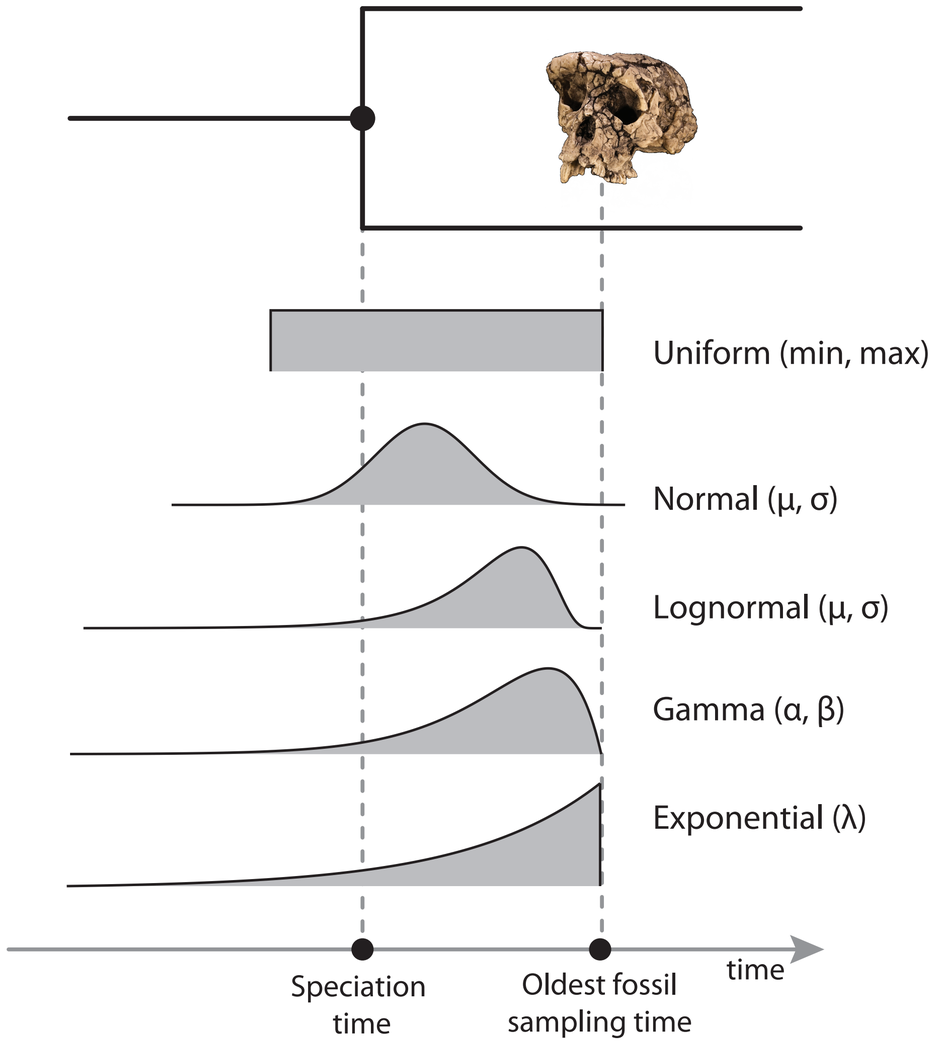
Prior

Likelihood

Based on the calibration times we can estimate the rate over time

Once we have the rate we can transform evolutionary rates in genetic distance

Node dating



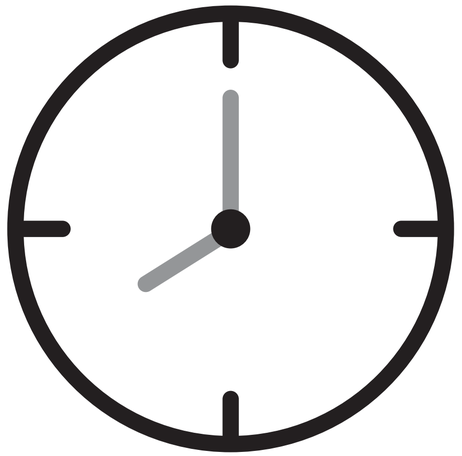
We can use a calibration density to constrain internal node ages

We typically use a birth-death process model to describe the tree generating process

Adapted from *Heath* 2012. Sys Bio

clock model

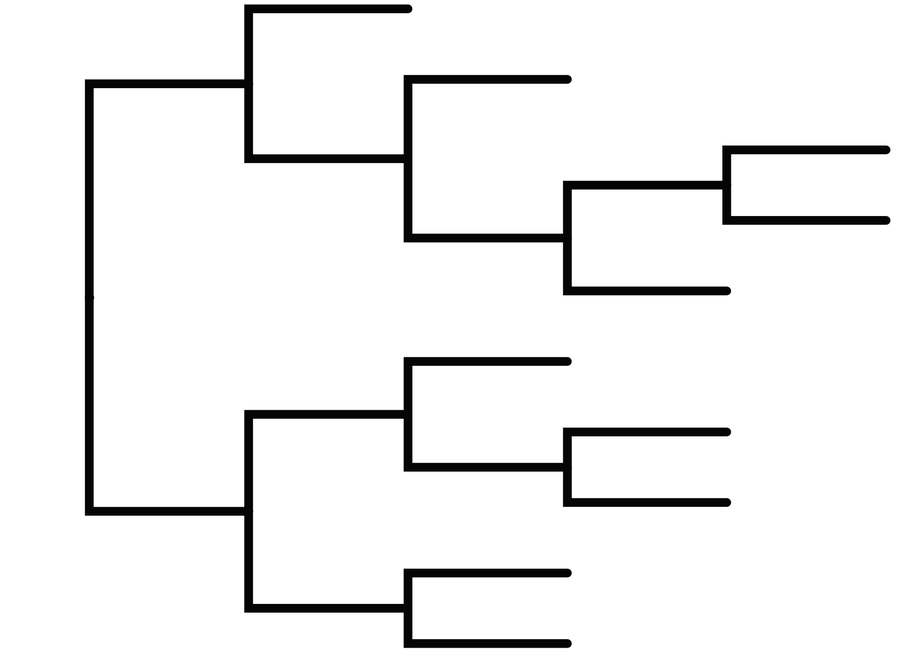
The clock model describes how evolutionary rates vary (or not) across the tree



The strict / constant molecular clock model

Assumptions

- The substitution rate is constant over time
- All lineages share the same rate



branch length = substitution rate

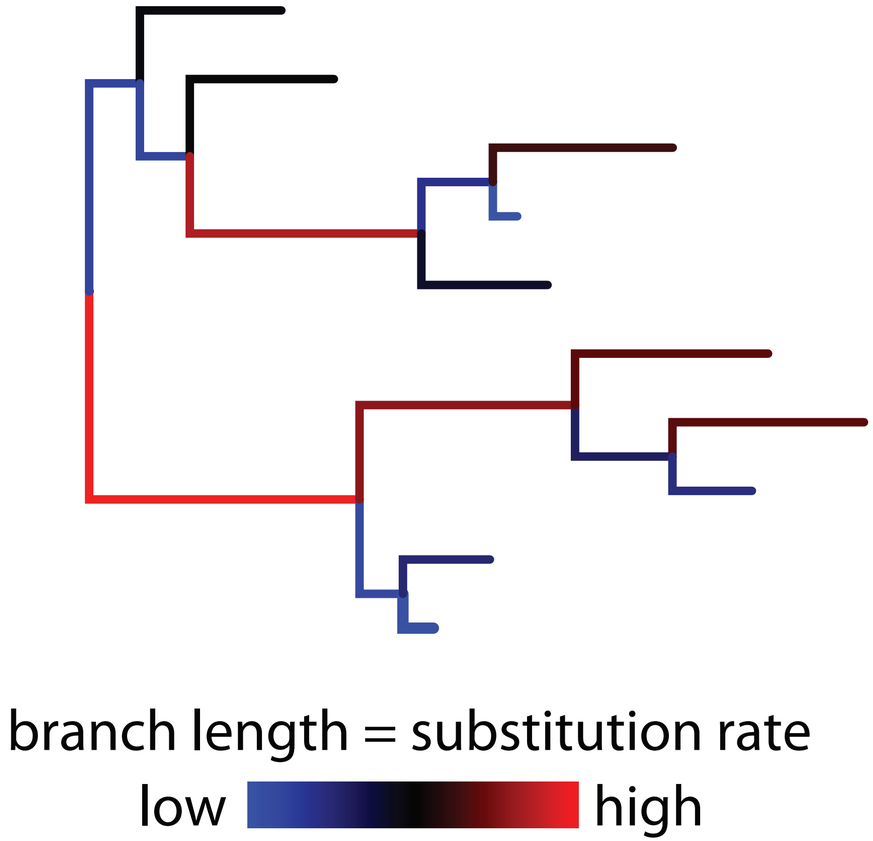
low high

Relaxed clock models

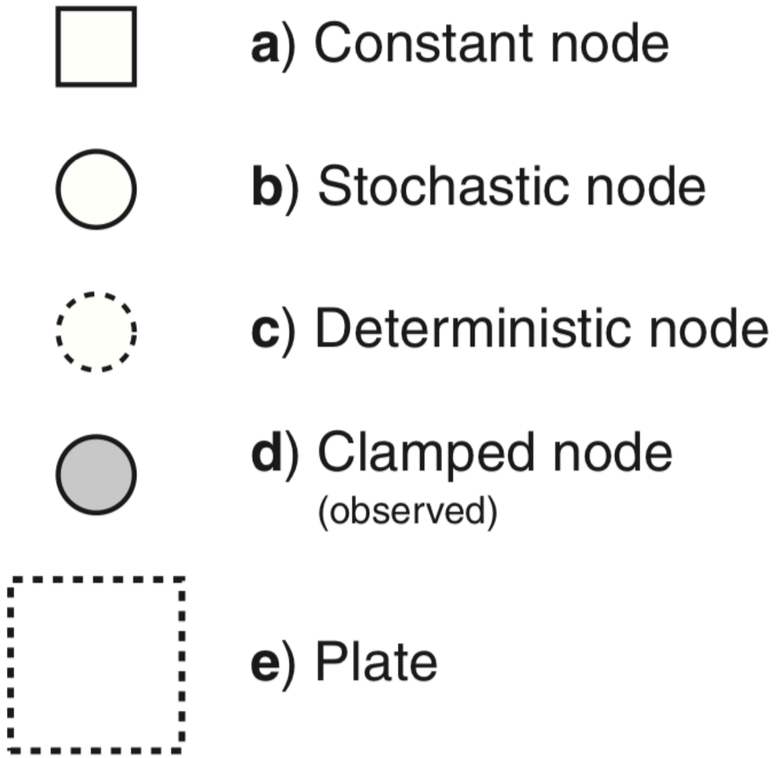
Assumptions

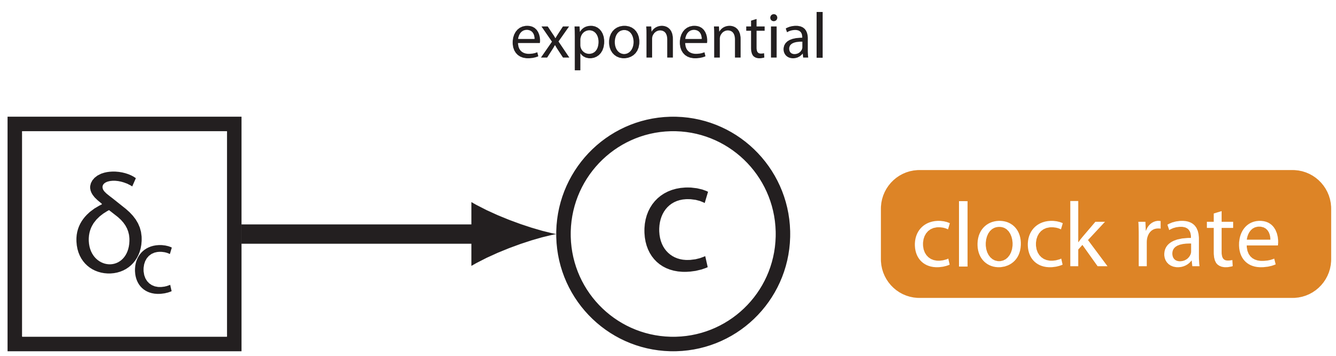
- Lineage-specific rates

- The rate assigned to each branch is drawn from some underlying distribution

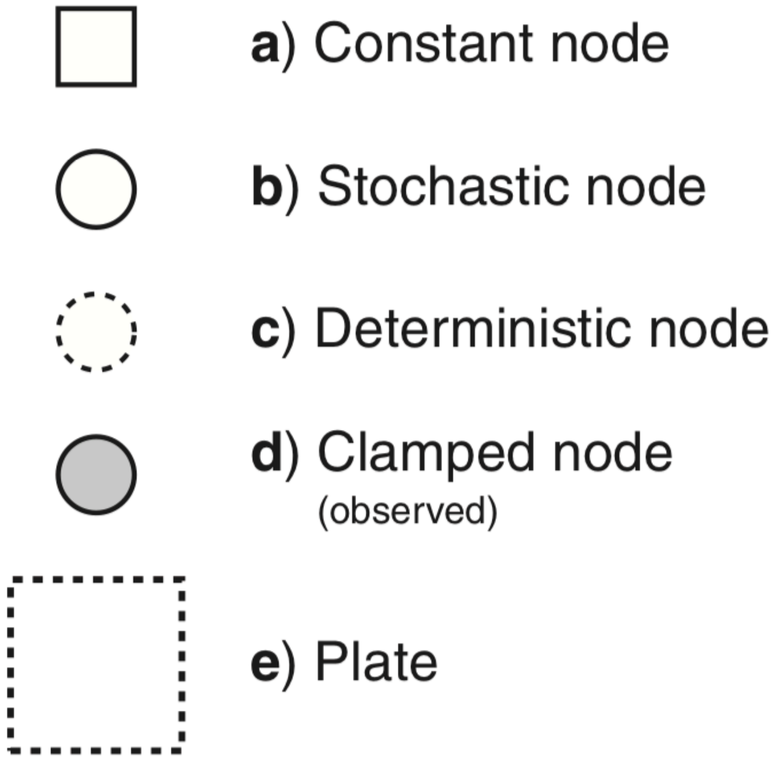


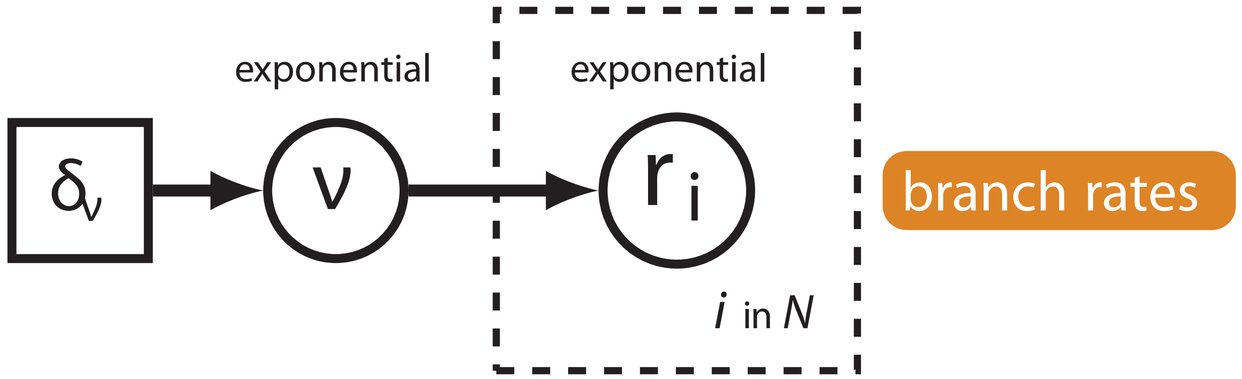
Graphical models: strict clock model





Graphical models: relaxed clock model





There are many different clock models

- Strict clock
- Uncorrelated or independent clock (= the favourite)
- Autocorrelated clock
- Local clocks
- Mixture models

The changing face of the molecular evolutionary clock. *Ho*. 2014
See also: *Warnock, Wright*. 2020. Elements of Paleontology

Exercise (demo only)

Applications of node dating

Hypothesis testing: did the break up of Gondwana drive the radiation of cichlids?

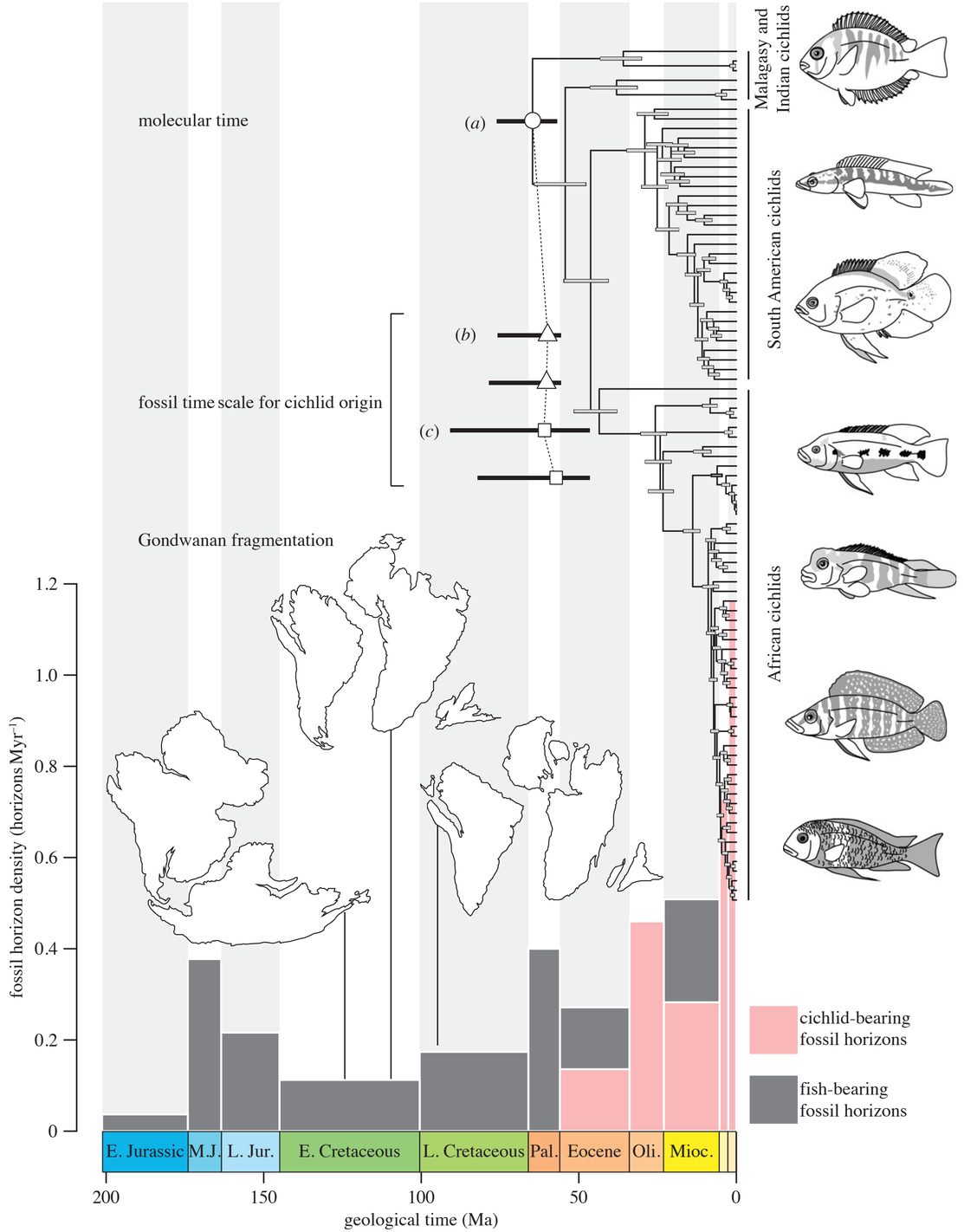
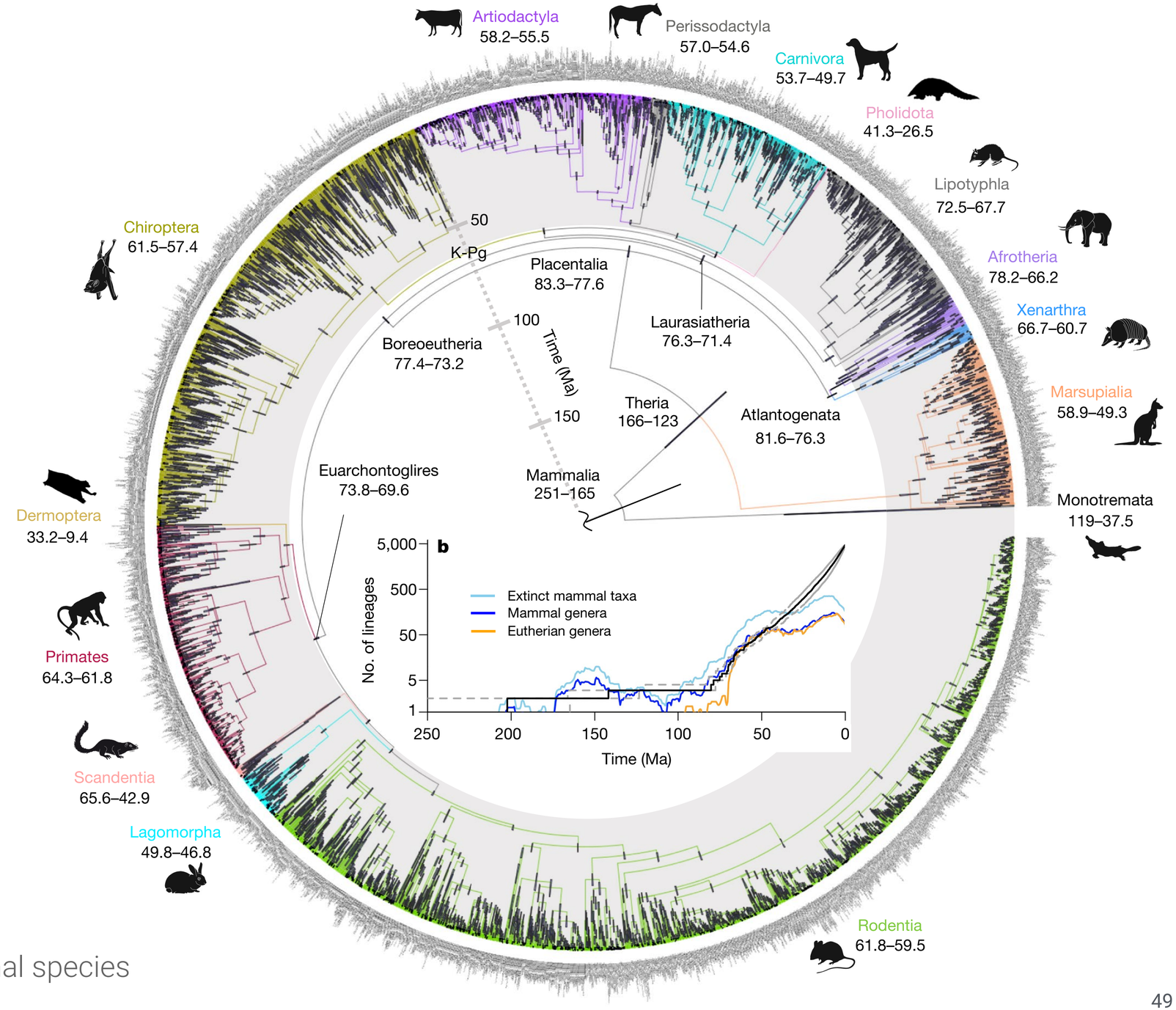


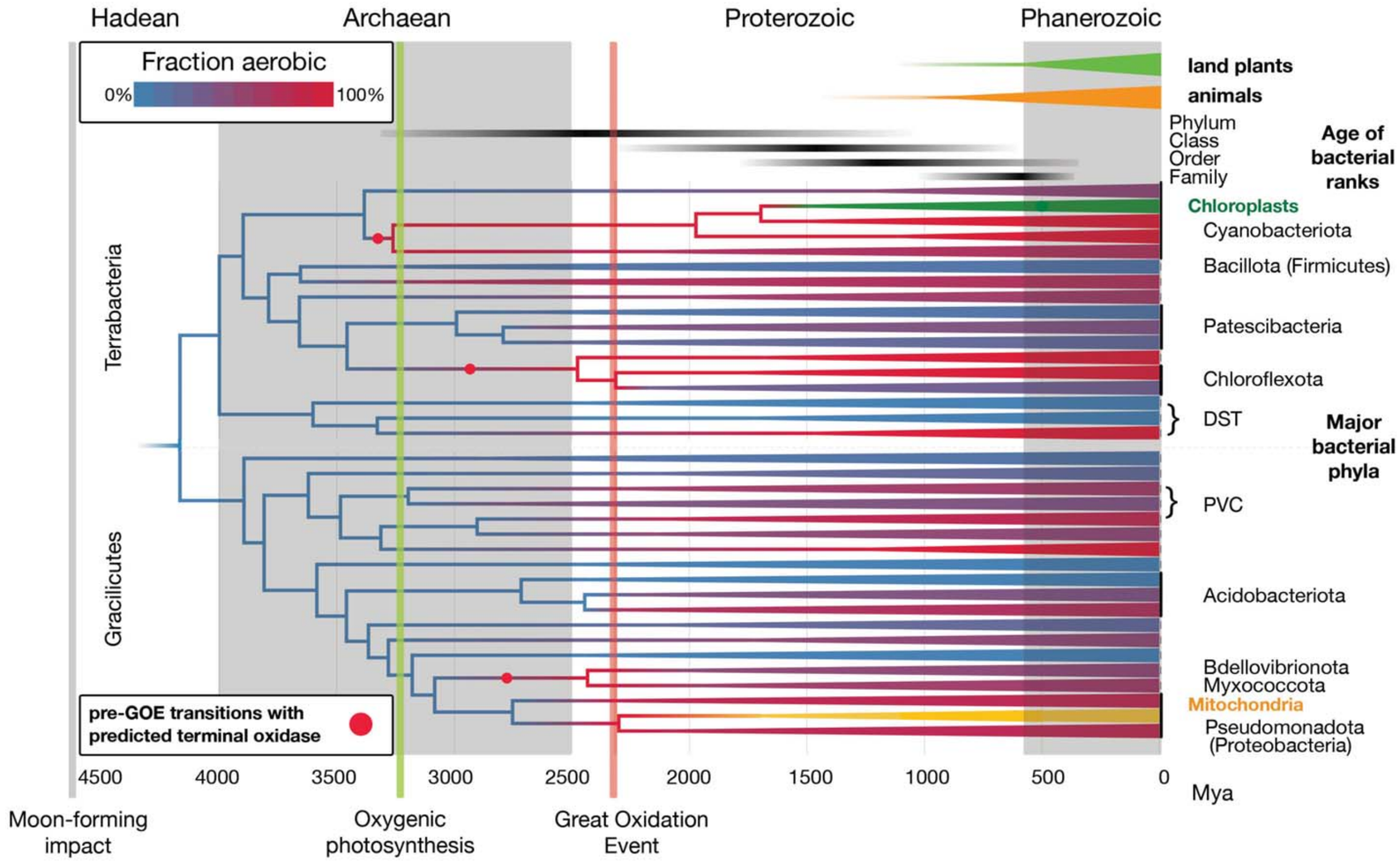
Image adapted from *Friedmann et al.* (2013)

Very large trees can only be time-calibrated using a node dating approach



Álvarez-Carretero et al. (2021) — 4,705 mammal species

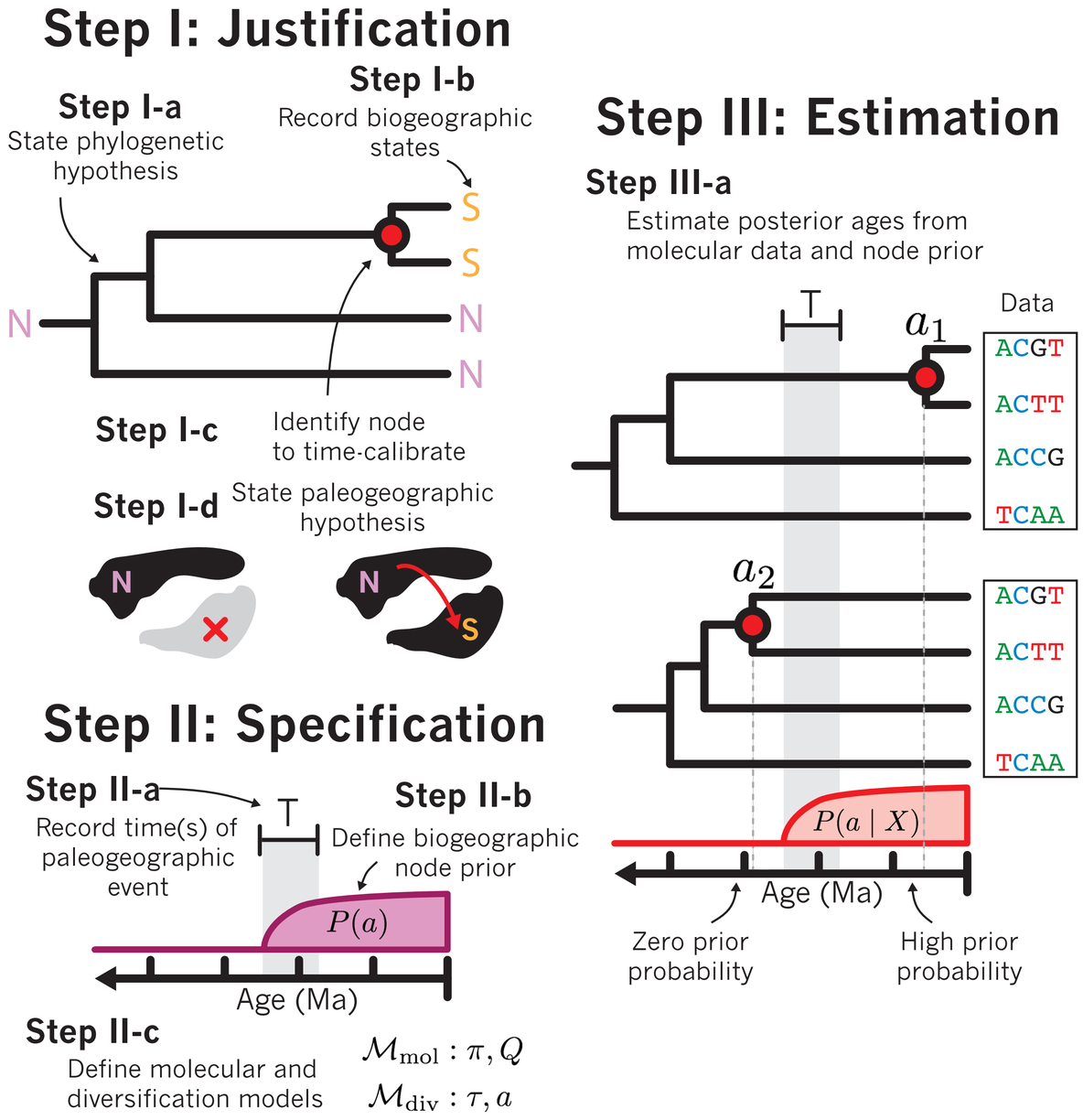
Dating with sparse calibration information



Davín et al. (2025)

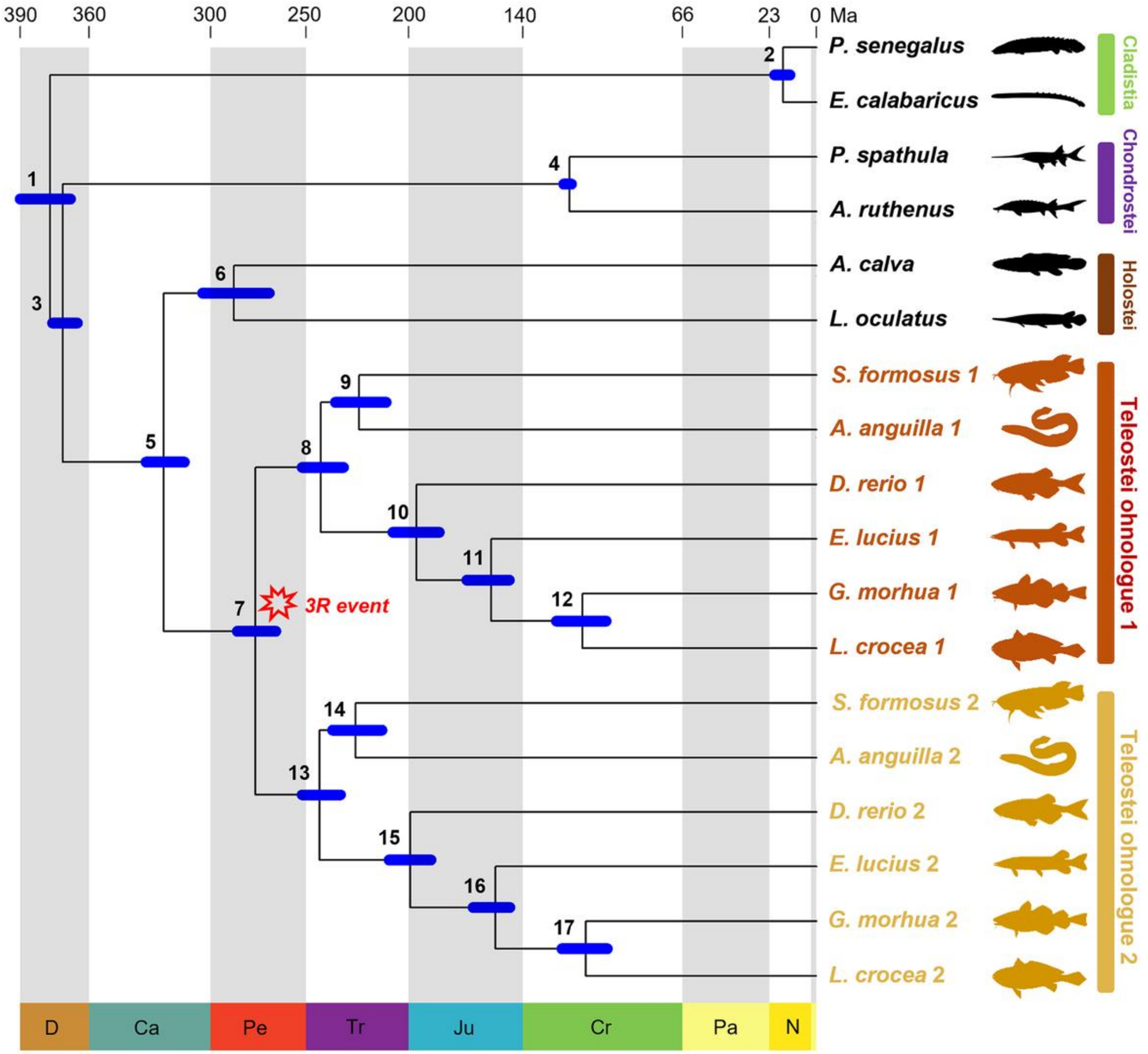
A geological timescale for bacterial evolution, calibrated using atmospheric oxygenation and the spread of aerobic metabolism

Biogeographic calibrations



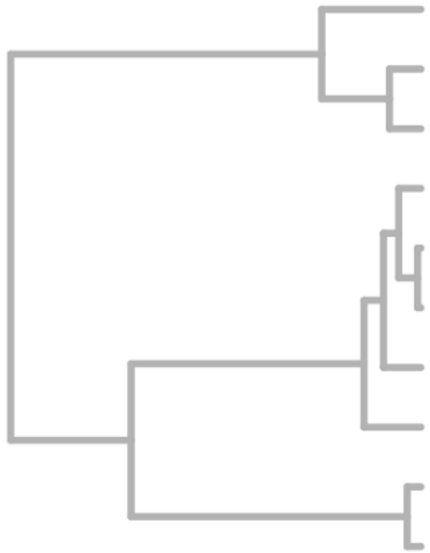
Landis (2017, 2021)

Dating gene or genome duplication events



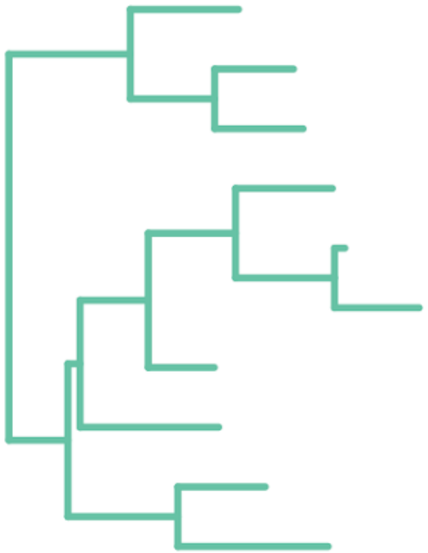
Álvarez-Carretero et al. (2021) — 4,705 mammal species

Times and rates are not fully identifiable!



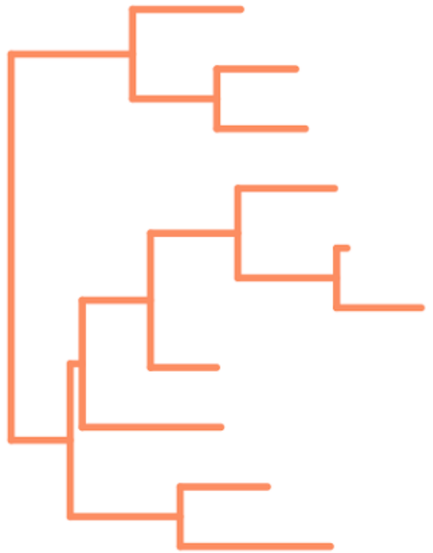
Time (year)

X



Clock rate (subs/site/year)

=



Genetic distance (subs/site)

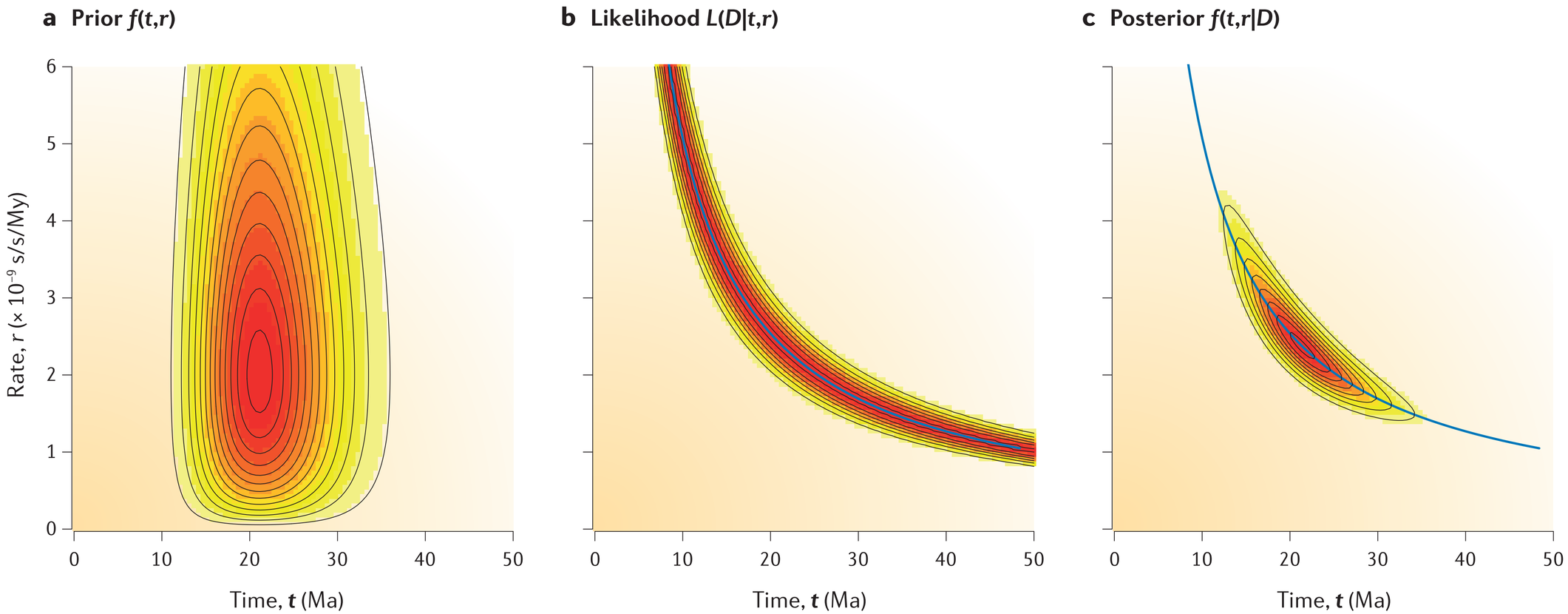
Prior

Prior

Likelihood

Slide adapted from Sebastian Duchene

The priors will always influence the results



dos Reis et al. 2015. Nature Reviews Genetics