

Phylogenetic dating

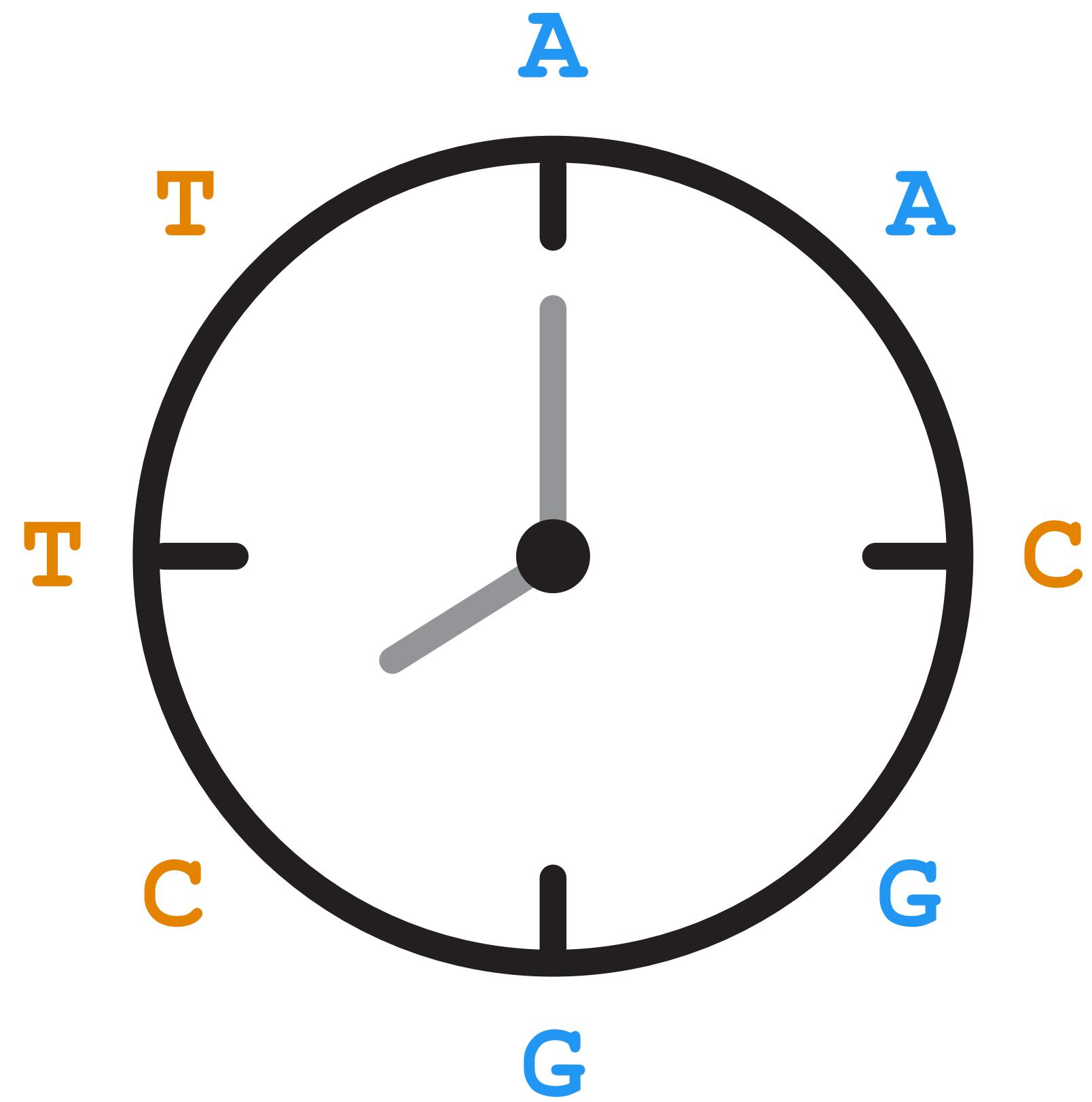
Analytical Paleobiology

Rachel Warnock
14.08.24



Today's objectives

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



Recap

Bayes' theorem

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

Bayes' theorem

Likelihood

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

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

Bayes' theorem

Priors

This represents our prior knowledge of the model parameters

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

Bayes' theorem

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

Marginal probability

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

Bayes' theorem

posterior

Reflects our combined knowledge based on the likelihood and the priors

$\Pr(\text{model} \mid \text{data}) =$

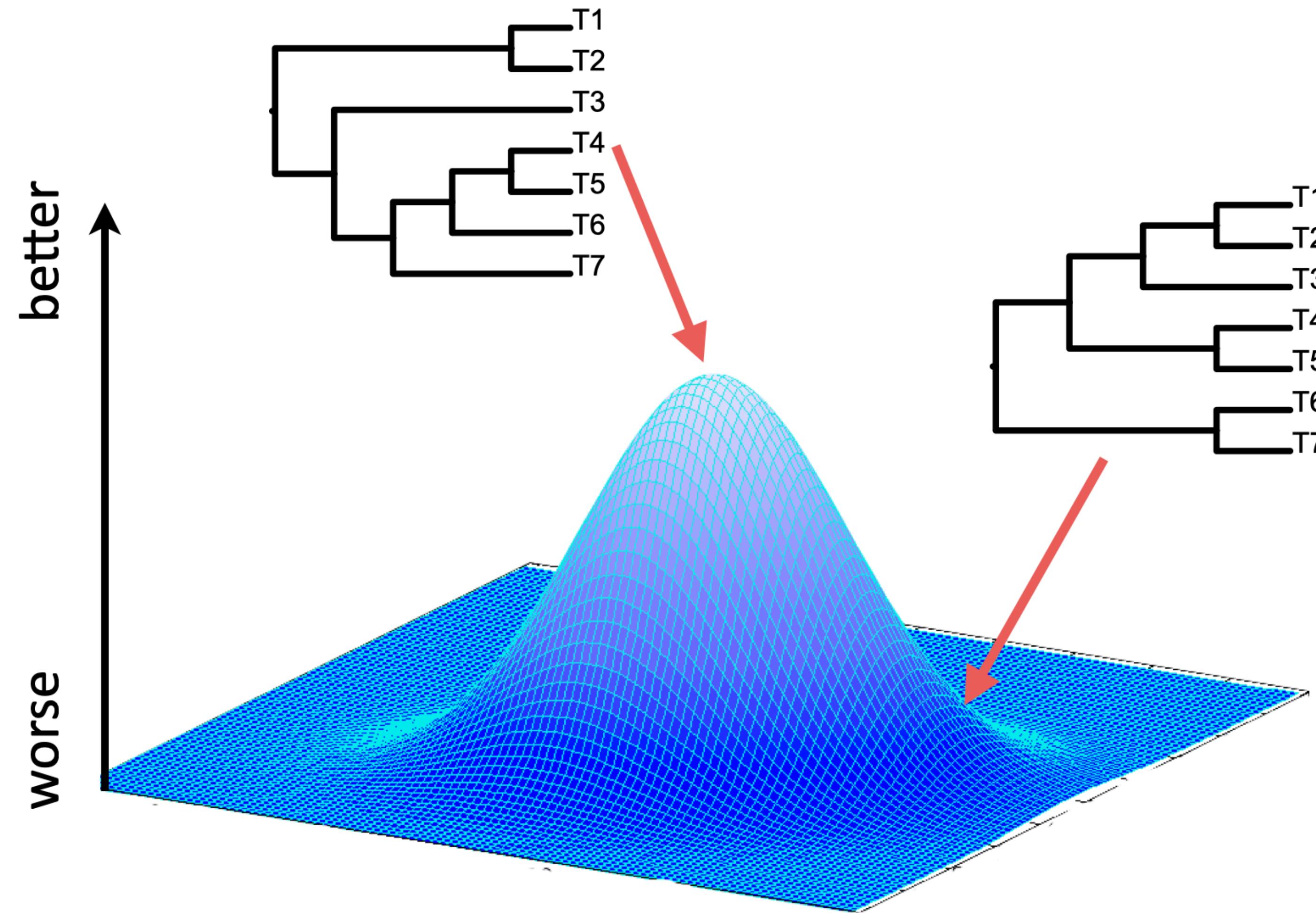
$$\frac{\Pr(\text{data} \mid \text{model}) \Pr(\text{model})}{\Pr(\text{data})}$$

Bayesian tree inference

$$\text{posterior} \quad P(E \mid \text{0101...}, \text{1101...}, \text{0100...}) = \frac{\text{likelihood} \quad P(\text{0101...}, \text{1101...}, \text{0100...} \mid E, \text{prior})}{\text{priors} \quad P(E)}$$

marginal probability

How do we find the ‘best’ tree?



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

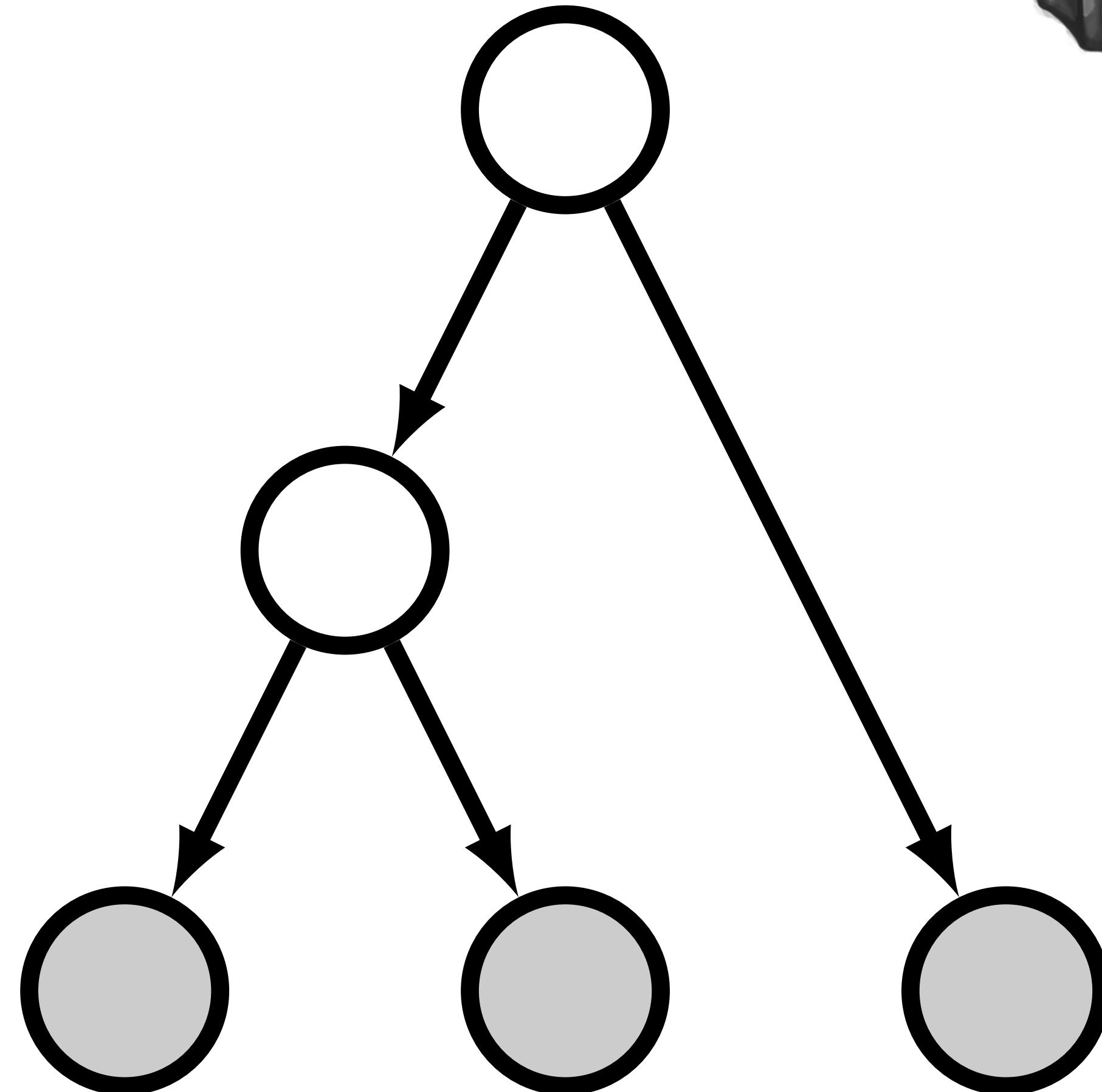
Note these are not the only approaches to tree-building but they are the most widely used

Graphical models

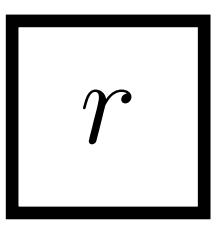


Provide tools for visually and computationally representing complex, parameter-rich models

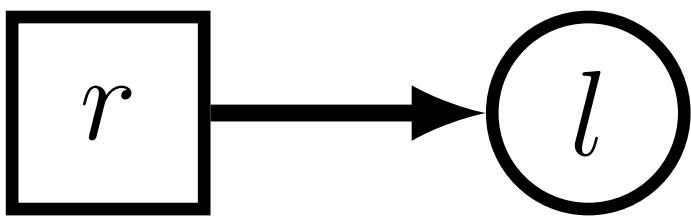
Depict the conditional dependence structure of parameters and other random variables



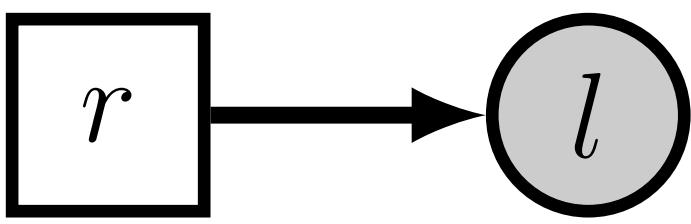
a)



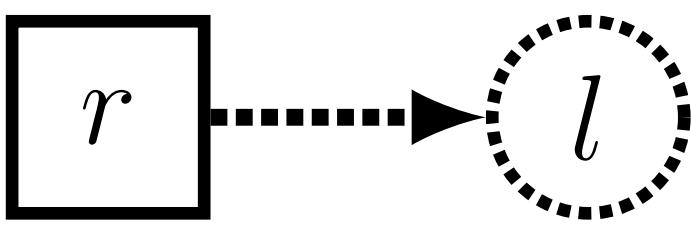
b)



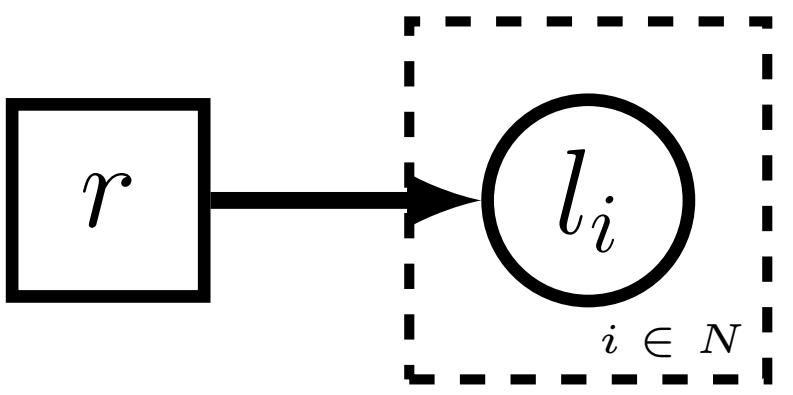
c)



d)



e)



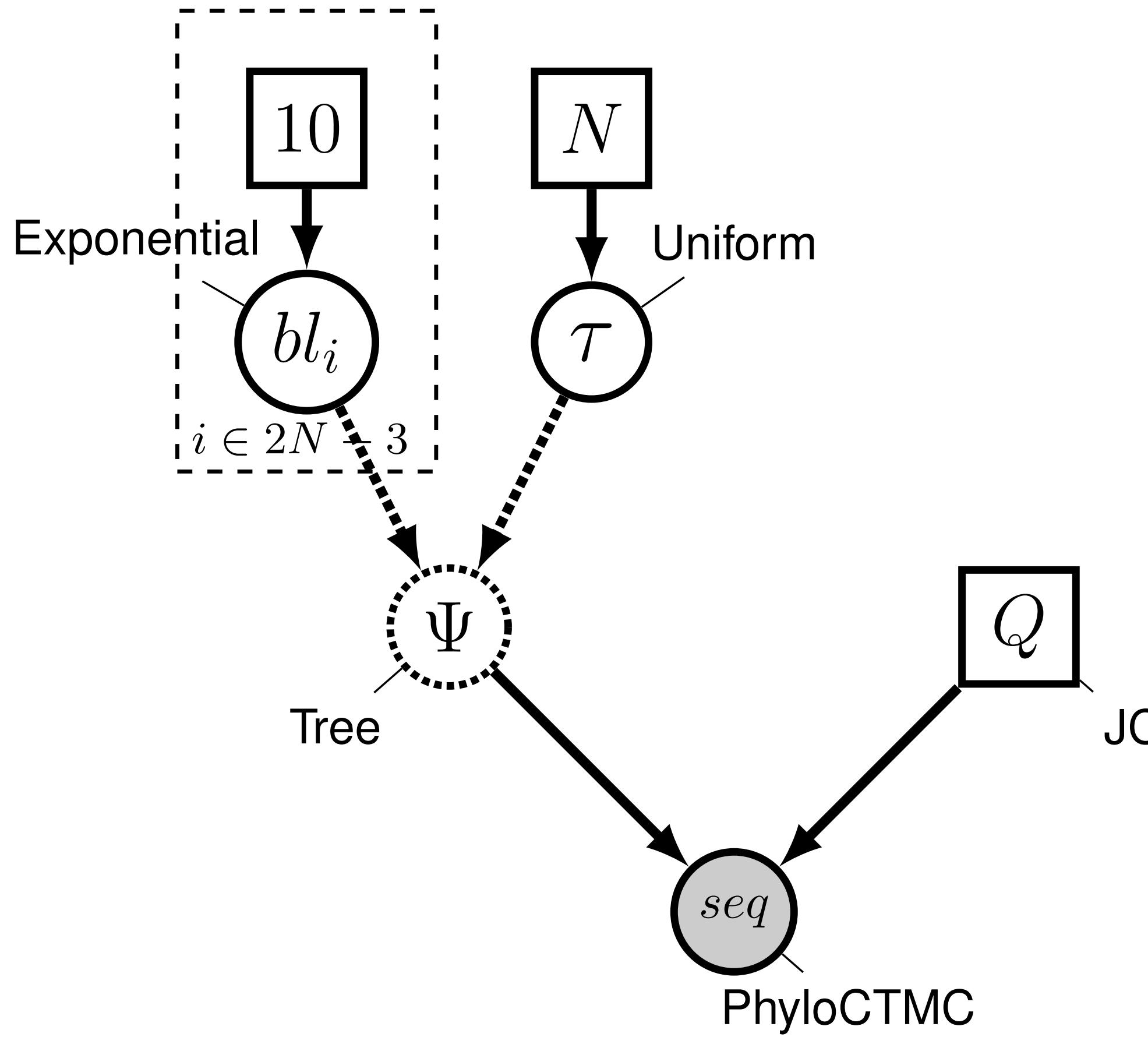
```
# constant node  
r <- 10
```

```
# stochastic node  
l ~ dnExp(r)
```

```
# stochastic node (observed)  
l.clamp(0.1)
```

```
# deterministic node  
l := exp(r)
```

```
# stochastic nodes (iid)  
for (i in 1:N) {  
  l[i] ~ dnExp(r)  
}
```



```

for (I in 1:n_branches) {
  bl[I] ~ dnExponential(10.0)
}
topology ~ dnUniformTopology(taxa)
psi := treeAssembly(topology, bl)

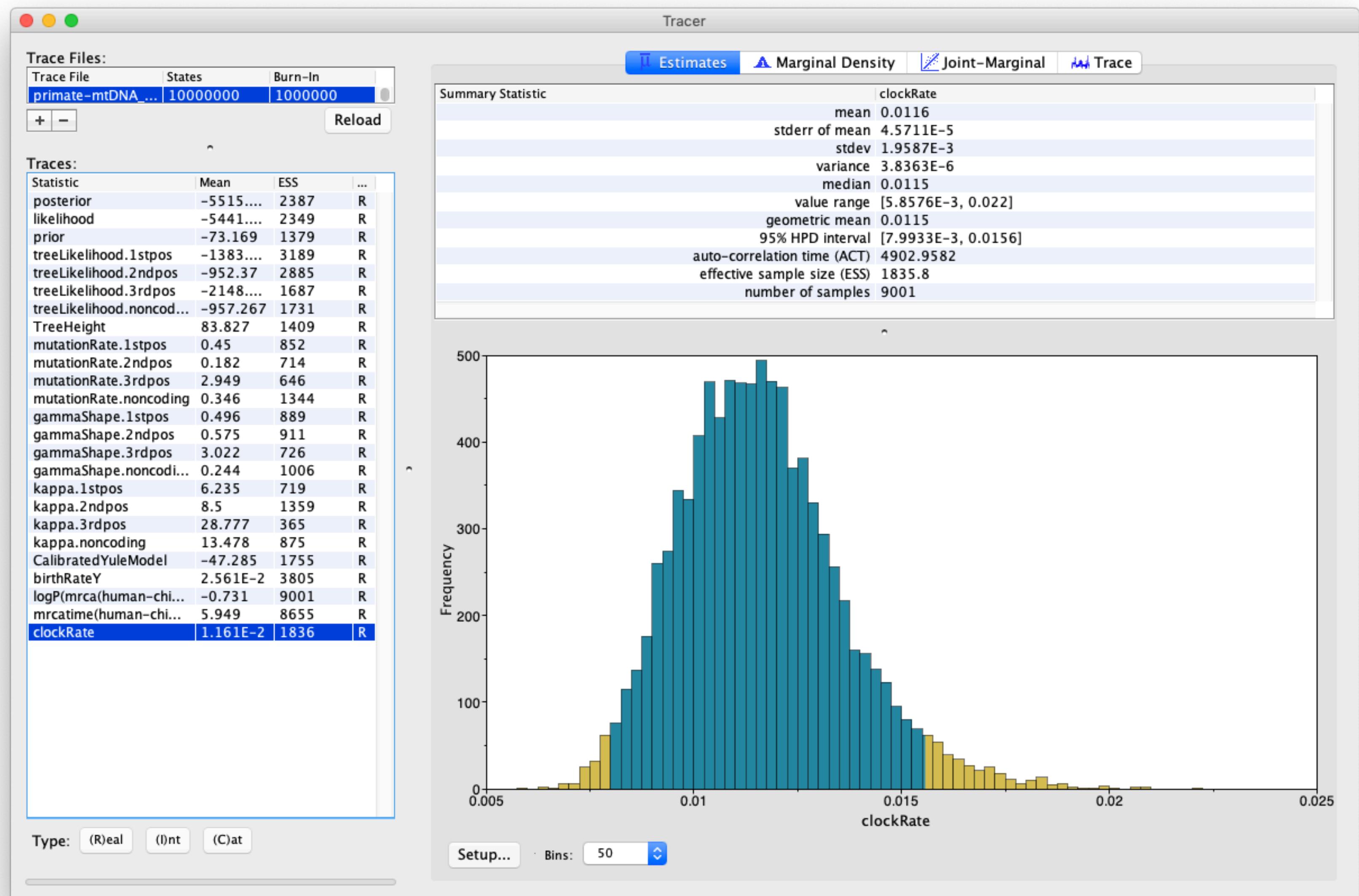
Q_morpho <- fnJC(2)

phyMorpho ~ dnPhyloCTMC( tree=psi,
siteRates=rates_morpho, Q=Q_morpho,
type="Standard", coding="variable" )
phyMorpho.clamp( data )

```

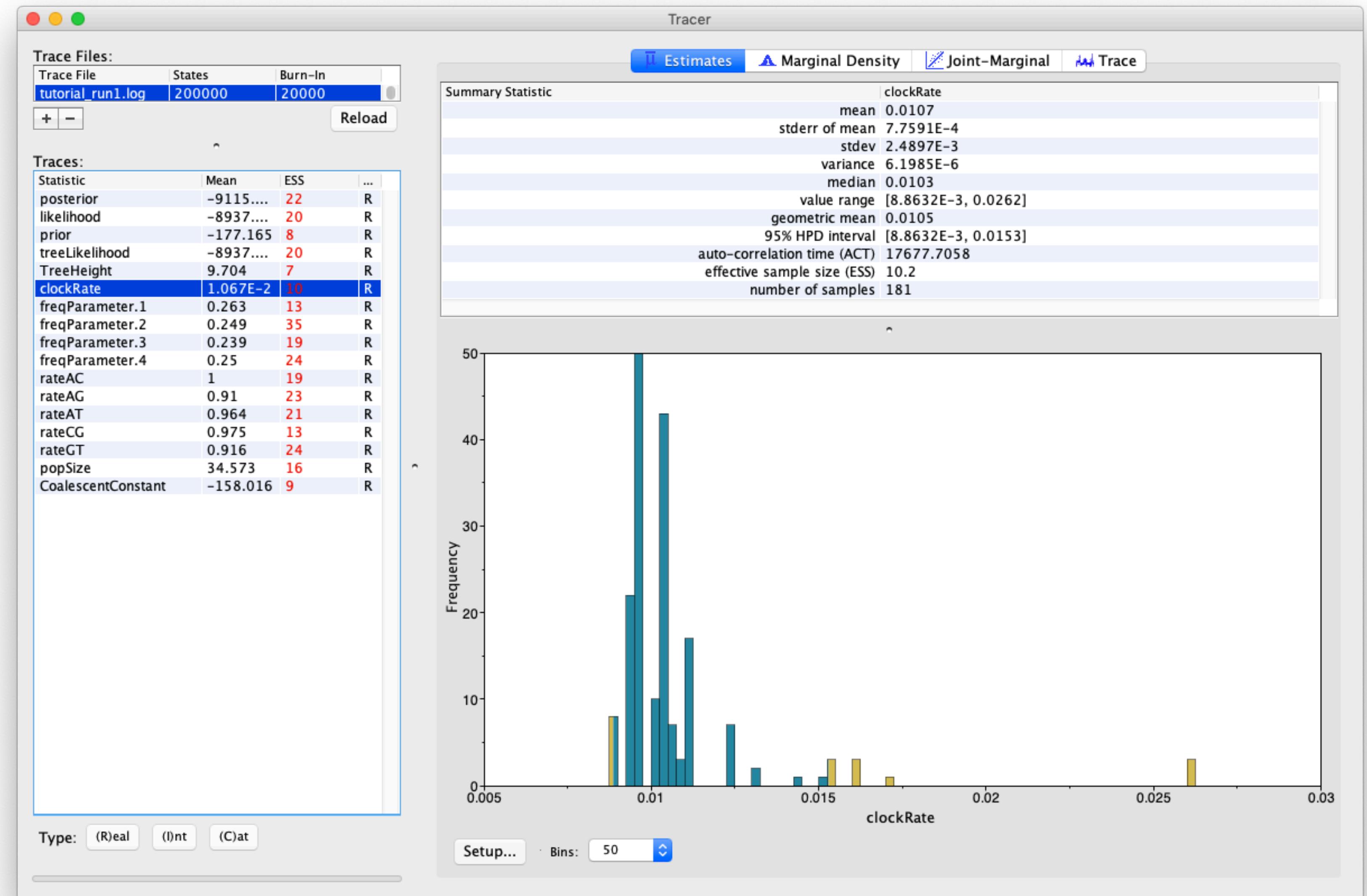
Summarising the posterior

Tracer is an amazing program for exploring MCMC output



Summarising the posterior

Tracer is an amazing program for exploring MCMC output

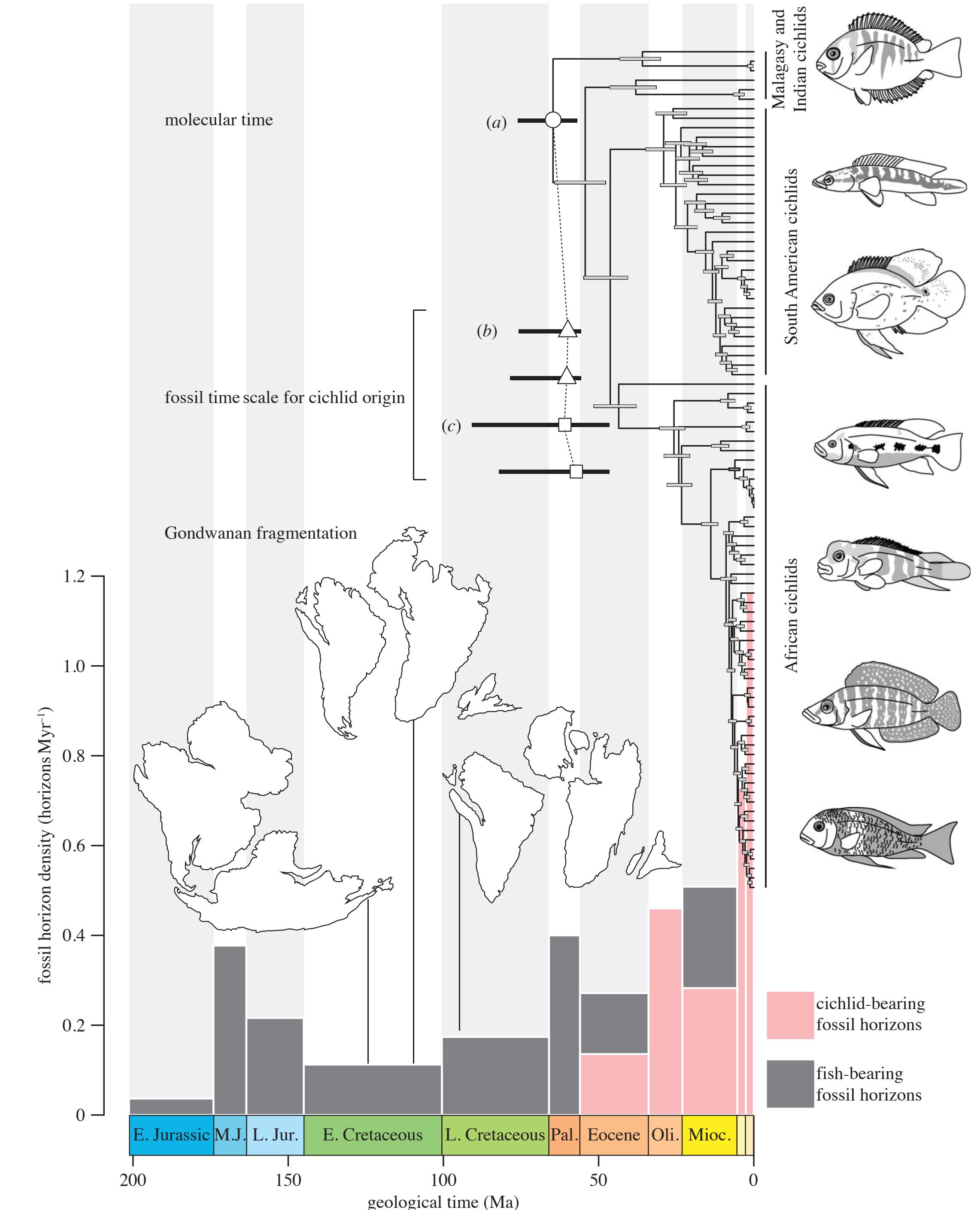


Introduction to molecular dating

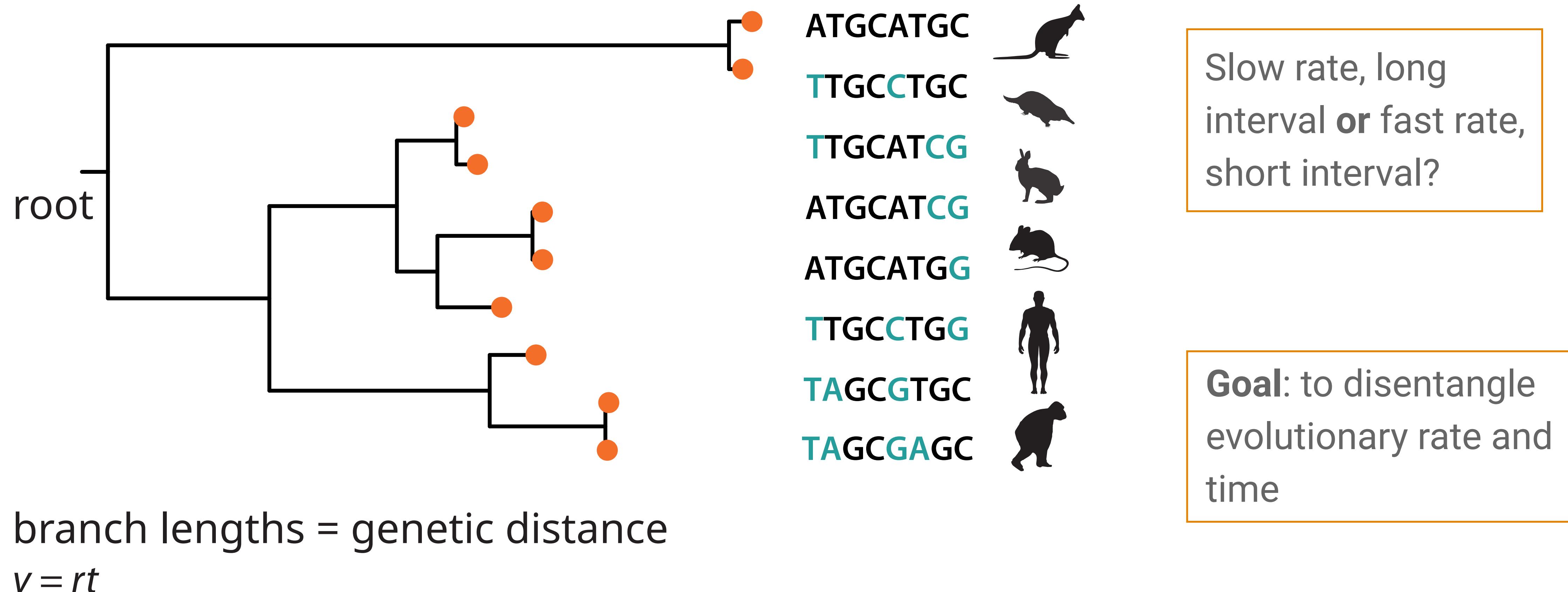
What can we learn from trees?

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

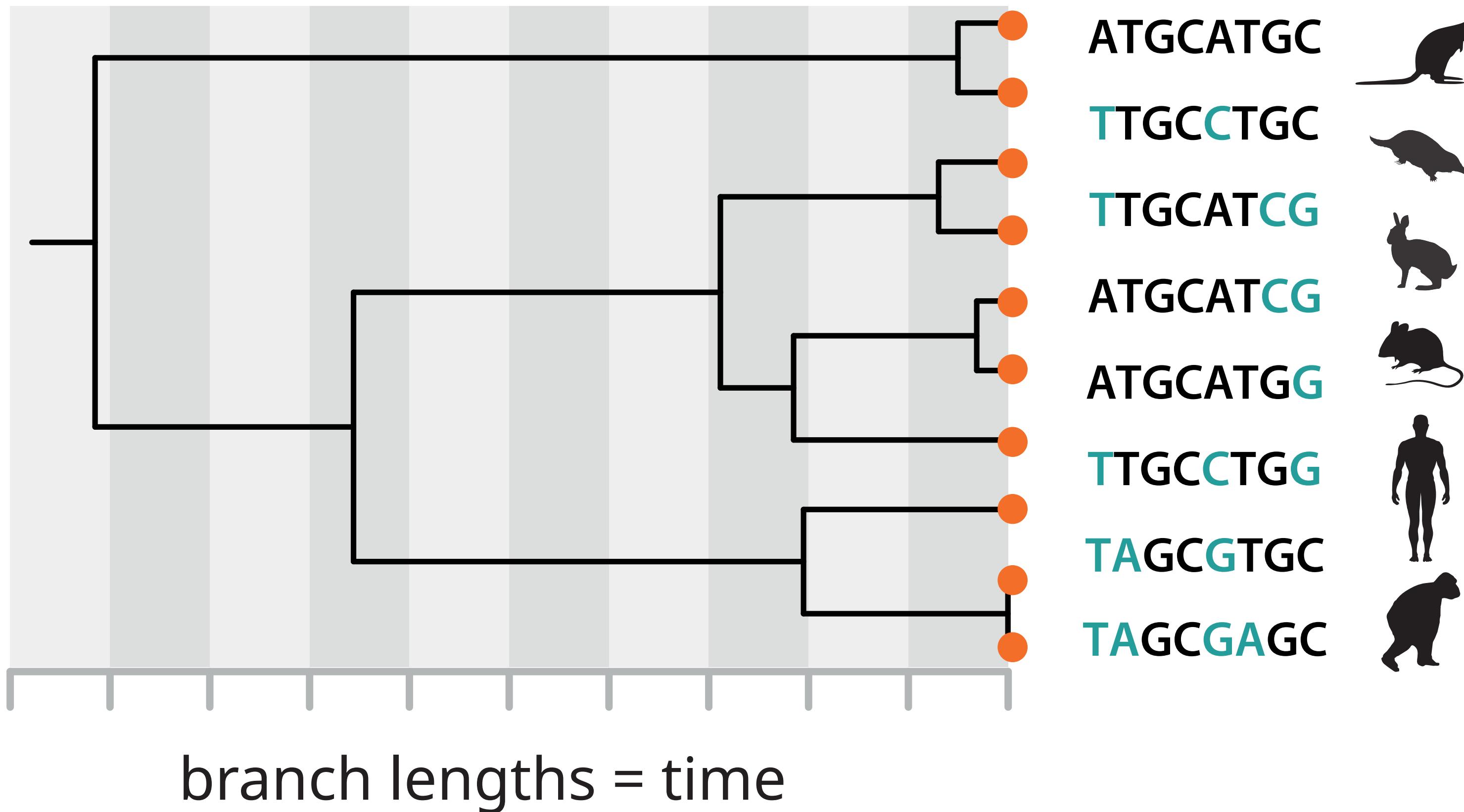
Image adapted from Friedmann et al. (2013)



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

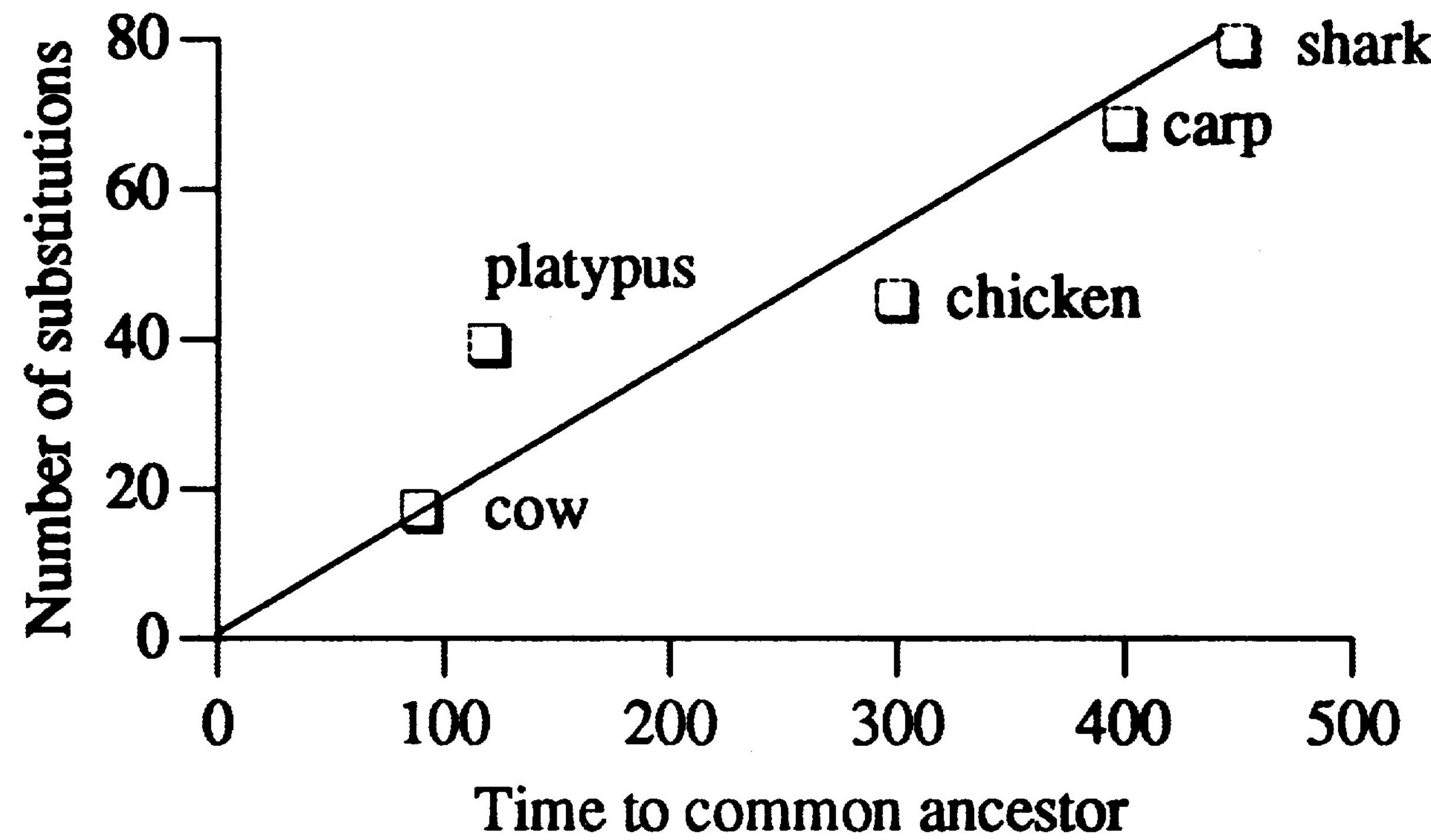


Molecular (or morphological) characters are not independently informative about time



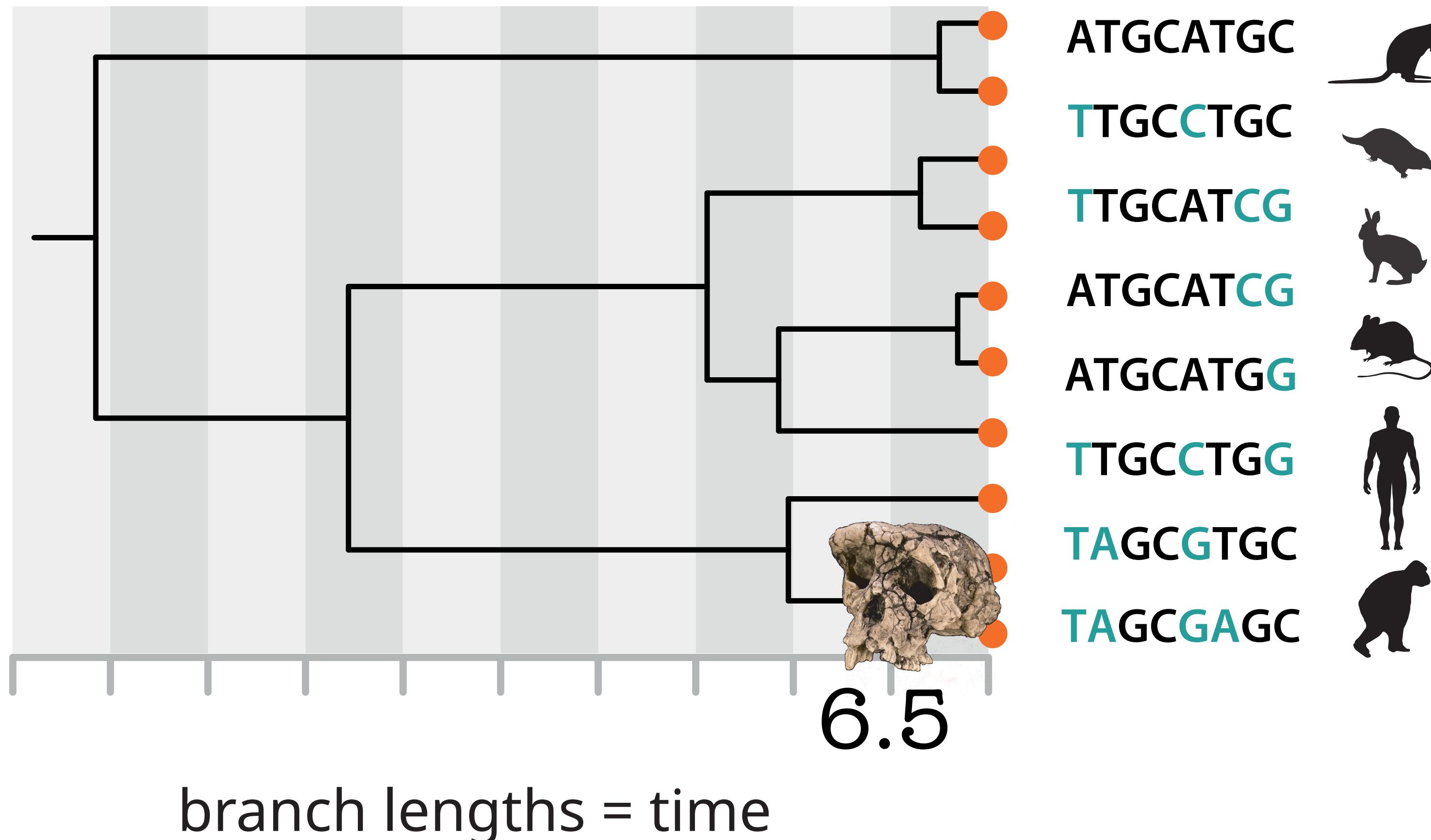
Goal: to disentangle
evolutionary rate and
time

The molecular clock hypothesis



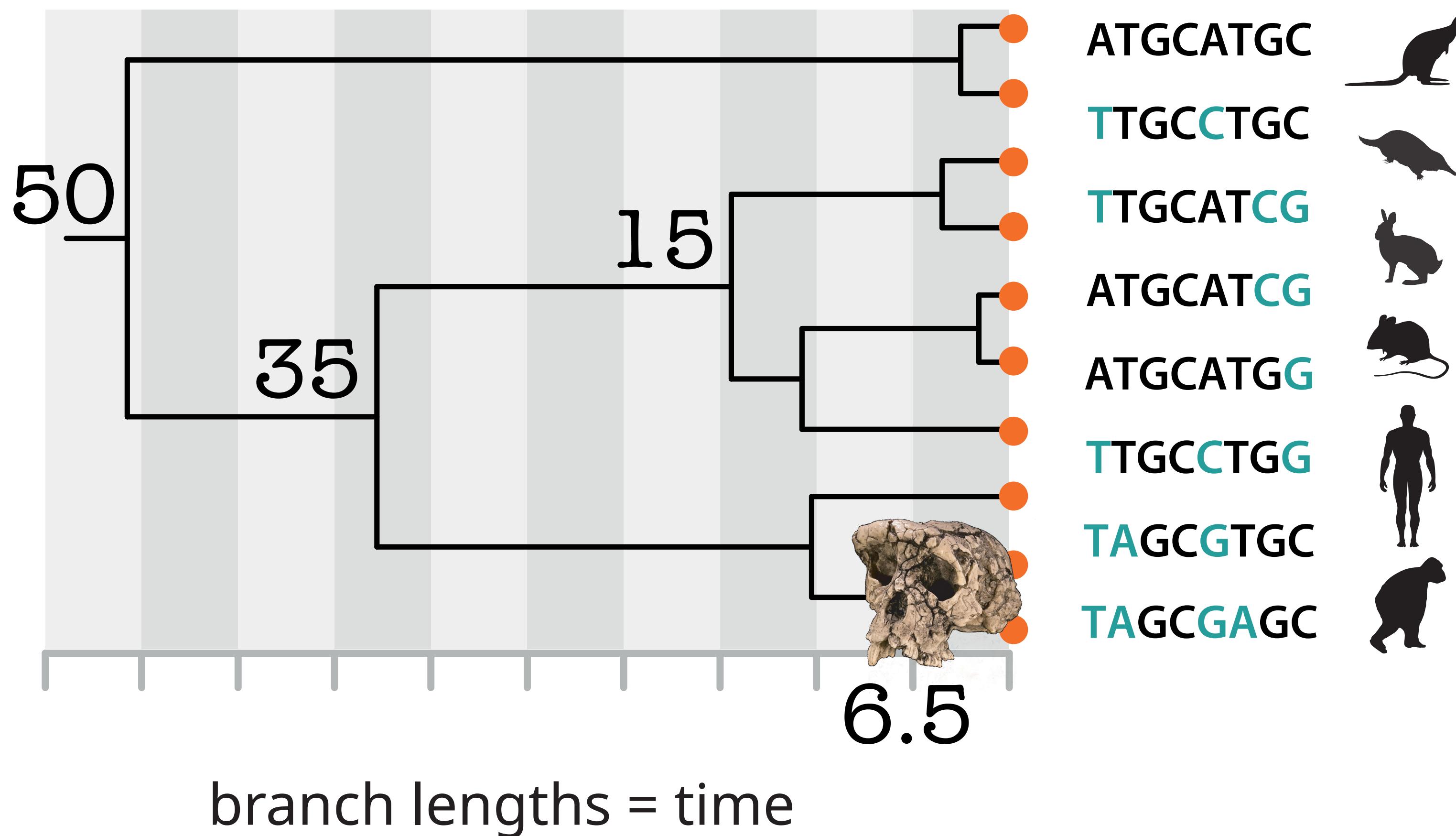
Molecules as documents of evolutionary history Zuckerkandl & Pauling ([1965](#))
A history of the molecular clock Morgan ([1998](#))

Calibrating the substitution rate



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

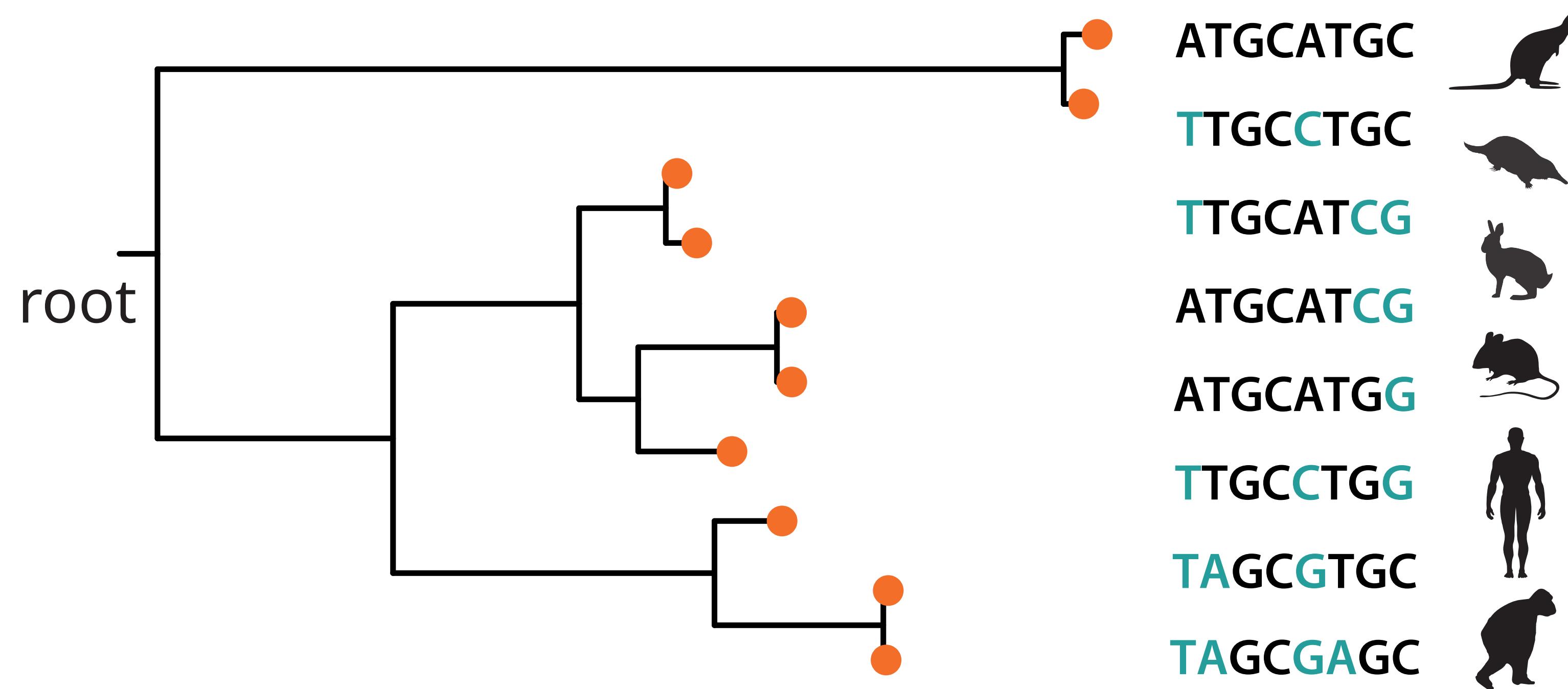
Calibrating the substitution rate



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

Molecular dating: challenges

Rate and time are not fully identifiable!

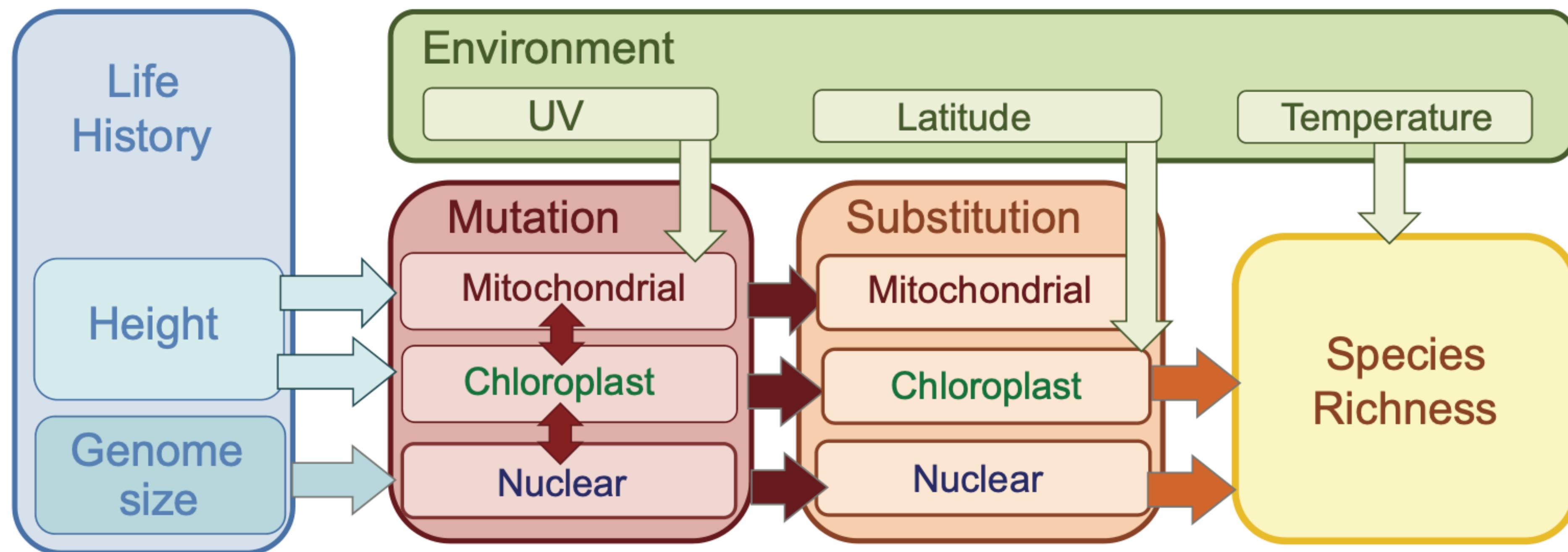


branch lengths = genetic distance

$$v = rt$$

Molecular dating: challenges

Many variables contribute to variation in the substitution rate



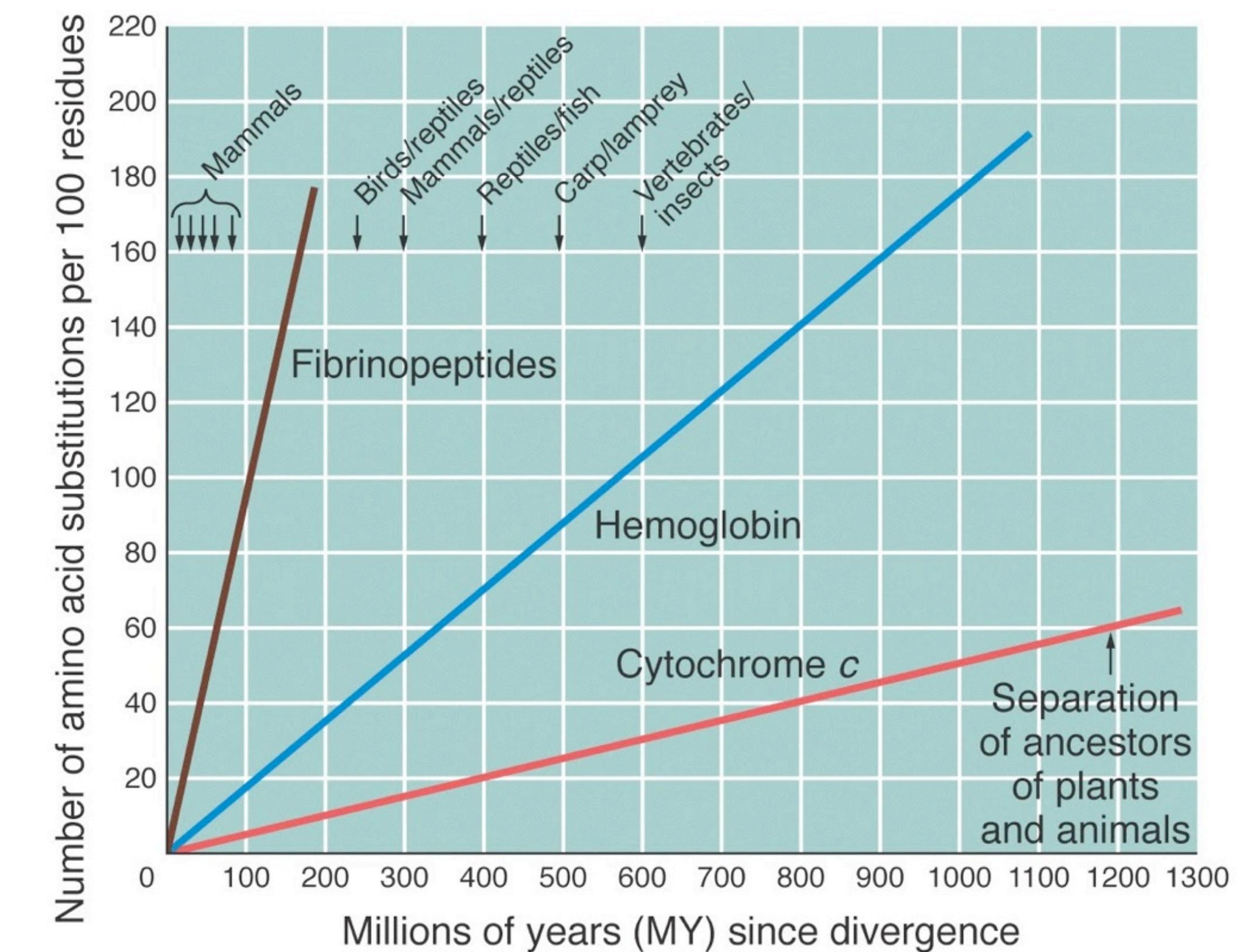
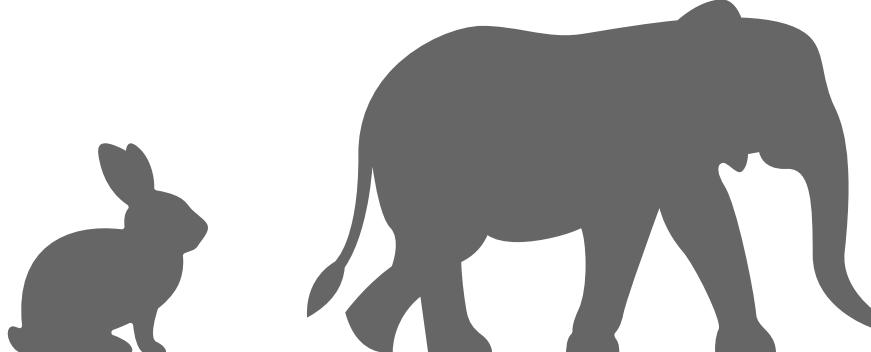
Molecular dating: challenges

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

The molecular clock is not constant

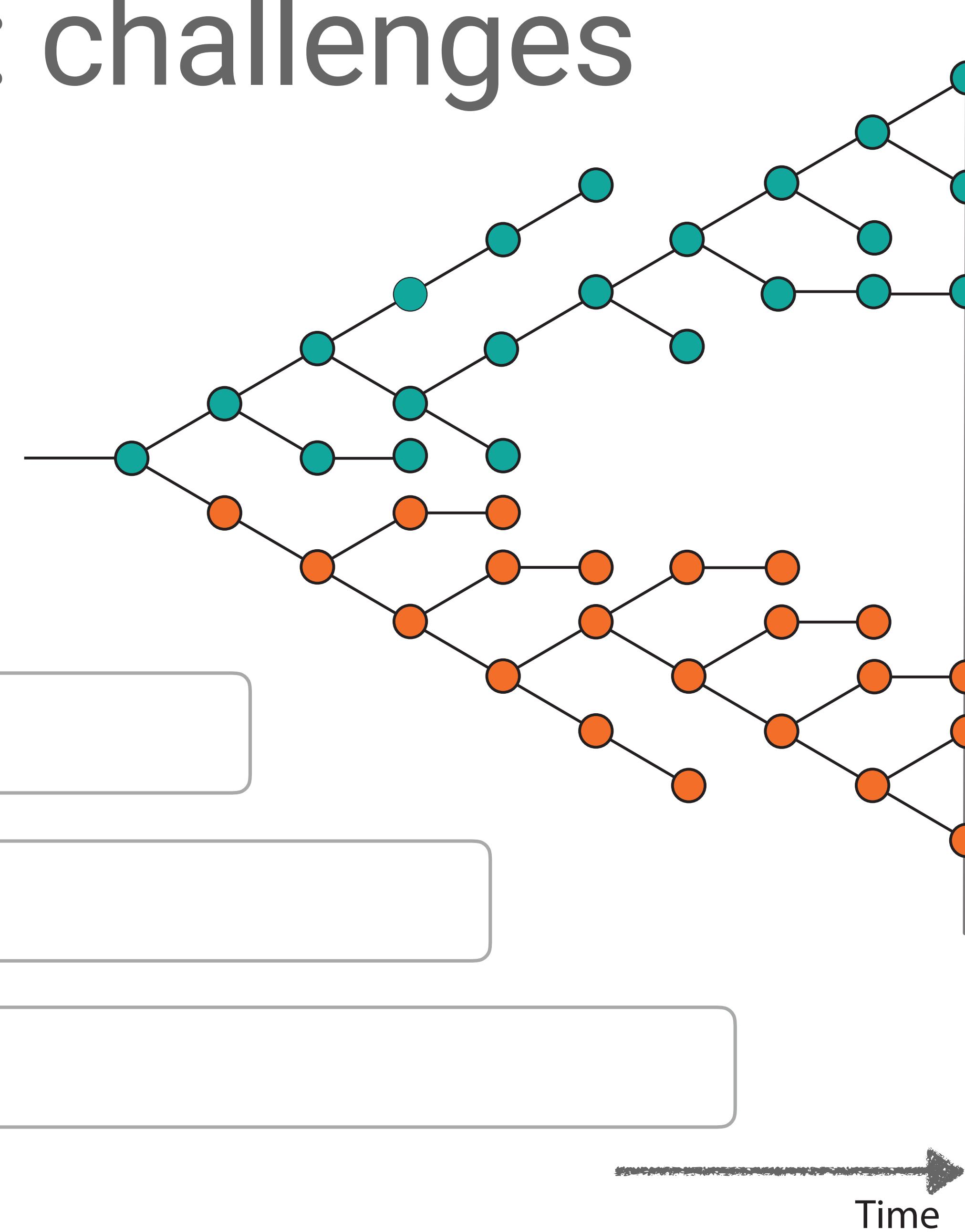
Rates vary across:

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



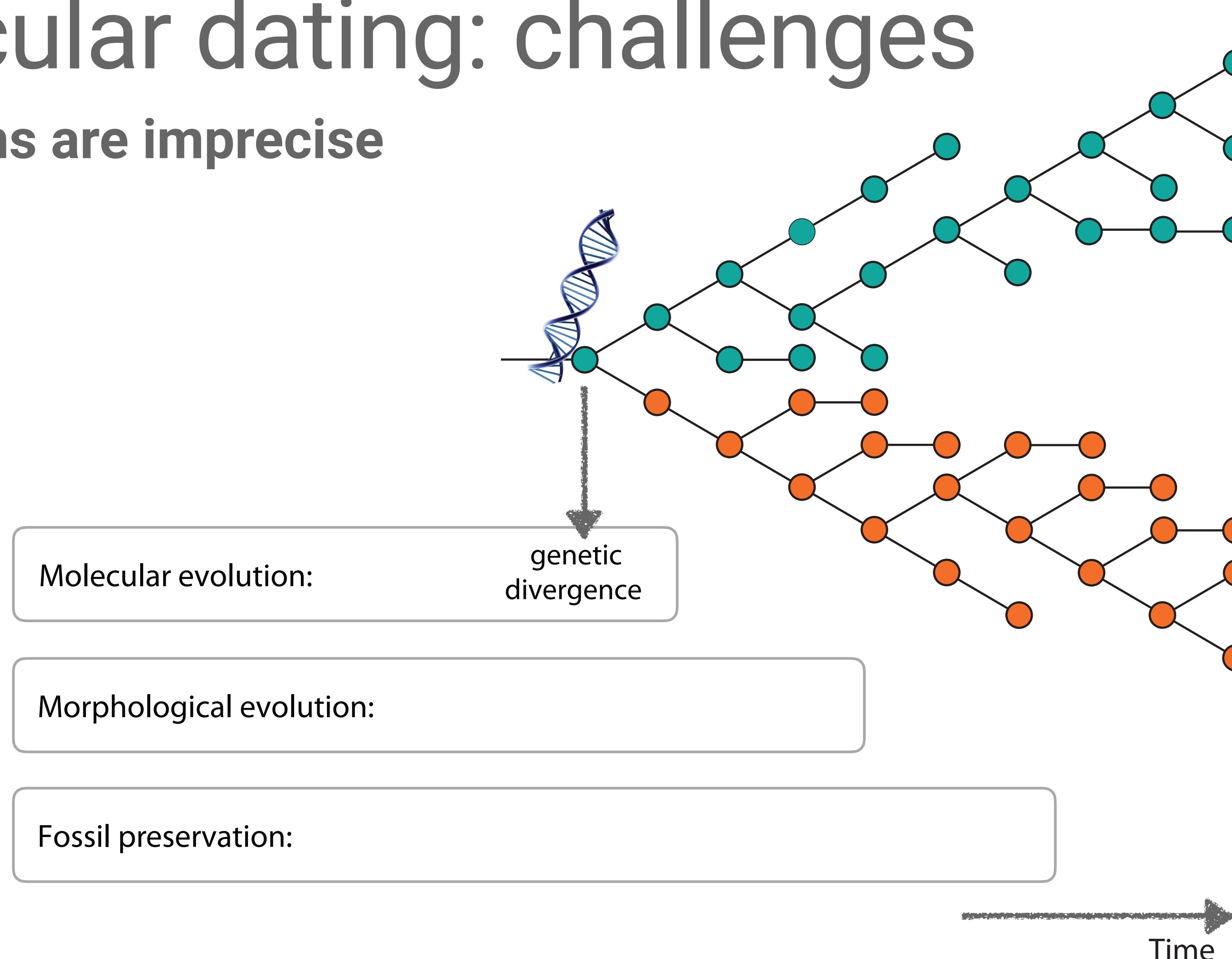
Molecular dating: challenges

Calibrations are imprecise



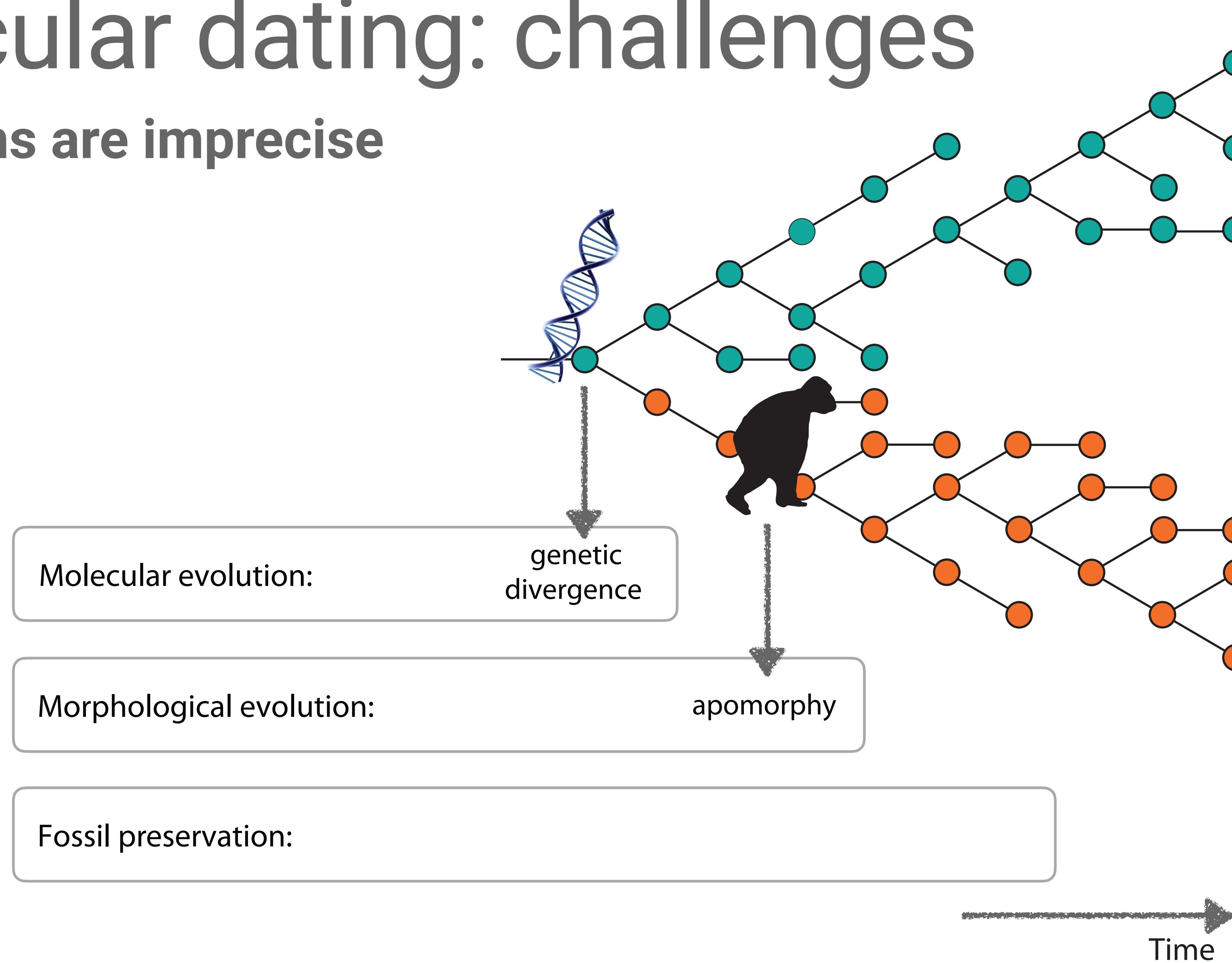
Molecular dating: challenges

Calibrations are imprecise



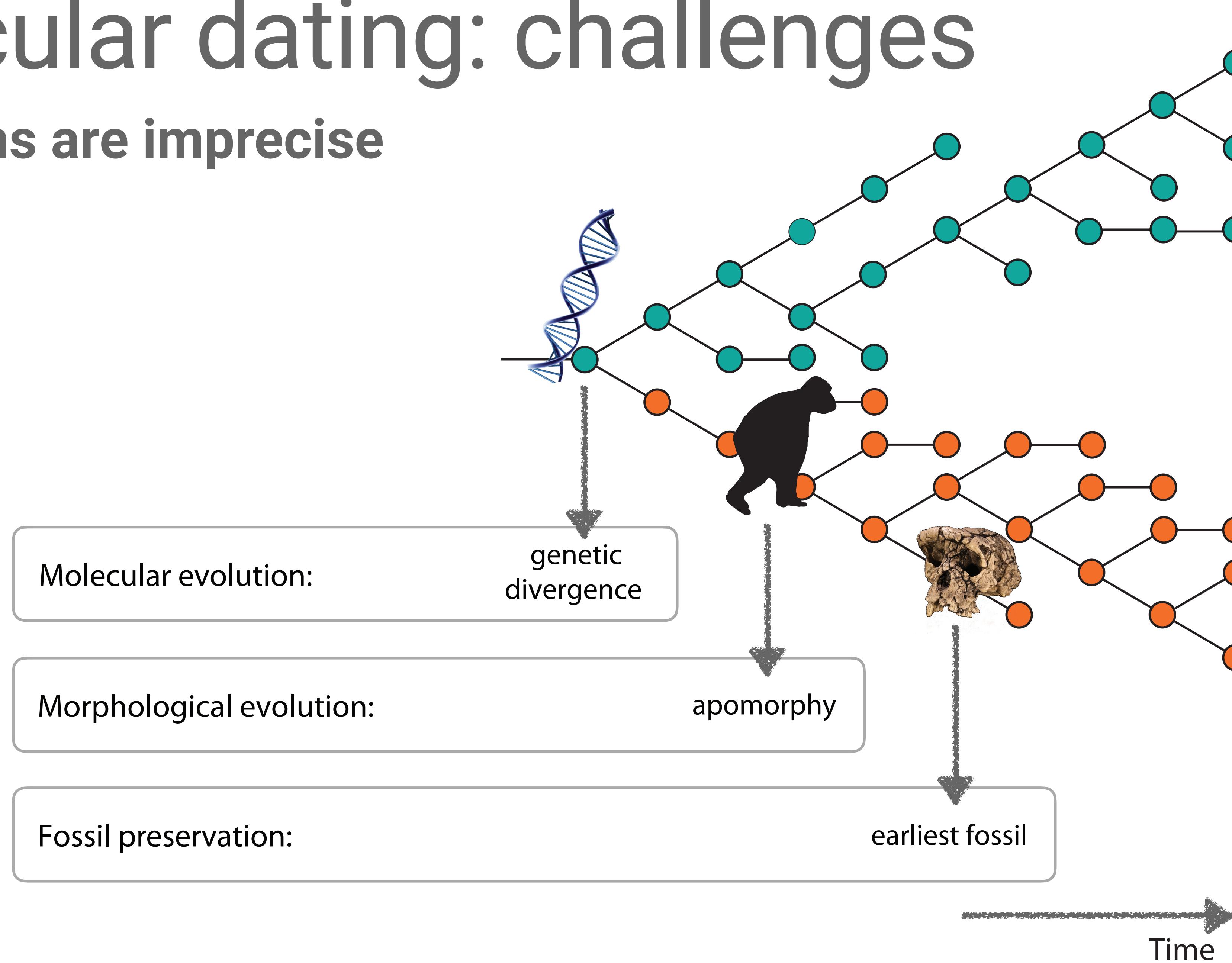
Molecular dating: challenges

Calibrations are imprecise



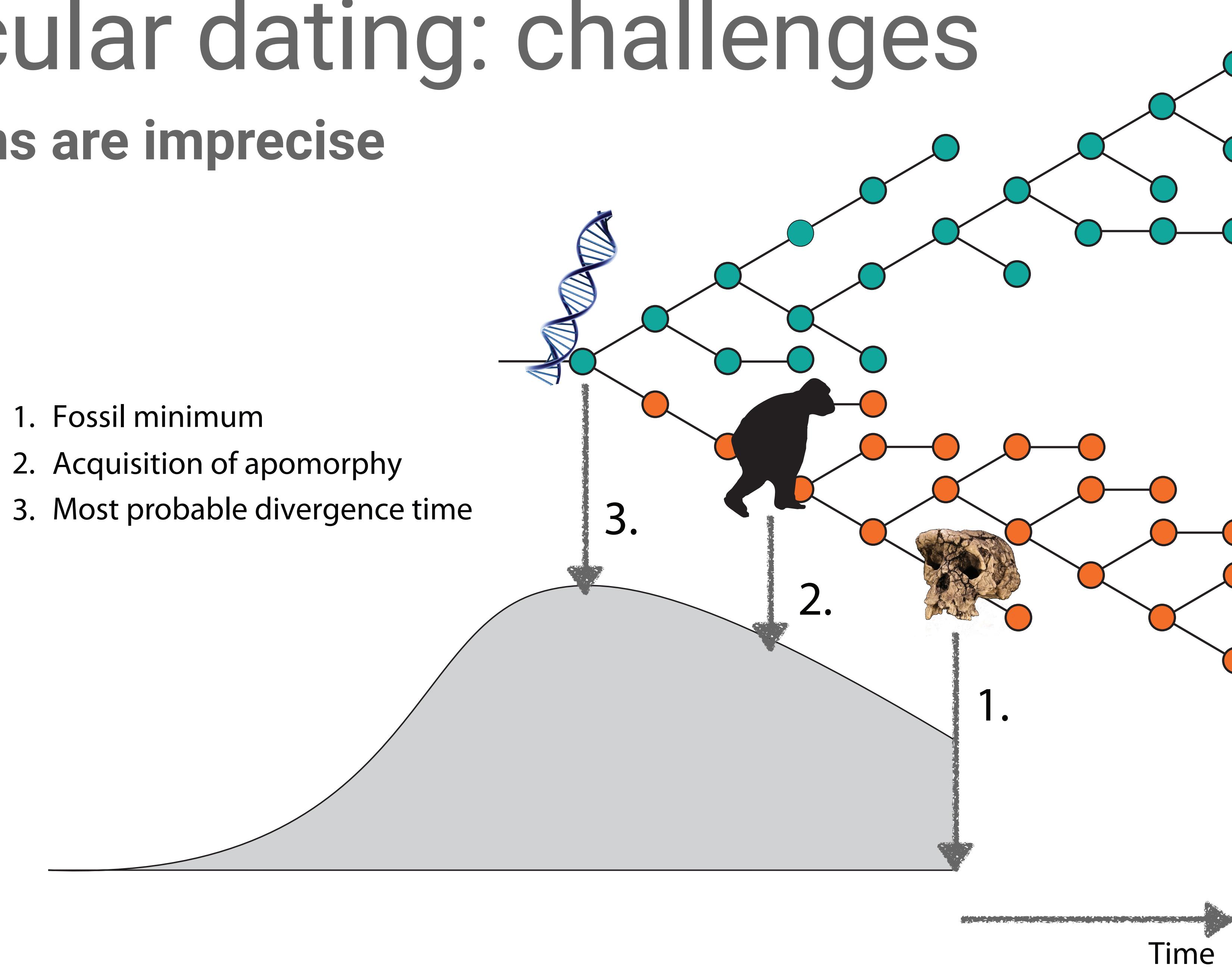
Molecular dating: challenges

Calibrations are imprecise



Molecular dating: challenges

Calibrations are imprecise



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

$$P(\text{ model } | \text{ data }) = \frac{P(\text{ data } | \text{ model }) P(\text{ model })}{P(\text{ data })}$$

likelihood

priors

posterior

marginal probability of the data

Bayesian divergence time estimation

The data

AND/OR

0101... ATTG...

1101... TTGC...

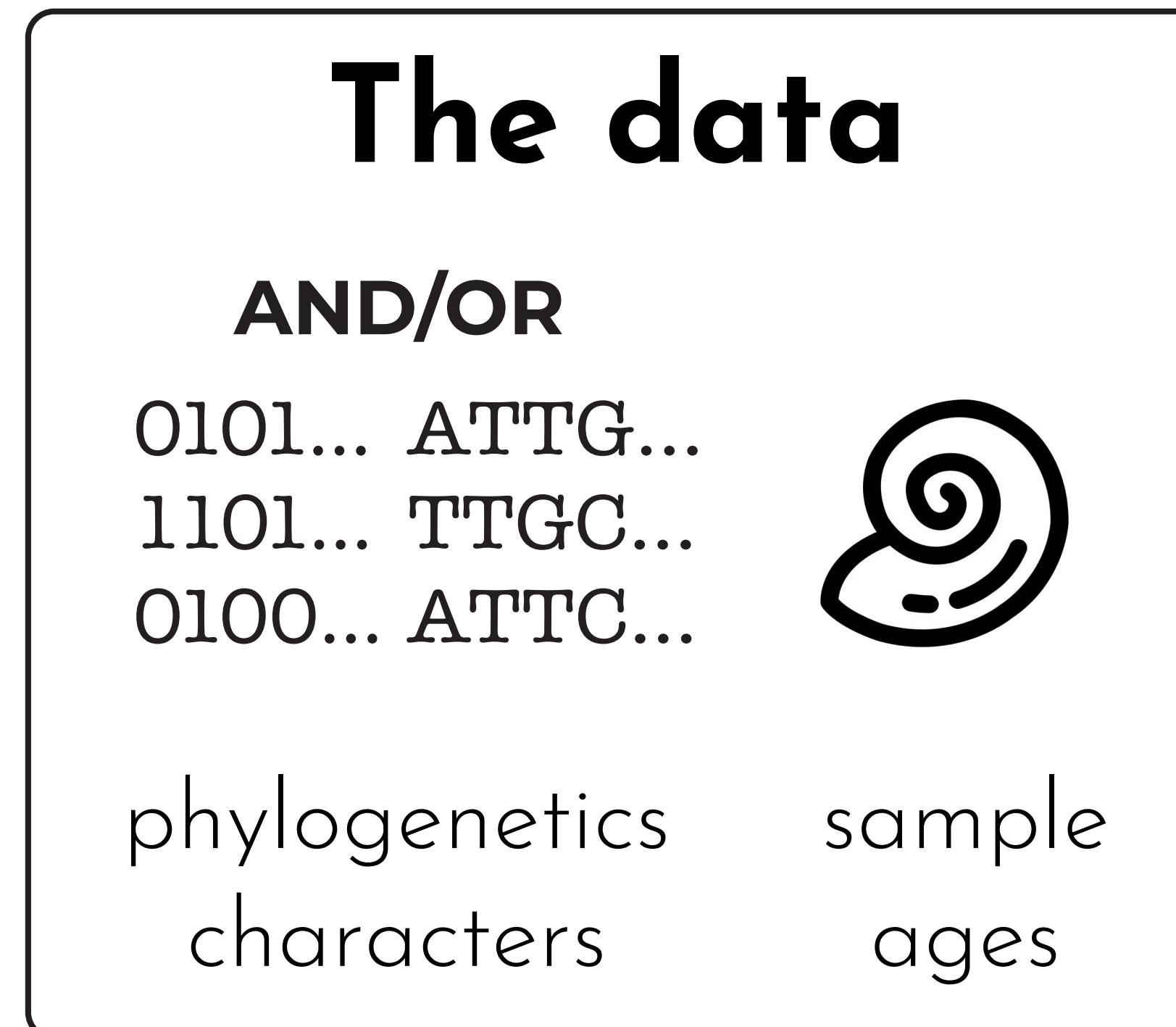
0100... ATTC...



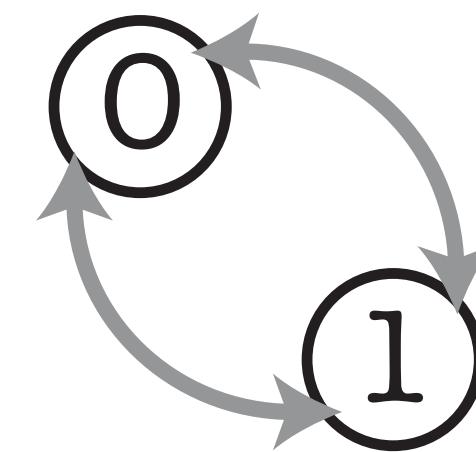
phylogenetics
characters

sample
ages

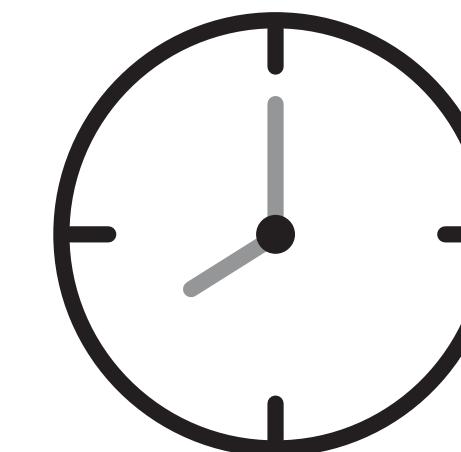
Bayesian divergence time estimation



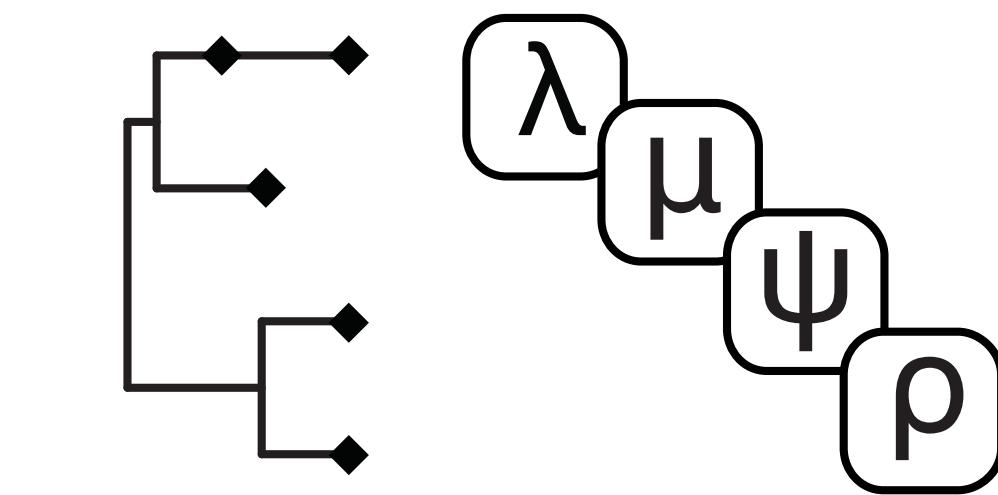
3 model components



substitution
model

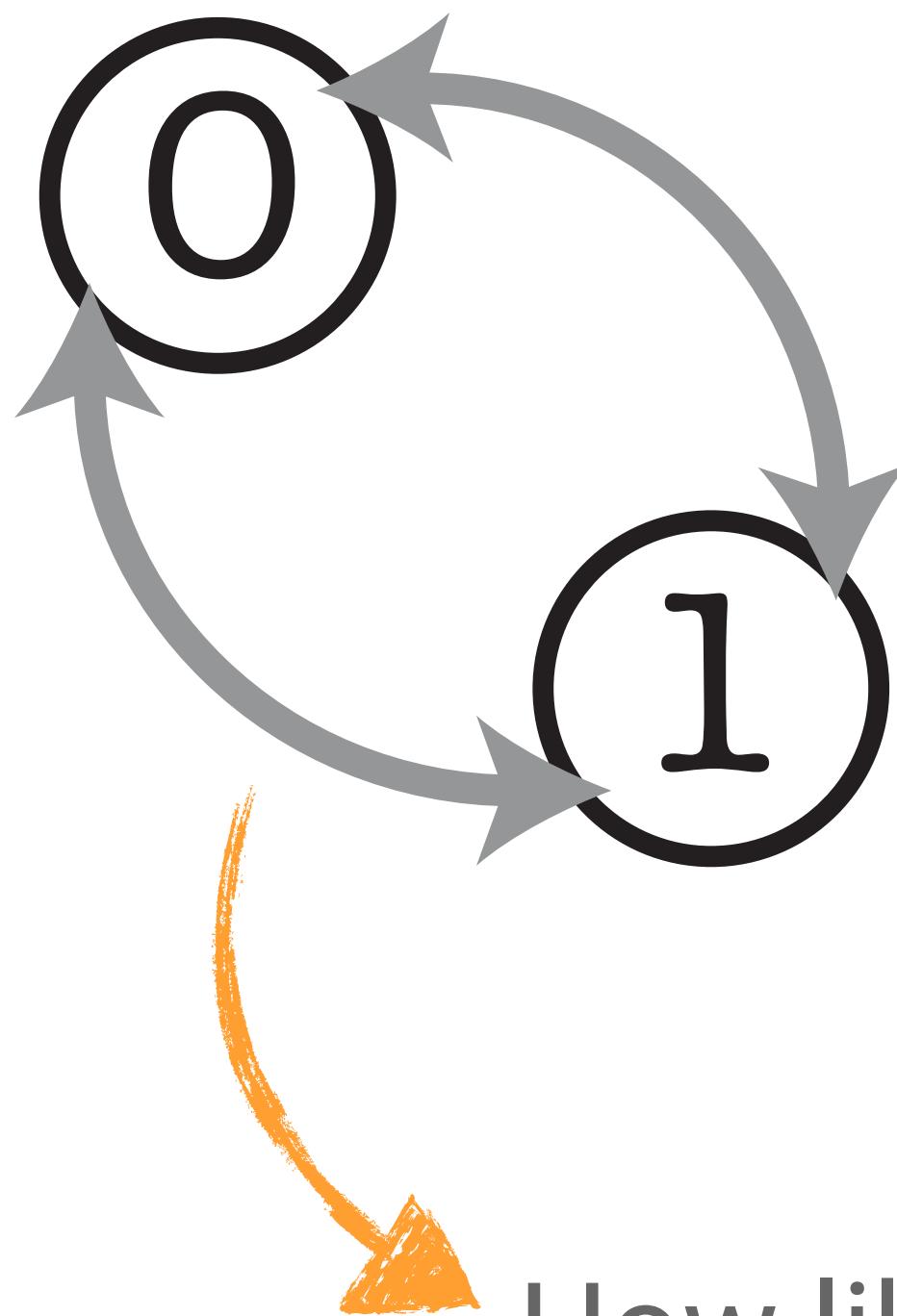


clock
model

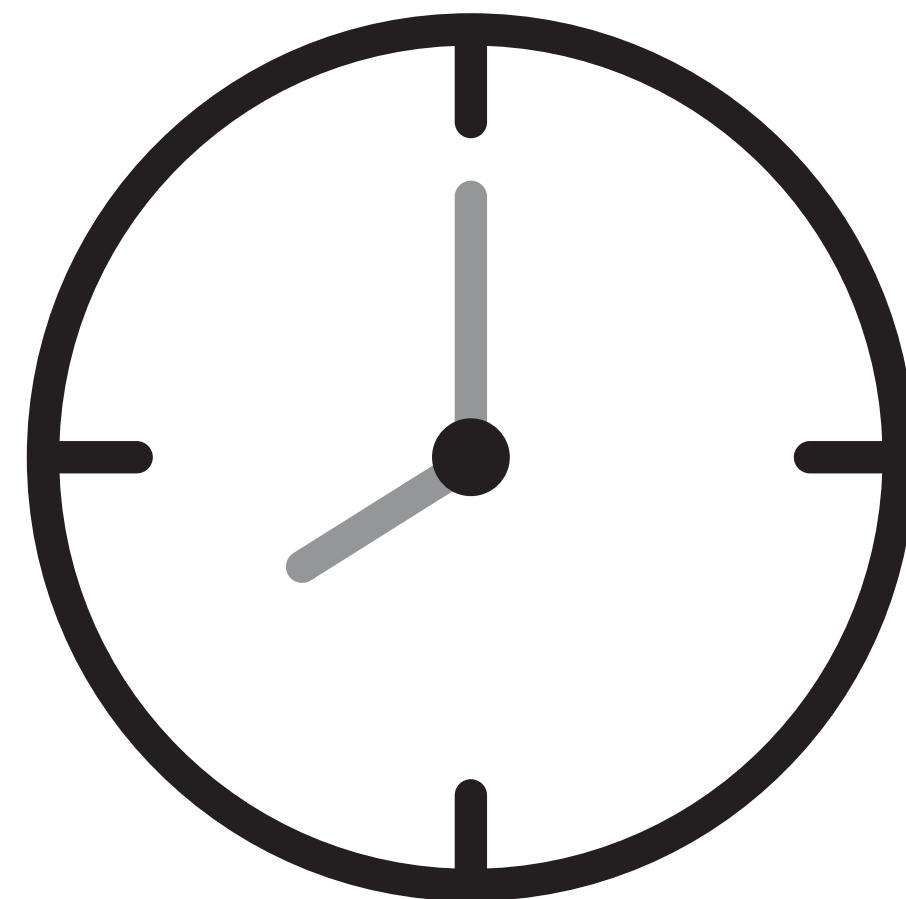


tree and
tree model

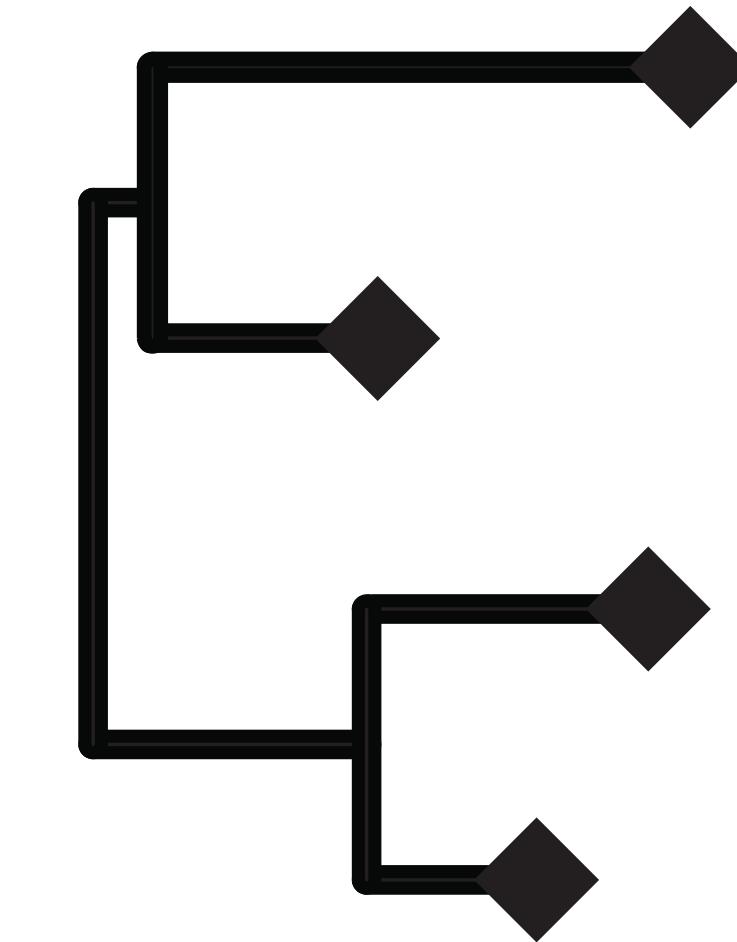
substitution model



clock model

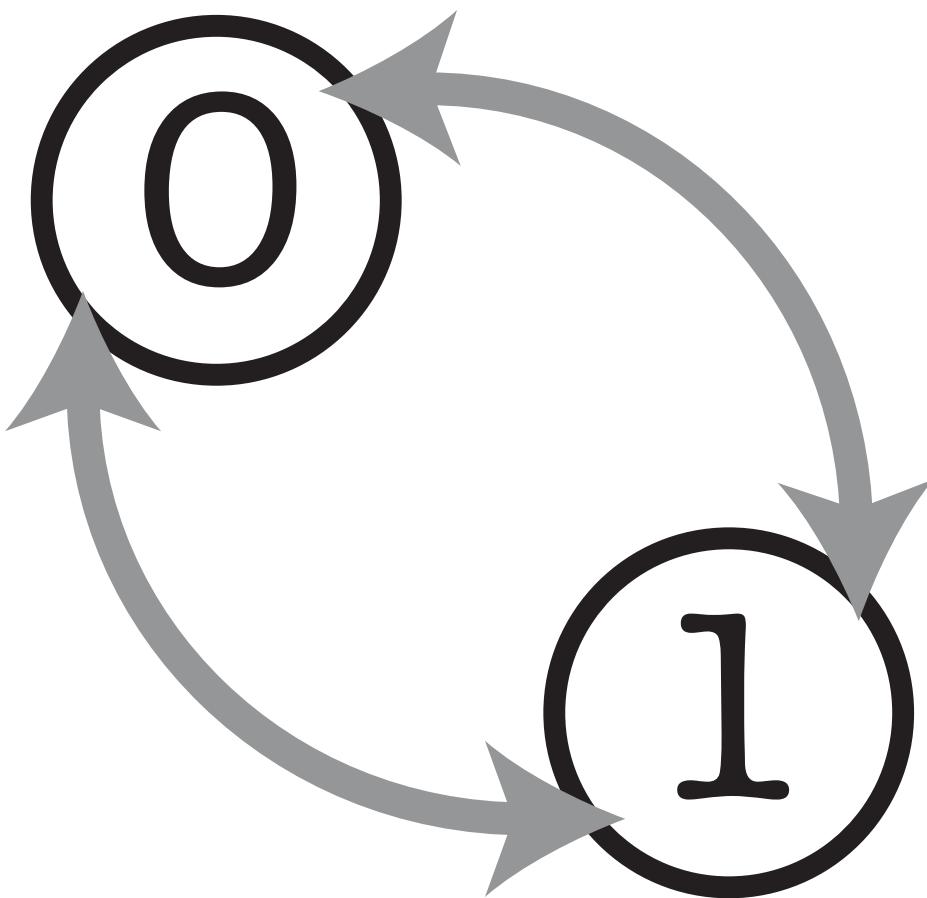


tree model

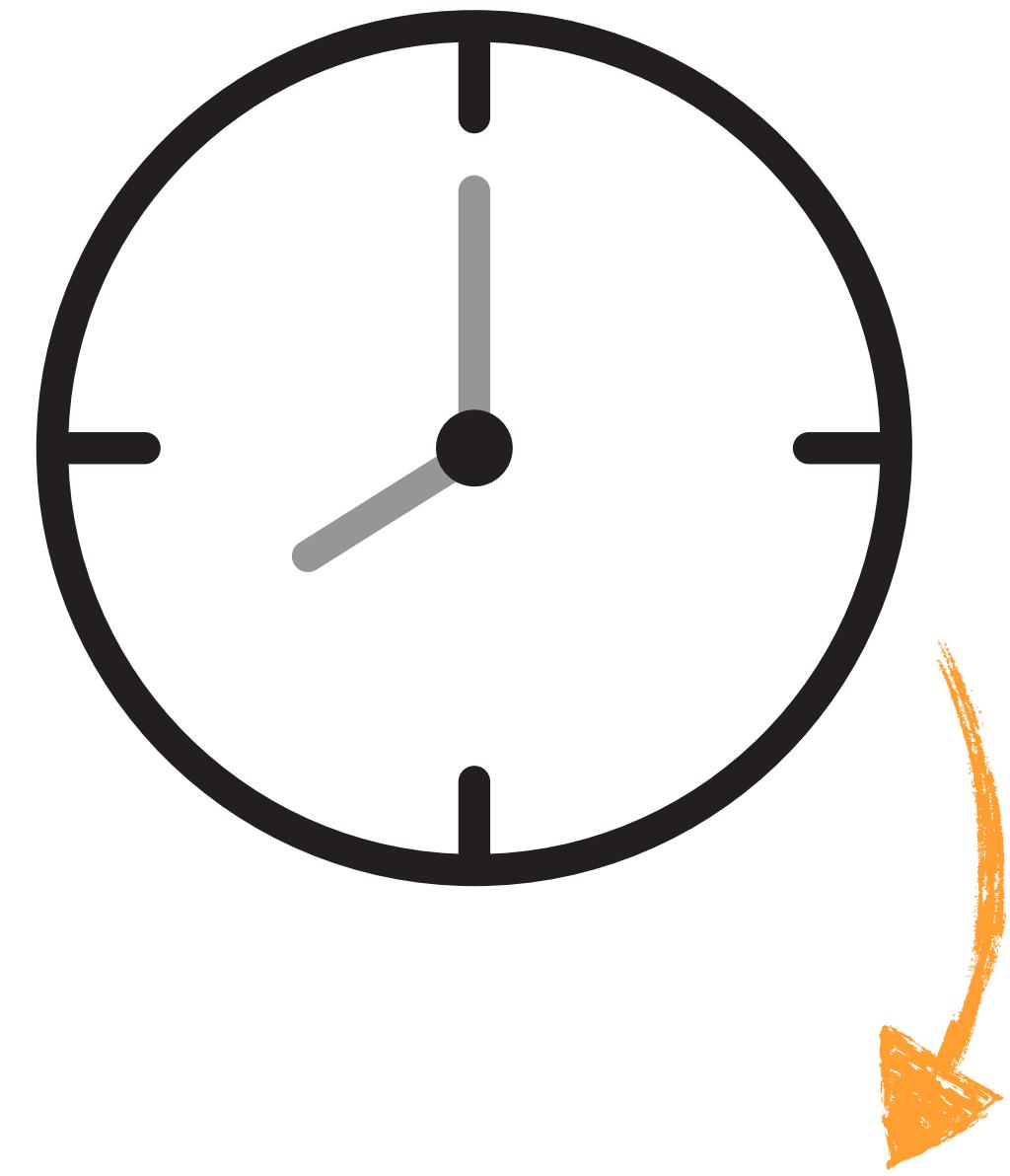


How likely are we to observe a change
between character states? e.g., A → T

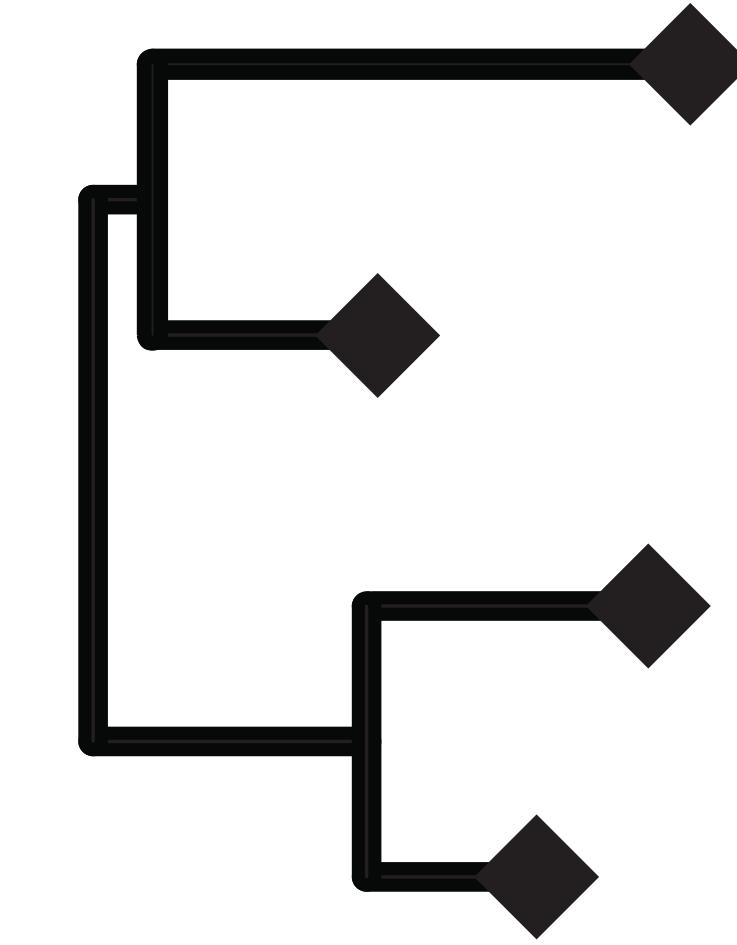
substitution model



clock model

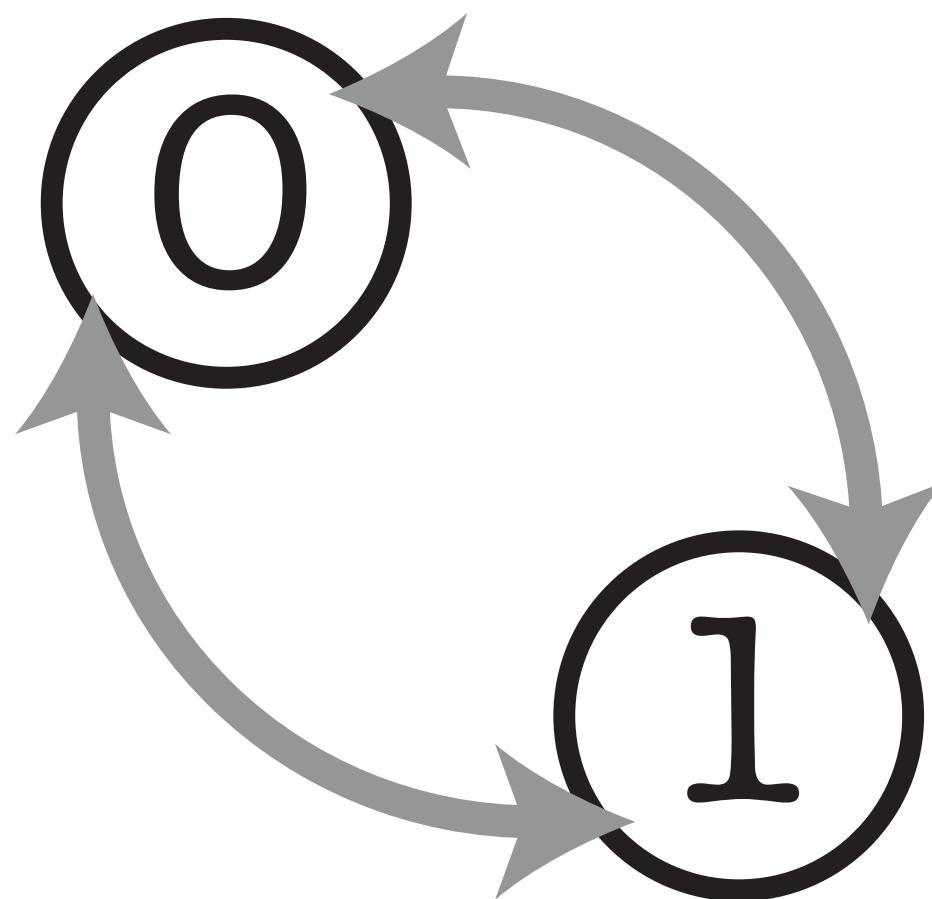


tree model

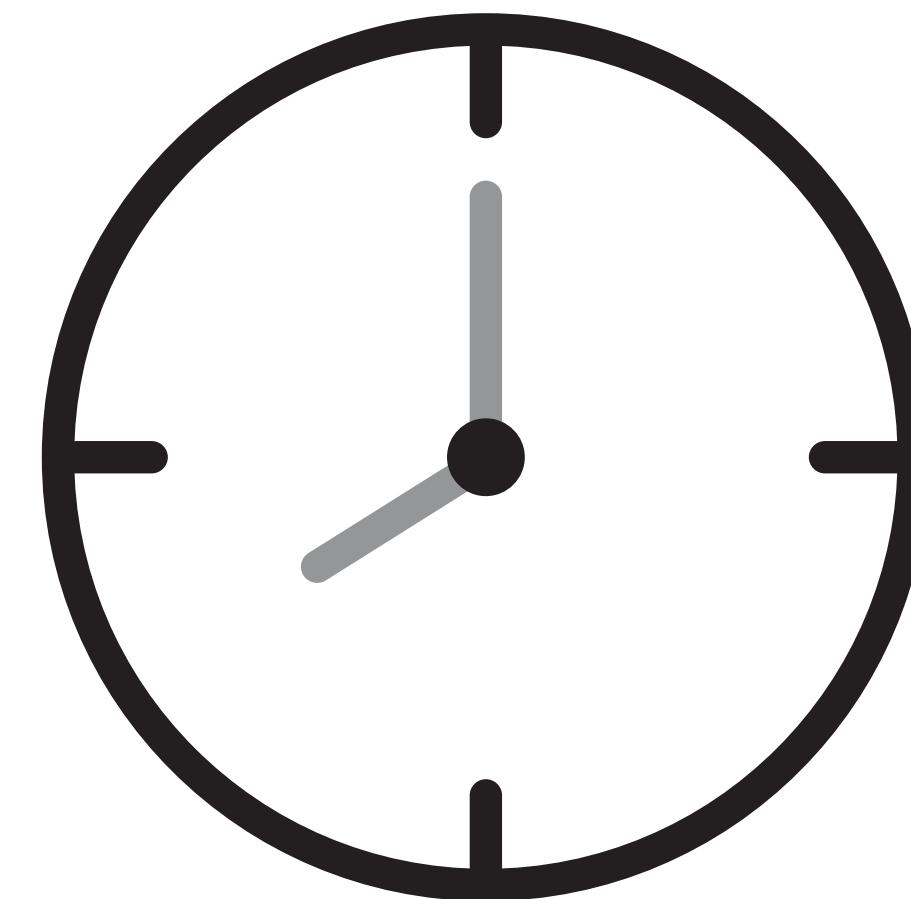


How have rates of evolution varied (or not)
across the tree?

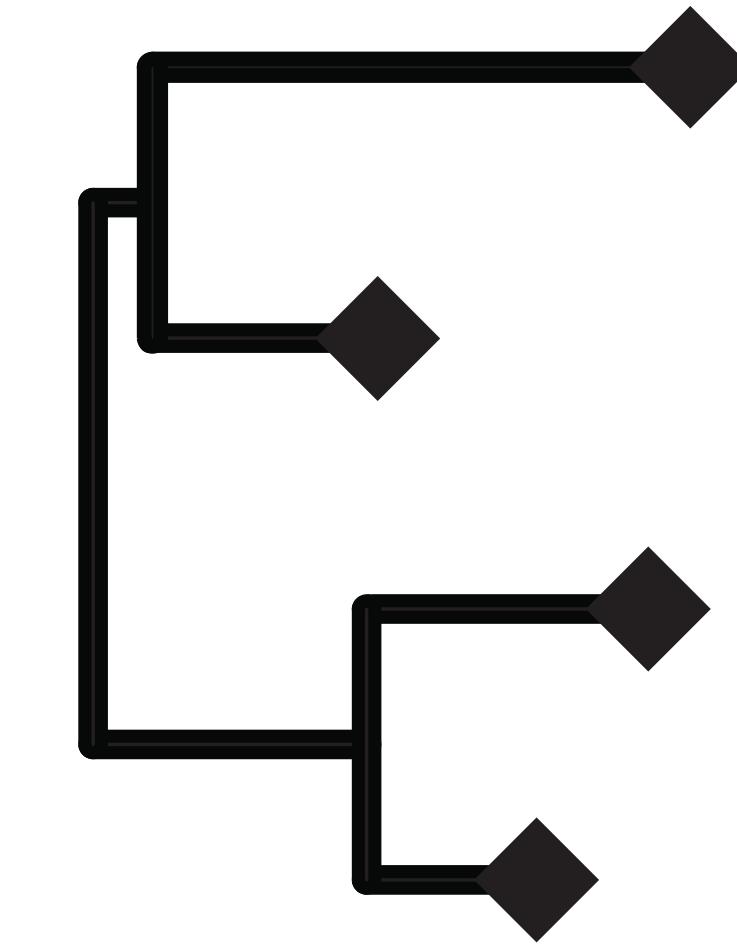
substitution model



clock model



tree model



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

Note: the tree model is often referred to as the **tree prior** even though the fossil sampling times are also data. See [May & Rothfels 2023](#)

Bayesian divergence time estimation

posterior

$$P(E \mid \lambda, \mu, \psi, p, O, t \mid 0101\dots, 1101\dots, 0100\dots, \text{snail}) =$$

likelihood

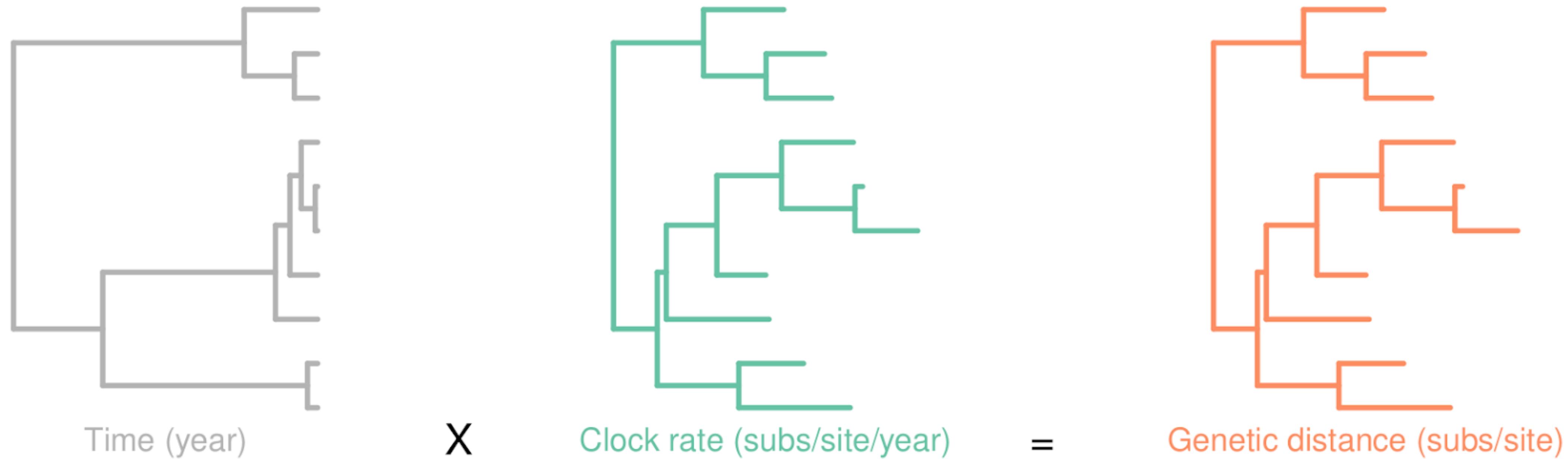
probability of the
time tree

priors

$$P(0101\dots, 1101\dots, 0100\dots \mid E) P(E \mid \lambda, \mu, \psi, p, O, t) = P(O \mid \lambda, \mu, \psi, p) P(O \mid t) P(t)$$

marginal pr of the data

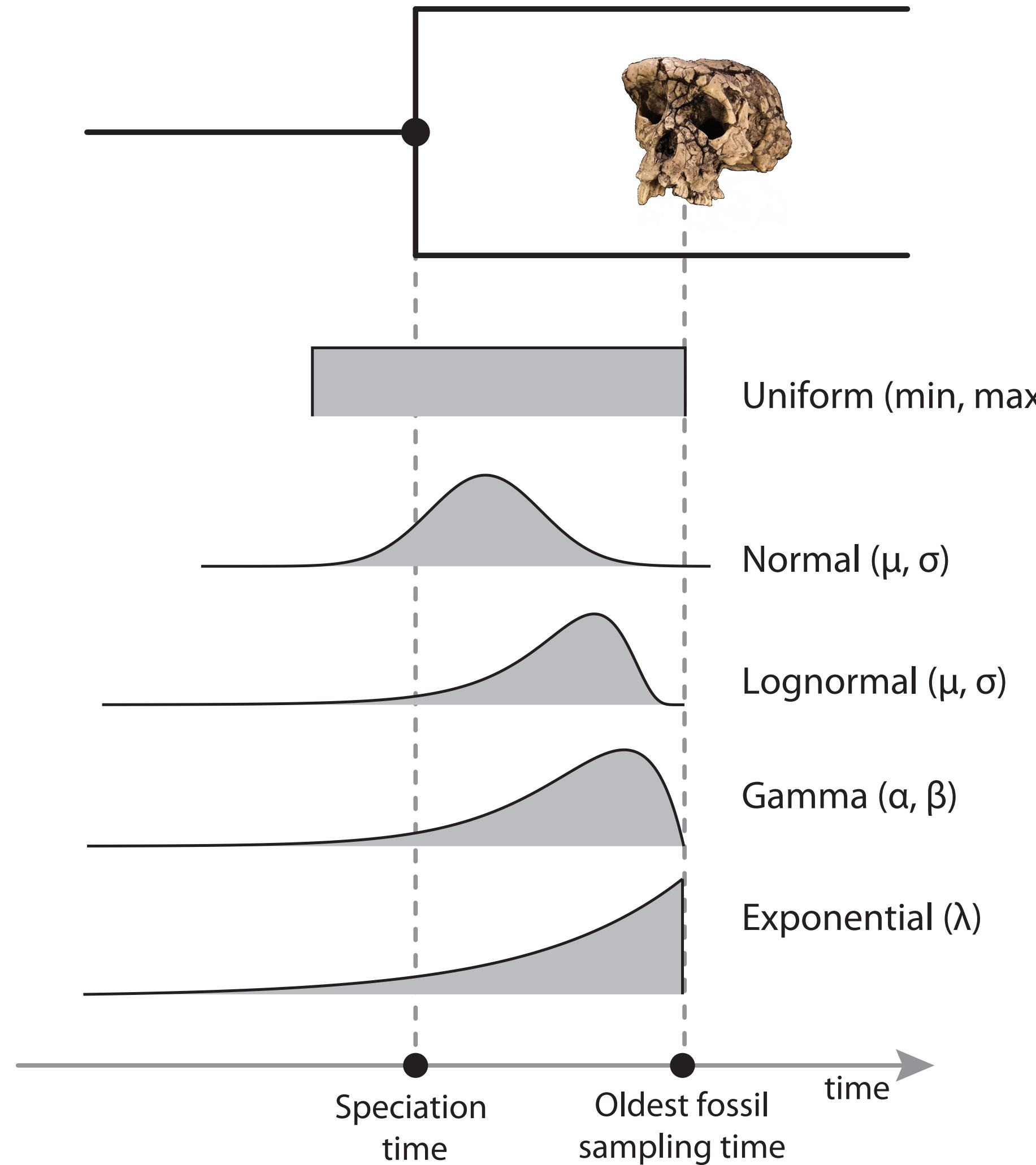
Calculating the likelihood



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



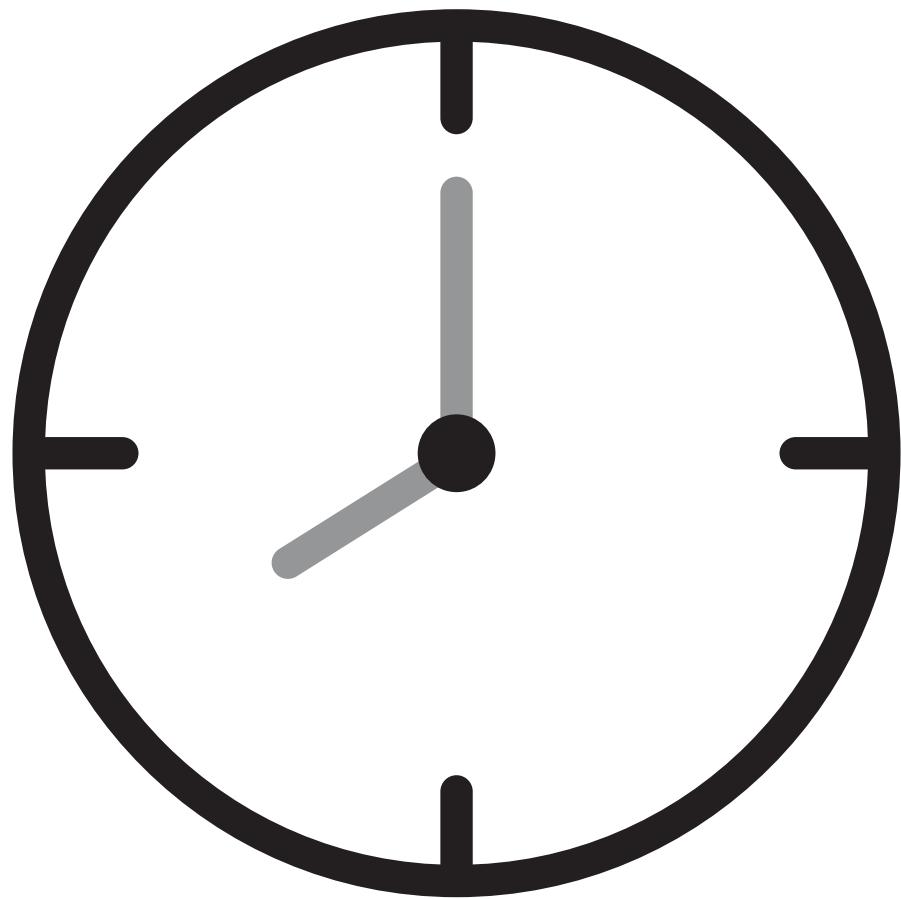
We can use a **calibration density** to constrain internal node ages

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

Adapted from Heath 2012. Sys Bio

clock model

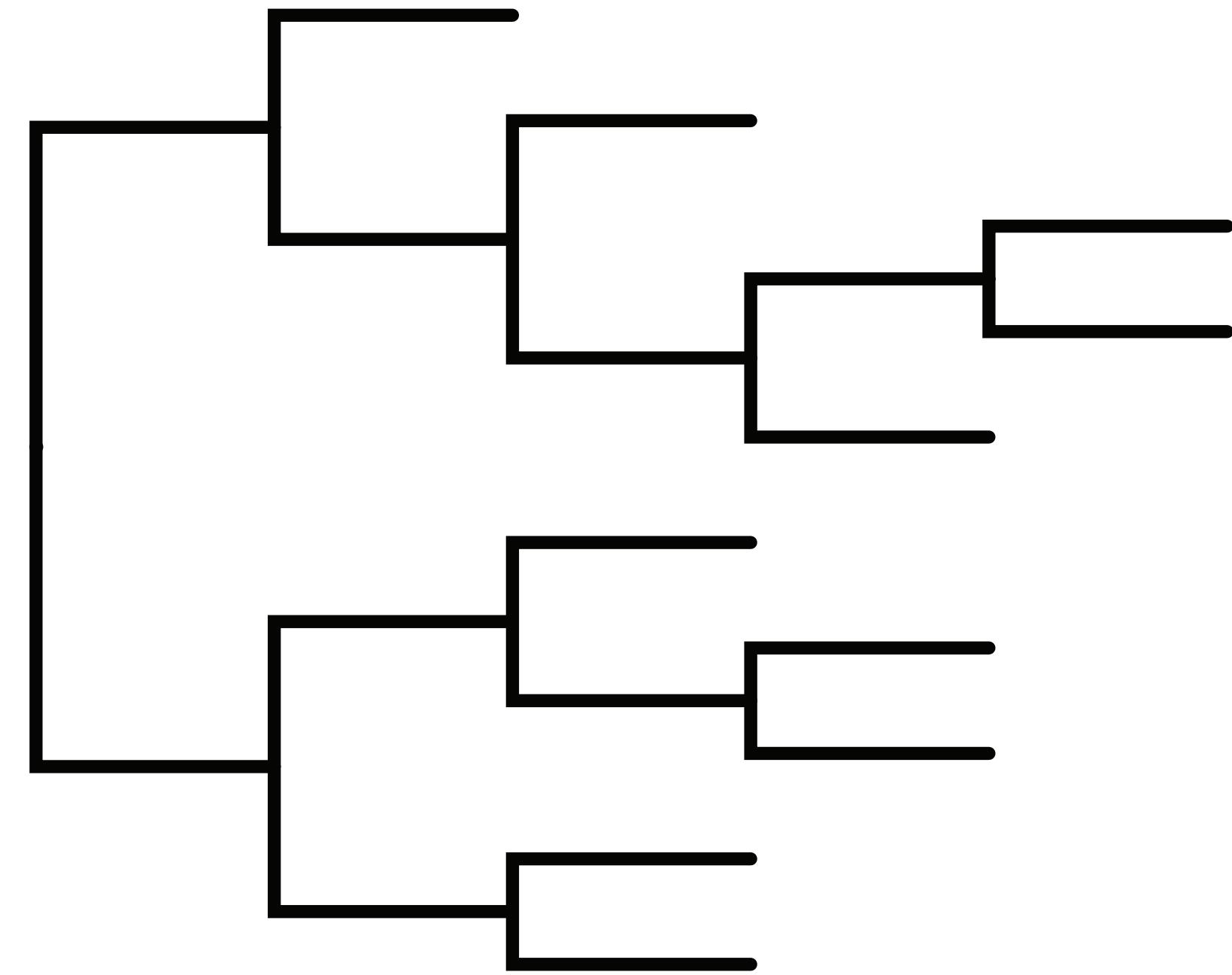
The clock model describes how evolutionary rates vary (or not) across the tree



The strict / constant molecular clock model

Assumptions

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



branch length = substitution rate

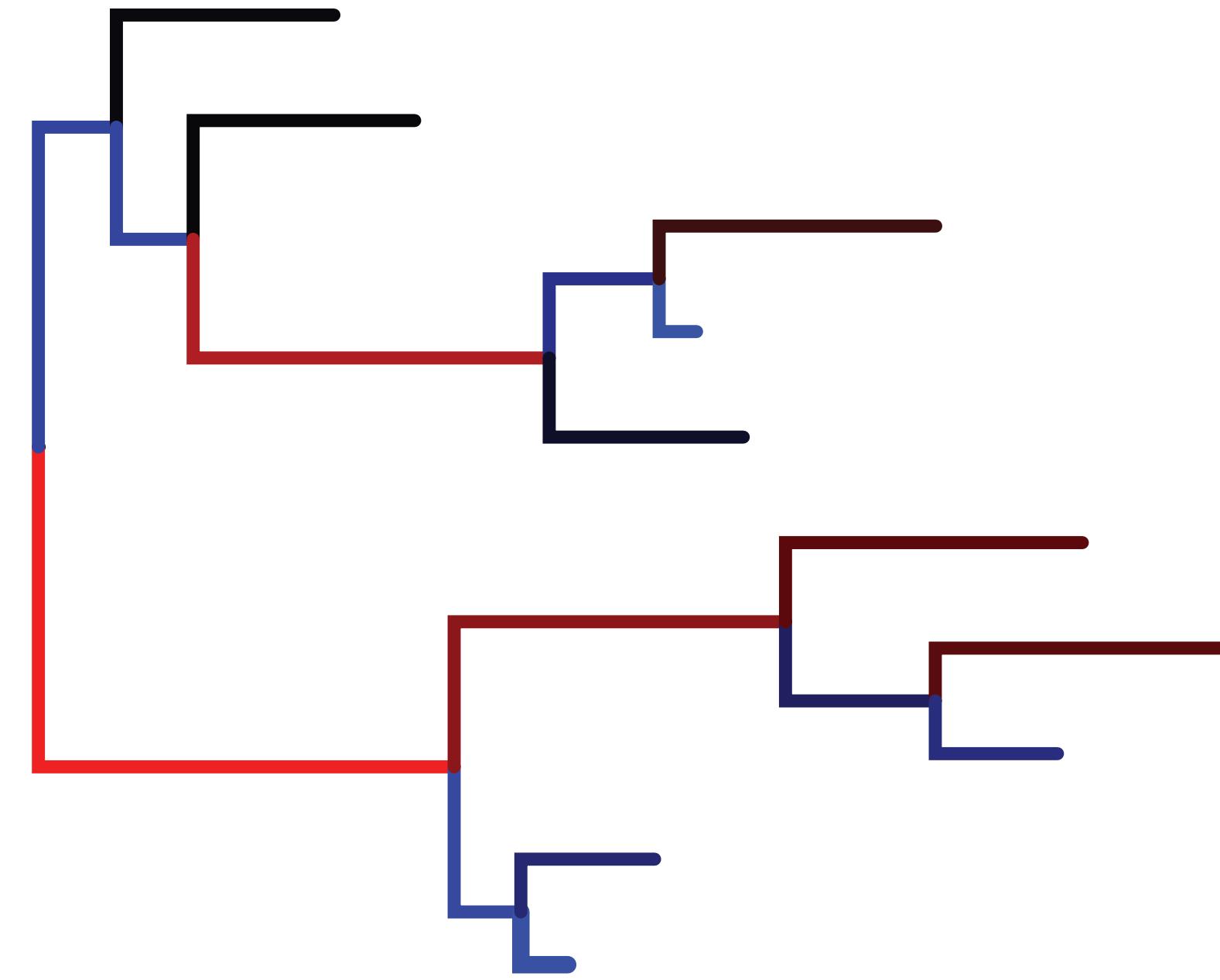


low high

Relaxed clock models

Assumptions

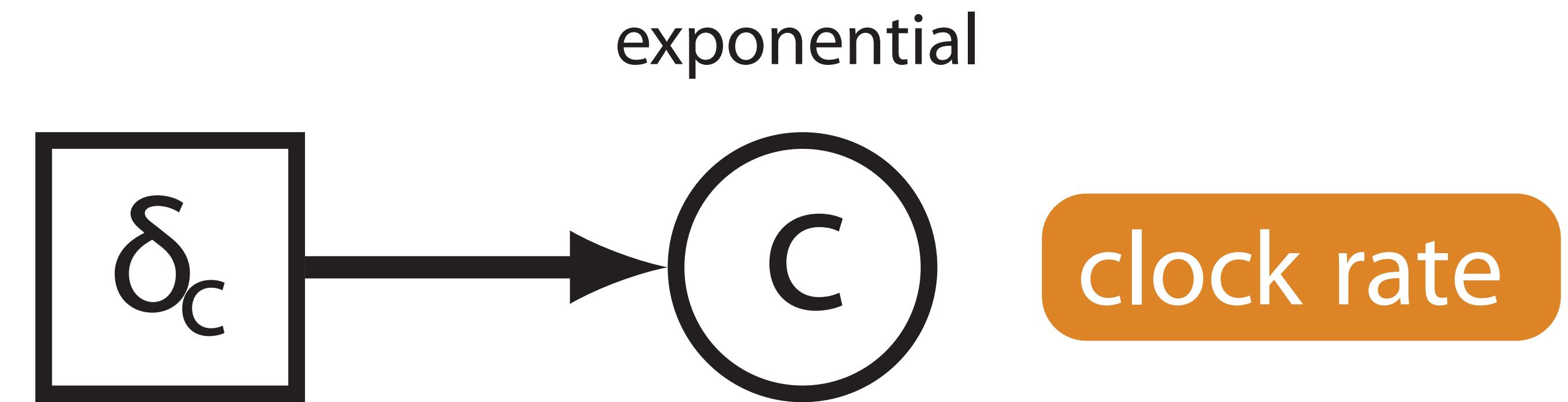
- Lineage-specific rates
- The rate assigned to each branch is drawn from some underlying distribution



branch length = substitution rate
low  high

Graphical models: strict clock model

- a) Constant node
- b) Stochastic node
- c) Deterministic node
- d) Clamped node
(observed)
- e) Plate



Graphical models: relaxed clock model



a) Constant node



b) Stochastic node



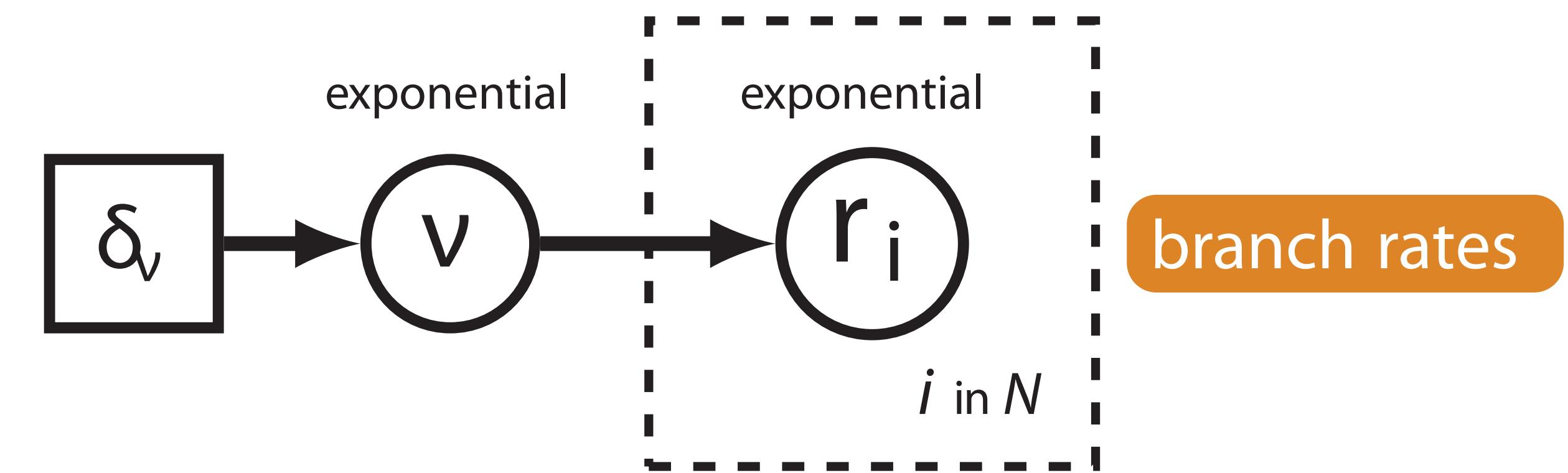
c) Deterministic node



d) Clamped node
(observed)



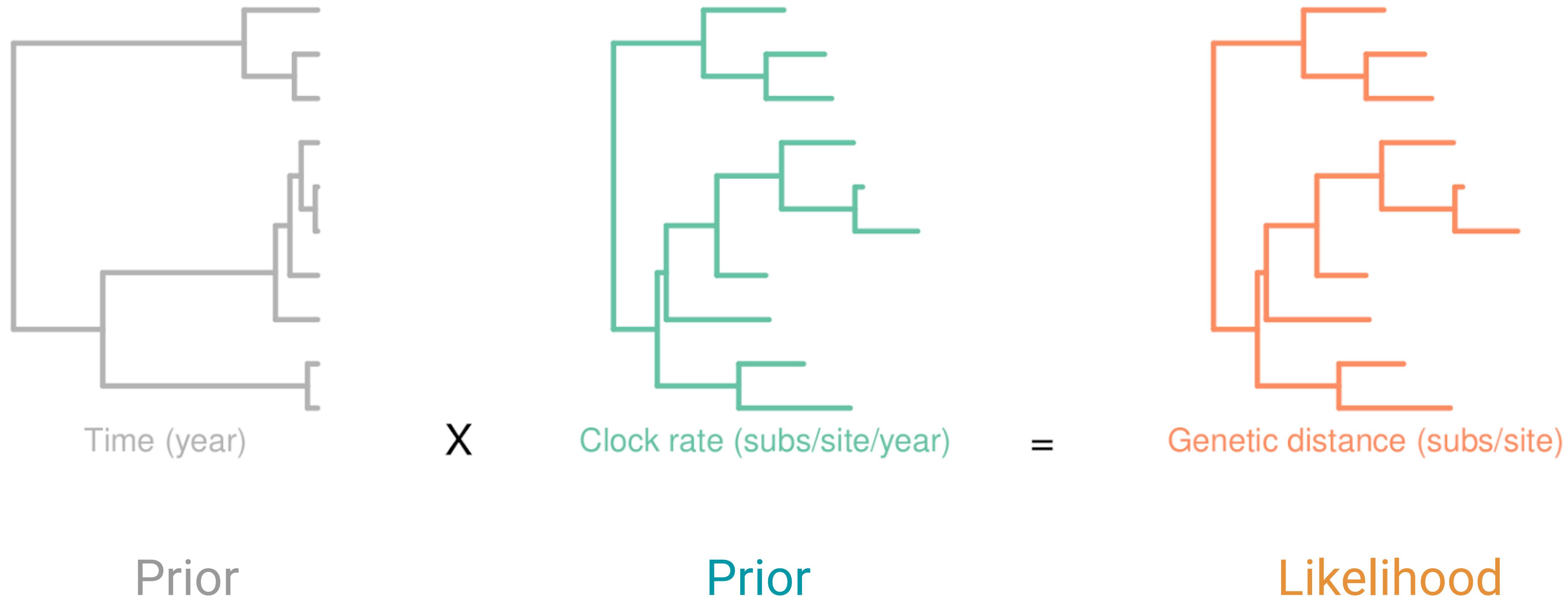
e) Plate



There are many different clock models

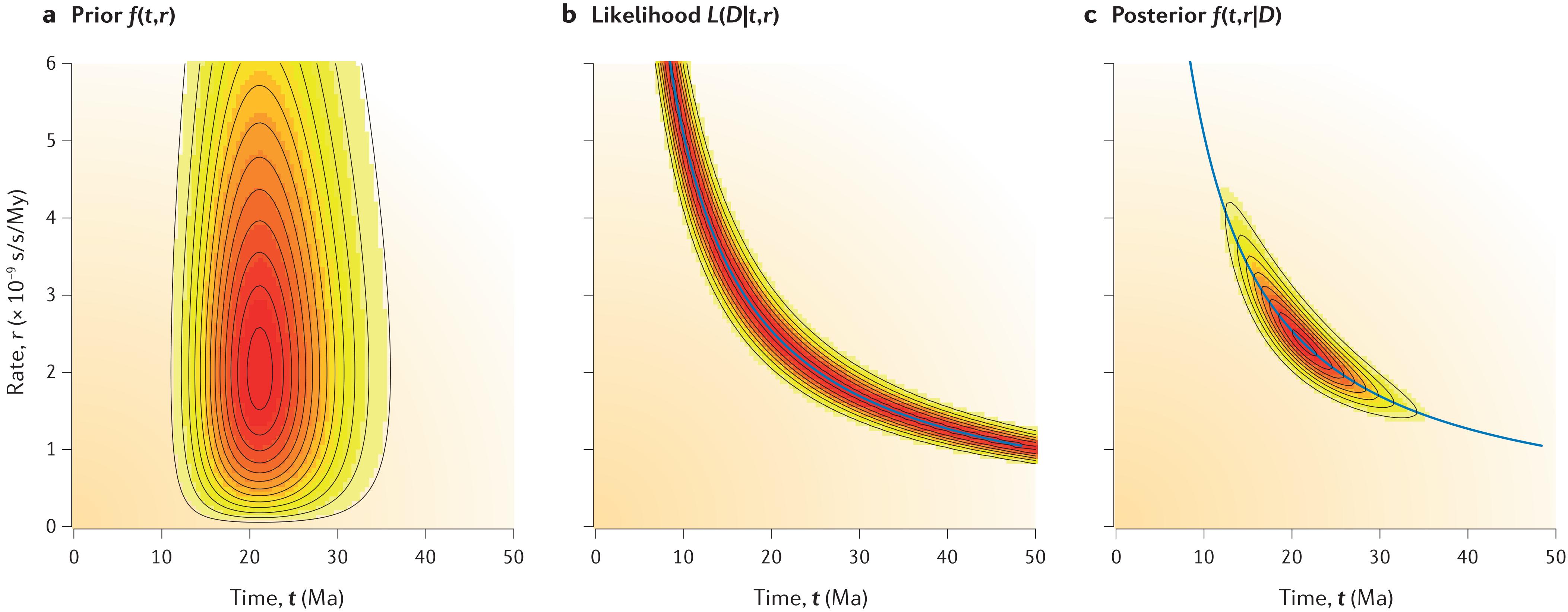
- Strict clock
- Uncorrelated or independent clock (= the favourite)
- Autocorrelated clock
- Local clocks
- Mixture models

Times and rates are not fully identifiable!



Slide adapted from Sebastian Duchene

The priors will always influence the results



Exercise