Phylogenetics

Introduction to molecular dating

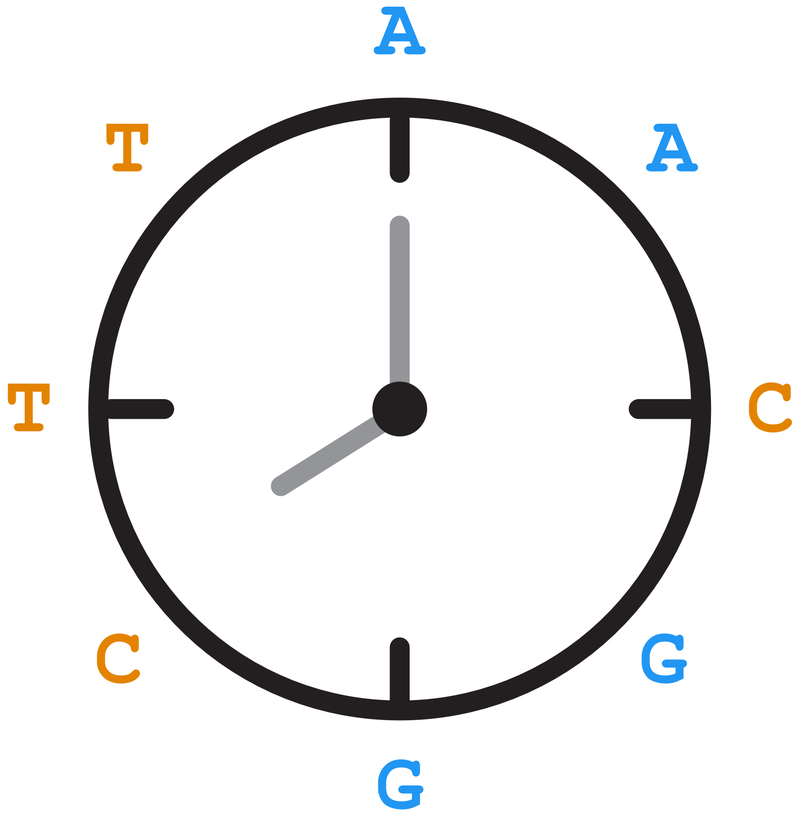
RL-V3 MPP

Rachel Warnock

07.05.24

Today’s objectives

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



Homework

Install the software software FigTree and Tracer

Phylogenetics primer part 2: tree likelihood and rate heterogeneity
Paul Lewis

→ See the question guide

Q&A Phylogenetics primer part 2 by *Paul Lewis*

(the answers provided here are my interpretation of these concepts — answers may vary!)

1. What are the assumptions of the following substitution models? Consider the rate of change between **character states** and the **state frequencies**.

JC →

HKY →

GTR →

1. What are the assumptions of the following substitution models? Consider the rate of change between **character states** and the **state frequencies**.

JC → equal frequencies, equal rates

HKY → unequal frequencies, unequal rates between transversions & transitions

GTR → unequal frequencies, unequal rates

transversions: A ↔ T or G ↔ C, transitions: A ↔ G or C ↔ T

2. Can you briefly describe the following approaches to account for rate variation among characters?

2a. Site specific rates

→ assign sites to separate partitions and allow each partition to have its own set of parameters

2b. Invariant sites model

→ assign a subset of sites to a “constant” (i.e., non-variable) category

2. Can you briefly describe the following approaches to account for rate variation among characters?

2c. Discrete Gamma model

→ calculate the likelihood assuming there are discrete rate categories, e.g., 4. Variation in rate categories is represented by a gamma distribution and the parameters of the gamma distribution are calculated from the data

3. For a given substitution model, what do the values in the *Q* matrix and the *P* matrix represent?

→ the *Q* matrix is the instantaneous rate matrix, i.e., the instantaneous rates of change between each character state combination

→ the *P* matrix is the transition probability matrix, i.e., the probabilities of change between character states after relative time *t*, which is represented by the branch lengths

Recap

How do we find the ‘best’ tree?

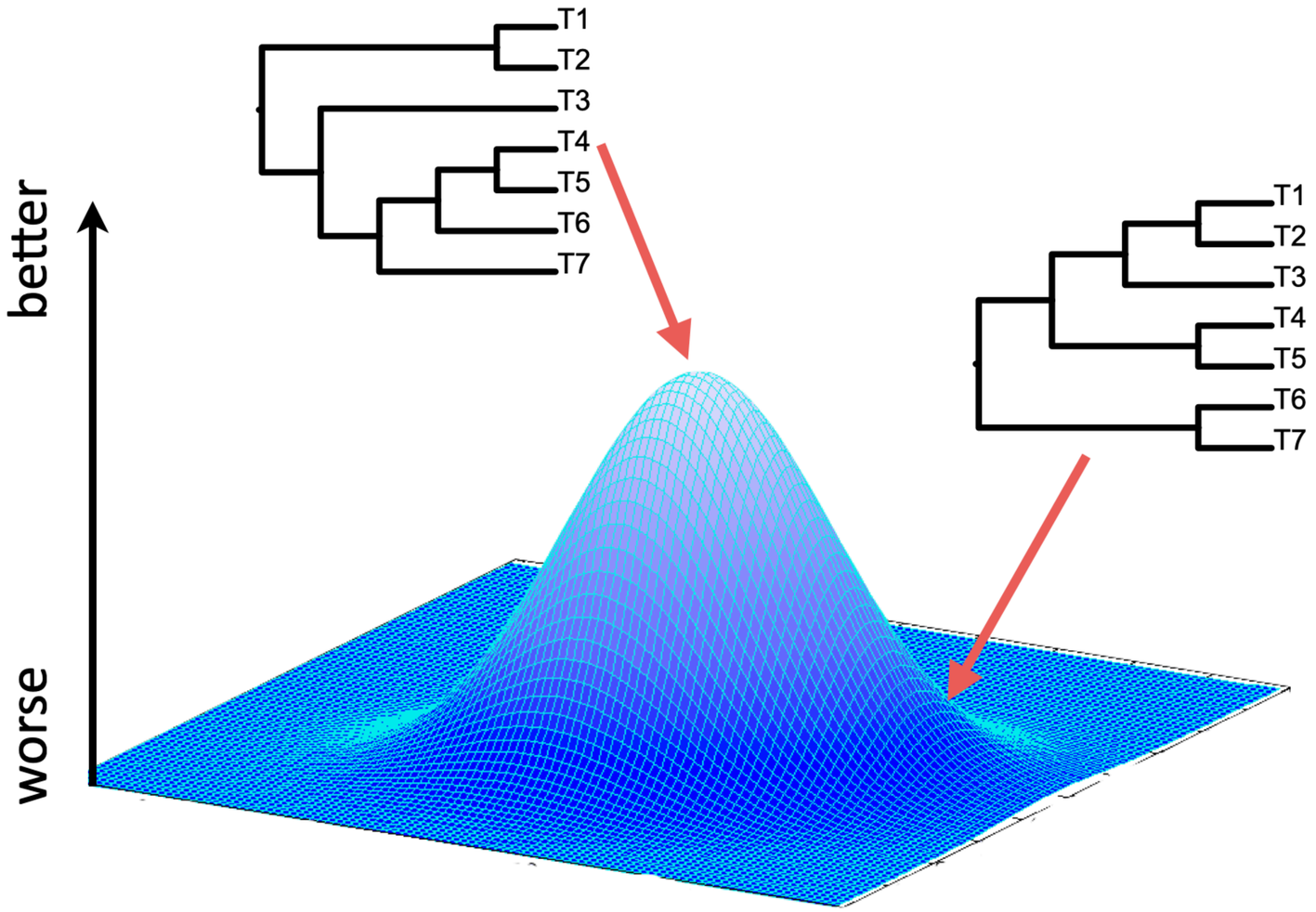


Image source Tracy Heath

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

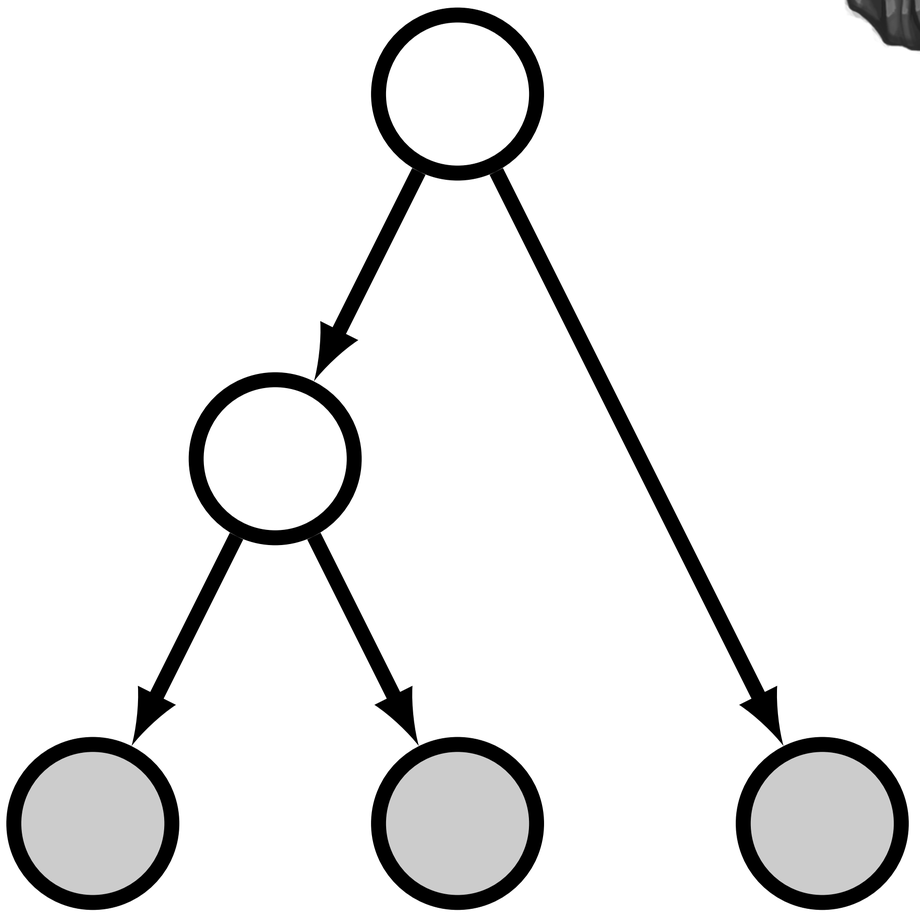
Both maximum likelihood and Bayesian inference are model-based approaches

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

Graphical models

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

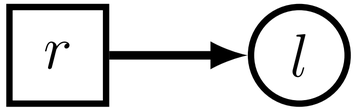
Depict the conditional dependence structure of parameters and other random variables



a)

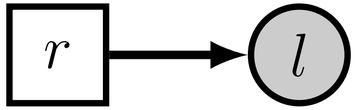

```
# constant node
r <- 10
```

b)



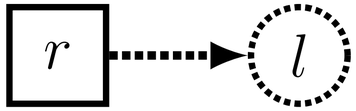

```
# stochastic node
l ~ dnExp(r)
```

c)



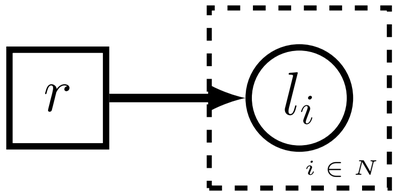

```
# stochastic node (observed)
l.clamp(0.1)
```

d)




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

e)




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

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

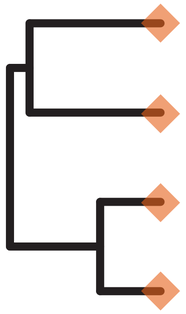
$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model})\,\Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Components used to infer trees

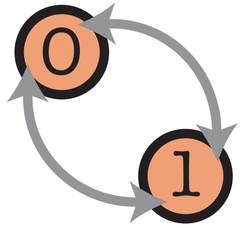
without considering time

0101...
1101...
0100...

data
sequences or characters

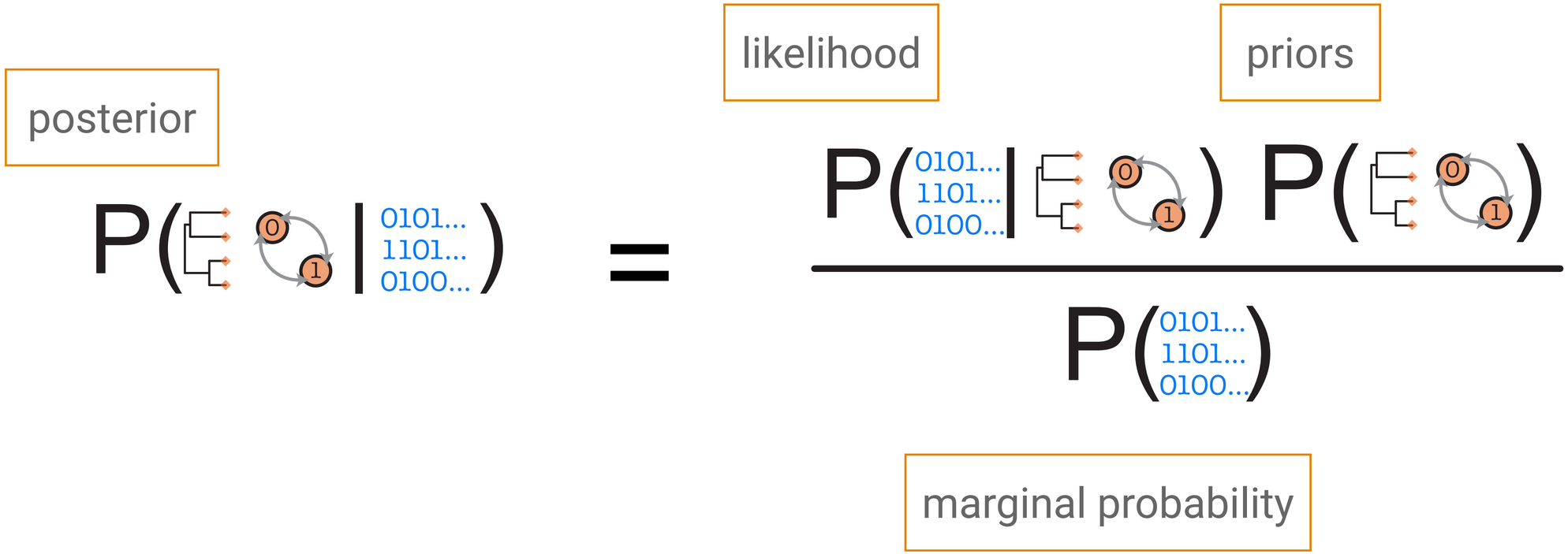


tree
topology and branch lengths

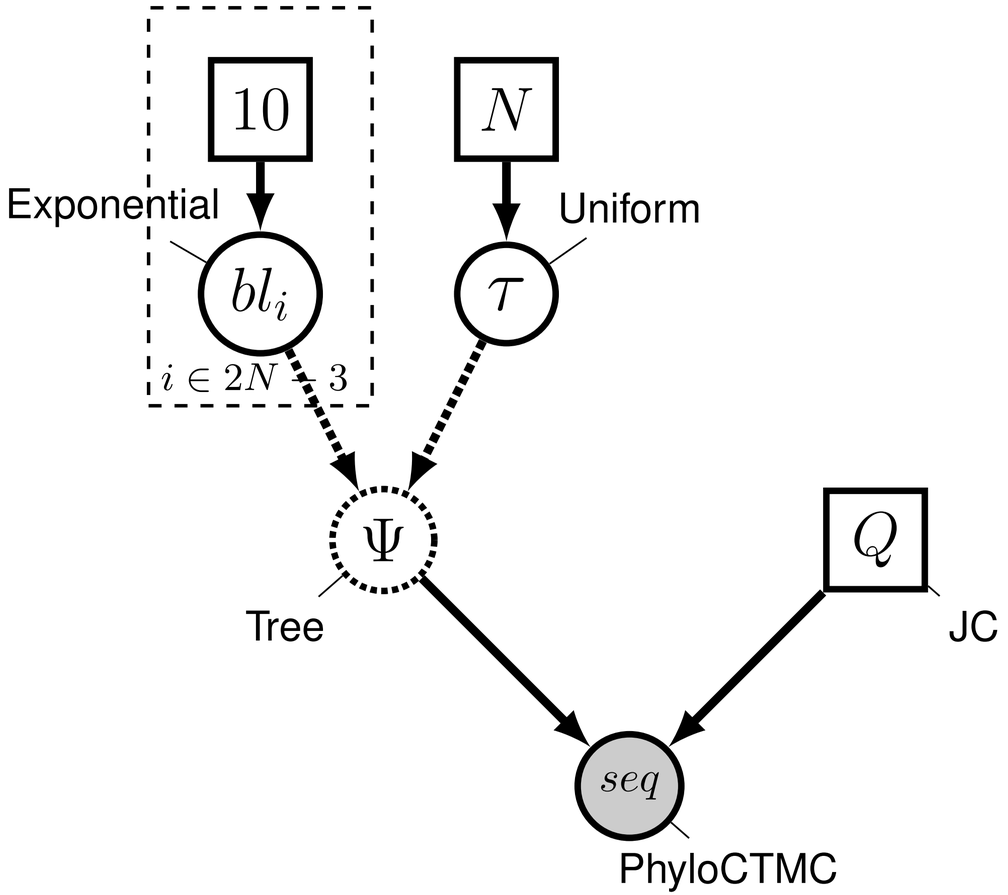


substitution model

Bayesian tree inference

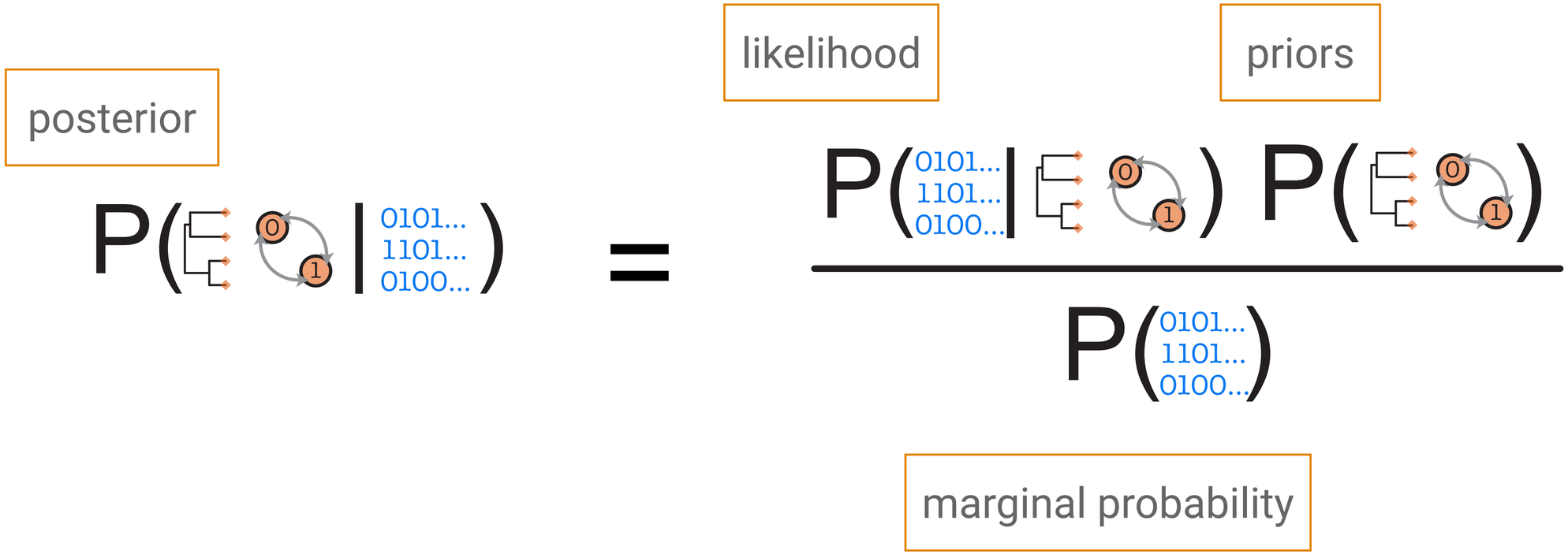


$

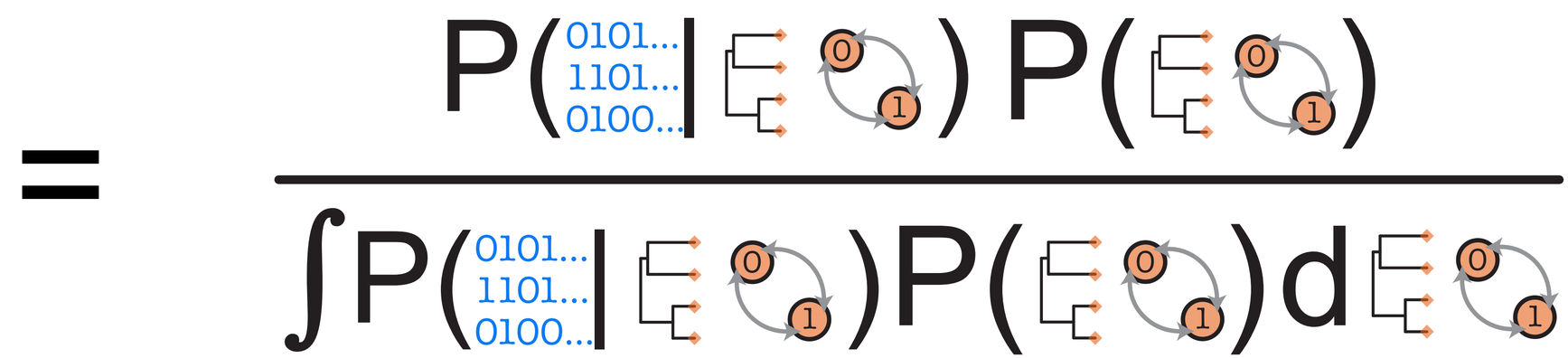



```
for (I in 1:n_branches) {
        bl[I] ~ dnExponential(10.0)
}
topology ~ dnUniformTopology(taxa)
psi := treeAssembly(topology, bl)


Q_morpho <- fnJC(2)


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

phyMorpho.clamp( data )
```

Bayesian tree inference

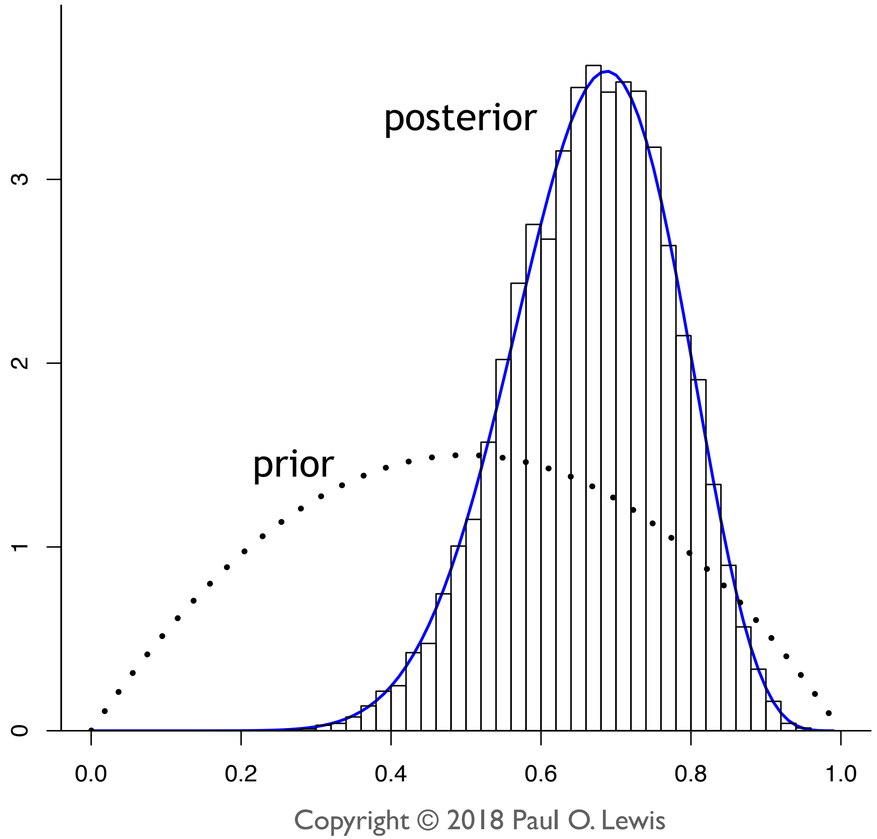


Bayesian tree inference



this part is incredibly difficult to calculate!

What is Markov chain Monte Carlo (MCMC)?



The aim is to produce a histogram that provides a good approximation of the posterior

Hastings ratio

new parameter values

$$R = \frac{P(\text{tree}^{*} \mid \text{data})}{P(\text{tree} \mid \text{data})}$$

$$= \frac{P(\text{data} \mid \text{tree}^{*})P(\text{tree}^{*})}{P(\text{data})} \Big/ \frac{P(\text{data} \mid \text{tree})\,P(\text{tree})}{P(\text{data})}$$

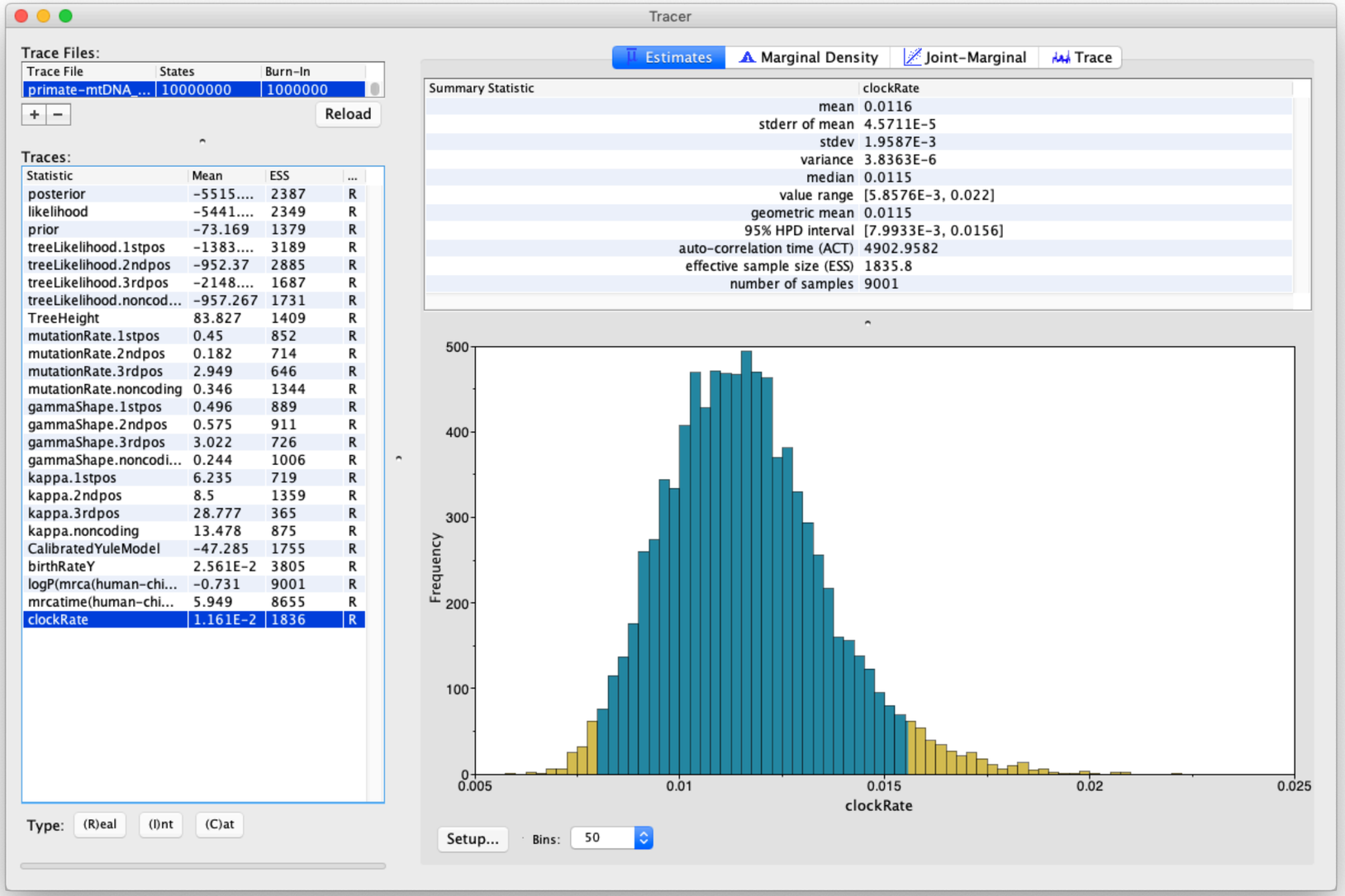
The marginal probability of the data cancels out

$$= \frac{P(\text{data} \mid \text{tree}^{*})P(\text{tree}^{*})}{P(\text{data} \mid \text{tree})\,P(\text{tree})}$$

All we're left to calculate is the likelihood ratio and the prior odds ratio

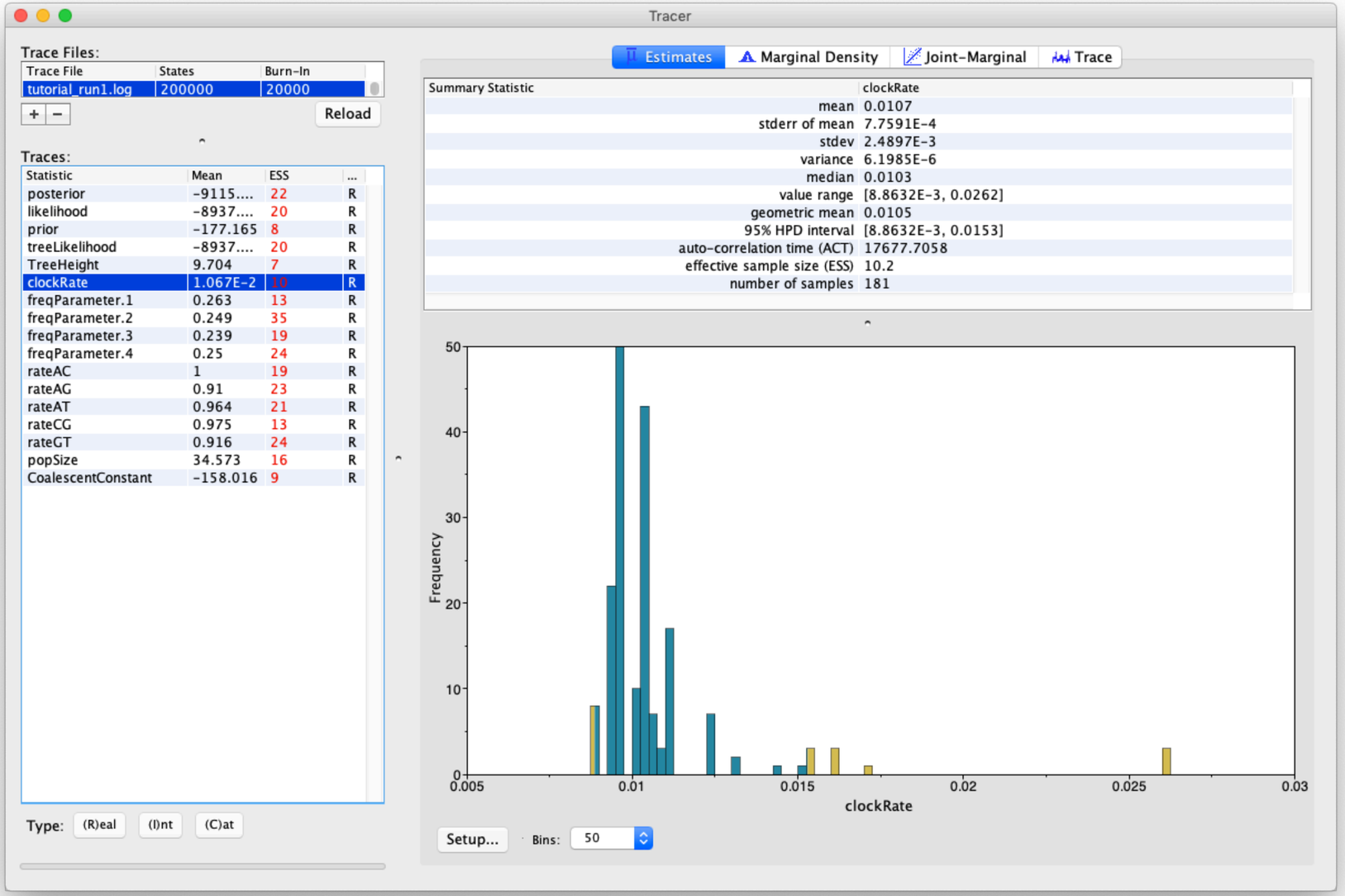
Summarising the posterior

Tracer is an amazing program for exploring MCMC output



Summarising the posterior

Tracer is an amazing program for exploring MCMC output

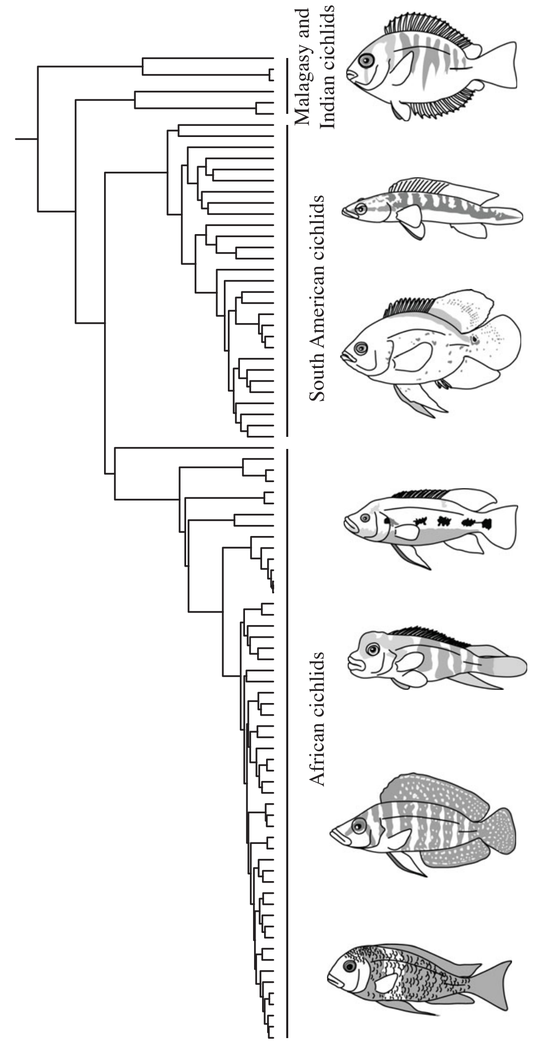


Exercise

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

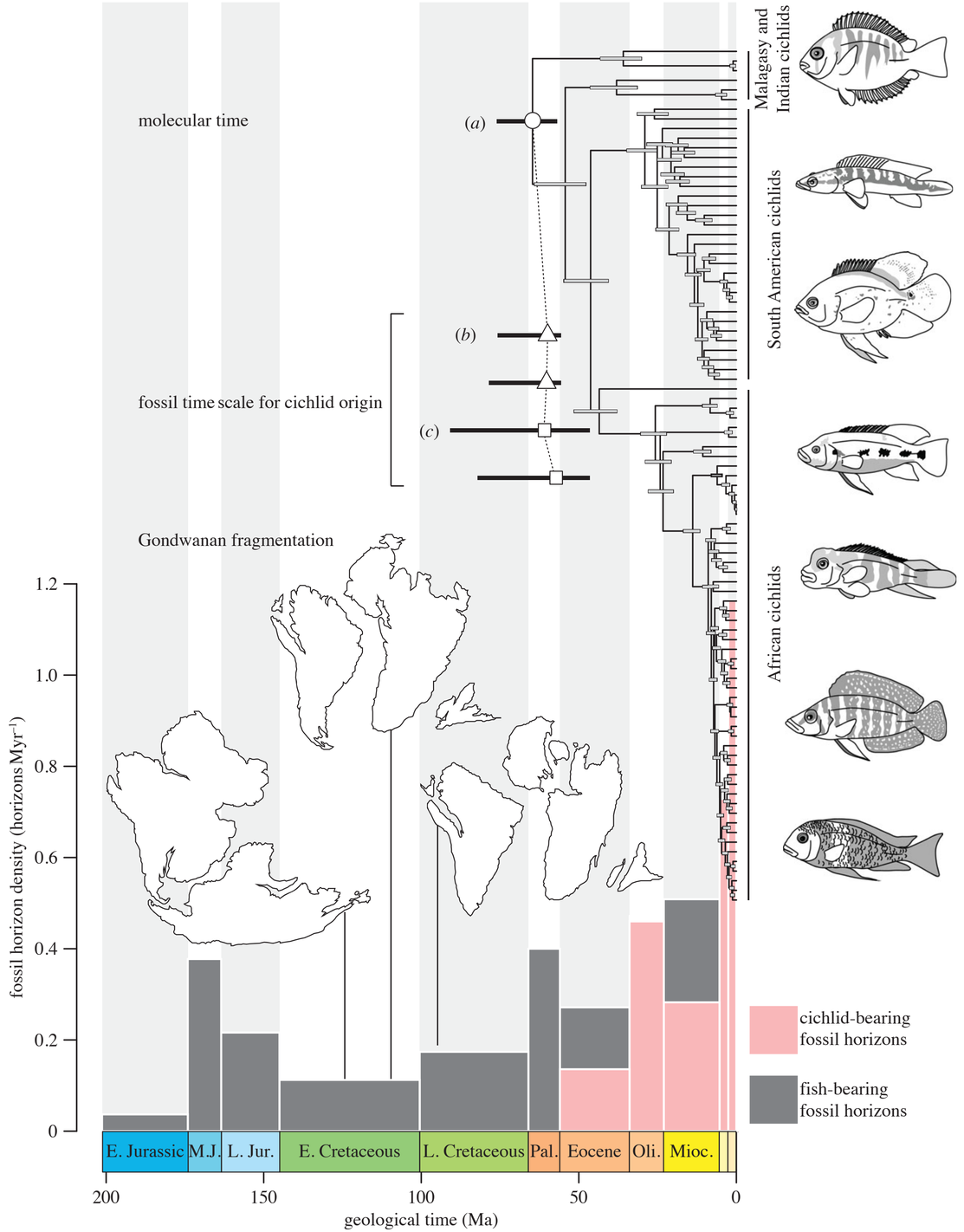
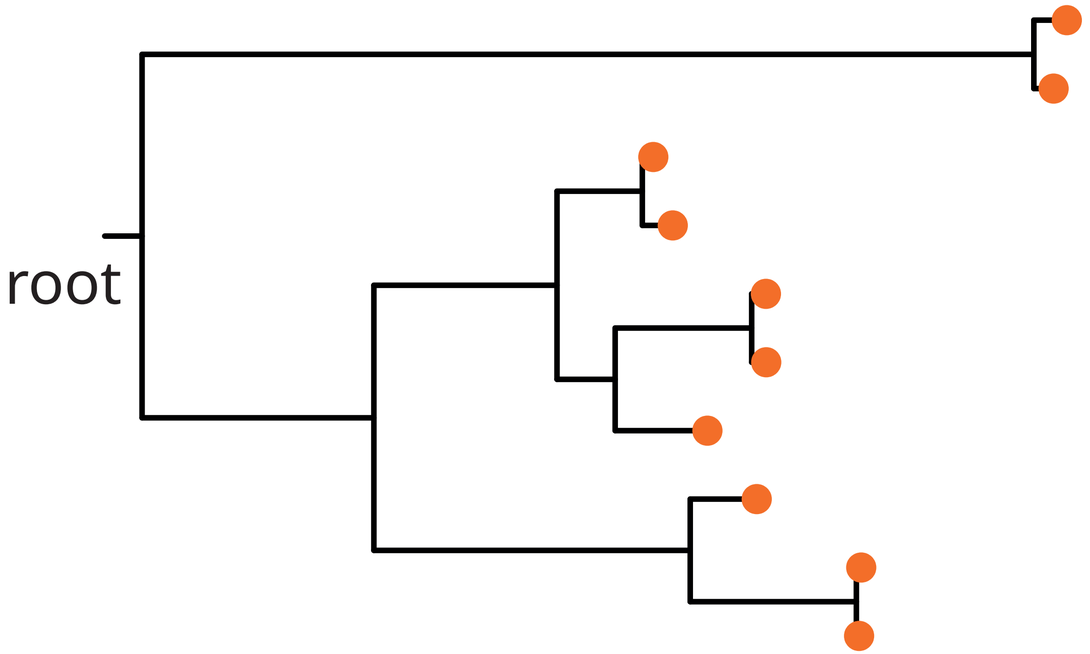


Image adapted from *Friedmann et al.* (2013)

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



ATGCATGC
TTGCCTGC
TTGCATCG
ATGCATCG
ATGCATGG
TTGCCTGG
TAGCGTGC
TAGCGAGC

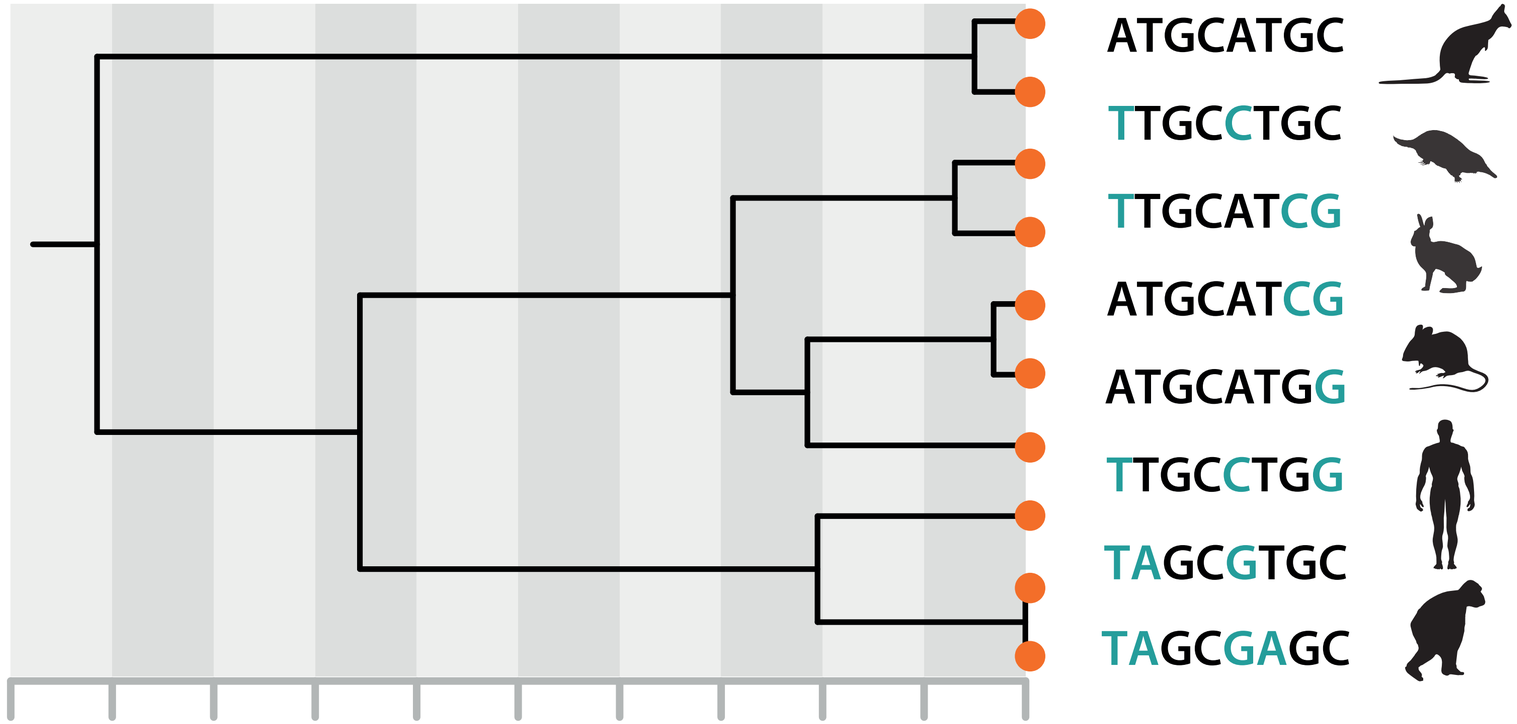
Slow rate, long interval **or** fast rate, short interval?

Goal: to disentangle evolutionary rate and time

branch lengths = genetic distance

$v = rt$

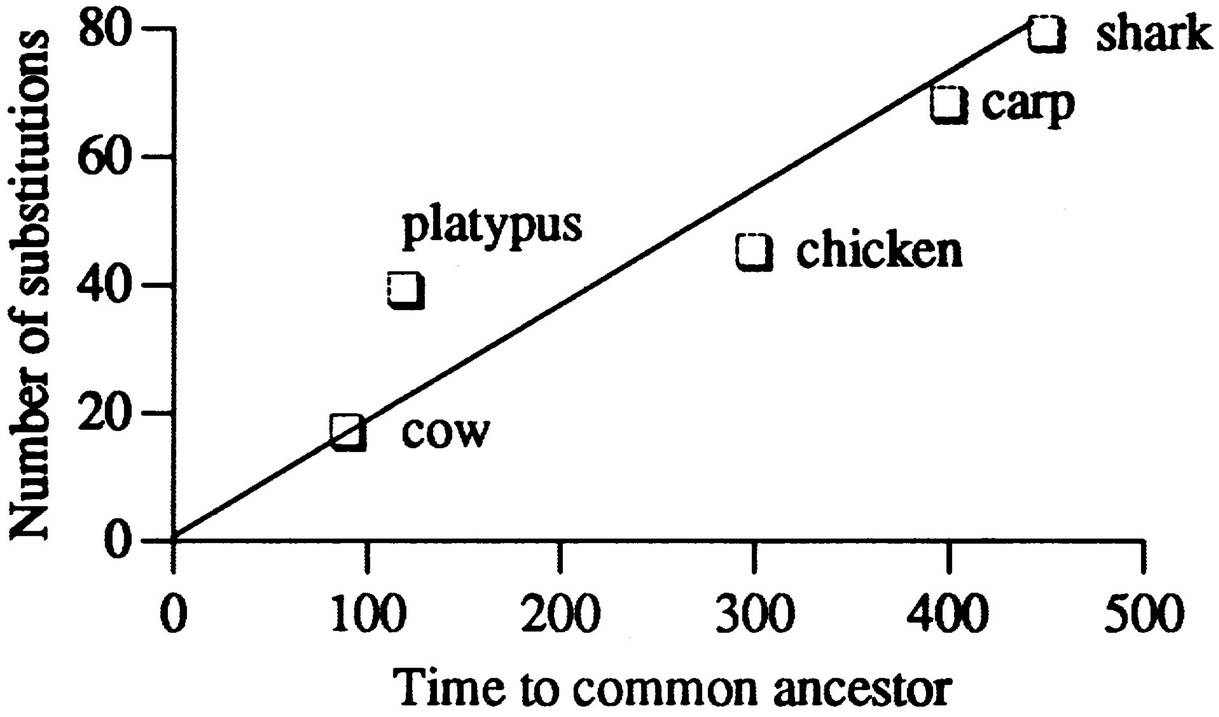
Molecular (or morphological) characters are not independently informative about time



branch lengths = time

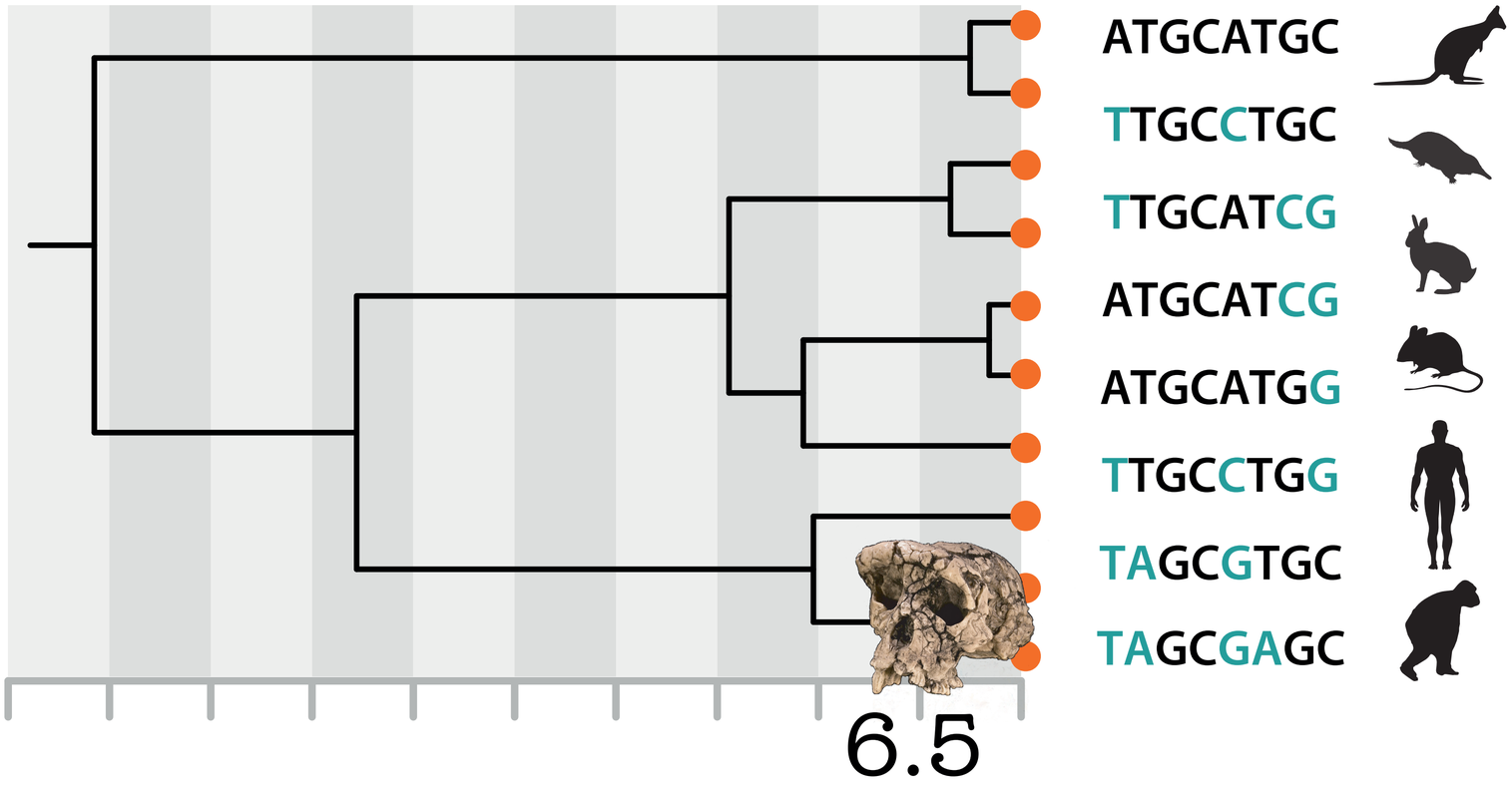
Goal: to disentangle evolutionary rate and time

The molecular clock hypothesis



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

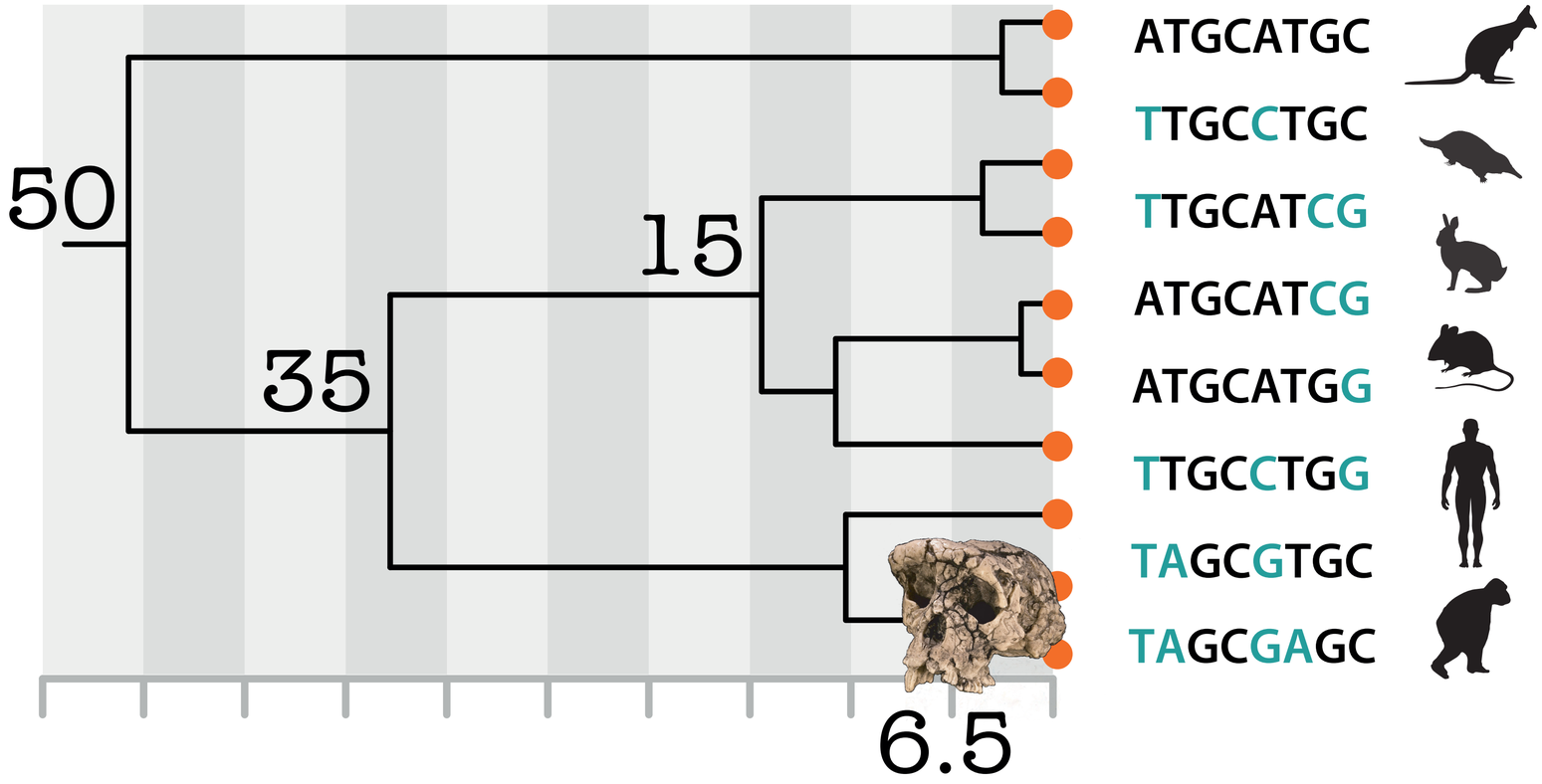
Calibrating the substitution rate



branch lengths = time

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

Calibrating the substitution rate

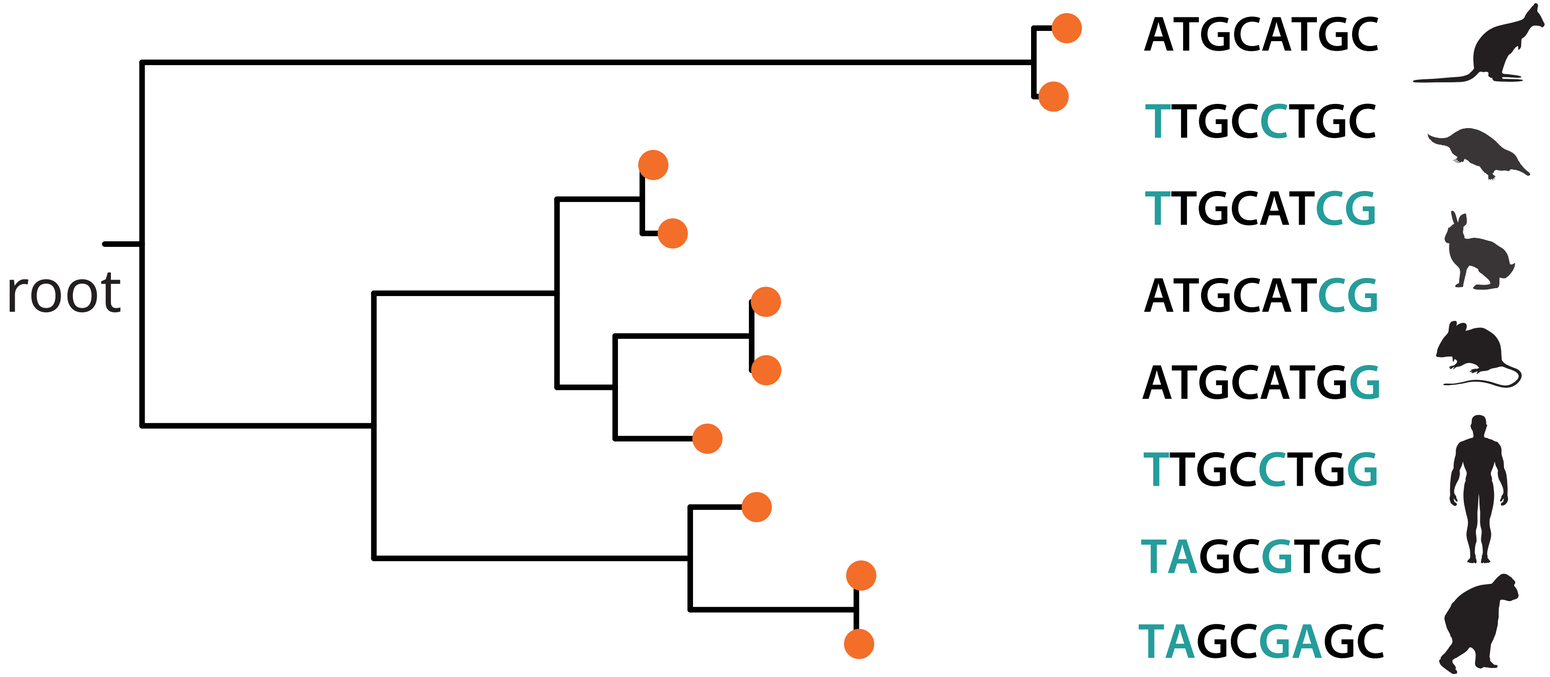


branch lengths = time

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

Molecular dating: challenges

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

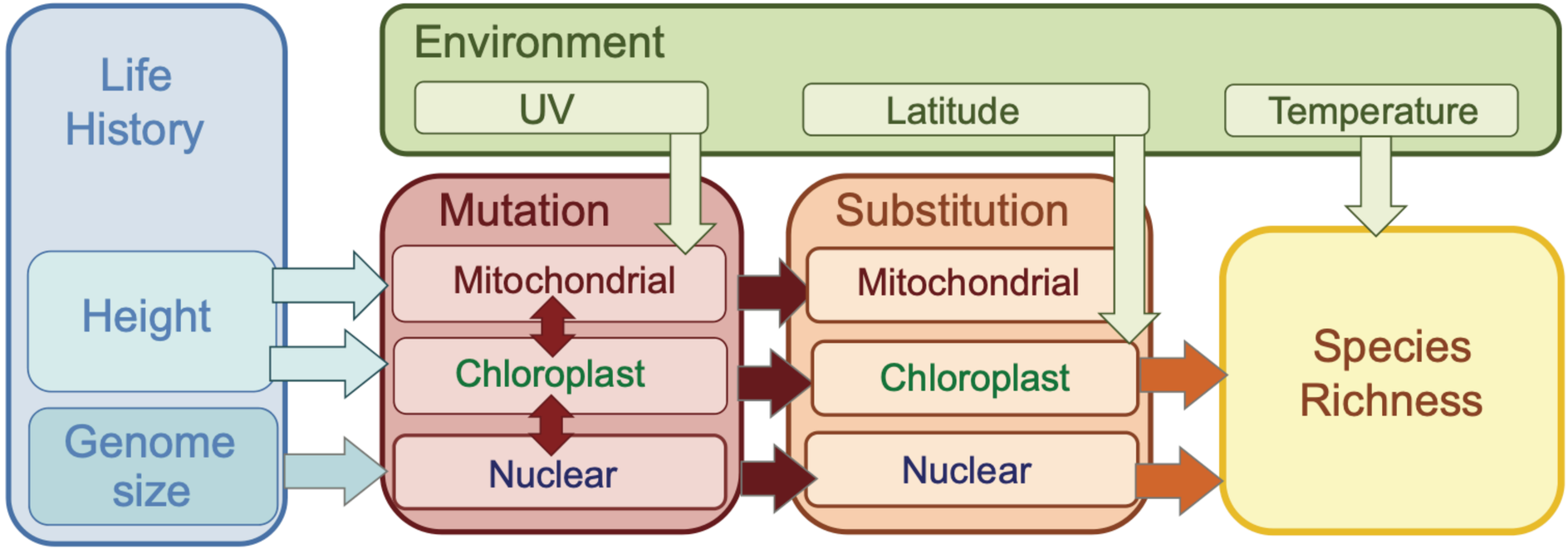


branch lengths = genetic distance

$v = rt$

Molecular dating: challenges

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



Bromham et al. (2015)

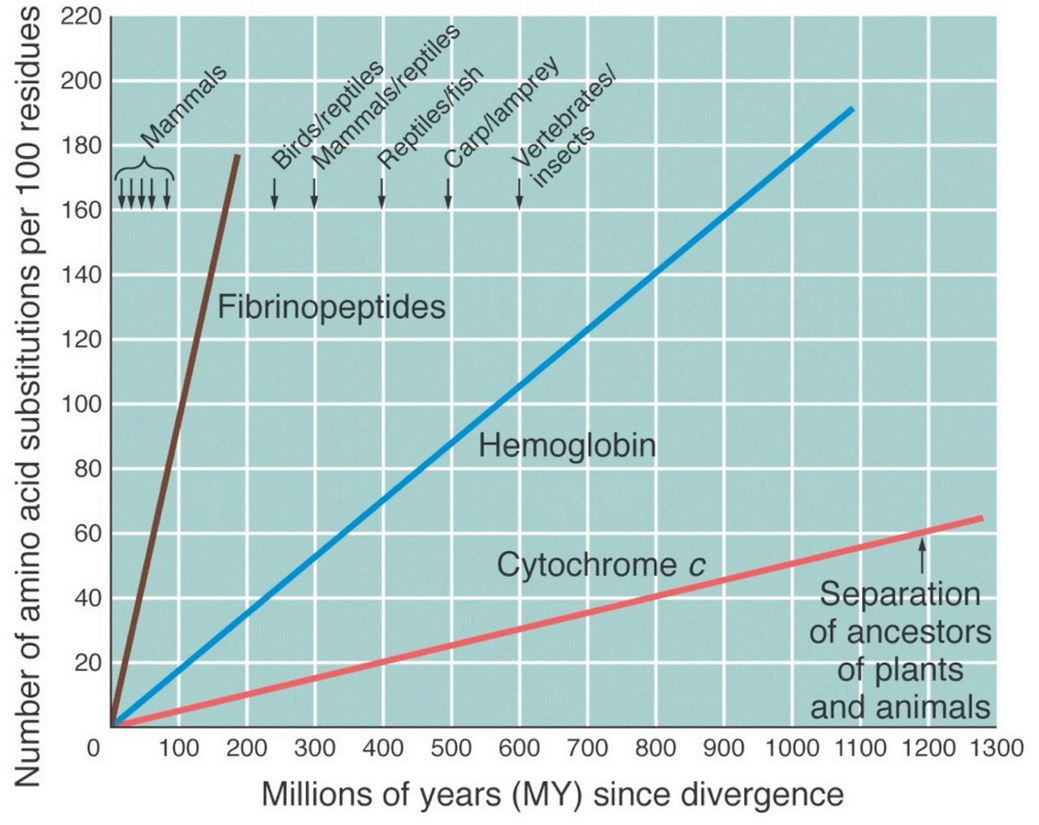
Molecular dating: challenges

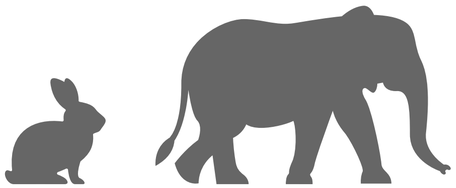
Many variables contribute to **variation in the substitution rate**

The molecular clock is not constant

Rates vary across:

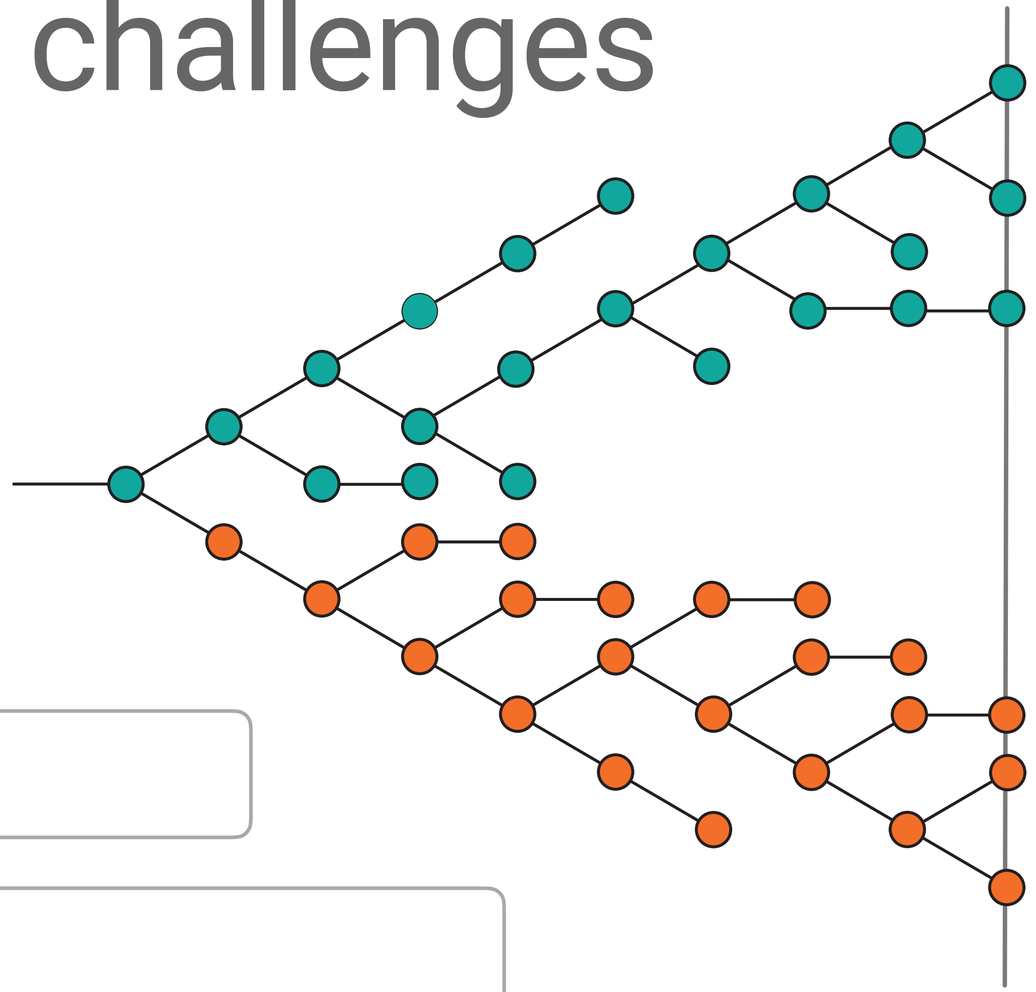
- taxa
- time
- genes
- sites within the same gene





Molecular dating: challenges

Calibrations are imprecise



Molecular evolution:

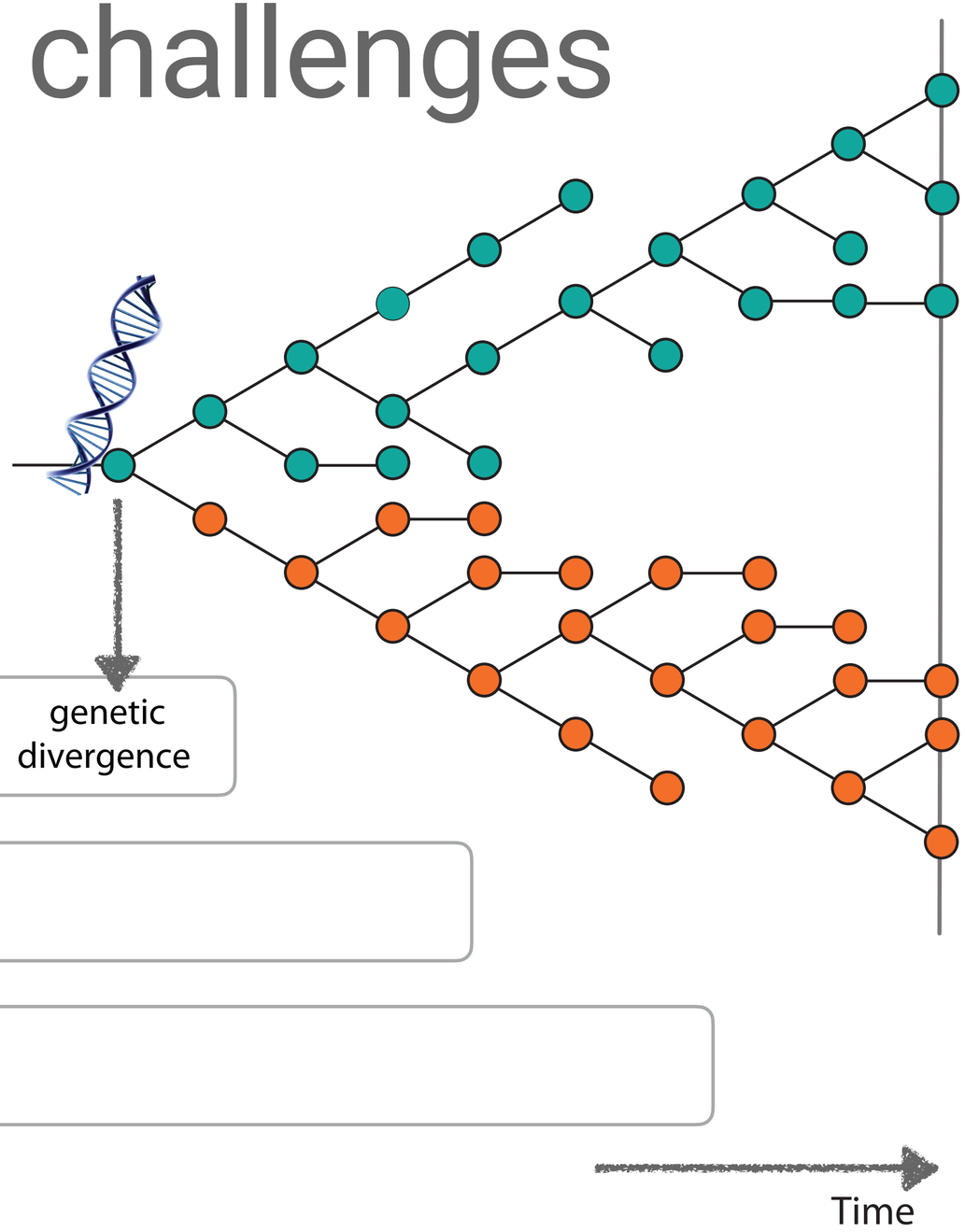
Morphological evolution:

Fossil preservation:

Time

Molecular dating: challenges

Calibrations are imprecise



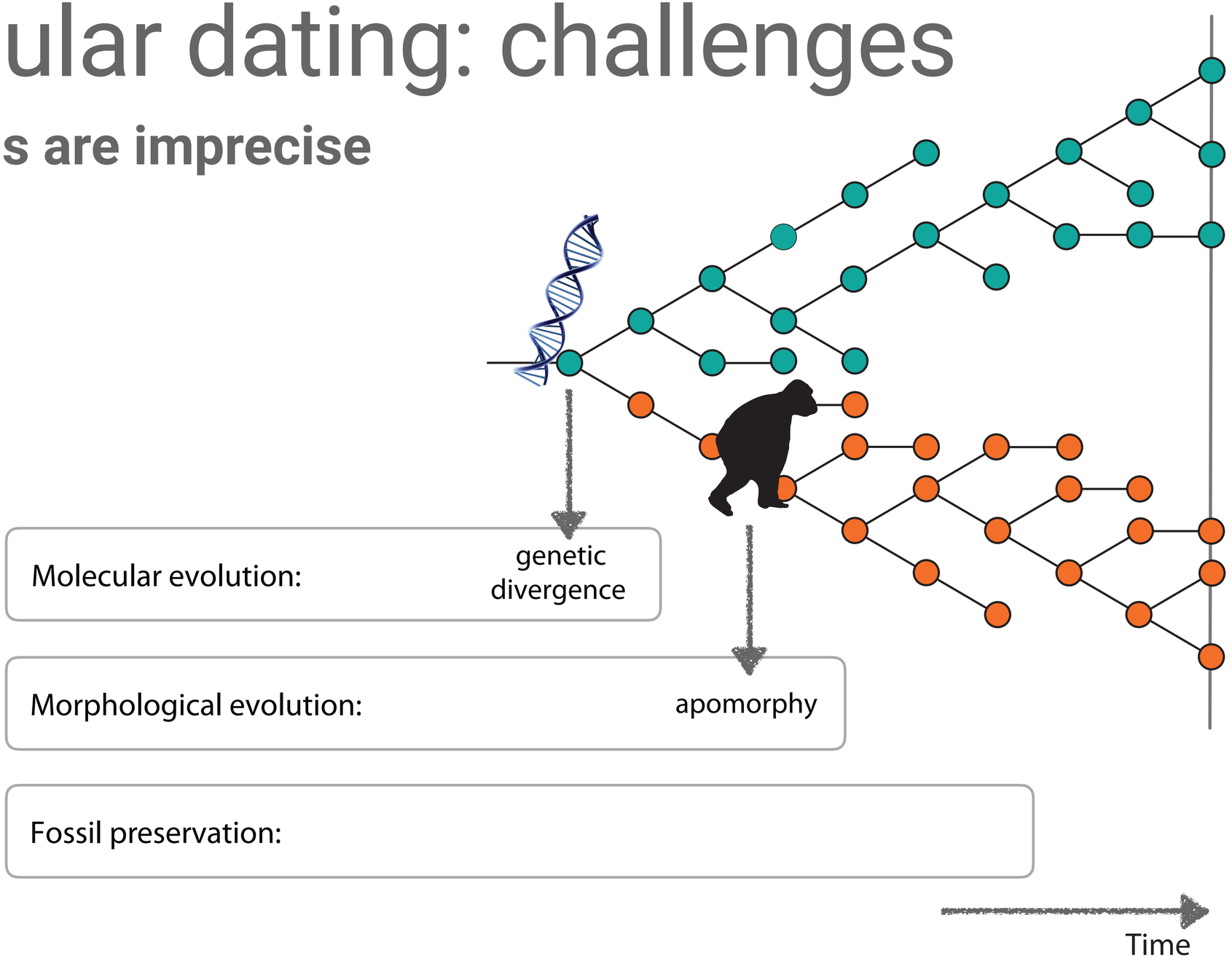
Molecular evolution:

Morphological evolution:

Fossil preservation:

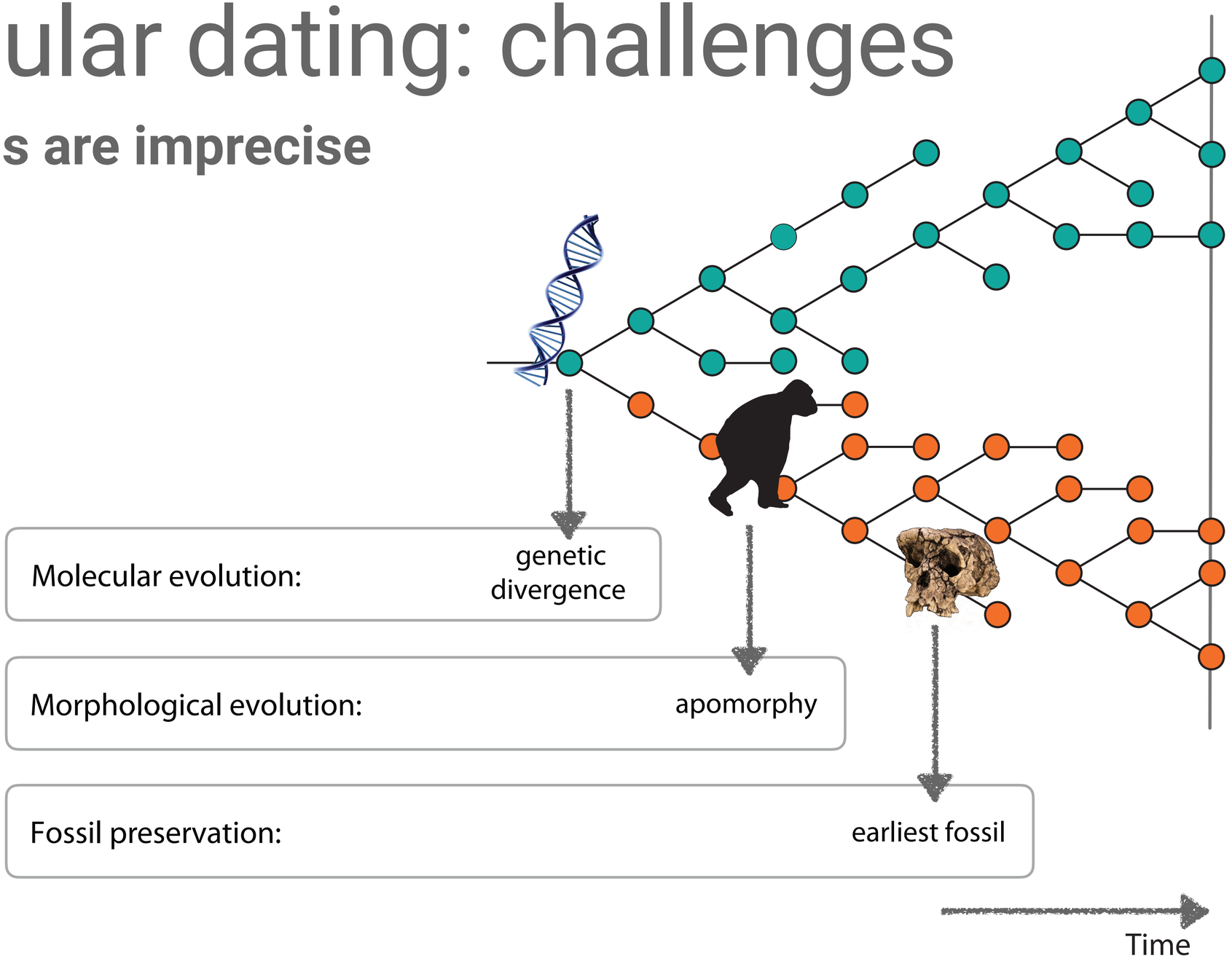
Molecular dating: challenges

Calibrations are imprecise



Molecular dating: challenges

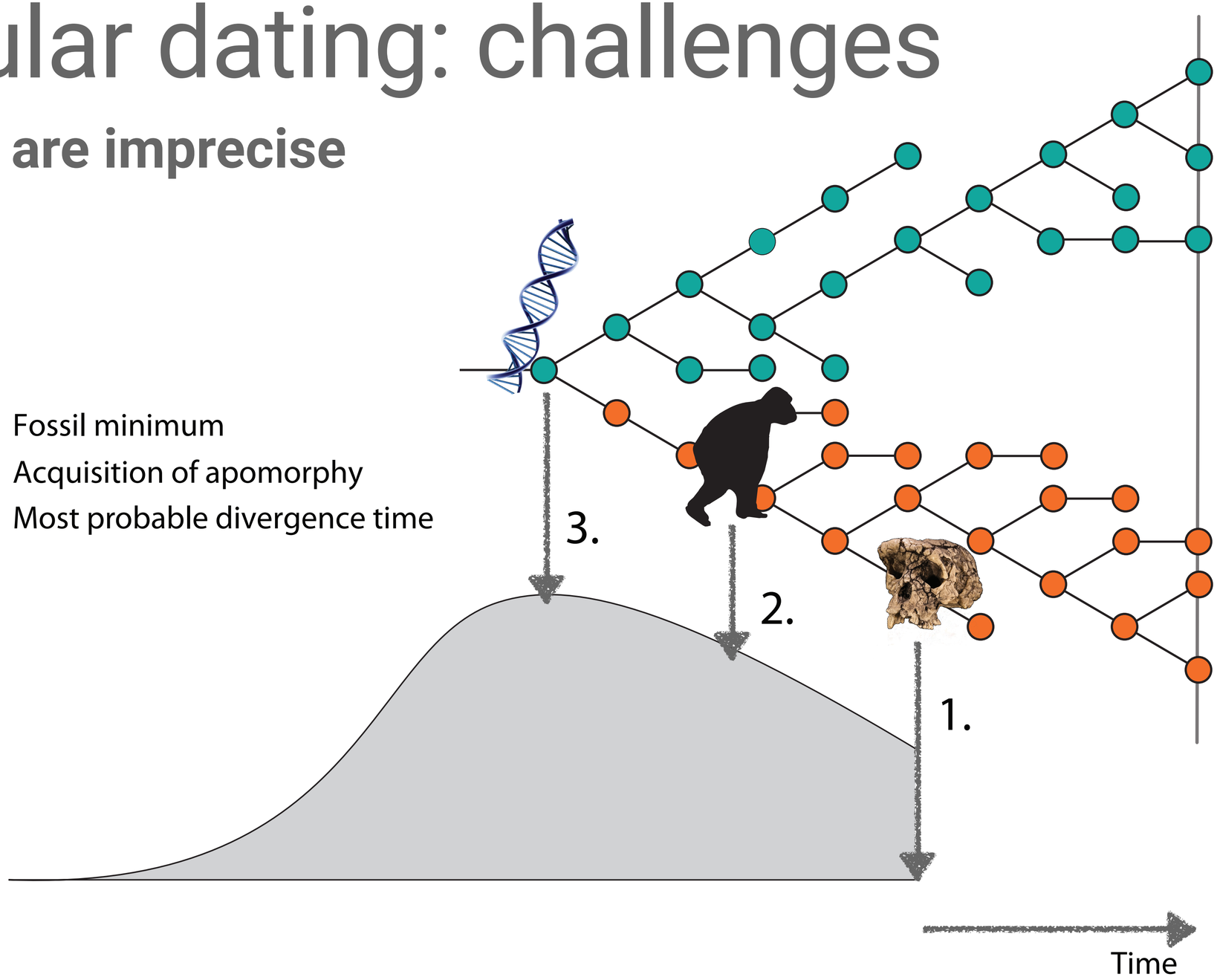
Calibrations are imprecise



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!

Homework

Phylogenetics primer part 3a: Introduction to Bayesian statistics *Paul Lewis*

See previous week's reading