Phylogenetic dating

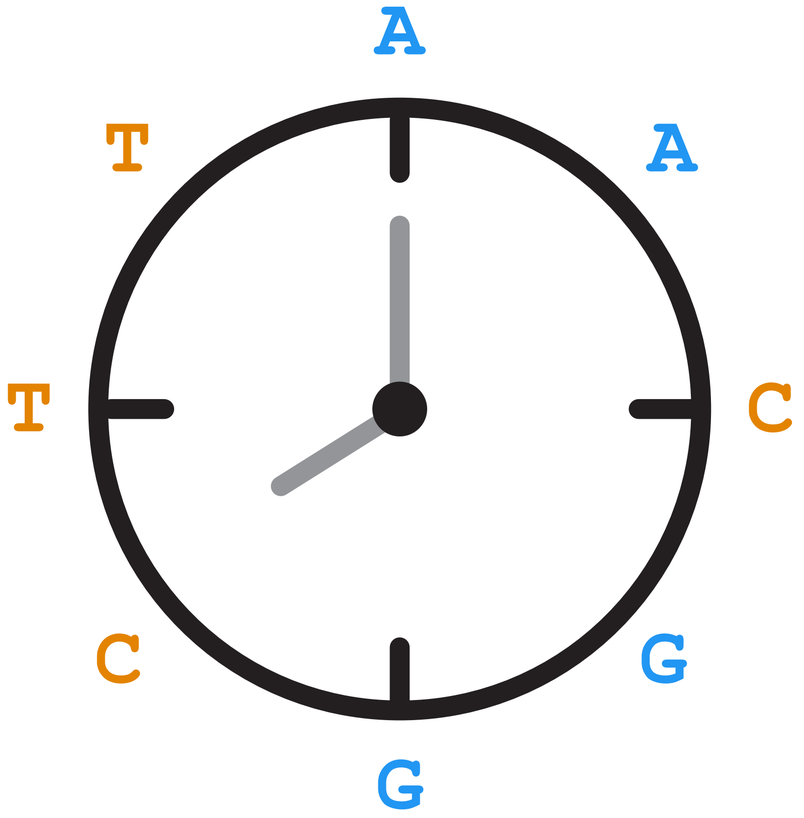
Analytical Paleobiology

Rachel Warnock

14.08.24

Today’s objectives

- **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

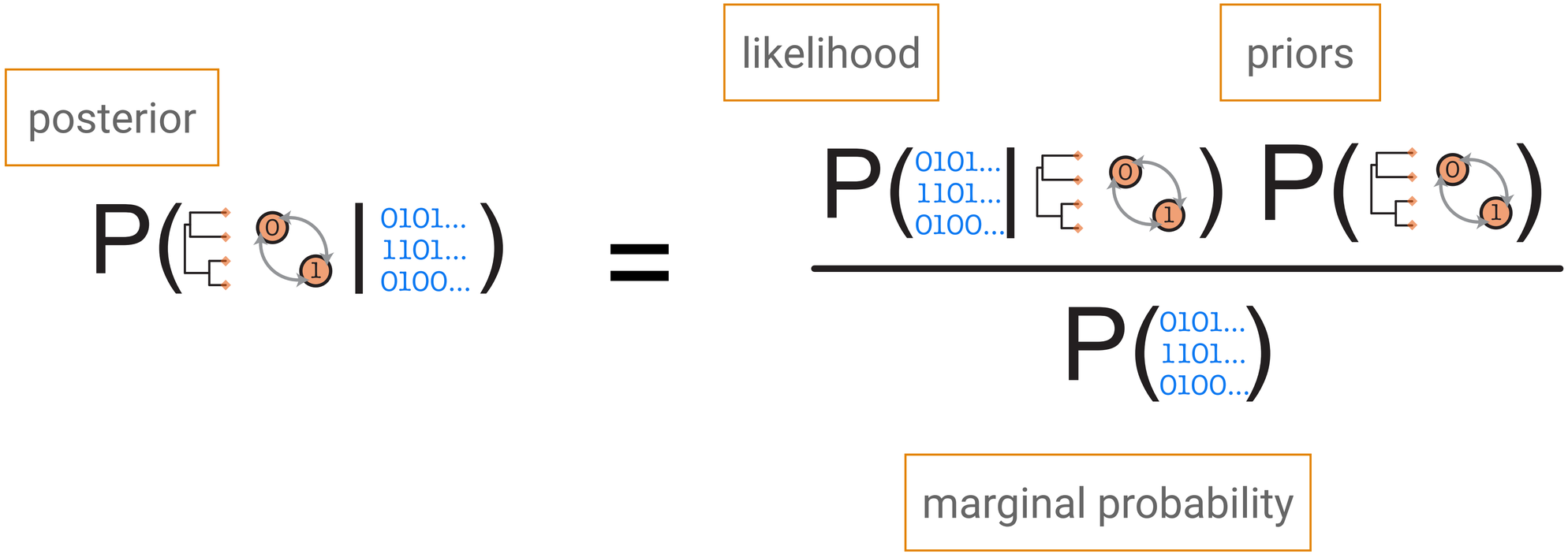
Bayes' theorem

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



How do we find the ‘best’ tree?

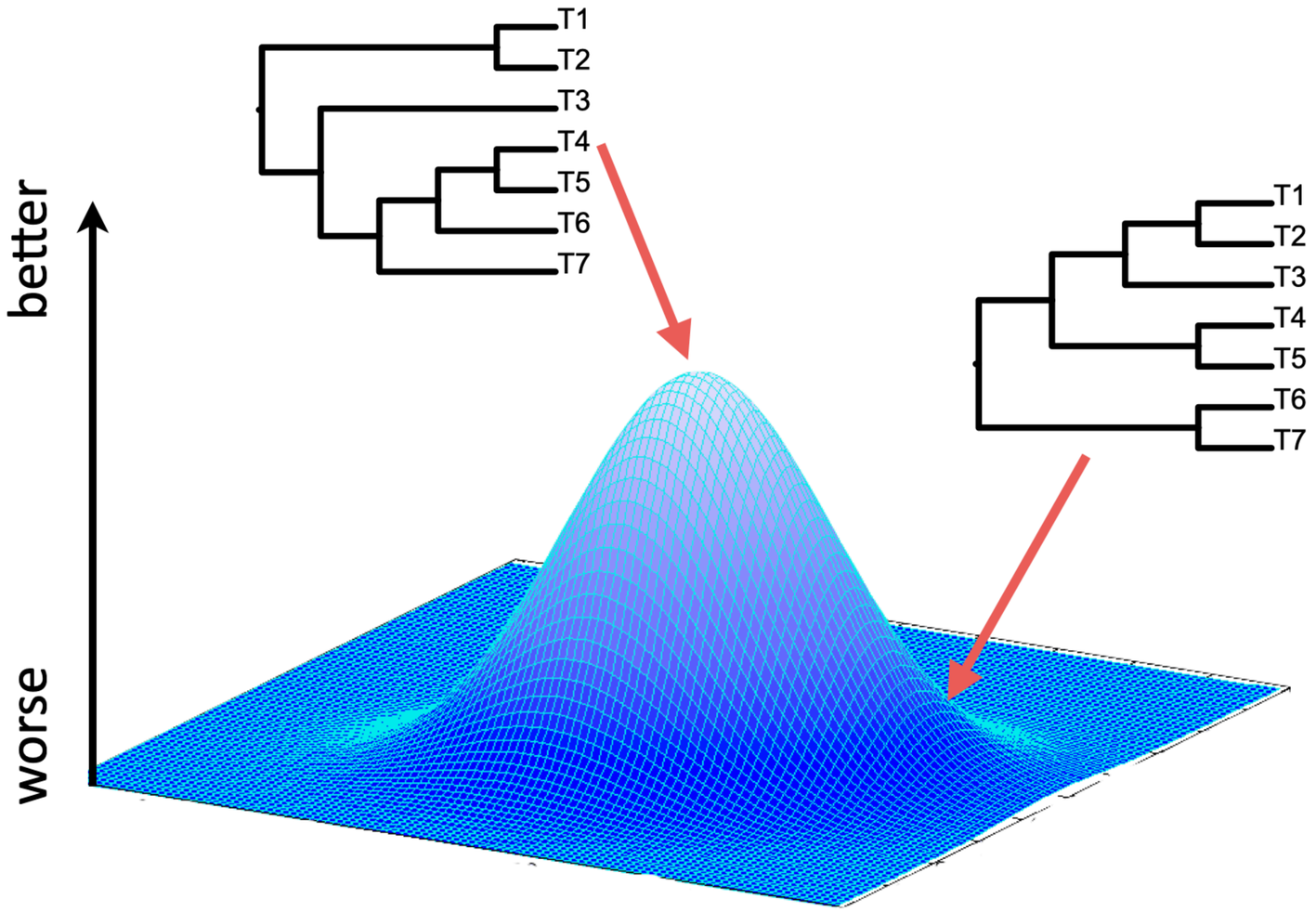


Image source Tracy Heath

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

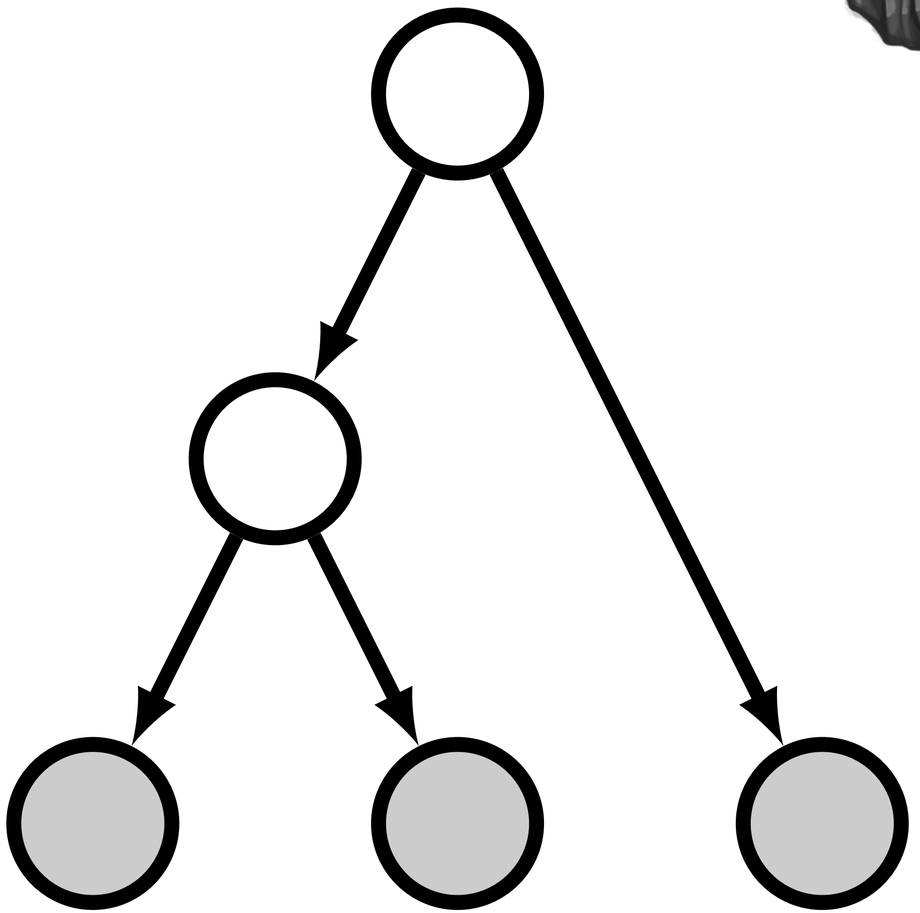
Both maximum likelihood and Bayesian inference are model-based approaches

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

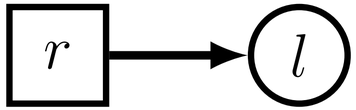
Depict the conditional dependence structure of parameters and other random variables



a)

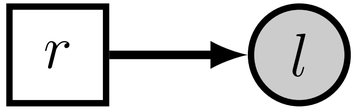

```
# constant node
r <- 10
```

b)



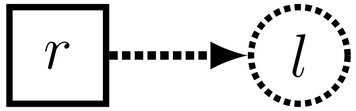

```
# stochastic node
l ~ dnExp(r)
```

c)



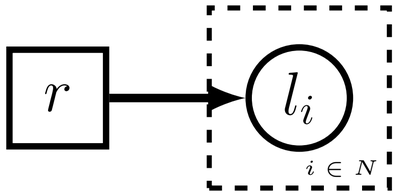

```
# stochastic node (observed)
l.clamp(0.1)
```

d)



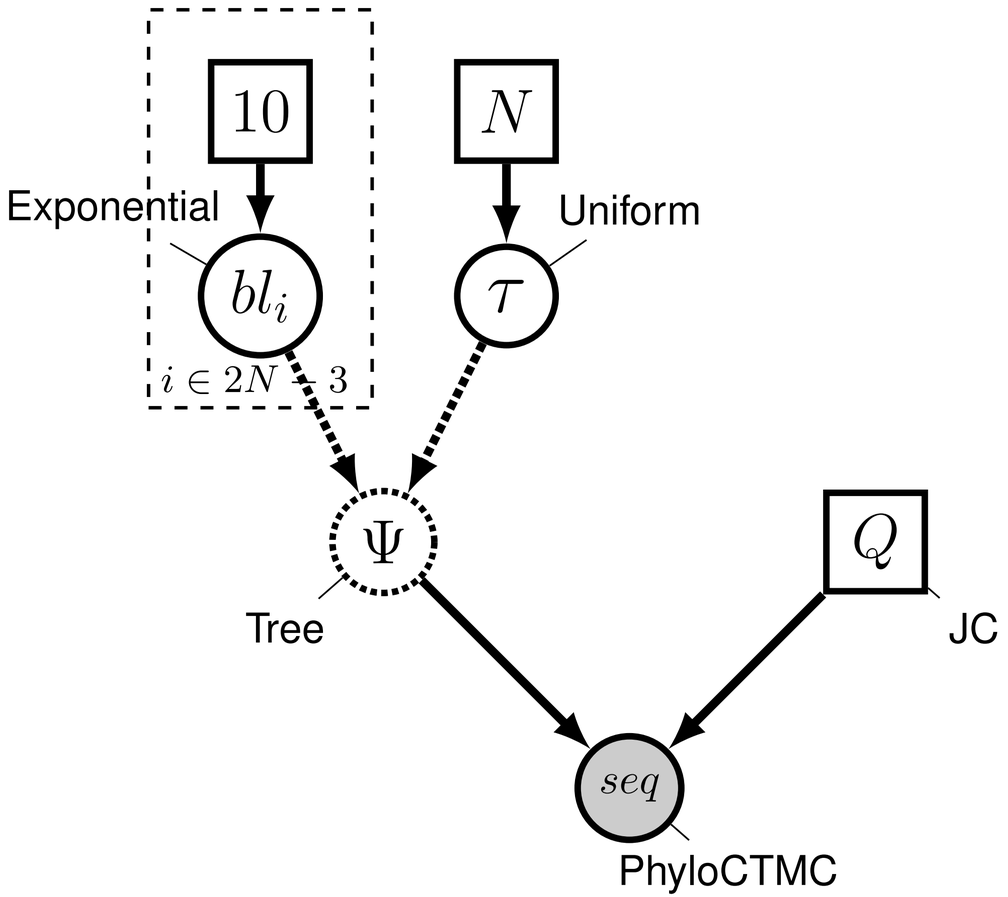

```
# deterministic node
l := exp(r)
```

e)




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

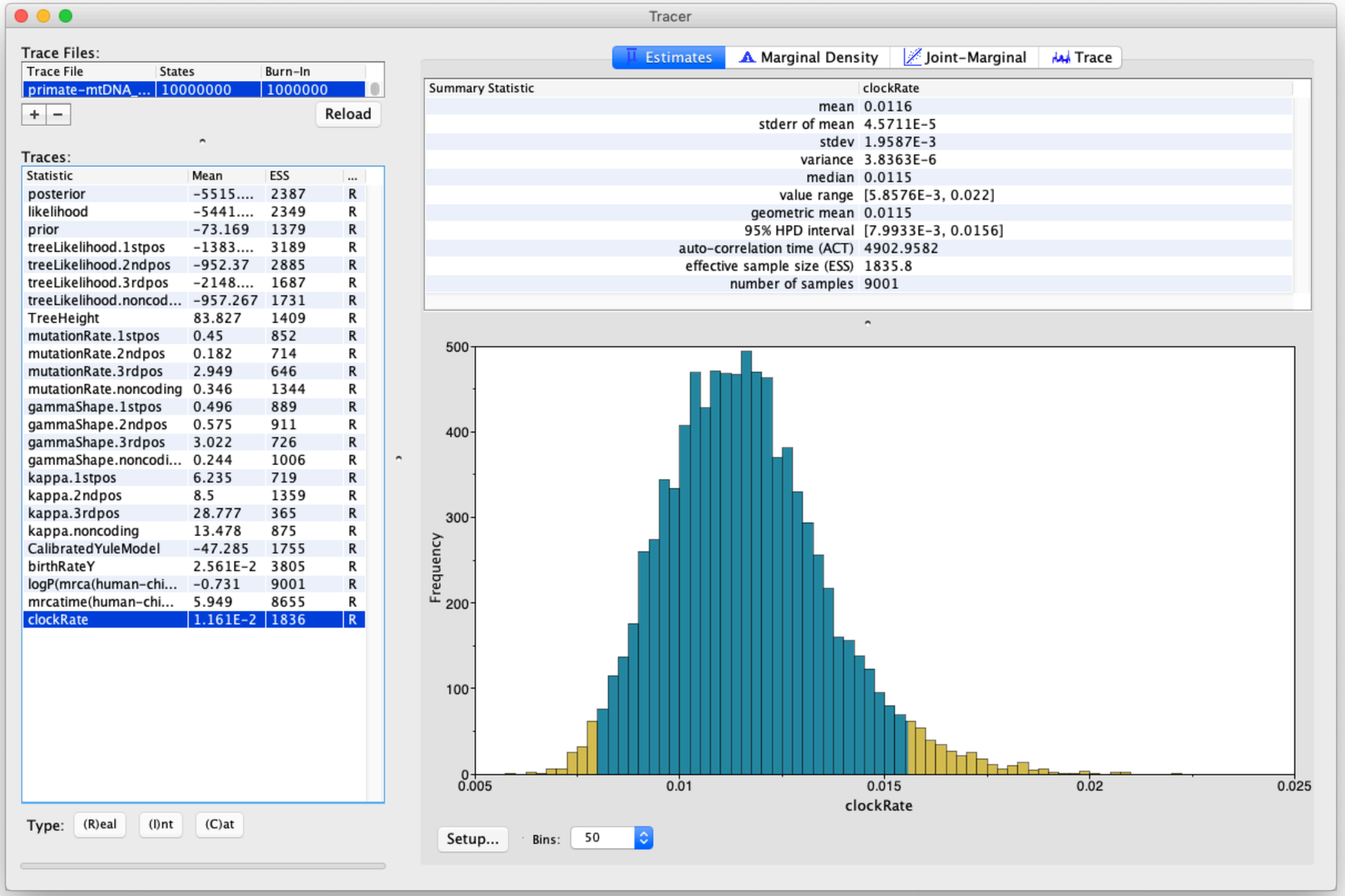
$




```
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 )
```

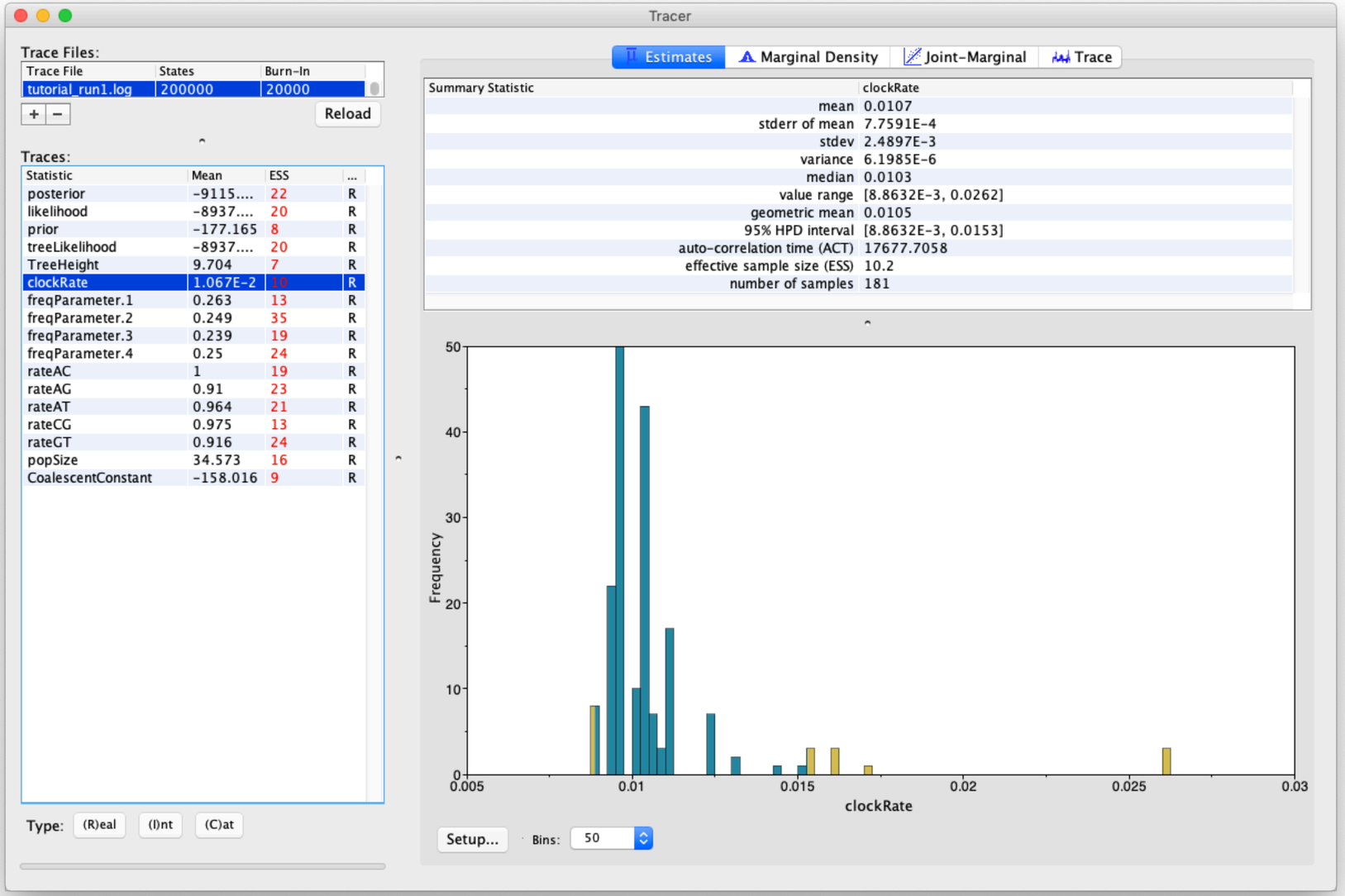
Summarising the posterior

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

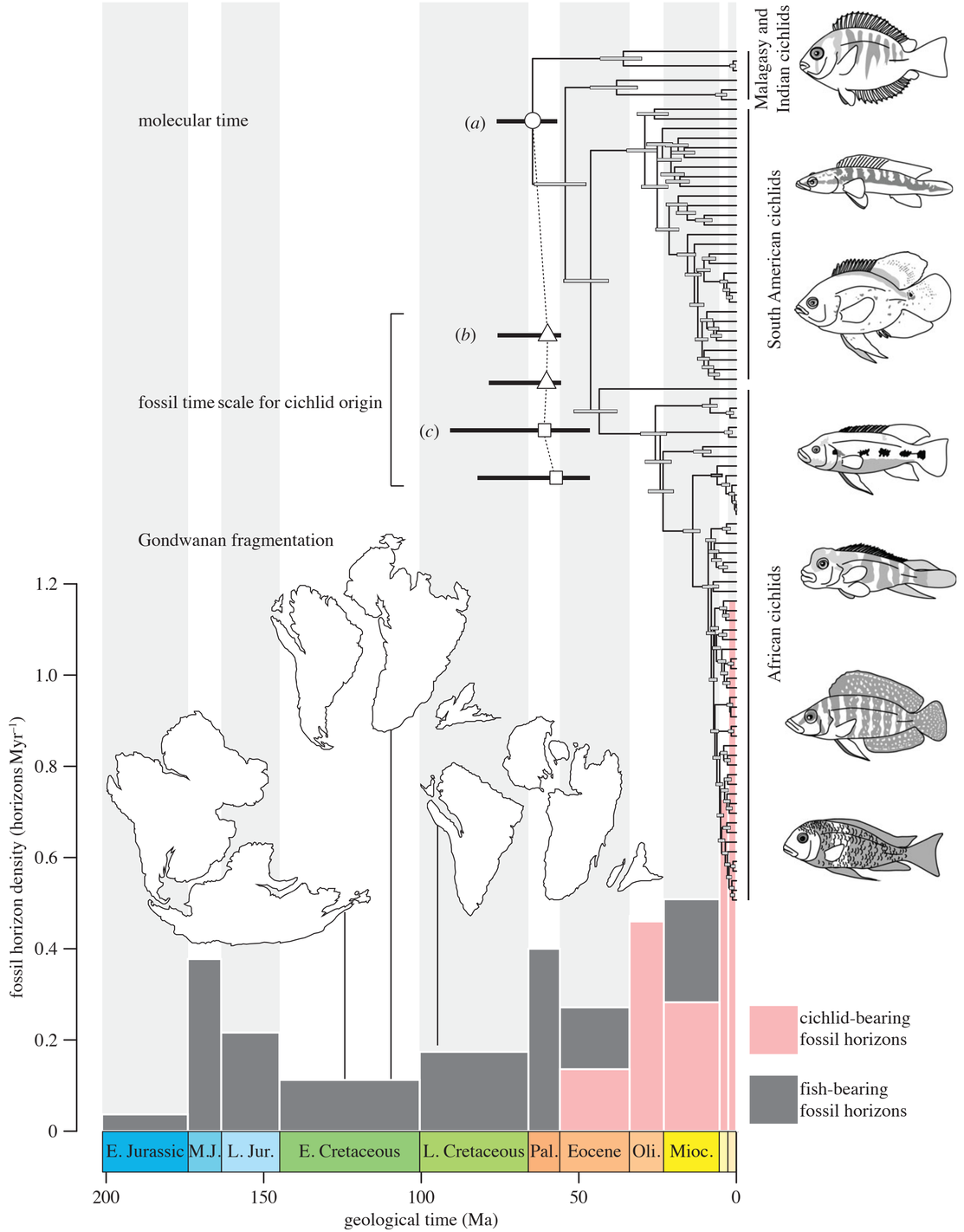
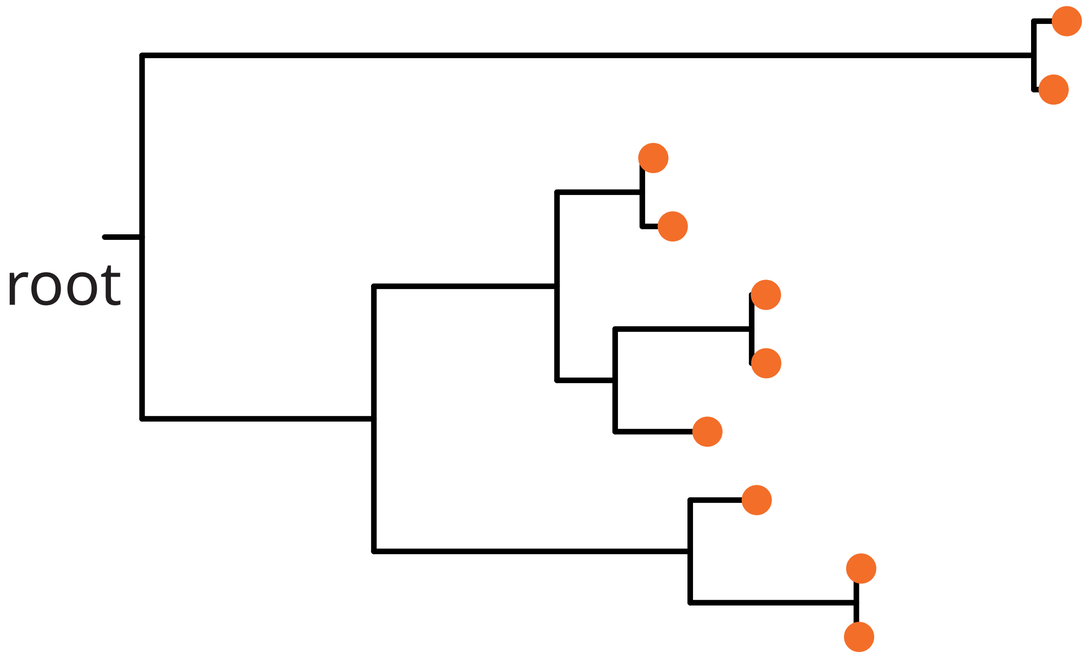


Image adapted from *Friedmann et al.* (2013)

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



ATGCATGC
TTGCCTGC
TTGCATCG
ATGCATCG
ATGCATGG
TTGCCTGG
TAGCGTGC
TAGCGAGC

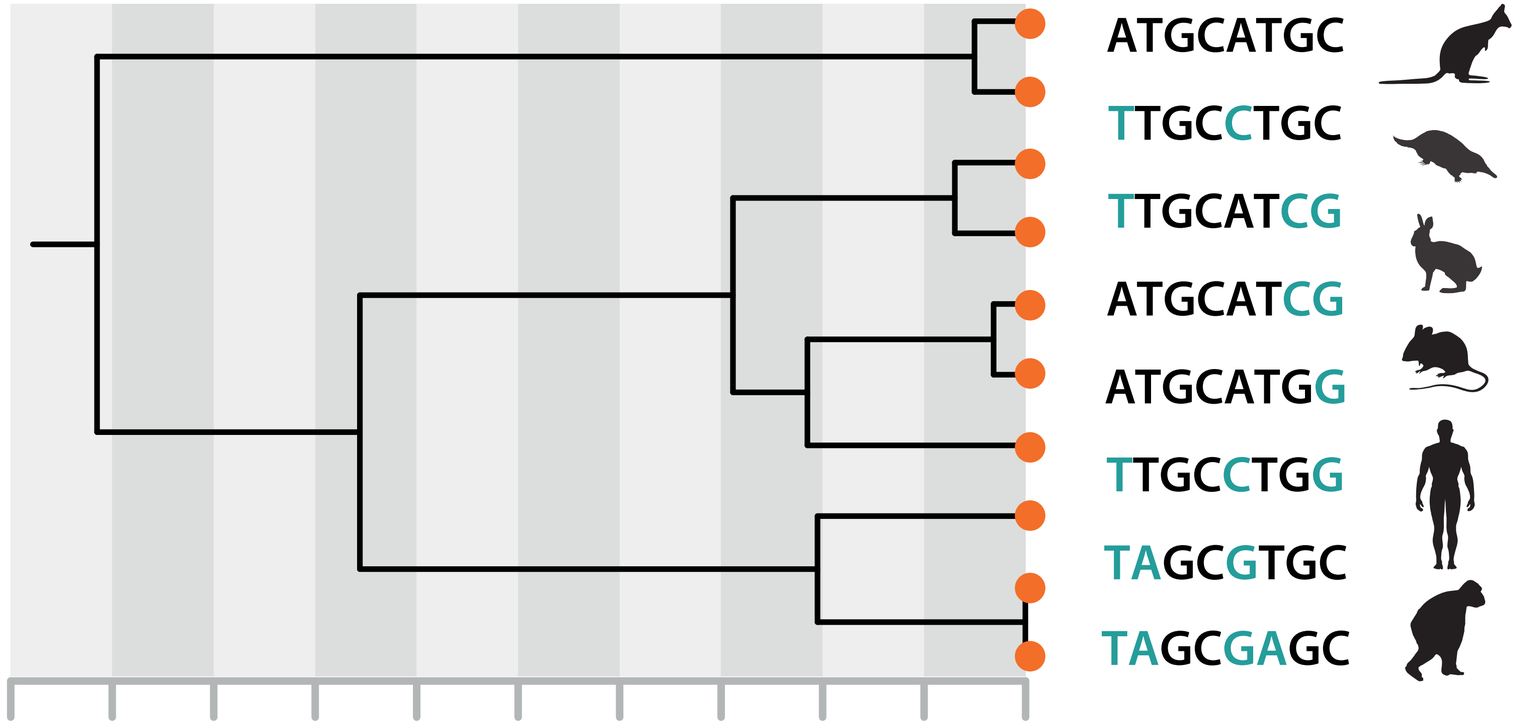
Slow rate, long interval **or** fast rate, short interval?

Goal: to disentangle evolutionary rate and time

branch lengths = genetic distance

$v = rt$

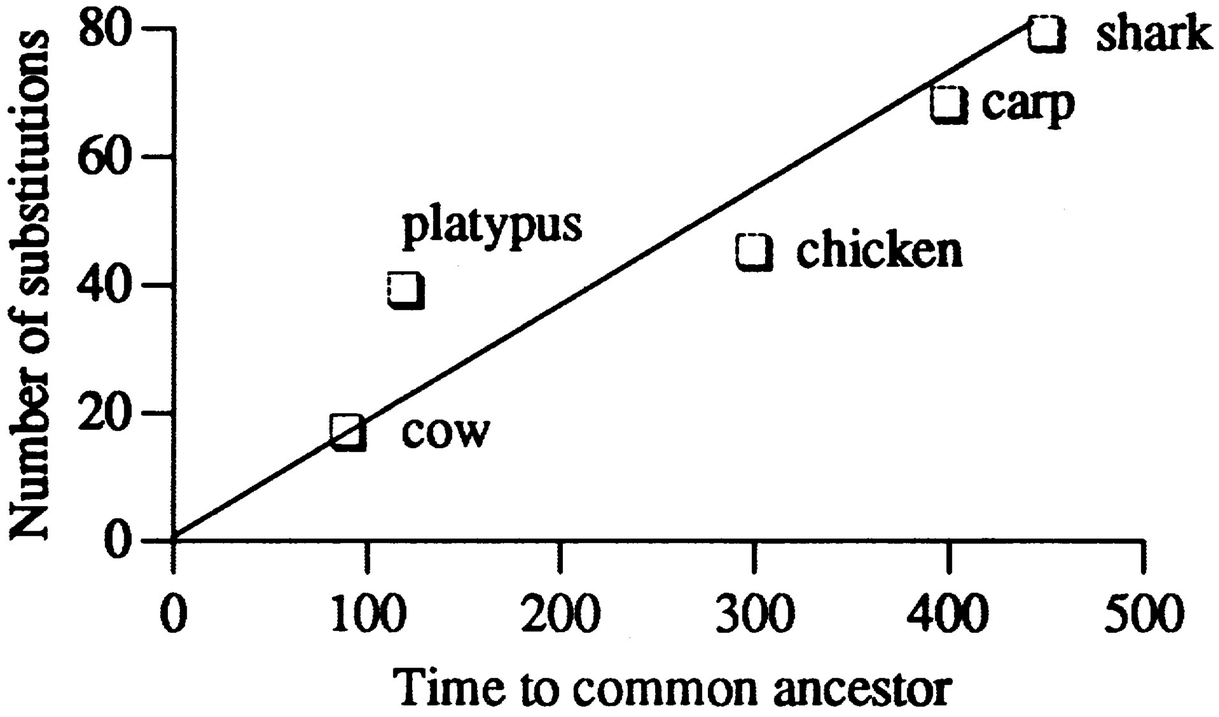
Molecular (or morphological) characters are not independently informative about time



branch lengths = time

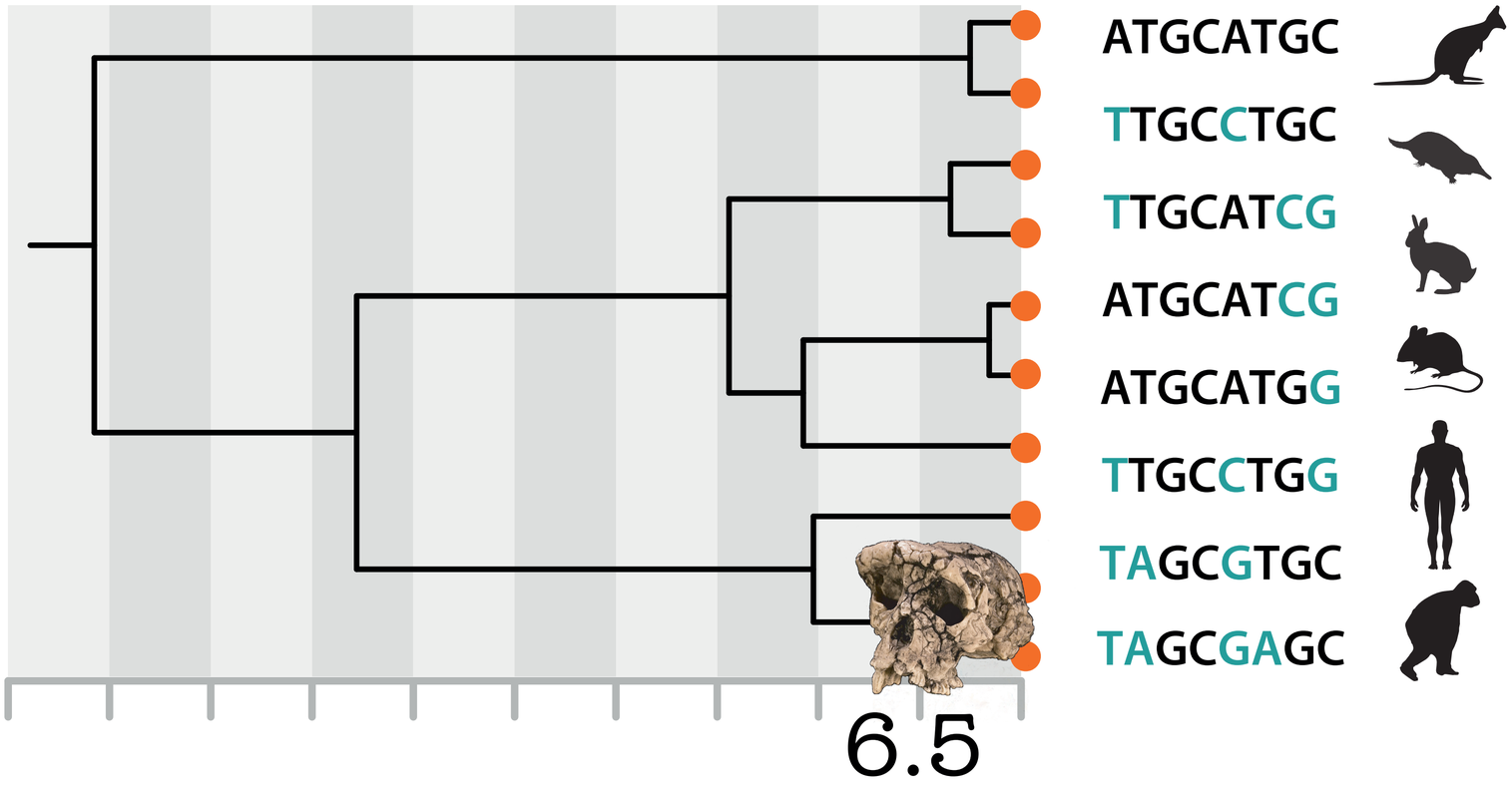
Goal: to disentangle evolutionary rate and time

The molecular clock hypothesis



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

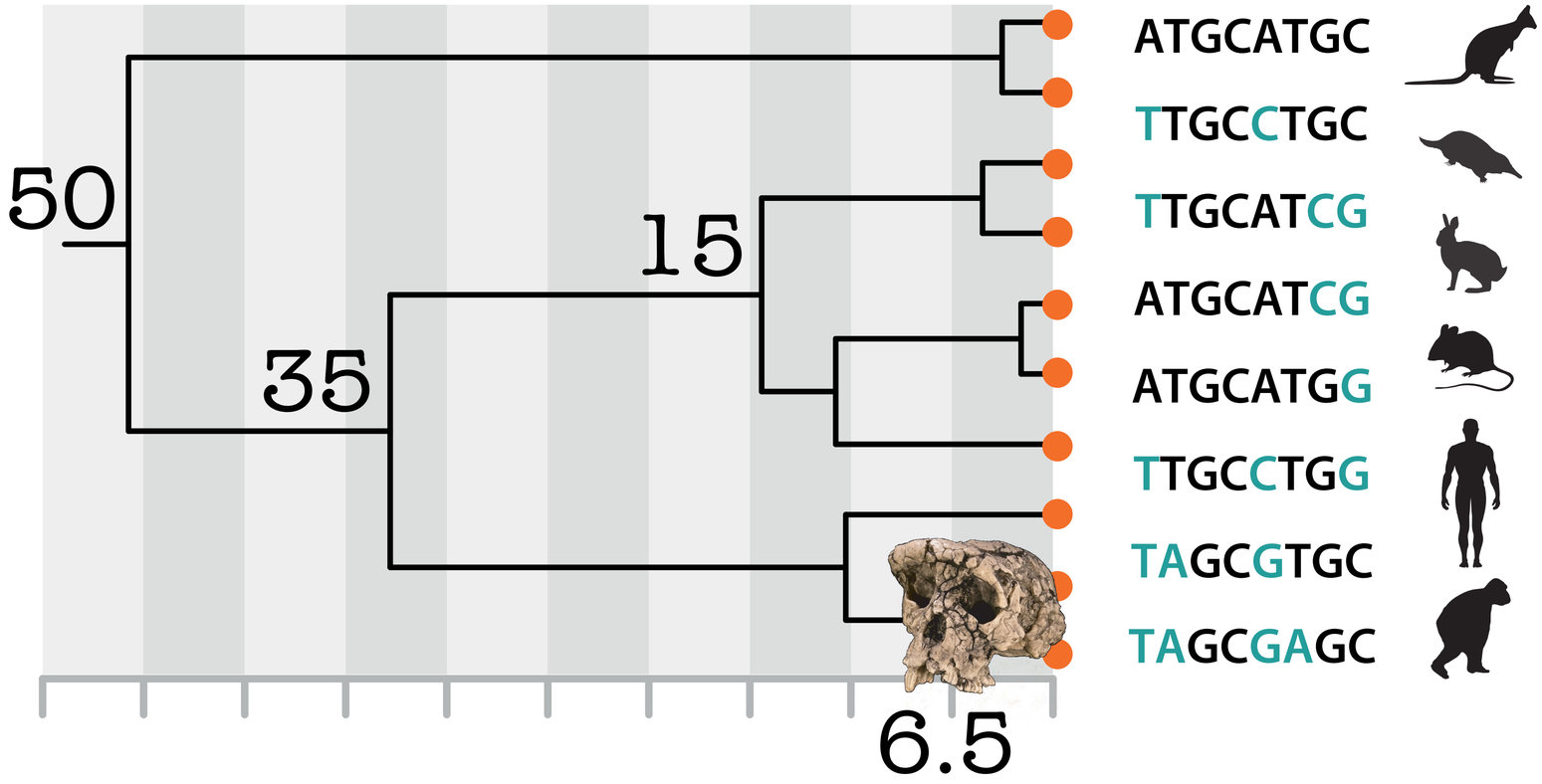
Calibrating the substitution rate



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

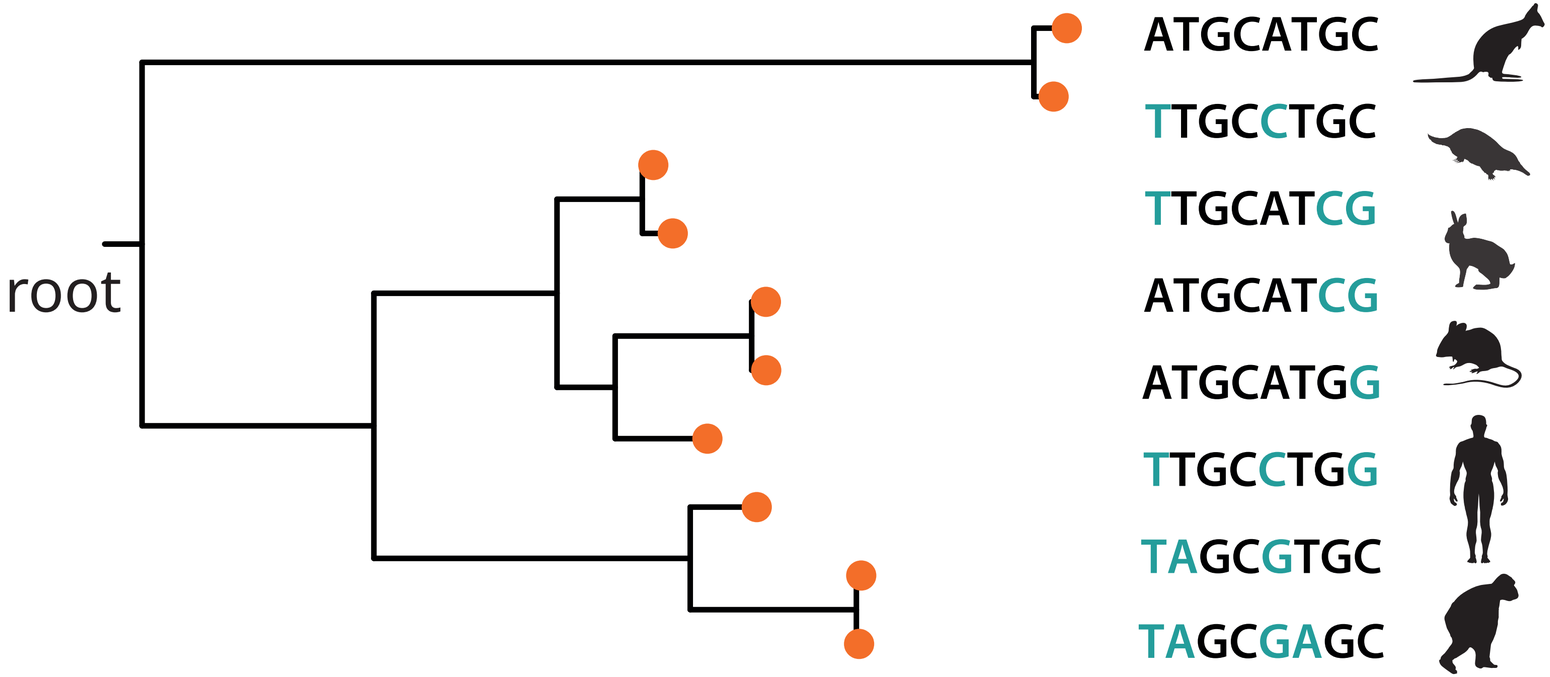


branch lengths = time

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

Molecular dating: challenges

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

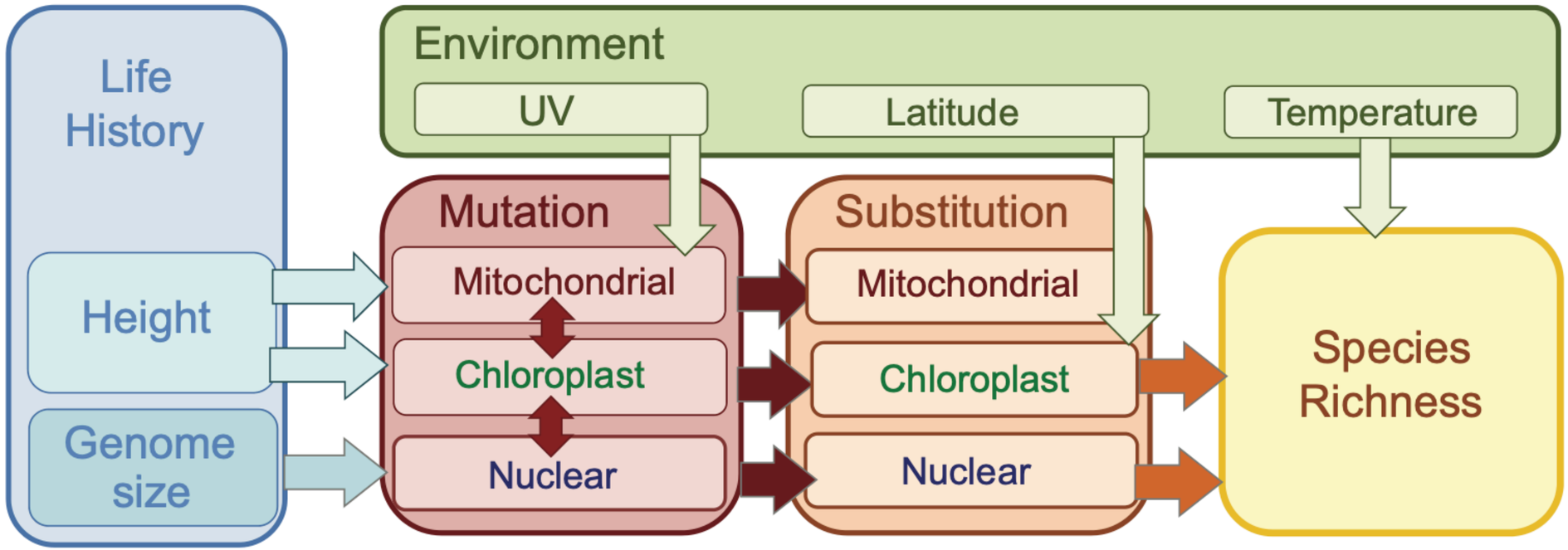


branch lengths = genetic distance

$v = rt$

Molecular dating: challenges

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



Bromham et al. (2015)

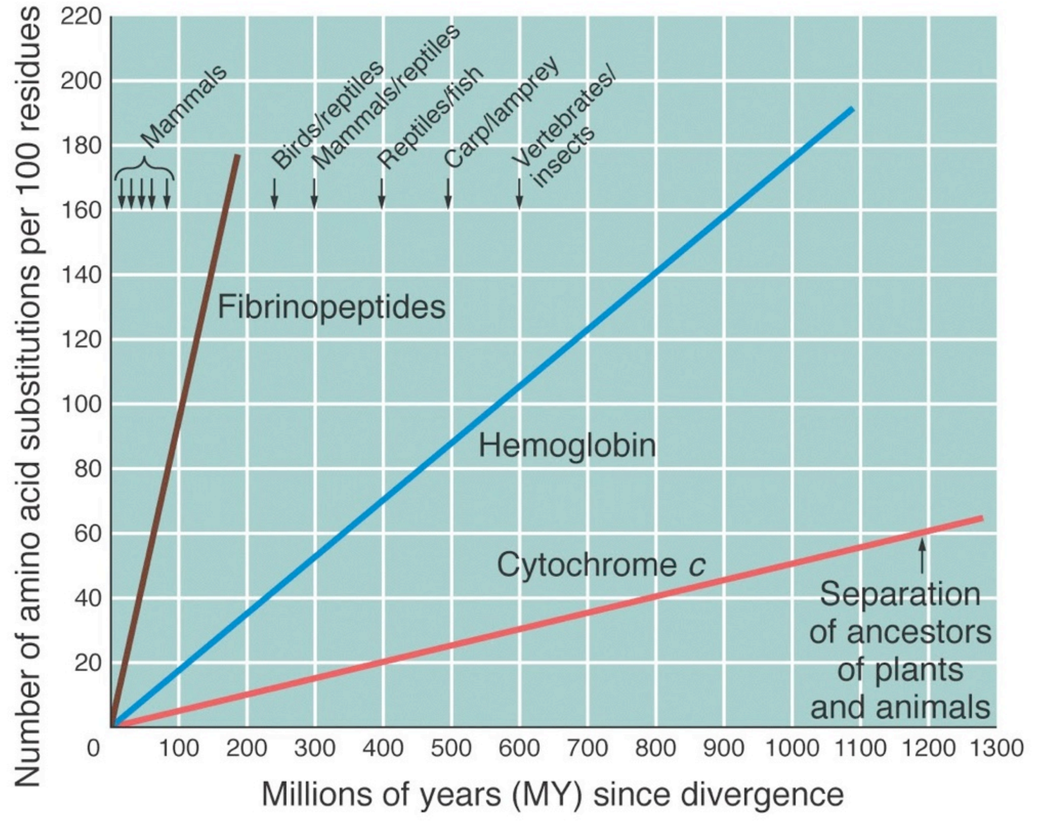
Molecular dating: challenges

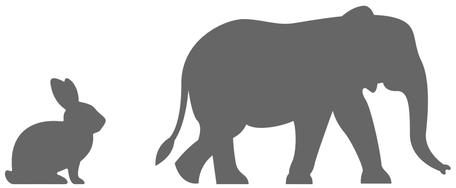
Many variables contribute to **variation in the substitution rate**

The molecular clock is not constant

Rates vary across:

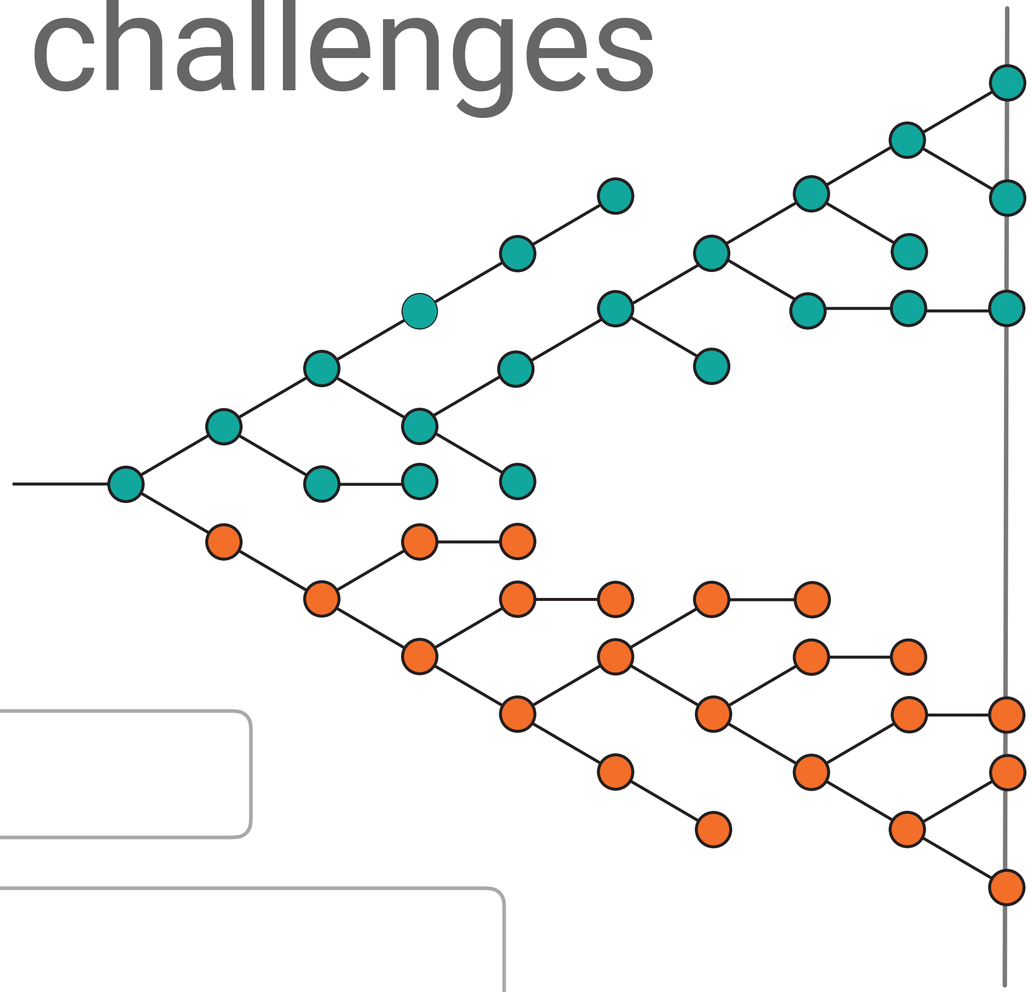
- taxa
- time
- genes
- sites within the same gene





Molecular dating: challenges

Calibrations are imprecise



Molecular evolution:

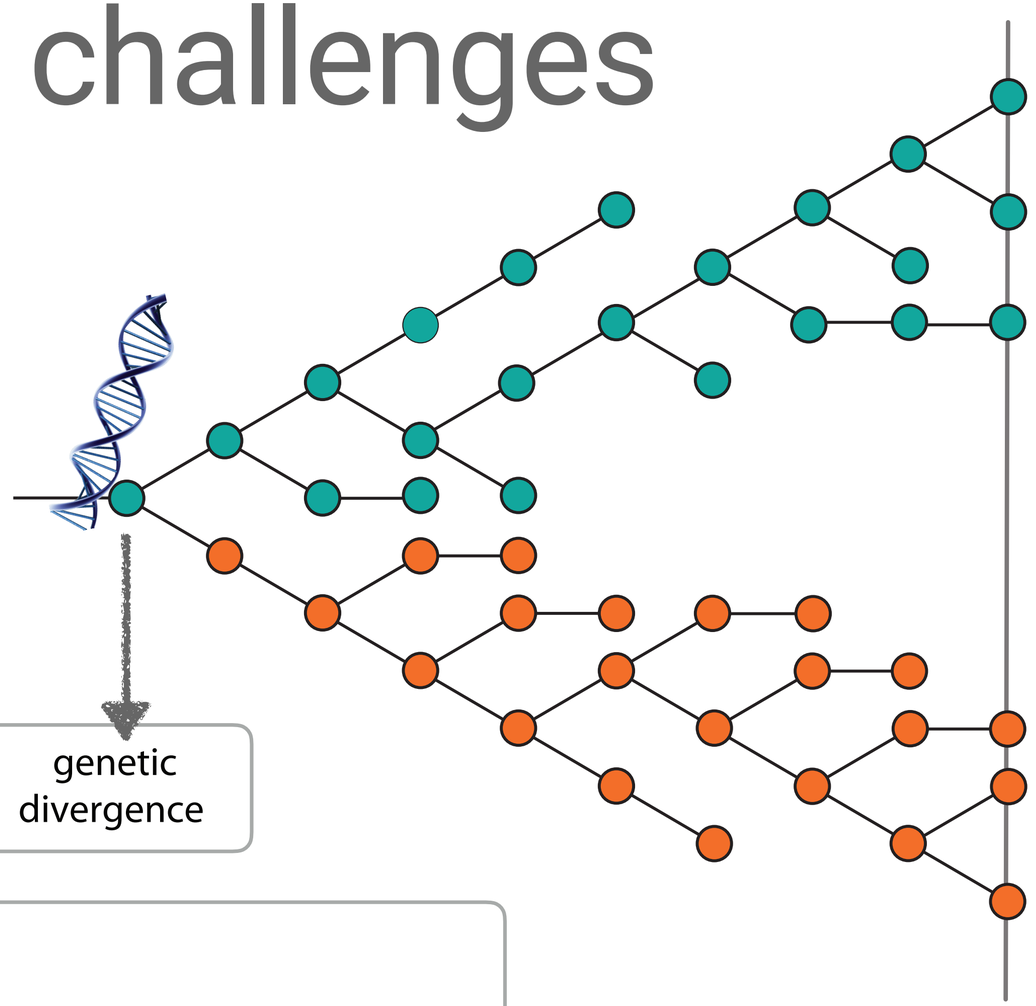
Morphological evolution:

Fossil preservation:

Time

Molecular dating: challenges

Calibrations are imprecise



Molecular evolution: genetic divergence

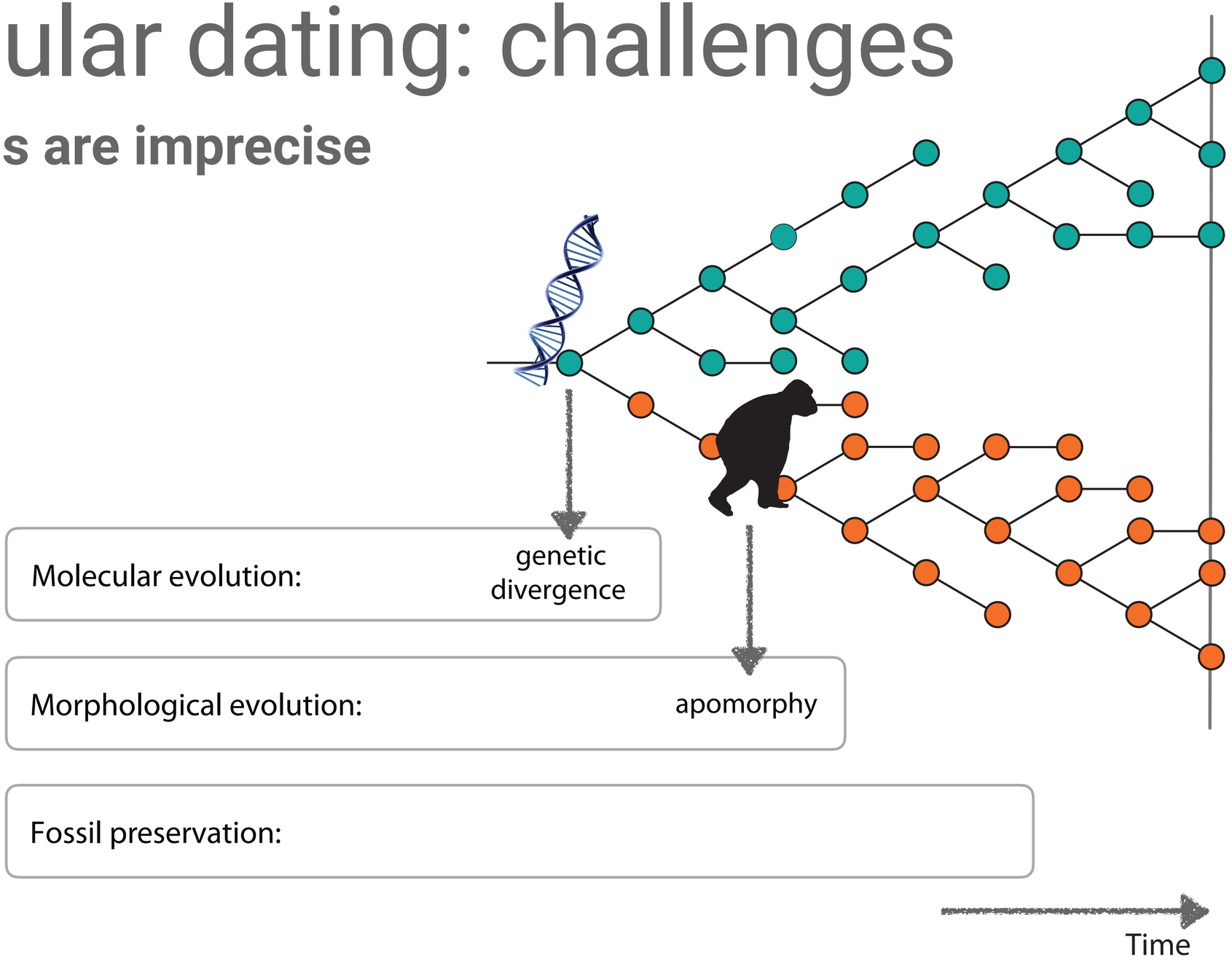
Morphological evolution:

Fossil preservation:

Time

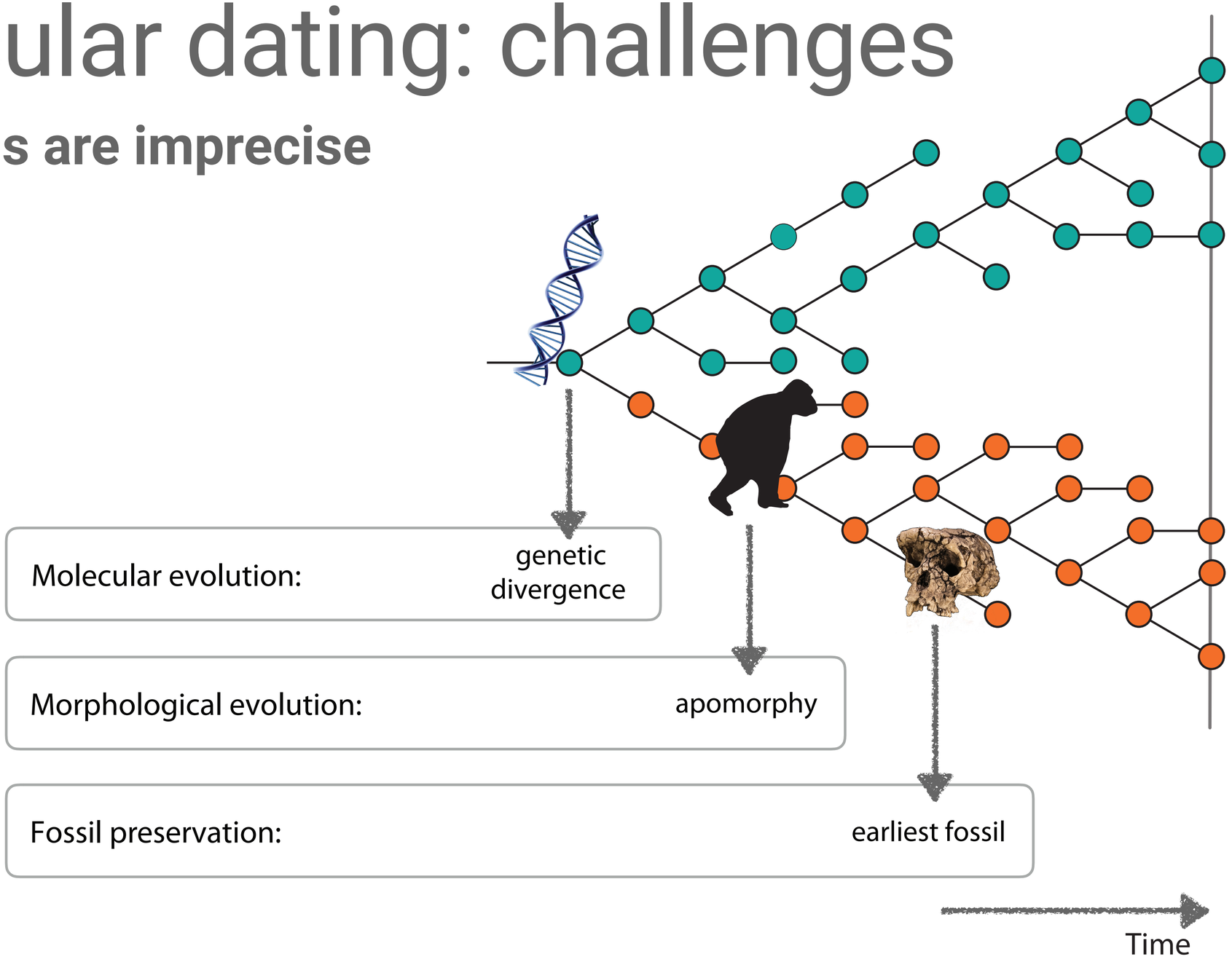
Molecular dating: challenges

Calibrations are imprecise



Molecular dating: challenges

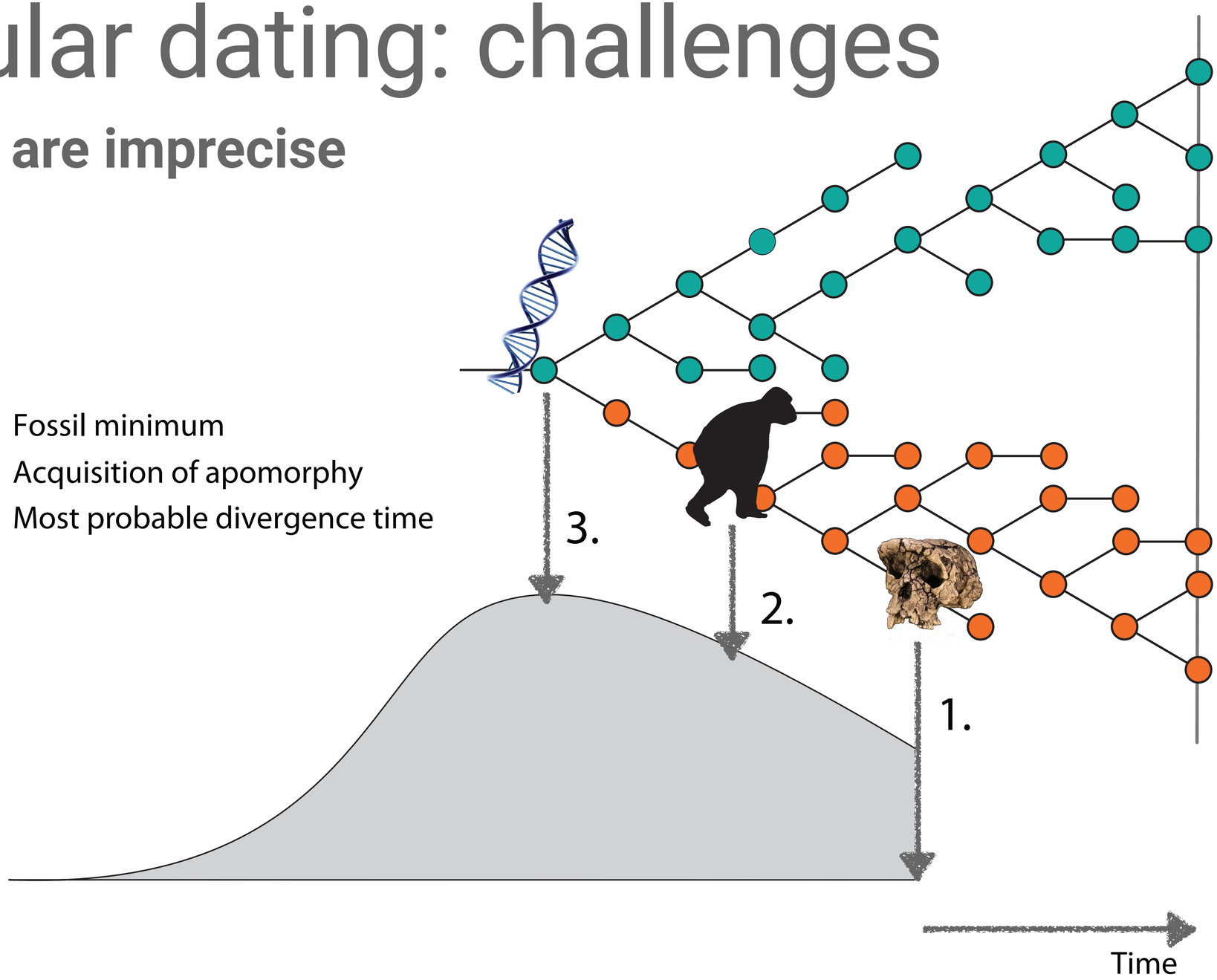
Calibrations are imprecise



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!

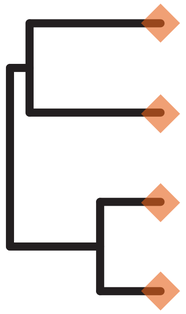
Bayesian divergence time estimation

Components used to infer trees

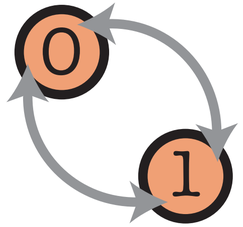
without considering time

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

data
sequences or characters

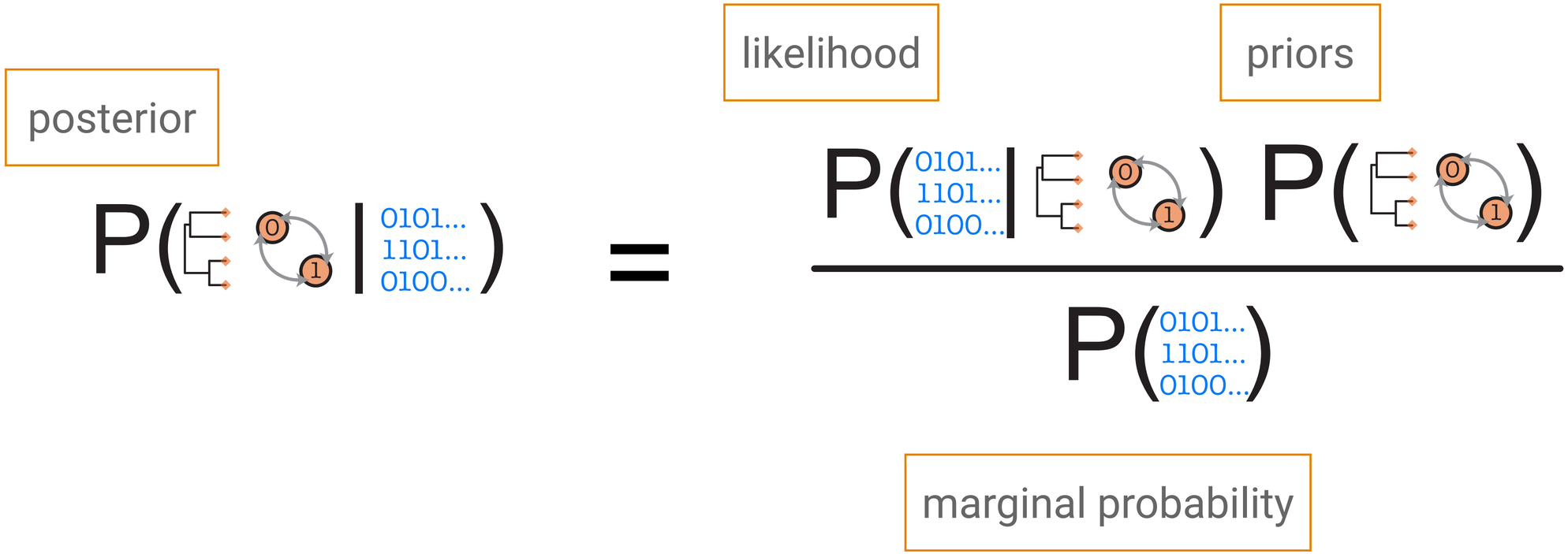


tree
topology and branch lengths

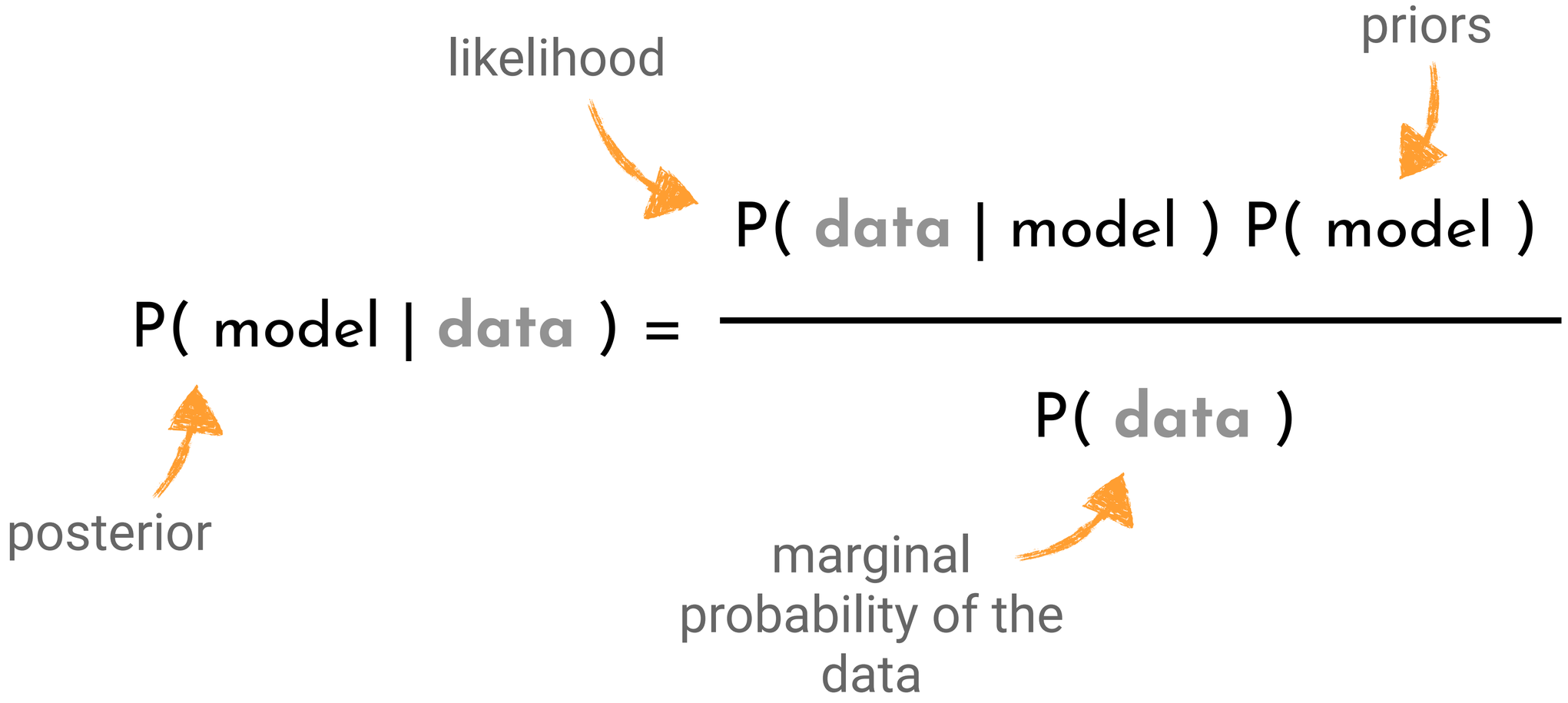


substitution model

Bayesian tree inference



We use a Bayesian framework



Bayesian divergence time estimation

The data

AND/OR

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

phylogenetics characters

sample ages

Understanding the tripartite approach to Bayesian divergence time estimation
Warnock, Wright. 2020. Elements of Paleontology

Bayesian divergence time estimation

The data

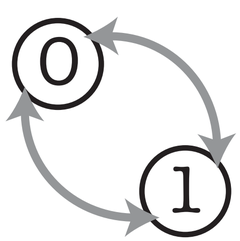
AND/OR

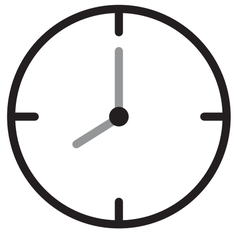
0101... ATTG...
1101... TTGC...
0100... ATTC...

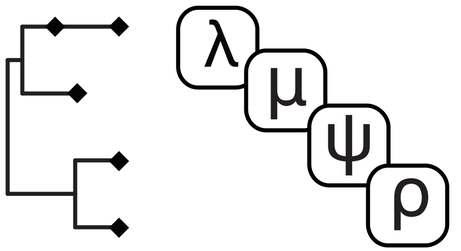
phylogenetics characters

sample ages

3 model components







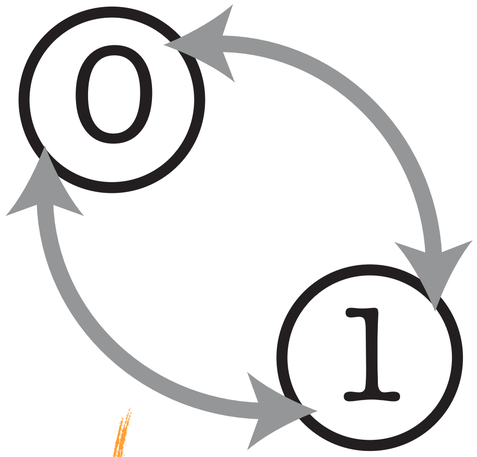
substitution model

clock model

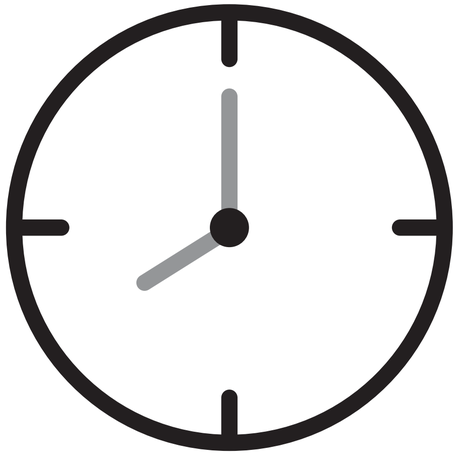
tree and tree model

Understanding the tripartite approach to Bayesian divergence time estimation
Warnock, Wright. 2020. Elements of Paleontology

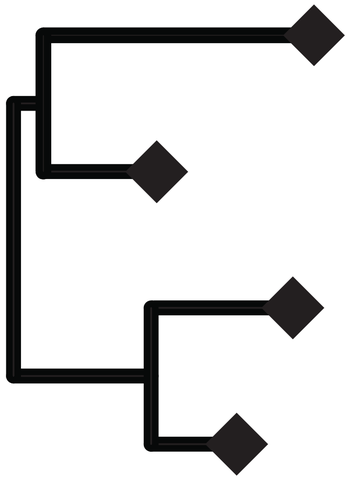
substitution model



clock model

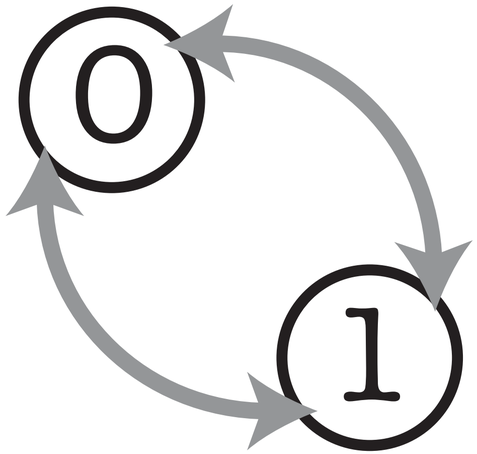


tree model

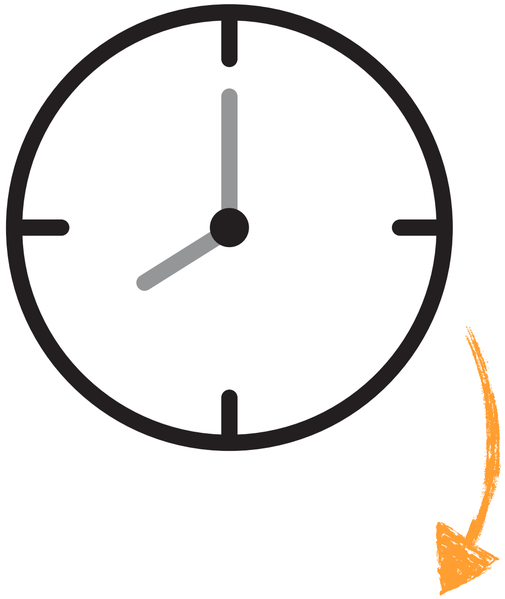


How likely are we to observe a change between character states? e.g., A → T

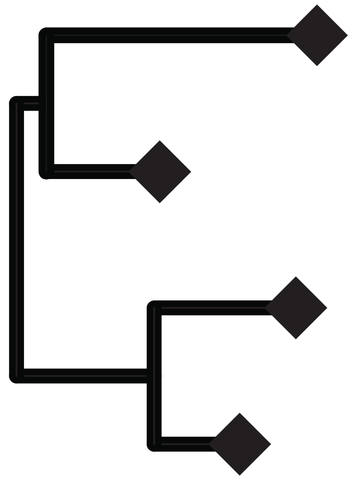
substitution model



clock model

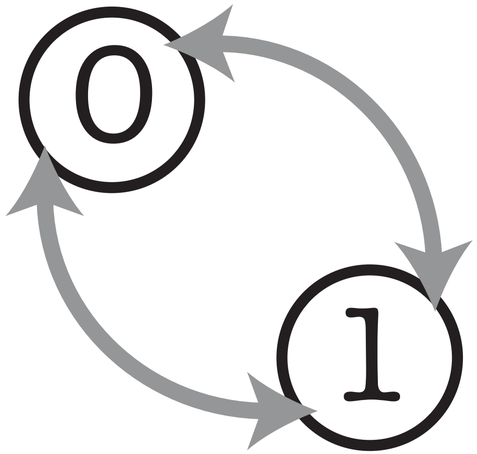


tree model

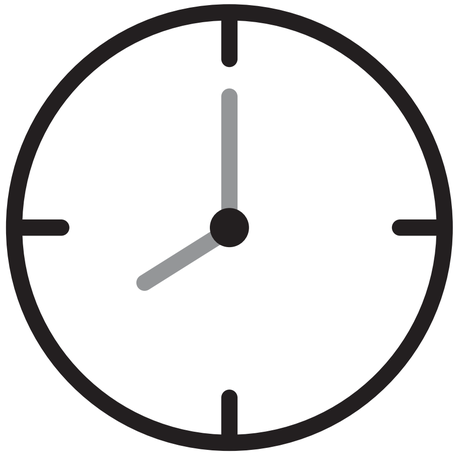


How have rates of evolution varied (or not) across the tree?

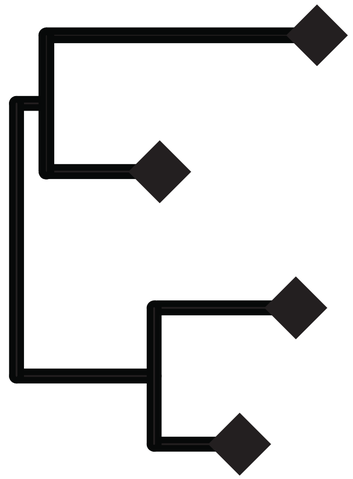
substitution model



clock model



tree model

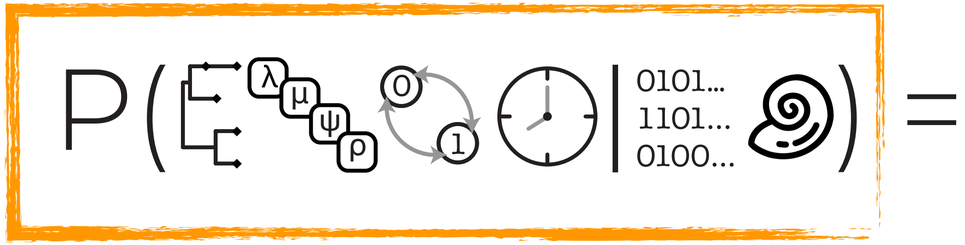


How have species originated, gone extinct and been sampled through time?

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

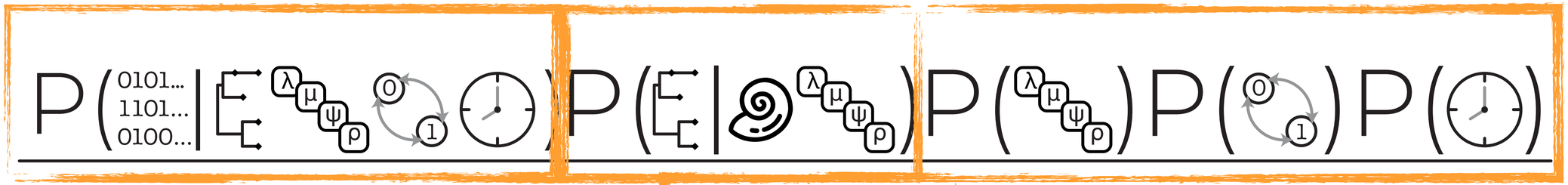
posterior

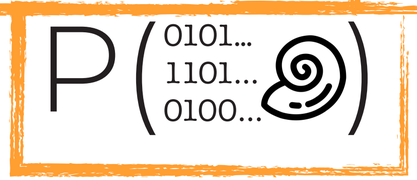


likelihood

probability of the time tree

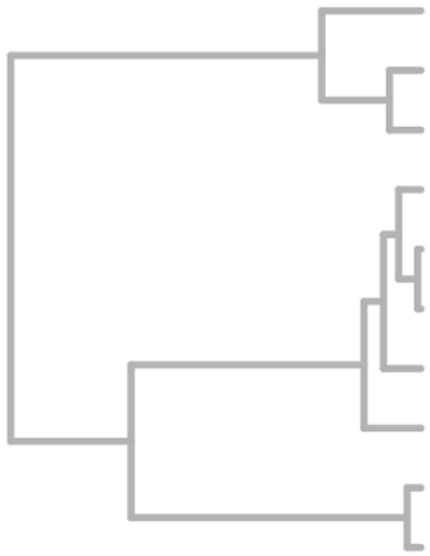
priors





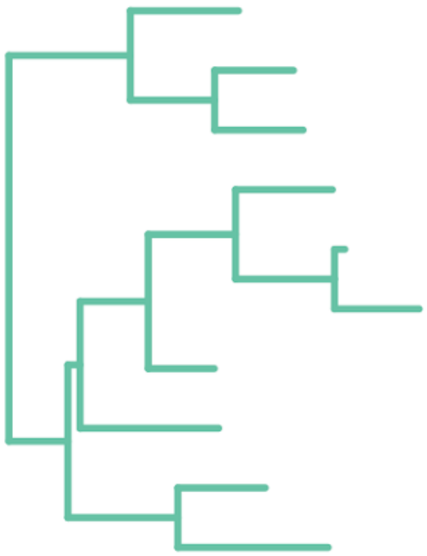
marginal pr of the data

Calculating the likelihood



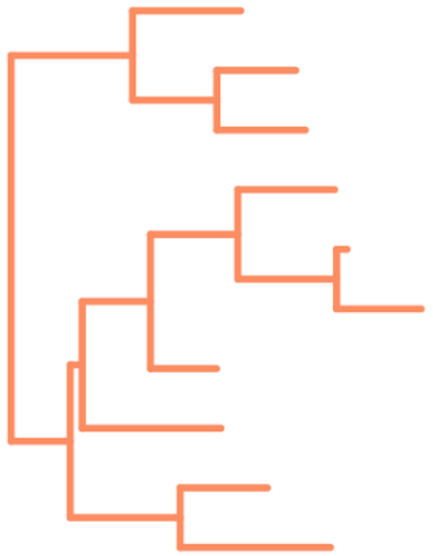
Time (year)

X



Clock rate (subs/site/year)

=



Genetic distance (subs/site)

Prior

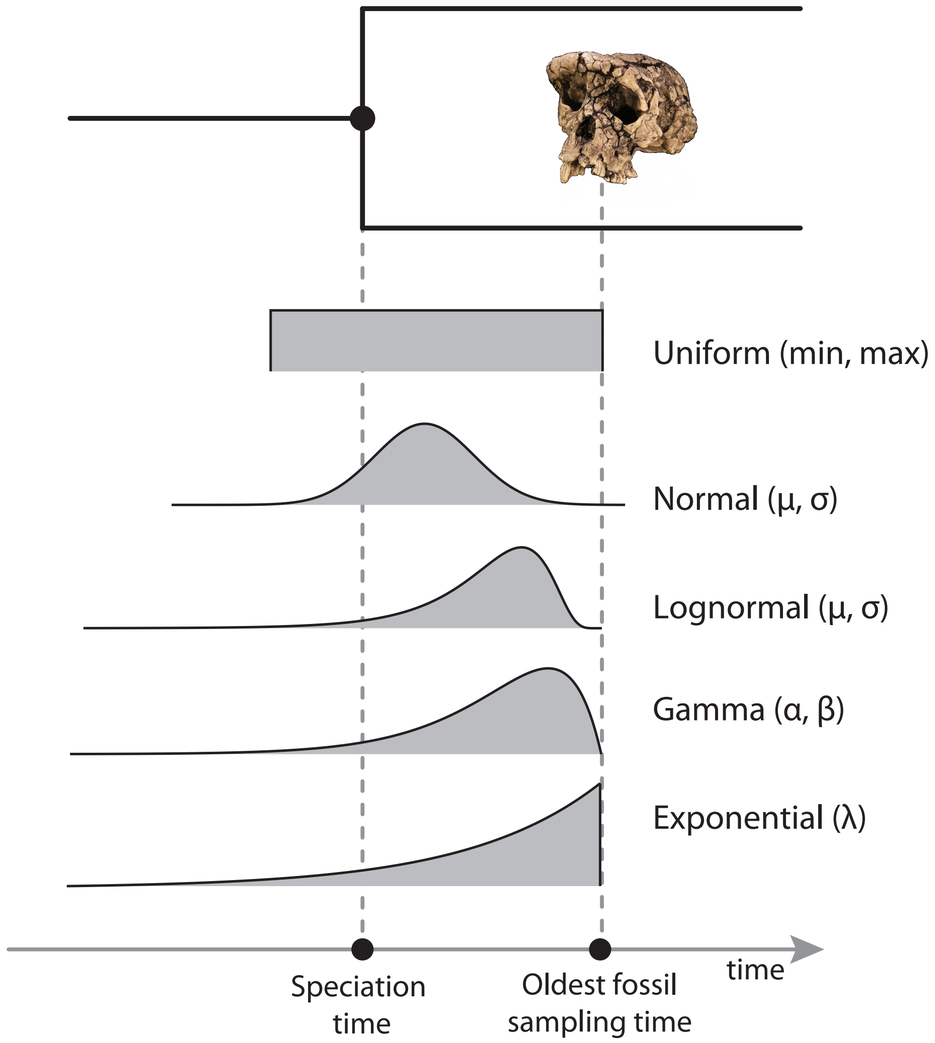
Prior

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



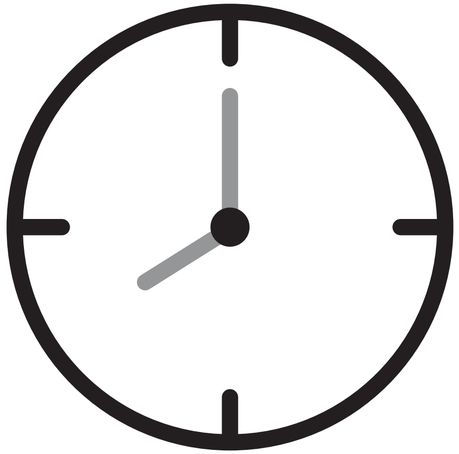
We can use a calibration density to constrain internal node ages

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

Adapted from *Heath* 2012. Sys Bio

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

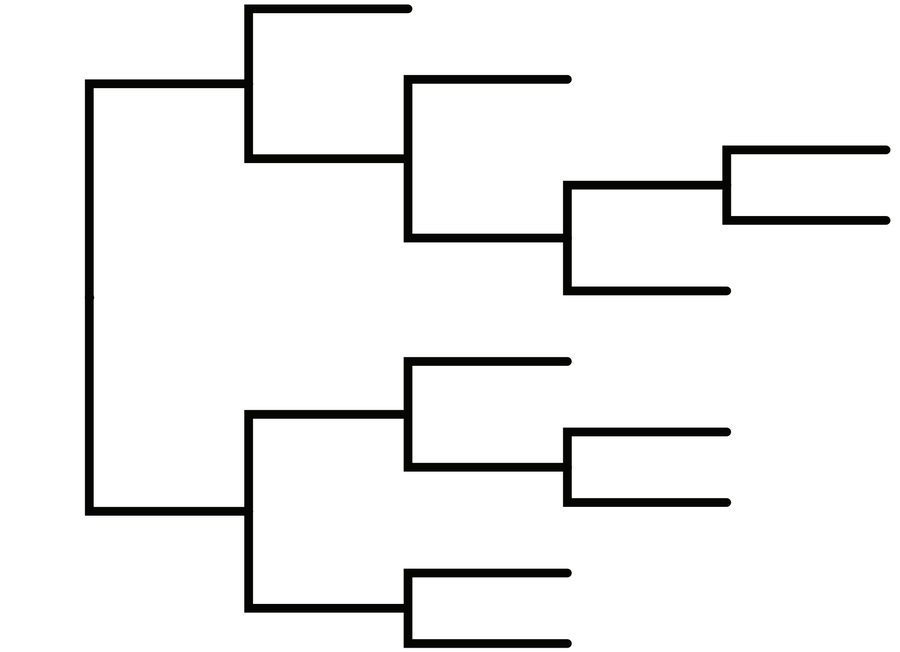
clock model



The strict / constant molecular clock model

Assumptions

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



branch length = substitution rate

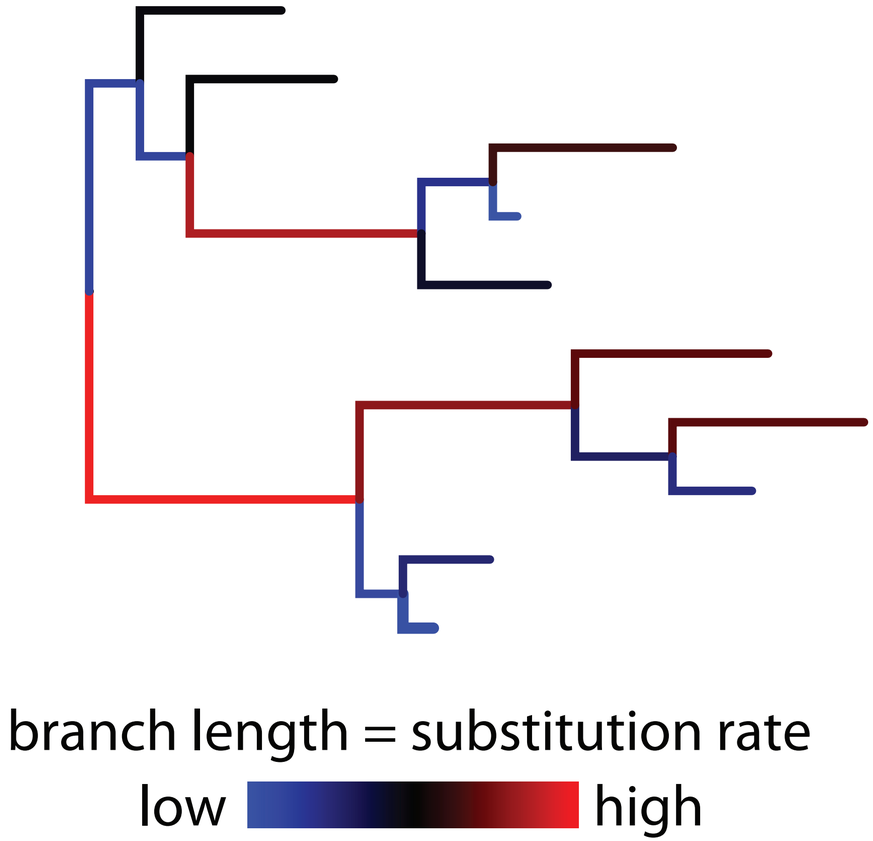
low high

Relaxed clock models

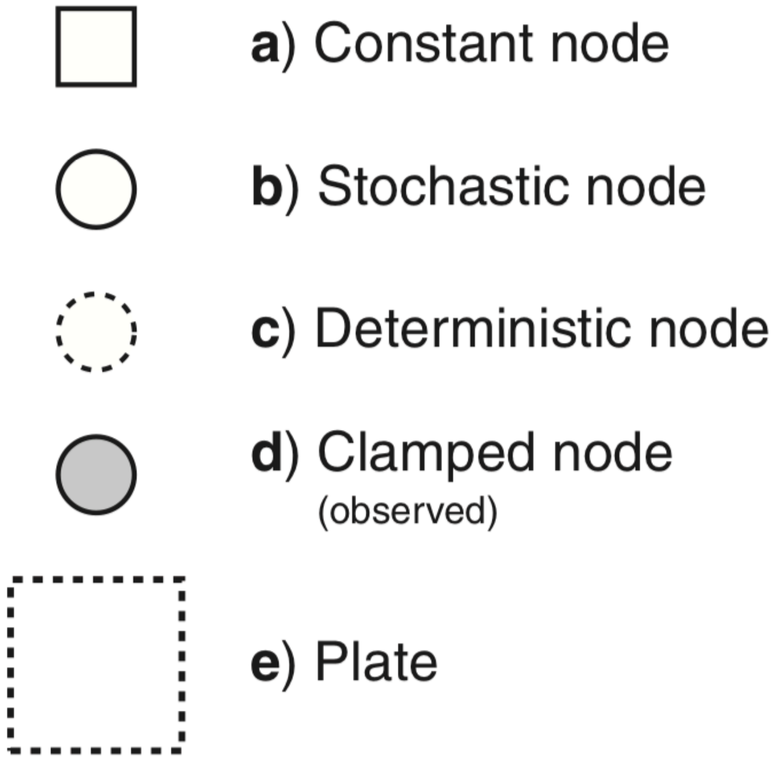
Assumptions

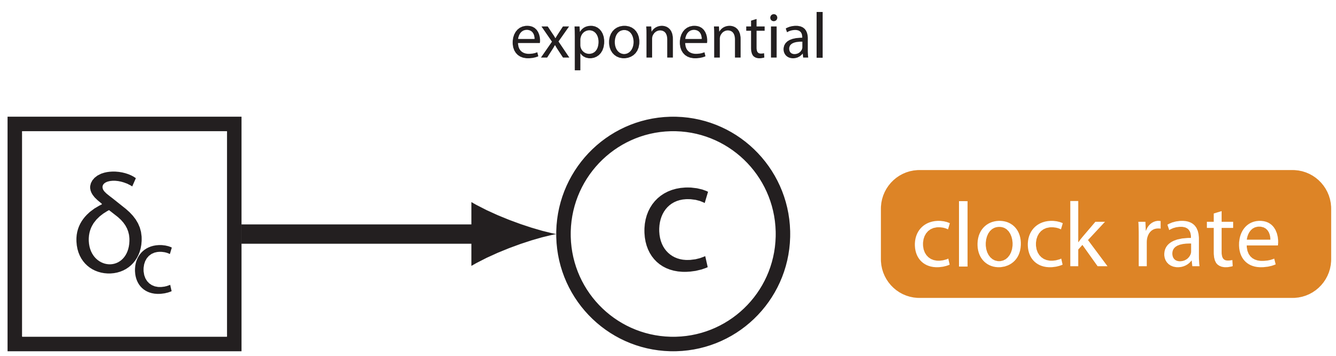
- Lineage-specific rates

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

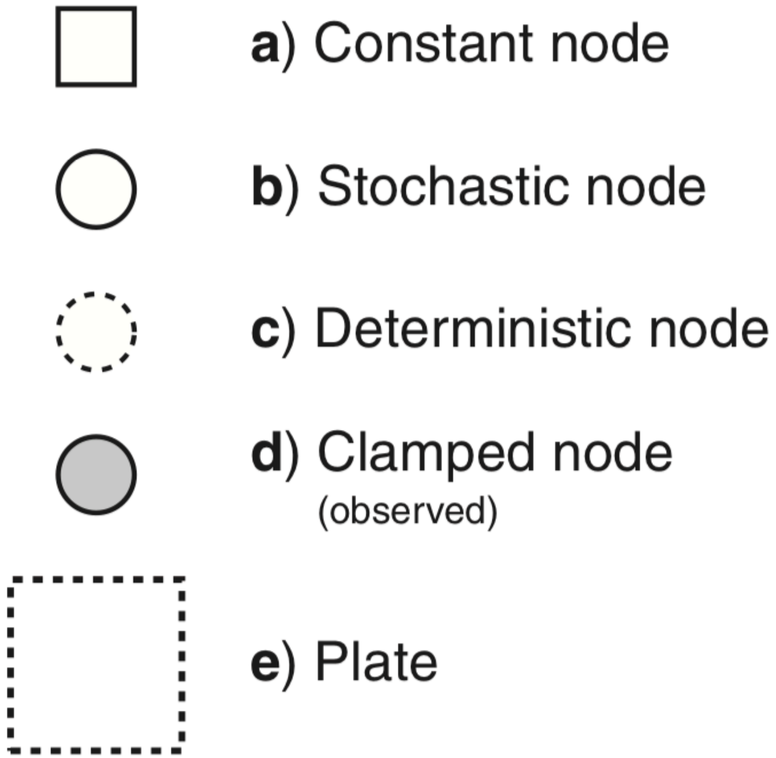


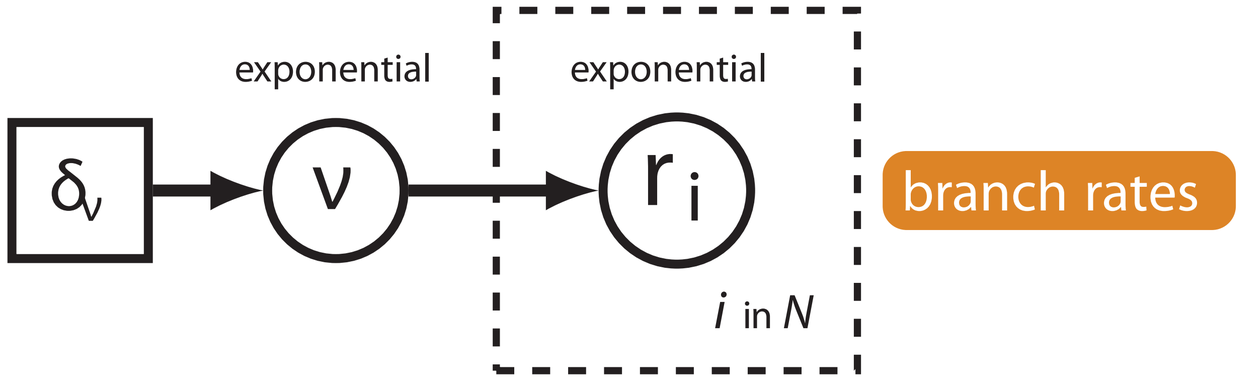
Graphical models: strict clock model





Graphical models: relaxed clock model



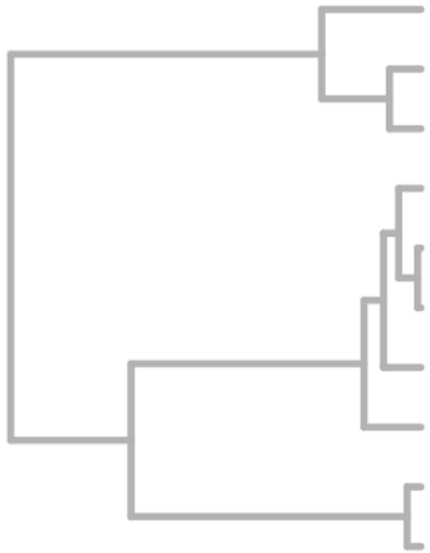


There are many different clock models

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

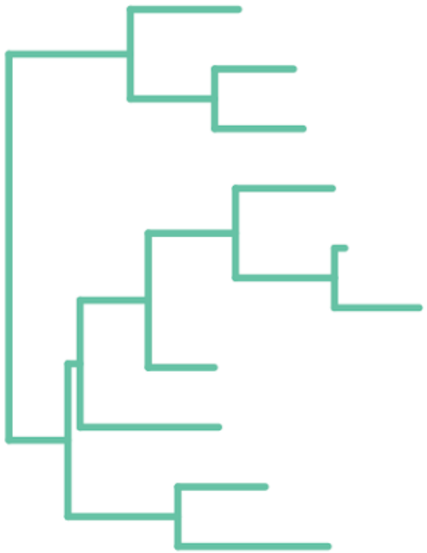
The changing face of the molecular evolutionary clock. *Ho*. 2014
See also: *Warnock, Wright*. 2020. Elements of Paleontology

Times and rates are not fully identifiable!



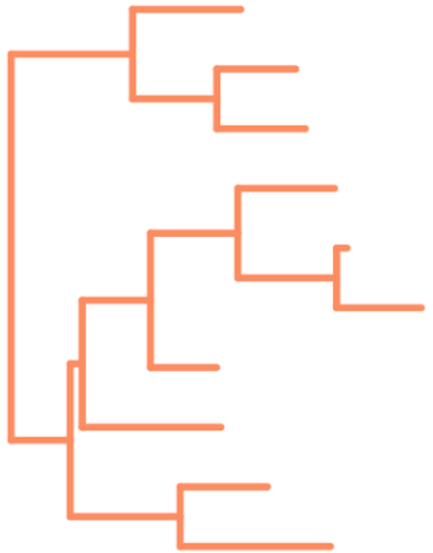
Time (year)

X



Clock rate (subs/site/year)

=



Genetic distance (subs/site)

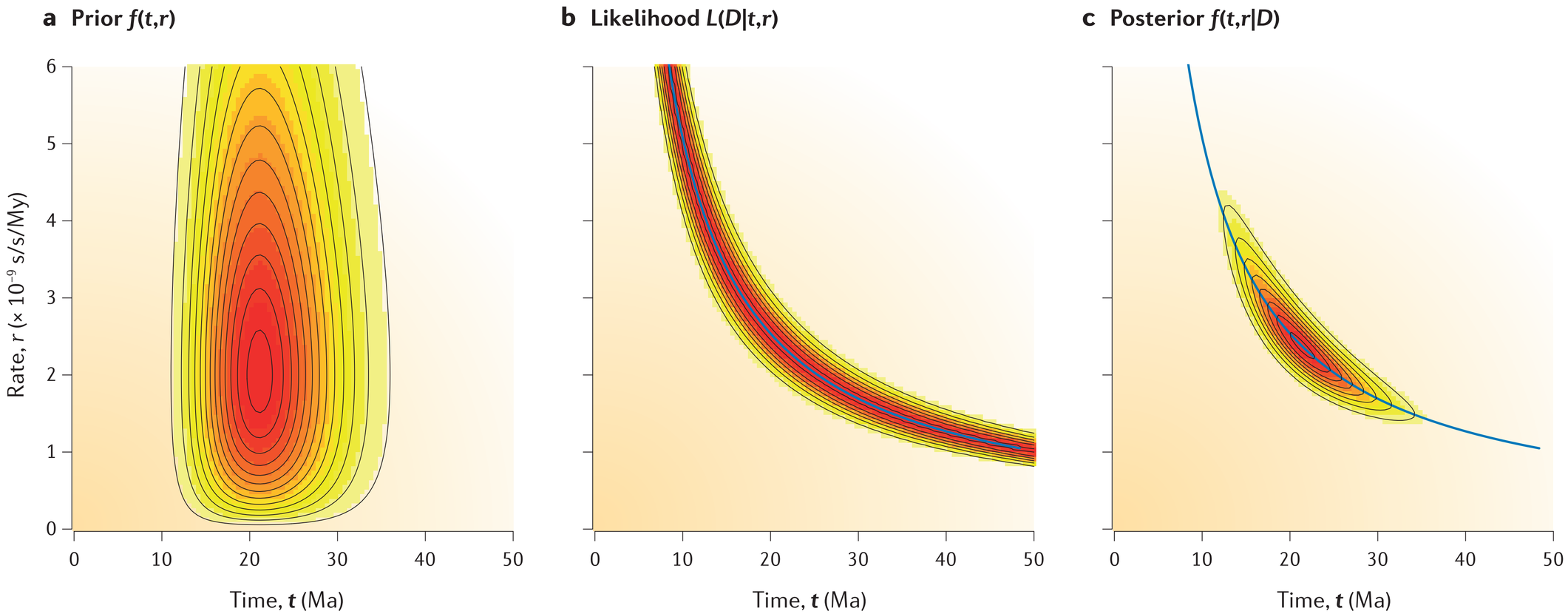
Prior

Prior

Likelihood

Slide adapted from Sebastian Duchene

The priors will always influence the results

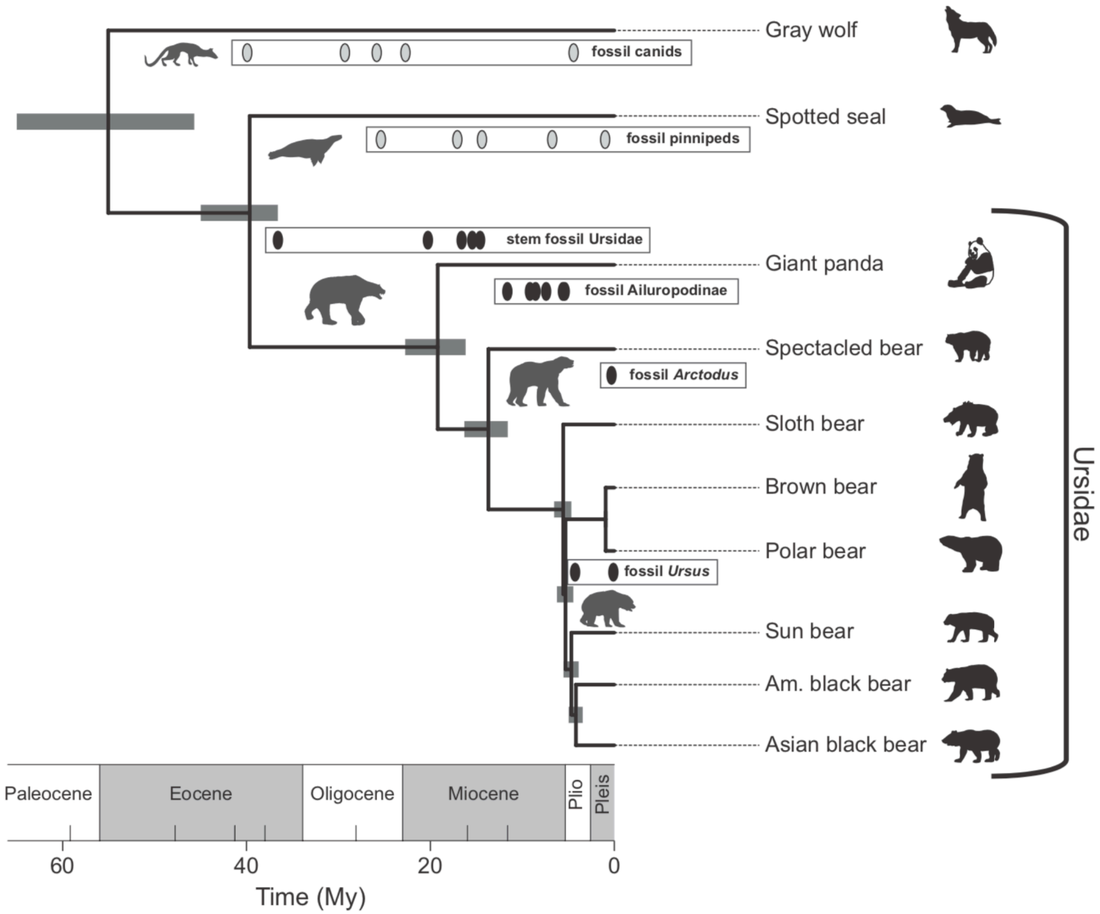


dos Reis et al. 2015. Nature Reviews Genetics

Exercise

Next objectives

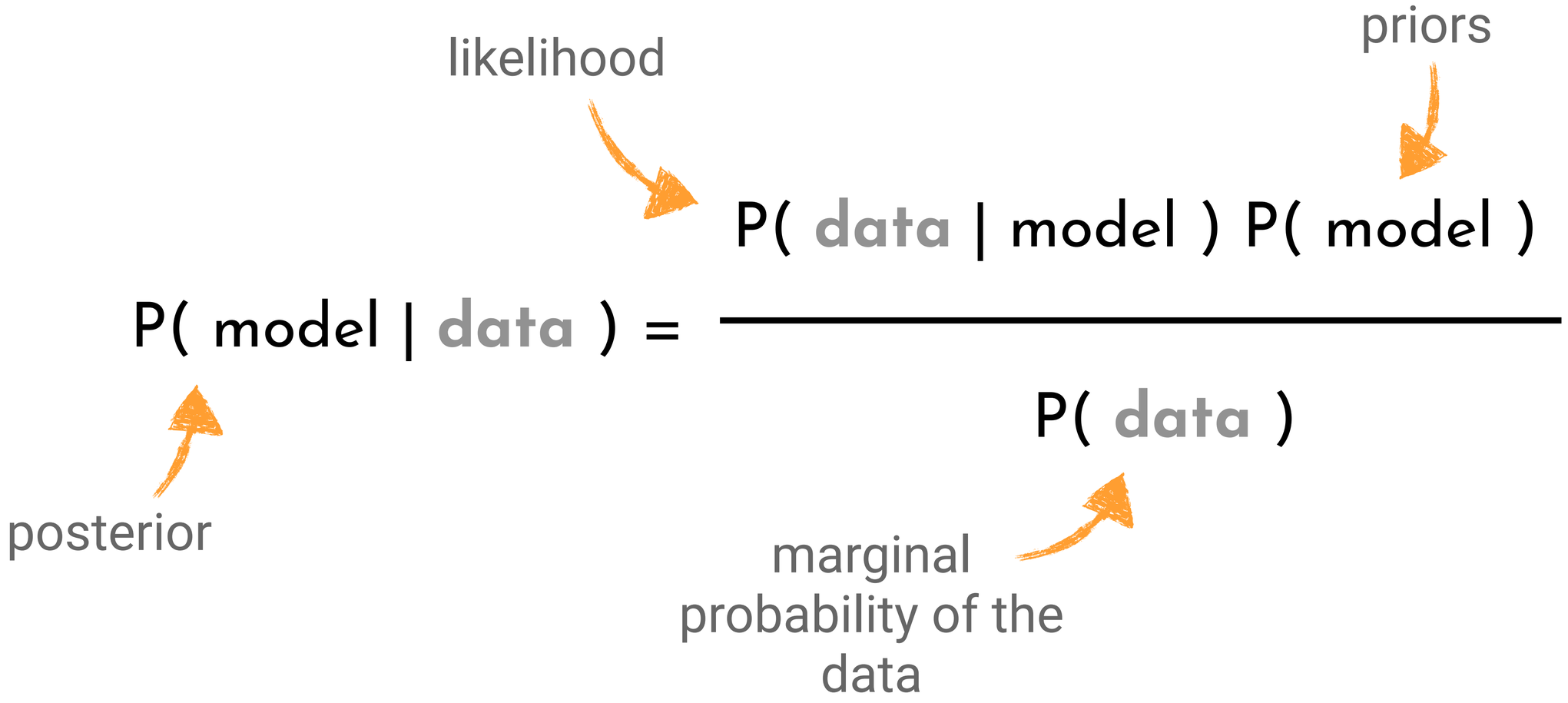
- **Recap**
  - Tripartite framework
- The fossilised birth-death process
- Total-evidence dating
- Phylodynamics



Bayesian divergence time estimation

Recap

We use a Bayesian framework



Bayesian divergence time estimation

The data

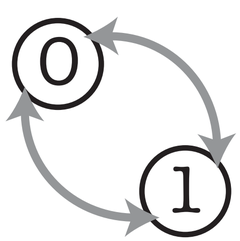
and / or

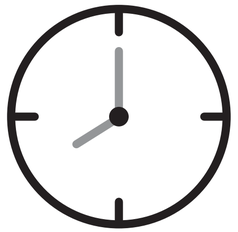
```
0101... ATTG...
1101... TTGC...
0100... ATTC...
```

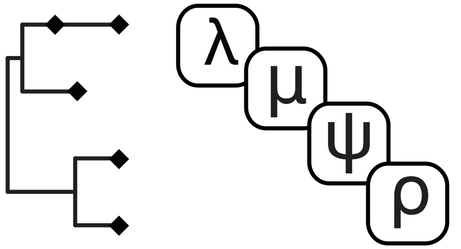
phylogenetics characters

sample ages

3 model components







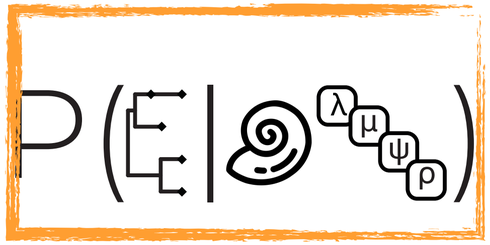
substitution model

clock model

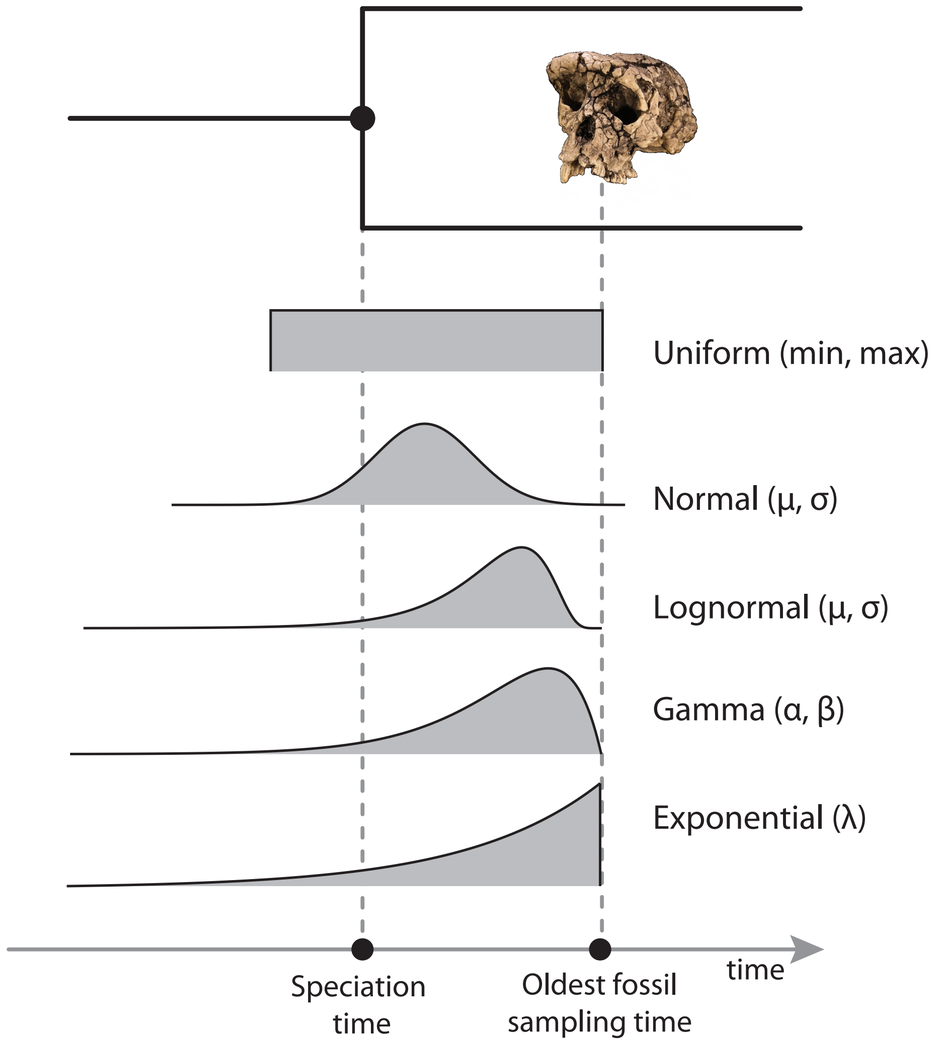
tree and tree model

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

probability of the time tree



Recap: Node dating



We can use a calibration density to constrain internal node ages

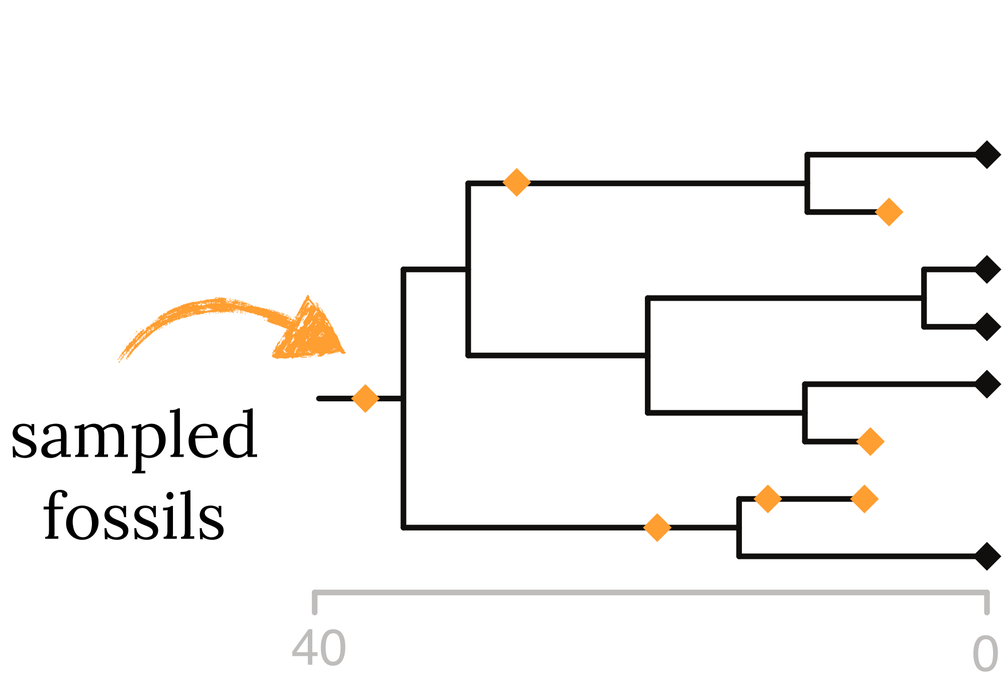
We typically use a birth-death process model to describe the tree generating process

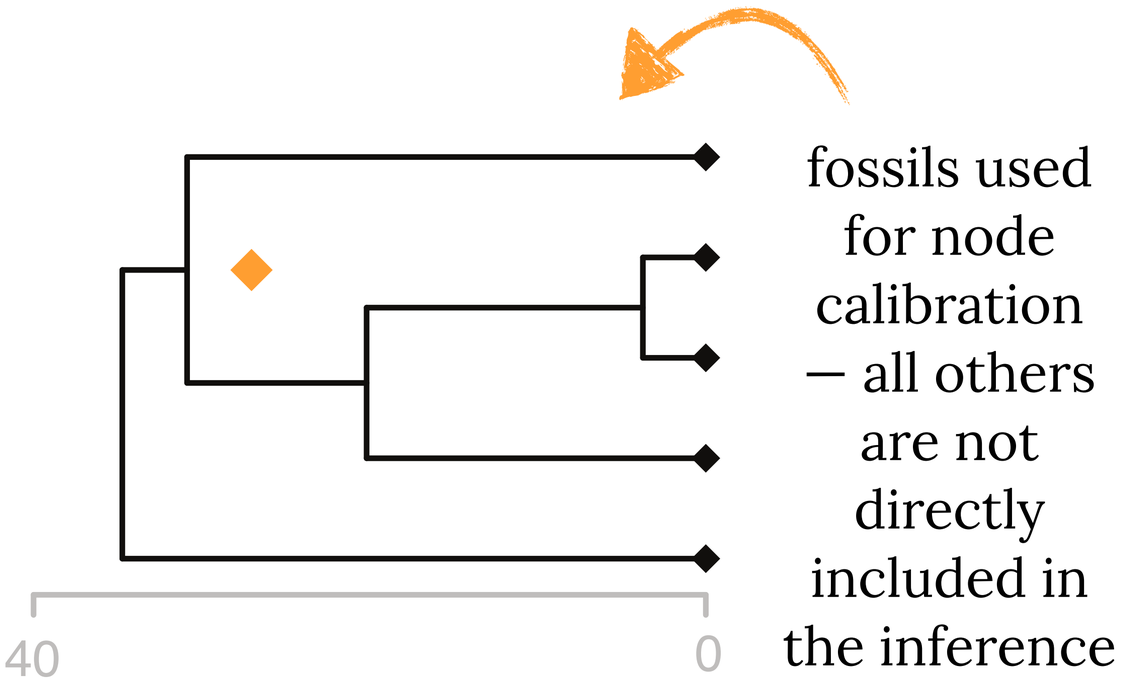
Adapted from *Heath* (2012). Sys Bio

Node dating: potential issues

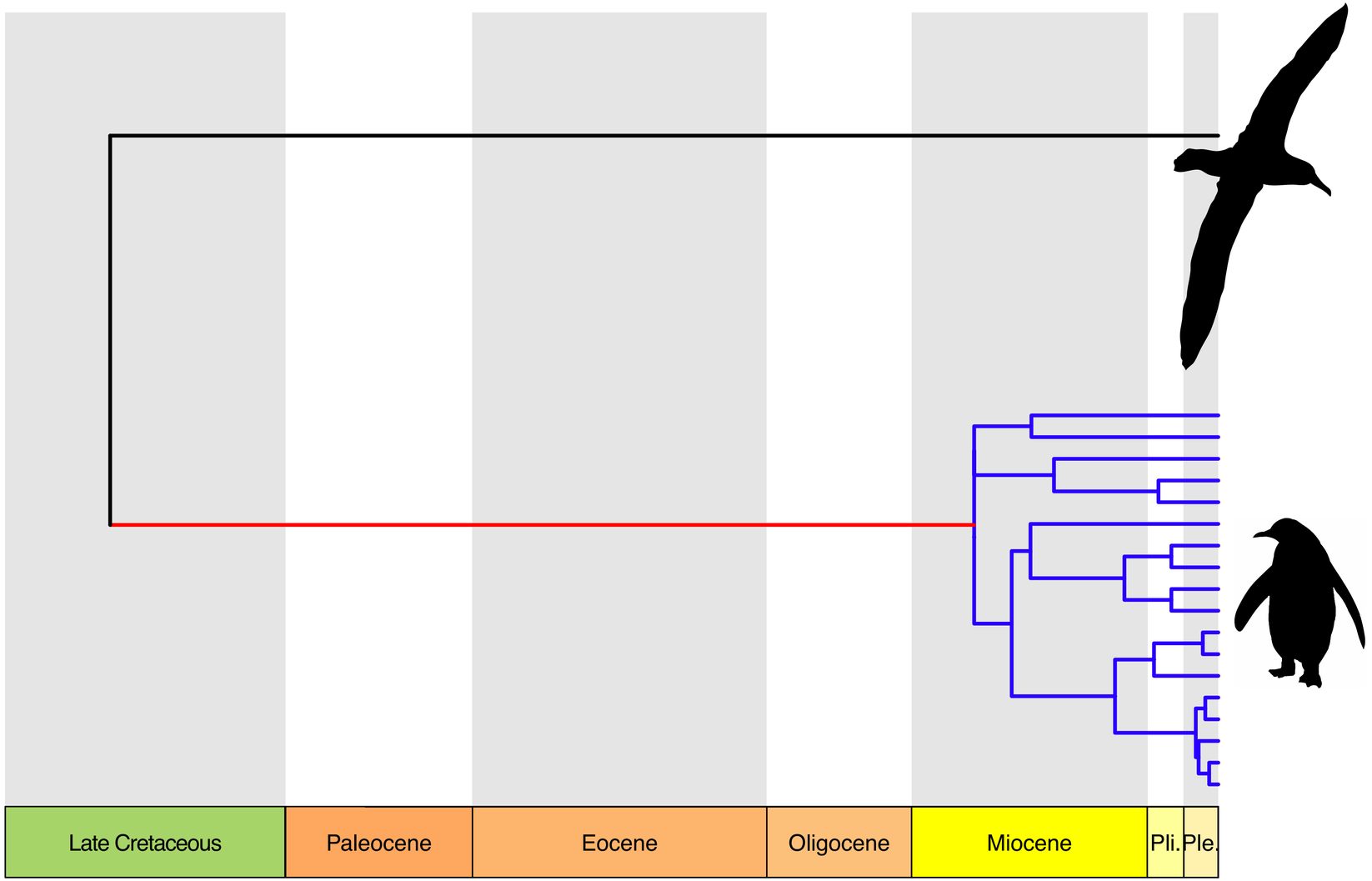
There are many!

A lot of information is excluded, since typically we assign one fossil per calibration node



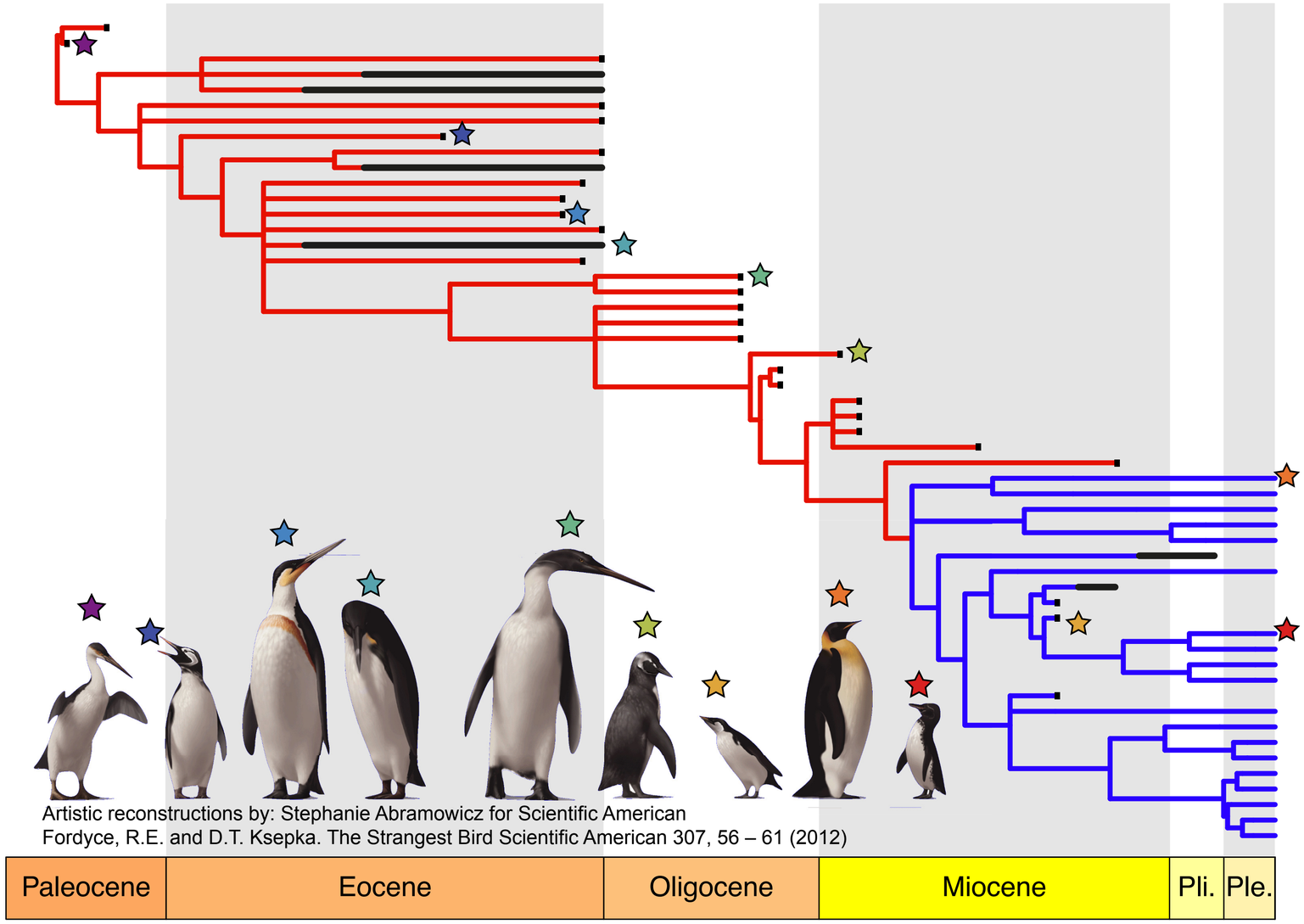


Example: living penguins



Nearest living relative is the group containing falcons - separated by ~60 Ma

Example: living penguins



Artistic reconstructions by: Stephanie Abramowicz for Scientific American
Fordyce, R.E. and D.T. Ksepka. The Strangest Bird Scientific American 307, 56 – 61 (2012)

But penguins have a rich fossil record!

Node dating: potential issues

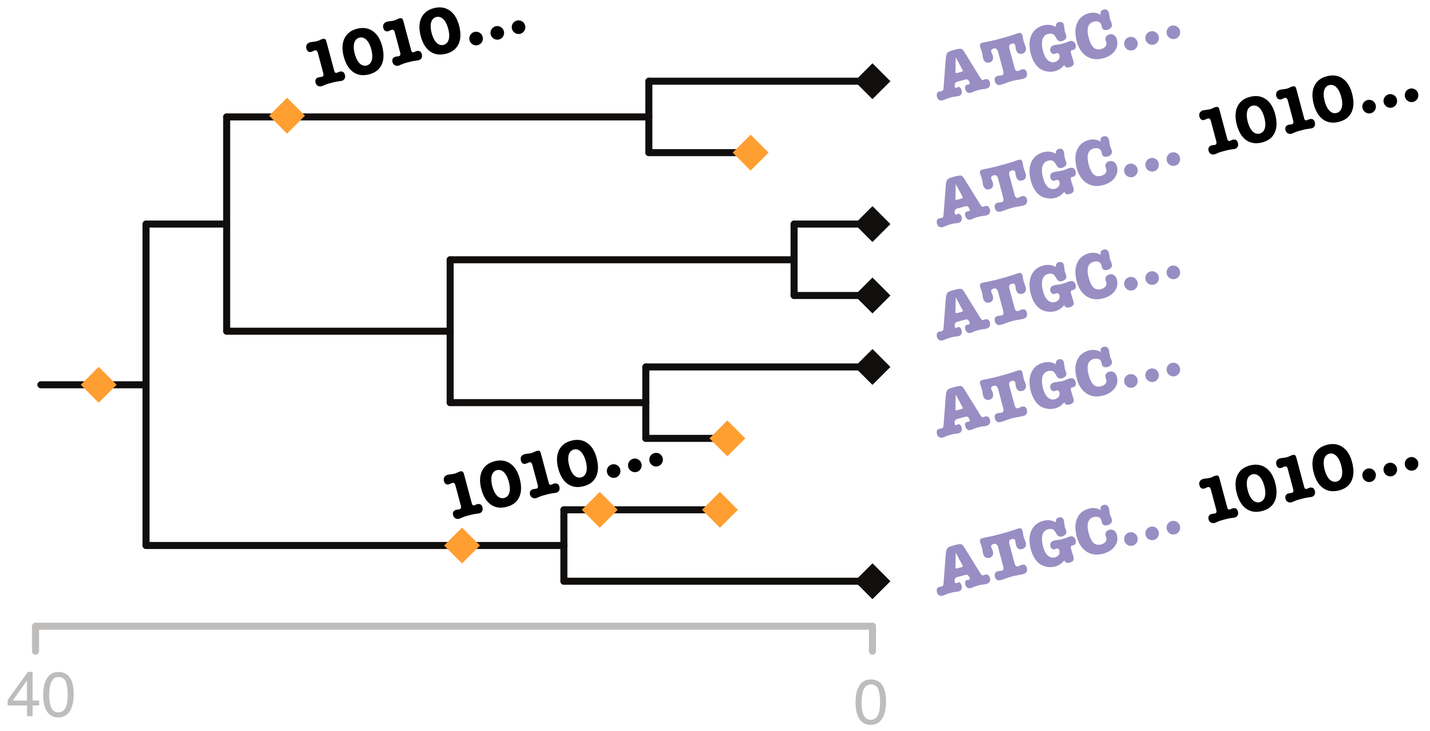
The model doesn’t describe the process that generated the fossil sampling times, meaning the model is **statistically incoherent**

The calibration priors are difficult to specify objectively and can have a massive impact on the divergence times. They can also interact with each other and / or the birth-death process prior in unintuitive ways

Some references on issues with specified vs effective priors
Yang and Rannala. 2006. MBE
Heled and Drummond. 2012. Sys Bio
Warnock et al. 2012, 2015

Total-evidence dating

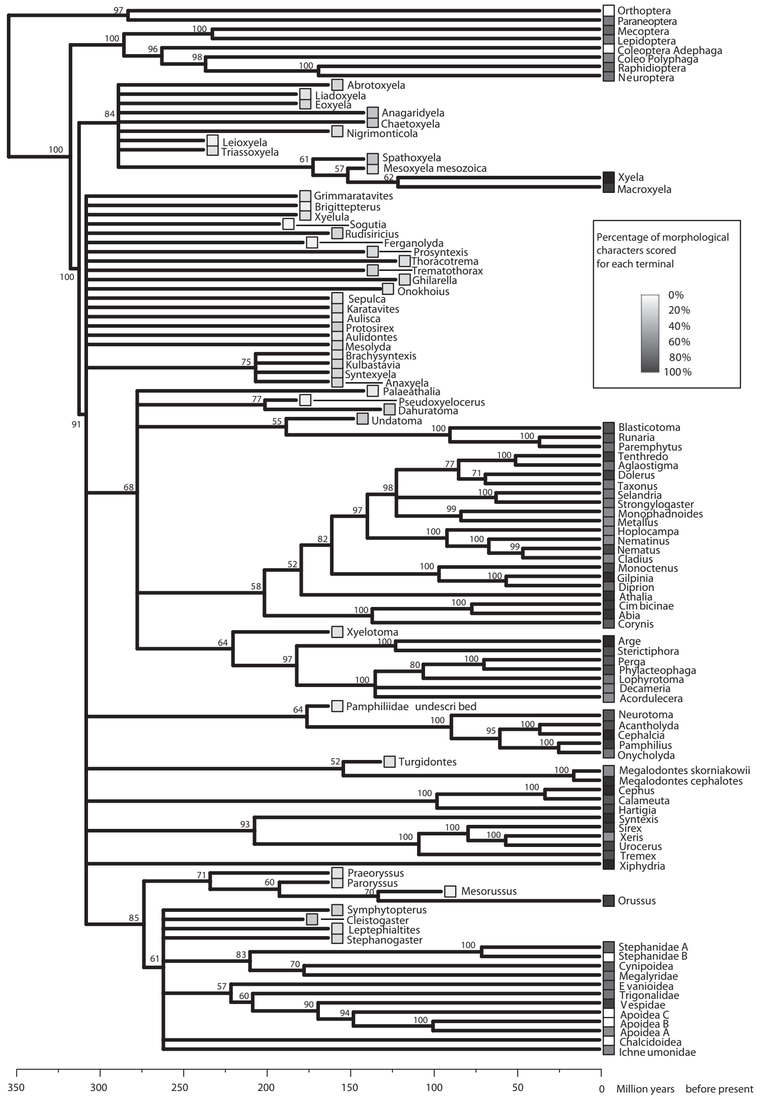
Tip-dating or “total-evidence” dating



We have DNA for living species. We have morphology for living *and* fossil species

Fossils can be positioned on the basis of morphology
→ accounts for uncertainty in fossil placement

The uniform tree prior

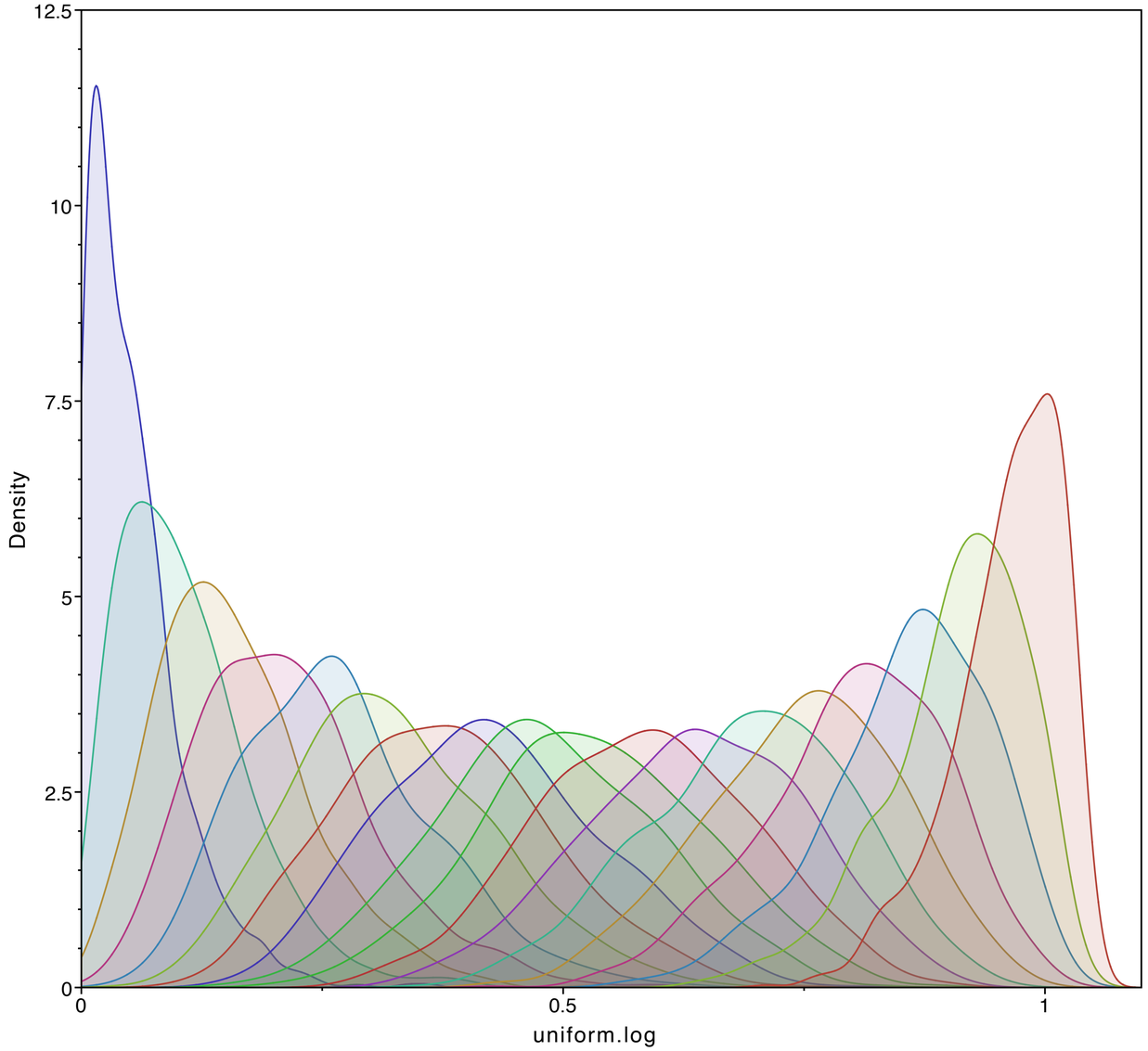


Dated tree of Hymenoptera

The uniform tree prior assumes all trees and branch lengths are equally likely within the bounds of the fossil ages (+ a max upper bound)

It does not explicitly account for the fossil sampling process

Ronquist et al. 2012 Sys Bio



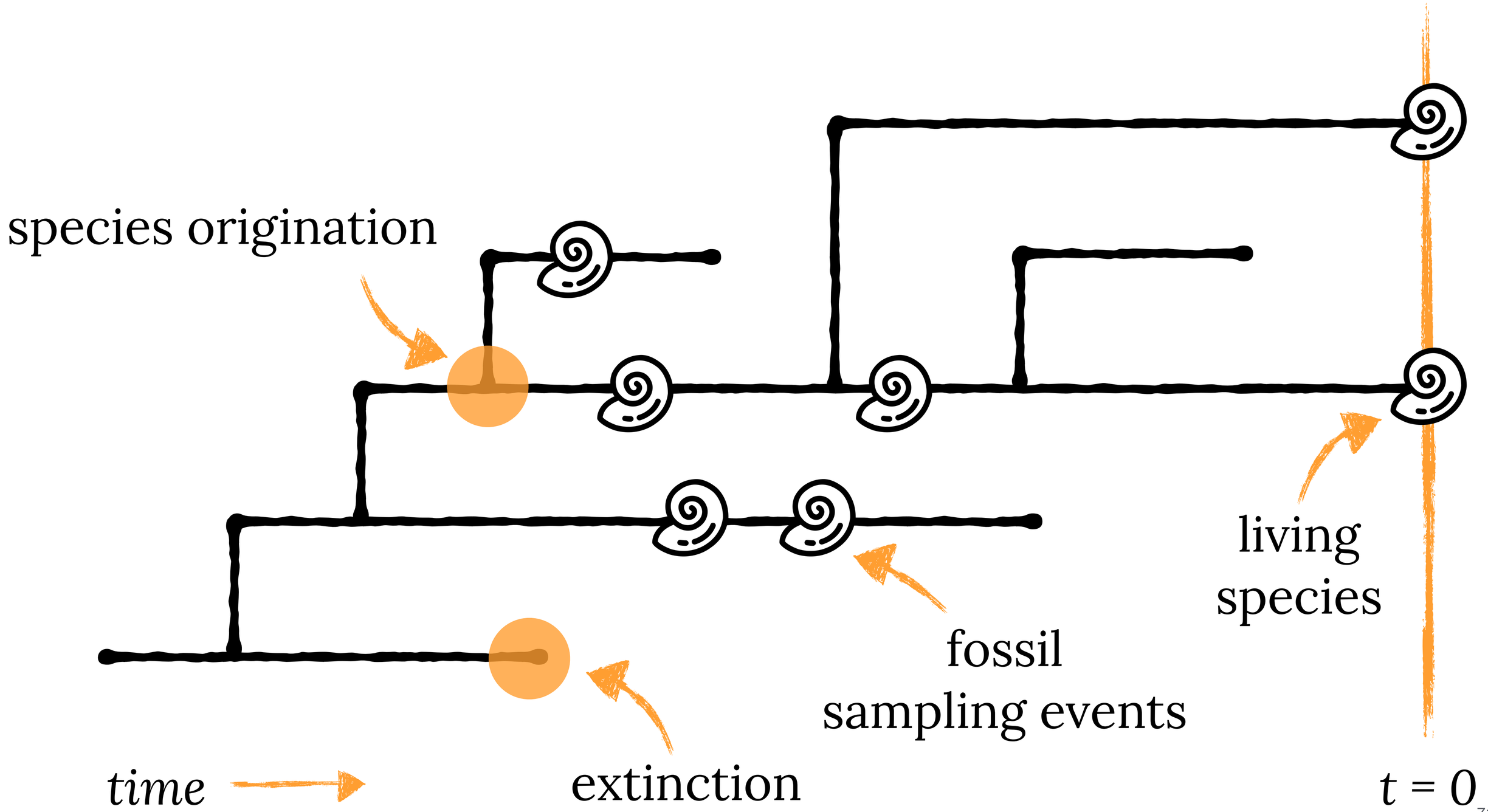
A uniform tree prior implies time till the next split is independent of how many lineages there are present

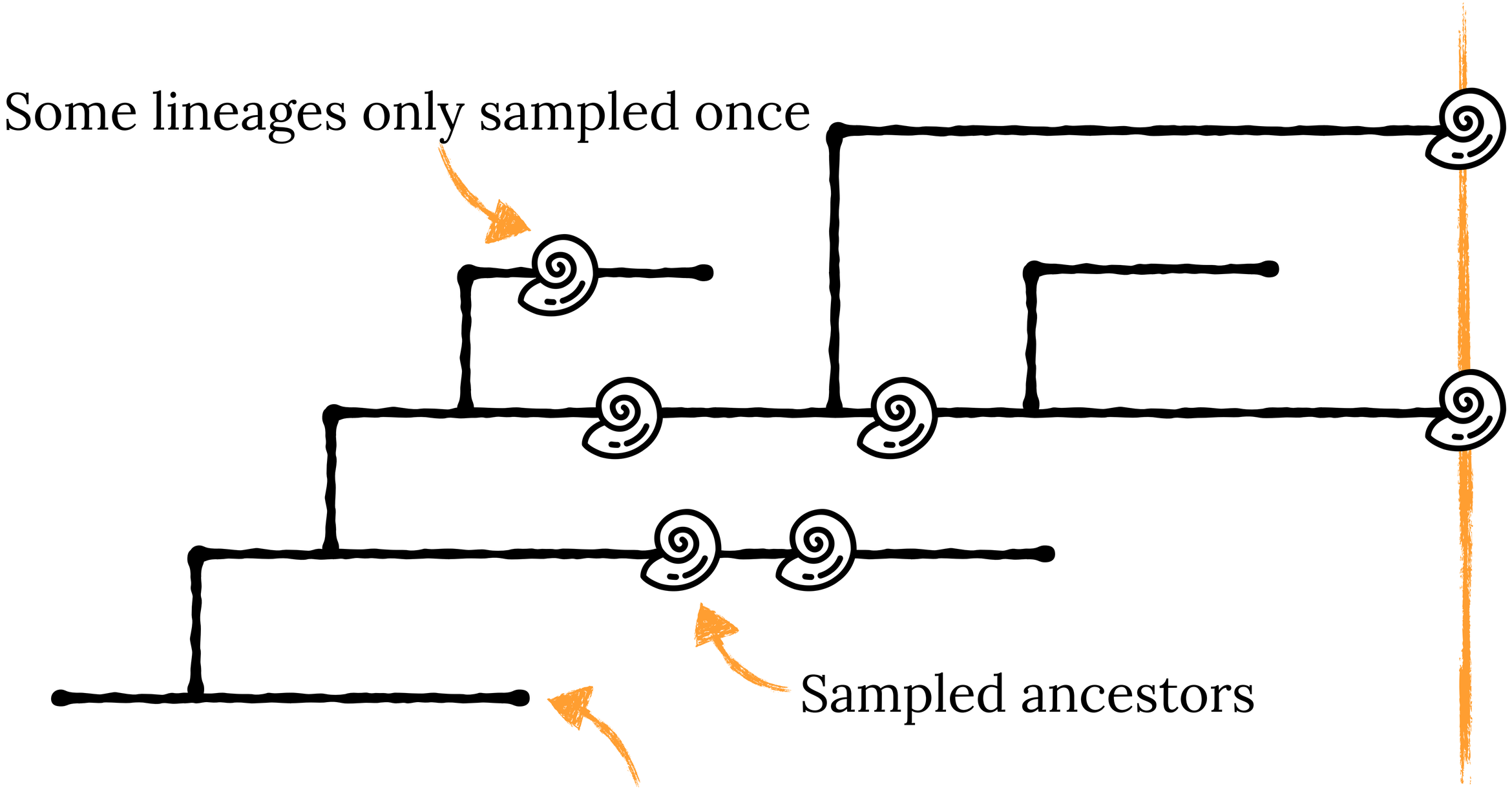
This is in contrast to birth-death processes, where more lineages mean a higher chance of observing a split in one of these lineages

[Uniform tree priors - why not use them?](#)
Remco Bouckaert

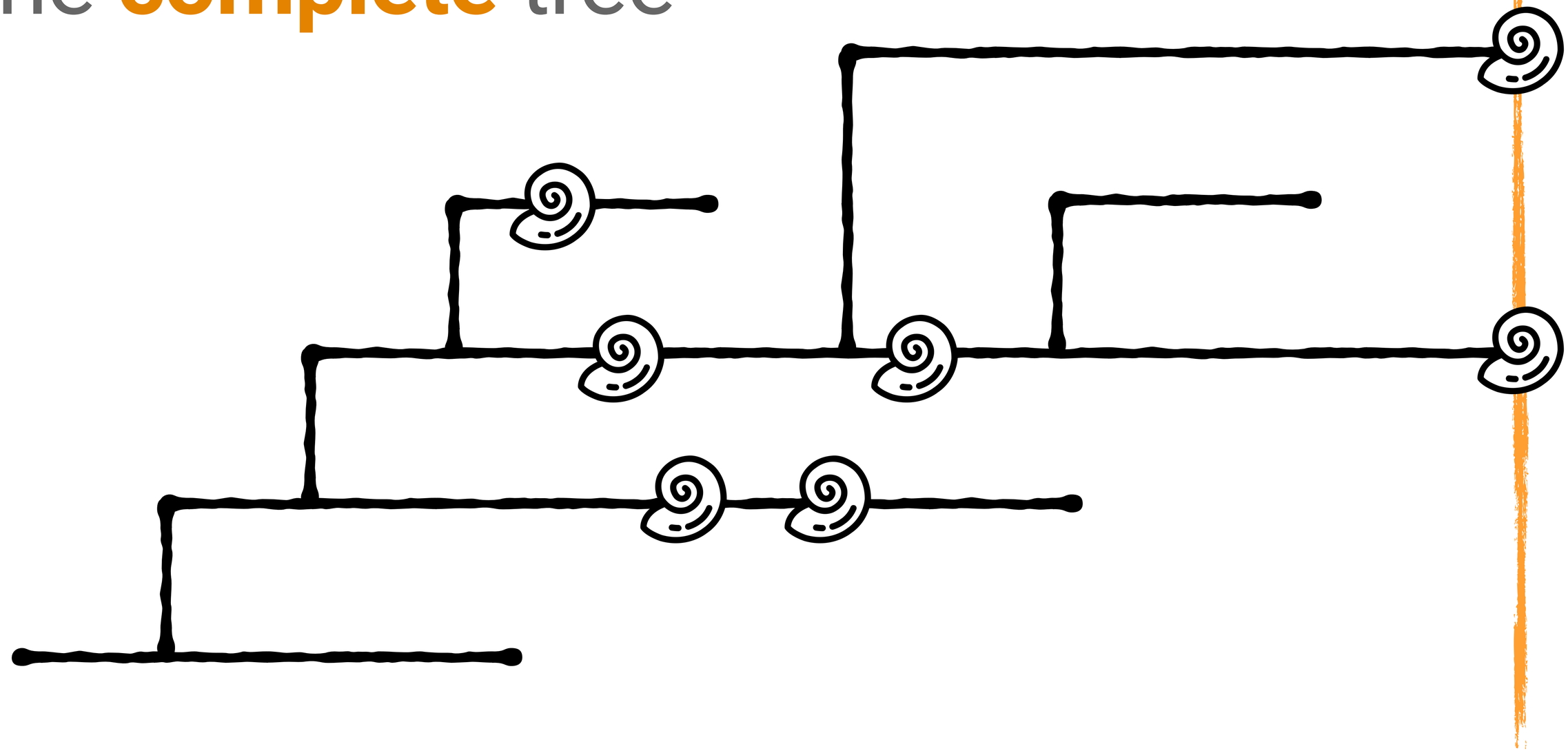
What does a generating prior for the fossil record look like?

The fossilised birth-death process


species origination
living species
fossil sampling events
extinction
time
t = 0


Some lineages only sampled once
Sampled ancestors
Some lineages go completely unsampled
$t = 0$

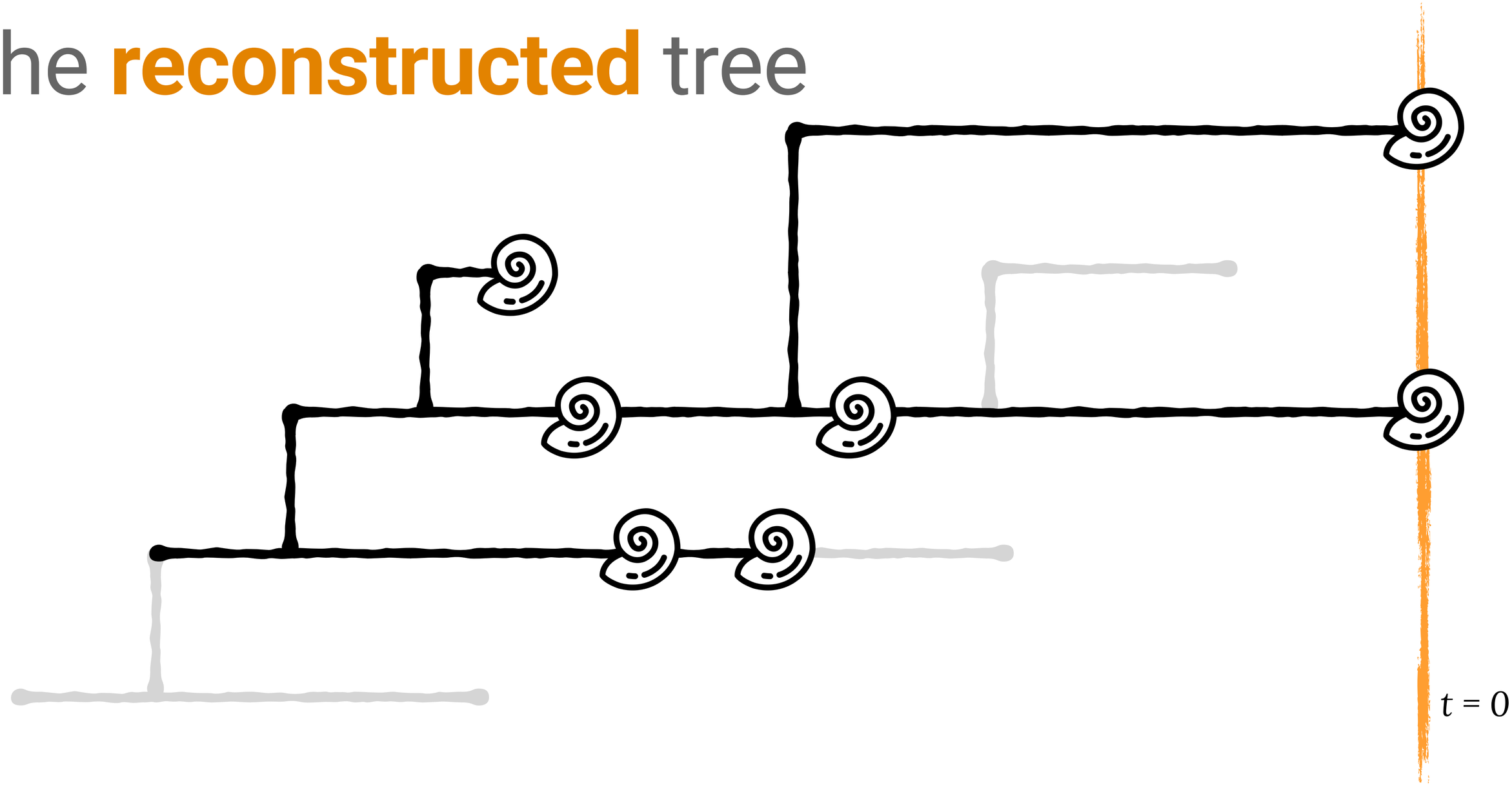
The **complete** tree



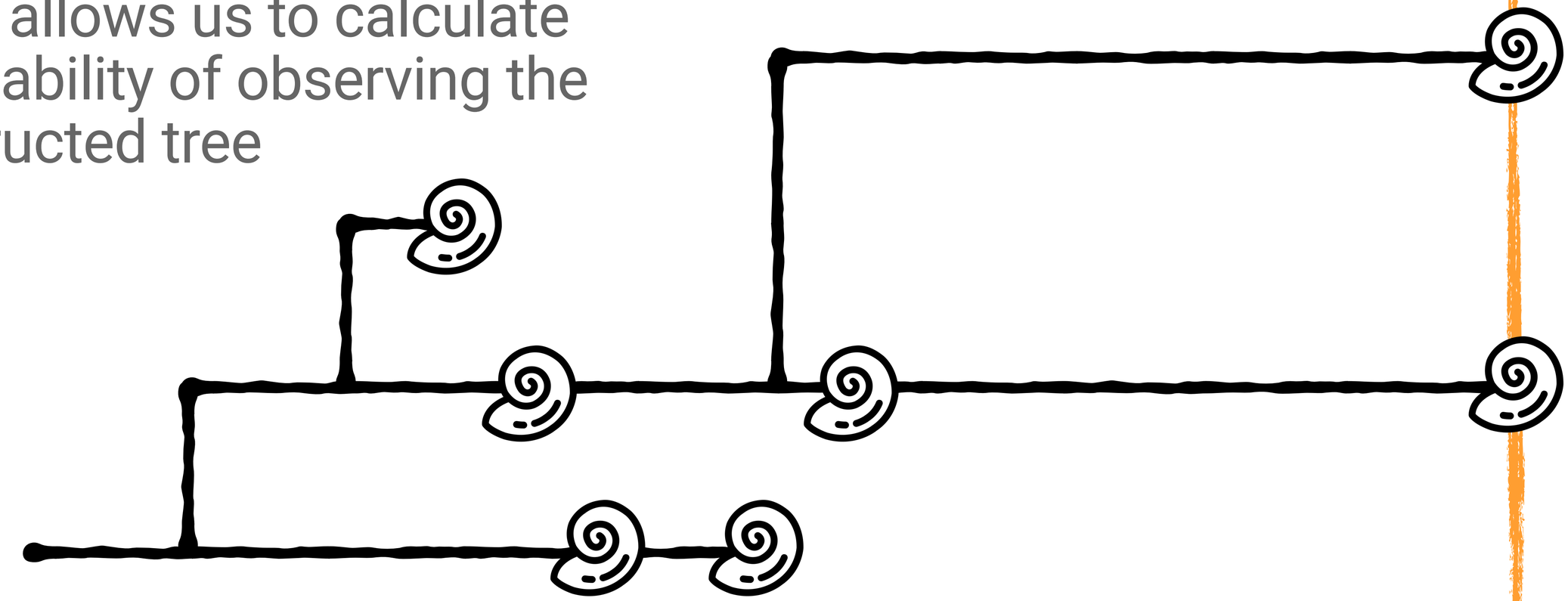
$time$

$t = 0$

The **reconstructed** tree



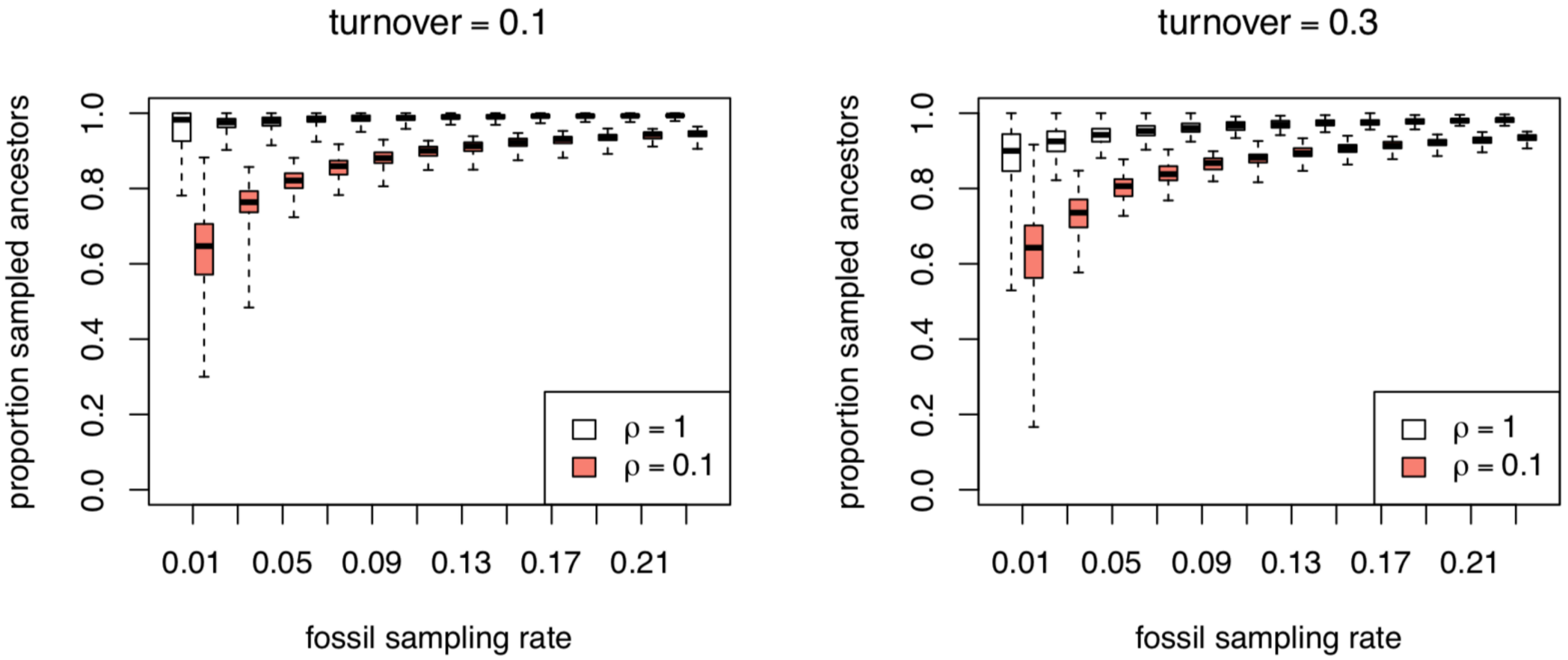
The **fossilised birth-death (FBD) process** allows us to calculate the probability of observing the reconstructed tree



Sampling-through-time in birth-death trees. Stadler. (2010)
First implemented: Heath et al. (2014) and *Gavryushkina et al.* (2014)

Sampled ancestors

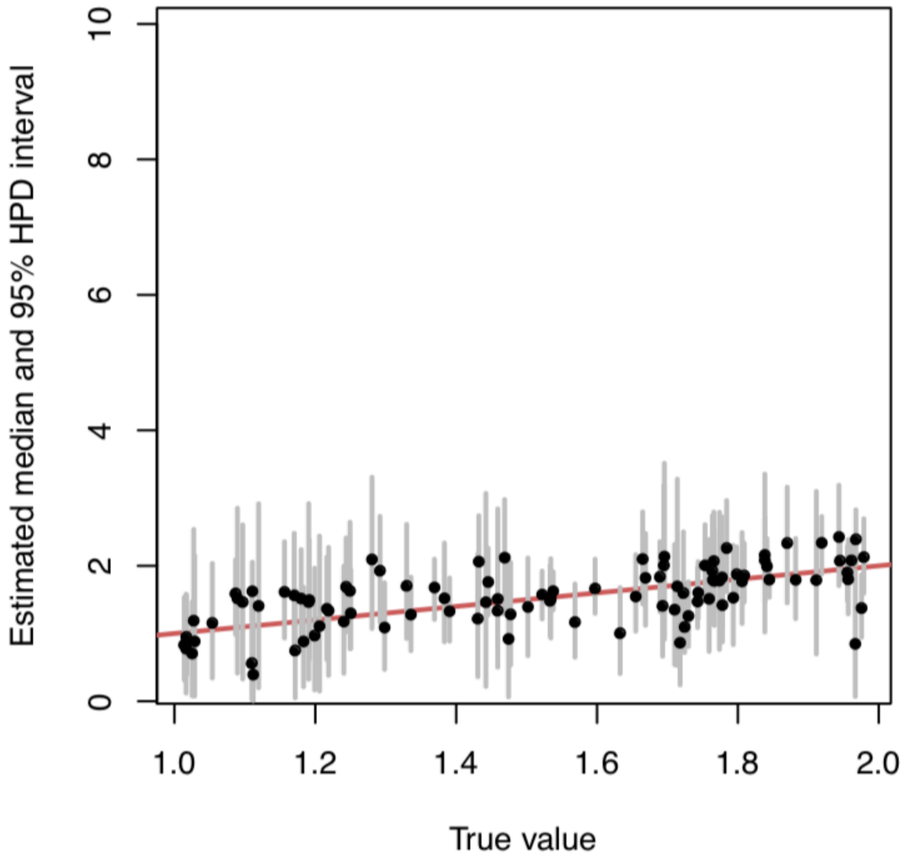
The proportion increases with higher turnover (birth - death) or higher sampling

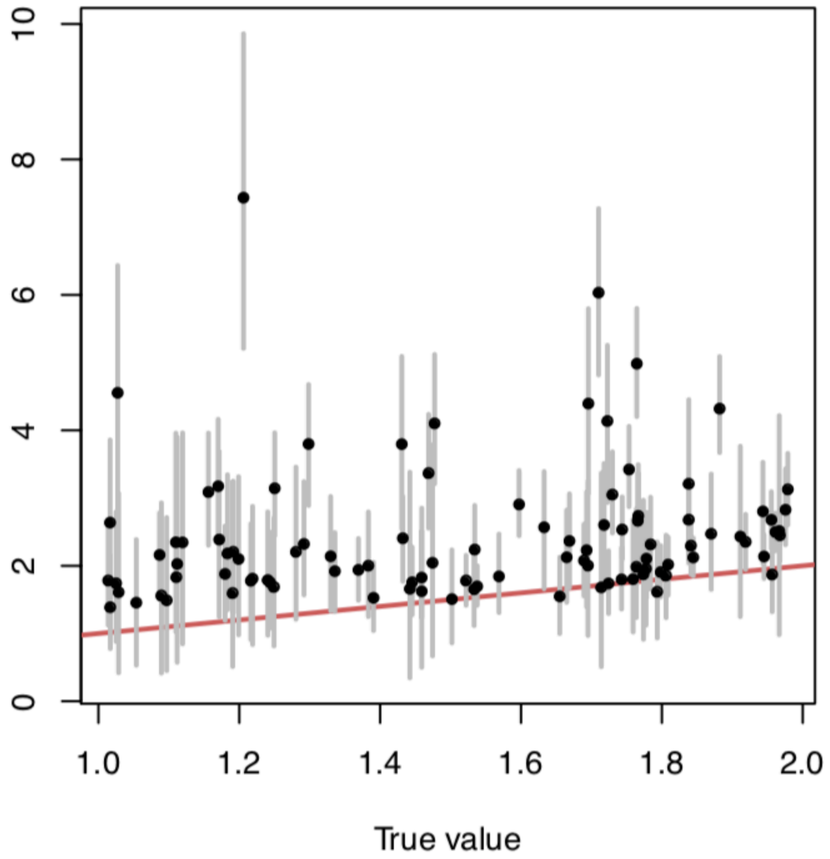


Walker, Heath. 2020. Phylogenetics in the Genomic Era

Sampled ancestors

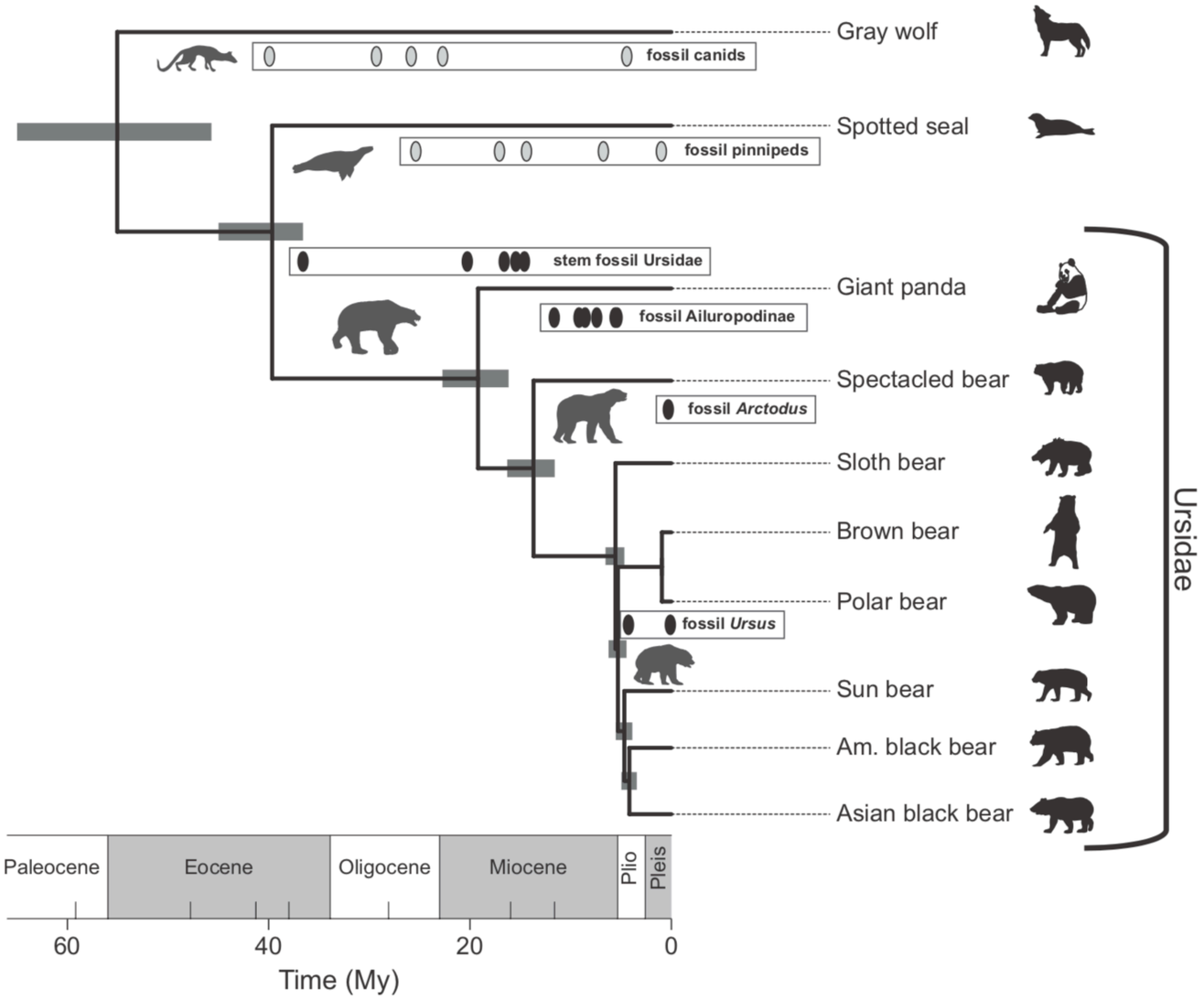
Ignoring sampled ancestors can lead to inaccurate parameter estimates





Gavryushkina et al. 2014 PLoS Comp Bio

Time calibrated tree of living and fossil bears



First application of the FBD model.

Fossils are incorporated via constraints, not character data. Their precise placement can not be inferred, but this uncertainty will be reflected in the posterior

Heath et al. 2014. PNAS

Exercise

Fossils can be incorporated via **taxonomy** or **character data** (total-evidence)

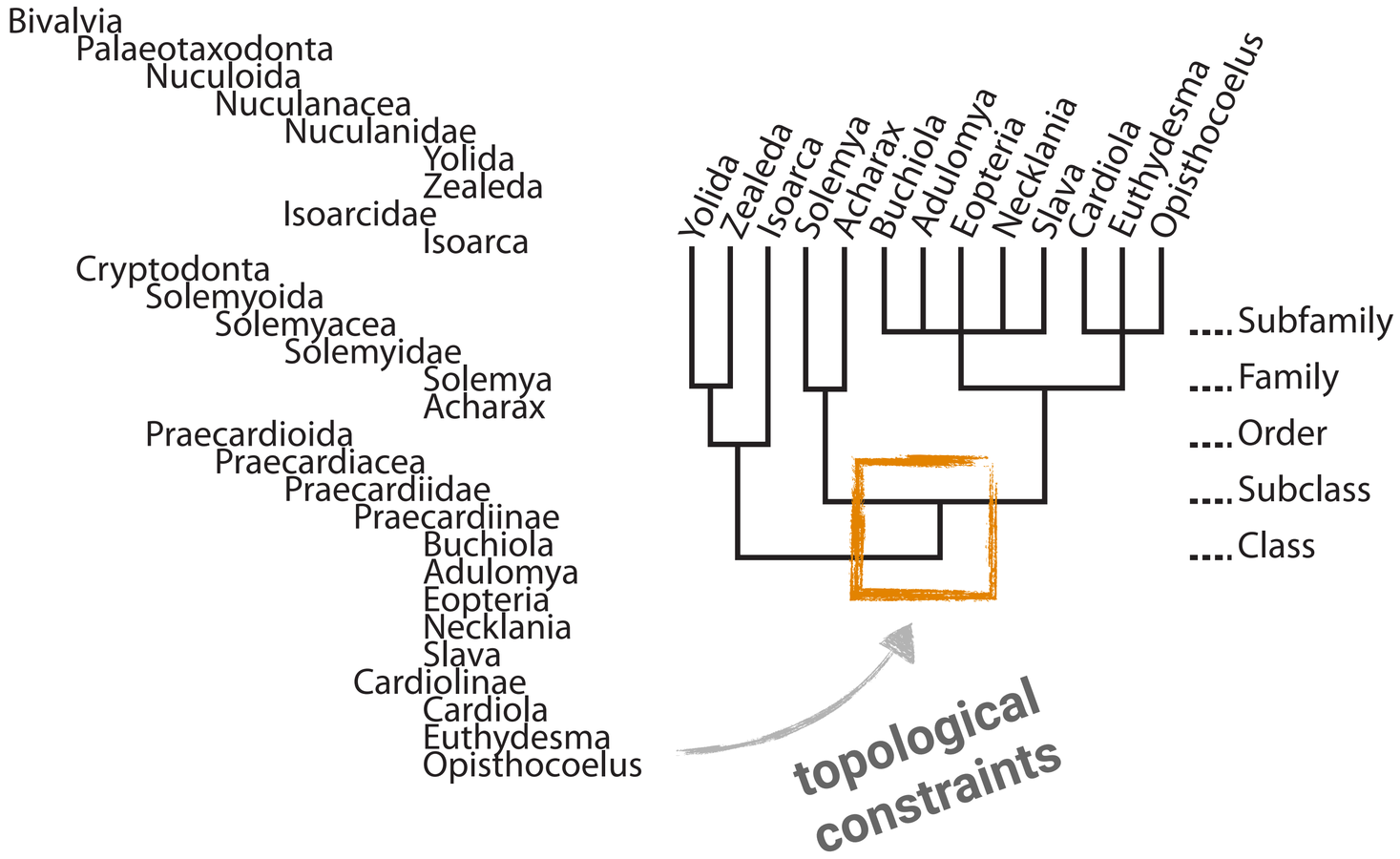
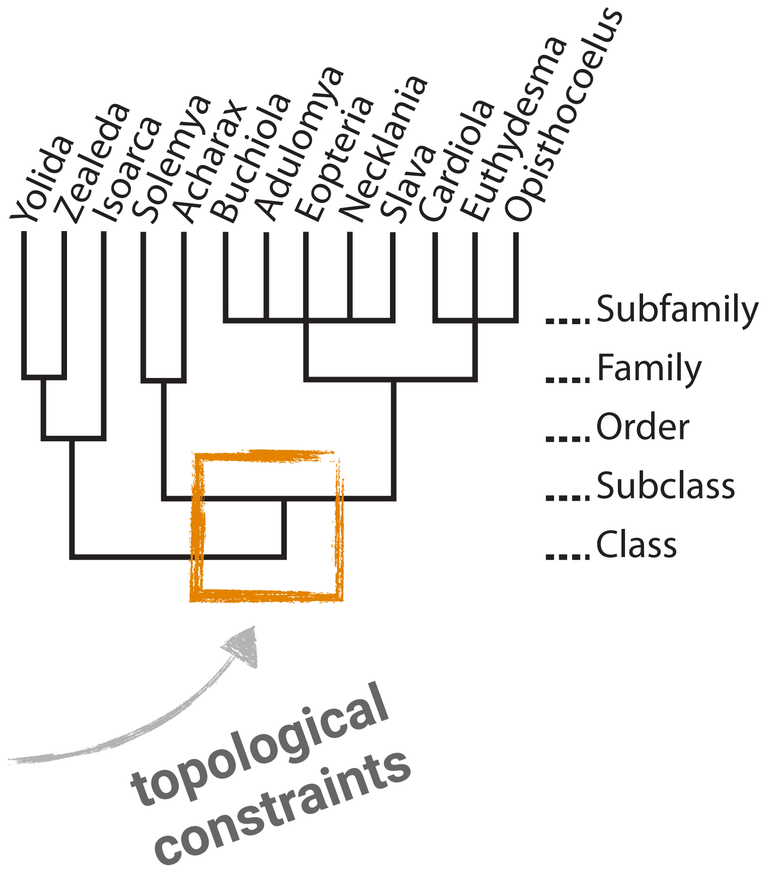


Image source Soul & Friedman (2015)

Fossils can be incorporated via **taxonomy** or **character data** (total-evidence)



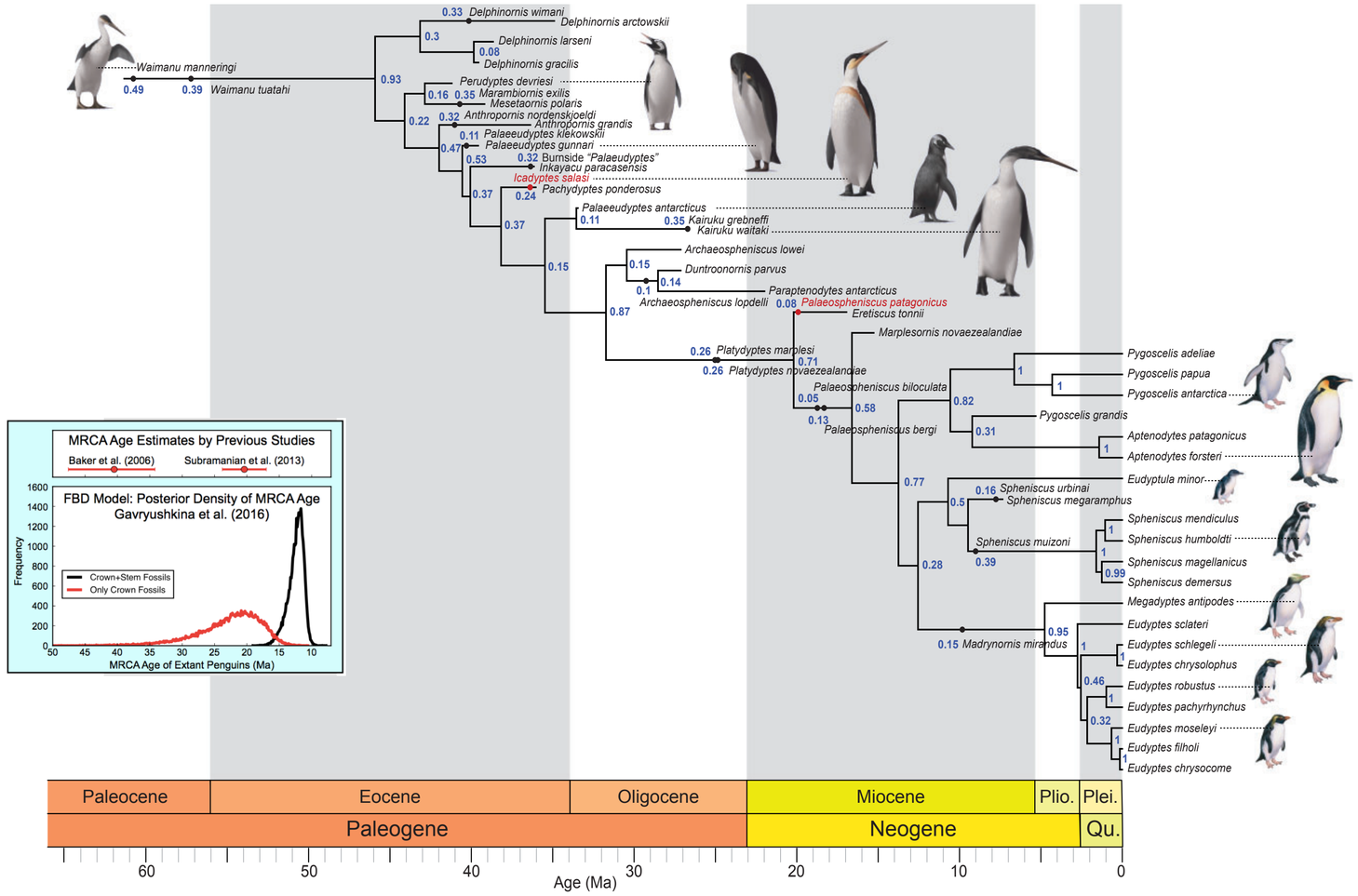
ATAT...
TCAC...
????...

OR

1001...
1101...
0100...

Image source Soul & Friedman (2015)

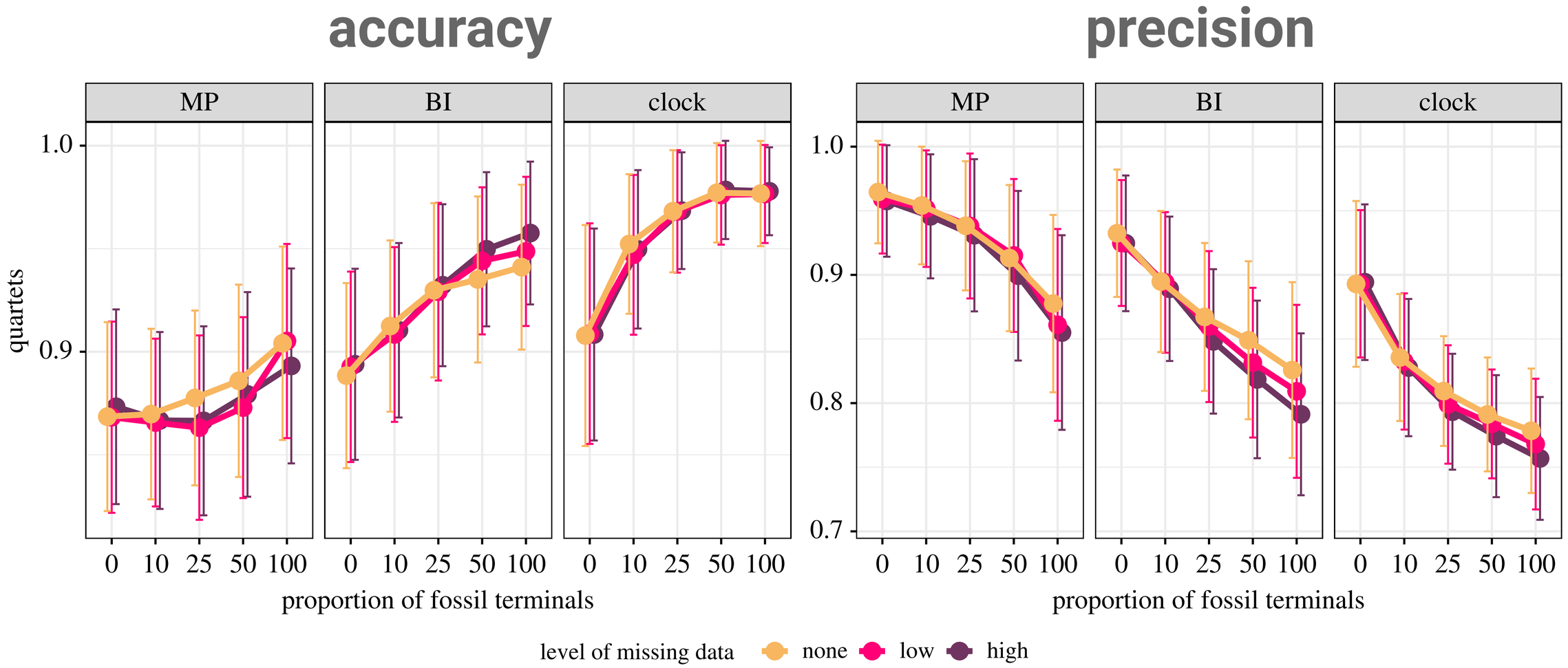
Time calibrated tree of living and fossil penguins



First application of total evidence dating using the FBD model

Fossils are incorporated using character data

Gavryushkina et al. (2016)



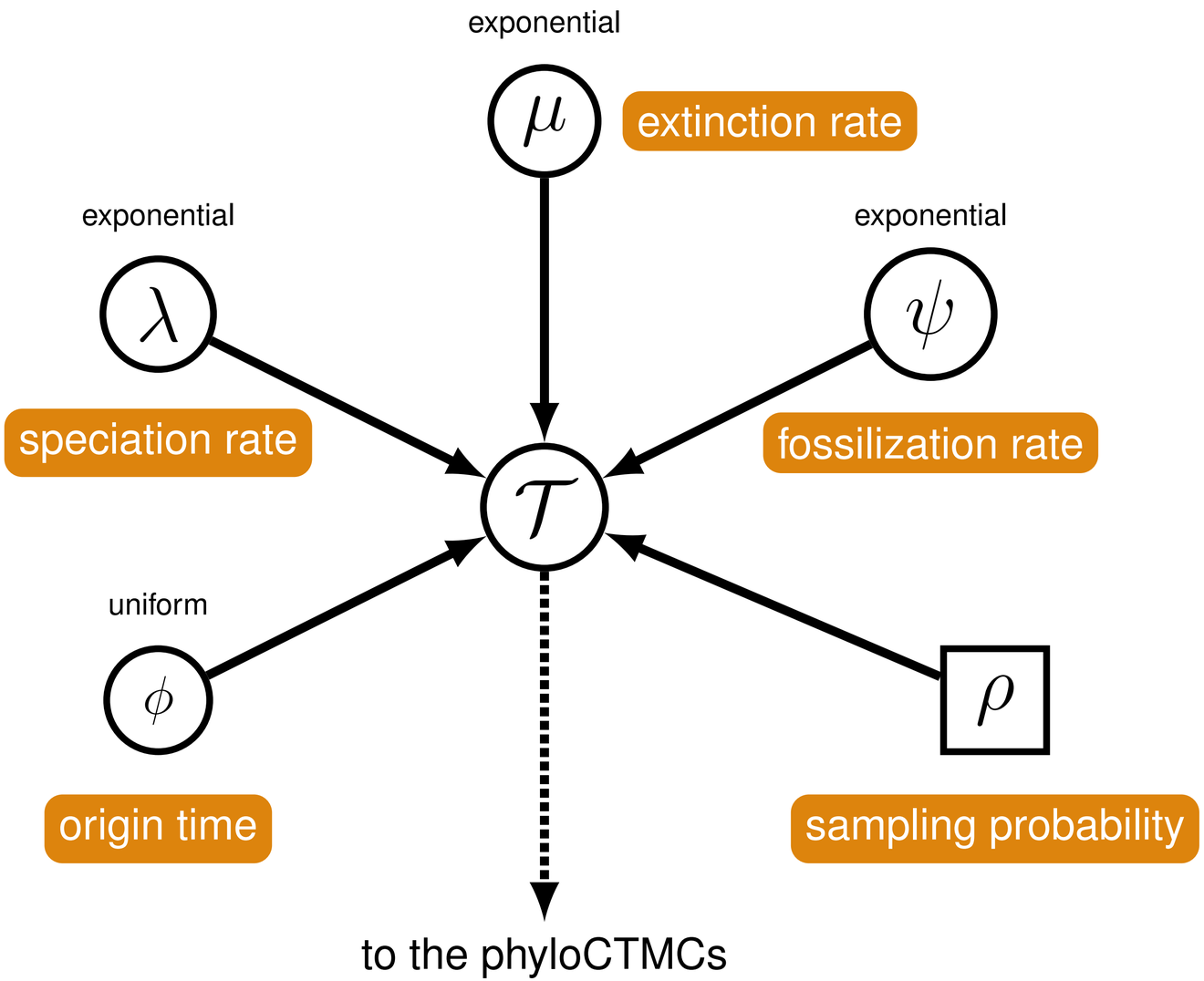
Fossils improve phylogenetic analyses of morphological characters
Koch, Garwood, Parry. 2020. Proc B

Some notes

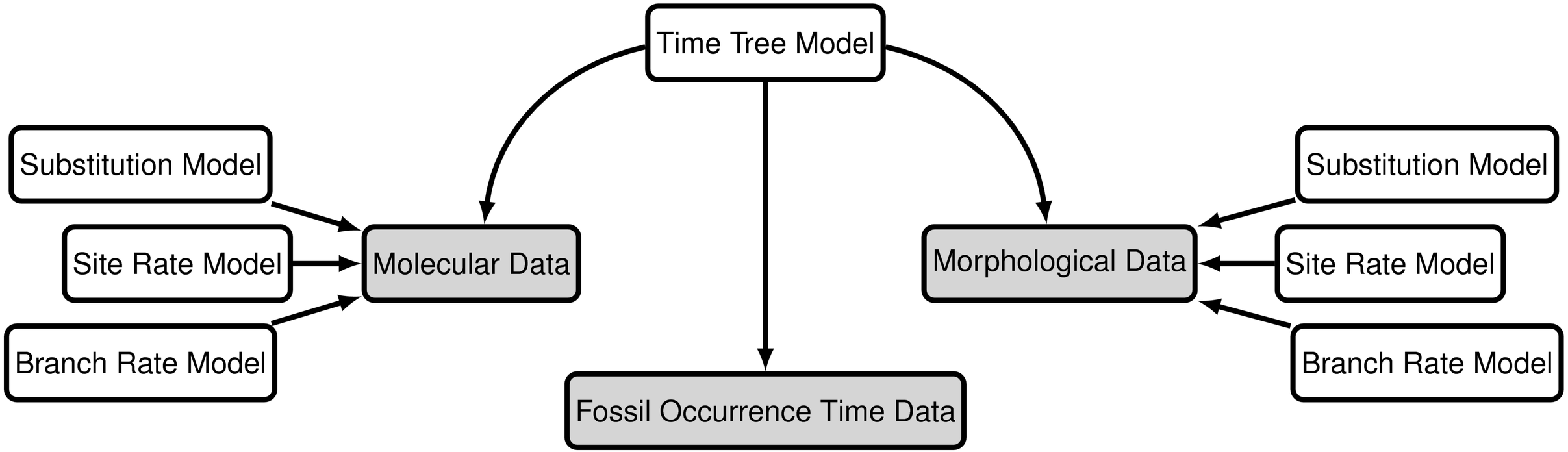
- The topology of extant taxa is largely unaffected by how fossils are incorporated
- Fossils *and* age information help inform topology
- Divergence times are much more sensitive to errors in fossil placement and model misspecification
- Total-evidence dating is more robust to model misspecification

Barido-Sottani et al. (2023) *Putting the F in FBD analyses: tree constraints or morphological data? Palaeontology*

Graphical model representation of the FBDP



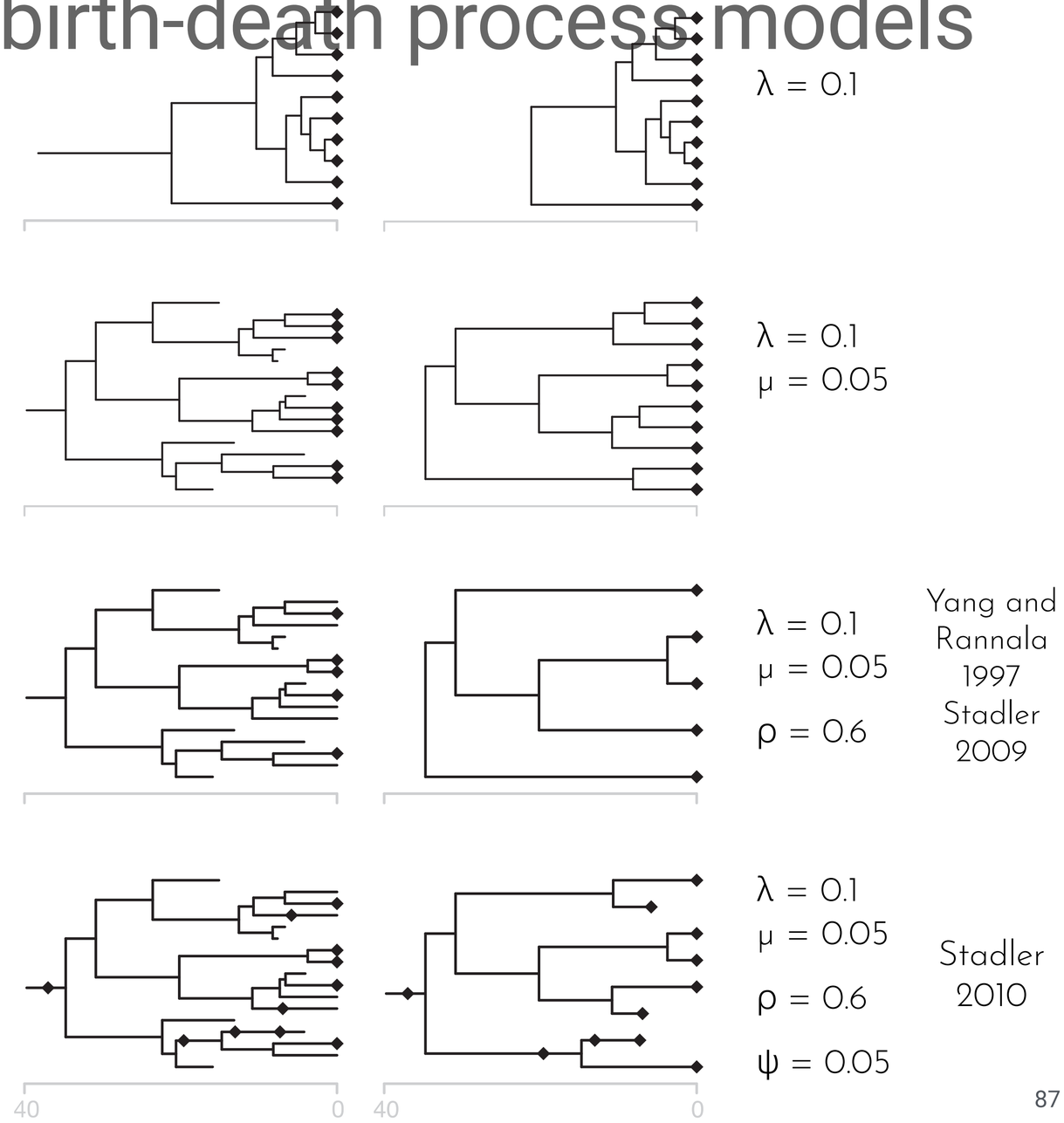
Bringing everything together



complete vs. reconstructed trees

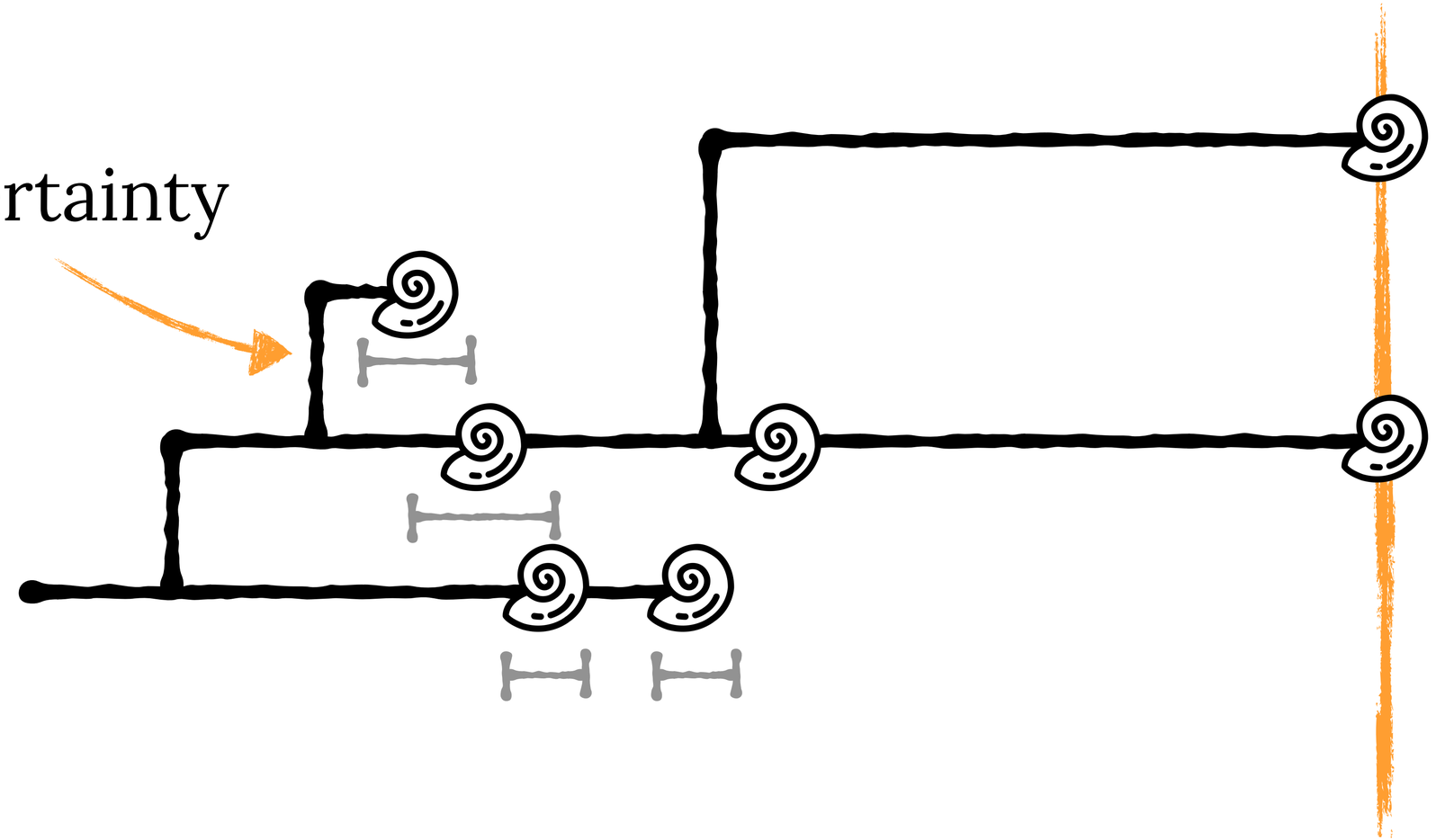
Relationship to (some) other birth-death process models

These models are special cases of the FBD process, with fossil sampling (ψ) = zero.

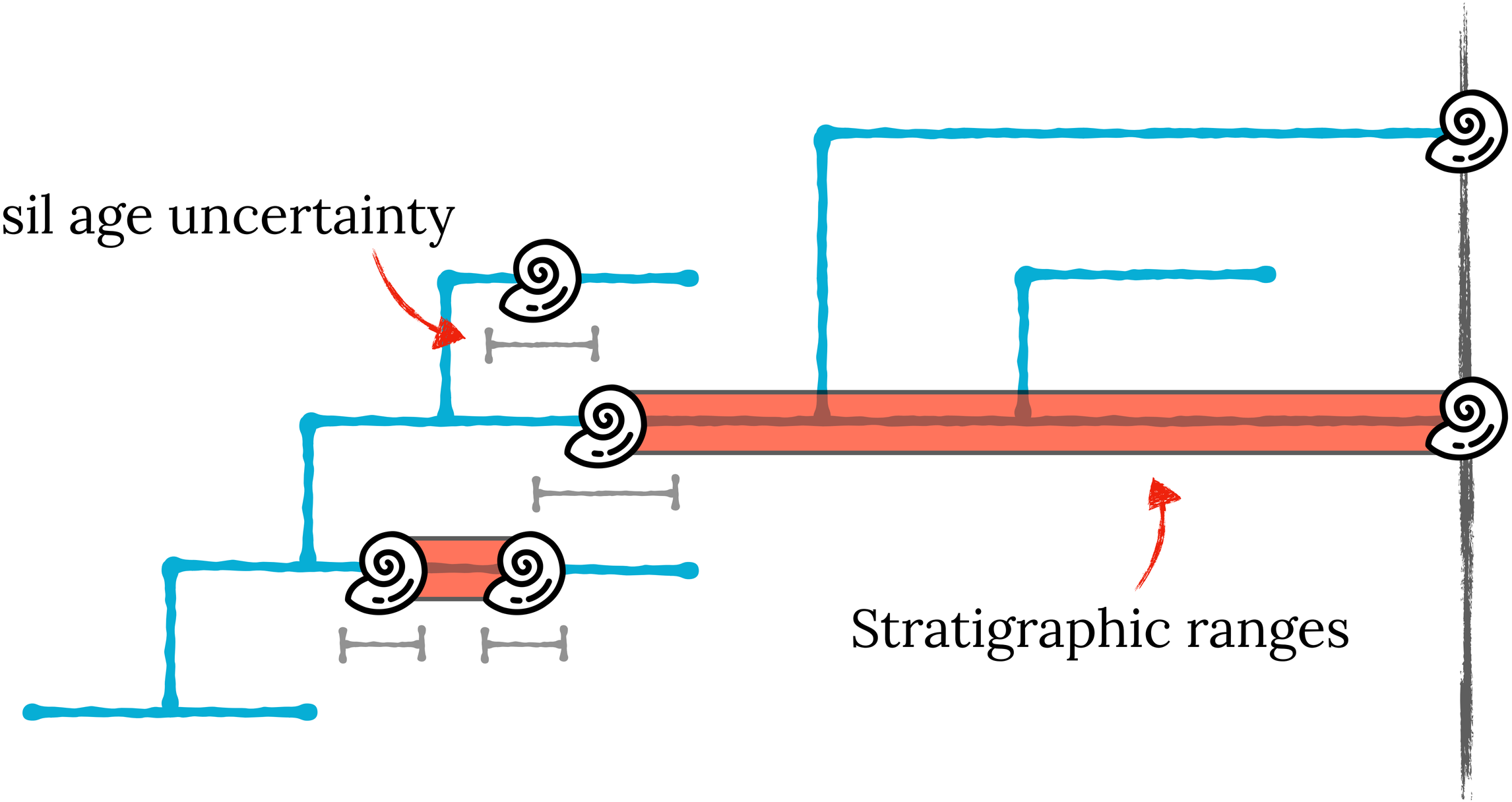


Stadler et al. 2012
See also: Stadler and Yang 2013

Sample age uncertainty

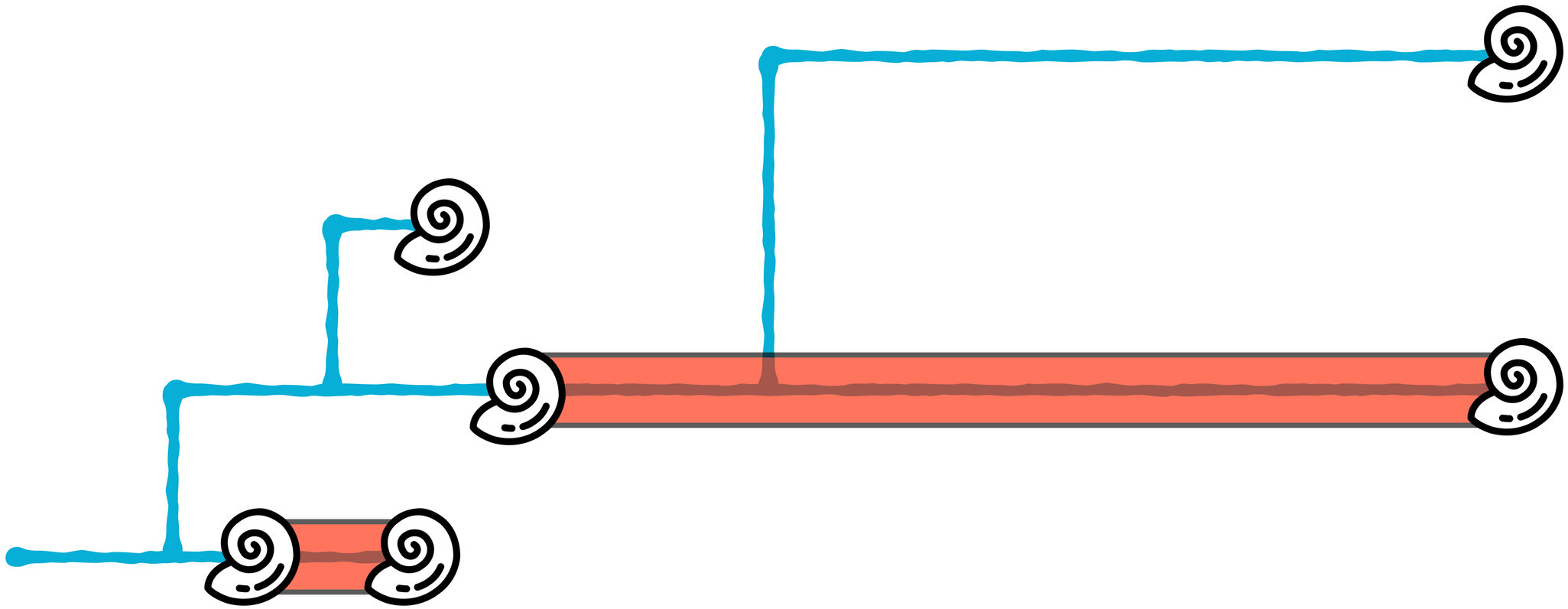


Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology in Time Calibrated Tree Inference
Barido-Sottani et al. 2018, 2020
Putting the F in FBD analyses: tree constraints or morphological data? Barido-Sottani et al. 2023

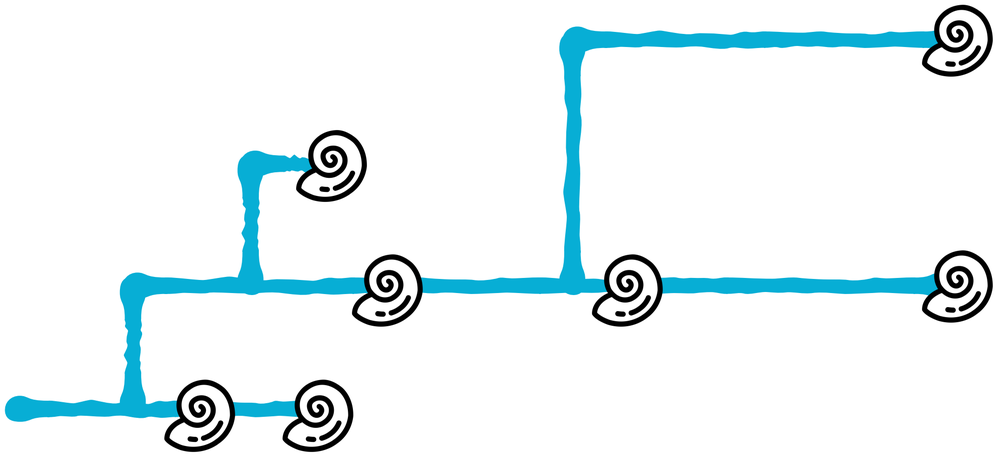


The inseparability of sampling and time and its influence on attempts to unify the molecular & fossil records
Hopkins et al. 2018. Paleobiology

The fossilised birth-death range process



The fossilised birth-death model for the analysis of stratigraphic range data under different speciation modes
Stadler et al. 2018. JTB

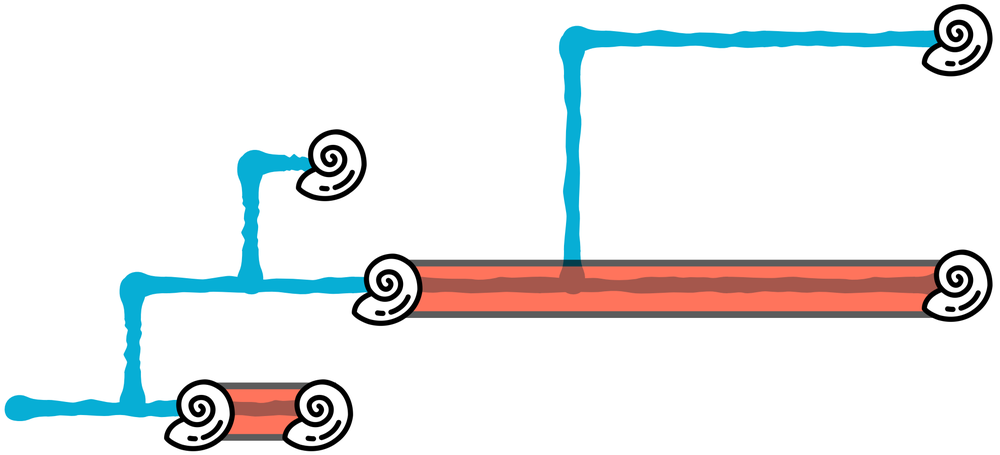


FBD process for analysis of specimen level data

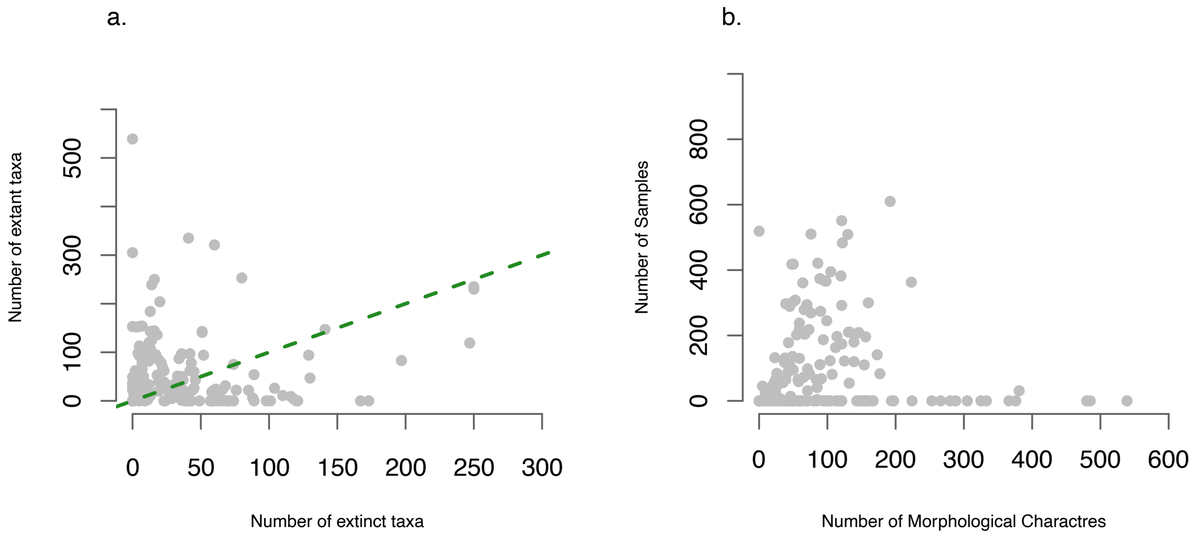
$$P(\Psi \mid \lambda, \mu, \psi, \rho)$$

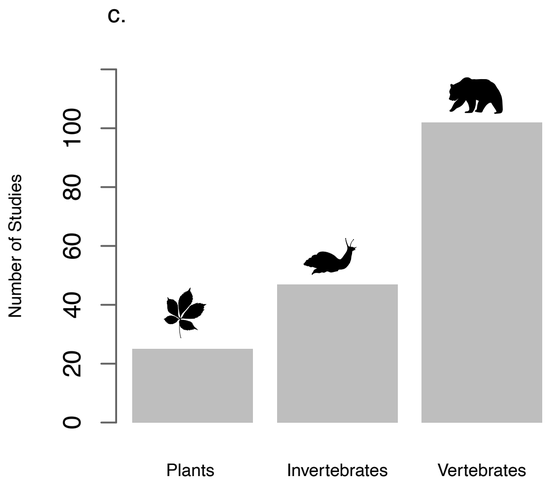
The FBD range process for analysis of stratigraphic ranges

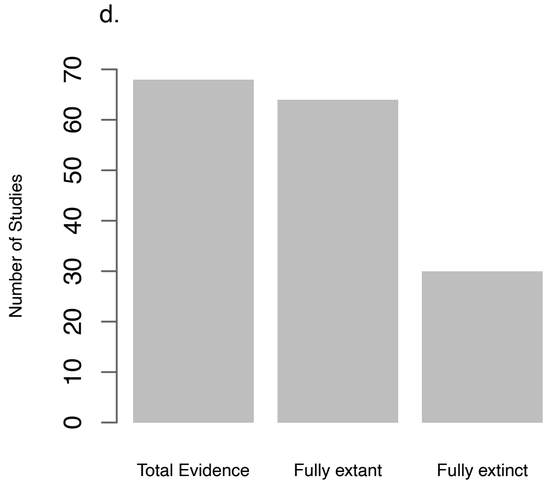
$$Pr(\Psi \mid \lambda, \mu, \psi)$$



When can you use the FBD process?







Used >170 times to data (mainly in macroevolution)

Can be applied to a wide range of scenarios

Upper limit is about 500 samples/tips

Not practical with large alignments

Computational cost comes from sampling tree space

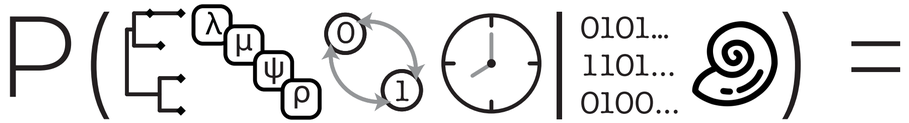
Mulvey et al. in review

Exercise

Phylodynamics

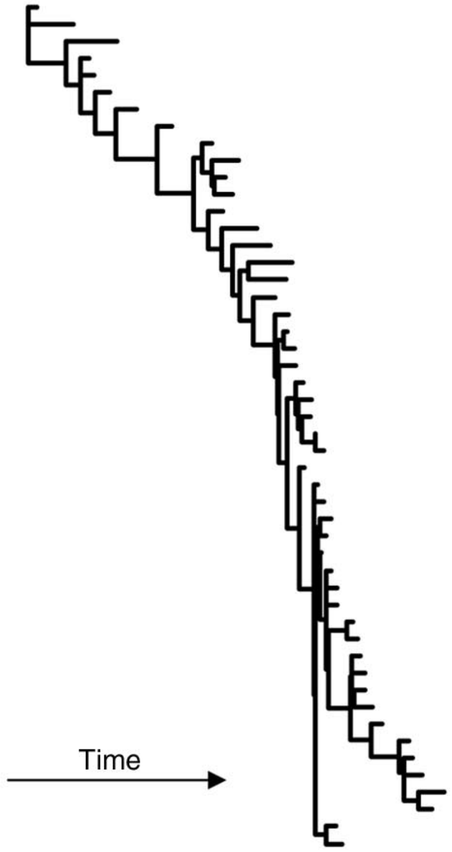
Diversification rate estimation

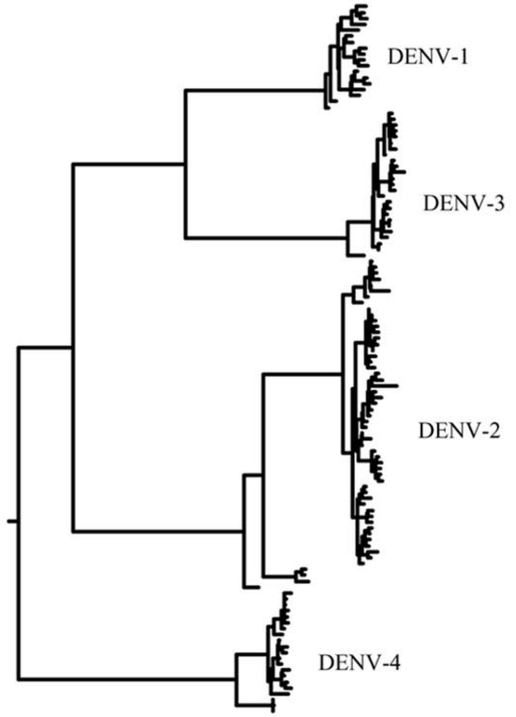
Bayesian divergence time estimation

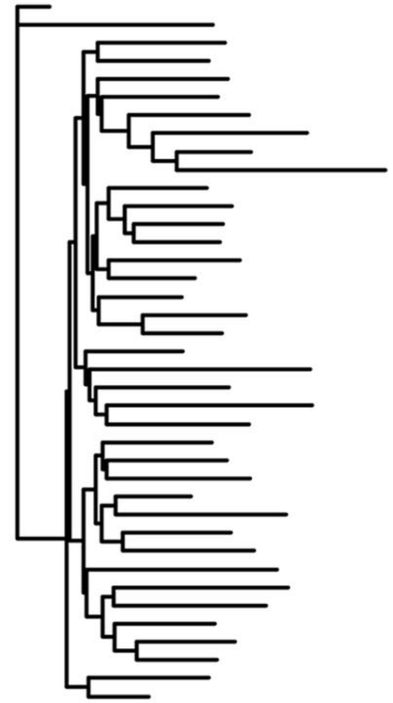


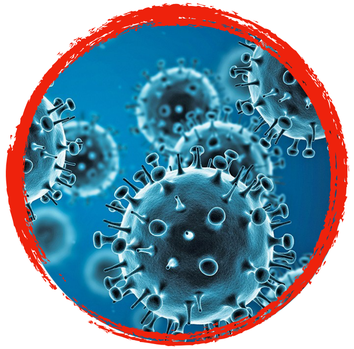
probability of the time tree

Tree shape is informative about underlying dynamics





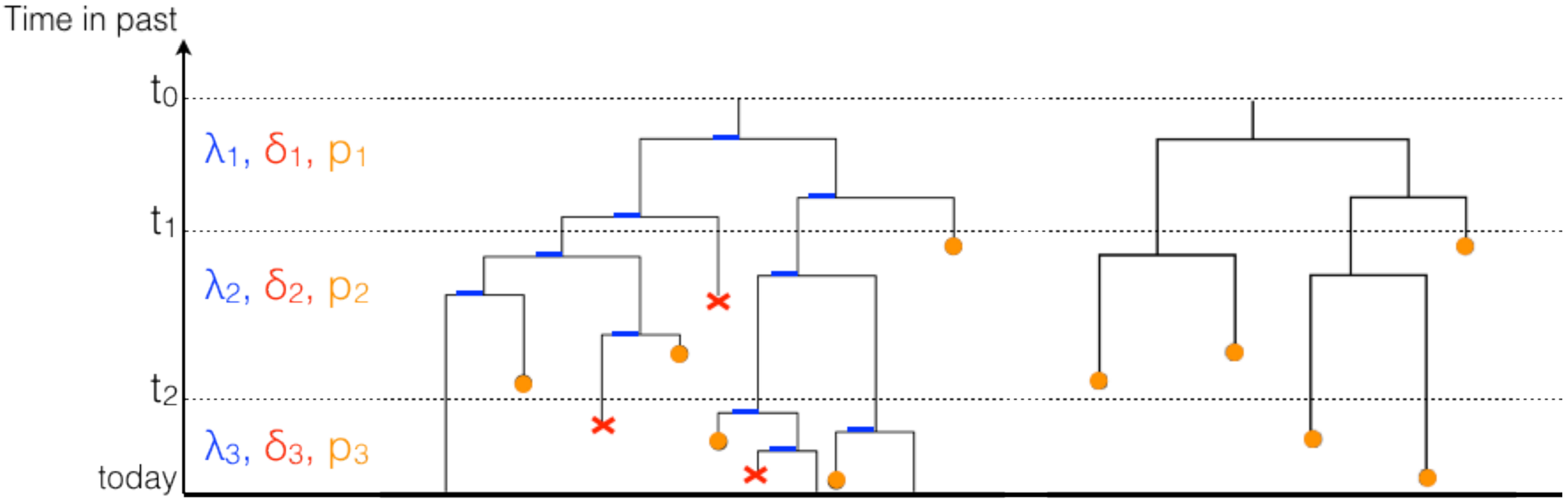




This paper coined the term phylodynamics
Grenfell et al. 2004. Science

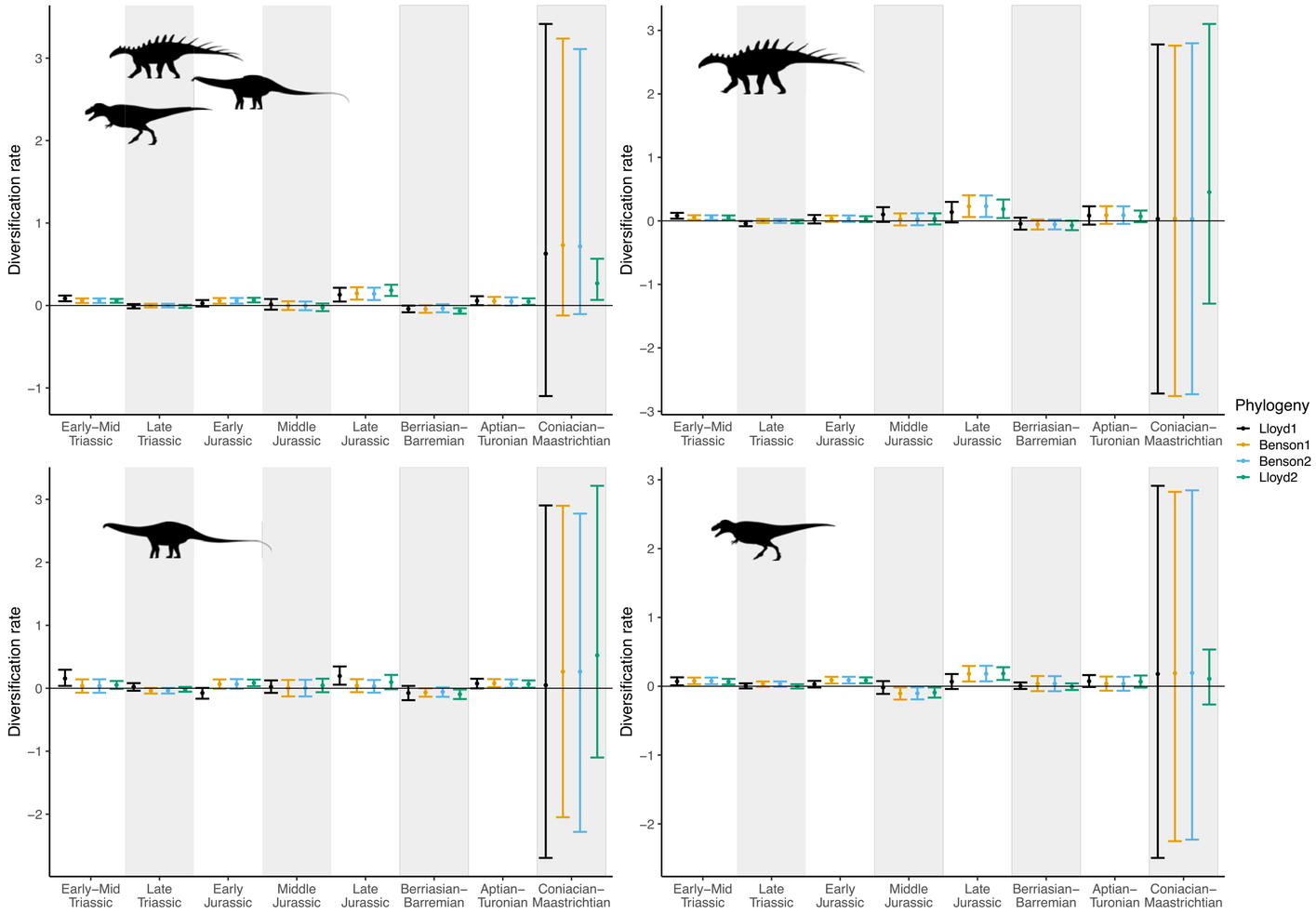
The skyline birth-death process

First used for tracking the spread of infectious diseases



Stadler et al. 2012. PNAS
Gavryushkina et al. 2014. PLoS Comp Bio

Macroevolutionary case study

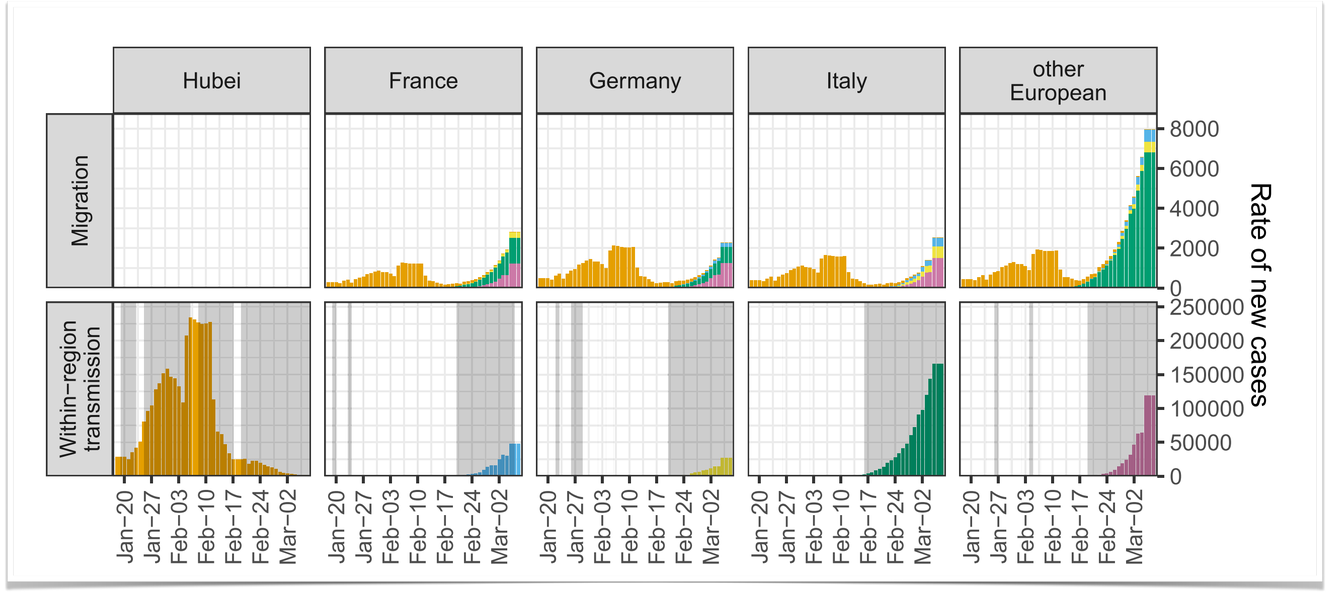


Phylogenies have been used to argue dinosaurs were incline prior to the KPg

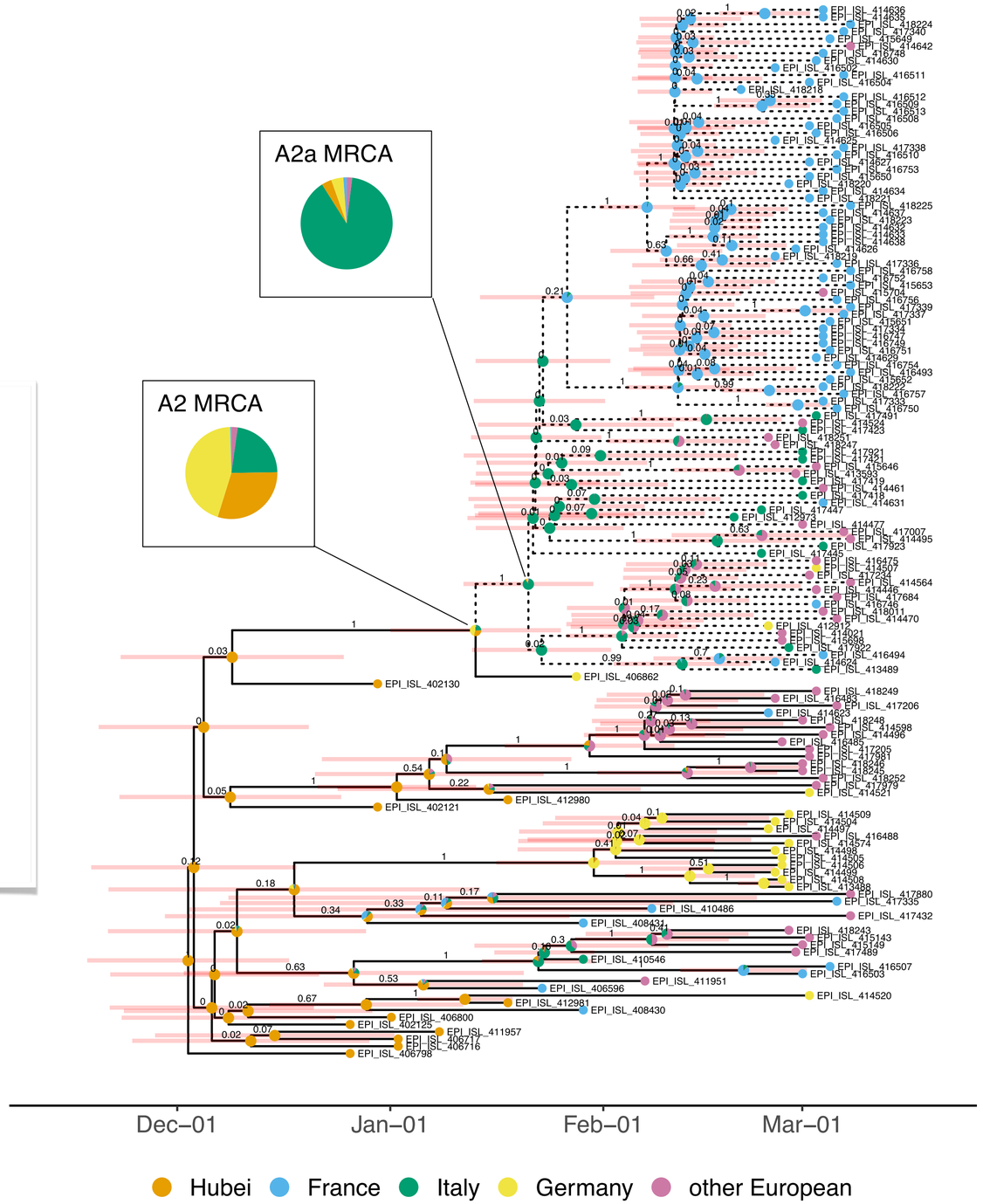
FBD analyses suggest that we can not currently answer that question using phylogenies

Allen et al. 2024. Extinction

Models that include migration



The origin and early spread of SARS-CoV-2 in Europe
Nadeau et al. 2021. PNAS



Bayesian divergence time estimation

The data

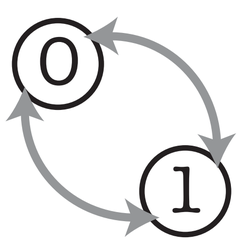
and / or

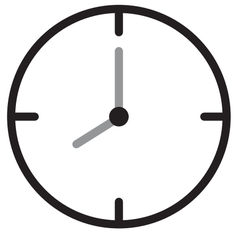
```
0101... ATTG...
1101... TTGC...
0100... ATTC...
```

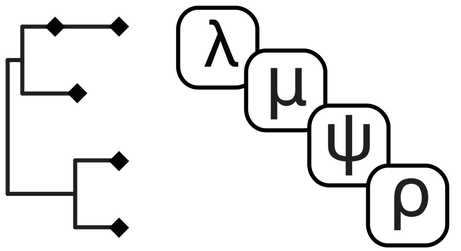
phylogenetics characters

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)

Using PCMs for dating

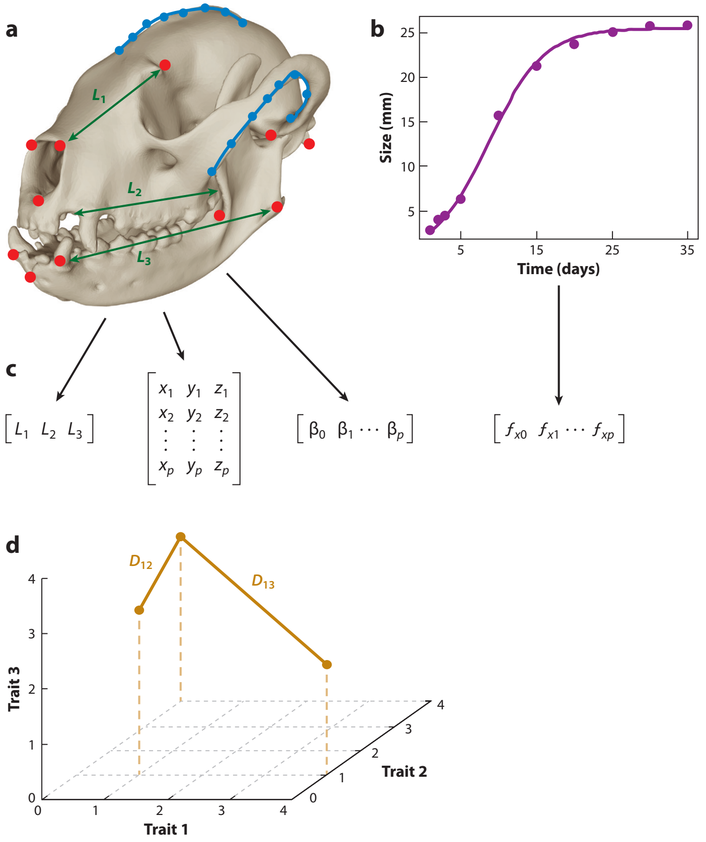
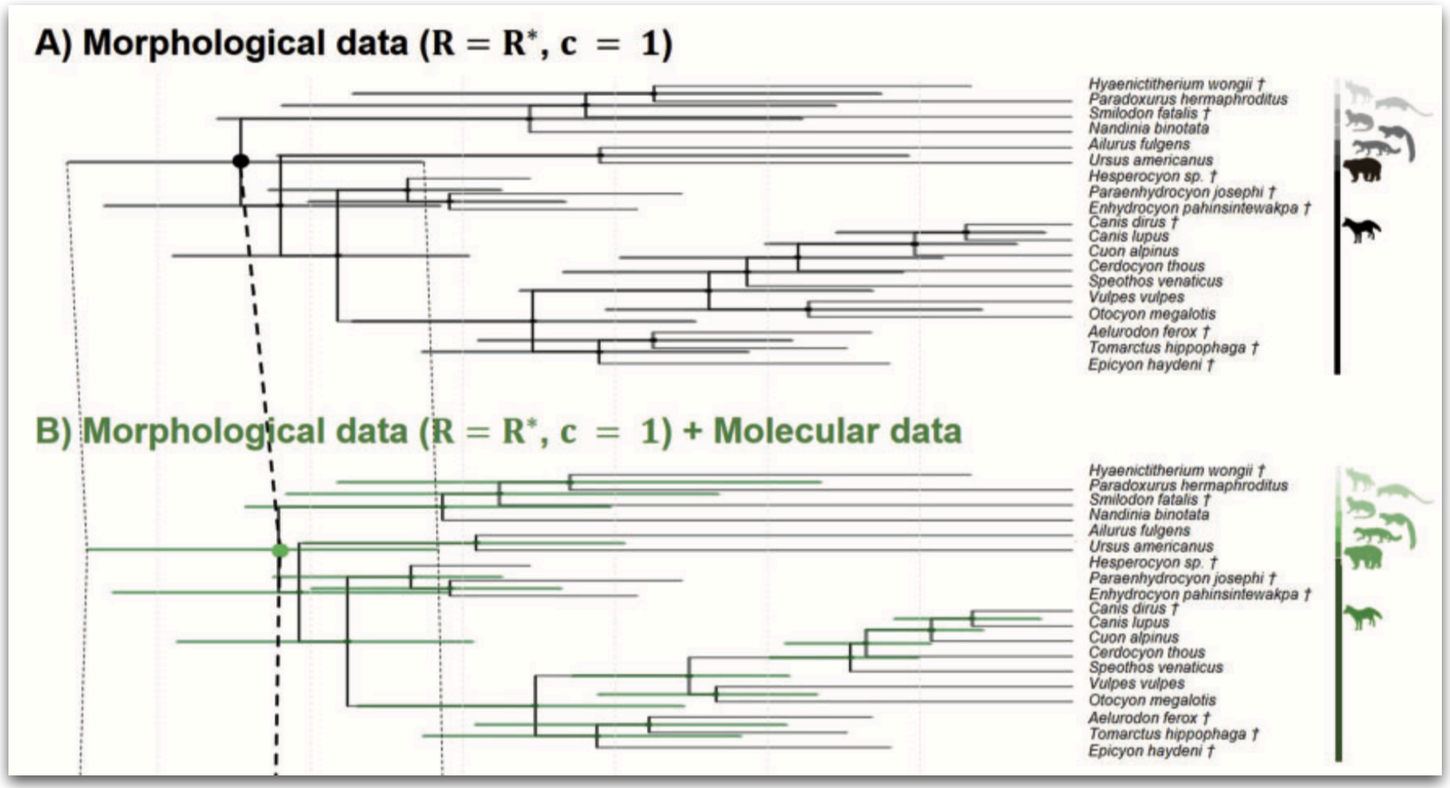
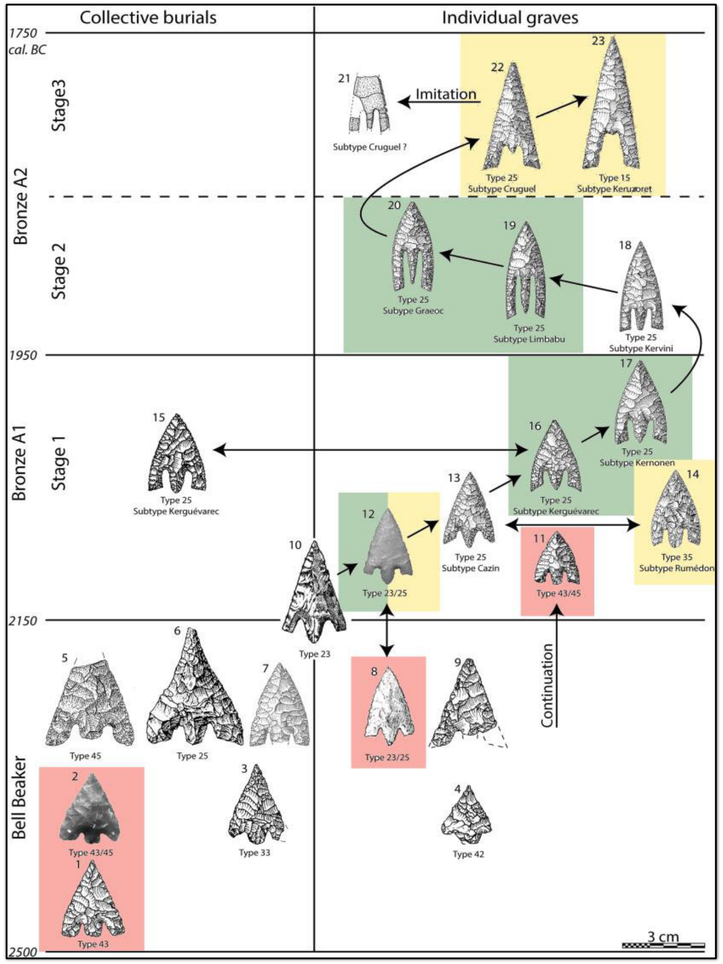


Image source *Adams & Collyer* (2019)



Álvarez-Carretero et al. (2019) *Bayesian Estimation of Species Divergence Times Using Correlated Quantitative Characters*

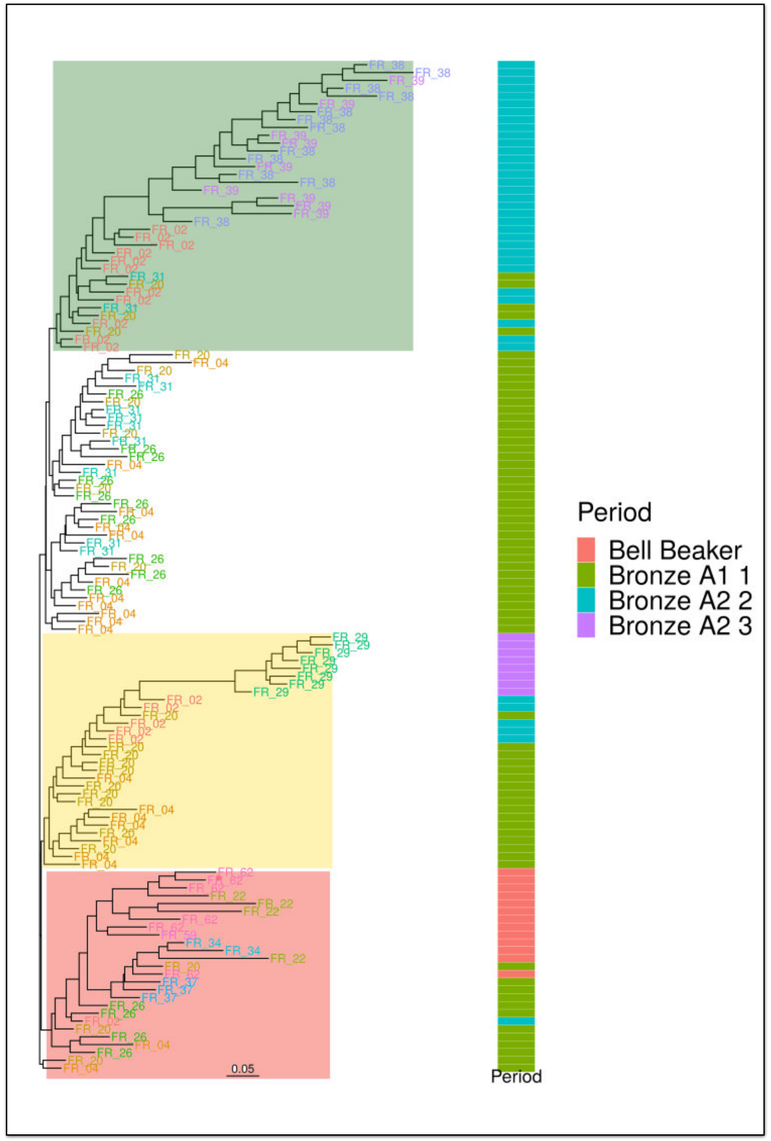
Cultural evolution



After Nicolas (2017)

←Typo-Chronology of Palaeolithic stone tools

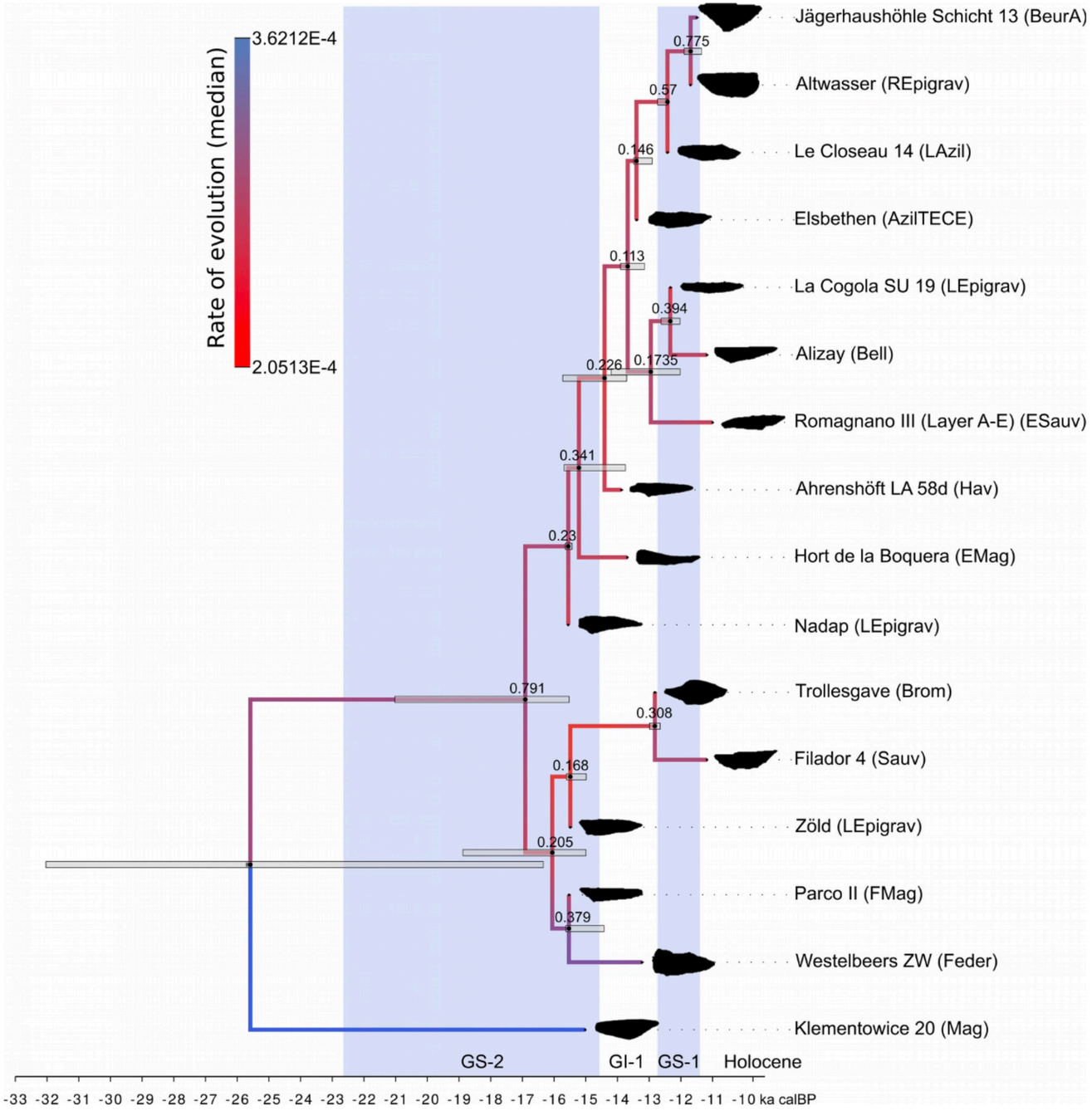
Outline based NJ tree →



Matzig et al. 2021.

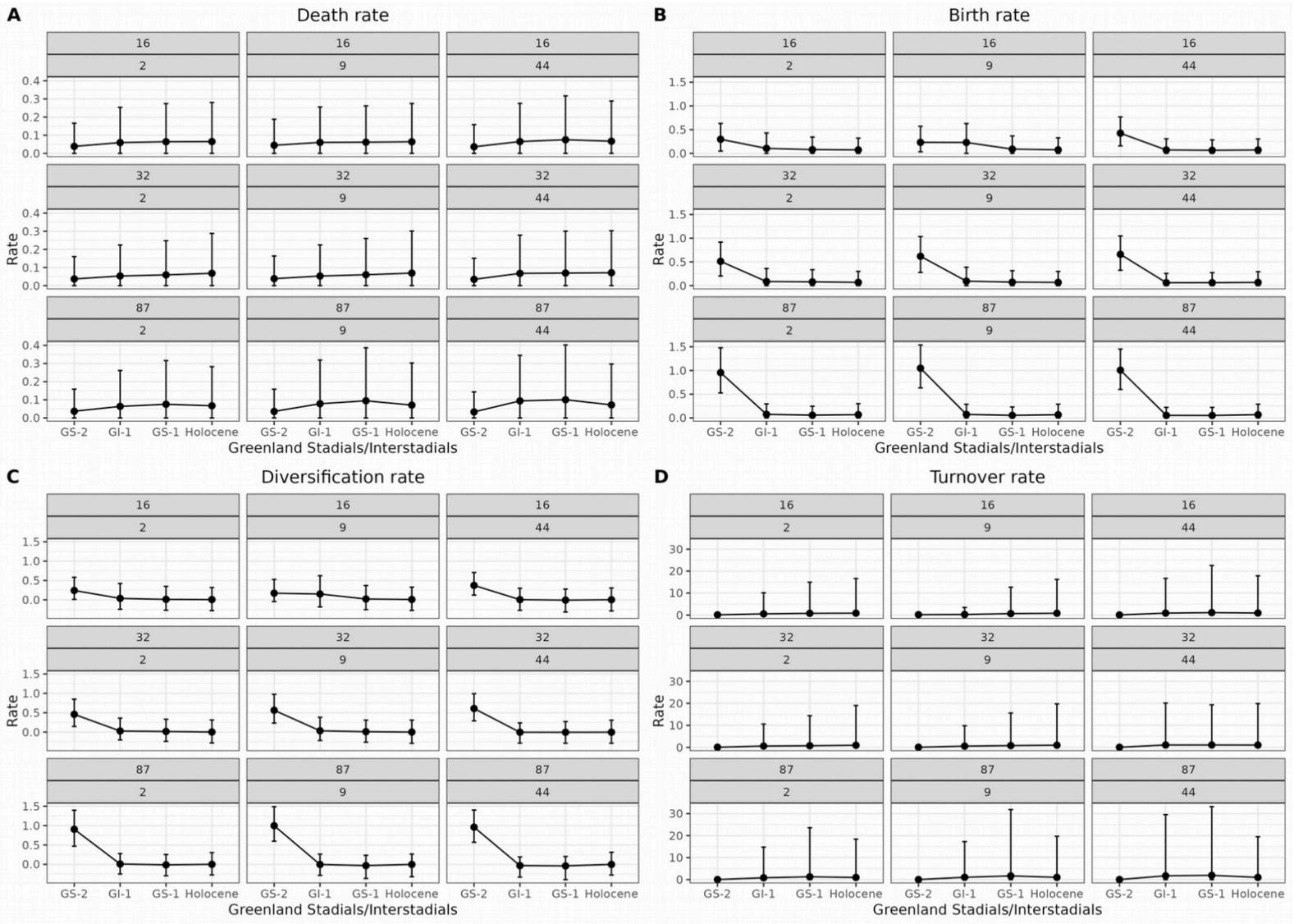
The tree topology of stone tools exhibits a lot of uncertainty

Matzig et al. (in review) *A macroevolutionary analysis of European Late Upper Palaeolithic stone tool shape using a Bayesian phylodynamic framework* (preprint available)



Sensitivity analyses

Birth, death, and sampling rates are impacted by trait and taxon sampling



Joint phylogenetic estimation of geographic movements and biome shifts

Landis et al. (2023) *Systematic Biology*

Global diversity of *Viburnum*

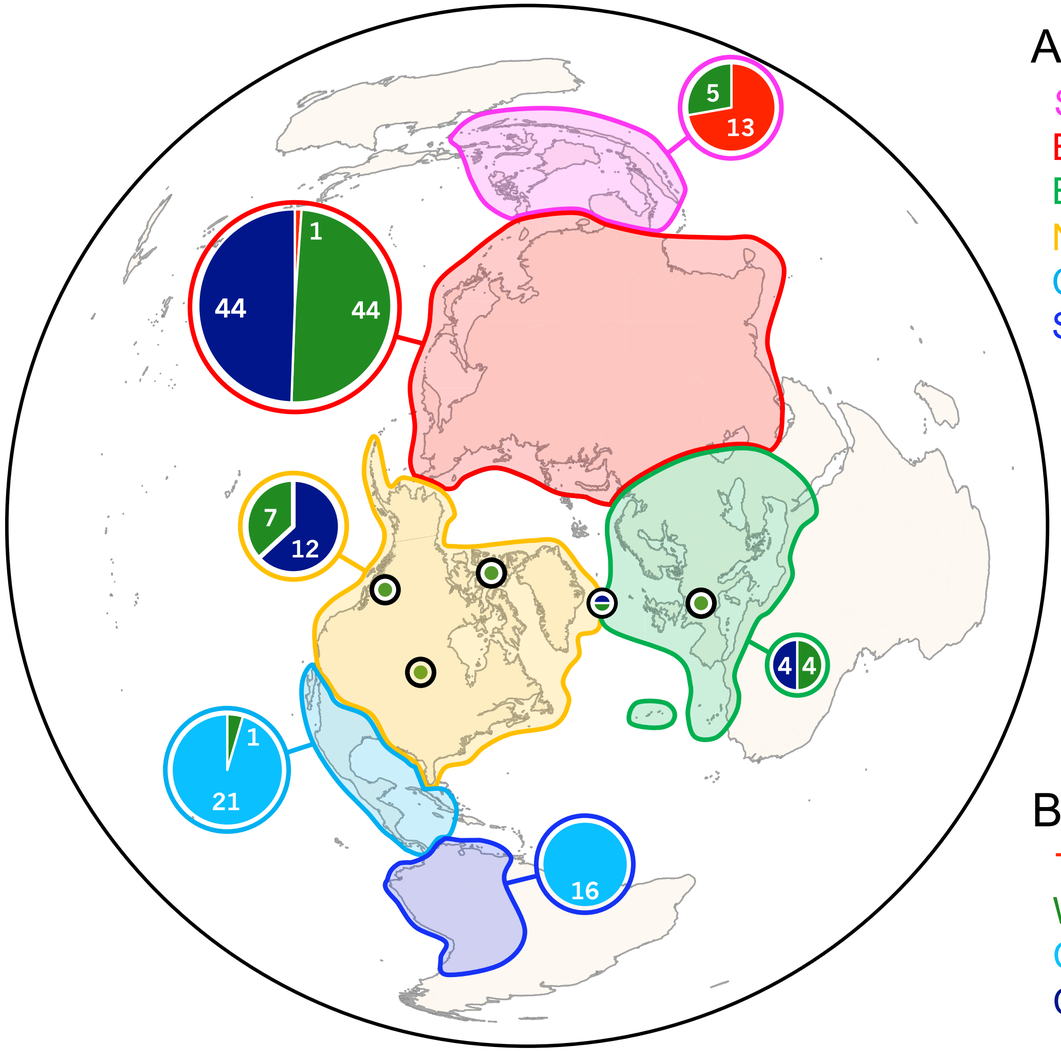
Data

163 extant species (127 with DNA)

5 fossils (with taxonomic constraints)

6 geographic areas

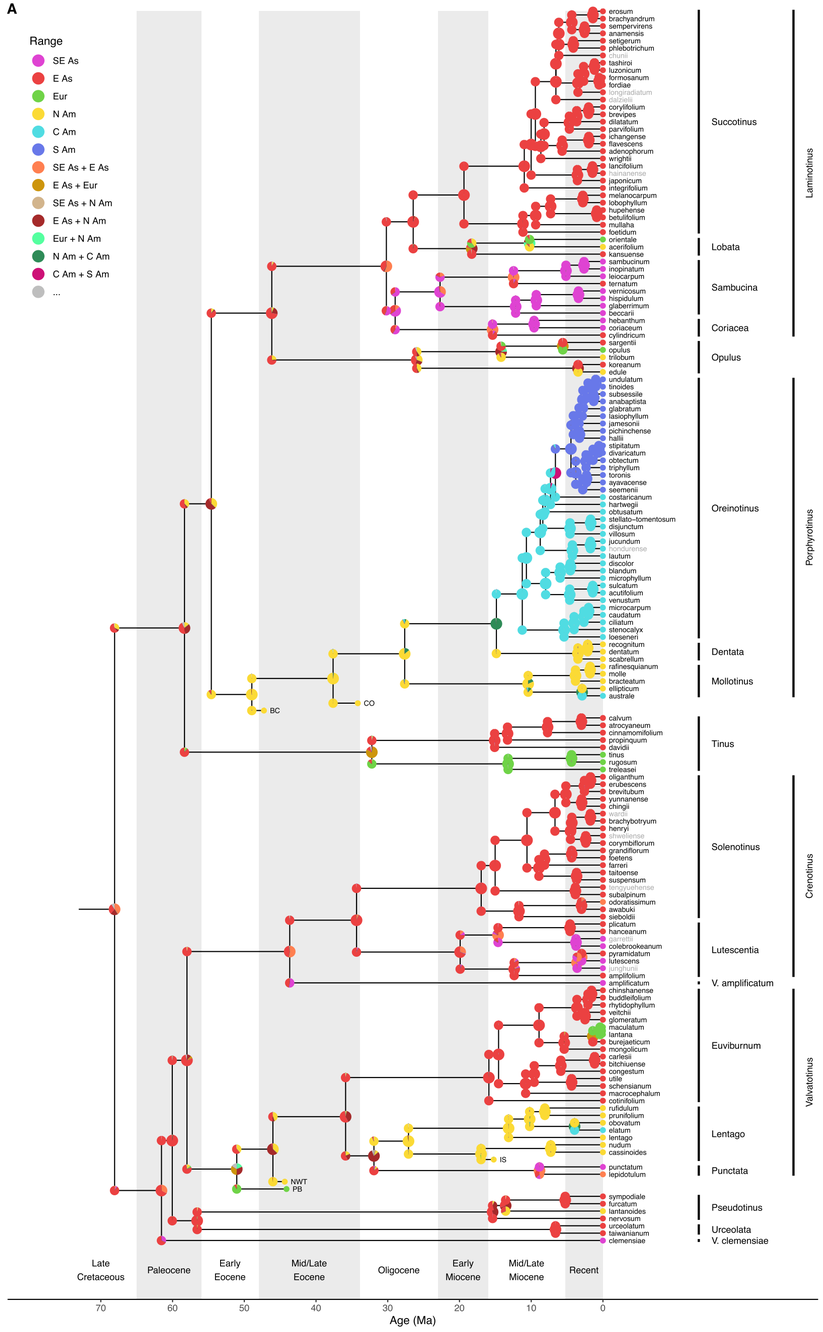
4 biomes

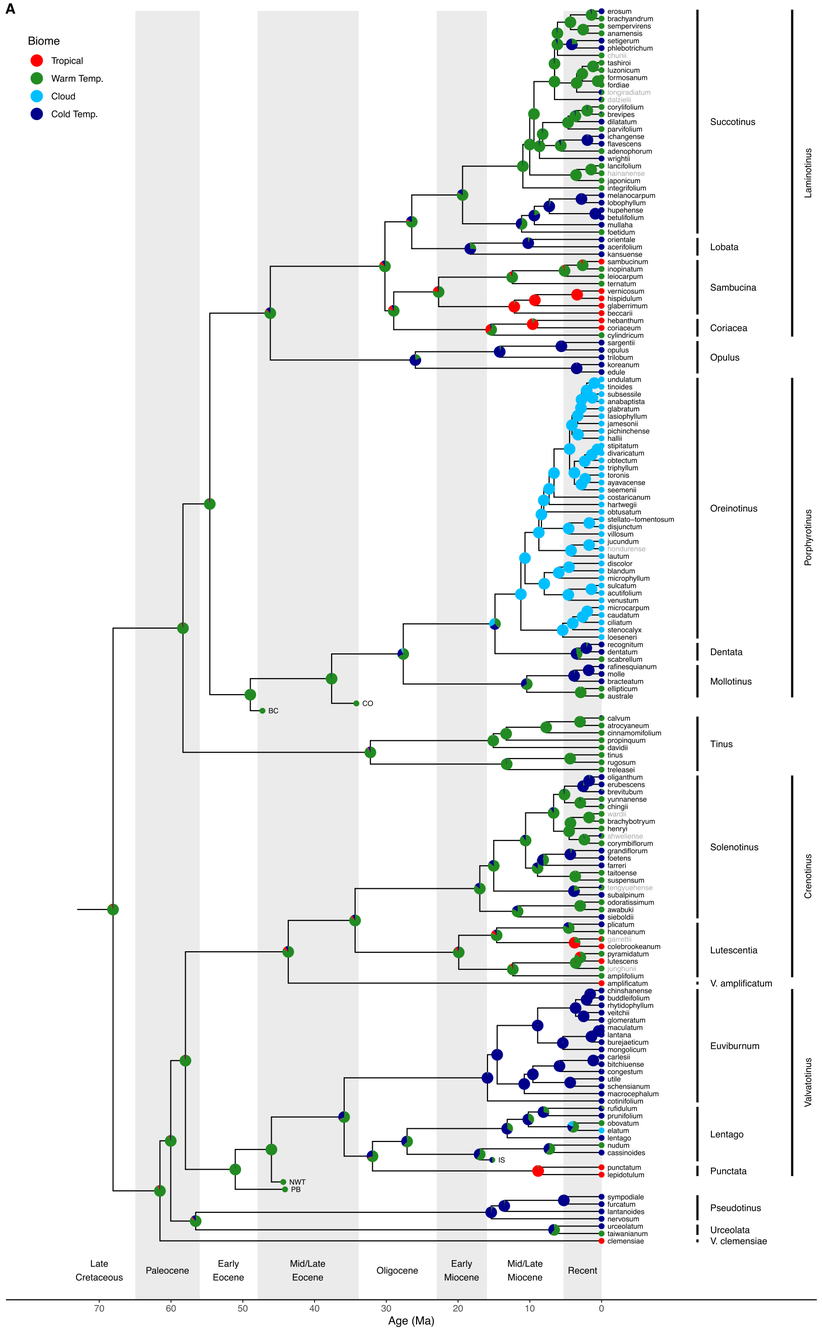


- Can we integrate **geographic range** and **biome data** into analysis using the FBD model?
- What can we learn about the **diversification history** of the *Viburnum*?

Analysis

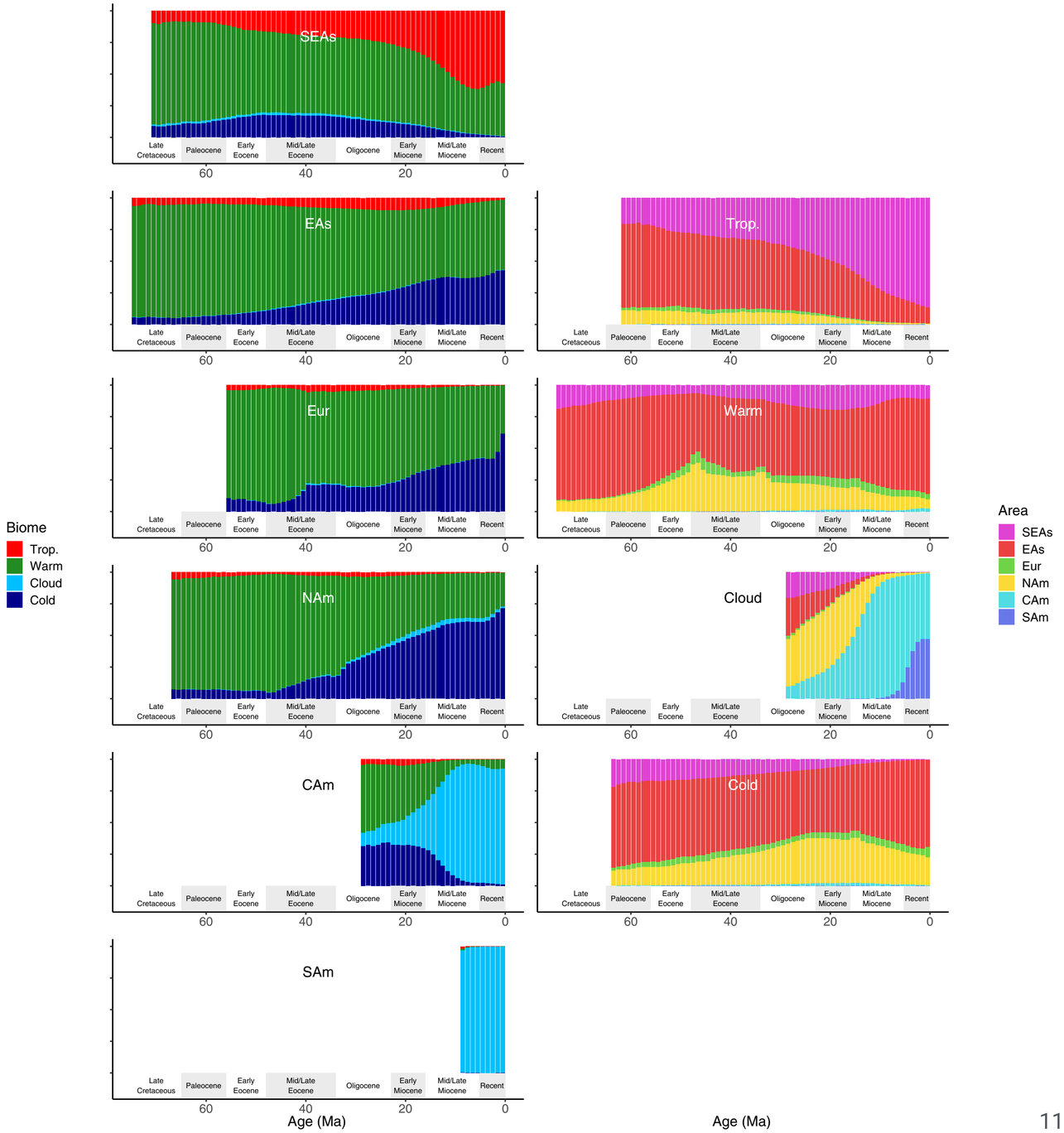
1. Estimate the extant topology using maximum likelihood
2. Joint inference divergence times, biogeographic and biome history (normally we *first* infer a dated tree, and *then* separately infer biogeographic history)
3. Ancestral state reconstruction
4. + various sensitivity analyses


A
Range
SE As
E As
Eur
N Am
C Am
S Am
SE As + E As
E As + Eur
SE As + N Am
E As + N Am
Eur + N Am
N Am + C Am
C Am + S Am
...
Succotinus
Lobata
Sambucina
Coriacea
Opulus
Oreinotinus
Dentata
Mollotinus
Tinus
Solenotinus
Lutescentia
V. amplificatum
Euviburnum
Lentago
Punctata
Pseudotinus
Urceolata
V. clemensiae
Laminotinus
Porphyrotinus
Crenotinus
Valvatotinus
BC
CO
IS
NWT
PB
Late Cretaceous
Paleocene
Early Eocene
Mid/Late Eocene
Oligocene
Early Miocene
Mid/Late Miocene
Recent
70
60
50
40
30
20
10
0
Age (Ma)


A
Biome
Tropical
Warm Temp.
Cloud
Cold Temp.
Succotinus
Lobata
Sambucina
Coriacea
Opulus
Oreinotinus
Dentata
Mollotinus
Tinus
Solenotinus
Lutescentia
V. amplificatum
Euviburnum
Lentago
Punctata
Pseudotinus
Urceolata
V. clemensiae
Laminotinus
Porphyrotinus
Crenotinus
Valvatotinus
BC
CO
IS
NWT
PB
Late Cretaceous
Paleocene
Early Eocene
Mid/Late Eocene
Oligocene
Early Miocene
Mid/Late Miocene
Recent
70
60
50
40
30
20
10
0
Age (Ma)

Results summary

- Joint inference using the FBD and biogeographic models allows us to estimate a rich diversification history
- Major lineages of *Viburnum* likely originated in warm / temperate regions and later adapted to the cold
- Fossils can change the results



*“It is, it must be admitted, a **humbling** task to infer ancient events, and the results in many cases are tenuous at best. Given the obvious limitations of working with extant species and few, if any, fossils, **it is necessary to integrate all of the available sources of evidence** if we hope to produce assuring answers.”*

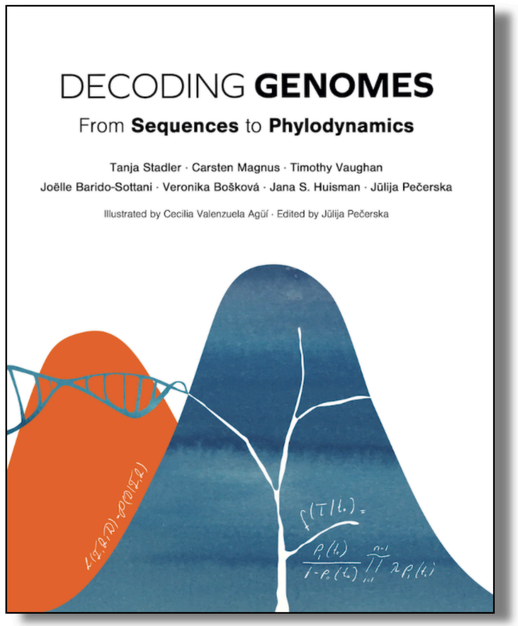
Landis et al. (2023) *Systematic Biology*
Joint phylogenetic estimation of geographic movements and biome shifts

Decoding Genomes: From Sequences to Phylodynamics

Tanja Stadler, Carsten Magnus, Timothy Vaughan, Joëlle Barido-Sottani, Veronika Bošková, Jana S. Huisman, Jūlija Pečerska

Illustrated by *Cecilia Valenzuela Agüí*

Edited by *Jūlija Pečerska*



Obtaining the book

You will shortly be able to purchase a hard copy from Amazon. (Quality testing currently in progress.)

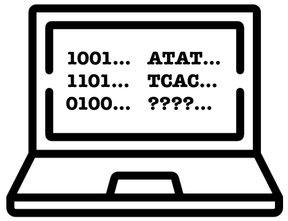
Alternatively, you can Download the complete PDF of the book free of charge. (See below for license information.)

About the book

Decoding Genomes demonstrates how to uncover information about past evolutionary and population dynamic processes based on genomic samples. The last decades have seen considerable theoretical and methodological advances in this area. These enable the assessment of critical scientific questions such as the impact of environmental changes on biodiversity and the evolution of pathogens during recent epidemics. The book gives the reader a detailed understanding of the whole process: from genome sampling to obtaining biological insights by applying sophisticated statistical and computational analyses. In particular, sequencing of genomic samples, the alignment of sequences, molecular evolution models, phylogenetics, and phylodynamics are core topics. Statistical and computational approaches discussed include dynamic programming, maximum likelihood, Bayesian statistics, and model selection, to name a few. The concepts introduced and applied throughout the book enable readers to answer

A few notes on software

for Bayesian time tree estimation. All open source.



- MCMCTree — BDSS process, continuous trait models. Best option for large sequence alignments and trees. Requires a fixed tree. Language: C
- PhyloBayes — for extant time tree inference. Good for amino acid data. C++
- MrBayes — FBD model, some unique clock models. Easy to use. C++

For increased modularity & flexibility:

- BEAST2 — FBD model, lots of flexible tree and character evolution models. More widely used in epidemiology. Java. (Sister software BEAST 1.8)

- RevBayes — FBD model, lots of flexible tree and character evolution models. C++. Uses graphical models. Developed by folk closer to macroevolution

When do use different software?

Scenario	Software
Large datasets of extant taxa and node calibrations	MCMCTree
If want (or have to) fix the tree topology	MCMCTree
If fossil sampling is sparse or complex	MCMCTree
If you have abundant fossil data, or are interested in the topological position of fossils	BEAST2, RevBayes
If you’re interested in the phylodynamic parameters	BEAST2, RevBayes
If you want to use a specific model	BEAST2, RevBayes, MCMCTree