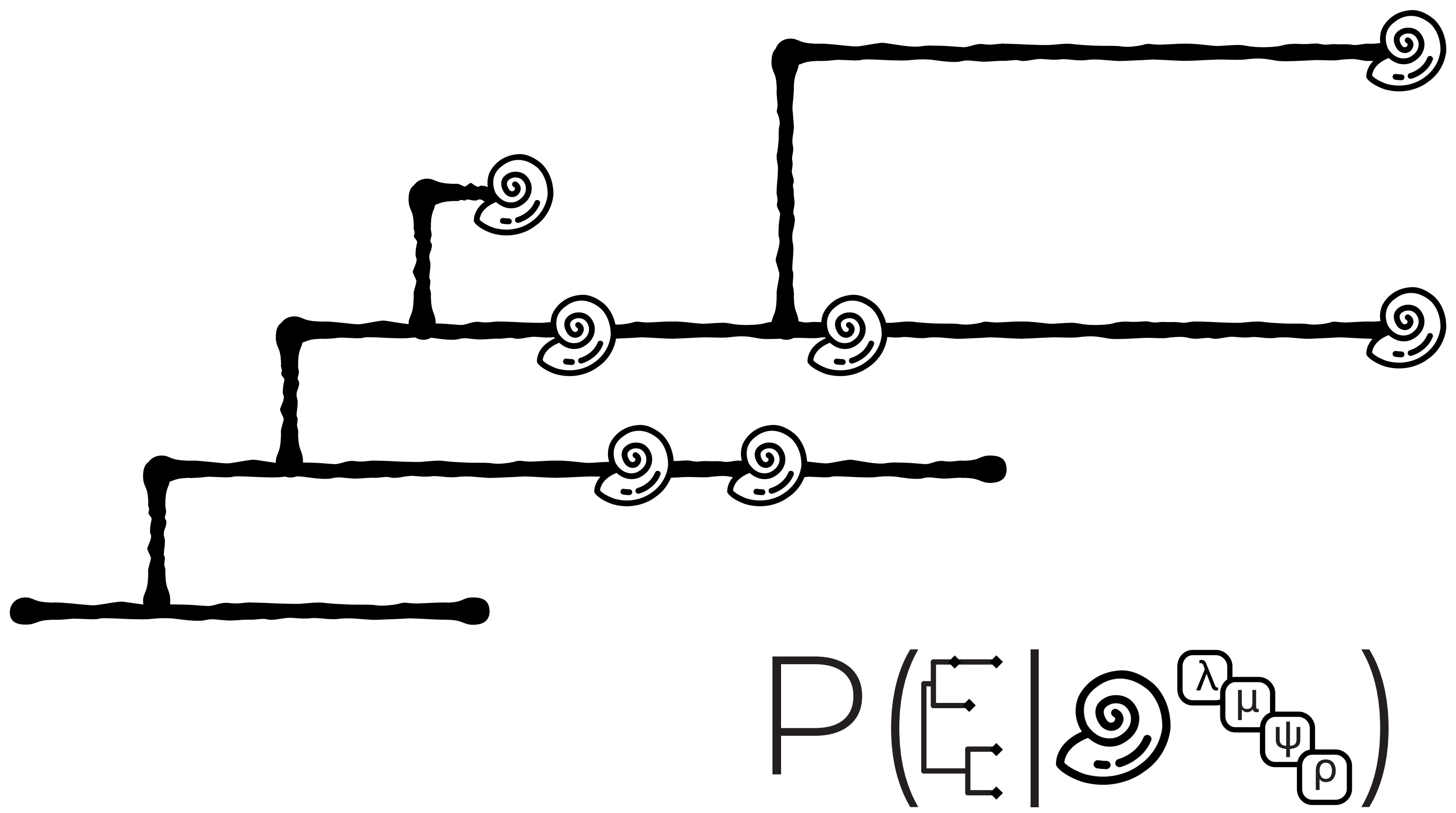
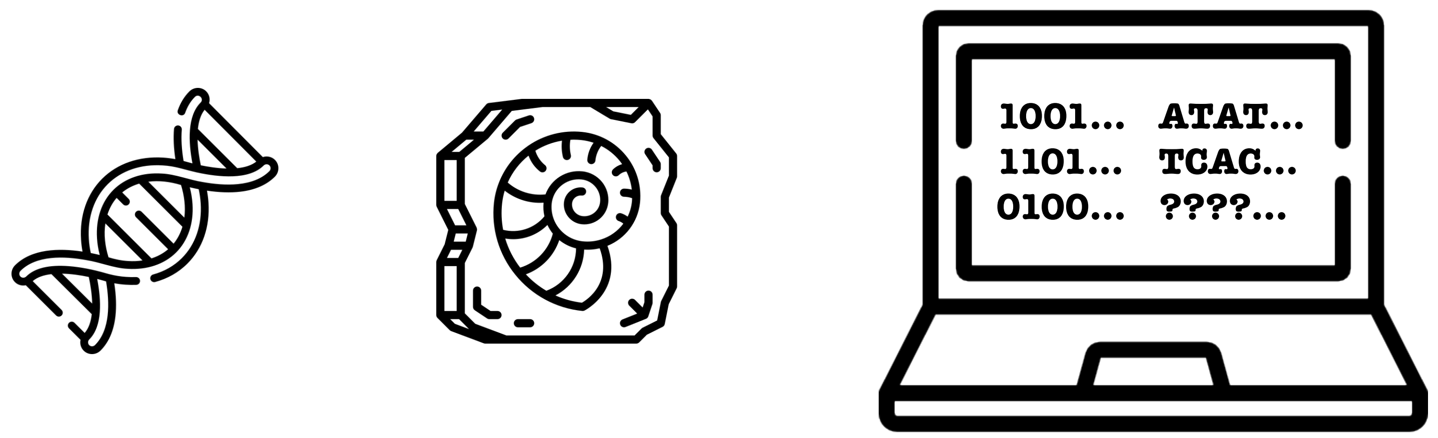


Estimating divergence times using the fossilised birth-death process

Rachel Warnock

APW 2023



Objectives

Lecture

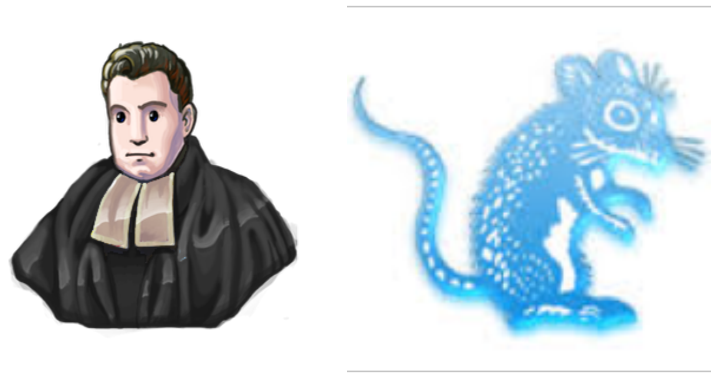
Potential issues with node dating (in brief)

Dating with sampling through time (the fossilised birth-death process)

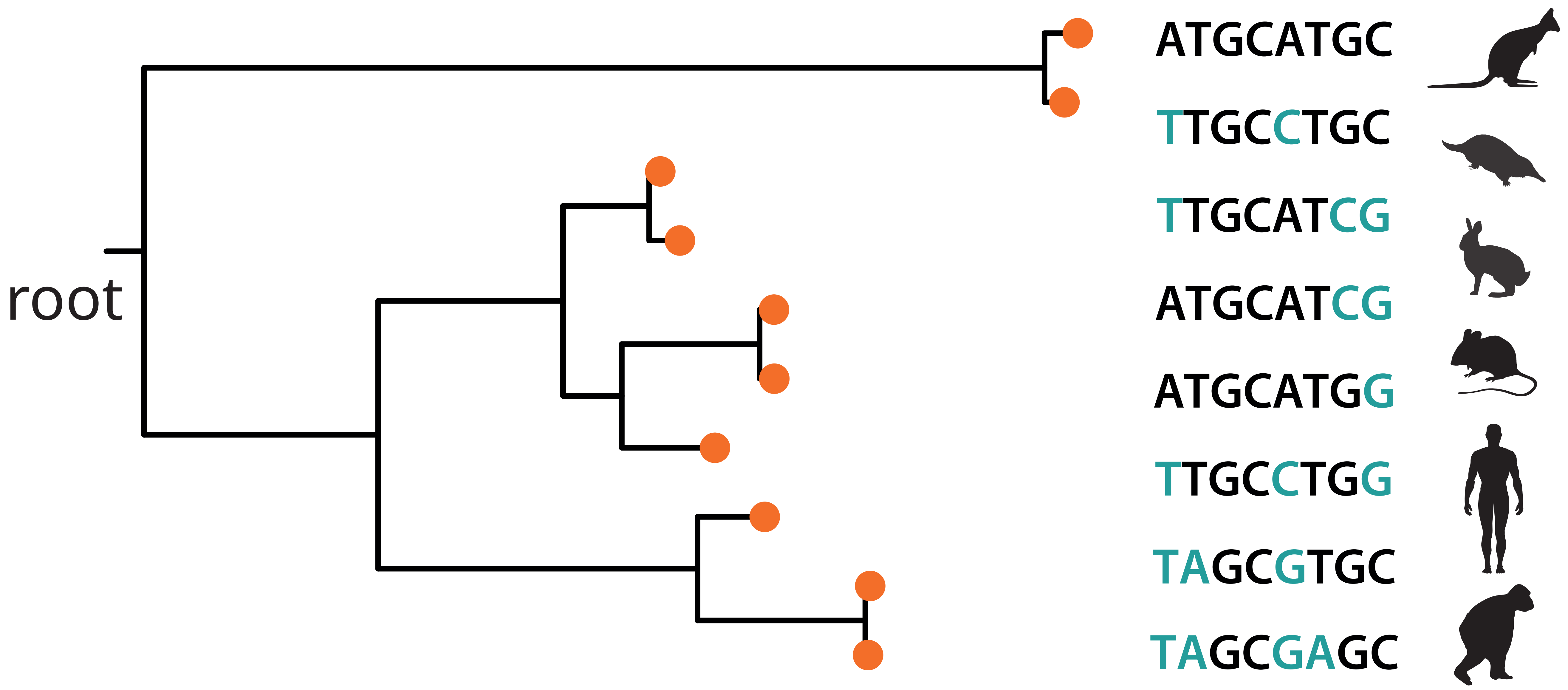
A few notes about available software

Tutorial

Divergence dating under the fossilised birth-death process



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

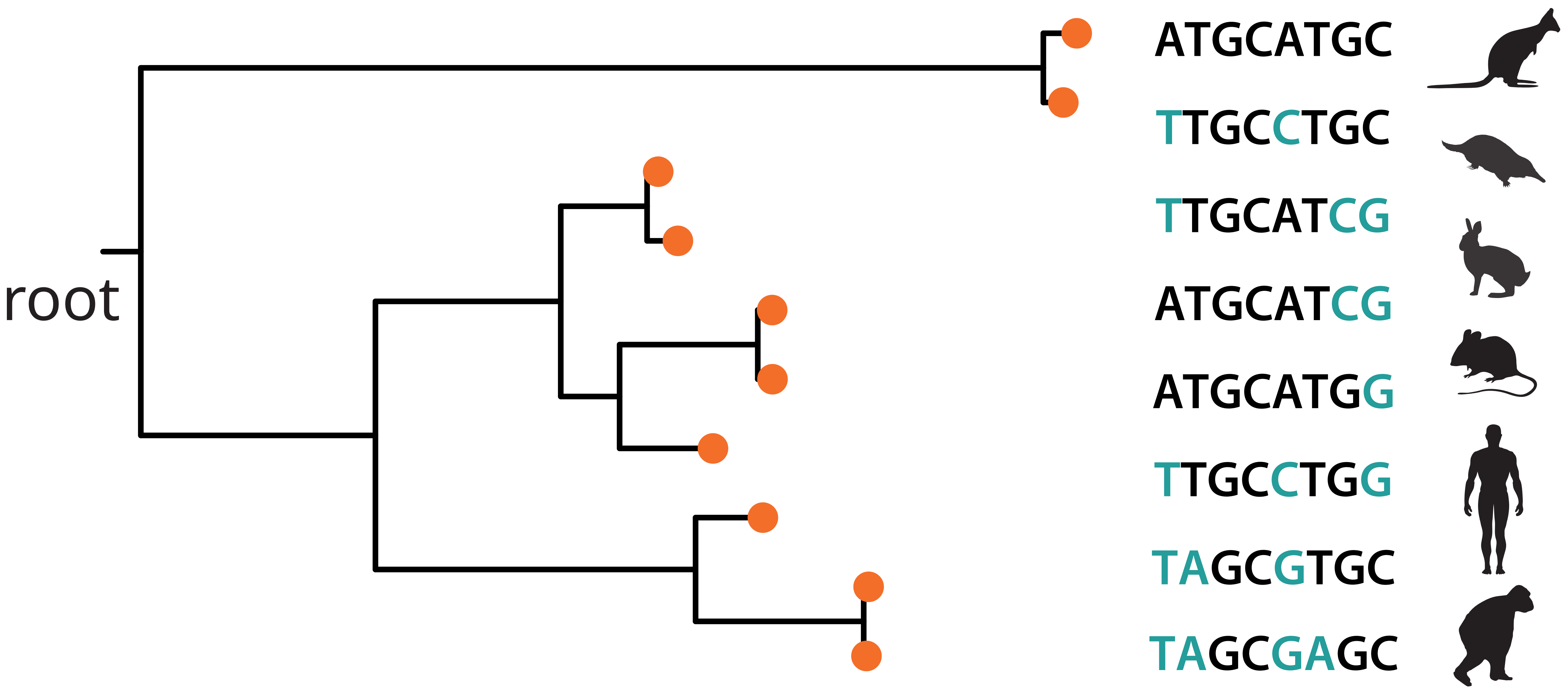


ATGCATGC
TTGCCTGC
TTGCATCG
ATGCATCG
ATGCATGG
TTGCCTGG
TAGCGTGC
TAGCGAGC

Slow rate, long interval OR fast rate, short interval?

branch lengths = genetic distance
$v = rt$

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

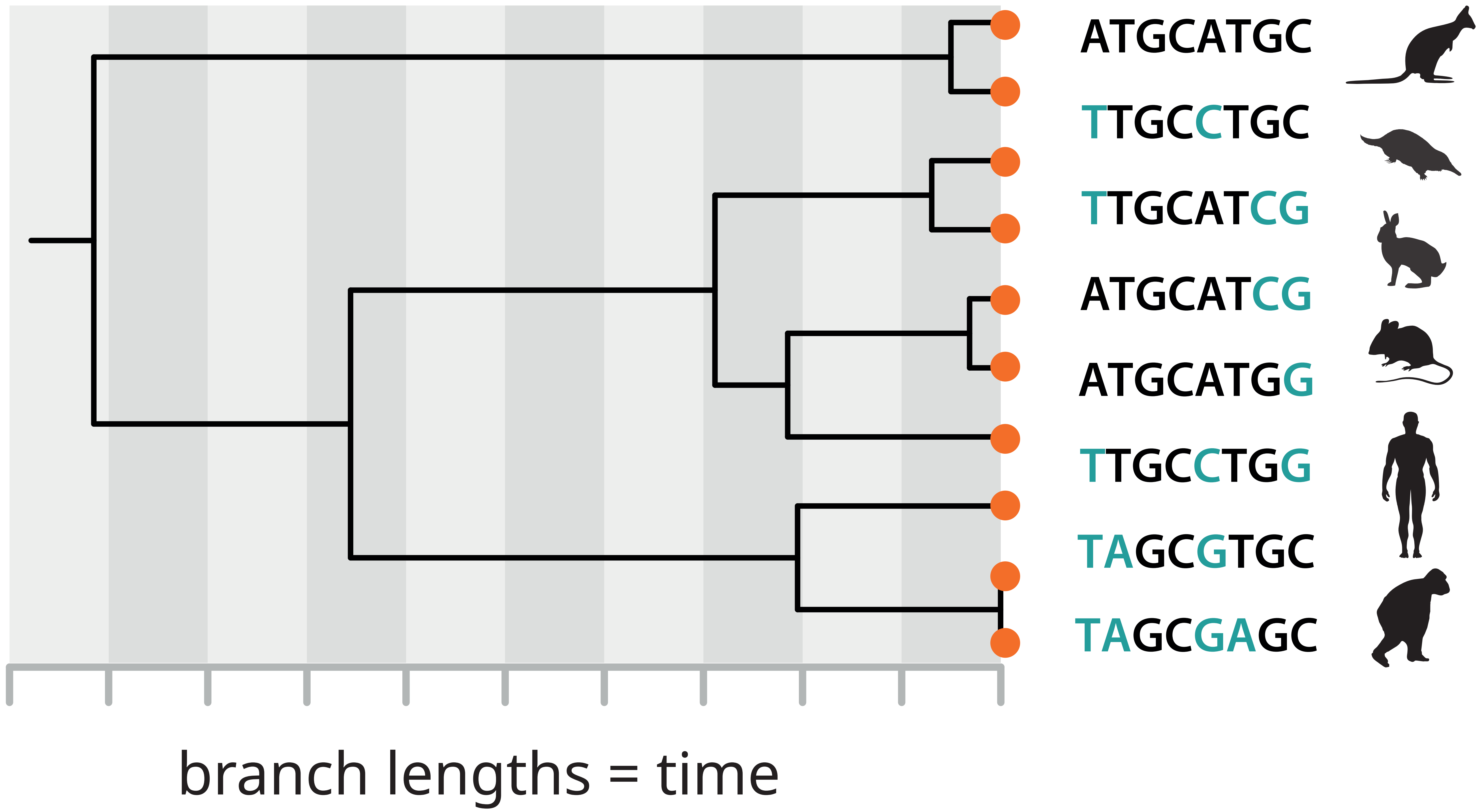


ATGCATGC
TTGCCTGC
TTGCATCG
ATGCATCG
ATGCATGG
TTGCCTGG
TAGCGTGC
TAGCGAGC

branch lengths = genetic distance
$v = rt$

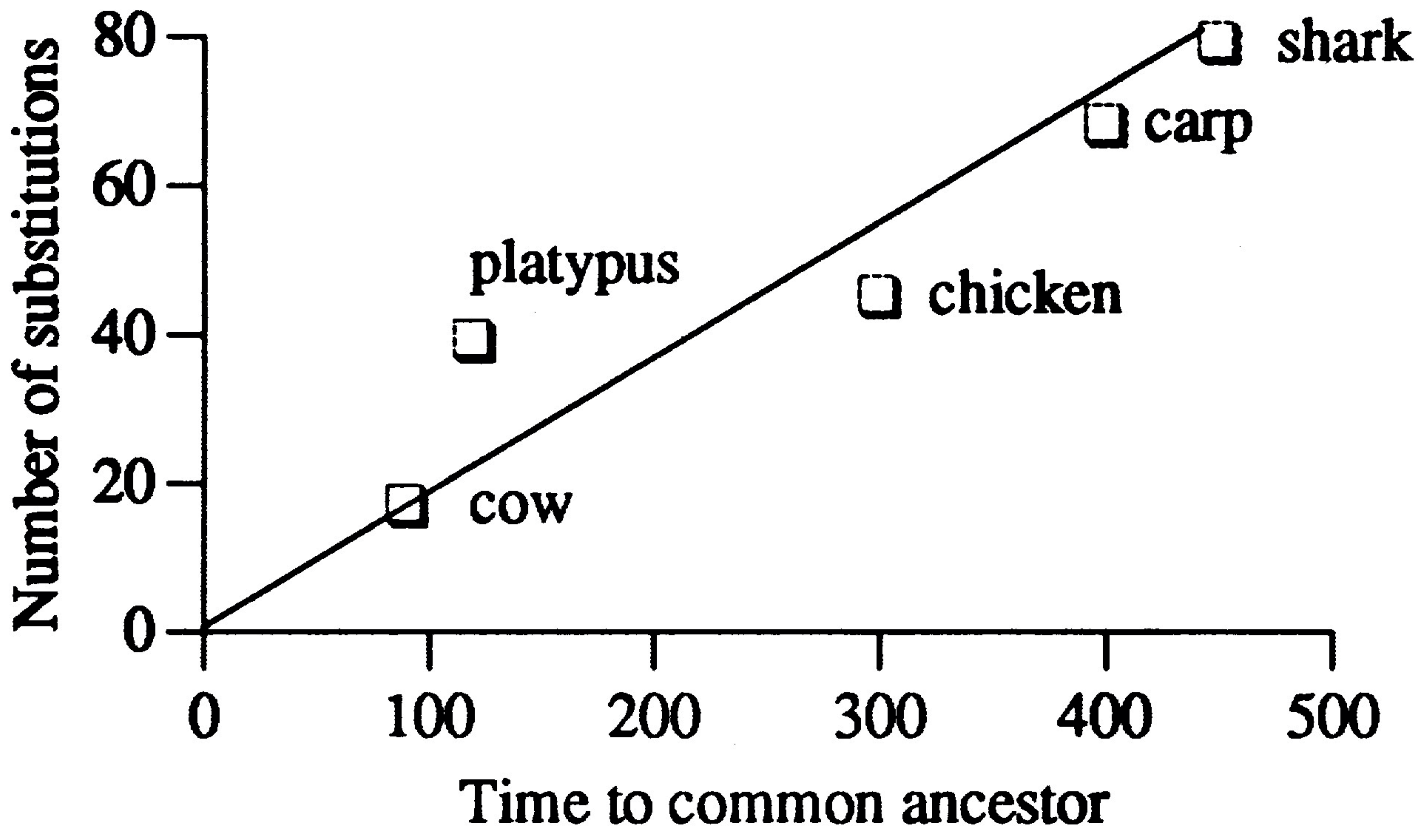
Goal: to disentangle evolutionary rate and time.

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



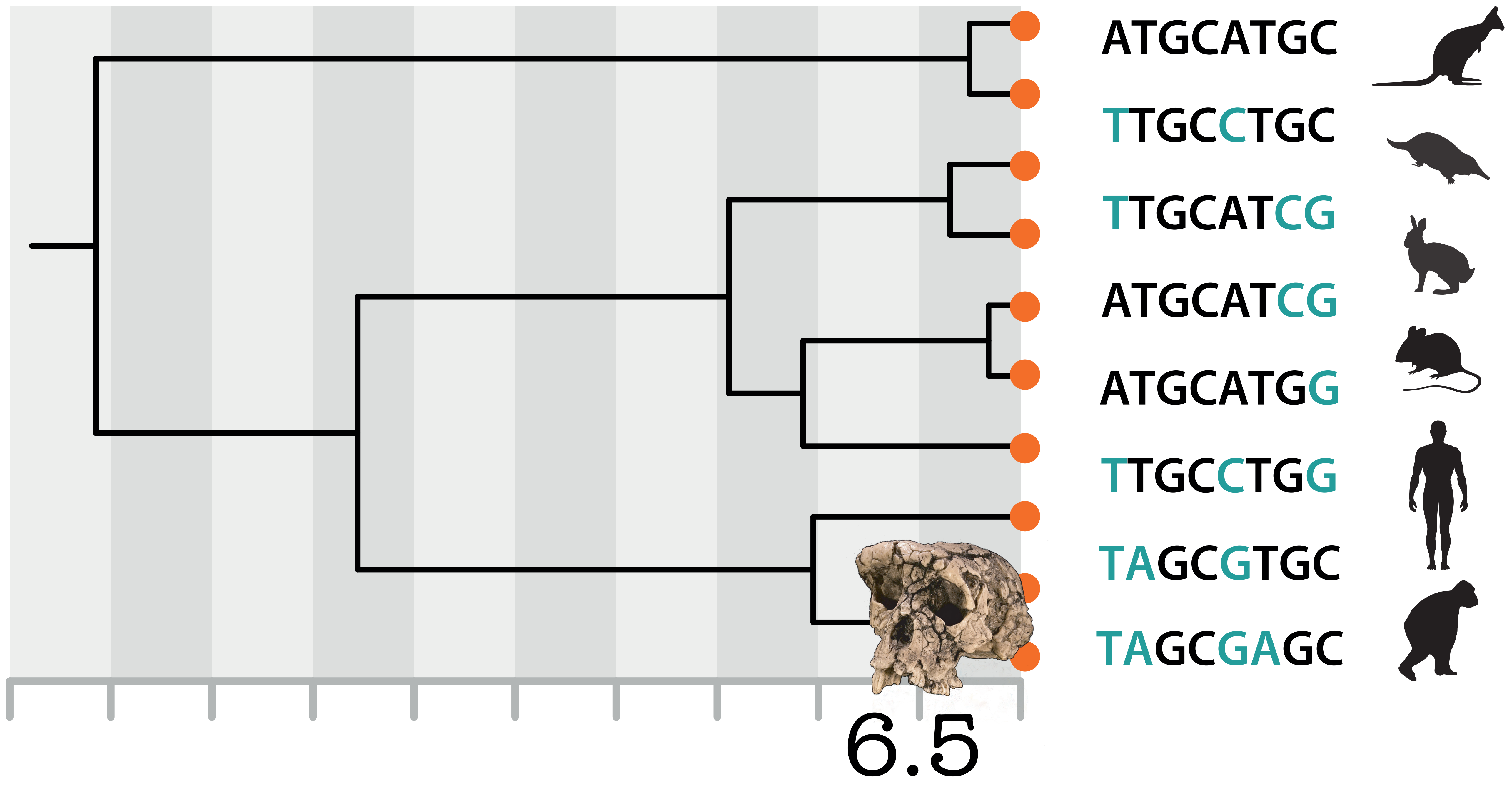
Goal: to disentangle evolutionary rate and time.

The molecular clock hypothesis



Zuckerkandl & Pauling (1965) — Molecules as documents of evolutionary history.

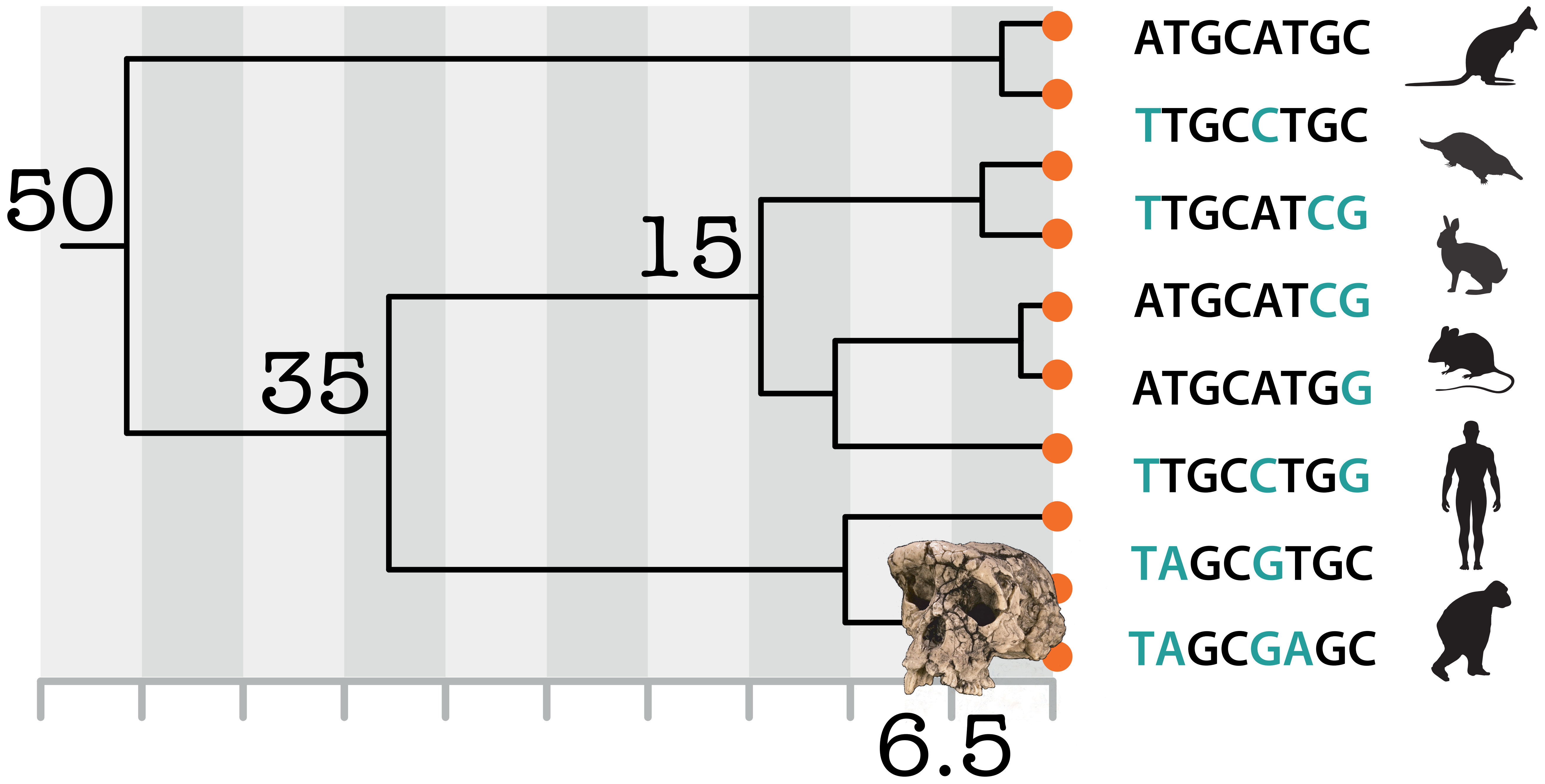
If we have independent evidence of time, we can calibrate the substitution rate



branch lengths = time

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

If we have independent evidence of time, we can calibrate the substitution rate

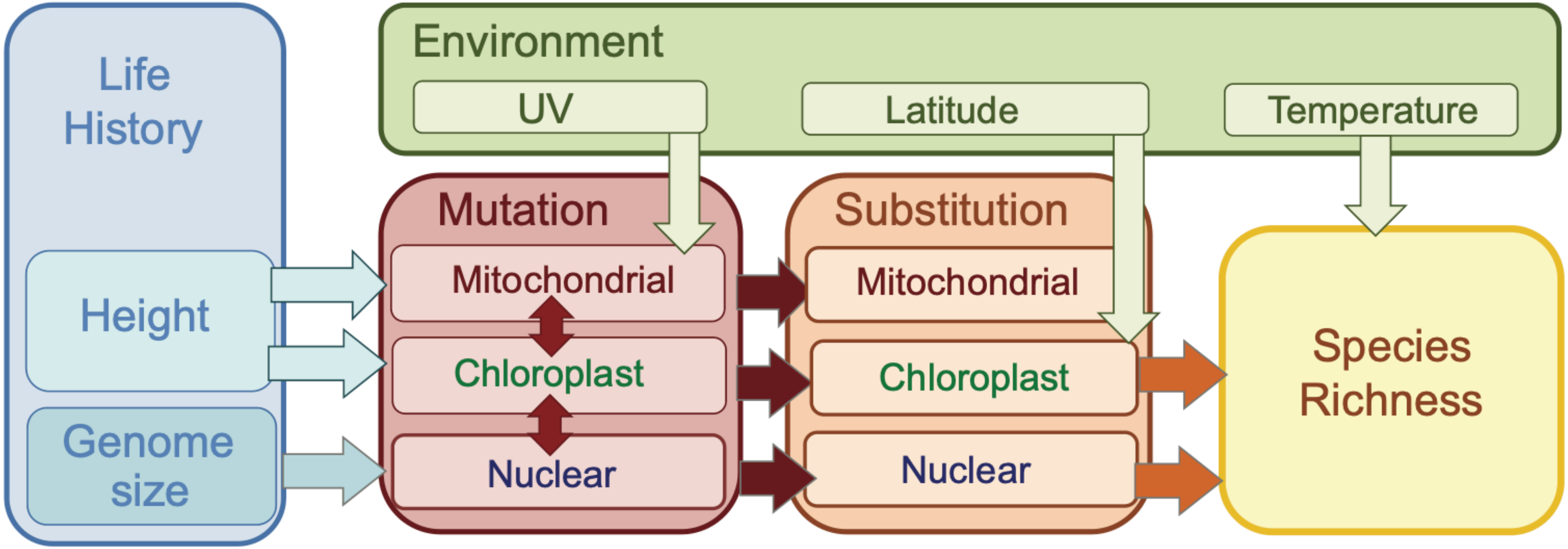


branch lengths = time

...and use this to extrapolate the divergence times for other species pairs.

The molecular clock: challenges

Many variables contribute to variation in the substitution rate.



Bromham et al. (2015)

The molecular clock: challenges

The molecular clock is not constant over time.

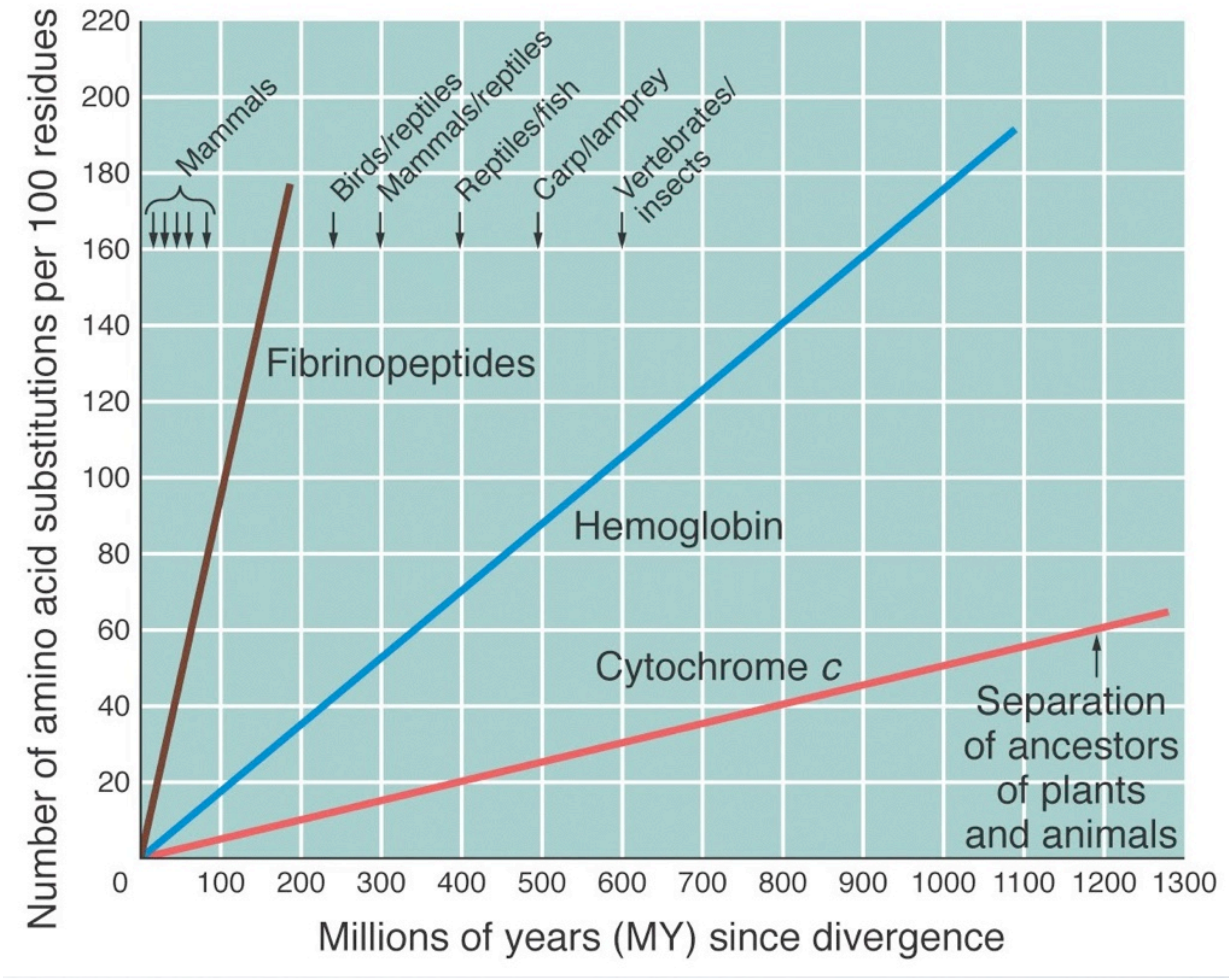
- Rates vary across taxa / time / genes / sites within the same gene

Bromham et al. (2015)

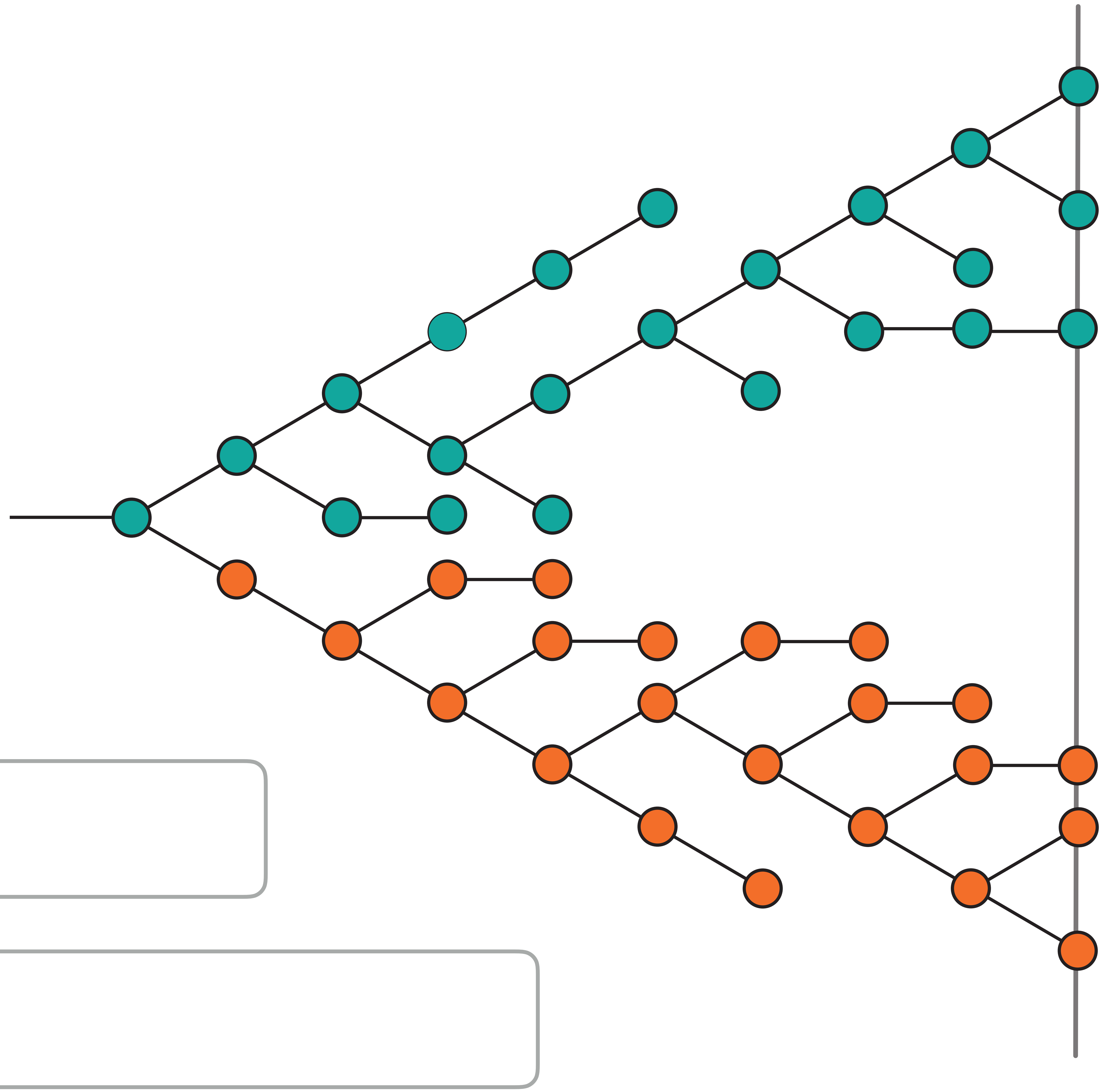
The molecular clock: challenges

The molecular clock is not constant over time.

- Rates vary across taxa / time / genes / sites within the same gene



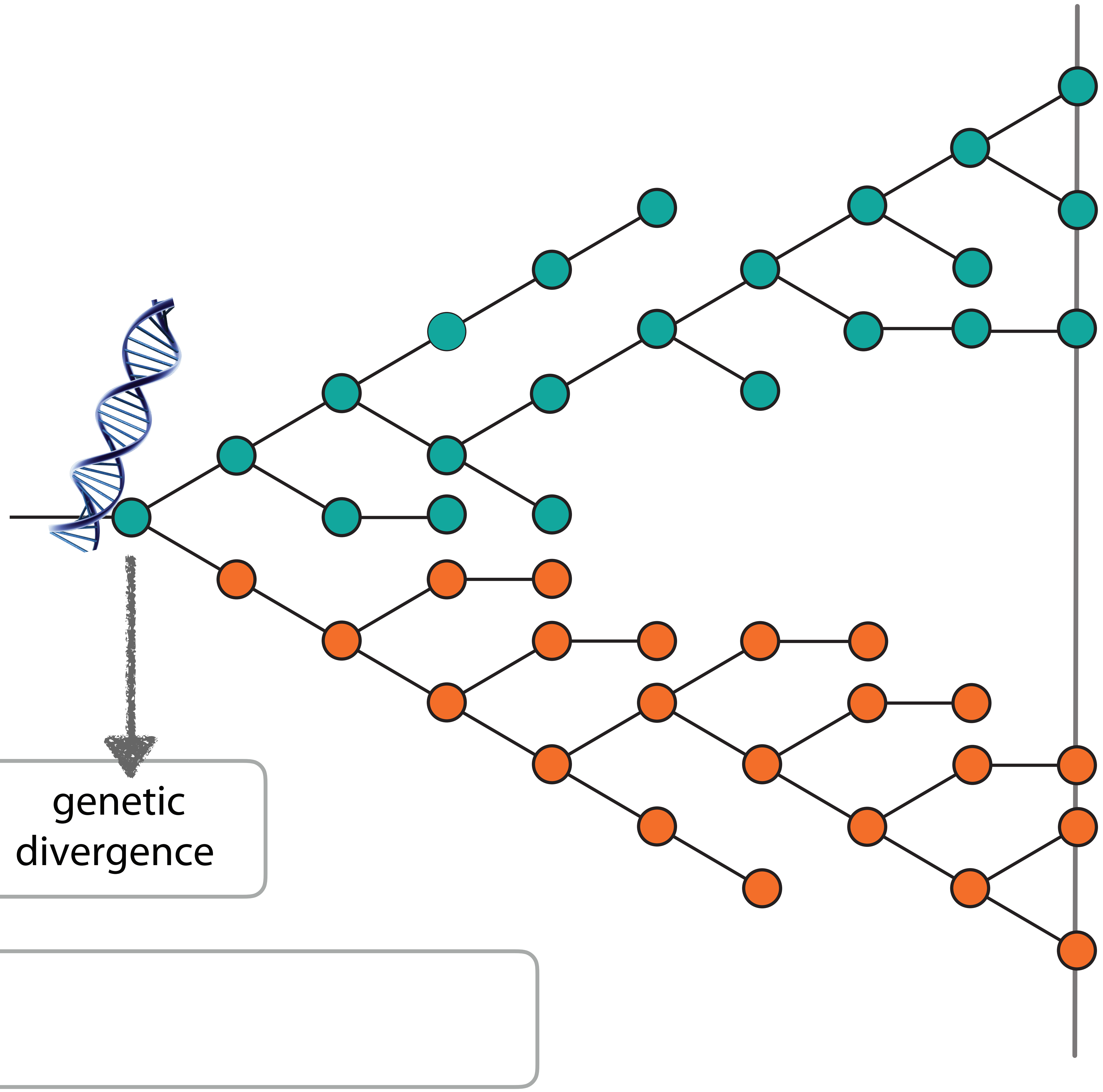
Variation in rate makes different genes useful for different timescales.



Molecular evolution:

Morphological evolution:

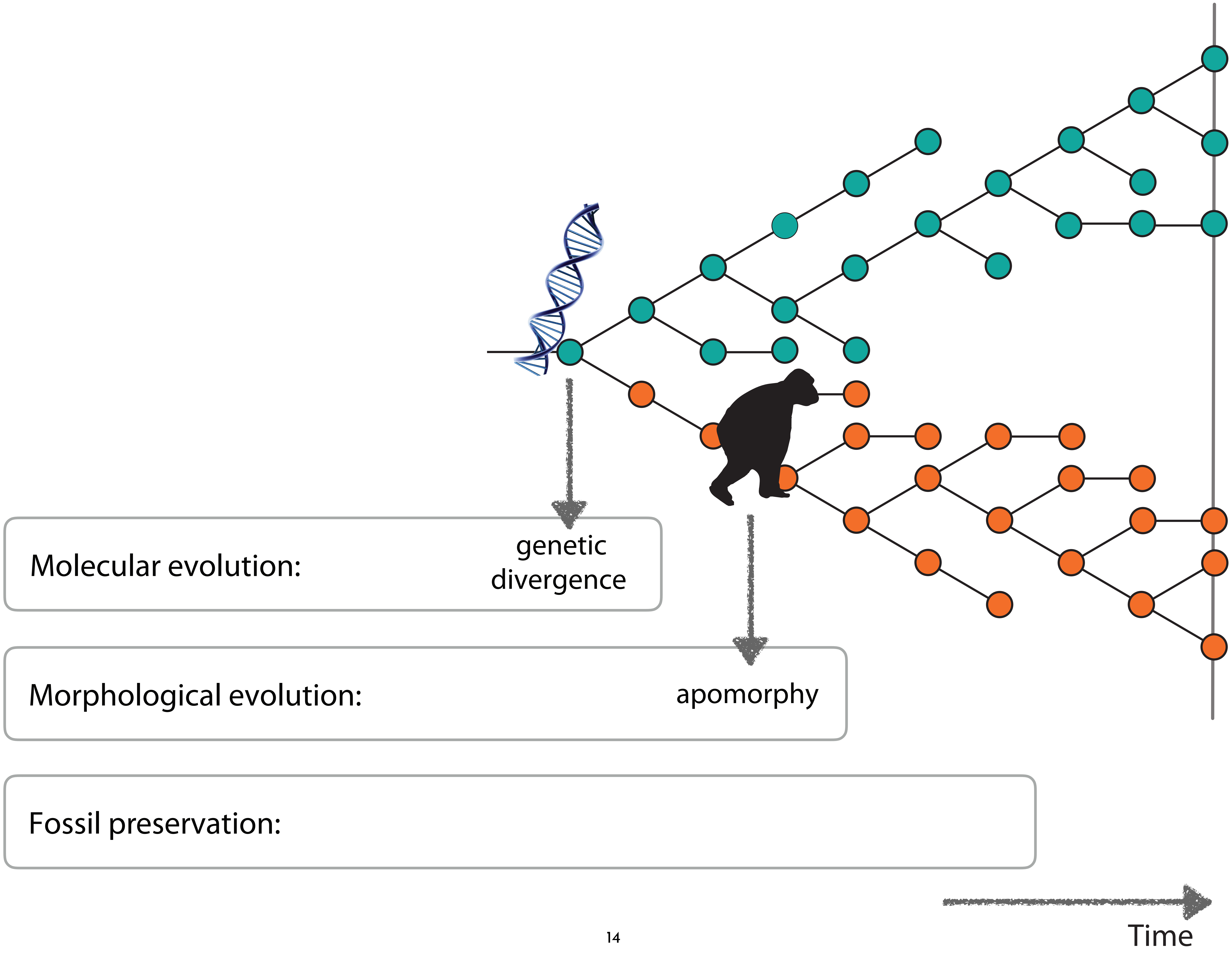
Fossil preservation:



Molecular evolution: genetic divergence

Morphological evolution:

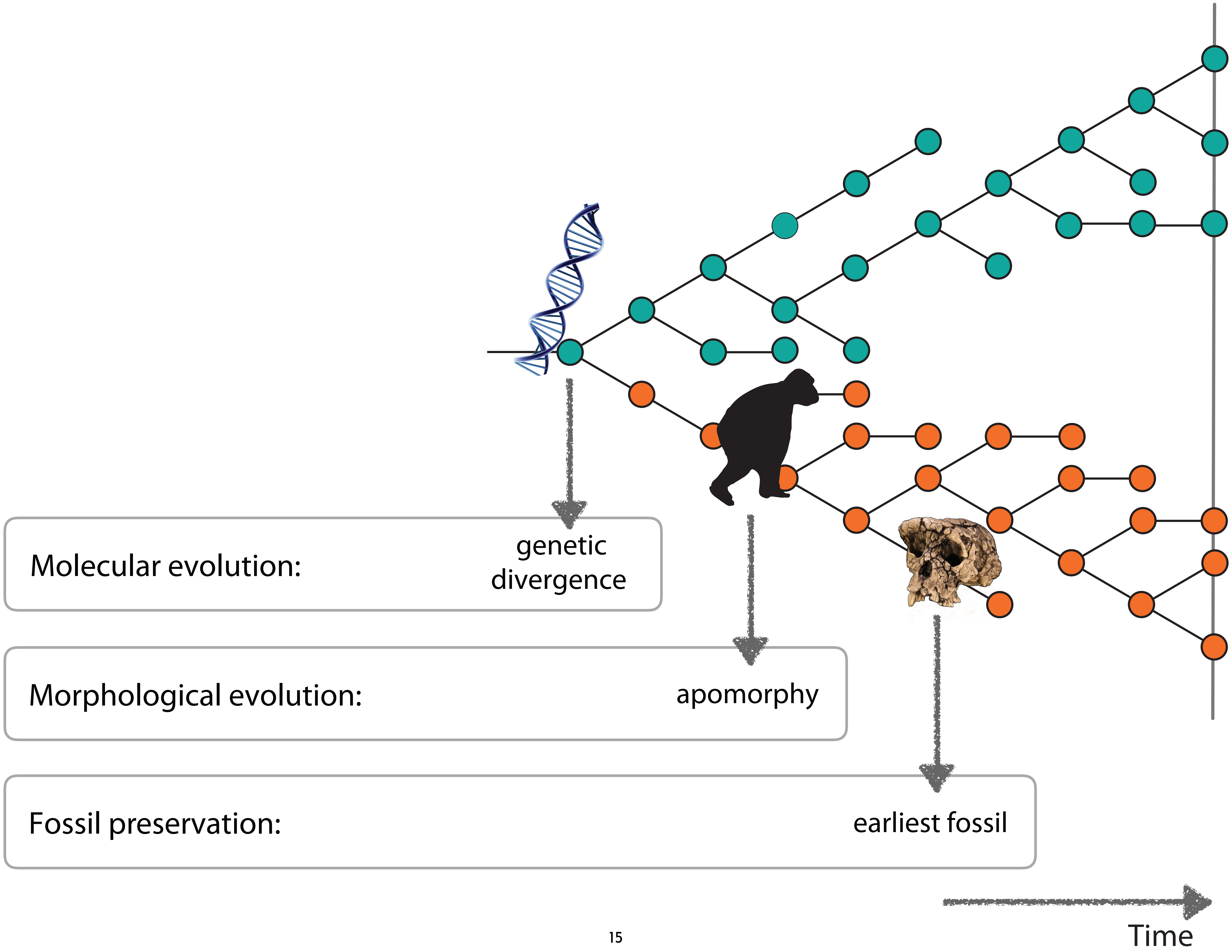
Fossil preservation:



Molecular evolution: genetic divergence

Morphological evolution: apomorphy

Fossil preservation:

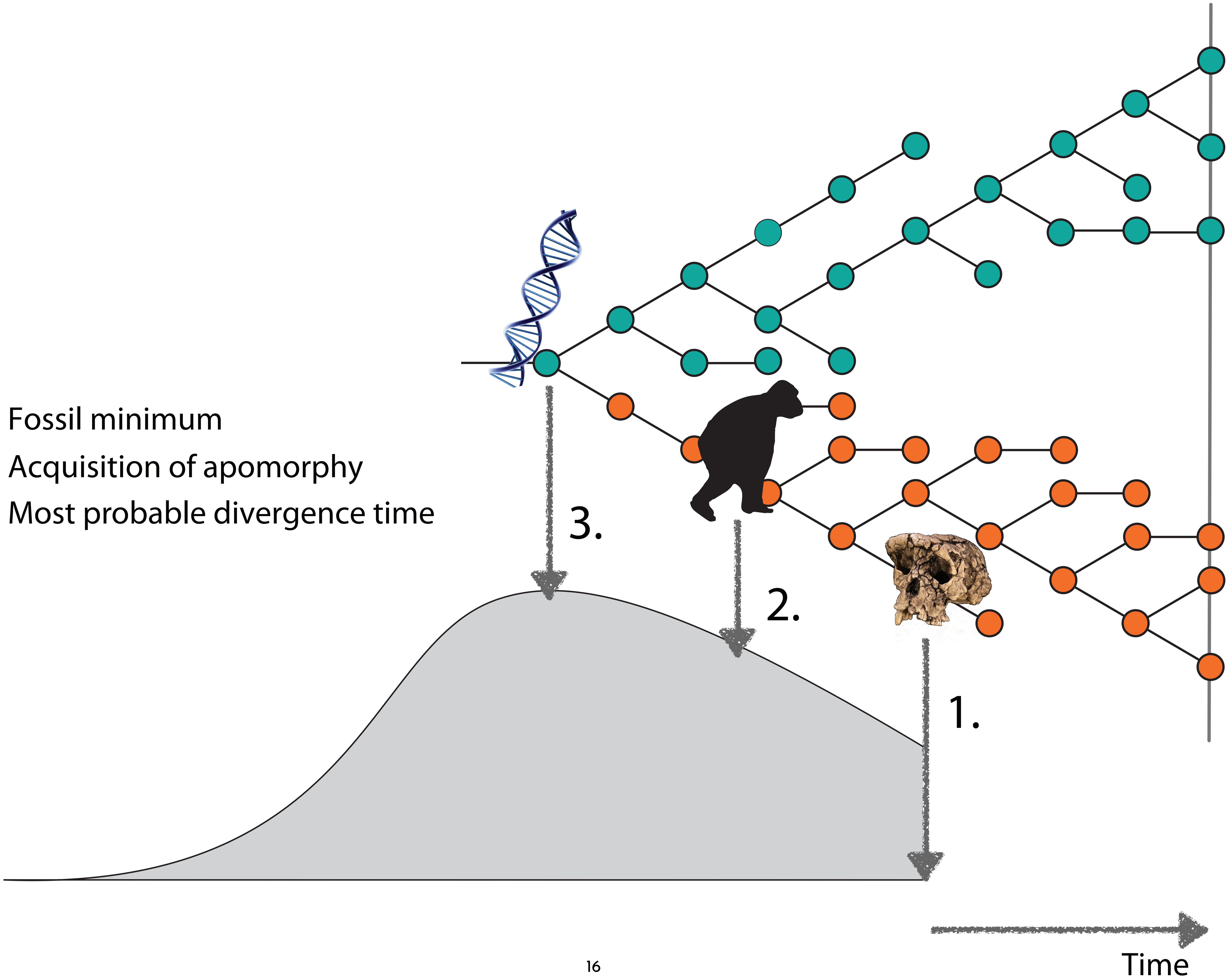


Molecular evolution: genetic divergence

Morphological evolution: apomorphy

Fossil preservation: earliest fossil

Time



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

The molecular clock: challenges

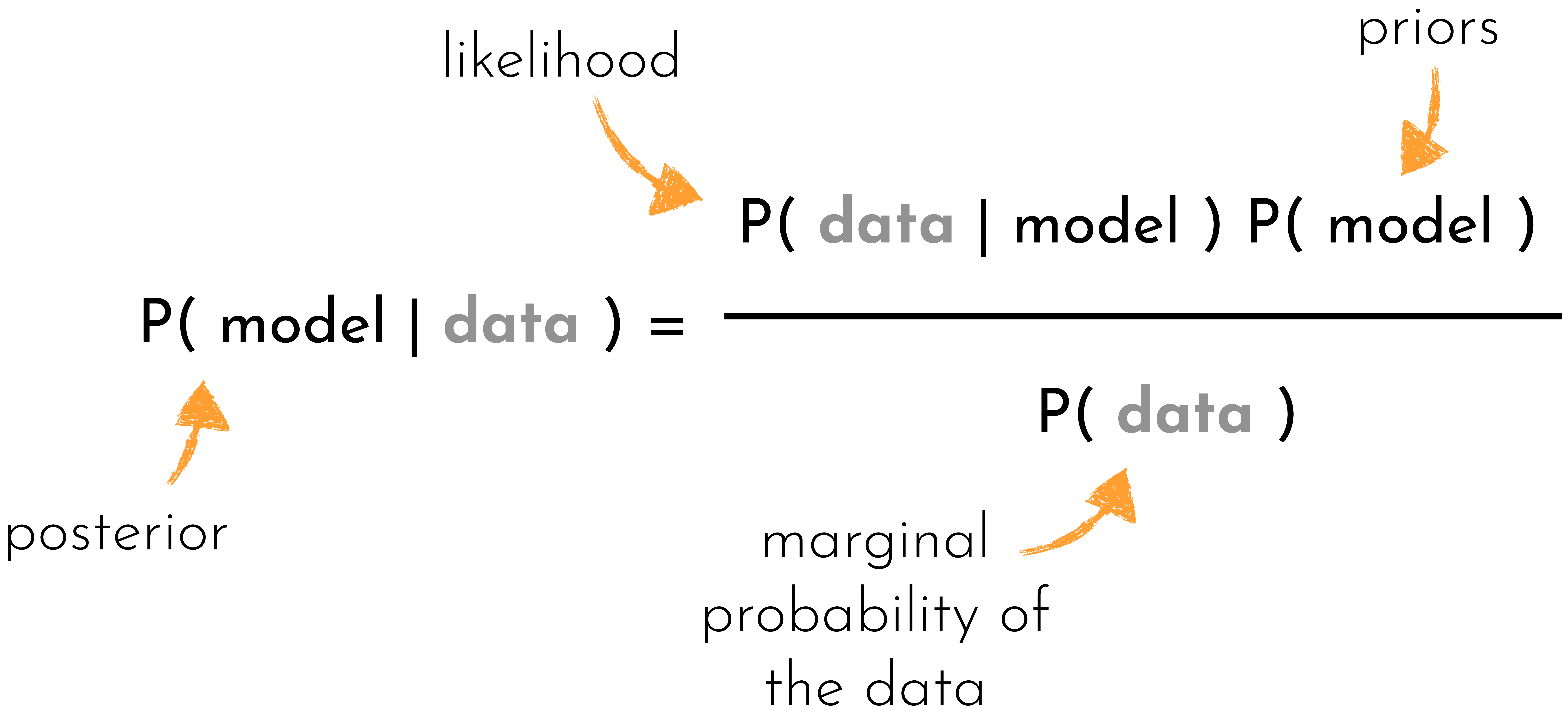
The molecular clock is not constant over time.

- Rates vary across taxa / time / genes / sites within the same gene

Calibrations are rarely known precisely.

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

We use a Bayesian framework



likelihood

priors

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

posterior

marginal probability of the data

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

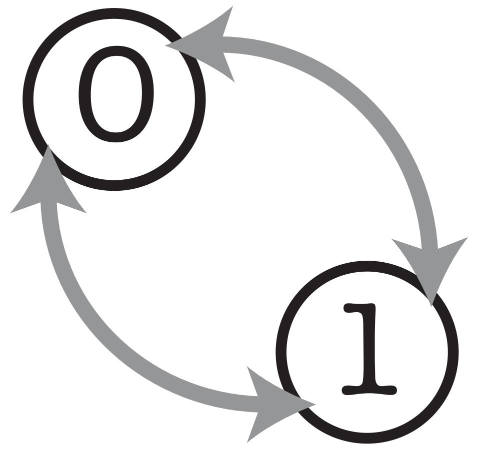
AND/OR

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

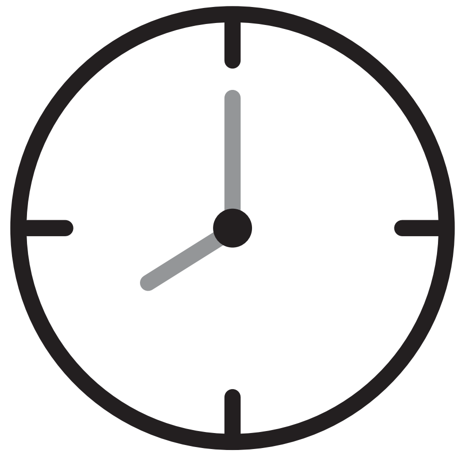
phylogenetics characters

sample ages

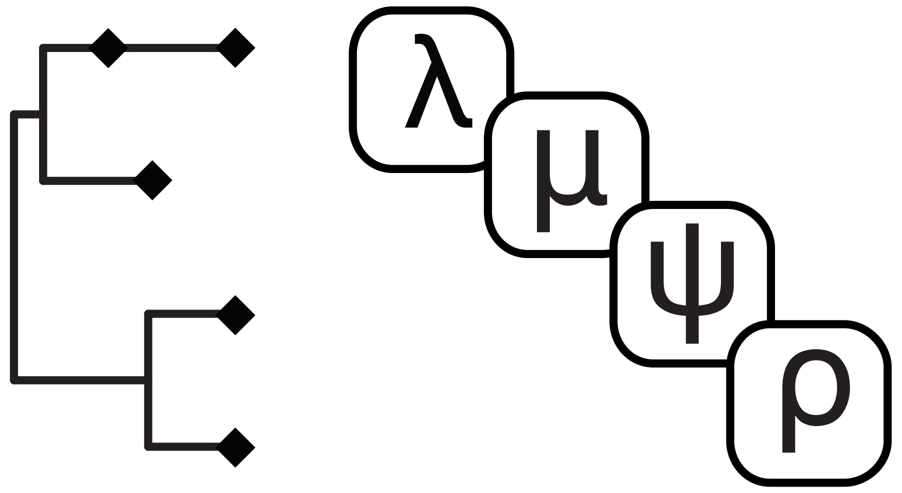
3 model components



substitution model

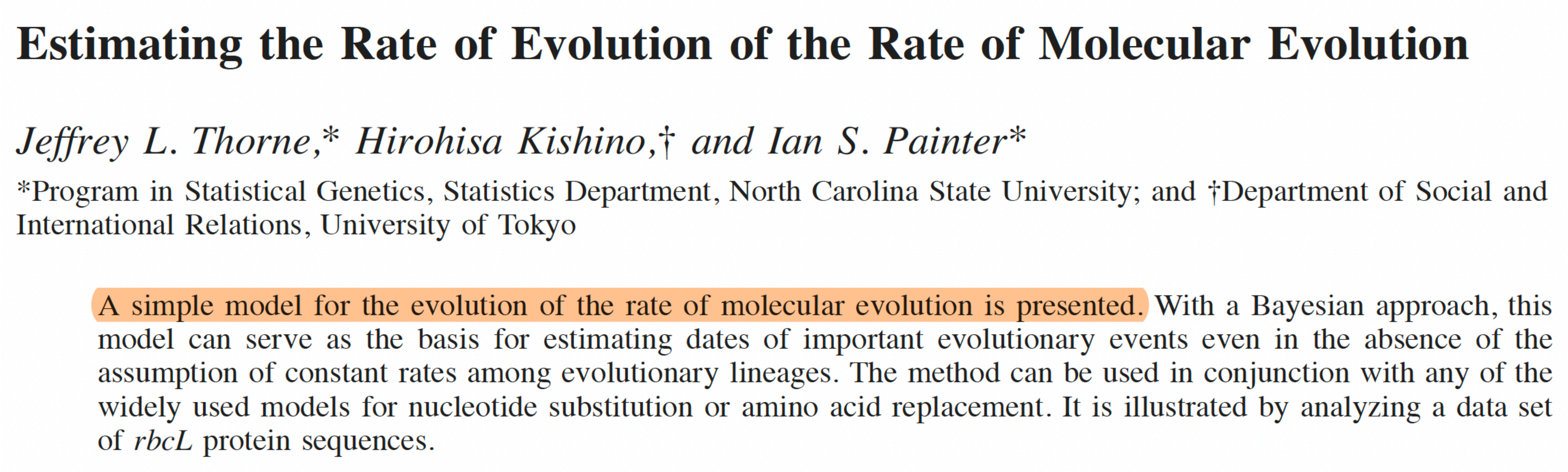


clock model



tree and tree model

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



Estimating the Rate of Evolution of the Rate of Molecular Evolution

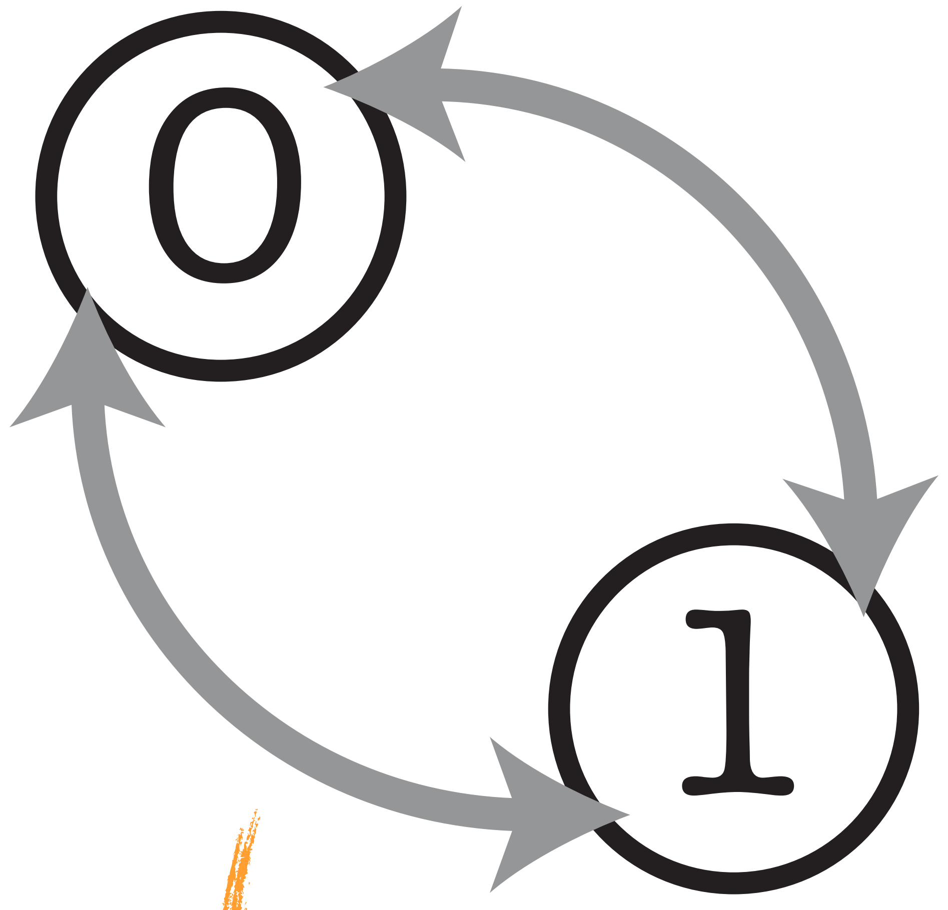
Jeffrey L. Thorne, Hirohisa Kishino,† and Ian S. Painter**

*Program in Statistical Genetics, Statistics Department, North Carolina State University; and †Department of Social and International Relations, University of Tokyo

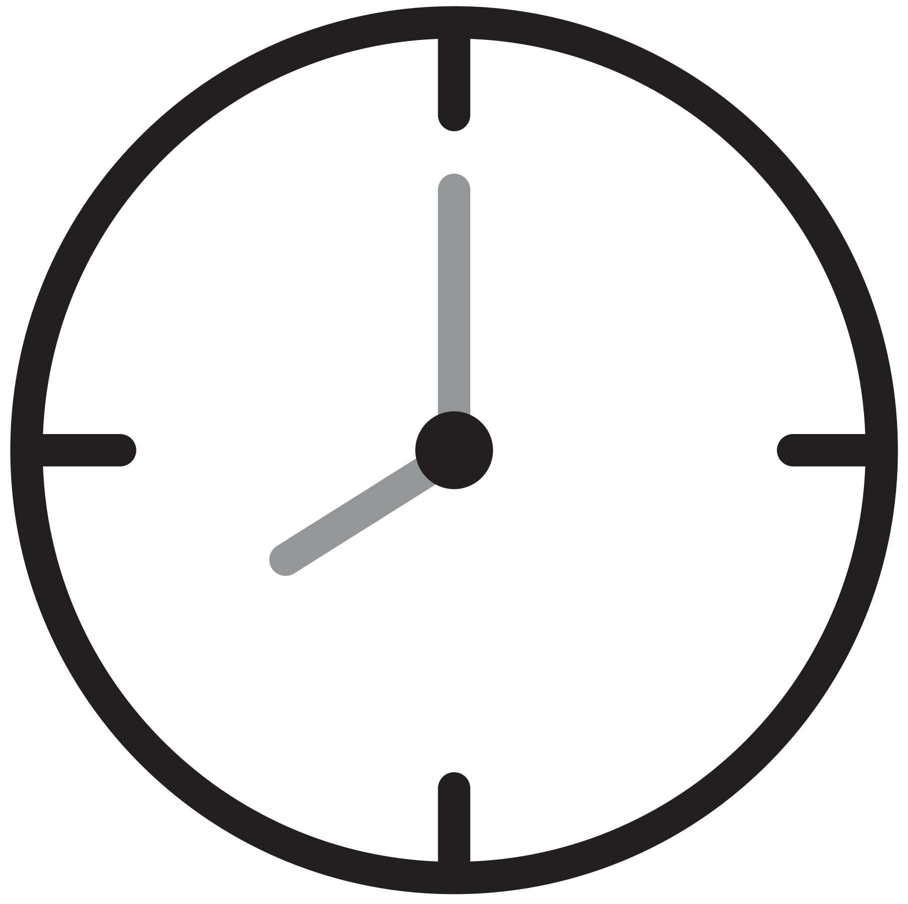
A simple model for the evolution of the rate of molecular evolution is presented. With a Bayesian approach, this model can serve as the basis for estimating dates of important evolutionary events even in the absence of the assumption of constant rates among evolutionary lineages. The method can be used in conjunction with any of the widely used models for nucleotide substitution or amino acid replacement. It is illustrated by analyzing a data set of *rbcL* protein sequences.

Thorne, Kishino, Painter. 1998. MBE
See also: Kishino, Thorne, Bruno. 2001. MBE

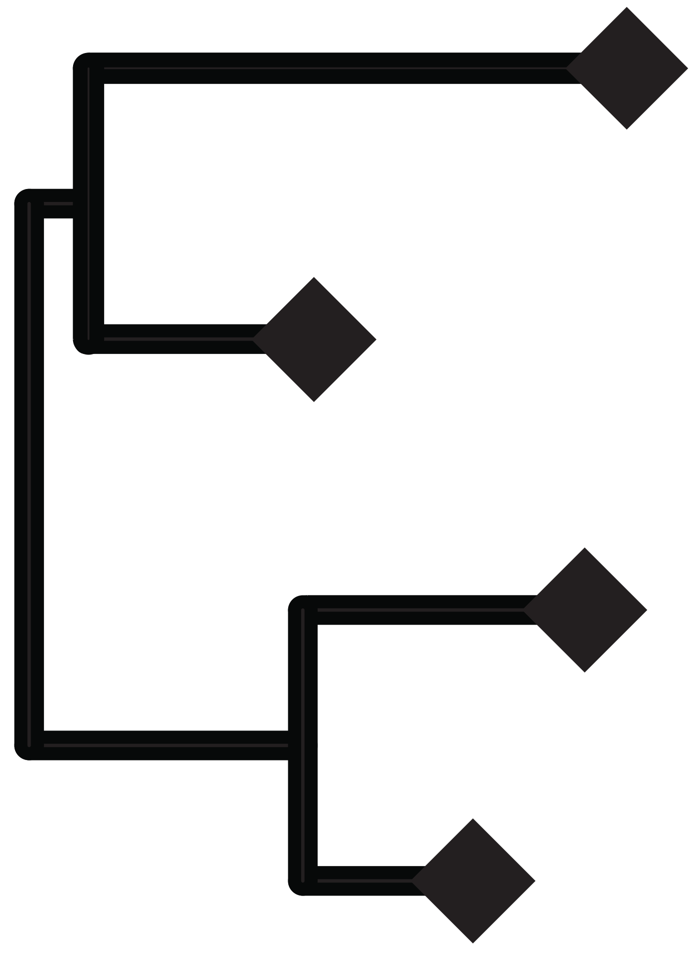
substitution model



clock model

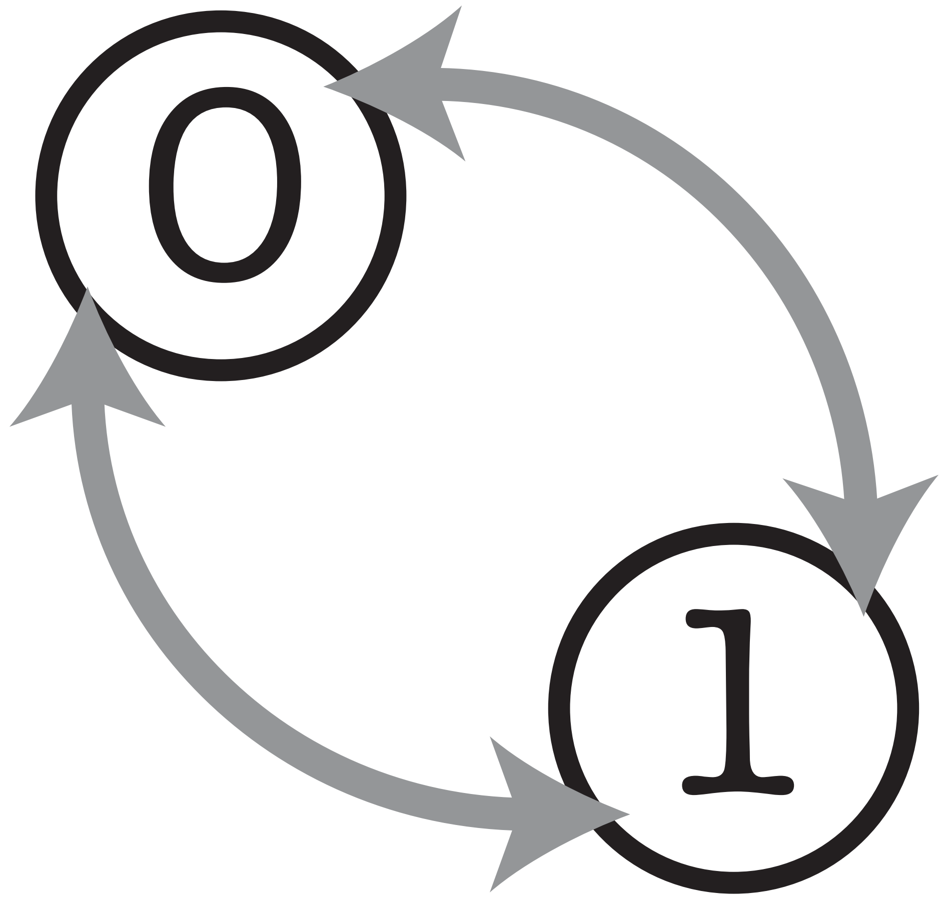


tree model

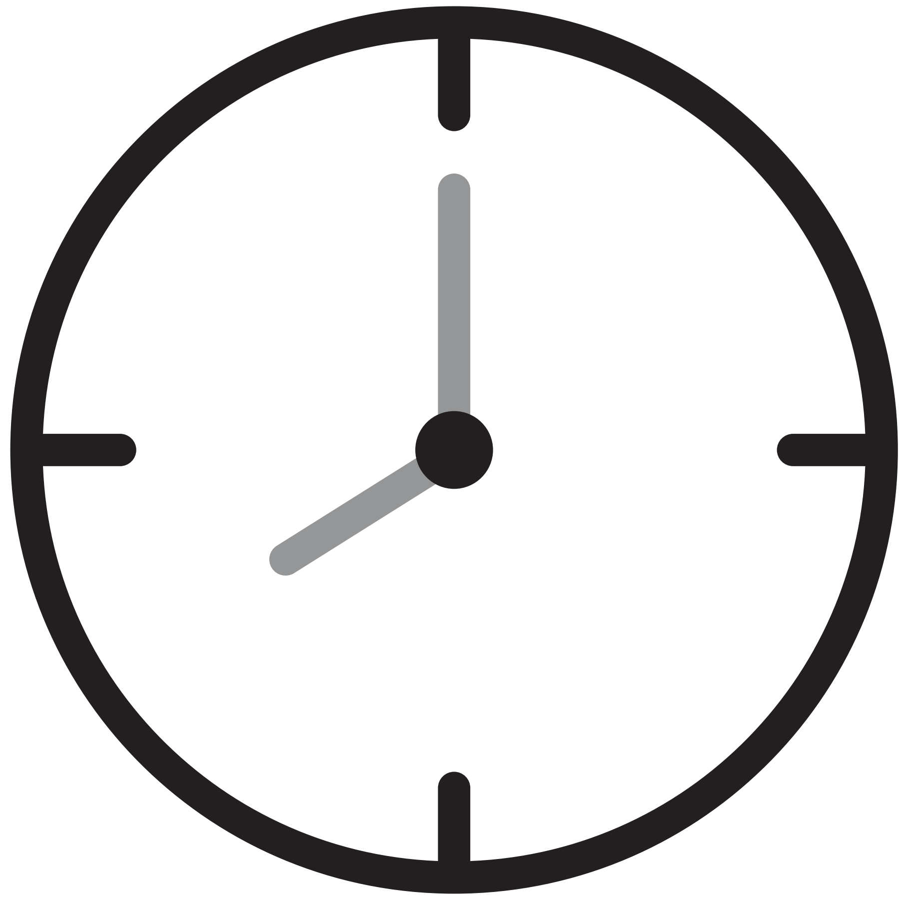


How likely are we to observe a change between character states? e.g., A → T

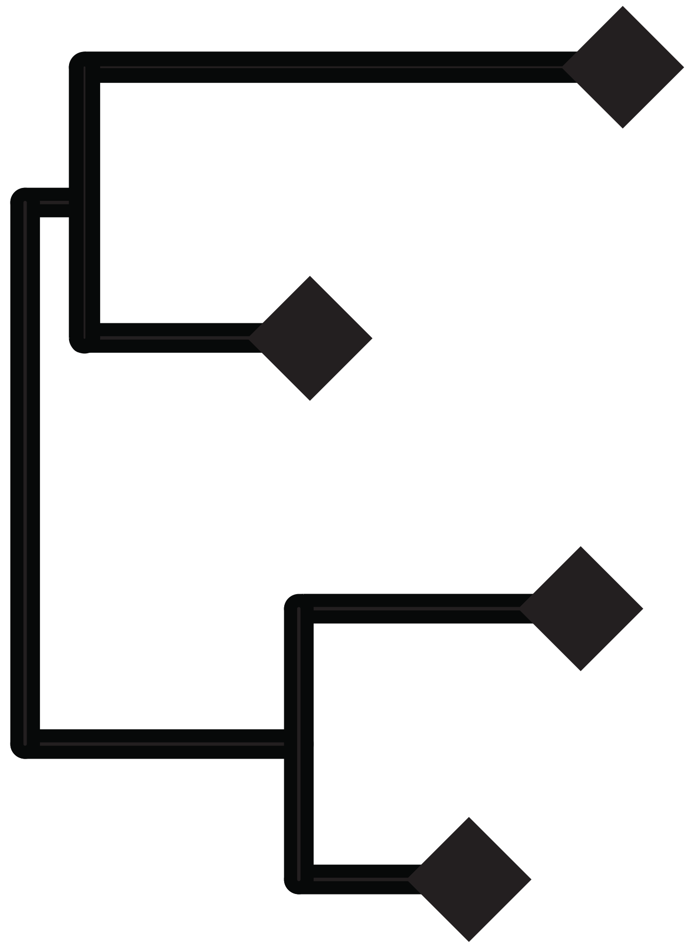
substitution model



clock model



tree model

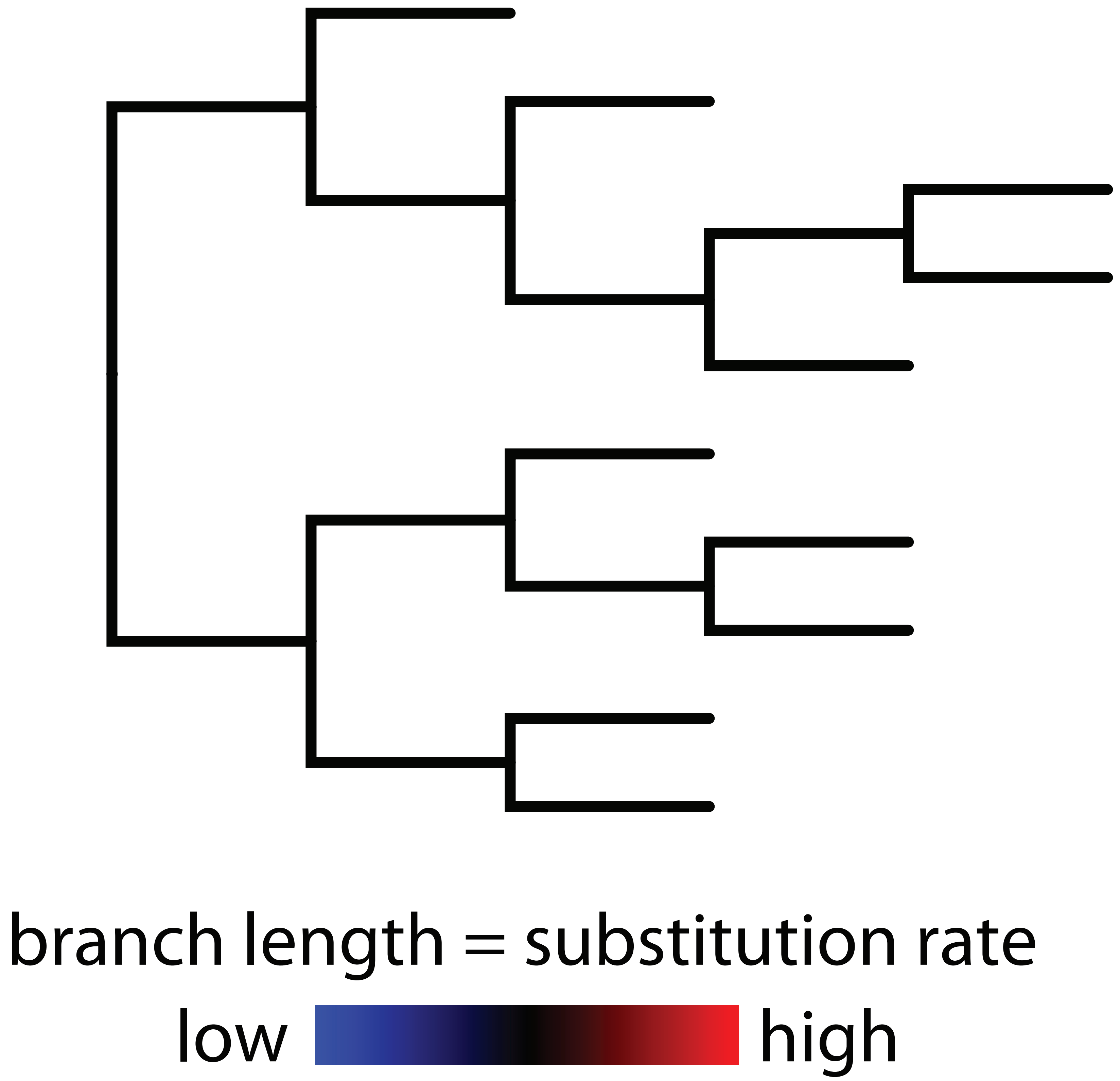


How have rates of evolution varied (or not) across the tree?

The strict molecular clock model

Assumptions:

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



branch length = substitution rate

low ▨ high

Image source: Tracy Heath

Relaxed clock models

Assumptions:

- Lineage-specific rates are independent (i.e., uncorrelated).
- The rate assigned to each branch is drawn independently from the underlying distribution.

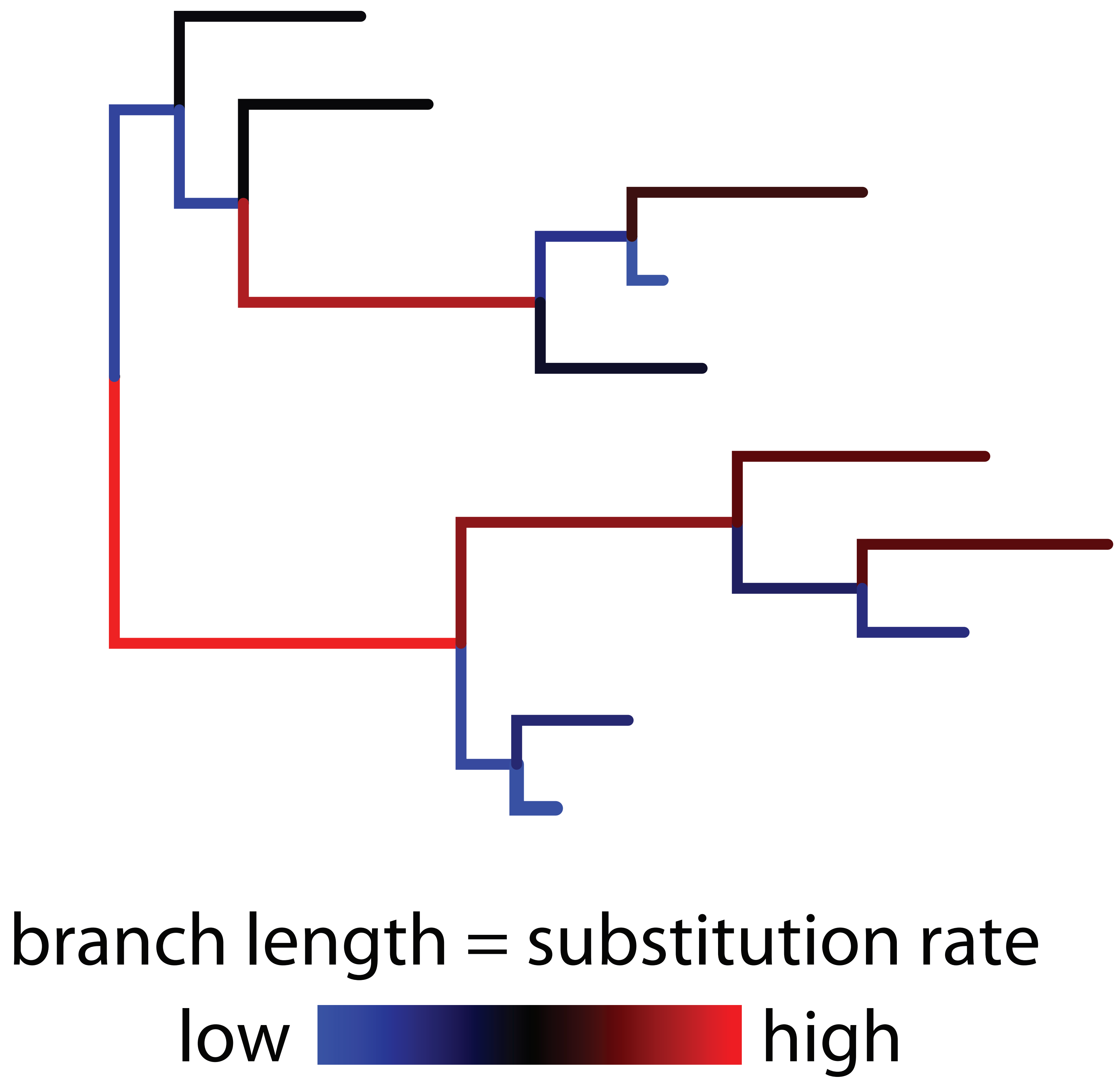
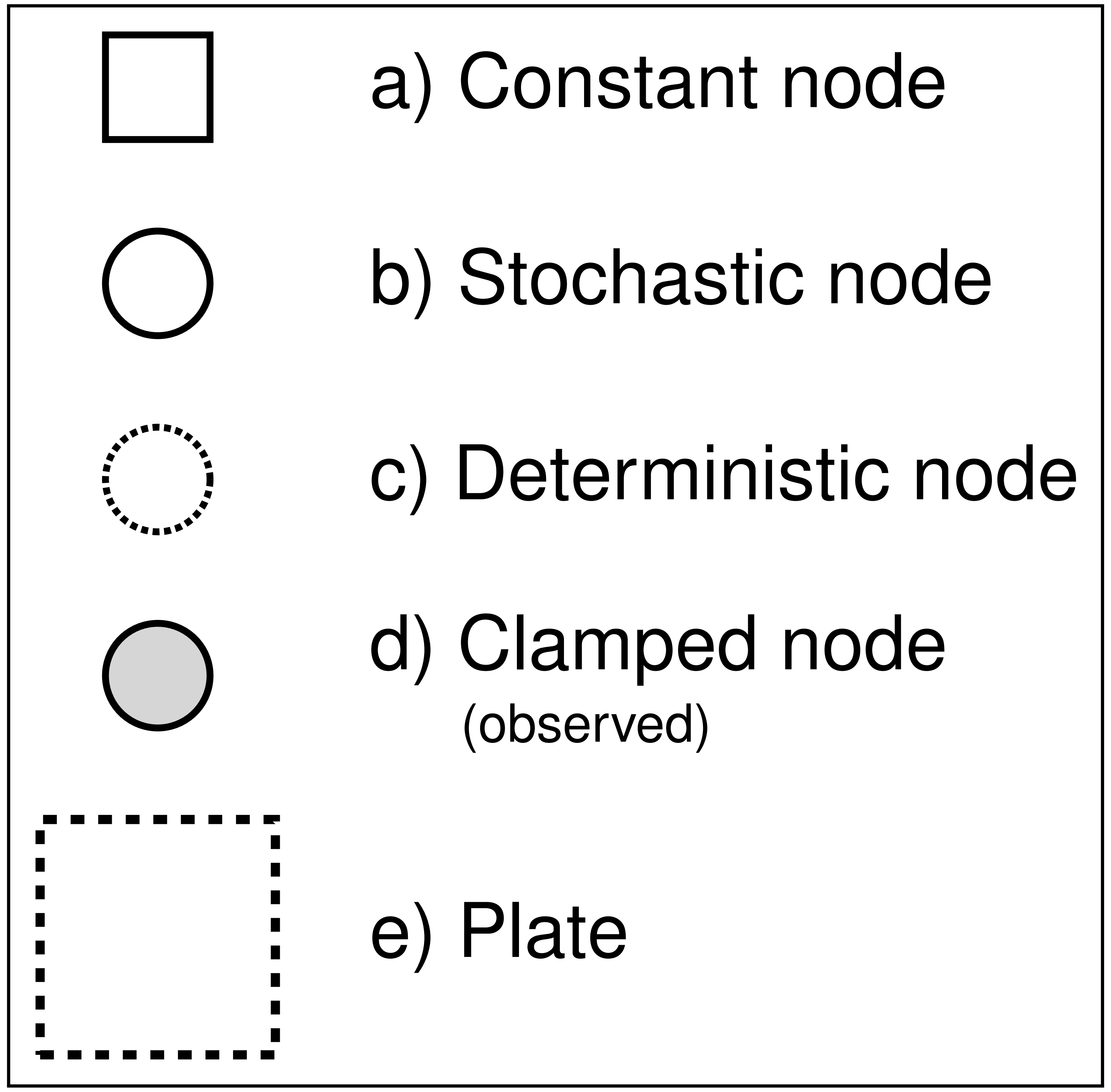
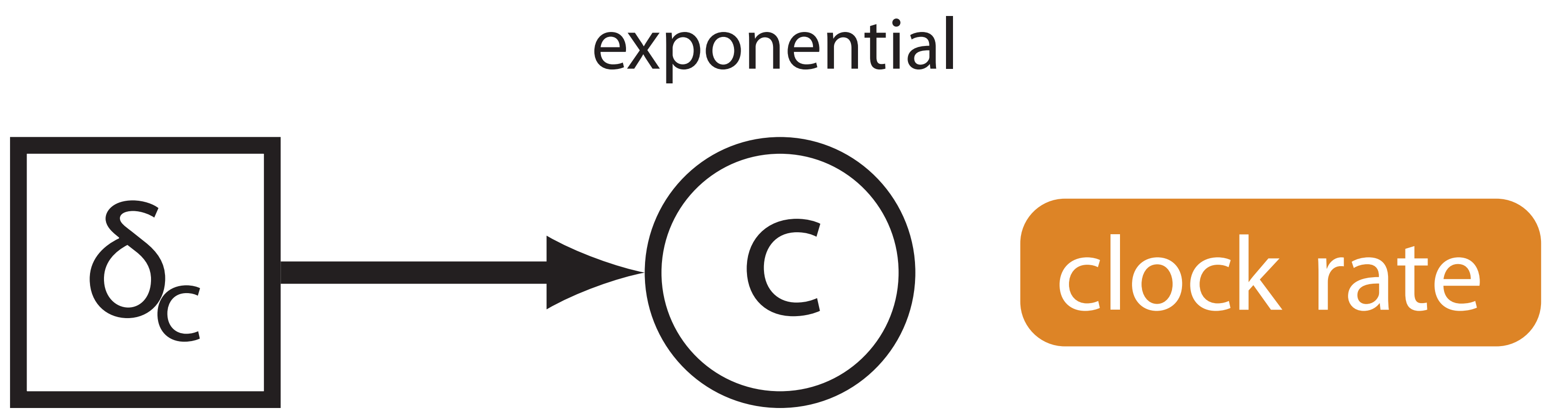


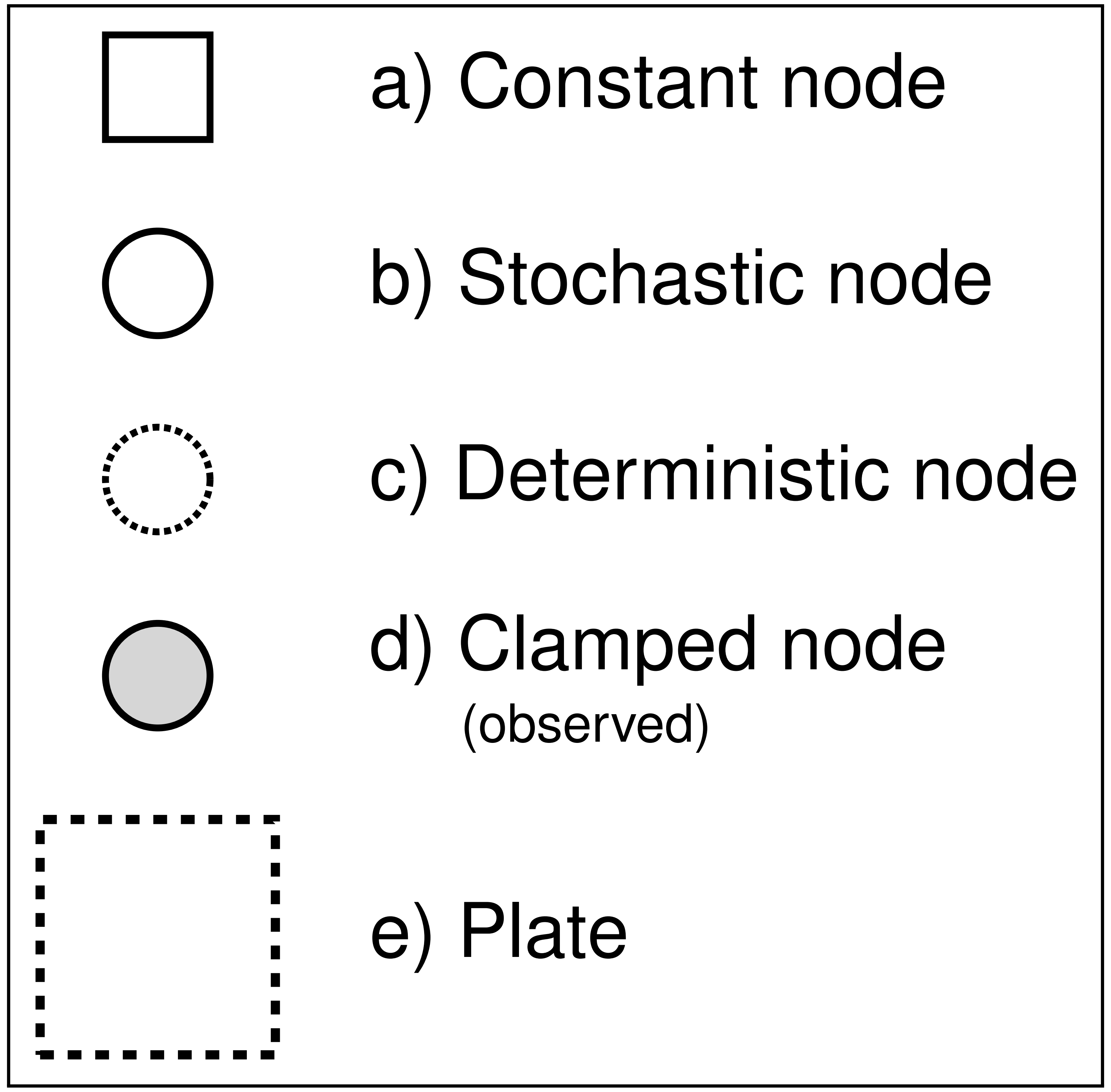
Image source: Tracy Heath

Graphical models: strict clock





Graphical models: exponential relaxed clock

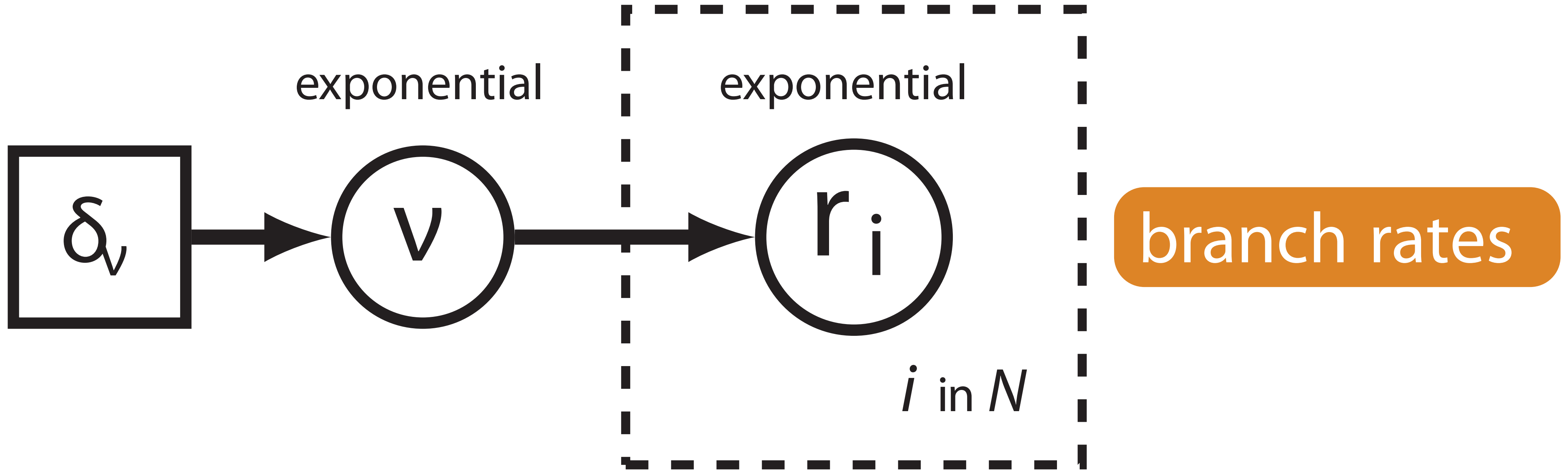

a) Constant node

b) Stochastic node

c) Deterministic node

d) Clamped node
(observed)

e) Plate



Many different clock models

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

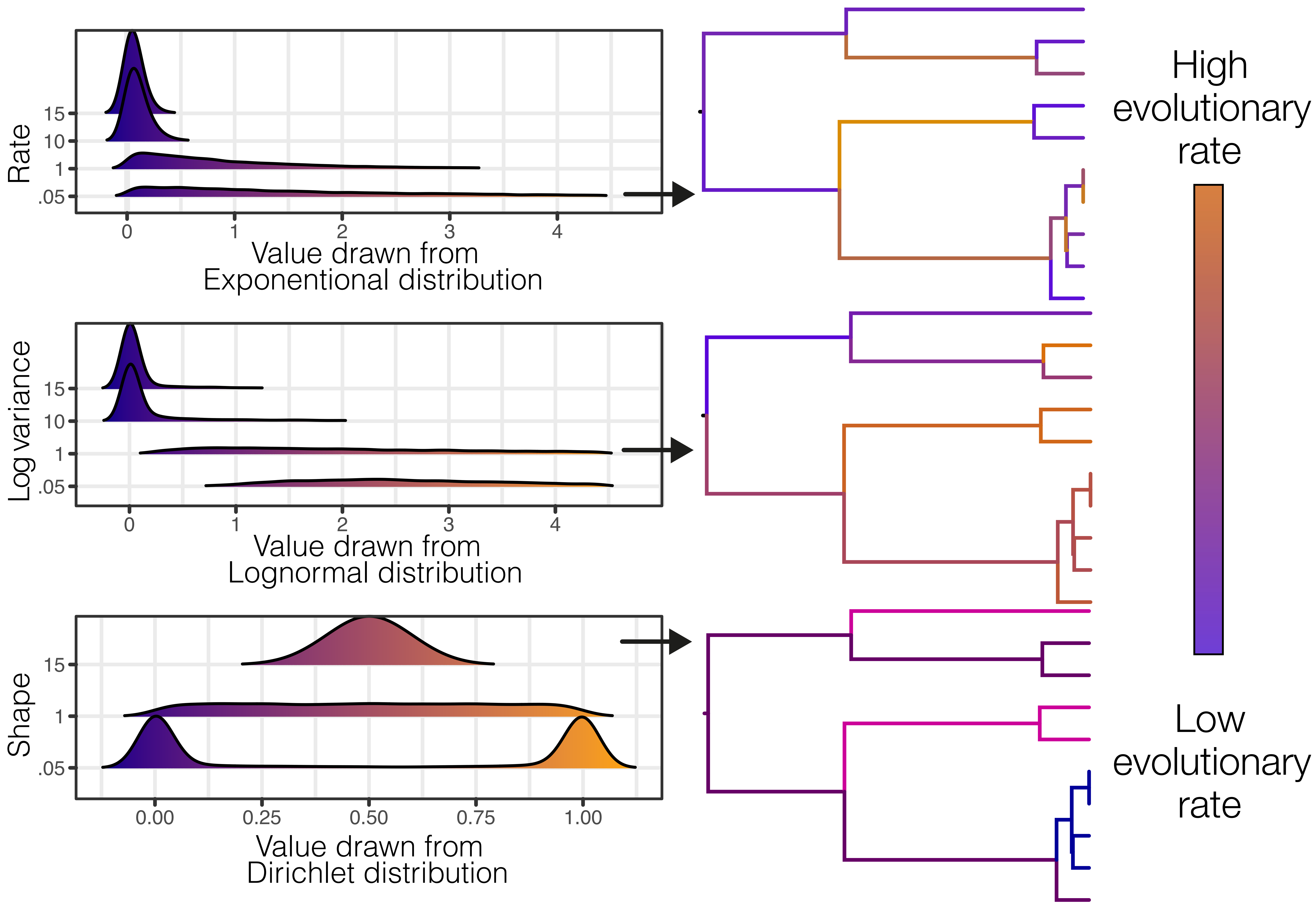
Understanding the tripartite approach to Bayesian divergence time estimation
Warnock, Wright. 2020. Elements of Paleontology

Clock type | Sample distribution | Sample tree

Uncorrelated

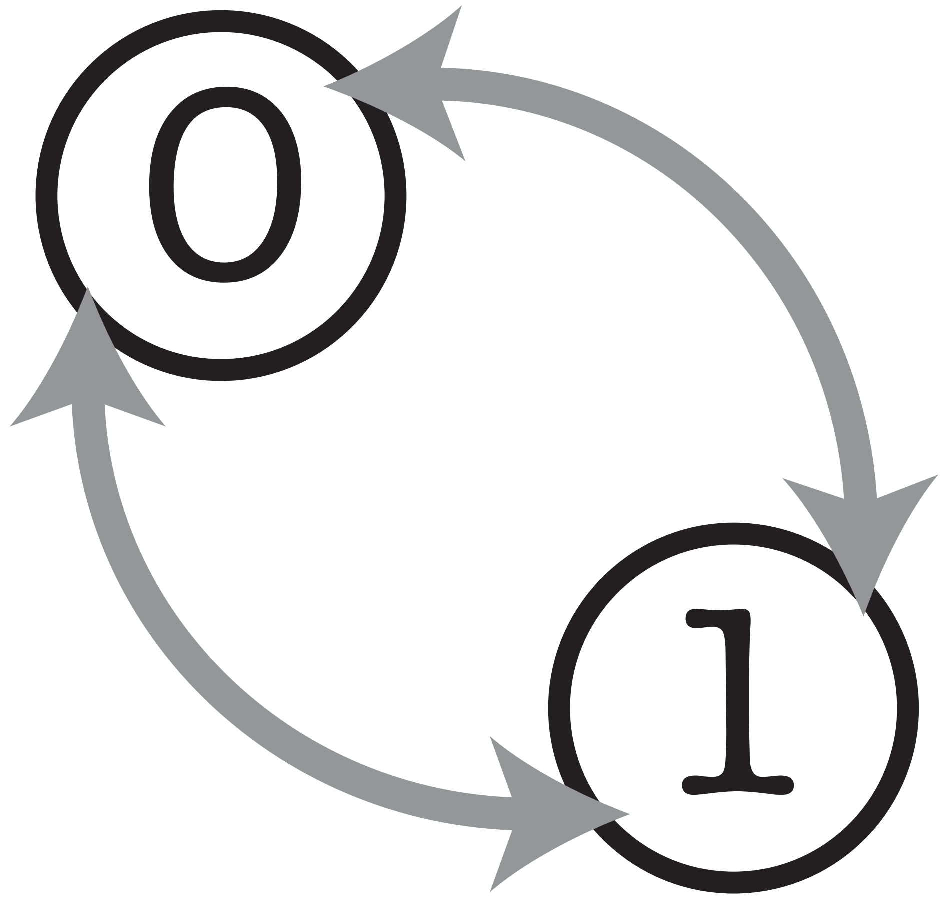
Auto-correlated

Dirichlet Process Prior

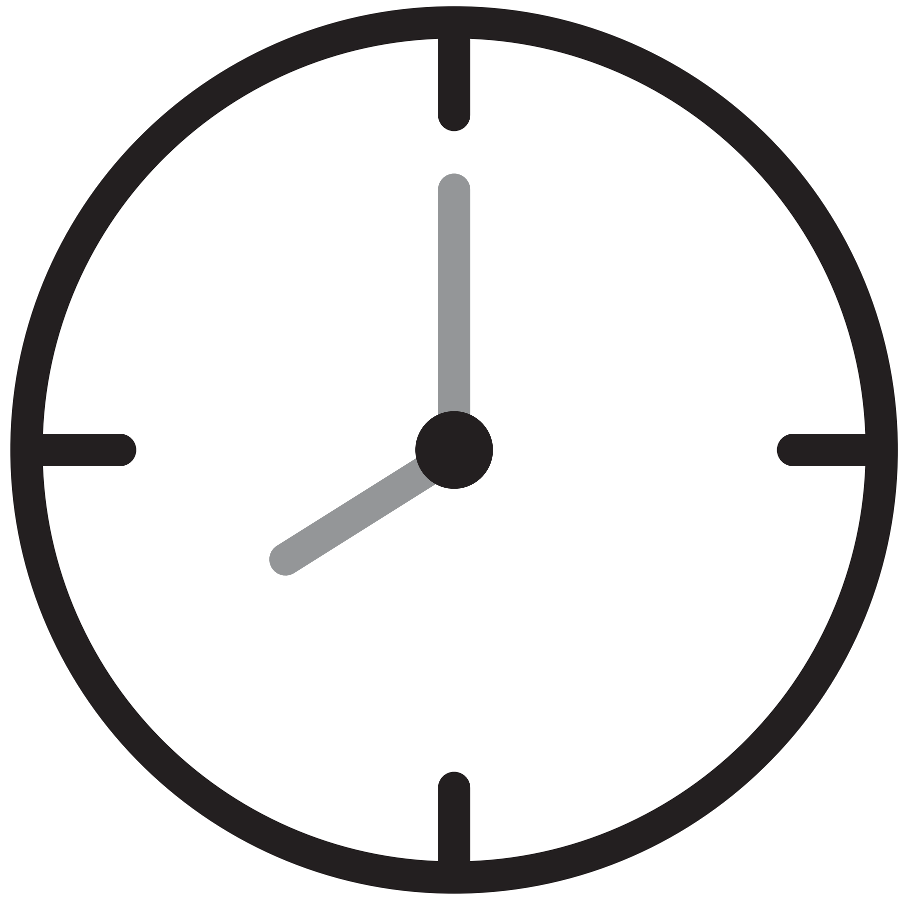


Warnock, Wright. 2020. Elements of Paleontology

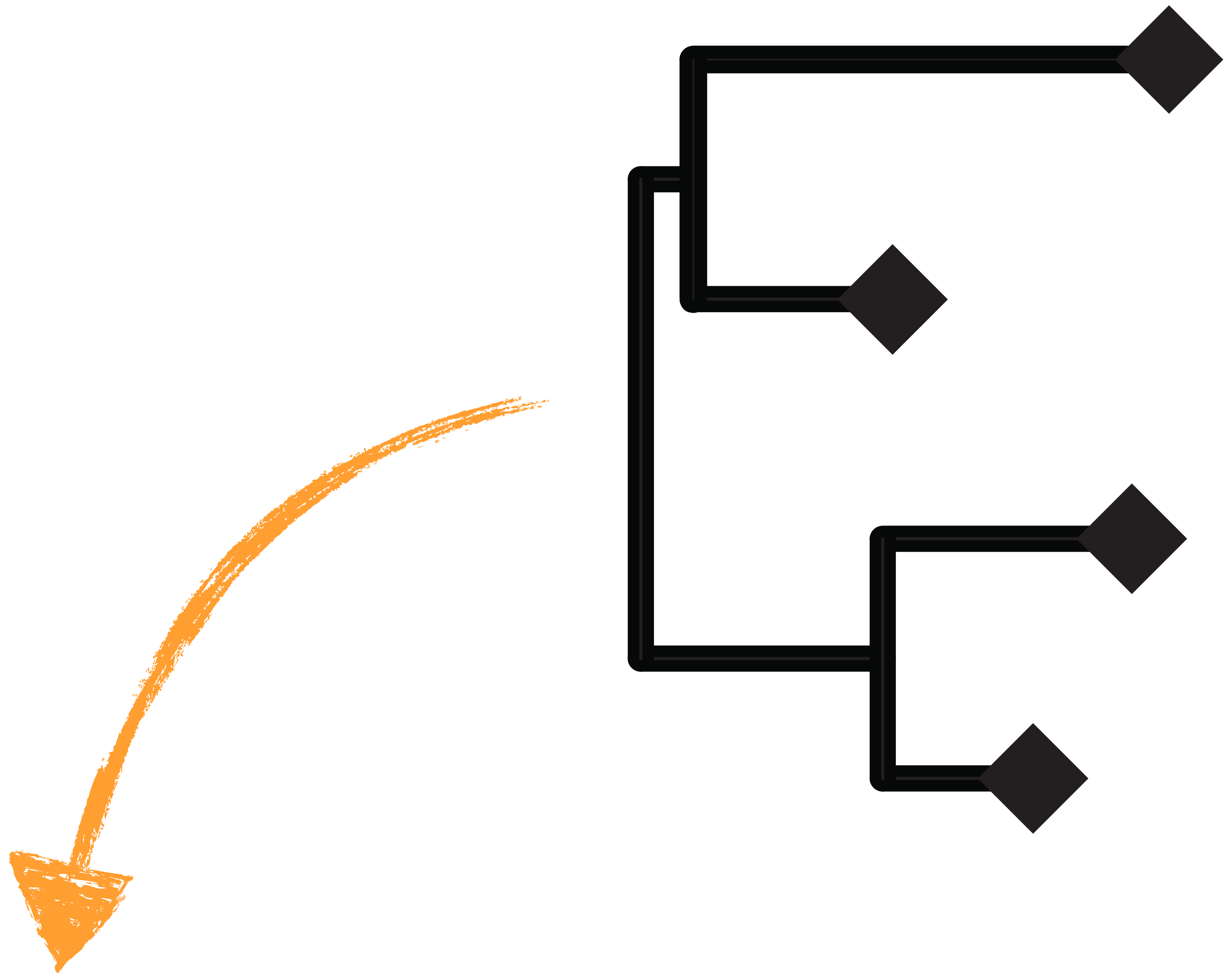
substitution model



clock model



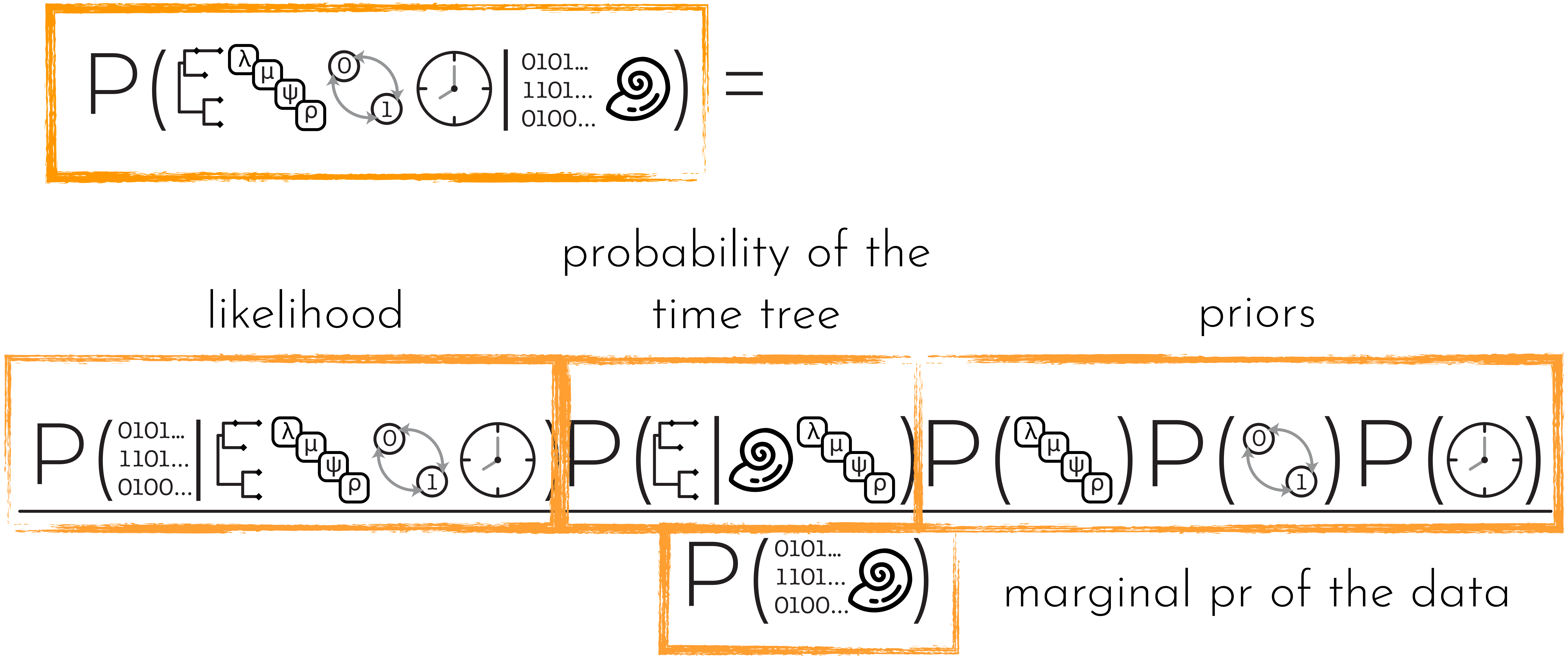
tree model



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

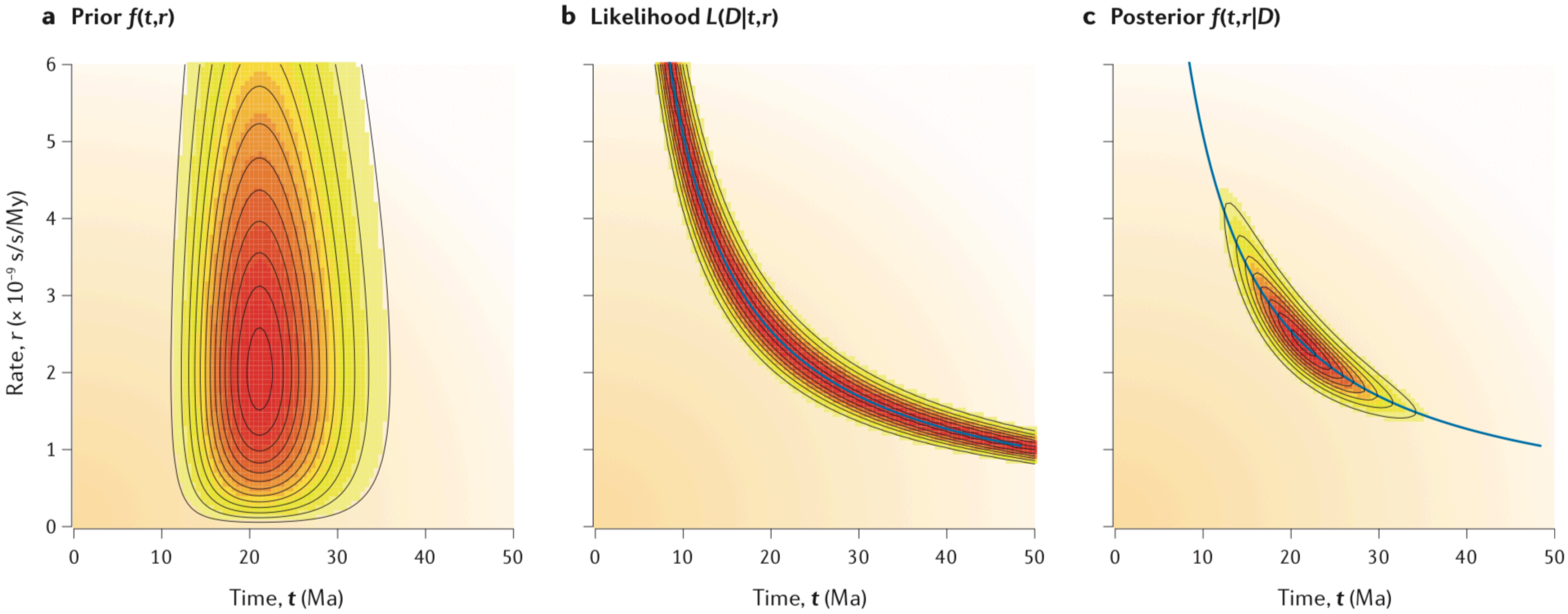
probability of the time tree

priors

marginal pr of the data

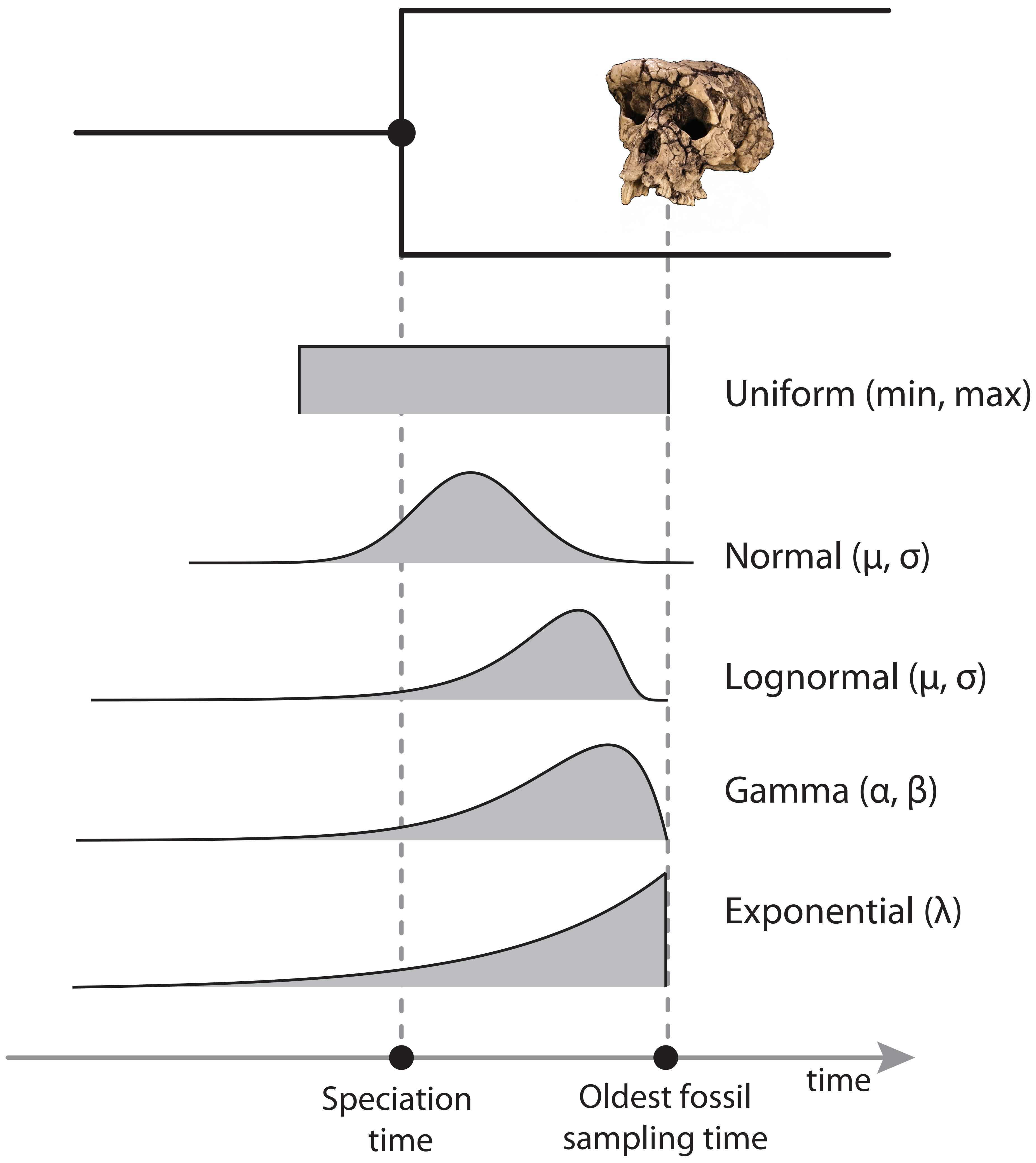
Rate and time are non-identifiabile

This mean we need relatively informative priors on the rates and times



dos Reis et al. 2016. Nature Genetics Reviews

Node dating



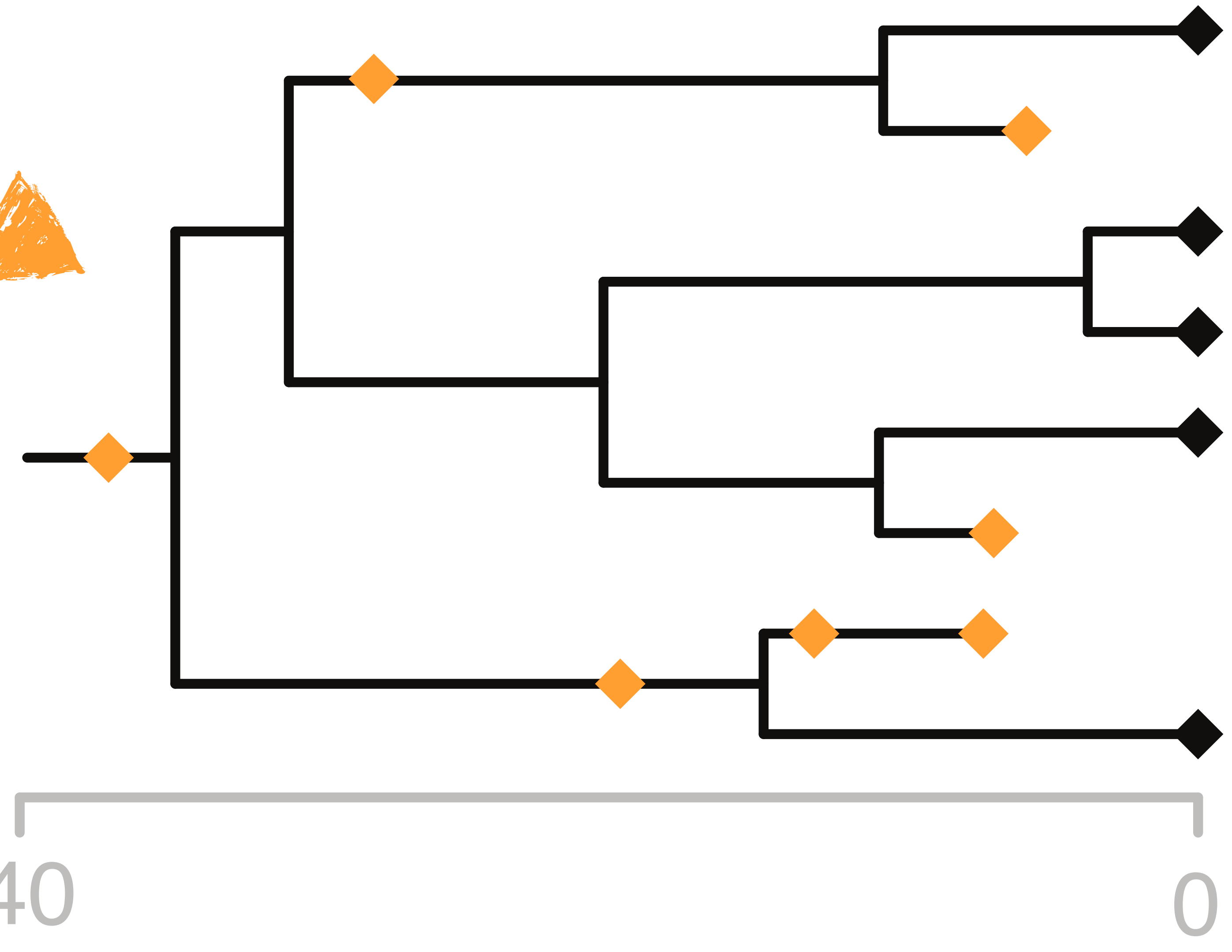
In a node dating context, we typically use a birth-death model to describe the tree generating process, given we observe <u>extant species only</u>.

Then we separately apply a calibration density to constrain internal node ages.

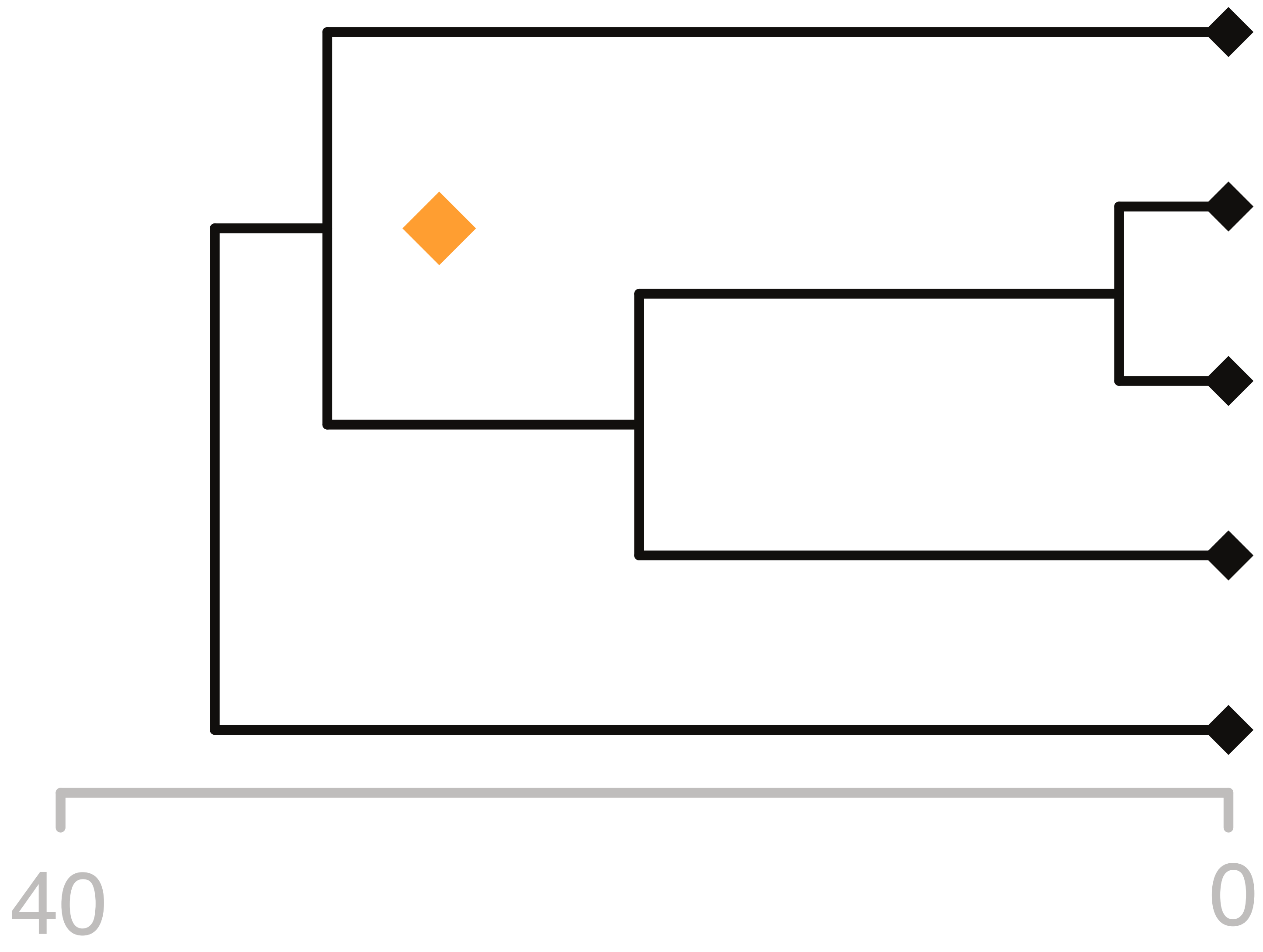
Adapted from Heath 2012. Sys Bio

Node dating: potential issues

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



sampled fossils



fossils used for node calibration – all others provide redundant info

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 Biology Letters

Is there another way?

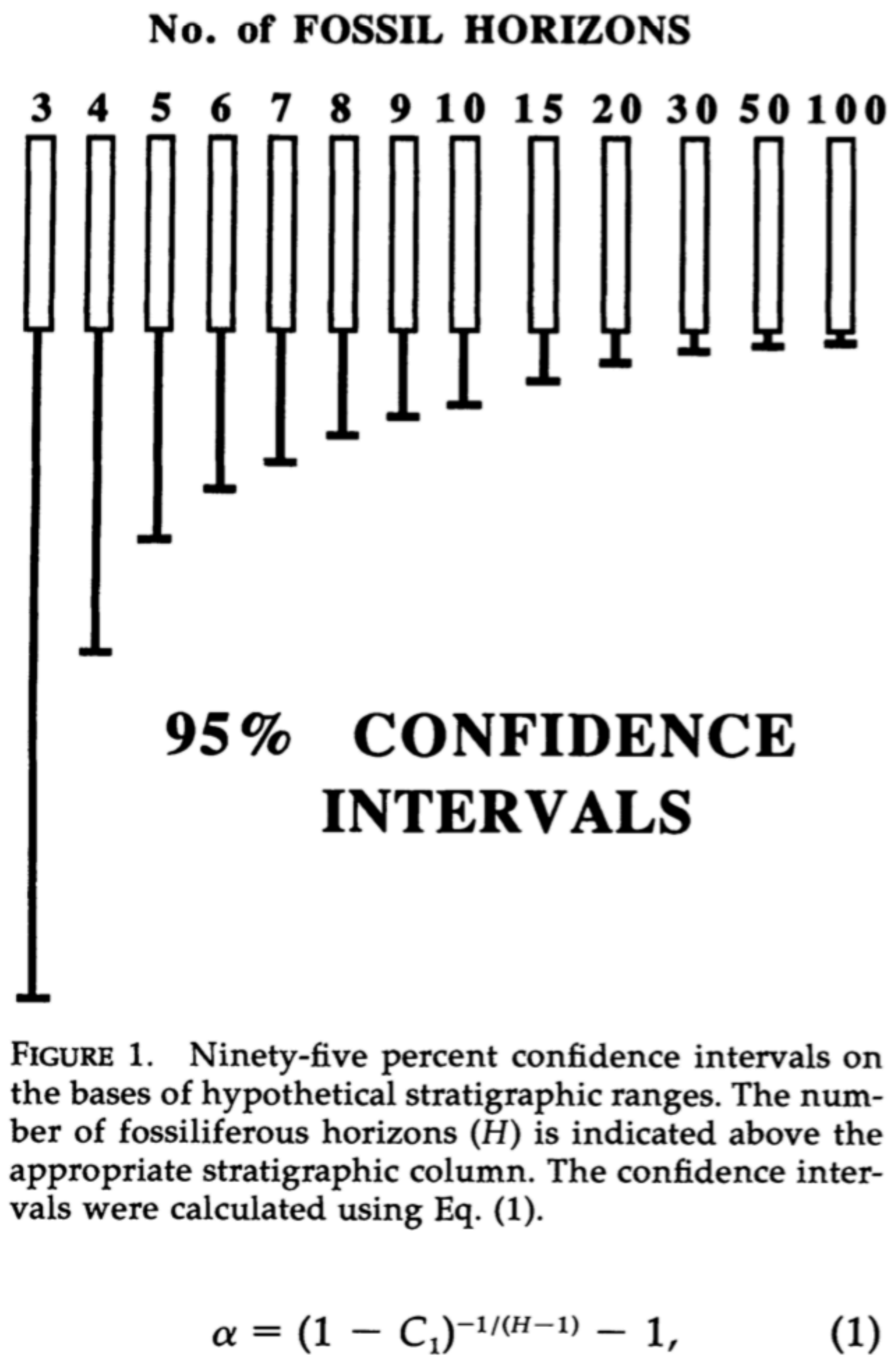
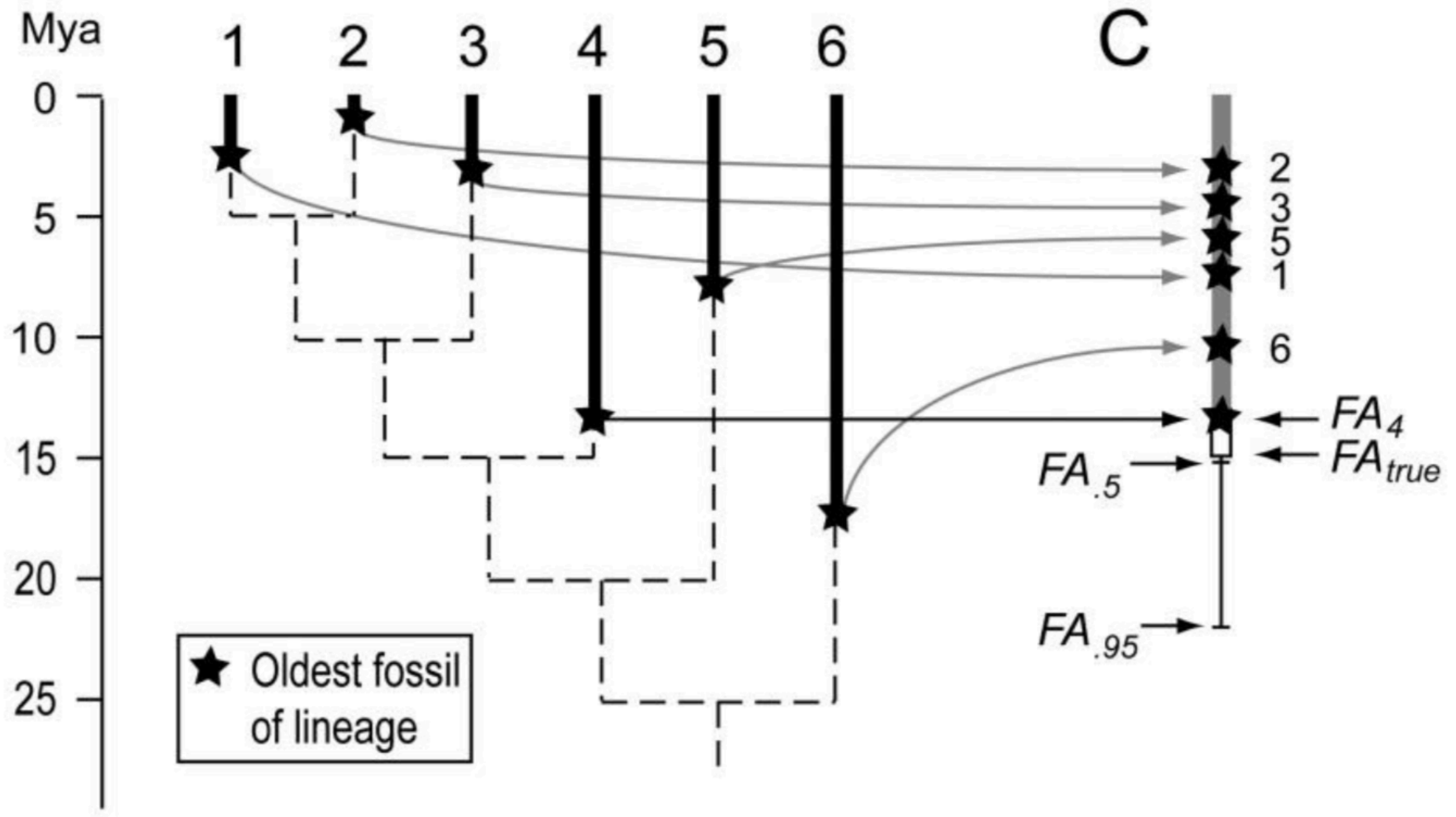


FIGURE 1. Ninety-five percent confidence intervals on the bases of hypothetical stratigraphic ranges. The number of fossiliferous horizons (H) is indicated above the appropriate stratigraphic column. The confidence intervals were calculated using Eq. (1).

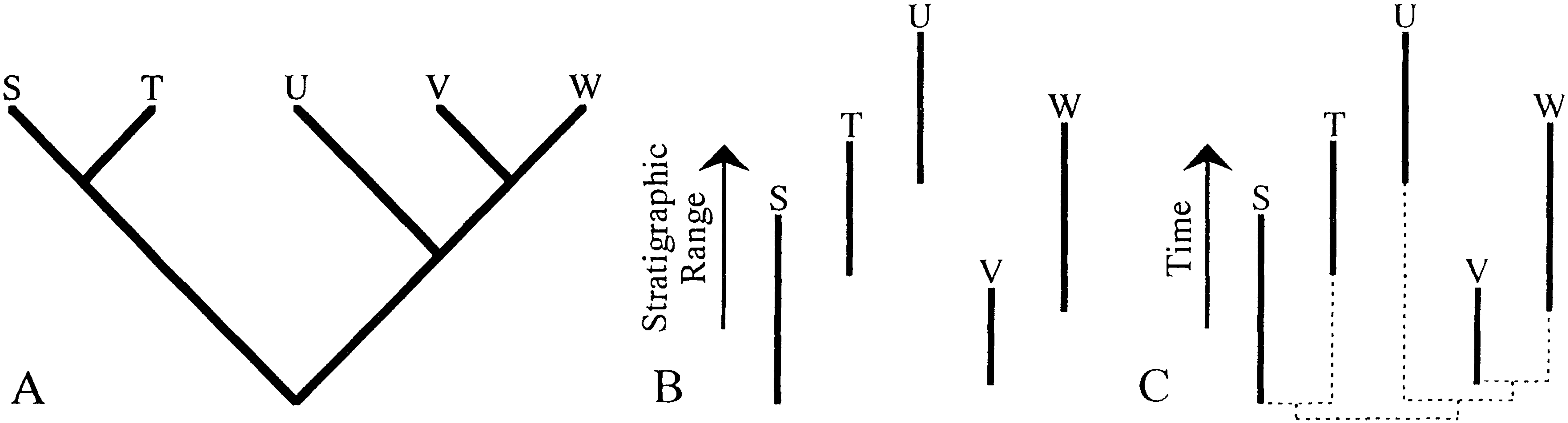
$$\alpha = (1 - C_1)^{-1/(H-1)} - 1, \qquad (1)$$

Marshall (1990) *Paleobiology*

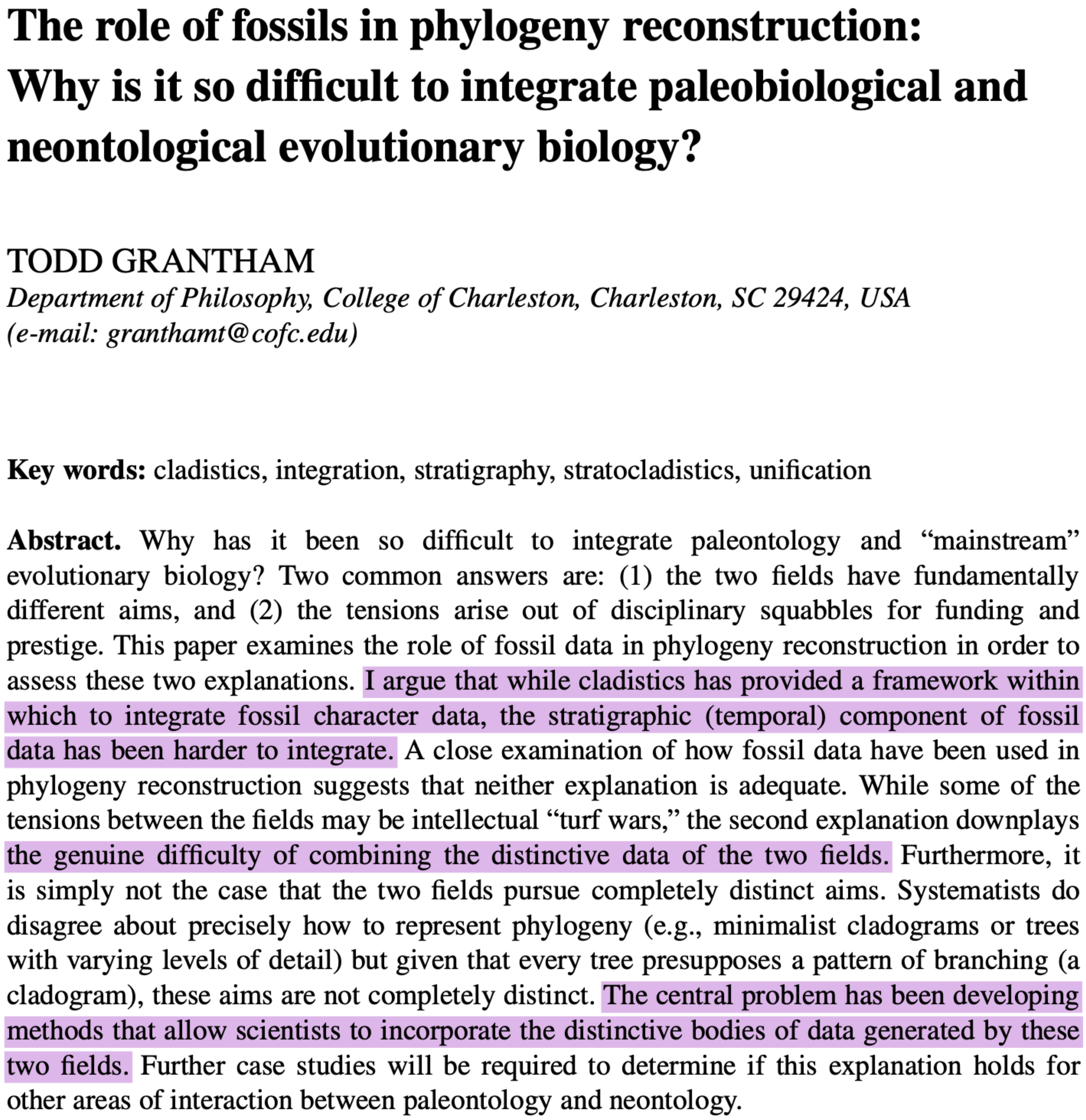


Marshall (2008) *American Naturalist*

Phylogeny, time and the stratigraphic record



Stratigraphic tests of cladistic hypotheses
Wagner. 1995. Paleobiology



The role of fossils in phylogeny reconstruction: Why is it so difficult to integrate paleobiological and neontological evolutionary biology?

TODD GRANTHAM
Department of Philosophy, College of Charleston, Charleston, SC 29424, USA
(e-mail: granthamt@cofc.edu)

Key words: cladistics, integration, stratigraphy, stratocladistics, unification

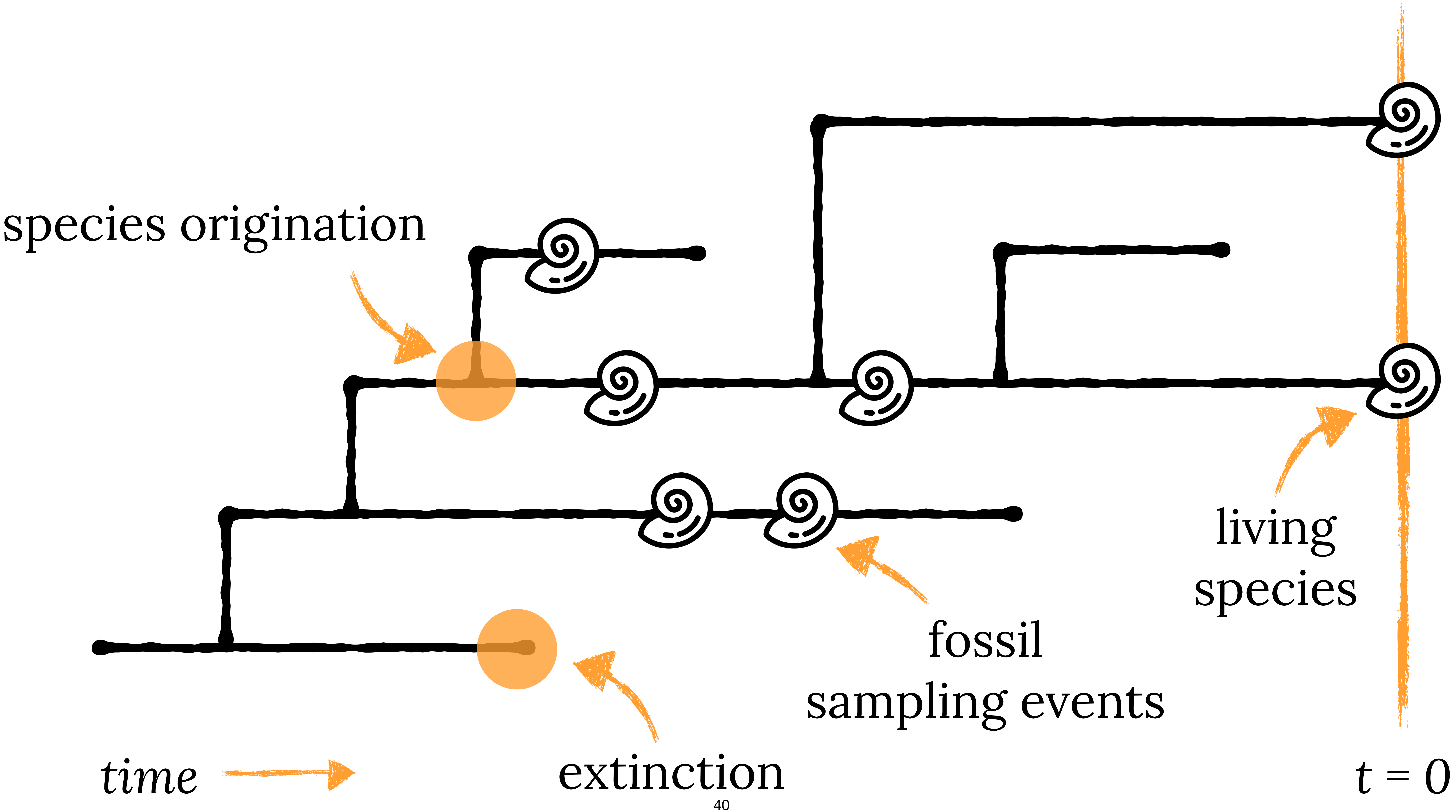
Abstract. Why has it been so difficult to integrate paleontology and "mainstream" evolutionary biology? Two common answers are: (1) the two fields have fundamentally different aims, and (2) the tensions arise out of disciplinary squabbles for funding and prestige. This paper examines the role of fossil data in phylogeny reconstruction in order to assess these two explanations. I argue that while cladistics has provided a framework within which to integrate fossil character data, the stratigraphic (temporal) component of fossil data has been harder to integrate. A close examination of how fossil data have been used in phylogeny reconstruction suggests that neither explanation is adequate. While some of the tensions between the fields may be intellectual "turf wars," the second explanation downplays the genuine difficulty of combining the distinctive data of the two fields. Furthermore, it is simply not the case that the two fields pursue completely distinct aims. Systematists do disagree about precisely how to represent phylogeny (e.g., minimalist cladograms or trees with varying levels of detail) but given that every tree presupposes a pattern of branching (a cladogram), these aims are not completely distinct. The central problem has been developing methods that allow scientists to incorporate the distinctive bodies of data generated by these two fields. Further case studies will be required to determine if this explanation holds for other areas of interaction between paleontology and neontology.

Until very recently, we lacked statistically coherent models that unified phylogenetic and temporal observations.

Grantham. 2004. Biology and Philosophy

So what would a generating model for fossil data look like?

→ Ideally, we want to use a tree model that describes the probability of observing the sampled tree given the speciation (birth), extinction (death) and sampling processes.



species origination

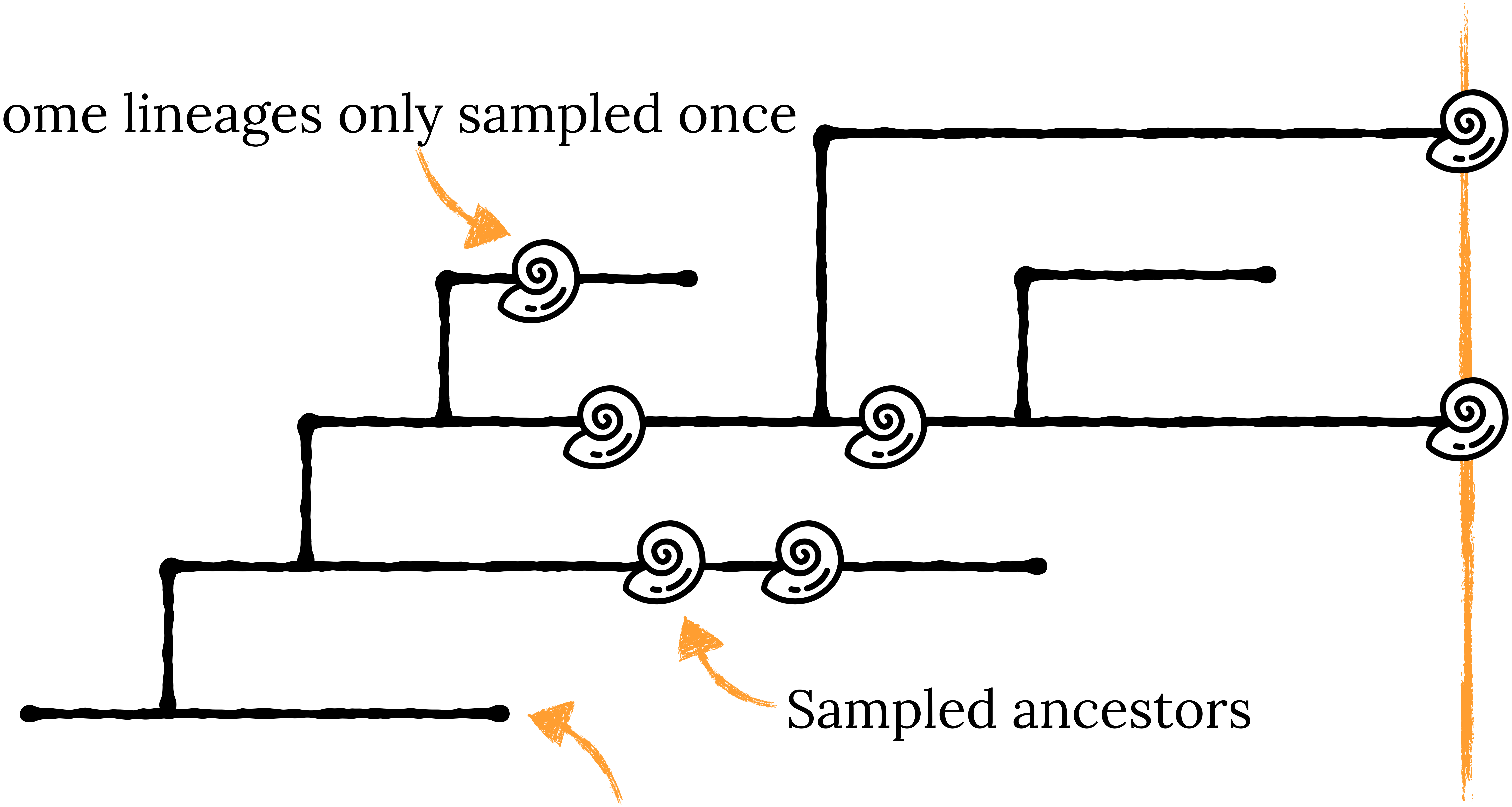
living species

fossil sampling events

extinction

time →

$t = 0$



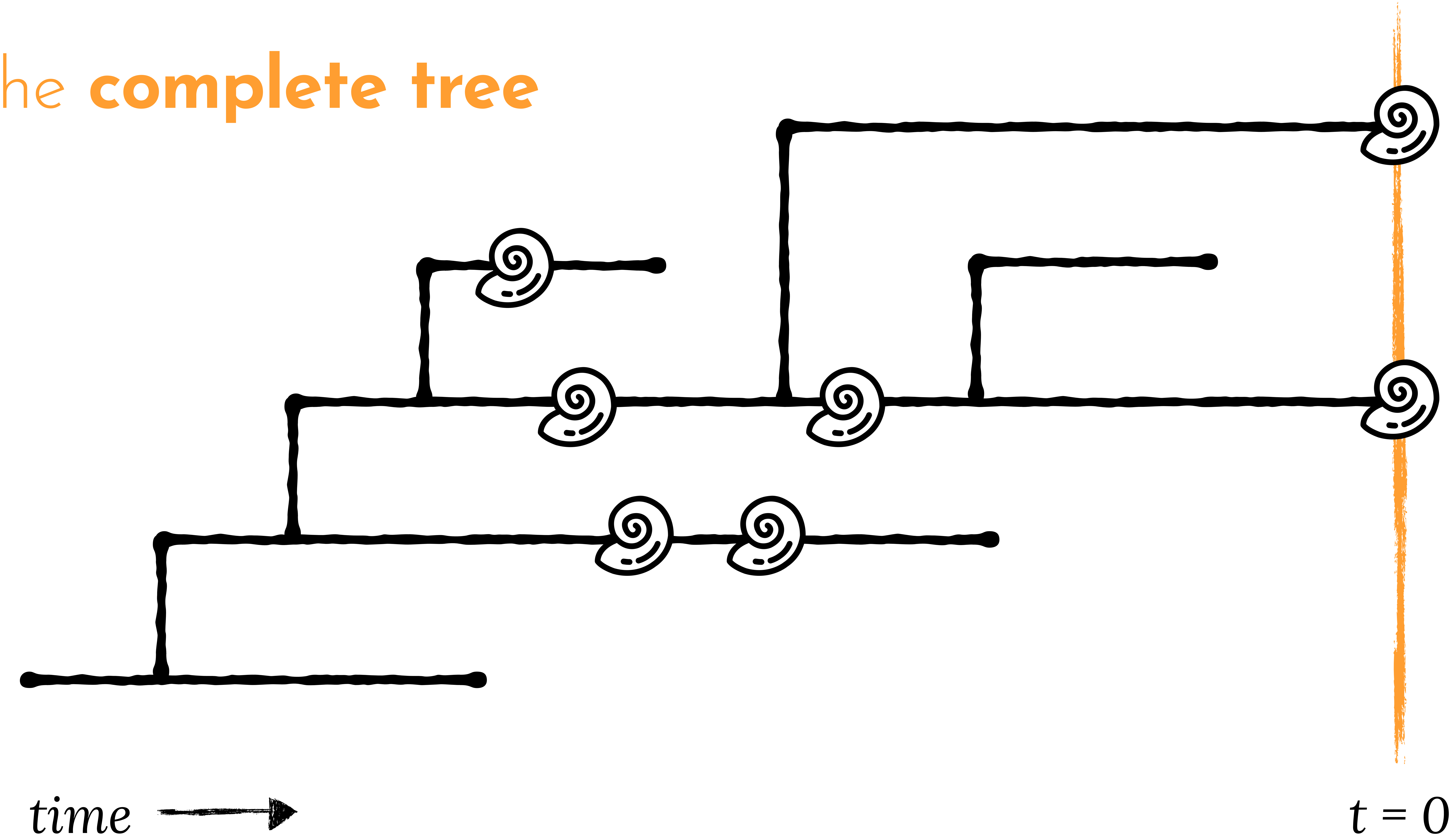
Some lineages only sampled once

Sampled ancestors

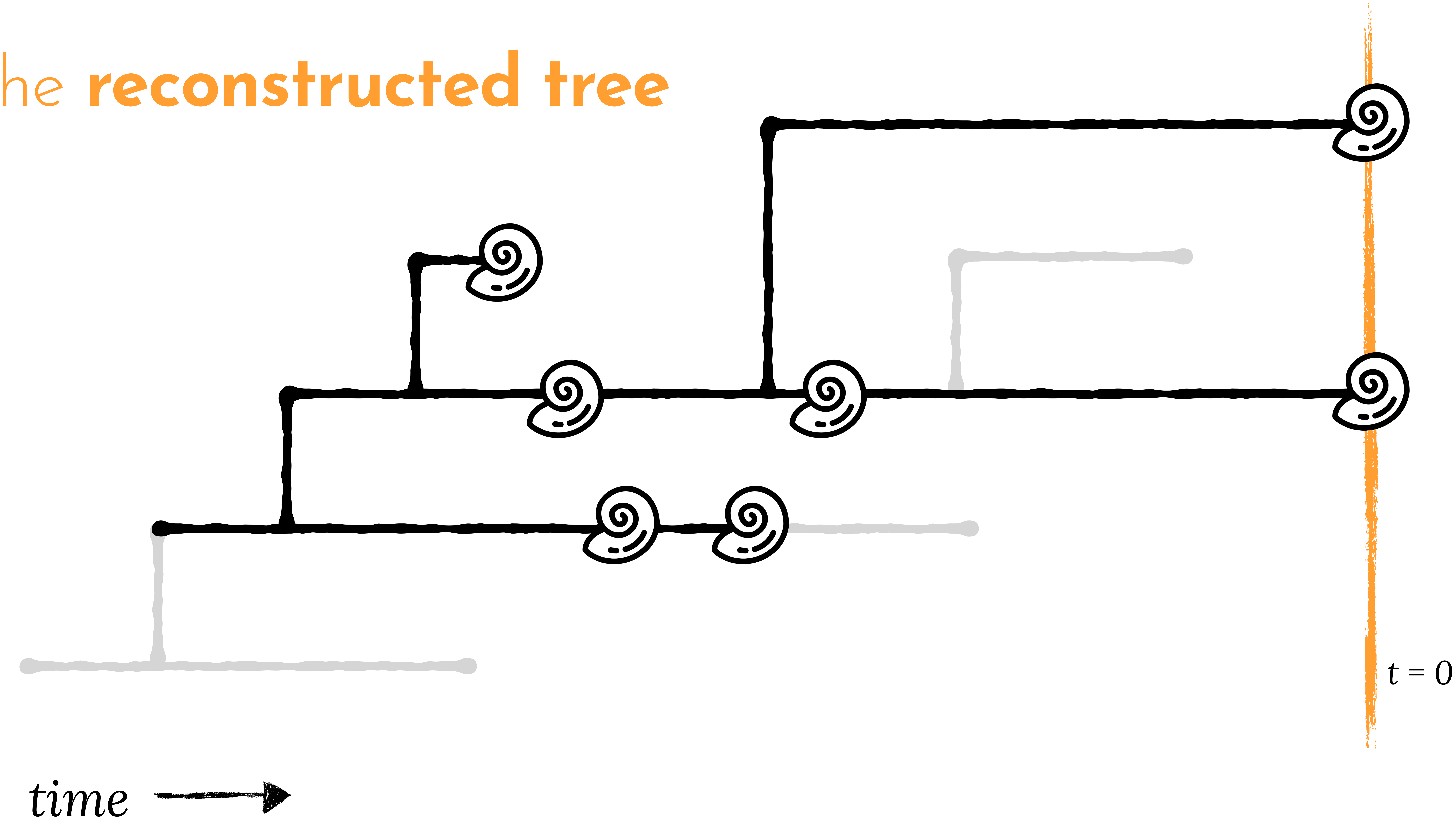
Some lineages go completely unsampled

$t = 0$

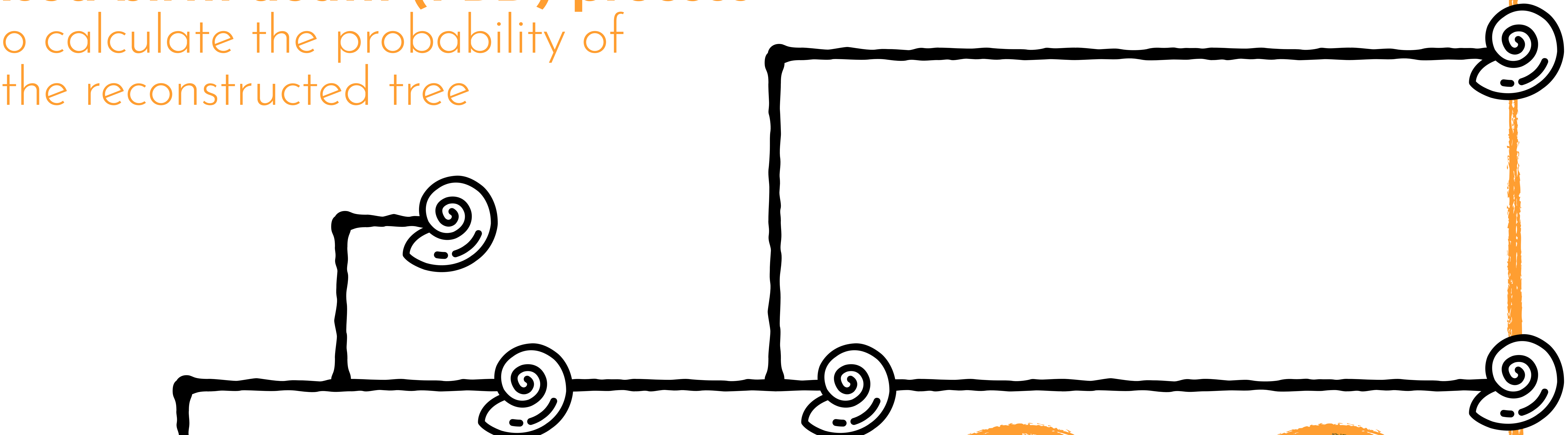
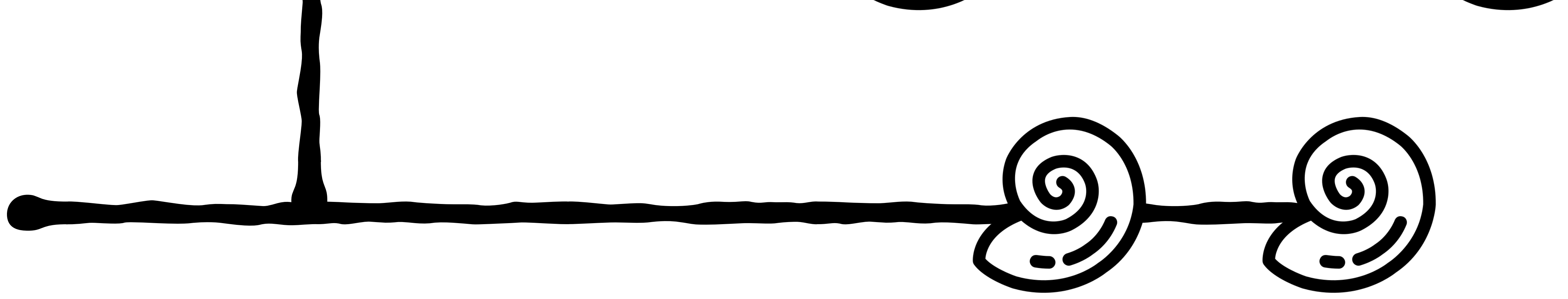
The **complete tree**

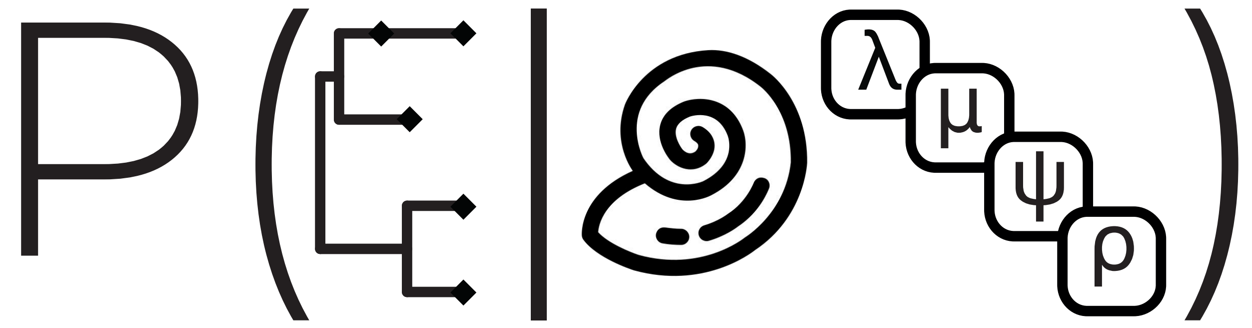


The **reconstructed tree**



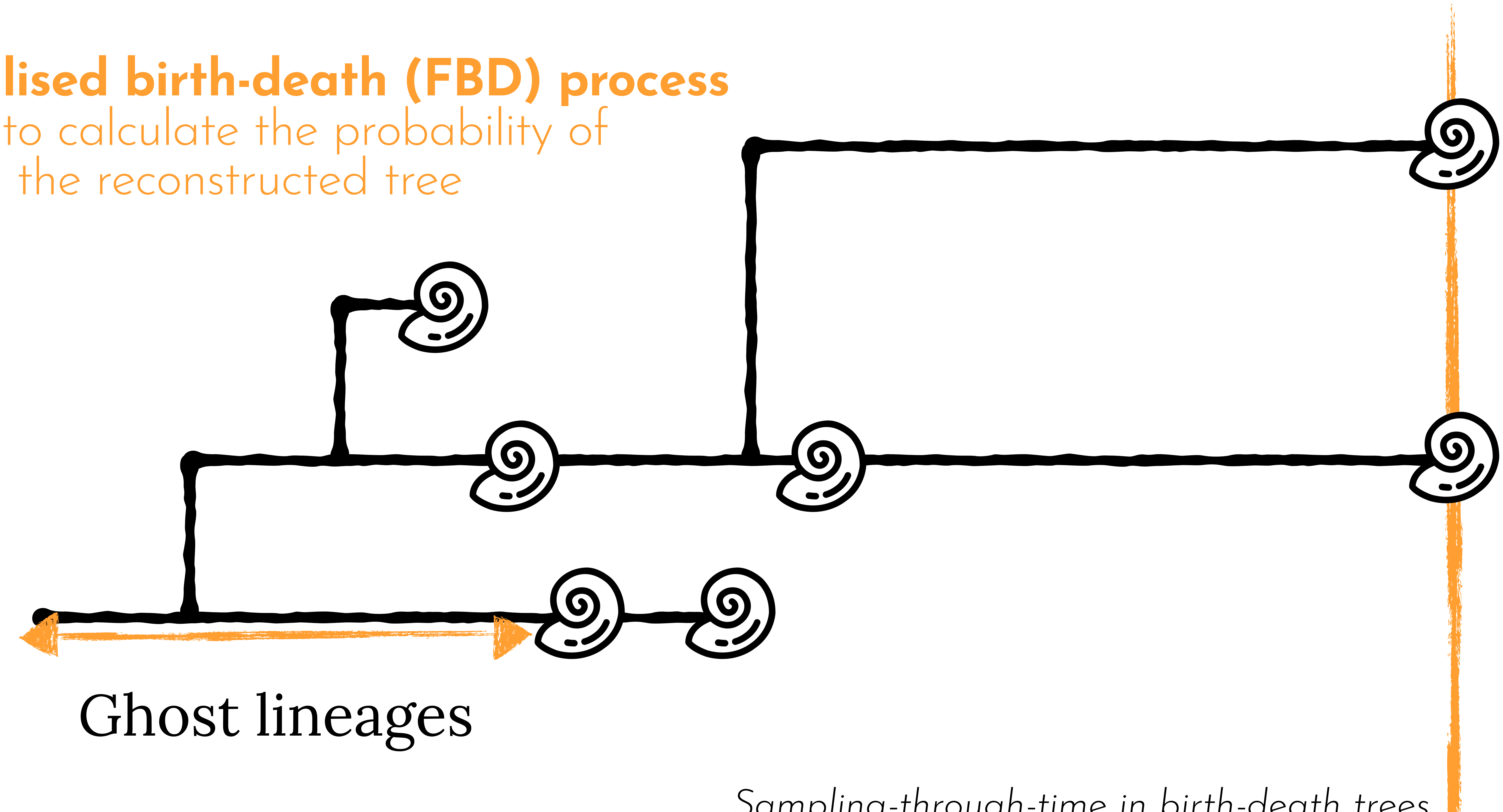
The **fossilised birth-death (FBD) process** allows us to calculate the probability of observing the reconstructed tree

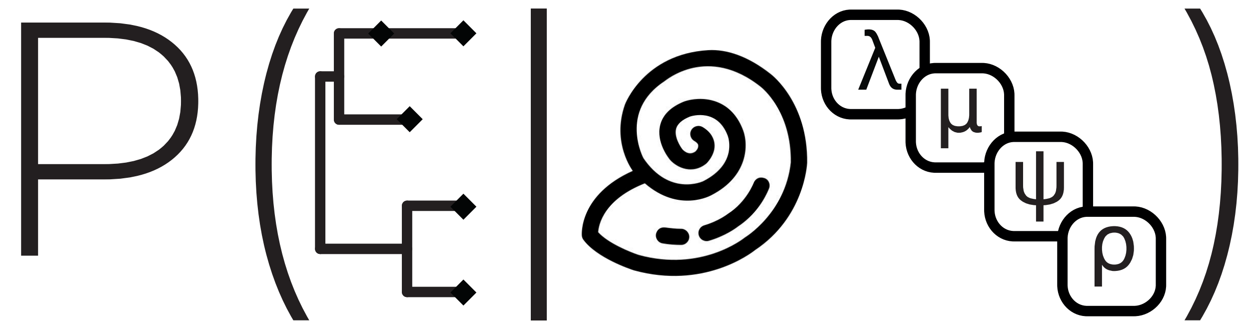





Sampling-through-time in birth-death trees
Stadler. 2010. JTB
First implemented: Heath et al. 2014. PNAS
Gavryushkina et al. 2014. PLoS Comp Bio

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. JTB
First implemented: Heath et al. 2014. PNAS
Gavryushkina et al. 2014. PLoS Comp Bio

Why is the FBD model so important?

→ There are many statistical advantages to having a generative model for time tree inference:

- greater accuracy
- better reflection of uncertainty
- increased flexibility

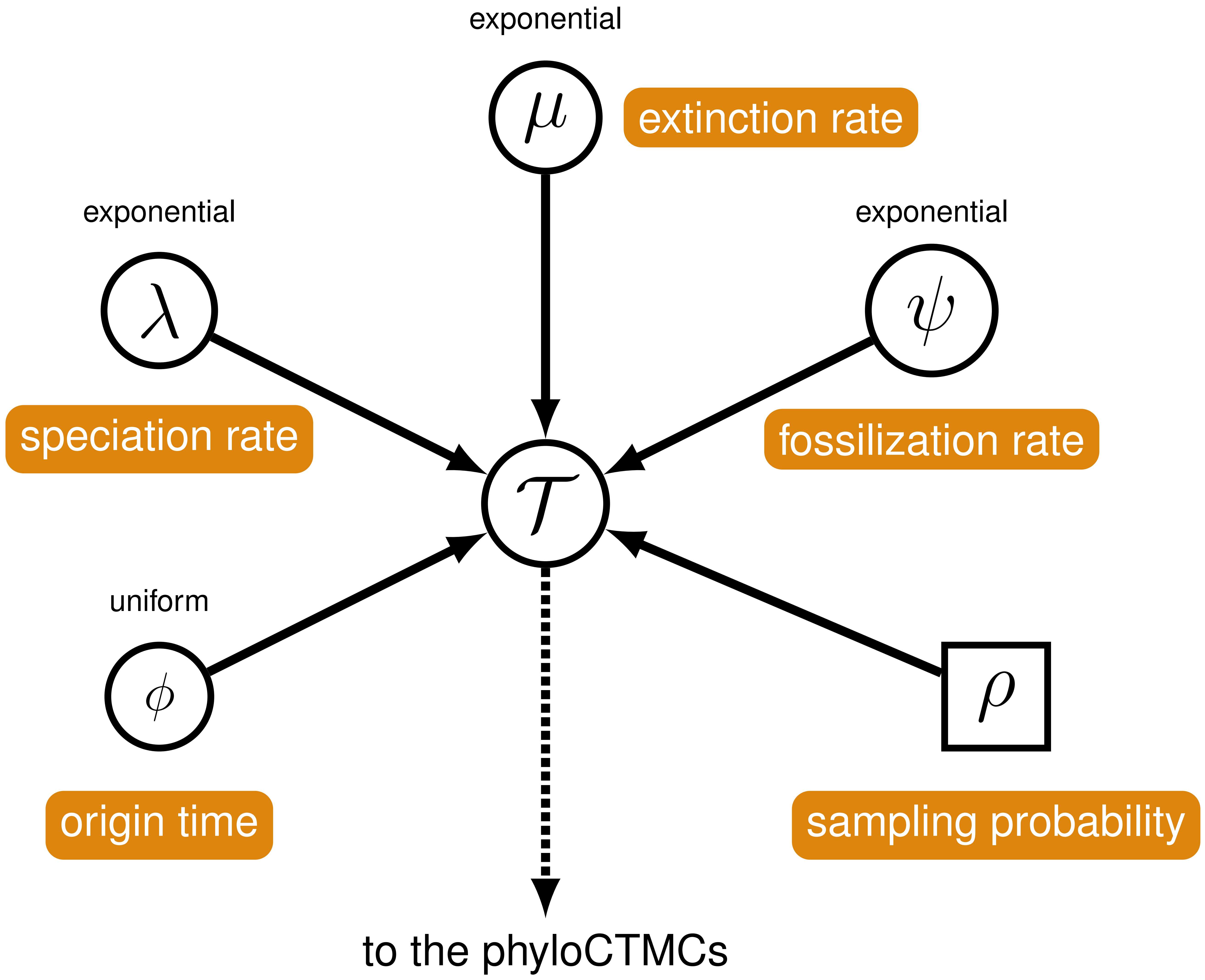
→ We can include fossils directly in the tree + much more fossil data.

→ We can combine data in ways that weren't previously possible and even link the model parameters to abiotic processes.

Grantham. 2004. Biology and Philosophy

Exercise 5: Simulating under the FBD process

Graphical representation of the FBD model



☐ **a**) Constant node

○ **b**) Stochastic node

Speciation, extinction and sampling are instantaneous rates – in this set up sampled from an exponential prior distribution, but could be constrained in many alternative ways.

Adapted from Barido-Sottani et al. 2020

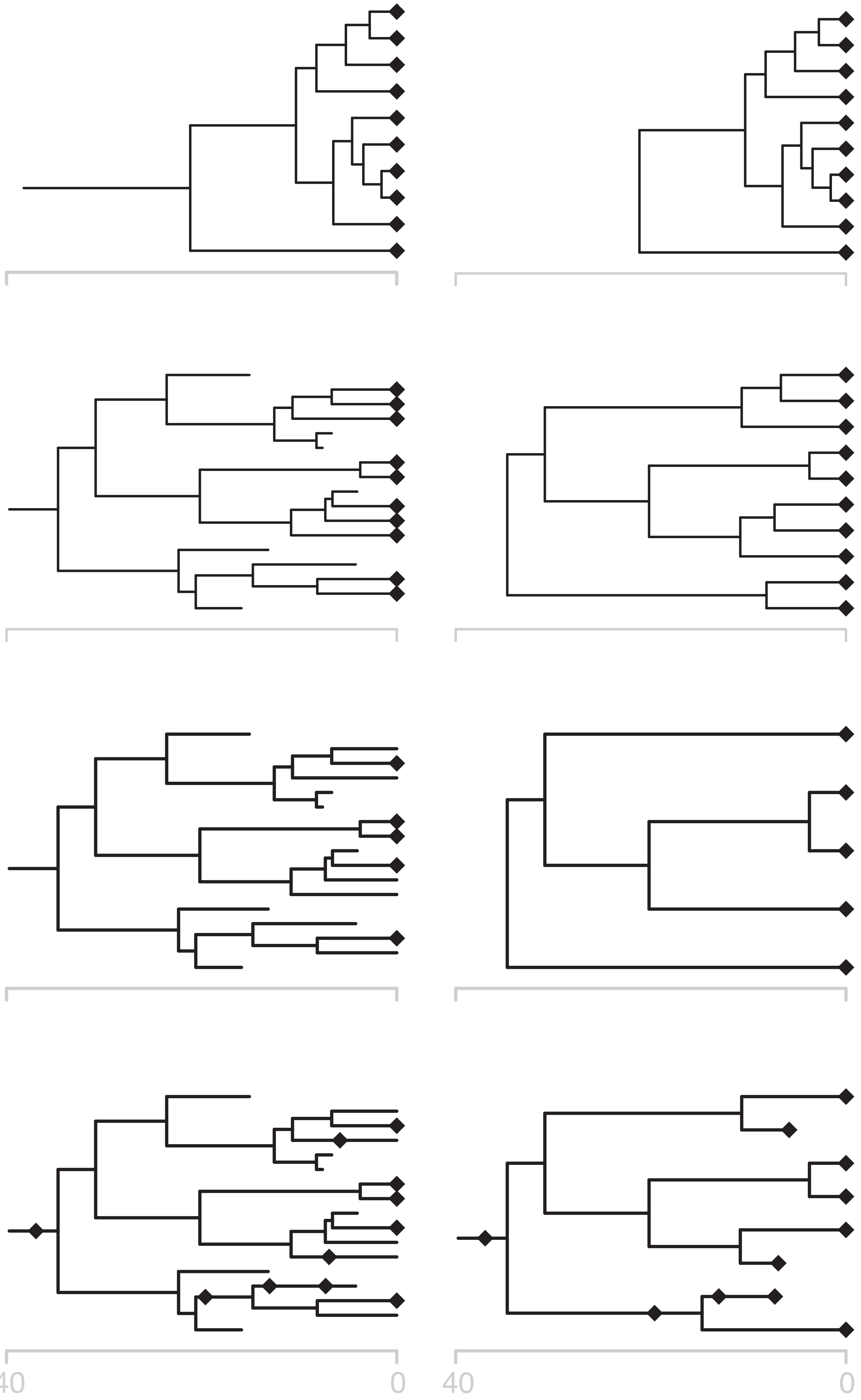
Relationship to (some) other birth-death process models

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

We can also use ρ at $t > 0$ to model serial sampling.

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

complete vs. reconstructed trees



$\lambda = 0.1$

$\lambda = 0.1$
$\mu = 0.05$

$\lambda = 0.1$
$\mu = 0.05$
$\rho = 0.6$

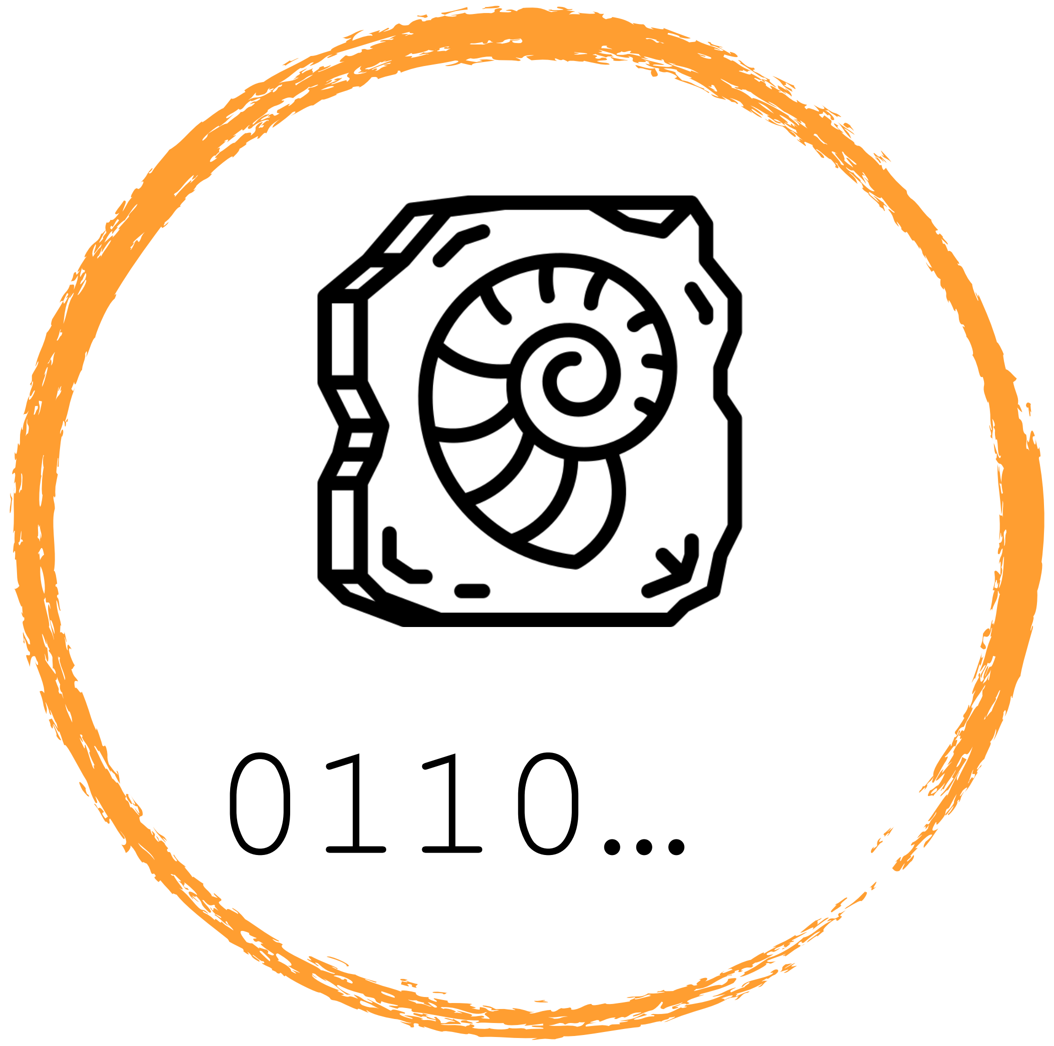
Yang and Rannala 1997
Stadler 2009

$\lambda = 0.1$
$\mu = 0.05$
$\rho = 0.6$
$\psi = 0.05$

Stadler 2010

For living species, **DNA or protein sequences** are the primary data used for phylogenetics.

For fossils, **morphology** and **sampling times** are the primary data.



Calculating the phylogenetic likelihood

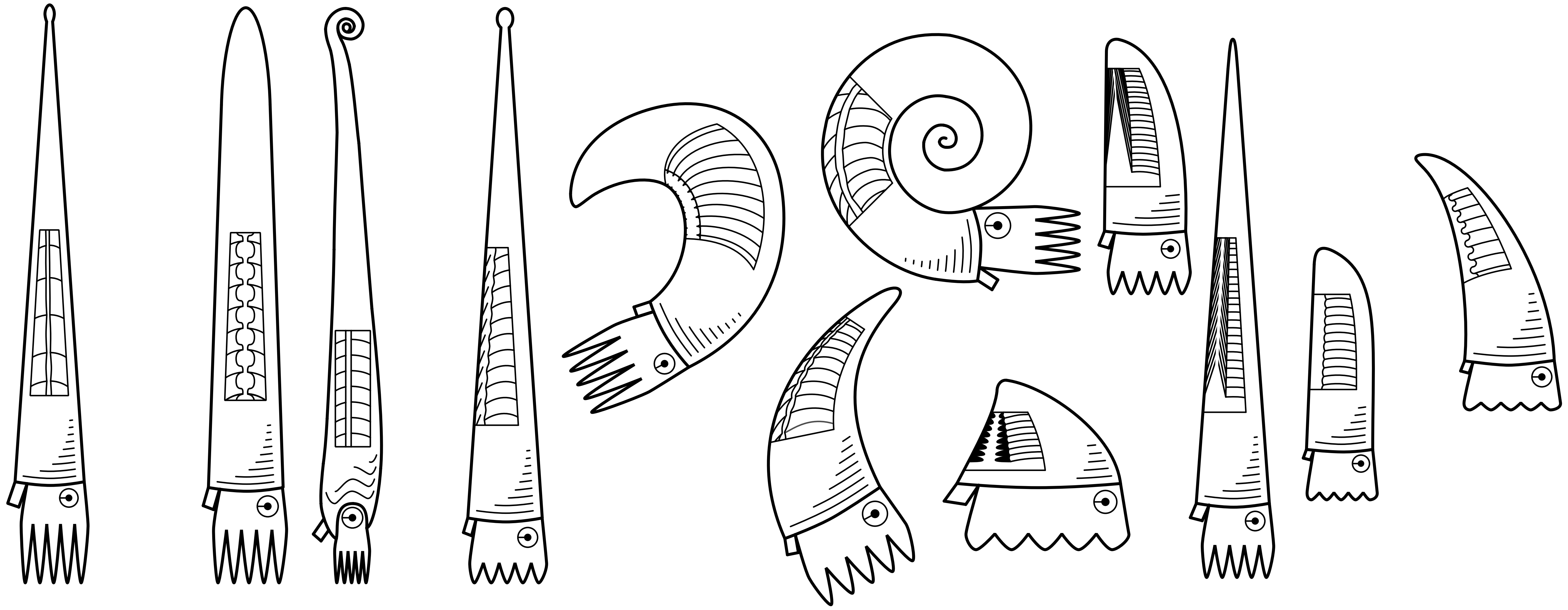
Molecular data → many options available (see previous lectures)

Discrete morphological character data → Mk model (generalisation of the JC model for *k* states)

Continuous trait data → brownian motion, other models that come from phylogenetic comparative methods

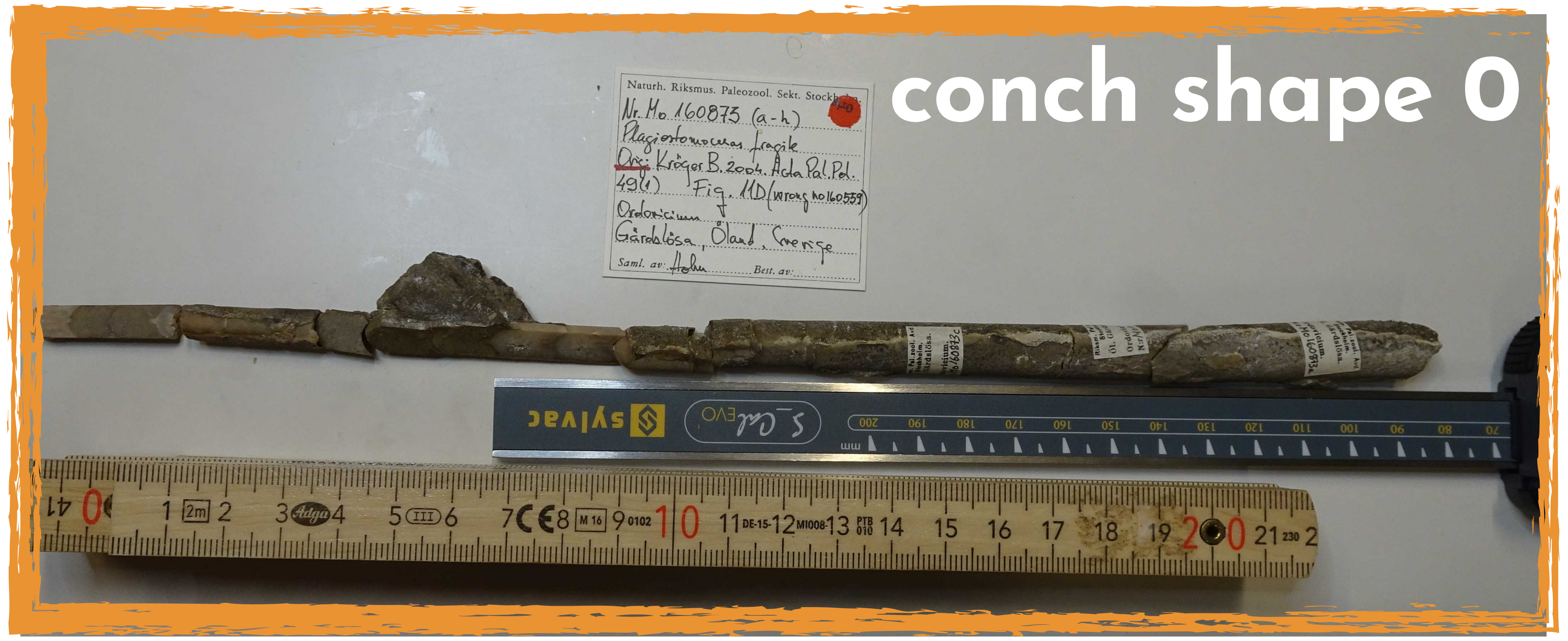
Great review on models for morphology: Wright. 2019
Great online book on PCMs: Harmon. 2019

Constructing a morphological matrix



0 or 1

1 = presence of prominent ribs (annulations)


conch shape 0


conch shape 1

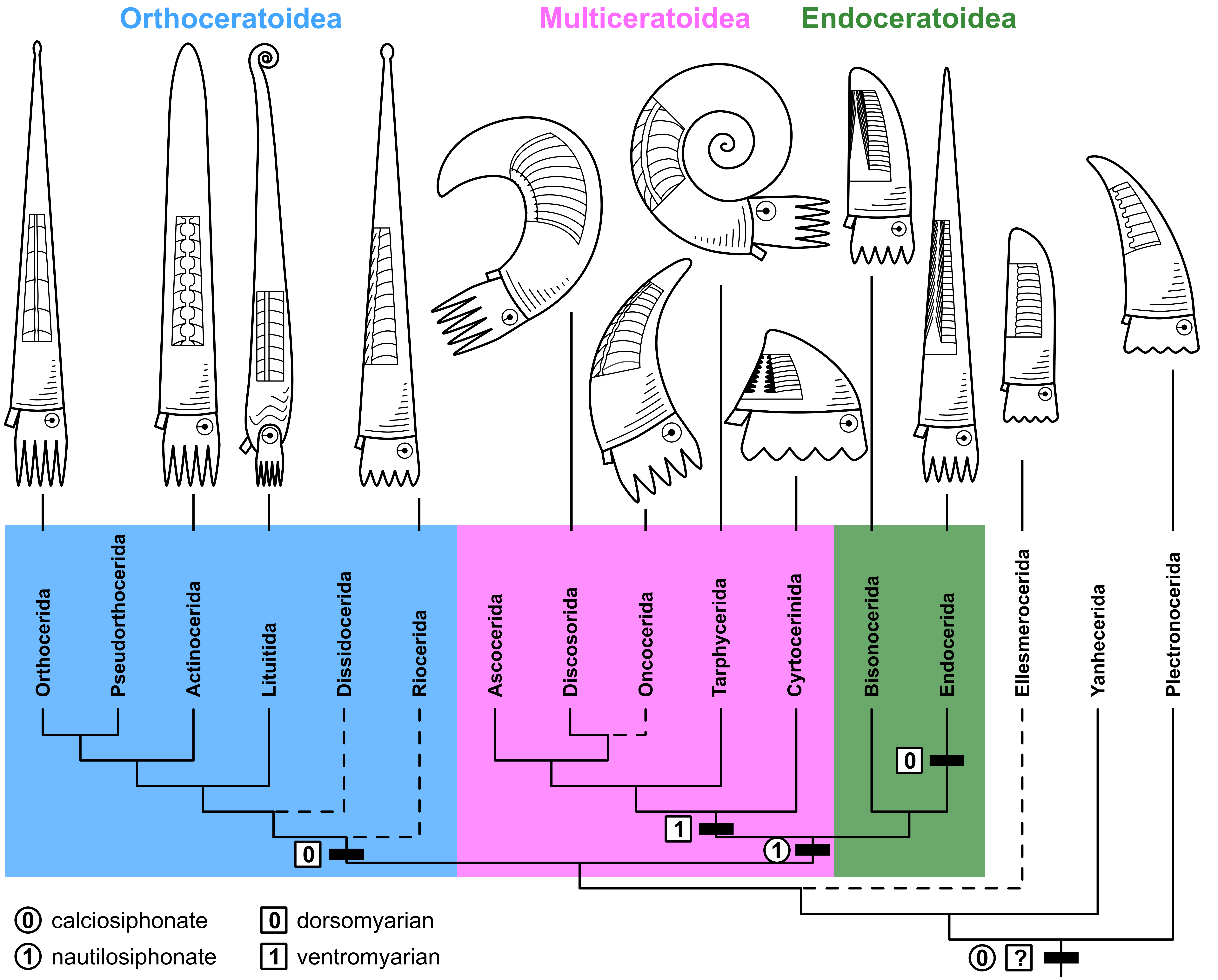

conch shape 2


conch shape 3

Images: Pics: Alex Pohle

	characters
species 1	001510010?00-100--0000000000
species 2	000400010?200100--0010010000
species 3	002500010?200100--0?10010000
species 4	00?5?0010?200100?-0???010110

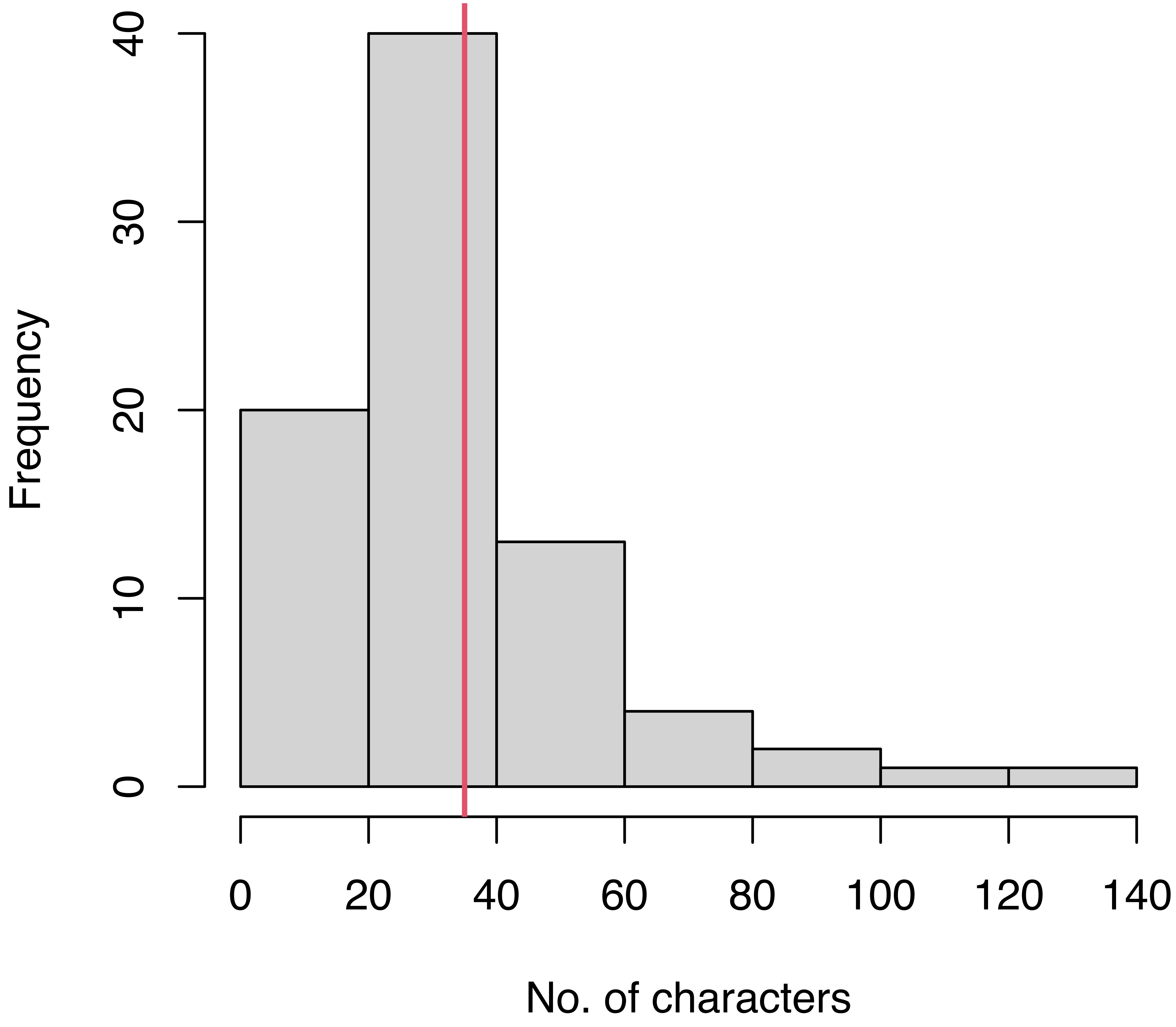
Integers are used to represent different morphological states.



Tree based on **141** morphological characters

Early cephalopod evolution clarified through Bayesian phylogenetic inference
Pohle et al. 2022. BMC Biology

The average matrix for Palaeozoic (541 – 252 Ma) invertebrates has 35 characters



Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology in Time Calibrated Tree Inference
Barido-Sottani et al. 2020. Frontiers in Ecology & Evolution

Continuous trait measurement data

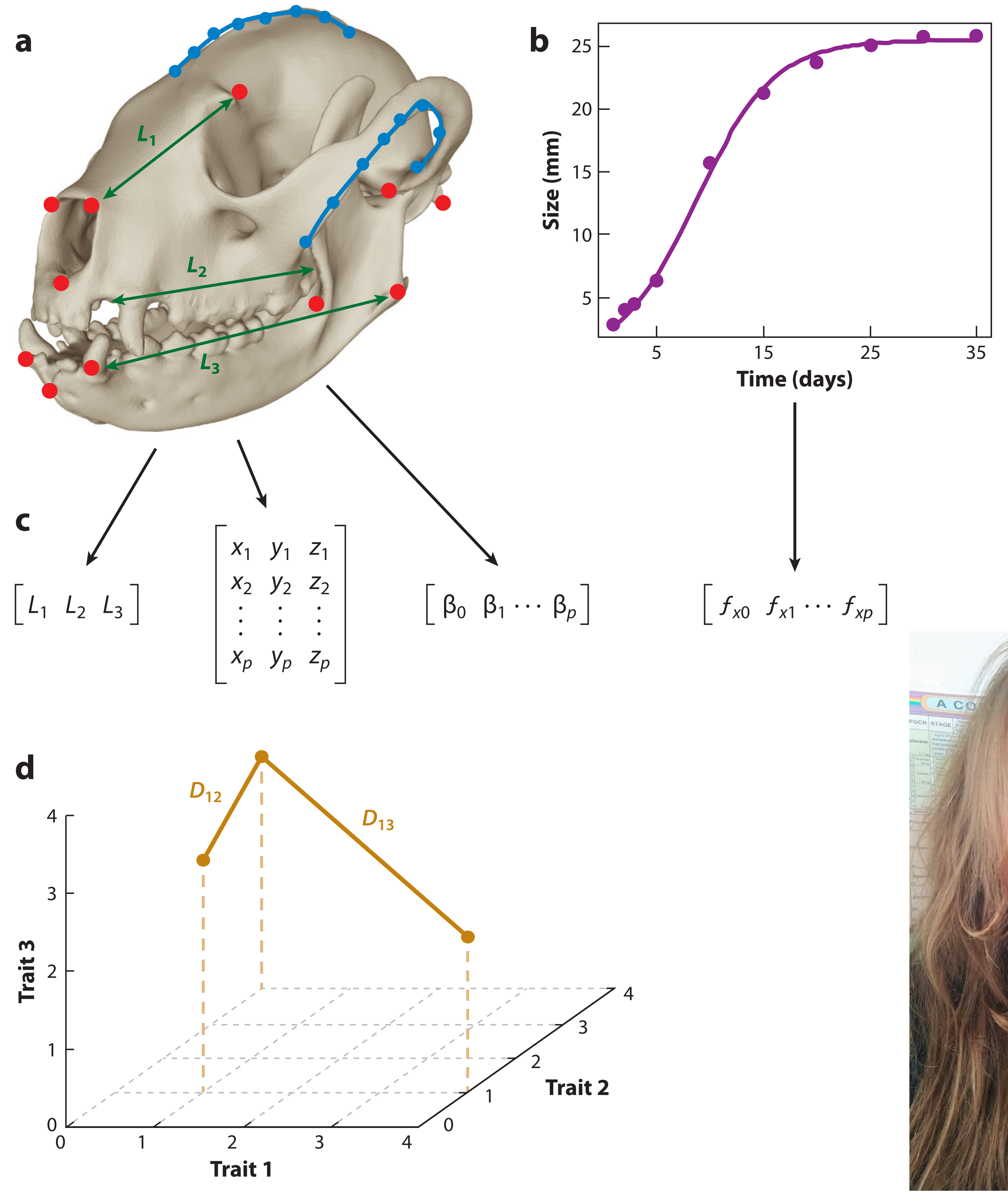
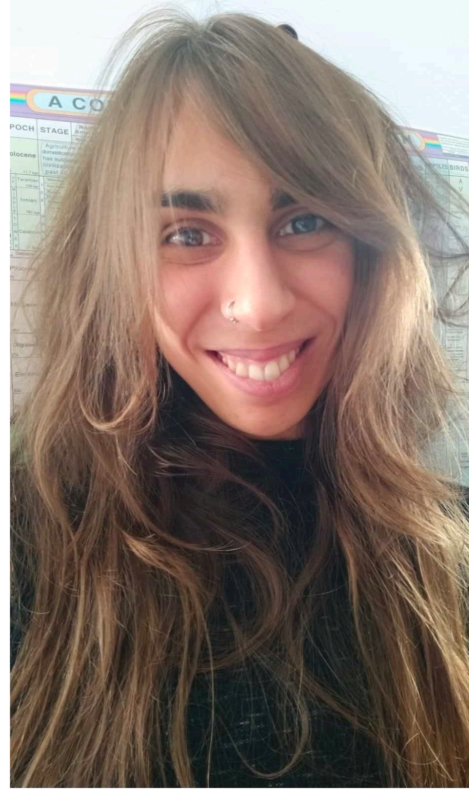
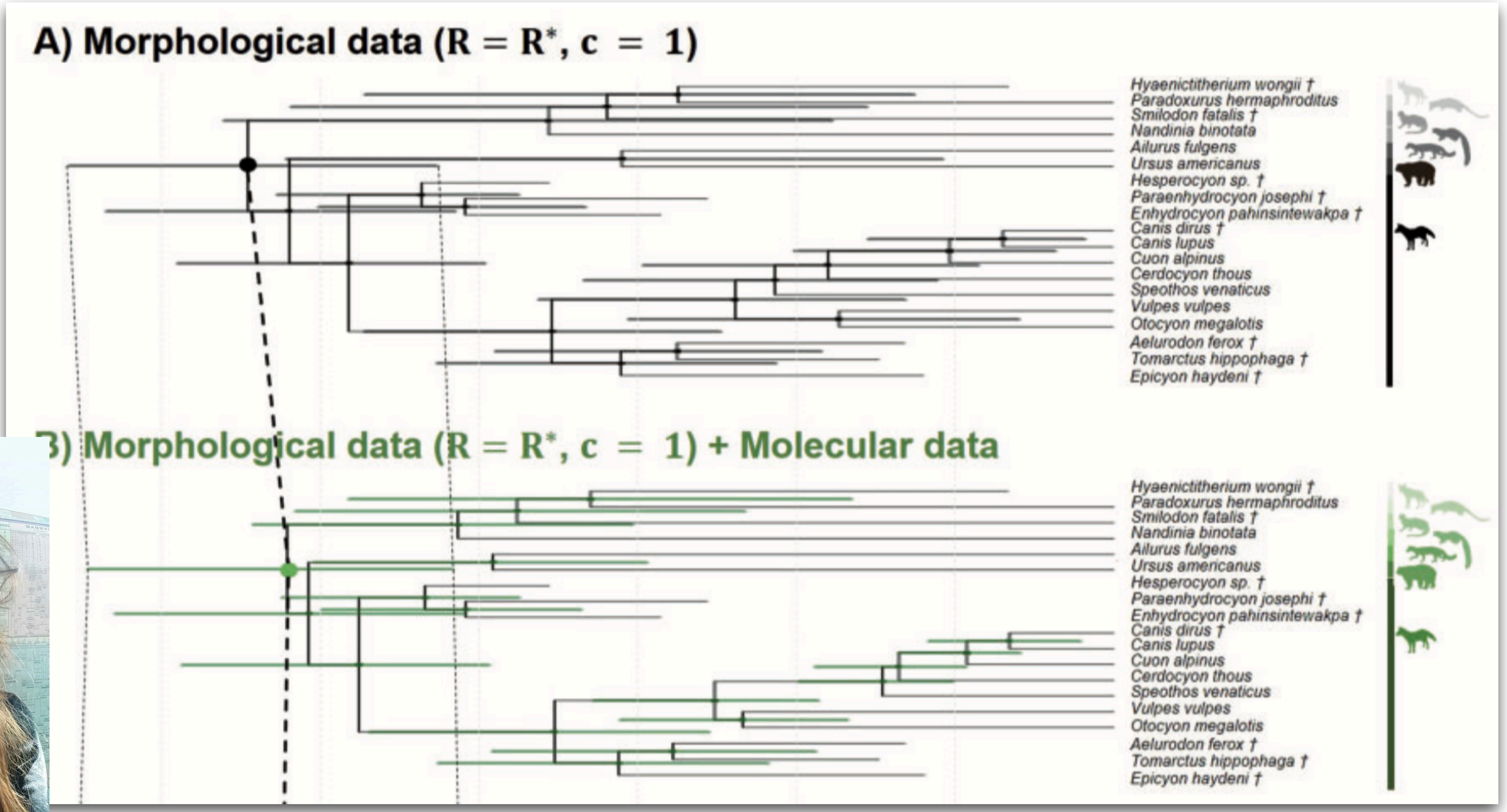


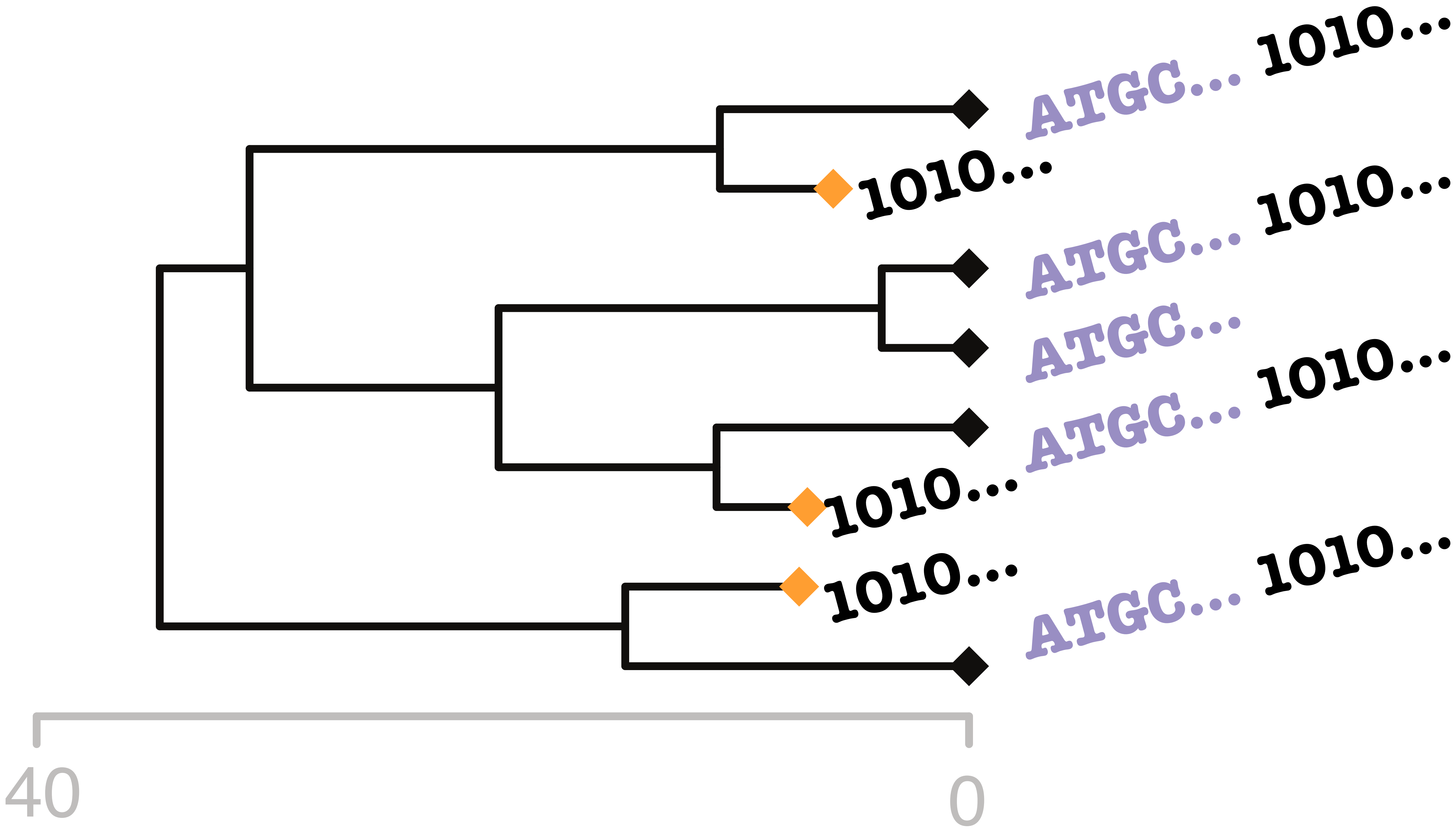
Image: Adams & Collyer. 2019.





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

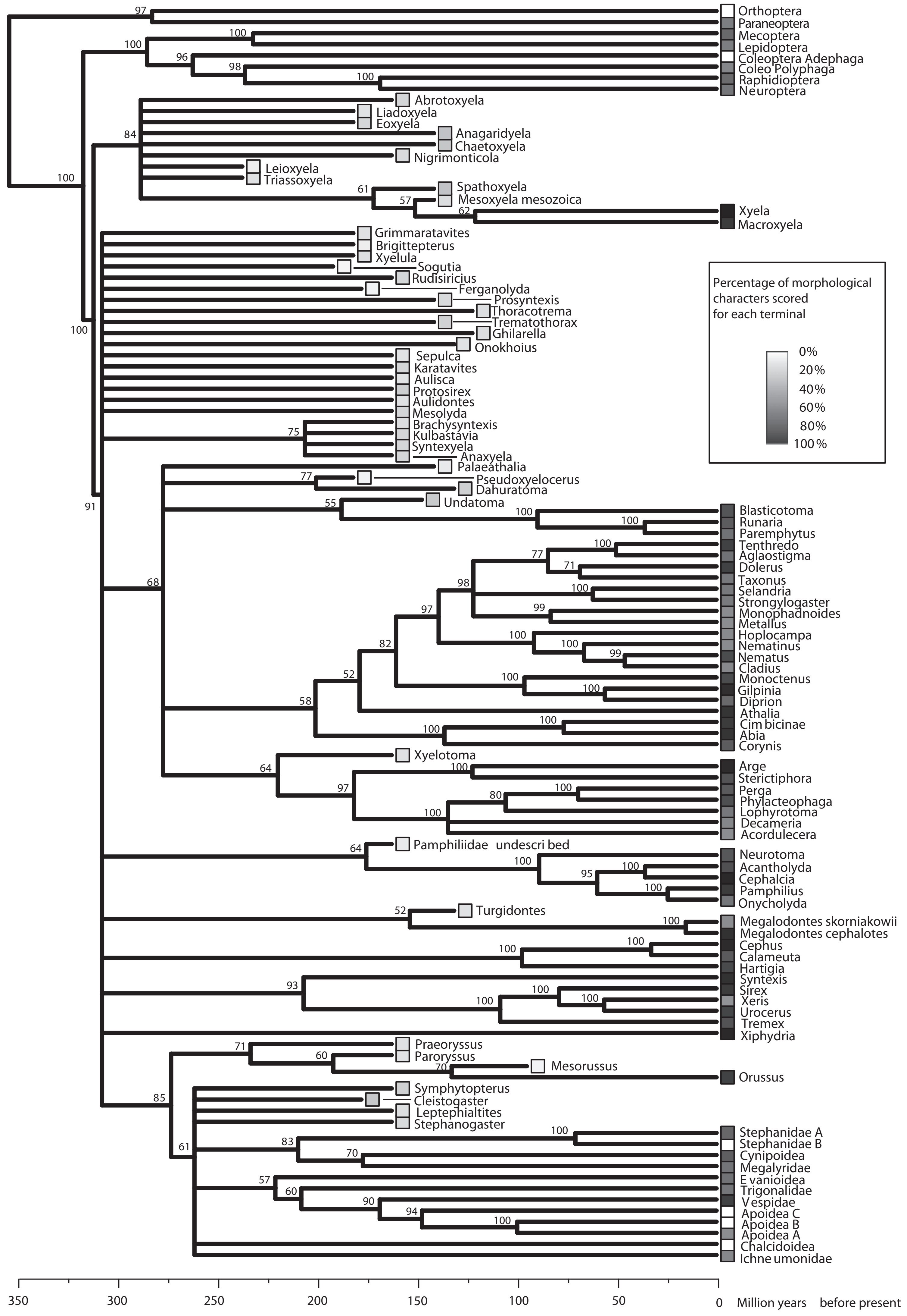
“Total-evidence” dating uses all available morphological & molecular data



Living species can have DNA & morphological data, while fossils are positioned on the basis of morphology only.

This approach has the advantage of accounting for uncertainty in fossil placement.

“Total-evidence” dating under the uniform tree model



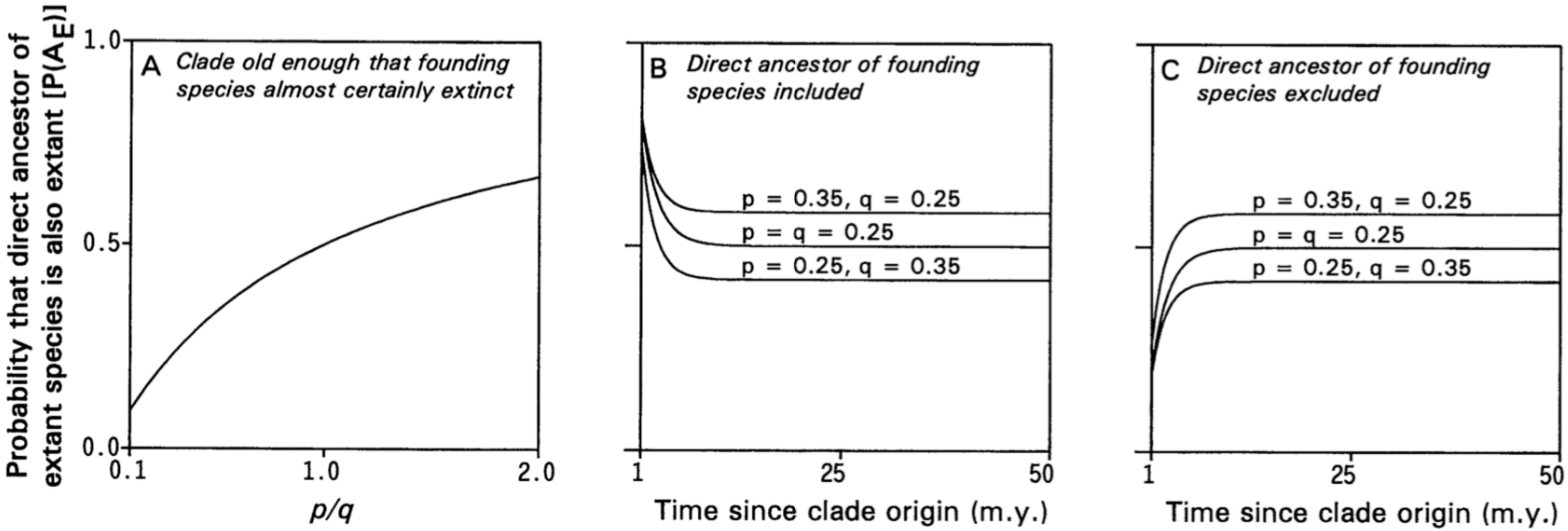
Dated tree of Hymenoptera

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.

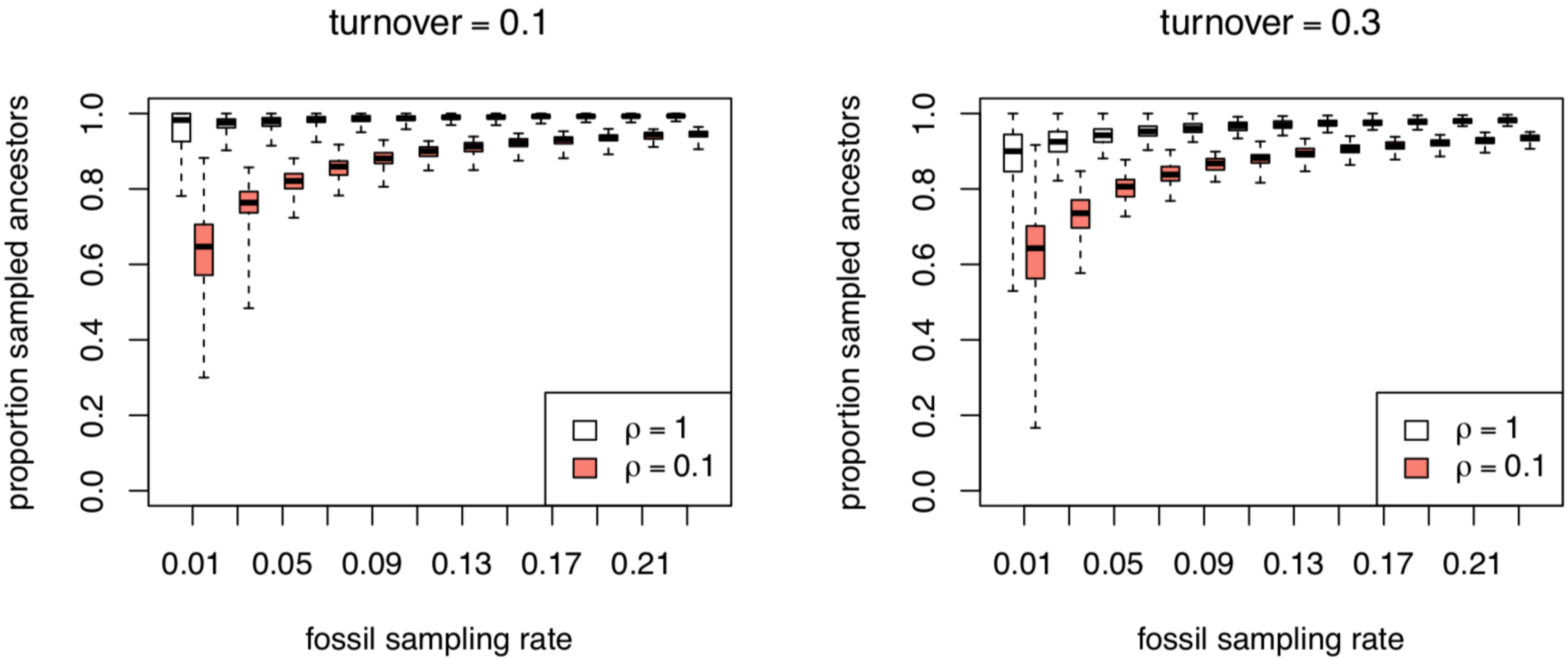
Ronquist et al. 2012 Sys Bio

Sampled ancestors



Sampled ancestors

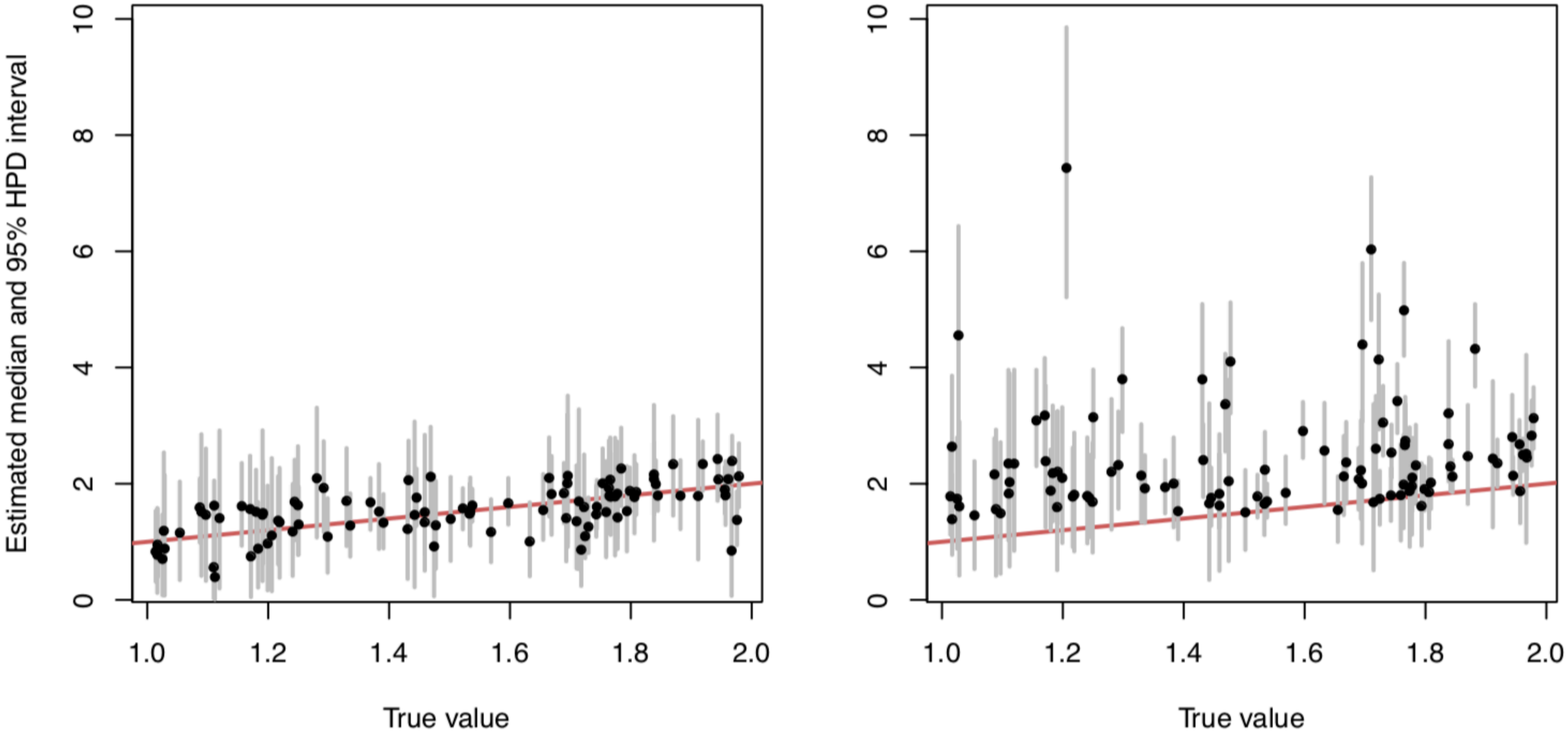
The proportion increases with higher turnover (birth - death) or higher sampling.



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

Under the FBD process fossils can be incorporated via **character data** (total-evidence) <u>OR</u> **topological constraints**

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

OR

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

- Bivalvia
  - Palaeotaxodonta
    - Nuculoida
      - Nuculanacea
        - Nuculanidae
            - Yolida
            - Zealeda
        - Isoarcidae
            - Isoarca
  - Cryptodonta
    - Solemyoida
      - Solemyacea
        - Solemyidae
            - Solemya
            - Acharax
    - Praecardioida
      - Praecardiacea
        - Praecardiidae
          - Praecardiinae
            - Buchiola
            - Adulomya
            - Eopteria
            - Necklania
            - Slava
          - Cardiolinae
            - Cardiola
            - Euthydesma
            - Opisthocoelus

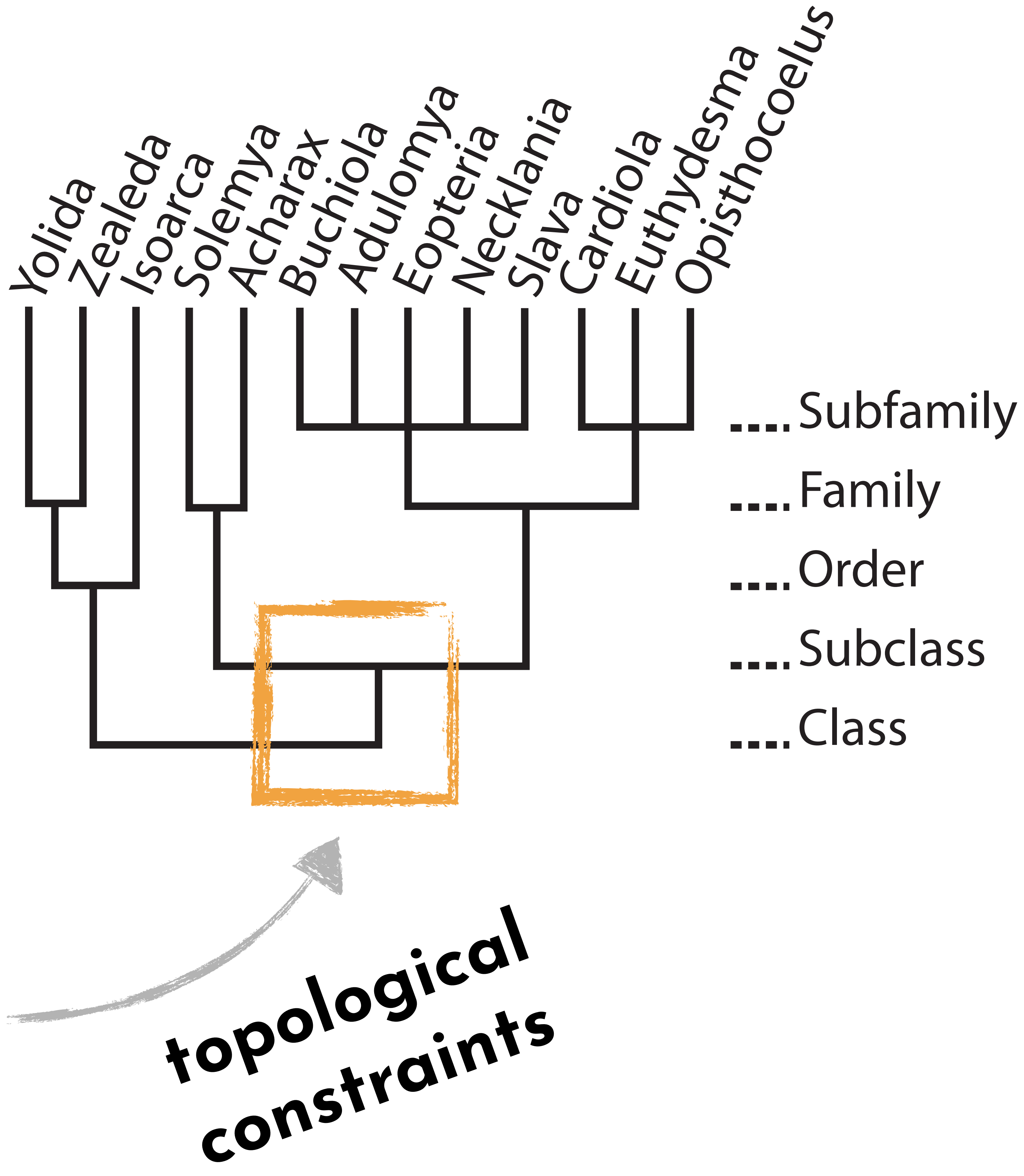
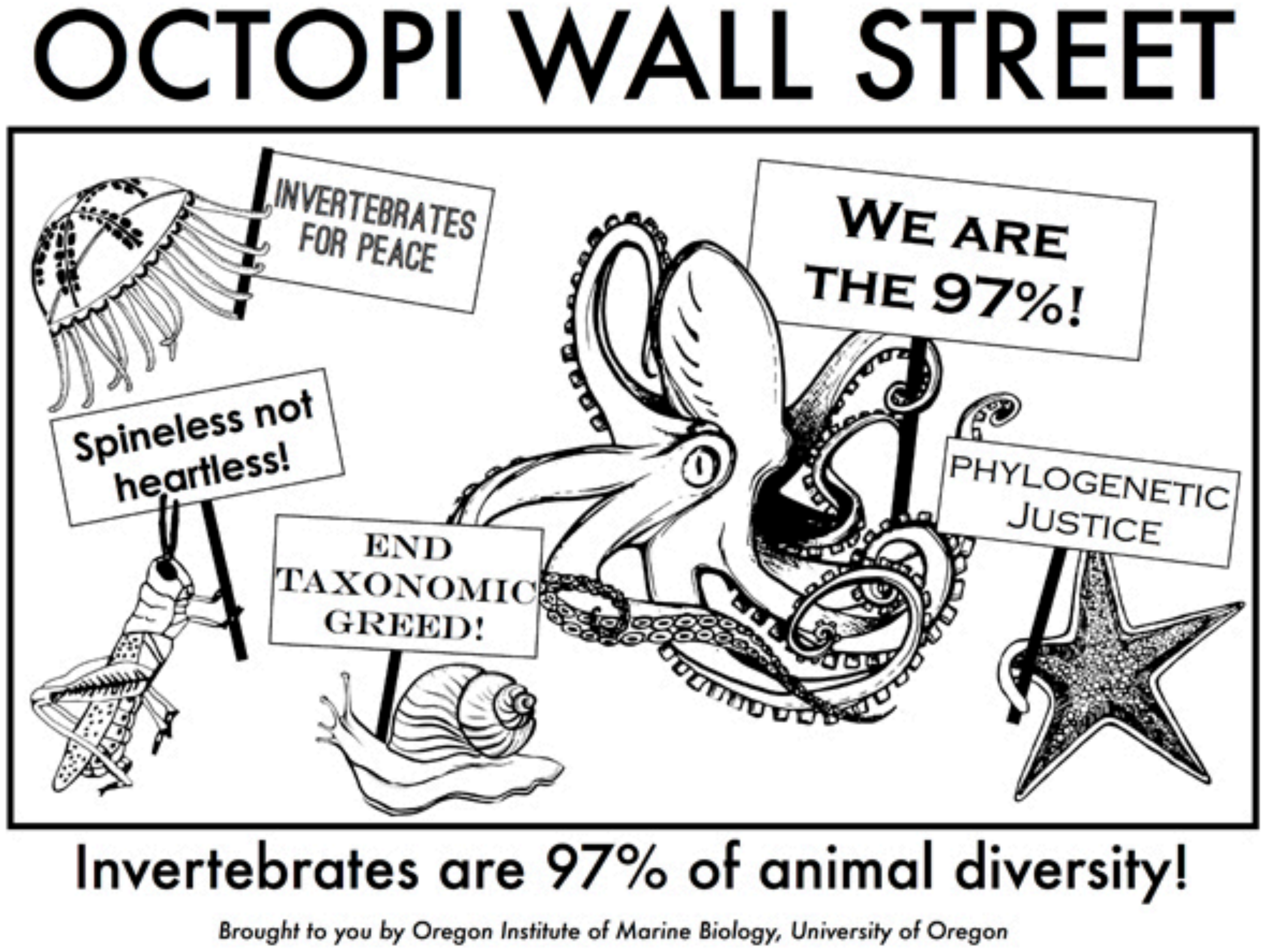


Image: Soul & Friedman 2015, Sys Bio

Phylogenetic analyses need taxonomists!

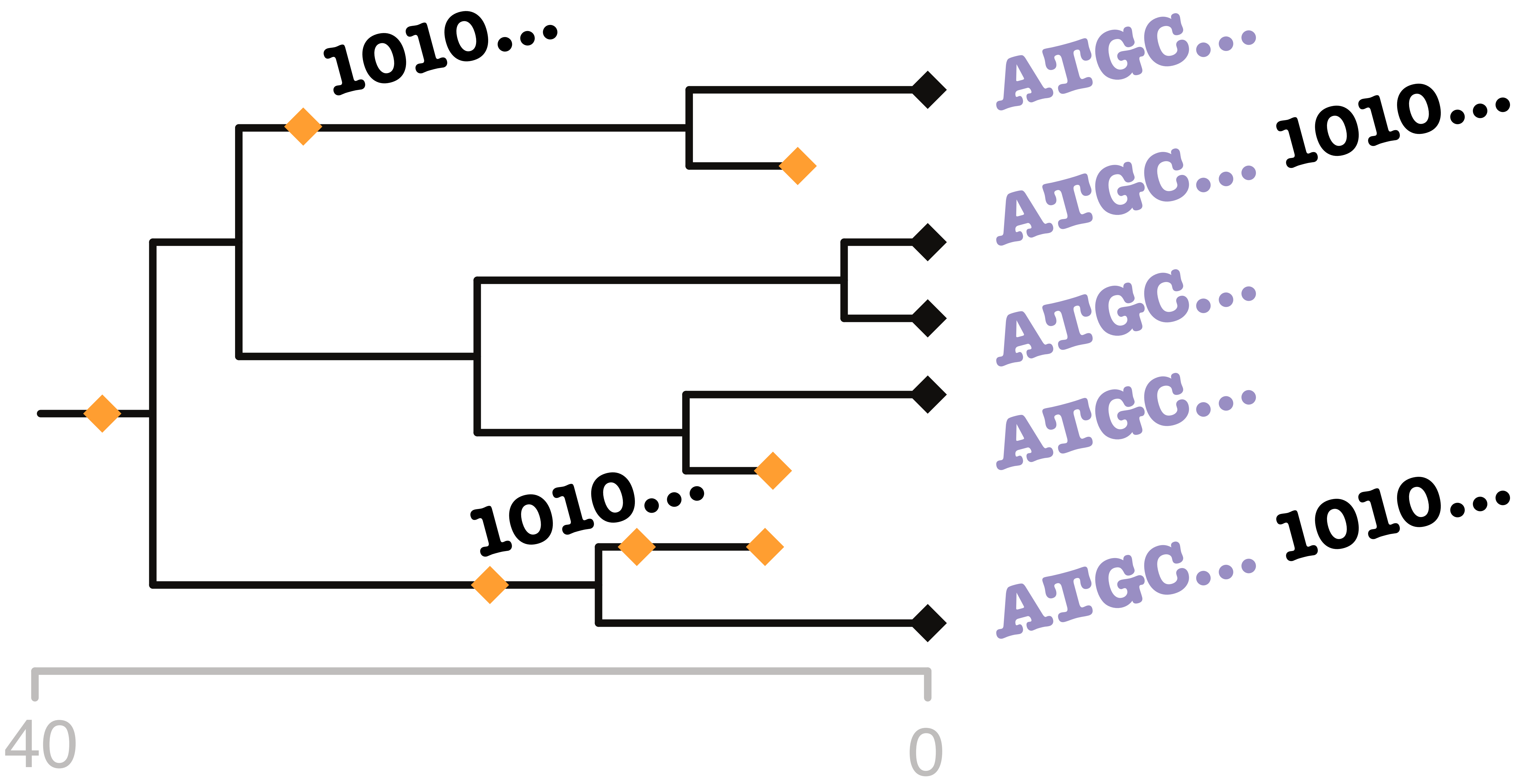


OCTOPI WALL STREET

Invertebrates are 97% of animal diversity!

Brought to you by Oregon Institute of Marine Biology, University of Oregon

Under the FBD process fossils can be included **with** and **without** character data

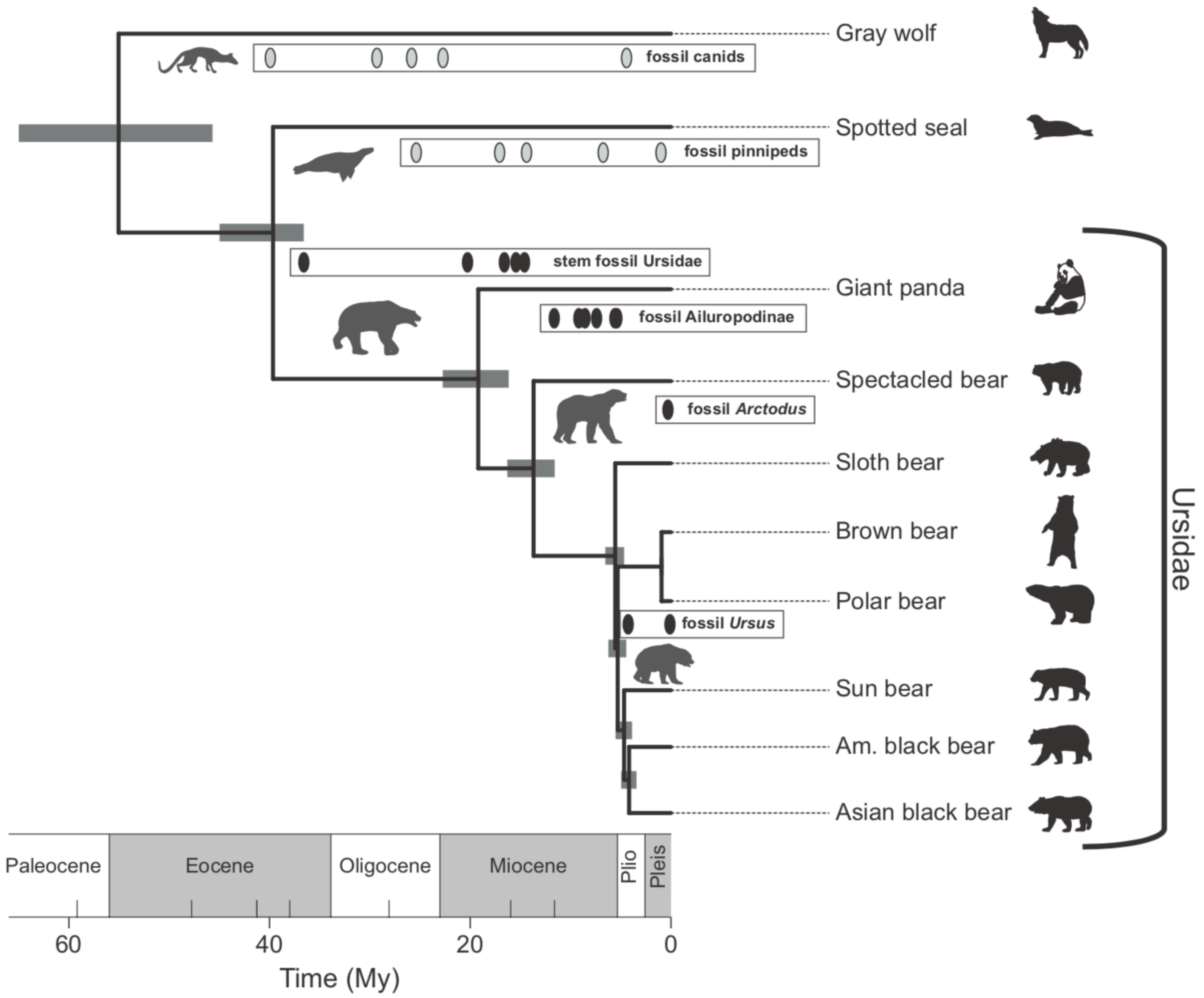


Note: For fossils without character <u>we can't infer the precise placement</u>, but we can take advantage of the additional age information, since this helps inform the FBD model parameters.

The signal for the extant tree topology largely comes from the molecular alignment. The signal for the topology inc. fossils comes from the morphological matrix *and* fossil ages.

The signal for diversification and sampling rates (& consequently the divergence times) comes from the fossil sampling times. Much more dependent the birth-death-sampling process

Putting the F in FBD analyses: tree constraints or morphological data? Barido-Sottani et al. bioRiv
See also Barido-Sottani et al. 2018, 2020.

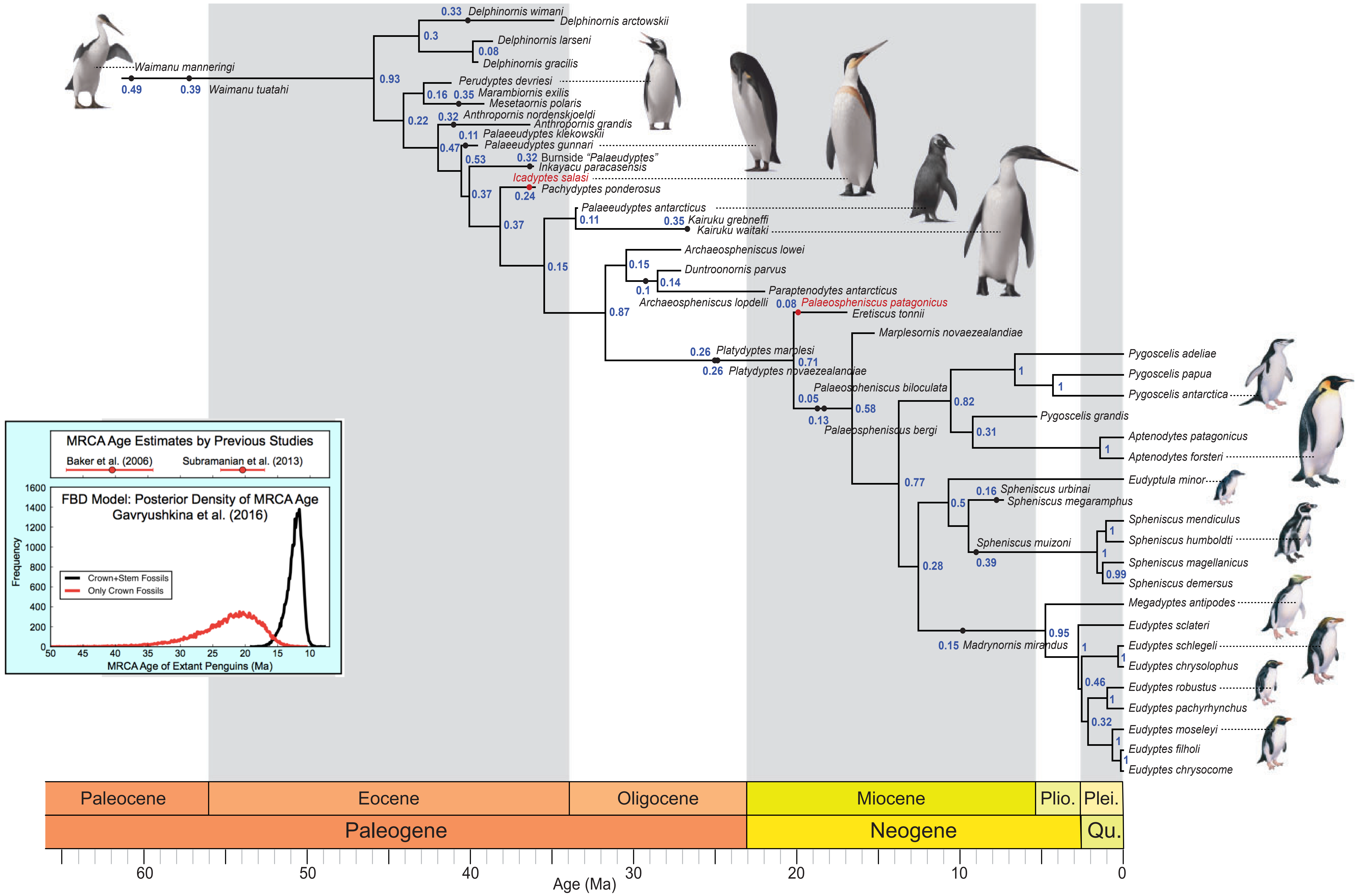


Time calibrated tree of living and fossil bears

First application of the FBD model.

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

Heath et al. 2014. PNAS

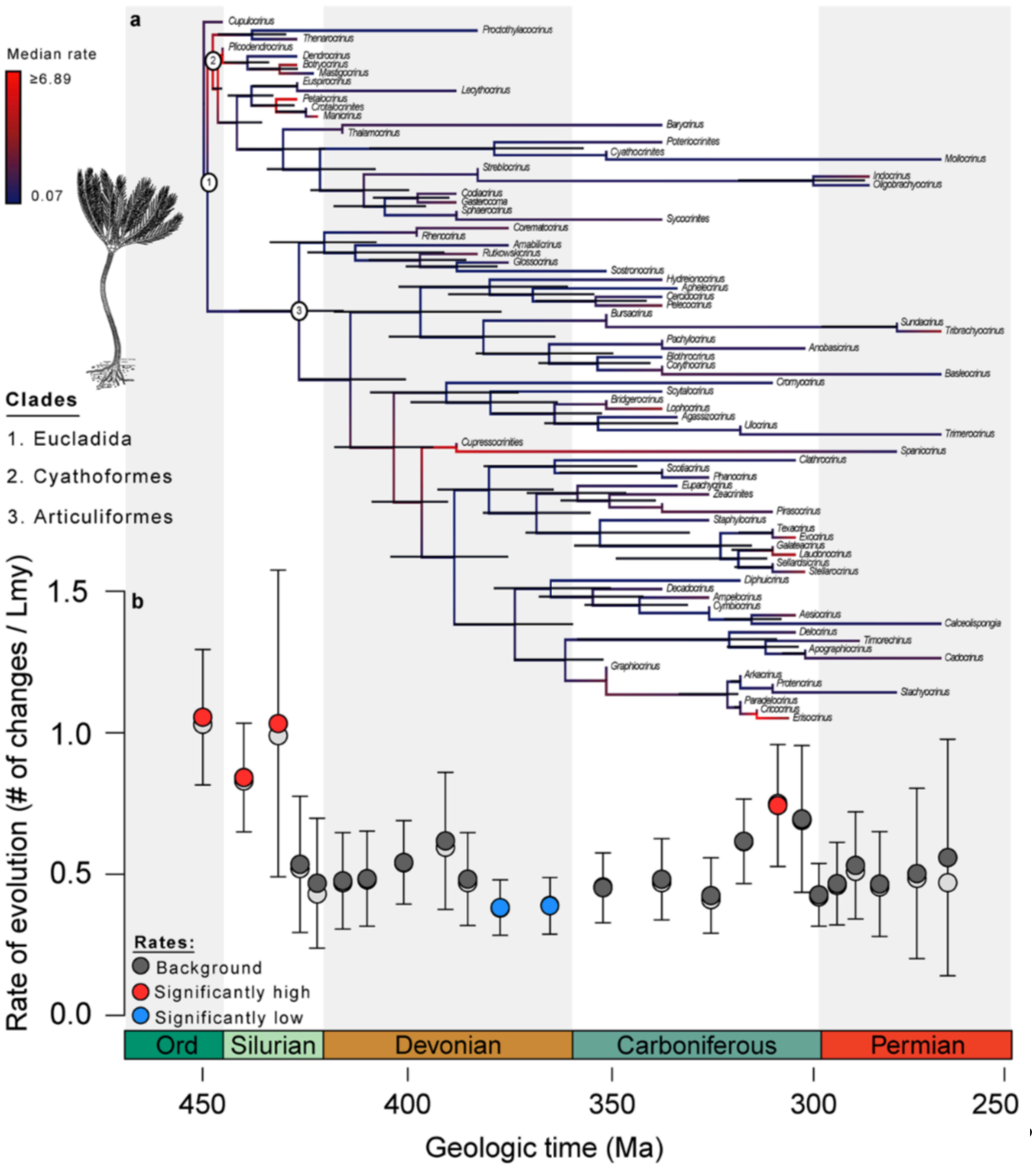


Time calibrated tree of living and fossil penguins

First application of total evidence dating using the FBD model.

Fossils are incorporated using character data, via a total evidence approach. Their placement can be inferred.

Gavryushkina et al. 2016. Sys Bio
See also: Zhang et al. 2016. Sys Bio

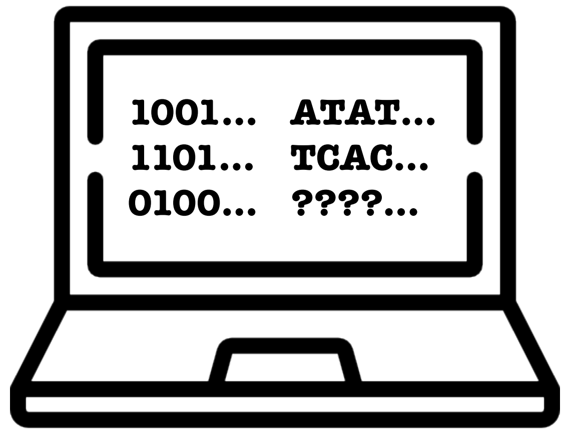


Analysis of fully extinct clades under the FBD process

Example using crinoids
Wright (2017) Sci Reports

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)

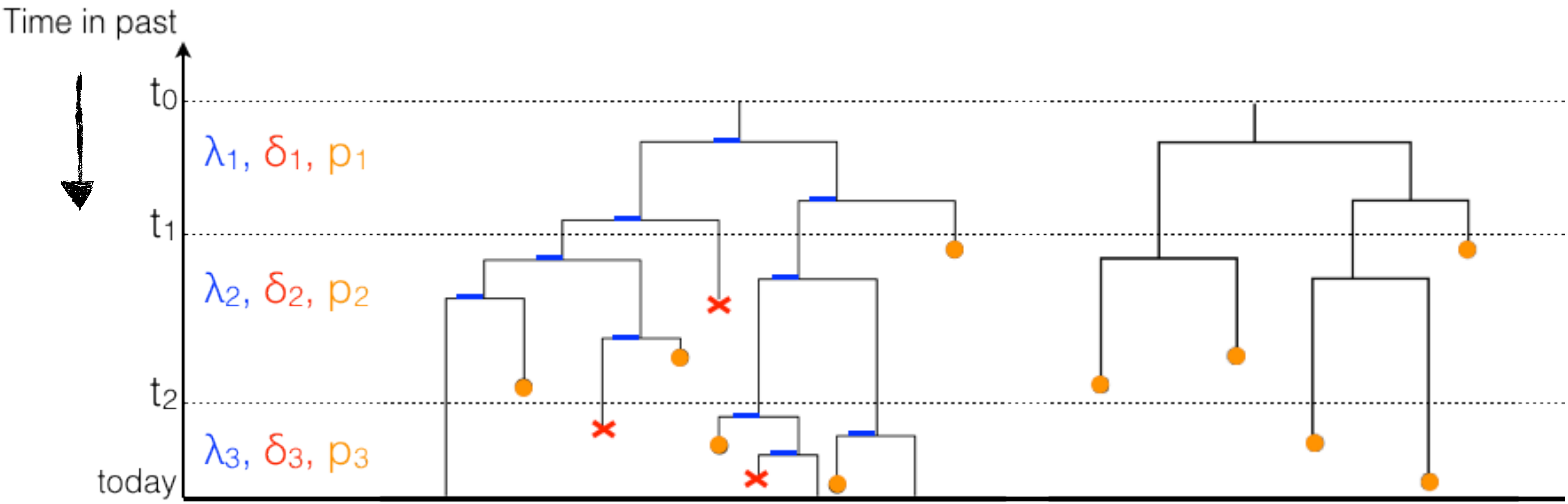
- RevBayes – FBD model, lots of flexible tree and character evolution models. C++. Uses graphical models. Developed by folk closer to macroevolution.

Exercise 6: Inference under the FBD process

A few additional points, if we have time.

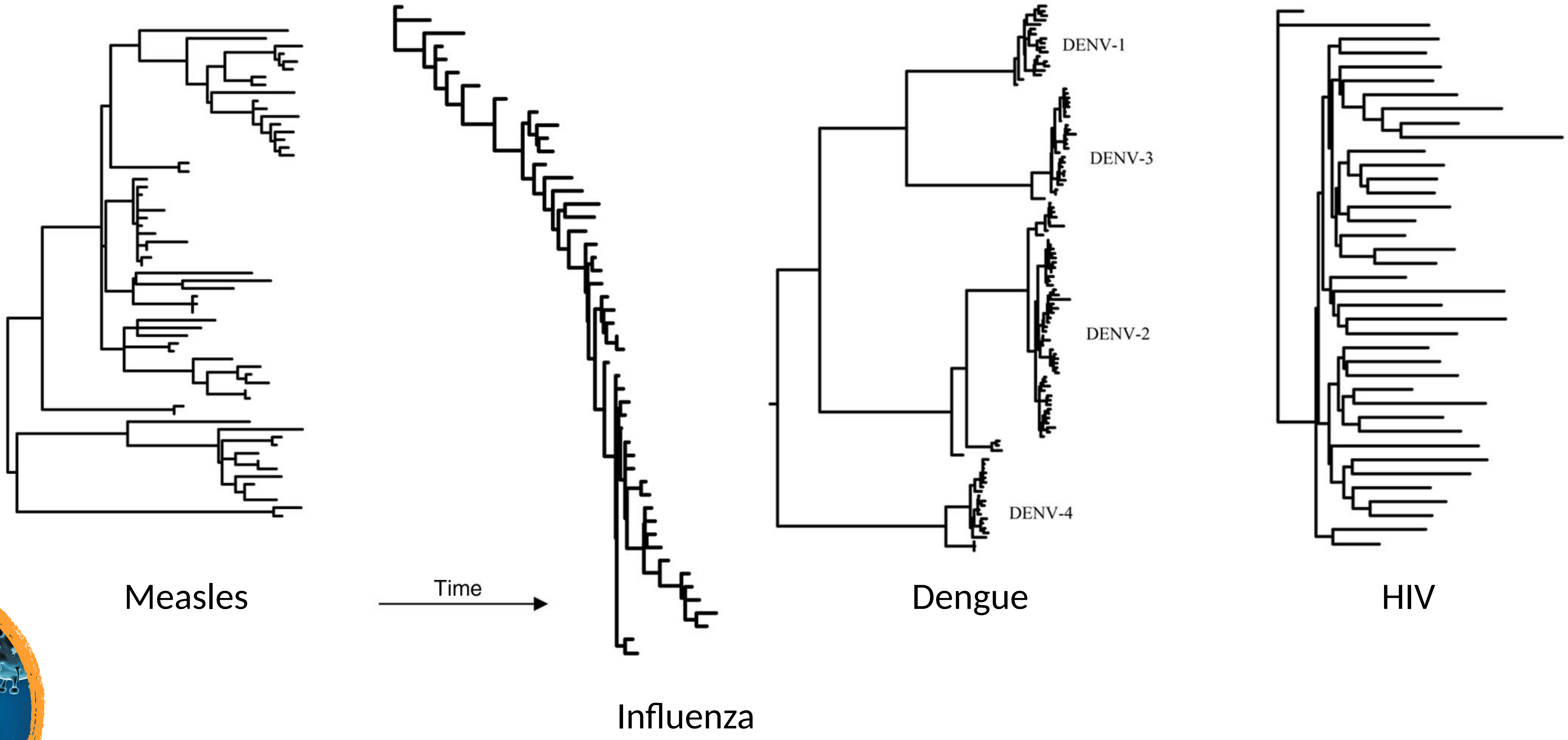
Skyline birth-death models

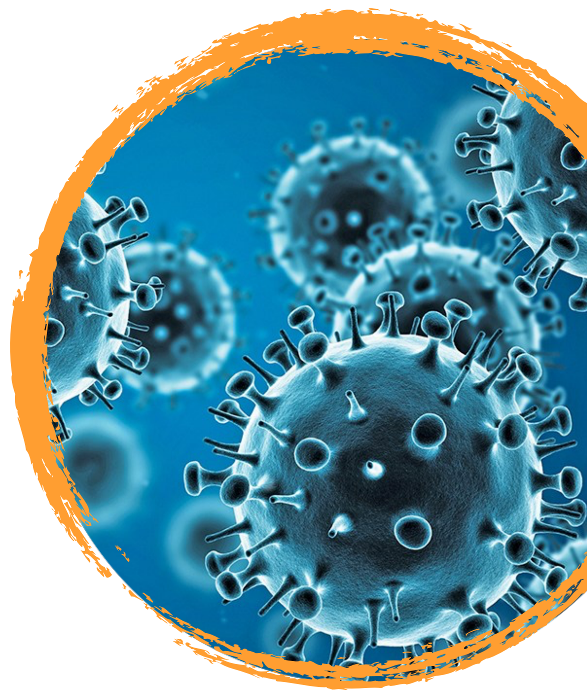
The skyline model incorporates piecewise constant rate variation



Stadler et al. 2012. PNAS.

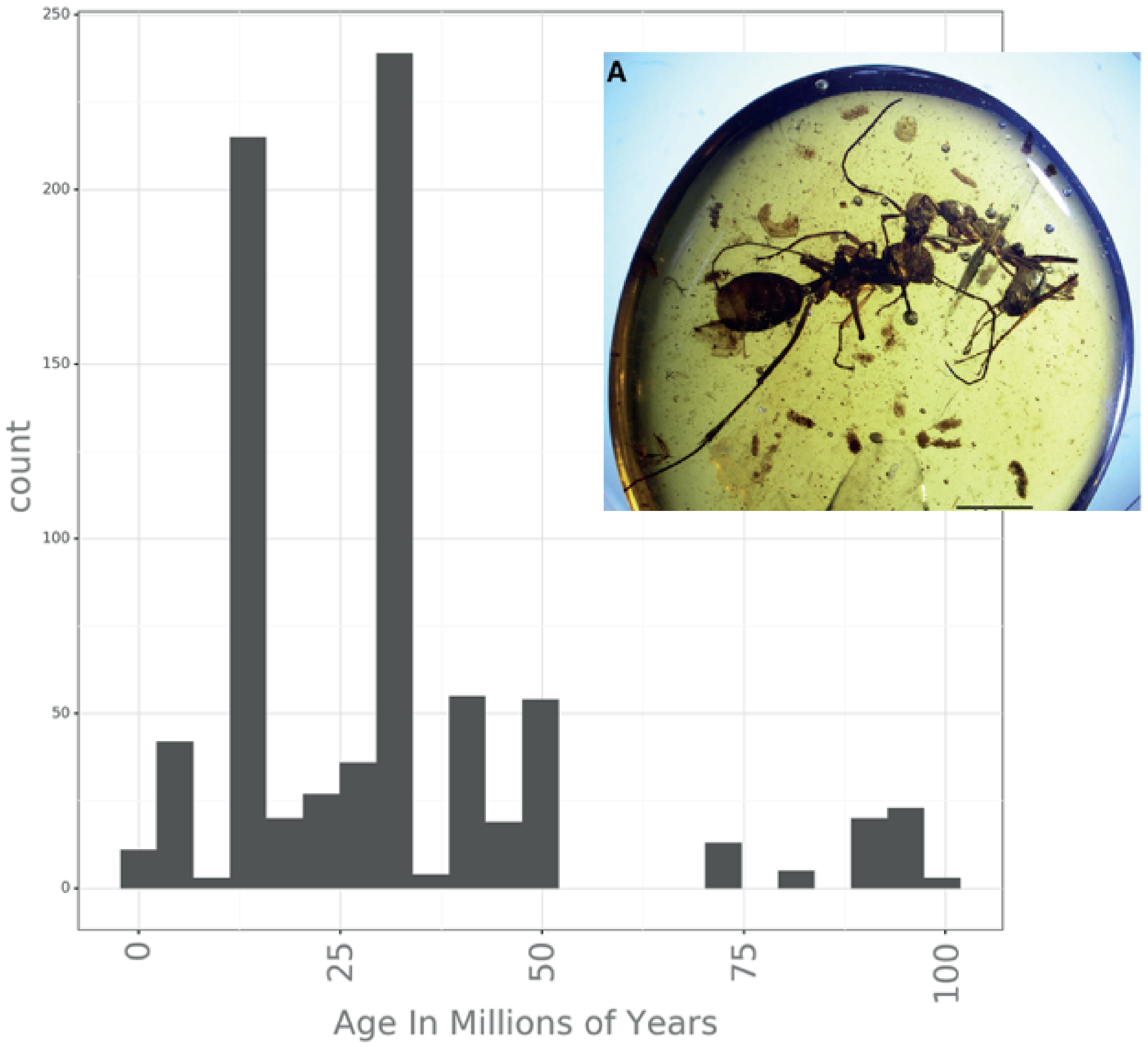
Tree shape is informative about underlying dynamics





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

Estimating parameters in macroevolution

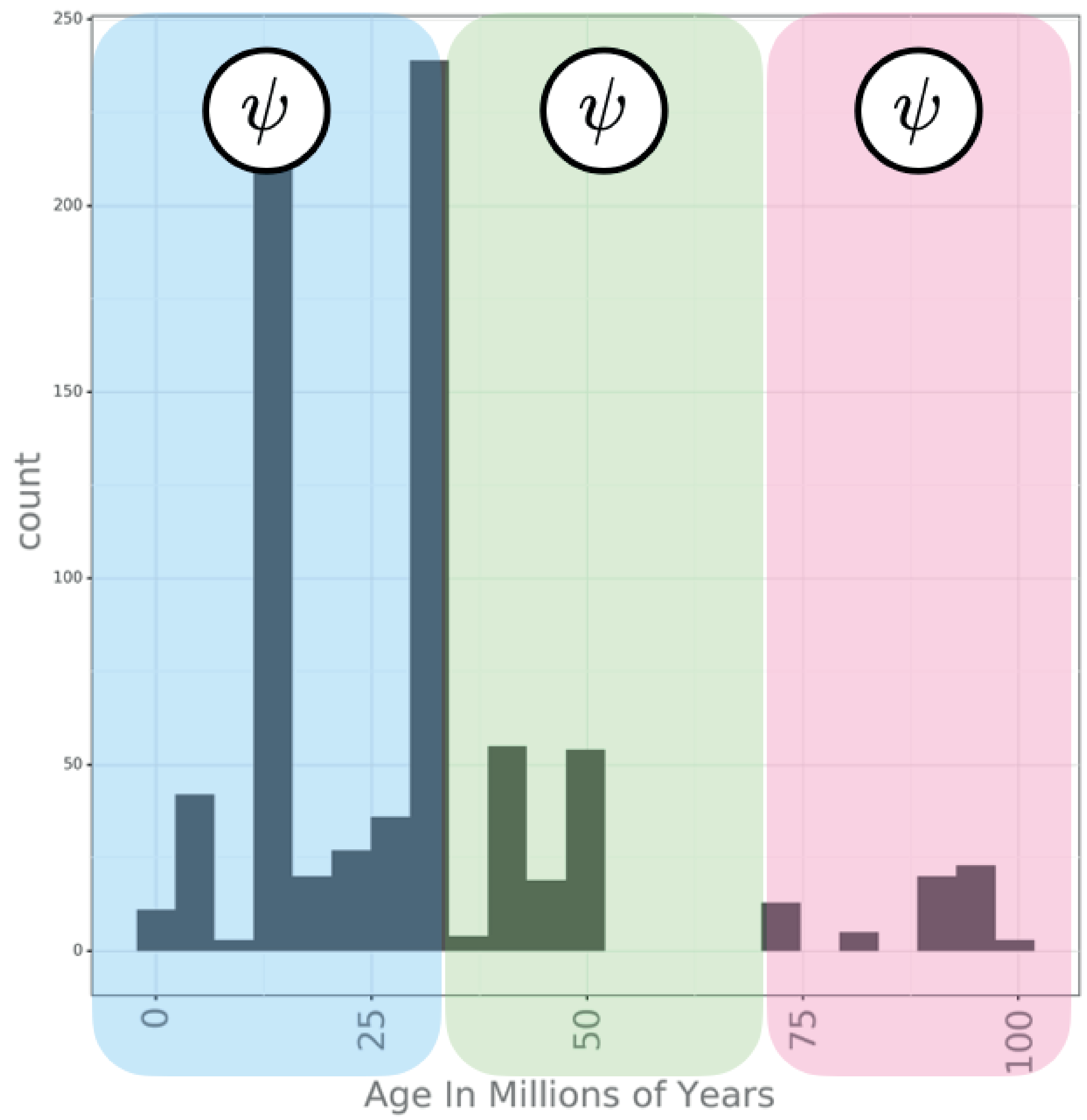


Ants have very variable fossil sampling over time.

We can take this into account using the FBD skyline model.

Images: April Wright

Estimating parameters in macroevolution

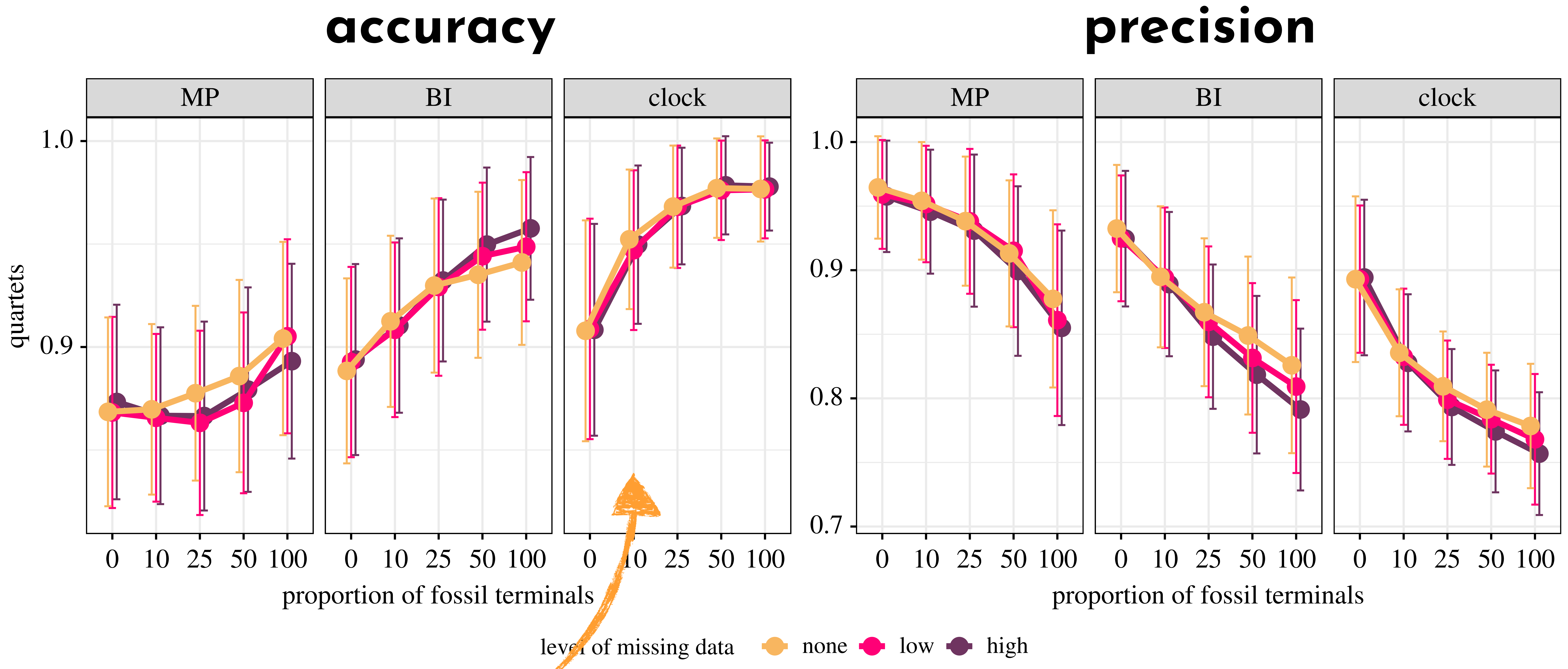


The oldest fossils are around 100 Ma.

Different assumptions about the fossil sampling process produce different results.

Skyline models recover an older age estimate for the origin of ants (= 140 Ma).

Images: April Wright



clock = inference with fossil ages

Time informs topology

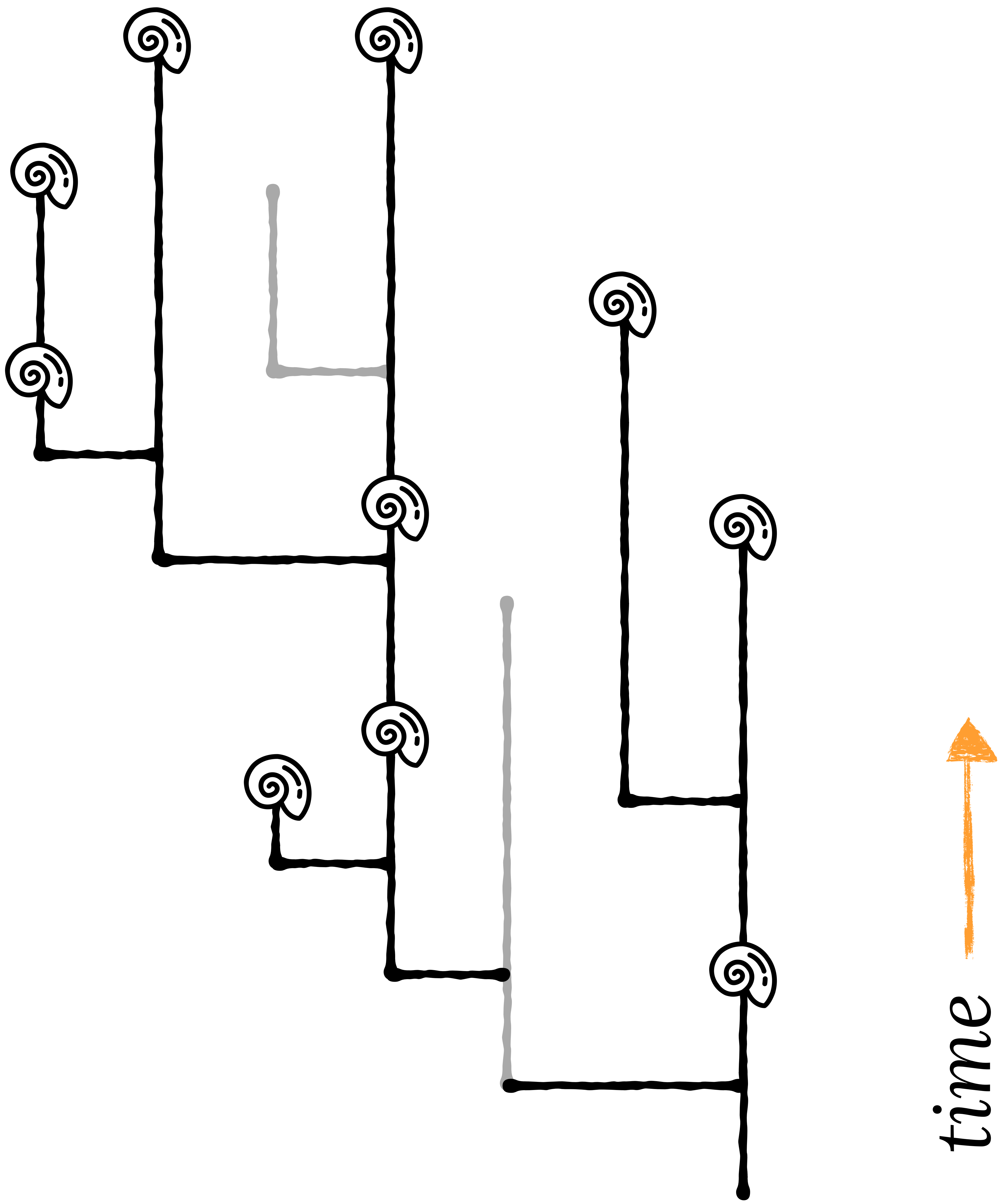
Fossils improve phylogenetic analyses of morphological characters
Koch, Garwood, Parry. 2020. Proc B

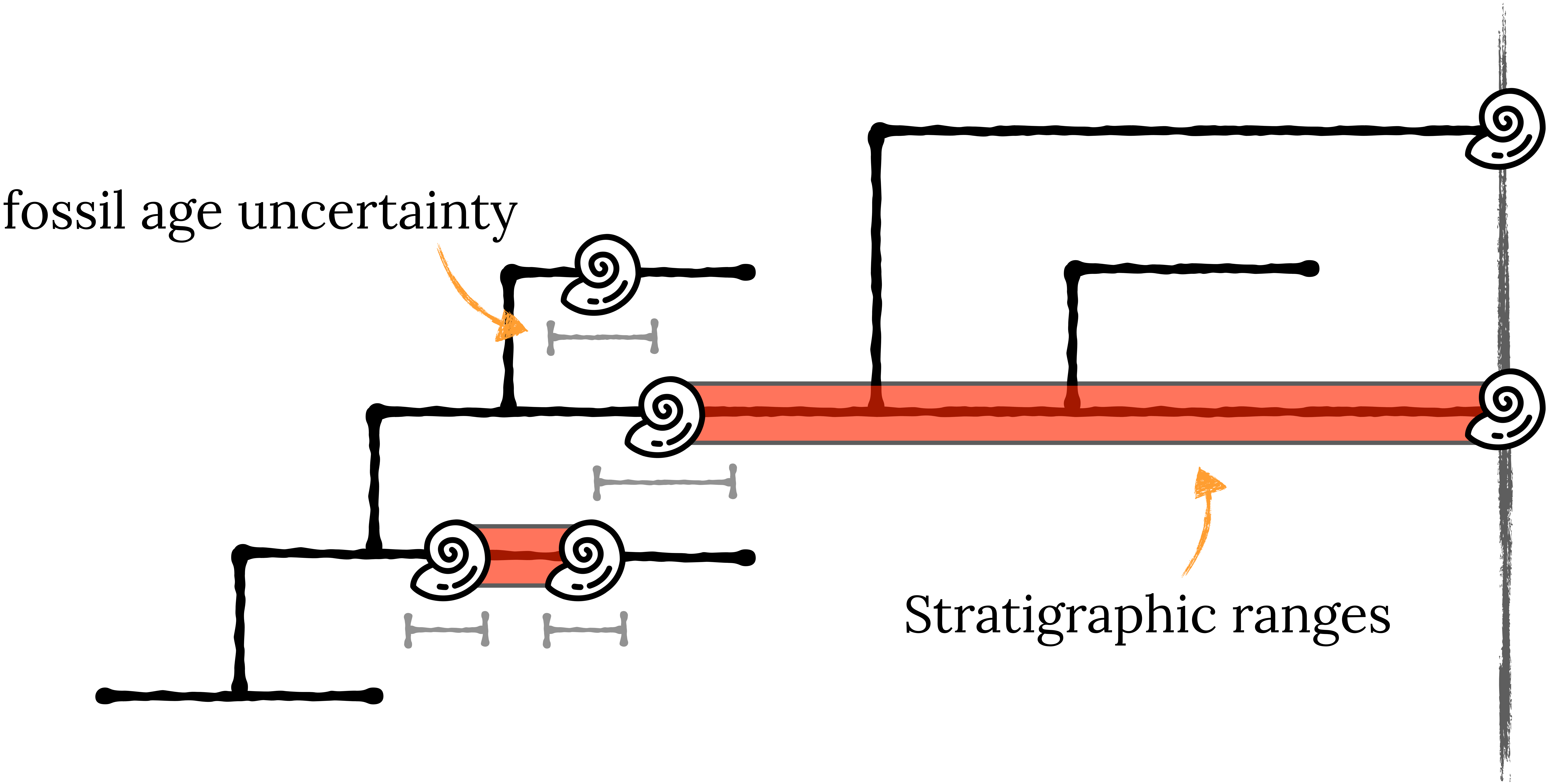
What is a **temporal observation** in the context of the FBD model? 🤔

Sampling events occur with instantaneous sampling rate ψ, assuming a Poisson sampling process.

This means, there's an exponential waiting time between events.

At any given point in time, a lineage (branch) can only be represented once.



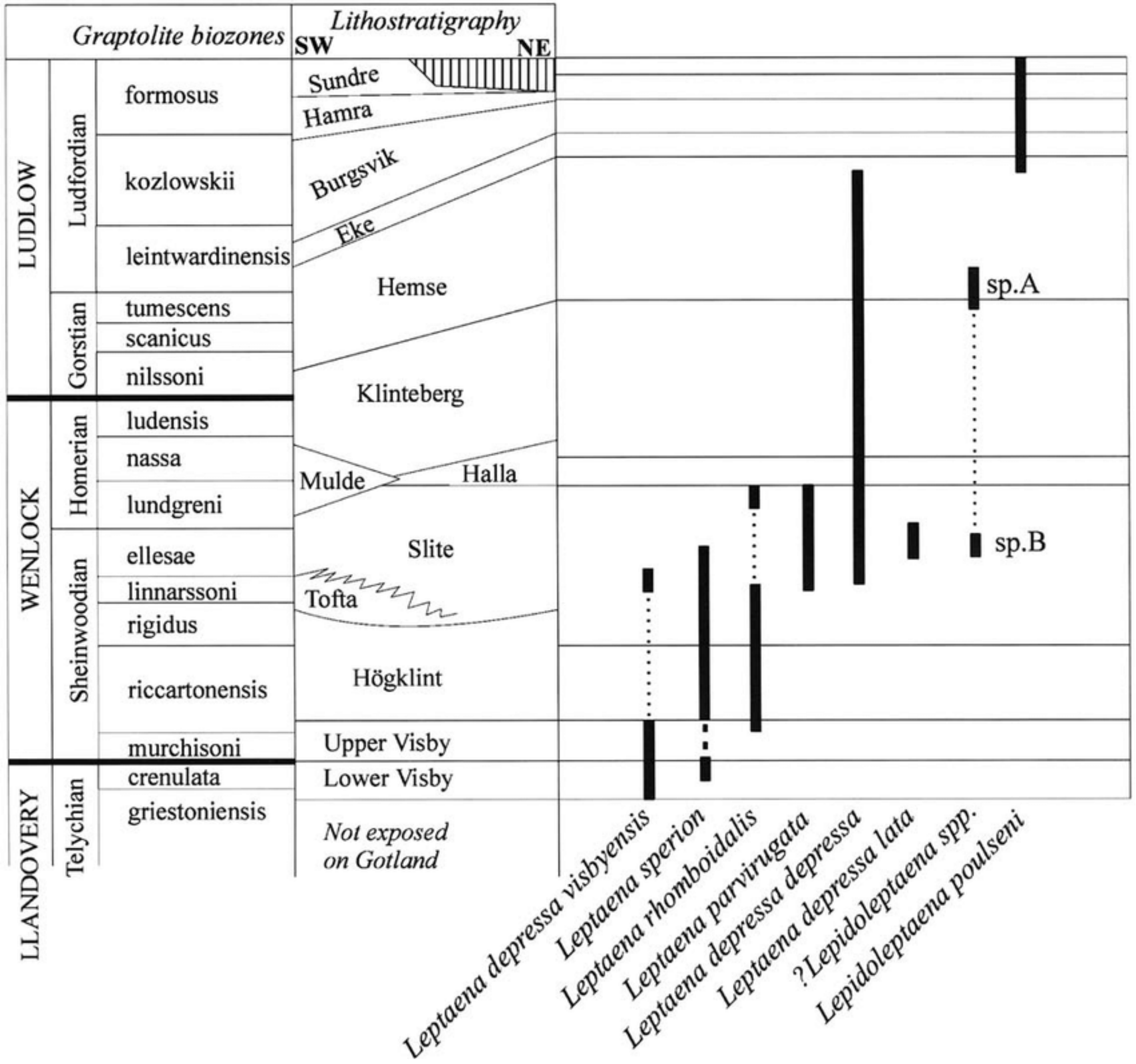


Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology in Time Calibrated Tree Inference
Barido-Sottani et al. 2018, 2020.

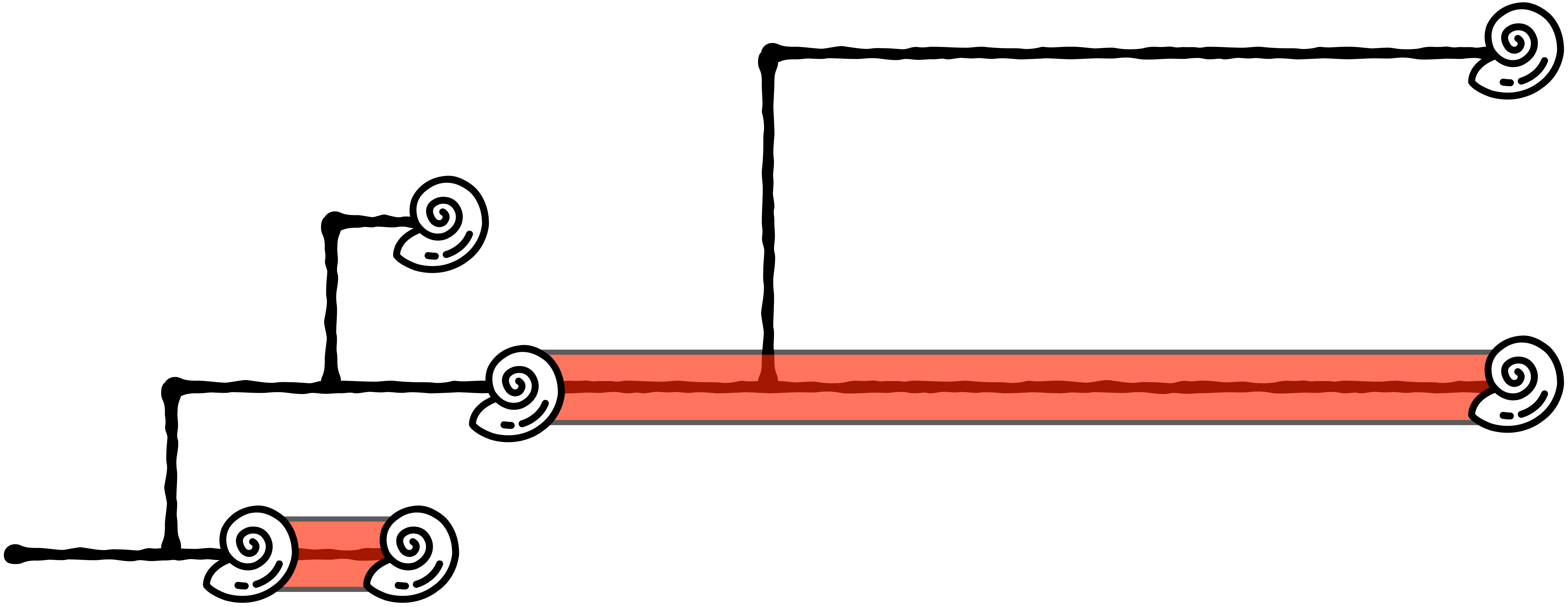
Stratigraphic range

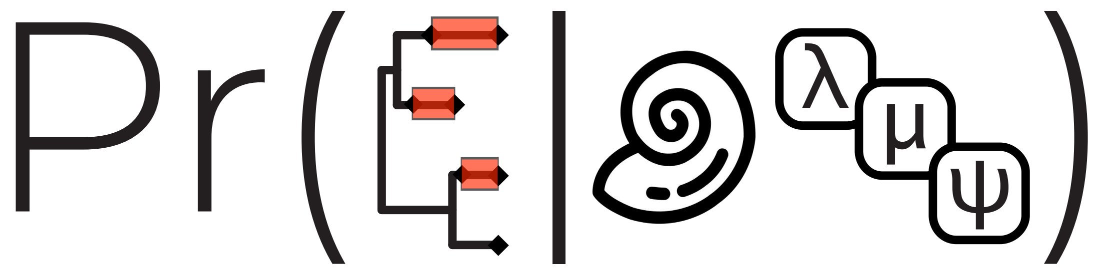
Fossil taxa are associated with first

Brachiopod ranges from Gotland

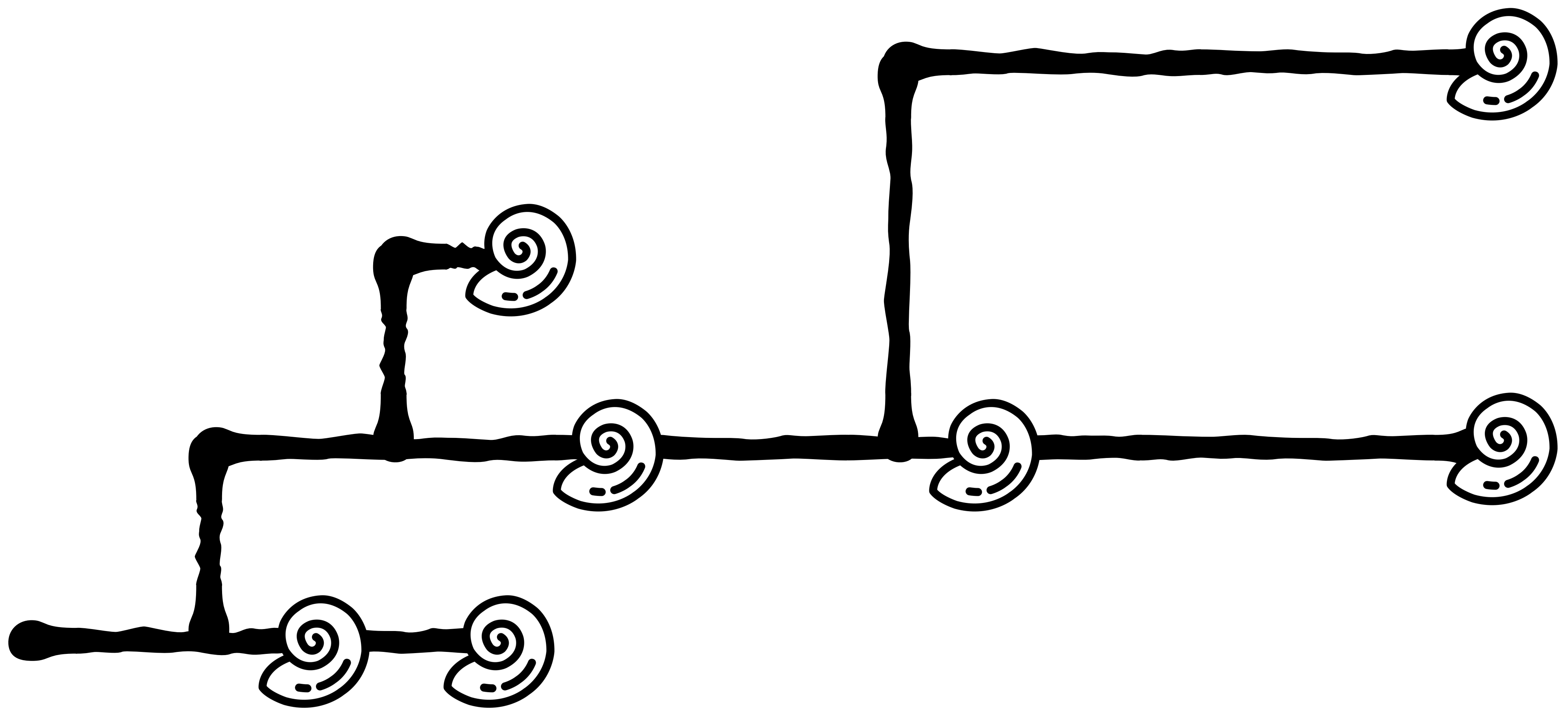


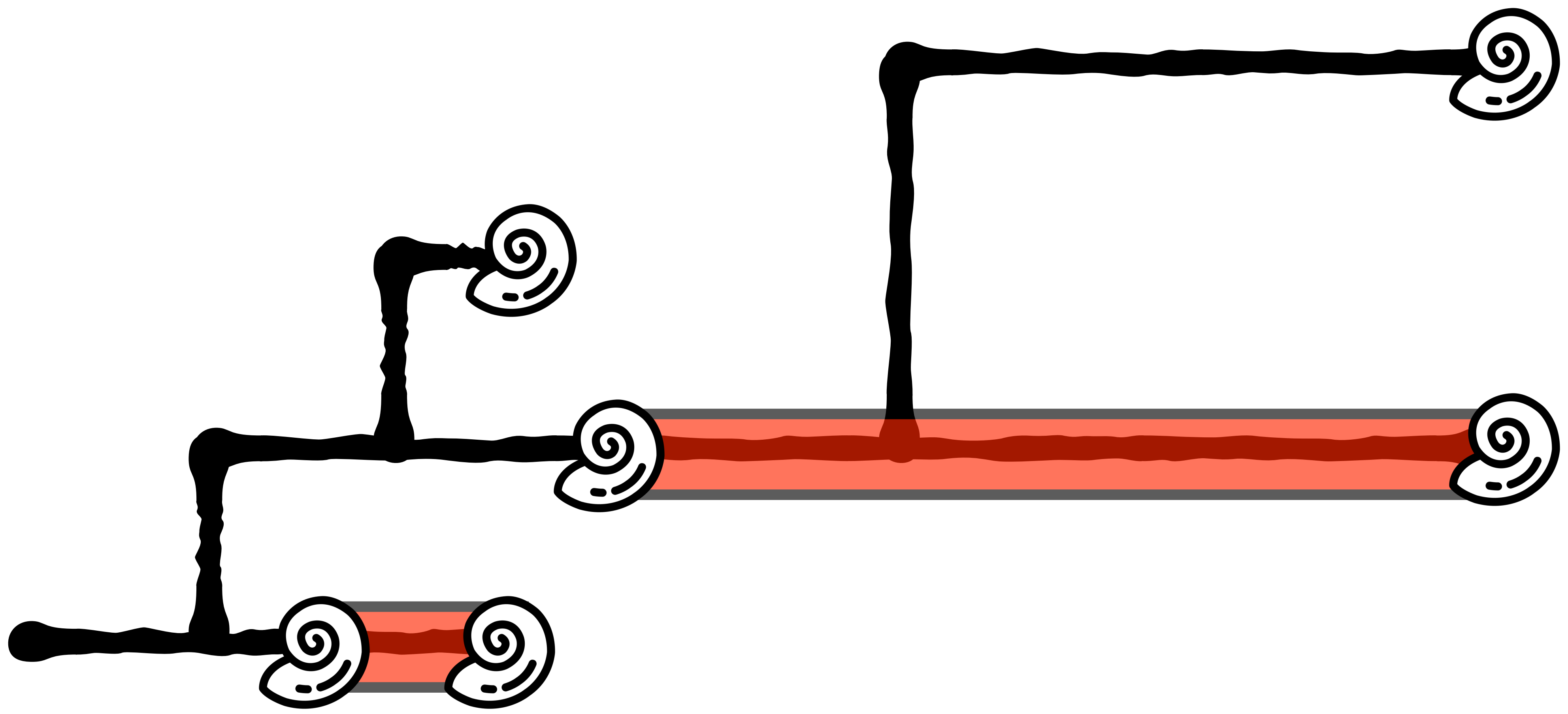
The FBD process for the analysis of **stratigraphic ranges**





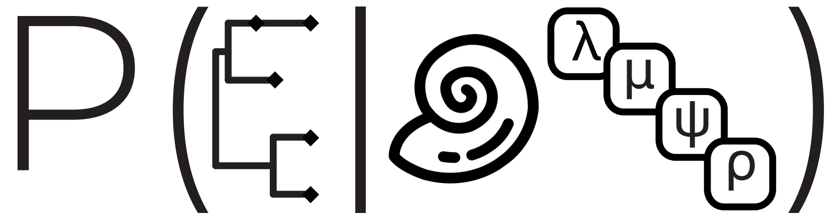
Analysis of stratigraphic range data
Stadler et al. 2018. JTB
See also Warnock, Heath, Stadler. 2020. Paleobiology

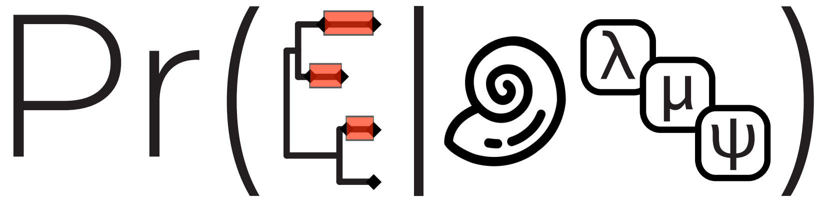




specimen level FBD process

morpho taxon FBD process





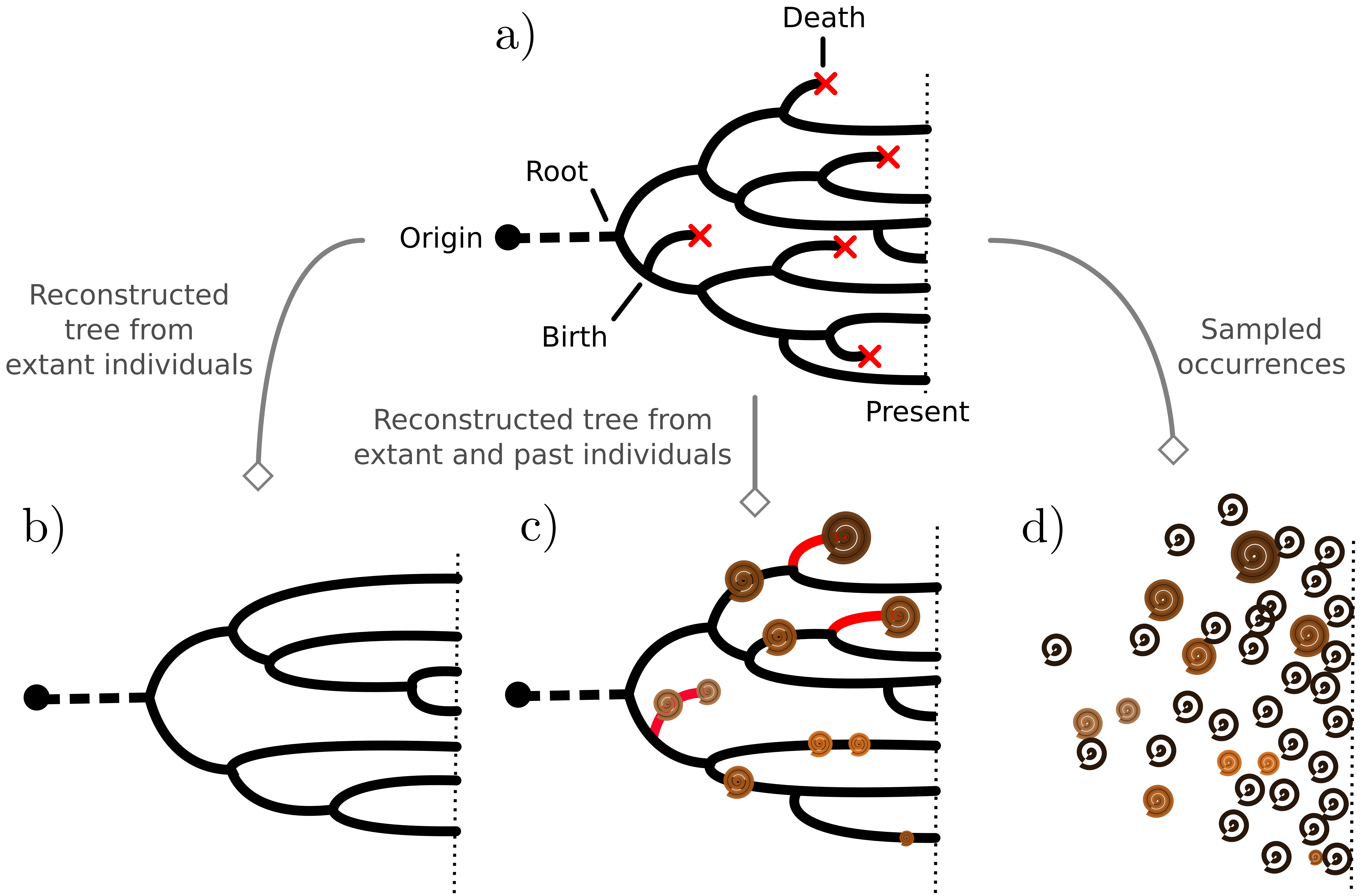
These belong to a growing family of models.

Parameterisation of the process

Parameter	**Transformation**
Net diversification	$d = \lambda - \mu$
Turnover	$v = \mu/\lambda$
Sampling proportion	$s = \psi/(\mu + \psi)$
Speciation	$\lambda = d/(1 - v)$
Extinction	$\mu = (vd)(1 - v)$
Sampling	$\psi = (s/(1 - s))((vd)/(1 - v))$

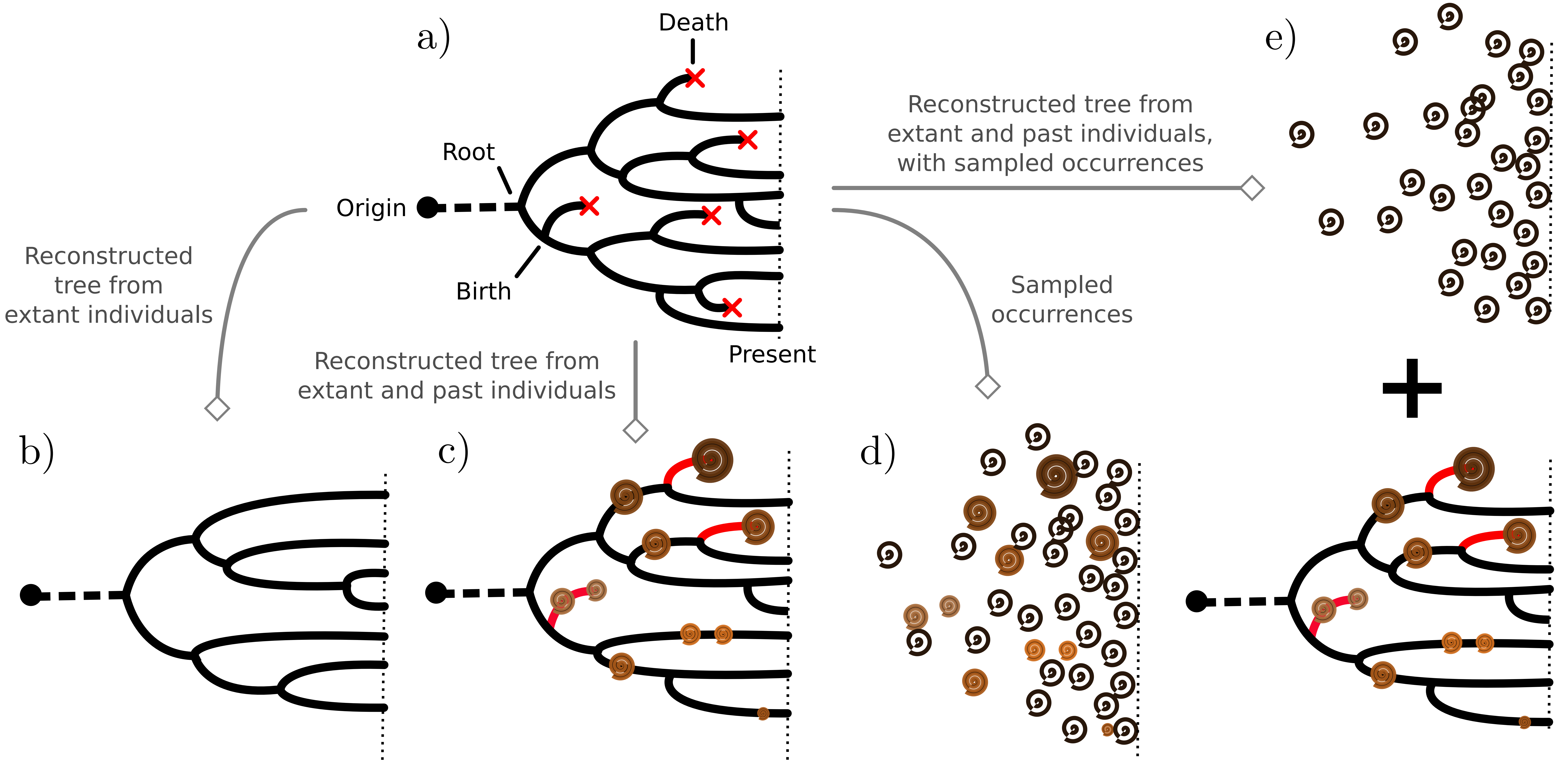
We can also transform fossil recovery rate into a per-interval sampling probability.

The occurrence birth-death process models



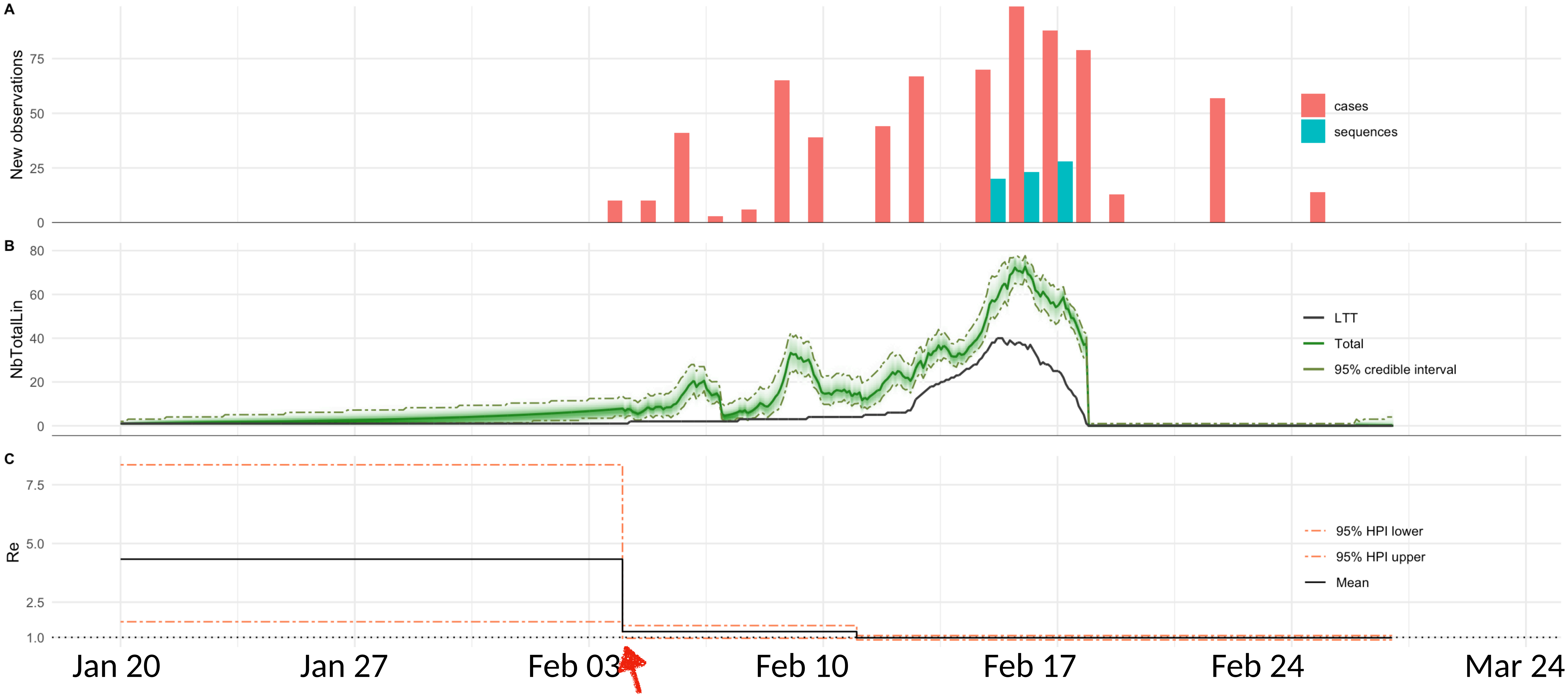
Andréoletti, Zwaans, et al. bioRxiv.

The occurrence birth-death process



Andréoletti, Zwaans, et al. bioRxiv.

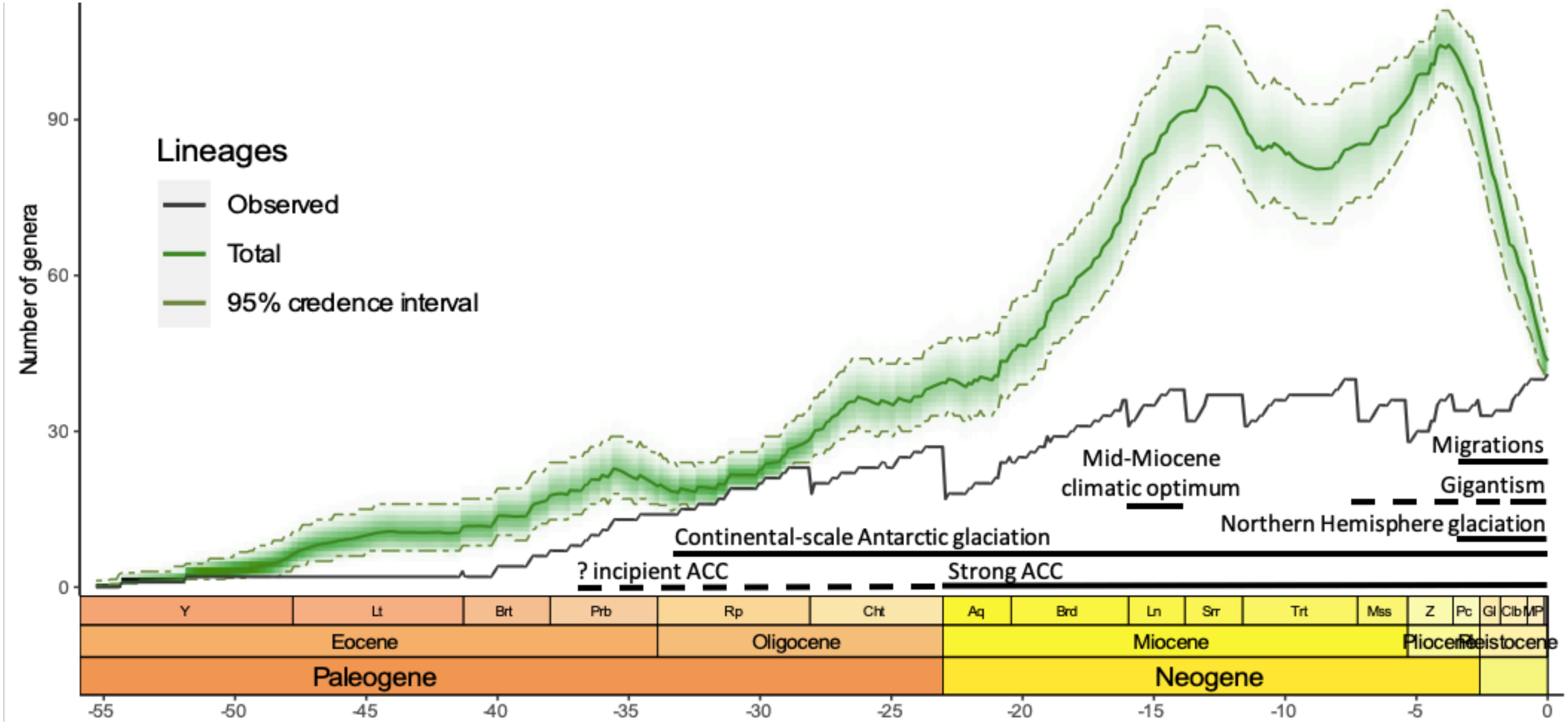
Case study: COVID-19 aboard the Diamond Princess



Cabin quarantine

Andréoletti, Zwaans, et al. bioRxiv.

Casestudy: cetaceans



Andréoletti, Zwaans, et al. bioRxiv.

These models aren't without caveats 😬

Phylogenetic dating is a kerfuffle. The theory is complex and inference is expensive.

Model identifiably

Louca & Pennell 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature.

Louca et al. 2021. Fundamental Identifiability Limits in Molecular Epidemiology. MBE.

It's not the end of the world!! See e.g., Morlon et el. 2023 (TREE), Kopperud et al. 2023 (PNAS).

Model selection

May & Rothfels 2023. Diversification models conflate likelihood and prior, and cannot be compared using conventional model-comparison tools. Sys Bio.