

Phylogenetics

Introduction to
phyldynamics models
RL-V3 MPP

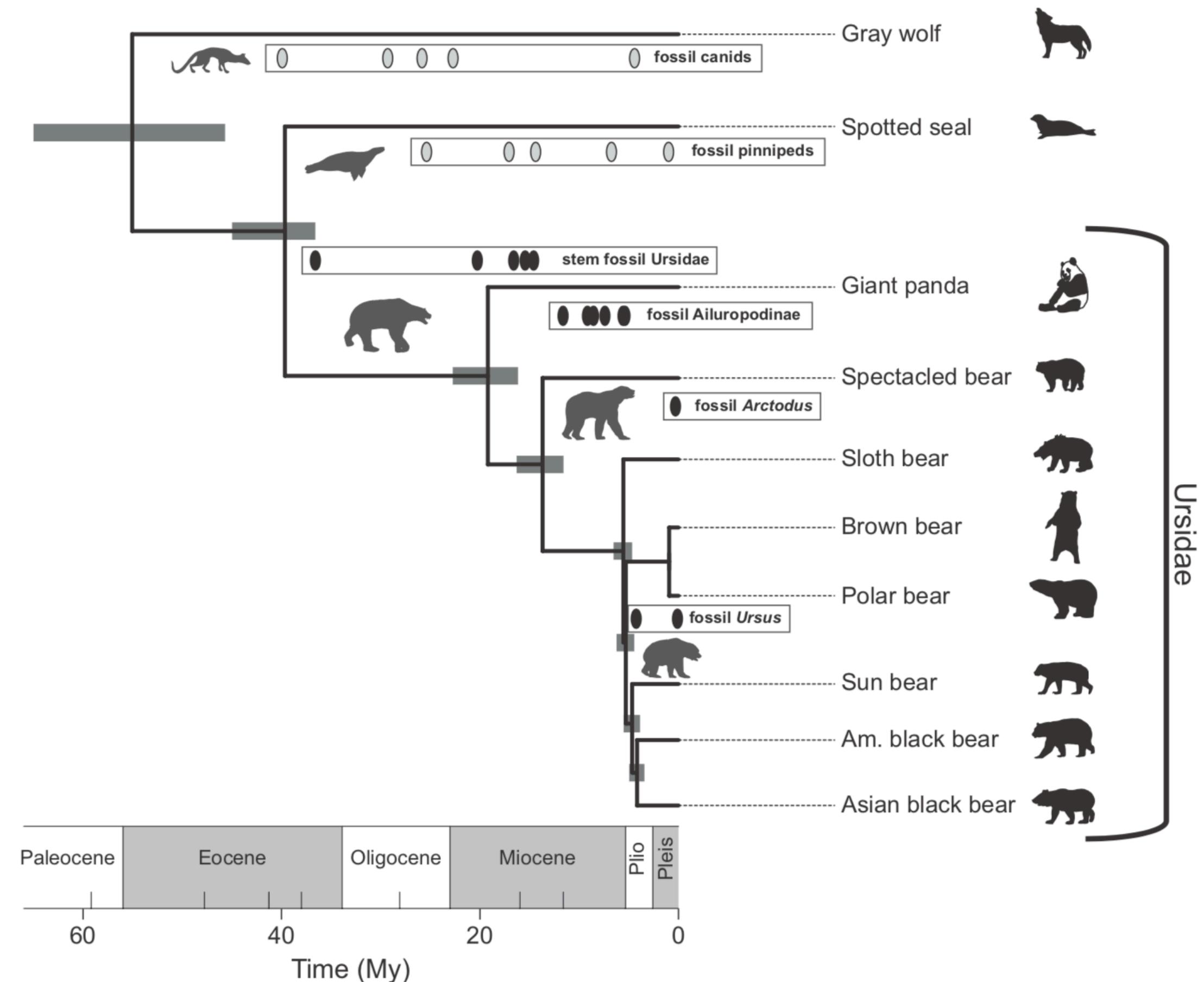
Rachel Warnock

16.04.2025



Objectives

- Recap: tripartite framework
- The fossilised birth-death process
- Total-evidence dating
- Phylodynamics



Bayesian divergence time estimation

Recap

We use a Bayesian framework

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

likelihood

priors

posterior

marginal probability of the data

Bayesian divergence time estimation

The data

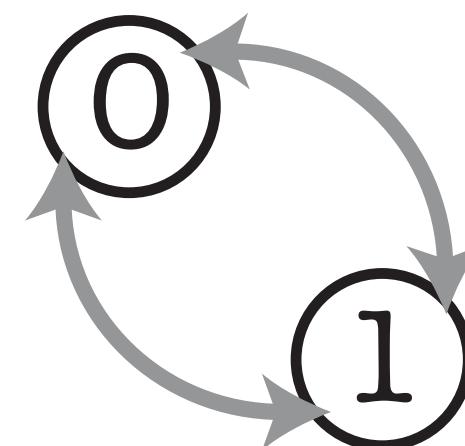
and / or
0101... ATTG...
1101... TTGC...
0100... ATTC...



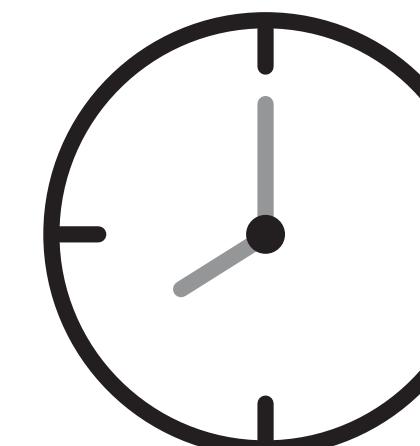
phylogenetics
characters

sample
ages

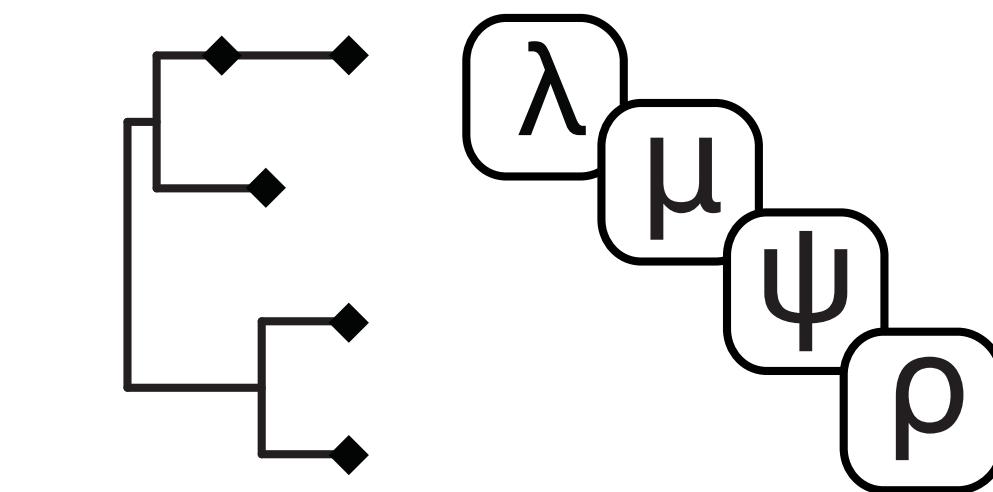
3 model components



substitution
model



clock
model



tree and tree
model

Bayesian divergence time estimation

posterior

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

likelihood

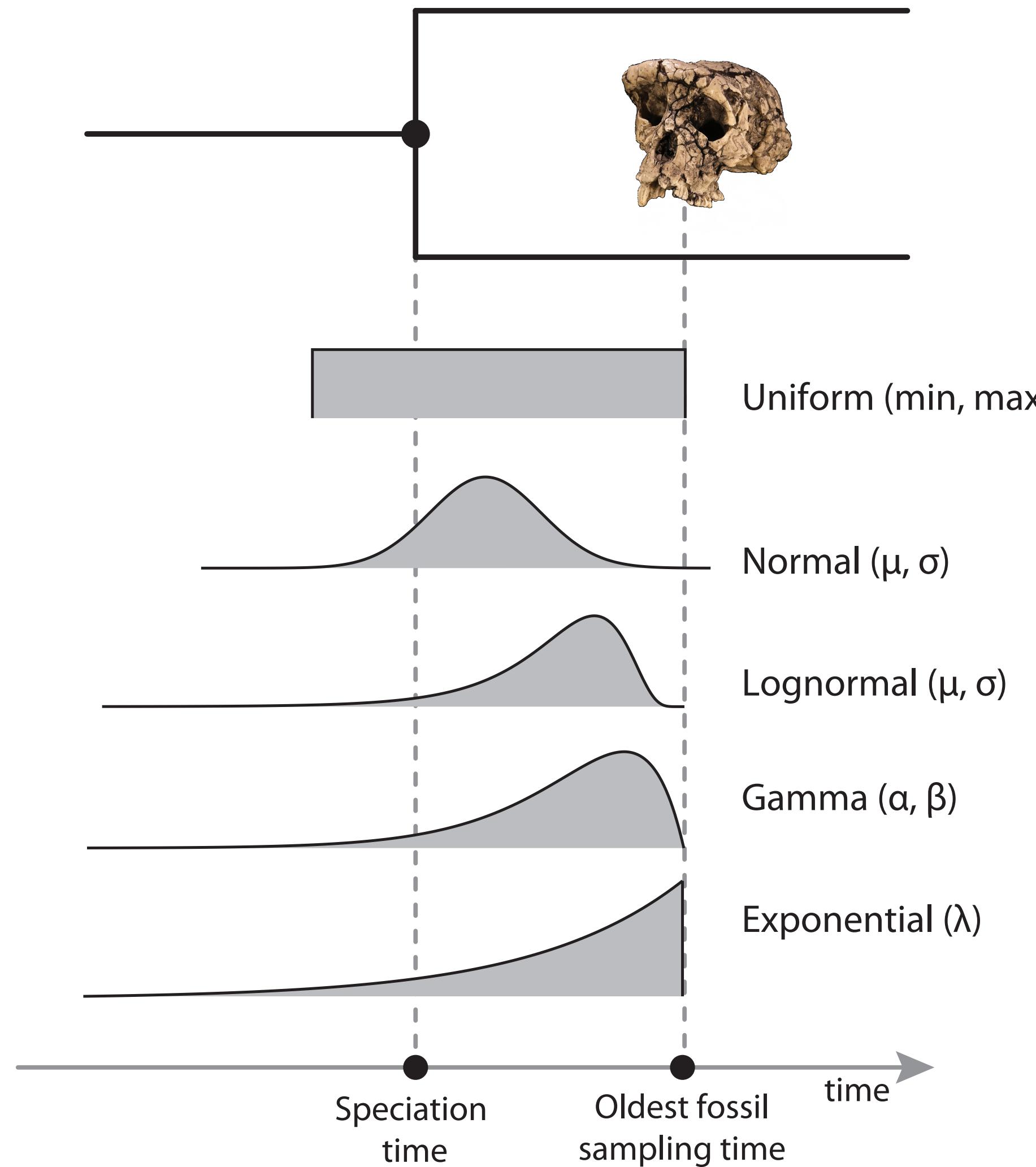
probability of the
time tree

priors

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

marginal pr of the data

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

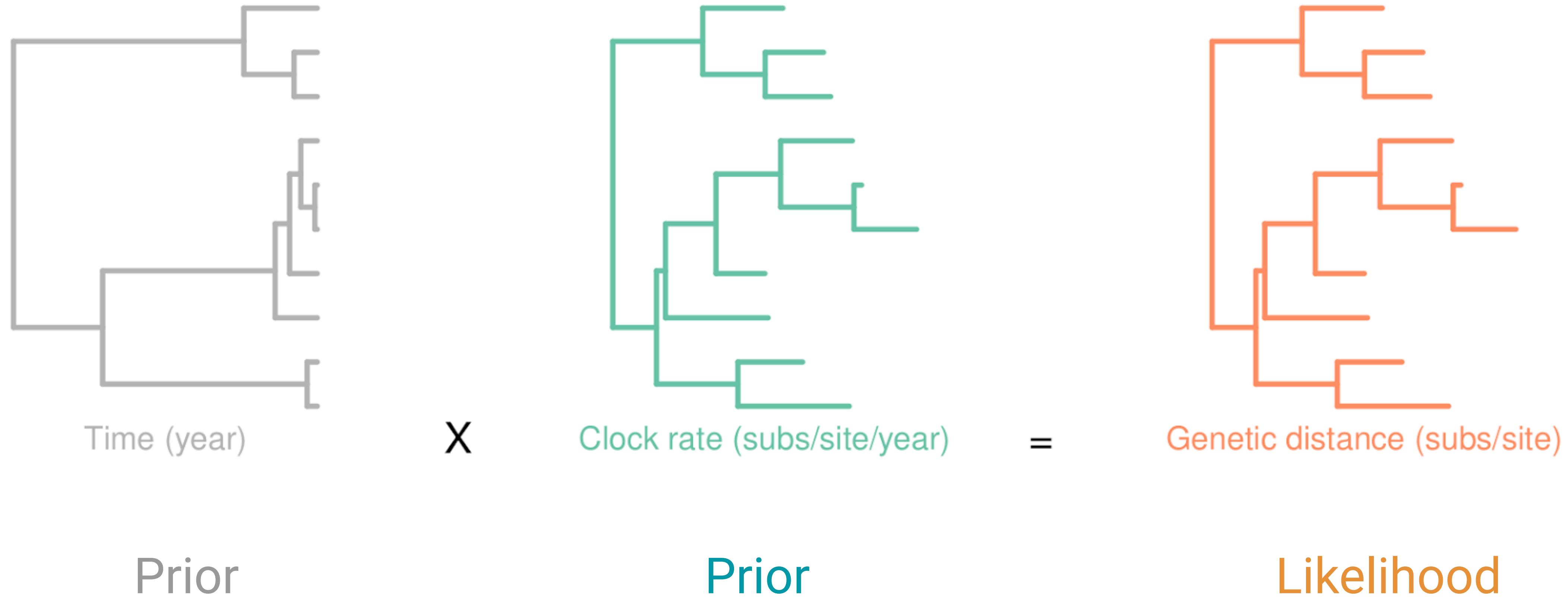
Adapted from Heath (2012). Sys Bio

Issues with node dating

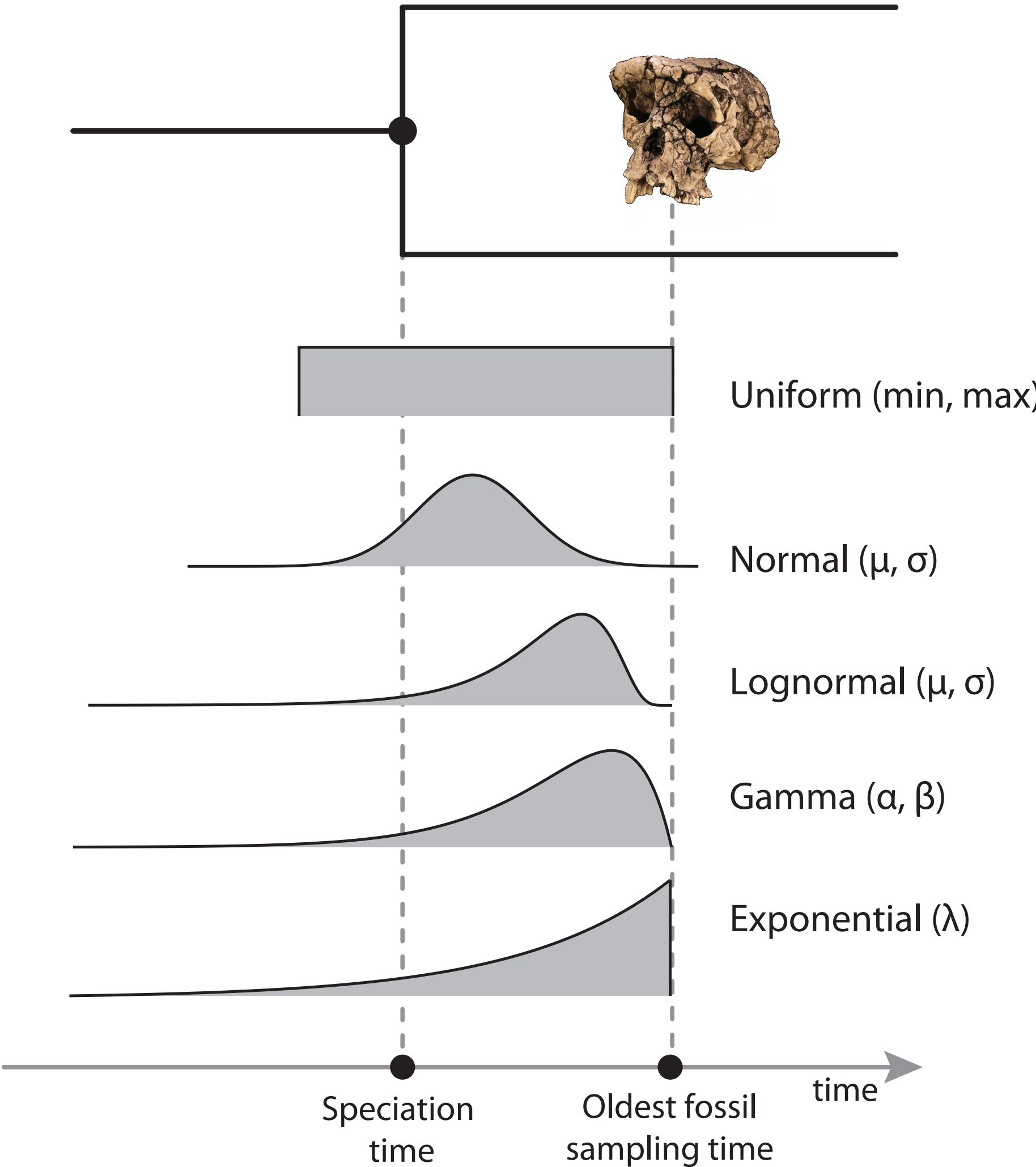
1. Rate and time is non-identifiable
2. Calibrations are hard to define objectively
3. The “effective priors” do not match the user specified priors
4. We’re potentially excluding a lot of information

The priors will always influence our results

Rate and time are not identifiable - not a conventional Bayesian problem



Minimum and maximum constraints



Hard minimum bounds are based on first appearances

Soft maximum bounds are based on more tenuous evidence – typically the 97.5% limit of the calibration density

The upper bounds often influence the results

Maximum constraints

Best Practices for Justifying Fossil Calibrations

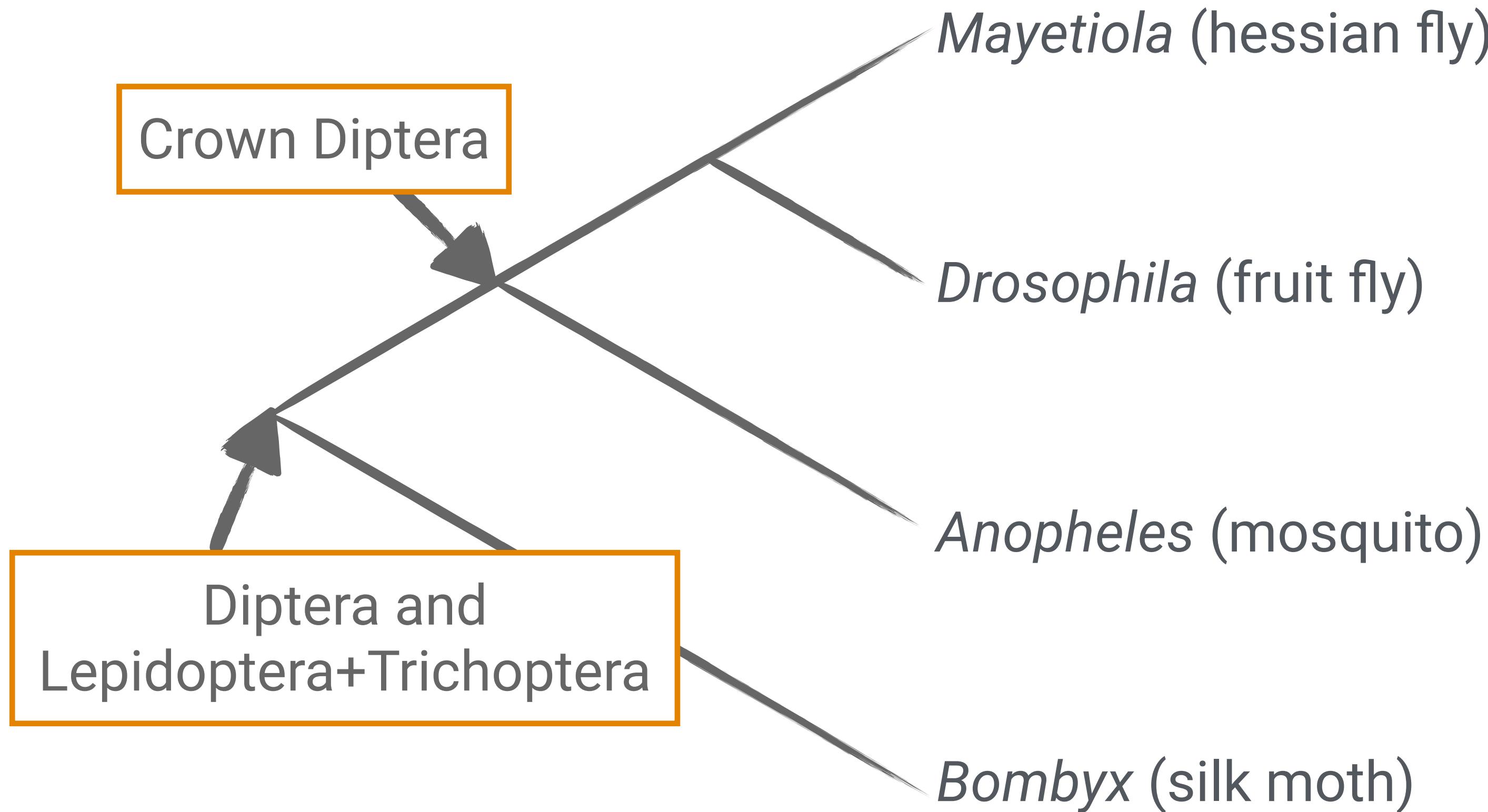
JAMES F. PARHAM^{1,2,*}, PHILIP C. J. DONOGHUE³, CHRISTOPHER J. BELL⁴, TYLER D. CALWAY⁵,
JASON J. HEAD⁶, PATRICIA A. HOLROYD⁷, JUN G. INOUE⁸, RANDALL B. IRMIS⁹,
WALTER G. JOYCE¹⁰, DANIEL T. KSEPKA^{11,12}, JOSÉ S. L. PATANÉ¹³, NATHAN D. SMITH^{14,15},
JAMES E. TARVER^{3,16}, MARCEL VAN TUINEN¹⁷, ZIHENG YANG¹⁸, KENNETH D. ANGIELCZYK¹⁵,
JENNY M. GREENWOOD³, CHRISTY A. HIPSLEY^{19,20}, LOUIS JACOBS²¹, PETER J. MAKOVICKY¹⁵,
JOHANNES MÜLLER¹⁹, KRISTER T. SMITH²², JESSICA M. THEODOR²³, RACHEL C. M. WARNOCK³,
AND MICHAEL J. BENTON³

Goal - to make calibration choices transparent and explicit

"The [soft] maximum constraint is established as older than all the oldest possible records, extending back to encompass a time when the ecologic, biogeographic, geologic, and taphonomic conditions for the existence of the lineage are met, but no records are known."

Min & max constraints

Example from insects



Minimum (all nodes)

- 238.5 Ma
- Triassic Grès-a-Voltzia Frm, France
- Earliest (non-controversial) evidence for all 4 lineages

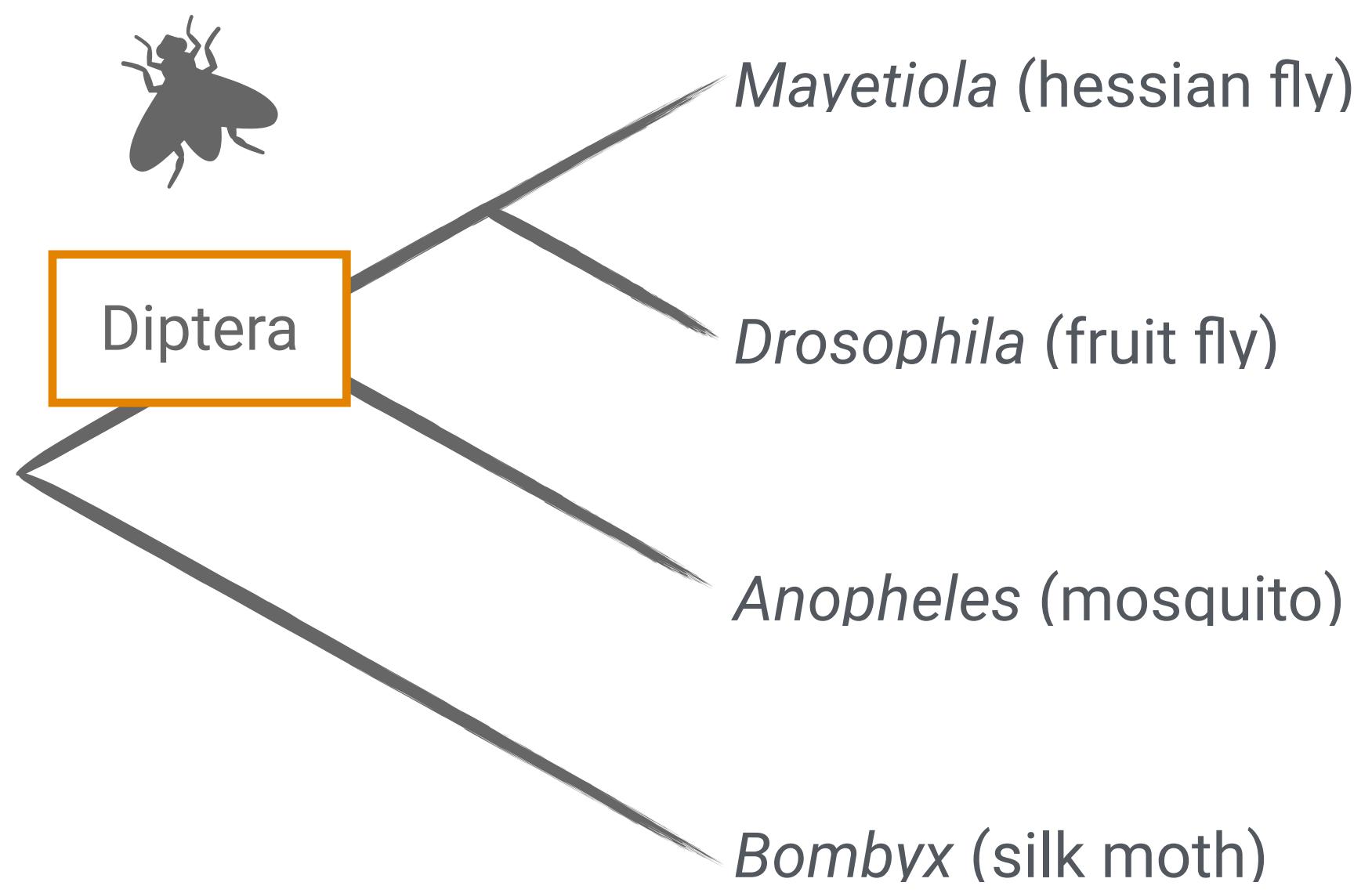
Maximum (all nodes)

- 295.4 Ma
- Boskovice Furrow, Moravia, Czechia
- Huge diversity of insects described from here - no members of even total group Diptera from here or younger deposits

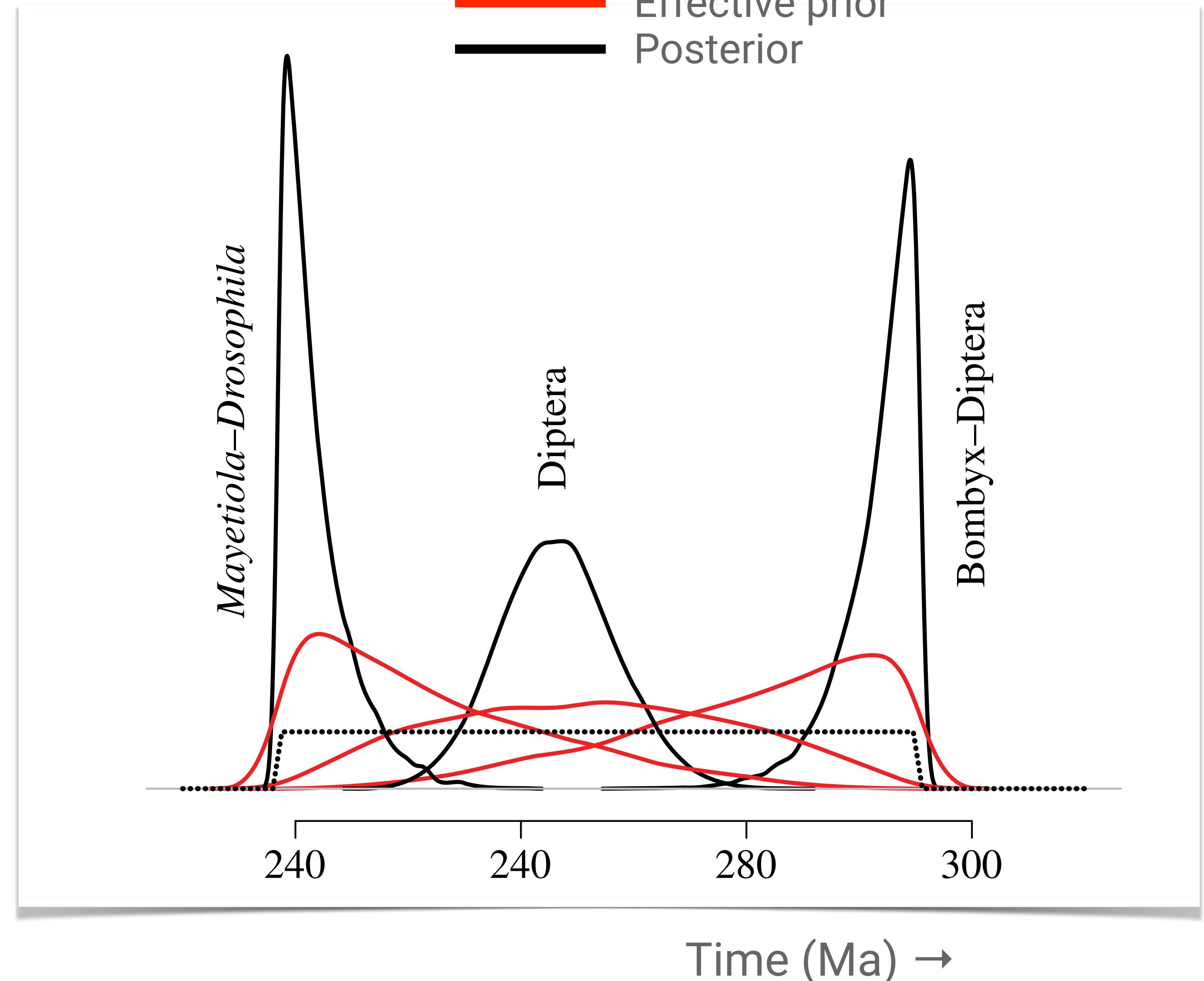
The user “specified priors” will not (always) match the “effective priors” used during analysis

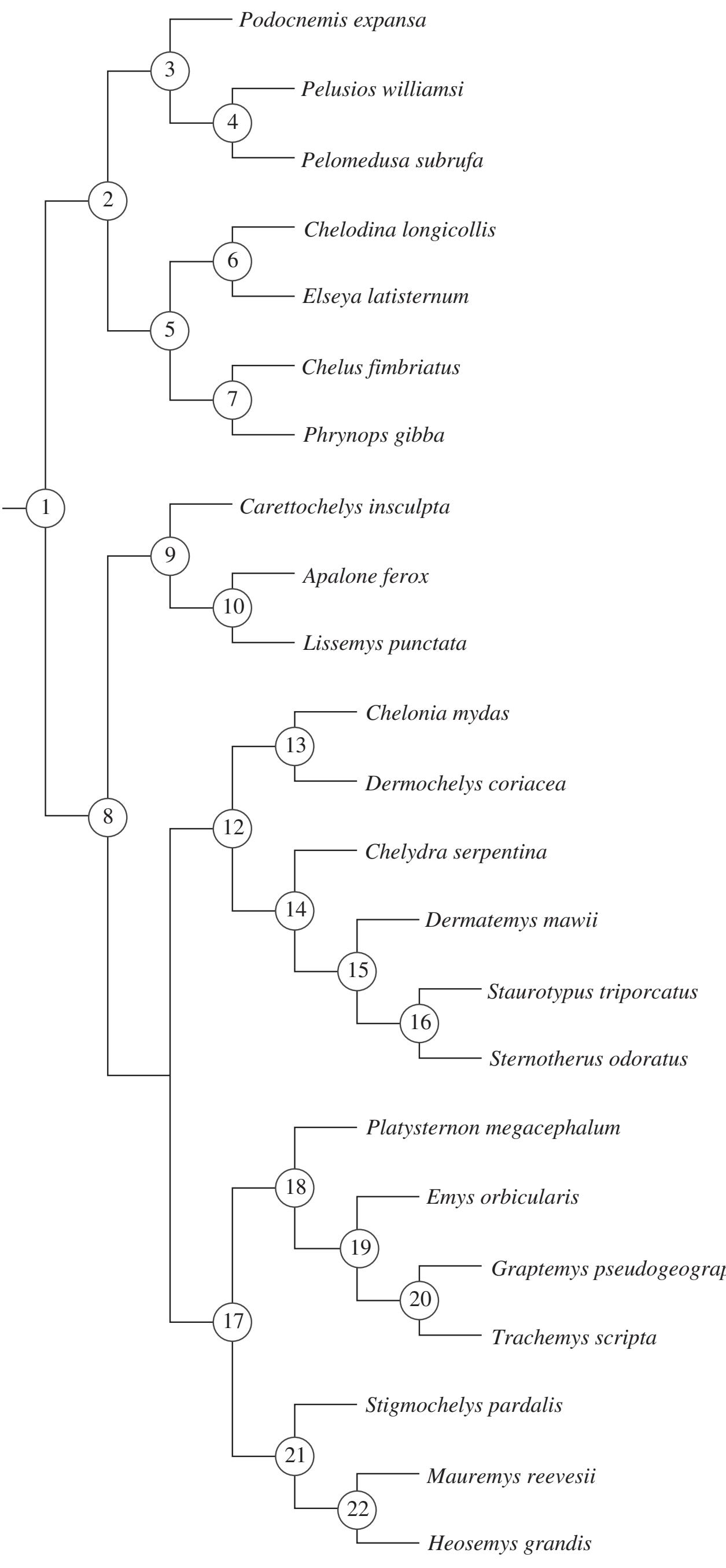
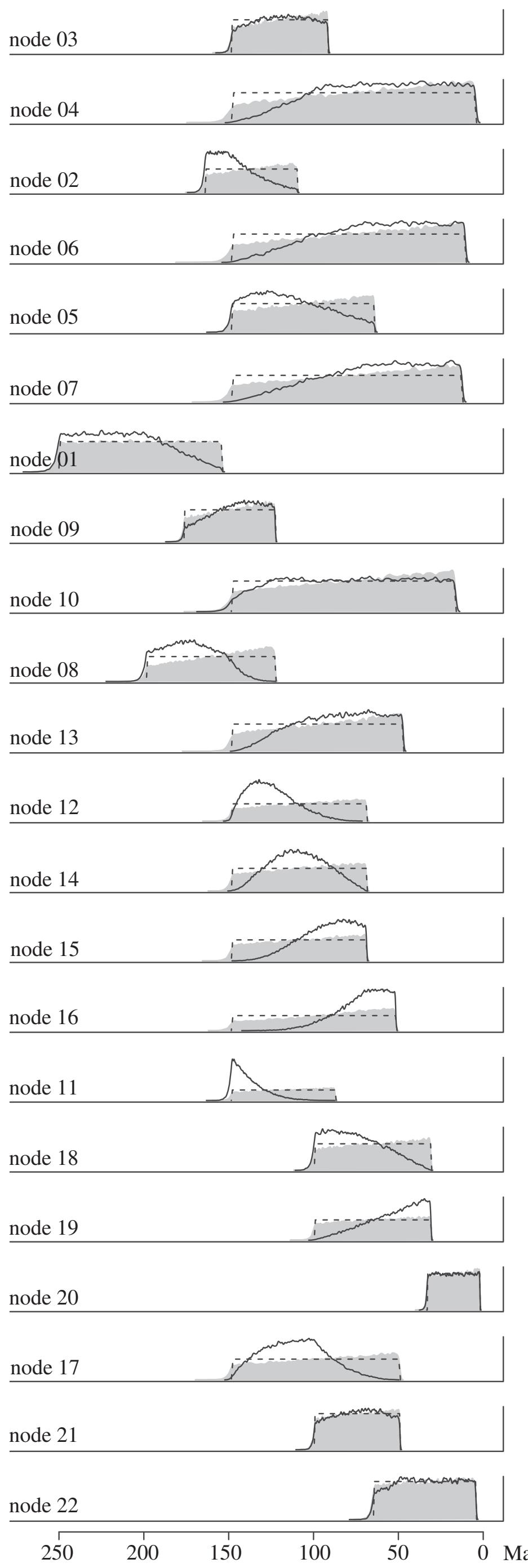


We can run our analysis “under the prior” (we ignore the sequence data)
to see how the software actually constructs the prior density
– accounts for the interaction between nodes



Posteriors look suspiciously like
the effective priors....





An example from turtles



Dashed lines = the specified priors

Grey shaded area = effective priors
for one calibration, analysed alone

Black shaded area = effective
priors for all calibrations together

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

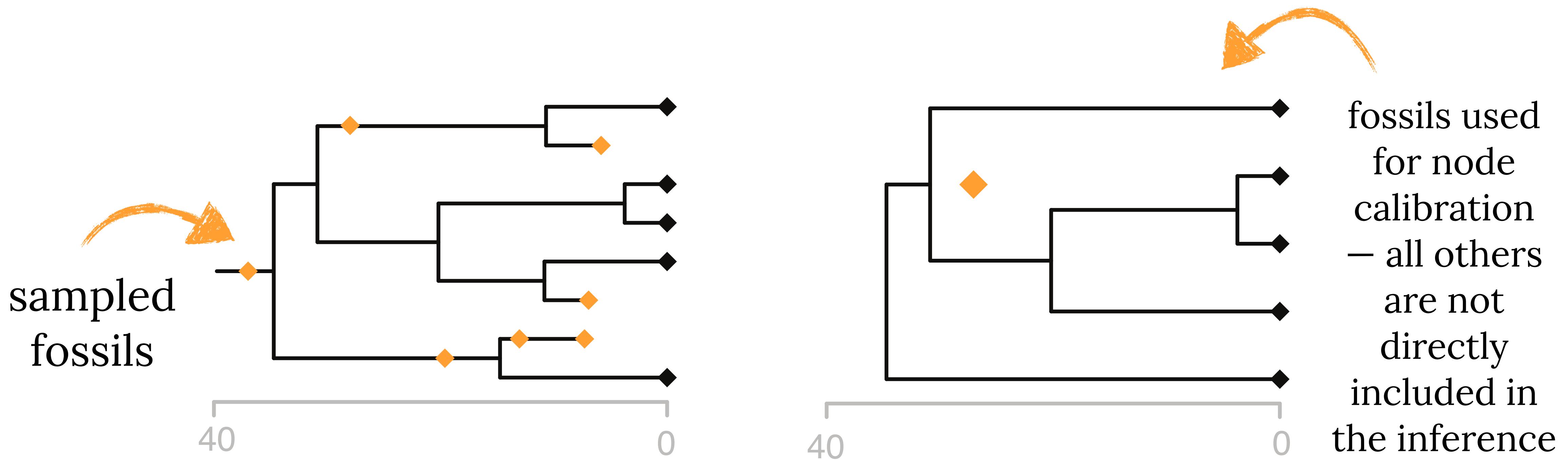
Heled and Drummond. [2012](#). Sys Bio

Warnock et al. [2012, 2015](#)

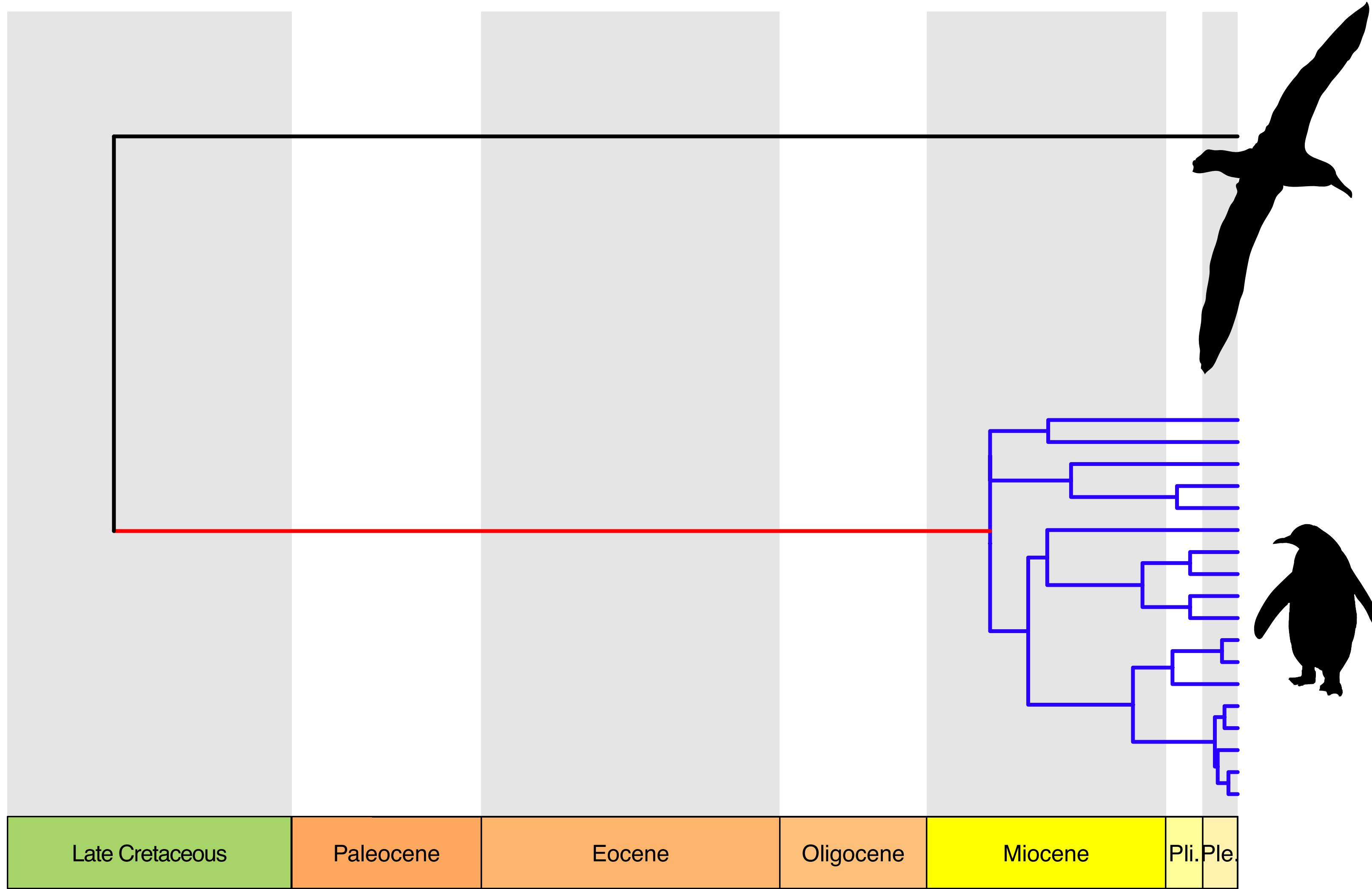
Node dating: potential issues

There are many!

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

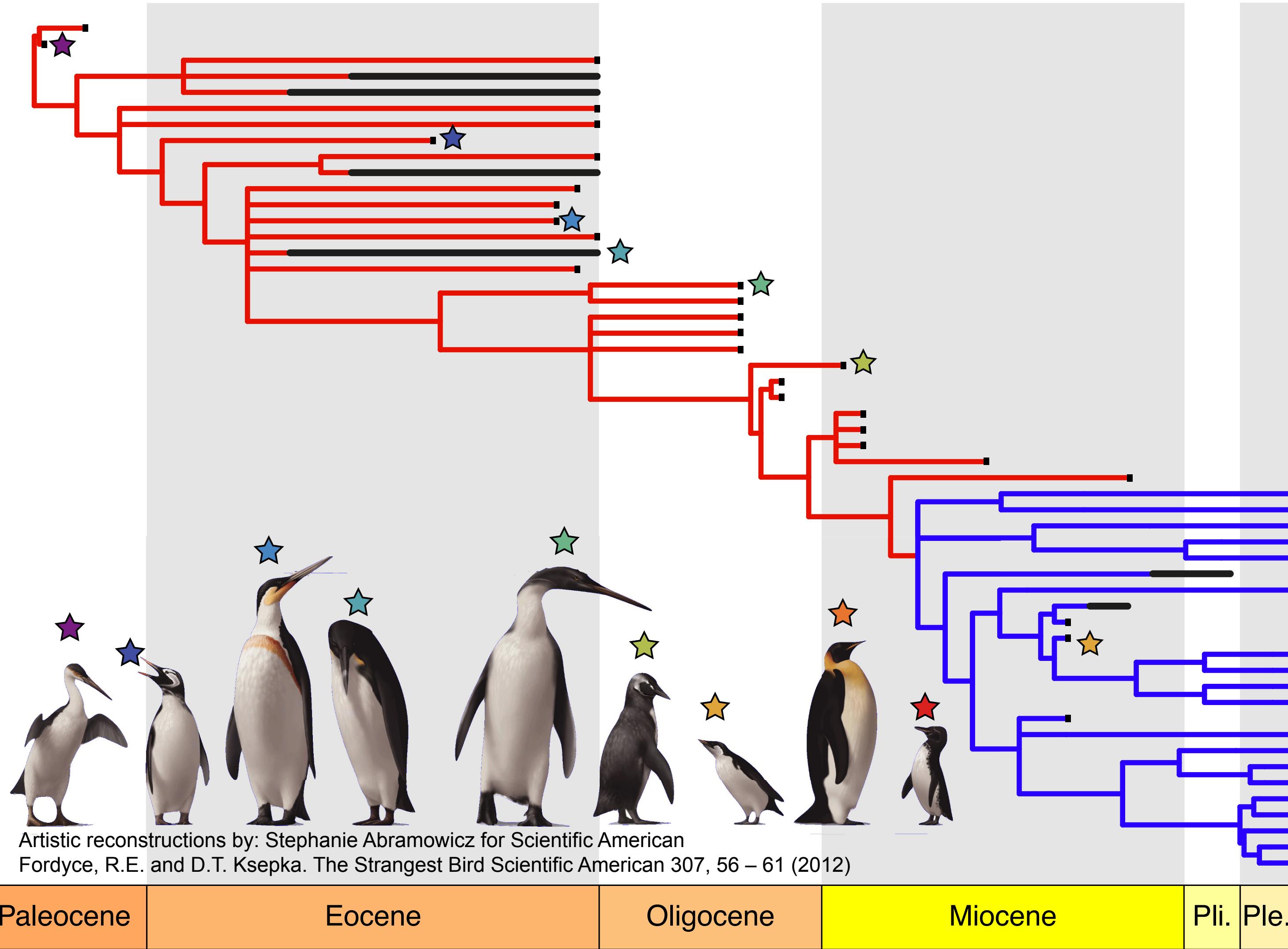


Example: living penguins



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

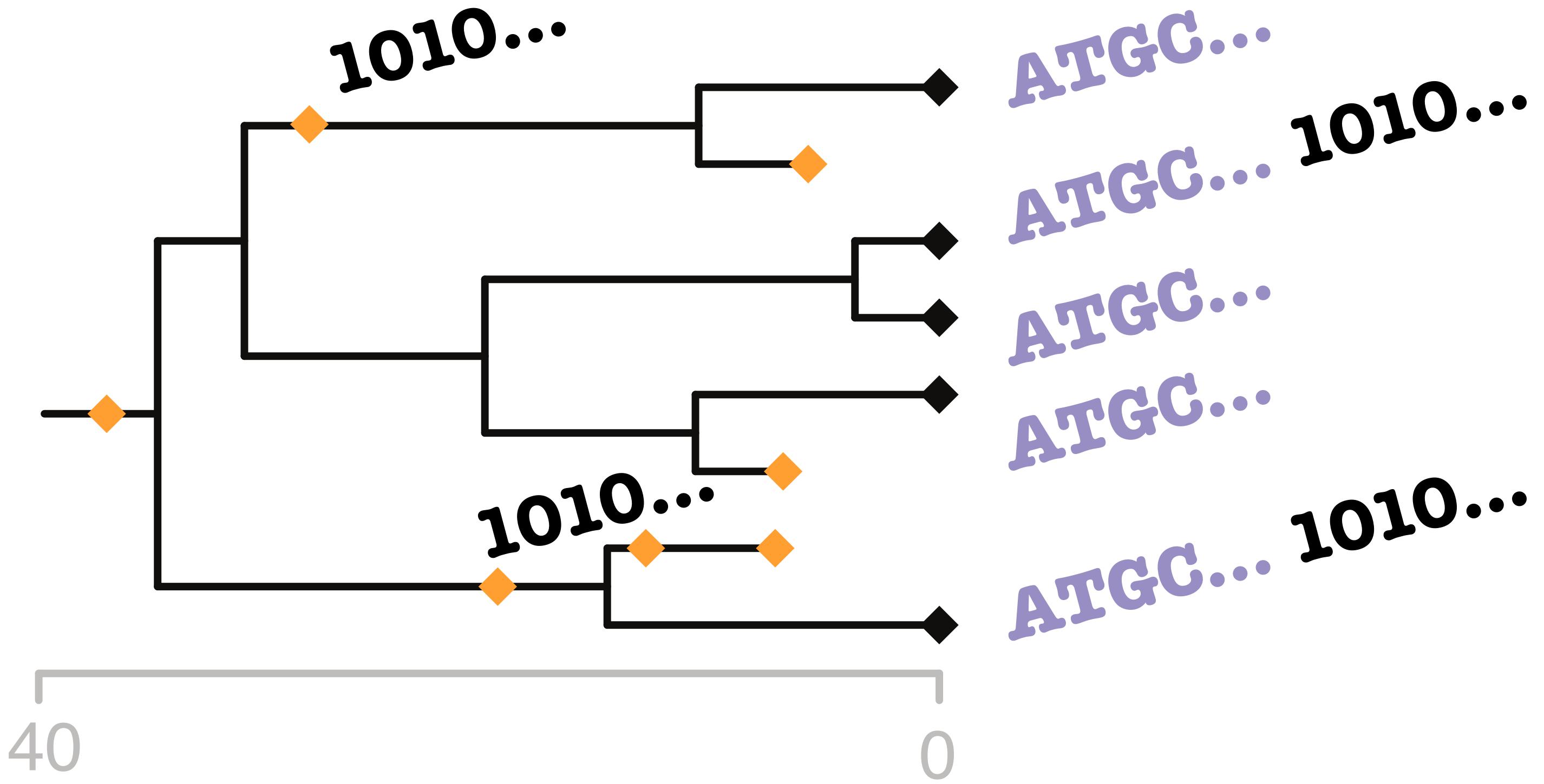
Example: living penguins



But penguins
have a rich
fossil record!

Total-evidence dating

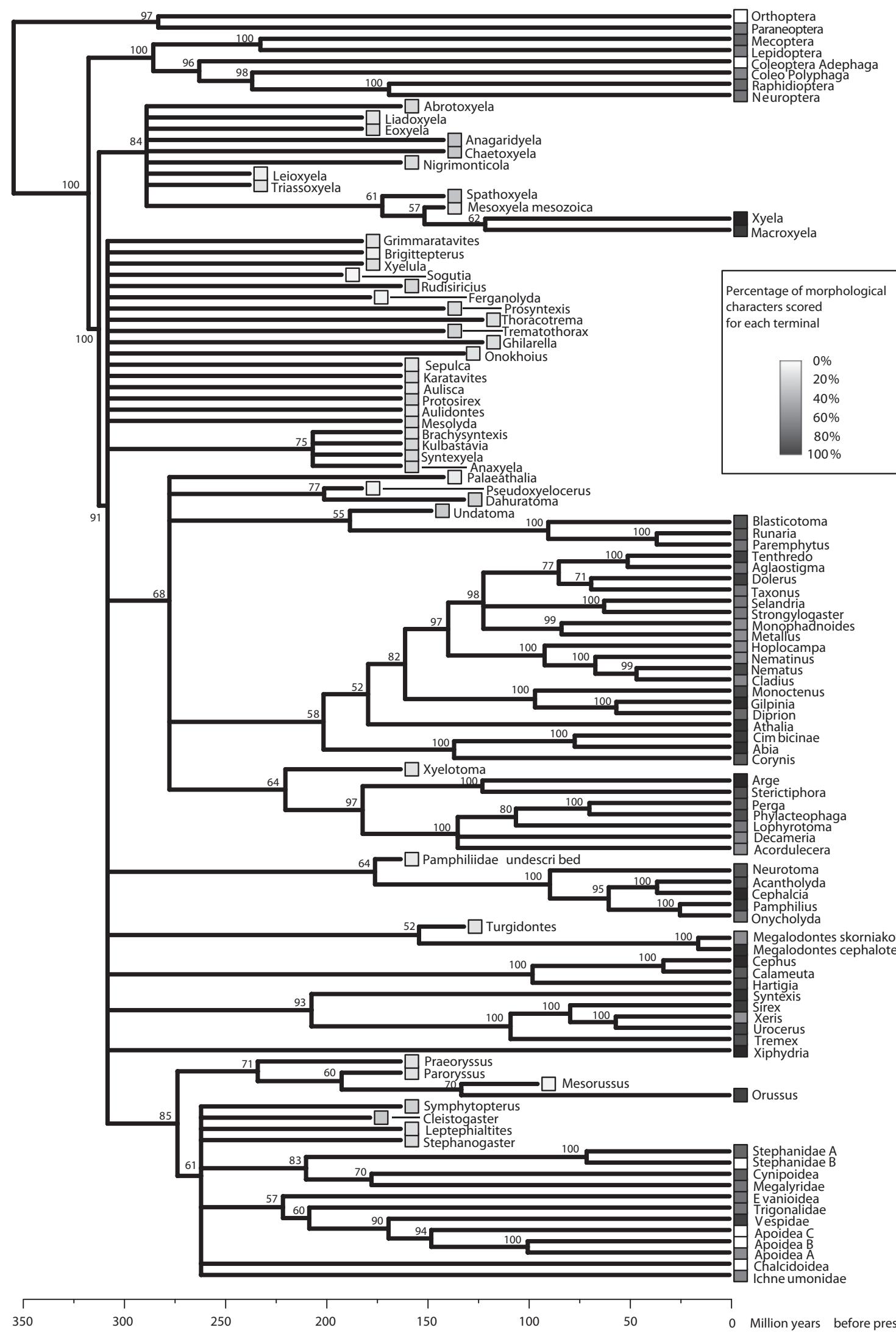
Tip-dating or “total-evidence” dating



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

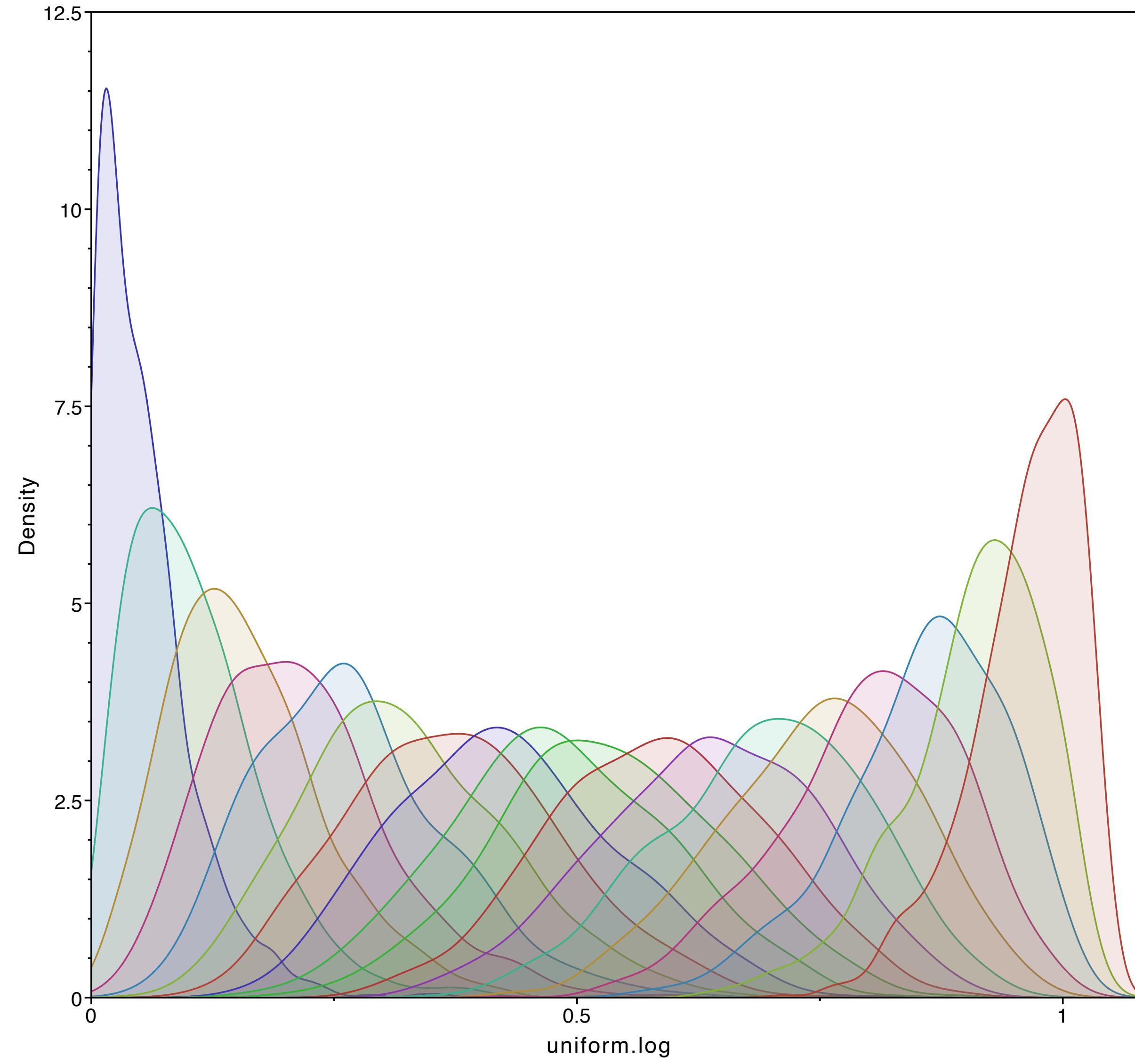


Dated tree of Hymenoptera



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



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

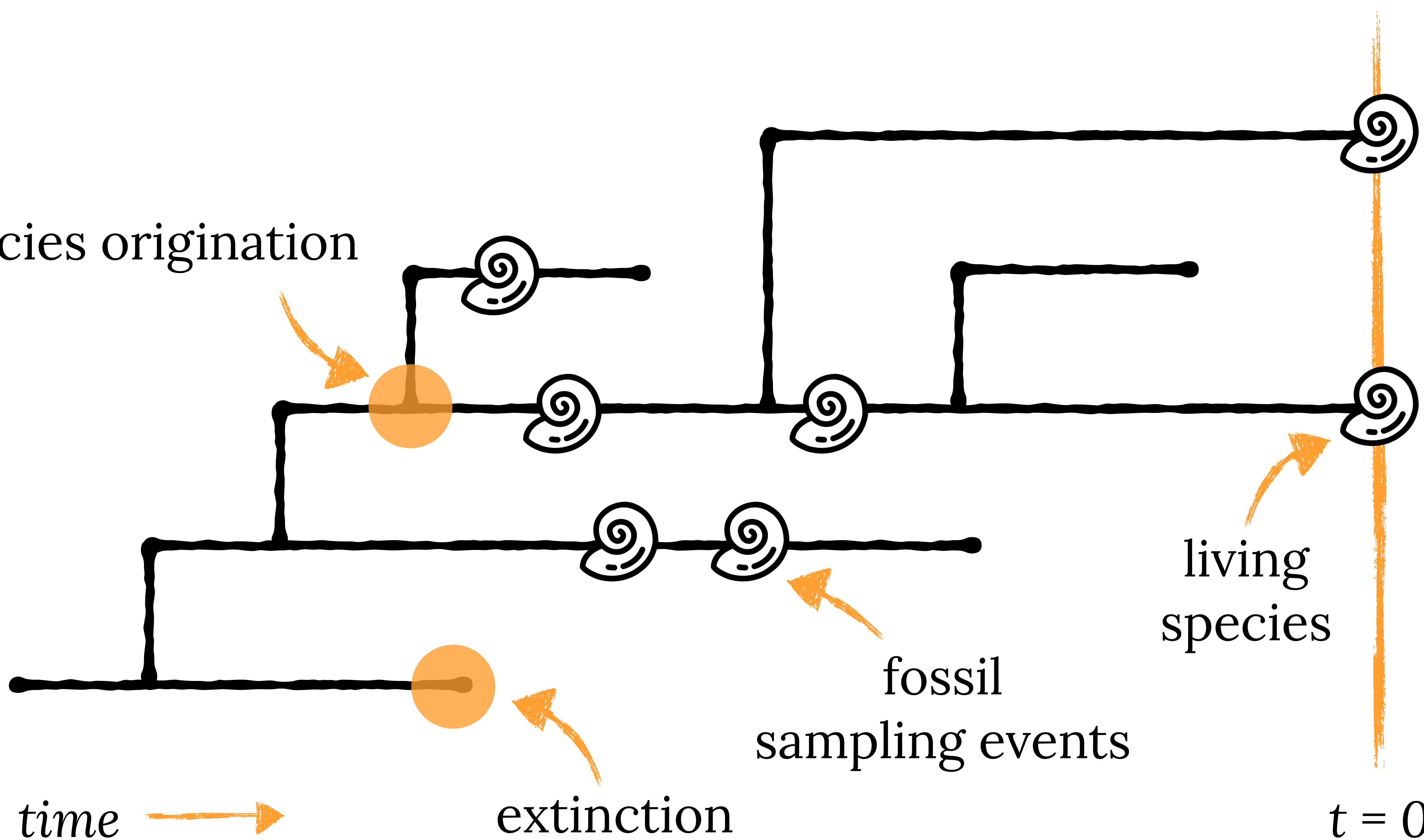
[Uniform tree priors - why not use them?](#)

Remco Bouckaert

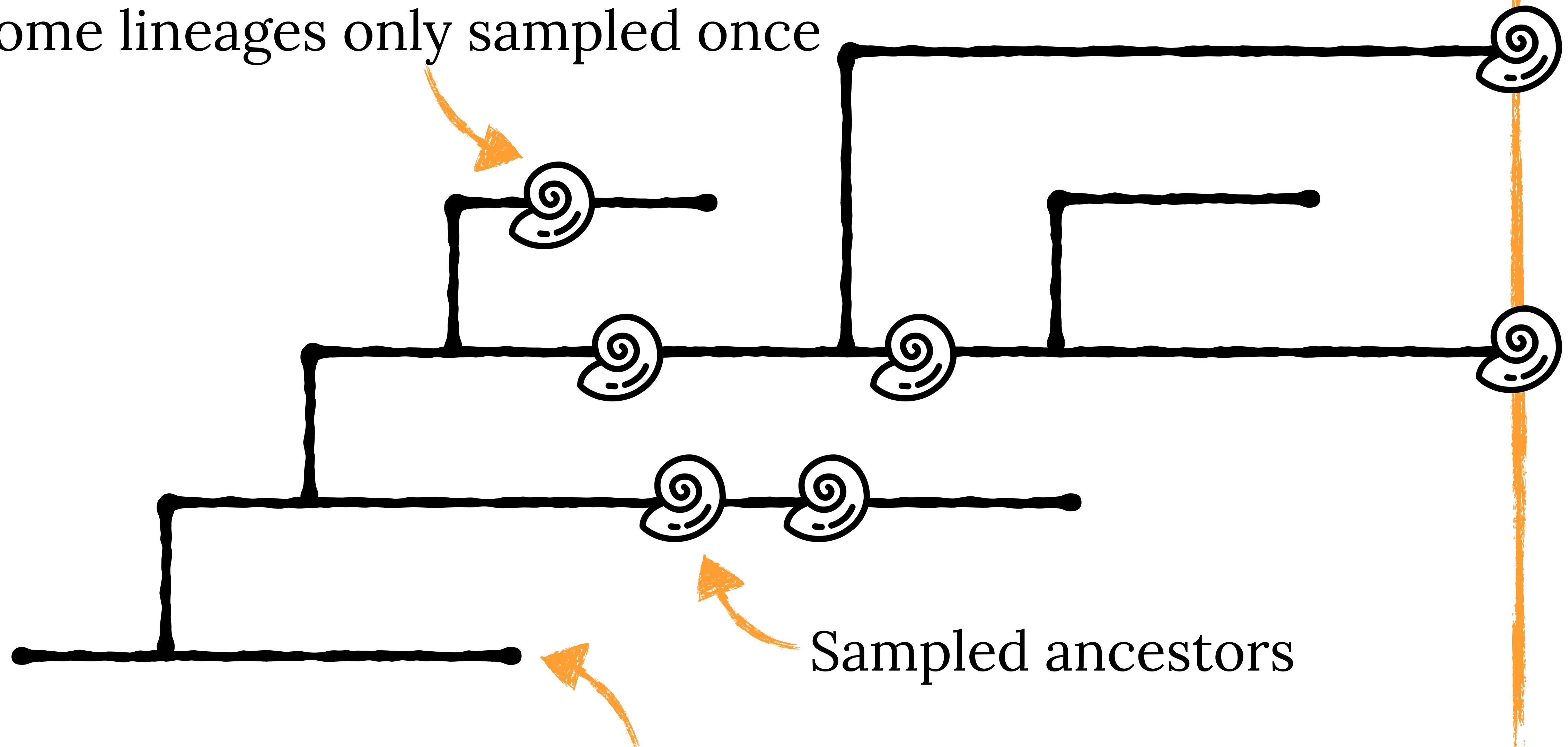
What does a generating prior for the fossil record look like?

The fossilised birth-death process

species origination



Some lineages only sampled once

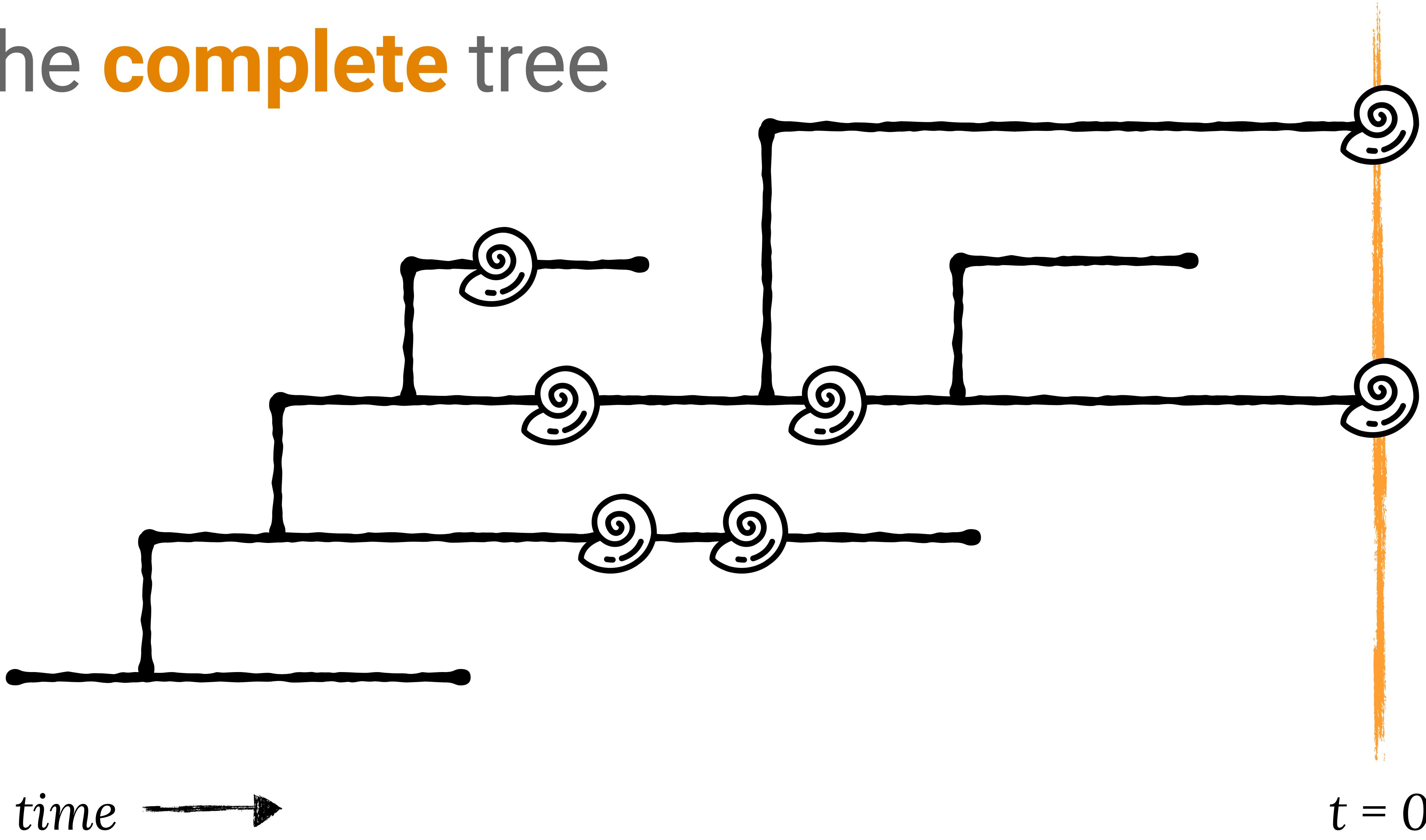


Sampled ancestors

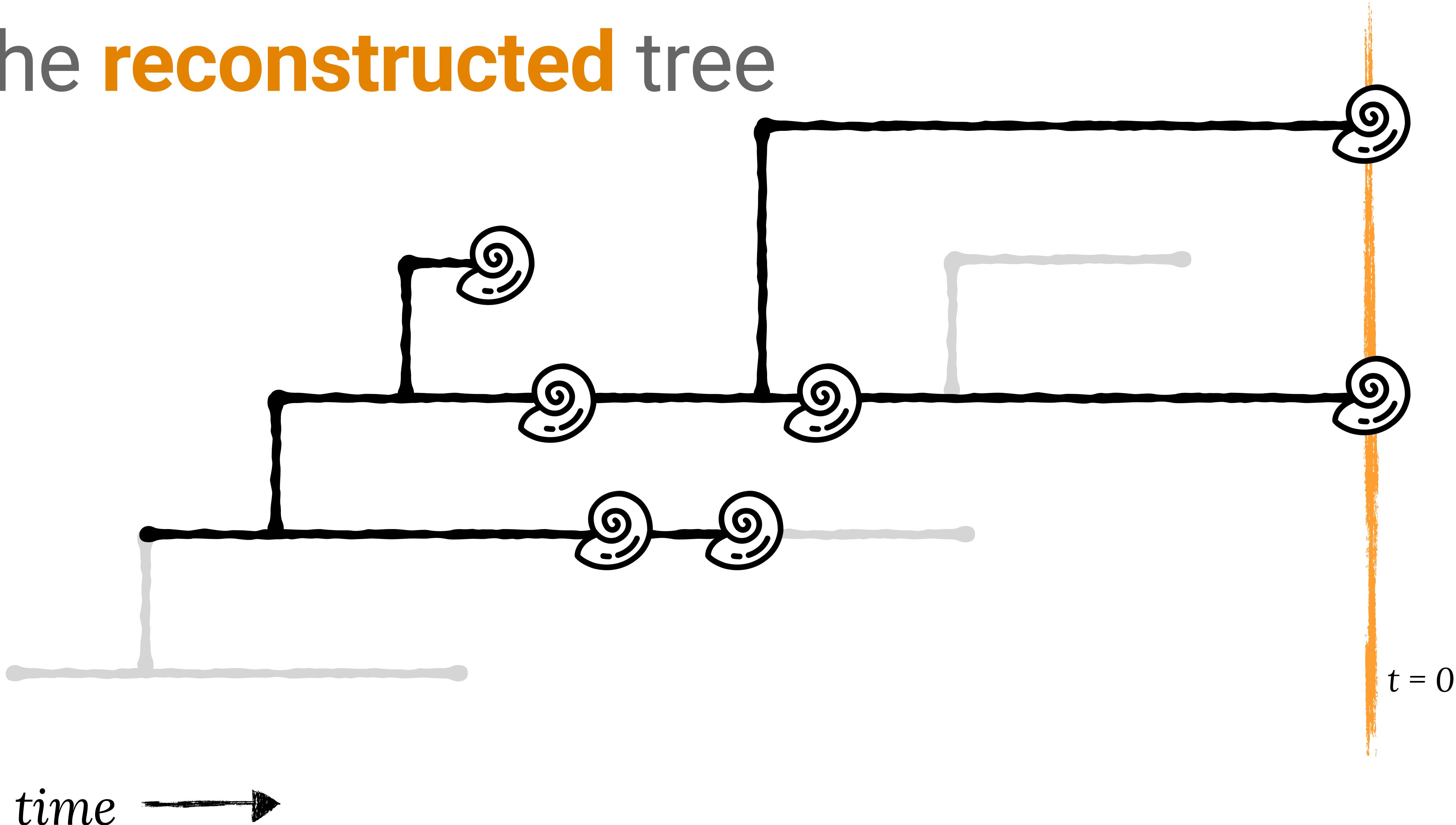
Some lineages go completely unsampled

$t = 0$

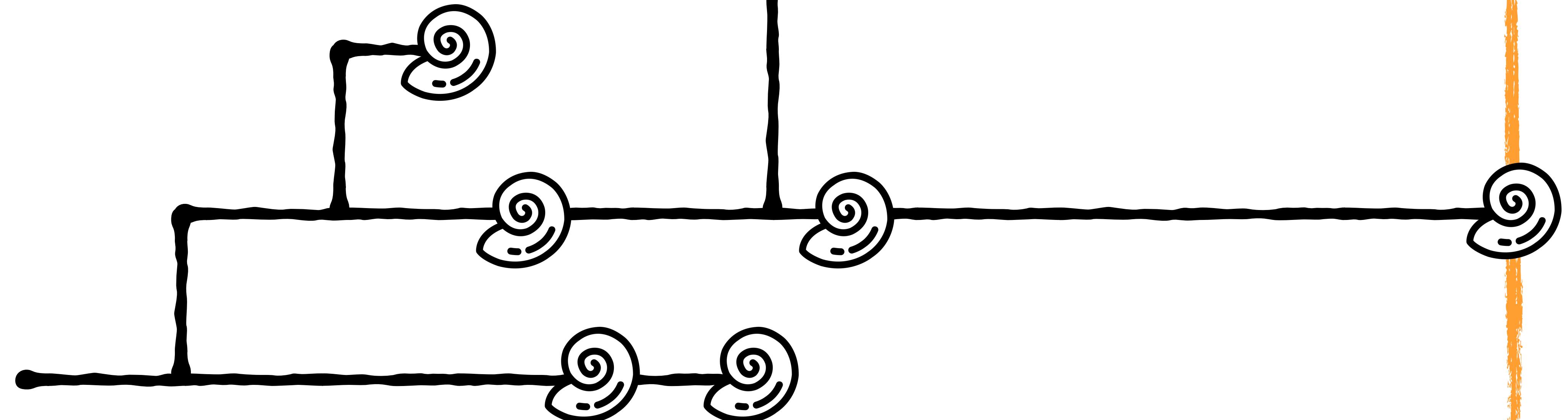
The complete tree



The reconstructed tree



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

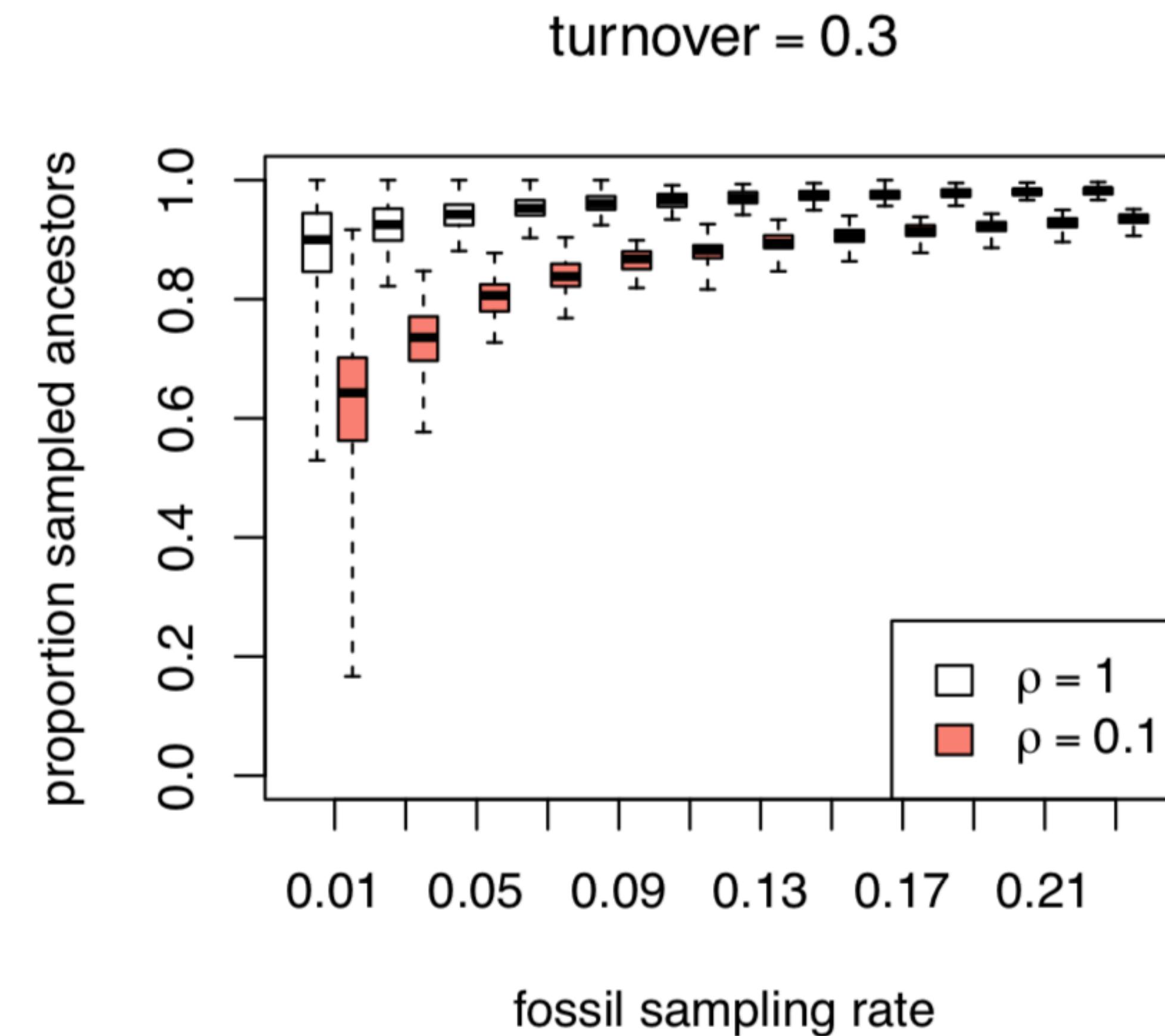
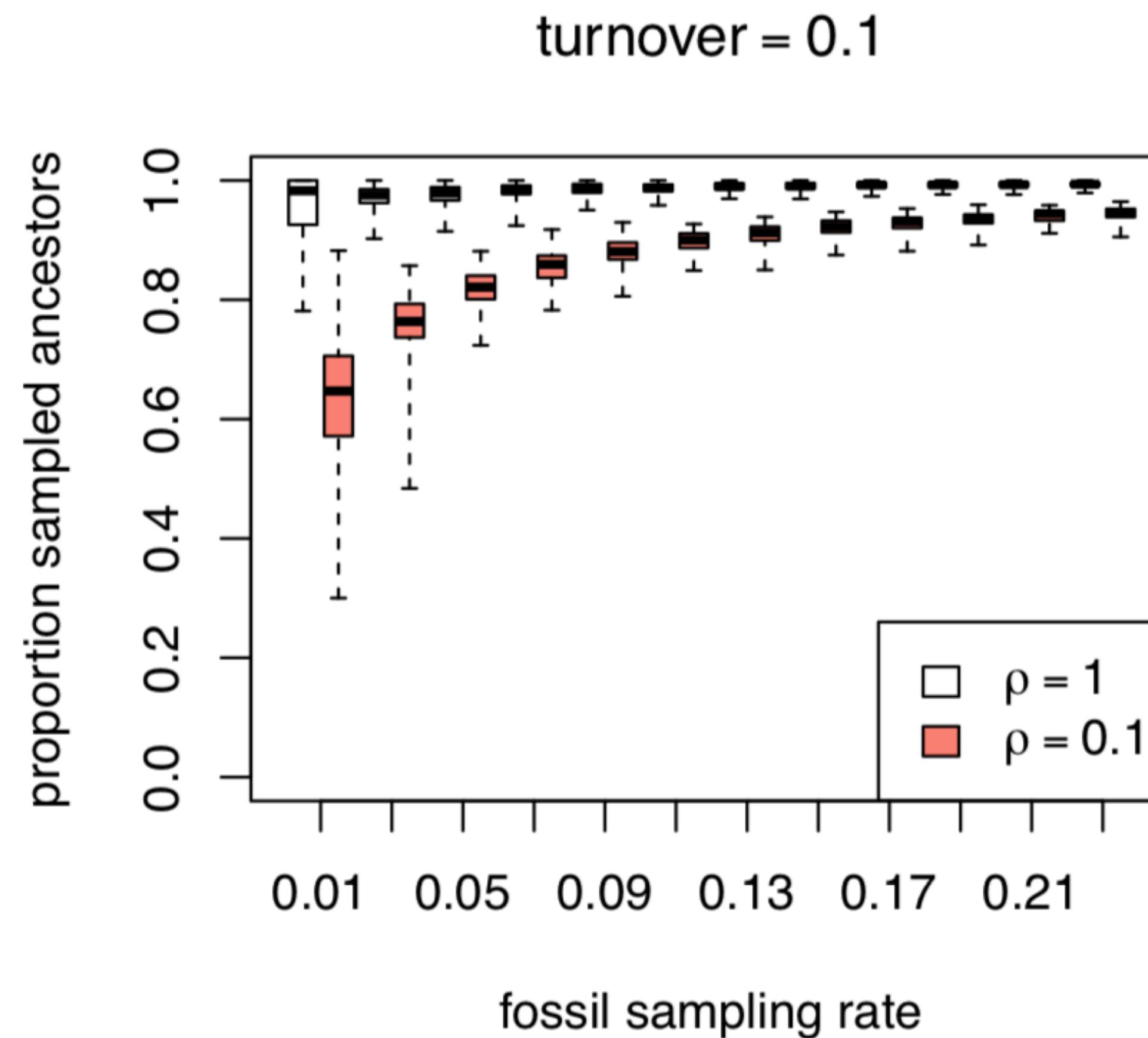


$$P(E | \text{snail}, \lambda, \mu, \psi, \rho)$$

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

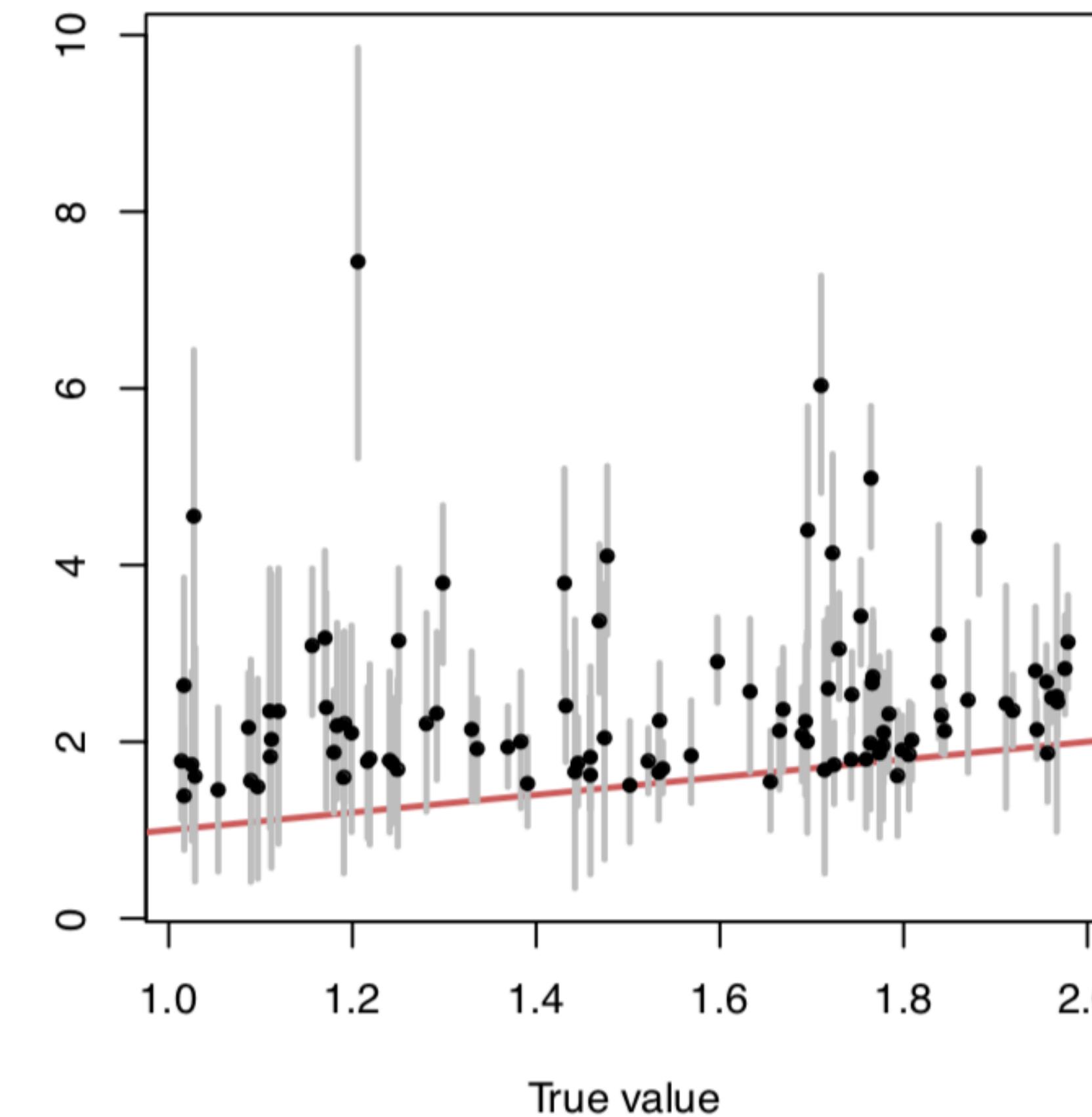
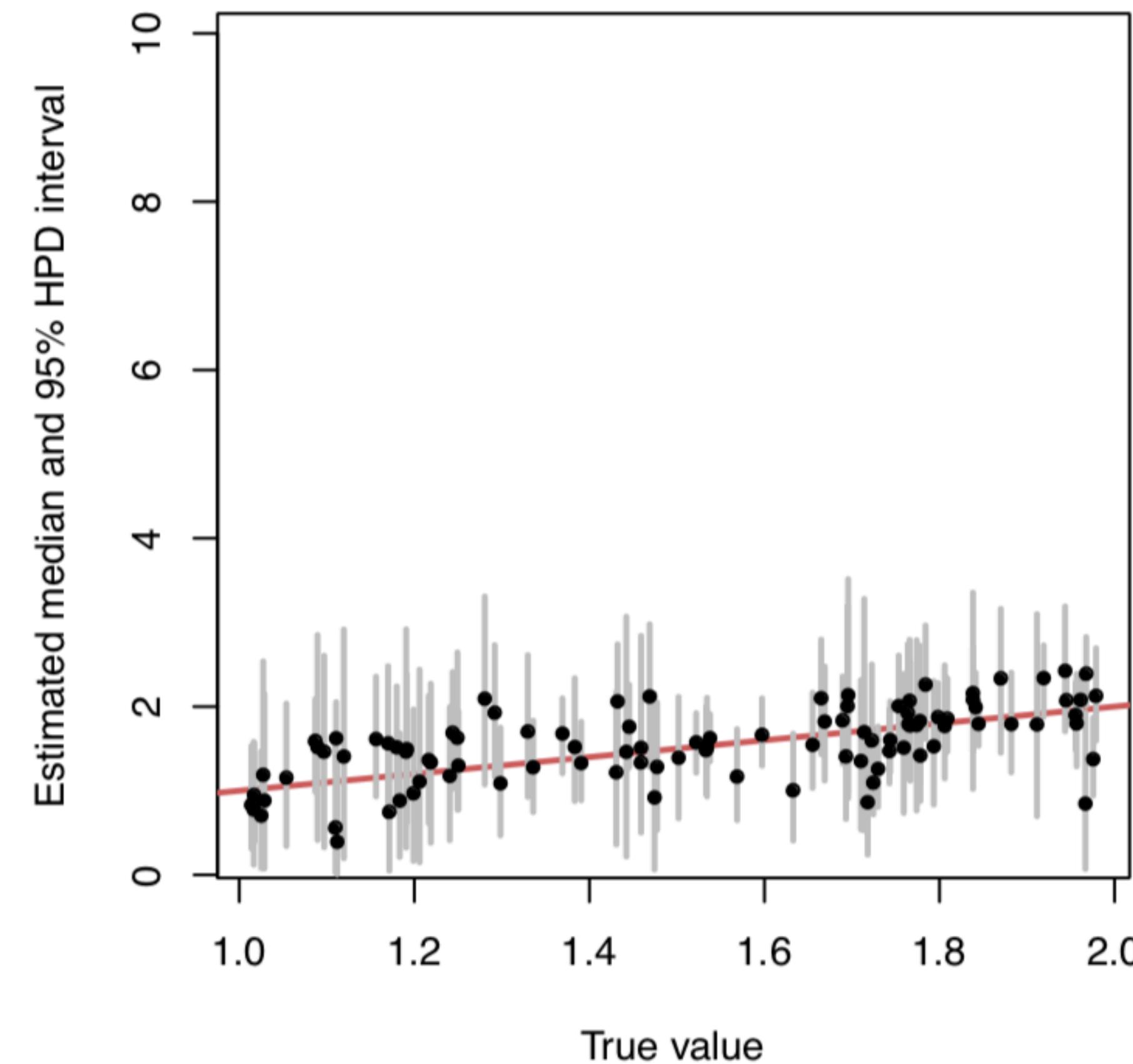
Sampled ancestors

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

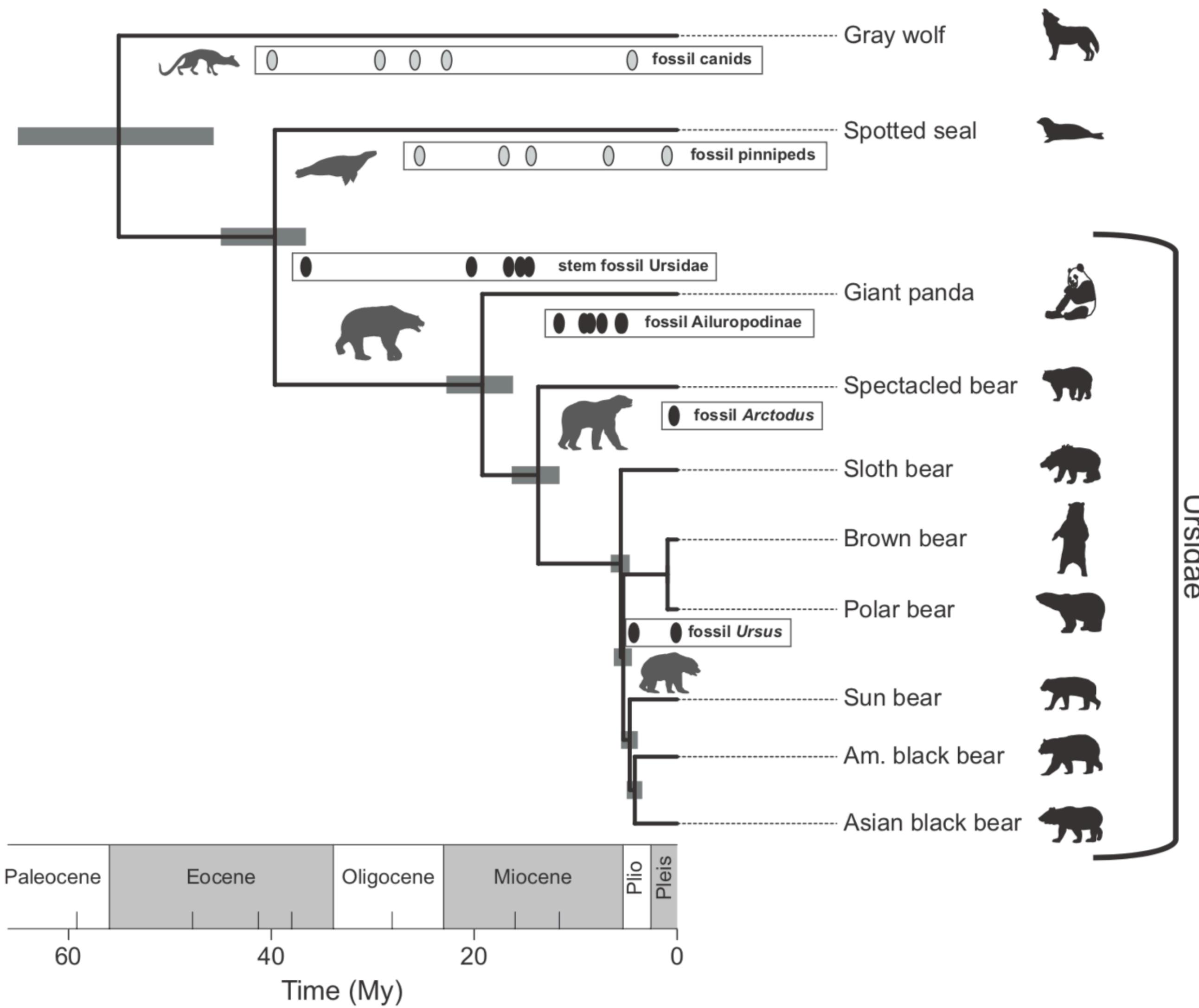


Sampled ancestors

Ignoring sampled ancestors can lead to inaccurate parameter estimates



Time calibrated tree of living and fossil bears

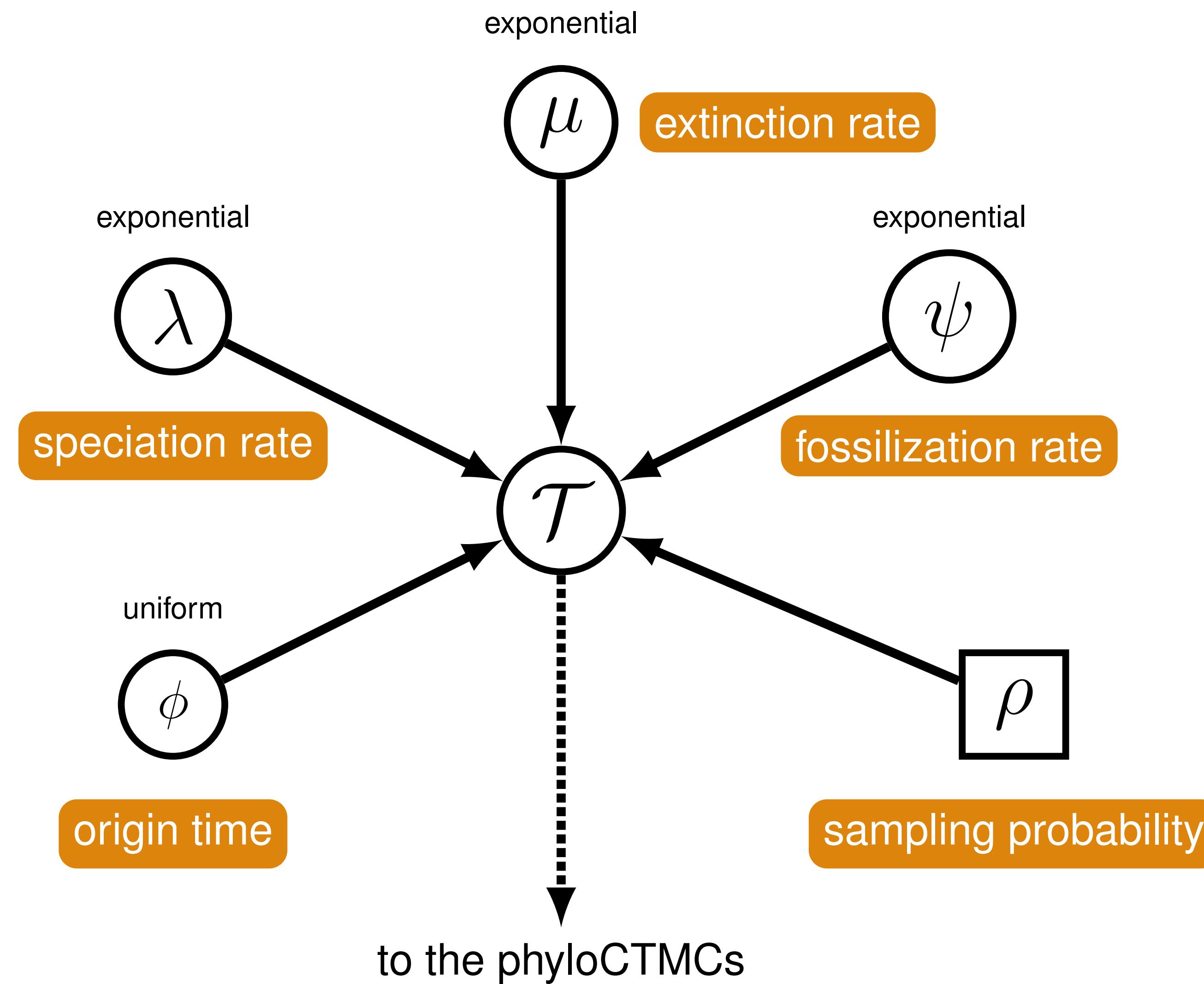


First application of the FBD model.

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

Exercise

Graphical model representation of the FBDP



Relationship to (some) other birth-death process models

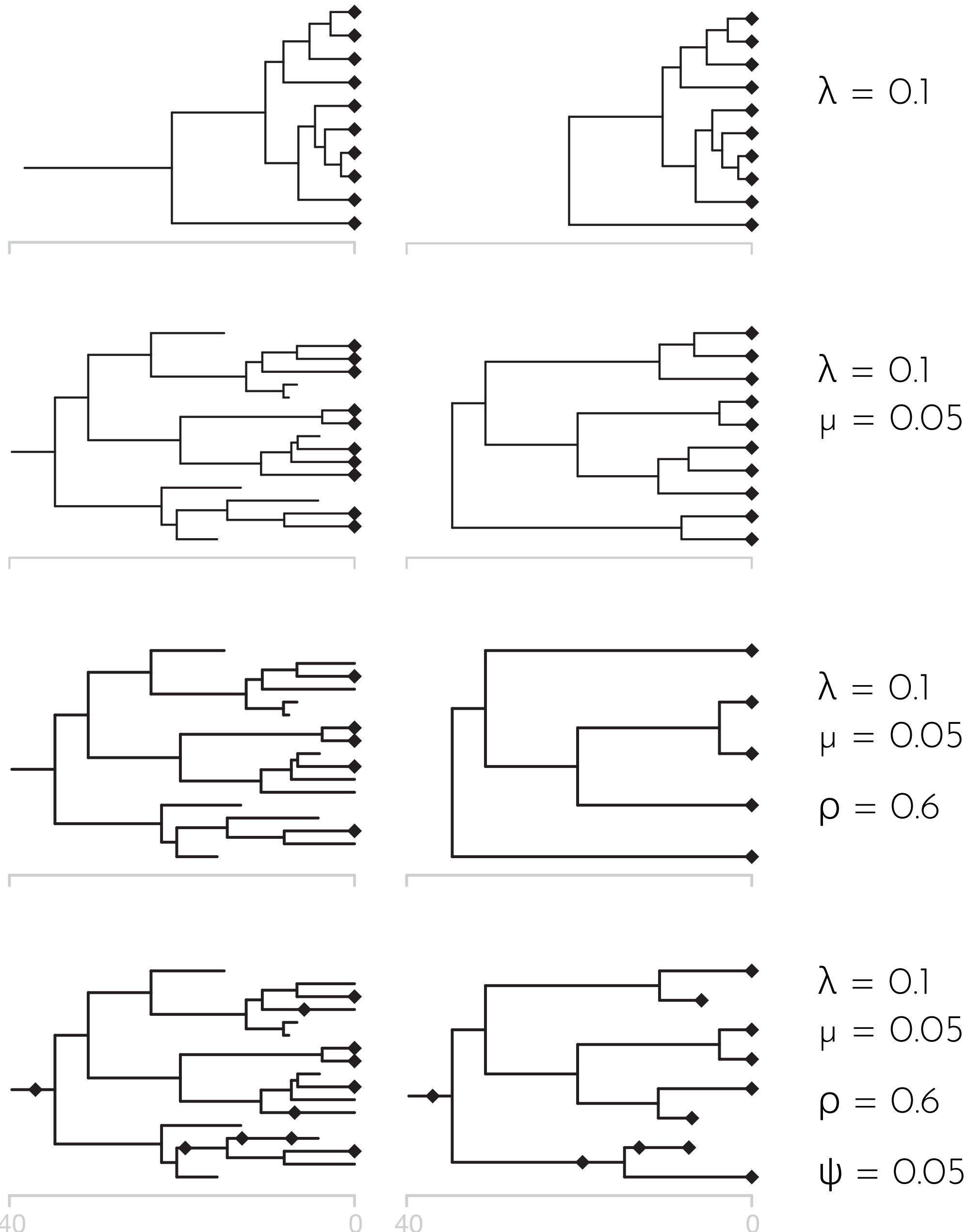
These models are special cases of the FBD process, with fossil sampling (ψ) = zero.

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



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

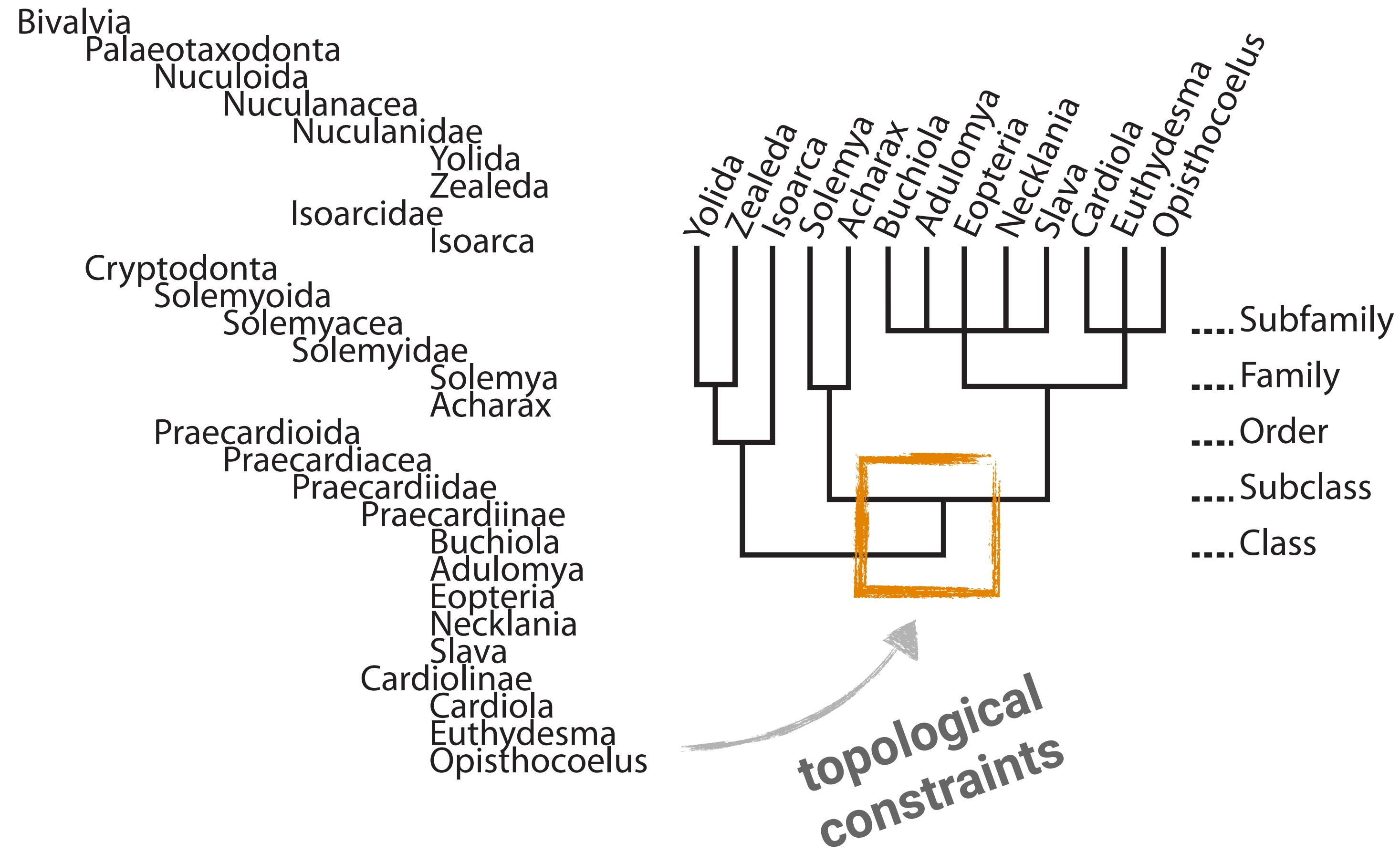
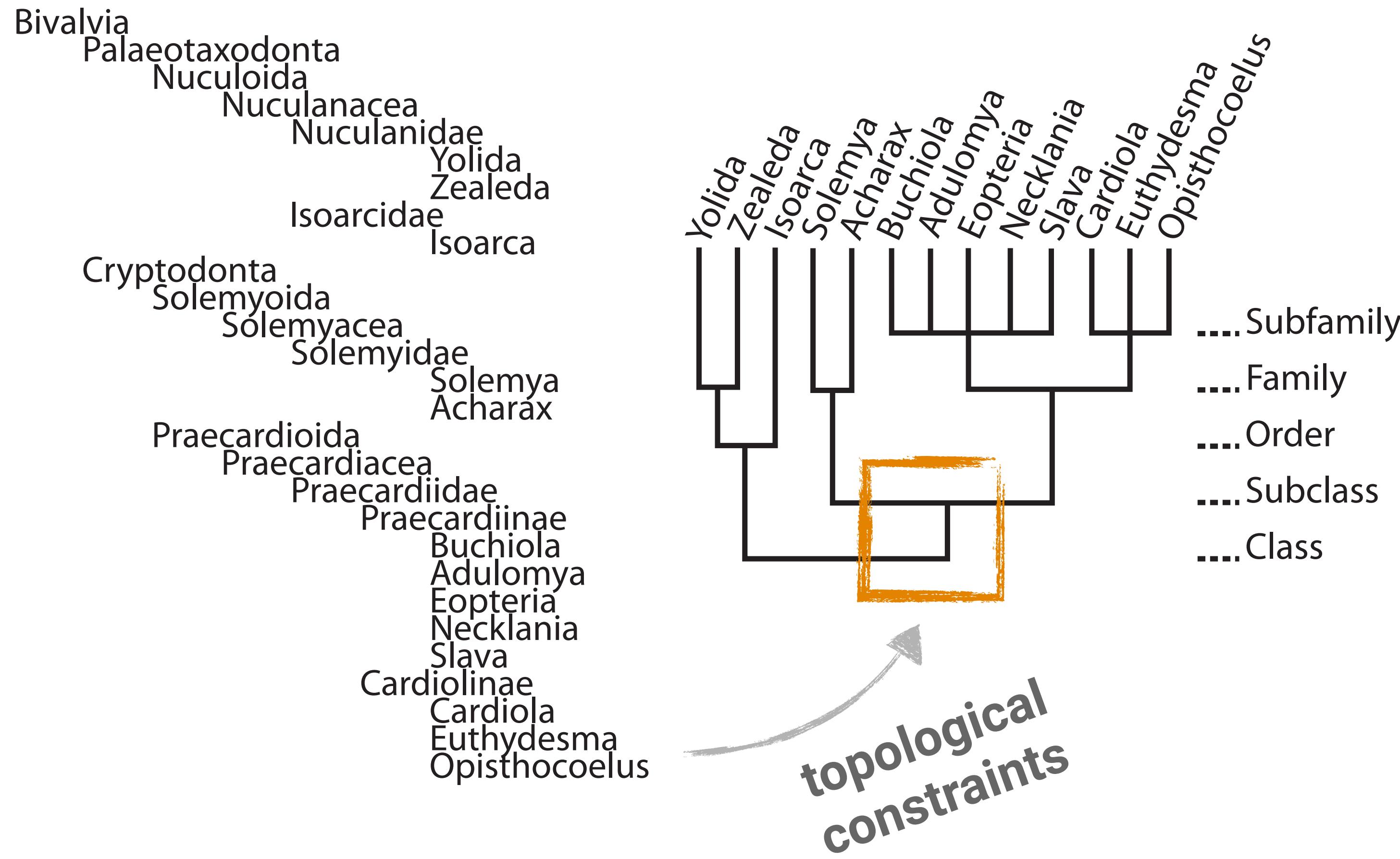


Image source Soul & Friedman (2015)

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



ATAT...

TCACT...

?????...

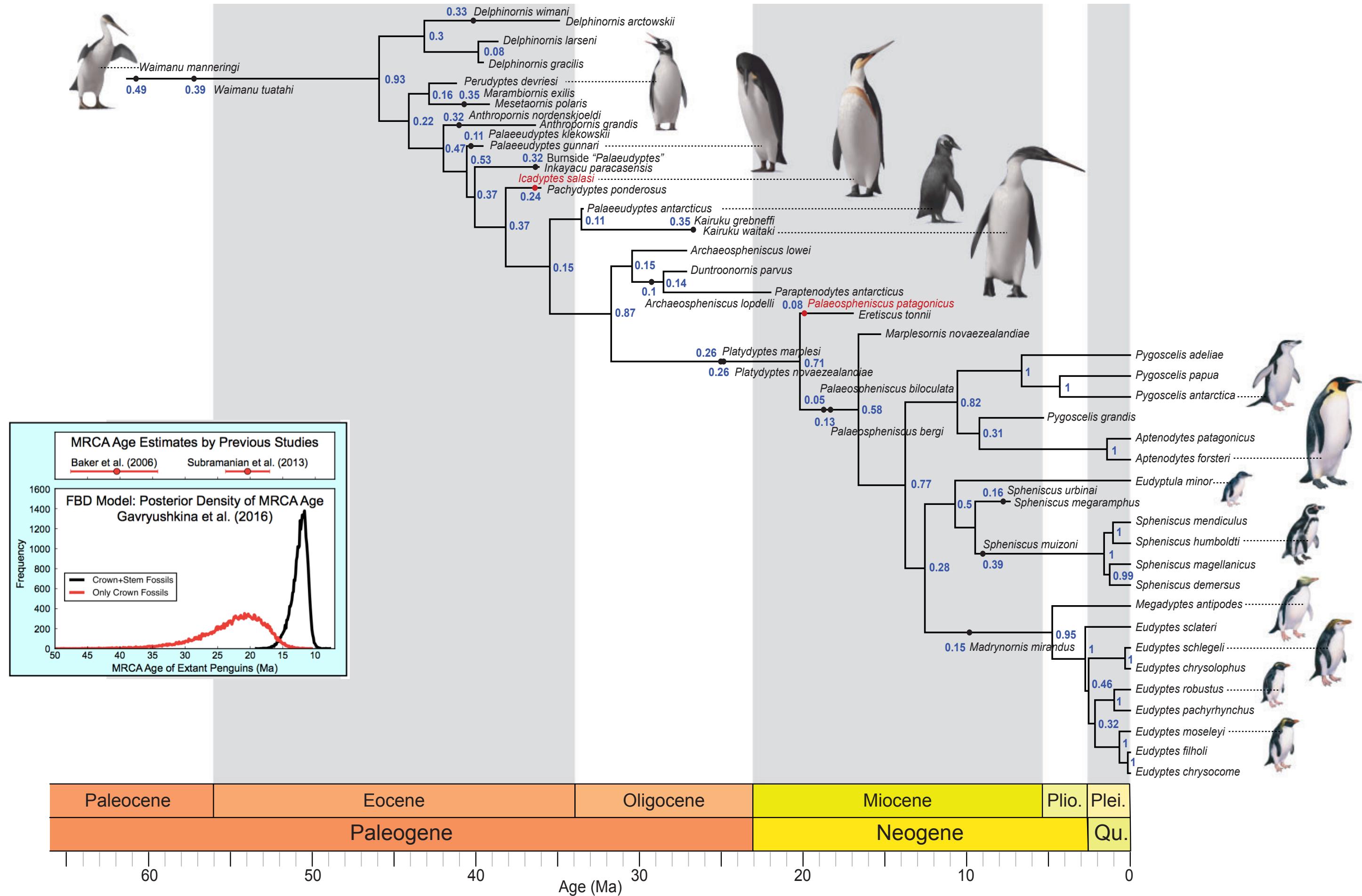
OR

1001...

1101...

0100...

Time calibrated tree of living and fossil penguins



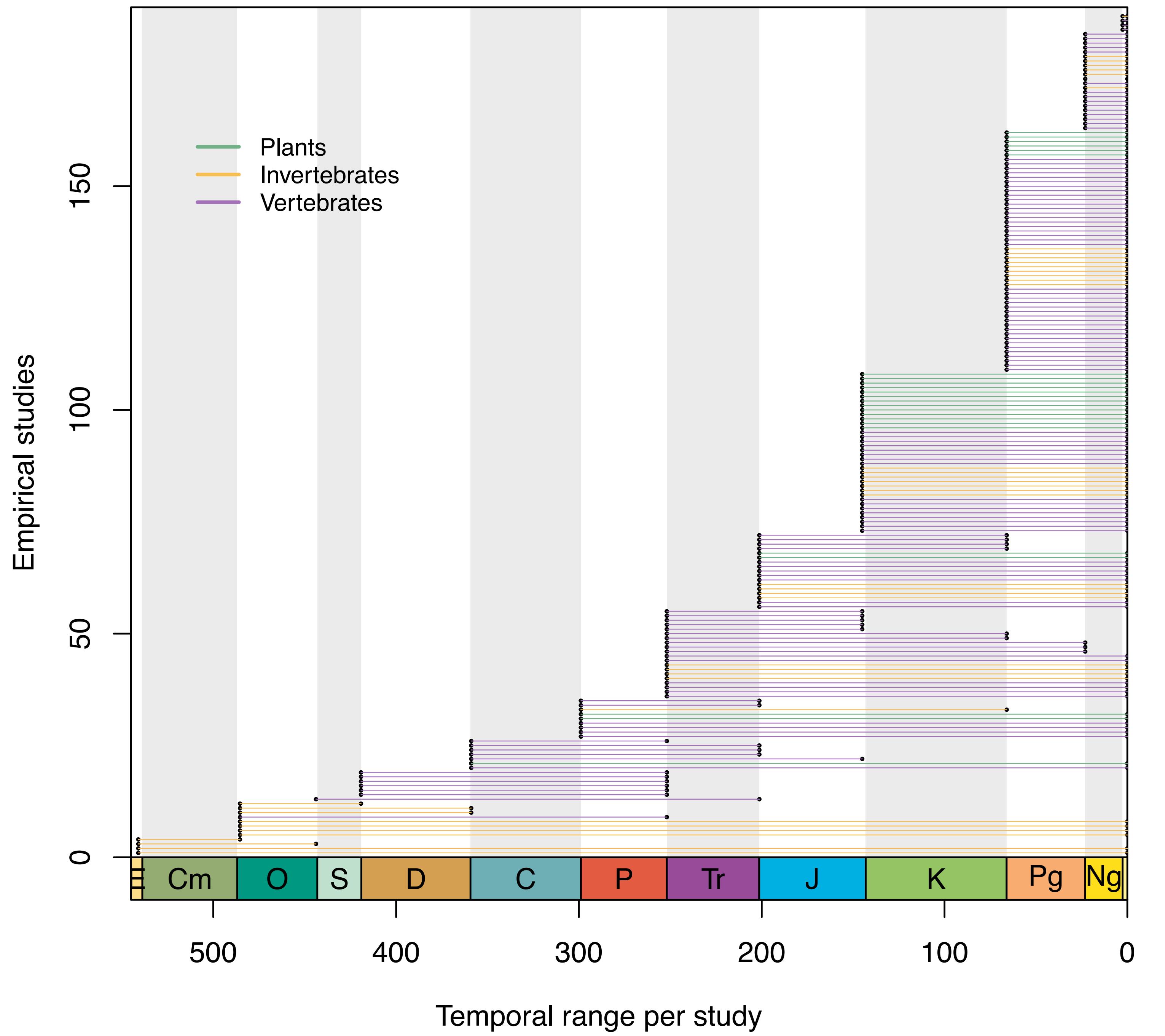
First application of total evidence dating using the FBD model

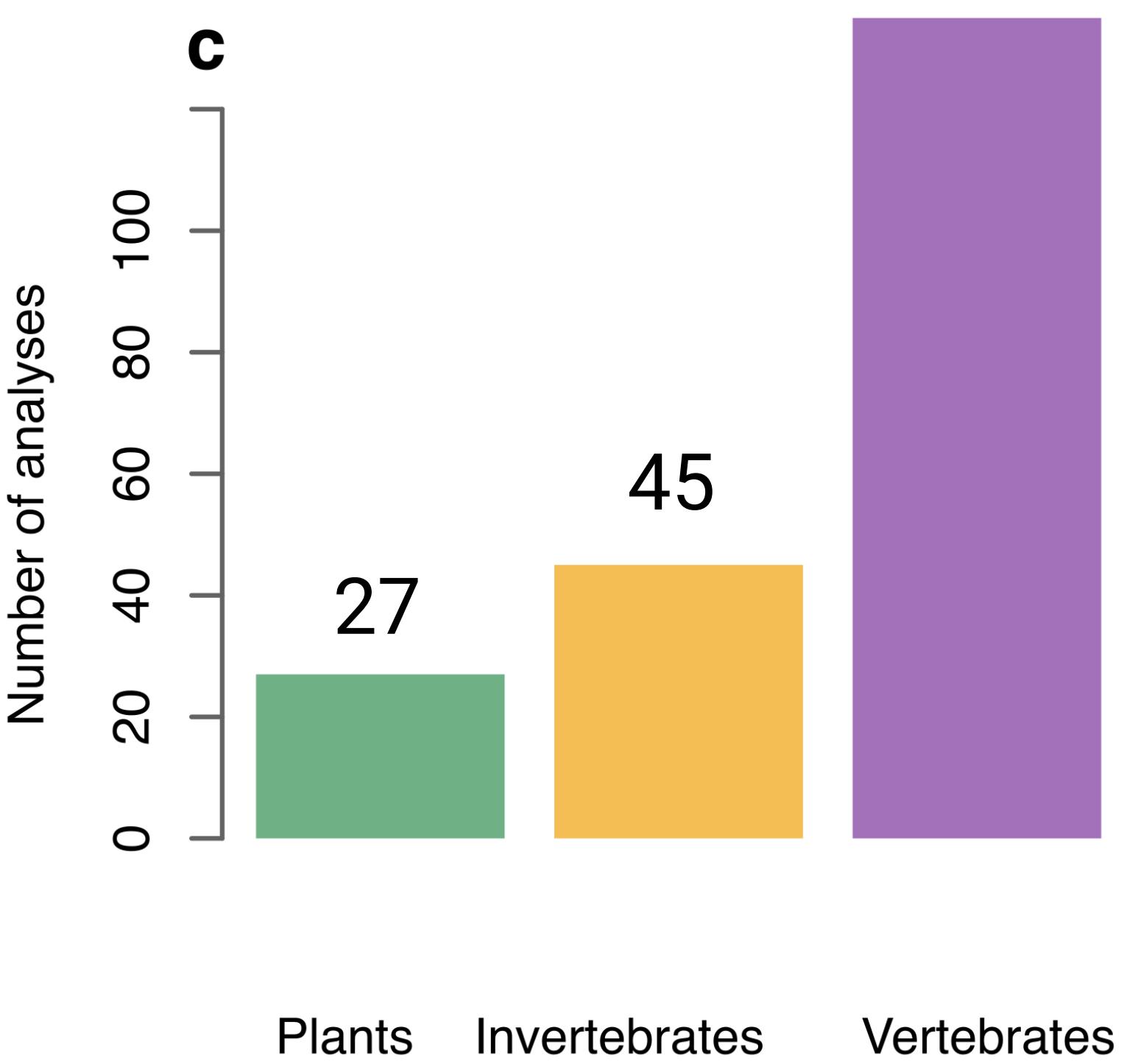
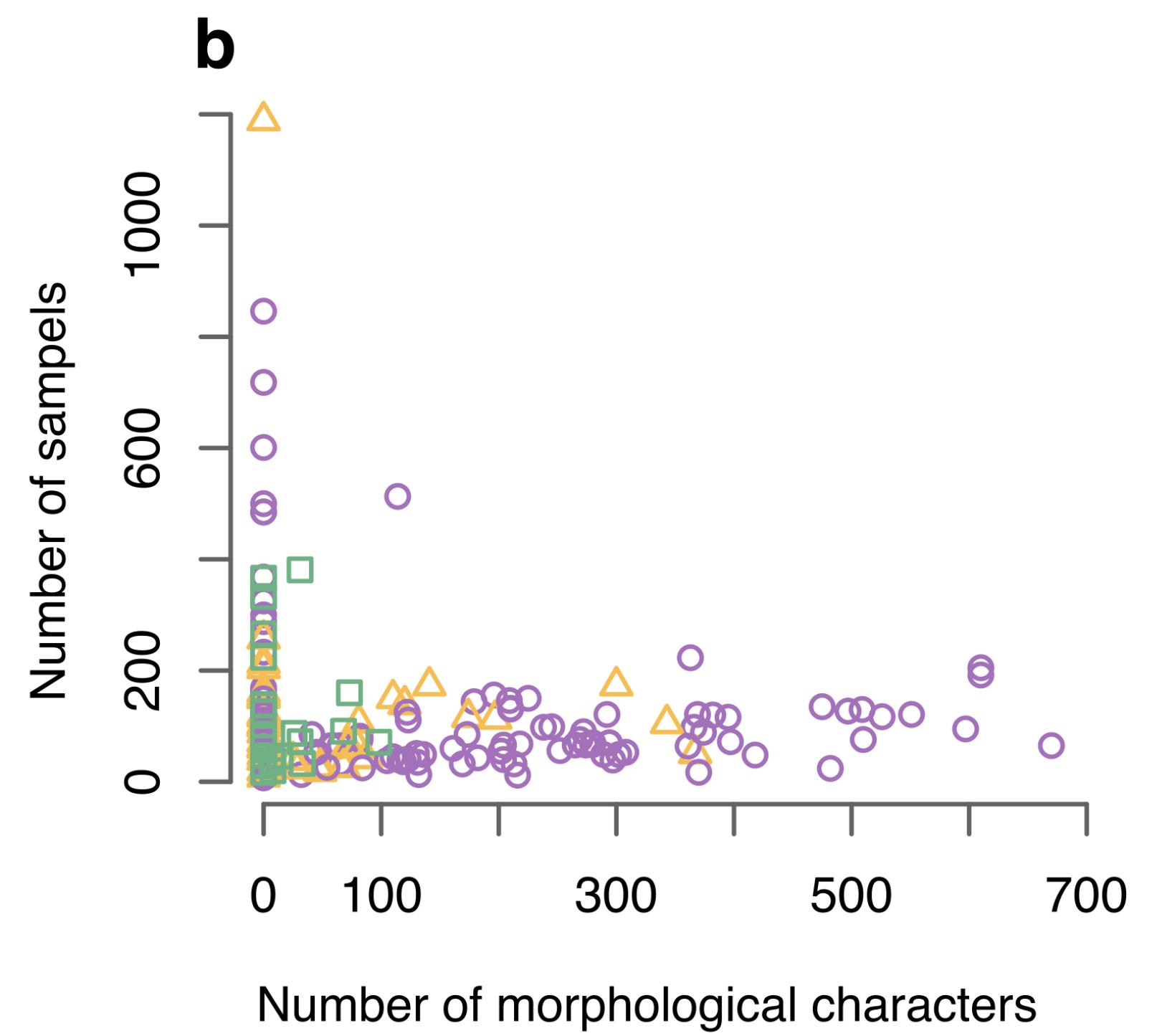
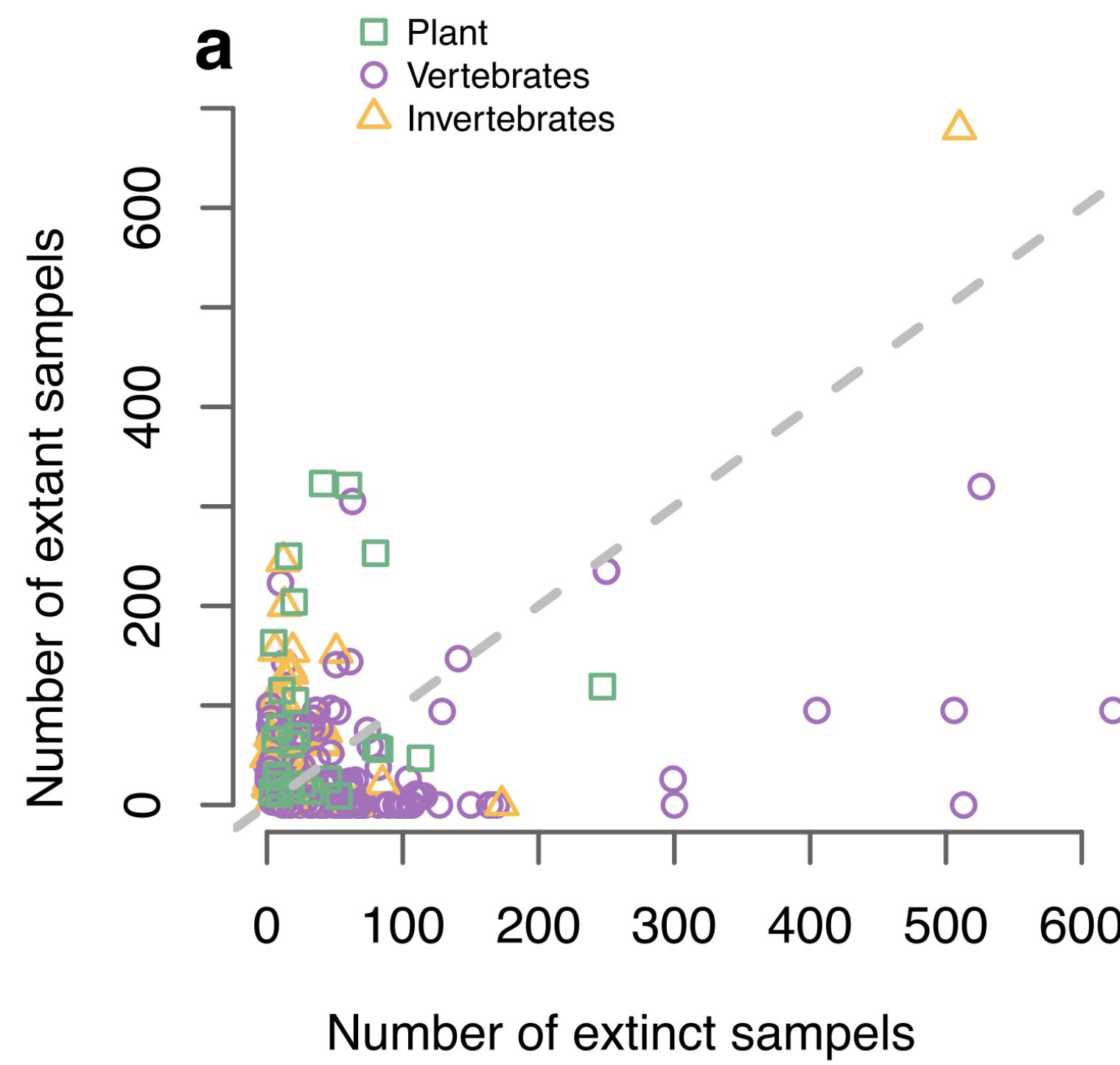
Fossils are incorporated
using character data

Gavryushkina et al. (2016)

Literature survey: FBD use to date

- 176 studies, with 208 empirical analyses, applying the FBD process (since February 2024)
- Used across a huge range of time intervals, data sets (mean extant samples = 74, mean extinct samples = 60)
- 109 studies used BEAST2, 93 used MrBayes, 10 used RevBayes

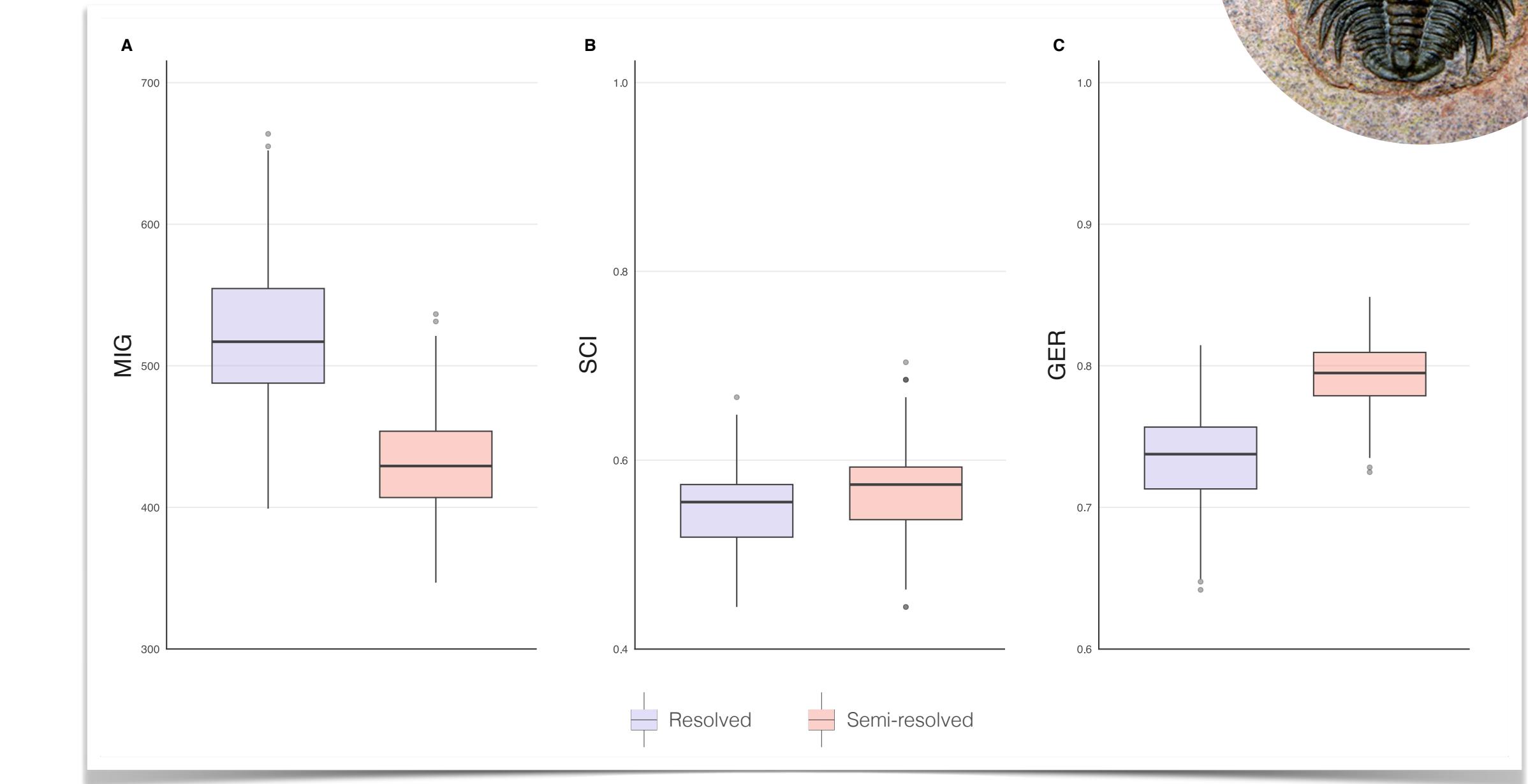
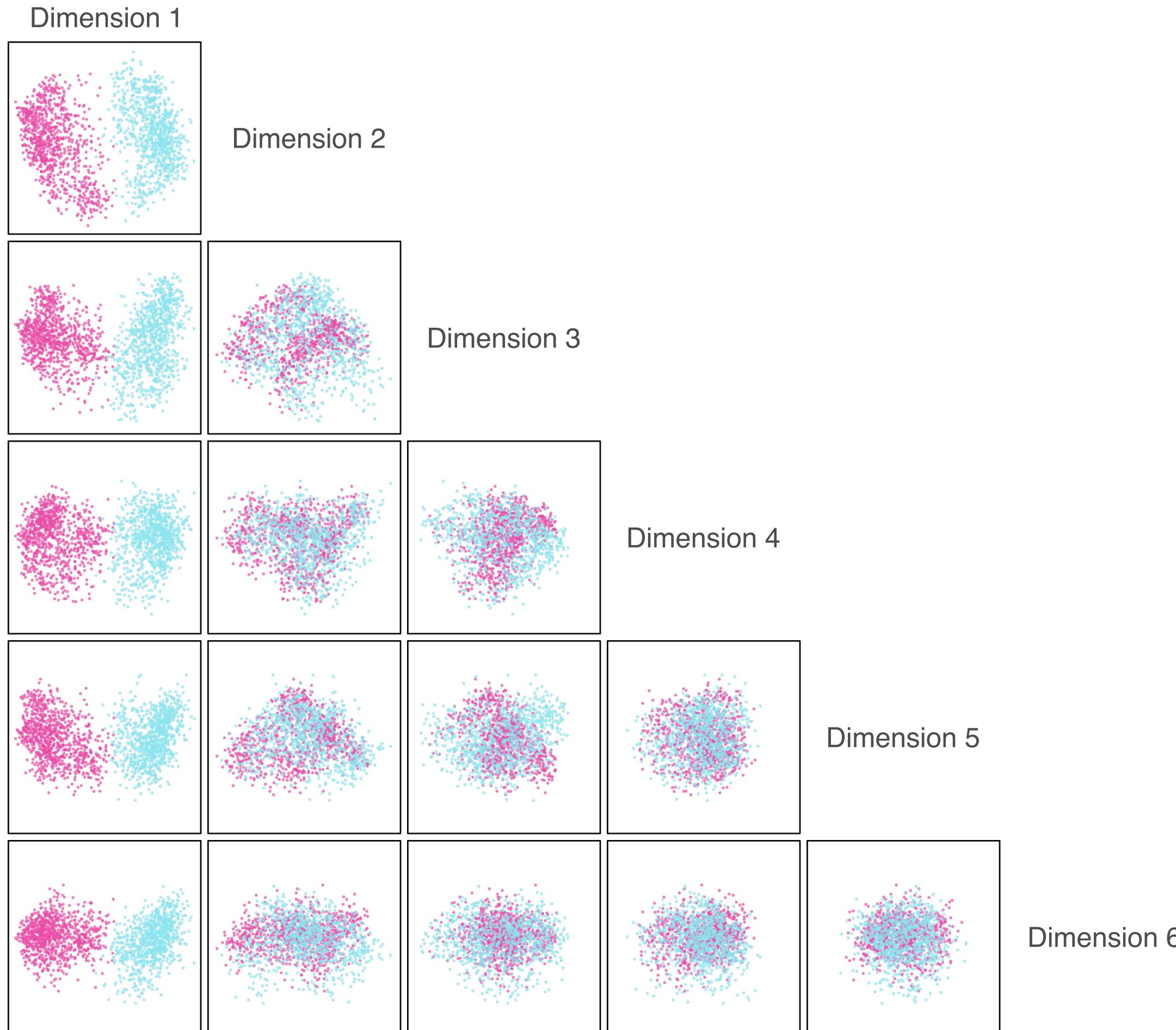




Phylogenetic data

Analysis type	Molecular	Morphology	Morphology [†]	No. of analyses
Total evidence	✓	✓	✓	53
Extant only	✓			78
Morphology		✓	✓	26
Extinct only			✓	35
No phylogenetic data				16

Including samples with no character data improves inference



And show worse / better correspondence
with the fossil record

Trees occupy different regions of tree space

Some lessons learned from simulations

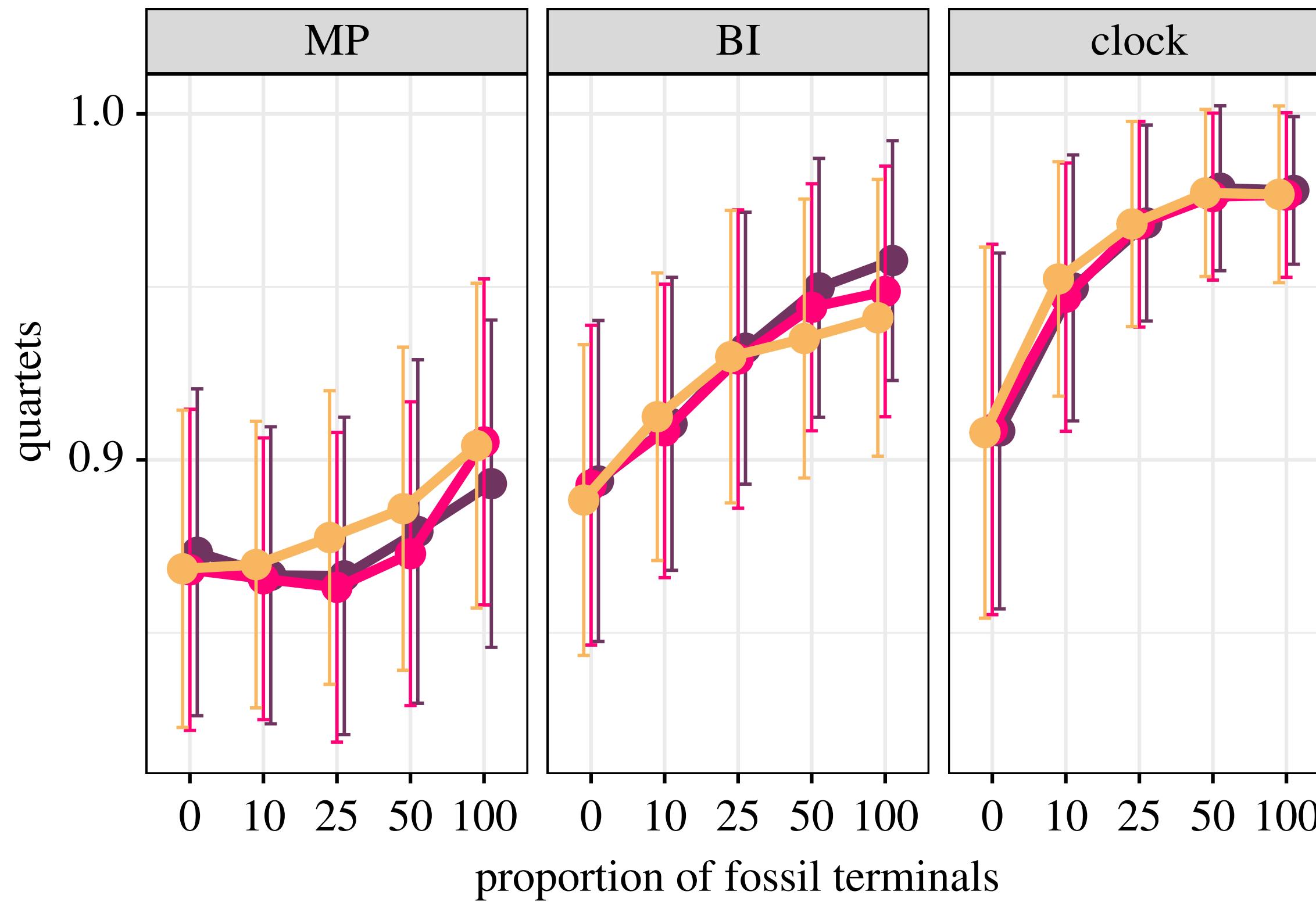
- The topology of extant taxa is largely unaffected by how fossils are incorporated
- Fossils and age information help inform topology
- Divergence times are much more sensitive to errors in fossil placement and model misspecification
- Total-evidence dating is more robust to model misspecification

Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times and topology

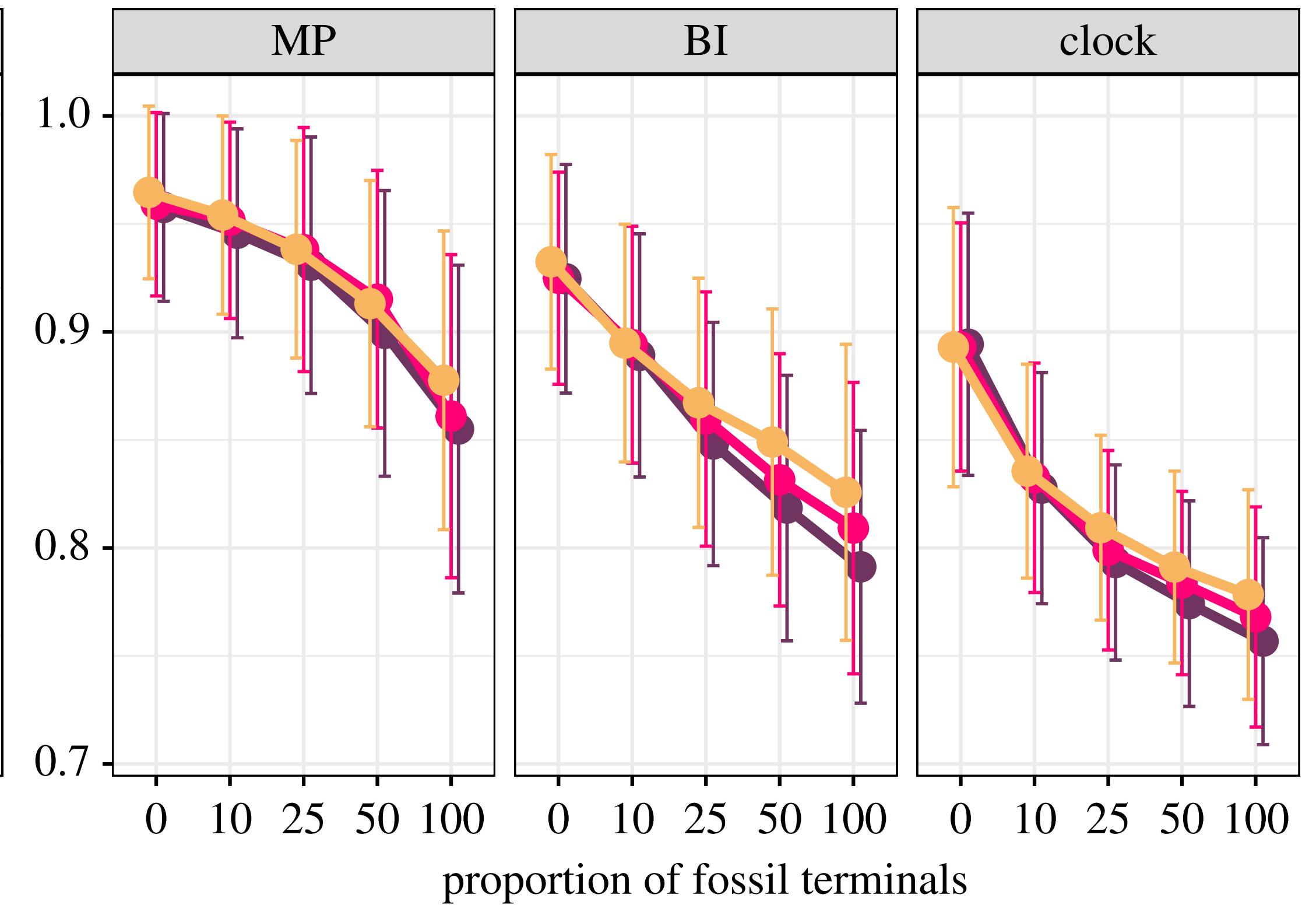
Barido-Sottani et al. (2019, 2020). *Proc Royal Soc B; Frontiers in Ecology and Evolution*.

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

accuracy



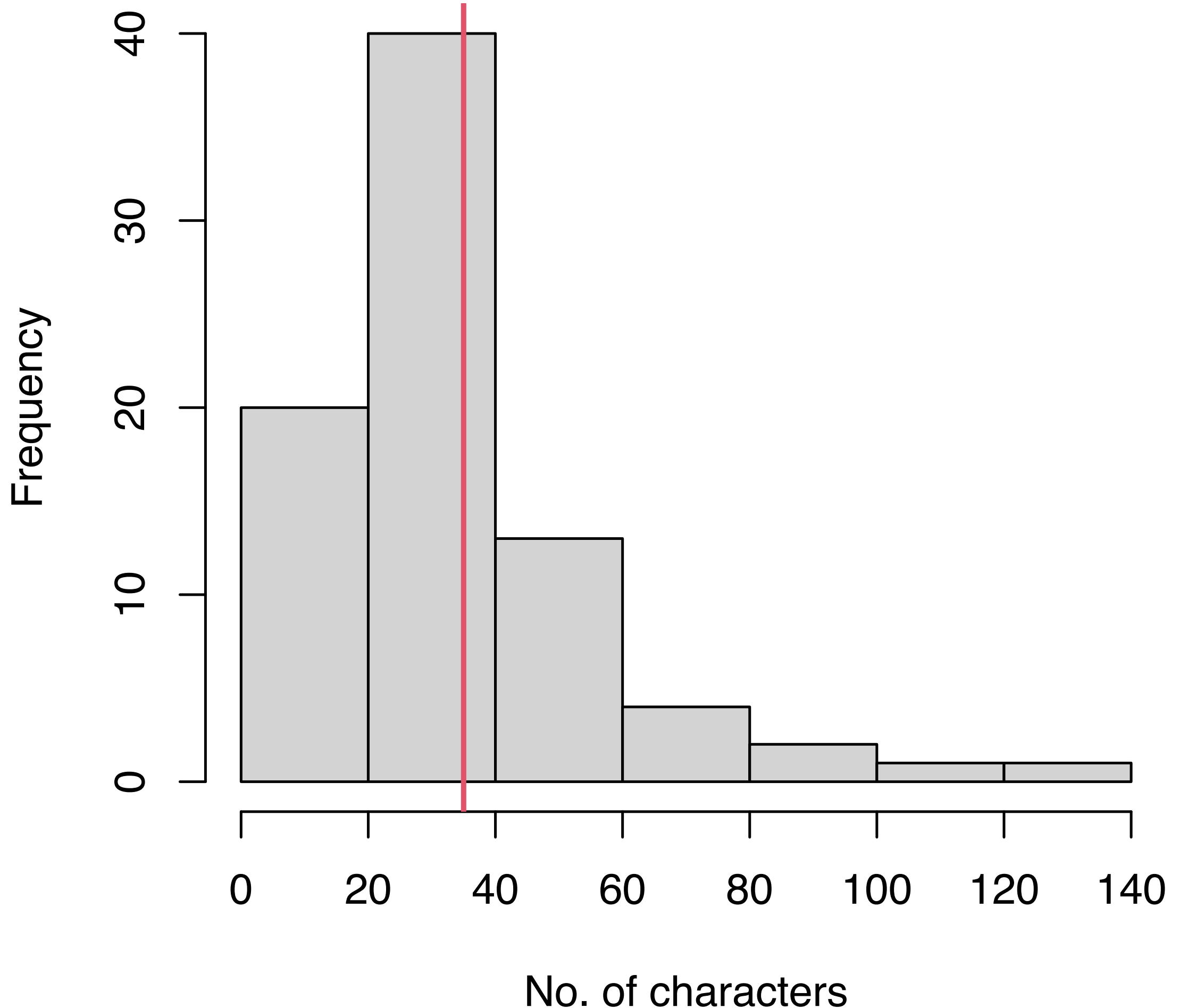
precision



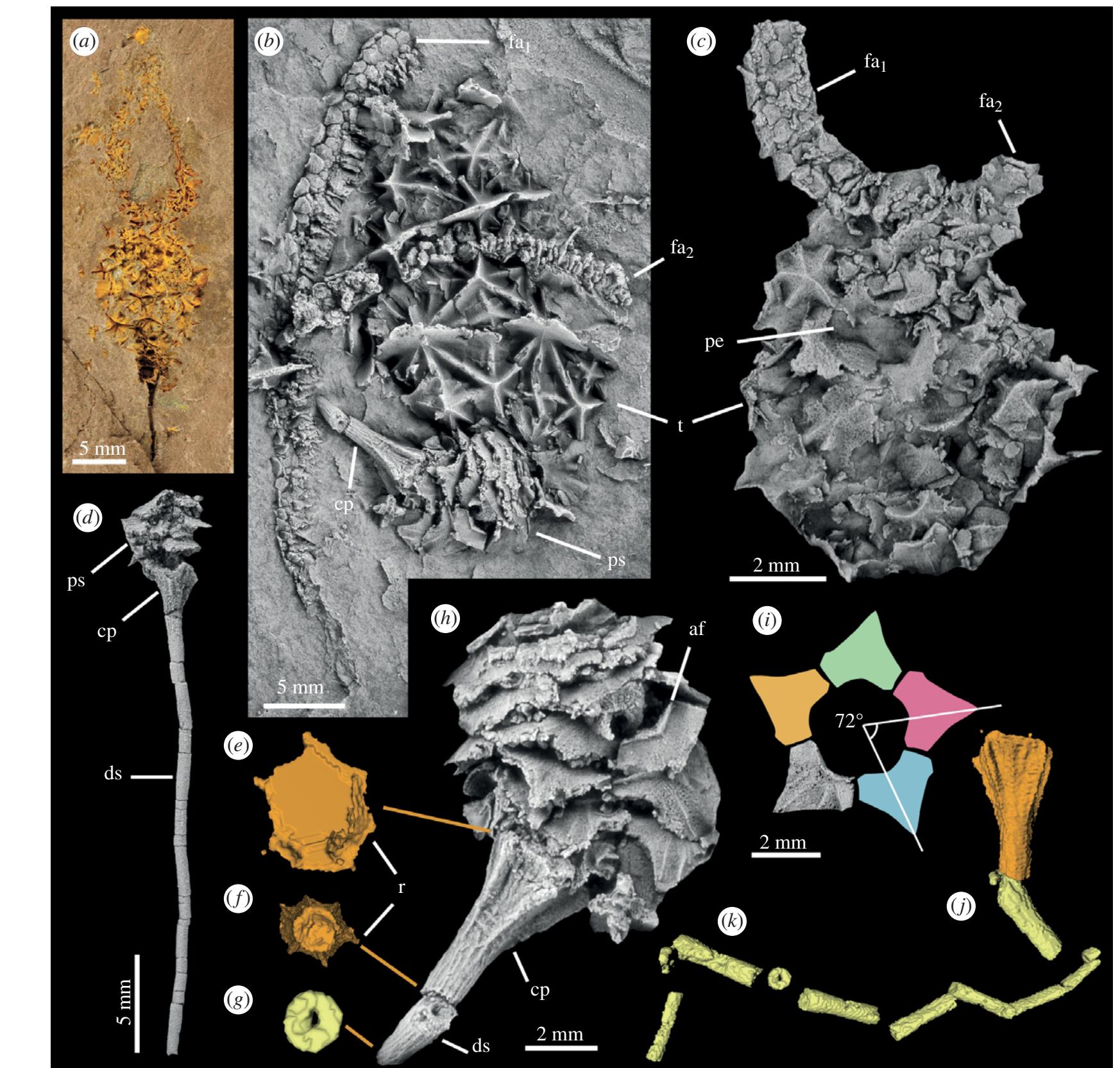
level of missing data ● none ● low ● high

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

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



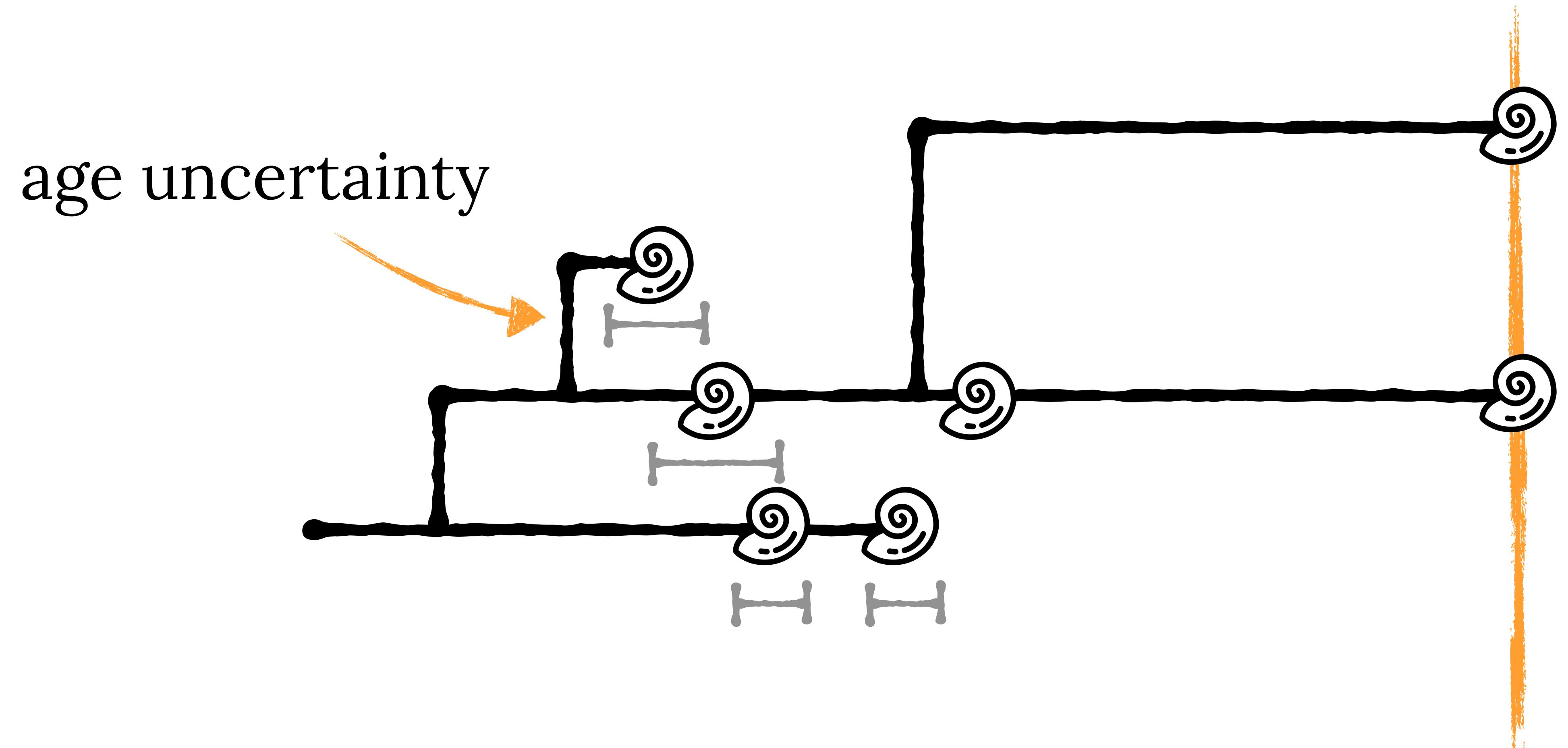
001510010?00-100--0000000000
 000500010?200100--0010010000
 002500010?200100--0?10010000
 00?5?0010?200100?-0??010110
 0015000101201000430100011111
 0015000101201010440111011111
 ??050?????201000440?11011111
 01050?010-210000?501??010110
 00020001002101003-1110010110
 0002000100211001441121011111
 000201111-210010?-?11011121
 ?103?0?11?1001104-0000010000
 1005002110100010--0?00110?20
 1005002000101010540?00110020



Dibrachicystis purujoensis

*Cambrian stalked echinoderms show
unexpected plasticity of arm construction*
 Zamora & Smith. (2012)

Sample age uncertainty



FBD analyses
need taxonomists!

OCTOPI WALL STREET

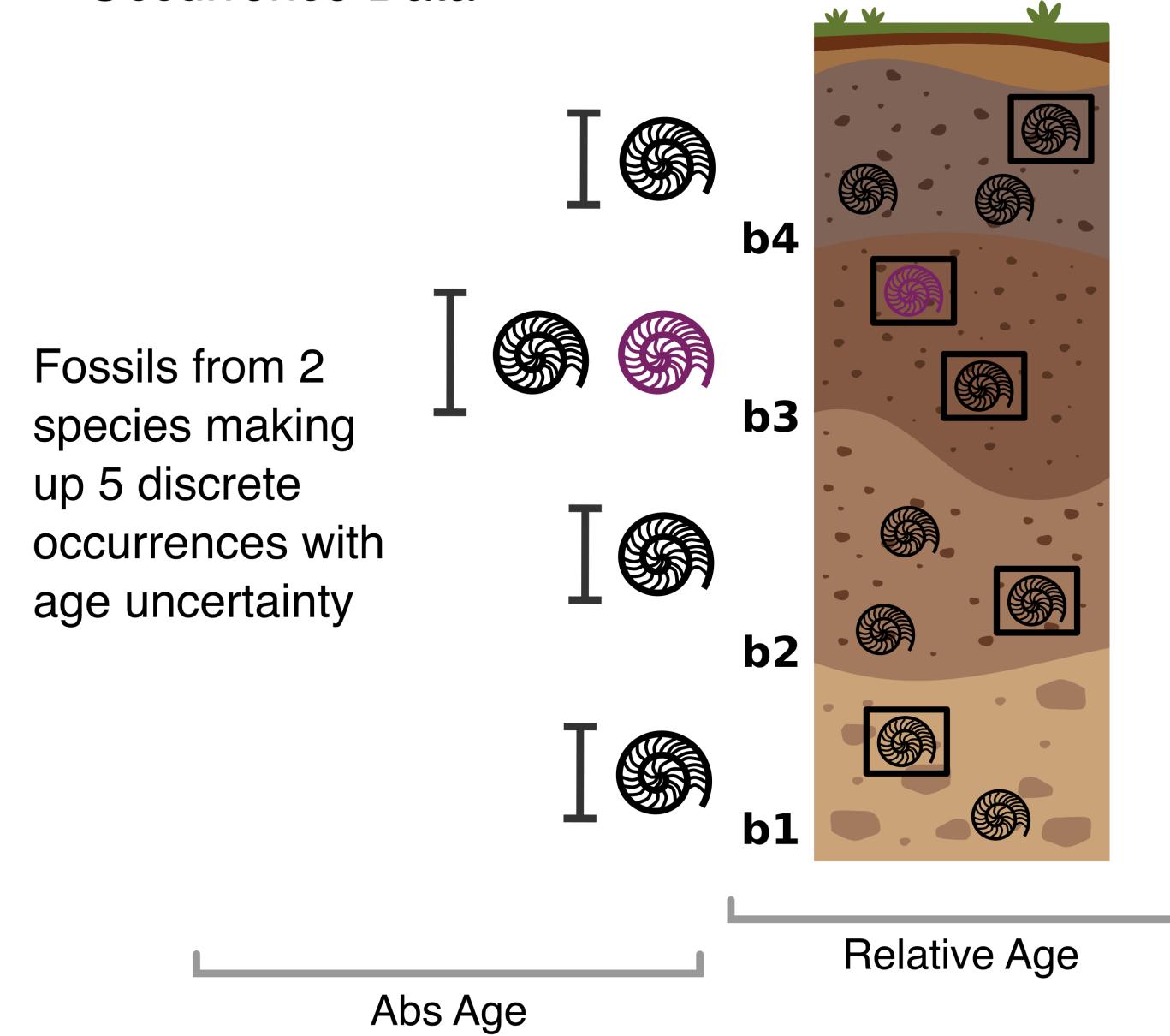


Invertebrates are 97% of animal diversity!

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

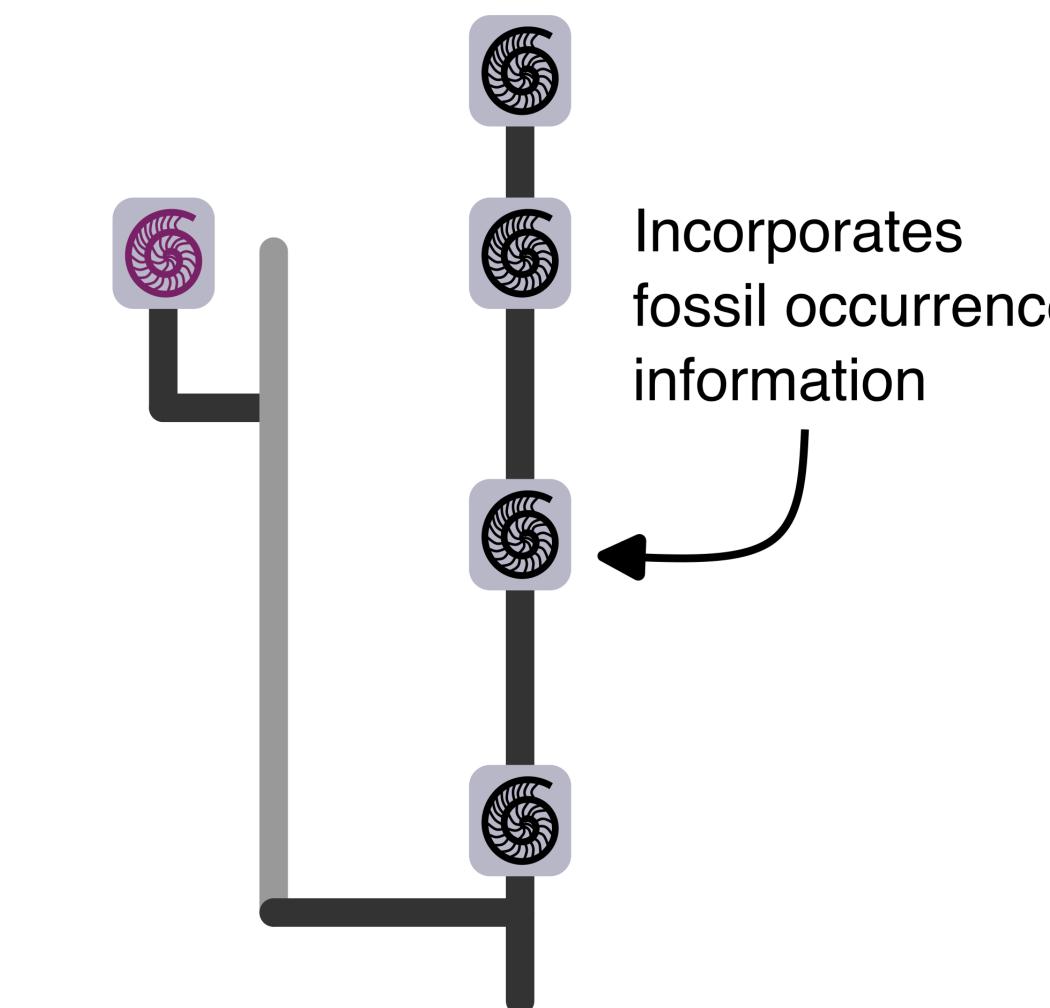
A Fossil Ages

Occurrence Data

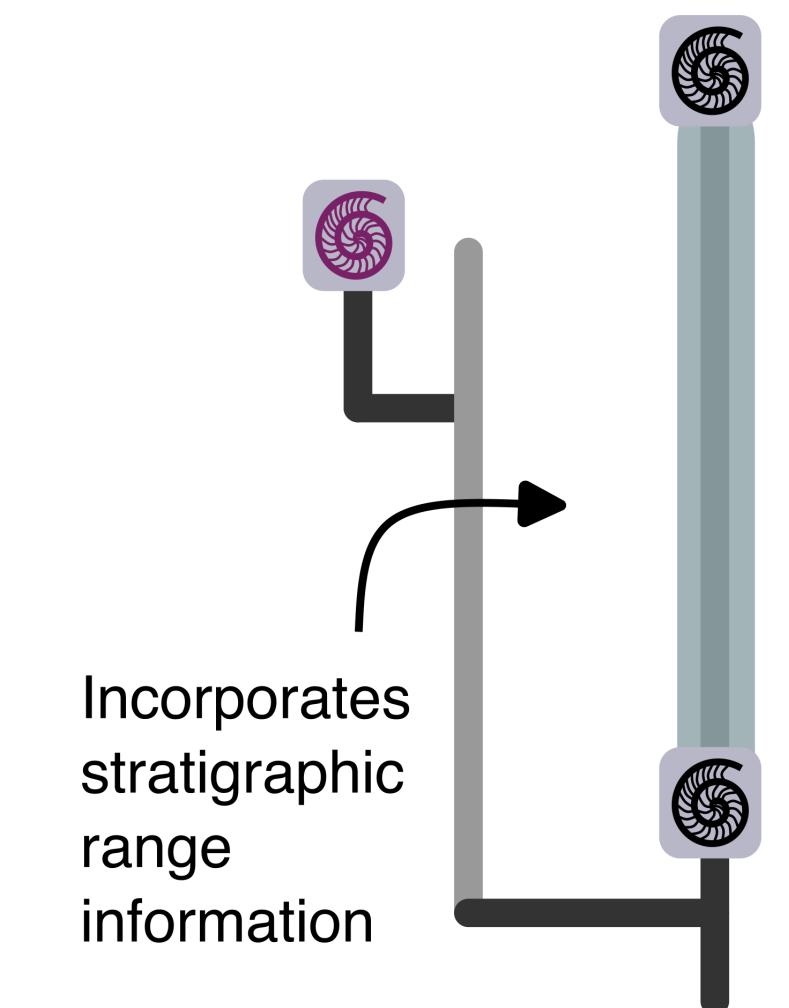


B FBD Models

FBD specimen model



FBD range model



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

Exercise

Phylogenetics

Diversification rate estimation

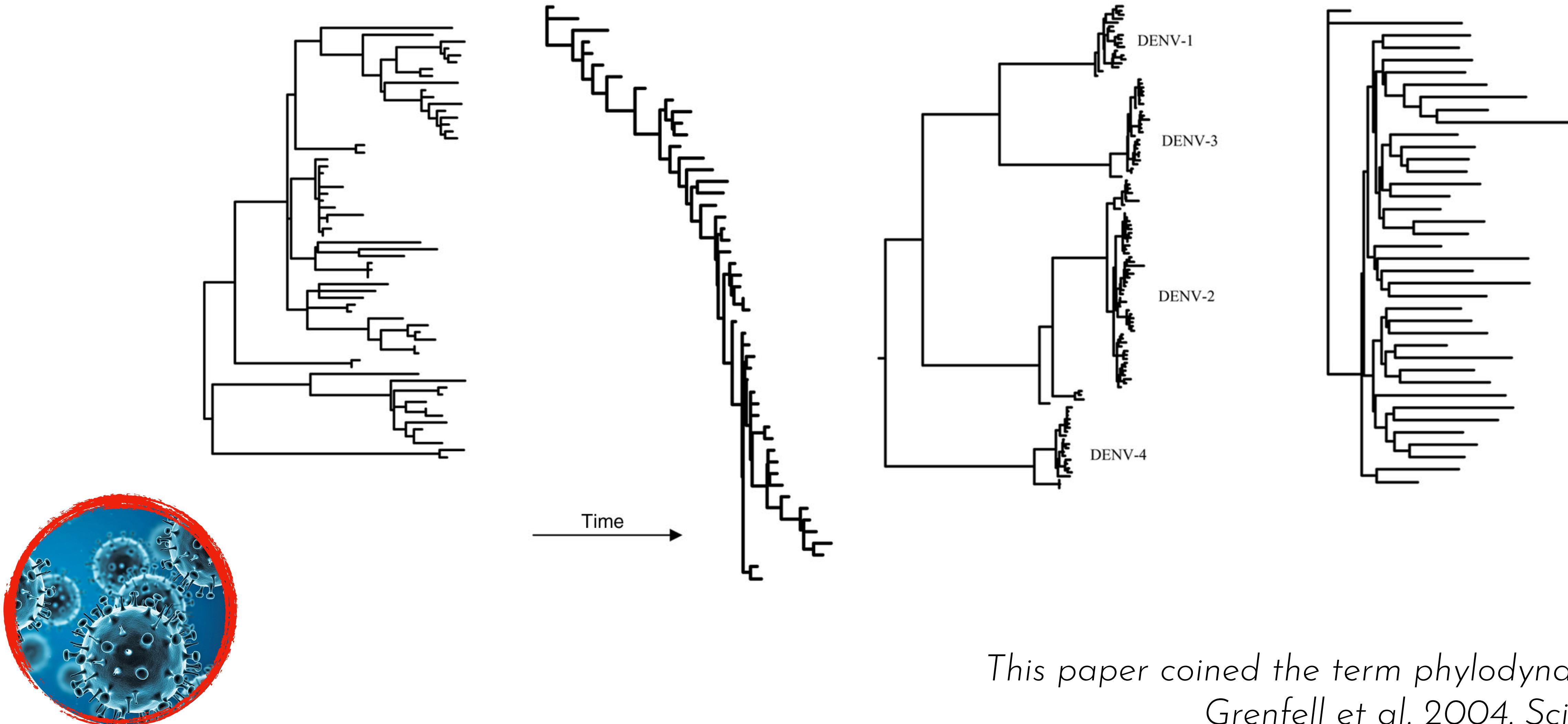
Bayesian divergence time estimation

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

probability of the
time tree

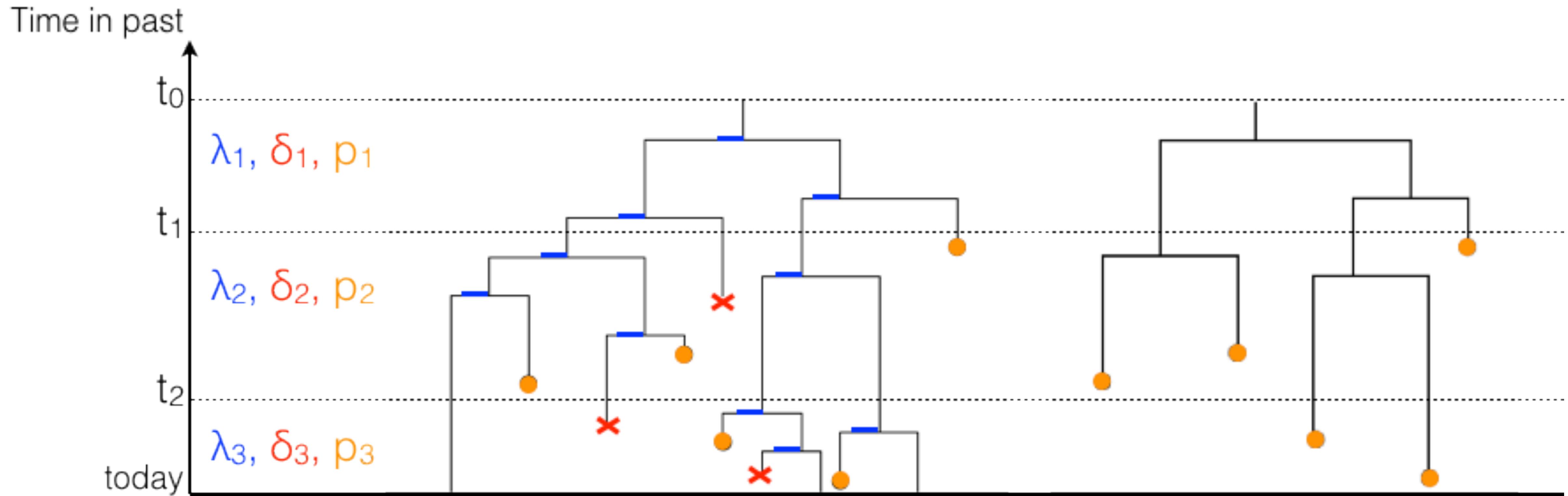
$$\frac{P(0101\dots, 1101\dots, 0100\dots \mid E) P(E \mid \lambda, \mu, \psi, p, O, t) P(\lambda, \mu, \psi, p) P(O) P(t)}{P(0101\dots, 1101\dots, 0100\dots, \text{snail})}$$

Tree shape is informative about underlying dynamics

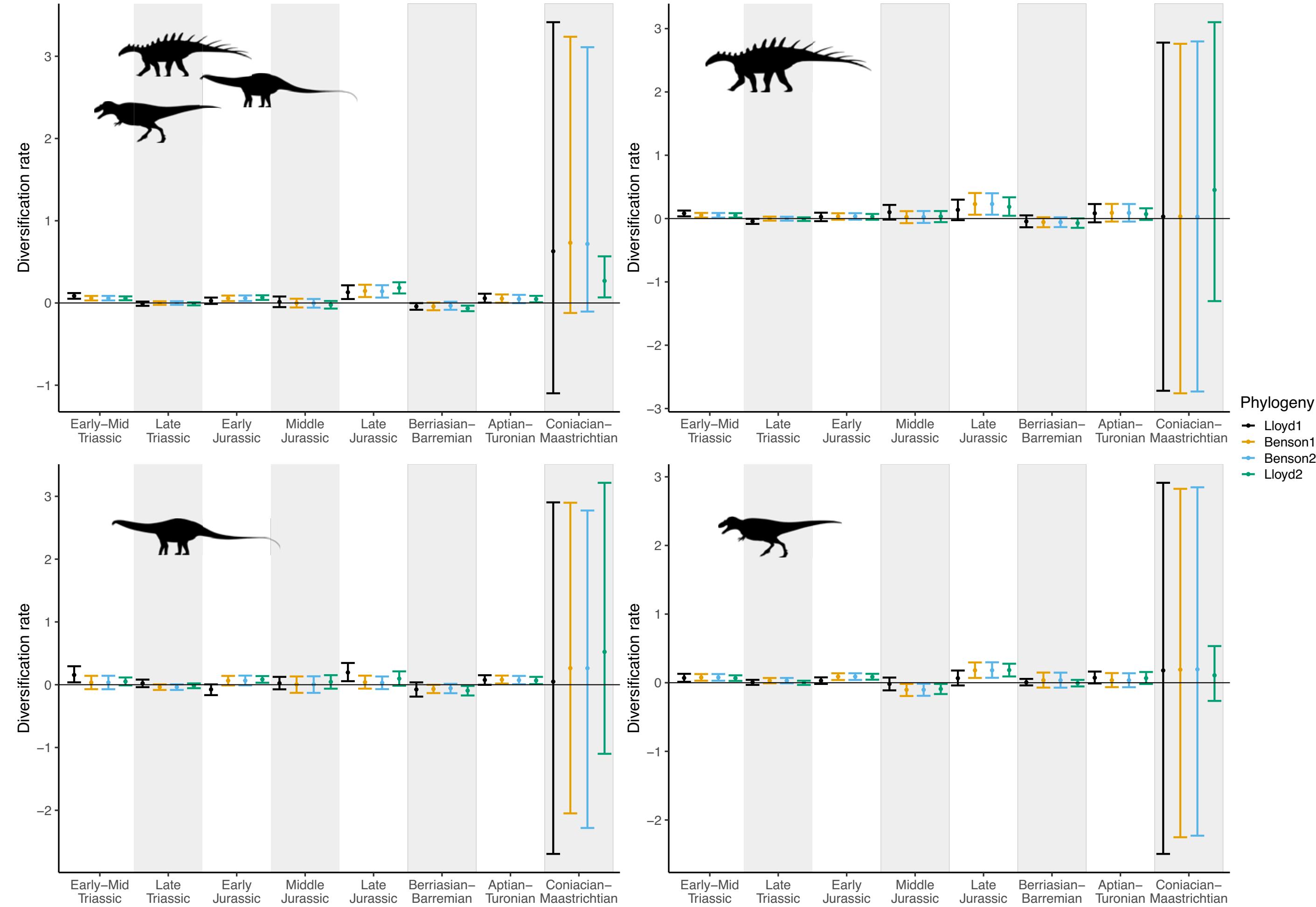


The skyline birth-death process

First used for tracking the spread of infectious diseases



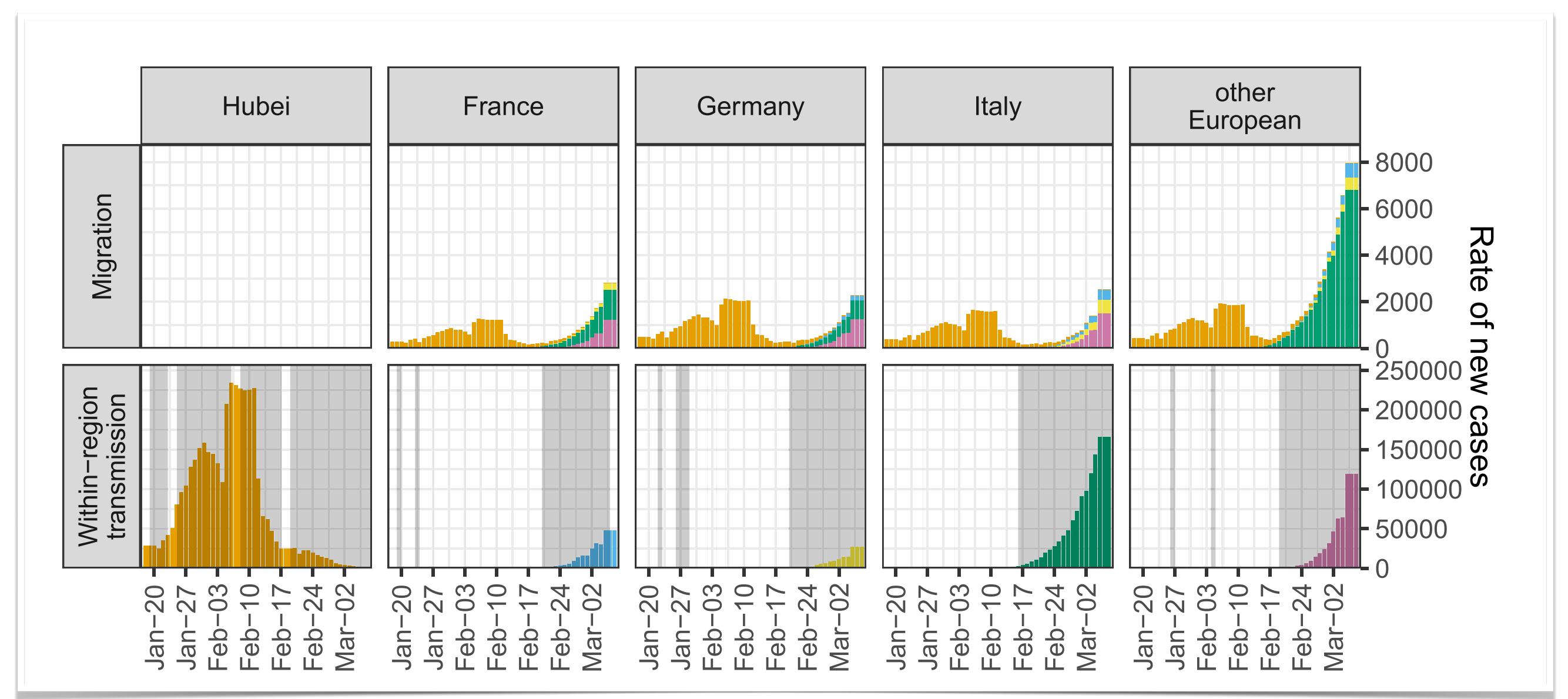
Macroevolutionary case study



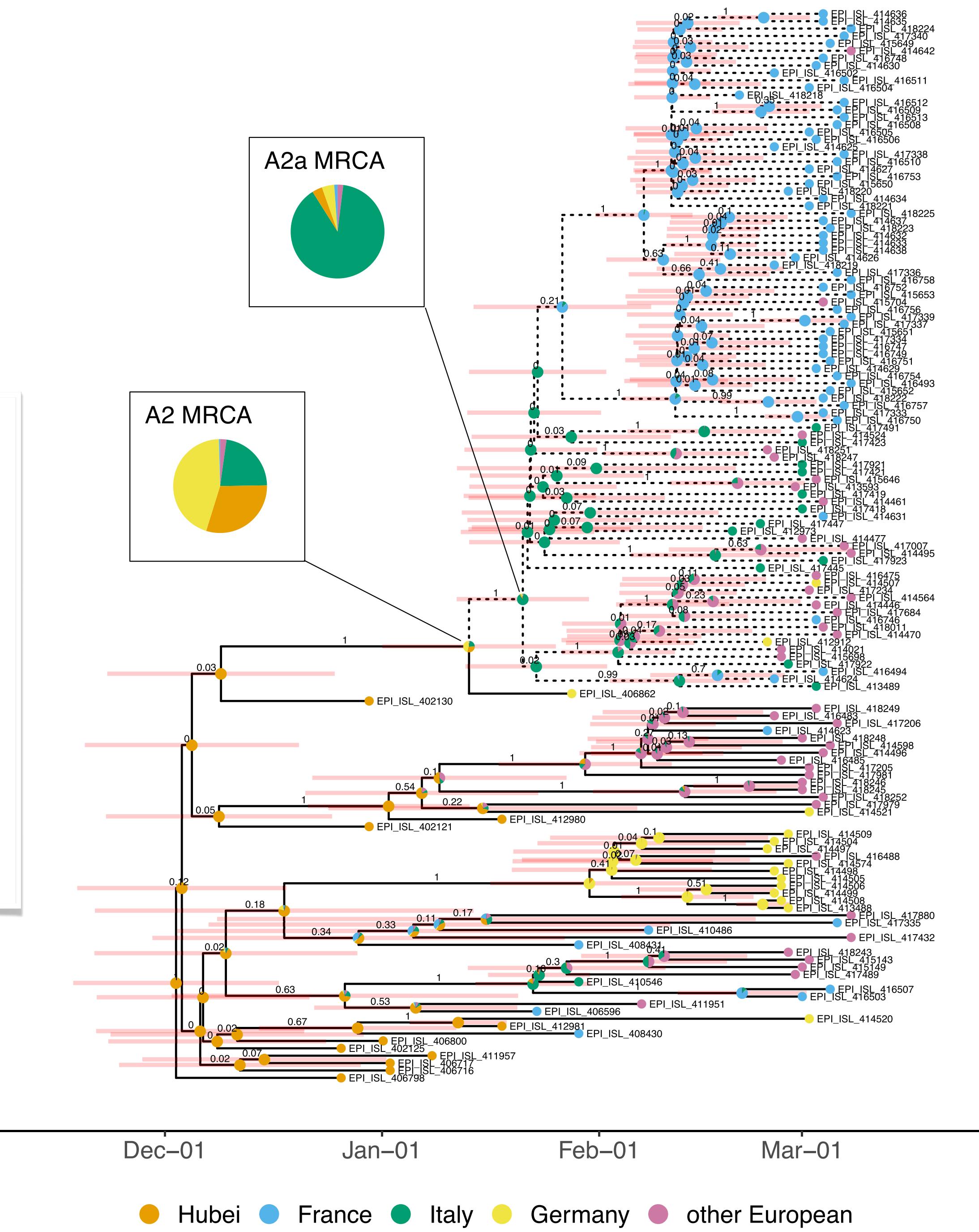
Phylogenies have been used to argue dinosaurs were incline prior to the KPg

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

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

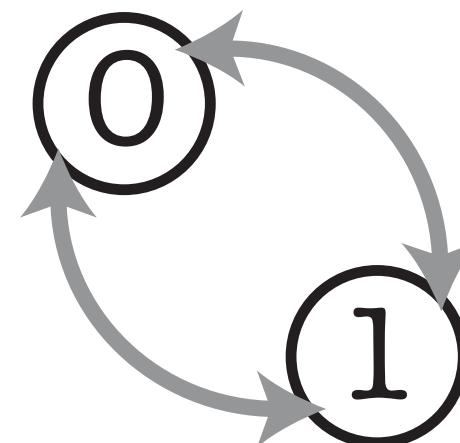
and / or
0101... ATTG...
1101... TTGC...
0100... ATTC...



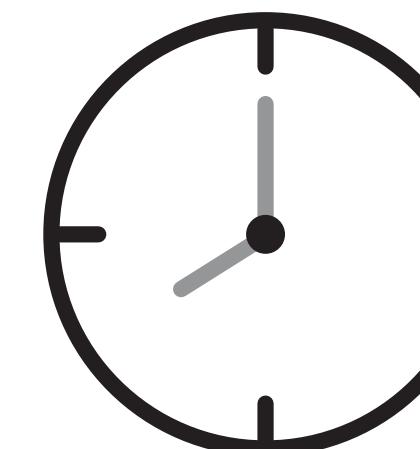
phylogenetics
characters

sample
ages

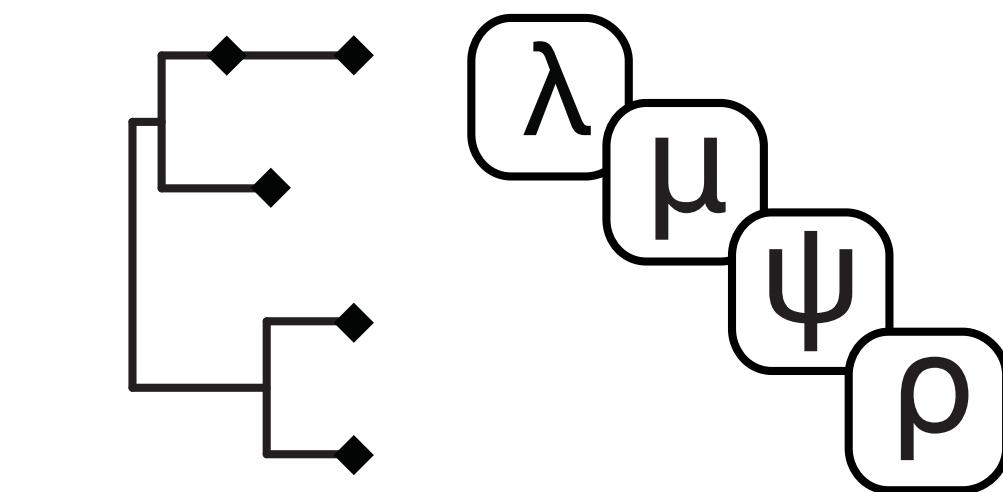
3 model components



substitution
model



clock
model



tree and tree
model

Using PCMs for dating

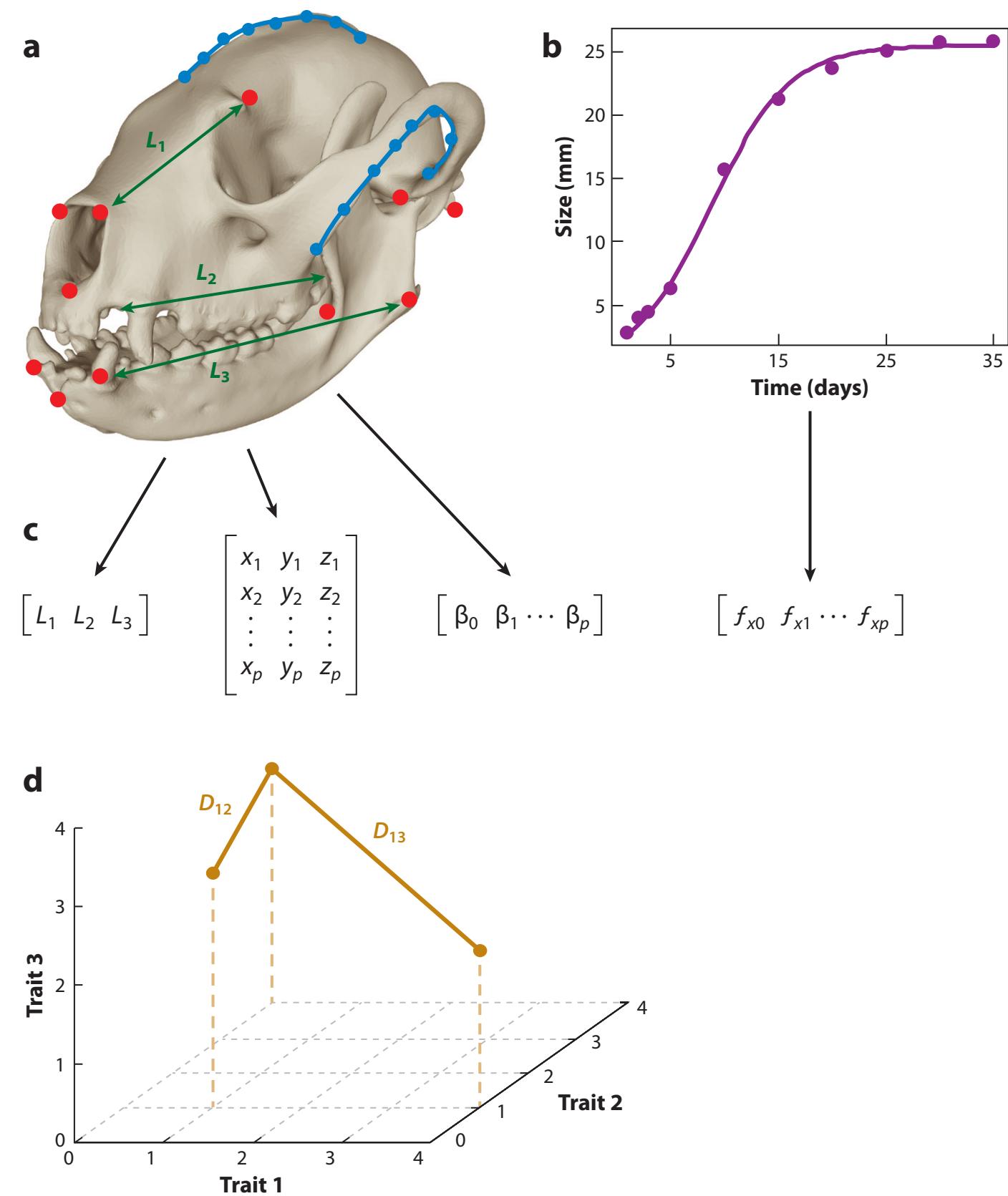
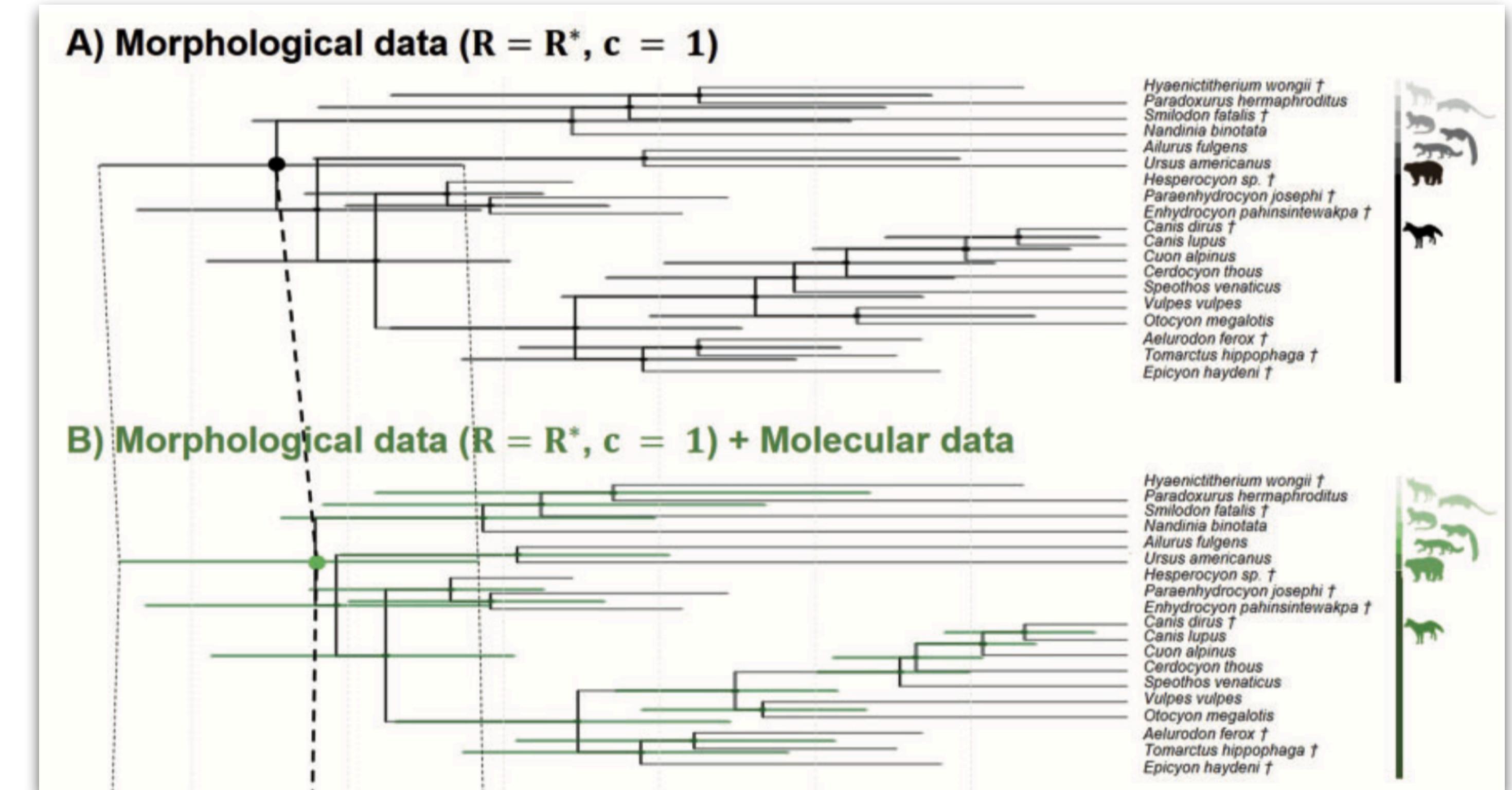
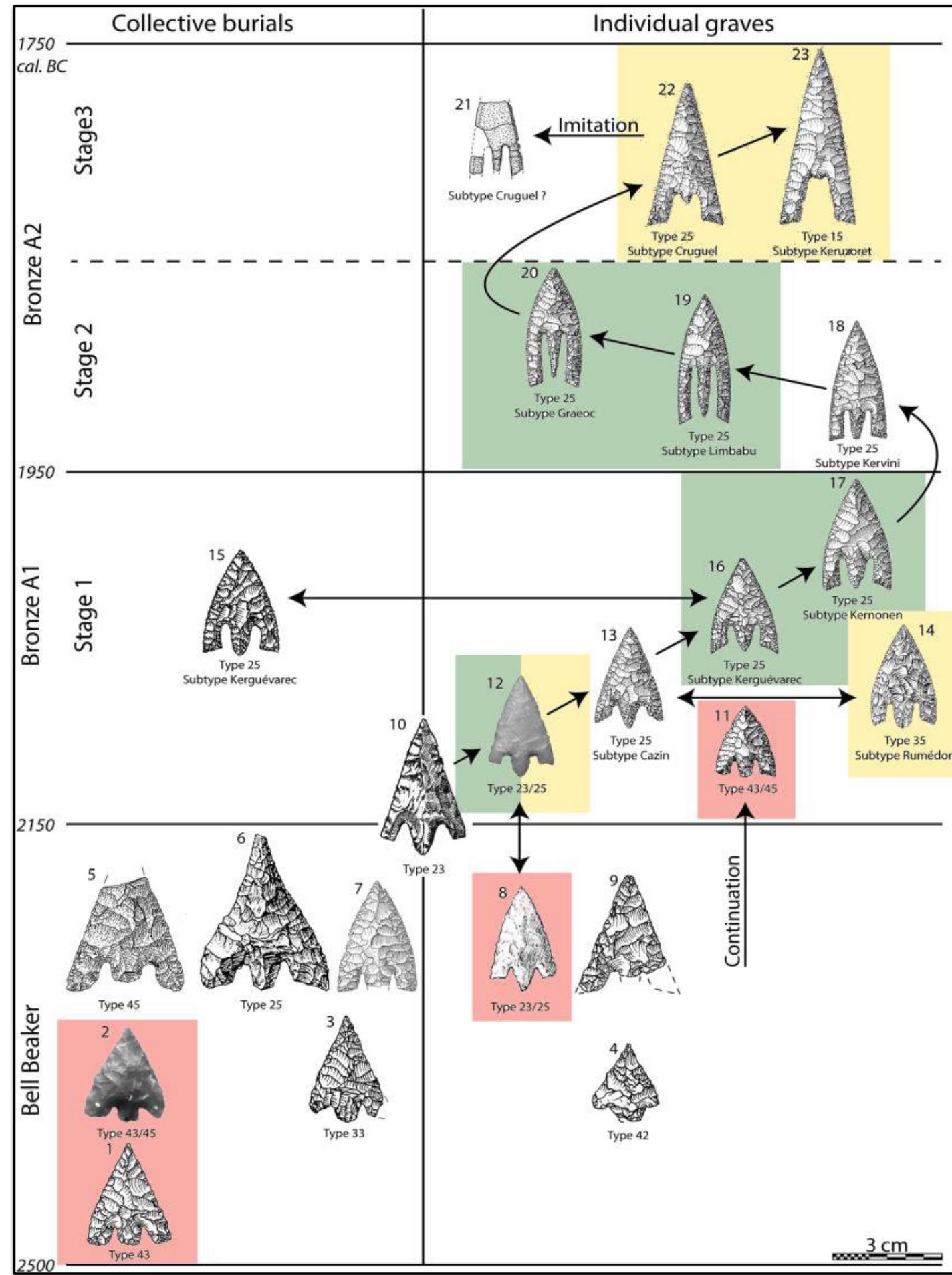


Image source Adams & Collyer ([2019](#))



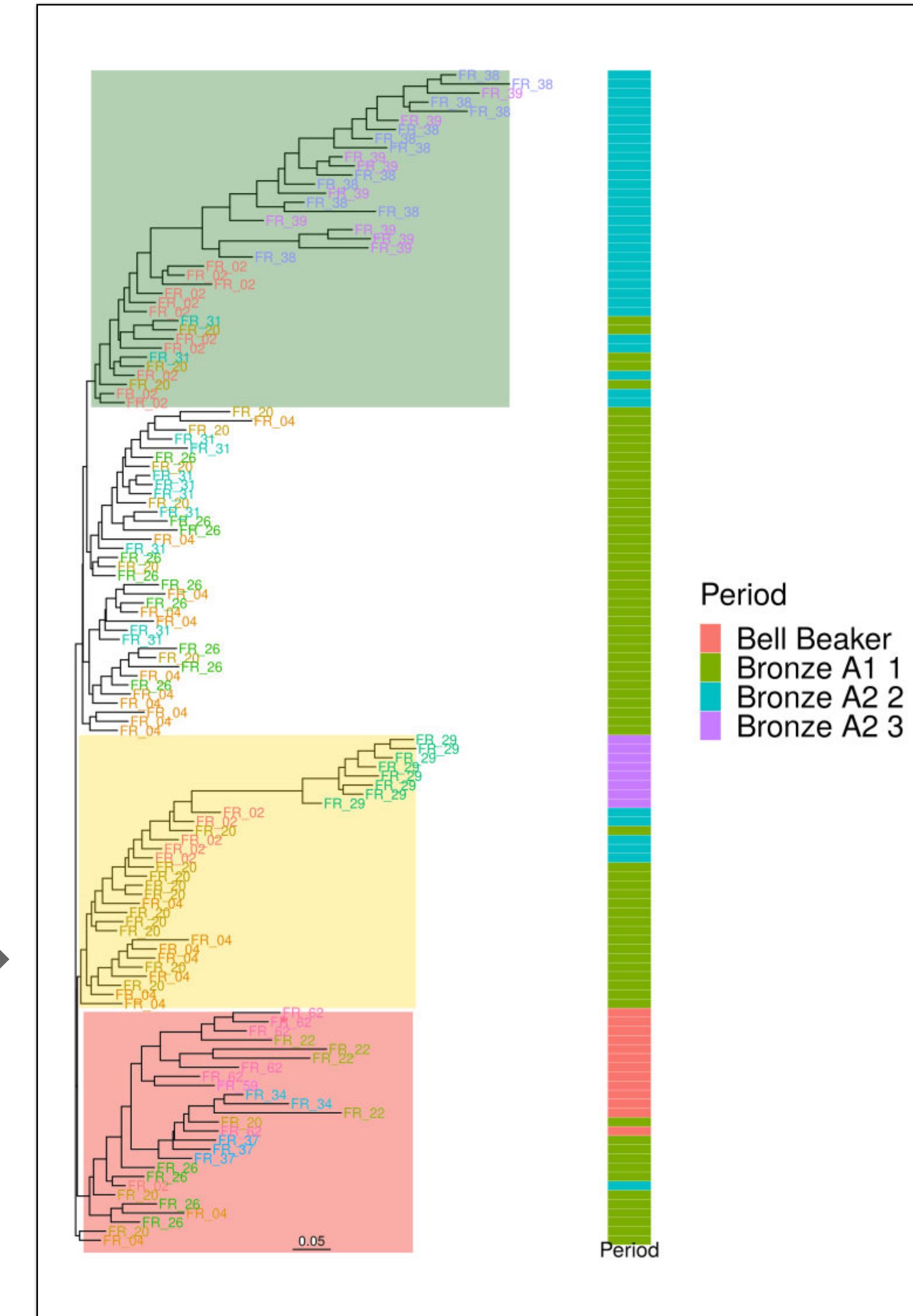
Álvarez-Carretero et al. ([2019](#)) Bayesian Estimation of Species Divergence Times Using Correlated Quantitative Characters

Cultural evolution



← Typo-Chronology of
Palaeolithic stone tools

Outline based NJ tree →

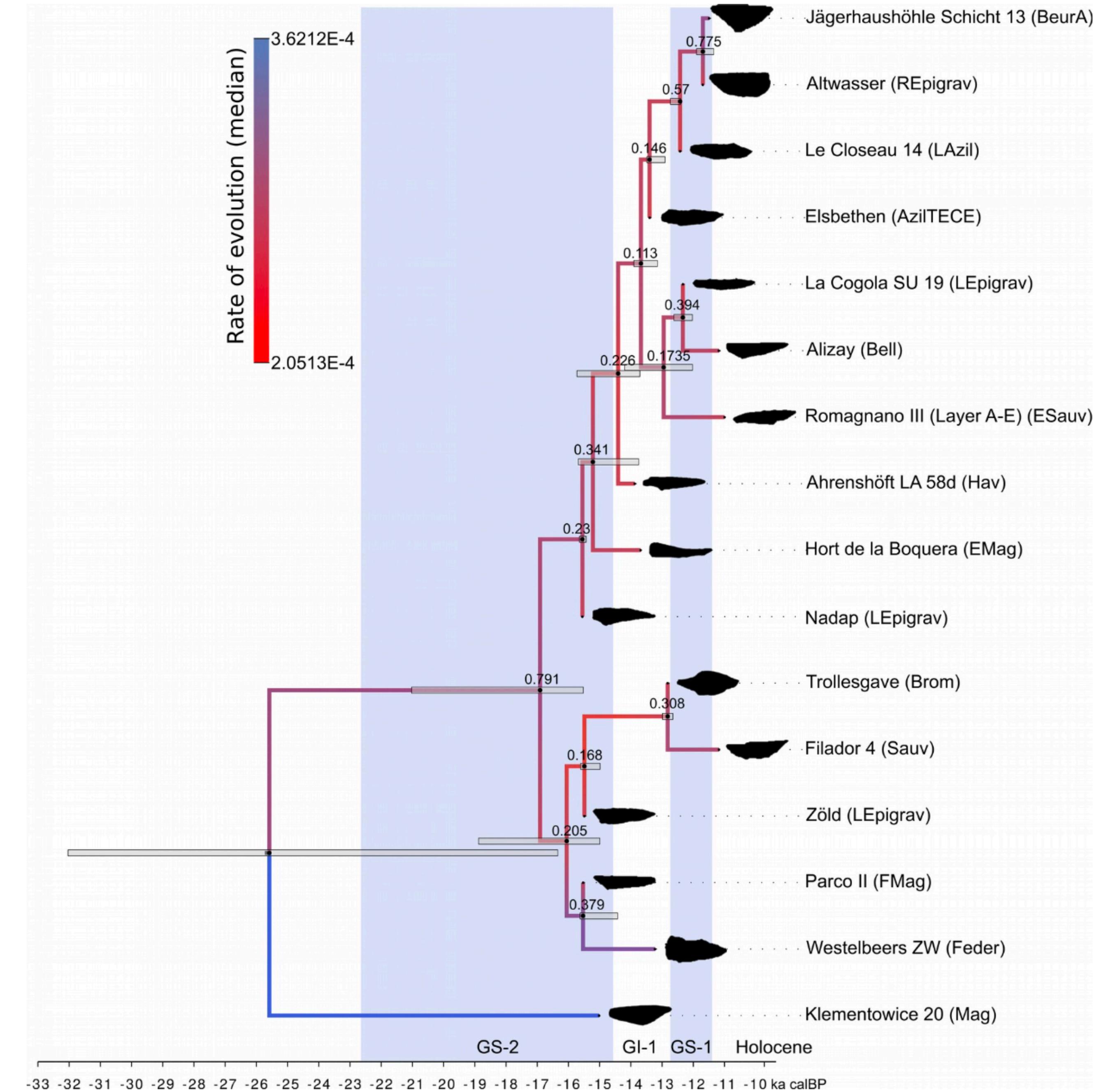


Matzig et al. 2021.

After Nicolas (2017)

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)



Decoding Genomes: From Sequences to Phyldynamics

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

Illustrated by Cecilia Valenzuela Agúí

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 phyldynamics 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

