Phylogenetics

Introduction to phylodynamics models

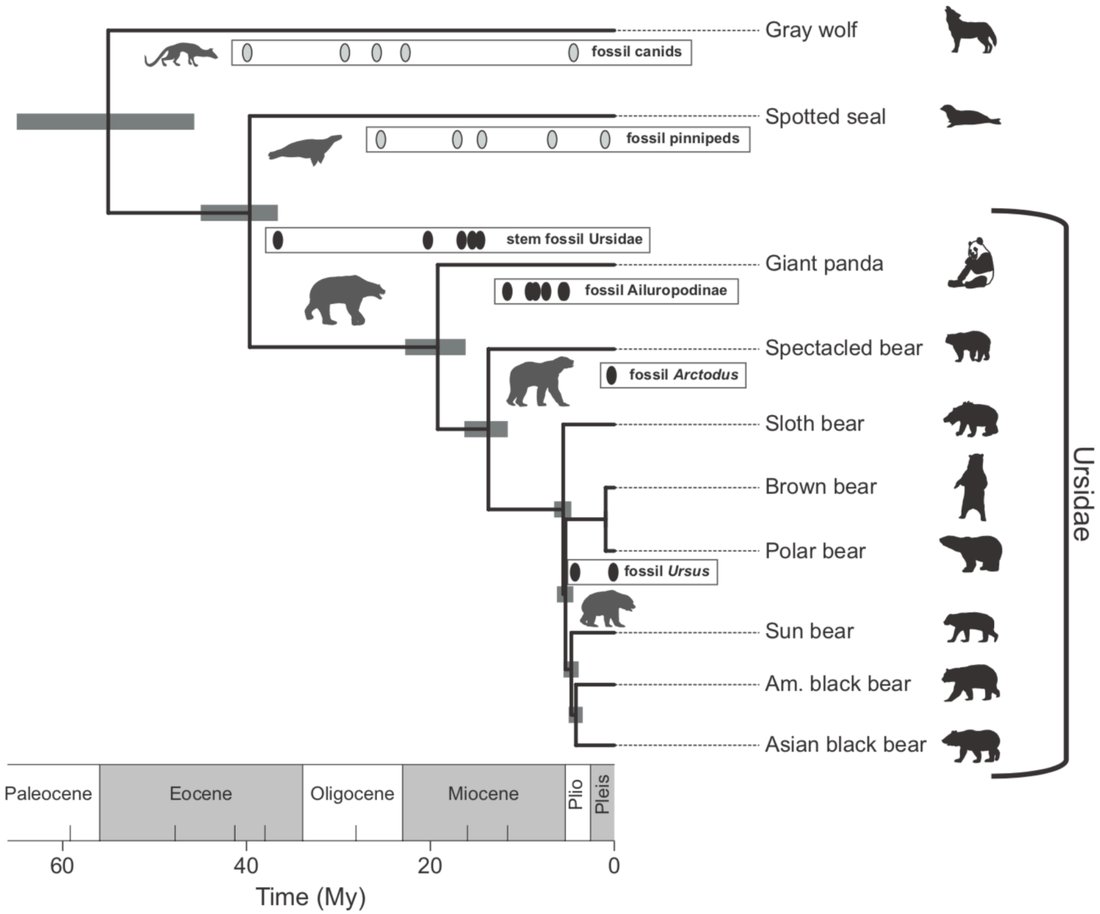
RL-V3 MPP

Rachel Warnock

16.04.2025

Objectives

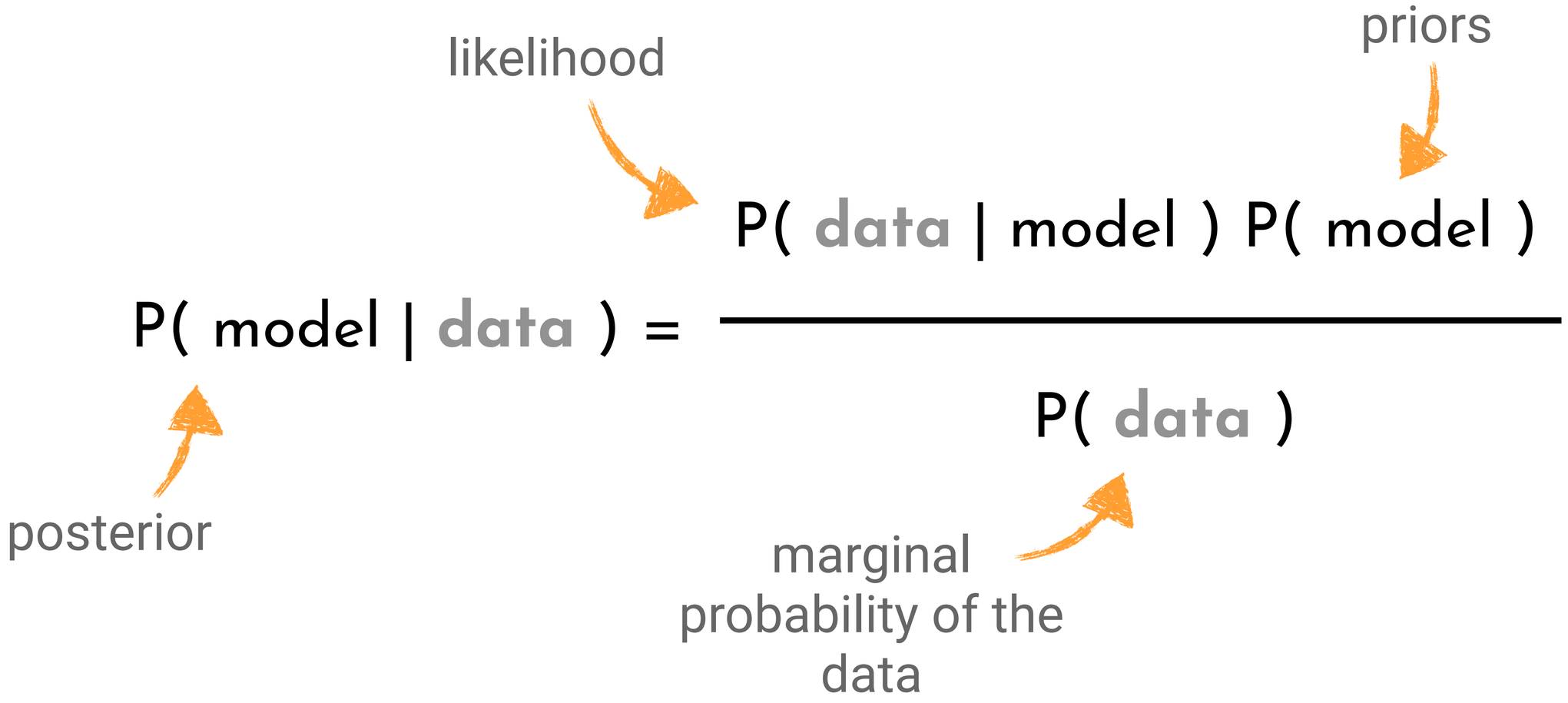
- **Recap**: tripartite framework
- The fossilised birth-death process
- Total-evidence dating
- Phylodynamics



Bayesian divergence time estimation

Recap

We use a Bayesian framework



Bayesian divergence time estimation

The data

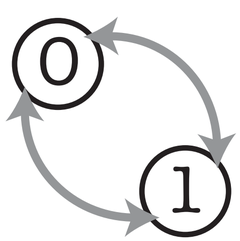
and / or

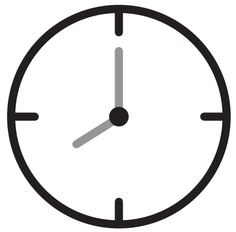
```
0101... ATTG...
1101... TTGC...
0100... ATTC...
```

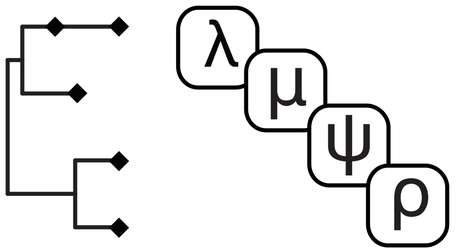
phylogenetics characters

sample ages

3 model components







substitution model

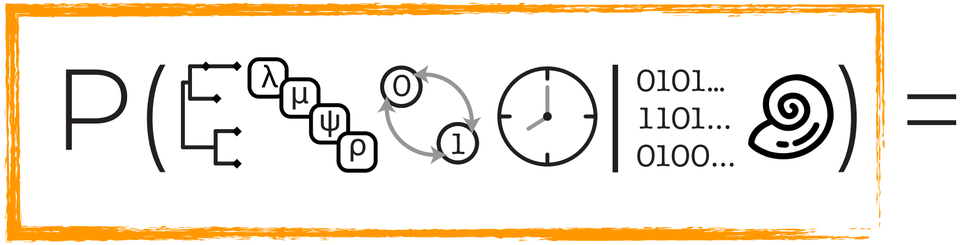
clock model

tree and tree model

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

Bayesian divergence time estimation

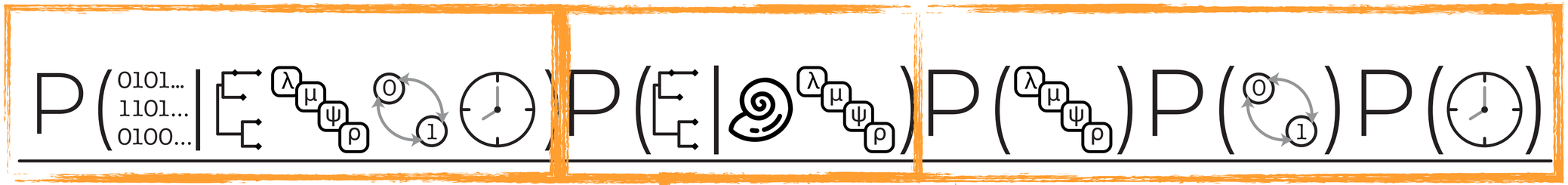
posterior

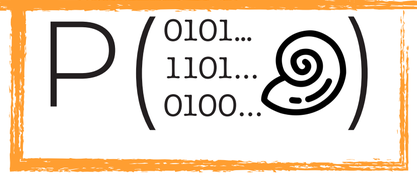


likelihood

probability of the time tree

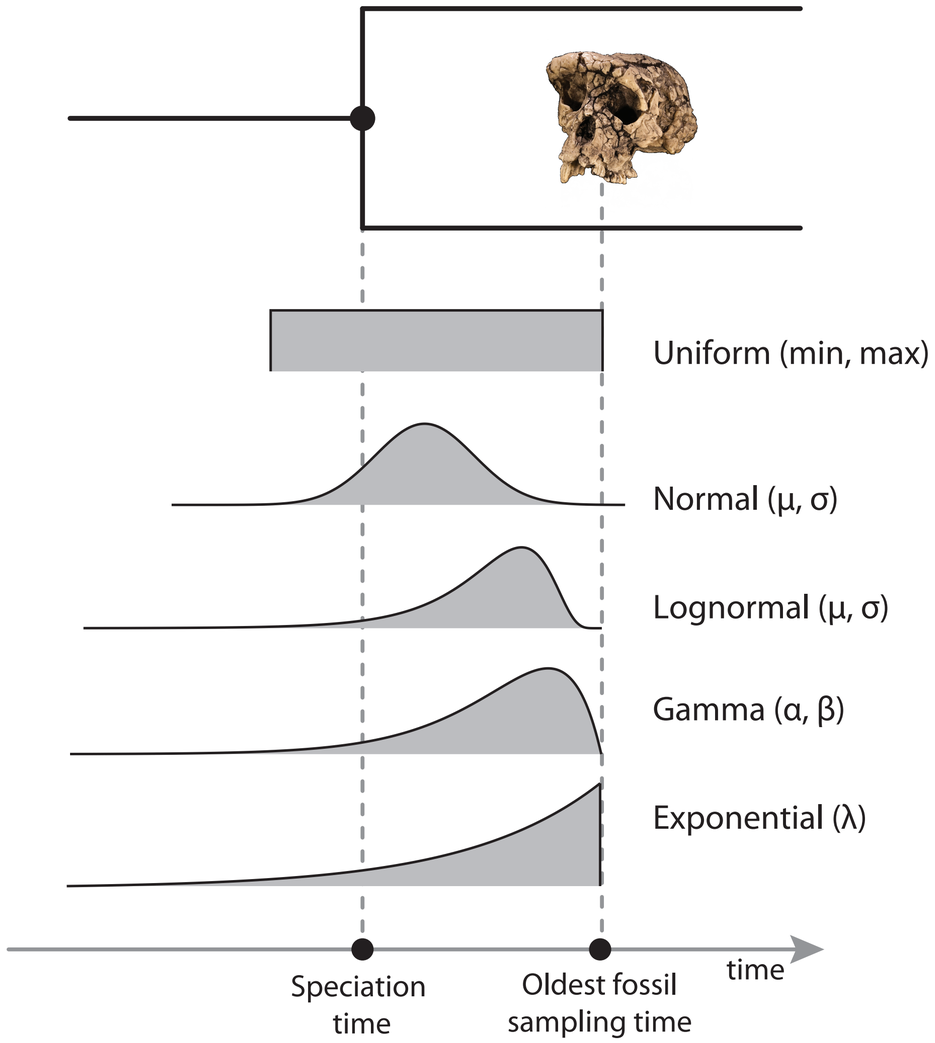
priors





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

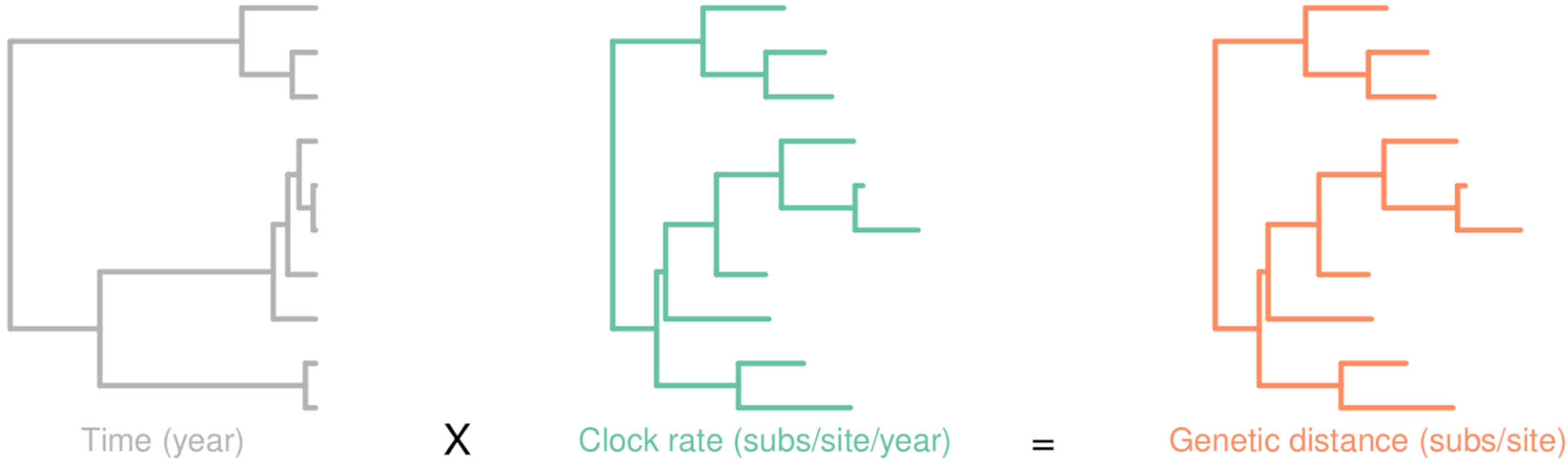
Adapted from *Heath* (2012). Sys Bio

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



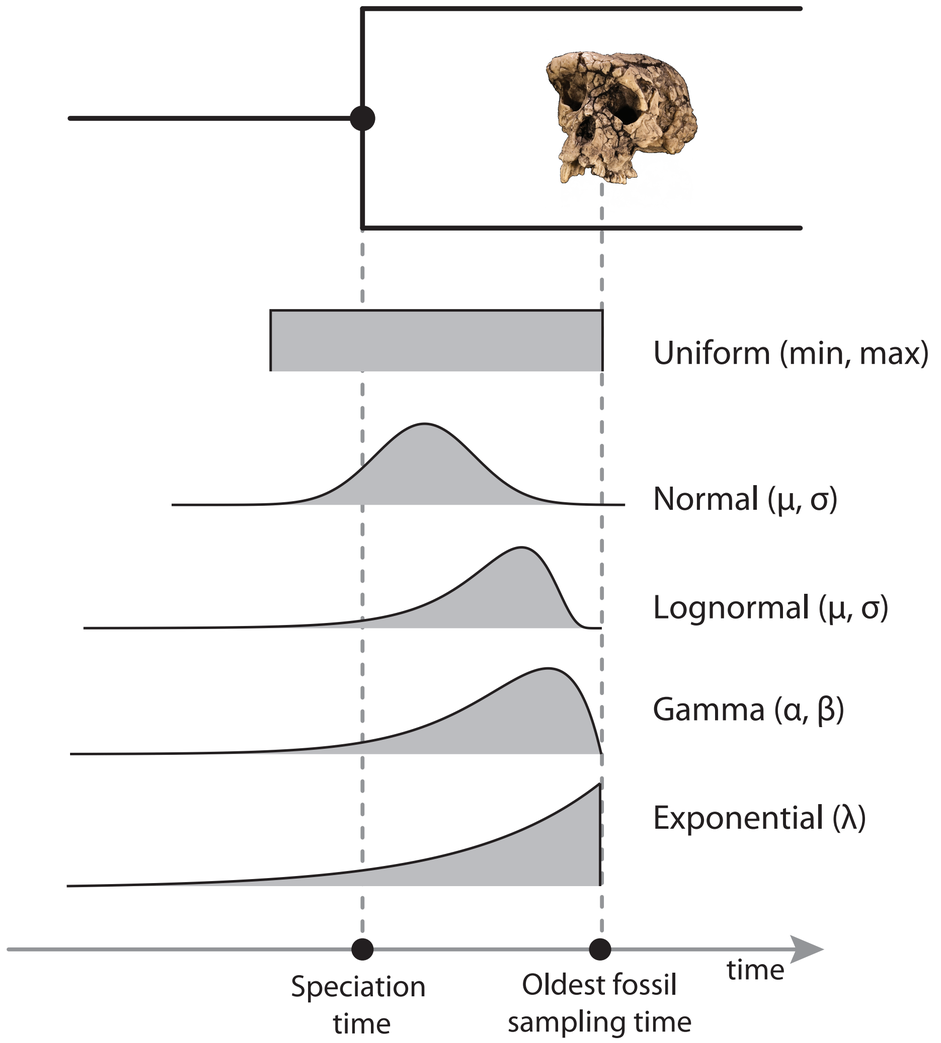
Prior

Prior

Likelihood

1.

Minimum and maximum constraints



Hard minimum bounds are based on first appearances

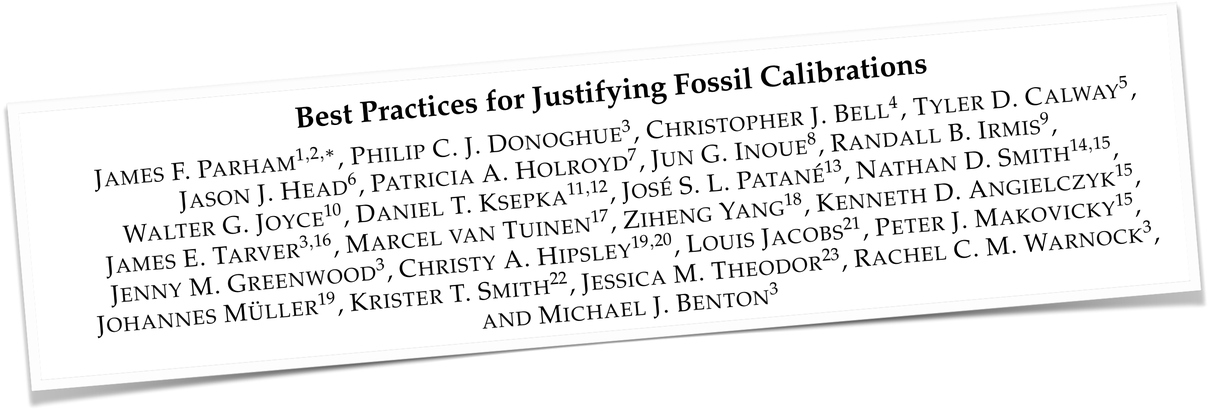
Soft maximum bounds are based on more tenuous evidence — typically the 97.5% limit of the calibration density

The upper bounds often influence the results

2.

Adapted from *Heath* 2012. Sys Bio

Maximum constraints



Best Practices for Justifying Fossil Calibrations

JAMES F. PARHAM[1,2,*], PHILIP C. J. DONOGHUE[3], CHRISTOPHER J. BELL[4], TYLER D. CALWAY[5], JASON J. HEAD[6], PATRICIA A. HOLROYD[7], JUN G. INOUE[8], RANDALL B. IRMIS[9], WALTER G. JOYCE[10], DANIEL T. KSEPKA[11,12], JOSÉ S. L. PATANÉ[13], NATHAN D. SMITH[14,15], JAMES E. TARVER[3,16], MARCEL VAN TUINEN[17], ZIHENG YANG[18], KENNETH D. ANGIELCZYK[15], JENNY M. GREENWOOD[3], CHRISTY A. HIPSLEY[19,20], LOUIS JACOBS[21], PETER J. MAKOVICKY[15], JOHANNES MÜLLER[19], KRISTER T. SMITH[22], JESSICA M. THEODOR[23], RACHEL C. M. WARNOCK[3], AND MICHAEL J. BENTON[3]

Goal - to make calibration choices transparent and explicit

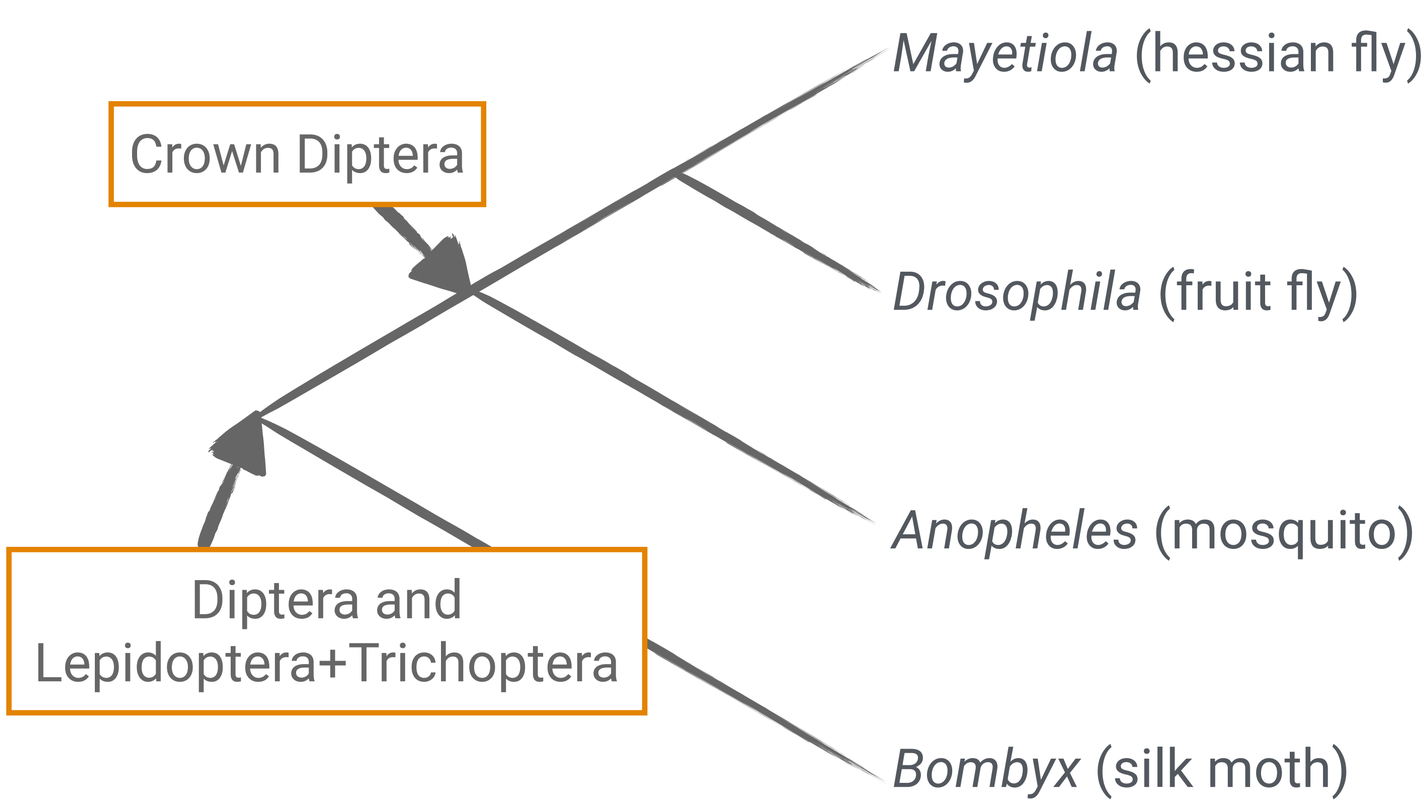
"The [soft] maximum constraint is established as older than all the oldest possible records, extending back to encompass a time when the ecologic, biogeographic, geologic, and taphonomic ***conditions for the existence of the lineage are met, but no records are known****."*

2.

Parham et al. 2011. Sys Bio

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

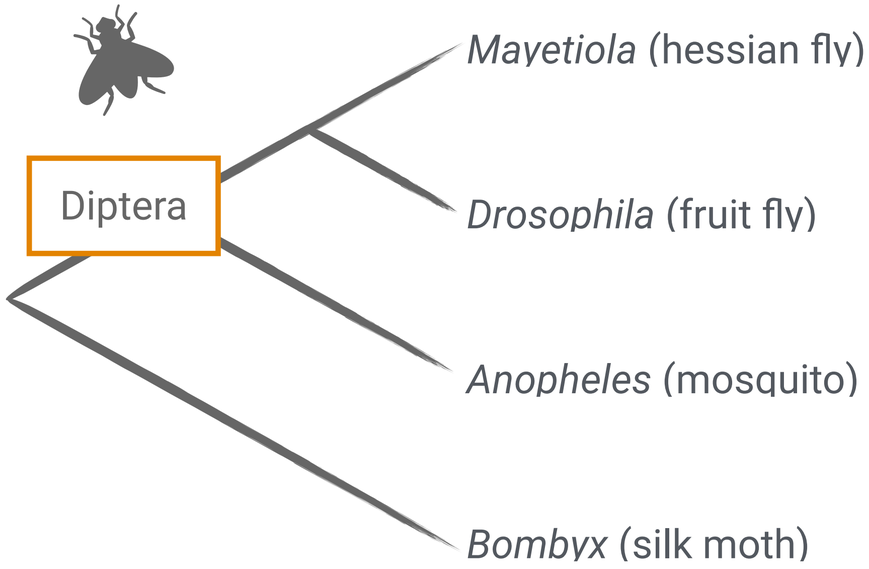
2.

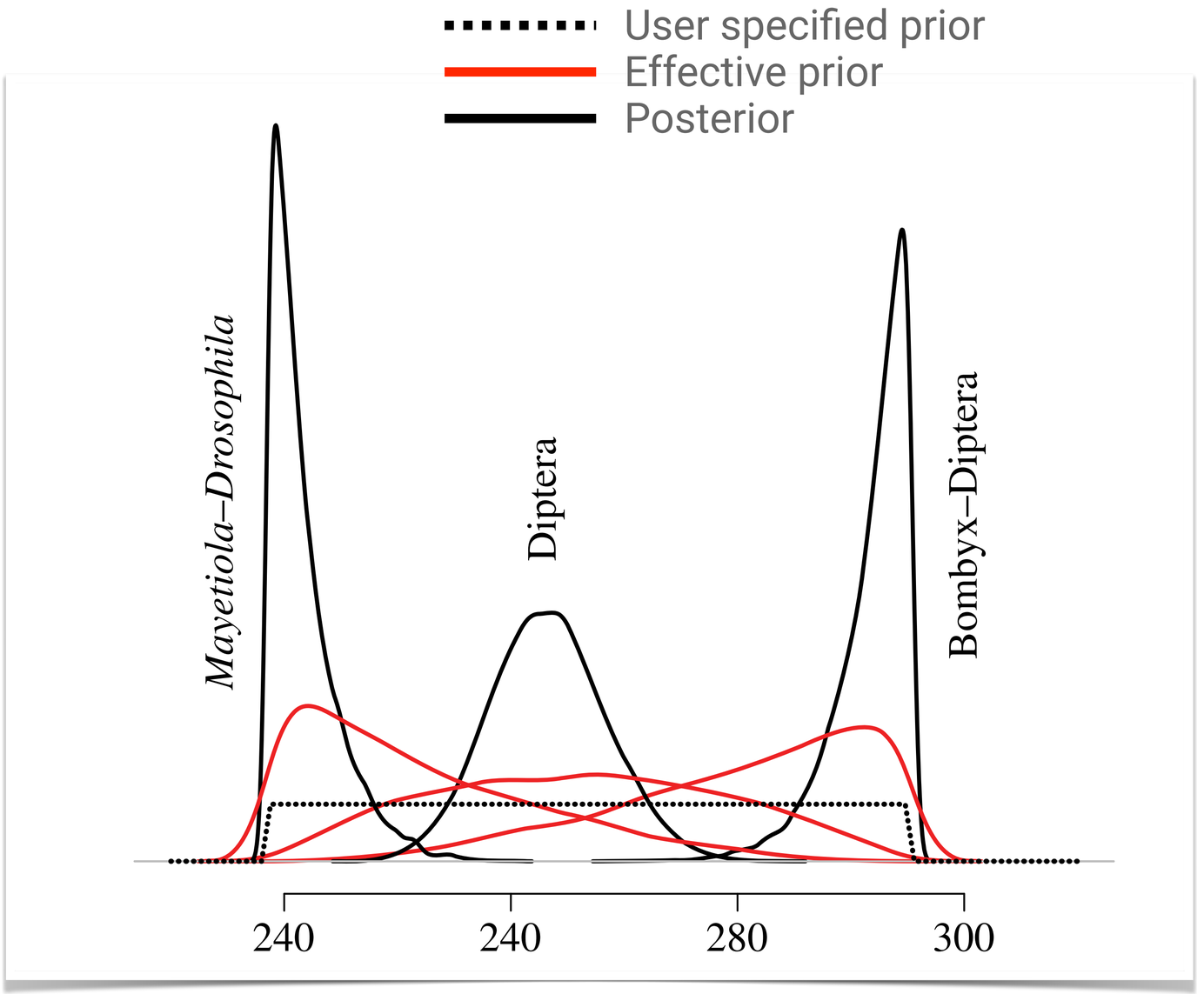
Warnock et al. 2011

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

3.

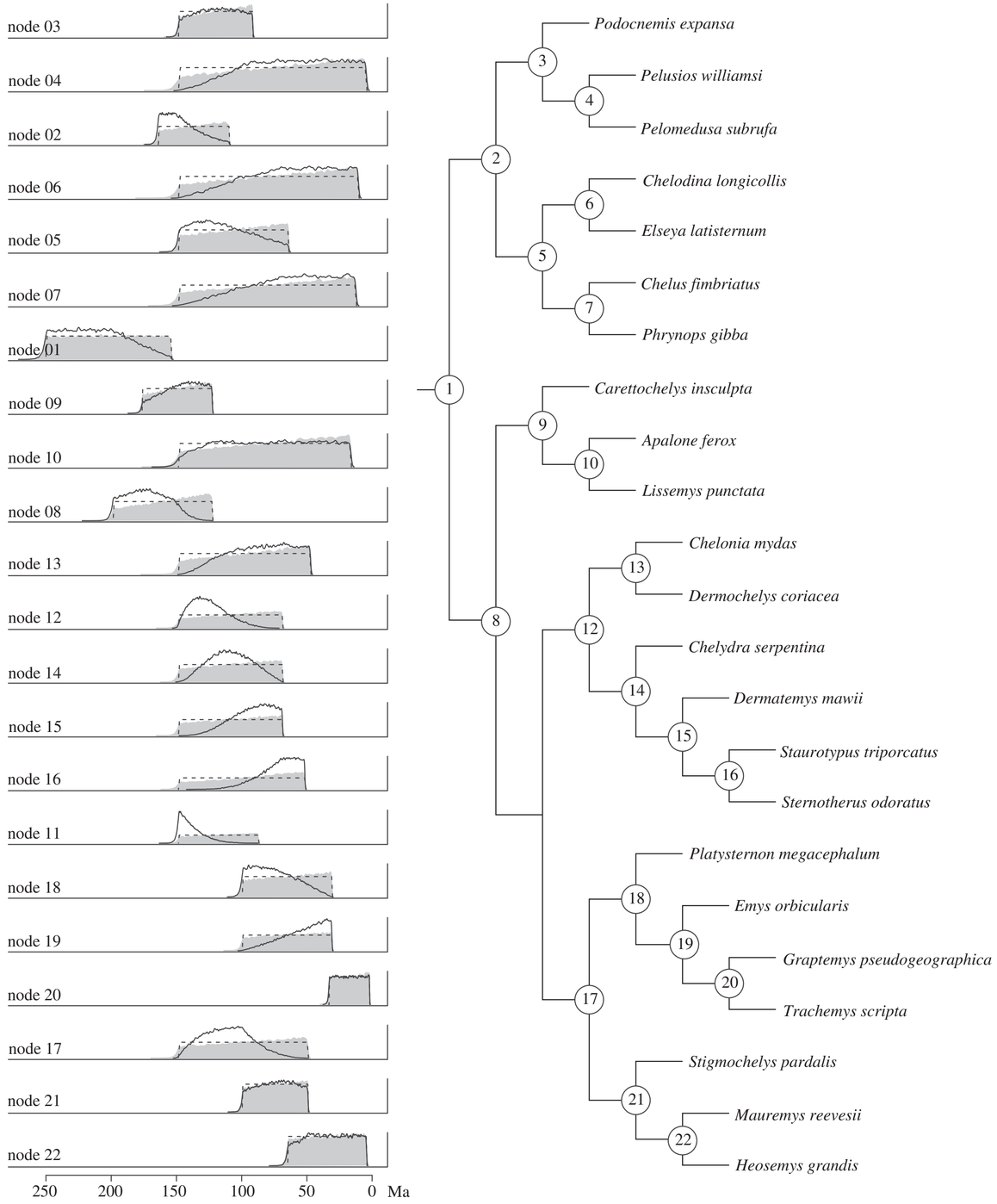




Posteriors look suspiciously like the effective priors….

3.

Warnock et al. 2011



3.

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

Warnock et al. 2015. *Proc B*

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

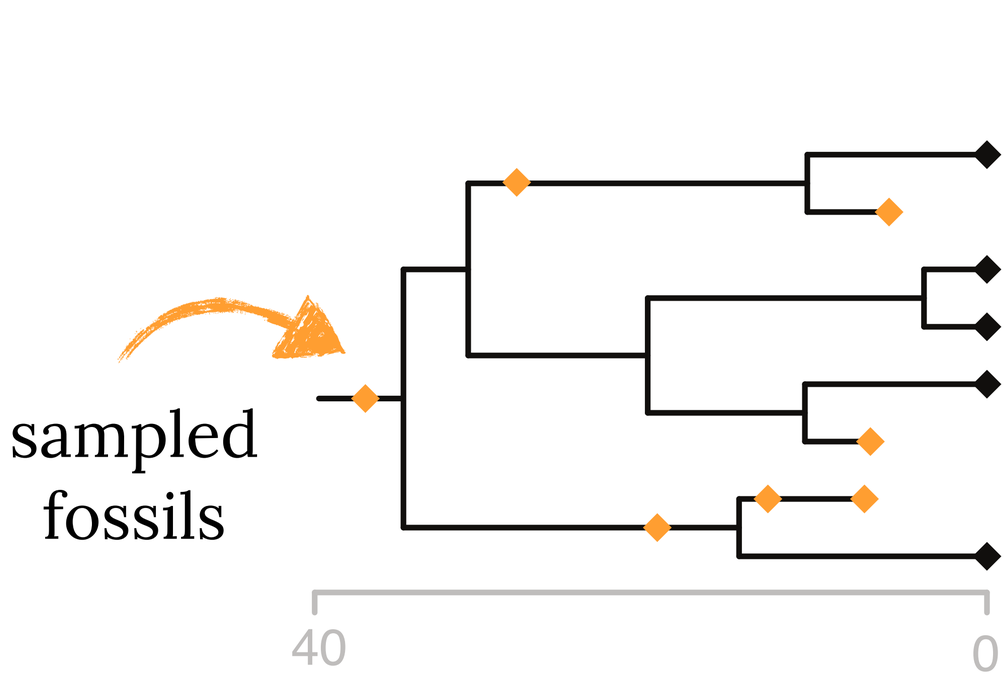
The calibration priors are difficult to specify objectively and can have a massive impact on the divergence times. They can also interact with each other and / or the birth-death process prior in unintuitive ways

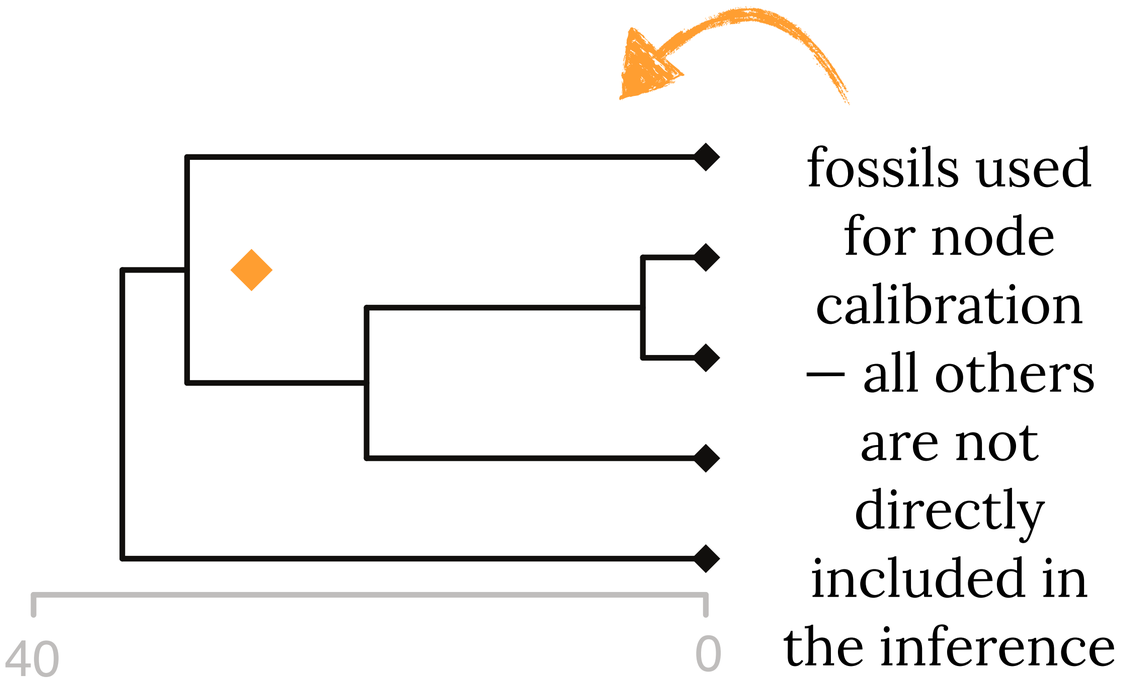
Some references on issues with specified vs effective priors
Yang and Rannala. 2006. MBE
Heled and Drummond. 2012. Sys Bio
Warnock et al. 2012, 2015

Node dating: potential issues

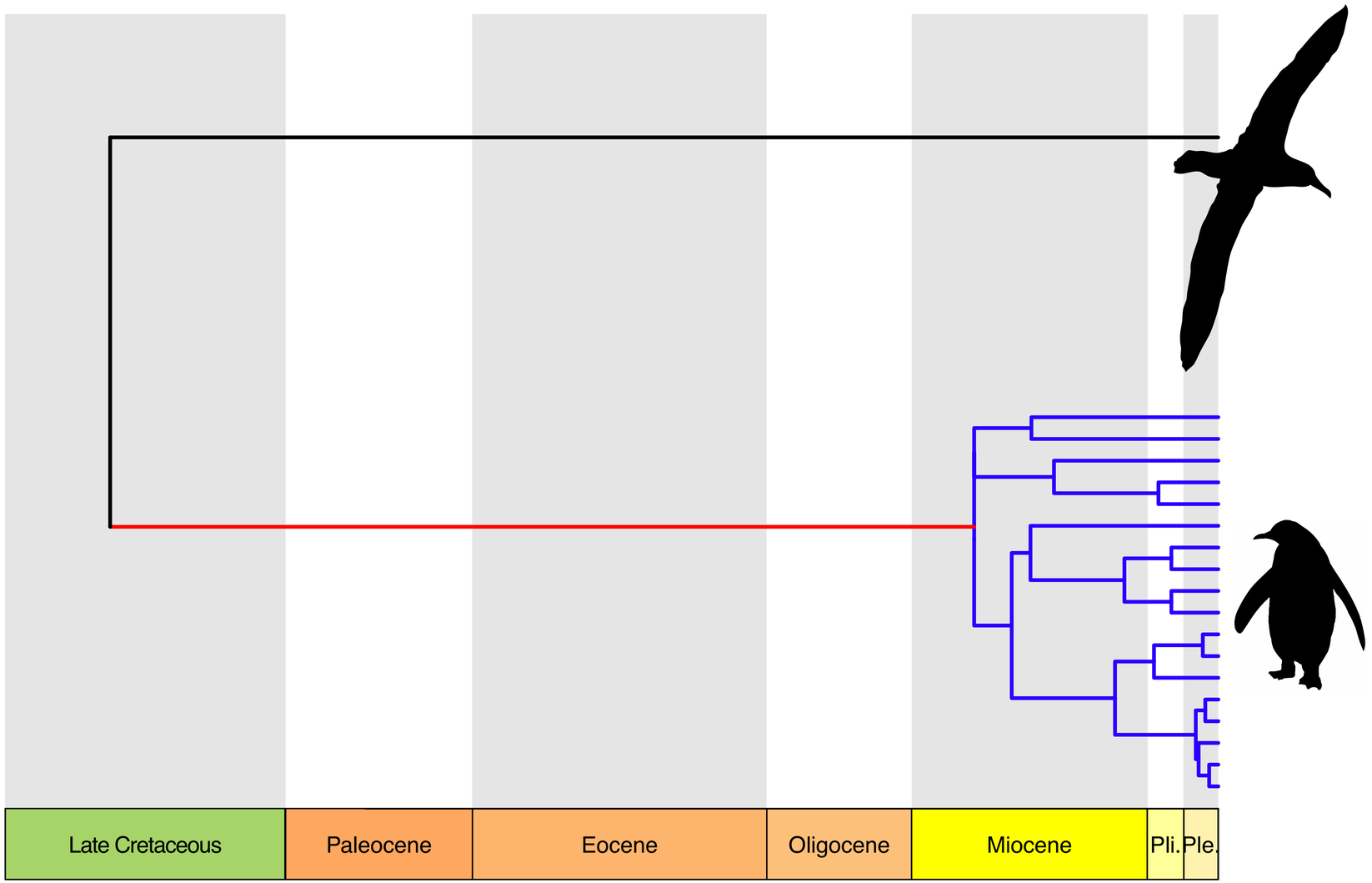
There are many!

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



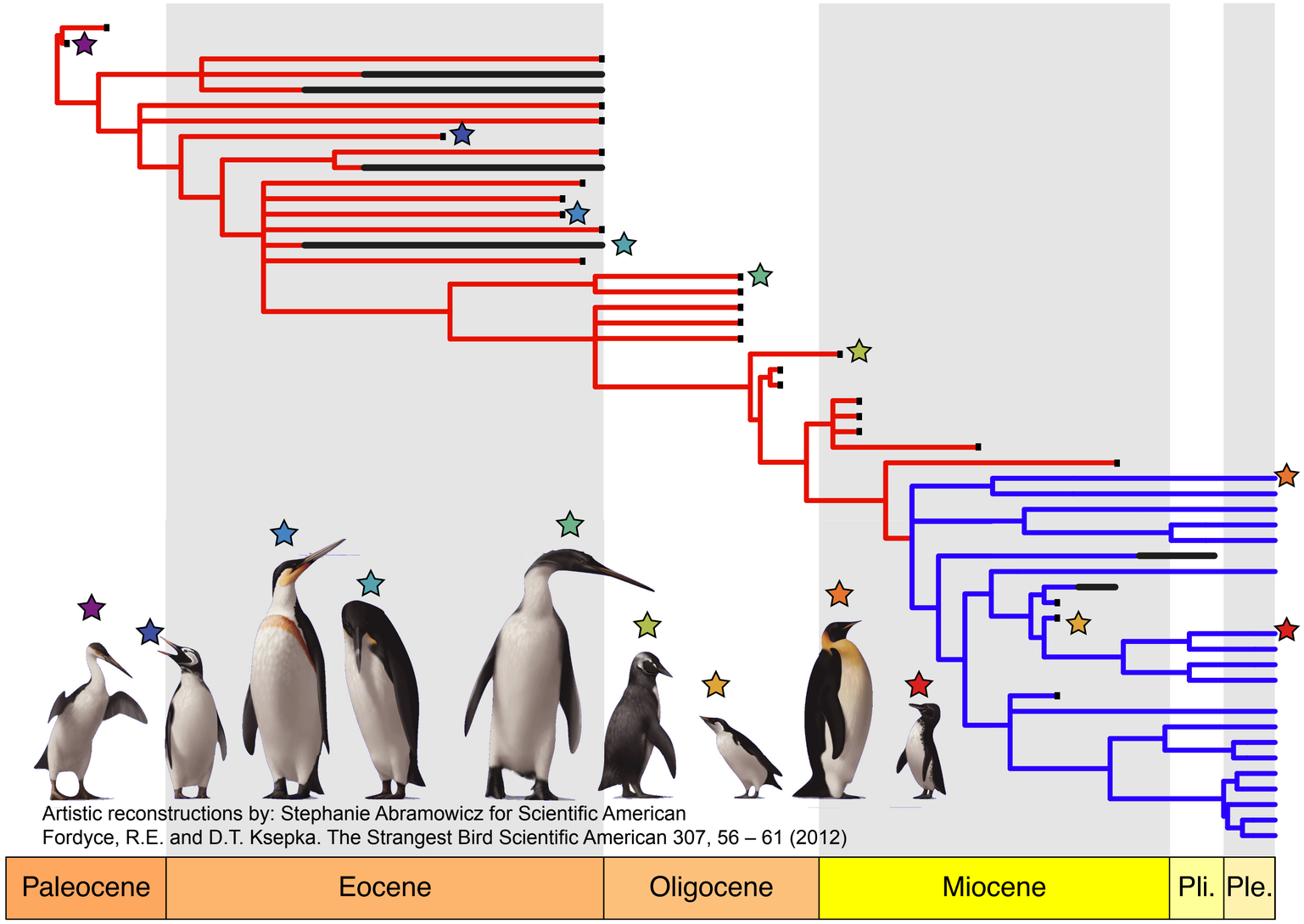


Example: living penguins



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

Example: living penguins

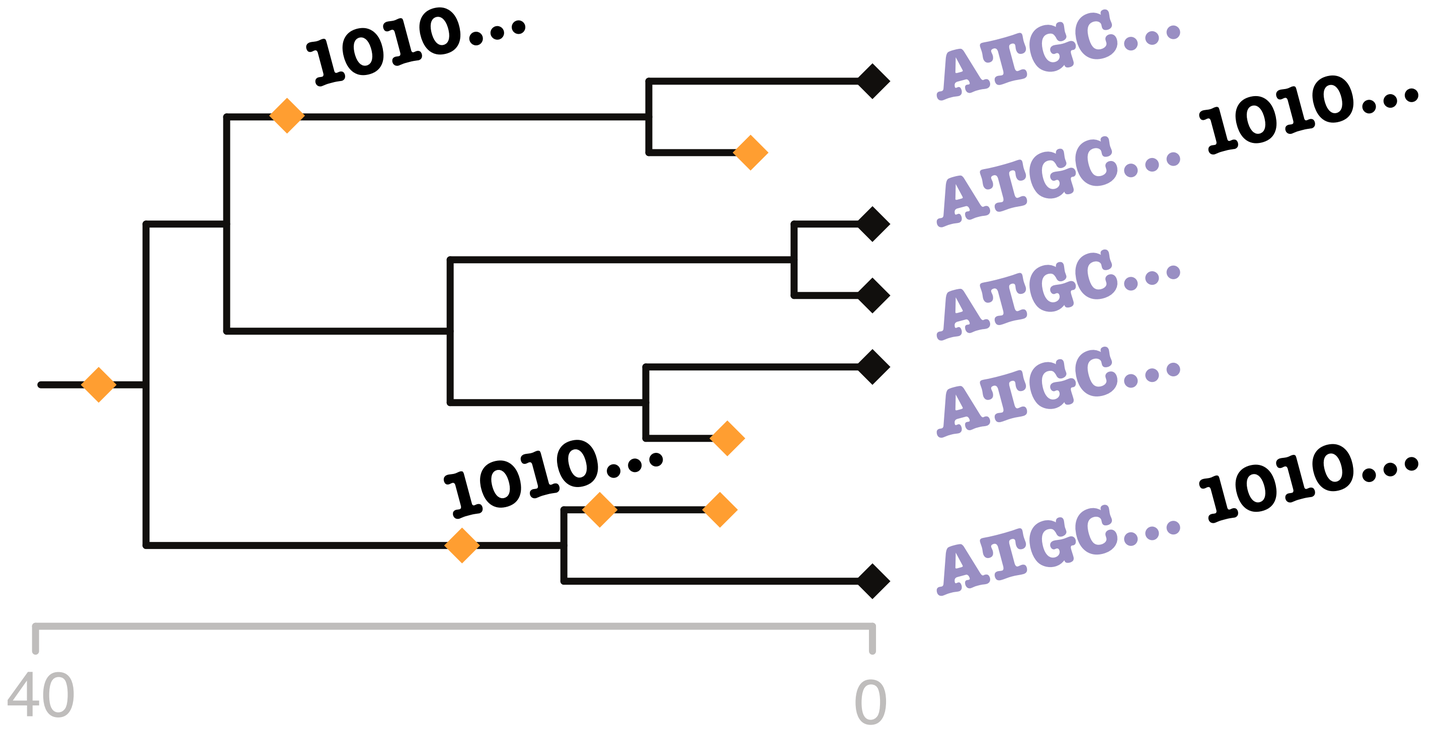


Artistic reconstructions by: Stephanie Abramowicz for Scientific American
Fordyce, R.E. and D.T. Ksepka. The Strangest Bird Scientific American 307, 56 – 61 (2012)

But penguins have a rich fossil record!

Total-evidence dating

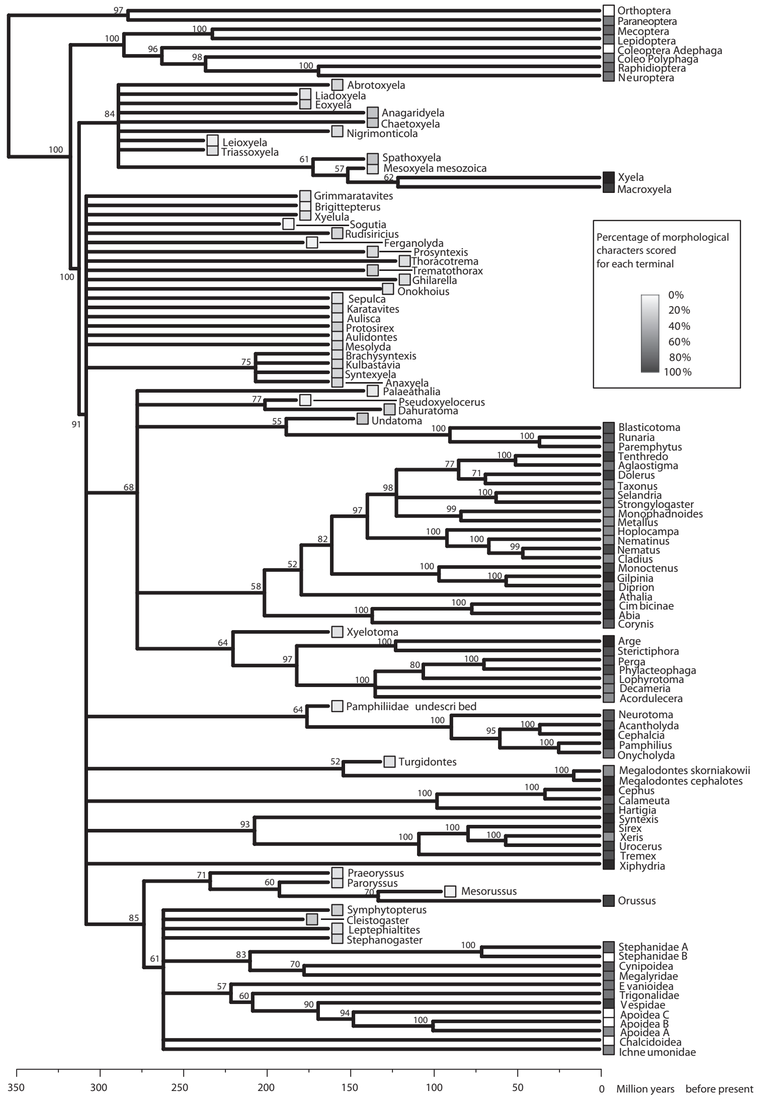
Tip-dating or “total-evidence” dating



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

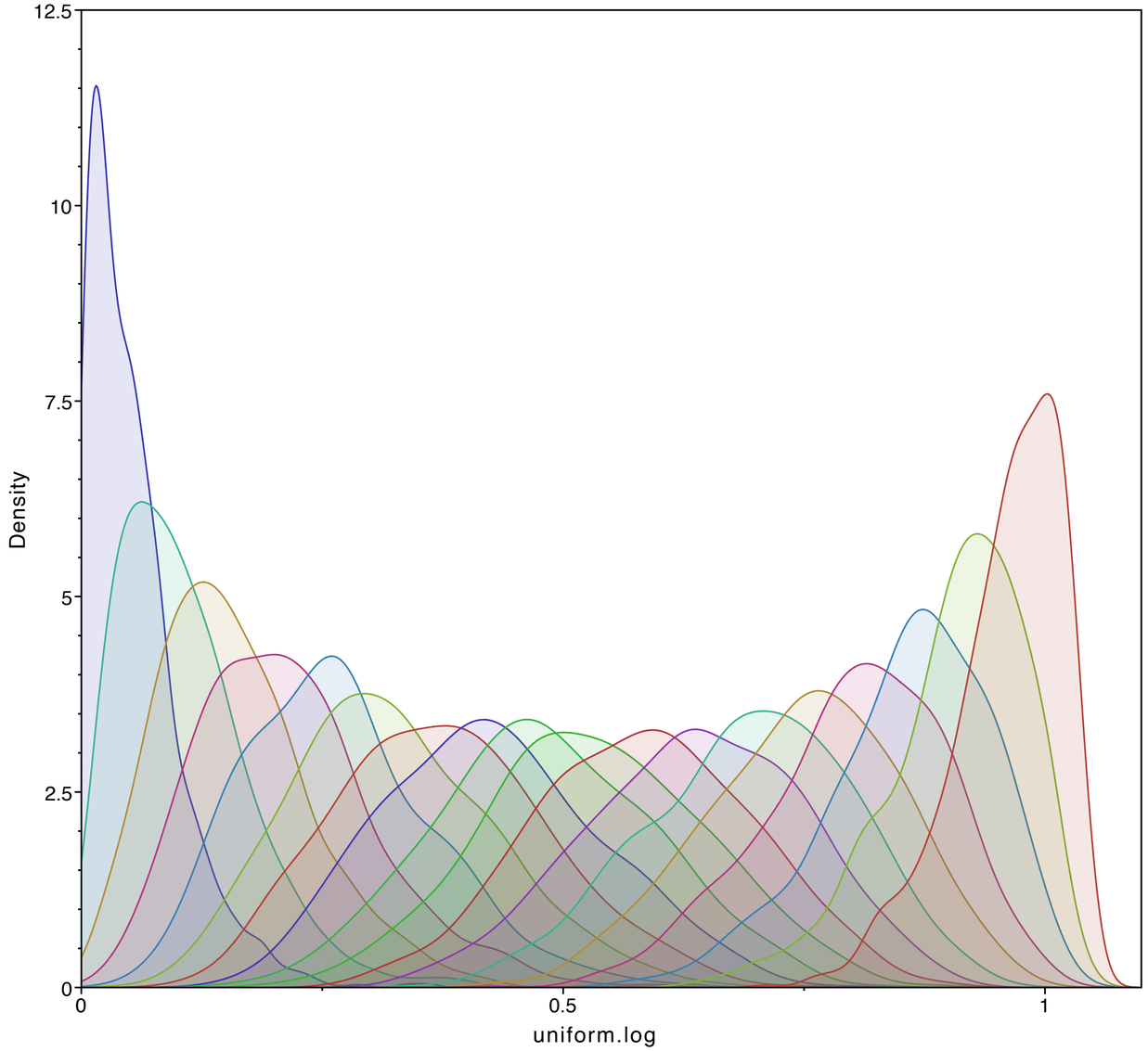


Dated tree of Hymenoptera

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



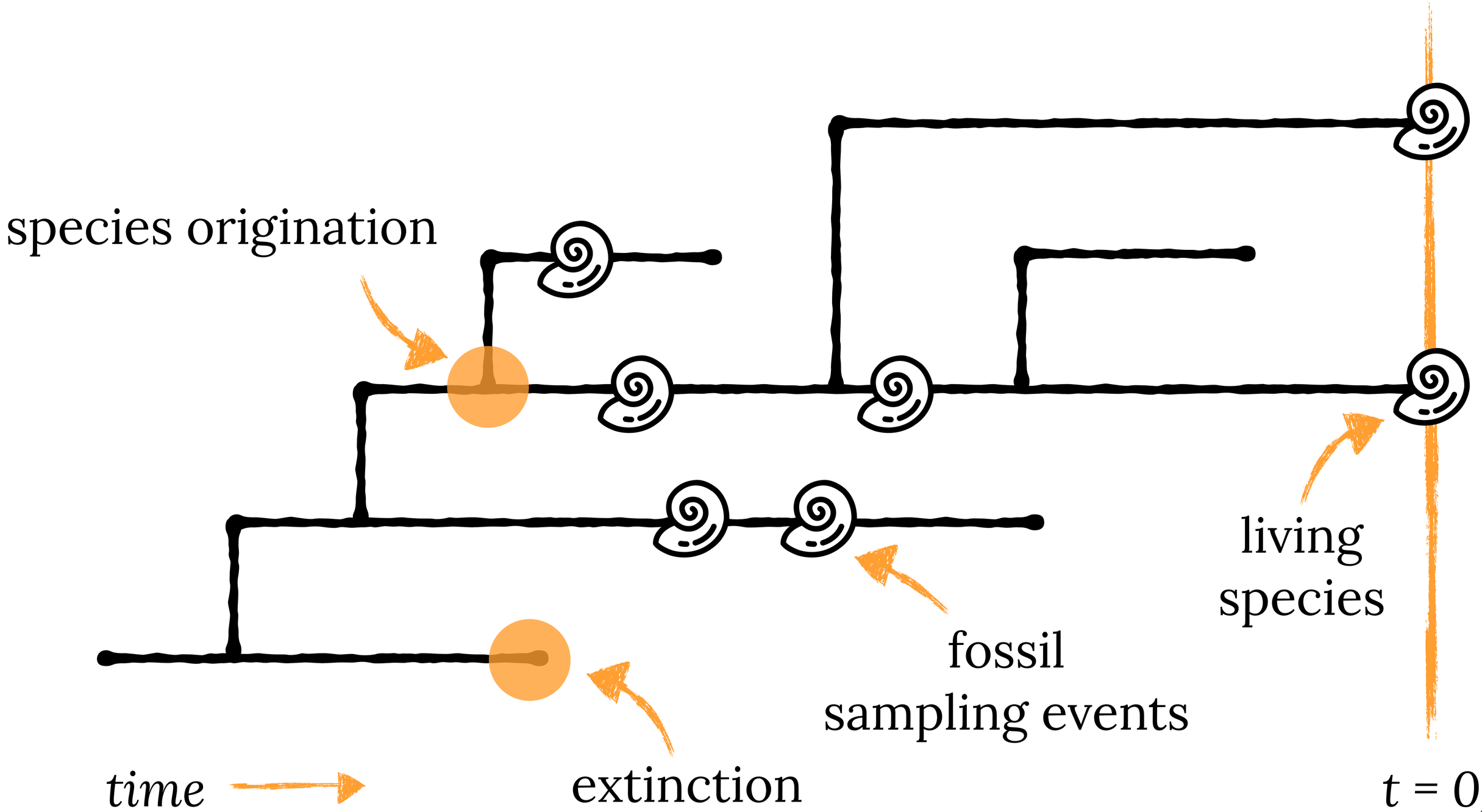
A uniform tree prior implies time till the next split is independent of how many lineages there are present

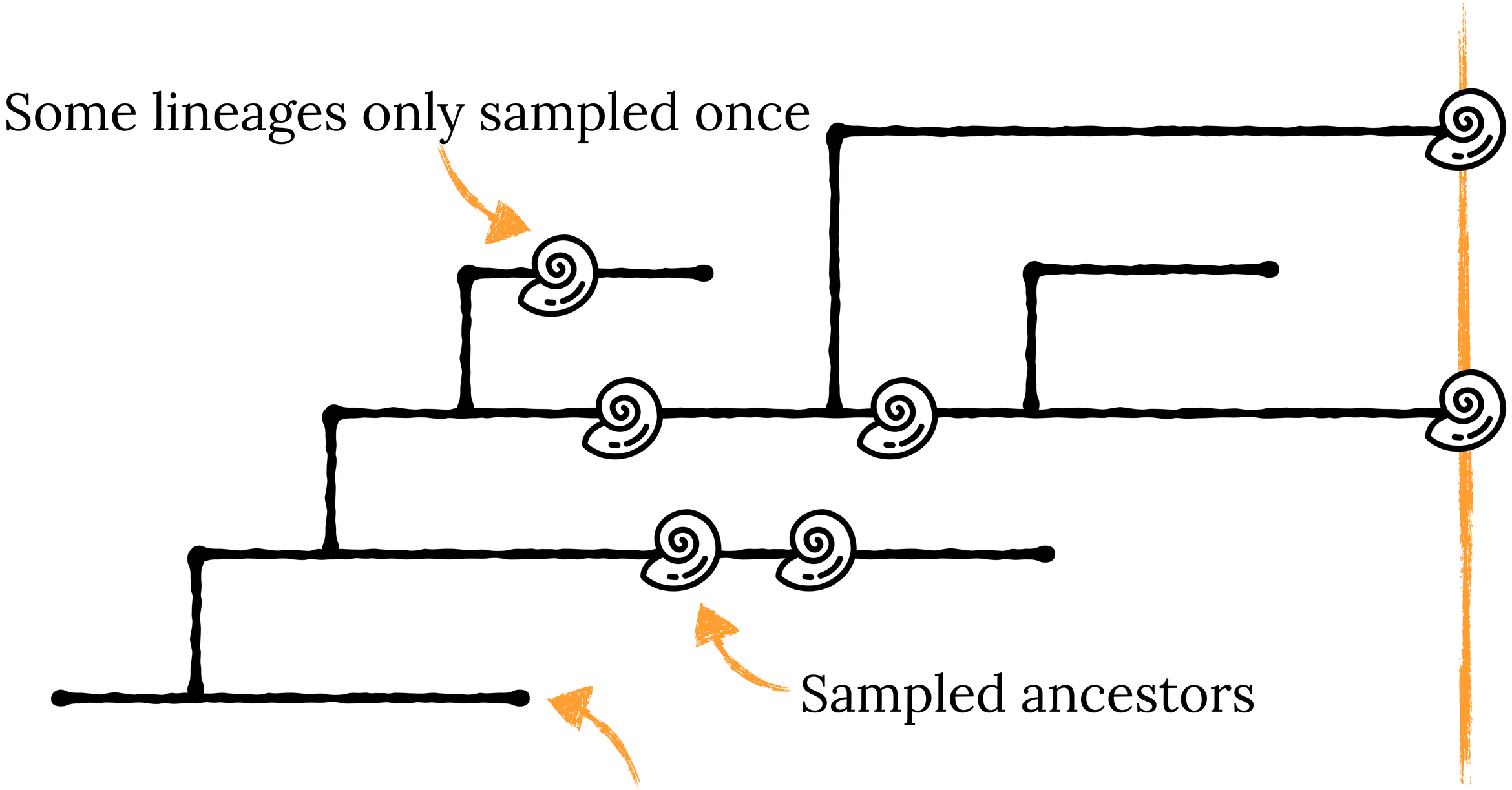
This is in contrast to birth-death processes, where more lineages mean a higher chance of observing a split in one of these lineages

Uniform tree priors - why not use them?
Remco Bouckaert

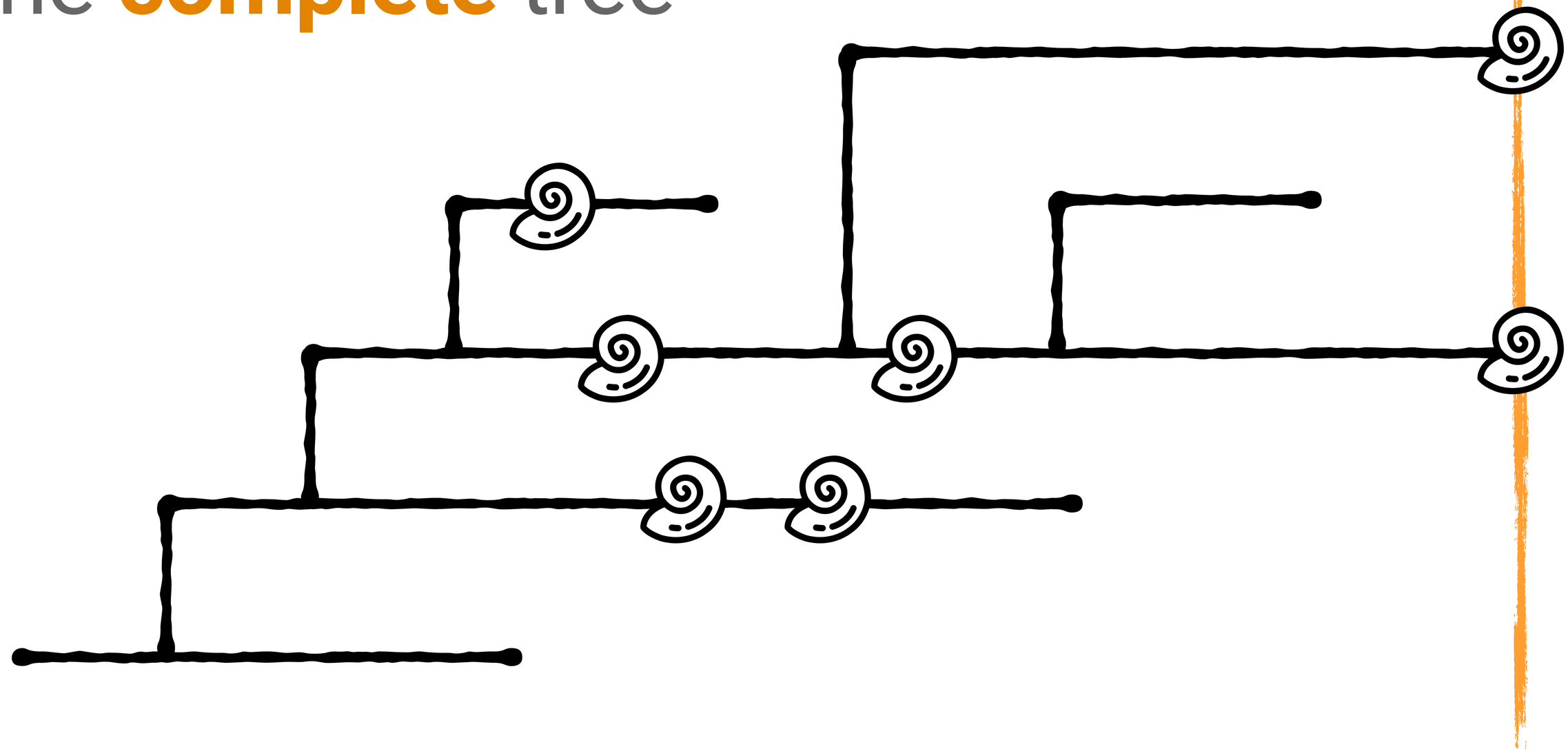
What does a generating prior for the fossil record look like?

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
$t = 0$

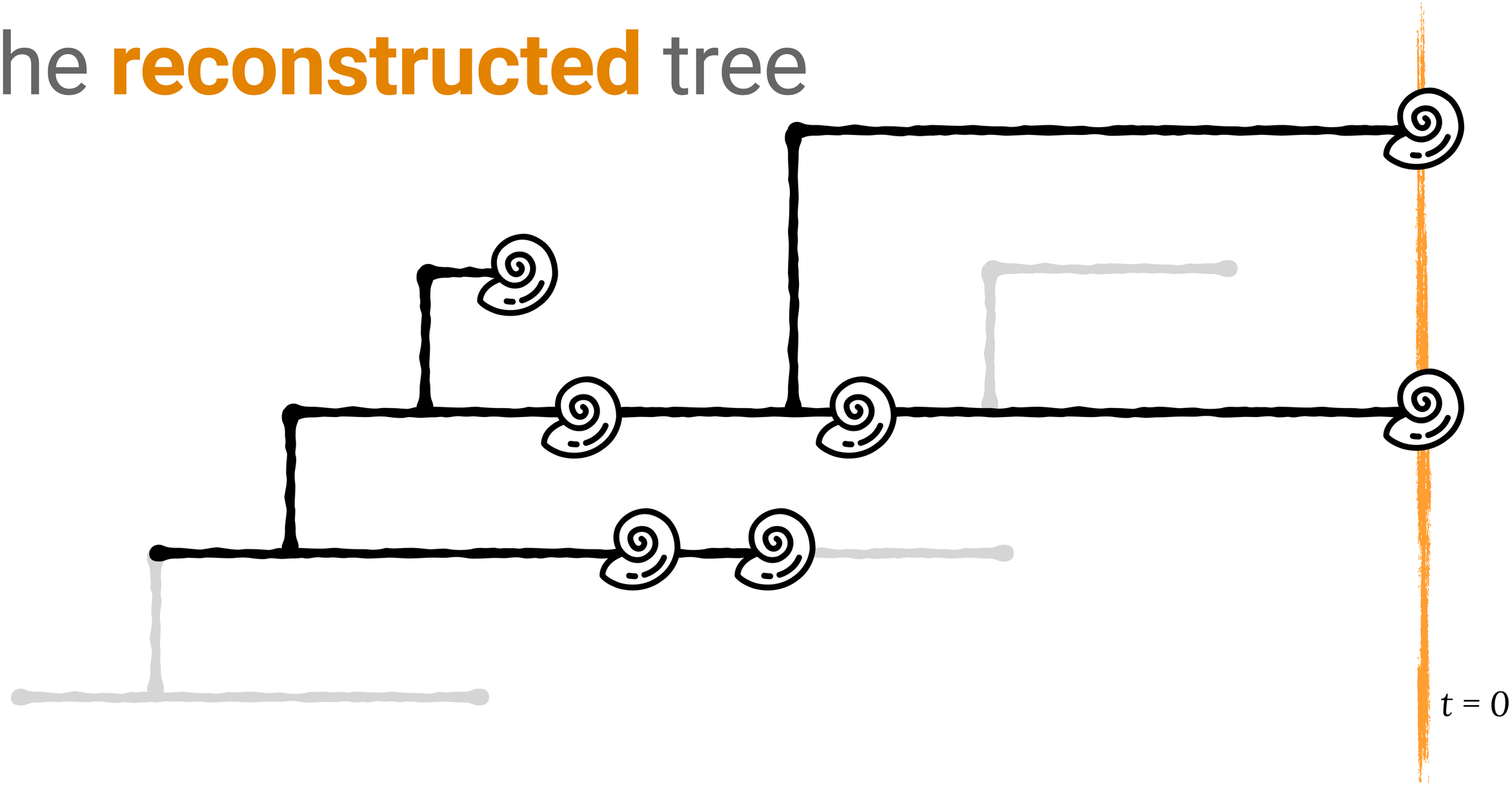
The **complete** tree



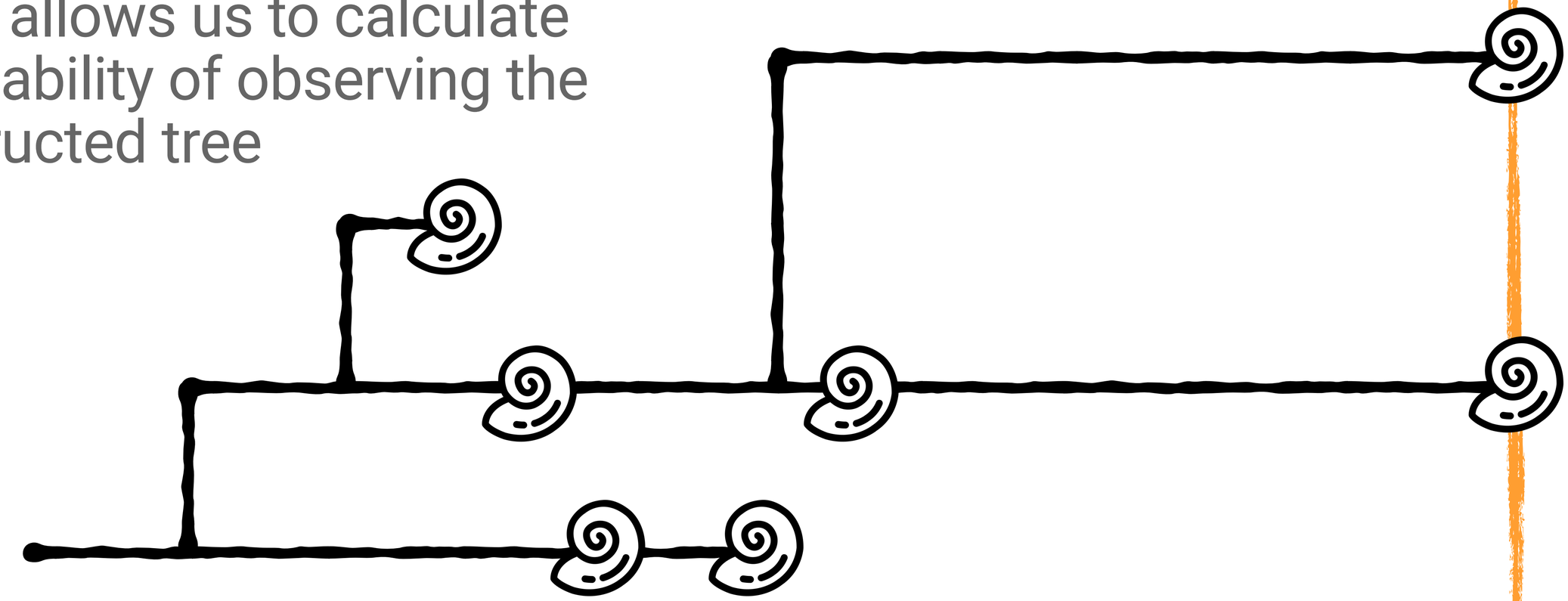
$time$ ⟶

$t = 0$

The **reconstructed** tree



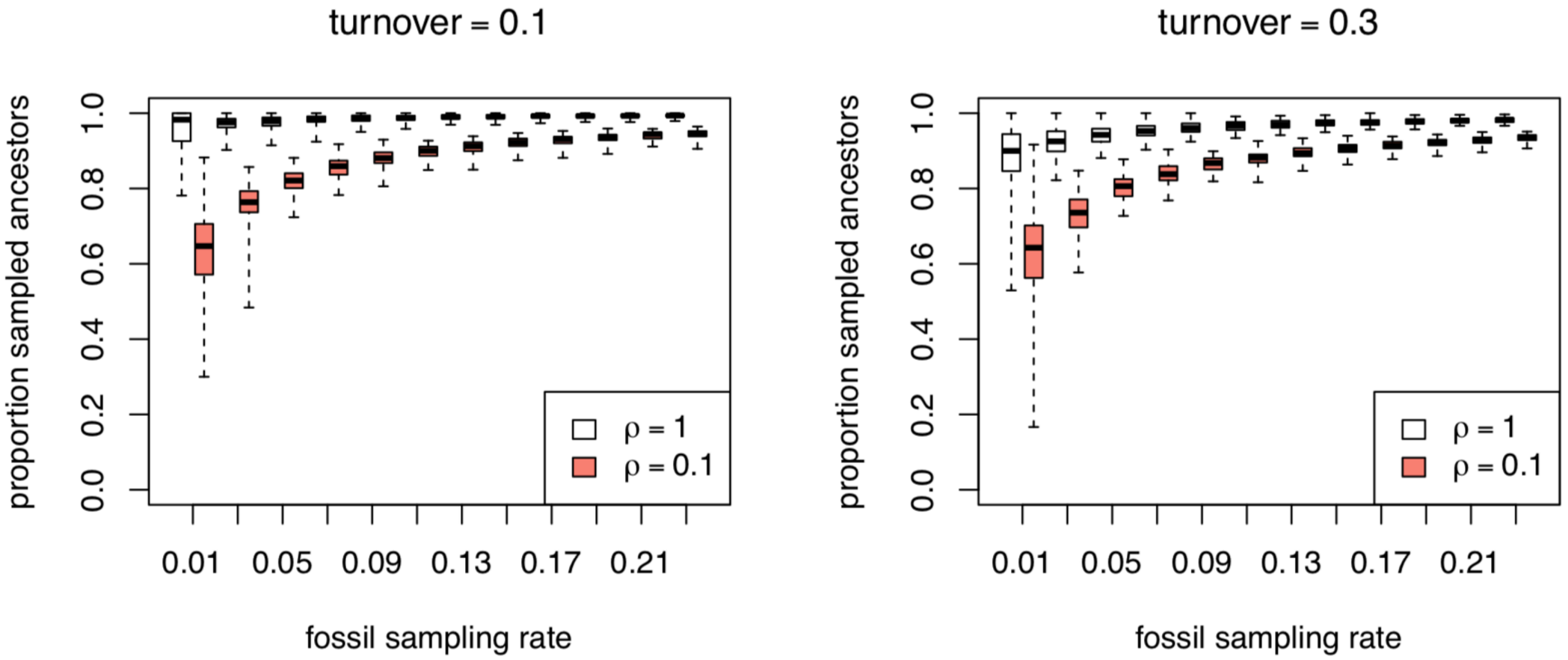
The **fossilised birth-death (FBD) process** allows us to calculate the probability of observing the reconstructed tree



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

Sampled ancestors

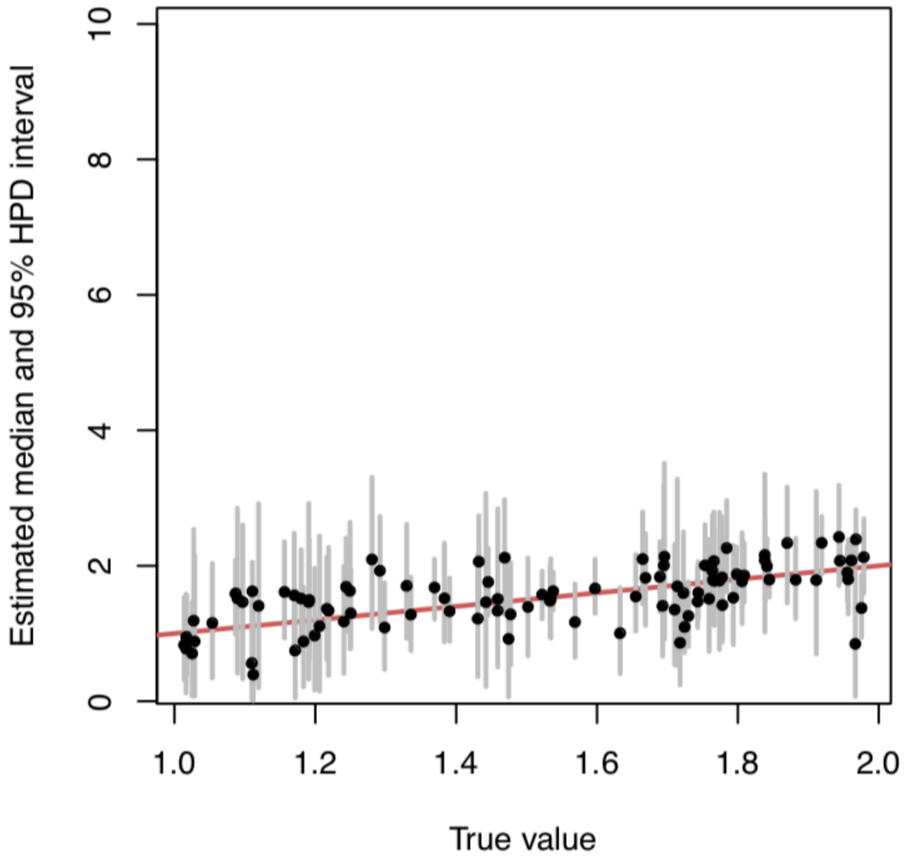
The proportion increases with higher turnover (birth - death) or higher sampling

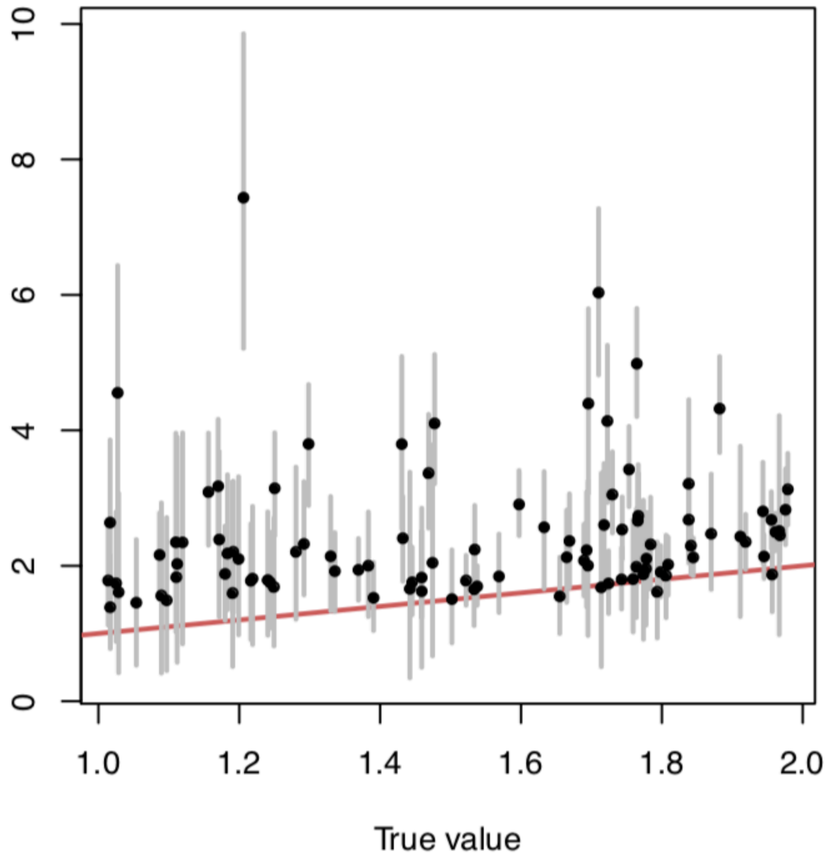


Walker, Heath. 2020. Phylogenetics in the Genomic Era

Sampled ancestors

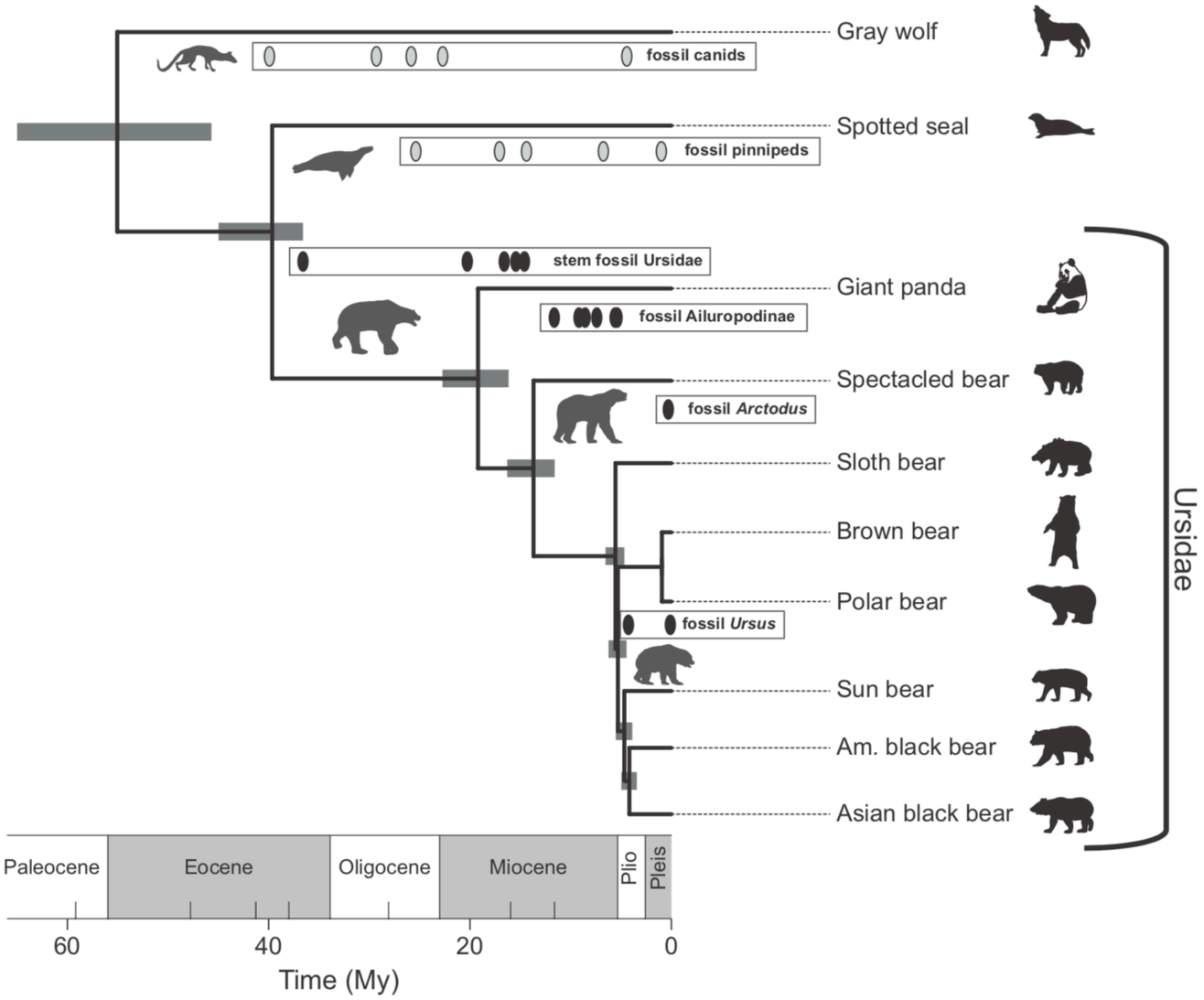
Ignoring sampled ancestors can lead to inaccurate parameter estimates





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)

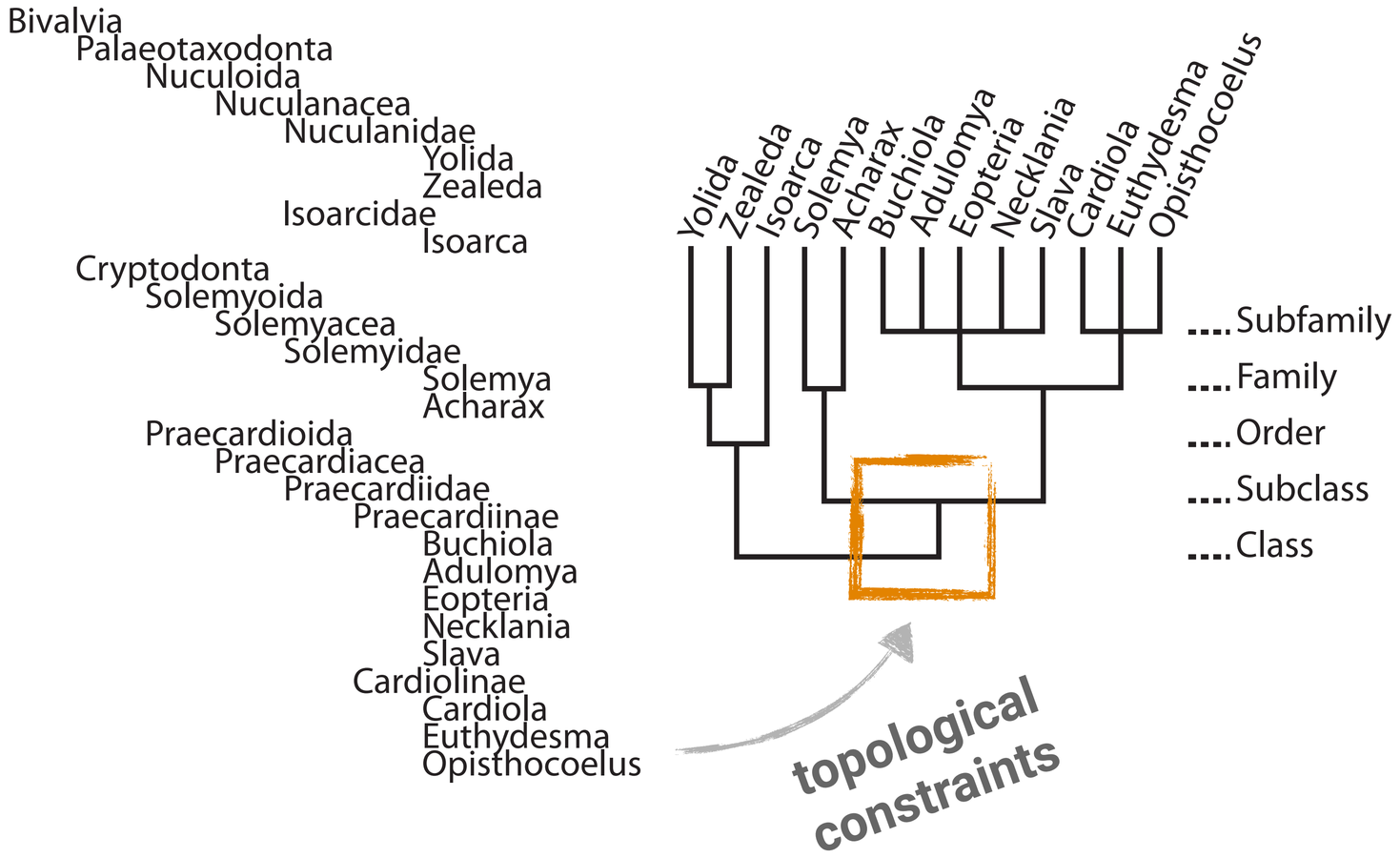
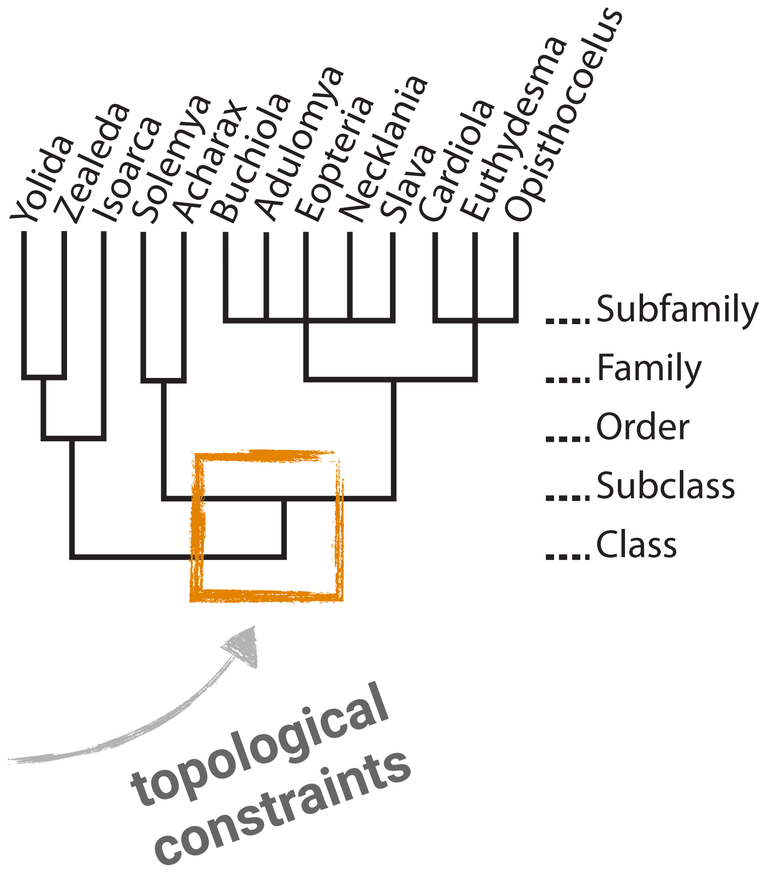


Image source Soul & Friedman (2015)

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



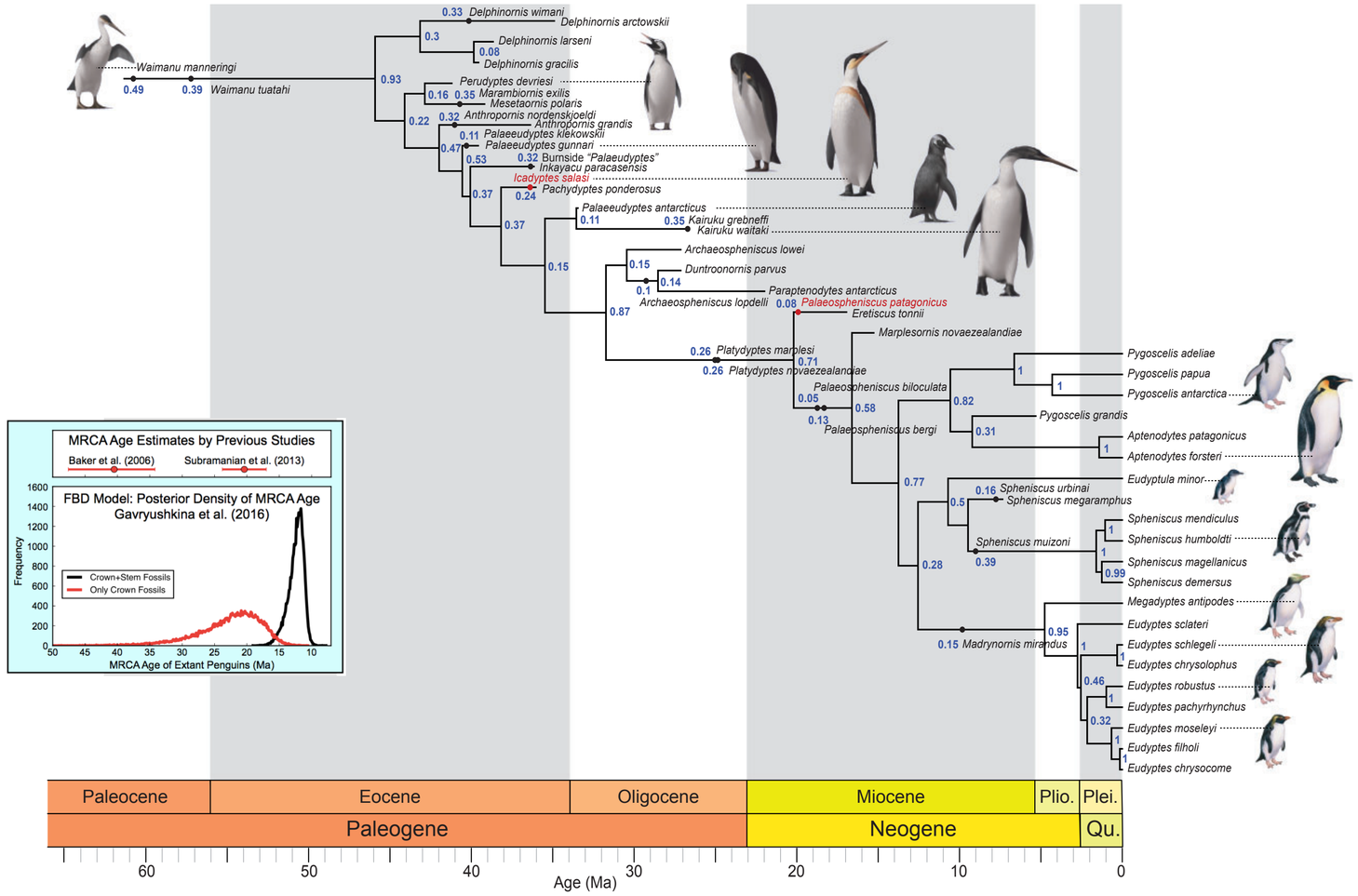
ATAT...
TCAC...
????...

OR

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

Image source Soul & Friedman (2015)

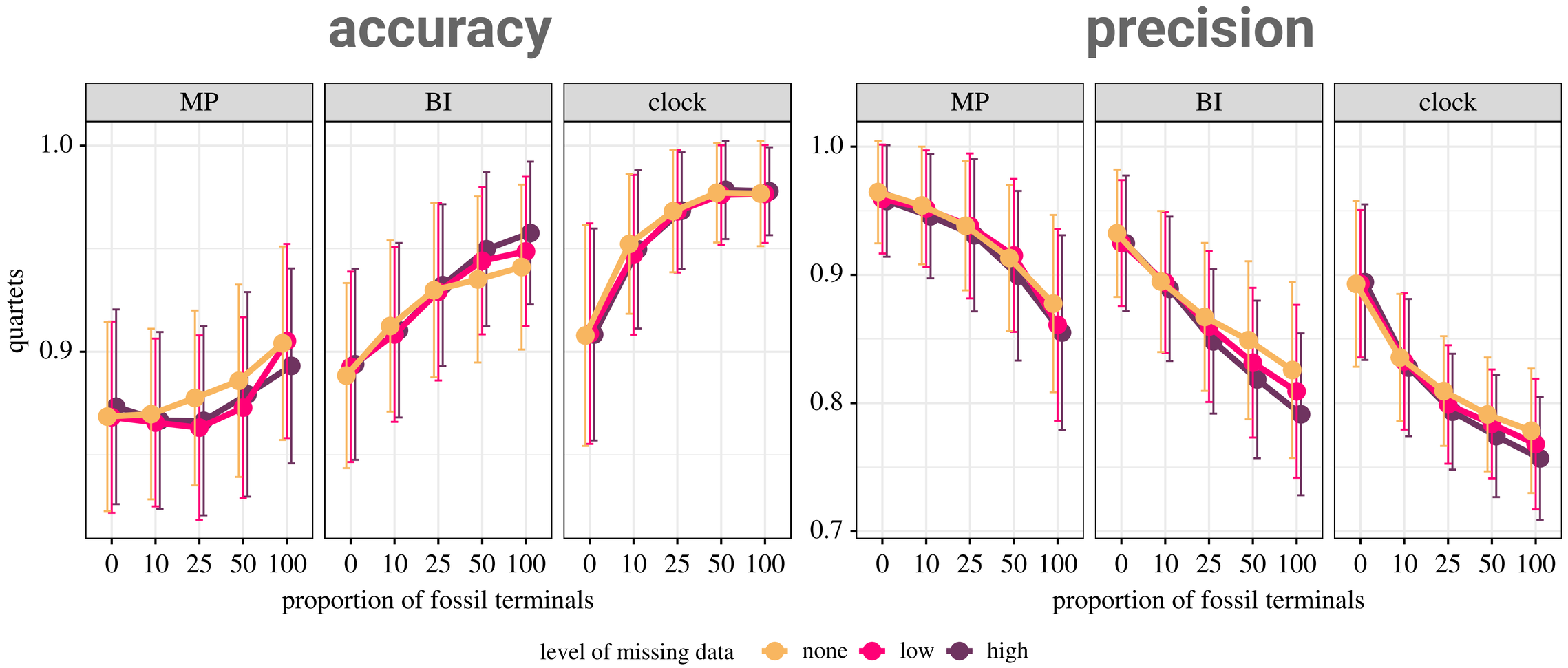
Time calibrated tree of living and fossil penguins



First application of total evidence dating using the FBD model

Fossils are incorporated using character data

Gavryushkina et al. (2016)



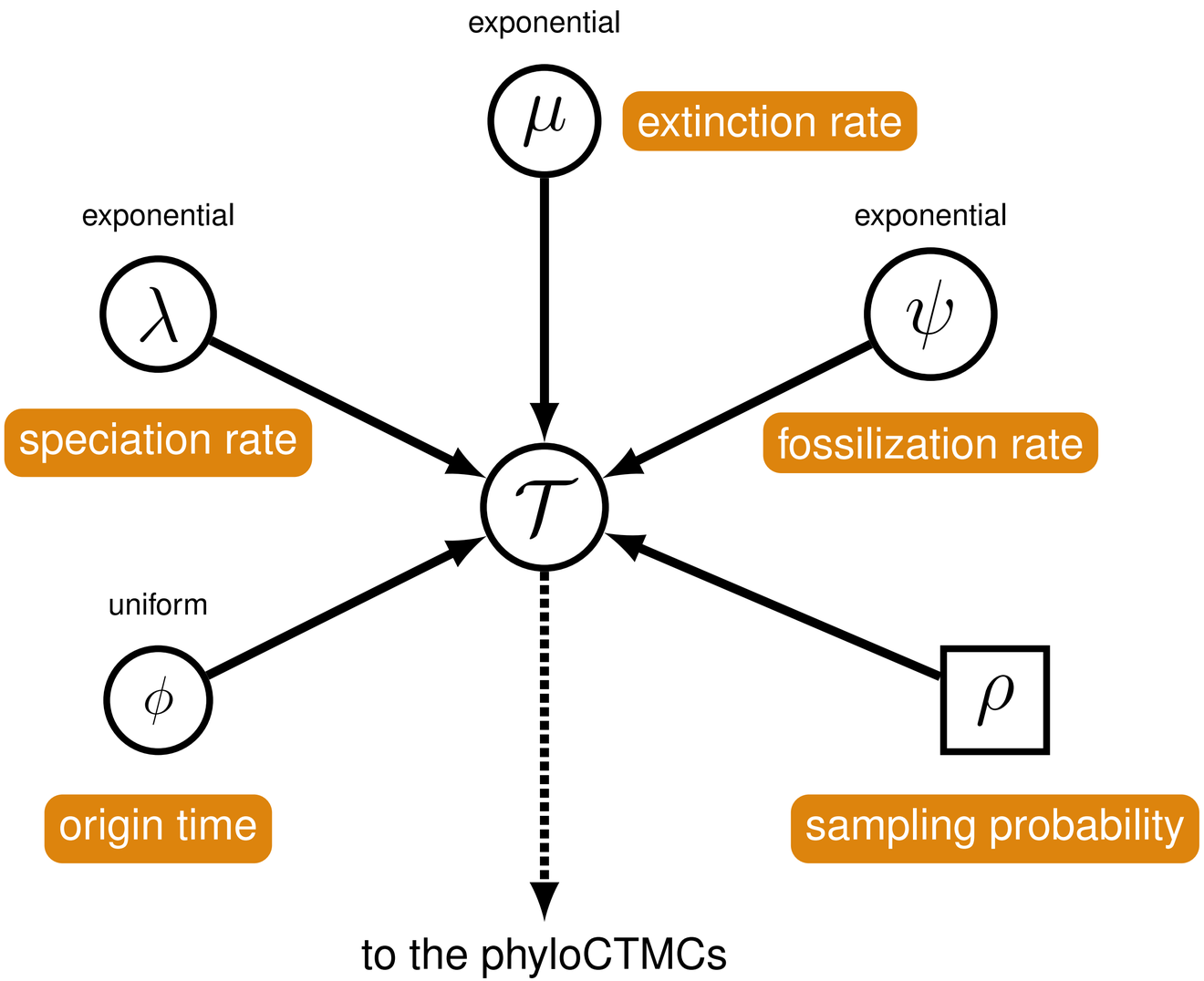
Fossils improve phylogenetic analyses of morphological characters
Koch, Garwood, Parry. 2020. Proc B

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

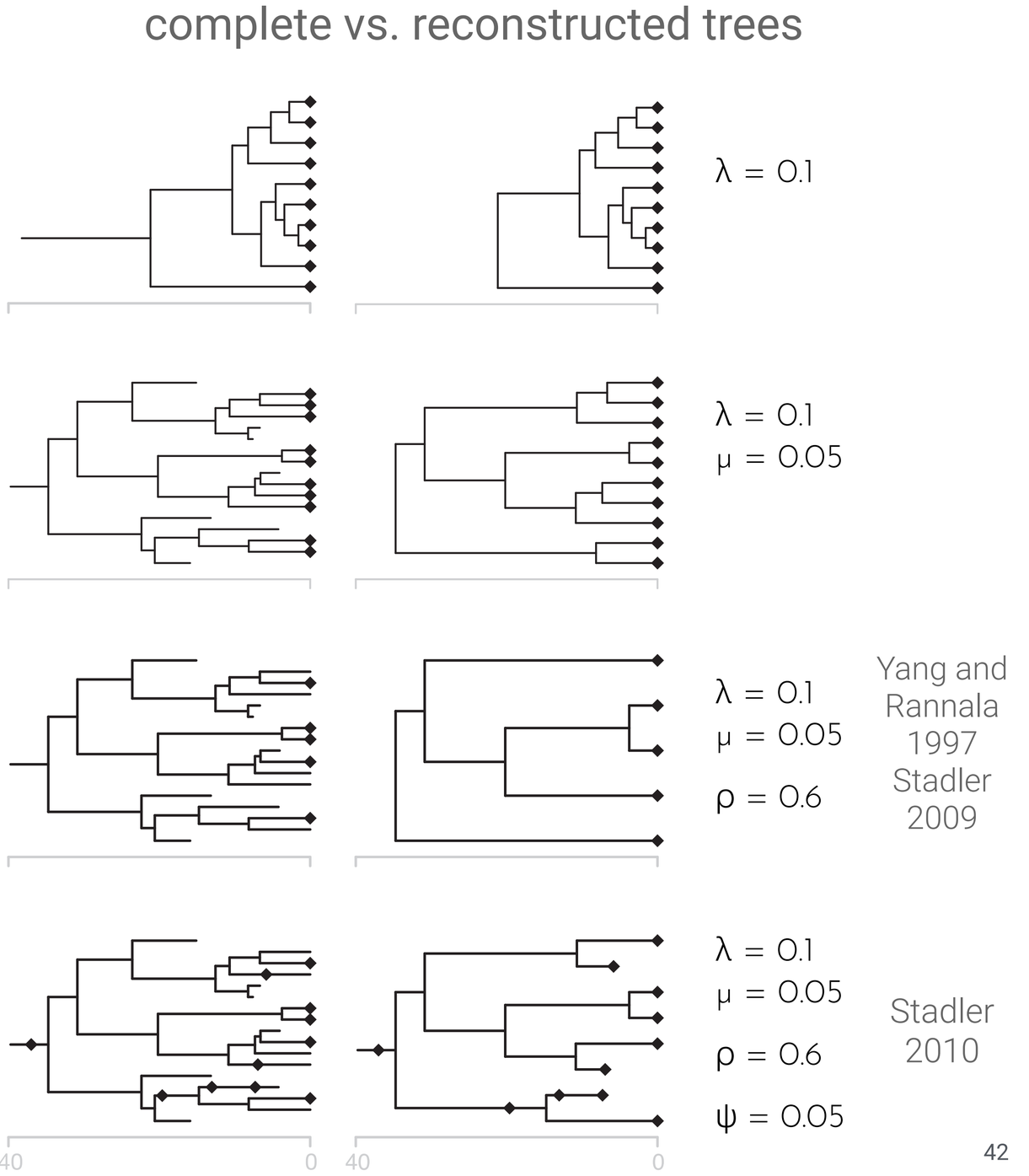


Relationship to (some) other birth-death process models

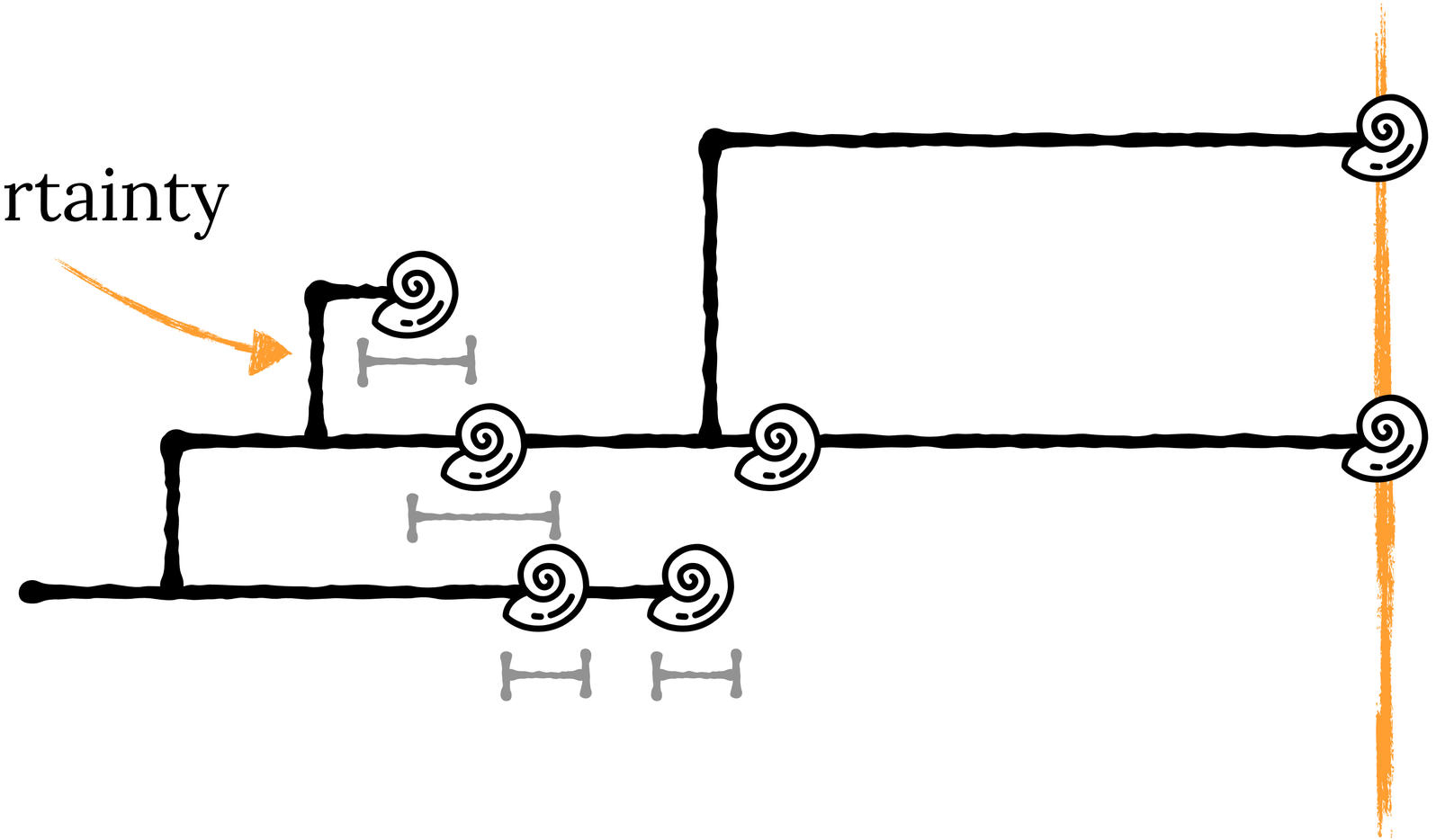
These models are special cases of the FBD process, with fossil sampling (ψ) = zero.

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

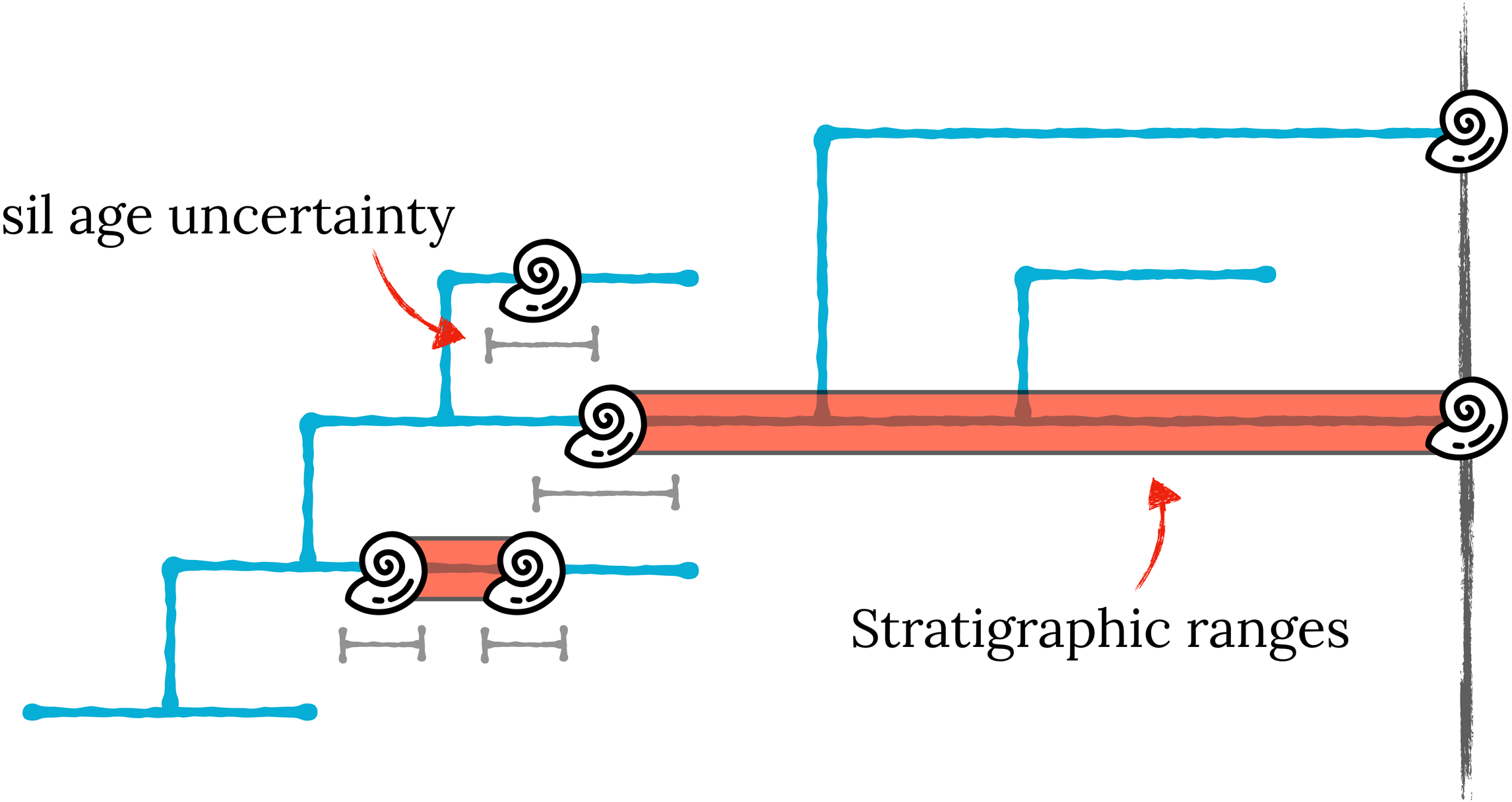
Stadler et al. 2012
See also: *Stadler and Yang* 2013



Sample age uncertainty

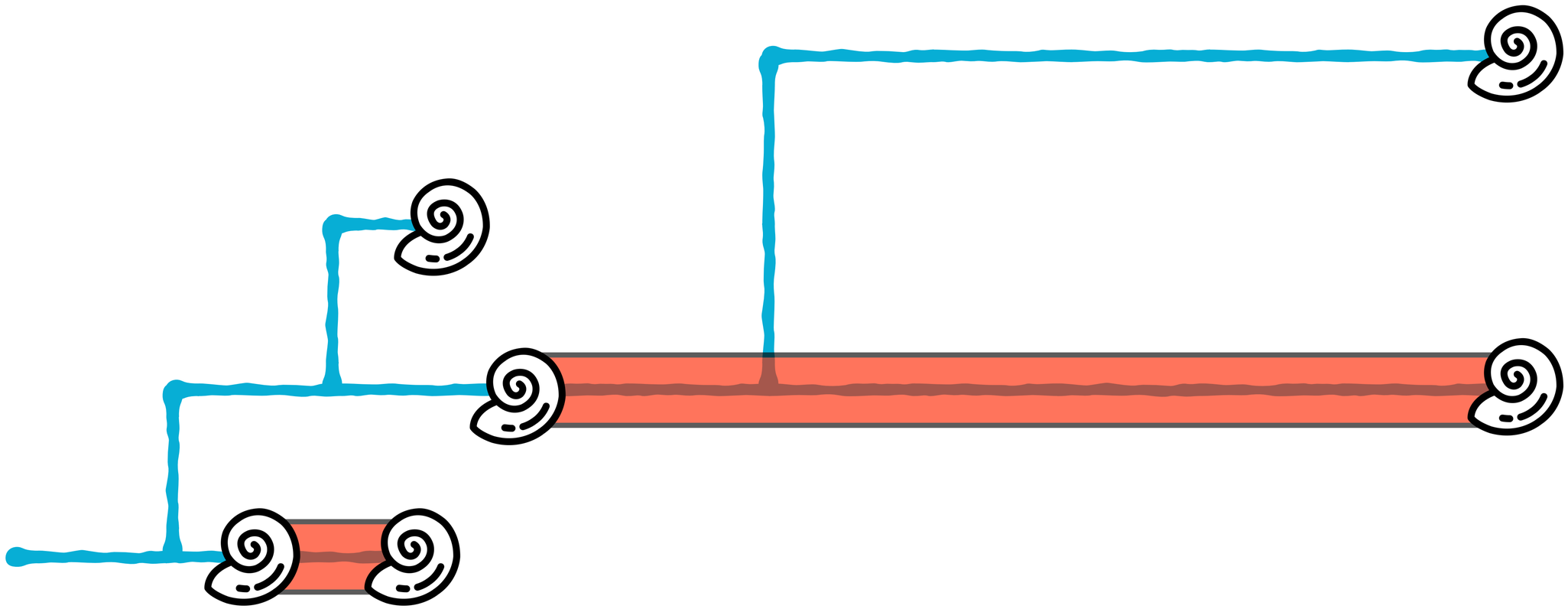


Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology in Time Calibrated Tree Inference
Barido-Sottani et al. 2018, 2020
Putting the F in FBD analyses: tree constraints or morphological data? Barido-Sottani et al. 2023

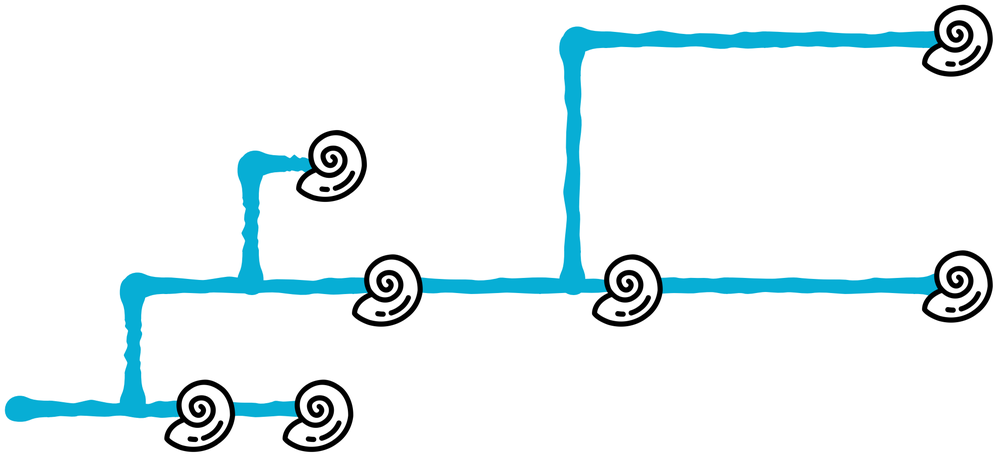


The inseparability of sampling and time and its influence on attempts to unify the molecular & fossil records
Hopkins et al. 2018. Paleobiology

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

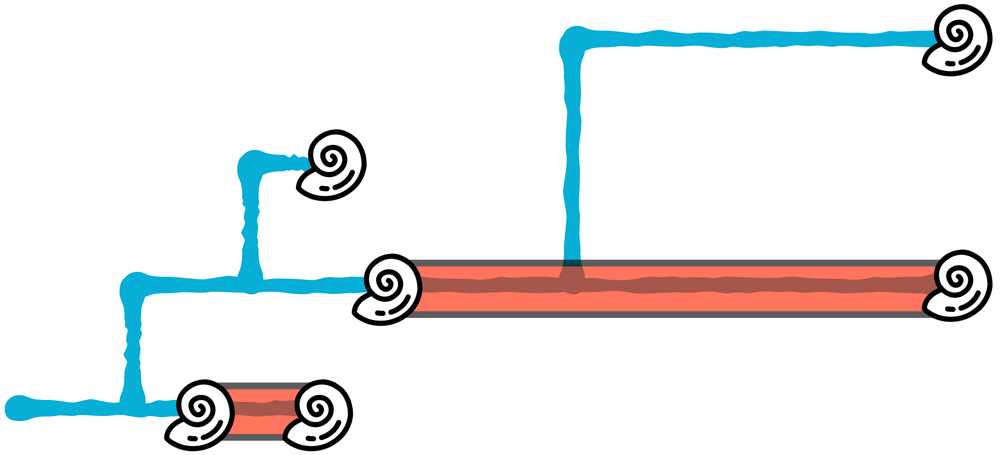


FBD process for analysis of specimen level data

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

The FBD range process for analysis of stratigraphic ranges

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

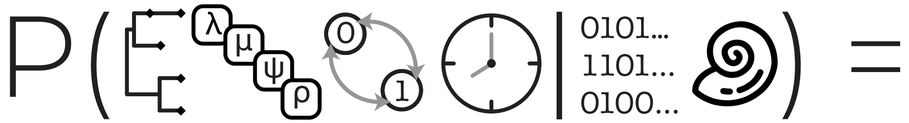


Exercise

Phylodynamics

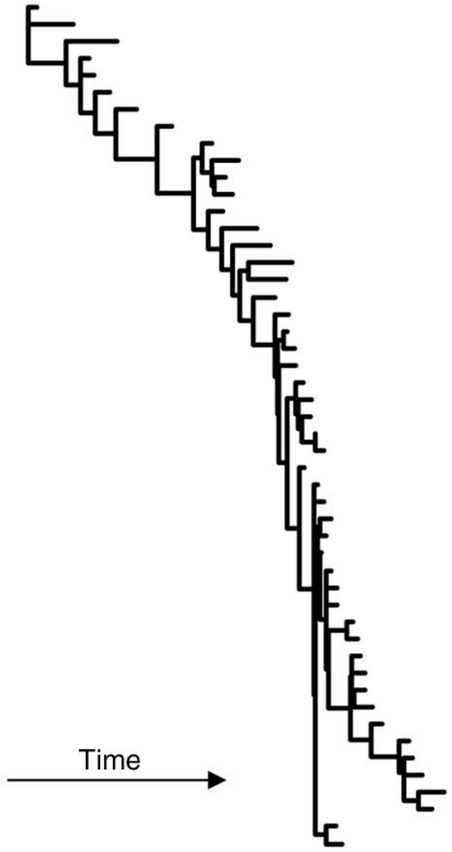
Diversification rate estimation

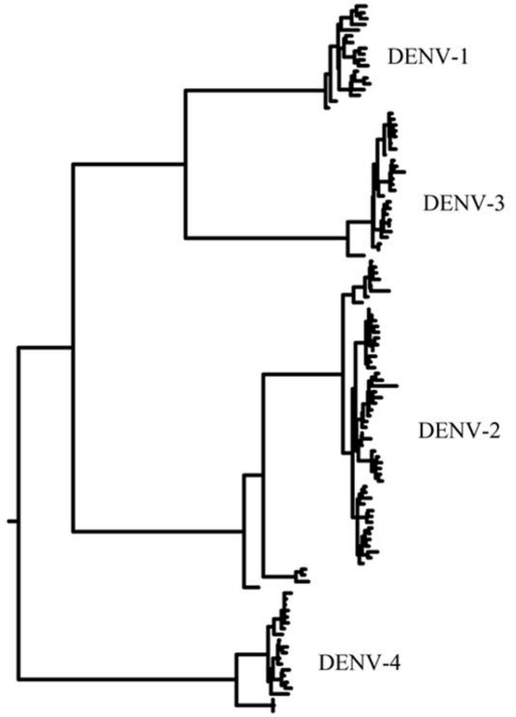
Bayesian divergence time estimation

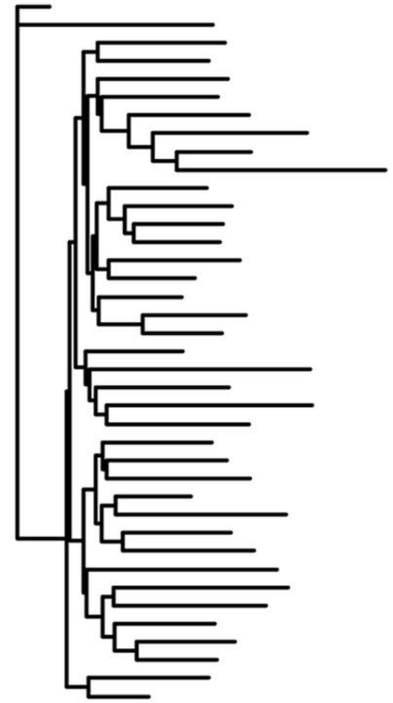


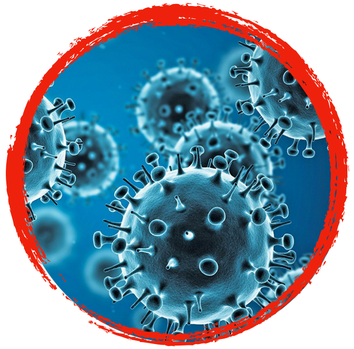
probability of the time tree

Tree shape is informative about underlying dynamics





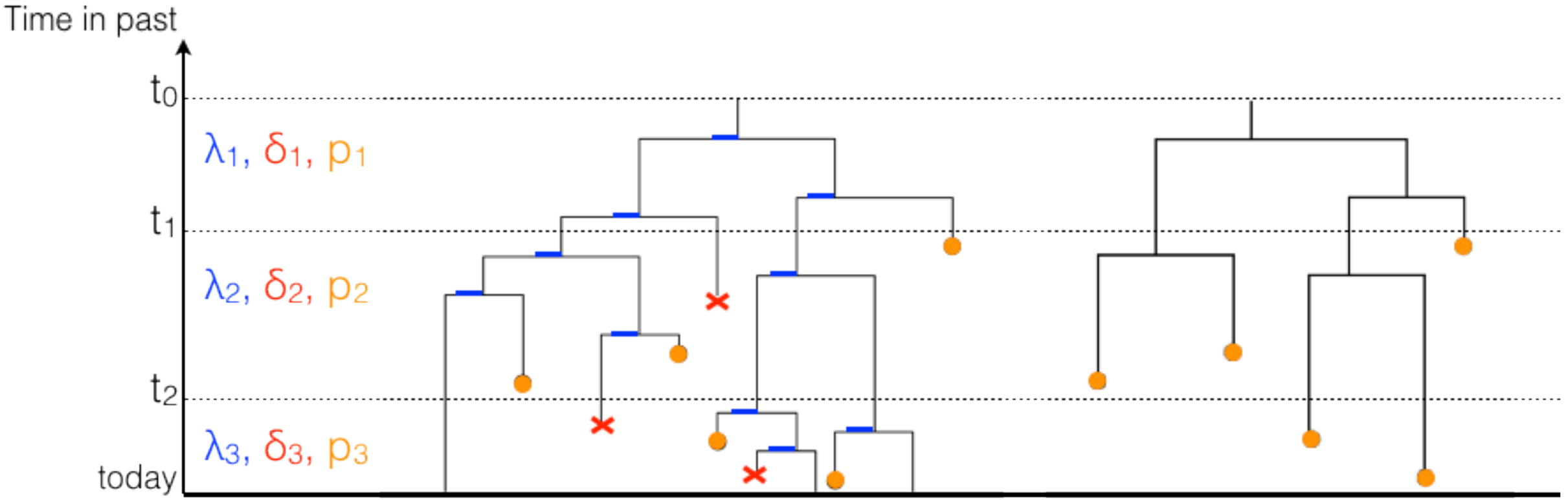




This paper coined the term phylodynamics
Grenfell et al. 2004. Science

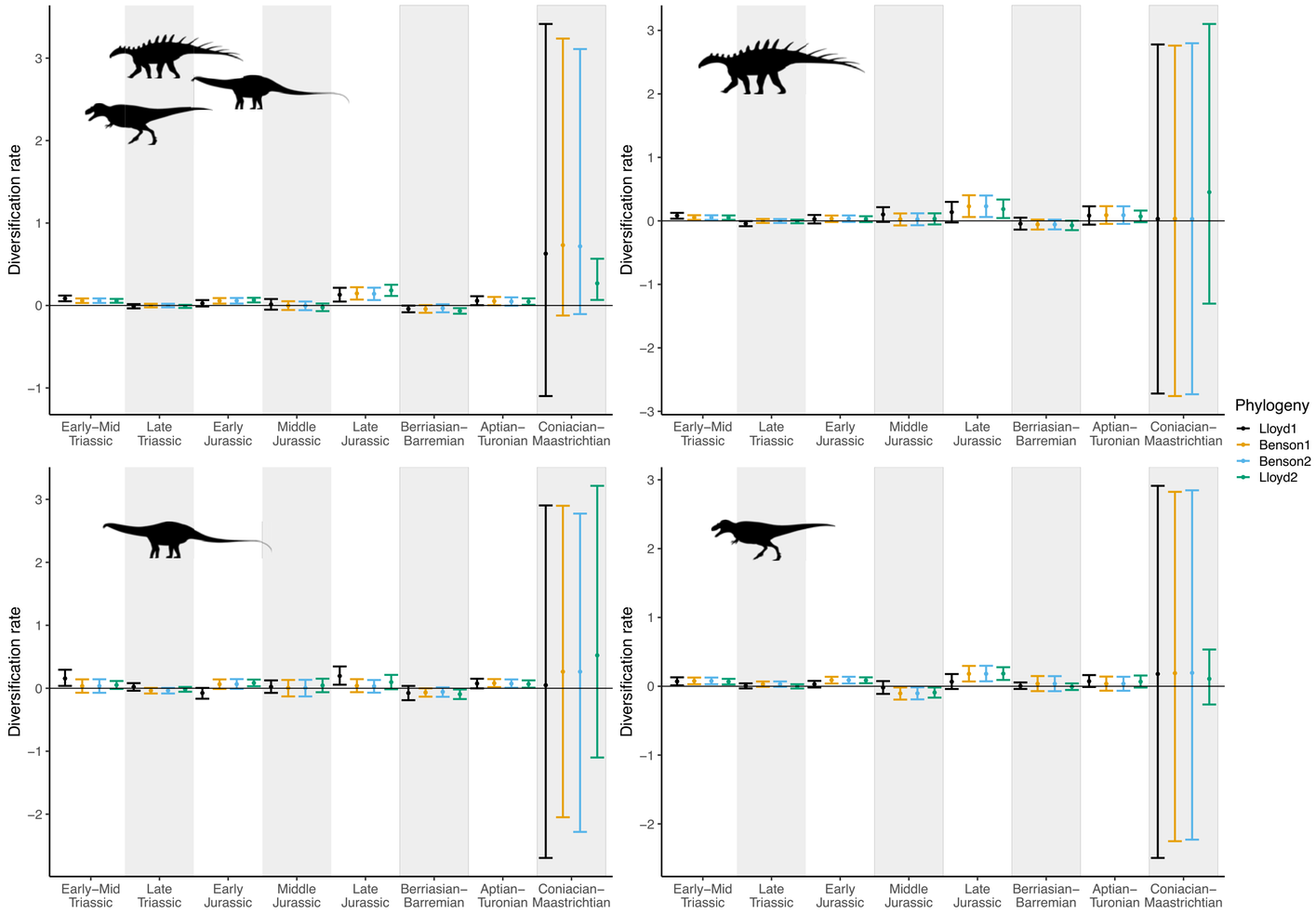
The skyline birth-death process

First used for tracking the spread of infectious diseases



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

Macroevolutionary case study

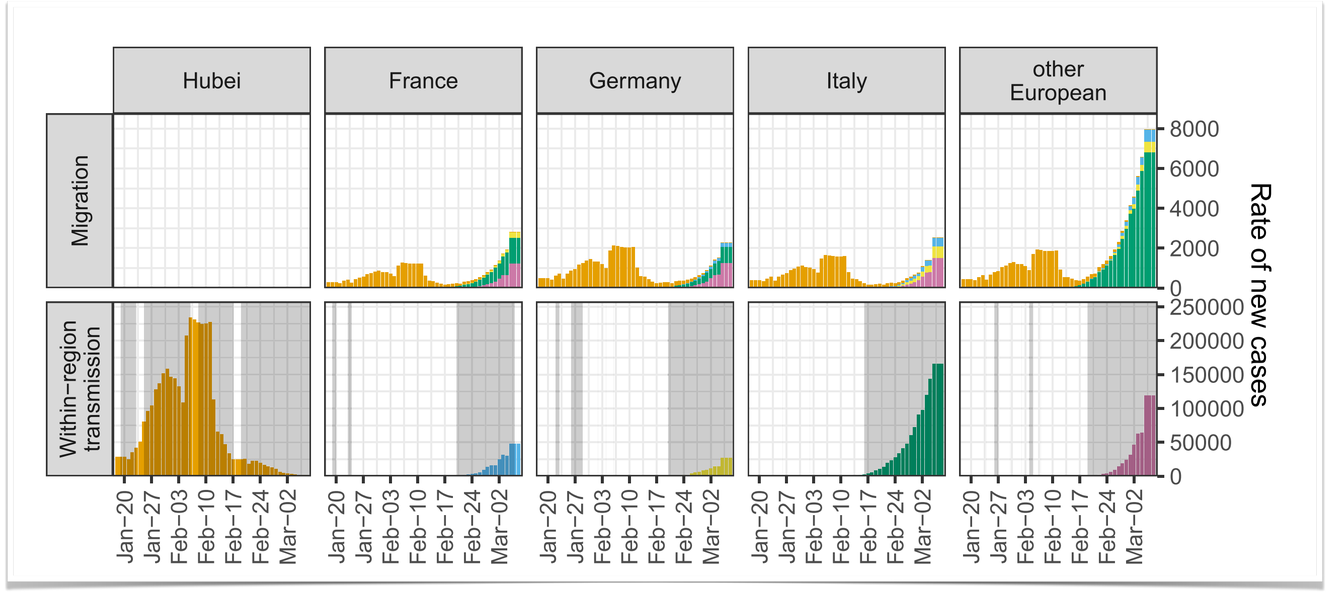


Phylogenies have been used to argue dinosaurs were incline prior to the KPg

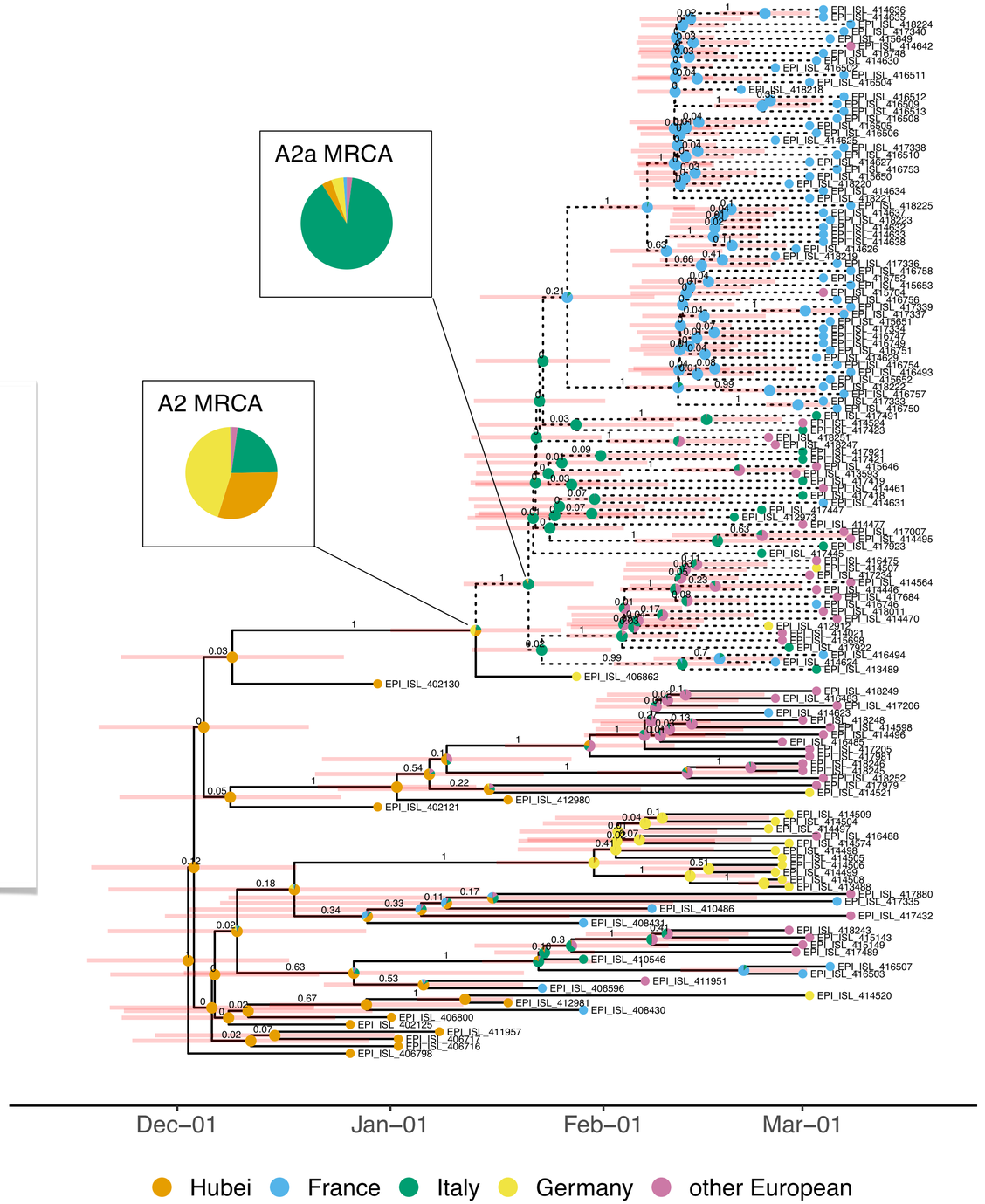
FBD analyses suggest that we can not currently answer that question using phylogenies

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

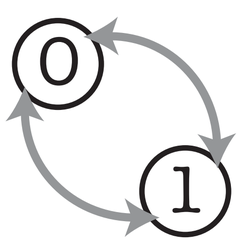
and / or

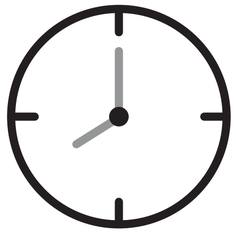
```
0101... ATTG...
1101... TTGC...
0100... ATTC...
```

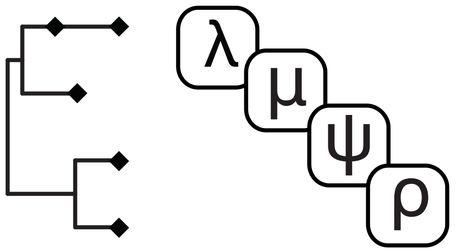
phylogenetics characters

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

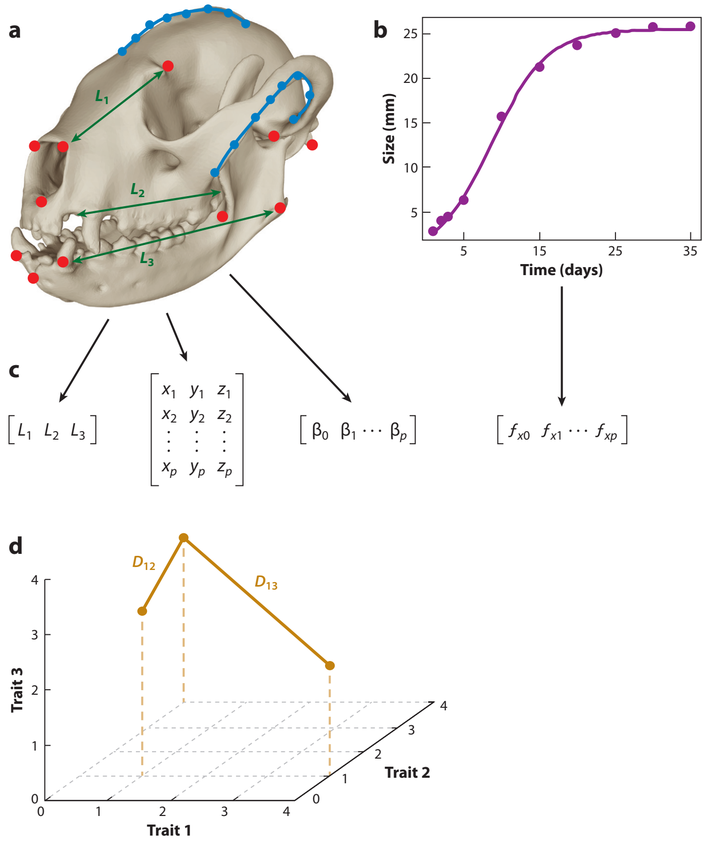
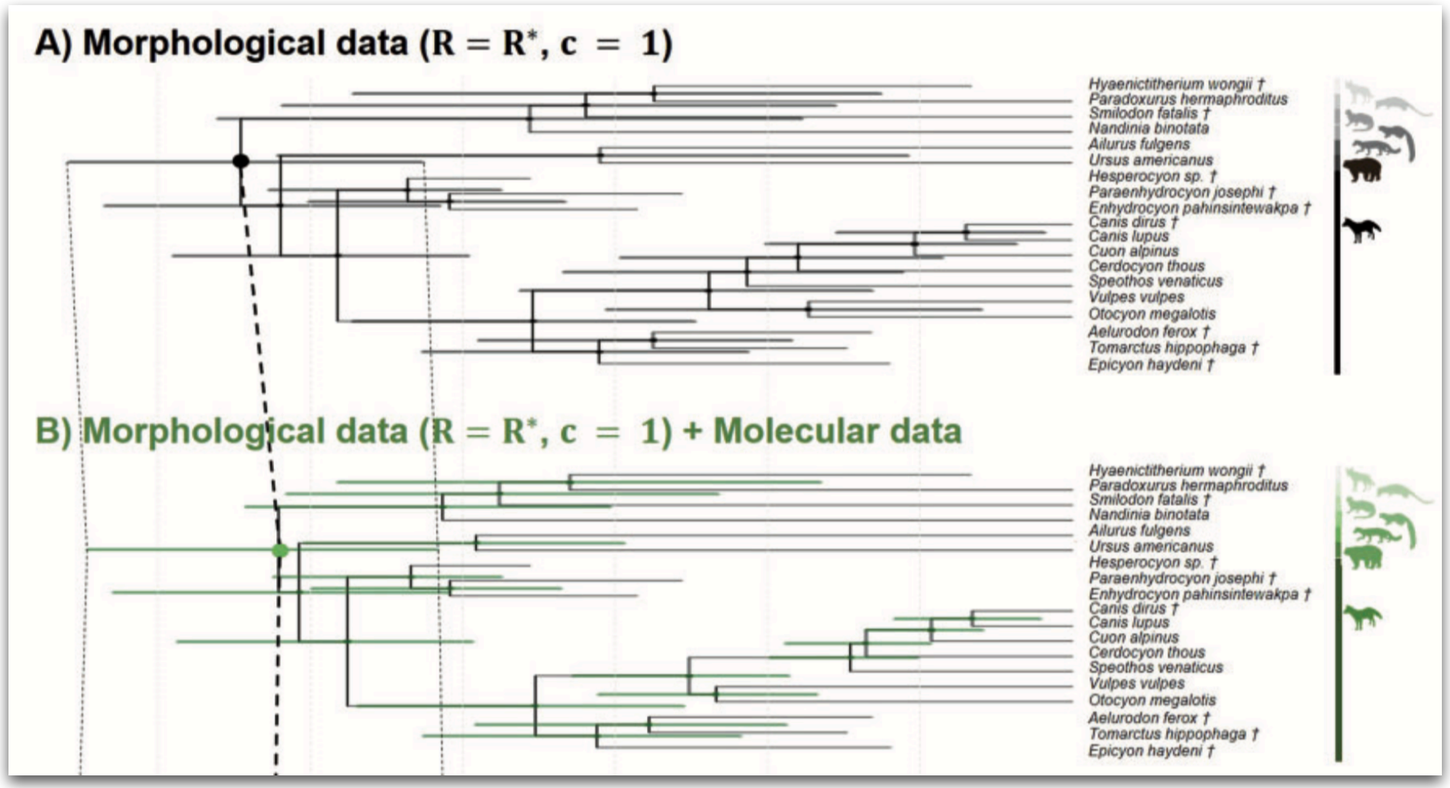
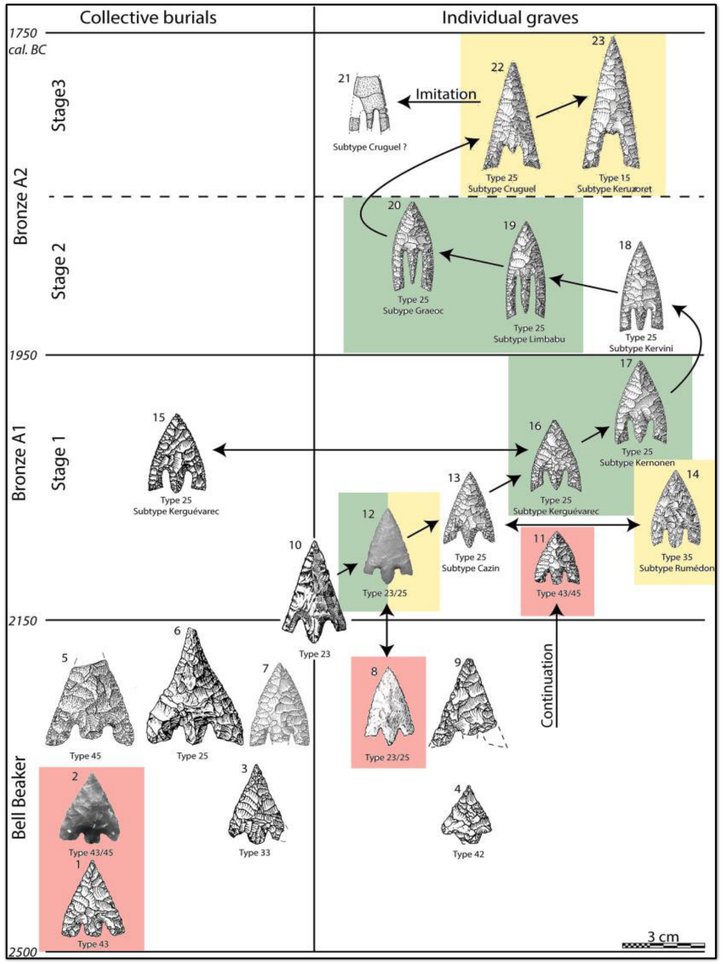


Image source *Adams & Collyer* (2019)



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

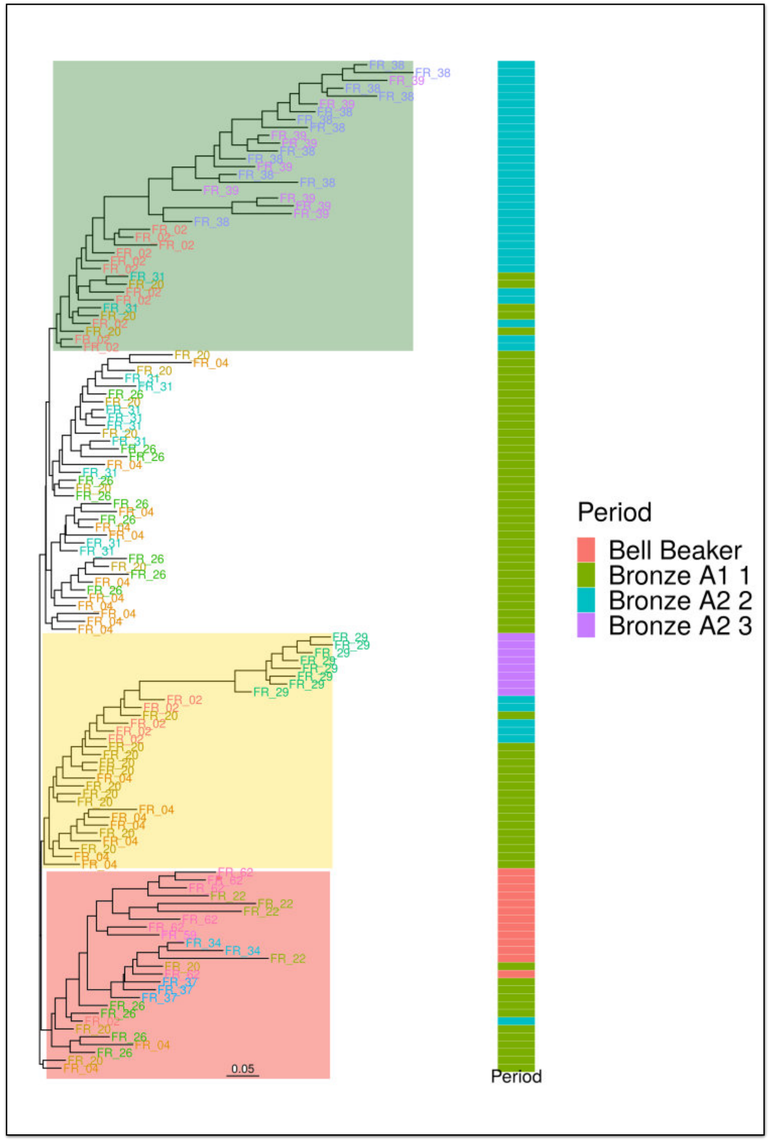
Cultural evolution



After Nicolas (2017)

←Typo-Chronology of Palaeolithic stone tools

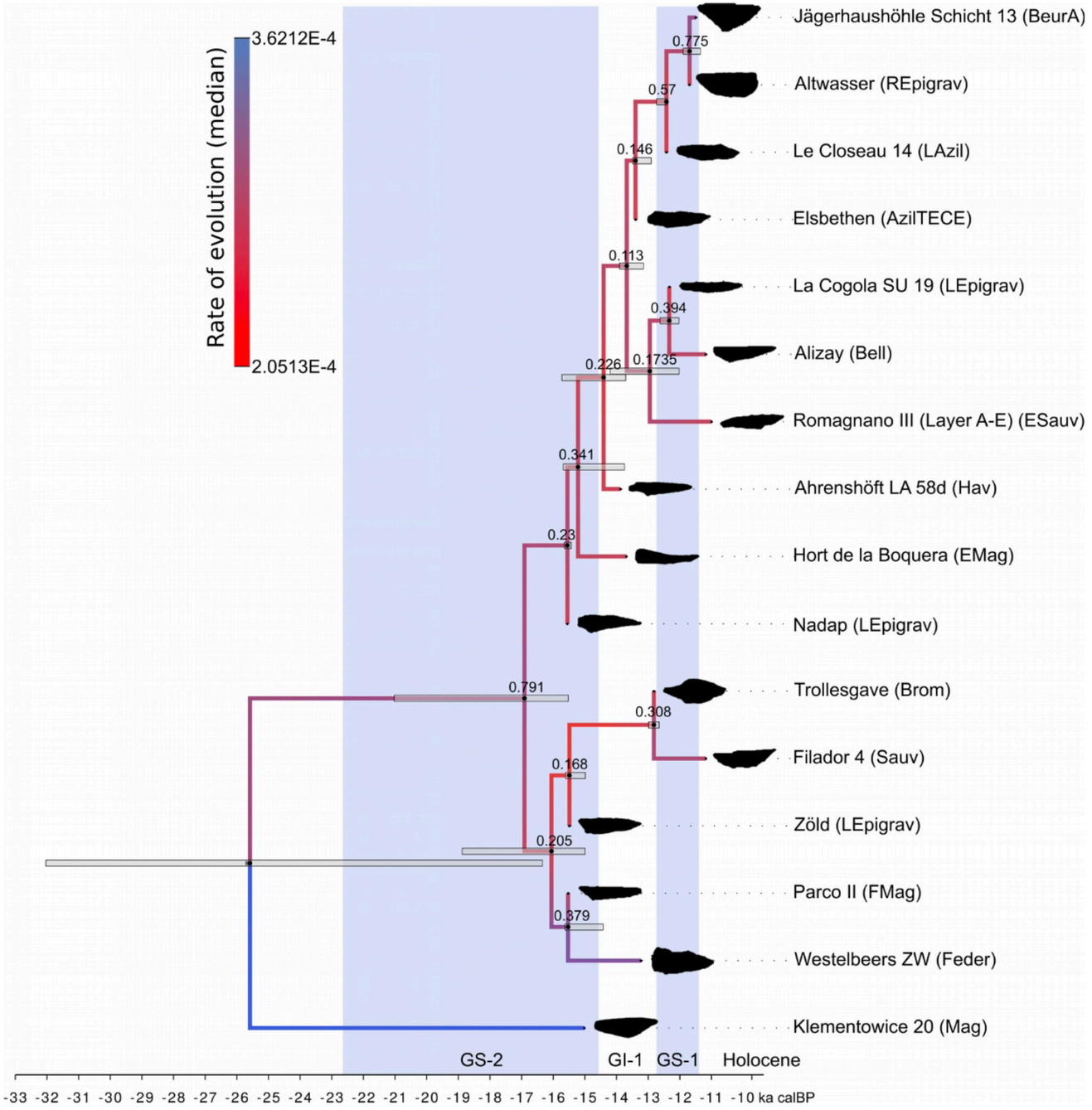
Outline based NJ tree →



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)

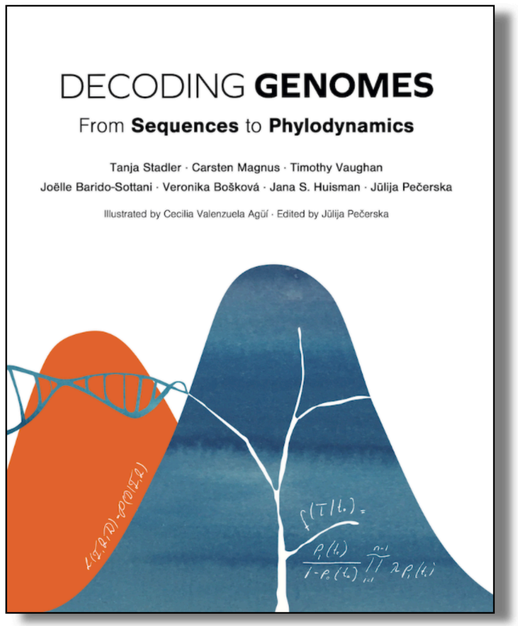


Decoding Genomes: From Sequences to Phylodynamics

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