Phylogenetics

The fossilised birth-death process

Rachel Warnock, Laura Mulvey
rachel.warnock@fau.de, laura.l.mulvey@fau.de

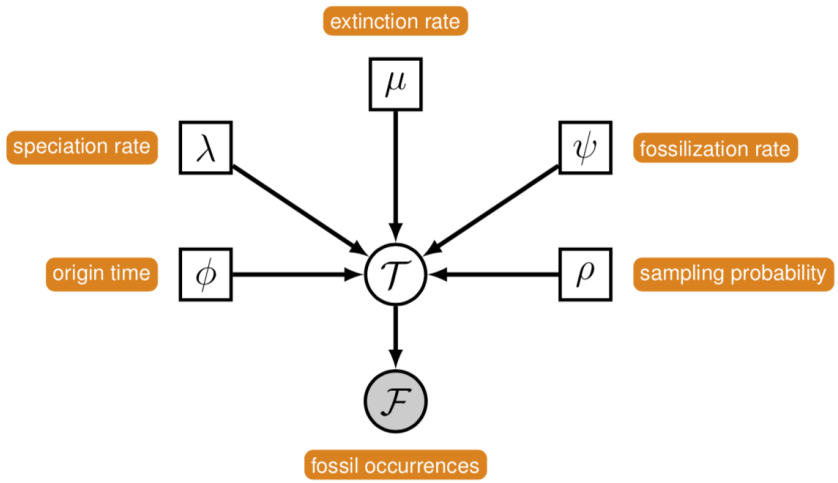
September 12, 2022

Analytical Paleobiology Workshop, Erlangen 2022

Part 6 objectives

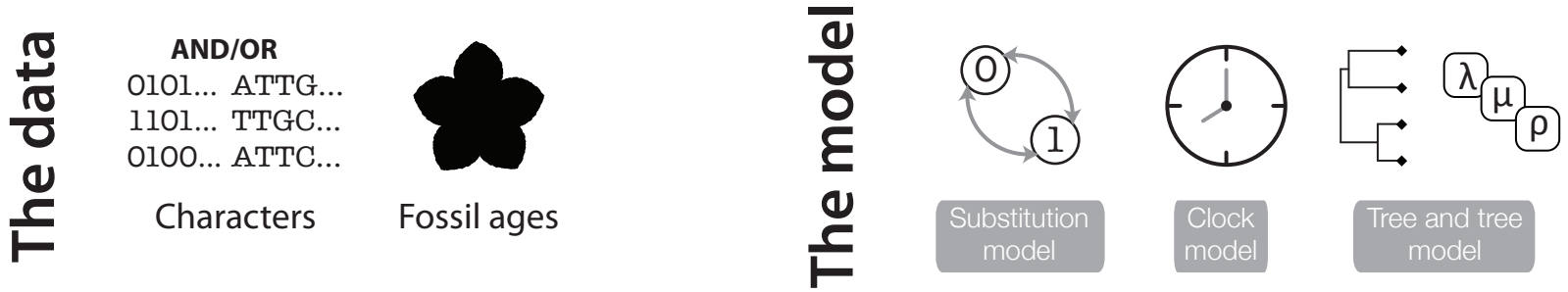
challenges with node dating

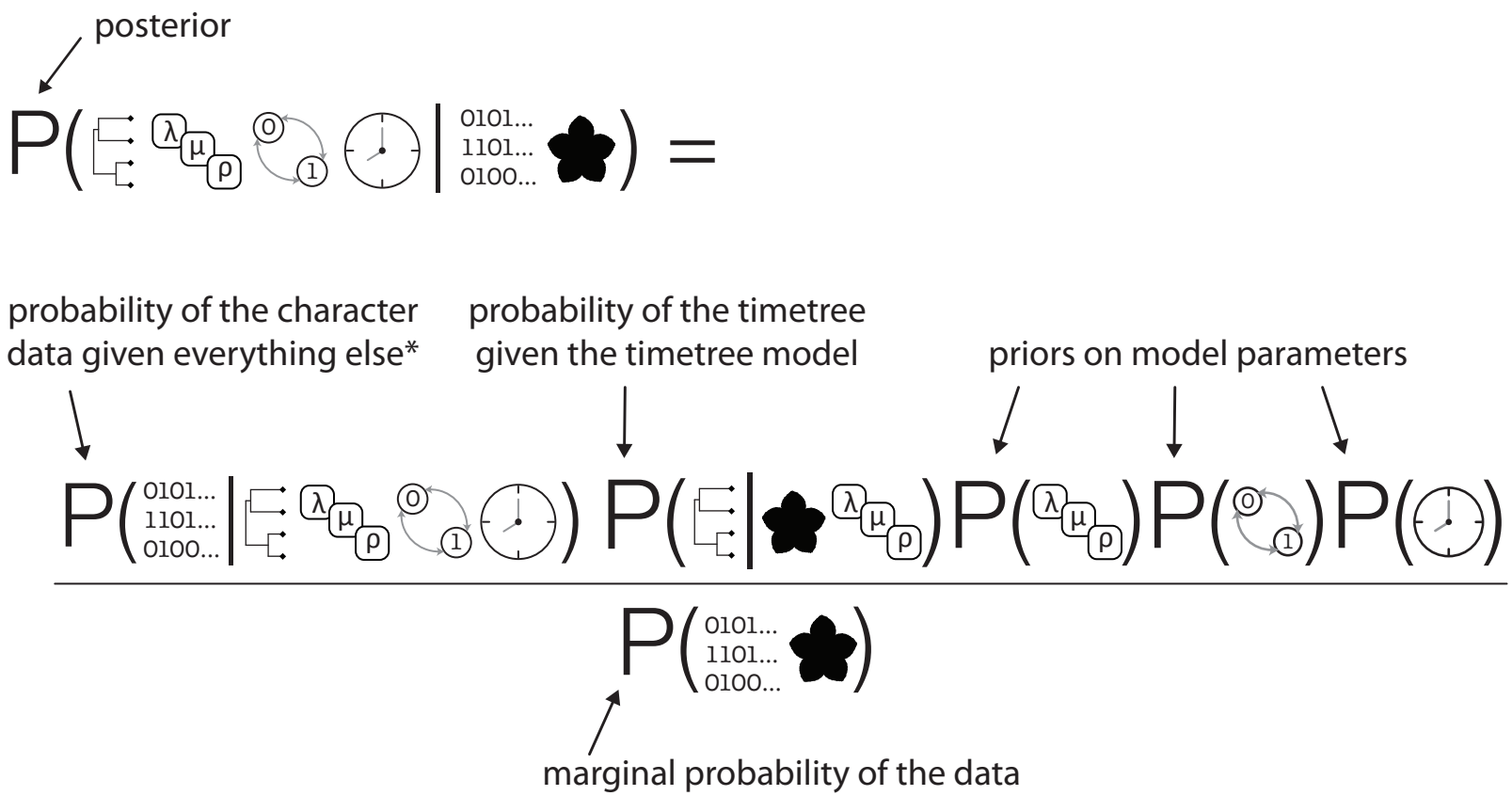
the fossilised birth-death process



Another quick recap

Bayesian phylogenetic dating

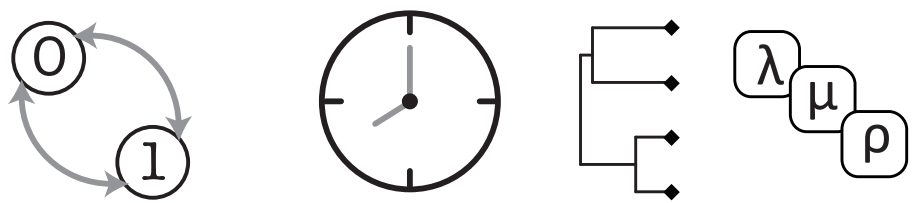




*the timetree, the parameters and the tripartite model

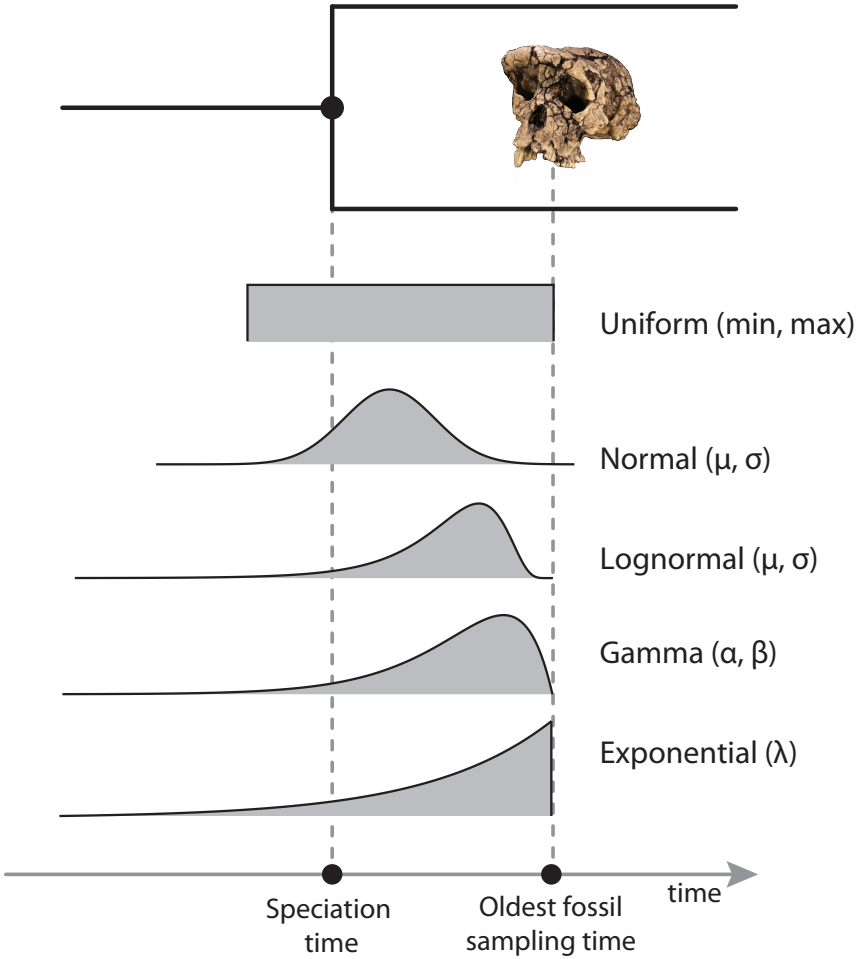
Recap: Bayesian phylogenetic dating requires three model components

- The **substitution model** ← describes how sites evolve over time.
- The **clock model** ← describes how evolutionary rates vary across the tree.
- The **tree model** ← describes how trees grow over time. Temporal evidence is included here.



Substitution model | Clock model | Tree and tree model

Recap: Node dating



- We used a birth-death model to describe the tree generating process, given we only observe extant species.
- Then we separately apply a calibration density to constrain internal node ages.

Image adapted from Heath (2012) *Systematic Biology*

Challenges with node dating

Taxonomic uncertainty

Early crown vs. stem group taxa can be hard to distinguish.

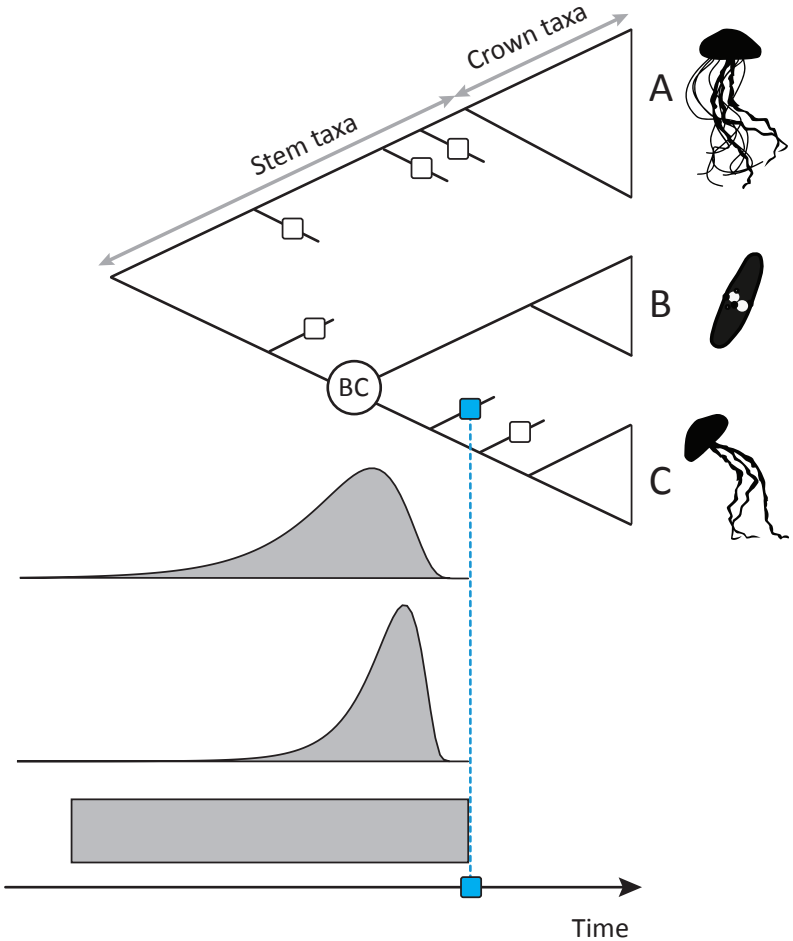
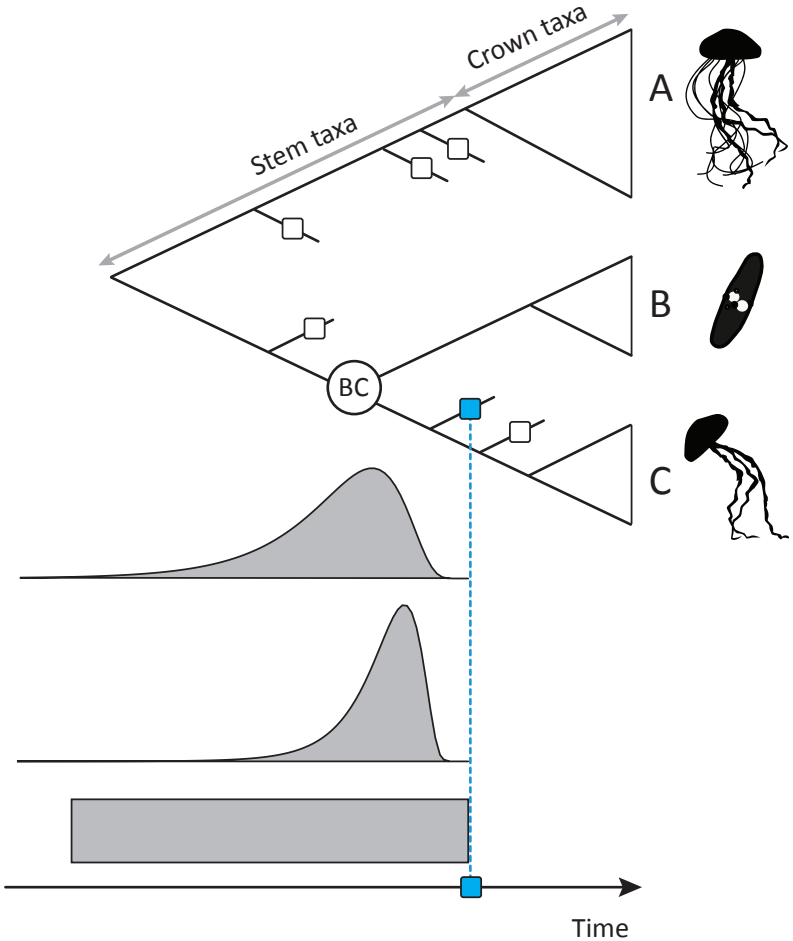


Image from Warnock, Engelstädter (in press)

Taxonomic uncertainty

Early crown vs. stem group taxa can be hard to distinguish.



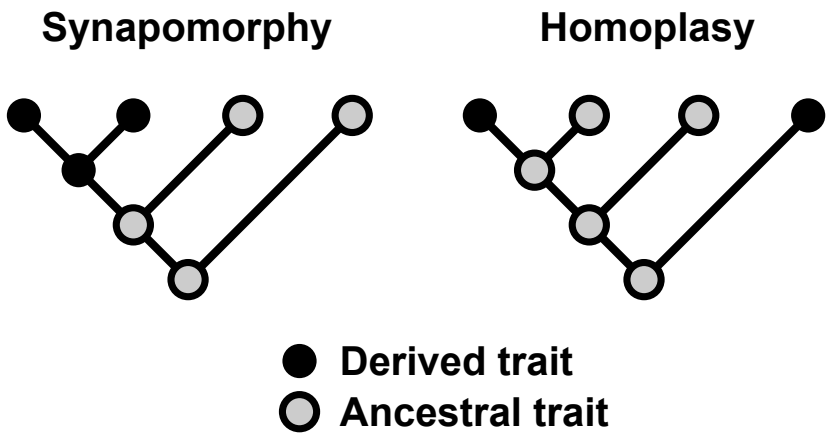
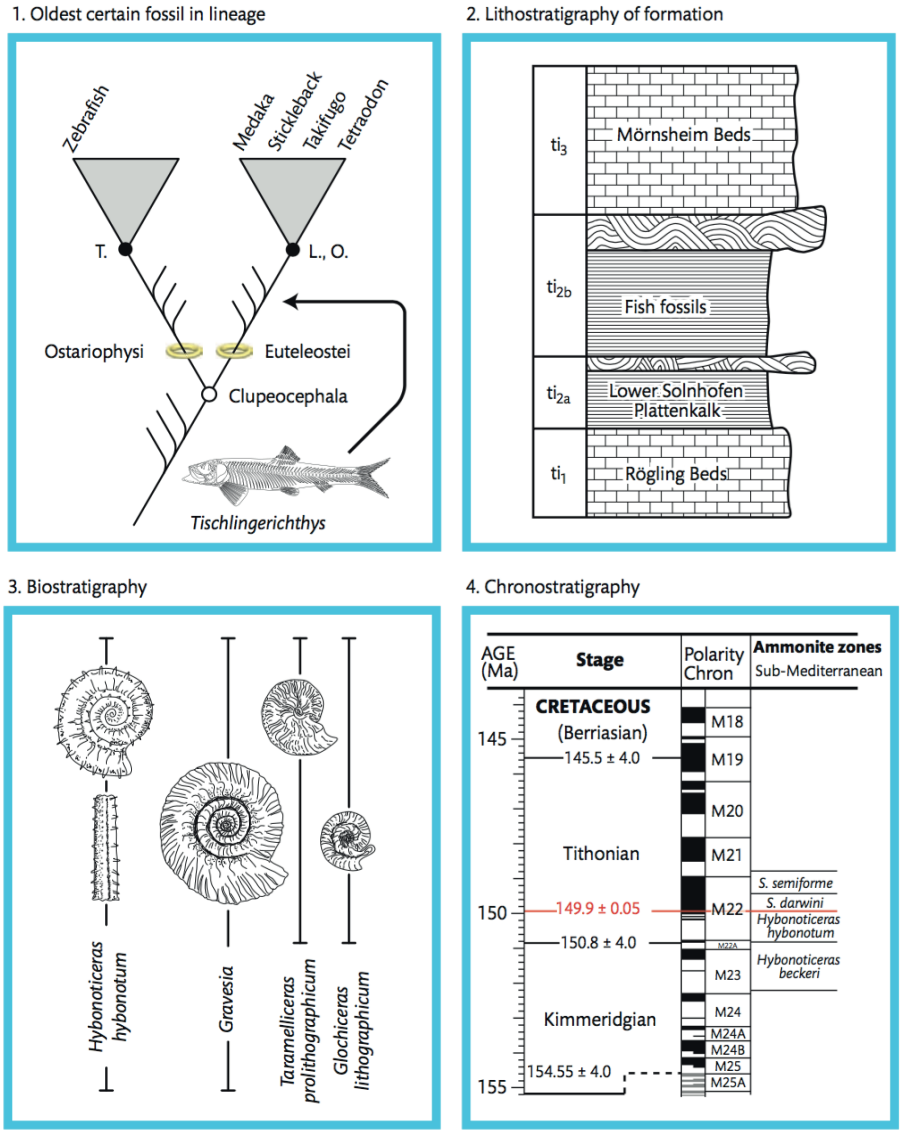


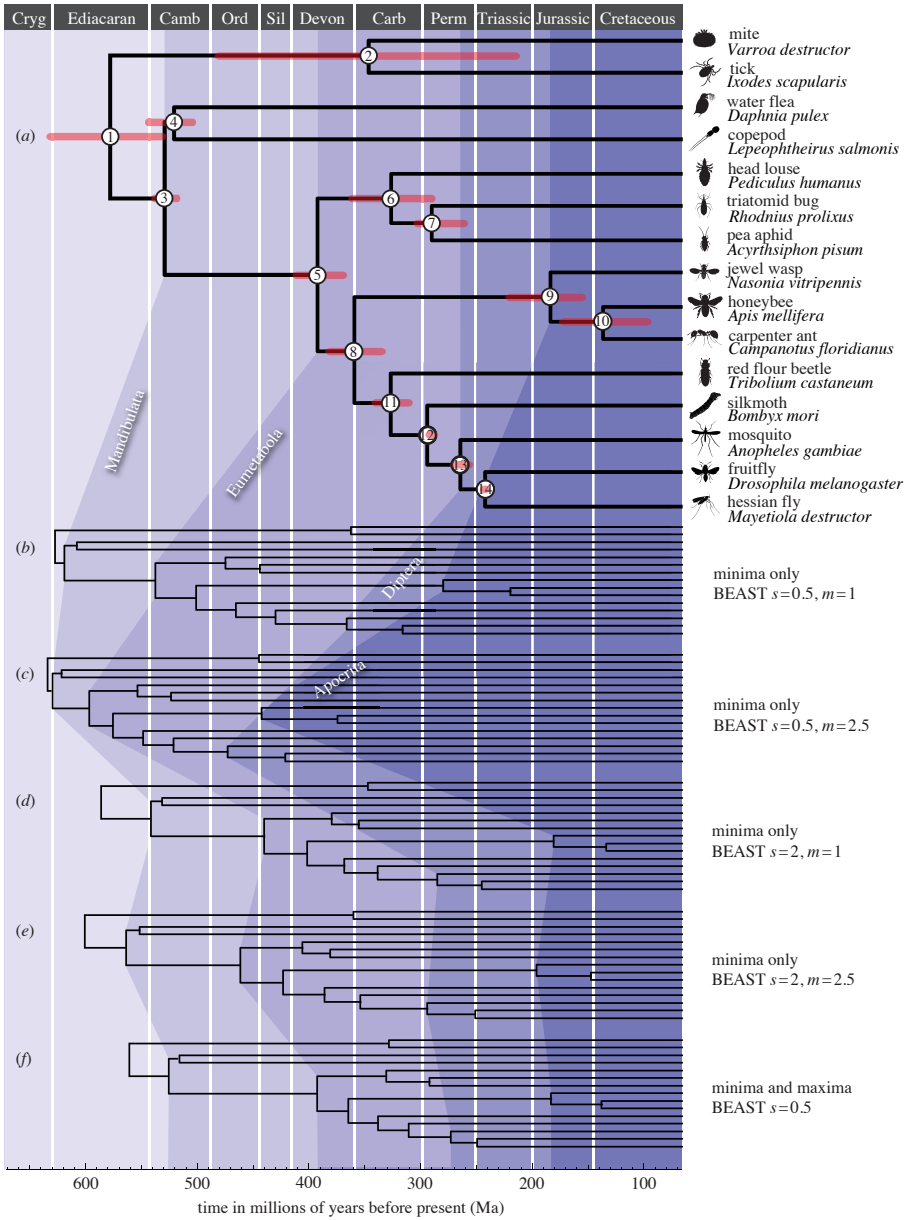
Image from Warnock, Engelstädter (in press)

Stratigraphic age uncertainty



Benton et al. (2009) *The Timetree of Life*

Calibration priors have a large impact

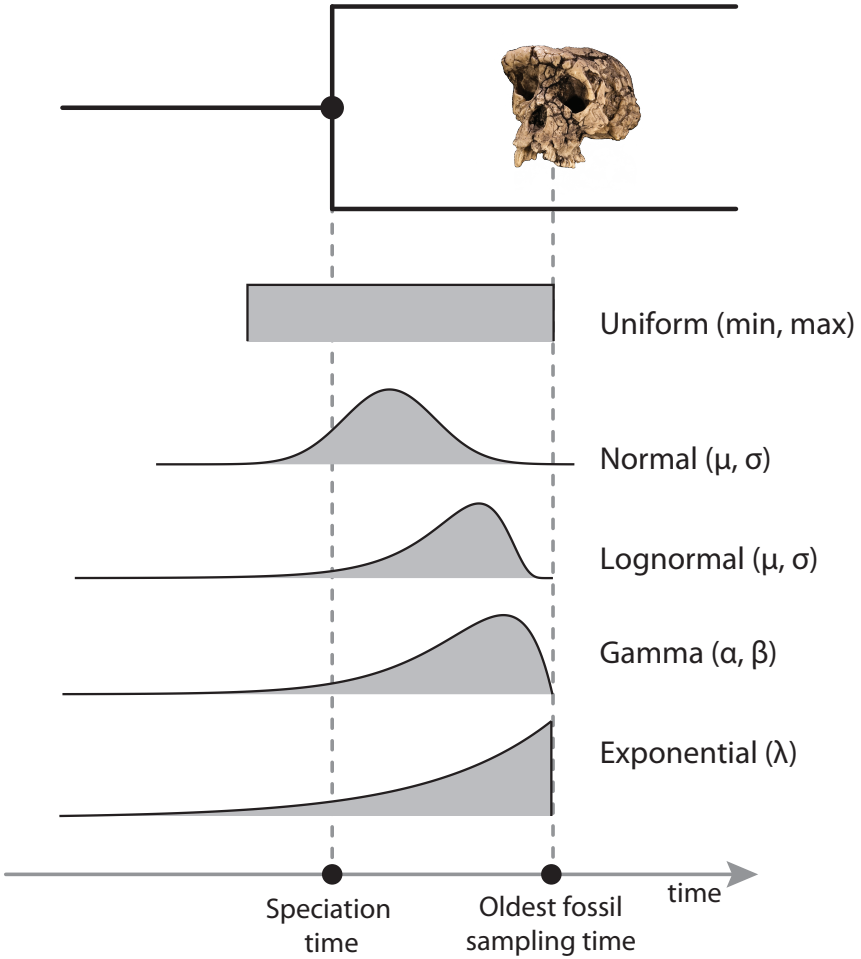


Warnock et al. (2011) *Biology Letters*

← Small differences in the prior parameters can have a huge impact.

We also need (loose) maximum constraints on divergence times.

Specifying calibration distributions

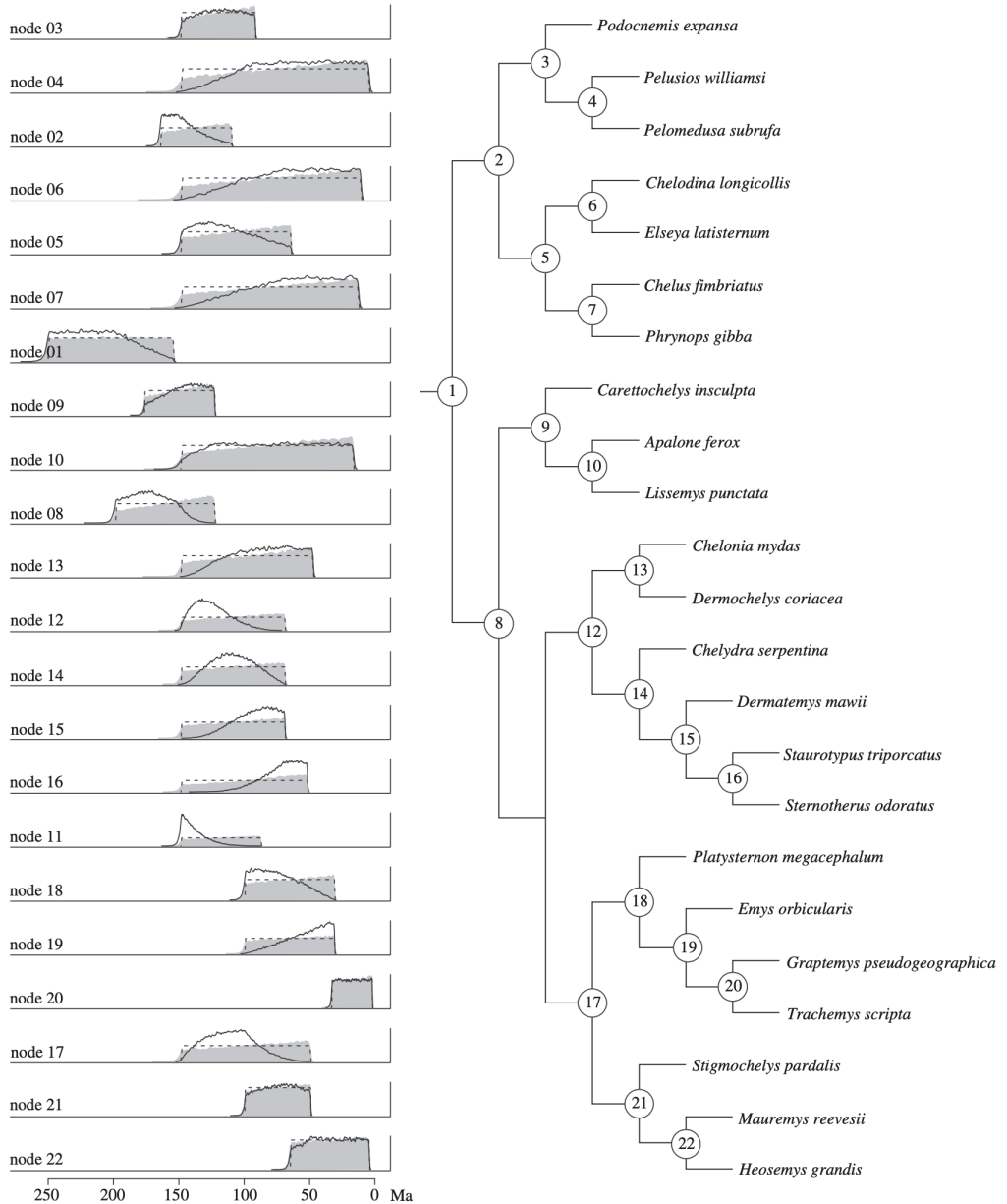


- It potentially excludes a lot of information, since typically we only assign one fossil per calibration node.
- The model doesn't describe the process that generated the fossil sampling times, leading to **statistically incoherency**.

Image adapted from Heath (2012) *Systematic Biology*

Statistical inconsistency

The user specified priors (dashed lines) don't necessarily match the effective priors (black lines).



Warnock et al. (2015)

Challenges associated with node dating: summary

It requires assuming the phylogenetic position of fossils is known without error.

Specifying calibration densities is tricky.

It potentially excludes a lot of information, since typically we only assign one fossil per calibration node.

The model doesn't describe the process that generated the fossil sampling times, leading to **statistically incoherency**.

See Warnock et al. (2015) for more on this topic.

How do we include more information from the fossil record?

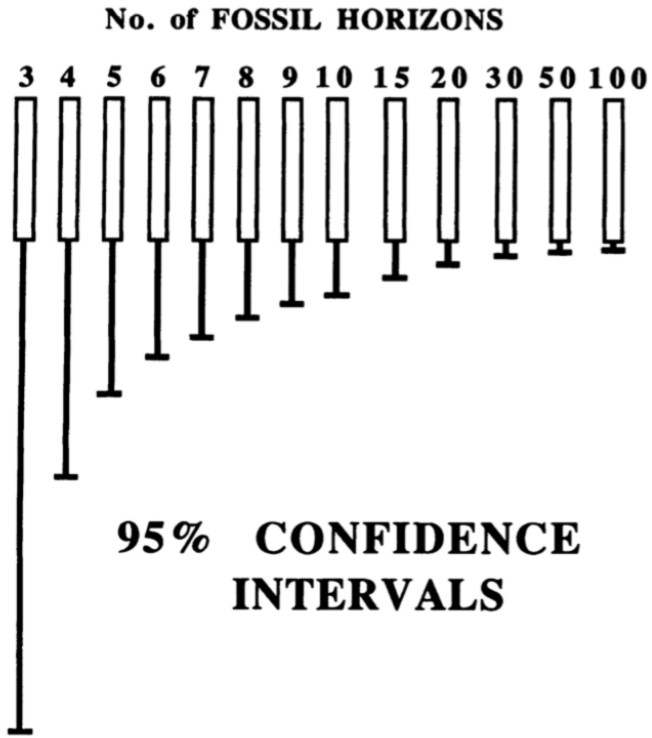
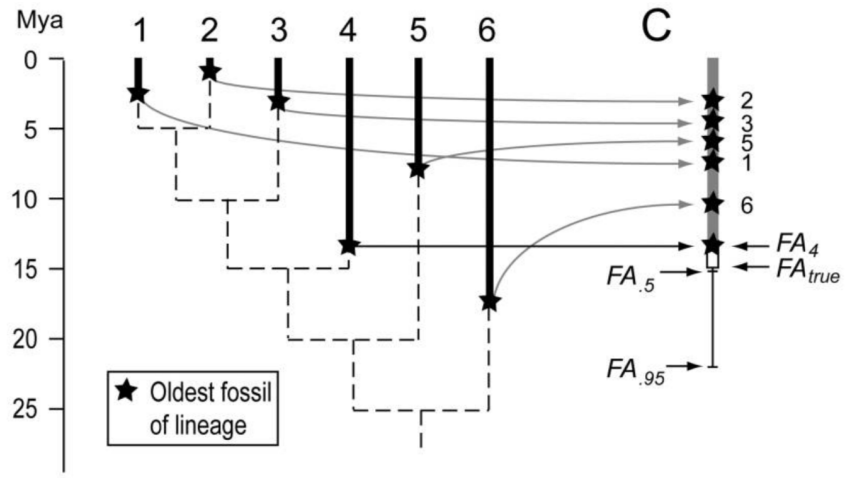


FIGURE 1. Ninety-five percent confidence intervals on the bases of hypothetical stratigraphic ranges. The number of fossiliferous horizons (H) is indicated above the appropriate stratigraphic column. The confidence intervals were calculated using Eq. (1).

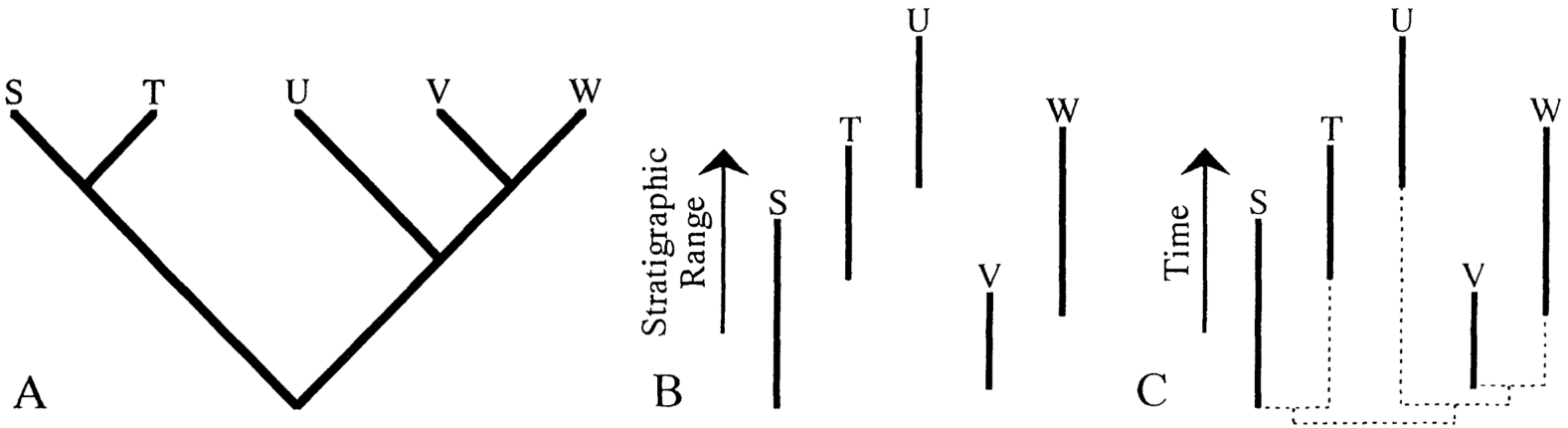
$$\alpha = (1 - C_1)^{-1/(H-1)} - 1, \qquad (1)$$

Marshall (1990) *Paleobiology*



Marshall (2008) *American Naturalist*

Is there another way?

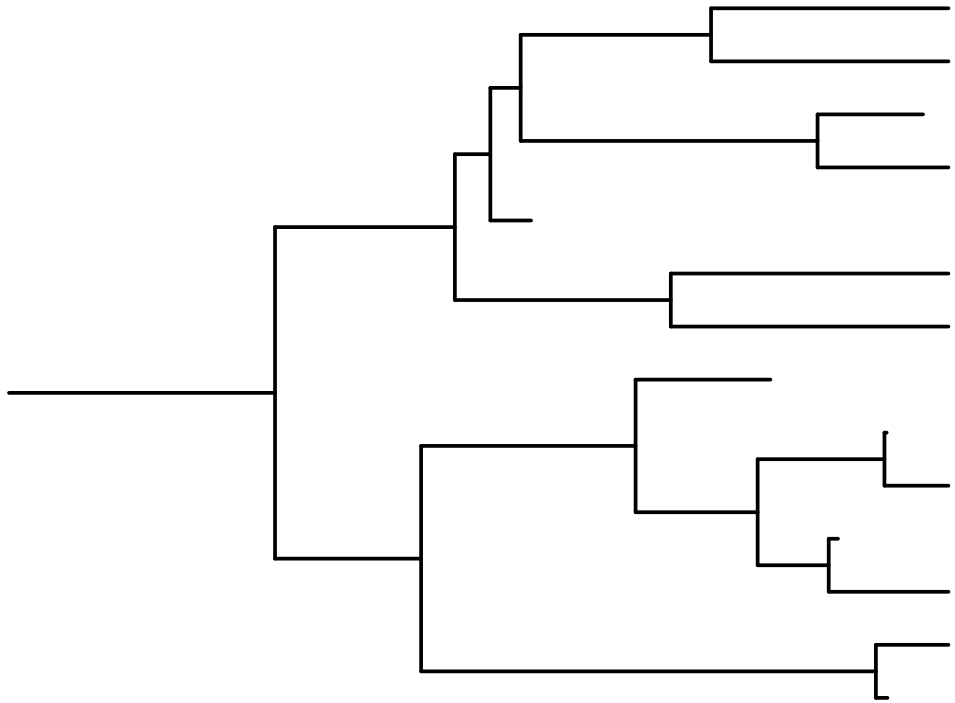


Stratigraphic tests of cladistic hypotheses
Wagner. 1995. Paleobiology

The fossilised birth-death process

Speciation and extinction

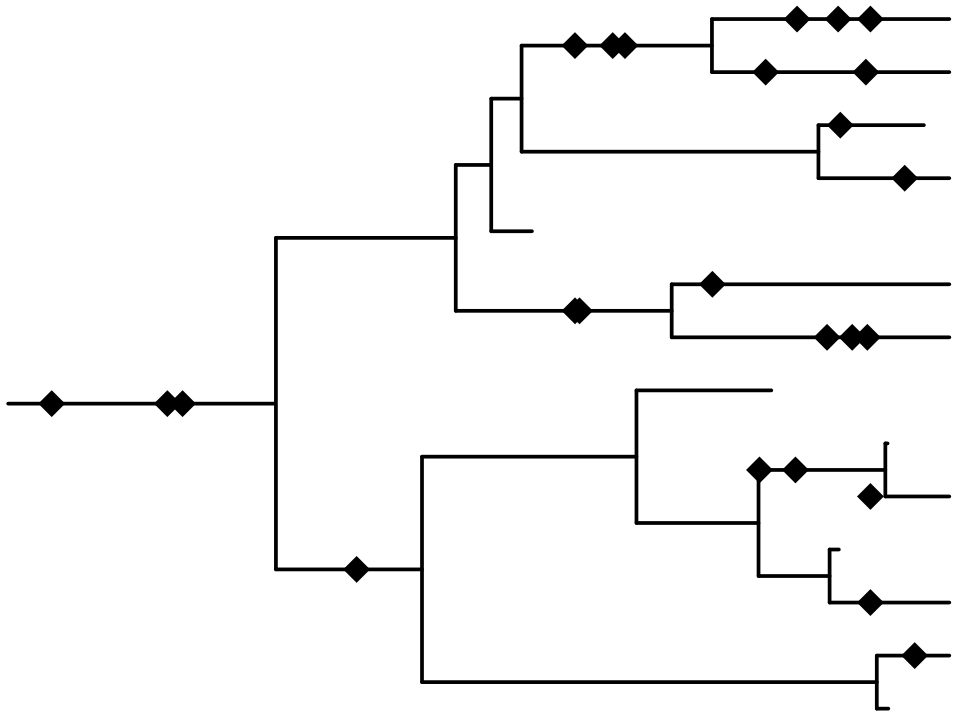
New lineages arise with speciation rate λ and lineages terminate with extinction rate μ.



Different combinations of λ and μ produce different tree shapes.

Extinct and extant (living) species sampling

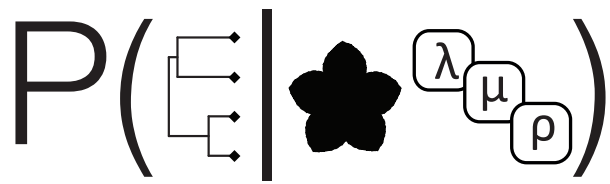
Fossils are sampled along lineages with fossil recovery rate ψ and extant (living) species are sampled at the present ($t = 0$) with probability ρ.



Different combinations of λ, μ and ψ produce different distributions of fossil sampling times.

The fossilised birth-death process

For statistically coherent phylogenetic inference we need an expression for the probability of observing the sampled tree given the speciation, extinction, living species and **fossil sampling** processes.

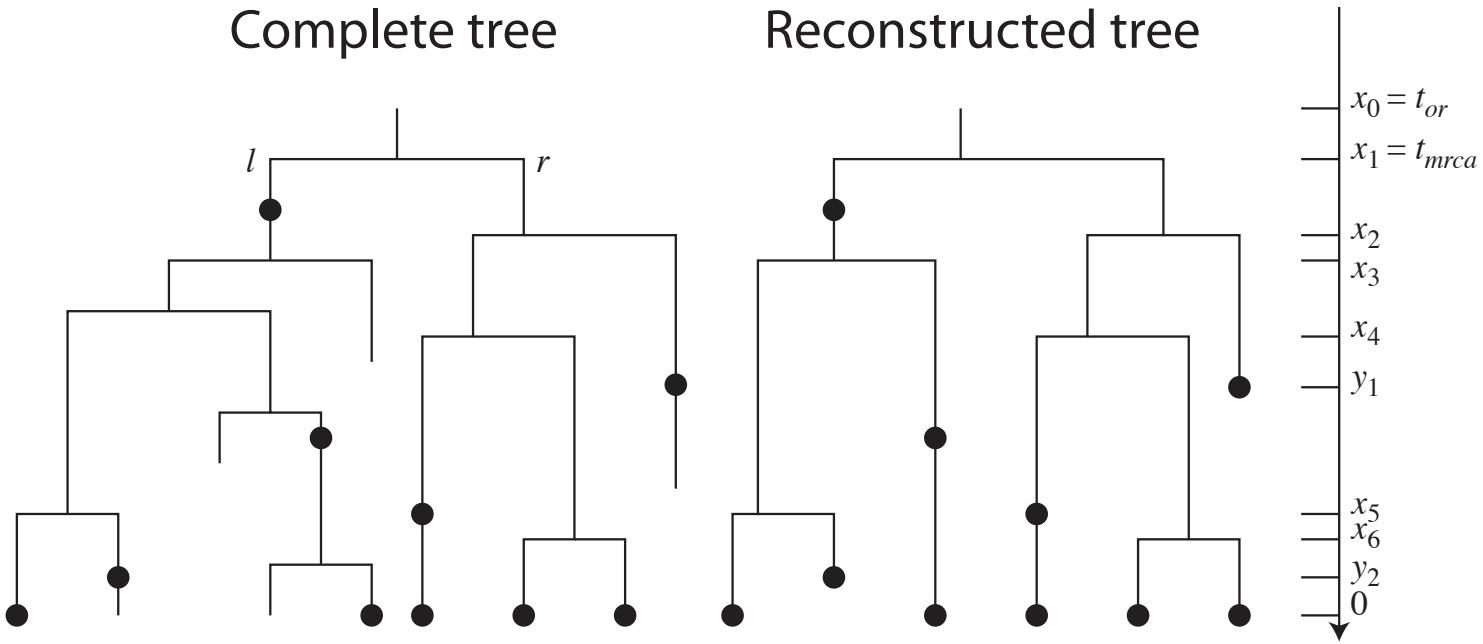


Previously we've talked about tree models that don't incorporate extinct species sampling.

- Pure birth $P(\mathcal{T}|\lambda)$
- Birth death $P(\mathcal{T}|\lambda, \mu)$
- Birth death sampling $P(\mathcal{T}|\lambda, \mu, \rho)$

Calibration information is combined with these models in a way that doesn't capture the fossil sampling process.

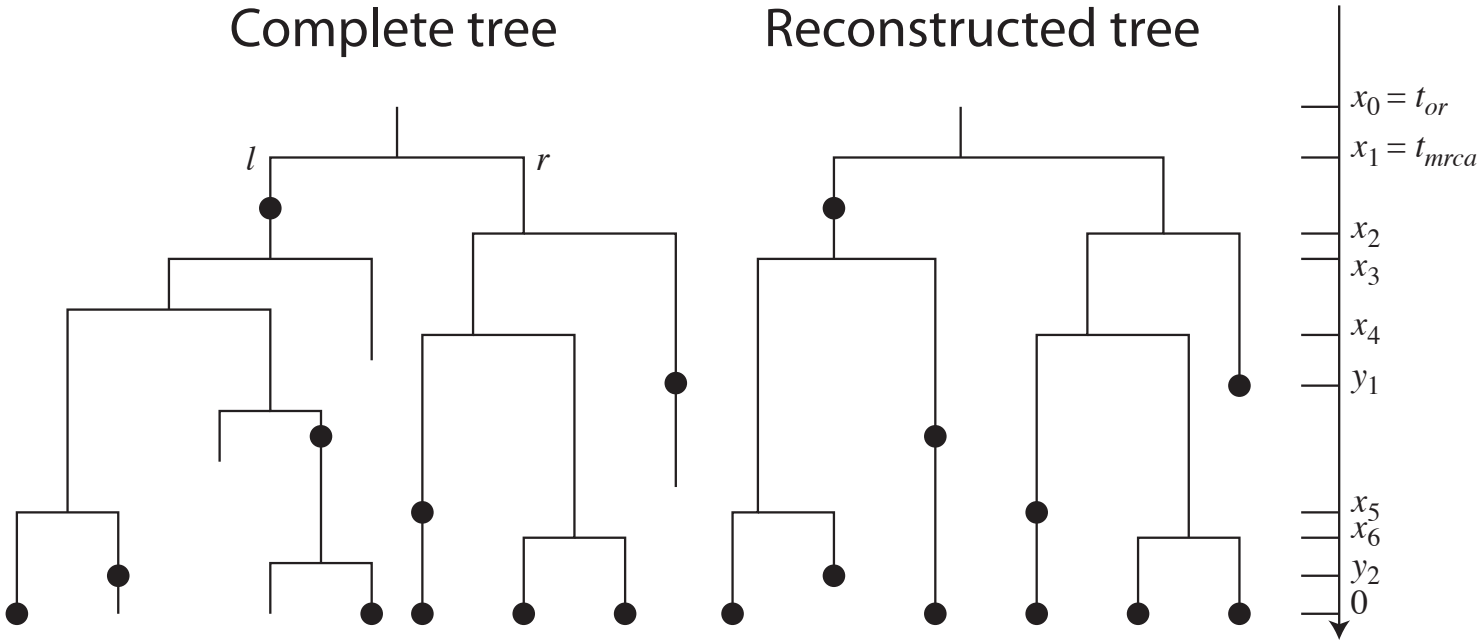
The fossilised birth-death process



Note that some samples fall along internal branches of the sampled tree. These are known as **sampled ancestors**.

Expression first derived Stadler (2010) *JTB*

The fossilised birth-death process



The FBD process describes the probability of observing the sampled tree, i.e. $P(\mathcal{T}|\lambda, \mu, \rho, \psi)$.

We can use this model as a prior on the tree topology and divergence times.

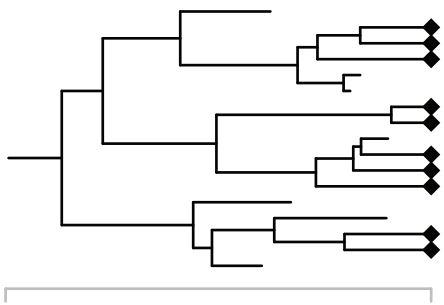
Expression first derived Stadler (2010) *JTB*

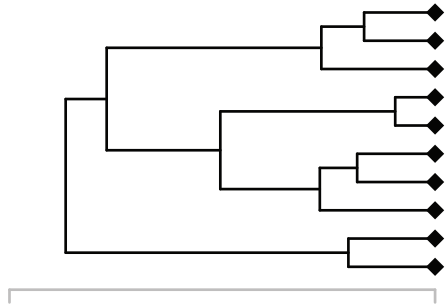
Complete versus reconstructed tree

The complete outcome of the diversification and sampling processes

The reconstructed tree

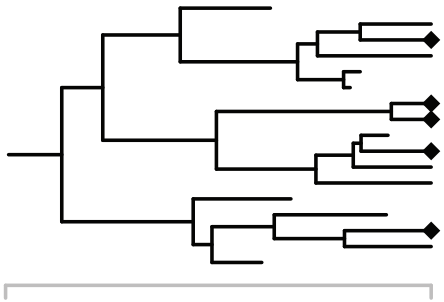
Model parameters

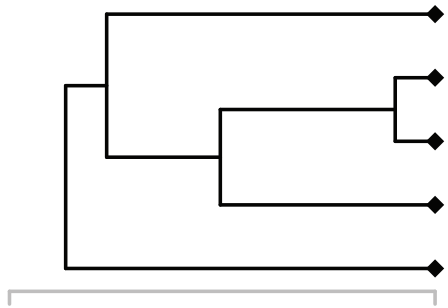




speciation (λ) = 0.1
extinction (μ) = 0.05

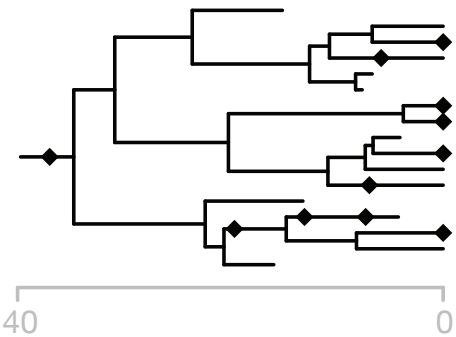
Birth-death process

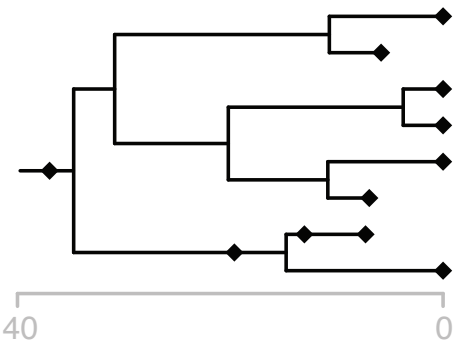




speciation (λ) = 0.1
extinction (μ) = 0.05
extant sampling (ρ) = 0.6

Birth-death sampling process





speciation (λ) = 0.1
extinction (μ) = 0.05
extant sampling (ρ) = 0.6
fossil recovery (ψ) = 0.05

Fossilized birth-death process

The fossilised birth-death process

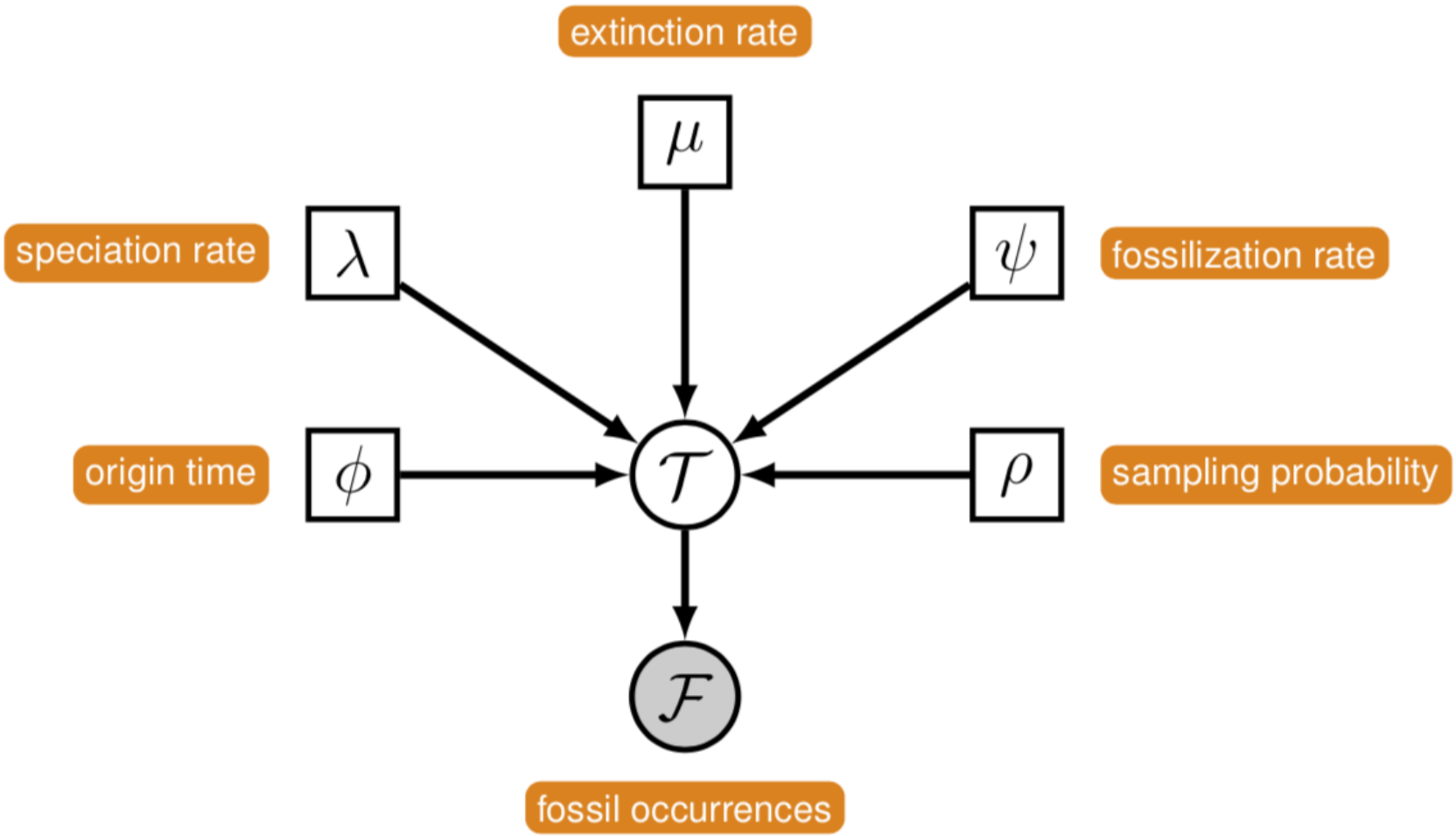


Image adapted from Walker, Heath (2020) *Phylogenetics in the Genomic Era*.

The fossilised birth-death process: advantages

The model gives rise to statistically coherent priors, i.e. the model describes the underlying data generating processes.

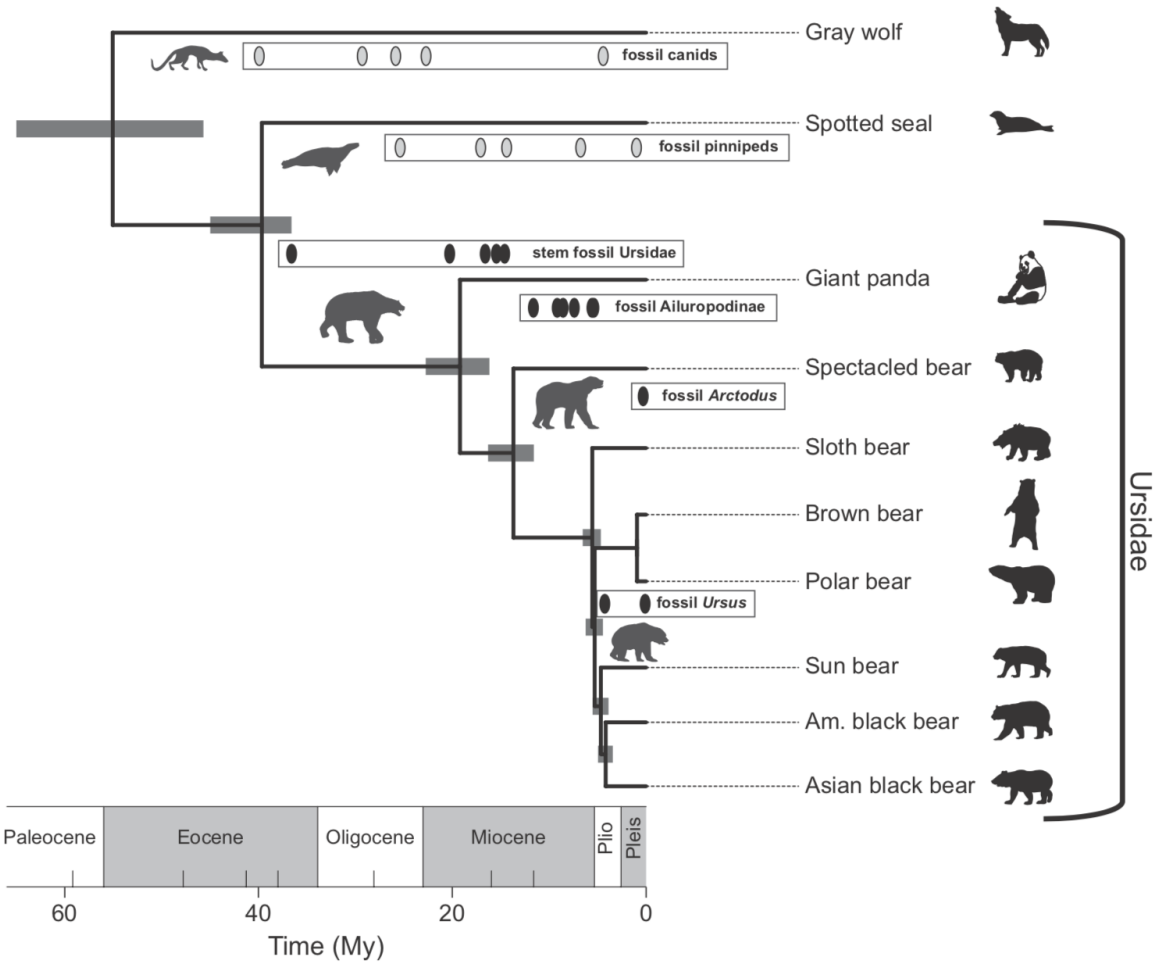
Fossils are directly considered as part of the tree $\rightarrow$ we can include much more information, not only first appearances but all available fossils, including stem fossils.

We can include fossils with and without character data + account for phylogenetic uncertainty.

We can account for sampled ancestors.

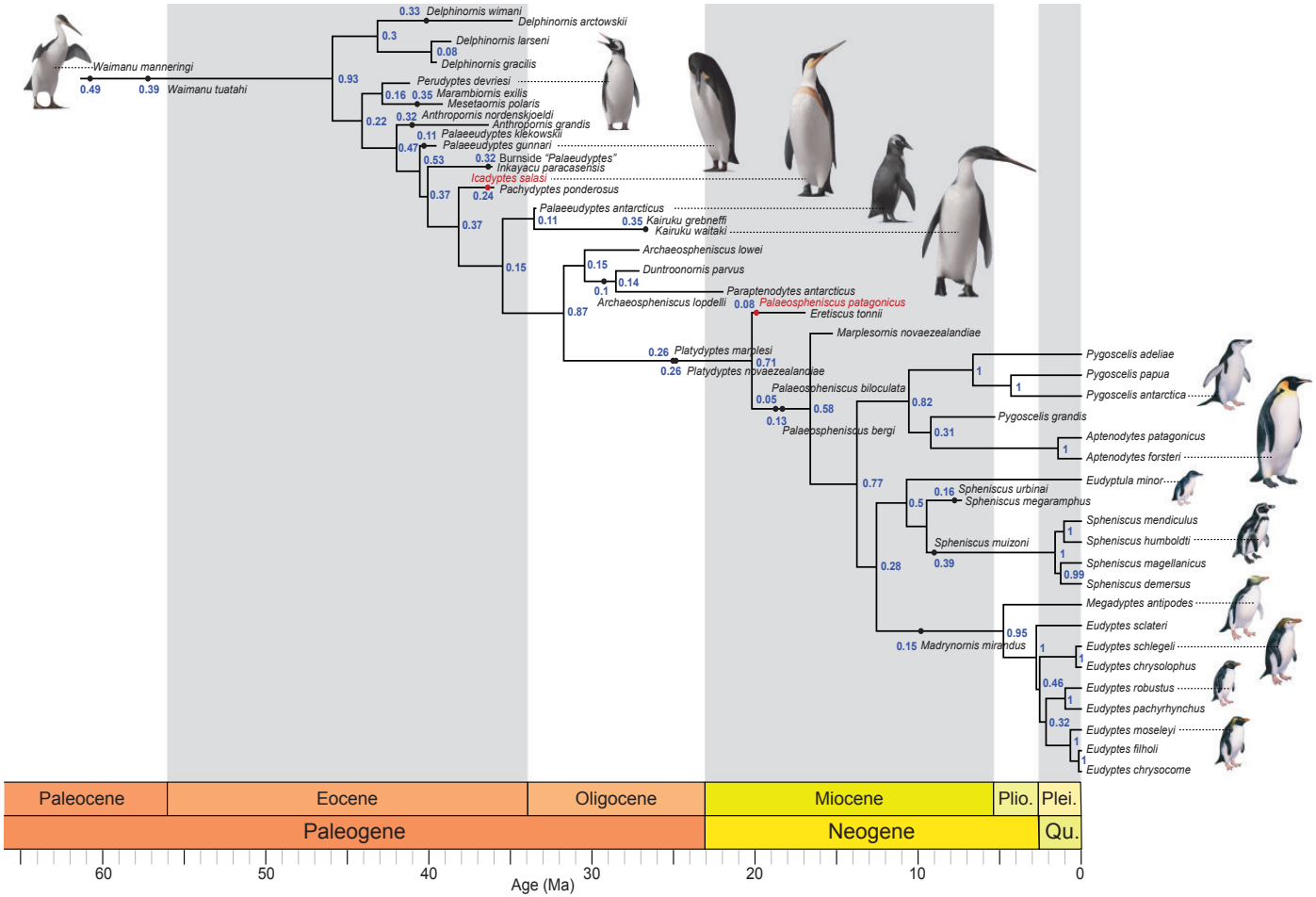
The model provides the basis for a very flexible framework (e.g. we can relax the assumption of constant sampling).

Phylogenetic dating under the fossilised birth-death process



First implementation of the FBD model Heath et al. (2014) *PNAS* and shortly after Gavryushkina et al. (2014) *PLoS Comp Bio*

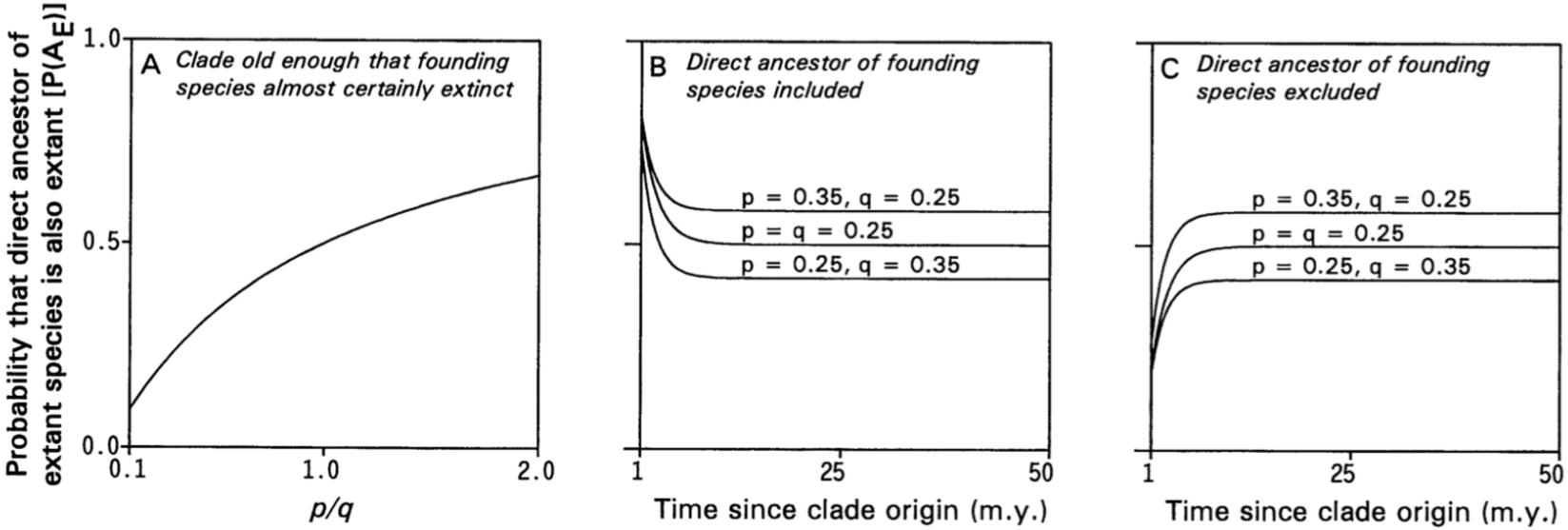
"Total" evidence dating under the fossilised birth-death process



Gavryushkina et al. (2017) *Sys Bio*, see also Zhang et al. (2017) *Sys Bio*

What about sampled ancestors?

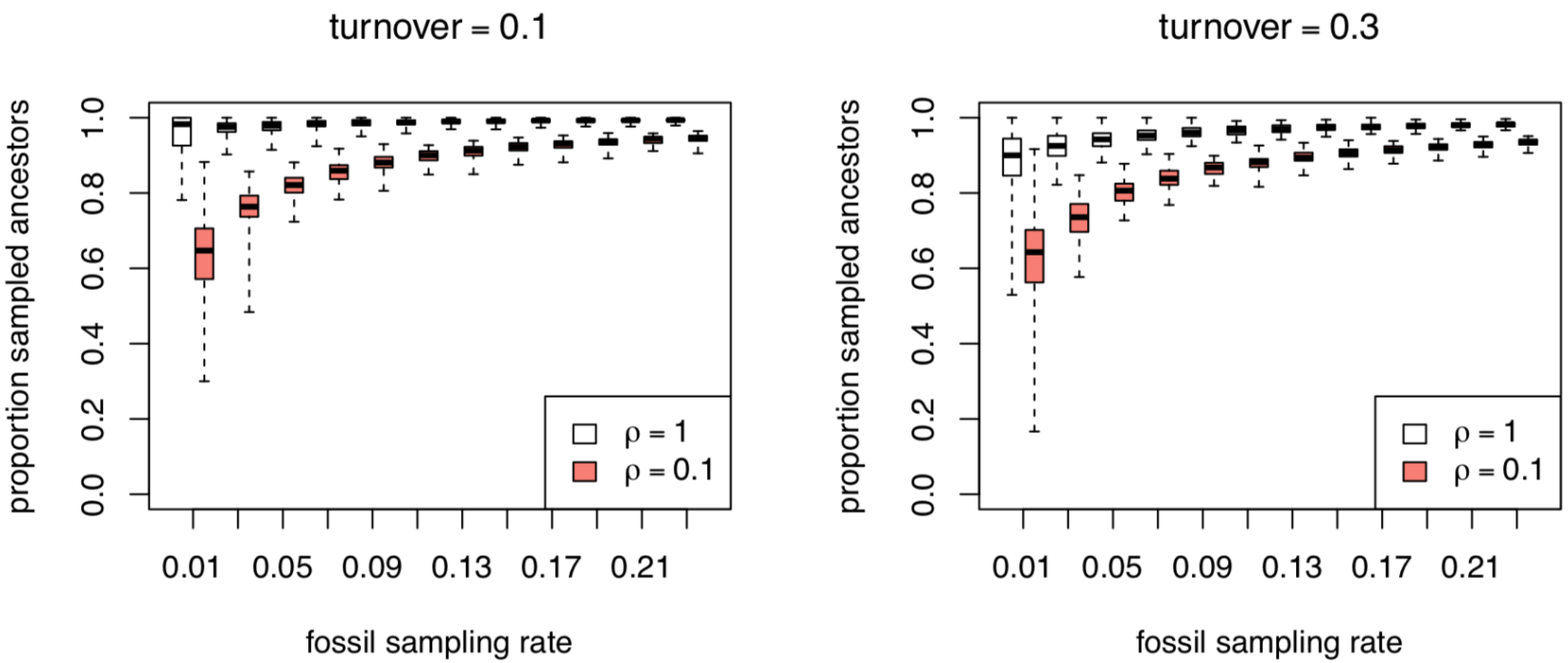
Depending on λ, μ and ψ, the probability of a sample having a sampled ancestor can be quite high.



Foote 1996 *Paleobiology*.

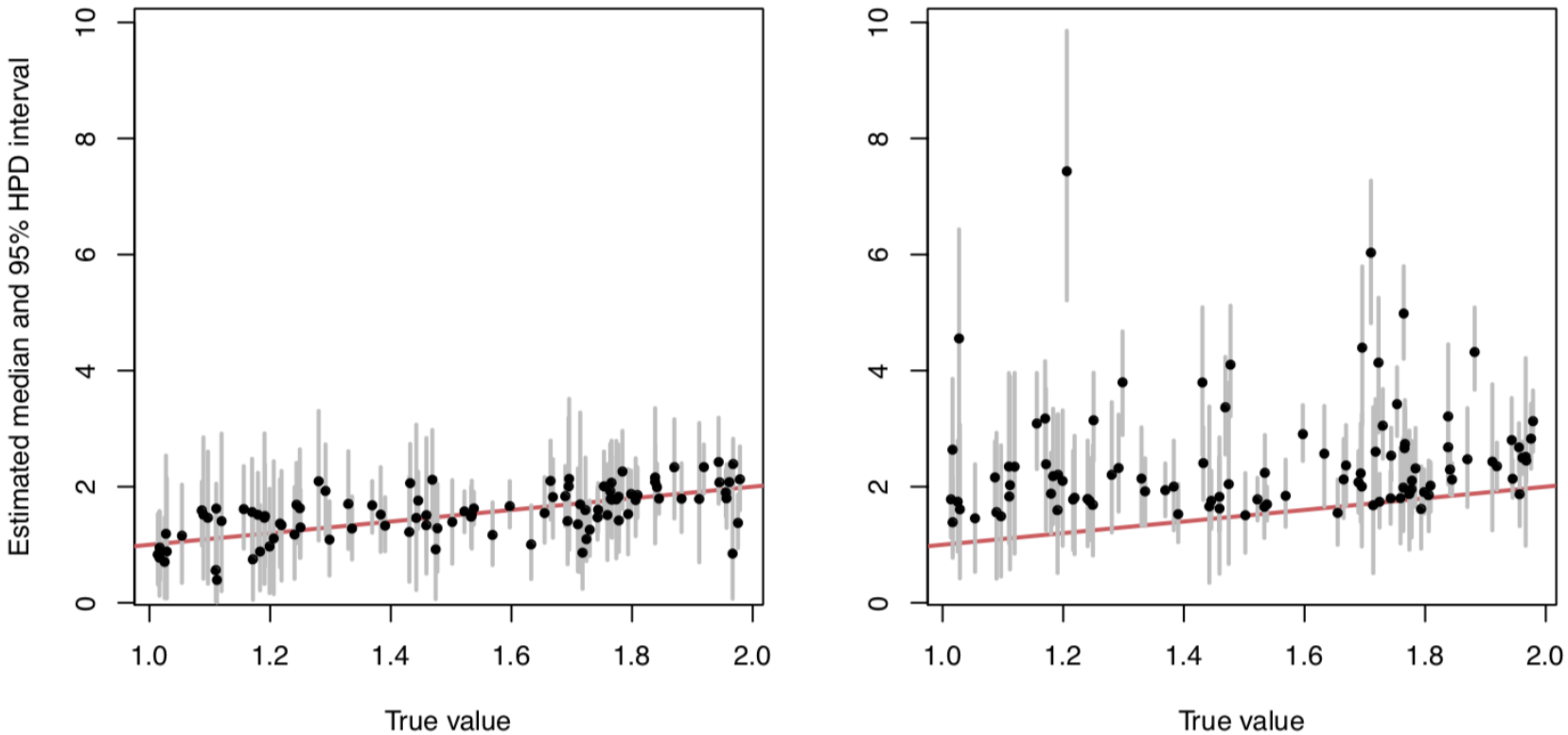
What about sampled ancestors?

Depending on λ, μ and ψ, the probability of a sample having a sampled ancestor can be quite high.



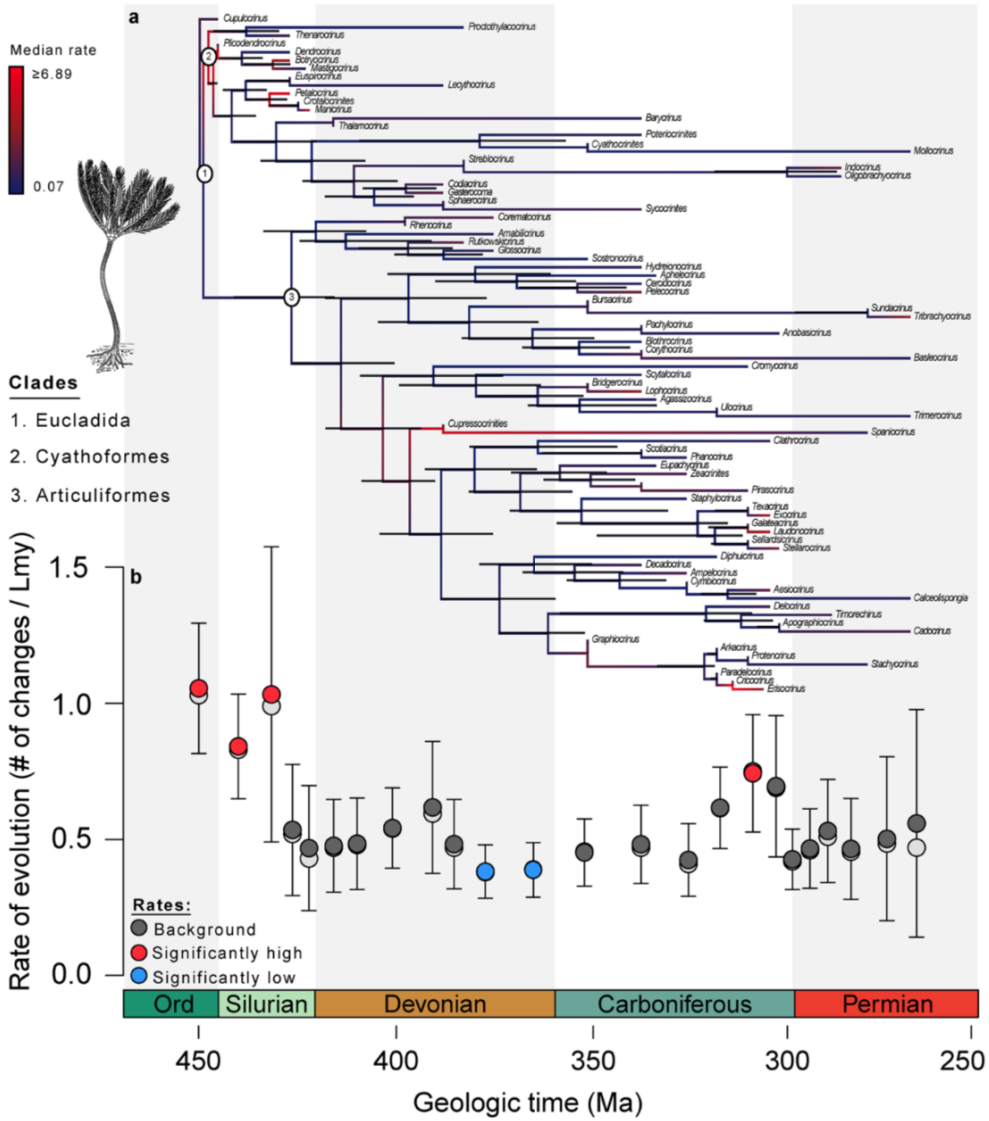
Walker, Heath (2020) *Phylogenetics in the Genomic Era*.

Ignoring sampled ancestors can produce inaccurate parameter estimates



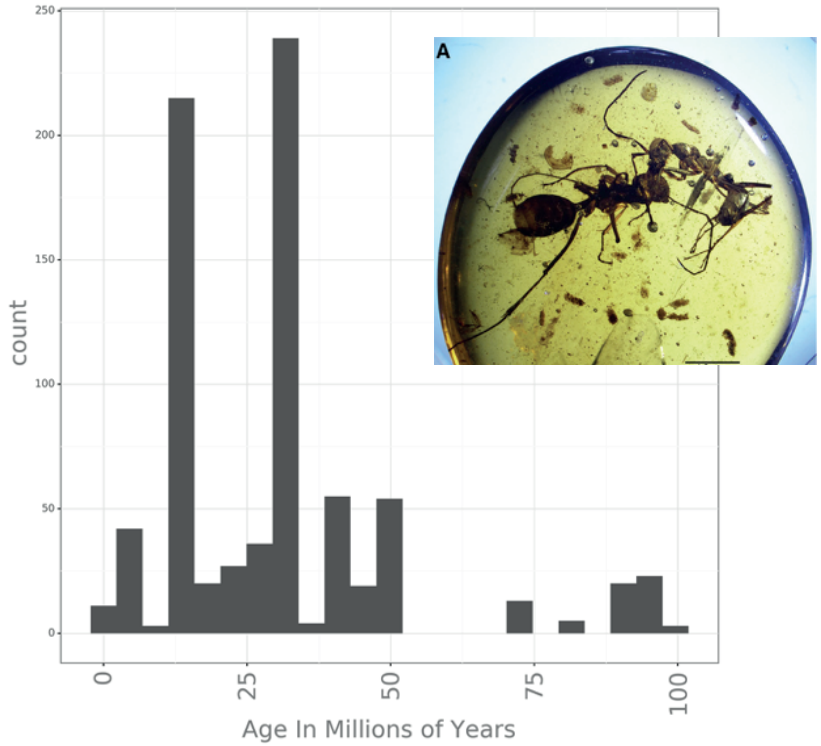
Gavryushkina et al. (2014) *PLoS Comp Bio*

Analysis of fully extinct clades under the FBD process



Example using crinoids Wright (2017) *Sci Reports*

Estimating parameters in macroevolution

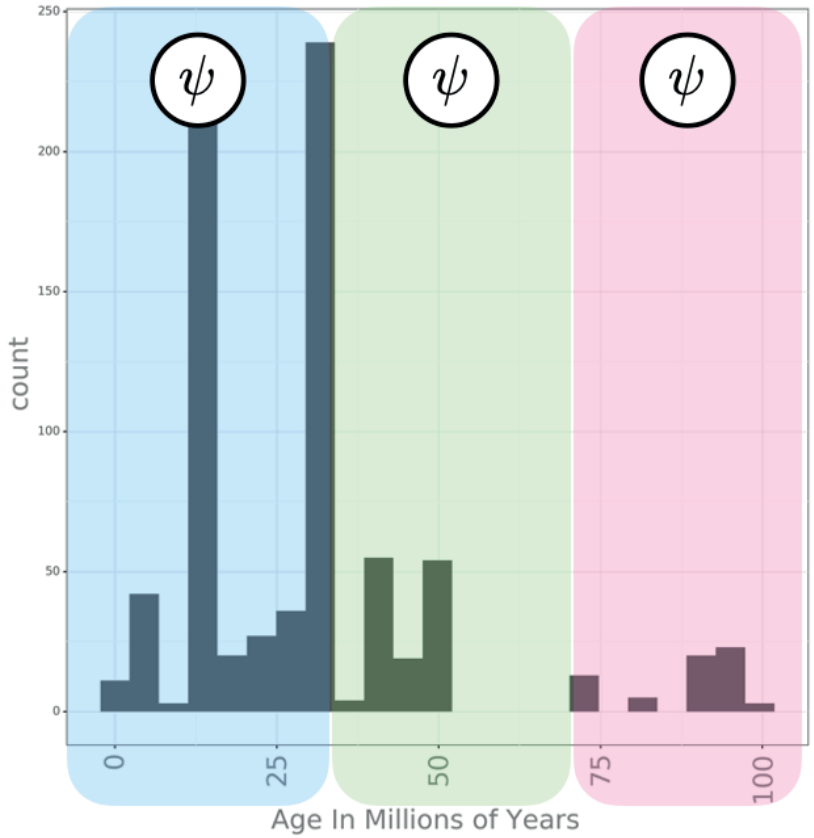


Ants have very variable fossil sampling over time.

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

Images borrowed from April Wright.

Estimating parameters in macroevolution



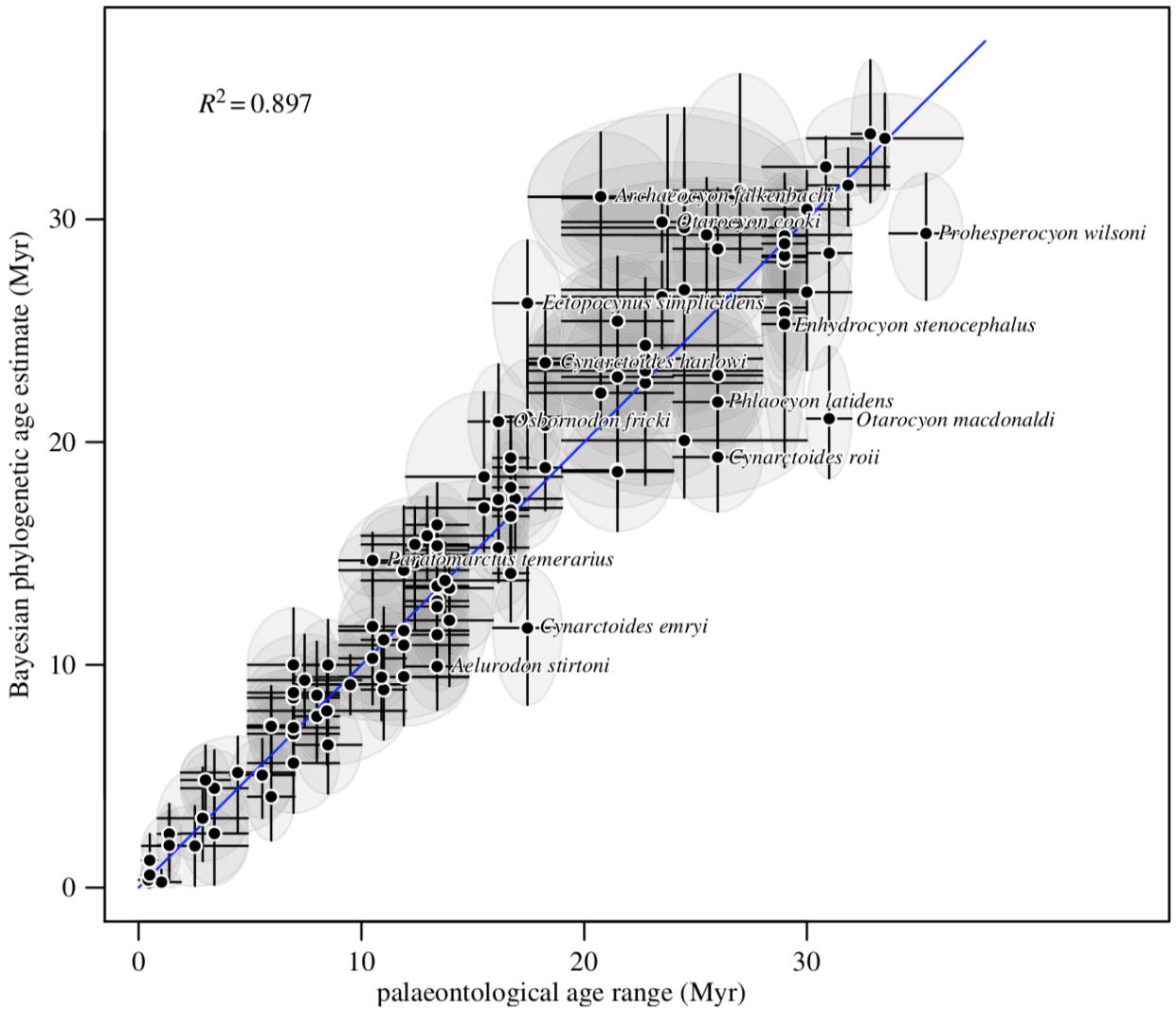
The oldest fossils are around 100 Ma.

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 borrowed from April Wright.

Estimating fossil ages under the FBD process



Example using canids in Drummond, Stadler (2016) *Phil Trans* — Bayesian estimation of fossil ages.

Ignoring stratigraphic age uncertainty leads to the wrong results

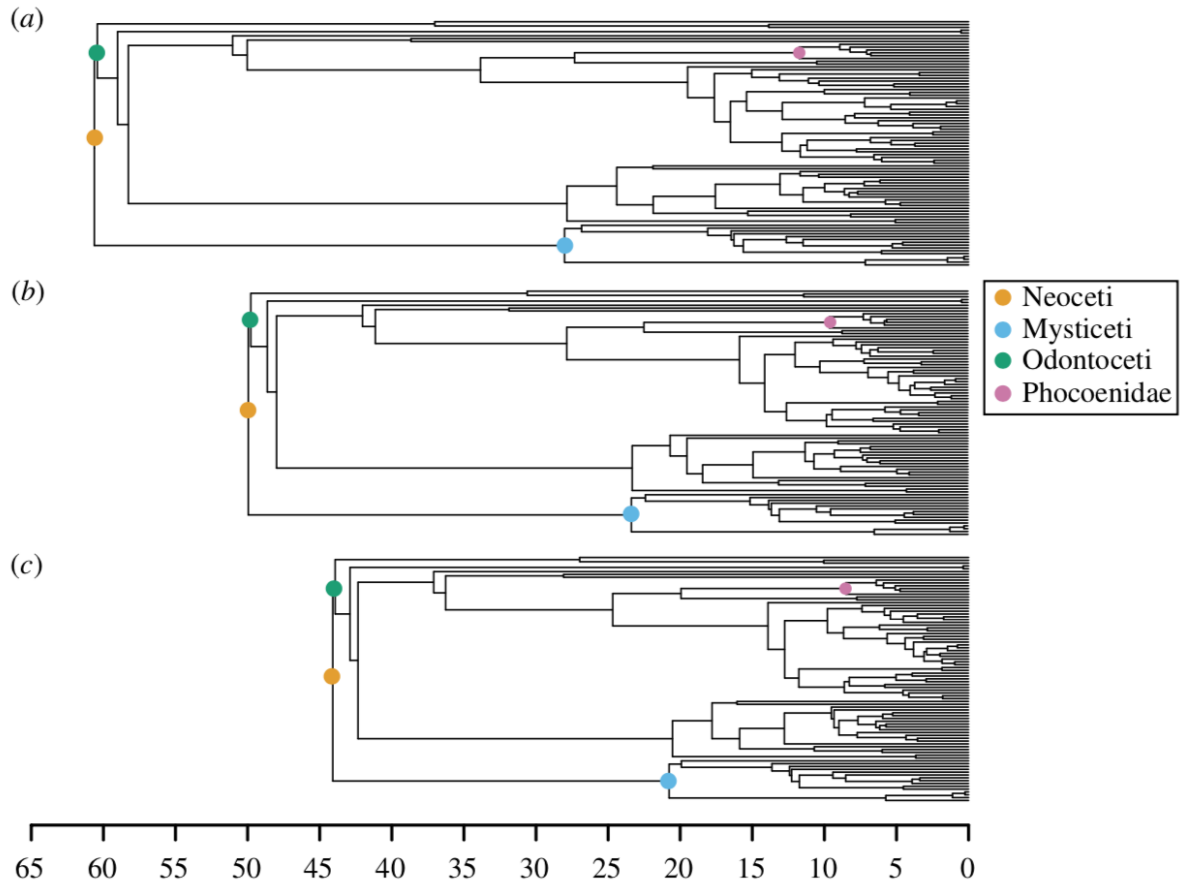


Figure 4. MCC trees inferred for the Cetacea dataset using the FBD process with fossil ages fixed to (*a*) median ages, (*b*) random ages or (*c*) sampled within the known interval of uncertainty. The major clades and the clade shown in figure 5 are highlighted.

Example using cetaceans in Barido-Sottani et al. (2019) *Proc B*

Take homes

The fossilised birth-death process provides a mechanistic framework for phylogenetic dating that has several advantages compared to traditional node dating approaches.

One of the main advantages is that we can incorporate a lot more fossil evidence directly during inference.

We need to carefully consider the underlying assumptions and what we consider data.

Suggested reading

Integrating fossil observations into phylogenetics using the fossilized birth-death model — Wright et al. (2022) → this review provides an overview of the FBD process (use cases and challenges) plus some broader scientific context.

Exercise