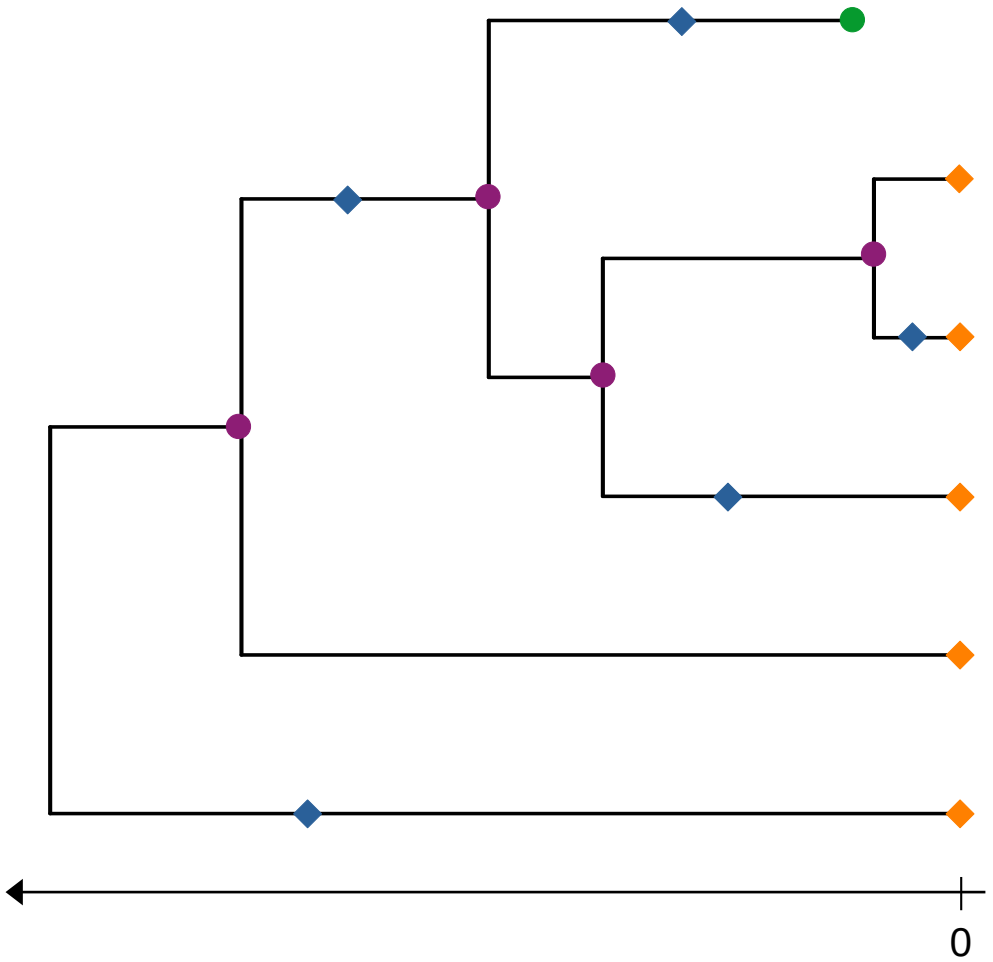
Integrating fossil evidence into phylogenetic dating Extensions and limits

Joëlle Barido-Sottani

The fossilized birth-death (FBD) model

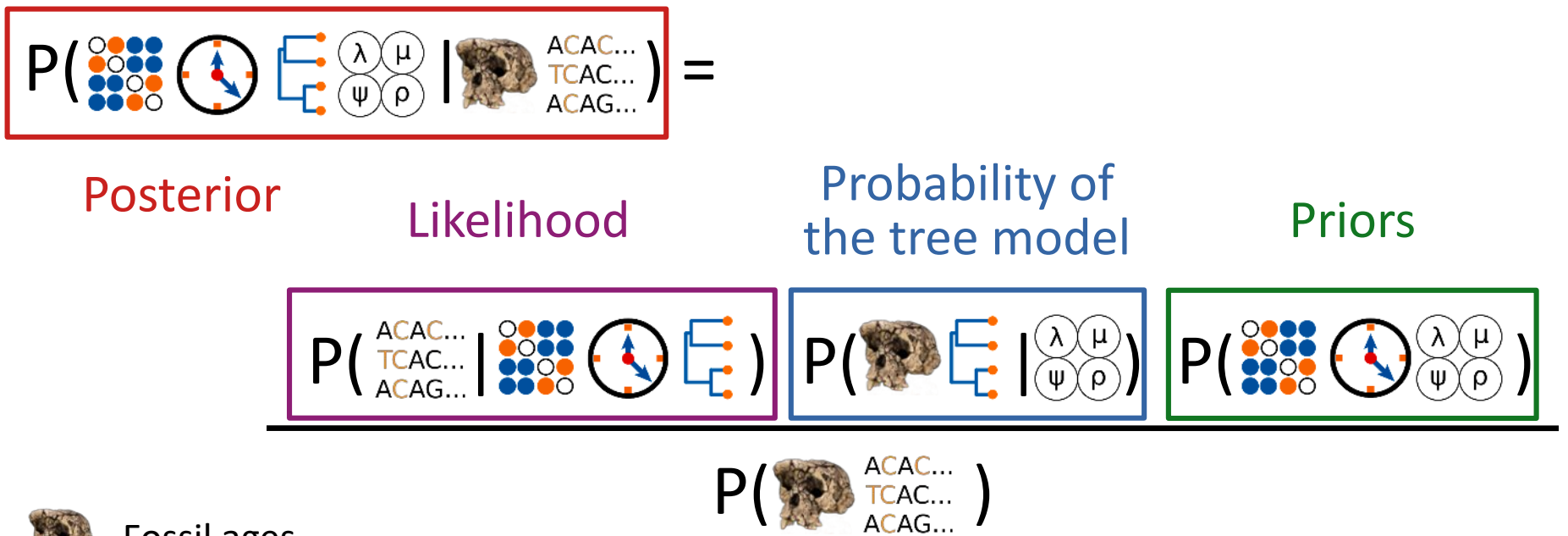


Parameters:

- λ — birth rate
- μ — death rate
- ψ — fossil sampling rate
- ρ — extant species sampling probability

Stadler et al. (2010)

Fossils in Bayesian inference



Fossil ages

Molecular alignment

Substitution model

Clock model

Time tree

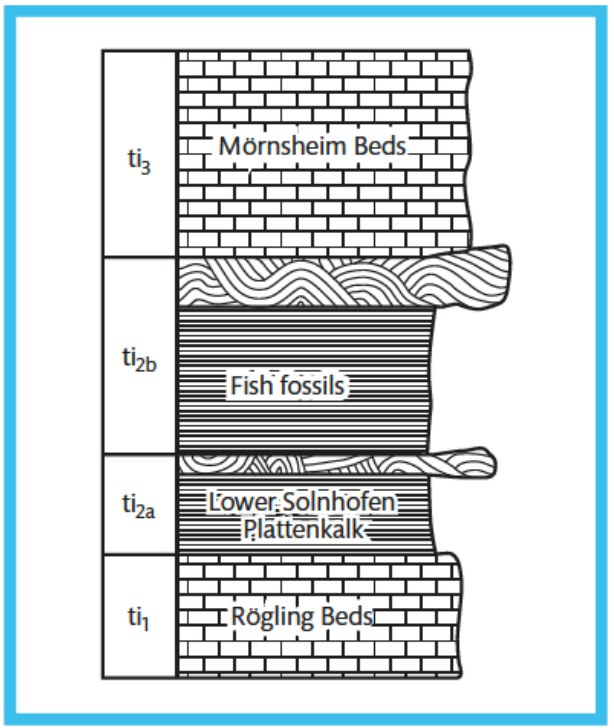
Tree model

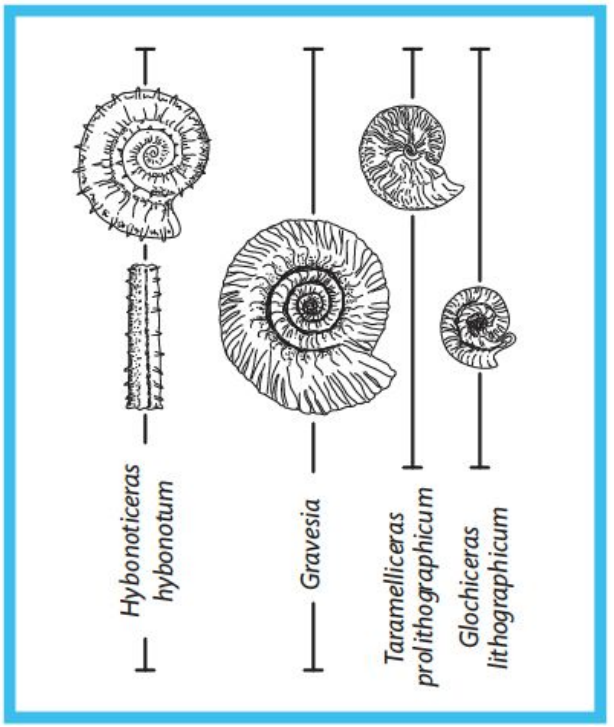
Heath et al. (2014)

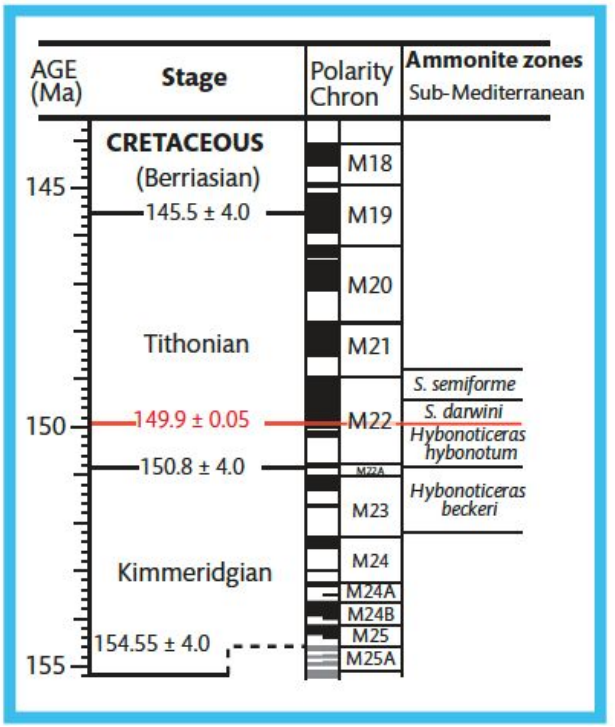
Outline

- Integrating new sources of information
  - Fossil age uncertainty
  - Occurrence data
  - Diversified sampling for large clades
- Correcting and extending model representations
  - Stratigraphic ranges
  - Variations through time
  - Rate heterogeneity

Fossil age uncertainty

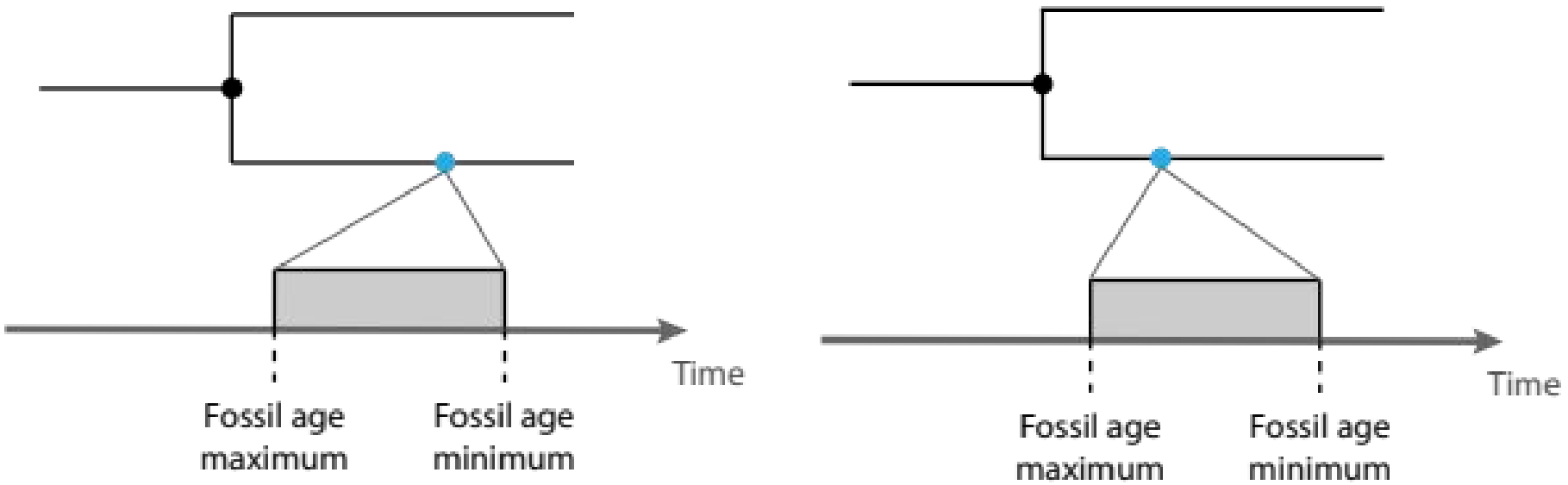






Benton et al. (2009)

Integrating the uncertainty



Fossil age uncertainty can be sampled as part of the MCMC

Drummond & Stadler (2016)

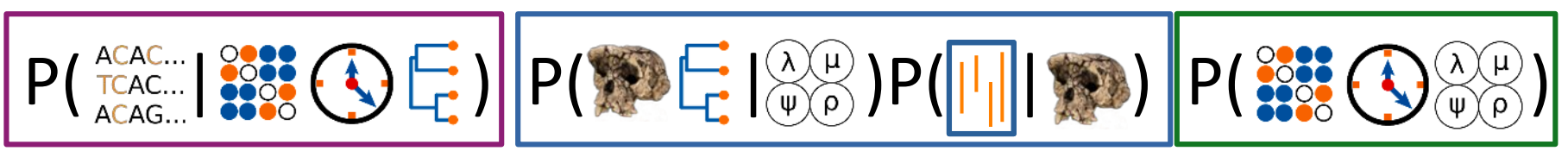
Age uncertainty in Bayesian inference

Posterior

Likelihood

Probability of the tree model

Priors



Fossil age ranges

Drummond & Stadler (2016)

Integrating the uncertainty

Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process

Joëlle Barido-Sottani[1,2,3], Gabriel Aguirre-Fernández[4], Melanie Hopkins[5], Tanja Stadler[1,2] and Rachel Warnock[1,2,4]

Fossil age uncertainty **should** be sampled as part of the MCMC

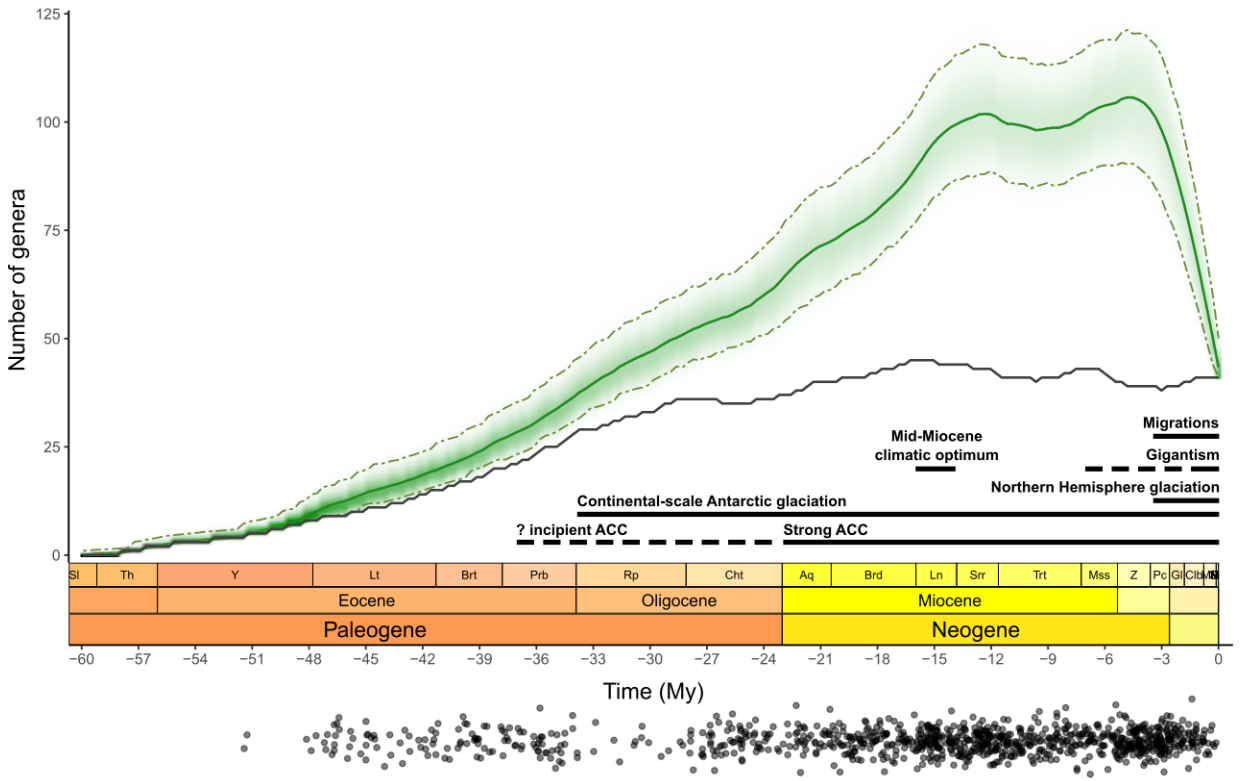
Integrating occurrences

Occurrence = sample without associated sequence information

=> Database of fossil specimens

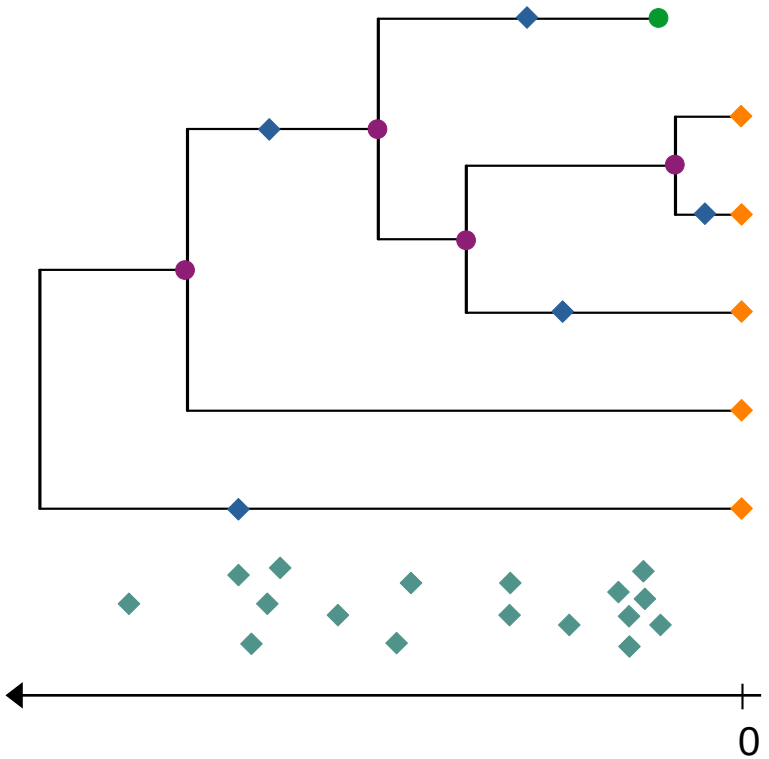
The Paleobiology Database

revealing the history of life



Occurrence birth-death process (OBDP)

Using occurrences to infer the total number of lineages through time and inform the population parameter estimates



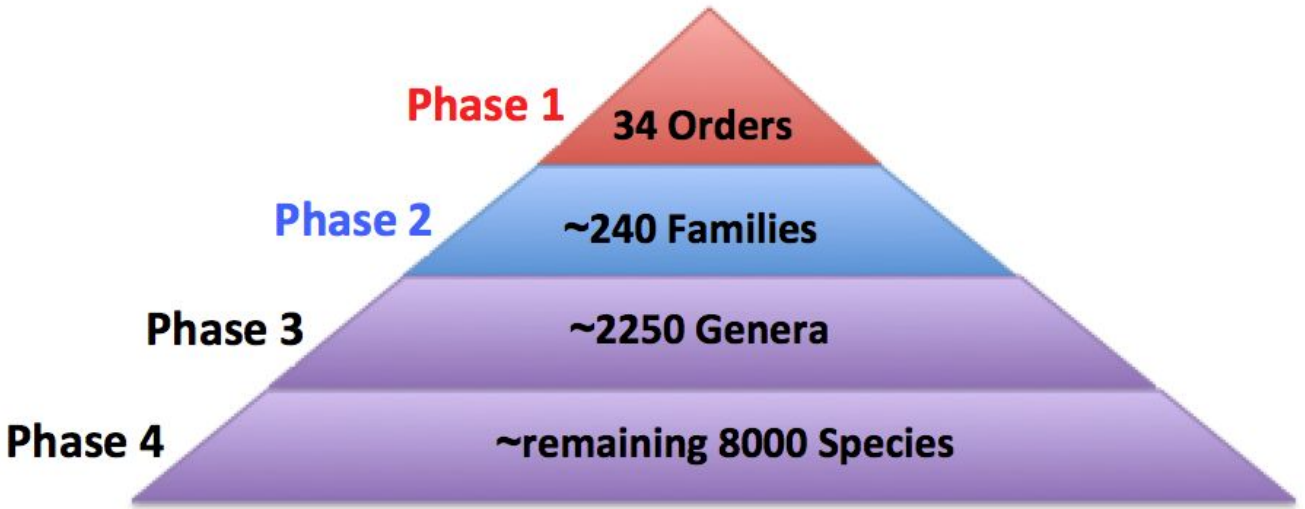
- λ — birth rate
- μ — death rate
- ψ — fossil sampling rate
- ρ — extant species sampling probability
- ω – occurrence sampling rate

Gupta et al. (2020), Manceau et al. (2021), Andréoletti et al. (2022), Zarebski et al. (2023)

The problem of size

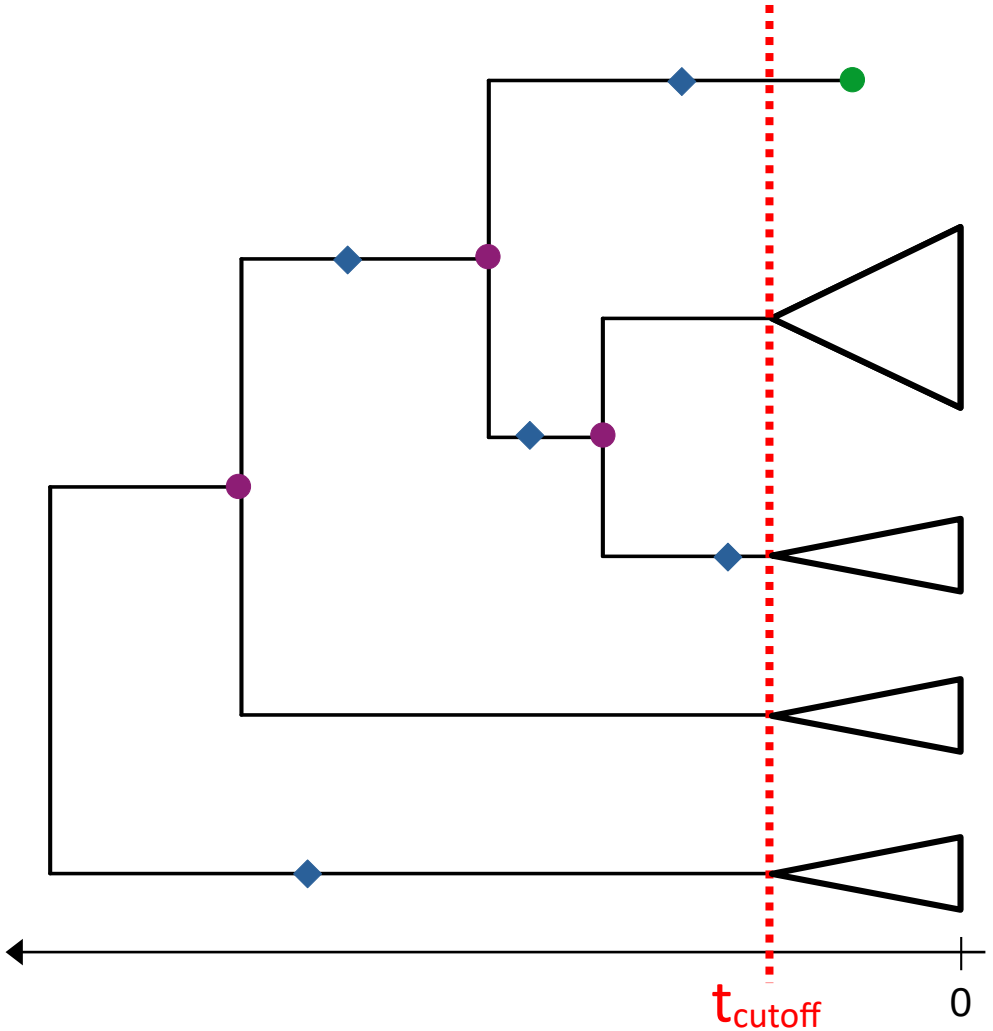
Bayesian FBD inference : up to ≈500-800 samples

VS



Source : Bird 10,000 Genomes (B10K) Project

FBD model with diversified sampling

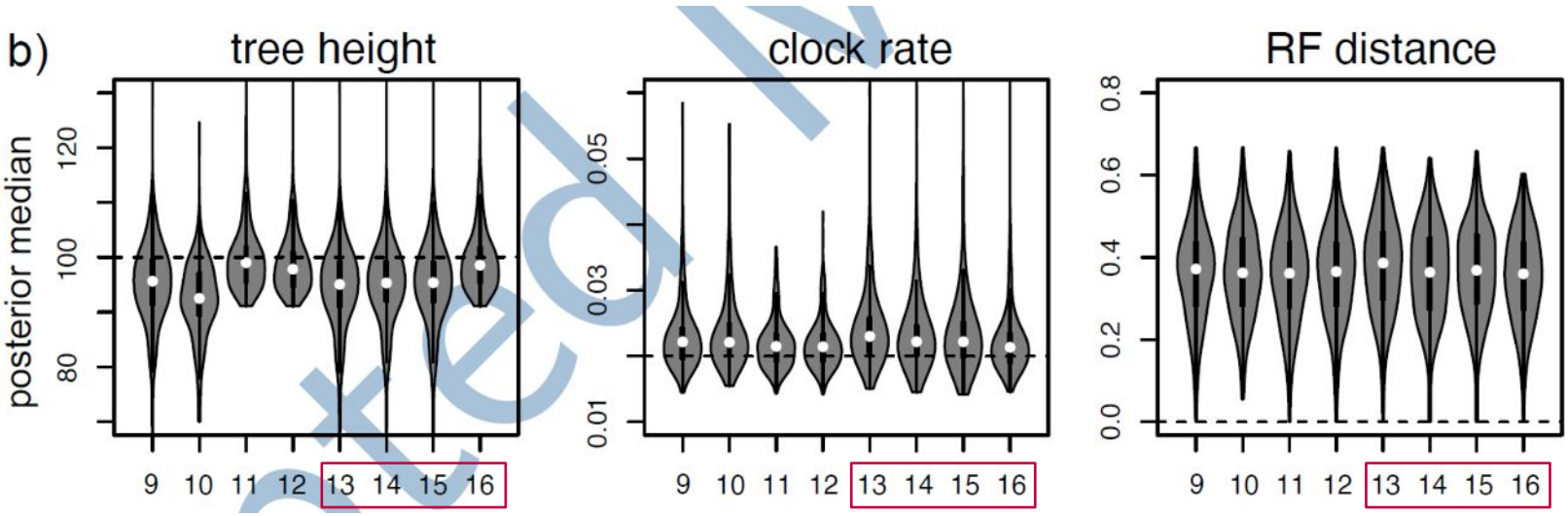


Parameters:

- λ — birth rate
- μ — death rate
- ψ — fossil sampling rate
- N — total number of extant species

Zhang et al. (2016)

Does the sampling process matter?



Simulated cutoff sampling

Mismatched inference

Zhang et al. (2023)

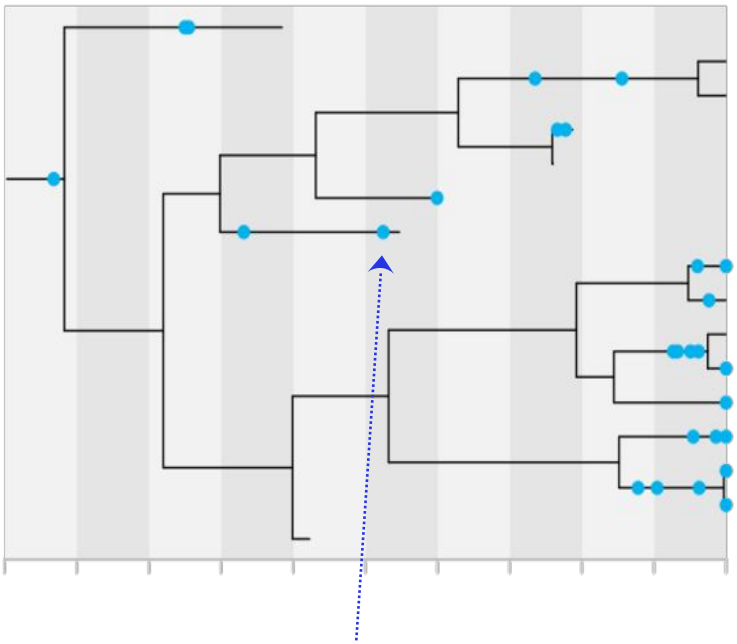
Caveats:

- Diversification rates are biased by sampling mismatch
- Simulated trees tend to be more balanced than real trees

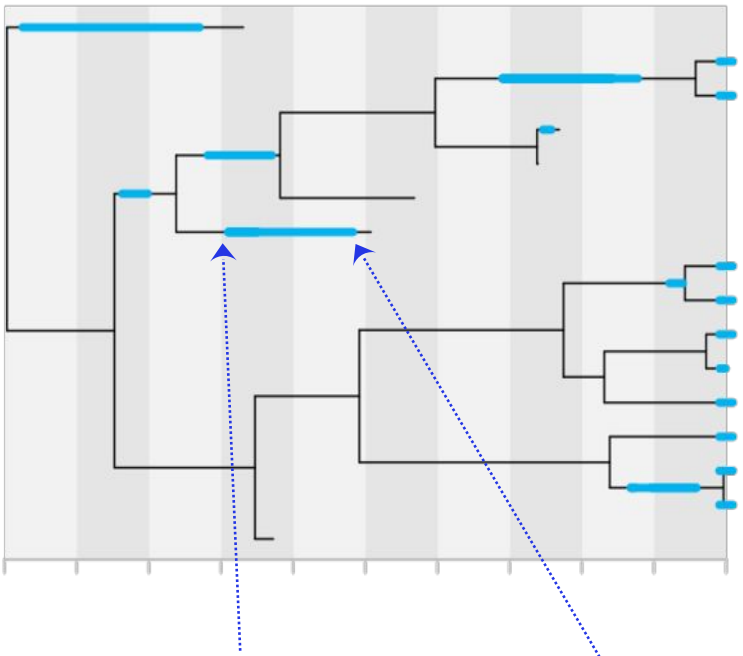
Outline

- Integrating new sources of information
  - Fossil age uncertainty
  - Occurrence data
  - Diversified sampling for large clades
- Correcting and extending model representations
  - Stratigraphic ranges
  - Variations through time
  - Rate heterogeneity

Specimen-level data vs range data



Specimen (one occurrence)

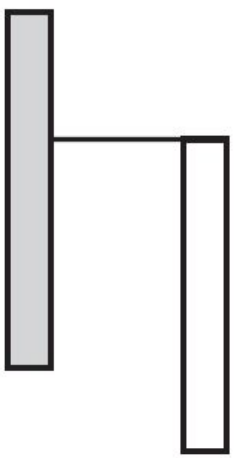


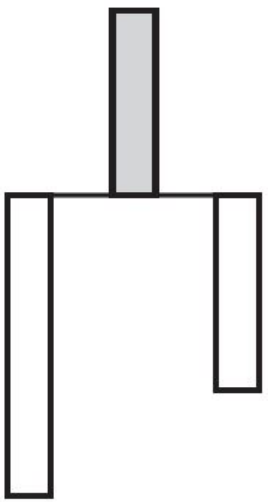
First occurrence

Last occurrence

Fossil species ≠ Tree lineages

(i) asymmetric speciation (ii) symmetric speciation (iii) anagenetic speciation





1 birth event

1 speciation event

1 birth event

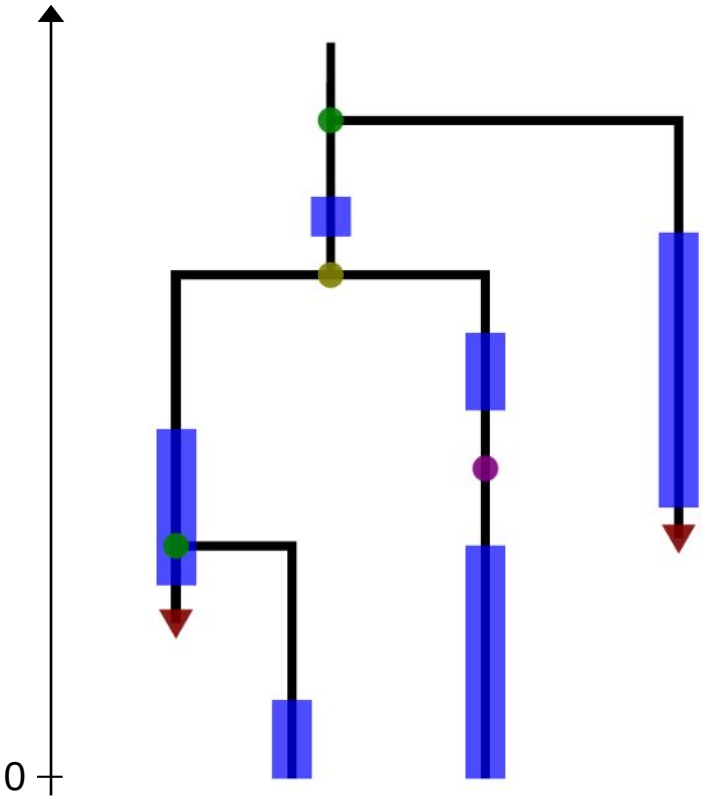
2 speciation events
1 extinction event

0 birth event

1 speciation event
1 extinction event

Stadler et al. (2018)

The FBD for stratigraphic ranges



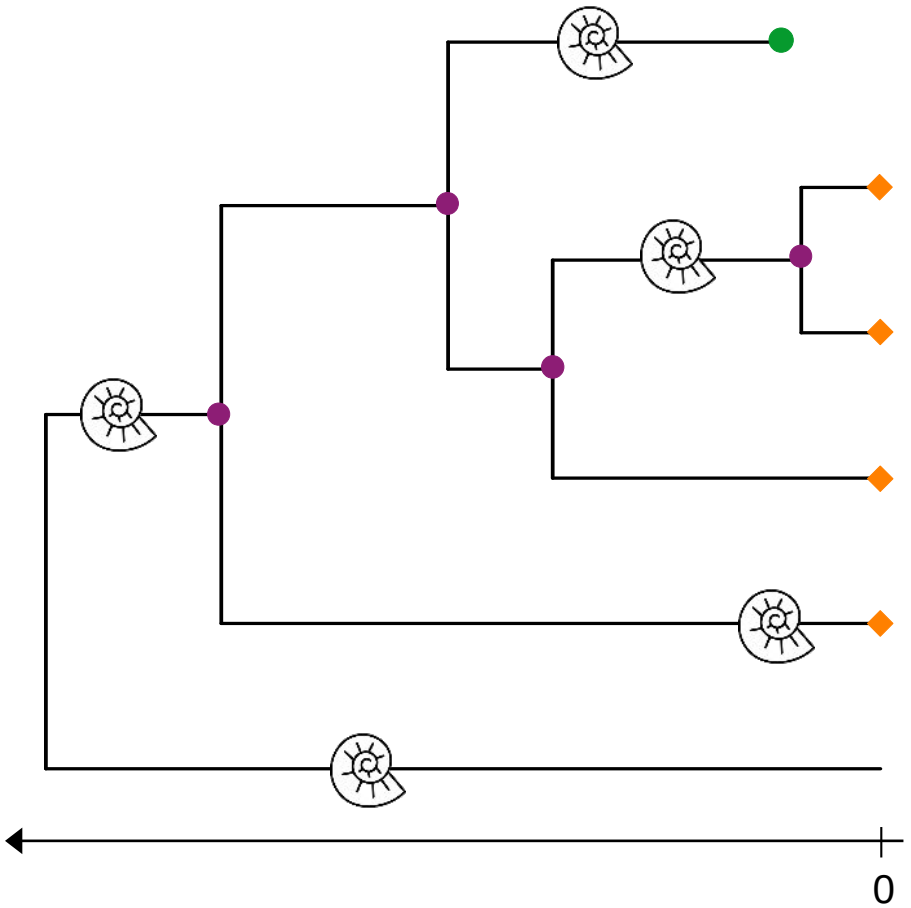
FBD-range model

- Budding speciation
- Bifurcating speciation
- Anagenetic speciation
- Extinction

Remaining challenges

- Full implementation
- Identifiability of the different processes
- Combination with morphological data
- ……

The (homogeneous) FBD process

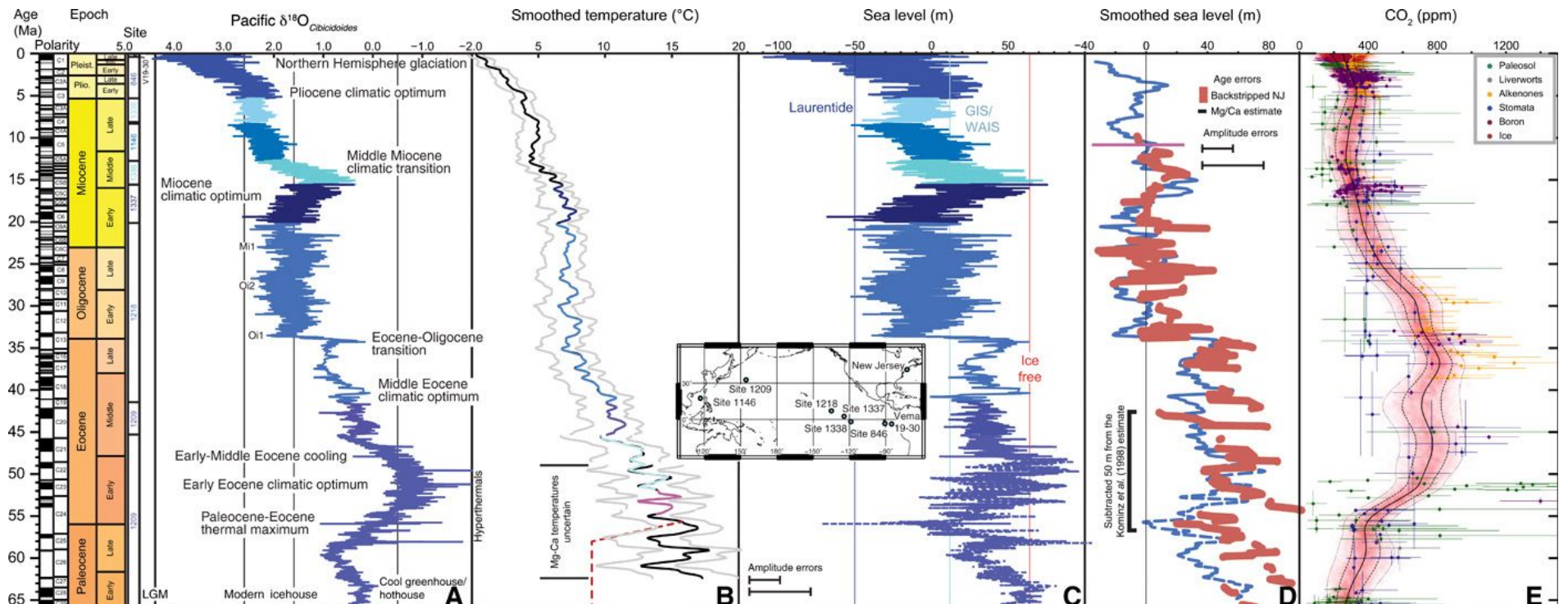


Parameters:

- λ — (constant) birth rate
- μ — (constant) death rate
- ψ — (constant) fossil sampling rate
- ρ — extant species sampling probability

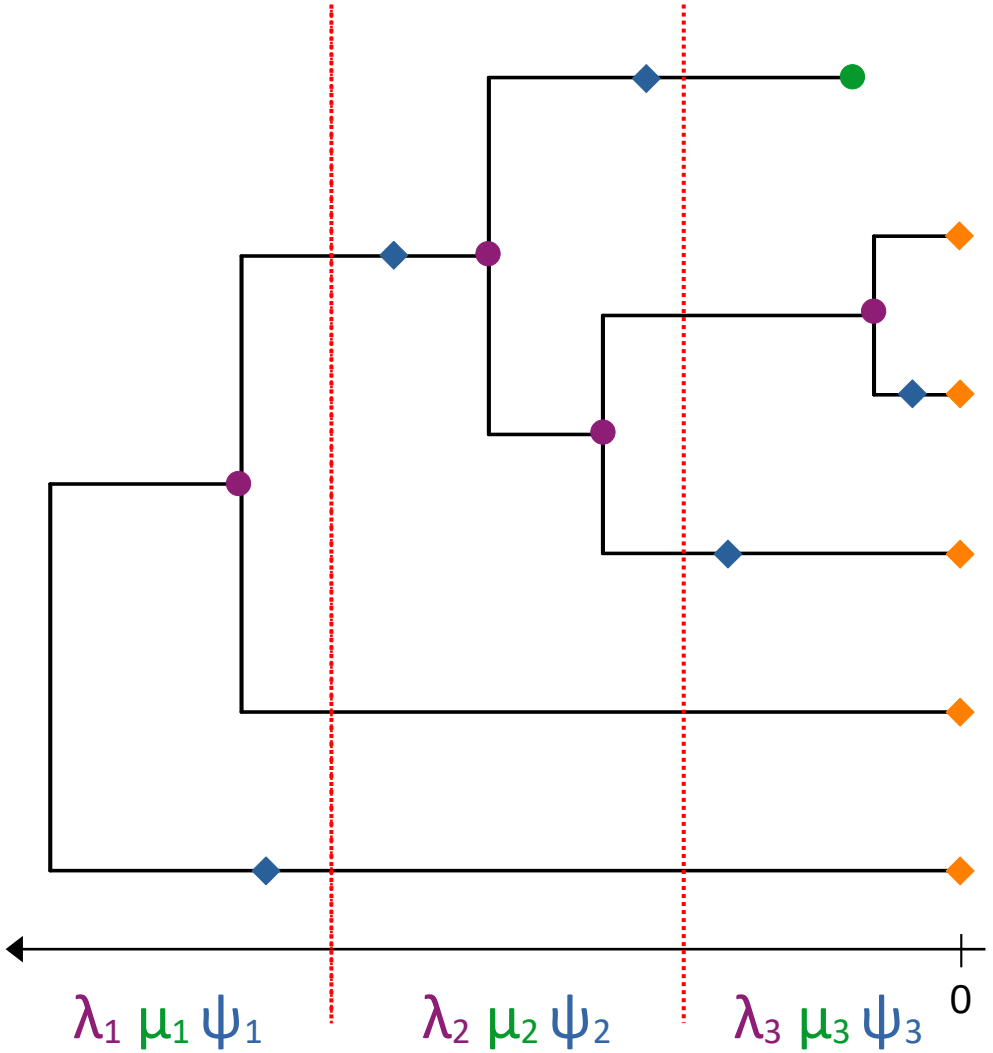
Variations in time

Many factors change through time: continents, climatic conditions, sea levels, etc.



Miller *et al.* (2020)

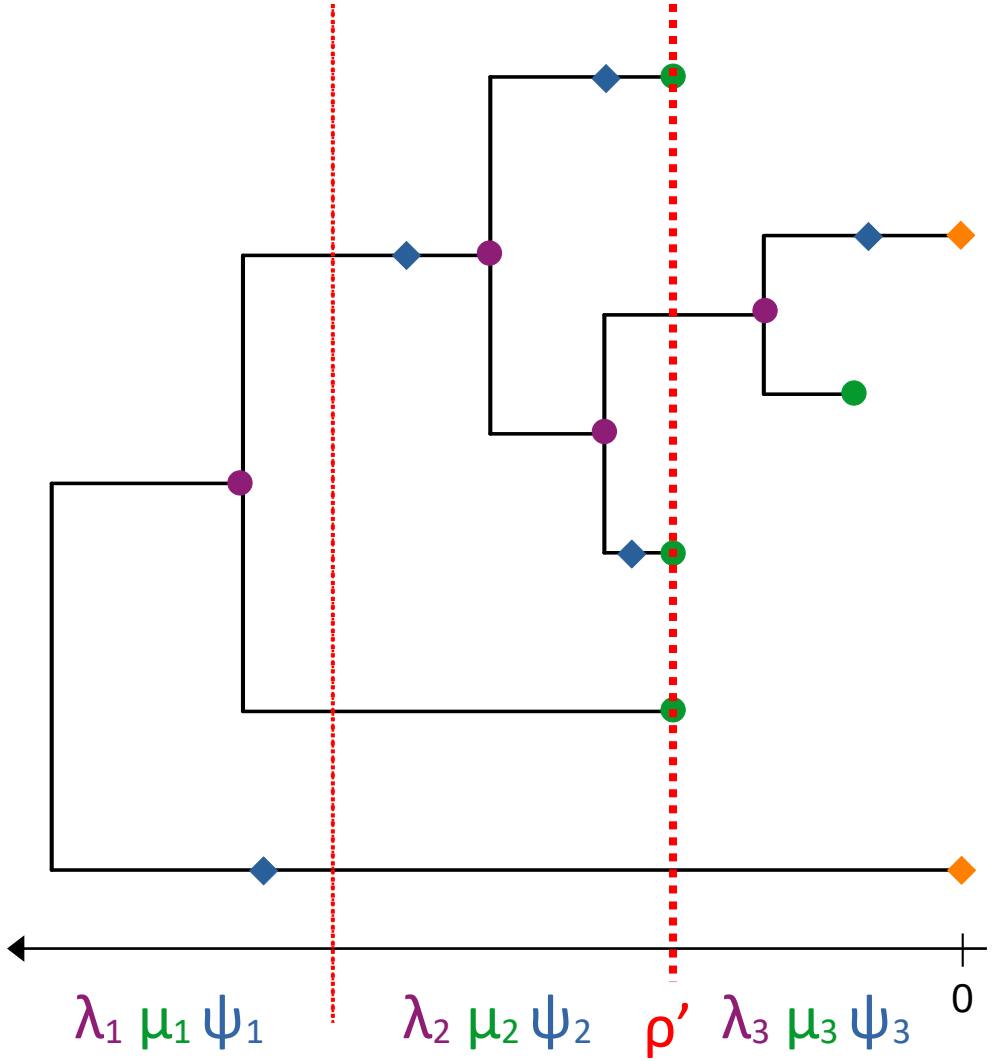
Skyline FBD model



- Allows for punctual shifts in rates through time
  => piecewise-constant rates
- Interval boundaries can be
  - fixed to specific ages
  - fixed to specific events (e.g. tree origin)
  - inferred by the method

Stadler (2011)

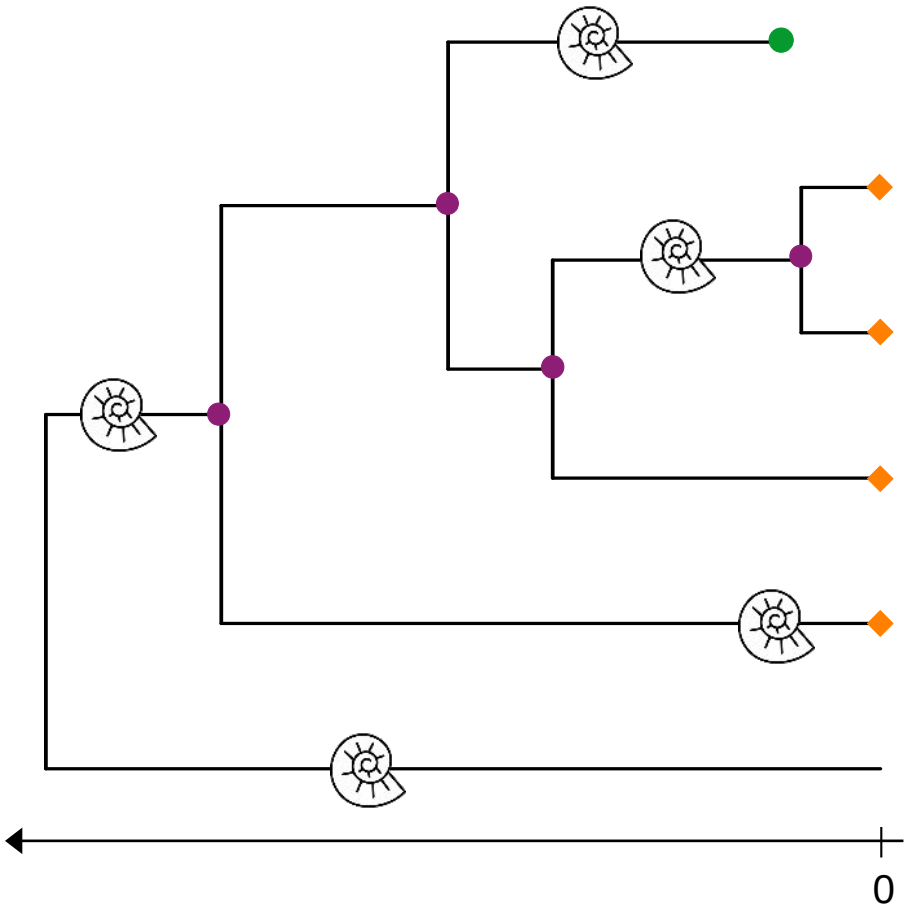
Mass extinction events



Parameters:

- λ — birth rate
- μ — death rate
- ψ — fossil sampling rate
- ρ — extant species sampling probability
- **ρ' — mass extinction probability**

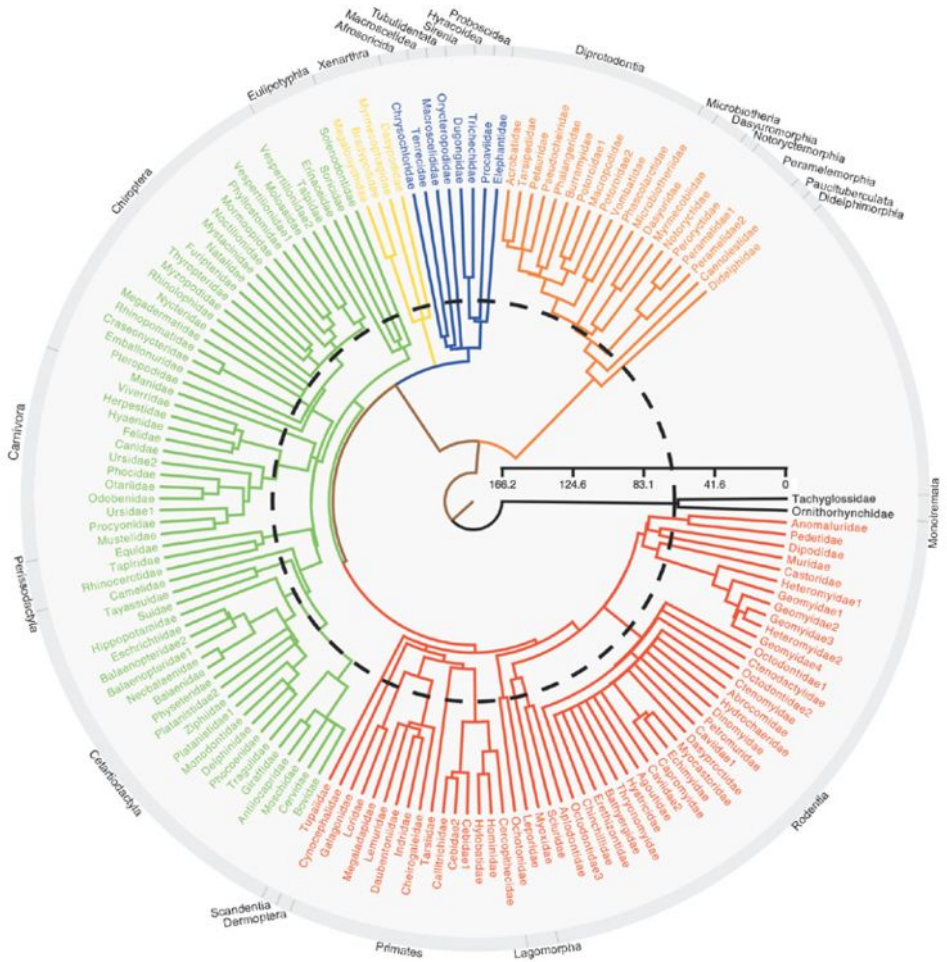
The (homogeneous) FBD process



Parameters:

- λ — (constant) birth rate
- μ — (constant) death rate
- ψ — (constant) fossil sampling rate
- ρ — extant species sampling probability

Diversification – an homogeneous process ?



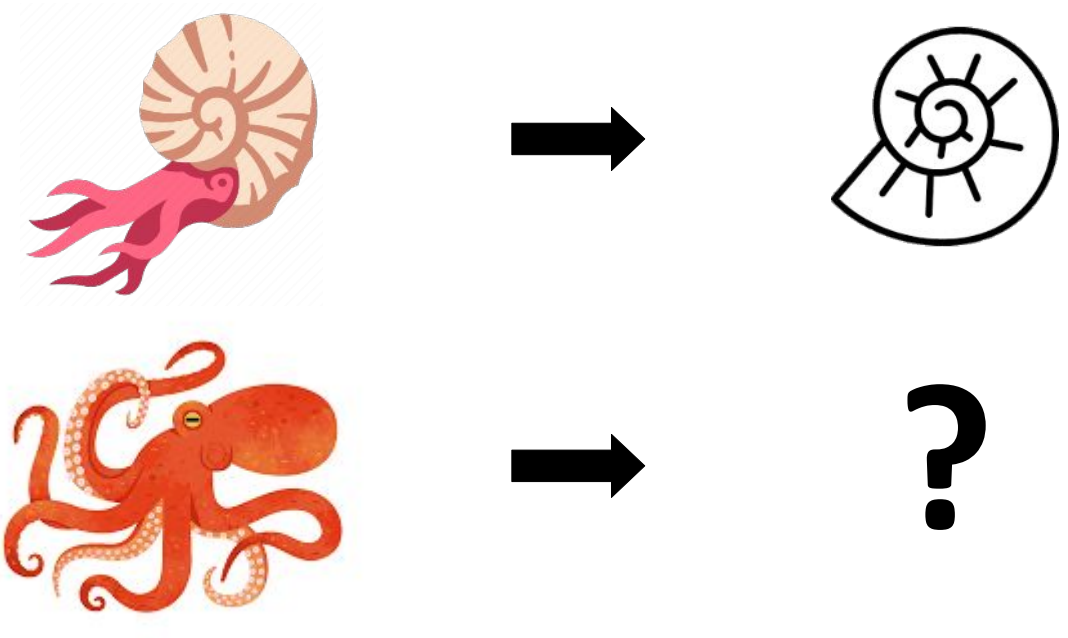
Bininda-Emonds *et al.* **Nature 2007**

Traits proposed as driving rate variations :

- body size
- environment
- mating system

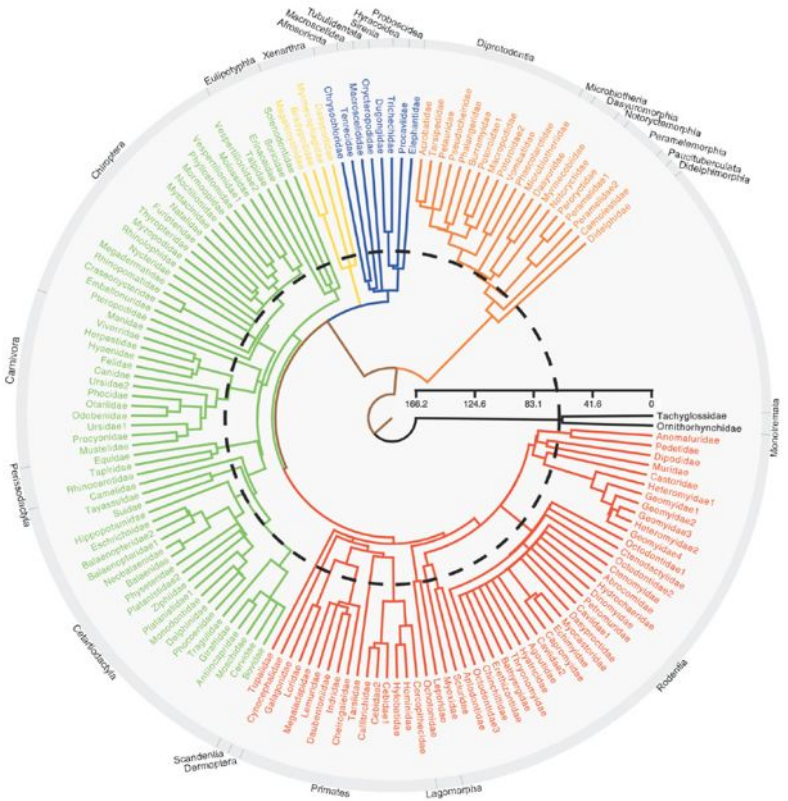
And many others……..

Fossilization – an homogeneous process ?



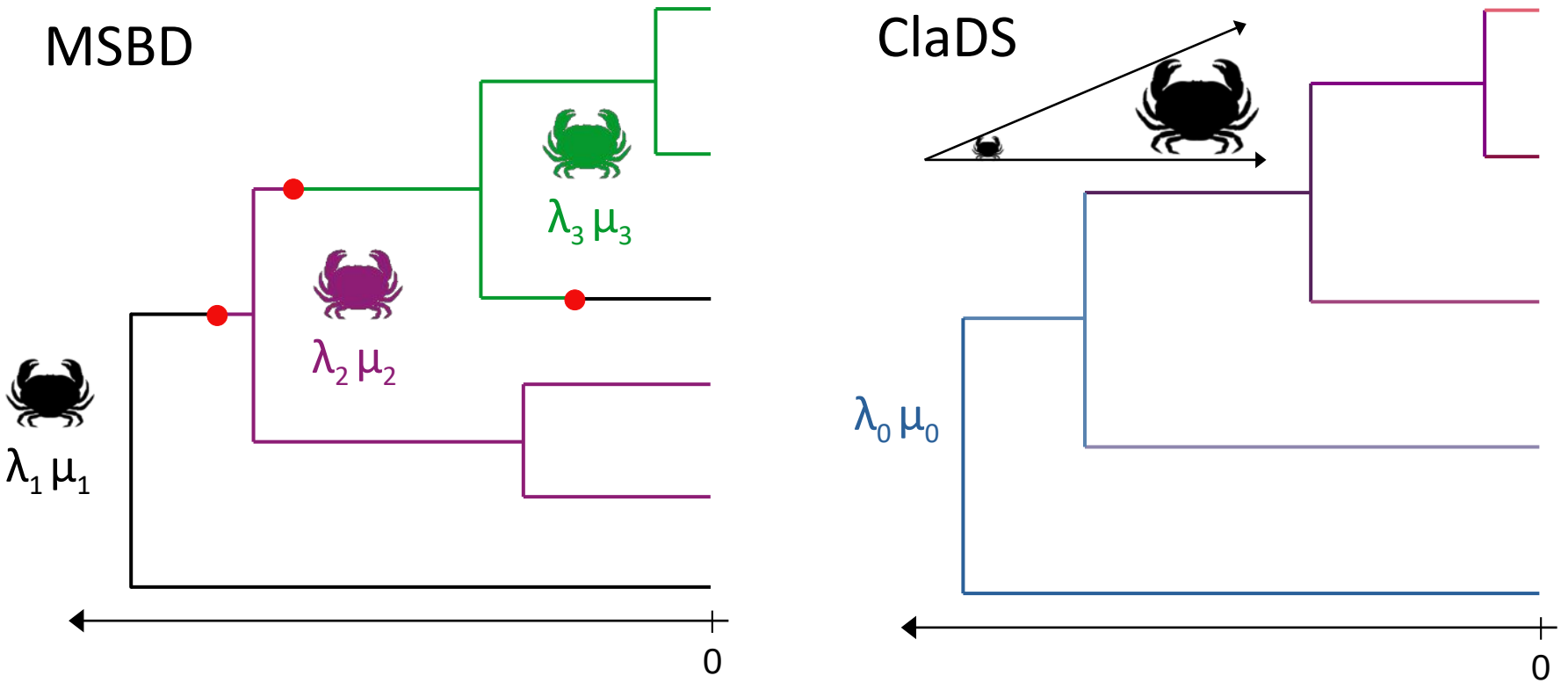
- Taphonomic biases: soft tissue degradation, environmental and time differences in sedimentation
- Sampling biases: geographical differences, identification issues

How can we integrate these variations in our inferences ?



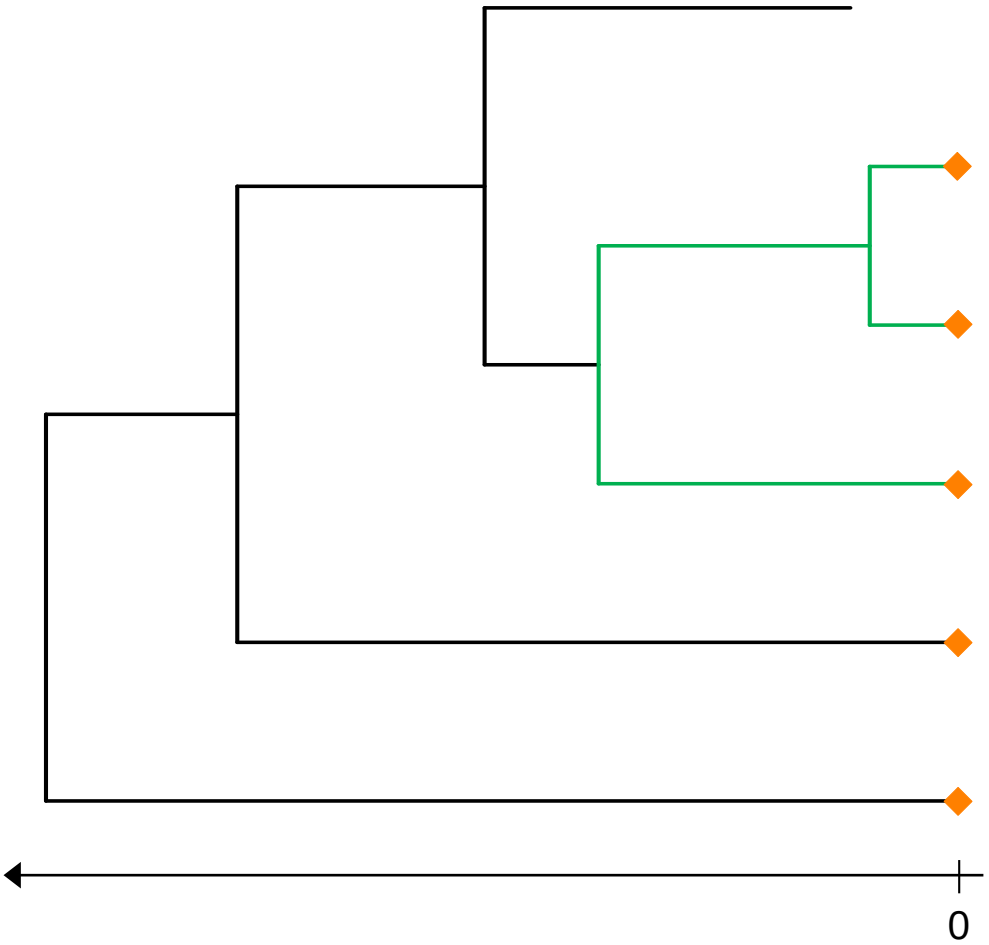
?

Heterogeneous BD models



Barido-Sottani *et al.* Sys. Bio. 2020, Barido-Sottani & Morlon. Sys. Bio. 2023

Multi-type birth-death (MTBD) process



2 types, type 1 & type 2

λ_1 & λ_2 — birth rates

μ_1 & μ_2 — death rates

ρ — extant species sampling probability

Character-dependent or independent ?

λ_1 λ_2 λ_3

μ_1 μ_2 μ_3

Type 1 Type 2 Type 3

Character-dependent

? ? ?

λ_1 λ_2 λ_3

μ_1 μ_2 μ_3

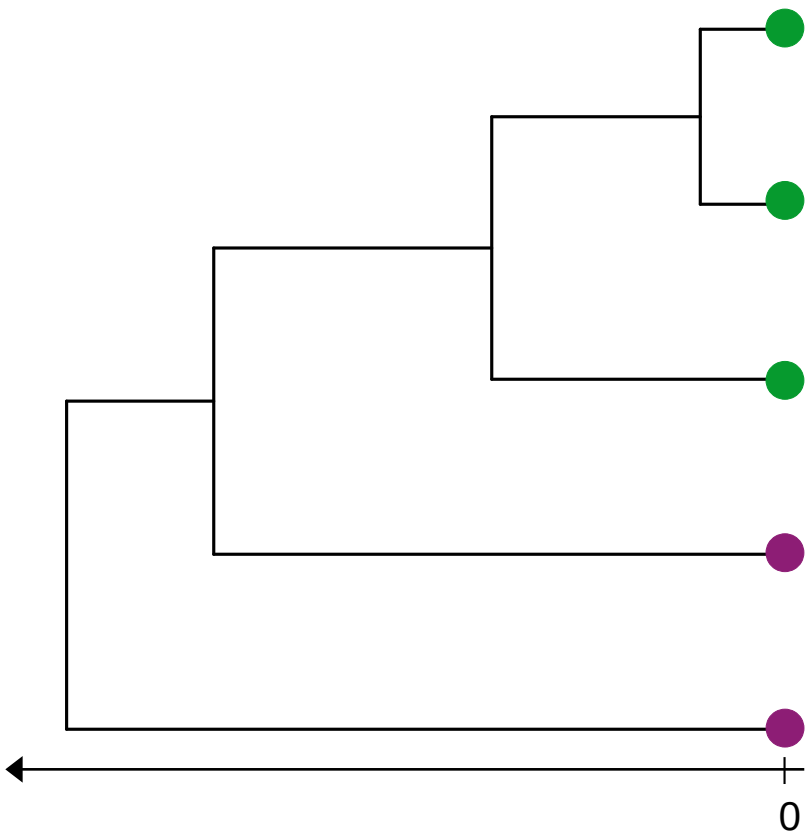
Type 1 Type 2 Type 3

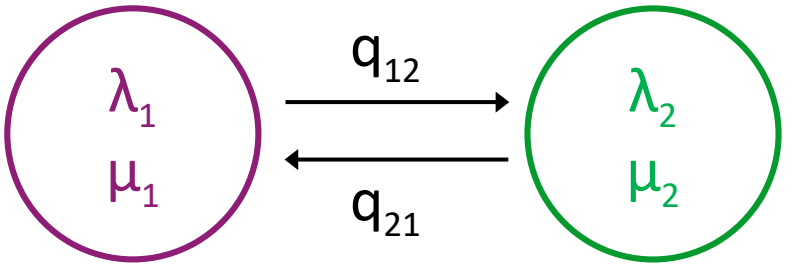
Character-independent

In a character-dependent model :

- The number of types is known
- The type at the tips is known

The BiSSE/MuSSE/BDMM model



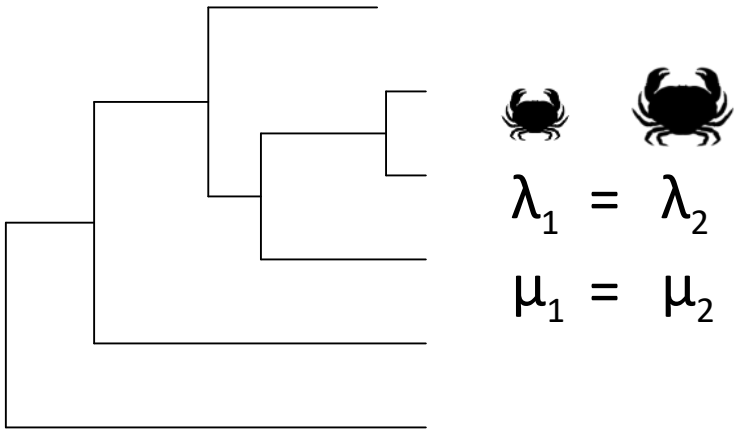


Parameters of the model:

- λ_i – birth rates
- μ_i – death rates
- q_{ij} – transition rates
- ρ/p – sampling probability

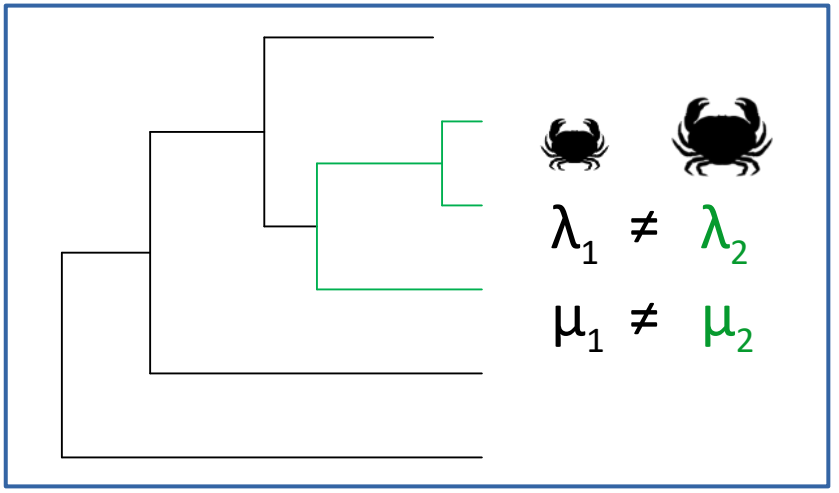
Maddison *et al.* **Sys. Bio. 2007**
Fitzjohn *et al.* **Sys. Bio. 2009**
Kühnert et al. **MBE 2016**

Model selection issues



$\lambda_1 = \lambda_2$

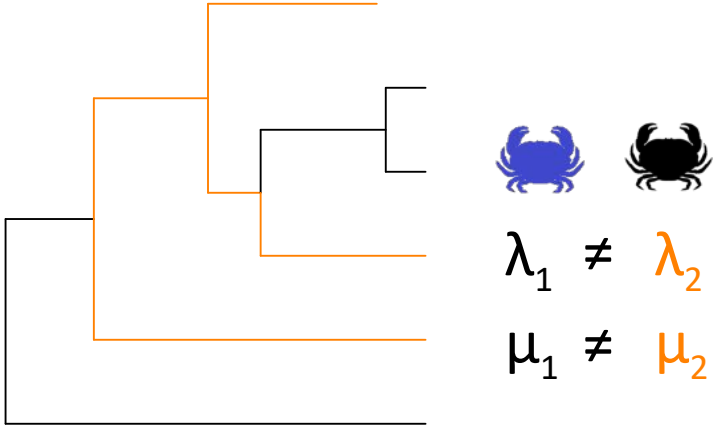
$\mu_1 = \mu_2$



$\lambda_1 \neq \lambda_2$

$\mu_1 \neq \mu_2$

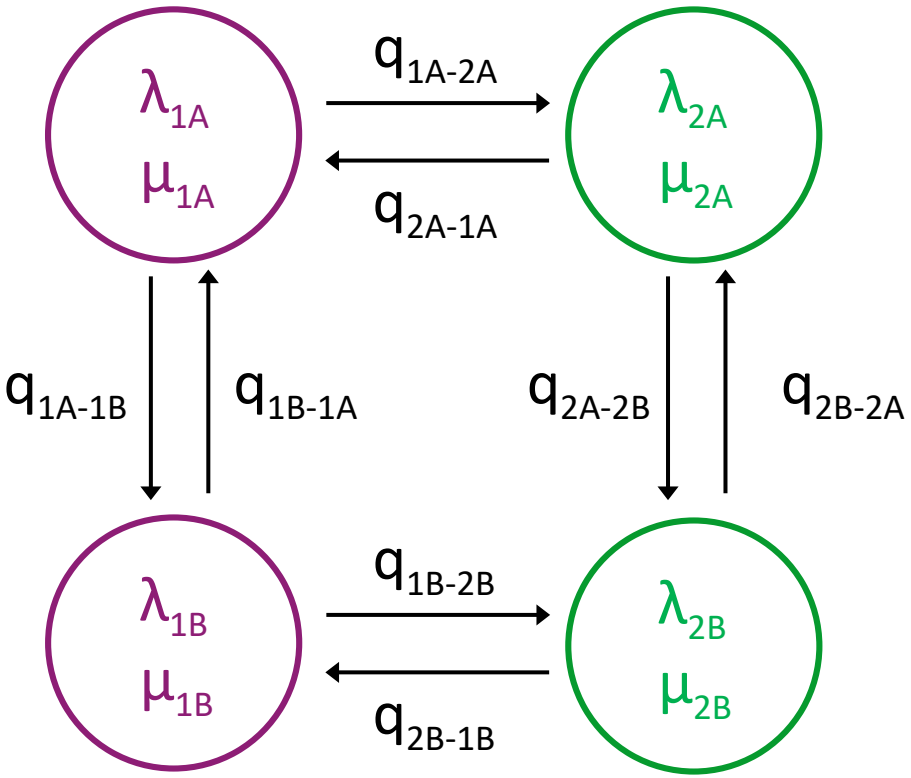
Rabosky & Goldberg 2015, Sys. Bio.



$\lambda_1 \neq \lambda_2$

$\mu_1 \neq \mu_2$

The HiSSE model



- Hybrid model with a hidden trait (A/B) added to the user-defined trait (1/2)
- Allows to distinguish whether the user-chosen character is linked to the rate variation
- Remaining issues:
  - The number of values for the hidden trait is chosen by the user
  - Higher complexity of the model

Beaulieu & O'Meara **Sys. Bio. 2016**

Character-dependent or independent ?

λ_1 λ_2 λ_3

μ_1 μ_2 μ_3

Type 1 Type 2 Type 3

Character-dependent

? ? ?

λ_1 λ_2 λ_3

μ_1 μ_2 μ_3

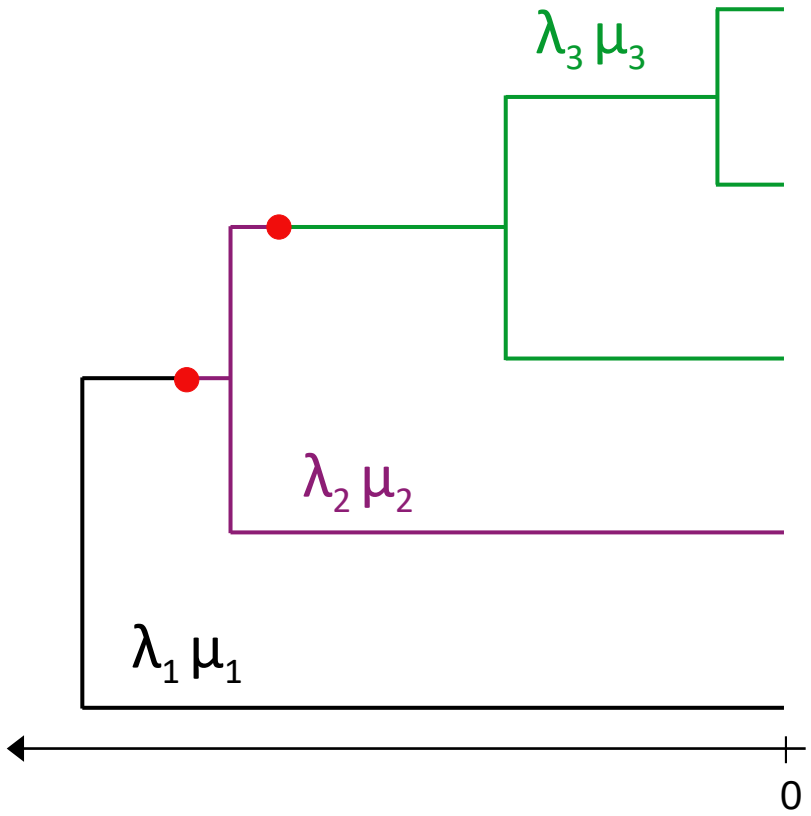
Type 1 Type 2 Type 3

Character-independent

In a character-dependent model :

- The number of types is known
- The type at the tips is known

BAMM/MSBD model

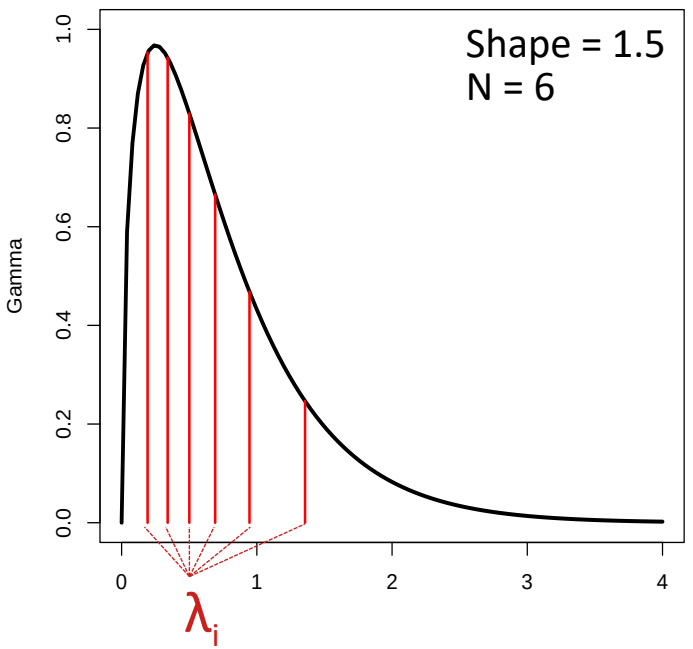


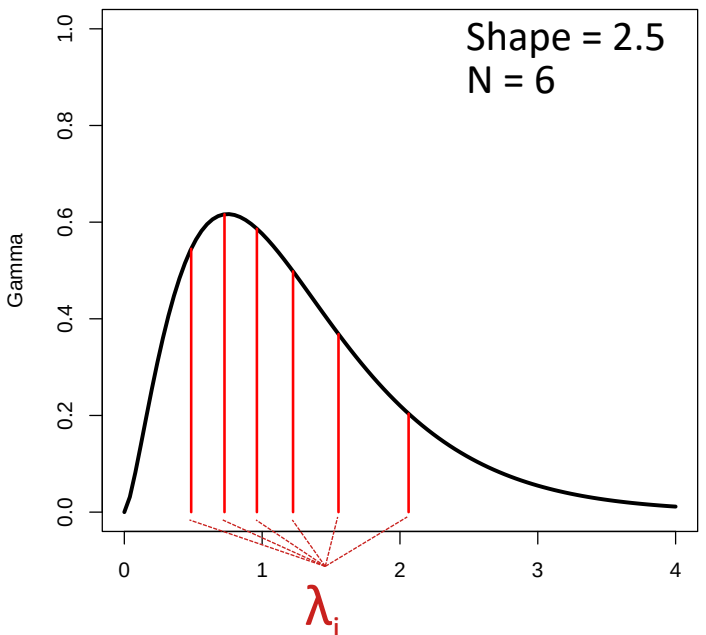
- Character-independent version of SSE
- New estimated parameters:
  - N total number of types
  - Types of edges and tips
- Simplified transition process: ●
  - Each transition is a new type (BAMM)
  - Constant transition rate ɣ (MSBD)

Rabosky *et al.* **Nat. Comm. 2013**
Barido-Sottani *et al.* **Sys. Bio. 2020**

RevBayes model

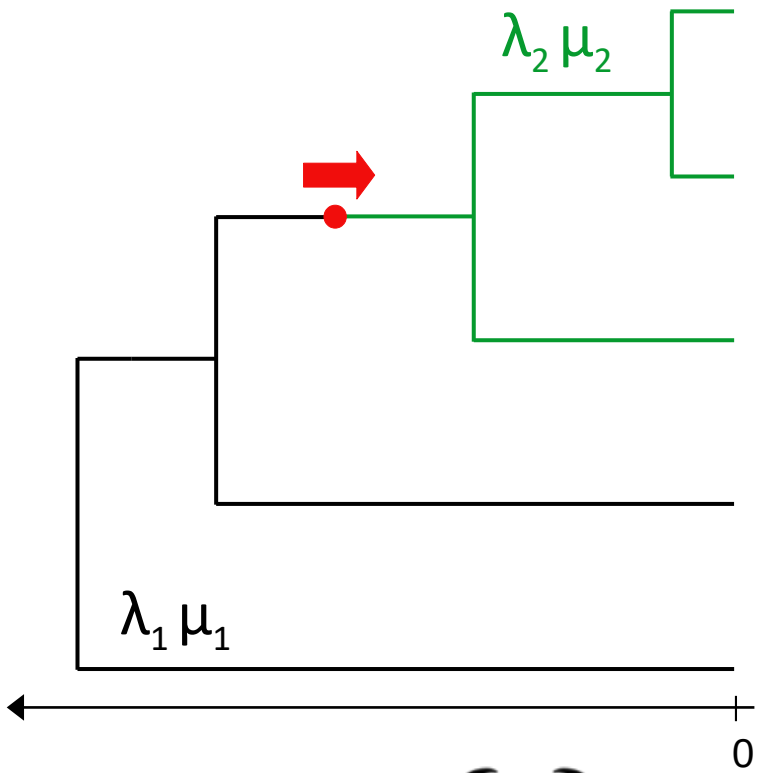
- Ordered types based on a Gamma distribution
- Fixed number of types N
- Simplified model: rates are not estimated, but determined by the shape of the Gamma distribution

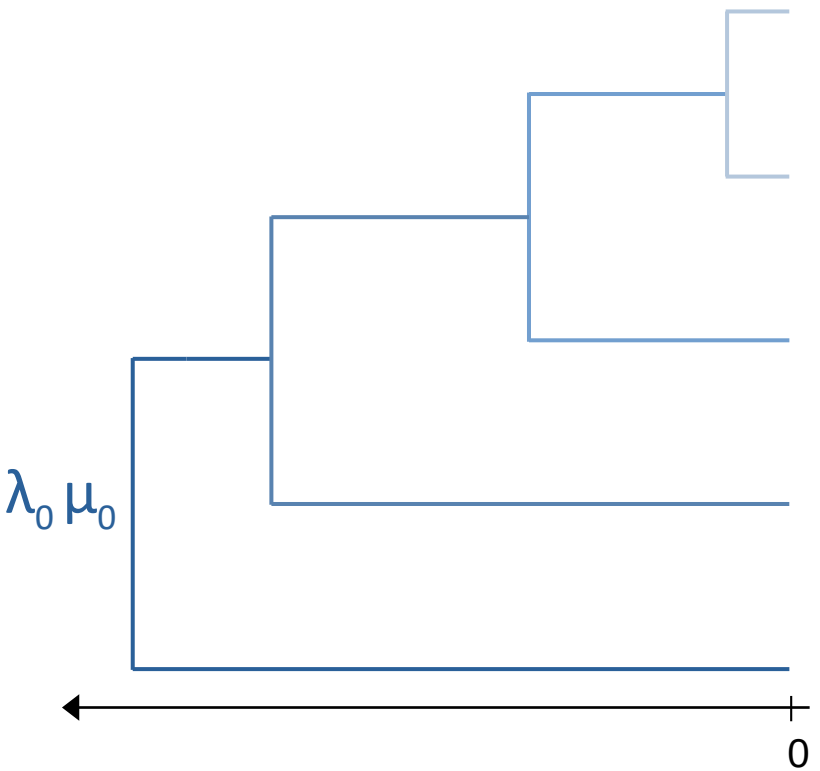


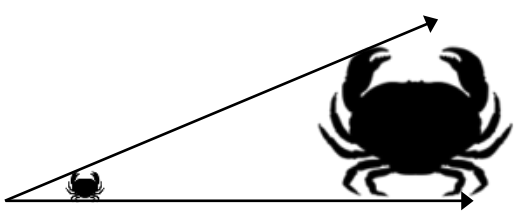


Höhna *et al.* **BiorXiv 2019**

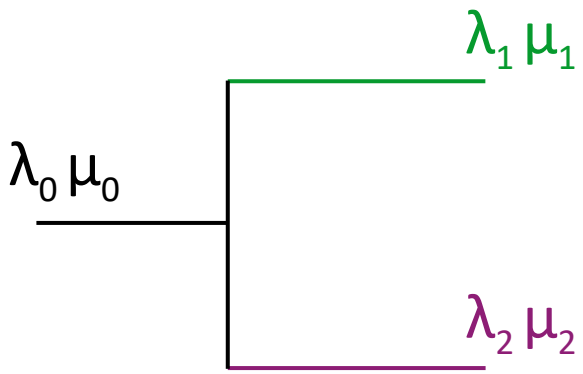
Going beyond types







ClaDS model



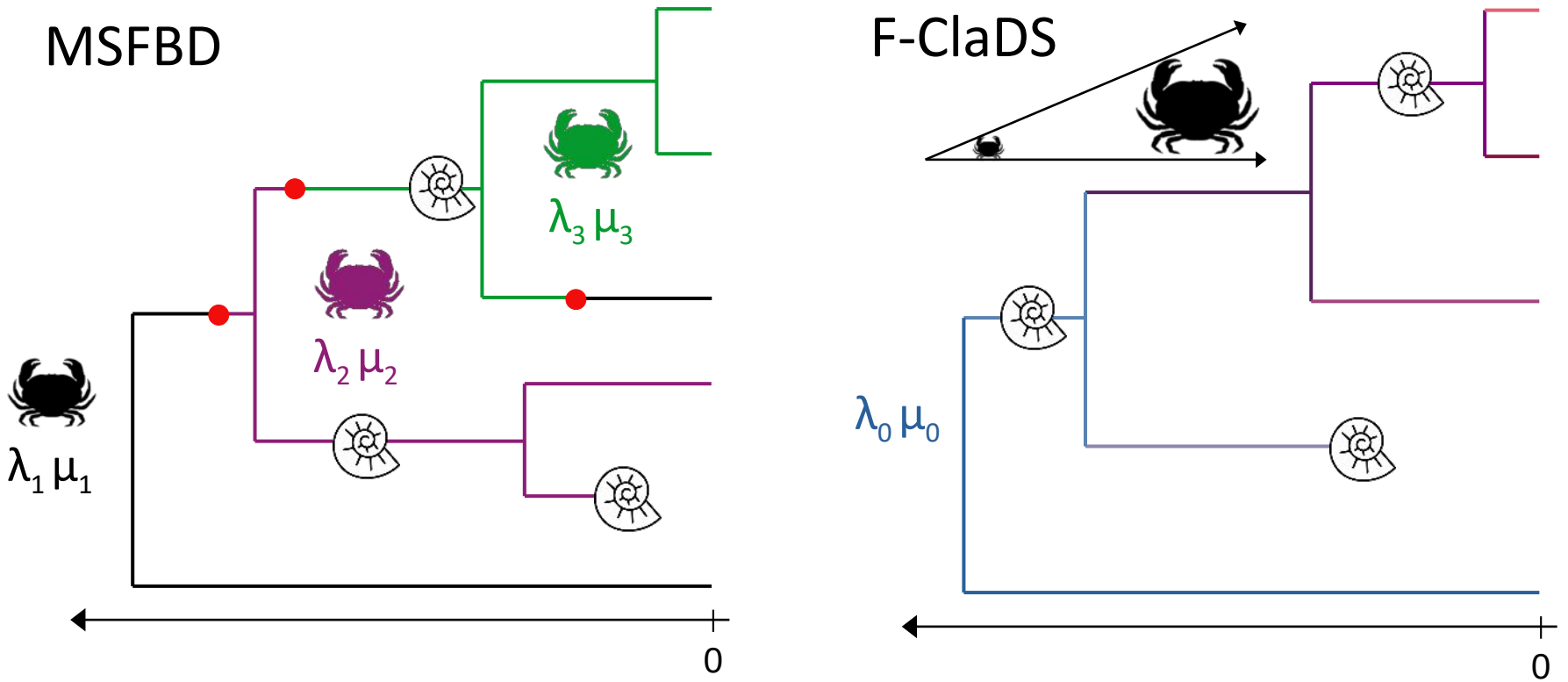
λ_1 = LogNormal (λ_0 x α, σ)

λ_2 = LogNormal (λ_0 x α, σ)

Maliet *et al.* **Nat. Eco. Evo. 2019**
Maliet & Morlon **Sys. Bio. 2021**
Barido-Sottani & Morlon **Sys. Bio. 2023**

- Continuous evolution process driven by a lognormal distribution
- Inheritance / autocorrelation process: descendant rates partially inherited from the ancestral rates
- New estimated parameters:
  - Initial rates at the root λ_0 and μ_0
  - Lognormal parameters α and σ
  - Birth rates for each edge λ_i

Heterogeneous FBD models



Barido-Sottani & Morlon. Sys. Bio. *In revision*, Chabrol *et al.* Sys. Bio. *submitted*

Summary

- FBD phylogenetic inference is a **powerful** and **flexible** tool for integrating fossil samples/species into phylogenies
- FBD phylogenetic inference is an **active** area of development
  - More complete integration of existing data
  - More realistic models of evolutionary processes
- Important challenges remain
  - Computational performance
  - Identifiability of models faced with limited amounts of data
  - Accurate modelling of morphological information