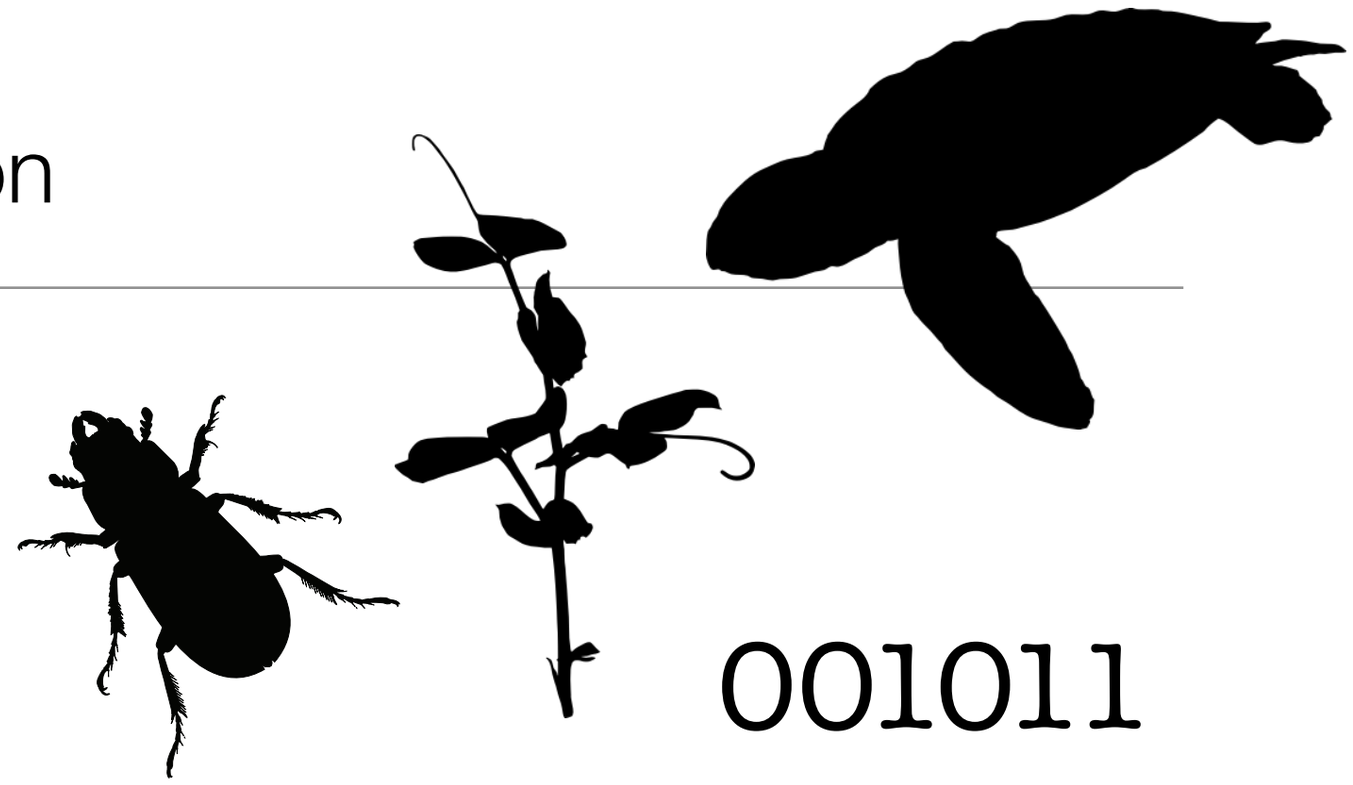
Divergence time estimation

Rachel Warnock



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE

CEVO
Computational Evolution
http://www.bsse.ethz.ch/cevo

Testing hypotheses in evolutionary biology & macroevolution

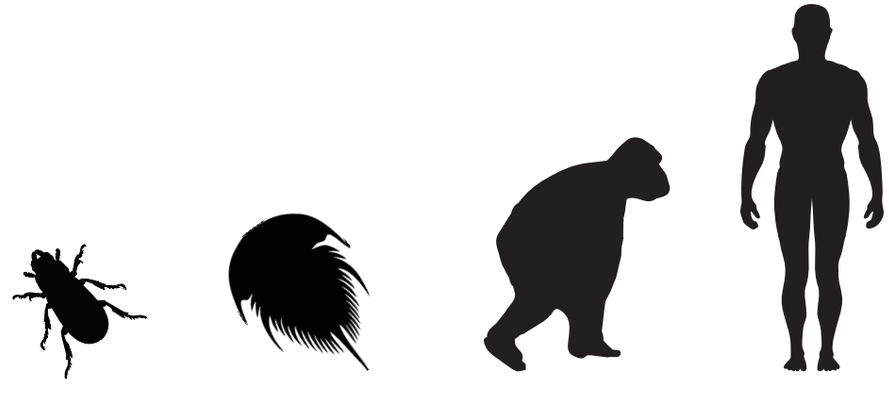
---

t speciation or extinction times

r rates of morphological or molecular evolution

λ rate of speciation

μ rate of extinction

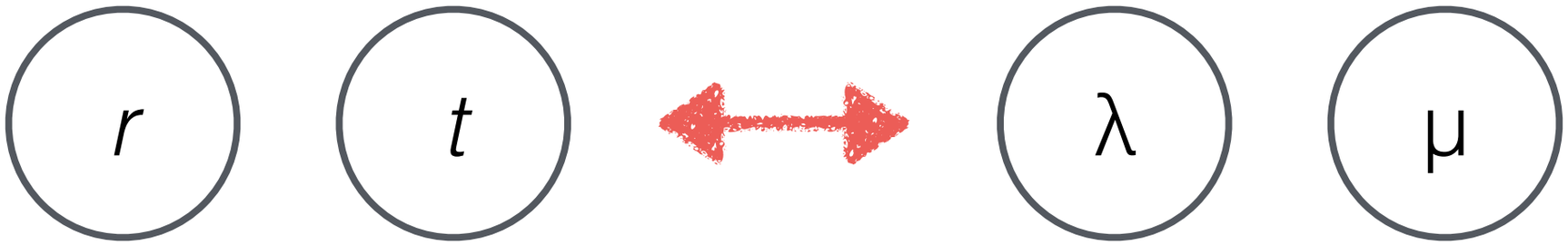


Macroevolutionary parameters of interest are phylogenetic parameters

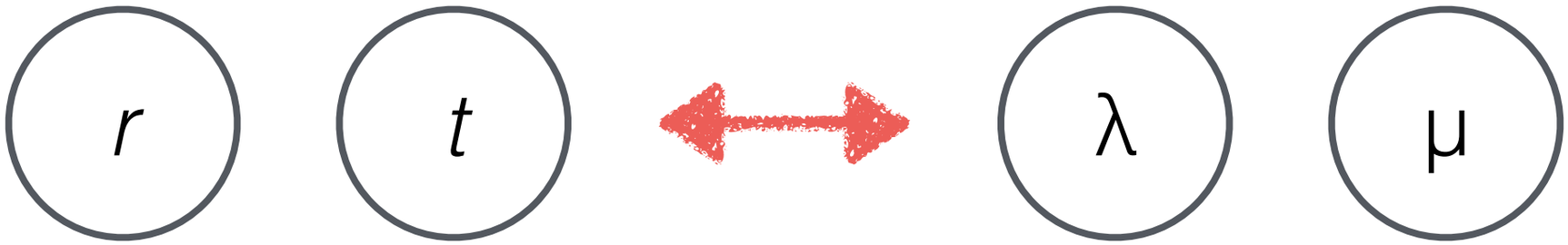
---

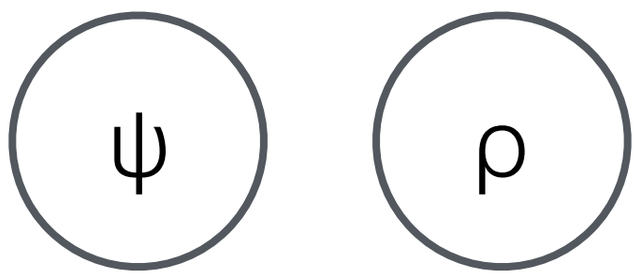
r t λ μ

Macroevolutionary parameters of interest are phylogenetic parameters



Macroevolutionary parameters of interest are phylogenetic parameters

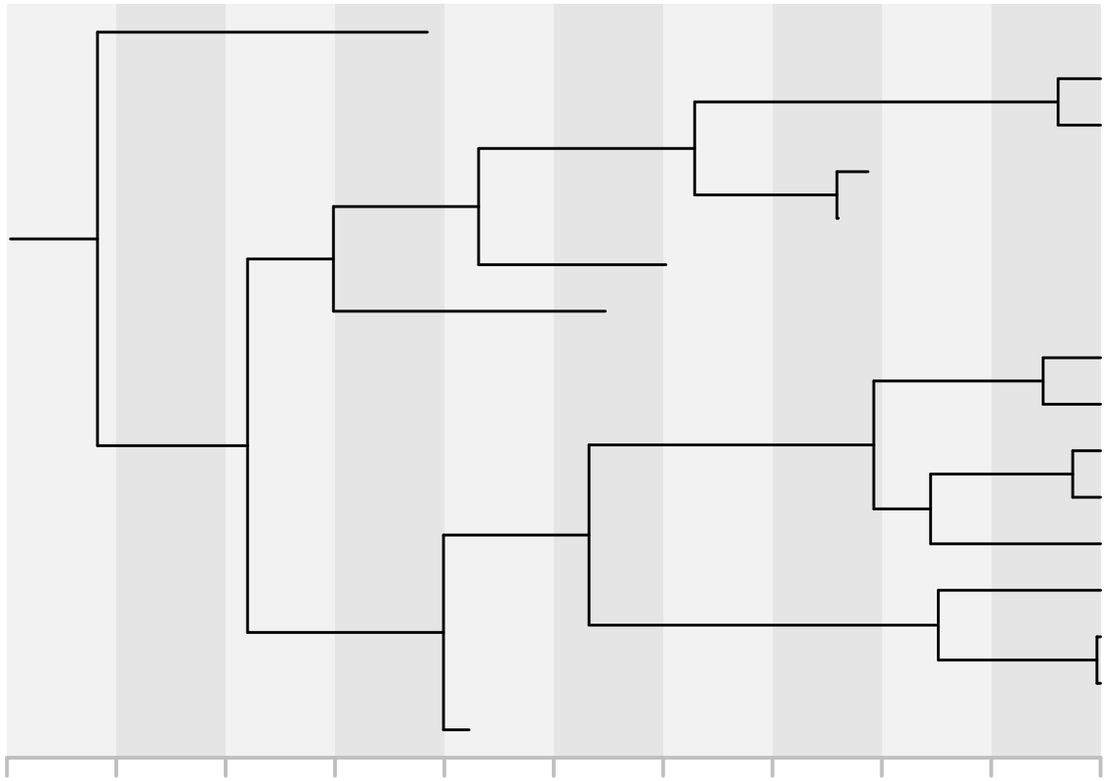




fossil sampling rate

extant species sampling

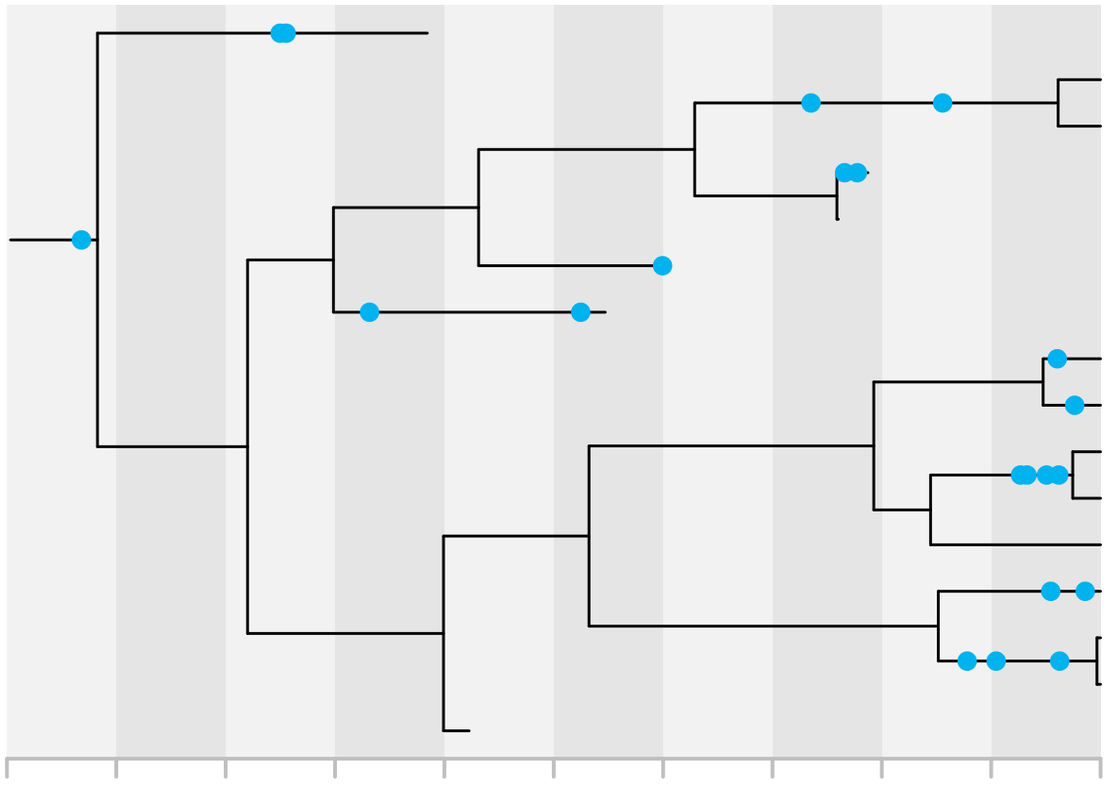
A straightforward model of evolution and sampling



λ — speciation rate

μ — extinction rate

A straightforward model of evolution and sampling

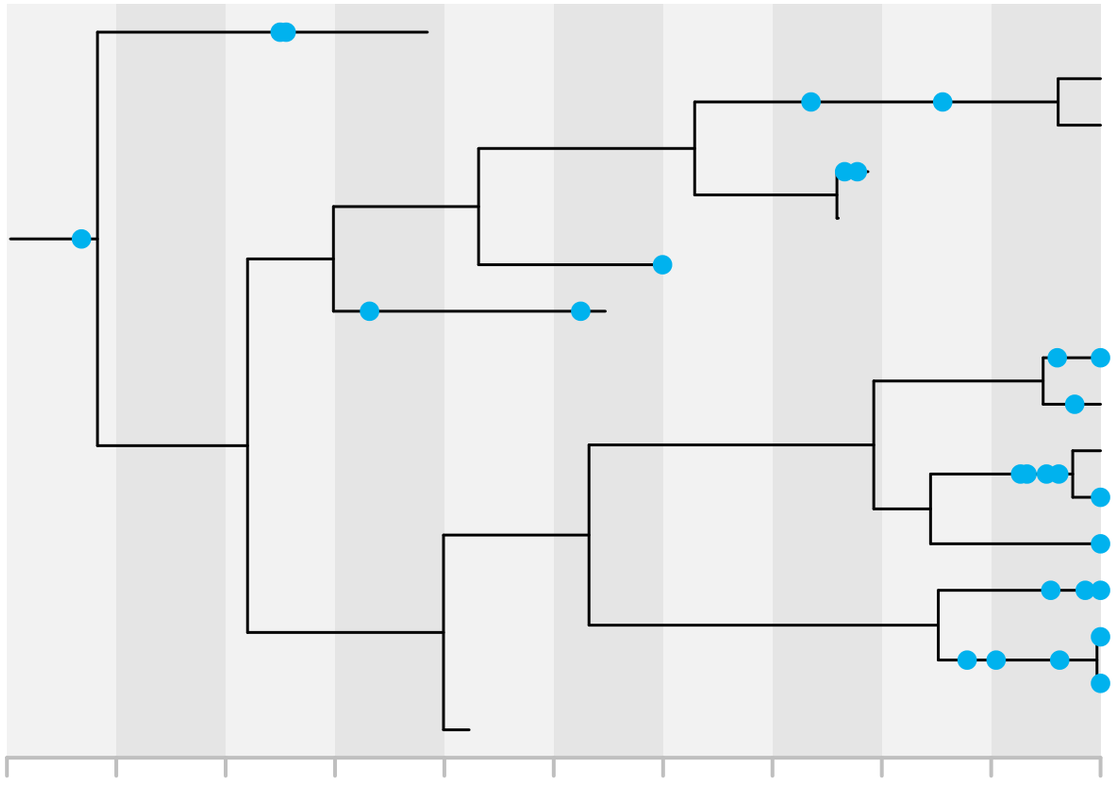


λ — speciation rate

μ — extinction rate

ψ — fossil sampling rate

A straightforward model of evolution and sampling



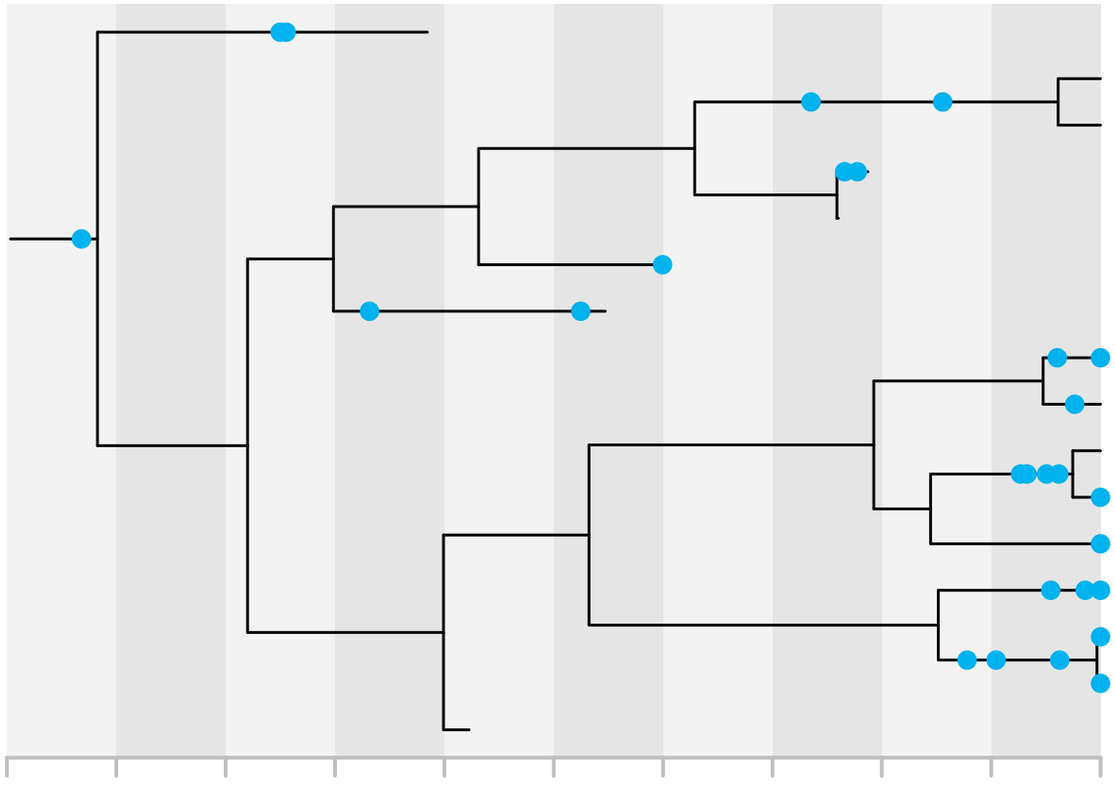
λ — speciation rate

μ — extinction rate

ψ — fossil sampling rate

ρ — extant species sampling

The fossilized birth-death process



λ — speciation rate

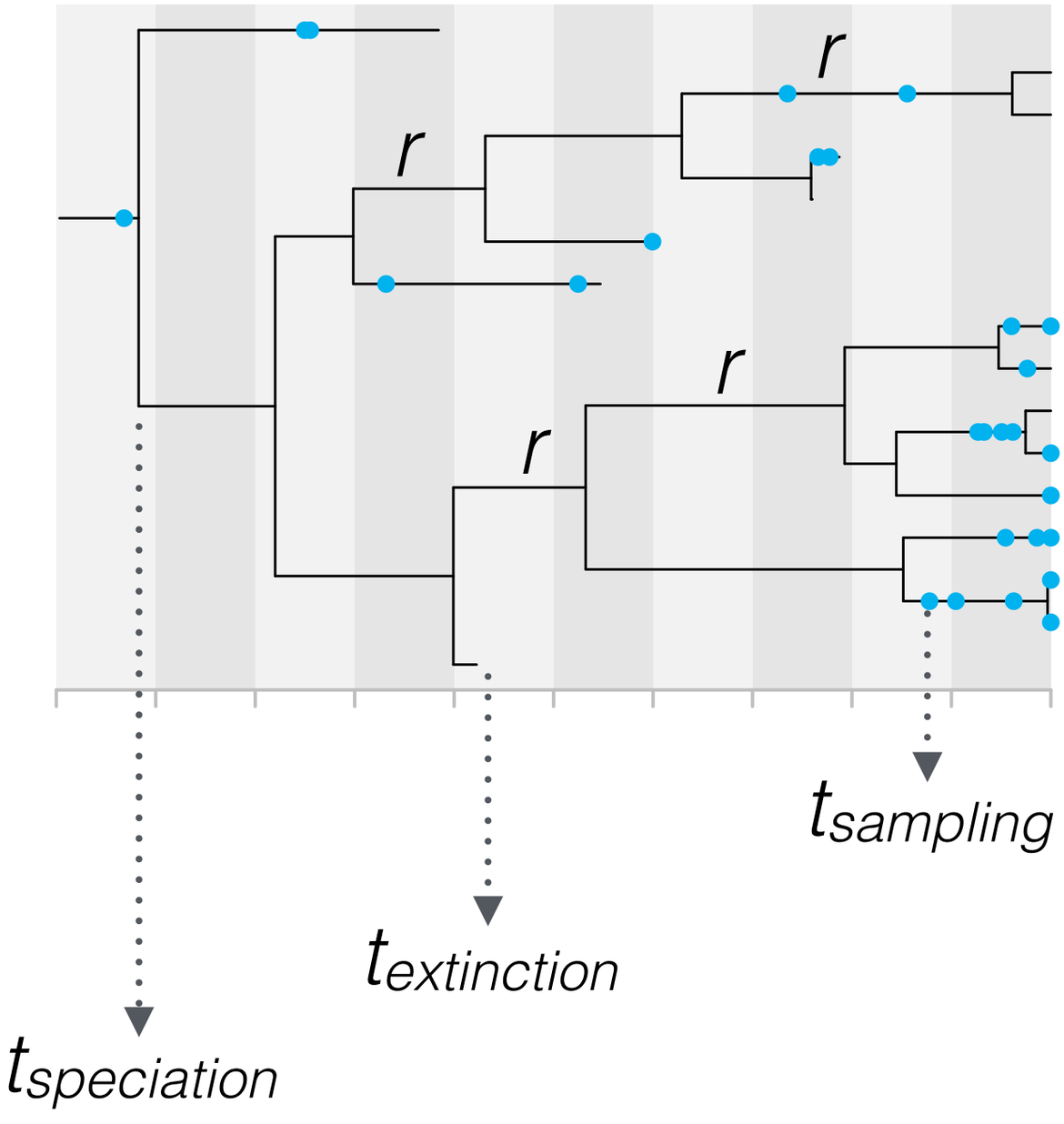
μ — extinction rate

ψ — fossil sampling rate

ρ — extant species sampling

Stadler, 2010

A straightforward model of evolution and sampling



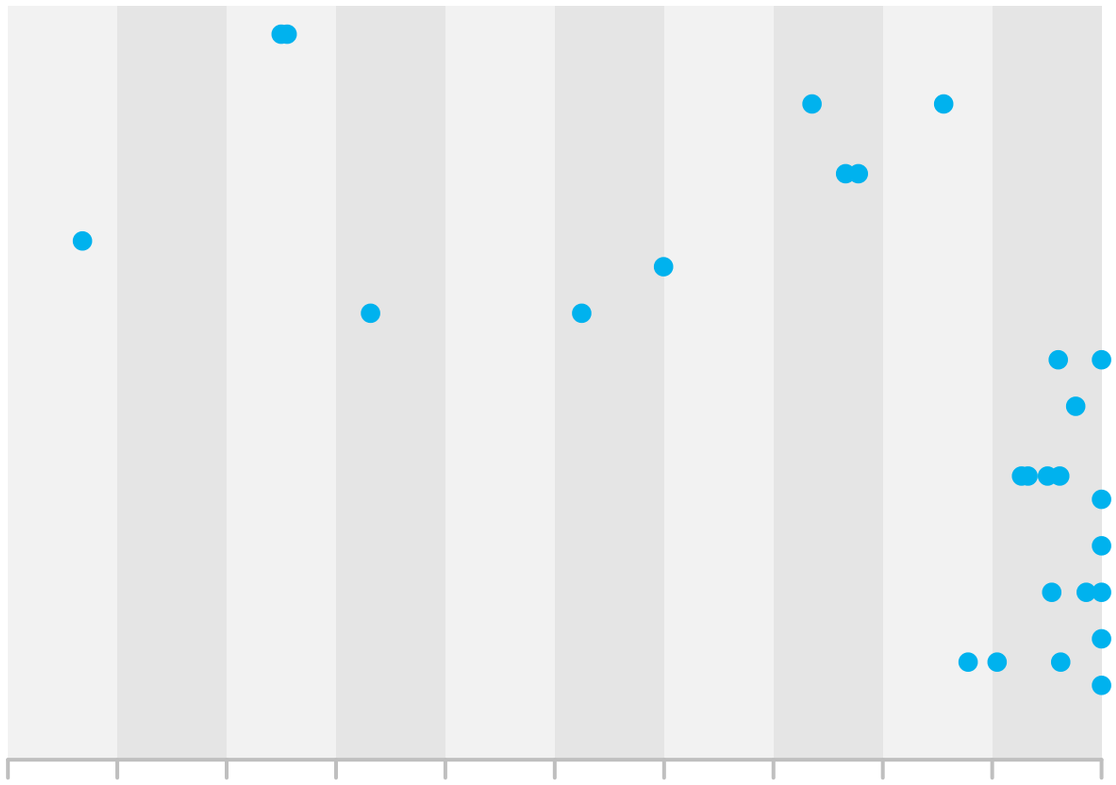
λ — speciation rate

μ — extinction rate

ψ — fossil sampling rate

ρ — extant species sampling

A straightforward model of evolution and sampling



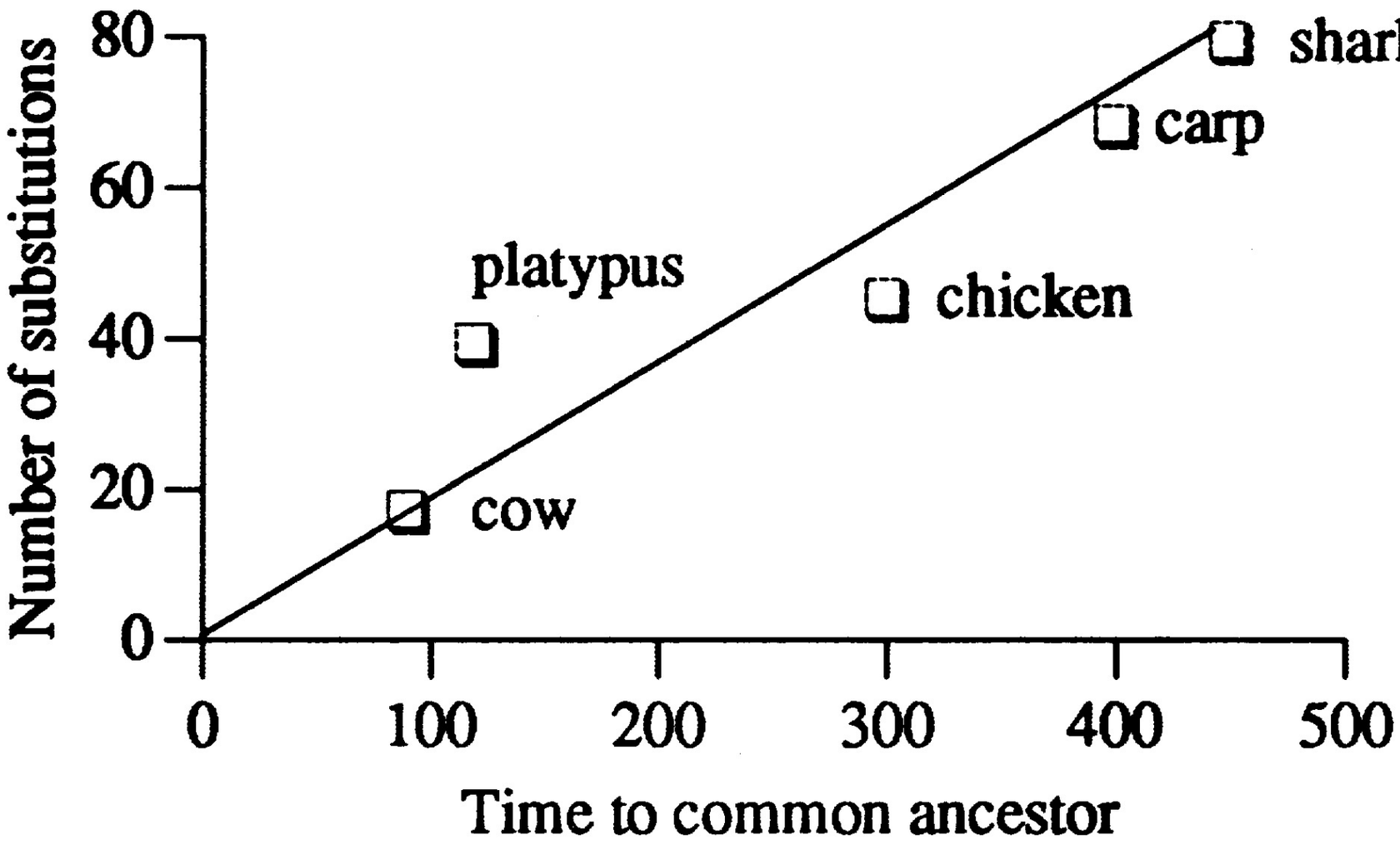
λ — speciation rate

μ — extinction rate

ψ — fossil sampling rate

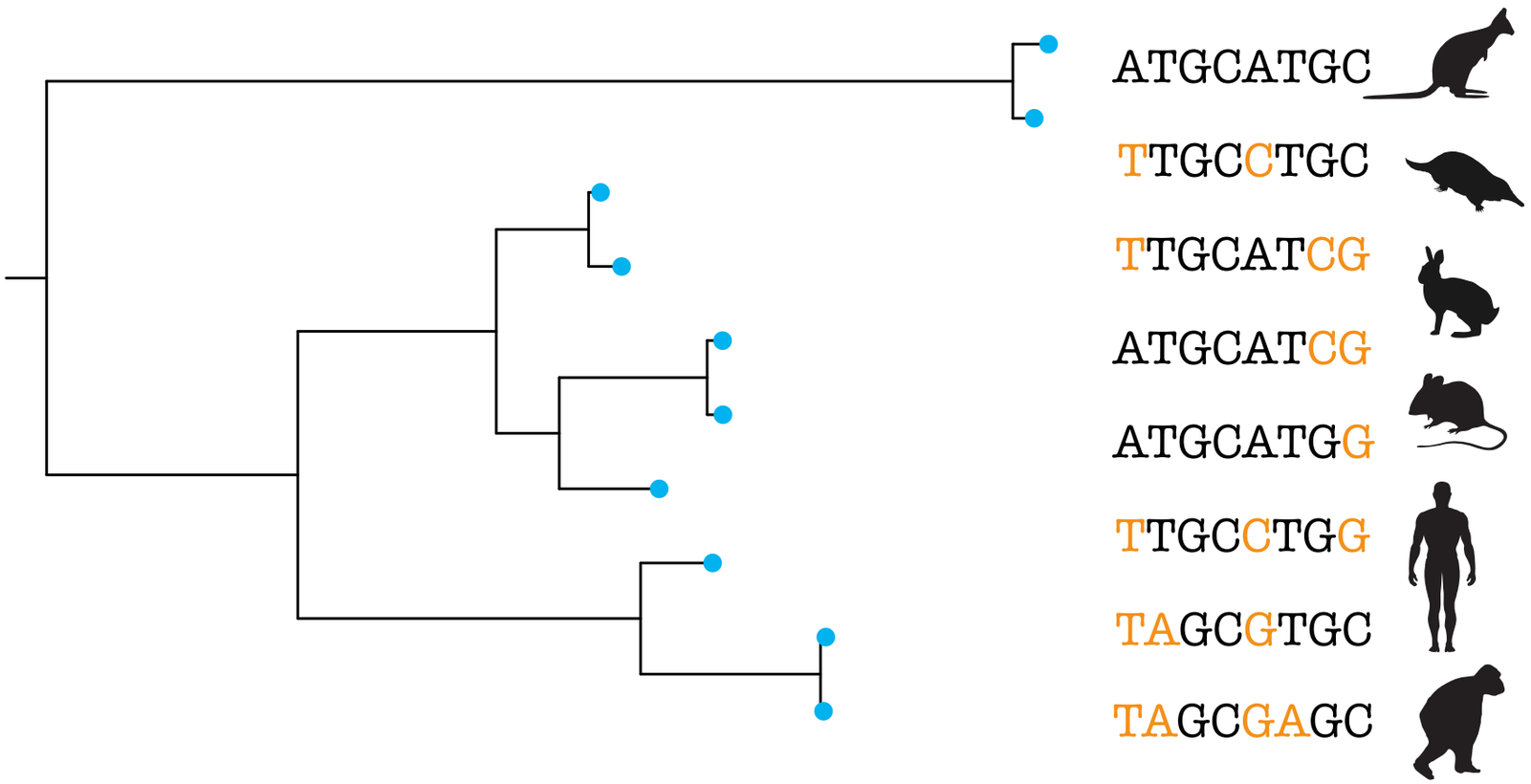
ρ — extant species sampling

The molecular clock hypothesis



Zuckerkandl, Pauling. 1962, 1965

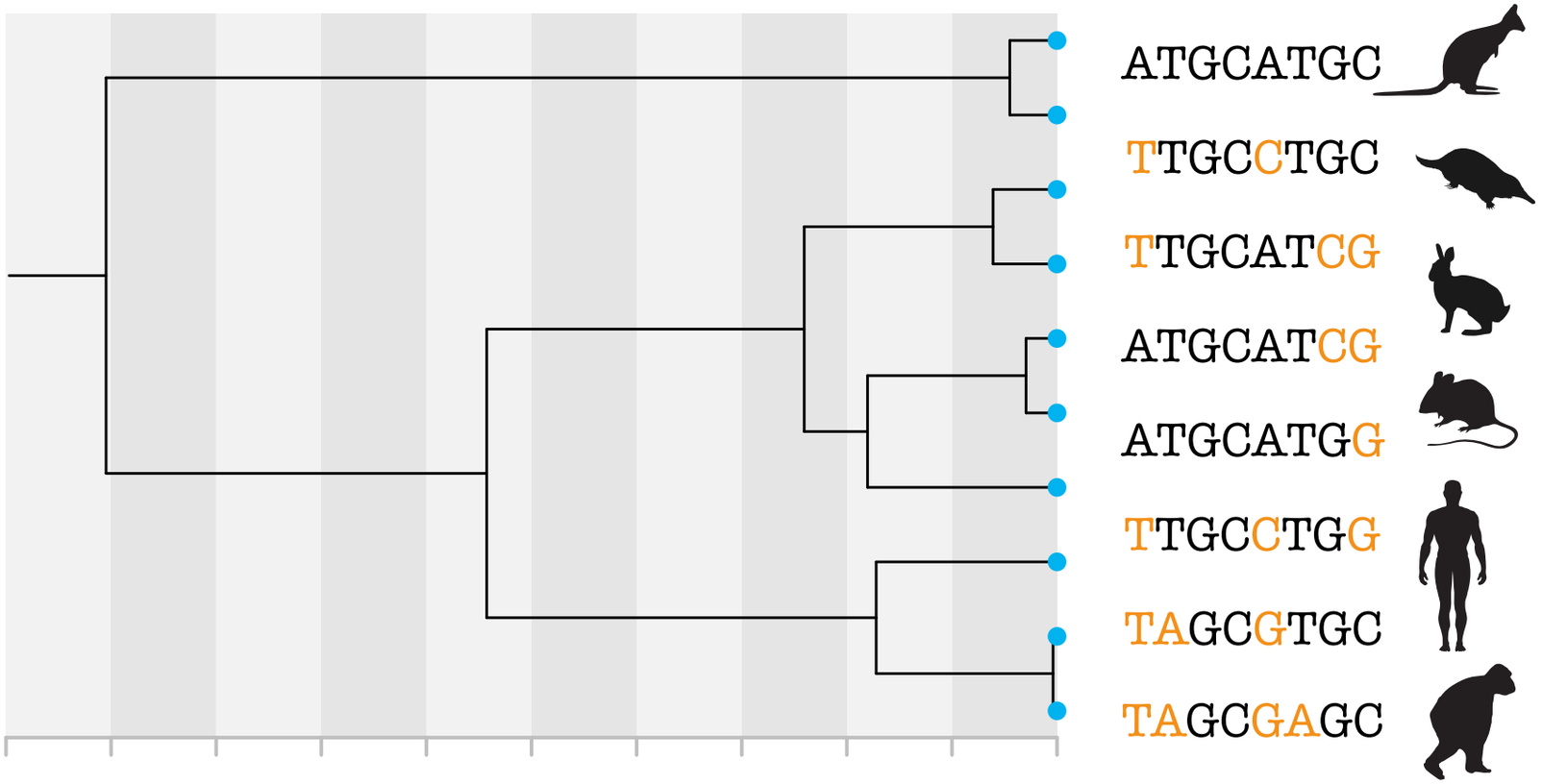
Building the tree of life



branch lengths = rate x time

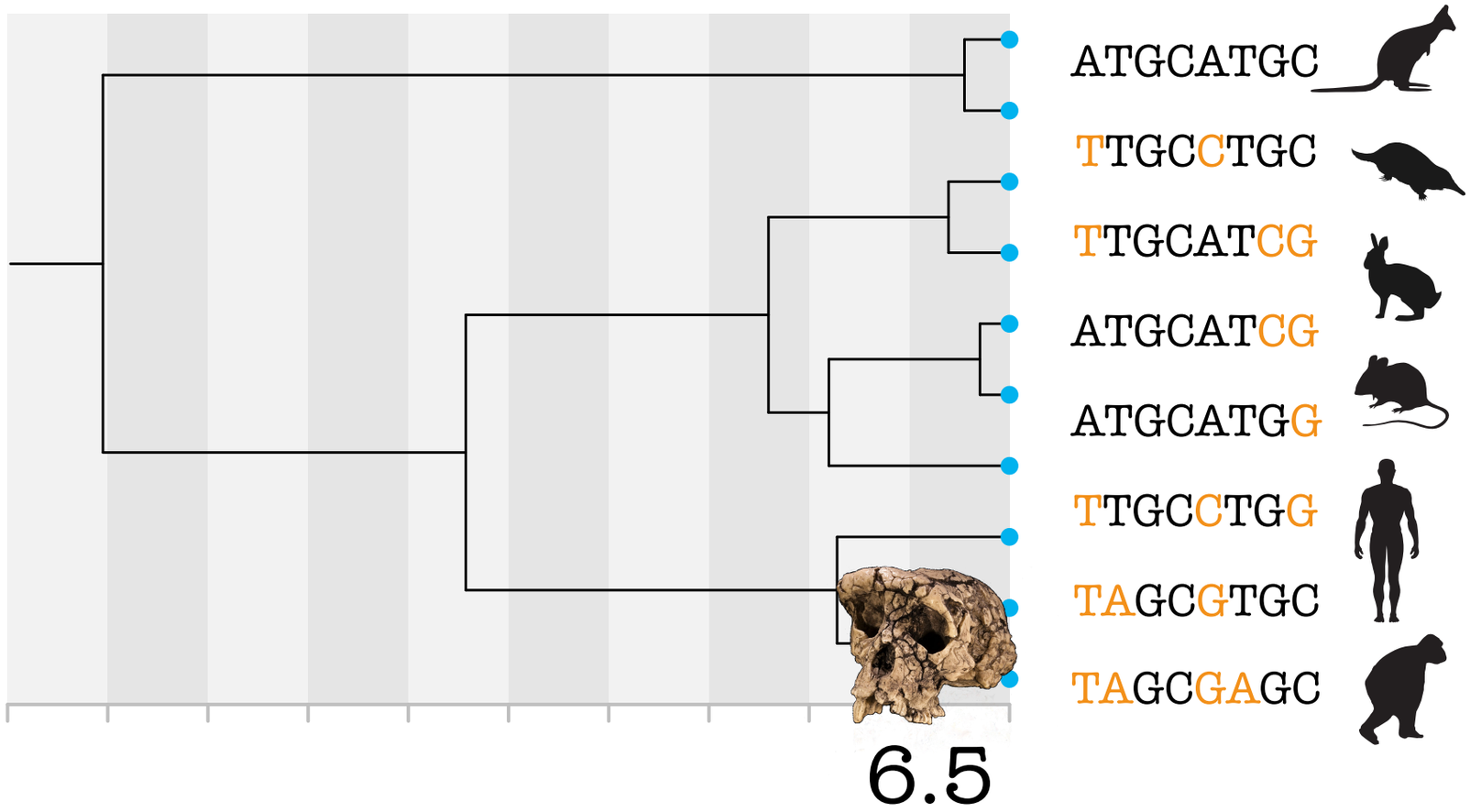
$v = rt$

Dating the tree of life



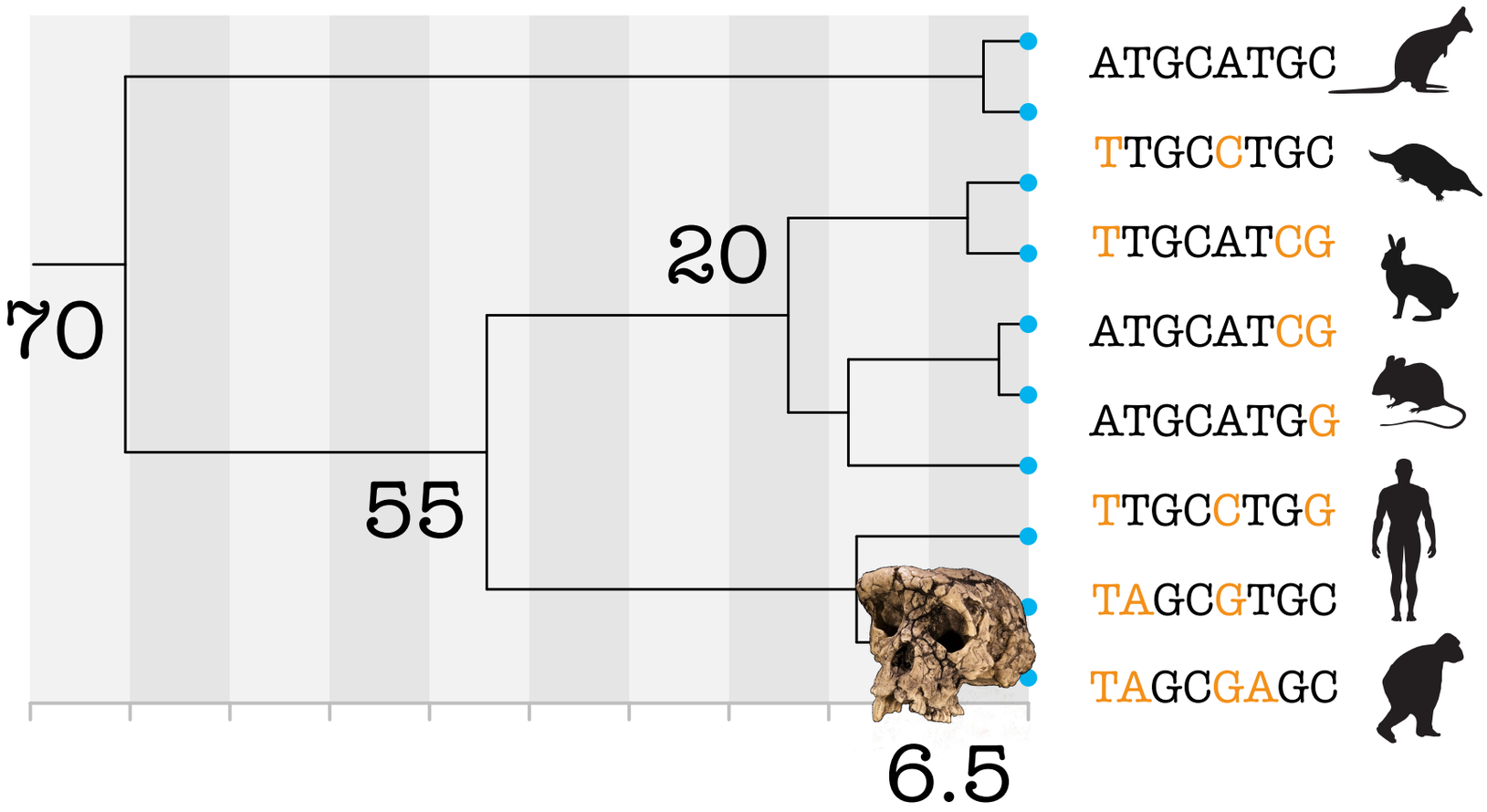
branch lengths = time

Calibrating the molecular clock



branch lengths = time

Calibrating the molecular clock



branch lengths = time

Bayesian divergence time estimation

$$P(\textbf{model} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{model})\, P(\textbf{model})}{P(\textbf{data})}$$

Slides adapted from Louis du Plessis Taming the BEAST Workshop 2016

Bayesian divergence time estimation

likelihood

priors

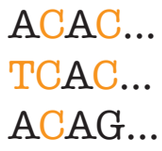
$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\,P(\text{model})}{P(\text{data})}$$

posterior

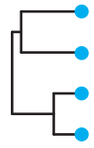
marginal likelihood of the data

Slides adapted from Louis du Plessis Taming the BEAST Workshop 2016

Bayesian divergence time estimation



alignment



tree

tree model

calibration priors

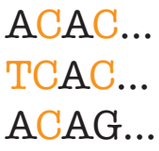
substitution model

clock model

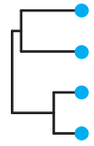
$$P(\textbf{model} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{model})\,P(\textbf{model})}{P(\textbf{data})}$$

Slides adapted from Louis du Plessis Taming the BEAST Workshop 2016

Bayesian divergence time estimation



alignment



tree

tree model

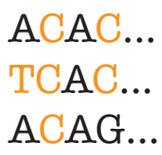
calibration priors

substitution model

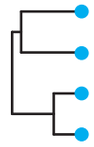
clock model

Slides adapted from Louis du Plessis Taming the BEAST Workshop 2016

Bayesian divergence time estimation



alignment



tree

tree model

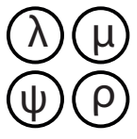
calibration priors

substitution model

clock model

$$P\left(\begin{array}{l}ACAC...\\TCAC...\\ACAG...\end{array}\right) = \frac{P\left(\begin{array}{l}ACAC...\\TCAC...\\ACAG...\end{array}\right) P\left(\right)}{P\left(\begin{array}{l}ACAC...\\TCAC...\\ACAG...\end{array}\right)}$$

Coming up…



Slides adapted from Louis du Plessis Taming the BEAST Workshop 2016

Bayesian divergence time estimation

- Sequence data do not contain information about absolute time

alignment tree tree model calibration priors substitution model clock model

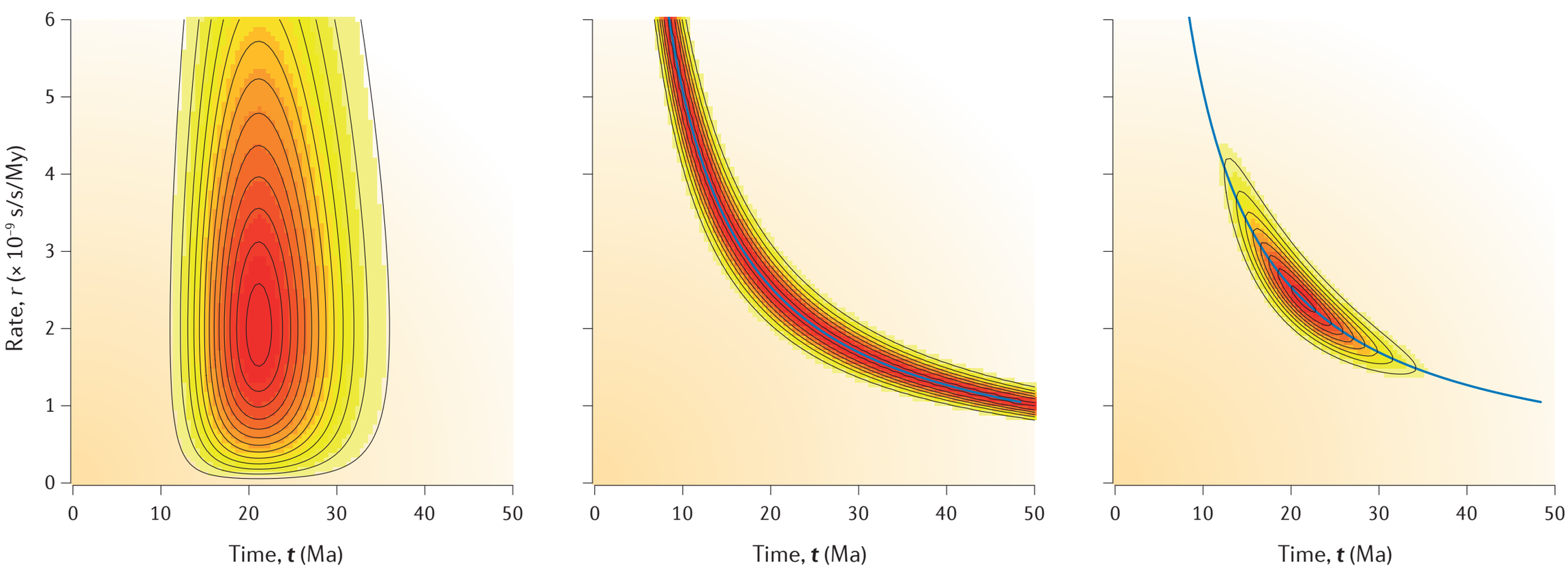
$$P(\text{tree}, \text{model}, \text{clock} \mid \text{ACAC...}) = \frac{P(\text{ACAC...} \mid \text{tree}, \text{model}, \text{clock})\, P(\text{tree}, \text{model}, \text{clock})}{P(\text{ACAC...})}$$

Coming up…

Bayesian divergence time estimation

- Sequence data do not contain information about absolute time
- This has several important consequences:
  - We need strong prior information about the divergence times or the substitution rate
  - Model selection cannot be used to select among possible calibration strategies
  - If there is uncertainty in the calibrations, even an infinite amount of sequence data won't completely eliminate uncertainty in the posteriors

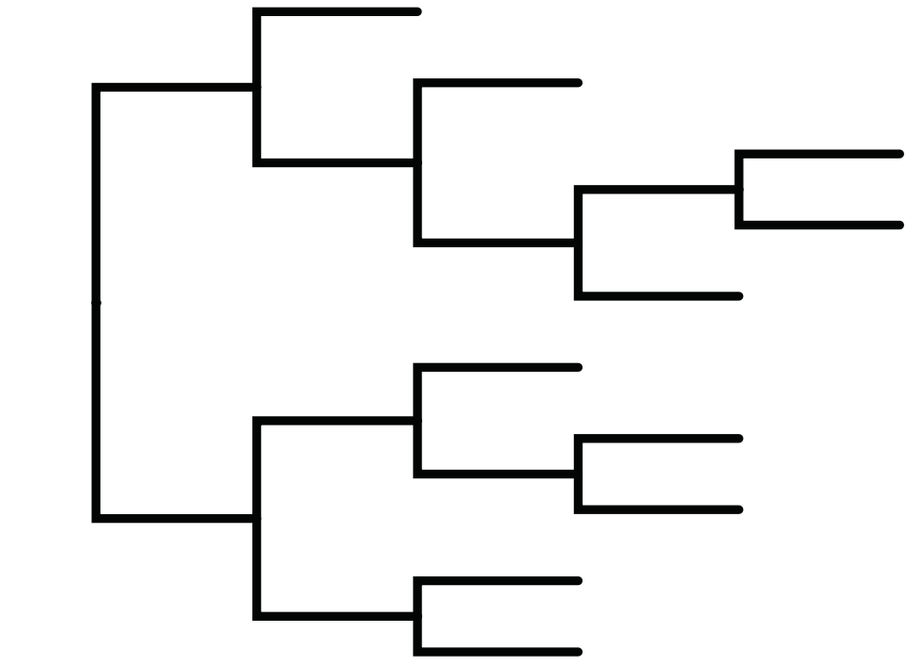
Rate and time are only semi-identifiable



dos Reis et al. 2016. Nature Genetics Reviews

The global molecular clock model

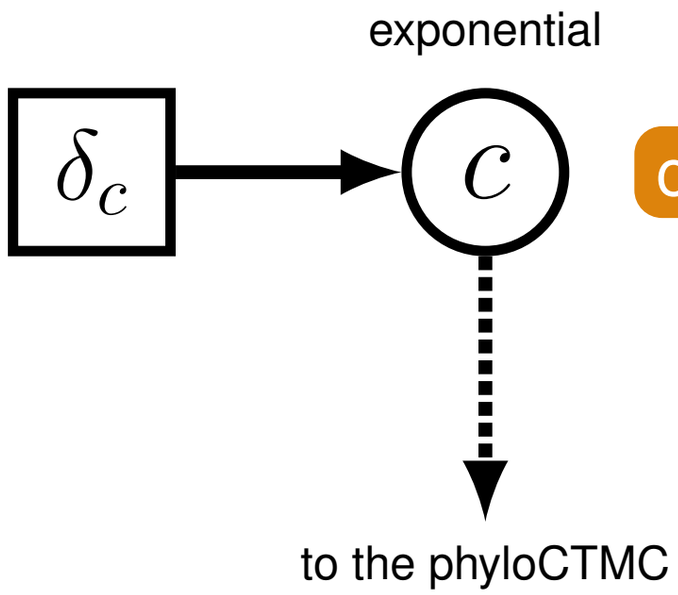
- The substitution rate is constant over time
- All lineages share the same rate

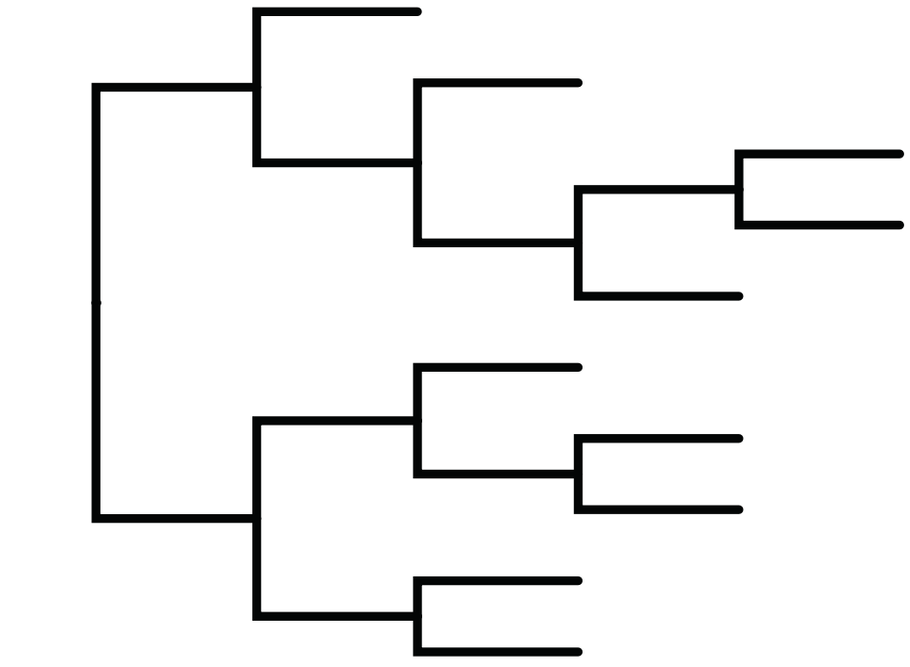


branch length = substitution rate

low high

The global molecular clock model

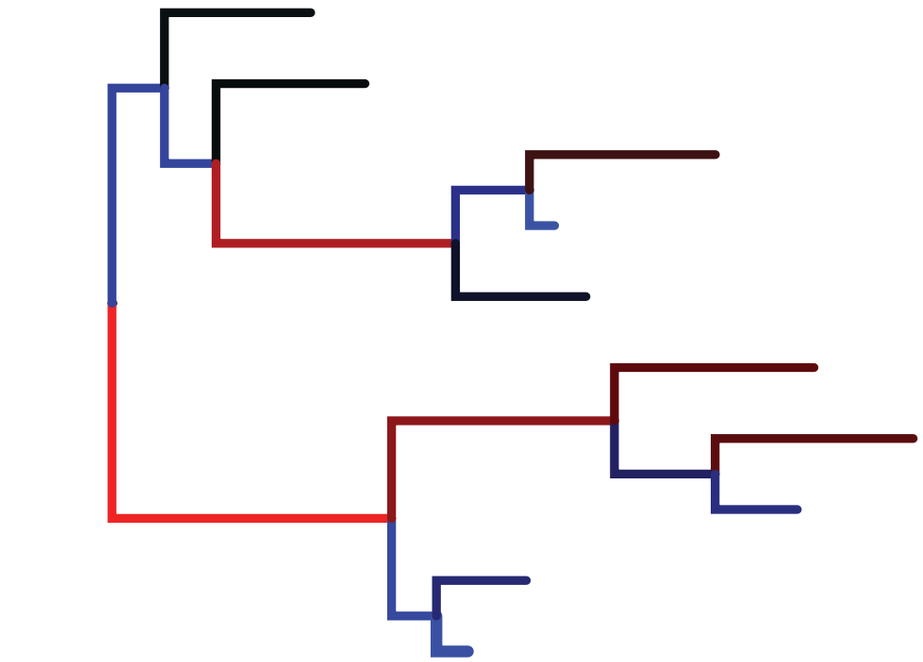




The independent uncorrelated rates model

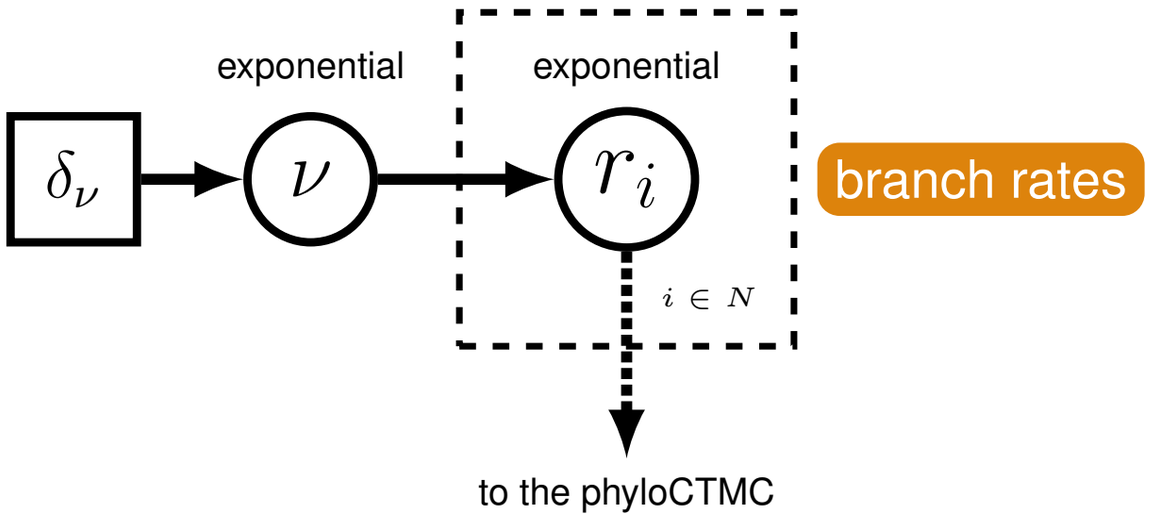
- Lineage-specific rates are uncorrelated

- The rate assigned to each branch is drawn independently from an underlying distribution

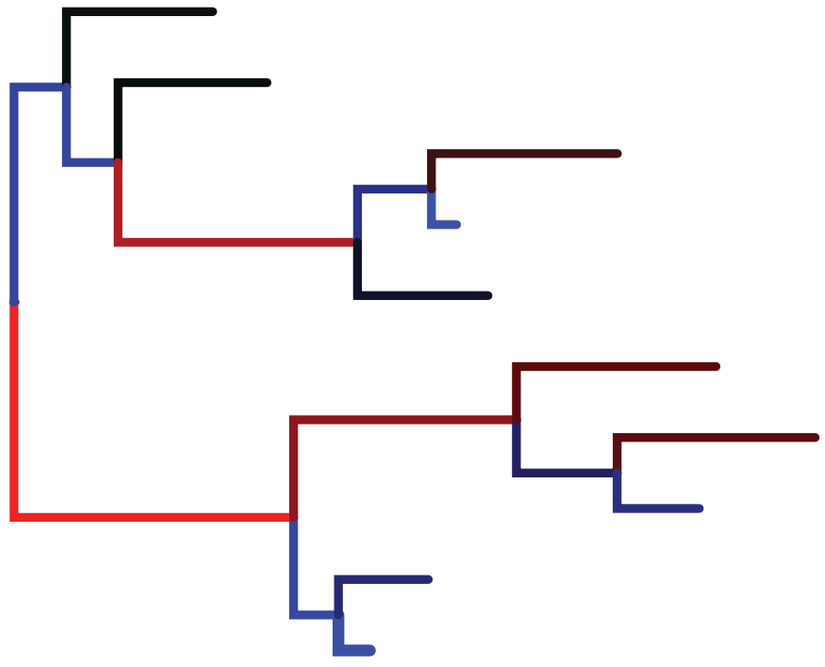


Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007

The independent uncorrelated rates model



branch rates



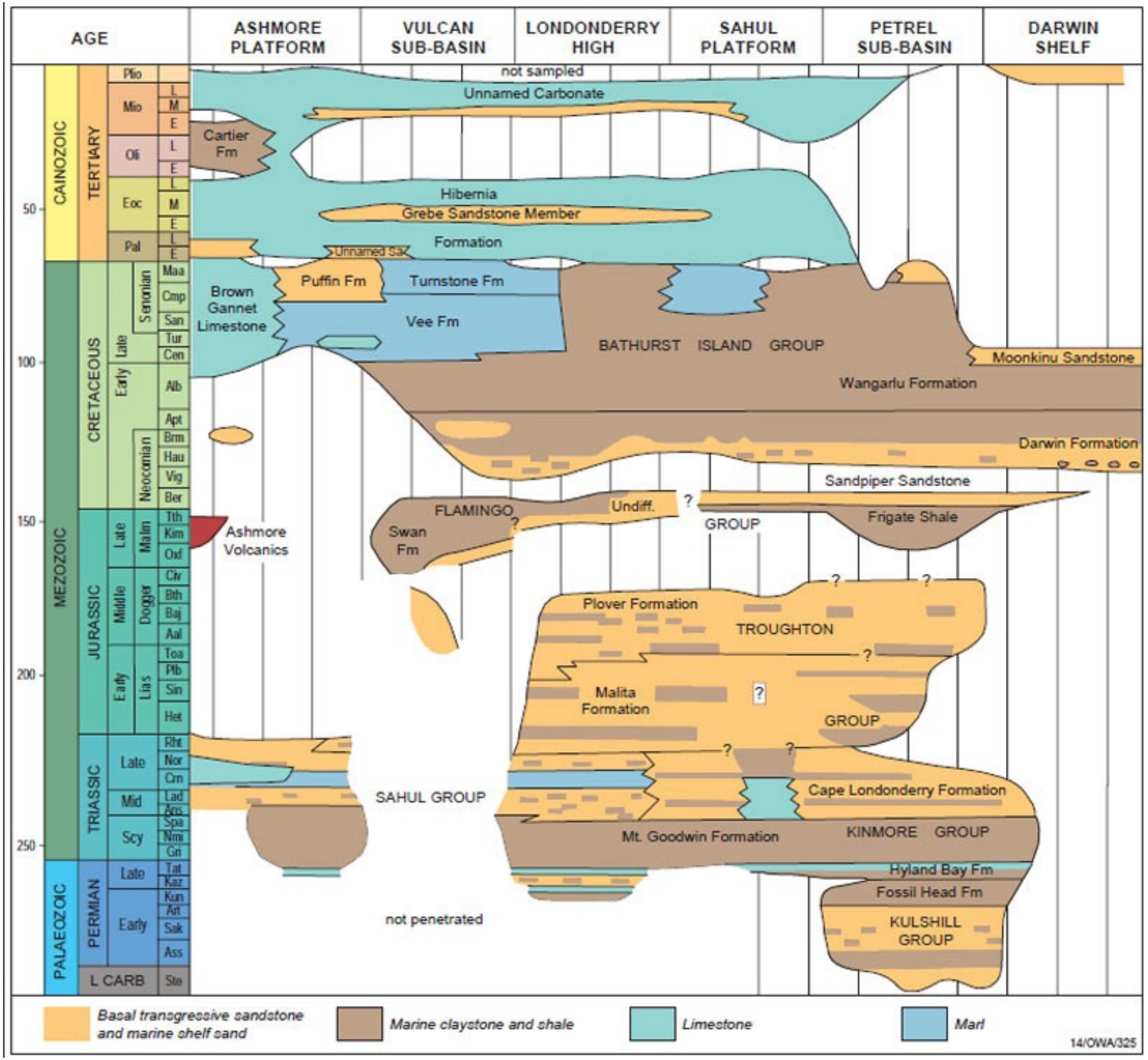
branch length = substitution rate

low high

Many molecular clock models

- Global clock (Zuckerkandl & Pauling, 1962)
- Uncorrelated/independent rates models (Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)
- Log-normally distributed autocorrelated rates (Thorne, Kishino & Painter 1998; Kishino, Thorne & Bruno 2001; Thorne & Kishino 2002)
- Local clocks (Hasegawa, Kishino & Yano 1989; Kishino & Hasegawa 1990; Yoder & Yang 2000; Yang & Yoder 2003, Drummond and Suchard 2010)
- Mixture models on branch rates (Heath, Holder, Huelsenbeck 2012)
- Punctuated rate change model (Huelsenbeck, Larget and Swoford 2000)

Vagaries of the rock and fossil records



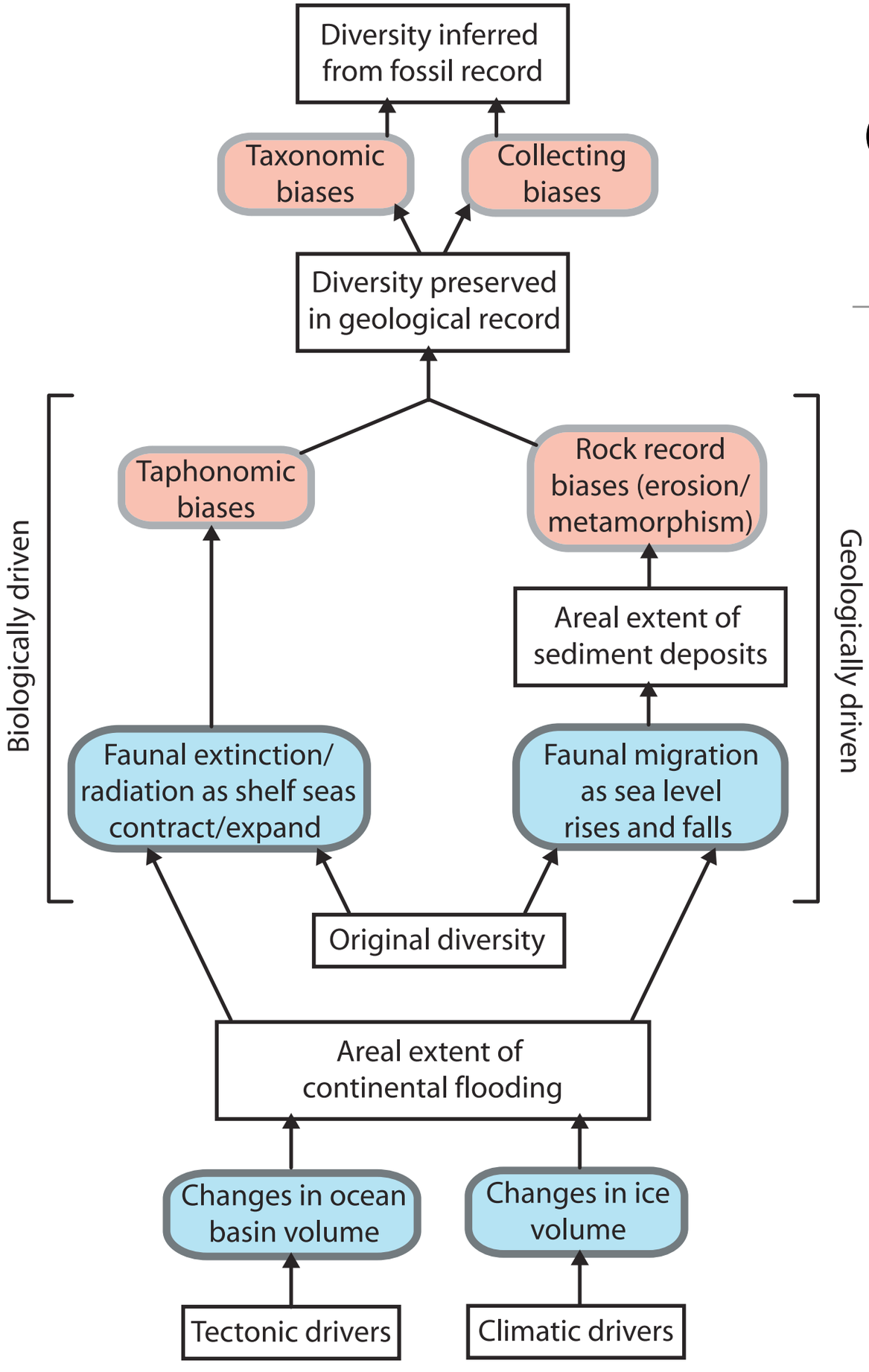
Bonaparte Basin, AU

Vagaries of the rock and fossil records

- The fossil record is extremely incomplete & highly uneven
- Preservation gets (approximately) worse as you go further back in time
- % of original habitat area that survives today at outcrop:
  - Marine shelf habitats 2%
  - Terrestrial habitats 0.5% (of which Cretaceous = 20% and Neogene =30% of the total)

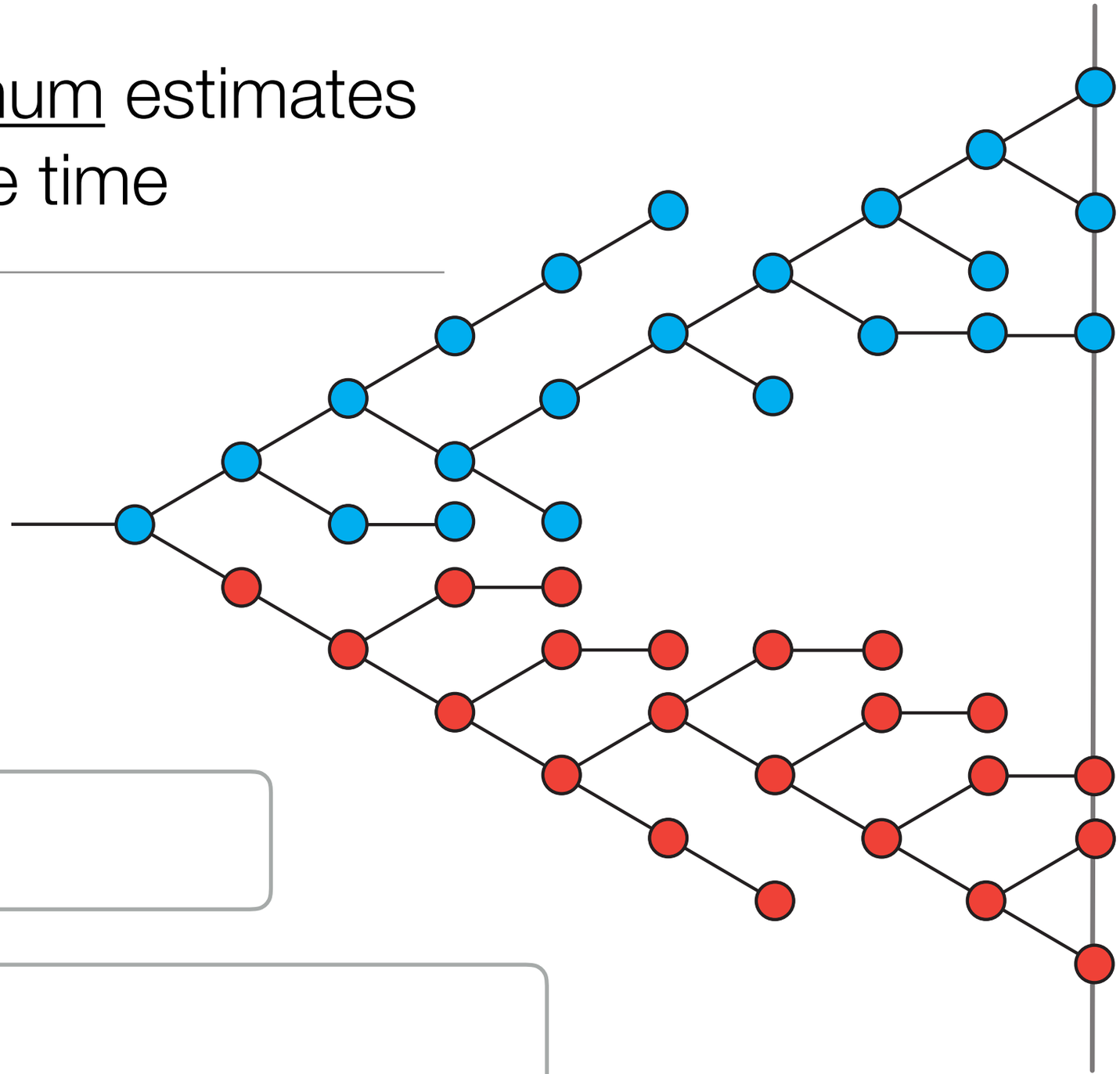
Wall et al. 2011. Geological Society Special Publication

Controls on the probability of preservation



Modified from Smith, 2007. J. of the Geological Society

Fossils provide minimum estimates of species divergence time



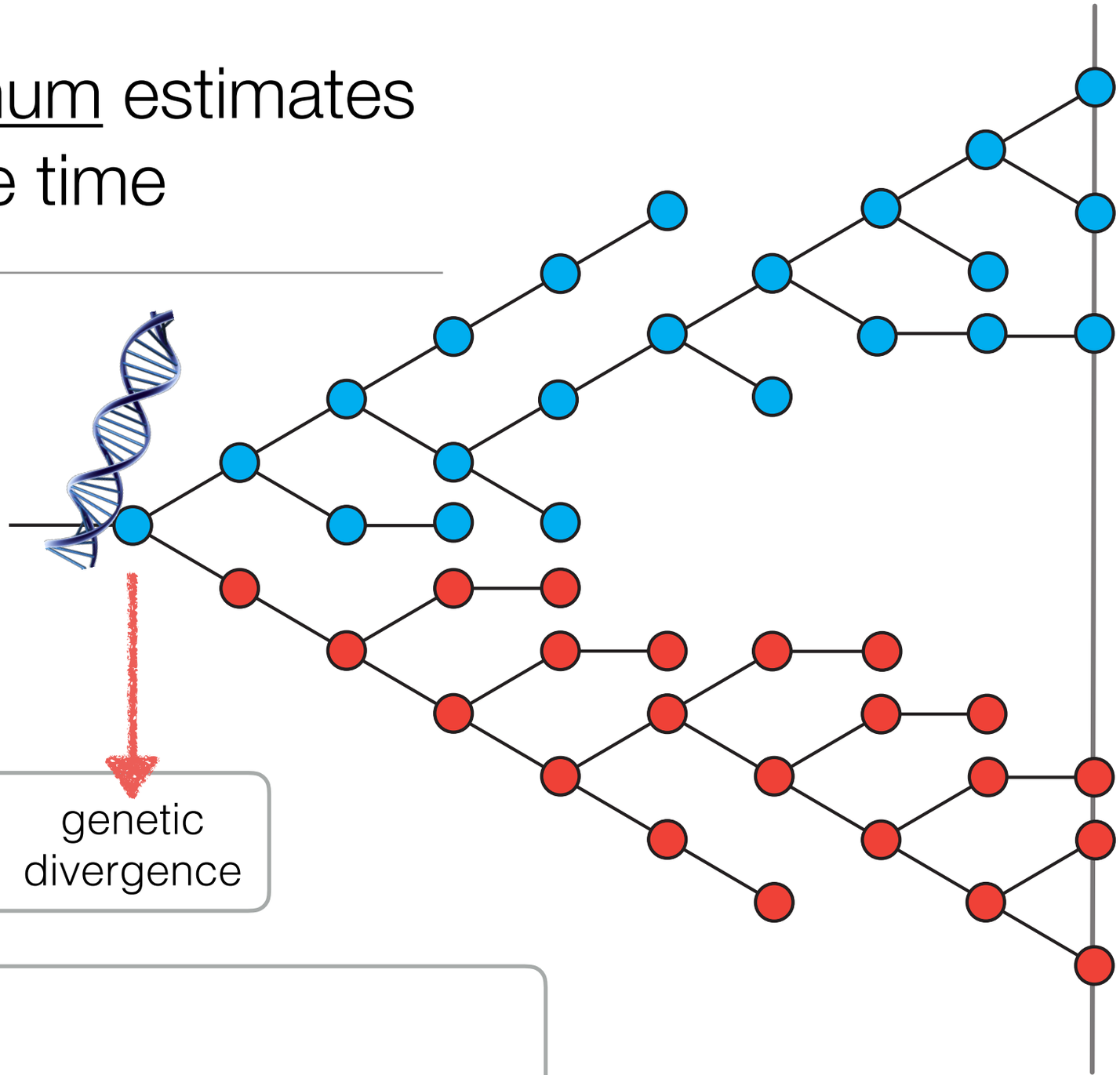
Molecular evolution:

Morphological evolution:

Fossil preservation:

Time

Fossils provide <u>minimum</u> estimates of species divergence time



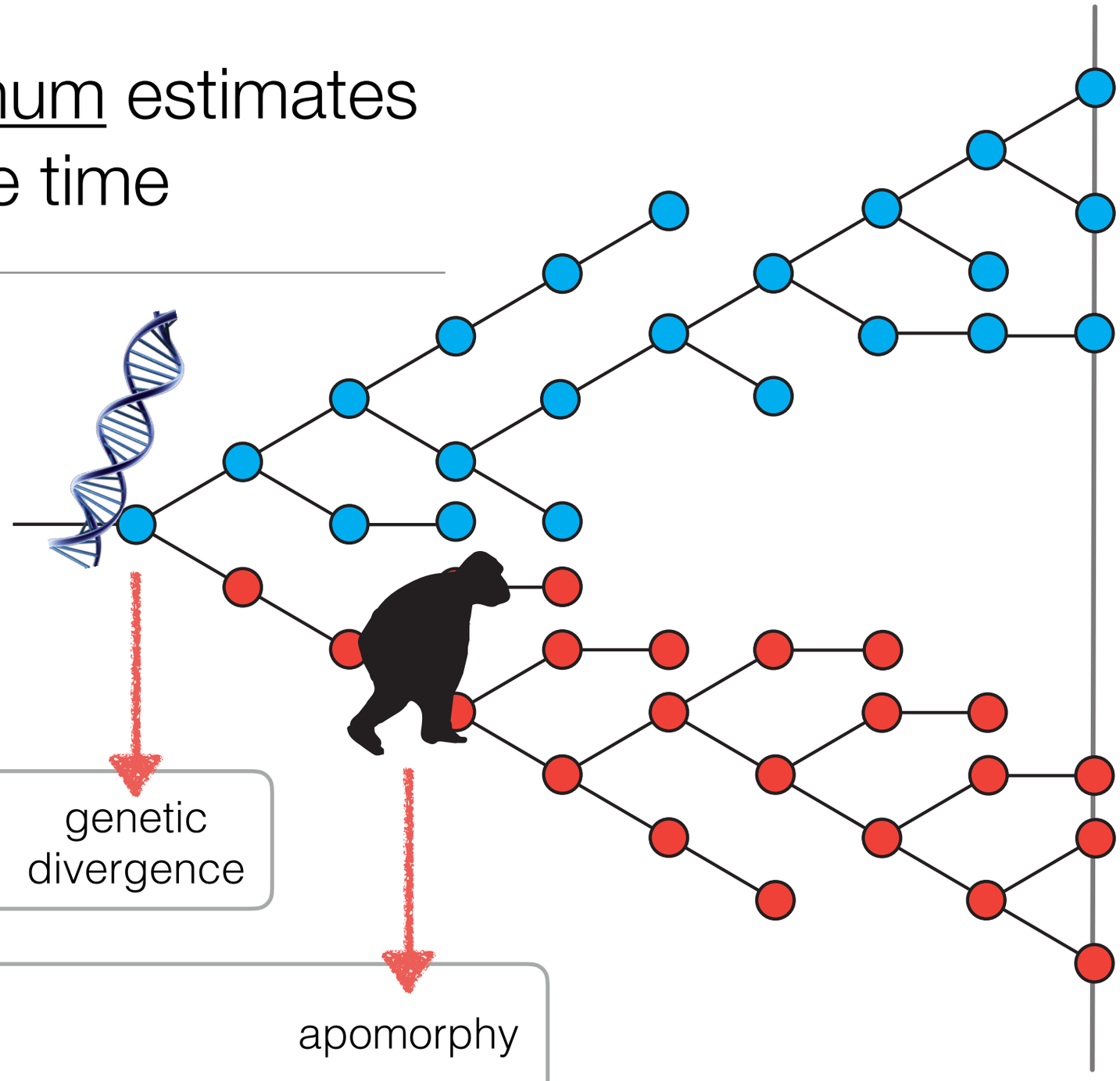
Molecular evolution: genetic divergence

Morphological evolution:

Fossil preservation:

Time

Fossils provide minimum estimates of species divergence time



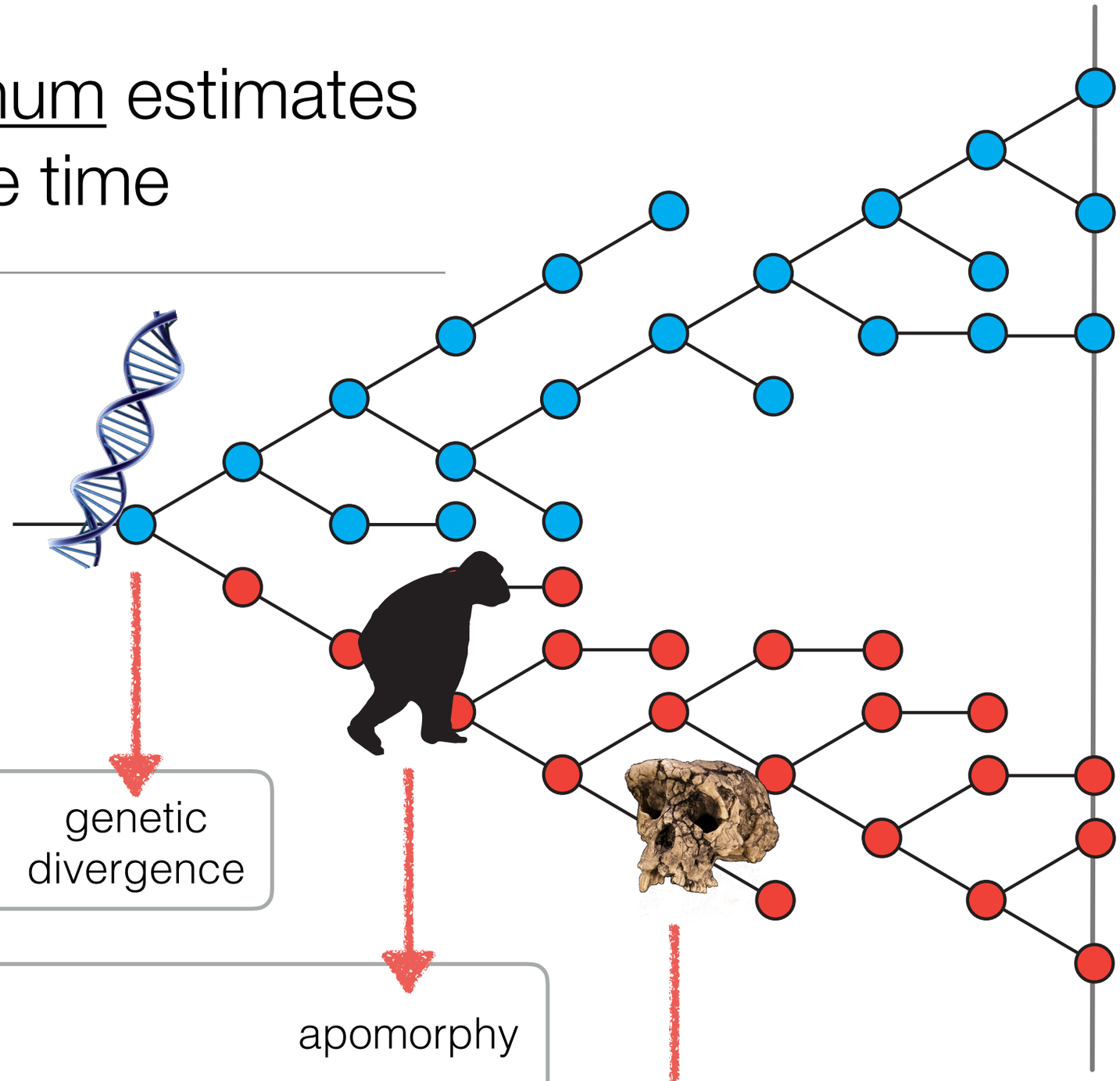
Molecular evolution: genetic divergence

Morphological evolution: apomorphy

Fossil preservation:

Time

Fossils provide minimum estimates of species divergence time



Molecular evolution: genetic divergence

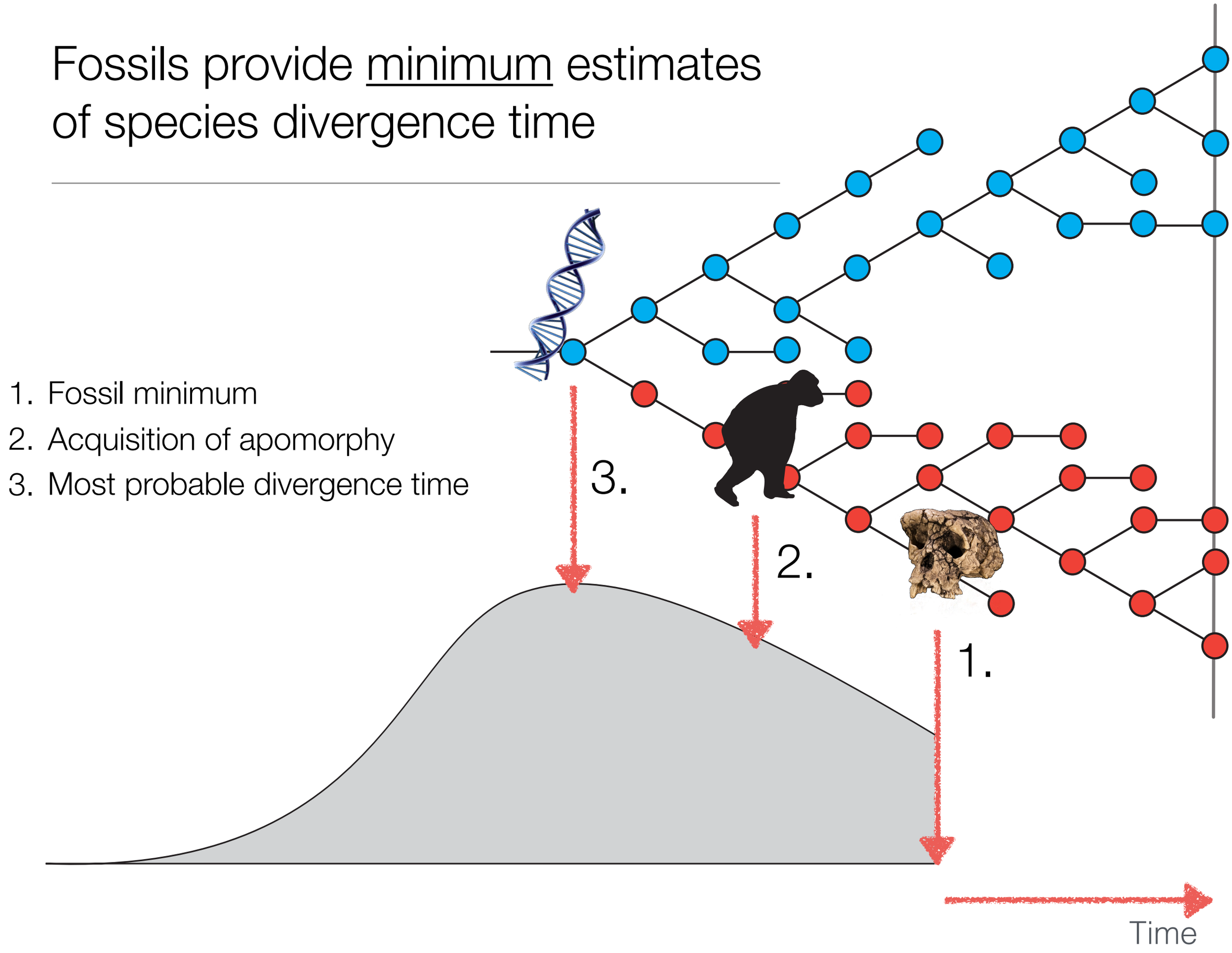
Morphological evolution: apomorphy

Fossil preservation: earliest fossil

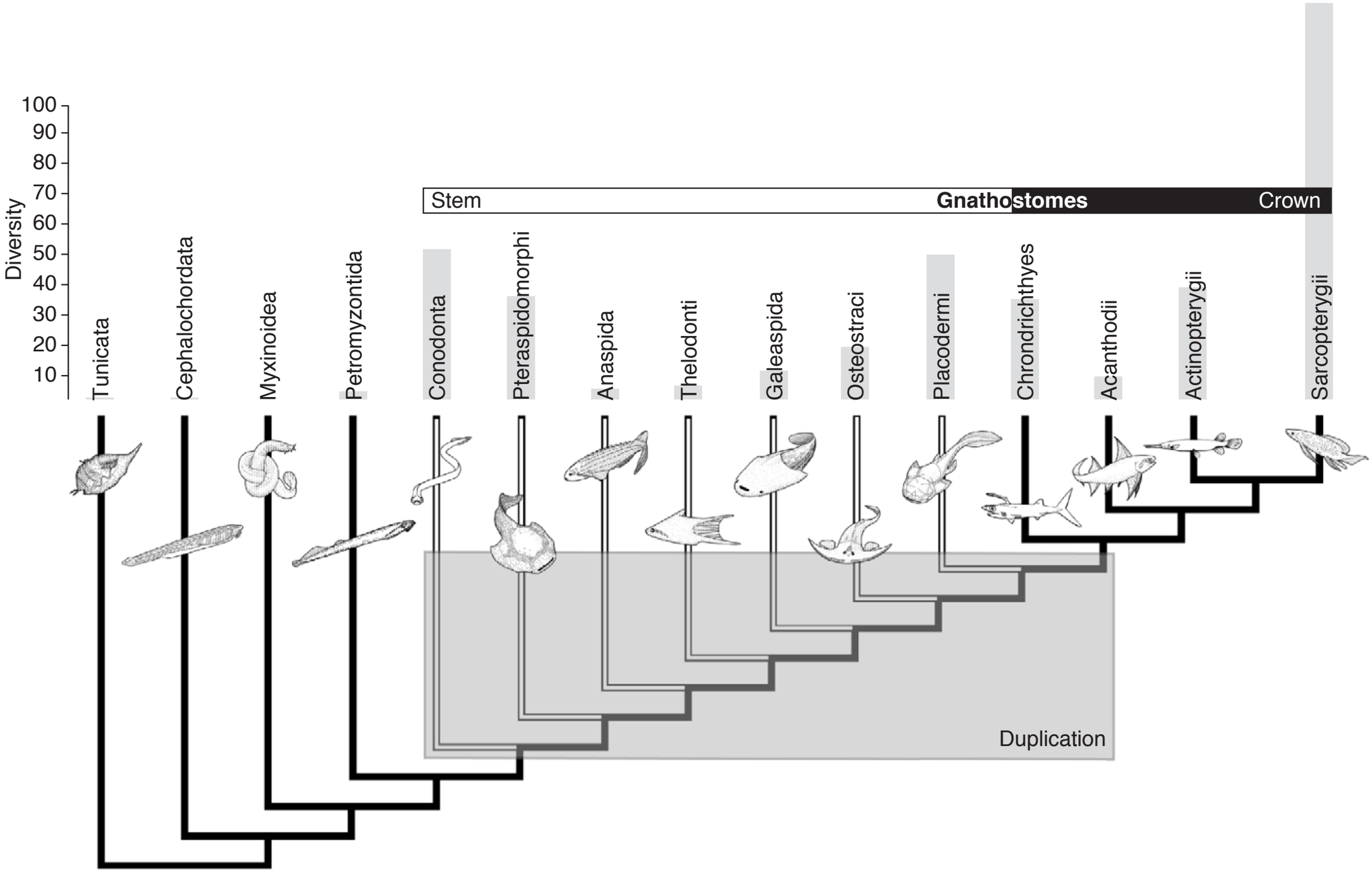
Time

Fossils provide <u>minimum</u> estimates of species divergence time

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



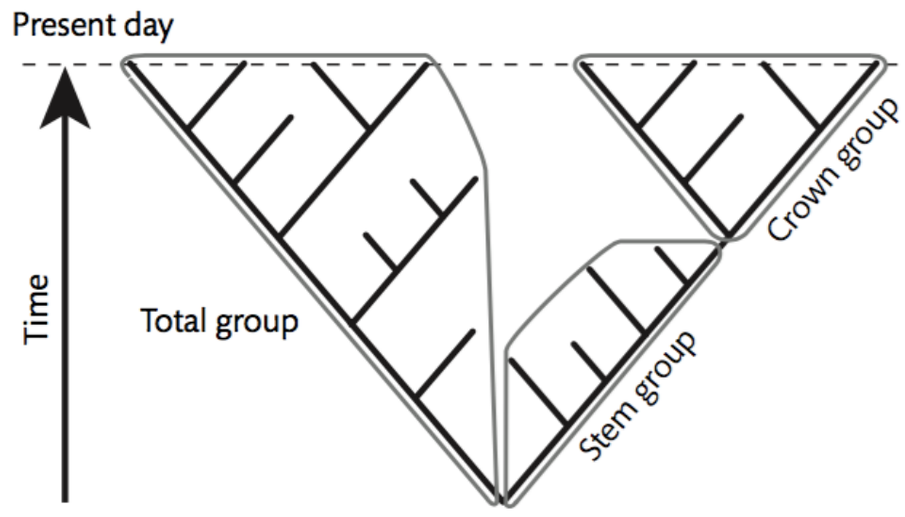
Taxonomic uncertainty: crown versus stem groups



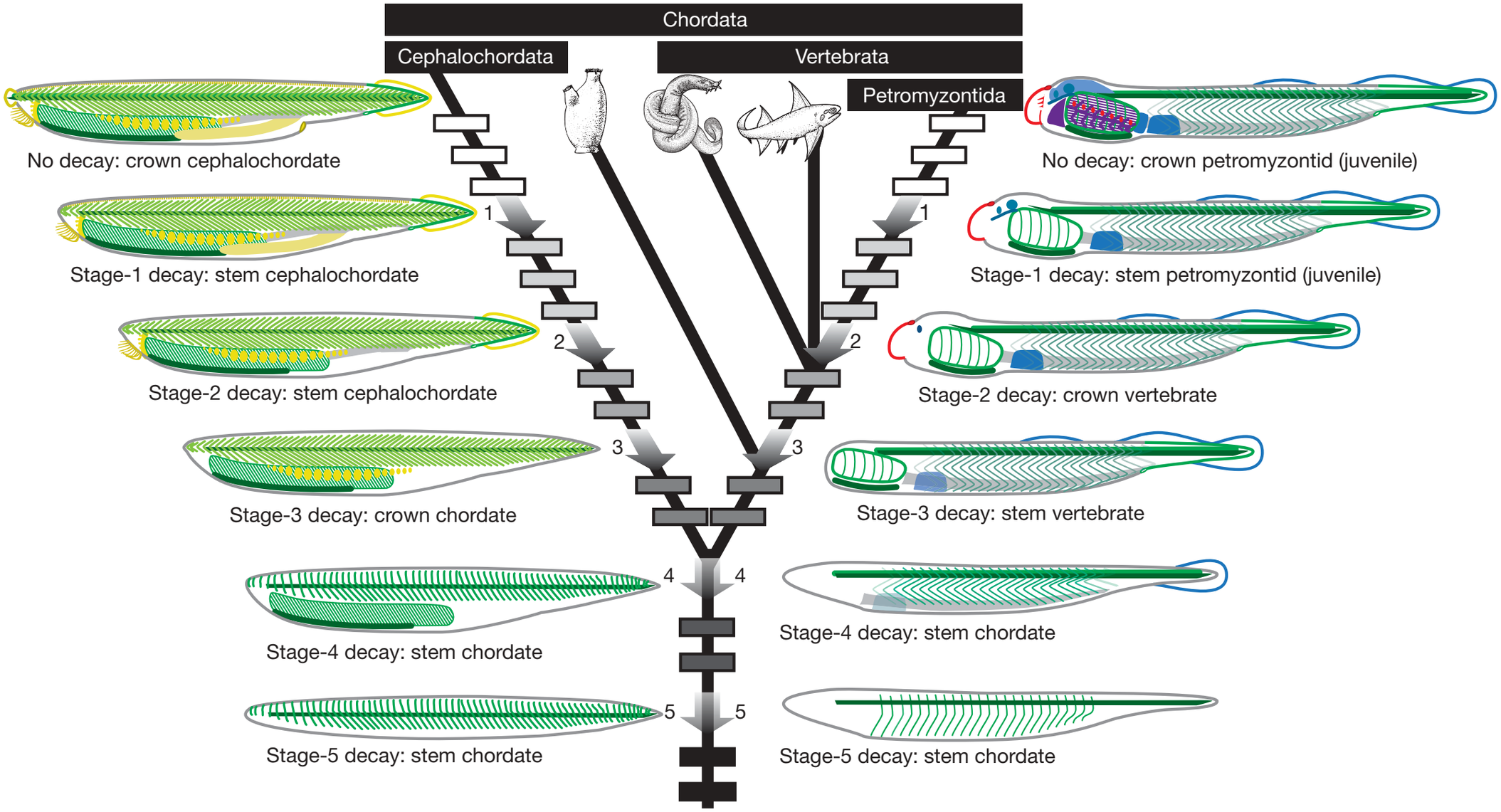
Donoghue, Purnell. 2005. TRENDS

Taxonomic uncertainty: crown versus stem groups

- Crown group: all descendants of the last common ancestor of the living members of a group
- Stem group: all species more closely related to the living members of a group than to any other
- Total group: stem + crown group members
- Early crown and stem group members may be difficult to distinguish



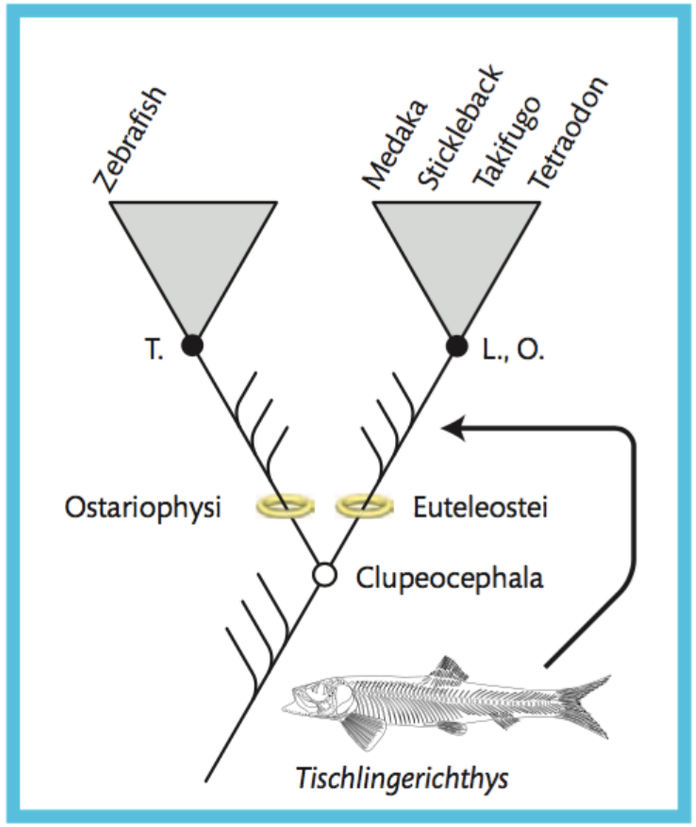
Taxonomic uncertainty: preservation biases



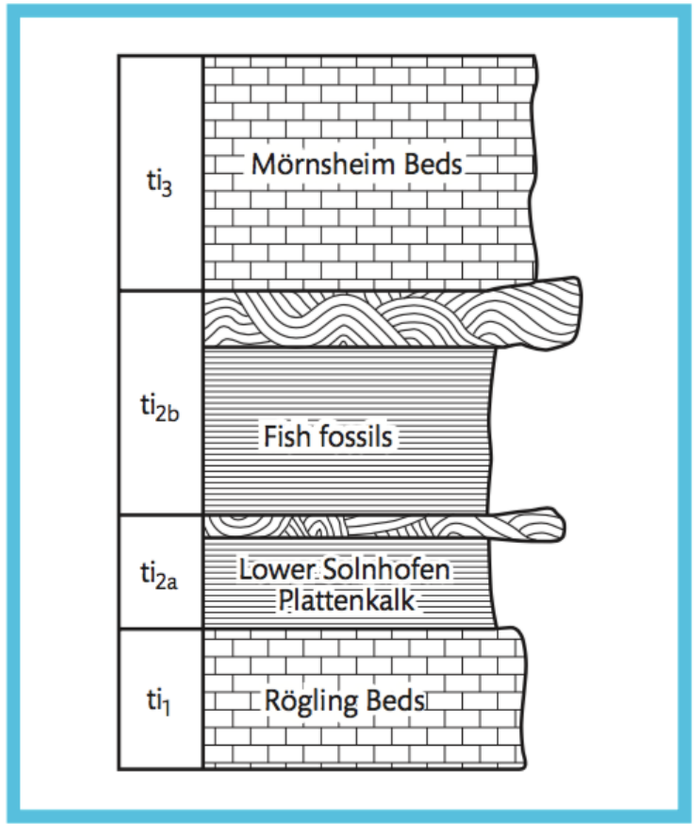
Samson et al. 2010. Nature

Stratigraphic age uncertainty

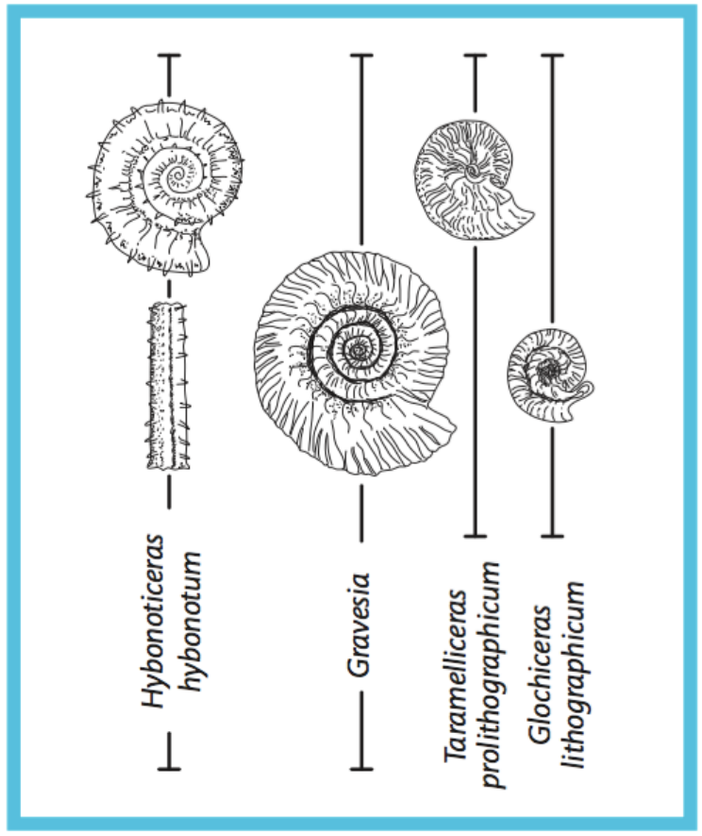
1. Oldest certain fossil in lineage



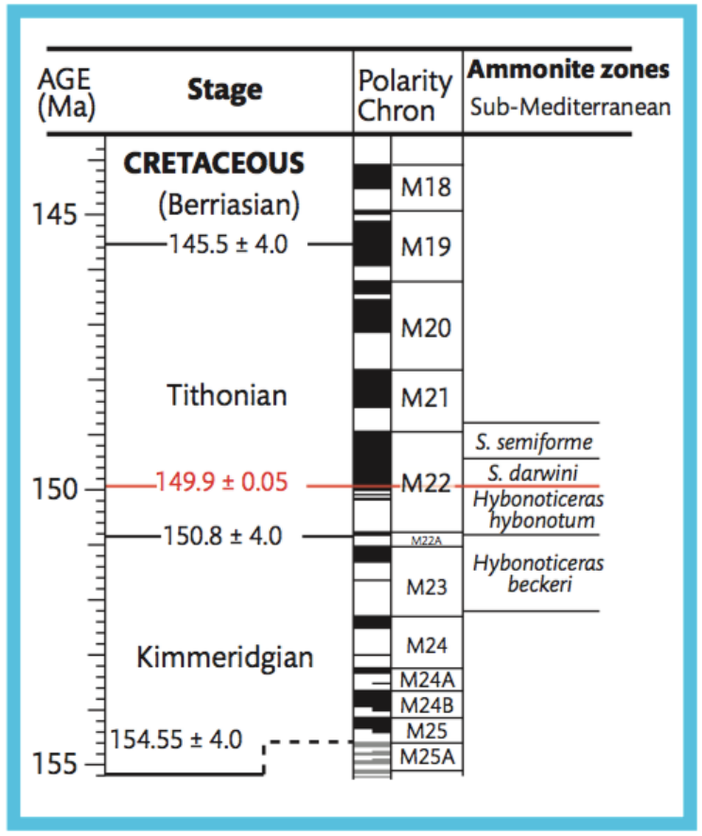
2. Lithostratigraphy of formation



3. Biostratigraphy

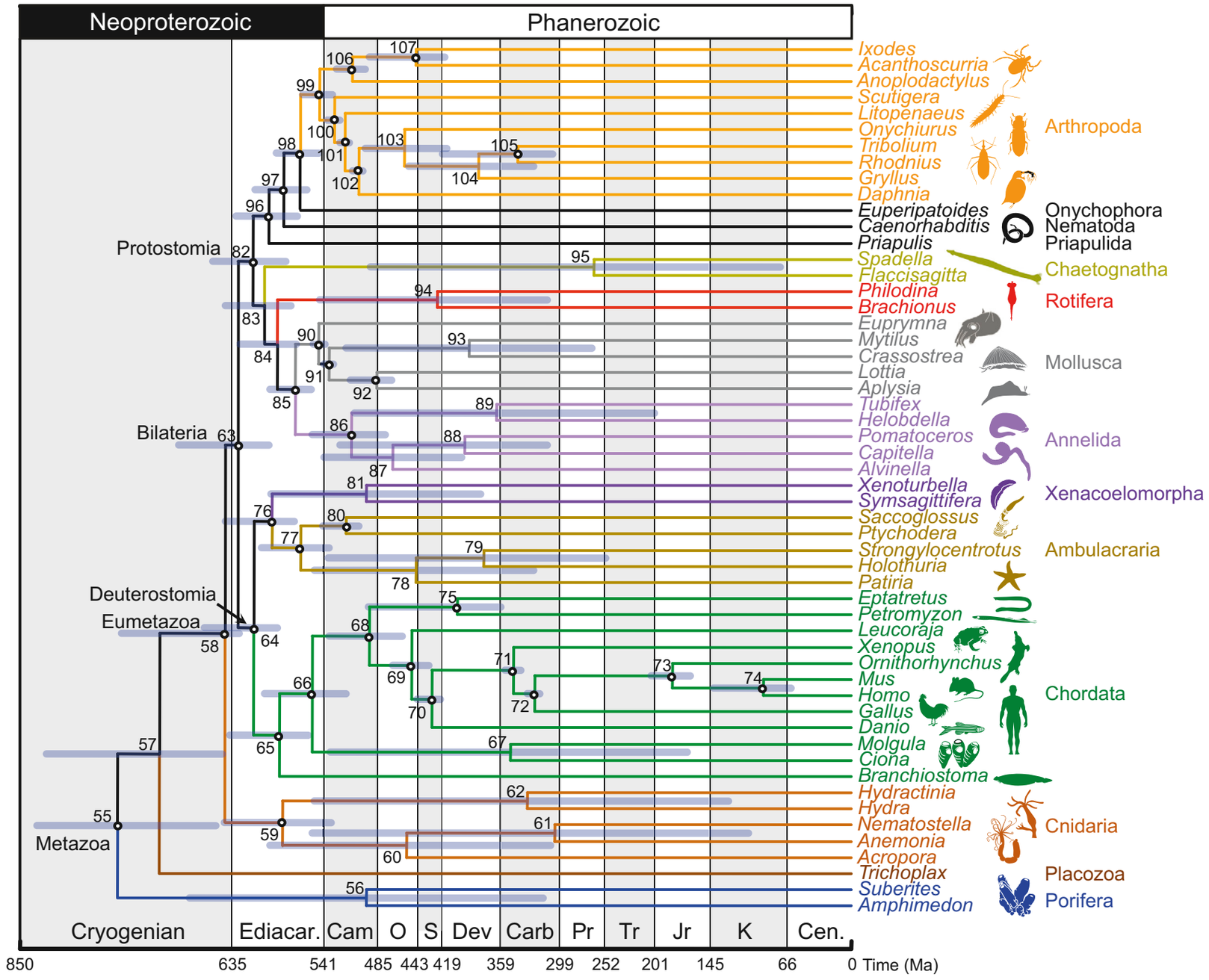


4. Chronostratigraphy



Benton et al. 2009. TimeTree of Life

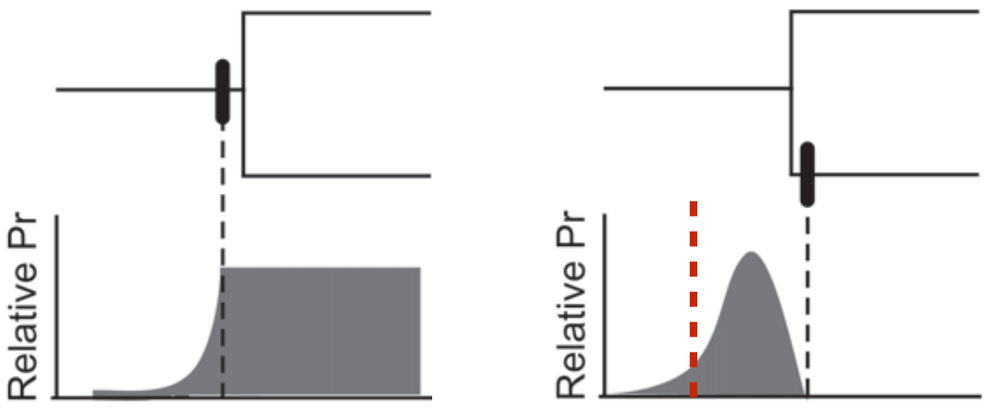
Soft maximum constraints on divergence times



dos Reis et al. 2015. Current Biology

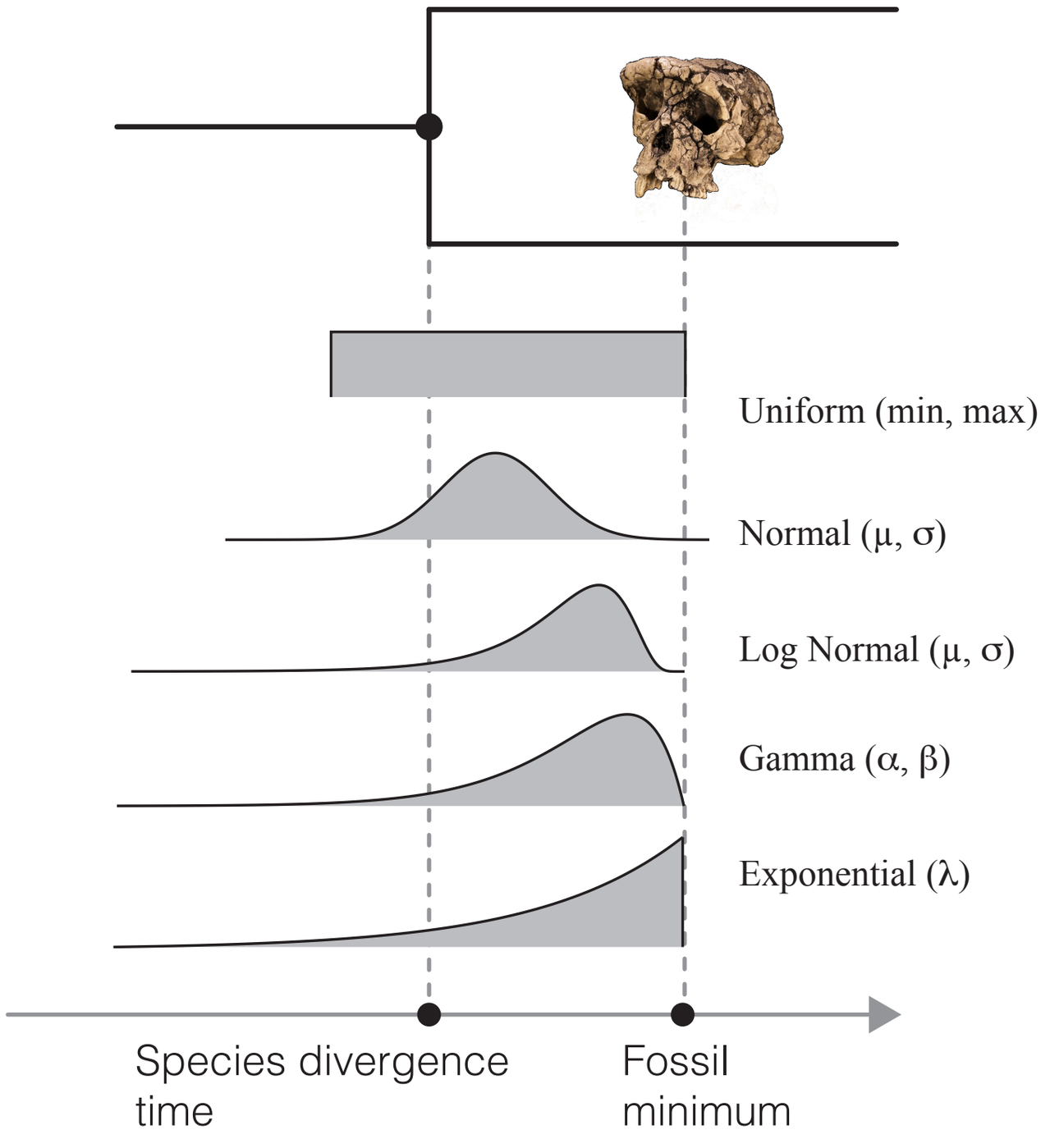
Soft maximum constraints on divergence times

- Minimum constraint: 552.85 Ma Kimberella, interpreted as a protostome
- Maximum constraint: 833 Ma Lagerstätten in the Bitter Springs and Svanbergfjellet Formations, preserve in 3 dimensions at the cellular level prokaryotes, sphaeromorph acritarchs, multicellular algae, but nothing that could be interpreted as a total group metazoan



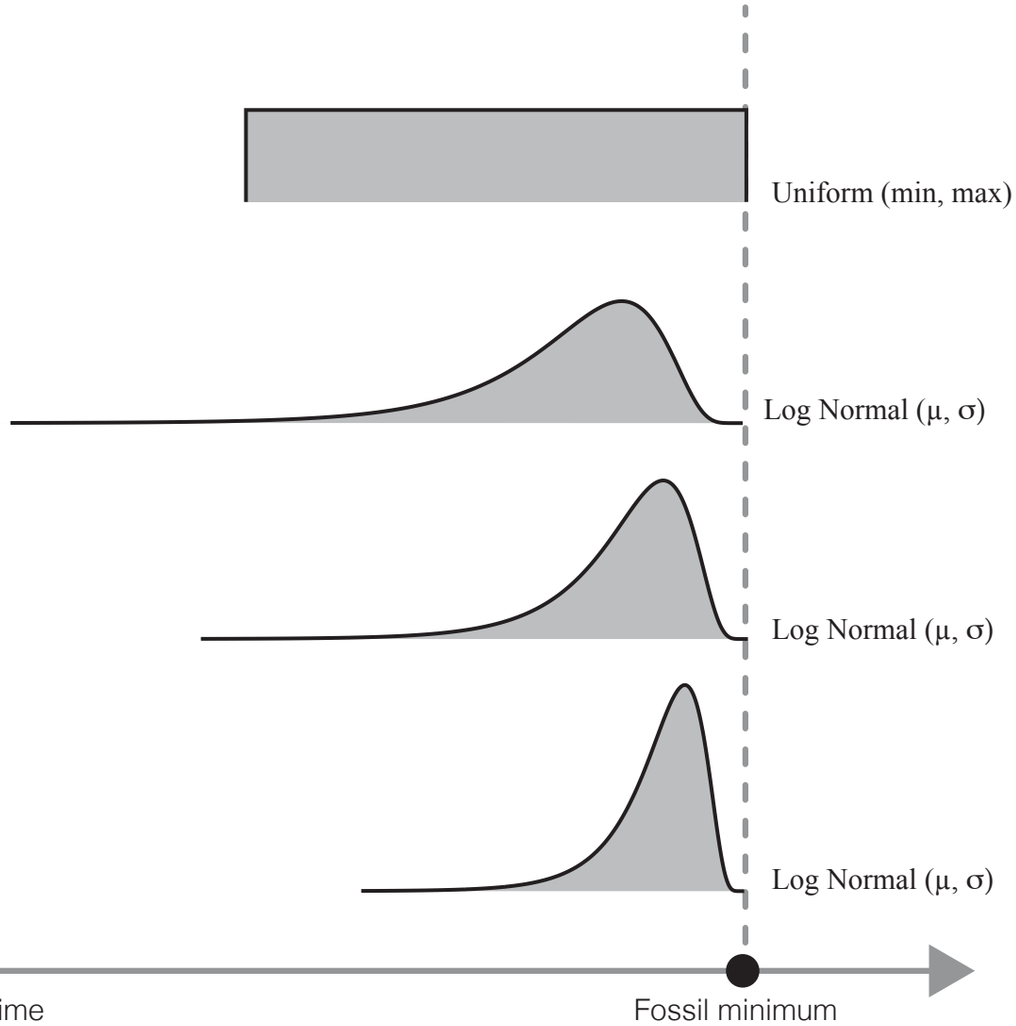
Calibrations: Benton et al. 2015. Palaeontologica Electronica. Plots: Ho, Phillips. 2009, Sys Bio

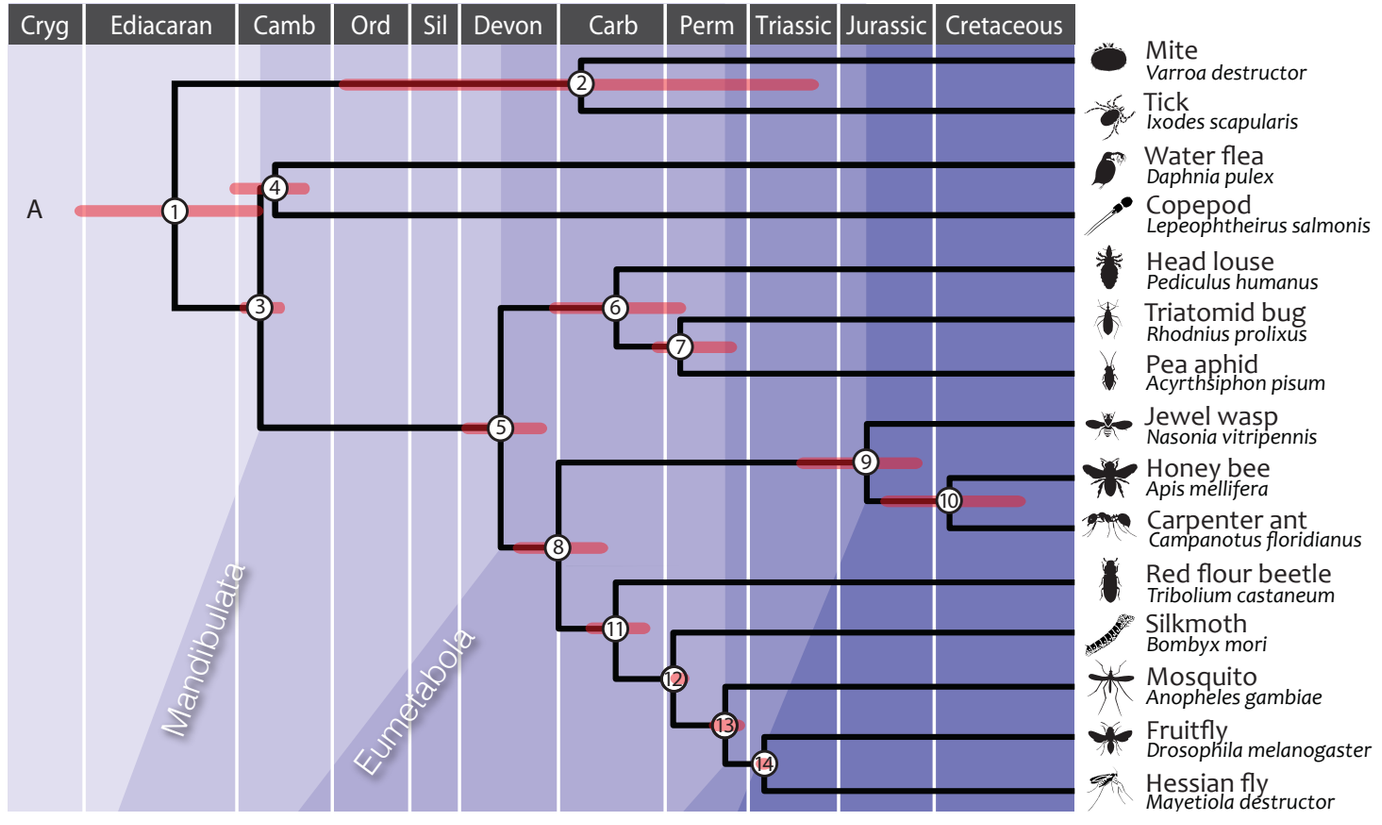
Probabilistic divergence times priors



Modified from Heath 2012. Sys Bio

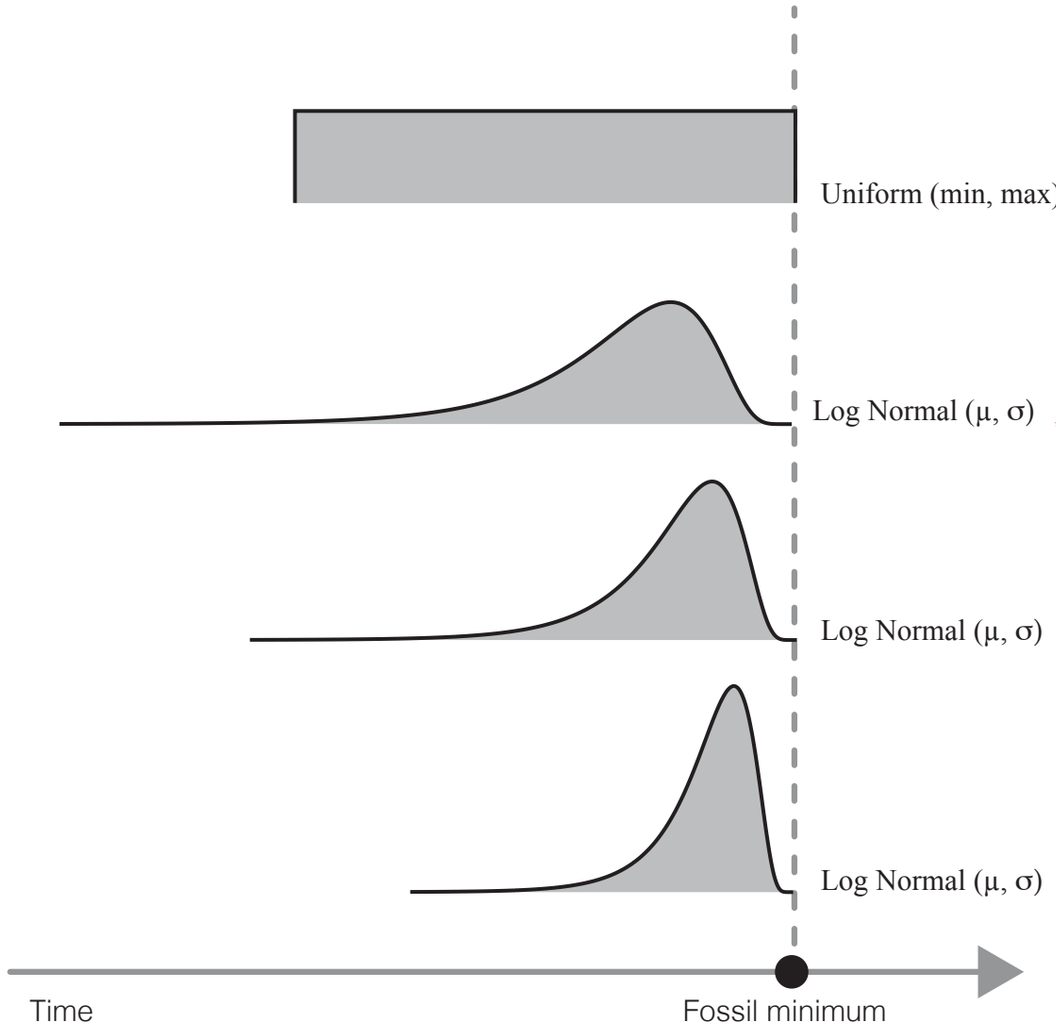
The impact of different calibration priors

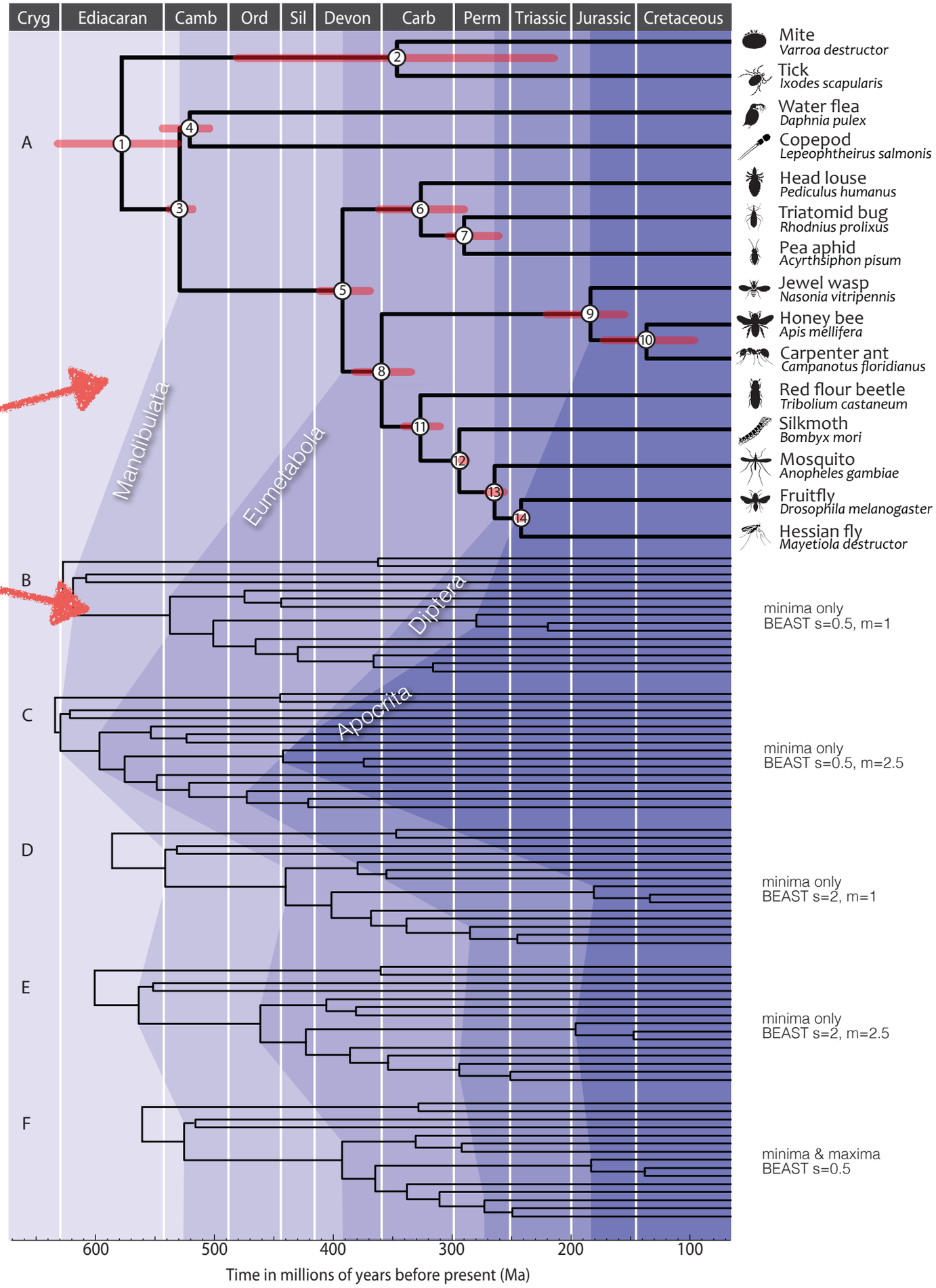




Warnock et al. 2012. Biology Letters

The impact of different calibration priors



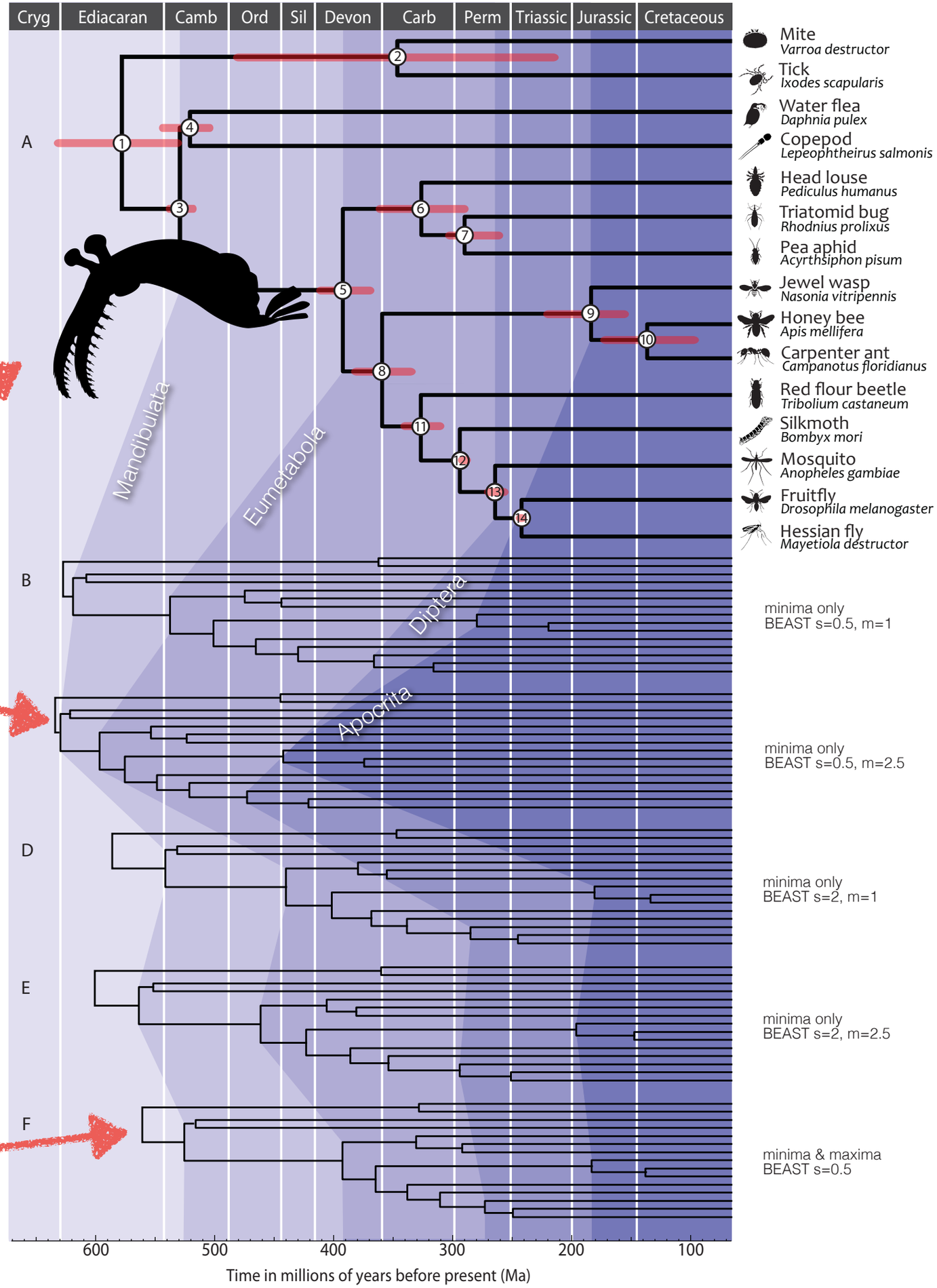


The impact of different calibration priors

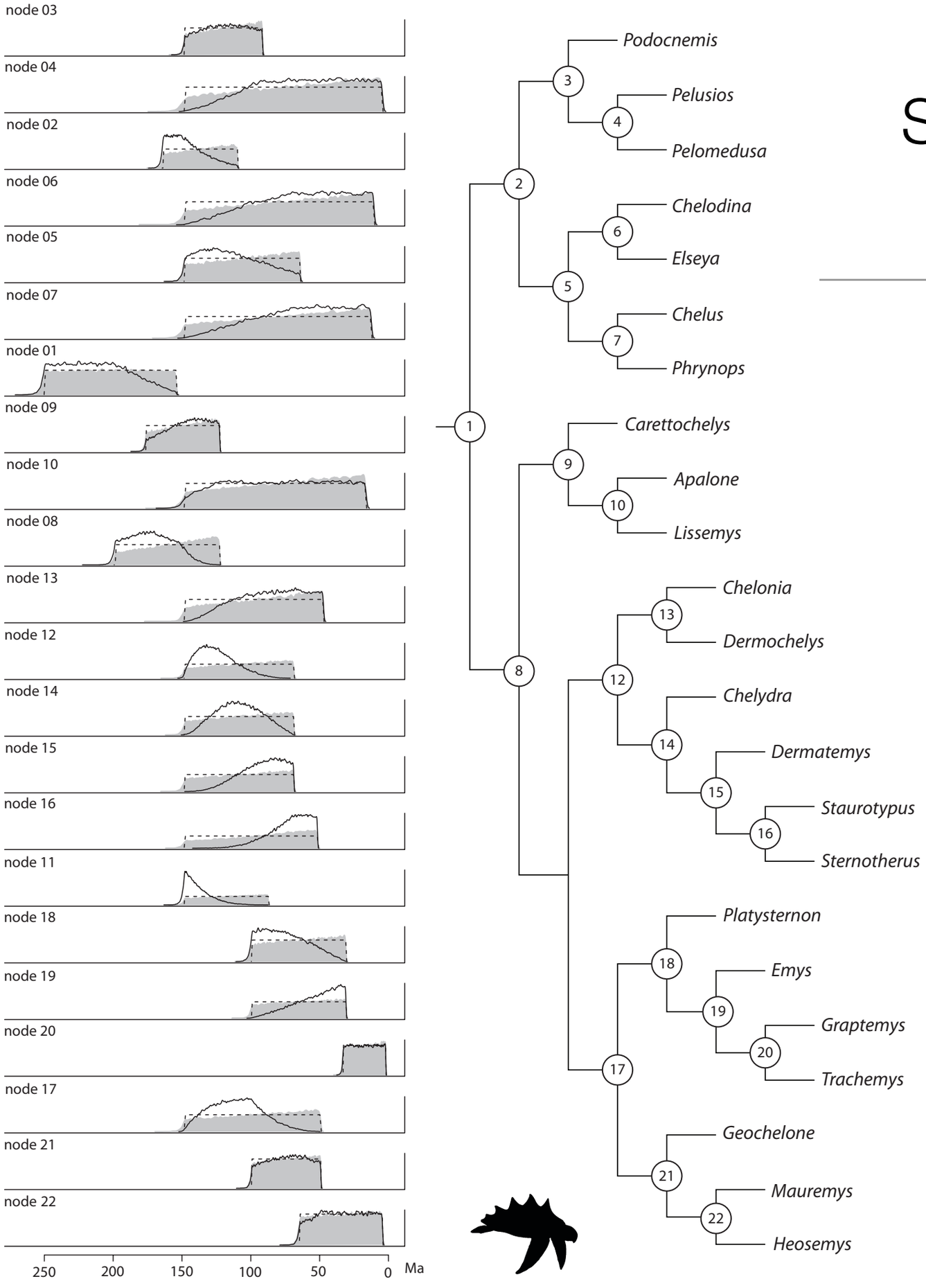
Cambrian explosion

Cryogenian

Ediacaran



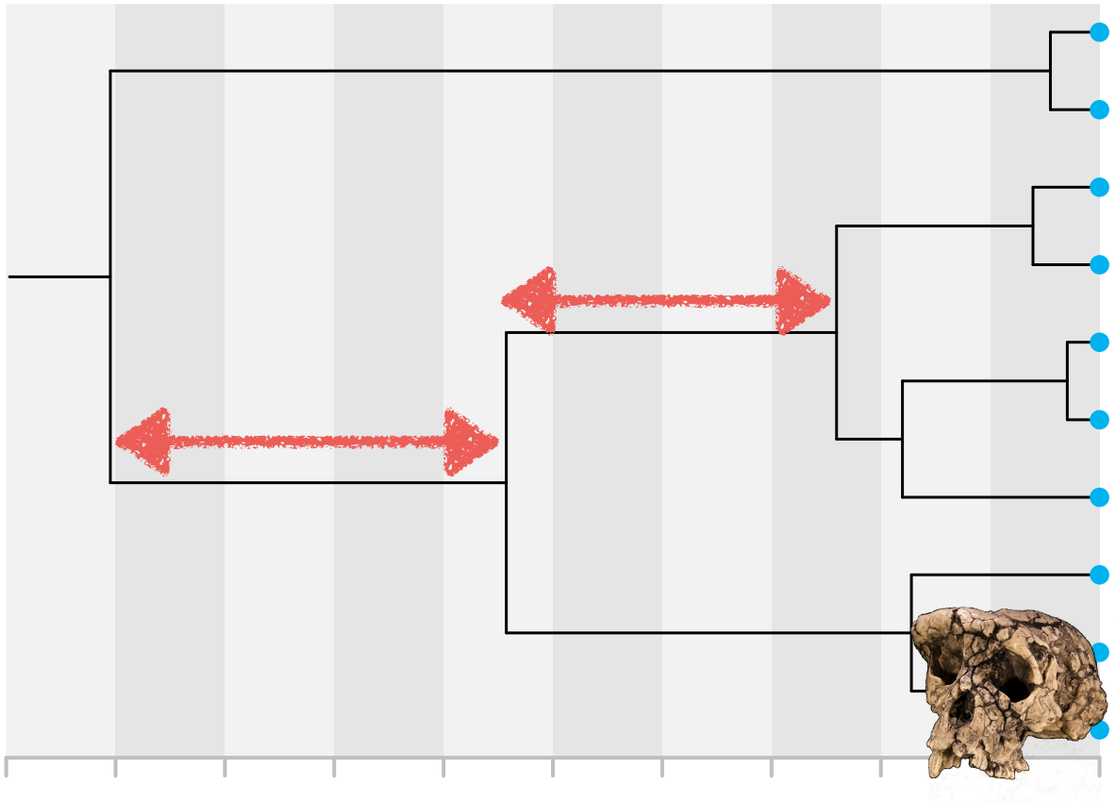
Specified versus effective priors



Warnock et al. 2015. Biology Letters

A prior for the non-fossil calibrated nodes

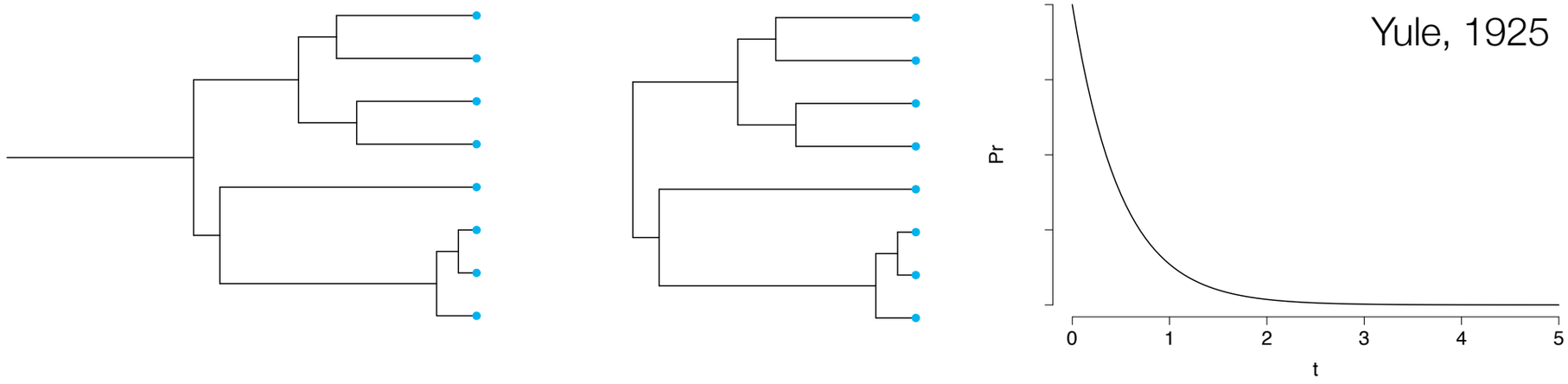
Extant species phylogeny



A prior for the uncalibrated nodes

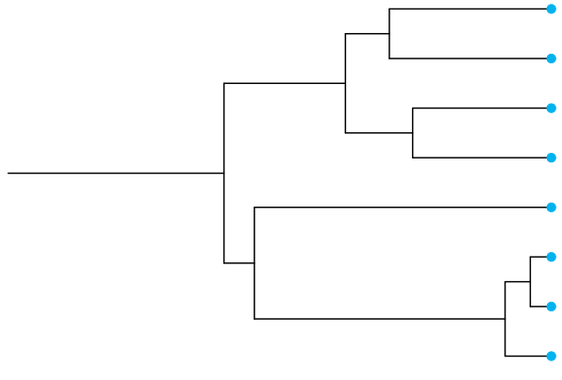
A prior for the non-fossil calibrated nodes: the tree prior

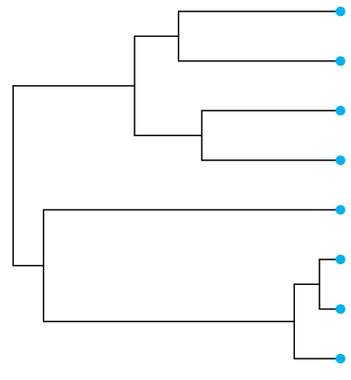
$\lambda = 1$

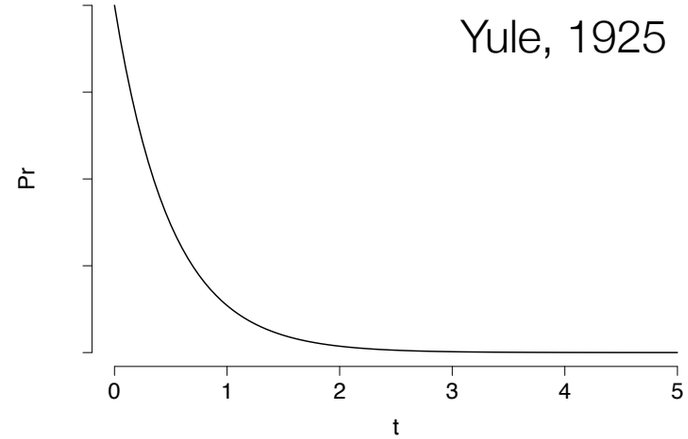


A prior for the non-fossil calibrated nodes: the tree prior

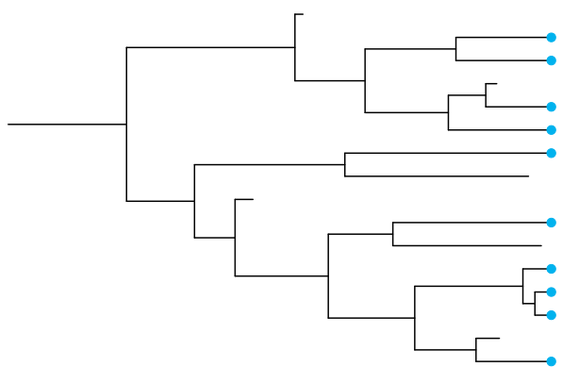
λ = 1

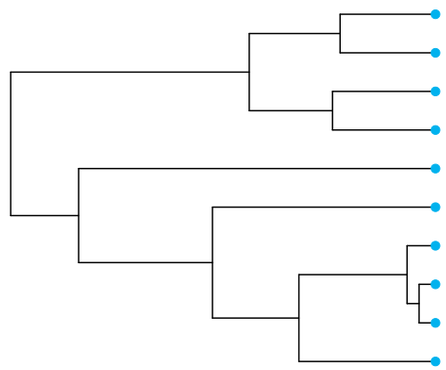


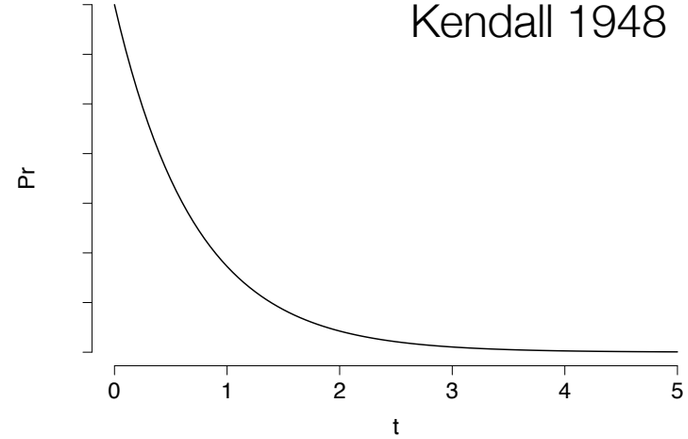




λ = 1, μ = 0.3



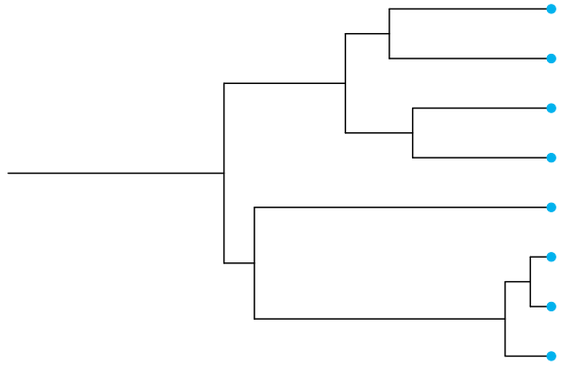


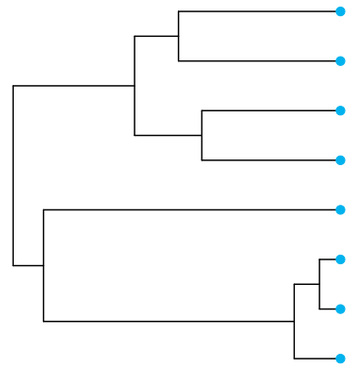


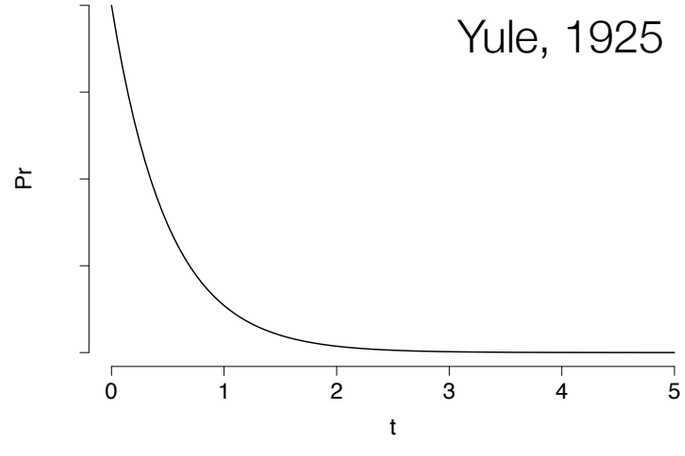
- Different combinations of λ and μ produce different tree shapes

A prior for the non-fossil calibrated nodes: the tree prior

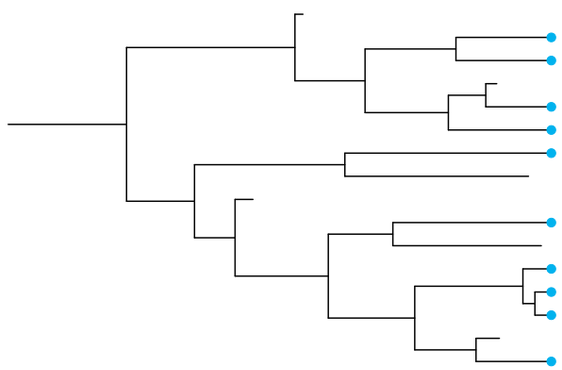
$\lambda = 1$

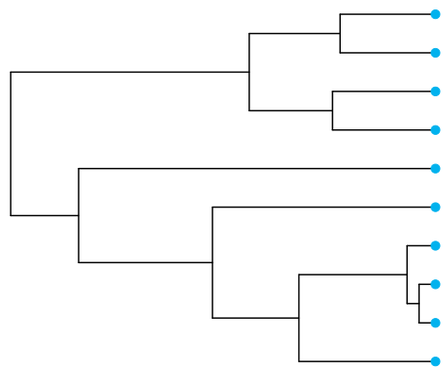


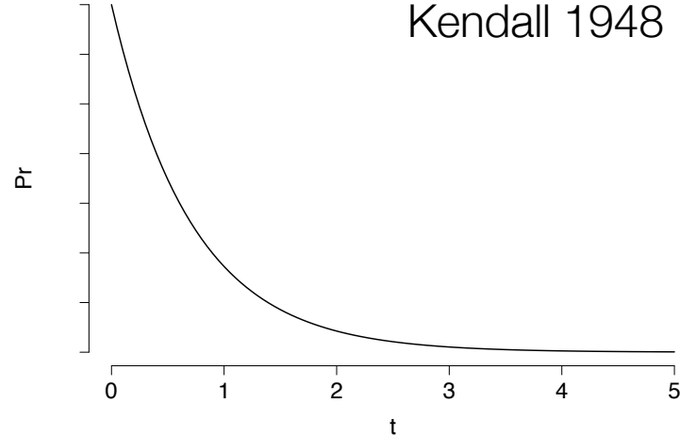




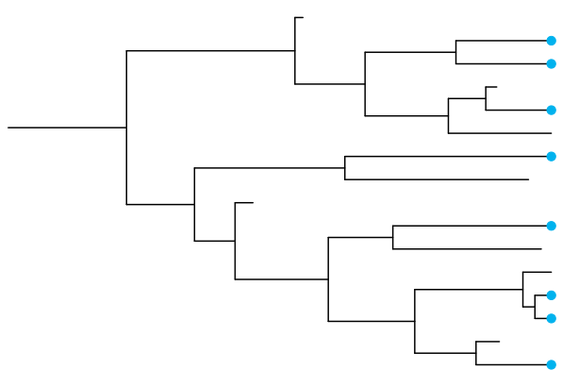
$\lambda = 1, \mu = 0.3$

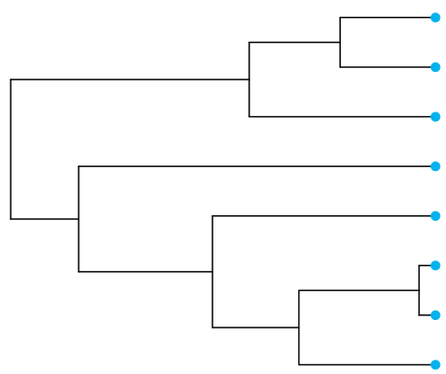


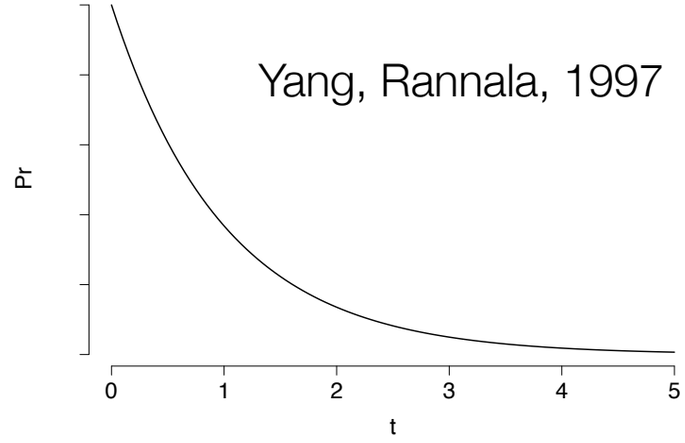




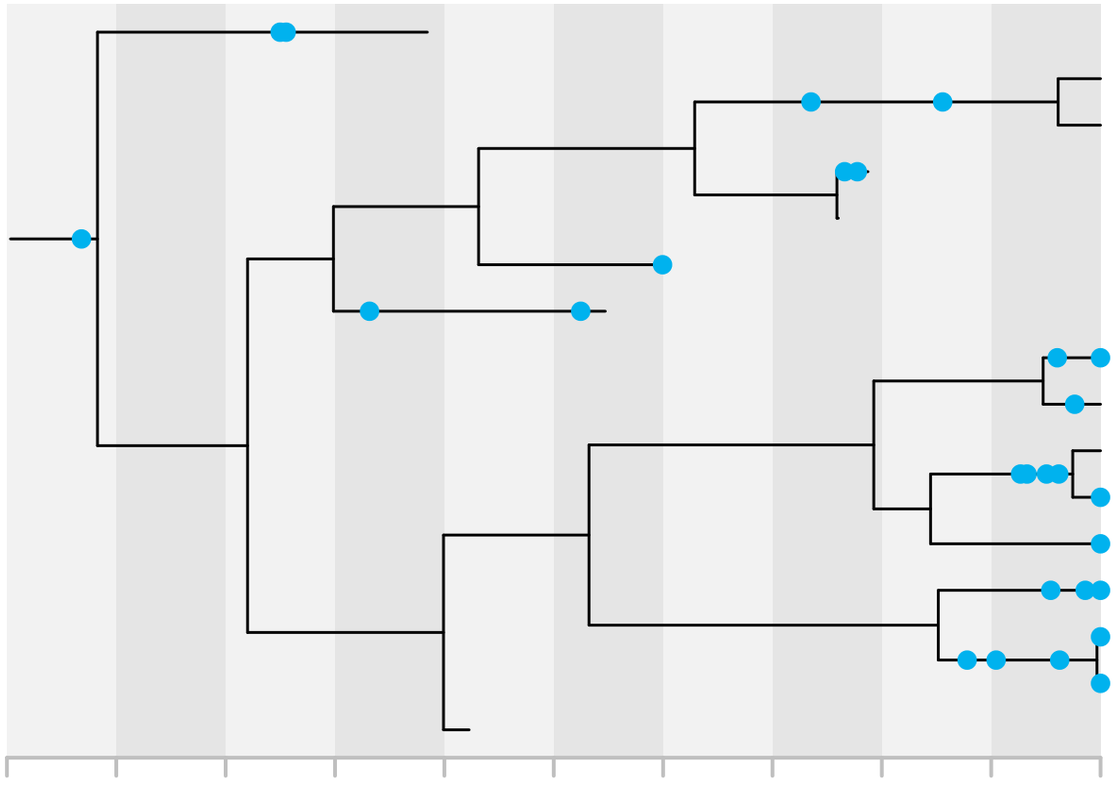
$\lambda = 1, \mu = 0.3,$
$\rho = 0.5$







The fossilized birth-death process



λ — speciation rate

μ — extinction rate

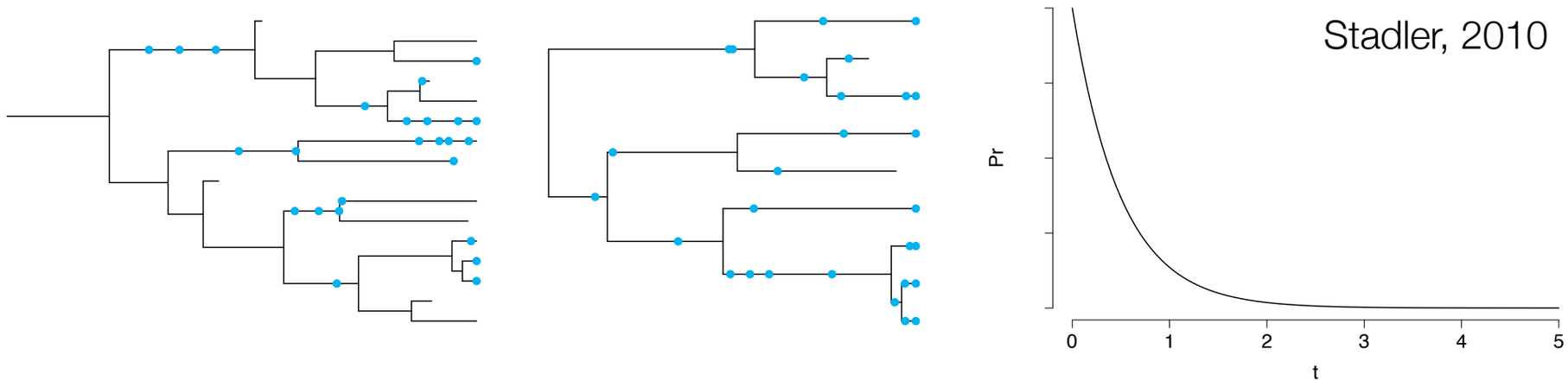
ψ — fossil sampling rate

ρ — extant species sampling

Stadler, 2010

Incorporating fossils into the tree prior

$\lambda = 1, \mu = 0.3,$
$\rho = 0.5, \psi = 1$



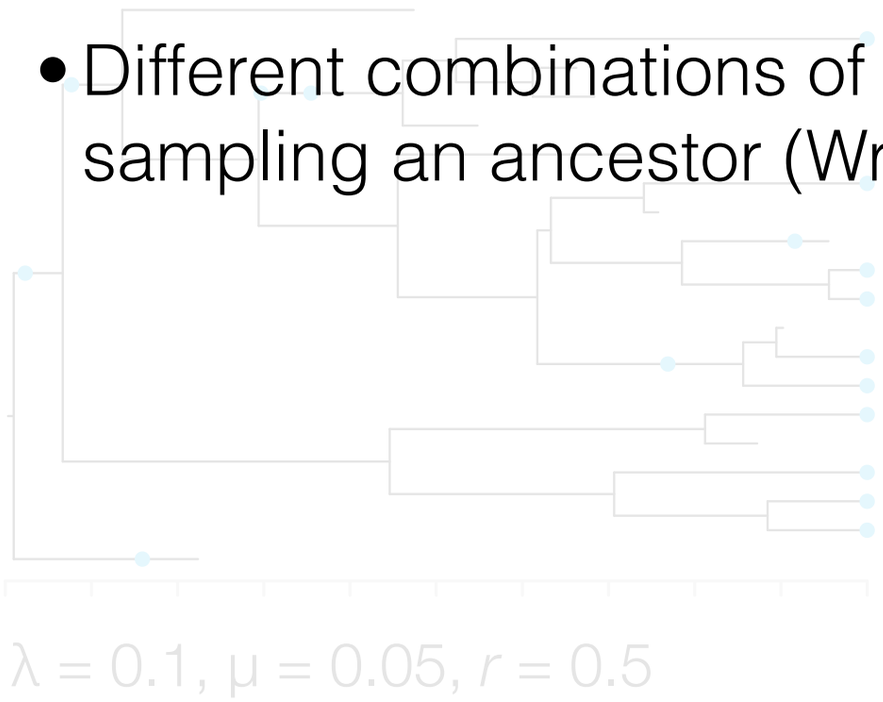
- The simpler models can be considered special cases of the FBD model
- Traditional node dating may depend on a small number of fossils
- Molecular clock analyses will be sensitive to both the calibration priors & the tree prior

Sampled ancestors

- The probability of sampling an ancestor in the fossil record is not zero (Foote, 1996. Paleobiology)

- Bayesian inference of sampled ancestor trees now possible in BEAST2, MrBayes, RevBayes, DPPDiv

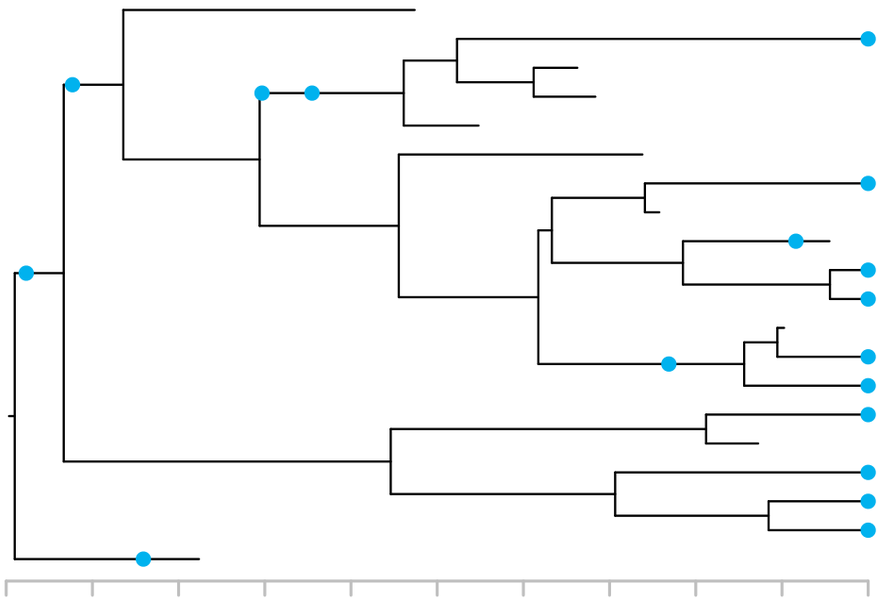
- Different combinations of λ, μ, ρ, ψ affect the probability of having a sampling an ancestor (Wright, Heath. in prep)



Sampled ancestors

$\lambda = 0.1$, $\mu = 0.03$, $r = 0.3$

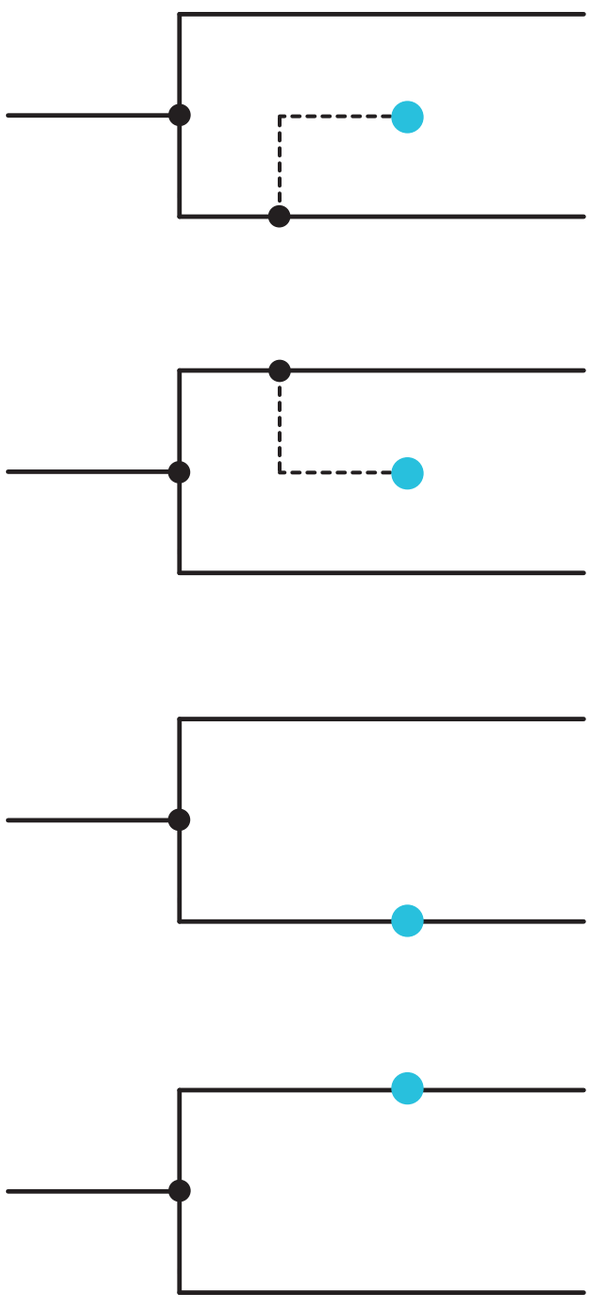
$\lambda = 0.1$, $\mu = 0.08$, $r = 0.8$



$\lambda = 0.1$, $\mu = 0.05$, $r = 0.5$

Wright et al. in prep.

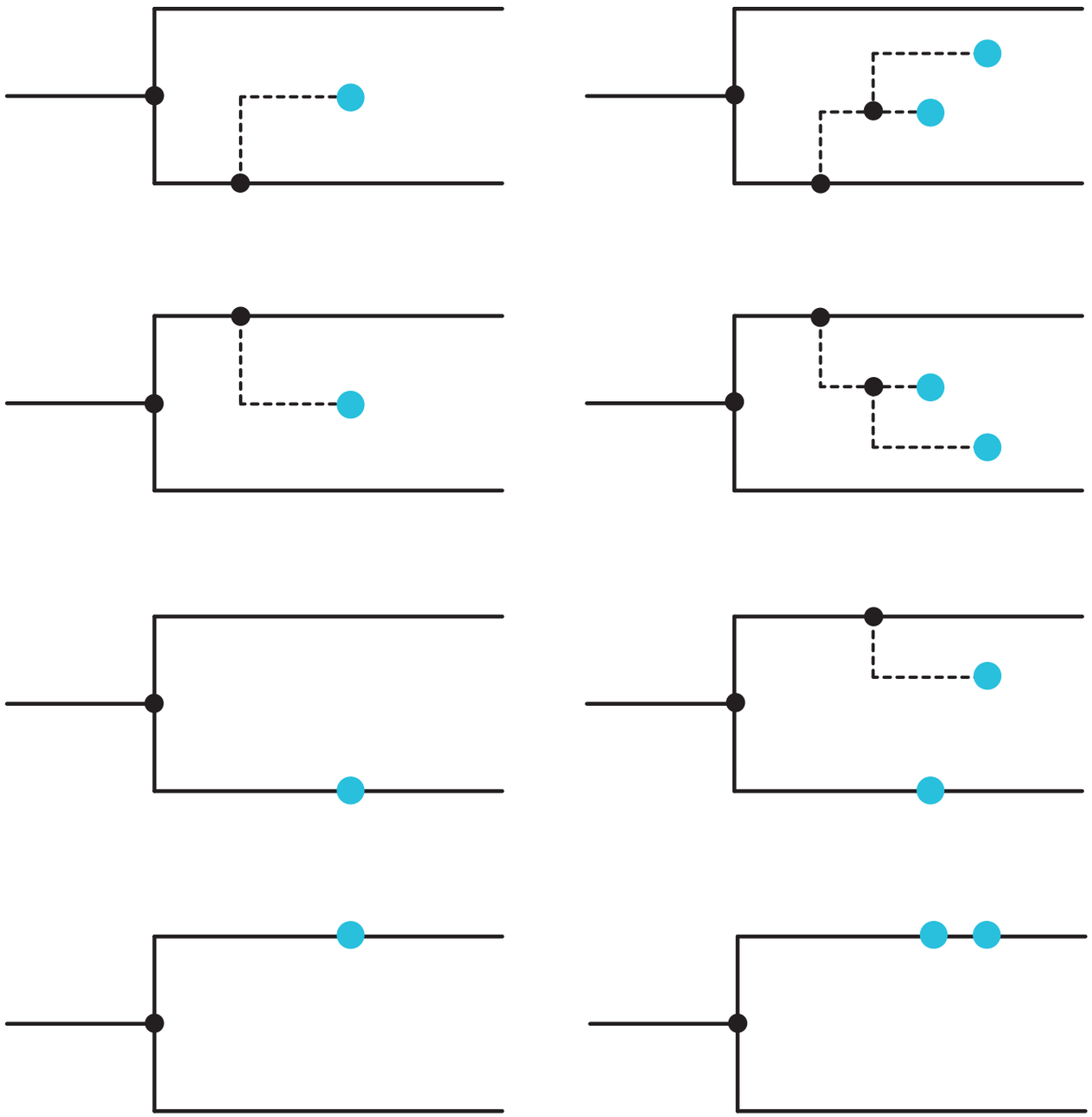
The fossilised birth-death process: taxonomic uncertainty



- MCMC is used to propose possible fossil placements
- rjMCMC is used to propose sampled ancestor placements

Fossil occurrence
Speciation event

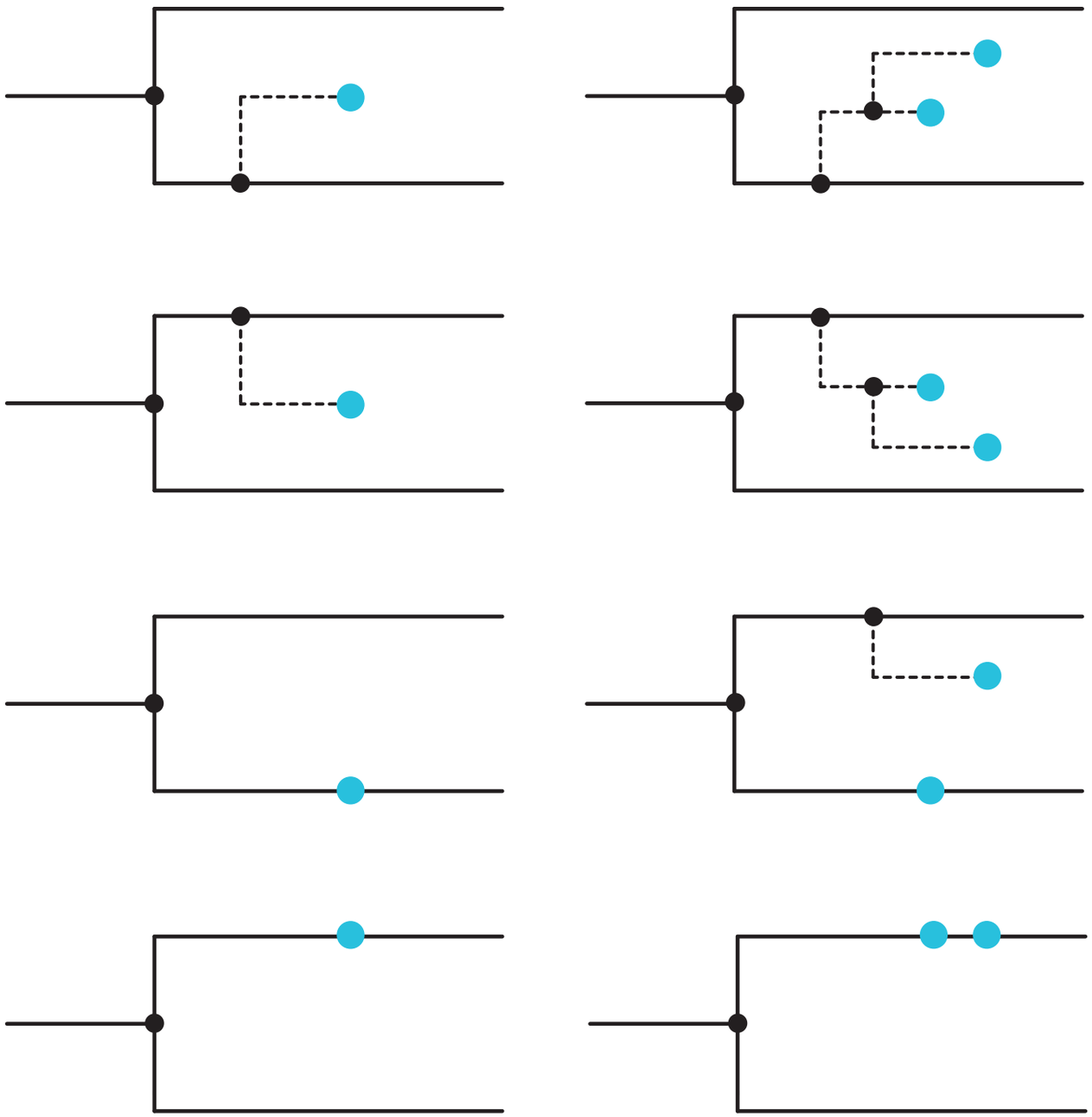
The fossilised birth-death process: taxonomic uncertainty



- Each fossil can attach anywhere along the tree, including along unobserved branches

Fossil occurrence
Speciation event

The fossilised birth-death process: taxonomic uncertainty



- The probability of any given realisation of the FBD process is conditional on the model parameters: λ, μ, ρ, ψ

Fossil occurrence
Speciation event

Application with or without character data for fossils

- Three alternative scenarios to applying to FBD model in divergence time estimation
  - Molecular data for extant taxa only

    ACAC...
    TCAC...
    ACAG...
  - Molecular data for extant taxa + morphological data for both extant & extinct taxa

    ACAC... 0101...
    TCAC... 1101...
    ACAG... 0100...
  - Morphological data for both extant and extinct taxa or extinct taxa only

    0101...
    1101...
    0100...

Application with or without character data for fossils

- Three alternative scenarios to applying to FBD model in divergence time estimation

  - Molecular data for extant taxa only

    ACAC...
    TCAC...
    ACAG...

  - Molecular data for extant taxa + morphological data for both extant & extinct taxa

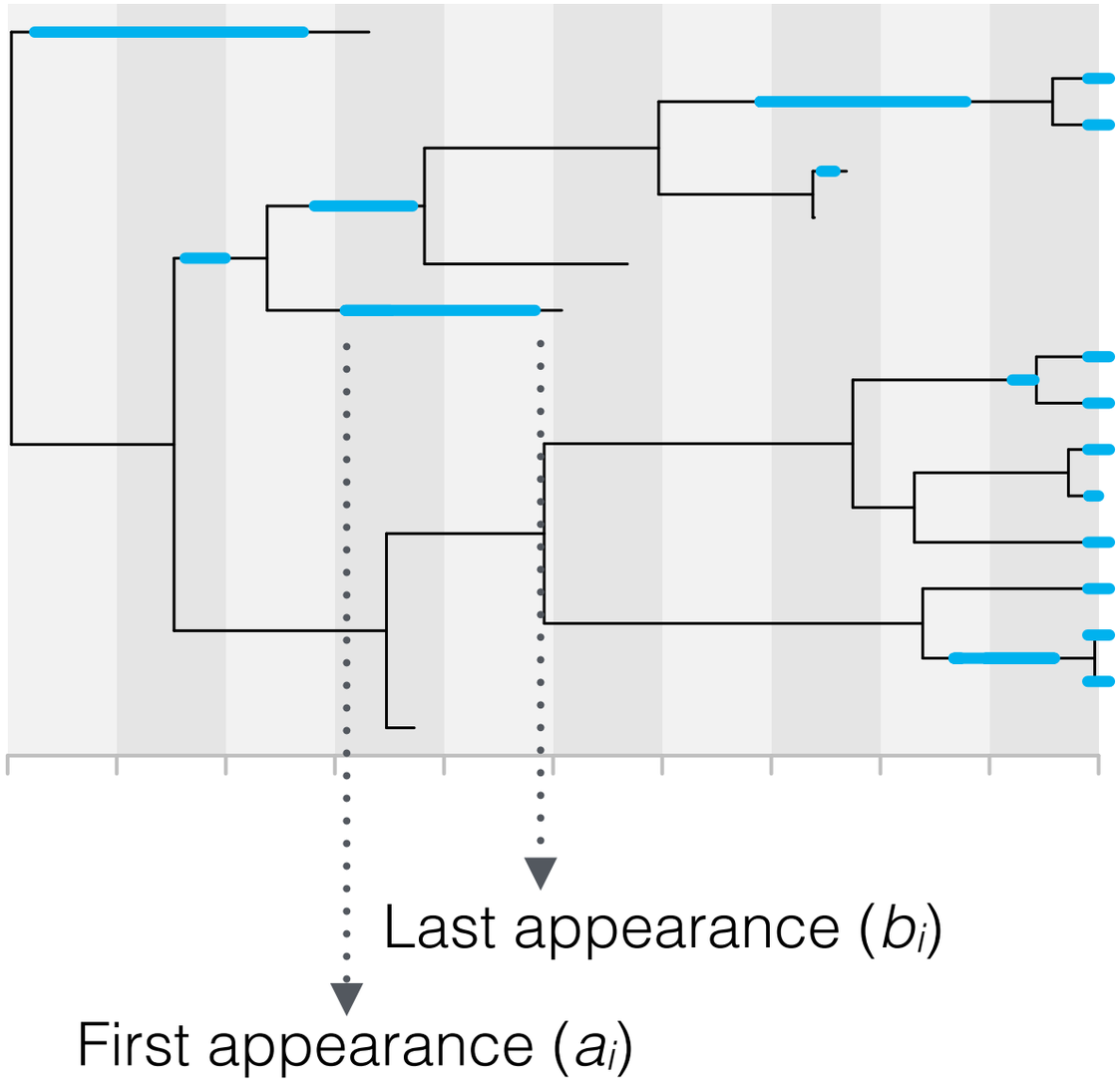
    ACAC... 0101...
    TCAC... 1101...
    ACAG... 0100...

  - Morphological data for both extant and extinct taxa or extinct taxa only

    0101...
    1101...
    0100...

  - The model can also be applied to estimate macroevolutionary parameters when you have no character data

Stratigraphic range data



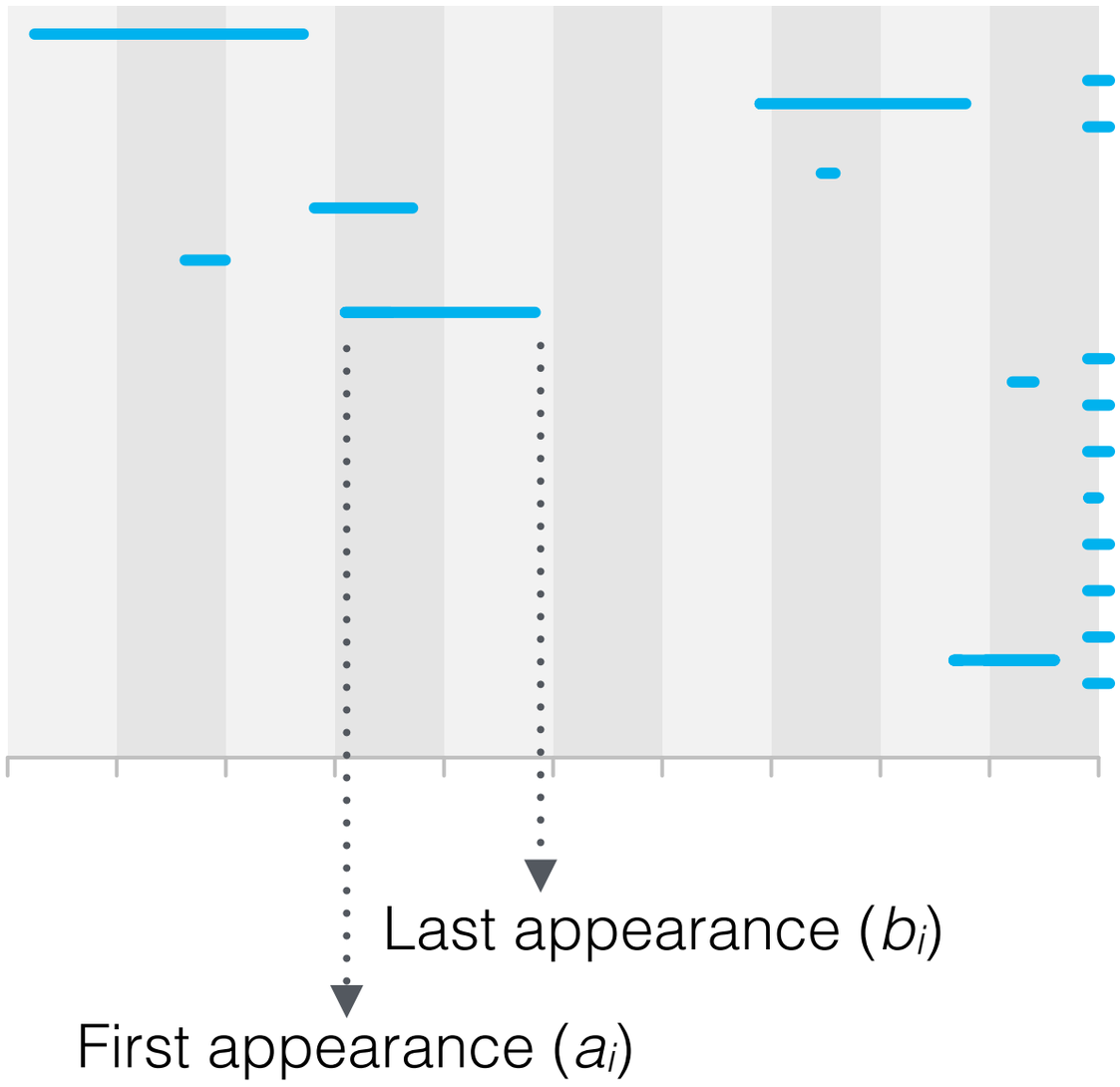
λ — speciation rate

μ — extinction rate

ψ — fossil sampling rate

ρ — extant species sampling

Stratigraphic range data



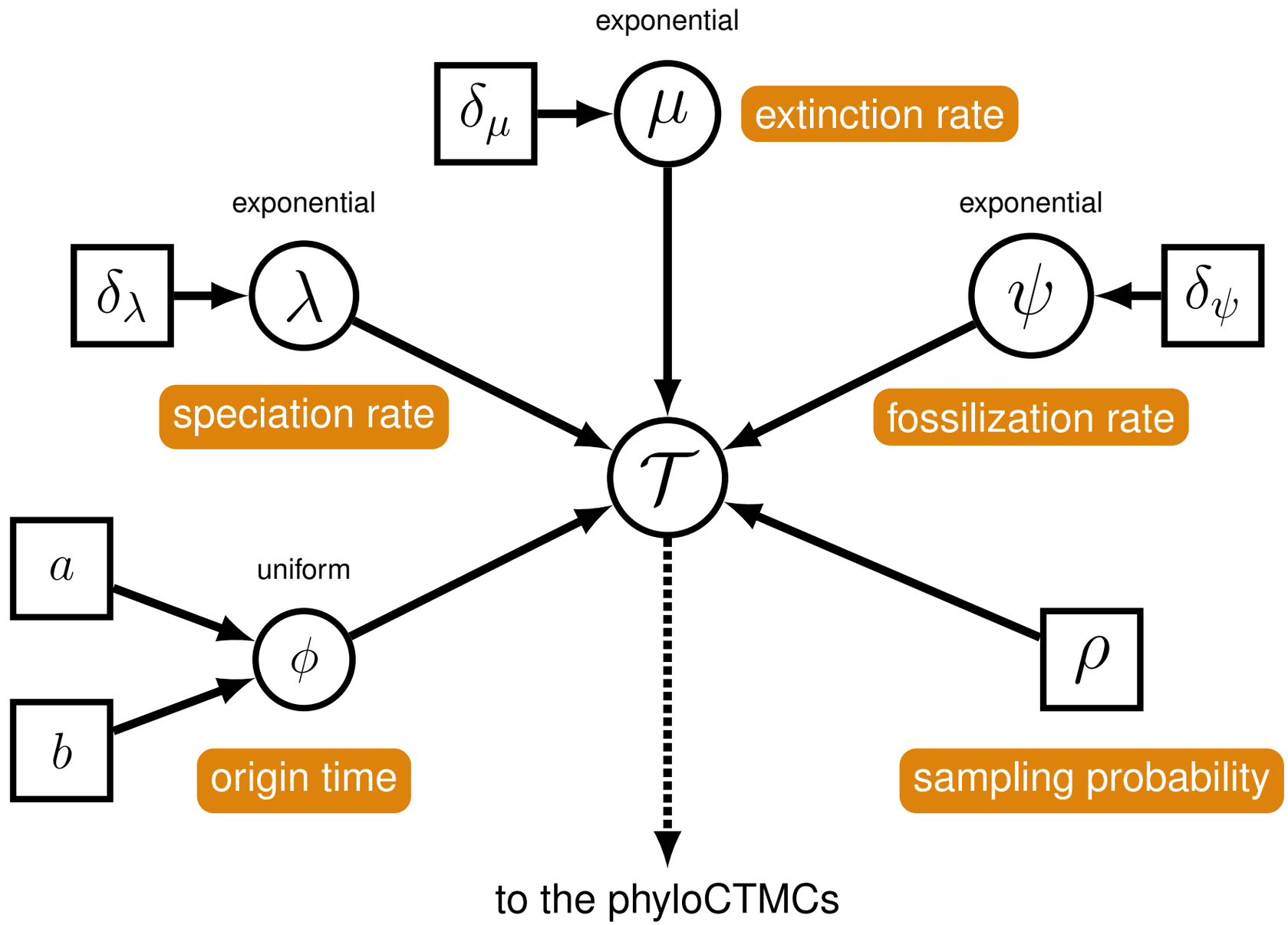
λ — speciation rate

μ — extinction rate

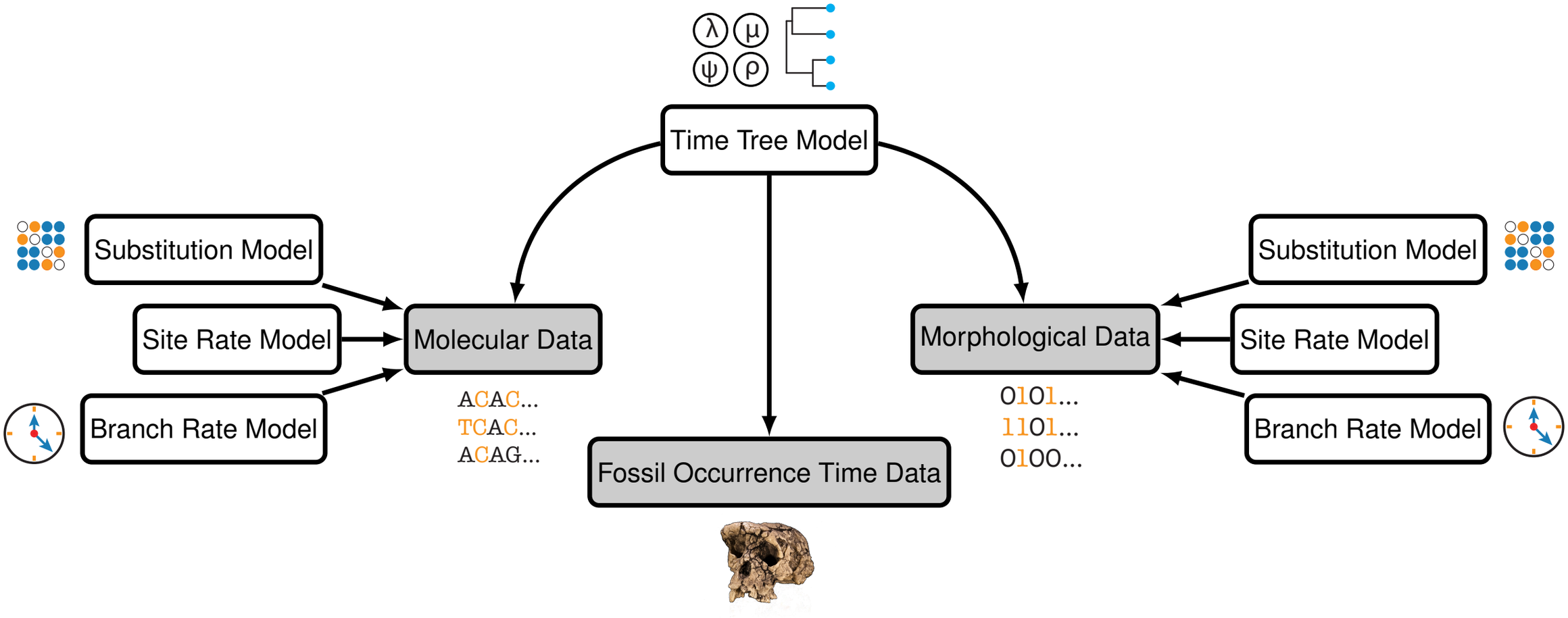
ψ — fossil sampling rate

ρ — extant species sampling

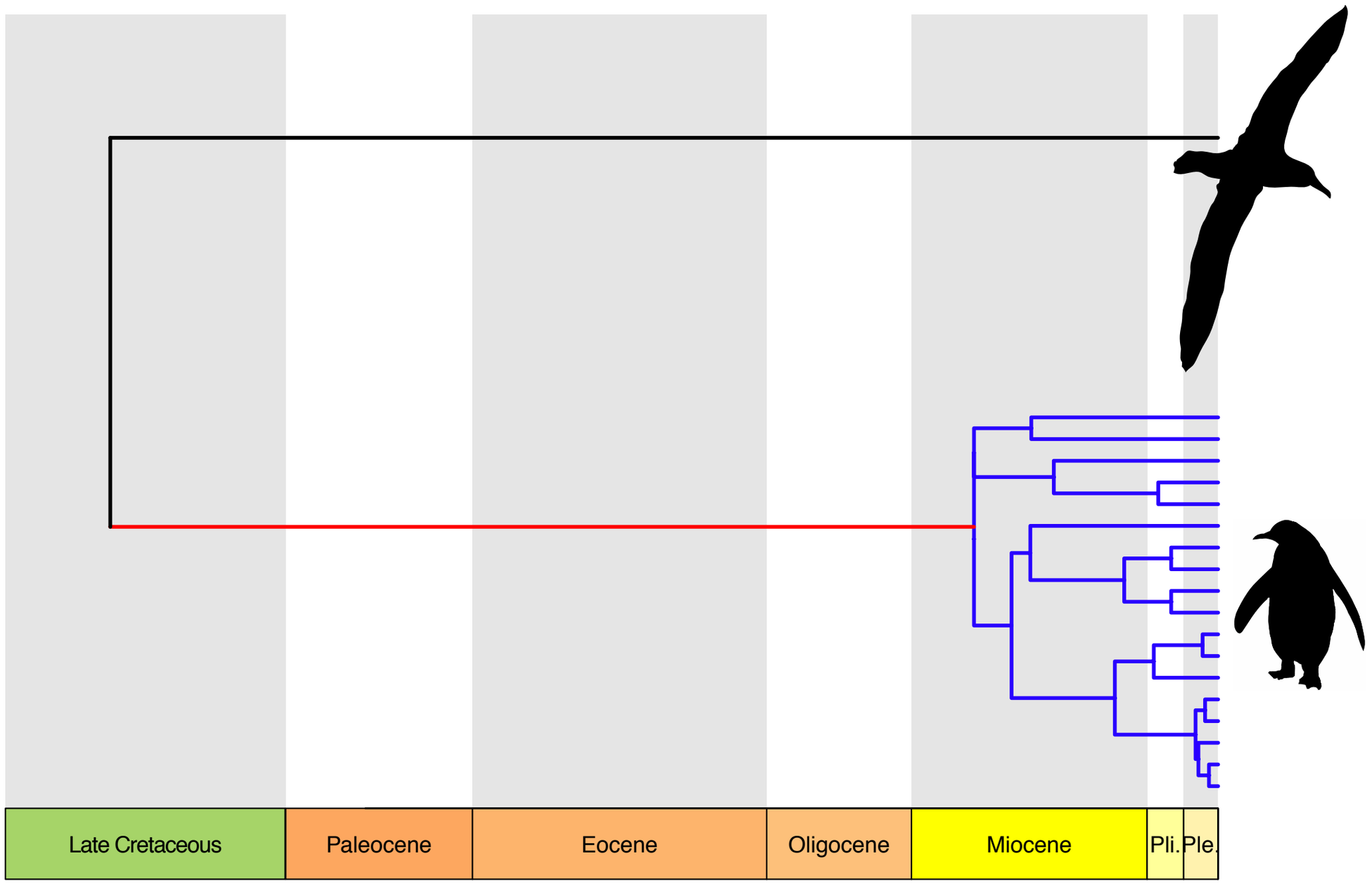
Putting things in a graphical modelling framework



Putting things in a graphical modelling framework

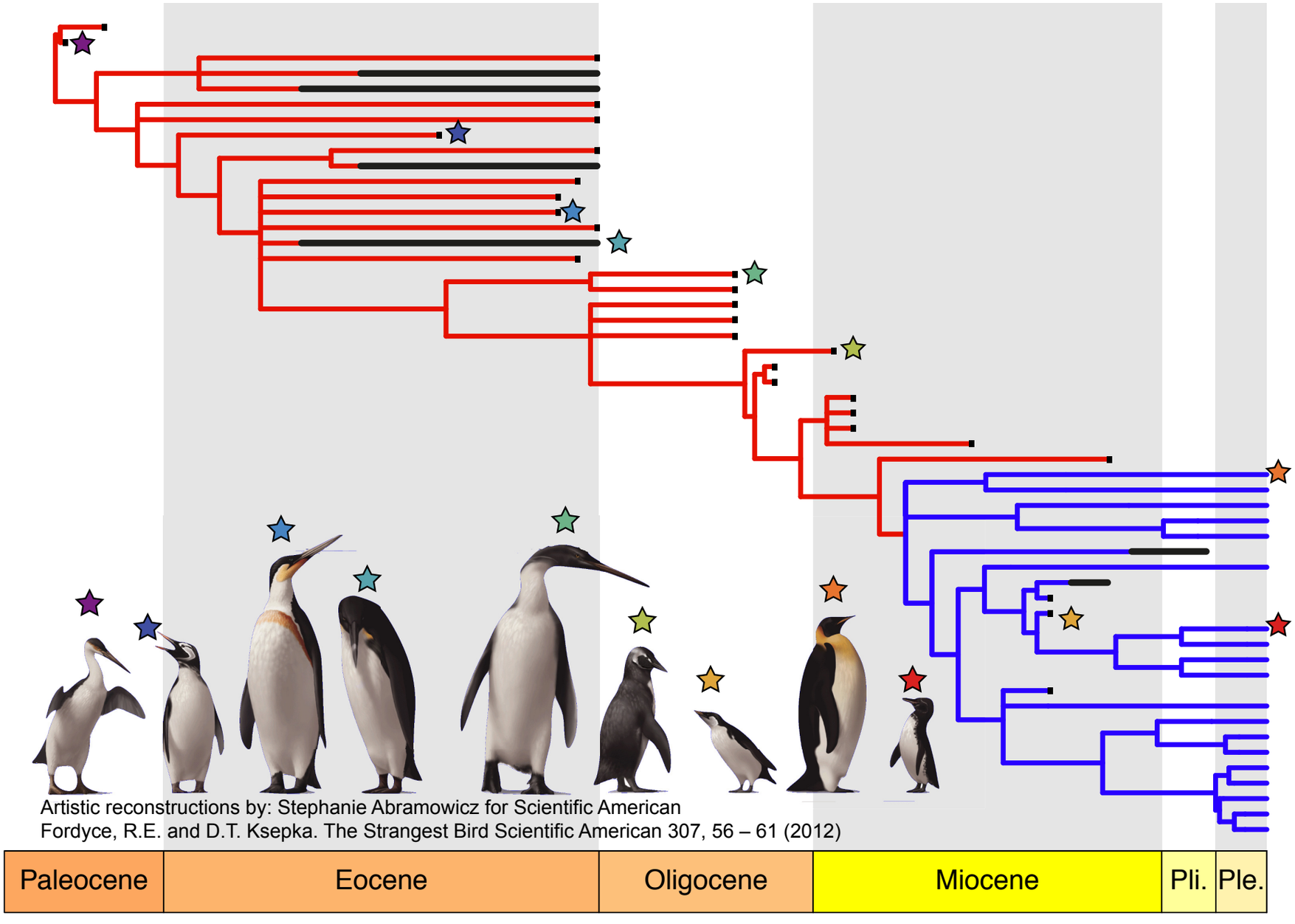


Dating the origin of crown penguins



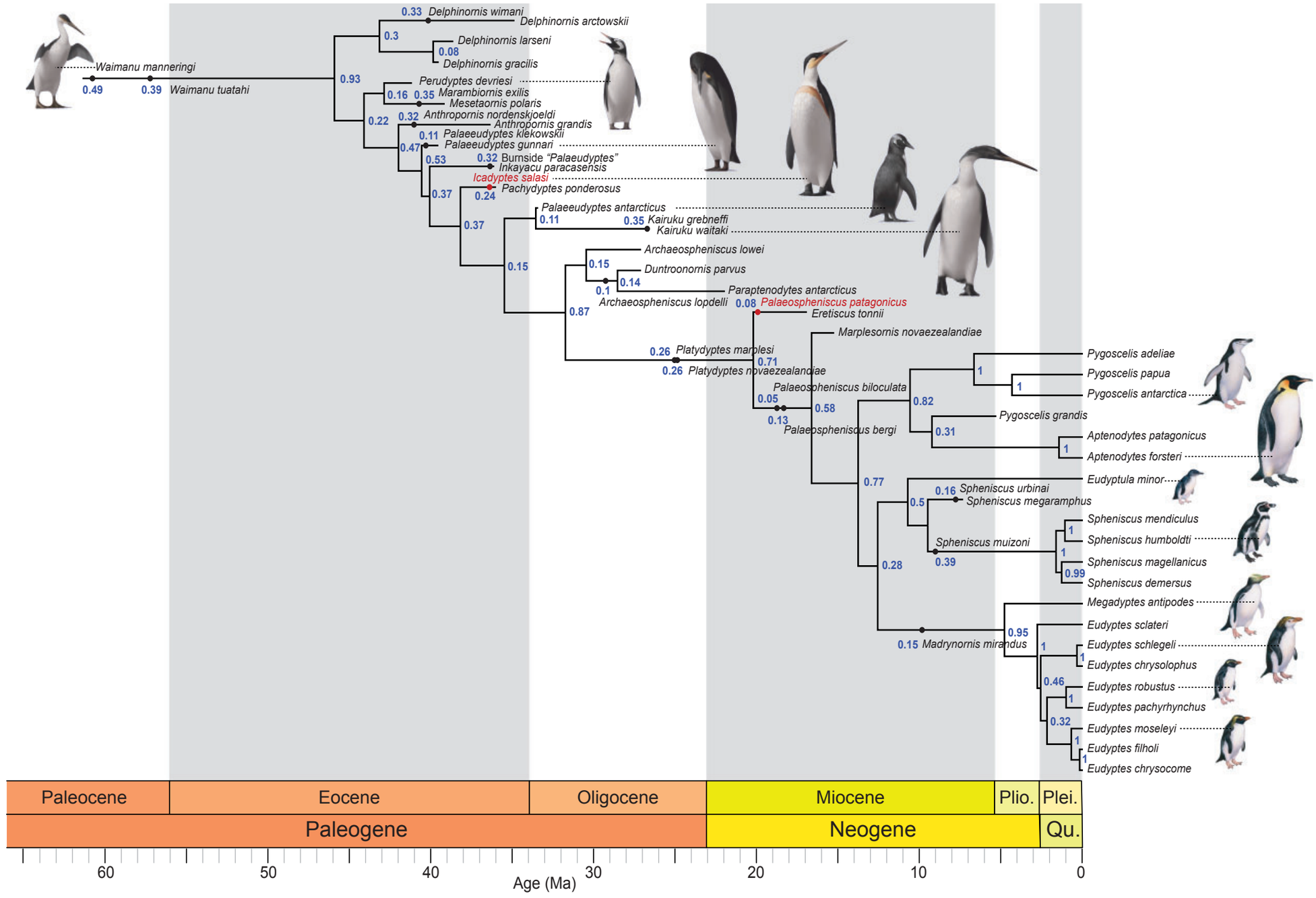
Gavryushkina et al. 2016. *Sys Bio*

Dating the origin of crown penguins



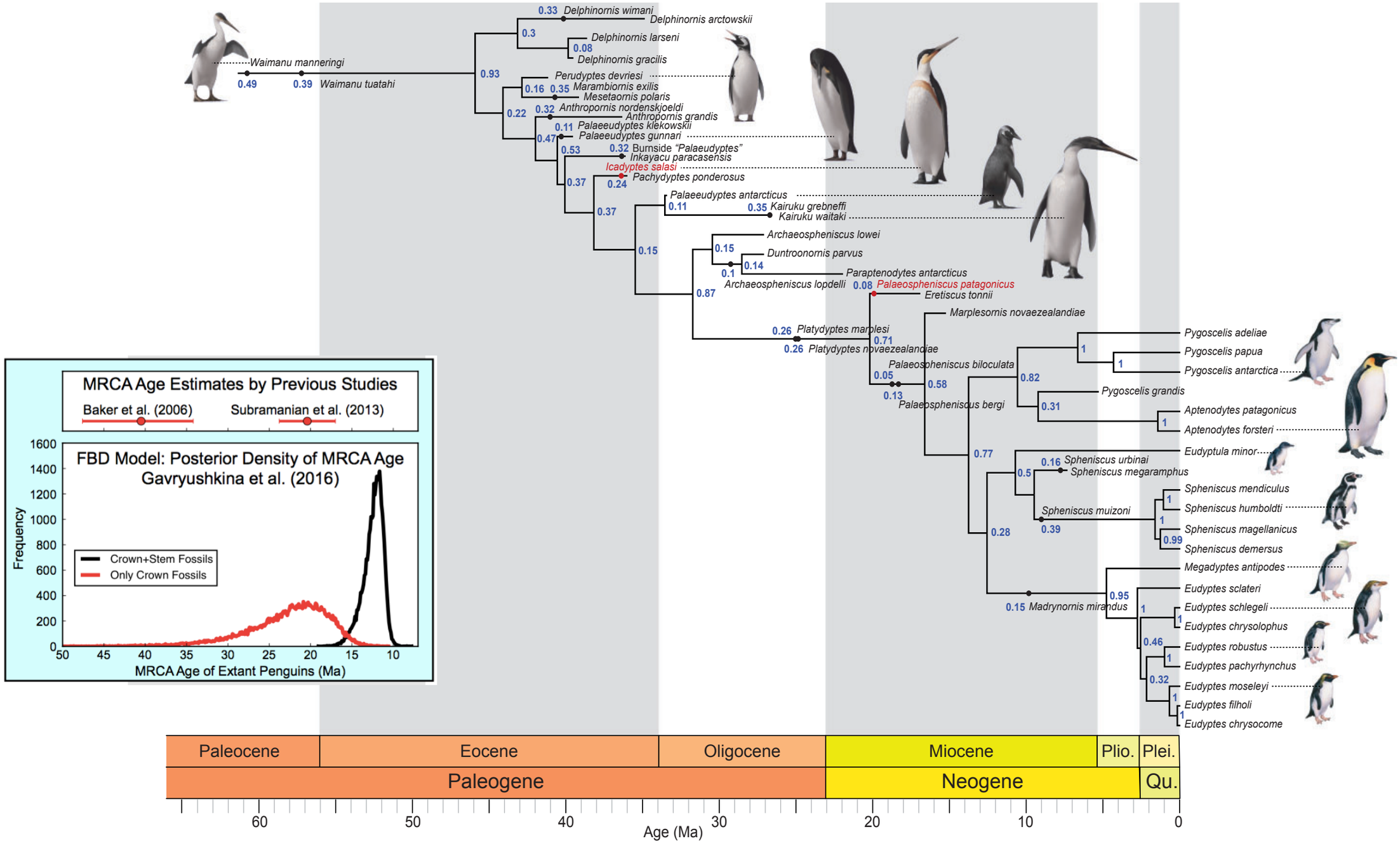
Gavryushkina et al. 2016. *Sys Bio*

Dating the origin of crown penguins



Gavryushkina et al. 2016. *Sys Bio*

Dating the origin of crown penguins



Gavryushkina et al. 2016. *Sys Bio*