

ESTIMATING SPECIES DIVERGENCE TIMES

Tracy A. Heath

University of Kansas
University of California, Berkeley
Iowa State University (in 2015)

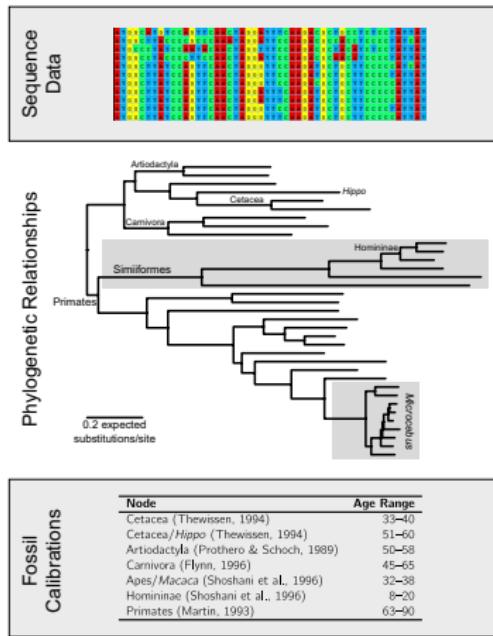
Tanja Stadler

ETH Zürich

2014 Phylogenetic Analysis in RevBayes
NESCent Academy

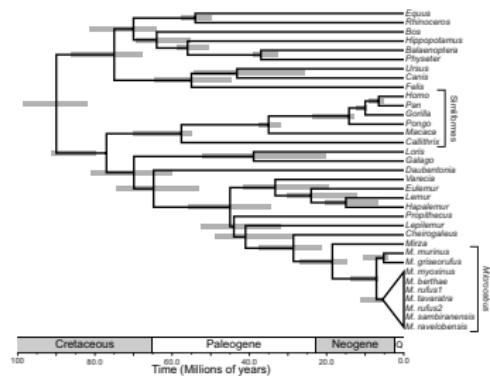
A TIME-SCALE FOR EVOLUTION

Phylogenetic trees can provide both topological information and temporal information



Did Simiiformes experience accelerated rates of molecular evolution?

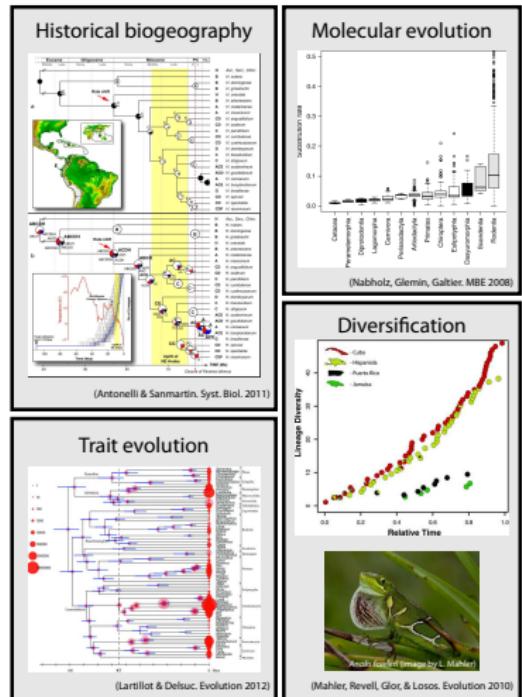
What is the age of the MRCA of mouse lemurs (*Microcebus*)?



A TIME-SCALE FOR EVOLUTION

Phylogenetic divergence-time estimation

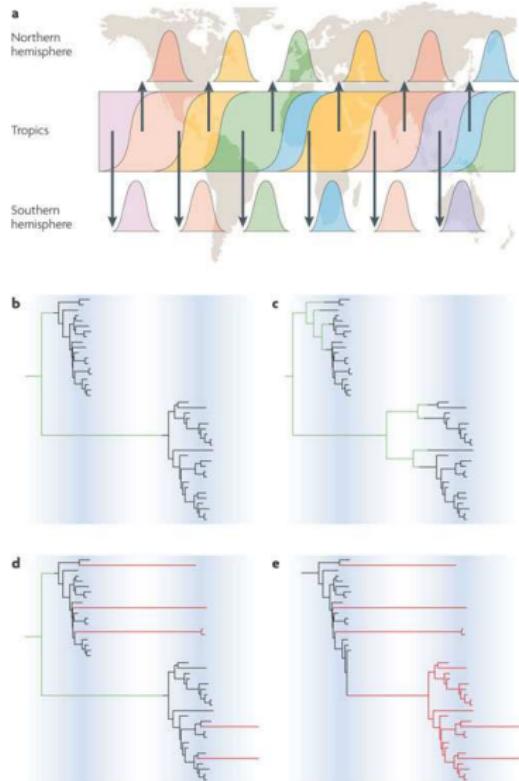
- What was the spacial and climatic environment of ancient angiosperms?
- Did the uplift of the Patagonian Andes drive the diversity of Peruvian lilies?
- How has mammalian body-size changed over time?
- Is diversification in Caribbean anoles correlated with ecological opportunity?
- How has the rate of molecular evolution changed across the Tree of Life?



THE DYNAMICS OF INFECTIOUS DISEASES

Divergence time estimation of rapidly evolving pathogens provide information about spatial and temporal dynamics of infectious diseases

Sequences sampled at different time horizons impose a temporal structure on the tree by providing ages for non-contemporaneous tips



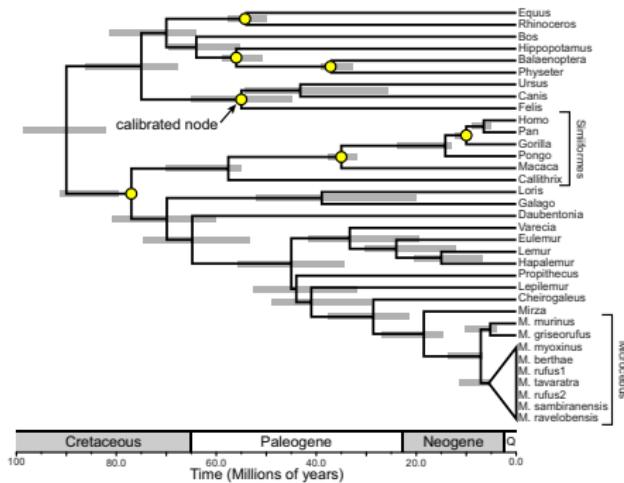
DIVERGENCE TIME ESTIMATION

Goal: Estimate the ages of interior nodes to understand the timing and rates of evolutionary processes

Model how rates are distributed across the tree

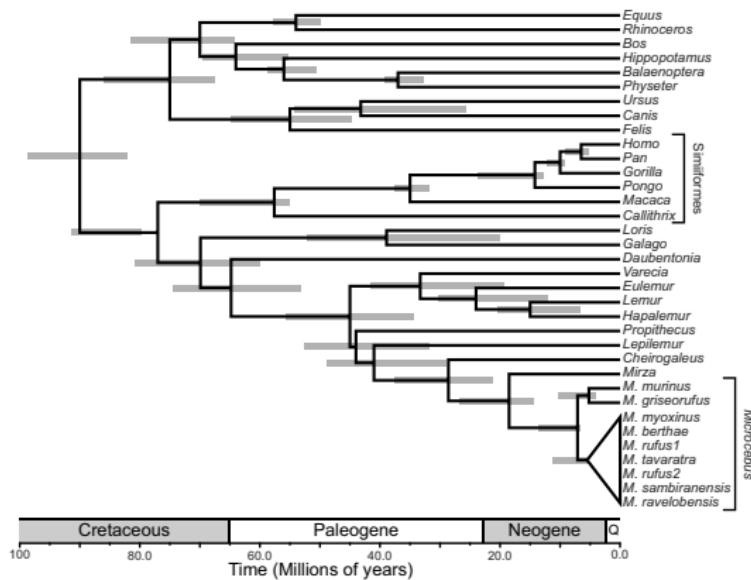
Describe the distribution of speciation events over time

External calibration information for estimates of absolute node times



A TIME-SCALE FOR EVOLUTION

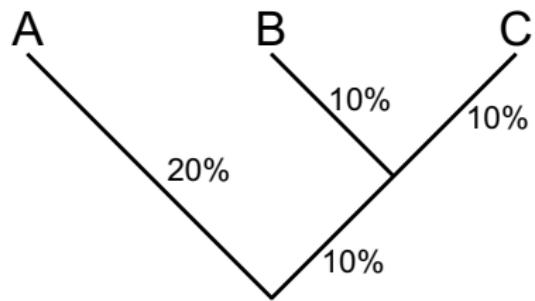
Phylogenetic trees can provide both topological information and temporal information



THE GLOBAL MOLECULAR CLOCK

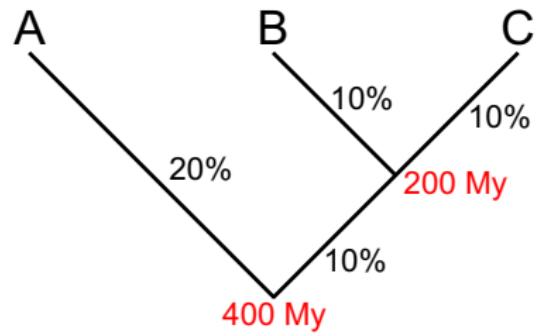
Assume that the rate of evolutionary change is constant over time

(branch lengths equal percent sequence divergence)



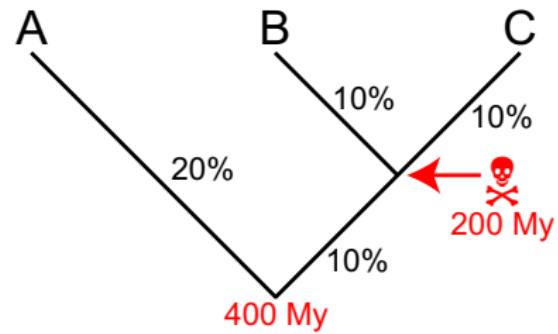
THE GLOBAL MOLECULAR CLOCK

We can date the tree if we know the rate of change is 1% divergence per 10 My



THE GLOBAL MOLECULAR CLOCK

If we found a fossil of the MRCA of **B** and **C**, we can use it to calculate the rate of change & date the root of the tree



REJECTING THE GLOBAL MOLECULAR CLOCK

Rates of evolution vary across lineages and over time

Mutation rate:

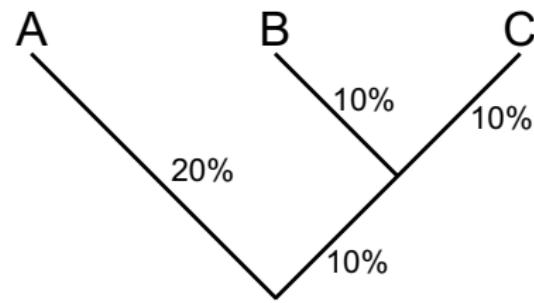
Variation in

- metabolic rate
- generation time
- DNA repair

Fixation rate:

Variation in

- strength and targets of selection
- population sizes

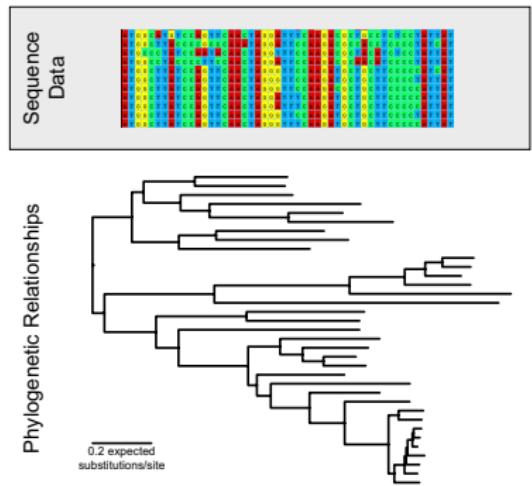


UNCONSTRAINED ANALYSIS

Sequence data provide information about **branch lengths**

In units of **the expected # of substitutions per site**

$$\text{branch length} = \text{rate} \times \text{time}$$



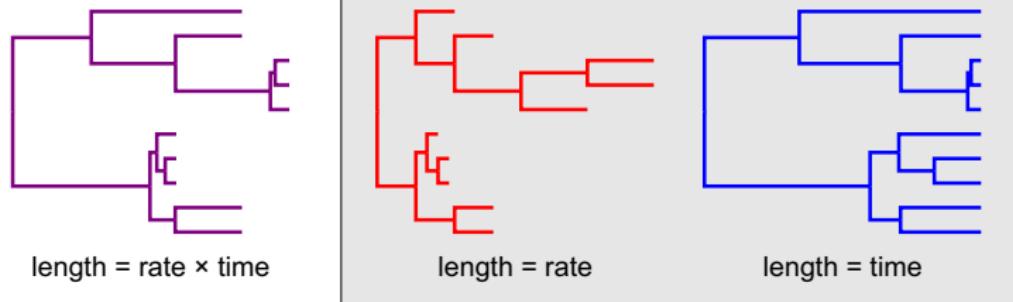
PHYLOGENETIC LIKELIHOOD

$$f(D | \mathcal{V}, \theta_s, \Psi)$$

- \mathcal{V} Vector of branch lengths
- θ_s Sequence model parameters
- D Sequence data
- Ψ Tree topology

RATE AND TIME

The **expected # of substitutions/site** occurring along a branch is the product of the **substitution rate** and **time**

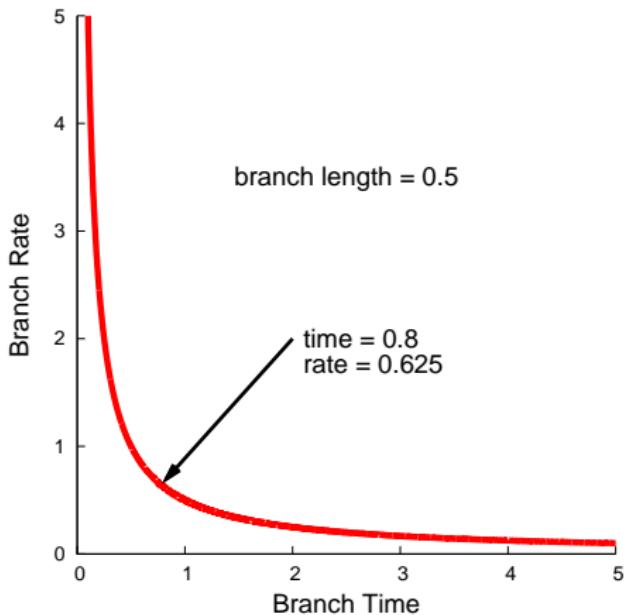


Methods for dating species divergences estimate the **substitution rate** and **time** separately

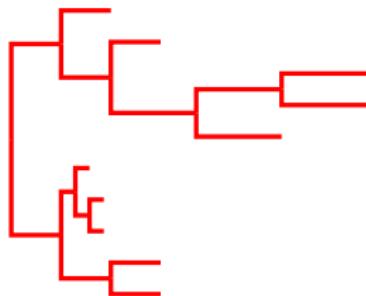
RATES AND TIMES

The sequence data provide information about branch length

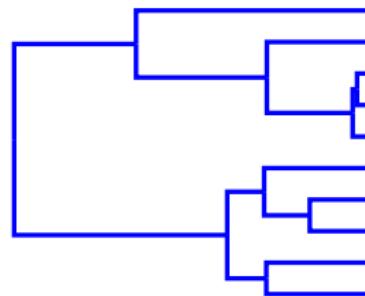
for any possible rate, there's a time that fits the branch length perfectly



BAYESIAN DIVERGENCE TIME ESTIMATION



length = rate



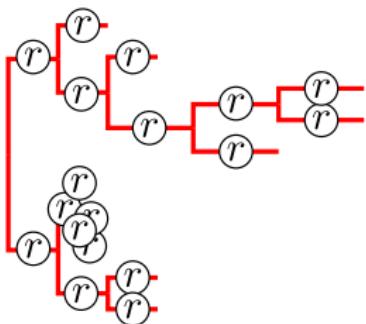
length = time

$$\mathcal{R} = (r_1, r_2, r_3, \dots, r_{2N-2})$$

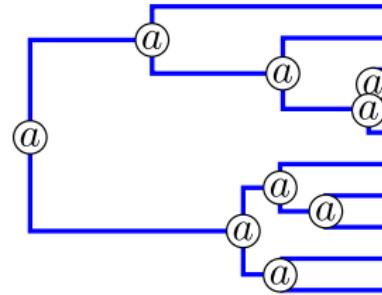
$$\mathcal{A} = (a_1, a_2, a_3, \dots, a_{N-1})$$

$$N = \text{number of tips}$$

BAYESIAN DIVERGENCE TIME ESTIMATION



length = rate



length = time

$$\mathcal{R} = (r_1, r_2, r_3, \dots, r_{2N-2})$$

$$\mathcal{A} = (a_1, a_2, a_3, \dots, a_{N-1})$$

$$N = \text{number of tips}$$

BAYESIAN DIVERGENCE TIME ESTIMATION

Posterior probability

$$f(\mathcal{R}, \mathcal{A}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s | D, \Psi)$$

\mathcal{R} Vector of rates on branches

\mathcal{A} Vector of internal node ages

$\theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s$ Model parameters

D Sequence data

Ψ Tree topology

BAYESIAN DIVERGENCE TIME ESTIMATION

$$f(\mathcal{R}, \mathcal{A}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s | D) =$$

$$\frac{f(D | \mathcal{R}, \mathcal{A}, \theta_s) \ f(\mathcal{R} | \theta_{\mathcal{R}}) \ f(\mathcal{A} | \theta_{\mathcal{A}}) \ f(\theta_s)}{f(D)}$$

$f(D \mathcal{R}, \mathcal{A}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s)$	Likelihood
$f(\mathcal{R} \theta_{\mathcal{R}})$	Prior on rates
$f(\mathcal{A} \theta_{\mathcal{A}})$	Prior on node ages
$f(\theta_s)$	Prior on substitution parameters
$f(D)$	Marginal probability of the data

BAYESIAN DIVERGENCE TIME ESTIMATION

Estimating divergence times relies on 2 main elements:

- Branch-specific rates: $f(\mathcal{R} | \theta_{\mathcal{R}})$
- Node ages: $f(\mathcal{A} | \theta_{\mathcal{A}}, \mathcal{C})$

MODELING RATE VARIATION

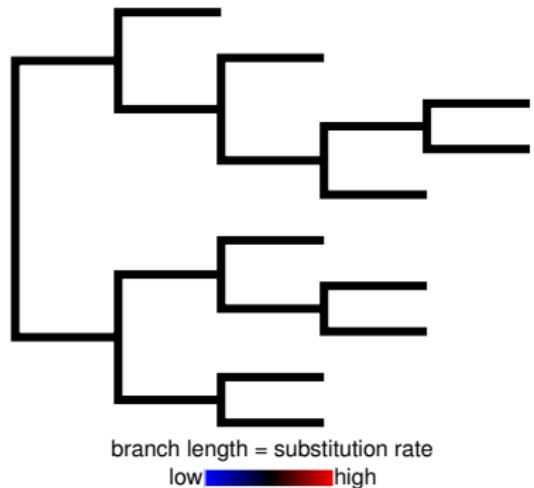
Some models describing lineage-specific substitution rate variation:

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

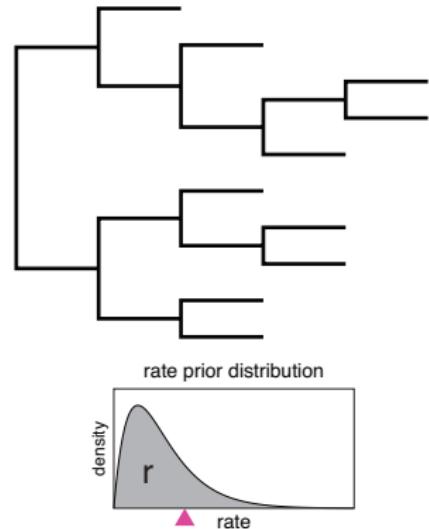
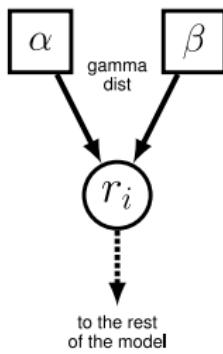
GLOBAL MOLECULAR CLOCK

The substitution rate is constant over time

All lineages share the same rate

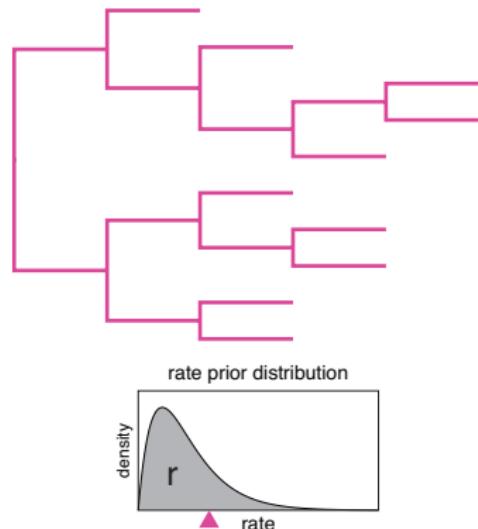


GLOBAL MOLECULAR CLOCK



GLOBAL MOLECULAR CLOCK

The sampled rate is applied
to every branch in the tree



REJECTING THE GLOBAL MOLECULAR CLOCK

Rates of evolution vary across lineages and over time

Mutation rate:

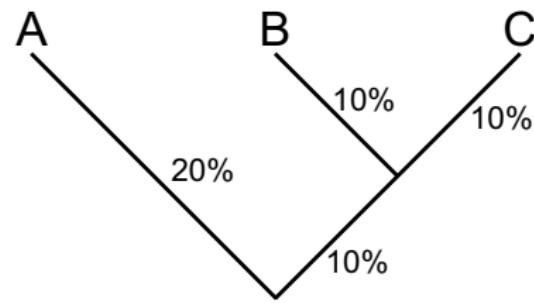
Variation in

- metabolic rate
- generation time
- DNA repair

Fixation rate:

Variation in

- strength and targets of selection
- population sizes



RELAXED-CLOCK MODELS

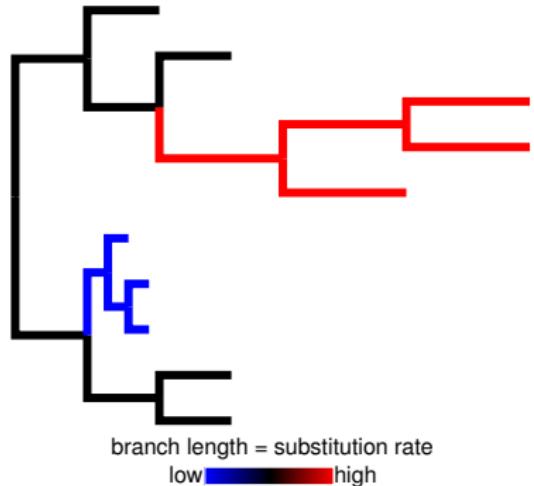
To accommodate variation in substitution rates
'relaxed-clock' models estimate lineage-specific substitution rates

- Local molecular clocks
- Punctuated rate change model
- Log-normally distributed autocorrelated rates
- Uncorrelated/independent rates models
- Mixture models on branch rates

LOCAL MOLECULAR CLOCKS

Rate shifts occur infrequently over the tree

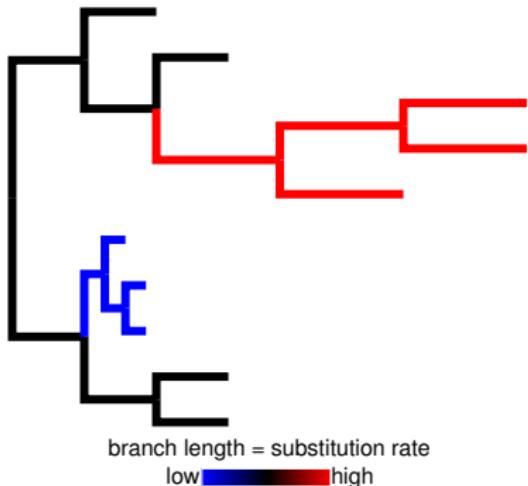
Closely related lineages have equivalent rates (clustered by sub-clades)



LOCAL MOLECULAR CLOCKS

Most methods for estimating local clocks required specifying the number and locations of rate changes *a priori*

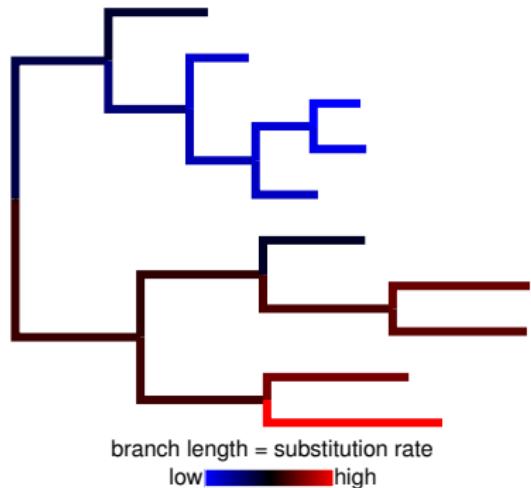
Drummond and Suchard (2010) introduced a Bayesian method that samples over a broad range of possible *random local clocks*



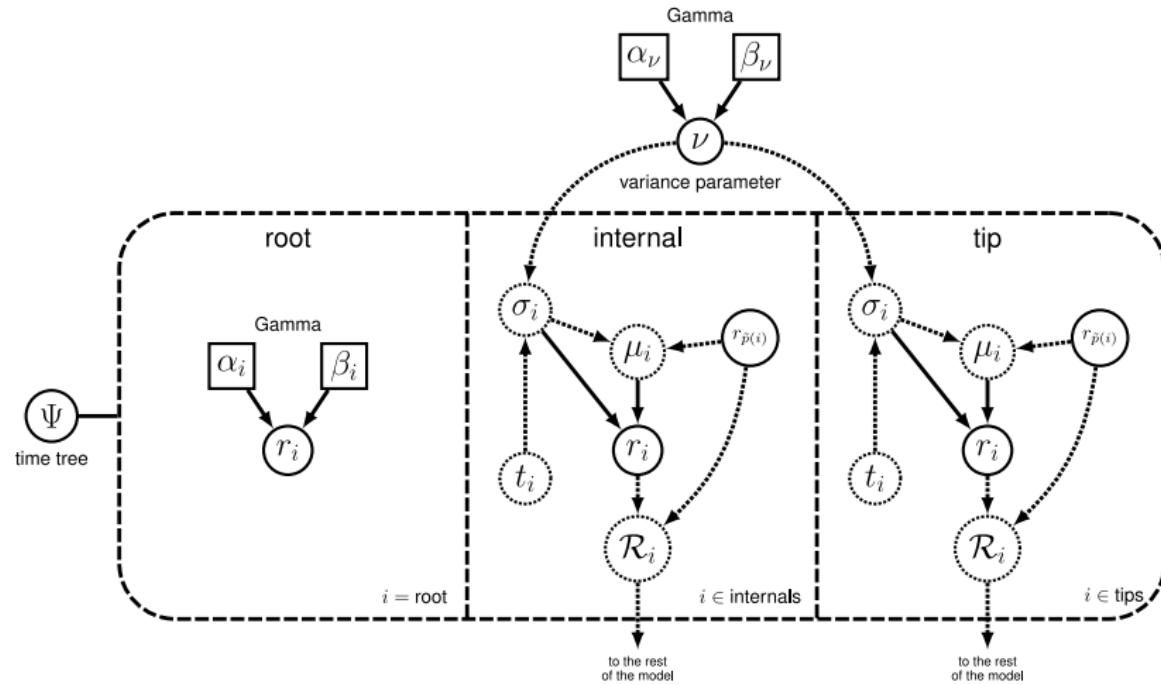
AUTOCORRELATED RATES

Substitution rates evolve gradually over time – closely related lineages have similar rates

The rate at a node is drawn from a lognormal distribution with a mean equal to the parent rate



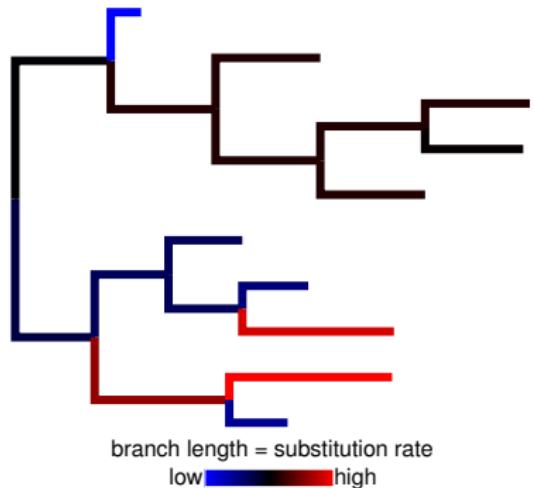
AUTOCORRELATED RATES



PUNCTUATED RATE CHANGE

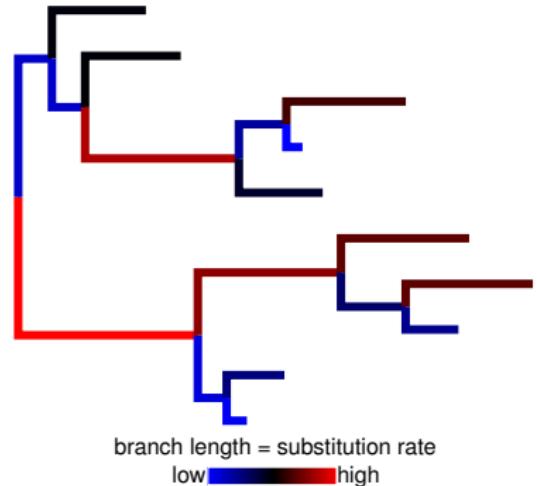
Rate changes occur along lineages according to a point process

At rate-change events, the new rate is a product of the parent's rate and a Γ -distributed multiplier

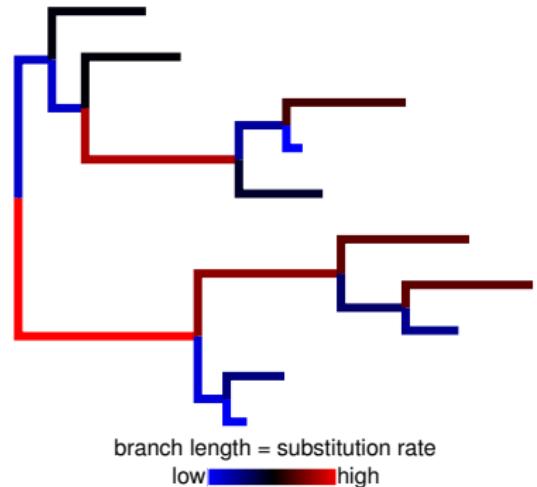
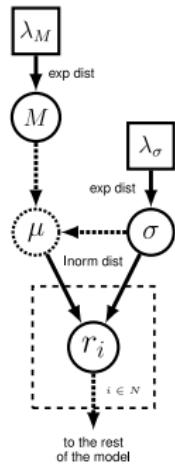


INDEPENDENT/UNCORRELATED RATES

Lineage-specific rates are uncorrelated when the rate assigned to each branch is independently drawn from an underlying distribution



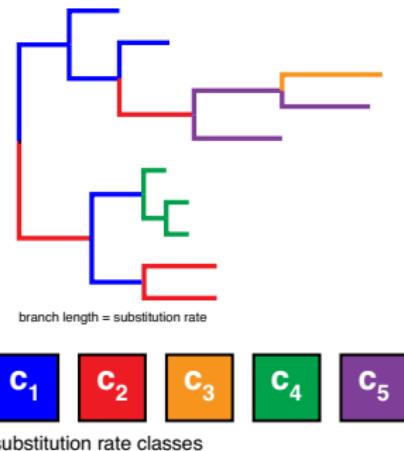
INDEPENDENT/UNCORRELATED RATES



INFINITE MIXTURE MODEL

Dirichlet process prior:

Branches are partitioned
into distinct rate categories

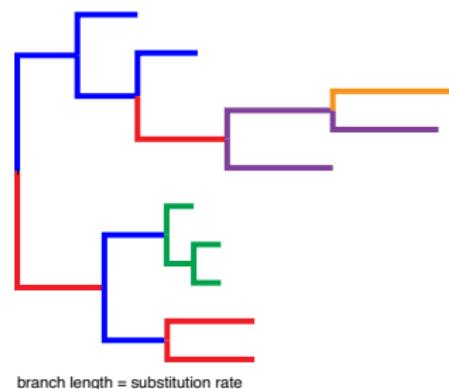


THE DIRICHLET PROCESS PRIOR (DPP)

A stochastic process that models data as a mixture of distributions and can identify latent classes present in the data

Branches are assumed to form distinct substitution rate clusters

Efficient Markov chain Monte Carlo (MCMC) implementations allow for inference under this model

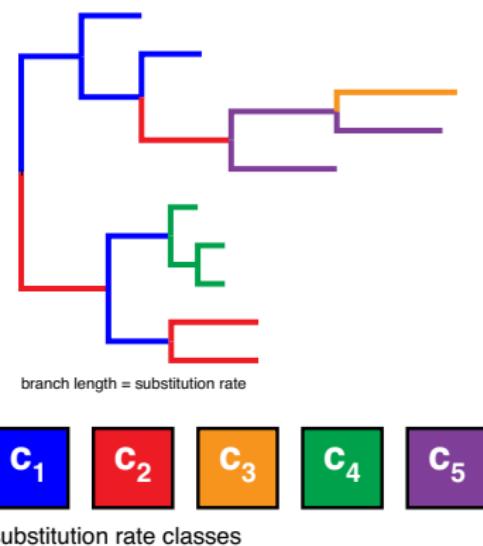


THE DIRICHLET PROCESS PRIOR (DPP)

A stochastic process that models data as a mixture of distributions and can identify latent classes present in the data

Random variables under the DPP informed by the **data**:

- the number of rate classes
- the assignment of branches to classes
- the rate value for each class

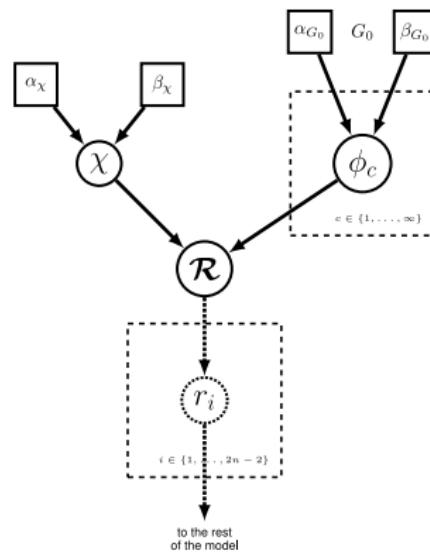


THE DIRICHLET PROCESS PRIOR (DPP)

A stochastic process that models data as a mixture of distributions and can identify latent classes present in the data

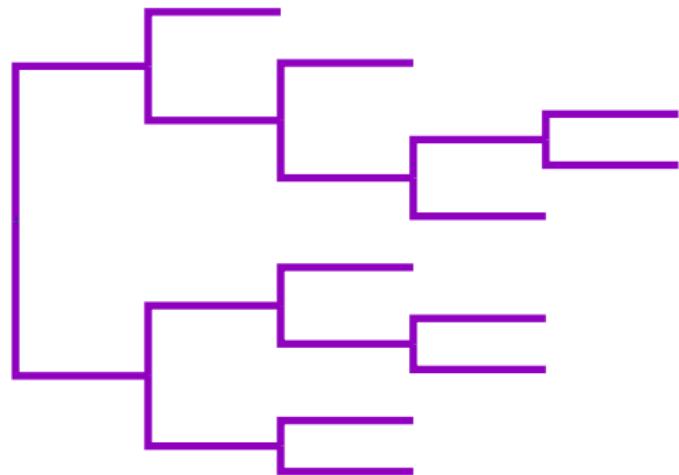
Random variables under the DPP informed by the **data**:

- the number of rate classes
- the assignment of branches to classes
- the rate value for each class



DPP FOR CLUSTERING PROBLEMS

branch length = substitution rate



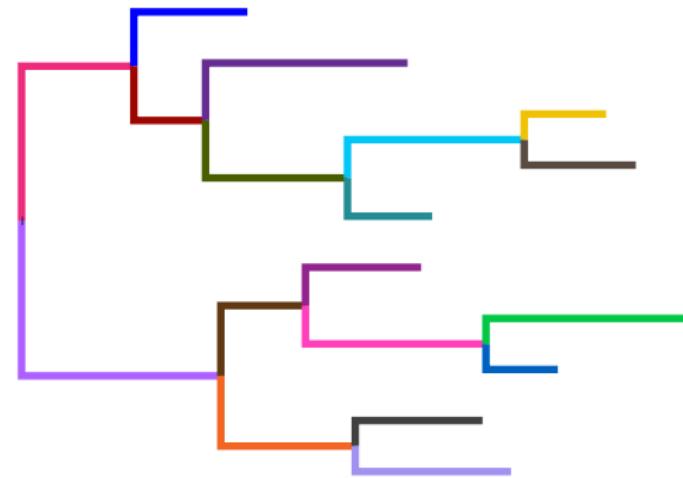
18

c₁

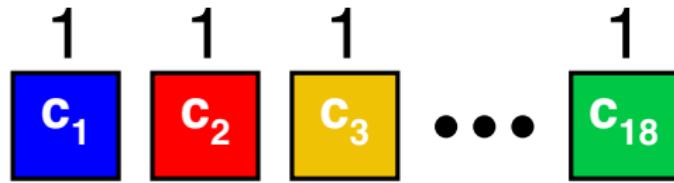
rate classes

DPP FOR CLUSTERING PROBLEMS

branch length = substitution rate



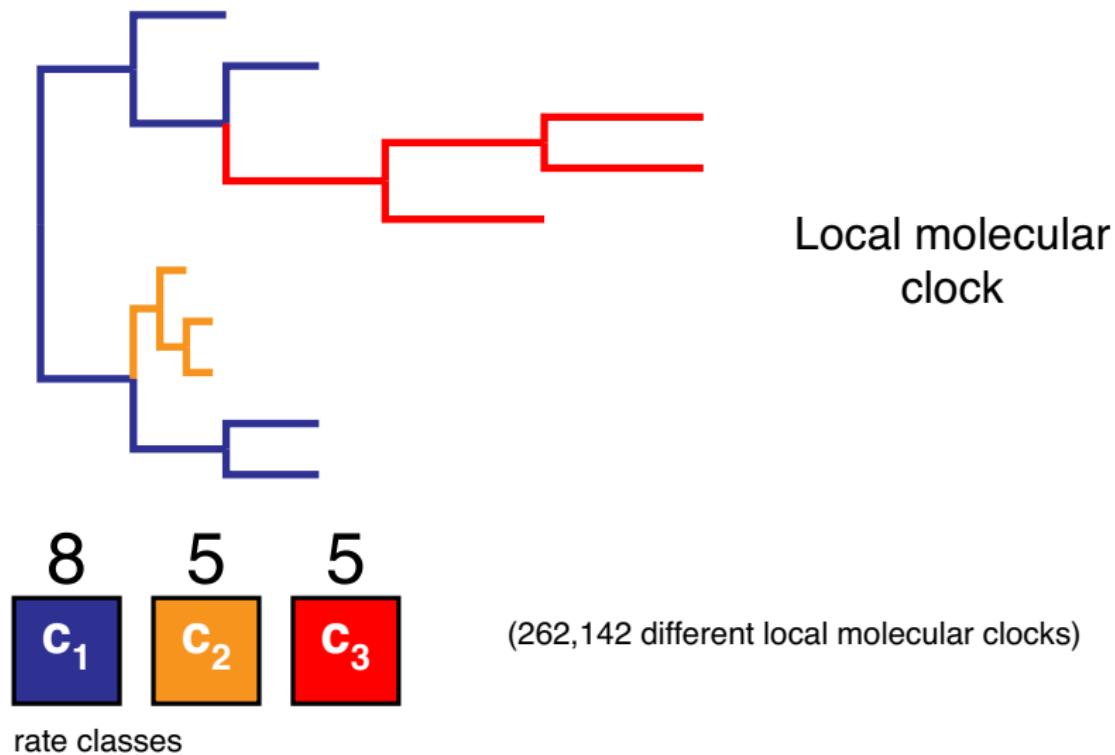
Independent
rates



rate classes

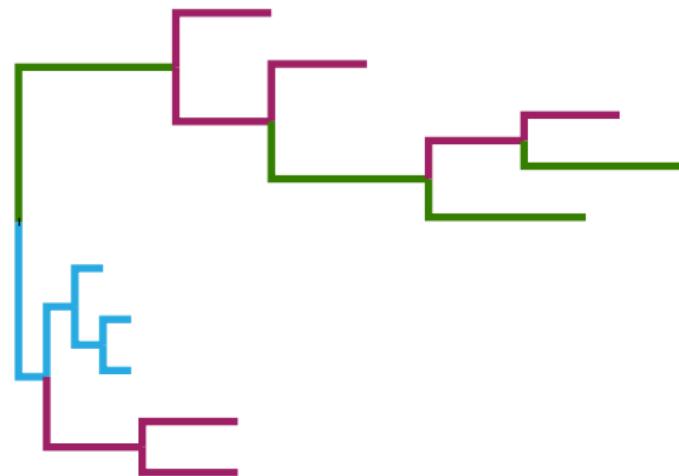
DPP FOR CLUSTERING PROBLEMS

branch length = substitution rate

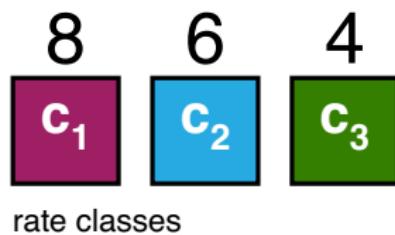


DPP FOR CLUSTERING PROBLEMS

branch length = substitution rate



Each of the
682,076,806,159
configurations
has a prior weight



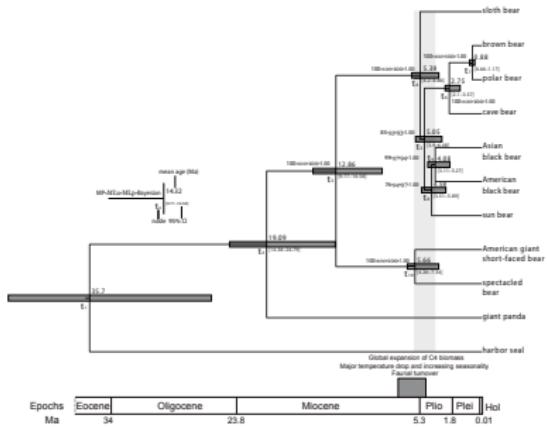
MODELING RATE VARIATION

These are only a subset of the available models for branch-rate variation

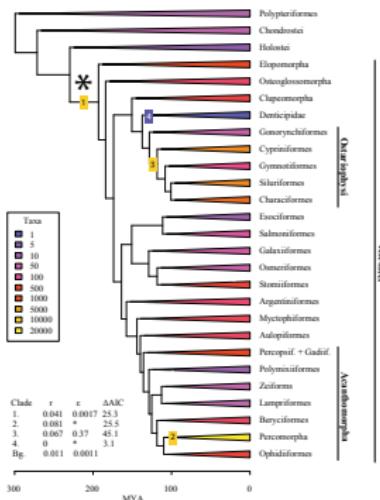
- Global molecular clock
- Local molecular clocks
- Punctuated rate change model
- Log-normally distributed autocorrelated rates
- Uncorrelated/independent rates models
- Dirichlet process prior

MODELING RATE VARIATION

Are our models appropriate across all data sets?



Krause et al., 2008. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evol. Biol.* 8.



Santini et al., 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evol. Biol.* 9.

MODELING RATE VARIATION

These are only a subset of the available models for branch-rate variation

- Global molecular clock
- Local molecular clocks
- Punctuated rate change model
- Log-normally distributed autocorrelated rates
- Uncorrelated/independent rates models
- Dirichlet process prior

Model selection and model uncertainty are **very** important for Bayesian divergence time analysis



BAYESIAN DIVERGENCE TIME ESTIMATION

Estimating divergence times relies on 2 main elements:

- Branch-specific rates: $f(\mathcal{R} | \theta_{\mathcal{R}})$
- Node ages: $f(\mathcal{A} | \theta_{\mathcal{A}}, \mathcal{C})$

<http://bayesiancook.blogspot.com/2013/12/two-sides-of-same-coin.html>

PRIORS ON NODE TIMES

Sequence data are only informative on *relative* rates & times

Node-time priors cannot give precise estimates of *absolute* node ages

We need external information (like fossils) to *calibrate* or scale the tree to absolute time

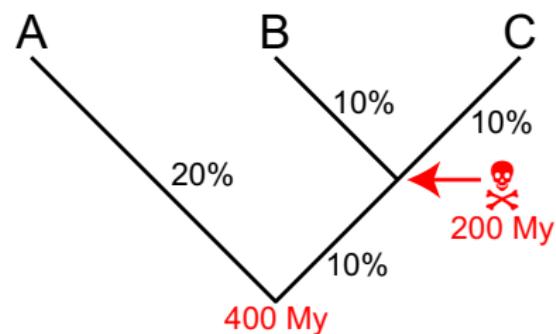


CALIBRATING DIVERGENCE TIMES

Fossils (or other data) are necessary to estimate *absolute* node ages

There is **no information** in the sequence data for absolute time

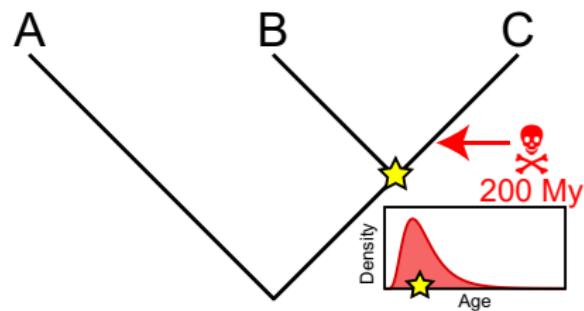
Uncertainty in the placement of fossils



CALIBRATION DENSITIES

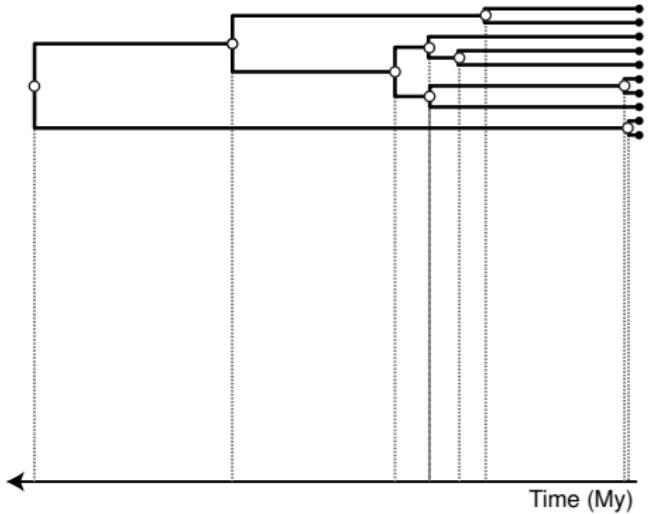
Bayesian inference is well suited to accommodating uncertainty in the age of the calibration node

Divergence times are calibrated by placing parametric densities on internal nodes offset by age estimates from the fossil record



Fossil Calibration

Fossil and geological data can be used to estimate the absolute ages of ancient divergences



Fossil Calibration



The ages of extant taxa
are known

Time (My)

Fossil Calibration

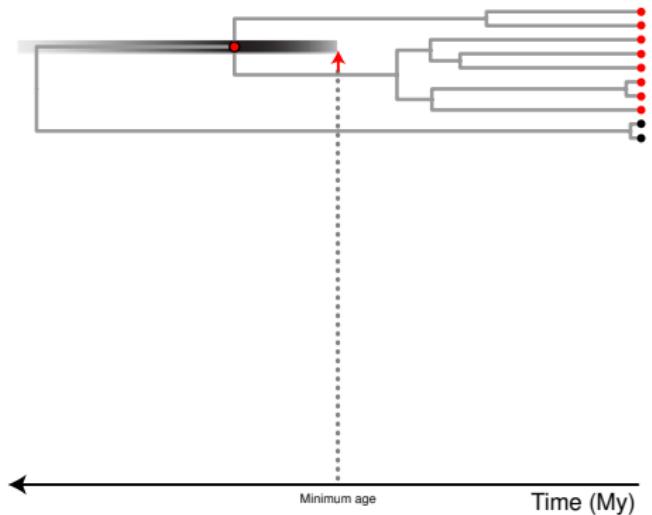


Fossil taxa are assigned to monophyletic clades



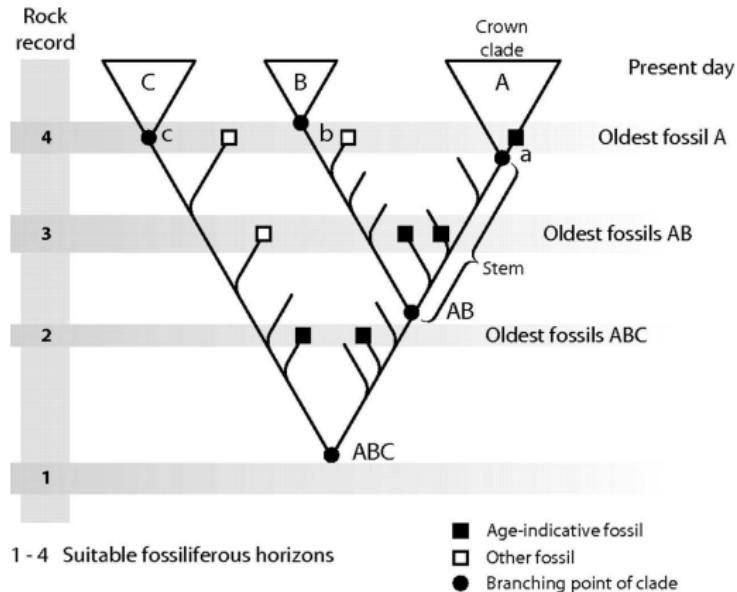
Fossil Calibration

Fossil taxa are assigned to monophyletic clades and constrain the age of the MRCA



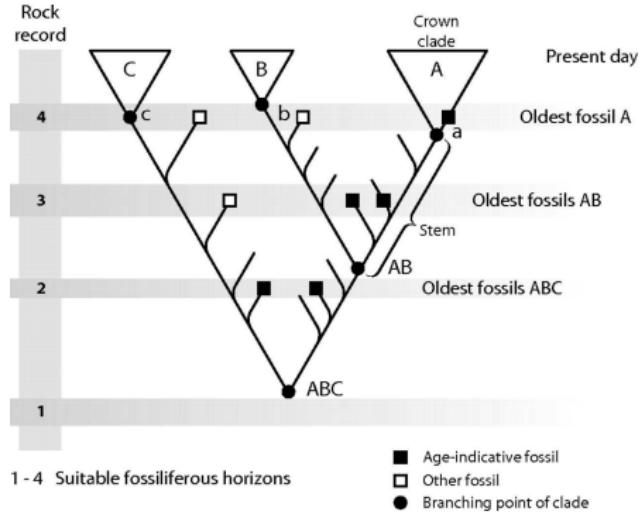
ASSIGNING FOSSILS TO CLADES

Misplaced fossils can affect node age estimates throughout the tree – if the fossil is older than its presumed MRCA



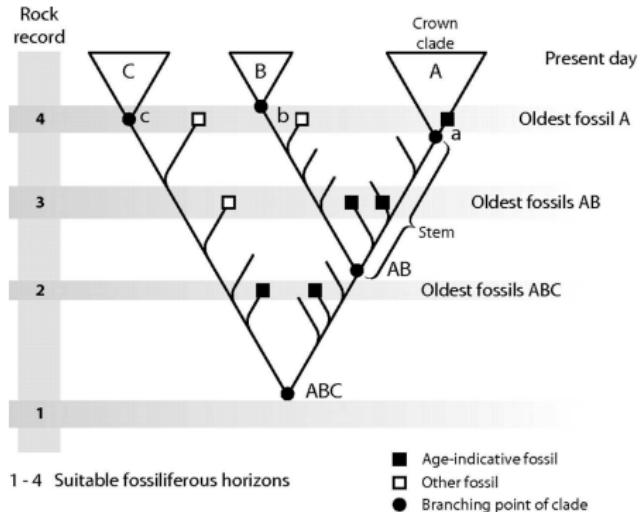
ASSIGNING FOSSILS TO CLADES

Crown clade: all living species and their most-recent common ancestor (MRCA)



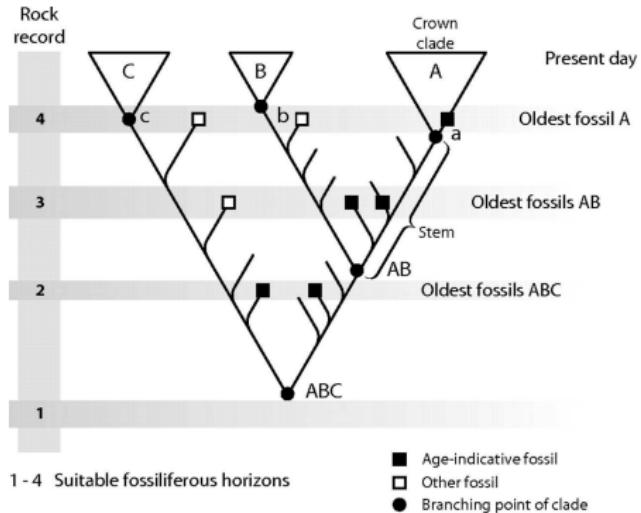
ASSIGNING FOSSILS TO CLADES

Stem lineages:
purely fossil forms
that are closer to
their descendant
crown clade than
any other crown
clade



ASSIGNING FOSSILS TO CLADES

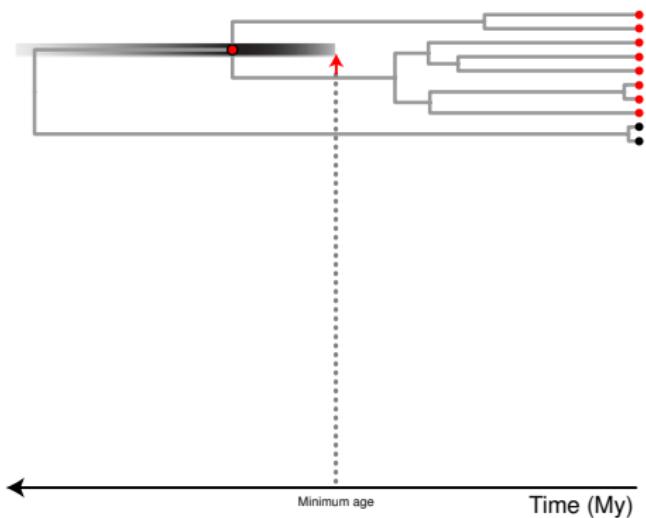
Fossiliferous horizons: the sources in the rock record for relevant fossils



Fossil Calibration

Age estimates from fossils can provide **minimum** time constraints for internal nodes

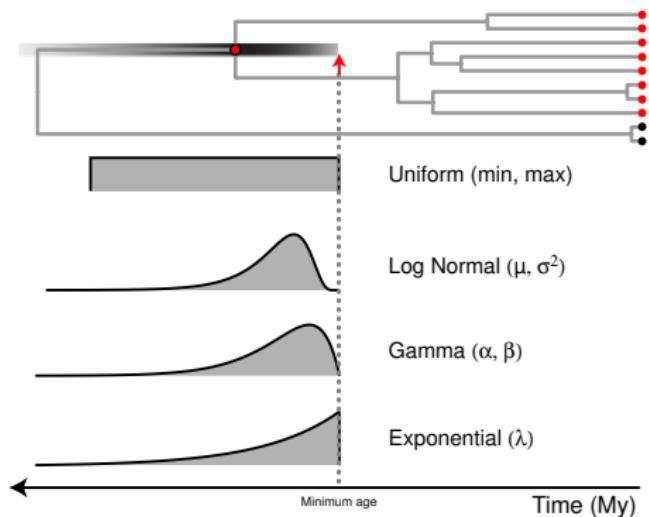
Reliable **maximum** bounds are typically unavailable



PRIOR DENSITIES ON CALIBRATED NODES

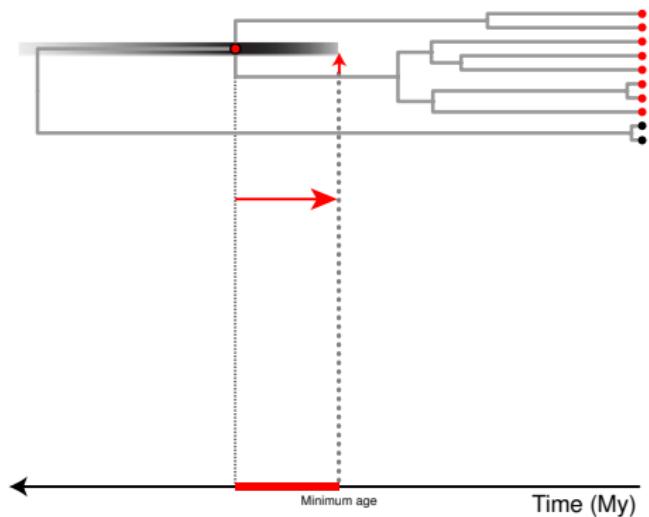
Parametric distributions are typically off-set by the age of the oldest fossil assigned to a clade

These prior densities do not (necessarily) require specification of maximum bounds



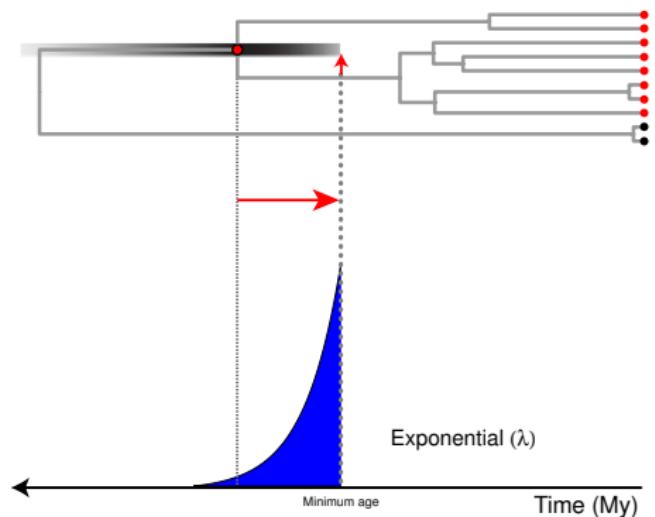
PRIOR DENSITIES ON CALIBRATED NODES

Describe the waiting time
between the divergence
event and the age of the
oldest fossil



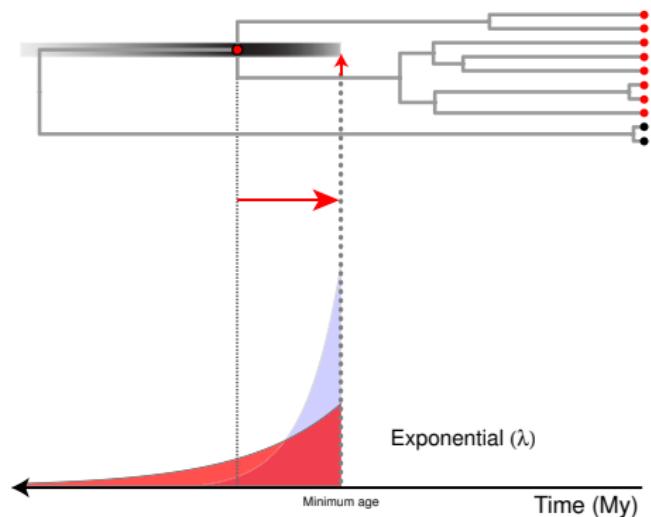
PRIOR DENSITIES ON CALIBRATED NODES

Overly **informative** priors
can bias node age
estimates to be too young



PRIOR DENSITIES ON CALIBRATED NODES

Uncertainty in the age of the MRCA of the clade relative to the age of the fossil may be better captured by **vague** prior densities

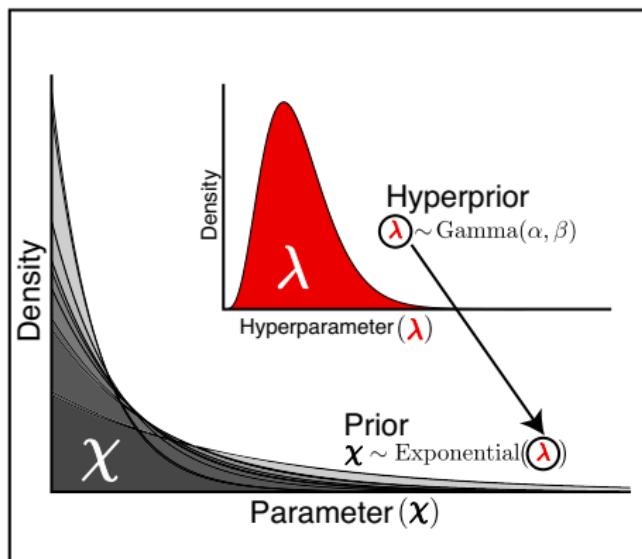


PUT A PRIOR ON IT

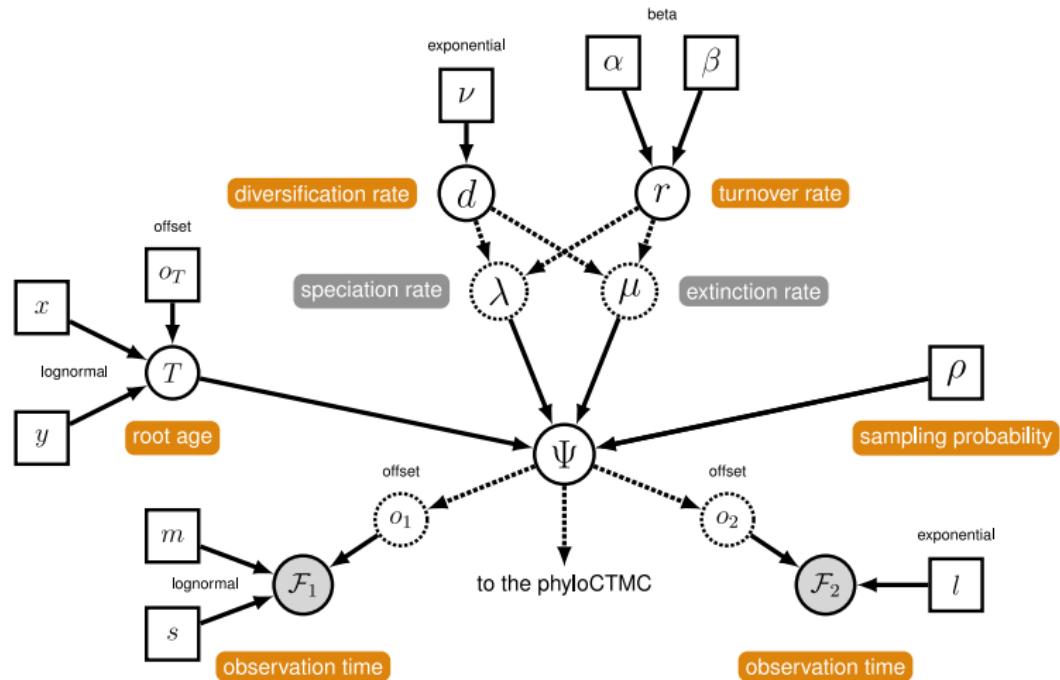
Hyperprior: place an higher-order prior on the parameter of a prior distribution

Sample the time from the MRCA to the fossil from a mixture of different exponential distributions

Account for uncertainty in values of λ



CALIBRATION PRIORS IN RevBAYES

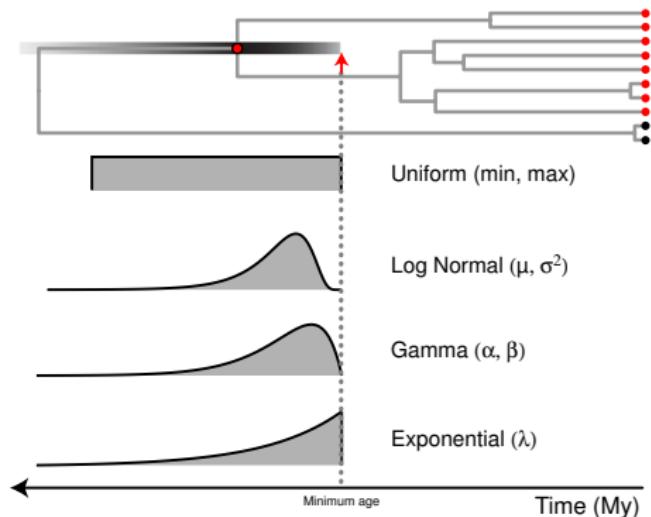


PRIOR DENSITIES ON CALIBRATED NODES

Common practice in Bayesian divergence-time estimation:

Estimates of absolute node ages are driven primarily by the calibration density

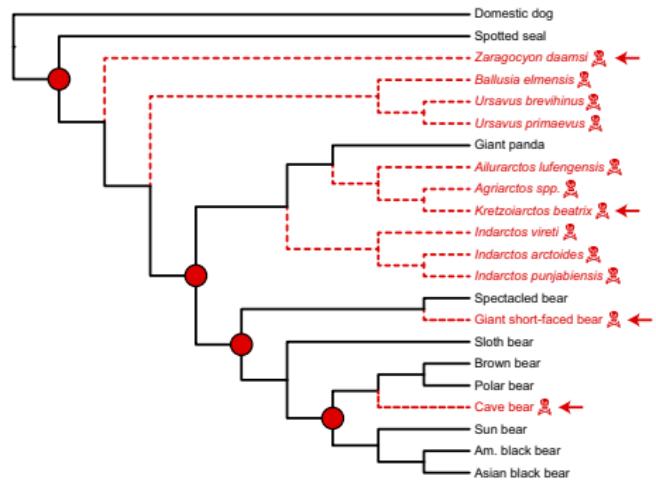
Specifying appropriate densities is a challenge for most molecular biologists



IMPROVING FOSSIL CALIBRATION

We would prefer to eliminate the need for *ad hoc* calibration prior densities

Calibration densities do not account for diversification of fossils

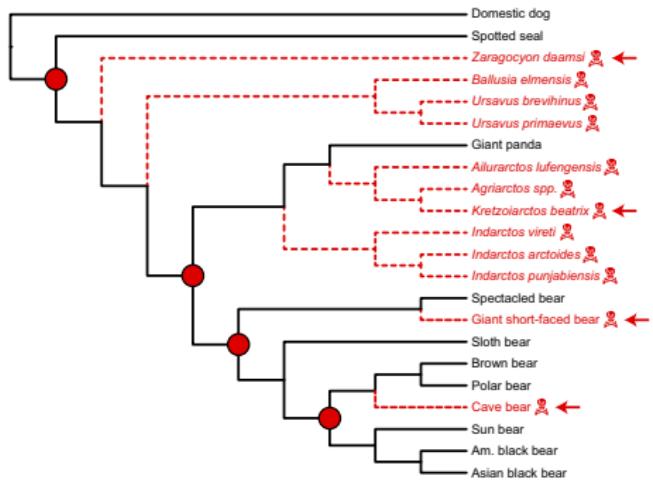


IMPROVING FOSSIL CALIBRATION

We want to use all of the available fossils

Example: Bears

12 fossils are reduced to 4 calibration ages with calibration density methods

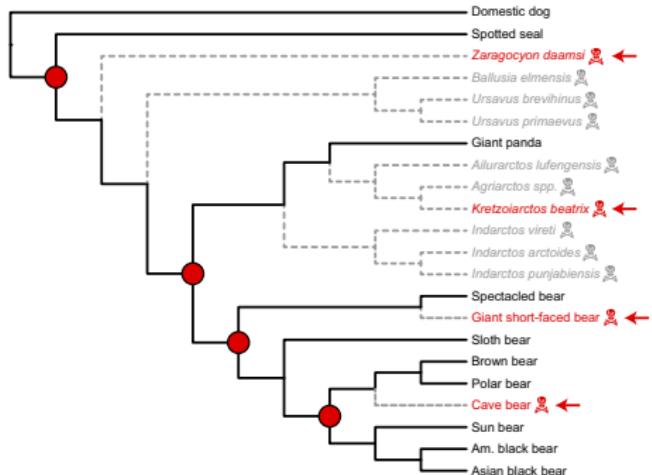


IMPROVING FOSSIL CALIBRATION

We want to use all of the available fossils

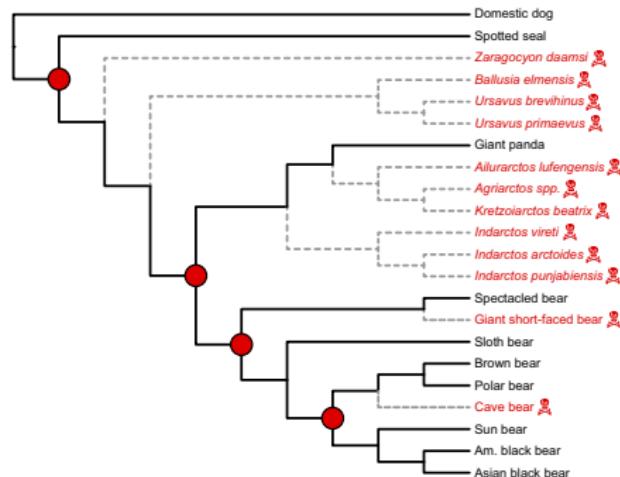
Example: Bears

12 fossils are reduced to 4 calibration ages with calibration density methods



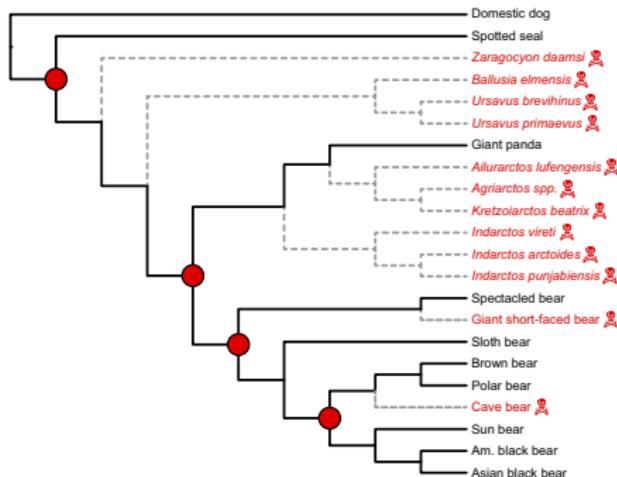
IMPROVING FOSSIL CALIBRATION

Because fossils are part of the diversification process, we can combine fossil calibration with birth-death models



IMPROVING FOSSIL CALIBRATION

This relies on a branching model that accounts for **speciation, extinction, and rates of fossilization, preservation, and recovery**



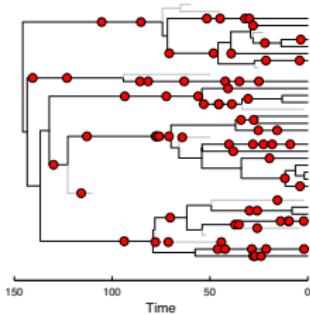
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Improving statistical inference of absolute node ages

Eliminates the need to specify arbitrary calibration densities

Better capture our statistical uncertainty in species divergence dates

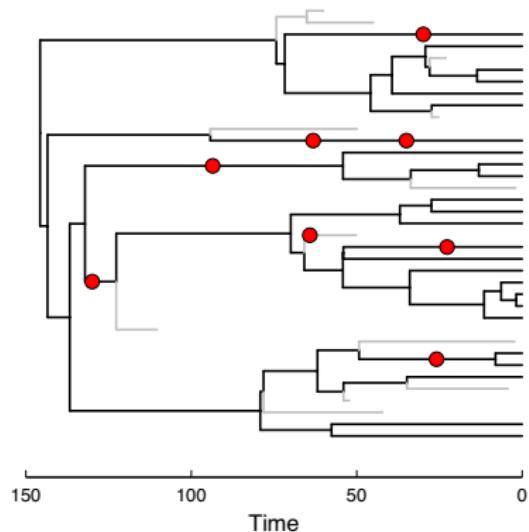
All reliable fossils associated with a clade are used



Heath, Huelsenbeck, & Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence time estimates. *PNAS* 111(29):E2957–E2966.
<http://www.pnas.org/content/111/29/E2957.abstract>

THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Recovered fossil specimens provide historical observations of the diversification process that generated the tree of extant species



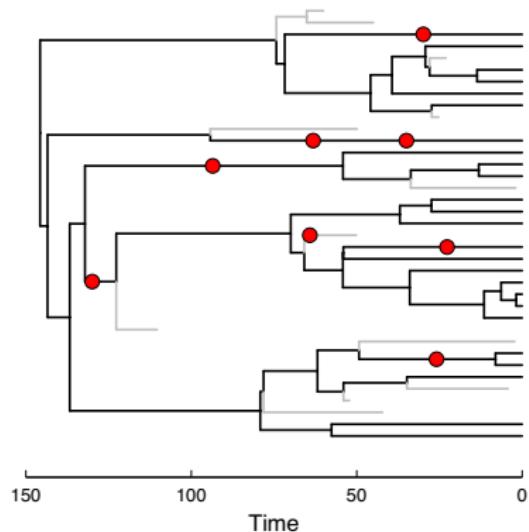
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

The probability of the tree and fossil observations under a birth-death model with rate parameters:

λ = speciation

μ = extinction

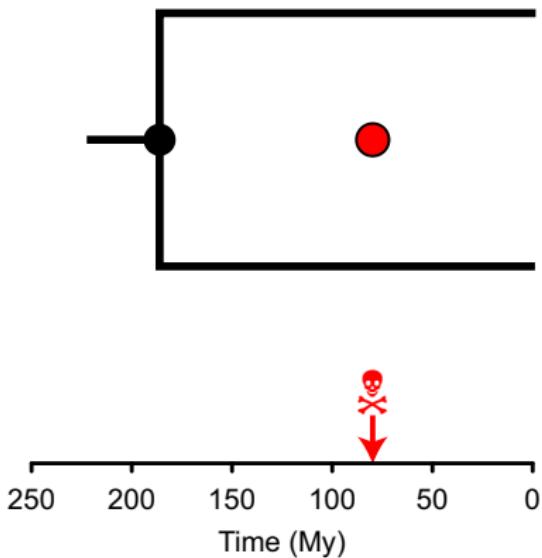
ψ = fossilization/recovery



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

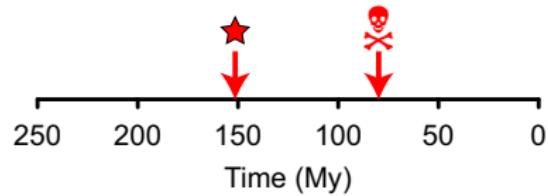
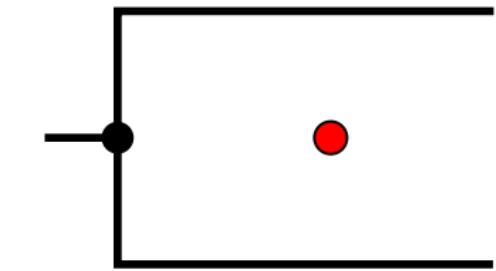
We assume that the fossil is a descendant of a specified calibrated node

The time of the fossil:  indicates an observation of the birth-death process after the age of the node



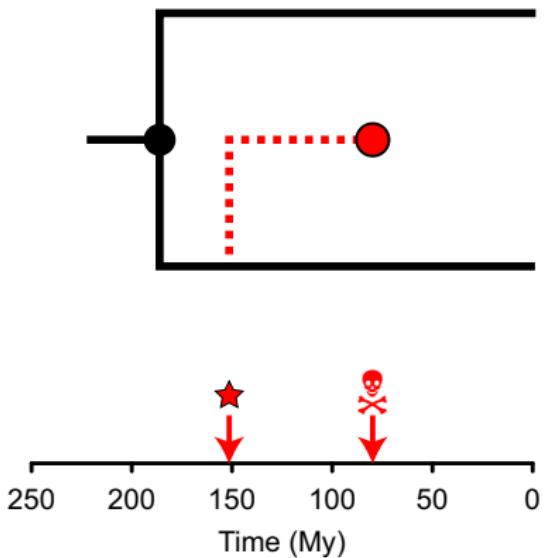
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

The fossil must attach to the tree at some time: ★



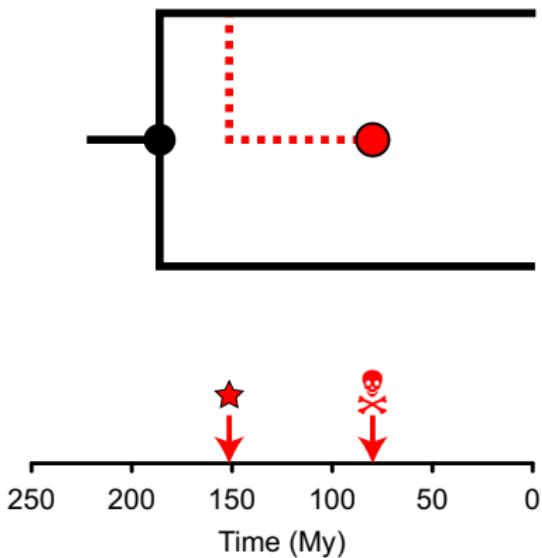
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

If it is the descendant of an unobserved lineage, then there is a speciation event at time \star on one of the 2 branches



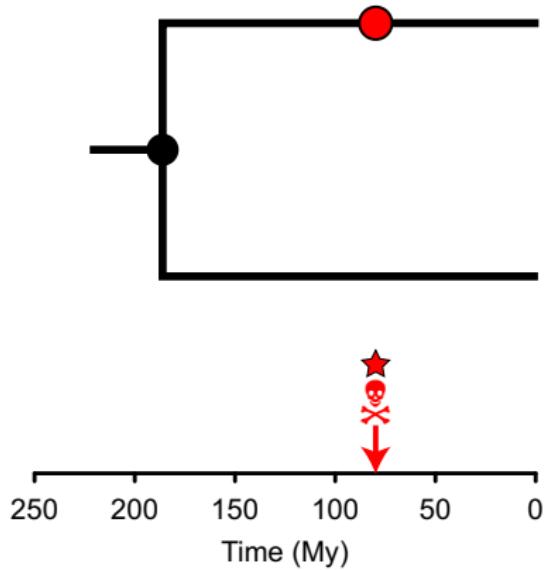
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

If it is the descendant of an unobserved lineage, then there is a speciation event at time \star on one of the 2 branches



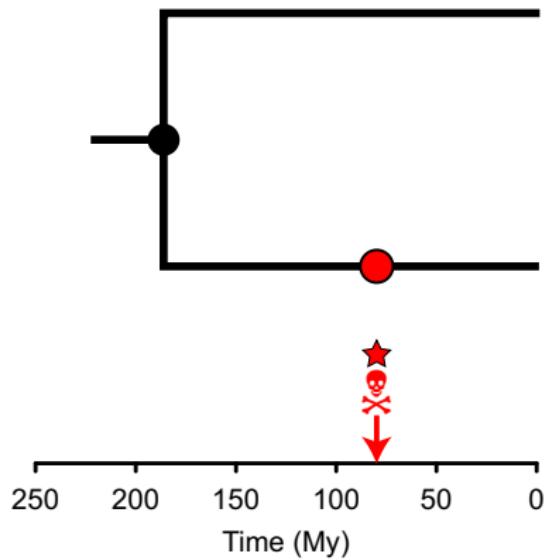
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

If $\star = \text{💀}$, the fossil is an observation of a lineage ancestral to the extant species



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

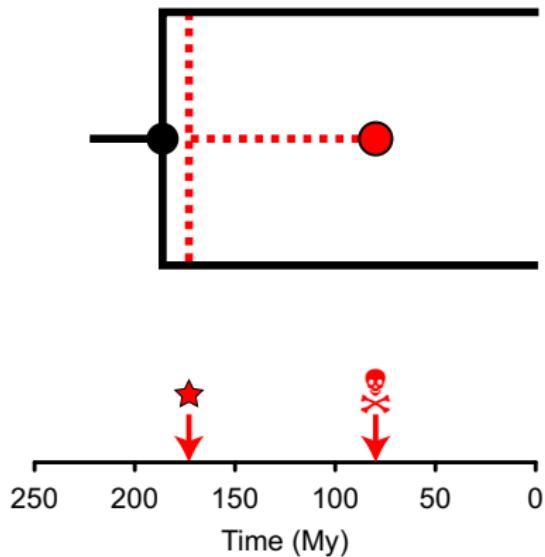
If $\star = \text{💀}$, the fossil is an observation of a lineage ancestral to the extant species



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

The probability of this realization of the diversification process is conditional on:

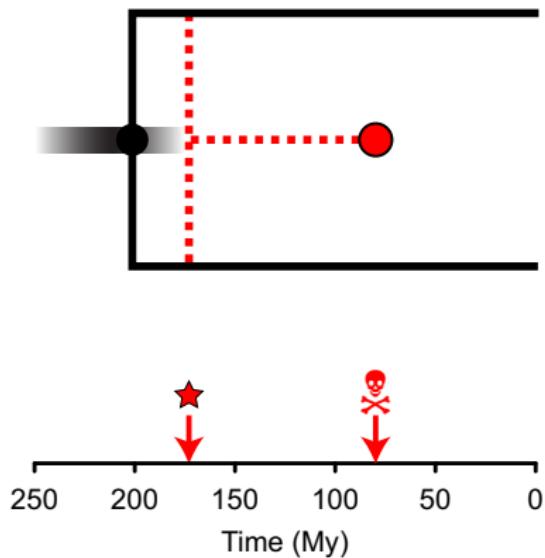
λ , μ , and ψ



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

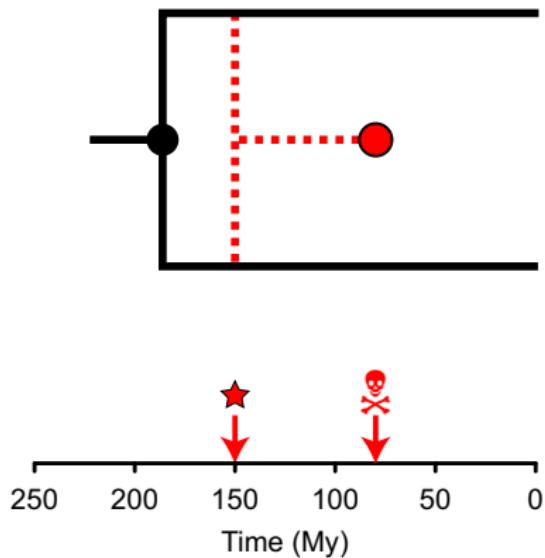
Using MCMC, we can sample the age of the calibrated node ● while conditioning on

λ , μ , and ψ
other node ages
💀 and ★



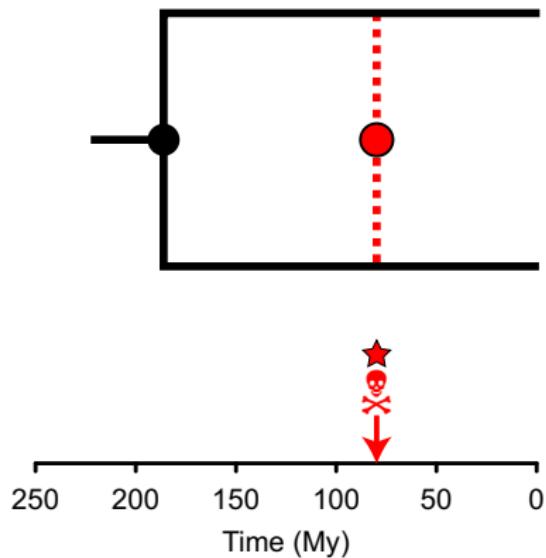
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

MCMC allows us to consider all possible values of \star (marginalization)



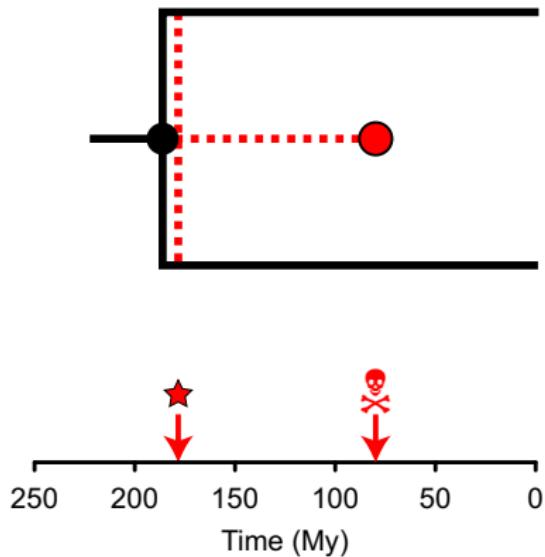
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

MCMC allows us to consider all possible values of  (marginalization)



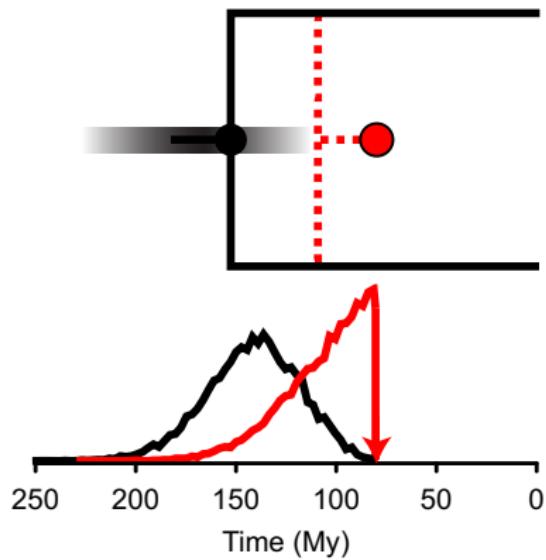
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

MCMC allows us to consider all possible values of \star (marginalization)



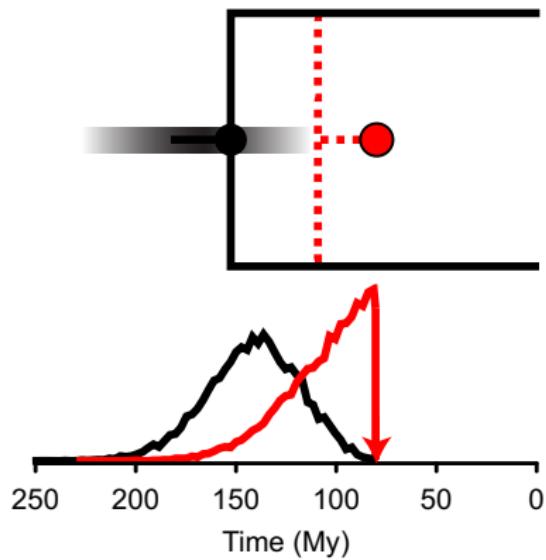
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

MCMC allows us to
consider all possible values
of \star (marginalization)



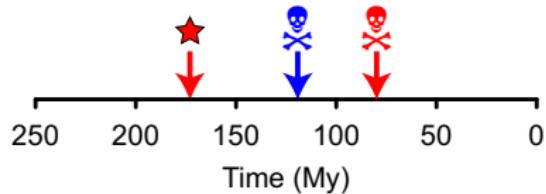
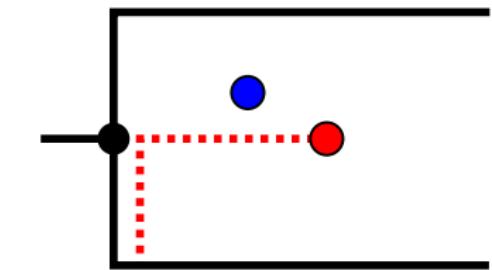
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

The posterior samples of the calibrated node age are informed by the fossil attachment times



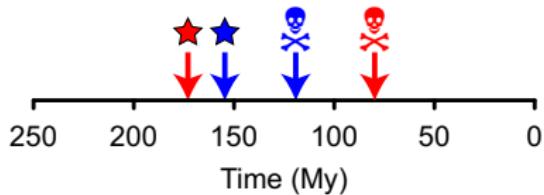
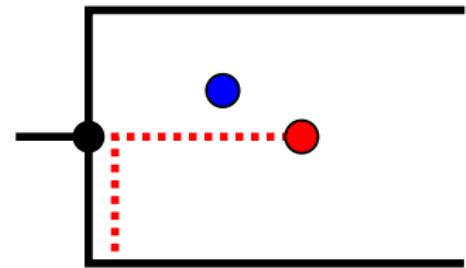
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

The **FBD** model allows multiple fossils to calibrate a single node



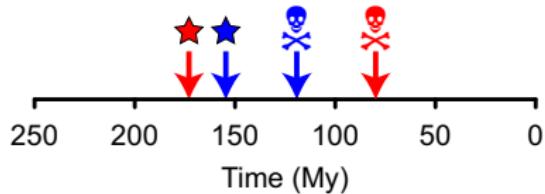
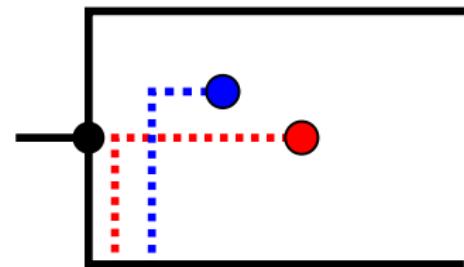
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Given  and , the new fossil can attach to the tree via speciation along either branch in the extant tree



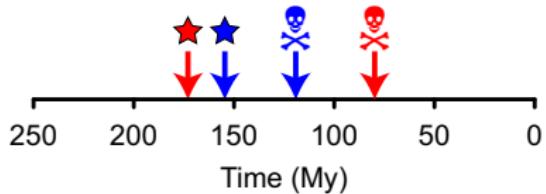
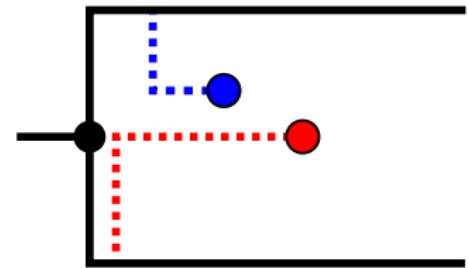
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Given  and , the new fossil can attach to the tree via speciation along either branch in the extant tree



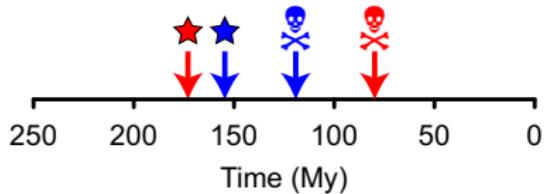
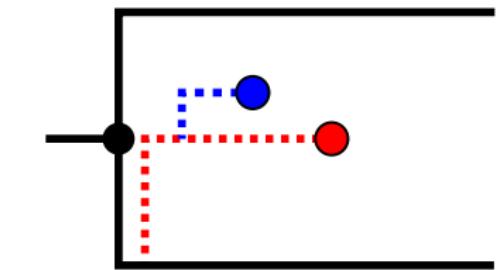
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Given  and , the new fossil can attach to the tree via speciation along either branch in the extant tree



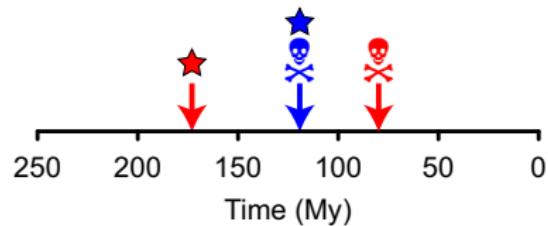
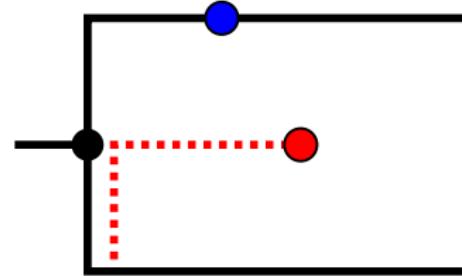
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Or the unobserved branch leading to the other calibrating fossil



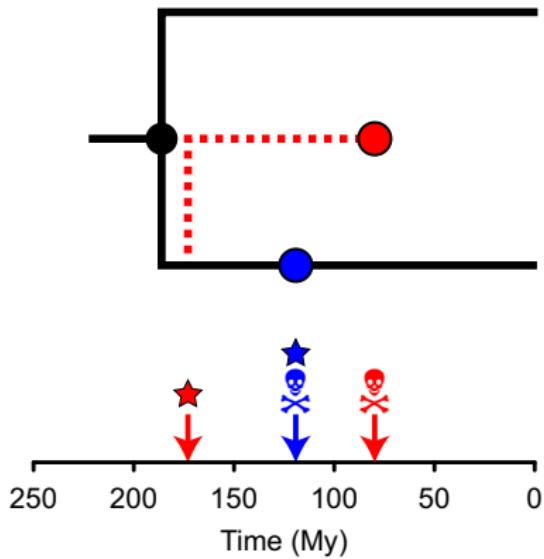
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

If $\star = \text{💀}$, then the new fossil lies directly on a branch in the extant tree



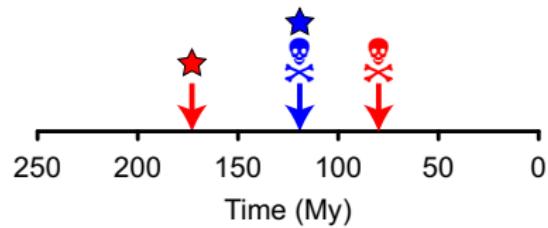
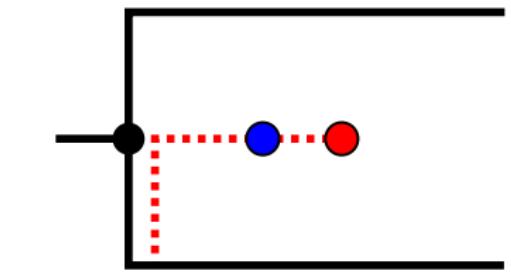
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

If $\star = \text{💀}$, then the new fossil lies directly on a branch in the extant tree



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

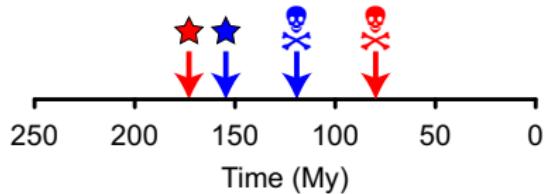
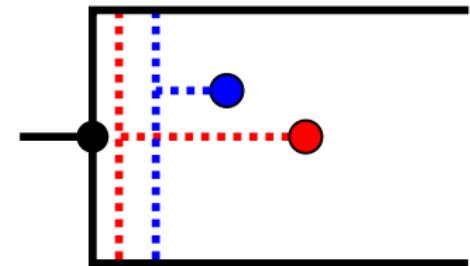
Or it is an ancestor of the other calibrating fossil



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

The probability of this realization of the diversification process is conditional on:

λ , μ , and ψ



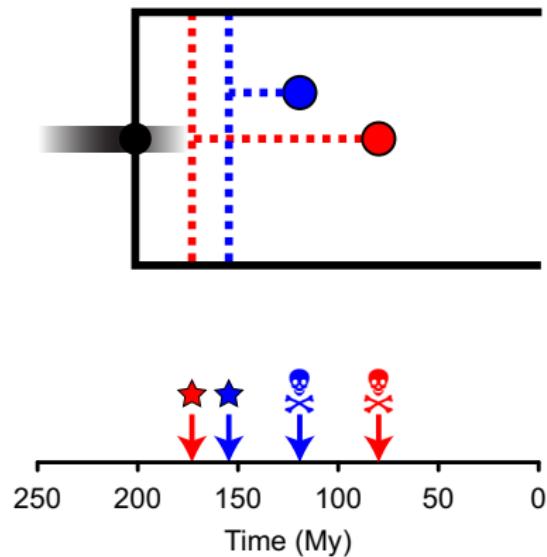
THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Using MCMC, we can sample the age of the calibrated node while conditioning on

λ , μ , and ψ
other node ages

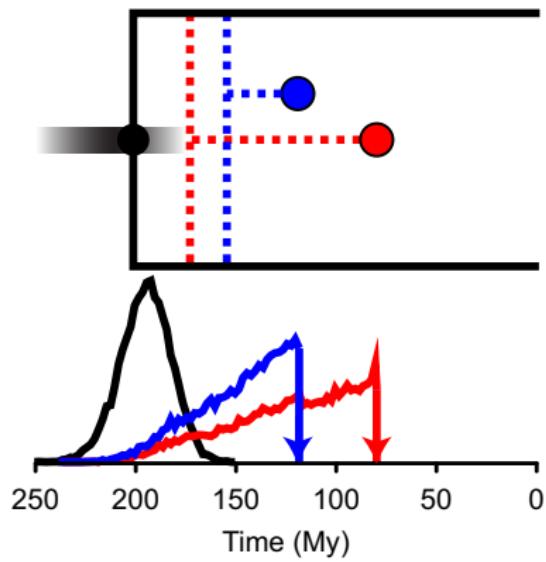
💀 and ★

💀 and ★



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

MCMC allows us to consider all possible values of \star (marginalization)



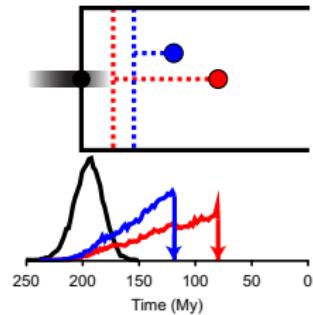
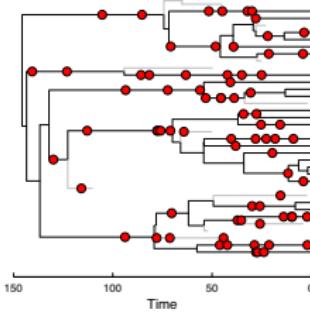
BAYESIAN INFERENCE UNDER THE FBD

Implemented in:

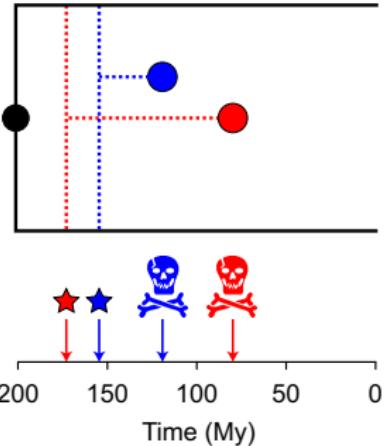
- DPPDiv:
github.com/trayc7/FDPPDIV
- BEAST2:
beast2.cs.auckland.ac.nz/

Available soon:

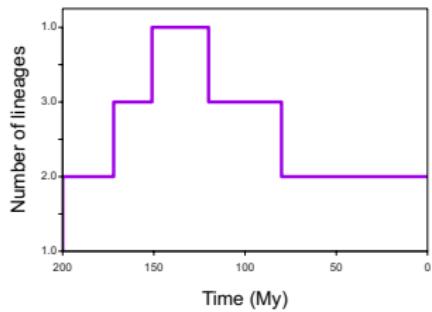
- RevBayes:
github.com/revbayes/revbayes



LINEAGES THROUGH TIME



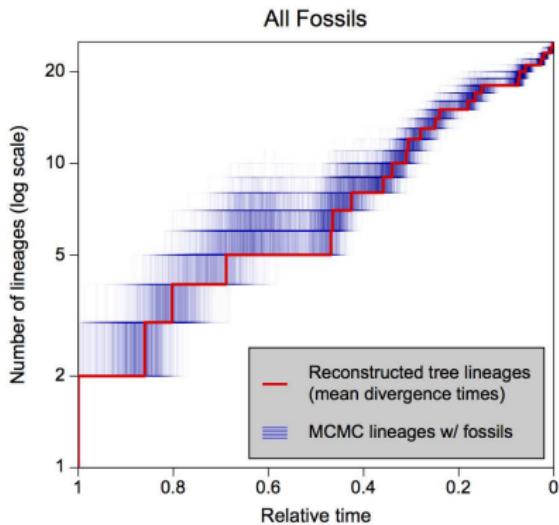
The FBD model provides an estimate of the number of lineages over time



LINEAGES THROUGH TIME

Lineage diversity over time with fossils

MCMC samples the times of lineages in the reconstructed tree **and** the times of the fossil lineages

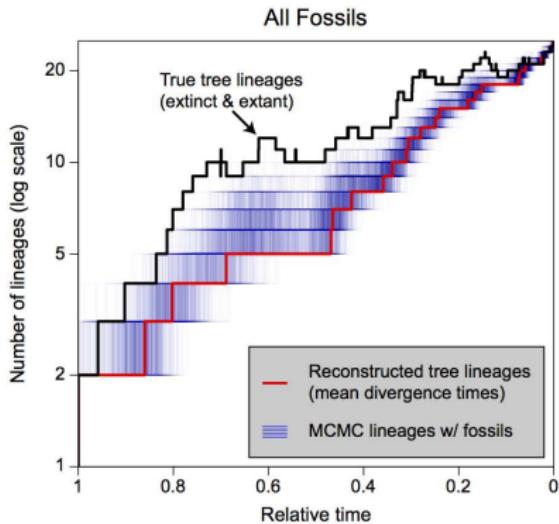


(for 1 simulation replicate with 10% random sample of fossils)

LINEAGES THROUGH TIME

Lineage diversity over time with fossils

Visualize extant and sampled fossil lineage diversification when using **all** available fossils (21 total)

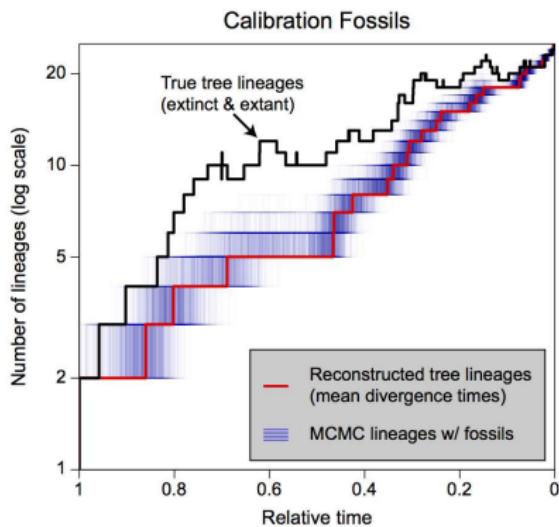


(for 1 simulation replicate with 10% random sample of fossils)

LINEAGES THROUGH TIME

Lineage diversity over time with fossils

Choosing only the oldest (calibration) fossils reduced the set of sampled fossils from 21 to 12, giving less information about diversification over time



(for 1 simulation replicate with 10% random sample of fossils)

BEARS: DIVERGENCE TIMES

Sequence data for extant species:

- 8 Ursidae
- 1 Canidae (gray wolf)
- 1 Phocidae (spotted seal)

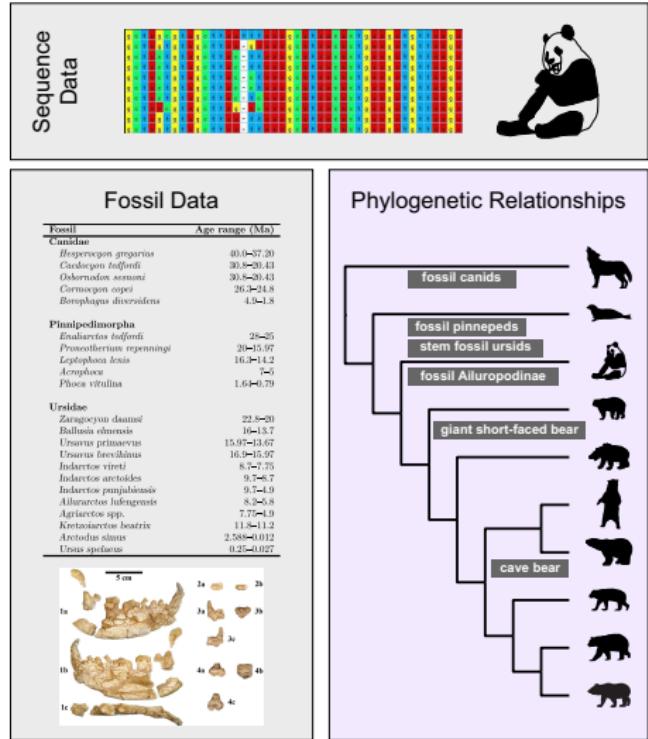
Fossil ages:

- 12 Ursidae
- 5 Canidae
- 5 Pinnipedimorpha

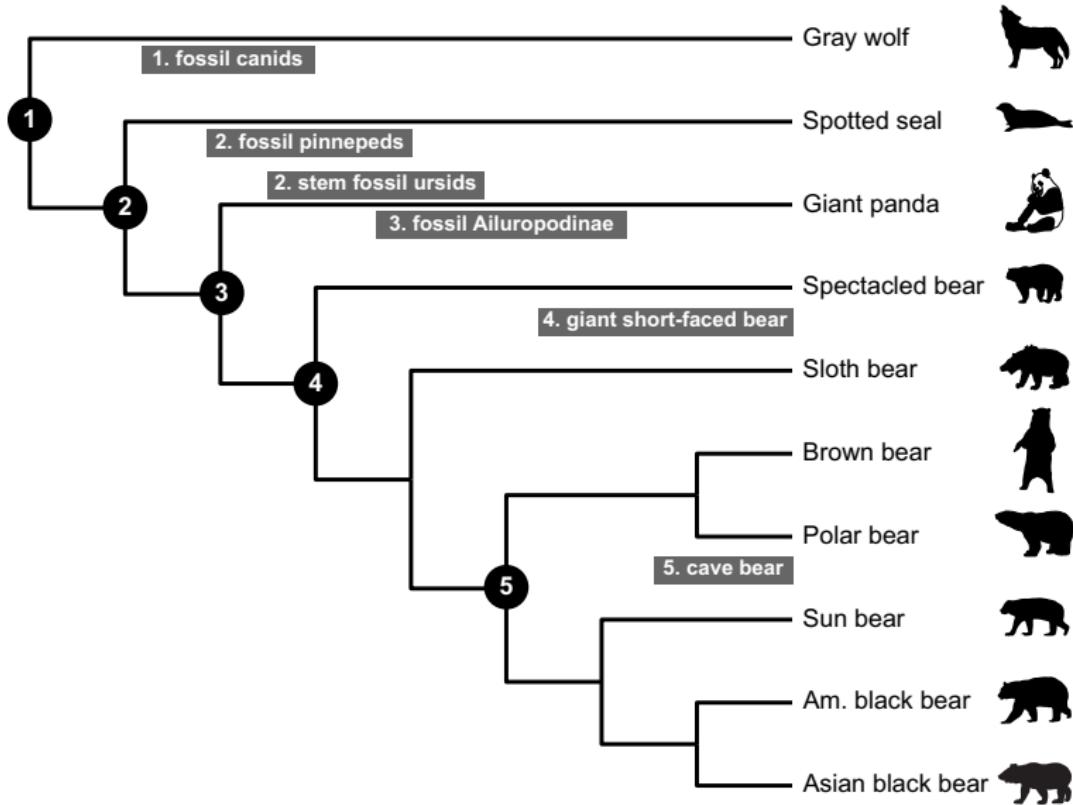
DPP relaxed clock model

(Heath, Holder, Huelsenbeck *MBE* 2012)

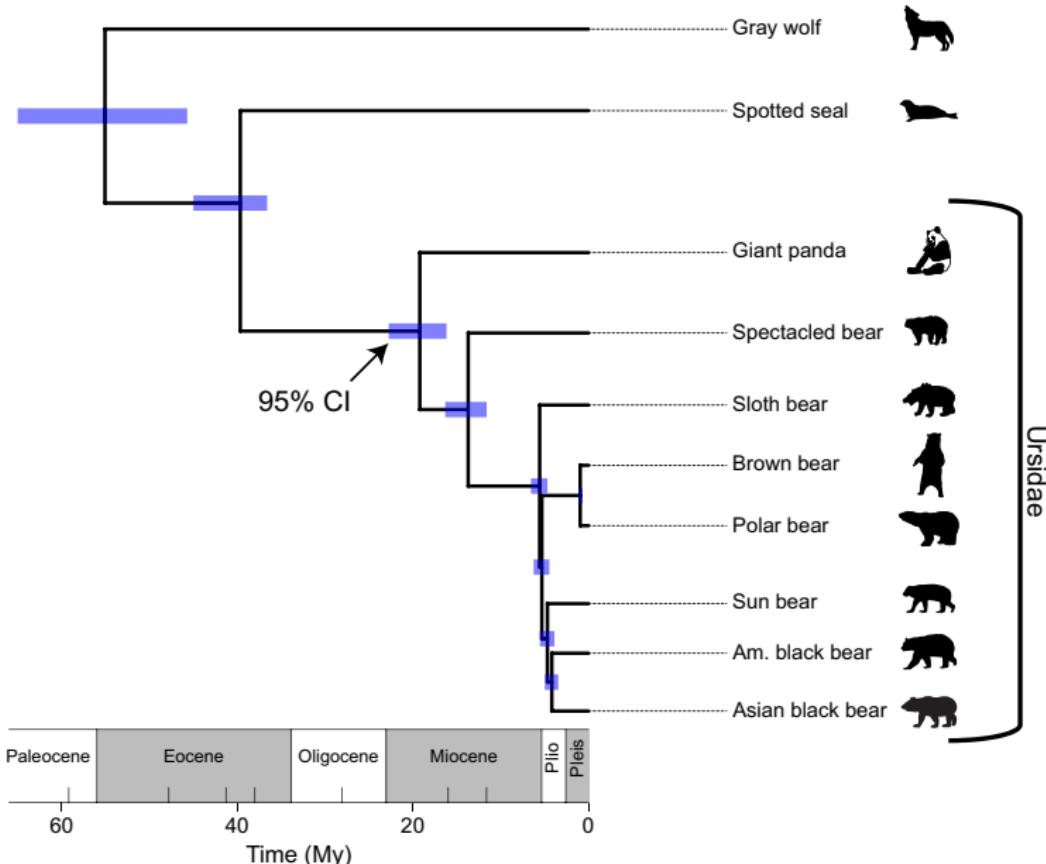
Fixed tree topology



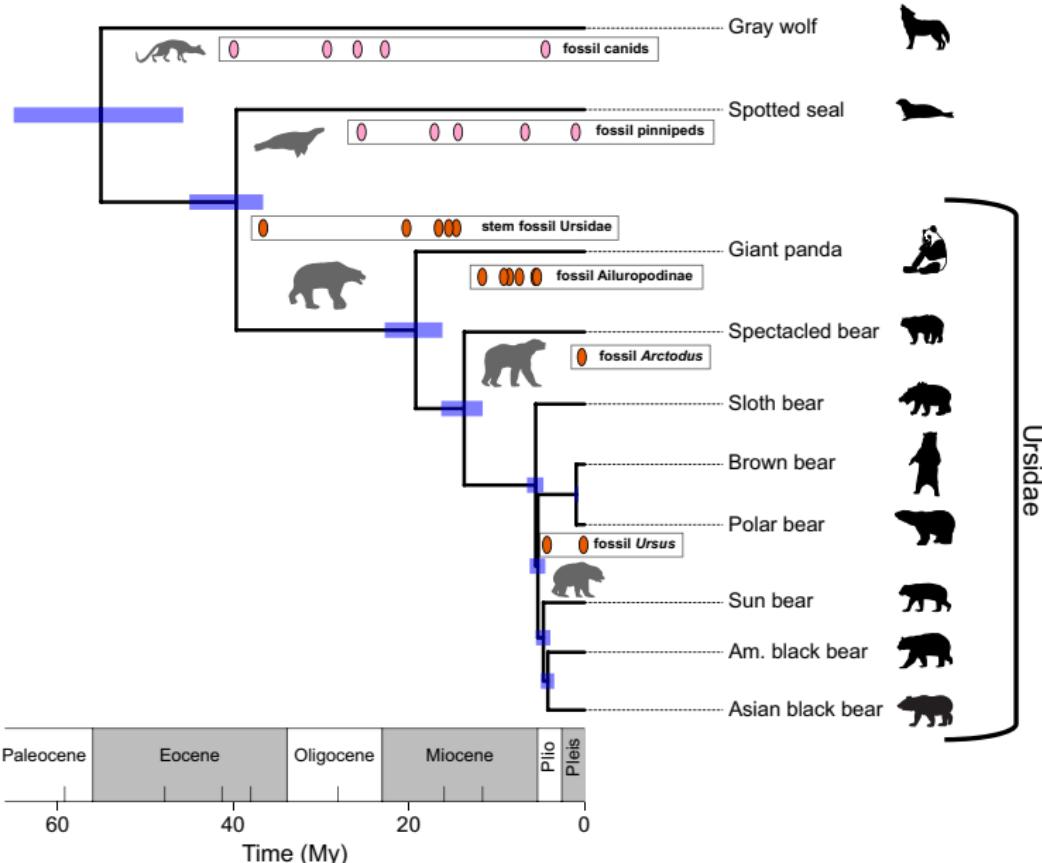
BEARS: DIVERGENCE TIMES



BEARS: DIVERGENCE TIMES



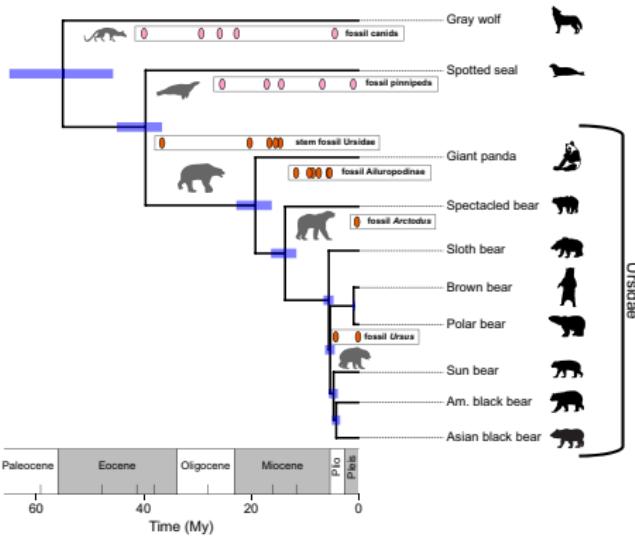
BEARS: DIVERGENCE TIMES



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Improved statistical inference of absolute node ages

Biologically motivated
models can better
capture statistical
uncertainty

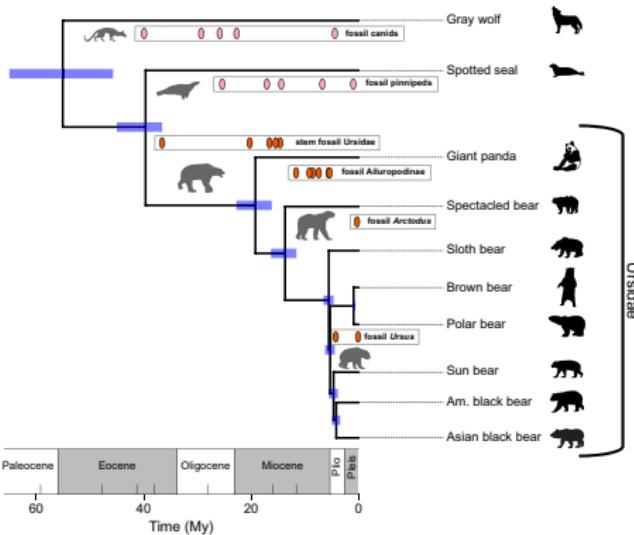


THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Improved statistical inference of absolute node ages

Use all available fossils

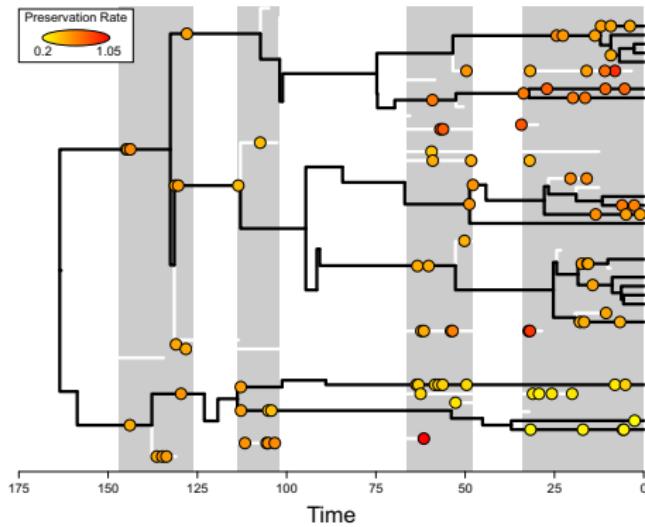
Eliminates arbitrary choice of calibration priors



THE FOSSILIZED BIRTH-DEATH PROCESS (FBD)

Improved statistical inference of absolute node ages

Extensions of the FBD can account for stratigraphic sampling of fossils and shifts in rates of speciation and extinction



Fossil Total Evidence DATING

Combining extant and fossil species

