Bayesian Divergence Time Estimation

Tracy Heath

Ecology, Evolution, & Organismal Biology
Iowa State University

@trayc7
http://phyloworks.org

2015 Workshop on Molecular Evolution
Woods Hole, MA USA

Outline

Overview of divergence time estimation

- Relaxed clock models — accounting for variation in substitution rates among lineages
- Tree models — lineage diversification and sampling

break

BEAST v2.2.0 Tutorial — Divergence-time estimation under birth-death processes

`http://phyloworks.org/workshops/divtime.html`

Choose one:

- Dating Bear Divergence Times with the Fossilized Birth-Death Process
- Estimating Epidemiological Parameters of an Ebola Outbreak

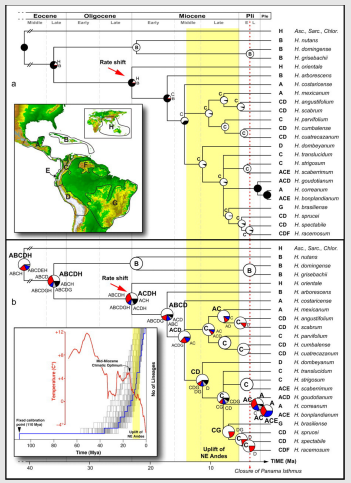
lobstah!

A Time-Scale for Evolution

Phylogenetic divergence-time estimation

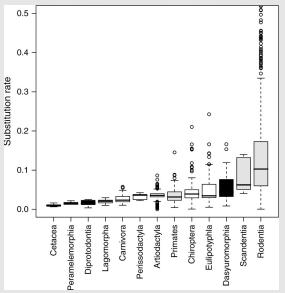
- What was the spacial and climatic environment of ancient angiosperms?
- How has mammalian body-size changed over time?
- How has the infection rate of HCV in Egypt changed over time?
- Is diversification in Caribbean anoles correlated with ecological opportunity?
- How has the rate of molecular evolution changed across the Tree of Life?

Historical biogeography



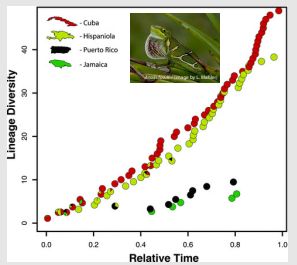
(Antonelli & Sanmartin. Syst. Biol. 2011)

Molecular evolution



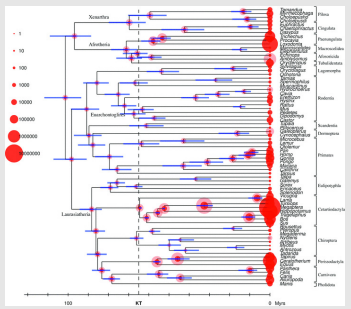
(Nabholz, Glemin, Galtier. MBE 2008)

Diversification



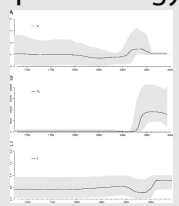
(Mahler, Revell, Glor, & Losos. Evolution 2010)

Trait evolution



(Lartillot & Delsuc. Evolution 2012)

Epidemiology



(Stadler et al. PNAS 2013)

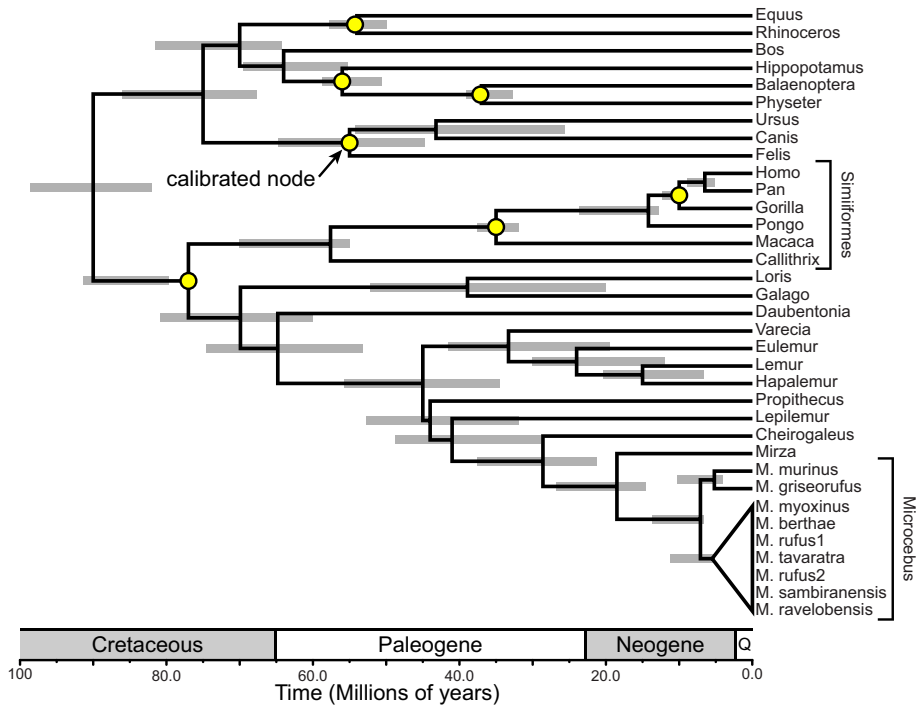
Divergence Time Estimation

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

Model how rates are distributed across the tree

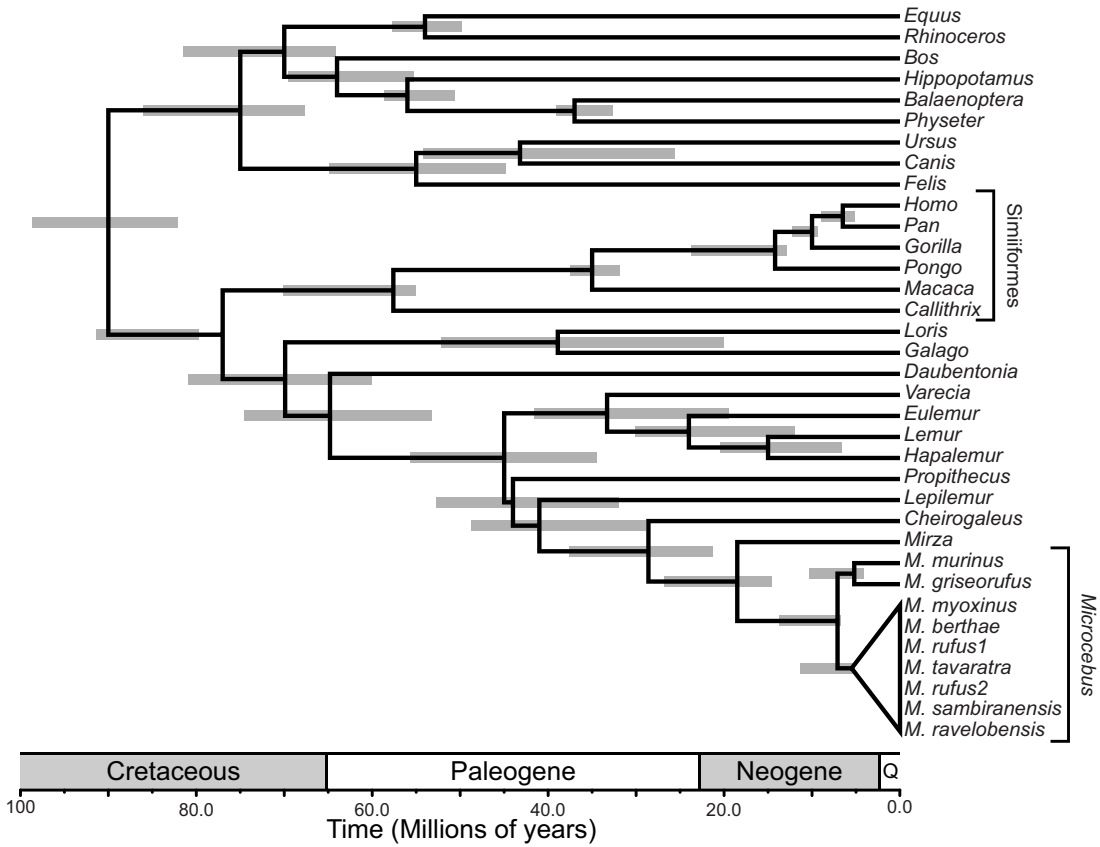
Describe the distribution of speciation events over time

External calibration information for estimates of absolute node times



A Time-Scale for Evolution

Phylogenetic trees can provide both topological information and temporal information

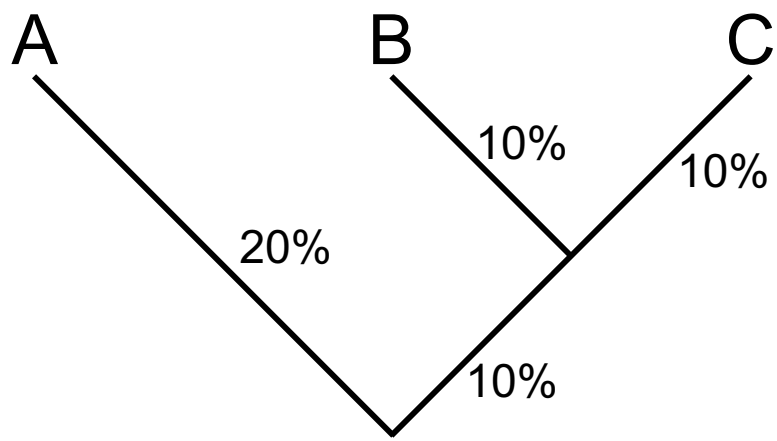


(Yang & Yoder *Syst. Biol.* 2003; Heath et al. *MBE* 2012)

The Global Molecular Clock

Assume that the rate of evolutionary change is constant over time

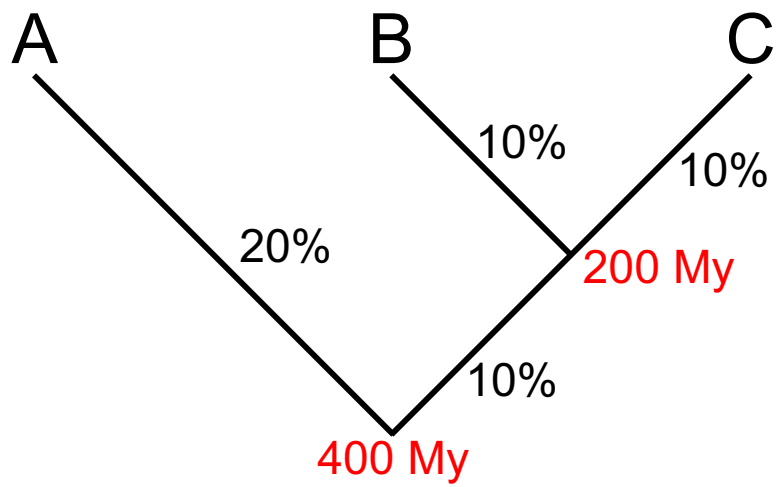
(branch lengths equal percent sequence divergence)



(Based on slides by Jeff Thorne; http://statgen.ncsu.edu/thorne/compmolevo.html)

The Global Molecular Clock

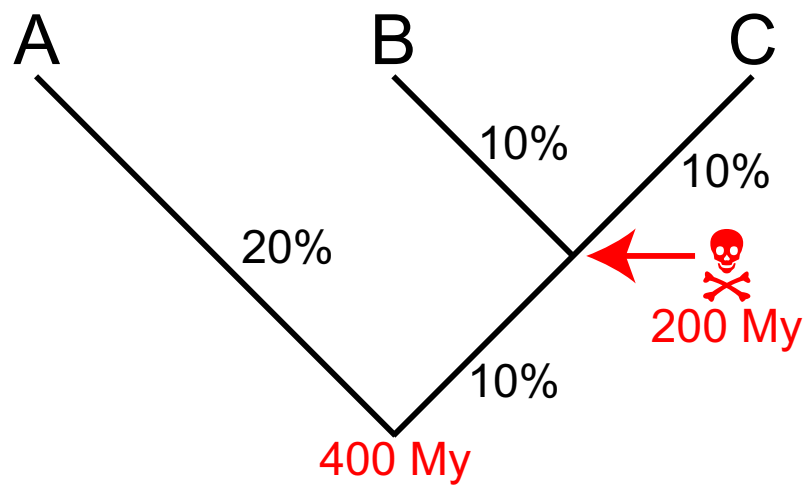
We can date the tree if we know the rate of change is 1% divergence per 10 My



(Based on slides by Jeff Thorne; http://statgen.ncsu.edu/thorne/compmolevo.html)

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



(Based on slides by Jeff Thorne; http://statgen.ncsu.edu/thorne/compmolevo.html)

Rejecting the Global Molecular Clock

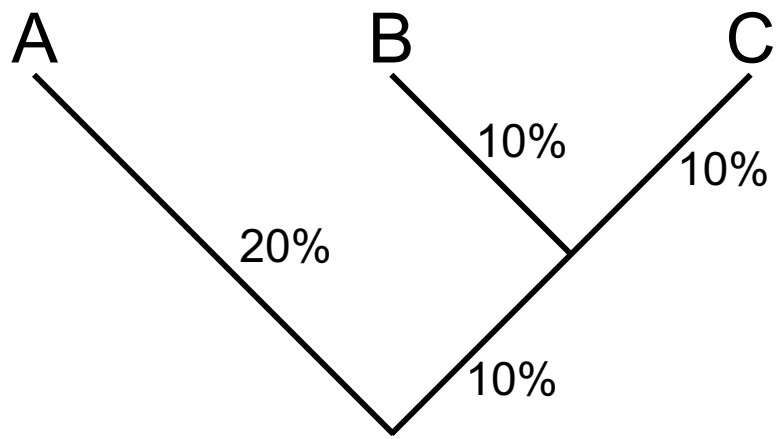
Rates of evolution vary across lineages and over time

Mutation rate:
Variation in

- metabolic rate
- generation time
- DNA repair

Fixation rate:
Variation in

- strength and targets of selection
- population sizes

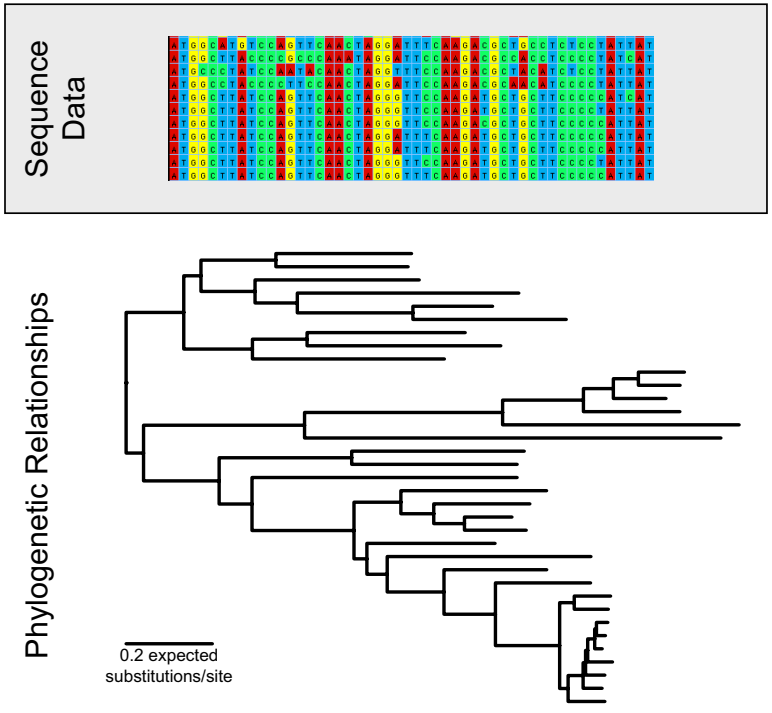


Unconstrained Analysis

Sequence data provide information about **branch lengths**

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

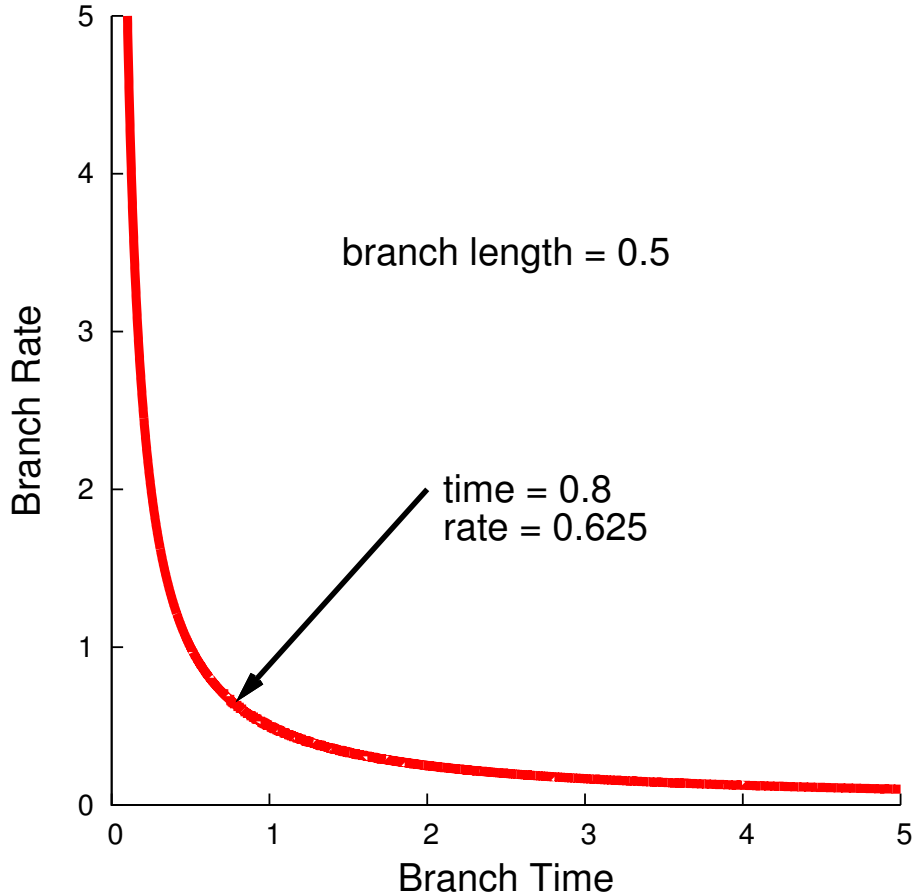
branch length = rate × time



Rate and Time

The sequence data provide information about branch length

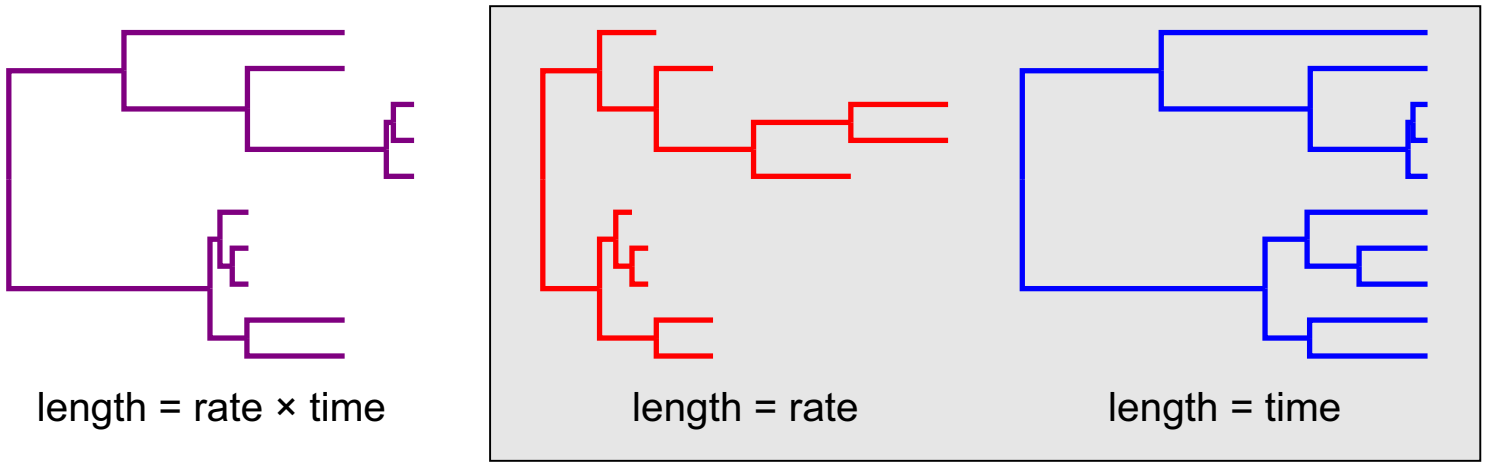
for any possible rate, there's a time that fits the branch length perfectly



(based on Thorne & Kishino, 2005)

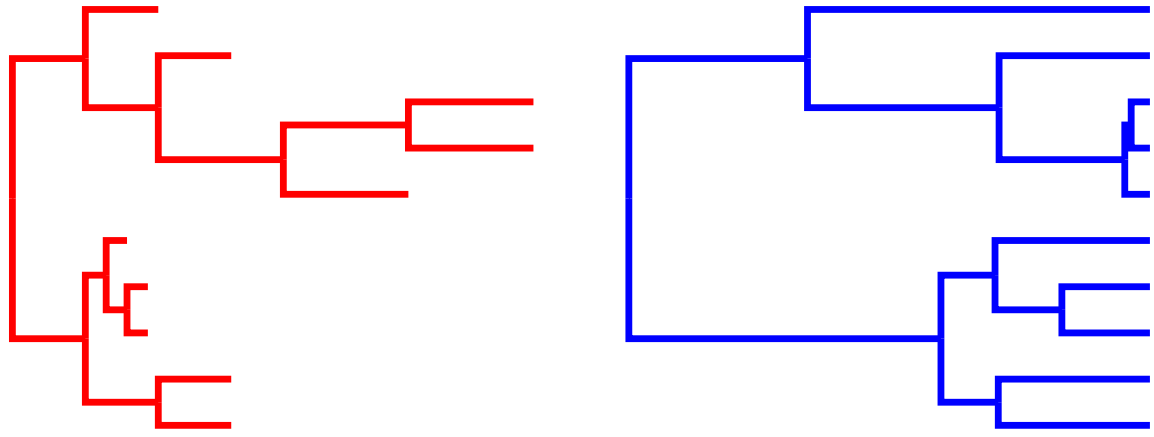
Rate and Time

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

Bayesian Divergence Time Estimation



length = rate

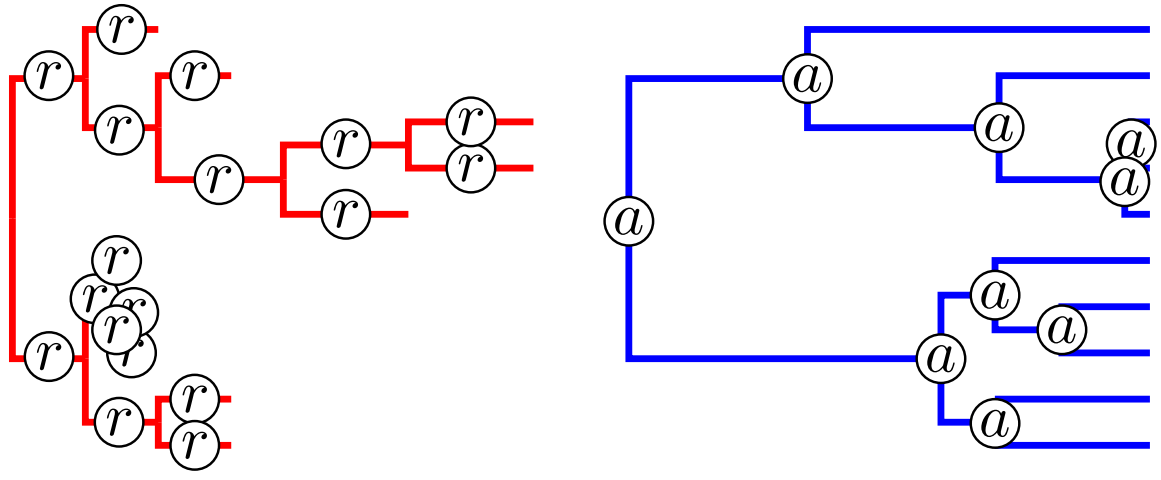
length = time

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

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

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

Bayesian Divergence Time Estimation



length = rate

length = time

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

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

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

Bayesian Divergence Time Estimation

Posterior probability

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

$\boldsymbol{\mathcal{R}}$	Vector of rates on branches
$\boldsymbol{\mathcal{A}}$	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

$$f(\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s \mid D) = \frac{f(D \mid \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}}, \theta_s)\, f(\boldsymbol{\mathcal{R}} \mid \theta_{\mathcal{R}})\, f(\boldsymbol{\mathcal{A}} \mid \theta_{\mathcal{A}})\, f(\theta_s)}{f(D)}$$

$f(D \mid \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s)$	Likelihood
$f(\boldsymbol{\mathcal{R}} \mid \theta_{\mathcal{R}})$	Prior on rates
$f(\boldsymbol{\mathcal{A}} \mid \theta_{\mathcal{A}})$	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: $f(\boldsymbol{\mathcal{R}} \mid \theta_{\mathcal{R}})$
- Node ages: $f(\boldsymbol{\mathcal{A}} \mid \theta_{\mathcal{A}}, \boldsymbol{\mathcal{C}})$

Modeling Rate Variation

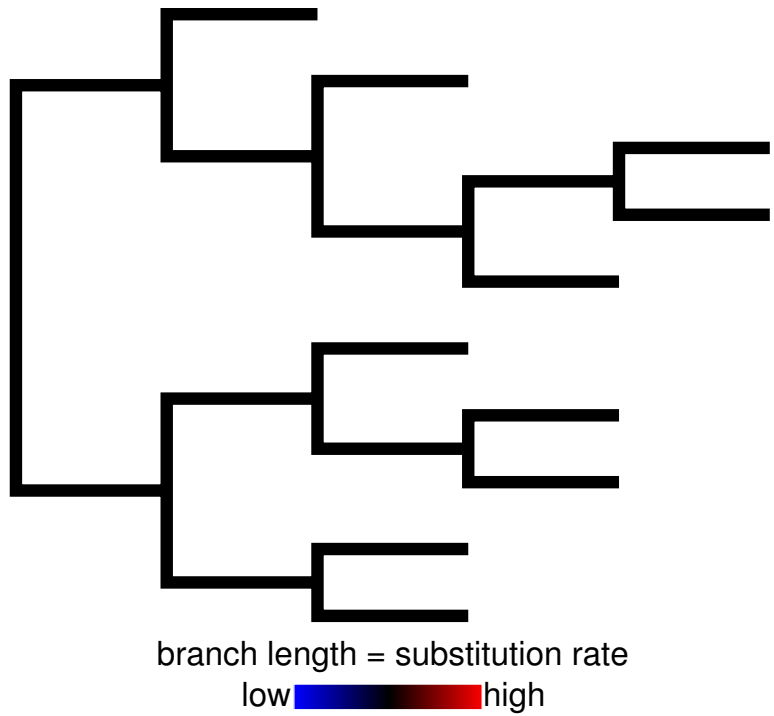
Some models describing lineage-specific substitution rate variation:

- **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

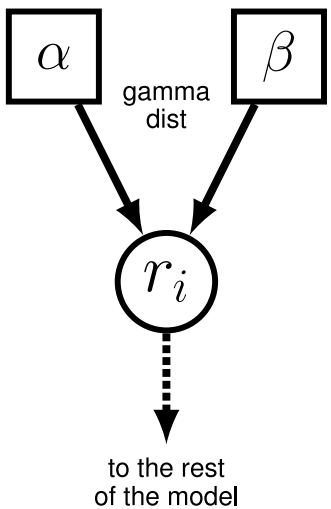
The substitution rate is constant over time

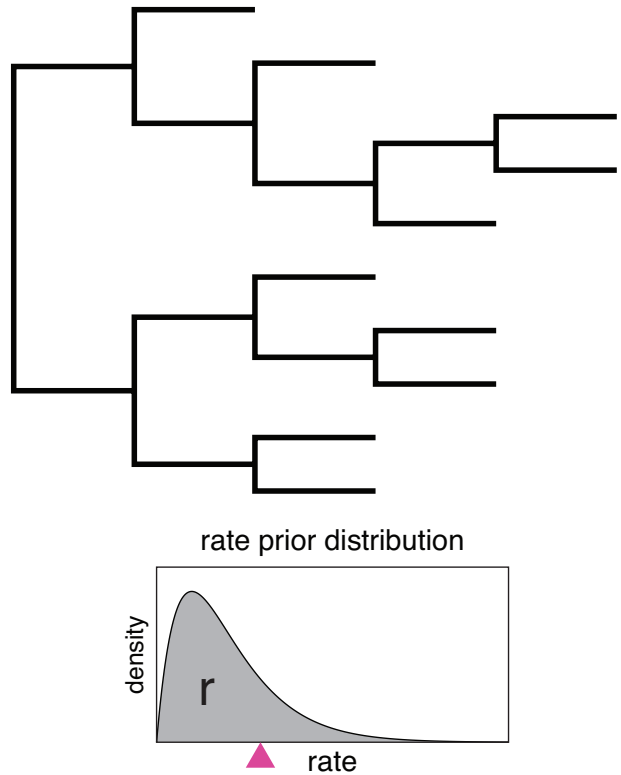
All lineages share the same rate



(Zuckerkandl & Pauling, 1962)

Global Molecular Clock





(Zuckerkandl & Pauling, 1962)

Relaxed-Clock Models

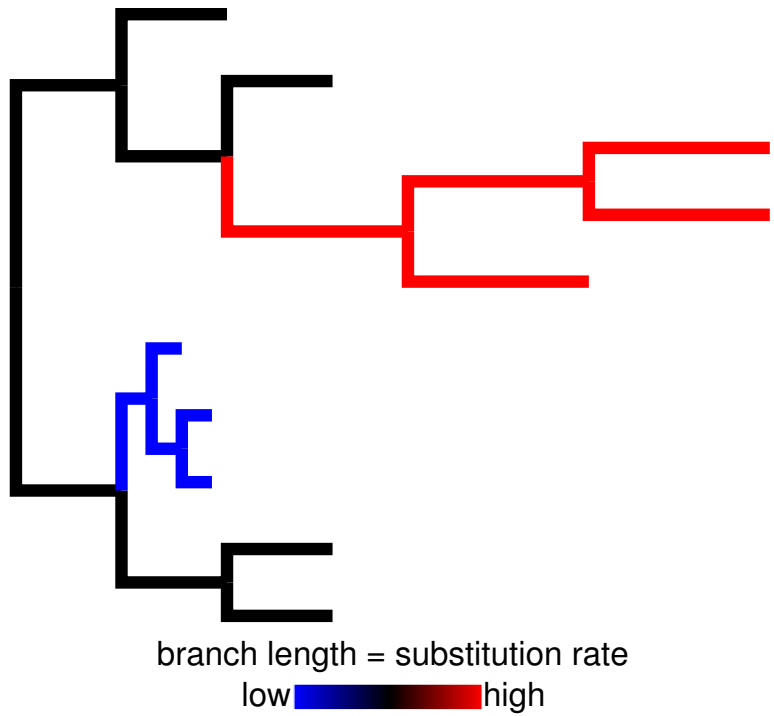
To accommodate variation in substitution rates 'relaxed-clock' models estimate lineage-specific substitution rates

- **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)

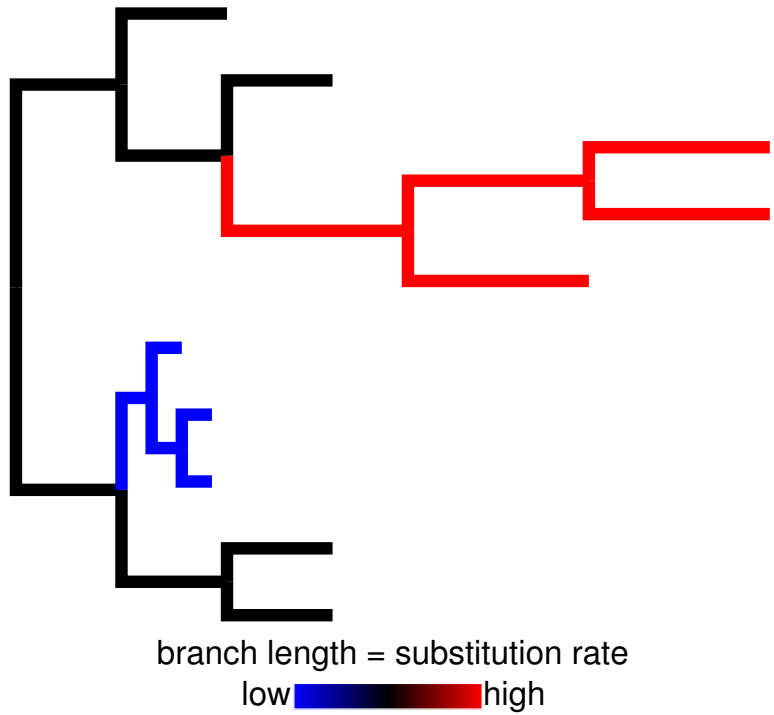


(Yang & Yoder 2003, Drummond and Suchard 2010)

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*

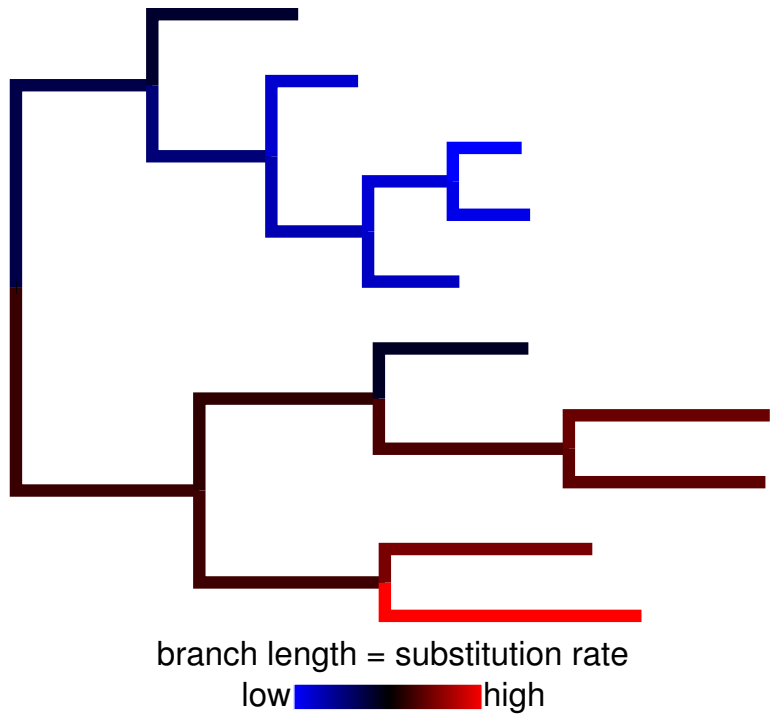


(Yang & Yoder 2003, Drummond and Suchard 2010)

Autocorrelated Rates

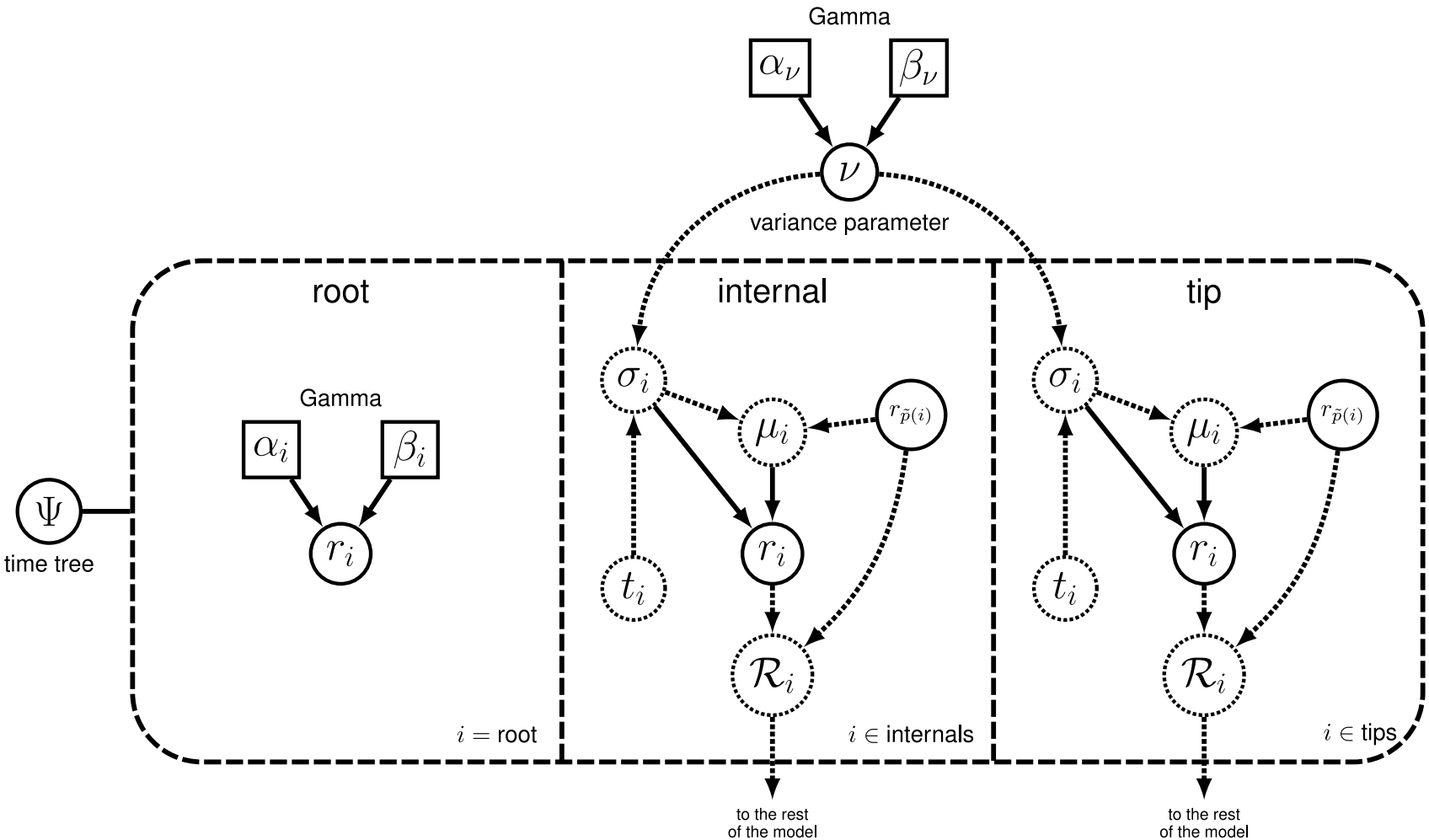
Substitution rates evolve gradually over time – closely related lineages have similar rates

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



(Thorne, Kishino & Painter 1998; Kishino, Thorne & Bruno 2001)

Autocorrelated Rates

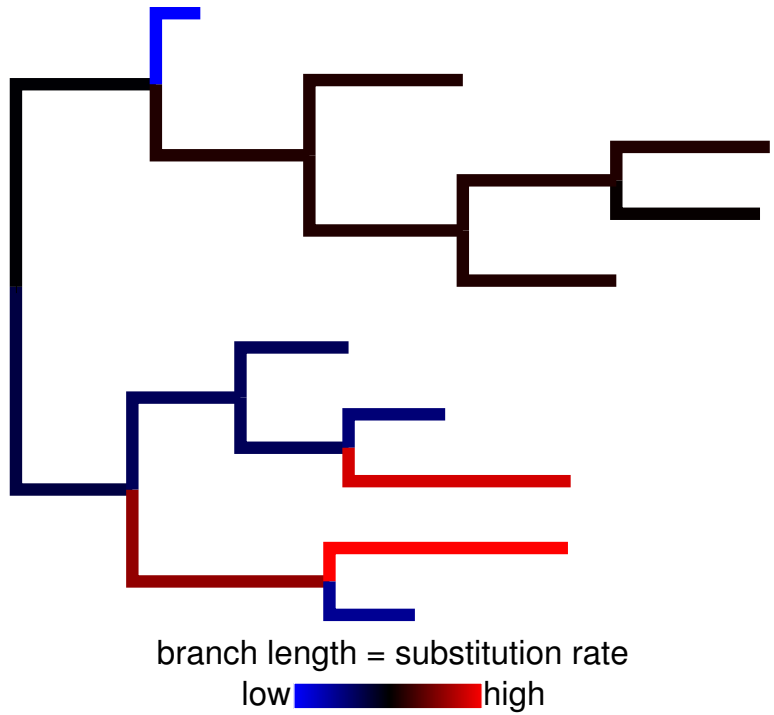


(Thorne, Kishino & Painter 1998; Kishino, Thorne & Bruno 2001)

Punctuated Rate Change

Rate changes occur along lineages according to a point process

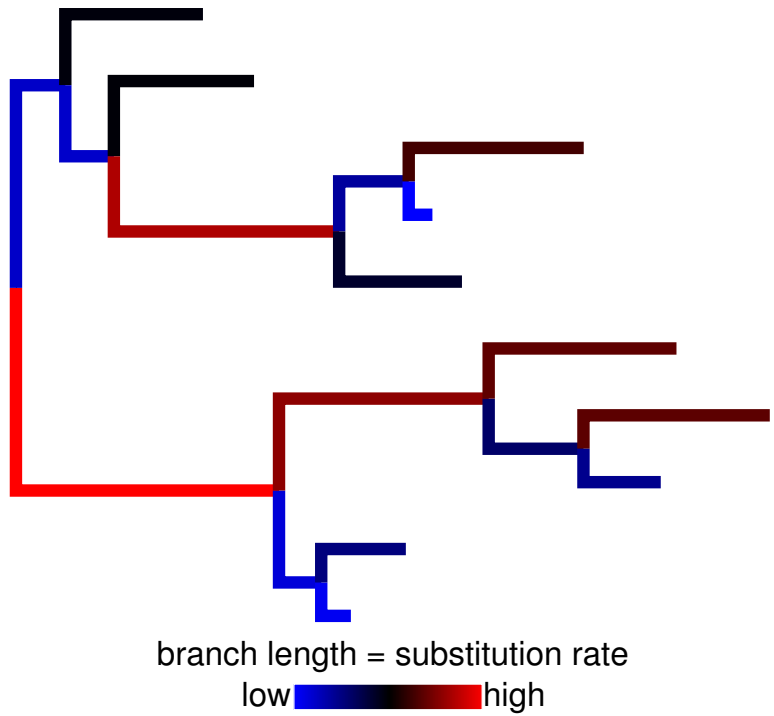
At rate-change events, the new rate is a product of the parent's rate and a Γ-distributed multiplier



(Huelsenbeck, Larget and Swofford 2000)

Independent/Uncorrelated Rates

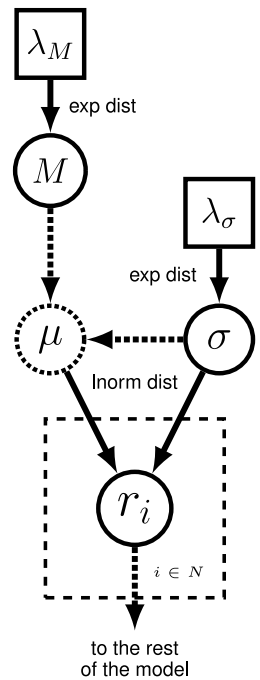
Lineage-specific rates are uncorrelated when the rate assigned to each branch is independently drawn from an underlying distribution



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

Independent/Uncorrelated Rates

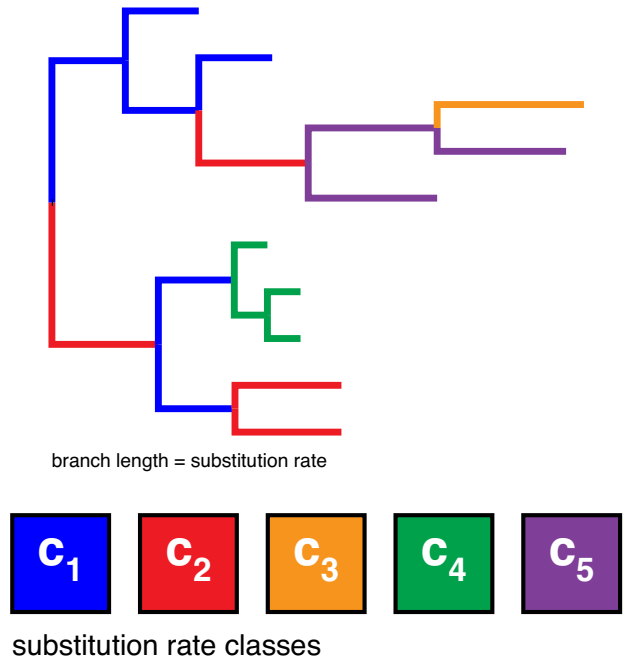
Lineage-specific rates are uncorrelated when the rate assigned to each branch is independently drawn from an underlying distribution



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

Infinite Mixture Model

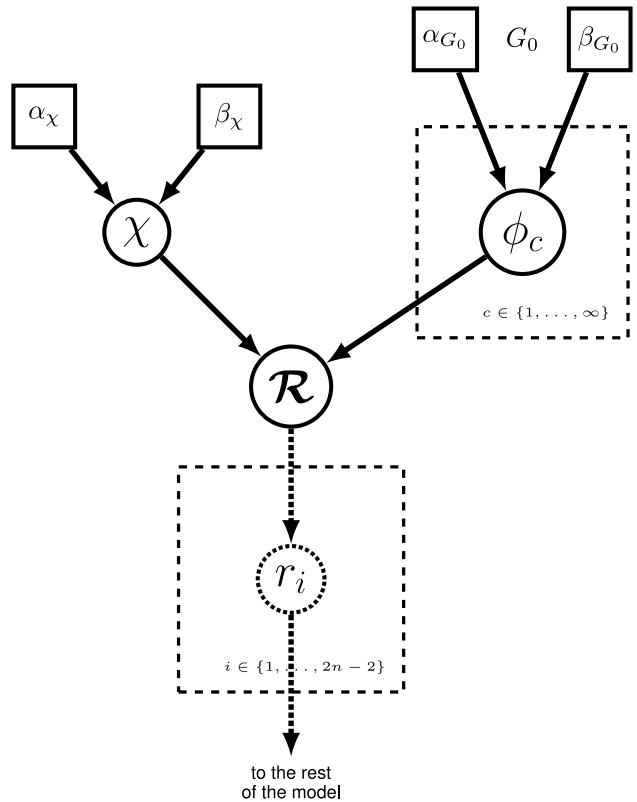
Dirichlet process prior:
Branches are partitioned into distinct rate categories



(Heath, Holder, Huelsenbeck. 2012 *MBE*)

Infinite Mixture Model

Dirichlet process prior:
Branches are partitioned into distinct rate categories



(Heath, Holder, Huelsenbeck. 2012 *MBE*)

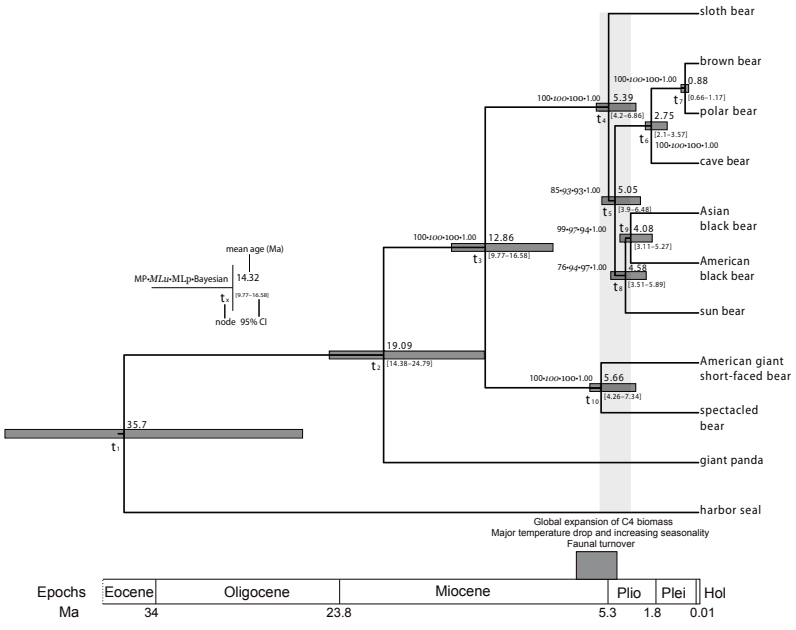
Modeling Rate Variation

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

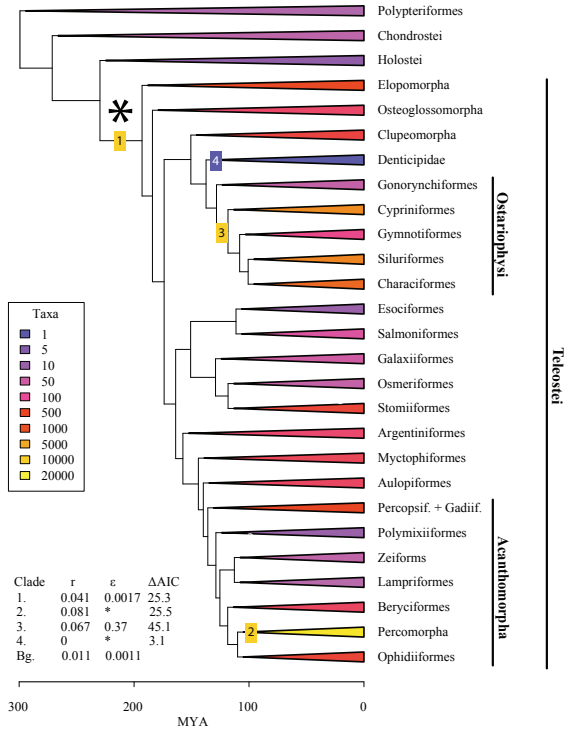
- **Global molecular clock**
- **Local molecular clocks**
- **Punctuated rate change model**
- **Log-normally distributed autocorrelated rates**
- **Uncorrelated/independent rates models**
- **Dirchlet process prior**

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**
- **Dirchlet process prior**

Considering model selection, uncertainty, & plausibility is **very** important for Bayesian divergence time analysis

Bayesian Divergence Time Estimation

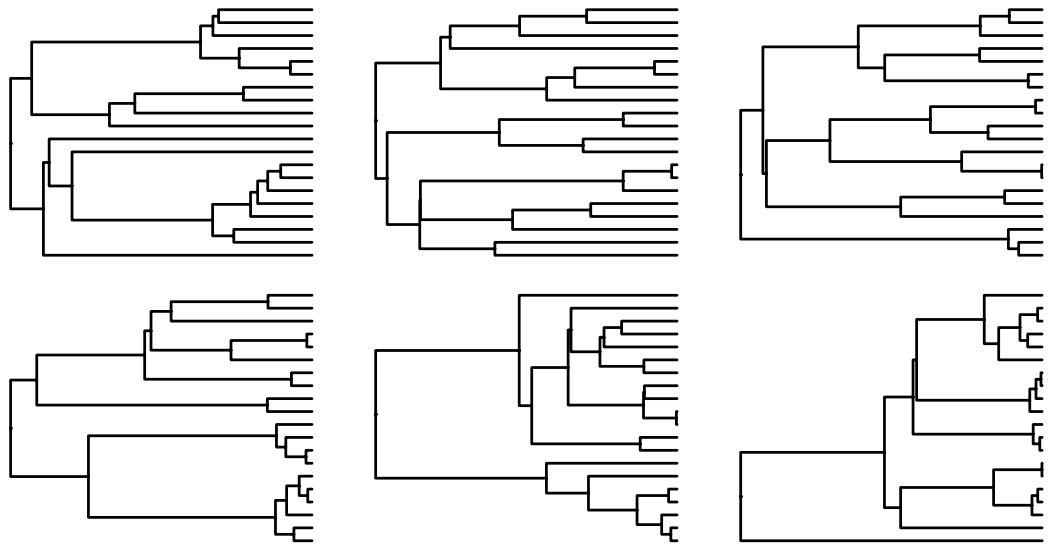
Estimating divergence times relies on 2 main elements:

- Branch-specific rates: $f(\boldsymbol{\mathcal{R}} \mid \theta_{\mathcal{R}})$
- Node ages: $f(\boldsymbol{\mathcal{A}} \mid \theta_{\mathcal{A}})$

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

Priors on the Tree and Node Ages

Relaxed clock Bayesian analyses require a prior distribution on time trees



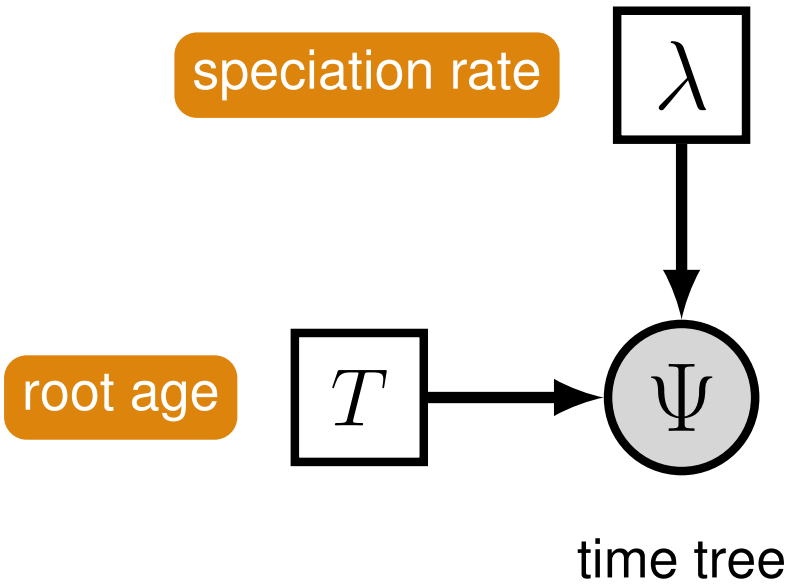
Different node-age priors make different assumptions about the timing of divergence events

Stochastic Branching Processes

Node-age priors based on stochastic models of lineage diversification

Yule process: assumes a constant rate of speciation, across lineages

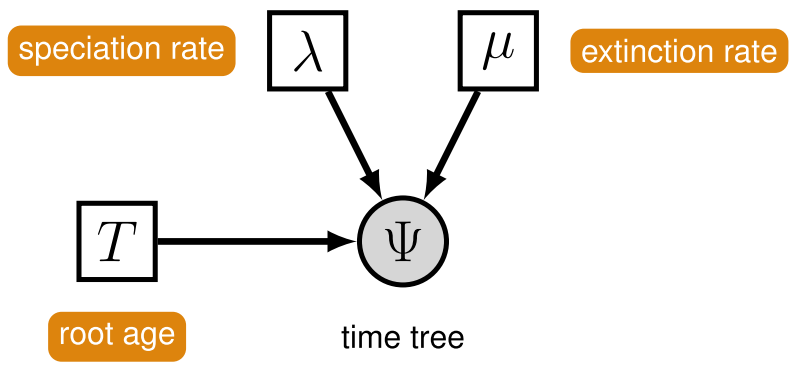
A pure birth process—every node leaves extant descendants (no extinction)



Stochastic Branching Processes

Node-age priors based on stochastic models of lineage diversification

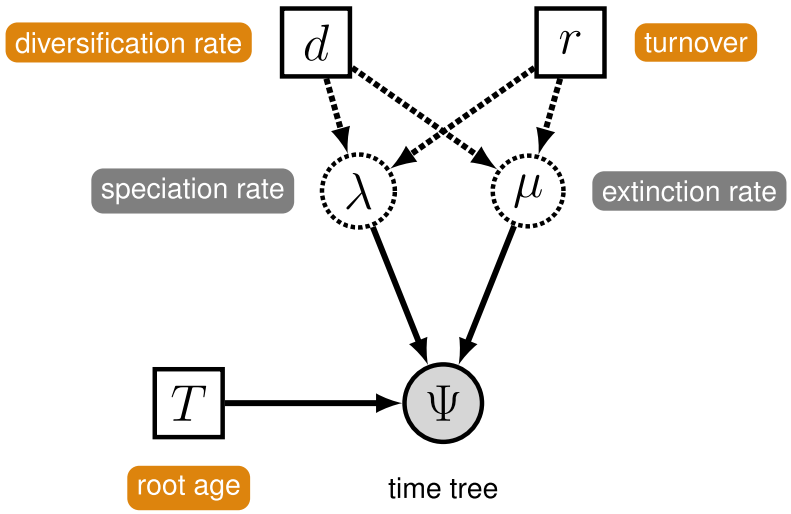
Constant-rate birth-death process: at any point in time a lineage can speciate at rate λ or go extinct with a rate of μ



Stochastic Branching Processes

Node-age priors based on stochastic models of lineage diversification

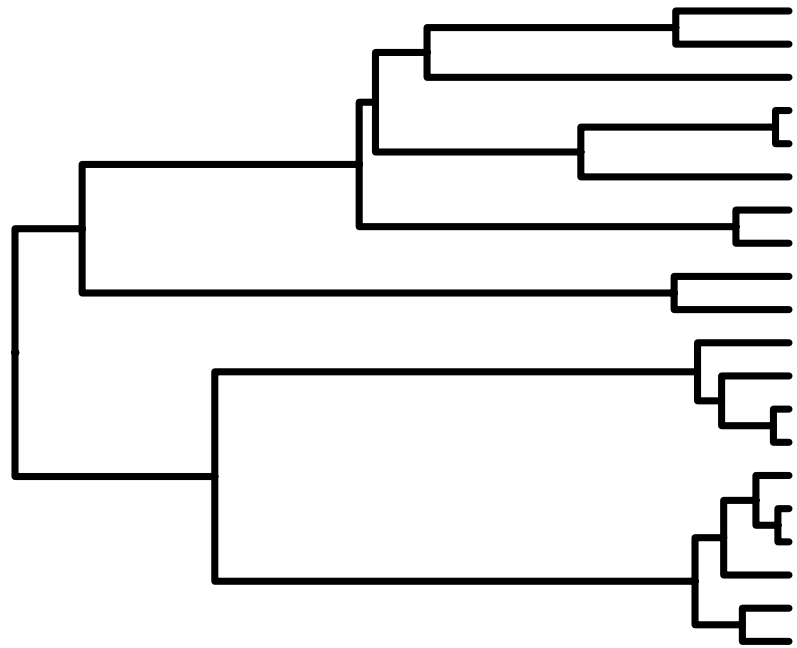
Constant-rate birth-death process: at any point in time a lineage can speciate at rate λ or go extinct with a rate of μ



Stochastic Branching Processes

Node-age priors based on stochastic models of lineage diversification

Constant-rate birth-death process: at any point in time a lineage can speciate at rate λ or go extinct with a rate of μ

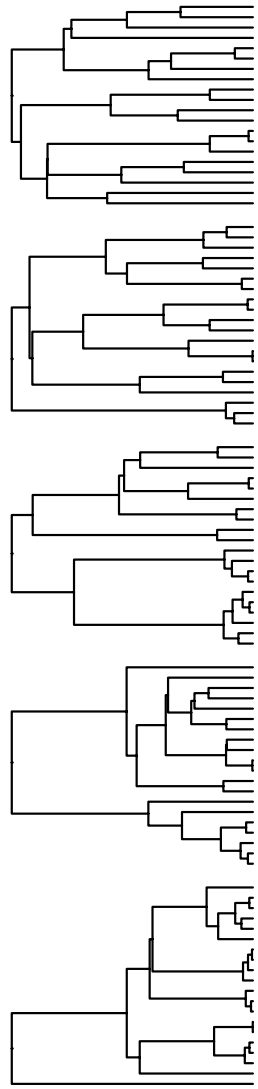


Stochastic Branching Processes

Different values of λ and μ lead to different trees

Bayesian inference under these models can be very sensitive to the values of these parameters

Using hyperpriors on λ and μ accounts for uncertainty in these hyperparameters

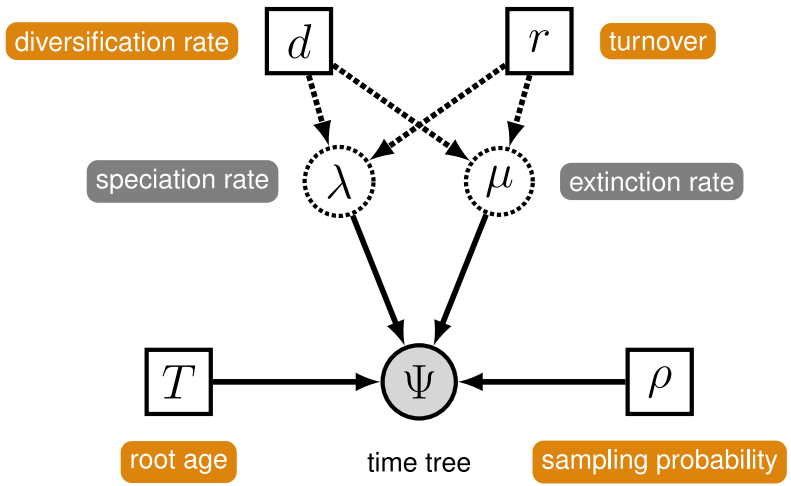


Stochastic Branching Processes

Node-age priors based on stochastic models of lineage diversification

Birth-death-sampling process: an extension of the constant-rate birth-death model that accounts for random sampling of tips

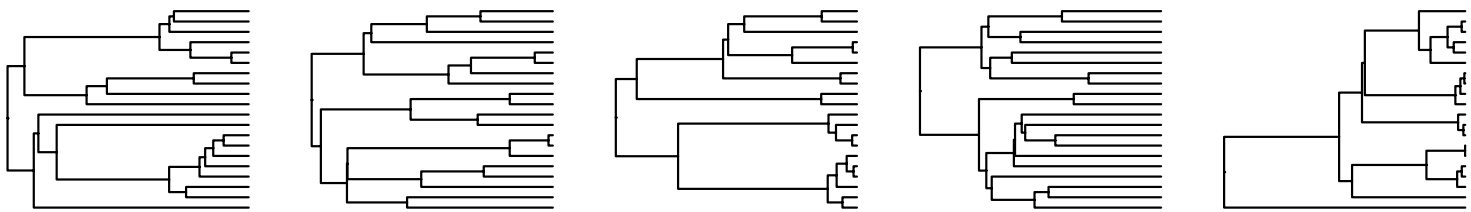
Conditions on a probability of sampling a tip, ρ



Priors on Node Times

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

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



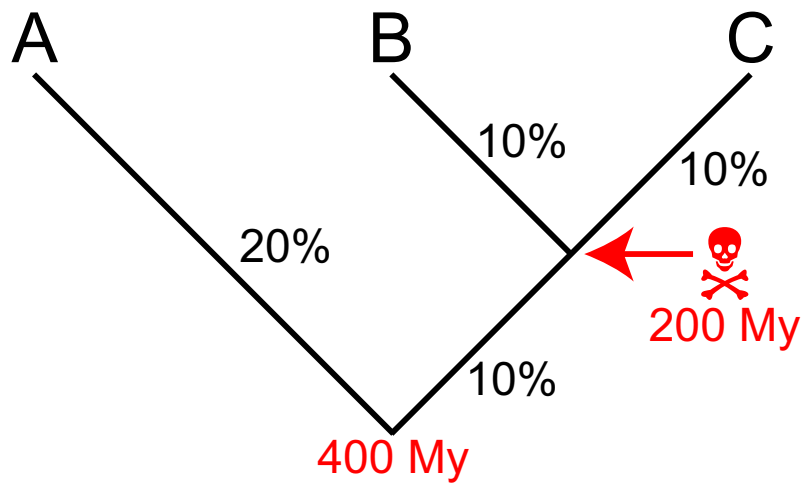
We need external information (like fossils) to provide absolute time scale

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

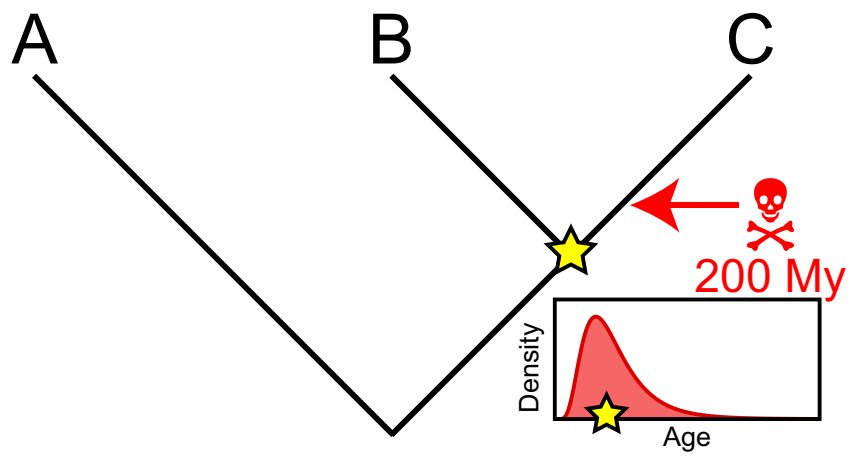
Uncertainty in the placement of fossils



Calibration Densities

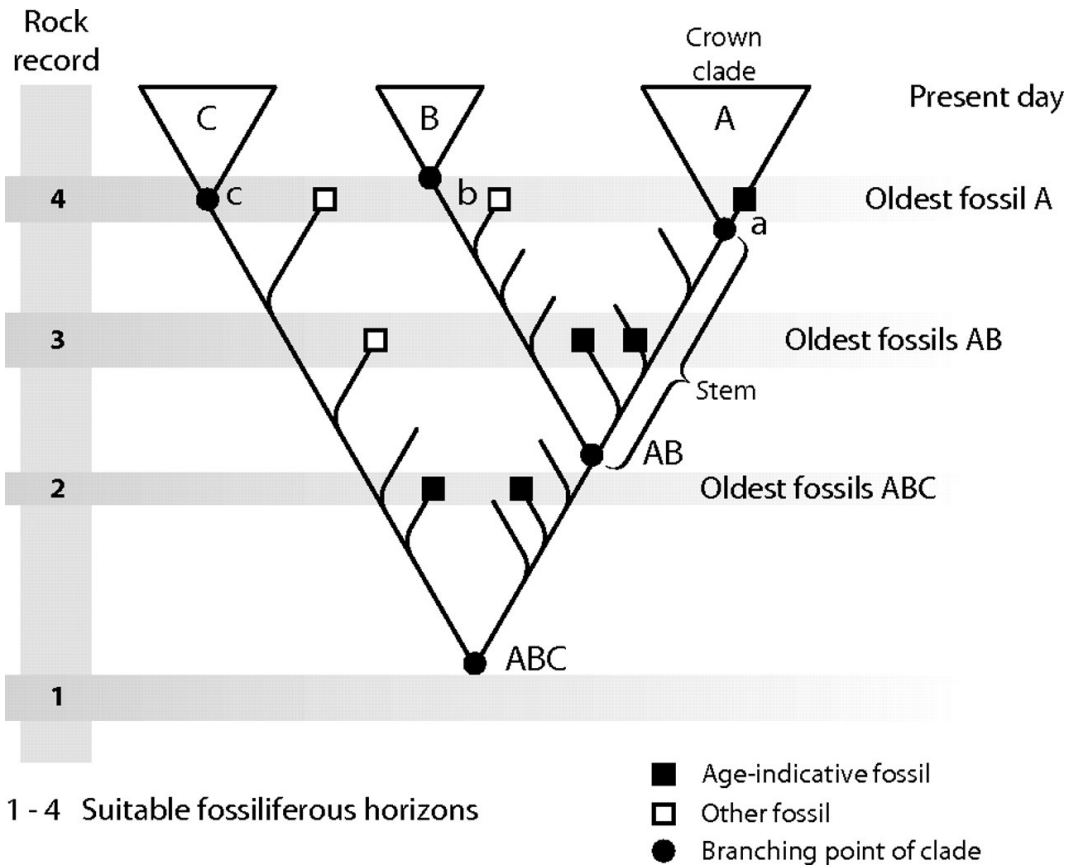
Bayesian inference is well suited to accommodating uncertainty in the age of the calibration node

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



Assigning Fossils to Clades

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

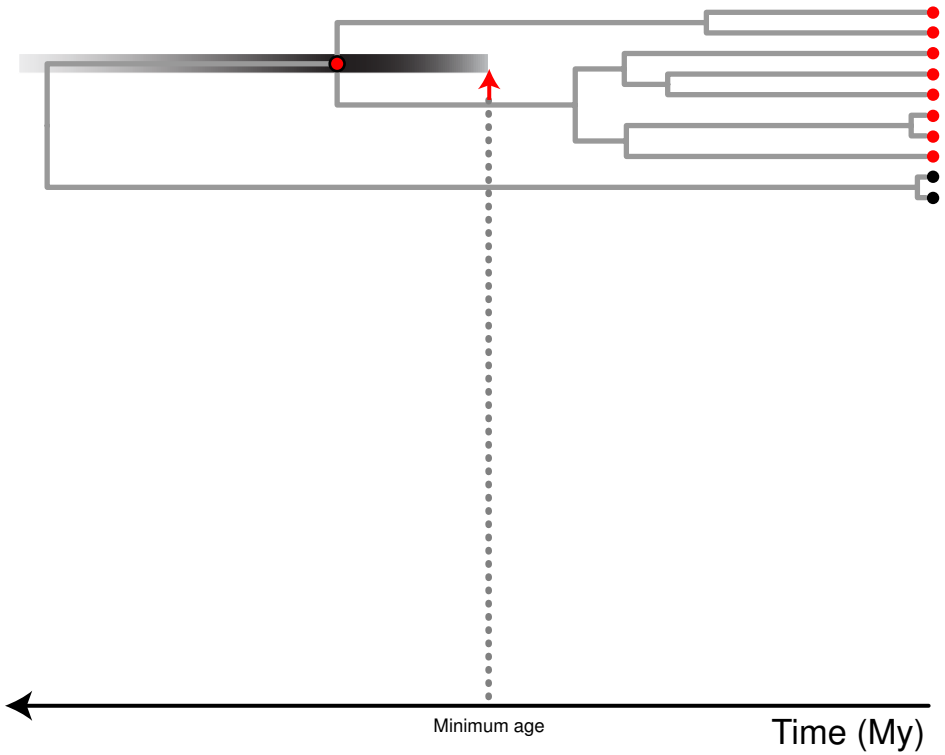


(figure from Benton & Donoghue *Mol. Biol. Evol.* 2007)

Fossil Calibration

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

Reliable **maximum** bounds are typically unavailable

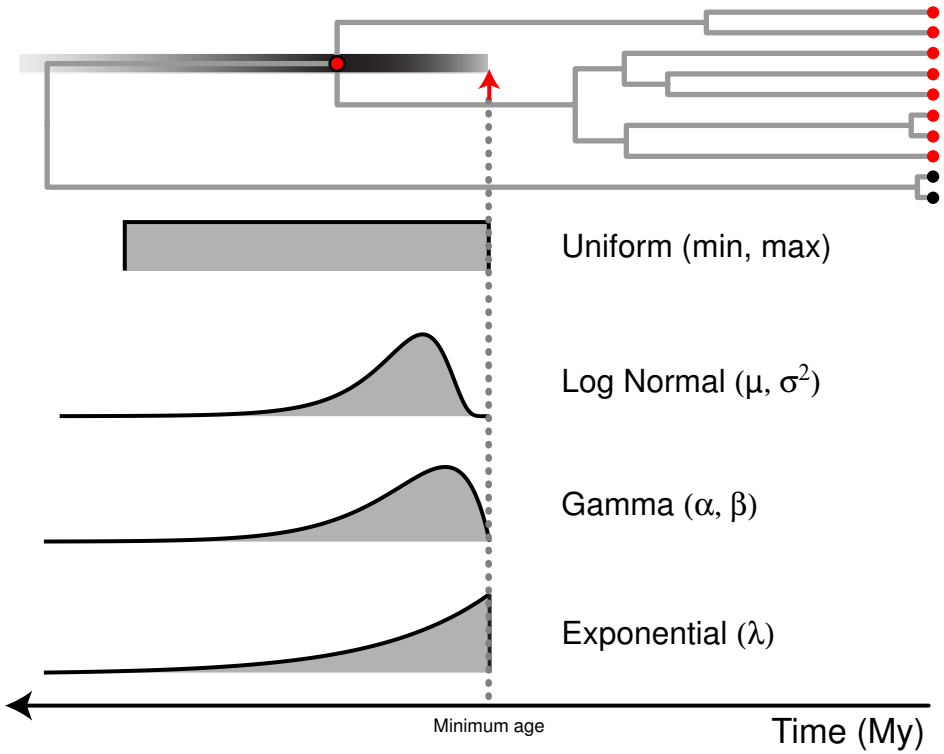


Prior Densities on Calibrated Nodes

Common practice in Bayesian divergence-time estimation:

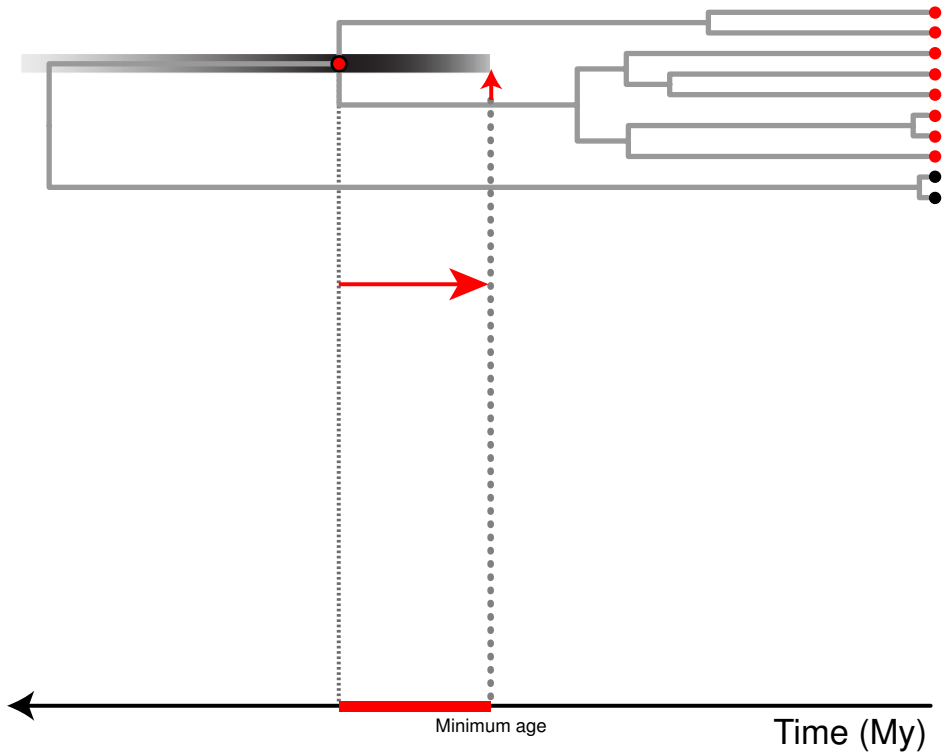
Parametric distributions are typically off-set by the age of the oldest fossil assigned to a clade

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



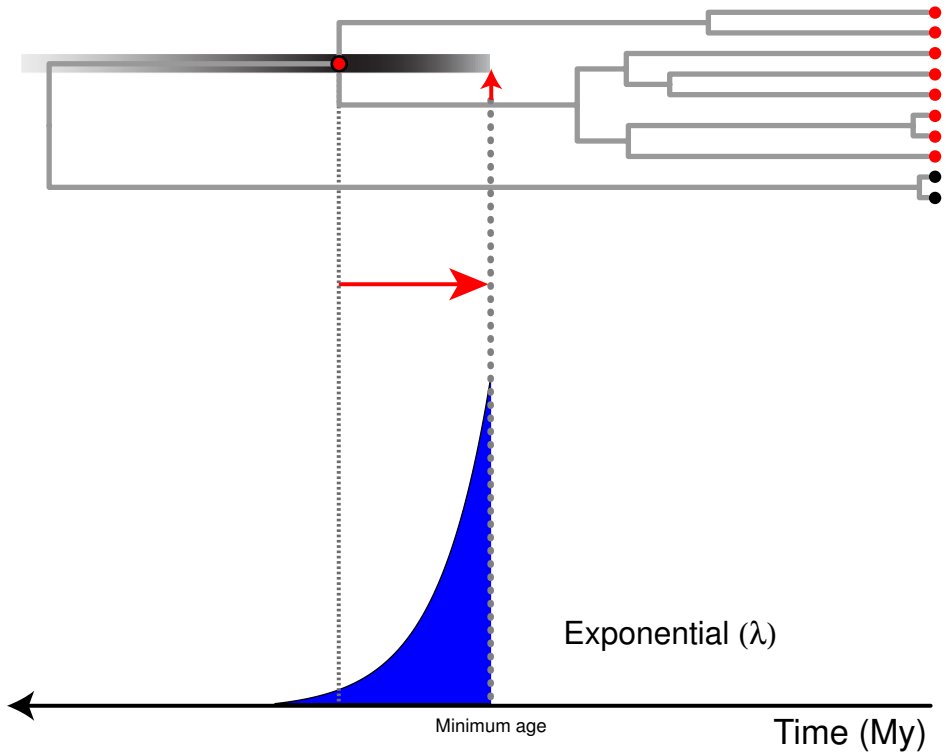
Prior Densities on Calibrated Nodes

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



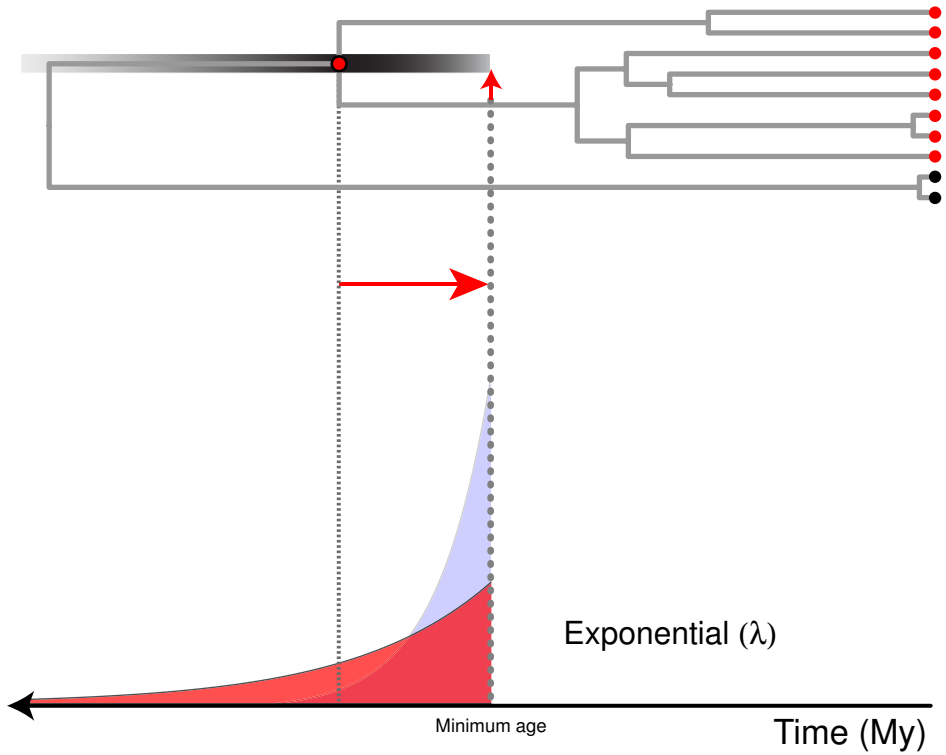
Prior Densities on Calibrated Nodes

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

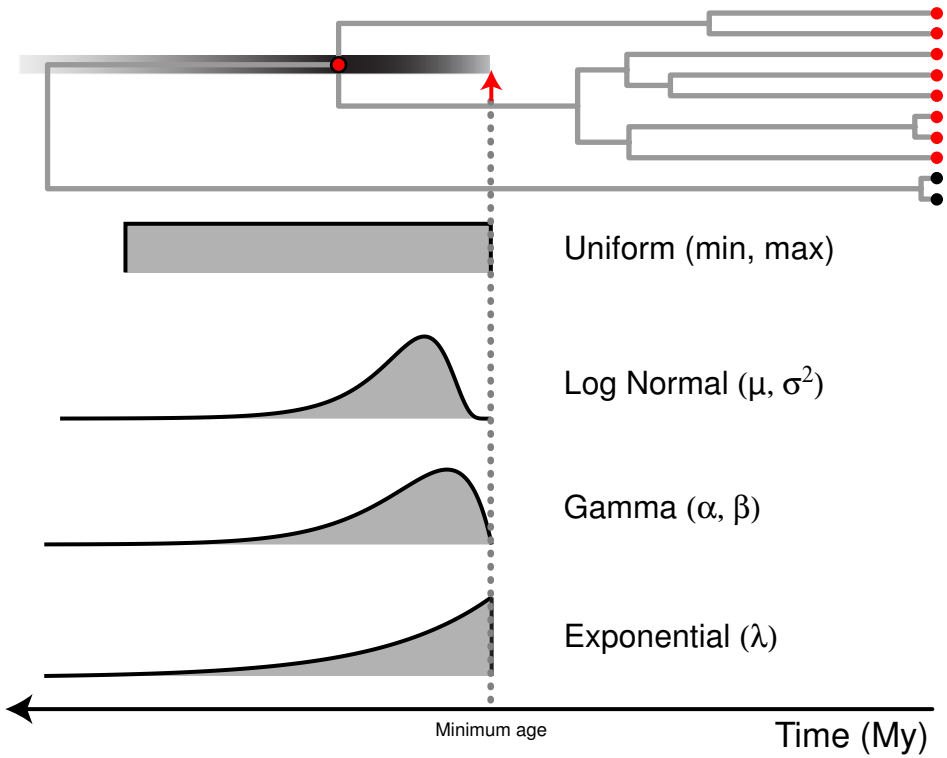


Prior Densities on Calibrated Nodes

Common practice in Bayesian divergence-time estimation:

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

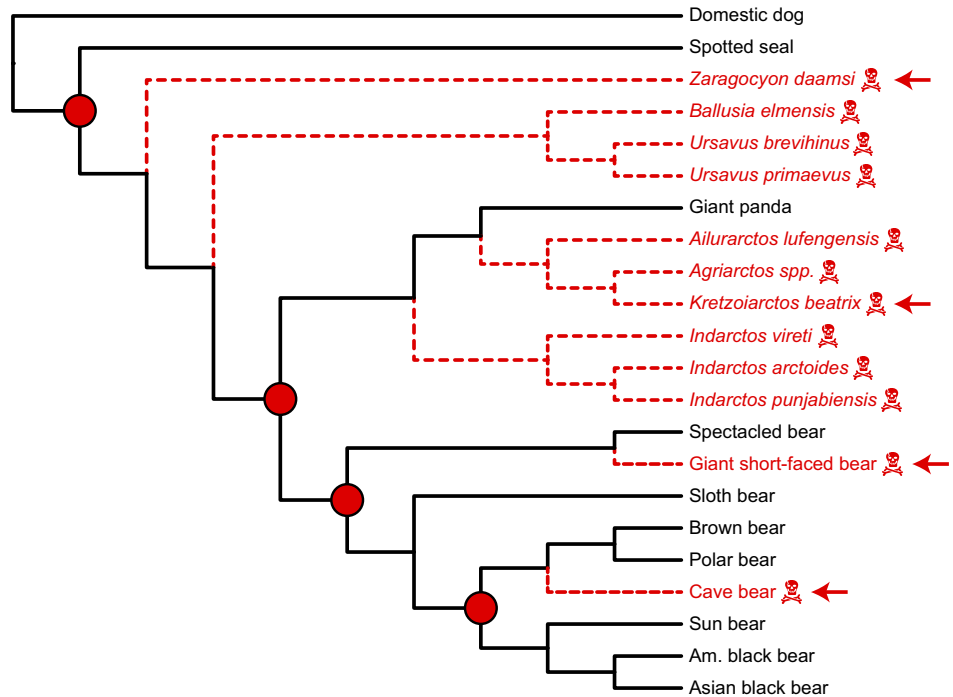
Specifying appropriate densities is a challenge for most molecular biologists



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

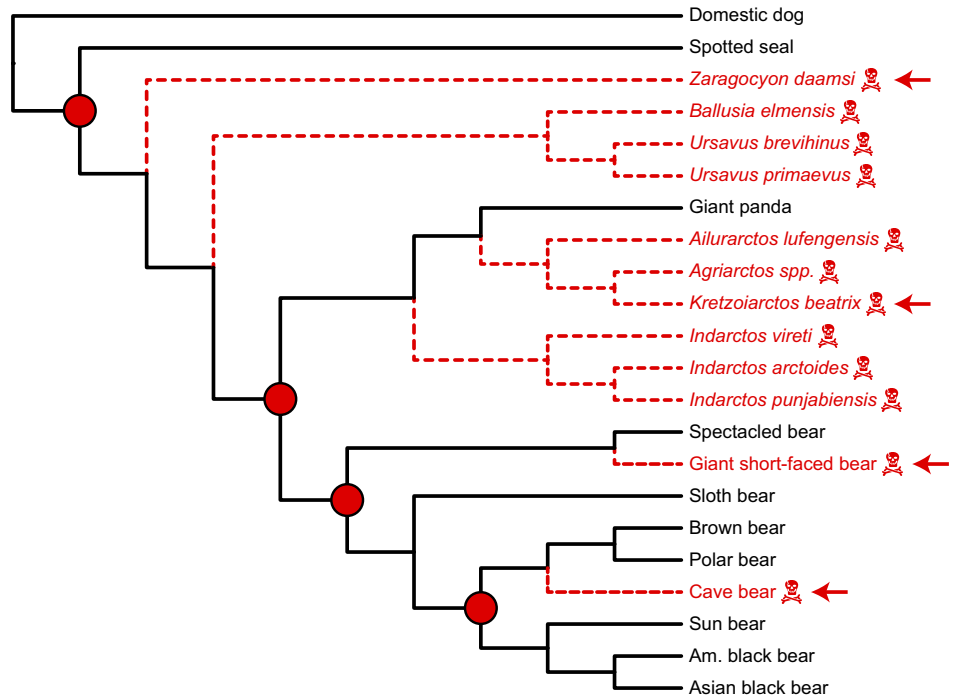


(Krause et al. *BMC Evol. Biol.* 2008; Abella et al. *PLoS ONE* 2012)

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

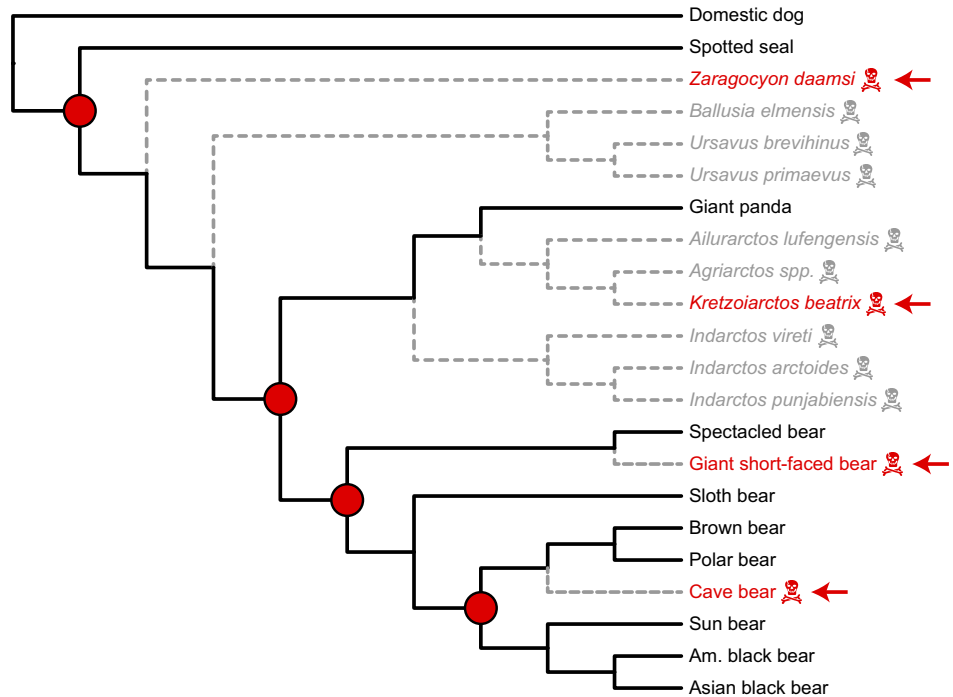


(Krause et al. *BMC Evol. Biol.* 2008; Abella et al. *PLoS ONE* 2012)

Improving Fossil Calibration

We want to use all of the available fossils

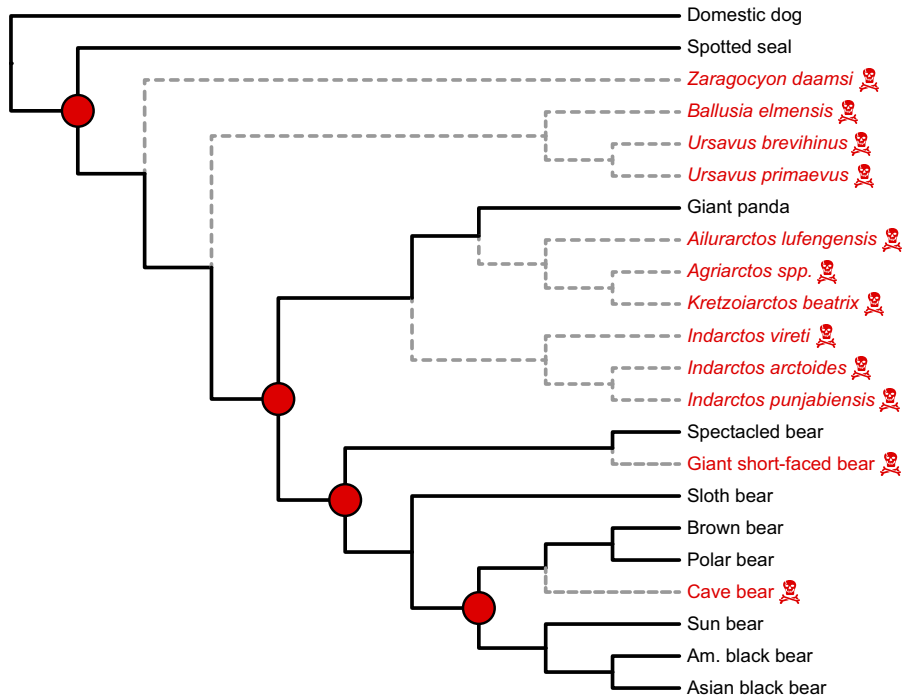
Example: Bears
12 fossils are reduced to 4 calibration ages with calibration density methods



(Krause et al. *BMC Evol. Biol.* 2008; Abella et al. *PLoS ONE* 2012)

Improving Fossil Calibration

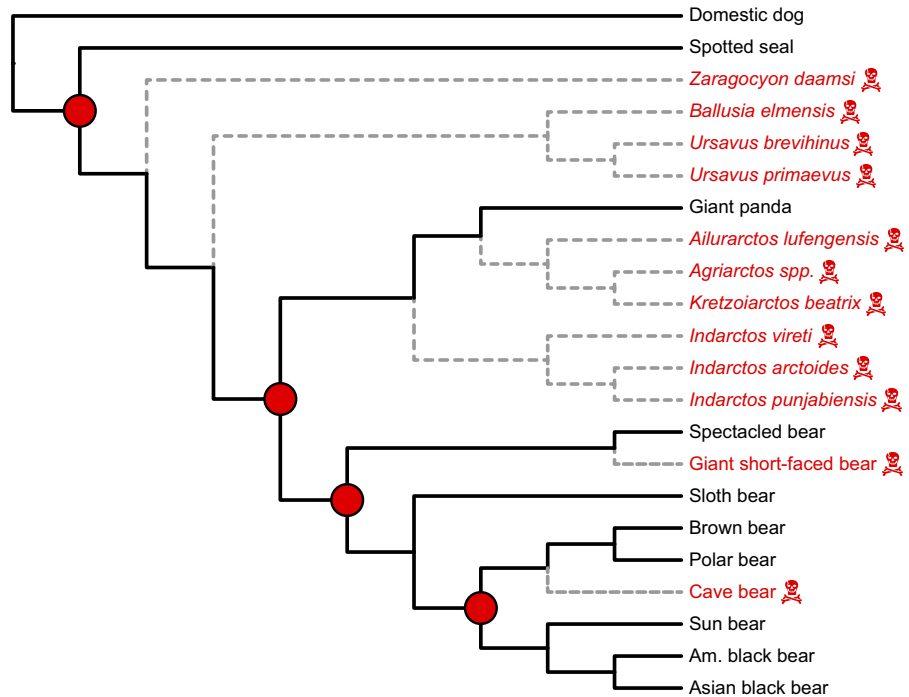
Because fossils are part of the diversification process, we can combine fossil calibration with birth-death models



(Krause et al. *BMC Evol. Biol.* 2008; Abella et al. *PLoS ONE* 2012)

Improving Fossil Calibration

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



(Krause et al. *BMC Evol. Biol.* 2008; Abella et al. *PLoS ONE* 2012)

The Fossilized Birth-Death Process (FBD)

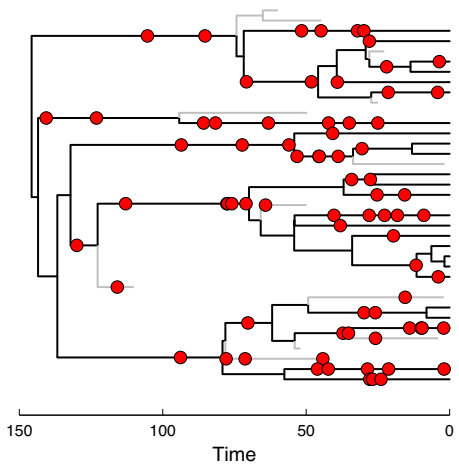
Improving statistical inference of absolute node ages

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

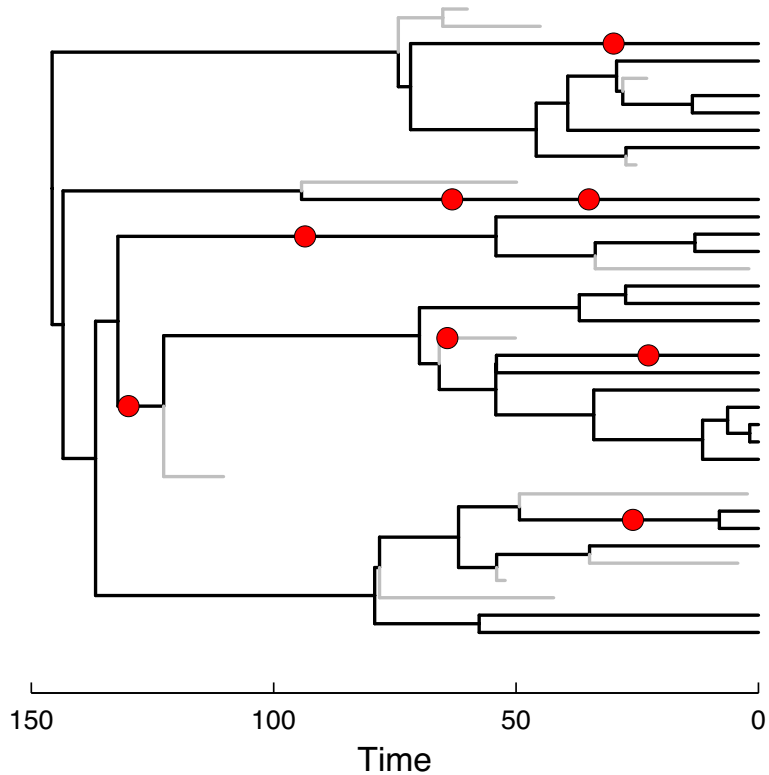
Useful for calibration or 'total-evidence' dating



(Heath, Huelsenbeck, Stadler. *PNAS* 2014)

The Fossilized Birth-Death Process (FBD)

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



(Heath, Huelsenbeck, Stadler. *PNAS* 2014)

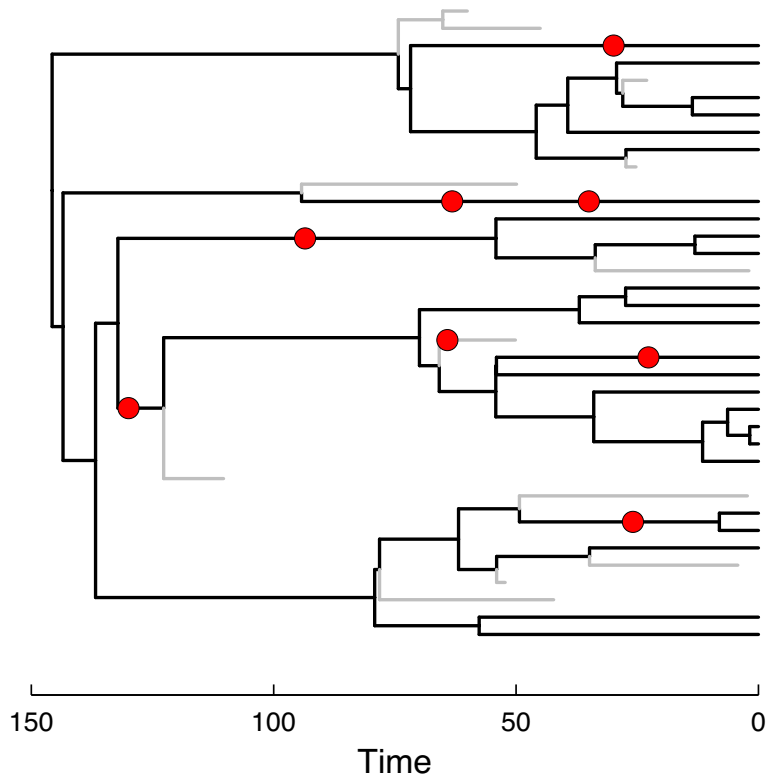
The Fossilized Birth-Death Process (FBD)

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

$\lambda =$ speciation

$\mu =$ extinction

$\psi =$ fossilization/recovery



(Heath, Huelsenbeck, Stadler. *PNAS* 2014)

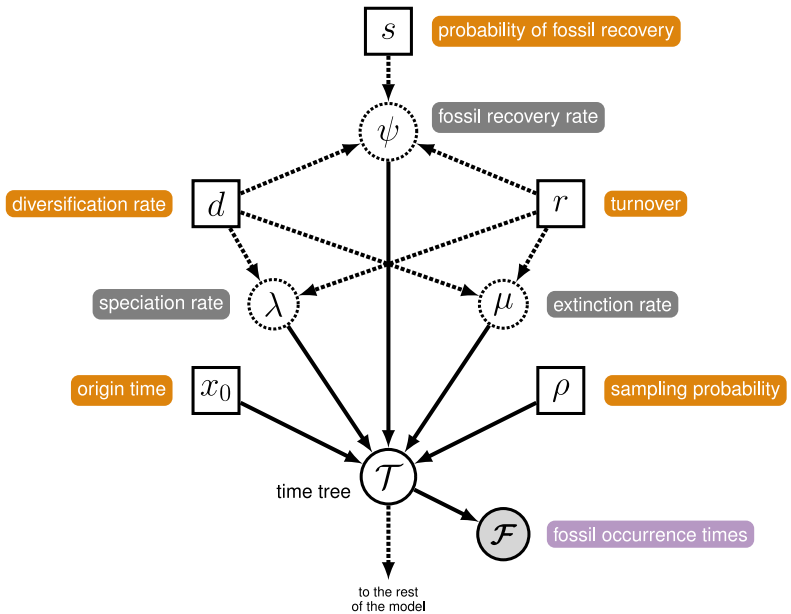
The Fossilized Birth-Death Process (FBD)

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

$\lambda =$ speciation

$\mu =$ extinction

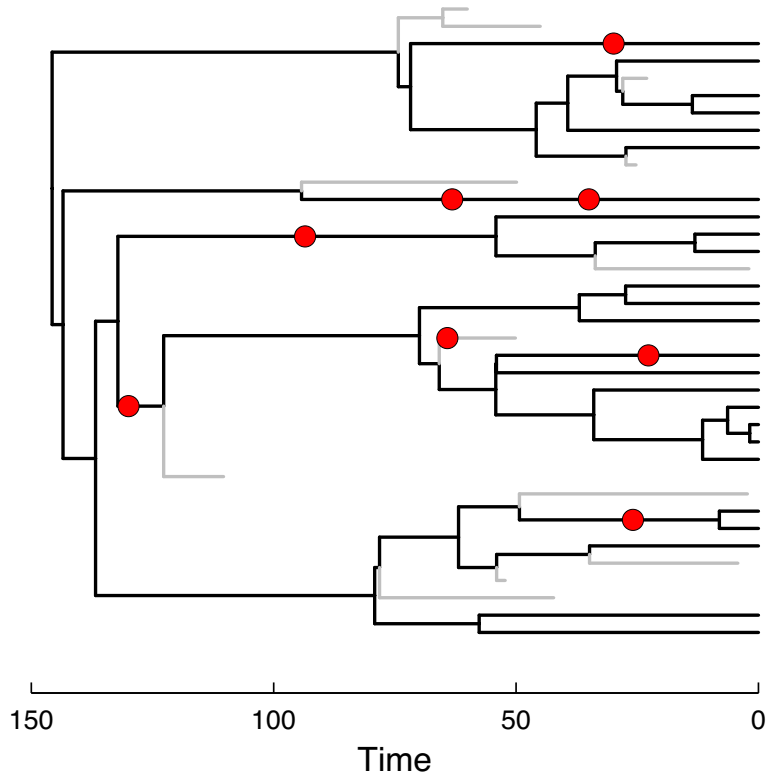
$\psi =$ fossilization/recovery



(Heath, Huelsenbeck, Stadler. *PNAS* 2014)

The Fossilized Birth-Death Process (FBD)

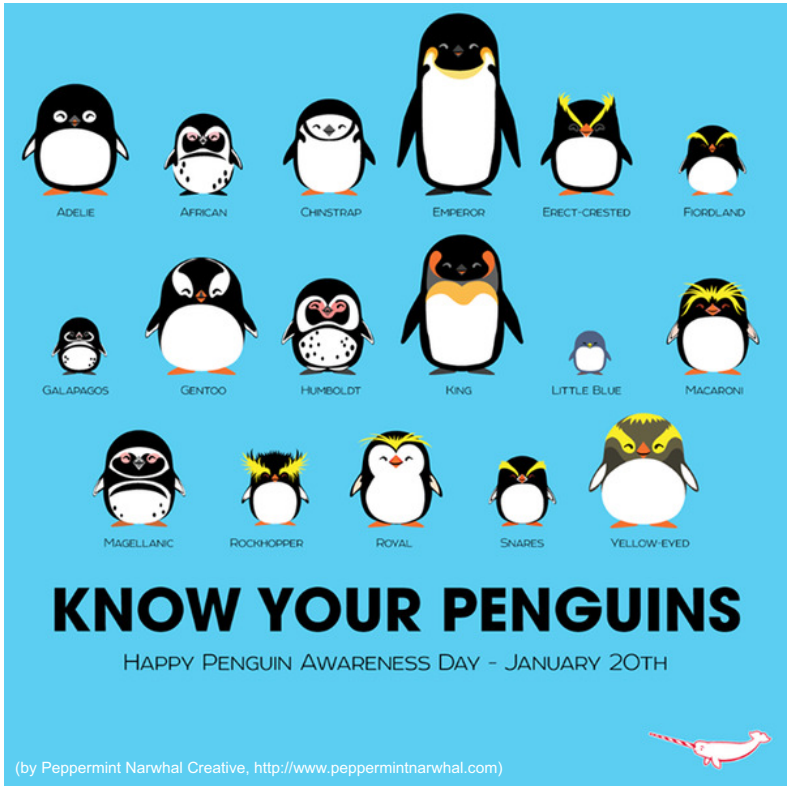
We use MCMC to sample realizations of the diversification process, integrating over the topology—including placement of the fossils—and speciation times

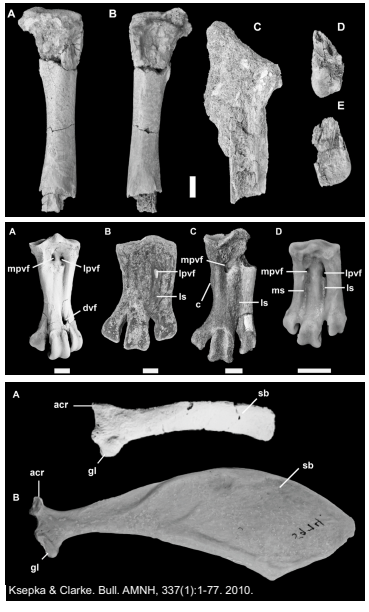


(Heath, Huelsenbeck, Stadler. *PNAS* 2014)

Penguin Diversity in Deep Time

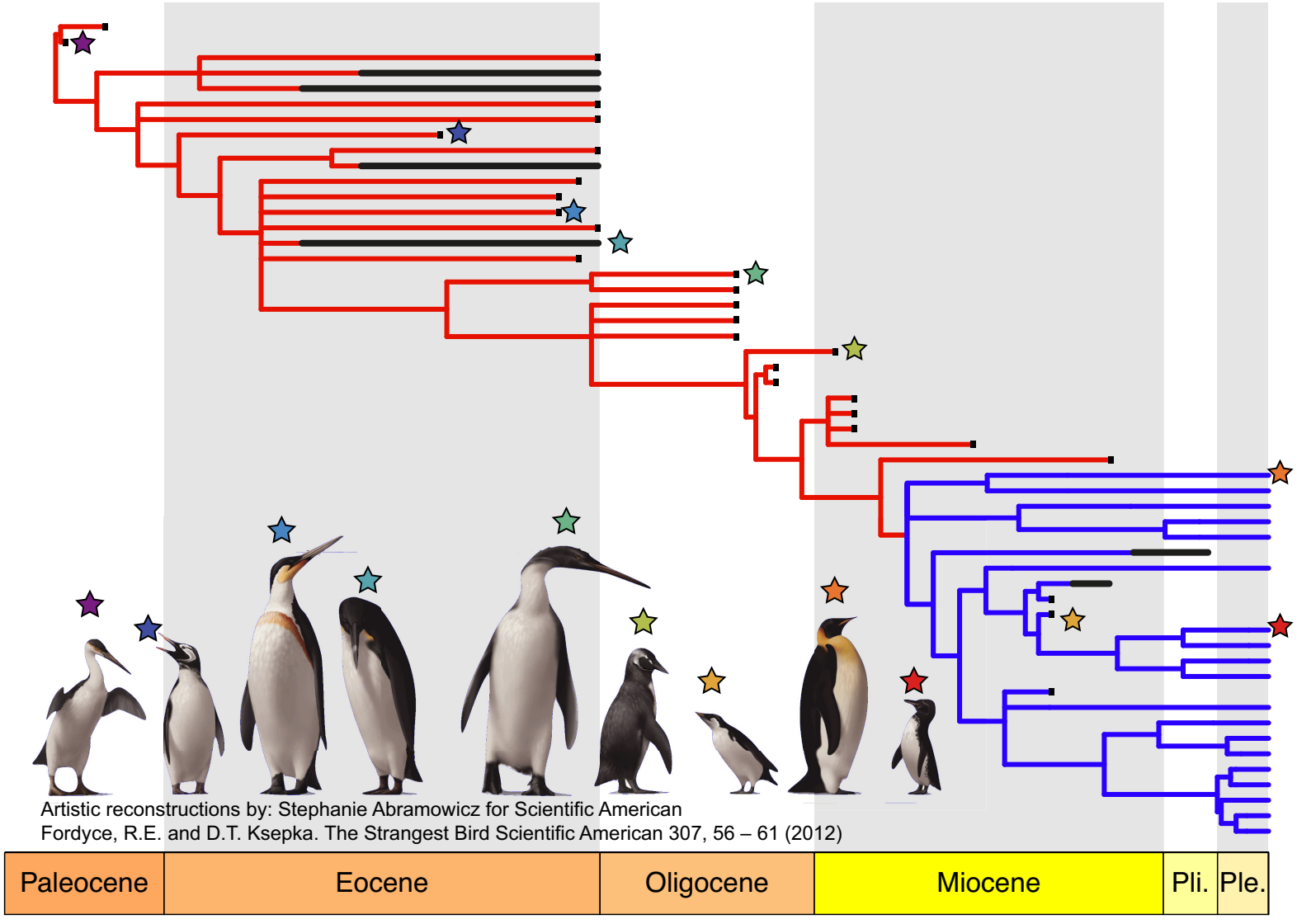
Can we improve our understanding of penguin evolution by considering both extant and fossil taxa?





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

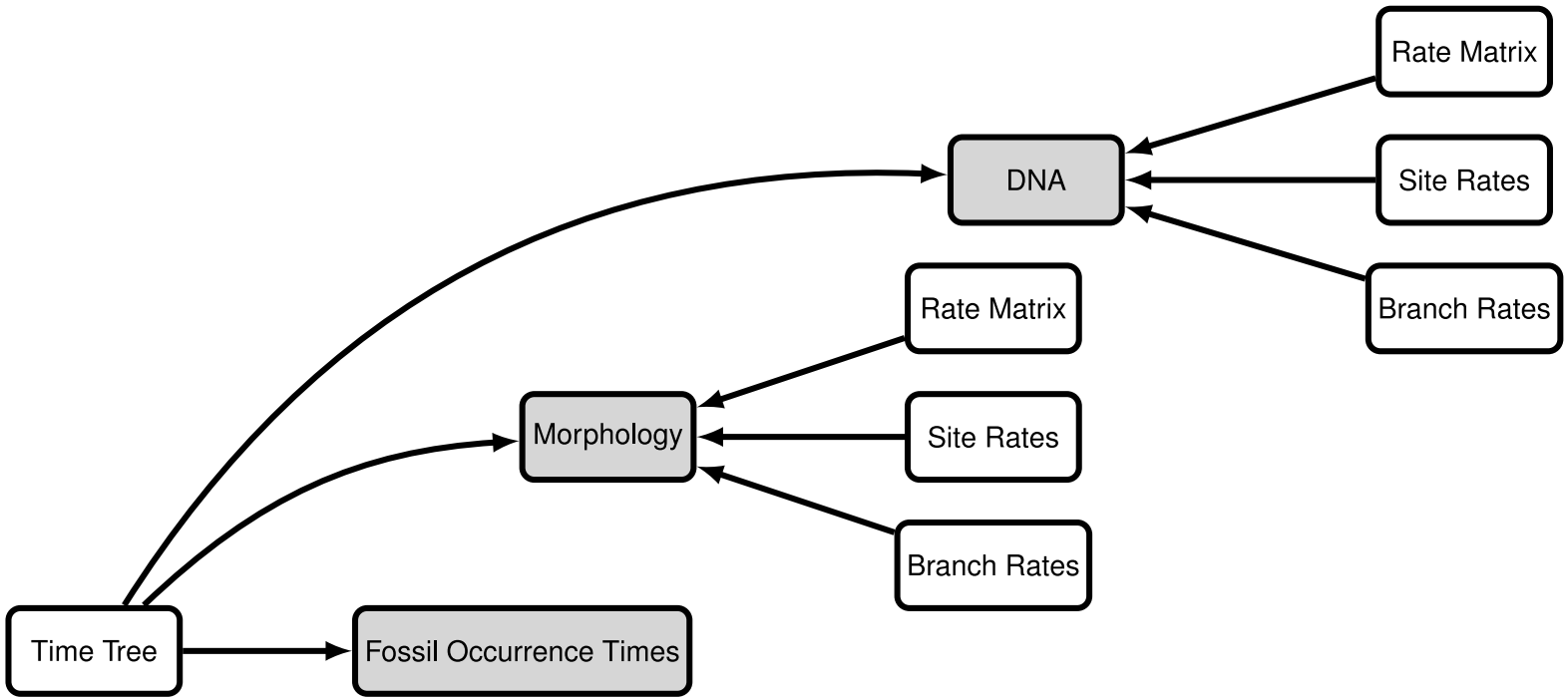
Penguin Diversity in Deep Time



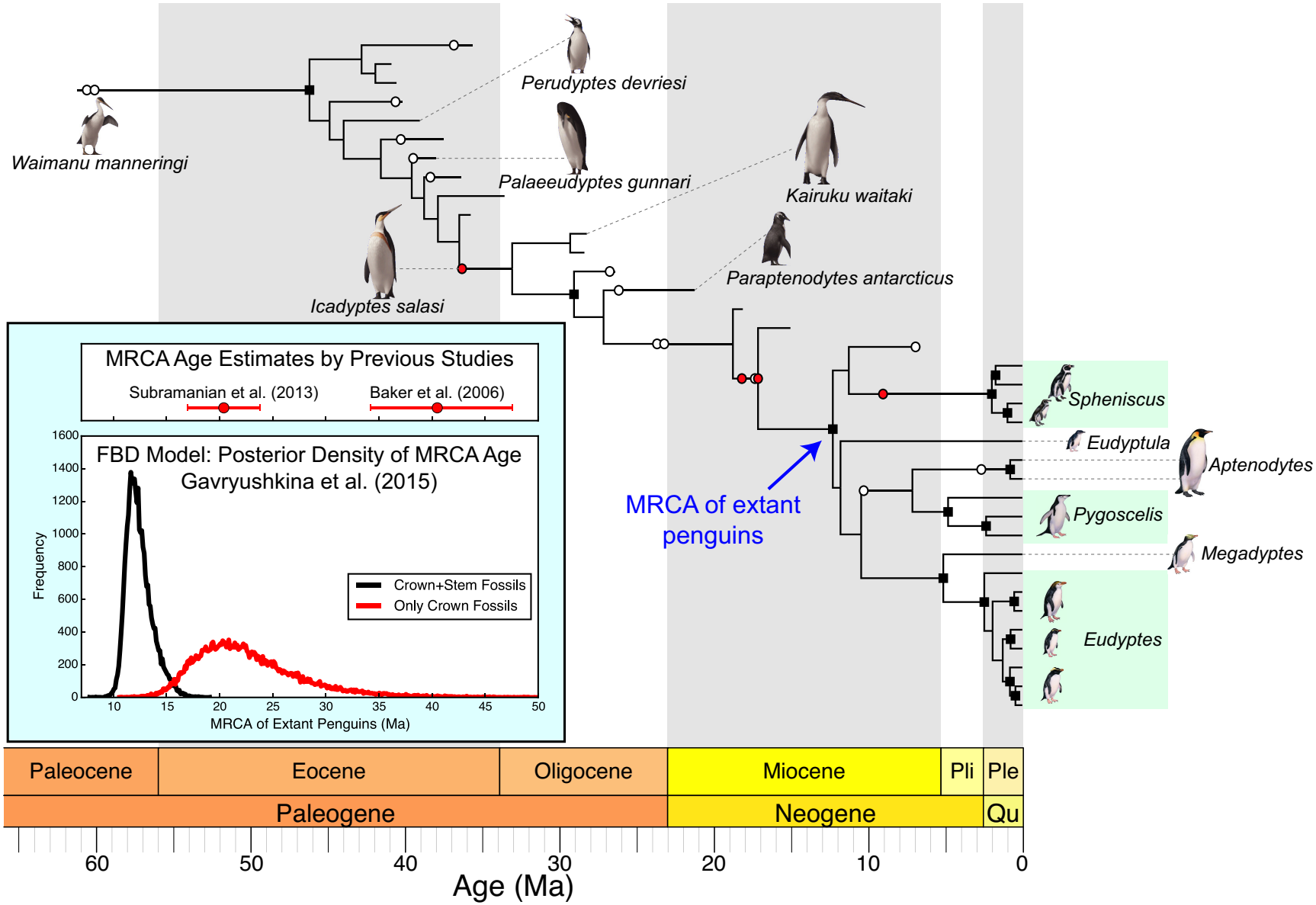
Artistic reconstructions by: Stephanie Abramowicz for Scientific American
Fordyce, R.E. and D.T. Ksepka. The Strangest Bird Scientific American 307, 56 – 61 (2012)

Integrative Bayesian Inference

Combine models for DNA sequence evolution, morphological change, and fossil recovery over time to jointly estimate the tree topology, divergence times, and lineage diversification rates



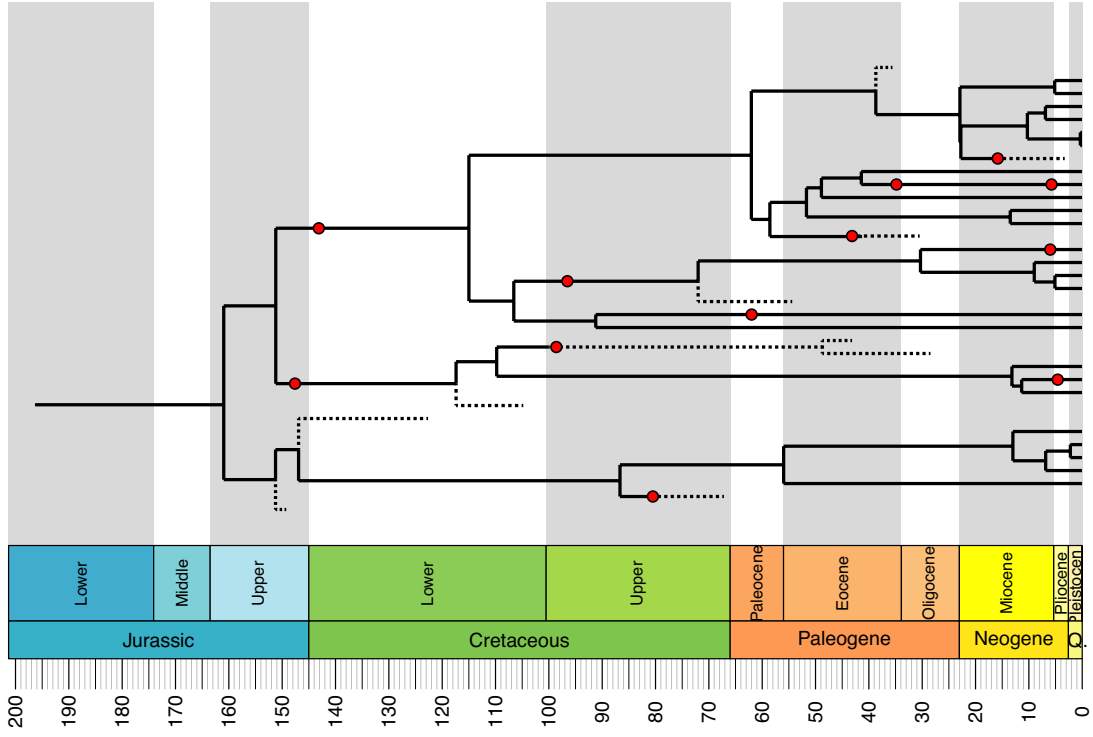
Penguin Diversity in Deep Time



(Gavruyshkina, Heath, Ksepka, Welch, Stadler, Drummond. 2015. http://arxiv.org/abs/1506.04797)

Inferring FBD Trees

Extensions of the fossilized birth-death process accommodate variation in fossil sampling, non-random species sampling, & shifts in diversification rates.

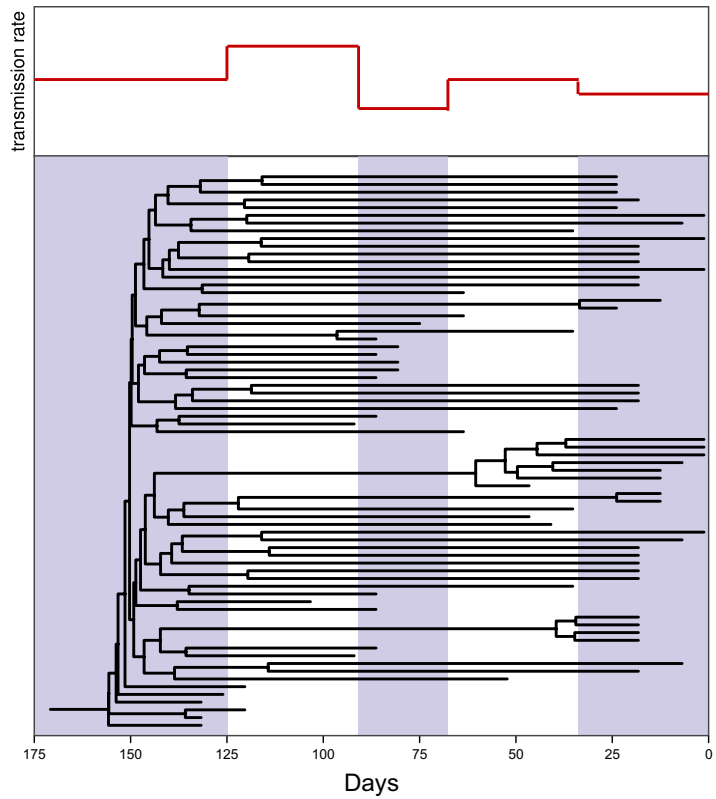


With character data for both fossil & extant species, we account for uncertainty in fossil placement

Skyline Birth-Death Process

A piecewise shifting model where parameters change over time

Used to estimate epidemiological parameters of an outbreak



Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)

Tanja Stadler[a,1,2], Denise Kühnert[b,c,1], Sebastian Bonhoeffer[a], and Alexei J. Drummond[b,c]

(see Stadler et al. *PNAS* 2013 and Stadler et al. PLoS Currents Outbreaks 2014)

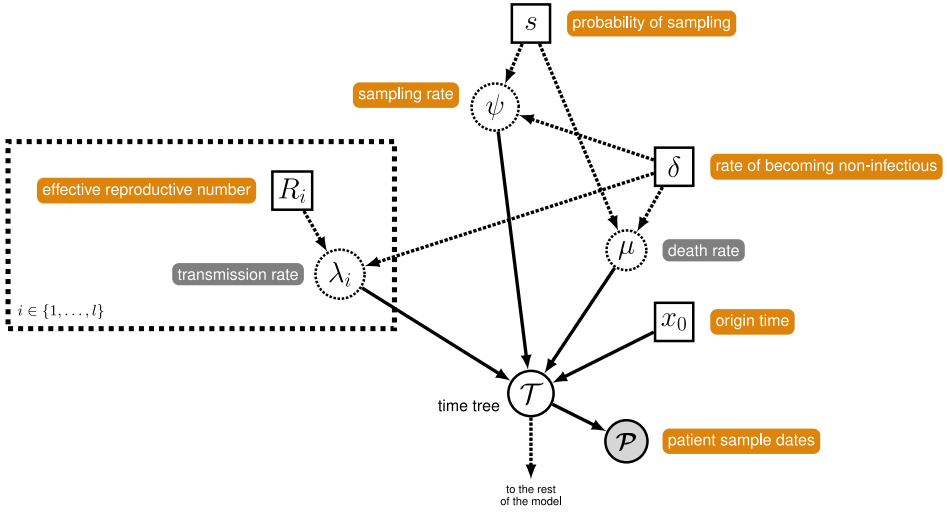
Skyline Birth-Death Process

l is the number of parameter intervals

R_i is the effective reproductive number for interval $i \in l$

δ is the rate of becoming non-infectious

s is the probability of sampling an individual after becoming non-infectious



$$R_i = \frac{\lambda_i}{\mu + \psi}, \quad \delta = \mu + \psi, \quad s = \frac{\psi}{\mu + \psi}$$

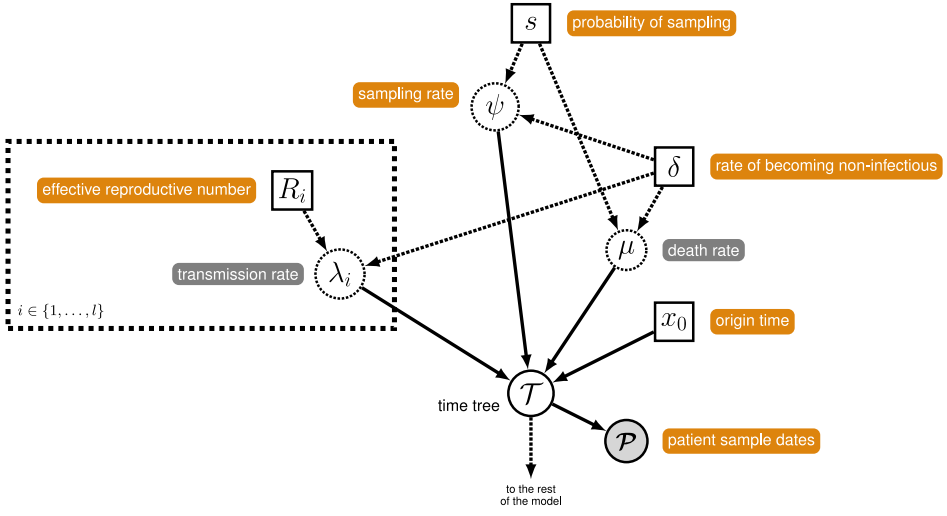
Skyline Birth-Death Process

l is the number of parameter intervals

λ_i is the transmission rate for interval $i \in l$

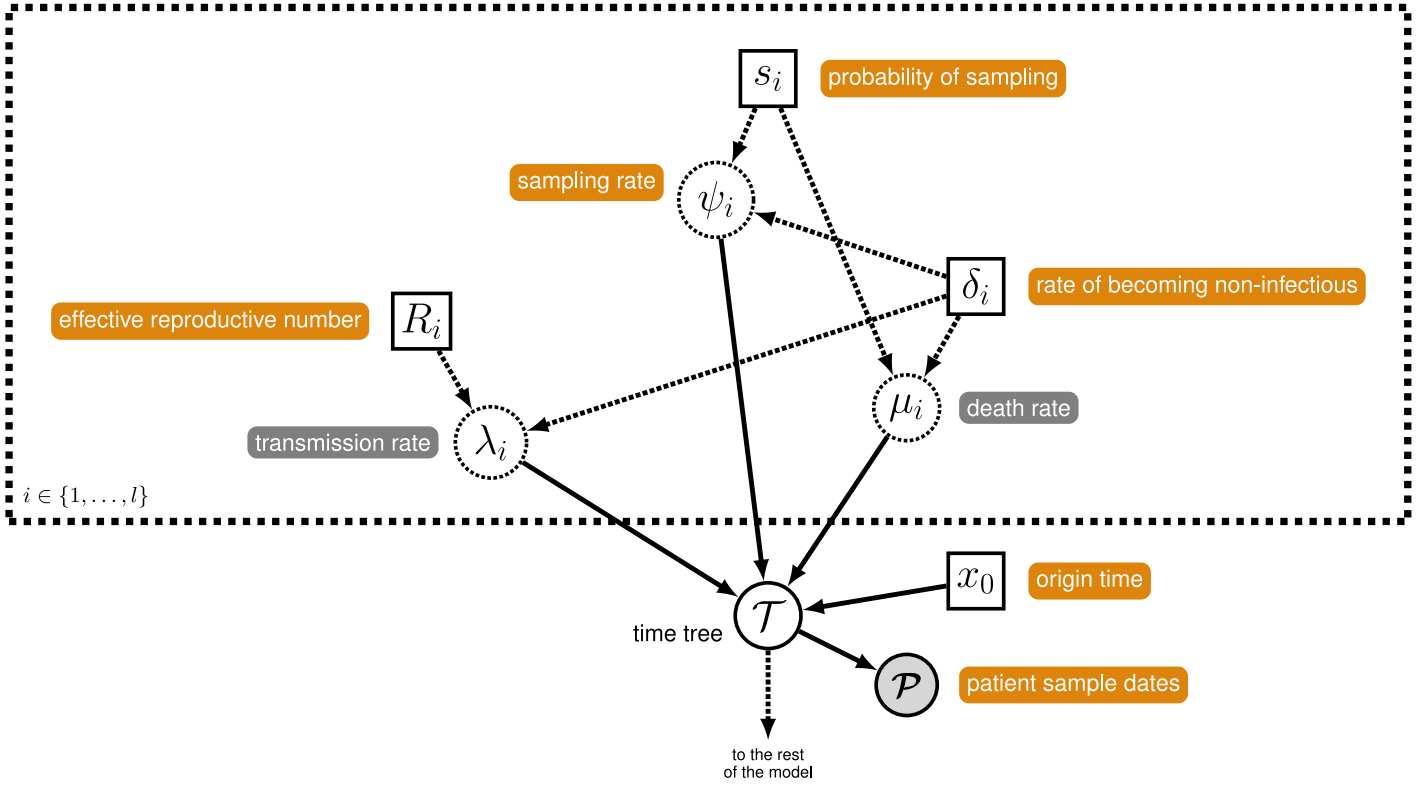
μ is the viral lineage death rate

ψ is the rate each individual is sampled



$$\lambda_i = R_i\delta, \quad \mu = \delta - s\delta, \quad \psi = s\delta$$

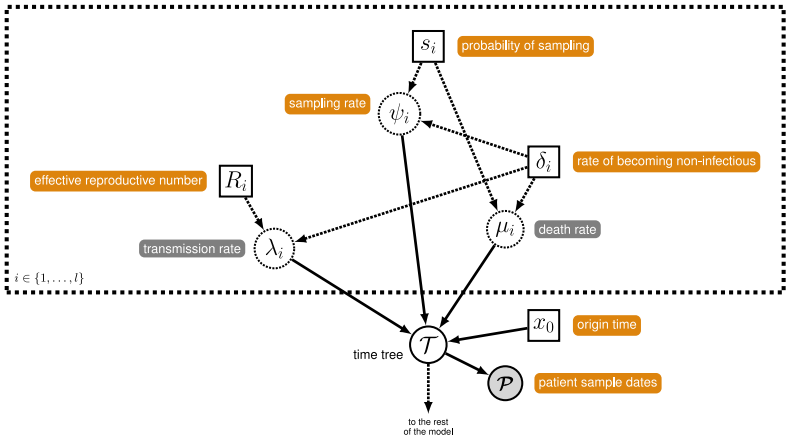
Skyline Birth-Death Process

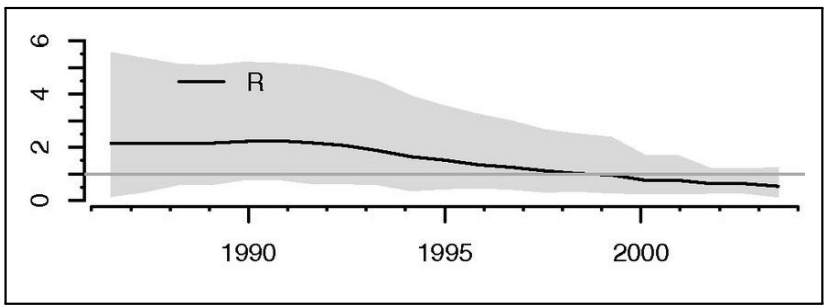


Skyline Birth-Death Process

A decline in *R* over the history of HIV-1 in the UK is consistent with the introduction of effective drug therapies

After 1998 *R* decreased below 1, indicating a declining epidemic



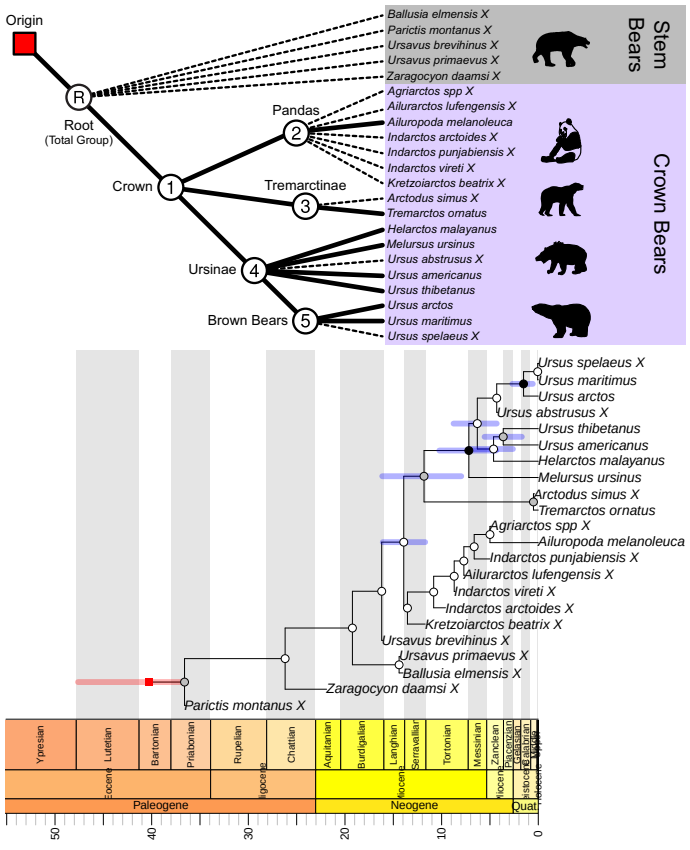


(Stadler et al. *PNAS* 2013)

Questions?

Exercises: Choose Your Own Adventure

Dating Bear Divergence Times with the Fossilized Birth-Death Process



Estimating Epidemiological Parameters of an Ebola Outbreak