# Phylogenetic Inference using RevBayes

*Diversification Rate Estimation with Missing Taxa*

Sebastian Höhna

## 1 Overview: Diversification Rate Estimation

Models of speciation and extinction are fundamental to any phylogenetic analysis of macroevolutionary processes (*e.g.,* divergence time estimation, diversification rate estimation, continuous and discrete trait evolution, and historical biogeography). First, a prior model describing the distribution of speciation events over time is critical to estimating phylogenies with branch lengths proportional to time. Second, stochastic branching models allow for inference of speciation and extinction rates. These inferences allow us to investigate key questions in evolutionary biology.

Diversification-rate parameters may be included as nuisance parameters of other phylogenetic models—*i.e.,* where these diversification-rate parameters are not of direct interest. For example, many methods for estimating species divergence times—such as `BEAST` (Drummond et al. 2012), `MrBayes` (Ronquist et al. 2012), and `RevBayes` (Höhna et al. 2016)—implement 'relaxed-clock models' that include a constant-rate birth-death branching process as a prior model on the distribution of tree topologies and node ages. Although the parameters of these 'tree priors' are not typically of direct interest, they are nevertheless estimated as part of the joint posterior probability distribution of the relaxed-clock model, and so can be estimated simply by querying the corresponding marginal posterior probability densities. In fact, this may provide more robust estimates of the diversification-rate parameters, as they accommodate uncertainty in the other phylogenetic-model parameters (including the tree topology, divergence-time estimates, and the other relaxed-clock model parameters). More recent work, *e.g.,* Heath et al. (2014), uses macroevolutionary models (the fossilized birth-death process) to calibrate phylogenies and thus to infer dated trees.

In these tutorials we focus on the different types of macroevolutionary models to study diversification processes and thus the diversification-rate parameters themselves. Nevertheless, these macroevolutionary models should be used for other evolutionary questions, when an appropriate prior distribution on the tree and divergence times is needed.

### 1.1 Types of Hypotheses for Estimating Diversification Rates

Many evolutionary phenomena entail differential rates of diversification (speciation − extinction); *e.g.,* adaptive radiation, diversity-dependent diversification, key innovations, and mass extinction. The specific study questions regarding lineage diversification may be classified within three fundamental categories of inference problems. Admittedly, this classification scheme is somewhat arbitrary, but it is nevertheless useful, as it allows users to navigate the ever-increasing number of available phylogenetic methods. Below, we describe each of the fundamental questions regarding diversification rates.

**(1) Diversification-rate through time estimation** *What is the (constant) rate of diversification in my study group?* The most basic models estimate parameters of the stochastic-branching process (*i.e.,* rates of speciation and extinction, or composite parameters such as net-diversification and relative-extinction rates) under the assumption that rates have remained constant across lineages and through time; *i.e.,* under a constant-rate birth-death stochastic-branching process model. Extensions to the (basic) constant-rate models include diversification-rate variation through time. First, we might ask whether there is evidence of an episodic, tree-wide increase in diversification rates (associated with a sudden increase in speciation rate and/or decrease in extinction rate), as might occur during an episode of adaptive radiation. A second question asks whether there is evidence of a continuous/gradual decrease in diversification rates through time (associated with decreasing speciation rates and/or increasing extinction rates), as might occur because of diversity-dependent diversification (*i.e.,* where competitive ecological interactions among the species of a growing tree decrease the opportunities for speciation and/or increase the probability of extinction). A final question in this category asks whether our study tree was impacted by a mass-extinction event (where a large fraction of the standing species diversity is suddenly lost). The common theme of these studies is that the diversification process is tree-wide, that is, all lineages of the study group have the exact same rates at a given time.

**(2) Diversification-rate variation across branches estimation** *Is there evidence that diversification rates have varied significantly across the branches of my study group?* Models have been developed to detect departures from rate constancy across lineages; these tests are analogous to methods that test for departures from a molecular clock—*i.e.,* to assess whether substitution rates vary significantly across lineages. These models are important for assessing whether a given tree violates the assumptions of rate homogeneity among lineages. Furthermore, these models are important to answer questions such as: *What are the branch-specific diversification rates?*; and *Have there been significant diversification-rate shifts along branches in my study group, and if so, how many shifts, what magnitude of rate-shifts and along which branches?*

**(3) Character-dependent diversification-rate estimation** *Are diversification rates correlated with some variable in my study group?* Character-dependent diversification-rate models aim to identify overall correlations between diversification rates and organismal features (binary and multi-state discrete morphological traits, continuous morphological traits, geographic range, etc.). For example, one can hypothesize that a binary character, say if an organism is herbivorous/carnivorous or self-compatible/self-incompatible, impact the diversification rates. Then, if the organism is in state 0 (*e.g.,* is herbivorous) it has a lower (or higher) diversification rate than if the organism is in state 1 (*e.g.,* carnivorous).

# 2 Models

We begin this section with a general introduction to the stochastic birth-death branching process that underlies inference of diversification rates in `RevBayes`. This primer will provide some details on the relevant theory of stochastic-branching process models. We appreciate that some readers may want to skip this somewhat technical primer; however, we believe that a better understanding of the relevant theory provides a foundation for performing better inferences. We then discuss a variety of specific birth-death models, but emphasize that these examples represent only a tiny fraction of the possible diversification-rate models that can be specified in `RevBayes`.

## 2.1 The birth-death branching process

Our approach is based on the *reconstructed evolutionary process* described by Nee et al. (1994); a birth-death process in which only sampled, extant lineages are observed. Let $N(t)$ denote the number of species at time $t$. Assume the process starts at time $t_1$ (the 'crown' age of the most recent common ancestor of the study group, $t_{\text{MRCA}}$) when there are two species. Thus, the process is initiated with two species, $N(t_1) = 2$. We condition the process on sampling at least one descendant from each of these initial two lineages; otherwise $t_1$ would not correspond to the $t_{\mathrm{MRCA}}$ of our study group. Each lineage evolves independently of all other lineages, giving rise to exactly one new lineage with rate $b(t)$ and losing one existing lineage with rate $d(t)$ (Figure 1 and Figure 2). Note that although each lineage evolves independently, all lineages share both a common (tree-wide) speciation rate $b(t)$ and a common extinction rate $d(t)$ (Nee et al. 1994; Höhna 2015). Additionally, at certain times, $t_{\mathbb{M}}$, a mass-extinction event occurs and each species existing at that time has the same probability, $\rho$, of survival. Finally, all extinct lineages are pruned and only the reconstructed tree remains (Figure 1).

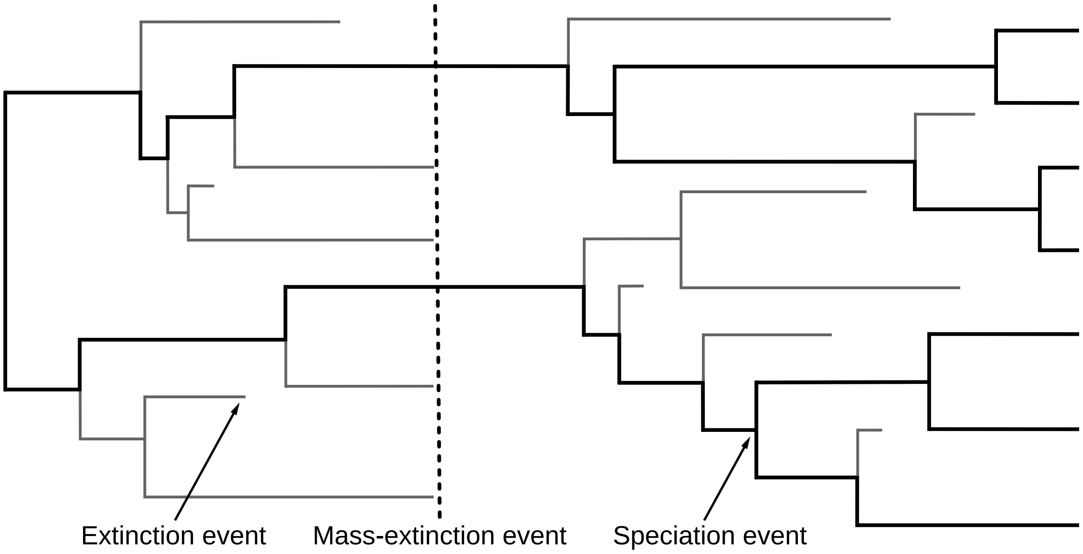


Figure 1: A realization of the birth-death process with mass extinction. Lineages that have no extant or sampled descendant are shown in gray and surviving lineages are shown in a thicker black line.

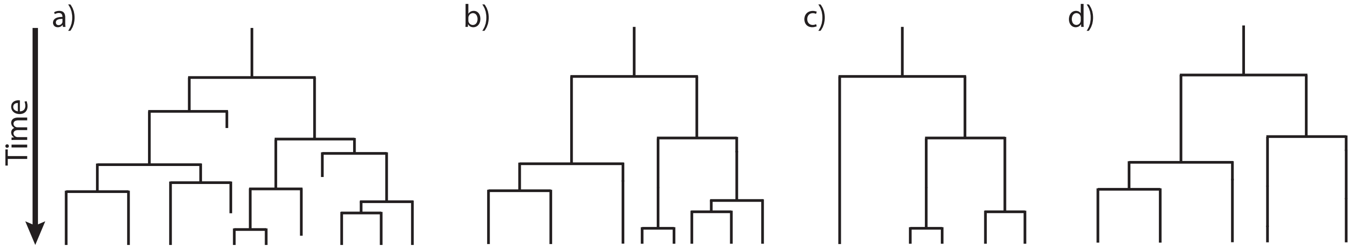


Figure 2: **Examples of trees produced under a birth-death process.** The process is initiated at the first speciation event (the ‘crown-age’ of the MRCA) when there are two initial lineages. At each speciation event the ancestral lineage is replaced by two descendant lineages. At an extinction event one lineage simply terminates. (A) A complete tree including extinct lineages. (B) The reconstructed tree of tree from A with extinct lineages pruned away. (C) A *uniform* subsample of the tree from B, where each species was sampled with equal probability, $\rho$. (D) A *diversified* subsample of the tree from B, where the species were selected so as to maximize diversity.

To condition the probability of observing the branching times on the survival of both lineages that descend from the root, we divide by $P(N(T) > 0|N(0) = 1)^2$. Then, the probability density of the branching times,

$\mathbb{T}$, becomes

$$P(\mathbb{T}) = \frac{\overbrace{P(N(T)=1 \mid N(0)=1)^2}^{\text{both initial lineages have one descendant}}}{\underbrace{P(N(T)>0 \mid N(0)=1)^2}_{\text{both initial lineages survive}}} \times \prod_{i=2}^{n-1} \overbrace{i \times b(t_i)}^{\text{speciation rate}} \times \overbrace{P(N(T)=1 \mid N(t_i)=1)}^{\text{lineage has one descendant}},$$

and the probability density of the reconstructed tree (topology and branching times) is then

$$\begin{aligned} P(\Psi) = & \; \frac{2^{n-1}}{n!(n-1)!} \times \left( \frac{P(N(T)=1 \mid N(0)=1)}{P(N(T)>0 \mid N(0)=1)} \right)^2 \\ & \times \prod_{i=2}^{n-1} i \times b(t_i) \times P(N(T)=1 \mid N(t_i)=1) \end{aligned} \tag{1}$$

We can expand Equation (1) by substituting $P(N(T) > 0 \mid N(t) = 1)^2 \exp(r(t,T))$ for $P(N(T) = 1 \mid N(t) = 1)$, where $r(u,v) = \int_u^v d(t) - b(t)dt$; the above equation becomes

$$\begin{aligned} P(\Psi) = & \; \frac{2^{n-1}}{n!(n-1)!} \times \left( \frac{P(N(T)>0 \mid N(0)=1)^2 \exp(r(0,T))}{P(N(T)>0 \mid N(0)=1)} \right)^2 \\ & \times \prod_{i=2}^{n-1} i \times b(t_i) \times P(N(T)>0 \mid N(t_i)=1)^2 \exp(r(t_i,T)) \\ = & \; \frac{2^{n-1}}{n!} \times \Big( P(N(T)>0 \mid N(0)=1) \exp(r(0,T)) \Big)^2 \\ & \times \prod_{i=2}^{n-1} b(t_i) \times P(N(T)>0 \mid N(t_i)=1)^2 \exp(r(t_i,T)). \end{aligned} \tag{2}$$

For a detailed description of this substitution, see Höhna (2015). Additional information regarding the underlying birth-death process can be found in (Thompson 1975; Equation 3.4.6) and Nee et al. (1994) for constant rates and Höhna (2013; 2014; 2015) for arbitrary rate functions.

To compute the equation above we need to know the rate function, $r(t,s) = \int_t^s d(x) - b(x)dx$, and the probability of survival, $P(N(T)\!>\!0|N(t)\!=\!1)$. Yule (1925) and later Kendall (1948) derived the probability that a process survives ($N(T) > 0$) and the probability of obtaining exactly $n$ species at time $T$ ($N(T) = n$) when the process started at time $t$ with one species. Kendall's results were summarized in Equation (3) and Equation (24) in Nee et al. (1994)

$$P(N(T)\!>\!0|N(t)\!=\!1) \;=\; \left(1 + \int_t^T \Big(\mu(s) \exp(r(t,s))\Big) ds \right)^{-1} \tag{3}$$

$$\begin{aligned} P(N(T)\!=\!n|N(t)\!=\!1) \;=\; & (1 - P(N(T)\!>\!0|N(t)\!=\!1) \exp(r(t,T)))^{n-1} \\ & \times P(N(T)\!>\!0|N(t)\!=\!1)^2 \exp(r(t,T)) \end{aligned} \tag{4}$$

An overview for different diversification models is given in Höhna (2015).

***Sidebar: Phylogenetic trees as observations***

The branching processes used here describe probability distributions on phylogenetic trees. This probability distribution can be used to infer diversification rates given an "observed" phylogenetic tree. In reality we never observe a phylogenetic tree itself. Instead, phylogenetic trees themselves are estimated from actual observations, such as DNA sequences. These phylogenetic tree estimates, especially the divergence times, can have considerable uncertainty associated with them. Thus, the correct approach for estimating diversification rates is to include the uncertainty in the phylogeny by, for example, jointly estimating the phylogeny and diversification rates. For the simplicity of the following tutorials, we take a shortcut and assume that we know the phylogeny without error. For publication quality analysis you should always estimate the diversification rates jointly with the phylogeny and divergence times.

# 3 Estimating Speciation & Extinction Rates Through Time

## 3.1 Outline

This tutorial describes how to specify different models of incomplete taxon sampling (Höhna et al. 2011; Höhna 2014) for estimating diversification rates in RevBayes (Höhna et al. 2016). Incomplete taxon sampling, if not modeled correctly, severely biases diversification-rate parameter estimates (Cusimano and Renner 2010; Höhna et al. 2011). Specifically, we will discuss *uniform*, *diversified*, and *empirical* taxon sampling. All analyses in this tutorial will focus on diversification rate estimation through-time and use a birth-death process where diversification rates vary episodically which we model by piecewise constant rates RevBayes (Höhna 2015; May et al. 2016). The probabilistic graphical model is given only once for this tutorial as an overview. The model itself does not change between the different analyses; only the assumptions of incomplete taxon sampling. For each analysis you will estimate speciation and extinction rates through-time using Markov chain Monte Carlo (MCMC) and assess the impact of incomplete taxon sampling as well as the sampling scheme.

## 3.2 Requirements

We assume that you have read and hopefully completed the following tutorials:

- Getting started
- Rev basics
- Basic Diversification Rate Estimation
- Diversification Rates Through Time

Note that the Rev basics tutorial introduces the basic syntax of Rev but does not cover any phylogenetic models. You may skip the Rev basics tutorial if you have some familiarity with R. We tried to keep this tutorial very basic and introduce all the language concepts and theory on the way. You may only need the Rev basics tutorial for a more in-depth discussion of concepts in Rev.

For this tutorial is especially important that you have read the two tutorials on diversification rate estimation: Basic Diversification Rate Estimation tutorial and Diversification Rates Through Time tutorial.

Specifically the Diversification Rates Through Time tutorial present the underlying diversification model and thus foundation for this tutorial. Here we will build on the tutorial by modifying the assumptions of incomplete taxon sampling in different ways.

## 4 Data and files

We provide the data file(s) which we will use in this tutorial. You may want to use your own data instead. In the **`data`** folder, you will find the following file

- **`primates.tre`**: Dated primates phylogeny including 23 out of 377 species.

Note that we use here the small primate phylogeny including only 23 of the 377 taxa instead of the much more complete primate phylogeny from Springer et al. (2012). This choice was solely made to emphasize the point and impact of incomplete taxon sampling, which is a very prominent feature in many large scale phylogenies.

→ Open the tree **`data/primates.tre`** in `FigTree`.

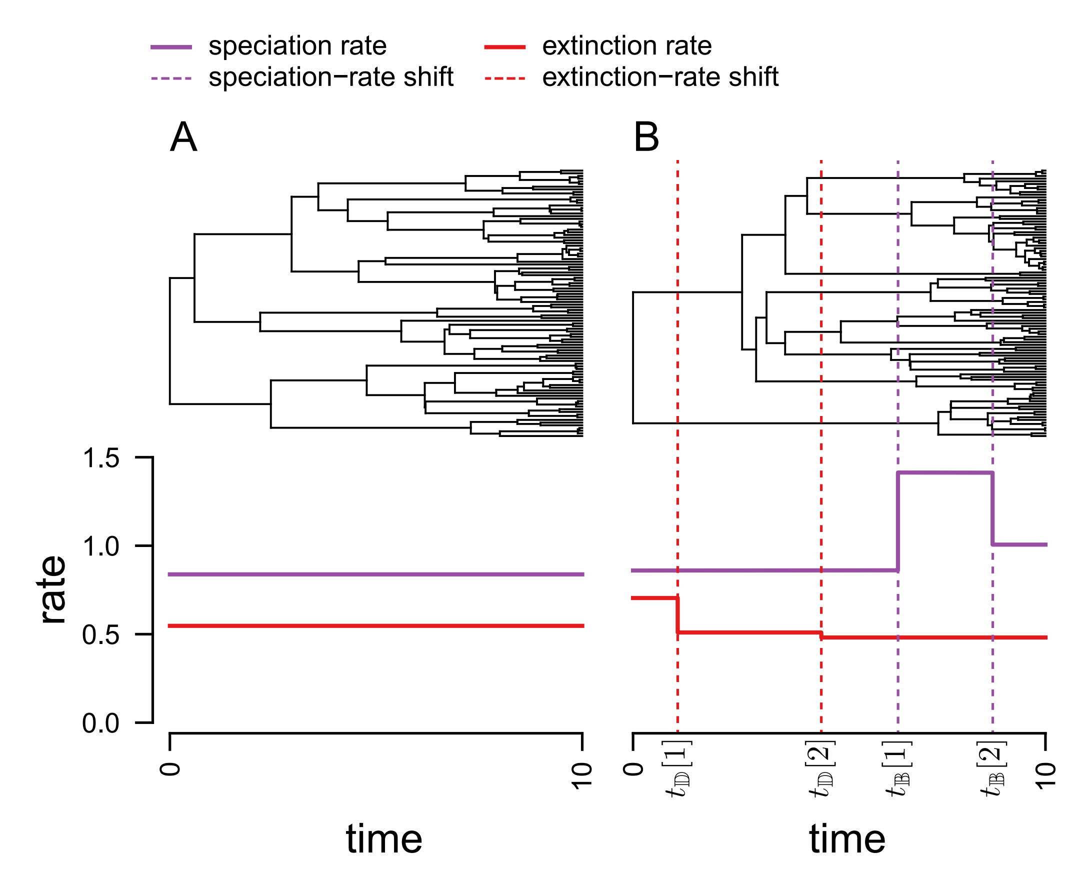
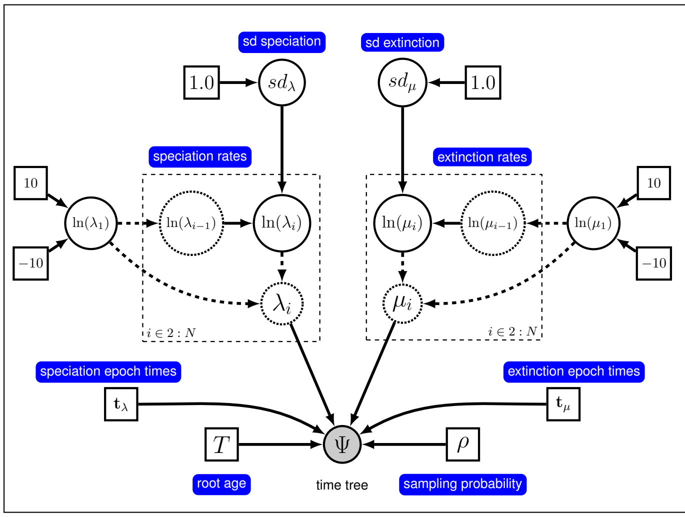


Figure 3: Two scenarios of birth-death models. On the left we show constant diversification. On the right we show an example of an episodic birth-death process where rates are constant in each time interval (epoch). The top panel of this figure shows example realization under the given rates.

# 5 Episodic Birth-Death Model

Here we study the impact of incomplete taxon sampling by estimating diversification rates through time. The goal is to compare the impact of different taxon sampling strategies rather than the description of the diversification-rate model itself. The episodic birth-death model used here is equivalent to the model described in our previous tutorial. Please read the Diversification Rates Through Time tutorial for more detailed information about the model.

We have included in Figure 3 again the cartoon of episodic birth-death process with piecewise constant diversification rates. As mentioned above, diversification rate estimates are biased when only a fraction of the species is included and the sampling scheme is not accommodated appropriately (Cusimano and Renner 2010; Höhna et al. 2011; Cusimano et al. 2012; Höhna 2014). Hence, the diversification rate through time model will be an excellent example to study the impact of the assumed incomplete sampling strategy on diversification rates.



```
speciation_sd ~ dnExponential(1.0)
extinction_sd ~ dnExponential(1.0)

log_speciation[1] ~ dnUniform(-10.0,10.0)
log_extinction[1] ~ dnUniform(-10.0,10.0)

speciation[1] := exp( log_speciation[1] )
extinction[1] := exp( log_extinction[1] )

for (i in 1:NUM_INTERVALS) {
     index = i+1

     log_speciation[index] ~ dnNormal( mean=log_speciation[i], sd=speciation_sd )
     log_extinction[index] ~ dnNormal( mean=log_extinction[i], sd=extinction_sd )

     speciation[index] := exp( log_speciation[index] )
     extinction[index] := exp( log_extinction[index] )
}

times_speciation <- T.rootAge() * (1:NUM_INTERVALS) / (NUM_INTERVALS) * 0.8
times_extinction <- T.rootAge() * (1:NUM_INTERVALS) / (NUM_INTERVALS) * 0.8

rho <- T.ntips()/377

timetree ~ dnEpisodicBirthDeath(rootAge=T.rootAge(),
                                lambdaRates=speciation, lambdaTimes=times_speciation,
                                muRates=extinction, muTimes=times_extinction,
                                rho=rho, samplingStrategy="uniform",
                                condition="time", taxa=taxa)

timetree.clamp(T)
```

Figure 4: A graphical model with the outline of the Rev code. On the left we see the graphical model describing the correlated (Brownian motion) model for rate-variation through time. On the right we show the correspond Rev commands to instantiate this model in computer memory. This figure gives a complete overview of the model that we use here in this analysis.

We additionally include the graphical model representing the episodic birth-death process with autocorrelated diversification rates. This graphical model shows you which variables are included in the model, and the dependency between the variables. Thus, it makes the structure and assumption of the model clear and visible instead of a black-box (Höhna et al. 2014). Here we will focus only on the variable **rho**, the sampling probability, to model incomplete taxon sampling.

## 5.1 Specifying the model in Rev

We will give a very brief and compressed version of the model with fewer comments and explanation. The more detailed explanation can be found in the Diversification Rates Through Time tutorial. Any attempt from us to present the full description here would only be a duplication/copy of the original tutorial with the additional to be less complete and less up to date.

Here are the summarized steps for running the episodic birth-death model in Rev.

```
#######################
# Reading in the Data #
#######################

### Read in the "observed" tree
T <- readTrees("data/primates.tre")[1]

# Get some useful variables from the data. We need these later on.
taxa <- T.taxa()

# set my move index
mvi = 0
mni = 0

NUM_INTERVALS = 10



####################
# Create the rates #
####################

# first we create the standard deviation of the rates between intervals
# draw the sd from an exponential distribution
speciation_sd ~ dnExponential(1.0)
moves[++mvi] = mvScale(speciation_sd,weight=5.0)

extinction_sd ~ dnExponential(1.0)
moves[++mvi] = mvScale(extinction_sd,weight=5.0)


# create a random variable at the present time
log_speciation[1] ~ dnUniform(-10.0,10.0)
log_extinction[1] ~ dnUniform(-10.0,10.0)


# apply moves on the rates
moves[++mvi] = mvSlide(log_speciation[1], weight=2)
moves[++mvi] = mvSlide(log_extinction[1], weight=2)


speciation[1] := exp( log_speciation[1] )
extinction[1] := exp( log_extinction[1] )


for (i in 1:NUM_INTERVALS) {
    index = i+1
    # specify normal priors (= Brownian motion) on the log of the rates
    log_speciation[index] ~ dnNormal( mean=log_speciation[i], sd=speciation_sd )
    log_extinction[index] ~ dnNormal( mean=log_extinction[i], sd=extinction_sd )

    # apply moves on the rates
    moves[++mvi] = mvSlide(log_speciation[index], weight=2)
    moves[++mvi] = mvSlide(log_extinction[index], weight=2)

    # transform the log-rate into actual rates
    speciation[index] := exp( log_speciation[index] )
    extinction[index] := exp( log_extinction[index] )

}

moves[++mvi] = mvVectorSlide(log_speciation, weight=10)
moves[++mvi] = mvVectorSlide(log_extinction, weight=10)

moves[++mvi] = mvShrinkExpand( log_speciation, sd=speciation_sd, weight=10 )
moves[++mvi] = mvShrinkExpand( log_extinction, sd=extinction_sd, weight=10 )


interval_times <- T.rootAge() * (1:NUM_INTERVALS) / (NUM_INTERVALS) * 0.8


### rho is the probability of sampling species at the present
### fix this to 23/377, since there are ~377 described species of primates
### and we have sampled 23
rho <- T.ntips()/377

timetree ~ dnEpisodicBirthDeath(rootAge=T.rootAge(), lambdaRates=speciation,
    lambdaTimes=interval_times, muRates=extinction, muTimes=interval_times, rho=rho,
    samplingStrategy="uniform", condition="survival", taxa=taxa)

### clamp the model with the "observed" tree
timetree.clamp(T)



#############
# The Model #
#############


### workspace model wrapper ###
mymodel = model(timetree)
```

```
### set up the monitors that will output parameter values to file and screen
monitors[++mni] = mnModel(filename="output/primates_uniform.log",printgen=10,
    separator = TAB)
monitors[++mni] = mnFile(filename="output/primates_uniform_speciation_rates.log",
    printgen=10, separator = TAB, speciation)
monitors[++mni] = mnFile(filename="output/primates_uniform_speciation_times.log",
    printgen=10, separator = TAB, interval_times)#
monitors[++mni] = mnFile(filename="output/primates_uniform_extinction_rates.log",
    printgen=10, separator = TAB, extinction)
monitors[++mni] = mnFile(filename="output/primates_uniform_extinction_times.log",
    printgen=10, separator = TAB, interval_times)
monitors[++mni] = mnScreen(printgen=1000, extinction_sd, speciation_sd)



################
# The Analysis #
################

### workspace mcmc ###
mymcmc = mcmc(mymodel, monitors, moves)

### pre-burnin to tune the proposals ###
mymcmc.burnin(generations=10000,tuningInterval=200)

### run the MCMC ###
mymcmc.run(generations=50000)
```

This `Rev` code shows the template for estimating episodic diversification rates. In the next sections we will tweak the script for the different sampling schemes.

## 6 Uniform Taxon Sampling

In our first analysis we will assume *uniform* taxon sampling (see Höhna et al. 2011; Höhna 2014). Uniform taxon sampling is the oldest and most basic technique to include incomplete taxon sampling (Nee et al. 1994; Yang and Rannala 1997). Uniform taxon sampling corresponds to the assumption that every species has the same probability $\rho$ to be included (*i.e.,* sampled) in our study. Imagine flipping a coin that has the probability $\rho$ to show up heads. For every species you flip the coin and are going to include the species, for example by sequencing it, in your study. This is what the assumption of uniform taxon sampling means.

For our study, we know that we have sampled 23 out of 377 living primate species. To account for this we can set the sampling parameter as a constant variable with a value of 23/377.

```
rho <- T.ntips()/377
```

Note that in principle you could specify a hyperprior distribution on the sampling probability **rho**. However, all three parameters (speciation rate, extinction rate, and sampling probability) are not identifiable (Stadler 2009). Nevertheless, we could specify informative priors on the sampling fraction if, for example, we know that the true diversity is in same range.

Moreover, we specify the *uniform* sampling scheme by setting **samplingStrategy="uniform"** in the birth-death process.

```
timetree ~ dnEpisodicBirthDeath(rootAge=T.rootAge(), lambdaRates=speciation,
    lambdaTimes=interval_times, muRates=extinction, muTimes=interval_times, rho=rho,
    samplingStrategy="uniform", condition="survival", taxa=taxa)
```

This is exactly what we did in the Rev script above.

→ The Rev file for performing this analysis: mcmc_uniform.Rev.

## 6.1 Exercise 1

- Run an MCMC simulation to estimate the posterior distribution of the speciation rate and extinction rate through time assuming *uniform* taxon sampling. You can use the script **mcmc_uniform.Rev** to run the analysis.
- Visualize the rate through time using R and RevGadgets.

## 6.2 Summarizing and plotting diversification rates through time

When the analysis is complete, you will have the monitored files in your output directory. You can then visualize the rates through time using R using our package RevGadgets. If you don't have the R-package RevGadgets installed, or if you have trouble with the package, then please read the separate tutorial about the package.

Just start R in the main directory for this analysis and then type the following commands:

```
library(RevGadgets)

tree <- read.nexus("data/primates.tre")
files <- c("output/primates_uniform_speciation_times.log", "output/
    primates_uniform_speciation_rates.log", "output/primates_uniform_extinction_times.
    log", "output/primates_uniform_extinction_rates.log")

rev_out <- rev.process.output(files,tree,burnin=0.25,numIntervals=100)

pdf("uniform.pdf")
par(mfrow=c(2,2))
rev.plot.output(rev_out)
dev.off()
```

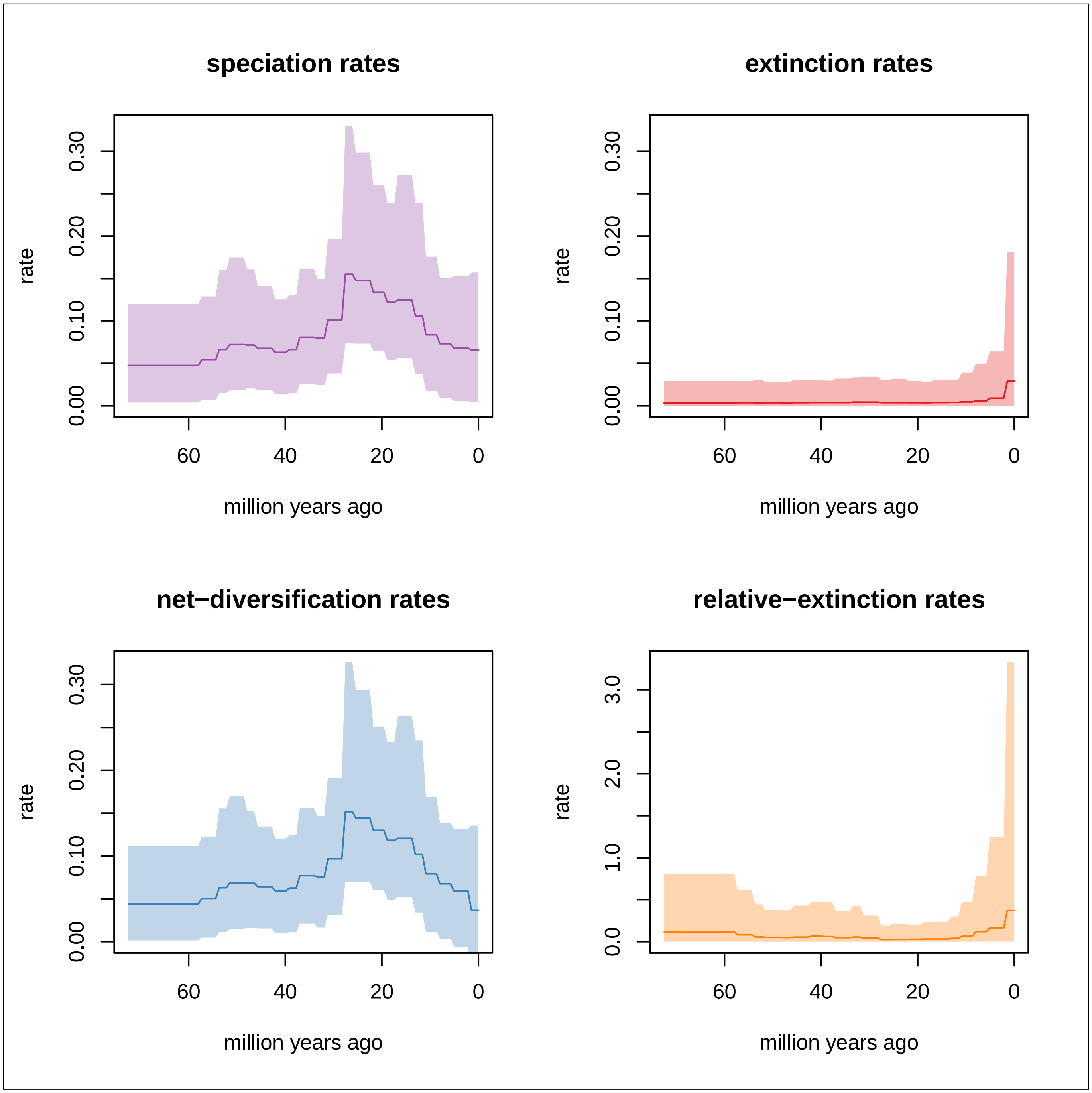


Figure 5: Resulting diversification rate estimations when using 20 intervals and assuming uniform taxon sampling. You should create similar plots for the other sampling schemes and compare the rates through time.

You can see the resulting plot in Figure 5.

## 7 Diversified Taxon Sampling

In the previous analysis we assumed that species were sampled uniformly. However, this assumption is very often violated (Cusimano and Renner 2010; Höhna et al. 2011). For example, the primate phylogeny that we use in this tutorial includes one species for almost all genera. Thus, we had selected the species for the study neither uniformly nor randomly but instead by including one species per genera and hence maximizing diversity. This sampling scheme is called *diversified* taxon sampling (Höhna et al. 2011).

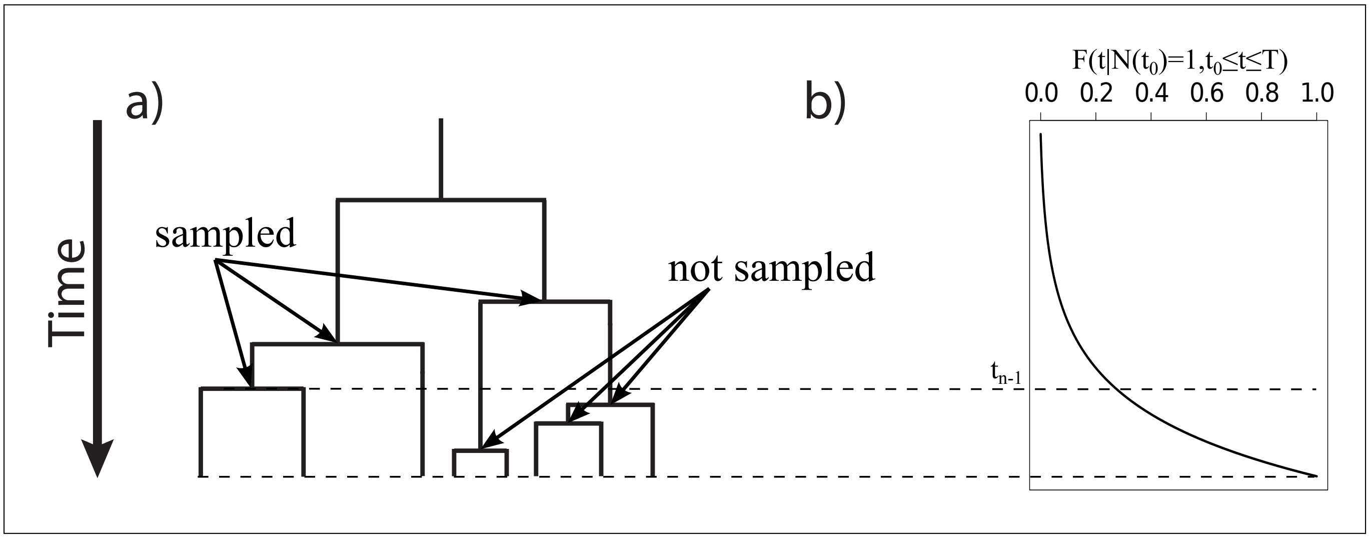


Figure 6: Example of diversified taxon sampling. a) An example phylogeny showing that all species after a certain time are not sampled. b) The cumulative probability of a speciation event occurring as a function of time. Here we see that the highest probability for a speciation event is more recently.

Figure 6 shows an example of diversified sampling. The example shows the same tree as in Figure 2 where 5 species are sampled. In fact, here we sampled 5 species so that every group is included and the most recent speciation events are excluded (not sampled).

In RevBayes we can specify *diversified* taxon sampling in the same way as we did *uniform* taxon sampling. First, we specify a constant variable for the sampling fraction **rho** which we set to the number of included (sampled) taxa divided by the number of total taxa in the group.

```
rho <- T.ntips()/377
```

Then, we specify that our sampling strategy was diversified by setting the argument of the birth-death process to**samplingStrategy="diversified"**.

```
timetree ~ dnEpisodicBirthDeath(rootAge=T.rootAge(), lambdaRates=speciation,
   lambdaTimes=interval_times, muRates=extinction, muTimes=interval_times, rho=rho,
   samplingStrategy="diversified", condition="time", taxa=taxa)
```

This is all we needed to do to change our previous script to model *diversified* taxon sampling.

## 7.1 Exercise 2

- Copy the `Rev` script **`mcmc_uniform.Rev`** and name it **`mcmc_diversified.Rev`**.
- Make the changes so that you assume now *diversified* taxon sampling.
- Change the file names in the monitors from **`uniform`** to **`diversified`**.
- Run the analysis and plot the diversification rates.
- How does the new sampling assumption influence your estimated rates?

## 8 Empirical Taxon Sampling

Unfortunately, *diversified* taxon sampling was derived under a strict mathematical concept that assumes all species that speciated before a given time were included and all other species were discarded (not sampled); see Figure 6. The *diversified* sampling strategy is clearly to restrictive to be realistic and can bias parameter estimates too (Höhna 2014). As another alternative we apply an *empirical* taxon sampling strategy that uses empirical information on the clade relationships and speciation times of the missing species. For example, in the primate phylogeny we know the crown age of Hominoidea and know that 19 additional speciation events must have happened between the crown age of the Hominoidea and the present time to accommodate the 19 missing species (see Figure 7). In fact, we can obtain for all larger groups the crown ages and the number of missing species and thus narrow down with empirical evidence the times when these missing speciation events have happened.

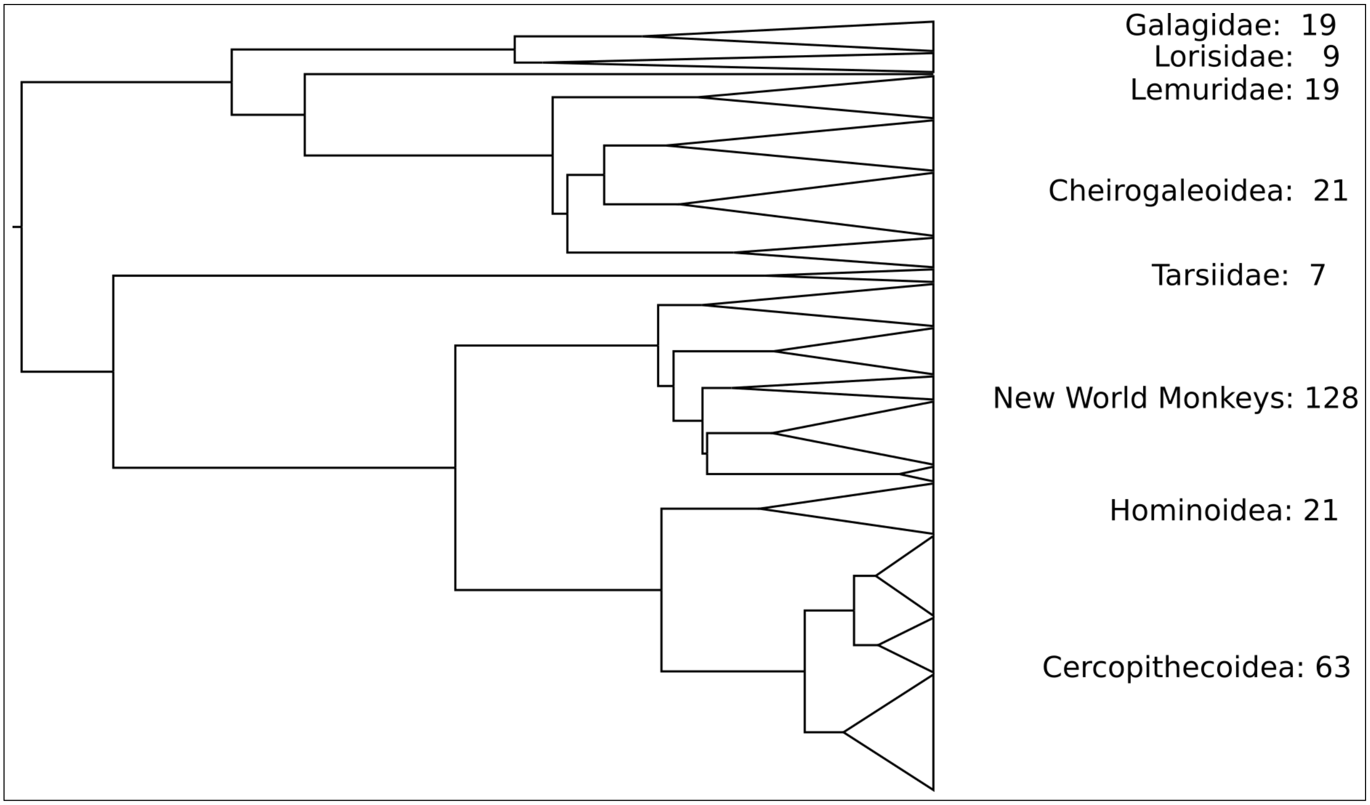


Figure 7: Cartoon of empirical taxon sampling. The triangle in the phylogeny depict clades with missing species. To illustrate the point we have written the names of higher taxa on the right with the number of species belonging to them. From this number of taxa in the clade we can compute how many species are missing per clade and which crown age the clade.

In your phylogeny you can count the number of species belonging to a given clade and thus compute how many species are missing for this clade. Then, you can pick two or more species to define the clade. These species will be used to compute the crown age. For example, we use *Pan_paniscus Hylobates_lar* to define the *Hominoidae* clade. If we would have *Homo_sapiens* sampled as well then we could additionally include it in the clade but we could not leave out *Hylobates_lar*.

In `Rev` we specify several clades and give the number of missing species.

```
Galagidae           = clade("Galago_senegalensis", "Otolemur_crassicaudatus", missing=
   17)
Lorisidae           = clade("Perodicticus_potto", "Loris_tardigradus", "
   Nycticebus_coucang", missing=6)
Cheirogaleoidea     = clade("Cheirogaleus_major", "Microcebus_murinus", missing= 19)
Lemuridae           = clade("Lemur_catta", "Varecia_variegata_variegata", missing=17)
Lemuriformes        = clade(Lemuridae, Cheirogaleoidea, missing=29)
Atelidae_Aotidae    = clade("Alouatta_palliata", "Aotus_trivirgatus", missing=30)
NWM                 = clade(Atelidae_Aotidae,"Callicebus_donacophilus", "
   Saimiri_sciureus", "Cebus_albifrons", missing=93)
Hominoidea          = clade("Pan_paniscus", "Hylobates_lar", missing=19)
Cercopithecoidea    = clade("Colobus_guereza", "Macaca_mulatta", "Chlorocebus_aethiops
   ", missing=60)
```

Next, we combine all clades into a single vector for later use.

```
missing_species_per_clade = v(Galagidae, Lorisidae, Cheirogaleoidea, Lemuridae,
   Lemuriformes, Atelidae_Aotidae, NWM, Hominoidea, Cercopithecoidea)
```

In the birth-death model we include these missing speciation events by integrating over the known interval when these speciation events have happened (between the crown age and the present). This integral of the probability density of a speciation event is exactly the same as one minus the cumulative distribution function of a speciation event,

$$F(t|N(t_1)=1, t_1 \leq t \leq T) = 1 - \frac{1 - P(N(T)>0|N(t)=1)\exp\left(r(t,T)\right)}{1 - P(N(T)>0|N(t_1)=1)\exp\left(r(t_1,T)\right)} \tag{5}$$

which was previously derived by Höhna (2014; Equation (6)) (see also Yang and Rannala (1997; Equation (3)) for constant rates and Höhna (2013; Equation (8))).

Let us define $\mathbb{K}$ as the set of missing species and $|\mathbb{K}|$ the number of clades with missing species. In our example we have $|\mathbb{K}| = 9$ clades. Additionally, we define $c_i$ as the time of most recent common ancestor of the $i^{th}$ clade.

Then, the joint probability density of the sampled reconstructed tree and the empirically informed missing speciation times is

$$\begin{aligned} f(\Psi, \mathbb{K}|N(t_1\!=\!0)\!=\!2, S(2,t_1\!=\!0,T)) &= f(\Psi|N(t_1\!=\!0)\!=\!2, S(2,t_1\!=\!0,T)) \\ &\quad \times \prod_{i=1}^{|\mathbb{K}|} \left(1 - F(t|N(c_i)=1, c_i \leq t \leq T)\right)^{k_i} \end{aligned} \tag{6}$$

Equation (6) is actually proportional to the original equation under the birth-death process times the probabilities of the missing species. There are two things to consider when specifying empirical taxon sampling in `RevBayes`. First, we set the "traditional" sampling fraction to one (**`rho=1.0`**). Second, we provide the clades with missing species as an argument of the birth-death model (**`incompleteClades=missing_species_per_clade`**).

```
timetree ~ dnEpisodicBirthDeath(rootAge=T.rootAge(), lambdaRates=speciation,
   lambdaTimes=interval_times, muRates=extinction, muTimes=interval_times, rho=1.0,
   taxa=taxa, incompleteClades=missing_species_per_clade, condition="time")
```

These are the only necessary changes to perform a diversification rate analysis under *empirical* taxon sampling.

## 8.1 Exercise 3

- Copy the `Rev` script **`mcmc_uniform.Rev`** and name it **`mcmc_empirical.Rev`**.
- Make the changes so that you assume now *empirical* taxon sampling.
- Change the file names in the monitors from **`uniform`** to **`empirical`**.
- Run the analysis and plot the diversification rates.
- How does the new sampling assumption influence your estimated rates?

# References

Cusimano, N. and S. Renner. 2010. Slowdowns in diversification rates from real phylogenies may not be real. Systematic Biology 59:458.

Cusimano, N., T. Stadler, and S. S. Renner. 2012. A new method for handling missing species in diversification analysis applicable to randomly or nonrandomly sampled phylogenies. Systematic Biology 61:785–792.

Drummond, A., M. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with beauti and the beast 1.7. Molecular Biology and Evolution 29:1969–1973.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Höhna, S. 2013. Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes. Bioinformatics 29:1367–1374.

Höhna, S. 2014. Likelihood Inference of Non-Constant Diversification Rates with Incomplete Taxon Sampling. PLoS One 9:e84184.

Höhna, S. 2015. The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. Journal of Theoretical Biology 380:321–331.

Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck. 2014. Probabilistic Graphical Model Representation in Phylogenetics. Systematic Biology 63:753–771.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65:726–736.

Höhna, S., T. Stadler, F. Ronquist, and T. Britton. 2011. Inferring speciation and extinction rates under different species sampling schemes. Molecular Biology and Evolution 28:2577–2589.

Kendall, D. G. 1948. On the generalized "birth-and-death" process. The Annals of Mathematical Statistics 19:1–15.

May, M. R., S. Höhna, and B. R. Moore. 2016. A Bayesian Approach for Detecting the Impact of Mass-Extinction Events on Molecular Phylogenies When Rates of Lineage Diversification May Vary. Methods in Ecology and Evolution 7:947–959.

Nee, S., R. M. May, and P. H. Harvey. 1994. The Reconstructed Evolutionary Process. Philosophical Transactions: Biological Sciences 344:305–311.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Springer, M. S., R. W. Meredith, J. Gatesy, C. A. Emerling, J. Park, D. L. Rabosky, T. Stadler, C. Steiner, O. A. Ryder, J. E. Janečka, et al. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS One 7:e49521.

Stadler, T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. Journal of Theoretical Biology 261:58–66.

Thompson, E. 1975. Human evolutionary trees. Cambridge University Press Cambridge.

Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution 14:717–724.

Yule, G. 1925. A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213:21–87.

Version dated: September 5, 2016