# Phylogenetic Inference using `RevBayes`

*Environmental Correlated Diversification Rate Estimation*

Sebastian Höhna and Luis Palazzesi

## 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 $t_{\text{MRCA}}$ 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).

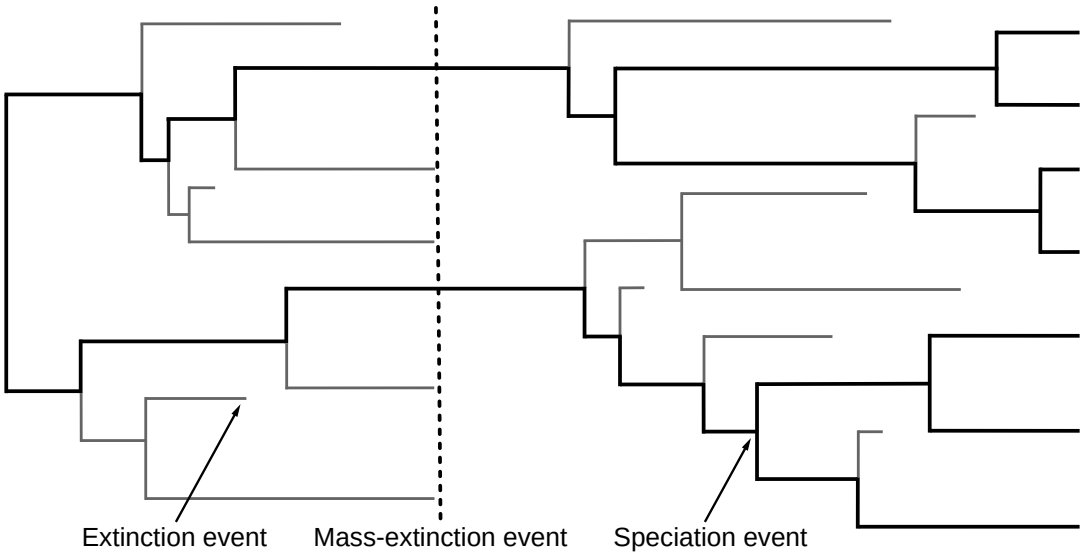


Figure 1: A realization of the birth-death process with mass extinction. Lineages that have no extant or sampled descendant are shown in gray and surviving lineages are shown in a thicker black line.

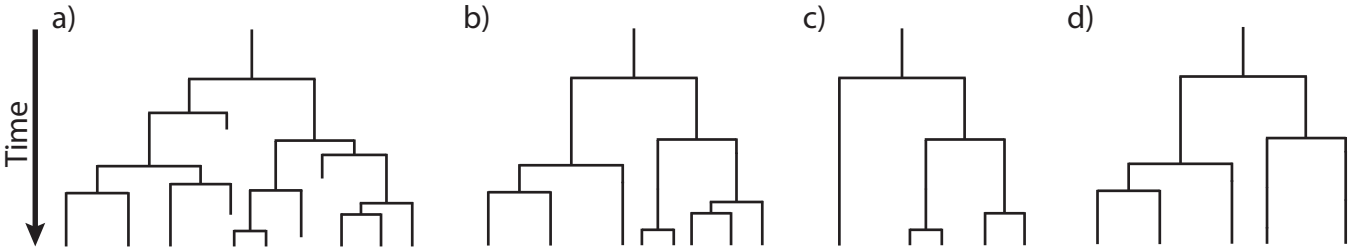


Figure 2: **Examples of trees produced under a birth-death process.** The process is initiated at the first speciation event (the 'crown-age' of the MRCA) when there are two initial lineages. At each speciation event the ancestral lineage is replaced by two descendant lineages. At an extinction event 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

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

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

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

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)

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

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

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 Environmental-dependent Speciation & Extinction Rates

## 3.1 Outline

This tutorial describes how to specify a branching-process model with diversification rate correlated with an environmental variable in `RevBayes`. Diversification rates are assumed to be equal among all lineages but vary through time correlated with an environmental predictor variable. Thus, this model can be used to test for correlations between diversification rates and environmental variables, such as $CO_2$ and temperature. However, these tests are only to establish a correlation, not a causality.

As usual, we provide the probabilistic graphical model at the beginning of this tutorial. Hopefully this will help you to get a better idea of all the variables in the model and their dependencies. Our goal in this tutorial is to estimate the correlation coefficient between speciation and extinction rates to historical $CO_2$ measurements using Markov chain Monte Carlo (MCMC).

## 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 it is particularly 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 episodic diversification rate tutorial by modifying the prior model on diversification rates through time to depend on some environmental variable.

## 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 files

- **primates_springer.tre**: Dated primates phylogeny including 369 out of 377 species from Springer et al. (2012).

→ Open the tree **data/primates_springer.tre** in FigTree.

## 5 Environmental-dependent Diversification Rates

The fundamental idea of this model is the question if diversification rates are correlated with an environmental variable. Examples of environmental variables are $CO_2$ and temperature. Have a look at Figure 3

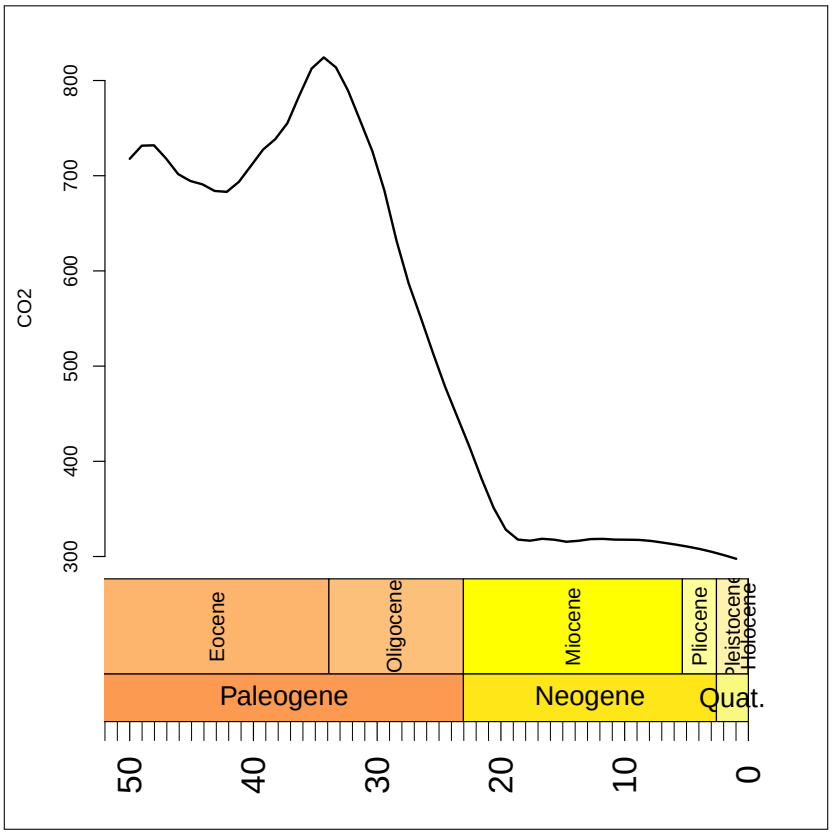


Figure 3: Estimates of historical $CO_2$ values. These estimates are obtained from XXX. The unit of $CO_2$ represents XXX.

which shows the historical value $CO_2$ in the last 50 million years. We can clearly see that the $CO_2$ dropped drastically around 30 million years ago.

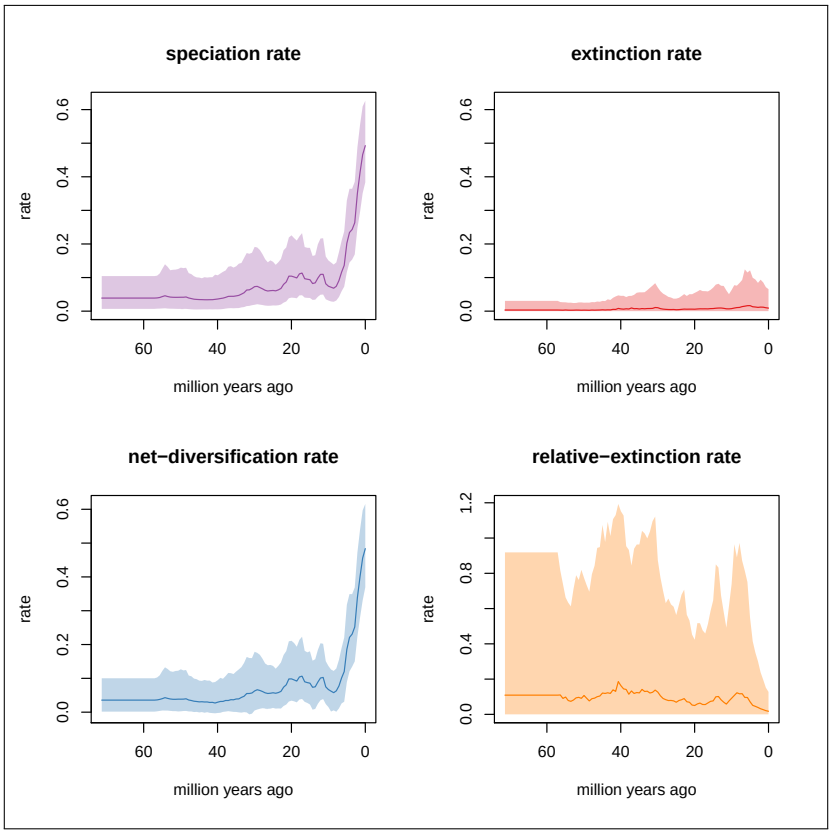


Figure 4: Estimated diversification rates through time. These estimates are taken from the episodic birth-death model with autocorrelated (Brownian motion) rate as described in the Diversification Rates Through Time tutorial.

In our previous Diversification Rates Through Time tutorial we estimated diversification as shown in Figure 4. We clearly see that diversification rates were not constant through time. Now we wonder if perhaps the diversification rates are correlated with $CO_2$.

We want to build on our episodic birth-death model so that our environmental correlation model collapses to the episodic birth-death model if there is no correlation. Recall that we used a Brownian motion model on the log-transformed rates. Hence, we assumed that the rates in the next time interval (epoch) have the current value as their expectation:

$$E[\log(\lambda(t))] = \log(\lambda(t - \Delta t)) \tag{5}$$

For the environmental dependent birth-death model, we have additional observation from the environmental variable. Thus, we know how much the environmental variable changed between time intervals (epochs). We can compute this change by taking the ratio between two consecutive measurements: $\frac{\mathrm{CO}_2(t)}{\mathrm{CO}_2(t-\Delta t)}$. Hence, if the $CO_2$ double from one epoch to the next we would compute a change of 2. This has the clear advantage that our computation is less sensitive to the unit and magnitude of the environmental variable.

Now let us assume that our diversification rates shift synchronously with the environmental variable if they are actually correlated. Then we can express our expectation of the log-transformed diversification rate in the next time interval (epoch) as being equal the log-transform diversification rate in the current time interval plus the log-transformed change in the environmental variable:

$$E[\log(\lambda(t))] = \log(\lambda(t - \Delta t)) + \beta \times \log\left(\frac{\mathrm{CO}_2(t)}{\mathrm{CO}_2(t - \Delta t)}\right) \ . \tag{6}$$

Here we denote the correlation coefficient by $\beta$. If $\beta > 0$ then there is a positive correlation between the speciation rate and $CO_2$, that is, if the $CO_2$ increases then the speciation increases also. If $\beta < 0$ then there is a negative correlation between the speciation rate and $CO_2$, that is, if the $CO_2$ increases then the speciation decreases. Finally, if $\beta = 0$ then there is no correlation and our model collapses to the episodic birth-death model.

In summary, we use a regression-like prior model for the speciation and extinction rate where the environmental variable (here $CO_2$) is the predictor variable. Specifically, we use a Brownian motion model for the log-transformed speciation and extinction rates where the expectation depends on the shift in the environmental variable. Thus, our model can be considered as a Brownian motion model with drift where the drift parameter is the environmental variable.

We will now walk you through setting up this analysis in `RevBayes`.

## 5.1 Read the tree

Begin by reading in the "observed" tree.

```
T <- readTrees("data/primates_springer.tre")[1]
```

From this tree, we get some helpful variables, such as the taxon information which we need to instantiate the birth-death process.

```
taxa <- T.taxa()
```

Additionally, we initialize an iterator variable for our vector of moves and monitors.

```
mvi = 0
mni = 0
```

## 5.2 Set up the environmental data

We take the $CO_2$ measurement from XXX and store the values on a vector; one measurement (value) per interval.

```
var <- v(297.6, 301.36, 304.84, 307.86, 310.36, 312.53, 314.48, 316.31, 317.42,
   317.63, 317.74, 318.51, 318.29, 316.5, 315.49, 317.64, 318.61, 316.6, 317.77,
   328.27, 351.12, 381.87, 415.47, 446.86, 478.31, 513.77, 550.74, 586.68, 631.48,
   684.13, 725.83, 757.81, 789.39, 813.79, 824.25, 812.6, 784.79, 755.25, 738.41,
   727.53, 710.48, 693.55, 683.04, 683.99, 690.93, 694.44, 701.62, 718.05, 731.95,
   731.56, 717.76)
```

Then we specify the maximum age of the measurements. This corresponds to the time of the last interval.

```
MAX_VAR_AGE = 50
```

We will later use this maximum age to compute the times for each interval by assuming that each interval is equal in time.

Finally, we create a helper variable that specifies the number of intervals.

```
NUM_INTERVALS = var.size()-1
```

This variable will help us to create the episodic diversification rate using a **for**-loop.

### 5.2.1 Setting up the time intervals

In RevBayes you actually have the possibility to specify unequal time intervals or even different intervals for the speciation and extinction rate. This is achieved by providing a vector of times when each interval ends. However, here we assume for simplicity that each interval has the same length because this is how we obtained our environmental data.

```
interval_times <- MAX_VAR_AGE * (1:NUM_INTERVALS) / NUM_INTERVALS
```

This vector of times will be used for both the speciation and extinction rates. Also, remember that the times of the intervals represent ages going backwards in time.

## 5.3 Specifying the model

### 5.3.1 Priors on amount of rate variation

We follow here exactly the prior specification as in the Diversification Rates Through Time tutorial because we want our model to collapse to the episodic birth-death if there is no correlation.

We start by specifying prior distributions on the rates. Each interval-specific speciation and extinction rate will be drawn from a normal distribution. Thus, we need a parameter for the standard deviation of those normal distributions. We use an exponential hyperprior with rate 1.0 to estimate the standard deviation, but assume that all speciation rates and all extinction rates share the same standard deviation. The motivation for an exponential hyperprior is that it has the highest probability density at 0 which would make the variance of rates between consecutive time intervals 0 and thus represent a constant rate process. The data will tell us if there should be much variation in rates through time. (You may want to experiment with this hyperprior if you are interested.)

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

We apply a simple scaling move on each prior parameter.

```
moves[++mvi] = mvScale(speciation_sd,weight=5.0)
moves[++mvi] = mvScale(extinction_sd,weight=5.0)
```

### 5.3.2 Specifying the correlation coefficients

Then we specify normal prior distributions on the correlation coefficient $\beta$ for the speciation and extinction rate. Again, out total lack of prior knowledge, we will assume that the standard deviation of $\beta$ is 1.0 and you may want to modify this value. Nevertheless, this normal prior distribution is motivated by being centered at 0.0 (no correlation) and gives equal weight to positive and negative correlations.

```
beta_speciation ~ dnNormal(0,1.0)
beta_extinction ~ dnNormal(0,1.0)
```

We apply simple sliding-window moves for the two correlation coefficients because they are defined on the whole real line.

```
moves[++mvi] = mvSlide(beta_speciation,delta=1.0,weight=10.0)
moves[++mvi] = mvSlide(beta_extinction,delta=1.0,weight=10.0)
```

Additionally, we might be interested in the posterior probability that there is a positive correlation, $\mathbb{P}(\beta > 0)$, or a negative correlation, $\mathbb{P}(\beta < 0)$, respectively. We achieve this using a deterministic variable that is 1 if $\beta < 0$

```
speciation_corr_neg_prob := ifelse(beta_speciation < 0.0, 1, 0)
extinction_corr_neg_prob := ifelse(beta_extinction < 0.0, 1, 0)
speciation_corr_pos_prob := ifelse(beta_speciation > 0.0, 1, 0)
extinction_corr_pos_prob := ifelse(beta_extinction > 0.0, 1, 0)
```

Note that in this model the probability of $\beta$ being 0.0 ($\mathbb{P}(\beta = 0) = 0$) because we are working with a prior and posterior *density* on $\beta$ and thus any specific value, *e.g.,* 0.0, has a probability of 0.0. We will circumvent this issue in the next chapter when we use reversible-jump MCMC to set $\beta$ specifically to 0.0. Here you can also check that the posterior probability of **`speciation_corr_pos_prob`** equals 1-**`speciation_corr_neg_prob`**.

### 5.3.3 Specifying correlated rates

As we mentioned before, we will apply normal distributions as priors for each log-transformed rate. We begin with the rate at the present which is our initial rate parameter. The rates at the present will be specified slightly differently because they are not correlated to any previous rates. This is because we are actually modeling rate-changes backwards in time and there is no previous rate for the rate at the present.

We use a uniform distribution between -10 and 10 because of our lack of prior knowledge on the diversification rate. This actually means that we allow speciation and extinction rates between $e^{-10}$ and $e^{1}0$ we should clearly cover the true values. (Note that for diversification rate estimates $e^{-10}$ is virtually 0 since the rate is so slow).

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

Notice that we store the diversification rate variables in vectors. Storing the rate parameters in vectors will be useful and important later when we pass the rates into the birth-death process.

We apply simple sliding window moves for the rates. Normally we would use scaling moves but in this case we work on the log-transformed parameters and thus sliding moves perform better. (If you are keen you can test the differences.)

```
moves[++mvi] = mvSlide(log_speciation[1], weight=2)
moves[++mvi] = mvSlide(log_extinction[1], weight=2)
```

Now we transform the diversification rate parameters into actual rates.

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

Next, we specify the speciation and extinction rates for each time interval (*i.e.,* epoch). This can be done efficiently using a **for**-loop. We will use a specific index variable so that we can easier refer to the rate at the previous interval. Remember that we want to model the rates as a Brownian motion, which we achieve by specify a normal distribution as the prior distribution on the rates centered around the previous rate plus the change in the environmental variable (*i.e.,* the mean is equal to the previous rate plus the log-transformed ratio of the environmental variable divided by the previous value).

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

    expected_speciation[index] := log_speciation[i] + beta_speciation * ln( var[index]
       / var[i] )
    expected_extinction[index] := log_extinction[i] + beta_extinction * ln( var[index]
       / var[i] )

    log_speciation[index] ~ dnNormal( mean=expected_speciation[index], sd=speciation_sd
        )
    log_extinction[index] ~ dnNormal( mean=expected_extinction[index], sd=extinction_sd
        )
```

```
    moves[++mvi] = mvSlide(log_speciation[index], weight=2)
    moves[++mvi] = mvSlide(log_extinction[index], weight=2)

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

}
```

Finally, we apply moves that slide all values in the rate vectors, *i.e.,* all speciation or extinction rates. We will use an **mvVectorSlide** move.

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

Additionally, we apply a **mvShrinkExpand** move which changes the spread of several variables around their mean.

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

Both moves considerably improve the efficiency of our MCMC analysis.

### 5.3.4 Incomplete Taxon Sampling

We know that we have sampled 367 out of 377 living primate species. To account for this we can set the sampling parameter as a constant node with a value of 367/377. For simplicity, and since almost all species have been sampled, we assume *uniform* taxon sampling (Höhna et al. 2011; Höhna 2014),

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

### 5.3.5 Root age

The birth-death process requires a parameter for the root age. In this exercise we use a fix tree and thus we know the age of the tree. Hence, we can get the value for the root from the Springer et al. (2012) tree.

```
root_time <- T.rootAge()
```

#### 5.3.6 The time tree

Now we have all of the parameters we need to specify the full episodic birth-death model. We initialize the stochastic node representing the time tree.

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

You may notice that we explicitly specify that we want to condition on survival. It is possible to change this condition to the *time of the process* or *the number of sampled taxa* too.

Then we attach data to the **timetree** variable.

```
timetree.clamp(T)
```

Finally, we create a workspace object of our whole model using the **model()** function.

```
mymodel = model(speciation)
```

The **model()** function traversed all of the connections and found all of the nodes we specified.

### 5.4 Running an MCMC analysis

#### 5.4.1 Specifying Monitors

For our MCMC analysis, we need to set up a vector of *monitors* to record the states of our Markov chain. First, we will initialize the model monitor using the **mnModel** function. This creates a new monitor variable that will output the states for all model parameters when passed into a MCMC function.

```
monitors[++mni] = mnModel(filename="output/primates_EBD_Corr.log",printgen=10,
    separator = TAB)
```

Additionally, we create four separate file monitors, one for each vector of speciation and extinction rates and for each speciation and extinction rate epoch (*i.e.,* the times when the interval ends). We want to have the speciation and extinction rates stored separately so that we can plot them nicely afterwards.

```
monitors[++mni] = mnFile(filename="output/primates_EBD_Corr_speciation_rates.log",
    printgen=10, separator = TAB, speciation)
monitors[++mni] = mnFile(filename="output/primates_EBD_Corr_speciation_times.log",
    printgen=10, separator = TAB, interval_times)
```

```
monitors[++mni] = mnFile(filename="output/primates_EBD_Corr_extinction_rates.log",
   printgen=10, separator = TAB, extinction)
monitors[++mni] = mnFile(filename="output/primates_EBD_Corr_extinction_times.log",
   printgen=10, separator = TAB, interval_times)
```

Finally, create a screen monitor that will report the states of specified variables to the screen with **mnScreen**:

```
monitors[++mni] = mnScreen(printgen=1000, beta_speciation, beta_extinction)
```

### 5.4.2 Initializing and Running the MCMC Simulation

With a fully specified model, a set of monitors, and a set of moves, we can now set up the MCMC algorithm that will sample parameter values in proportion to their posterior probability. The **mcmc()** function will create our MCMC object:

```
mymcmc = mcmc(mymodel, monitors, moves)
```

First, we will run a pre-burnin to tune the moves and to obtain starting values from the posterior distribution.

```
mymcmc.burnin(generations=10000,tuningInterval=200)
```

Now, run the MCMC:

```
mymcmc.run(generations=50000)
```

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.tree("data/primates_Springer.tre")

# the CO2 values as a reference in our plot
```

```
co2 <- c(297.6, 301.36, 304.84, 307.86, 310.36, 312.53, 314.48, 316.31, 317.42,
    317.63, 317.74, 318.51, 318.29, 316.5, 315.49, 317.64, 318.61, 316.6, 317.77,
    328.27, 351.12, 381.87, 415.47, 446.86, 478.31, 513.77, 550.74, 586.68, 631.48,
    684.13, 725.83, 757.81, 789.39, 813.79, 824.25, 812.6, 784.79, 755.25, 738.41,
    727.53, 710.48, 693.55, 683.04, 683.99, 690.93, 694.44, 701.62, 718.05, 731.95,
    731.56, 717.76)

MAX_VAR_AGE = 50
NUM_INTERVALS = length(co2)
co2_age <- MAX_VAR_AGE * (1:NUM_INTERVALS) / NUM_INTERVALS
predictor.ages <- co2_age
predictor.var <- co2

rev_out <- rev.process.div.rates(speciation_times_file = "output/
    primates_EBD_Corr_speciation_times.log",
                                 speciation_rates_file = "output/
                                    primates_EBD_Corr_speciation_rates.log",
                                 extinction_times_file = "output/
                                    primates_EBD_Corr_extinction_times.log",
                                 extinction_rates_file = "output/
                                    primates_EBD_Corr_extinction_rates.log",
                                 tree,
                                 burnin=0.25,numIntervals=100)
pdf("EBD_Corr.pdf")
par(mfrow=c(2,2))
rev.plot.div.rates(rev_out, predictor.ages=co2_age, predictor.var=co2, use.geoscale=
    TRUE)
dev.off()
```

→ The `Rev` file for performing this analysis: `mcmc_EBD.Rev`.

## 5.5 A brief discussion on estimated diversification rates

Figure 5 shows the estimated diversification rates through time and the $CO_2$. If you compare these estimates with Figure 4 then you may notice that the diversification rate estimate are virtually identical. This is a good sign for the analysis because it shows that the information in the estimates comes from the data (the tree in this case) and not from the assumed model. Thus, we are not artificially forcing the diversification rates to follow our environmental variable but instead estimate if there is a correlation. Small deviation between the estimated rates under the different analyses are expected because there will be some interaction between the environmental variable and the diversification rate estimates. Additionally, the uncertainty in estimated diversification rates through time is large und minor changes are within this uncertainty.

## 5.6 Exercise 1

- Run an MCMC simulation to estimate the posterior distribution of the speciation rate and extinction rate.

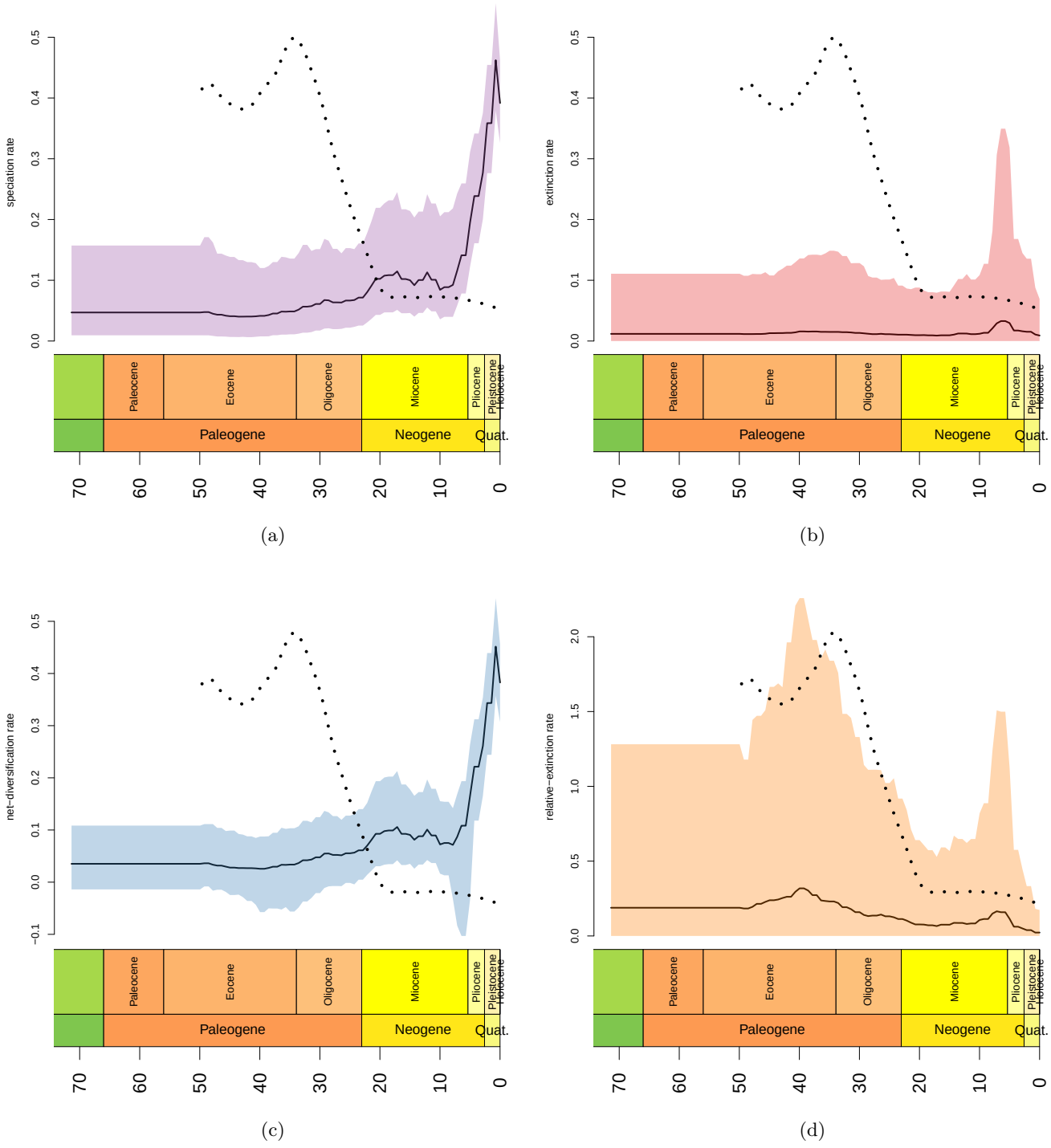


Figure 5: Resulting diversification rate estimations

- Visualize the rate through time using `R`.
- Open the file **`output/primates_EBD_Corr.log`** in `Tracer`. What is the estimated probability that $\beta < 0$? You'll find the estimate in the variable **`speciation_corr_neg_prob`**.
- We specified a normal prior with mean 0 on $\beta$ and thus used a prior probability of 0.5 that $\beta < 0$. Now you can use the posterior ratio divided by the prior ratio to compute the Bayes factor. What is the Bayes factor support for or against a positive correlation between the speciation rate and $CO_2$?
- Similarly, is there support for a positive or negative correlation between the extinction rate and $CO_2$?

## 6 Testing for correlation using reversible-jump MCMC

In the previous exercise we wanted that our model collapses to the episodic birth-death process if there is no environmental correlation. We achieved this by setting up our prior model so that if $\beta = 0$ the model collapses. However, we also used a normal prior distribution with mean 0.0 and standard deviation 1.0 for $\beta$. Thus, we implicitly specified that $\beta$ being exactly 0.0 has probability 0.0 because every specific value of a continuous distribution has a 0.0 probability despite having a positive probability density. For example, you might notice that you will never sample in your MCMC run the value 0.0 exactly although we might sample values that are close to 0.0.

Now we want to use reversible jump MCMC to test specifically if the hypothesis $\beta = 0$ is rejected. Remember that reversible jump MCMC can estimate the posterior probability for different models. The first model will be that $\beta = 0$ and the second model will be that $\beta \sim \text{norm}(0, 1)$. Then we can simply compute Bayes factors by computing the posterior ratio divided by the prior ratio to assess the support for either model.

In `RevBayes` we have a very flexible way to specify a reversible-jump MCMC. We can provide any constant value and distribution to the distribution **`dnReversibleJumpMixture`**. This will mean that the value, **`beta_speciation`** and **`beta_extinction`**, will either take on the constant value or drawn from the base-distribution.

```
beta_speciation ~ dnReversibleJumpMixture(constantValue=0.0, baseDistribution=dnNormal
    (0,1.0), p=0.5)
beta_extinction ~ dnReversibleJumpMixture(constantValue=0.0, baseDistribution=dnNormal
    (0,1.0), p=0.5)
```

Additionally we also need a specific move that switches if the value is equal to the constant value or drawn from the base-distribution. This is where we use the reversible-jump move **`mvRJSwitch`**.

```
moves[++mvi] = mvRJSwitch(beta_extinction, weight=5)
```

Now we can also monitor for convenience what the probability of **`beta_speciation`** and **`beta_extinction`** being 0.0 is. We will set this up by a deterministic variable that will be 1.0 if $\beta \neq 0$ and will be 0.0 if $\beta = 0.0$.

Thus the two variables **speciation_corr_prob** and **extinction_corr_prob** represent the probability that there is a correlation between the speciation rate or the extinction rate and $CO_2$.

```
speciation_corr_prob := ifelse(beta_speciation == 0.0, 0, 1)
extinction_corr_prob := ifelse(beta_extinction == 0.0, 0, 1)
```

These are the only necessary changes to the above analysis to run a reversible-jump MCMC.

### 6.1 Exercise 2

- Make a copy of the script **mcmc_EBD_Corr.Rev** and call it **mcmc_EBD_Corr_RJ.Rev**.
- Replace the prior distribution on **beta_speciation** and **beta_extinction** to use the **dnReversibleJumpMixture** instead.
- Also add the new moves **mvRJSwitch** and deterministic variables **speciation_corr_prob** and **extinction_corr_**
- Don't forget to change the output filenames in the monitors, *e.g.,* add **_RJ** to the name.
- Run the reversible-jump MCMC analysis.
- Open the file **output/primates_EBD_Corr_RJ.log** in **Tracer**. What is the estimated probability that $\beta = 0$? You'll find the estimate in the variable **speciation_corr_prob**.
- We specified a prior probability of 0.5 on $\beta$ being fixed to 0.0. Now you can use the posterior ratio divided by the prior ratio to compute the Bayes factor. What is the Bayes factor support for or against any correlation between the speciation rate and $CO_2$?
- Similarly, is there support for a positive or negative correlation between the extinction rate and $CO_2$?

## References

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., 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.

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.

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

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: October 18, 2016