Divergence Time Estimation using BEAST v2.∗
Dating Species Divergences with the Fossilized Birth-Death Process

Tracy A. Heath

Central among the questions explored in biology are those that seek to understand the timing and rates of evolutionary processes. Accurate estimates of species divergence times are vital to understanding historical biogeography, estimating diversification rates, and identifying the causes of variation in rates of molecular evolution.

This tutorial will provide a general overview of divergence time estimation and fossil calibration using a stochastic branching process and relaxed-clock model in a Bayesian framework. The exercise will guide you through the steps necessary for estimating phylogenetic relationships and dating species divergences using the program BEAST v2.∗.

1 Background

Estimating branch lengths in proportion to time is confounded by the fact that the rate of evolution and time are intrinsically linked when inferring genetic differences between species. A model of lineage-specific substitution rate variation must be applied to tease apart rate and time. When applied in methods for divergence time estimation, the resulting trees have branch lengths that are proportional to time. External node age estimates from the fossil record or other sources are necessary for inferring the real-time (or absolute) ages of lineage divergences (Figure 1).

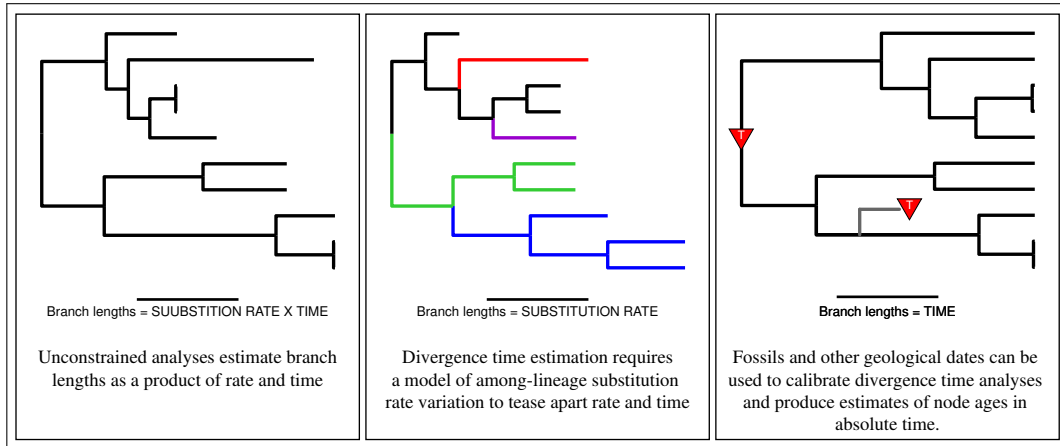| Branch lengths = SUUBSTITION RATE X TIME	Branch lengths = SUBSTITUTION RATE	Branch lengths = TIME
Unconstrained analyses estimate branch lengths as a product of rate and time	Divergence time estimation requires a model of among-lineage substitution rate variation to tease apart rate and time	Fossils and other geological dates can be used to calibrate divergence time analyses and produce estimates of node ages in absolute time.

Figure 1: Estimating branch lengths in units of time requires a model of lineage-specific rate variation, a model for describing the distribution of speciation events over time, and external information to calibrate the tree.

Ultimately, the goal of Bayesian divergence time estimation is to estimate the joint posterior probability, $\mathbb{P}(\mathcal{R}, \mathcal{T} | \mathcal{S}, \mathcal{C})$, of the branch rates ($\mathcal{R}$) and times ($\mathcal{T}$) given a set of sequences ($\mathcal{S}$) and calibration information ($\mathcal{C}$):

$$\mathbb{P}(\mathcal{R}, \mathcal{T} \mid \mathcal{S}, \mathcal{C}) = \frac{\mathbb{P}(\mathcal{S} \mid \mathcal{R}, \mathcal{T}) \; \mathbb{P}(\mathcal{R}) \; \mathbb{P}(\mathcal{T} \mid \mathcal{C})}{\mathbb{P}(\mathcal{S} \mid \mathcal{C})},$$

where $\mathbb{P}(\mathcal{S}|\mathcal{R}, \mathcal{T})$ is the likelihood, $\mathbb{P}(\mathcal{R})$ is the prior probability of the rates, $\mathbb{P}(\mathcal{T}|\mathcal{C})$ is the prior probability of the times, and $\mathbb{P}(\mathcal{S}|\mathcal{C})$ is the marginal probability of the data. We use numerical methods—Markov

chain Monte Carlo (MCMC)—to eliminate the difficult task of calculating the marginal probability of the data. Thus, our primary focus, aside from the tree topology, is devising probability distributions for the prior on the rates, $\mathbb{P}(\mathcal{R})$, and the prior on the times, $\mathbb{P}(\mathcal{T}|\mathcal{C})$.

1.1   Modeling lineage-specific substitution rates

Many factors can influence the rate of substitution in a population such as mutation rate, population size, generation time, and selection. As a result, many models have been proposed that describe how substitution rate may vary across the Tree of Life.

The simplest model, the molecular clock, assumes that the rate of substitution remains constant over time (Zuckerkandl and Pauling 1962). However, many studies have shown that molecular data (in general) violate the assumption of a molecular clock and that there exists considerable variation in the rates of substitution among lineages.

Several models have been developed and implemented for inferring divergence times without assuming a strict molecular clock and are commonly applied to empirical data sets. Many of these models have been applied as priors using Bayesian inference methods. The implementation of dating methods in a Bayesian framework provides a flexible way to model rate variation and obtain reliable estimates of speciation times, provided the assumptions of the models are adequate. When coupled with numerical methods, such as MCMC, for approximating the posterior probability distribution of parameters, Bayesian methods are extremely powerful for estimating the parameters of a statistical model and are widely used in phylogenetics.

Some models of lineage-specific rate variation:

- Global molecular clock: a constant rate of substitution over time (Zuckerkandl and Pauling 1962)
- Local molecular clocks (Kishino et al. 1990; Rambaut and Bromham 1998; Yang and Yoder 2003; Drummond and Suchard 2010)
    – Closely related lineages share the same rate and rates are clustered by sub-clades
- Compound Poisson process (Huelsenbeck et al. 2000)
    – Rate changes occur along lineages according to a point process and at rate-change events, the new rate is a product of the old rate and a Γ-distributed multiplier.
- Autocorrelated rates: substitution rates evolve gradually over the tree
    – Log-normally distributed rates: the rate at a node is drawn from a log-normal distribution with a mean equal to the parent rate (Thorne et al. 1998; Kishino et al. 2001; Thorne and Kishino 2002)
    – Cox-Ingersoll-Ross Process: the rate of the daughter branch is determined by a non-central χ^2 distribution. This process includes a parameter that determines the intensity of the force that drives the process to its stationary distribution (Lepage et al. 2006).
- Uncorrelated rates
    – The rate associated with each branch is drawn from a single underlying parametric distribution such as an exponential or log-normal (Drummond et al. 2006; Rannala and Yang 2007; Lepage et al. 2007).
- Mixture model on branch rates
    – Branches are assigned to distinct rate categories according to a Dirichlet process (Heath et al. 2012).

The variety of models for relaxing the molecular clock assumption presents a challenge for investigators interested in estimating divergence times. Some models assume that rates are heritable and autocorrelated

over the tree, others model rate change as a step-wise process, and others assume that the rates on each branch are independently drawn from a single distribution. Furthermore, studies comparing the accuracy (using simulation) or precision of different models have produced conflicting results, some favoring uncorrelated models (Drummond et al. 2006) and others preferring autocorrelated models (Lepage et al. 2007). Because of this, it is important for researchers performing these analyses to consider and test different relaxed clock models (Lepage et al. 2007; Ronquist et al. 2012; Li and Drummond 2012; Baele et al. 2013). It is also critical to take into account the scale of the question when estimating divergence times. For example, it might not be reasonable to assume that rates are autocorrelated if the data set includes very distantly related taxa and low taxon sampling. In such cases, it is unlikely that any signal of autocorrelation is detectible.

1.2   Priors on node times

There are many component parts that make up a Bayesian analysis of divergence time. One that is often overlooked is the prior on node times, often called a *tree prior*. This model describes how speciation events are distributed over time. When this model is combined with a model for branch rate, Bayesian inference allows you to estimate *relative* divergence times. Furthermore, because the rate and time are confounded in the branch-length parameter, the prior describing the branching times can have a strong effect on divergence time estimation.

We can separate the priors on node ages into different categories:

- **Phenomenological**—models that make no explicit assumptions about the biological processes that generated the tree. These priors are conditional on the age of the root.
    - Uniform distribution: This simple model assumes that internal nodes are uniformly distributed between the root and tip nodes (Lepage et al. 2007; Ronquist et al. 2012).
    - Dirichlet distribution: A flat Dirichlet distribution describes the placement of internal nodes on every path between the root and tips (Kishino et al. 2001; Thorne and Kishino 2002).
- **Mechanistic**–models that describe the biological processes responsible for generating the pattern of lineage divergences.
    - Population-level processes—models describing demographic processes (suitable for describing differences among individuals in the same species/population)
        * Coalescent—These demographic models describe the time, in generations, between coalescent events and allow for the estimation of population-level parameters (Kingman 1982c; Kingman 1982a; Kingman 1982b; Griffiths and Tavare 1994).
    - Species-level processes—stochastic branching models that describe lineage diversification (suitable for describing the timing of divergences between samples from different species)
        * Yule (pure-birth) process: The simplest branching model assumes that, at any given point in time, every living lineage can speciate at the same rate, λ. Because the speciation rate is constant through time, there is an exponential waiting time between speciation events Yule (1924); Aldous (2001). The Yule model does not allow for extinction.
        * Birth-death process: An extension of the Yule process, the birth-death model assumes that at any point in time every lineage can undergo speciation at rate λ or go extinct at rate μ (Kendall 1948; Thompson 1975; Nee et al. 1994; Rannala and Yang 1996; Yang and Rannala 1997; Popovic 2004; Aldous and Popovic 2005; Gernhard 2008). Thus, the Yule process is a special case of the birth-death process where $\mu = 0$.

In BEAST, the available tree priors for divergence time estimation using inter-species sequences are variants of the birth-death prior. Extensions of the birth-death model include the calibrated Yule (Heled and Drummond 2012), the birth-death model with incomplete species sampling (Rannala and Yang 1996; Yang and Rannala 1997; Stadler 2009), and serially-sampled birth-death processes (Stadler 2010). Other programs also offer speciation priors as well as some alternative priors such as a uniform prior (*PhyloBayes*, *MrBayes v3.2*, *DPPDiv*), a Dirichlet prior (*multidivtime*), and a birth-death prior with species sampling (*MCMCTree*).

Tree priors based on the coalescent which are intended for population-level analyses virus data are also available in BEAST. The effect of different node-time priors on estimates of divergence times is not well understood and appears to be dataset-dependent (Lepage et al. 2007). Accordingly, it is important to account for the characteristics of your data when choosing a tree prior. If you know that your sequences are from extant species, each from different genera, then it is unlikely that a coalescent model adequately reflects the processes that generated those sequences. And since you do not have any samples from lineages in the past, then you should not use the serial-sampled birth-death model. Furthermore, if you have prior knowledge that extinction has occurred, then a pure-birth (Yule) prior is not appropriate.

1.3 Calibration to absolute time

Without external information to calibrate the tree, divergence time estimation methods can only reliably provide estimates of relative divergence times and not absolute node ages. In the absence of adequate calibration data, relative divergence times are suitable for analyses of rates of continuous trait evolution or understanding relative rates of diversification. However, for some problems, such as those that seek to uncover correlation between biogeographical events and lineage diversification, an absolute time scale is required. Calibration information can come from a variety of sources including "known" substitution rates (often secondary calibrations estimated from a previous study), dated tip sequences from serially sampled data (typically time-stamped virus data), or geological date estimates (fossils or biogeographical data).

Age estimates from fossil organisms are the most common form of divergence time calibration information. These data are used as age constraints on their putative ancestral nodes. There are numerous difficulties with incorporating node age estimates from fossil data including disparity in fossilization and sampling, uncertainty in dating, and correct phylogenetic placement of the fossil. Thus, it is critical that careful attention is paid to the paleontological data included in phylogenetic divergence time analyses. With an accurately dated and identified fossil in hand, further consideration is required to determine how to apply the node-age constraint. If the fossil is truly a descendant of the node it calibrates, then it provides a reliable minimum age bound on the ancestral node time. However, maximum bounds are far more difficult to come by. Bayesian methods provide a way to account for uncertainty in fossil calibrations. Prior distributions reflecting our knowledge (or lack thereof) of the amount of elapsed time from the ancestral node to its calibrating fossil are easily incorporated into these methods.

A nice review paper by Ho and Phillips (2009) outlines a number of different parametric distributions appropriate for use as priors on calibrated nodes.

Uniform distribution – Typically, you must have both maximum and minimum age bounds when applying a uniform calibration prior (though some methods are available for applying uniform constraints with soft bounds). The minimum bound is provided by the fossil member of the clade. The maximum bound may come from a bracketing method or other external source. This distribution places equal probability across all ages spanning the interval between the lower and upper bounds.
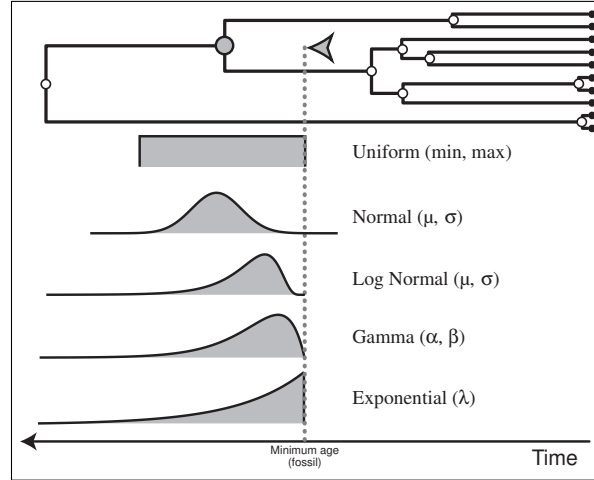
Figure 2: Five different parametric distributions that can be applied as priors on the age of a calibrated node. (figure adapted from Heath 2012)

Normal distribution – The normal distribution is not always appropriate for calibrating a node using fossil information (though some methods allow for assigning a truncated normal prior density). When applying a biogeographical date (e.g. the Isthmus of Panama) or a secondary calibration (a node age estimate from a previous study), the normal distribution can be a useful calibration prior. This distribution is always symmetrical and places the greatest prior weight on the mean (μ). Its scale is determined by the standard deviation parameter (σ).

Probability distributions restricted to the interval $[0, \infty)$, such as the log-normal, exponential, and gamma are appropriate for use as zero-offset calibration priors. When applying these priors on node age, the fossil age is the origin of the prior distribution. Thus, it is useful to consider the fact that the prior is modeling the amount of time that has elapsed since the divergence event (ancestral node) until the time of the descendant fossil (Figure 3).

Gamma distribution – The gamma distribution is commonly used as a prior on scalar variables in Bayesian inference. It relies on 2 parameters: the scale parameter (α) and a rate parameter (λ). More specifically, the gamma distribution is the sum of α independently and identical exponentially distributed random variables with rate λ. As α becomes very large ($\alpha > 10$), this distribution approaches the normal distribution.

Exponential distribution – The exponential distribution is a special case of the gamma distribution and is characterized by a single rate parameter (λ) and is useful for calibration if the fossil age is very close to the age of its ancestral node. The expected (mean) age difference under this distribution is equal to λ^{-1} and the median is equal to $\lambda^{-1} * ln(2)$. Under the exponential distribution, the greatest prior weight is placed on node ages very close to the age of the fossil with diminishing probability to ∞. As λ is increased, this prior density becomes strongly informative, whereas very low values of λ result in a fairly non-informative prior (Figure 3a).

Log-normal distribution – An offset, log-normal prior on the calibrated node age places the highest probability on ages somewhat older than the fossil, with non-zero probability to ∞. If a variable is log-normally distributed with parameters μ and σ, then the natural log of that variable is normally distributed with a mean of μ and standard deviation of σ. The median of the lognormal distribution is equal to e^{μ} and the mean is equal to $e^{\mu + \frac{\sigma^2}{2}}$ (Figure 3b).
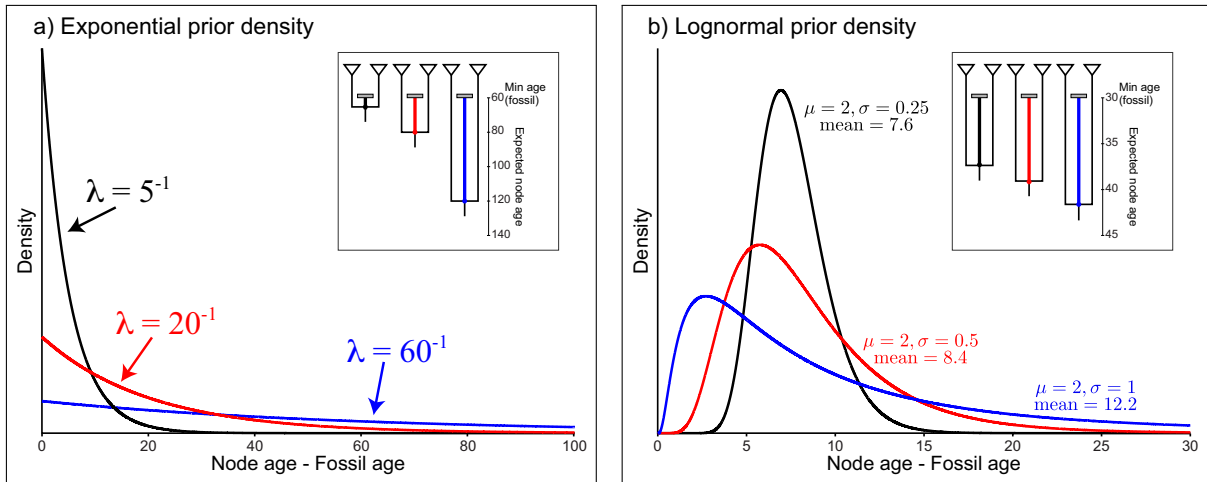
Figure 3: Two common prior densities for calibrating node ages. a) The exponential distribution with three different values for the rate parameter, λ. As the value of the λ rate parameter is decreased, the prior becomes less informative (the blue line is the least informative prior, $\lambda = 60^{-1}$). The inset shows an example of the three different priors and their expected values placed on the same node with a minimum age bound of 60. b) The lognormal distribution with 3 different values for the shape parameter, σ. For this distribution, even though μ is equal to 2.0 for all three, expected value (mean) is dependent on the value of σ. The inset shows an example of the three different priors and their expected values placed on the same node with a minimum age bound of 30.

1.4 Integrating Fossil Occurrence Times in the Speciation Model

Calibrating Bayesian divergence-time estimates using parametric densities (as described in the previous section: Sec. 1.3) is ultimately a difficult and unsatisfactory approach, particularly if the calibration information comes from fossil occurrence times. The calibration densities are typically applied in a multiplicative manner such that the prior probability of a calibrated node age is the product o the probability coming from the tree-wide speciation model and the probability under the calibration density (Heled and Drummond 2012; Warnock et al. 2012; Warnock et al. 2015). This approach leads to an incoherence and induces a prior that is inconsistent with the described calibration density. This statistical incoherence has been corrected by conditional tree prior models (Yang and Rannala 2006; Heled and Drummond 2012; Heled and Drummond 2013). These conditional approaches are an important contribution to the field, particularly when non-fossil data are used to calibrate an analysis. However, when using fossil information, it is more appropriate to account for the fact that the fossils are part of the same diversification process (*i.e.*, birth-death model) that generated the extant species.

1.4.1 The Fossilized Birth-Death Process

The exercise outlined in this tutorial demonstrates how to calibrate species divergence using the *fossilized birth-death* (FBD) model described in Stadler (2010) and Heath et al. (2014). This model simply treats the fossil observations as part of the prior on node times, such as the birth-death models outlined in Section 1.2 of this document. The fossilized birth-death process provides a model for the distribution of speciation times, tree topology, and distribution of lineage samples before the present (i.e., non-contemporaneous samples like fossils or viruses). Importantly, this model can be used *with or without* character data for the historical samples. Thus, it provides a reasonable prior distribution for analyses combining morphological or DNA data for both extant and fossil taxa—i.e, the so-called 'total-evidence' approaches described by Ronquist et al. (2012) (also see Pyron 2011). When matrices of discrete morphological characters for

both living and fossil species are unavailable, the fossilized birth-death model imposes a time structure on the tree by *marginalizing* over all possible attachment points for the fossils on the extant tree (Heath et al. 2014), therefore, some prior knowledge of phylogenetic relationships is important, much like for calibration-density approaches.

The FBD model describes the probability of the tree and fossils conditional on the birth-death parameters: $f[\mathcal{T} \mid \lambda, \mu, \rho, \psi, x_c]$, where $\mathcal{T}$ denotes the tree topology, divergence times, fossil occurrence times, and the times at which the fossils attach to the tree. The parameters of the model are:

λ	speciation rate
μ	extinction rate
ρ	probability of sampling extant species
ψ	fossil recovery rate
x_c	the starting time of the process, either x_0 or x_1

Figure 4A shows the probabilistic graphical model of the FBD process. Additionally, an example FBD tree is shown in Figure 4B, where the diversification process originates at time x_0, giving rise to $n = 20$ species in the present. All of the lineages represented in Figure 4B (both solid and dotted lines) show the *complete tree*. This is the tree of all extant *and* extinct lineages generated by the process. The complete tree is distinct from the *reconstructed tree* which is the tree representing only the sampled *extant* lineages. Fossil observations (red circles in Figure 4B) are recovered over the process along the lineages of the complete tree. If a lineage does not have any descendants sampled in the present, it is lost and cannot be observed, these are the dotted lines in Figure 4B. The probability must be conditioned on the starting time of the process x_c. This can be one of two different nodes in the tree, either the origin time x_0 or the root age x_1 (Figure 4B). The origin (x_0) of a birth death process is the starting time of the *stem* lineage, thus this conditions on a single lineage giving rise to the tree. Alternatively, a birth-death process can be conditioned on the age of the root (x_1), which is the time of the most-recent-common ancestor (MRCA) of the sampled lineages. Here, the model assumes that the branching process starts with two lineages, each of which has the same starting time.

An important characteristic of the FBD model is that it accounts for the probability of sampled ancestor-descendant pairs (Foote 1996). Given that fossils are sampled from lineages in the diversification process, the probability of sampling fossils that are ancestors to taxa sampled at a later date is correlated with the turnover rate ($r = \mu/\lambda$) and the fossil recovery rate (ψ). This feature is important, particularly for datasets with many sampled fossils. In the example (Figure 4B), several of the fossils have sampled descendants. These fossils have solid black lines leading to the present.

Recently, Gavryushkina et al. (2014) extended the fossilized birth-death model to allow more flexibility in the assignment of fossils to clades and conditioning parameters and implemented this version in the BEAST2 package called **SA** ("Sampled Ancestors", by Alexandra Gavryushkina). Notably, this implementation also allows you to use the FBD model as a tree prior for datasets combining molecular data from extant taxa with morphological data for both fossil and extant species. In this implementation of the FBD model, instead of the parameters λ, μ, and ψ, the following parameters are used in MCMC optimization:

$d = \lambda - \mu$	Net diversification rate
$r = \mu/\lambda$	Turnover
$s = \psi/(\mu + \psi)$	Probability of fossil observation prior to species extinction

Figure 4: The *fossilized birth-death process*. (A) The probabilistic graphical model of the FBD process for divergence-time estimation. All of the relevant parameters are labeled, with square-shaped nodes representing constant parameters, circles with solid borders indicate stochastic variables, deterministic nodes are shown in circles with dotted borders, and shaded nodes indicate observed variables. This model connects to the rest of the phylogenetic continuous-time Markov model through the time-tree variable. (B) A FBD tree generated by single realization of the process; simulated with the parameters: $\lambda = 0.02, \mu = 0.01, \psi = 0.01, \rho = 1$. The red circles represent recovered fossil occurrence times. Sampled lineages are shown with solid black lines and unobserved lineages with dotted lines.

The parameters λ, μ, and ψ needed for the FBD probability are recovered via:

$$\lambda = \frac{d}{1-r}, \quad \mu = \frac{rd}{1-r}, \quad \psi = \frac{s}{1-s}\frac{rd}{1-r}.$$

Conveniently, under the d, r, s, ρ parameterization, three parameters $r, s, \rho \in [0, 1]$, with only d on the interval $(0, \infty)$, thus, this allows for easier prior distribution construction under this model (Heath et al. 2014). Furthermore, it is important to carefully consider the relevant parameters of the birth-death process and the choice of the d, r, s parameterization allows for estimation of the diversification rate and turnover rate. Typically, macroevolutionary studies are interested in these parameters.

The fossilized birth-death process provides a way to integrate fossil occurrence times into the tree prior and is currently the most appropriate way to calibrate divergence time estimates when the calibration dates represent fossil occurrence times. However, this approach is not suitable for dating trees using biogeographical information or secondary calibration dates. In these cases, it is better to use conditioned birth-death processes (Yang and Rannala 2006; Heled and Drummond 2012; Heled and Drummond 2013). Furthermore, advances in methods for modeling historical biogeography (Landis et al. 2013) also present some exciting potential for accommodating this information in methods for dating species divergences.

2   Programs used in this Exercise

BEAST – Bayesian Evolutionary Analysis Sampling Trees

BEAST is a free software package for Bayesian evolutionary analysis of molecular sequences using MCMC and strictly oriented toward inference using rooted, time-measured phylogenetic trees (Drummond et al. 2006; Drummond and Rambaut 2007; Bouckaert et al. 2014). The development and maintenance of BEAST is a large, collaborative effort and the program includes a wide array of different types of analyses:

- Phylogenetic tree inference under different models for substitution rate variation
  - Constant rate molecular clock (Zuckerkandl and Pauling 1962)
  - Uncorrelated relaxed clocks (Drummond et al. 2006)
  - Random local molecular clocks (Drummond and Suchard 2010)
- Estimates of species divergence dates and fossil calibration under a wide range of branch-time models and calibration methods
- Analysis of non-contemporaneous sequences
- Heterogenous substitution models across data partitions
- Population genetic analyses
  - Estimation of demographic parameters (population sizes, growth/decline, migration)
  - Bayesian skyline plots
  - Phylogeography (Lemey et al. 2009)
- Gene-tree/species-tree inference (∗BEAST; Heled and Drummond 2010)
- and more...

BEAST is written in java and its appearance and functionality are consistent across platforms. Inference using MCMC is done using the BEAST program, however, there are several utility applications that assist in the preparation of input files and summarize output (BEAUti, LogCombiner, and TreeAnnotator are all part of the BEAST software bundle).

There are currently two available versions of the BEAST package:

   BEAST v1.8 http://beast.bio.ed.ac.uk (BEAST 1)
   BEAST v2∗ http://www.beast2.org (BEAST 2).

BEAST 2 is a complete re-write of BEAST 1, with different design choices (Bouckaert et al. 2014). The BEAST 2 package allows for implementation and distribution of new models and methods through *add-ons* (also called "plugins"). Add-ons include SNAPP (phylogenetic analysis using SNP and AFLP data) and BDSSM (a birth-death skyline model for serially-sampled data), as well as several others that are available or in development. It is important to note, however, that the set of analyses and models available in the BEAST 2 package do not completely overlap with the set of analyses with BEAST 1 (though this should not be the case in the near future). I strongly encourage you to learn more about BEAST and the BEAST v2 software by reading the book provided online by the developers: *Bayesian Evolutionary Analysis with BEAST 2* (Drummond and Bouckaert 2014).

BEAUti – Bayesian Evolutionary Analysis Utility

BEAUti is a utility program with a graphical user interface for creating BEAST and *BEAST input files which must be written in the eXtensible Markup Language (XML). This application provides a clear way to specify priors, partition data, calibrate internal nodes, etc.

LogCombiner – When multiple (identical) analyses are run using BEAST (or MrBayes), LogCombiner can be used to combine the parameter log files or tree files into a single file that can then be summarized using Tracer (log files) or TreeAnnotator (tree files). However, it is important to ensure that all analyses reached convergence and sampled the same stationary distribution before combining the parameter files.

TreeAnnotator – TreeAnnotator is used to summarize the posterior sample of trees to produce a maximum clade credibility tree and summarize the posterior estimates of other parameters that can be easily visualized on the tree (e.g. node height). This program is also useful for comparing a specific tree topology and branching times to the set of trees sampled in the MCMC analysis.

Tracer – Tracer is used for assessing and summarizing the posterior estimates of the various parameters sampled by the Markov Chain. This program can be used for visual inspection and assessment of convergence and it also calculates 95% credible intervals (which approximate the 95% highest posterior density intervals) and effective sample sizes (ESS) of parameters (http://tree.bio.ed.ac.uk/software/tracer).

FigTree – FigTree is an excellent program for viewing trees and producing publication-quality figures. It can interpret the node-annotations created on the summary trees by TreeAnnotator, allowing the user to display node-based statistics (e.g. posterior probabilities) in a visually appealing way (http://tree.bio.ed.ac.uk/software/figtree).

3 The eXtensible Markup Language

The eXtensible Markup Language (XML) is a general-purpose markup language, which allows for the combination of text and additional information. In BEAST, the use of the XML makes analysis specification very flexible and readable by both the program and people. The XML file specifies sequences, node calibrations, models, priors, output file names, etc. BEAUti is a useful tool for creating an XML file for many BEAST analyses. However, typically, dataset-specific issues can arise and some understanding of the BEAST-specific XML format is essential for troubleshooting. Additionally, there are a number of interesting models and analyses available in BEAST that cannot be specified using the BEAUti utility program. Refer to the BEAST web page (http://beast.bio.ed.ac.uk/XML_format) for detailed information about the BEAST XML format. Box 1 shows an example of BEAST XML syntax for specifying a birth-death prior on node times.

```
<!- An exponential prior distribution on the gamma shape parameter of the irbp gene   ->
<prior id="GammaShapePrior.s:irbp" name="distribution" x="@gammaShape.s:irbp">
    <Exponential id="Exponential.01" name="distr">
        <parameter id="RealParameter.01" lower="0.0" name="mean" upper="0.0">1.0</↩
            parameter>
    </Exponential>
</prior>
```

Box 1: BEAST 2 XML specification of an exponential prior density on the shape of a gamma distribution.

4 Practical: Divergence Time Estimation

This tutorial will walk you through an analysis of the divergence times of the bears. The occurrence times of 14 fossil species are integrated into the tree prior to impose a time structure on the tree and calibrate the analysis to absolute time. Additionally, an uncorrelated, lognormal relaxed clock model is used to describe the branch-specific substitution rates.

Make sure you have already downloaded the software listed in Section 2.

4.1 The Data

Download data and pre-cooked output files from:

https://taming-the-beast.github.io/tutorials/FBD-tutorial

The analysis in this tutorial includes data from several different sources. We have molecular sequence data for eight extant species, which represent all of the living bear taxa. The sequence data include interphotoreceptor retinoid-binding protein (irbp) sequences (in the file **bears_irbp_fossils.nex**) and 1000 bps of the mitochondrial gene cytochrome b (in the file **bears_cytb_fossils.nex**). If you open either of these files in your text editor or alignment viewer, you will notice that there are 22 taxa listed in each one, with most of these taxa associated with sequences that are entirely made up of missing data (i.e., **?????**). The NEXUS files contain the names of 14 fossil species, that we will include in our analysis as calibration information for the fossilized birth-death process. Further, we must provide an occurrence time for each taxon sampled from the fossil record. For the fossil species, this information is obtained from the literature or fossil databases like the Fossilworks PaleoDB or the Fossil Calibration Database, or from your own paleontological expertise. The 14 fossil species used in this analysis are listed in Table 1 along with the age range for the specimen and relevant citation. For this exercise, we will fix the ages to a value within the age range provided in Table 1. In BEAST2, it is possible to use MCMC to sample the occurrence time for a fossil conditional on a prior distribution. However, the options to do this are not currently available in BEAUti, and will not be covered by this tutorial. The age of each taxon is encoded in the taxon name following the last '_' character. For example, the fossil panda *Kretzoiarctos beatrix* has an age of approximately 11.7 Mya, thus, the taxon name in the alignment files is: **Kretzoiarctos_beatrix_11.7**. Similarly, since the polar bear, *Ursus maritimus*, represents an extant species, its occurrence time is 0.0 Mya, which makes its taxon name: **Ursus_maritimus_0**. By including the tip ages in the taxon names, we can easily import these values into BEAUti while setting up the XML file. This is simply easier than entering them in by hand (which is also possible).

The final source of data required for our analysis is some information about the phylogenetic placement of the fossils. This prior knowledge can come from previous studies of morphological data and taxonomy. Ideally, we would know exactly where in the phylogeny each fossil belongs. However, this is uncommon for most groups in the fossil record. Often, we can place a fossil with reasonable resolution by assigning it to some total group. For example, if a fossil specimen has all of the identifying characters of a bear in the subfamily Ursinae, then, we can create a monophyletic group of all known Ursinae species and our fossil. Here, we would be assigning the fossil to the *total group* Ursinae, meaning that the fossil can be a crown or stem fossil of this group. For some fossils, we may have very little data to inform their placement in

Table 1: Fossil species used for calibrating divergence times under the FBD model. Modified from Table
S.3 in the supplemental appendix of Heath et al. (2014).

Fossil species	Age range (My)	Citation
Parictis montanus	33.9–37.2	Clark and Guensburg 1972; Krause et al. 2008
Zaragocyon daamsi	20–22.8	Ginsburg and Morales 1995; Abella et al. 2012
Ballusia elmensis	13.7–16	Ginsburg and Morales 1998; Abella et al. 2012
Ursavus primaevus	13.65–15.97	Andrews and Tobien 1977; Abella et al. 2012
Ursavus brevihinus	15.97–16.9	Heizmann et al. 1980; Abella et al. 2012
Indarctos vireti	7.75–8.7	Montoya et al. 2001; Abella et al. 2012
Indarctos arctoides	8.7–9.7	Geraads et al. 2005; Abella et al. 2012
Indarctos punjabiensis	4.9–9.7	Baryshnikov 2002; Abella et al. 2012
Ailurarctos lufengensis	5.8–8.2	Jin et al. 2007; Abella et al. 2012
Agriarctos spp.	4.9–7.75	Abella et al. 2011; Abella et al. 2012
Kretzoiarctos beatrix	11.2–11.8	Abella et al. 2011; Abella et al. 2012
Arctodus simus	0.012–2.588	Churcher et al. 1993; Krause et al. 2008
Ursus abstrusus	1.8–5.3	Bjork 1970; Krause et al. 2008
Ursus spelaeus	0.027–0.25	Loreille et al. 2001; Krause et al. 2008

the tree and perhaps we may only know that it falls somewhere within our group of interest. In this case,
we can account for our uncertainty in the relationship of the fossil and all other taxa and allow MCMC to
sample all possible places where the fossil can attach in the tree. For the bear species in our analysis, we
have some prior knowledge about their relationships, represented as an unresolved phylogeny in Figure 5.
Four out of five of the clades shown in Figure 5 are defined in the NEXUS file `bears_cytb_fossils.nex`
as `taxsets` in the `sets` block. The one clade that is not included in the data file will be created using the
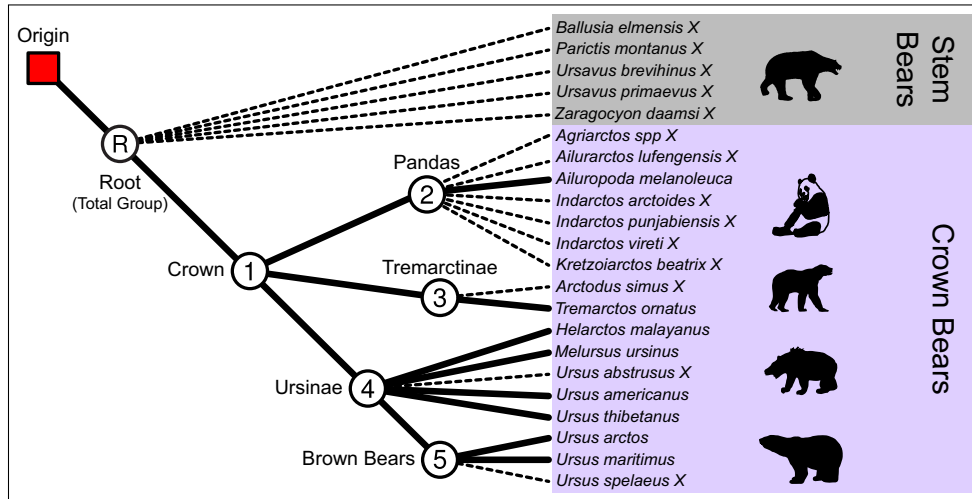BEAUti options.

Figure 5: The phylogenetic relationships of crown and stem bears based on taxonomy and morphological data. The
resolution of monophyletic clades is based on well-supported previous analyses. Monophyletic clades are indicated
with labeled circles. In addition to the root node (R), there are 5 nodes defining clades containing both fossil and
extant species. The origin of the tree is indicated with a red square. The time of this node represents the start of
the diversification process that generated these linages. The extant lineages are shown with heavy, solid lines and
the fossil lineages are dotted lines.

4.2   Creating the Analysis File with BEAUti

Creating a properly-formatted BEAST XML file from scratch is not a simple task. However, BEAUti provides a simple way to navigate the various elements specific to the BEAST XML format.

> Begin by executing the BEAUti program

Be sure that this is the version that came from the BEAST 2 download from: http://beast2.org. For Mac OSX and Windows, you can do this by double clicking on the application. For Unix systems (including Mac OSX), it is convenient to add the entire **BEAST/bin** directory to your path.

4.2.1   Install BEAST 2 Plug-Ins

Next, we have to install the BEAST 2 packages (also called "plug-ins" or "add-ons") that are needed for this analysis. The package that we will use is called **SA**.

> Open the **BEAST 2 Package Manager** by navigating to **File→Manage Packages** in the menu. [Figure 6]

In the package manager, you can install all of the available plug-ins for BEAST 2. These include a number of packages for analyses such as species delimitation (DISSECT, STACEY), population dynamics (MASTER), the phylodynamics of infectious disease (BDSKY, phylodynamics), etc.

> Install the **SA** package by selecting it and clicking the **Install/Upgrade** button. [Figure 6]

Name	Status/Version	Latest	Dependencies	Detail
bacter	un–installed	1.0.0–pre6		ClonalOrigin ARG inference.
BASTA	un–installed	2.1.0		Bayesian structured coalescent approxi…
BDSKY	1.2.2	1.2.2		birth death skyline – handles serially sa…
BEAST_CLASSIC	un–installed	1.2.1	BEASTLabs	BEAST classes ported from BEAST 1 in …
BEASTLabs	un–installed	1.3.2		BEAST utilities, such as multi threaded t…
BEASTShell	un–installed	1.1.1		BEAST Shell – BeanShell scripting for BE…
bModelTest	un–installed	0.1.3		Bayesian model test for nucleotide subs…
CA	un–installed	1.1.0		CladeAge aPackage for fossil calibrations
DISSECT	un–installed	1.2.0		Species delimitation with *BEAST
GEO_SPHERE	un–installed	0.1.4		Whole world phylogeography
MASTER	un–installed	4.1.3		Stochastic population dynamics simulation
MGSM	un–installed	0.1.3		Multi–gamma and relaxed gamma site …
MODEL_SELECTION	un–installed	1.1.3		Select models through path sampling/st…
morph–models	1.0.3	1.0.3		Enables models of morphological chatac…
MultiTypeTree	un–installed	5.5.0		Structured coalescent inference
phylodynamics	un–installed	1.1.2	BDSKY	birth death skyline model
PoMo	un–installed	0.1.1		PoMo, a substitution model that separat…
RBS	un–installed	1.2.5		Reversible–jump Based substitution mo…
SA	1.1.3	1.1.3		Sampled ancestor trees
SCOTTI	un–installed	1.0.1		Structured COalescent Transmission Tre…
SNAPP	un–installed	1.2.5		SNP and AFLP Phylogenies
STACEY	un–installed	1.0.5		Species delimitation and species tree es…
SubstBMA	un–installed	1.2.0		Substitution Bayesian Model Averaging

Figure 6: The BEAST2 Package Manager.

It is important to note that you only need to install a BEAST 2 package once, thus if you close BEAUti, you don't have to load **SA** the next time you open the program. However, it is worth checking the package manager for updates to plug-ins, particularly if you update your version of BEAST 2.

> Close the **BEAST 2 Package Manager** and <u>**restart BEAUti**</u> to fully load the **SA** package.

4.2.2   Import Alignments

> Navigate to the **Partitions** window (you should already be here).

Next we will load the alignment files for each of our genes. Note that separate loci can be imported as separate files or in a single NEXUS file with partitions defined using the `ASSUMPTIONS` command.

> Using the menu commands **File→Import Alignment**, import the data files:
> `bears_irbp_fossils.nex` and `bears_cytb_fossils.nex`

Now that the data are loaded into BEAUti, we can unlink the site models, link the clock models, link the trees and rename these variables.

> Highlight both partitions (using shift+click) and click on **Unlink Site Models** to assume different models of sequence evolution for each gene (the partitions are typically already unlinked by default).
>
> Now click the **Link Clock Models** button so that the two genes have the same relative rates of substitution among branches.
>
> Finally click **Link Trees** to ensure that both partitions share the same tree topology and branching times.

It is convenient to rename some of the variables in the **Partitions** window. By doing this, the parameters associated with each partition that are written to file are a bit more intuitively labeled.

> Double click on the site model for the cytochrome b gene, it is currently called `bears_cytb_fossils`. Rename this: `cytb`. (Note that you may have to hit the return or enter key after typing in the new label for the new name to be retained.)
>
> Do the same for the site model for the other gene, calling it `irbp`.
>
> Rename the clock model `bearsClock`.
>
> Rename the tree `bearsTree`. [Figure 7]

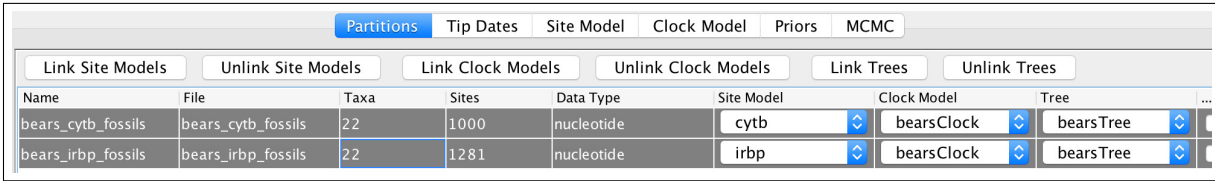Figure 7 shows how the final **Partitions** window should look.

Figure 7: The **Partitions** window after unlinking the site models, linking the clock models, linking the trees, and renaming the XML variables.

4.2.3 Set Tip Dates

Navigate to the **Tip Dates** panel.

We must indicate that we have sequentially sampled sequences. When performing an analysis without dated tips or any fossil information, you can skip this window, and BEAST will assume that all of your samples are contemporaneous.

Toggle on the **Use tip dates** option.

The next step involves specifying how the dates are oriented on the tree and the units they are in. We will indicate that the dates are in **years**, even though they are in fact in units of *millions of years*. This is because the units themselves are arbitrary and this scale difference will not matter. Additionally, we will tell BEAUti that the zero time of our tree is the present and the ages we are providing are the number of years **Before the present**.

Change: **Dates specified as: year Before the present** [Figure 8]

For some types of analyses, such as serially sampled viruses, the dates given are relative to some time in the past, thus this option is available as well.
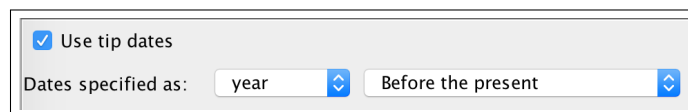
Figure 8: Specifying the units and reference point of the fossil dates.

When inputing the dates for each tip species, one option is to enter each one by hand. This may be quite onerous if you have many fossils or many sequences sampled back in time. Conveniently, these dates can be included in the taxon names so that BEAUti can easily extract them for us using the **Guess** option.

Click on the **Guess** button.

This will open a window where you can specify the pattern in the taxon names from which the tip ages can be extracted. Obviously, it's better to make this a fairly simple code that doesn't require multiple iterations of searches. Moreover, if this is straightforward, then you will be able to easily eliminate these dates when creating figures from your final summary tree.

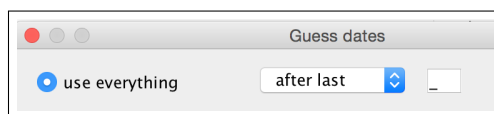Tell BEAUti to **use everything** after the **last** '**_**', then click **OK**. [Figure 9]

Figure 9: Specify the text pattern in the taxon names that identifies the tip's age using the **Guess dates** option.

You should now see that the tip ages have been filled in for all of the fossil taxa and that the same age is listed in the **Date** and **Height** column for each species.

4.2.4 Specify the Sites Model

Navigate to the **Site Model** window.

The molecular sequence data sampled for each extant bear species are from two different genes: the mitochondrial cytochrome b gene (`cytb`) and the nuclear interphotoreceptor retinoid-binding protein gene (`irbp`). We will partition these loci into two separate alignments and apply different models of sequence evolution to each one. For the `cytb` gene, we will apply a general-time reversible model with homogeneous rates across sites: GTR. For the nuclear gene `irbp`, we will assume a 2-rate model where transitions and transversions happen at different rates, and that the rates vary across the alignment according to a mean-one gamma distribution: HKY+Γ.

Select the `cytb` gene and change the **Subst Model** to **GTR**.

Toggle on **estimate** for the **Substitution Rate**. [Figure 10]

By changing the substitution model, we have now introduced additional parameters for the GTR exchangeability rates. We will construct priors for these parameters later on.

Select the `irbp` gene and change the **Subst Model** to **HKY**.

To indicate gamma-distributed rates, set the **Gamma Category Count** to 4.

Then switch the **Shape** parameter to **estimate**.

Toggle on **estimate** for the **Substitution Rate**. [Figure 11]

Figure 10: The fully specified site model for the `cytb` gene: GTR.

Figure 11: The fully specified site model for the `irbp` gene: HKY+Γ.

Now both models are fully specified for the unlinked genes. Note that **Fix mean substitution rate** is always specified and we also have indicated that we wish to **estimate** the **Substitution Rate** for each gene. This means that we are estimating the *relative* substitution rates for our two loci.

4.2.5 The Clock Model

Navigate to the **Clock Model** window.

Here, we can specify the model of lineage-specific substitution rate variation. The default model in BEAUti is the **Strict Clock** with a fixed substitution rate equal to **1**. Three models for relaxing the assumption of a constant substitution rate can be specified in BEAUti as well. The **Relaxed Clock Log Normal** option assumes that the substitution rates associated with each branch are independently drawn from a single, discretized lognormal distribution (Drummond et al. 2006). Under the **Relaxed Clock Exponential** model, the rates associated with each branch are exponentially distributed (Drummond et al. 2006). The **Random Local Clock** uses Bayesian stochastic search variable selection to average over random local molecular clocks (Drummond and Suchard 2010). For this analysis we will use the uncorrelated, lognormal model of branch-rate variation.

Change the clock model to **Relaxed Clock Log Normal**.

The uncorrelated relaxed clock models in BEAST2 are discretized for computational feasibility. This means that for any given parameters of the lognormal distribution, the probability density is discretized into some number of discrete rate bins. Each branch is then assigned to one of these bins. By default, BEAUti sets the **Number Of Discrete Rates** to **-1**. This means that the number of bins is equal to the number of branches.

The fully specified **Clock Model** assumes that the rates for each branch are drawn independently from a single lognormal distribution. The mean of the rate distribution will be estimated, thus we can account for uncertainty in this parameter by placing a prior distribution on its value. Note that there is an option to **Normalize** the average clock rate. We will leave this unchecked.

4.2.6   Priors on Parameters of the Site Models

Navigate to the **Priors** window.

In the **Priors** window, all of the parameters and hyperparameters (and hyper-*hyper*parameters, etc.) specific to the models defined in the **Site Model** and **Clock Model** windows are listed. Here you can set up the prior distributions on these parameters, as well as define calibration nodes and calibration densities and specify a tree model. One convenient feature of BEAUti is that the list of parameters changes dynamically as you change the models. Thus, if you missed a step along the way, you would notice at this point because something might be missing here. For example, if you did not change the substitution mode for **cytb** from *JC69* to *GTR* in the **Site Model** window, then you would not see the exchangeability rates and base frequency parameters listed for **cytb**. [Figure 12]
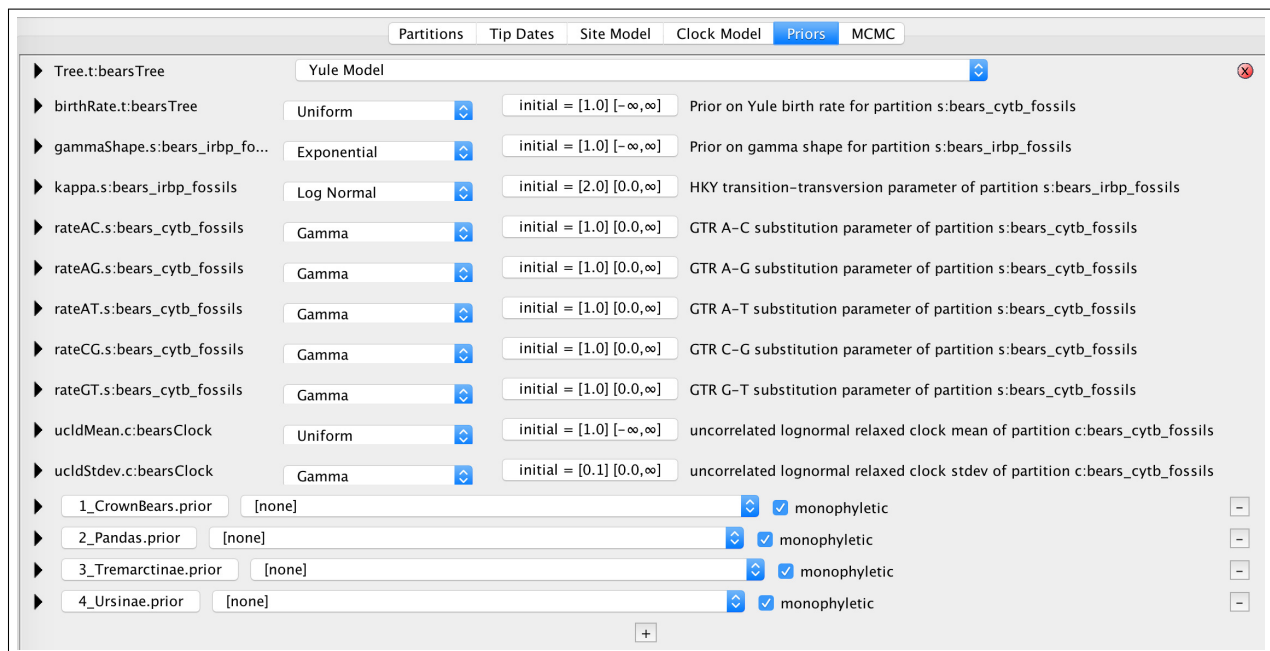

Figure 12: The **Priors** window with default (unmodified) settings.

In the **Priors** panel we will begin by specifying priors for the parameters associated with the sites models. Since we partitioned the two genes, there are parameters for the two different models:

- **cytb**: exchangeability rates for the GTR model (*rateAC.s:cytb*, *rateAG.s:cytb*, ...)
- **irbp**: the transition-transversion rate ratio (*kappa.s:irbp*) and the shape parameter of the gamma distribution on site rates (*gammaShape.s:irbp*)

Note that the base frequencies for each of these models are not listed in the **Priors** window, though they are estimated.

We will keep the default priors for the HKY model on the evolution of **irbp**. The default gamma priors on the GTR exchangeability rates for the **cytb** gene place a lot of prior density on very small values. For some datasets, the sequences might not be informative for some of the rates, consequentially the MCMC may propose values very close to zero and this can induce long mixing times. Because of this problem, we will alter the gamma priors on the exchangeability rates. For each one, we will keep the expected values as in the default priors. The default priors assume that transitions (A↔G or C↔T) have an expected rate of 1.0. Remember that we fixed the parameter *rateCT* to equal 1.0 in the **Site Model** window, thus this parameter isn't in the **Priors** window. For all other rates, transversions, the expected value of the priors is lower: 0.5. In BEAST, the gamma distribution is parameterized by a shape parameter (*Alpha*) and a scale parameter (*Beta*). Under this parameterization, expected value for any gamma distribution is: $\mathbb{E}(x) = \alpha\beta$. To reduce the prior density on very low values, we can increase the shape parameter and then we have to adjust the scale parameter accordingly.

> Begin by changing the gamma prior on the transition rate *rateAG.s:cytb*. Clicking on the ▶ next to this parameter name to reveal the prior options. Change the parameters: *Alpha* = 2 and *Beta* = 0.5. [Figure 13, bottom]
>
> Then change all of the other rates: *rateAC.s*, *rateAT.s*, *rateCG.s*, *rateGT.s*. For each of these, change the parameters to: *Alpha* = 2 and *Beta* = 0.25. [Figure 13, top]

Figure 13: Gamma prior distributions on two of the five relative rates of the GTR model.

4.2.7   Priors for the Clock Model

Since we are assuming that the branch rates are drawn from a lognormal distribution, this induces two hyperparameters: the mean and standard deviation (**ucldMean.c** and **ucldStdev.c** respectively). By default, the prior distribution on the **ucldMean.c** parameter is an improper, uniform distribution on the interval $(0, \infty)$. Note that this type of prior is called improper because the prior density of a *uniform* distribution with infinite bounds does not integrate to 1. Although improper priors can sometimes lead to proper posterior distributions, they may also have undesired effects and cause problems with mixing and convergence.

Reveal the options for the prior on **ucldMean.c** by clicking on the ▶. Change the prior density to an *Exponential* with a mean of `10.0`. [Figure 14]

Figure 14: The exponential prior distribution on the mean of the log normal relaxed clock model.

The other parameter of our relaxed-clock model is, by default assigned a gamma prior distribution. However, we have a strong prior belief that the variation in substitution rates among branches is low, since some previous studies have indicated that the rates of molecular evolution in bears is somewhat clock-like (Krause et al. 2008). Thus, we will assume an exponential prior distribution with 95% of the probability density on values less than 1 for the **ucldStdev.c** parameter.

Reveal the options for the prior on **ucldStdev.c** by clicking on the ▶. Change the prior density to an *Exponential* with a mean of `0.3337`. [Figure 15]

Figure 15: The exponential prior distribution on the standard deviation of the log normal relaxed clock model.

4.2.8   The Tree Prior

Next we will specify the prior distribution on the tree topology and branching times. You should notice an error notification (a red circle with an "X" in it) in the **Priors** panel to the left of **Tree.t:bearsTree** (Figure 12). If you mouse over this notification, you will se a message telling you that the default **Yule Model** is not appropriate for non-contemporaneous tips and that you must choose a different tree prior. Thus, here is where we specify the *fossilized birth-death process.*

Change the tree model for **Tree.t:bearsTree** to **Fossilized Birth Death Model**.

Reveal the options for the prior on **Tree.t** by clicking on the ▶.

Origin Time — In Section 1.4.1, the parameters of the FBD model are given. Remember that this model, like any branching process (i.e., constant rate birth-death, Yule) can be conditioned on either the origin time or the root age. Depending on the available prior information or the type of data available, it makes sense to condition on one or the other (but not both, obviously). If you know that all of the fossils in your dataset are *crown* fossils—descendants of the MRCA of all the extant taxa—and you have some prior knowledge of the age of the clade, then it is reasonable to condition the FBD on the root. Alternatively, if the fossils in your analysis are stem fossils, or can only reliably be assigned to your total group, then it is appropriate to condition on the origin age.

For this analysis, we have several bear fossils that are considered stem fossils, thus we will condition on the origin age. Previous studies (dos Reis et al. 2012) estimated an age of approximately 45.5 My for the MRCA of seals and bears. We will use this time as a starting value for the origin.

Set the starting value of the **Origin** to 45.5 and specify that this parameter will be estimated by checking the **estimate** box. (You may have to expand the width of the BEAUti window to see the check-boxes for these parameters.) [Figure 16]

Figure 16: The initial values and conditions for the fossilized birth-death process (Stadler 2010; Heath et al. 2014; Gavryushkina et al. 2014)

Since we are estimating the origin parameter, we must assign a prior distribution to it (unless we wish to keep the default Uniform(0,∞) prior). We will assume that the origin time is drawn from a lognormal distribution with an expected value (mean) equal to 45.5 My and a standard deviation of 1.0.

Reveal the options for the prior on ***originFBD.t*** by clicking on the ▶.

Change the prior distribution to ***Log Normal***.

Check the box marked ***Mean In Real Space*** and set the mean ***M*** equal to `8.5` and the standard deviation ***S*** to `1.0`.

Set the ***Offset*** of the lognormal distribution to equal the age of the oldest fossil: `37.0`. [Figure 17]

Figure 17: The lognormal prior distribution on the origin time.

Notice that the options for the ***Log Normal*** prior distribution allow you to specify ***Mean in Real Space***. If you choose this option, then the mean value you enter is the expected value of the lognormal distribution. We used this option above to specify an expected value of 8.5 My. You will create the exact same prior distribution if you uncheck the ***Mean in Real Space*** option and enter the location parameter μ of the distribution and give it a value of `1.640066`. It is important that you are very careful when specifying these parameters. If, for example, ***Mean in Real Space*** was not checked and you provided a value of `8.5` for ***M***, then your calibration prior would be extremely diffuse, with a *mean* value of 8,140.084!

Diversification Rate — In Section 1.4.1, we discussed that the diversification rate is: $d = \lambda - \mu$. Generally, we think that this value is fairly small, particularly since we have few extant species and many fossils. Therefore, an exponential distribution is a reasonable prior for this parameter as it places the highest probability on zero. For this analysis we will assume $d \sim$ Exponential(1.0). The exponential distribution's only parameter is called the *rate* parameter (denoted ν) and it controls both the mean and variance of the distribution (here $\nu = 1$). The mean of an exponential distribution is the inverse of the rate: $\mathbb{E}(d) = \nu^{-1} = 1$. Importantly, in BEAUti/BEAST, when specifying an exponential prior, you provide the *mean* and not the rate parameter.

Reveal the options for the prior on ***diversificationRateFBD.t*** by clicking on the ▶.

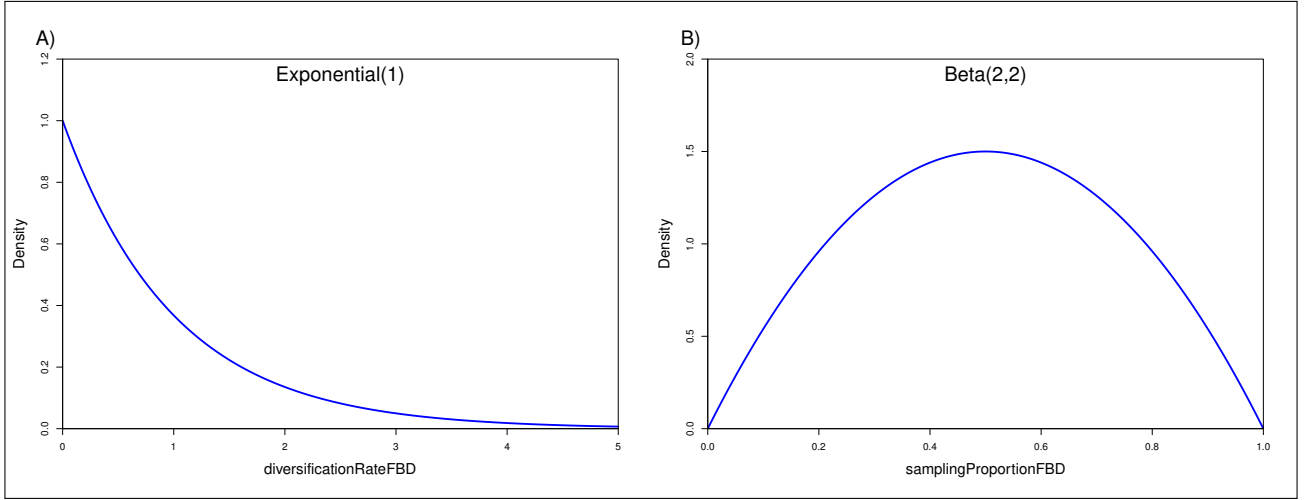Change the prior distribution to ***Exponential*** with a ***Mean*** equal to `1.0`. [Figure 18A]

Figure 18: Prior distributions on FBD parameters. (A) An exponential prior with a mean of 1 describes the distribution on the diversification rate ($d = \lambda - \mu$). (B) The sampling proportion is the probability of observing a fossil prior to the lineage extinction ($s = \psi/(\mu + \psi)$). Because this parameter is on the interval [0,1], we assume a beta prior density with $\alpha = \beta = 2$.

SAMPLING PROPORTION — The sampling proportion is the probability of observing a lineage as a fossil before that lineage goes extinct. This parameter is a function of the extinction rate (μ) and fossil recovery rate (ψ): $s = \psi/(\mu + \psi)$. Let's say that we have prior knowledge that this parameter is approximately equal to 0.5, and that we wish extreme values (very close to 0 or 1) to have low probability. This prior density can be described with a beta distribution. The beta distribution is a probability density over values between 0 and 1 and is parameterized by two values, called α and β. A beta distribution with $\alpha = \beta = 1$ is equivalent to a uniform distribution between 0 and 1. By changing the parameters, we can assign higher probability to values closer to 1 or 0. The mean of the beta distribution on s is: $\mathbb{E}(s) = \frac{\alpha}{\alpha+\beta}$. Thus, if $\alpha = \beta$, then $\mathbb{E}(s) = 0.5$. For this prior we will set $\alpha = \beta = 2$.

> Reveal the options for the prior on ***samplingProportionFBD.t*** by clicking on the ▶.
>
> Change the prior distribution to ***Beta*** with ***Alpha*** equal to `2.0` and ***Beta*** equal to `2.0`. [Figure 18B]

THE PROPORTION OF SAMPLED EXTANT SPECIES — The parameter ρ (***Rho***) represents the probability of sampling a tip in the present. For most birth-death processes, it is helpful to be able to fix or place a very strong prior on one of the parameters (λ, μ, ρ, ψ) because of the strong correlations that exist among them. Typically, we may have the most prior knowledge about the proportion of sampled extant species (ρ). The diversity of living bears is very well understood and we know that there are only 8 species around today. Since we have sequence data representing each species, we can then fix $\rho = 1$, thus we do not need to specify a prior for this parameter.

TURNOVER — This parameter represents the relative rate of extinction: $r = \mu/\lambda$. For most implementations of the birth-death process, we assume that $\mu < \lambda$, thus this parameter is always $r < 1$. Large values of r that are close to 1.0, indicate high extinction, and values close to 0, indicate very little extinction. It is more challenging to define an appropriate prior for the turnover parameter and we will simply assume that all values on the interval [0,1] have equal probability. Thus, we can leave the default prior, a Uniform(0,1), on this parameter.

4.2.9 Creating Taxon Sets

Since some of the relationships of the fossil and living bears are well understood (from previous analyses of molecular and morphological data), we can incorporate this prior information by creating monophyletic taxon sets in BEAUti. If we do not impose any phylogenetic structure on the fossil lineages, they will have equal probability of attaching to any branch in the tree. Given that previous studies have provided information about the relationships of fossil bears, we can limit the MCMC to only sample within the known groups. For example, morphological analysis of fossil taxa place the sequoias *Kretzoiarctos beatrix* and several others in the subfamily Ailuropodinae, which includes pandas. Thus, if we create a monophyletic taxon set containing these taxa (*Ailuropoda melanoleuca, Indarctos vireti, Indarctos arctoides, Indarctos punjabiensis, Ailurarctos lufengensis, Agriarctos spp., Kretzoiarctos beatrix*) the prior probability that *K. beatrix* will attach to any lineage outside of this group is equal to 0.

There are five distinct clades within the phylogeny of bears that we can define (see Figure 5), and four out of five of these clades are defined as taxon sets in one of the nexus files containing our sequences (**bears_cytb_fossils.nex**). The taxa represented in our dataset are in the *total group* of bears. This includes all of the fossils that diverged before the most-recent-common-ancestor (MRCA) of all *living* bears. These early diverging fossils are *stem* lineages.

The first taxon set is the crown bears shown in Figure 19. The crown bears include all living species of bears and all the fossils that are descended from the MRCA of living taxa (node 1 in Figure 5). The MRCA of all crown bears and stem lineages is represented by the root node of our tree (node R in Figure 5). We do not have to specify a taxon set for the root node.
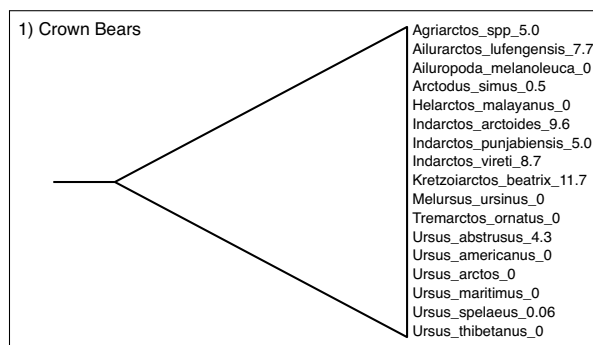

Figure 19: The species belonging to the Crown Bears clade.

If you view the file **bears_cytb_fossils.nex** in your text editor, you will see the four taxon sets defined in the **sets** block of the **NEXUS** file.

```
begin sets;
taxset 1_CrownBears = Agriarctos_spp_5.0 ... Ursus_thibetanus_0;
taxset 2_Pandas = Agriarctos_spp_5.0 ... Kretzoiarctos_beatrix_11.7;
taxset 3_Tremarctinae = Arctodus_simus_0.5 Tremarctos_ornatus_0;
taxset 4_Ursinae = Helarctos_malayanus_0 ... Ursus_thibetanus_0;
end;
```

When you have a **taxset** in your data file, this creates a defined clade in BEAUti that is constrained to be monophyletic.

You can also open the BEAUti taxon set for the crown bears by clicking the **1_CrownBears.prior** button. This will bring up the **Taxon set editor** where you can modify this taxon set (don't do that, though). [Figure 20]

Figure 20: The **Taxon set editor** used to create the clade containing the crown bear species.

The second node defined in our `NEXUS` file is the MRCA of all species within the subfamily Ailuropodinae. This group includes the giant panda (*Aliuropoda melanoleuca*) and six fossil relatives (Figure 21).

Figure 21: The species belonging to the Panda clade.

The subfamily Tremarctinae includes only the extant spectacled bear (*Tremarctos ornatus*) and the short-faced bear (*Arctodus simus*), shown in Figure 22. The occurrence time of the short-faced bear is only 500,000 years ago since it is known from the Pleistocene.

Figure 22: The two species in the clade Tremarctinae.

The subfamily Ursinae comprises all of the species in the genus *Ursus* (including two fossil species) as well as the sun bear (*Helarctos malayanus*) and the sloth bear (*Melursus ursinus*). These species are listed in Figure 23.

Finally, multiple studies using molecular data have shown that the polar bear (*Ursus maritimus*) and the brown bear (*Ursus arctos*) are closely related. Furthermore, phylogenetic analyses of ancient DNA from Pleistocene sub-fossils concluded that the cave bear (*Ursus spelaeus*) is closely related to the polar bear

Figure 23: The species belonging to the subfamily Ursinae.

and the brown bear (Figure 24). This taxon set is not included in our `NEXUS` file, so we must define it using BEAUti.

Figure 24: Three closely related *Ursus* species in the "Brown Bears" clade.

Create a new taxon set for node 5 by clicking the $\boxed{+}$ and select **MRCA prior** in the pop-up option box.

Label this taxon set `5_BrownBears`.

Move all of the taxa listed in this clade (Figure 24) to the right-hand column, click **OK** [Figure 25].

Back in the **Priors** window, check the box labeled **monophyletic** for node 5.

Figure 25: The **Taxon set editor** used to create the clade containing the brown bear, polar bear, and cave bear.

At this point, there should be five, monophyletic taxon sets listed in the **Priors** window.

4.2.10   Other BEAUti Options

There are two additional windows that are hidden in BEAUti by default. You can reveal them by selecting **View**→**View All** from the pull-down menu above. This will reveal the **Initialization** and **Operators** panels. The **Initialization** options allow you to change the starting values for the various parameters and specify if you want them estimated or fixed. The **Operators** menu contains a list of the parameters and hyperparameters that will be sampled over the course of the MCMC run. In this window, it is possible to turn off any of the elements listed to fix a given parameter to its starting value. For example, if you would like to estimate divergence times on a fixed tree topology (using a starting tree that you provided), then disable proposals operating on the `Tree`. For this exercise, leave both windows unmodified.

4.2.11   Set MCMC Options and Save the XML File

> Navigate to the **MCMC** window.

Now that you have specified all of your data elements, models, priors, and operators, go to the **MCMC** tab to set the length of the Markov chain, sample frequency, and file names. By default, BEAST sets the number of generations to 10,000,000.

> Since we have a limited amount of time for this exercise, change the **Chain Length** to 2,000,000. (Runtimes may vary depending on your computer, if you have reason to believe that this may take a very long time, then change the run length to something smaller.)

Next we can set the filenames and output frequency.

> Reveal the options for the **tracelog** using the ▶ to the left.

The frequency parameters are sampled and logged to file can be altered in the box labeled **Log Every**. In general, this value should be set relative to the length of the chain to avoid generating excessively large output files. If a low value is specified, the output files containing the parameter values and trees will be very large, possibly without gaining much additional information. Conversely, if you specify an exceedingly large sample interval, then you will not get enough information about the posterior distributions of your parameters.

> Change **Log Every** to `100`. Also specify a name for the log file by changing **File Name** to `bearsDivtime_FBD.log`.

Next, we will specify how the trees are written to file. Analyses in BEAST often estimate parameters that are associated with each branch in the tree (i.e, substitution rate). Additionally, the tree topology and branching times for each MCMC iteration can be written to file. These parameters are stored in a different file with the trees written in extended Newick format. The default name for this file is `$(tree).trees`.

The string "`$(tree)`" indicates the name of the tree we defined in the ***Partitions*** panel. For this analysis we relabeled this variable: `bearsTree` and our file name will begin with `bearsTree` even if we do not alter the name of this file.

Reveal the options for the ***treelog.t:bearsTree*** file. Keep the ***File Name*** `$(tree).trees` and ***Log Every*** to `100`.

An important part of any MCMC analysis is that multiple, independent runs are executed starting from different initial states for the various parameters. To do this, one can create multiple files in BEAUti, ensuring that the log and trees files have different names; or you can simply copy the XML file and alter the file names and starting states in a text editor. Given the time available for this practical, it isn't feasible to run multiple chains, but the output will be provided for you to evaluate this.

Now we are ready to save the XML file!

In the pull-down menu save the file by going to ***File→Save As*** and save the file: `bearsDivtime_FBD.xml`.

For the last step in BEAUti, create an XML file that will run the analysis by sampling under the prior. This means that the MCMC will ignore the information coming from the sequence data and only sample parameters in proportion to their prior probability. The output files produced from this run will provide a way to visualize the marginal prior distributions on each parameter.

Check the box labeled ***Sample From Prior*** at the bottom of the ***MCMC*** panel. We will want to change the names of the output files as well, so change the ***tracelog* – *File Name*** to `bearsDivtime_FBD.prior.log` and the ***treelog.t:bearsTree* – *File Name*** to `bearsDivtime_FBD.prior.trees`.

Save these changes by going to ***File→Save As*** and name the file `bearsDivtime_FBD.prior.xml`.

4.3 Making changes in the XML file

BEAUti is a great tool for generating a properly-formatted XML file for many types of BEAST analyses. However, you may encounter errors that require modifying elements in your input file and if you wish to make small to moderate changes to your analysis, altering the input file is far less tedious than generating a new one using BEAUti. Furthermore, BEAST is a rich program and all of the types of analyses, models, and parameters available in the core cannot be specified using BEAUti. Thus, some understanding of the BEAST XML format is essential.

Open the `bearsDivtime_FBD.xml` file generated by BEAUti in your text editor and glance over the contents. BEAUti provides many comments describing each of the elements in the file.

As you look over the contents of this file, you will notice that the components are specified in an order similar to the steps you took in BEAUti. The XML syntax is very verbose. This feature makes it fairly easy to understand the different elements of the BEAST input file. If you wished to alter your analysis or realized that you misspecified a prior parameter, changing the XML file is far simpler than going through all of the steps in BEAUti again. For example, if you wanted to change the mean of the exponential prior distribution on the mean clock rate (**ucldMean.c**), this can be done easily by altering this value in the XML file (Box 4), though leave this at **10.0** for this exercise.

```
<prior id="MeanRatePrior.c:bearsClock" name="distribution" x="@ucldMean.c:bearsClock">
    <Exponential id="Exponential.02" name="distr">
        <parameter id="RealParameter.020" estimate="false" name="mean">10.0</parameter>
    </Exponential>
</prior>
```

Box 1: BEAST 2 XML syntax for specifying an exponential prior distribution on the mean clock rate. Changing the expected value of this prior is simply done by altering the XML file.

Although running multiple, independent analyses is an important part of any Bayesian analysis, BEAST does not do this by default. However, setting this up is trivial once you have a complete XML file in hand and only requires that you make a copy of the input file and alter the names of the output files in the XML (it's also best to change the initial states for all of your parameters, including the starting tree).

If you need to return to your analysis specification in BEAUti, you can load your saved XML file when you reopen the program using the ***File→Load*** menu options.

4.4   Running BEAST 2

Now you are ready to start your BEAST analysis.

Execute **bearsDivtime_FBD.prior.xml** and **bearsDivtime_FBD.xml** in BEAST. You should see the screen output every 1,000 generations, reporting the likelihood and several other statistics.

4.5   Summarizing the output

Once the run reaches the end of the chain, you will find three new files in your analysis directory. The MCMC samples of various scalar parameters and statistics are written to the file called **bearsDivtime_FBD.log**. The tree-state at every sampled iteration is saved to **bearsTree.1.trees**. The tree strings in this file are all annotated in extended Newick format with the substitution rate from the uncorrelated lognormal model at each node. The file called **bearsDivtime_FBD.xml.states** summarizes the performance of the proposal mechanisms (operators) used in your analysis, providing information about the acceptance rate for each move. Reviewing this file can help identify operators that might need adjustment if their acceptance probabilities are too high.

The main output files are the `.log` file and `.trees` file. It is not feasible to review the data contained in these files by simply opening them in a spreadsheet program or a tree viewing program. Fortunately, the developers of BEAST have also written general utility programs for summarizing and visualizing posterior samples from Bayesian inference using MCMC. Tracer is a cross-platform, java program for summarizing posterior samples of scaler parameters. This program is necessary for assessing convergence, mixing, and determining an adequate burn-in. Tree topologies, branch rates, and node heights are summarized using the program TreeAnnotator and visualized in FigTree.

4.5.1 Tracer

This section will briefly cover using Tracer and visual inspection of the analysis output for MCMC convergence diagnostics.

> Open Tracer and import the `bearsDivtime_FBD.log` file in the *File→Import New Trace File*.

The fist statistic loaded will be the ***posterior***, in the ***Estimates*** tab, and you can see the list of statistics and variables that you can navigate through and visualize the summaries. You may notice that items in the ***ESS*** column are colored red or gold. The MCMC runs you have performed today are far too short to produce adequate posterior estimates of divergence times and substitution model parameters and this is reflected in the ESS values. The ESS is the *effective sample size* of a parameter. The value indicates the number of effectively independent draws from the posterior in the sample. This statistic can help to identify autocorrelation in your samples that might result from poor mixing. It is important that you run your chains long enough and sufficiently sample the stationary distribution so that the ESS values of your parameters are all high (over 200 or so).

> Click on a parameter with a low ESS and explore the various windows in Tracer. It is clear that we must run the MCMC chain longer to get good estimates.

Provided with the files for this exercise are the output files from analyses run for 50,000,000 iterations. The files can be found in the `output` directory and are all labeled with the file stem: `bearsDivtime_FBD.*.log` and `bearsDivtime_FBD.*.trees`.

> Close `bearsDivtime_FBD.log` in Tracer using the ⊟ button and open `bearsDivtime_FBD.1.log`, `bearsDivtime_FBD.2.log` and `bearsDivtime_FBD.prior.log`.

These log files are from much longer runs and since we ran two independent, identical analyses, we can compare the log files in Tracer and determine if they have converged on the same stationary distribution. Additionally, analyzing samples from the prior allows you to compare your posterior estimates to the prior distributions used for each parameter.

> Select and highlight all three files (**bearsDivtime_FBD.1.log**, **bearsDivtime_FBD.2.log** and **bearsDivtime_FBD.prior.log**) in the ***Trace Files*** pane (do not include ***Combined***). This allows you to compare all three runs simultaneously. Click on the various parameters and view how they differ in their estimates and 95% credible intervals for those parameters

The 95% credible interval is a Bayesian measure of uncertainty that accounts for the data. If we use the 95% credible interval, this means that the probability the true value of a parameter lies within this interval is 0.95, given our model and data. This measure is often used to approximate the 95% highest posterior density region (HPD).

> Find the parameter ***ucldStdev*** and compare the estimates of the standard deviation of the uncorrelated log-normal distribution.

The ***ucldStdev*** indicates the amount of variation in the substitution rates across branches. Our prior on this parameter is an exponential distribution with $\nu = 2.997$ ($mean = 0.3337$). Thus, there is a considerable amount of prior weight on ***ucldStdev*** $= 0$. A standard deviation of 0 indicates support for no variation in substitution rates and the presence of a molecular clock.

> With ***ucldStdev*** highlighted for all three runs, go to the ***Marginal Prob Distribution*** window, which allows you to compare the marginal posterior densities for each parameter. (By default Tracer gives the kernel density estimate (***KDE***) of the marginal density. You can change this to a ***Histogram*** using the options at the bottom of the window.)
>
> Color (or "colour") the densities by ***Trace File*** next to ***Colour by*** at the bottom of the window (if you do not see this option, increase the size of your Tracer window). You can also add a ***Legend*** to reveal which density belongs to which run file. [Figure 26]

The first thing you will notice from this plot is that the marginal densities from each of our analysis runs (**bearsDivtime_FBD.1.log** and **bearsDivtime_FBD.2.log**) are nearly identical. If you click through the other sampled parameters, these densities are the same for each one. This indicates that both of our runs have converged on the same stationary distribution. If some of the other parameters might not have mixed well, we may want to run them longer, but we can have good confidence that our runs have sampled the same distribution.

Second, it is likely for the UCLD rates that the parameters of this model appear to be somewhat sensitive to the priors. That is because we have several lineages in our tree without any data since we are using fossils to calibrate under the FBD model. For these branches, the clock rate assigned to them is sampled from the prior. The branches with extant descendants, however have rates that are informed by DNA sequence data. Because of this implementation, it is difficult to visualize the branch rates and associated parameters from the log files. If these parameters are of interest, then it may be necessary to extract the rates associated with the extant branches from the MCMC samples of trees. When fossils are used without character data, alternative implementations of the FBD model consider the fossils separately and thus, the branch rates can be summarized for only the extant lineages (Heath et al. 2014).
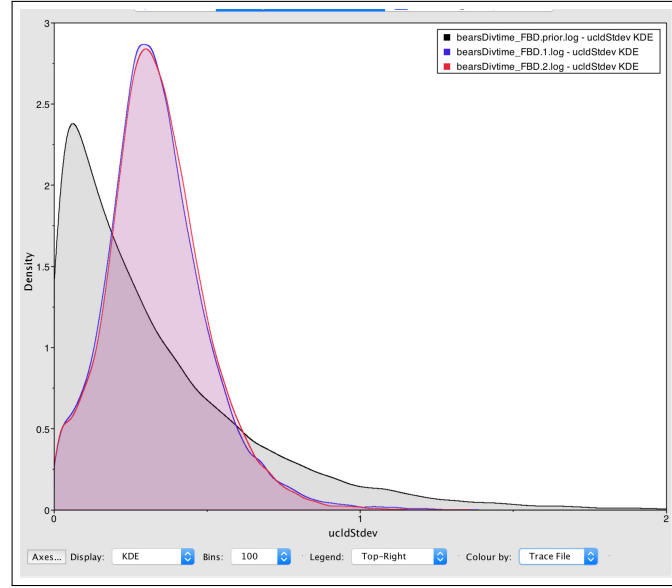
Figure 26: Comparing the marginal densities (using the kernel density estimate) of the **ucldStdev** parameter from 2 independent runs (red and blue) and the prior (gray) in Tracer.

Next, we will look at the **Marginal Prob Distribution** for the turnover parameter (**turnoverFBD**).

Select all trace files for the **turnoverFBD** parameter and go to the **Marginal Prob Distribution** window.

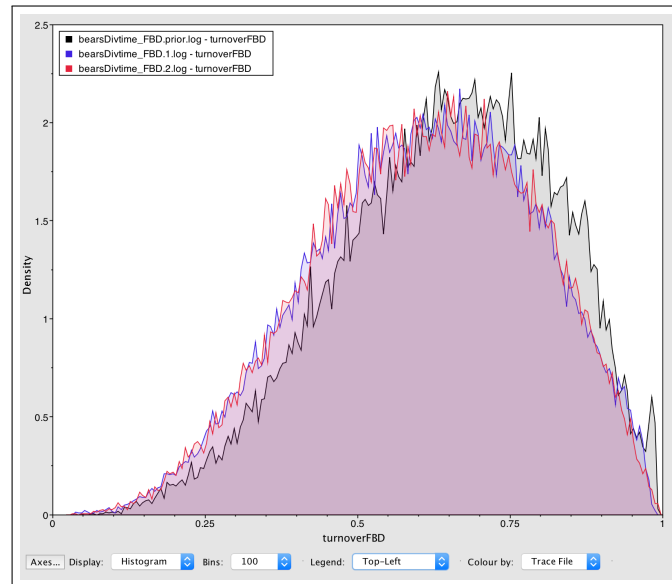Color the densities by **TraceFile** and add a legend. [Figure 27]

Figure 27: Comparing the marginal densities (using the histogram) of the **turnoverFBD** parameter from 2 independent runs (red and blue) and the prior (gray) in Tracer.

When comparing the MCMC with data to those sampled just under the prior, we can see that this prior and posterior densities are nearly identical. Additionally, you may be alarmed that the prior density looks nothing like the Uniform(0,1) prior we applied in BEAUti. However, it's important to understand that when specifying **Sample From Prior** in BEAUti we are telling BEAST to ignore the likelihood coming from the *sequence data*, but the MCMC continues to account for the other data—fossil occurrence times— that we have input. Thus, since we have a good number of fossils, the estimates of turnover, diversification, and the sampling proportion are highly informed by these data. If we used fewer fossils, the posterior and prior densities would be less congruent because the fossil data are still considered when "sampling from the prior" in this case. By contrast, if the information from the fossil occurrence times was ignored when sampling under the prior, then the posterior and prior densities for the FBD parameters would deviate more when many fossils are used.

Continue examining the options in Tracer. This program is very, *very* useful for exploring many aspects of your analysis.

4.5.2 Summarizing the Trees in Treeannotator

After reviewing the trace files from the two independent runs in Tracer and verifying that both runs converged on the posterior distributions and reached stationarity, we can combine the sampled trees into a single tree file and summarize the results.

Open the program LogCombiner and set the **File type** to **Tree Files**. Next, import the two tree files in the **output** directory (**bearsDivtime_FBD.1.trees** and **bearsDivtime_FBD.2.trees**) using the ⊞ button.

Set a burn-in percentage of 20 for each file, thus discarding the first 20% of the samples in each tree file.

Both of these files have 50,000 trees, so it is helpful to thin the tree samples and summarize fewer states (to avoid hitting the maximum memory allotted for this program). Turn on **Resample states at lower frequency** and set this value to 2000.

Click on the **Choose file ...** button to create an output file and run the program. Name the file: **bearsDivtime.combined.trees**.

Once LogCombiner has terminated, you will have a file containing 40,000 trees which can be summarized using TreeAnnotator. TreeAnnotator takes a collection of trees and summarizes them by identifying the topology with the best support, calculating clade posterior probabilities, and calculating 95% credible intervals for node-specific parameters. All of the node statistics are annotated on the tree topology for each node in the Newick string.

Open the program TreeAnnotator. Since we already discarded a set of burn-in trees when combining the tree files, we can leave **Burnin** set to **0** (though, if TreeAnnotator is taking a long time to load the trees, set the burnin to 10–60% to reduce the number of trees).

For the ***Target tree type***, choose ***Maximum clade credibility tree***.

The ***Maximum clade credibility tree*** is the topology with the highest product of clade posterior probabilities across all nodes. Alternatively, you can select the ***Maximum sum of clade credibilities*** which sums all of the clade posteriors. Or you can provide a target tree from file. The ***Posterior probability limit*** option applies to summaries on a target tree topology and only calculates posteriors for nodes that are above the specified limit.

Choose ***Median heights*** or ***Mean heights*** for ***Node heights*** which will set the node heights of the output tree to equal the median or mean height for each node in the sample of trees.

Choose `bearsDivtime.combined.trees` as your ***Input Tree File***. Then name the ***Output File:*** `bearsDivtime_FBD.summary.tre` and click ***Run***.

4.5.3   Visualizing the Dated Tree

The tree file produced by TreeAnnotator contains the maximum clade credibility tree and is annotated with summaries of the various parameters.

Open `bearsDivtime_FBD.summary.tre` in your text editor. The tree is written in NEXUS format. Look at the tree string and notice the annotation. Each node in the tree is labeled with comments using the `[&parameter_name=<value>]` format.

The summary tree and it's annotations can be visualized in the program FigTree.

Execute FigTree and open the file `bearsDivtime_FBD.summary.tre`.

Explore the options for viewing different summary statistics on the tree.

The tree you are viewing in FigTree has several fossil taxa with zero-length branches, e.g., *Parictis montanus* and *Ursus abstrusus*. These branches actually indicate fossil taxa with a significant probability of representing a sampled ancestor. However, it is difficult to represent this in typical tree-viewing programs. A tree with sampled ancestors properly represented will have two-degree nodes — i.e., a node with only one descendant. The web-based tree viewer IcyTree (Vaughan 2017) is capable of plotting such trees from BEAST2 analyses (Figure 28).

Although the tree in Figure 28 (and also Figure 30) looks awesome, the topology can be misleading for this particular analysis. Because we did not provide character data for our 14 fossil taxa, the tree topology does not adequately illustrate the degree of uncertainty in the phylogenetic relationships of the fossil lineages. **Without morphological data, the fossil lineages can attach to any lineage on the tree that is consistent with the monophyletic clades we specified and the fossil occurrence times**
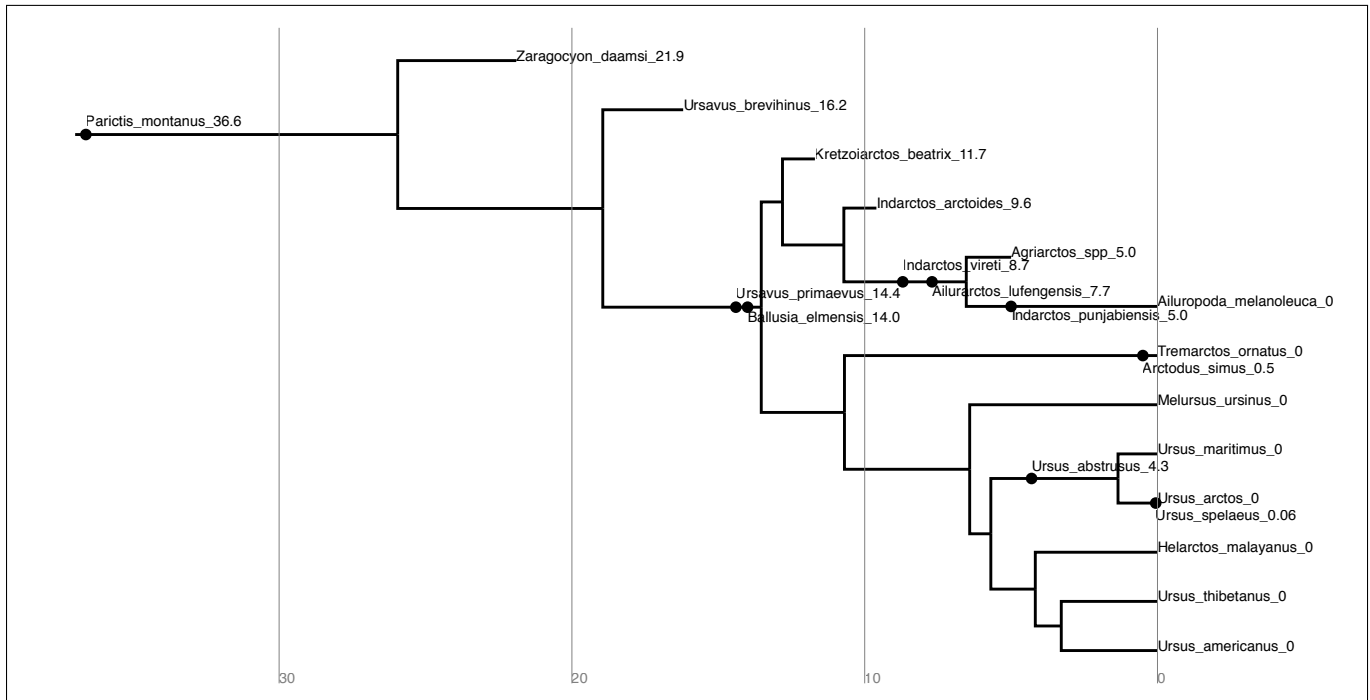
Figure 28: A sampled-ancestor tree exported from IcyTree. The sampled-ancestor nodes are revealed by selecting the ***Mark singletons*** and ***Internal node text*** set to ***Label***. The tree was exported to a SVG file and minor manipulations were conducted in a vector graphics editing program (Inkscape or Adobe Illustrator).

with equal probability. Thus, the relationships shown here are not reliable as phylogenetic inferences. However, if we provided morphological character data for these fossils, then many would consider this tree to be an adequate summary our MCMC sample of trees.

Since the fossil taxa were used in this analysis to only inform the fossilized birth-death model, we can prune off all of the fossil lineages and plot the tree with the geological time scale. BEAUti has a few accessory applications that are for post-processing files from certain types of analyses. The first one we will use is called ***FullToExtantTreeConverter***. This program can be used to prune the fossil lineages off of the MCMC sample of trees (in file `bearsDivtime_FBD.combined.trees`), then use TreeAnnotator to summarize the extant-only trees.

> Open BEAUti and launch the accessory apps in the file menu: ***File→Launch Apps***.
>
> This will open a window with a few applications, launch ***FullToExtantTreeConverter***.
>
> In the file specification window for ***Trees*** provide the file called `bearsDivtime_FBD.combined.trees`.
>
> And give the ***Output*** file the name `bearsDivtime_FBD.extant.trees`.
>
> Run TreeAnnotator on the file called `bearsDivtime_FBD.extant.trees`, selecting the same options as you did for the file with complete trees (see Section 4.5.2).
>
> Name the summary tree file `bearsDivtime_FBD.extant_summary.tre`

When the fossil lineages are removed, much of the information about the history of this group is lost (Figure 29). However, alternative approaches for summarizing fossilized birth-death trees are currently under development.
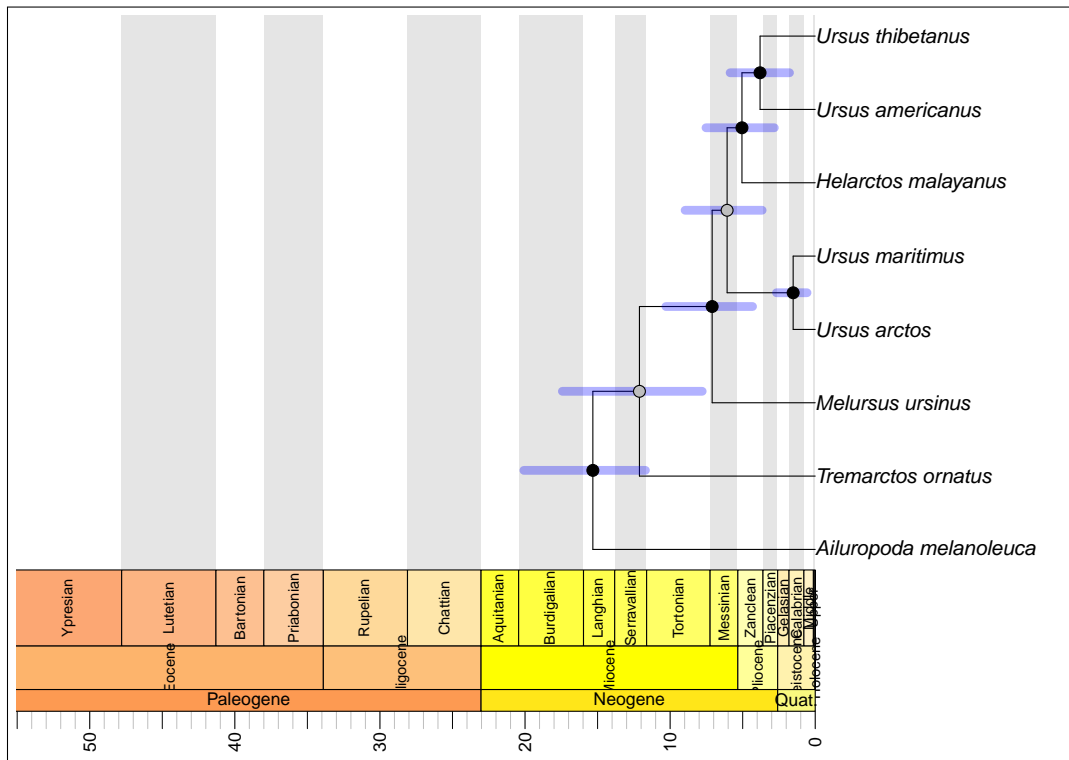

Figure 29: The maximum clade credibility tree of *extant* bears summarized by TreeAnnotator and plotted against stratigraphy using the **strap** package in **R** (see how to do this in Section 4.5.5 below). The internal nodes of the tree are indicated with circles, where circles mark nodes with posterior probability: ● ≥ 0.95, 0.95 > ◐ ≥ 0.75, 0.75 > ○. The 95% credible intervals for node ages are shown with transparent blue bars.

4.5.4   Analysis of Sampled Ancestors

Another tool that is available in the BEAUti accessory applications, is ***SampledAncestorTreeAnalyser***. If you run this app from the ***Launch apps*** menu in BEAUti on the file with all the MCMC samples of the complete tree (**bearsDivtime_FBD.combined.trees**), you will get a report that summarizes the clades/nodes in the sampled ancestor tree. In particular, it gives a quantitative representation of "ancestral-ness" in the posterior sample of trees. This is represented by listing how often a particular node is ancestral to other nodes. The summary is generated and automatically opens in your web browser.

4.5.5   Visualizing the Dated Tree on a Geological Time Scale*

*Note that some of the **R** packages required for this section may be difficult to download.

FigTree and IcyTree are great tree-viewing programs and also allow you to produce publication-quality tree figures. However, viewing a dated phylogenetic tree with a unit-less timescale is not as meaningful as plotting the tree with a geological (stratigraphic) time scale. The **R** package **strap** (Bell and Lloyd 2014) offers several nice functions for visualizing time-calibrated phylogenies in the context of the rock record.

Install R Packages for Viewing and Plotting Trees

For this exercise we will use some `R` packages to visualize the summary tree with a geological timescale. If you do not already have `R` installed, please download the current version: http://www.r-project.org

strap – Viewing dated phylogenies with an arbitrary time-scale removes the context of geological time and the fossil record from the analysis. The package **strap** in `R` provides a set of functions to plot trees and stratigraphic information against geologic time, with scales provided by different sources including the International Commission on Stratigraphy (Bell and Lloyd 2014). A detailed tutorial for using the functions in **strap** is available here: http://datadryad.org/resource/doi:10.5061/dryad.4k078. To install **strap**, execute the following command in `R`:

```
> install.packages("strap",dependencies=TRUE)
```

phytools – In `R`, there are many packages available for performing phylogenetic comparative methods, among them **phytools** is one of the richest, providing functions for a wide range of different analyses and for visualizing evolutionary processes in the context of phylogenetic relationships (Revell 2012). To install **phytools** in `R`, execute the following command:

```
> install.packages("phytools",dependencies=TRUE)
```

phyloch – The trees sampled by the MCMC in BEAST contain valuable information about the sampled divergence times and branch rates. Additionally, the summary trees produced by the BEAST accessory program TreeAnnotator have information about the 95% credible intervals for the ages and rates. The package **phyloch** provides functions for reading in data files written by BEAST and its accessory programs (Heibl 2008). This package, however, is not available for download from CRAN. Instead, it is hosted on the developer's website. To install the **phyloch** package in `R`, first load the **devtools** package and install **phyloch** directly from the URL:

```
> library("devtools")
> install_url("http://www.christophheibl.de/phyloch_1.5-5.tgz")
```

Plot the Tree in `R`

Begin by opening an `R` instance and load the **strap** package.

```
> library("strap")
```

Now we can read in the tree using the **ape** function `read.nexus()` (note that you might have to type in the whole file path to your tree).

```
> tree <- read.nexus("bearsDivtime.summary.tre")
```

In order to use the `geoscalePhylo()` function of **strap**, we have to set a value for the variable `tree$root.time`, which is the age of the root. We can compute this from the tree using the

dist.nodes() function from **ape**:

```
> tree$root.time <- max(dist.nodes(tree))
```

Now plot the tree:

```
> geoscalePhylo(tree=ladderize(tree,right=FALSE),label.offset=0)
```

You will notice that the plotted figure might need some work to make it easier to read taxon labels, etc. Additionally, we don't get the node bars or other summary statistics for our tree. If we wish to plot these values, then we need to do a bit more in **R**. With additional packages and functions, we can produce a summary tree that includes the credible intervals on the node ages (only for nodes in the extant tree, the origin time, and symbols representing the posterior probabilities of the tree bipartitions). This tree is shown in Figure 30 and the **R** syntax for producing this figure is provided in the **output** directory in the download files (**plot_geoscaled_tree.R**).

5   Useful Links

- Taming the BEAST – tutorials and workshops for learning BEAST 2: https://taming-the-beast.github.io
- BEAST 2 website and documentation: http://www.beast2.org
- BEAST 1 website and documentation: http://beast.bio.ed.ac.uk
- Join the BEAST user discussion: http://groups.google.com/group/beast-users
- RevBayes: https://github.com/revbayes/code
- DPPDiv: https://github.com/trayc7/FDPPDIV
- PhyloBayes: www.phylobayes.org/
- multidivtime: http://statgen.ncsu.edu/thorne/multidivtime.html
- MCMCtree (PAML): http://abacus.gene.ucl.ac.uk/software/paml.html
- BEAGLE: http://code.google.com/p/beagle-lib/
- A list of programs: http://evolution.genetics.washington.edu/phylip/software.html
- The Paleobiology Database: http://www.paleodb.org
- The Fossil Calibration Database: http://fossilcalibrations.org

Version dated: February 8, 2018

Relevant References

Abella, J, P Montoya, and J Morales. 2011. Una nueva especie de *Agriarctos* (Ailuropodinae, Ursidae, Carnivora) en la localidad de Nombrevilla 2 (Zaragoza, España). *Estudios Geológicos* 67: 187–191.
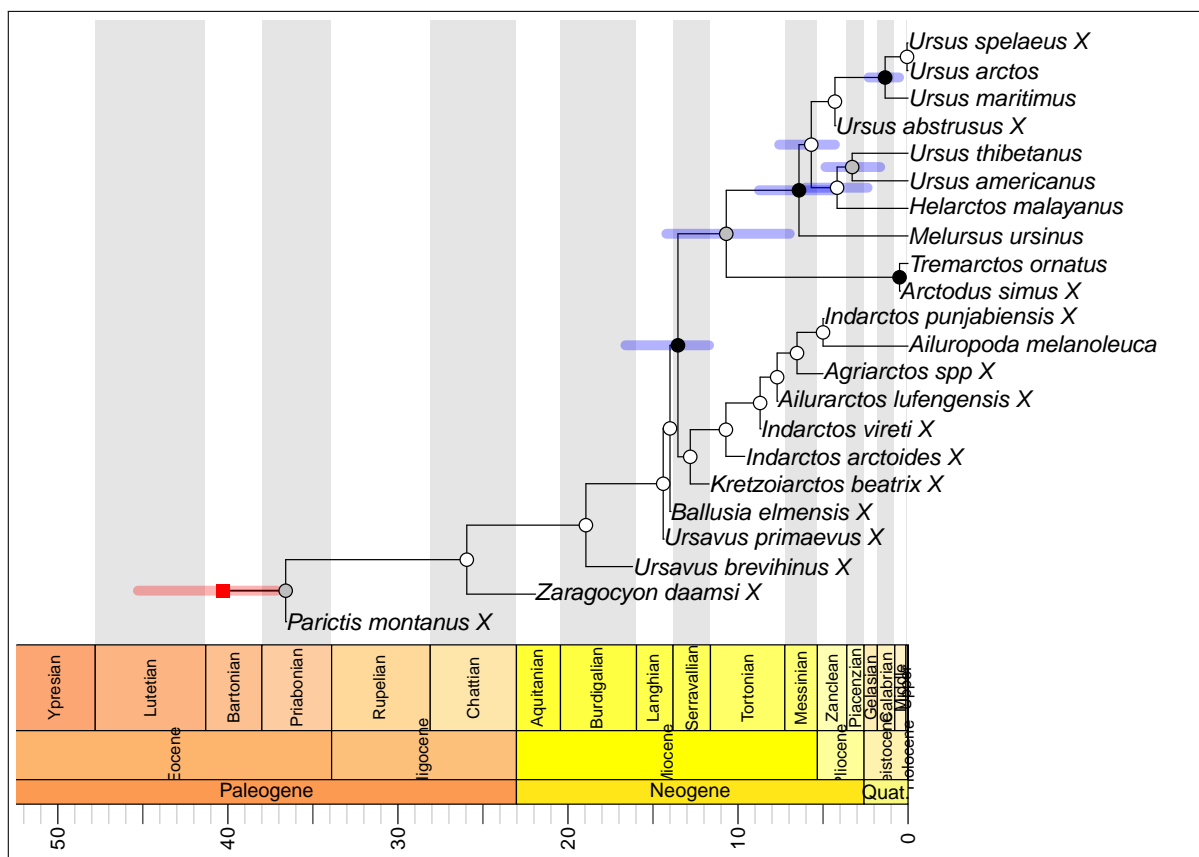
Figure 30: The maximum clade credibility tree summarized by TreeAnnotator and plotted with a geological time scale using the **strap** package in **R**. The fossil taxa are all indicated with an *X* in the taxon names. The ■ represents the mean origin time. The remaining internal nodes of the tree are indicated with circles, where circles mark nodes with posterior probability: ● ≥ 0.95, 0.95 > ◉ ≥ 0.75, 0.75 > ○. The 95% credible intervals for node ages are shown with transparent bars, for only nodes that are represented in the *extant* tree. (The **R** code to produce this figure is in the **output** folder in the file **plot_geoscaled_tree.R**.)

Abella, J, DM Alba, JM Robles, A Valenciano, C Rotgers, R Carmona, P Montoya, and J Morales. 2012. *Kretzoiarctos* gen. nov., the oldest member of the giant panda clade. *PLoS One* 17: e48985.

Aldous, D. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* 16: 23–34.

Aldous, D and L Popovic. 2005. A critical branching process model for biodiversity. *Advances in Applied Probability* 37: 1094–1115.

Andrews, P and H Tobien. 1977. New Miocene locality in turkey with evidence on the origin of *Ramapithecus* and *Sivapithecus. Nature* 268: 699.

Baele, G, WLS Li, AJ Drummond, MA Suchard, and P Lemey. 2013. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* 30: 239–243.

Baryshnikov, GF. 2002. Late Miocene *Indarctos punjabiensis atticus* (Carnivora, Ursidae) in Ukraine with survey of *Indarctos* records from the former USSR. *Russian J. Theriol* 1: 83–89.

Bell, MA and GT Lloyd. 2014. Strap: an R package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence. *Palaeontology*

Bjork, PR. 1970. The Carnivora of the Hagerman local fauna (late Pliocene) of Southwestern idaho. *Transactions of the American Philosophical Society* 60: 3–54.

Bouckaert, R, J Heled, D Kühnert, T Vaughan, C-H Wu, D Xie, MA Suchard, A Rambaut, and AJ Drummond. 2014. Beast 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology* 10: e1003537.

Churcher, C, A Morgan, and L Carter. 1993. *Arctodus simus* from the Alaskan Arctic slope. *Canadian Journal of Earth Sciences* 30: 1007–1013.

Clark, J and TE Guensburg. 1972. *Arctoid Genetic Characters as Related to the Genus* Parictis. Vol. 1150. Field Museum of Natural History, Chicago, Ill.

dos Reis, M, J Inoue, M Hasegawa, R Asher, P Donoghue, and Z Yang. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences* 279: 3491–3500.

Drummond, AJ and A Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 214.

Drummond, AJ and MA Suchard. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology* 8: 114.

Drummond, AJ, SY Ho, MJ Phillips, and A Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4: e88.

Drummond, AJ and RR Bouckaert. 2014. *Bayesian evolutionary analysis with BEAST 2*. Cambridge University Press,

Foote, M. 1996. On the probability of ancestors in the fossil record. *Paleobiology* 22: 141–151.

Gavryushkina, A, D Welch, T Stadler, and AJ Drummond. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology* 10: e1003919.

Geraads, D, T Kaya, S Mayda, et al. 2005. Late miocene large mammals from Yulafli, Thrace region, Turkey, and their biogeographic implications. *Acta Palaeontologica Polonica* 50: 523–544.

Gernhard, T. 2008. The conditioned reconstructed process. *Journal of Theoretical Biology* 253: 769–778.

Ginsburg, L and J Morales. 1995. *Zaragocyon daamsi n. gen. sp. nov.*, ursidae primitif du Miocène inférieur d'Espagne. *Comptes Rendus de l'Académie des Sciences. Série 2. Sciences de la Terre et des Planètes* 321: 811–815.

Ginsburg, L and J Morales. 1998. Les Hemicyoninae (Ursidae, Carnivora, Mammalia) et les formes apparentées du Miocène inférieur et moyen d'Europe occidentale. *Annales de Paléontologie.* 84: pp. 71–123.

Griffiths, R and S Tavare. 1994. Simulating probability distributions in the coalescent. *Theoretical Population Biology* 46: 131–159.

Heath, TA. 2012. A hierarchical Bayesian model for calibrating estimates of species divergence times. *Systematic Biology* 61: 793–809.

Heath, TA, MT Holder, and JP Huelsenbeck. 2012. A Dirichlet process prior for estimating lineage-specific substitution rates. *Molecular Biology and Evolution* 29: 939–255.

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

Heibl, C. 2008. Phyloch: R language tree plotting tools and interfaces to diverse phylogenetic software packages, http://www.christophheibl.de/Rpackages.html.

Heizmann, E, L Ginsburg, and C Bulot. 1980. *Prosansanosmilus peregrinus*, ein neuer machairodontider Felidae aus dem Miozän Deutschlands und Frankreichs. *Stuttgarter Beitr. Naturk. B* 58: 1–27.

Heled, J and AJ Drummond. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27: 570–580.

— 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology* 61: 138–149.

Heled, J and AJ Drummond. 2013. Calibrated birth-death phylogenetic time-tree priors for Bayesian inference. *arXiv preprint arXiv:1311.4921*

Ho, SYW and MJ Phillips. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology* 58: 367–380.

Huelsenbeck, JP, B Larget, and DL Swofford. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154: 1879–1892.

Jin, C, RL Ciochon, W Dong, RM Hunt, J Liu, M Jaeger, and Q Zhu. 2007. The first skull of the earliest giant panda. *Proceedings of the National Academy of Sciences* 104: 10932–10937.

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

Kingman, JFC. 1982. Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics.* (Ed.) Koch, G and Spizzichino, F. North-Holland, pp. 97–112.

— 1982. On the genealogy of large populations. *Essays in Statistical Science: Papers in Honour of P. A. P. Moran, Journal of Applied Probability, Special Volume 19A.* (Ed.) Gani, J and Hannan, EJ. Applied Probability Trust, pp. 27–43.

— 1982. The coalescent. *Stochastic Processes and their Applications* 13: 235–248.

Kishino, H, T Miyata, and M Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution* 31: 151–160.

Kishino, H, JL Thorne, and W Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* 18: 352–361.

Krause, J, T Unger, A Noçon, A-S Malaspinas, S-O Kolokotronis, M Stiller, L Soibelzon, H Spriggs, PH Dear, AW Briggs, et al. 2008. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evolutionary Biology* 8: 220.

Landis, MJ, NJ Matzke, BR Moore, and JP Huelsenbeck. 2013. Bayesian analysis of biogeography when the number of areas is large. *Systematic biology* 789–804.

Lemey, P, A Rambaut, AJ Drummond, and MA Suchard. 2009. Bayesian phylogeography finds its roots. *PLoS Computational Biology* 5: e1000520.

Lepage, T, D Bryant, H Philippe, and N Lartillot. 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* 24: 2669–2680.

Lepage, T, S Lawi, P Tupper, and D Bryant. 2006. Continuous and tractable models for the variation of evolutionary rates. *Mathematical Biosciences* 199: 216–233.

Li, WLS and AJ Drummond. 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* 29: 751–761.

Loreille, O, L Orlando, M Patou-Mathis, M Philippe, P Taberlet, and C Hänni. 2001. Ancient DNA analysis reveals divergence of the cave bear, *Ursus spelaeus*, and brown bear, *Ursus arctos*, lineages. *Current Biology* 11: 200–203.

Montoya, P, L Alcalá, and J Morales. 2001. *Indarctos* (Ursidae, Mammalia) from the Spanish Turolian (Upper Miocene). *Scripta Geologica* 122: 123–151.

Nee, S, RM May, and PH Harvey. 1994. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B* 344: 305–311.

Popovic, L. 2004. Asymptotic genealogy of a critical branching process. *Annals of Applied Probability* 14: 2120–2148.

Pyron, RA. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Systematic Biology* 60: 466–481.

Rambaut, A and L Bromham. 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution* 15: 442–448.

Rannala, B and Z Yang. 2007. Inferring speciation times under an episodic molecular clock. *Systematic Biology* 56: 453–466.

— 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* 43: 304–311.

Revell, LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3: 217–223.

Ronquist, F, S Klopfstein, L Vilhelmsen, S Schulmeister, DL Murray, and AP Rasnitsyn. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology* 61: 973–999.

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

— 2010. Sampling-through-time in birth-death trees. *Journal of Theoretical Biology* 267: 396–404.

Thompson, EA. 1975. *Human Evolutionary Trees.* Cambridge University Press, Cambridge, England.

Thorne, J and H Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* 51: 689–702.

Thorne, J, H Kishino, and IS Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15: 1647–1657.

Vaughan, TG. 2017. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* in press.

Warnock, RCM, Z Yang, and PCJ Donoghue. 2012. Exploring the uncertainty in the calibration of the molecular clock. *Biology Letters* 8: 156–159.

Warnock, RC, JF Parham, WG Joyce, TR Lyson, and PC Donoghue. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. *Proceedings of the Royal Society B: Biological Sciences* 282: 20141013.

Yang, Z and B Rannala. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution* 23: 212–226.

— 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution* 14: 717–724.

Yang, Z and AD Yoder. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology* 52: 705–716.

Yule, GU. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Wills, F. R. S. *Philosophical Transactions of the Royal Society of London, Biology* 213: 21–87.

Zuckerkandl, E and L Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. *Horizons in Biochemistry.* (Ed.) Kasha, M and Pullman, B. Academic Press, New York, pp. 189–225.