Tutorial using BEAST v2.7.7

contraband tutorial

Rong Zhang and Fábio K. Mendes

Total-evidence dating and trait-evolution evolutionary inference using phylogenetic multivariate Brownian motion models

1 Background

Bird's-eye view. This tutorial shows how to use the `contraband` package in `BEAST 2` to model continuous trait evolution along a phylogeny with Brownian motion. Unlike methods that assume a "known", fixed tree, `contraband` lets you estimate the tempo and mode of trait evolution simultaneously with both species relationships and divergence times.

1.1 What is contraband for

In this tutorial, we will walk you through running a simple analysis with the `contraband` (**con**tinuous **tra**its **b**rowni**an** mo**d**els) `BEAST 2` package. As the name suggests, `contraband` implements Brownian motion (BM) models for the evolution of continuous traits on a phylogeny.

To understand how these models can be useful to evolutionary biologists, let's put our X-ray goggles on and look at the core of the `contraband` package: the probability density function (pdf) of the multivariate Brownian motion model – the same pdf used for a multivariate normal distribution:

$$f(\mathbf{M}|\boldsymbol{V}, \boldsymbol{y_0}) = \frac{1}{(2\pi)^{nk/2}|\boldsymbol{V}|^{1/2}} \exp\left(-\frac{1}{2}(\text{vec}(\mathbf{M}) - \boldsymbol{y_0})^{\text{T}}\boldsymbol{V}^{-1}(\text{vec}(\mathbf{M}) - \boldsymbol{y_0})\right), \quad (1)$$

This equation simply gives us the probability of observing our data $\mathbf{M}$ – that is, one or more continuous traits – given two key parameters: (i) the expected value vector (or mean vector), $\boldsymbol{y_0}$, and (ii) the variance-covariance matrix, $\boldsymbol{V}$. If you have tried a few of the other Taming the BEAST tutorials, these two parameters are the quantities whose posterior probability distributions we want to approximate via Markov Chain Monte Carlo (MCMC).

In phylogenetics, $\boldsymbol{V}$ is typically decomposed as $\boldsymbol{V} = \boldsymbol{\Sigma} \otimes \boldsymbol{T}$, where $\boldsymbol{\Sigma}$ describes the variance and covariance structure of the traits, and $\boldsymbol{T}$ represents phylogenetic relatedness. In essence, $\boldsymbol{T}$ captures the phylogeny itself – the shared evolutionary history among species.

In many software tools, especially those implemented in `R` and using frequentist methods, the phylogeny ($\boldsymbol{T}$) is not estimated but instead fixed to a tree point estimate from the literature. The downside of this approach is that the continuous trait data can only inform our estimates of trait evolution parameters, $\boldsymbol{y_0}$ and $\boldsymbol{\Sigma}$ – not the phylogeny itself.

While it is possible to take this approach in `BEAST 2` as well, its hierarchical Bayesian framework allows us to go further: we can co-estimate $\boldsymbol{T}$ (i.e., the species tree or phylogeny) together with the parameters of trait evolution. This means we can infer trait-evolution parameters **alongside** the species divergence times and phylogenetic relationships captured in $\boldsymbol{T}$. In other words, `contraband` is a tool not only for studying how continuous traits evolve, but also for estimating the topology and divergence times of phylogenies.

The estimation of divergence times using multiple types of data – for example, molecular sequences combined with discrete and/or continuous morphological traits – is known as *total-evidence dating* (TED; **ronquist12**). Among other things, `contraband` is a TED method. It is designed to help evolutionary biologists leverage continuous traits to reconstruct species evolutionary histories, including both divergence times and the tempo and mode of phenotypic evolution.

1.2 A quick peek under the hood

Later in this tutorial, you will be placing prior distributions on a series of parameters, as well as making modeling decisions related to things like the correlation between traits, for example, or the intraspecific variance in trait values. Setting up such an analysis can quickly become overwhelming, so in this section we will introduce a few implementation and statistical details to help you understand what comes next.

While it is possible to directly compute the value of equation (1) via matrix algebra, this is computationally expensive. Instead, `contraband` saves us time by using an alternative mathematical formulation (**mitov20**) and a dynamic programming algorithm. The details do not matter for this tutorial, but it is important to re-write equation (1) as:

$$f(\mathbf{M}|\boldsymbol{V}, \boldsymbol{y_0}) = f(\mathbf{M}|\Phi, \boldsymbol{y_0}, \boldsymbol{r}, \boldsymbol{\rho}, c_m, \boldsymbol{b}_m, \boldsymbol{\theta}) \tag{2}$$

As mentioned above, $\mathbf{M}$ contains our continuous-character data, it is a matrix whose dimensions are the number of species $\times$ the number of characters. On the right-hand side of this equation, you should further recognize some of the terms that have direct counterparts in models used for molecular evolution, e.g., those involved in the morphological clock model. These are the morphological global clock rate (c_m) and relative branch rates ($\boldsymbol{b}_m$). Other parameters, however, are unique to multivariate Brownian models, like the character values from all characters at the root of Φ ($\boldsymbol{y_0}$), a vector containing all relative character-specific evolutionary rates ($\boldsymbol{r}$), and the between-character correlation matrix ($\boldsymbol{\rho}$). All of these parameters can in principle be estimated with MCMC, but the accuracy of and uncertainty about our estimates will be a function of our data set size, which include the number of traits and of species (more details can be found in **zhang24**), as well as analysis running times.

Among the most challenging parameters to estimate is $\boldsymbol{\rho}$. For example, attempting to estimate $\boldsymbol{\rho}$ with MCMC means we have a potentially very large number of parameters that will be very hard to identify unless one has a very large phylogeny (which in turn would make the analysis prohibitively slow). One thing we can do is obtain intraspecific character data from one of the species in the phylogeny, and then estimate character correlations from that. In short, one can obtain an estimate of $\boldsymbol{\rho}$, $\widehat{\boldsymbol{\rho}}$, from characters scored in many individuals of a single species, and then assume this estimate is true and constant across the phylogeny – i.e., there is no MCMC sampling of character correlation parameters.

Depending on the dimensions of $\widehat{\boldsymbol{\rho}}$, however, it can become unwieldly: it may become nearly singular, its determinant approaching 0, and its inverse blowing up. (Down the line, we obviously cannot compute $f(\mathbf{M}|\boldsymbol{V}, \boldsymbol{y_0})$!) Here, we can borrow a technique often referred to as "regularization": we can "shrink" $\widehat{\boldsymbol{\rho}}$ towards the identity matrix $\boldsymbol{I}$ – which represents the correlation matrix of a data set where characters are uncorrelated – and obtain what we call a ridge estimator of $\boldsymbol{\rho}$, $\boldsymbol{\rho}^*$. Doing so effectively adds some values to the off-diagonals of the correlation matrix, making it better conditioned; the extent to which we "shrink" $\widehat{\boldsymbol{\rho}}$ towards the identity matrix is captured in a tuning ("shrinkage") parameter, δ. (As you will see later in this tutorial, we will have to specify δ to run one of our inference analyses.) The more uncertain we are about character correlations, because we have way too many characters for way too few species, say, the

larger δ should be. At any rate, we will not have to worry too much about the details of how to obtain δ. There are methods for doing just that in the literature, and we will use them.

Overall, here is a list of the continuous-trait model parameters that we want to estimate, and for which we will need to place prior distributions on:

1) Character-specific evolutionary rates, $\boldsymbol{r}$,
2) Character correlations, $\boldsymbol{\rho}$,
3) Ancestral state values, $\boldsymbol{y_0}$,
4) Morphological clock parameters, relative branch rates $\boldsymbol{b}_m$ and global evolutionary rate c_m

In what follows, we will guide you through the explicit steps – including installation of dependencies and post-processing tools – that will (i) set up the analysis for inferring the above parameters, and (ii) help you process and visualize the results.

2 Programs used in this exercise

2.0.1 BEAST2 - Bayesian Evolutionary Analysis Sampling Trees2

BEAST2 (http://www.beast2.org) 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. This tutorial is written for BEAST v2.7.7 (**bouckaert2019beast**).

2.0.2 BEAUti2 - Bayesian Evolutionary Analysis Utility

BEAUti2 is the successor of BEAUti, a graphical user interface tool that makes it easy to generate BEAST2 XML configuration files (these files are necessary to specify and run MCMC analyses). It is provided as a part of the BEAST2 package so you do not need to install it separately. Both BEAST2 and BEAUti2 are written in Java, meaning that these programs can not only be integrated at their codebase level, but that they are also cross-platform: the exact same code runs on all platforms. Although the screenshots used in this tutorial have been taken on a Mac OS X computer, both BEAST2 and BEAUti2 will have the same layout and functionality under other operating systems like Windows and Linux.

2.0.3 TreeAnnotator

TreeAnnotator is a program we will use to produce a summary tree from a posterior sample of trees obtained via MCMC. We will also use this program to summarize and visualize the posterior estimates of other tree-related parameters (e.g., node heights). TreeAnnotator is also provided as a part of the BEAST2 package so you do not need to install it separately.

2.0.4 Tracer

Tracer (http://tree.bio.ed.ac.uk/software/tracer) is used to summarize the posterior estimates of the various parameters sampled via MCMC. This program can be used for visual inspection of MCMC chains and to assess their convergence. Tracer makes it easy to calculate parameter median estimates, their 95% highest posterior density (95%-HPD) intervals, their effective sample sizes (ESS), and their correlation with other parameters.

2.0.5 FigTree

The last program we will use is FigTree (http://tree.bio.ed.ac.uk/software/figtree). FigTree was designed so that users can easily visualize trees and draw publication-quality figures. FigTree interprets the annotations created by TreeAnnotator and associated to summary tree nodes; this allows the researcher to easily visualize and compare node-based statistics (e.g., posterior probabilities). We will use FigTree v1.4.4.

3 Setting up

3.1 Installing dependencies

Total-evidence dating of phylogenies is a complex task that requires a series of models, a few of which are implemented in their own BEAST2 packages. The main package for this tutorial is called `contraband` and it implements Brownian models for the evolution of multiple characters on phylogenetic trees.

In order to install `contraband`, we have to download it using the BEAUTi [2?] package manager. Open BEAUti2, go to *File >> Manage Packages*, and click on the `contraband` link (Figure 1). The package will become available in BEAUti2 once you close and restart the program.

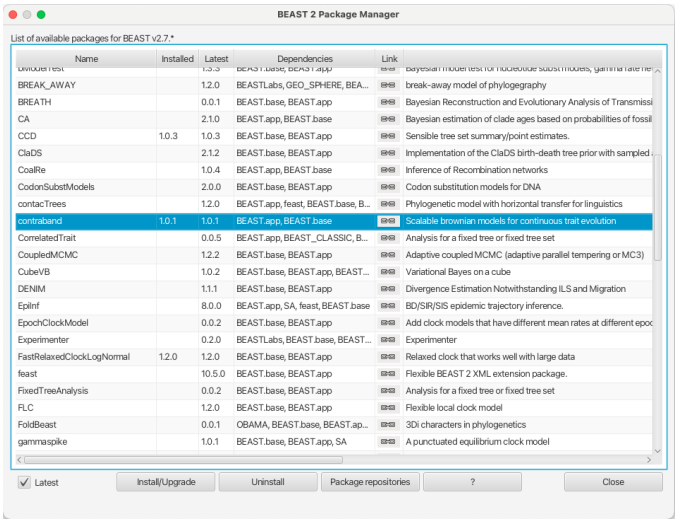

Figure 1: Download the contraband package.

This tutorial will also make use of a few other packages; these are `bdtree`, `sampled-ancestors`, and `morph-models`. The first two implement speciation and fossilization models (for evaluating the probability of phylogenies themselves), and the latter implements models for the evolution of discrete morphological characters. These packages have a dedicated tutorial (here [https://taming-the-beast.org/tutorials/Total-Evidence-Tutorial/]) and will not be discussed further in the present exercise.

3.2 Preparing the data

The data sets used in this tutorial include three types of data – molecular, discrete morphology, continuous morphology – scored fo up to 27 Carnivore species (11 of which are extinct and 16 of which are extant).

3.2.1 Continuous characters

Our TED analysis will leverage a published geometric-morphometric data set consisting of 29 three-dimensional (3D) cranium landmarks (**alvarez19**), each dimension of which will be treated as a separate continuous characters (i.e., we will have a total of 87 continuous characters). This data can be found in file `carnivora_continuous_27.nex` attached to this tutorial.

The same cranium landmarks have also been scored in 21 *Vulpes vulpes* (one of the focal carnivore species) individuals. This intraspecific data will be used in the analyses the bypasses the estimation of character correlations ($\boldsymbol{\rho}$), and can be found in attached file `vulpes_continuous_data.txt`.

3.2.2 Discrete characters

12 species of interest have discrete morphological characters that describe their basicranial, dental, postcranial anatomical features (carnivora_discrete_27.nex) (**barrett21**). There are 183 features in total and the number of character states ranges from 0 to 3.

3.2.3 Molecular sequences

The molecular sequences of 12 mitochondrial genes for 14 species of interest are collected from NCBI database and are further were concatenated, aligned using MAFFT (carnivora_dna_27.fasta).

4 Practical Part III: Parameter and State inference under Brownian motion model

4.0.1 Loading the Carnivoran Continuous data

The continuous characters can be found in the *data* folder named *carnivora_continuous_27.nex*. It can be either drag and dropped into BEAUti "Partitions" panel or added using BEAUti's menu system via *File >> Load Continuous Data*. Once the character are loaded successfully into BEAUTi, the panel will show

4.0.2 Get the fossil ages (Tip Dates)

Since the data set have fossil species, we will need to open the "Tip Dates" panel and then select the "Use tip dates" checkbox to specify the fossil ages. This can be done in multiple ways. In our case, we can obtain the date information from the species names. We can tell BEAUti to use these by clicking the *Auto-configure* button. The fossil ages appear following the second underscore "_" in the species name. To extract these times, select "use everything", then select "after last" from the drop-down box to the right, and input "_" (without the quotes) in the text box immediately to the right, as shown in the figure below Figure 2. Clicking "Ok" should now populate the table with the fossil ages extracted from the species names.

In the populated table, the two columns **Date** and **Height** should now have values between 0.0 and 35.55 in million years Figure 3.

4.0.3 Set the Brownian motion Model

As is introduced above, the parameters under the Brownian motion model include trait evolutionary rate (Sigmasq), trait correlations (Correlation) and ancestral states at the root (Root Values). Here we assume that all characters share one evolutionary rate. Therefore, we put a tick in the box in front of the "One

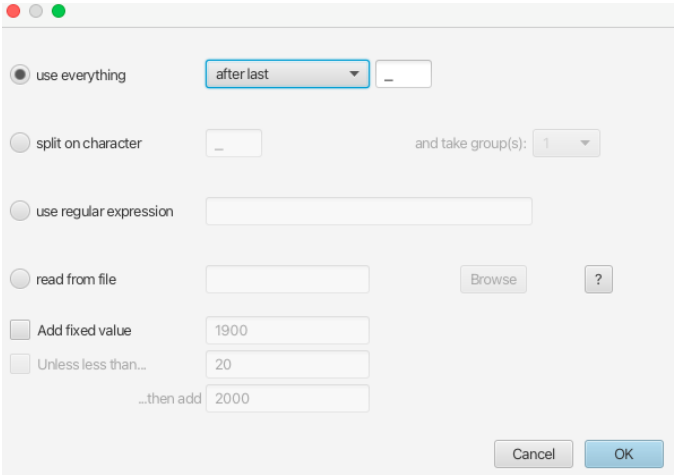

Figure 2: Guess sampling times.

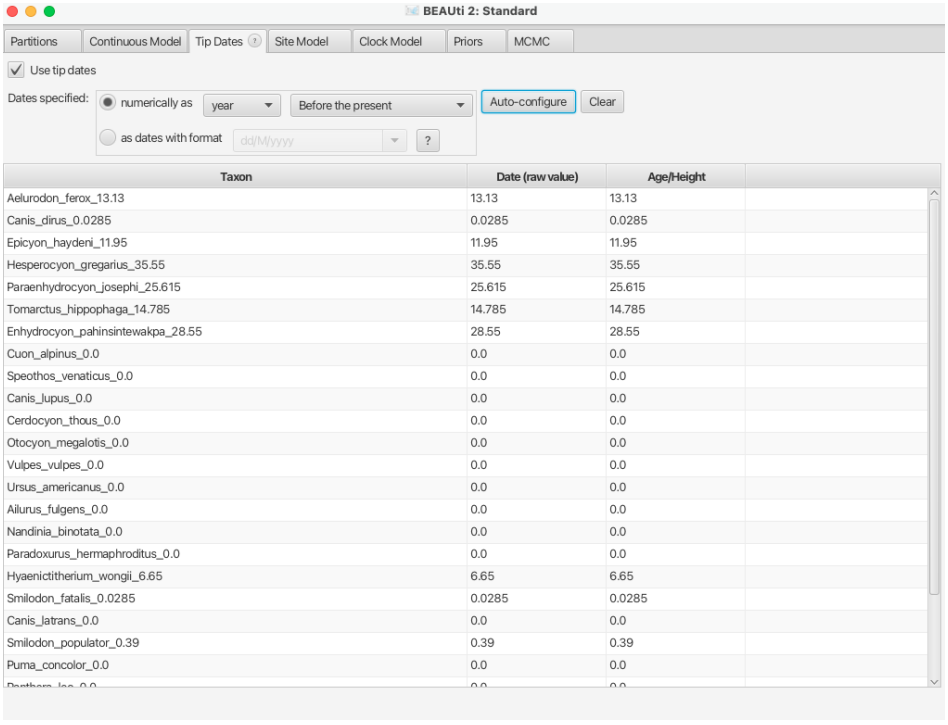

Figure 3: Fossil ages.

Rate Only".

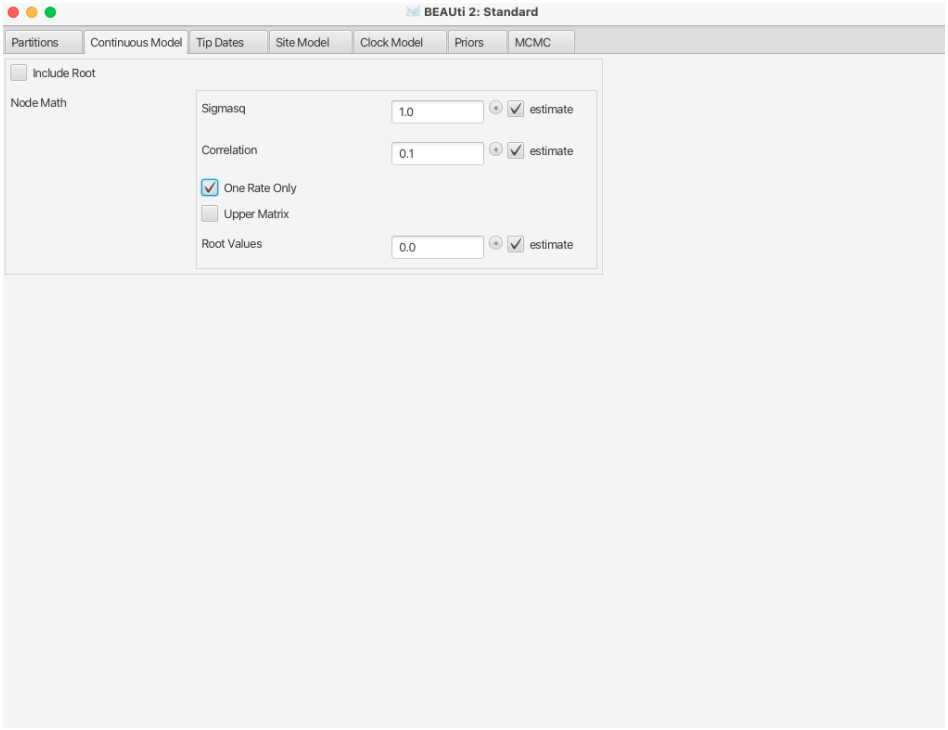

Figure 4: BM model parameter specifications.

4.0.4 Set the Clock model

We assume the relative branch-specific rates are independently distributed and follow a LogNormal distribution with a fixed mean of 1. Therefore, we specify a relaxed clock model by selecting "Optimised Relaxed Clock" in the drop-down menu, where the mean clock rate represents the global morphological clock rate that will be estimated by default. The detailed description of the model can be found in **douglas2021**.

4.0.5 Specify the priors

In the "Priors" panel, we select "Fossilized Birth Death Model" (**gavryushkina2014**) as the tree prior and leave the rest of the parameters having their default prior distributions.

4.0.6 Specify the MCMC chain length (MCMC)

Here we can set the length of the MCMC chain and after how many iterations the parameter and trees a logged. For this dataset, 2 million iterations should be sufficient. In order to have enough samples but not create too large files, we can set the logEvery to 2000, so we have 1001 samples overall. Next, we have to save the `*.xml` file under *File >> Save as*.

4.0.7 Run the Analysis using BEAST2

Run the `*.xml` using BEAST2 or use finished runs from the *precooked-runs* folder. The analysis should take about 6 to 7 minutes.

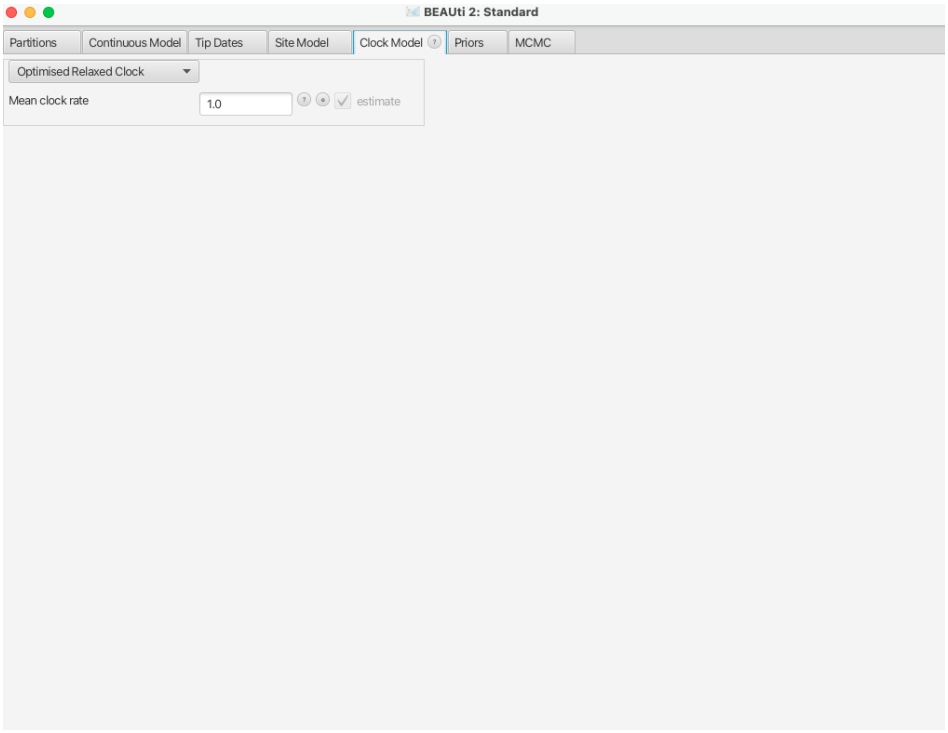

Figure 5: Set the initial clock rate.

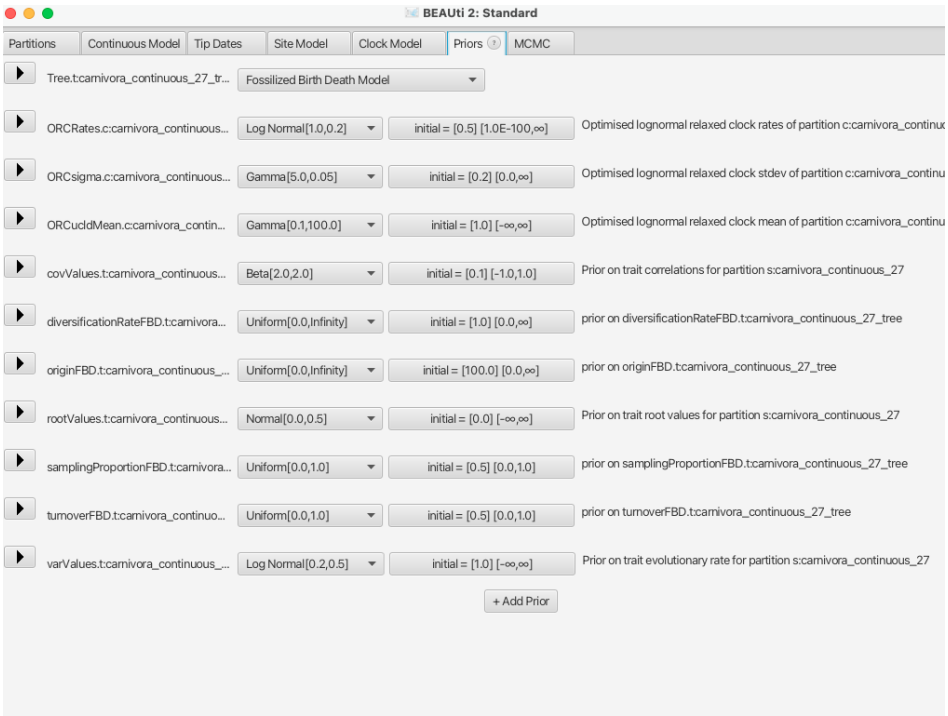

Figure 6: Set the tree model and priors on parameters.

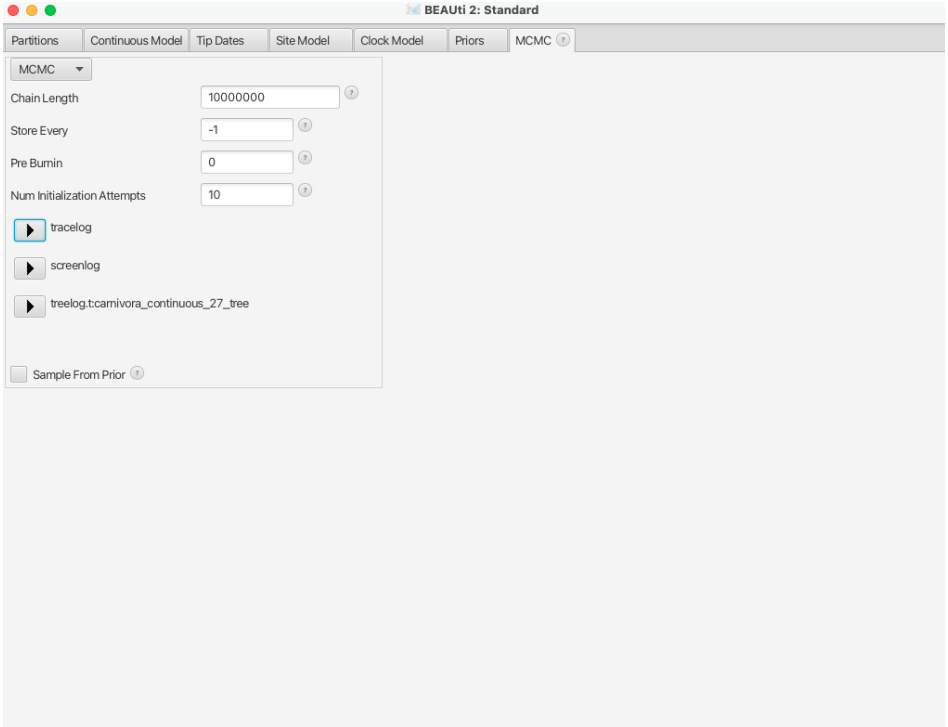

Figure 7: save the *.xml.

4.0.8 Post analysis

- Analyse the log file using Tracer
  First, we can open the `*.log` file in tracer to check if the MCMC has converged. The ESS value should be above 200 for almost all values and especially for the posterior estimates (Figure 8).
- The estimated evolutionary rate (Figure 9)
- The estimated character correlations (Figure 10)
- The estimated ancestral states at the root (Figure 11)
- The inferred morphological clock model (Figure 12)
- Make the summary tree using TreeAnnotator
  Open **TreeAnnotator** and then set the options including **Burnin percentage**, **Target tree type**, **Node heights**, **Input Tree File** and the **Output File**. Use the logged trees in the file `carnivora_continuous_27_tree_bm.trees` as **Input Tree File**. Name output file `carnivora_continuous_27_bm_mcc.tree`. After clicking **Run** the program should summarize the trees.

5 Practical Part IV: Parameter and State inference using combined data with Brownian motion model with shrinkage method

5.0.1 Loading the Carnivoran data sets

We first load the continuous data and parse the fossil ages as is mentioned in previous sections 4.0.1 and 4.0.2. Then, in the "Partitions" panel, we continue to load the Carnivoran molecular sequences via *File >> Import Alignment*. Finally, we add the discrete characters by *File >> Add Morphological Data*. As is shown in Figure 14,

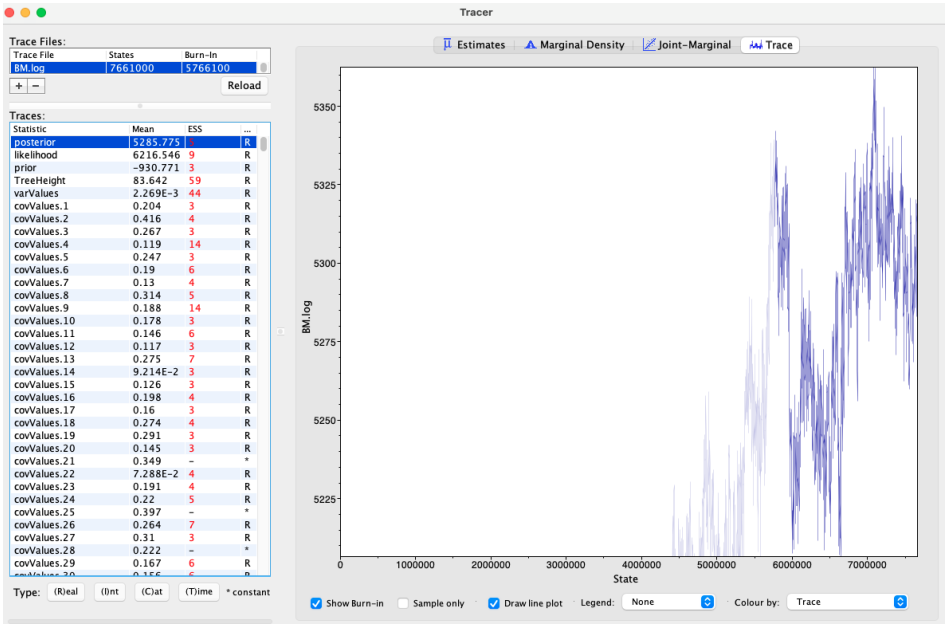

Figure 8: Check if the posterior converged.

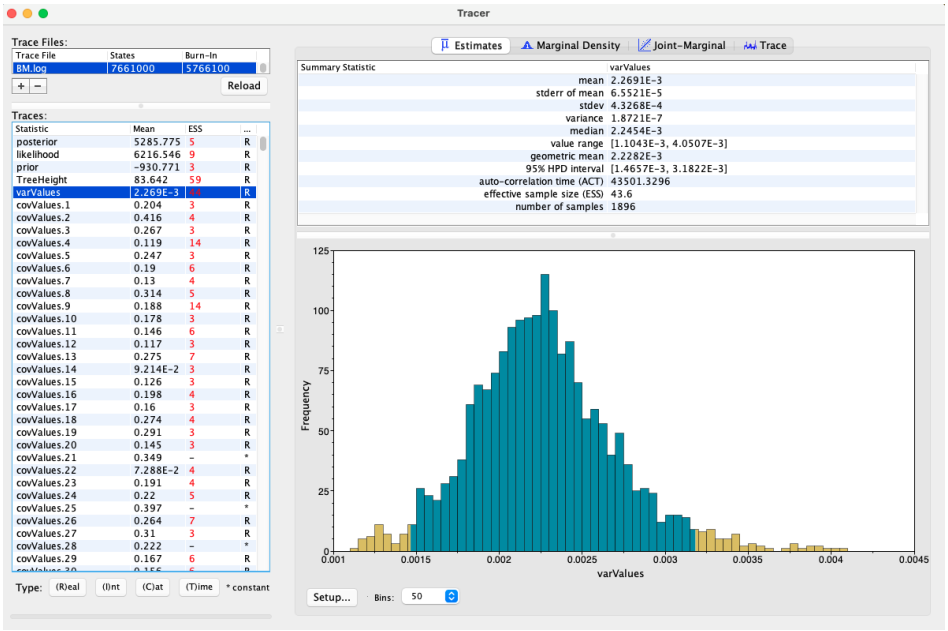

Figure 9: Evolutionary rate shared by 87 characters.

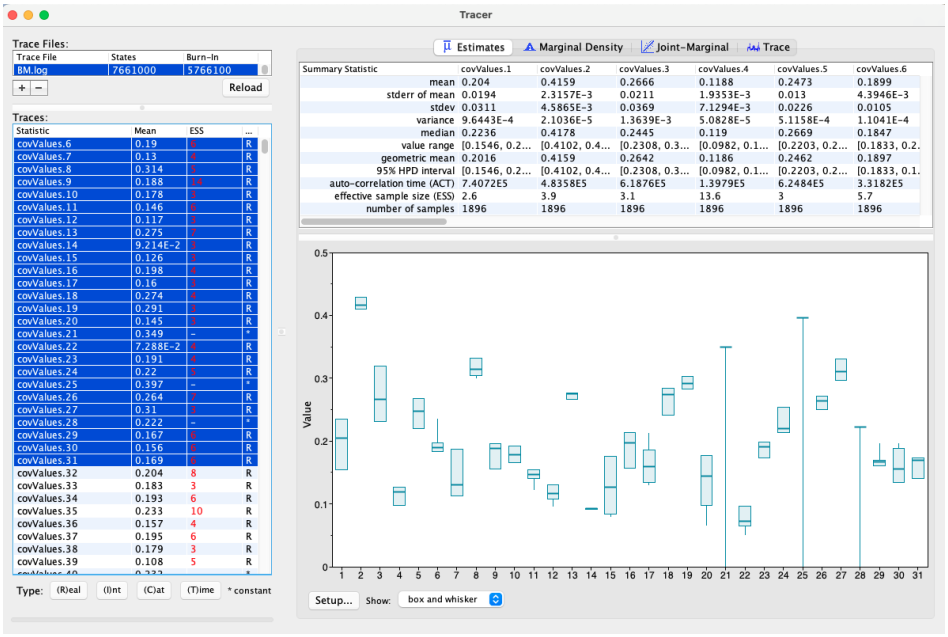

Figure 10: Character correlations among 87 characters.

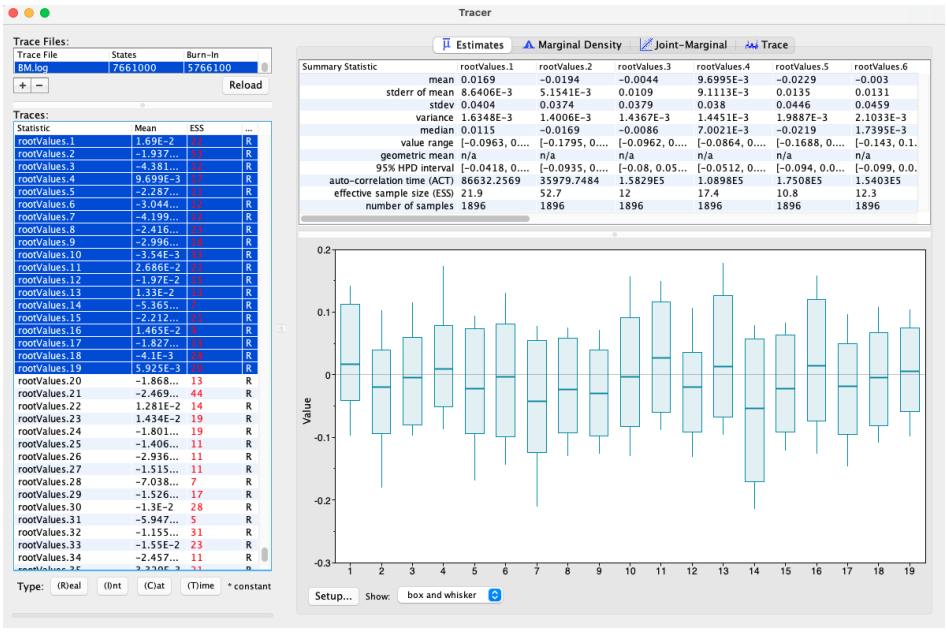

Figure 11: 87 trait values at the root of the tree.

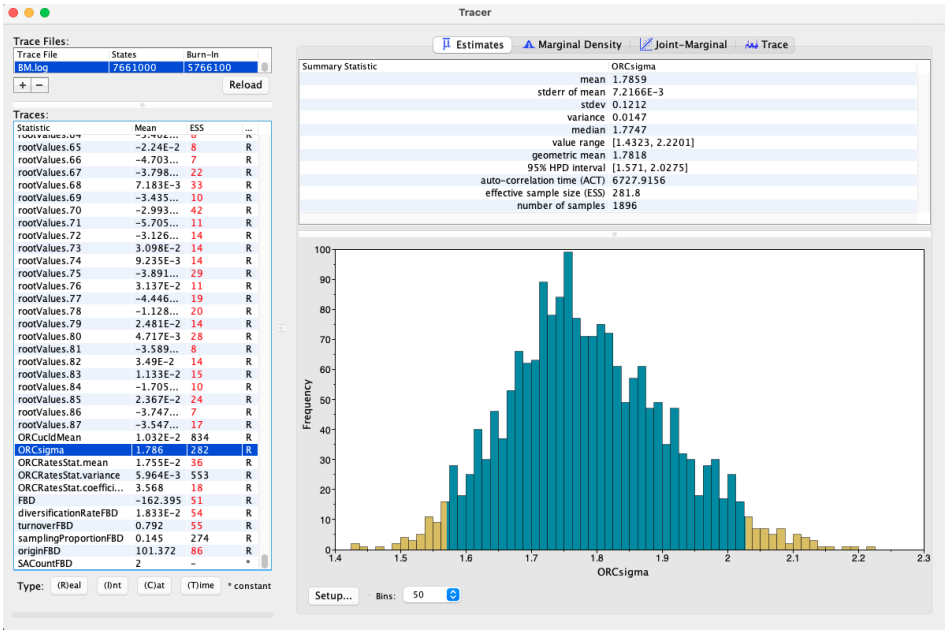

Figure 12: Compare the inferrred migration rates.

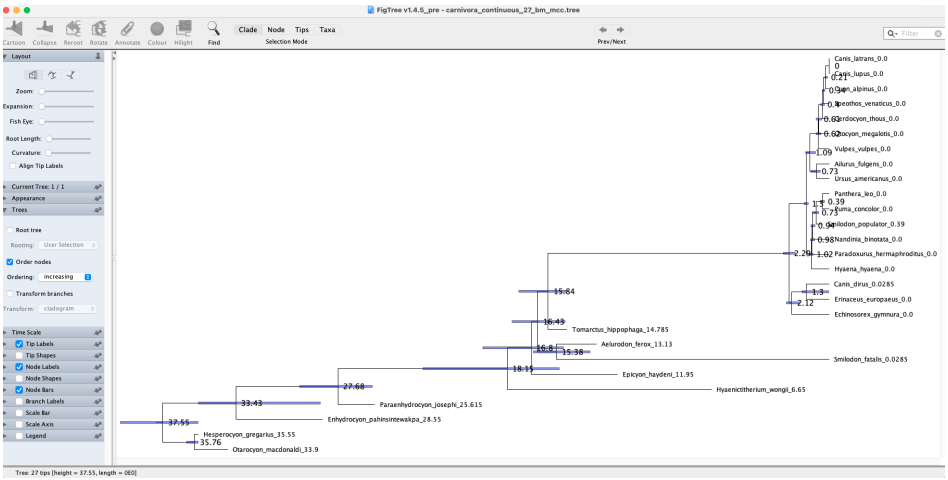

Figure 13: Summarised MCC tree of logged trees under Brownian motion model.

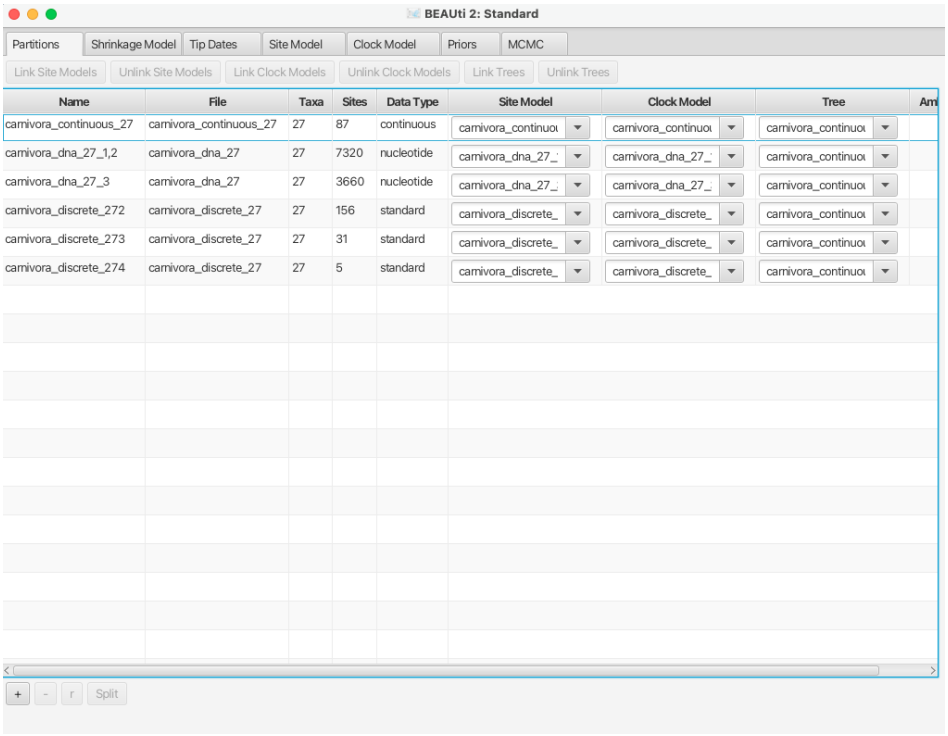

Figure 14: Load continuous characters, molecular sequences and discrete characters.

5.0.2 Set the Shrinkage Model

In the "Shrinkage Model" panel, we will need to fill in three components of the model. First, the shrinkage parameter is given by a constant value in the box to the right of "Delta". Second, the continuous characters from 21 *Vulpes vulpes* individuals are given in the block of "Population Traits". To be more specific, the trait data should be written in one-line data separated by spaces. In addition, the number of trait is given by "Minordimension" and should be consistent with the dimension of the continuous data in "Partitions" panel. Third, the added individual trait values are not only used for estimating correlations, but also normalizing the continuous data of the 19 carnivoran species. Therefore, we put a ✓ in the box in front of "Include Pop Var".

5.0.3 Set the Substitution Model

In the "Site Model" panel, we assume a HKY+Gamma for nucleotide substitutions by specifying 4 categories Figure 1. In addition, we assume Mk models (**lewis2001**) for discrete characters, as is shown in Figure 17.

5.0.4 Set the Clock model

Similar to what is mentioned in section subsubsection 4.0.4, we assume relaxed clock model for each data partition. The specifications are shown in Figure 18.

5.0.5 Specify the priors

First, we select "Fossilized Birth Death Model" from the drop-down menu and set it as the tree prior. Then we also keep the default priors for the rest of the parameters Figure 19.

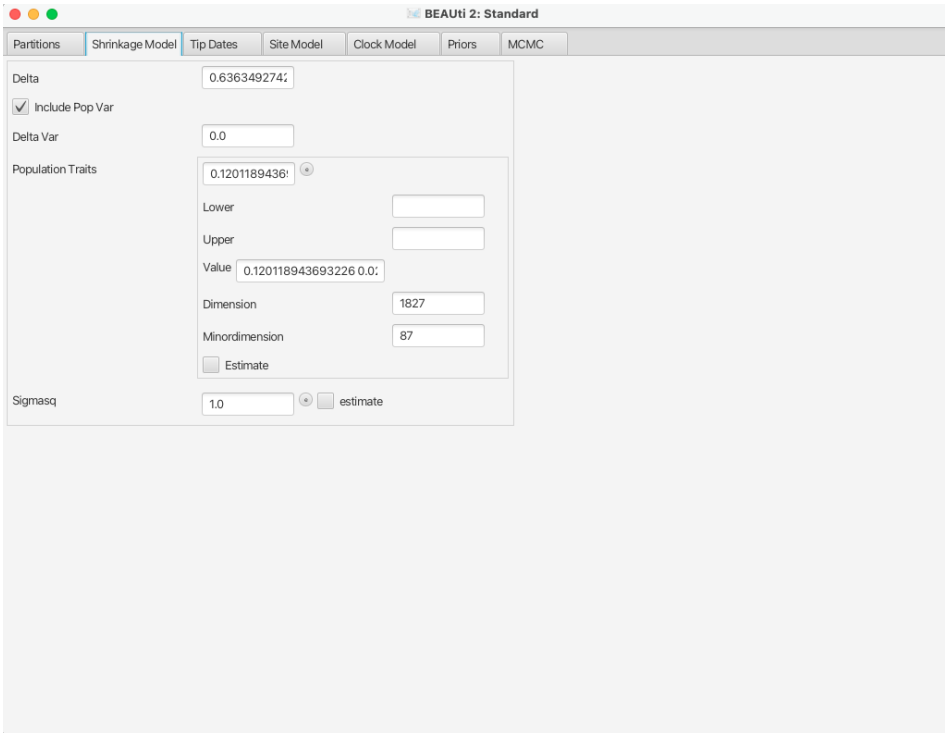

Figure 15: Set the shrinkage model.

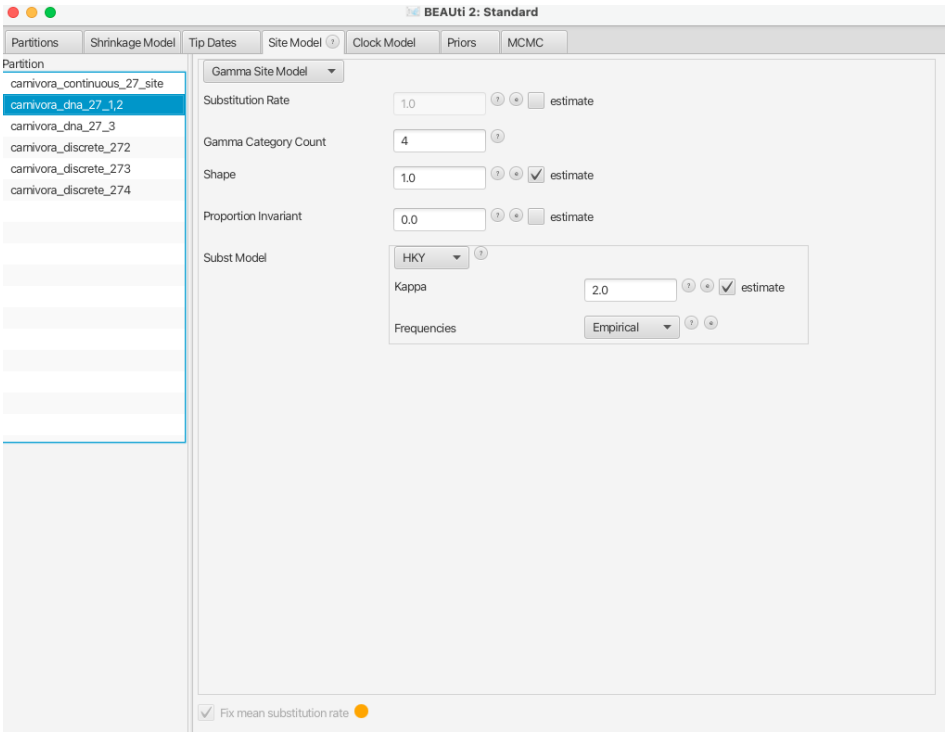

Figure 16: Set site models for molecular sequences and discrete characters.

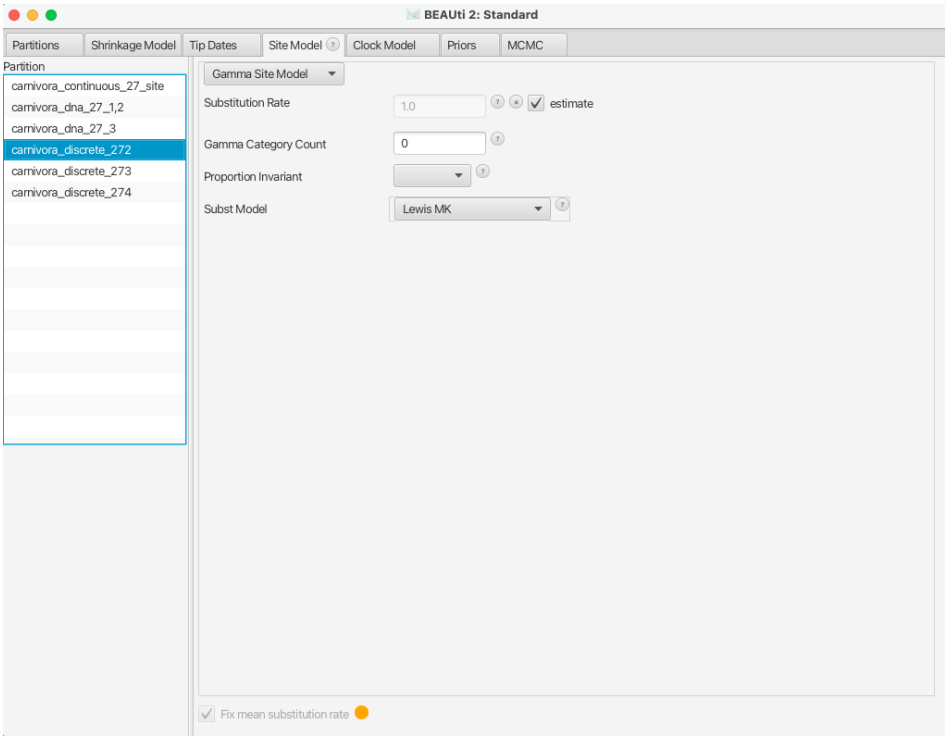

Figure 17: Set site models for molecular sequences and discrete characters.

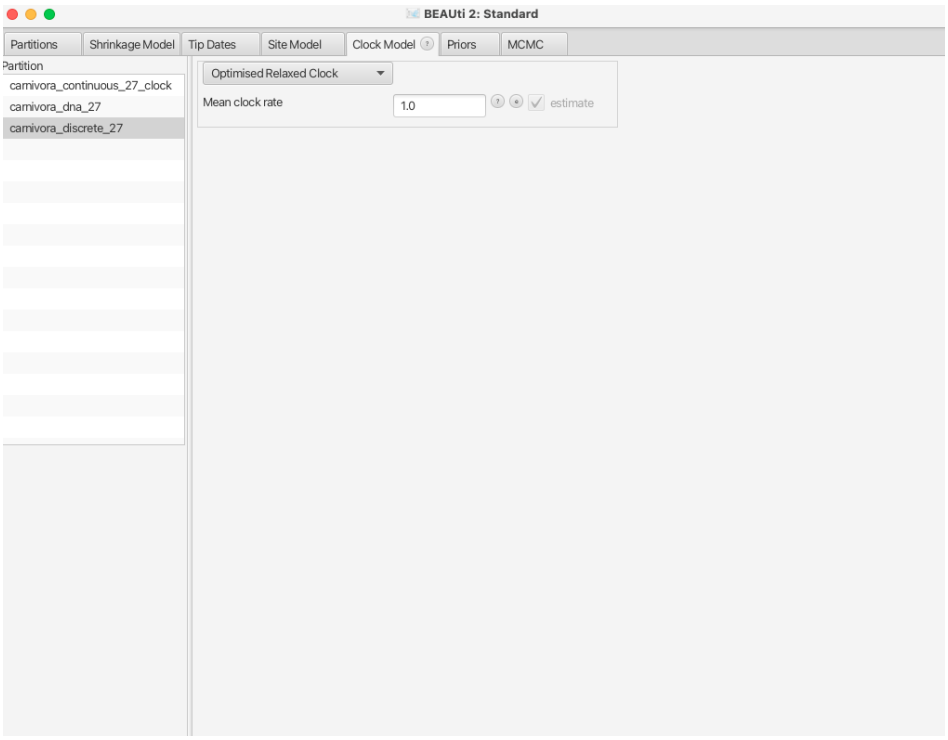

Figure 18: Set the initial clock models for continuous data, molecular data and discrete data.

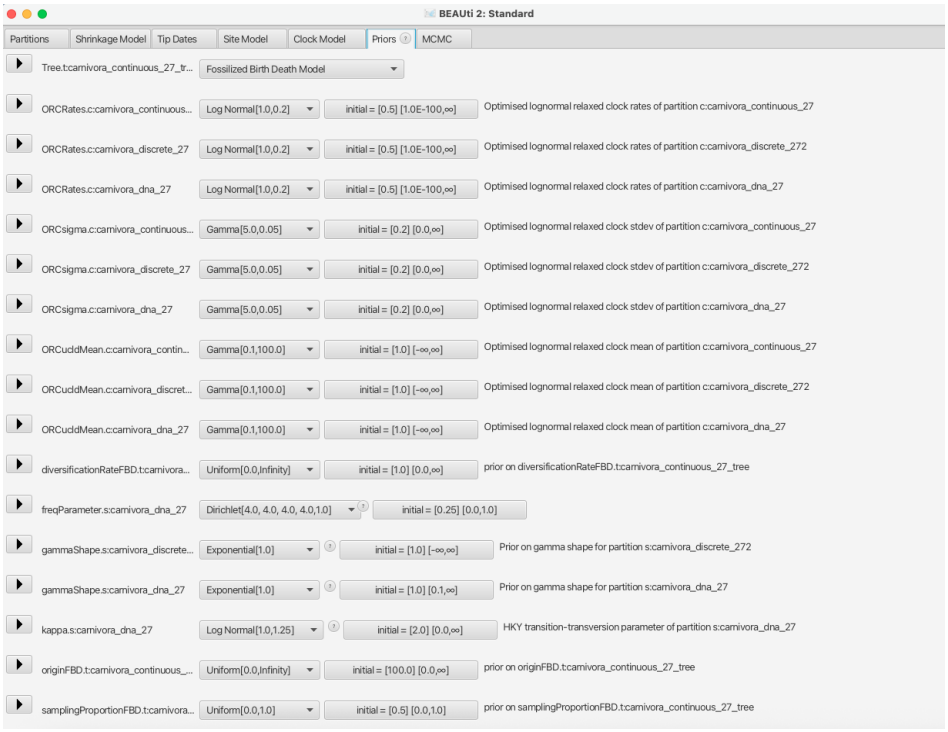

Figure 19: Set up tree model and the prior distributions.

5.0.6 Specify the MCMC chain length (MCMC)

Here we can set the length of the MCMC chain and after how many iterations the parameter and trees a logged. For this dataset, 2 million iterations should be sufficient. In order to have enough samples but not create too large files, we can set the logEvery to 2000, so we have 1001 samples overall. Next, we have to save the `*.xml` file under *File >> Save as*.

5.0.7 Run the Analysis using BEAST2

Run the `*.xml` using BEAST2 or use finished runs from the *precooked-runs* folder. The analysis should take about 6 to 7 minutes.

5.0.8 Post analysis

- Inferred clock models for continuous data, discrete data and molecular sequences (Figure 20 and Figure 21)
- Make the summary tree using TreeAnnotator (Figure 22)

6 Errors that can occur (Work in progress)

One of the errors message that can occur regularly is the following: *Infinity likelihood*

Negative branch length

Unequal likelihoods

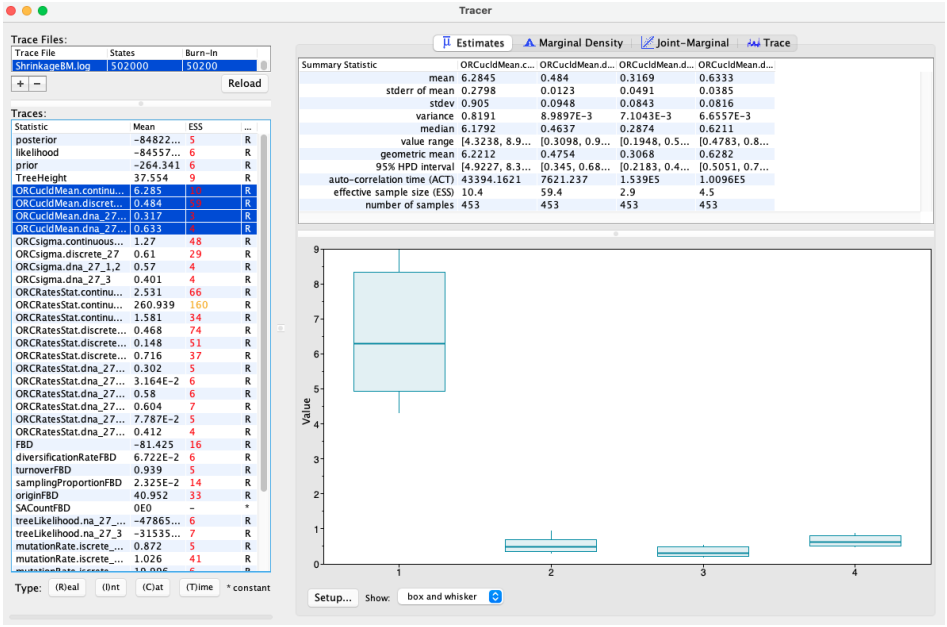

Figure 20: Estimated clock rates.

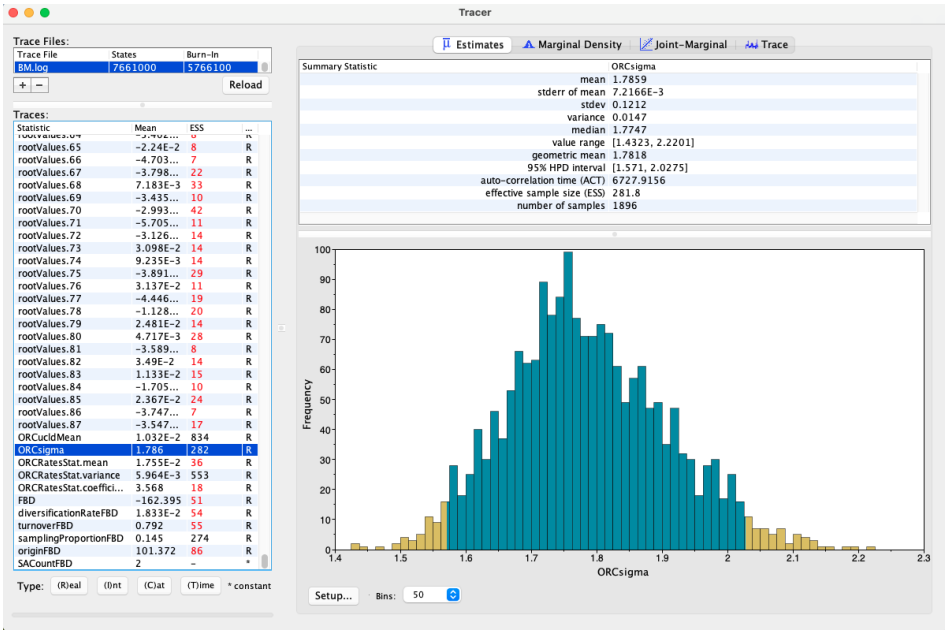

Figure 21: Estimated standard deviations.

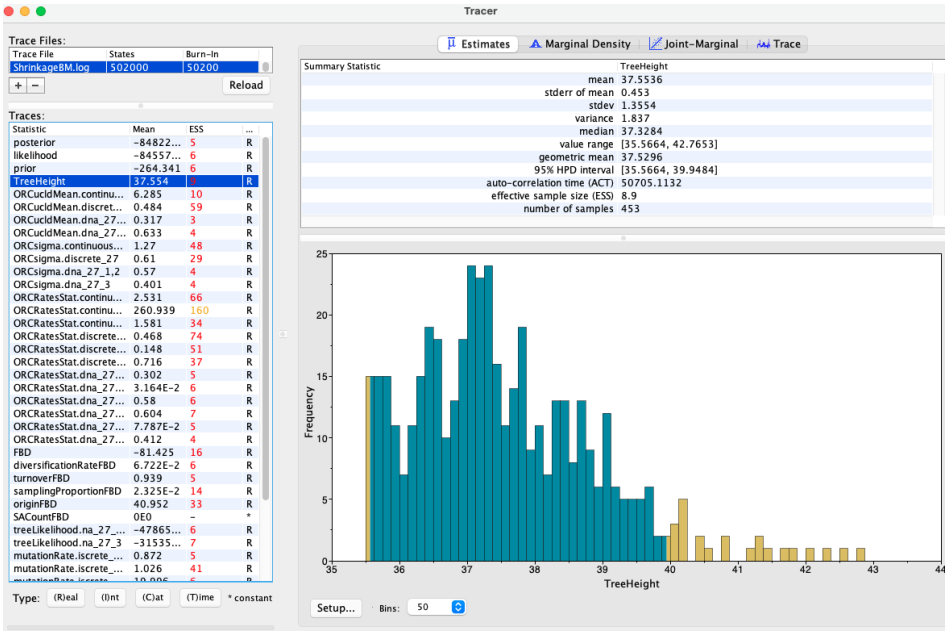

Figure 22: Summarised MCC trees estimated by combined data sets.

Version dated: July 13, 2025