StarBEAST2 v1.0.0 tutorial

Estimating species trees using StarBEAST2

Huw A. Ogilvie (huw.a.ogilvie@rice.edu)
Department of Computer Science, Rice University, Houston, Texas

1 Introduction

StarBEAST2 is a method for multilocus, multispecies coalescent phylogenetic inference (Ogilvie et al. 2017). Multilocus methods like StarBEAST2 take as input one or more multiple sequence alignments derived from genomic loci. There is assumed to be no recombination within each locus and free recombination between loci. Therefore the sequence alignments should be relatively short, and relatively distantly spaced along the genome. Genomic data sets which are typically a good fit to the StarBEAST2 model include RAD-seq loci (Ogilvie et al. 2016) or exons (Scornavacca and Galtier 2017).

StarBEAST2 can be used to estimate species tree topologies and divergence times, the topologies and coalescence times of gene trees, the substitution model rates and base frequencies for those gene trees, per-species population sizes and per-species molecular clock rates.

Newer versions of StarBEAST2 supports the integration of molecular sequences and morphological characters for so-called "total evidence" analyses, and the use of fossil data for tip-dating (**Ogilvie2021**). Fossil data can include ancient DNA, morphological characters, and the estimated ages of fossils. Tip-dating calibrates the trees in absolute time, typically millions of years.

This tutorial will cover using StarBEAST2 to estimate a species tree including per-species population sizes of the genus *Canis* from exon sequences (Section 3), to estimate per-species molecular clock rates (Section 4), and to conduct a total evidence tip-dating study (Section 5).

2 Preliminaries

Before doing anything, we need to install StarBEAST2, which is available as a package for the phylogenetic software platform BEAST2. This tutorial is for version 2.7 of the platform, which you can download from `http://beast2.org`. Make sure you have the latest update to BEAST2 version 2.7 before proceeding.

After downloading and installing BEAST2 on your computer, open BEAUti — the graphical user interface for BEAST2. To install StarBEAST2 (or any BEAST2 package), open the File menu and select Manage Packages (Figure 1).

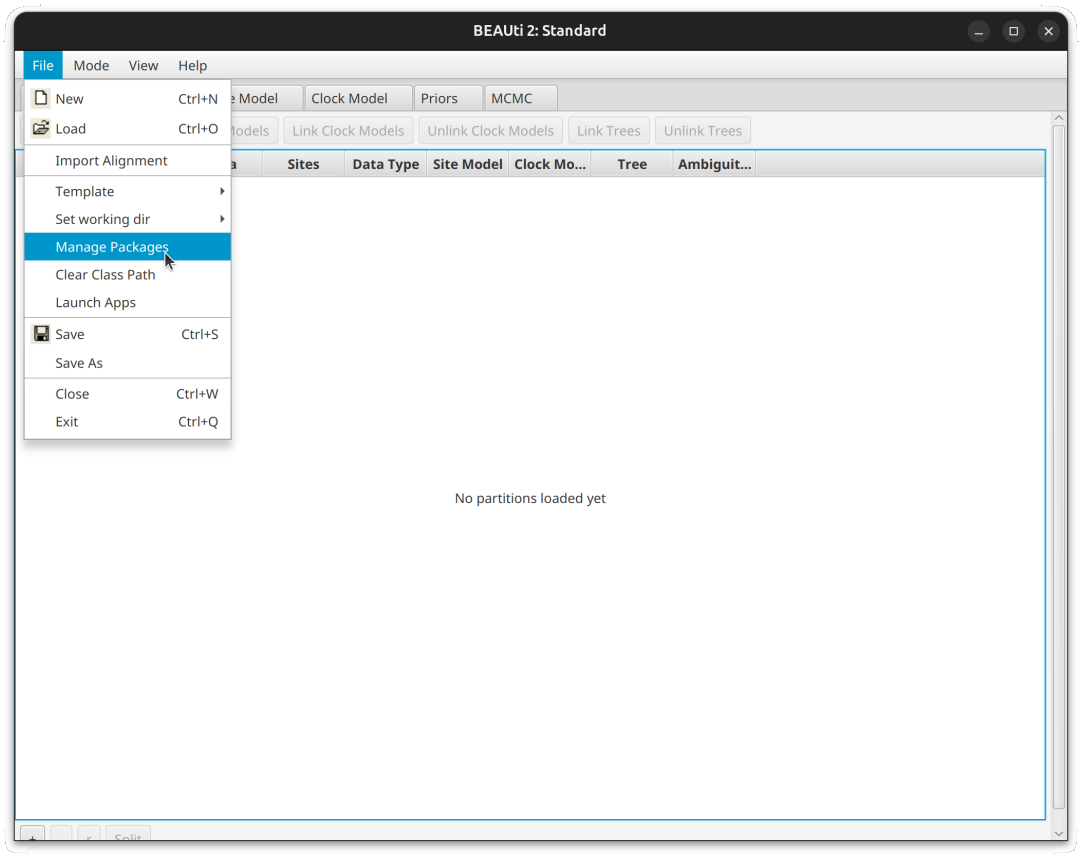

Figure 1: Opening the package manager

Select StarBEAST2 from the list of available packages and click the install button, which will take care of the installation for you (Figure 2). After the installation of StarBEAST2 is finished, just like for any BEAST2 package, you **must** quit and relaunch BEAUti.

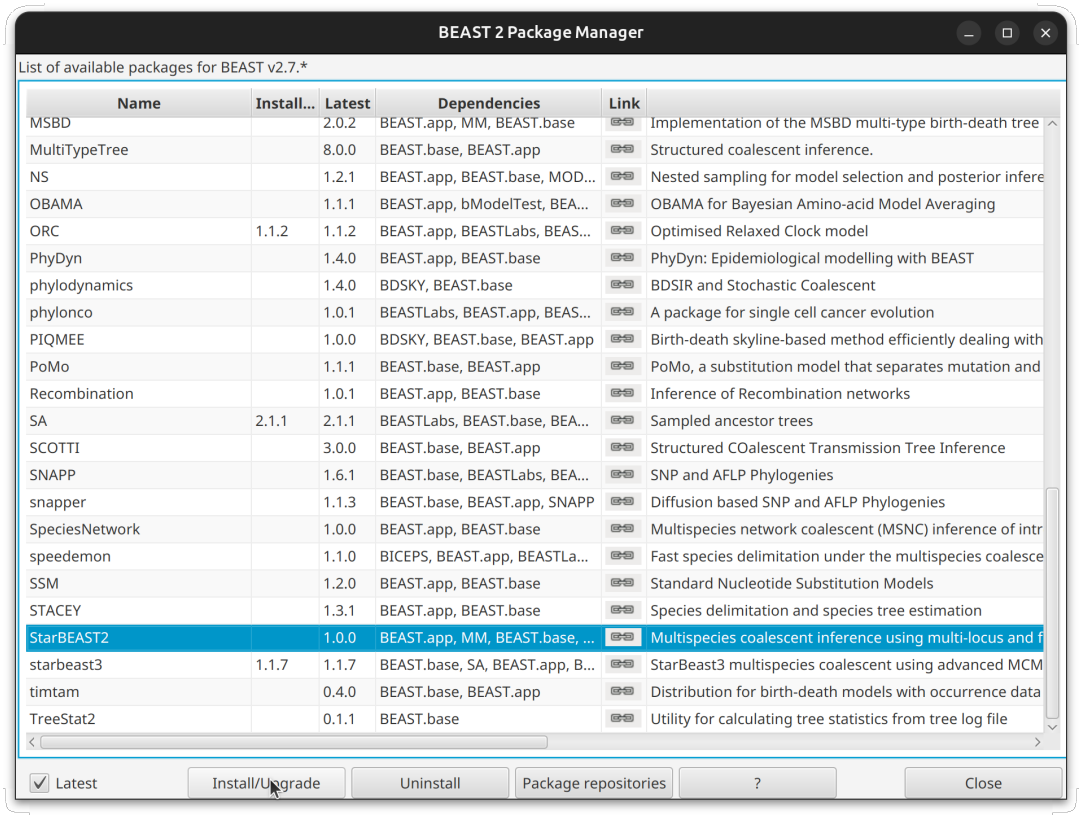

Figure 2: Installing StarBEAST2

After relaunching BEAUti, you will notice four new templates (Figure 3). The first is "StarBeast2", which is for a strict molecular clock or relaxed gene tree clocks uncorrelated with the species tree lineages. The others are "SpeciesTreeUCLN", "SpeciesTreeRLC" and "SpeciesTreeUCED" which enable per-species clock rates, as described by Ogilvie et al. 2016. The relaxed clock models used by those templates are uncorrelated log-normal (UCLN), relaxed local clock (RLC), and uncorrelated exponential distribution (UCED) respectively.

To visualize posterior distributions of trees, we will be using DensiTree (Bouckaert 2010) which is included with BEAST2. We will also use the web application IcyTree to view summary trees, which is available at `http://icytree.org/`. To analyze continuous parameters we will use Tracer, which can be downloaded from `https://beast.community/tracer`.

3 Reconstructing the species tree of *Canis*

The genus *Canis* includes species such as *Canis lupus* (wolves), *Canis latrans* (coyotes), and *Canis dirus* (dire wolves, extinct). Because the taxonomy of this genus is quite messy, it also includes *Cuon alpinus* and *Lycaon pictus*. In this section we will reconstruct the species tree of *Canis* using StarBEAST2.

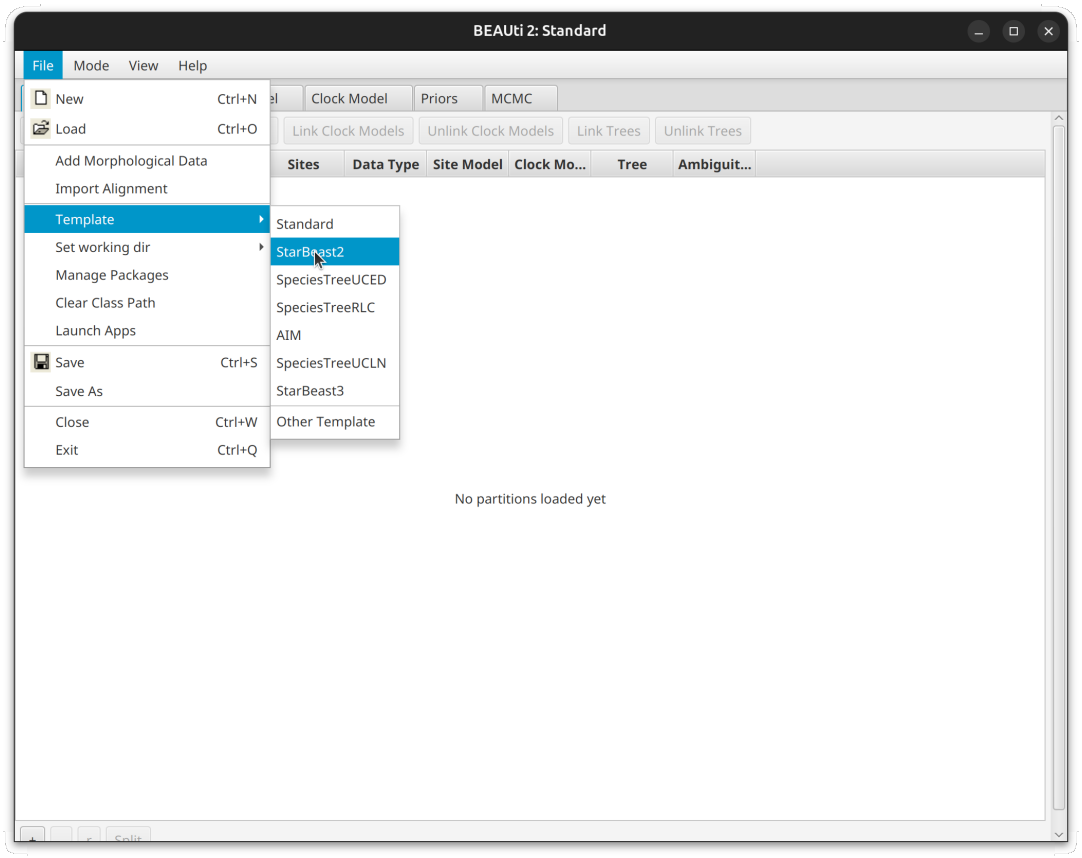

Figure 3: Selecting a StarBEAST2 template

3.1 Importing alignments

First up, create a new folder somewhere with a sensible name like "CanisPhylogeny". Relaunch BEAUti and load up the strict clock template by selecting "StarBeast2" from the Templates submenu of the File menu (Figure 3). Now we will import multiple sequence alignments of the 16 nuclear loci sequenced by Lindblad-Toh et al. 2005. Select Import Alignments from the File menu, and navigate to the "data" subfolder of the tutorial. Select all 16 FASTA files, and click OK to import (Figure 4). Do not select "morphology-canis.nexus" which is for the total evidence study.

After you click OK, you will be asked what type of sequences are in the files. As all 16 files are MSAs of nucleotide sequences, select "all are nucleotide" and click OK again to continue. There should now be 16 partitions listed in BEAUti.

3.2 Linking models

As a general rule, clock models should be linked when using the strict clock StarBEAST2 template. Rates between loci are still allowed to vary and the mean rate among all sites is fixed at 1. Therefore by linking the clocks, the average clock rate can be fixed at an *a priori* value to calibrate the analysis, or estimated when node-dating or tip-dating is used. Select all 16 loci and click the "Link Clock Models" button. Now all loci should have "APOBS1" as the clock model name. The site model, clock model and tree names are text fields that you can click on to edit. Change the name of the clock model to "molecular" and it should

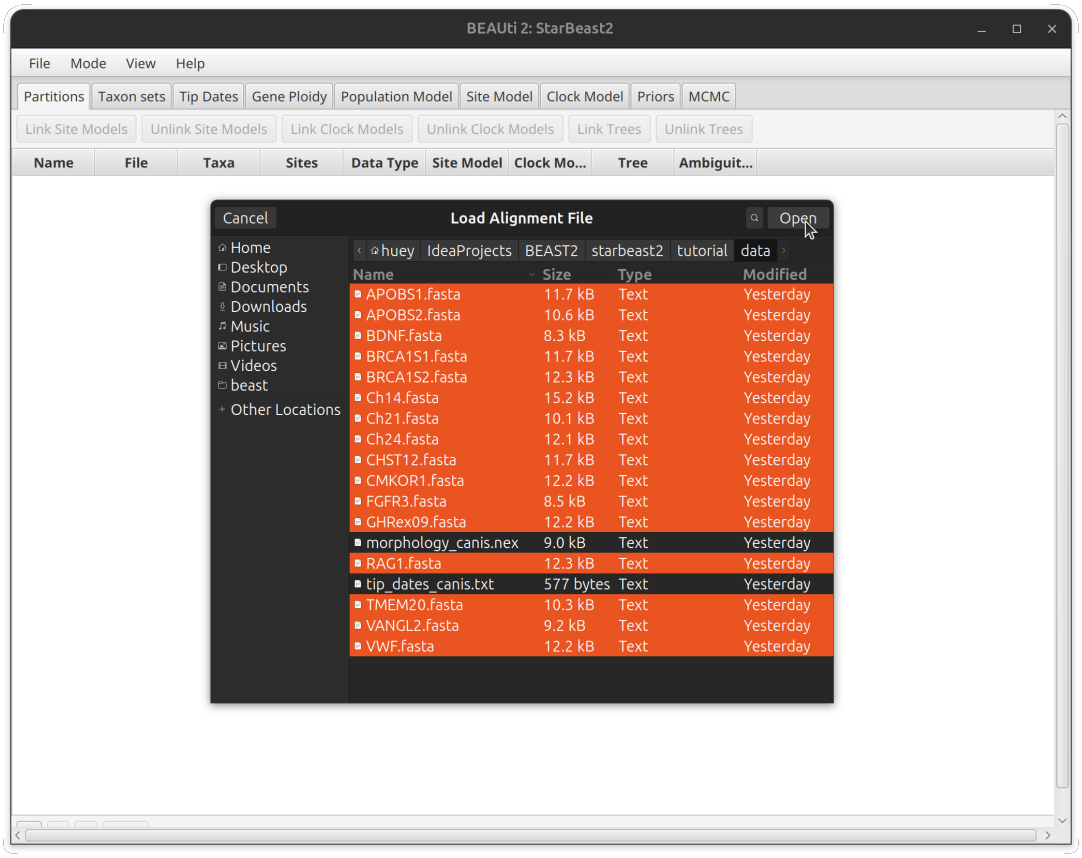

Figure 4: Importing *Canis* multiple sequence alignments

look like Figure 5.

Mixing nuclear and mitochondrial loci

StarBEAST2 can be used with a combination of mitochondrial and nuclear loci, but this requires careful thought when specifying the model. One possibility is to link the nuclear loci as above, but not to link the mitochondrial loci. Then when editing the site models, untick "estimate" for the substitution rate of each mitochondrial locus. Now separate clock rates will be used for each mitochondrial locus, in addition to the overall clock rate which applies to nuclear loci.

If the species tree is calibrated using an *a priori* nuclear or mitochondrial molecular clock rate, you can set that rate in the Clock Model panel. If the mitochondrial rate is fixed, the nuclear rate should probably be estimated and *vice versa*. If the species tree is calibrated using node or tip dating, then all clock rates should probably be estimated.

Make sure appropriate priors are used for all estimated clock rates. This could be a $1/X$ prior for rates where you genuinely have no prior knowledge. However in most circumstances you will have a general idea of the clock rate, e.g. within an order of magnitude. In that case, you can set a broad prior with a mean based on your prior knowledge. Broad priors include Log Normal with a large standard deviation (up to about 2), or a Gamma distribution with a small α (set "alpha" to 1 or 2).

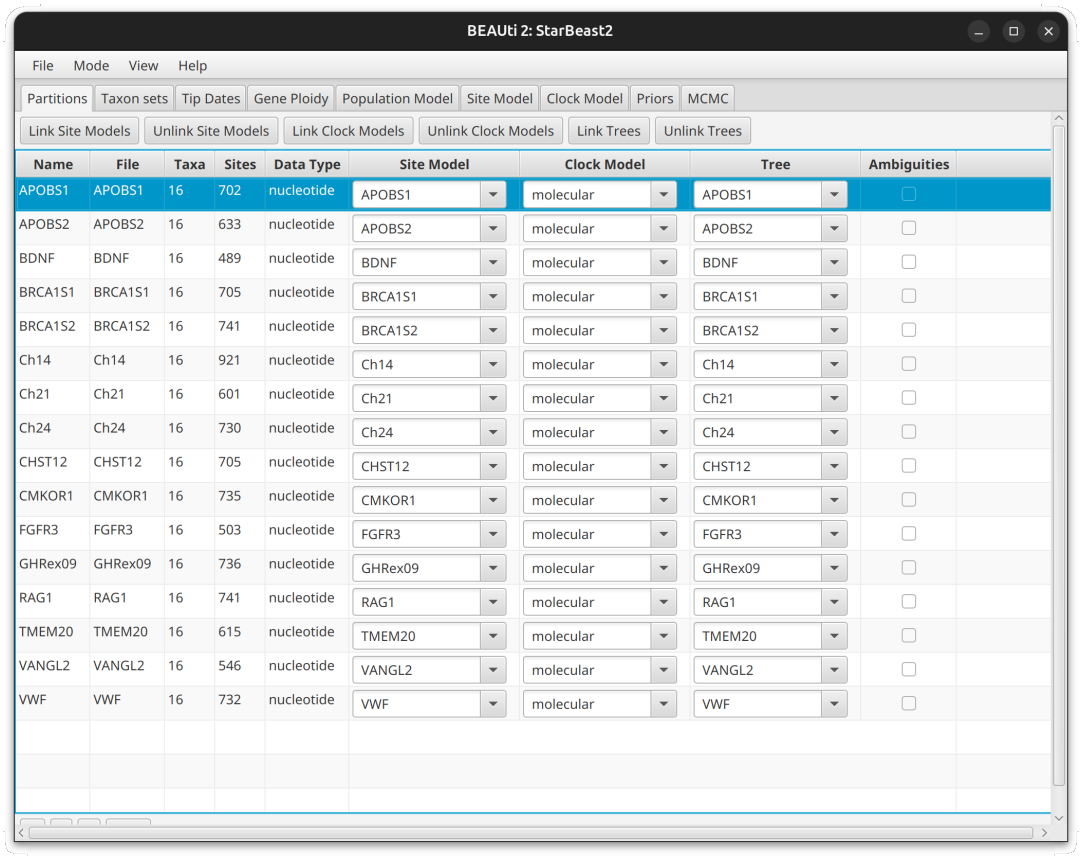

Figure 5: Linking clock models

Linking site models

When clock models are linked as is recommended by this tutorial, linking site models will result in every gene having exactly the same clock rate. As this is unrealistic, generally site models should **not** be linked using BEAUti when setting up a StarBEAST2 analysis.

3.3 Specifying species names

In the FASTA files, each sequence has a name like "Canis_anthus_a". Here *Canis anthus* is the binomial name, and "a" is a haplotype. For the data supplied with this tutorial, each species has two haplotypes "a" and "b", both from the same diploid individual. However in other studies multiple individuals may be sequenced, so an arbitrary number of haplotypes are available per species.

To assign haplotypes to species, select the Taxon Sets tab in BEAUti, then click the Guess button. To assign the correct names, keep "use everything" selected, but change "after first" to "before last". Leave the underscore in the text box and click OK (Figure 6). This way everything before the last underscore in the haplotype name will be used as the species name, so "Canis_anthus" will be the species name for "Canis_anthus_a".

Now each haplotype (Taxon) should have a corresponding species (Species/Population). Based on how we assigned the species names, there should be two haplotypes "a" and "b" for each species (Figure 7).

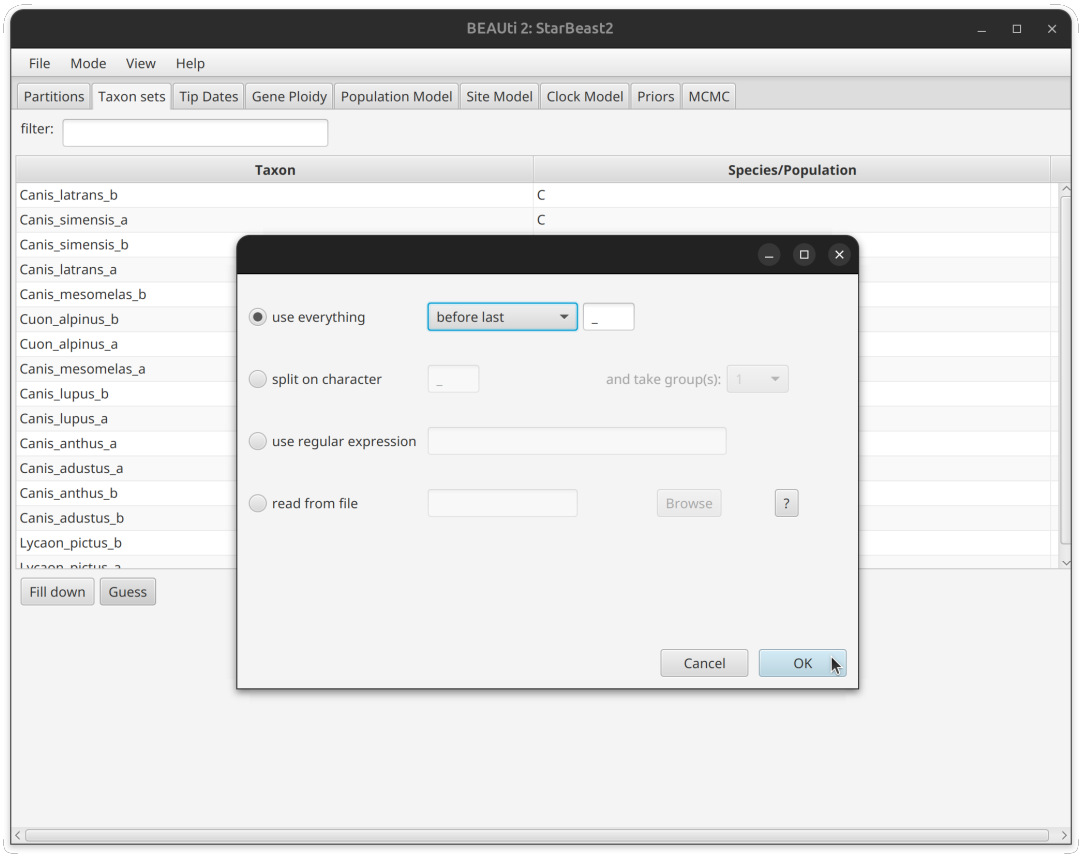

Figure 6: Guessing species names from haplotype names

3.4 Gene Ploidy and Population Model

Ignore the Tip Dates tab, which is for tip-dating and will be covered in section 5 of the tutorial. Open the Gene Ploidy tab, and observe that the default value for all loci is 2.0. This is because there are two copies of a locus in each individual for diploid populations, so we scale the effective population sizes by 2.0. If you use any mitochondrial or Y/W chromosomal loci, you should change their ploidy to 0.5, because there is effectively only half as many copies of those molecules as there are individuals (male mitochondrial genomes do not count, as they are not inherited by offspring). The ploidy of X/Z chromosomal loci should be set to 1.5 for the same reason.

Select the Population Model tab. By default analytical integration is used, which is slightly faster but does not produce estimates of per-species population sizes. Change the model to “Constant Populations”, which will add effective population sizes to the species tree output (Figure 8).

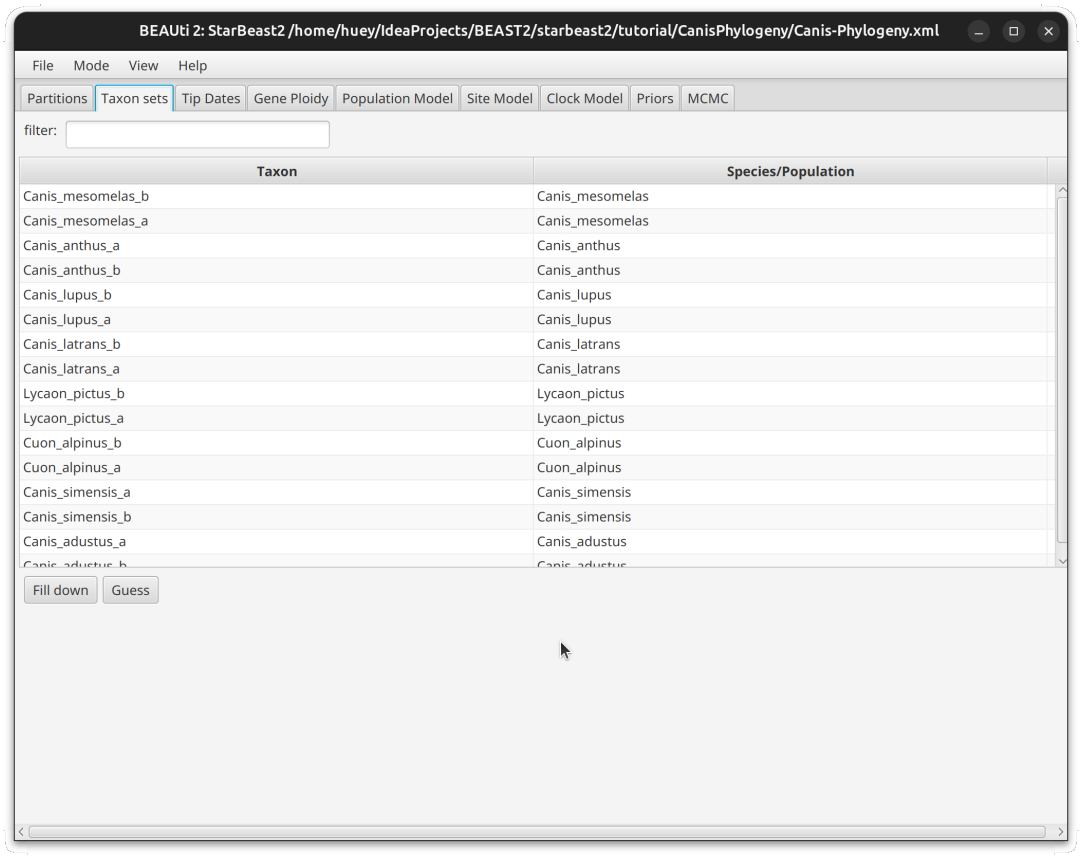

Figure 7: The assignment of haplotypes to species

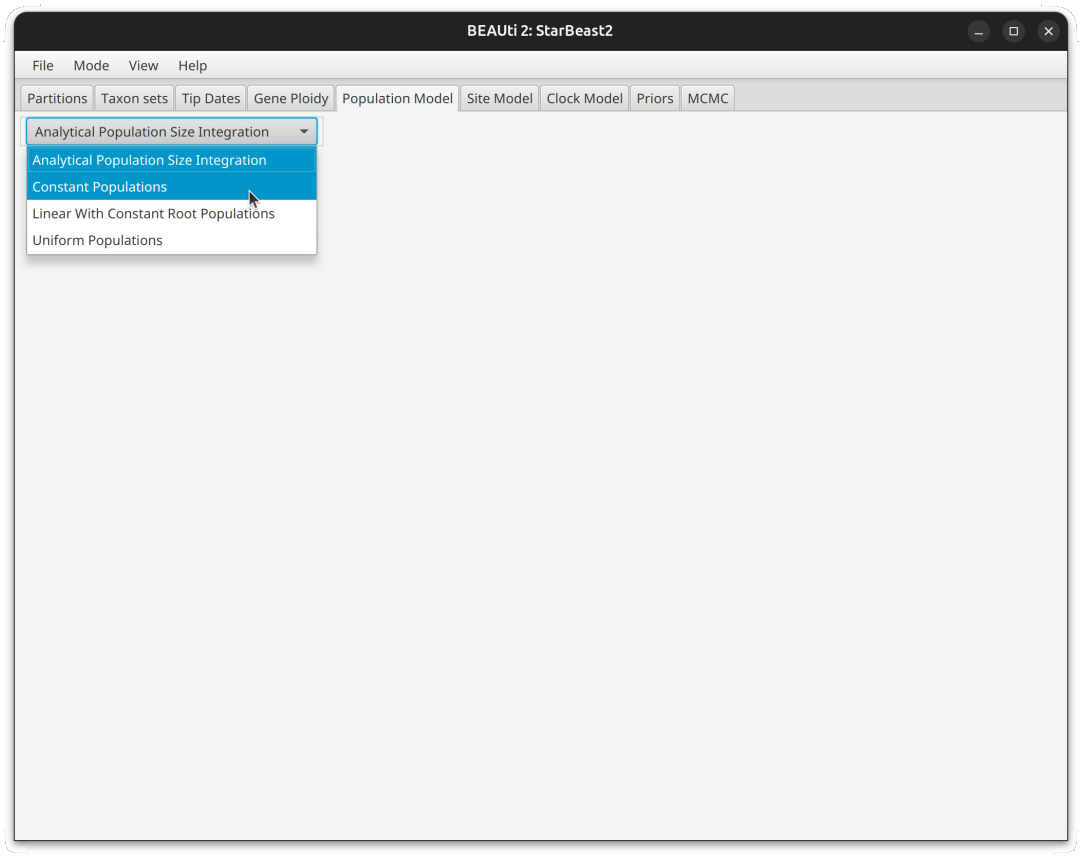

Figure 8: Changing the population model

3.5 Site Model

Select the Site Model tab, and you will see the site model for the first partition displayed. Change "JC69" to "HKY", and then set the frequencies to empirical (Figure 9). HKY is more flexible because it allows nucleotide transitions to have a different rate relative to transversions (Hasegawa et al. 1985).

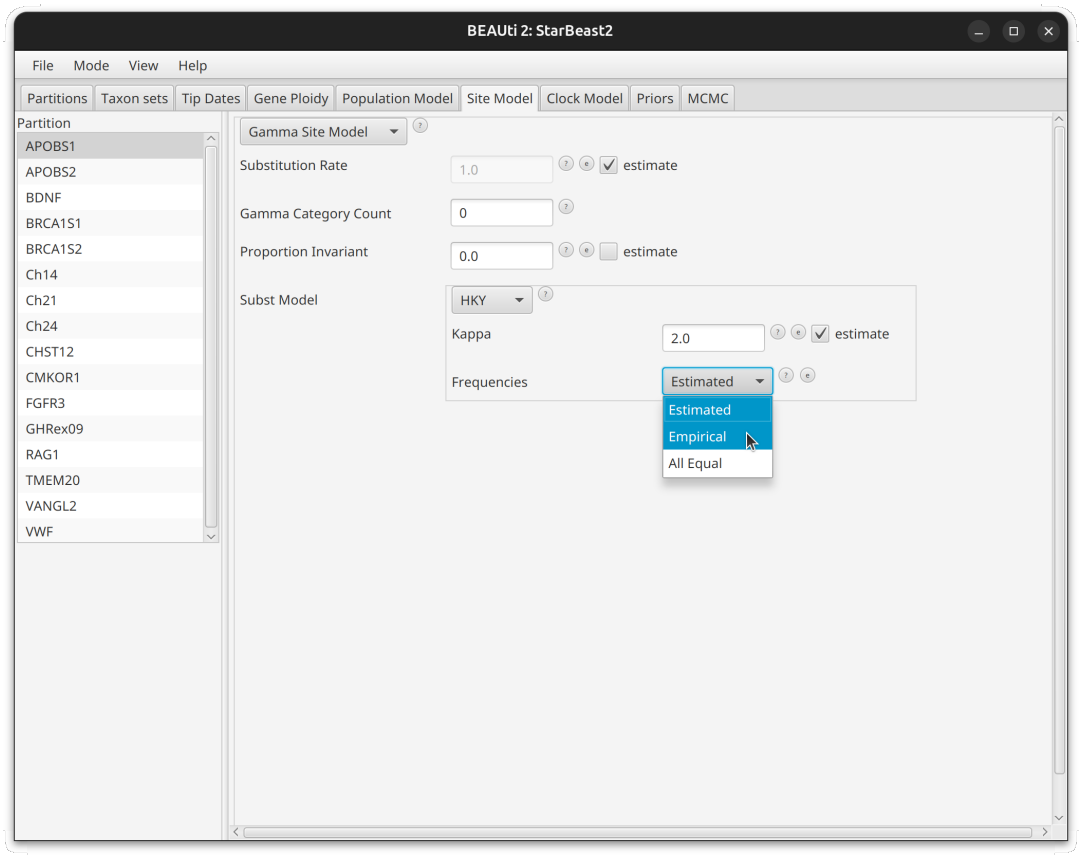

Figure 9: Setting the site model to HKY with empirical base frequencies

Now to change the site model for *all* loci to HKY with empirical base frequencies, select all of the partitions in the left hand column with the shift key. Choose to clone the site model you just configured (presumably for APOBS1), then click "OK" to apply the same site model used for the first locus to everything else (Figure 10).

In a more serious analysis, you might consider setting the number of gamma categories to 4. This allows for different sites to evolve at different rates, an obviously more realistic model (Yang 1994). However the phylogenetic likelihood must be calculated once for each category, so 4 categories will be be 4× slower. That likelihood calculation is a major part of the StarBEAST2 algorithm, so more gamma rate categories will require more computer time.

3.6 Clock Model

Hugall et al. 2007 estimated that the molecular clock rate for the RAG-1 nuclear coding gene in mammals is approximately 10^{-3} substitutions per site per million years. While we don't know the average rate across all loci in our data within *Canis*, we can use it as a **very approximate** calibration. Open the clock model

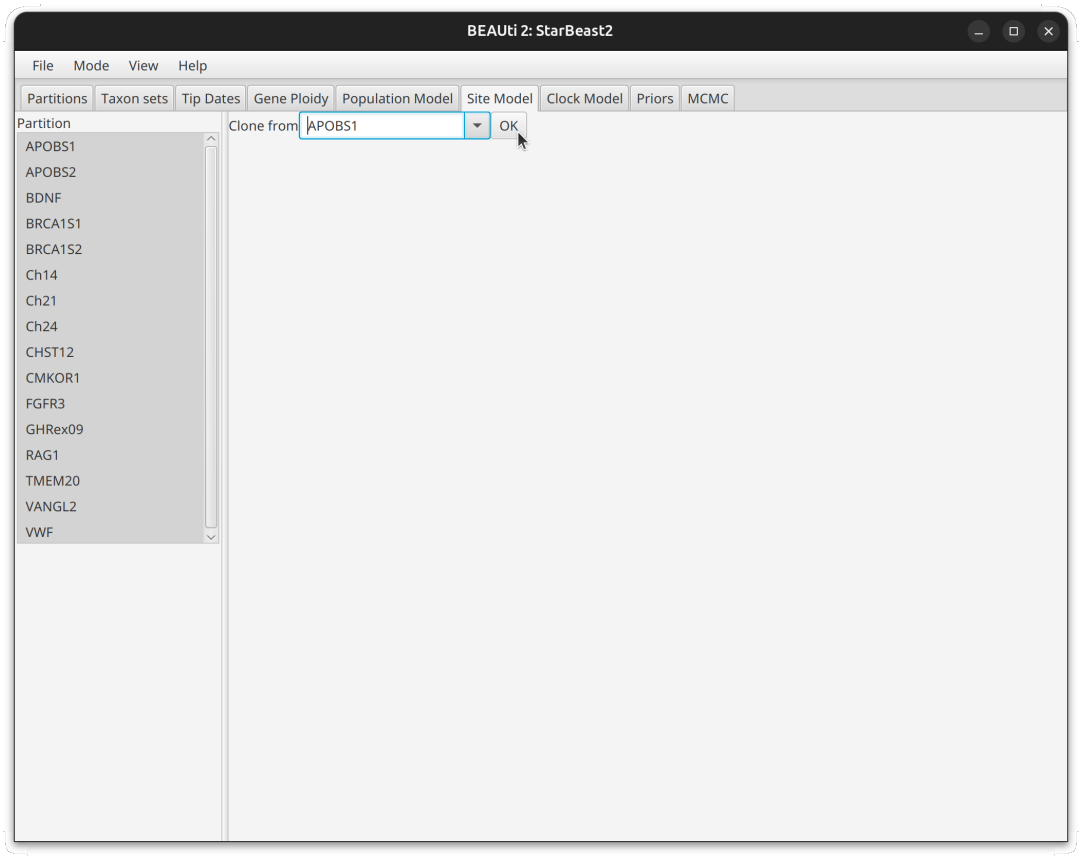

Figure 10: Cloning the site model

panel and set the rate to "0.001", to match the *a priori* estimate (Figure 11).

You can also use the scientific notation shorthand, 1e-3, which is equivalent to 0.001.

3.7 Priors and MCMC

Open the Priors panel and change "Yule Model" to "Birth Death Model". These models are identical, except the birth-death model allows for a non-zero rate of extinction.

StarBEAST2 is an MCMC method. These kind of methods do better at estimating the posterior distributions (of trees or other parameters) the longer they are run, although after a point there are diminishing returns. The default chain length in StarBEAST2 is 10 million states, but for this analysis we need a bit more for good estimates; change the chain length to 40 million (Figure 12)

Save your model as an XML file – in the folder you created before importing alignments – by clicking Save As in the File menu. Navigate to the folder and give your XML file a name like "CanisPhylogeny.xml".

3.8 Running BEAST

You can run BEAST from the command line or using the GUI. To run your XML from the command line, first navigate to the folder you saved the XML file into. Then run "/path/to/beast/bin/beast CanisPhylogeny.xml". Make sure to change /path/to/beast to match the folder where BEAST is installed on your

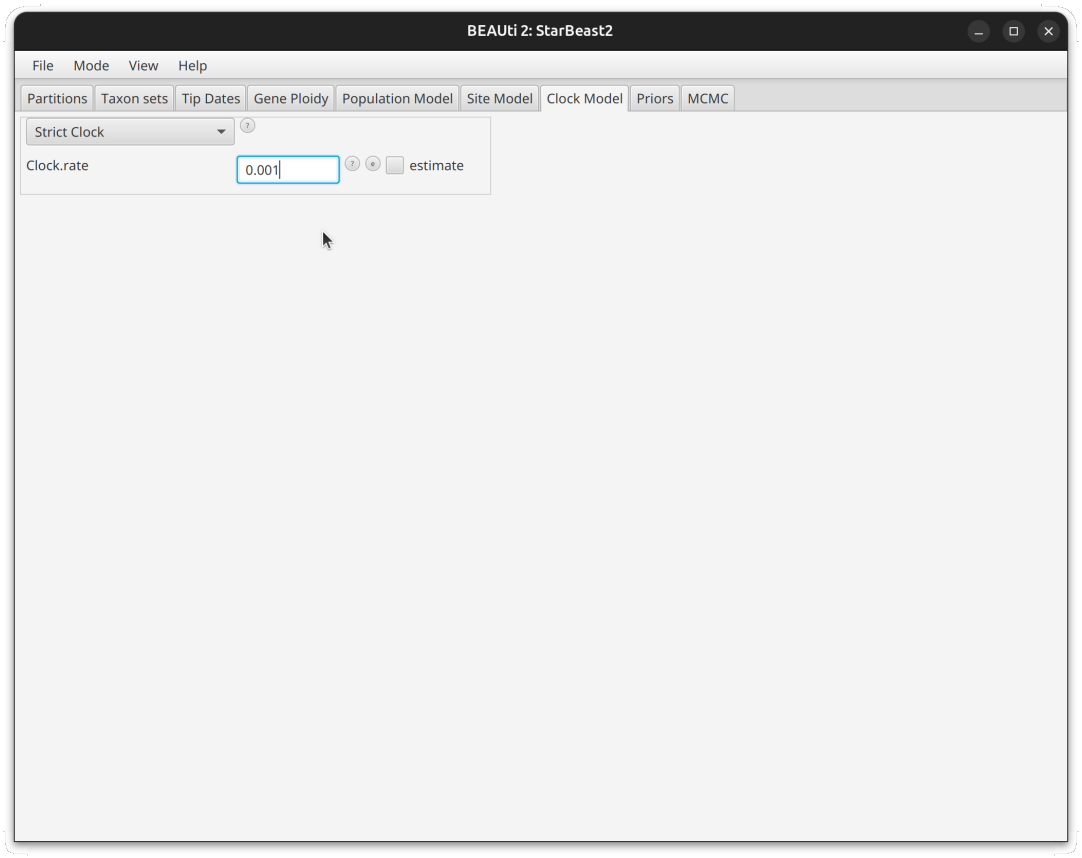

Figure 11: Using an *a priori* clock rate for calibration

computer.

BEAST should take about 5 to 10 minutes to finish the chain. Sampled statistics and various parameters will be saved to "starbeast.log", species trees to "species.trees", and separate gene tree files will be created for each locus. The command line output should start off looking something like what follows:

```
                BEAST v2.7.5 Prerelease, 2002-2023
          Bayesian Evolutionary Analysis Sampling Trees
                    Designed and developed by
Remco Bouckaert, Alexei J. Drummond, Andrew Rambaut & Marc A. Suchard

                Centre for Computational Evolution
                      University of Auckland
                    r.bouckaert@auckland.ac.nz
                     alexei@cs.auckland.ac.nz

                Institute of Evolutionary Biology
                     University of Edinburgh
                        a.rambaut@ed.ac.uk

                 David Geffen School of Medicine
```

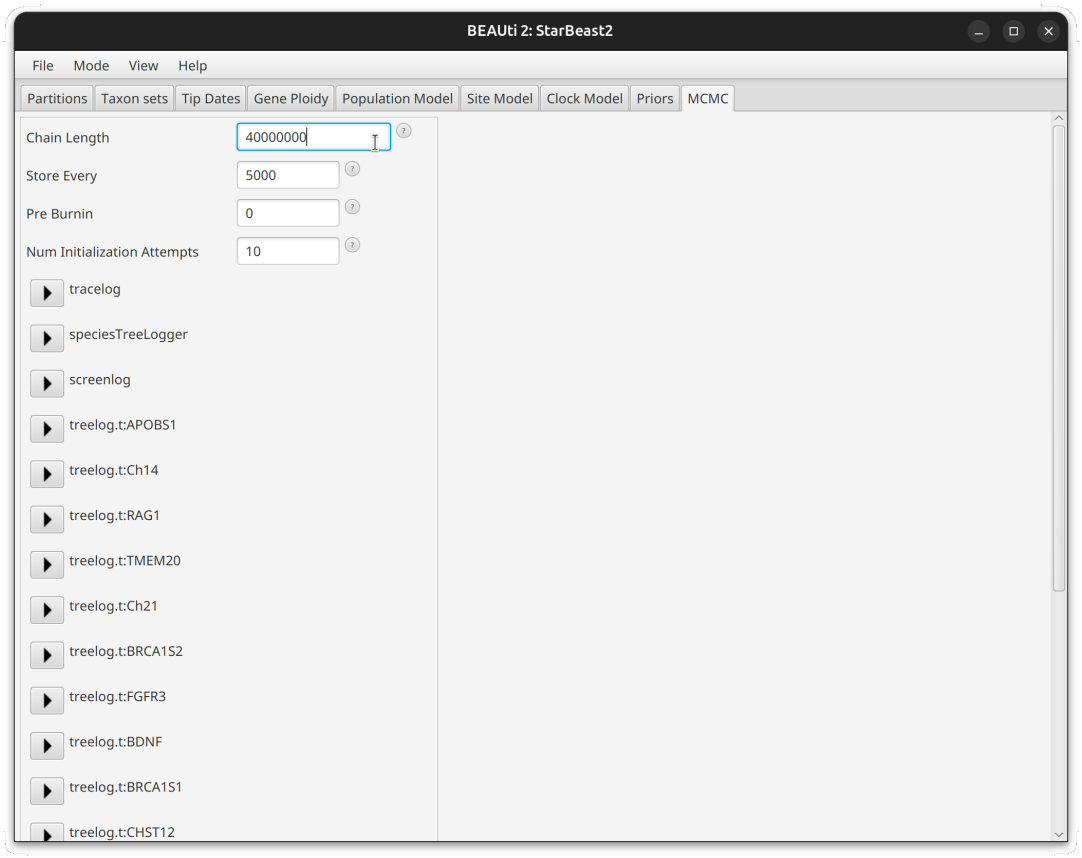

Figure 12: Setting a longer MCMC chain length of 40 million iterations

```
                University of California, Los Angeles
                          msuchard@ucla.edu

                     Downloads, Help & Resources:
                          http://beast2.org/

  Source code distributed under the GNU Lesser General Public License:
                       http://github.com/CompEvol/beast2

                              BEAST developers:
    Alex Alekseyenko, Trevor Bedford, Erik Bloomquist, Joseph Heled,
   Sebastian Hoehna, Denise Kuehnert, Philippe Lemey, Wai Lok Sibon Li,
Gerton Lunter, Sidney Markowitz, Vladimir Minin, Michael Defoin Platel,
          Oliver Pybus, Tim Vaughan, Chieh-Hsi Wu, Walter Xie

                                 Thanks to:
          Roald Forsberg, Beth Shapiro and Korbinian Strimmer

Random number seed: 1691867441007

File: Canis-Phylogeny.xml seed: 1691867441007 threads: 1
```

```
Loading package BEAST.app v2.7.5 SA v2.1.1 MM v1.2.1 ORC v1.1.2
starbeast2 v1.0.0 BEAST.base v2.7.5 BEASTLabs v2.0.1 starbeast3 v1.1.7
```

At the end of the output will be statistics describing the performance of all the MCMC operators. This will look something like below, except I needed to replace some of the output with ellipses so it would fit in this document:

```
Operator                                                          ...  Tuning    #accept    #reject
...
beast.base.inference.operator.UpDownOperator(clockUpDownOperator ... 0.80611      18039      52115
ScaleOperator(TreeScaler.t:Ch14)                                  ... 0.81774      18037      51748
ScaleOperator(TreeRootScaler.t:Ch14)                              ... 0.52148      14234      55718
Uniform(UniformOperator.t:Ch14)                                   ...       -     121376     228556
SubtreeSlide(SubtreeSlide.t:Ch14)                                 ... 0.65112     117291     232591
Exchange(Narrow.t:Ch14)                                           ...       -      21860     326864
Exchange(Wide.t:Ch14)                                             ...       -        804     347965
WilsonBalding(WilsonBalding.t:Ch14)                               ...       -       1394     347821
beast.base.inference.operator.kernel.BactrianDeltaExchangeOperat ... 0.39628      86822     284324
ScaleOperator(KappaScaler.s:APOBS1)                               ... 0.18949       6789      16652
ScaleOperator(KappaScaler.s:APOBS2)                               ... 0.21016       7722      15778
ScaleOperator(KappaScaler.s:BDNF)                                 ... 0.18970       7620      15658
ScaleOperator(KappaScaler.s:BRCA1S1)                              ... 0.18986       7245      15817
ScaleOperator(KappaScaler.s:BRCA1S2)                              ... 0.19130       6763      16591
ScaleOperator(KappaScaler.s:Ch14)                                 ... 0.30988       6979      16290
ScaleOperator(KappaScaler.s:Ch21)                                 ... 0.25278       6882      16419
ScaleOperator(KappaScaler.s:Ch24)                                 ... 0.23802       7303      16019
ScaleOperator(KappaScaler.s:CHST12)                               ... 0.18101       7118      16184
ScaleOperator(KappaScaler.s:CMKOR1)                               ... 0.18940       7070      16249
ScaleOperator(KappaScaler.s:FGFR3)                                ... 0.20670       7454      15770
ScaleOperator(KappaScaler.s:GHRex09)                              ... 0.19213       7583      15839
ScaleOperator(KappaScaler.s:RAG1)                                 ... 0.22845       7414      15823
ScaleOperator(KappaScaler.s:TMEM20)                               ... 0.21491       7438      15926
ScaleOperator(KappaScaler.s:VANGL2)                               ... 0.19990       7590      15886
ScaleOperator(KappaScaler.s:VWF)                                  ... 0.21476       7301      16098
starbeast2.NodeReheight2(Reheight.t:Species)                      ...       -      98775    1438750
starbeast2.CoordinatedUniform(coordinatedUniform.t:Species)       ...       -     354356     415005
starbeast2.CoordinatedExponential(coordinatedExponential.t:Speci ... 0.07335     453543     315251
beast.base.inference.operator.UpDownOperator(updownAll:Species)   ... 0.66322      74117     233719
starbeast2.RealCycle(constPopSizesSwap.Species)                   ... 2.00000      30178     123980
ScaleOperator(constPopSizesScale.Species)                         ... 0.28757      37835     115633
ScaleOperator(constPopMeanScale.Species)                          ... 0.44194      16611      34495
ScaleOperator(netDiversificationRateScale.t:Species)              ... 0.20702      13692      37872
ScaleOperator(ExtinctionFractionScale.t:Species)                  ... 0.11933       5801      19945
beast.base.inference.operator.UniformOperator(ExtinctionFraction ...       -      13197      12145
SubtreeSlide(bdSubtreeSlide.t:Species)                            ... 0.06648     236813     532106
WilsonBalding(bdWilsonBalding.t:Species)                          ...       -        495     770580
Exchange(bdWide.t:Species)                                        ...       -       4258     764403
```

```
Exchange(bdNarrow.t:Species)                          ...       -      38229     730568
Uniform(bdUniformOperator.t:Species)                  ...       -      59042     710796
ScaleOperator(bdTreeRootScaler.t:Species)             ... 0.87763       7185     147379
ScaleOperator(bdTreeScaler.t:Species)                 ... 0.98773      27551     126232

     Tuning: The value of the operator's tuning parameter, or '-' if the operator can't be optimiz
    #accept: The total number of times a proposal by this operator has been accepted.
    #reject: The total number of times a proposal by this operator has been rejected.
      Pr(m): The probability this operator is chosen in a step of the MCMC (i.e. the normalized we
  Pr(acc|m): The acceptance probability (#accept as a fraction of the total proposals for this ope


Total calculation time: 375.843 seconds
End likelihood: -16950.572066013454
Done!
```

3.9 Checking the log file

To check parameters other than species tree topologies and branch values, or to verify that the chain has been run long enough to reliably represent the posterior distribution, we will use Tracer. Start the Tracer app and then open the "starbeast.log" file. The first statistic that will be displayed is a histogram of the log posterior probability (Figure 13).

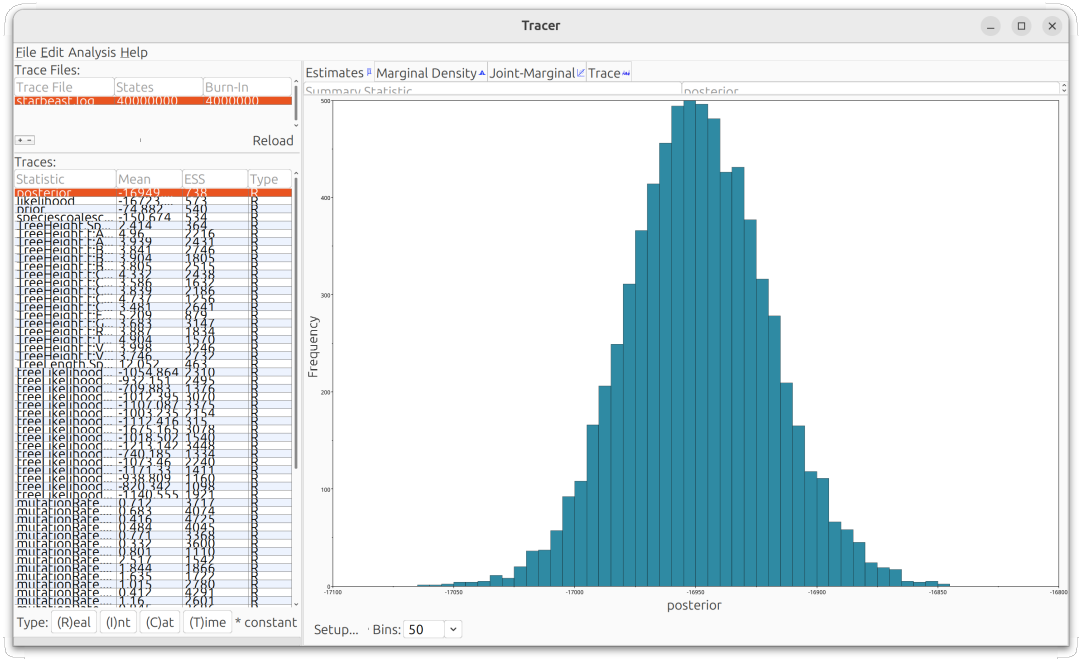

Figure 13: Opening a log file in Tracer

The posterior probability is the sum of the likelihood (which is the sum of log phylogenetic likelihoods for all sites for all loci), the prior probability (which is the sum of log prior probabilities for all parameters), and the speciescoalescent (which is the sum of log coalescent probabilities for all gene trees). TreeHeight.Species is the height of the root node of the species tree, and TreeLength.Species is the sum of all branch lengths

in the species tree.

Tracer computes effective sample sizes (ESS) for each logged statistic and parameter. As a rule, ESS values should be at least 200, particularly for the important statistics just noted and for parameters of interest. Because of the stochastic nature of MCMC algorithms, your values **will be different** to those in Figure 13.

3.10 Checking the species trees

Start the DensiTree app (included with BEAST2), and then open the "species.trees" file. Under the Show panel, enable the Root Canal tree to get an idea of the most plausible species tree. Then open the Grid panel, and enable the full grid (Figure 14). If you want to check the clade posterior probabilities, select "View clade toolbar" from the Window menu.

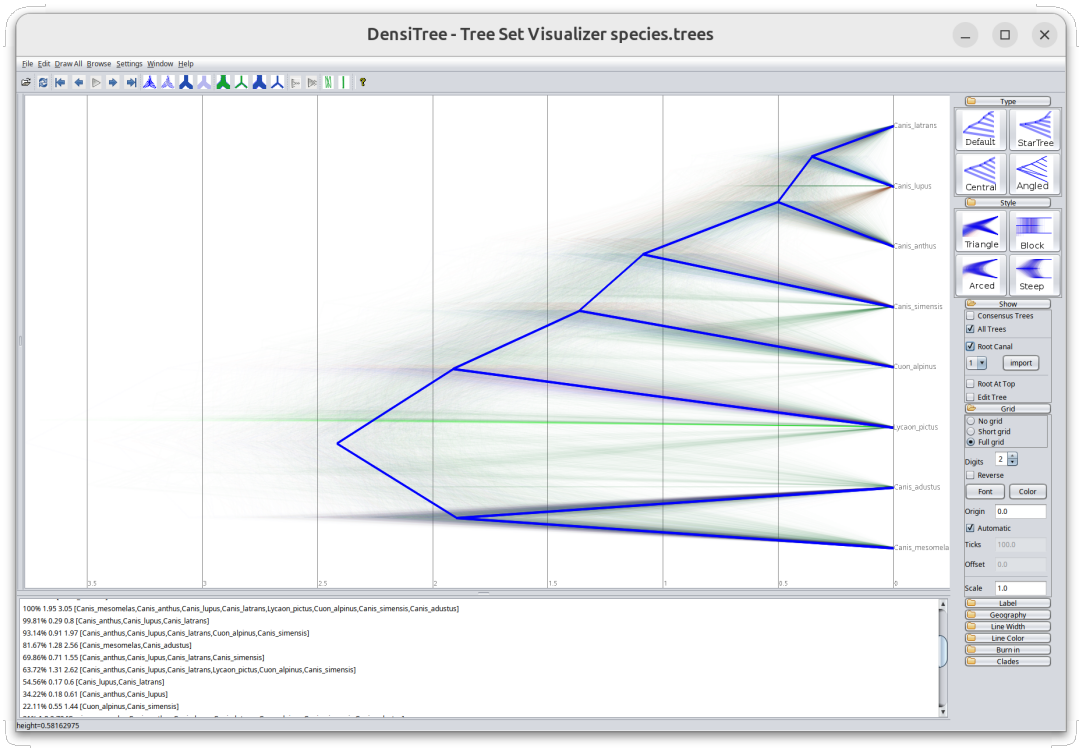

Figure 14: Viewing the species trees in DensiTree

You can see that using a fixed clock rate of 10^{-3} substitutions per site per year, the split between *Canis latrans* (coyotes) and *Canis lupus* (wolves) is probably less than 500,000 years ago. The age of the most recent common ancestor (MRCA) of extant *Canis* taxa is approximately 2.5 million years ago, more recent than the split between humans and chimpanzees (Prado-Martinez et al. 2013).

3.11 Generating a summary tree

Summary trees are a way of reducing a posterior distribution of tree topologies and times to a single tree. This is often more readable than the cloud of trees that DensiTree displays, but researchers should be cautious not to give it too much emphasis because for most clades and data sets there is substantial uncertainty in the posterior distribution of topologies and times.

To generate a summary tree, open the Tree Annotator app included with BEAST2. Set the burnin

percentage to 10 and choose the "species.trees" file generated by StarBEAST2 as the input file. Specify the output file to be something sensible like "summary.tree" and click Run to generate the summary tree (Figure 15).

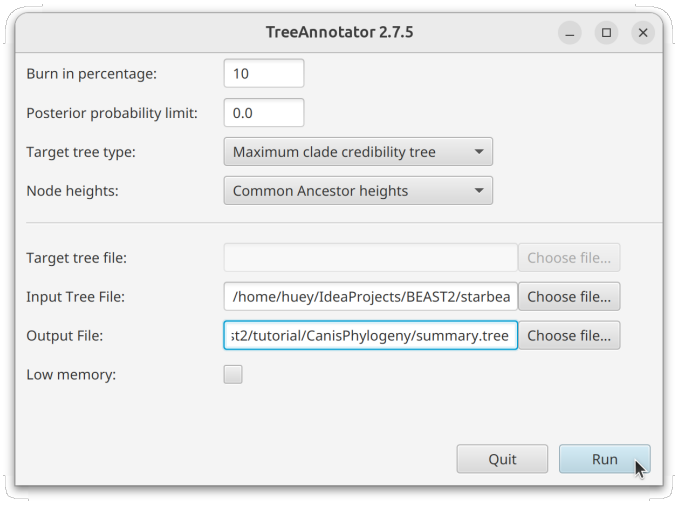

Figure 15: Running Tree Annotator

Load http://icytree.org/ in your web browser and open the summary tree file using the "Load from file" option in the File menu. In the Style menu, under "Node height error bars," choose "CAheight_95%_HPD" to display 95% highest posterior density (HPD) intervals of species divergence times on each branch. Also in the Style menu, under "Internal node text," choose "dmv1" to display estimates of ancestral effective population size. Enable the time axis by typing the letter "a" or by going through the Style menu again (Figure 16).

In StarBEAST2 and many other BEAST packages, effective population sizes are scaled by generation time. That means if generation times (the average number of years from zygote to zygote) vary between species in your analysis, the effective population sizes of different branches cannot be directly compared.

If the average generation time is 5 years, and our tree is scaled in millions of years, then the generation time is 5×10^{-6}. To get effective population sizes in numbers of individuals, divide each value by the generation time. So if the scaled effective population size is 1, that will correspond to 200,000 individuals.

4 Estimating per-species clock rates

We can estimate molecular clock rates separately for each species using a relaxed clock model, for example the uncorrelated log-normal (UCLN) model. First create a new folder for this analysis with a sensible name, something like "CanisUCLN". Relaunch BEAUti, and select "SpeciesTreeUCLN" from the Templates submenu of the File menu (Figure 17).

Import the same FASTA files as in subsection 3.1 (Figure 4). For any of the species tree relaxed clock templates, do **NOT** link the clock models. This will break the inference of per-species clock rates.

Assign the same species names as in subsection 3.3 (Figure 6, 7). This time we will keep the default setting for the population model (Analytical Integration), so that our analysis will run slightly faster. Set all the site models to HKY with empirical frequencies as in subsection 3.5 (Figure 9, 10).

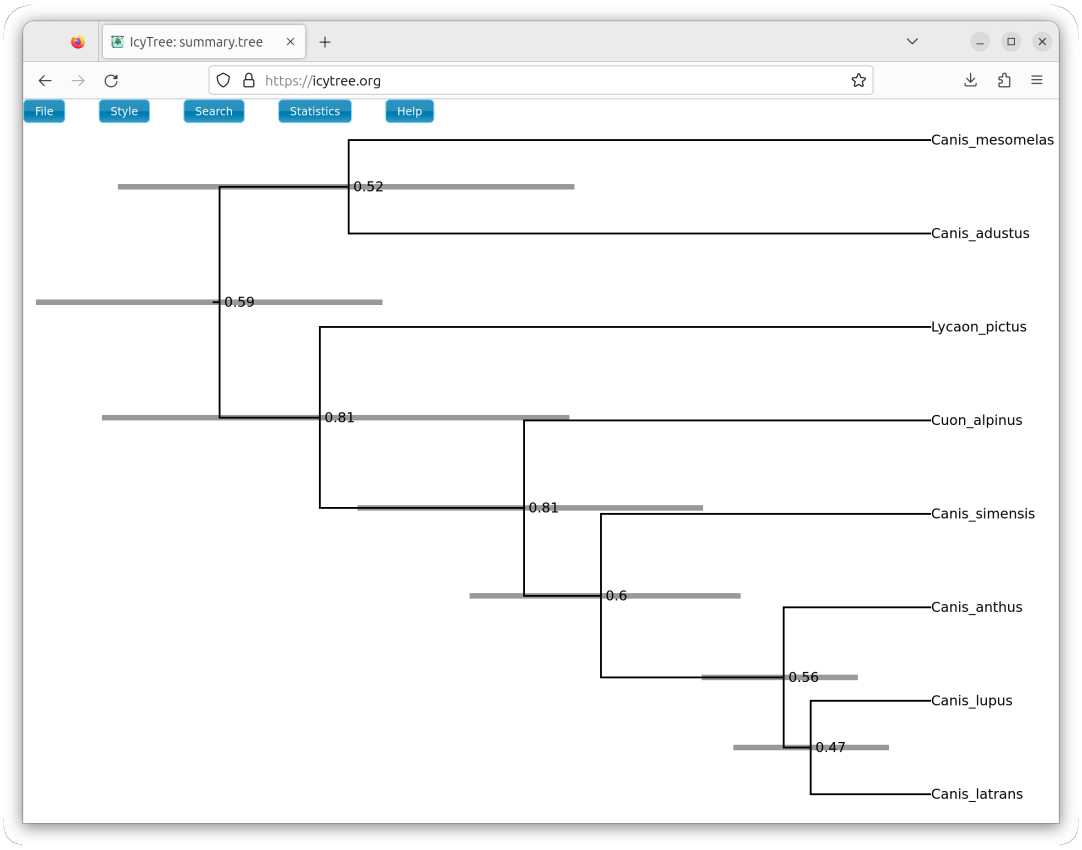

Figure 16: Running IcyTree. This tree been made more legible by increasing font size and limiting numbers to two significant figures, further options under IcyTree's "Style" menu.

Open the Clock Model tab, and manually set the Clock.rate to 0.001 (or equivalently 1e-3) for **every** locus (Figure 18). This is necessary because for technical reasons we cannot link the clock models for the species tree relaxed clock templates.

Open the Priors tab and change the species tree prior from Yule to Birth-Death to allow for extinction. Finally, open the MCMC tab and change the length of the chain to 40 million, as in subsection 3.7 (Figure 12). Save the file in the folder you created previously for this analysis, and give it a sensible name like "Canis-UCLN.xml".

Run the MCMC chain by opening the XML file in BEAST, or running it from the command line as in subsection 3.8. This will take about twice as long (10 to 20 minutes) as the fixed clock analysis, because relaxed clocks are more computationally intensive.

After BEAST has finished, open the "starbeast.log" file in Tracer to check that the important statistics have ESS values of at least 200. Select the branchRatesStdev.Species parameter (Figure 19).

This parameter models the spread of molecular clock rates among branches in the species tree. The default prior for this parameter in StarBEAST2 is a log-normal distribution with a mean of 1, but the posterior distribution of this parameter has a mean of about 0.2. This means that the data is pulling this parameter lower, probably because there is very little variation in clock rates between species. Indeed the mode of the posterior distribution is about zero, suggesting that a strict clock may be more appropriate (Figure 19).

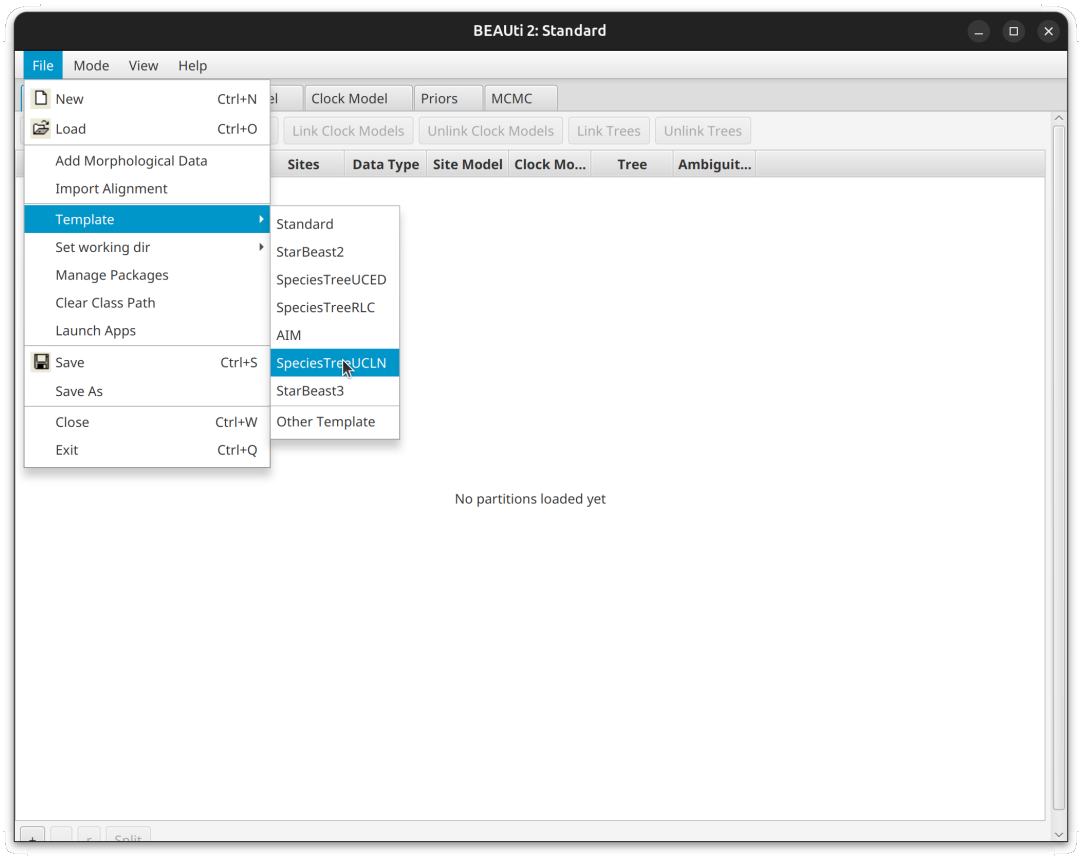

Figure 17: Starting a UCLN relaxed clock analysis

5 Total evidence tip-dating

StarBEAST is able to combine tip-dating using the fossilized birth death (FBD) process with multispecies coalescent (MSC) inference of species and gene trees. We call this integrative model "FBD-MSC". To begin your analysis, creat a new folder for this purpose with a sensible name, something like "CanisFBD".

When setting up an FBD-MSC analysis in BEAUti, it is necessary to import morphological data after specifying the taxon sets and the tree model, but before specifying the tip dates. The following steps will be consistent with that ordering. First, repeat section 3 from subsection 3.1 to and including 3.3. Leave the population model set at the default (Analytical Integration) to save time, then set all site models to HKY with empirical frequencies as in subsection 3.5.

Before adding the morphological data, go to the Priors panel and change the prior on the species tree from Yule to FBD (Figure 20). This is necessary because the inclusion of fossil taxa means our species tree will be *serially sampled* rather than *ultrametric*. StarBEAST2 gives you the option of various birth-death models – Yule, Calibrated Yule, Birth Death and FBD – but only FBD is valid for serially sampled trees.

Now go back to the Partitions tab, and from the File menu select "Add Morphology to Species Tree" (Figure 21). Navigate to the data folder of the tutorial and import the "morphology_canis.nex" file.[1]

[1]This file contains recorded states for 50 morphological characters of *Canis* species and some related species, taken from Slater 2015.

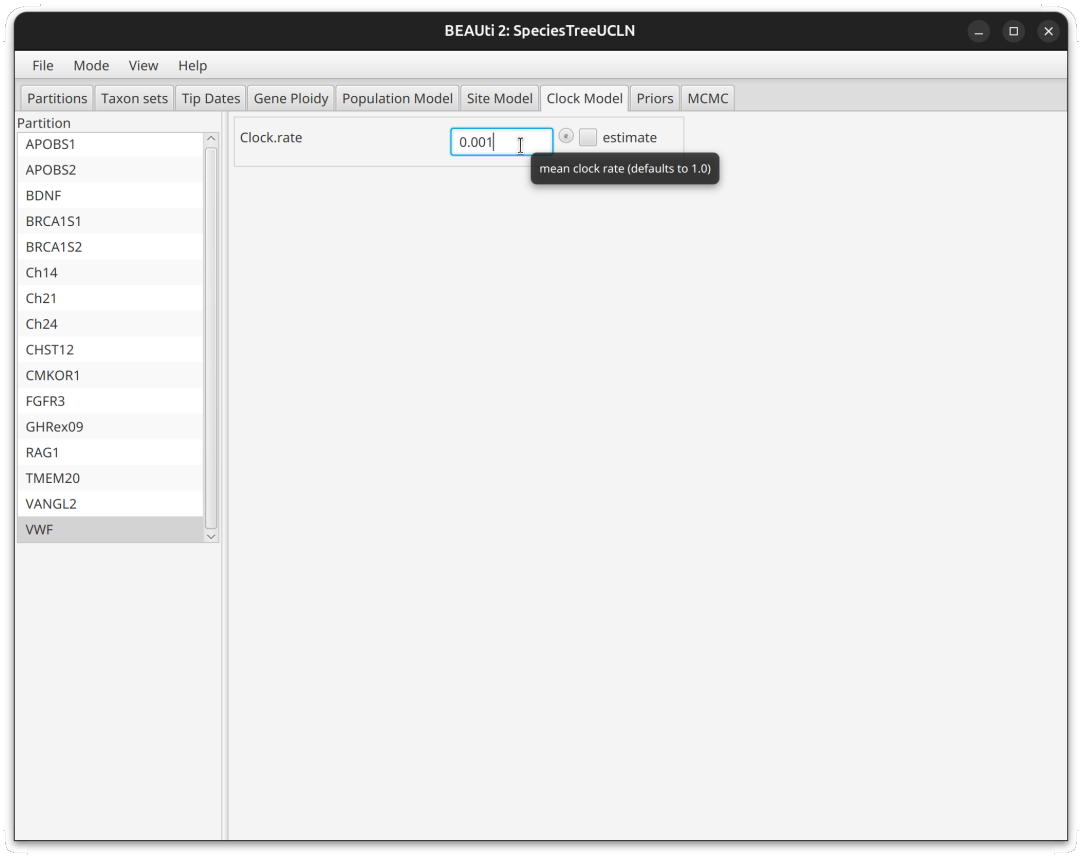

Figure 18: Manually setting every clock rate

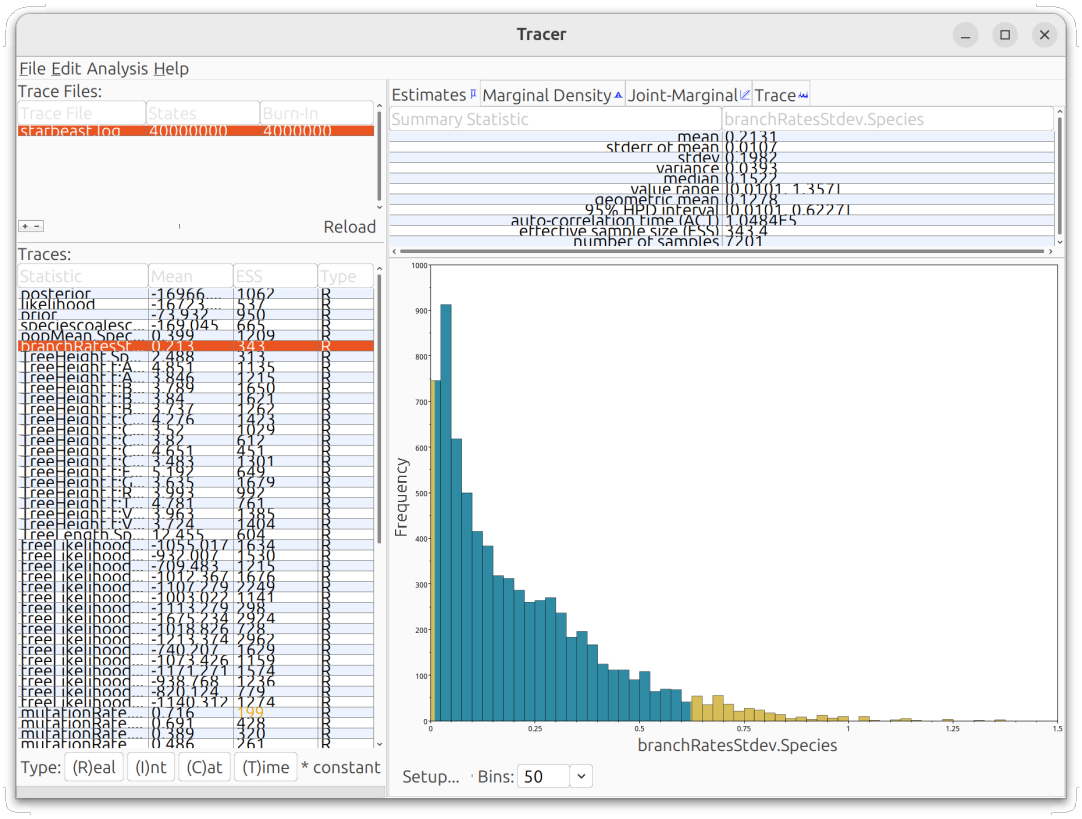

Figure 19: Checking the species tree relaxed clock analysis in Tracer

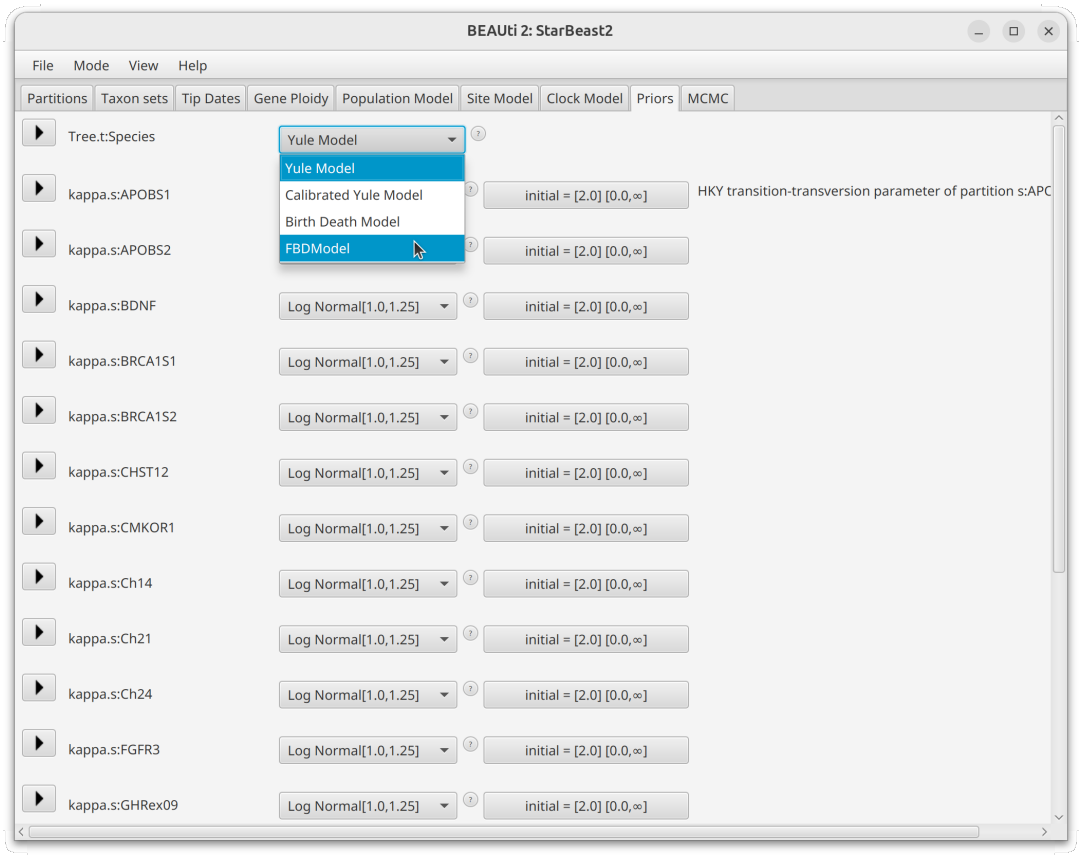

Figure 20: Change the tree model from Yule to FBD.

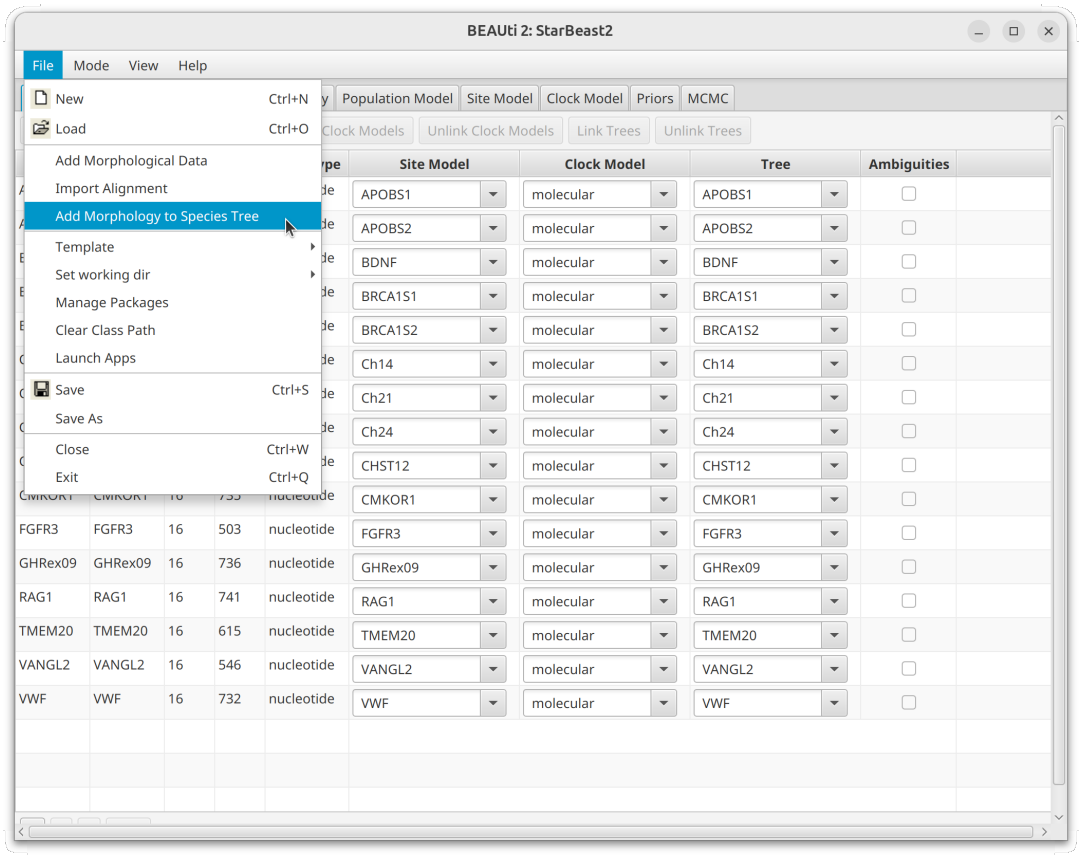

Figure 21: The option to add morphological data to the species tree

Select Yes to condition on recording variable characters only (Mkv). Mkv is used when only characters that vary between species have been included in the character matrix, as is the case for this tutorial's data set. You should now see four morphology partitions (Figure 22), one for each number of character states. For example, morphology_canis2 is for binary characters.

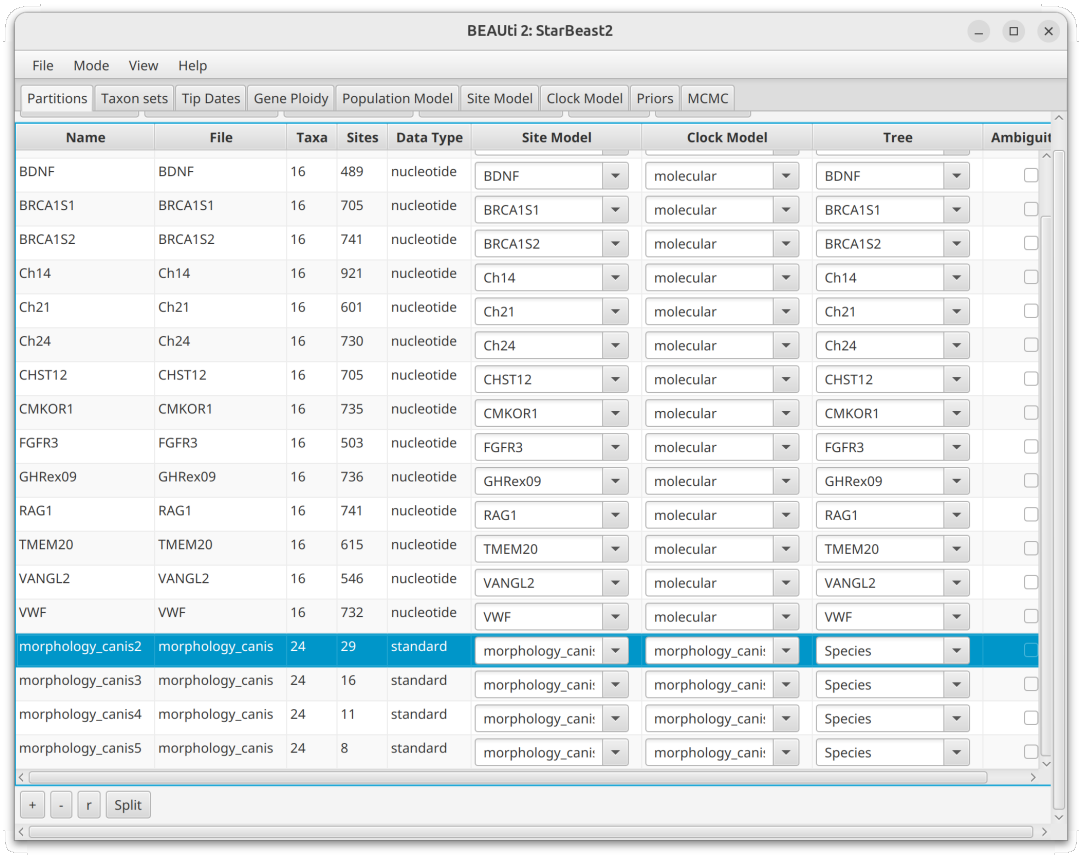

Figure 22: After morphological data has been added to the species tree

Open the Tip Dates panel and enable "Use tip dates". Change "Since some time in the past" to "before the present". Click on Auto-configure and select "read from file". Choose the "tip_dates_canis.dat" file in the data folder of the tutorial. Click OK, and your tip dates should look like Figure 23.

Go to the Clock Model panel, and enable "estimate" for the molecular clock (by default called APOBS1 after the first locus, although this can be changed in the Partitions tab). Then select the morphological data partition and enable "estimate" so that the morphological clock rate will also be estimated (Figure 24). Make sure you have enabled "estimate" for **both** clocks, since we are using fossil data to calibrate our tree.

Next, go back to the Priors panel, scroll down to the strictClockRate.c:molecular parameter, and expand that prior. Change the mean to 0.001, the *a priori* molecular clock rate estimate from section 3. Leave the standard deviation set at 1, a moderately informative prior that will still allow the rate to be guided by the fossil data. Then change the prior distribution for the strictClockRate.c:morphology_canis prior to 1/X (Figure 25). This is an improper, uninformative prior so the inferred morphological clock rate will be estimated solely from the data.

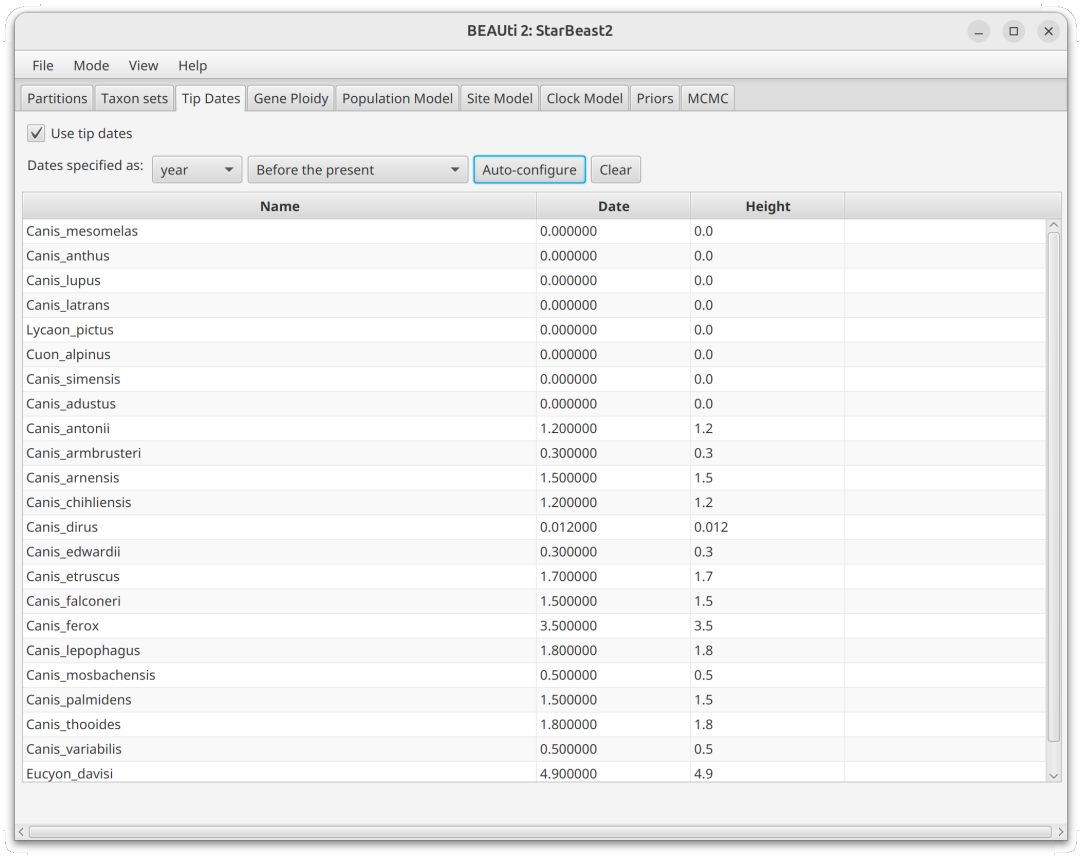

Figure 23: Configured tip dates

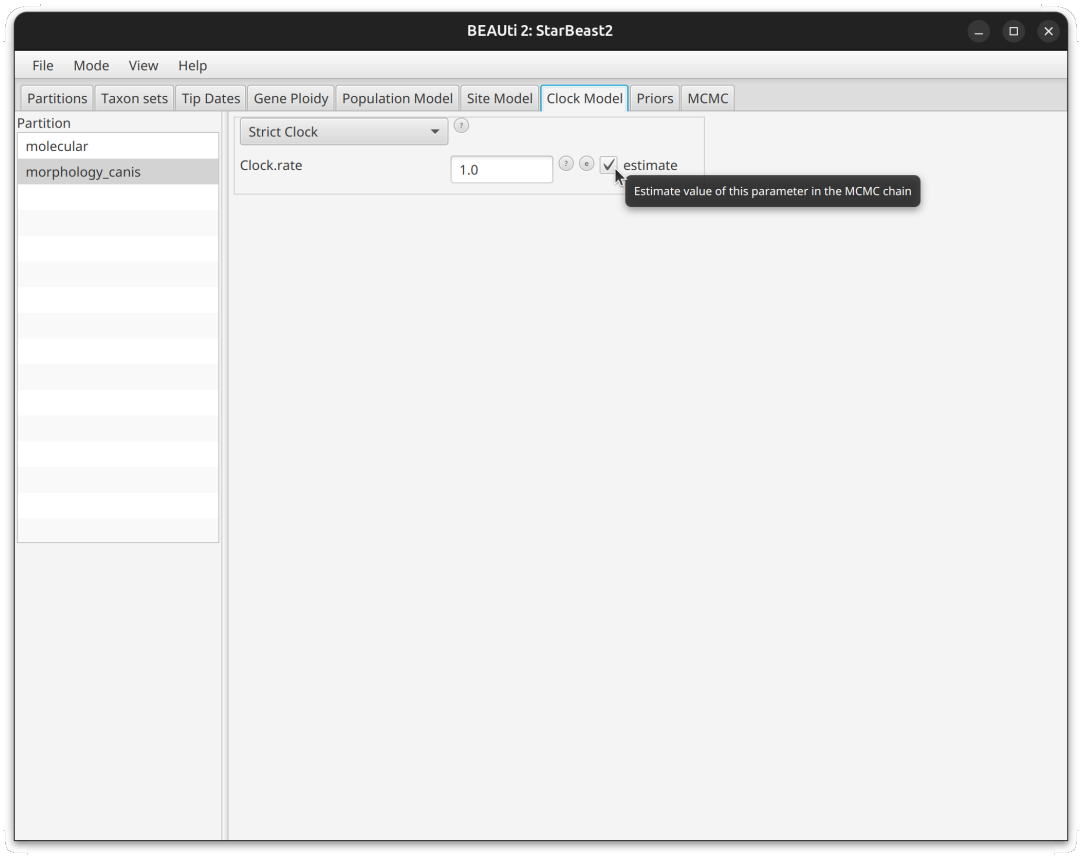

Figure 24: Estimating the clock rates

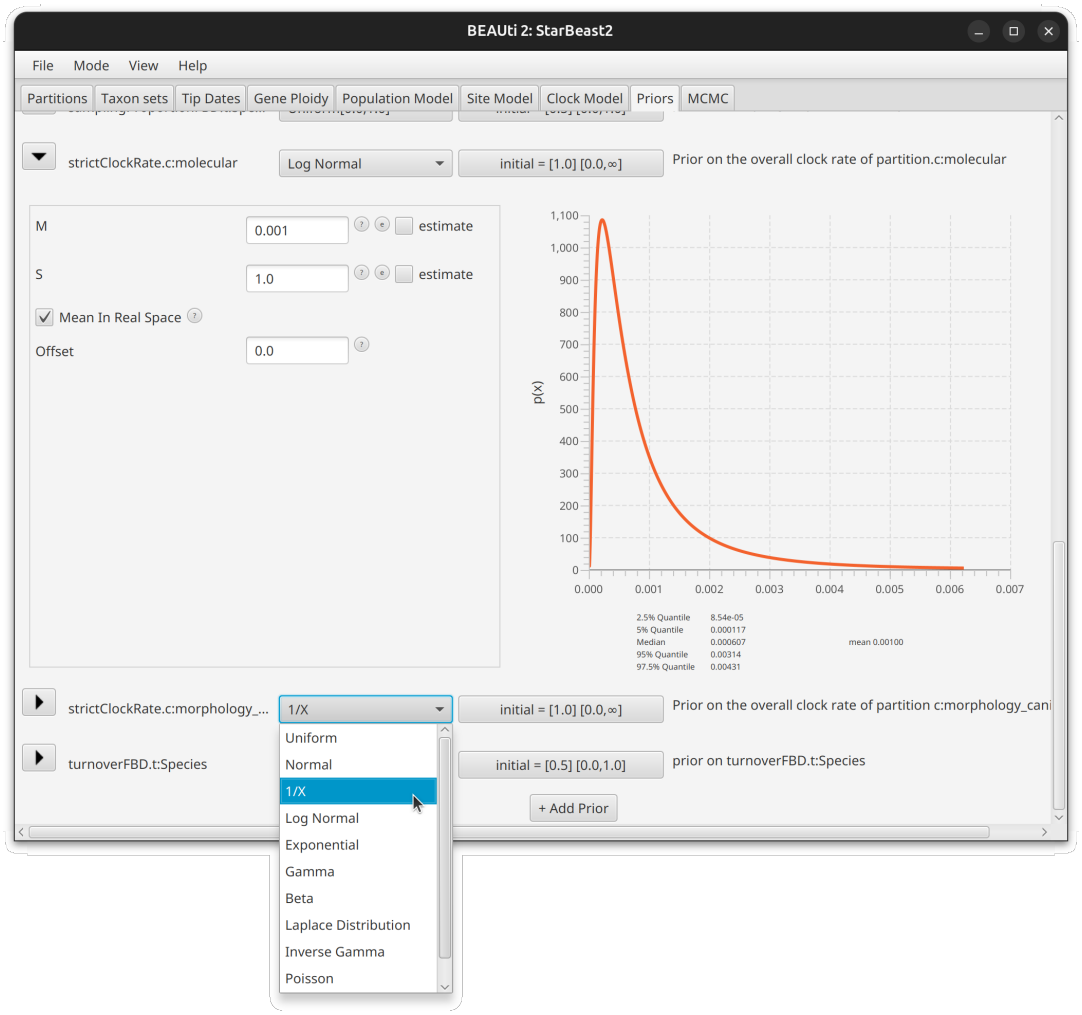

Figure 25: Configuring the clock prior distributions

The FBD-MSC model is quite complex and requires very long chains for reliable sampling. Open the MCMC tab and change the chain length to 200 million. We will need to reduce the sampling rate so that the output files remain managably small, so change the Store Every value to 50,000. Expand the tracelog, speciesTreeLogger, screenlog and every gene tree log, and set the Log Every value to 50,000 for each of them. This will limit the number of samples to 4,000 in each log file (Figure 26).

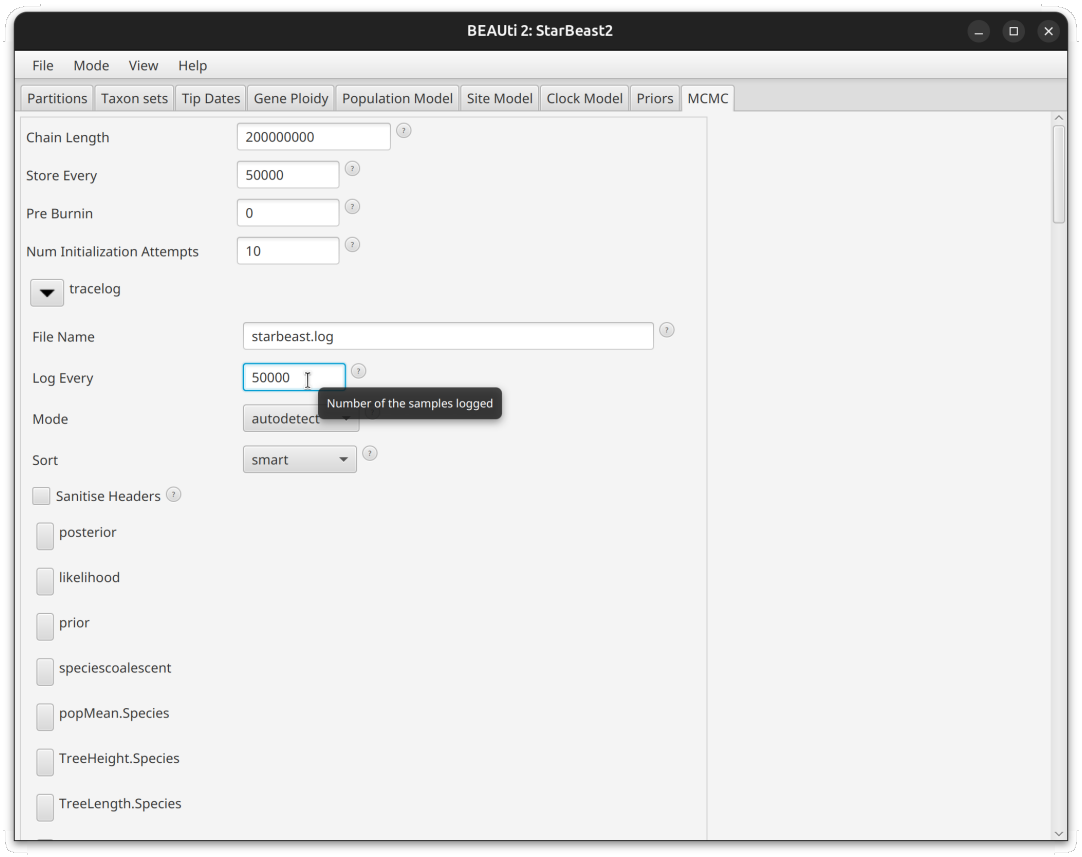

Figure 26: Changing the sampling rates after settting a chain length of 200 million

Save the XML file in the folder you created for this analysis, and name it something like "CanisFBD.xml". Run the MCMC chain by opening the XML file in BEAST, or running it from the command line as in subsection 3.8. The chain should take between 1 and 2 hours to finish, depending on the speed of your computer.

Editing BEAST XML files

BEAUti is a wonderful tool, but sometimes it can be faster to edit XML files directly rather than using BEAUti to configure every aspect of your model. Setting the sampling rate for gene tree log files one at a time is a pain, especially if you have 100 or more loci in your analysis!
Instead of changing the sampling rates in BEAUti, you can leave them at their default values, and use a text editor to change them all at once. After opening your XML file in the editor, find-and-replace all occurences of the sampling rate attribute and its default value, e.g. `logEvery="5000"`, with the value changed to the one desired, e.g. `logEvery="50000"`.
You may even want to design a model in a way that is impossible to configure with BEAUti, in which case you *must* edit the XML file directly. However, that sort of thing is beyond the scope of this tutorial.

After BEAST has finished, open the "starbeast.log" file in Tracer to check on the ESS values. These values are likely insufficient given a chain length of only 200 million for such a complex model, and a real analysis might require around 2 billion states, possibly spread across multiple chains.

Select the strictClockRate.c:molecular parameter (Figure 27). The 95% HPD interval for this parameter spans approximately 0.0005 to 0.0013 substitutions per site per million years, in agreement with the *a priori* rate of 0.001.

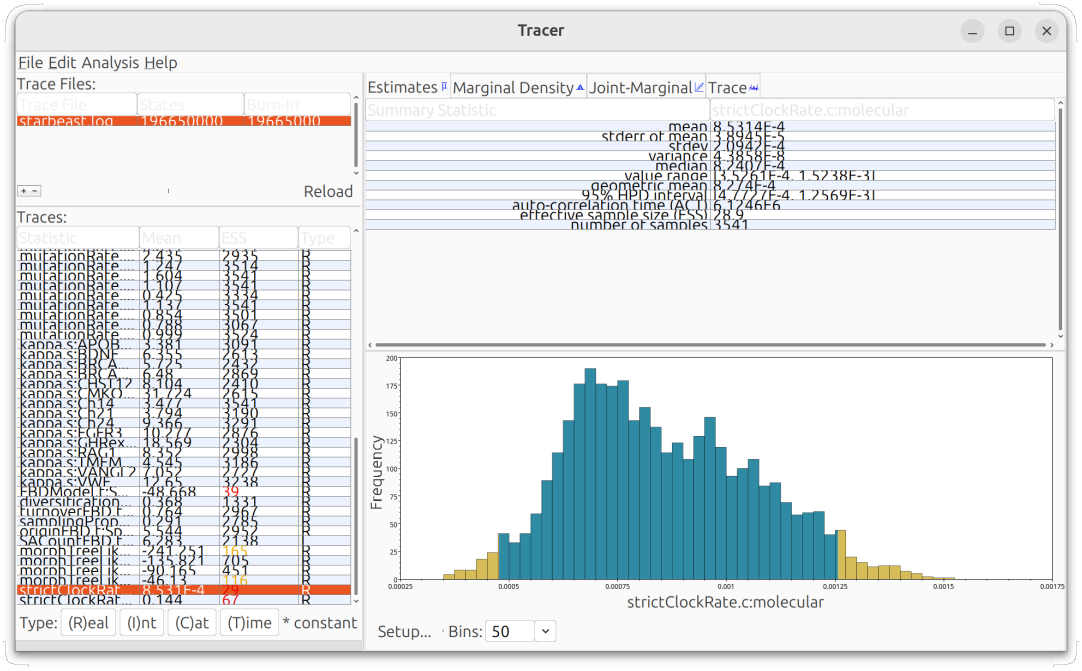

Figure 27: The estimated posterior distribution of the molecular clock rate

Now open the "Trace" tab to see how this value wandered around over the course of the MCMC chain. With only 200 million states this is unlikely to look like the ideal "hairy catepillar," and will instead exhibit slow transitions between modes of the distribution (Figure 28).

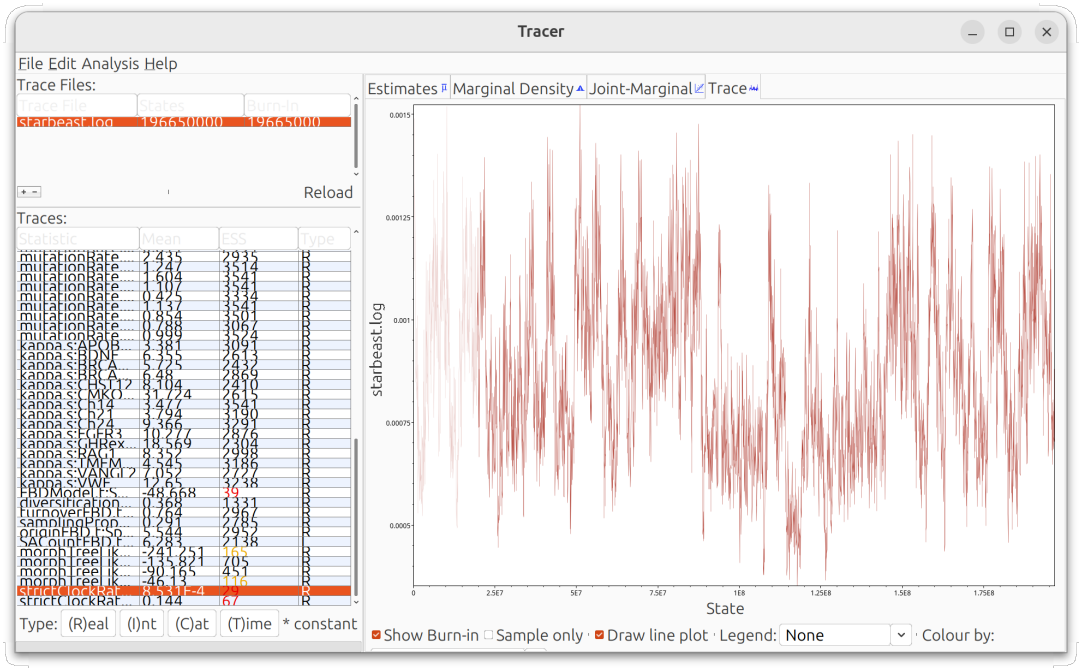

Figure 28: The MCMC trace of the molecular clock rate

Start the DensiTree app again, and then open the "species.trees" file. Under the Show panel, enable the Root Canal tree to get an idea of the most plausible species tree. Then open the Grid panel, and enable the full grid. Compared with the fixed clock analysis (Figure 14), the divergence time between *Canis lupus* and *Canis latrans* is a little older, and the age of the MRCA of extant species is a little younger (Figure 29).

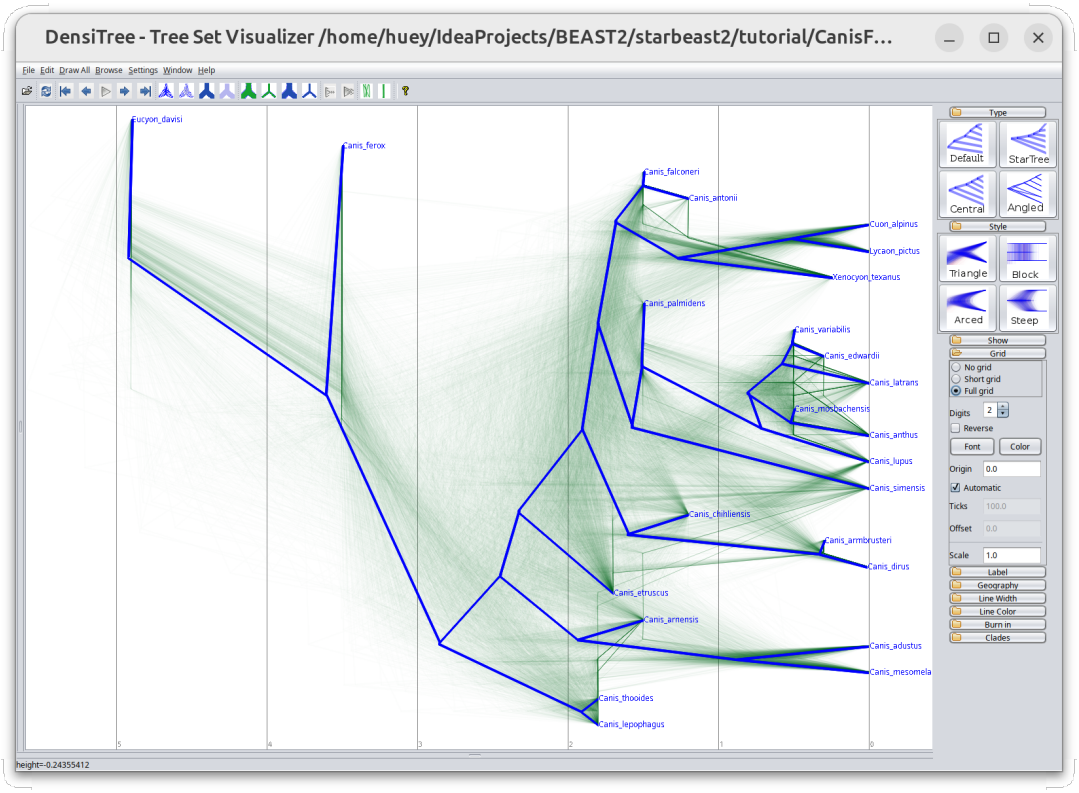

Figure 29: Cloud of FBD-MSC species trees

Open the Tree Annotator app included with BEAST2, and set the burnin percentage to 10. Trees inferred by the FBD-MSC model as implemented in StarBEAST2 can include sampled ancestors (Gavryushkina et al. 2014), which are incompatible with Common Ancestor heights (Heled and Bouckaert 2013) as implemented in Tree Annotator. So in order to generate a summary tree, change "Node heights" to "Mean heights" (Figure 30). Choose the "species.trees" file as the input file, and specify the output file to be something sensible like "summary.tree", then click Run.

Go back to IcyTree (`http://icytree.org/`) in your web browser and open the summary tree file using the "Load from file" option in the File menu. In the Style menu, under "Node height error bars," choose "height_95%_HPD" to display 95% highest posterior density (HPD) intervals of species divergence times on each branch. This time show the posterior support of each clade by choosing "posterior" in "Internal node text" from the Style menu. Enable the time axis again through the Style menu or by typing the letter "a" (Figure 31).

The tree topology is similar to the fixed clock analysis (Figure 16), except that *Cuon alpinus* and *Lycaon pictus* are now a clade with 100% support. A disagreement between molecular data which contradicts this clade and morphological data which supports it has been reported previously (Zrzavý and Řičánková 2004). This shows how morphological characters and fossil data can influence species tree estimates.

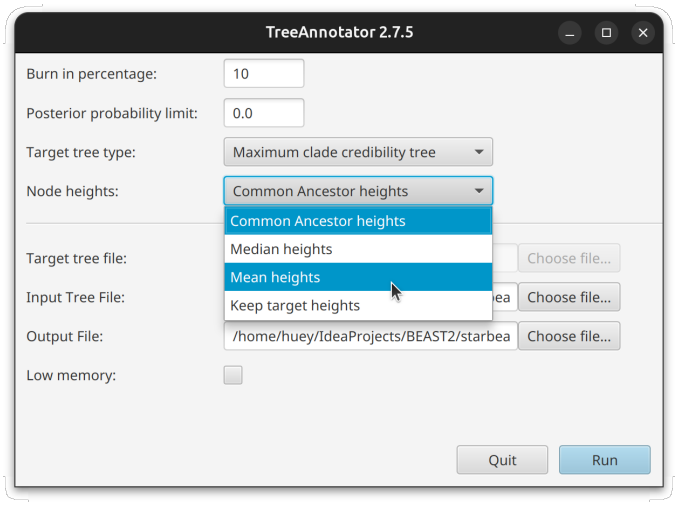

Figure 30: Running Tree Annotator for sampled ancestor trees

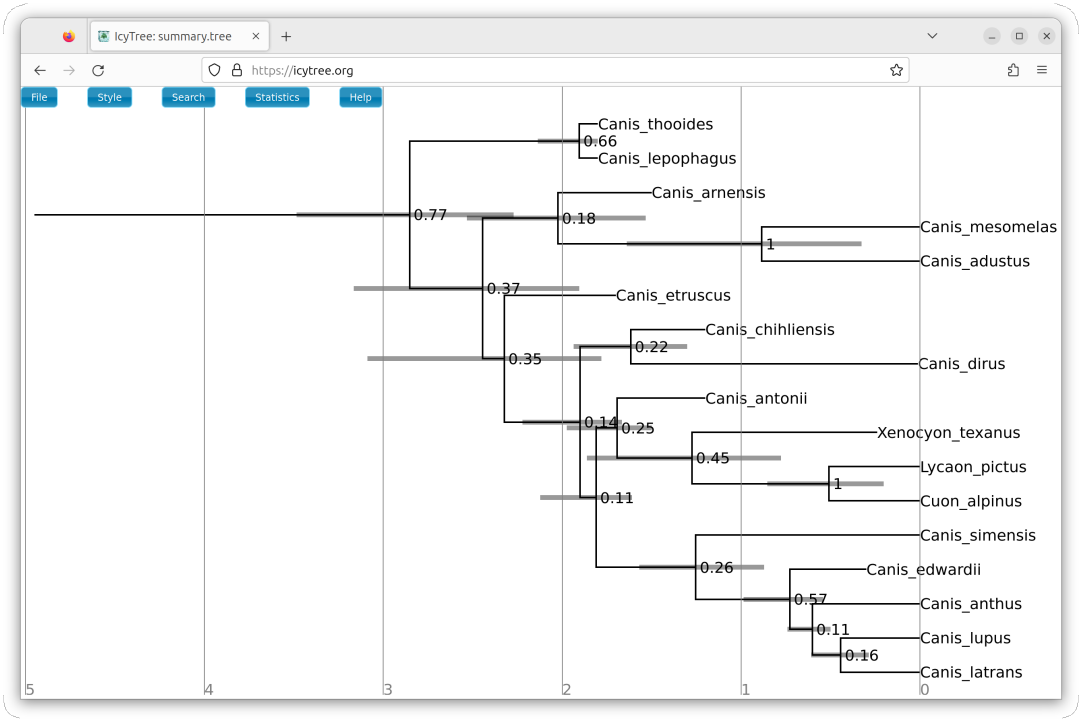

Figure 31: Showing posterior support for clades in IcyTree

6 Wrapping up

That concludes the tutorial. One final note; when runnning any MCMC method, it is possible than the chain has gotten stuck in a region of parameter space with high local probability. So for any kind of analysis intended for publication, it is a good idea to run at least two MCMC chains. Open both chains in Tracer, and check that the traces for the logged statistics and parameters have the same distributions.

Version dated: August 15, 2023

Relevant References

Bouckaert, RR. 2010. Densitree: making sense of sets of phylogenetic trees. *Bioinformatics* 26: 1372–1373.

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.

Hasegawa, M, H Kishino, and T Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22: 160–174.

Heled, J and RR Bouckaert. 2013. Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology* 13: 221.

Hugall, AF, R Foster, MSY Lee, and M Hedin. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Systematic Biology* 56: 543–563.

Lindblad-Toh, K, CM Wade, TS Mikkelsen, EK Karlsson, DB Jaffe, M Kamal, M Clamp, JL Chang, EJ Kulbokas, MC Zody, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438: 803–819.

Ogilvie, HA, RR Bouckaert, and AJ Drummond. 2017. Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution* 34: 2101–2114.

Ogilvie, HA, J Heled, D Xie, and AJ Drummond. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. *Systematic Biology* 65: 381–396.

Prado-Martinez, J et al. 2013. Great ape genetic diversity and population history. *Nature* 499: 471–475.

Scornavacca, C and N Galtier. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology* 66: 112–120.

Slater, GJ. 2015. Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. *Proceedings of the National Academy of Sciences* 112: 4897–4902.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39: 306–314.

Zrzavý, J and V Řičánková. 2004. Phylogeny of recent Canidae (Mammalia, Carnivora): relative reliability and utility of morphological and molecular datasets. *Zoologica Scripta* 33: 311–333.