Calibrated Species Trees

Estimating species divergence times using StarBEAST2

Huw A. Ogilvie

1 Background

This tutorial assumes you have completed the first StarBEAST2 tutorial, “Species Trees with Relaxed Molecular Clocks.” In that tutorial, the data suggested that zero shifts in substitution rates have occured within the pocket gopher genus *Thomomys*. No priors on node heights were used in the first tutorial, and the expected clock rate of each gene was 1.0, so the estimated speciation times were all in expected number of substituions.

We will improve on our previous analysis by calibrating the tree, so that we can estimate how many millions of years ago (ma) different *Thomomys* lineages diverged.

2 BEAUti

Run BEAUti by double clicking on its icon, or by launching the BEAUTi executable file from the command line in Linux.

Beginning a strict clock analysis

Building on our previous results, we will use a strict clock for this tutorial. Compared to a relaxed clock, this will take less computational time and produce narrower credible ranges of divergence times. To set up a StarBEAST2 analysis using a strict clock, choose the **File/Template/StarBEAST2** menu item (Figure 1).

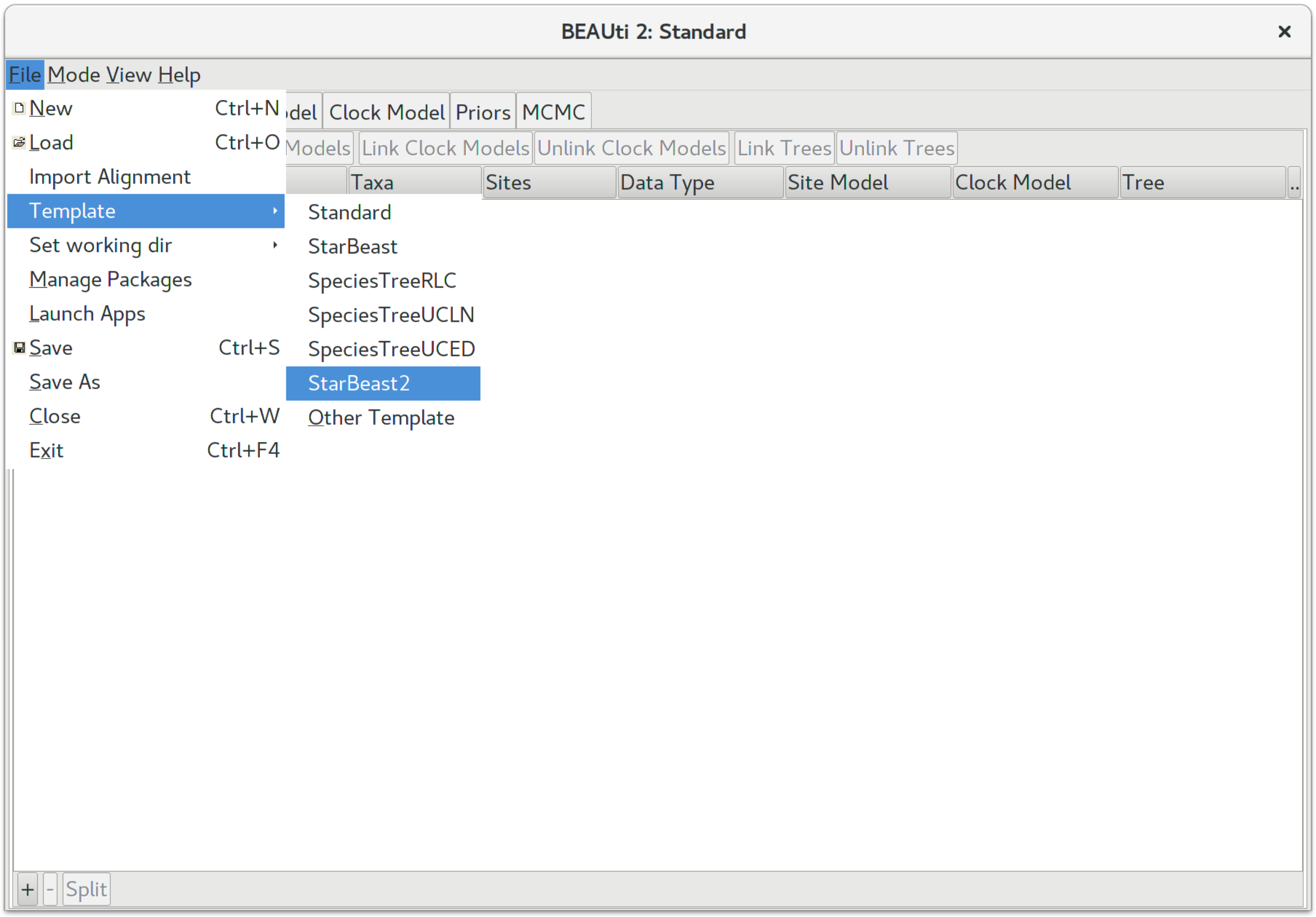


Figure 1: Select a species tree template in BEAUti.

Allow clock rates to vary

By default BEAUTi fixes the clock rate of the first partition to 1, so that the rates of other loci are estimated relative to the first locus. This is generally inappropriate for StarBEAST2 analyses, so it is very important to disable this behaviour by deselecting the **Mode/Automatic set clock rate** menu item (Figure 2).

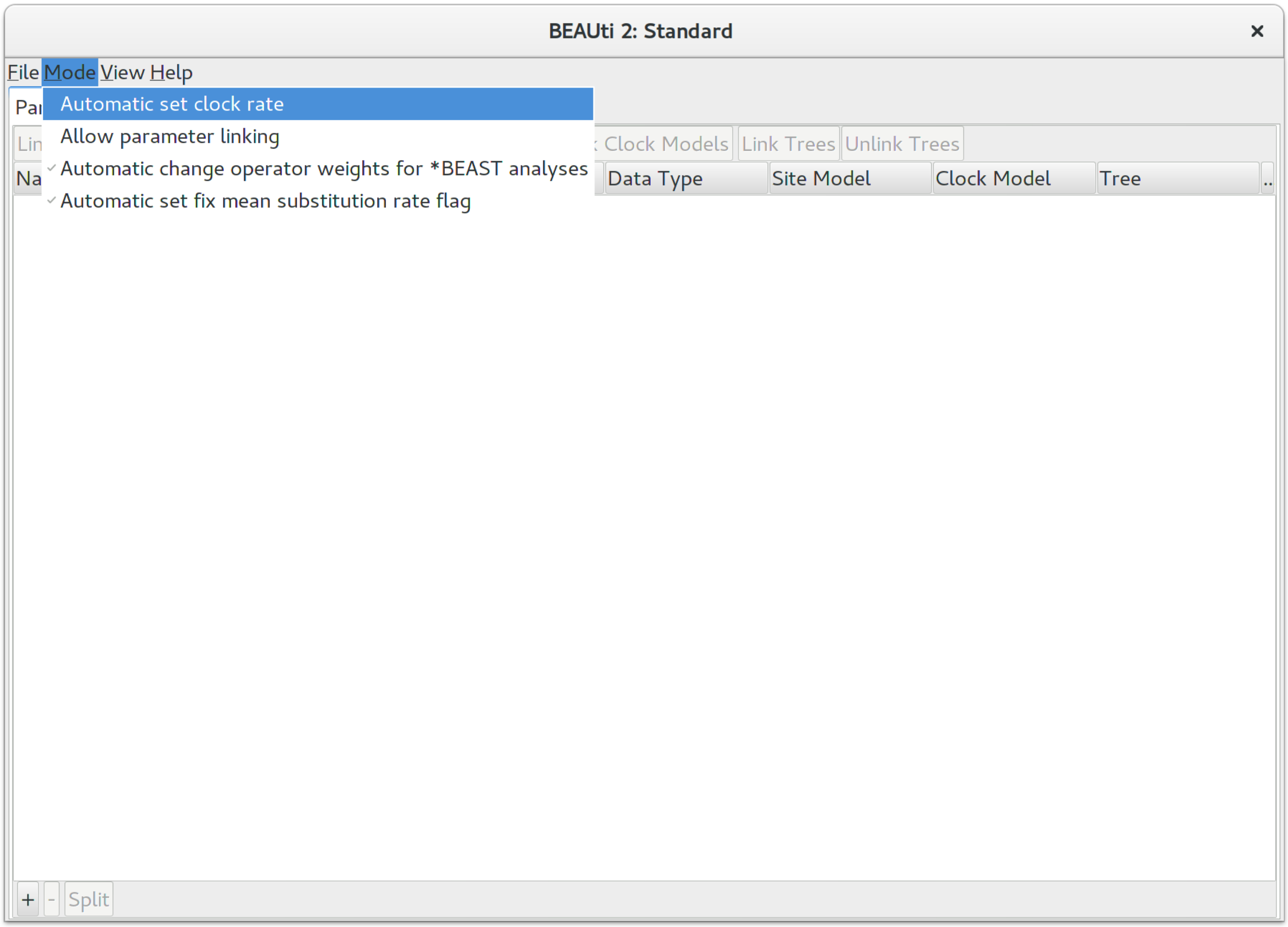
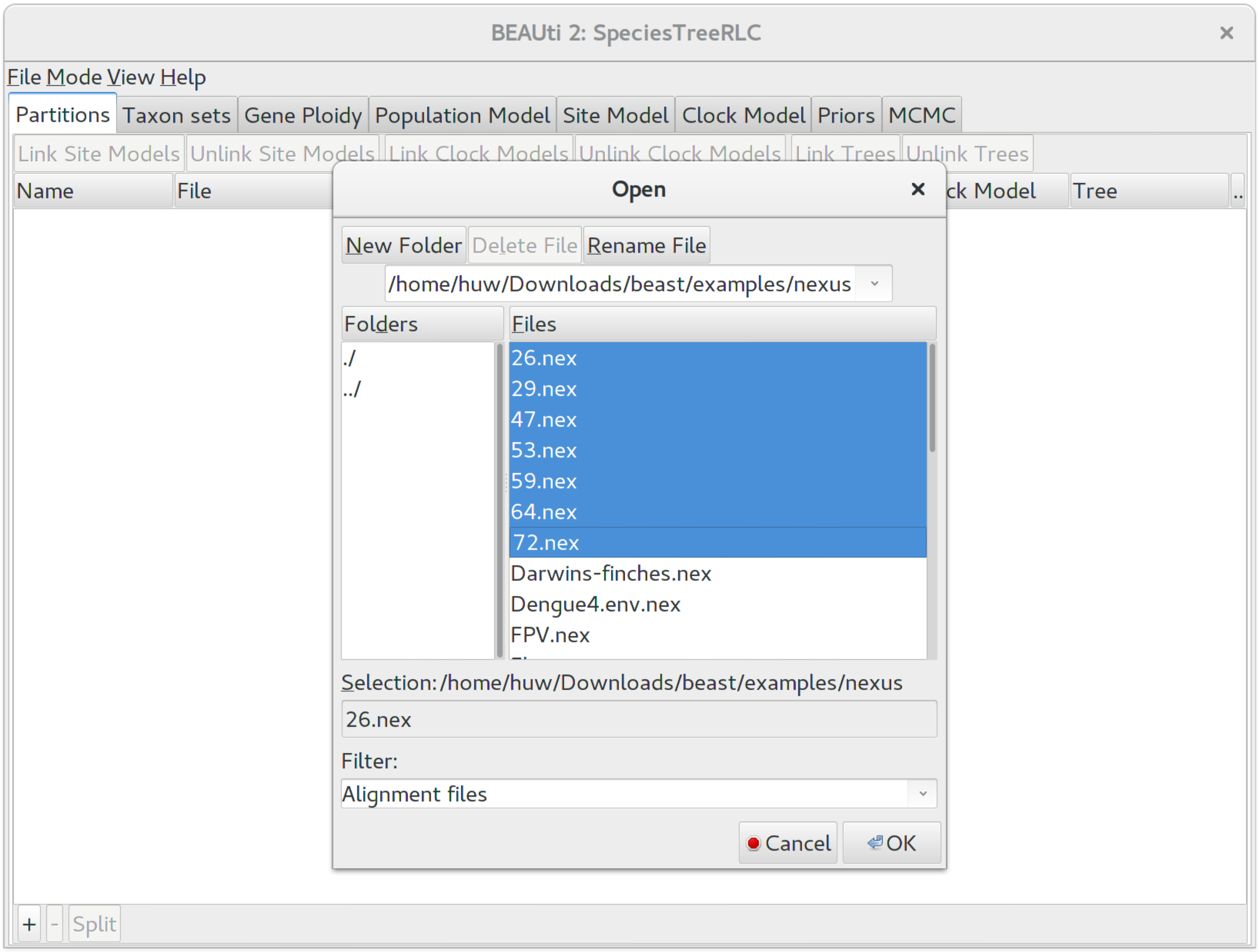


Figure 2: Disable automatic setting of clock rates.

Loading the NEXUS files

To load a NEXUS format alignment, click the button with the plus symbol (+) in the lower left corner of the main **Partitions** tab. We will be using the same sequence files as in the previous tutorial; navigate to the **examples/nexus** subfolder inside the **beast** application folder, and select all of the first seven NEXUS files. They should be numbered 26, 29, 47, 53, 59, 64, and 72 (Figure 3).



/home/huw/Downloads/beast/examples/nexus

26.nex
29.nex
47.nex
53.nex
59.nex
64.nex
72.nex
Darwins-finches.nex
Dengue4.env.nex
FPV.nex

Selection: /home/huw/Downloads/beast/examples/nexus

26.nex

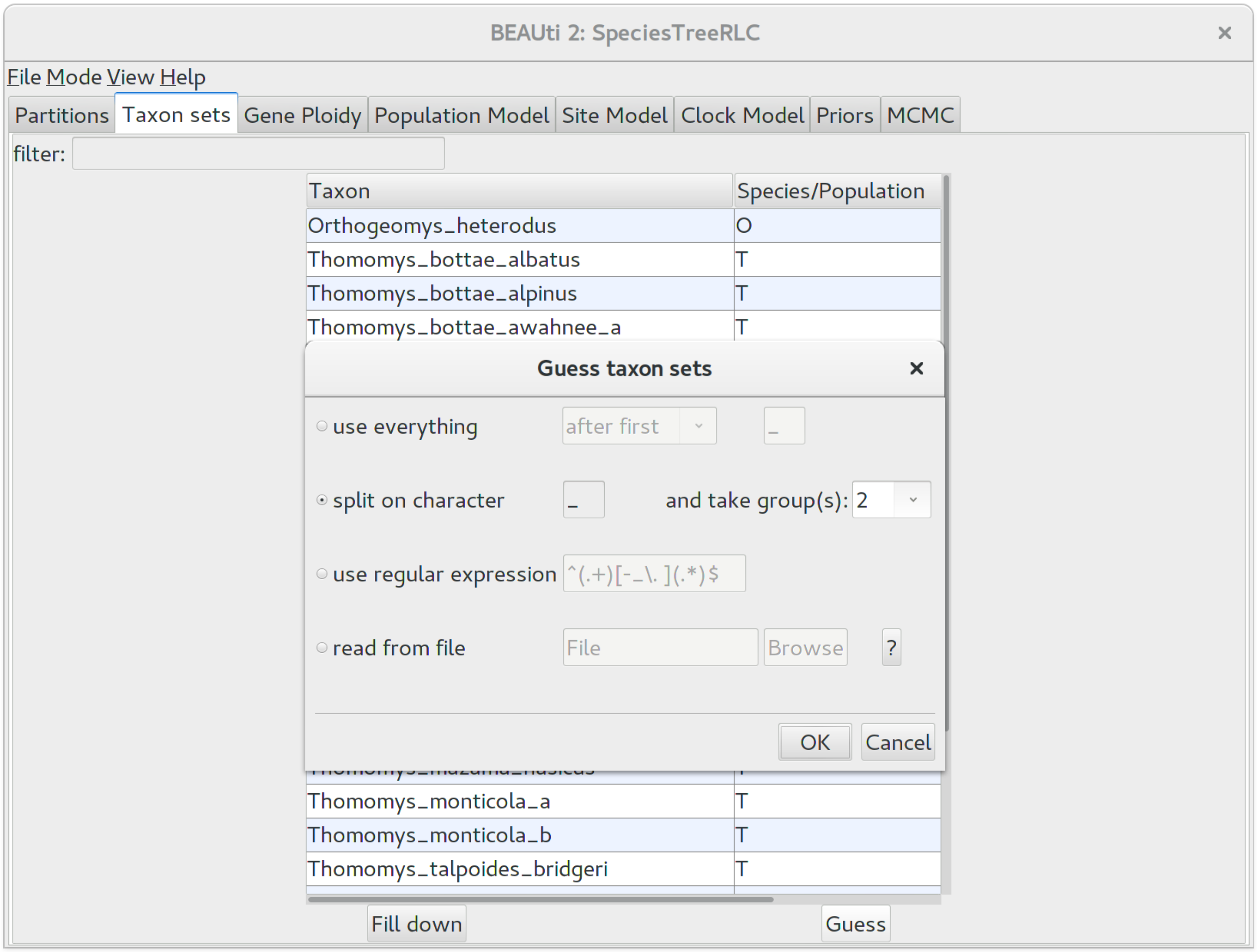
Filter:

Alignment files

Figure 3: Selecting NEXUS alignment files to import.

Assigning the correct species to each sequence

As in the previous tutorial, splitting the name on the underscore character '_' and selecting the second group will give us the mapping that we need (Figure 4).



Taxon	Species/Population
Orthogeomys_heterodus	O
Thomomys_bottae_albatus	T
Thomomys_bottae_alpinus	T
Thomomys_bottae_awahnee_a	T
Thomomys_monticola_a	T
Thomomys_monticola_b	T
Thomomys_talpoides_bridgeri	T

○ use everything — after first — _

● split on character — _ — and take group(s): 2

○ use regular expression — ^(.+)[-_\.](.*)$

○ read from file — File

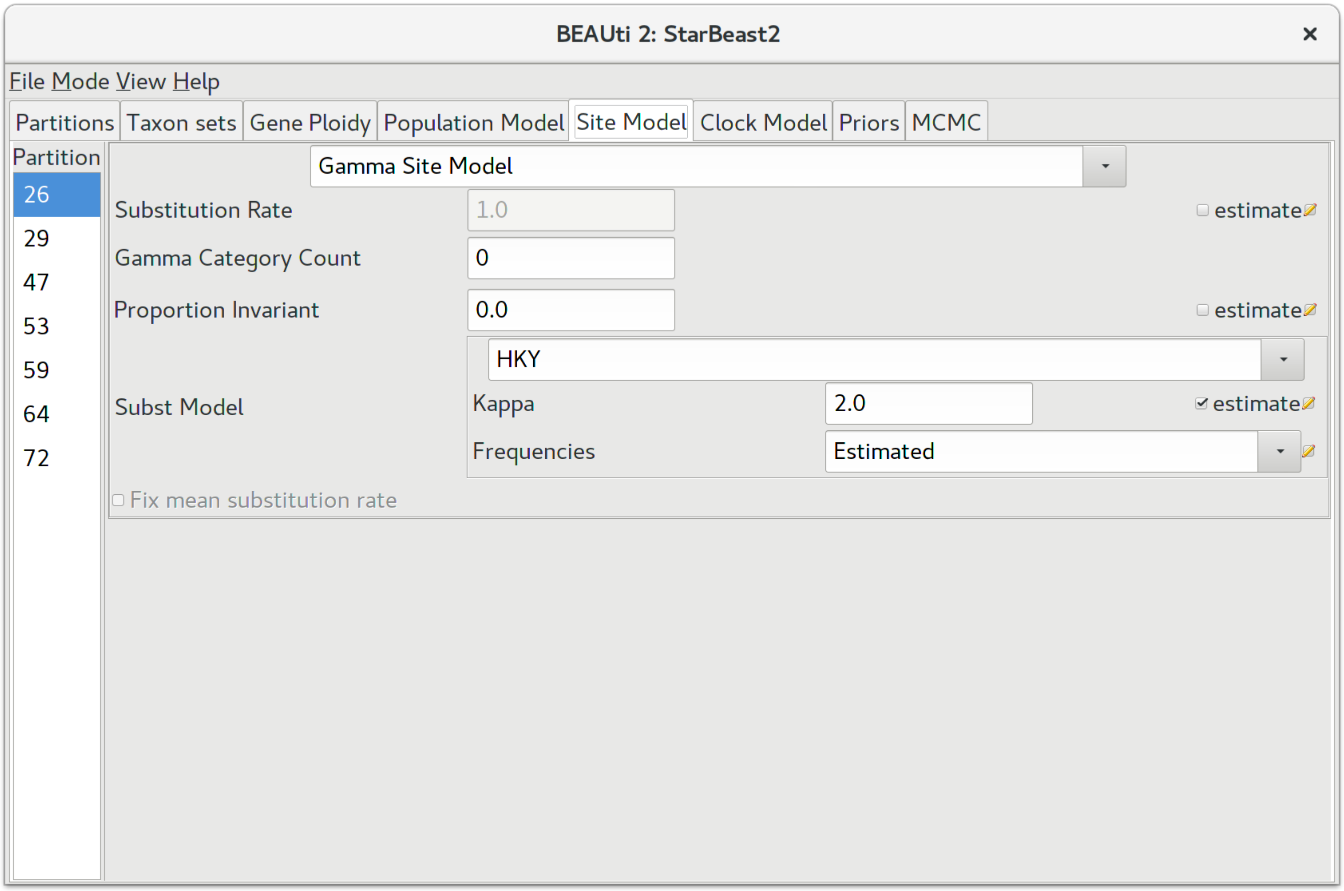
Figure 4: Assigning species to sequences in BEAUti using the guess dialog.

Ploidy and population models

As in the previous tutorial, leave the ploidy of each gene and the population model at their default values.

Setting the substitution model

For this tutorial we will also use the HKY substitution model, but will use a shortcut to set it for all genes at once. For the first gene (which should be “26”) select “HKY” for substitution model (**Subst Model** in Figure 5) Now select the other six genes using the shift key. The right panel will now allow you to set the same substitution model for those six genes by cloning the parameters from the first gene. Click **OK** to proceed (Figure 6).



Gamma Site Model

Substitution Rate 1.0 □ estimate

Gamma Category Count 0

Proportion Invariant 0.0 □ estimate

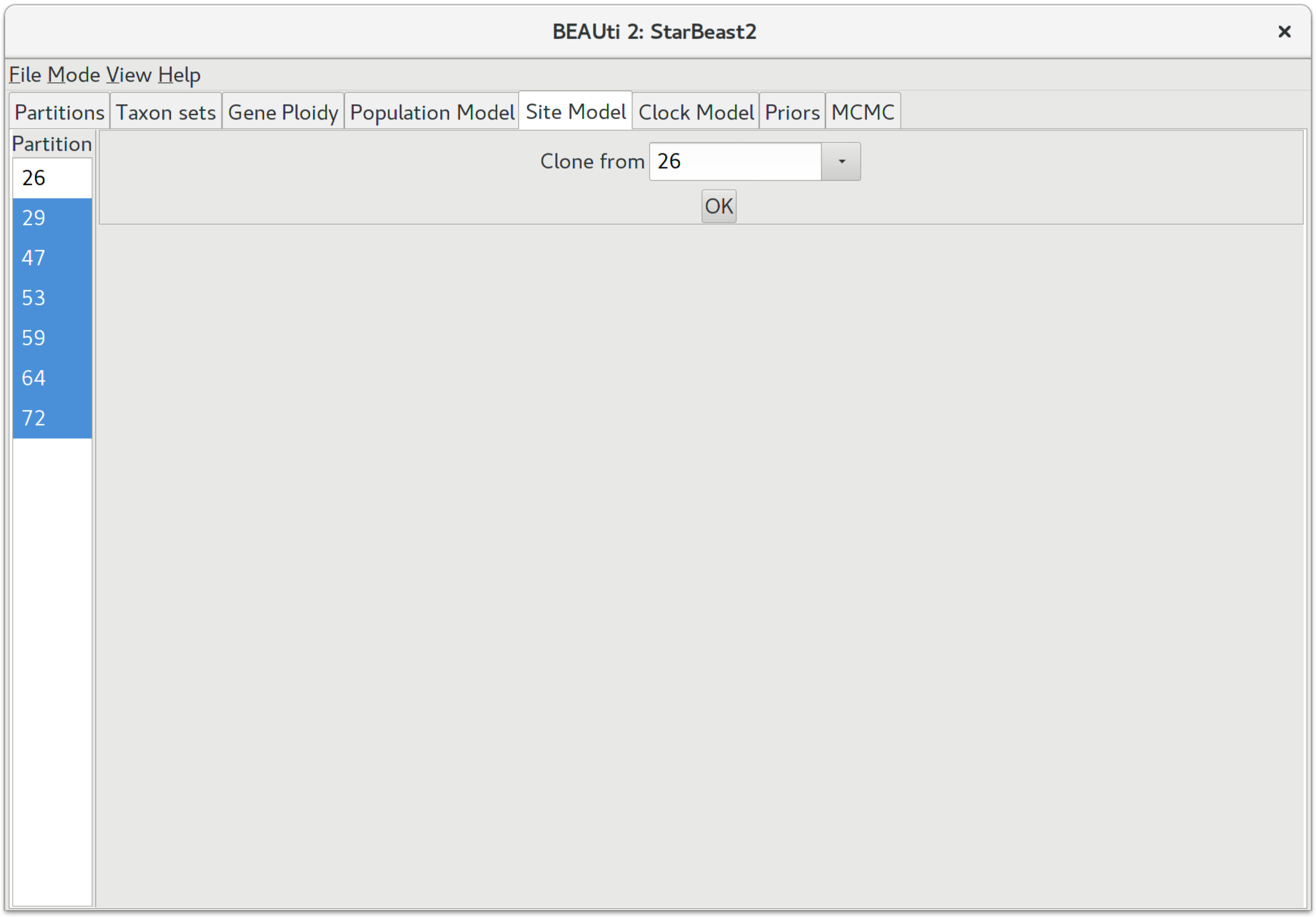
Subst Model HKY

Kappa 2.0 ☑ estimate

Frequencies Estimated

□ Fix mean substitution rate

Figure 5: Setting up substitution and site models for the gopher alignments.



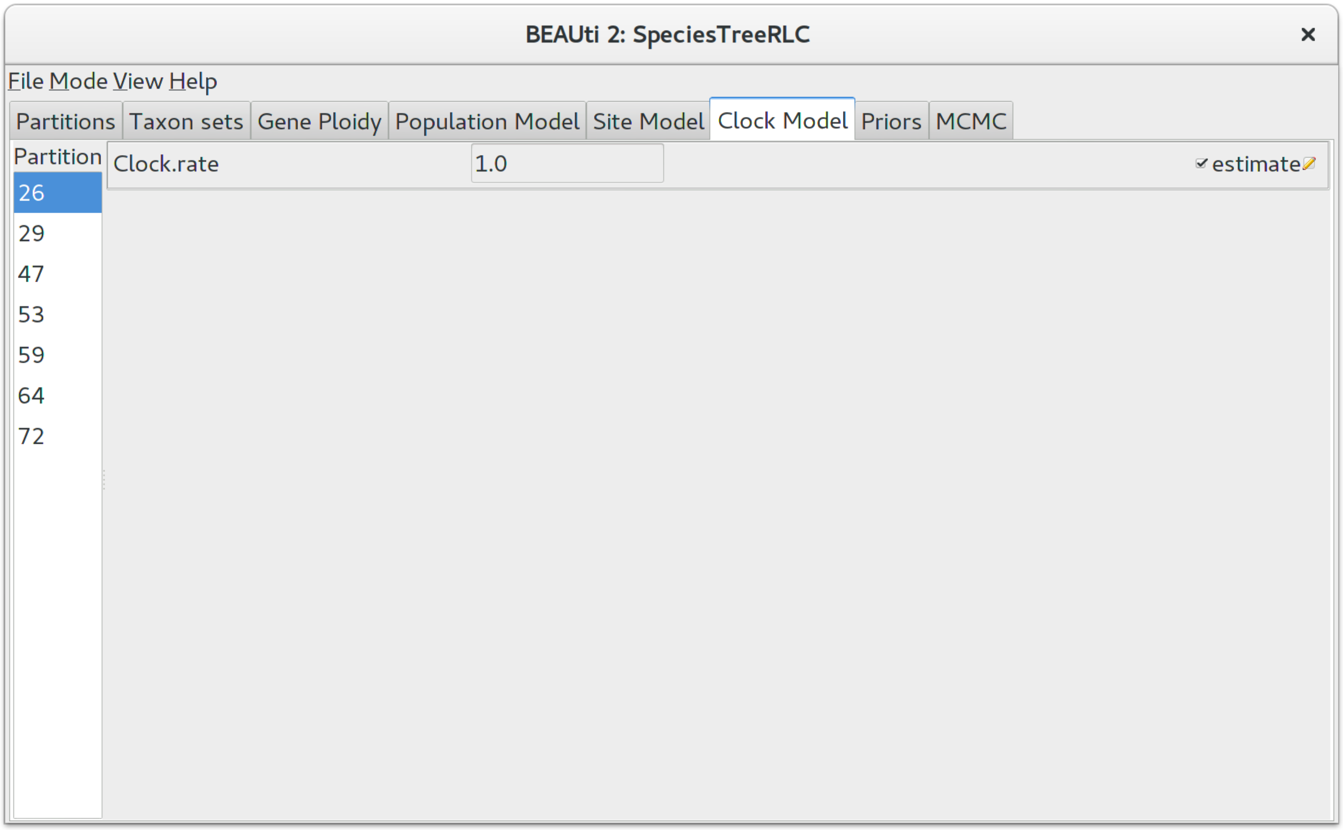
Partition
26
29
47
53
59
64
72

Clone from 26

Figure 6: Setting multiple substitution and site models at once.

Setting the clock model

Click on the **Clock Model** tab at the top of the main window. In this panel you can configure the mean clock rate for each locus. If you followed the earlier instructions to disable automatic setting of clock rates, the mean clock rate “Clock.rate” of all partitions should have the **estimate** box ticked (Figure 7).



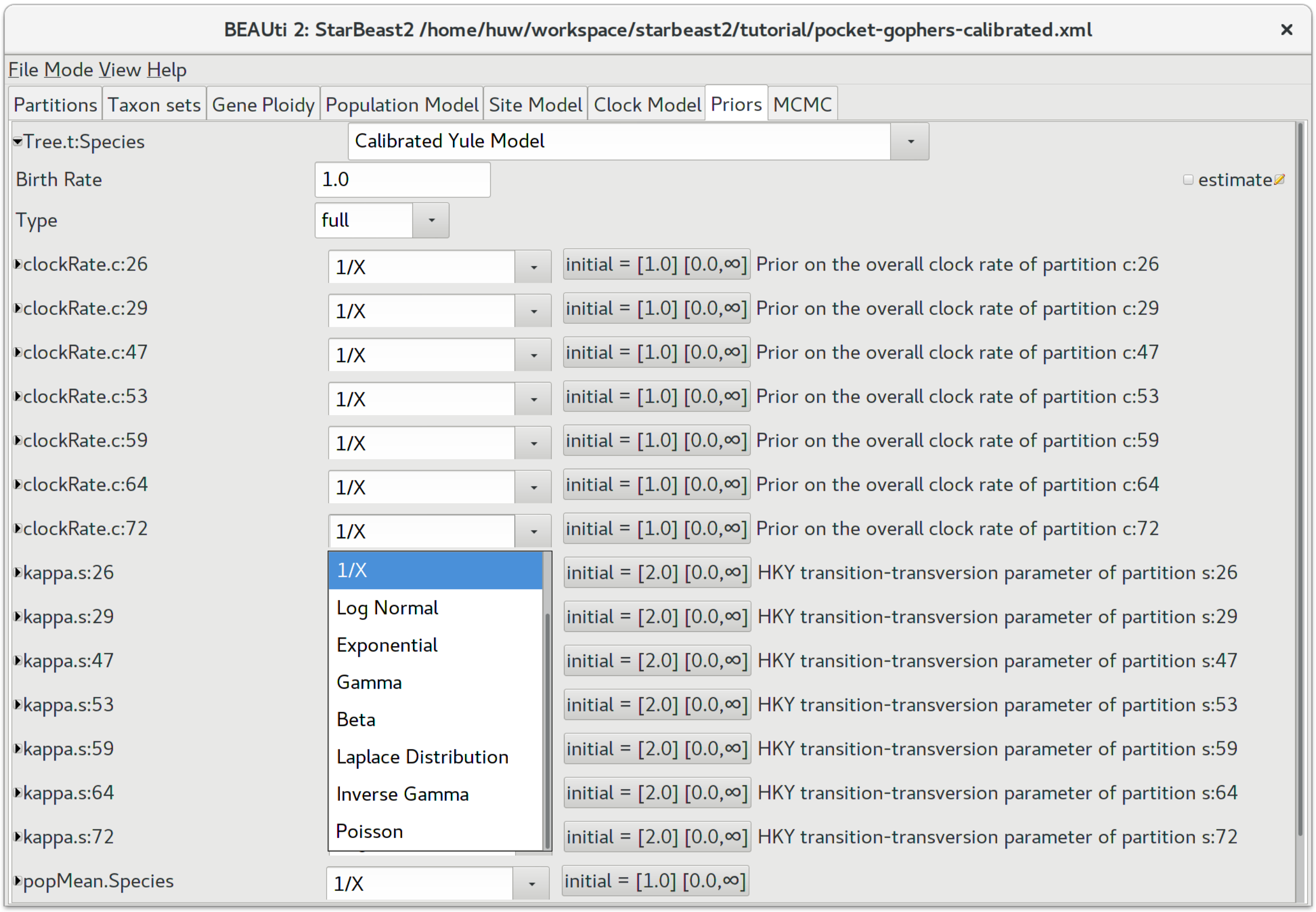
Clock.rate 1.0 ☑ estimate

Figure 7: The default when automatic clock rate setting is disabled.

Priors

Before applying the date calibration first change the prior on the species tree, “Tree.t:Species”, to **calibrated Yule** (Heled and Drummond 2012). Click the top-leftmost arrow to expand the options available for the calibrated Yule model, leave the Birth Rate at 1.0, and uncheck the **estimate** box to make this a fixed parameter. For real analyses you should almost certainly estimate this value, but a fixed value will help us complete the tutorial in a reasonable time frame.

The default prior for mean clock rates in StarBEAST2 is a lognormal distribution with a mean (in real space) of 1. This will not be appropriate when using fossil (or other external) calibrations. For this tutorial we will use the 1/X prior, which is uninformative and will allow the calibration(s) to guide the clock rates. For each clockRate.c:*gene* parameter, set the prior distribution to be 1/X (Figure 8)



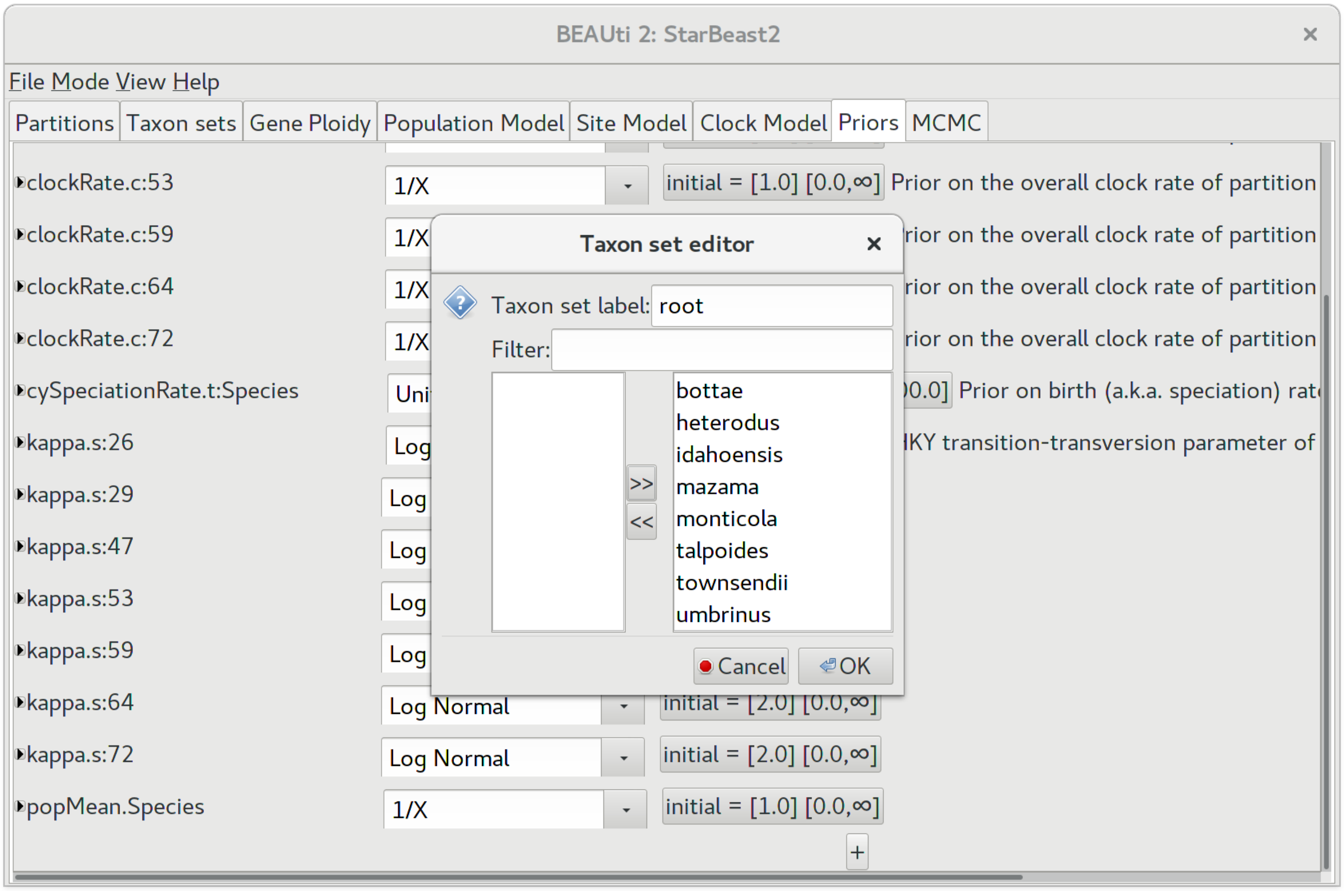
Tree.t:Species	Calibrated Yule Model		
Birth Rate	1.0		□ estimate
Type	full		
clockRate.c:26	1/X	initial = [1.0] [0.0,∞]	Prior on the overall clock rate of partition c:26
clockRate.c:29	1/X	initial = [1.0] [0.0,∞]	Prior on the overall clock rate of partition c:29
clockRate.c:47	1/X	initial = [1.0] [0.0,∞]	Prior on the overall clock rate of partition c:47
clockRate.c:53	1/X	initial = [1.0] [0.0,∞]	Prior on the overall clock rate of partition c:53
clockRate.c:59	1/X	initial = [1.0] [0.0,∞]	Prior on the overall clock rate of partition c:59
clockRate.c:64	1/X	initial = [1.0] [0.0,∞]	Prior on the overall clock rate of partition c:64
clockRate.c:72	1/X	initial = [1.0] [0.0,∞]	Prior on the overall clock rate of partition c:72
kappa.s:26		initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:26
kappa.s:29		initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:29
kappa.s:47		initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:47
kappa.s:53		initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:53
kappa.s:59		initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:59
kappa.s:64		initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:64
kappa.s:72		initial = [2.0] [0.0,∞]	HKY transition-transversion parameter of partition s:72
popMean.Species	1/X	initial = [1.0] [0.0,∞]	

Figure 8: Setting an uninformative prior on clock rates.

We can calibrate this analysis by applying a date calibration to the root of the species tree. This is equivalent to the split between *Thomomys* and *Orthogeomys*, estimated to have occured between 6.5ma and 6.8ma (Belfiore et al. 2008).

To add the new date calibration, click the button with the plus symbol (+) at the very bottom of the list of priors, and click OK to add a new MRCA (most recent common ancestor) prior. Select “Tree.t:Species” to apply this prior to the species tree.

Because this prior applies to the root node, move all the species from the left hand column to the right hand column using the >> button. Give this taxon set a sensible name, for example "root" (Figure 9).



clockRate.c:53 1/X initial = [1.0] [0.0,∞] Prior on the overall clock rate of partition

clockRate.c:59 1/X rior on the overall clock rate of partition

clockRate.c:64 1/X rior on the overall clock rate of partition

clockRate.c:72 1/X rior on the overall clock rate of partition

cySpeciationRate.t:Species Uni 0.0] Prior on birth (a.k.a. speciation) rat

kappa.s:26 Log IKY transition-transversion parameter of

kappa.s:29 Log

kappa.s:47 Log

kappa.s:53 Log

kappa.s:59 Log

kappa.s:64 Log Normal initial = [2.0] [0.0,∞]

kappa.s:72 Log Normal initial = [2.0] [0.0,∞]

popMean.Species 1/X initial = [1.0] [0.0,∞]

Taxon set label: root

Filter:

bottae
heterodus
idahoensis
mazama
monticola
talpoides
townsendii
umbrinus

Figure 9: Setting an uninformative prior on clock rates.

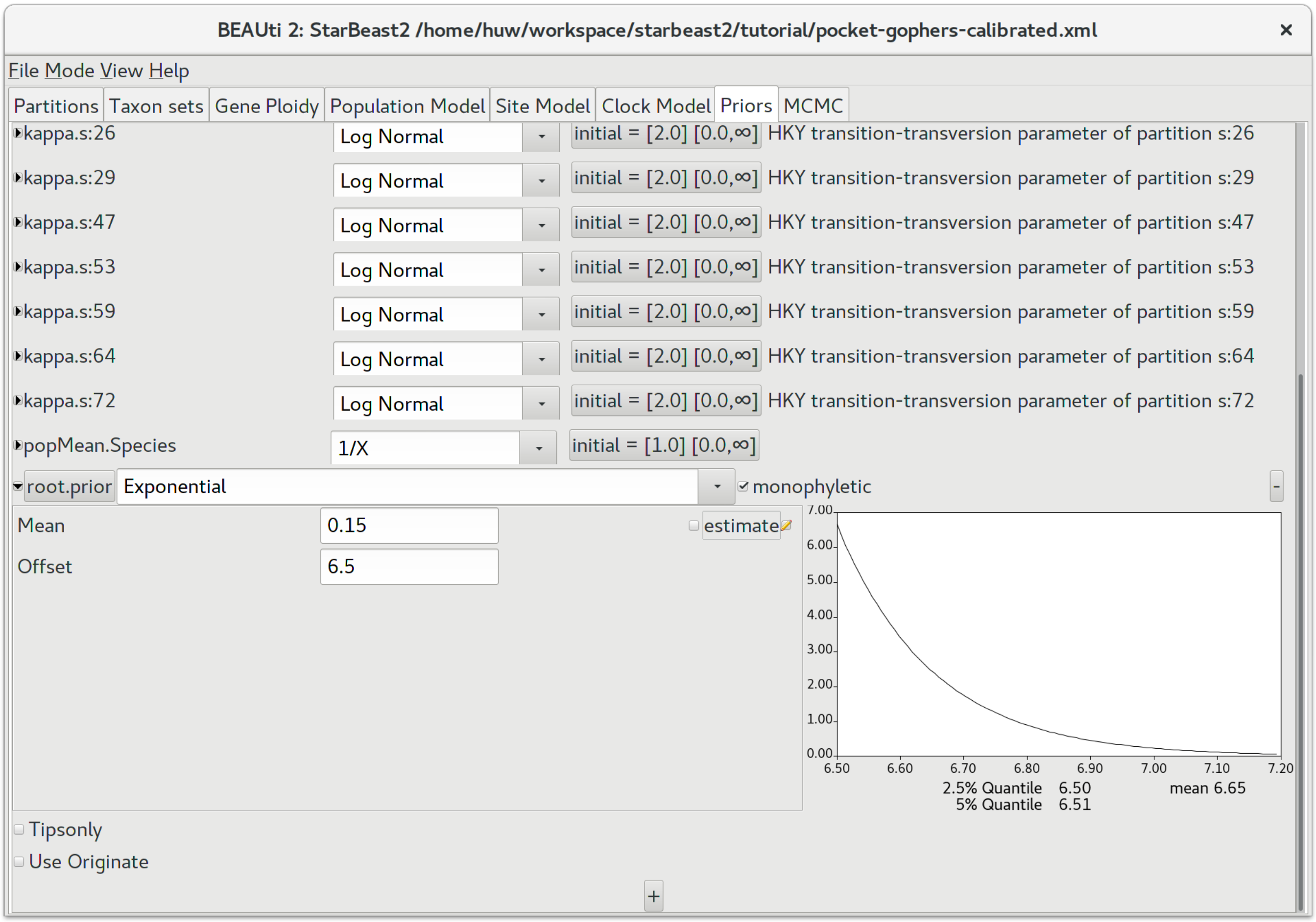
Make this prior monophyletic by checking the "monophyletic" tick box. If 6.5ma is a lower bound, we can use an exponential prior with an offset of 6.5 to set a lower limit on the root node height. Set the mean of the exponential distribution to 0.15, so that the mean of the prior is equal to 6.65ma, the midpoint of the estimated range (Figure 10).

Generating the BEAST XML file

As in the previous tutorial, stick to the default settings in the MCMC tab. We are now ready to create the BEAST XML file. To do this, either select the **File/Save** or **File/Save As** menu options. Save the file with an appropriate name (we usually end the filename with ".xml", e.g. "pocket-gophers-calibrated.xml"). We are now ready to run the file through BEAST.

Running BEAST

Now run BEAST and when it asks for an input file, provide your newly created XML file as input by clicking **Choose File...**, and then click **Run**. In Linux BEAST will immediately launch a file opening dialog box, which is to select the BEAST XML to run. BEAST will then run until it has finished reporting information to the screen. The actual results files are saved to the disk in the same location as your input file.



kappa.s:26 Log Normal initial = [2.0] [0.0,∞] HKY transition-transversion parameter of partition s:26

kappa.s:29 Log Normal initial = [2.0] [0.0,∞] HKY transition-transversion parameter of partition s:29

kappa.s:47 Log Normal initial = [2.0] [0.0,∞] HKY transition-transversion parameter of partition s:47

kappa.s:53 Log Normal initial = [2.0] [0.0,∞] HKY transition-transversion parameter of partition s:53

kappa.s:59 Log Normal initial = [2.0] [0.0,∞] HKY transition-transversion parameter of partition s:59

kappa.s:64 Log Normal initial = [2.0] [0.0,∞] HKY transition-transversion parameter of partition s:64

kappa.s:72 Log Normal initial = [2.0] [0.0,∞] HKY transition-transversion parameter of partition s:72

popMean.Species 1/X initial = [1.0] [0.0,∞]

root.prior Exponential ☑ monophyletic

Mean 0.15 □ estimate

Offset 6.5

2.5% Quantile 6.50
5% Quantile 6.51
mean 6.65

□ Tipsonly

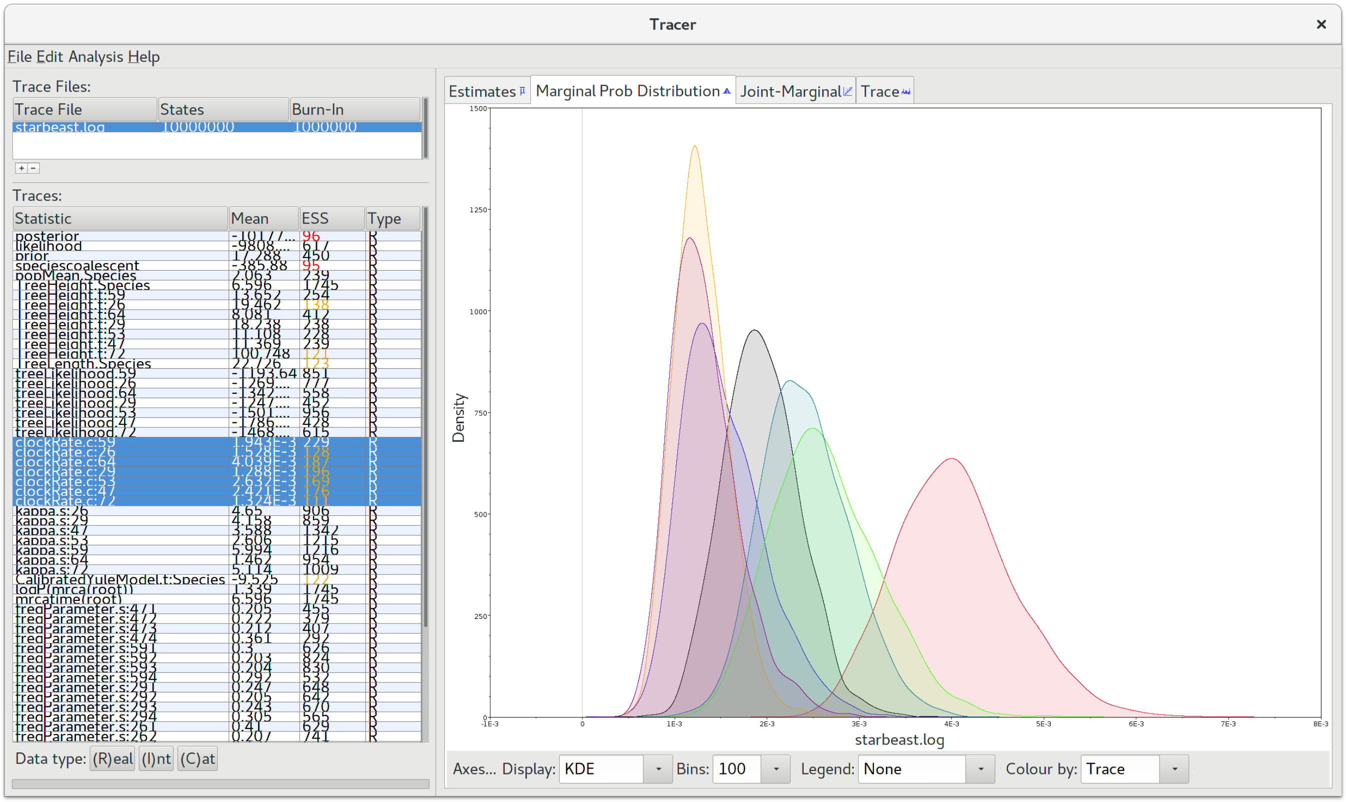
□ Use Originate

Figure 10: Setting an uninformative prior on clock rates.

3 Analyzing the results

Run the program called **Tracer** to analyze the output of BEAST. When the main window has opened, choose **Import Trace File...** from the **File** menu and select the file that BEAST has created called “starbeast.log”. Select the clock rates for all genes using the shift key, then click on the “Marginal Prob Distribution” tab. You should now see a window like in Figure 11.

This display shows the marginal distributions of clock rates for all loci. Remember that MCMC is a stochastic algorithm so the actual distributions will not be exactly the same. The clock rates (in units of substitutions per million years) range from about 0.001 for genes 29 and 72 to 0.004 for locus 64, demonstrating that the loci sequenced by Belfiore et al. 2008 have slower and faster rates of molecular evolution respectively.



Trace Files:

Trace File	States	Burn-In
starbeast.log	10000000	1000000

Traces:

Statistic	Mean	ESS	Type
posterior	-10177...	96	R
likelihood	-9808....	617	R
prior	17.288	450	R
speciescoalescent	-385.88	95	R
popMean.Species	2.063	239	R
TreeHeight.Species	6.596	1745	R
TreeHeight.t:59	13.652	254	R
TreeHeight.t:26	19.462	138	R
TreeHeight.t:64	8.081	412	R
TreeHeight.t:29	18.238	238	R
TreeHeight.t:53	11.108	228	R
TreeHeight.t:47	11.369	239	R
TreeHeight.t:72	100.748	121	R
TreeLength.Species	22.726	123	R
treeLikelihood.59	-1193.64	851	R
treeLikelihood.26	-1269....	777	R
treeLikelihood.64	-1342....	558	R
treeLikelihood.29	-1247....	452	R
treeLikelihood.53	-1501....	956	R
treeLikelihood.47	-1786....	428	R
treeLikelihood.72	-1468....	615	R
clockRate.c:59	1.943E-3	229	R
clockRate.c:26	1.528E-3	128	R
clockRate.c:64	4.039E-3	187	R
clockRate.c:29	1.288E-3	196	R
clockRate.c:53	2.632E-3	169	R
clockRate.c:47	2.421E-3	176	R
clockRate.c:72	1.324E-3	111	R
kappa.s:26	4.65	906	R
kappa.s:29	4.158	859	R
kappa.s:47	3.588	1342	R
kappa.s:53	2.606	1215	R
kappa.s:59	5.994	1216	R
kappa.s:64	1.462	954	R
kappa.s:72	5.114	1009	R
CalibratedYuleModel.t:Species	-9.525	122	R
logP(mrca(root))	1.339	1745	R
mrcatime(root)	6.596	1745	R
freqParameter.s:471	0.205	455	R
freqParameter.s:472	0.222	379	R
freqParameter.s:473	0.212	407	R
freqParameter.s:474	0.361	292	R
freqParameter.s:591	0.3	626	R
freqParameter.s:592	0.203	824	R
freqParameter.s:593	0.204	830	R
freqParameter.s:594	0.292	532	R
freqParameter.s:291	0.247	648	R
freqParameter.s:292	0.205	642	R
freqParameter.s:293	0.243	670	R
freqParameter.s:294	0.305	565	R
freqParameter.s:261	0.41	629	R
freqParameter.s:262	0.207	741	R

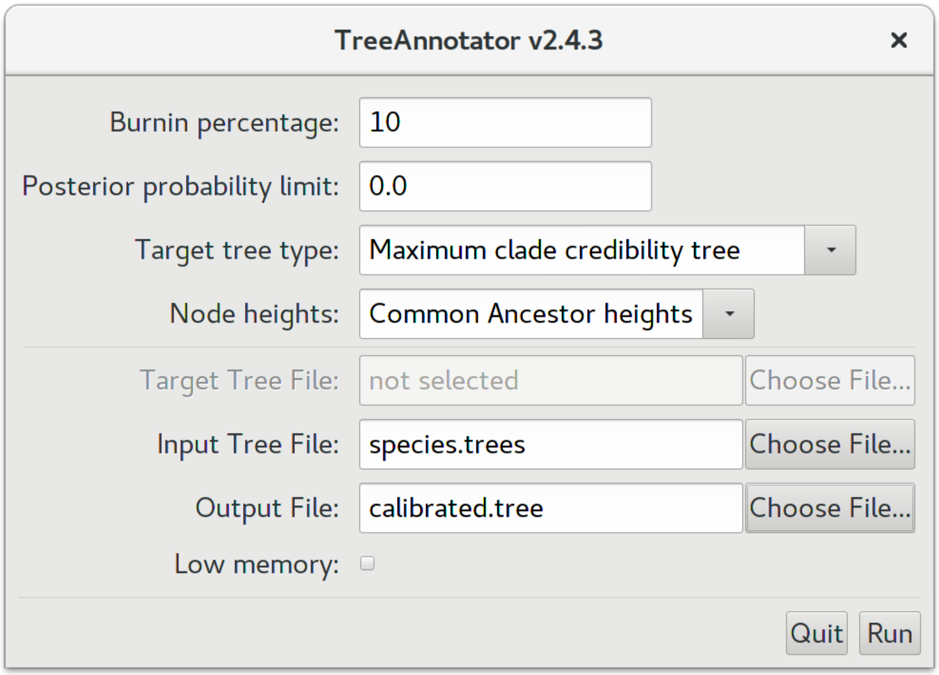
Data type: (R)eal (I)nt (C)at

Display: KDE   Bins: 100   Legend: None   Colour by: Trace

Figure 11: Tracer with the gopher data.

Obtaining an estimate of the phylogenetic tree

Summarize the posterior sample produced by BEAST by running the **TreeAnnotator** program and setting it up to look like in Figure 12.



TreeAnnotator v2.4.3

Burnin percentage: 10

Posterior probability limit: 0.0

Target tree type: Maximum clade credibility tree

Node heights: Common Ancestor heights

Target Tree File: not selected

Input Tree File: species.trees

Output File: calibrated.tree

Low memory: □

Figure 12: Using TreeAnnotator to summarise the tree set.

The **Burnin percentage** is the proportion of trees to remove from the start of the sample; for this tutorial, set a 10% burnin as shown in Figure 12.

The **Posterior probability limit** option specifies a limit such that if a node is found at less than this frequency in the sample of trees (i.e., has a posterior probability less than this limit), it will not be annotated.

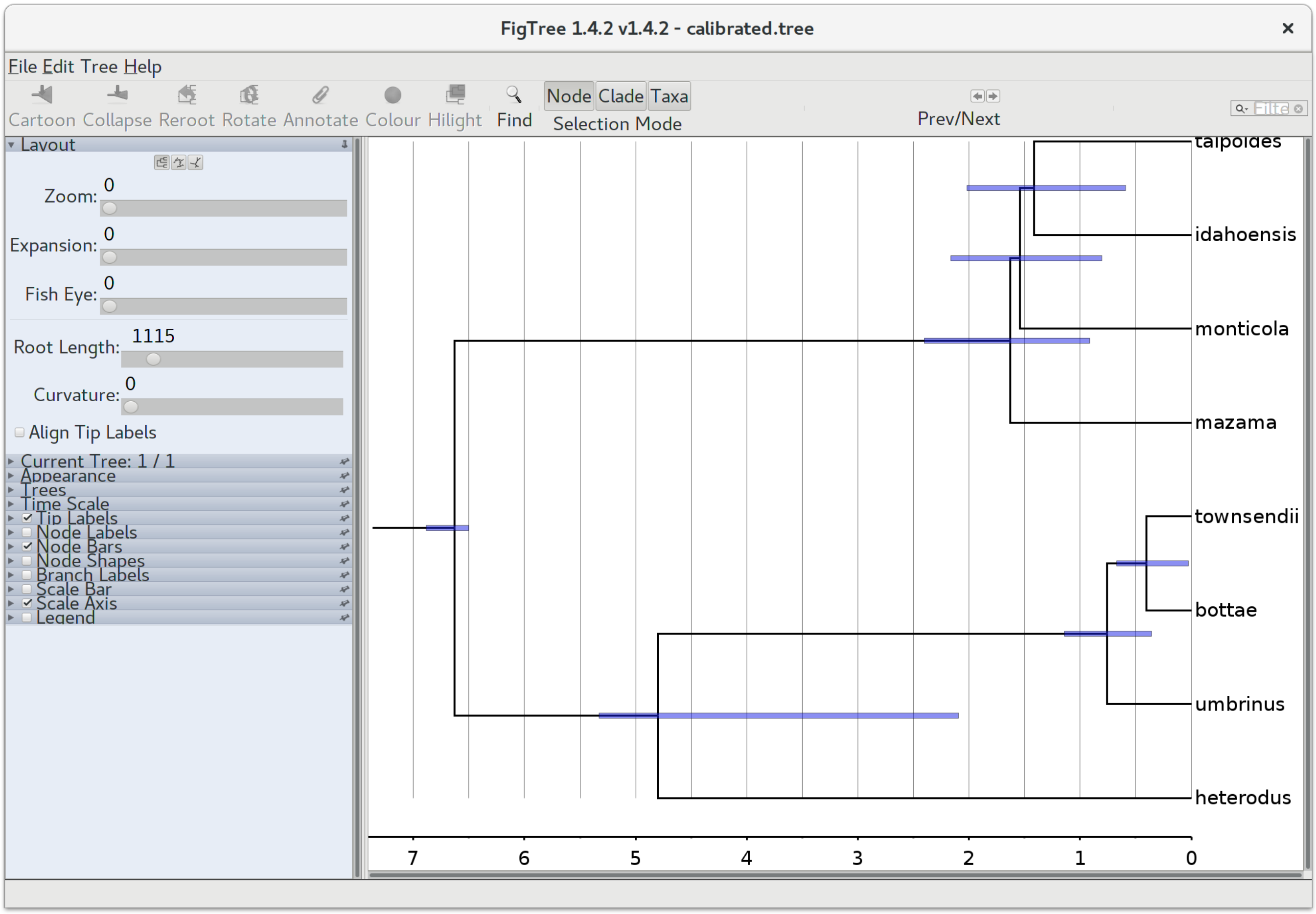
For **Target tree type** you can either choose a specific tree from a file or ask TreeAnnotator to find a tree in your sample. The default option, **Maximum clade credibility tree**, finds the tree with the highest product of the posterior probability of all its nodes.

Keep “Common Ancestor heights” for **Node heights**. This sets the heights (ages) of each node in the tree to the mean height of the most recent common ancestor across the entire set of trees in the posterior.

For the input file, select the trees file that BEAST created (by default this will be called “species.trees”) and select a file for the output (here we have called it “calibrated.tree”). Now press **Run** and wait for the program to finish.

Viewing the species tree(s)

Run **FigTree** and open the "calibrated.tree" file by using the Open command in the File menu. The tree should appear. Select **Node Bars** to get node age error bars. Disable **Scale Bar** and select **Scale Axis** to indicate the ages when speciation events occured. You should end up with something like Figure 13.



Layout

Zoom: 0

Expansion: 0

Fish Eye: 0

Root Length: 1115

Curvature: 0

□ Align Tip Labels

Current Tree: 1 / 1
Appearance
Trees
Time Scale
☑ Tip Labels
□ Node Labels
☑ Node Bars
□ Node Shapes
□ Branch Labels
□ Scale Bar
☑ Scale Axis
□ Legend

Figure 13: Figtree representation of the species tree.

According to our reanalysis, the radiation of *T. townsendii*, *T. bottae*, and *T. umbrinus* occured quite recently, within the last million years. On the other hand timing of the *T. heterodus* split is uncertain, and could have occured more than 5ma or as recently as 2ma.

Version dated: November 2, 2023

Relevant References

Belfiore, NM, L Liu, and C Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus thomomys (rodentia: geomyidae). *Systematic Biology* 57: 294–310.

Heled, J and AJ Drummond. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology* 61: 138–149.