Relaxed Phylogenetics and Dating with Confidence

Alexei J Drummond, Andrew Rambaut and Walter Xie

February 10, 2012

Introduction

This practical will introduce the BEAST software for Bayesian evolutionary analysis, with a focus on estimating phylogenies and divergence times when you have calibration information from fossil evidence or other prior knowledge.

You will need the following software at your disposal:

- **BEAST** - this package contains the BEAST program, BEAUti, TreeAnnotator and other utility programs. This tutorial is written for BEAST v1.6.x, which has support for multiple partitions. It is available for download from `http://beast.bio.ed.ac.uk/`.
- **Tracer** - this program is used to explore the output of BEAST (and other Bayesian MCMC programs). It graphically and quantitively summarizes the distributions of continuous parameters and provides diagnostic information. At the time of writing, the current version is v1.5. It is available for download from `http://beast.bio.ed.ac.uk/`.
- **FigTree** - this is an application for displaying and printing molecular phylogenies, in particular those obtained using BEAST. At the time of writing, the current version is v1.3.1. It is available for download from `http://tree.bio.ed.ac.uk/`.

Rates and dates

This tutorial will guide you through the analysis of an alignment of sequences sampled from twelve primate species. The goal is to estimate the phylogeny as well as the rate of evolution on each lineage based on dates of divergence of their host species.

The first step will be to convert a NEXUS file with a DATA or CHARACTERS block into a BEAST XML input file. This is done using the program BEAUti (this stands for Bayesian Evolutionary Analysis Utility). This is a user-friendly program for setting the evolutionary model and options for the MCMC analysis. The second step is to actually run BEAST using the input file that contains the data, model and settings.

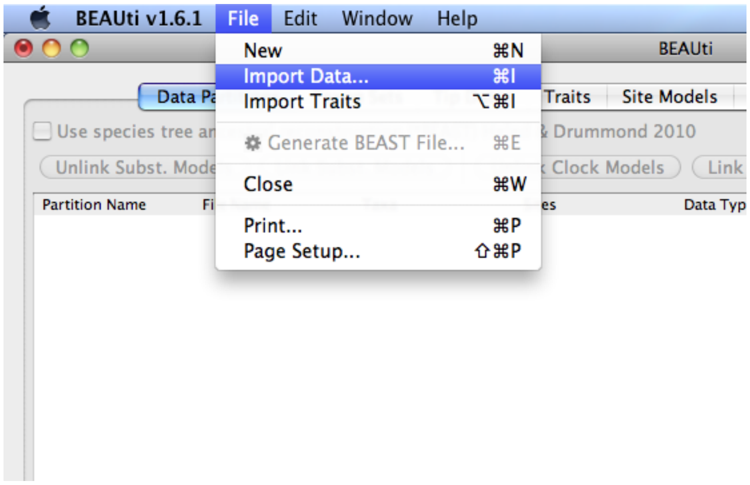
The final step is to explore the output of BEAST in order to diagnose problems and to summarize the results.

BEAUti

The program BEAUti is a user-friendly program for setting the model parameters for BEAST. Run BEAUti by double clicking on its icon.

Loading the NEXUS file

To load a NEXUS format alignment, simply select the `Import Alignment...` option from the File menu:

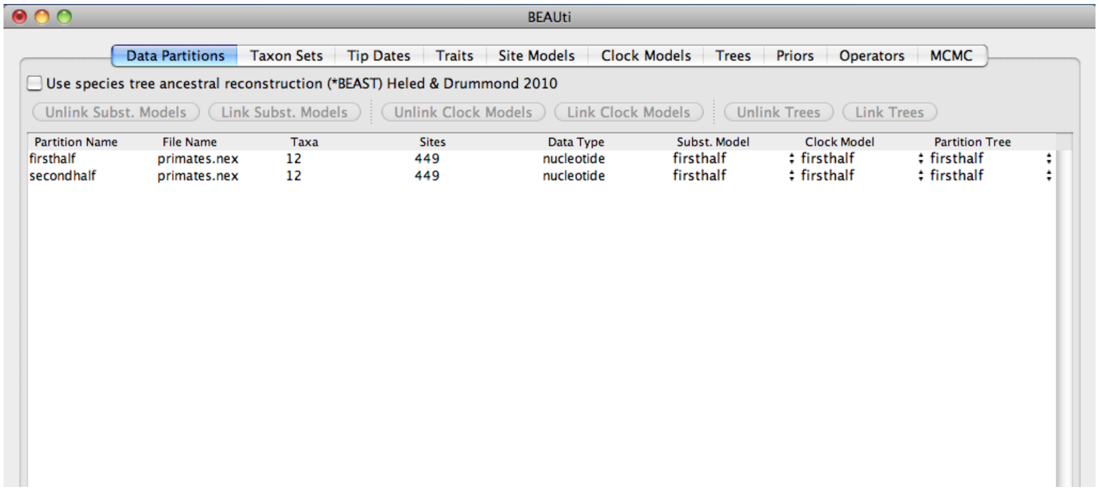


Select the file called `primates.nex`. This file contains an alignment of sequences of 12 species of primates. It looks like this (the lines have been truncated):

```
#NEXUS
begin data;
dimensions ntax=12 nchar=898;
format datatype=dna interleave=no gap=-;
matrix
Tarsius_syrichta  AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACC
Lemur_catta       AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACA
Homo_sapiens      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACA
Pan               AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACA
Gorilla           AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACA
Pongo             AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACA
Hylobates         AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACC
Macaca_fuscata    AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACC
M_mulatta         AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACC
M_fascicularis    AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACC
```

```
M_sylvanus        AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACC
Saimiri_sciureus  AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACT
;
end;

begin assumptions;
charset firsthalf = 1-449;
charset secondhalf = 450-898;
end;
end;
```

Once loaded, the two character partitions are displayed in the main panel:



☐ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

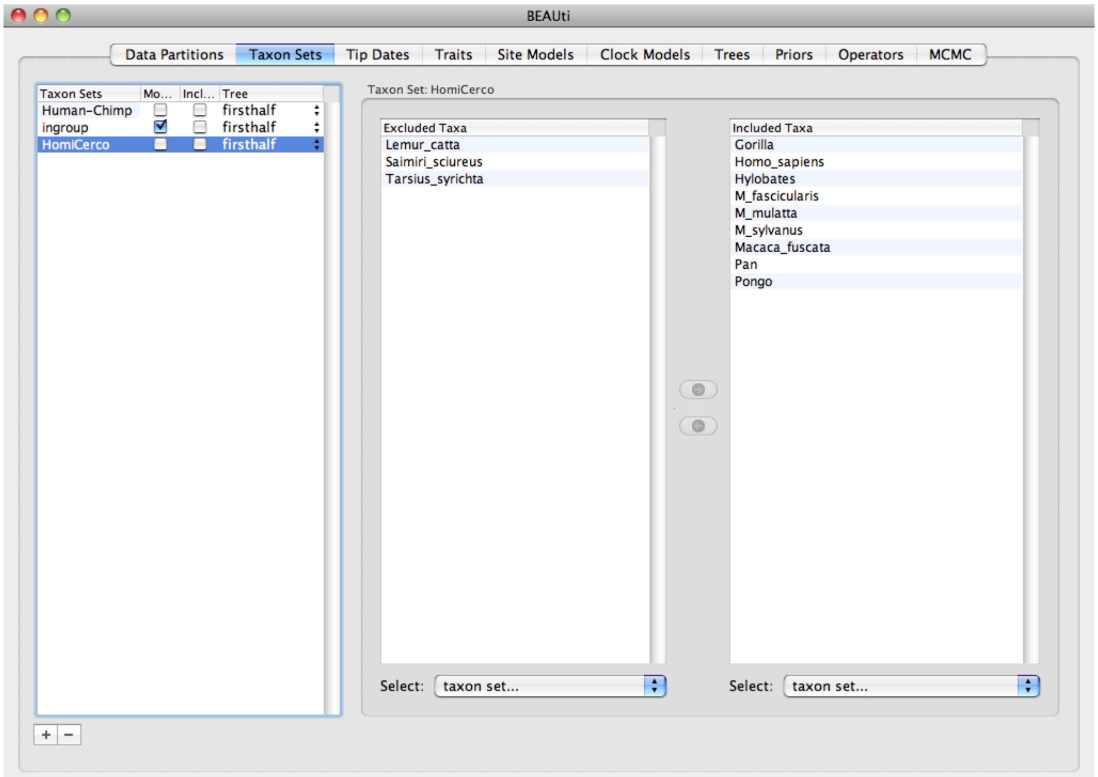
Partition Name	File Name	Taxa	Sites	Data Type	Subst. Model	Clock Model	Partition Tree
firsthalf	primates.nex	12	449	nucleotide	firsthalf	firsthalf	firsthalf
secondhalf	primates.nex	12	449	nucleotide	firsthalf	firsthalf	firsthalf

Defining the calibration nodes

Select the **Taxon Sets** tab at the top of the main window. You will see the panel that allows you to create sets of taxa. Once you have created a taxa set you will be able to add calibration information for its most recent common ancestor (MRCA) later on. Press the small "plus" button at the bottom left of the panel. This will create a new taxon set.

Rename it by double-clicking on the entry that appears (it will initially be called `untitled1`). Call it `ingroup` (it will contain all taxa except the lemur, which will form the outgroup). In the next table along you will see the available taxa. Select all taxa and press the green arrow button. Move the `Lemur` back into the excluded taxa set. Since we know that lemur is the outgroup, we will set select the checkbox in the `Monophyletic?` column. This will ensure that the ingroup is kept monophyletic during the course of the MCMC analysis.

Now repeat the whole procedure creating a set called `Human-Chimp` that contains only `Homo_sapiens` and `Pan` taxa. The screen should look like this:



Taxon Sets	Mo...	Incl...	Tree
Human-Chimp	☐	☐	firsthalf
ingroup	☑	☐	firsthalf
HomiCerco	☐	☐	firsthalf

Taxon Set: HomiCerco

Excluded Taxa
Lemur_catta
Saimiri_sciureus
Tarsius_syrichta

Select: taxon set...

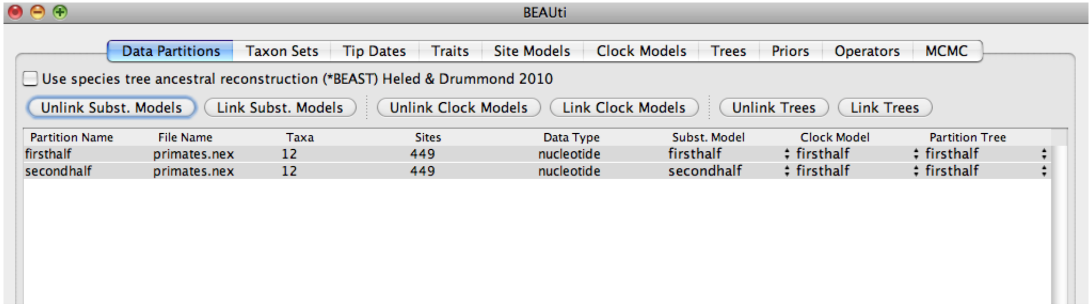
Included Taxa
Gorilla
Homo_sapiens
Hylobates
M_fascicularis
M_mulatta
M_sylvanus
Macaca_fuscata
Pan
Pongo

Select: taxon set...

Finally, create a taxon group that contains everything under the hominoid/cercopithecoid split (i.e. everything except `Lemur`, `Saimiri` and `Tarsius`). Call this taxon set something like `HomiCerco`.

Unlink partition models

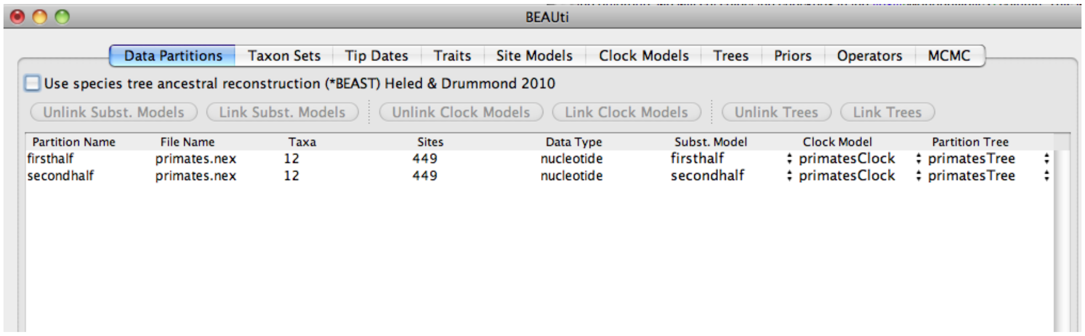
At this point we will need to unlink the substitution model so that each parameter is estimated separately for the two partitions. To do this return to **Data Partitions** panel, select both partitions in the table and click the `Unlink Subst Models` button.



☐ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

Partition Name	File Name	Taxa	Sites	Data Type	Subst. Model	Clock Model	Partition Tree
firsthalf	primates.nex	12	449	nucleotide	firsthalf	firsthalf	firsthalf
secondhalf	primates.nex	12	449	nucleotide	secondhalf	firsthalf	firsthalf

And you can also change the partition model name in its corresponding panel (e.g. **Clock Models** panel), and make the final partitions as illustrated below:



□ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

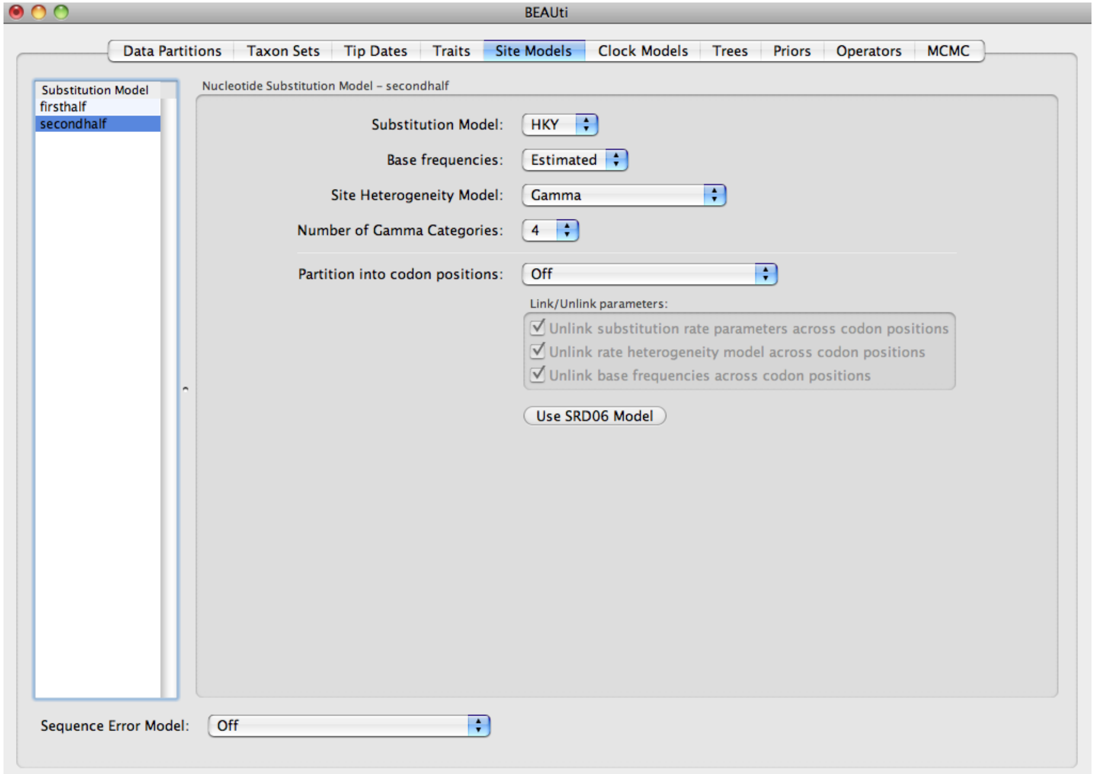
Partition Name	File Name	Taxa	Sites	Data Type	Subst. Model	Clock Model	Partition Tree
firsthalf	primates.nex	12	449	nucleotide	firsthalf	primatesClock	primatesTree
secondhalf	primates.nex	12	449	nucleotide	secondhalf	primatesClock	primatesTree

Setting the substitution model

The next thing to do is to click on the **Site Models** tab at the top of the main window. This will reveal the evolutionary model settings for BEAST. Exactly which options appear depend on whether the data are nucleotides, or amino acids, or binary data, or general data. The settings that will appear after loading the Primates data set will be the default values so we need to make some changes.

Most of the models should be familiar to you. For this analysis, we will **respectively** configure two substitution models by clicking each of them **one by one** on the left side to make the same change: select **Gamma** under the **Site Heterogeneity Model** menu which will allow rate variation between sites in the associated alignment.

Note: Do not forget to change ALL substitution models.



Nucleotide Substitution Model - secondhalf

Substitution Model: HKY

Base frequencies: Estimated

Site Heterogeneity Model: Gamma

Number of Gamma Categories: 4

Partition into codon positions: Off

☑ Unlink substitution rate parameters across codon positions

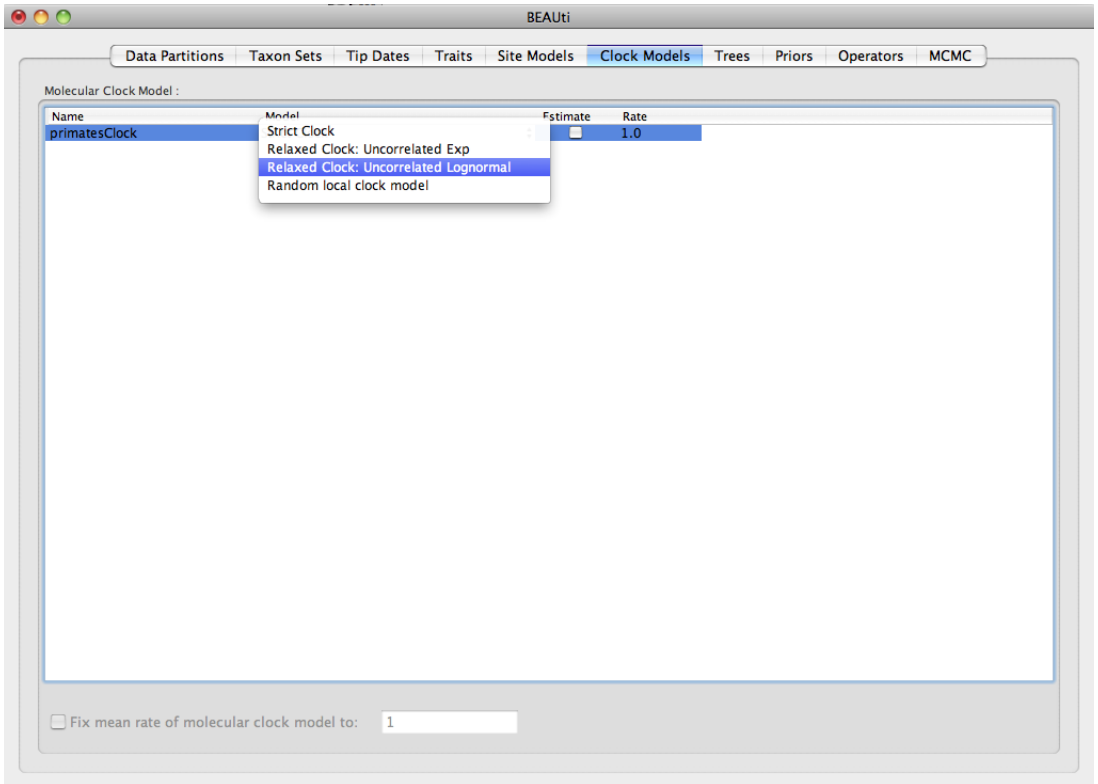
☑ Unlink rate heterogeneity model across codon positions

☑ Unlink base frequencies across codon positions

Sequence Error Model: Off

Setting the clock model

Second, we will do is to click on the **Clock Models** tab at the top of the main window, and to change the molecular clock model to **Relaxed Clock: Uncorrelated Lognormal** so as to account for lineage-specific rate heterogeneity. Your model options should now look like this:



Molecular Clock Model :

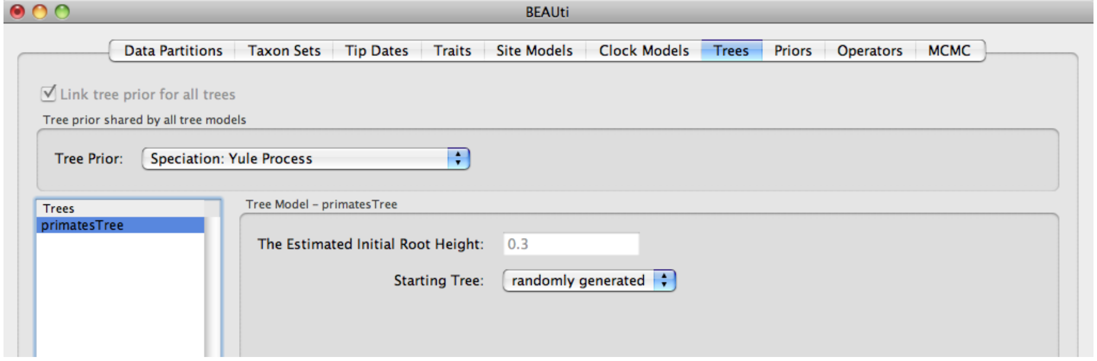
Name	Model	Estimate	Rate
primatesClock	Relaxed Clock: Uncorrelated Lognormal	☐	1.0

☐ Fix mean rate of molecular clock model to: 1

The **Estimate** check box is required to be checked, because we wish to estimate the clock rate (and in doing so the divergence times). But this will be automatically checked, in this case, when we put a proper prior on **tmcra** statistics appeared in **Priors** panel.

Trees

The **Trees** tab allows priors to be specified for each parameter in the model. The first thing to do is to specify that we wish to use the **Yule** model as the tree prior. This is a simple model of speciation that is generally more appropriate when considering sequences from different species. Select this from the **Tree prior** dropdown menu.



☑ Link tree prior for all trees

Tree prior shared by all tree models

Tree Prior: Speciation: Yule Process

Tree Model – primatesTree

The Estimated Initial Root Height: 0.3

Starting Tree: randomly generated

Priors

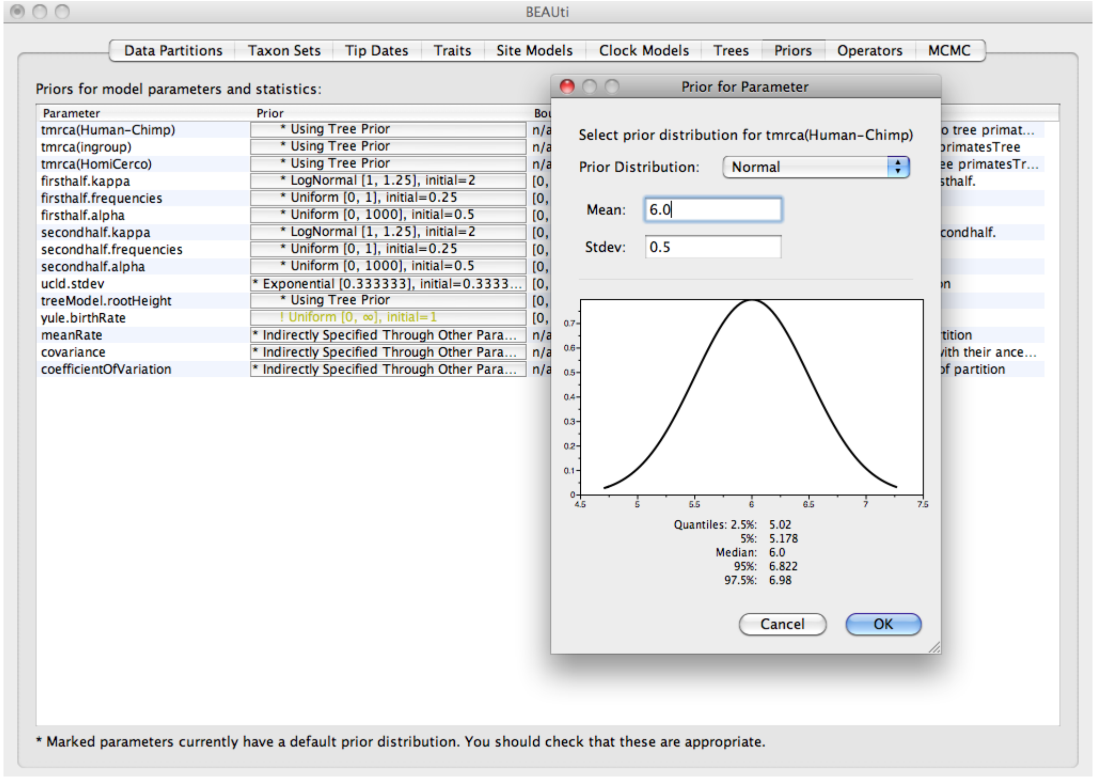
The **Priors** tab allows priors to be specified for each parameter in the model. The first thing to do is to specify that we wish to use the Yule model as the tree prior. This is a simple model of speciation that is generally more appropriate when considering sequences from different species. Select this from the **Tree prior** dropdown menu.

We now need to specify a prior distribution for some of the divergence times, based on our prior fossil knowledge. This is known as calibrating our tree. We will actually use two calibrations in this analysis. Click on the button in the table next to `tmrca(human-chimp)`, A dialog box will appear allowing you to specify a prior for the MRCA of species.

Select the **Normal** distribution. We are going to assume a normal distribution centered at 6 million years with a standard deviation of 0.5 million years. This will give a central 95% range of about 5-7 My. This corresponds to the consensus estimate of the date of the most recent common ancestor of humans and chimps.

Following the same procedure set a calibration of 24 +/- 0.5 million (stdev) for the hominoid-cercopithecoid split.

Although we created a taxon set for the ingroup (`tmrca(ingroup)` in the prior table), we are not going to put an informative prior on this. We can then estimate this divergence time based on the other calibrations.



Priors for model parameters and statistics:

Parameter	Prior	Bou	
tmrca(Human-Chimp)	* Using Tree Prior	n/a	o tree primat...
tmrca(ingroup)	* Using Tree Prior	n/a	primatesTree
tmrca(HomiCerco)	* Using Tree Prior	n/a	ee primatesTr...
firsthalf.kappa	* LogNormal [1, 1.25], initial=2	[0,	sthalf.
firsthalf.frequencies	* Uniform [0, 1], initial=0.25	[0,	
firsthalf.alpha	* Uniform [0, 1000], initial=0.5	[0,	
secondhalf.kappa	* LogNormal [1, 1.25], initial=2	[0,	condhalf.
secondhalf.frequencies	* Uniform [0, 1], initial=0.25	[0,	
secondhalf.alpha	* Uniform [0, 1000], initial=0.5	[0,	
ucld.stdev	* Exponential [0.333333], initial=0.3333...	[0,	n
treeModel.rootHeight	* Using Tree Prior	[0,	
yule.birthRate	! Uniform [0, ∞], initial=1	[0,	
meanRate	* Indirectly Specified Through Other Para...	n/a	tition
covariance	* Indirectly Specified Through Other Para...	n/a	ith their ance...
coefficientOfVariation	* Indirectly Specified Through Other Para...	n/a	f partition

Prior for Parameter

Select prior distribution for tmrca(Human-Chimp)

Prior Distribution: Normal

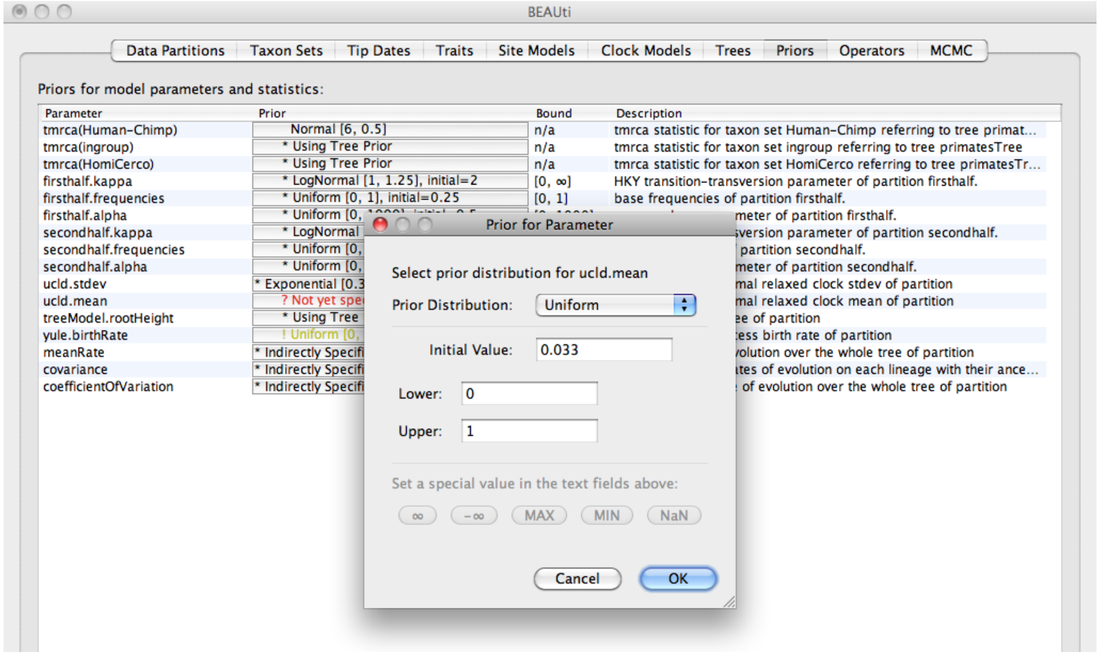
Mean: 6.0

Stdev: 0.5

```
Quantiles: 2.5%:   5.02
             5%:   5.178
         Median:   6.0
            95%:   6.822
          97.5%:   6.98
```

* Marked parameters currently have a default prior distribution. You should check that these are appropriate.

And the clock model parameters will appear when the clock rate is estimated. The priors table should now look like this:



Priors for model parameters and statistics:

Parameter	Prior	Bound	Description
tmrca(Human-Chimp)	Normal [6, 0.5]	n/a	tmrca statistic for taxon set Human-Chimp referring to tree primat...
tmrca(ingroup)	* Using Tree Prior	n/a	tmrca statistic for taxon set ingroup referring to tree primatesTree
tmrca(HomiCerco)	* Using Tree Prior	n/a	tmrca statistic for taxon set HomiCerco referring to tree primatesTr...
firsthalf.kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY transition-transversion parameter of partition firsthalf.
firsthalf.frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	base frequencies of partition firsthalf.
firsthalf.alpha	* Uniform [0,		meter of partition firsthalf.
secondhalf.kappa	* LogNormal		sversion parameter of partition secondhalf.
secondhalf.frequencies	* Uniform [0,		partition secondhalf.
secondhalf.alpha	* Uniform [0,		meter of partition secondhalf.
ucld.stdev	* Exponential [0.3		mal relaxed clock stdev of partition
ucld.mean	? Not yet spe		mal relaxed clock mean of partition
treeModel.rootHeight	* Using Tree		ee of partition
yule.birthRate	! Uniform [0,		ess birth rate of partition
meanRate	* Indirectly Specifi		volution over the whole tree of partition
covariance	* Indirectly Specifi		tes of evolution on each lineage with their ance...
coefficientOfVariation	* Indirectly Specifi		of evolution over the whole tree of partition

Prior for Parameter

Select prior distribution for ucld.mean

Prior Distribution: Uniform

Initial Value: 0.033

Lower: 0

Upper: 1

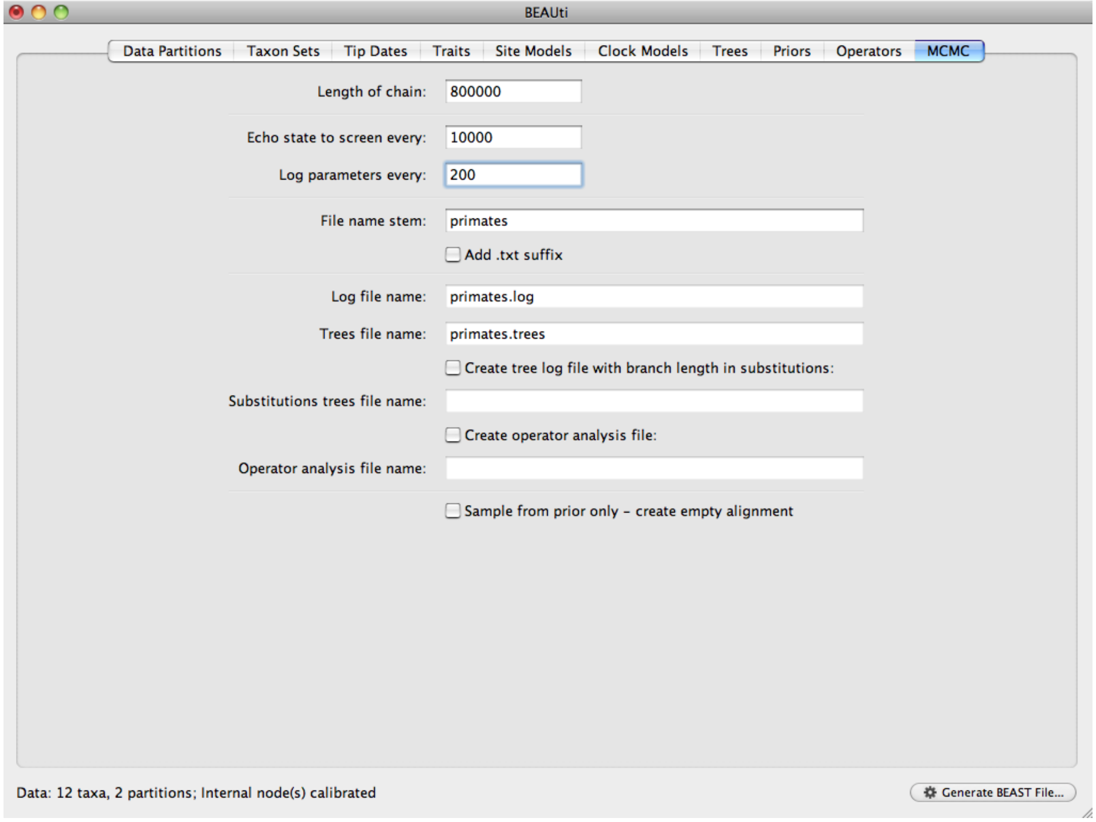
Set a special value in the text fields above:

Setting the MCMC options

Ignore the **Operators** tab as this just contains technical settings effecting the efficiency of the MCMC program (see Notes for details).

The next tab, **MCMC**, provides more general settings to control the length of the MCMC and the file names.

Firstly we have the **Length of chain**. This is the number of steps the MCMC will make in the chain before finishing. How long this should be depends on the size of the data set, the complexity of the model and the quality of answer required. The default value of 10,000,000 is entirely arbitrary and should be adjusted according to the size of your data set. For this data set let's initially set the chain length to 800,000 as this will run reasonably quickly on most modern computers (a few minutes).



Length of chain: 800000

Echo state to screen every: 10000

Log parameters every: 200

File name stem: primates

□ Add .txt suffix

Log file name: primates.log

Trees file name: primates.trees

□ Create tree log file with branch length in substitutions:

Substitutions trees file name:

□ Create operator analysis file:

Operator analysis file name:

□ Sample from prior only – create empty alignment

Data: 12 taxa, 2 partitions; Internal node(s) calibrated

The next options specify how often the parameter values in the Markov chain should be displayed on the screen and recorded in the log file. The screen output is simply for monitoring the programs progress so can be set to any value (although if set too small, the sheer quantity of information being displayed on the screen will actually slow the program down). For the log file, the value should be set relative to the total length of the chain. Sampling too often will result in very large files with little extra benefit in terms of the precision of the analysis. Sample too infrequently and the log file will not contain much information about the distributions of the parameters. You probably want to aim to store no more than 10,000 samples so this should be set to no less than chain length / 10000.

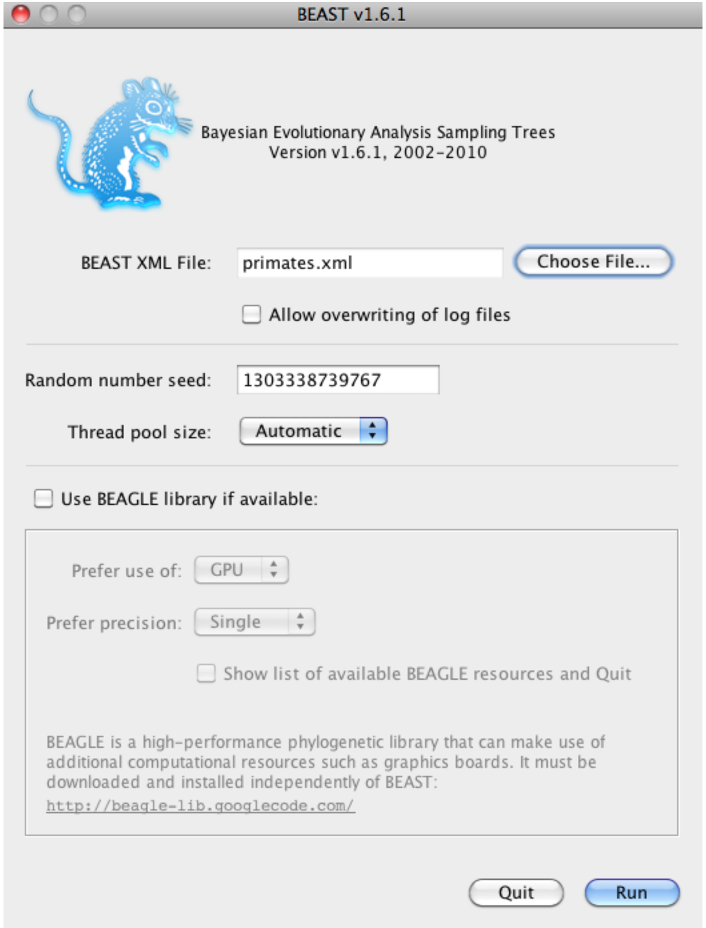
For this exercise we will set the screen log to 10,000 and the file log to 200. The final two options give the file names of the log files for the sampled parameters and the trees. These will be set to a default based on the name of the imported NEXUS file.

- If you are using windows then we suggest you add the suffix `.txt` to both of these (so, `Primates.log.txt` and `Primates.trees.txt`) so that Windows recognizes these as text files.

Generating the BEAST XML file

We are now ready to create the BEAST XML file. To do this, either select the **Generate BEAST File...** option from the **File** menu or click the similarly labelled button at the bottom of the window. Check the default priors, and save the file with an appropriate name (we usually end the filename with `.xml`, i.e., `Primates.xml`). We are now ready to run the file through BEAST.

Running BEAST



Bayesian Evolutionary Analysis Sampling Trees
Version v1.6.1, 2002-2010

BEAST XML File: primates.xml

□ Allow overwriting of log files

Random number seed: 1303338739767

Thread pool size: Automatic

□ Use BEAGLE library if available:

Prefer use of: GPU

Prefer precision: Single

□ Show list of available BEAGLE resources and Quit

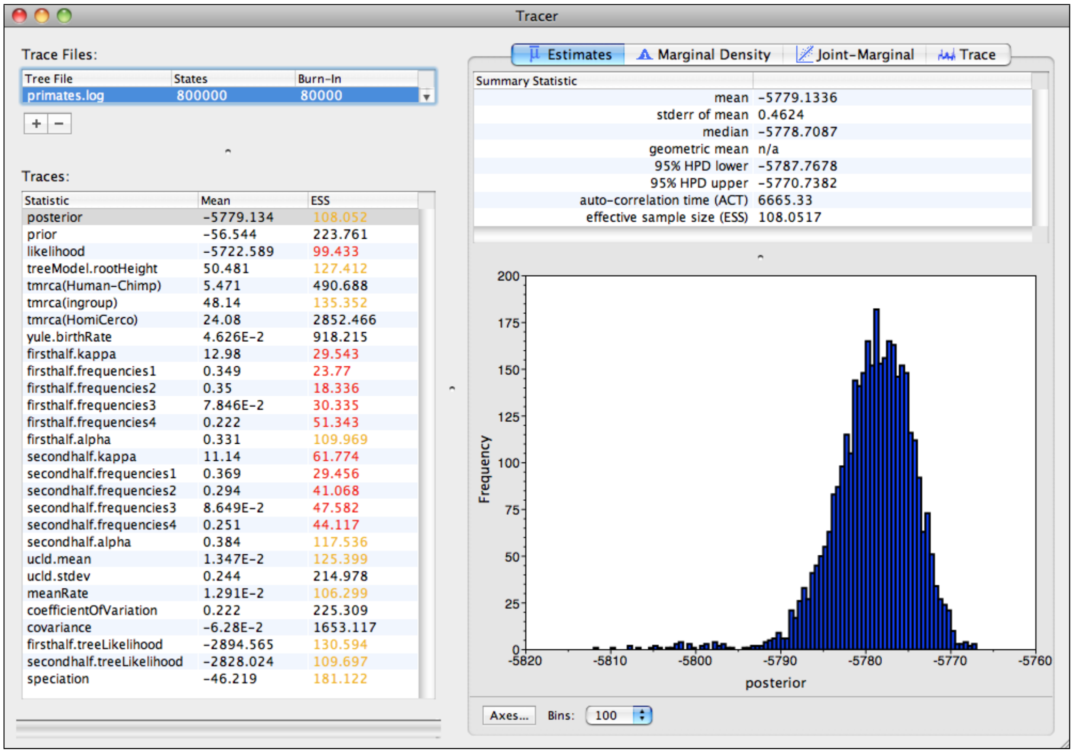
BEAGLE is a high-performance phylogenetic library that can make use of additional computational resources such as graphics boards. It must be downloaded and installed independently of BEAST:
http://beagle-lib.googlecode.com/

Now run BEAST and when it asks for an input file, provide your newly created XML file as input. BEAST will then run until it has finished reporting information to the screen. The actual results files are save to the disk in the same location as your input file. The output to the screen will look something like this:

```
                   BEAST v1.6.1, 2002-2010
       Bayesian Evolutionary Analysis Sampling Trees
                 Designed and developed by
   Alexei J. Drummond, Andrew Rambaut and Marc A. Suchard

               Department of Computer Science
                   University of Auckland
                  alexei@cs.auckland.ac.nz

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

              David Geffen School of Medicine
           University of California, Los Angeles
                     msuchard@ucla.edu

                Downloads, Help & Resources:
                  http://beast.bio.ed.ac.uk

Source code distributed under the GNU Lesser General Public License:
             http://code.google.com/p/beast-mcmc

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

                         Thanks to:
     Roald Forsberg, Beth Shapiro and Korbinian Strimmer


Random number seed: 1303338739767

Parsing XML file: primates.xml
  File encoding: MacRoman
Read alignment: alignment
  Sequences = 12
      Sites = 898
   Datatype = nucleotide
Site patterns 'firsthalf.patterns' created from positions 1-449 of alignment 'alignment'
  pattern count = 227
Site patterns 'secondhalf.patterns' created from positions 450-898 of alignment 'alignment'
  pattern count = 231
Using Yule prior on tree
Creating the tree model, 'treeModel'
  initial tree topology = ((((((Gorilla,Tarsius_syrichta),Pan),M_sylvanus),(M_mulatta,Macaca_fuscata)),((((Homo_sapiens,Saimi
  tree height = 311.73866128365125
Using discretized relaxed clock model.
  over sampling = 1
  parametric model = logNormalDistributionModel
   rate categories = 22
Creating state frequencies model: Initial frequencies = {0.25, 0.25, 0.25, 0.25}
Creating HKY substitution model. Initial kappa = 2.0
Creating site model.
  4 category discrete gamma with initial shape = 0.5
Creating state frequencies model: Initial frequencies = {0.25, 0.25, 0.25, 0.25}
Creating HKY substitution model. Initial kappa = 2.0
Creating site model.
  4 category discrete gamma with initial shape = 0.5
TreeLikelihood(treeModel) using native nucleotide likelihood core
  Ignoring ambiguities in tree likelihood.
  With 227 unique site patterns.
Branch rate model used: discretizedBranchRates
TreeLikelihood(treeModel) using native nucleotide likelihood core
  Ignoring ambiguities in tree likelihood.
  With 231 unique site patterns.
Branch rate model used: discretizedBranchRates
Creating swap operator for parameter branchRates.categories (weight=10.0)
Likelihood is using -1 threads.
Creating the MCMC chain:
  chainLength=800000
  autoOptimize=true
  autoOptimize delayed for 8000 steps
# BEAST v1.6.1, Build r3651
# Generated Thu Apr 21 16:27:15 NZST 2011 [seed=1303338739767]
state Posterior    Prior        Likelihood   rootHeight   ucld.mean
0 -115044.0238 -105244.7369 -9799.2869   311.739      3.3E-2       -
10000 -6019.0942   -48.1460     -5970.9482   31.2013      1.13161E-2   -
20000 -5932.1777   -53.7968     -5878.3809   58.3426      7.80428E-3   0.11 hours/million states
30000 -5864.1836   -55.6386     -5808.5450   54.7400      9.7303E-3    0.09 hours/million states
40000 -5845.2874   -58.9120     -5786.3753   82.1531      6.86446E-3   0.08 hours/million states
50000 -5798.2689   -60.2217     -5738.0472   87.4058      7.84354E-3   0.08 hours/million states
60000 -5779.7566   -57.1298     -5722.6268   49.1015      1.17545E-2   0.07 hours/million states
70000 -5787.4056   -60.5876     -5726.8180   81.5675      1.03161E-2   0.07 hours/million states
80000 -5785.8794   -58.9768     -5726.9027   70.4040      9.80333E-3   0.07 hours/million states
90000 -5778.6354   -59.2618     -5719.3737   77.9236      1.0054E-2    0.07 hours/million states
100000 -5780.4106   -60.2610     -5720.1496   78.7802      8.80928E-3   0.07 hours/million states

... ...

790000 -5772.8359   -59.0540     -5713.7819   77.1089      9.1281E-3    0.06 hours/million states
800000 -5774.2917   -57.2419     -5717.0498   61.2605      9.77329E-3   0.06 hours/million states

Operator analysis
Operator                                          Tuning   Count      Time     Time/Op  Pr(accept)  Performance suggestion
scale(firsthalf.kappa)                            0.536   701        232      0.33     0.2596      good
firsthalf.frequencies                             0.059   677        206      0.3      0.2851      good
scale(firsthalf.alpha)                            0.598   719        220      0.31     0.3255      good
scale(secondhalf.kappa)                           0.555   685        222      0.32     0.3109      good
secondhalf.frequencies                            0.058   661        216      0.33     0.3238      good
scale(secondhalf.alpha)                           0.598   680        210      0.31     0.3015      good
scale(ucld.mean)                                  0.697   21192      7410     0.35     0.2786      good
scale(ucld.stdev)                                 0.274   21350      7342     0.34     0.3526      good
subtreeSlide(treeModel)                           4.405   105795     18763    0.18     0.337       good
Narrow Exchange(treeModel)                                106685     19150    0.18     0.0004      very low
Wide Exchange(treeModel)                                  21301      2071     0.1      0.0         very low
wilsonBalding(treeModel)                                  21229      3431     0.16     0.0         very low
scale(treeModel.rootHeight)                       0.853   21224      1334     0.06     0.218       good
uniform(nodeHeights(treeModel))                           213114     43645    0.2      0.222       good
scale(yule.birthRate)                             0.268   21594      697      0.03     0.2761      good
up:ucld.mean down:nodeHeights(treeModel)          0.596   21448      7551     0.35     0.2417      slightly high
Try setting scaleFactor to about 0.586
swapOperator(branchRates.categories)                      70928      16433    0.23     0.6553      high No suggestions
randomWalkInteger(branchRates.categories)                 70818      13509    0.19     0.9432      very high
Try increasing windowSize to about 2.0
uniformInteger(branchRates.categories)                    71199      14008    0.2      0.7518      high

3.028016666666667 minutes
```

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 `Primates.log`. You should now see a window like the following:



Tree File	States	Burn-In
primates.log	800000	80000

Statistic	Mean	ESS
posterior	-5779.134	108.052
prior	-56.544	223.761
likelihood	-5722.589	99.433
treeModel.rootHeight	50.481	127.412
tmrca(Human-Chimp)	5.471	490.688
tmrca(ingroup)	48.14	135.352
tmrca(HomiCerco)	24.08	2852.466
yule.birthRate	4.626E-2	918.215
firsthalf.kappa	12.98	29.543
firsthalf.frequencies1	0.349	23.77
firsthalf.frequencies2	0.35	18.336
firsthalf.frequencies3	7.846E-2	30.335
firsthalf.frequencies4	0.222	51.343
firsthalf.alpha	0.331	109.969
secondhalf.kappa	11.14	61.774
secondhalf.frequencies1	0.369	29.456
secondhalf.frequencies2	0.294	41.068
secondhalf.frequencies3	8.649E-2	47.582
secondhalf.frequencies4	0.251	44.117
secondhalf.alpha	0.384	117.536
ucld.mean	1.347E-2	125.399
ucld.stdev	0.244	214.978
meanRate	1.291E-2	106.299
coefficientOfVariation	0.222	225.309
covariance	-6.28E-2	1653.117
firsthalf.treeLikelihood	-2894.565	130.594
secondhalf.treeLikelihood	-2828.024	109.697
speciation	-46.219	181.122

Summary Statistic	
mean	-5779.1336
stderr of mean	0.4624
median	-5778.7087
geometric mean	n/a
95% HPD lower	-5787.7678
95% HPD upper	-5770.7382
auto-correlation time (ACT)	6665.33
effective sample size (ESS)	108.0517

Remember that MCMC is a stochastic algorithm so the actual numbers will not be exactly the same.

On the left hand side is a list of the different quantities that BEAST has logged. There are traces for the posterior (this is the log of the product of the tree likelihood and the prior probabilities), and the continuous parameters. Selecting a trace on the left brings up analyses for this trace on the right hand side depending on tab that is selected. When first opened, the 'posterior' trace is selected and various statistics of this trace are shown under the Estimates tab. In the top right of the window is a table of calculated statistics for the selected trace.

Select `meanRate` to look at the rate of evolution averaged over the whole tree. Tracer will plot a (marginal posterior) distribution for the selected parameter and also give you statistics such as the mean and median. The 95% HPD stands for *highest posterior density interval* and represents the most compact interval on the selected parameter that contains 95% of the posterior probability. It can be thought of as a Bayesian analog to a confidence interval.

Questions

What is the rate of molecular evolution in Primates (include the HPD interval)?

What sources of error does this estimate include?

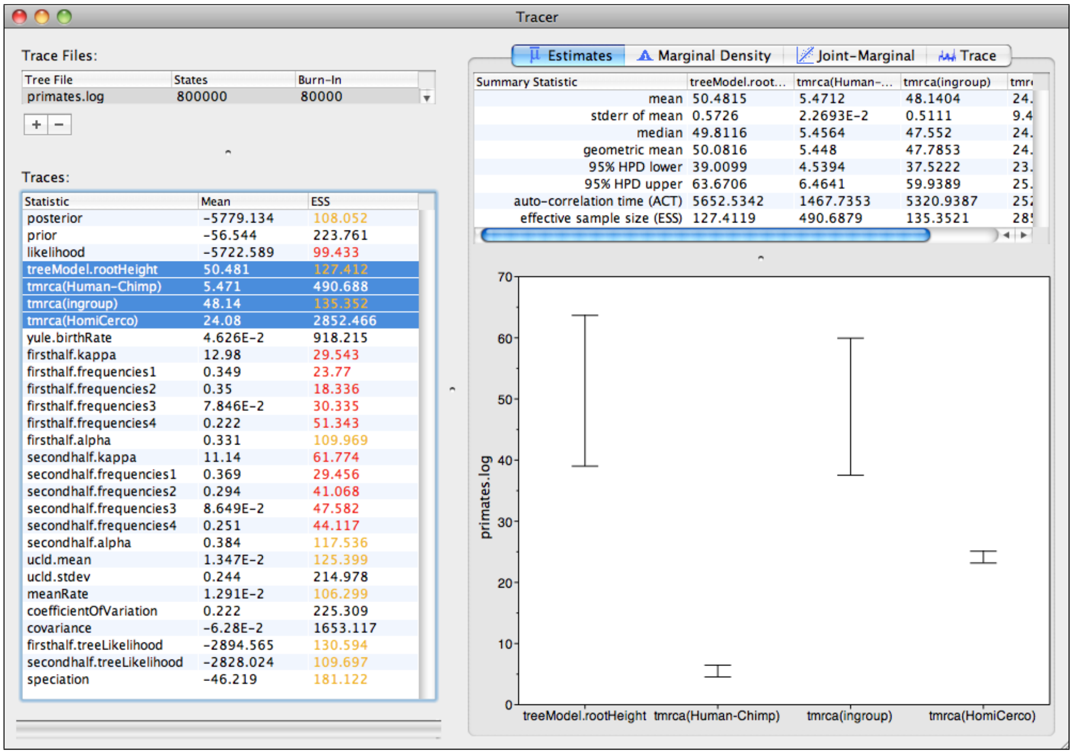
The `coefficientOfVariation` statistic gives a summary of how much the rate of evolution varies from lineage to lineage (expressed as a proportion of the mean rate).

Does the rate of evolution differ substantially amongst different lineages in the tree?

Selecting the `treeModel.rootHeight` parameter gives the marginal posterior distribution of the age of the root of entire tree.

How old is the root of the tree (give the mean and the HPD range)?

Select the `treeModel.rootHeight` parameter and the next three (hold shift whilst selecting). This will show a display of the age of the root and the three MRCAs we specified in BEAUti. The parameter that we used to calibrate the tree (`tmrca(human-chimp)`) will have posterior distributions very similar to the prior distributions that we specified.



Tracer

Trace Files:

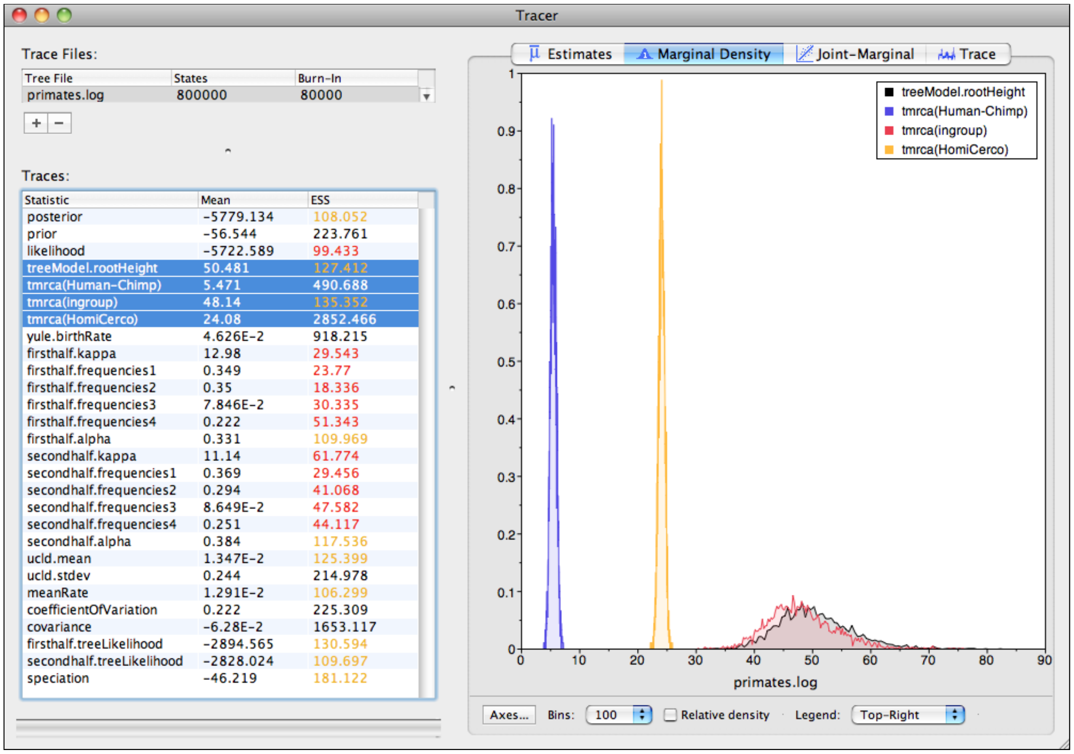
Tree File	States	Burn-In
primates.log	800000	80000

Traces:

Statistic	Mean	ESS
posterior	-5779.134	108.052
prior	-56.544	223.761
likelihood	-5722.589	99.433
treeModel.rootHeight	50.481	127.412
tmrca(Human-Chimp)	5.471	490.688
tmrca(ingroup)	48.14	135.352
tmrca(HomiCerco)	24.08	2852.466
yule.birthRate	4.626E-2	918.215
firsthalf.kappa	12.98	29.543
firsthalf.frequencies1	0.349	23.77
firsthalf.frequencies2	0.35	18.336
firsthalf.frequencies3	7.846E-2	30.335
firsthalf.frequencies4	0.222	51.343
firsthalf.alpha	0.331	109.969
secondhalf.kappa	11.14	61.774
secondhalf.frequencies1	0.369	29.456
secondhalf.frequencies2	0.294	41.068
secondhalf.frequencies3	8.649E-2	47.582
secondhalf.frequencies4	0.251	44.117
secondhalf.alpha	0.384	117.536
ucld.mean	1.347E-2	125.399
ucld.stdev	0.244	214.978
meanRate	1.291E-2	106.299
coefficientOfVariation	0.222	225.309
covariance	-6.28E-2	1653.117
firsthalf.treeLikelihood	-2894.565	130.594
secondhalf.treeLikelihood	-2828.024	109.697
speciation	-46.219	181.122

Summary Statistic	treeModel.root...	tmrca(Human-...	tmrca(ingroup)	tmr
mean	50.4815	5.4712	48.1404	24.
stderr of mean	0.5726	2.2693E-2	0.5111	9.4
median	49.8116	5.4564	47.552	24.
geometric mean	50.0816	5.448	47.7853	24.
95% HPD lower	39.0099	4.5394	37.5222	23.
95% HPD upper	63.6706	6.4641	59.9389	25.
auto-correlation time (ACT)	5652.5342	1467.7353	5320.9387	252
effective sample size (ESS)	127.4119	490.6879	135.3521	285

If you switch the tab at the top of the window to **Marginal Density** then you will get a plot of the marginal posterior densities of each of these date estimates overlayed:



Tracer

Trace Files:

Tree File	States	Burn-In
primates.log	800000	80000

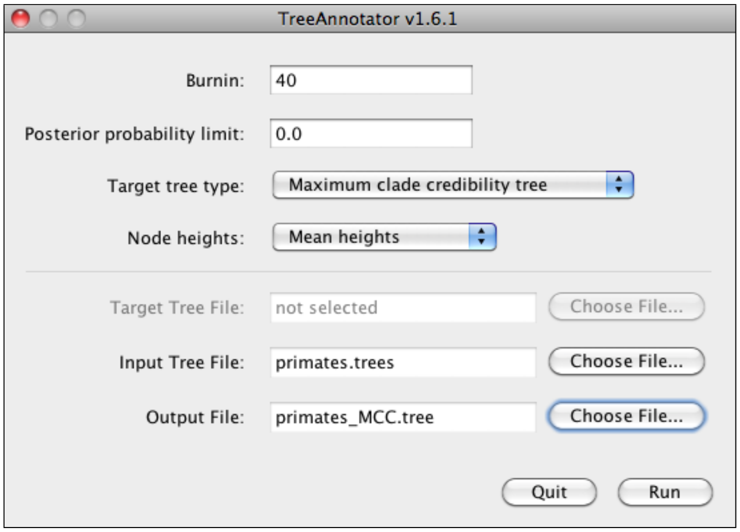
Traces:

Statistic	Mean	ESS
posterior	-5779.134	108.052
prior	-56.544	223.761
likelihood	-5722.589	99.433
treeModel.rootHeight	50.481	127.412
tmrca(Human-Chimp)	5.471	490.688
tmrca(ingroup)	48.14	135.352
tmrca(HomiCerco)	24.08	2852.466
yule.birthRate	4.626E-2	918.215
firsthalf.kappa	12.98	29.543
firsthalf.frequencies1	0.349	23.77
firsthalf.frequencies2	0.35	18.336
firsthalf.frequencies3	7.846E-2	30.335
firsthalf.frequencies4	0.222	51.343
firsthalf.alpha	0.331	109.969
secondhalf.kappa	11.14	61.774
secondhalf.frequencies1	0.369	29.456
secondhalf.frequencies2	0.294	41.068
secondhalf.frequencies3	8.649E-2	47.582
secondhalf.frequencies4	0.251	44.117
secondhalf.alpha	0.384	117.536
ucld.mean	1.347E-2	125.399
ucld.stdev	0.244	214.978
meanRate	1.291E-2	106.299
coefficientOfVariation	0.222	225.309
covariance	-6.28E-2	1653.117
firsthalf.treeLikelihood	-2894.565	130.594
secondhalf.treeLikelihood	-2828.024	109.697
speciation	-46.219	181.122

Bins: 100
□ Relative density
Legend: Top-Right

Obtaining an estimate of the phylogenetic tree

BEAST also produces a sample of plausible trees along with its sample of parameter estimates. These need to be summarized using the program **TreeAnnotator** (see Notes for details). This will take the set of trees and find the best supported one. It will then annotate this summary tree with the mean ages of all the nodes and the HPD ranges. It will also calculate the posterior clade probability for each node. Run the TreeAnnotator program and set it up to look like this:



TreeAnnotator v1.6.1

Burnin: 40

Posterior probability limit: 0.0

Target tree type: Maximum clade credibility tree

Node heights: Mean heights

Target Tree File: not selected

Input Tree File: primates.trees

Output File: primates_MCC.tree

The burnin is the number of trees to remove from the start of the sample. Unlike **Tracer** which specifies the number of steps as a burnin, in **TreeAnnotator** you need to specify the actual number of trees. For this run, you specified a chain length of 800,000 steps sampling every 200 steps. Thus the trees file will contain 4000 trees and so to specify a 1% burnin use the value 40.

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. The default of 0.5 means that only nodes seen in the majority of trees will be annotated. Set this to zero to annotate all nodes.

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.

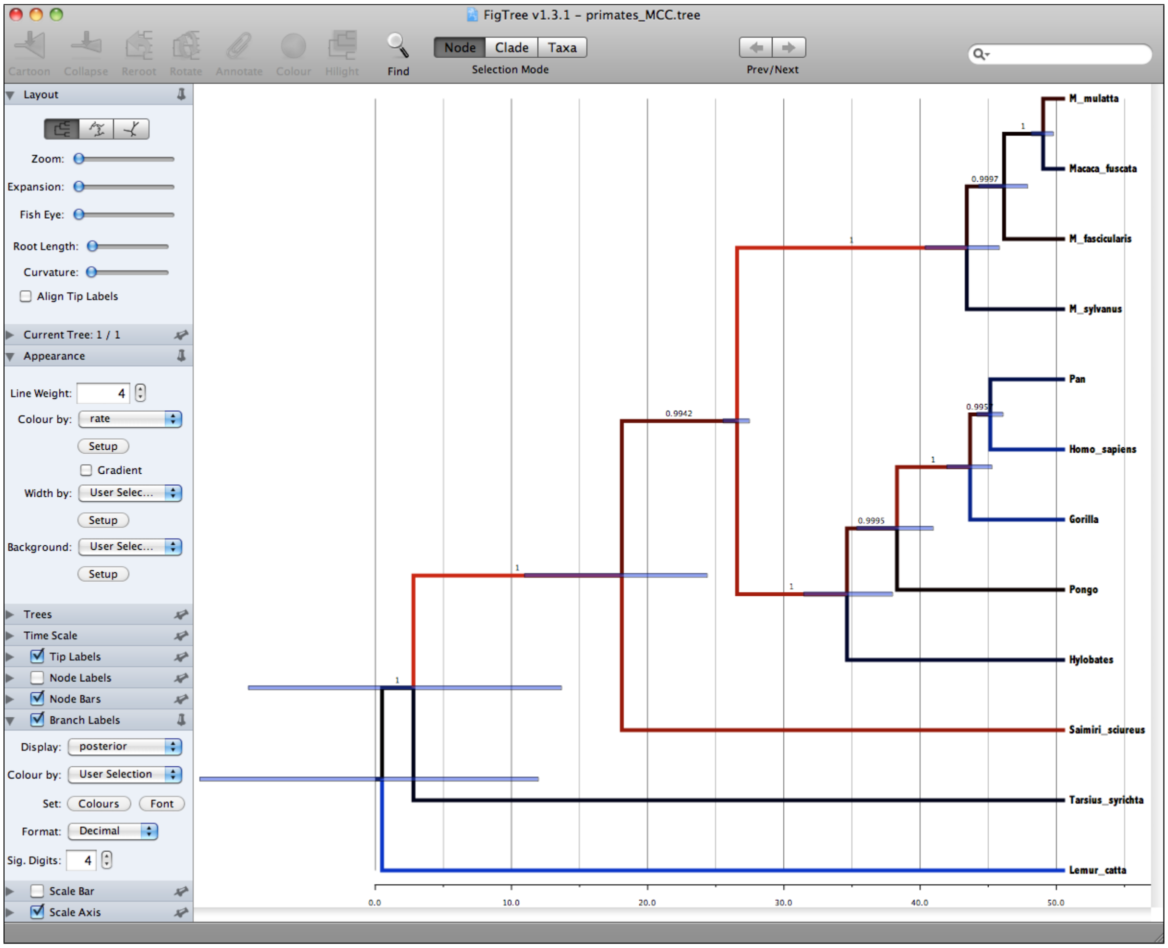
Choose **Mean heights** for node heights. This sets the heights (ages) of each node in the tree to the mean height across the entire sample of trees for that clade.

For the input file, select the trees file that BEAST created (by default this will be called `Primates.trees`) and select a file for the output (here we called it `Primates_MCC.tree`).

Now press Run and wait for the program to finish.

Viewing the Tree

Finally, we can look at the tree in another program called **FigTree**. Run this program, and open the `Primates.MCC.tree` file by using the Open command in the File menu. The tree should appear. You can now try selecting some of the options in the control panel on the left. Try selecting **Node Bars** to get node age error bars. Also turn on **Branch Labels** and select **posterior** to get it to display the posterior probability for each node. Under **Appearance** you can also tell FigTree to colour the branches by the rate. You should end up with something like this:



FigTree v1.3.1 – primates_MCC.tree

Align Tip Labels ☐

Current Tree: 1 / 1

Line Weight: 4

Colour by: rate

☐ Gradient

Width by: User Selec...

Background: User Selec...

☑ Tip Labels

☐ Node Labels

☑ Node Bars

☑ Branch Labels

Display: posterior

Colour by: User Selection

Format: Decimal

Sig. Digits: 4

☐ Scale Bar

☑ Scale Axis

Which branch has the fastest rate of evolution and what is the estimated rate?

Which branch has the slowest rate of evolution and what is the estimated rate?

Are these two rate estimates significantly different? How would you answer this question?

Comparing your results to the prior

Using BEAUti, set up the same analysis but under the MCMC options, select the **Sample from prior only** option. This will allow you to visualize the full prior distribution in the absence of your sequence data. Summarize the trees from the full prior distribution and compare the summary to the posterior summary tree.