

Estimating rates and dates from time-stamped sequences

A hands-on practical

This chapter provides a step-by-step tutorial for analyzing a set of virus sequences which have been isolated at different points in time (heterochronous data). The data are 71 sequences from the *prM/E* gene of yellow fever virus (YFV) from African and South American countries with isolation dates ranging from 1940-2009. The sequences represent a subset of the data set analyzed by Bryant *et al.* (Bryant JE, Holmes EC, Barrett ADT, 2007 Out of Africa: A Molecular Perspective on the Introduction of Yellow Fever Virus into the Americas. PLoS Pathog 3(5): e75. doi:10.1371/journal.ppat.0030075).

The most commonly cited hypothesis of the origin of yellow fever virus (YFV) in the Americas is that the virus was introduced from Africa, along with *Aedes aegypti* mosquitoes, in the bilges of sailing vessels during the slave trade. Although the hypothesis of a slave trade introduction is often repeated, it has not been subject to rigorous examination using gene sequence data and modern phylogenetic techniques for estimating divergence times. The aim of this exercise is to obtain an estimate of the rate of molecular evolution, an estimate of the date of the most recent common ancestor and to infer the phylogenetic relationships with appropriate measures of statistical support.

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.

To undertake this tutorial, you will need to download three software packages in a format that is compatible with your computer system (all three are available for Mac OS X, Windows and Linux/UNIX operating systems):

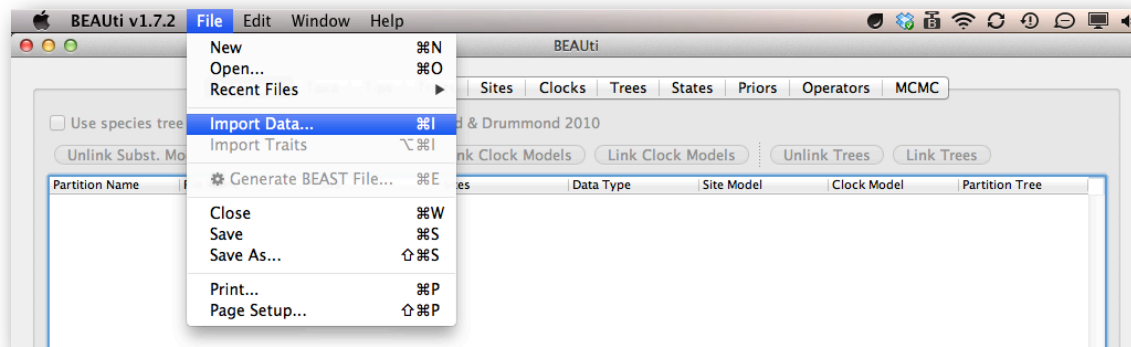
- **BEAST** - this package contains the BEAST program, BEAUti and a couple of utility programs. At the time of writing, the current version is v1.7.2. It is available for download from <http://beast.bio.ed.ac.uk/>.
- **Tracer** - this program is used to explore the output of **BEAST** (and other Bayesian MCMC programs). It graphically and quantitatively summarizes the empirical 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/>.

Running BEAUti

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

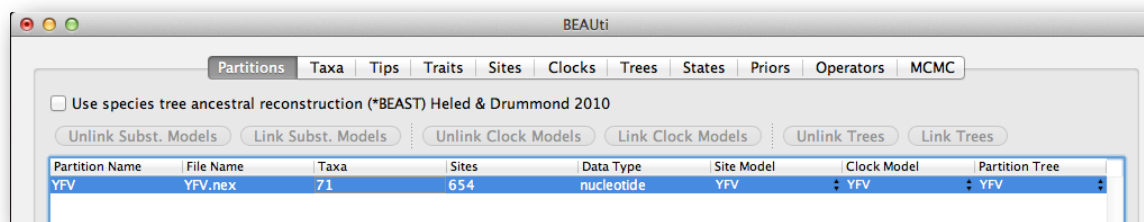
Loading the NEXUS file

To load a NEXUS format alignment, simply select the **Import Data...** option from the **File** menu.

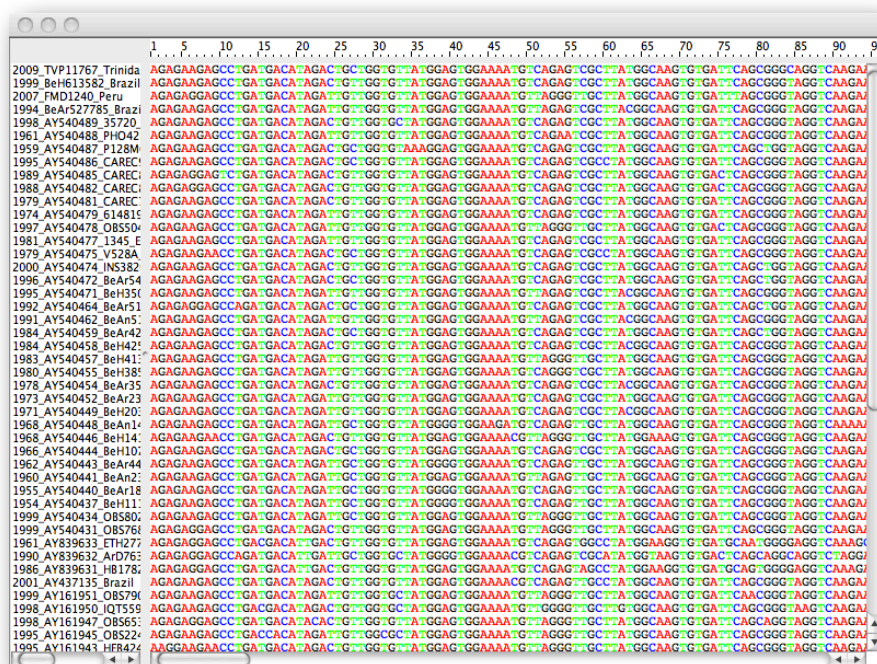


The NEXUS alignment

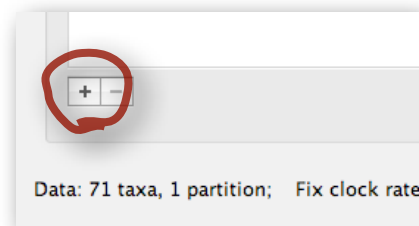
Select the file called **YFV.nex**. This file contains an alignment of 71 sequences from the *prM/E* gene of YFV, 654 nucleotides in length. Once loaded, the sequence data will be listed under **Data Partitions**:



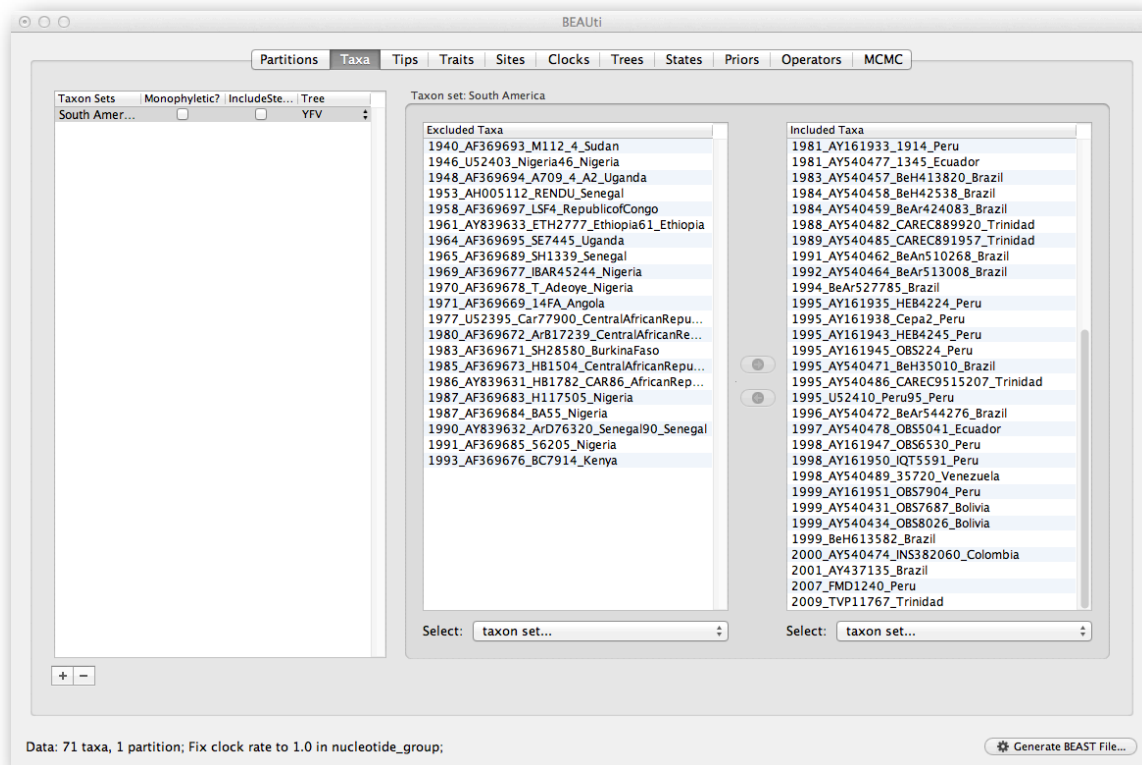
Double clicking on the YFV partition (or on "Show"), will display the alignment in a separate window:



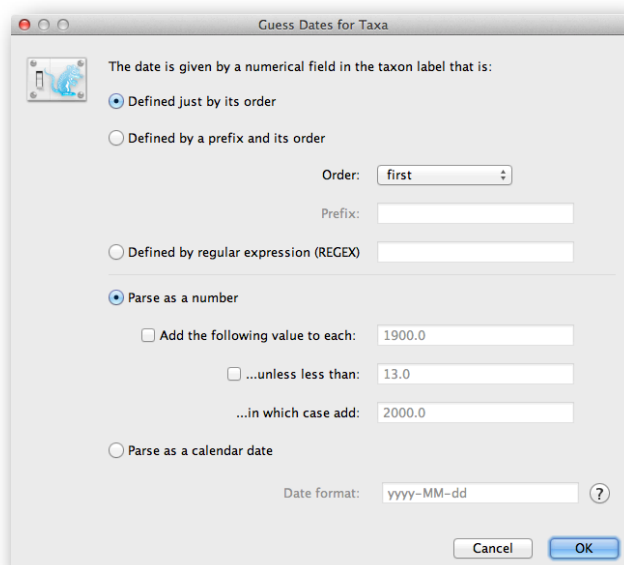
Under the Taxon Sets menu, we can define sets of taxa for which we would like to obtain particular statistics, enforce a monophyletic constraint, or put calibration information on. Let's define a "South America" taxon set by pressing the small "plus" button at the bottom left of the panel:



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 **South America**. Do not enforce monophyly using the "monophyletic?" option because we will evaluate the support for this cluster. We do not opt for the "includeStem?" option either because we would like to estimate the TRMCA for the South American viruses and not for the parent node leading to this clade. In the next table along you will see the available taxa. Select a South American taxon such as **1954_AY540437_BeH111_Brazil** and press the green arrow button to move it into the included taxa set. Repeat this until all South American taxa are included. Note that multiple taxa can be selected simultaneously holding down the cmd/ctrl button on a Mac/PC. The screen should look like this:



To inform BEAUti/BEAST about the sampling dates of the sequences, go to the **Tips** menu and select the "Use tip dates" option. By default all the taxa are assumed to have a date of zero (i.e. the sequences are assumed to be sampled at the same time). In this case, the YFV sequences have been sampled at various dates going back to the 1940s. The actual year of sampling is given in the name of each taxon and we could simply edit the value in the Date column of the table to reflect these. However, if the taxa names contain the calibration information, then a convenient way to specify the dates of the sequences in BEAUti is to use the "Guess Dates" button at the top of the Data panel. Clicking this will make a dialog box appear:



This operation attempts to guess what the dates are from information contained within the taxon names. It works by trying to find a numerical field within each name. If the taxon names contain more than one numerical field (such as the some YFV sequences, above) then you can specify how to find the one that corresponds to the date of sampling. You can (1) specify the order that the date field comes (e.g., first, last or various positions in between) or (2) specify a prefix (some characters that come immediately before the date field in each name) and the order of the field, or (3) define a regular expression (REGEX). For the YFV sequences you can keep the default ‘Defined just by its order’ and ‘Order: first’.

When parsing a number, you can ask BEAUti to add a fixed value to each guessed date. For example, the value “1900” can be added to turn the dates from 2 digit years to 4 digit. Any dates in the taxon names given as “00” would thus become “1900”. However, if these ‘00’ or ‘01’, etc. represent sequences sampled in 2000, 2001, etc., ‘2000’ needs to be added to those. This can be achieved by selecting the “unless less than: ..” and “..in which case add:..” option adding for example 2000 to any date less than 10. There is also an option to parse calendar dates. Both are not necessary in our case as all dates are specified in a four digit format. So, we can press “OK”. At the top of the window you can set the units that the dates are given in (years, months, days) and whether they are specified relative to a point in the past (as would be the case for years such as 1984) or backwards in time from the present (as in the case of radiocarbon ages).

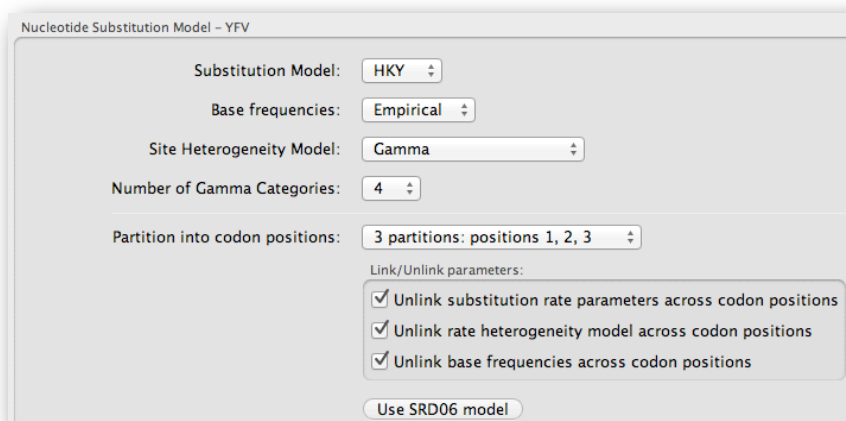
Name	Date	Height
2009_TVP11767_Trinidad	2009.0	0.0
1999_BeH613582_Brazil	1999.0	10.0
2007_FMD1240_Peru	2007.0	2.0
1994_BeAr527785_Brazil	1994.0	15.0
1998_AY540489_35720_Venezuela	1998.0	11.0
1961_AY540488_PHO42H_Venezuela	1961.0	48.0
1959_AY540487_P128MC_Venezuela	1959.0	50.0
1995_AY540486_CAREC9515207_Trinidad	1995.0	14.0
1989_AY540485_CAREC891957_Trinidad	1989.0	20.0
1988_AY540482_CAREC889920_Trinidad	1988.0	21.0
1979_AY540481_CAREC797984_Trinidad	1979.0	30.0
1974_AY540479_614819_Panama	1974.0	35.0

Setting the evolutionary model

The next thing to do is to click on the **Sites** 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 traits). This tutorial assumes that you are familiar with the evolutionary models available; however there are a couple of points to note about selecting a model in BEAUti:

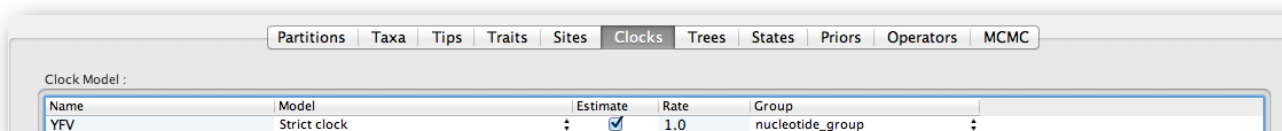
- Selecting the **Partition into codon positions** option assumes that the data are aligned as codons. This option will then estimate a separate rate of substitution for each codon position, or for 1+2 versus 3, depending on the setting.
- Selecting the **Unlink substitution model across codon positions** will specify that BEAST should estimate a separate transition-transversion ratio or general time reversible rate matrix for each codon position.
- Selecting the **Unlink rate heterogeneity model across codon positions** will specify that BEAST should estimate set of rate heterogeneity parameters (gamma shape parameter and/or proportion of invariant sites) for each codon position.

For this tutorial, select the **3 partitions: codon positions 1, 2 & 3** option so that each codon position has its own rate of evolution, **Estimated** base frequencies, and **Gamma**-distributed rate variation among sites:



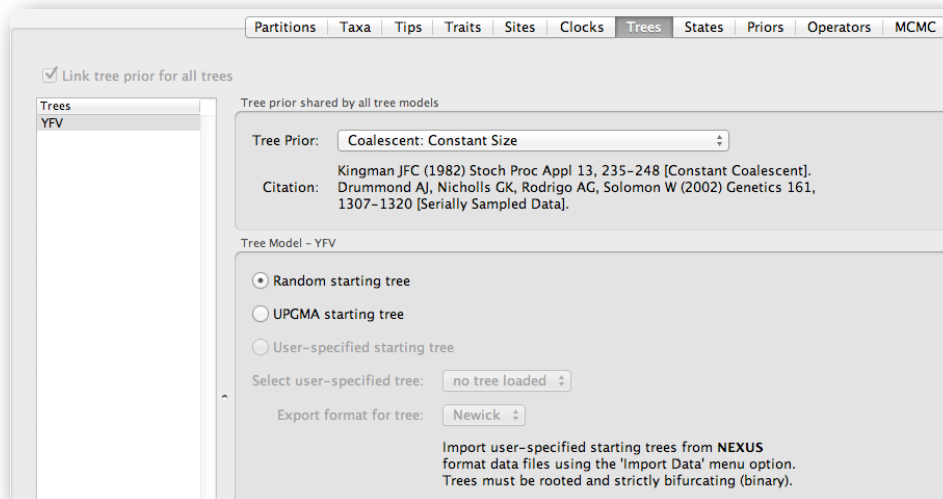
Setting the clock model

Click on the **Clocks** tab at the top of the main window. We will perform our initial run using the (default) strict molecular clock model:



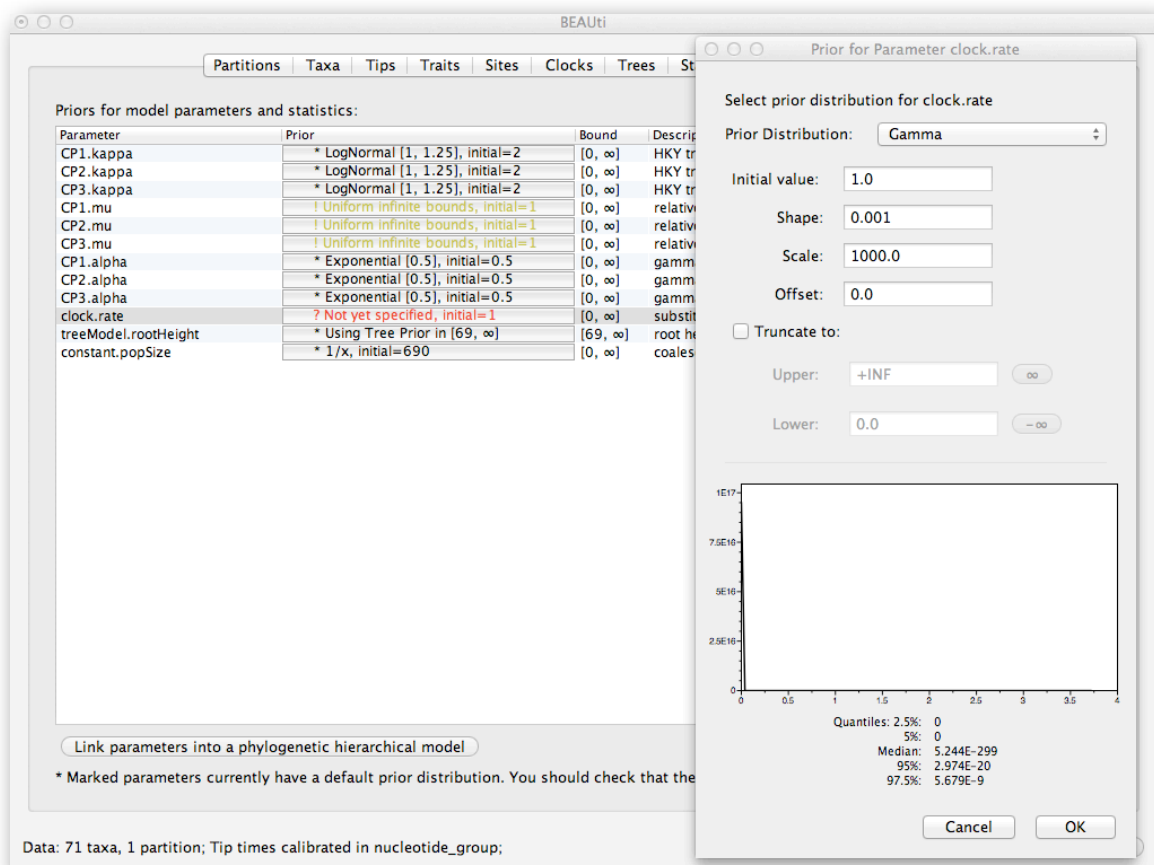
Setting the starting tree and tree prior

Click on the **Trees** tab at the top of the main window. We keep a default random starting tree and a constant size coalescent prior. The tree priors (coalescent and other models) will be explained in other lectures.



Setting up the priors

Review the prior settings under the **Priors** tab. Although some of the default marginal priors are improper (e.g. indicated in yellow), with sufficiently informative data the posterior becomes proper. Priors that have not been set yet appear in red (e.g. clock.rate). Click on the prior for this parameter and a prior selection window will appear. Set the prior to a gamma distribution with shape = 0.001 and scale = 1000. The graphical representation of this prior distribution indicates that most prior mass is put on small values, but the density remains sufficiently diffuse. Notice that the prior setting turns black after confirming this setting by clicking "OK".

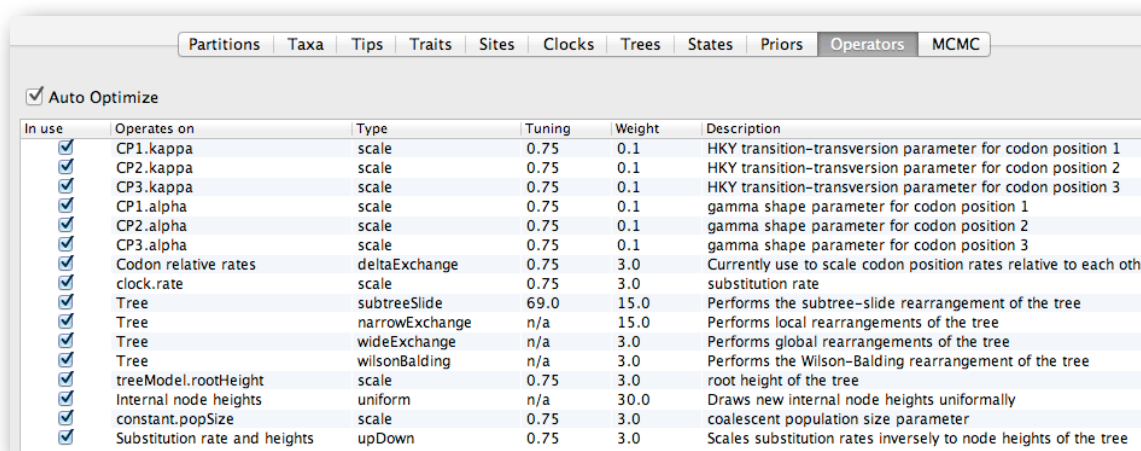


Setting up the operators

Each parameter in the model has one or more “operators” (these are variously called moves and proposals by other MCMC software packages such as MrBayes and LAMARC). The operators specify how the parameters change as the MCMC runs. The operators tab in BEAUti has a table that lists the parameters, their operators and the tuning settings for these operators. In the first column are the parameter names. These will be called things like **CP1.kappa** which means the HKY model's kappa parameter (the transition-transversion bias) for the first codon position. The next column has the type of operators that are acting on each parameter. For example, the scale operator scales the parameter up or down by a random proportion and the uniform operator simply picks a new value uniformly within a range. Some parameters relate to the tree or to the divergence times of the nodes of the tree and these have special operators.

The next column, labelled **Tuning**, gives a tuning setting to the operator. Some operators don't have any tuning settings so have **n/a** under this column. The tuning parameter will determine how large a move each operator will make which will affect how often that change is accepted by the MCMC which will affect the efficiency of the analysis. For most operators (like the subtree slide operator) a larger tuning parameter means larger moves. However for the scale operator a tuning parameter value closer to 0.0 means bigger moves. At the top of the window is an option called **Auto Optimize** which, when selected, will automatically adjust the tuning setting as the MCMC runs to try to achieve maximum efficiency. At the end of the run a table of the operators, their performance and the final values of these tuning settings can be written to standard output.

The next column, labelled **Weight**, specifies how often each operator is applied relative to the others. Some parameters tend to be sampled very efficiently - an example is the kappa parameter - these parameters can have their operators down-weighted so that they are not changed as often.



The screenshot shows the BEAUti interface with the 'Operators' tab selected. A checkbox for 'Auto Optimize' is checked. Below it is a table with the following columns: In use, Operates on, Type, Tuning, Weight, and Description.

In use	Operates on	Type	Tuning	Weight	Description
<input checked="" type="checkbox"/>	CP1.kappa	scale	0.75	0.1	HKY transition-transversion parameter for codon position 1
<input checked="" type="checkbox"/>	CP2.kappa	scale	0.75	0.1	HKY transition-transversion parameter for codon position 2
<input checked="" type="checkbox"/>	CP3.kappa	scale	0.75	0.1	HKY transition-transversion parameter for codon position 3
<input checked="" type="checkbox"/>	CP1.alpha	scale	0.75	0.1	gamma shape parameter for codon position 1
<input checked="" type="checkbox"/>	CP2.alpha	scale	0.75	0.1	gamma shape parameter for codon position 2
<input checked="" type="checkbox"/>	CP3.alpha	scale	0.75	0.1	gamma shape parameter for codon position 3
<input checked="" type="checkbox"/>	Codon relative rates	deltaExchange	0.75	3.0	Currently use to scale codon position rates relative to each oth.
<input checked="" type="checkbox"/>	clock.rate	scale	0.75	3.0	substitution rate
<input checked="" type="checkbox"/>	Tree	subtreeSlide	69.0	15.0	Performs the subtree-slide rearrangement of the tree
<input checked="" type="checkbox"/>	Tree	narrowExchange	n/a	15.0	Performs local rearrangements of the tree
<input checked="" type="checkbox"/>	Tree	wideExchange	n/a	3.0	Performs global rearrangements of the tree
<input checked="" type="checkbox"/>	Tree	wilsonBalding	n/a	3.0	Performs the Wilson-Balding rearrangement of the tree
<input checked="" type="checkbox"/>	treeModel.rootHeight	scale	0.75	3.0	root height of the tree
<input checked="" type="checkbox"/>	Internal node heights	uniform	n/a	30.0	Draws new internal node heights uniformly
<input checked="" type="checkbox"/>	constant.popSize	scale	0.75	3.0	coalescent population size parameter
<input checked="" type="checkbox"/>	Substitution rate and heights	upDown	0.75	3.0	Scales substitution rates inversely to node heights of the tree

Setting the MCMC options

The **MCMC** tab in BEAUti provides settings to control the MCMC chain. 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 depend on the size of the dataset, the complexity of the model and the precision of the answer required. The default value of 10,000,000 is entirely arbitrary and should be adjusted according to the size of your dataset. We will see later how the resulting log file can be analyzed using Tracer in order to examine whether a particular chain length is adequate.

The next couple of options specify how often the current parameter values should be displayed on the screen and recorded in the log file. The screen output is simply for monitoring the program's progress so can be set to any value (although if set too small, the sheer quantity of information being displayed on the screen will 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 estimates. 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 something $\geq \text{chain length} / 10,000$.

For this dataset let's initially set the chain length to 100,000 as this will run reasonably quickly on most modern computers. Although the suggestion above would indicate a lower sampling frequency, in this case set both the sampling frequencies to 100.

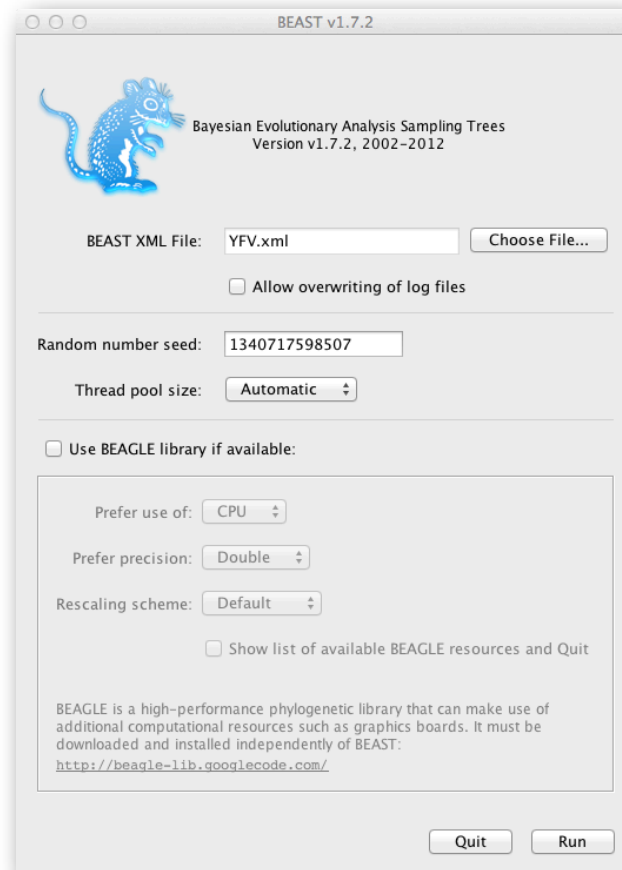
The next option allows the user to set the File stem name; if not set to 'YFV' by default, you can type this in here. The next two options give the file names of the log files for the parameters and the trees. These will be set to a default based on the file stem name. Let's also create an operator analysis file by selecting the relevant option. Finally, an option is available to sample from the prior only, which can be useful to evaluate how divergent our posterior estimates are when information is drawn from the data. Here, we will not select this option, but analyze the actual data.

At this point we are ready to generate a BEAST XML file and to use this to run the Bayesian evolutionary analysis. 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. BEAUti will ask you to review the prior settings one more time before saving the file (and indicate that some are improper). Continue and choose a name for the file (for example, **YFV.xml**) and save the file..

For convenience, leave the BEAUti window open so that you can change the values and re-generate the BEAST file as required later in this tutorial.

Running BEAST

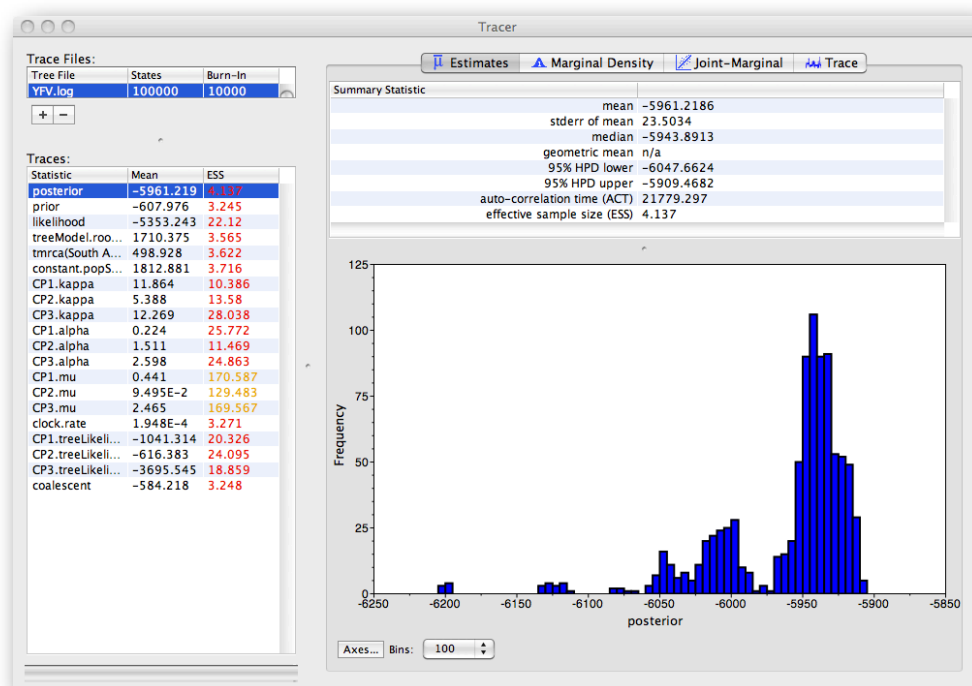
Once the BEAST XML file has been created the analysis itself can be performed using BEAST. The exact instructions for running BEAST depends on the computer you are using, but in most cases a standard file dialog box will appear in which you select the XML file: If the command line version is being used then the name of the XML file is given after the name of the BEAST executable. When pressing "Run", the analysis will be performed with detailed information about the progress of the run being written to the screen. When it has finished, the log file and the trees file will have been created in the same location as your XML file.



Analysing the BEAST output

To analyze the results of running BEAST we are going to use the program **Tracer**. The exact instructions for running Tracer differs depending on which computer you are using. Please see the README text file that was distributed with the version you downloaded. Once running, Tracer will look similar irrespective of which computer system it is running on.

Select the **Import Trace File...** option from the **File** menu. If you have it available, select the log file that you created in the previous section. The file will load and you will be presented with a window similar to the one below. Remember that MCMC is a stochastic algorithm so the actual numbers will not be exactly the same.



On the left hand side is the name of the log file loaded and the traces that it contains. There are traces for a quantity proportional to posterior (this is the product of the data likelihood and the prior probabilities, on the log-scale), 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. The statistics and their meaning are described in the table below.

Mean - The mean value of the samples (excluding the burn-in).

Stdev of mean - The standard error of the mean. This takes into account the effective sample size so a small ESS will give a large standard error.

Median - The median value of the samples (excluding the burn-in).

Geometric mean - The central tendency or typical value of the set of samples (excluding the burn-in).

95% HPD Lower - The lower bound of the highest posterior density (HPD) interval. The HPD is the shortest interval that contains 95% of the sampled values.

95% HPD Upper - The upper bound of the highest posterior density (HPD) interval.

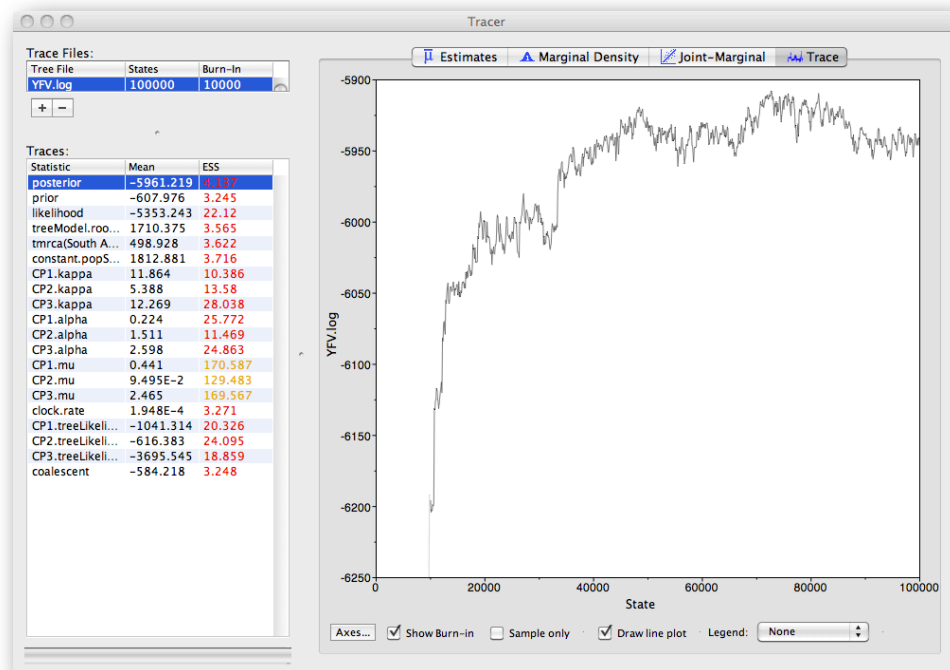
Auto-Correlation Time (ACT) - The average number of states in the MCMC chain that two samples have to be separated by for them to be uncorrelated (i.e. independent samples from the posterior). The ACT is estimated from the samples in the trace (excluding the burn-in).

Effective Sample Size (ESS) - The effective sample size (ESS) is the number of independent samples that the trace is equivalent to. This is calculated as the chain length (excluding the burn-in) divided by the ACT.

Note that the effective sample sizes (ESSs) for all the traces are small (ESSs less than 100 are highlighted in red by Tracer and values > 100 but < 200 are in yellow). This is not good. A low ESS means that the trace contained a lot of correlated

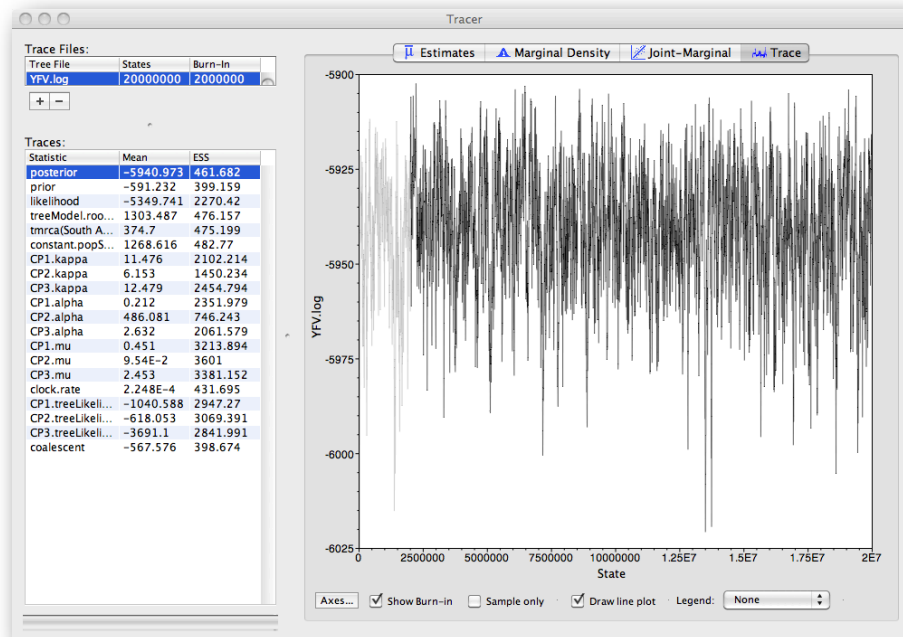
samples and thus may not represent the posterior distribution well. In the bottom right of the window is a frequency plot of the samples which is expected given the low ESSs is extremely rough.

If we select the tab on the right-hand-side labelled 'Trace' we can view the raw trace, that is, the sampled values against the step in the MCMC chain.



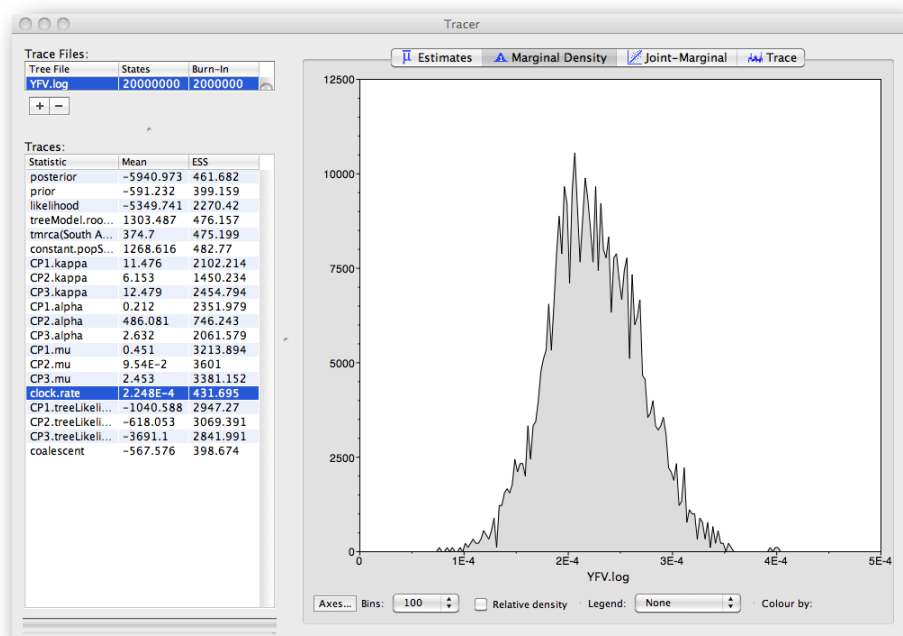
Here you can see how the samples are correlated. There are 1000 samples in the trace (we ran the MCMC for 100,000 steps sampling every 100) but it is clear that adjacent samples often tend to have similar values. The ESS for the age of the root (**treeModel.rootHeight** is about 3.5 so we are only getting 1 independent sample to every 285 actual samples). It also seems that the default burn-in of 10% of the chain length is inadequate (the posterior values are still increasing over most of the chain). Not excluding enough of the start of the chain as burn-in will bias the results and render estimates of ESS unreliable.

The simple response to this situation is that we need to run the chain for longer. Go back to the **MCMC Options** section, above, and create a new BEAST XML file with a longer chain length (e.g. 10,000,000). We can also incorporate the suggestions regarding the operator performance. At the end of the run, BEAST performs an operator analysis and suggests how operator settings could be improved. These suggested modifications can be set in the 'Operators' tab. Now run BEAST and load the new log file into Tracer (you can leave the old one loaded for comparison). To continue the tutorial without having to wait for this run to complete, you can also make use of the log files provided with this tutorial (chain length of 20,000,000 and logged every 5000 sample). Import the log file for the strict clock analysis and click on the **Trace** tab and look at the raw trace plot.



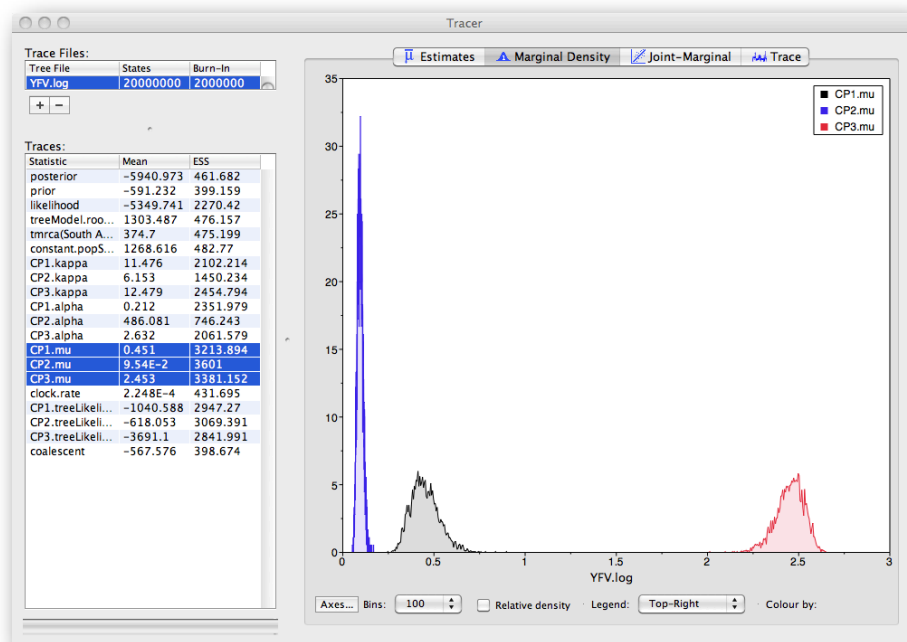
The log file provided contains 4000 samples and with an ESS of about 400 there is still auto-correlation between the samples but 400 effectively independent samples will now provide an reasonably adequate estimate of the posterior distribution. There are no obvious trends in the plot which would suggest that the MCMC has not yet converged, and there are no large-scale fluctuations in the trace which would suggest poor mixing.

As we are happy with the behavior of posterior probability we can now move on to one of the parameters of interest: substitution rate. Select **clock.rate** in the left-hand table. This is the average substitution rate across all sites in the alignment. Now choose the density plot by selecting the tab labeled Density. This shows a plot of the posterior probability density of this parameter. You should see a plot similar to this:

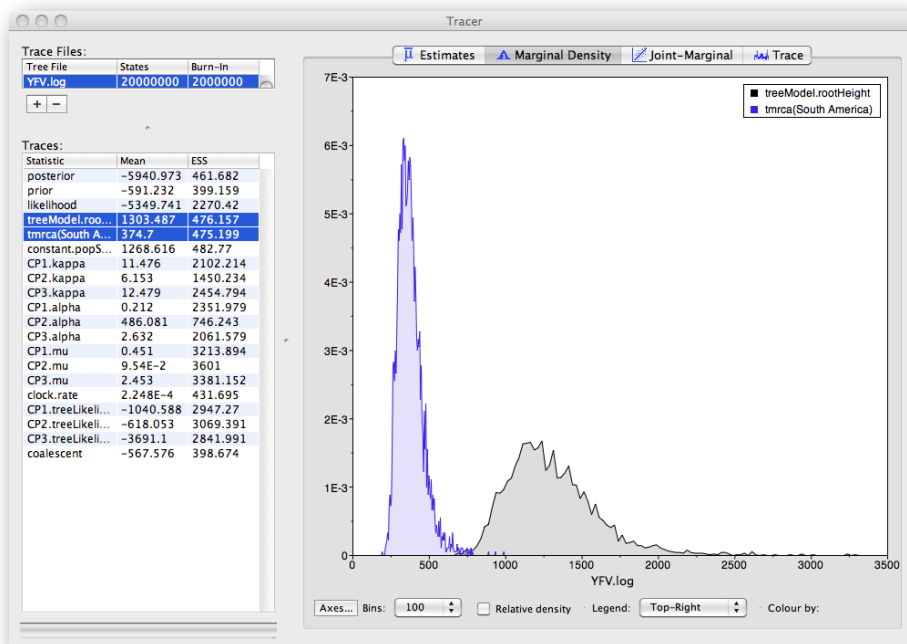


As you can see the posterior probability density is roughly bell-shaped. There is some sampling noise which would be reduced if we ran the chain for longer but we already have a reasonable estimate of the mean and HPD interval. You can overlay the density plots of multiple traces in order to compare them (it is up to the user to determine whether they are comparable on the the same axis or not). Select the relative substitution rates for all three codon positions in the table to the

left (labelled CP1.mu, CP2.mu and CP3.mu. You will now see the posterior probability densities for the relative substitution rate at all three codon positions overlaid:



Note that the three rates are markedly different, what does this tell us about the selective pressure on this gene? Now, let's have a look at the time to the most recent common ancestor (tmrca) for the South American strains relative to the general tmrca:



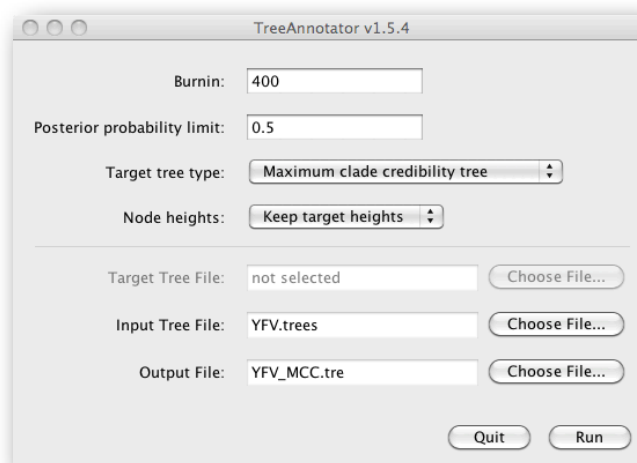
This indicates that the South American tmrca is significantly younger than the root height and argues for more recent origin of YFV in South America.

Summarizing the trees

We have seen how we can diagnose our MCMC run using Tracer and produce estimates of the marginal posterior distributions of parameters of our model. However, BEAST also samples trees (either phylogenies or genealogies) at the same time as the other parameters of the model. These are written to a separate file called the 'trees' file. This file is a standard NEXUS format file. As such it can easily be loaded into other software in order to examine the trees it contains. One

possibility is to load the trees into a program such as PAUP* and construct a consensus tree in a similar manner to summarizing a set of bootstrap trees. In this case, the support values reported for the resolved nodes in the consensus tree will be the posterior probability of those clades.

In this tutorial, however, we are going to use a tool that is provided as part of the BEAST package to summarize the information contained within our sampled trees. The tool is called **TreeAnnotator** and once running, you will be presented with a window like the one below.



TreeAnnotator takes a single 'target' tree and annotates it with the summarized information from the entire sample of trees. The summarized information includes the average node ages (along with the HPD intervals), the posterior support and the average rate of evolution on each branch (for models where this can vary). The program calculates these values for each node or clade observed in the specified 'target' tree.

- **Burnin** - This is the number of trees in the input file that should be excluded from the summarization. This value is given as the number of trees rather than the number of steps in the MCMC chain. Thus for the example above, with a chain of 20,000,000 steps, sampling every 5000 steps, there are 4000 trees in the file. To obtain a 10% burnin, set this value to 400.
- **Posterior probability limit** - This is the minimum posterior probability for a node in order for TreeAnnotator to store the annotated information. The default is 0.5 so only nodes with this posterior probability or greater will have information summarized (the equivalent to the nodes in a majority-rule consensus tree). Set this value to 0.0 to summarize all nodes in the target tree.
- **Target tree type** - This has two options "Maximum clade credibility" or "User target tree". For the latter option, a NEXUS tree file can be specified as the Target Tree File, below. For the former option, TreeAnnotator will examine every tree in the Input Tree File and select the tree that has the highest sum of the posterior probabilities of all its nodes.
- **Node heights** - This option specifies what node heights (times) should be used for the output tree. If the "Keep target heights" is selected, then the node heights will be the same as the target tree. The other two options give node heights as an average (Mean or Median) over the sample of trees. Keep the default median node heights for the time being.
- **Target Tree File** - If the "User target tree" option is selected then you can use "Choose File..." to select a NEXUS file containing the target tree.
- **Input Tree File** - Use the "Choose File..." button to select an input trees file. This will be the trees file produced by BEAST.
- **Output File** - Select a name for the output tree file (e.g., YFV_MCC.tre).

Once you have selected all the options above, press the "Run" button. TreeAnnotator will analyze the input tree file and write the summary tree to the file you specified. This tree is in standard NEXUS tree file format so may be loaded into any tree drawing package that supports this. However, it also contains additional information that can only be displayed using the FigTree program.

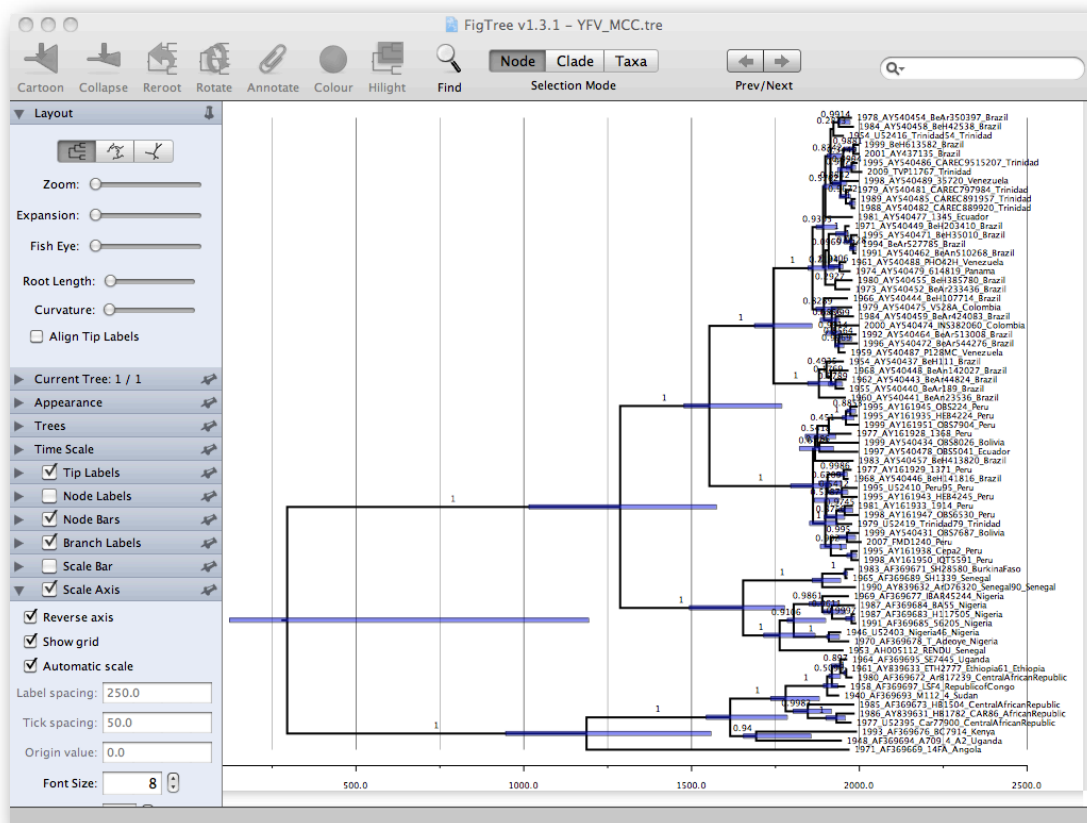
Viewing the annotated tree

Run FigTree now and select the **Open...** command from the **File** menu. Select the tree file you created using TreeAnnotator in the previous section. The tree will be displayed in the FigTree window. On the left hand side of the window are the options and settings which control how the tree is displayed. In this case we want to display the posterior probabilities of each of the clades present in the tree and estimates of the age of each node. In order to do this you need to change some of the settings.

First open the **Branch Labels** section of the control panel on the left. Now select **posterior** from the **Display popup** menu. The posterior probabilities won't actually be displayed until you tick the check-box next to the **Branch Labels** title.

We now want to display bars on the tree to represent the estimated uncertainty in the date for each node. TreeAnnotator will have placed this information in the tree file in the shape of the 95% highest posterior density (HPD) intervals (see the description of HPDs, above). Open the **Node Bars** section of the control panel and you will notice that it is already set to display the 95% HPDs of the node heights so all you need to do is to select the check-box in order to turn the node bars on. We can also plot a time scale axis for this evolutionary history (select '**Scale Axis**' and deselect '**Scale bar**'). For appropriate scaling, open the '**Time Scale**' section of the control panel, set the '**Offset**' to 2009.0, the scale factor to -1.0. and '**Reverse Axis**' under '**Scale Axis**'.

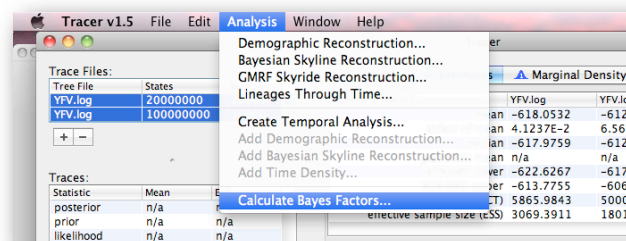
Finally, open the **Appearance** panel and alter the **Line Weight** to draw the tree with thicker lines. None of the options actually alter the tree's topology or branch lengths in anyway so feel free to explore the options and settings (Highlight or collapse for example the South American clade). You can also save the tree and this will save all your settings so that when you load it into FigTree again it will be displayed exactly as you selected.



How do the South American viruses cluster relative to the African viruses and what conclusions can we draw from the inferred time scale?

Evaluating rate variation

To investigate lineage-specific rate heterogeneity in this data set and its impact on divergence date estimates, a log and trees file is available for an analysis using an uncorrelated lognormal relaxed clock. Import this log file in Tracer in addition to previously imported strict clock log file. Investigate the posterior density for the lognormal standard deviation; if this density excludes zero (= no rate variation), it would suggest that the strict clock model can be rejected in favor of the relaxed clock model. A more formal test can be performed using a marginal likelihood estimator (MLE, Suchard et al., 2001, MBE 18: 1001-1013), which in turn employs a mixture of model prior and posterior samples (Newton and Raftery 1994). The ratio of marginal likelihoods defines a Bayes factor, which measures the relative fit for two different models given the data at hand. To obtain such estimates using a harmonic mean estimator (HME), select both trace files in Tracer and choose “calculate Bayes factors” from the “Analysis” menu.



Keep the default likelihood traces, set the bootstrap replicates to '0' (as this does not provide a good estimate of the uncertainty of the MLE estimate anyway) of the and press 'OK'. After a few seconds, log marginal likelihood estimates and \log_{10} Bayes factors will appear in a Bayes Factors window. Which model is favored according to the log marginal likelihood estimates? The \log_{10} Bayes factors are relatively convincing in this case, but it should be noted that the HME is not yield the best MLEs. More accurate MLE estimates can be obtained using computationally more demanding approaches, such as path sampling (sometimes also referred to as 'thermodynamic integration') and stepping-stone sampling, which have recently been implemented in beast (Baele et al., 2012, MBE, doi: 10.1093). To set up such analyses, we refer to the relevant BEAST tutorial: http://beast.bio.ed.ac.uk/Model_selection.

Conclusion and Resources

This chapter only scratches the surface of the analyses that are possible to undertake using BEAST. It has hopefully provided a relatively gentle introduction to the fundamental steps that will be common to all BEAST analyses and provide a basis for more challenging investigations. BEAST is an ongoing development project with new models and techniques being added on a regular basis. The BEAST website provides details of the mailing list that is used to announce new features and to discuss the use of the package. The website also contains a list of tutorials and recipes to answer particular evolutionary questions using BEAST as well as a description of the XML input format, common questions and error messages.

- The BEAST website: <http://beast.bio.ed.ac.uk/>
- Tutorials: <http://beast.bio.ed.ac.uk/Tutorials/>
- Frequently asked questions: <http://beast.bio.ed.ac.uk/FAQ/>