P. Lemey, M. Suchard and A. Rambaut

1. 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 Africa and the Americas 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 v1.8.2. It is available for download from http://beast.bio.ed.ac.uk/, or when unavailable at https://github.com/beast-dev/beast-mcmc.

- **Tracer** - this program is used to explore the output of **BEAST** (and other Bayesian MCMC programs). It graphically and quantitively summarizes the empirical distributions of continuous parameters and provides diagnostic information. At the time of writing, the current version is v1.6. It is available for download from http://beast.bio.ed.ac.uk/.

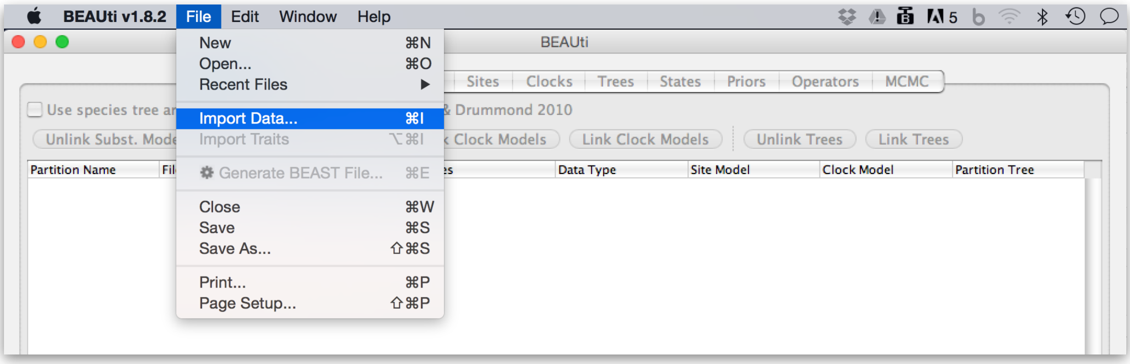
- **FigTree** - this is an application for displaying and printing molecular phylogenies, in particular those obtained using **BEAST**. At the time of writing, the current version is v1.4.2. It is available for download from http://tree.bio.ed.ac.uk/.

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.

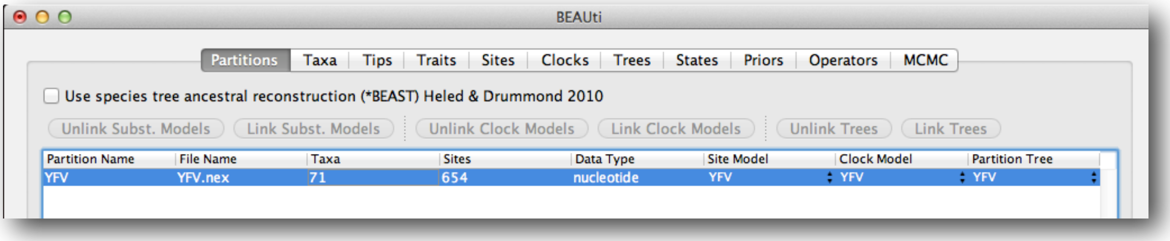
Loading the NEXUS file

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



The NEXUS alignment

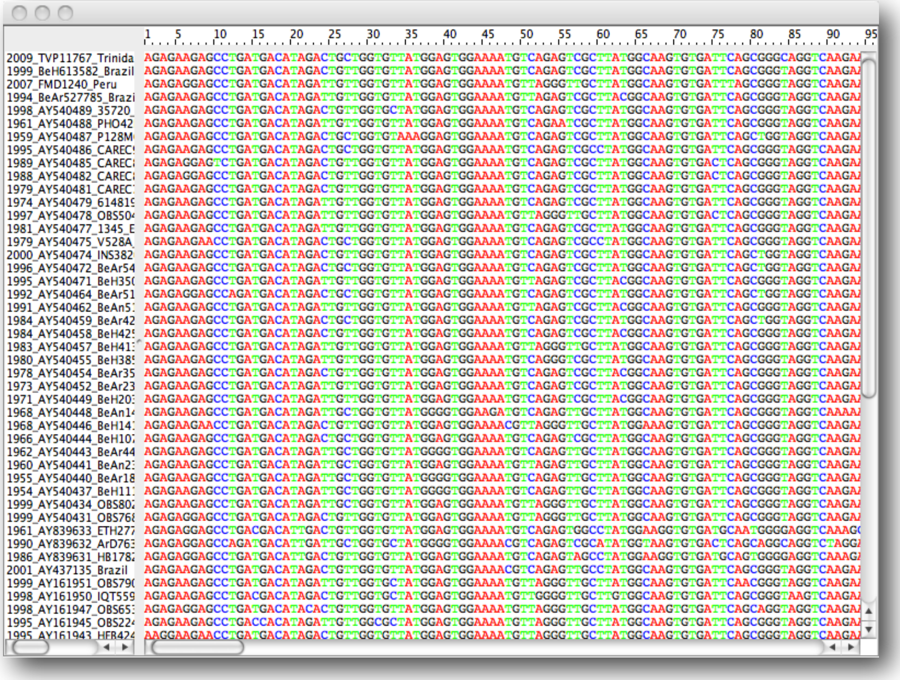
Select the file called `YFV.nex`. This file contains an alignment of 71 sequences from the *prM/E* gene of YFV, 654 nucleotides in length. Once loaded, the sequence data will be listed under **Partitions**:



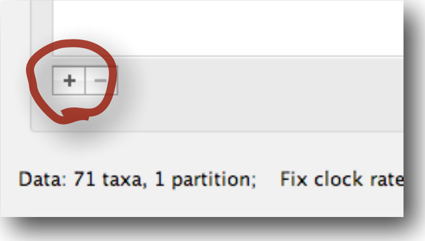
□ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

Partition Name	File Name	Taxa	Sites	Data Type	Site Model	Clock Model	Partition Tree
YFV	YFV.nex	71	654	nucleotide	YFV	YFV	YFV

Double clicking on the YFV.nex File Name will display the alignment in a separate window:

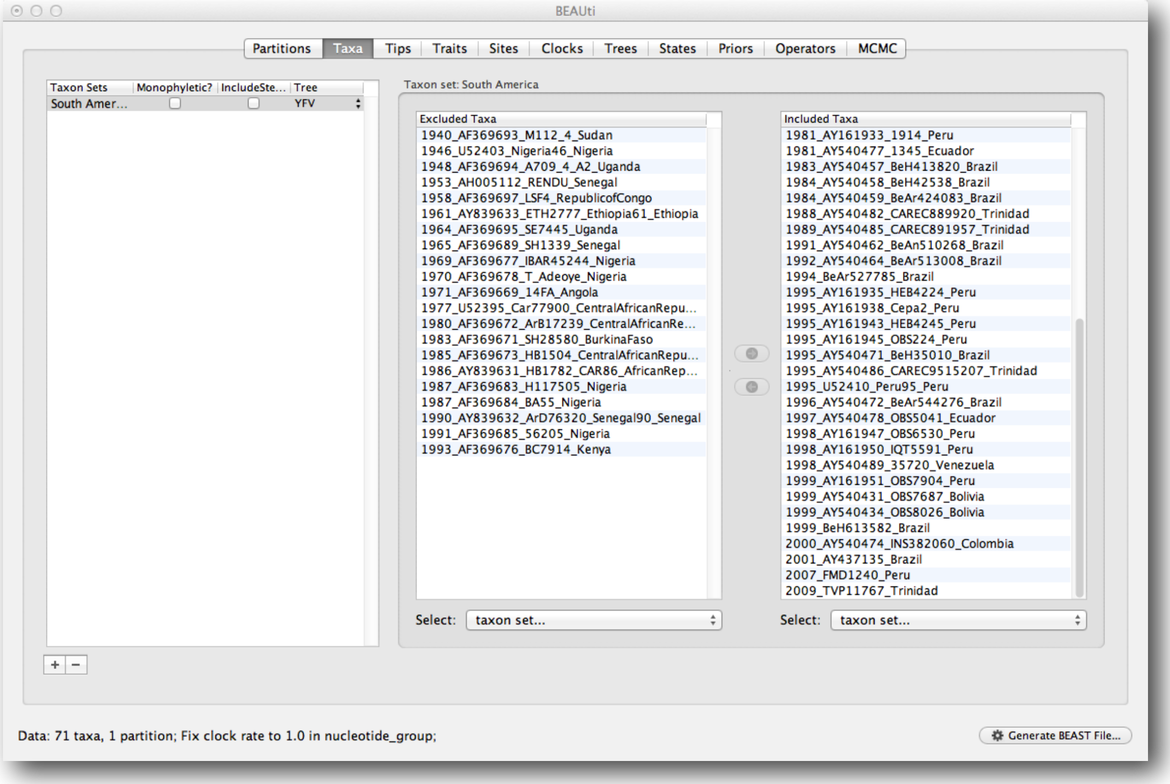


Under the Taxon Sets menu, we can define sets of taxa for which we would like to obtain particular statistics, enforce a monophyletic constraint, or put calibration information on. Let's define a "Americas" taxon set by pressing the small "plus" button at the bottom left of the panel:



Data: 71 taxa, 1 partition; Fix clock rate

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 `Americas`. 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 viruses from the Americas and not for the parent node leading to this clade. In the next table along you will see the available taxa. Select a taxon from the Americas such as `1954_AY540437_BeH111_Brazil` and press the green arrow button to move it into the included taxa set. Repeat this until all taxa from the Americas 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:



Taxon Sets	Monophyletic?	IncludeSte...	Tree
South Amer...	☐	☐	YFV

Taxon set: South America

Excluded Taxa
1940_AF369693_M112_4_Sudan
1946_U52403_Nigeria46_Nigeria
1948_AF369694_A709_4_A2_Uganda
1953_AH005112_RENDU_Senegal
1958_AF369697_LSF4_RepublicofCongo
1961_AY839633_ETH2777_Ethiopia61_Ethiopia
1964_AF369695_SE7445_Uganda
1965_AF369689_SH1339_Senegal
1969_AF369677_IBAR45244_Nigeria
1970_AF369678_T_Adeoye_Nigeria
1971_AF369669_14FA_Angola
1977_U52395_Car77900_CentralAfricanRepu...
1980_AF369672_ArB17239_CentralAfricanRe...
1983_AF369671_SH28580_BurkinaFaso
1985_AF369673_HB1504_CentralAfricanRepu...
1986_AY839631_HB1782_CAR86_AfricanRep...
1987_AF369683_H117505_Nigeria
1987_AF369684_BA55_Nigeria
1990_AY839632_ArD76320_Senegal90_Senegal
1991_AF369685_56205_Nigeria
1993_AF369676_BC7914_Kenya

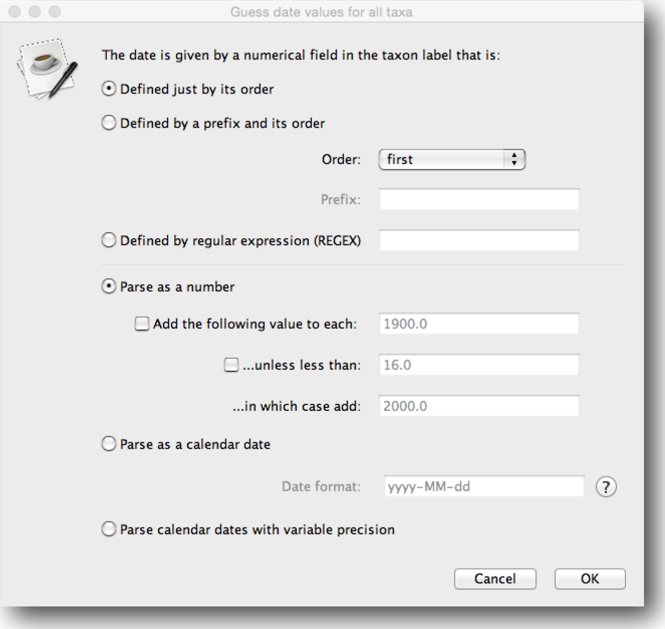
Select: taxon set...

Included Taxa
1981_AY161933_1914_Peru
1981_AY540477_1345_Ecuador
1983_AY540457_BeH413820_Brazil
1984_AY540458_BeH42538_Brazil
1984_AY540459_BeAr424083_Brazil
1988_AY540482_CAREC889920_Trinidad
1989_AY540485_CAREC891957_Trinidad
1991_AY540462_BeAn510268_Brazil
1992_AY540464_BeAr513008_Brazil
1994_BeAr527785_Brazil
1995_AY161935_HEB4224_Peru
1995_AY161938_Cepa2_Peru
1995_AY161943_HEB4245_Peru
1995_AY161945_OBS224_Peru
1995_AY540471_BeH35010_Brazil
1995_AY540486_CAREC9515207_Trinidad
1995_U52410_Peru95_Peru
1996_AY540472_BeAr544276_Brazil
1997_AY540478_OBS5041_Ecuador
1998_AY161947_OBS6530_Peru
1998_AY161950_IQT5591_Peru
1998_AY540489_35720_Venezuela
1999_AY161951_OBS7904_Peru
1999_AY540431_OBS7687_Bolivia
1999_AY540434_OBS8026_Bolivia
1999_BeH613582_Brazil
2000_AY540474_INS382060_Colombia
2001_AY437135_Brazil
2007_FMD1240_Peru
2009_TVP11767_Trinidad

Select: taxon set...

Data: 71 taxa, 1 partition; Fix clock rate to 1.0 in nucleotide_group;

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; BEAST considers the present or most recent sampling time as time 0). 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:



The date is given by a numerical field in the taxon label that is:

● Defined just by its order

○ Defined by a prefix and its order

Order: first

Prefix:

○ Defined by regular expression (REGEX)

● Parse as a number

□ Add the following value to each: 1900.0

□ ...unless less than: 16.0

...in which case add: 2000.0

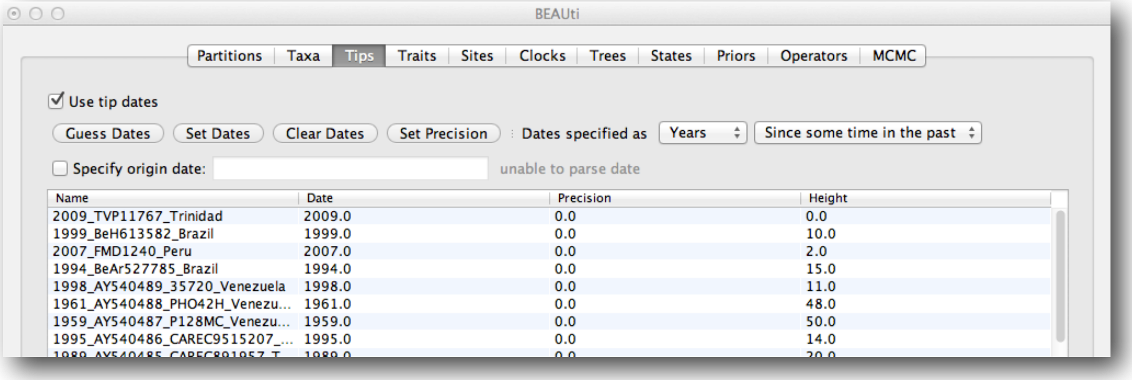
○ Parse as a calendar date

Date format: yyyy-MM-dd

○ Parse calendar dates with variable precision

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 and one for calendar dates with various precisions. Because all dates are specified in a four digit format in this case, no additional settings are needed. 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). The “Height” column lists the ages of the tips relative to time 0 (in our case 2009).



☑ Use tip dates

Dates specified as Years Since some time in the past

□ Specify origin date: unable to parse date

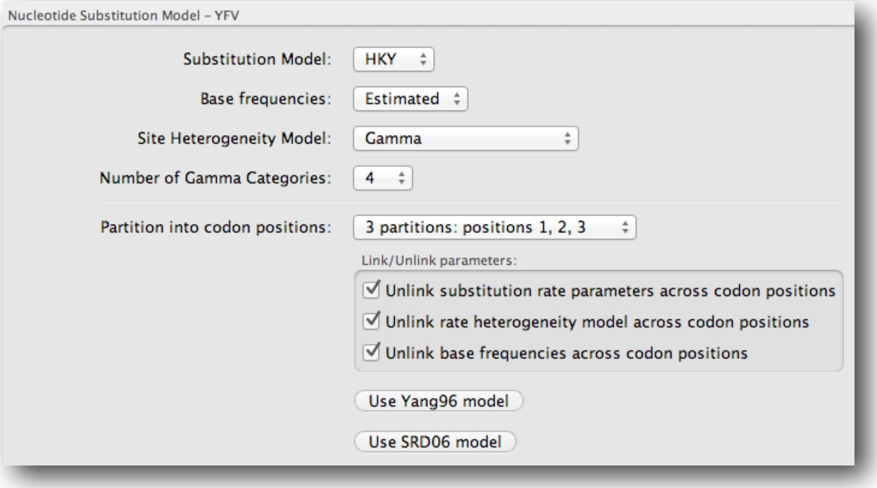
Name	Date	Precision	Height
2009_TVP11767_Trinidad	2009.0	0.0	0.0
1999_BeH613582_Brazil	1999.0	0.0	10.0
2007_FMD1240_Peru	2007.0	0.0	2.0
1994_BeAr527785_Brazil	1994.0	0.0	15.0
1998_AY540489_35720_Venezuela	1998.0	0.0	11.0
1961_AY540488_PHO42H_Venezu...	1961.0	0.0	48.0
1959_AY540487_P128MC_Venezu...	1959.0	0.0	50.0
1995_AY540486_CAREC9515207_...	1995.0	0.0	14.0
1989_AY540485_CAREC891957_T...	1989.0	0.0	20.0

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:



Nucleotide Substitution Model – YFV

Substitution Model: HKY

Base frequencies: Estimated

Site Heterogeneity Model: Gamma

Number of Gamma Categories: 4

Partition into codon positions: 3 partitions: positions 1, 2, 3

Link/Unlink parameters:

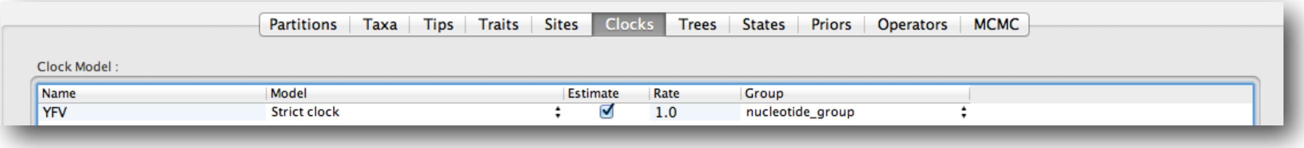
☑ Unlink substitution rate parameters across codon positions

☑ Unlink rate heterogeneity model across codon positions

☑ Unlink base frequencies across codon positions

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:



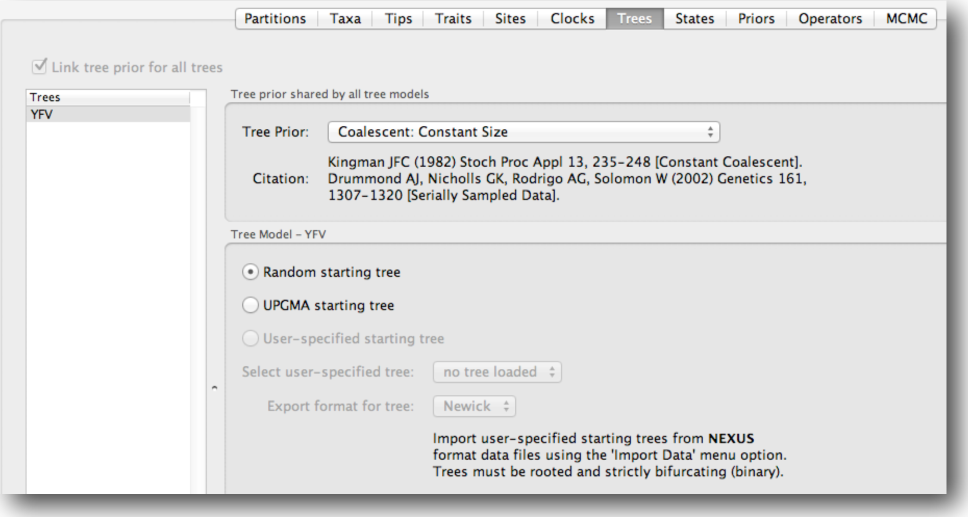
Clock Model :

Name	Model	Estimate	Rate	Group
YFV	Strict clock	☑	1.0	nucleotide_group

If there are no dates for the sequences (they are contemporaneous) then you can specify a fixed mean substitution rate obtained from another source (or better, specify a prior distribution later in the Priors tab to incorporate the uncertainty on this 'external' rate). Setting this to 1.0 will result in the ages of the nodes of the tree being estimated in units of substitutions per site (i.e. the normal units of branch lengths in popular packages such as *MrBayes*).

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 (simple) constant size coalescent prior. The tree priors (coalescent and other models) will be explained in other lectures.



☑ Link tree prior for all trees

Trees
YFV

Tree prior shared by all tree models

Tree Prior: Coalescent: Constant Size

Citation: Kingman JFC (1982) Stoch Proc Appl 13, 235–248 [Constant Coalescent]. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Genetics 161, 1307–1320 [Serially Sampled Data].

Tree Model - YFV

● Random starting tree

○ UPGMA starting tree

○ User-specified starting tree

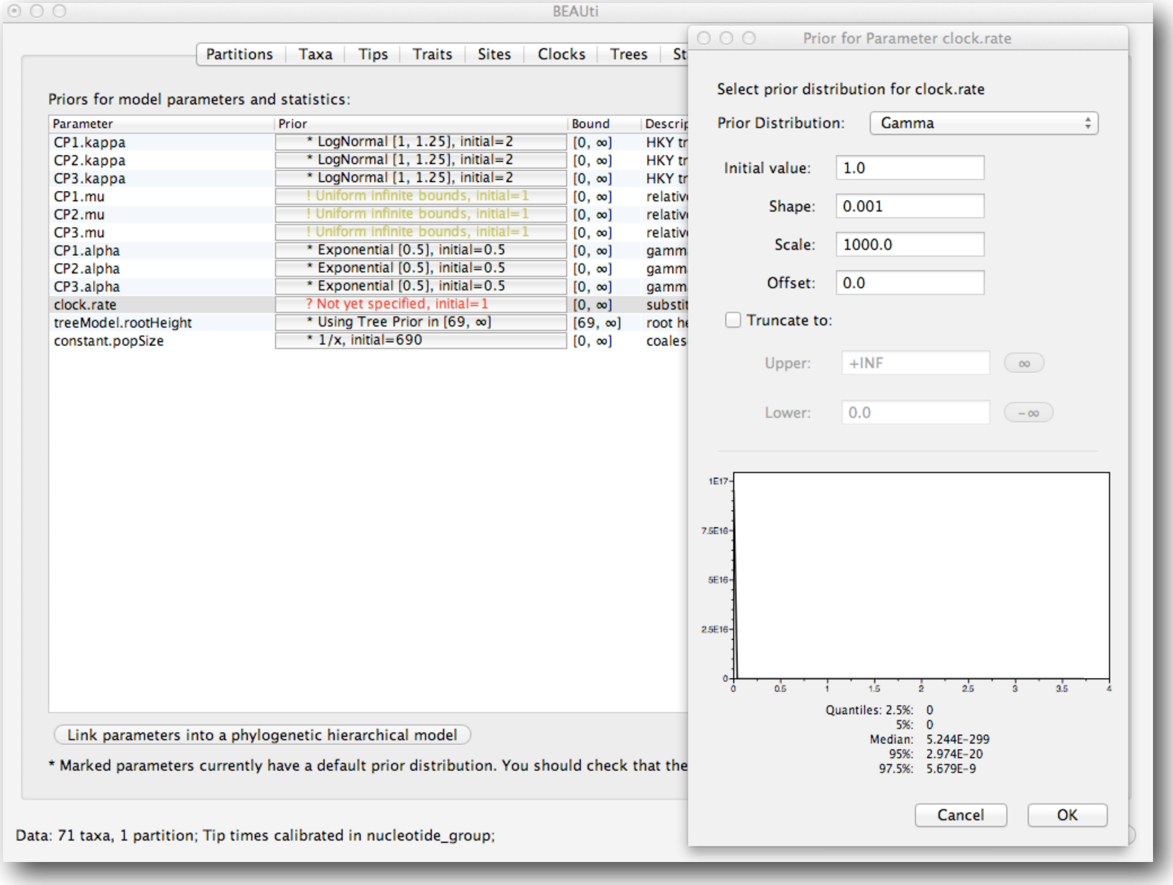
Select user-specified tree: no tree loaded

Export format for tree: Newick

Import user-specified starting trees from **NEXUS** format data files using the 'Import Data' menu option. Trees must be rooted and strictly bifurcating (binary).

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".



Priors for model parameters and statistics:

Parameter	Prior	Bound	Descrip
CP1.kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY tr
CP2.kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY tr
CP3.kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY tr
CP1.mu	! Uniform infinite bounds, initial=1	[0, ∞]	relativ
CP2.mu	! Uniform infinite bounds, initial=1	[0, ∞]	relativ
CP3.mu	! Uniform infinite bounds, initial=1	[0, ∞]	relativ
CP1.alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamm
CP2.alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamm
CP3.alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamm
clock.rate	? Not yet specified, initial=1	[0, ∞]	substit
treeModel.rootHeight	* Using Tree Prior in [69, ∞]	[69, ∞]	root he
constant.popSize	* 1/x, initial=690	[0, ∞]	coales

Link parameters into a phylogenetic hierarchical model

* Marked parameters currently have a default prior distribution. You should check that the

Data: 71 taxa, 1 partition; Tip times calibrated in nucleotide_group;

Select prior distribution for clock.rate

Prior Distribution: Gamma

Initial value: 1.0

Shape: 0.001

Scale: 1000.0

Offset: 0.0

□ Truncate to:

Upper: +INF

Lower: 0.0

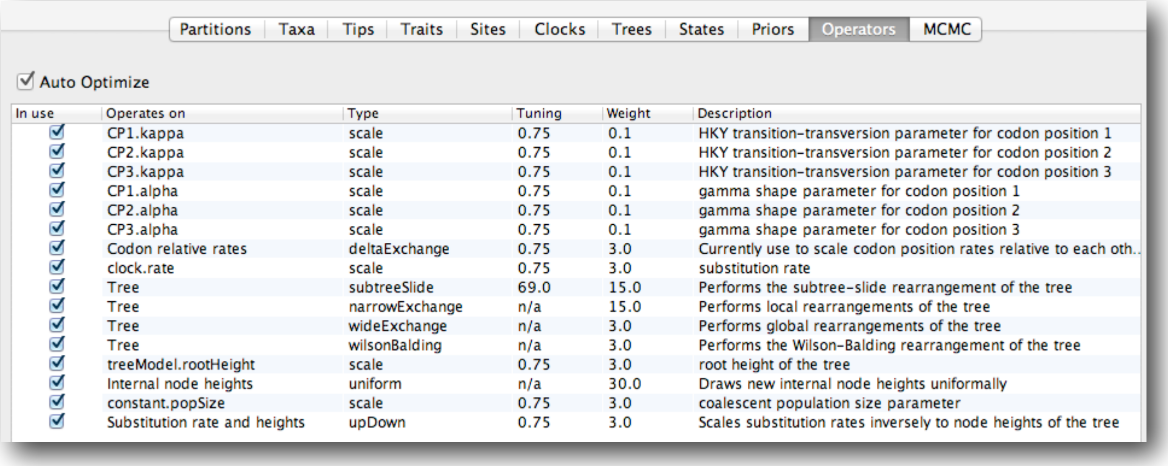
Quantiles: 2.5%: 0
5%: 0
Median: 5.244E-299
95%: 2.974E-20
97.5%: 5.679E-9

Setting up the operators

Each parameter in the model has one or more “operators” (these are variously called moves, proposals or transition kernels 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. We will start by using the default settings for this analysis.



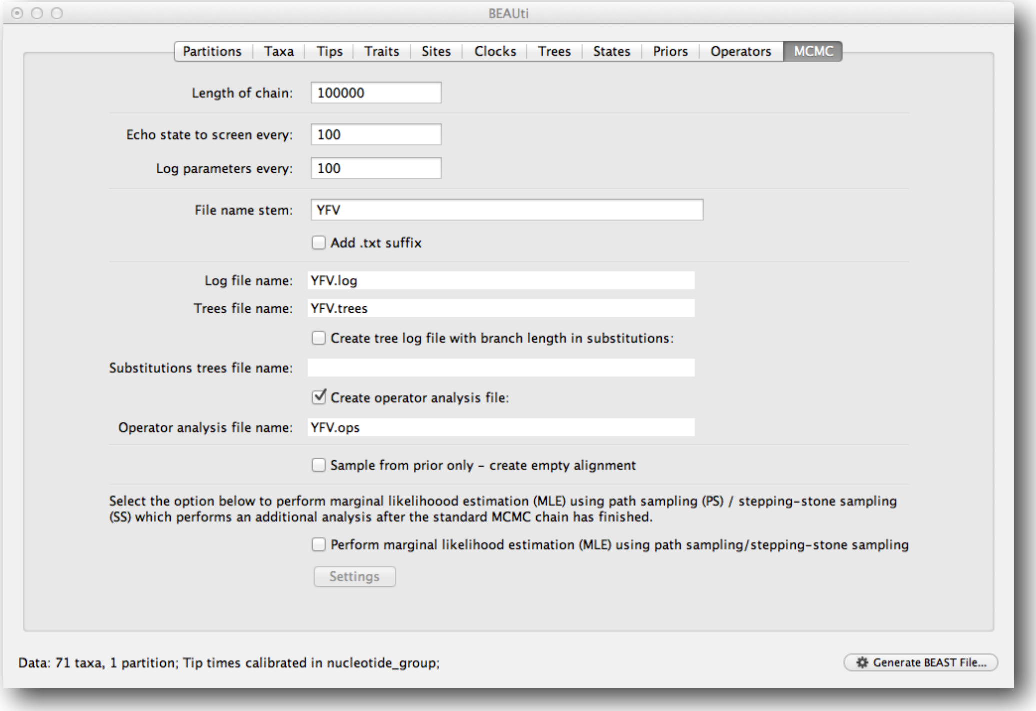
☑ Auto Optimize

In use	Operates on	Type	Tuning	Weight	Description
☑	CP1.kappa	scale	0.75	0.1	HKY transition-transversion parameter for codon position 1
☑	CP2.kappa	scale	0.75	0.1	HKY transition-transversion parameter for codon position 2
☑	CP3.kappa	scale	0.75	0.1	HKY transition-transversion parameter for codon position 3
☑	CP1.alpha	scale	0.75	0.1	gamma shape parameter for codon position 1
☑	CP2.alpha	scale	0.75	0.1	gamma shape parameter for codon position 2
☑	CP3.alpha	scale	0.75	0.1	gamma shape parameter for codon position 3
☑	Codon relative rates	deltaExchange	0.75	3.0	Currently use to scale codon position rates relative to each oth..
☑	clock.rate	scale	0.75	3.0	substitution rate
☑	Tree	subtreeSlide	69.0	15.0	Performs the subtree-slide rearrangement of the tree
☑	Tree	narrowExchange	n/a	15.0	Performs local rearrangements of the tree
☑	Tree	wideExchange	n/a	3.0	Performs global rearrangements of the tree
☑	Tree	wilsonBalding	n/a	3.0	Performs the Wilson-Balding rearrangement of the tree
☑	treeModel.rootHeight	scale	0.75	3.0	root height of the tree
☑	Internal node heights	uniform	n/a	30.0	Draws new internal node heights uniformally
☑	constant.popSize	scale	0.75	3.0	coalescent population size parameter
☑	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 >= chain length / 10,000.



Length of chain: 100000

Echo state to screen every: 100

Log parameters every: 100

File name stem: YFV

□ Add .txt suffix

Log file name: YFV.log

Trees file name: YFV.trees

□ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☑ Create operator analysis file:

Operator analysis file name: YFV.ops

□ Sample from prior only - create empty alignment

Select the option below to perform marginal likelihoood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) which performs an additional analysis after the standard MCMC chain has finished.

□ Perform marginal likelihood estimation (MLE) using path sampling/stepping-stone sampling

Data: 71 taxa, 1 partition; Tip times calibrated in nucleotide_group;

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.

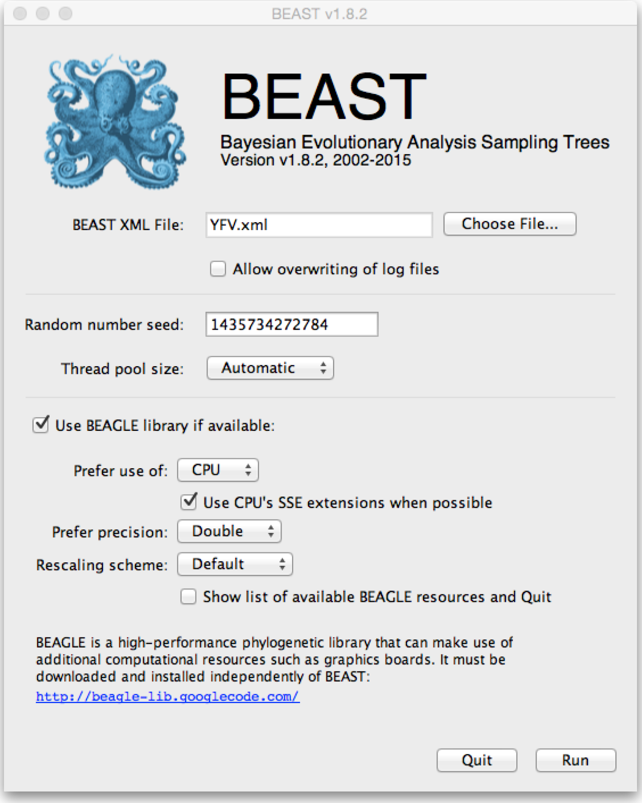
The next option allows the user to set the File stem name; if not set to 'YFV' by default, you can type this in here. The next two options give the file names of the log files for the parameters and the trees. These will be set to a default based on the file stem name. Let's also create an operator analysis file by selecting the relevant option. An option is also 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. Finally, one can select to perform marginal likelihood estimation to assess model fit; we will return to this later in this tutorial.

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 you have installed the BEAGLE library (https://github.com/beagle-dev/beagle-lib), you can use this in conjunction with BEAST to speed up the calculations. If not installed, unselect the use of the BEAGLE library. 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.


BEAST v1.8.2
BEAST
Bayesian Evolutionary Analysis Sampling Trees
Version v1.8.2, 2002-2015
Choose File...
Quit
Run

BEAST XML File: YFV.xml

□ Allow overwriting of log files

Random number seed: 1435734272784

Thread pool size: Automatic

☑ Use BEAGLE library if available:

Prefer use of: CPU

☑ Use CPU's SSE extensions when possible

Prefer precision: Double

Rescaling scheme: Default

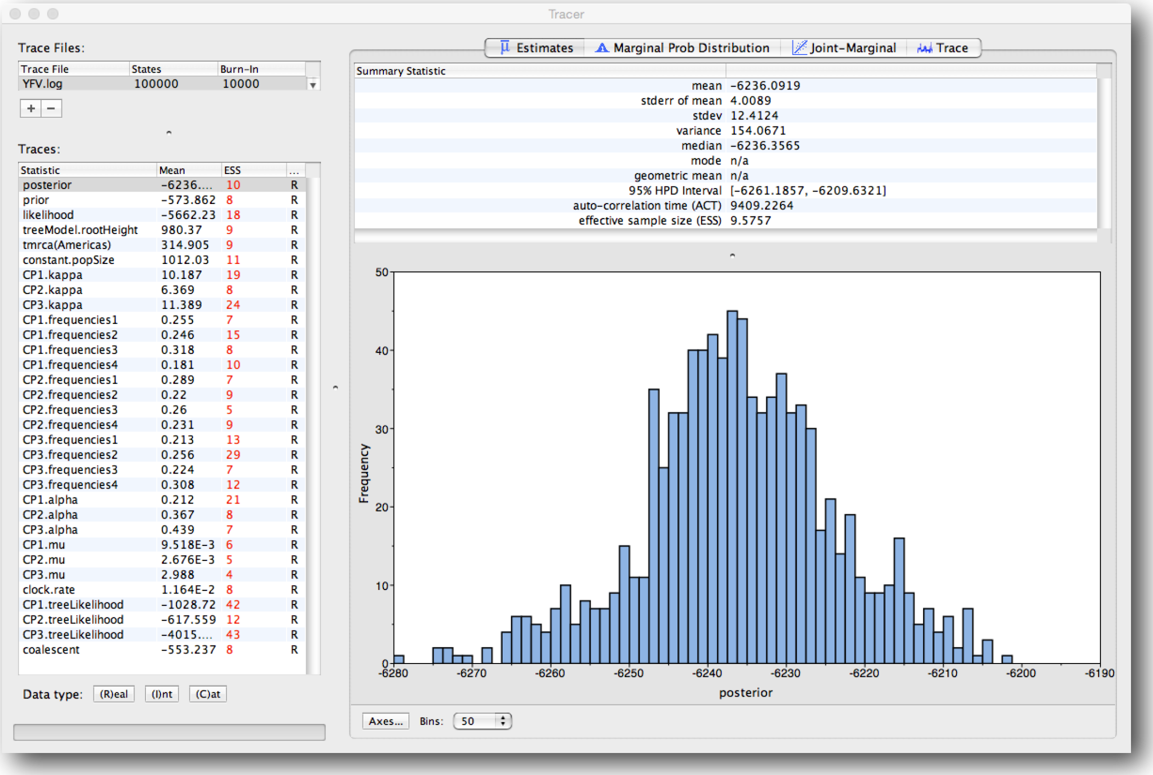
□ Show list of available BEAGLE resources and Quit

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/

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.



Trace Files:

Trace File	States	Burn-In
YFV.log	100000	10000

Traces:

Statistic	Mean	ESS	...
posterior	-6236....	10	R
prior	-573.862	8	R
likelihood	-5662.23	18	R
treeModel.rootHeight	980.37	9	R
tmrca(Americas)	314.905	9	R
constant.popSize	1012.03	11	R
CP1.kappa	10.187	19	R
CP2.kappa	6.369	8	R
CP3.kappa	11.389	24	R
CP1.frequencies1	0.255	7	R
CP1.frequencies2	0.246	15	R
CP1.frequencies3	0.318	8	R
CP1.frequencies4	0.181	10	R
CP2.frequencies1	0.289	7	R
CP2.frequencies2	0.22	9	R
CP2.frequencies3	0.26	5	R
CP2.frequencies4	0.231	9	R
CP3.frequencies1	0.213	13	R
CP3.frequencies2	0.256	29	R
CP3.frequencies3	0.224	7	R
CP3.frequencies4	0.308	12	R
CP1.alpha	0.212	21	R
CP2.alpha	0.367	8	R
CP3.alpha	0.439	7	R
CP1.mu	9.518E-3	6	R
CP2.mu	2.676E-3	5	R
CP3.mu	2.988	4	R
clock.rate	1.164E-2	8	R
CP1.treeLikelihood	-1028.72	42	R
CP2.treeLikelihood	-617.559	12	R
CP3.treeLikelihood	-4015....	43	R
coalescent	-553.237	8	R

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

Summary Statistic	
mean	-6236.0919
stderr of mean	4.0089
stdev	12.4124
variance	154.0671
median	-6236.3565
mode	n/a
geometric mean	n/a
95% HPD Interval	[-6261.1857, -6209.6321]
auto-correlation time (ACT)	9409.2264
effective sample size (ESS)	9.5757

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.

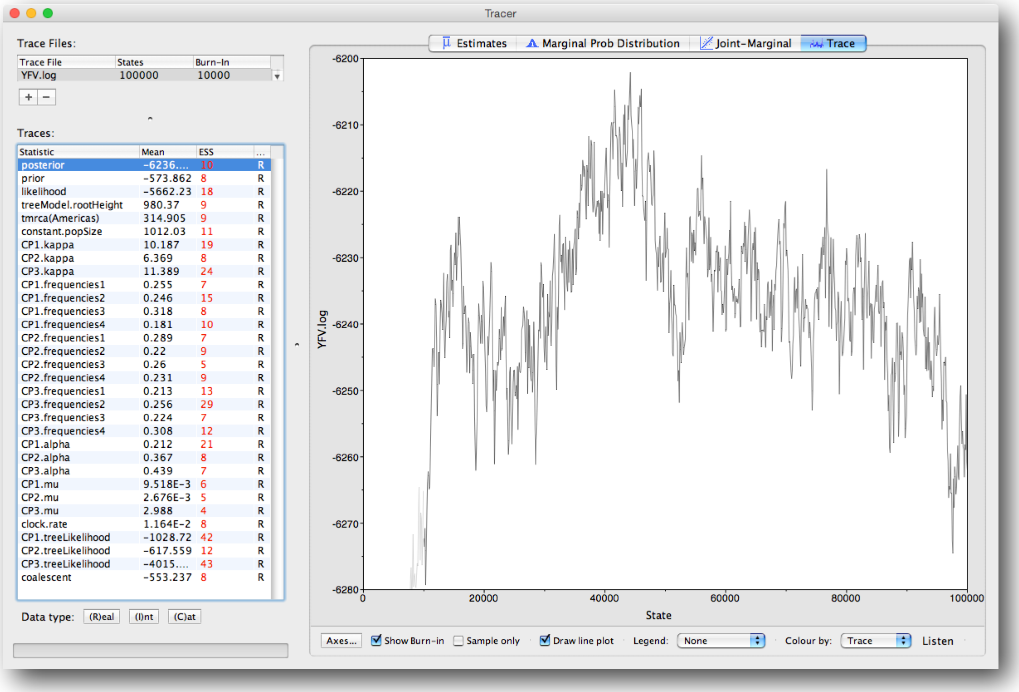
95% HPD Upper - The upper bound of the highest posterior density (HPD) interval.

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 - as 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.



Trace Files:

Trace File	States	Burn-In
YFV.log	100000	10000

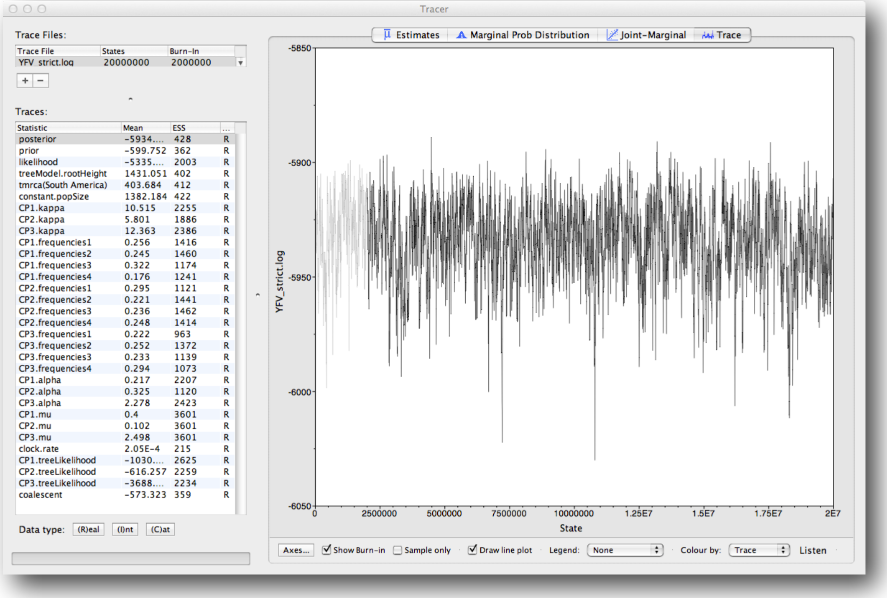
Traces:

Statistic	Mean	ESS	...
posterior	-6236....	10	R
prior	-573.862	8	R
likelihood	-5662.23	18	R
treeModel.rootHeight	980.37	9	R
tmrca(Americas)	314.905	9	R
constant.popSize	1012.03	11	R
CP1.kappa	10.187	19	R
CP2.kappa	6.369	8	R
CP3.kappa	11.389	24	R
CP1.frequencies1	0.255	7	R
CP1.frequencies2	0.246	15	R
CP1.frequencies3	0.318	8	R
CP1.frequencies4	0.181	10	R
CP2.frequencies1	0.289	7	R
CP2.frequencies2	0.22	9	R
CP2.frequencies3	0.26	5	R
CP2.frequencies4	0.231	9	R
CP3.frequencies1	0.213	13	R
CP3.frequencies2	0.256	29	R
CP3.frequencies3	0.224	7	R
CP3.frequencies4	0.308	12	R
CP1.alpha	0.212	21	R
CP2.alpha	0.367	8	R
CP3.alpha	0.439	7	R
CP1.mu	9.518E-3	6	R
CP2.mu	2.676E-3	5	R
CP3.mu	2.988	4	R
clock.rate	1.164E-2	8	R
CP1.treeLikelihood	-1028.72	42	R
CP2.treeLikelihood	-617.559	12	R
CP3.treeLikelihood	-4015....	43	R
coalescent	-553.237	8	R

☑ Show Burn-in □ Sample only ☑ Draw line plot Legend: None Colour by: Trace

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 9 so we are only getting 1 independent sample to every 111 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 the first part 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.



Trace Files:

Trace File	States	Burn-In
YFV_strict.log	20000000	2000000

Traces:

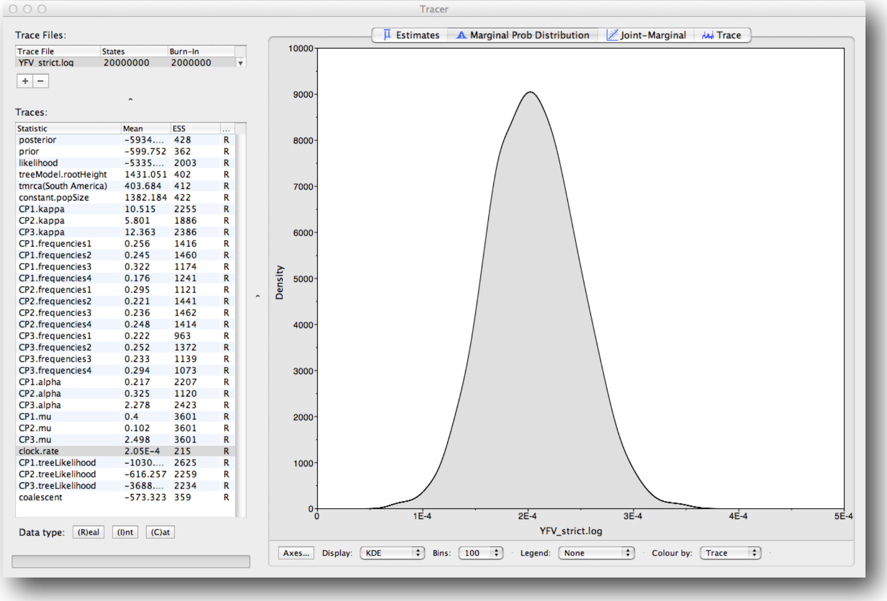
Statistic	Mean	ESS	...
posterior	-5934....	428	R
prior	-599.752	362	R
likelihood	-5335....	2003	R
treeModel.rootHeight	1431.051	402	R
tmrca(South America)	403.684	412	R
constant.popSize	1382.184	422	R
CP1.kappa	10.515	2255	R
CP2.kappa	5.801	1886	R
CP3.kappa	12.363	2386	R
CP1.frequencies1	0.256	1416	R
CP1.frequencies2	0.245	1460	R
CP1.frequencies3	0.322	1174	R
CP1.frequencies4	0.176	1241	R
CP2.frequencies1	0.295	1121	R
CP2.frequencies2	0.221	1441	R
CP2.frequencies3	0.236	1462	R
CP2.frequencies4	0.248	1414	R
CP3.frequencies1	0.222	963	R
CP3.frequencies2	0.252	1372	R
CP3.frequencies3	0.233	1139	R
CP3.frequencies4	0.294	1073	R
CP1.alpha	0.217	2207	R
CP2.alpha	0.325	1120	R
CP3.alpha	2.278	2423	R
CP1.mu	0.4	3601	R
CP2.mu	0.102	3601	R
CP3.mu	2.498	3601	R
clock.rate	2.05E-4	215	R
CP1.treeLikelihood	-1030....	2625	R
CP2.treeLikelihood	-616.257	2259	R
CP3.treeLikelihood	-3688....	2234	R
coalescent	-573.323	359	R

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

☑ Show Burn-in ☐ Sample only ☑ Draw line plot Legend: None Colour by: Trace

The log file provided contains 4000 samples and with an ESS of about 200 for the **clock.rate** there is still auto-correlation between the samples but 200 effectively independent samples will now provide an reasonable 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 Marginal Prob Distribution. This plot shows a kernel density estimate (KDE) of the posterior probability density of this parameter. You should see a plot similar to this:



Trace Files:

Trace File	States	Burn-In
YFV_strict.log	20000000	2000000

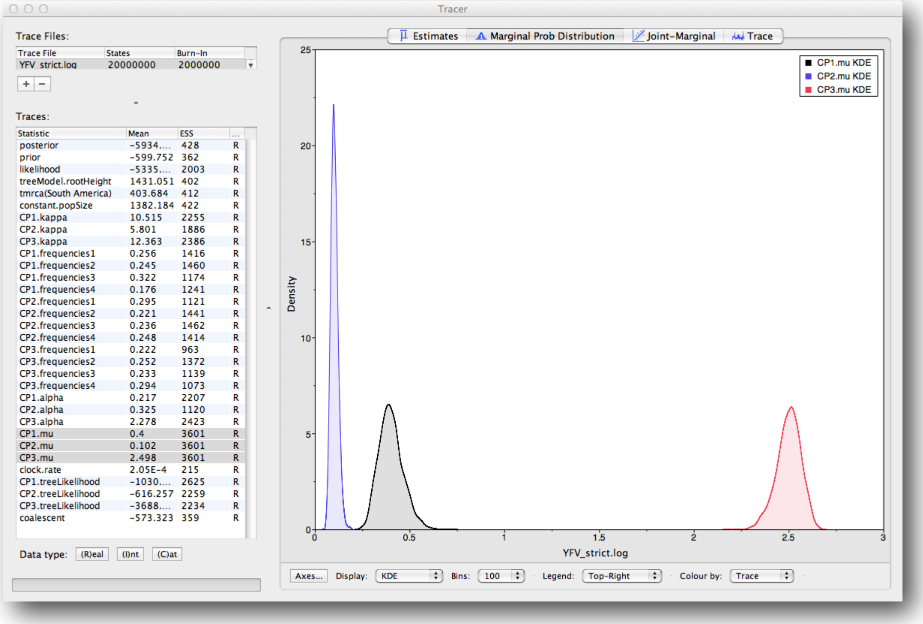
Traces:

Statistic	Mean	ESS	...
posterior	-5934....	428	R
prior	-599.752	362	R
likelihood	-5335....	2003	R
treeModel.rootHeight	1431.051	402	R
tmrca(South America)	403.684	412	R
constant.popSize	1382.184	422	R
CP1.kappa	10.515	2255	R
CP2.kappa	5.801	1886	R
CP3.kappa	12.363	2386	R
CP1.frequencies1	0.256	1416	R
CP1.frequencies2	0.245	1460	R
CP1.frequencies3	0.322	1174	R
CP1.frequencies4	0.176	1241	R
CP2.frequencies1	0.295	1121	R
CP2.frequencies2	0.221	1441	R
CP2.frequencies3	0.236	1462	R
CP2.frequencies4	0.248	1414	R
CP3.frequencies1	0.222	963	R
CP3.frequencies2	0.252	1372	R
CP3.frequencies3	0.233	1139	R
CP3.frequencies4	0.294	1073	R
CP1.alpha	0.217	2207	R
CP2.alpha	0.325	1120	R
CP3.alpha	2.278	2423	R
CP1.mu	0.4	3601	R
CP2.mu	0.102	3601	R
CP3.mu	2.498	3601	R
clock.rate	2.05E-4	215	R
CP1.treeLikelihood	-1030....	2625	R
CP2.treeLikelihood	-616.257	2259	R
CP3.treeLikelihood	-3688....	2234	R
coalescent	-573.323	359	R

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

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

As you can see the posterior probability density is nicely bell-shaped. When looking at the equivalent histogram in the Estimates panel, there is some sampling noise which is smoothened by the KDE; this 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 and select 'Top-Right' under Legend. You will now see the posterior probability densities for the relative substitution rate at all three codon positions overlaid:



Trace Files:

Trace File	States	Burn-In
YFV_strict.log	20000000	2000000

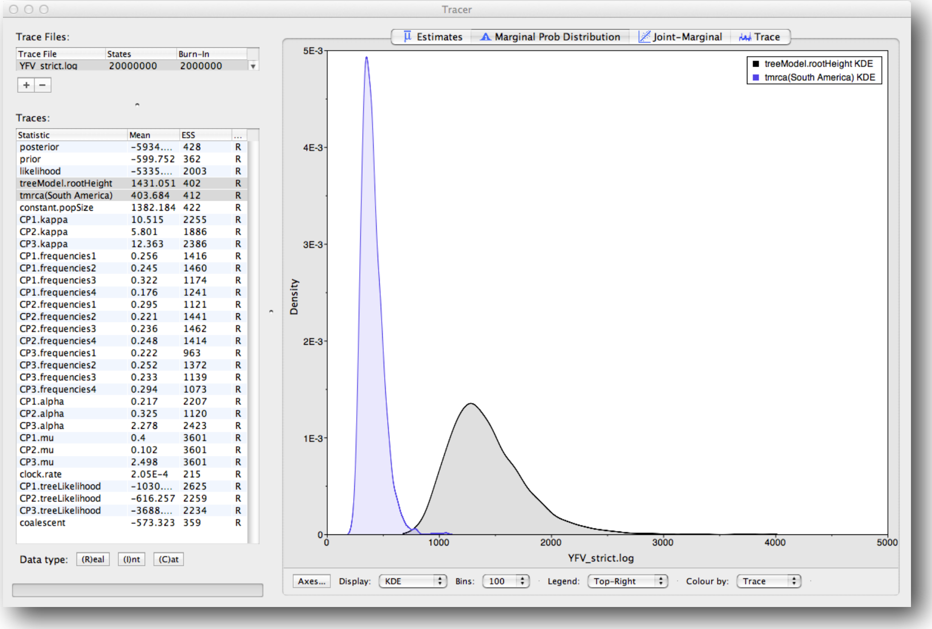
Traces:

Statistic	Mean	ESS	...
posterior	-5934....	428	R
prior	-599.752	362	R
likelihood	-5335....	2003	R
treeModel.rootHeight	1431.051	402	R
tmrca(South America)	403.684	412	R
constant.popSize	1382.184	422	R
CP1.kappa	10.515	2255	R
CP2.kappa	5.801	1886	R
CP3.kappa	12.363	2386	R
CP1.frequencies1	0.256	1416	R
CP1.frequencies2	0.245	1460	R
CP1.frequencies3	0.322	1174	R
CP1.frequencies4	0.176	1241	R
CP2.frequencies1	0.295	1121	R
CP2.frequencies2	0.221	1441	R
CP2.frequencies3	0.236	1462	R
CP2.frequencies4	0.248	1414	R
CP3.frequencies1	0.222	963	R
CP3.frequencies2	0.252	1372	R
CP3.frequencies3	0.233	1139	R
CP3.frequencies4	0.294	1073	R
CP1.alpha	0.217	2207	R
CP2.alpha	0.325	1120	R
CP3.alpha	2.278	2423	R
CP1.mu	0.4	3601	R
CP2.mu	0.102	3601	R
CP3.mu	2.498	3601	R
clock.rate	2.05E-4	215	R
CP1.treeLikelihood	-1030....	2625	R
CP2.treeLikelihood	-616.257	2259	R
CP3.treeLikelihood	-3688....	2234	R
coalescent	-573.323	359	R

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

Display: KDE Bins: 100 Legend: Top-Right Colour by: Trace

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 strains from the Americas relative to the general tmrca:



Trace Files:

Trace File	States	Burn-In
YFV_strict.log	20000000	2000000

Traces:

Statistic	Mean	ESS	...
posterior	-5934....	428	R
prior	-599.752	362	R
likelihood	-5335....	2003	R
treeModel.rootHeight	1431.051	402	R
tmrca(South America)	403.684	412	R
constant.popSize	1382.184	422	R
CP1.kappa	10.515	2255	R
CP2.kappa	5.801	1886	R
CP3.kappa	12.363	2386	R
CP1.frequencies1	0.256	1416	R
CP1.frequencies2	0.245	1460	R
CP1.frequencies3	0.322	1174	R
CP1.frequencies4	0.176	1241	R
CP2.frequencies1	0.295	1121	R
CP2.frequencies2	0.221	1441	R
CP2.frequencies3	0.236	1462	R
CP2.frequencies4	0.248	1414	R
CP3.frequencies1	0.222	963	R
CP3.frequencies2	0.252	1372	R
CP3.frequencies3	0.233	1139	R
CP3.frequencies4	0.294	1073	R
CP1.alpha	0.217	2207	R
CP2.alpha	0.325	1120	R
CP3.alpha	2.278	2423	R
CP1.mu	0.4	3601	R
CP2.mu	0.102	3601	R
CP3.mu	2.498	3601	R
clock.rate	2.05E-4	215	R
CP1.treeLikelihood	-1030....	2625	R
CP2.treeLikelihood	-616.257	2259	R
CP3.treeLikelihood	-3688....	2234	R
coalescent	-573.323	359	R

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

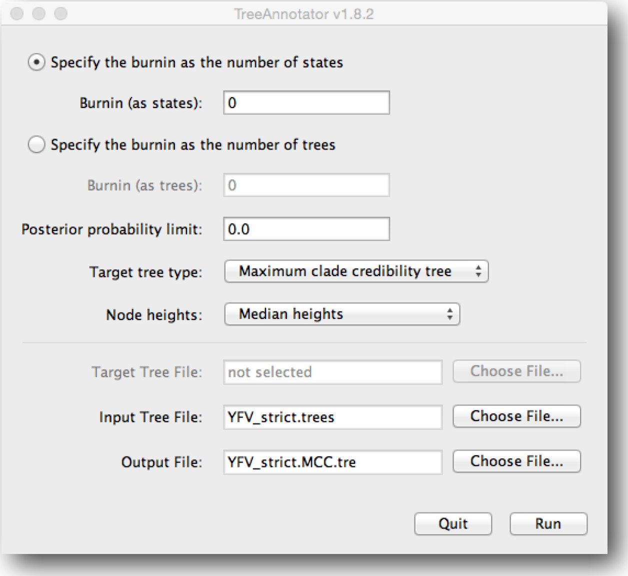
Display: KDE Bins: 100 Legend: Top-Right Colour by: Trace

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

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.



● Specify the burnin as the number of states

Burnin (as states): 0

○ Specify the burnin as the number of trees

Burnin (as trees): 0

Posterior probability limit: 0.0

Target tree type: Maximum clade credibility tree

Node heights: Median heights

Target Tree File: not selected

Input Tree File: YFV_strict.trees

Output File: YFV_strict.MCC.tre

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 steps in the MCMC chain, Burnin (as states), or the number of trees, Burnin (as trees), that should be excluded from the summarization. For the example above, with a chain of 20,000,000 steps, a 10% burnin corresponds to 2,000,000 steps. Alternatively, sampling every 5,000 steps results in 4,000 trees in the file, and to obtain at the same 10% burnin, the number of trees needs to be set to 400.

- Posterior probability limit - This is the minimum posterior probability for a node in order for TreeAnnotator to store the annotated information. Keep this value to the default of 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 product 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_strict.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.

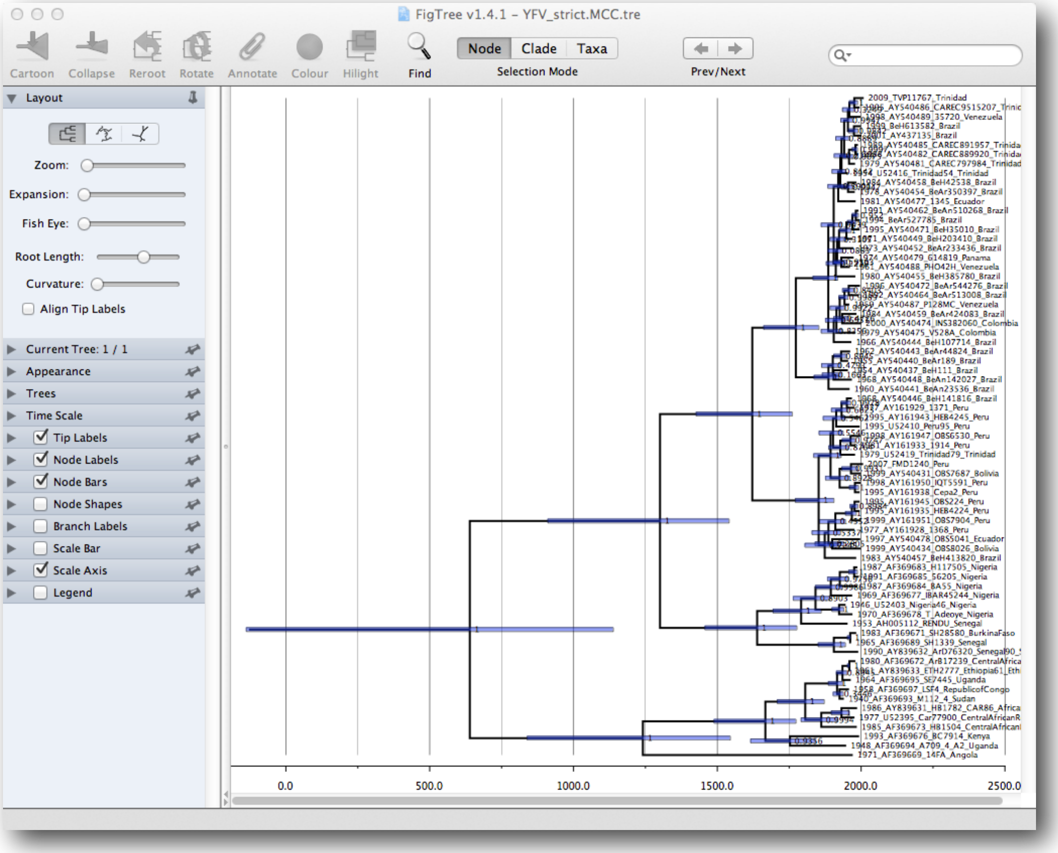
Viewing the annotated tree

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, re-order the node order by Increasing Node Order under the Tree Menu. Click on **Branch Labels** in the control panel on the left and open its section by clicking on the arrow on the left. Now select **posterior** under the **Display option**.

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). Select **Node Bars** in the control panel and open this section; select 'height_95%_HPD' to display the 95% HPDs of the node heights. It might be that this makes the tree move to the right almost entirely out of the display window. In that case, increase the Root Length under Layout in the control panel. 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 Americas clade). You can also save the tree and this will save most of your settings so that when you load it into FigTree again it will be displayed almost exactly as you selected. The tree can also be exported to a graphics file (pdf, eps, etc.).



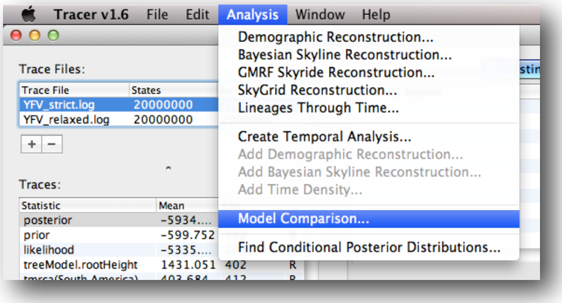
FigTree v1.4.1 – YFV_strict.MCC.tre

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

How do the viruses from the Americas 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. Note that despite the 5 times longer run for the relaxed clock analysis, the standard deviation for the lognormal distribution used in the relaxed clock model is still below 100 and therefore not optimal. This appears to be due to poor mixing between two different modes, resulting in a high autocorrelation. Investigate further 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. A frequently used method to obtain marginal likelihood estimates is the harmonic mean estimator (HME), which can be calculated using Tracer. Select both trace files in Tracer and choose “Model Comparison...” from the “Analysis” menu.


Tracer v1.6
File
Edit
Analysis
Window
Help
Demographic Reconstruction...
Bayesian Skyline Reconstruction...
GMRF Skyride Reconstruction...
SkyGrid Reconstruction...
Lineages Through Time...
Create Temporal Analysis...
Add Demographic Reconstruction...
Add Bayesian Skyline Reconstruction...
Add Time Density...
Model Comparison...
Find Conditional Posterior Distributions...
+
−

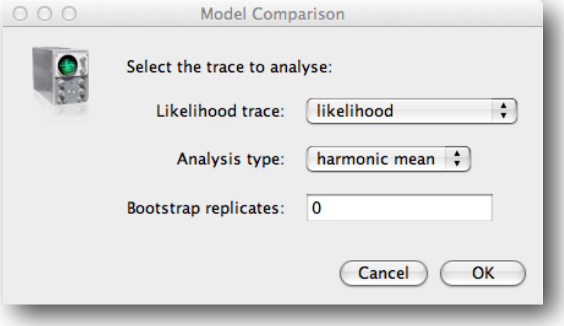
Trace Files:

Trace File	States
YFV_strict.log	20000000
YFV_relaxed.log	20000000

Traces:

Statistic	Mean		
posterior	-5934....		
prior	-599.752		
likelihood	-5335....		
treeModel.rootHeight	1431.051	402	R

Keep the default likelihood traces, select ‘harmonic mean’ as Analysis type and set the bootstrap replicates to ‘0’ (as this does not provide a good estimate of the uncertainty of the MLE estimate anyway) and press ‘OK’.



Select the trace to analyse:

Likelihood trace: likelihood

Analysis type: harmonic mean

Bootstrap replicates: 0

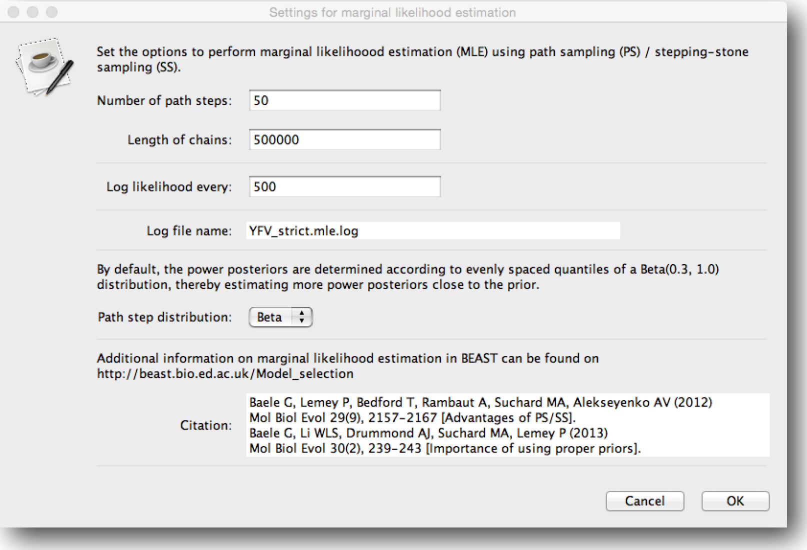
After a few seconds, log marginal likelihood estimates and $\log_{10}$ Bayes factors will appear in a Bayes Factors window. Which model is favored according to the log marginal likelihood estimates? The $\log_{10}$ Bayes factors are relatively convincing in this case, but it should be noted that the HME does not yield the accurate MLEs and is therefore deemed to be obsolete.

More accurate MLE estimates can be obtained using computationally more demanding approaches, such as:

- path sampling (PS): proposed in the statistics literature over 2 decades ago (sometimes also referred to as ‘thermodynamic integration’), this approach has only recently been introduced in phylogenetics (Lartillot and Philippe 2006). Rather than only using samples from the posterior distribution, samples are required from a series of power posteriors between posterior and prior. As the power posterior gets closer to the prior, there is less and less data available for the posterior, which may lead to improper results when improper priors have been specified. The use of proper priors is hence of primary importance for PS.

- stepping-stone sampling (SS): following essentially the same approach as PS and using the same collection of samples from a series of power posteriors, stepping-stone sampling (SS) yields faster-converging results compared to PS. As was the case for PS, the use of proper priors is critical for SS.

These approaches have recently been implemented in beast (Baele at al., 2012, MBE, doi:10.1093/molbev/mss084). Typically, PS/SS model selection is performed after doing a standard MCMC analysis. PS and SS sampling can then start where the MCMC analysis has stopped (i.e. you should have run the MCMC analysis long enough so that it converged towards the posterior), thereby eliminating the need for PS and SS to first converge towards the posterior. To set up such analyses, we can return to BEAUti and select "Perform marginal likelihood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) which performs and additional analysis after the standard MCMC chain has finished." in the MCMC panel. Click on "settings" to specify the PS/SS settings.



Set the options to perform marginal likelihoood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS).

Number of path steps: 50

Length of chains: 500000

Log likelihood every: 500

Log file name: YFV_strict.mle.log

By default, the power posteriors are determined according to evenly spaced quantiles of a Beta(0.3, 1.0) distribution, thereby estimating more power posteriors close to the prior.

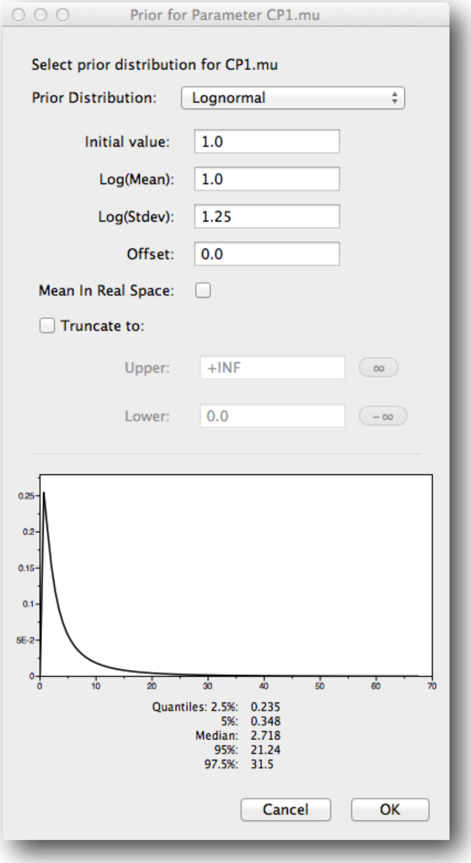
Path step distribution: Beta

Additional information on marginal likelihood estimation in BEAST can be found on http://beast.bio.ed.ac.uk/Model_selection

Citation: Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV (2012) Mol Biol Evol 29(9), 2157-2167 [Advantages of PS/SS].
Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P (2013) Mol Biol Evol 30(2), 239-243 [Importance of using proper priors].

We need to set the number of steps (in this case 50) for the path between the posterior and the prior, the length of the MCMC chain (in this case 500,000) for each step that estimates a specific power posterior, and the logging frequency for each MCMC sampling. Note that using these settings, the marginal likelihood estimation will take approximately the time it takes to complete a standard MCMC run of 25,000,000 generations for this data. The powers for the different power posteriors are defined using evenly spaced quantiles of a Beta(α,1.0) distribution, with α here equal to 0.30, as suggested in the stepping-stone sampling paper (Xie et al. 2011) since this approach is shown to outperform a uniform spreading suggested in the path sampling paper (Lartillot and Philippe 2006). Note that currently additional research is being performed on the number of power posteriors needed and how long they need to run in order to estimate accurate (log) marginal likelihoods, which is why it is sometimes suggested to perform the calculations twice to get an idea of the variation on these estimators with the assumed settings (Baele et al. 2012).

As noted above, the relative substitution rates did not have proper priors in our standard MCMC analysis, which was not a major issue because with sufficiently informative data theirs posterior will become proper. For the PS/SS procedure however, we need to sample from the prior at the end of the series of the power posteriors, which will be problematic without proper priors (Baele at al., 2013, MBE, doi:10.1093/molbev/mss243). In fact, BEAST will return to the prior settings when trying to write to xml the PS/SS analysis without proper priors. Therefore, we will set proper priors for the relative substitution rates in the Prior panel before proceeding with this run. For CP1.mu, CP2.mu and CP3.mu, specify the same prior as for the kappa parameters, which is a lognormal distribution with log(mean) 1.0 and log(stdev) 1.25:



Prior for Parameter CP1.mu

Select prior distribution for CP1.mu

Prior Distribution: Lognormal

Initial value: 1.0

Log(Mean): 1.0

Log(Stdev): 1.25

Offset: 0.0

Mean In Real Space: □

□ Truncate to:

Upper: +INF ∞

Lower: 0.0 −∞

```
Quantiles: 2.5%:  0.235
             5%:  0.348
         Median:  2.718
            95%:  21.24
          97.5%:  31.5
```

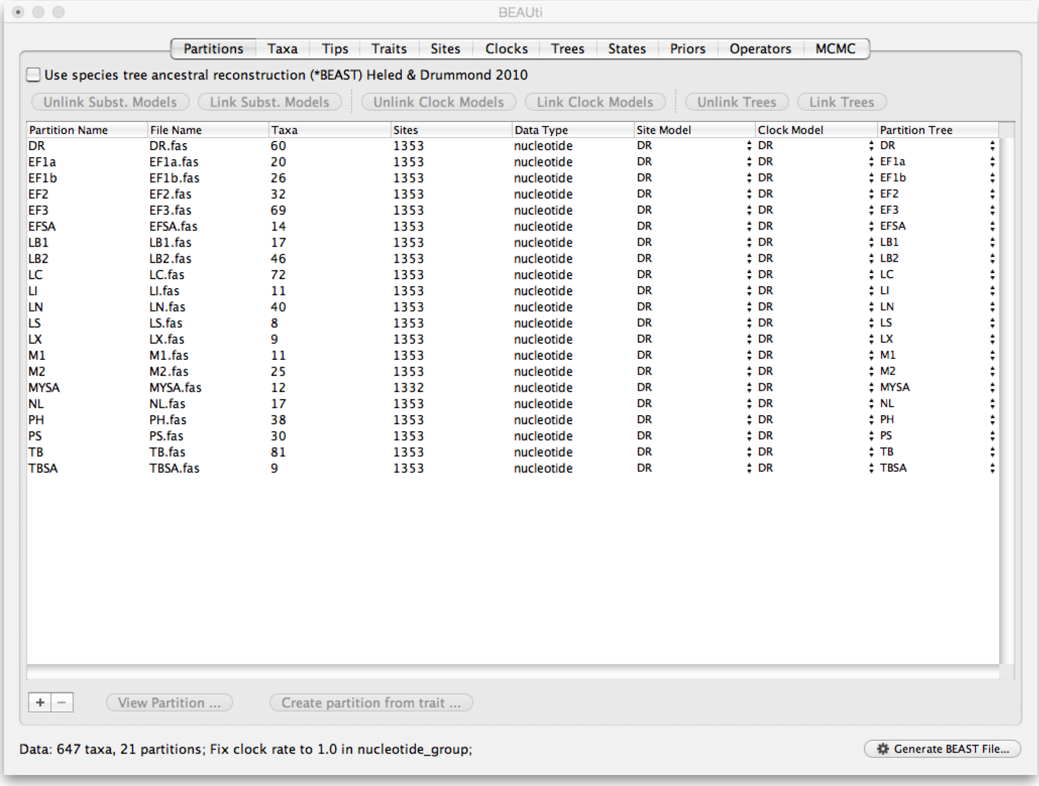
Having set the PS/SS settings and proper priors, we can write to xml and run the analysis in BEAST. However, the output and results of these analyses has been made available. For the strict clock analysis we arrive at -5678.97 and -5678.31 for PS and SS respectively. For the relaxed clock (ucld) analysis we get -5651.93 and 5652.36 for the same estimators. How do the PS/SS MLEs compare to those obtained by the HME, and the Bayes factors resulting from these different estimators?

2. Testing evolutionary rate hypotheses in bat rabies viruses

This part briefly describes how to set up an BEAST analysis aimed at identifying the factors responsible for evolutionary rate variation in bat rabies viruses (RABV) based on a data set previously analysed by Streicker et al. (2012). Bat rabies viruses in the Americas have established host-associated lineages through sequential host jumping followed by successful transmission in the new bat species (see Streicker et al., 2010). This provides a rare occasion to test the impact of host factors (physiological, environmental or ecological) on virus evolutionary rates. We will test these factors by setting up a mixed effects model for the evolutionary rate, which requires manual xml editing. We will first set up an xml using BEAUti that specifies hierarchical prior distributions for the clock rates and substitution model parameters (which models the random effects) and then manually introduce the fixed effects by editing the xml.

A hierarchical phylogenetic model

Separate alignments in fasta format are available for each host-associated lineage (DR.fas, EF1a.fas, EF1b.fas, EF2.fas, EF3.fas, EFSA.fas, LB1.fas, LB2.fas, LC.fas, LI.fas, LN.fas, LS.fas, LX.fas, M1.fas, M2.fas, MYSA.fas, NL.fas, PH.fas, PS.fas, TB.fas, TBSA.fas). Start BEAUti, select all the fasta files and drag them into the Partitions panel:



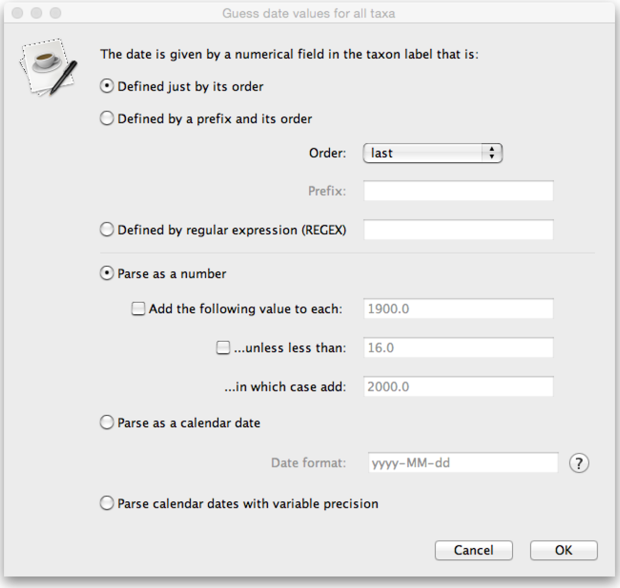
□ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

Partition Name	File Name	Taxa	Sites	Data Type	Site Model	Clock Model	Partition Tree
DR	DR.fas	60	1353	nucleotide	DR	DR	DR
EF1a	EF1a.fas	20	1353	nucleotide	DR	DR	EF1a
EF1b	EF1b.fas	26	1353	nucleotide	DR	DR	EF1b
EF2	EF2.fas	32	1353	nucleotide	DR	DR	EF2
EF3	EF3.fas	69	1353	nucleotide	DR	DR	EF3
EFSA	EFSA.fas	14	1353	nucleotide	DR	DR	EFSA
LB1	LB1.fas	17	1353	nucleotide	DR	DR	LB1
LB2	LB2.fas	46	1353	nucleotide	DR	DR	LB2
LC	LC.fas	72	1353	nucleotide	DR	DR	LC
LI	LI.fas	11	1353	nucleotide	DR	DR	LI
LN	LN.fas	40	1353	nucleotide	DR	DR	LN
LS	LS.fas	8	1353	nucleotide	DR	DR	LS
LX	LX.fas	9	1353	nucleotide	DR	DR	LX
M1	M1.fas	11	1353	nucleotide	DR	DR	M1
M2	M2.fas	25	1353	nucleotide	DR	DR	M2
MYSA	MYSA.fas	12	1332	nucleotide	DR	DR	MYSA
NL	NL.fas	17	1353	nucleotide	DR	DR	NL
PH	PH.fas	38	1353	nucleotide	DR	DR	PH
PS	PS.fas	30	1353	nucleotide	DR	DR	PS
TB	TB.fas	81	1353	nucleotide	DR	DR	TB
TBSA	TBSA.fas	9	1353	nucleotide	DR	DR	TBSA

Data: 647 taxa, 21 partitions; Fix clock rate to 1.0 in nucleotide_group;

Select all partitions, and click on ‘Unlink Subst. Models’ to allow each lineage to evolve according to different substitution parameters (check that the Site Model has changed and become specific to each partition). Also select ‘Unlink Clock Models’ to allow each lineage to evolve under different rates of evolution (check that the Clock Model has become specific to each partition).

In the Tips panel, select ‘Use tip dates’ and click on ‘Guess Dates’. In the new ‘Guess date values for all taxa’, set the Order to ‘last’ and click OK.



The date is given by a numerical field in the taxon label that is:

● Defined just by its order

○ Defined by a prefix and its order

Order: last

Prefix:

○ Defined by regular expression (REGEX)

● Parse as a number

□ Add the following value to each: 1900.0

□ ...unless less than: 16.0

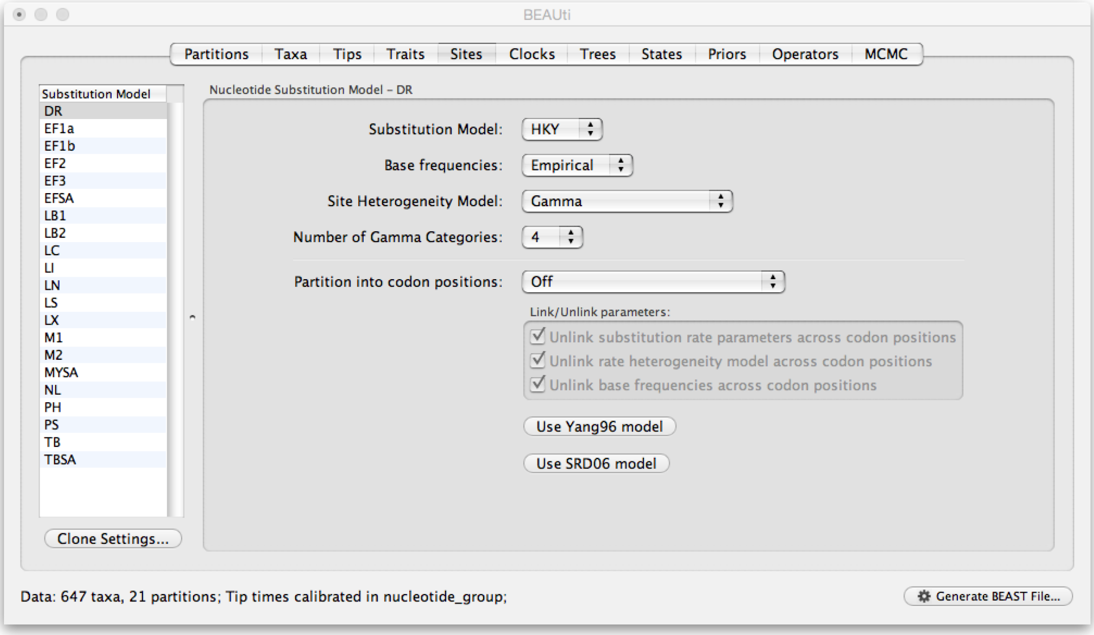
...in which case add: 2000.0

○ Parse as a calendar date

Date format: yyyy-MM-dd

○ Parse calendar dates with variable precision

In the Sites panel, keep the default HKY model, but set the ‘Base Frequencies’ to ‘Empirical’ and ‘Site heterogeneity Model’ to ‘Gamma’ for the DR partition:



Substitution Model
DR
EF1a
EF1b
EF2
EF3
EFSA
LB1
LB2
LC
LI
LN
LS
LX
M1
M2
MYSA
NL
PH
PS
TB
TBSA

Nucleotide Substitution Model – DR

Substitution Model: HKY

Base frequencies: Empirical

Site Heterogeneity Model: Gamma

Number of Gamma Categories: 4

Partition into codon positions: Off

Link/Unlink parameters:

☑ Unlink substitution rate parameters across codon positions

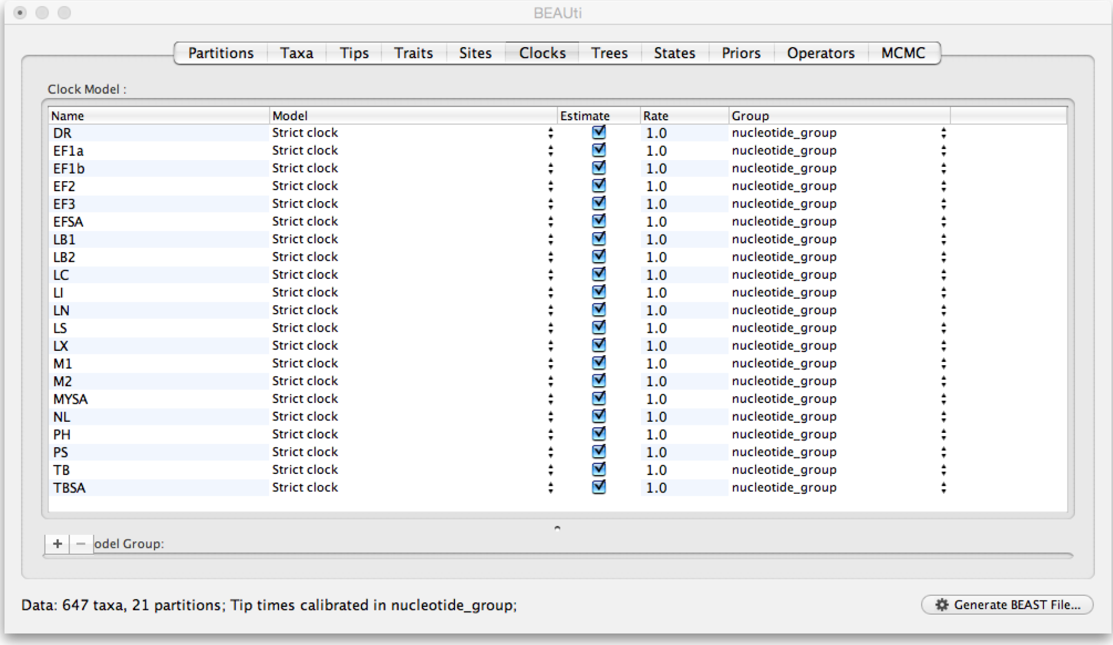
☑ Unlink rate heterogeneity model across codon positions

☑ Unlink base frequencies across codon positions

Data: 647 taxa, 21 partitions; Tip times calibrated in nucleotide_group;

To apply the same settings to all other partitions, select all partitions and click on the ‘Clone Settings’ button and keep the default ‘DR’ partition as source model for the cloning.

Keep the default ‘Strict Clock’ model in the Clocks panel:

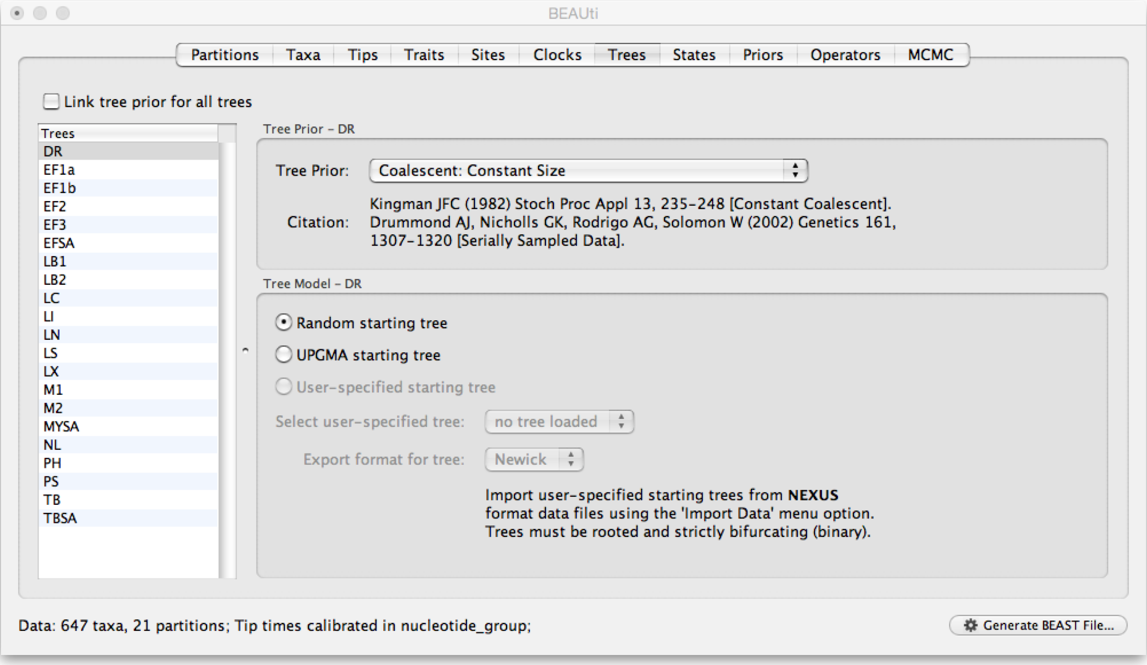


Clock Model :

Name	Model	Estimate	Rate	Group
DR	Strict clock	☑	1.0	nucleotide_group
EF1a	Strict clock	☑	1.0	nucleotide_group
EF1b	Strict clock	☑	1.0	nucleotide_group
EF2	Strict clock	☑	1.0	nucleotide_group
EF3	Strict clock	☑	1.0	nucleotide_group
EFSA	Strict clock	☑	1.0	nucleotide_group
LB1	Strict clock	☑	1.0	nucleotide_group
LB2	Strict clock	☑	1.0	nucleotide_group
LC	Strict clock	☑	1.0	nucleotide_group
LI	Strict clock	☑	1.0	nucleotide_group
LN	Strict clock	☑	1.0	nucleotide_group
LS	Strict clock	☑	1.0	nucleotide_group
LX	Strict clock	☑	1.0	nucleotide_group
M1	Strict clock	☑	1.0	nucleotide_group
M2	Strict clock	☑	1.0	nucleotide_group
MYSA	Strict clock	☑	1.0	nucleotide_group
NL	Strict clock	☑	1.0	nucleotide_group
PH	Strict clock	☑	1.0	nucleotide_group
PS	Strict clock	☑	1.0	nucleotide_group
TB	Strict clock	☑	1.0	nucleotide_group
TBSA	Strict clock	☑	1.0	nucleotide_group

Data: 647 taxa, 21 partitions; Tip times calibrated in nucleotide_group;

In the ‘Trees’ panel, make sure to unlink the tree prior (in the top left), and keep the default constant population size model as a simple tree prior:



☐ Link tree prior for all trees

Trees
DR
EF1a
EF1b
EF2
EF3
EFSA
LB1
LB2
LC
LI
LN
LS
LX
M1
M2
MYSA
NL
PH
PS
TB
TBSA

Tree Prior – DR

Tree Prior: Coalescent: Constant Size

Citation: Kingman JFC (1982) Stoch Proc Appl 13, 235–248 [Constant Coalescent].
Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Genetics 161, 1307–1320 [Serially Sampled Data].

Tree Model – DR

● Random starting tree

○ UPGMA starting tree

○ User-specified starting tree

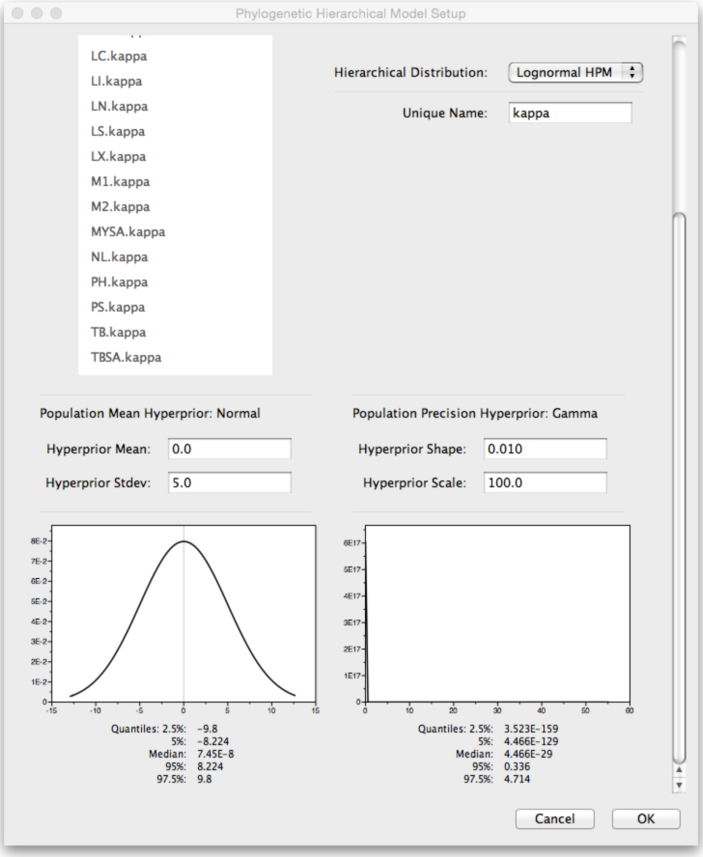
Select user-specified tree: no tree loaded

Export format for tree: Newick

Import user-specified starting trees from **NEXUS**
format data files using the 'Import Data' menu option.
Trees must be rooted and strictly bifurcating (binary).

Data: 647 taxa, 21 partitions; Tip times calibrated in nucleotide_group;

In the ‘Priors’ panel, we will specify the hierarchical priors for the substitution model parameters and the clock rates. We can achieve this by cmd-selecting all the equivalent parameters across the partitions and using the ‘Link parameters into a hierarchical model’ button at the bottom of the window. First cmd-select all kappa parameters, and select ‘Link parameters into a hierarchical model’. In the new Phylogenetic Hierarchical Model Setup window, enter a Unique Name (e.g. kappa), set the Normal Hyperprior Stdev to 5.0 and the Gamma Hyperprior Shape and Scale to 0.01 and 100.0 respectively and click OK.



LC.kappa
LI.kappa
LN.kappa
LS.kappa
LX.kappa
M1.kappa
M2.kappa
MYSA.kappa
NL.kappa
PH.kappa
PS.kappa
TB.kappa
TBSA.kappa

Hierarchical Distribution: Lognormal HPM

Unique Name: kappa

Population Mean Hyperprior: Normal

Hyperprior Mean: 0.0

Hyperprior Stdev: 5.0

Population Precision Hyperprior: Gamma

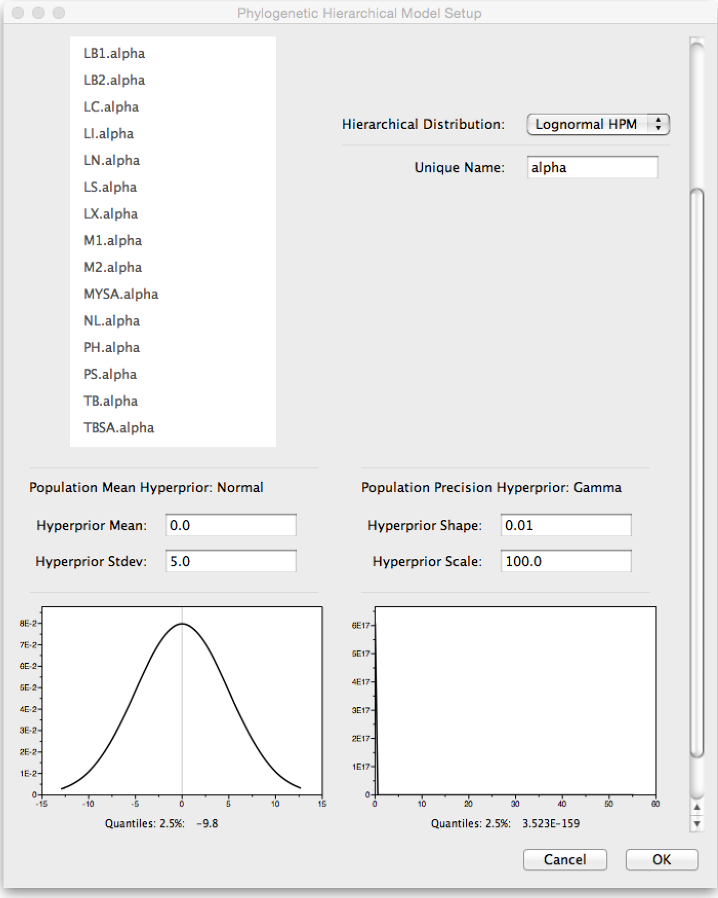
Hyperprior Shape: 0.010

Hyperprior Scale: 100.0

```
Quantiles: 2.5%:   -9.8
             5%:   -8.224
         Median:   7.45E-8
            95%:   8.224
          97.5%:   9.8
```

```
Quantiles: 2.5%:   3.523E-159
             5%:   4.466E-129
         Median:   4.466E-29
            95%:   0.336
          97.5%:   4.714
```

Repeat this operation for the alpha’s with the same hyperprior settings and alpha as Unique Name:



LB1.alpha
LB2.alpha
LC.alpha
LI.alpha
LN.alpha
LS.alpha
LX.alpha
M1.alpha
M2.alpha
MYSA.alpha
NL.alpha
PH.alpha
PS.alpha
TB.alpha
TBSA.alpha

Hierarchical Distribution: Lognormal HPM

Unique Name: alpha

Population Mean Hyperprior: Normal

Hyperprior Mean: 0.0

Hyperprior Stdev: 5.0

Population Precision Hyperprior: Gamma

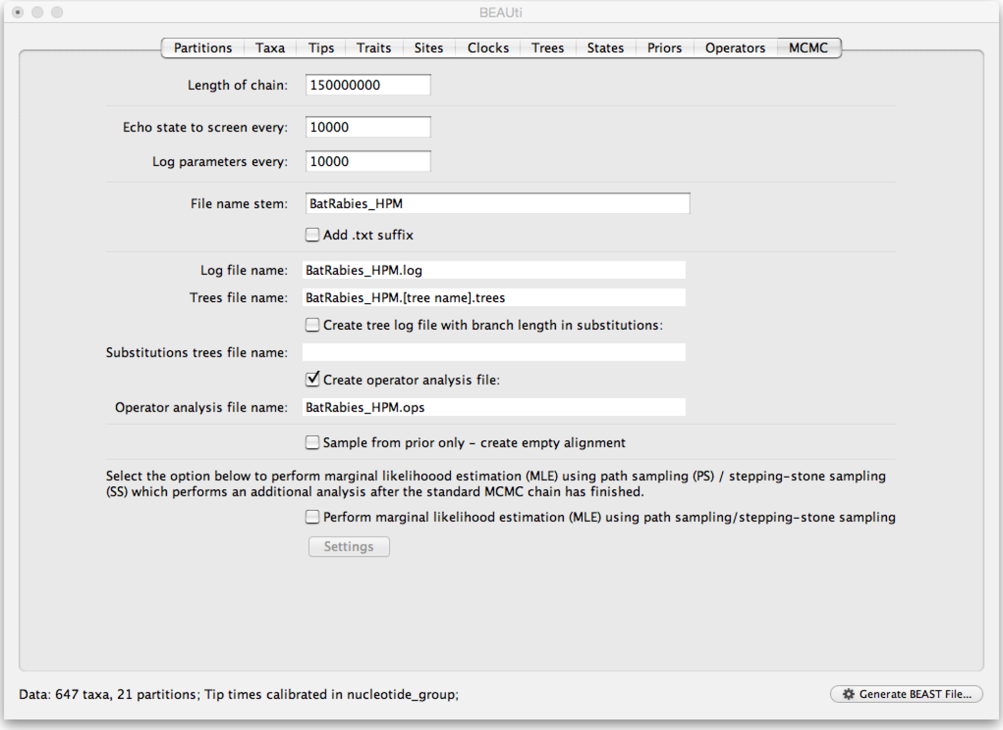
Hyperprior Shape: 0.01

Hyperprior Scale: 100.0

```
Quantiles: 2.5%:   -9.8
```

```
Quantiles: 2.5%:   3.523E-159
```

Finally, shift select the first and the last clock.rate to select all clock.rate parameters and ‘Link parameters into a hierarchical model’. Provide clock.rate as unique name and set the Normal Hyperprior Stdev to 100.0 and the Gamma Hyperprior Shape and Scale to 0.001 and 1000.0 respectively and click OK. Keep the default pop size priors. In the mcmc panel, set the chain frequency to 150,000,000 and the logging frequency to 10,000 and provide ‘BatRabies_HPM’ as File stem name:



Length of chain: 150000000

Echo state to screen every: 10000

Log parameters every: 10000

File name stem: BatRabies_HPM

□ Add .txt suffix

Log file name: BatRabies_HPM.log

Trees file name: BatRabies_HPM.[tree name].trees

□ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☑ Create operator analysis file:

Operator analysis file name: BatRabies_HPM.ops

□ Sample from prior only - create empty alignment

Select the option below to perform marginal likelihood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) which performs an additional analysis after the standard MCMC chain has finished.

□ Perform marginal likelihood estimation (MLE) using path sampling/stepping-stone sampling

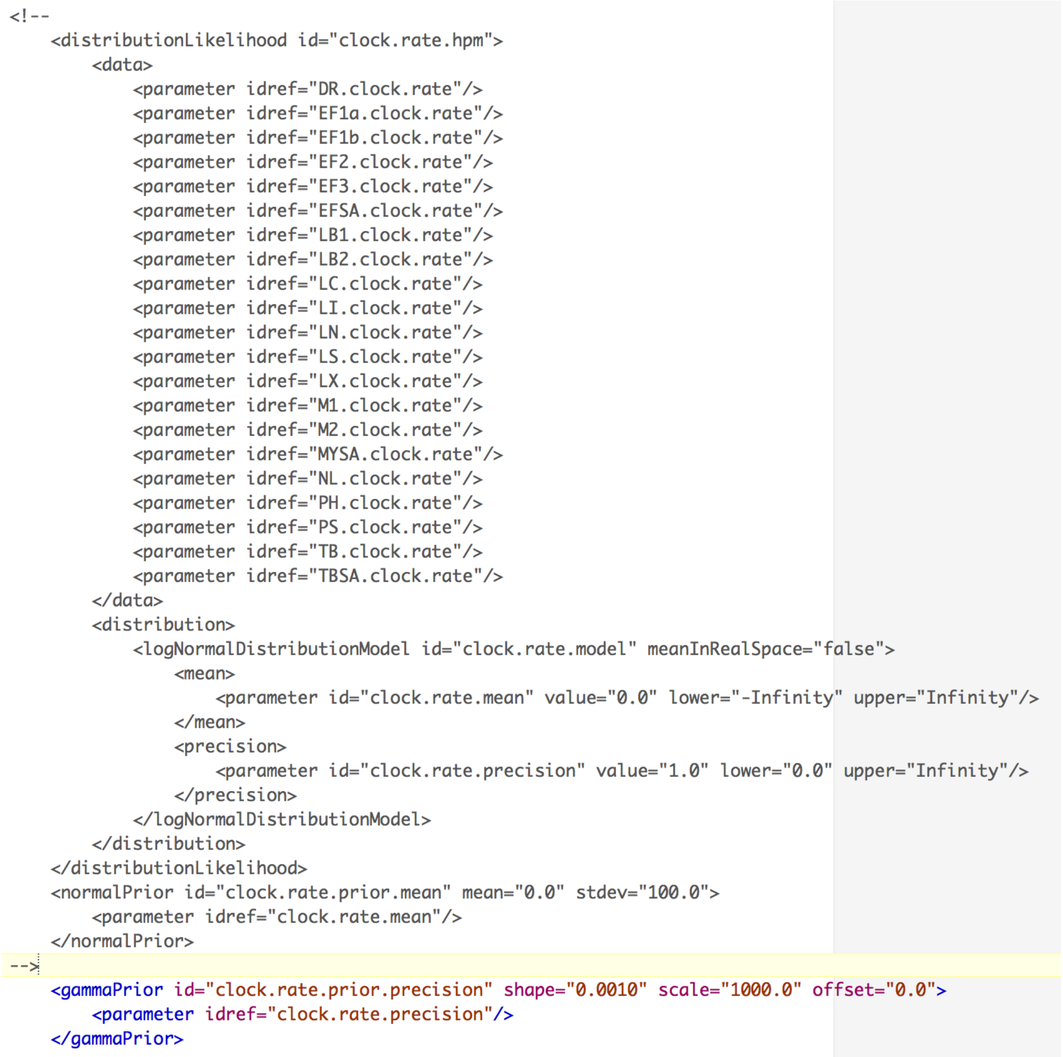
Data: 647 taxa, 21 partitions; Tip times calibrated in nucleotide_group;

Check that the xml file runs in BEAST, but there is no real need to run this to completion.

Introducing fixed effects

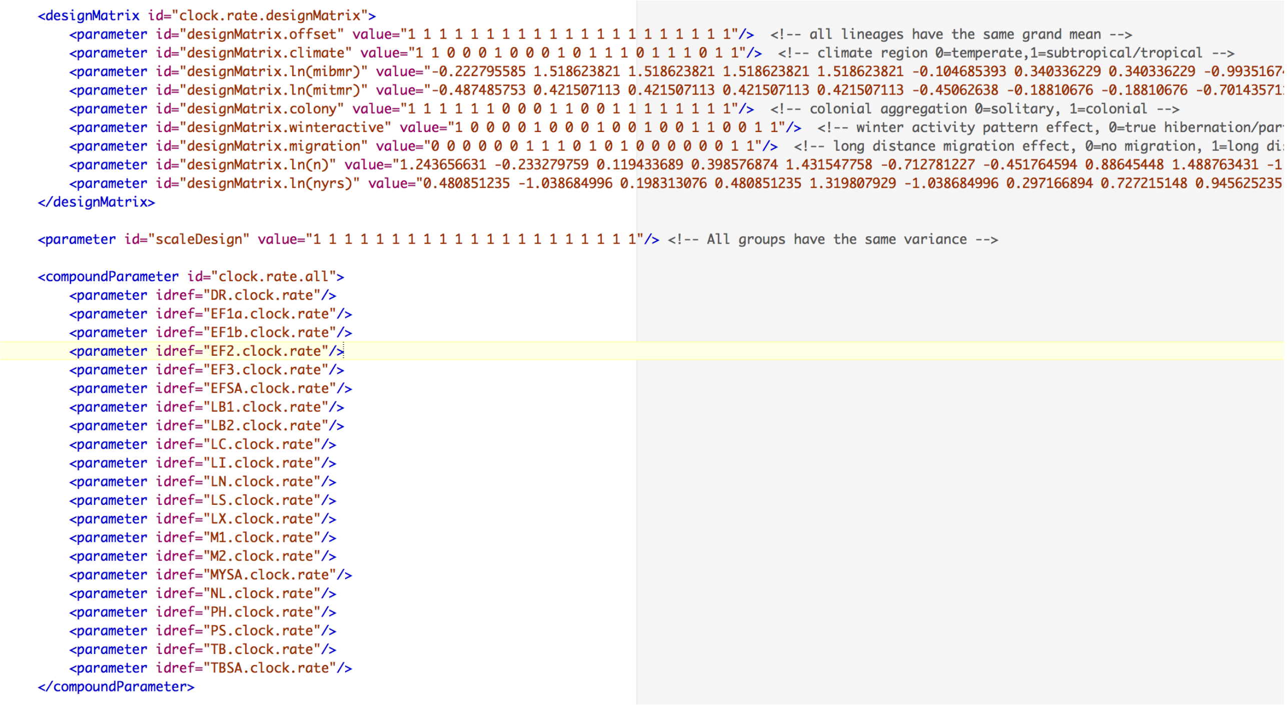
In the next section, the xml edits will be highlighted that are required for introducing fixed effects. We will use this to test several potential predictors of evolutionary rate variation among bat RABV lineages. An xml that includes these edits is provided for guidance (look for start and end mixed effect modelling edit comments).

First comment out the clock.rate.hpm distributionLikelihood and clock.rate.prior.mean normalPrior, but keep the clock.rate.prior.precision gammaPrior:



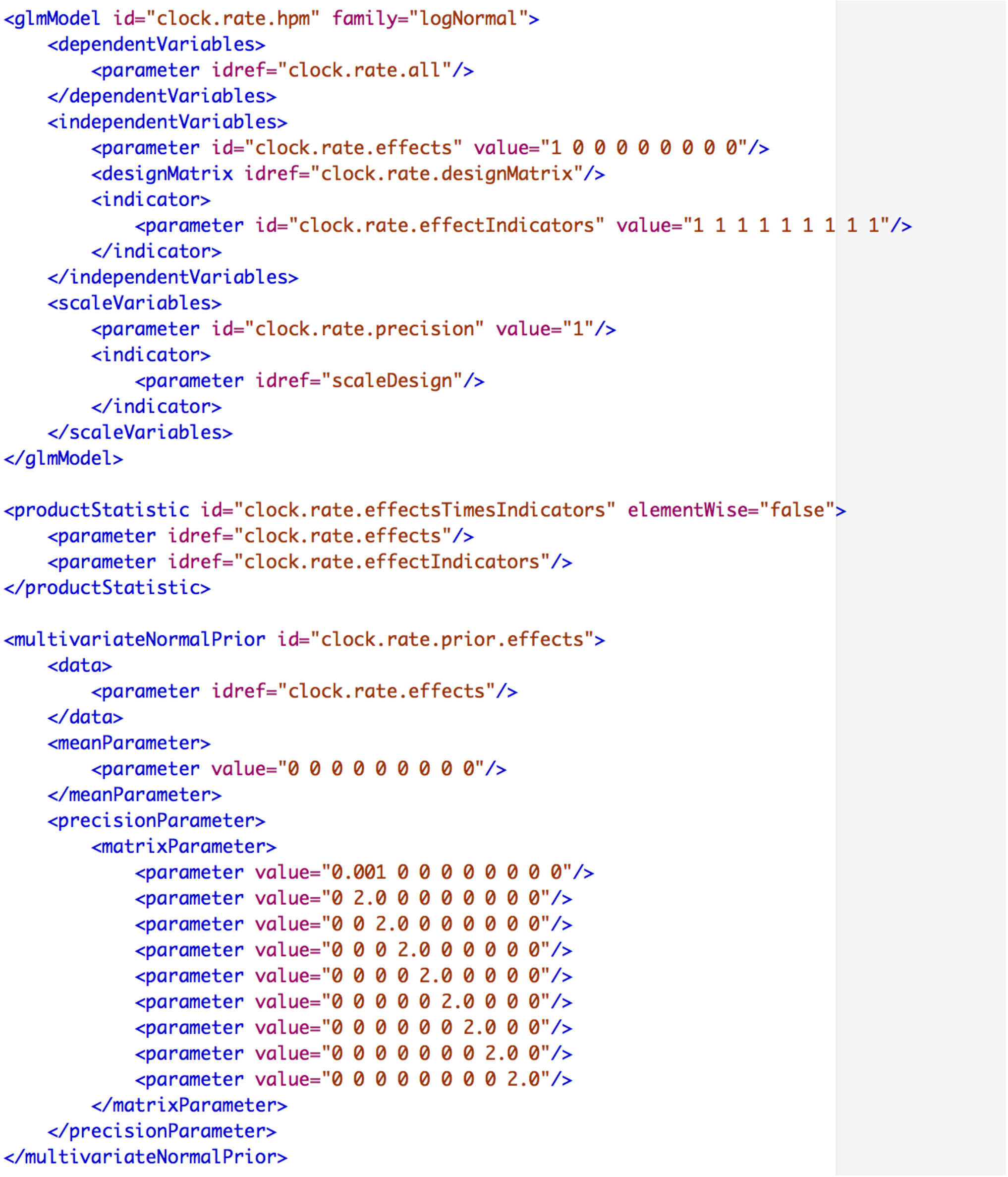
```
<!--
    <distributionLikelihood id="clock.rate.hpm">
        <data>
            <parameter idref="DR.clock.rate"/>
            <parameter idref="EF1a.clock.rate"/>
            <parameter idref="EF1b.clock.rate"/>
            <parameter idref="EF2.clock.rate"/>
            <parameter idref="EF3.clock.rate"/>
            <parameter idref="EFSA.clock.rate"/>
            <parameter idref="LB1.clock.rate"/>
            <parameter idref="LB2.clock.rate"/>
            <parameter idref="LC.clock.rate"/>
            <parameter idref="LI.clock.rate"/>
            <parameter idref="LN.clock.rate"/>
            <parameter idref="LS.clock.rate"/>
            <parameter idref="LX.clock.rate"/>
            <parameter idref="M1.clock.rate"/>
            <parameter idref="M2.clock.rate"/>
            <parameter idref="MYSA.clock.rate"/>
            <parameter idref="NL.clock.rate"/>
            <parameter idref="PH.clock.rate"/>
            <parameter idref="PS.clock.rate"/>
            <parameter idref="TB.clock.rate"/>
            <parameter idref="TBSA.clock.rate"/>
        </data>
        <distribution>
            <logNormalDistributionModel id="clock.rate.model" meanInRealSpace="false">
                <mean>
                    <parameter id="clock.rate.mean" value="0.0" lower="-Infinity" upper="Infinity"/>
                </mean>
                <precision>
                    <parameter id="clock.rate.precision" value="1.0" lower="0.0" upper="Infinity"/>
                </precision>
            </logNormalDistributionModel>
        </distribution>
    </distributionLikelihood>
    <normalPrior id="clock.rate.prior.mean" mean="0.0" stdev="100.0">
        <parameter idref="clock.rate.mean"/>
    </normalPrior>
-->
    <gammaPrior id="clock.rate.prior.precision" shape="0.0010" scale="1000.0" offset="0.0">
        <parameter idref="clock.rate.precision"/>
    </gammaPrior>
```

Before this clock.rate.prior.precision gammaPrior, add a designMatrix, a scaleDesign and a compoundParameter including all the rate parameter. The designMatrix includes the offset (the same grand mean for all lineages) and the different predictors. The latter includes boolean indicators (for climatic regions, colonial aggregation, winter activity, long-distance migration) and log-transformed, standardised continuous predictors (for physiological traits: basal metabolic rate and torpid metabolic rate, and for sampling characteristics: number of samples and sampling time interval for each viral lineage).



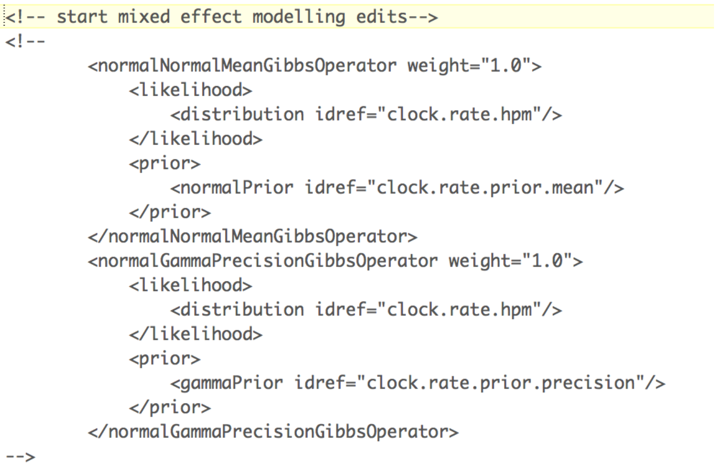
```
<designMatrix id="clock.rate.designMatrix">
    <parameter id="designMatrix.offset" value="1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1"/>  <!-- all lineages have the same grand mean -->
    <parameter id="designMatrix.climate" value="1 1 0 0 0 1 0 0 0 1 0 1 1 1 0 1 1 1 0 1 1"/>  <!-- climate region 0=temperate,1=subtropical/tropical -->
    <parameter id="designMatrix.ln(mibmr)" value="-0.222795585 1.518623821 1.518623821 1.518623821 1.518623821 -0.104685393 0.340336229 0.340336229 -0.9935167
    <parameter id="designMatrix.ln(mitmr)" value="-0.487485753 0.421507113 0.421507113 0.421507113 0.421507113 -0.45062638 -0.18810676 -0.18810676 -0.70143571
    <parameter id="designMatrix.colony" value="1 1 1 1 1 1 0 0 0 1 1 0 0 1 1 1 1 1 1 1 1"/>  <!-- colonial aggregation 0=solitary, 1=colonial -->
    <parameter id="designMatrix.winteractive" value="1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 1 1 0 0 1 1"/>  <!-- winter activity pattern effect, 0=true hibernation/par
    <parameter id="designMatrix.migration" value="0 0 0 0 0 0 1 1 1 0 1 0 1 0 0 0 0 0 0 1 1"/>  <!-- long distance migration effect, 0=no migration, 1=long di
    <parameter id="designMatrix.ln(n)" value="1.243656631 -0.233279759 0.119433689 0.398576874 1.431547758 -0.712781227 -0.451764594 0.88645448 1.488763431 -1
    <parameter id="designMatrix.ln(nyrs)" value="0.480851235 -1.038684996 0.198313076 0.480851235 1.319807929 -1.038684996 0.297166894 0.727215148 0.945625235
</designMatrix>

<parameter id="scaleDesign" value="1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1"/> <!-- All groups have the same variance -->

<compoundParameter id="clock.rate.all">
    <parameter idref="DR.clock.rate"/>
    <parameter idref="EF1a.clock.rate"/>
    <parameter idref="EF1b.clock.rate"/>
    <parameter idref="EF2.clock.rate"/>
    <parameter idref="EF3.clock.rate"/>
    <parameter idref="EFSA.clock.rate"/>
    <parameter idref="LB1.clock.rate"/>
    <parameter idref="LB2.clock.rate"/>
    <parameter idref="LC.clock.rate"/>
    <parameter idref="LI.clock.rate"/>
    <parameter idref="LN.clock.rate"/>
    <parameter idref="LS.clock.rate"/>
    <parameter idref="LX.clock.rate"/>
    <parameter idref="M1.clock.rate"/>
    <parameter idref="M2.clock.rate"/>
    <parameter idref="MYSA.clock.rate"/>
    <parameter idref="NL.clock.rate"/>
    <parameter idref="PH.clock.rate"/>
    <parameter idref="PS.clock.rate"/>
    <parameter idref="TB.clock.rate"/>
    <parameter idref="TBSA.clock.rate"/>
</compoundParameter>
```

Under the compoundParameter and before the clock.rate.prior.precision gammaPrior, add the glmModel, a productStatistic (optional) and a multivariateNormalPrior for the fixed effects. The dimensions of the clock.rate.effects and clock.rate.effectIndicators in the glommed should match the number of effects in the designMatrix. The same hold true for the multivariateNormalPrior:



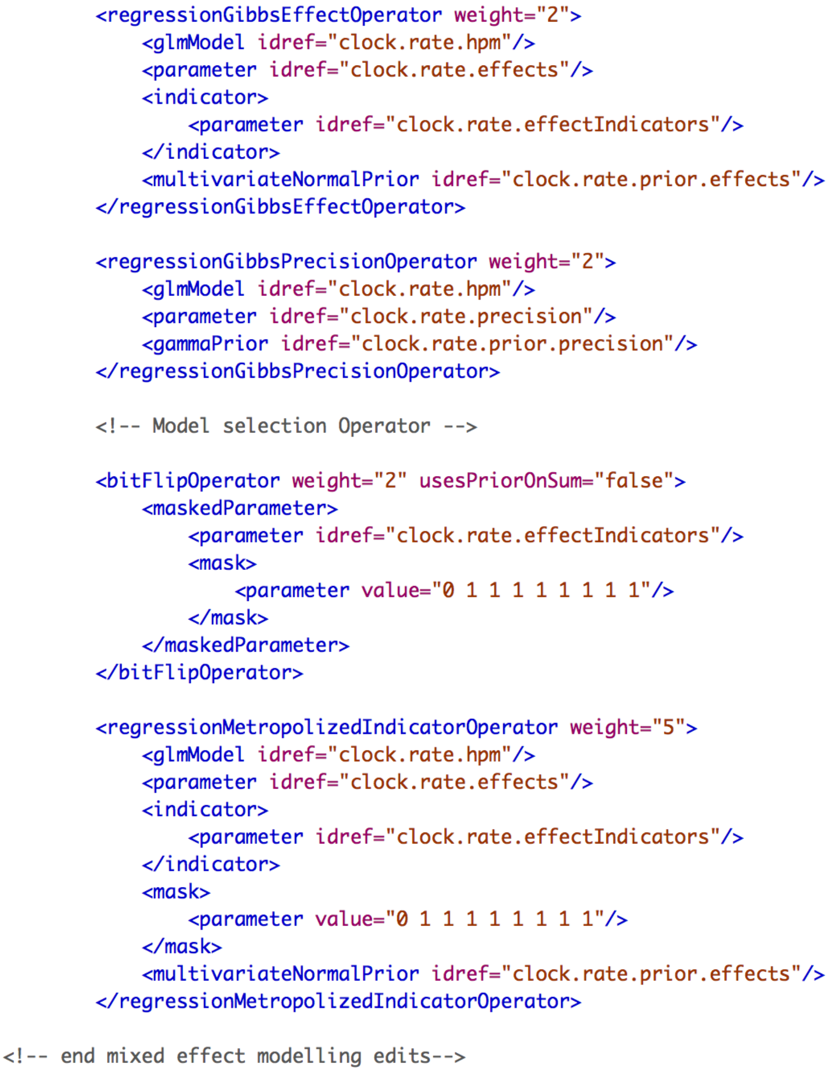
```
<glmModel id="clock.rate.hpm" family="logNormal">
    <dependentVariables>
        <parameter idref="clock.rate.all"/>
    </dependentVariables>
    <independentVariables>
        <parameter id="clock.rate.effects" value="1 0 0 0 0 0 0 0 0"/>
        <designMatrix idref="clock.rate.designMatrix"/>
        <indicator>
            <parameter id="clock.rate.effectIndicators" value="1 1 1 1 1 1 1 1 1"/>
        </indicator>
    </independentVariables>
    <scaleVariables>
        <parameter id="clock.rate.precision" value="1"/>
        <indicator>
            <parameter idref="scaleDesign"/>
        </indicator>
    </scaleVariables>
</glmModel>

<productStatistic id="clock.rate.effectsTimesIndicators" elementWise="false">
    <parameter idref="clock.rate.effects"/>
    <parameter idref="clock.rate.effectIndicators"/>
</productStatistic>

<multivariateNormalPrior id="clock.rate.prior.effects">
    <data>
        <parameter idref="clock.rate.effects"/>
    </data>
    <meanParameter>
        <parameter value="0 0 0 0 0 0 0 0 0"/>
    </meanParameter>
    <precisionParameter>
        <matrixParameter>
            <parameter value="0.001 0 0 0 0 0 0 0 0"/>
            <parameter value="0 2.0 0 0 0 0 0 0 0"/>
            <parameter value="0 0 2.0 0 0 0 0 0 0"/>
            <parameter value="0 0 0 2.0 0 0 0 0 0"/>
            <parameter value="0 0 0 0 2.0 0 0 0 0"/>
            <parameter value="0 0 0 0 0 2.0 0 0 0"/>
            <parameter value="0 0 0 0 0 0 2.0 0 0"/>
            <parameter value="0 0 0 0 0 0 0 2.0 0"/>
            <parameter value="0 0 0 0 0 0 0 0 2.0"/>
        </matrixParameter>
    </precisionParameter>
</multivariateNormalPrior>
```

In the operator block, comment out the normalNormalMeanGibbsOperator on the clock.rate.prior.mean and the normalGammaPrecisionGibbsOperator on the normalGammaPrecisionGibbsOperator:



```
<!-- start mixed effect modelling edits-->
<!--
        <normalNormalMeanGibbsOperator weight="1.0">
            <likelihood>
                <distribution idref="clock.rate.hpm"/>
            </likelihood>
            <prior>
                <normalPrior idref="clock.rate.prior.mean"/>
            </prior>
        </normalNormalMeanGibbsOperator>
        <normalGammaPrecisionGibbsOperator weight="1.0">
            <likelihood>
                <distribution idref="clock.rate.hpm"/>
            </likelihood>
            <prior>
                <gammaPrior idref="clock.rate.prior.precision"/>
            </prior>
        </normalGammaPrecisionGibbsOperator>
-->
```

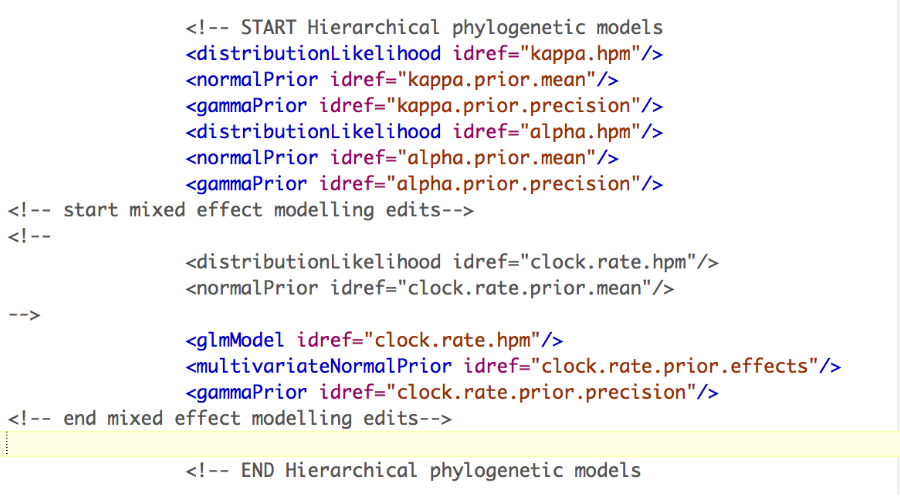
And add in a regressionGibbsEffectOperator, a regressionGibbsPrecisionOperator, a bitFlipOperator and a regressionMetropolizedIndicatorOperator:



```
        <regressionGibbsEffectOperator weight="2">
            <glmModel idref="clock.rate.hpm"/>
            <parameter idref="clock.rate.effects"/>
            <indicator>
                <parameter idref="clock.rate.effectIndicators"/>
            </indicator>
            <multivariateNormalPrior idref="clock.rate.prior.effects"/>
        </regressionGibbsEffectOperator>

        <regressionGibbsPrecisionOperator weight="2">
            <glmModel idref="clock.rate.hpm"/>
            <parameter idref="clock.rate.precision"/>
            <gammaPrior idref="clock.rate.prior.precision"/>
        </regressionGibbsPrecisionOperator>

        <!-- Model selection Operator -->

        <bitFlipOperator weight="2" usesPriorOnSum="false">
            <maskedParameter>
                <parameter idref="clock.rate.effectIndicators"/>
                <mask>
                    <parameter value="0 1 1 1 1 1 1 1 1"/>
                </mask>
            </maskedParameter>
        </bitFlipOperator>

        <regressionMetropolizedIndicatorOperator weight="5">
            <glmModel idref="clock.rate.hpm"/>
            <parameter idref="clock.rate.effects"/>
            <indicator>
                <parameter idref="clock.rate.effectIndicators"/>
            </indicator>
            <mask>
                <parameter value="0 1 1 1 1 1 1 1 1"/>
            </mask>
            <multivariateNormalPrior idref="clock.rate.prior.effects"/>
        </regressionMetropolizedIndicatorOperator>

<!-- end mixed effect modelling edits-->
```

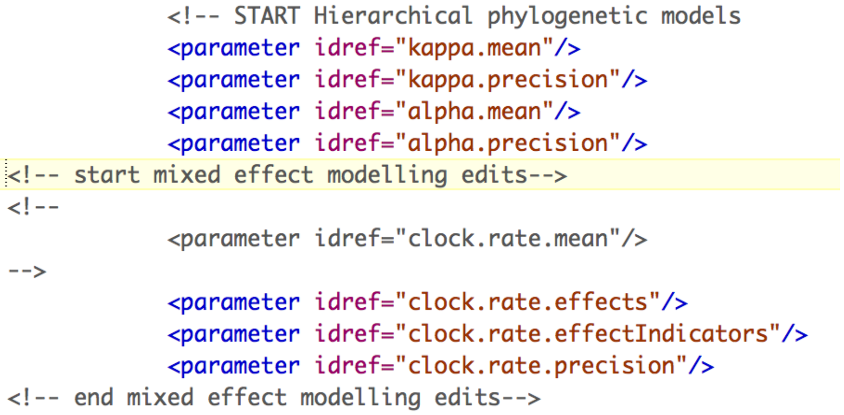
Note that we use a mask on the indicator vector when operating on the indicators to avoid operating on the grand mean (which should always be included).

In the prior block, comment out the distributionLikelihood on the clock.rate.hpm and the normalPrior on the clock.rate.prior.mean and refer to the glmModel and multivariateNormalPrior on the clock.rate.prior.effects instead. Keep the gammaPrior on the clock.rate.prior.precision:



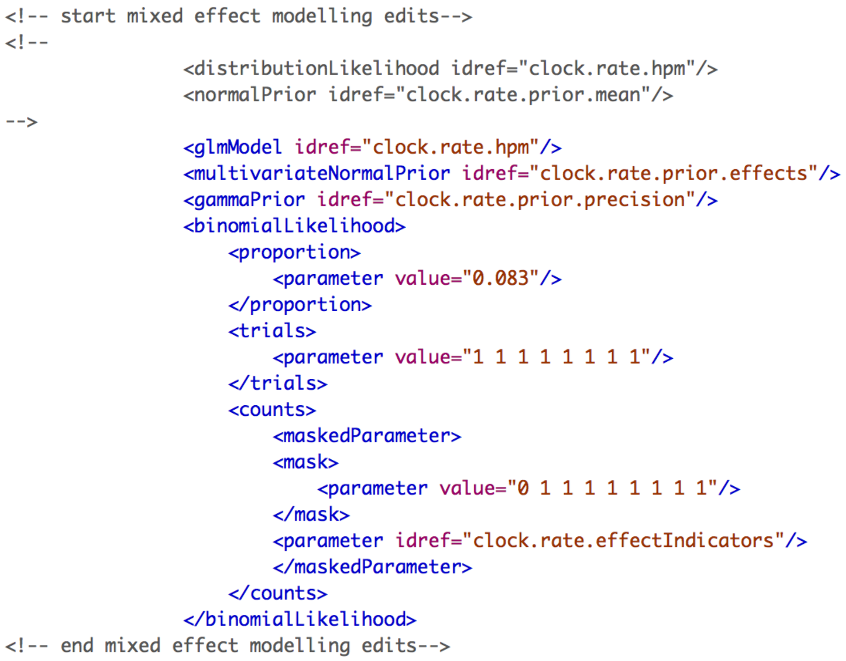
```
                <!-- START Hierarchical phylogenetic models
                <distributionLikelihood idref="kappa.hpm"/>
                <normalPrior idref="kappa.prior.mean"/>
                <gammaPrior idref="kappa.prior.precision"/>
                <distributionLikelihood idref="alpha.hpm"/>
                <normalPrior idref="alpha.prior.mean"/>
                <gammaPrior idref="alpha.prior.precision"/>
<!-- start mixed effect modelling edits-->
<!--
                <distributionLikelihood idref="clock.rate.hpm"/>
                <normalPrior idref="clock.rate.prior.mean"/>
-->
                <glmModel idref="clock.rate.hpm"/>
                <multivariateNormalPrior idref="clock.rate.prior.effects"/>
                <gammaPrior idref="clock.rate.prior.precision"/>
<!-- end mixed effect modelling edits-->

                <!-- END Hierarchical phylogenetic models
```

Finally, in the fileLog, comment out the clock.rate.mean parameter, keep the clock.rate.precision and add in the clock.rate.effects and clock.rate.effectIndicators:



```
            <!-- START Hierarchical phylogenetic models
            <parameter idref="kappa.mean"/>
            <parameter idref="kappa.precision"/>
            <parameter idref="alpha.mean"/>
            <parameter idref="alpha.precision"/>
<!-- start mixed effect modelling edits-->
<!--
            <parameter idref="clock.rate.mean"/>
-->
            <parameter idref="clock.rate.effects"/>
            <parameter idref="clock.rate.effectIndicators"/>
            <parameter idref="clock.rate.precision"/>
<!-- end mixed effect modelling edits-->
```

Check that the xml file runs in BEAST. Output fils are provided for the complete analysis. Are there any predictors that help to explain the rate variation among the host-associated bat RABV lineages. What support do they get? Note that we assumed implicitly a 0.5 prior success probability for each predictor. It may be useful to explore a different prior that prefers a smaller inclusion probability a priori. For example, we could assigning independent Bernoulli prior probability distributions on the indicators and use a small prior probability on each predictor's inclusion that reflects a 50% prior probability on no predictors being included, e.g.:



```
<!-- start mixed effect modelling edits-->
<!--
                <distributionLikelihood idref="clock.rate.hpm"/>
                <normalPrior idref="clock.rate.prior.mean"/>
-->
                <glmModel idref="clock.rate.hpm"/>
                <multivariateNormalPrior idref="clock.rate.prior.effects"/>
                <gammaPrior idref="clock.rate.prior.precision"/>
                <binomialLikelihood>
                    <proportion>
                        <parameter value="0.083"/>
                    </proportion>
                    <trials>
                        <parameter value="1 1 1 1 1 1 1 1"/>
                    </trials>
                    <counts>
                        <maskedParameter>
                        <mask>
                            <parameter value="0 1 1 1 1 1 1 1 1"/>
                        </mask>
                        <parameter idref="clock.rate.effectIndicators"/>
                        </maskedParameter>
                    </counts>
                </binomialLikelihood>
<!-- end mixed effect modelling edits-->
```

What effect will this have on the support for the predictors?

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**/ or **https://github.com/beast-dev/beast-mcmc**
- Tutorials: **http://beast.bio.ed.ac.uk/Tutorials/**
- Frequently asked questions: **http://beast.bio.ed.ac.uk/FAQ/**

References

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4.

Bryant, J. E., E. C. Holmes, and A. D. Barrett. 2007. Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. PLoS pathogens 3:e75.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution 29:1969-1973.

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular biology and evolution 29:2157-2167.

Baele, G., W. L. Li, A. J. Drummond, M. A. Suchard, and P. Lemey. 2013. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Molecular biology and evolution 30:239-243.

Streicker, D. G., P. Lemey, A. Velasco-Villa, and C. E. Rupprecht. 2012. Rates of viral evolution are linked to host geography in bat rabies. PLoS pathogens 8:e1002720.

Streicker, D. G., A. S. Turmelle, M. J. Vonhof, I. V. Kuzmin, G. F. McCracken, and C. E. Rupprecht. 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science 329:676-679.