BEAST

Bayesian Evolutionary Analysis Sampling Trees

Revealing the evolutionary dynamics of influenza

Introduction

This tutorial provides a step-by-step explanation on how to reconstruct the evolutionary dynamics of influenza based on a set of virus sequences which have been isolated at different points in time ('heterochronous' data) using *BEAST*. We will focus on influenza A virus evolution, in particular on the emergence of swine-origin influenza A (H1N1) virus in 2009 (H1N1/09) and on the epidemic dynamics of H3N2 in the New York State. The H1N1/09 data set is a subset of an analyzed set genomes in a study that provides insights into the origins and evolutionary genomics of this outbreak (Smith et al., 2009). The H3N2 data is a subset of a comprehensive data set spanning several epidemic seasons in the New York state, which has been used to unravel the genomic and epidemiological dynamics of this virus (Rambaut et al., 2008). In the first exercise, the aim is to obtain an estimate of the rate of molecular evolution, an estimate of the date of the most recent common ancestor, an estimate of the H1N1/09 epidemic growth or the H1N1/09 basic reproductive number. In the second exercise, we will examine how H3N2 diversity fluctuates through time.

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 v1.8.4. *BEAST* releases are generally available for download from http://beast.bio.ed.ac.uk/, but the latest (pre-)releases can also be found at https://github.com/beast-dev/beast-mcmc/releases.

- **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 v1.4.3. It is available for download from http://tree.bio.ed.ac.uk/.

EXERCISE 1: The swine-origin influenza A outbreak

Running *BEAUti*

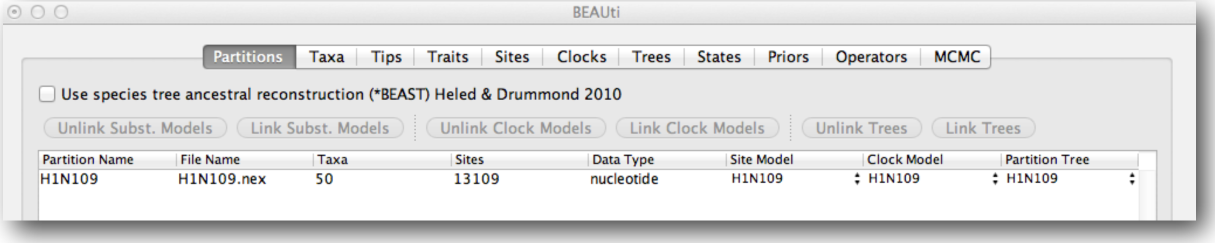
The program *BEAUti* is a user-friendly program for setting the model parameters for *BEAST*. Run *BEAUti* by double clicking on its icon.

Loading the NEXUS file

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

The NEXUS alignment

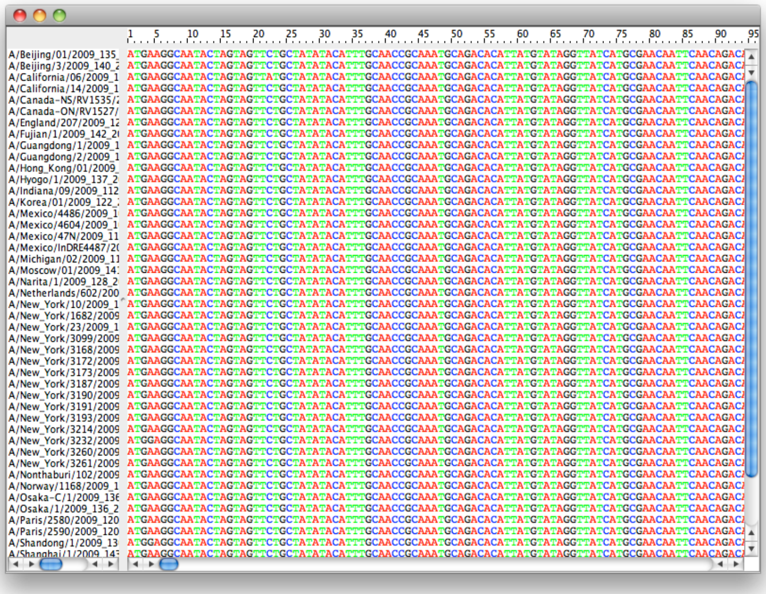
Select the file called H1N109.nex. This file contains an alignment of 50 genomes (concatenated segments), 13109 nucleotides in length. Once loaded, the new data will be listed under **Partitions** as shown in the figure:



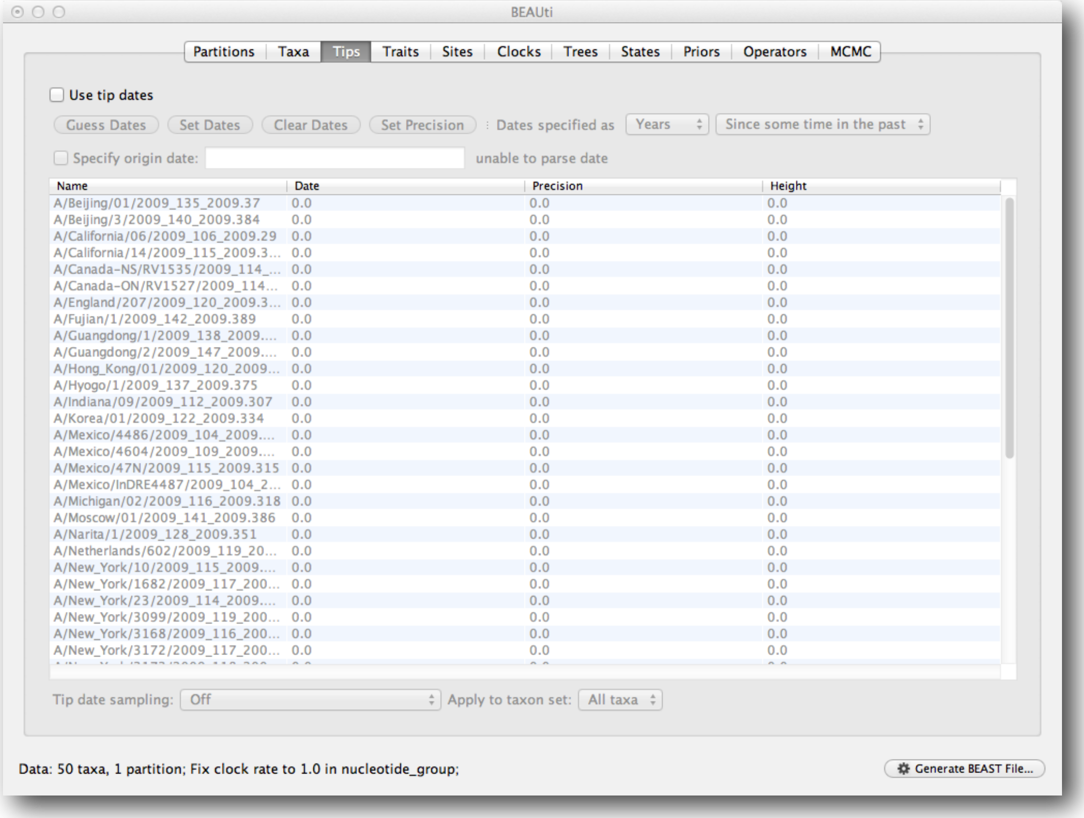
□ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

Partition Name	File Name	Taxa	Sites	Data Type	Site Model	Clock Model	Partition Tree
H1N109	H1N109.nex	50	13109	nucleotide	H1N109	H1N109	H1N109

Double-click on the row of the table (but not on Partition Name) to display the actual sequence alignment:



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 sequences have been sampled from the H1N1/09 epidemic between March and May 2009. To set these dates switch to the ‘Tips’ panel using the tabs at the top of the window:



□ Use tip dates

Dates specified as: Years; Since some time in the past

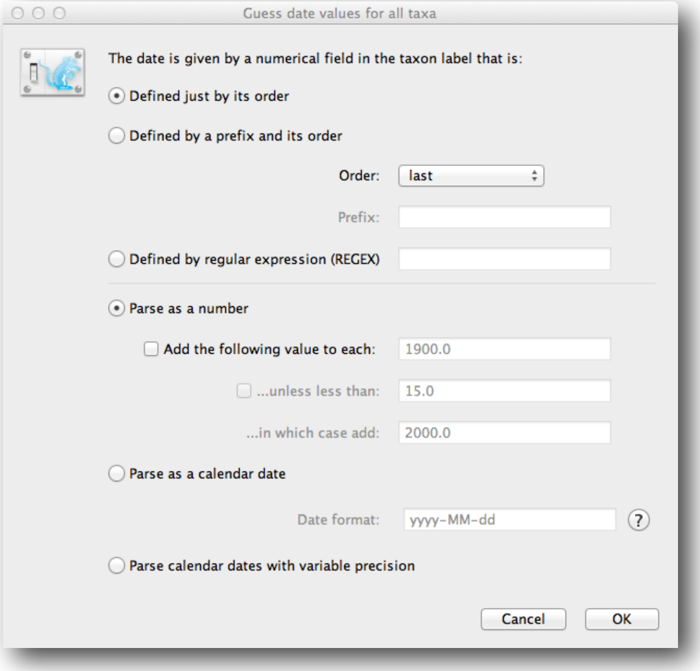
□ Specify origin date: unable to parse date

Name	Date	Precision	Height
A/Beijing/01/2009_135_2009.37	0.0	0.0	0.0
A/Beijing/3/2009_140_2009.384	0.0	0.0	0.0
A/California/06/2009_106_2009.29	0.0	0.0	0.0
A/California/14/2009_115_2009.3...	0.0	0.0	0.0
A/Canada-NS/RV1535/2009_114_...	0.0	0.0	0.0
A/Canada-ON/RV1527/2009_114...	0.0	0.0	0.0
A/England/207/2009_120_2009.3...	0.0	0.0	0.0
A/Fujian/1/2009_142_2009.389	0.0	0.0	0.0
A/Guangdong/1/2009_138_2009....	0.0	0.0	0.0
A/Guangdong/2/2009_147_2009....	0.0	0.0	0.0
A/Hong_Kong/01/2009_120_2009...	0.0	0.0	0.0
A/Hyogo/1/2009_137_2009.375	0.0	0.0	0.0
A/Indiana/09/2009_112_2009.307	0.0	0.0	0.0
A/Korea/01/2009_122_2009.334	0.0	0.0	0.0
A/Mexico/4486/2009_104_2009....	0.0	0.0	0.0
A/Mexico/4604/2009_109_2009....	0.0	0.0	0.0
A/Mexico/47N/2009_115_2009.315	0.0	0.0	0.0
A/Mexico/InDRE4487/2009_104_2...	0.0	0.0	0.0
A/Michigan/02/2009_116_2009.318	0.0	0.0	0.0
A/Moscow/01/2009_141_2009.386	0.0	0.0	0.0
A/Narita/1/2009_128_2009.351	0.0	0.0	0.0
A/Netherlands/602/2009_119_20...	0.0	0.0	0.0
A/New_York/10/2009_115_2009....	0.0	0.0	0.0
A/New_York/1682/2009_117_200...	0.0	0.0	0.0
A/New_York/23/2009_114_2009....	0.0	0.0	0.0
A/New_York/3099/2009_119_200...	0.0	0.0	0.0
A/New_York/3168/2009_116_200...	0.0	0.0	0.0
A/New_York/3172/2009_117_200...	0.0	0.0	0.0

Tip date sampling: Off Apply to taxon set: All taxa

Data: 50 taxa, 1 partition; Fix clock rate to 1.0 in nucleotide_group;

Select the box labelled '**Use tip dates**'. The actual sampling in fractional years is encoded 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: last

Prefix:

○ Defined by regular expression (REGEX)

● Parse as a number

□ Add the following value to each: 1900.0

□ ...unless less than: 15.0

...in which case add: 2000.0

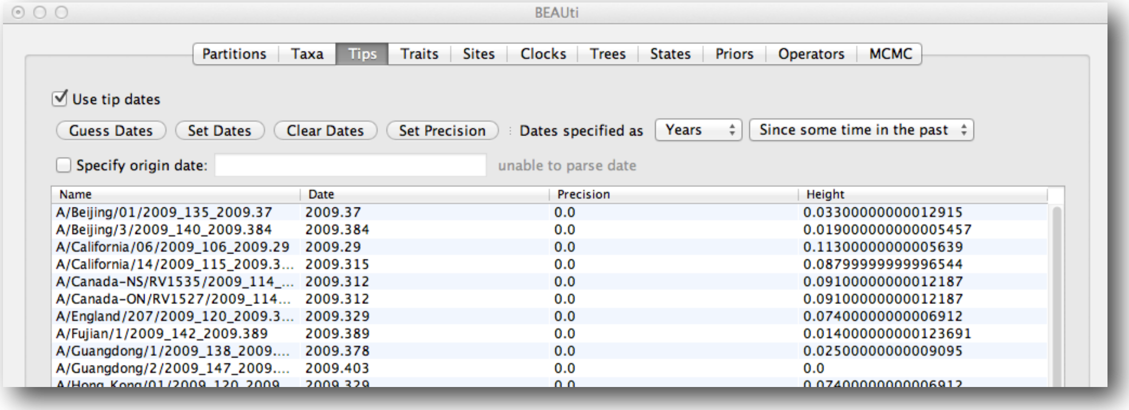
○ Parse as a calendar date

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 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).

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. These operations are not necessary in our case since the dates are fully specified at the end of the sequence names. There is also an option to parse calendar dates and one for calendar dates with various precisions. For the H1N1/09 sequences you can keep the default 'Defined just by its order' and select 'last' from the drop-down menu for the order and press 'OK'. The dates will appear in the appropriate column of the main window. You can then check these and edit them manually as required. 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 2009) or backwards in time from the present (as in the case of radiocarbon ages).



☑ Use tip dates

Dates specified as Years · Since some time in the past

☐ Specify origin date: unable to parse date

Name	Date	Precision	Height
A/Beijing/01/2009_135_2009.37	2009.37	0.0	0.03300000000012915
A/Beijing/3/2009_140_2009.384	2009.384	0.0	0.019000000000005457
A/California/06/2009_106_2009.29	2009.29	0.0	0.11300000000005639
A/California/14/2009_115_2009.3...	2009.315	0.0	0.08799999999996544
A/Canada-NS/RV1535/2009_114_...	2009.312	0.0	0.09100000000012187
A/Canada-ON/RV1527/2009_114...	2009.312	0.0	0.09100000000012187
A/England/207/2009_120_2009.3...	2009.329	0.0	0.07400000000006912
A/Fujian/1/2009_142_2009.389	2009.389	0.0	0.014000000000123691
A/Guangdong/1/2009_138_2009....	2009.378	0.0	0.02500000000009095
A/Guangdong/2/2009_147_2009....	2009.403	0.0	0.0
A/Hong_Kong/01/2009_120_2009	2009.329	0.0	0.07400000000006912

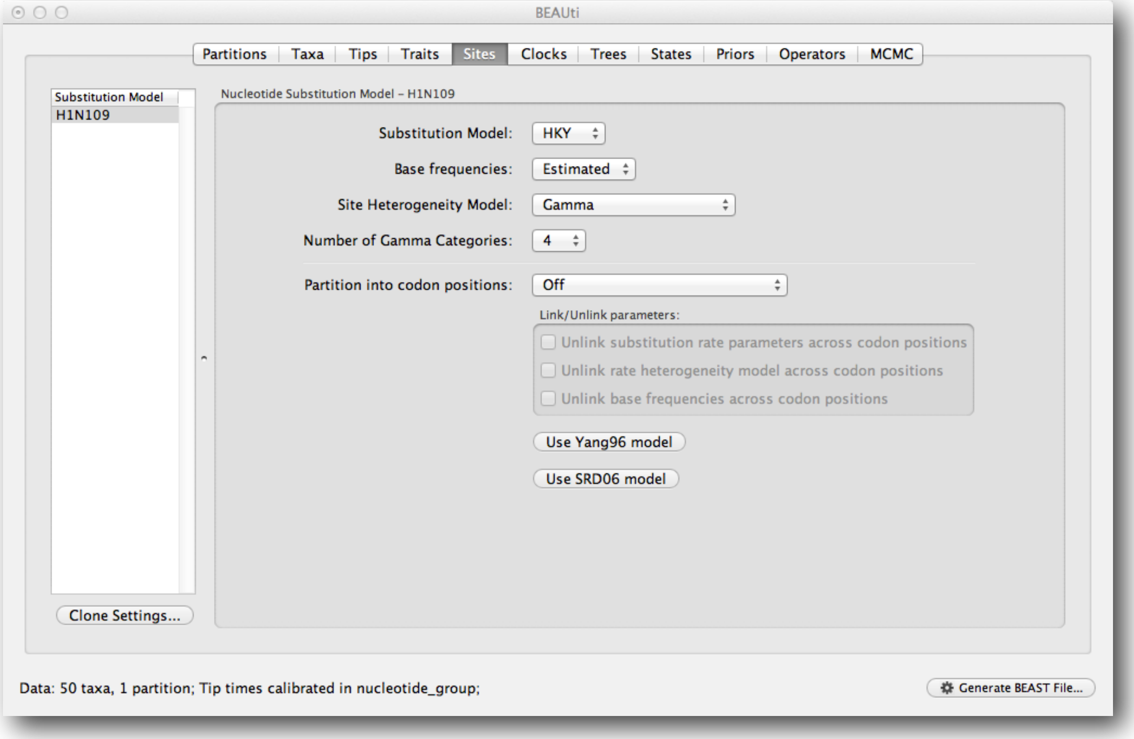
The "Height" column lists the ages of the tips relative to time 0 (in our case 2009.403). The Precision column allows specifying with what precision the sampling time is know. To include taxa only known up to the year of sampling for example (e.g., 2009), a precision of 1 year can be set and the age of those tips can be integrated over the time interval of 1 year using the Tip date sampling option at the bottom left of the Tips panel.

Setting the substitution 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.

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 a set of rate heterogeneity parameters (gamma shape parameter and/or proportion of invariant sites) for each codon position.



Substitution Model
H1N109

Nucleotide Substitution Model – H1N109

Substitution Model: HKY

Base frequencies: Estimated

Site Heterogeneity Model: Gamma

Number of Gamma Categories: 4

Partition into codon positions: Off

Link/Unlink parameters:

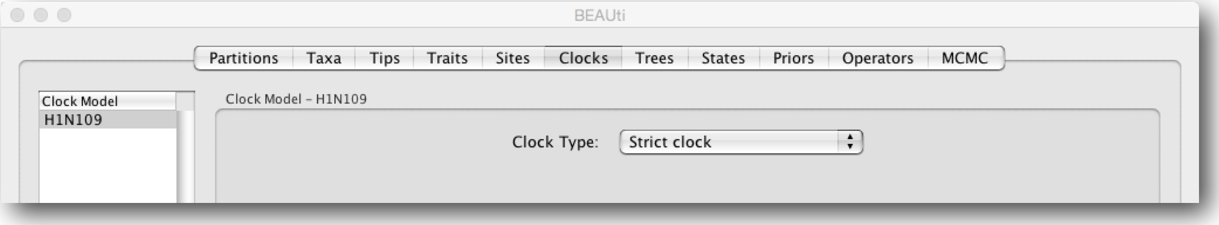
□ Unlink substitution rate parameters across codon positions
□ Unlink rate heterogeneity model across codon positions
□ Unlink base frequencies across codon positions

Data: 50 taxa, 1 partition; Tip times calibrated in nucleotide_group;

For this tutorial, keep the default ‘HKY’ model, the default ‘Estimated’ base frequencies and select ‘Gamma’ as ‘Site Heterogeneity Model’ (with 4 discrete categories) before proceeding to the ‘Clocks’ tab.

Setting the ‘molecular clock’ model

The ‘Molecular Clock Model’ options allows us to choose between a strict and a relaxed (uncorrelated lognormal or uncorrelated exponential) clock. Because of the low diversity data we analyze here, a relaxed clock would probably be over-parameterization. Hence, we keep a strict clock setting.



Clock Model
H1N109

Clock Model – H1N109

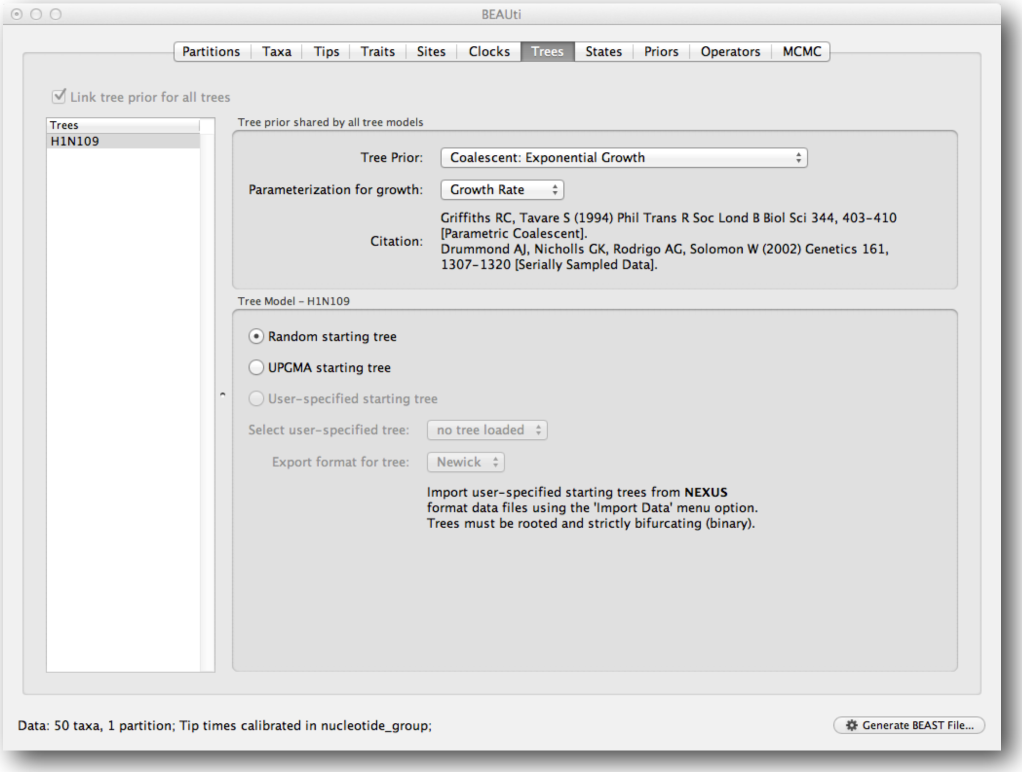
Clock Type: Strict clock

Now move on to the ‘Trees’ panel.

Setting the tree prior

This panel contains settings about the tree. Firstly the starting tree is specified to be ‘randomly generated’. The other main setting here is to specify the ‘Tree prior’ which describes how the population size is expected to change over time for coalescent models. The default tree prior is set to a constant size coalescent prior.

To estimate the epidemic growth rate, we will change this demographic model to an exponential growth coalescent prior, which is intuitively appealing for viral outbreaks. Switch the option for ‘Tree Prior’ to ‘Coalescent: Exponential Growth’.



☑ Link tree prior for all trees

H1N109

Tree prior shared by all tree models

Tree Prior: Coalescent: Exponential Growth

Parameterization for growth: Growth Rate

Citation: Griffiths RC, Tavare S (1994) Phil Trans R Soc Lond B Biol Sci 344, 403–410 [Parametric Coalescent].
Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Genetics 161, 1307–1320 [Serially Sampled Data].

Tree Model – H1N109

● Random starting tree

○ UPGMA starting tree

○ User-specified starting tree

Select user-specified tree: no tree loaded

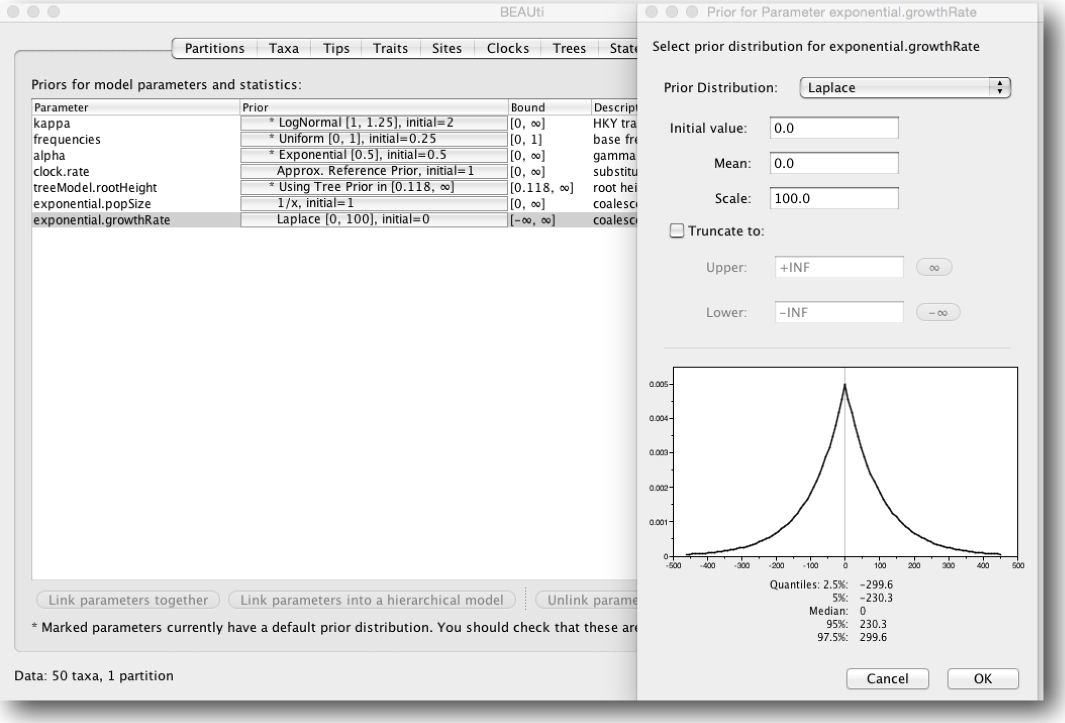
Export format for tree: Newick

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: 50 taxa, 1 partition; Tip times calibrated in nucleotide_group;

Setting up the priors

Now switch to the ‘Priors’ tab. This panel has a table showing every parameter of the currently selected model and what the prior distribution is for each. A prior allows the user to ‘inform’ the analysis by selecting a particular distribution. Although some of the default priors may be improper, with sufficiently informative data the posterior becomes proper. If priors or not set, they will appear in red.



Priors for model parameters and statistics:

Parameter	Prior	Bound	Descript
kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY tra
frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	base fre
alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamma
clock.rate	Approx. Reference Prior, initial=1	[0, ∞]	substitu
treeModel.rootHeight	* Using Tree Prior in [0.118, ∞]	[0.118, ∞]	root hei
exponential.popSize	1/x, initial=1	[0, ∞]	coalesc
exponential.growthRate	Laplace [0, 100], initial=0	[−∞, ∞]	coalesc

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

Data: 50 taxa, 1 partition

Select prior distribution for exponential.growthRate

Prior Distribution: Laplace

Initial value: 0.0

Mean: 0.0

Scale: 100.0

□ Truncate to:

Upper: +INF ∞

Lower: −INF −∞

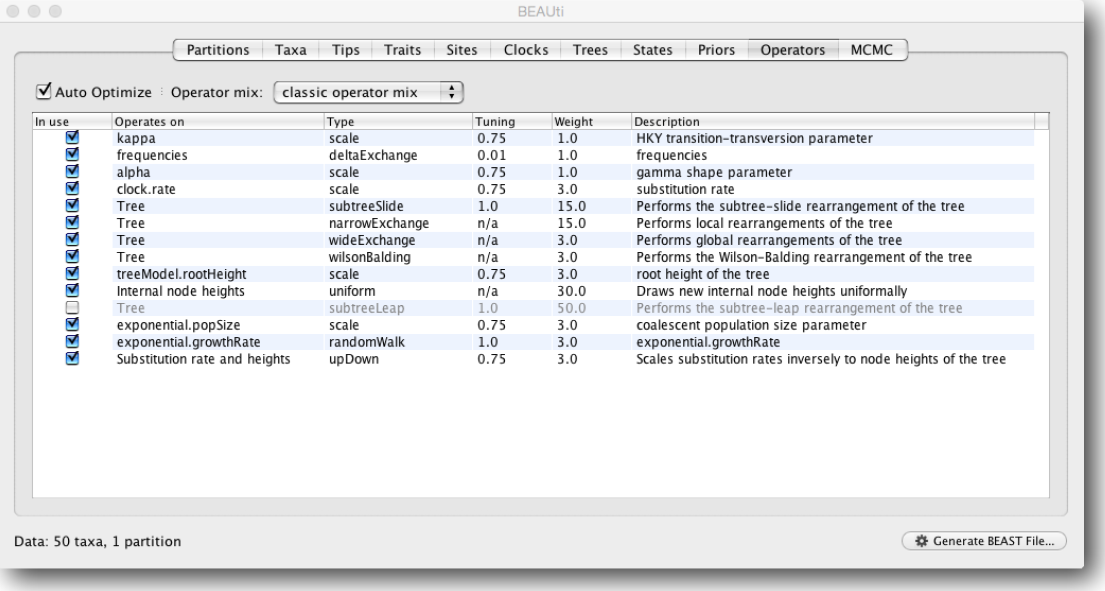
Quantiles: 2.5%: −299.6
5%: −230.3
Median: 0
95%: 230.3
97.5%: 299.6

Note that a prior distribution must be specified for every parameter and whilst *BEAUti* provides default options these are not necessarily tailored to the problem and data being analyzed. In this case, the default Laplace prior prefers relatively small growth rates while this parameter take can take on relatively large values on this epidemic scale. Therefore, we will increase the variance of this prior distribution by setting the scale to 100.

The default prior on the rate of evolution (clock.rate) is an approximation of a conditional reference prior (`Approx. Reference Prior`) (Ferreira and Suchard, 2008). If the sequences are not associated with different sampling dates (they are contemporaneous), or when the sampling time range is trivial for the evolutionary scale of the taxa, the substitution rate can be fixed to a value based on another source, or better, a prior distribution can be specified to also incorporate the uncertainty of this 'external' rate. Fixing the rate 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*). Note that when selecting to fix the rate to a value, the transition kernel(s) on this parameter ('Operators' panel, see next section) will be automatically unselected.

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:



☑ Auto Optimize Operator mix: classic operator mix

In use	Operates on	Type	Tuning	Weight	Description
☑	kappa	scale	0.75	1.0	HKY transition-transversion parameter
☑	frequencies	deltaExchange	0.01	1.0	frequencies
☑	alpha	scale	0.75	1.0	gamma shape parameter
☑	clock.rate	scale	0.75	3.0	substitution rate
☑	Tree	subtreeSlide	1.0	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
☐	Tree	subtreeLeap	1.0	50.0	Performs the subtree-leap rearrangement of the tree
☑	exponential.popSize	scale	0.75	3.0	coalescent population size parameter
☑	exponential.growthRate	randomWalk	1.0	3.0	exponential.growthRate
☑	Substitution rate and heights	upDown	0.75	3.0	Scales substitution rates inversely to node heights of the tree

Data: 50 taxa, 1 partition

In the first column are the parameter names. These will be called things like `kappa` which means the HKY model's kappa parameter (the transition-transversion bias). 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 proportion, the random walk operator adds or subtracts an amount to the parameter 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.

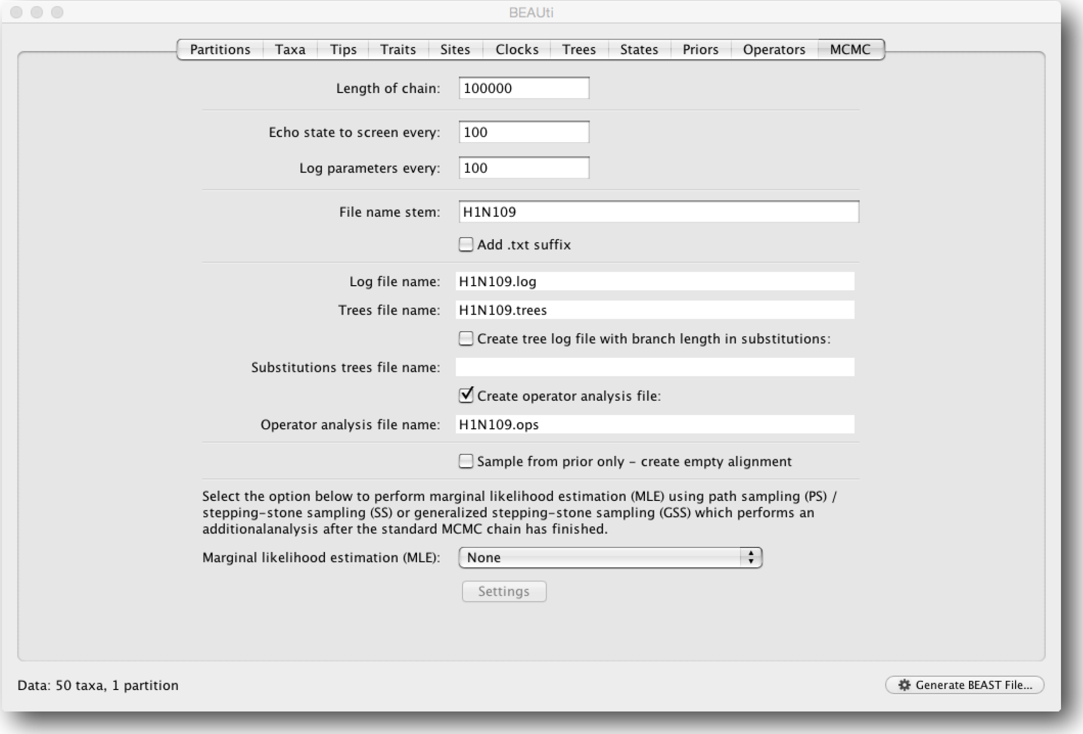
The next column, labelled '`Tuning`', gives a tuning setting to the operator. Some operators don't have any tuning settings so have n/a under this column. The tuning parameter will determine how large a move each operator will make which will affect how often that change is accepted by the MCMC which will in turn affect the efficiency of the analysis. For most operators (like random walk and subtree slide operators) 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 will be written to standard output. These can then be used to set the starting tuning settings in order to minimize the amount of time taken to reach optimum performance in subsequent runs.

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. As of BEAST v1.8.4, different options are available w.r.t. exploring tree space. In this tutorial, we will use the 'classic operator mix', which consists of of set of tree transition kernels that propose changes to the tree. There is also an option to fix the tree topology as well as a 'new experimental mix', which is currently under development with the aim to improve mixing for large phylogenetic trees.

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 be depends 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 analysed using *Tracer* in order to examine whether a particular chain length is adequate. Change the chain length to 1,000,000 for our initial test run.



Length of chain: 100000

Echo state to screen every: 100

Log parameters every: 100

File name stem: H1N109

□ Add .txt suffix

Log file name: H1N109.log

Trees file name: H1N109.trees

□ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☑ Create operator analysis file:

Operator analysis file name: H1N109.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) or generalized stepping-stone sampling (GSS) which performs an additionalanalysis after the standard MCMC chain has finished.

Marginal likelihood estimation (MLE): None

Settings

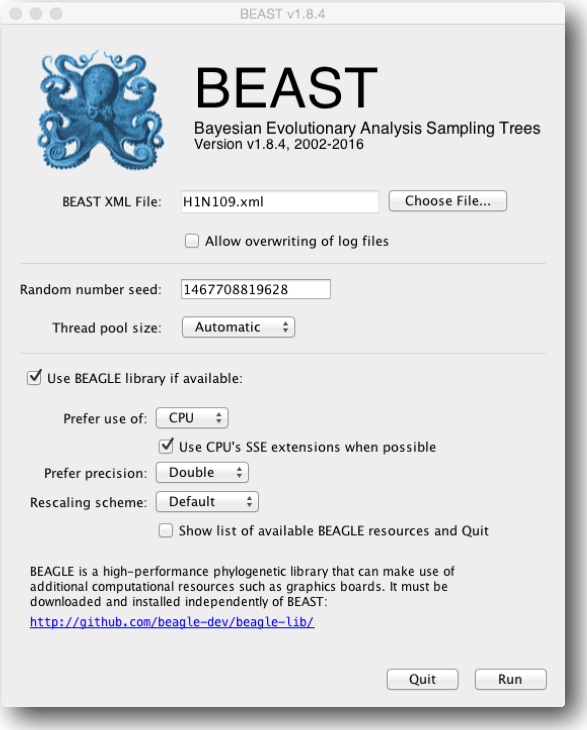
Data: 50 taxa, 1 partition

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 the 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. A useful exercise could be to examine the sensitity of the growth rate estimates to different scale values for this prior distribution (e.g. scale = 1, 10, 100).

The next option allows the user to set the File stem name; if not set to 'H1N109' 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 based on the file stem name. You can also log the operator analysis to a file. 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, which is not needed in this exercise. So, 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. BEAST will ask you to review the prior settings one more time before saving the file. Continue and keep the default name for the file, add the xml extension (H1N109.xml) and save the file.

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 dialog box will appear in which you select the XML file:



BEAST v1.8.4

BEAST

Bayesian Evolutionary Analysis Sampling Trees
Version v1.8.4, 2002-2016

BEAST XML File: H1N109.xml

□ Allow overwriting of log files

Random number seed: 1467708819628

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

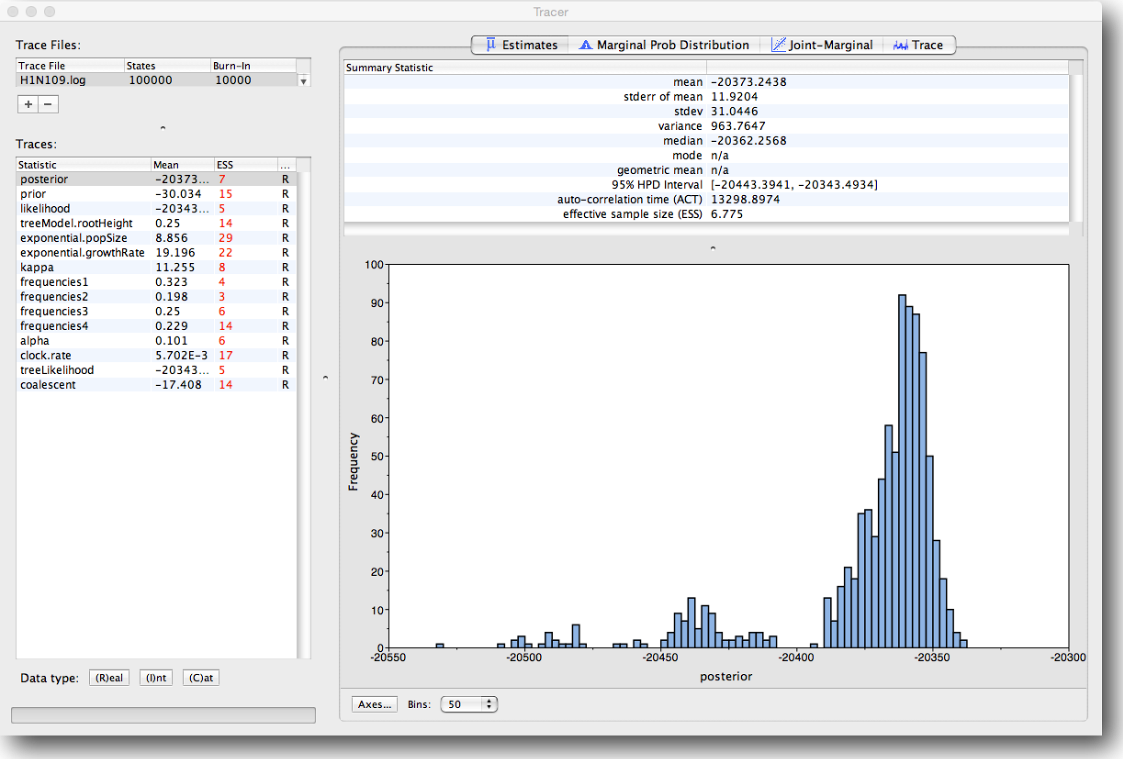
BEAGLE is a high-performance phylogenetic library that can make use of additional computational resources such as graphics boards. It must be downloaded and installed independently of BEAST:
http://github.com/beagle-dev/beagle-lib/

Press the 'Choose File' button and select the XML file you just created and press 'Run'. 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. If the command line version of *BEAST* is being used then the name of the XML file is given after the name of the *BEAST* executable. The analysis will then 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.

Analyzing 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. Double click on the *Tracer* icon; 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
H1N109.log	100000	10000

Traces:

Statistic	Mean	ESS	...
posterior	-20373...	7	R
prior	-30.034	15	R
likelihood	-20343...	5	R
treeModel.rootHeight	0.25	14	R
exponential.popSize	8.856	29	R
exponential.growthRate	19.196	22	R
kappa	11.255	8	R
frequencies1	0.323	4	R
frequencies2	0.198	3	R
frequencies3	0.25	6	R
frequencies4	0.229	14	R
alpha	0.101	6	R
clock.rate	5.702E-3	17	R
treeLikelihood	-20343...	5	R
coalescent	-17.408	14	R

Summary Statistic	
mean	-20373.2438
stderr of mean	11.9204
stdev	31.0446
variance	963.7647
median	-20362.2568
mode	n/a
geometric mean	n/a
95% HPD Interval	[-20443.3941, -20343.4934]
auto-correlation time (ACT)	13298.8974
effective sample size (ESS)	6.775

On the left hand side is the name of the log file loaded and the traces that it contains. There are traces for the posterior (this is the log of the product of the tree likelihood and the prior probabilities), and the continuous parameters. Selecting a trace on the left brings up analyses for this trace on the right hand side depending on tab that is selected. When first opened, the `posterior' trace is selected and various statistics of this trace are shown under the Estimates tab.

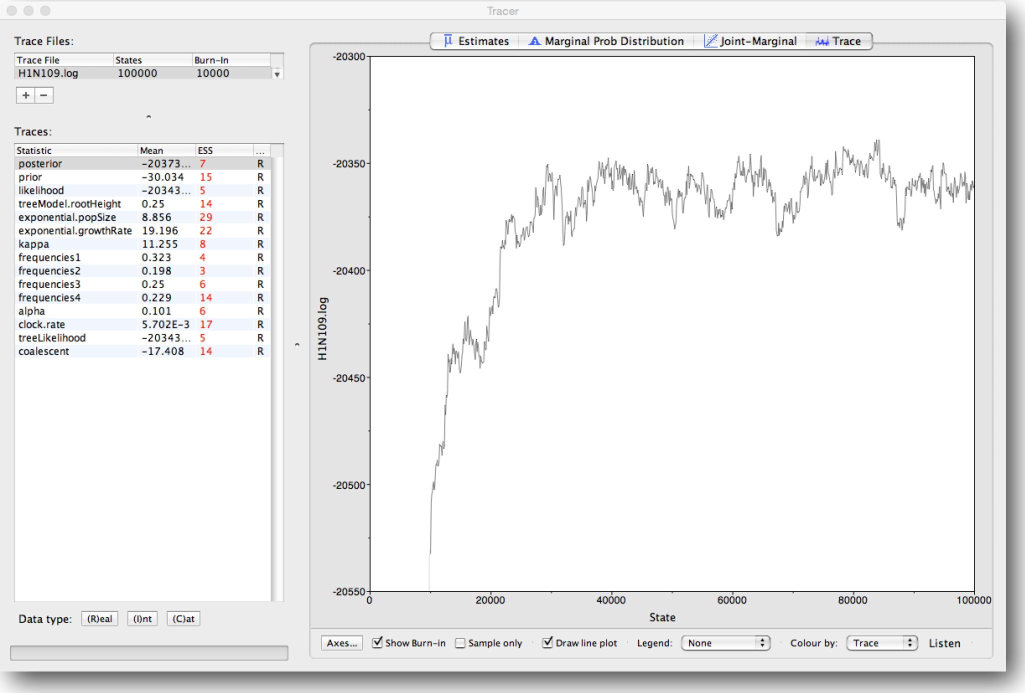
In the top right of the window is a table of calculated statistics for the selected trace. The statistics and their meaning are described in the table below.

- **Mean** - The mean value of the samples (excluding the burn-in).
- **Stdev** - 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).
- **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 200 and 100 are highlighted in yellow and red respectively by *Tracer*). 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 (e.g for treeModel.rootHeight), that is, the sampled values against the step in the MCMC chain.



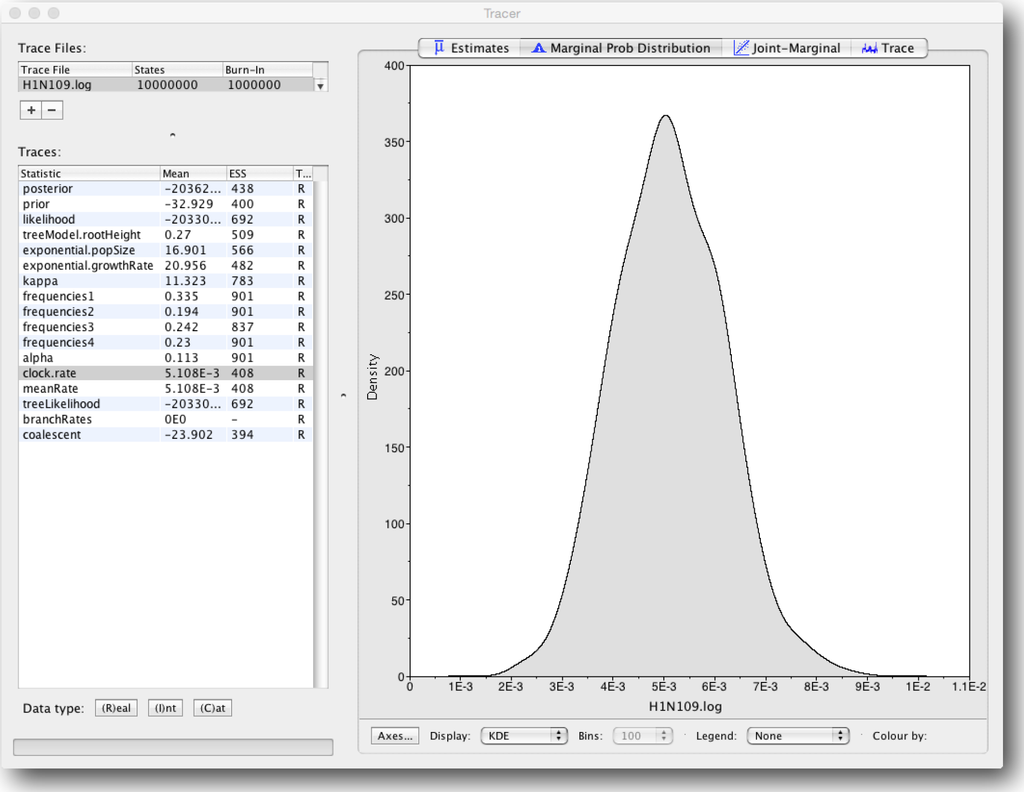
Trace File	States	Burn-In
H1N109.log	100000	10000

Statistic	Mean	ESS	...
posterior	-20373...	7	R
prior	-30.034	15	R
likelihood	-20343...	5	R
treeModel.rootHeight	0.25	14	R
exponential.popSize	8.856	29	R
exponential.growthRate	19.196	22	R
kappa	11.255	8	R
frequencies1	0.323	4	R
frequencies2	0.198	3	R
frequencies3	0.25	6	R
frequencies4	0.229	14	R
alpha	0.101	6	R
clock.rate	5.702E-3	17	R
treeLikelihood	-20343...	5	R
coalescent	-17.408	14	R

Here you can see how the samples are correlated. There are 1000 samples in the trace (we ran the MCMC for 1,000,000 steps sampling every 1000) but it is clear that adjacent samples often tend to have similar values. The ESS for the clock.rate is about 17, so we are only getting 1 independent sample to every 60 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 analysis needs to be run longer. The lowest ESS of about 3 suggests that we have to run it at least 35 times longer to get ESSs that are >100. However, it would be better to aim higher (e.g. a chain length of 10,000,000 and sampling every 10,000 generations). If the previous analysis ran reasonably fast and if time permits, you can can go back to *BEAUti* and set up and run this longer analysis, but it is probably advisable to proceed with summarizing the longer runs that are provided with this tutorial. Load the new log file into *Tracer* (you can leave the old one loaded for comparison). Click on the Trace tab and look at the raw trace plot.

Again we have chosen options that produce 1000 samples and with an ESS of > 400 for the coalescent model parameters there is little auto-correlation between the samples. 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 satisfied with the behavior of the MCMC 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 shows a plot of the posterior probability density of this parameter. You should see a plot similar to this:



Trace Files:

Trace File	States	Burn-In
H1N109.log	10000000	1000000

Traces:

Statistic	Mean	ESS	T...
posterior	-20362...	438	R
prior	-32.929	400	R
likelihood	-20330...	692	R
treeModel.rootHeight	0.27	509	R
exponential.popSize	16.901	566	R
exponential.growthRate	20.956	482	R
kappa	11.323	783	R
frequencies1	0.335	901	R
frequencies2	0.194	901	R
frequencies3	0.242	837	R
frequencies4	0.23	901	R
alpha	0.113	901	R
clock.rate	5.108E-3	408	R
meanRate	5.108E-3	408	R
treeLikelihood	-20330...	692	R
branchRates	0E0	-	R
coalescent	-23.902	394	R

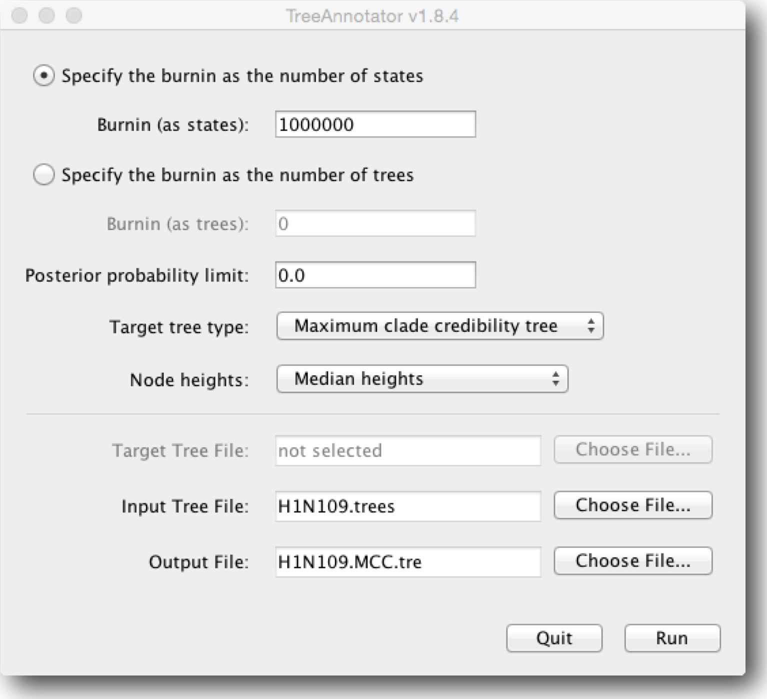
As you can see the posterior probability density is roughly bell-shaped. When looking at the equivalent histogram in the Estimates panel, there is some sampling noise which is smoothened by the KDE; this would be reduced if we ran the chain for longer but we already have a reasonable estimate of the mean and HPD interval. The **treeModel.rootHeight** parameter provides an estimate of the time to the most recent common ancestor since the most recent sampling data (in our case: 2009.403). What would be the mean estimate for the date of the MRCA?

The **exponential.growthRate** (r) provides an estimate of the epidemic growth of H1N1/09. Given that $N_t = N_0\,e^{-rt}$ (with N_0 being the population size at present), the doubling time for $r = 21$ is about 0.03 years or 12 days. Interestingly, it has been shown that the basic reproductive ratio (R_0) is related to the growth rate (see http://tree.bio.ed.ac.uk/wiki/pages/t769F5D1/Relationship_between_R0_and_the_epidemic_growth_rate.html). However, the basic reproductive number is dependent not just on an estimate of *r*, but also a good estimate of the generation time distribution, which reflects the time between successive infections in a chain of transmission. If we assume a generation time distribution that follows the gamma distribution, then $R_0 = (1 + \ r / b)$ ^*a*, where *a* and *b* are the parameters of the gamma distribution (and $a = \mu^2 / \sigma^2$, $b = \mu / \sigma^2$). Taking μ = 3 days and σ = 2 days, what would be the mean estimate for the H1N1/09 R_0?

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 v1.8.4

● Specify the burnin as the number of states

Burnin (as states): 1000000

○ 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: H1N109.trees

Output File: H1N109.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 relaxed clock 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 10,000,000 steps, a 10% burnin corresponds to 1,000,000 steps. Alternatively, sampling every 10,000 steps results in 1000 trees in the file, and to obtain at the same 10% burnin, the number of trees needs to be set to 100.

- **Posterior probability limit** - This is the minimum posterior probability for a node in order for *TreeAnnotator* to store the annotated information. The default is 0.0 so all nodes will have information summarized independent of their posterior probability.

- **Target tree type** - This has three options ‘Maximum clade credibility tree’ or ‘User target tree’ For the latter option, a NEXUS tree file can be specified as the Target Tree File, below. Select the first 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. Node heights can also be summarised as a Mean or a Median over the sample of trees. Sometimes a mean or median height for a node may actually be higher than the mean or median height of its parental node (because particular ancestral-descendent relationships in the MCC tree may still be different compared to a large number of other tree sampled). This will result in artifactual negative branch lengths, but can be avoided by the ‘Common Ancestor heights’ option. Let’s use the default Median heights for our summary tree.

- **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.

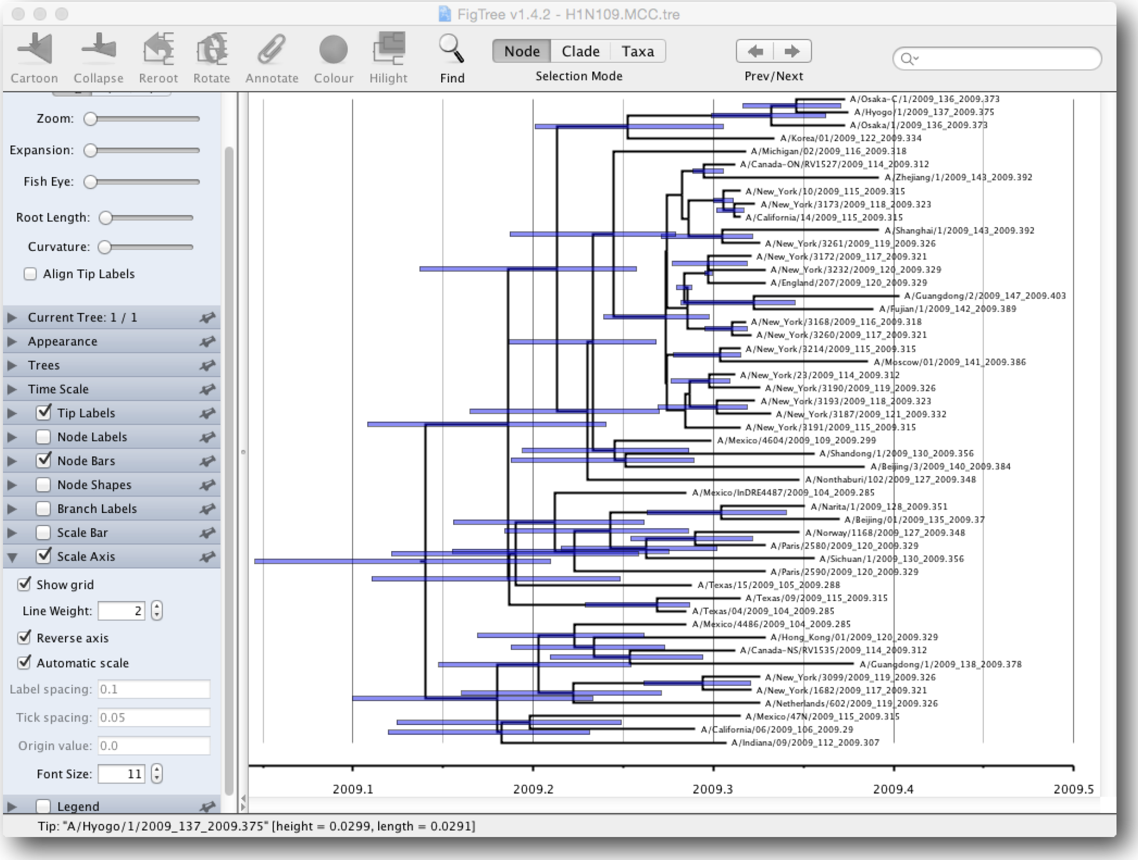
- **Input Tree File** - Use the ‘Choose File...’ button to select an input trees file. This will be the trees file produced by *BEAST*.

- **Output File** - Select a name for the output tree file (e.g., H1N109.MCC.tre).

Once you have selected all the options, above, press the ‘**Run**’ button. *TreeAnnotator* will analyse 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.


FigTree v1.4.2 - H1N109.MCC.tre
Cartoon
Collapse
Reroot
Rotate
Annotate
Colour
Hilight
Find
Node
Clade
Taxa
Selection Mode
Prev/Next
Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
A/Osaka-C/1/2009_136_2009.373
A/Hyogo/1/2009_137_2009.375
A/Osaka/1/2009_136_2009.373
A/Korea/01/2009_122_2009.334
A/Michigan/02/2009_116_2009.318
A/Canada-ON/RV1527/2009_114_2009.312
A/Zhejiang/1/2009_143_2009.392
A/New_York/10/2009_115_2009.315
A/New_York/3173/2009_118_2009.323
A/California/14/2009_115_2009.315
A/Shanghai/1/2009_143_2009.392
A/New_York/3261/2009_119_2009.326
A/New_York/3172/2009_117_2009.321
A/New_York/3232/2009_120_2009.329
A/England/207/2009_120_2009.329
A/Guangdong/2/2009_147_2009.403
A/Fujian/1/2009_142_2009.389
A/New_York/3168/2009_116_2009.318
A/New_York/3260/2009_117_2009.321
A/New_York/3214/2009_115_2009.315
A/Moscow/01/2009_141_2009.386
A/New_York/23/2009_114_2009.312
A/New_York/3190/2009_119_2009.326
A/New_York/3193/2009_118_2009.323
A/New_York/3187/2009_121_2009.332
A/New_York/3191/2009_115_2009.315
A/Mexico/4604/2009_109_2009.299
A/Shandong/1/2009_130_2009.356
A/Beijing/3/2009_140_2009.384
A/Nonthaburi/102/2009_127_2009.348
A/Mexico/InDRE4487/2009_104_2009.285
A/Narita/1/2009_128_2009.351
A/Beijing/01/2009_135_2009.37
A/Norway/1168/2009_127_2009.348
A/Paris/2580/2009_120_2009.329
A/Sichuan/1/2009_130_2009.356
A/Paris/2590/2009_120_2009.329
A/Texas/15/2009_105_2009.288
A/Texas/09/2009_115_2009.315
A/Texas/04/2009_104_2009.285
A/Mexico/4486/2009_104_2009.285
A/Hong_Kong/01/2009_120_2009.329
A/Canada-NS/RV1535/2009_114_2009.312
A/Guangdong/1/2009_138_2009.378
A/New_York/3099/2009_119_2009.326
A/New_York/1682/2009_117_2009.321
A/Netherlands/602/2009_119_2009.326
A/Mexico/47N/2009_115_2009.315
A/California/06/2009_106_2009.29
A/Indiana/09/2009_112_2009.307
2009.1
2009.2
2009.3
2009.4
2009.5

☐ Align Tip Labels

▶ Current Tree: 1 / 1
▶ Appearance
▶ Trees
▶ Time Scale
▶ ☑ Tip Labels
▶ ☐ Node Labels
▶ ☑ Node Bars
▶ ☐ Node Shapes
▶ ☐ Branch Labels
▶ ☐ Scale Bar
▼ ☑ Scale Axis

☑ Show grid
Line Weight: 2
☑ Reverse axis
☑ Automatic scale
Label spacing: 0.1
Tick spacing: 0.05
Origin value: 0.0
Font Size: 11

▶ ☐ Legend

Tip: "A/Hyogo/1/2009_137_2009.375" [height = 0.0299, length = 0.0291]

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. 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 2009.403, 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. 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.).

A quick how-to summary for Exercise 1

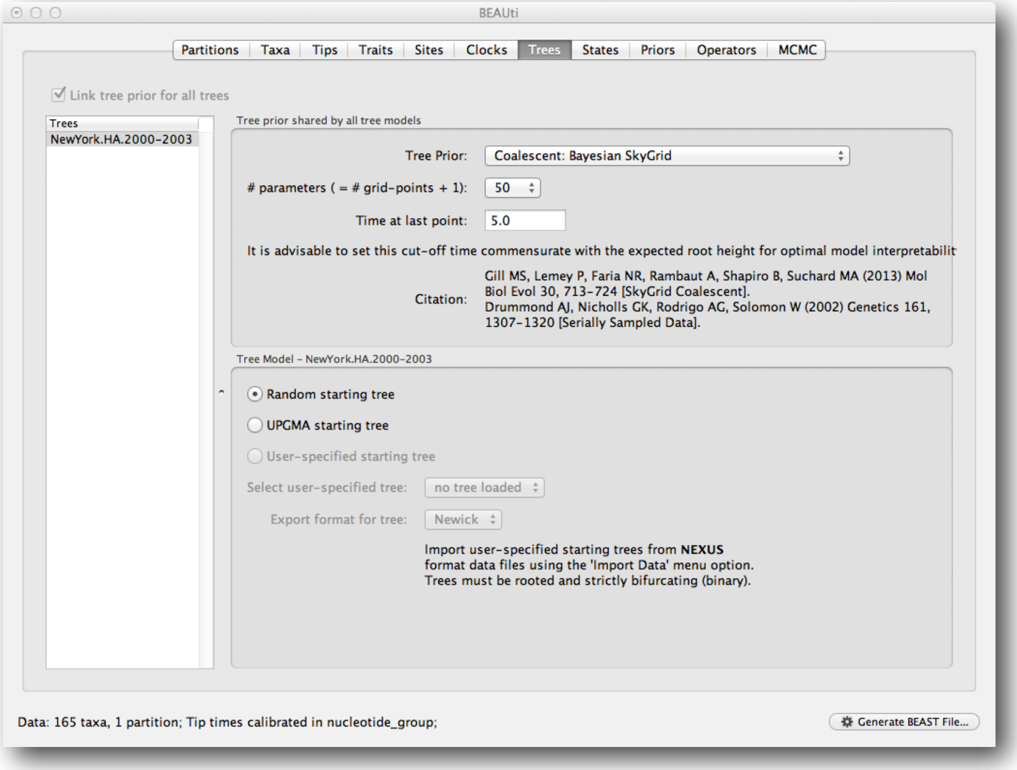
- Run **BEAUti**.
  - Load a NEXUS format alignment by selecting the Import Data... option from the File menu. Select the file called H1N109.nex.
  - In the Tips tab, select the box labelled Use tip dates and click the Guess Dates button. Keep the default Defined just by its order and select last from the drop-down menu for the order and press OK.
  - In the Sites tab, keep the default HKY substitution model for the nucleotide data and base frequencies as Estimated. Select 'Gamma' as 'Site Heterogeneity Model' (with 4 discrete categories) before proceeding to the 'Clocks' tab.
  - In the Clocks tab, keep a strict clock model.
  - In the Trees tab, set the option for Tree Prior to Coalescent: Exponential Growth..
  - In the MCMC tab, set the chain length to 1,00,000 and both the sampling frequencies to 100. Set the File name stem to H1N109 and generate the beast file (H1N109.xml).
- Run **BEAST** and load the xml file.
- Analyze the output using **Tracer**. Analyze the output file for the longer runs.
- Calculate the growth rate for pandemic influenza H1N1 (see page 12 of this tutorial).
- Summarize the trees of the longer run using **treeAnnotator** (burn-in = 500,000 states or 100 trees).
- Visualize the tree in **FigTree**.

EXERCISE 2: reconstructing H3N2 epidemic dynamics in the New York state.

In this exercise, we will reconstruct a *Bayesian skygrid* of H3N2 spread during three epidemic seasons. The data set, NewYork.HA.2000-2003.nex, contains 165 Hemagglutinin genes and takes more time to run in *BEAST* than available during a practical session. Therefore, this tutorial will discuss how to set up this analysis and how to summarize the results based on runs that have already been performed.

Running BEAUti

Run *BEAUti*, load the nexus file (NewYork.HA.2000-2003.nex) and set the dates to the last numerical field in the taxa names as previously. Set the same evolutionary model (including gamma distributed rate variation) and clock model as in the previous exercise. In the 'Trees' tab, select a 'Coalescent: Bayesian SkyGrid' as the 'Tree Prior'. We will construct a grid of 50 intervals over 5 years (Time at point) prior to the most recent sampling dates (2003.98 in our case, so going back to about 1999) requiring us to estimate 10 population sizes by year.


BEAUti
Partitions | Taxa | Tips | Traits | Sites | Clocks | Trees | States | Priors | Operators | MCMC
Generate BEAST File...

☑ Link tree prior for all trees

Trees
NewYork.HA.2000-2003

Tree prior shared by all tree models

Tree Prior: Coalescent: Bayesian SkyGrid

parameters (= # grid-points + 1): 50

Time at last point: 5.0

It is advisable to set this cut-off time commensurate with the expected root height for optimal model interpretabilit

Citation: Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA (2013) Mol Biol Evol 30, 713-724 [SkyGrid Coalescent].
Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Genetics 161, 1307-1320 [Serially Sampled Data].

Tree Model - NewYork.HA.2000-2003

● Random starting tree

○ UPGMA starting tree

○ User-specified starting tree

Select user-specified tree: no tree loaded

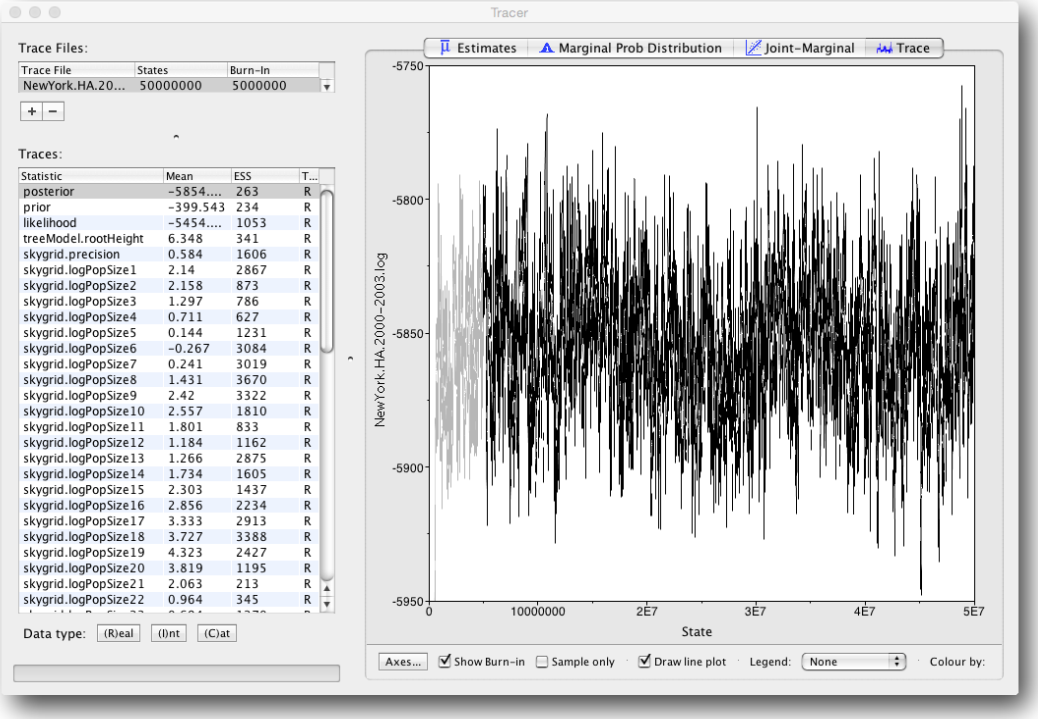
Export format for tree: Newick

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: 165 taxa, 1 partition; Tip times calibrated in nucleotide_group;

Analyzing the BEAST output

Using *Tracer*, we can analyze the run based on the output files provided (load the file called 'NewYork.HA.2000-2003.log'):



Trace Files:

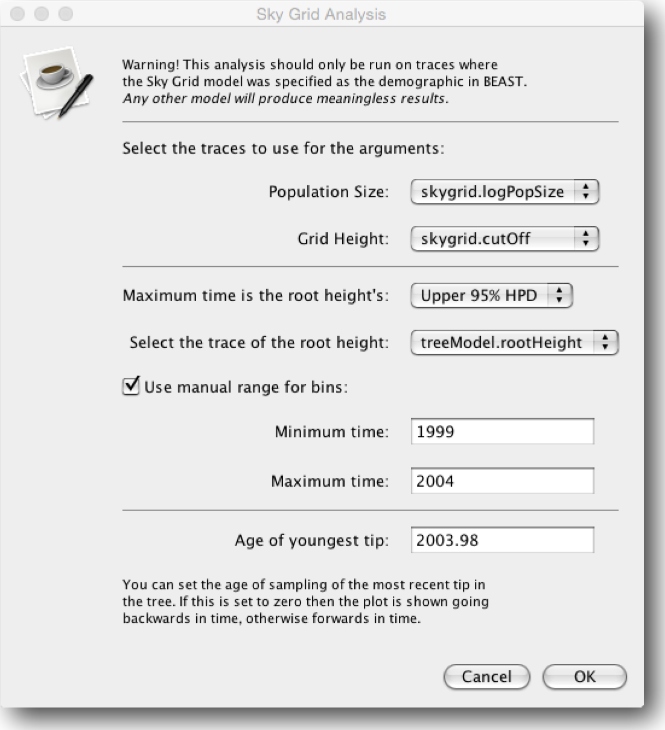
Trace File	States	Burn-In
NewYork.HA.20...	50000000	5000000

Traces:

Statistic	Mean	ESS	T...
posterior	-5854....	263	R
prior	-399.543	234	R
likelihood	-5454....	1053	R
treeModel.rootHeight	6.348	341	R
skygrid.precision	0.584	1606	R
skygrid.logPopSize1	2.14	2867	R
skygrid.logPopSize2	2.158	873	R
skygrid.logPopSize3	1.297	786	R
skygrid.logPopSize4	0.711	627	R
skygrid.logPopSize5	0.144	1231	R
skygrid.logPopSize6	-0.267	3084	R
skygrid.logPopSize7	0.241	3019	R
skygrid.logPopSize8	1.431	3670	R
skygrid.logPopSize9	2.42	3322	R
skygrid.logPopSize10	2.557	1810	R
skygrid.logPopSize11	1.801	833	R
skygrid.logPopSize12	1.184	1162	R
skygrid.logPopSize13	1.266	2875	R
skygrid.logPopSize14	1.734	1605	R
skygrid.logPopSize15	2.303	1437	R
skygrid.logPopSize16	2.856	2234	R
skygrid.logPopSize17	3.333	2913	R
skygrid.logPopSize18	3.727	3388	R
skygrid.logPopSize19	4.323	2427	R
skygrid.logPopSize20	3.819	1195	R
skygrid.logPopSize21	2.063	213	R
skygrid.logPopSize22	0.964	345	R

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

To reconstruct the Bayesian skygrid plot, select ‘SkyGrid reconstruction...’ under the Analysis window. The following window should appear:



Warning! This analysis should only be run on traces where
the Sky Grid model was specified as the demographic in BEAST.
Any other model will produce meaningless results.

Select the traces to use for the arguments:

Population Size: skygrid.logPopSize

Grid Height: skygrid.cutOff

Maximum time is the root height's: Upper 95% HPD

Select the trace of the root height: treeModel.rootHeight

☑ Use manual range for bins:

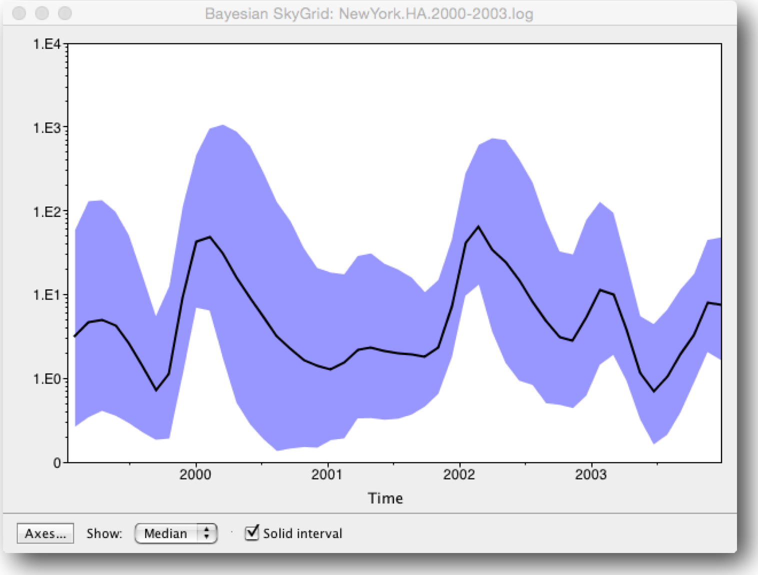
Minimum time: 1999

Maximum time: 2004

Age of youngest tip: 2003.98

You can set the age of sampling of the most recent tip in
the tree. If this is set to zero then the plot is shown going
backwards in time, otherwise forwards in time.

Set the manual bin range from 1999 to 2004 and specify ‘2003.98’ as the ‘Age of the youngest tip’ at the bottom. Press ‘OK’ and after some time, the following Bayesian skyGrid reconstruction should appear (with solid interval selected):



Output files for a Bayesian skyline plot analysis are also provided for comparison. To reconstruct a Bayesian skyline plot based on these, select '**Bayesian Skyline reconstruction**' under the Analysis window.

Some Questions

- What type of dynamics does the H3N2 skyride plot suggest? Would you expect to see the similar dynamics for H3N2 sampled in a southern hemisphere location?

- Is the H1N1/09 evolutionary rate similar to the H3N2 evolutionary rate? If not, what could explain their differences.

- Based on the H1N1/09 tree inferred from a limited sampling, how many H1N1/09 introductions in New York would you conclude for this sample?

A quick how-to summary

- Run **BEAUti**.
  - Load a NEXUS format alignment by selecting the **Import Data...** option from the **File** menu. Select the file called **NewYork.HA.2000-2003.nex**.
  - In the **Tips** tab, select the box labelled **Use tip dates** and click the **Guess Dates** button. Keep the default **Defined just by its order** and select **last** from the drop-down menu for the order and press **OK**.
  - In the **Sites** tab, keep the default **HKY** substitution model for the nucleotide data and base frequencies as **Estimated**. Select '**Gamma**' as '**Site Heterogeneity Model**' (with 4 discrete categories) before proceeding to the '**Clocks**' tab.
  - In the **Clocks** tab, keep a **strict clock** model.
  - In the **Trees** tab, set the option for **Tree Prior** to **Coalescent: Bayesian SkyGrid** with 50 parameters and a 5 year cut-off (**Time at last point**).

- In the **Priors** tab, keep the default prior settings.
- In the **MCMC** tab, set the chain length to 25,000,000 and both the sampling frequencies to 5000. Set the **File name stem** to NewYork.HA.2000-2003 and generate the beast file (NewYork.HA.2000-2003.xml).

- Run **BEAST** and load the xml file.
- Analyze the output using **Tracer**. Analyze the output file for the longer runs.
- Reconstruct a Bayesian skyride plot by selecting '**SkyGrid reconstruction...**' under the **Analysis** window (bin range = 1999-2004 and age of the youngest tip = 2003.98).

References

- Drummond AJ, Rambaut A (2007) *BEAST*: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**: 214.
- Drummond AJ, Ho SYW, Phillips MJ & Rambaut A (2006) *PLoS Biology* 4, e88.
- Drummond AJ, Rambaut A & Shapiro B and Pybus OG (2005) *Mol Biol Evol* **22**, 1185-1192.
- Drummond AJ, Nicholls GK, Rodrigo AG & Solomon W (2002) *Genetics* **161**, 1307-1320.
- Ferreira, M. A. R. and M. A. Suchard. 2008. Bayesian analysis of elapsed times in continuous-time Markov chains. Can J Statistics, 36: 355–368. doi: 10.1002/cjs.5550360302
- Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, and Suchard MA (2013) Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol* **30**, 713-724.
- Minin VN, Bloomquist EW and Suchard MA (2008) Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution* **25**:1459-1471; doi:10.1093/molbev/msn090.
- Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC (2008) The genomic and epidemiological dynamics of human influenza A virus. *Nature*, **453**: 615-9.
- Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, Peiris JSM, Guan Y & Rambaut A (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122-1125.

Help and documentation

- The BEAST software download: http://beast-mcmc.googlecode.com/
- 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/
- H1N1/09: http://tree.bio.ed.ac.uk/wiki/projects/influenza/