Revealing the evolutionary dynamics of influenza

Summary: This tutorial provides a step-by-step explanation on how to reconstruct the evolutionary dynamics of influenza based on a set of virus sequences w
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 pandemic influenza A (H1N1) virus in 2009 (H1N1pdm) and on the seasonal dynamics of H3N2 in the New York State. The H1N1pdm data set is a
subset of an analyzed set genomes in a study that provides insights into the origins and evolutionary genomics of this pandemic (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 date of the origin of the epidemic and a
estimate of the H1N1pdm epidemic growth and basic reproductive number. In the second exercise, we will examine how H3N2 diversity fluctuates through time

Table of Contents

Introduction

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 us
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 exp
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 syster

BEAST - this package contains the BEAST (beast) program, BEAUti (beauti) and a couple of utility programs. At the time of writing, the current version is v1.10.4. BEAST releases are available for downloa
from https://github.com/beast-dev/beast-mcmc/releases ⬈.
(beast)

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 v1.7.1. It is available for download from https://github.com/beast-dev/tracer/ ⬈.
(tracer)

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/software/figtree/ ⬈.
(figtree)

ⓘ **Note:** This tutorial builds on the Estimating rates and dates from time-stamped sequences (workshop_rates_and_dates) which should be completed before starting this one.

⬇ All the files needed for this tutorial can be downloaded from here (/tutorials/workshop_influenza_phylodynamics/files/influenzaTutorialFiles.zip). If you download this zipped folder, there is no need to download
other files/folders linked further in the tutorial.

EXERCISE 1: The swine-origin influenza A outbreak

⬇ The data file is called 'H1N1pdm_2009.nex' and can be downloaded from here (/tutorials/workshop_influenza_phylodynamics/files/H1N1pdm_2009.nex).
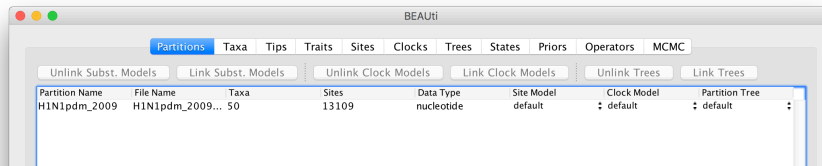
Running BEAUti

Run BEAUti (beauti) by double clicking on its icon. BEAUti is an interactive graphical application for designing your analysis and generating the control file (a BEAST XML file) which BEAST will use to run the
analysis.

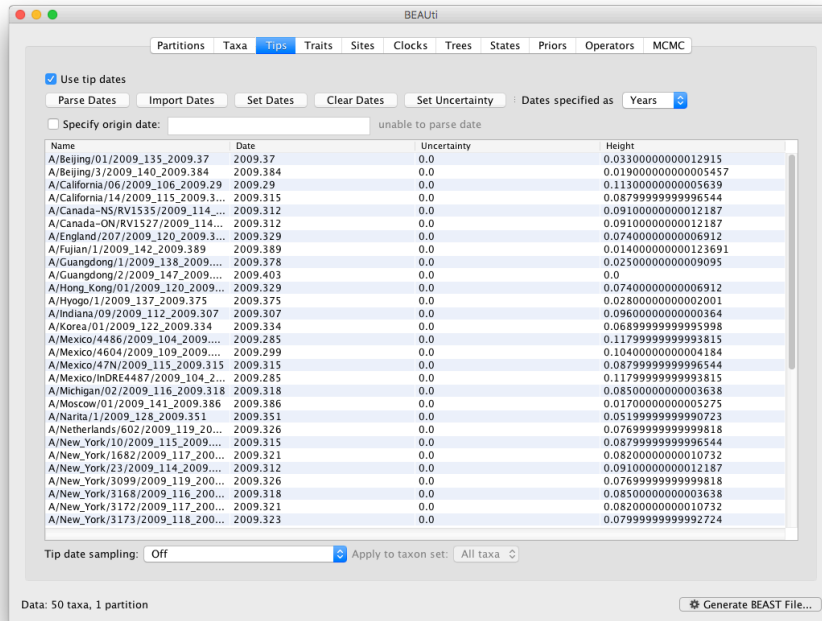To load a NEXUS format alignment, simply select the `Import NEXUS...` option from the `File` menu.

Select the file called `H1N1pdm_2009.nex` . This file contains an alignment of 50 genomes (all 8 genomic segments concatenated), 13109 nucleotides in length. Once loaded, the new data will be listed under Partition
shown in the figure:

Setting the tip dates

To undertake a phylodynamic analysis we need to specify the dates that the individual viruses were collected. In this case, the sequences were sampled from the H1N1 2009 pandemic between March and May 2009. To set these dates switch to the `Tips` panel using the tabs at the top of the window.
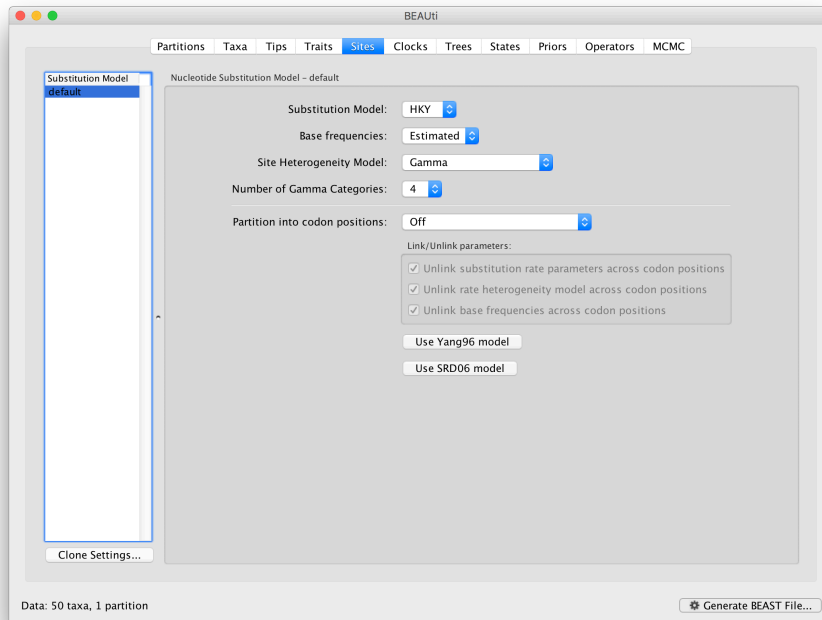
Select the box labelled `Use tip dates`. The actual sampling time in fractional years is encoded in the name of each taxon and we can use the `Parse Dates` button at the top of the panel to extract these. For the H1N1pdm_2009 sequences you can keep the default `Defined just by its order` and select `last` 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.

☑ **Tip:** There are many other options for reading and parsing tip dates in different formats. See this page for a more detailed description of these options. (tip_dates.html)
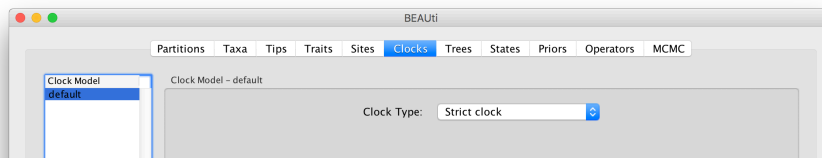
Setting the substitution model

The next thing to do is to click on the `Sites` tab at the top of the main window to specify the evolutionary model settings for BEAST:

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 `Clock` panel 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 pro be over-parameterization. Hence, we keep a strict clock setting.

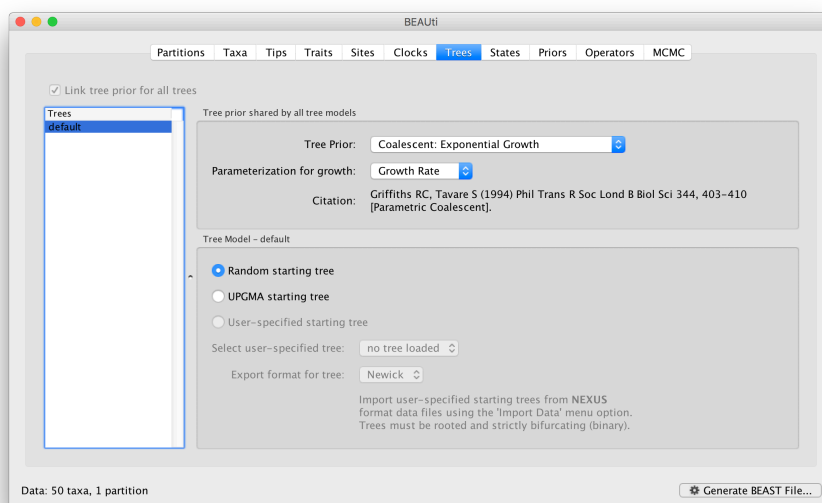Now move on to the `Trees` panel.
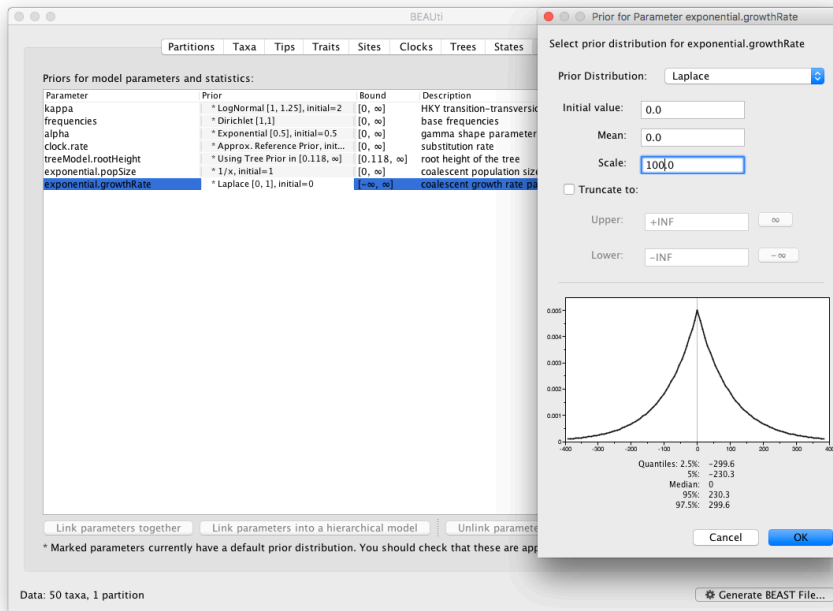
Setting the tree prior

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 change over time for coalescent models. The default tree prior is set to a constant size coalescent prior. The range of different tree priors (coalescent and other models) are described on this page (tree_priors).

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`:

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 strong prior allows the user to 'inform' the analysis by sel a particular distribution with a small variance. Alternatively we can select a weak (diffuse) prior to try to minimise the effect on the analysis. Note that a prior distribution must be specified for every parameter and whils BEAUti provides default options these are not necessarily tailored to the problem and data being analyzed.
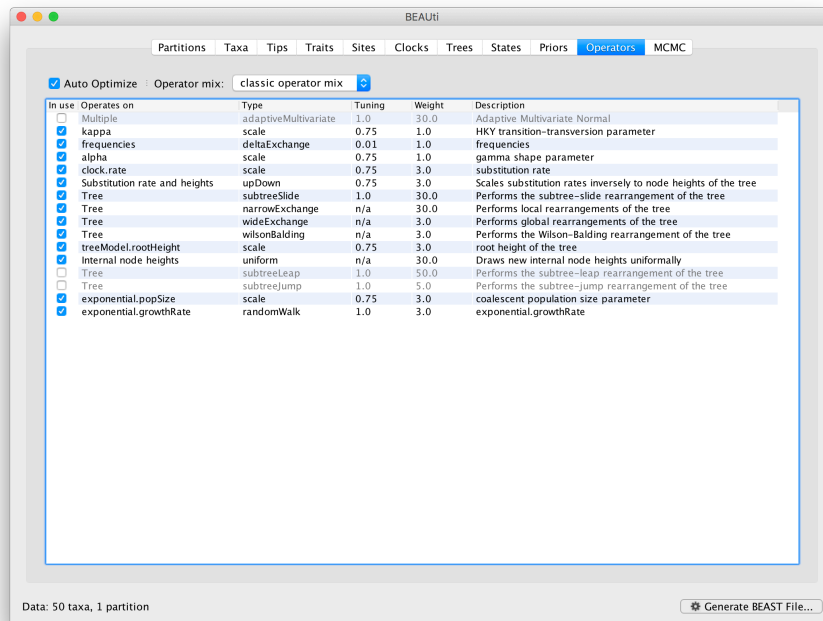
In this case, the default prior for the exponential growth rate (the Laplace distribution) prefers relatively small growth rates because the of the default scale (`1.0`). However, on this epidemic scale, the growth rate para take could take on relatively large values. Therefore, we will increase the variance of this prior distribution by setting the scale to `100`. A useful exercise could be to examine the sensitivity of the growth rate estimates to different scale values for this prior distribution (e.g. scale = 1, 10, 100).

The other priors can be left at their default options.
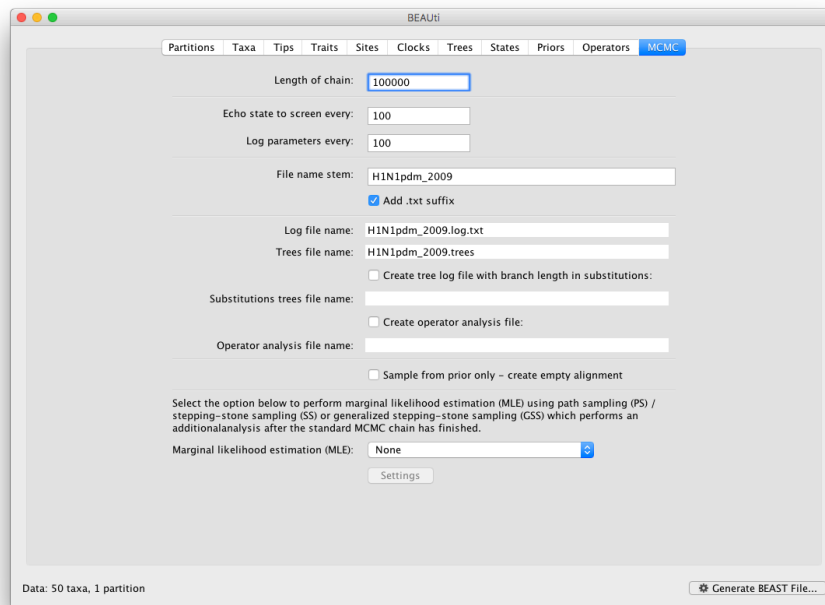
Setting up the operators

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 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:

Notice that the coalescent growth rate parameter (`exponential.growthrate`) has a `randomWalk` operator. This is appropriate for a parameter that can take both positive and negative values (parameters that ar strictly positive can use a scale operator). No changes are required in this table.

Setting the MCMC options

The `MCMC` tab in BEAUti provides settings to control the MCMC chain and the log files that get produced.

For this dataset let's initially set the chain length to `100,000` and both the sampling frequencies to `100`. The `File name stem:` should already be set to `H1N1pdm_2009` but you can adjust this (perhaps add more indications about the analysis).

We are now ready to create the BEAST XML file. Select `Generate XML...` from the `File` menu (or the button at the bottom of the window). BEAUti will ask you to review the prior settings one more time before s the file (and will indicate if any are improper). Continue and choose a name for the file — it will offer the name you gave it in the MCMC panel and we usually end the filename with '.xml' (although on Windows machin may want to give the file the extension '.xml.txt').
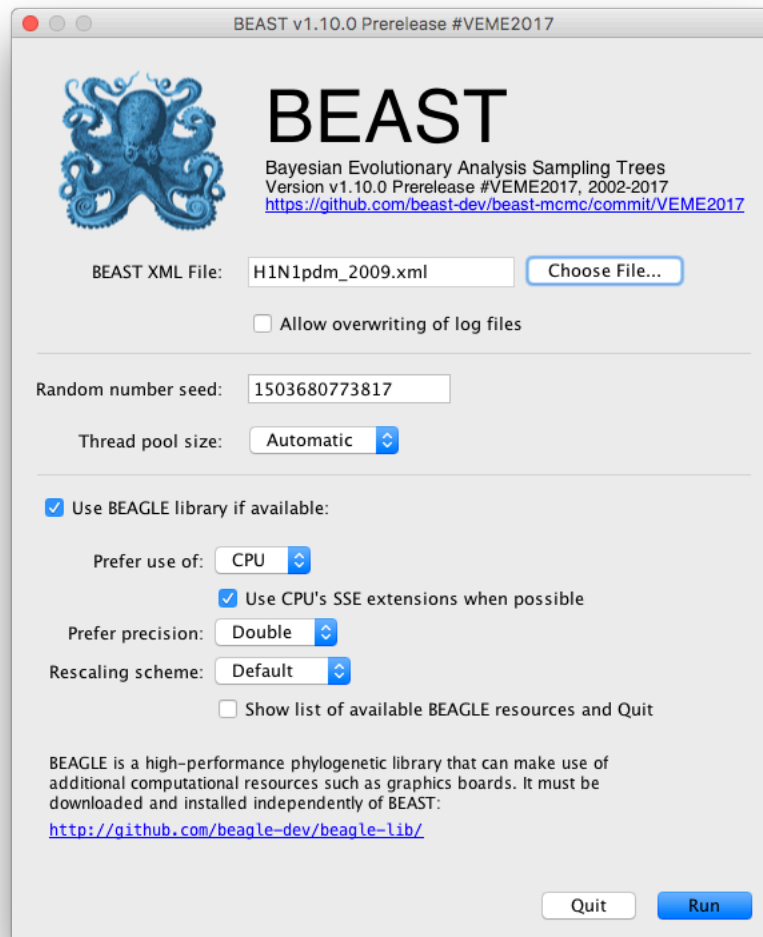
☑ **Tip:** For convenience, leave the BEAUti window open so that you can change the values and re-generate the BEAST file as required later in this tutorial.

Running BEAST

Once the BEAST XML file has been created the analysis itself can be performed using BEAST.

Run BEAST (beast) by double-clicking on the BEAST icon.

Once BEAST has started a dialog box will appear in which you select the XML file:

Press the `Choose File...` button and select the XML file you just created and press `Run`. The analysis will then be performed with detailed information about the progress of the run being written to the screen. V has finished, the log file and the trees file will have been created in the same location as your XML file.
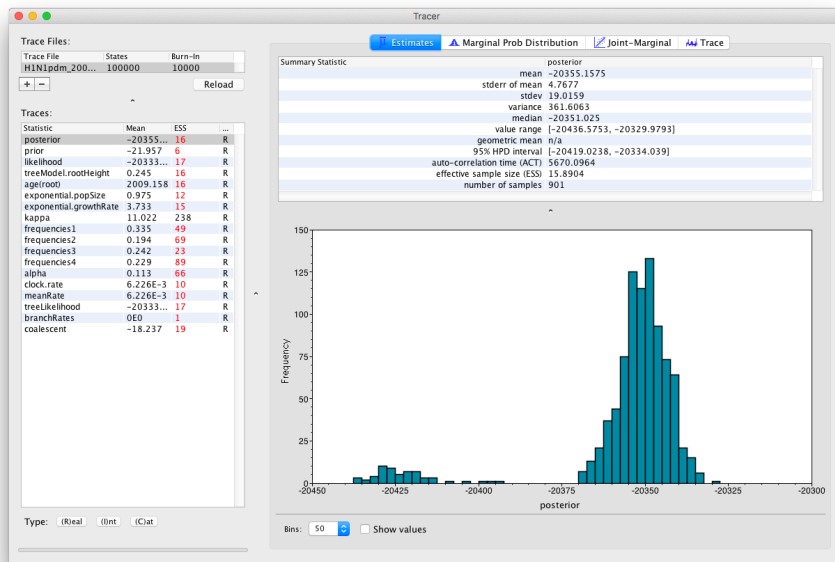
For more information about the other options in the BEAST dialog box see this page (beast).
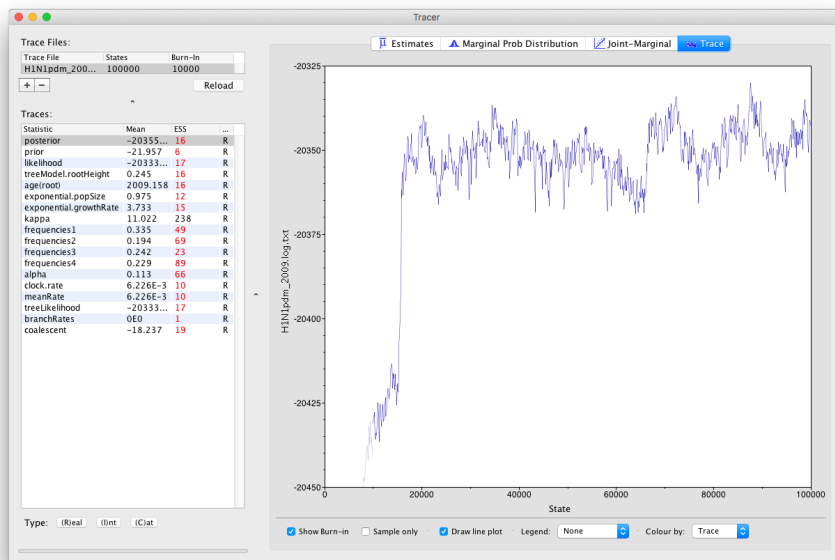
Analyzing the BEAST output

To analyze the results of running BEAST we are going to use the program Tracer (tracer). Run Tracer by double clicking on the Tracer icon.

Select the `Import Trace File...` option from the `File` menu. Select the log file, `H1N1pdm_2009.log`, that you created in the previous section. The file will load and you will be presented with a window simila the one below.

Similarly to the previous tutorial (workshop_rates_and_dates) the effective sample sizes (ESSs) for all the traces are small (ESSs less 100 are highlighted in red respectively by Tracer). In the bottom right of the window frequency plot of the samples, which for the `posterior` trace in the above figure, has multiple peaks.

If we select the tab on the right-hand-side labelled `Trace` we can view the raw trace — the sampled values against the step in the MCMC chain:
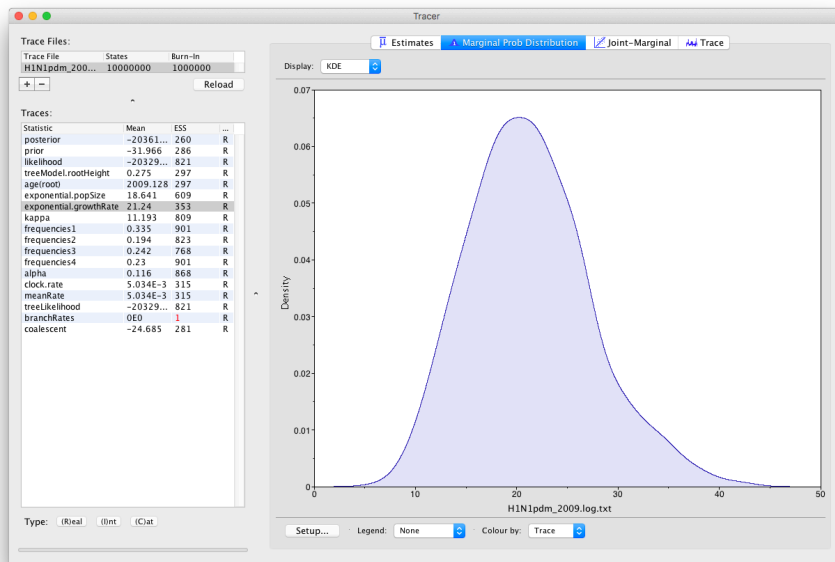

Here it is clear default burn-in of 10% of the chain length is inadequate (the posterior values are still increasing over the first part of the chain). Double-click on the `Burn-In` column in the top left and edit (in the case above, a minimum of `20,000` is needed). However, it is still clear that a chain length of `100,000` was in adequate. Looking at the ESS values (generally in the low double-digits) suggests that a chain length of `10,000,000` would be more appropriate. On a modern computer this would probably only take about 20 minutes but we have provided the output of a run of this length which you can use for the rest of this section.

⬇ The log and trees files for a longer H1N1 exponential growth rate run can be downloaded from here (/tutorials/workshop_influenza_phylodynamics/files/Long_Run_H1N1pdm_Exponential.zip).

Load the new log file (`H1N1pdm_2009.log`) 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 > 300 for the coalescent model parameters there is little auto-correlation between the samples. There are no obvious trends in the plot would suggest that the MCMC has not yet converged, and there are no large-scale fluctuations in the trace which would suggest poor mixing.
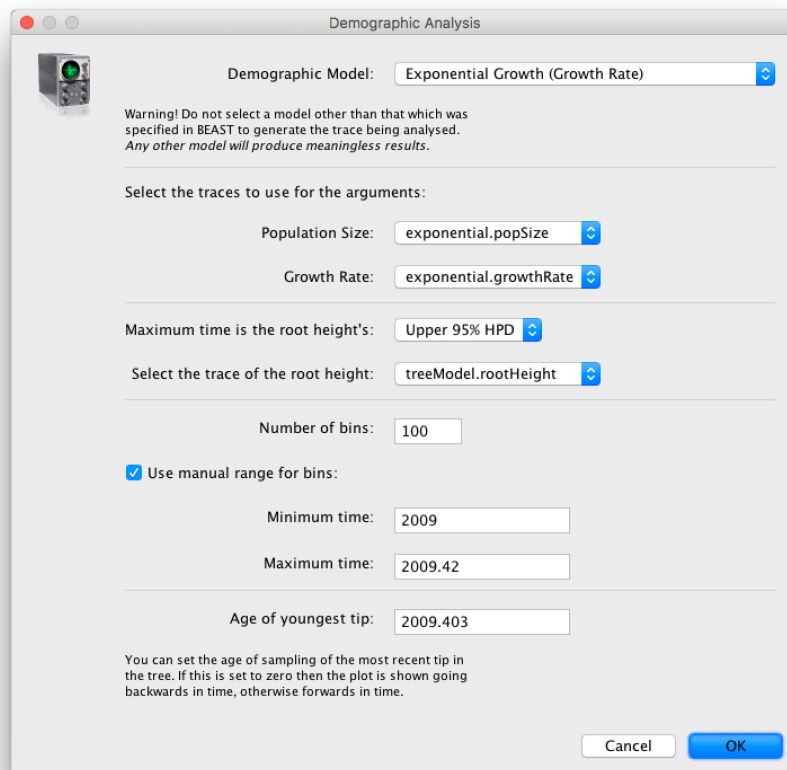
As we are satisfied with the behavior of the MCMC we can now move on to one of the parameters of interest: exponential growth rate for the coalescent model we chose as the tree prior. Select `exponential.growthRate` in the left-hand table. Now choose the density plot by selecting the tab labeled `Marginal Prob Distribution` . This shows a plot of the posterior probability density of this parameter should see a plot similar to this:
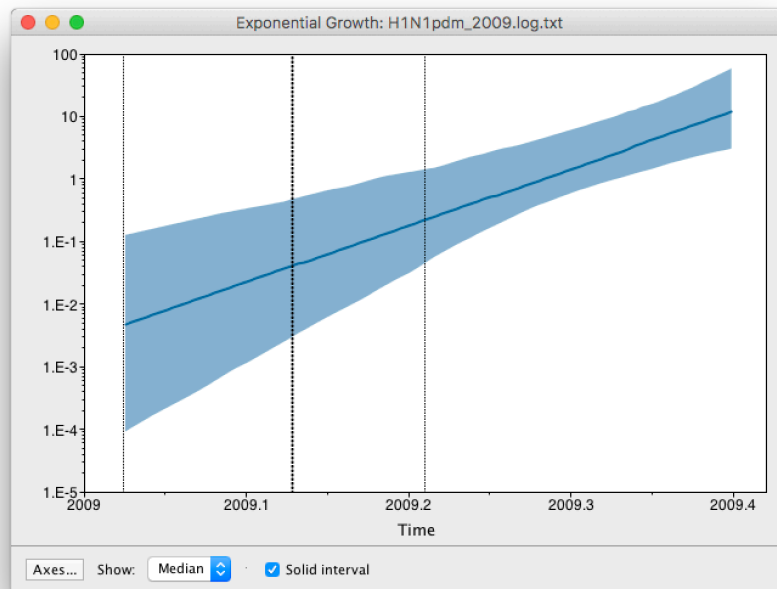
As you can see the posterior probability density is roughly bell-shaped. The default is to show the kernel density estimate (KDE) which is smoothed probability density fitted to the data. Switch the `Display:` option at the top to `Histogram` to see the unsmoothed frequency plot. There is still a lot of noise here but it is a good estimate of the distribution.

> The `age(root)` statistic provides an estimate of the time of the most recent common ancestor of the entire tree. In this case it may be a reasonable estimate of the start of the epidemic, when the virus jumped from pigs into humans. What is the mean estimate and 95% HPDs for the date of the MRCA?

You can visualize the growth estimate using the `Demographic Reconstruction...` option in the `Analysis` menu. Select this option and set up the dialog box that appears like this:

Select `Demographic Model: Exponential Growth (Growth Rate)` — note, you must select the tree prior you picked in BEAUti, you can't change this here. Tracer will automatically identify the parameters of the model (`exponential.popSize` and `exponential.growthRate`). The option `Maximum time is the root height's:` pick the `Upper 95% HPD`. This means it will extend the reconstruction back to the extreme of the root age credible interval. Set the `Age of youngest tip:` to `2009.403` (the date of the most recently sampled virus). You can also `Use manual range for bins:` to make the time-scale a bit cleaner — e.g. `2009.0` to `2009.42`. Then press `OK` and this window will appear:

This shows the exponential growth line for the median growth rate and the 95% HPD intervals for this growth as a solid area. It is on a log scale so is a straight line. You can play with the axis settings using the `Setu` button. The dotted vertical lines represent the 95% HPD for the date of the root of the tree.

The exponential.growthRate (r) provides an estimate of the epidemic growth of H1N1pdm 2009. Given that $Nt = N_0 e^{-rt}$ (with N_0 being the population size at present), the doubling time for $r = 21$ is about 0. years or 12 days. Interestingly, it has been shown that the basic reproductive ratio (R_0) is related to the growth rate — see this page for details (/estimating_R0.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 $\mu = 3$ days and $\sigma = 2$ days, what would be the mean estimate of R_0 for the H1N1pdm 2009 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. Next we can use the TreeAnnotator (treeannotator) tool tha provided as part of the BEAST package to summarize the information contained within our sampled trees.

Run TreeAnnotator by double clicking on the icon.

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). 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.

Use a `Burnin (as states):` of `1,000,000`. This is 10% of the chain and we confirmed that this was adequate in Tracer, above. Use the defaults for the rest of the options — `Posterior probability limit` `Target tree type: Maximum clade credibility tree`, and `Node heights: Median`.

Use the `Choose File...` button to select an input trees file, `H1N1pdm_2009.trees`.

☑ **Tip:** In most of the BEAST package, if there is a button to select a file, you can also simply drag the file into this area.

Select a name for the output tree file (e.g., `H1N1pdm_2009.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 s be loaded into any tree drawing package that supports this. However, it also contains additional information that can only be displayed using the FigTree (figtree) program.

Viewing the annotated tree

FigTree (figtree) is a user-friendly, graphical program for viewing trees and the associated information provided by BEAST. Double-click on the FigTree icon to run it.
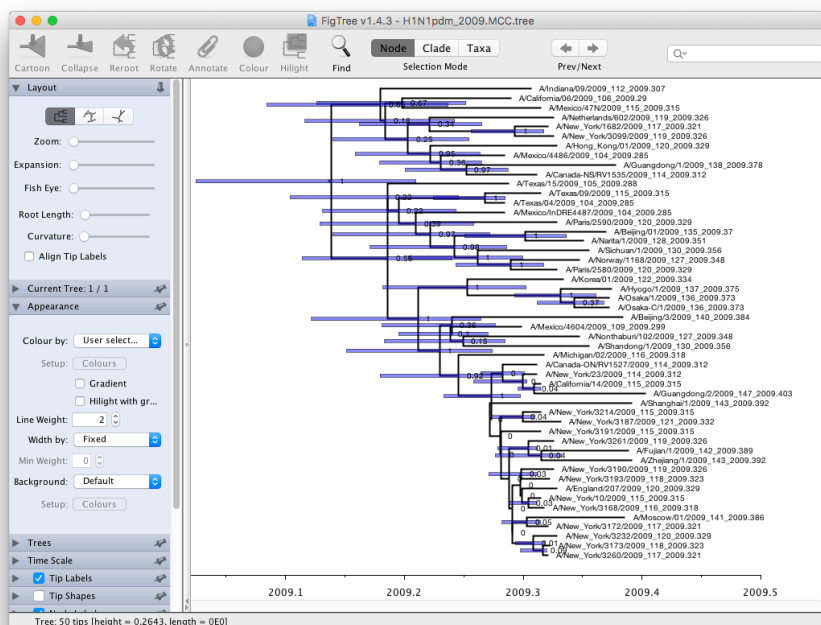
Run FigTree now and select the `Open...` command from the `File` menu. Select the tree file you created using TreeAnnotator in the previous section. The tree will be displayed in the FigTree window. On the left ha side of the window are the options and settings which control how the tree is displayed. 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 e 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. Switch 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. Reduce `Sig. Digits` to 2.

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) credible intervals. Switch on `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. Switch on `Scale Axis` (and switch off `Scale bar`) and select `Reverse Axis` in the `Scale Axis` options (you can also increase the font size a b appropriate scaling, open the `Time Scale` section of the control panel, set the `Offset:` to `2009.403` (the date of our most recently sampled virus).

Finally, open the `Appearance` panel and alter the `Line Weight` to draw the tree with thicker lines. The resulting tree will look like this:

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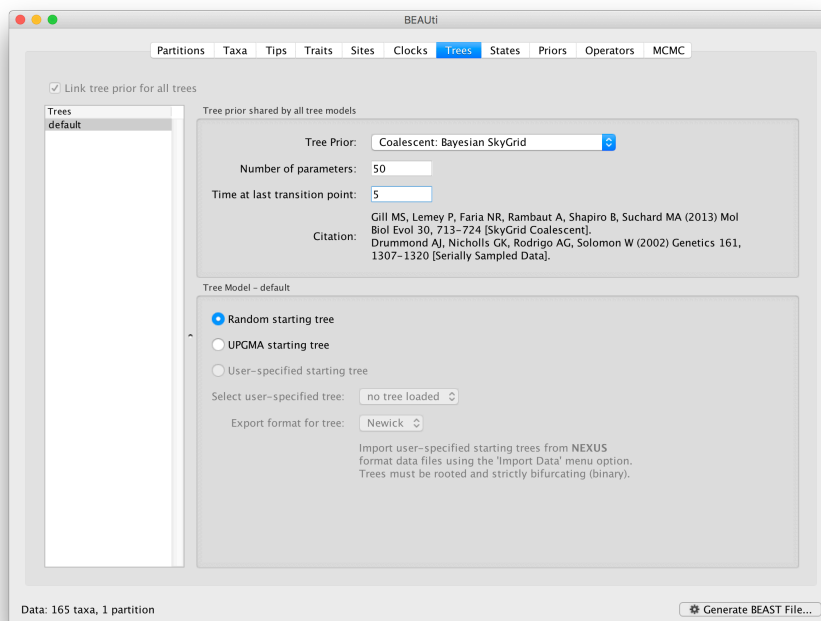
EXERCISE 2: reconstructing H3N2 epidemic dynamics in the New York state.

In this tutorial, we will reconstruct a Bayesian skygrid (tree_priors#skygrid) of human influenza A, H3N2 subtype, over three northern hemisphere epidemic seasons. The data set contains 165 Hemagglutinin gene sequences 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 alre been performed.

> ⬇ The data file is called '`NewYork.HA.2000-2003.nex`' and can be downloaded from here (/tutorials/workshop_influenza_phylodynamics/files/NewYork.HA.2000-2003.nex).

Running BEAUti

Run BEAUti, load the nexus file (`NewYork.HA.2000-2003.nex`) and set the `Tips` panel's `Parse Dates` to the last numerical field in the sequence names as previously. Set the same evolutionary model (includi 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 ye (`Time at last point: 5`) before the most recent sampling date (`2003.98` in our case, so going back to about 1999) thus estimating 10 population sizes per year. The panel should look like this:
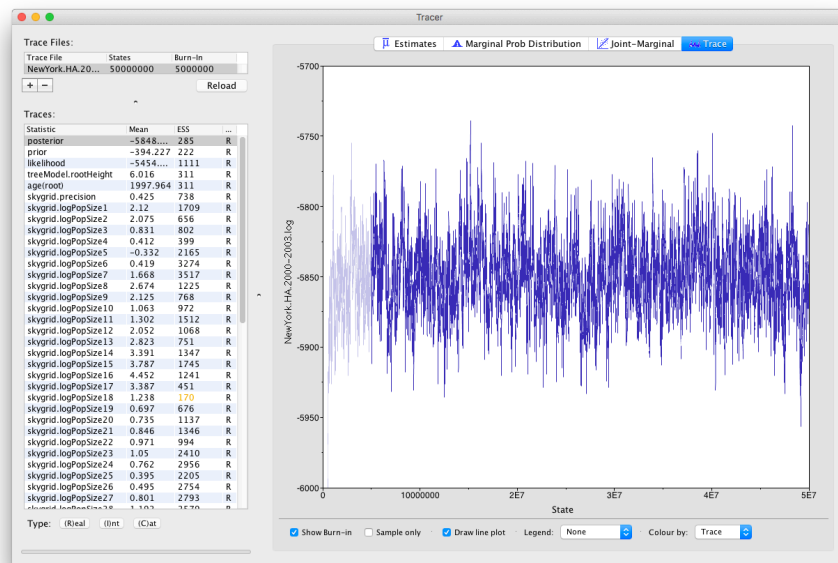
At this point we would usually generate the BEAST XML file, load it into BEAST and run it. However, this data set is a bit bigger than before and the model is a bit more computationally intensive so rather than waiting around for it to finish, you can go straight on and analyse some results files that have already been run.

⬇ The log and trees files for a longer H3N2 SkyGrid run

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`). This has been run with a chain length of 50,000,000 sampling every 5,000 steps so a to 10,000 samples:

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

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 appea solid interval selected):

By default the y-axis is in a log scale. Press the `Axes` button and turn off `Log axis` for the `Y axis`. You will also need to set the `Manual range` because the upper HPD bounds will be very large. Set this to `0` and you get the following:

Here you can see that the seasonal peaks very strong (but that the uncertainty denoted by the credible intervale is also very large).

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?

What happened in 2001?

Is the H1N1pdm 2009 evolutionary rate similar to the seasonal H3N2 evolutionary rate? If not, what could explain their differences?

Based on the H1N1pdm 2009 tree inferred from a limited sampling, how many H1N1pdm introductions in New York would you conclude for this sample?

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 H1N1 influenza A epidemic. Nature 459, 1122-1125.

Help and documentation

The BEAST website: http://beast.community

Tutorials: http://beast.community/tutorials

Frequently asked questions: http://beast.community/faq

Tags: tutorial (tag_tutorial.html) workshop (tag_workshop.html)