**Molecular dating in BEAST**

Sebastian Duchene

**Data set**

* Sequence alignment in fasta format of H1N1 flu data from the 2009 epidemic in North America:

NorthAm.Nov.fasta

**Software**

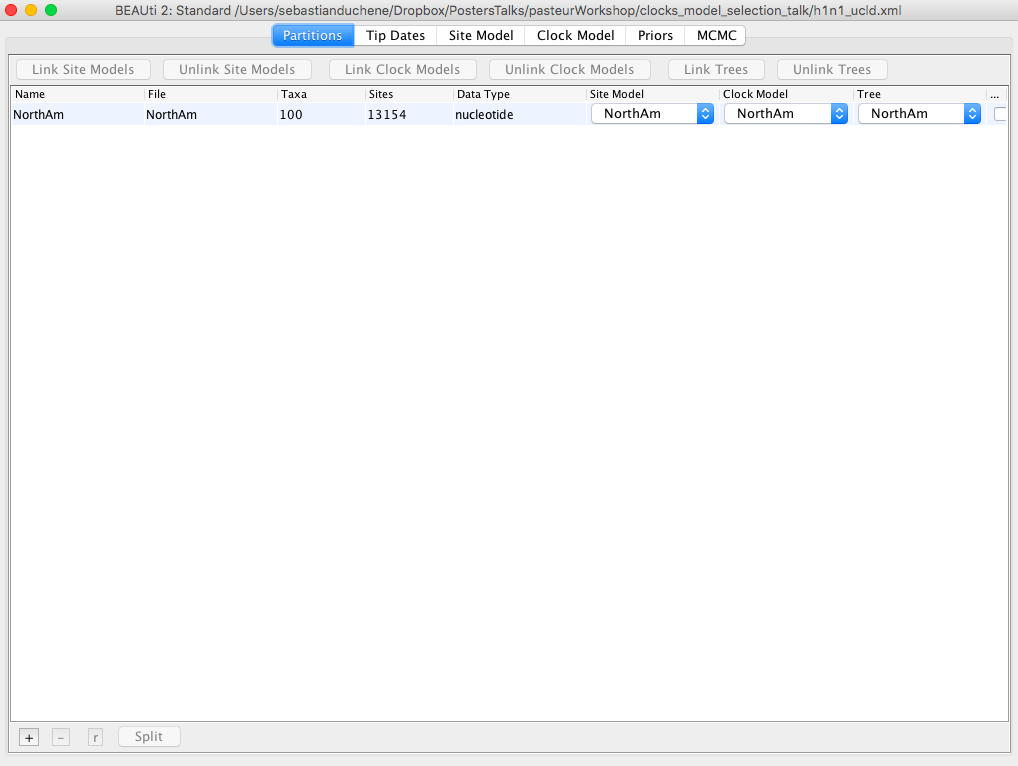
* BEAST 2.4 (beast2.org)
* Package bModelTest. For installation instructions:

<http://www.beast2.org/managing-packages/>

* FigTree or Icytree.org

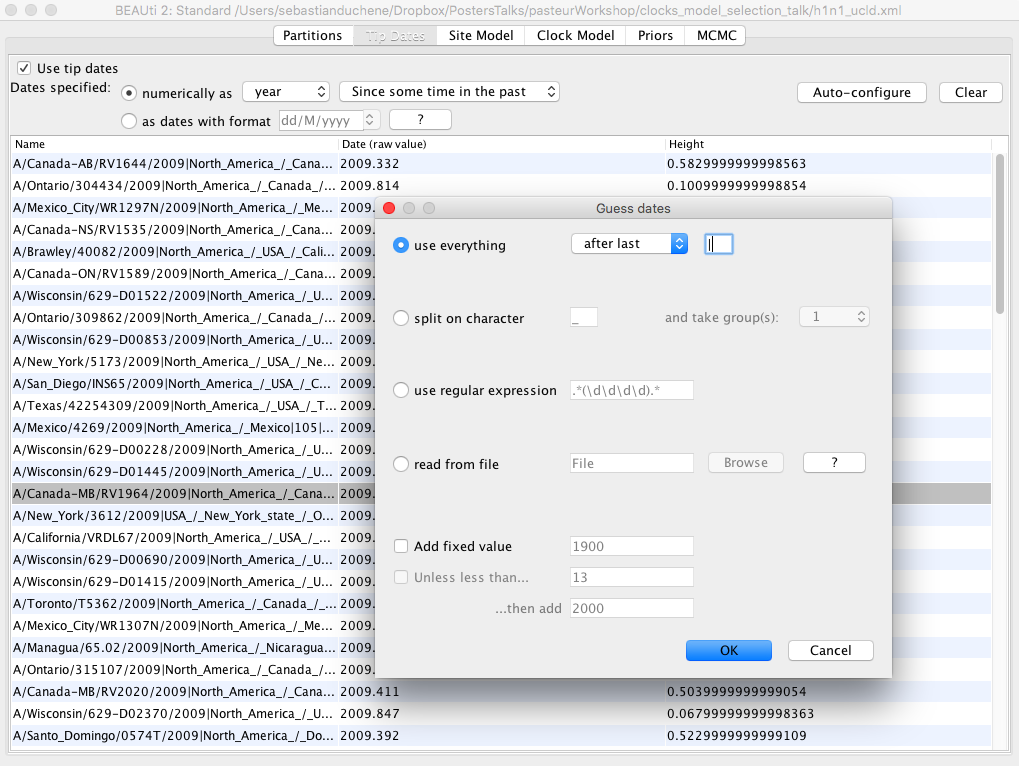
Optional exercises are shown in blue font. They are not required to complete this tutorial, but they will be useful to understand some advanced concepts.

We will be using BEAST2 for the Bayesian part of the workshop. This program requires the data and model specified in an xml format, which can be done using the program BEAUTI. Open BEAUTI and drag the alignment (NorthAm.Nov.fasta) to this window (Fig 1). Note that there are several tabs (Partitions, Tip Dates, Site Model, Clock Model, Priors, and MCMC).



**Fig 1.** BEAUTI with h1n1 sequence alignment loaded.

Click on the *Tip Dates* tab and check the box *Use tip dates* (Fig 2.).

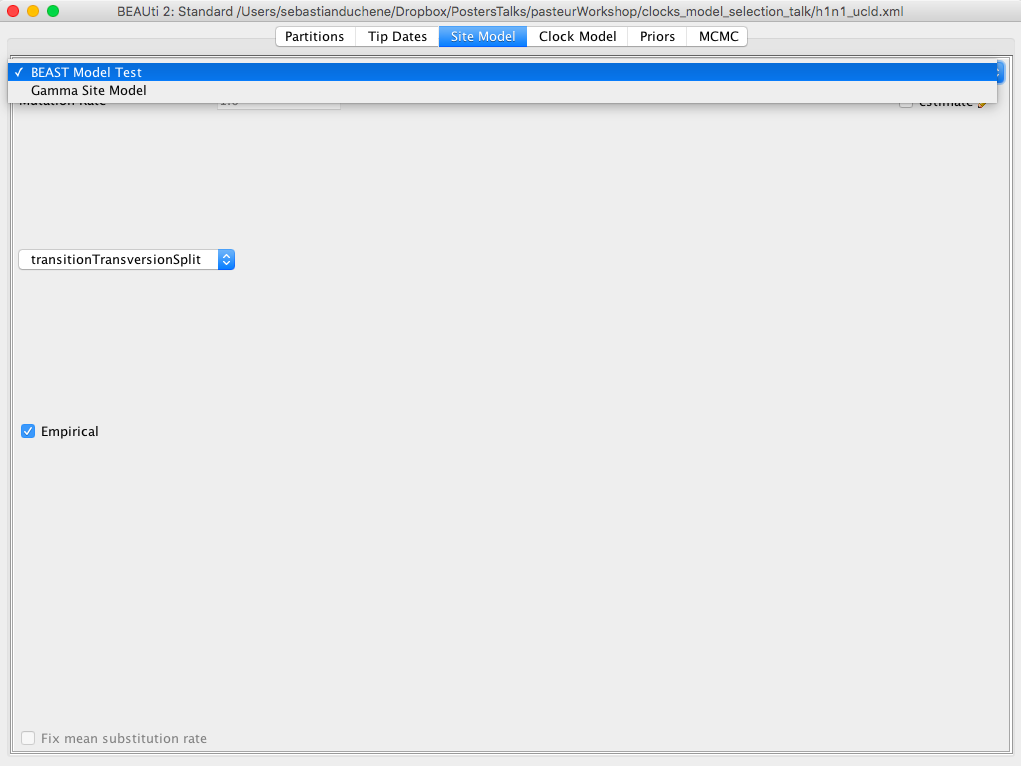


**Fig 2.** Tip dates enabled in BEAUTI.

To use the tip dates as calibrations, click on the *Auto-configure* button. Check the first box (*use everything*) and in the dropdown menu, select *after last*, and type in a vertical line (|) as shown in Fig 2.

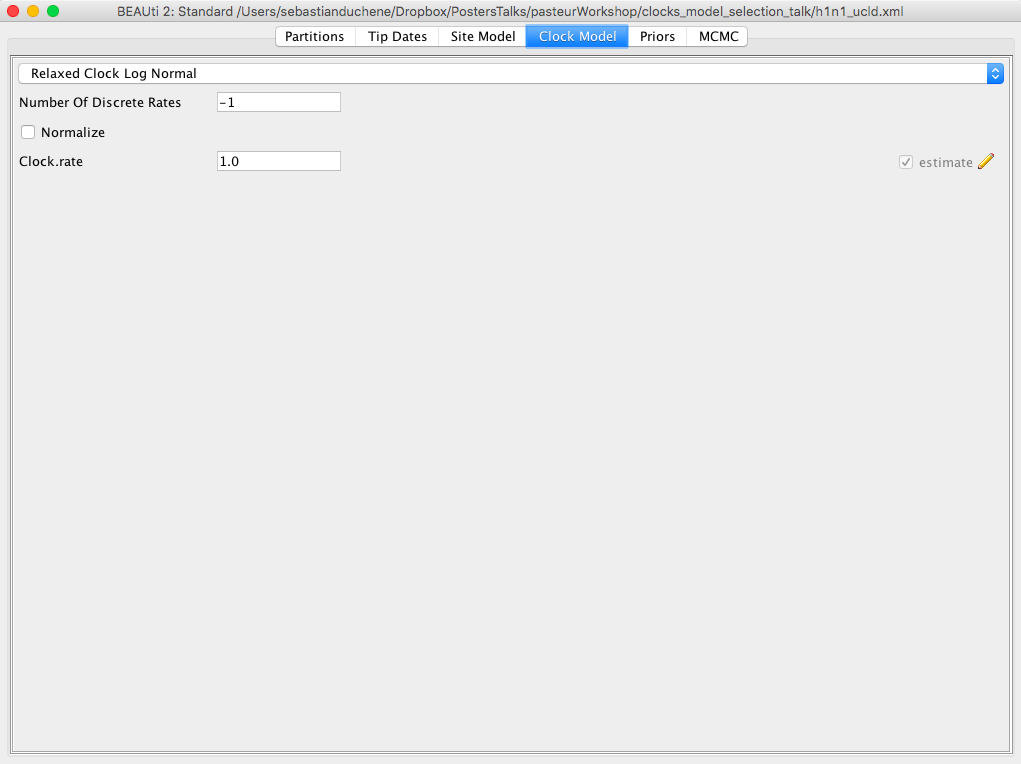
The BEAUTI window should now display the dates for each of the sequences under the column *date*.

Click on the *Site Model* tab. Instead of using a single substitution model, we will average over those that account for differences in the number of transitions to transversions. In the first drop-down menu select *BEAST Model Test*. There is a second drop-down menu to select the range of models that we will sample during the MCMC. Select transitionToTransversionSpit to limit our search to those that allow for differences in transitions to transversions. Click on the box *Empirical* to use the empirical base frequencies. These options should look like those in Fig 5.



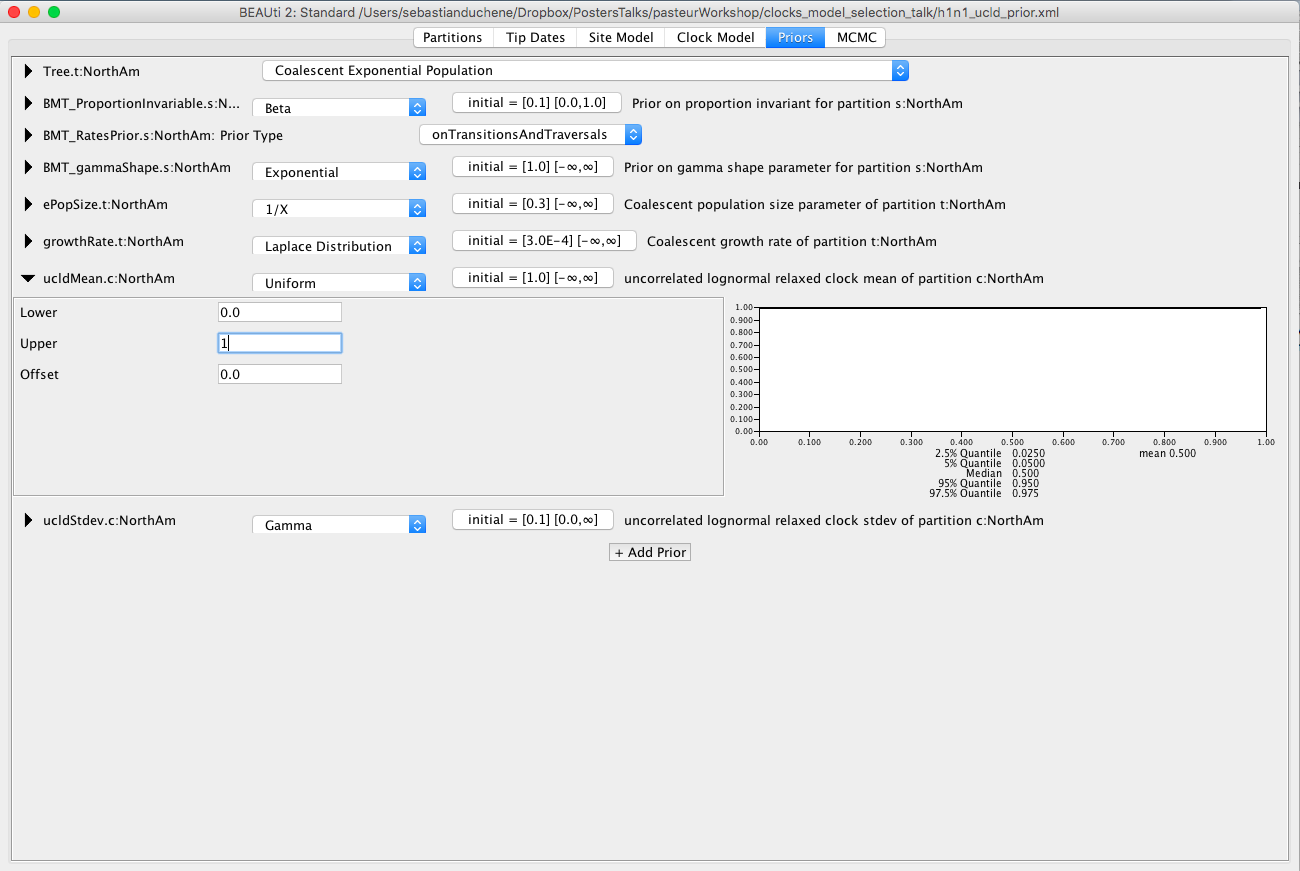
**Fig 5.** Substitution model set up in BEAUTI.

Click on the *Clock Model* tab. In the dropdown menu, select Relaxed Clock Lognormal (Fig 6). The other default options are fine.



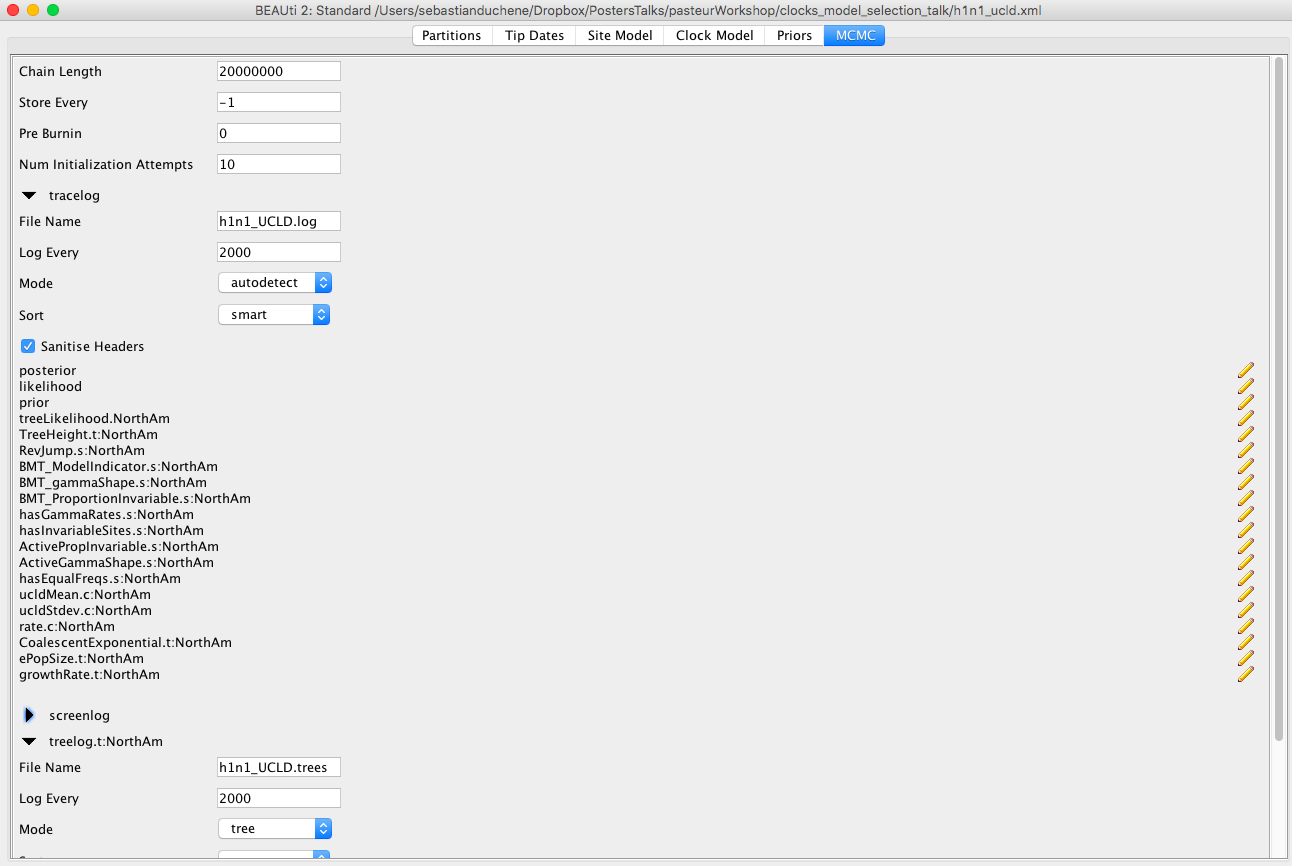
**Fig 6.** Molecular clock model set up in BEAUTI.

Click on the *Priors* tab. Select the Coalescent Exponential Population model (Fig 7). Most of the remaining priors are fine for our analyses, but it is a good exercise to inspect these distributions. In particular, set a Uniform distribution for the mean of the lognormal distribution for the rate with lower and upper bounds of 0 and 1, respectively. This is based on our knowledge that flu probably does not evolve at a rate that is faster than 1 subs/site/year.



**Fig 7.** Priors tab in BEAUTI with the Coalescent Exponential Population prior.

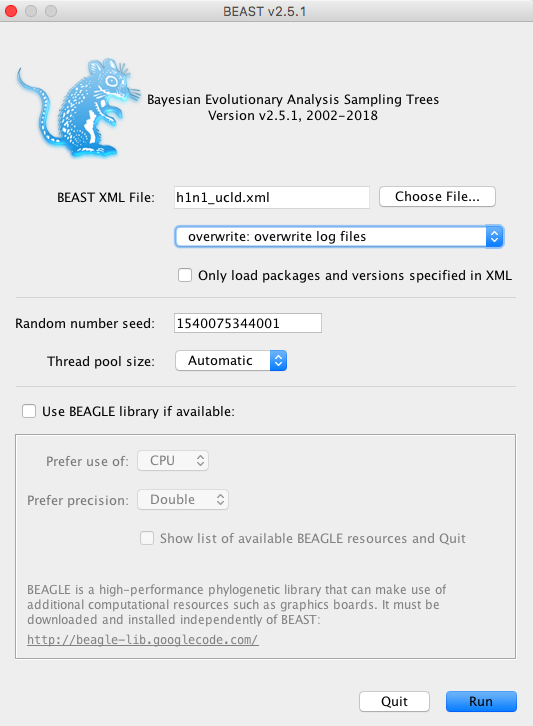
Click on the MCMC tab. Here, we can select different options for the MCMC. The chain length should be 20000000. Click on *tracelog* and *treelog* to specify the BEAST output, which is a set of trees and parameters values, sampled from the posterior. Ensure that the *Log Every* box is 2000 for the trees and log files. Name the log and tree files h1n1\_UCLD.log and h1n1\_UCLD.trees, respectively (Fig 8). We use the extension \_ucld to refer to the uncorrelated lognormal clock model.



**Fig 8.** MCMC set up in BEAST.

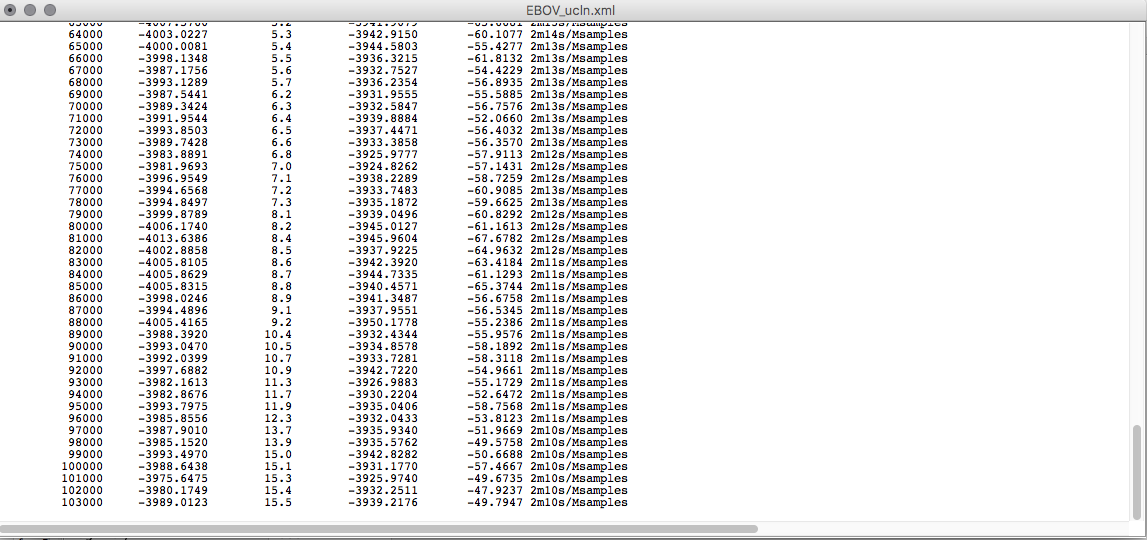
Our BEAST input file is ready. To save it, click on *File*, *Save*, and name it h1n1\_ucld.xml. Do not close BEAUTI.

To run BEAST, double-click on the BEAST2 icon. A window with some options will appear (Fig 9).



**Fig 9.** BEAST starting window.

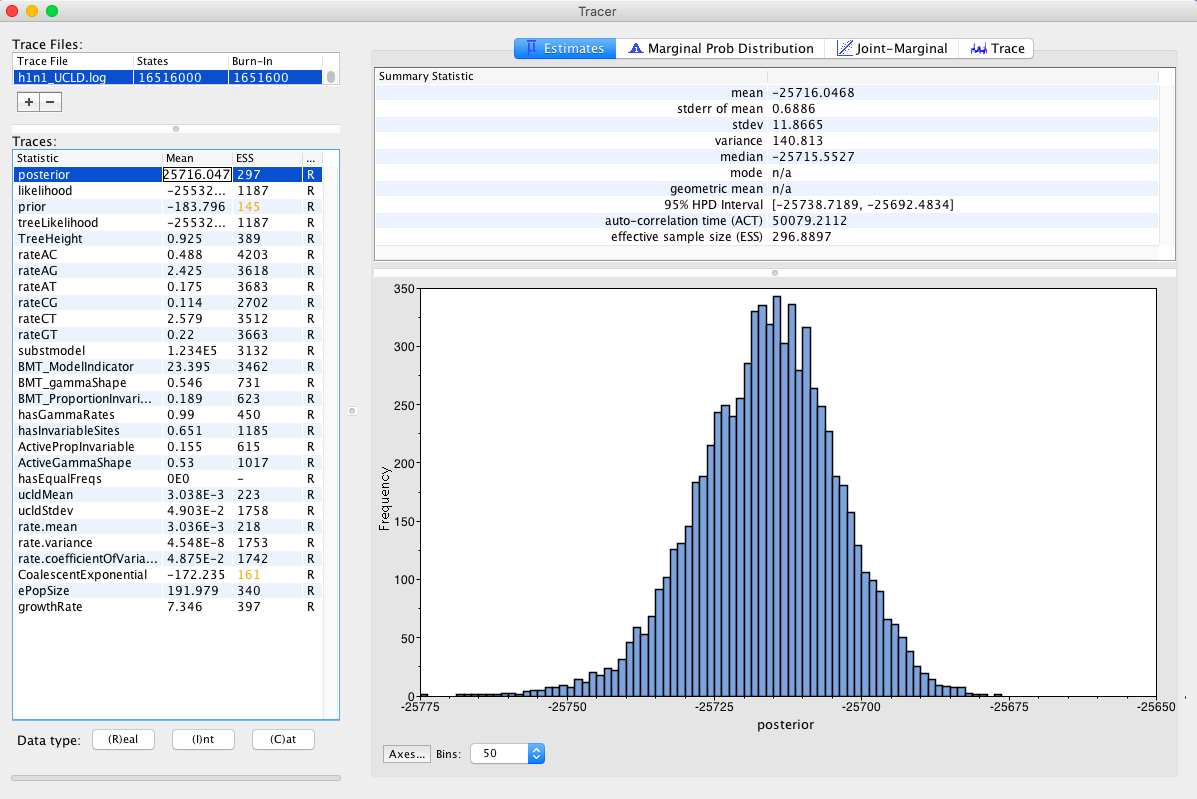
Click on *Choose File...* and select the xml file that we created in BEAUTI. Click Run. The MCMC will start running (Fig 10).



**Fig 10.** BEAST MCMC sampling.

Note that two files have been created in the folder where we saved the xml file, these are the .trees and .log files. This analysis can take up to two hours to complete, but we can inspect the log file much earlier.

After letting BEAST run for about 30 minutes open Tracer(Fig 11), and drag the h1n1\_UCLD.log file to the right pane of the Tracer window.

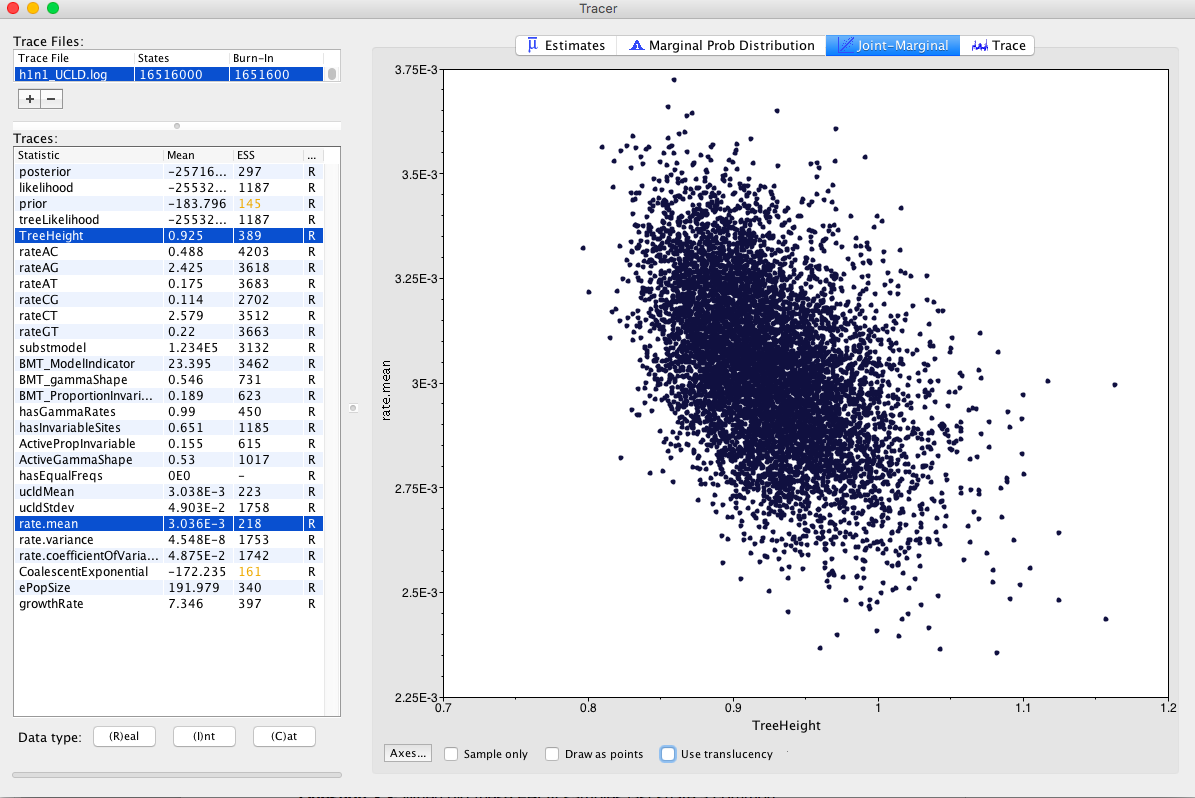
**Fig 11.** Tracer window.

Select  tab. This shows how the MCMC has sampled in parameter space.

**Question:** Inspect the trace for TreeHeight, and the clock model parameters (rate.mean and rate.variance). Does it appear that we have sufficient sampling from the stationary distribution? What do these parameters mean?

An other diagnostic of MCMC sampling is the effective sample size, shown in Tracer as ESS. This is the estimated number of independent samples obtained. A rule of thumb is to ensure that ESS is at least 200 for all parameters.

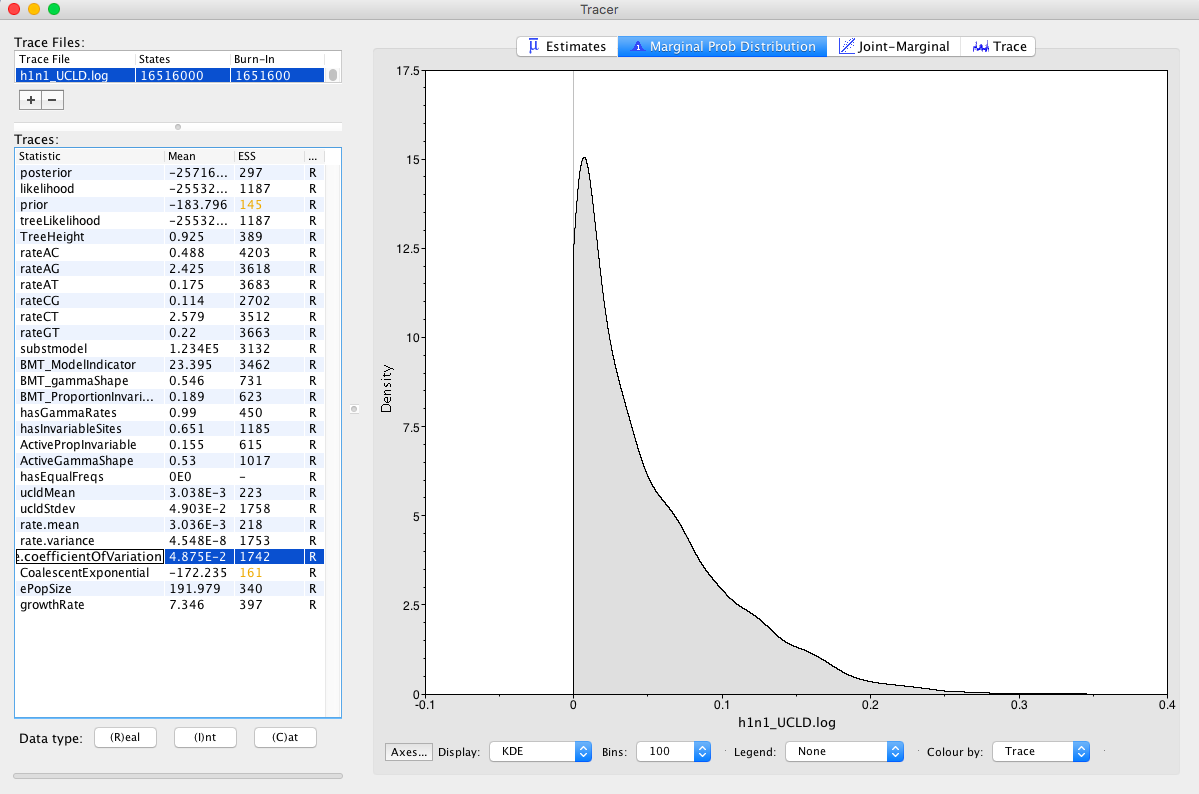
Select the TreeHeight and rate.mean parameters (you might need to use the command or control key to select them at the same time), then select the  tab. TreeHeight is the age of the root of the tree, while rate.mean is the mean substitution rate in the model. If these parameters were independent, we would expect them to form a cloud along the x and y axes. However, these two parameters are naturally correlated. In particular, high rates typically lead to more recent timescales for the root, while lower rates lead to older root ages (Fig 12).



**Fig 12.** Joint marginal plot of TreeHeight and rate.mean.

Check rate.coefficientOfVariation parameter. This is the standard deviation of branch rates divided by the mean rate. Typically, if this parameter is abutting zero, the data have low rate variation, such that they can follow a strict clock.

**Question:** Does this data set appear to follow a strict clock, or does it display substantial rate variation among lineages?

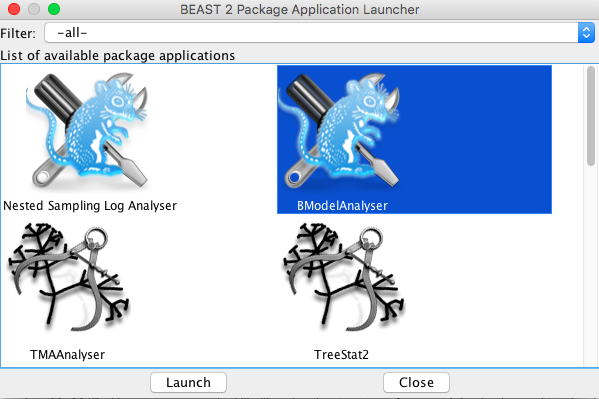


**Fig 13.** Posterior density of coefficient of rate variation.

**Question:** When did these h1n1 samples last share a common ancestor? Is it consistent with our knowledge of the 2009 pandemic?

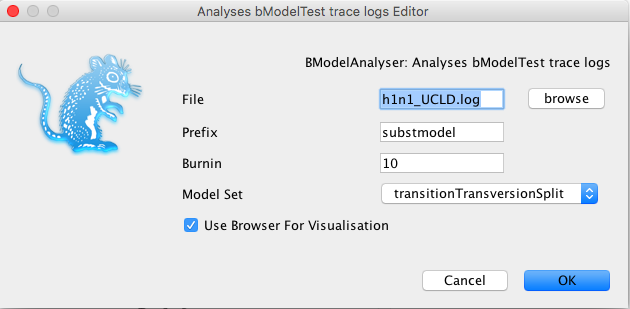
Select the BMT\_ModelIndicator trace. These are the models that were sampled using an index as shown in **Table 1**.

**Optional exercise:** The BEAST2 app BModelAnalyser provides a visual inspection of substitution model averaging. To use it, open the BEAST2 folder and find the AppLauncher icon (). Click it to obtain a list of BEAST2 apps (Fig S1).



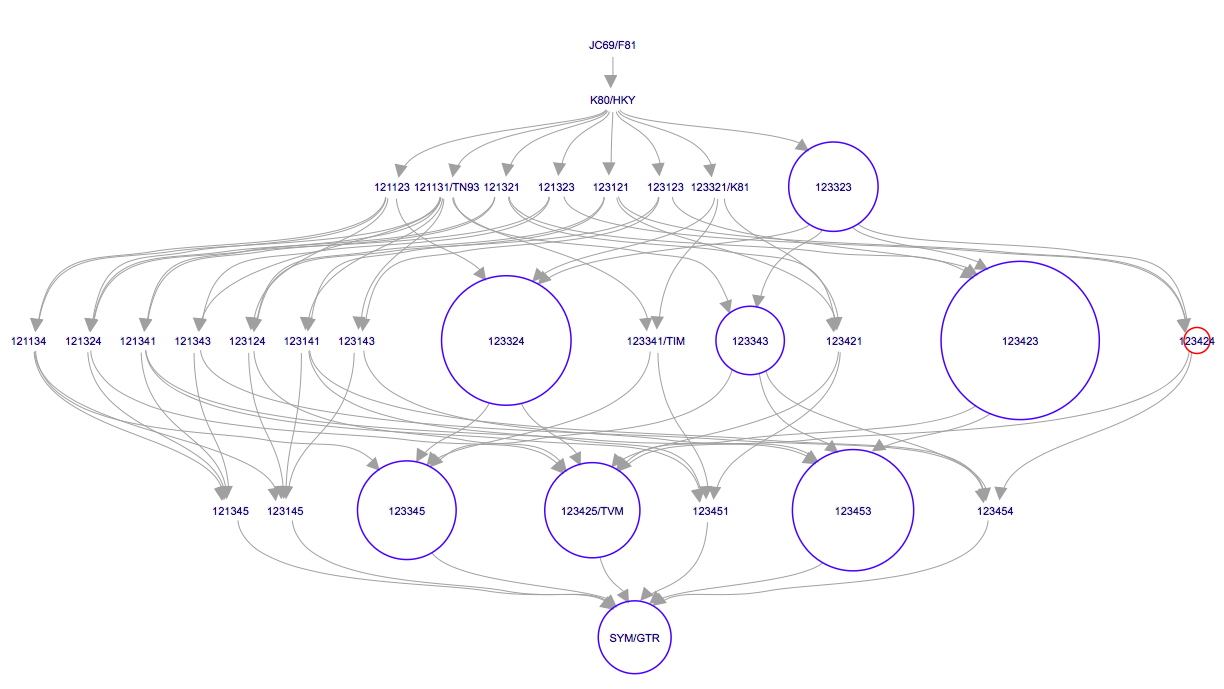
**Fig S1.** AppLauncher for BEAST2.

Select BModelAnalyser and click *Launch*. A window will pop up with a few options. Click on *browse* and find the log file for the h1n1 data and set the remaining options as shown in Fig S2.



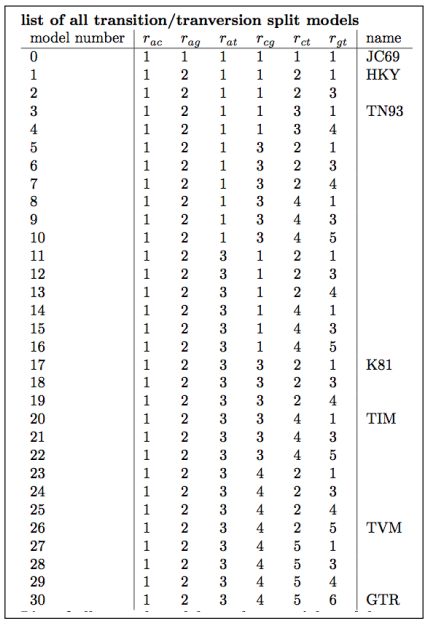
**Fig S2.** Options for BModelAnalyser.

Click OK. A browser with a figure similar to Fig S3 should appear in your default browser.



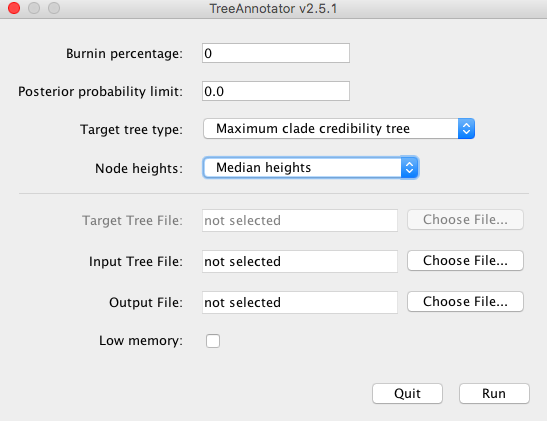
**Fig S3.** Models visited in BeastModelTest. Those in blue circles correspond to the 95% credible set of models, with their size proportional to their posterior probability. Those in red are outside the credible set, and those with no circles have less than 0.43 posterior support.

**Table 1.** Index of models in BeastModelTest (from the bModelTest tutorial by Ramussen et al.).



**Question:** Which model has the highest posterior probability? Does this model include among site rate heterogeneity? (hint: check the hasGammaRates and hasInvariableSites traces)

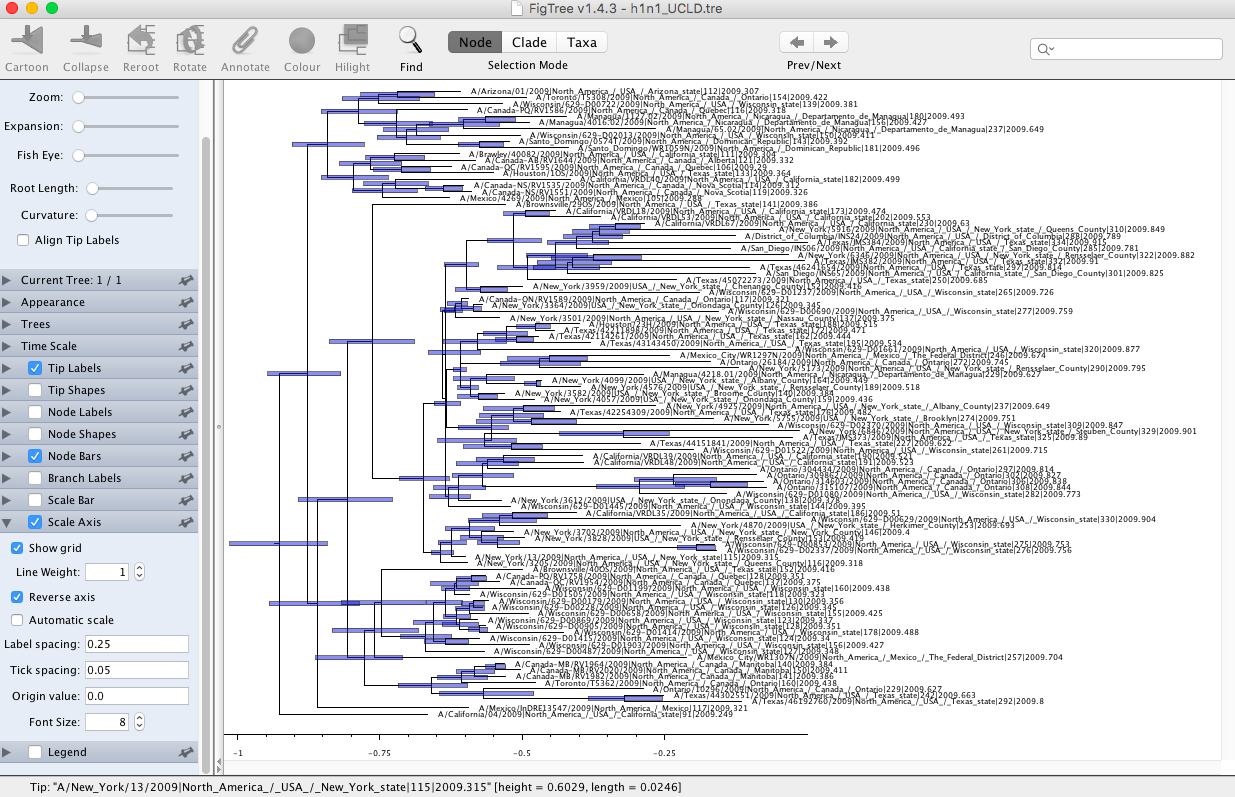
The .trees file contains trees sampled from the posterior. We can summarise them by using TreeAnnotator, which is distributed with the BEAST package. Double-click the TreeAnnotator icon. The window in Fig 14 will appear.



**Fig 14.** TreeAnnotator input window.

Type 10 for *Burnin percentage* and choose the same settings for *Target tree type* and *node heights* as shown in Fig. 14. In *Input Tree File* click on *Choose File...*, and select h1n1\_ucld.trees. In *Output File* click on *Choose File...* and type h1n1\_ucld.tre. **Note that we use the .tre extension for the output file**. Click on *Run*.

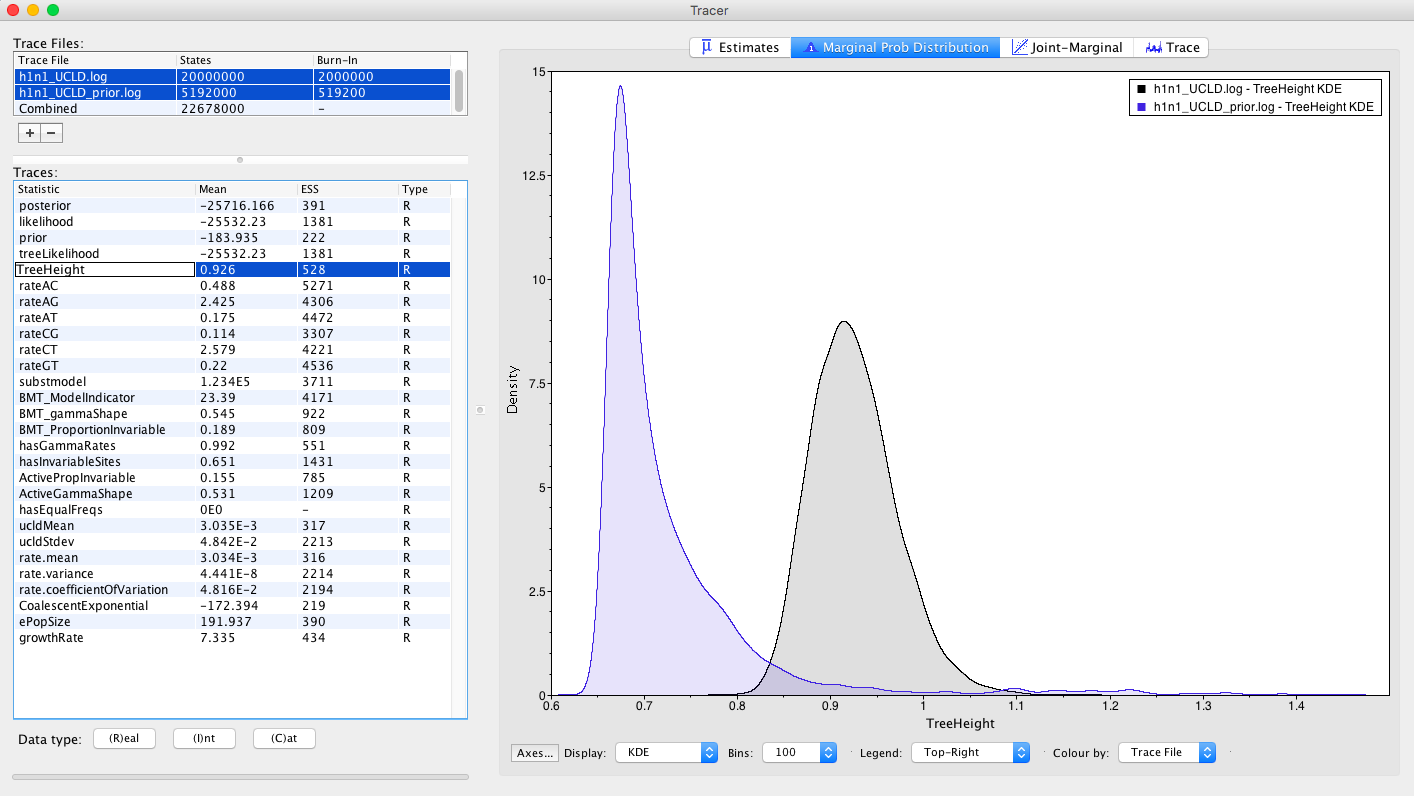
After the program has run, find the h1n1\_ucld.tre and open it in FigTree. Tick the *Node Labels* box and select *Display: node ages*. Tick the Node Bars box and select *Display: height 95% HPD* (Fig 15). Tick the *Scale Axis* box and tick *Reverse Axis*. The axis now represents time before the most recently collected sample, which is November 2009.



**Fig 15.** H1N1 tree shown in FigTree. The branch lengths correspond to time, and the blue error bars represent the uncertainty in the node age estimates.

**Optional exercise 1:** Use the BEAUTI window, which we left open, to sample from the prior distribution. This is useful to assess whether the data are informative about parameters of interest. To do this, go to the MCMC tab and tick the *SampleFromPrior* box. Change the names of the output log and trees files to *h1n1\_ucld\_prior.log* and *h1n1\_ucld\_prior.trees* and go to File, Save as, and save it as *h1n1\_UCLD\_prior.xml*. This analysis will run much faster because it does not need to calculate the phylogenetic likelihood. After it has run, load the log file with that from the posterior.

**Question:** Does it seem like our data are driving our estimates of evolutionary rates and timescales (hint: compare the prior and the posterior for the tree height, as in Fig S4, and for the *rate.mean* parameters).



**Fig S4.** Comparing the prior and posterior for the tree height.

**Optional exercise 2:** Use the BEAUTI window, which we left open, to set up a strict clock. To do this go to the *Clock Model* tab and select *Strict Clock*. In the MCMC tab change the output file names to h1n1\_sc.log, h1n1\_sc.trees. Save it as h1n1\_sc.xml and run it in BEAST. Compare the rate and node age estimates to those from the relaxed clock used here.