**Practical 3: Molecular dating in BEAST**

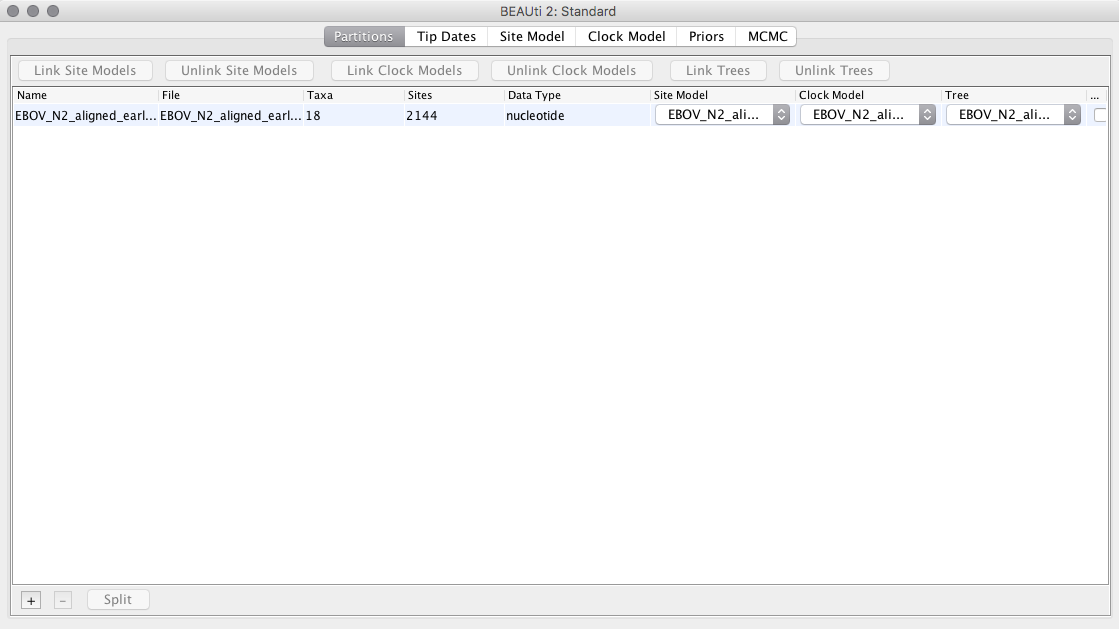
**Data set**

* Sequence alignment in fasta format of samples collected before the 2013-2016 epidemic: EBOV\_N2\_aligned\_early\_samples.fasta

**Software**

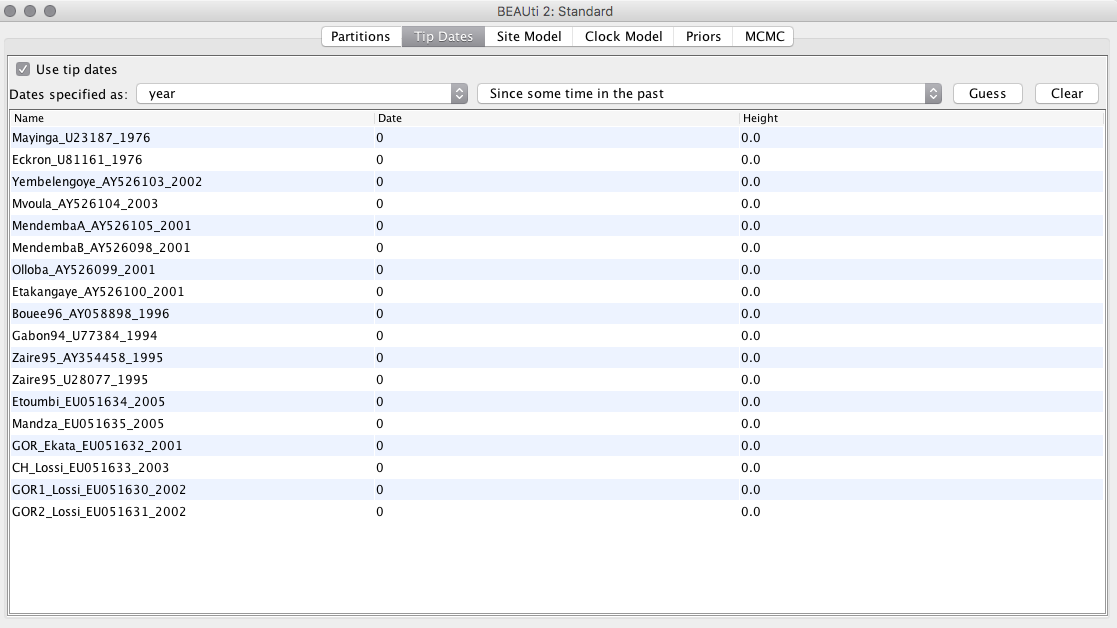
* BEAST 2.4 (beast2.org)
* BEAST2 addons: BDSKY and Phylodynamics

For the rest of the workshop, we will be using BEAST2. 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 (EBOV\_N2\_aligned\_early\_samples.fasta) to this window (Fig 1). Note that there are several tabs (Partitions, Tip Dates, Site Model, Clock Model, Priors, and MCMC).



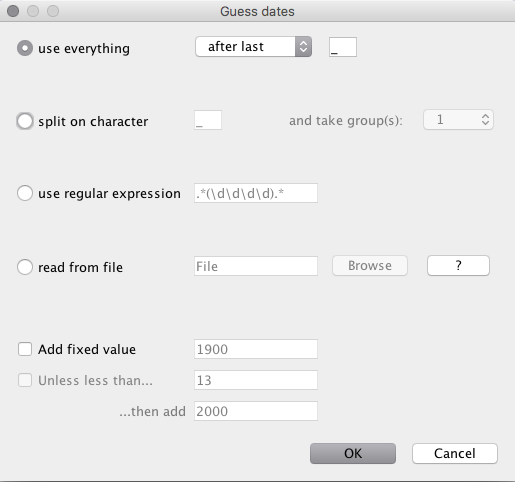
**Fig 1.** BEAUTI with EBOV alignment loaded.

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



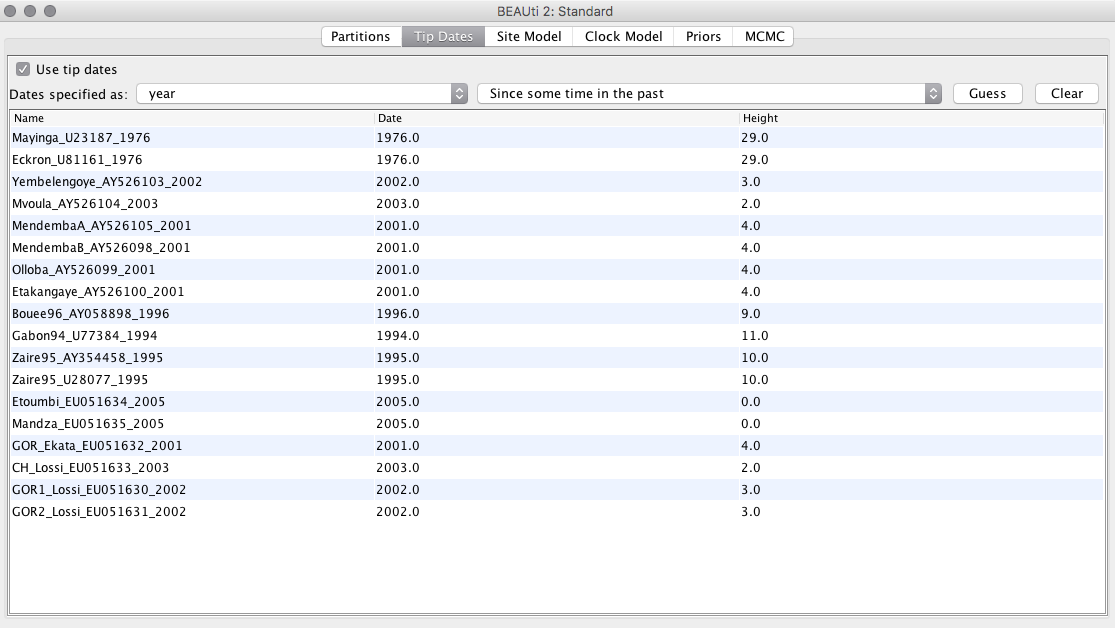
**Fig 2.** Tip dates enabled in BEAUTI.

To use the tip dates as calibrations, click on the  button. Check the first box (*use everything*) and in the dropdown menu, select *after last*, as shown in Fig 3.



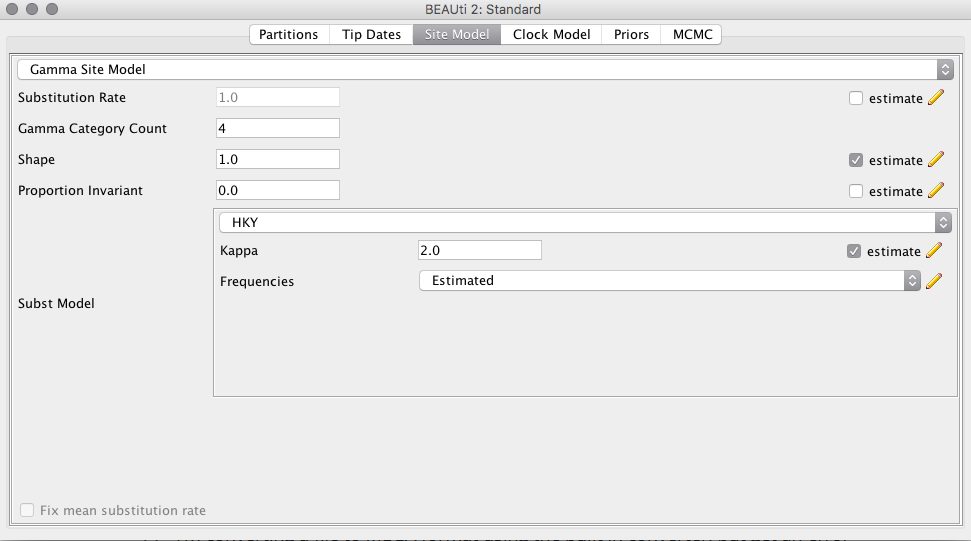
**Fig 3.** Specifying tip dates from sequence names in BEAUTI.

The BEAUTI window should now display the dates for each of the sequences (Fig 4).



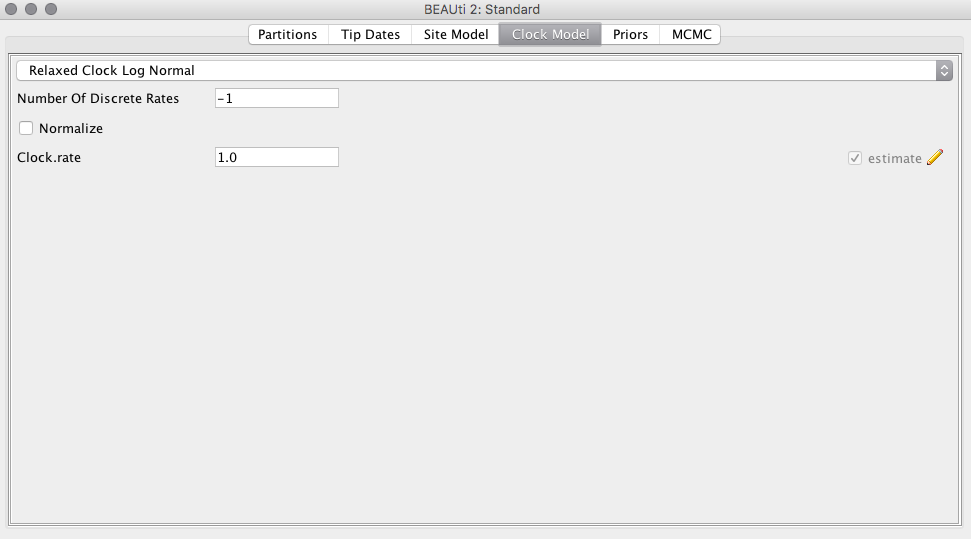
**Fig 4.** Dates assigned to tips in BEAUTI.

Click on the *Site Model* tab. In *Gamma category count* type 4, tick the box to estimate the shape, and select the HKY model from the *Subst Model* dropdown menu. These options should look like those in Fig 5.



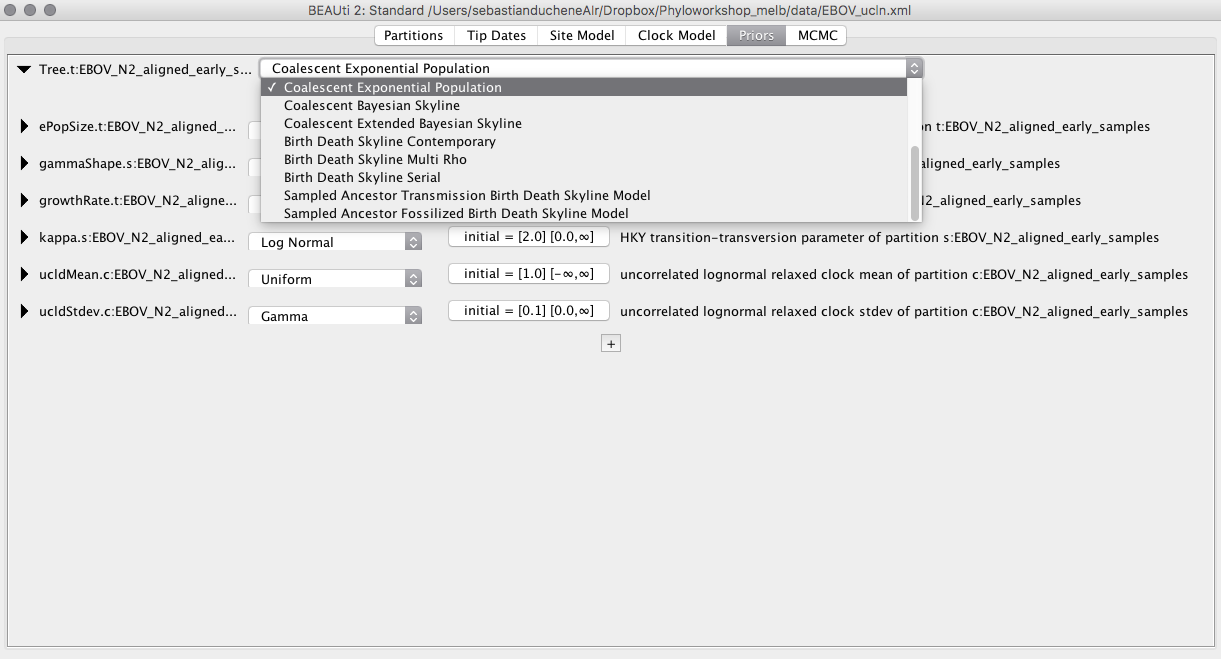
**Fig 5.** Substitution model set up in BEAUTI.

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



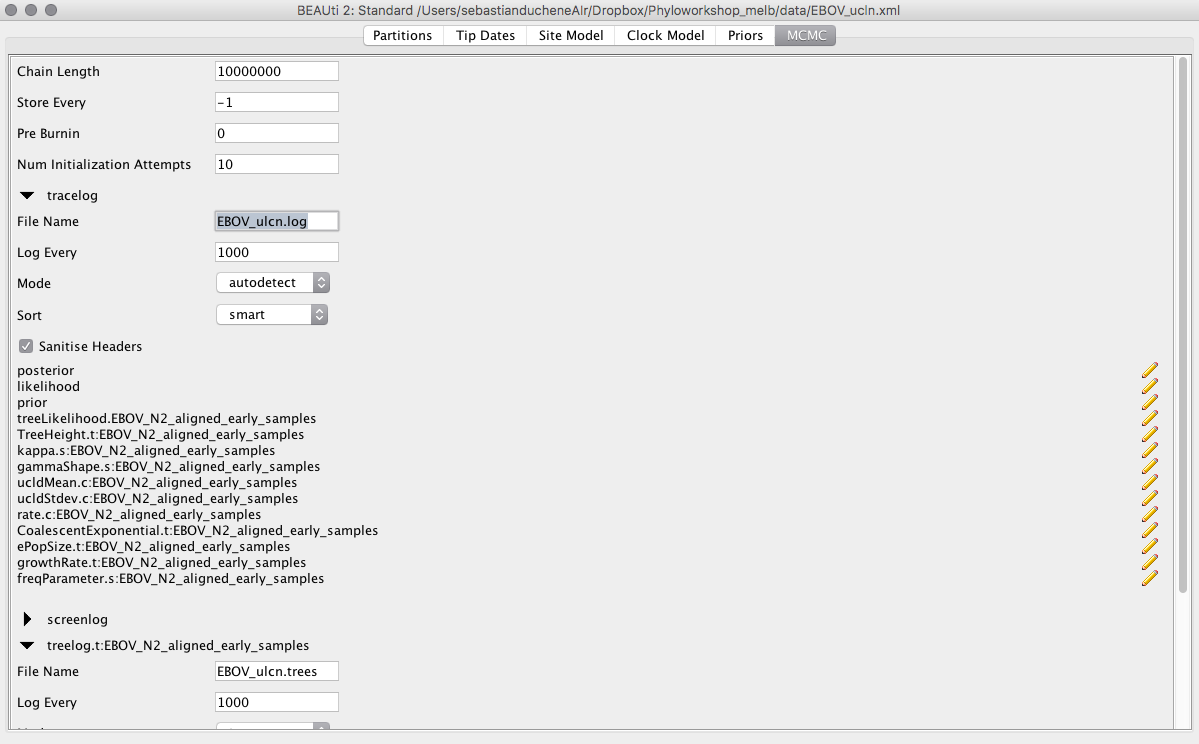
**Fig 6.** Molecular clock model set up in BEAUTI.

Click on the *Priors* tab. Select the Coalescent Exponential Population model (Fig 7). The remaining priors are fine for our analyses, but it is a good exercise to inspect these distributions.



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

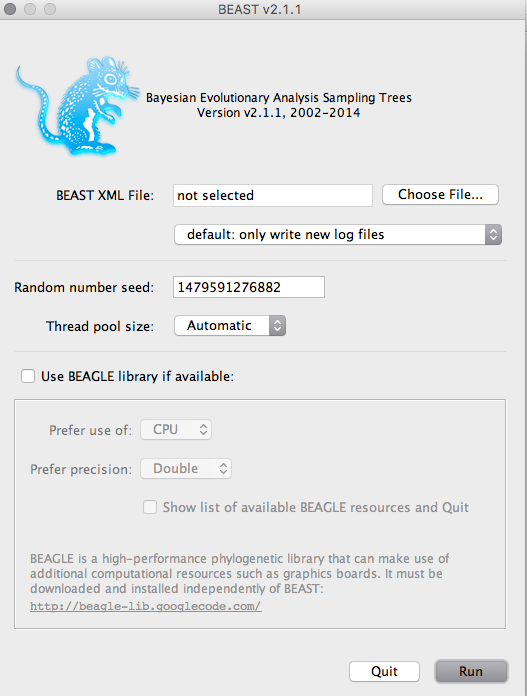
Click on the MCMC tab. Here, we can select different options for the MCMC. The chain length (number of iterations) is fine for these analyses. Click on *tracelog* and *treelog* to specify the BEAST output, which is a set of trees and parameters values, sampled from the posterior. Name the log and tree files EBOV\_ulcn.log and EBOV\_ulcn.trees, respectively (Fig 8). We use the extension \_ucln to refer to the uncorrelated lognormal clock model.



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

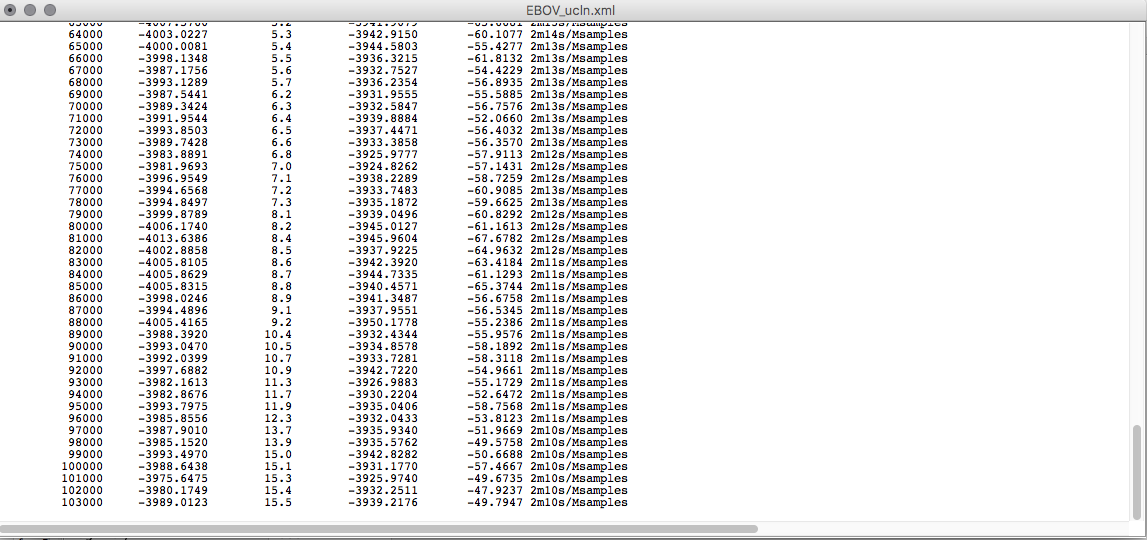
Our BEAST input file is ready. To save it, click on *File*, *Save*, and name it EBOV\_ucln.xml. Do not close BEAUTI.

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



**Fig 9.** BEAST starting window.

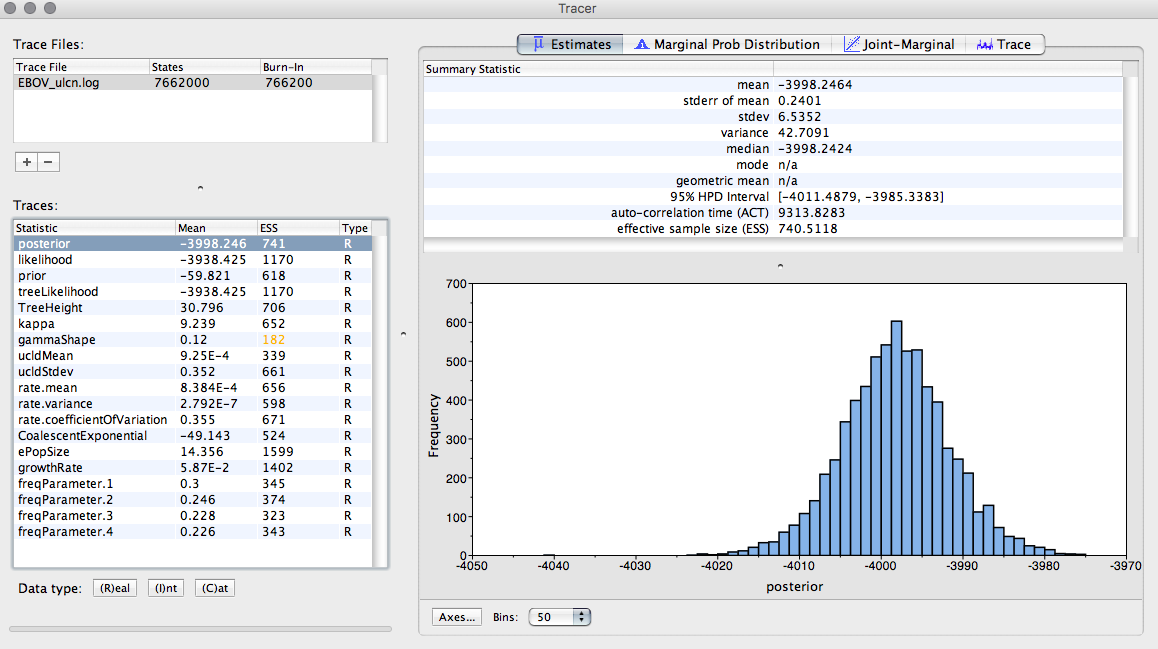
Click on *Choose File...* and select the xml file that we created in BEAUTI. Click Run. The MCMC will start running (Fig 10).



**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 10 to 15 minutes, but in large data sets or more complex models, it can take much longer.

After the program has run open Tracer(Fig 11), and drag the EBOV\_ucln.log file to the right pane of the Tracer window.

****

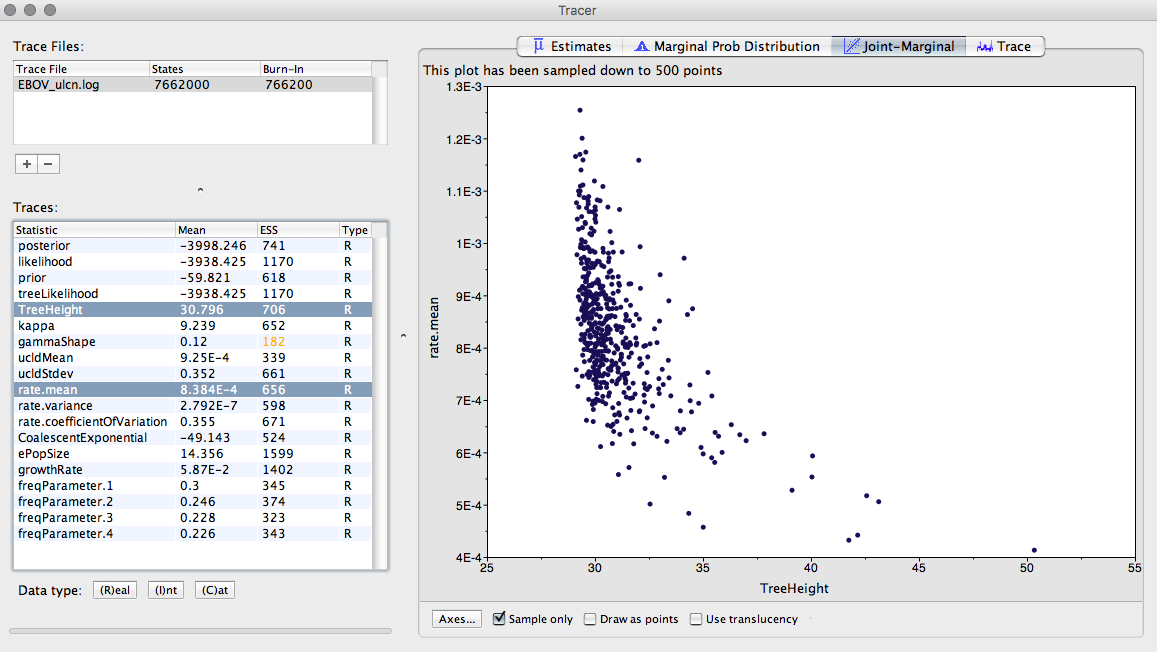
**Fig 11.** Tracer window.

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

**Question 3.1:** 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?

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).



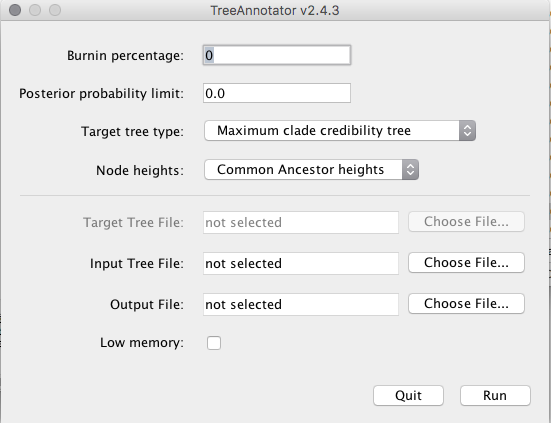
**Fig 12.** Joint marginal plot of TreeHeight and rate.mean.

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 3.2:** Does this data set appear to follow a strict clock, or does it display substantial rate variation among lineages? Is this consistent with what we observed in TempEst?

**Question 3.3:** When did these EBOV samples last share a common ancestor? Is this consistent with the first description of the virus in 1975?

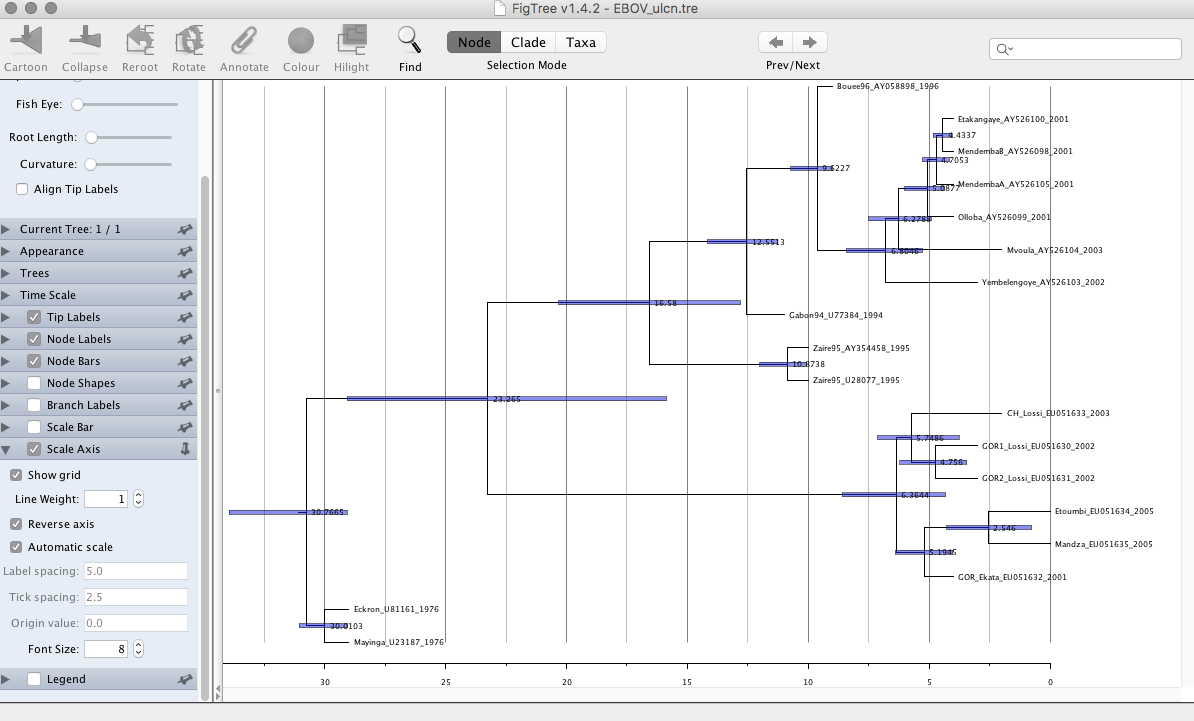
The .trees file contains trees sampled from the posterior. We can summarise them by using TreeAnnotator, which is distributed with the BEAST package. Double-click the TreeAnnotator icon. The window in Fig 13 will appear.



**Fig 13.** TreeAnnotator input window.

Type 10 for *Burnin percentage*. In *Input Tree File* click on *Choose File...*, and select EBOV\_ucln.trees. In *Output File* click on *Choose File...* and type EBOV\_ucln.tre. **Note that we use the .tre extension for the output file**. Click on *Run*.

After the program has run, find the EBOV\_ucln.tre and open it in FigTree (which we used yesterday for the maximum likelihood prac). Tick the *Node Labels* box and select *Display: node ages*. Tick the Node Bars box and select *Display: height 95% HPD* (Fig 14). Tick the *Scale Axis* box and tick *Reverse Axis*. The axis now represents time before the most recently collected sample, which is 2005.



**Fig 14.** EBOV tree shown in FigTree. The branch lengths correspond to time, and the blue error bars represent the uncertainty around the node ages.

**Optional exercise:** 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 EBOV\_sc.log, EBOV\_sc.trees. Save it as EBOV\_sc.xml and run it in BEAST. Compare the rate and node age estimates to those from the relaxed clock used here.