**Practical 4: Molecular epidemiology in BEAST**

**Data set**

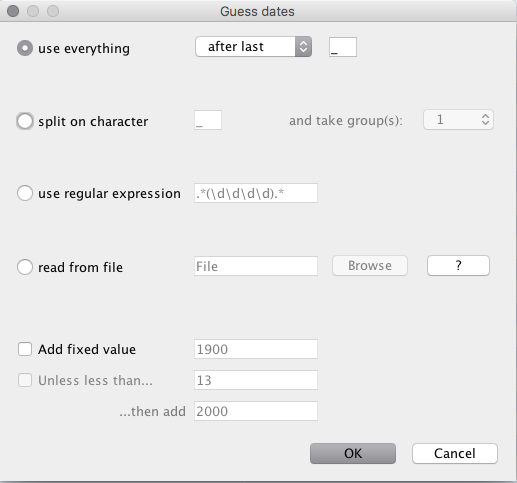
* Sequence alignment in xml format of samples collected in the early stages of the 2013-2016 epidemic: EBOV\_2014\_sequence\_data.xml

**Software**

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

In this session we will estimate Re for the early stages of the West African EBOV outbreak. We will use a set of genomes from Stadler et al. (2014). Importantly, because these data were obtained over a very short timeframe, they do not contain a sufficient number of substitutions. For this reason, we need to calibrate this analysis using a previous rate estimate. In their study, Stadler et al. set the rate to 1.9×10-3 subs/site/year. In our previous analyses, we estimated it at around 9.28×10-4 subs/site/year, which is quite similar.

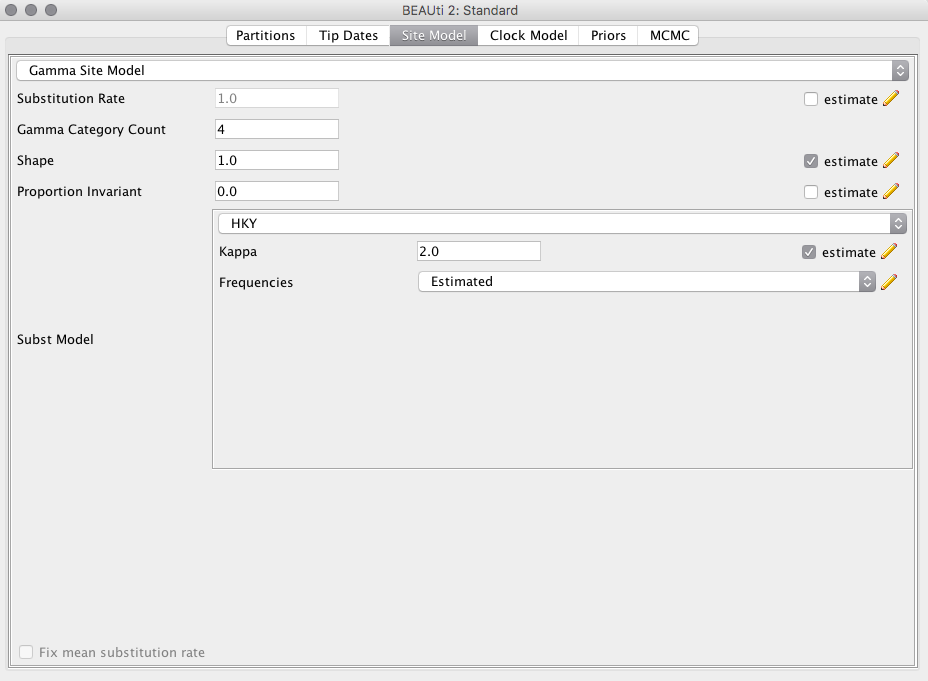
Open BEAUTI and drag and drop the EBOV\_2014\_sequence\_data.xml file. Go to the *Tip Dates* tab and check *Use Tip Dates*. Click on *Guess* and set up the options as shown in Fig 1.



**Fig 1.** Specifying tip dates in BEAUTI.

You should see a column with the year assigned to each sample. Note that they are specified as decimal years.

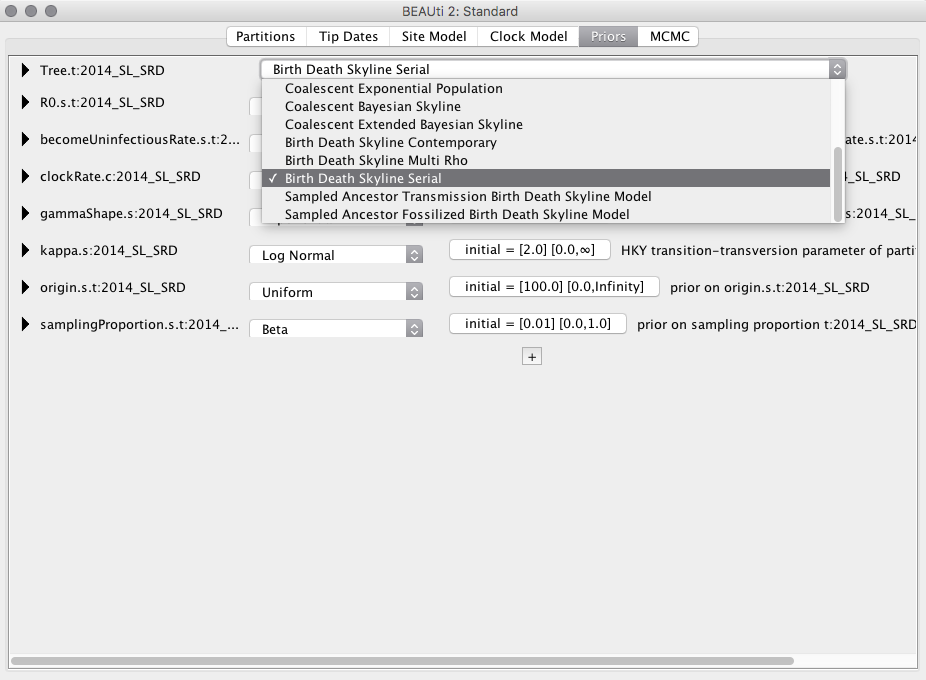
Go to the *Site Model* tab. Use 4 categories for the Gamma distribution, check the box to estimate the shape parameter, and choose the HKY model. Your Site model set up should look like that in Fig 2.



**Fig 2.** Set up of the site model in BEAUTI.

Go to the *Clock Model* tab. Make sure that the strict clock model is selected. Go to *Priors* tab.

Select the Birth-Death Skyline Serial tree prior. This model differs from the Birth-Death Skyline Contemporary in that it allows the samples to have different ages (Fig 3).



**Fig 3.** Select the Birth-Death Skyline Serial tree prior.

There are a number of priors that we need to specify as follows:

**R0.s.t:2014\_SL\_SRD**

Log Normal M = 0.0, S = 1.0 Do not check the *estimate* boxes.

**becomeUninfectiousRate.s.t:2014\_SL\_SRD**

Gamma, alpha = 0.5, beta = 61.0 Do not check the *estimate* boxes.

**clockRate.c:20014\_SL\_SRD** (This will act as our calibration)

Normal, mean = 0.001984, sigma = 4.592E-4

**gammaShape.s:2014\_SL\_SRD**

Exponential, mean = 0.5

**Kappa.s:2014\_SL\_SRD**

Use default Log Normal

**Origin.s.t:2014\_SL\_SRD** (this parameter is the time or origin of the outbreak, must have been sometime before the most recent common ancestors of the samples)

Uniform, upper=100.0

**samplingProportion.s.t:2014\_SL\_SRD** (this distribution needs to be informative because sampling proportion, R0, and the rate of becoming uninfecious are non-identifiable. It corresponds to a mean sampling proportion of about 0.65)

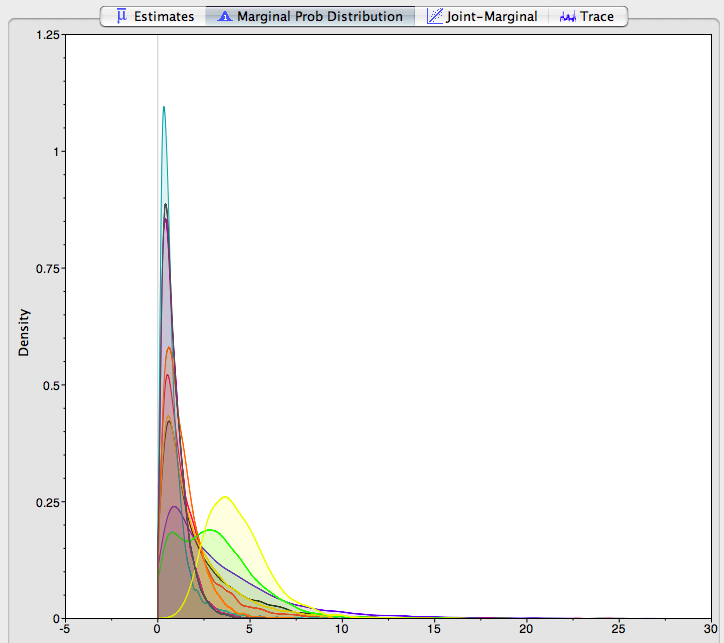
Beta, alpha = 10.0, beta = 6.0

Because these data are substantially larger and our model has more parameters than in our previous prac, we need to modify the MCMC. Go to the *MCMC* tab. Set the chain length to 25000000. In *tracelog*, change the file name to EBOV\_BDSky.log, and *Log Every* to 2500. Do the same for the *treelog* section.

Go to *File*, *Save*, and save the file as EBOV\_BDSky.xml.

Run BEAST as we did in the previous prac, but do not close BEAUTI. This analysis might take 15 minutes or more. However, you can open *Tracer* and open the log file as it runs. This will allow you to monitor the progress of the analysis.

When the analysis is complete, open the log file in *Tracer* as we did in the previous prac. Select all the Re estimates, which correspond to Re for different coalescent intervals in the tree. Click on the tab *Marginal Prob Distribution*, and at the bottom of the *Tracer* window select *Colour by: Trace* (Fig 5).

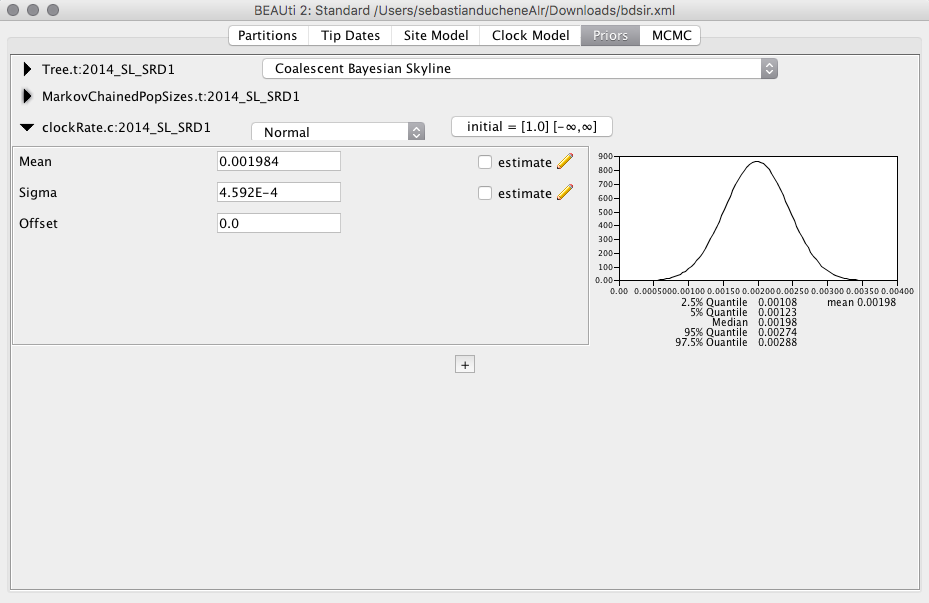


**Fig 4.** Probability density of R0 in *Tracer*.

**Question 1:** Go to the *Estimates* tab while selecting all the Re estimates. Does the uncertainty in our estimates tend to include previous estimates of R0 of 2.18?

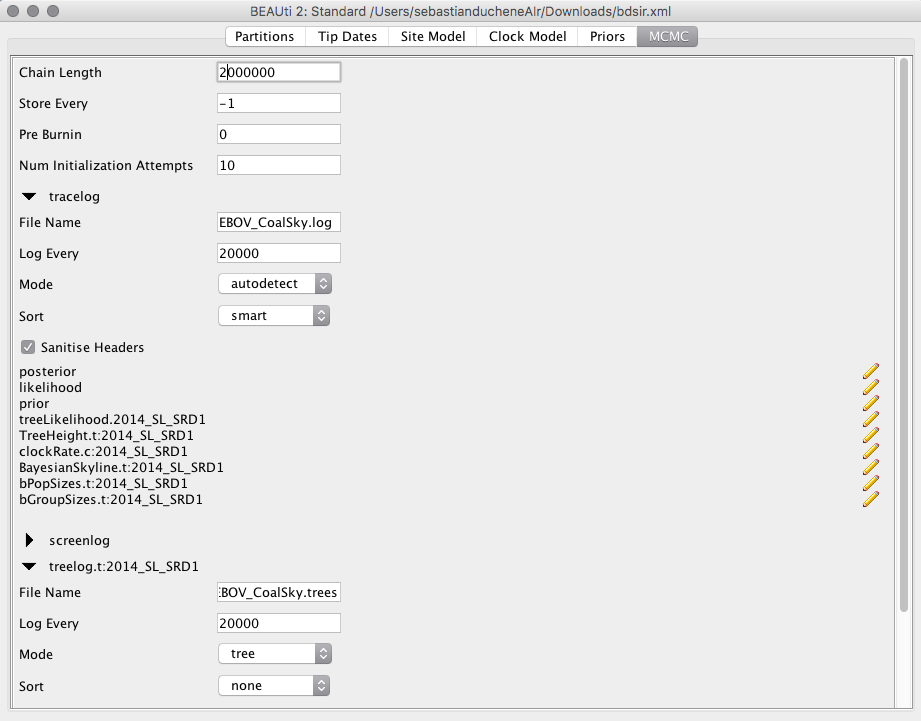
**Skyline plots**

We can visualise changes in pathogen population size over time. To do this, go back to the BEAUTI window and change the tree prior from Birth-Death Skyline to Coalescent Bayesian Skyline (Fig 5). The clock rate prior should be the same.

****

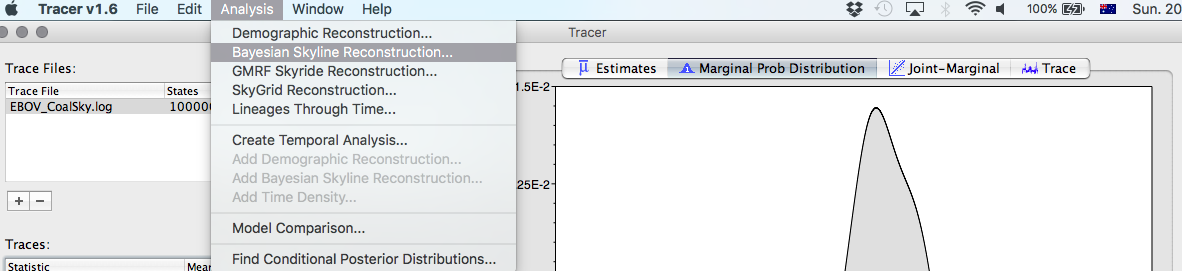
**Fig 5.** Setting up the Coalescent Bayesian Skyline.

Go to the MCMC tab and change the names of the log and trees files to EBOV\_CoalSky.log, and EBOV\_CoalSky.trees. Set the *Chain Length* to 2000000, and *Log Every*, to 2000 for the trees and log files (Fig 6). Save this file as EBOV\_CoalSky.xml. Run this new file in BEAST.



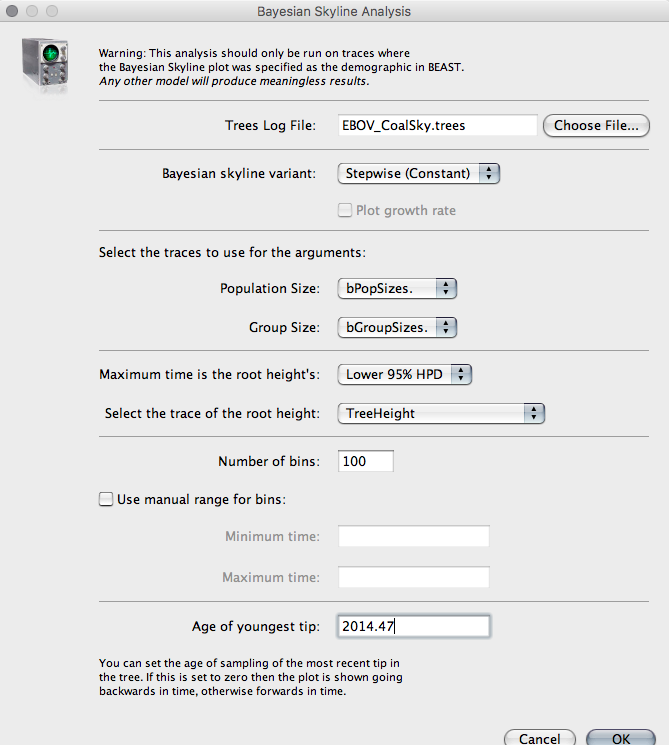
**Fig 6.** MCMC settings for Coalescent Skyline.

When the analysis has run, open *Tracer* and load the log file (drag and drop). Go to the Analysis drop-down menu and select *Bayesian Skyline Reconstruction* (Fig 7).



**Fig 7.** Select the Bayesian Skyline Reconstruction.

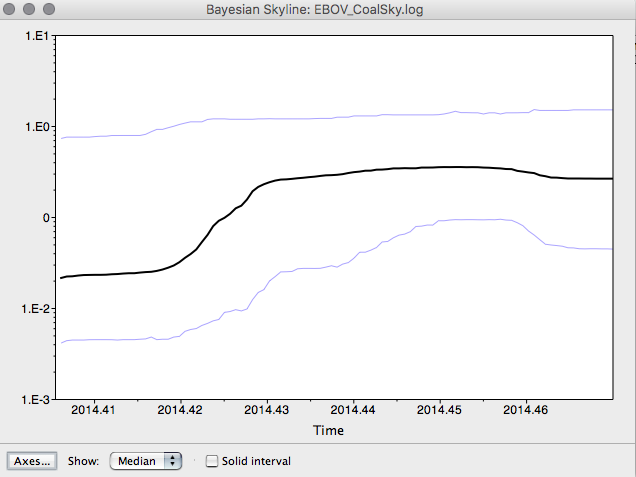
A window with some options will appear. In *Tree Log File* click on *Choose File..* and select the .trees files from the BEAST analysis in the previous step. For the *Age of youngest tip* type 2014.47, which is the age of the most recently sampled sequence. The remaining default options are fine for this analysis (Fig 8).



**Fig 8.** Setting up Bayesian Skyline reconstruction.

Click OK. After a few seconds, you will see a plot of the log10 population size of the virus as a function of time (Fig 9). Note that population size here does nt correspond to the number of virions or infected individuals. This parameter corresponds to the product of individuals and the virus generation time. Thus, to obtain the effective numbers of individuals this quantity must be divided by the generation time.

**Question 2:** Does the skyline plot support a rapid growth in the population size at this early stage?

****

**Fig 9.** Skyline plot of EBOV sequence data. The black line is the mean population size, while the blue lines represent the upper and lower 95% highest posterior densities.