Estimating Epidemiological Parameters of an Ebola Outbreak using BEAST2

Tutorial by Tracy Heath and Tanja Stadler

This tutorial uses the sequence data from 72 Ebola patients in Sierra Leone. The viral sequences were first presented in Gire et al. (Science 2014). These data were re-analyzed under complex birth-death processes to estimate key epidemiological parameters by Stadler et al. (PLoS Currents Outbreaks 2014).

In this exercise, we will perform a simplified analysis similar to one conduced by Stadler et al. (2014). We will use the birth-death process with serial sampling and piecewise shifts in rates (the “BD” model in the Stadler et al. 2014 study). Please refer to Stadler et al. (2014) and Stadler et al. (2013) for detailed descriptions of these models and methods.

For more information about divergence-time estimation in general and BEAST v2.2, please refer the resources and tutorials on: http://phyloworks.org/workshops/divtime.html. In particular see the detailed tutorial on estimating speciation times using extant and fossil data and the links and references therein: http://treethinkers.org/tutorials/divergence-time-estimation-using-beast/.

Important parameters for infectious disease dynamics:

- R is the **effective reproductive number**. Stadler et al. (2013) states that this parameter is the "number of expected secondary infections of an infected individual. The effective reproductive number is closely related to the basic reproductive number (Anderson and May 1979): the latter additionally assumes a completely susceptible population, and thus the two quantities are equal at the start of an epidemic outbreak." [called "R" in BEAST2]
- δ is the **rate of becoming non-infectious**. An individual may become non-infectious if they are cured or treated, their behavior changes, or they die. [called "becomeUninfectiousRate" in BEAST2]
- s is the **probability of sampling an individual upon becoming non-infectious**. [called "samplingProportion" in BEAST2]
- λ is the **rate of transmission (birth rate)**. [called "birth" in BEAST2]
- μ is the **viral lineage death rate**. [called "death" in BEAST2]
- ψ is the **rate each individual is sampled**. [called "sampling" in BEAST2]

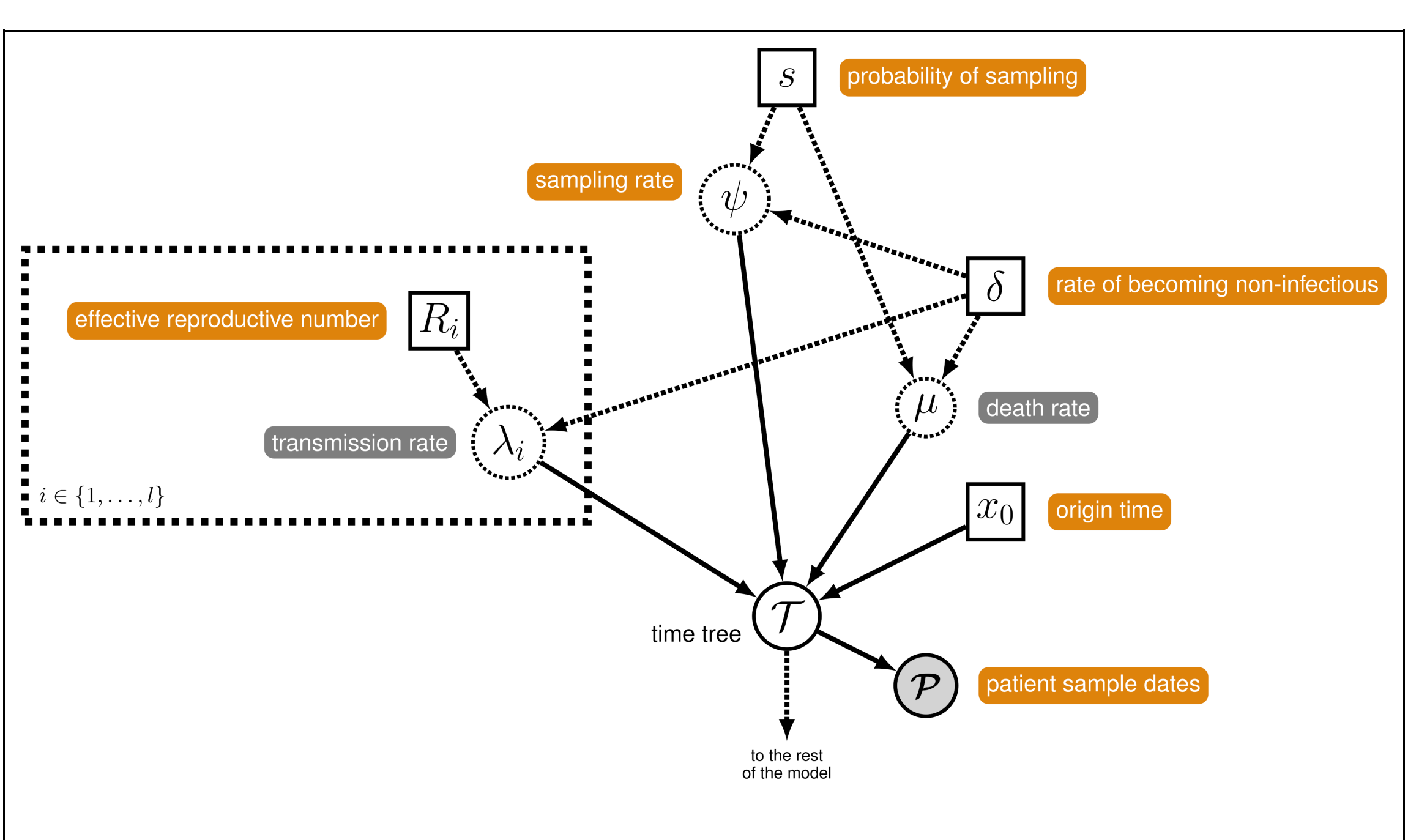
In the analyses, we will estimate R, δ, and s. The remaining parameters listed above can be calculated if we have the first three because they are defined:

$$R = \frac{\lambda}{\mu + \psi}, \quad \delta = \mu + \psi, \quad s = \frac{\psi}{\mu + \psi}$$

Thus we can compute λ, μ, and Ψ given:

$$\lambda = R\delta, \quad \mu = \delta - s\delta, \quad \psi = s\delta$$

In this exercise, we also assume that there are l intervals, evenly spaced, over the tree. These intervals are delineated by shifts in the value of R and because λ is determined by R, λ also changes over time. Below is the probabilistic graphical model depicting the conditional dependence structure of all of the parameters in our skyline birth-death model. It is also possible to consider a model the other parameters (δ, s, μ, Ψ) also change over time as in Stadler et al. (2013), however, we will not consider that more complex model in this exercise.



Probabilistic graphical model of the skyline birth-death process. The parameters R, δ, s, λ, μ, and Ψ are defined above, as is l. The origin time x_0, is the start of the epidemic. This graph shows that the time tree $\mathcal{T}$ is conditionally dependent on all of the parameters with arrows pointing to it. The list of patient sample dates, $\mathcal{P}$, are treated as observed data and assumed to be generated by the skyline birth-death process. Additionally, the arrow extending from the time tree to the rest of the model indicates connects this graph to a larger set of relationships defining the entire phylogenetic continuous-time Markov model (including the relaxed-clock model, sequence evolution model, etc.). For this model we are using the notation described in Höhna et al. (2014).

Exercise

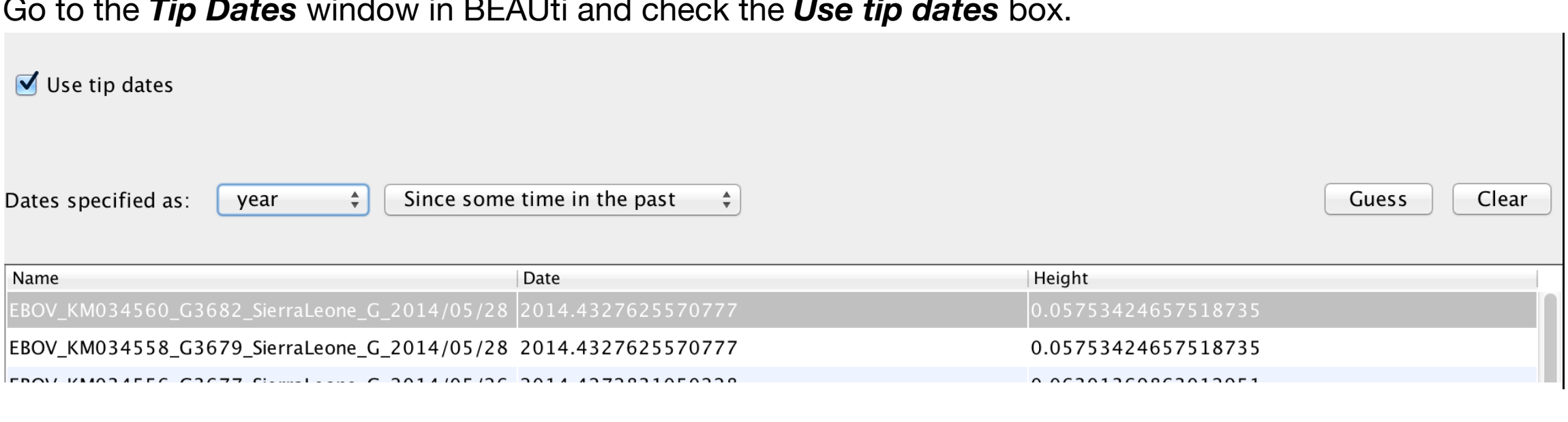
This exercise requires ***BEAST v2.3*** (http://beast2.org/). It also requires that you install the ***BDSKY*** package in the package manager found in BEAUti. Open the package manager using the menu options: ***File→Manage Packages***. Re-launch BEAUti *after* installing the package before proceeding with the tutorial.

Download the sequence data from: http://bit.ly/1JjoSaP. This file is called "`ebola_simple.nex`".

Open BEAUti and import the sequence from `ebola_simple.nex`. To do this go to the menu: ***File→Import Alignment***.

Note that the sequences are all given explicit names indicating the patient, location, and date of the sample: '`EBOV_KM034560_G3682_SierraLeone_G_2014/05/28`'. These serially sampled sequences need to be given dates. The dates and tip ages can be extracted from their names.

Go to the ***Tip Dates*** window in BEAUti and check the ***Use tip dates*** box.

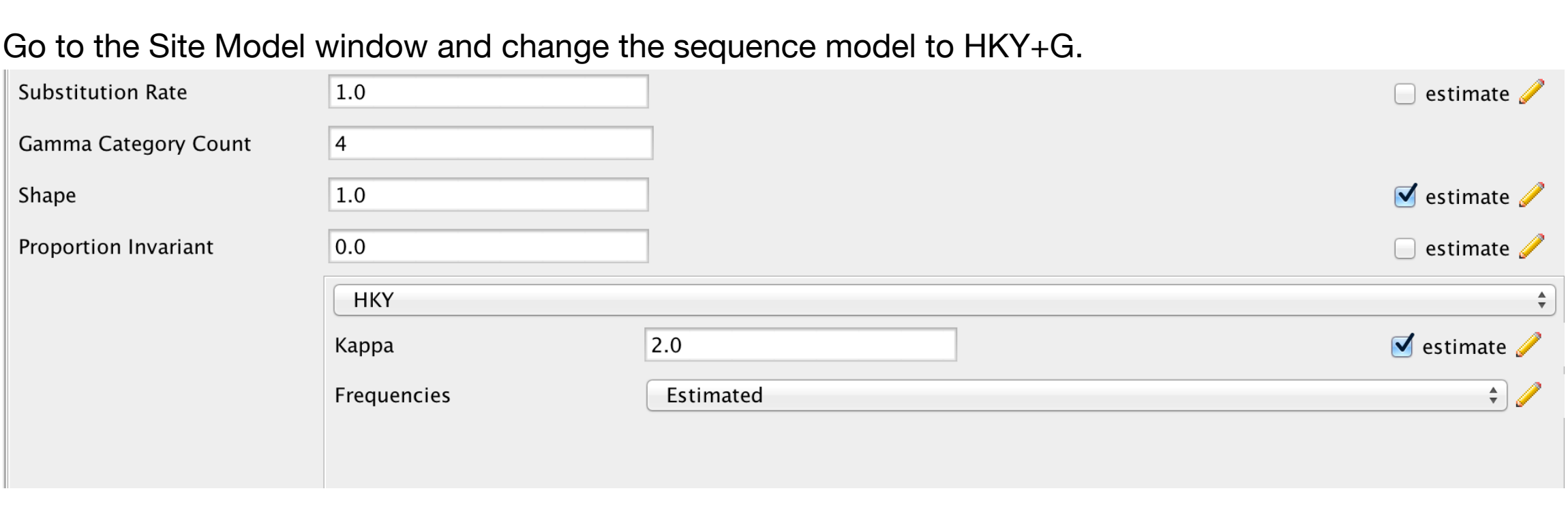


With the dates specified as ***year*** and ***Since some time in the past***, click on the button called ***Guess***.

BEAUti will extract the dates from the sequence names, given that they are provided using a specific text pattern. Here, the dates all follow the last "_" character. Indicate that you want to ***use everything*** after this character and click ***OK***.

You can see now that each sequence is given a date that is a value relative to year 0. Thus, the sequence sampled on 28 May 2014 is 2014.4327625570777. The tip heights are computed relative to the most recent sample, which has a height of 0.0. The units are in years, thus one day is a difference in tip height of 1/365 = 0.002739726.

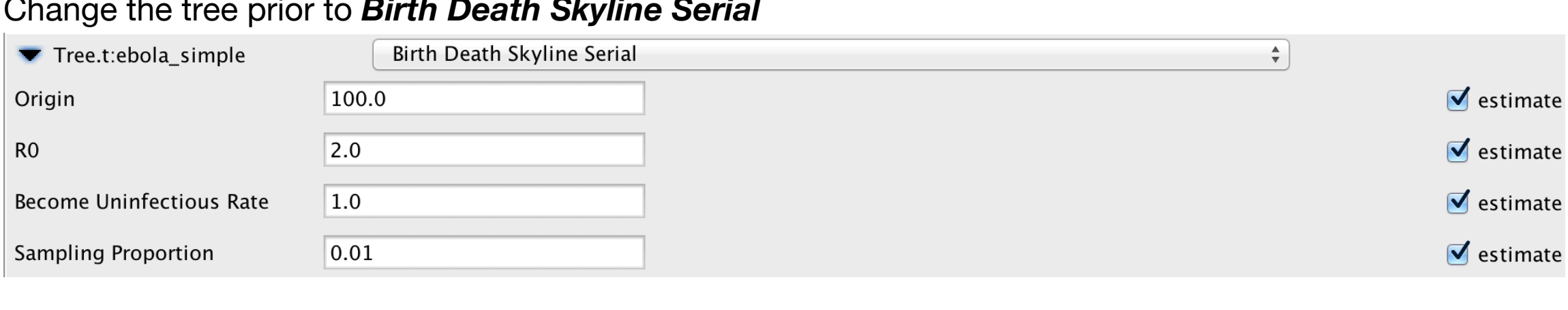
Go to the Site Model window and change the sequence model to HKY+G.



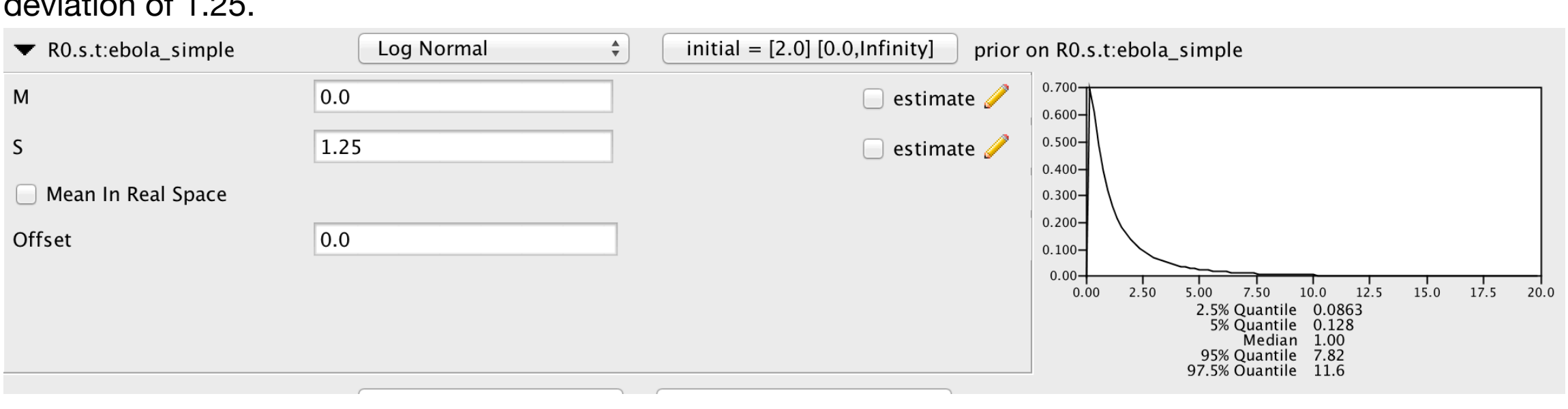
For this analysis, in the ***Clock Model*** window we will leave the default ***Strict Clock*** and estimate the clock rate.

Go to the ***Priors*** window.

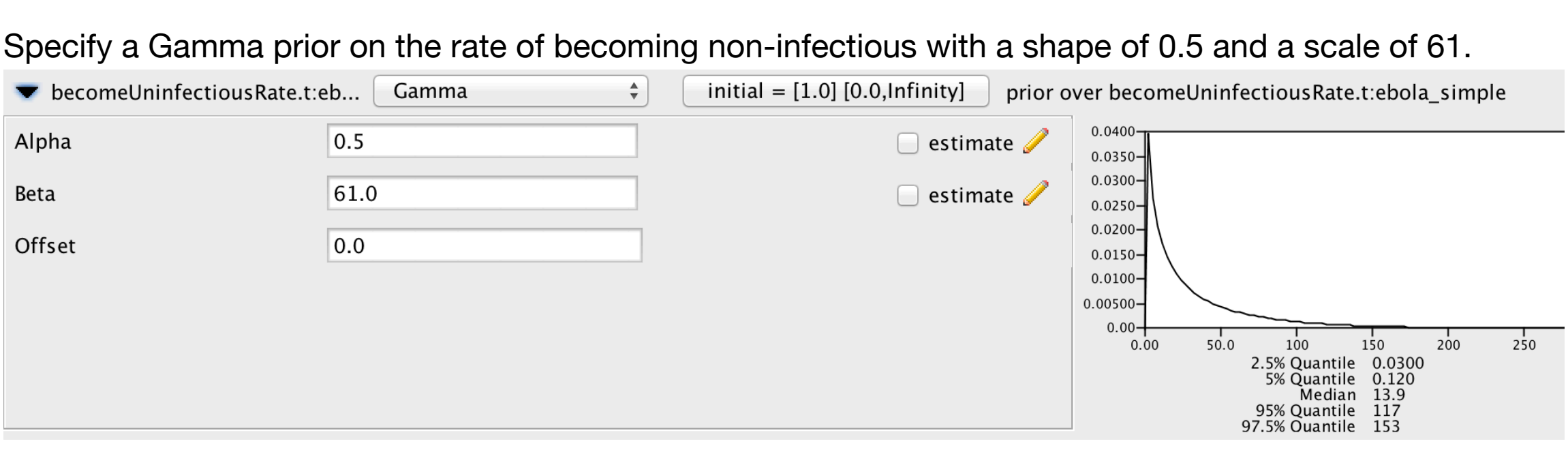
Change the tree prior to ***Birth Death Skyline Serial***



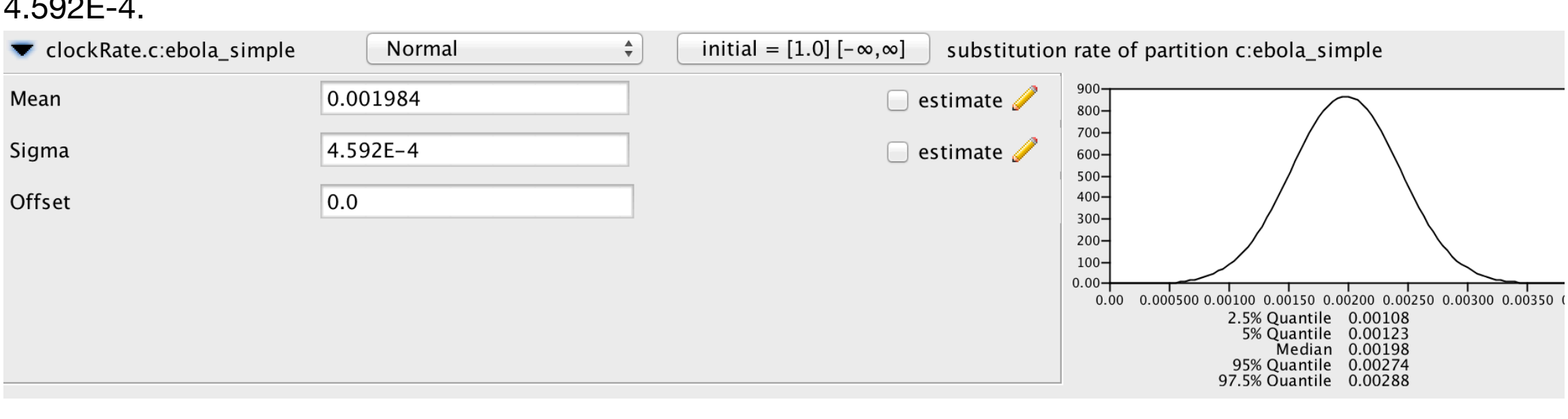
Set the prior on the effective reproductive rate (R) to a lognormal with a log-mean of 0.0 and standard deviation of 1.25.



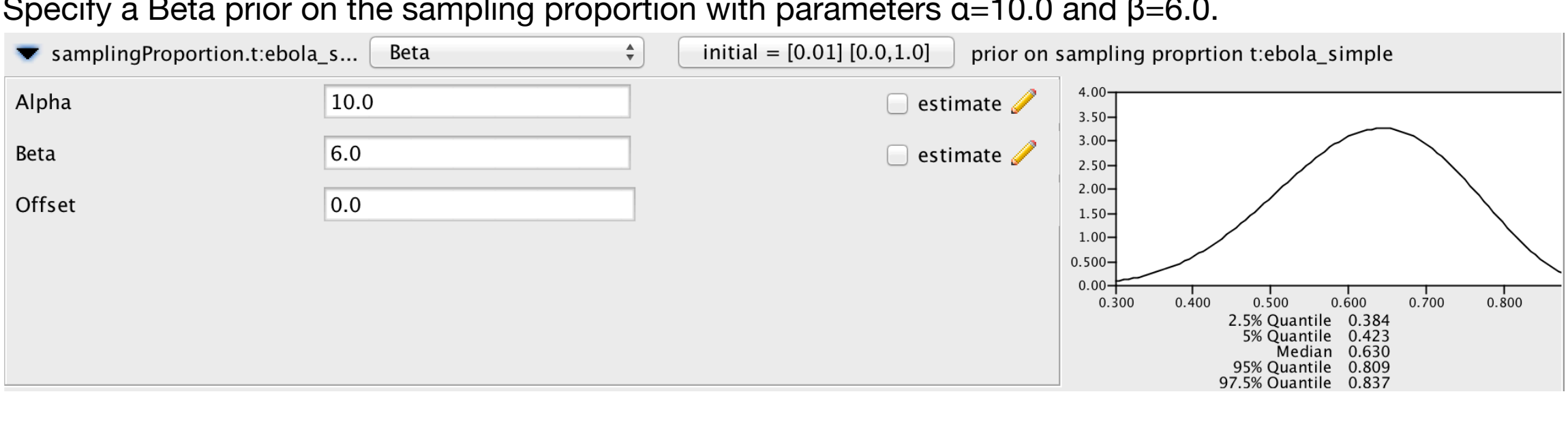
Specify a Gamma prior on the rate of becoming non-infectious with a shape of 0.5 and a scale of 61.



Set the prior on the clock rate to a normal distribution with a mean of 0.001984 and a standard deviation of 4.592E-4.



Specify a Beta prior on the sampling proportion with parameters α=10.0 and β=6.0.



Now go to the Initialization menu. To reveal this, you have to go to ***View→Show Initialization panel***.

By default, the birth-death skyline model assumes that there are 9 shifts in the R, δ, and s parameters over time. We want to change this.

For ***R***, set the ***Dimension*** to "**3**". This means that we are assuming a model where the effective reproductive number changed two times after the start of the epidemic.

Make sure the ***Dimension*** is set to "**1**" for the **samplingProportion** and **becomeUninfectiousRate**. Here, we are assuming that these parameters remain constant over time, even though the effective reproductive number changed over the course of the epidemic.

Go to the ***MCMC*** panel to set the sampling frequency and chain length.

Change the ***Chain Length*** to 20,000,000.

Set the sampling frequency (***Log Every***) to 2,000 iterations for both the ***tracelog*** and ***treelog***.

Save the XML file by going to ***File→Save As***. (Perhaps name the file "ebola_simple.xml".)

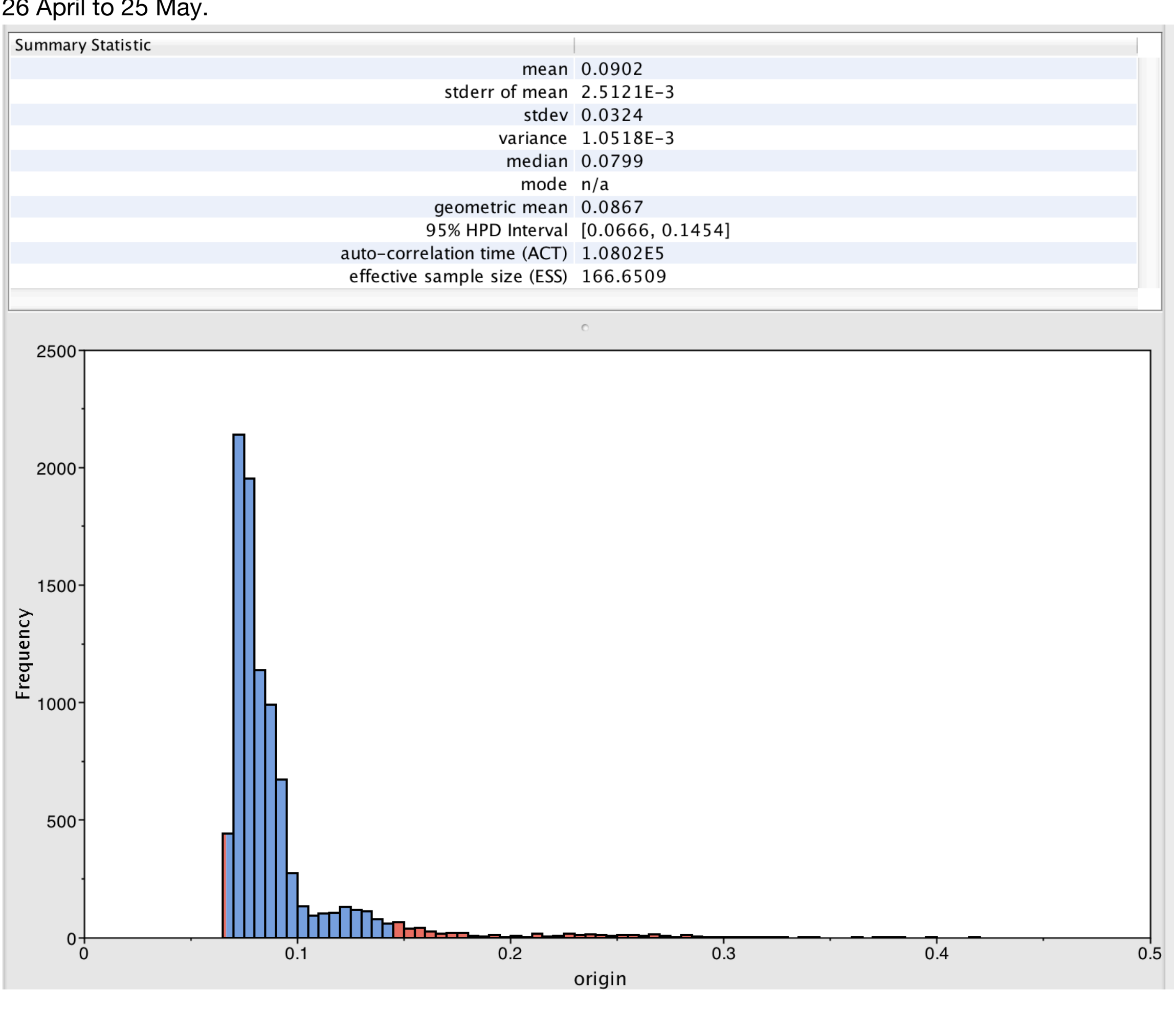
This file should now run in BEAST.

When the run terminates, open the log file in ***Tracer*** and view the summaries of each parameter.

Look at the estimate of the ***origin*** time. The origin is the time of infection of the first person in the outbreak (Stadler et al. 2014). Below is a histogram of the origin time from an analysis of these data. The mean estimated time is 0.0902. We can compute the origin time in terms of the number of days before the last sample since we know that our time is in units of years. Thus, the origin time is 32.9 days before the most recently sampled sequence:

$$\frac{0.0902}{\frac{1}{365}} = 32.9$$

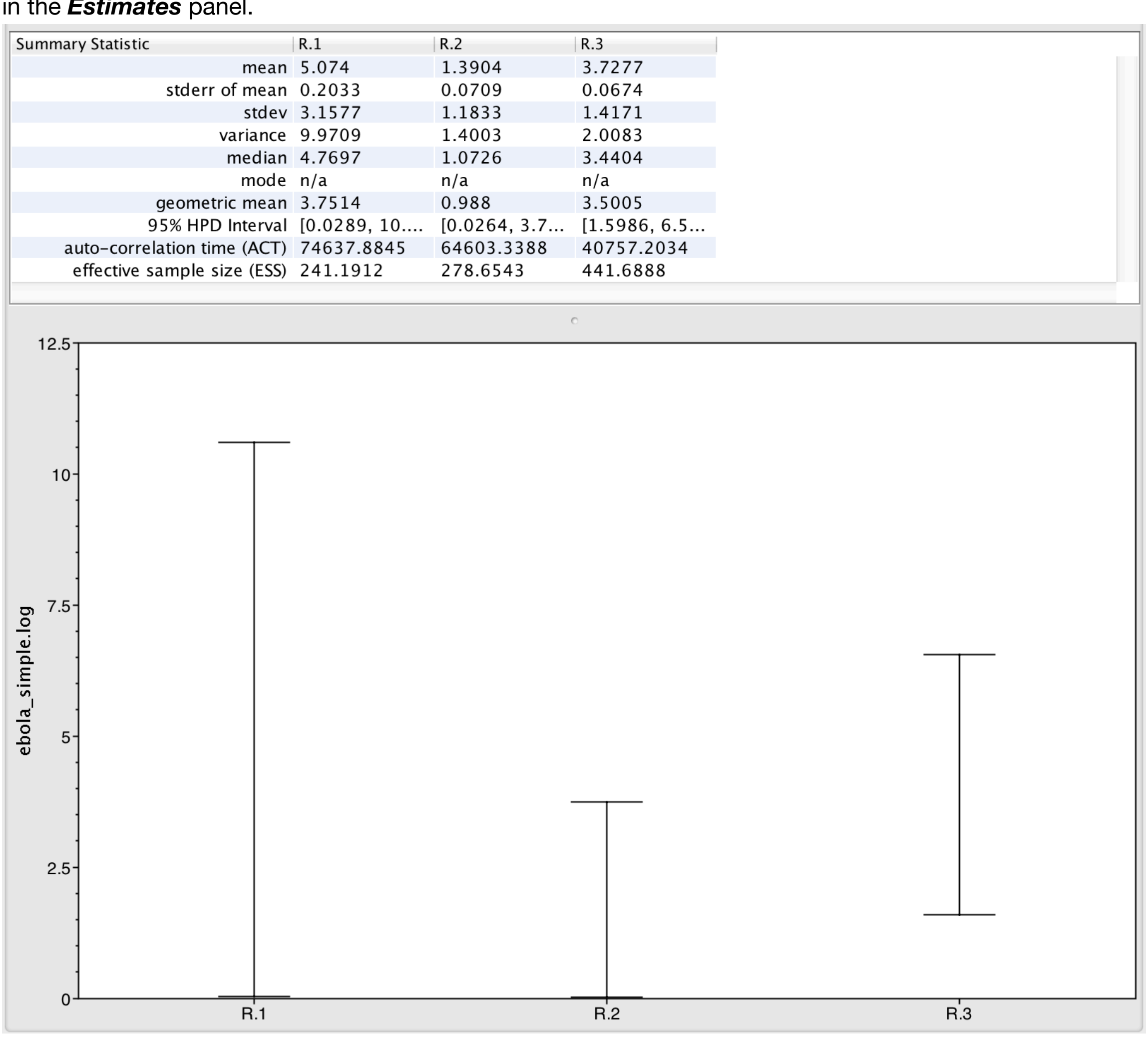
The last sequences were sampled on 18 June 2014, which is the 169th day of 2014. Thus, the mean origin time corresponds to 17 May 2014 (the 137th day of 2014), with a 95% highest posterior density interval of 26 April to 25 May.



Summary Statistic	
mean	0.0902
stderr of mean	2.5121E−3
stdev	0.0324
variance	1.0518E−3
median	0.0799
mode	n/a
geometric mean	0.0867
95% HPD Interval	[0.0666, 0.1454]
auto-correlation time (ACT)	1.0802E5
effective sample size (ESS)	166.6509

Calculate the mean and 95% HPD interval dates for the ***TreeHeight*** parameter using the formulas given above. This parameter is the date of the root or most-recent-common-ancestor of all sampled sequences.

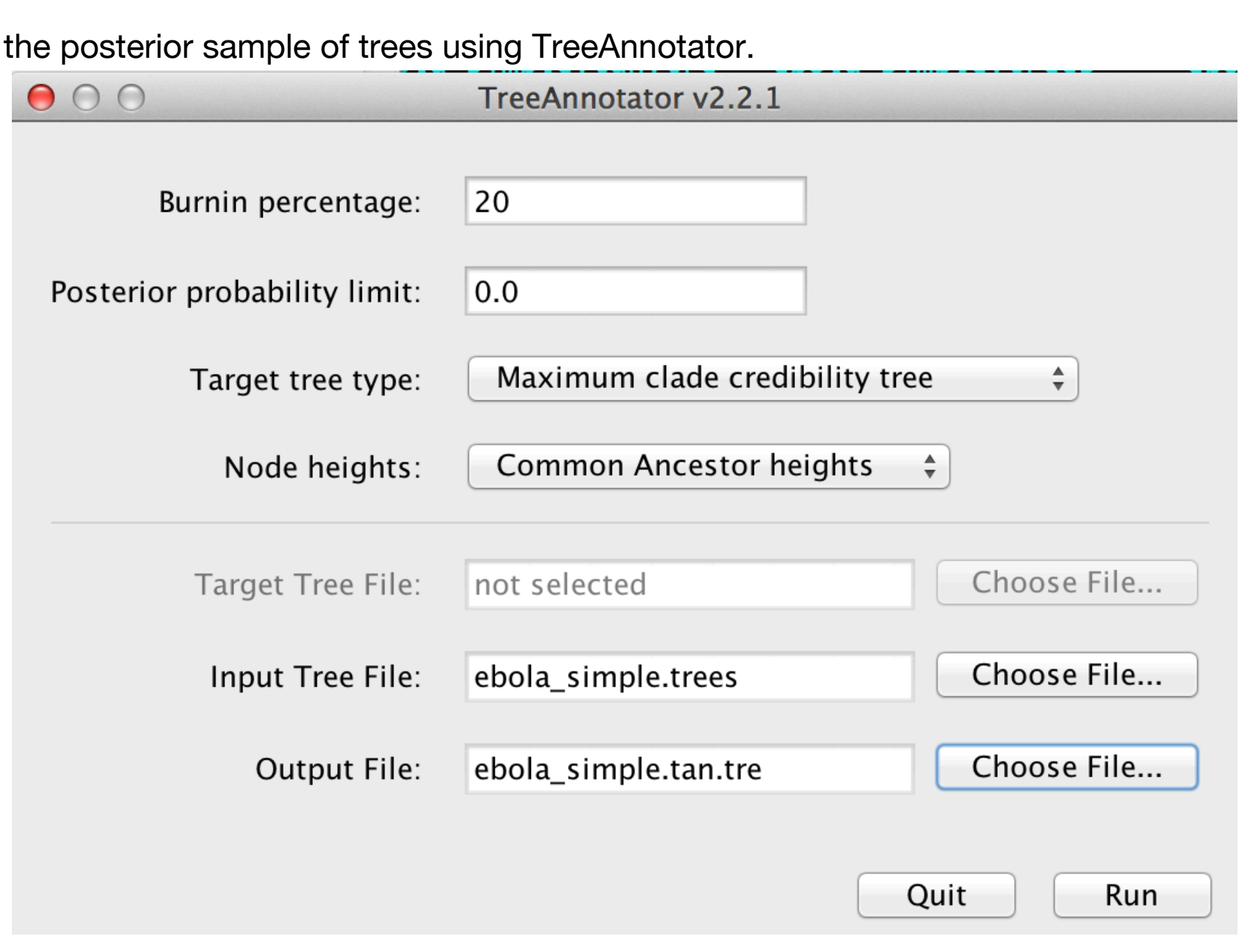
You can visualize the change in the effective reproductive number over time by selecting ***R.1***, ***R.2***, and ***R.3*** in the ***Estimates*** panel.



Summary Statistic	R.1	R.2	R.3
mean	5.074	1.3904	3.7277
stderr of mean	0.2033	0.0709	0.0674
stdev	3.1577	1.1833	1.4171
variance	9.9709	1.4003	2.0083
median	4.7697	1.0726	3.4404
mode	n/a	n/a	n/a
geometric mean	3.7514	0.988	3.5005
95% HPD Interval	[0.0289, 10....	[0.0264, 3.7...	[1.5986, 6.5...
auto-correlation time (ACT)	74637.8845	64603.3388	40757.2034
effective sample size (ESS)	241.1912	278.6543	441.6888

The estimates of the reproductive number and origin times differ from those reported in the Stadler et al. (2014) study. This is because we are using slightly different priors and a reduced dataset. It isn't currently possible to replicate the model used in their study by simply setting up the analysis in BEAUti. This more complex model, that includes variable sampling through time and other extensions of the piecewise birth-death process discussed here, one must alter the XML file directly.

Summarize the posterior sample of trees using TreeAnnotator.



The maximum clade credibility tree has very low support for these data.