



Inferring epidemiological parameters with BEAST 2.3.0

Louis du Plessis and Veronika Boskova

This tutorial is an introduction to using the BEAST software package for Bayesian evolutionary analysis for inferring epidemiological parameters. The aim of this tutorial is to learn how to set up a BEAST analysis using BEAUti, run the analysis and then analyse the results using Tracer and TreeAnnotator. This tutorial is written for BEAST version 2.3.0, which is available at <http://beast2.org>. The BEAST 2.3.0 installation includes BEAUti and TreeAnnotator. Tracer is available at <http://beast.bio.ed.ac.uk/>.

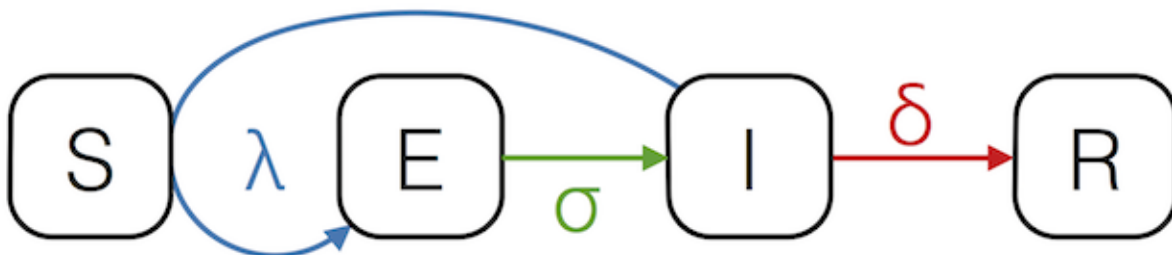
BEAST takes as input an XML file containing the model specification and parameter settings. BEAUti is a user friendly tool that provides a graphical user interface for easily creating BEAST input files. Tracer is a program that can be used to quickly visualize the results of a BEAST run (.log file) and TreeAnnotator can be used to create the MCC tree of a run (.trees file).

This tutorial is based on a tutorial by Tracy Heath and Tanja Stadler and some parts of it have been copied verbatim. The original tutorial is available at <http://treethinkers.org/tutorials/divergence-time-estimation-using-beast/> and <http://phyloworks.org/workshops/divtime.html>.

The 2014-2015 West African Ebola epidemic

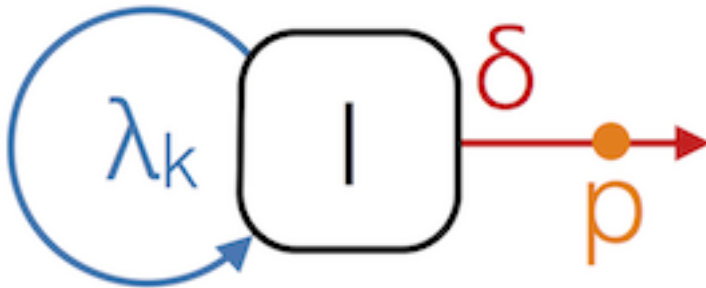
In this tutorial we will analyse sequence data from the 2014-2015 West African Ebola epidemic. The data we will use was first published by [Carroll et al.](#) in June 2015 and in total comprises Ebola genomes sampled from 179 patients in Guinea and Liberia. We extracted the complete coding regions of all the sequences and aligned them using Prank. In this tutorial we will mostly concentrate on only the coding regions of the NP and GP genes from 50 randomly selected patients. The data are available at <https://github.com/cevo-public/2015-GEMP-Phylogenetics/tree/master/TutorialData>.

Patients exposed to Ebola virus first undergo an incubation period of 2-21 days before becoming infectious. Once infectious, patients either die between days 6 and 16 or may begin to recover between days 6 and 11. Thus, an SEIR model is appropriate for an Ebola epidemic.



Because it is computationally expensive to calculate the likelihood, and because the likelihood surface is complex under an SEIR model, we will approximate the epidemic with a simple birth-death skyline model.

This model assumes no incubation period. Additionally, the period of the epidemic is divided into n equally spaced rate shifts, where the effective reproduction number, R_e (or any of the other model parameters) may change. Between rate shifts all model parameters remain constant. This results in a piecewise constant approximation of the epidemic dynamics.

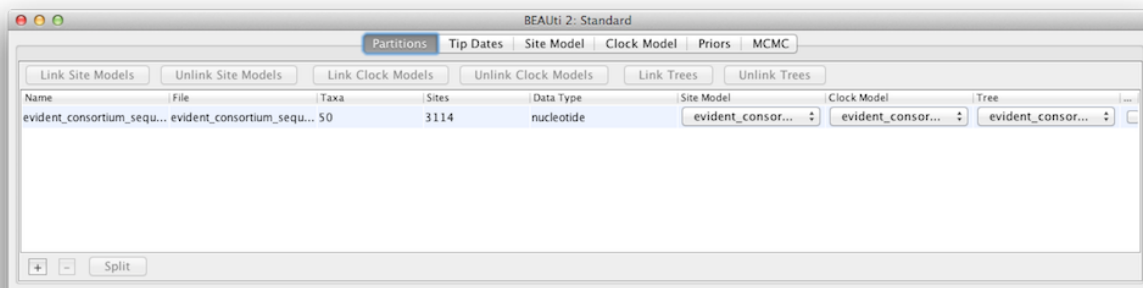


Question 1: A simple BEAST analysis

In the first question we will only use the *NP* and *GP* genes from 50 of the genomes presented by Carroll et al. The data is available at <https://github.com/cevo-public/2015-GEMP-Phylogenetics/tree/master/TutorialData> in the file `evident_consortium_sequences_NPGP_50.fas`.

Importing the sequence data

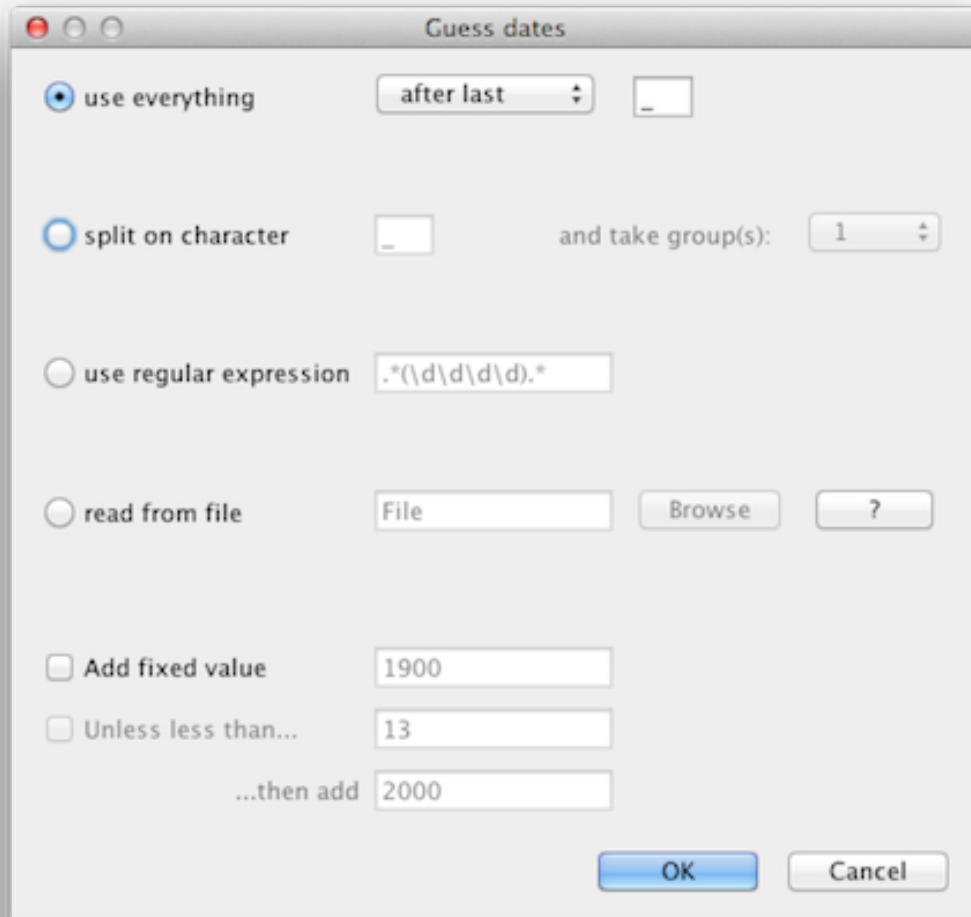
1. Open **BEAUi** and import the sequence from `filename.nex`. To do this go to **File** → **Import Alignment**
2. Inside the **Partitions** panel (open by default in BEAUi) you should see the alignment has 50 sequences with 3114 sites each.



3. Double click on the alignment to look at it and at the sequence identifiers.

Note that every sequence has an identifier followed by the collection date, e.g. `KR817113_22-Jul-2014`. By default BEAUi assumes that all tip dates are at the present, however we want to make use the collection dates of the serially sampled sequences.

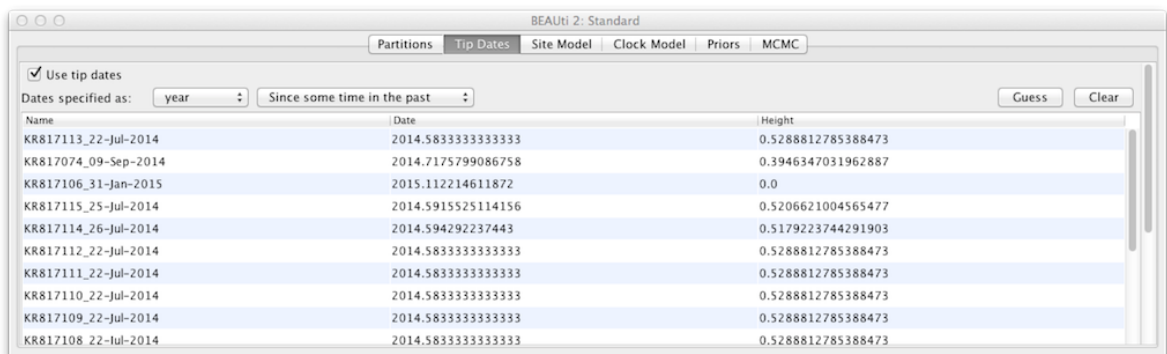
- Click on the **Tip Dates** panel and check the **Use tip dates** box. With the dates specified as **year** and **Since some time in the past** click on **Guess**.
- Beauti will extract the dates from the sequence names, given that they are provided with a specific text pattern. Here the dates all follow the last “_” character. Indicate that you want to **use everything** after this character and press **OK**.



The "Guess dates" dialog box is shown with the following settings:

- ☒ **use everything**: The dropdown is set to "after last" and the character is "_".
- ☐ **split on character**: The character is "_" and "and take group(s)" is set to "1".
- ☐ **use regular expression**: The regular expression is ".*(\d\d\d\d\d).*".
- ☐ **read from file**: The "File" field is empty, with "Browse" and "?" buttons.
- ☐ **Add fixed value**: The value is "1900".
- ☐ **Unless less than...**: The value is "13".
- ...then add**: The value is "2000".

Buttons at the bottom: **OK** and **Cancel**.



The "Tip Dates" panel in BEAUTI 2: Standard is shown with the following settings:

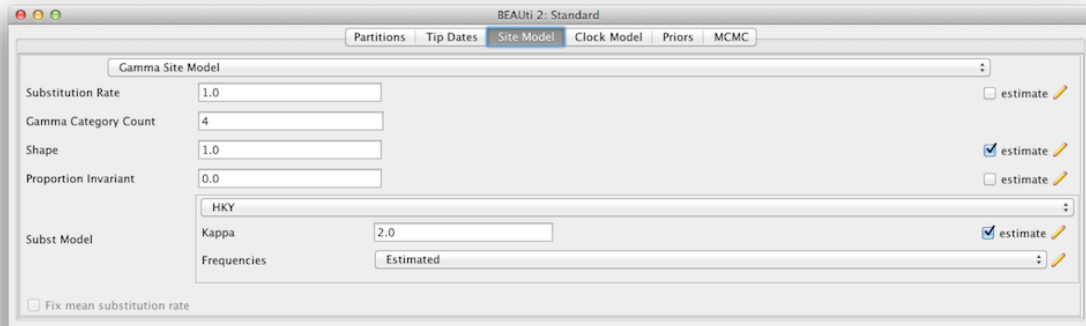
- ☒ **Use tip dates**
- Dates specified as: **year** and **Since some time in the past**
- Guess** and **Clear** buttons

Name	Date	Height
KR817113_22-Jul-2014	2014.583333333333	0.5288812785388473
KR817074_09-Sep-2014	2014.7175799086758	0.3946347031962887
KR817106_31-Jan-2015	2015.112214611872	0.0
KR817115_25-Jul-2014	2014.5915525114156	0.5206621004565477
KR817114_26-Jul-2014	2014.594292237443	0.5179223744291903
KR817112_22-Jul-2014	2014.583333333333	0.5288812785388473
KR817111_22-Jul-2014	2014.583333333333	0.5288812785388473
KR817110_22-Jul-2014	2014.583333333333	0.5288812785388473
KR817109_22-Jul-2014	2014.583333333333	0.5288812785388473
KR817108_22-Jul-2014	2014.583333333333	0.5288812785388473

You can now see that each sequence is given a date that is a value relative to year 0. Thus, the sequence sampled on **9 September 2014** is **2014.7175799086758**. 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 $1/365 = 0.002739726$.

Setting the sequence and clock models

1. Click on the **Site model** panel and change the model to **HKY+Gamma**, with **4** rate categories. Note that we are only estimating the Gamma shape parameter and Kappa.

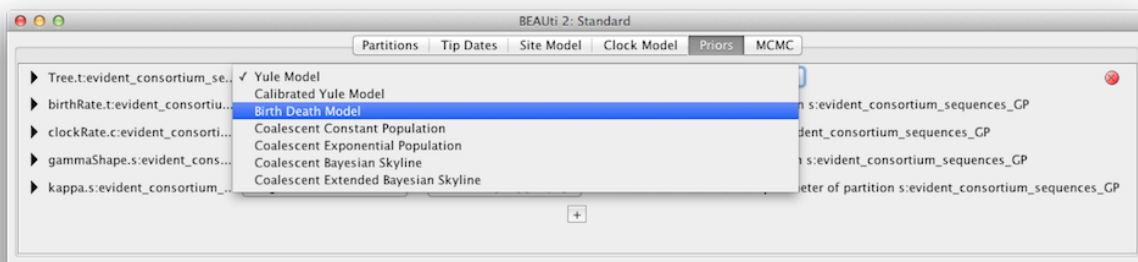


2. Click on the **Clock model** panel. For this analysis we will leave the default **Strict clock** and estimate the **Clock rate**. Thus, nothing needs to be changed here.

Setting the priors

We want to use the Birth-Death Skyline model as a tree-prior. The model is part of the **BDSKY** package, however it is not installed by default.

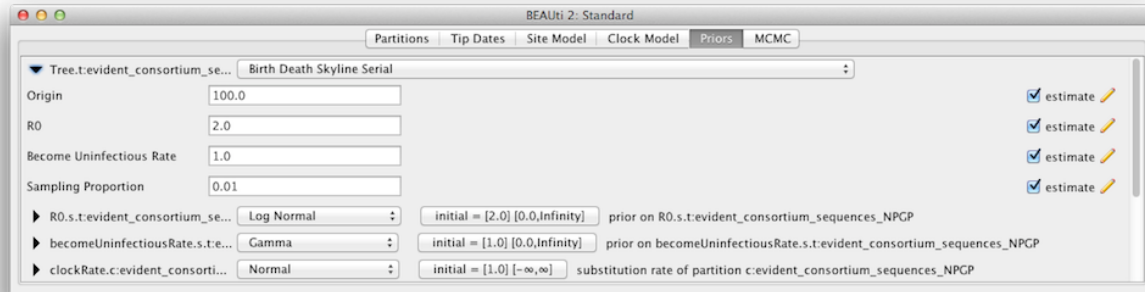
1. Click on the **Priors** panel and look at the available tree-priors to verify that it is not there.



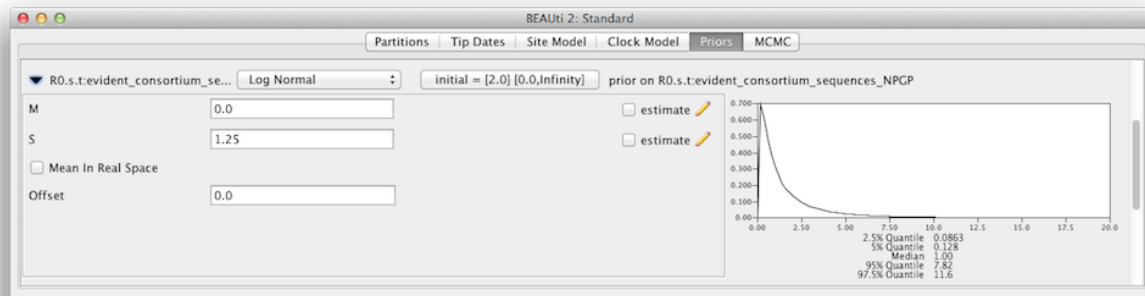
2. After you have installed the package you have to restart **Beauti** before it is available. Therefore it is best to save your configuration now and then load it again once you have installed BDSKY. You can save your configuration by going to **File** → **Save As**.
3. To install the **BDSKY** package open the package manager by going to **File** → **Manage Packages**.
4. Once you have installed **BDSKY** load the configuration file you saved in step 2 and go to the **Priors** panel.

5. **Birth Death Skyline Serial** should now be available as a tree-prior.

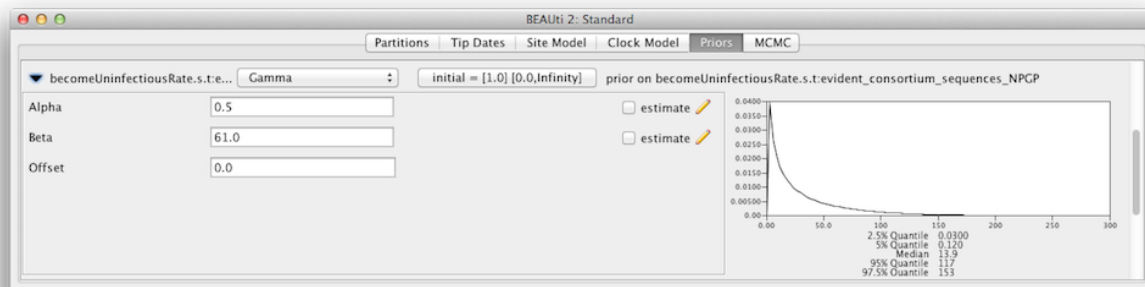
For the analysis we will use a Birth Death Skyline Serial tree-prior where we will estimate all 4 parameters, **Origin**, R_0 , δ (become uninfected rate) and p (sampling proportion).



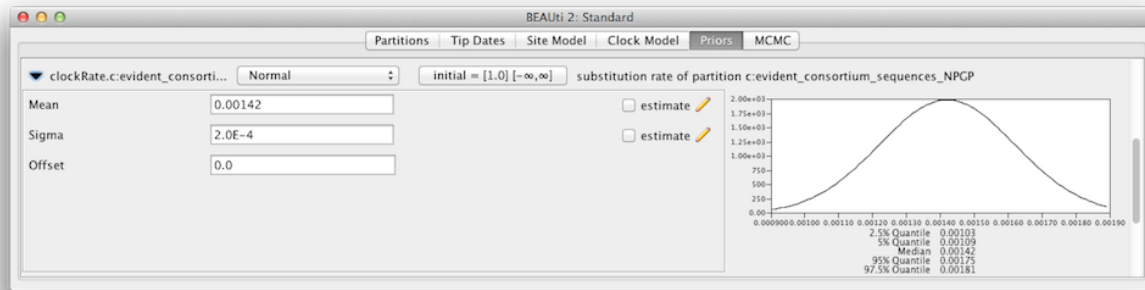
6. Set the prior on the effective reproductive rate to a lognormal with a log-mean of **0.0** and standard deviation of **1.25**.



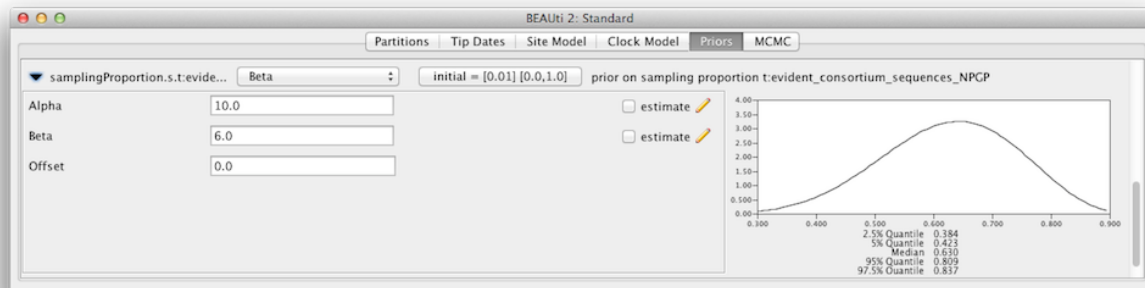
7. Specify a Gamma prior on the rate of becoming non-infectious with a shape of **0.5** and a scale of **61**.



8. Set the prior on the clock rate to a normal distribution with a mean of **0.00142** and a standard deviation of **2E-4**.



9. Specify a Beta prior on the sampling proportion with parameters $\alpha = 10.0$ and $\beta = 6.0$.



10. Leave the rest of the priors (Gamma shape parameter, Origin and Kappa) on their default values.

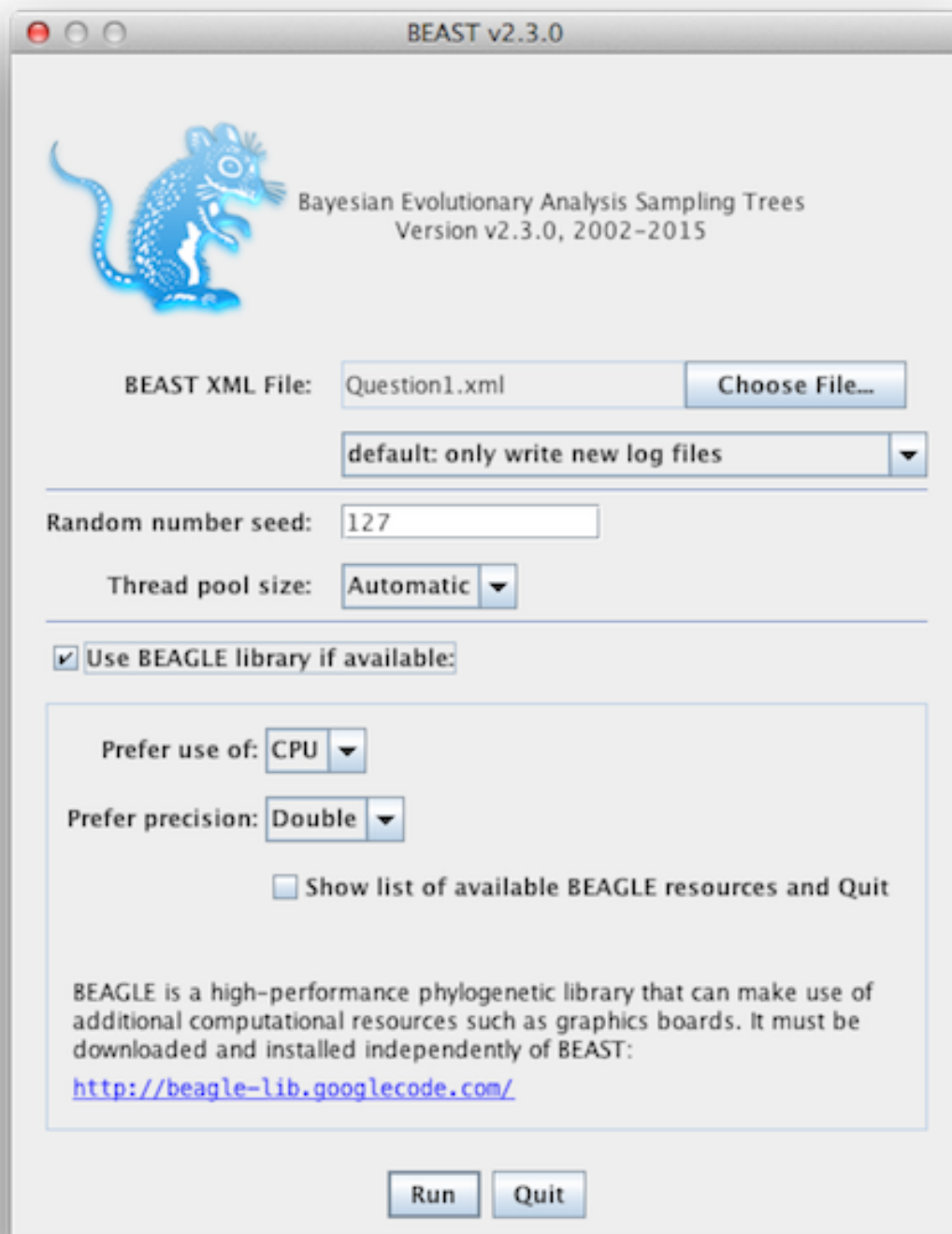
Note that we are not estimating any hyperparameters (parameters of the priors) of the tree-prior.

By default BDSKY assumes 10 equally spaced rate shifts in R_e over time and no rate shifts in δ (becoming uninfected rate) and s (sampling proportion).

11. To change the number of rate shifts we need to go to the **Initialization** panel, which is not displayed by default. To display it go to **View** → **Show Initialization panel**.
12. Change the number of rate shifts for R_e to **5** by changing its dimension.
13. Verify that become uninfected rate (δ) and the sampling proportion (s) both have a dimension of **1**.

Running the model

1. Navigate to the MCMC panel. Set the chain length to **10,000,000**
2. Set the sampling frequency (**Log Every**) to **1,000** for both **tracelog** and **treeelog** and **10,000** for **screenlog**.
3. Now save your configuration file and open **BEAST 2**.
4. Load the configuration file in **BEAST 2**. If you want your analysis to be repeatable set the random seed to some number and write it down. (In the part that follows we used a random seed of **127**)
5. Run the analysis! (Depending on how fast your machine is this may take a while, so now is a good time to have a coffee break).



Analysing the results

Once the run has finished, open the log file in **Tracer**.

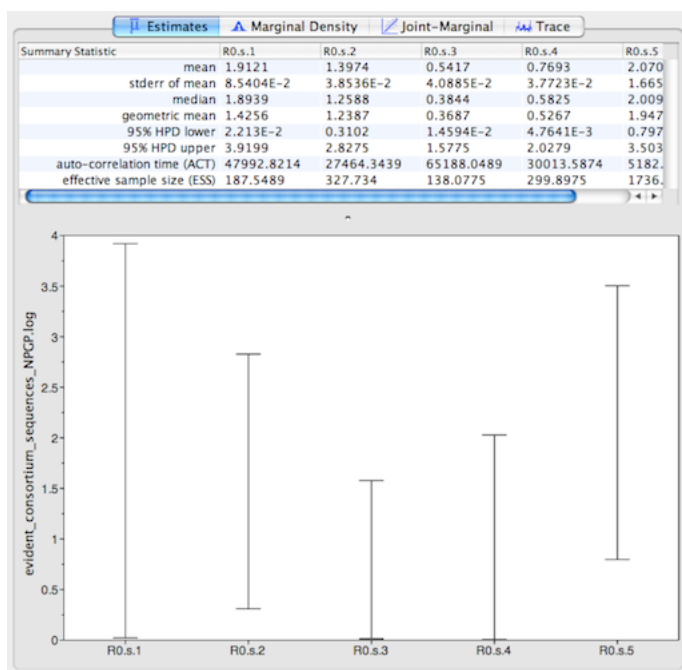
1. Look at the traces, ESSs and summaries of each parameter. Do you think the run mixed well?
2. Look at the estimate of the origin time. The origin is the time of infection of the first person in the outbreak. Below is a histogram of the origin time from an analysis of these data. The mean estimated time is **0.7316**. 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 **267** days before the most recently sampled sequence:

$$\frac{0.7316}{\frac{1}{365}} = 267.034$$

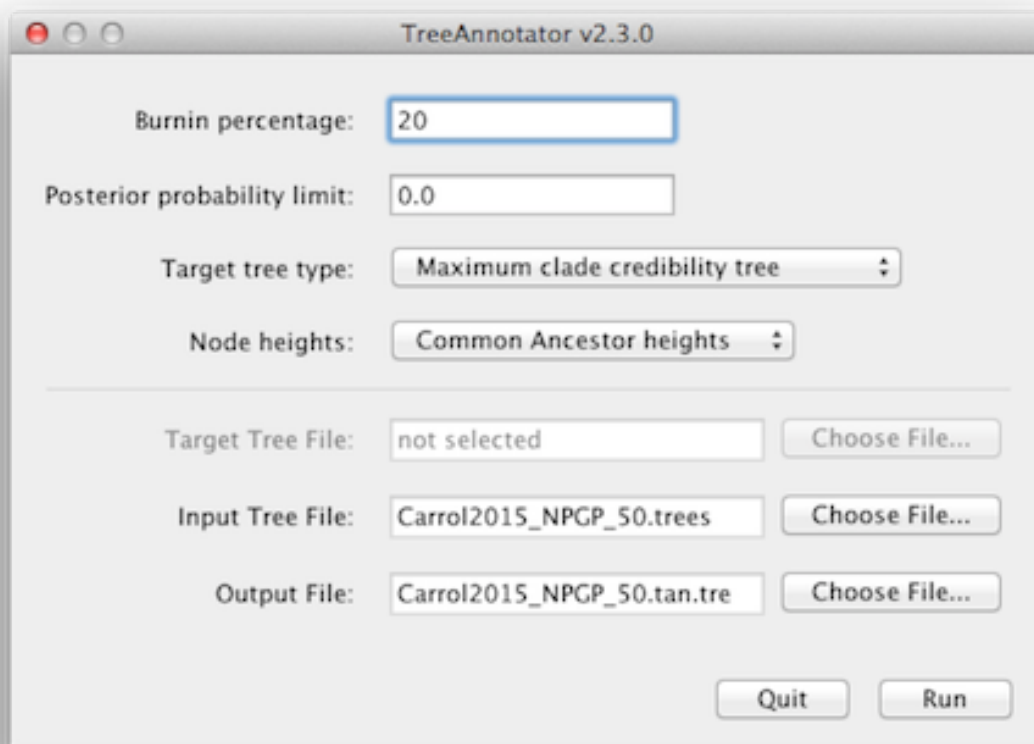
![] (Figures/Origin.png)

The last sequence was sampled on 31 January 2015. Thus, the mean origin time corresponds to the 129th day of 2014, which corresponds to 9 May, with a 95% highest posterior density interval between 7 January and 23 May 2015.

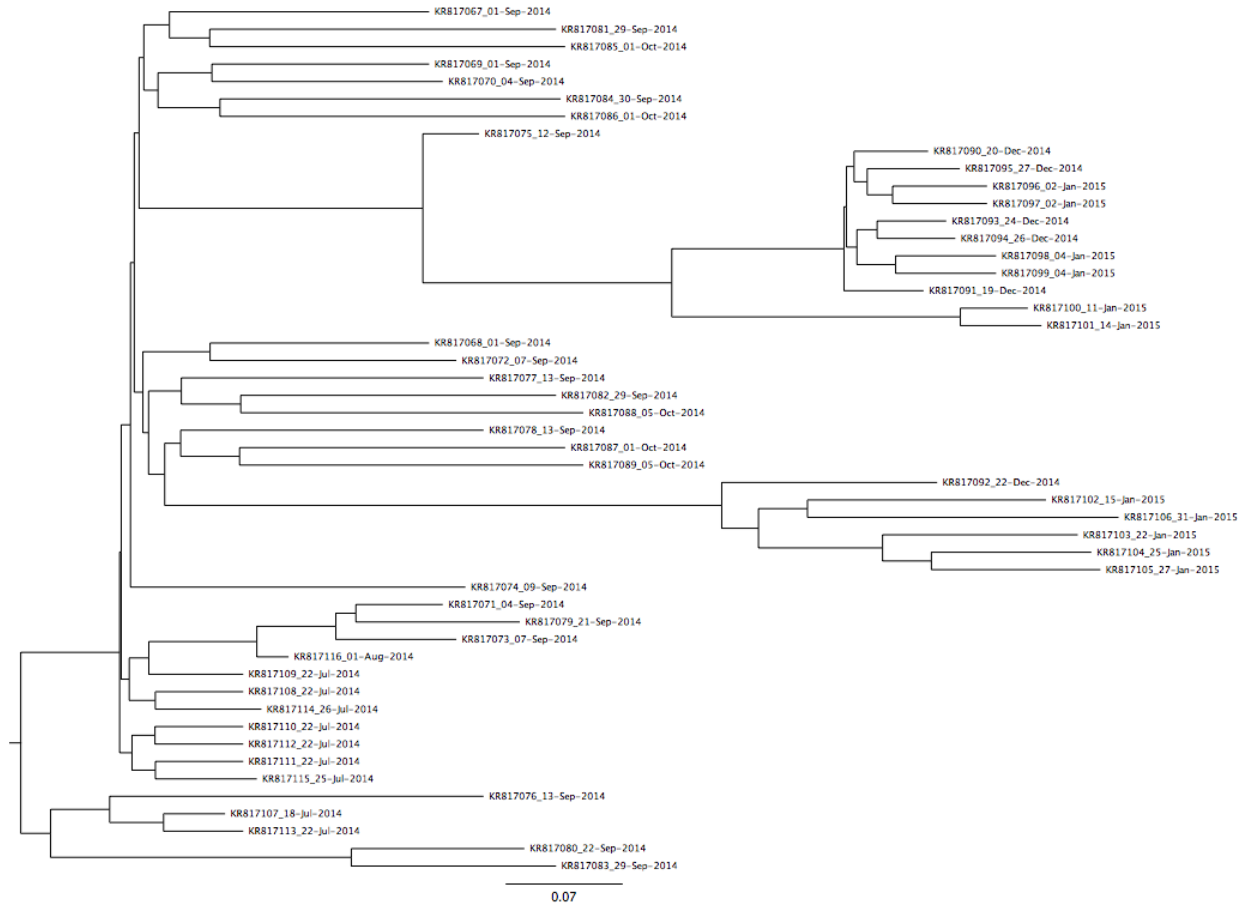
3. You can visualize the change in the effective reproductive number (R_e) over time by selecting **** **** in the **Estimates** panel.



4. Summarize the posterior sample of trees using **TreeAnnotator**.



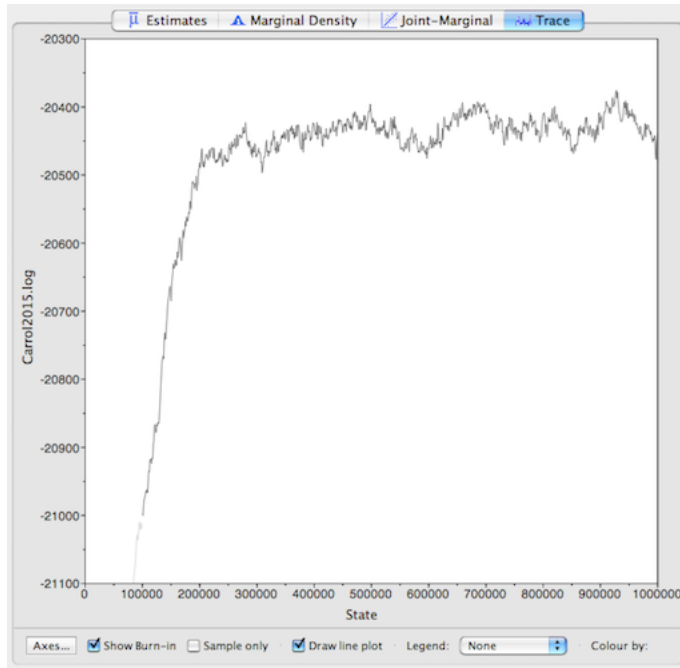
The maximum clade credibility tree for the sequences is shown below:



Question 2: Mixing it up

In this question we will use the full dataset of 179 sequences and the coding sequences of all 6 genes (13,419 sites in the alignment). The data is available at <https://github.com/cevo-public/2015-GEMP-Phylogenetics/tree/master/TutorialData> in the file `evident_consortium_sequences_NPGP_50.fas`.

1. Create a **BEAUi** configuration file for this dataset using the same parameter settings as in Question 1.
2. Change the chain length to **1,000,000** so the run does not take forever!
3. Run the analysis in **BEAST** (have another coffee) and analyse the results in **Tracer**.
4. Did the run mix? Would increasing the chain length help?



Question 3: Of Priors and Posteriors (optional)

If the priors used in the analysis are too strong or if the data do not contain enough information then the posterior distribution will resemble the prior distribution. Evaluate if this is the case by plotting the prior distributions for parameters over the histograms of their posterior distributions. If you are using **R** you can import the **BEAST** log files using the `read.table` command.