Tutorial using BEAST v2.5.x

Inference of population dynamics

Julia Palacios

This tutorial is adapted from taming-the-beast.org and from beast-community.

1 Background

This tutorial provides a step-by-step illustration on how to use BEAST to reconstruct the evolutionary dynamics of hepatitis C virus in Egypt (Example 1) and Influenza H3N2 in the New York State (Example 2). For example 1, we will first place a Skyline prior on Ne(t) with 4 change points and a coalescent prior on the isochronous genealogy (Drummond et al. 2005). In this case, mutation rate and effective population size are not identifiable because there is no time sampling heterogeneity. Hence we will fix the clock rate at a known value and infer Ne(t). We will then generate the MCCT tree from the posterior run and do a Bayesian nonparametric estimation of the effective population size from a fixed tree (Palacios and Minin 2013). The third method places a Birth-Death prior on the genealogy and a Skyline prior on the reproductive number (Stadler et al. 2013) based on the birth-death model. Finally, we will analyze human influenza A H3N2 virus in New York.

2 Programs used in this Exercise

2.0.1 BEAST2 - Bayesian Evolutionary Analysis Sampling Trees 2 (http://www.beast2.org)

2.0.2 BEAUti - Bayesian Evolutionary Analysis Utility

2.0.3 TreeAnnotator

2.0.4 Tracer (http://beast.community/tracer)

2.0.5 R / RStudio packages phylodyn and bdskytools R

.

3 Practical 1: Bayesian and Birth-Death Skyline Plots of HCV in Egypt

In this tutorial we will estimate the dynamics of the Egyptian Hepatitis C epidemic from genetic sequence data collected in 1993.

The aims of this tutorial are to:

- Learn how to infer population dynamics.
- Get to know how to choose the set-up of a skyline analysis.
- Get to know how to set-up a birth-death skyline analysis.

3.1 The Data

The dataset consists of an alignment of 63 Hepatitis C sequences sampled in 1993 in Egypt (Ray et al. 2000). This dataset has been used multiple times to test the performance of developed methods (Drummond et al. 2005; Stadler et al. 2013)

With an estimated 15-25%, Egypt has the highest Hepatits C prevalence in the world. In the mid 20^{th} century, the prevalence of Hepatitis C increased drastically. We will try to infer this increase from sequence data.

The alignment file is hcv.nexus in the Data folder of this tutorial.

3.2 Setting up the Coalescent Bayesian Skyline analysis

To start we have to import the alignment into BEAUti.

> In the **Partitions** panel, import the nexus file with the alignment by navigating to **File > Import Alignment** in the menu and then finding the file on your computer **or** simply drag and drop the file into the **BEAUti** window.

BEAUti will recognize the sequences from the file as nucleotide data. It will do so for sequence files with the character set of **A | C | G | T | N**, where **N** indicates an unknown nucleotide. As soon as other non-gap characters are included (e.g. using **R** or **Y** to indicate purines and pyramidines) BEAUti will not recognize the data as nucleotides anymore (unless the type of data is specified in the file) and open a dialogue box to confirm the data type.

> Skip the **Tip Dates** panel and navigate to the **Site Model** panel.

The next step is to specify the model of nucleotide evolution (the site model). We will be using the HKY model with estimated frequencies and we will fix the clock rate to a previously estimated value (Pybus et al. 2001). In this case, all the samples are contemporaneous (sampled at the same time) and the clock rate is simply a scaling of the estimated tree branch lengths (in substitutions/site) into calendar time.

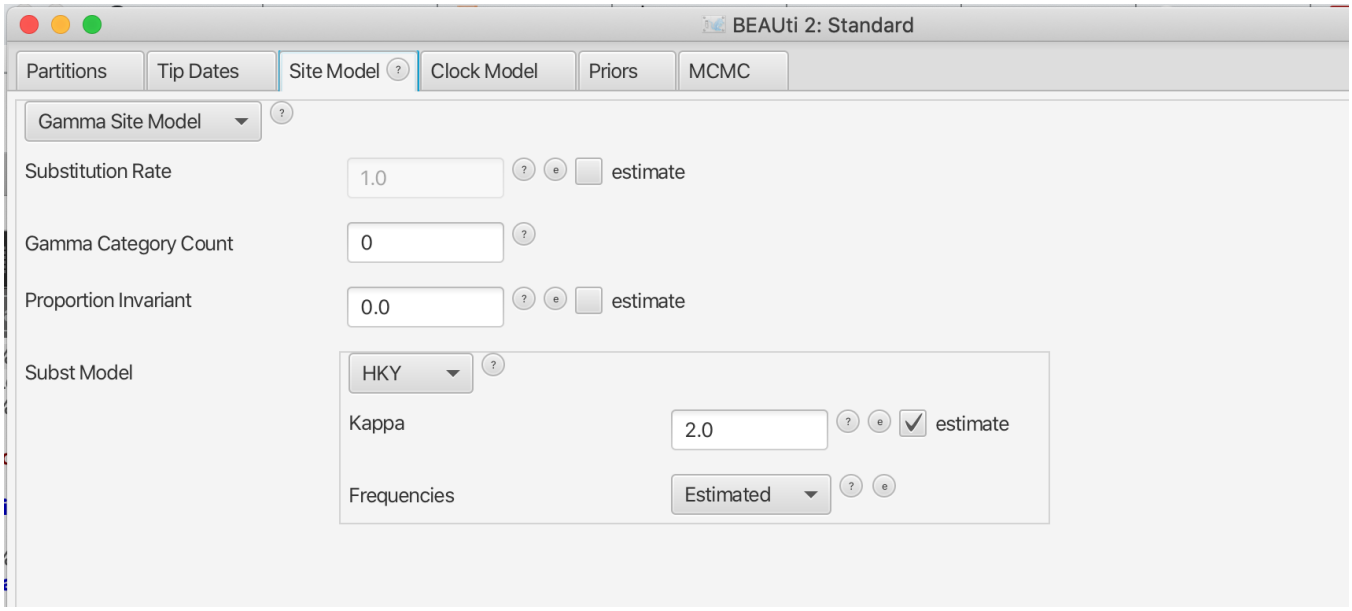

Figure 1: Set the HKY substitutions model.

Navigate to the **Clock Model** panel.

Leave the clock model as a **Strict Clock** and set **Clock.rate** to 0.00079 s/s/y (Figure 2). (Note that BEAUti is smart enough to know that the clock rate cannot be estimated on this dataset and grays out the estimate checkbox).

Now we are ready to set up the Coalescent Bayesian Skyline as a tree-prior.

Navigate to the **Priors** panel and select **Coalescent Bayesian Skyline** as the tree prior (Figure 18).

The Coalescent Bayesian Skyline divides the time between the present and the root of the tree (the tMRCA) into segments, and estimates a piece-wise constant effective population size $N_e(t)$ for each segment. The endpoints of segments are tied to the branching times (also called coalescent events) in the tree, and the size of segments is measured in the number of coalescent events included in each segment. The Coalescent Bayesian Skyline groups coalescent events into segments and jointly estimates the $N_e(t)$ (**bPopSizes** parameter in BEAST) and the size (**bGroupSizes** parameter) of each segment. To set the number of segments we have to change the dimension of **bPopSizes** and **bGroupSizes** (note that the dimension of both parameters has to be the same). Note that the length of a segment is not fixed, but dependent on the timing of coalescent events in the tree, as well as the number of events contained within a segment (**bGroupSizes**).

To change the number of segments we have to navigate to the **Initialialization** panel, which is by default not visible. Navigate to **View > Show Initialization Panel** to make it visible and navigate to it (Figure 4).

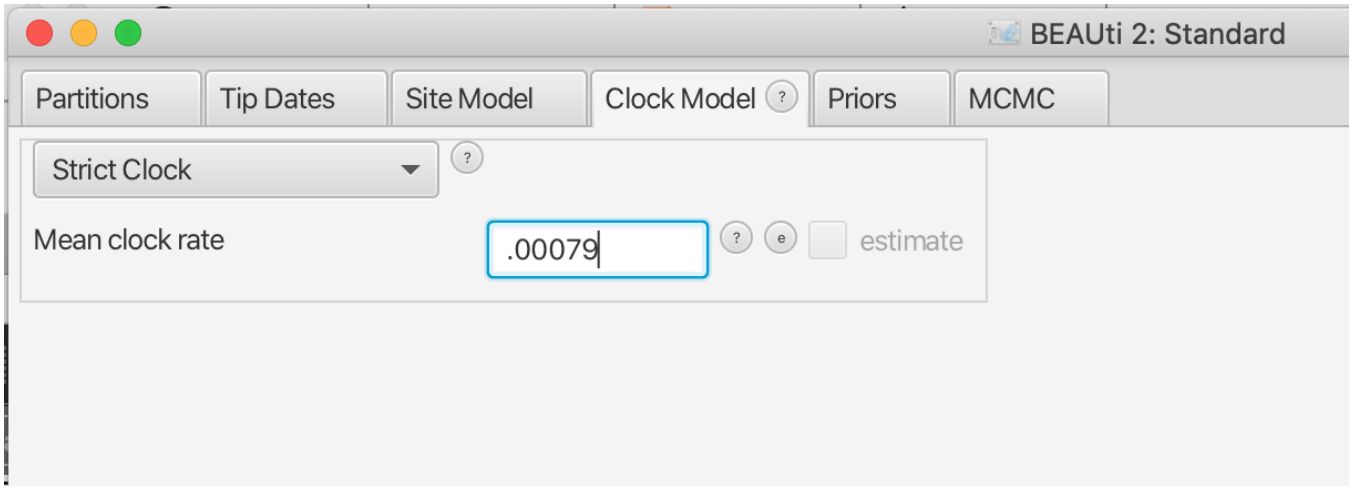

Figure 2: Set the clock rate to 0.00079 s/s/y.

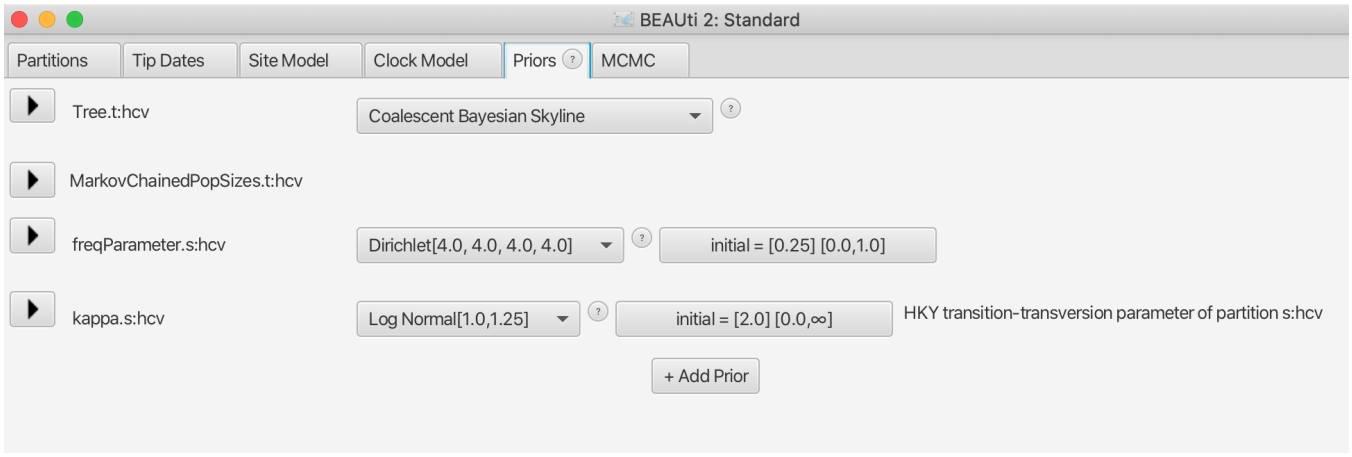

Figure 3: Choose the Coalescent Bayesian Skyline as a tree prior.

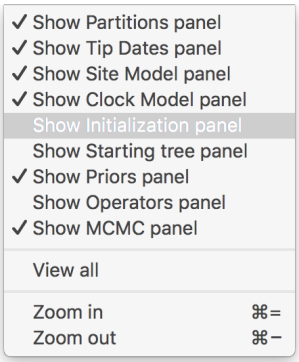

Figure 4: Show the initialization panel.

Set the dimension of **bPopSizes** and **bGroupSizes** to 4 (the default value is 5) (Figure 5).

This sets the number of segments equal to 4 (the parameter dimension), which means $N_e(t)$ will be allowed to change 3 times between the tMRCA and the present (if we have d segments, N_e is allowed to change $d-1$ times).

We can leave the rest of the priors as they are and save the XML file. We want to shorten the chain length and decrease the sampling frequency so the analysis completes in a reasonable time and the output files stay small. (Keep in mind that it will be necessary to run a longer chain for parameters to mix properly).

Navigate to the **MCMC** panel.

Change the **Chain Length** from 10'000'000 to 3'000'000.

Click on the arrow next to the **tracelog** and change the **File Name** and set the **Log Every** to 3'000. To avoid confusion use the same name for the log and trees files

Click on the arrow next to the **treelog** and change the **File Name** and set the **Log Every** to 3'000.

Leave all other settings at their default values and save the file.

Now we are ready to run the analysis.

Start **BEAST2** and choose the file `hcv.xml`.

If you have **BEAGLE** installed tick the box to **Use BEAGLE library if available**, which will make the run faster.

Hit **Run** to start the analysis.

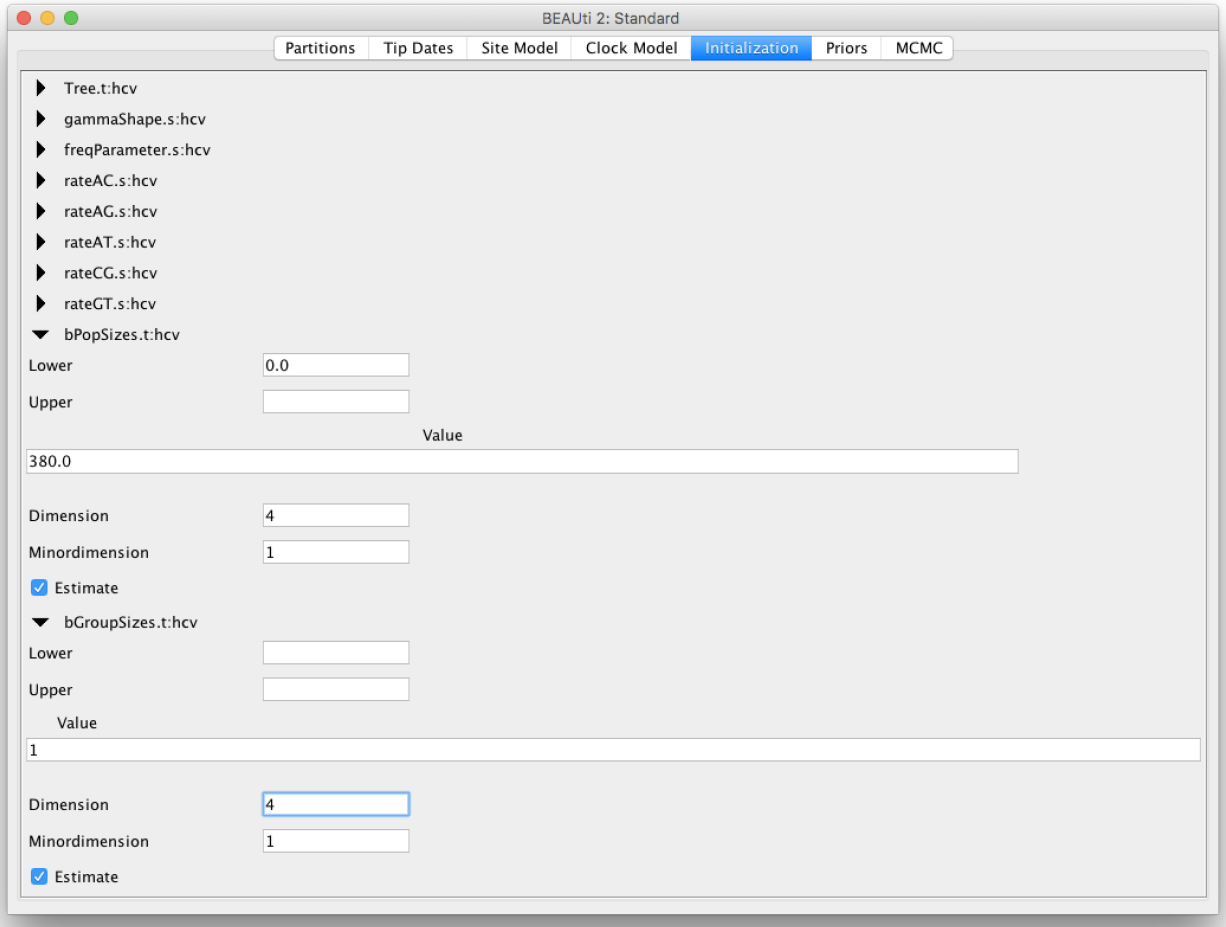

Figure 5: Set the dimension of bPopSizes and bGroupSizes to 4.

The analysis will take about 10 minutes to complete. Read through the next section while waiting for your results or start preparing the XML file for the birth-death skyline analysis.

3.2.1 The Coalescent Bayesian Skyline parameterization

In the standard coalescent, the effective population size ($N_e(t)$) is the inverse of the rate at which lineages merge λ. The larger $N_e(t)$ is the less likely lineages are to coalesce. Thus, intervals in a sampled tree with many branching events often coincide with periods when the population size was small. Similarly, few coalescent events occur during periods of large population size. (Note that these results are conditioned on sampling only a small fraction of the population).

$$\lambda = \frac{1}{N_e} \tag{1}$$

The Coalescent Bayesian Skyline model allows $N_e(t)$ to change over time as a piece-wise constant trajetory with specified number of change points. Choosing the dimension for the Bayesian Skyline can be rather arbitrary. If the dimension is chosen too low, not all population size changes are captured, but if it is chosen too large, there may be too little information in a segment to support a robust estimate. In the extended Bayesian skyline model, the number of change points is not specified but jointly inferred with model parameters. More flexible alternatives are implemented in BEAST such as nonparametric models in which the number of parameters grows with the number of samples such as Bayesian Skyride (Minin et al. 2008) and other Gaussian Process based methods (Palacios and Minin 2013).

3.2.2 Exploring the results of the Coalescent Bayesian Skyline analysis

For the reconstruction of the population dynamics, we need two files, the log file and the trees file. The log file contains the information about the group sizes and population sizes of each segment, while the trees file is needed for the times of the coalescent events.

Load the logfile into **Tracer** to check mixing and parameter estimates (Figure 19).

Because we shortened the chain most parameters have very low ESS values. If you like, you can compare your results with the example results we obtained with identical settings and a chain of 30,000,000 (`hcv_coal_30M.log`).

Navigate to **Analysis > Bayesian Skyline Reconstruction**. From there open the trees file. To get the correct dates in the analysis we should specify the **Age of the youngest tip**. In our case it is 1993, the year where all the samples were taken. If the sequences were sampled at different times (heterochronous data), the age of the youngest tip is the time when the most recent sample was collected.

Press **OK** to reconstruct the past population dynamics (Figure 7).

The output will have the years on the x-axis and the effective population size on the y-axis. By default, the y-axis is on a log-scale. If everything worked as it is supposed to work you will see a sharp increase in the effective population size in the 1920s.

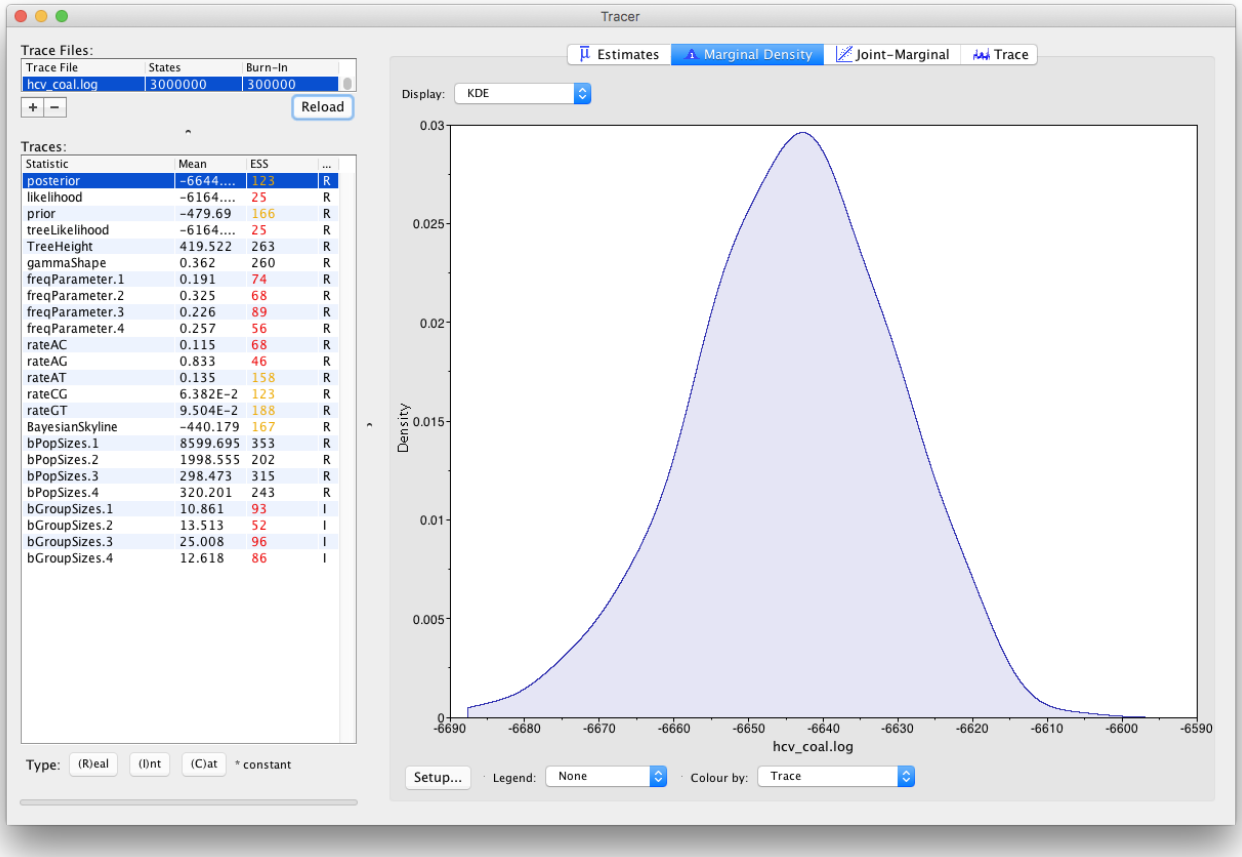

Figure 6: Loading the log file into Tracer.

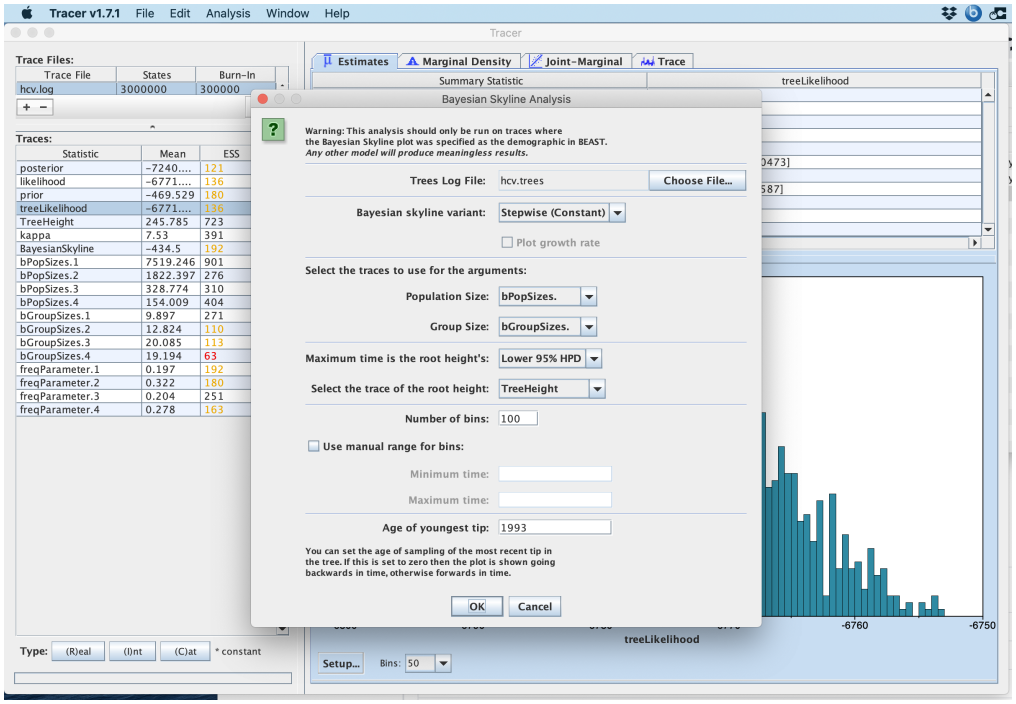

Figure 7: Reconstructing the Bayesian Skyline plot in Tracer.

(Note that the reconstruction will only work if the log and trees files contain the same number of states and both files were logged at the same frequency).

There are two ways to save the analysis, it can either be saved as a pdf for display purposes or as a tab delimited file.

> Navigate to **File > Export Data Table**.
>
> Enter the filename and save the file.

The exported file will have five rows, the time, the mean, median, lower and upper boundary of the 95% BCI interval of the estimates, which you can use to plot the data with other software (R, Matlab, etc).

3.2.3 Exercise 1:

From your trees file, use TreeAnnotator to obtain the Maximum Clade Credibility Tree. You can visualize your tree and estimate Ne(t) using a non-parametric Gaussian processes based method implemented in R package `phylodyn` (Karcher et al. 2017). Some packages need to be installed directly over GitHub. The R markdown phylodyn.Rmd and pdf files are available in the folder's tutorial.

3.3 Setting up the Birth-Death Skyline analysis

3.3.1 Install BEAST 2 packages

While the coalescent-based Bayesian Skyline Plot is integrated in the BEAST2 core, we need to install the BDSKY package, which contains the Birth-Death Skyline model. Installation of packages is done using

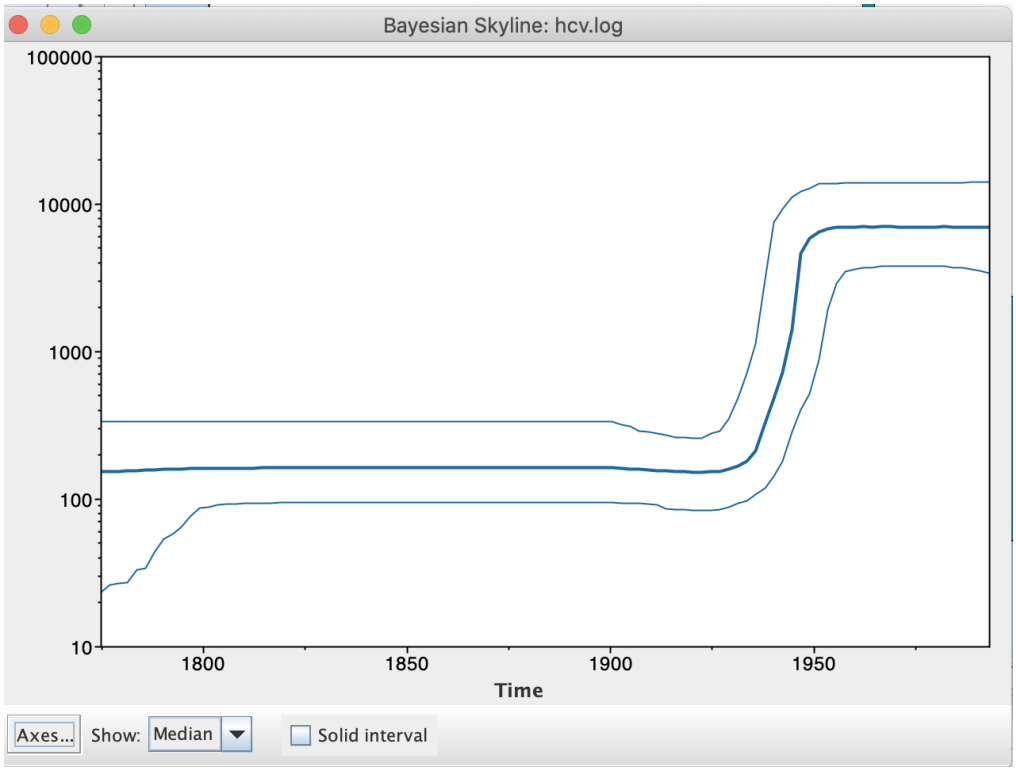

Figure 8: Coalescent Bayesian Skyline analysis output. The black line is the posterior median of the effective population size (can be changed to the mean). The two blue lines are the upper and lower bounds of the 95% BCI. The x-axis is the time in years and the y-axis is on a log-scale.

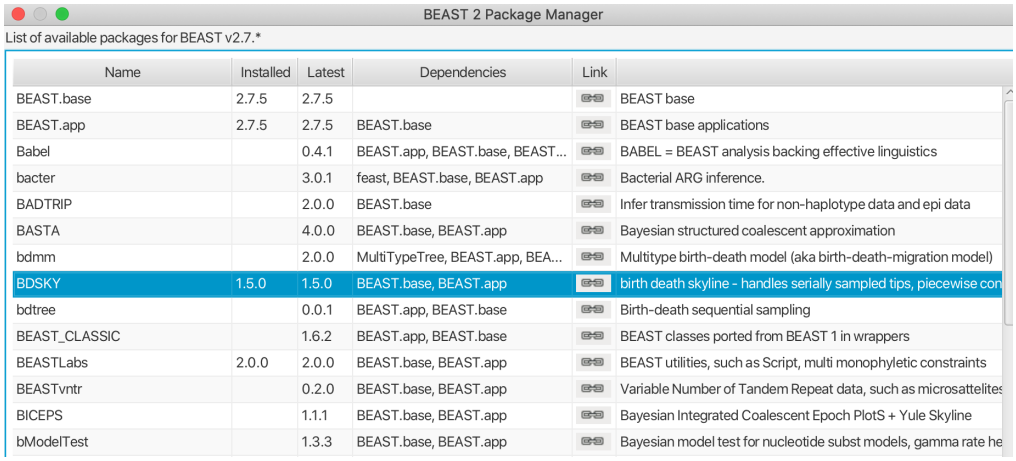

Figure 9: Install the BDSKY package which contains the Birth-Death Skyline model.

the package manager, which is integrated into BEAUti.

Open the **BEAST2 Package Manager** by navigating to **File > Manage Packages**.

Install the **BDSKY** package by selecting it and clicking the **Install/Upgrade** button (Figure 9).

After the installation of a package, the program is on your computer, but BEAUti is unable to load the template files for the newly installed model unless it is restarted. So, let's restart BEAUti to make sure we have the **BDSKY** model at hand.

Close the **BEAST2 Package Manager** and ***restart*** BEAUti to fully load the **BDSKY** package.

In the first analysis, we used the coalescent approach to estimate population dynamics. We now want to repeat the analysis using the Birth-Death Skyline model. We will use the same model setup as in the previous analysis and only change the tree prior.

Restart **BEAUti**, load `hcv.nexus` as before and set up the same site and clock model as in the Coalescent Bayesian Skyline analysis.

We will need to set the prior to **Birth Death Skyline Contemporary**, since the sequences were all sampled at the same point in time. For heterochronous data (sequences sampled at different times), we would use **Birth Death Skyline Serial**. As with the Coalescent Bayesian Skyline, we need to set the number of dimensions. Here we set the dimension for R_e, the effective reproduction number, which denotes the average number of secondary infections caused by an infected person at a given time during the epidemic, i.e. an R_e of 2 would mean that every infected person causes two new infections on average. In other words, an R_e above 1 means that the number of cases are increasing, therefore the disease will cause an exponentially growing epidemic, and an R_e below 1 means that the epidemic will die out.

Navigate to the **Priors** panel and select **Birth Death Skyline Contemporary** as the tree prior (Figure 10).

Then, click on the button where it says **initial = [2.0] [0.0, Infinity]** next to **reproductiveNumber**. A pop-up window will open which allows us to change the dimension of the parameter (Figure 11). In this case we will keep the default dimension of 10.

Press **OK** to close the pop-up window.

This means that R_e will be allowed to change at 9 times equally spaced between the origin of the epidemic and the present time. Choosing this dimension can again be arbitrary and may require the testing of a few different values. Too few intervals and not all rate shifts are captured. Too many intervals and the intervals may not contain enough information to infer parameters. (As with setting the dimension of the Coalescent Bayesian Skyline the dimension of $R_e$can also be set in the initialization panel).

Besides R_e (**reproductiveNumber**), the **Birth Death Skyline Contemporary** model has 3 more parameters, **becomeUninfectiousRate** (the rate at which infected patients become uninfectious, δ, through recovery, death or isolation), **rho** (the proportion of lineages sampled in the present, ρ) and the **origin** (the time at which the index case became infected, which is always earlier than the tMRCA of the tree).

We will use a lognormal prior for R_e. If an epidemic is neither growing or declining, it has an R_e of 1, which we will use as a null hypothesis, by setting a prior on R_e centered around 1 (we assume that if there isn't a strong signal in an interval for an epidemic to grow or decline that R_e, i.e. the epidemic size stays constant). Note that this prior is used for each of the R_e intervals (the Birth-Death Skyline assumes that R_e is independent in each of the intervals).

Select a **Log Normal** distribution for the **reproductiveNumber** prior.

Click on the arrow to the left of **reproductiveNumber** to open all the options for R_e settings

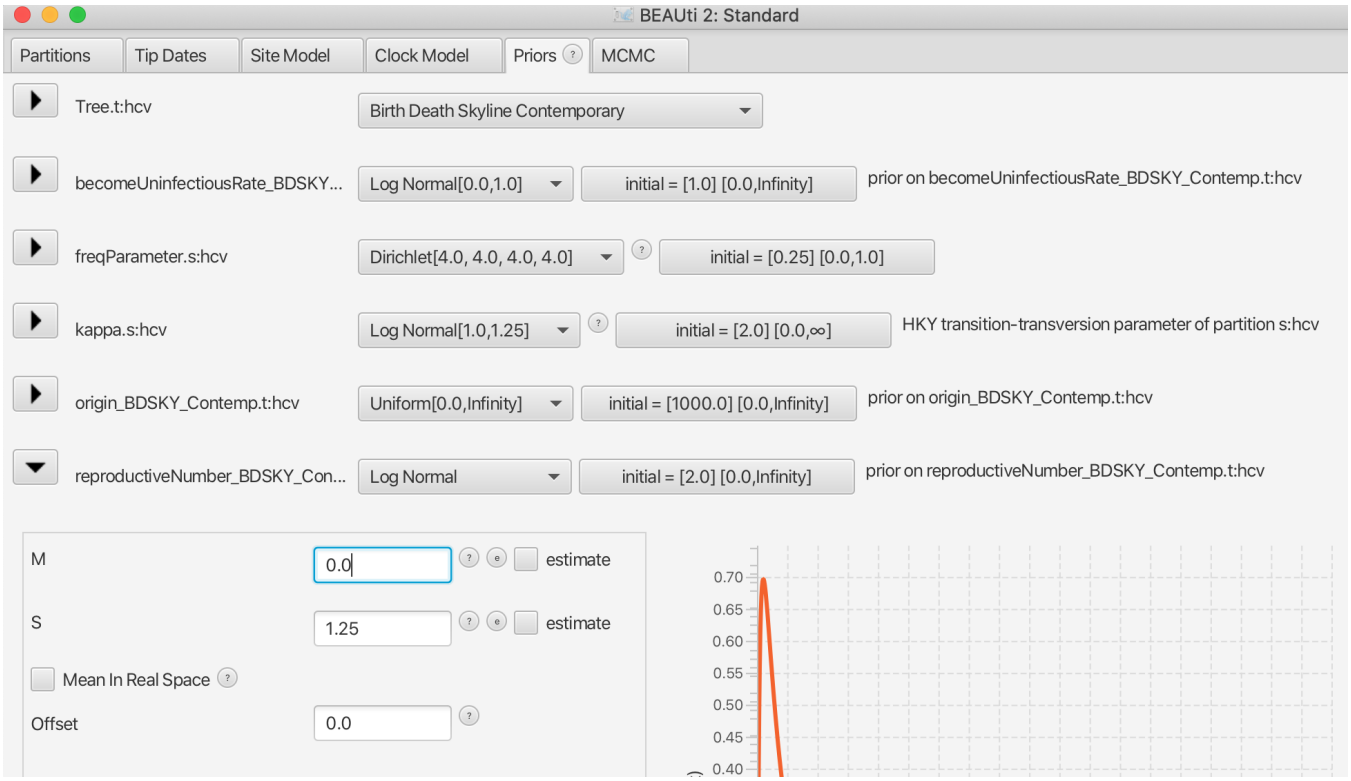

Figure 10: Setting the prior on the tree to the Birth-Death Skyline.

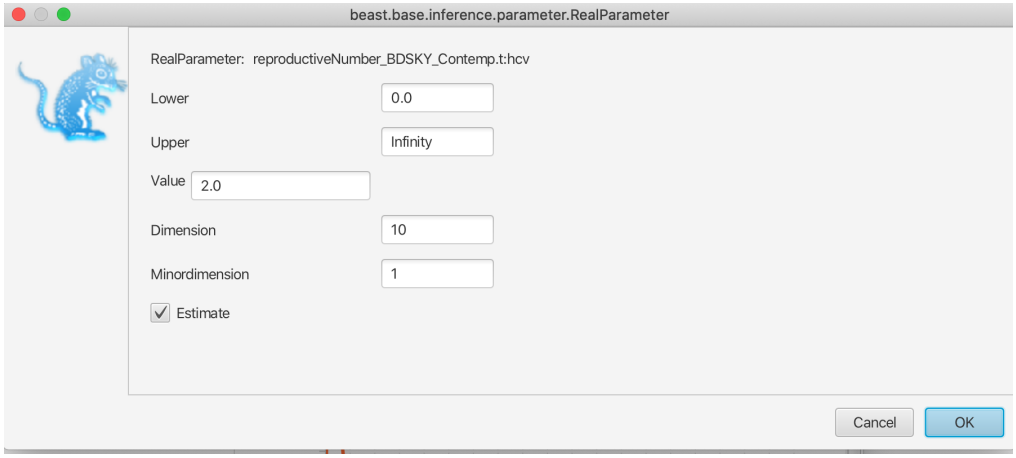

Figure 11: Setting the dimension of the reproductiveNumber parameter.

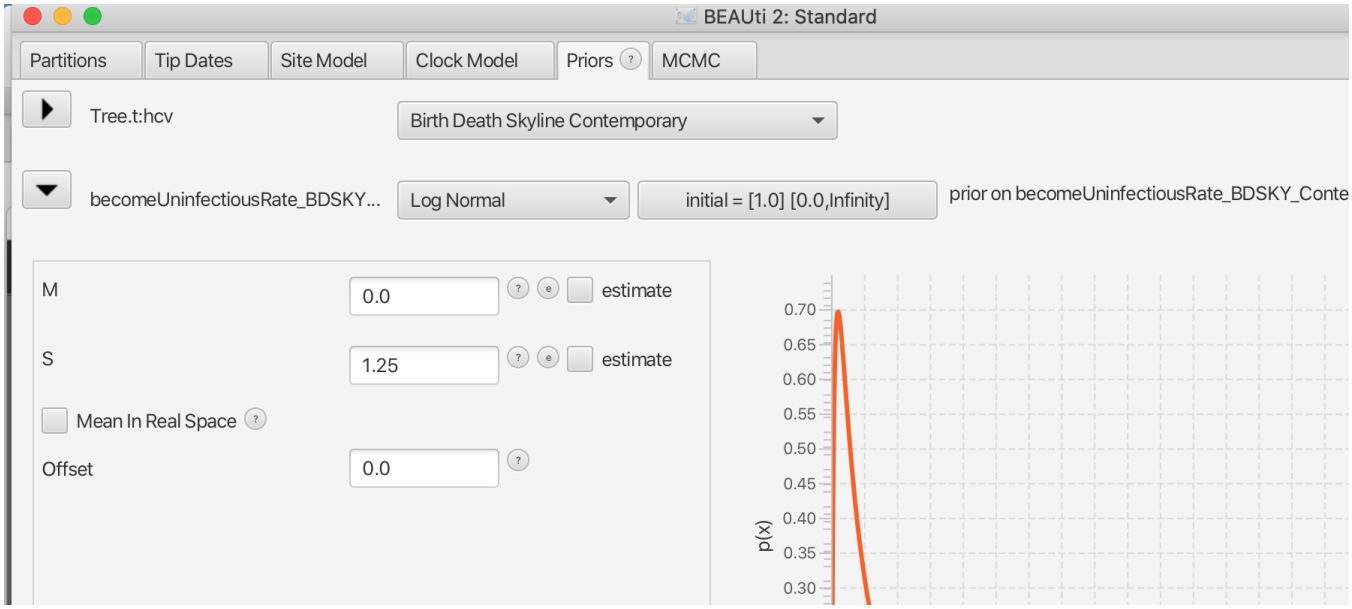

Figure 12: Setting the becoming uninfectious rate prior.

Set **M** to 0, which results in a median of 1. We set the variance to 1.25, which places most weight below 7.82 (95% quantile). (Figure 10).

For the becoming uninfectious rate we will again use a log normal prior. The inverse of the becoming uninfectious rate is the average infectious period. In some patients an HCV infection only lasts a few weeks, while in others it is a chronic infection lasting for many years. Setting $M = 0$ and $S = 1.25$ results in the same prior we used for the R_e. In terms of the becoming uninfectious rate, this translates to the 95% quantiles for the infectious period falling between 0.0862 years (31.5 days) and 11.59 years, with a median of 1 year. We will see later that there is a strong signal in the data for a longer becoming uninfectious period.

Set the same prior for **becomeUninfectiousRate** as for **reproductiveNumber** (Log Normal, with M=0.0, S=1.25) (Figure 12)

The sampling proportion, ρ, represents the proportion of HCV cases in Egypt in 1993 that are included in the analysis. In 1993 Egypt had a population of roughly 60 million people, and with a prevalence of at least 15% this translates into millions of cases, while we only have 63 sequences.

We will use a beta distribution for the prior on ρ. Beta distributions are a very flexible class of distributions that are only defined between 0 and 1, making them ideal to use for proportions.

Select a **Beta** distribution for the **rho** prior.

Click on the arrow to the left of **rho** to open all the options for the prior settings.

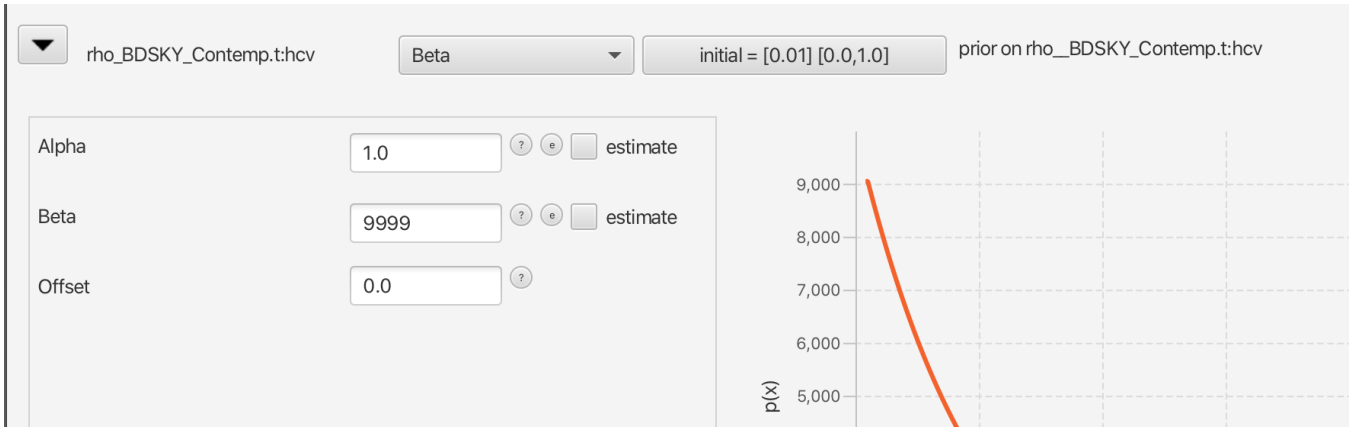

Figure 13: Setting the prior on 'ρ'.

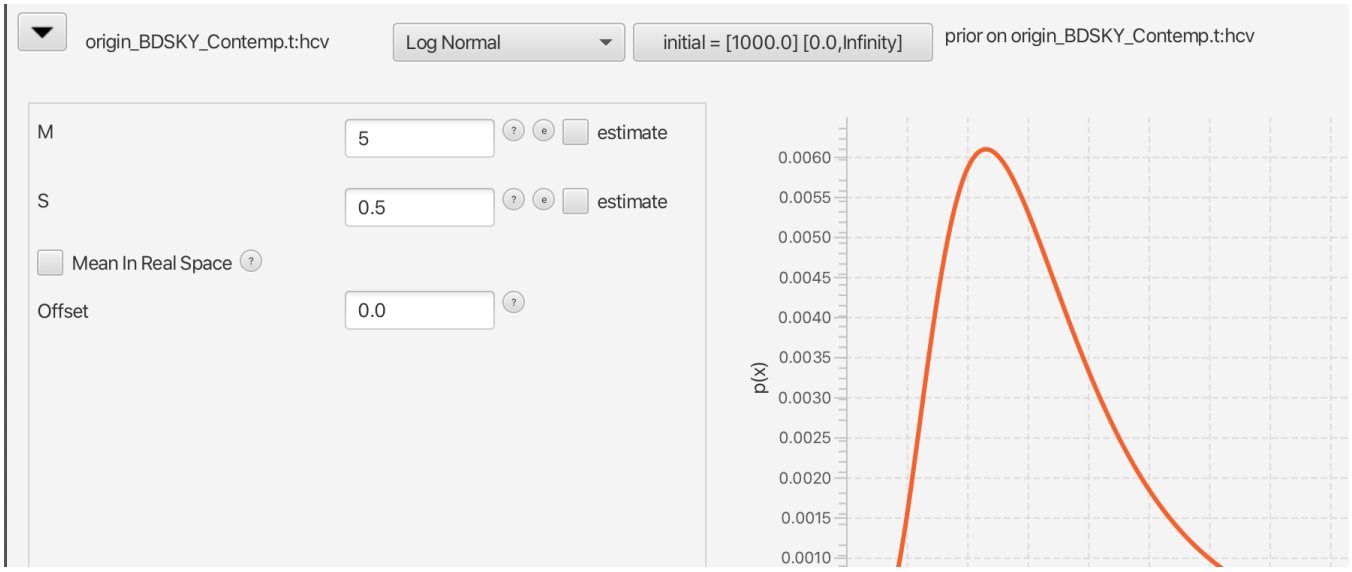

Figure 14: Setting the prior on the origin of the epidemic.

Alpha to 1 and Beta to 9999, reflecting our prior knowledge that our dataset represents only a miniscule fraction of cases (Figure 13).

Finally, we need to set a prior for the origin of the epidemic. We will once again use a log normal distribution for this parameter. Note that the origin also has to be positive and needs to be bigger than the MRCA of the tree. We know that HCV has been circulating in Egypt for at least a hundred years, so we set a prior with a median value greater than 100.

Set a **Log Normal** prior for **origin** with $\mathbf{M = 5}$ and $\mathbf{S = 0.5}$ (Figure 14), resulting in a median prior estimate for the origin of 148 years.

The rest of the priors pertain to the site model parameters and we can leave them as they are.

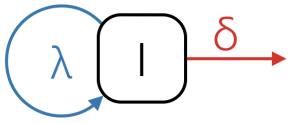

Figure 15: A schematic of the BDSKY model.

Navigate to the **MCMC** panel.

Change the **Chain Length** from 10'000'000 to 3'000'000.

Click on the arrow next to the **tracelog**, change the **File Name** and set the **Log Every** to 3'000. Use the same name for the log and trees files to avoid confusion.

Click on the arrow next to the **treelog** and change the **File Name** of the log file and set the **Log Every** to 3'000.

Leave all other settings at their default values and save the file as *hcv_bdsky.xml*.

Now we are ready to run the analysis.

Start **BEAST2** and choose your xml file. If you have **BEAGLE** installed tick the box to **Use BEAGLE library if available**, which will make the run faster.

Hit **Run** to start the analysis.

Read through the next section while waiting for the analysis to finish.

3.3.2 The Birth-Death Skyline parameterization

The birth-death model is parameterized very differently from the coalescent model. A basic birth-death model has a birth rate (λ), the rate of branching, and a death rate (δ), the rate at which lineages are removed from the tree (Figure 15). In an infectious disease epidemic λ can be thought of as the transmission rate, the rate at which rate infected individuals infect susceptibles, while δ can be thought of as the becoming uninfectious rate, the rate at which infected individuals recover, die or are isolated. In species tree inferences these rates can be thought of in terms of speciation and extinction.

The **Birth Death Skyline Contemporary** model we used was parameterized in terms of R_e and δ. Recall that $R_e > 1$means that an epidemic will keep growing. We can see this from the definition of R_e as the ratio of the birth and death rates.

$$R_e = \frac{\lambda}{\delta} \tag{2}$$

$$\begin{array}{ll} \text{if } \lambda > \delta \text{ then } R_e > 1 & \text{epidemic grows} \\ \text{if } \lambda = \delta \text{ then } R_e = 1 & \text{epidemic stays constant} \\ \text{if } \lambda < \delta \text{ then } R_e < 1 & \text{epidemic declines} \end{array}$$

We used this parameterization simply because it is often easier to specify priors for R_e than the transmission rate, and because R_e is often more informative for prevention efforts. In addition, the model also has a sampling probability (ρ) parameter, which in our analysis describes how likely it is that a person infected with HCV in Egypt in 1993 was sampled in our dataset. The final parameter is the origin. Whereas coalescent models work backward-in-time from the sampled sequences, birth-death models work forward-in-time from the origin. Hence, the model needs an origin time, which can also be jointly estimated along with the other parameters. The origin will always be bigger than the tMRCA of the sampled tree, since the sampled tree is by definition smaller than the complete tree.

You may have noticed that there are many Birth-Death Skyline models available in BEAUti. For example, the **Birth Death Skyline Contemporary BDSParam** model is parameterized in terms of λ, δ and ρ and is usually more appropriate for macro-evolutionary studies. The **Birth Death Skyline Serial** model assumes that the data are heterochronous (sampled at different times). It assumes that:

$$\delta = \psi + \mu \tag{3}$$

where ψ is the rate at which lineages are sampled through time and μ is the rate at which lineages are removed from the tree for any other reason (death, recovery, extinction etc.). (In this case the ρ parameter is no-longer available, because samples are collected through time, and not just at one timepoint). By default, the model is parameterized in terms of R_e, δ and p, the sampling proportion:

$$p = \frac{\psi}{\psi + \mu} \tag{4}$$

The sampling proportion is the proportion of all removed lineages that were sampled, and can be used to obtain a rough estimate of the total population size (such as the case of preferential sampling). This model is useful for studying infectious disease dynamics, because samples are often collected over the course of an epidemic. It can also be used for macro-evolutionary studies, when fossil data (morphological traits or ancient DNA) are incorporated. In that case a parameterization in terms of λ, μ and ψ is preferable.

You can also see that the model **Birth Death Skyline Serial** assumes that upon sampling a lineage is removed from the tree (e.g. in a disease model the sampled individual cannot transmit the disease after sampling). The consequence for the phylogeny is that a sampled lineage cannot be a direct ancestor of any other lineage in the tree.

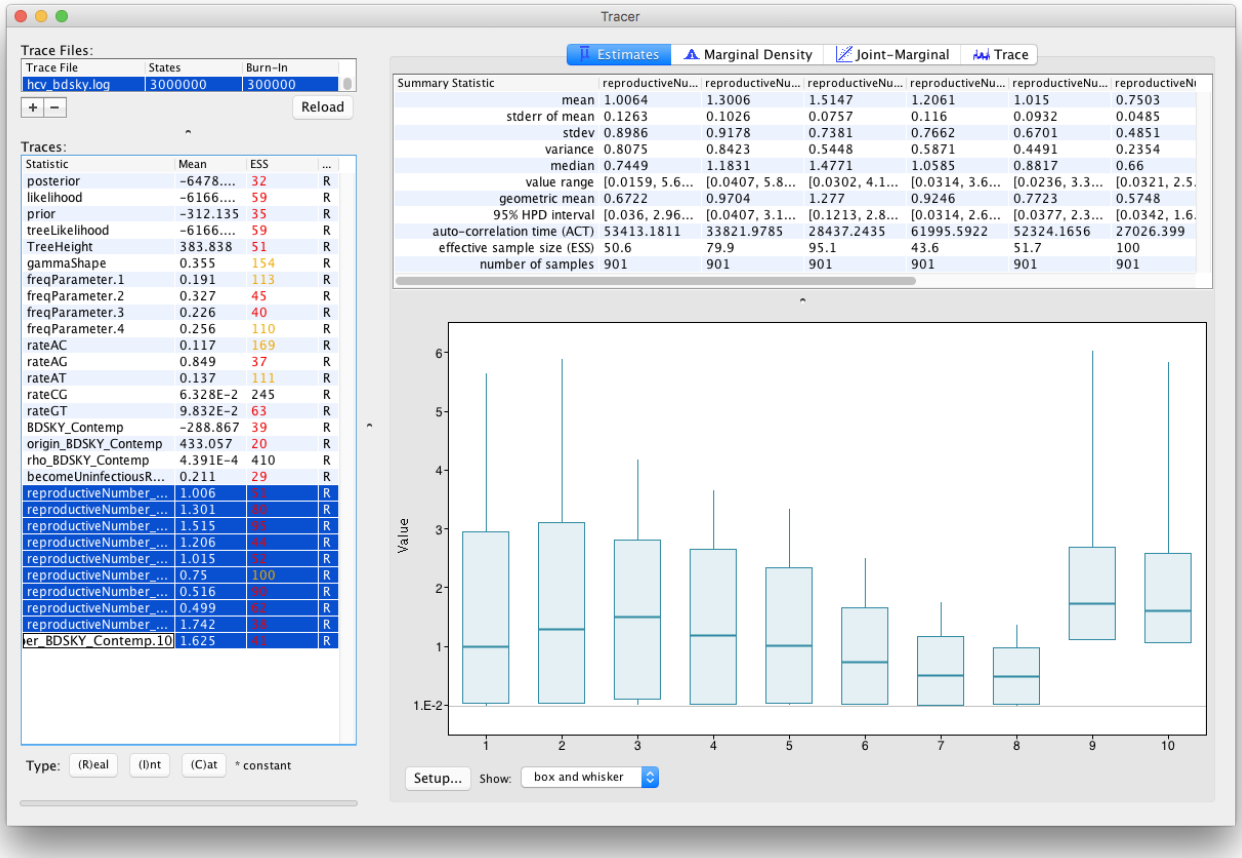

Figure 16: Estimated population dynamics by BDSKY in Tracer.

3.3.3 Visualizing the Birth-Death Skyline Output

There is no equivalent visualization of the skyline plot of a Birth-Death Skyline (BDSKY) analysis in Tracer as there is for the Coalescent Bayesian Skyline. We need to do some extra post-processing to plot the skyline estimator. We can do this with the R-package `bdskytools`. The R Markdown and pdf files are in this tutorial's folder.

4 Practical 2: Bayesian Skyline Plots of Influenza H3N2 in New York State

In this tutorial we will estimate the dynamics of human influenza A H3N2 subtype in New York. The data set contains 165 Hemagglutinin gene sequences. To save time, we will show how to setup the analysis and summarize precomputed runs.

The aim of this tutorial is to infer population dynamics from heterochronously sampled data. The data file is called "NewYork.HA.2000-2003" and can be download here: https://beast.community/tutorials/workshop_influenza_phylodynamics/files/NewYork.HA.2000-2003.nex.

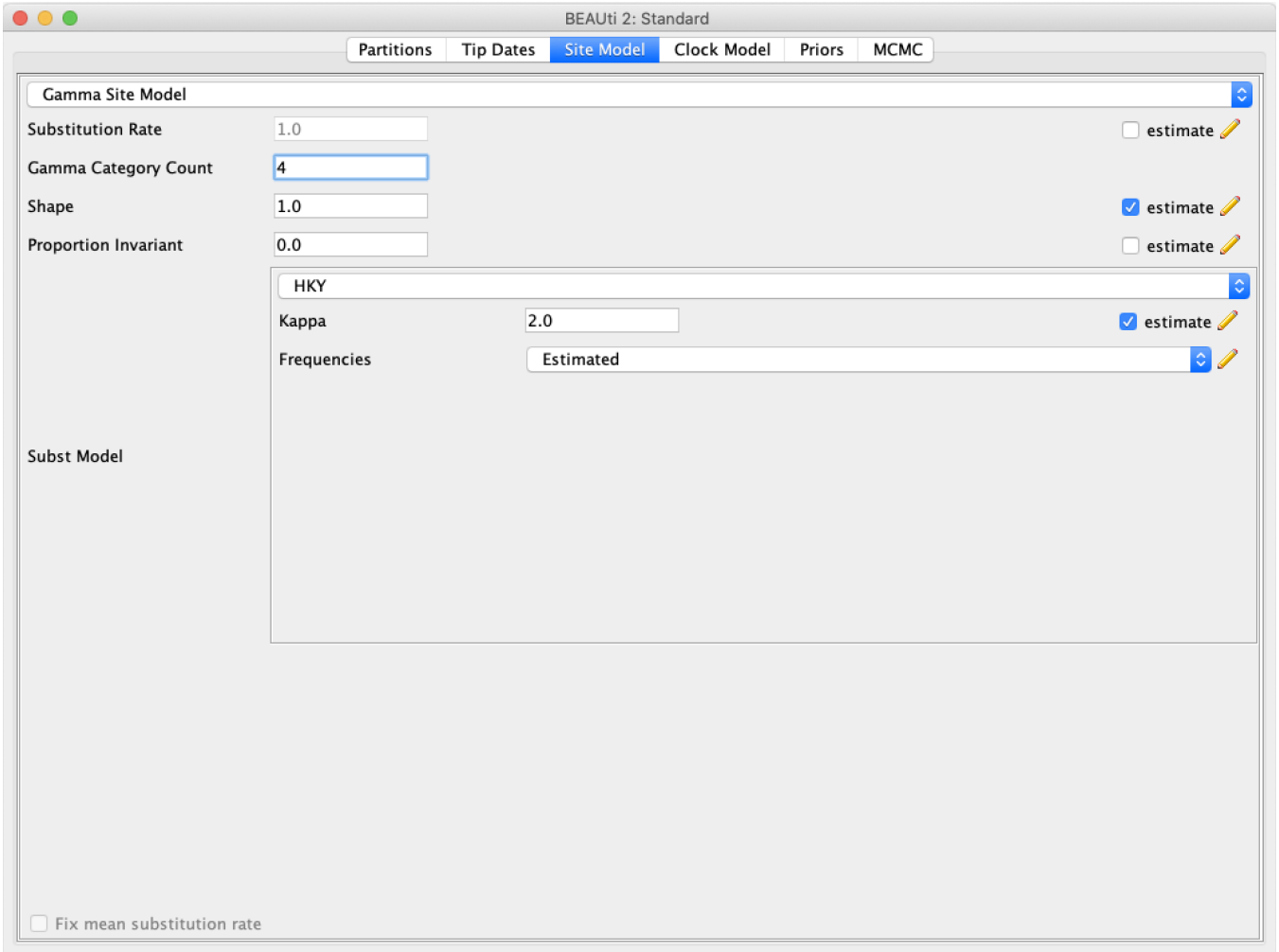

Figure 17: Set the HKY substitutions model.

4.1 Setting up the Coalescent Bayesian Skyline analysis

To start we have to import the alignment into BEAUti and follow the same steps as Practical 1 Bayesian Skyline, except that now we will set the tip dates.

In the **Tip Dates** panel, select "Use tip dates and click Auto-configure. Select use everything after last "_" in the **BEAUti** window.

This time we place the HKY model prior with a Gamma count of 4 and fix the clock rate.

For the tree prior, we will select the Coalescent Bayesian Skyline with 50 change points.

The XML file is called NewYork.xml and available in the tutorial folder. The estimated Skyline Plot is

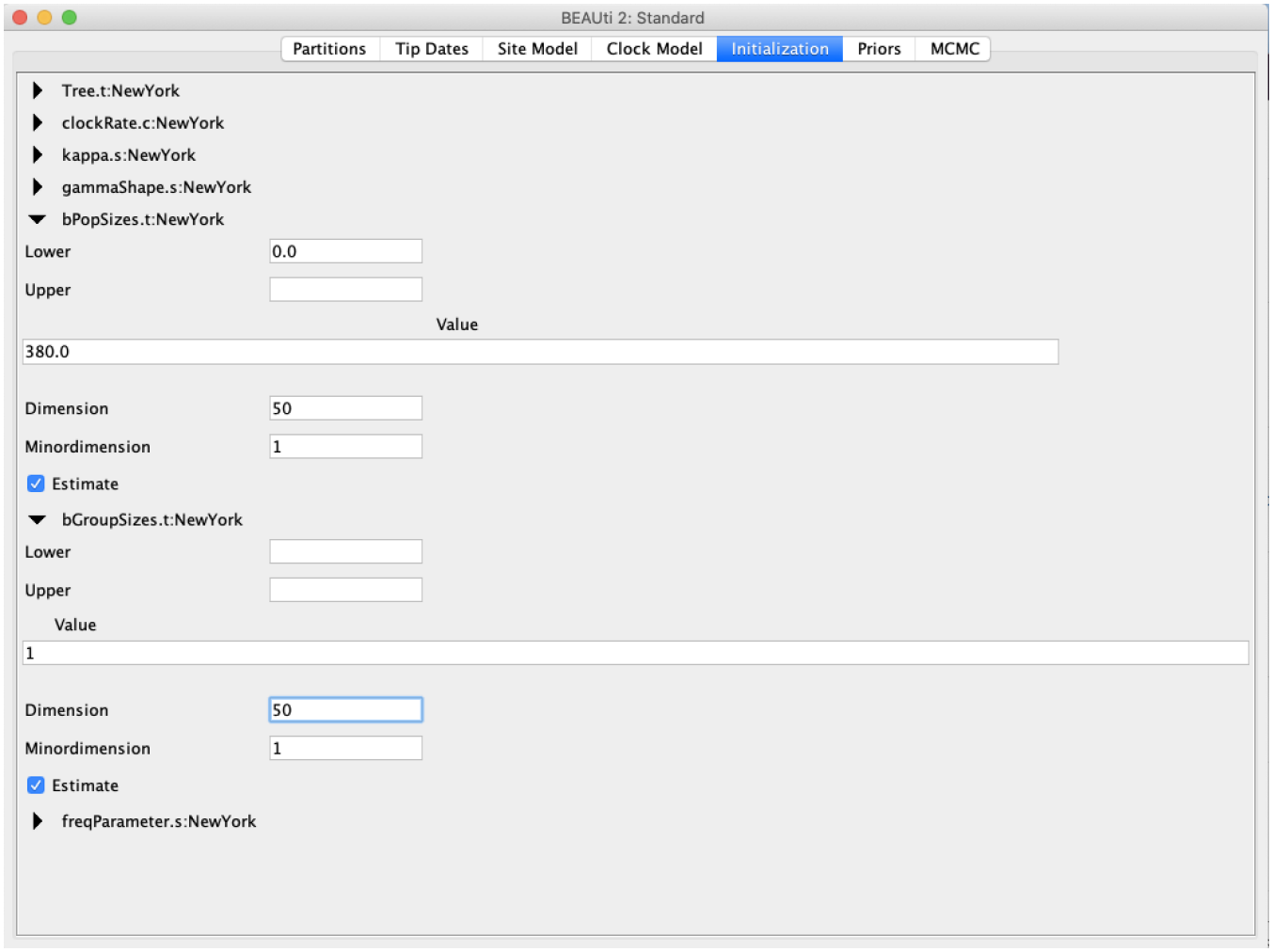

Figure 18: Choose the Coalescent Bayesian Skyline as a tree prior.

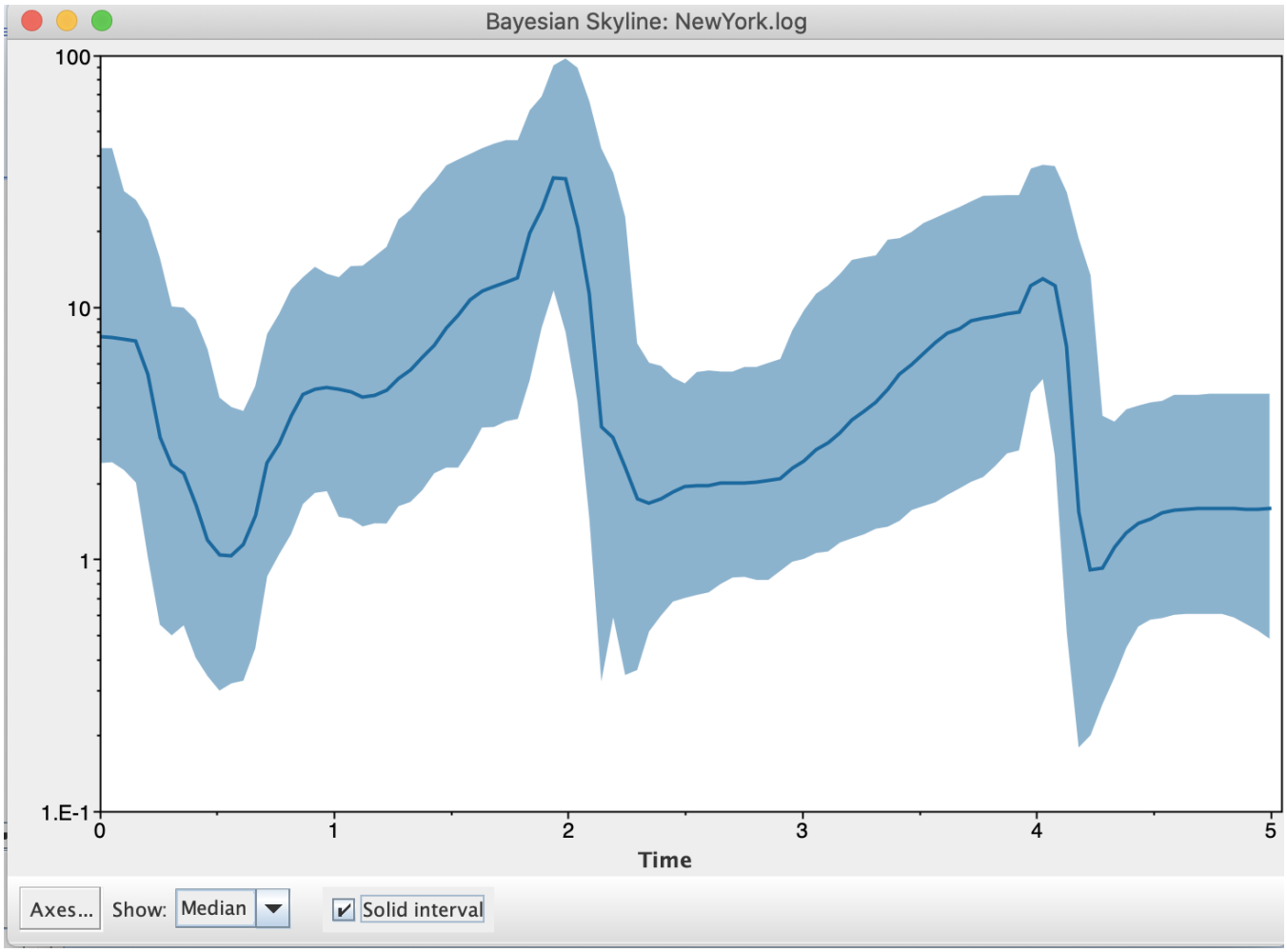

Figure 19: Estimated effective population size trajectory. We did not set present time and instead present time is shown as 0.

4.2 Accounting for preferential sampling

We can generate the Maximum Clade Credibility Tree and use it to estimate the effective population size with Gaussian Process priors with and without preferential sampling. The R markdown file is available in this tutorial's folder.

5 Useful Links

- Bayesian Evolutionary Analysis with BEAST 2 (Drummond and Bouckaert 2014)
- BEAST 2 website and documentation: http://www.beast2.org/
- BEAST 1 website and documentation: http://beast.bio.ed.ac.uk
- Join the BEAST user discussion: http://groups.google.com/group/beast-users
- bdskytools: An R-package for post-processing Birth-Death Skyline analyses
- phylodyn: An R-package for Bayesian nonparametric inference of effective population sizes from a fixed phylogeny.

Relevant References

Drummond, AJ, A Rambaut, B Shapiro, and OG Pybus. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution* 22: 1185–92.

Drummond, AJ and RR Bouckaert. 2014. *Bayesian evolutionary analysis with BEAST 2*. Cambridge University Press,

Karcher, MD, JA Palacios, S Lan, and VN Minin. 2017. Phylodyn: an R package for phylodynamic simulation and inference. *Molecular ecology resources* 17: 96–100.

Minin, VN, EW Bloomquist, and MA Suchard. 2008. Smooth skyride through a rough skyline: bayesian coalescent-based inference of population dynamics. *Molecular biology and evolution* 25: 1459–71.

Palacios, JA and VN Minin. 2013. Gaussian process-based bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics* 69: 8–18.

Pybus, OG et al. 2001. The epidemic behavior of the hepatitis c virus. *Science (New York, N.Y.)* 292: 2323–5.

Ray, S, R Arthur, A Carella, J Bukh, and D Thomas. 2000. Genetic epidemiology of hepatitis c virus throughout egypt. *The Journal of Infectious Diseases* 182: 698–707.

Stadler, T, D Kuhnert, S Bonhoeffer, and AJ Drummond. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences* 110: 228–233.