Tutorial using BEAST v2.4.7

MASCOT Tutorial

Nicola F. Müller

Parameter and State inference using the approximate structured coalescent

1 Background

Phylogeographic methods can help reveal the movement of genes between populations of organisms. This has been widely used to quantify pathogen movement between different host populations, the migration history of humans, and the geographic spread of languages or the gene flow between species using the location or state of samples alongside sequence data. Phylogenies therefore offer insights into migration processes not available from classic epidemiological or occurrence data alone.

The structured coalescent allows to coherently model the migration and coalescent process, but struggles with complex datasets due to the need to infer ancestral migration histories. Thus, approximations to the structured coalescent, which integrate over all ancestral migration histories, have been developed. This tutorial gives an introduction into how a MASCOT analysis in BEAST2 can be set-up. MASCOT is short for **M**arginal **A**pproximation of the **S**tructured **CO**alescen**T** and implements a structured coalescent approximation (Müller et al. 2017). This approximation doesn't require migration histories to be sampled using MCMC and therefore allows to analyse phylogenies with more than three or four states.

2 Programs used in this Exercise

2.0.1 BEAST2 - Bayesian Evolutionary Analysis Sampling Trees 2

BEAST2 (http://www.beast2.org) is a free software package for Bayesian evolutionary analysis of molecular sequences using MCMC and strictly oriented toward inference using rooted, time-measured phylogenetic trees. This tutorial is written for BEAST v2.4.7 (Drummond and Bouckaert 2014).

2.0.2 BEAUti2 - Bayesian Evolutionary Analysis Utility

BEAUti2 is a graphical user interface tool for generating BEAST2 XML configuration files.

Both BEAST2 and BEAUti2 are Java programs, which means that the exact same code runs on all platforms. For us it simply means that the interface will be the same on all platforms. The screenshots used in this tutorial are taken on a Mac OS X computer; however, both programs will have the same layout and functionality on both Windows and Linux. BEAUti2 is provided as a part of the BEAST2 package so you do not need to install it separately.

2.0.3 TreeAnnotator

TreeAnnotator is used to summarise the posterior sample of trees to produce a maximum clade credibility tree. It can also be used to summarise and visualise the posterior estimates of other tree parameters (e.g. node height).

TreeAnnotator is provided as a part of the BEAST2 package so you do not need to install it separately.

2.0.4 Tracer

Tracer (http://tree.bio.ed.ac.uk/software/tracer) is used to summarise the posterior estimates of the various parameters sampled by the Markov Chain. This program can be used for visual inspection and to assess convergence. It helps to quickly view median estimates and 95% highest posterior density intervals of the parameters, and calculates the effective sample sizes (ESS) of parameters. It can also be used to investigate potential parameter correlations. We will be using Tracer v1.6.0.

2.0.5 FigTree

FigTree (http://tree.bio.ed.ac.uk/software/figtree) is a program for viewing trees and producing publication-quality figures. It can interpret the node-annotations created on the summary trees by TreeAnnotator, allowing the user to display node-based statistics (e.g. posterior probabilities). We will be using FigTree v1.4.2.

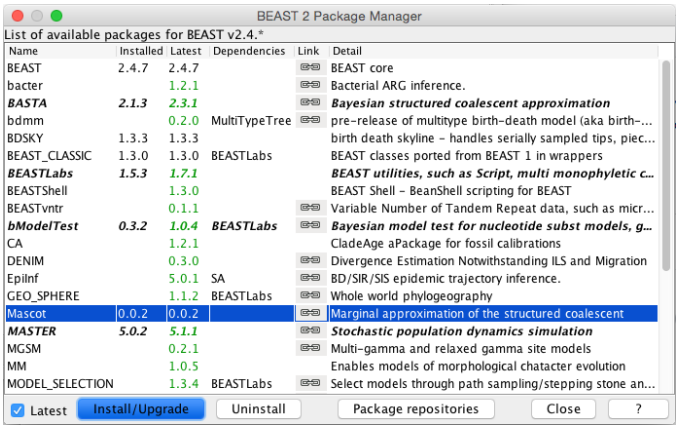

Figure 1: Download the MASCOT package.

3 Practical: Parameter and State inference using the approximate structured coalescent

In this tutorial we will estimate migration rates, effective population sizes and locations of internal nodes using the marginal approximation of the structured coalescent implemented in BEAST2, MASCOT.

The aim is to:

- Learn how to infer structure from trees with sampling location
- Get to know how to choose the set-up of such an analysis
- Learn how to read the output of a MASCOT analysis

3.1 Setting up an analysis in BEAUti

3.1.1 Download MASCOT

First, we have to download the package MASCOT using the BEAUTi package manager. Go to *File >> Manage Packages* and download the package MASCOT.

MASCOT will only be available in BEAUti once you close and restart the program.

3.1.2 Loading the template

Next, we have to load the BEAUTi template from *File*, select *Template >> Mascot*.

3.1.3 Loading the Influenza A/H3N2 Sequences (Partitions)

The sequences from the *data* folder name *H3N2.nexus* can be either drag and dropped into BEAUti or added by going to *File >> Import Alignment*. Once the sequences are added, we need to specify the sampling dates and locations.

3.1.4 Get the sampling times (Tip Dates)

After clicking the *Auto-configure* button, the sampling times can be guessed. The sampling times are encoded in the sequence names and are in the third group after splitting on the vertical bar "|". The first

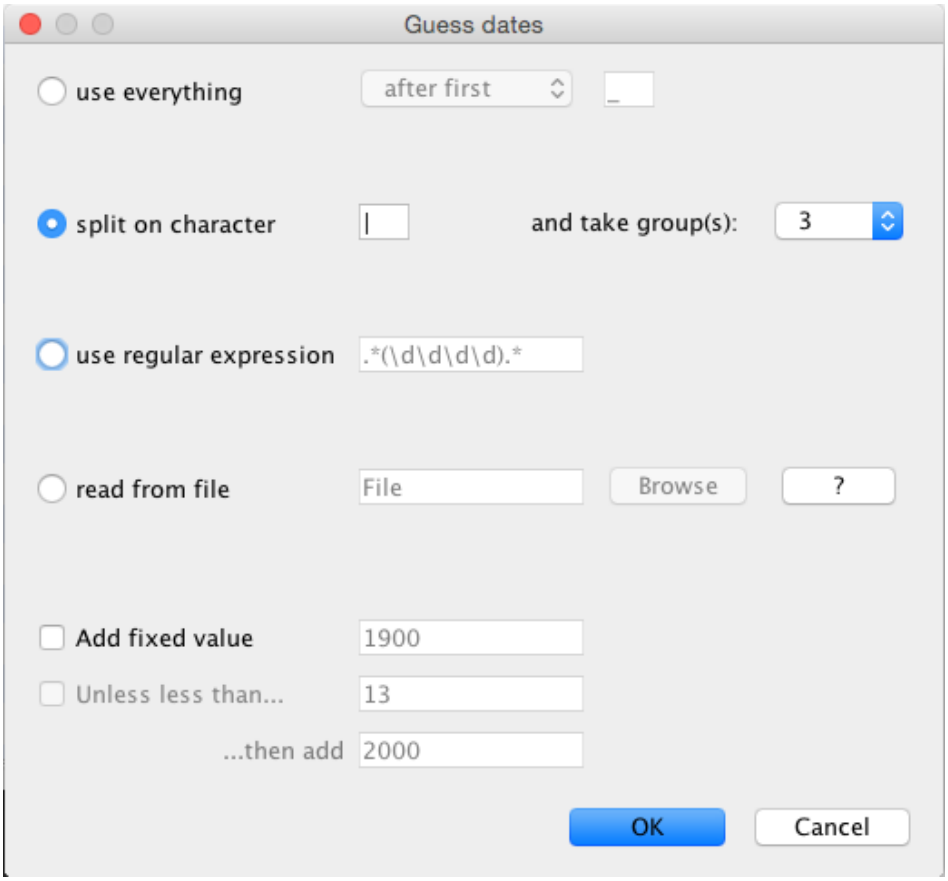

Figure 2: Guess sampling times.

group after splitting is the name of the sequence, the second group contains the accession numbers. The third are the sampling times and the fourth are the sampling location.

After guessing the sampling times, the column **Date** should now have values between 2000 and 2002 and the column **Height** should have values from 0 to 2. The heights denote the time difference from a sequence to the most recently sampled sequence. If everything is specified correctly, the sequence with Height 0.0 should have Date 2001.9. Next, the sampling locations need to be specified.

3.1.5 Get the sampling locations (Tip Locations)

As for the sampling times, they can be guessed from the sequence names. Initially the column **Location** should be NOT_SET for every sequence. After clicking the *Guess* button, you can split the sequence on the vertical bar "|" again. As said before, the locations are in the fourth group. After clicking the *OK* button, the window should now look like in the figure below:

3.1.6 Specify the Site Model (Site Model)

Next, we have to specify the site model. For Influenza Hemagluttanin sequences as we have here, HKY is the most commonly used model of nucleotide evolution. It allows for difference in transversion and transition rates. Meaning that changes between bases that are chemically closer related (transitions) are allowed to have a different rate than changes between bases that chemically more distinct (transversion).

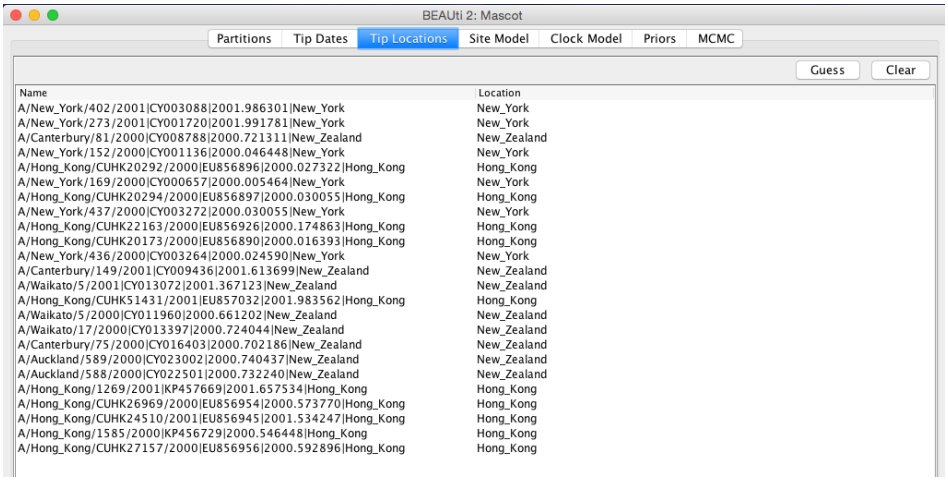

Figure 3: Guess sampling locations.

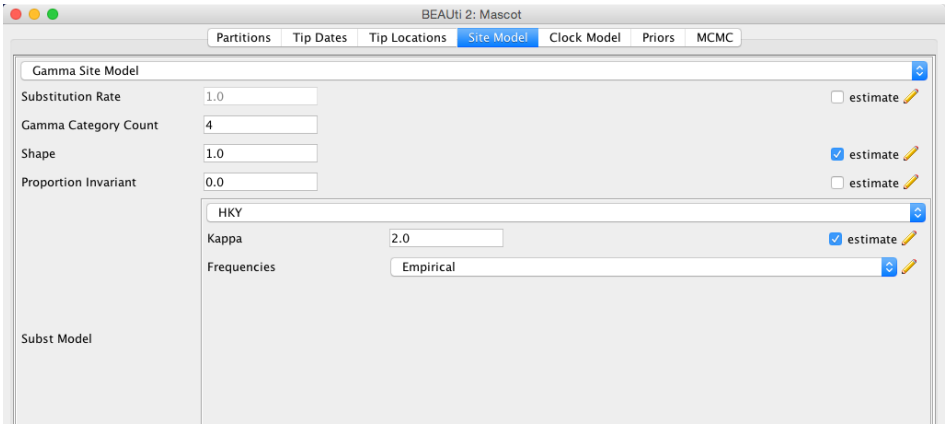

Figure 4: Set the site model.

Additionally, we should allow for different rate categories for different sires in the alignment. This can be done by setting the *Gamma Category Count* to 4, which is just a value that has typically been used. Make sure that estimate is checked next to the shape parameter. To reduce the number of parameters we have to estimate, we can set Frequencies to Empirical.

3.1.7 Set the clock model (Clock Model)

For rapidly evolving viruses, the assumption of a strict molecular clock is often made, meaning that the molecular clock is the same on each branch of the phylogeny. To decrease the burnin phase, we can set the initial value to 0.005.

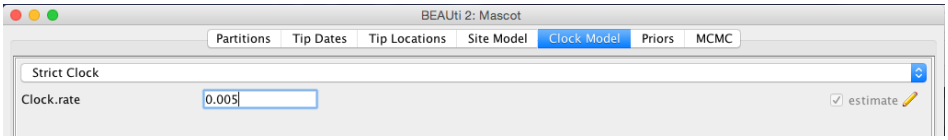

Figure 5: Set the initial clock rate.

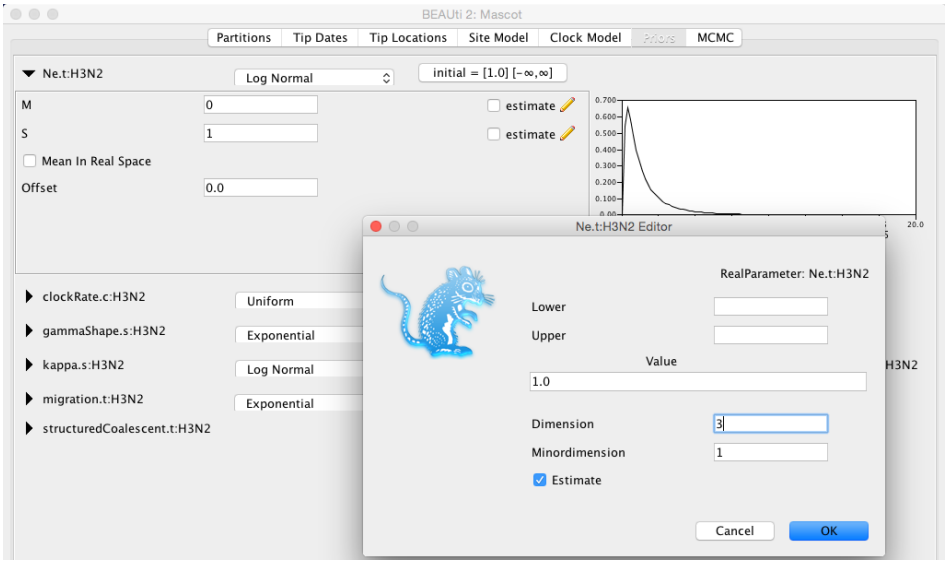

Figure 6: Set the dimension for the effective population sizes to 3.

3.1.8 Specify the priors and set dimensions (Priors)

Now, we need to set the priors as well as the dimensions of the effective population sizes and the migration rates. For this example we have sequences from Hong Kong, New Zealand and, New York . Overall we have three different locations, meaning that we need an effective population size for each of these locations. You can set the dimension of the effective population size by pressing the *initial* button. A window will then appear where you can set the dimension to 3. Next, we can change the prior to a Log Normal prior with M=0 and S=1. Since we have only a few samples per location, meaning little information about the different effective population sizes, we will need an informative prior.

Next, we have to set the dimension of the migration rate parameter. A lineage from any of the 3 locations can migrate to 2 (3-1) other locations. Overall, we therefore have to estimate 3*(3-1) migration rates and have to set the dimension accordingly. The exponential distribution as a prior on the migration rate puts much weight on lower values while not prohibiting larger ones. For migration rates, a prior that prohibits too large values while not greatly distinguishing between very small and very very small values (such as the inverse uniform) is generally a good choice.

Next, we have to set a prior for the clock rate. Since we only have a narrow time window of less than a year and only 24 sequences, there isn't much information in the data about the clock rate. We have however a good idea about it for Influenza A/H3N2 Hemagglutinin. We can therefore set the prior to be normally distributed around 0.005 substitution per site and year with a variance of 0.0001. (At this point we could also just fix the rate)

3.1.9 Specify the MCMC chain length (MCMC)

Here we can set the length of the MCMC chain and after how many iterations the parameter and trees a logged. For this dataset, 2 million iterations should be sufficient. In order to have enough samples but not create too large files, we can set the logEvery to 2500, so we have 801 samples overall. Next, we have to save the `*.xml` file under *File >> Save as*.

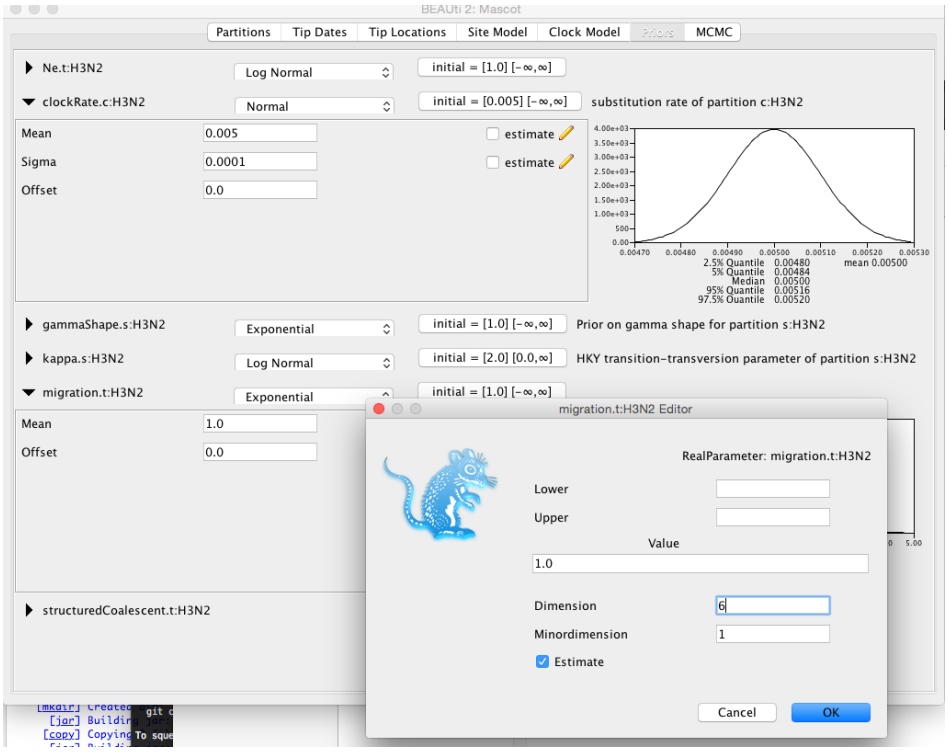

Figure 7: Set the dimension of the migration rates to 6.

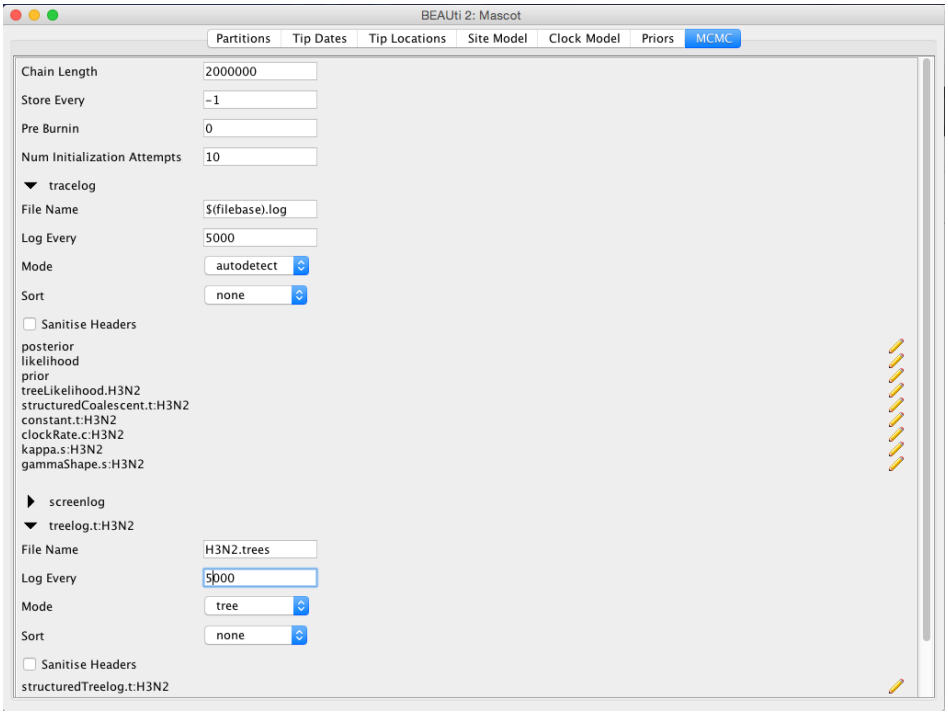

Figure 8: save the *.xml.

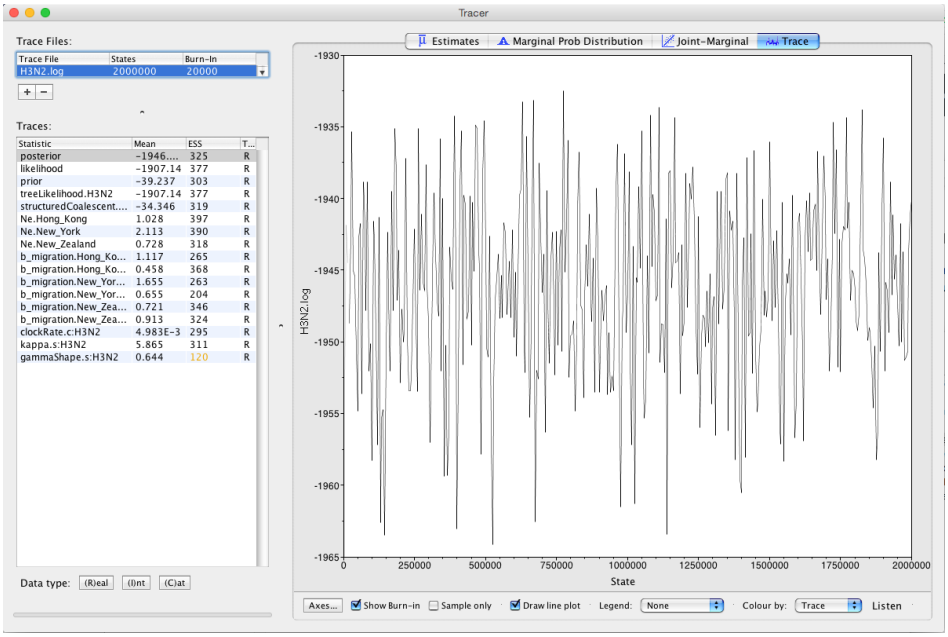

Figure 9: Check if the posterior converged.

3.1.10 Run the Analysis using BEAST2

Run the `*.xml` using BEAST2 or use finished runs from the *precooked-runs* folder. The analysis should take about 6 to 7 minutes.

3.1.11 Analyse the log file using Tracer

First, we can open the `*.log` file in tracer to check if the MCMC has converged. The ESS value should be above 200 for almost all values and especially for the posterior estimates. The burnin taken by Tracer is 10%, but for this analysis 1% is enough.

Next, we can have a look at the inferred effective population sizes. New York is inferred to have the largest effective population size before Hong Kong and New Zealand. This tells us that two lineages that are in the New Zealand are expected to coalesce quicker than two lineages in Hong Kong or New York.

In this example, we have relatively little information about the effective population sizes of each location. This can lead to estimates that are greatly informed by the prior. Additionally, there can be great differences between median and mean estimates. The median estimates are generally more reliable since they are less influence by extreme values.

We can then look at the inferred migration rates. The migration rates have the label b_migration.*, meaning that they are backwards in time migration rates. The highest rates are from New York to Hong Kong. Because they are backwards in time migration rates, this means that lineages from New York are inferred to be likely from Hong Kong if we're going backwards in time. In the inferred phylogenies, we should therefore make the observation that lineages ancestral to samples from New York are inferred to be from the Hong Kong backwards.

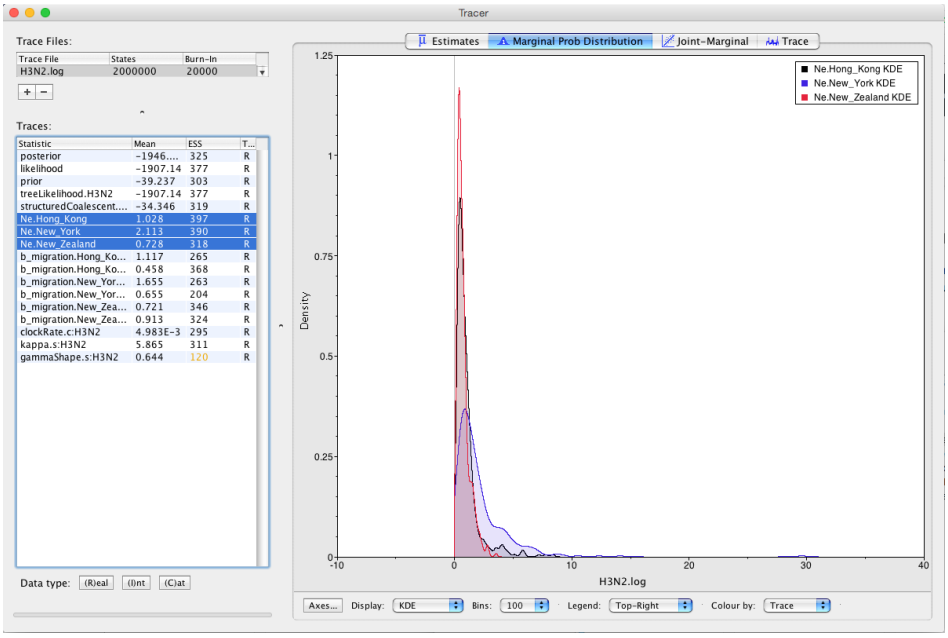

Figure 10: Compare the different inferred effective population sizes.

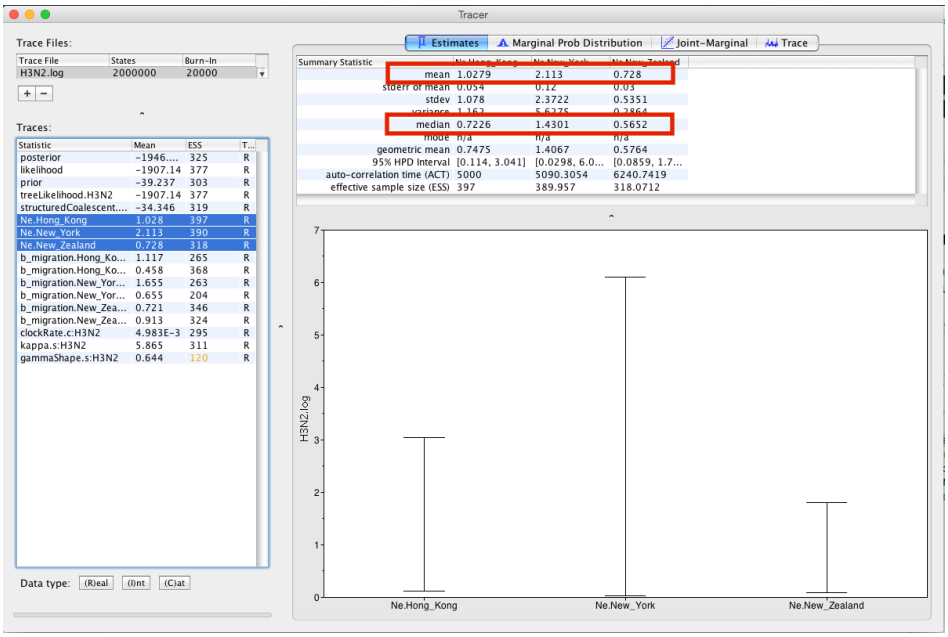

Figure 11: Differences between Mean and Meadian estimates.

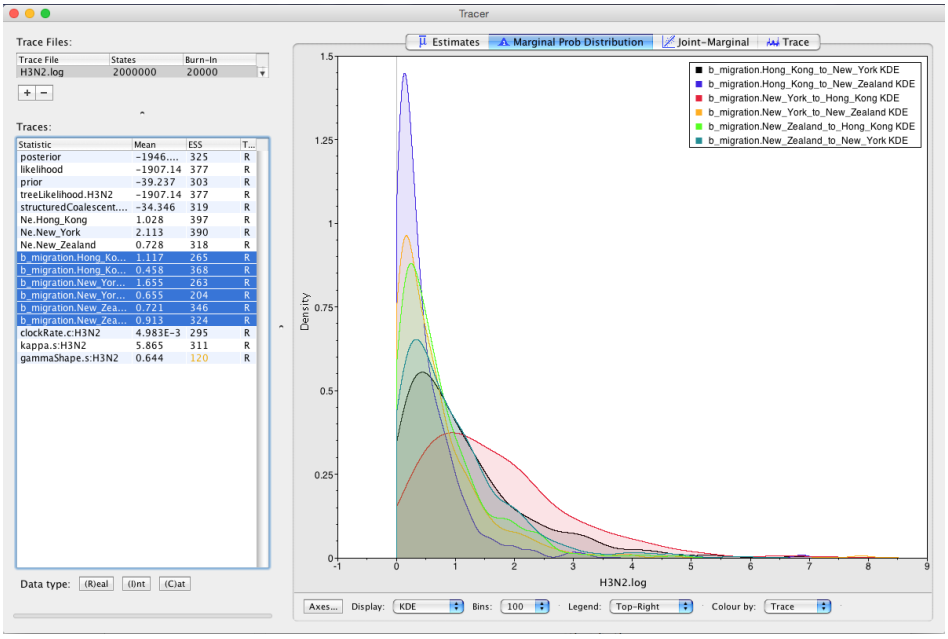

Figure 12: Compare the inferrred migration rates.

3.1.12 Make the MCC tree using TreeAnnotator

Next, we want to summarize the trees. This we can do using treeAnnotator. Open the programm and then set the options as below. You have to specify the *Burnin precentage*, the *Node heights*, *Input Tree File* and the *Output File* after clicking *Run* the programm should summarize the trees.

3.1.13 Check the MCC tree using FigTree

We can now open the MCC tree using FigTree. The output contains several things. Each node has several traits. Among them are those called Hong_Kong, New_York and New_Zealand. The value of those traits is the probability of that node being in that location as inferred using MASCOT.

We can now check if lineages ancestral to samples from New York are actually inferred to be from Hong Kong, or the probability of the root being in any of the locations. It should here be mentioned that the inference of nodes being in a particular location makes some simplifying assumptions, such as that there are no other locations where lineages could have been.

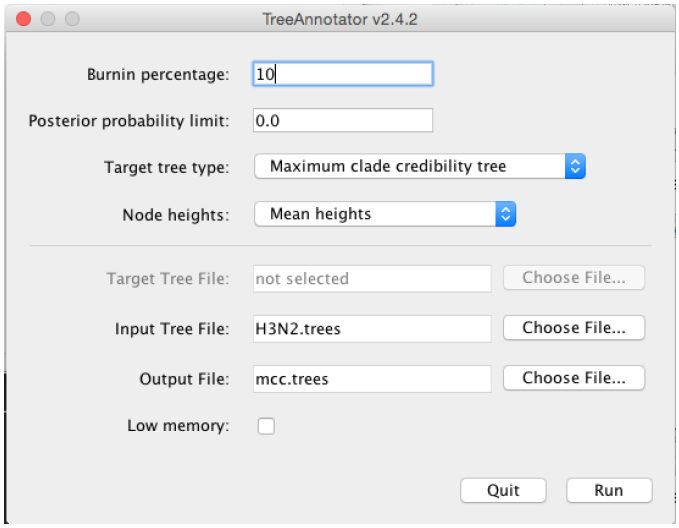

Figure 13: Make the maximum clade credibility tree.

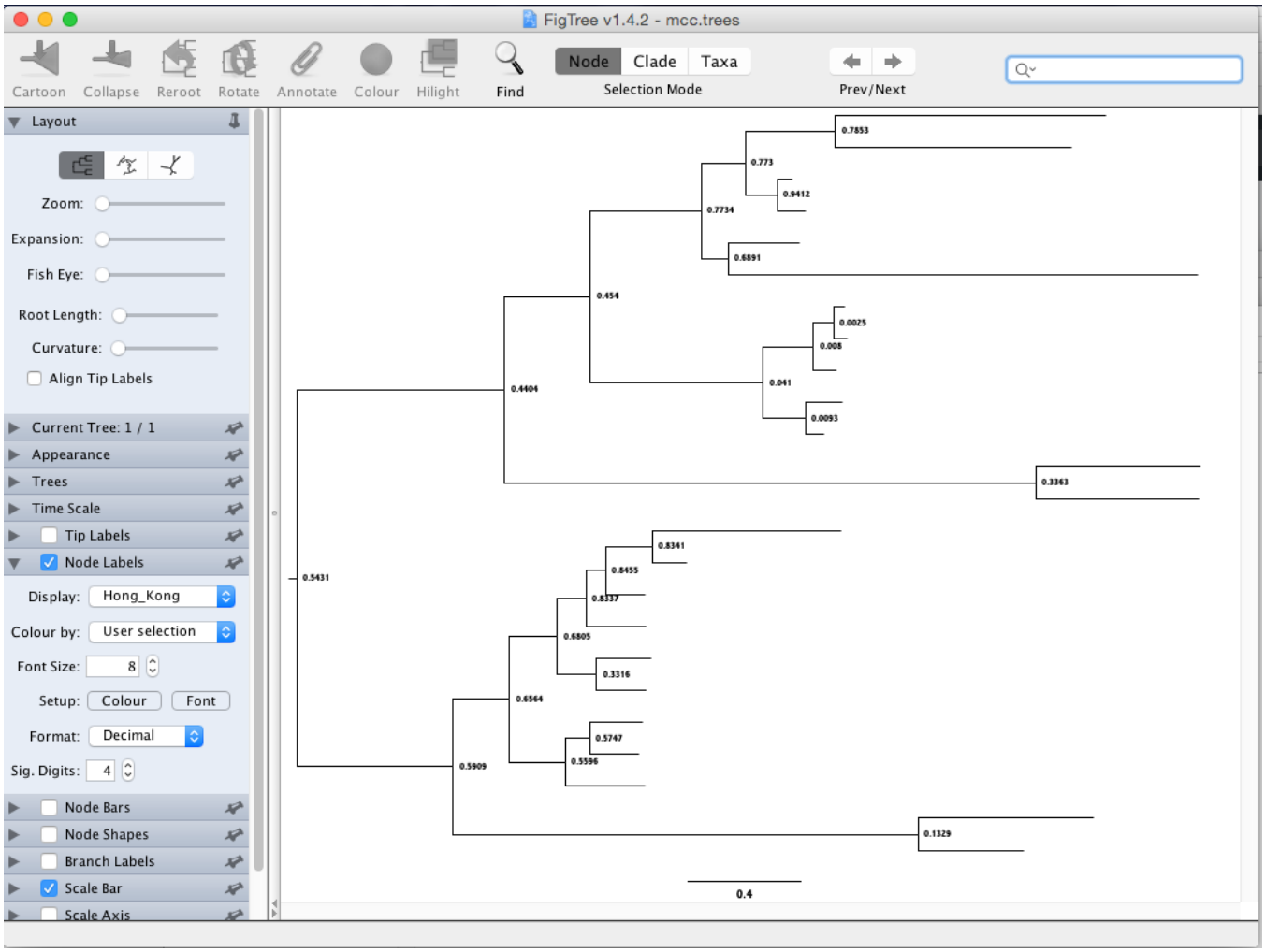

Figure 14: Compare the inferred node probabilities.

4 Useful Links

- MASCOT source code: https://github.com/nicfel/Mascot
- Bayesian Evolutionary Analysis with BEAST 2 (Drummond and Bouckaert 2014)
- BEAST 2 website and documentation: http://www.beast2.org/
- Join the BEAST user discussion: http://groups.google.com/group/beast-users

Version dated: July 21, 2017

Relevant References

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

Müller, NF, DA Rasmussen, and T Stadler. 2017. The structured coalescent and its approximations. *Molecular Biology and Evolution* msx186.