Fundamental Tree Priors

Linking the SIR, Birth-Death, and Coalescent Exponential

Leo A. Featherstone

1 Background

This tutorial covers the two key tree priors upon which most other phylodynamics applications are built. We begin by laying out the key epidemiology that each tree-prior models before moving into an application of each to an empirical dataset. We analyse a a set of sequences collected in North America from the 2009 H1N1 Influenza pandemic. These samples were collected in May, when the outbreak was still growing exponentially. This fits our assumption of exponential population growth well. (Hedge et al. 2013).

Our goal is that by the end of this tutorial, you will have understood the process of analysing sequence datasets under each key model, and would feel confident conducting similar analyses on other datasets using this tutorial as a reference.

2 Programs used in this Exercise

This tutorial uses the BEAST2 version 2.6.6, but any similar version will work. Just make sure you have the BDSKY package installed. You can do this though BEAUti before restarting it to use the package.

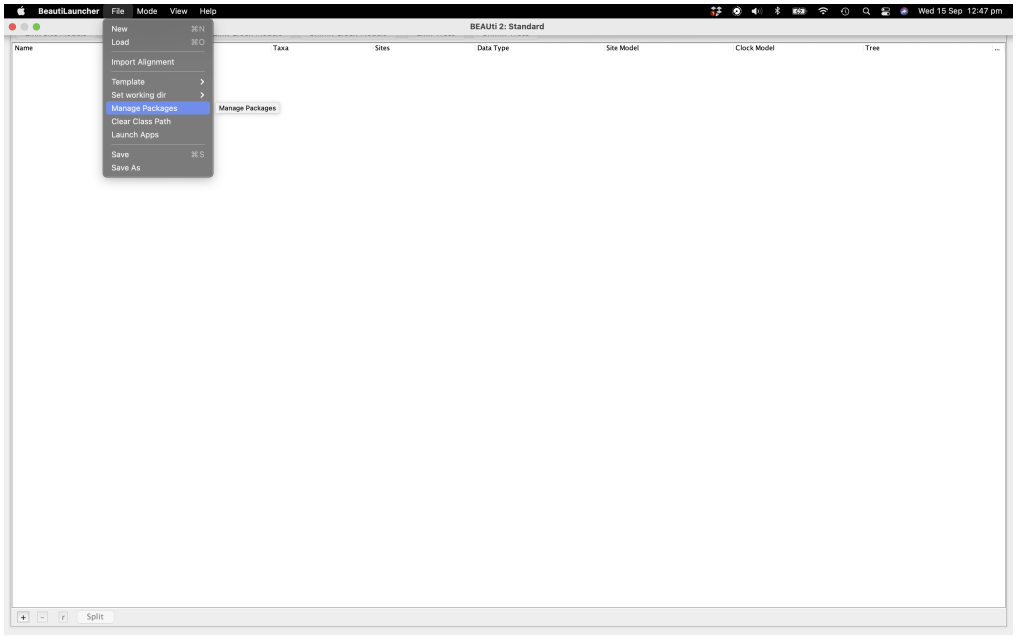

Figure 1: Navigating to 'Manage Packages' in BEAUti

3 Practical

3.1 Epidemiology Background

We covered the SIR model in the presentation earlier. To briefly recap, it models how people move from the susceptible, to infected, to recovered subsets of the population. In the way this plays out, we always initially observe an exponential growth in the number of infections because most people are still susceptible. This offers us a the chance to reliably assume and infer a constant rate of infection, which specifies exponential increase in infections (Given we start with only one infection!) $I(t) = e^{rt}$.

This growth rate, r, is the conduit between the SIR and birth-death or coalescent exponential models. It is the rate of new infections arising, taken as the difference between the per-capita rates of infection and removal ($r = \lambda - \delta$). Note finally that although we specify a 'Recovered' compartment, this really encompasses removals from the Infected compartment. From our point of view, this can be due to death, or sampling which we tend to assume immediately precedes removal from infectious compartment.

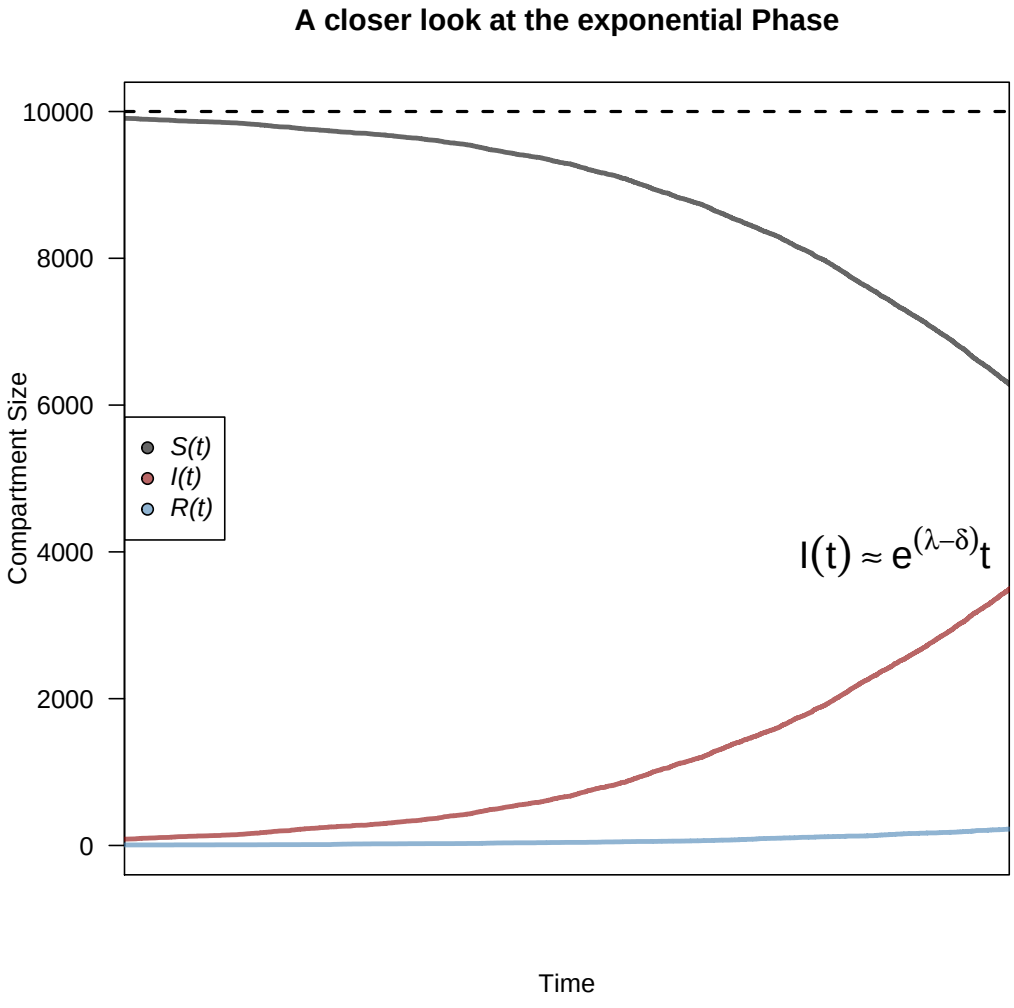

Figure 2: A closer look at the 'exponential phase'.

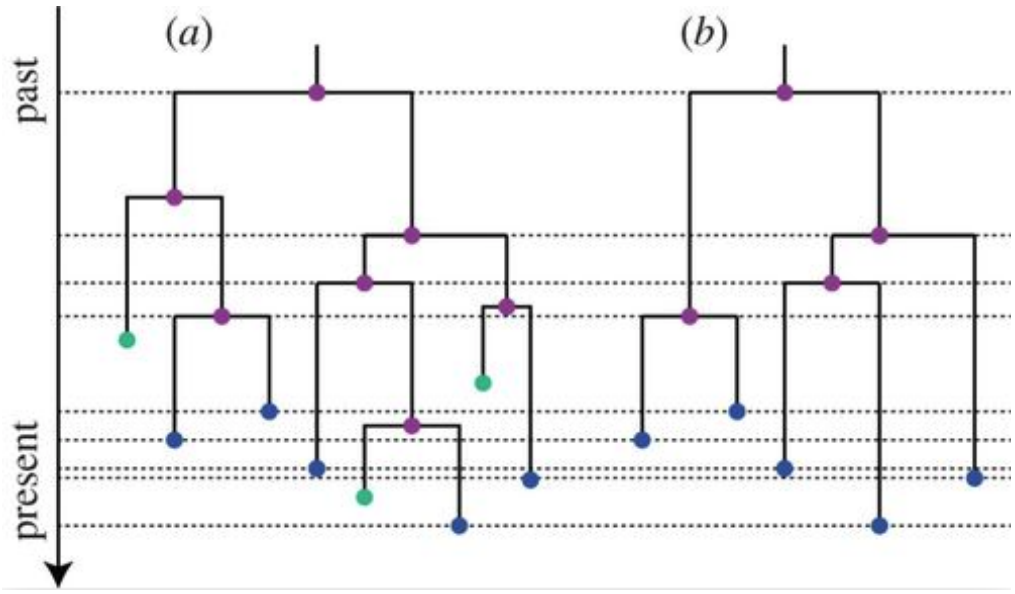

Figure 3: a) Transmission tree. b) Isolate tree. Purple dots are transmission events, green for death, and blue for sampling. Adapted from (Kühnert et al. 2014).

3.2 Coalescent Exponential

Now we're going to analyse our data using the coalescent exponential. To briefly recap, this model has two parameters. They are:

1. r, the epidemic growth rate, which is an approximation of the SIR's exponential growth rate above
2. ϕ, the scaled effective population size. This isn't really some thing we can perceive, but it's a crucial part of the frame work here. Formally,

$$\phi = \frac{I(0)}{\lambda} \tag{1}$$

   Recall that the coalescent measures time backwards, so $I(0)$ is prevalence at the time of the most recent sample, and λ is the coalescence rate (backwards infection rate). This means ϕ captures both incidence and prevalence, so these can thankfully both influence our model. It's a nuisance parameter as far as interpretation is concerned, but it's critical to the coalescent exponential's effectiveness.

After running the coalescent exponential, we will obtain a posterior mean estimate for r. We will use this to calculate critical communicable parameters: R_0, doubling time (t_d), and estimates number of cases at the time of the final sample $I(0)$.

$$R_0 = rD + 1, \text{ where D = average duration of infection } (D = \frac{1}{\delta}) \tag{2}$$

$$t_d = \frac{\ln 2}{r} \tag{3}$$

$$I(0) = e^{rT}, \text{ where T is the age of the root} \tag{4}$$

Now we can move on to setting up this analysis in BEAUti. Bear in mind that the first few steps of the process include placing priors on factors like the the substitution model that aren't part of the tree prior.

3.2.1 Setting Up the Coalescent Exponential

1. Import the sequence data in BEAUti. Navigate to wherever you have NorthAM.May.fasta stored. Be sure to select nucleotide form the drop-down menu.

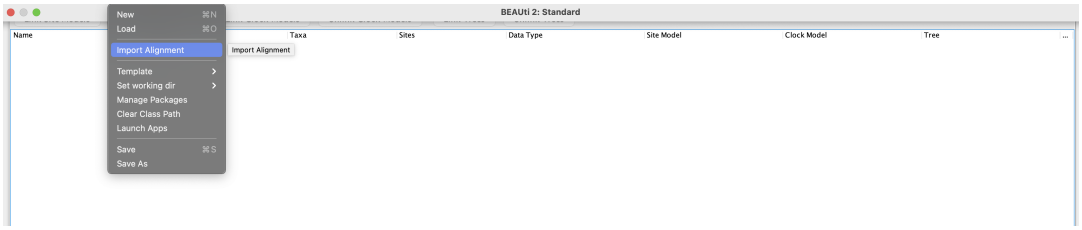

Figure 4: Importing the alignment

2. Next, click on the 'Tip Dates' atop the window and select 'Use tip dates'. This should load in the tip names as per Figure 4. Go to 'Auto-configure' to the right of the window and tell BEAUti to grab the dates from after the last '|' in each tip name. This should then populate the 'Date (raw value)' and Height columns. Heights refer to the difference in time between the youngest tip and all others. BEAST uses these heights in its likelihood calculations.

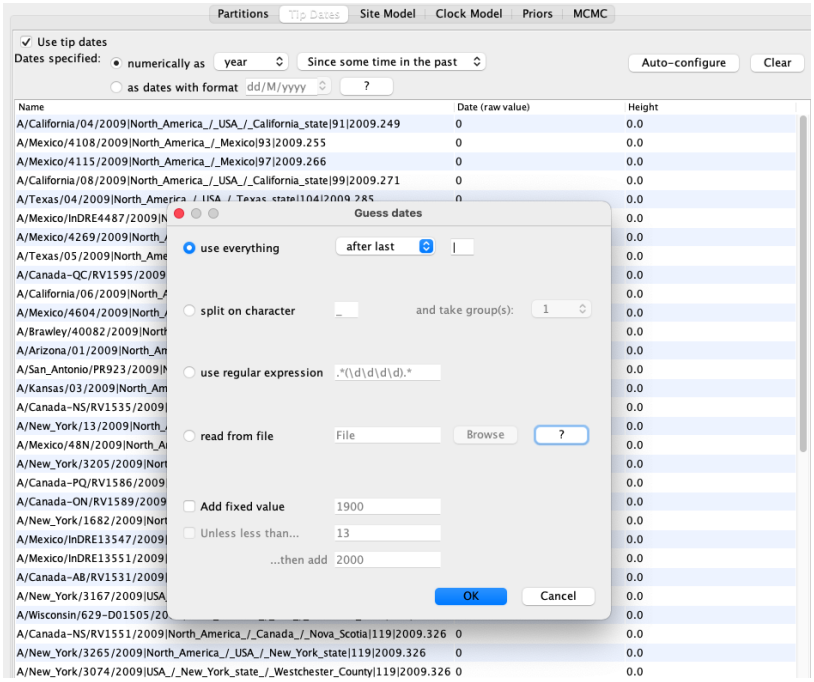

Figure 5: Importing tip-dates

3. Click on the site model tab. Select the HKY+G model as shown in Figure 5. This model will account for rate heterogeneity among sites and for the transition to tranversion bias.

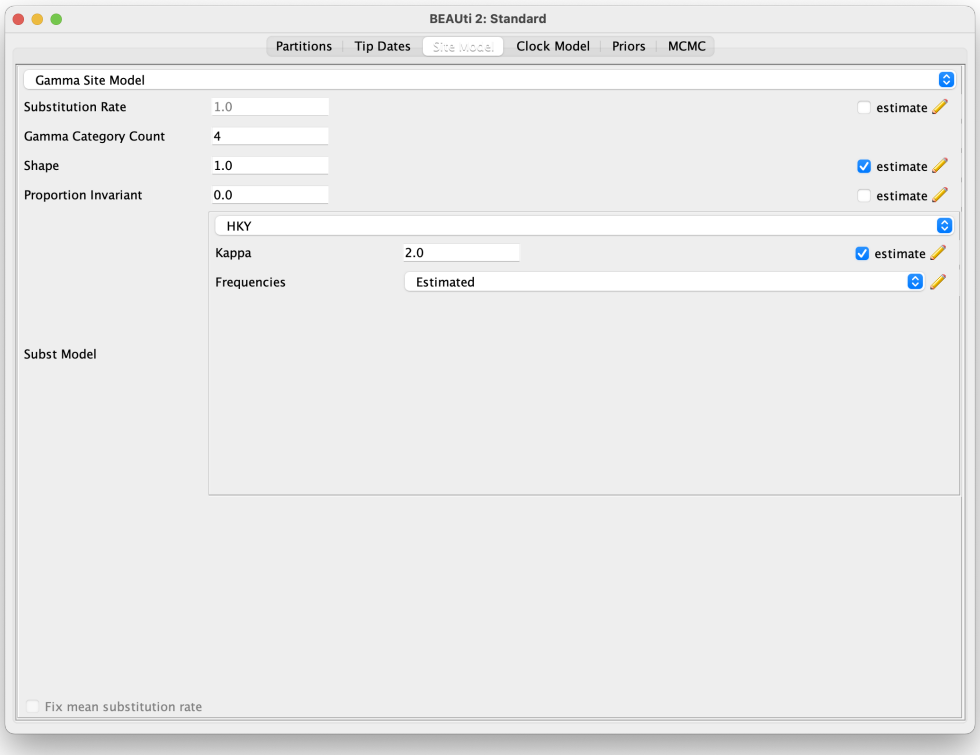

Figure 6: Setting up the site model

4. Next, select the Clock model tab and ensure that the strict clock model is selected. This is the default, so there is probably no need to change anything. It should like the same as in Figure 6.

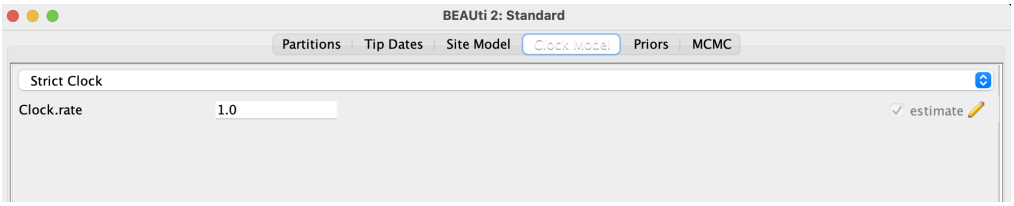

Figure 7:

5. Select the Priors tab. For the tree prior, select the **Coalescent Exponential Population** as shown in Fig 7. Note two key parameters here:
   - 'ePopSize:t.NorthAm', the scaled effective population size ϕ
   - 'growthRate:t.NorthAm', the growth rate r

   These are the key Coalescent parameters defined above. There others are relevant to the substitution model and clock rate. You can click on the arrow to the left to see each prior placed on the two parameters. The defaults are fine for our purposes here, but one should always check this before using them!

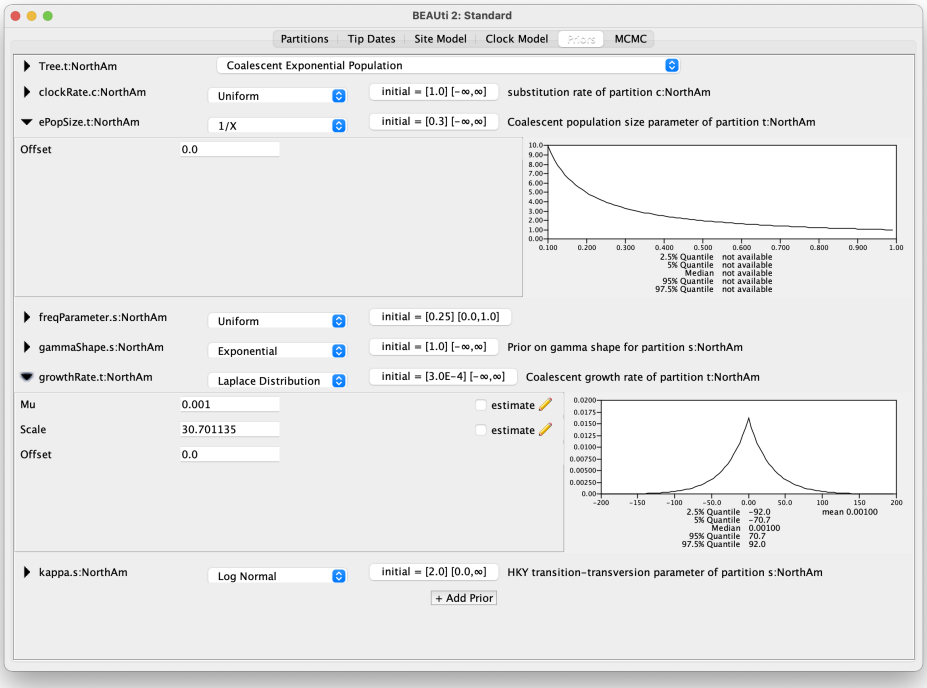

Figure 8: Priors for the exponential growth coalescent in BEAUTI.

6. Click on the MCMC tab. Make sure that your settings match those in Fig 8. be sure to rename the .log and .trees outputs.

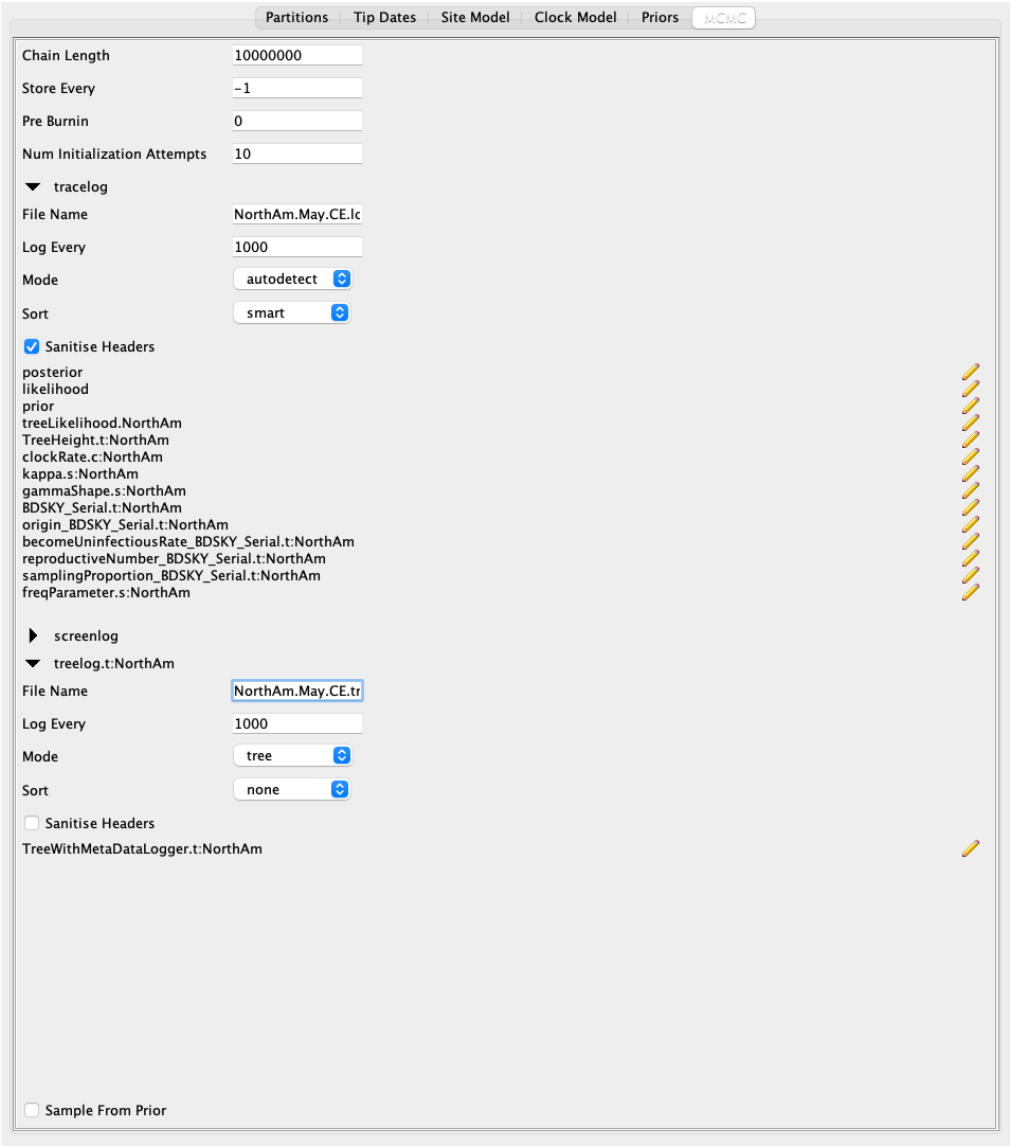

Figure 9: Setting MCMC specifications and output file names.

7. Go to File, Save and name the file NorthAm.May.CE.xml. Run it in BEAST as in earlier workshops. The analysis will take about 20 minutes. While that runs, we'll move onto the birth death and set that up too.

3.3 Birth Death

As before, let's start with a brief recap of the birth-death as is relevant to our task. The birth death allows for stochastic population growth, but it is still exponential on average. It has three parameters: the transmission rate λ, the becoming uninfectious rate δ, and the sampling proportion p. In practice though, BEAST parametrises it with R_0, δ, and p because these are easier parameters to think of. Just like the coalescent, we can get R_0 (given), doubling time (t_d), and estimate number of cases at the time of the final sample $I(T)$. Recall that the birth-death works forwards in time, so $I(T)$ is identical to $I(0)$ under the coalescent.

$$r = \delta R_0 - \delta \tag{5}$$

$$I(T) = e^{rT}, \text{ where } T \text{ is the age of the root} \tag{6}$$

$$t_d = \frac{\ln 2}{r} \tag{7}$$

3.3.1 Setting Up the Birth-Death

1. Steps 1-4 are identical here because they don't pertain to swapping from coalescent exponential to birth death tree priors. You can skip to step 2.
2. Go back to the Priors tab. For the tree prior, select the **Birth Death Skyline Serial** as shown in Fig 9. Note three key parameters here:
   - becomeUninfectiousRate_BDSKY:t.NorthAm, δ
   - 'reproductiveNumber_BDSKY:t.NorthAm', R_0
   - 'samplingProportion_BDSKY:t.NorthAm', p

   These are the key birth-death parameters defined above. There others are relevant to the substitution model and clock rate. You can click on the arrow to the left to see each prior placed on the two parameters. From here, we will set priors on δ and R_0.

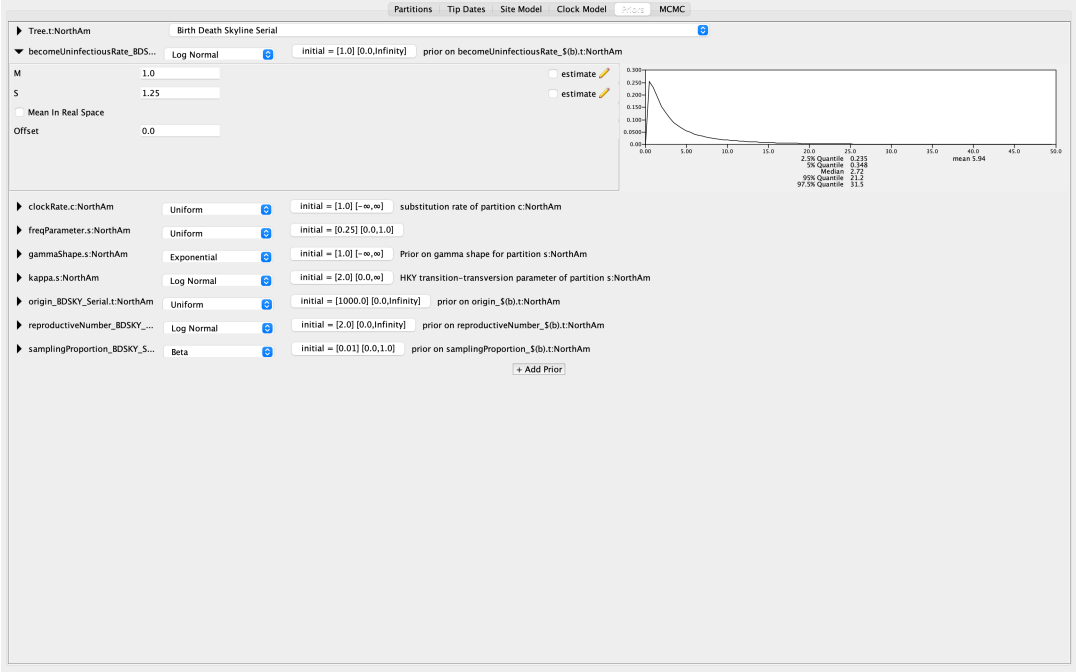

Figure 10: Priors for the birth-death in BEAUTI.

3. Note will place a prior on δ. We know the duration of infection for flu (D) is about 2.6 days, which corresponds to $\frac{365}{2.6} \approx 140$ infections in a year - i.e. the becoming-uninfectious rate (Cauchemez et al. 2009). The range of D is from 2-8 days, placing a likewise placing a range on δ of $[\frac{365}{8}, \frac{365}{2}] = [46.625, 182.5]$. To capture this, we will place a $Normal(\mu = 140, \sigma = 1.3)$ prior on δ to capture its central-tendency, as in Figure 10. Strictly speaking, a normal distribution isn't appropriate because its domain is $(-\infty, \infty)$ while δ cannot be negative. But, our prior with mean 140 and standard deviation 1.3 means it places almost no probability on negative values, so we can get away with its

simplicity here. It's necessary to have this narrow prior on δ because we usually need one or two narrow priors on birth-death parameters to get sharp estimates on our parameter of interest - R_0 in this case.

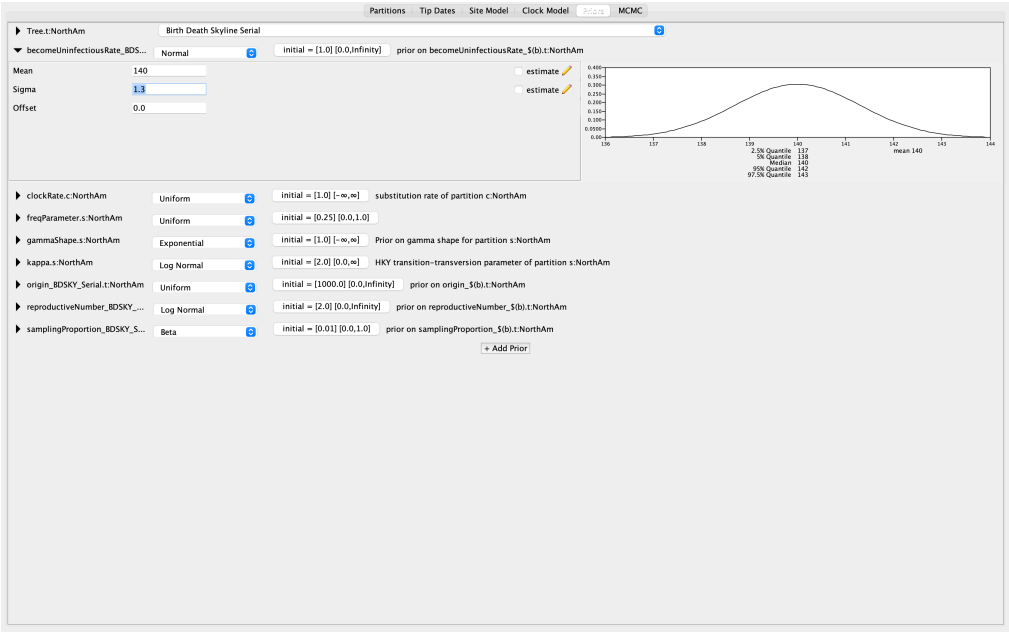

Figure 11: Placing a prior on δ

The remaining priors are fine, but it is important to inspect them. For example look at the prior for the sampling proportion, which has a beta distribution as a default. It is fairly flat between 0 and 1, which is appropriate for our data because we have no information of the sampling strategy. Since we want to infer R_0, and hence don't want to bias it with narrow prior, and have little information on p, it makes sense that we used a narrow prior on δ to compensate.

4. The last model-related task we have to do is change the 'Dimension' of R_0 as in Figure 11. The Dimension is the number of estimates of R_0 along the tree. You see, the constant rate birth-death that we are considering here is really just the simplest case of the birth-death-skyline, where we allow parameters to change in a series of sequential windows along the tree. The default is 10, meaning that for each proposed tree, the timespan is cut into 10 intervals where individual estimates of R_0 are made. Since we are only considering the exponential phase here, we only need one estimate of R_0, so we click on the 'initial' table next to the reproductiveNumber prior and change the dimension to 1.

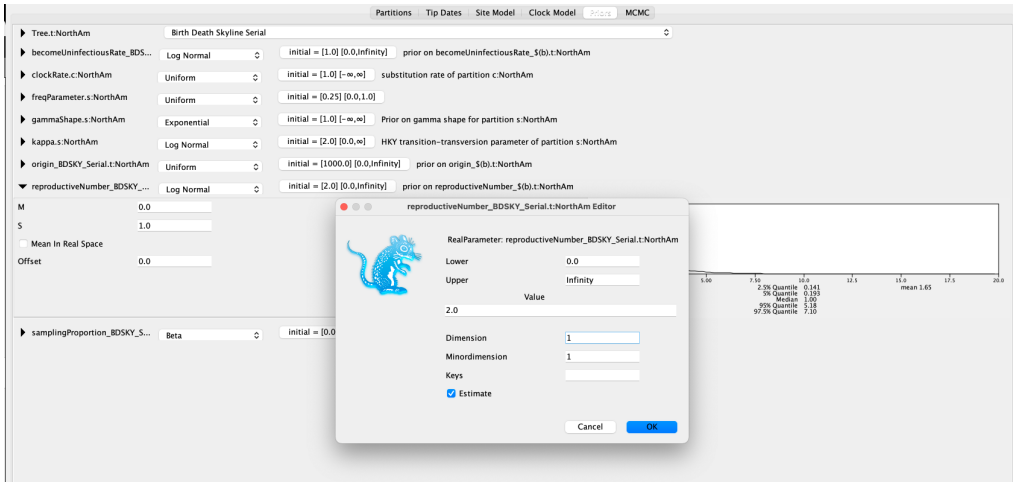

Figure 12: Changing the R_0 dimension to 1

5. And finally we can navigate to the MCMC tab to change the name of the .log and .trees output files, as in Figure 12. All the other default setting will be fine here. If you're doing this for your own dataset and you're having trouble getting high enough ESS values, then it's best to make your chain longer AND sample less frequently from it. Save as NothAm.May.BD.xml

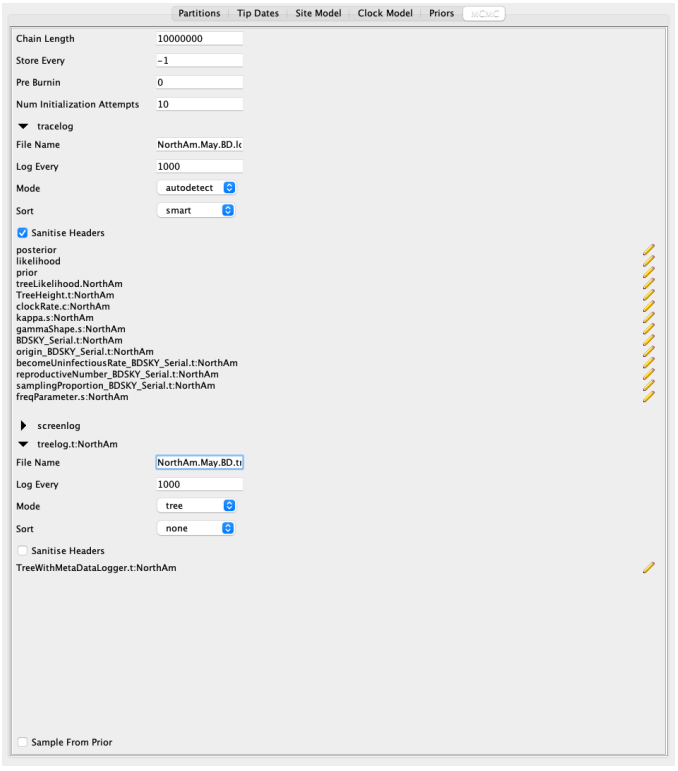

Figure 13: Setting MCMC specifications and output file names.

3.4 Exercises

3.4.1 Coalescent Exponential Exercises

To see the output of the MCMC, load the NorthAM.May.CE.log file into tracer as in Figure 13

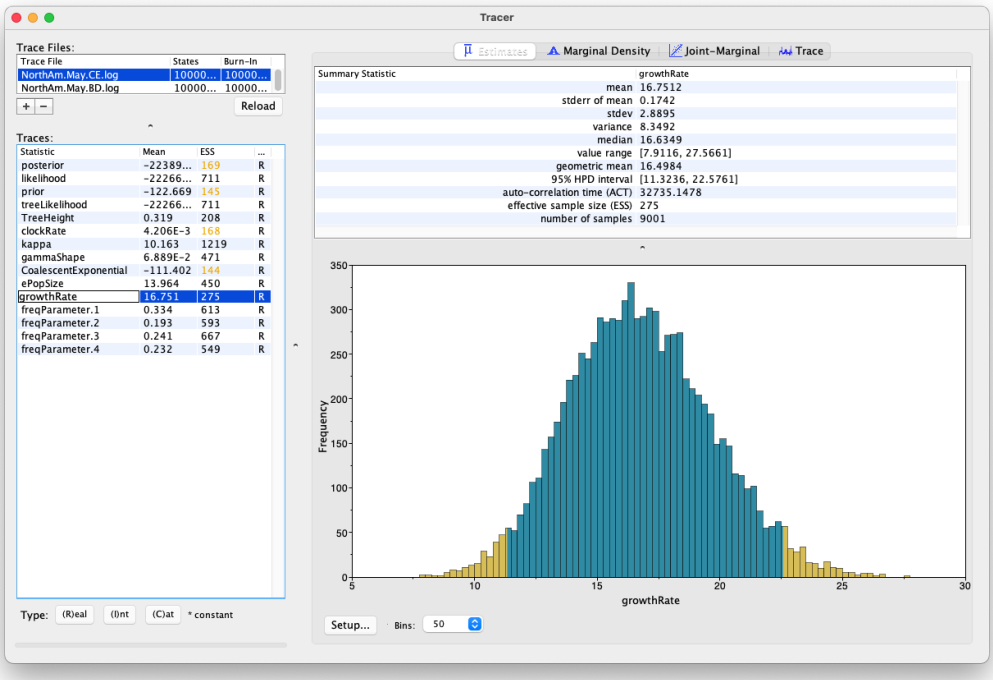

Figure 14: Using Tracer

1. What is the substitution rate for this data set?

2. What is the age of the root of the H1N1 data set? (hint: you can subtract the TreeHeight trace in the log file from the age of the most recently collected sample, 2009.414)

3. Do the same calculations for the upper and lower 95% credible interval (or HPD) of the growth rate. What is the credible interval of R_0? Are they consistent with the disease tending to spread at this stage? (hint: $R_0 > 1$ for a disease to continue spreading)
   We can now do some algebra to estimate R_0. We mentioned earlier that the duration of infection ranges from 2 to 8 days, but the mean is closer to about 2.6 days. Here I will use my estimates of growth rate and effective population size, but yours might vary slightly:

$$\text{Duration of infection in years: } = \frac{2.6}{365} = 0.0071 \text{ (in years)}$$

$$\text{Becoming uninfectious rate: } \delta = \frac{1}{0.0071} = 140.85$$

$$\text{Mean growth rate: } r = 16.75 = \lambda\text{-}\delta$$

$$\lambda = 16.75 + 140.85 = 157.6$$

$$R_0 = \frac{\lambda}{\delta} = \frac{157.6}{140.85} = 1.12$$

4. Can you use the estimates above and the equations a the start of this document to estimate the number of infected individuals in May? (hint: use r)

3.4.2 Birth Death Exercises

To see the output of the MCMC, load the NorthAM.May.BD.log file into tracer as in Figure 14. You can drag and drop it as you did for the coalescent .log file.

5. Does the reproductive number R_0 from this model match what we obtained for the exponential growth coalescent?

6. What is the sampling proportion for this analysis? Is this estimate reasonable?

7. The birth and becoming uninfectious rates are equivalent to the infection rate λ and δ in the exponential coalescent. Do these match our estimates above?

8. One of the parameters in this model is the origin of the outbreak, which is the time when the outbreak started. How much earlier is it from the age of the root of the tree?

9. Is the age of the root of the tree in the birth-death model similar to that from the exponential growth? (hint: load both log files in Tracer, select them both, and select TreeHeight. The distributions can be compared using the tab Marginal Prob Distribution.

4 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`

Relevant References

Cauchemez, S, CA Donnelly, C Reed, AC Ghani, C Fraser, CK Kent, L Finelli, and NM Ferguson. 2009. Household transmission of 2009 pandemic influenza a (h1n1) virus in the united states. *New England Journal of Medicine* 361: PMID: 20042753, 2619–2627.

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

Hedge, J, SJ Lycett, and A Rambaut. 2013. Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology letters* 9: 20130331.

Kühnert, D, T Stadler, TG Vaughan, and AJ Drummond. 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death sir model. *Journal of The Royal Society Interface* 11: 20131106.