**Practical 3a: Phylodynamics in BEAST2 using coalescent and birth-death models**

For the remaining of the workshop we will use two sequences data sets of the 2009 H1N1 influenza pandemic. The first sequence data set consists of sequences collected in North America from the start of the pandemic to May 2009 (Hedge et al. 2013). At this time, we expect that the pandemic has been growing exponentially, such that we can use the sequence data to estimate some epidemiological parameters of interest.

**The exponential growth coalescent**

The exponential growth coalescent model has two key parameters in its implementation in BEAST2; the growth rate (*r*) and the effective population size (*Φ*). The *r* determines how quickly the population size is increasing, and the *Φ* is proportional to the number of infected individuals at present. These two parameters can be expressed in terms of the infection rate (*λ*) and the rate of becoming uninfectious (*δ*):

*r* = *λ – δ*

and

*Φ =* # of infected individuals at present / 2\* *λ*

The basic reproductive number (*R0*) is *λ / δ.* As such, to estimate this parameter using *r* and *Φ* we require additional information, such as an estimate of the current number of infected individuals. We can also use information about the average duration of the infection, which ranges between 2 to 8 days for flu (Cauchemez et al. 2009) (Note that this is the period over which one is infectious, and not necessarily the time over which one suffers from symptoms). In fact, the duration of infection is 1/ *δ*, so we can do some simple algebra to estimate *R0*.

*λ = r – 1/*duration of infection

*R0 = λ / (1/duration of infection)*

**The constant birth-death**

The constant birth-death model describes a similar process of exponential growth in the population (see Stader et al. 2011 and Boskova et al. 2014 for more details). It has the following key parameters in BEAST2; *R0*, *λ*, and *δ.* However, *δ* is known to involve death (or recovery) and the sampling rate (assuming that sampling coincides with recovery). As such, it is can be parameterised as:

*δ = μ - ψ*

where *μ* is the death rate and *ψ* is the sampling rate. In this model, we typically include information about the sampling rate or sampling proportion (*p*; *p* = *ψ / δ*), to estimate *R0*.

In practice, for the exponential growth coalescent and the constant birth-death the tree likelihood depends on two compound parameters; *λ – δ*, and *λ\*δ\*p.* This implies that we will always need to include additional information about *λ, δ,* or *p* to estimate *R0*. Similar to what we did with the exponential growth coalescent, we will include information about the duration of infection to obtain *R0*  because we have no independent information about the sampling strategy for these data.

**Estimating *R0* with the exponential growth coalescent**

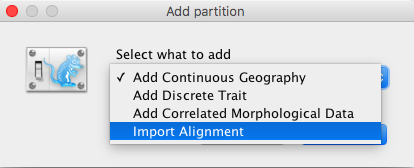
**Data set**

* Sequence alignment in fasta format of samples collected in the early stages of the 2009 H1N1 pandemic in North America: NorthAm.May.fasta

**Software**

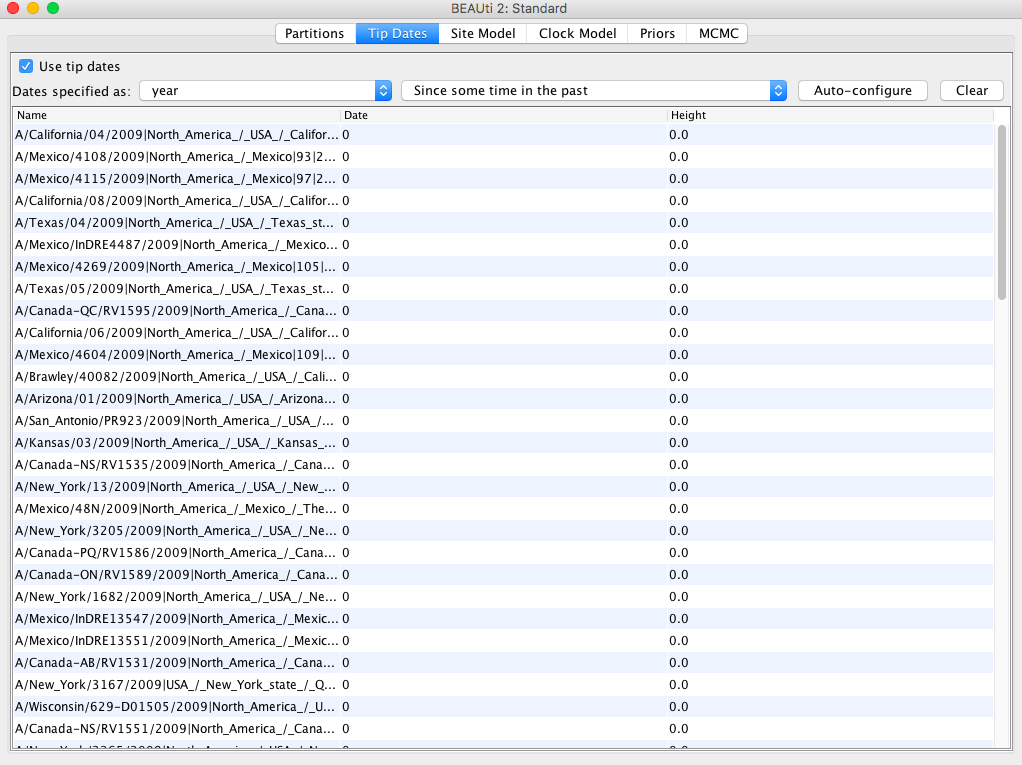
* BEAST 2.4 (beast2.org)

**1.** Open BEAUTI and drag and drop the alignment file. You might get a pop-up window asking you what kind of data this is (Fig 1). Select ‘import alignment’ and click ‘OK’. You might get an other prompt asking you to select aminoacid or nucleotide sequences. Select nucleotide and click ‘OK’.

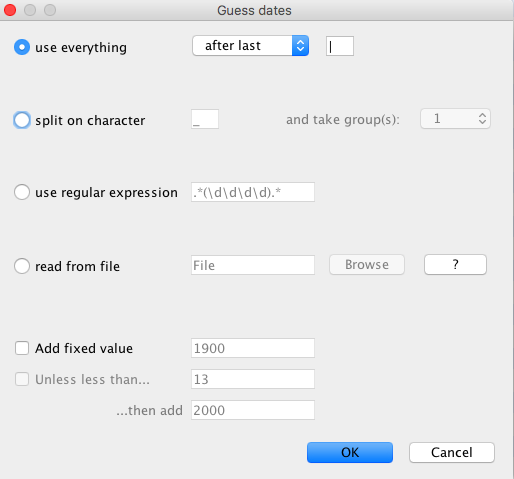


**Fig 1.**

**2.** The data will appear in the ‘Partitions’ tab. It should correspond to 100 taxa and 13,154 sites. Click on the ‘Tip Dates’ tab. You should see a blank window with a  box. Check the box, and you should see the table in Fig 2 appear. Click on ‘Auto-configure’. You should a window to set the dates, as shown in Fig 3. Select ‘use everything’, ‘after last’ and type in a | symbol. Click OK. The table from Fig 2 should be updated to list the sampling times for all the samples.

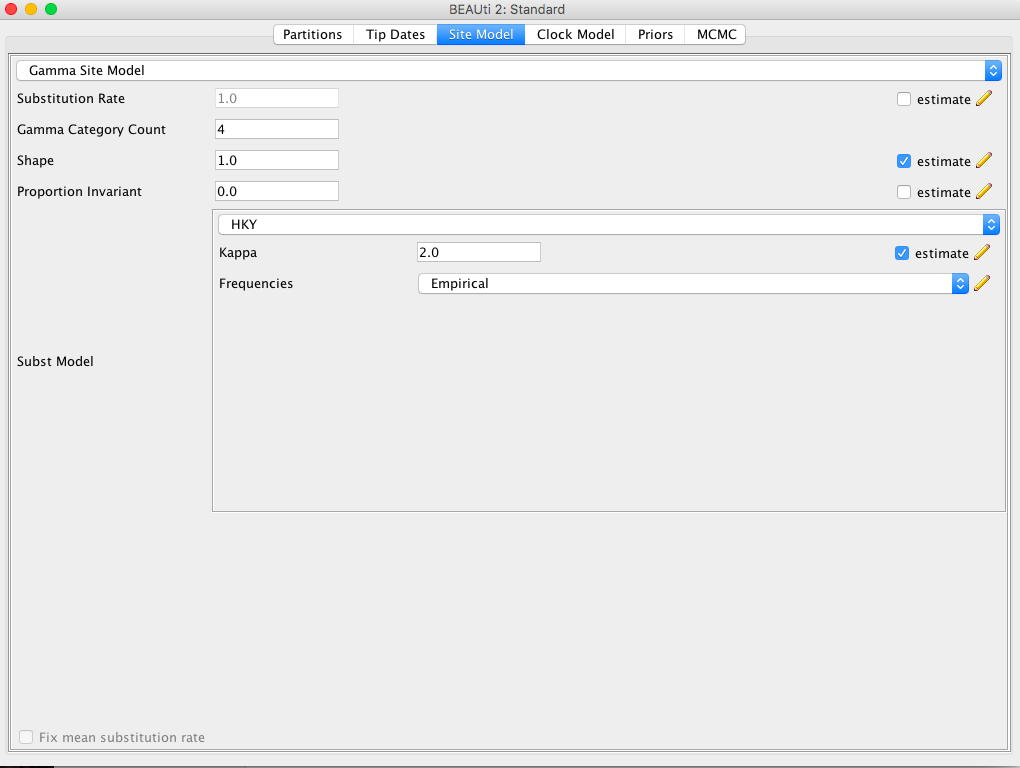


**Fig 2.** Tip dates tab in BEAUTI.



**Fig 3.** Specifying tip dates in BEAUTI.

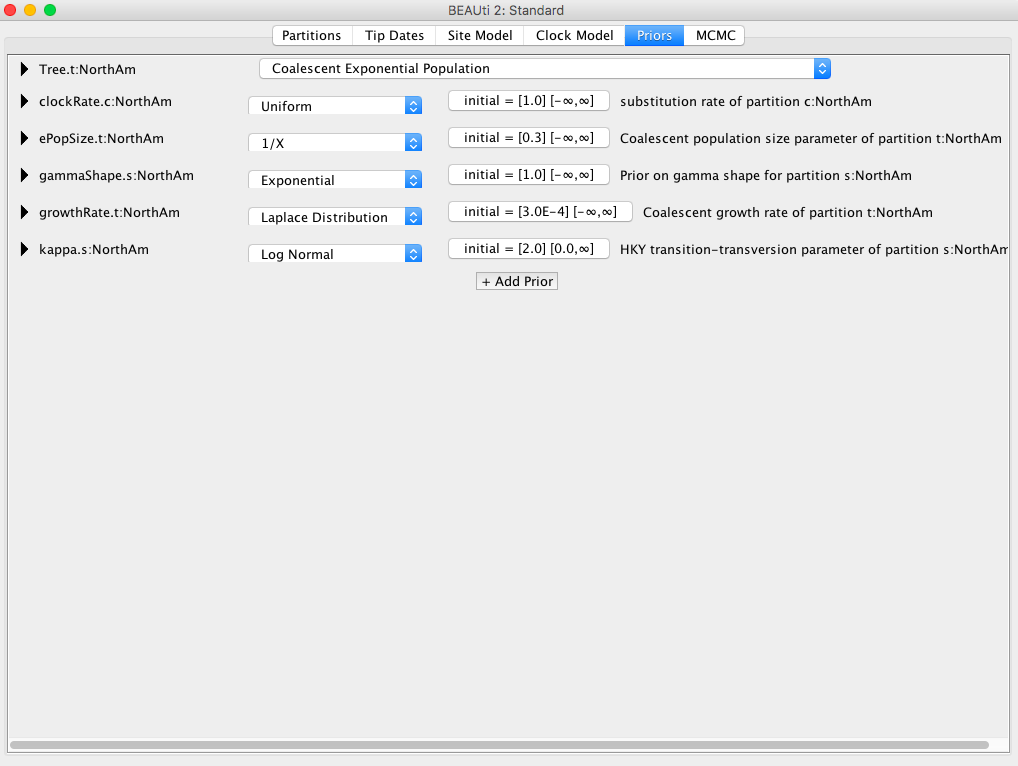
**3.** Click on the ‘site model’ tab. Select the HKY+G model as shown in Fig 4. This model will account for rate heterogeneity among sites and for the transition to tranversion bias.



**Fig 4.** Setting up the HKY+G model in BEAUTI.

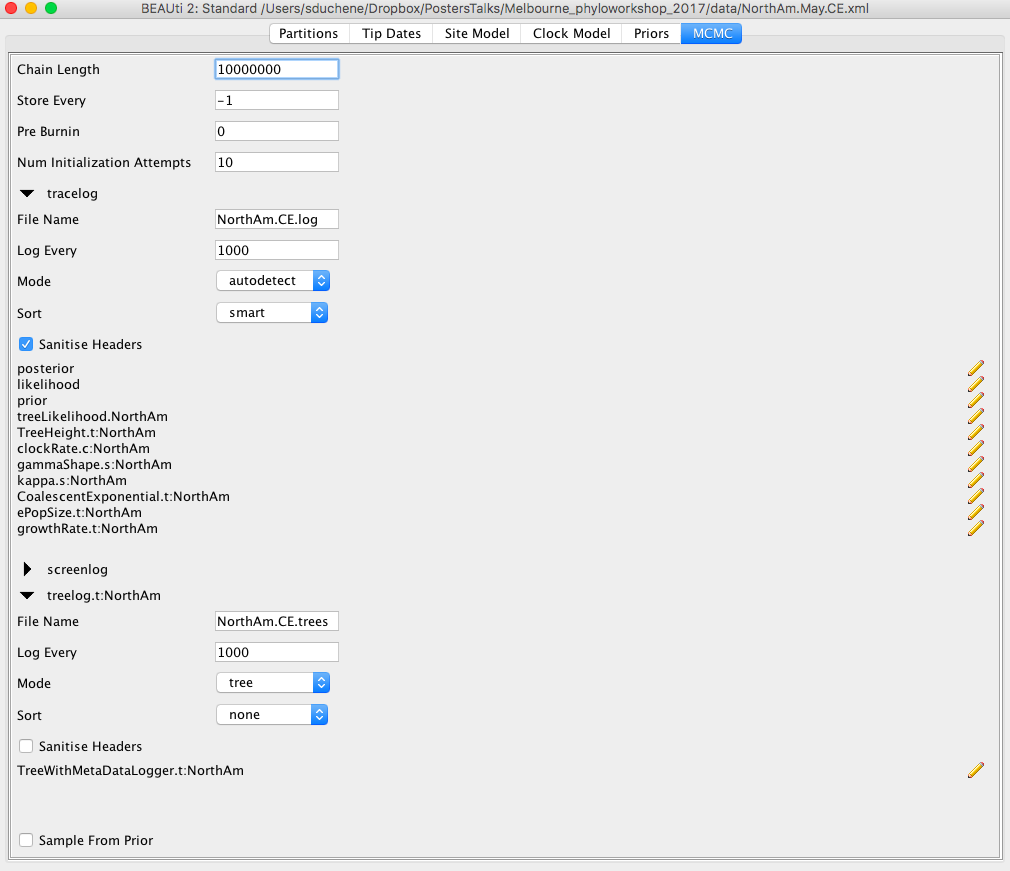
**4.** Next, select the ‘Clock model’ tab and ensure that the strict clock model is selected. This is the default, so there is probably no need to change anything.

**5.** Select the ‘Priors’ tab. For the tree prior, select the ‘Coalescent Exponential Population’ as shown in Fig 5. The rest of the priors are fine for this analysis.



**Fig 5.** Priors for the exponential growth coalescent in BEAUTI.

**6.** Click on the MCMC tab. Make sure that your settings match those in Fig 6.

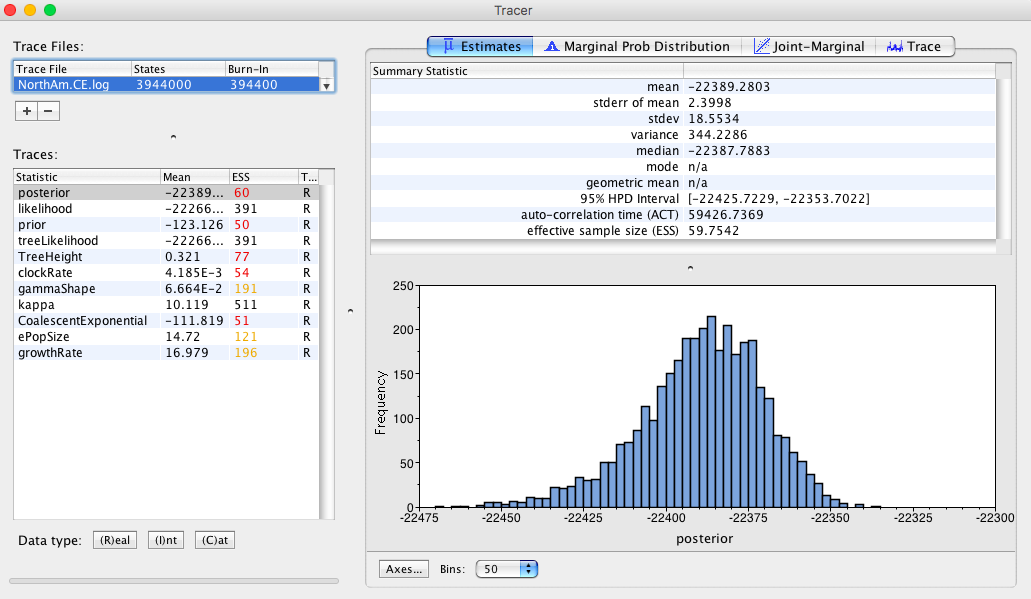


**Fig 6.** MCMC settings for the exponential growth coalescent in BEAUTI.

Go to ‘File’, ‘Save’ and name the file NorthAm.May.CE.xml. Run it in BEAST like as we did in the molecular dating prac.

The analysis will take about 20 minutes. Instead of waiting, go the next section (the constant birth-death model) and return here when the analysis is complete.

**7.** Open Tracer and drag and drop the log file (NorthAm.May.CE.log) to the panel on the left. It should look something like Fig 7, but note that your ESS values should be much higher, as this is a shorter version of the run.



**Fig 7.** Tracer with the exponential growth analysis of H1N1.

**Question 1:** What is the substitution rate for this data set? How does it compare to what we estimated for Ebola yesterday?

**Question 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)

We can now do some algebra to estimate *R0*. 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:

Duration of infection in years = 2.6 days / 365 days = 0.0071 (in years)

Become uninfectious rate (*δ*) = 1 / 0.0071 = 140.85

Mean growth rate (*r*) = 16.98 = *λ – δ*

*λ =* 16.98 + 140.85 = 157.83

*R0*  = *λ / δ =* 157.83 */* 140.85 = 1.12

**Question 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 *R0*? Are they consistent with the disease tending to spread at this stage? (hint: *R0* > 1 for a disease to continue spreading)

**Question 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 *Φ =* # of infected individuals at present / 2\* *λ*)

**Estimating *R0* with the constant birth-death model**

**Data set**

* Sequence alignment in fasta format of samples collected in the early stages of the 2009 H1N1 pandemic in North America: NorthAm.May.fasta

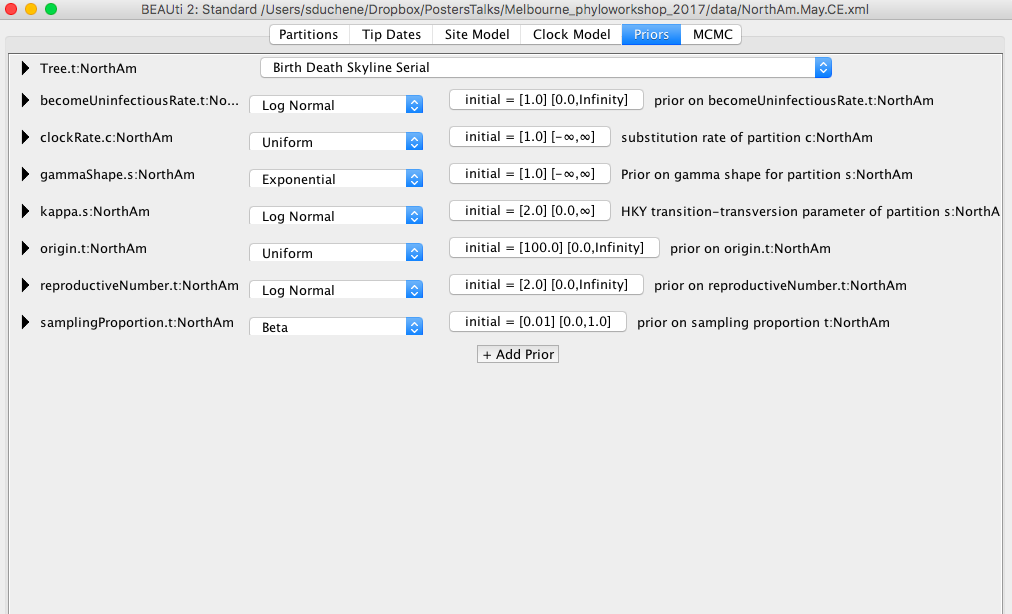
**Software**

* BEAST 2.4 (beast2.org)
* Phylodynamics addon

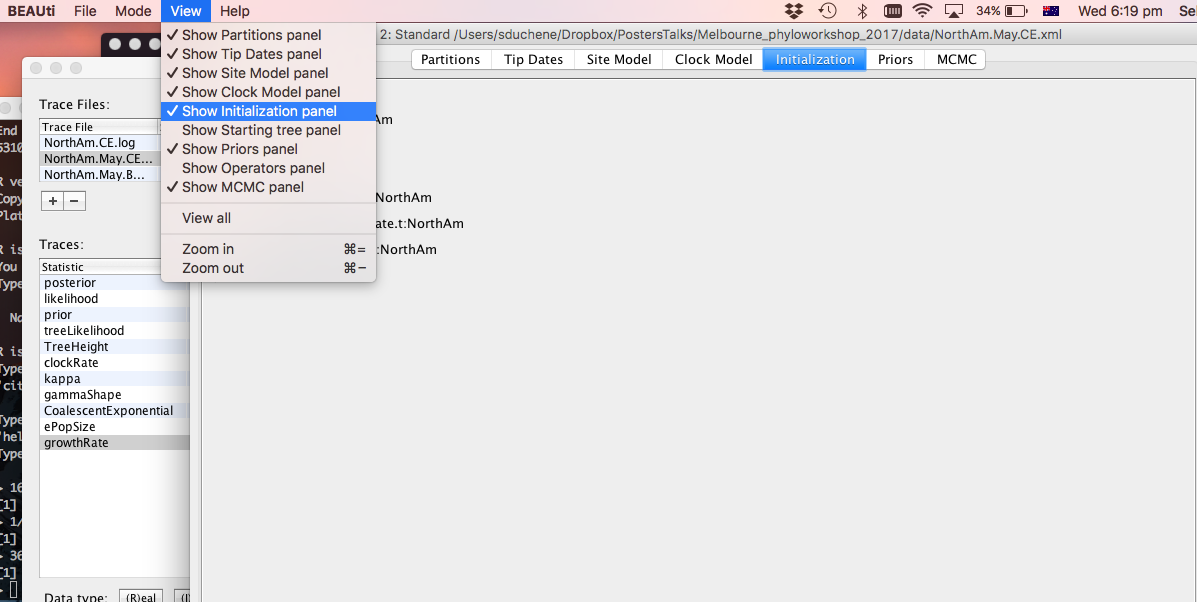
**1.** Follow steps 1 through 4 in the previous section. You sholud have set up tip dates, the site model, and the clock model.

**2.** Select the ‘Priors’ tab. Select the ‘Birth Death Skyline Serial’ model for the tree prior (Fig 8).

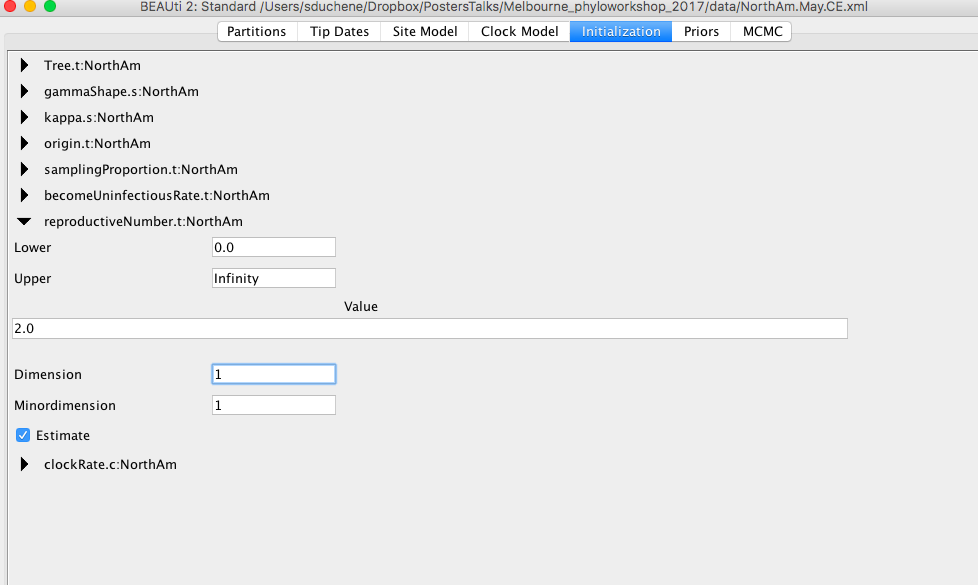
**3.** In the ‘View’ menu above the window, select ‘Show initialization panel’ (Fig 9). This will take you to the ‘Initalization’ panel. Click on the arrow next to reproductiveNumber, and change the dimensions from 10 to 1 (Fig 10). This is because we are using the birth-death skyline model, but we want to use a single skyline interval, such that the model collapses to a constant birth-death.



**Fig 8.** Selecting the birth death skyline model in BEAUTI.



**Fig 9.** Selecting initialisation panel in BEAUTI.



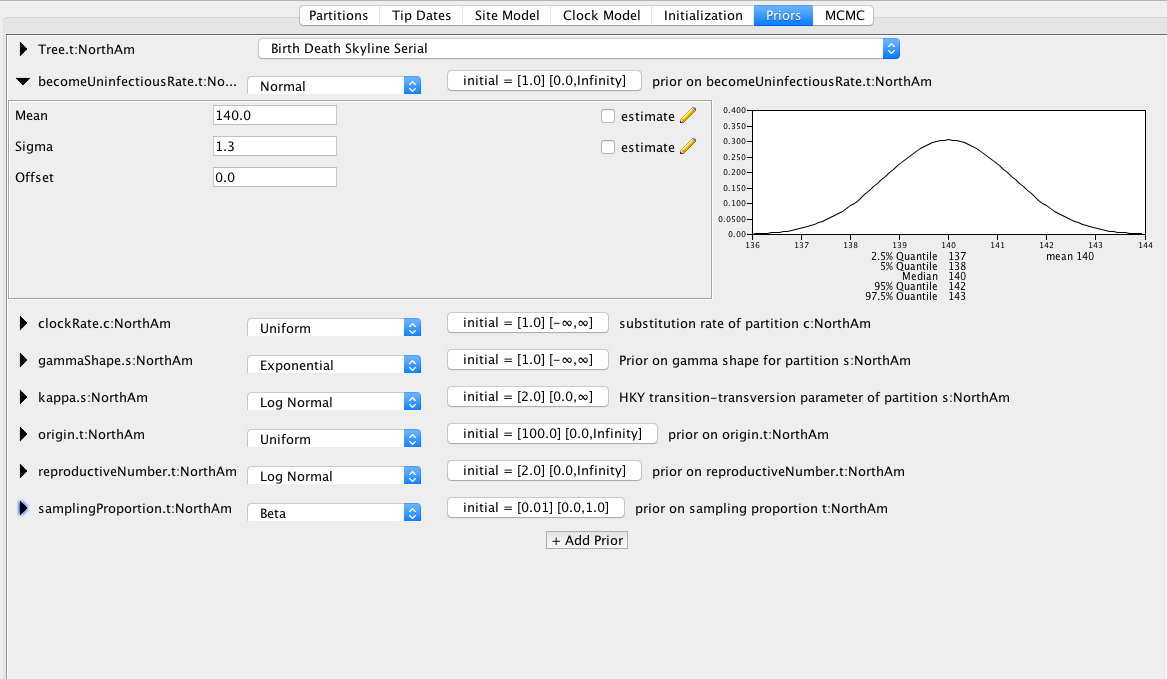
**Fig 10.** Selecting one dimension for the reproductive number in the birth death skyline.

**4.** Go back to the ‘Priors’ tab. We will leave some priors with their default values, but we will need to change one as follows:

**becomeUninfectiousRate:NorthAm**

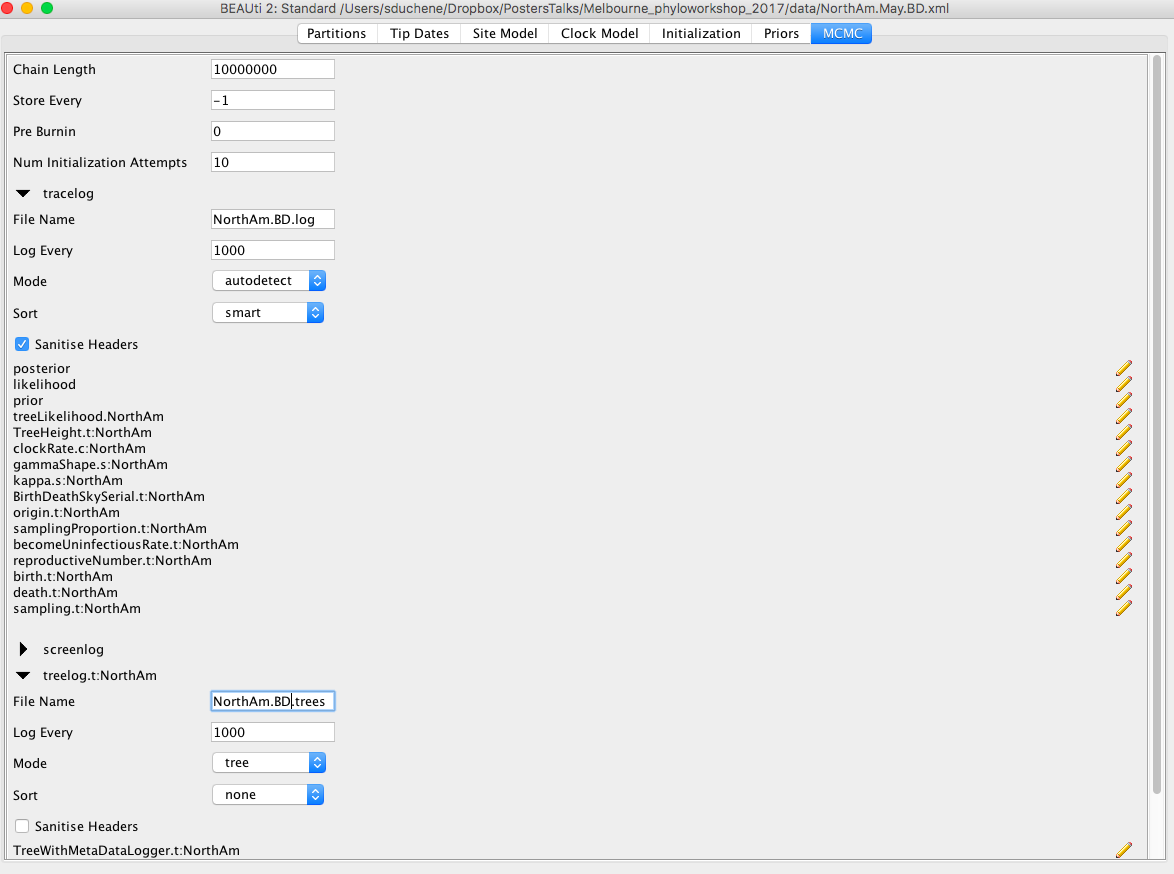
This is similar to the *δ* parameter discussed earlier. Use a normal distribution with mean of 140 and standard deviation of 1.3. This roughly corresponds to an infectious period of 2.6 days (Fig 11).

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.



**Fig 11.** Setting the becomeUninfectiousRate prior.

**5.** Select the MCMC tab and set it as shown in Fig 12.



**Fig 12.** MCMC set up for the birth-death model.

**6.** Save the file by clicking ‘File’ and ‘Save’. Name it NorthAm.May.BD.xml.

**7.** Run this analysis in BEAST as you have done in the previous section.

**8.** Open the log file in Tracer to answer questions 5 through 8.

**Question 5:** Does the reproductive number (*R0*) from this model match what we obtained for the exponential growth coalescent?

**Question 6:** What is the sampling proportion for this analysis? Is this estimate reasonable?

**Question 7:** The birth and become uninfectious rates are equivalent to the infection rate (*λ*) and *δ* in the exponential coalescent. Do these match our estimates above?

**Question 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?

**Question 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’.

**Question 10:** Use tree annotator as you have done in the molecular dating prac to obtain summary trees for the exponential growth coalescent and the birth-death models. Are there any major differences in the trees?

**References**

Boskova, V., Bonhoeffer, S., & Stadler, T. (2014). Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLoS computational biology*, *10*(11), e1003913.

Cauchemez, S., Donnelly, C. A., Reed, C., Ghani, A. C., Fraser, C., Kent, C. K., ... & Ferguson, N. M. (2009). Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *N Engl J Med*, *2009*(361), 2619-2627.

Hedge, J., Lycett, S. J., & Rambaut, A. (2013). Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology letters*, *9*(5), 20130331.

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., ... & Günthard, H. F. (2011). Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution*, *29*(1), 347-357.