January 30th: Week 4 Tutorial

***Objective*:** After today's lecture and lab, you should understand the concepts of how gene trees differ from species trees. You should also know how to perform a species tree analysis in the program ***BEAST** by starting with sequence alignments for multiple genes. Just as in last week, you should be able to determine if your Bayesian analysis has **converged** by using the program **Tracer**.

Lab Assignment

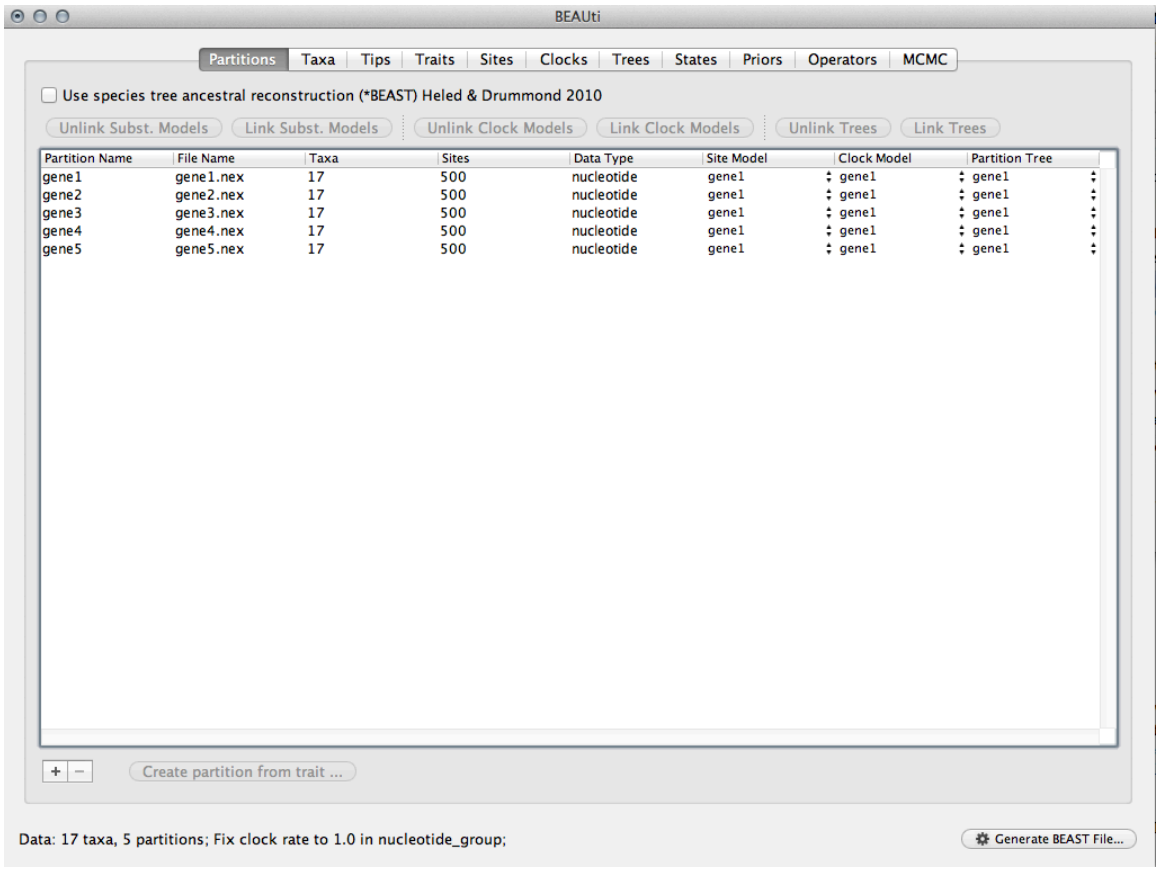
We are going to infer the species tree, also known as the population-level history, for great apes, given multiple independent loci. Maximum likelihood approaches exist to infer the species tree, but we will be using a Bayesian program today. The program, ***BEAST**, is part of the **BEAST** package (stands for Bayesian Evolutionary Analysis by Sampling Trees), and is currently one of the most popular programs for inferring the species-level phylogeny.

Before we begin, download the (5) single gene alignments from the Google Drive folder (Week 4 > gene1-5.nex).

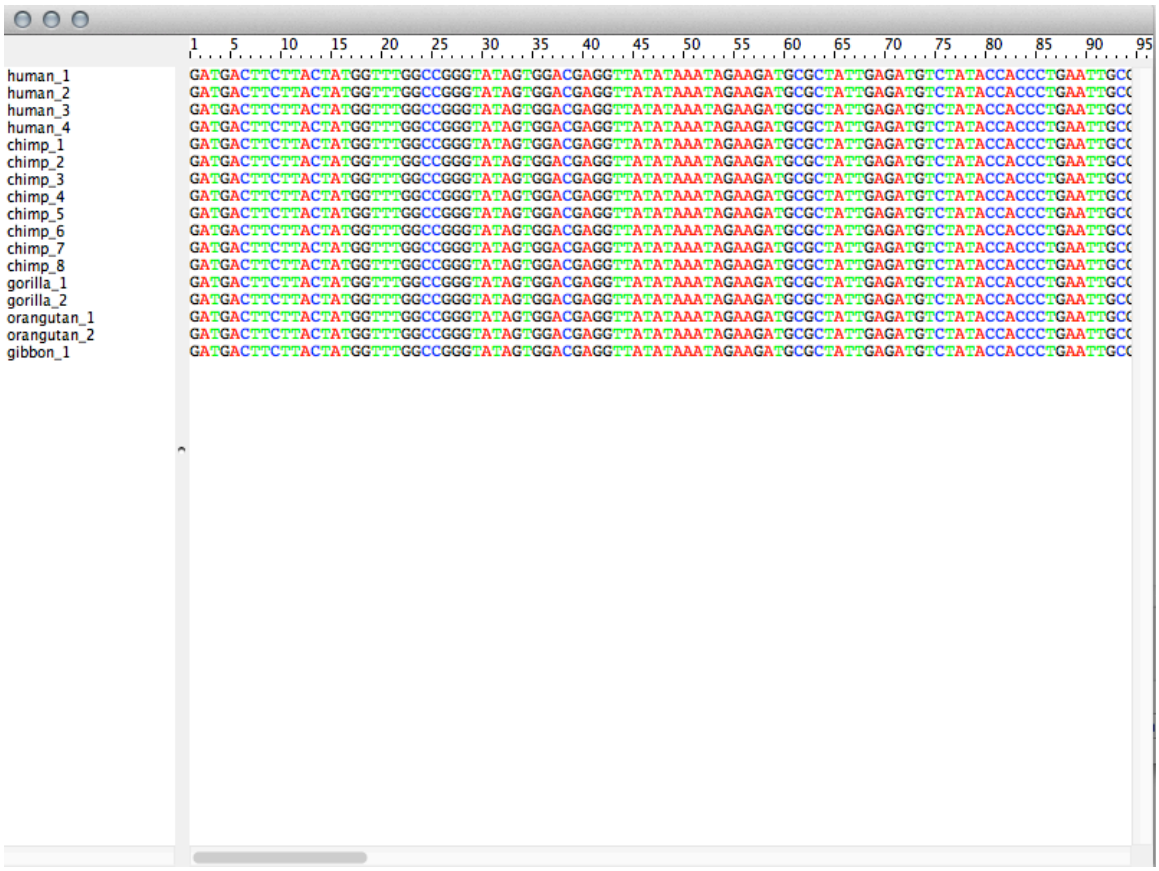
Before we get into *BEAST, some tips:

1) Taxon names must be identical across gene files (can verify in TextWrangler)
2) Each individual in the single gene alignments will need to be assigned to a species before analysis. Using the species name in the taxon label in a consistent way will make assigning samples to species easier.

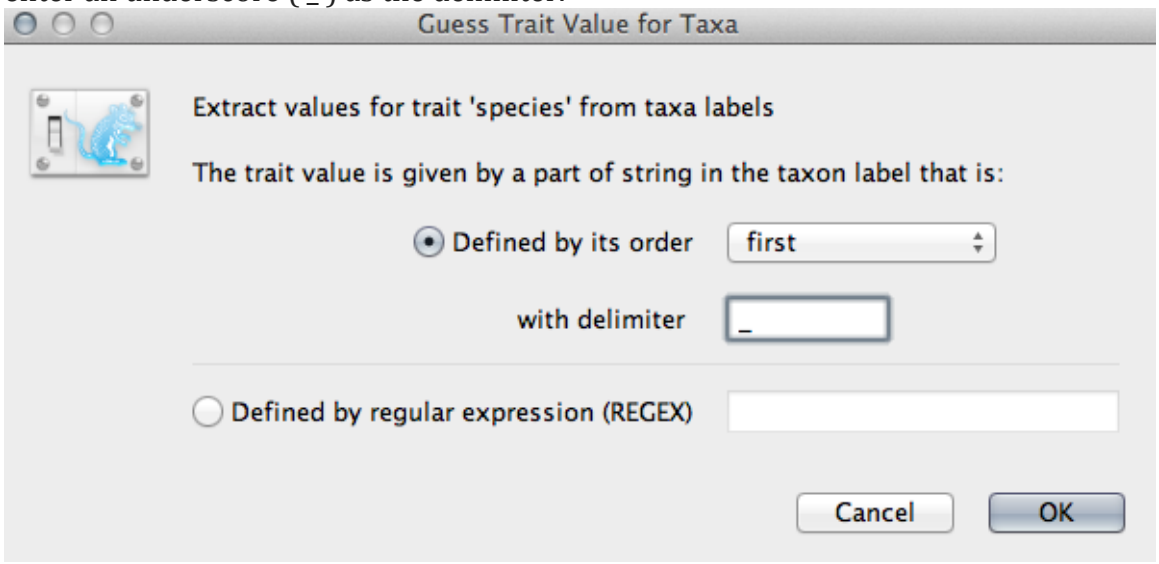
Now, to the programs! Although we are using the program BEAST to perform the phylogenetic analysis, we will use the program **BEAUti** to set up the analysis file. Go to Phylogenetic Analysis > BEASTv1.7.5 (not v1.6 !), and then double-click on the BEAUti program. Now, load your files by hitting the "+" button in the bottom left of the window. Alternatively, you can drag-and-drop your files into the main BEAUti window. At this point, you should see the # of taxa, sites, etc.



Next, double-click on a file name and examine what the data files look like. This is a good checkpoint to make sure your data are properly aligned (which you would have already done in another program!).

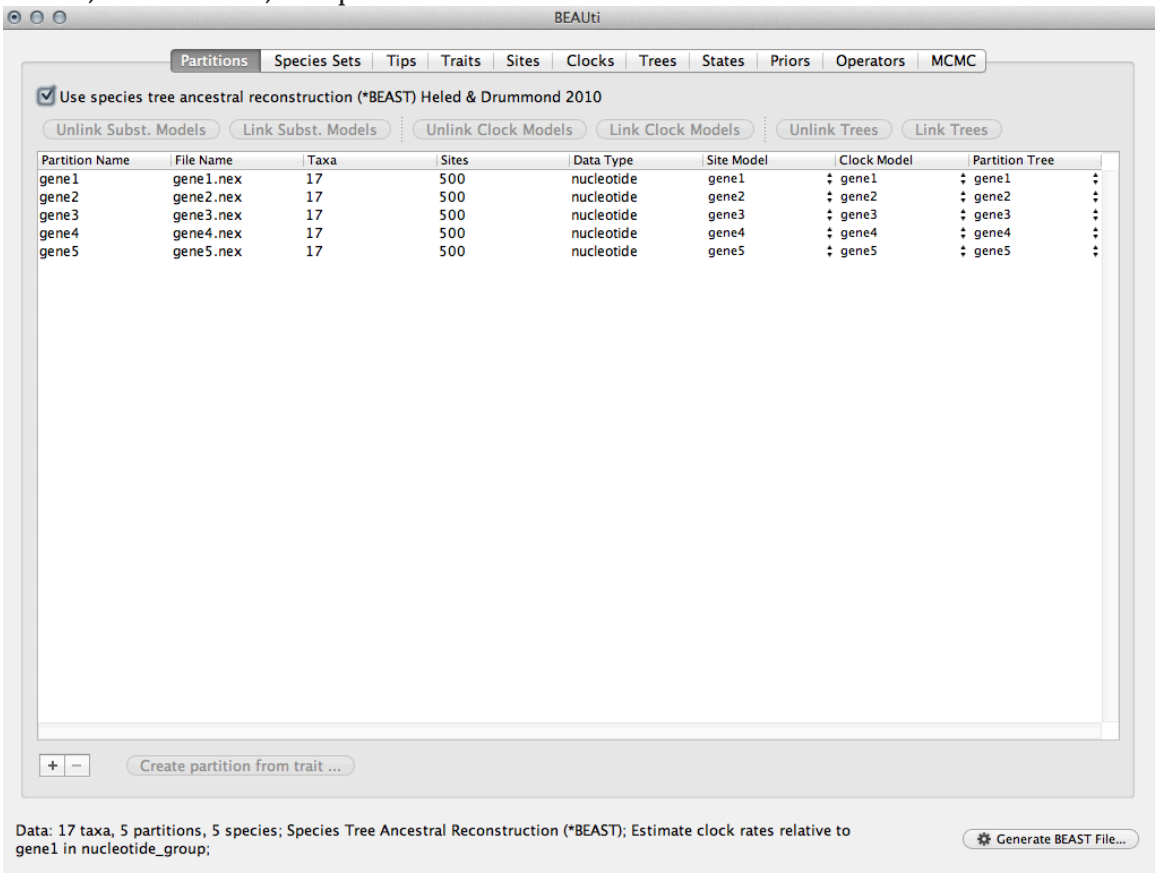


Now, click on the “Use species tree ancestral reconstruction (*BEAST) button in the top left. This will ask you to create or import traits; click OK. Now, you are in the “Traits” tab. Click the “Guess trait values” button just below the “Traits” tab, and enter an underscore (_) as the delimiter.



You will now see the species name as the “Value” for each taxon; this is the species assignment.

Now, go back to the “Partitions” tab and verify that "Site Models", "Clock Models", and "Partition Trees" are unlinked. If this is the case, gene will have its own site model, clock model, and partition tree.



Next, click the “Species Sets” tab. This is where you could add a constraint to the species tree.

Now, click on the “Sites” tab. In this tab you choose the model of substitution for each partition of your data (each gene). This is where you’d specify the model that jModelTest determined, but BEAST can only use a subset of the models that we tested in jModelTest. The default model of HKY will be appropriate for these data. These genes are simple and do not require a site heterogeneity model, but this is where you could specify a gamma or invariant sites model for site rate heterogeneity. Change the base frequencies to "All equal" for gene1, then do the same for the remaining 4 genes.

Click the “Clocks” tab. Clock models for these genes will follow a "strict clock" and we will estimate the rate for all genes with a fixed mean in the nucleotide group by checking the "Estimate" boxes for each gene, then the "Fix Mean" box in the sub-window below the Clock Model window.

Next, click on the “Trees” tab. The top boxes deal with the species tree prior and population size model. The species tree prior default is a Yule process and the population size will be Piecewise linear and constant root; **change** the Population Size Model to “Piecewise Constant”. The Tree model section is for each gene. You can designate Ploidy and what type of starting tree is desired. All genes in this analysis are nuclear, set the ploidy appropriately.

Now, click on the “Priors” tab. This tab lists all the parameters that have priors for parameters, which the user can change. The priors for the clock.rate parameter for each gene are not set by default and must be selected (thus they are listed in red!). Click on the clock.rate parameter in red. The popup box will allow you to choose an appropriate distribution for this prior. Explore the possible distributions to answer the following:

Questions to answer (more to follow below)
1.) Which priors are "improper priors"?

2.) What does "improper prior" mean?

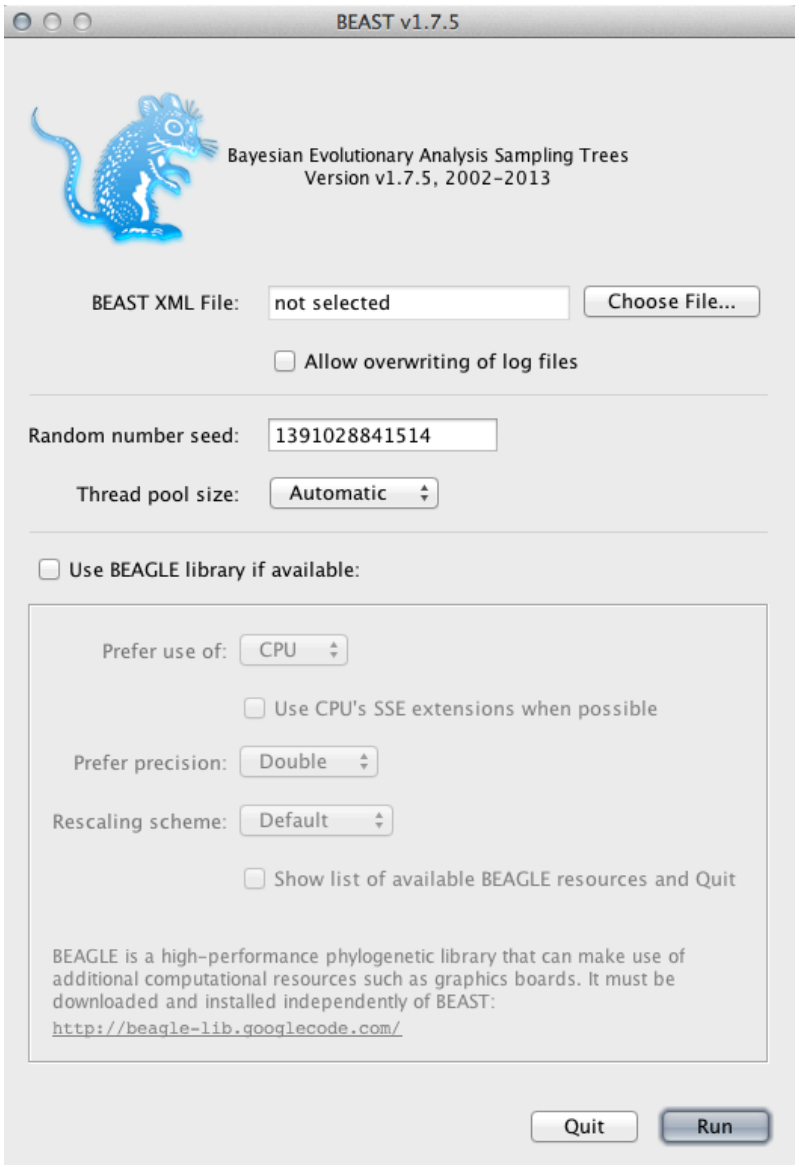
When done exploring, select the exponential prior with mean of 1.

3.) Why is the exponential an appropriate prior for the clock.rate?

Next, click on the “MCMC” tab. Leave the length of the chain at 10 million and change the “Echo state to screen every” to 5000. Choose a "Log parameters every" value that will result in 1000 samples from the posterior (do the math!). Change File name stem to your last name.

Now, your file is ready to go! Click the “Generate BEAST File” button in the bottom right. You will get a warning that some parameters are still default; this is OK. If you get other warnings, you have made an error in your settings and should go back and find the mistake. Save the .xml file to a new folder on the desktop labeled Lastname_run1.

To start your analysis, go back to the BEASTv1.7.5 folder and find the BEAST program. Execute the BEASTv1.7.5 application.



Click the "Choose File" button and locate your .xml file you just created, then click Run; this analysis should take ~20 minutes. Tell us if you have generated any errors while trying to execute your file. After the analysis begins, you will have six files that end in .trees, and one .log file; when the analysis is complete you should also have a .ops file. At any point during the run, you can look at the .log file in Tracer to see how your run is going.

While your analysis is going, we are going to load some files into Tracer from runs we previously performed for this same dataset. Go to the Week 4 folder in Google Drive, then go into the "Previously_Done" folder; download to your folder the two .log files. Open these two files in Tracer. You can select between the two files in the Trace Files box. There is also a "combined" listed that joins both posteriors together. You will notice, in the Traces tab, that nearly all parameters have converged in both independent runs and only one item in the ESS column is orange. These analyses

were run for 30,000,000 generations, 3x the length of the chain you are currently running, so we'd expect the ESS values to be better in these files. The MCMC run you are performing today may be too short to produce adequate posterior estimates for many of the parameters. A run of 30–40 million generations would be needed for convergence for a dataset of this size and complexity. It is important to run your analyses long enough to sufficiently sample the posterior and obtain ESS values of 200 or higher for all parameters. It is also important to run multiple independent analyses to ensure that your posterior sample is consistent.

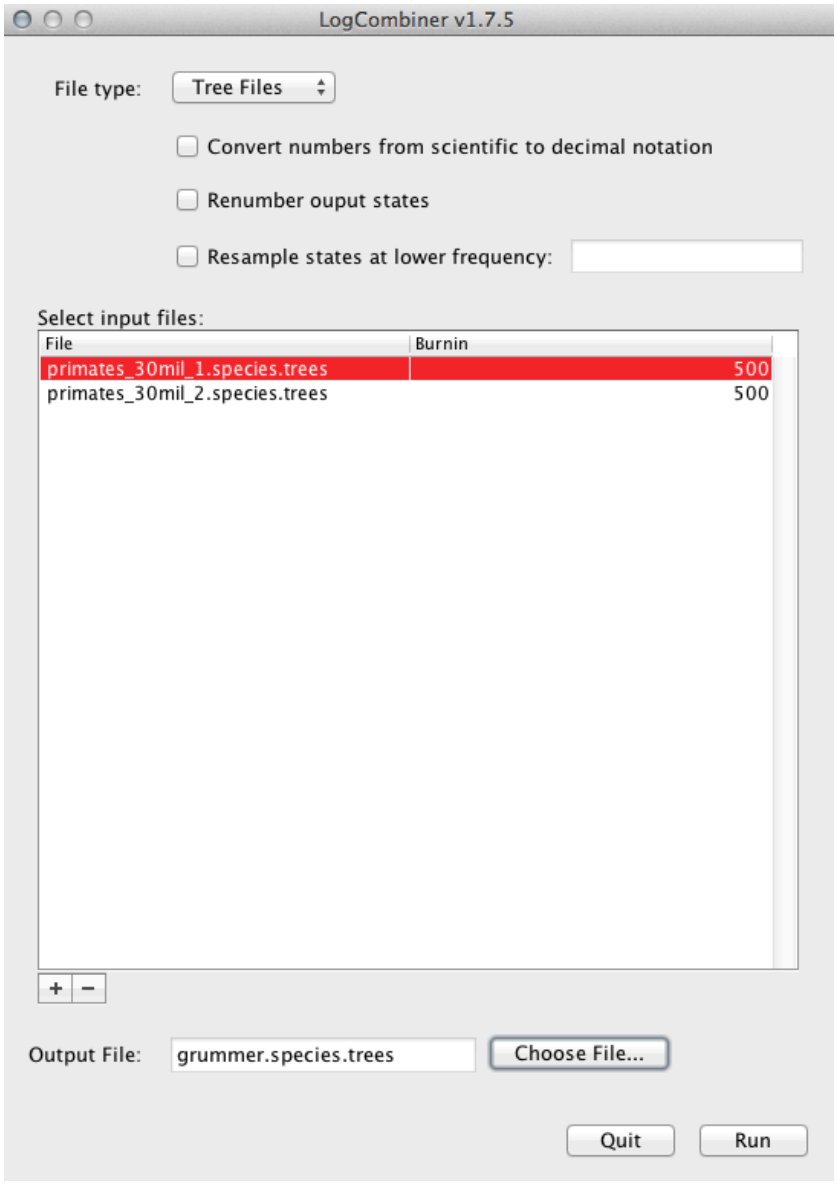
To determine if these two runs have converged, highlight both Tree file names in the Trace Files box. Click on Marginal Density tab and change the "colour by" to Trace file. If the distributions are overlapping the parameters have converged. Another way to check is that the ESS values for all parameters are above 200 when looking at the combined trace file.

Once your analysis has finished, load the .log file into Tracer and see if the values for your estimated parameters are converging. You should be able to recognize some of these parameters based on the knowledge you've gained over the past three weeks in this course!

Now, we are going to combine the tree files generated during the previous analyses. These tree files are located in the same folder as where you just downloaded the .log files, and end in "species.trees". We are doing this to increase our sample size and number of trees in our analysis.

Go to the BEASTv1.7.5 folder and Open the program LogCombiner. Change "File type" at the top to Tree Files. Press the "+" button in the bottom left to load the two primates_30mil_#.species.trees files. Alternatively, you can drag-and-drop them into this window. For each file set the Burnin to 500. Designate an output file name that ends in ".species.trees" and click Run.

4.) What percentage of the sample of trees are we discarding as burnin?



We are now going to summarize the posterior distribution of trees (from two runs) into a single tree. This was also done in MrBayes with the last command in the Bayes block. Open the program TreeAnnotator in the BEASTv1.7.5 folder. The "Burnin" value will stay zero because we already removed burnin during combining tree files in LogCombiner. Leave the values for "Target tree type" and "Node heights" at their default settings. Your Input Tree File will be the file that was output from LogCombiner (grummer.species.trees in the above screen shot). Select the Output File name and end the file name with .tre . Now click Run.

When completed, find the new tree file TreeAnnotator just generated and open it in FigTree.

<u>Questions to Answer</u> (along with the ones from above)

5.) What is the resulting species tree?

6.) What are the posterior probability values for the nodes in the tree?

The gene trees for each of the sampled genes (gene1–5) are in the primates_genetrees.tre file, located in the Previously_Done folder in Google Drive. Open this file in FigTree. Scroll between the different trees in the file with the arrow buttons in the FigTree window. These trees were simulated and therefore don't have support values, but you can assume all nodes are fully supported.

7.) Which gene trees show a different topology than the species tree?

8.) In which ways are the gene tree topologies different from the species tree topology?