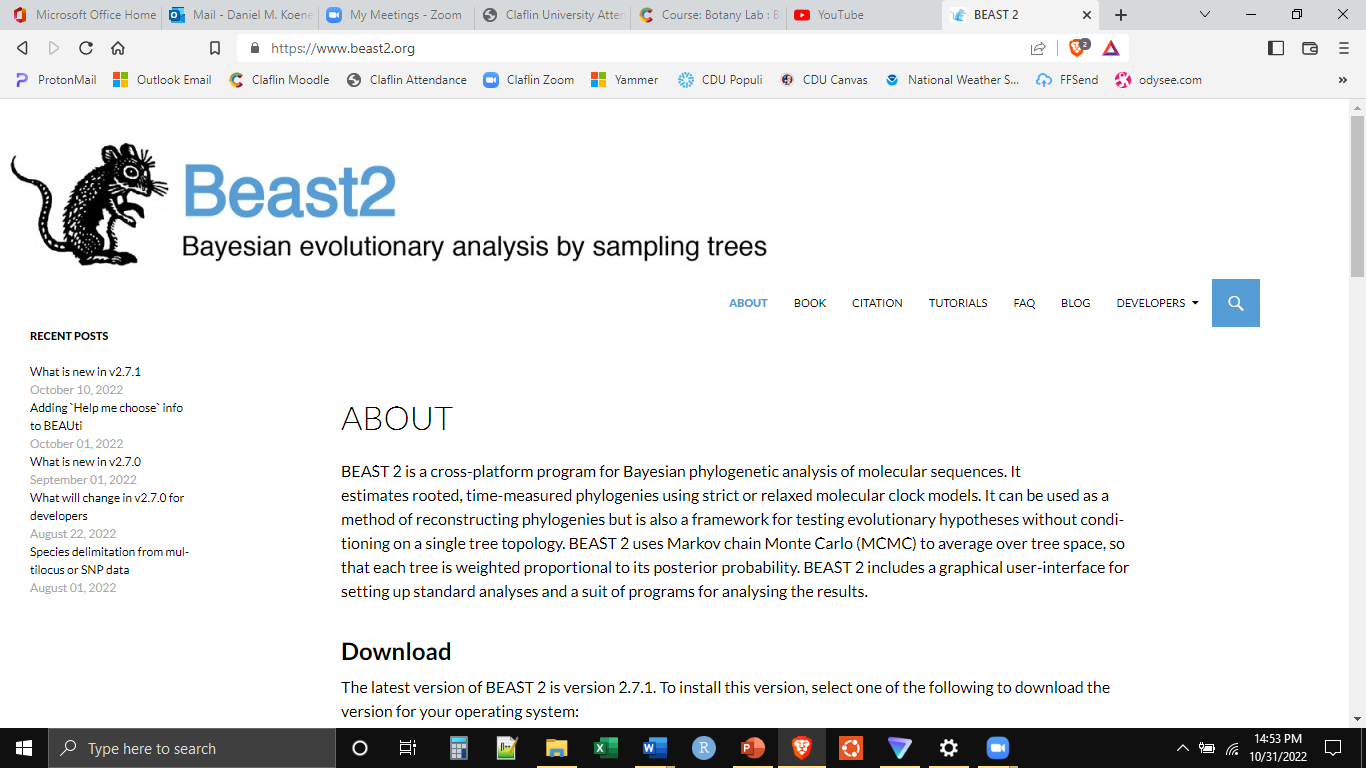
**Bioinformatics: Phylogenetics**

**Introduction**

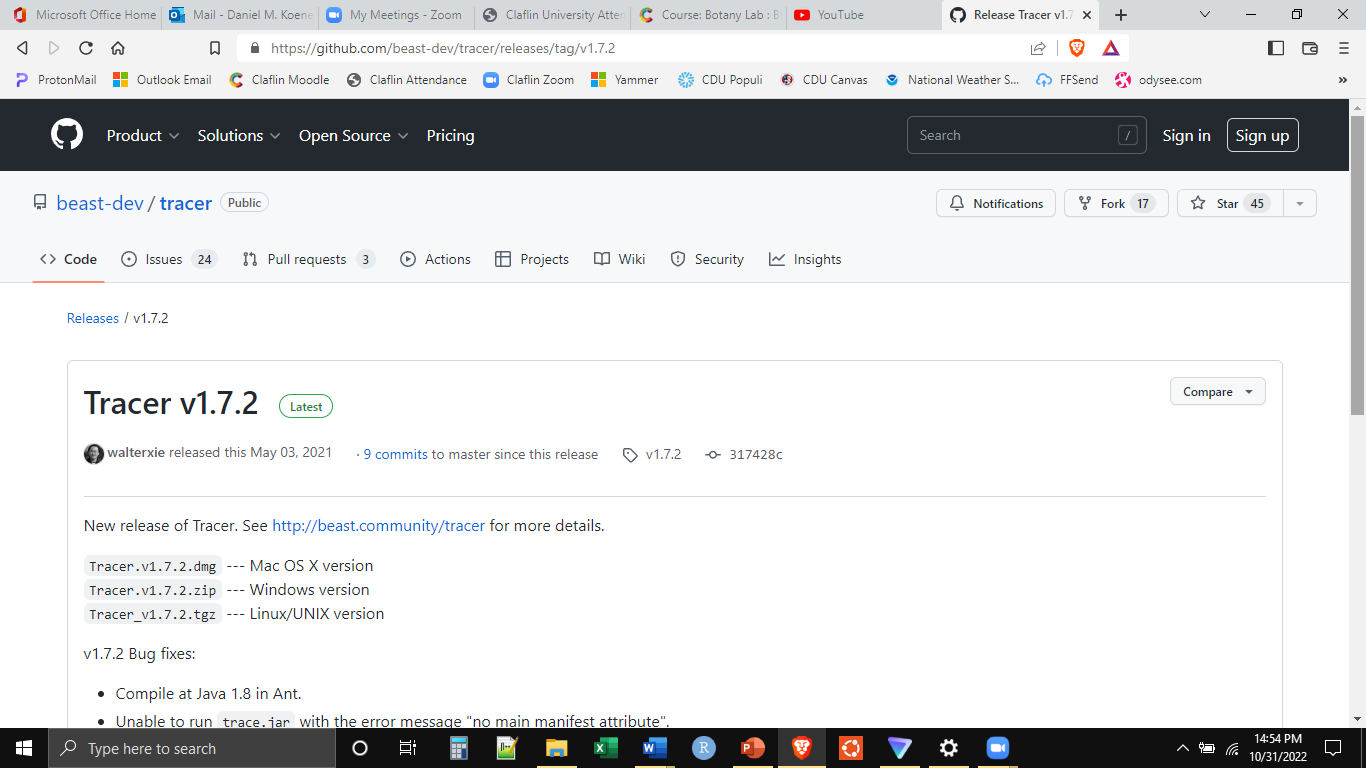
Once we have our annotated chloroplast sequences, we can use them for a variety of things. One thing that they can be used for is building phylogenies, or hypotheses of evolutionary relationships. There are a variety of methods for doing this, including parsimony (which we did in class), or more complicated model-based methods such as maximum likelihood or Bayesian inference, where we aren’t interested in the simplest explanation, but rather the one that makes the most sense in light of a particular model of evolution. Many of these tools can be accessed and operated only through a command line interface, which is why we have tried to build a skillset in that area. Some, however, can be operated both in the command line and GUI. One such phylogenetic estimation tool is BEAST, which we will use today. BEAST estimates phylogenies using Bayesian inference.

**Download BEAST**

The first step will be downloading BEAST and its supporting programs. BEAST can be downloaded from its website: https://www.beast2.org. No installation should be necessary, we will simply run the program out of its own folder.



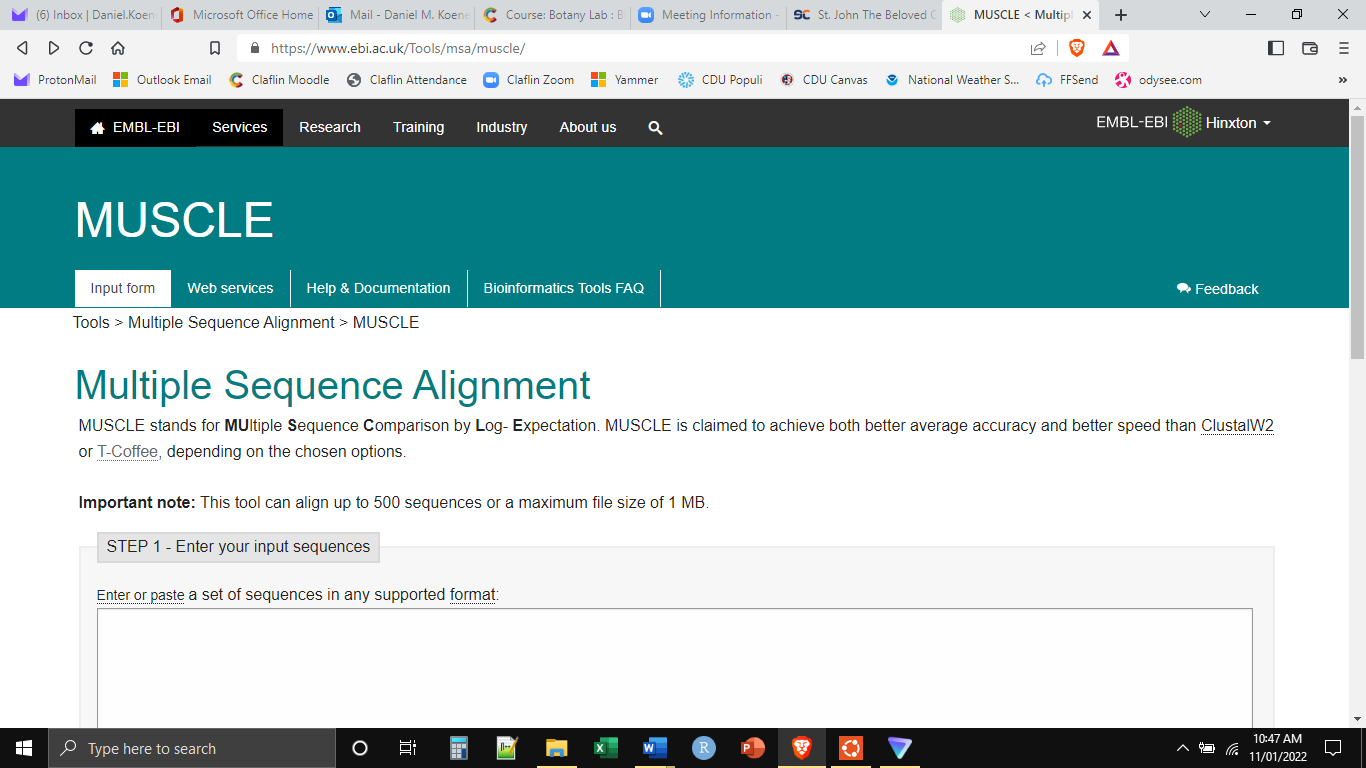
We will also need the program Tracer, to analyze and evaluate the solution that BEAST gives us. This can be downloaded here: www.github.com/beast-dev/tracer/releases/tag/v1.7.2.



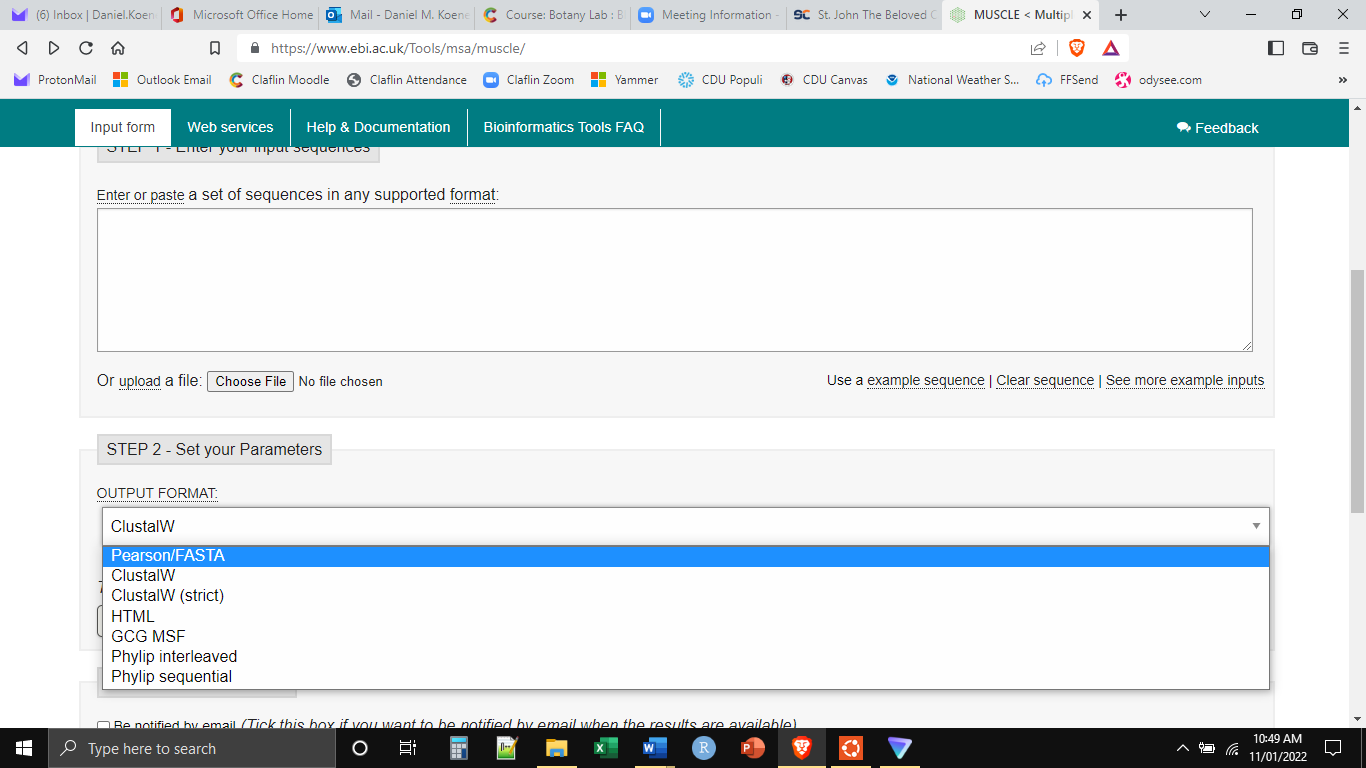
**Running BEAST**

Once we have the programs that we need, now we can start the process of building the phylogeny. The first step is to align the sequences. The sequence files we will be using today can be found on Moodle (AllSeqs.fasta). These are portions of the chloroplast genome for 12 species. The whole chloroplast genome is too big and the processes would run for too long. Please download this file. Do not open the file, especially if you are on a Mac, as it will change the file format.

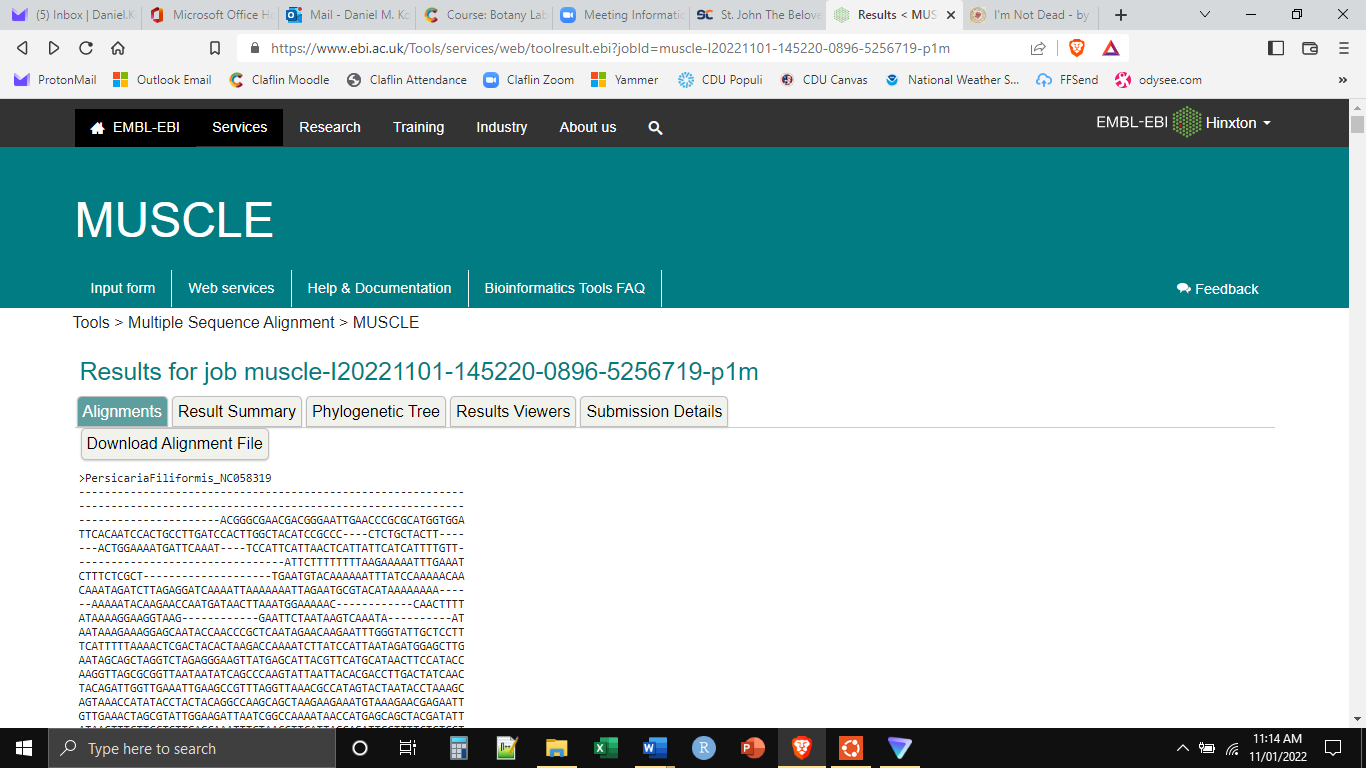
We will use a web platform to align the sequences so that all of the same nucleotide positions are in the same column for all of the species. The platform we will use is MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/). Please navigate to the MUSCLE website.



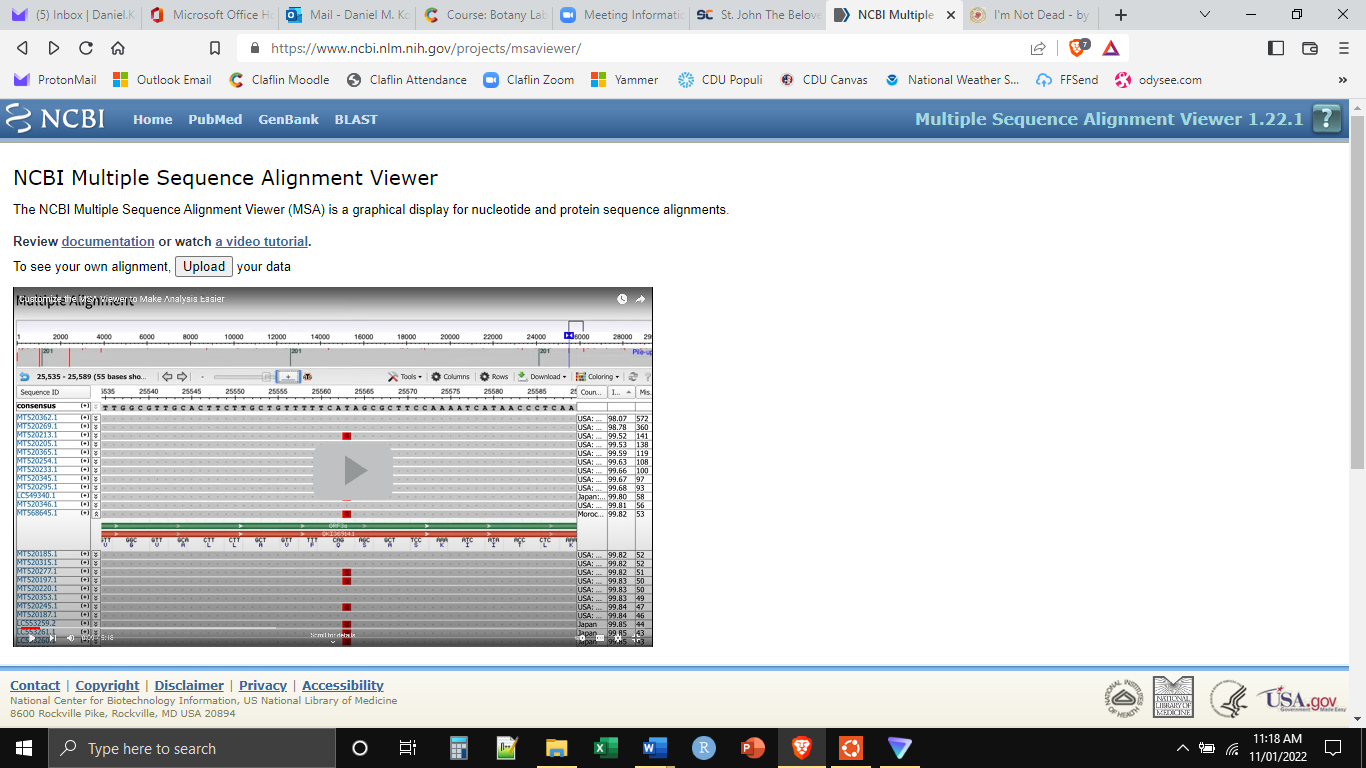
Scroll down, and we want to change two parameters. First, we want to upload our sequences. Select “Choose File” in STEP 1. The file we will select is the AllSeqs.fasta we downloaded from Moodle. In STEP 2, we want to select the output format to be Pearson/FASTA.



Then scroll to the bottom and click “Submit”. Even though these sequences are substantially shortened, they still may take a while to align. Be patient. When it finishes, we want to download the aligned FASTA file.

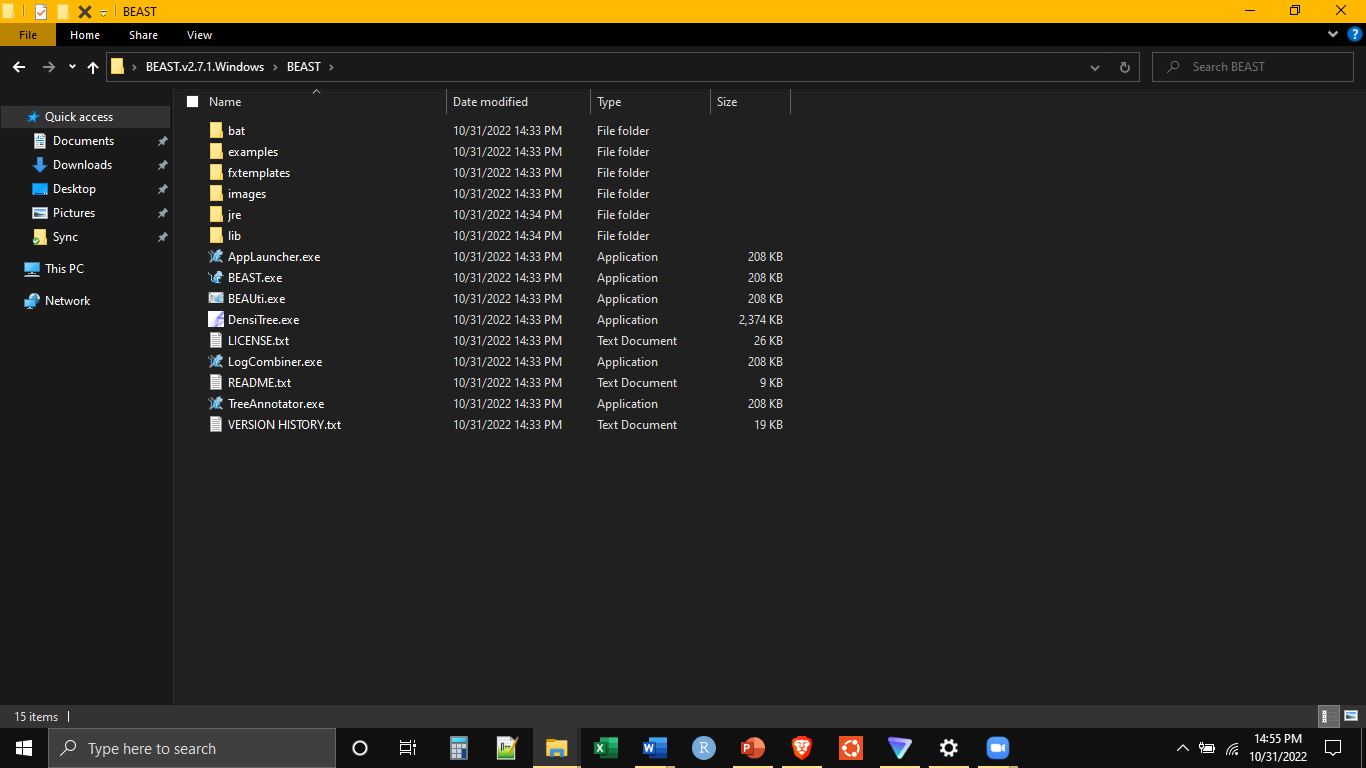


You will likely have to copy and paste the output into a simple text editor, and save it as “AllSeqs\_ALN.fasta”. We can view the completed alignment through the NCBI Alignment viewer (https://www.ncbi.nlm.nih.gov/projects/msaviewer/). Unfortunately, using this tool is a bit complicated. Navigate to the page and we will walk through it together.

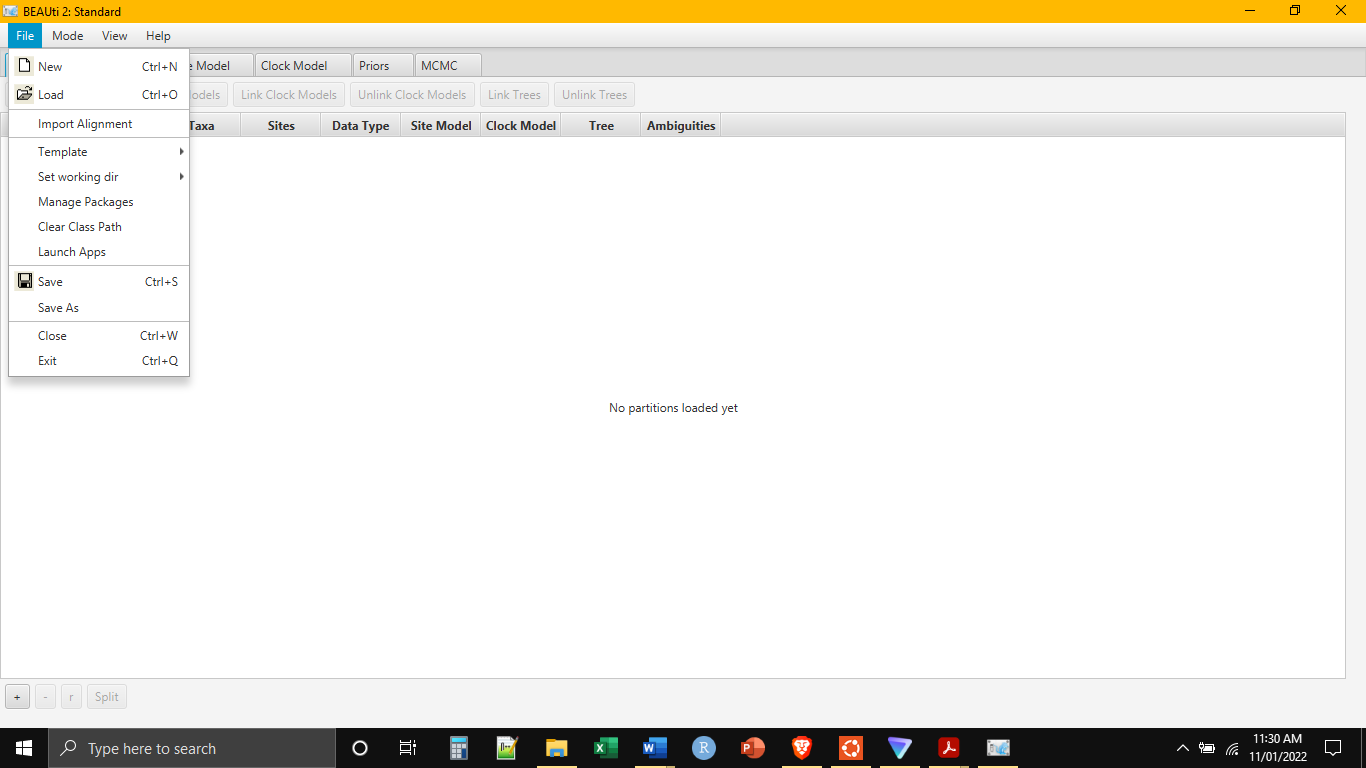


With the alignment in-hand, we are now ready to build the phylogeny. Using BEAST is a multi-step process. First, the input file must be built. Second, BEAST must run. Third, we have to assess the output. And fourth, we have to build the consensus tree.

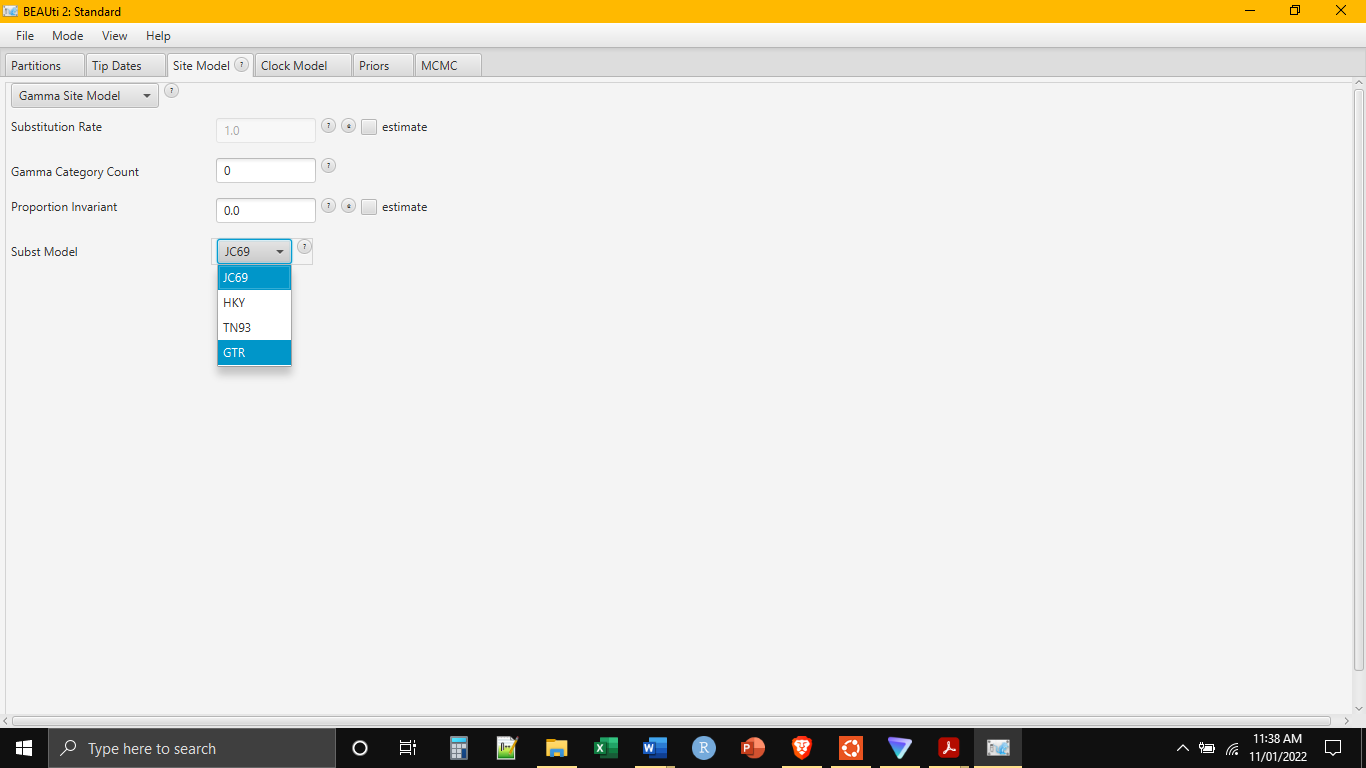
The first step, building the input file, we do in BEAUti. This can be found in the BEAST folder we downloaded and unpacked. Click on the BEAUti.exe to open the program.



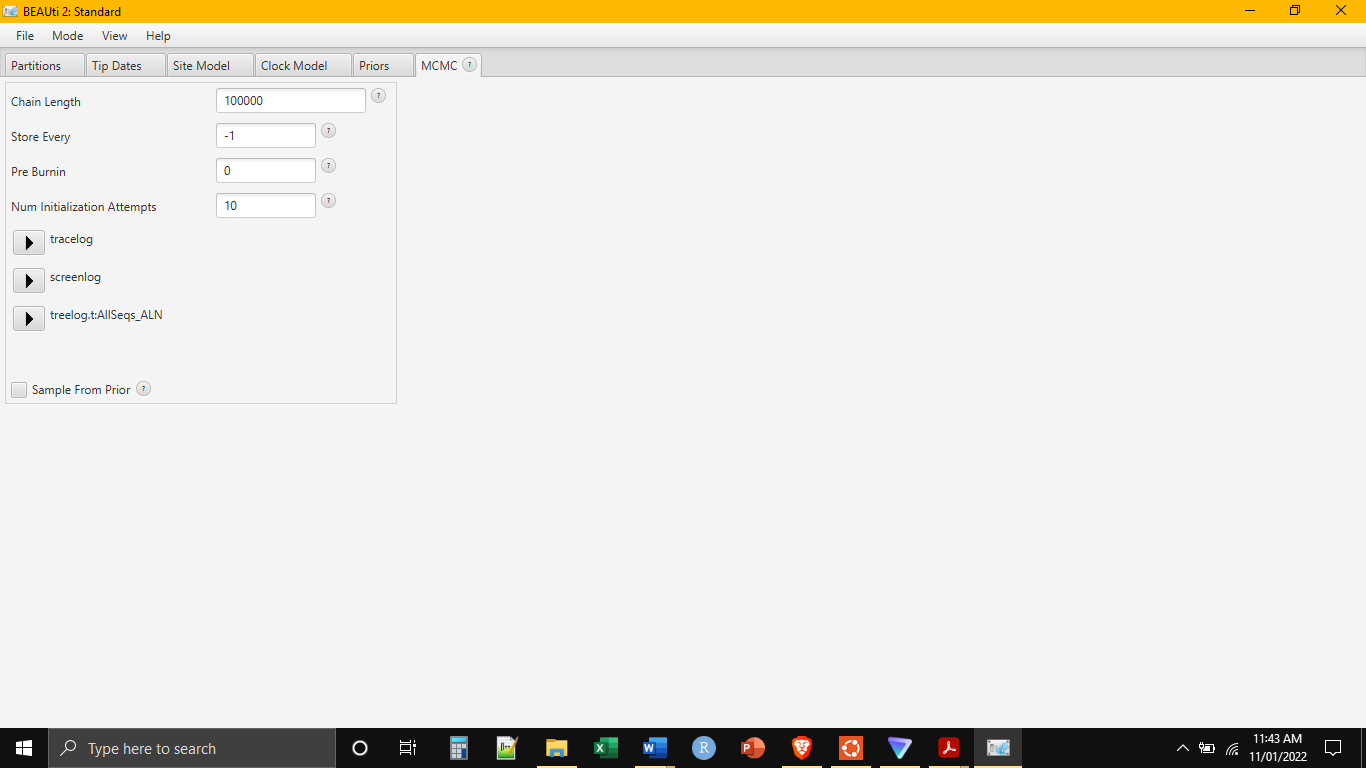
The first thing we will need to do in BEAUti is to import our alignment. This is the alignment file we just created with MUSCLE. It will ask for the data type, select “nucleotide”.



BEAST is what we call “highly parameterized”, that is, there are a lot of options that can be specified in reconstructing the phylogeny. We will do a pretty basic run, but will have to wade through some of the parameter selection process. At the top of the BEAUti window are a series of tabs. The “Partitions” tab should already be selected. We will skip the “Tip Dates” tab and instead select the “Site Model” tab. Where the program indicates Subst. Model, choose GTR from the dropdown model. The GTR (general time reversible) model is a versatile model of nucleotide evolution allowing independent rates for all nucleotide mutations.



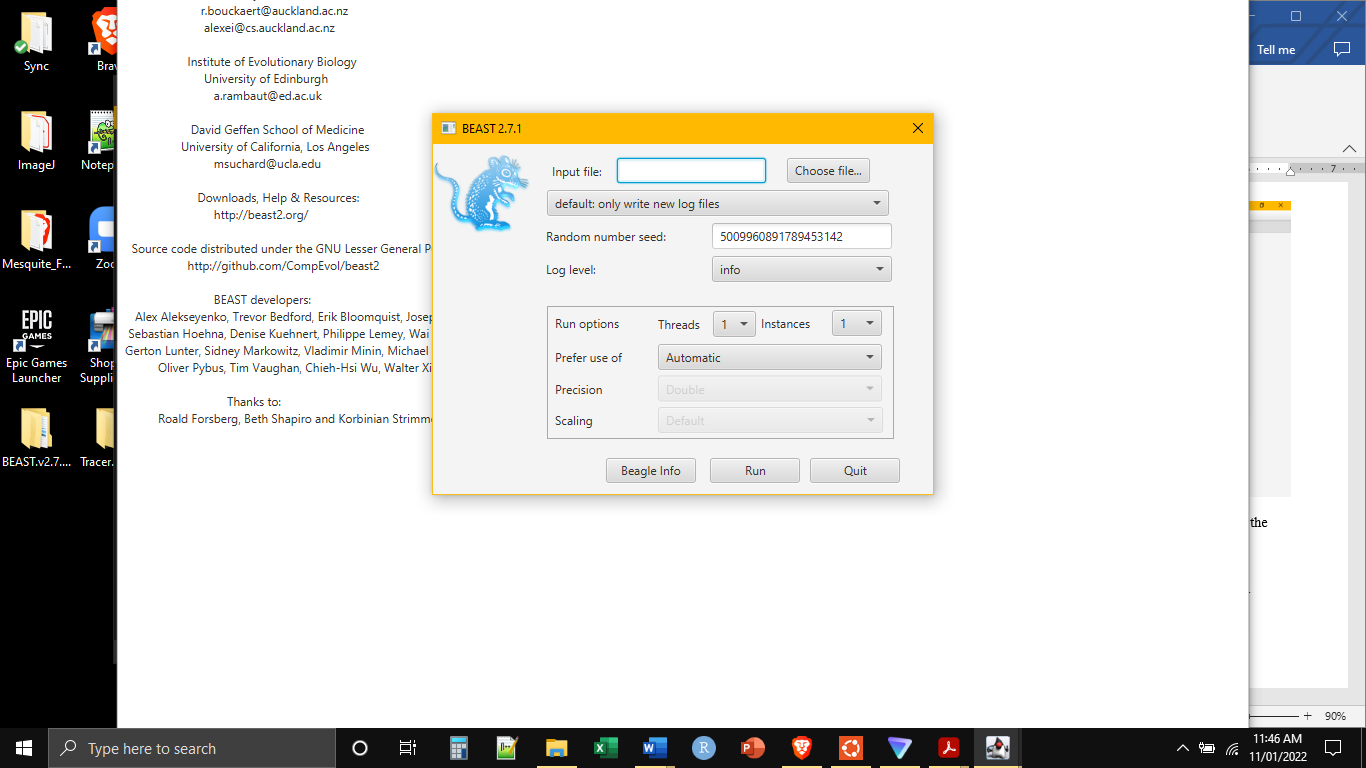
We will skip the “Clock Model” and “Priors” tabs. These would be used if we had fossil information that we wanted to include in the analysis. We do not. Thus, open the MCMC tab. Here we will choose how long we would like the analysis to run. Set the chain length to 1,000,000 and leave everything else as the default.



We then save the file as an XML. Go to File-SaveAs. Save the file with the rest of the files for today and name it AllSeqs.xml. Close the BEAUti program.

**BEAST**

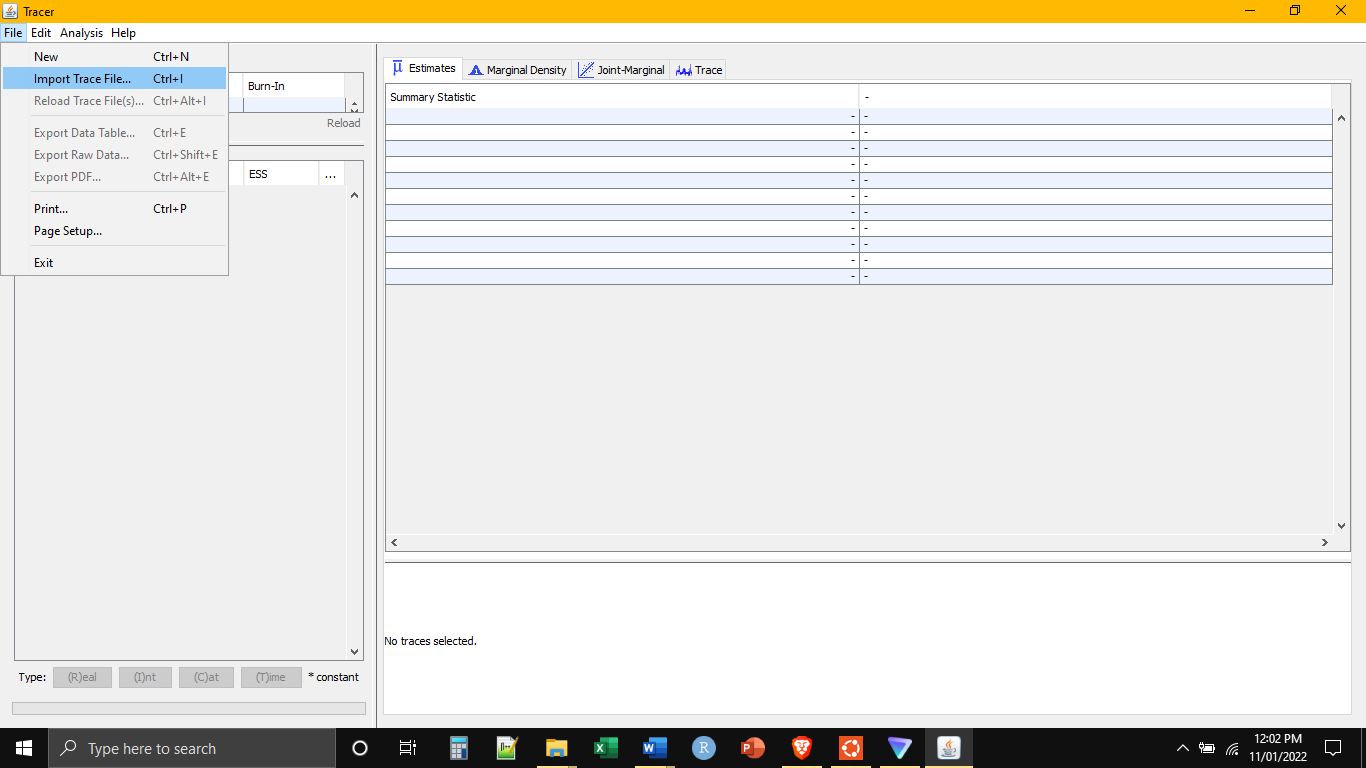
Now go back into the BEAST folder and click on the BEAST.exe to open BEAST. It will open a popup window that will prompt you for an input file. This input file is the XML file we just made in BEAUti. Click “Choose File” and select the XML file.



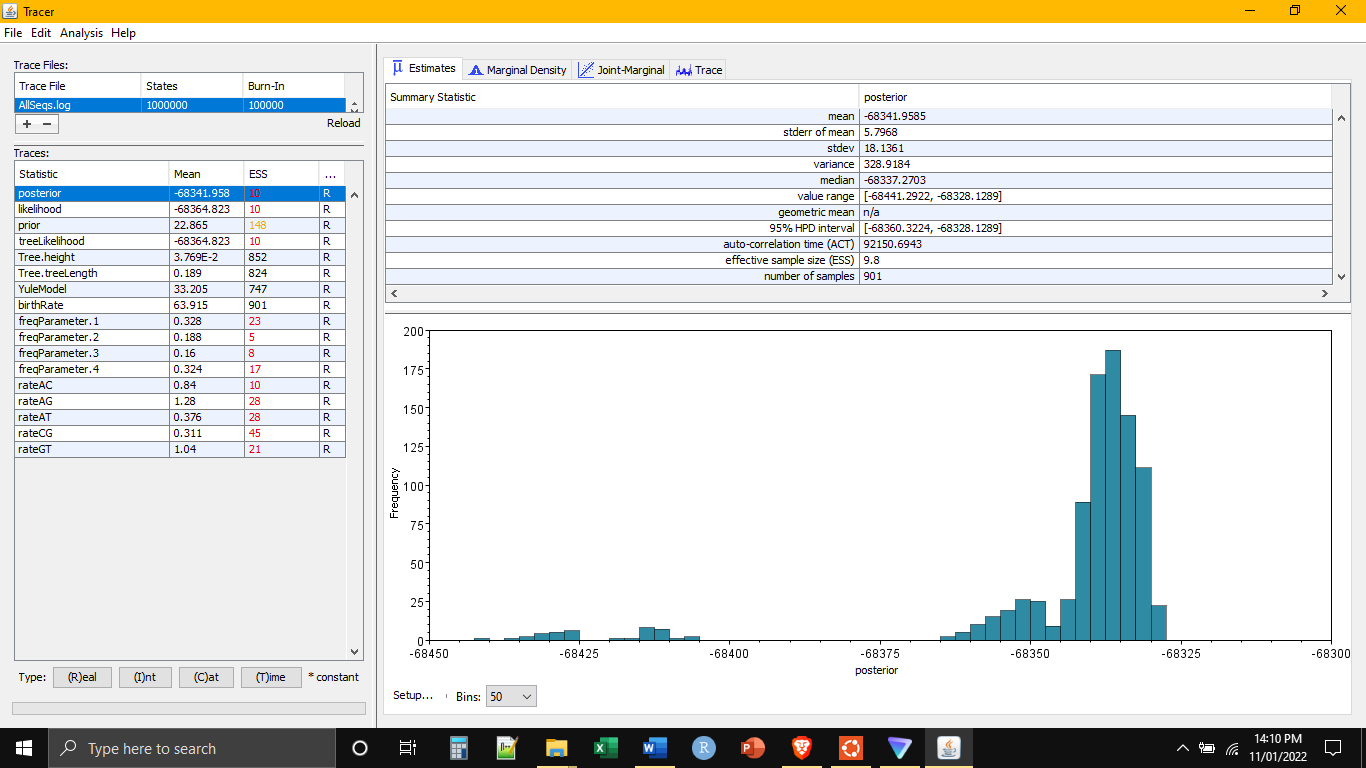
Leave everything else as the default and click “Run”. We have set BEAST to run for one million generations. This will take a few minutes but you should be able to see the log counting up. When it finishes, three new files will have been generated in the folder with the XML file.

**Tracer**

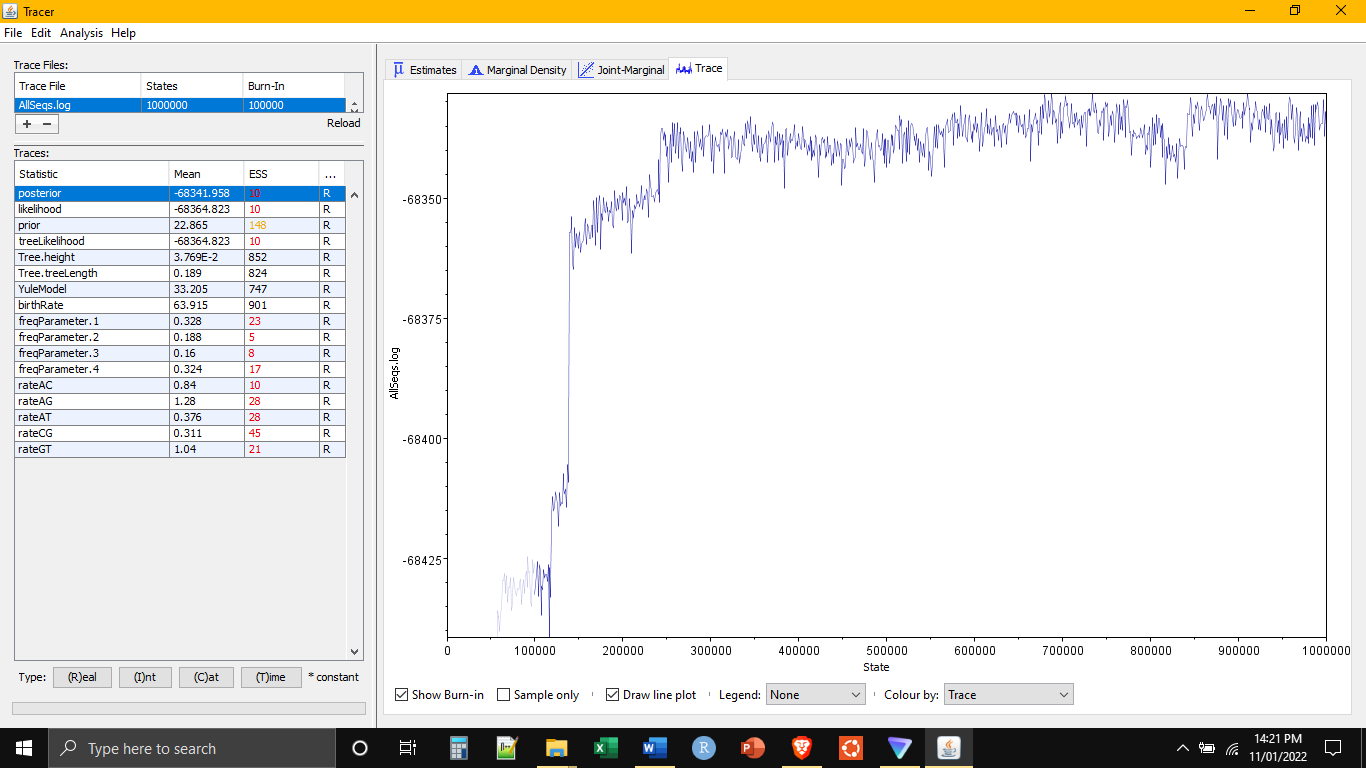
The next step is to assess the output. This is done through tracer. Open the Tracer folder and click on the “Tracer v1.7.2.exe” to open it. Then go to File-ImportTraceFile.



Choose the file named “AllSeqs.log” and open it. One of the major challenges of phylogenetic reconstruction is that there are a very large number of possible trees to choose from. Programs like BEAST are designed to search for the most probable tree among all of the possible trees. Tracer helps us to see how well BEAST has done at this task.



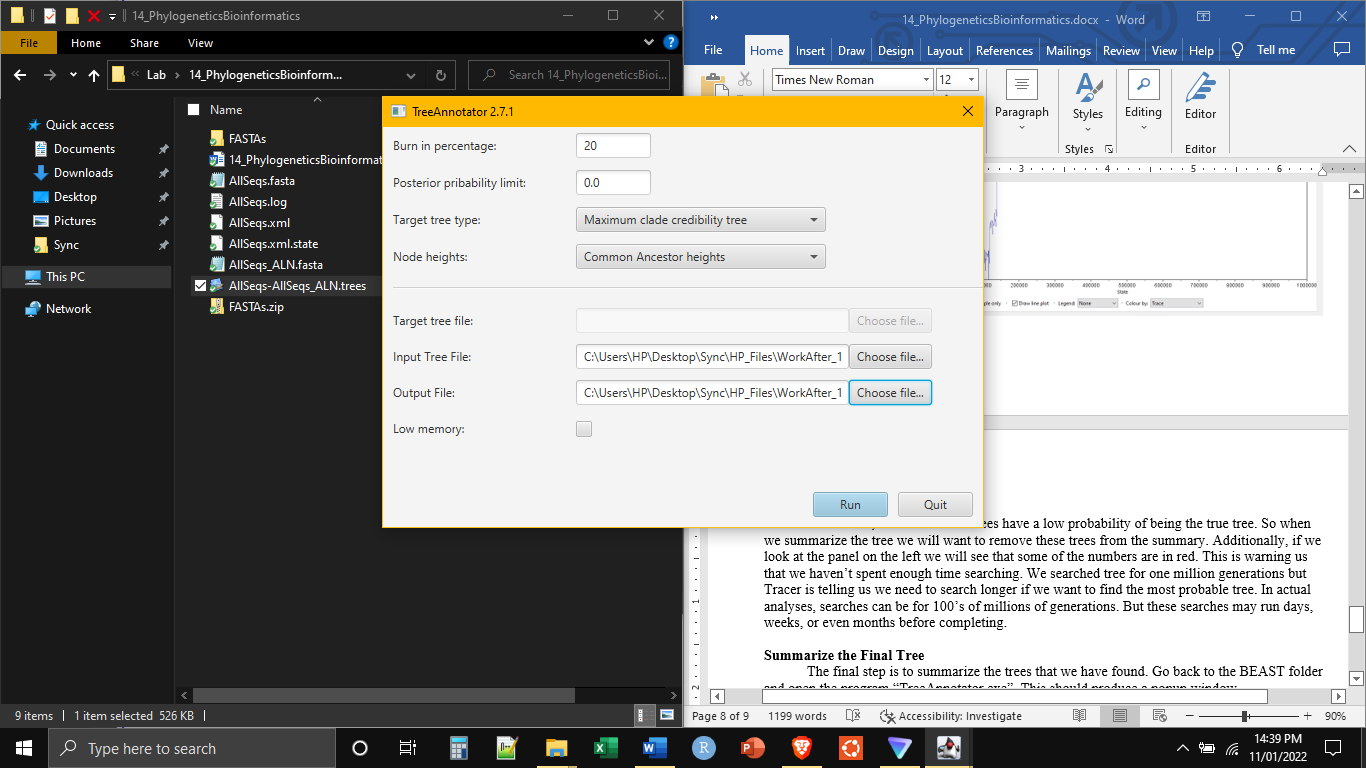
The process of finding the most probable tree takes some time. The early trees that BEAST tries are usually very poor estimations of the true tree. We ran the tree search for one million generations. When we summarize the tree search, we want to get rid of these early poor estimations. We can estimate this burn-in by viewing the trace. It is the fourth tab at the top of the screen.



In this case, the first 20% of trees have a low probability of being the true tree. So when we summarize the tree we will want to remove these trees from the summary. Additionally, if we look at the panel on the left we will see that some of the numbers are in red. This is warning us that we haven’t spent enough time searching. We searched tree for one million generations but Tracer is telling us we need to search longer if we want to find the most probable tree. In actual analyses, searches can be for 100’s of millions of generations. But these searches may run days, weeks, or even months before completing.

**Summarize the Final Tree**

The final step is to summarize the trees that we have found. Go back to the BEAST folder and open the program “TreeAnnotator.exe”. This should produce a popup window. We want to specify three things. First, set the burn in percentage to 20 percent. Second, choose the input trees file. This will be called “AllSeqs-AllSeqs\_ALN.trees”. Finally, select a name and location where the output tree file will be saved. I am saving this file in the same folder as all the other files and naming it “Final.tre”. Then click “Run”.



This will generate the “Final.tre” file in the folder you selected. Close the window. We can view this tree file using any number of web applications or desktop programs. For our purposes, let’s use IcyTree (https://icytree.org/) . Navigate to the website and drag and drop the “Final.tre” file into the upload box.

