Lab 5. Phylogenomics Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The purpose of today’s lab is to use genomic data to test phylogenomic hypotheses. Mitochondrial DNA suggests that the Green-winged Teal is paraphyletic with respect to the South American Yellow-billed Teal (*Anas flavirostris*); specifically, *A. c. carolinensis* is sister to *A. flavirostris*, not *A. c. crecca*. However, analysis of morphological characters suggests that *A. c. carolinensis* and *A. c. crecca* are sister taxa. We will use nuclear DNA to test these two phylogenetic hypotheses. We will use the program starBEAST2, which is a Bayesian implementation of the multispecies coalescent that jointly infers gene and species trees directly from multiple sequence alignments.

**starBEAST2**

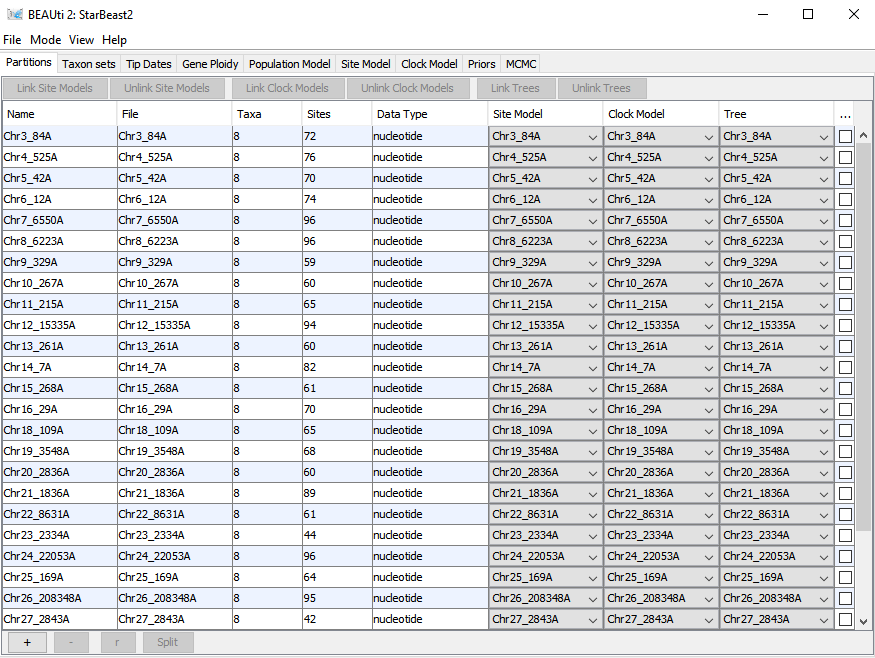
1. See the document **GettingStarted.2022.docx** for instruction on downloading and installing starBEAST2.

**Input File**

1. Because starBEAST2 is computationally intensive, we will focus our analyses on locus from each of 29 chromosomes. The names of the files provide the chromosome number (chr#) and a locus number (an arbitrary number assigned to the locus when assembling the data using the DaCosta-Sorenson pipeline). Note that one locus is linked to the Z sex-chromosome.
2. Open **Chr1\_15A.fas** by right clicking on the file, selecting Open with > WordPad
   1. This file contains sequences from 4 individuals; because they are diploid, each individual has two alleles (a & b).
   2. The samples are as follows:
      1. DDG1707 = Anas acuta (used as an outgroup)
      2. JLP016 = Anas crecca crecca (the Eurasian Green-winged Teal)
      3. KSW3040 = Anas crecca carolinensis (the American Green-winged Teal)
      4. KGM267 = Anas flavirostris (South American Yellow-billed Teal)

**Creating an Input File using BEAUti**

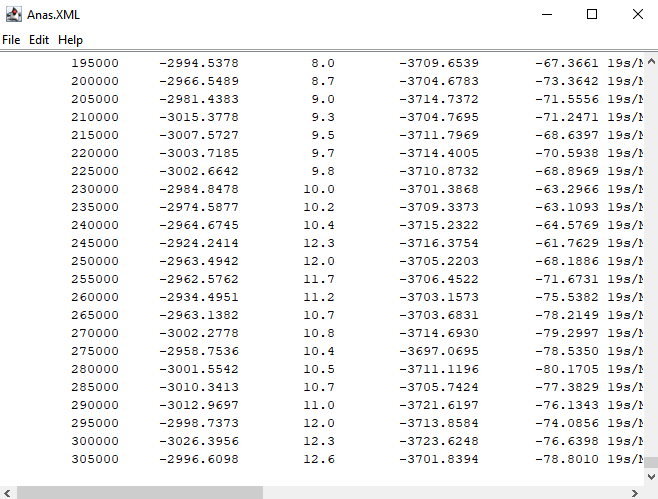
1. Go to your BEAST folder and open BEAUti.exe.
2. Select **File** > **Template** > **starBEAST2**
3. Drag your 29 \*.fas files (from the **InputFiles** folder) onto your BEAUti window.
   1. A window (Add Partition) will appear. Select **Import Alignment** from the dropdown menu.
   2. Another window will appear for each locus. Make sure **nucleotide** is selected and click OK (you will do this 29 times).
   3. Your BEAUti window show now look like the screenshot below.



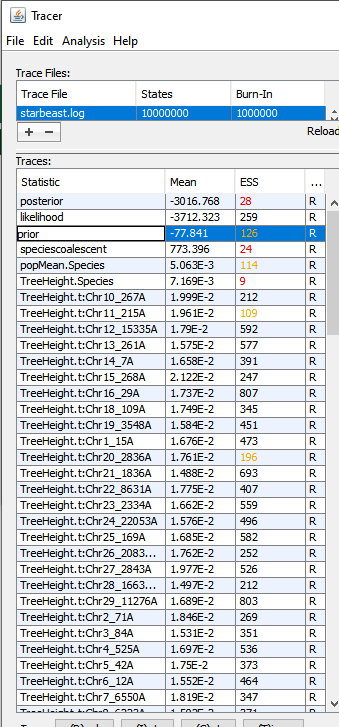
1. For simplicity, we are going to use a strict molecular clock for this analysis. In doing so, we allow the rate of molecular change to vary among loci but it is fixed among taxa. If we were going to publish these results, we would want to also explore relaxed clocks.
   1. Select all the files from your BEAUti window (click on the first locus, scroll down to the last locus, hold down the shift key and click on the last locus).
   2. Click on Link Clock Models near the top of the window.
2. Select the **Taxon sets** tab. Here we will define which sequences belong to which taxon.
   1. Under Species/Population, double-click on the “D” for DDG1707a and replace it with **A\_acuta**.
   2. Copy and paste A\_acuta into the “D” cell for DDG1707b
   3. Replace “J” for JLP016a & JLP016b with **A\_c\_crecca**
   4. Replace “K” for KGM267a & KGM267b with **A\_flavirostris**
   5. Replace “K” for KSW3040a & KSW3040b with **A\_c\_carolinensis**
3. Select the **Gen Ploidy** tab.
   1. Set **geneTree.t:ChrZ\_1639Z** to 1.5 to indicate that this locus is sex-linked (for every 2 copies of an autosomal locus, there are 1.5 copies of Z-linked loci).
   2. Leave all others as is (2.0) to indicate that they are autosomal (diploid) loci
4. Select the **Site Model** tab.
   1. Set **Subst Model** to HKY and **Frequencies** to Empirical. For more complex data sets, we might want to estimate nucleotide frequencies, but here, we can use the actual frequencies.
   2. To apply this site model to all loci, select all of the partitions in the left hand column (use the shift key) and click “OK”.
5. Select the **Priors** tab. Here we will tell the MCMC algorithm how to explore parameter space.
   1. For Tree.t:Species, select the **Birth Death Model**.
   2. Leave the others the same.
6. Select the **MCMC** tab. Normally, we would want to run longer chains to better explore parameter space, but given time constraints, we will leave the default run parameters.
7. Select File > Save As > Anas.XML

**Running BEAST**

1. Execute BEAST.exe
   1. If you get this window, click on **More Info** and **Run anyway**
   2. 
2. When the BEAST window opens, click on **Choose File** and Select **Anas.XML**
3. Click Run - you see a lot of information scrolling through the BEAST window. BEAST is reading your input file created in BEAUti.
4. When the run actually starts, you should see the following scrolling through. These are the steps of the MCMC and the data that are being collected.



1. The amount of time needed for the analysis to complete will depend on your computer speed. Mine completed in about 5 minutes.
2. Once the run is complete, return to the folder where you stored your Anas.XLM file. Here you will find a variety of new files that were generated.
3. We will start by checking how well the MCMC performed
   1. Open Tracer from the Lab\_5 folder.
   2. Find **starbeast.log** and drag it into the Tracer window and drop it under Trace File.
   3. You should now see a list of parameters that were estimated during the BEAST run.



* 1. The ESS column gives the Effective Sample Size. Each step in the MCMC is highly dependent on the preceding step. The ESS provides an estimate of how many times our steps could be considered independent from earlier steps. Ideally, we want this number to be greater than 200, but 100 is generally sufficient.
     1. You should notice that several of ESS values are quite low (and highlighted with red font). This indicates that we need to run our analyses for many for steps before trusting the data.
     2. Click on TreeHeight.Species to visualize an extreme case of the issue. Make sure Trace is selected in the panel on the right.
     3. This plot shows the parameter estimates that were obtained throughout the run. Although we see plenty of fluctuating values, you should notice peaks and valleys resulting from the non-independence. Thus, the over tree height was not explored (or estimated) very well.
     4. Now click on one of the parameters with ESS > 500. You show notice greater independence among estimates.
     5. So, bottom line, we really need to run this analysis for many more steps (perhaps 100 million or 1 billion). Regardless, we can still learn some things from these results.
  2. Look through some of the other tabs to see other information provided by tracer.

1. Open DensiTree from your BEAST folder.
   1. Select File > Load and choose **species.trees** from your list of files. Click Open.
   2. This DensiTree plot shows all the trees that were visited during the BEAST run.
      1. Blue lines show the most common topology recovered; red the second most common; green the third.
   3. From the menu on the right, select Show and check Root Canal to see the most plausible tree.
   4. What does this tree suggest about phylogenetic relationships (in particular, is A. c. carolinensis more similar to A. c. crecca or A. flavirostris? (We will discuss this plot as a group.)
2. To generate a summary tree, open TreeAnnotator from your BEAST folder.
   1. Define the burnin as 10% (this will discard the first 10% of trees).
   2. For input tree file, choose **species.trees** (same file used in DensiTrees).
   3. Name the output **species.trees.summary**
   4. To visualize the summary tree, open FigTree from the Lab 5 folder.
   5. File > Open > species.trees.summary
   6. You should see the same topology as you saw when viewing the most plausible tree in DensiTrees.
   7. To view support for each of the relationships, check the box for **Branch Labels** from the menu on the left, then choose **posterior** from the **Display:** dropdown menu.
   8. 1.0 indicates that there is 100% posterior support for the relationship. 0.9 would indicate 90% support, and so on.

Questions for group discussion

How well supported are the relationships in this tree?

Does this tree agree with mitochondrial DNA or morphology?

Why might there be conflicting gene trees based on different markers?

How can we improve our confidence in the results?