Lab 3. PopGenome Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The purpose of today’s lab is to use genomic data to calculate genetic summary statistics for the two subspecies of Green-winged Teal (*Anas crecca crecca* and *Anas crecca carolinensis*) that we have been working with. We will use the R script, PopGenome, to calculate nucleotide diversity, Tajima’s D, and Fst. Because ddRAD-seq generates a lot of data, we will analyze a subset of those data and focus on Chromosome 2.

1. Open R
2. If you haven’t done so already, you need to install the PopGenome library.
   1. Type **install.packages("PopGenome")** on the command line.
   2. Alternatively, you can choose **Packages** from the toolbar, **Load Packages…**, then find **PopGenome** in the list of packages.
3. Next we need to tell R that we want to use the PopGenome library – type **library(PopGenome)** on the command line.
4. Input files
   1. PopGenome requires FASTA files, which contain aligned DNA sequences, for analysis. The format of a FASTA file is:

>sample\_name\_1

AGGGCATTATGCCATGCC

>sample\_name\_2

AGGGCATTATACCATGCC

.

.

.

>sample\_name\_n

AGGGCATTATGCCATGCC

* + 1. Open your folder labeled FASTA—this folder contains DNA sequences from 396 loci that map to Chromosome 2 in the duck genome. There is also a file called **Chr2\_concat.fasta**, which contains all 396 loci combined (concatenated).
    2. Open **20A.fasta** using WordPad or Notepad (right-click on the file, select **Open with…**, then choose the program from the list).
    3. This file contains 192 sequences from 96 individuals.
    4. The aligned sequences are 147 nucleotides long.
    5. Scroll to the right, and you see that some samples have dashes (------) rather than nucleotides (A, G, C, T). This indicates the presence of an indel (or insertion-deletion)—the dashes signify a gap in the alignment.

1. Running PopGenome
   1. Return to R.
   2. Type **setwd(“*type the path of the directory in which you want to work”*)**.
      1. For example, my path would be **setwd("C:/Users/w048jlp/Desktop/PopGen/Lab3\_PopGenome")**
      2. Helpful hint: if your PopGen folder is open, you can click on the address bar to copy the path and paste it into R after setwd(“. However, you need to replace the back slash (\) with a forward slash (/)
      3. Type **getwd()** to confirm that you now have the correct working directory
   3. Once you are in the correct directory, we need to load the folder of fasta files into R:

**GWTE <- readData("FASTA", include.unknown=TRUE, progress\_bar\_switch=FALSE, SNP.DATA=FALSE)**

Note that we are telling PopGenome to include sites with missing data in the analysis (include.unknown), not use the progress bar (if the program freezes, this option allows us to see which file caused the problem), and that we are not using SNP (single nucleotide polymorphism data).

* 1. Once the data are loaded, we need to define our populations. The command for defining populations is set.populatons. We will call our defined groups GWTEpops:

**GWTEpops <- set.populations(GWTE,list(**

* 1. Now assign individual sequences to each of two populations, **carolinensis** (*N* = 26 individuals) and **crecca** (*N* = 39 individuals). It will be easiest to copy (Highlight the lines and hit Ctrl-C) the following lines and paste them into R (right-click, paste):

**carolinensis <- c("CRAB18a", "CRAB18b", "CRAB22a", "CRAB22b", "CRAB47a", "CRAB47b", "CSW4487a", "CSW4487b", "DDG1732a", "DDG1732b", "JJW978a", "JJW978b", "JMM448a", "JMM448b", "KGM009a", "KGM009b", "KGM1483a", "KGM1483b", "KGM1485a", "KGM1485b", "KGM1487a", "KGM1487b", "KGM1488a", "KGM1488b", "KGM1489a", "KGM1489b", "KGM1490a", "KGM1490b", "KGM1494a", "KGM1494b", "KGM1496a", "KGM1496b", "KGM1497a", "KGM1497b", "KGM1498a", "KGM1498b", "KGM1503a", "KGM1503b", "KGM1504a", "KGM1504b", "KGM1506a", "KGM1506b", "KGM1507a", "KGM1507b", "KSW3040a", "KSW3040b", "REW483a", "REW483b", "REW554a", "REW554b", "UAMX203a", "UAMX203b", "UAMX2138a", "UAMX2138b", "UAMX5069a", "UAMX5069b"),**

**crecca <- c("BKS3223a", "BKS3223b", "BKS3277a", "BKS3277b", "BM52620a", "BM52620b", "BM61324a", "BM61324b", "DDG1704a", "DDG1704b", "DDG1884a", "DDG1884b", "DDG1885a", "DDG1885b", "G78a", "G78b", "G79a", "G79b", "G80a", "G80b", "G83a", "G83b", "G85a", "G85b", "G86a", "G86b", "JLP016a", "JLP016b", "JLP021a", "JLP021b", "JLP022a", "JLP022b", "JLP026a", "JLP026b", "JLP027a", "JLP027b", "KGM020a", "KGM020b", "KGM021a", "KGM021b", "KSW4951a", "KSW4951b", "KSW4961a", "KSW4961b", "PR108a", "PR108b", "PR113a", "PR113b", "PR115a", "PR115b", "PR117a", "PR117b", "PR118a", "PR118b", "PR120a", "PR120b", "PR125a", "PR125b", "SAR6109a", "SAR6109b", "SVD1161a", "SVD1161b", "SVD1204a", "SVD1204b", "SVD1205a", "SVD1205b", "SVD1210a", "SVD1210b", "SVD1380a", "SVD1380b", "SVD469a", "SVD469b", "SVD481a", "SVD481b", "SVD482a", "SVD482b", "SVD483a", "SVD483b")))**

* 1. Now that we have our populations defined, we can calculate our genetic summary statistics.
     1. First, we need to tell R how to handle the data using this command:

**F\_ST.stats(GWTEpops,new.populations=FALSE,subsites=FALSE,detail=TRUE,mode="nucleotide",only.haplotype.counts=FALSE,FAST=TRUE)**

* + 1. This command will generate a list of statistics that can be calculated. We will only use a few of the many options.
  1. We will calculate two sets of statistics – those related to Fst (including nucleotide diversity) and tests of neutrality (specifically Tajima’s D). Type (or copy and paste) the following two commands

**GWTE.fst <- F\_ST.stats(GWTEpops,FAST=TRUE)**

**GWTE.neutr <- neutrality.stats(GWTEpops)**

* 1. We can now extract the statistics from R.
     1. We might want to record the number of aligned nucleotides in each file:

**write.table(GWTE.fst@n.sites, "GWTE.n.sites.txt")**

[We are telling R to write a file (write.table) that includes n.sites (number of sites for each locus) that was calculated in our GWTE.fst calculations.]

* + 1. However, because of gaps and other missing data, the aligned nucleotides might not reflect the number used in analysis. To get those numbers, we want n.valid.sites:

**write.table(GWTE.fst@n.valid.sites, "GWTE.n.valid.sites.txt")**

* + 1. The following commands will create files containing our values for Fst, nucleotide diversity within each population, and nucleotide diversity between populations (how different are the populations), respectively:

**write.table(GWTE.fst@nuc.F\_ST.pairwise, "GWTE.F\_ST.pairwise.txt")**

**write.table(GWTE.fst@nuc.diversity.within, "GWTE.nuc.diversity.within.txt")**

**write.table(GWTE.fst@nuc.diversity.between, "GWTE.nuc.diversity.between.txt")**

* + 1. Finally, we want to examine our values of Tajima’s D to test for population expansions (strongly negative values)

**write.table(GWTE.neutr@Tajima.D, "GWTE.TajD.txt")**

1. Examining the results
   1. First, let’s compare nucleotide between the three subspecies.
   2. Open your file **GWTE.nuc.diversity.within.txt**
   3. Copy the data and paste them into column A of an Excel spreadsheet.
   4. To separate the data into separate columns,
      1. select **Data**, then **Text to Columns** from the toolbar.
      2. Select **Delimited** from the popup window, hit **Next**
      3. Select **Space** under **Delimiters**
      4. Click **Finish**
      5. Type **nucl.div.within** in cell A1, **carolinensis** in B1, **crecca** in C1
   5. Nucleotide diversity was calculated as the average number of pairwise differences among all sequences sampled from the same populations. Notice that for Chr2\_concat, there 100s of differences, but for 20A there are approximately 2-4 differences. Therefore, nucleotide diversity is sensitive to the length of the sequences and is not directly comparable among loci. We can use n.valid.sites to calculate the average proportion of differences, which makes the values comparable.
      1. Open your file **GWTE.n.valid.sites.txt**
      2. Copy the data and paste them into column F of your excel file.
      3. Since we already selected space as a delimiter, the data should copy into two columns. If not, highlight column F and repeat step d above.
      4. Type **n.valid.sites** into cell F1
      5. Copy cells B1, C1, and D1, and paste the subspecies names in I1, J1, and K1.
      6. Type **=B2/F2** into cell H2, **=C2/F2** into cell I2. These commands divide the number of differences by the number of sites to yield the proportion of differences.
      7. Copy H2 and I2, then paste the formulas into H3-H398, I3-I398
   6. We can now directly compare values of nucleotide diversity. The values for Chr2\_concat are the overall (or composite) values calculated for the full data set. What is nucleotide diversity for,

carolinensis\_\_\_\_\_\_\_\_\_\_\_\_\_\_

crecca\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Interpret: Which subspecies has higher genetic diversity for chromosome 2? \_\_\_\_\_\_\_\_\_\_\_\_

* 1. Open your file GWTE.nuc.diversity.between.txt and copy the data into a new sheet in excel (note that unlike the previous data files, the loci are listed in columns rather than rows—I haven’t been able to figure out how to fix this)
     1. Copy the data that you just pasted into excel and paste it into column K of your original spreadsheet as follows: highlight cell K1, choose **Home** > **Paste** > **Paste Special…** from the toolbar. Choose **transpose** from the popup window and click on OK. This flips your table around so that loci are in the rows.
     2. Highlight cell K1, and from the toolbar, choose **Home** > **Insert** > **Insert Cells…** > **Shift cells down**, so that the locus names are lined up with the correct data.
     3. Label cell K1 as **nuc.div.between**
  2. Repeat step 6g for your Fst values (**GWTE.F\_ST.pairwise.txt**), pasting the data into cell N1 (after following those steps, label cell N1 as **Fst**)
  3. Click Ctrl-H, and replace pop1 with carolinensis and pop2 with crecca.

Based on the composite Fst value, how much of the total genetic variation is explained by differences between crecca and carolinensis? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What is the highest Fst value among all 396 loci? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (hint: type **=MAX(O:O)** into any cell find the maximum Fst value)

Question for group discussion: How does the maximum compare to the overall values?

* 1. Let’s take a closer look at Fst and what it means. As we discussed in lecture, Fst measures the proportion of genetic variation that is explained by differences between populations. A value of 0 (or < 0) indicates no population structure; a value of 1 indicates that there are fixed nucleotide differences between the populations but no diversity within populations (nucl.div.between = 1; nucl.div.within = 0).
     1. To calculated Fst, we simply subtract the average nucleotide diversity within populations (𝜋s) from the value for nucleotide diversity between populations (𝜋d) and divide by 𝜋d.
     2. To do this in excel, in cell P1, type **=(L2-AVERAGE(B2:C2))/L2**

*L2 is 𝜋d, average(B2:C2) is 𝜋s.*

* + 1. Copy the formula into cells P3-P398
    2. How do these values compare with those calculated using PopGenome?

1. Finally, let’s examine Tajima’s D. Recall that values close to zero suggest a stable population size, whereas negative values suggest a population expansion, and positive values suggest a population decline.
   1. Open your file **GWTE.TajD.txt**
   2. Copy the data and paste them into columns S-U in your excel spreadsheet.
   3. What do the values for **Chr2\_concat** suggest about population size changes in Green-winged Teal?
   4. Also recall that Tajima’s D can also be influenced by natural selection: values near zero suggest selective neutrality, negative values suggest positive selection, and positive values suggest balancing selection. We could compare the values from each locus with the overall values to test for selection. Recall that selection acts locally within the genome whereas population size changes affect the entire genome. We will return to these values during the next lab.
2. Save your excel spreadsheet in your PopGen folder. We will need some of these values later.

Questions for group discussion:

* + - 1. How does nucleotide diversity compare between *A. c. crecca* and *A. c. carolinensis*? What does this tell us about relative effective population sizes?
      2. What does the overall Fst suggest about population divergence? How does this compare to the deep divergence found for mtDNA (Fst = 0.9)?
      3. How does the maximum Fst value compare to the overall value for chromosome 2 and for mtDNA? What might explain this high locus-specific value of Fst?
      4. What does Tajima’s D suggest about population size changes in Green-winged Teal?