Lab 1. Population Genomics Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

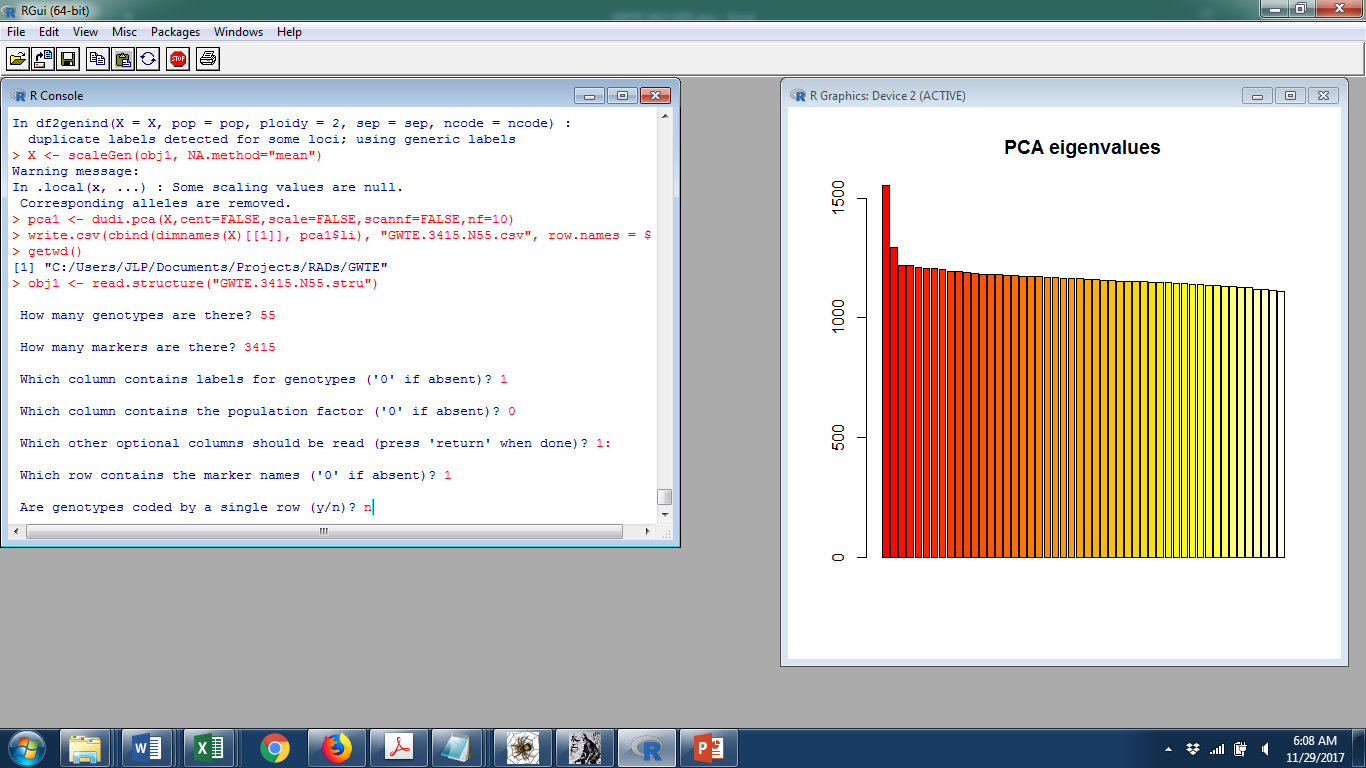
The purpose of today’s lab is to use genomic data to examine population structure in a species of duck that is found across the Holarctic (North America, Europe, and Asia). Two subspecies have been described, the Eurasian Green-winged Teal (*Anas crecca crecca*) and the American Green-winged Teal (*Anas crecca carolinensis*). We will use principal component analysis (PCA) to test for evidence of discrete population differences and gene flow. PCA will be conducted using the **adegenet** [http://adegenet.r-forge.r-project.org/images/R.png](http://www.r-project.org/)package, which is dedicated to the exploratory analysis of genetic data. It implements a set of tools ranging from multivariate methods to spatial genetics and genome-wise SNP data analysis.

**PCA**

1. Getting started.
   1. Go to <https://www.r-project.org/>
   2. Click “download R” and choose an appropriate mirror that is geographically close to Bolivia (Chile or Argentina will probably work best)
   3. Choose “Download R for Windows”, and then “install R for the first time”.
   4. Choose “Download R 3.4.2 for Windows” and Save File.
   5. Move the file “R-3.4.2-win.exe” to your desktop and execute the file.
   6. Go through the steps, leaving the default options checked, to install the program. Once it finishes, click “Finish”.
   7. Execute R
   8. Type **install.packages("adegenet")** on the command line to install the adegenet library in your R libraries
   9. Type **library(adegenet)** to tell R that you want to use the adegenet library for analyses
2. Input file
   1. Download the file “GWTE.3448A.N67.stru”, and move this file to your PopGen folder.
   2. This file contains 3448 ddRAD-seq loci from 67 individuals of Green-winged Teal.
   3. Open the data file in Microsoft Excel.
      1. The first row contains locus names. Each locus is named by the chromosome number followed by an arbitrary number.
      2. The 2nd and 3rd rows give the name of an individual sample followed by its genotype at each locus. Note that for this individual, there is an “a” & “b” entry. These are the two alleles sampled for that individual.
      3. The 4th and 5th rows provide the name of the next sample and its genotype at each locus.
      4. The genotypes are entered as arbitrary numbers. An individual that is homozygous for allele one will have a “1” entered for both the “a” & “b” entries. An individual that is heterozygous for alleles one and three will have “1” entered in one cell and “3” entered in the other. [An entry of “-9” indicates that data are missing for the locus for that individual.]
   4. What is the genotype of individual CRAB22 at locus 22?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   5. Name one individual/locus that is missing data.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   6. Close the file.
3. Conducting a PCA.
   1. Return to R.
   2. Type **getwd()** to determine what your default working directory is. You can either move your data file into this folder or you can type the command **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/Lab1-adegenet”)**
      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 working in the correct directory. Load the data by typing the command

obj1 <- read.structure("GWTE.3448A.N67.stru")

* 1. You will then be prompted to describe the data file. Add the information below. **(Note: For “Which other optional columns should be read…”, just hit return (we have no other optional columns).**



67

**Just hit “enter” or “return”**

3448

* 1. Once the data are loaded, type the following command:

**X <- scaleGen(obj1, NA.method="mean")**

[This command converts the missing data (-9) to the average allele frequency for the entire data set so that is it usable for PCA]

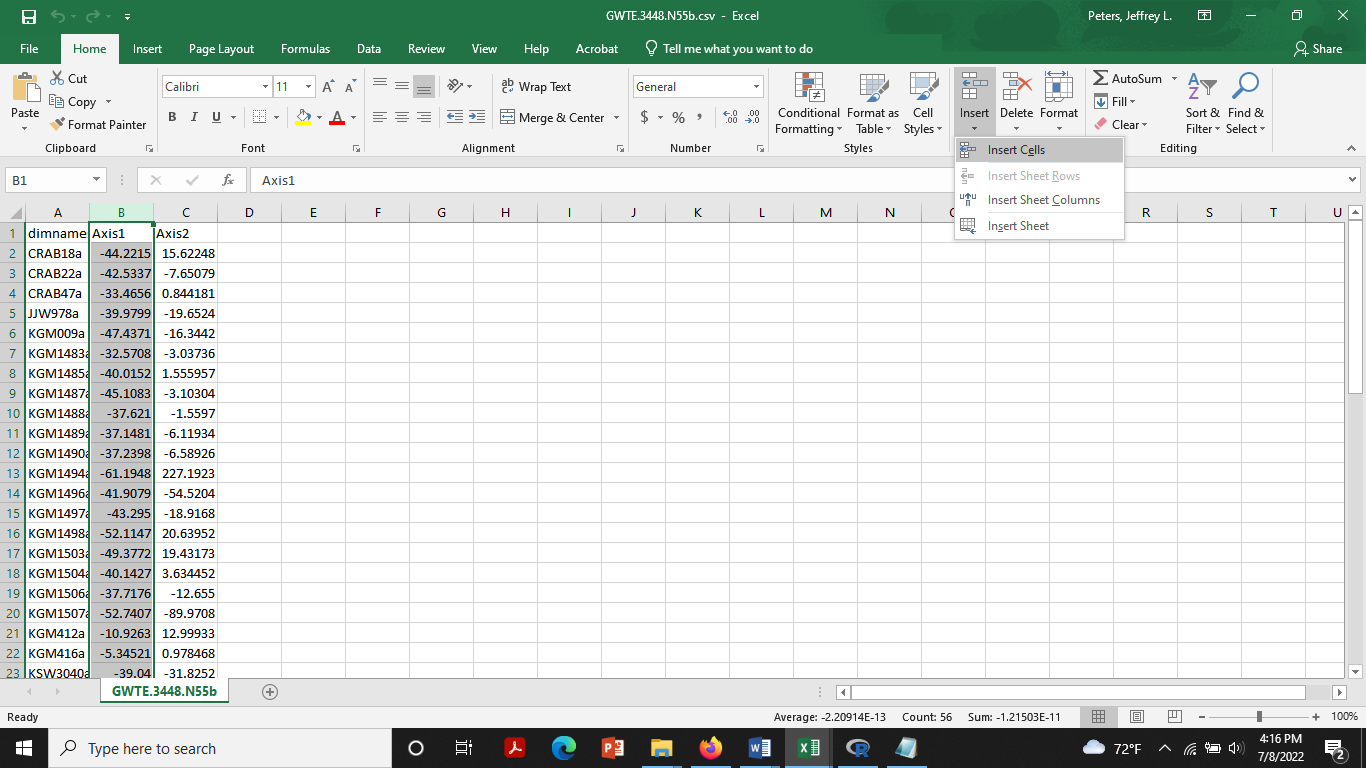
* 1. Tell the program to conduct a PCA by typing the following command.

**pca1 <- dudi.pca(X,cent=FALSE,scale=FALSE,scannf=FALSE,nf=2)**

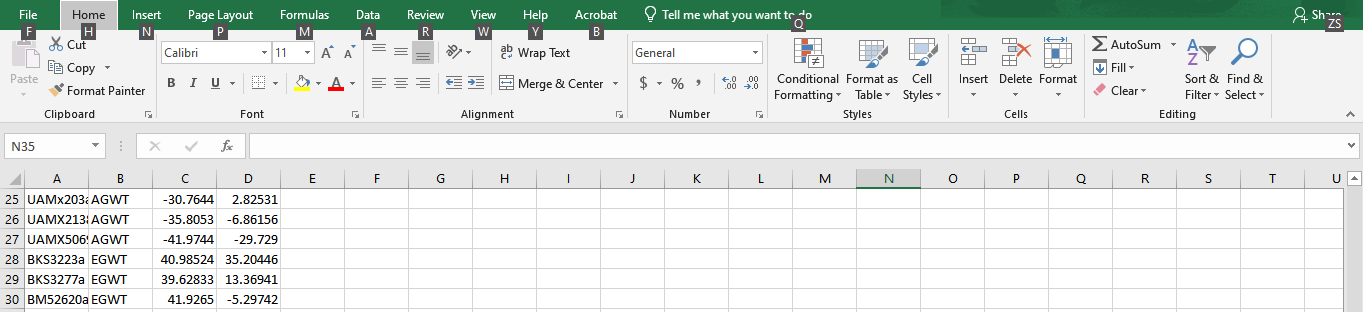
* 1. Export the results in a csv file by typing the following command:

**write.csv(cbind(dimnames(X)[[1]], pca1$li), "GWTE.3448.N67.csv", row.names = F)**

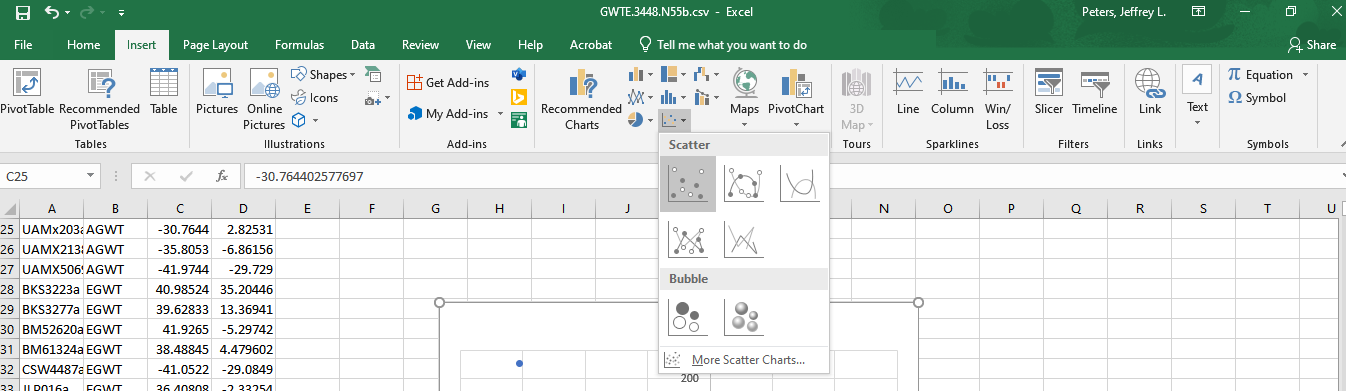
1. Examining the results.
   1. Open your results file in excel.
   2. The first column lists the 67 individuals sampled.
   3. The second column lists the values for principal component 1 (PC1; the axis explaining the highest percentage of variation among samples). The third column lists the values for PC2.
   4. The first 26 individuals (rows 2-27) were sampled in North America from widespread locations, including Alaska, Saskatchewan, North Dakota, Nova Scotia, etc.
   5. The remaining 29 individuals (rows 28-56) were sampled from Europe and Asia.
   6. To insert this information in your spreadsheet, highlight column B, then choose “Home” from the toolbar, click “Insert”, and then choose “Insert cells” from the dropdown menu.



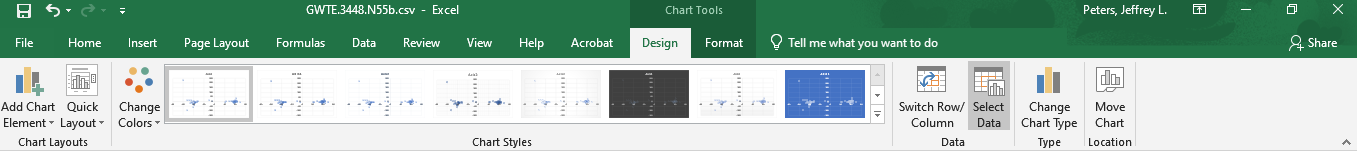
* 1. Add “AGWT” in column B for the first 26 individuals, and “EGWT” for the remaining individuals.



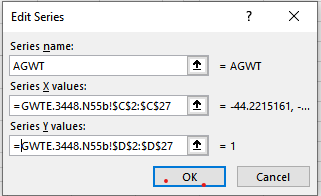
* 1. Make a scatterplot. The easiest way to do this is to:
     1. Highlight columns C & D.
     2. From the toolbar, choose the “Insert” tab.
     3. From the “Charts” panel, choose the scatterplot option (the icon with a bunch of dots), and then select the top-left most option from the dropdown menu.



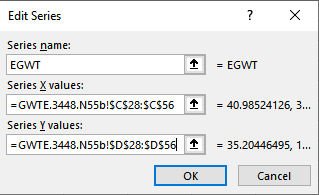
* + 1. Click on your chart, and then under the “Design” tab, choose, “Select Data”.



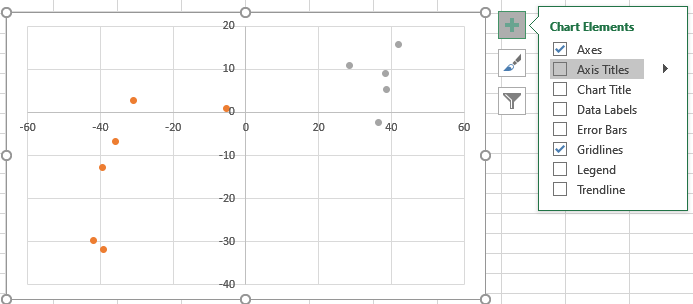
* + 1. Click “Add”
    2. For “Series name”, type “AGWT” (without the quotes)
    3. Series X values: choose the cells in column C (PC1) corresponding to your American Green-winged Teal (AGWT) [=GWTE.3448.N67!$C$2:$C$27]
    4. Series Y values: choose the cells in column D (PC2) corresponding to your AGWT [=GWTE.3448.N67!$D$2:$D$27]



* + 1. Click OK. You should now see color-points in your scatterplot corresponding to your AGWT samples.
    2. Click “Add”
    3. For “Series name”, type “EGWT” (without the quotes)
    4. Series X values: choose the cells in column C (PC1) corresponding to your Eurasian Green-winged Teal (EGWT) [=GWTE.3448.N67!$C$28:$C$68]
    5. Series Y values: choose the cells in column D (PC2) corresponding to your EGWT [=GWTE.3448.N67!$D$28:$D$68]



* + 1. Click OK. You should now see color-points in your scatterplot corresponding to your AGWT samples
    2. Label your axes: click on your chart, then choose the + icon and check “Axis Titles”. Change the x-axis to “PC1” and the y-axis to “PC2” by double-clicking on the default axes titles that were added.



* + 1. Copy your scatterplot and paste it here (click on your chart in excel and hit Ctrl-C; select the next (blank) line in this document and hit Ctrl-V):

1. Describe the results
   1. In the Word Document, provide a figure legend, briefly describing what the plot represents. “Figure 1. ….”
   2. Think about answers to the following questions (we will address these as a group):
      1. Is there evidence for population structure between American Green-winged Teals and Eurasian Green-winged Teals? Do they appear to be discrete populations? Explain.
      2. Is there evidence of gene flow (hybrids)? (Include the names of any individuals that appear to be hybrids.)
      3. Is there any evidence of dispersal between continents? (Include the names of any individuals that appear to have dispersed, and include the direction of dispersal (e.g., Eurasia to North America or vice versa)