Lab 2. Population assignment Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The purpose of today’s lab is to use multilocus genotype data to test for population structure and to assign individuals to each population. The program STRUCTURE does this by attempting to maximize Hardy-Weinberg Equilibrium and minimizing Linkage Disequilibrium by clustering samples into populations. The program does many iterations (we’ll do 200,000) of clustering to meet these conditions. STRUCTURE is available at: http://pritchardlab.stanford.edu/structure\_software/release\_versions/v2.3.4/html/structure.html

1. We will continue to work with the ddRAD-seq data obtained from green-winged teal. However, STRUCTURE is computationally intensive and requires several days to run the full data set. For this exercise, I have subsampled only 60 loci.
2. Input file = GWTE.N67.60A.stru
3. Open the data file in Microsoft Excel—the format is the same as the input used for adegenet.
	1. The first column contains the names of each individual that was genotyped. There are “a” & “b” entries for each individual. These are the two alleles sampled at each locus.
	2. The 2nd – 61st column contains allelic information for each locus with the header giving the name of the locus. Different numbers indicate different alleles.
		1. An entry of “-9” indicates that data are missing for the locus for that individual.
4. Open Structure and Create a “New Project” from the “File” dropdown menu.
5. A Project Window will appear.



1. “Name the project” with an informative name (e.g., “GWTE.N67.60A”).
2. “Select the directory” where you saved your data file.
3. “Choose data file” (GWTE.N67.60A.stru); click “Next”
4. Fill in the appropriate information.
	1. Number of individuals: 67
	2. Ploidy of data: 2
	3. Number of loci: 60
	4. Missing data value: -9
5. Click “Next”.
6. Check the box for “Row of marker names” to tell STRUCTURE that the name of each locus is provided. Click “Next”.
7. Check the box for “Individual ID for each individual”, because this information is given in the data file. Click “Finish”.
8. A pop-up window will appear that gives a summary of your project file that you just created. If all looks okay, click “Proceed”.



67

1. The Data should appear in the “Project Data” window. If the Project Data does not look like my screenshot below, close the project and try again. If it looks the same, save your project.



Parameter Set

1. Now we need to tell STRUCTURE what parameters to use in the analysis.
2. Click “Parameter Set” > “New…”
3. A pop-up window will appear labeled “New Parameter Set”.
4. Enter **50000** steps of Burnin and **250000** MCMC Reps after Burnin. (The burnin is used to find “good clusterings” before the actual data collection begins. Typically, we would want to run at least 100000 burnin and 500000 sampling steps; for the full ddRAD-seq data set, I would run 500000 burnin and 2000000 sampling steps.) DO NOT CLICK OK yet.
5. Click on “Ancestry Model”. Select the **Admixture Model**, which tests for mixed ancestry.
6. Click on “Allele Frequency Model”. There are two options here (ignoring “Infer Lambda, which is for more advanced analyses): “Allele Frequencies Correlated” & “Allele Frequencies Independent”. Select **Allele Frequencies Independent**; this model is simpler than the correlated frequencies model and works well for these data. If you have populations that have similar allele frequencies, then the correlated model may perform better.
7. Click OK.
8. Give your Parameter set a name. I usually do something like IND.ADM.200000 so I know it is the independent allele frequency + admixture model with 200,000 sampling steps. Click OK.
9. Save your Project.

Running the program

1. You are now ready to run STRUCTURE. Click “Project” > “Start a Job”
2. Select your parameter set and “Set K from” 1 to 3, where K is the number of populations. This tells STRUCTURE to examine 3 different models, assuming that the number of populations can range between 1 & 3. To be thorough, we should examine many more if analyzing data for a research project.
3. Click “Start”.
4. You should see numbers scrolling through the bottom window of your screen. These are the iterations that are being run as STRUCTURE is testing Hardy-Weingberg Equilibrium and Linkage Disequilibrium for various clusters of individuals. When this stops, your job is finished.
5. STRUCTURE will examine the k = 1 population model first, then it will automatically proceed to the k = 2 population model. It should take about 5 minutes or less to finish each run.
6. When the k = 1 model completes, a results folder will appear in the left top-left panel of your window. Choose the K=1 run (which will be proceeded by your name for your Parameter Set). This will cause the output to appear in the top, right panel.



1. Scroll through the output until you find “Estimated Ln Prob of Data” (this is the log likelihood). What is the value given? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. Next lists the assignment probabilities of each individual. They will all be 1.000 since you defined a 1 population model (each individual has a 100% probability of being from that population).
3. Wait for the K = 2 model to complete.
4. What is the “Estimated Ln Prob of Data”? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
5. The higher the “Estimated Ln Prob of Data”, the better supported the model.
6. Plot your Ln Prob scores on a graph for K = 1-3 populations, label the axes, and draw a line connecting the points.
7. The model with the highest (least negative) log likelihood is the best supported model. How many populations are supported by the data? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
8. Click on the results for the K = 2 model. At the top of the right panel, click on “Bar plot” > “Show” from the toolbar. The first 26 bars are assignment probabilities for American Green-winged Teal, and the final 41 bars are for Eurasian Green-winged Teal.

Questions for group discussion

1. Do any patterns emerge?
2. Individuals might receive low assignment probabilities because they have an admixed ancestry or because there is low power in the data for assignment. Given that we examined 60 loci in this data set, how do you interpret the evidence of mixed ancestry that you see across all individuals?
3. Are there any cases where you think admixed history (hybridization) is a better explanation than low power?
4. Is there any evidence of vagrants?
5. Examine the bar plot for K = 3. How does this differ from K = 2?