Lab 4. Islands of Divergence Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The purpose of today’s lab is to use genomic data to test for a role of selection in divergence. We will continue working with the Eurasian Green-winged Teal (*Anas crecca crecca*) and the American Green-winged Teal (*Anas crecca carolinensis*). We will use BayeScan to test for outlier loci (i.e., candidate loci for being under selection) that deviate from background levels of divergence. We will also construct Manhattan Plots that will be examined for islands of divergence. Because BayeScan is computationally intensive, we will restrict our analyses to chromosome 2 only.

**Bayescan-overview**

1. BayeScan can be downloaded from <http://cmpg.unibe.ch/software/BayeScan/>
2. In brief, the purpose of BayeScan is to identify candidate loci that are under natural selection from a sample of genetic data collected at multiple loci. It does so by comparing allele frequencies between two or more populations, and identifying loci that have higher levels of differentiation relative to overall levels of differentiation.
3. BayeScan assumes an island model of differentiation, where there are two or more populations connected by gene flow.
4. The statistic **alpha** is measure of how much the divergence at a given locus deviates from background levels of divergence. A positive value of alpha suggests diversifying (divergent) selection (more divergent than expected), whereas a negative value suggests balancing selection or **purifying selection** (less divergent than expected).

**Input File**

1. Because BayeScan is computationally intensive, we will focus our analyses on Chromosome 2 only.
2. Open **GWTE.chr2.bayescan** using WordPad. This file gives the alleles frequencies for each population and each locus.
	1. The first line gives the number of loci in the data set. How many loci?\_\_\_\_\_\_\_\_\_\_\_\_
	2. The next line gives the number of populations to be examined. How many populations?\_\_\_\_\_\_
	3. The following lines provide the allele frequencies for population 1 (*crecca*) and population 2 (*carolinensis*), separately. Population 1 is listed first; you will need to scroll down to about the middle of the document to find population 2 (or hit Ctrl-F and search for [pop]=2).
	4. The first column for the data gives the locus #. The second column gives the number of chromosomes that were typed (i.e., sample size or *n*, where *n* = 2*N* & *N* is he number of individuals genotyped). The third column gives the number of different alleles. The following columns are the number of samples for each allele.
	5. **What are the allele frequencies for locus 5 for population 1 and population 2?**

→ Population 1 (*crecca*):

→ Population 2 (*carolinensis*):

**Running BayeScan**

1. Execute BayeScan.
	1. Under “Input”, click on the icon with “**…**” next to “**Genotypes data**”. Find your file “**GWTE.chr2.bayescan**” and click “Open”.
	2. Under Output, change the “**Output files prefix**” to “**GWTE.chr2.out**”
	3. Click on the “**Options**” tab. This tab gives information about the run parameters. For example, you can change how many iterations (i.e., chains) to run. Typically, we want to run many more to fully explore the data, but given time constraints, we will use the default options. Note that there are parameters for the main chain (under “Sample size”) and for Pilot runs. The pilot runs help explore the data more efficiently.
	4. **How many chains will we run (Sample size)?**\_\_\_\_\_\_\_\_\_ **How many Pilot runs?**\_\_\_\_\_\_\_
	5. The program will now run through a series of iterations to calculate Fst, determine background levels of divergence, and calculate the probability that each locus deviates from background levels (i.e., evidence of selection).
	6. Return to the “**Main**” tab and click “**Start**”
	7. If the analysis is running properly, you should see a file called **GWTE.chr2\_Verif.txt** popup on your screen (you can close this file), and see the “Total time” estimate in the lower left of the BayeScan screen fluctuating.
	8. 
	9. This analysis will take about 30 minutes or more to complete, depending on your computer speed. If you find it takes much longer than that, I can give you the results file.
2. Examine the output file
	1. Open your output file “GWTE.chr2\_Fst.txt”
	2. Copy the contents of the file and paste them into Excel.
		1. Highlight column A.
		2. From the “Data” tab, select “Text to Columns”
		3. Make sure “Delimited” is selected and click “Next>”
		4. Under “Delimiters”, check “Space”, and click “Finish”
		5. Your results should now be six columns. In cell A1, type **Locus\_ID**.
		6. The next three columns provide statistics related to selection (i.e., how much better is the model with selection than the model without selection).
			1. “prob”: The posterior probability for the model including selection (0 indicates no evidence of selection at the locus, and 1 indicates nearly 100% probability of selection.)
			2. “log10(PO)”: how much better a model with selection is than a model without selection, where 1000 is equivalent to infinity.
			3. “qval”: q-value, which is equivalent to a p-value; the probability of obtaining the observed or more extreme alpha value if the locus is selectively neutral.
		7. The next two columns provide information about the strength of selection
			1. “alpha”: The estimated alpha coefficient indicating strength and direction of selection. A positive value suggests diversifying selection, whereas negative values suggest balancing or purifying selection. More extreme values suggest a stronger influence of selection.
			2. “fst”: The Fst value between the two populations. Note: this is not calculated in the same way that we calculated Fst earlier; it is an analog of Fst based on model averaging (it is correlated with our traditional Fst).
		8. Visualize the results by making a scatterplot showing the relationship between the posterior probability and alpha.
			1. Highlight the columns “prob” and “alpha”.
			2. Select the Insert tab and choose scatterplot:



* + - 1. In this plot, the probability of the model with selection (prob) is on the x-axis; points to the left indicate loci consistent with neutrality, and points to the right suggest an influence of selection. The direction and magnitude of selection (alpha) is on the y-axis.
			2. Label your axes by clicking on your chart, choosing the +, and checking the box for axes titles. You can then select the default titles and rename them directly on your graph.
			3. Click on figure and click Ctrl-C to copy; Paste it in the box below (Ctrl-V). Give your figure an informative figure legend.

Figure 1.

* + 1. *Group discussion: What can you conclude about the influence of selection in these ducks from this plot?*
		2. Sort your data by qvalue (Highlight all six columns, Data > Sort > qvalue).
		3. If we set our threshold at qvalue = 0.001 for determining significance, which loci (if any) appear to be under diversifying selection (a significantly positive alpha)?

→

* + 1. Which loci (if any) appear to be significantly influenced by purifying or balancing selection (a significantly negative alpha)?

→

* 1. Is there evidence of an island of divergence?
		1. Visualize the data by making a scatterplot of chromosome number (note that the loci in your list are ranked by chromosome position) and Fst. **Label the x-axis “chromosome position”, and the y-axis “Fst”.**
		2. Recolor loci that appear to be under selection.
			1. Click on your chart. Go to the “**Design**” tab, and click on “**Select Data**”. Click **Add**.
			2. For **Series Name**, type **Outlier**
			3. Click in the field for **Series x values**; from column A (Locus ID), select all the cells associated with a significantly positive alpha (e.g., cells A1-A9) [
			4. Click in the field for **Series x values** and delete the ={1}; then from column F (fst) select all the cells associated with a significantly positive alpha.
			5. Click OK. You should see that some of the data points (each of which represents a different locus) changes colors.
			6. Copy and paste your graph in the box below and give your figure an informative legend.

Figure 2.

Questions for group discussion

1. Is there evidence of an island of divergence?
2. What can you conclude about the influence of selection in these ducks?