F_{ST} Outliers:

A targeted approach to look for patterns of local selection is to look for FST outliers. This approach compares the level of differentiation at a given locus to levels of differentiation across the genome (or transcriptome) to determine whether there is evidence of selection. For example, a locus that is significantly more divergent than average has likely been affected by positive or directional selection. Similarly, lower than usual FST suggests either balancing or purifying selection. To conduct an FST outlier analysis we will use the program BayeScan, which uses a Bayesian framework to estimate the probability that each locus has been acted on by selection.

The Bayescan program and the script for making Bayescan input are located in your scripts folder on partition 2 on the high_mem mode.

a. Create input for BayeScan:

Script: make_bayescan_input.py
Input file: VQSR_PASS_SNPS.vcf
Output files: *bayes_input*.txt, snpkey.txt

As input, BayeScan requires allele frequency counts for each SNP in each population. To generate this input file we will use the script make_bayescan_input.py. This script extracts allele count data from a .vcf file. Because this analysis does not require individual level data, it is less critical to limit the analysis to SNPs with full coverage across all individuals. Therefore, by default this script will include any SNP that has high quality genotype information from at least 5 individuals in each population (the number of individuals required per population can be changed with an optional third argument on the command line). This allows us to increase the minimum genotype quality score while still maintaining a large number of SNPs.

Execute the script with the following command:

```
../scripts/make_bayescan_input.py VQSR_PASS_SNPS.vcf 20
```

The third argument on the command line is an integer specifying the minimum genotype quality for a SNP to be used in the analysis. Depending on the size of the .vcf file this process may take several minutes. This step will generate two files: 1) bayes_input.txt will include the information needed by BayeScan to conduct the analysis and 2) snpkey.txt, which is a reference for you to move from the SNP numbers used by BayeScan to the contig and base pair location of the SNP in the reference contigs. This script will also print to your Terminal window the names of your populations along with the number that each one was given in the input file for BayeScan. Go ahead and open these 2 files in nano or less to get an idea of their formats. Also, check the number of SNPs that met the specified quality criteria.

b. Run the analysis for FST outliers using BayeScan:

Input file: *bayes_input*.txt
Output files: *bayes_input_fst*.txt, *bayes_input.sel, bayes_input_AccRte.txt, bayes_input_Verif.txt*
Program: BayeScan

Execute the program with the following command:

```
../scripts/BayeScan2.1_linux64bits bayes_input.txt
```

The first argument on the command line is the executable program. The exact name may vary depending on your operating system and version.

This analysis will take a while. When it is done it will have created several output files. The main file of interest is bayes_input_fst.txt. One way to visualize the results is to use a function in R that is included with the BayeScan

distribution. It is distributed in the script plot_R.r (in the folder called R_scripts). To load this function into R, copy the file to your computer, open the R console and type the following command:

```
source ("path/to/plot_R.r")
```

Then, run the function on the BayeScan results using the following command:

```
plot_bayescan("path/to/bayes_input_fst.txt", FDR=0.05)
```

This function will calculate the Posterior Odds (PO) threshold leading to a false discovery rate of no more that 5%. Using the posterior odds value it will find and list the SNPs that are FST outliers. It will also produce a figure of FST vs. log(PO).

Additionally, you can look at what SNPs are nearly significant by opening the file bayes_input_fst.txt in Excel and sorting based on the second column, which corresponds to the probability that a SNP has been affected by selection. The largest values indicate the SNPs that are most likely to have been affected by selection.