# 2.6: Identification of FST outliers

In this section, we will visualize genotype data and summary information using R. For each locus, we will calculate the fixation index, FST, a measure of differentiation between two populations due to genetic structure. Based on our results, we will then choose several loci of interest to characterize further.

Note that there are many programs that could calculate FST for us quite simply, but as the calculation is actually rather straightforward, we will take advantage of this to reveal some underlying mechanics of how one could calculate a single value over many loci/samples using R.

### Objectives:

• Visualize and interpret genetic data in R  
• Characterize particular loci of interest

### Protocol:

1. Open R and load the data table structure\_4R.txt. This data table was created as one of the output files from the TabToStructuRe.pl script we ran yesterday. You will need to either specify the complete filepath, as shown here, or change your working directory to the folder containing your data file:

aip\_gts <- read.table("/Users/weissem/Desktop/AiptasiaPopGen/structtesting/structure\_4R.txt", header=T)

1. Take a look at this table. The head command will show us the first few rows of the aip\_gts table and the [,1:4] indicates that we would just like to view the first 4 columns of this file, rather than printing out hundreds of columns of SNPs!

The first column indicates the sample name, the second the population ID, and each remaining column gives information for a different SNP.

head(aip\_gts[,1:4])

## ID POP scaffold100size601457\_139826\_139861\_R\_7  
## 1 BTCA01 1 -9  
## 2 BTCA02 1 0  
## 3 BTCA03 1 0  
## 4 BTCA04 1 0  
## 5 BTCA05 1 0  
## 6 BTCA06 1 0  
## scaffold100size601457\_314922\_314957\_R\_2  
## 1 -9.0  
## 2 0.5  
## 3 0.0  
## 4 0.0  
## 5 0.0  
## 6 0.0

In this data file, there is only 1 line per sample, and genotypes are coded in the following way:

|  |  |
| --- | --- |
| Genotype | Value |
| Homozygous reference | 0 |
| Heterozygous | 0.5 |
| Homozygous alternate | 1 |
| Missing genotype | -9 |

1. How many loci are in this file?

One way we can find out is to store all the column names as a 'vector', essentially a sequence of data elements, and count the number of values in the vector using the length function:

loci <- colnames(aip\_gts)

head(loci)

## [1] "ID"   
## [2] "POP"   
## [3] "scaffold100size601457\_139826\_139861\_R\_7"   
## [4] "scaffold100size601457\_314922\_314957\_R\_2"   
## [5] "scaffold100size601457\_458335\_458370\_R\_11"  
## [6] "scaffold100size601457\_511844\_511879\_F\_4"

length(loci) - 2 ## subtract 2 for the population and sample ID columns ## [1] 1070

1. Now pick two populations to compare. Integers in the POP column correspond to the following populations:  
   (1)BTCA, Cayo de Agua; (2) BTCR, Cayo Roldan; (3) BTIC, Isla Colon; (4) CAPG/CAIL, Galeta; (5) Laboratory strains/Unknown

To filter the data table to include only your two populations of interest, you can use the subset function. I have picked populations BTIC and BTCR, but you can pick any two. Make a new table containing only information for each of these populations:

popA <- subset(aip\_gts, aip\_gts$POP == 2)  
popB <- subset(aip\_gts, aip\_gts$POP == 3)

1. Create a new empty vector to hold the 1070 values of FST that we are about to calculate:

Fst <- NULL

1. Now we will loop through all SNP columns in the table and calculate FST for each one. FST values near 0 indicate minimal population differentiation and values near 1 indicate almost complete population differentiation. FST is given by the following equation:

,

where *HT* is the expected heterozygosity of the total population under Hardy-Weinberg equilibrium (2*pq*), and *HS* is the average expected heterozygosity of each subpopulation. We will need several values including frequency of the reference allele in subpopulation A or B, expected heterozygosity in each subpopulation under Hardy-Weinberg, and the average allele frequencies for the total population.

If you are new to population genetics and/or computer coding, you don't need to worry too much about FST or what is going on in the loop below. Our focus is mostly on being able to visualize our genotype data in a meaningful way. But do ask if you are curious and would like to learn more ☺!

Run the following code as a single chunk to calculate FST for each SNP:

for(i in 3:length(loci)){  
#count genotypes for population A  
 gtcallsA <- popA[,i]  
 hom.RefA <- sum(gtcallsA == 0)  
 hom.AltA <- sum(gtcallsA == 1)   
 hetsA <- sum(gtcallsA == 0.5)   
  
#count genotypes for population B   
 gtcallsB <- popB[,i]   
 hom.RefB <- sum(gtcallsB == 0)  
 hom.AltB <- sum(gtcallsB == 1)  
 hetsB <- sum(gtcallsB == 0.5)  
  
#calculate allele frequencies for population A, B and total population  
 RefFreq.A <- ((hom.RefA \* 2) + hetsA)/((hom.RefA+hetsA+hom.AltA)\*2)  
 AltFreq.A <- 1-RefFreq.A #q = 1-p   
   
 RefFreq.B <- ((hom.RefB \* 2) + hetsB)/((hom.RefB+hetsB+hom.AltB)\*2)  
 AltFreq.B <- 1-RefFreq.B  
   
 AvgRef <- (RefFreq.A + RefFreq.B)/2   
 AvgAlt <- (AltFreq.A + AltFreq.B)/2  
   
#calcuate heterozygosity in population A, B and total population.  
 HsA <- 2\*RefFreq.A\*AltFreq.A #expected heterozygosity in popA = 2pq  
 HsB <- 2\*RefFreq.B\*AltFreq.B  
 HS <- (HsB+HsA)/2 # average subpopulation heterozygosity  
 HT <- 2\*AvgRef\*AvgAlt # average heterozygosity for the total population  
   
 Fst[i] <- (HT-HS)/HT # store the Fst calculated for each locus at position i  
}

1. Make a new data table, with the name of the locus in the first column, and the corresponding FST value in the second column:

FstData <- cbind.data.frame(loci[3:length(loci)], Fst[3:length(Fst)])

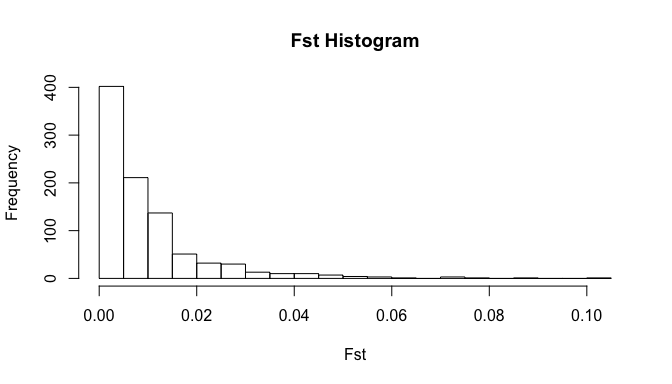
colnames(FstData) <- c("Locus","Fst")

head(FstData)

## Locus Fst  
## 1 scaffold100size601457\_139826\_139861\_R\_7 3.183865e-03  
## 2 scaffold100size601457\_314922\_314957\_R\_2 1.520103e-05  
## 3 scaffold100size601457\_458335\_458370\_R\_11 6.531290e-03  
## 4 scaffold100size601457\_511844\_511879\_F\_4 2.623708e-05  
## 5 scaffold100size601457\_99625\_99660\_R\_31 2.614853e-04  
## 6 scaffold1011size19531\_2710\_2745\_R\_9 1.575985e-04

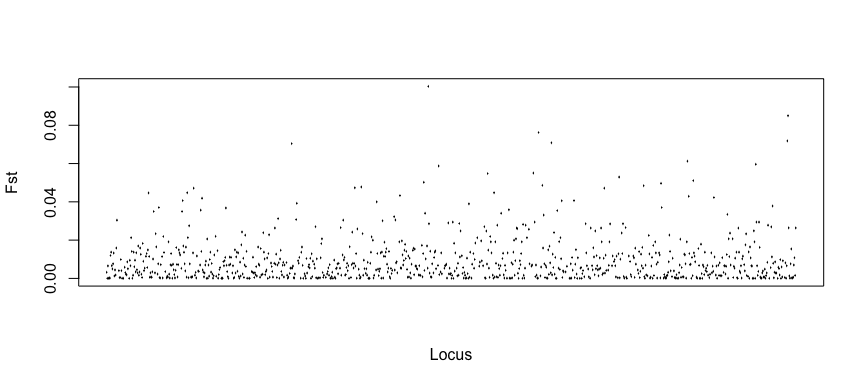
1. How are these values distributed? Are most polymorphisms highly differentiated between the two populations, or does there appear to be a lot of gene flow between populations?

hist(FstData$Fst, main="Fst Histogram", xlab="Fst", breaks=20)



1. Try running the hist code with a few different values of breaks. You can also see more information by running ?hist.
2. Now let's make a scatterplot of our FST values. We can make a basic scatterplot by running the following code:

plot(FstData$Locus, FstData$Fst, xaxt='n', xlab="Locus", ylab="Fst")



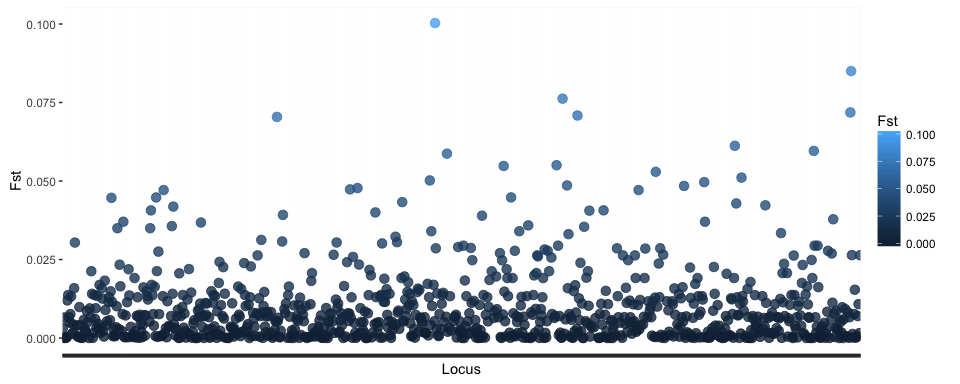
# the xaxt option removes the x-axis tick labels

1. Many people also really like the package ggplot2 for making nice looking graphs in R. To make a similar graph with ggplot, load the ggplot2 libary:

library(ggplot2)  
#you may need to install the ggplot package if it is not already installed by running install.packages("ggplot2")

1. And run:

ggplot(FstData, aes(Locus, Fst, colour=Fst)) + geom\_point(alpha=0.8, size=3) +xlab("Locus") + theme(axis.text.x = element\_blank())



1. Try changing some of the parameters to see what that does to the look of the graph.
2. Once you are finished, you may be interested in selecting all the loci above a certain FST value:

subset(FstData, Fst > 0.075)

## Locus Fst  
## 500 scaffold299size278964\_41549\_41584\_R\_27 0.10030564  
## 671 scaffold4050size1311\_1258\_1293\_R\_17 0.07622575  
## 1058 scaffold97size618721\_393953\_393988\_R\_30 0.08500758

1. Choose 3 loci of interest to investigate further on the Aiptasia Genome Browser at <http://aiptasia.reefgenomics.org/jbrowse/>.

*Does the SNP occur in a coding or non-coding portion of a gene or in intergenic DNA?*

1. To characterize the gene the SNP is associated with or closest to, we will perform a BLAST search to find other genes with similar nucleotide sequence.

* Click on the Aiptasia gene model
* Copy the fasta sequence for the gene model.
* In your web browser, navigate to NCBI's blast homepage http://blast.ncbi.nlm.nih.gov/Blast.cgi
* Click on Nucleotide Blast to blast DNA sequences against a DNA database, and paste the fasta sequence in the white box.
* You can try out different parameters, or BLAST with the default settings

*Did you get any good hits? How long was your best hit? What was its E-value? What organism was it from?*

* If you did not get any hits, you could also try a blastx search, which will translate the query sequence and search a protein sequence database. It takes a bit longer, but can detect homologous genes that are less similar at the DNA sequence level.