**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**San Diego State University**

**Dr. Arun Sethuraman**

**Lab 5 Population Genomics - Part 2**

For today’s lab, we’re going to switch gears a bit to work on a different system (diploid), so as to allow us to fully utilize population genomics functions and libraries that are available. We are working with a VCF file generated by my lab from global populations of *Harmonia axyridis*, or commonly known as the Harlequin lady beetle. Details on how these data were generated are in the manuscript (Li et al., in review) that I have posted on your Canvas page. We will briefly spend some time going over these details in lab prior to starting our lab. The VCF file that we will be utilizing in our analyses is also on Canvas (harmonia.vcf). NOTE: Please don’t share this file, or the manuscript anywhere - these are still in review, and will be made publicly available post the peer review process, but please don’t jeopardize this process! Thanks!

We will be utilizing a tool that’s pre-installed on your JetStream VM’s called VCFTools (<https://vcftools.github.io/man_latest.html>) for computing a lot of the statistics, and then we will use R to visualize them. So go ahead and create a new folder, called Week11 on your JetStream VM, then download the VCF file to that folder first. It might be easier to download it from Canvas inside your VM, since transfer does take a bit of time otherwise.

We do however need to install a couple of other tools that will allow us to do estimation of population structure.

So to install these:

#to install plink2, go to your Tools folder

mkdir plink2

cd plink2

wget <https://s3.amazonaws.com/plink2-assets/alpha2/plink2_linux_x86_64.zip>

unzip plink\*.zip

pwd

#Set the path to this folder

#to install ADMIXTURE, go back to your Tools folder

wget <https://dalexander.github.io/admixture/binaries/admixture_linux-1.3.0.tar.gz>

tar -xzvf admixture\*.gz

cd dist

cd admixture\_\*

pwd

#set the path again to this folder

**Exercise 1: Testing for Hardy-Weinberg Equilibrium and removing loci that are out of HWE**

Recall that loci that are in HWE are evolving neutrally - i.e. these are loci that are exhibiting characteristics of being derived from a large population that is randomly mating, under no selection, with no new mutations, no migration. In other words, since they’re not changing in allele frequencies over time, they are not evolving. So for most population genomic analyses, we would expect to keep loci that are in HWE, and toss out loci that aren’t.

To test for HWE, the function in vcftools is:

vcftools --vcf harmonia.vcf --hardy

Now look at the output file produced - what are the fields that are produced?

Now to create a “reduced” file with only loci that are in HWE at a p-value cutoff of 0.05:

vcftools --vcf harmonia.vcf --hwe 0.05 --recode --stdout > harmonia\_hwe.vcf

You can view how many sites were filtered by just using:

vcftools --vcf harmonia\_hwe.vcf

**Exercise 2: Estimating population structure**

To convert the VCF file to a format that is accessible by ADMIXTURE, we will use plink2.

plink2 --vcf harmonia\_hwe.vcf --make-bed --out harmonia\_hwe --allow-extra-chr

awk '{$1="0";print $0}' harmonia\_hwe.bim > harmonia\_hwe.bim.tmp

mv harmonia\_hwe.bim.tmp harmonia\_hwe.bim

Now to run admixture to test different number of subpopulations from 2-5:

admixture --cv harmonia\_hwe.bed 2 > log2.out

admixture --cv harmonia\_hwe.bed 3 > log3.out

admixture --cv harmonia\_hwe.bed 4 > log4.out

admixture --cv harmonia\_hwe.bed 5 > log5.out

To estimate which one has the least cross validation error:

grep "CV" \*out | awk '{print $3,$4}' | sed -e 's/(//;s/)//;s/://;s/K=//' > harmonia.cv.error

grep "CV" \*out | awk '{print $3,$4}' | cut -c 4,7-20 > harmonia.cv.error

awk '/CV/ {print $3,$4}' \*out | cut -c 4,7-20 > harmonia.cv.error

Now just do:

cat harmonia.cv.error

What do you notice? Which has the least error? What does this say about the global population structure of the species?

Let’s go ahead and plot this to see if we can make sense of this population structure. To do this, we will use R to make what are called stacked barplots. Open R (just type R and hit enter).

x<-read.table("harmonia\_hwe.2.Q",header=FALSE)

barplot(t(as.matrix(x)),col=rainbow(2),xlab="Individuals",ylab="Ancestry",border=NA)

Chart, bar chart, histogram

Description automatically generated

Let’s take some time to interpret these results. Let’s also visualize the K=3 results:

y<-read.table("harmonia\_hwe.3.Q",header=FALSE)

barplot(t(as.matrix(y)),col=rainbow(3),xlab="Individuals",ylab="Ancestry",border=NA)

Can we interpret these?

**Exercise 3: Estimating outlier loci (Tajima’s D, Pi, and Fst)**

Recall that in the original dataset (harmonia.vcf), if some loci that are out of HWE are exhibiting extreme characteristics (i.e. they have strongly positive or negative Tajima’s D, or extreme values of Pi, and Fst), they are likely under some non-neutral evolutionary demographic process like selection. So to test for these, and obtain outlier loci, we can use vcftools again.

vcftools --vcf harmonia.vcf --window-pi 100 --window-pi-step 100

vcftools --vcf harmonia.vcf --TajimaD 100 #computes Tajima’s D in windows of 100 bases

vcftools --vcf harmonia.vcf --weir-fst-pop asia.txt --weir-fst-pop europe.txt --weir-fst-pop americas.txt --fst-window-size 100 --fst-window-step 100

#little Unix magic to format the output of these to be agreeable to the qqman package in R

sed -i 's/Scaff/scaff/g' out.\*

sed -i 's/[a-z]\*\_[a-z]\*\_//g' out.\*

Now let’s visualize these using R.

Inside R:

install.packages(“qqman”)

library(qqman)

A picture containing scatter chart

Description automatically generatedtajimad<-read.table("out.Tajima.D",header=TRUE)

tajimad\_nomissing<-na.omit(tajimad)

manhattan(tajimad\_nomissing,chr="CHROM",bp="BIN\_START",p="TajimaD",snp="N\_SNPS",logp=TRUE,ylab="Tajima’s D",ylim=c(-1.0,3.0))

Chart, scatter chart

Description automatically generated

diversity<-read.table("out.windowed.pi",header=TRUE)

manhattan(diversity,chr="CHROM",bp="BIN\_START",p="PI",snp="N\_VARIANTS",logp=FALSE,ylab="Pi",ylim=c(0,0.01))

Similary, to plot the Fst values

fst<-read.table("out.windowed.weir.fst",header=TRUE)

manhattan(fst,chr="CHROM",bp="BIN\_END",p="MEAN\_FST",snp="BIN\_START",logp=FALSE,ylab="Fst",ylim=c(0.0,1.1))

Chart, scatter chart

Description automatically generated

Now let’s try to make sense of this plot in the context of the paper.