**P/BIO 381 Spring 2017 Laura Caicedo-Quiroga**

**Homework #3 04/05/2017**

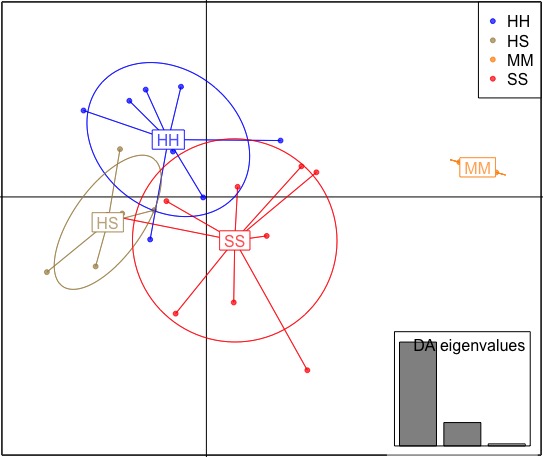
**Population genomic diversity and structure**

The main goal of this study is to compare the genetic structure of *Pisaster ochraceus* at various disease stages using SNPs from RNA sequencing. We aim to know if SNPs are useful for characterizing differences in health status. In addition, we will be using VCF tools for two filtering strategies on the SNPs and test whether different filters affect the structuring of our samples by health status. I used Discriminant Analysis of Principal Components, DAPC, to infer the clustering of samples of *P. ochraceus* based on their health status: SS, SH, HH, MM. DAPC aims to maximize discrimination between groups. This is a multivariate statistical approach that were variance is divided into between-group and within-group components (Jombart et al. 2010). DAPC first transforms data with a principal components analysis and then uses discriminant analysis to make clusters.

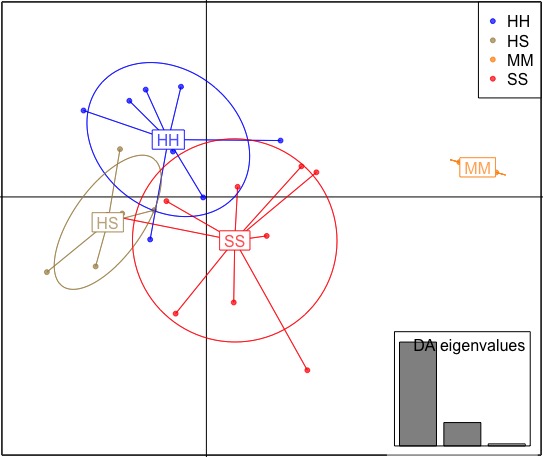
The vcf file used has 7.4 million SNPs that were previously called on a minimum read depth of 10. VCF tools was used to filter SNPs using two strategies. For the first filter, biallelic SNPs were kept, minor allele frequencies of 0.02 were selected, and I allowed for 30% of missing data. For the second filter, I kept all parameters with the change of minor allele frequencies of 0.05 considered low frequency variants. The two minor allele frequencies will be compared while keeping the analysis towards rare genetic variants that can reveal fine-scale genetic structuring and might reflect more recent mutations (Gompert et al. 2014). Below are the commands used:

* Filter 1: vcftools --gzvcf SSW\_by24inds.txt.vcf.gz --min-alleles 2 --max-alleles 2 --maf 0.02 --max-missing 0.7 --recode --out ~/scripts/SSW\_biallelic.MAF0.02.Miss0.7
* Filter 2: vcftools --gzvcf SSW\_by24inds.txt.vcf.gz --min-alleles 2 --max-alleles 2 --maf 0.05 --max-missing 0.7 --recode --out ~/scripts/SSW\_biallelic.MAF0.05.Miss0.7

Filter 1 retained 8452 (11%), while filter 2 kept 3418 (0.045%) of all SNPs respectively. The results from both analyses cluster the samples by their health status with DAPC, (Figures 1 and 2). Nevertheless, there is close clustering of HH, HS and SS with MM as the only cluster further from all.



**Figure 1**: DAPC Plot for Filter 1, allowing 30% of missing data and minor allele frequencies of 0.02. Individuals are represented as dots and ellipses as clusters. Each cluster is colored by the health status, blue: HH, green: HS, yellow: MM and red: SS. The lower left corner shows bar plots for 3 eigenvalues (ratio of variance between groups and within groups) of the discriminant analysis.



**Figure 2**: DAPC Plot for Filter 2, allowing 30% of missing data and minor allele frequencies of 0.05. Individuals are represented as dots and ellipses as clusters. Each cluster is colored by the health status, blue: HH, green: HS, yellow: MM and red: SS. The lower left corner shows bar plots for 3 eigenvalues (ratio of variance between groups and within groups) of the discriminant analysis.

While DAPC shows clustering of SNPs by health status, sick individuals (SS) and those that eventually became sick (HS) overlap with healthy individuals (HH). There are 15 SNPs that contribute the most to the clustering by health status for both filters.

I used a non-stringent filtering for SNPs with missing data (max-missing 0.7), allowing 30% for both filtering strategies. While loose, the filter for missing data might still be excluding SNPs that might account for variation between health status. My analysis allows a comparison between 0.02 and 0.05 as minor allele frequencies while still filtering for rare genetic variants. Less SNPs are kept when employing the second filter, nevertheless the results for the DAPC are mostly the same; therefore, differences in rare variant calling do not make a difference when searching for genetic clustering among individuals with different health status. Common variants might be responsible for clear clustering by health status and perhaps even location (intertidal vs. subtidal). The overlap between clusters, for both filtering strategies, might be indicative of loose clustering which might occur because of the sampling of individuals from one population.

It would be useful to increase sampling within the population as well as include other populations for comparisons between clustering by health status and by populations. As a next step in this study I would search for the 15 SNPs that contribute the most to the clustering observed and determine the function of the gene to which they belong.

**Literature cited:**

Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Molecular Ecology, 23:4555-4573.

Jombart, T., Devillard S., Balloux F. 2010. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC genetics* 11:94.

Code at: