**Assignment #3: Population genomic diversity and structure P/BIO 381 Spring 2017**

Kirsten Tyler

**Introduction:**

To compare the effects of different filtering strategies in VCFtools on our interpretations of population genomic structure in *Pisaster* ochraceus, a species of sea-star greatly suffering from sea-star wasting disease (SSWD), we altered two filters (the missing data filter and the minor allele frequency (MAF) filter) in our SNP analysis and plotted PCAs to visualize the differences.

**Methods:**

Using VCFtools, we chose the --remove-indv filter to remove individuals with >90% missing data (owing to potential low sequence quality: random noise) from the raw data because this might give us more powerful interpretations of analyses. By removing them, it might be possible to see potential similarities between analyzed groups that were hidden by random noise. Also, we chose to alter the MAF filter --maf to eliminate very rare SNPs (criteria <2% and more stringent <10%) in the data (owing to potential sequencing error). Plotting the principal component analysis (PCA), we observed any correlated genetic similarities between individuals grouped by locality and disease status.

To filter and remove individuals with >90% missing data from the raw dataset, we created a list with the –missing-indv command in VCFtools to visualize total missing data, then removed six individuals with the –remove-indv command (see <https://github.com/rkirstentyler/Ecological-Genomics-PBIO381/tree/master/Homework%20Assignments> for Rmd file with code for all analyses). After removing the six individuals, the MAF filter 0.02 and MAF filter 0.1 were applied to both data sets (original 24 and new 18 individuals) to create four new datasets (24 MAF 0.02; 24 MAF 0.1; 18 MAF 0.02; 18 MAF 0.1). We decided to keep the default filtering strategies used in our previous studies. Only biallelic sites were used (because <2 were probably errors) and we only kept sites where more than 80% of our samples have data (to reduce bias in analysis). Using Cyberduck, the files were downloaded onto a Mac hard drive then read into R. Vcf files were converted to genlight files and fields were updated for PCA analysis. We computed PCA analyses on the SNP genotypes for all four data sets then plotted the results with locality labels and disease status labels.

**Results:**

After applying the filters, we found a large difference in the number of SNPs kept between data sets (table 1). The data set with 18 individuals kept many more sites than the data set with 24 individuals regardless of the MAF filter used.

Comparing PCA plots between the data sets with 18 and 24 individuals, we found that generally the plots with 18 individuals showed tighter correlation between samples (figure 1). This finding is especially true for figure 1 comparing A and B and comparing E and F, which associates samples based on disease status. Individuals that started off healthy, then turned sick (HS) show a tighter correlation in the group with 18 samples than with 24 samples. Comparing PCA plots between MAF filters 0.2% and 1%, we found very little difference within or between groups. For example, when comparing figures 1 A and E for correlation in disease status, or comparing D and H for correlation in locality, although the structure varies between filters, there is little variation found in the correlation of samples.

**Figures:**

Table 1

|  |  |  |
| --- | --- | --- |
| # indv | MAF | # sites kept |
| 24 | 2 | 5317 |
| 18 | 2 | 12578 |
| 24 | 10 | 1100 |
| 18 | 10 | 3920 |

Number of sites kept per group after filtering raw data

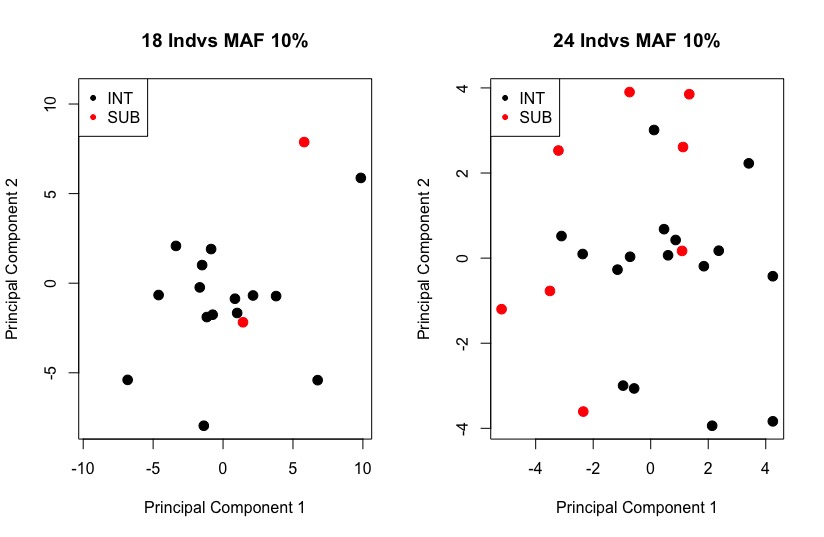
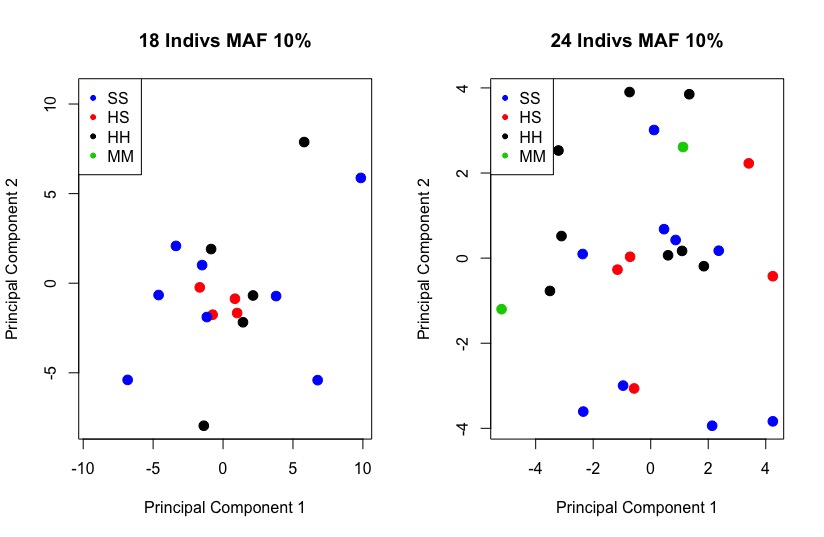


Fig. 1A Fig. 1B Fig. 1C Fig. 1D

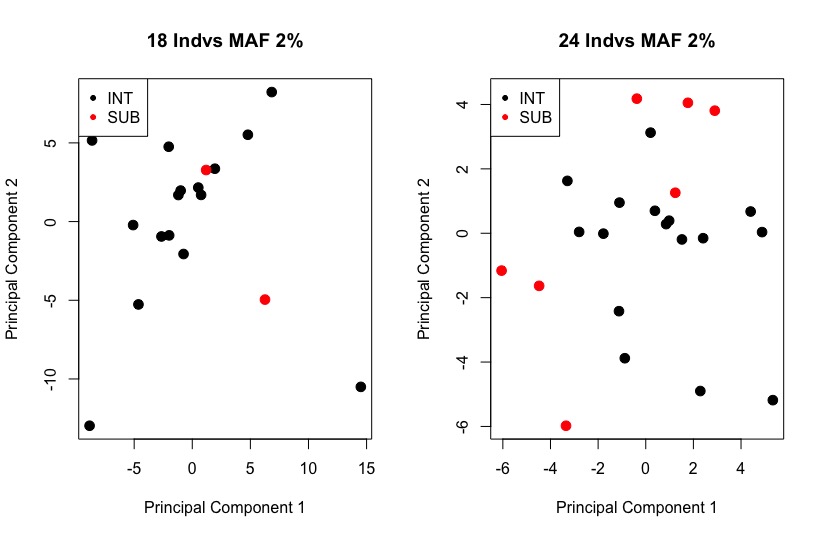
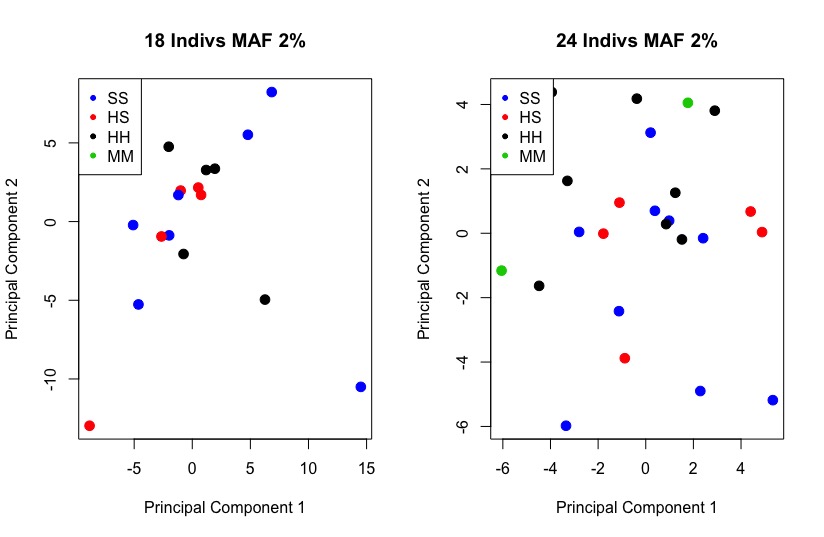


Fig. 1E Fig. 1F Fig 1G Fig 1H

Figure 1 A-H: PCA plots displaying the correlation between samples among groups in our study. The left four plots (A, B, E, F) exhibit findings for correlation among samples based on disease status. In these plots, colors represent disease status (SS = sick, HS = healthy turned sick, HH = healthy, MM = healthy turned sick then turned healthy). The right four plots (C, D, G, H) exhibit findings for correlation among samples based on locality (INT = intertidal, SUB = subtidal). The top row A, B, C, D exhibits MAF 10% between groups while the bottom row shows MAF 2%. The top and bottom rows are lined up by sample number for ease of comparison between sample numbers.

**Discussion:**

The filtering strategy --remove-indv may have removed some of the random noise from our raw data, allowing observation of a tighter correlation between samples in plots with 18 individuals (figure 1). This finding is also observed in the number of SNPs kept after applying the filters (table 1), leading us to believe that without the missing SNPs, we are able to see a clearer picture of the similarities between samples. Moreover, it may be possible that we can infer more genetic similarity between all individuals after removing the potential noise. As these individuals were collected from the same population, this is something we would expect. Applying a more stringent MAF filter did not result in more correlated samples (for example, comparing figure 1C with G) which may be the result of the –max-missing filter (which we used as a default filter in this analysis) removing those errors.

It is important to note that if this study were to be conducted again, we’d change our design in that we’d remove other filters to get a better grasp on how particular filters changed the data. Moving forward with this analysis, we suggest focusing on the functions of particular genes and how they relate to susceptibility or resistance to this unknown pathogen causing SSWD.