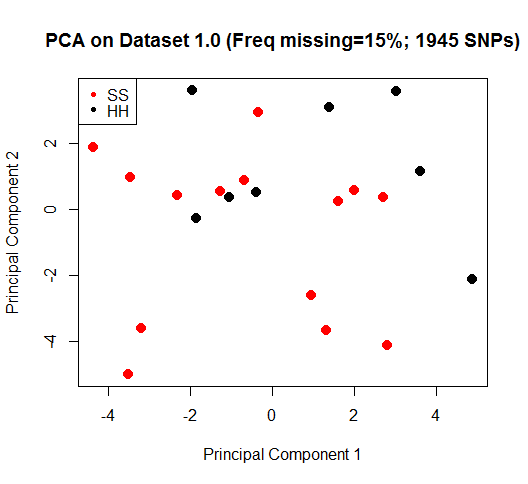
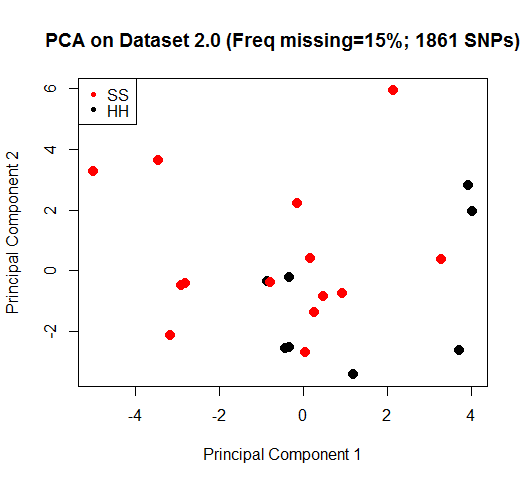
**P/BIO 381 Spring 2017**

**Assignment #3: Population genomic diversity and structure**

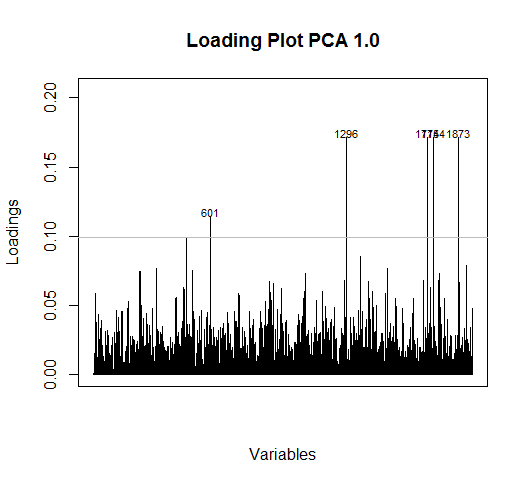
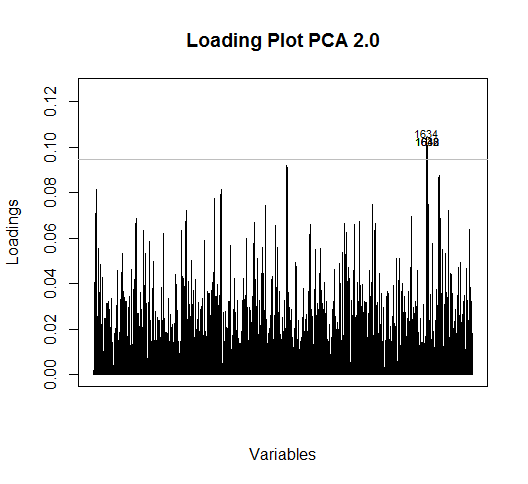
**Introduction**: The goal of this assignment was to demonstrate the impact of different filtering strategies on the interpretation of population genomic structure in our data. The population structure analyzed in this assignment, is the structure found within our sampled *Pisaster Ochraceus* individuals. I used two different filtering strategies on these data, resulting in two different sets of SNP data. I then computed PCAs in R for each set of SNP data1. These analyses facilitated my assessment of the level of population structure in our *Pisaster Ochraceus* samples, and the impact of filtering strategy on this assessment.

**Methods**: We assessed the quality of raw RNA-seq data using fastqc, and cleaned our raw reads using Trimmomatic. After removing adapters and low quality bases from our file, we again evaluated the quality using fastqc. We then assembled a de novo transcriptome using fastq files from four sea stars (both sick and healthy). Clean reads were mapped to our reference transcriptome, producing sequence alignment files. We then utilized the reads2snp method outlined in Gayral et al. (2013), using our SAM file to call SNPs and genotypes for each individual2. The variants called using this method were stored in a VCF file. I used two methods to filter this VCF file. Data set 1.0 was filtered for biallelic loci, a minor allele frequency of 0.04, and a maximum of 15% missing data. Data set 2.0 used the same criteria, while also removing sites that significantly deviated from Hardy-Weinberg Equilibrium. Additionally, I removed the MM individuals (n=2) from both data sets, allowing for the sole comparison of healthy versus sick individuals. I computed PCAs for each dataset, and identified individuals within each PCA plot by their health status (HH, or SS). For each PCA, I identified the SNPs that had loadings on PC1 in the 99.9th percentile.

**Results**: Our original VCF file contained 24 individuals, with 7,486,938 SNPs. My first method of filtering these data (biallelic loci, maf = 0.04, no MM, and a max. missing = 15%) yielded 22 individuals and 1,945 SNPs. My second method (biallelic loci, maf = 0.04, no MM, max. missing = 15%, HWE = 0.05) yielded 22 individuals 1,861 SNPs. PCA plots for data set 1.0 and 2.0, demonstrate a minimal amount of population structure in our samples (Figure 1). There are no clear clusters of individuals in either plot, and individuals do not group by health status. Loading plots for each data set identified five SNPs in data set 1.0, and two SNPs in data set 2.0 that have loadings in the 99.9th percentile on PC1 (Figure 2). Despite minor differences in filtering strategy, there are no highly loaded SNPs in common between dataset 1.0 and 2.0 (Table 1).

A. B. 

**Figure 1**. PCA plots depicting the variation in SNP genotypes for each individual sea star. Individuals are labeled by health status. A) Depicts results from the PCA performed on dataset 1.0. B) Depicts results from the PCA performed on dataset 2.0.

A. B. 

**Figure 2**. Plots depicting Loading for each SNP on PC1. The horizontal line depicts the 99.9th percentile. A) Depicts results from the PCA performed on dataset 1.0. B) Depicts results from the PCA performed on dataset 2.0.

|  |  |
| --- | --- |
| **Data Set** | **SNPs with PC1 Loading in 99.9th Percentile** |
| 1.0 | DN46269\_c0\_g1 |
| 1.0 | DN45155\_c27\_g1 |
| 1.0 | DN39079\_c3\_g1 |
| 1.0 | DN39696\_c4\_g1 |
| 1.0 | DN42225\_c1\_g1 |
| 2.0 | DN46193\_c5\_g2 |
| 2.0 | DN44693\_c5\_g1 |

**Table 1**. List of SNPs from dataset 1.0 and 2.0 with loadings in the 99.9th percentile on PC1.

**Discussion**: The use of different filtering strategies resulted in fine scale differences in my PCA plots, but did not impact my inference of the population structure in our sampled sea stars. Both analyses show no indication of population structure in our data. This outcome is unsurprising, considering the relatively high gene flow among sea star populations and the fact that all sampled individuals were collected from the same isolated region. The impact of filtering strategy is more evident when looking at the loadings of different SNPs. In this case, the additional filter of removing Hardy-Weinberg deviants changed my results for which SNPs load most strongly on the first Principal Component. In future analyses, I could improve my filtration strategy by removing individuals that have a large amount of missing data. Additionally, I could use my two datasets to look at the difference in SNP allele frequencies between healthy and sick individuals. This would demonstrate the sensitivity of my results to different filtering strategies.

**Works Cited**: 1. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2. Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, et al. (2013) Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. PLoS Genet 9(4): e1003457. doi:10.1371/journal.pgen.1003457