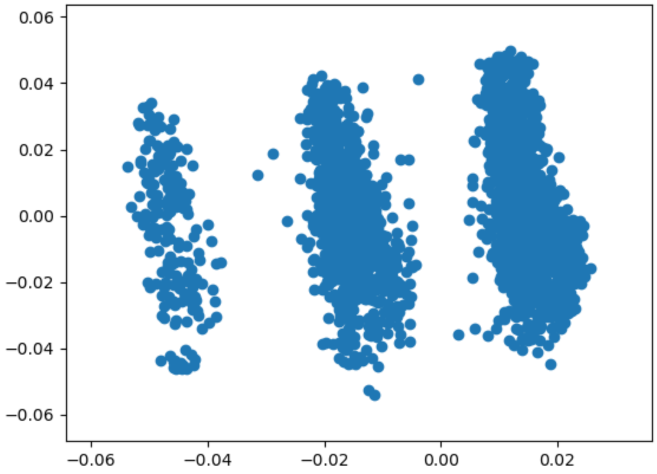
Shonak Shah
DSC 180A

PART 1

In this assignment, my pipeline used the sample Chromosome 22 test file as a basis to determine whether the created code works as intended. Many of the file transformations were done using the PLINK2 software and I connected to the software using Bash files that would run a certain order of commands to obtain a PCA plot in the end. This file was converted from a vcf.gz file into BED/BIM/FAM files that plink could work with more easily. After these files were created, I limited the size of the positions I was looking at through a combination of several filtering methods. First, I made the SNPs only bi-allelic through the --snps-only command and then made the allelic frequency 15% (0.15) so anything that appears less than that does not appear. I chose 0.15 as it was not as stringent as 0.20, but still filters a large amount of the data. The data was filtered again through the --geno and --mind flags and their values were also chosen as they were limiting the total dataset but not too limiting that the results were not useful.

The data was then recoded through the recode flag to make it faster and more efficient for PLINK2's commands. Then, the data was put through PLINK2's PCA flag to perform PCA analysis on the data with the chosen number of principal components as 2. This was done as we are looking at the locations of the subjects so the components would represent the longitude and latitude of the subject's origin and that would help more easily identify the origin of their lineage. The outliers would be removed by determining how far they are away from the center of the cluster and those further away that a certain threshold are removed as they would skew the results. The resulting PCA clumped around 3 clusters as shown in the picture below.


0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.06
−0.04
−0.02
0.00
0.02