**Population Stratification for GWAS**

**Principal Component Analyses & Logistic Regression**

I utilized Principal Component Analyses (PCA) as the method of control for population stratification. PCA is considered the “gold standard” to control for population stratification in GWAS (Hinrichs, Larkin, & Suarez, 2009).

PCA was only utilized in the file in which the stringent quality control was performed. The file with the less stringent quality control had a genomic inflation factor (lambda) = 1. The literature that it is not necessary to control for population stratification if lambda is equal to or below a value of 1 (Hinrichs et al., 2009).

The Plink commands used were:

plink --bfile smokescreen\_stringent\_p2\_s6 --pca --out smokescreen\_stringent\_pca

The eigenvalues generated by the “--pca” function were used to create a scree plot in R:

A screenshot of a cell phone

Description automatically generated

I considered PC2 the “elbow” point in the scree plot and therefore only utilized PC1 & PC2 as covariates in the logistic analyses (Team, 2018). Three logistic analyses were run: 1) PC1 as a covariate, 2) PC2 as a covariate, 3) PC1 and PC2 as covariates.

PCA 1:

plink --bfile smokescreen\_stringent\_p2\_s6 --logistic --covar smokescreen\_stringent\_pca.eigenvec --covar-number 1 --assoc --adjust --out smokescreen\_stringent\_logistic\_pca\_1

Genomic inflation est. lambda (based on median chisq) = 1.00064 (--adjust)

Significant SNP:



PCA 2:

plink --bfile smokescreen\_stringent\_p2\_s6 --logistic --covar smokescreen\_stringent\_pca.eigenvec --covar-number 2 --assoc --adjust --out smokescreen\_stringent\_logistic\_pca\_2

Genomic inflation est. lambda (based on median chisq) = 1 (--adjust).

Significant SNP:



PCA 1 & 2:

The “--covar-number” command on PLINK only allows for 1 selected covariate while “--covar” takes all the PCs in a file. In order to get a logistic analyses a file containing only PC1 and PC2 was created “smokescreen\_stringent\_pca\_first\_two.eigenvec”. This file allows us to only look at PC1 & PC2 instead of PC 1-20 in the original file.

plink --bfile smokescreen\_stringent\_p2\_s6 --logistic --covar smokescreen\_stringent\_pca\_first\_two.eigenvec --assoc --adjust --out smokescreen\_stringent\_logistic\_pca\_1\_2

Genomic inflation est. lambda (based on median chisq) = 1.00345 (--adjust)

Significant SNP:



**Principal Component Ethnicity Graph**

The first 5 principal components were plotted against each other utilizing R. To create this plot I had to cross reference the individuals left after the stringent QC to our subject data bank to find the self-reported ethnicity for each individual. The self-reported ethnicity data was used to create the plots below. This step provides a tool for the visualization of possible clusters based on reported ethnicity. These graphs are used to check for heterogeneity in the population.

**A screenshot of a cell phone

Description automatically generated**

Based on the plot above there doesn’t seem to be any principal components that cluster populations based on ethnicity. This is unusual in most GWAS analyses, but it is consistent with previous data found on Brazilian populations. The Brazilian population is made up of individuals with varying proportions of African, Native American, and European ancestry (Pena et al., 2011). Skin color is poorly correlated with genetic ancestry in the Brazilian population and therefore self-reported race can be inaccurate for genetic studies (Pena et al., 2011).

**Principal Component Case/Control Graph**

**A screenshot of a cell phone

Description automatically generated**

Based on the plot above there doesn’t seem to be any principal components that cluster populations based on case/control status.

**References**

Hinrichs, A. L., Larkin, E. K., & Suarez, B. K. (2009). Population stratification and patterns of linkage disequilibrium. *Genetic Epidemiology*, *33 Suppl 1*, S88-92. https://doi.org/10.1002/gepi.20478

Pena, S. D. J., Di Pietro, G., Fuchshuber-Moraes, M., Genro, J. P., Hutz, M. H., Kehdy, F. de S. G., … Suarez-Kurtz, G. (2011). The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PloS One*, *6*(2), e17063. https://doi.org/10.1371/journal.pone.0017063

Team, B. (2018, September 18). How to read PCA biplots and scree plots. Retrieved October 1, 2019, from Medium website: https://medium.com/@bioturing/how-to-read-pca-biplots-and-scree-plots-186246aae063