Complete Tutorial to Perform GWAS in Windows Operation System using the 1000 Genome Study data.

Student: Faizan Khalid Mohsin; Professor Lei Sun; Course: CHL5224 Statistical Genetics

November 22, 2020

1 Abstract

The objective of this paper is to conduct and report a GWAS as practice. To do this we will perform a GWAS study using a cleaned dataset of the 10000 Genome project. The cleaned dataset has 1755 unrelated/independent individuals and ~2M high-quality SNPs. Since, this dataset has no phenotype data we will randomly assign binary phenotype to the dataset for each individual assigning them either case or control with the same probability. We will then conduct a GWAS without any covariates first using the chi-square test and then using logistic regression Secondly, we will impute continuous phenotype to perform a GWAS using linear regression. Thirdly, we will perform PCA and use the first three principal components as covariates and perform an adjusted logistic regression. Last, we will perform six quality control steps on the already cleaned data set including removing SNPs with low minor allele frequency. This will reduce the number of individuals to 1230. All results from the four association tests - chi-square, logistic regression, linear regression, logistic regression with PCA covariates - will be visualized using Manhattan plots, histograms, and QQ plots. In our results we found that all our association tests showed SNP’s with associations. These are all false positives. However, when adjusted for population structure by using the principal components as covariates the false positive signal from the association test became much weaker. Hence, we can say that since this dataset has people from different regions the population structure acts as a confounder. This highlights the importance of adjusting for covariates and being aware of confounders. To conclude, it is very instructive to perform a GWAS with many important lessons to be learned in terms of statistics, programming, and genetics.

2 Introduction

Background and Motivation:

Genome-wide association studies (GWAS) have become very popular thanks to their rapid success in finding genetic variations associated with particular diseases. GWAS can become even more impactful than they already are as they can help take medication and treatment from a one size fits all approach to personalized medical care and treatment. Hence, for scientist in field of medicine, genetics, and statistics, it can be very advantageous to be able to perform such studies, not only from their own implementation and discovery but also for understanding GWAS studies and the applicability of its results (Marees, 2018).

Objective:

The objective of this paper is to conduct and report a GWAS as practice. We will perform the GWAS using the cleaned 1000 Genome dataset.

GWAS Method:

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

GWAS allows us to skip past the candidate gene approach, because most of the millions of candidates do not turn out to be statistically significant and enable us to say the whole set of genes are your candidates.

3 Methods and Materials

3.1 Overview of the Methods

We will run 4 GWAS using the cleaned 1000 Genome project data with 1755 individuals. First, we will perform a chi-square test of association, followed by a logistic regression, and then linear regression. For all three we will not use any covariates. For the final GWAS, we will perform an adjusted logistic regression with principal components as covariates. Since, the dataset has no phenotypes we will randomly impute both binary and continuous phenotypes. Finally, we will conduct quality control steps on the already cleaned data. Doing this will leave us with 1230 individuals in the dataset.

3.2 Data

The 1000 Genomes Project, an international collaboration, is sequencing the whole genome of approximately 2,000 individuals from different worldwide populations. The central goal of this project is to describe most of the genetic variation that occurs at a population frequency greater than 1%. The results of this project will allow scientists to identify genetic variation at an unprecedented degree of resolution and will also help improve the imputation methods for determining unobserved genetic variants that are not represented on current genotyping arrays. By identifying novel or rare functional genetic variants, researchers will be able to pinpoint disease-causing genes in genomic regions initially identified by association studies. This level of detailed sequence information will also improve our knowledge of the evolutionary processes and the genomic patterns that have shaped the human species as we know it today. The new data will also lay the foundation for future clinical applications, such as prediction of disease susceptibility and drug response (Roslin, 2016).

A cleaned set of 1000 genome project data was downloaded from the website: http://tcag.ca/tools/1000genomes.html (http://tcag.ca/tools/1000genomes.html) where it is available for free.

The cleaned dataset has 1755 unrelated/independent individuals and ~2M high-quality SNPs (1,986,922 SNPs to be exact).

The cleaned data set comes with three bfiles (indep.bed, indep.fam, indep.bid). The .bed file contains the genotyping results of all patients and healthy controls; the .fam file contains the subjectâ□□related data (family relationship with other participants in the study, sex, and clinical diagnosis); and the .bim file contains information on the physical position of the SNPs. Analysis using covariates requires a fourth file, containing the values of these covariates for each individual (Marees, 2018).

The the phenotypic data is missing in our .fam file so we randomly assign binary (cases and controls) and continuous phenotypes, then conduct using these GWAS.

3.3 Statistical Method and Model

Model

A chi-square test, a logistic regression, and a linear regression are used to perform tests of association in GWAS. Since the logistic regression allows us to correct for covariates and is the main model we use in this tutorial we will explain the logistic regression model briefly below.

A logistic regression is used to perform the test of association between SNP’s and binary outcome of phenotype trait of interest (Sun, 2020).

In particular an Additive Model is used to perform the GWAS:

$$logit(phenotype) = SNP_i + covariates \qquad where \quad i = 1\,, \ldots,\ 1,986,922$$

Where:

- phenotype = 1 if control, 2 if case.
- SNP_i = 0 if have both minor alleles AA, 1 if have one minor allele AT, and 2 if do not have any minor alleles (TT).
- covariates for our study will be the principal components. Though, for most of the analysis we will not include any covariates.

Just a note on covariates. In any GWAS study, the gender, the age and the age-squared should be used as covariates as these appear to be significant in most studies and can act as confounders if not included.

Mutliple Testing and P-Value Correction

To determine which p-value should be used for the statistical significance threshold level, various simulations have indicated that the widely used genomeâ□□wide significance threshold of 5 Ã□ 10â□□8 for studies on European populations adequately controls for the number of independent SNPs in the entire genome, regardless of the actual SNP density of the study (Dudbridge & Gusnanto, 2008). When testing African populations, more stringent thresholds are required due to the greater genetic diversity among those individuals (probably close to 1.0 Ã□ 10â□□8 ; Hoggart, Clark, De Lorio, Whittaker, & Balding, 2008). Three widely applied alternatives for determining genomeâ□□wide significance are the use of Bonferroni correction, Benjaminiâ□“Hochberg false discovery rate (FDR), and permutation testing. For simplicity, we will not use any of these alternatives (Marees, 2018).

3.4 Software

All the analysis and data cleaning was performed on a Windows 10 operating system. Further, Plink version 1.9 was used to run the GWAS and quality control steps. The R software version 3.6.6 was used to visualize the results in Manhattan plots, QQ-plots, and histogram.

It was actually very challenging to find out how to run a complete GWAS in the Windows 10 operating system (OS). However, once it was discovered it was straight forward to perform a complete GWAS study in Windows 10 OS from the quality control steps in Plink to running Rscripts on the command line for visualizations.

One of the challenges was that there is no centralized resource which guides one on how to perform a GWAS from start to finish in a Windows 10 OS and we have to piece together from multiple different resources and past programming experience on how to perform a PheWAS to be able to successfully run a GWAS.

3.5 Code for reproducibility: Guide on how to run Genome-Wide-Association-Study in Windows OS.

The difficult and unique thing to run GWAS: running plink and Rscripts from the command line, we need to use both the windows command prompt and we also need to download and use the Git Bash which is a Linux system command prompt software together. This is because Plink runs on the Windows command prompt but cannot run on the Git Bash command prompt, however, Rscript runs on the Git Bash command prompt however, it does not run on the Windows command prompt. Therefore, both must be used concurrently.

The first step to running a GWAS is to create our directory in which we will be working. We call this folder 1000GenomeProj and will refer to it as the working directory. The second step is to download Plink software in our folder. Third is to download the compressed data file from http://tcag.ca/tools/1000genomes.html (http://tcag.ca/tools/1000genomes.html) and unzip the file in the working directory. This will give the bfiles: indep.bed, indep.bid, indep.fam. The three files together are referred to as the bfiles and

Following this we move into the directory in both the windows command prompt and the Git Bash command prompt. Below are our command for each of them respectively.

3.5.1 Raondomly assigning case and control to the missing phenotype with equal probability.

First, we read in the fam file to assign random case control on last column. The -9 indicates missing phenotype

```
# Reading the fam file using the fread command from the data.table package as it is very fast.
indep.fam <- fread("indep_original_fam.fam", na.strings = -9   )
```

We now set seed of 2020 so that our results are reproducible. Further, we will use the sample() function in R to randomly assign 1 or 2 to the last column of the indep.fam file as it indicates the phenotype.

```
# Assign randomly the cases and control to V6 which is the binary phenotype.

set.seed(2020)

case_controls = sample(1:2, size = nrow(indep.fam), replace = T )

indep.fam$V6 = case_controls

write.table(indep.fam, file = "indep.fam", quote = F, row.names = F, col.names = F )
```

We print the first few rows. The last column shows if case or control was assigned. Again, control is 1 and case is 2. We can see even from these few lines there is no distinct pattern for the assignment.

Table 1. The first 10 rows of the indep.fam file with imputed binary case (2) and control (1) phenotype.

FID	IID	PID	MID	Sex	Phenotype
1328	NA06984	0	0	1	2
1328	NA06989	0	0	2	2
1330	NA12340	0	0	1	1
1330	NA12341	0	0	2	2
1330	NA12342	0	0	1	2
1330	NA12343	0	0	2	1
1331	NA07340	0	0	2	1
1334	NA12144	0	0	1	2
1334	NA12145	0	0	2	2
1334	NA12146	0	0	1	2

Note: The phenotype is imputed using the sample() function in R. Also, FID - Family ID, IID - Individual ID, PID - Parental ID, and MID - Maternal ID

We can see from the below table that there were almost equal number of cases and controls randomly assigned for the phenotype.

Table 2. The number of cases (2) and controls (1) for the phenotype.

	control	case
Var1	1	2
Freq	875	881

3.5.2 Running PLINK and Rscripts from the command line in Windows OS.

We will be using Windows Command Prompt and Git Bash Command Prompt to perform a GWAS in Windows OS.

For Windows backlashes are used to specify the path. For Git Bash forward slashes are required. Further, since one of the folder names has a space in the name then the entire folder path needs to be in double quotations.

```
cd OneDrive\OneDrive\Statistical Genetics\HW2\1000GenomeProj  #Run on Windows Command Prompt

cd "OneDrive/OneDrive/Statistical Genetics/HW2/1000GenomeProj"   # Run on Git Bash Command Promp
t
```

3.5.3 Running GWAS using Chi-square Association Test in Plink.

3.5.3.1 PLINK Commands for GWAS Chi-square Association Test

Since the data is already cleaned we can directly perform the logistic regression association tests between the SNP’s and the phenotype. We will perform this in Plink. However, Plink does not run on Git Bash command prompt hence we will run it on Windows command Prompt. To simply perform a simple chi-square test of association between the binary phenotype and SNP’s we can use the “–assoc” command in PLINK as below. Note that –assoc option does not allow to correct for covariates.

```
plink --bfile indep --assoc --out assoc_results # Run on Windows Command Prompt
```

This is all it takes to run a GWAS study and get results if the phenotype is binary and we have no covariates to adjust for.

We get the following output in the command line. You should also get something very similar.

```
PLINK v1.90b6.21 64-bit (19 Oct 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to assoc_results.log.
Options in effect:
  --assoc
  --bfile indep
  --out assoc_results

8113 MB RAM detected; reserving 4056 MB for main workspace.
1989184 variants loaded from .bim file.
1756 people (843 males, 898 females, 15 ambiguous) loaded from .fam.
Ambiguous sex IDs written to assoc_results.nosex .
1756 phenotype values loaded from .fam.
Warning: Ignoring phenotypes of missing-sex samples.  If you don't want those
phenotypes to be ignored, use the --allow-no-sex flag.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1756 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: Nonmissing nonmale Y chromosome genotype(s) present; many commands
treat these as missing.
Total genotyping rate is 0.997622.
1989184 variants and 1756 people pass filters and QC.
Among remaining phenotypes, 876 are cases and 865 are controls.  (15 phenotypes
are missing.)
Writing C/C --assoc report to assoc_results.assoc ... done.
```

3.5.3.2 Visualizing GWAS Results Using R.

Now to visualize the results we will use the following R code to create the Manhattan plot, the QQ-plot, and the histogram.

```
# Chi-square test

# We use the qqman package to create the manhattan and Q-Q plots. We use base R to create the hi
storgram plot.
library(qqman)
library(data.table)

# Finding the number of NA's in the p-values. Found 105 NA's.
# sum(is.na(results_as$P))

# R code to read in the data from the plink output into R. Use na.omit() to remove any NA's.
results_as <- na.omit(fread("assoc_results.assoc", head=TRUE))

# Create and save the Manhattan-plot.
jpeg("Manhattan_assoc.jpeg")
manhattan(results_as, ylim = c(0, 8.5), col = c("blue4", "orange3"),
    chr="CHR",bp="BP",p="P",snp="SNP", main = "Manhattan plot: assoc"
    , annotatePval = 0.000001
    )
dev.off()

# Create and save the QQ-plot.
jpeg("QQ-Plot_assoc.jpeg")
qq(results_as$P, main = "Q-Q plot of GWAS p-values : log")
dev.off()

# Create and save the Historgram-plot
jpeg("Histogram_assoc.jpeg")
hist(results_as$P, main = "Histogram plot of GWAS p-values : log")
dev.off()
```

3.5.4 Running GWAS using Logistic Regression Association Test in Plink.

3.5.4.1 PLINK Commands

We will now perform a test of association using logistic regression instead of chi-square test. Further, even though logistic regression can include covariates we will do that in a later analysis.

We will use the –logistic command in PLINK to run a logistic regression using the code below.

```
# logistic regression association test without any covariates.
plink --bfile indep --logistic --out logistic_results # Run on Windows Command Prompt
```

We will now remove NA values, those might give problems generating plots in later steps.

```
# Removing NA values.
awk '!/'NA'/' logistic_results.assoc.logistic > logistic_results.assoc_2.logistic # Run on Git B
ash Command Prompt
```

We will now visualize our results using the Manhattan plot, QQ-plot, and the histogram of the p-values.

3.5.4.2 Visualizing the GWAS Results.

```
# Logistic Regression

# We use the qqman package to create the manhattan and Q-Q plots. We use base R to create the historgram plot.
library(qqman)
library(data.table)
# R code to read in the data from the plink output into R.
results_logistic <- fread("logistic_results.assoc_2.logistic", head=TRUE)

# Create and save the Manhattan-plot.
jpeg("Manhattan_logistic.jpeg")
manhattan(results_logistic, ylim = c(0, 8.5), col = c("blue4", "orange3"),
    chr="CHR",bp="BP",p="P",snp="SNP", main = "Manhattan plot: logistic"
    , annotatePval = 0.000001
    )
dev.off()

# Create and save the QQ-plot.
jpeg("QQ-Plot_logistic.jpeg")
qq(results_logistic$P, main = "Q-Q plot of GWAS p-values : log")
dev.off()

# Create and save the Historgram-plot.
jpeg("Histogram_logistic.jpeg")
hist(results_logistic$P, main = "Histogram plot of GWAS p-values : log")
dev.off()
```

Again, this is all it takes to run a GWAS study and get results if we want to run a logistic regression test of association without covariates in in PLINK. Further, we can show that squared of the logistic regression statistic will give us the chi-square statistic. Hence, we expect very similar results from both association tests.

3.5.5 Running GWAS using Linear Regression Association Test for Continuous Phynotype in PLINK.

To run a GWAS for a continuous phenotype we will be using the –linear command in PLINK.

First, however, we have to impute the phenotype again but this time with a continuous phenotype. We will again being doing this randomly hence do not expect to find any significant results.

3.5.5.1 Creating Continuous Random Phenotype in R.

We create randomly sampled continuous phenotype by randomly sampling from the normal distribution with mean 0 and standard deviation 1. Below is the code to perform this in R.

```
# Read in the fam file to assign random case control on last column.
indep_linear.fam <- fread("indep_linear.fam", na.strings = "-9"   )

# Summary of all the vairiables.
str(indep_linear.fam )

# See if have anything but "NA". No we do not.
unique(indep_linear.fam$V6 )

# See how many NA's
sum(is.na(indep_linear.fam$V6))
```

```
# Now sample randomly form N(0,1) to create continuous phenotype.

set.seed(2020)

case_controls_linear = rnorm( n = nrow(indep_linear.fam ), mean = 0, sd = 1 )

indep_linear.fam$V6 = case_controls_linear

write.table(indep_linear.fam, file = "indep_linear.fam", quote = F, row.names = F, col.names = F
)
```

3.5.5.2 PLINK Commands

We will use the –linear command in Plink to run a linear regression.

```
# logistic regression association test without any covariates.
plink --bfile indep_linear --linear --out linear_results # Run on Windows Command Prompt
```

Following this, remove any NA values created.

```
# Remove NA values, those might give problems generating plots in later steps.
awk '!/'NA'/' linear_results.assoc.linear > linear_results.assoc_2.linear # Run on Git Bash Comm
and Prompt
```

This completes the linear regression association GWAS.

3.5.5.3 Visualizing the GWAS Results in R.

```
# Linear Regression

# We use the qqman package to create the manhattan and Q-Q plots. We use base R to create the historgram plot.
library(qqman)
library(data.table)
# R code to read in the data from the plink output into R.
results_linear <- fread("linear_results.assoc_2.linear", head=TRUE))

# Create and save the Manhattan-plot.
jpeg("Manhattan_linear.jpeg")
manhattan(results_linear, ylim = c(0, 8.5), col = c("blue4", "orange3"),
    chr="CHR",bp="BP",p="P",snp="SNP", main = "Manhattan plot: linear"
    , annotatePval = 0.000001
    )
dev.off()

# Create and save the QQ-plot.
jpeg("QQ-Plot_linear.jpeg")
qq(results_linear$P, main = "Q-Q plot of linear regression GWAS p-values : log")
dev.off()

# Create and save the Historgram-plot
jpeg("Histogram_linear.jpeg")
hist(results_linear$P, main = "Histogram plot of linear regression GWAS p-values : log")
dev.off()
```

3.5.6 Running GWAS using Logistic Regression Association Test with Principal Componants as Covariates in PLINK.

3.5.6.1 PLINK Commands

The following command performs PCA in PLINK calculating the first three principal components. The output of the command creates to files: indep.eigenval and indep.eigenvectors. The file indep.eigenvectors is used as the covariate file when performing a logistic regression adjusted for covariates.

3.5.6.1.1 Logistic regression with PCA components as covariates.

The below code runs logistic regression in PLINK with the principal components as covariates. The –covar options is used to indicate the covariate file to be used and the –logistic command tells PLINK to perform a logistic regression. The –hide-covar hides the p-values and output of the covariates.

The R code to visualize this is saved in the R script logistic_regression_pca_plots.R and can be run from the Git Bash Command Prompt using the following code.

With this we are done with performing a GWAS. However, majority of the work when conducting a GWAS is performing the quality control steps and data cleaning. We will reproduce the quality control steps performed in the PLINK tutorial which can be found at http://zzz.bwh.harvard.edu/plink/tutorial.shtml (http://zzz.bwh.harvard.edu/plink/tutorial.shtml).

3.6 Quality Control steps.

We reproduce the six quality control steps performed in the tutorial http://zzz.bwh.harvard.edu/plink/tutorial.shtml (http://zzz.bwh.harvard.edu/plink/tutorial.shtml) in detail with reproducible PLINK and R code in the Appendix Section. Here we will summarize the main QC steps (Marees, 2018).

The six main quality control steps are Step 1: Check Missingness of Genotypic Data and keeping only high quality SNP’s and individuals with little missing data, Step 2: Check for Sex Discrepancy and either remove them or impute using genotypic data, Step 3: Check Minor Allele Frequency and remove SNP’s with very low frequency, Step 4: Check Hardy-Weinberg Equilibrium for SNPâ□□s and remove the SNP’s that have p-value smaller than 0.000001, Step 5: Check Heterozygosity Rate and remove the non-compliant ones, and Step 6: Check Relatedness of individuals in the dataset and remove those that are closely related, typically $\phi > 0.2$ which is calculated using PLINK. We will present the QC results at the end of the results section.

Finally, this is not a quality control step, however, we also perform PCA in PLINK for population structure to use as covariates to adjust for people of different ethnicity, as people of different ethnicity have fundamentally different genomic structure.

4 Results

4.1 Chi-Square Association Test - PLINK COMMAND: assoc

4.1.1 Top 10 SNP associations with the Phenotype using chi-square test

Table 3. Top 10 SNP associations with the Phenotype using chi-square test

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
2	SNP2-209944191	210235946	C	0.28290	0.20710	T	26.97	2.0e-07	1.5110
6	SNP6-138762585	138720892	C	0.03600	0.07474	T	24.99	6.0e-07	0.4623
6	rs6902603	96384515	C	0.37800	0.29800	A	24.75	7.0e-07	1.4320
14	SNP14-58775223	59705470	C	0.41770	0.50120	T	24.38	8.0e-07	0.7141
3	rs2724697	137798155	C	0.34290	0.42290	T	23.53	1.2e-06	0.7120
17	SNP17-9554183	9613458	A	0.04525	0.08595	G	23.48	1.3e-06	0.5040
6	SNP6-96504996	96398275	G	0.36910	0.29200	T	23.36	1.3e-06	1.4190
6	rs10484741	96409424	T	0.38580	0.30890	C	22.54	2.1e-06	1.4050
11	SNP11-45190049	45233473	A	0.51940	0.43930	G	22.38	2.2e-06	1.3790
2	SNP2-11909316	11991865	G	0.01086	0.03472	A	22.31	2.3e-06	0.3051

4.1.2 Manahattan Plot

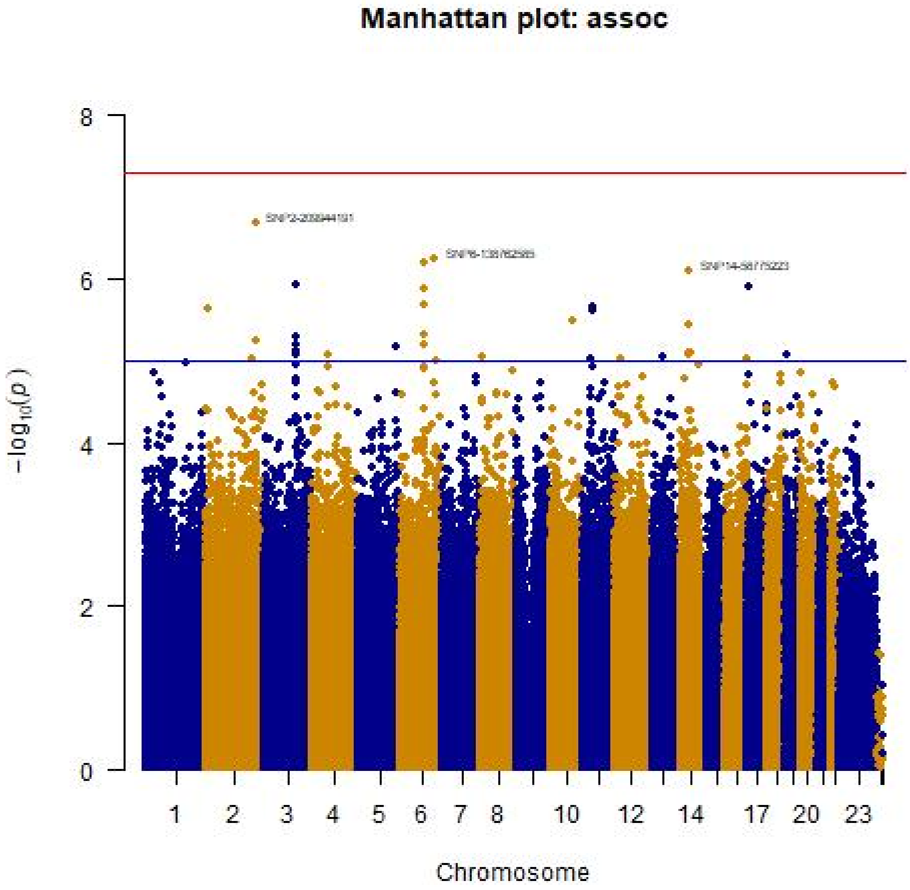


Figure 1. GWAS chi-square association for GWAS without any covariates. The red line is significance level of $5x10^{(} - 8)$. The blue line indicates significance of level of $1x10^{(} - 5)$.

From the Manhattan plot above in Figure 1 we can see that none of the SNPs reach the significance level of $5x10^{(} - 8)$ as they are all below the red line. However, we still see that the SNPs still show some signal of association with the phenotype as they are over the significance level of $1x10^{(} - 5)$ which is indicated by the blue line.

4.1.3 QQ-Plot

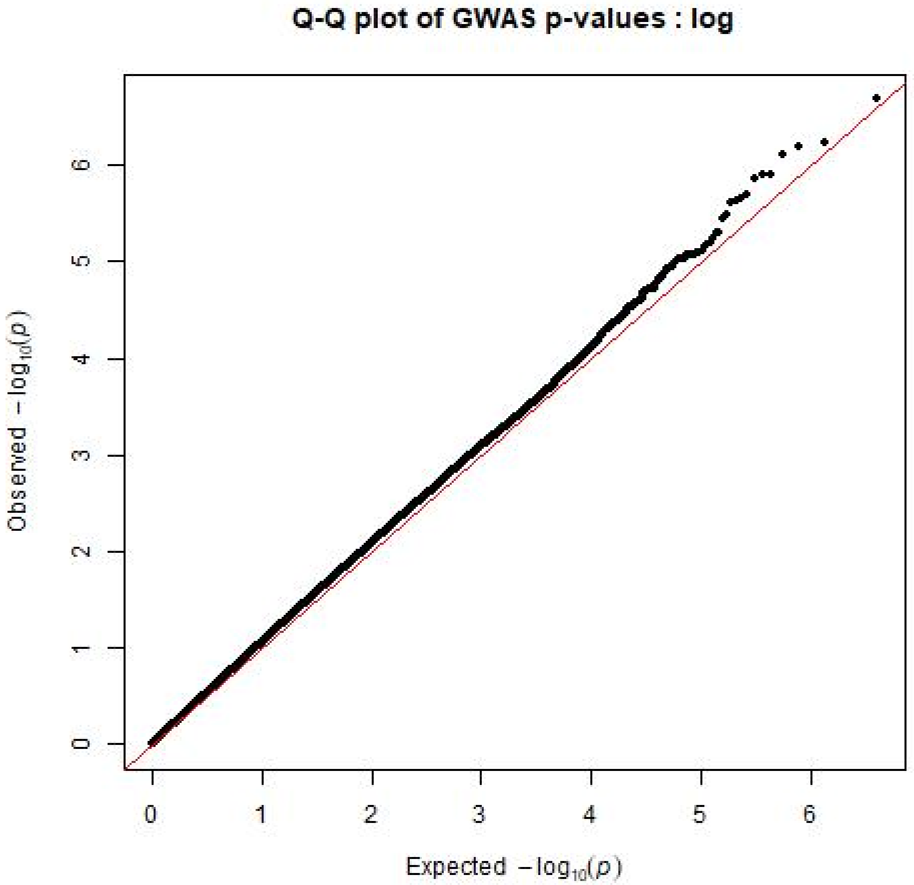


Figure 2. QQ-PLot of GWAS chi-square association without any covariates.

4.1.4 Historgram

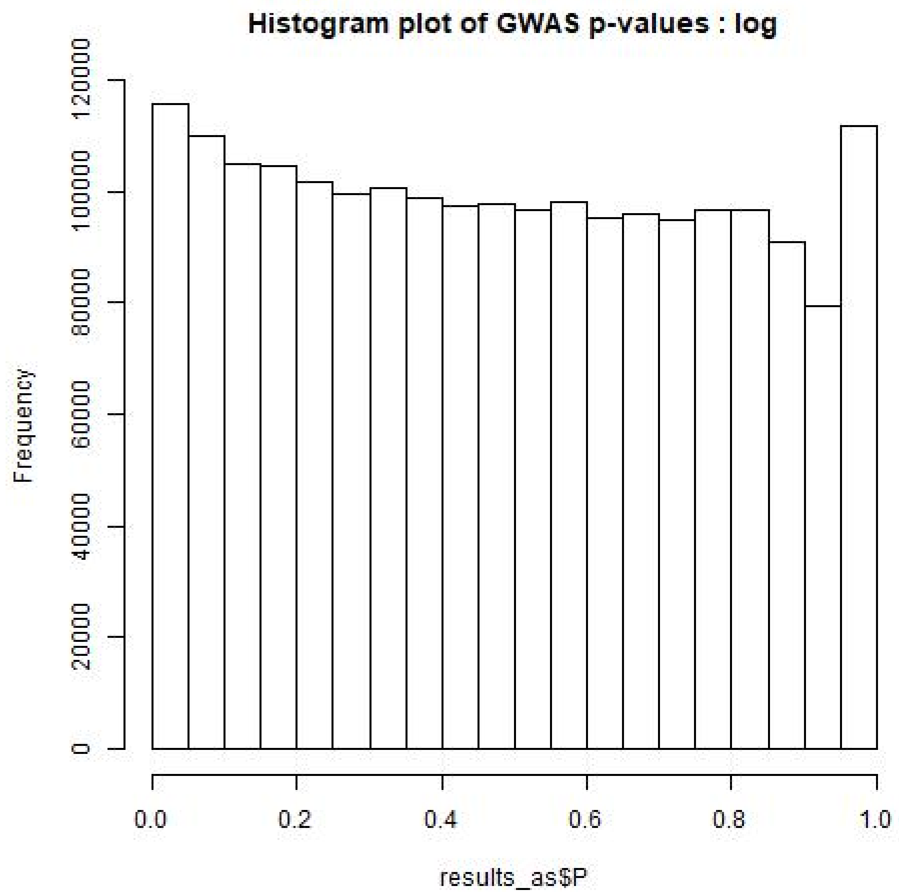


Figure 3. Histogram-Plot of GWAS chi-square association without any covariates.

4.2 Logistic Regression Association Test - PLINK COMMAND: logistic

4.2.1 Top 10 SNP Associated with the Phenotype using Logistic Regression.

Table 4. Top 10 SNP associated with the phenotype using logistic association regression.

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
2	SNP2-209944191	210235946	C	ADD	1735	1.4780	5.000	6.0e-07
14	SNP14-58775223	59705470	C	ADD	1739	0.7203	-4.836	1.3e-06
6	rs6902603	96384515	C	ADD	1732	1.4050	4.805	1.5e-06
3	rs2724697	137798155	C	ADD	1735	0.7256	-4.679	2.9e-06
6	SNP6-96504996	96398275	G	ADD	1738	1.3840	4.624	3.8e-06
6	rs10484741	96409424	T	ADD	1732	1.3760	4.568	4.9e-06
17	SNP17-9554183	9613458	A	ADD	1734	0.5266	-4.565	5.0e-06
14	SNP14-58781428	59711675	T	ADD	1741	0.7340	-4.545	5.5e-06
6	SNP6-138762585	138720892	C	ADD	1738	0.5120	-4.487	7.2e-06
2	SNP2-11909316	11991865	G	ADD	1739	0.3042	-4.464	8.1e-06

4.2.2 Manahattan Plot

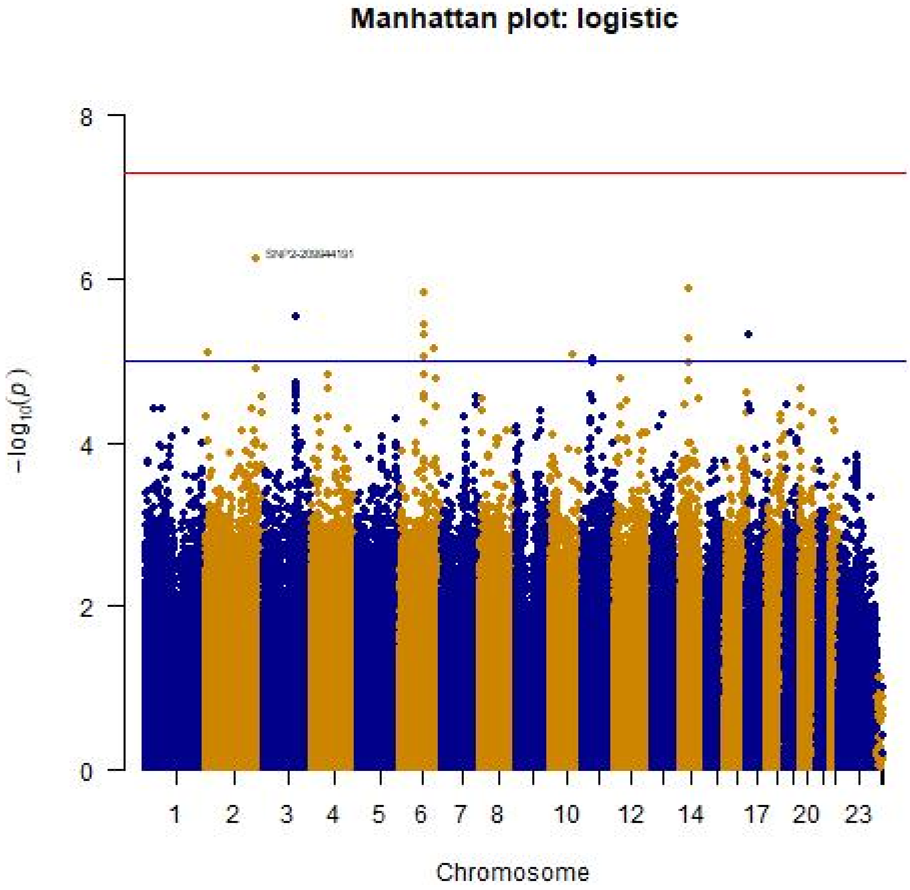


Figure 4. GWAS logistic regression association without any covariates. The red line indicates significance level of $5x10^(-8)$. The blue line indicates significance of level of $1x10^(-5)$

From the Manhattan plot above in Figure 1 we can see that even though

4.2.3 QQ-Plot

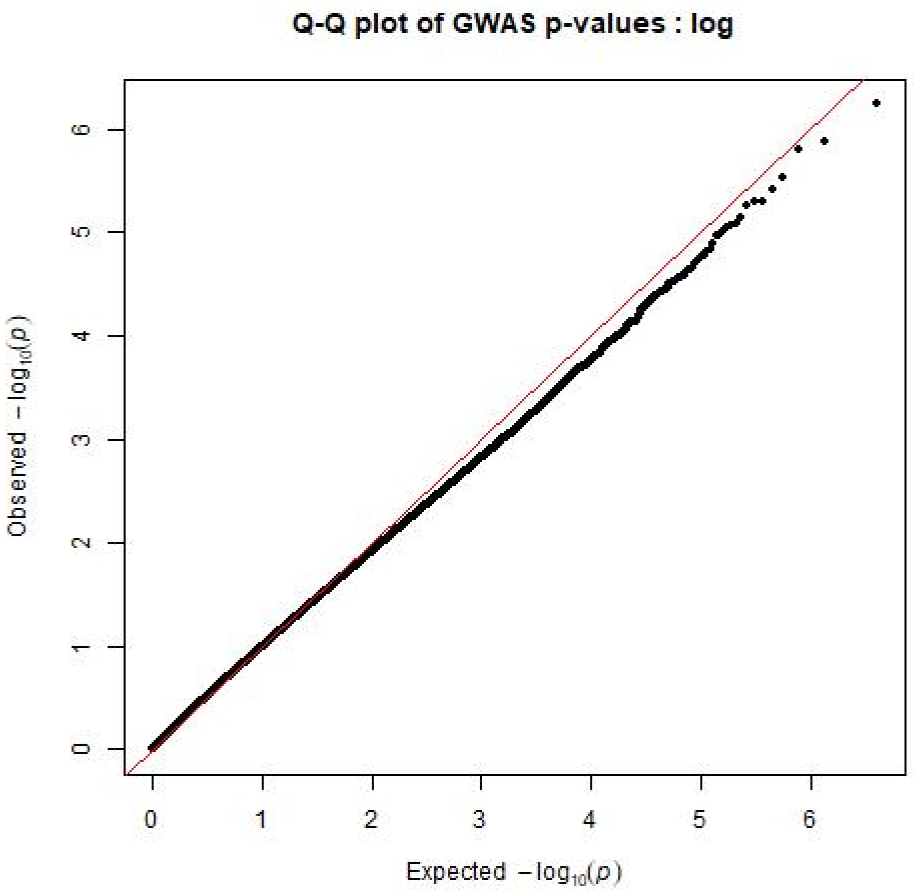


Figure 5. QQ-Plot of GWAS logistic regression association without any covariate.

4.2.4 Historgram

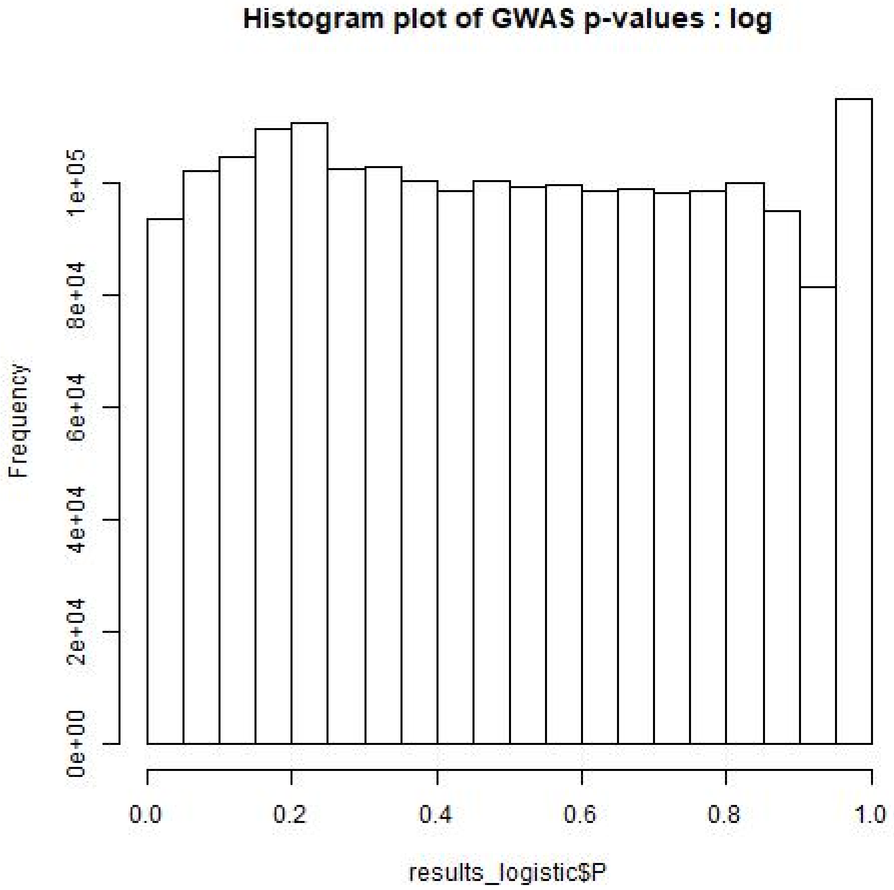


Figure 6. Histogram-Plot of GWAS logistic regression association without any covariates.

4.3 Linear Regression Association Test - PLINK COMMAND: linear

4.3.1 Top 10 SNP Associated with the Phenotype using Linear Regression.

Table 5. Top 10 SNP associated with the phenotype using linear association regression.

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
2	SNP2-209944191	210235946	C	ADD	1735	1.4780	5.000	6.0e-07
14	SNP14-58775223	59705470	C	ADD	1739	0.7203	-4.836	1.3e-06
6	rs6902603	96384515	C	ADD	1732	1.4050	4.805	1.5e-06
3	rs2724697	137798155	C	ADD	1735	0.7256	-4.679	2.9e-06
6	SNP6-96504996	96398275	G	ADD	1738	1.3840	4.624	3.8e-06
6	rs10484741	96409424	T	ADD	1732	1.3760	4.568	4.9e-06
17	SNP17-9554183	9613458	A	ADD	1734	0.5266	-4.565	5.0e-06
14	SNP14-58781428	59711675	T	ADD	1741	0.7340	-4.545	5.5e-06
6	SNP6-138762585	138720892	C	ADD	1738	0.5120	-4.487	7.2e-06
2	SNP2-11909316	11991865	G	ADD	1739	0.3042	-4.464	8.1e-06

4.3.2 Manahattan Plot

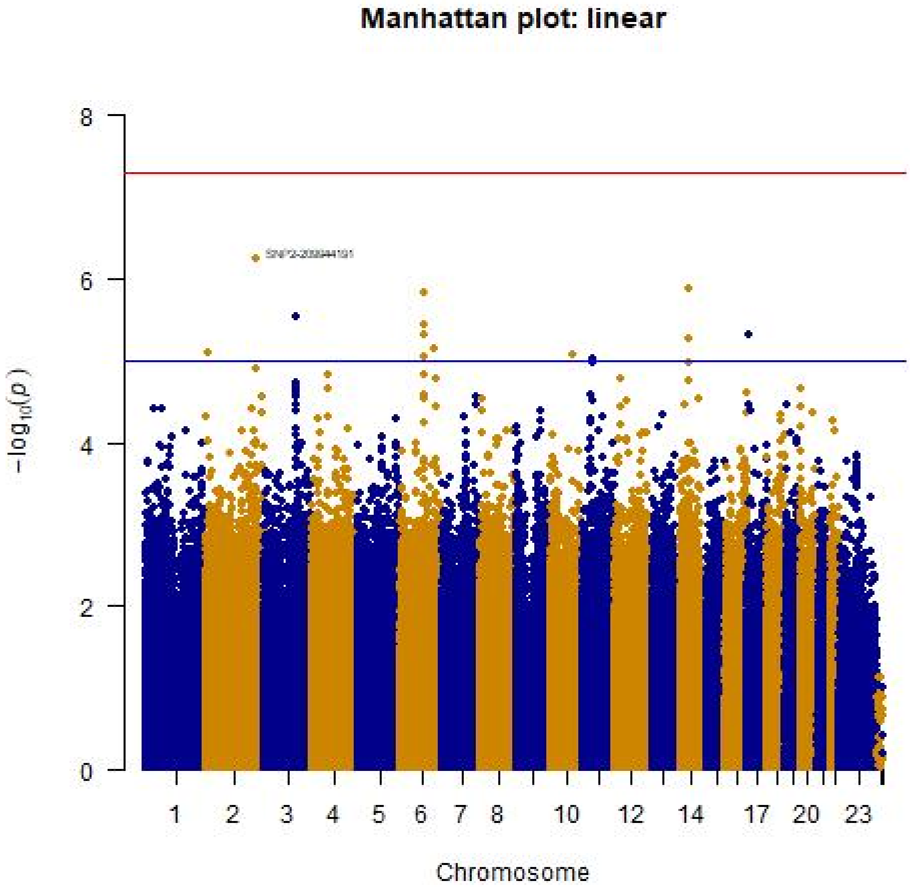


Figure 7. GWAS linear regression association without any covariates. The red line indicates significance level of $5x10^{(}-8)$. The blue line indicates significance of level of $1x10^{(}-5)$

From the Manhattan plot above in Figure 1 we can see that even though

4.3.3 QQ-Plot

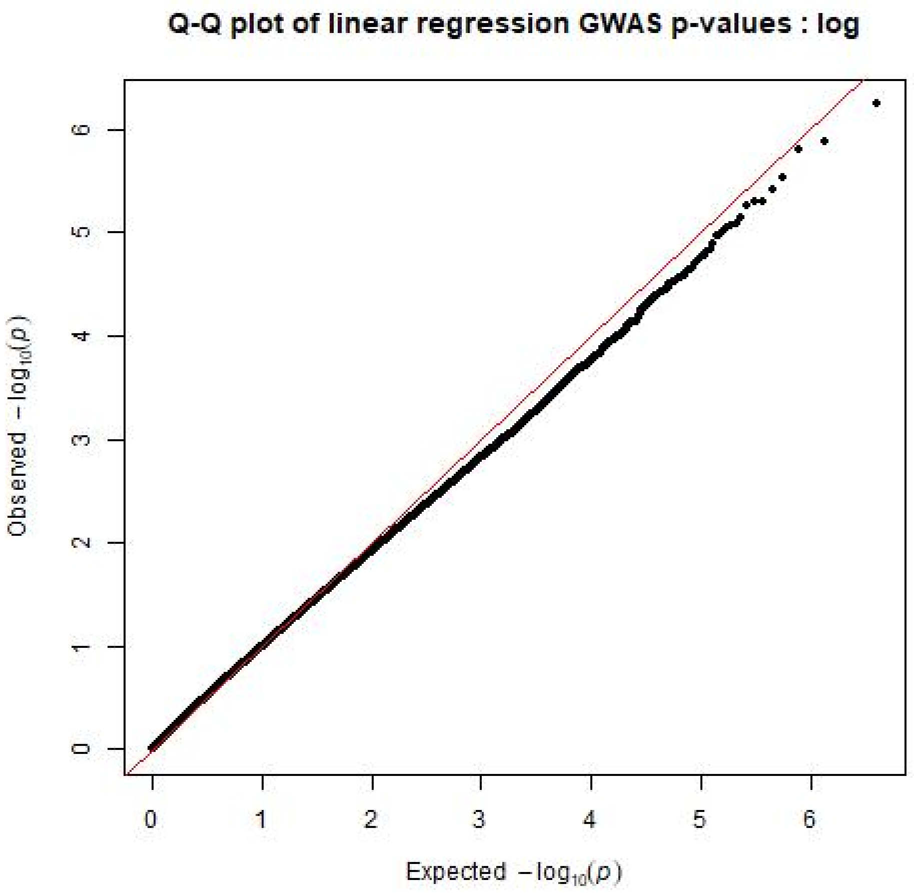


Figure 8. QQ-Plot of GWAS linear regression association without any covariates.

4.3.4 Historgram

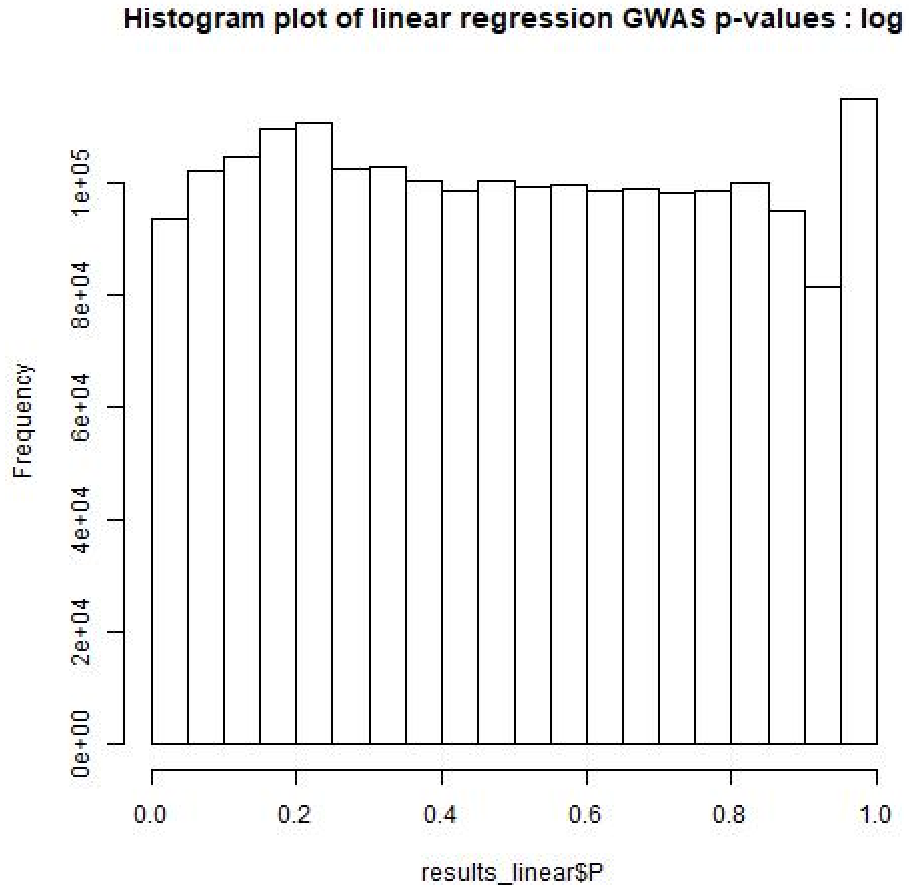


Figure 9. Histogram-Plot of GWAS linear regression association without any covariates.

4.4 Logistic Regression Association Test with PCA covariates - PLINK COMMAND: linear & pca

4.4.1 Top 10 SNP Associated with the Phenotype using logistic Regression with PCA covariates.

Table 6. Top 10 SNP associated with the phenotype using logistic association regression with first 3 PCA componants as covariates.

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
2	SNP2-209944191	210235946	C	ADD	1735	1.4780	5.000	6.0e-07
14	SNP14-58775223	59705470	C	ADD	1739	0.7203	-4.836	1.3e-06
6	rs6902603	96384515	C	ADD	1732	1.4050	4.805	1.5e-06
3	rs2724697	137798155	C	ADD	1735	0.7256	-4.679	2.9e-06
6	SNP6-96504996	96398275	G	ADD	1738	1.3840	4.624	3.8e-06
6	rs10484741	96409424	T	ADD	1732	1.3760	4.568	4.9e-06
17	SNP17-9554183	9613458	A	ADD	1734	0.5266	-4.565	5.0e-06
14	SNP14-58781428	59711675	T	ADD	1741	0.7340	-4.545	5.5e-06
6	SNP6-138762585	138720892	C	ADD	1738	0.5120	-4.487	7.2e-06
2	SNP2-11909316	11991865	G	ADD	1739	0.3042	-4.464	8.1e-06

4.4.2 Manahattan Plot

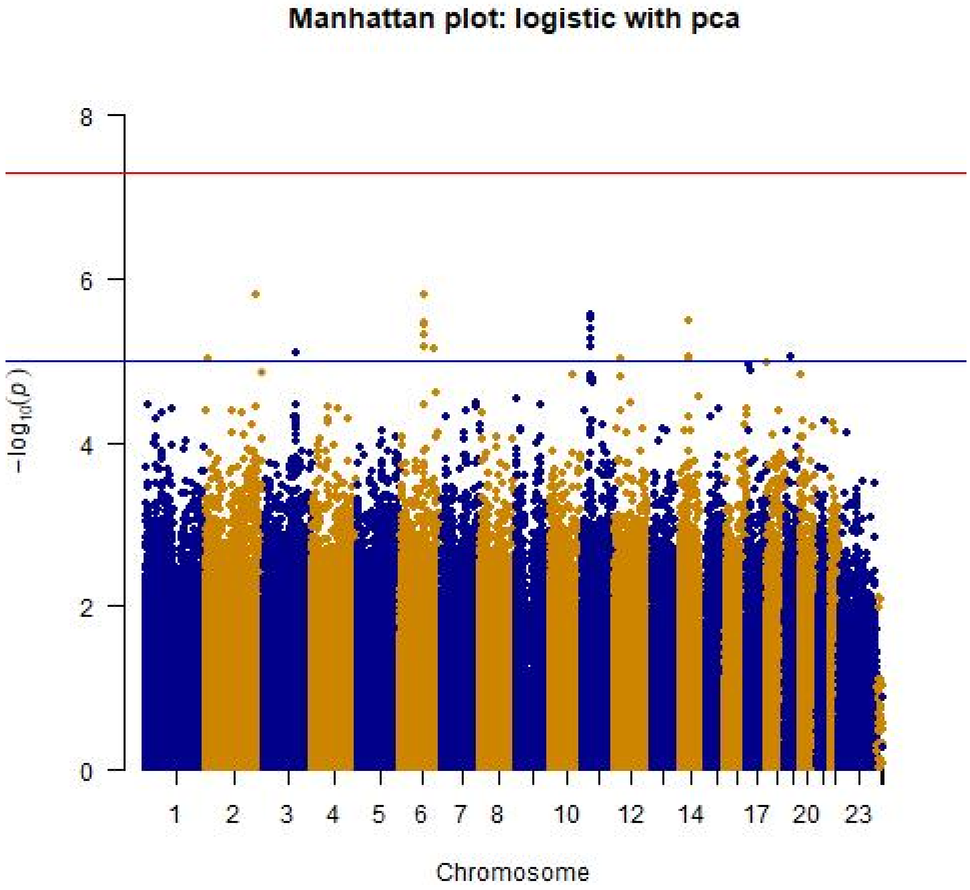


Figure 10. GWAS logistic regression association with first 3 PCA componants as covariates. The red line indicates significance level of $5x10^(- 8)$. The blue line indicates significance of level of $1x10^(- 5)$

From the Manhattan plot above in Figure 1 we can see that even though

4.4.3 QQ-Plot

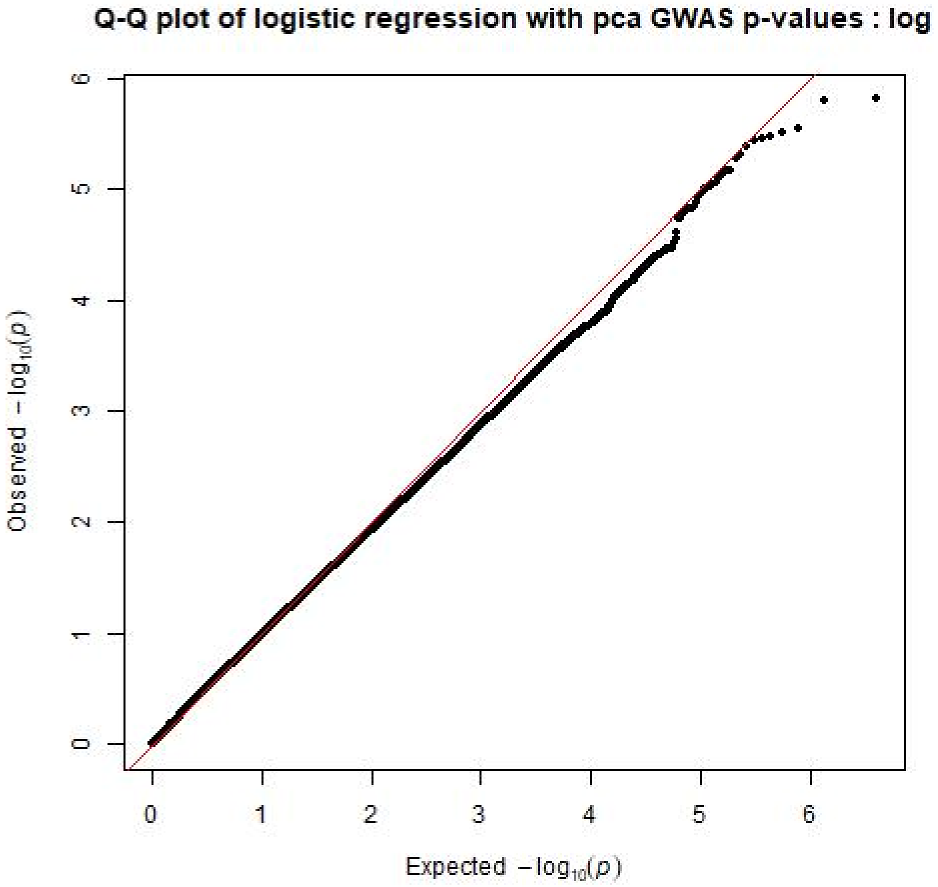


Figure 11. QQ-Plot of GWAS logistic regression association with first 3 PCA componants as covariates

4.4.4 Historgram

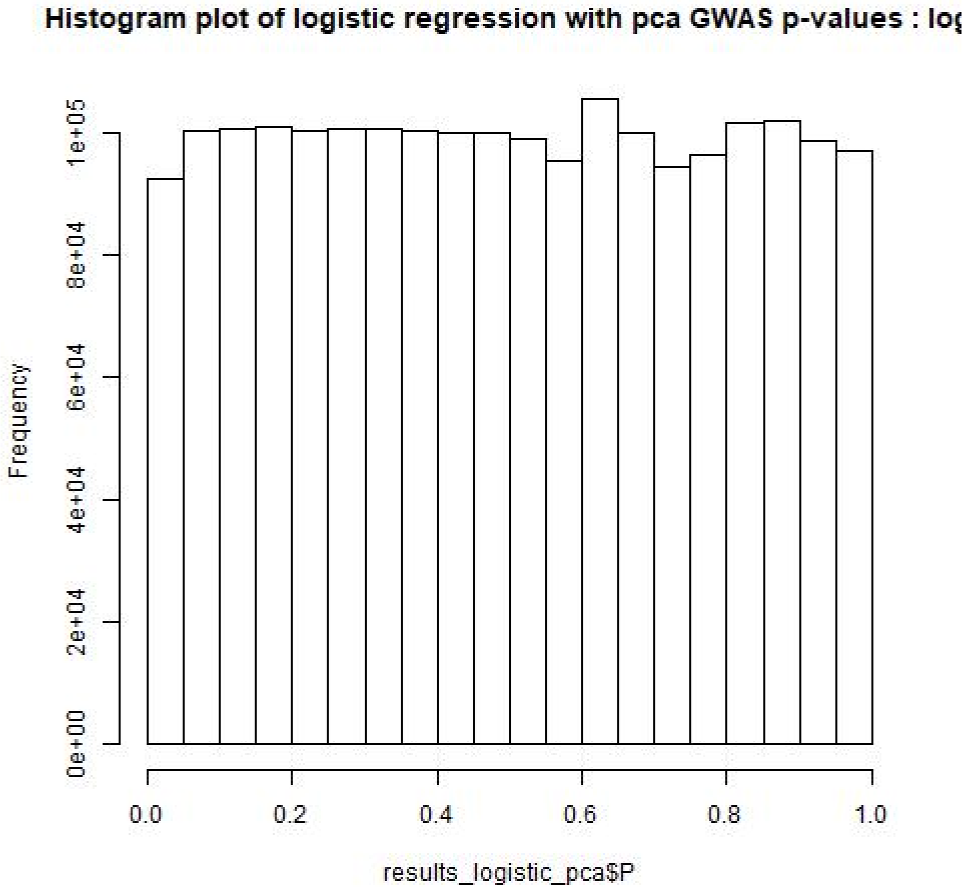
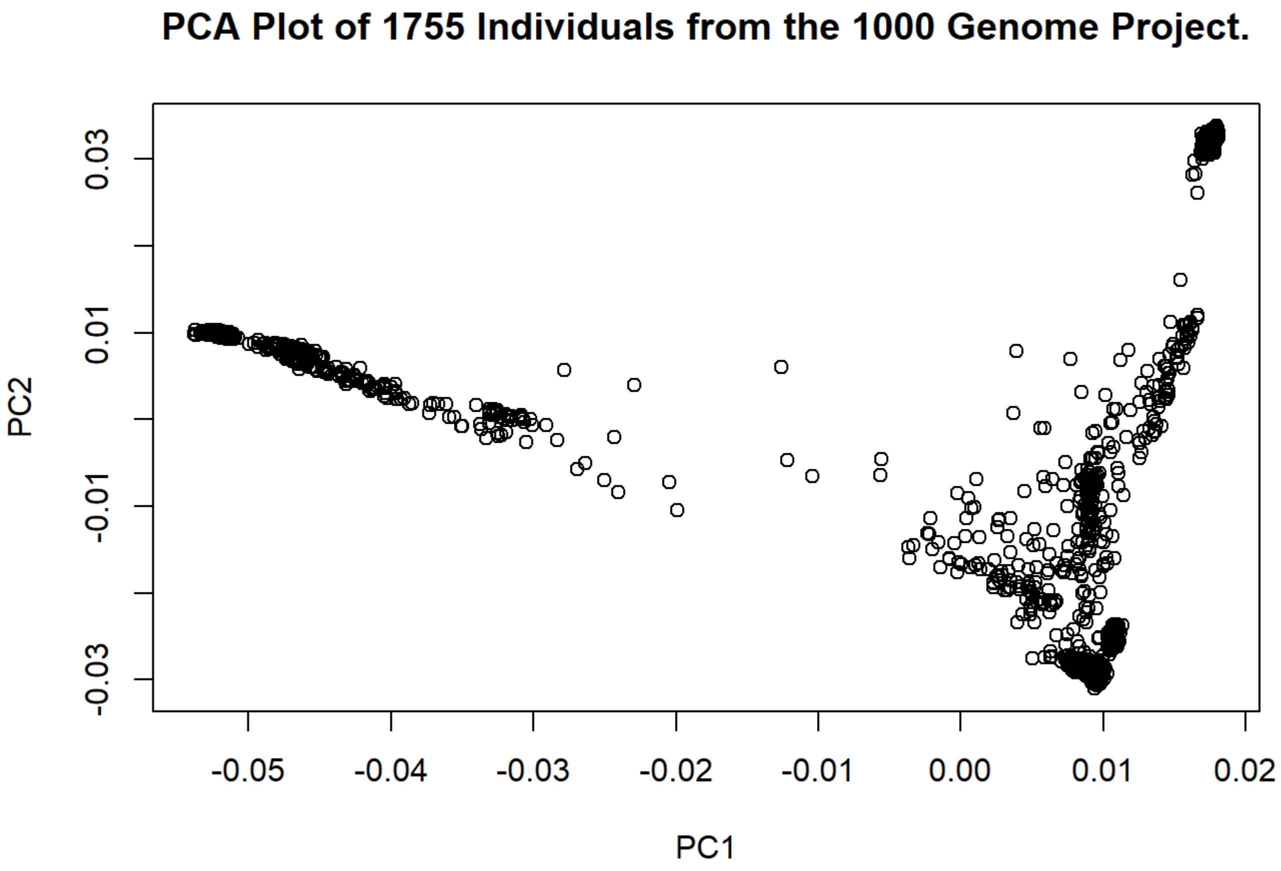


Figure 12. Histogram-Plot of GWAS logistic regression association with first 3 PCA componants as covariates

4.4.5 PCA



4.5 Quality Control Results

4.5.1 Missing Data Check.

The below histograms show respectively the proportion of missing SNPs per individual and the proportion of missing individuals per SNP. Using this we delete the SNPs and individuals with high levels of missingness.

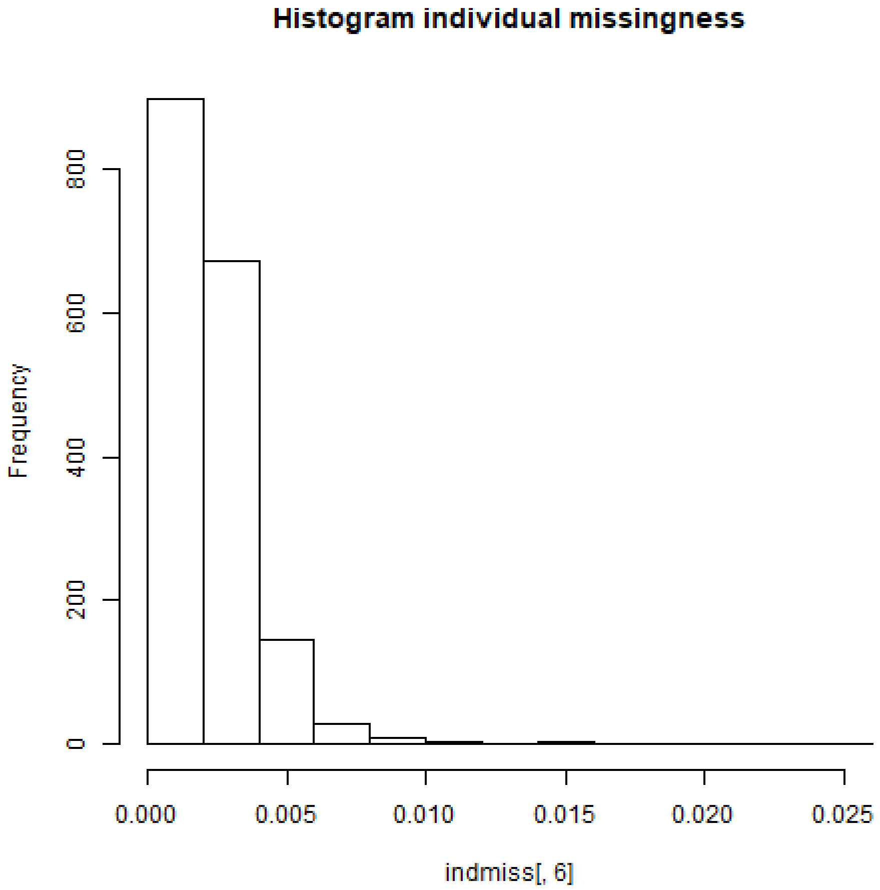


Figure 13. Histogram of missingness for individuals .

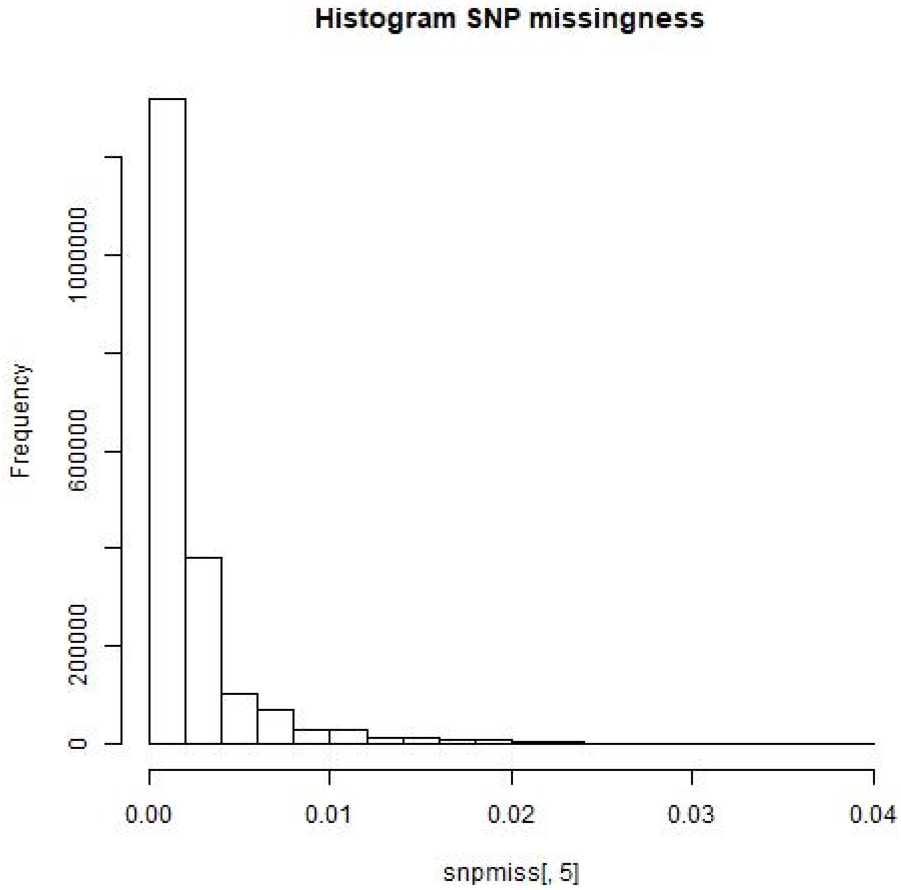


Figure 14. Histogram of missingness for SNP's for the 1755 individuals.

4.5.2 Minor Allele Frequency.

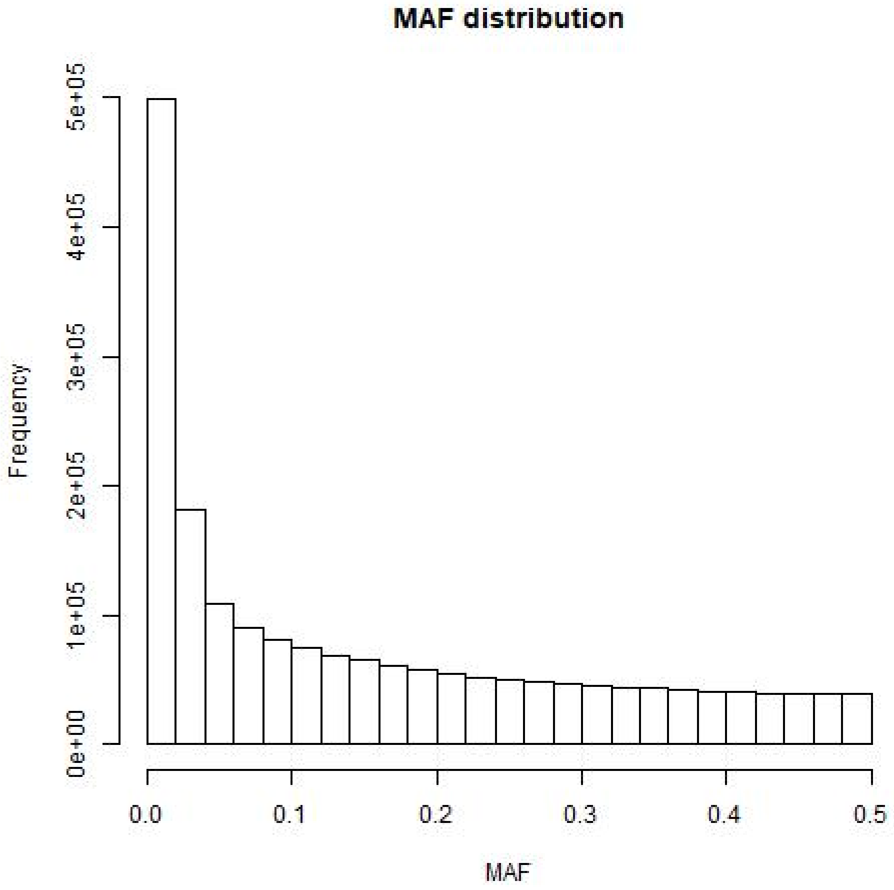


Figure 15. Minor allele frequency distribution for 1755 individuals.

4.5.3 Sex Discrepancy Check.

Subjects who were a priori determined as females must have a F value of <0.2, and subjects who were a priori determined as males must have a F value >0.8. This F value is based on the X chromosome inbreeding (homozygosity) estimate. Subjects who do not fulfill these requirements are flagged “PROBLEM” by PLINK and then removed from the dataset.

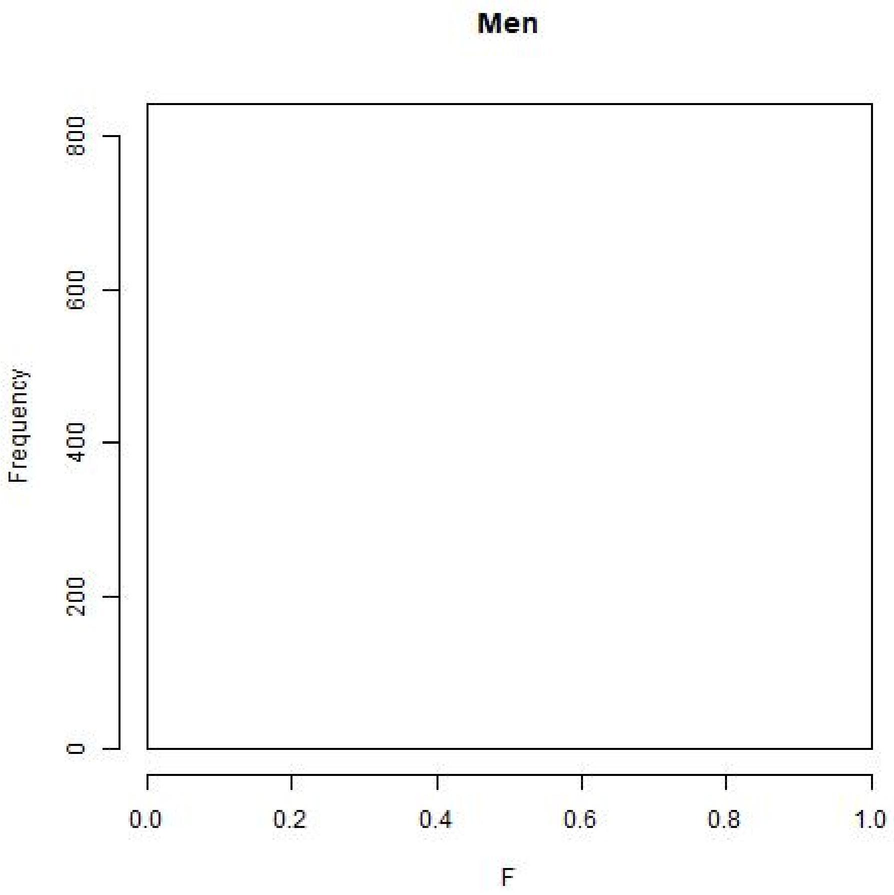


Figure 16. Histogram of the probability of being female based on the sex check for males.

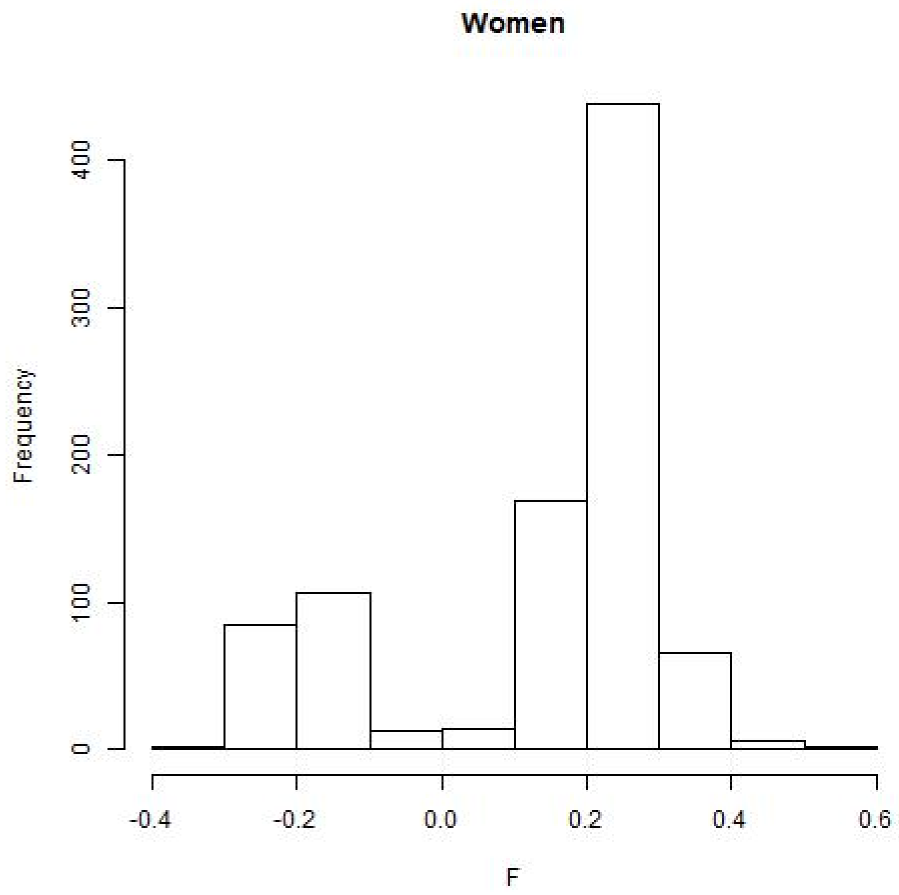


Figure 17. Histogram of the probability of being female based on the sex check for females.

We simply delete the individuals with a sex discrepancy. However, as an alternative we could have also impute the sex using the X-chromosome in PLINK instead of deleting them.

4.5.4 Hardy-Weinberg Equilibrium Check for SNPâ□□s.

We will check the distribution of HWE p-values of all SNPs and delete the SNPs which are not in Hardy-Weinberg equilibrium (HWE).

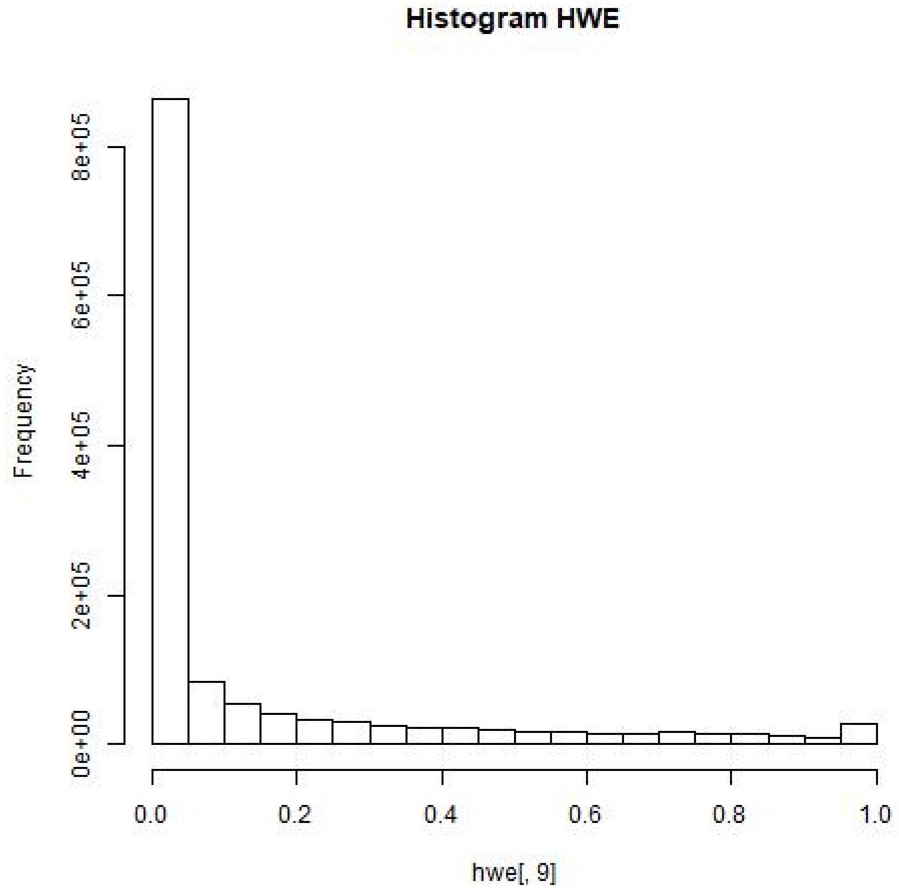


Figure 18. Distribution of HWE p-values of all SNPs.

Selecting SNPs with HWE p-value below 0.00001, required for one of the two plot generated by the next Rscript, allows to zoom in on strongly deviating SNPs.

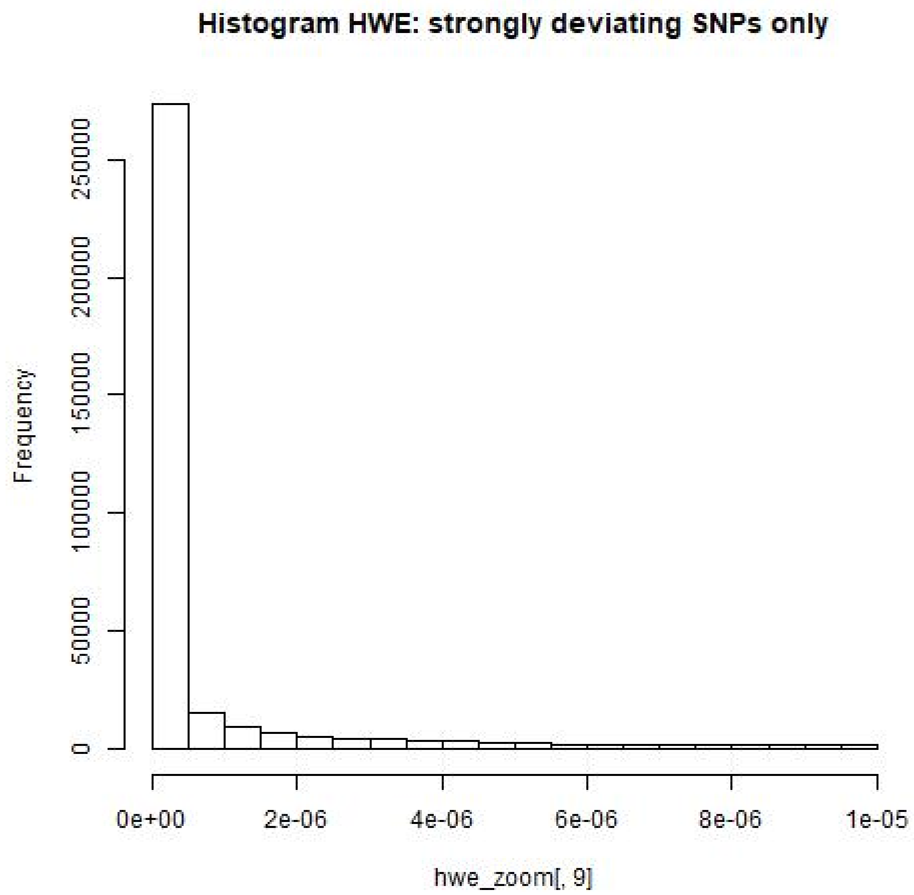


Figure 19. Distribution of the SNPs with HWE p-value below 0.00001.

4.5.5 Heterozygosity Rate Check.

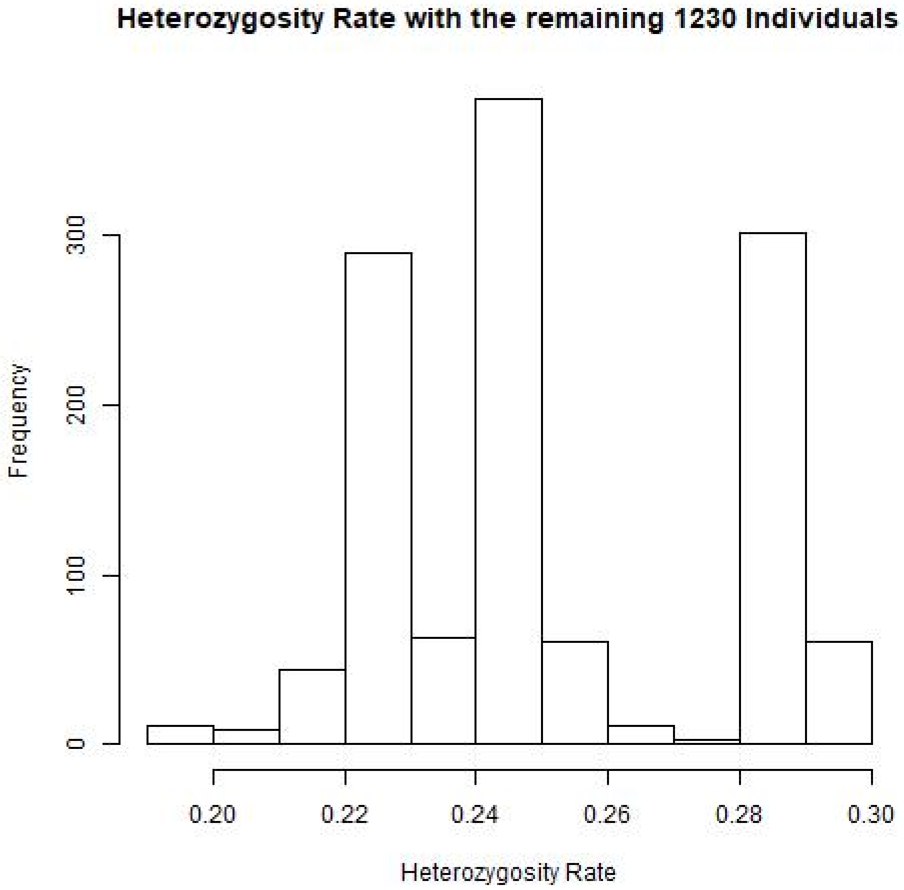


Figure 20. Distribution of the heterozygosity rate of our remaining 1230 individuals.

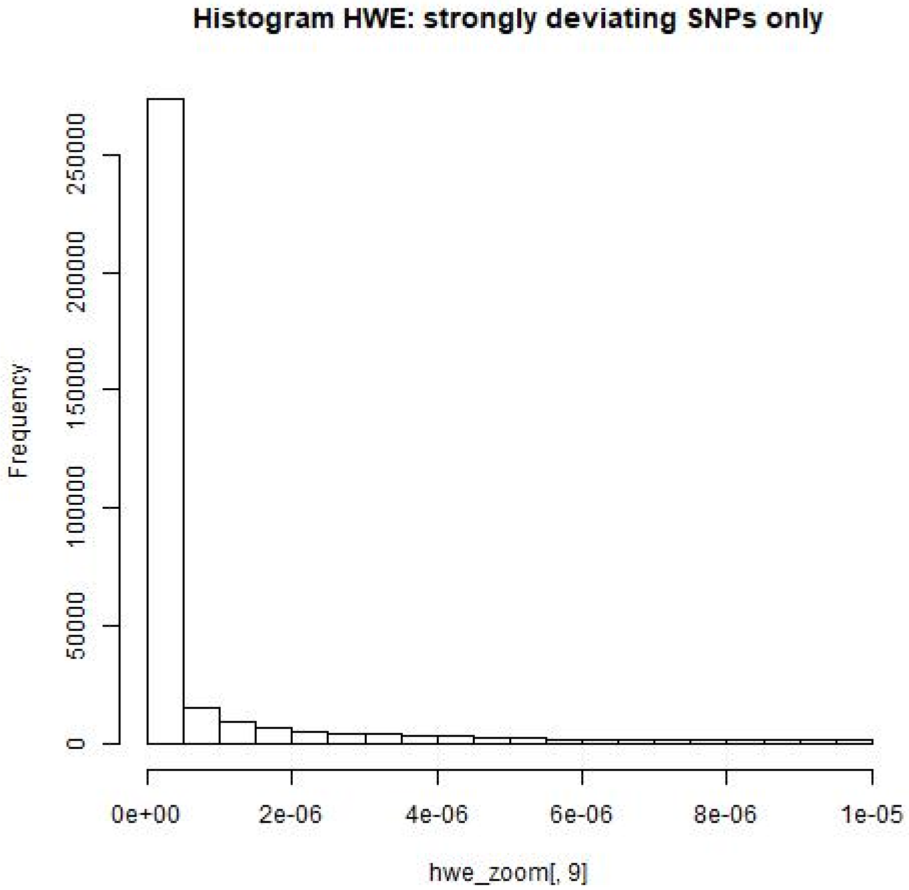


Figure 21. Distribution of the SNPs with HWE p-value below 0.00001 of the 1230 individuals.

4.5.6 Relatedness Check.

The below plot visualizes the parent-offspring relations of individuals and individuals that are unrelated using the z values. We see that there are no unrelated individuals in the dataset. There are all related.

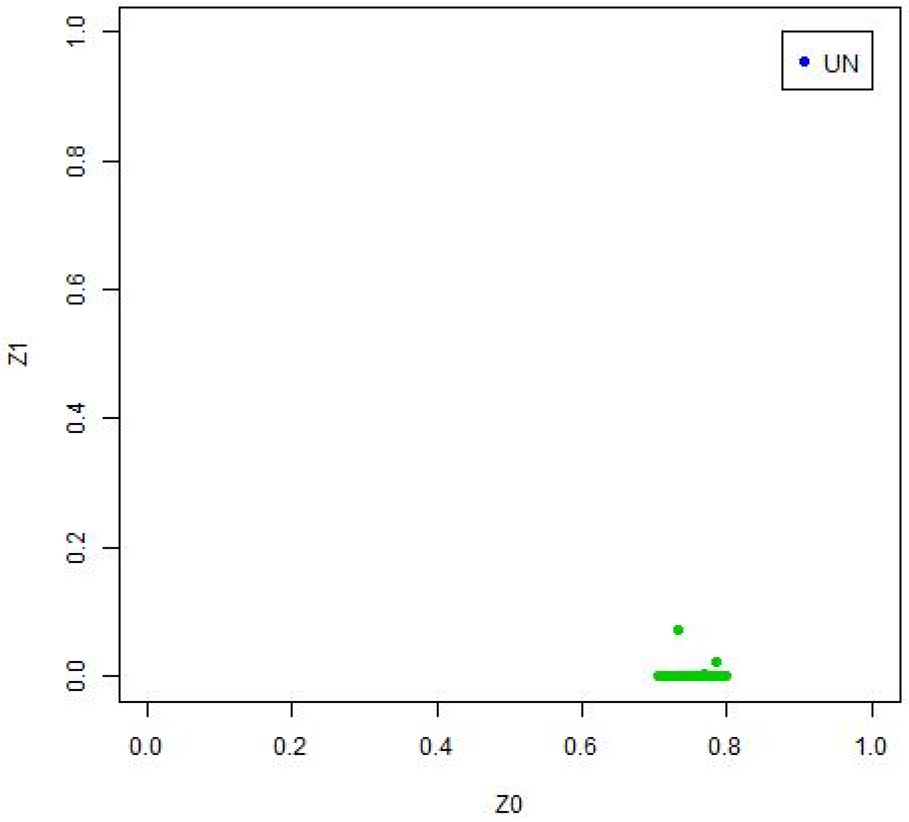


Figure 22. Visualization for parent-offspring relations and unrelated individuals using the z values; PO = parent-offspring, UN = unrelated individuals

To demonstrate that all of the relatedness was due to parent-offspring we only include founders (individuals without parents in the dataset) and plot that in which we see that no one is plotted meaning there are no unrelated individuals.

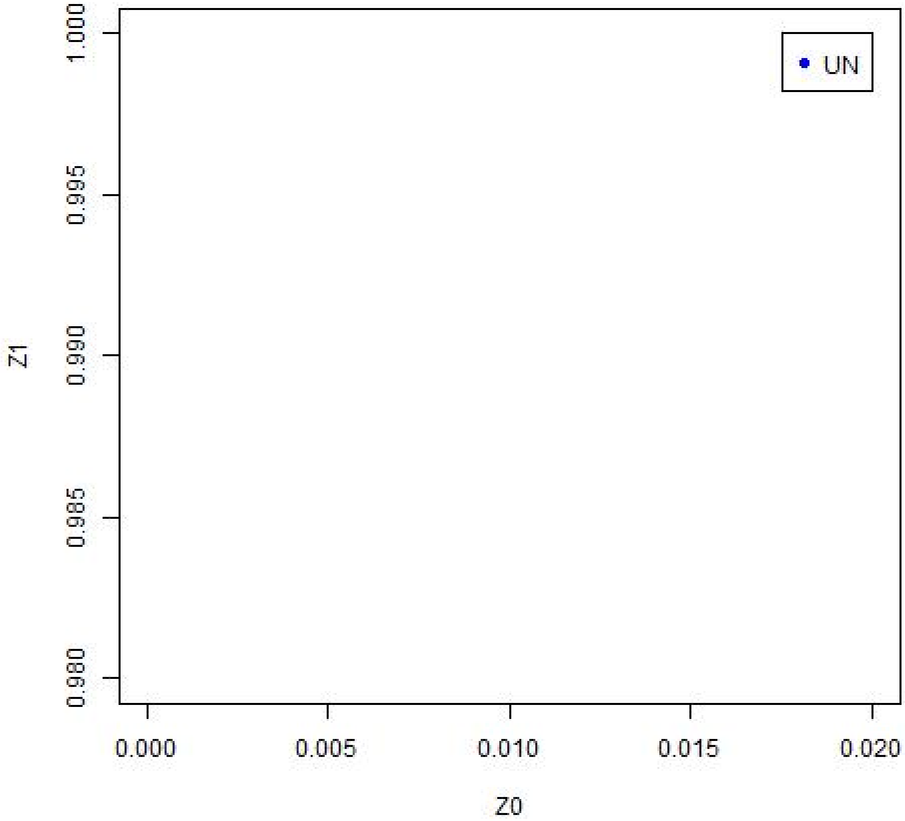


Figure 23. Visualization for parent-offspring relations and unrelated individuals using the z values for only founder individuals.

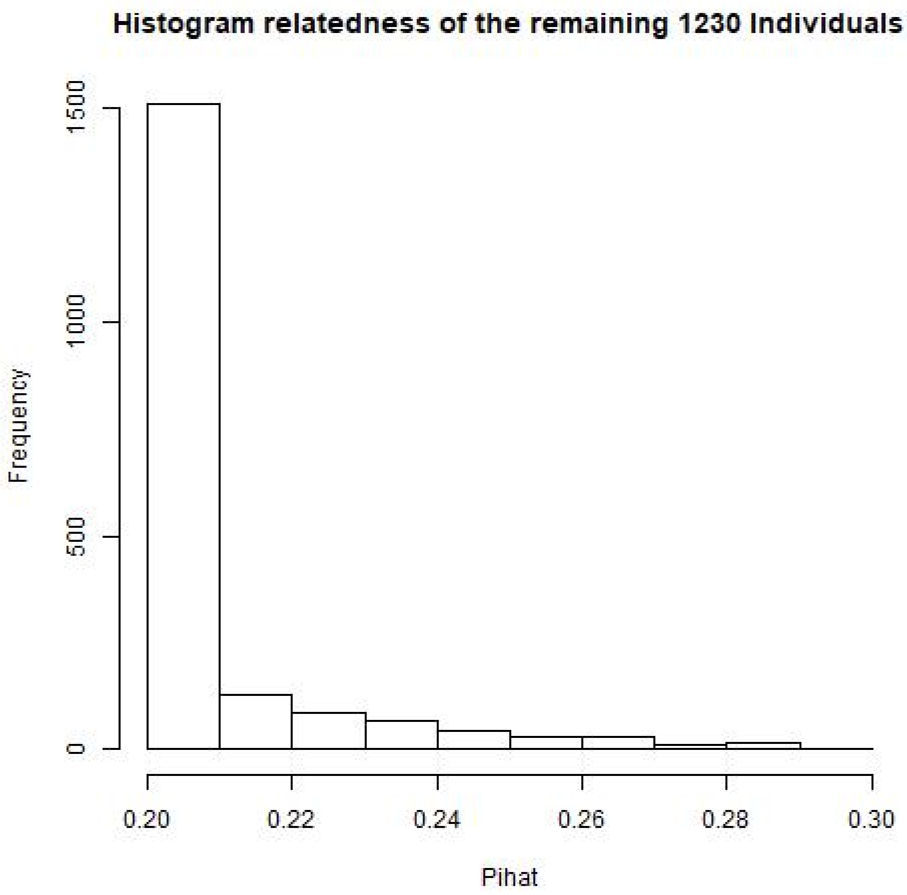


Figure 24. Histogram of relatedness indicator phi of the 1230 individuals.

5 Discussion

5.1 Main GWAS Results

First of all, the continuous and binary phenotypes were assigned completely randomly, hence, we expected to not see any significant results for the GWAS.

Now we observed that for all of the Association results and all the tests conducted for the G wash none of the snips were found to be statistically significant at the significance threshold of 5x10^(-8) that has been determined by are the paper cited. This being said, several snips in all four association tests were found to be statistically significant at the 10^(-5) significance level further. For the first three analysis in which covariates were not included even though not statistically significant at the five to the 5x10^(- 8) significance level several, SNIPs were found to have a reasonably strong signal with p-value of up to 10^(-6) and 10^(-7). Further, some of the SNIPs had stronger signal than others and clustering of significant SNIPs was also found. The clustering of significant SNIPs can be explained by linkage disequilibrium which we talk about in more detail later. Moreover, evidence of signal from the association study can be seen from the QQ plots and the histograms. The p-values in the histogram are not completely uniformly distributed and in the QQ plots the p-values do not follow exactly the diagonal line. The slight deviation from the diagonal line, especially the points deviating near the beginning of the plots, indicate a signal in the association tests. Though not a highly strong, the fact that even this much of signal was observed in the association study between SNIPs and phenotypes is a clear indication of false positive results as the phenotypes were completely randomly assigned, and no signal is expected to be observed. This is most likely due to confounding. The first three of these analysis did not include covariates: they were simply the chi-squared test, logistic regression test and the linear regression test without adjustment. However, when for the last analysis a logistic regression was performed with the principle components as covariates which adjusts for population structure and peopleâ□□s heritage the signal of the snips that were found to be strong became significantly weaker indicating that in the previous three analysis confounding by population structure was occurring. Further, when we adjust for population structure as the principal components covariates, we notice that in the QQ plot and the histogram plot in fingers 20 and 21 respectively, follow more closely to the diagonal line and deviate less from it and for the histogram the plot looks more uniformly distributed then the histogram of the three previous analysis which did not include covariates. Therefore, confounding by population structure is clearly the reason for getting false positives for this dataset. This is also very reasonable as the people in this study are from several distinct geographical region as mentioned in the introduction. Lastly, after speaking with the TA Jianhui, he said if people with only European heritage are kept, then we get completely see no signal with the QQ-plot being perfectly diagonal and the histogram also being perfectly, uniformly distributed as we would expect.

We also note that all tests of associations produce the same top ten SNPs for both continuous and binary phenotypes. Which further indicates that the signal is due to some confounder as this is unusual.

5.2 Population Structure and Effect of PCA Covariates.

Effect of covariates (population stratification, that is, controlling for genetic variation by population in different races).

We note that when we take into account population structure using principal components as covariates it is able to take into account the variation of the the data and and take into account the vibration. Hence, reducing the false negative rage. Here we can see that the population structure was acting as a confounder. Similarly, the individual’s age, age-squared and sex should also be included as covariates and should be adjusted for since they many also be potential confounders. Further, we only used three principal components as covariate, it is possible that if we had used more principal components such that which explained up to to 90% of the variation in the data this would correct for covariates and reduce false positive results even more. Even though we did perform PCA and used three principal components we did not know how much of the variation is explained by these three components. Therefore, it is also recommended that the SmartPca package in PLINK should be used as it automatically includes enough principal components as to capture most of the variation in the data but also not too many so to become overly computationally intensive

The reason we chose three principal components is because that two is the minimum that should be required and we thought choosing one above the minimum would be good for illustrative purposes and also be fairly less computationally intensive where as when we used five or six principal components the it was already taking a lot of computational time.

5.3 Effect of Minor Allele Frequency.

We should remove SNPs with a low MAF frequency because typically SNPs with low MAF will not have sufficient power to detect signal. One advantage is that we decrease the number of tests we perform. This is because the more SNPs the more tests we have therefore the higher the correction needed for multiple testing, hence want low correction therefore just removing these SNPs which we think we will not have enough sample and not enough power to detect statistically significant, is better to remove them and have lower correction coefficient. This also does not have any significant impact on the Manhattan plot.

5.4 Effect of Linkage Disequilibrium

We can see that when there is a signal for one SNP, high p-value, typically, the other SNP's close to it also appear to cluster around it and give of some signal. This clustering effect for significance is the consequence of Linkage Disequilibrium as SNP's close to the SNP with highest p-value (strongest signal) is close to the other SNP's and hence strongly correlated to them. Thus, since we are doing a tests of association, when one SNP has strong signal then the SNPs close to it which are highly correlated with it due to small distance also give off a signal, albeit less strong.

5.5 Quality Control Steps

The order of the steps of Quality Control is important. The particular order in which we did the quality control steps was the same as the one performed in the PLINK GWAS tutorial (ref). Though we recognize that more thought and reasoning should go into the order as it could have an important effect of the file results.

As mentioned in the data section this data has people related up to 3rd cousin. There are people who are related. Typically, we would remove related people in a GWAS unless if we wanted to a relatedness study. In such a case we would have to conduct appropriate analysis and it would not be correct to simply conduct a GWAS as we did here.

5.6 Conclusion

We found that it was very important to perform a GWAS study with actual data and going through the entire quality control steps to the visualizations of the results. This paper also provides a guide as to how to perform a GWAS study from start to finish in Windows 10 operation system.

6 Acknowledgements

I would like to acknowledge first and foremost Professor Lei Sun for the great experience and continued support for this project. Secondly, I would like to thank the TA Jianhui without whom I would not been able to run association test and PCA in PLINK. Finally, I would like to acknowledge Christina in helping me understand PLINK, genetics, and general support.

7 References

1. Sun, L. (2020, September 16). Lecture Module 6 - 14. University of Toronto, Department of Statistical Sciences, FA, Division of Biostatistics, Dalla Lana School of Public Health.

2. Marees, Andries T., et al. "A tutorial on conducting genomeâ□□wide association studies: Quality control and statistical analysis." International journal of methods in psychiatric research 27.2 (2018): e1608.

3. Roslin, Nicole M., Li Weili, Andrew D. Paterson, and Lisa J. Strug. "Quality control analysis of the 1000 Genomes Project Omni2. 5 genotypes." BioRxiv (2016): 078600.

8 Appendix

8.1 Detailed QC Steps with Reproducible Code.

Below are the command line commands to perform the QC steps in PLINK. All the Plink commands will be run in the Windows Command Prompt. All the Rscript command must be run in Git Hub Command Prompt.

Before running the QC steps we will copy the original data files and call them indep_1 and run the QC steps on this set of original bfiles, so we do not by mistake change the original files.

Need to first go to the 1000GenomeProj directory in both the Windows Command Prompt and the Git Bash Command Prompt. Below are the sample commands:

```
cd OneDrive\OneDrive\Statistical Genetics\HW2\1000GenomeProj  # Windows Command Prompt

cd "OneDrive/OneDrive/Statistical Genetics/HW2/1000GenomeProj"   # Git Bash Command Prompt
```

8.1.1 Step 1: Check Missingness of Genotypic Data.

First, we investigate the missingness of the data.

```
# Investigate missingness per individual and per SNP and make histograms.
plink --bfile indep_1 --missing
```

The output: plink.imiss and plink.lmiss, these files show respectively the proportion of missing SNPs per individual and the proportion of missing individuals per SNP.

```
# Generate plots to visualize the missingness results.
Rscript --no-save hist_miss.R
```

Delete SNPs and individuals with high levels of missingness.

The following two QC commands will not remove any SNPs or individuals. However, it is good practice to start the QC with these non-stringent thresholds.

```
# Delete SNPs with missingness >0.2. (even more than C 0.005) 1744 individuals
plink --bfile indep_1 --geno 0.2 --make-bed --out indep_2

# Delete individuals with missingness >0.2.
plink --bfile indep_2 --mind 0.2 --make-bed --out indep_3

# Delete SNPs with missingness >0.02. This is a 10 times more stringent critrion.
plink --bfile indep_3 --geno 0.02 --make-bed --out indep_4

# Delete individuals with missingness >0.02. This is a 10 times more stringent critrion.
plink --bfile indep_4 --mind 0.02 --make-bed --out indep_5
```

After this step we have 1971324 variants and 1755 people that pass filters and QC.

8.1.2 Step 2: Check for Sex Discrepancy.

We will now check for sex discrepancy.

Subjects who were a priori determined as females must have a F value of <0.2, and subjects who were a priori determined as males must have a F value >0.8. This F value is based on the X chromosome inbreeding (homozygosity) estimate. Subjects who do not fulfill these requirements are flagged “PROBLEM” by PLINK.

```
plink --bfile indep_5 --check-sex
```

We will generate plots to visualize the sex-check results.

```
Rscript --no-save gender_check.R
```

OUTPUT PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/ (C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to plink.log. Options in effect: –bfile indep_5 –check-sex

8113 MB RAM detected; reserving 4056 MB for main workspace. 1971324 variants loaded from .bim file. 1755 people (842 males, 898 females, 15 ambiguous) loaded from .fam. Ambiguous sex IDs written to plink.nosex . Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 1755 founders and 0 nonfounders present. Calculating allele frequencies… done. Warning: Nonmissing nonmale Y chromosome genotype(s) present; many commands treat these as missing. Total genotyping rate is 0.997838. 1971324 variants and 1755 people pass filters and QC. Note: No phenotypes present. –check-sex: 22651 Xchr and 0 Ychr variant(s) scanned, 525 problems detected. Report written to plink.sexcheck .

These checks indicate that there is one woman with a sex discrepancy, F value of 0.99. (When using other datasets often a few discrepancies will be found).

The following two scripts can be used to deal with individuals with a sex discrepancy. Note, please use one of the two options below to generate the bfile hapmap_r23a_6, this file we will use in the next step of this tutorial.

```
# 1) Delete individuals with sex discrepancy.
# The below code will not work in the Windows command prompt as we get the error: 'awk' is not recognized as an internal or external command, operable program or batch file.
grep "PROBLEM" plink.sexcheck| awk '{print$1,$2}'> sex_discrepancy.txt
# Hence, it must be run in the Git Bash command prompt

# This command generates a list of individuals with the status ?PROBLEM?.
plink --bfile indep_5 --remove sex_discrepancy.txt --make-bed --out indep_6
# This command removes the list of individuals with the status ?PROBLEM?.

# 2) impute-sex.
plink --bfile indep_5 --impute-sex --make-bed --out indep_6
# This imputes the sex based on the genotype information into your data set.
```

OUTPUT

PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/ (C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to indep_6.log. Options in effect: –bfile indep_5 –make-bed –out indep_6 –remove sex_discrepancy.txt

8113 MB RAM detected; reserving 4056 MB for main workspace. 1971324 variants loaded from .bim file. 1755 people (842 males, 898 females, 15 ambiguous) loaded from .fam. Ambiguous sex IDs written to indep_6.nosex . –remove: 1230 people remaining. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 1230 founders and 0 nonfounders present. Calculating allele frequencies… done. Total genotyping rate in remaining samples is 0.997799. 1971324 variants and 1230 people pass filters and QC. Note: No phenotypes present. –make-bed to indep_6.bed + indep_6.bim + indep_6.fam … done.

At this step we have 1971324 variants and 1755 people that pass filters and QC.

8.1.3 Step 3: Check Minor Allele Frequency.

We will generate a bfile with autosomal SNPs only and delete SNPs with a low minor allele frequency (MAF).

```
# Select autosomal SNPs only (i.e., from chromosomes 1 to 22).
awk '{ if ($1 >= 1 && $1 <= 22) print $2 }' indep_6.bim > snp_1_22.txt
# Got another error for awk: does not recognize it as an internal or external command, operable
 program or batch file.

plink --bfile indep_6 --extract snp_1_22.txt --make-bed --out indep_7

# Generate a plot of the MAF distribution.
plink --bfile indep_7 --freq --out MAF_check

Rscript --no-save MAF_check.R
```

OUTPUT

PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/ (C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to indep_7.log. Options in effect: –bfile indep_6 –extract snp_1_22.txt –make-bed –out indep_7

8113 MB RAM detected; reserving 4056 MB for main workspace. 1971324 variants loaded from .bim file. 1230 people (842 males, 388 females) loaded from .fam. –extract: 1948615 variants remaining. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 1230 founders and 0 nonfounders present. Calculating allele frequencies… done. Total genotyping rate is 0.997808. 1948615 variants and 1230 people pass filters and QC. Note: No phenotypes present. –make-bed to indep_7.bed + indep_7.bim + indep_7.fam … done.

After doing this we find 1948615 variants remain and all 1230 people are kept.

Remove SNPs with a low MAF frequency. #Minor allele Frequency: the effect of minor allele frequency # SO if we do not remove this step do we get something different in the Manhattan plots. Find which SNP's are rare and what are their p=values if they are included in the association analysis. Look at the p-values of these SNP's. Do we remove them because they skew the results? Do they have small p-values and do they affect the other SNP's p-values? = No. Is it because the more SNP's the more tests therefore the higher the correction needed for multiple testing, hence want low correction therefore just removing these snp's which we think we will not have enough sample (and not enough power to detect statistically significant), hence better to remove them and have lower correction coefficient. We can empirically verify this.

```
plink --bfile indep_7 --maf 0.03 --make-bed --out indep_8 #0.05 == 0.03
# 1073226 SNPs are left for 0.05
# A conventional MAF threshold for a regular GWAS is between 0.01 or 0.05, depending on sample s
ize.
```

PLINK v1.90b6.21 64-bit (19 Oct 2020) www.cog-genomics.org/plink/1.9/ (C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to indep_8.log. Options in effect: –bfile indep_7 –maf 0.03 –make-bed –out indep_8

8113 MB RAM detected; reserving 4056 MB for main workspace. 1948615 variants loaded from .bim file. 1230 people (842 males, 388 females) loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 1230 founders and 0 nonfounders present. Calculating allele frequencies… done. Total genotyping rate is 0.997808. 562794 variants removed due to minor allele threshold(s) (–maf/–max-maf/–mac/–max-mac). 1385821 variants and 1230 people pass filters and QC. Note: No phenotypes present. –make-bed to indep_8.bed + indep_8.bim + indep_8.fam … done.

562794 variants removed due to minor allele threshold(s). 1385821 variants and 1230 people pass filters and QC.

8.1.4 Step 4: Check Hardy-Weinberg Equilibrium for SNP’s

We will delete SNPs which are not in Hardy-Weinberg equilibrium (HWE).

```
# Check the distribution of HWE p-values of all SNPs.
plink --bfile indep_8 --hardy
# Selecting SNPs with HWE p-value below 0.00001, required for one of the two plot generated by t
he next Rscript, allows to zoom in on strongly deviating SNPs.
awk '{ if ($9 <0.00001) print $0 }' plink.hwe>plinkzoomhwe.hwe
Rscript --no-save hwe.R

# By default the --hwe option in plink only filters for controls.
# Therefore, we use two steps, first we use a stringent HWE threshold for controls, followed by
 a less stringent threshold for the case data.
plink --bfile indep_8 --hwe 1e-6 --make-bed --out indep_hwe_filter_step1

# The HWE threshold for the cases filters out only SNPs which deviate extremely from HWE.
# This second HWE step only focusses on cases because in the controls all SNPs with a HWE p-valu
e < hwe 1e-6 were already removed
plink --bfile indep_hwe_filter_step1 --hwe 1e-10 --hwe-all --make-bed --out indep_9
```

OUTPUT 1:

1385821 variants loaded from .bim file. 1230 people (842 males, 388 females) loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 1230 founders and 0 nonfounders present. Calculating allele frequencies… done. Total genotyping rate is 0.997641. –hardy: Writing Hardy-Weinberg report (founders only) to plink.hwe … done.

OUTPUT 2: plink –bfile indep_hwe_filter_step1 –hwe 1e-10 –hwe-all –make-bed –out indep_9

Note: –hwe-all flag deprecated. Use “–hwe include-nonctrl”. 8113 MB RAM detected; reserving 4056 MB for main workspace. 1170861 variants loaded from .bim file. 1230 people (842 males, 388 females) loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 1230 founders and 0 nonfounders present. Calculating allele frequencies… done. Total genotyping rate is 0.997654. –hwe: 0 variants removed due to Hardy-Weinberg exact test. 1170861 variants and 1230 people pass filters and QC. Note: No phenotypes present. –make-bed to indep_9.bed + indep_9.bim + indep_9.fam … done.

8.1.5 Step 5: Check Heterozygosity Rate

We will generate a plot of the distribution of the heterozygosity rate of your subjects and remove the individuals with a heterozygosity rate deviating more than 3 sd from the mean.

Checks for heterozygosity are performed on a set of SNPs which are not highly correlated. Therefore, to generate a list of non-(highly)correlated SNPs, we exclude high inversion regions (inversion.txt [High LD regions]) and prune the SNPs using the command r –indep-pairwise.

```
# The parameters ?50 5 0.2? stand respectively for: the window size, the number of SNPs to shift the window at each step, and the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously.

plink --bfile indep_9 --exclude inversion.txt --range --indep-pairwise 50 5 0.2 --out indepSNP
# Note, don't delete the file indepSNP.prune.in, we will use this file in later steps of the tutorial.

plink --bfile indep_9 --extract indepSNP.prune.in --het --out R_check
# This file contains your pruned data set.

# Plot of the heterozygosity rate distribution
Rscript --no-save check_heterozygosity_rate.R

# The following code generates a list of individuals who deviate more than 3 standard deviations from the heterozygosity rate mean.
# For data manipulation we recommend using UNIX. However, when performing statistical calculations R might be more convenient, hence the use of the Rscript for this step:
Rscript --no-save heterozygosity_outliers_list.R

# Output of the command above: fail-het-qc.txt .
# When using our example data/the HapMap data this list contains 2 individuals (i.e., two individuals have a heterozygosity rate deviating more than 3 SD's from the mean).
# Adapt this file to make it compatible for PLINK, by removing all quotation marks from the file and selecting only the first two columns.
sed 's/"// g' fail-het-qc.txt | awk '{print$1, $2}'> het_fail_ind.txt

# Remove heterozygosity rate outliers.
plink --bfile indep_9 --remove het_fail_ind.txt --make-bed --out indep_10
```

8.1.6 Step 6: Checking Relatedness.

It is essential to check datasets you analyse for cryptic relatedness. Assuming a random population sample, we should in general exclude all individuals above the pihat threshold of 0.2, which we do not do in this tutorial.

```
# Check for relationships between individuals with a pihat > 0.2.
plink --bfile indep_10 --extract indepSNP.prune.in --genome --min 0.2 --out pihat_min0.2

# The HapMap dataset is known to contain parent-offspring relations.
# The following commands will visualize specifically these parent-offspring relations, using the z values.
awk '{ if ($8 >0.9) print $0 }' pihat_min0.2.genome>zoom_pihat.genome

# Generate a plot to assess the type of relationship.
Rscript --no-save Relatedness.R
```

The generated plots show a considerable amount of related individuals (exploitation plot; PO = parent-offspring, UN = unrelated individuals). Normally, family based data should be analyzed using specific family based methods.

```
plink --bfile indep_10 --filter-founders --make-bed --out indep_11

# Now we will look again for individuals with a pihat >0.2.
plink --bfile indep_11 --extract indepSNP.prune.in --genome --min 0.2 --out pihat_min0.2_in_founders
# The file 'pihat_min0.2_in_founders.genome' shows that, after exclusion of all non-founders, only 1 individual pair with a pihat greater than 0.2 remains in the HapMap data.
# This is likely to be a full sib or DZ twin pair based on the Z values. Noteworthy, they were not given the same family identity (FID) in the HapMap data.

# For each pair of 'related' individuals with a pihat > 0.2, we recommend to remove the individual with the lowest call rate.
plink --bfile indep_11 --missing
# Use an UNIX text editor (e.g., vi(m) ) to check which individual has the highest call rate in the 'related pair'.

# Generate a list of FID and IID of the individual(s) with a Pihat above 0.2, to check who had the lower call rate of the pair.
# In our dataset the individual 13291  NA07045 had the lower call rate.
vi 0.2_low_call_rate_pihat.txt
i
13291  NA07045
# Press esc on keyboard!
:x
# Press enter on keyboard
# In case of multiple 'related' pairs, the list generated above can be extended using the same method as for our lone 'related' pair.

# Delete the individuals with the lowest call rate in 'related' pairs with a pihat > 0.2
plink --bfile indep_11 --remove low_call_rate_pihat_0.2.txt --make-bed --out indep_12

plink --bfile indep_11 --remove low_call_rate_pihat_0.2.txt --make-bed --out indep_12
```

8.1.7 Performing PCA in PLINK.

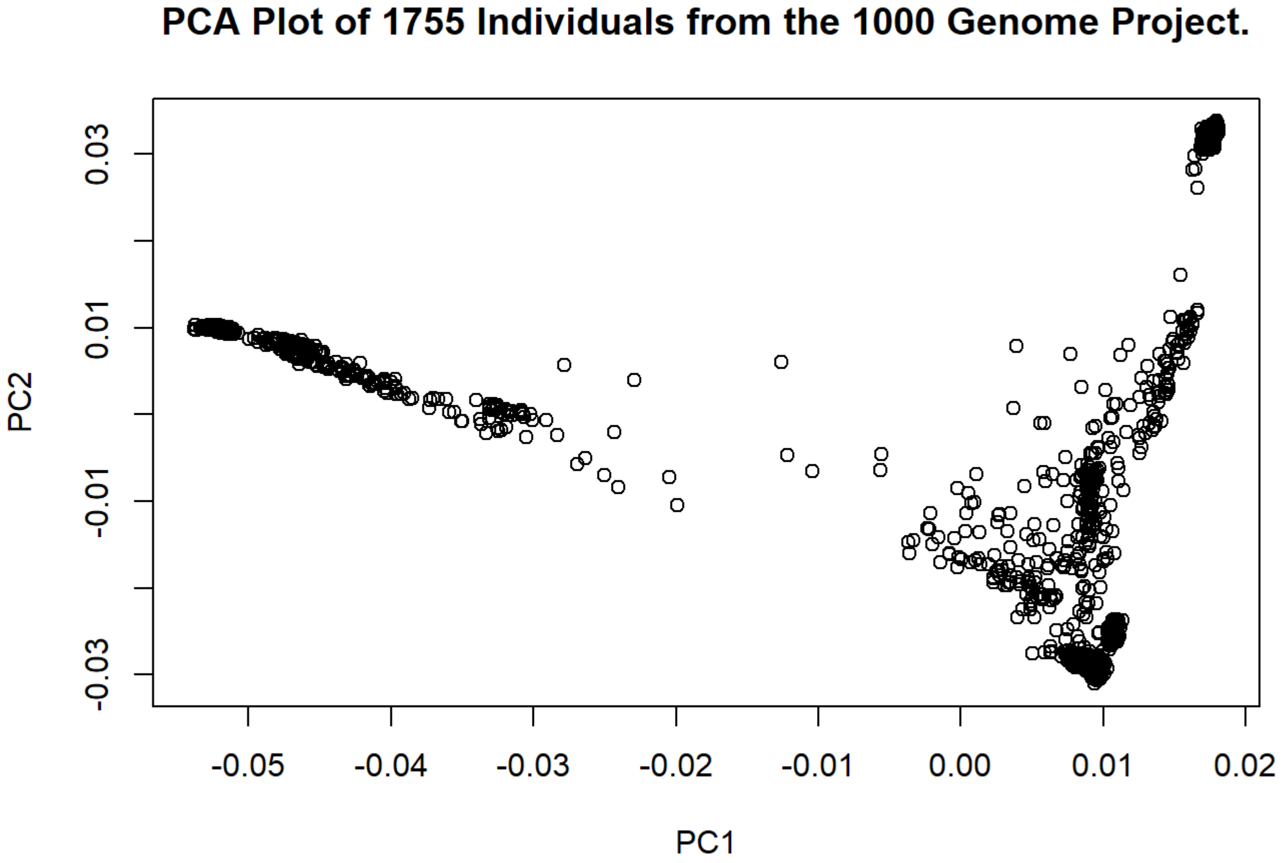
PLINK Code.

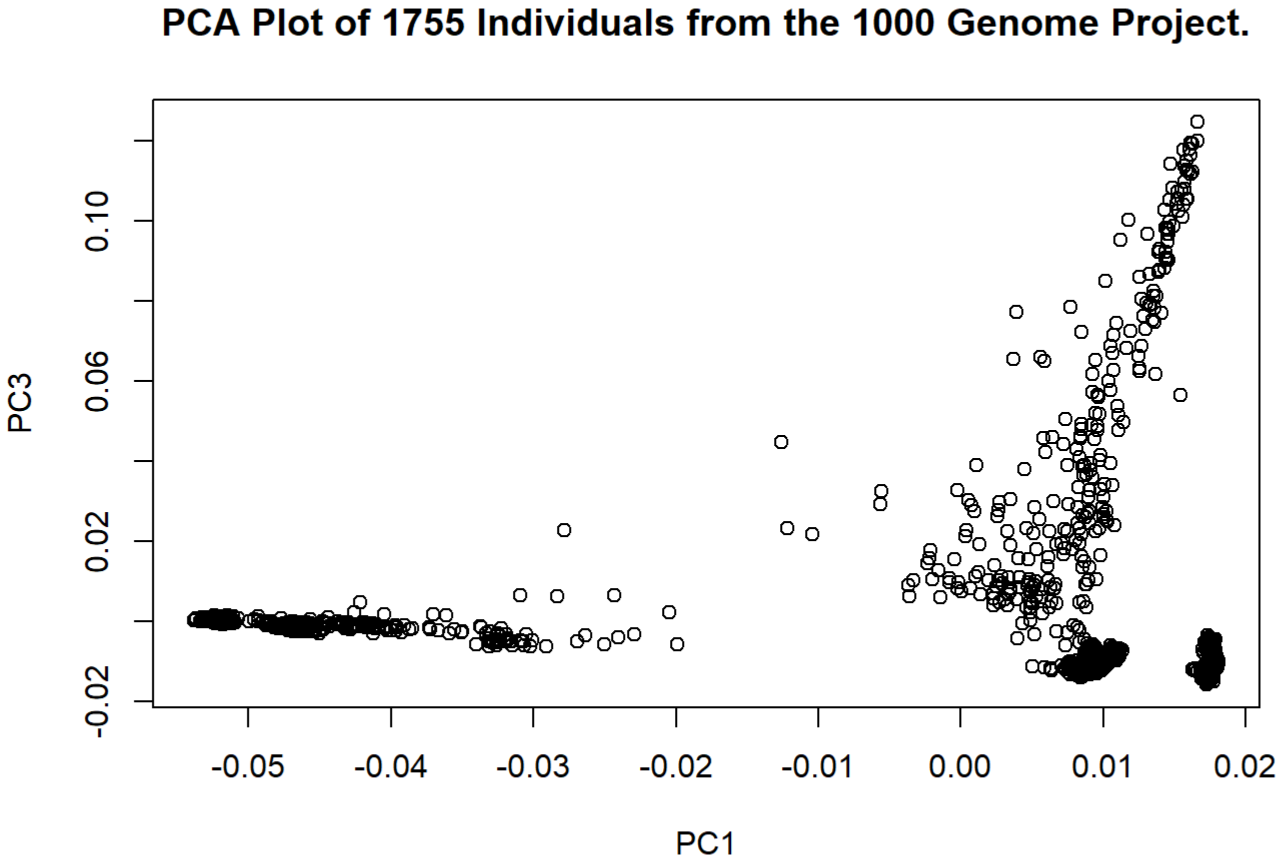
8.1.7.1 PLINK Commands

8.1.7.1.1 Performing PCA in PLINK.

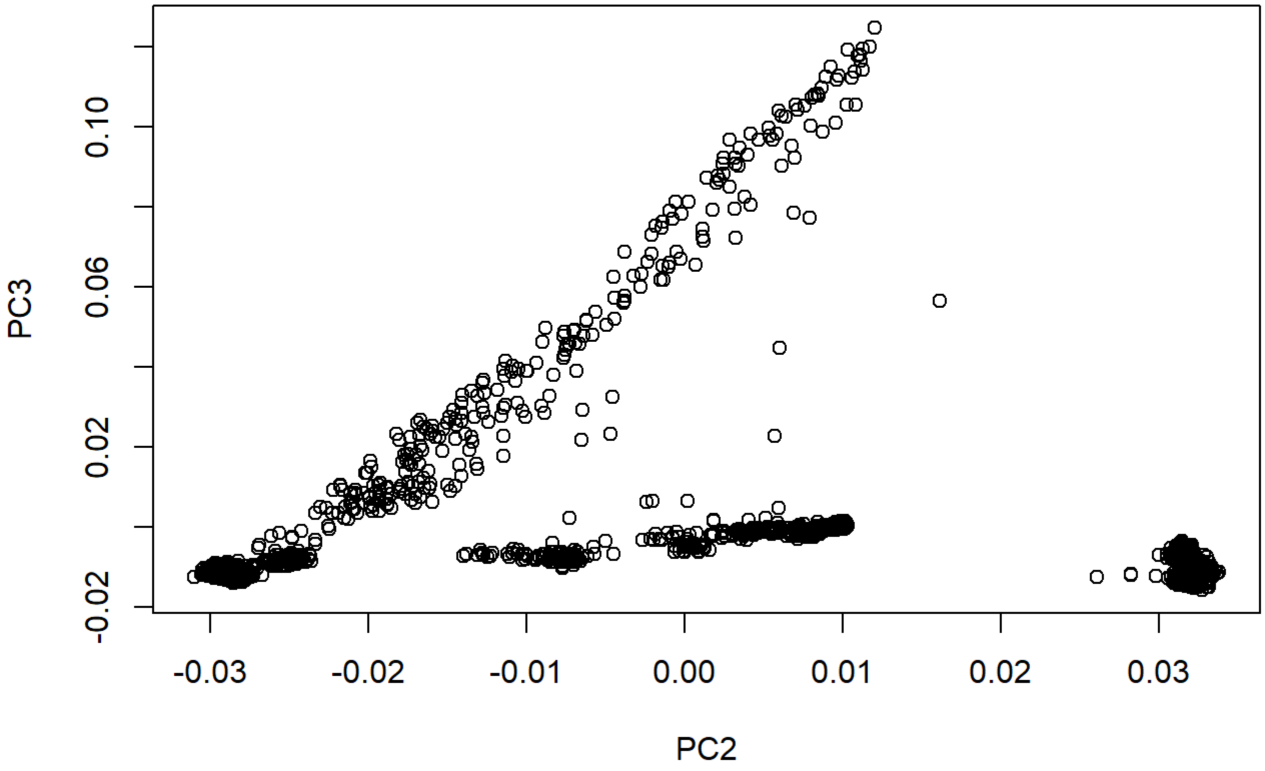
Run a PCA.

8.1.7.1.2 Logistic regression with PCA components as covariates.


PCA Plot of 1755 Individuals from the 1000 Genome Project.
PC2
0.03
0.01
-0.01
-0.03
PC1
-0.05
-0.04
-0.03
-0.02
-0.01
0.00
0.01
0.02


PCA Plot of 1755 Individuals from the 1000 Genome Project.
PC3
0.10
0.06
0.02
-0.02
PC1
-0.05
-0.04
-0.03
-0.02
-0.01
0.00
0.01
0.02

PCA Plot of 1755 Individuals from the 1000 Genome Project.


PC3
PC2