Introduction to GWAS

(assuming a general biology knowledge of genetics, practice in introductory statistics, and no previous computing knowledge)

What is a GWAS? (source)

A GWAS is a genome-wide association study that is used for investigating the genetic architecture of complex traits by analyzing DNA sequence variations across the genome for association with a phenotype, commonly for SNPs but now expanding to entire genes. These associations can be studied for a vast number of phenotypes, such as disease susceptibility to cholesterol levels to drug responses. However, finding strong associations is rare due to the multi-gene associations present with many symptoms, but it varies strongly by phenotype.

Definitions

Bonferroni correction (GWAS): corrects for multiple testing in simultaneous tests, which would normally have an erroneous amount of significance, assuming all tests are independent. In a GWAS, this threshold is $5*10^{-8}$.

Cohort: a collection of samples, usually individuals in GWAS, being studied

Covariate: an extraneous control variable that can affect the main analysis, and is run in the background to account for the difference

ggplot2: an R graphical package, in this example used to make a principal components analysis plot

Haplotype: genes in a region commonly inherited together

HapMap: a catalog of common genetic variants assorted into a haplotype by population around the world (ex. Chinese in Bejing, Yoruba in Nigeria) that can be imputed into study data

Heterozygosity (--het): observed and expected autosomal homozygous genotype counts for each sample, and method-of-moments F coefficient estimates

High density lipoproteins (HDL): “good” cholesterol; carries cholesterol back from the cells to the liver; helps remove LDL cholesterol from the arteries; higher levels are associated with lower risk of cardiovascular disease

Identity-by-descent (--genome): calculation of relatedness of individuals within a cohort

Imputation: Input of “missing” SNPs into study data using a reference panel, performed for free on the Michigan Imputation Server or the Sanger Imputation Server

Low density lipoproteins (LDL): “bad” cholesterol; carries cholesterol to cells; main source of cholesterol buildup and blockage in the arteries, with higher levels associated with higher risk of cardiovascular disease

Linkage disequilibrium (--indep-pairwise): a connection between SNPs due to their shared inheritance; can be pruned to analyze less SNPs while still remaining accurate

LocusZoom: a tool that plots regional association results from genome-wise association scans using PLINK .assoc.linear files as input

Manhattan plot: a plot of SNPs in a study, organized by chromosome, on the x-axis with their p-values on the y-axis, transformed into their negative logarithm. Commonly, the Bonferroni correction is plotted as a line as well.

PLINK: a program used to perform many processes on datasets, including statistics, editing, and other analyses.

PrediXcan: a program that tests the imputed gene expression with the phenotype studied to identify genes in association; tested with tissue-dependent prediction models

Principal component analysis: segregates populations by structure on a set of pruned data points, and plotted using the principal components with the highest association

Shapiro test (--shapiro): a test that the null hypothesis came from a normally distributed population

Triglycerides (TRIG): another “bad” cholesterol, with higher levels also associated with increased risk of cardiovascular dreams

Q-Q plot: associates the observed values between SNPs and the phenotypes between the expected null value, with any large deviations as possibly significant

Example

For the past semester, I have been working on a Yoruba cohort from the database of Genotypes and Phenotypes studying height and cholesterol levels, with an original population size of 1,264, later narrowed down to 1,163 after quality control. In this study, I have found two SNPs of significance, one associated with raising high density lipoprotein-cholesterol, colloquially known as “good” cholesterol, and the other with lowering low density lipoprotein-cholesterol, colloquially known as “bad” cholesterol. Most work is performed with PLINK or RStudio.

Quality control (plots)

The cohort studied after initial quality control consisted of 1,264 people, with 446 males and 818 females. Initial quality control included removing missing call rates > 0.01, leaving a total genotyping rate of 0.999244. The lowest p-value from Hardy-Weinberg test statistics was 0.002879, but there were only 9 founders present in the cohort. This may be because the population is heavily related, as 456 people out of 1,264 had identity by descent (IBD) > 0.125. Due to this high degree of relatedness in a large proportion of the population, the threshold for exclusion was set to a pi-hat value of > 0.25, removing 75 members of the original population to leave 1,189 people in the cohort. Five more outlier members of the population were excluded as they had plus or minus three standard deviations in heterozygosity, creating a working population of 1,184. Further quality control included performing a principal components analysis after generating a pruned subset of SNPs in approximate linkage equilibrium with each other using PLINK’s indep-pairwise function, which left 126,827 SNPs for analysis out of the original 1,522,836 SNPs. Those considered outliers were outside of plus or minus five standard deviations of the YRI HAPMAP3 population, reducing the cohort by 26 to leave a total of 1163 members in the working population for the study.

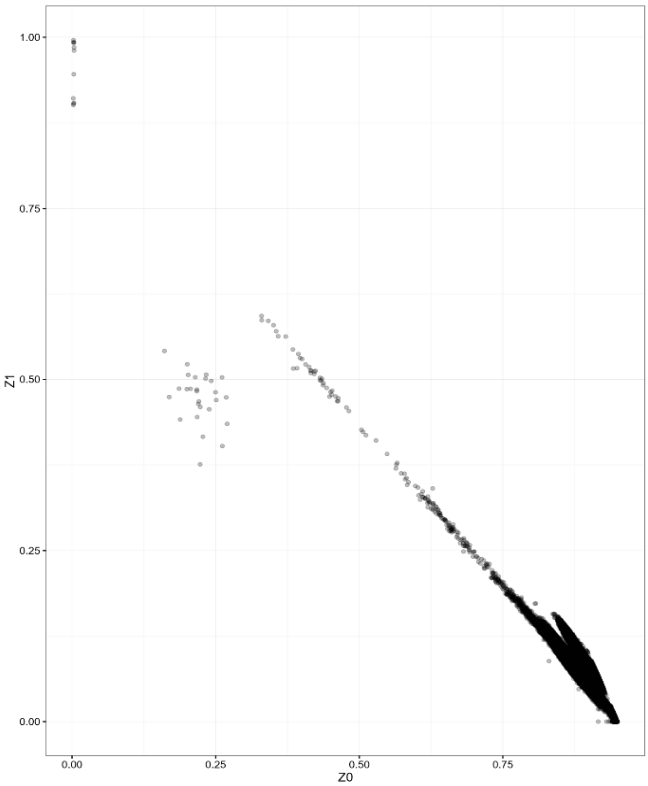


Figure 1: The study cohort experienced high levels of relatedness, seen in the strong linear relationship in this identity by descent graph. Due to this degree of relatedness, the pi-hat value for exclusion from analysis was set to > 0.25 instead of > 0.125.

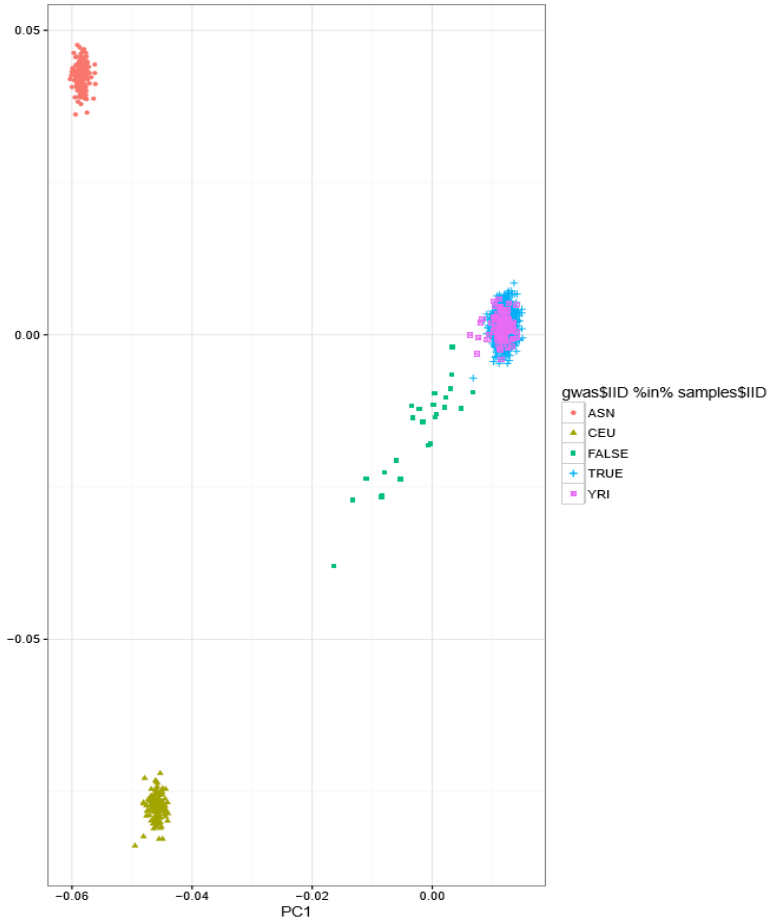


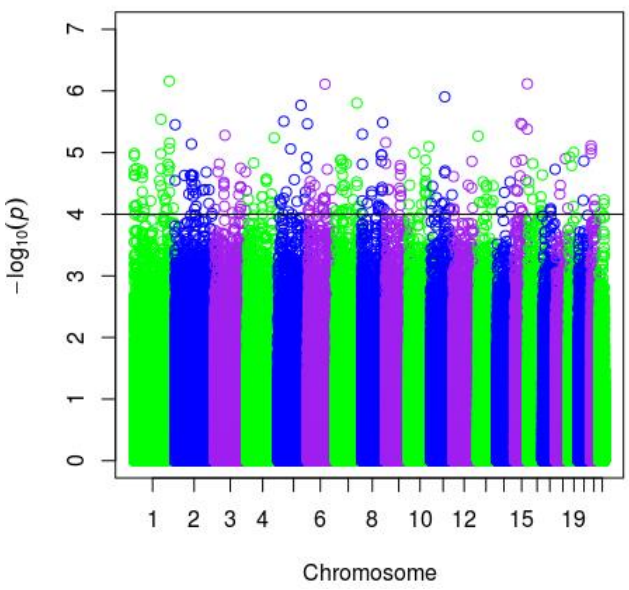
Figure 2: The study population was mapped against a 391 member HAPMAP3 unrelated reference population with a combined sample of Chinese in Bejing and Japanese in Tokyo (ASN), Utah residents with northern and western European ancestry (CEU), and Yoruba from Ibadan, Nigeria (YRI). To reduce skewing, members of the study population within five standard deviations of the HAPMAP3 YRI means were included, excluding 26 original members, depicted as green boxes. After this exclusion, 1,163 of the original 1,264 members of the study population were retained, portrayed as blue plus signs.

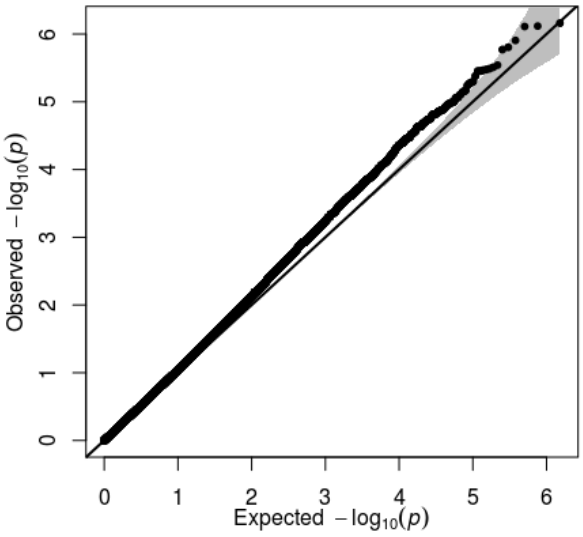
*Conducting the GWAS (source) *add link to pipeline when made*

The genome wide association study concerned five phenotypes: height (in), total cholesterol (mg/dL), high density lipoproteins (mg/dL), triglycerides (mg/dL), and low density lipoproteins (mg/dL). These studies were performed using mainly RStudio, as well as PLINK.

Height

Since height is a known polygenic trait, significant results from the phenotypic data of this population were not expected. Also, as sex is a larger factor in the determination of adult height, measured in inches, it was ran as a covariate in this study, which contained 361 males and 648 females. A Shapiro test for the original data resulted in W = 0.99327 and a p-value of $2.418*10^{-5}$, with no points surpassing the significance threshold of $5*10^{-8}$. The data was then transformed by taking the $\log_{10}$ of the data, resulting in a slight genomic inflation of 1.0519, and giving Shapiro test results of W = 0.99699 and p-value = 0.02002, slightly normalizing the data. However, there was again a lack of significant SNPs in this analysis, leading to no conclusive results in the analysis concerning on height.





	CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
98058	1	kgp5231448	215127260	A	ADD	1220	0.047040	4.991	6.885e-07
1267029	15	kgp11360062	99860960	G	ADD	1221	0.114500	4.971	7.630e-07
613724	6	rs6914292	106006902	G	ADD	1221	0.062650	4.968	7.741e-07

Figures 3, 4, and 5: In the analysis concerning height with sex as a covariate, no significance below $5*10^{-8}$ was found, the Q-Q plot is largely linear and normal, and there are no significant peaks in the Manhattan plot for the logarithmic variation of the data. The SNPs with the lowest p-value in the study only reach $6.885 * 10^{-7}$.

Total Cholesterol

As with height, total cholesterol is a broad combination of traits, so it was further subdivided into high density lipoproteins, triglycerides, and low density lipoproteins, which are all separately investigated later in this study. A Shapiro test on the original data resulting in W = 0.99298 and a p-value of $9.258*10^{-5}$, and the logarithmic version of the data produced a p-value of $8.722*10^{-9}$, so the study was carried on with the original data. All lipid studies in this experiment had 1,009 phenotypes available in the data available from the original population. Like height, no point reached significance below $5*10^{-8}$, but subsequent studies on two individual categories of lipids received significant SNPs.

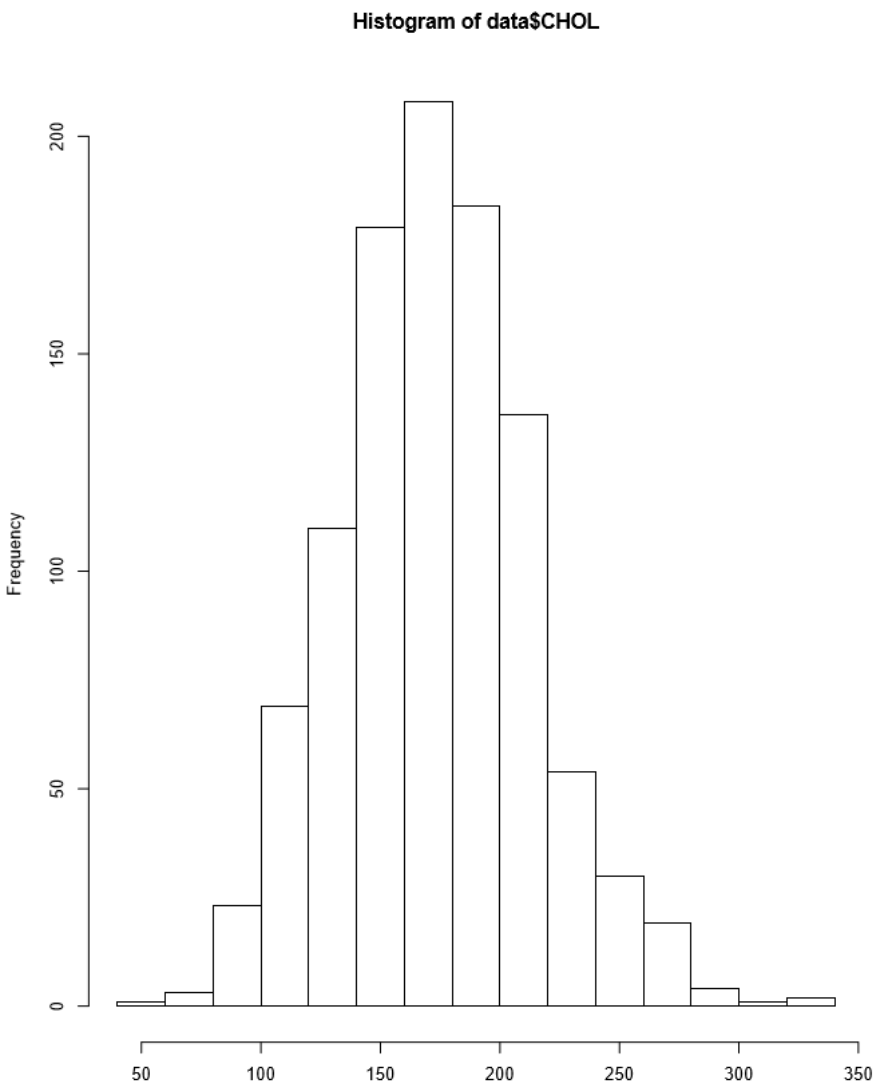
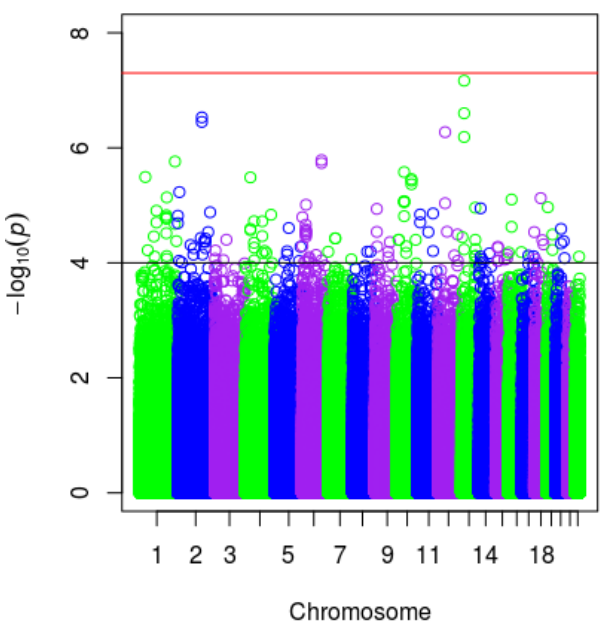
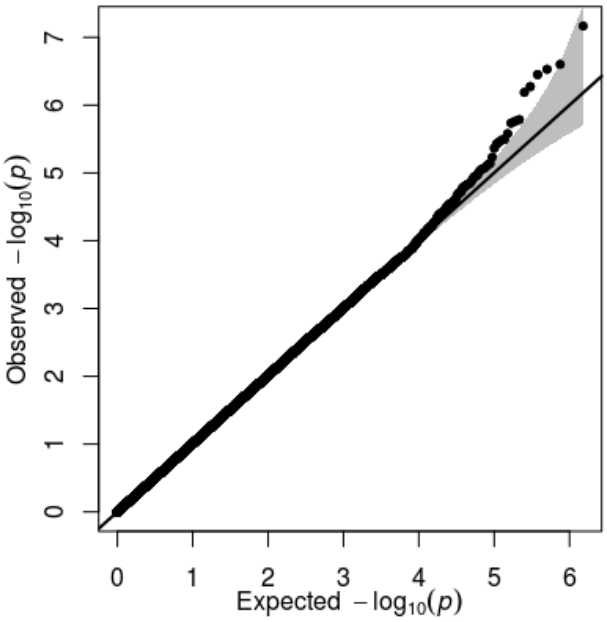


Figure 6: For overall cholesterol levels, the summary values were first quartile 146.20, median 171.60, mean 173.30, and third quartile 198.80, with levels measured in mg/dL. However, the p-value from the Shapiro test was low at $p = 9.258*10^{-5}$

	CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1127311	13	kgp1335700	38023590	A	ADD	1023	28.010	5.436	6.808e-08
1127341	13	kgp11884281	38066517	C	ADD	1023	22.560	5.192	2.505e-07
202351	2	kgp11341957	158333712	A	ADD	1023	18.650	5.161	2.958e-07

Figure 7: In the overall study of cholesterol levels, a broad trait, no one point reached significance under $5*10^{-8}$, with the lowest SNP almost reaching the threshold at $p = 6.808*10^{-8}$.





Figures 8 and 9: In a study composing all lipid traits combined, there is no individual SNP that reaches significance. However, separating the traits into individual types of cholesterol produces results that reach significance.

<u>High Density Lipoproteins</u>

High density lipoproteins are known as the "good" cholesterol, as they take cholesterol from the cells to the liver, and help remove low density lipoprotein cholesterol from the arteries. High levels of HDL are also associated with lower risk for cardiovascular disease. In analysis of these data, the Shapiro test resulted in W = 0.97683 and a p-value of $1.052*10^{-11}$, while the logarithmic variation of the data had p = $9.518*10^{-11}$, so the original data were used in analysis. There were also 1,009 phenotypes available from the study population, and it reached one point of significance.

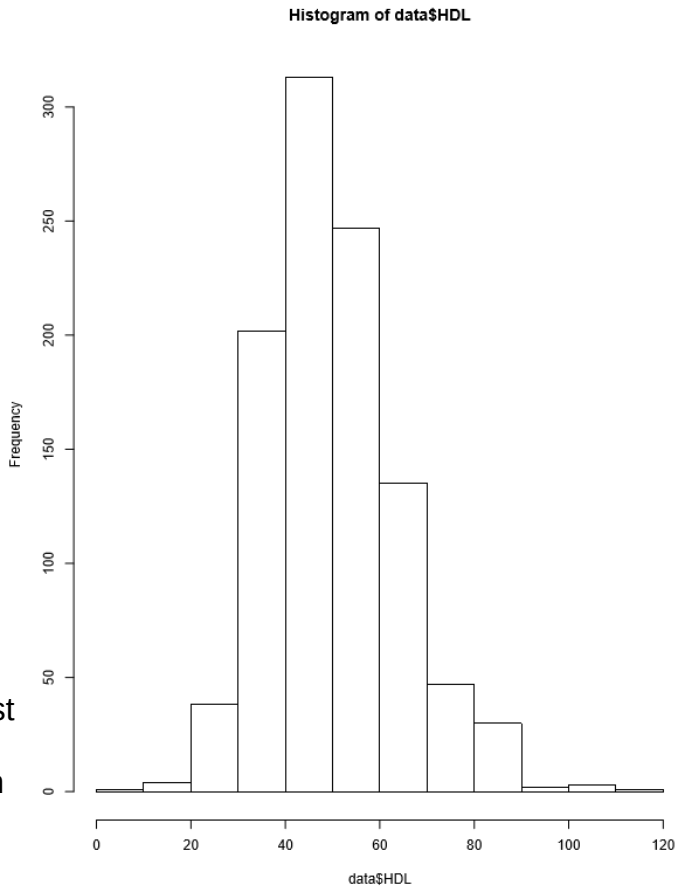


Figure 10: The data had a Shapiro test p-value of $1.052*10^{-11}$, a first quartile value of 40.57, median of 48.48, mean of 50.22, and a third quartile value of 58.24, all measured in mg/dL

	CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	p
1294590	16	kgp11992600	57003980	A	ADD	1023	5.971	5.689	1.666e-08
1294577	16	kgp12245826	56996288	A	ADD	1023	4.016	5.373	9.593e-08
1417374	19	kgp309217	13353401	G	ADD	1018	-3.326	-5.240	1.956e-07
1294567	16	kgp6695323	56991363	A	ADD	1023	3.744	5.231	2.043e-07
1294568	16	rs183130	56991363	A	ADD	1023	3.744	5.231	2.043e-07
1294588	16	kgp685469	57003723	C	ADD	1014	4.580	5.204	2.361e-07
1294573	16	rs4783961	56994894	A	ADD	1023	3.099	5.070	4.729e-07
1417379	19	kgp7054114	13356481	G	ADD	1020	3.233	5.036	5.631e-07
1417367	19	kgp12125483	13345526	A	ADD	1022	3.260	5.029	5.812e-07

Figure 11: Six out of the top nine SNPs were close to each other on chromosome 16, and the other three were close to each other on chromosome 19, leading to belief in high linkage disequilibrium in these regions.

The SNP kgp11992600, located at base pair 57003980 on chromosome 16, reached a p-value of $1.666*10^{-8}$, below the significance threshold of $5*10^{-8}$. Six of the top nine SNPs in the analysis were also from chromosome 16 and close to kgp11992600, and the other three SNPs from the top 9 were in chromosome 19 and also very close to each other, possibly indicating linkage disequilibrium in these regions. This SNP was tagged for further analysis in the future PrediXcan portion of the study.

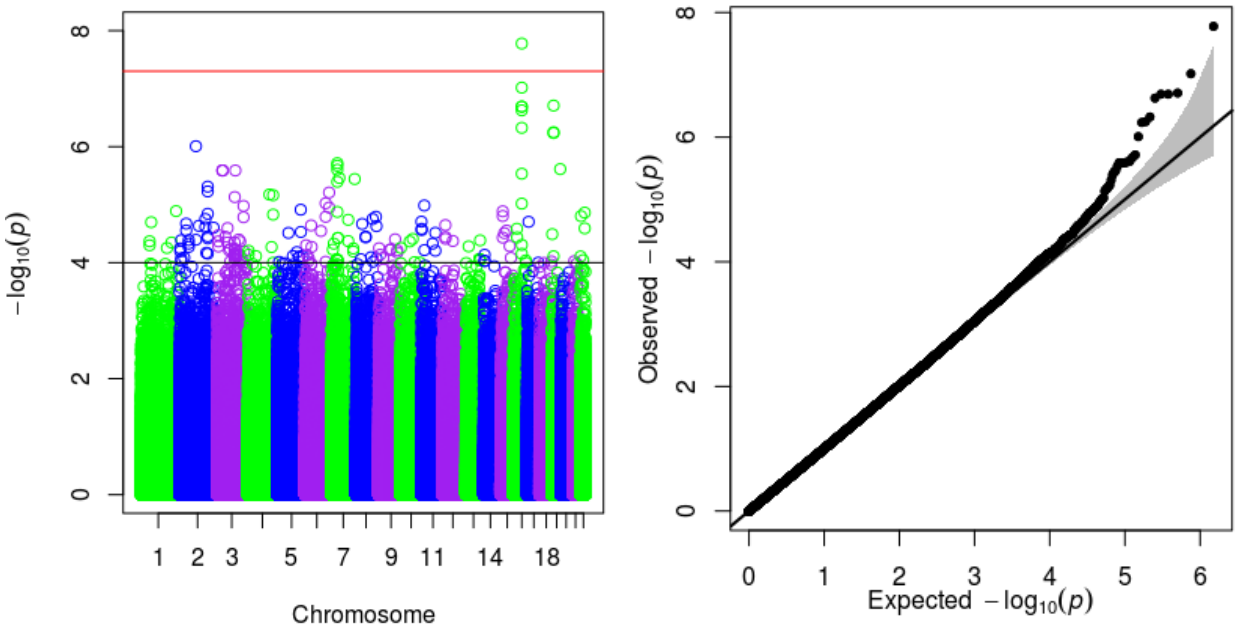
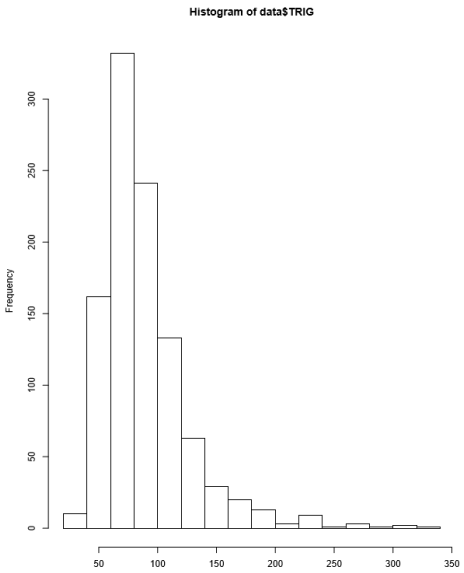


Figure 12: While only one point reached Bonferroni significance, many SNPs near that point had low p-values as well. In the Q-Q plot, many of the related points are deviate from of the 95% confidence interval.

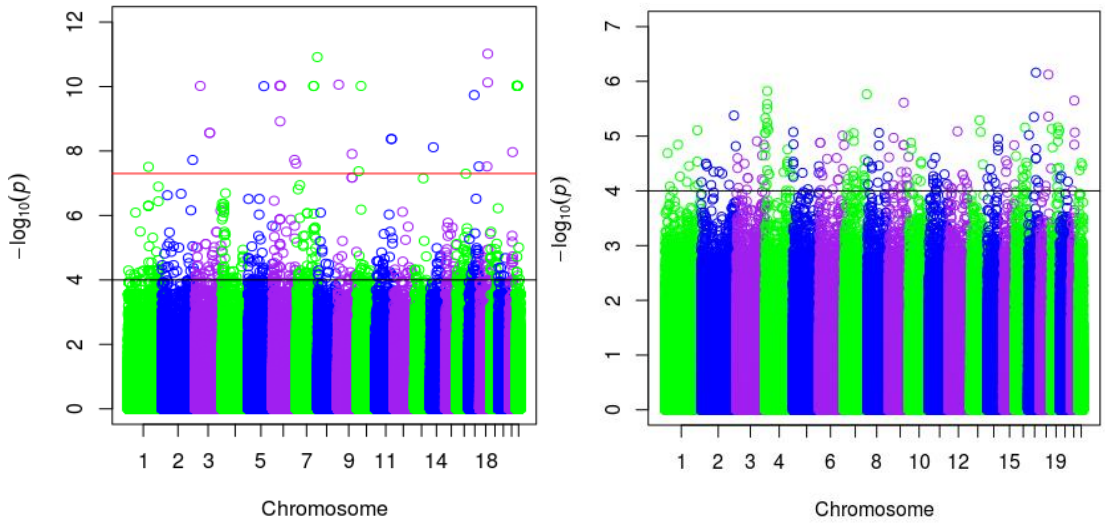
Triglycerides

A type of lipid, high triglyceride levels are linked to an increased risk of cardiovascular issues. In this population, there was a large skew in 1,009 individuals with high triglyceride levels, receiving Shapiro test results of W = 0.82521 and a p-value < $2.2*10^{-16}$. Analyses run on these data received thirty-five points of significance clustered together on chromosomes, indicating abnormal population structure due to some irregular individuals. Therefore, analysis was performed on a logarithmic variation of these data, which had a Shapiro test p-value of $2.727*10^{-11}$, and no SNPs reached significance, in contrast to the highly skewed original data's thirty-five points.



	CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1346504	17	kgp2445215	54300943	A	ADD	1023	-0.09060	-4.996	6.897e-07
1395960	18	rs17609410	58563363	A	ADD	1021	0.88620	4.979	7.485e-07
362991	4	rs7682816	14060248	A	ADD	1023	0.11980	4.840	1.500e-06

Figures 13 & 14: The original data for triglyceride levels in the study population was highly skewed, resulting in a Shapiro test p-value < $2.2*10^{-16}$, with a first quartile value of 64.72, median of 80.82, mean of 89.21, and third quartile value of 102.70, all measured in mg/dL. A disparity between median and mean is common in abnormal populations. In the original, highly skewed analysis, thirty-five highly clustered linked SNPs reached significance, while a logarithmic correction for abnormal population structure revealed no true significant SNPs for triglycerides in the population.



Figures 15 and 16: In contrast to the original SNP results, the logarithmic corrected data did not receive any SNPs with a p-value below the threshold of $5*10^{-8}$ (depicted as a red line).

Low Density Lipoproteins

Known colloquially as “bad cholesterol,” LDLs carry cholesterol to the cells and are the main source of cholesterol buildup in the arteries. Like triglycerides, high levels of LDLs are associated with an increased risk of cardiovascular issues. The original data of 1,009 phenotypes produced Shapiro test values of W = 0.99327, and p = $2.418*10^{-7}$, while the logarithmic variation of the data produced a p-value of $4.54*10^{-7}$, so the original data was used in analysis. Like the high density lipoprotein study, there was one point of significance in the data, kgp7807118.

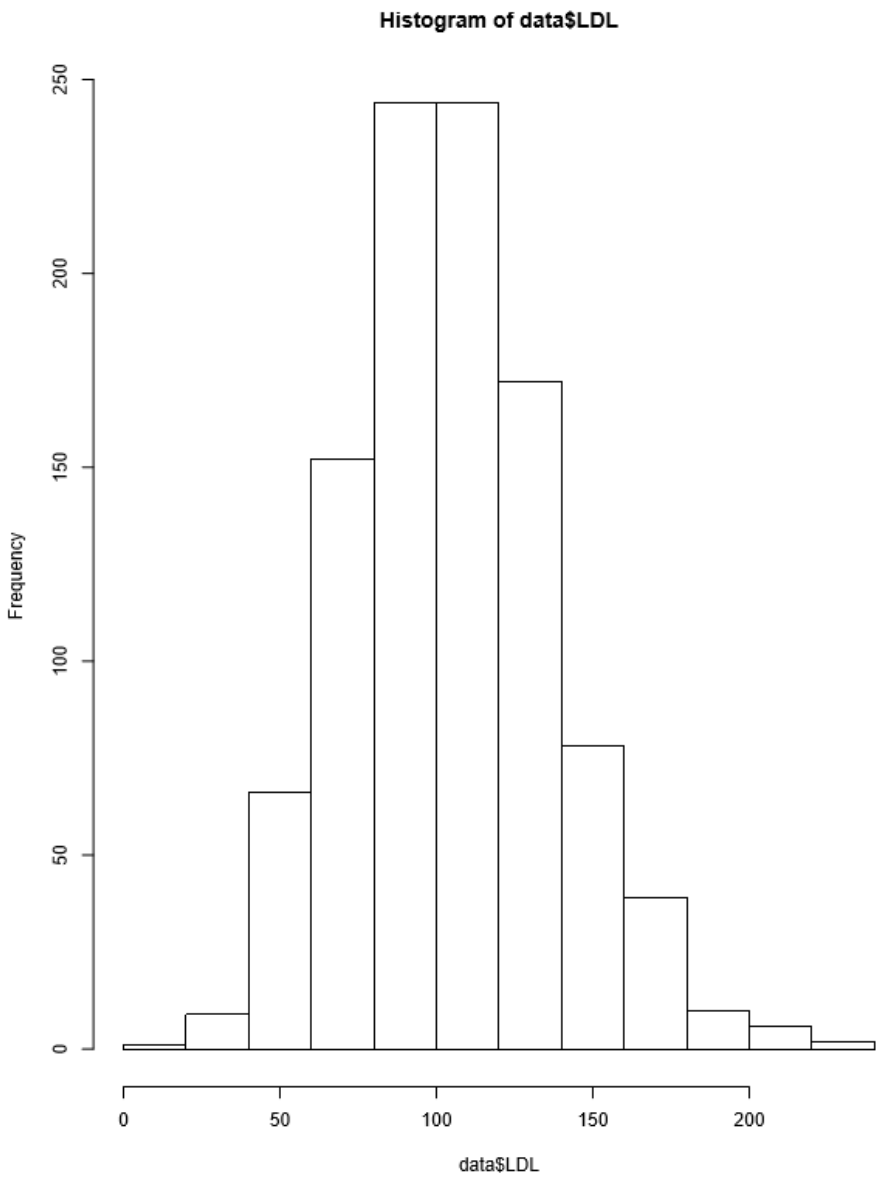
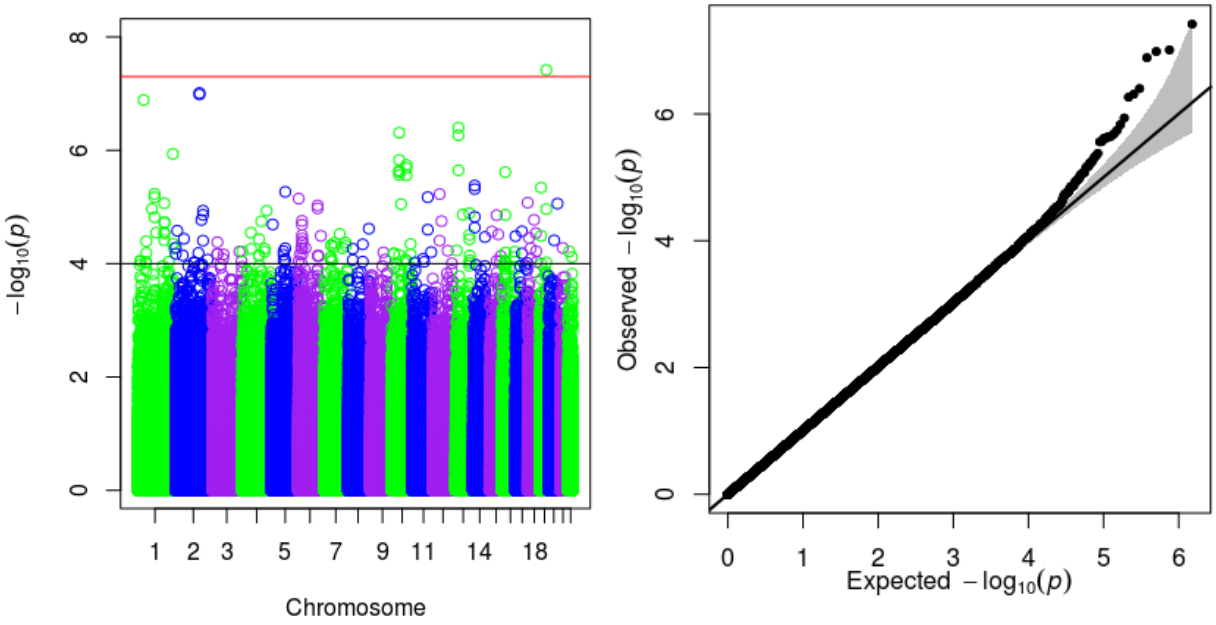


Figure 17: In these LDL data, the first quartile value was 82.84, the median was 102.90, mean 105.20, and third quartile 125.10. All measurements are in mg/dL.

	CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1430863	19	kgp7807118	45413576	A	ADD	1023	-11.130	-5.541	3.828e-08
202351	2	kgp11341957	158333712	A	ADD	1023	15.580	5.370	9.760e-08
202353	2	kgp14742183	158334468	A	ADD	1023	15.600	5.359	1.035e-07

Figure 18: In the study concerning low density lipoproteins, one SNP reached significance at a p-value of $3.828*10^{-8}$, surpassing the significance threshold of $5*10^{-8}$. Unlike the high density lipoprotein study, there were no nearby SNPs to this one.

The SNP of significance kgp7807118 is located on chromosome 16, and is base pair number 45413576. Like kgp11992600, this SNP was tagged for further analysis in the PrediXcan portion of the study, but had no obvious linkage disequilibrium with nearby SNPs.

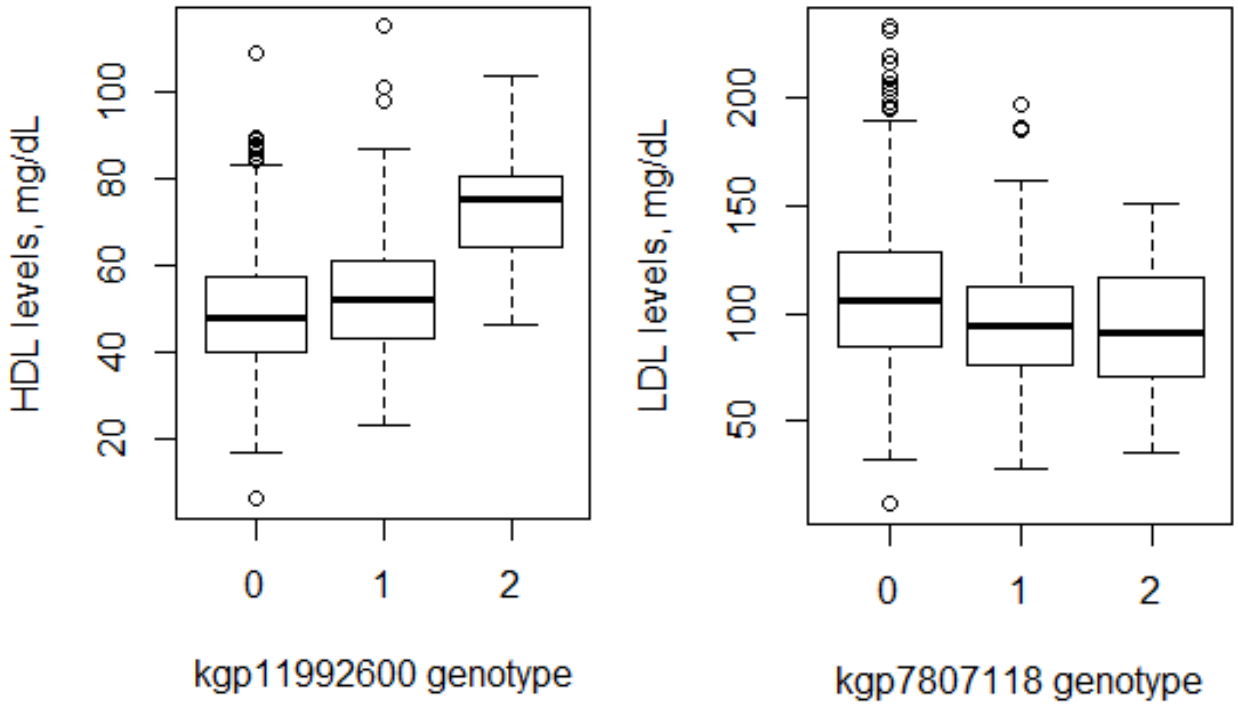


Figures 19 and 20: This analysis only yielded one SNP of significance, kgp7807118. There was no obvious pattern of linkage disequilibrium in the top ten SNPs, and the Q-Q plot reveals the independent nature of the significant SNP.

Discussion

In this study concerning database of Genotypes and Phenotypes data on height, total cholesterol levels, high density lipoproteins, triglycerides, and low density lipoproteins, two single nucleotide polymorphisms out of 1,522,836 reached genome-wide significance in two separate traits. For the high density lipoprotein study, the SNP kgp11992600, minor allele A, base pair 57003980 on chromosome 16, reached a p-value of $1.666*10^{-8}$, and in the low density lipoprotein study, the SNP kgp7807118 base pair 45413576, minor allele A, on chromosome 16 reached a p-value of $3.828*10^{-8}$, both SNPs lower than the significance threshold of $5*10^{-8}$. Homozygosity in the minor allele of kgp11992600 was found to be associated with a large increase in HDL levels, while homozygosity in the minor allele of kgp780118 was found to be associated with a slight decrease in LDL levels, both outcomes having a beneficial link to reducing risks of cardiovascular disease.

After performing these genome-wide association studies, the entire Yoruba population was imputed using the University of Michigan's imputation server and the haplotype reference consortium, mainly composed of a European cohort. Additionally, after fixing the reference alleles, the data was also imputed to the Sanger imputation server, which offers an African reference genome of 4,956 members, which will likely yield more accurate results considering the Yoruba population is based in Nigeria. These imputation results will be converted into a format applicable to LocusZoom by



Figures 21 and 22: In the HDL study, those homozygous for the minor allele A at kgp11992600 have an associated large increase in HDL levels, while those in the LDL study homozygous for the minor allele A at kgp7807118 have an associated slight decrease in LDL levels.

translating the kgp IDs to rs IDs to aid in revealing the genetic association behind these two significant SNPs, especially kgp11992600 as it is indicative of linkage disequilibrium in its locus. Additionally, these data will be applicable to help build PrediXcan's database of plausible genes associated with traits.

However, these results must be taken with caution, especially concerning the small to moderate population size, with the cholesterol studies having a working population of only 1,009 members, and the high degree of relatedness within the population as depicted in figure 1. These studies also all had p-values much lower than 0.05 for their Shapiro test values, indicating an abnormal population structure within the cohort, as well as some skewing, especially notable in the triglyceride study.

Imputation and PrediXcan