Genome-Wide Association Study of Coronary Artery Disease

Adrien Feillard, Sciper: 315921

1 Introduction/Background

Genome-wide association study (GWAS) is a powerful tool for identifying genetic variants associated with complex diseases. In this study, we aim to identify single nucleotide polymorphisms (SNPs) out from 861,473 SNPs associated with coronary artery disease (CAD) with a sample of 1401 patients. CAD is a leading cause of morbidity and mortality worldwide, and identifying the genetic factors that contribute to its risk is essential to understand and treat it.

This report presents the findings of a genome-wide association study (GWAS) aimed at identifying genetic variants associated with coronary artery disease (CAD). We performed quality control, imputation, and population structure analysis, followed by association testing. The results are visualized using Manhattan and Q-Q plots, and we discuss the significance of the results.

The primary variables that were used in this study are:

- **CAD**: Binary outcome indicating the presence (1) or absence (0) of coronary artery disease.
- **SNP Genotypes**: Genotypic data for each individual, coded as 0, 1, or 2, representing the number of minor alleles. Each genotypic data associated with a marker ID and the genetic data of the chromosomes associated with this marker.
- **Covariates**: Age, sex, hdl (HDL cholesterol level), ldl (LDL cholesterol level), tg (Triglycerides level) of the patients and the first five principal components (PCs) to adjust for population stratification.

2 Exploratory Data Analysis

Based on the provided plots, several conclusions can be drawn that highlight important considerations before performing a Genome-Wide Association Study (GWAS).

The distribution of CAD cases and controls (Figure 1) suggests a potential imbalance in the dataset, which could introduce bias in the analysis.

The histograms of SNP and sample missing data (Figures 2 and 3) indicate substantial missing data that need to be addressed, either through imputation or by excluding poorly genotyped SNPs and samples.

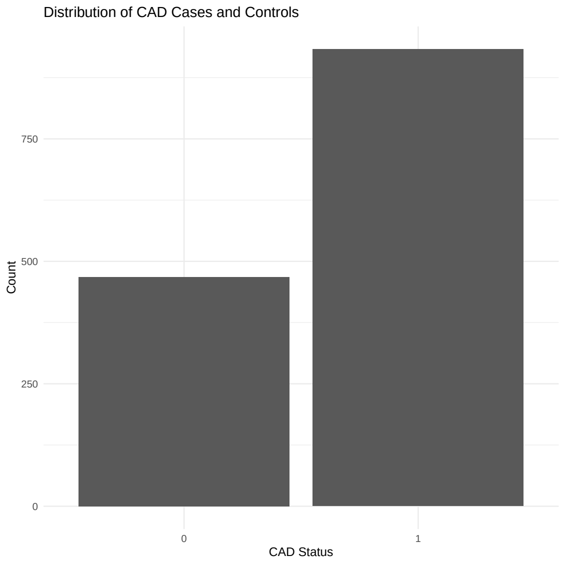

Figure 1: Distribution of CAD cases and controls.

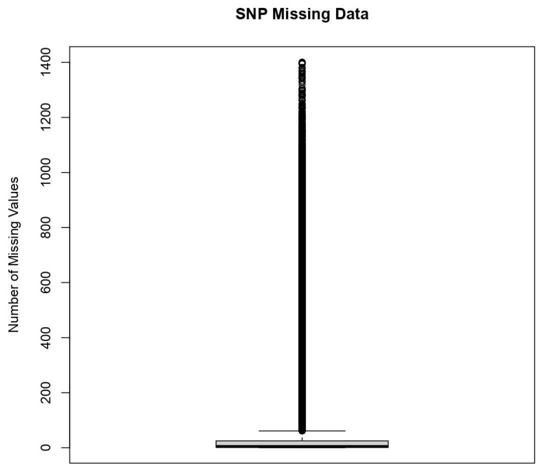

Figure 2: Amount of missing data per SNP.

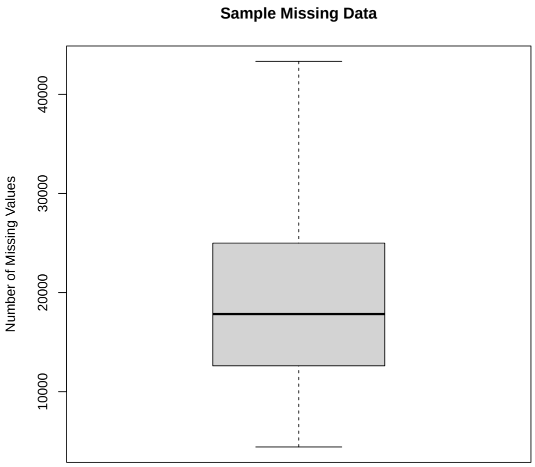

Figure 3: Amount of missing data per sample.

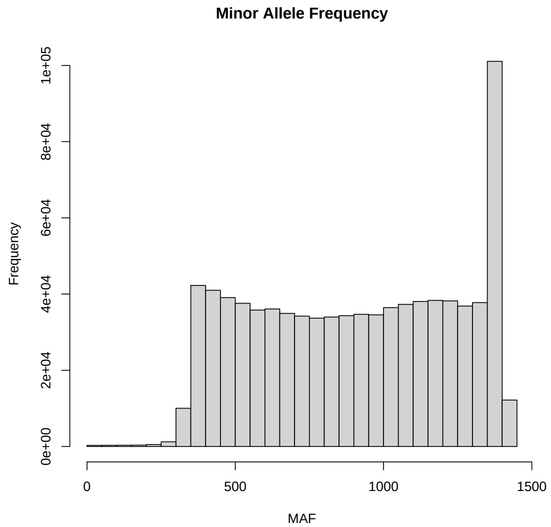

Figure 4: Distribution of minor allele frequencies across all SNPs.

The minor allele frequency (MAF) distribution (Figure 4) shows the frequency of less common alleles. Attention should be paid to SNPs with very low MAF, as they might have reduced statistical power and could potentially lead to unreliable results.

The heterozygosity histogram (Figure 5) illustrates the proportion of samples that are heterozygous at each SNP, providing insights into the genetic diversity of the population. Any deviations from expected heterozygosity rates might indicate population stratification or genotyping errors, which need to be addressed.

The distribution of age (Figure 7) shows the spread of ages among the samples. A well-

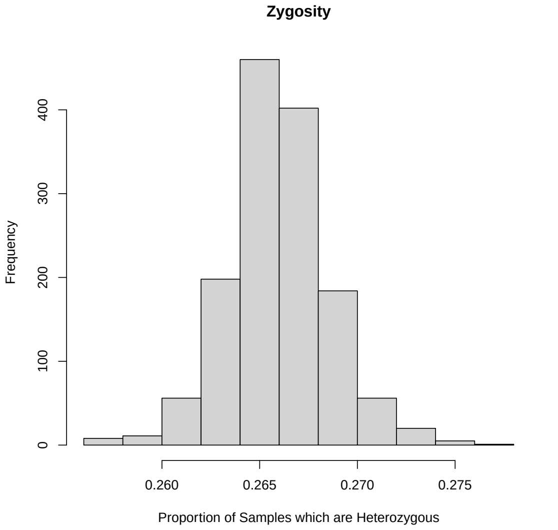

Figure 5: Distribution of heterozygosity rates per sample.

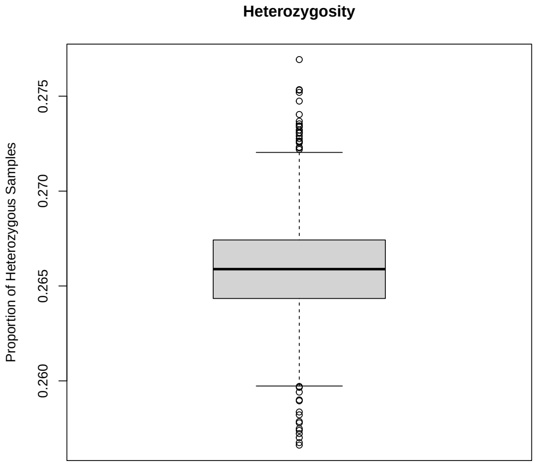

Figure 6: Proportion of heterozygous samples.

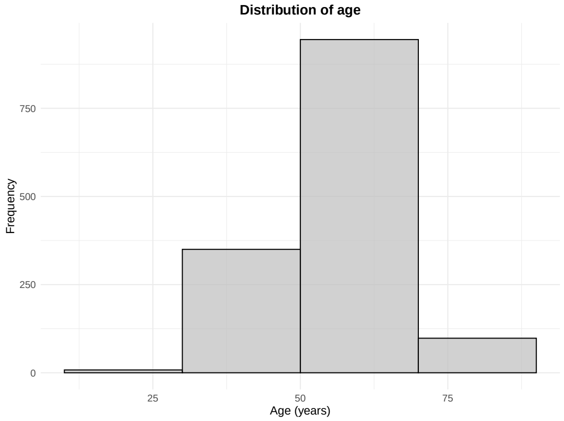

Figure 7: Distribution of age among the samples.

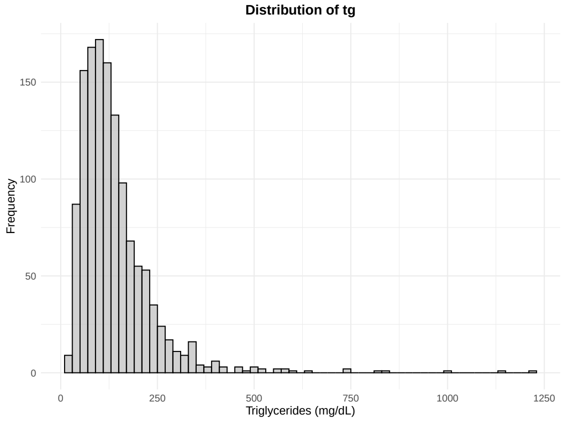

Figure 8: Distribution of triglycerides (TG) levels.

distributed age range ensures that the results are not biased towards a particular age group.

The triglycerides (TG) distribution (Figure 8) illustrates the spread and frequency of TG levels in the sample population. Understanding the distribution of TG levels is important for identifying any outliers or anomalies that could affect the analysis.

The HDL (Figure 9) and LDL (Figure 10) cholesterol distributions provide insights into the range and frequency of cholesterol levels among the participants. These distributions help in identifying any potential outliers or data quality issues that need to be addressed.

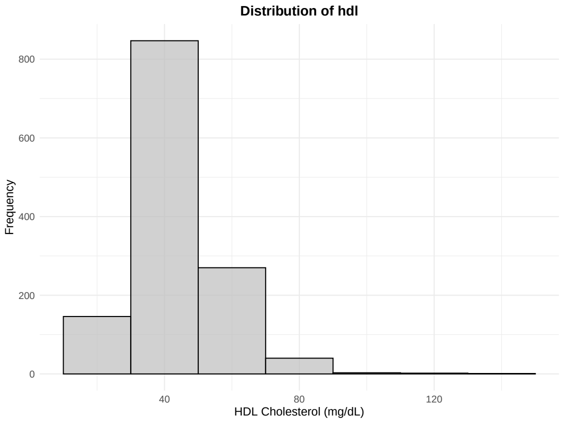

Figure 9: Distribution of HDL cholesterol levels.

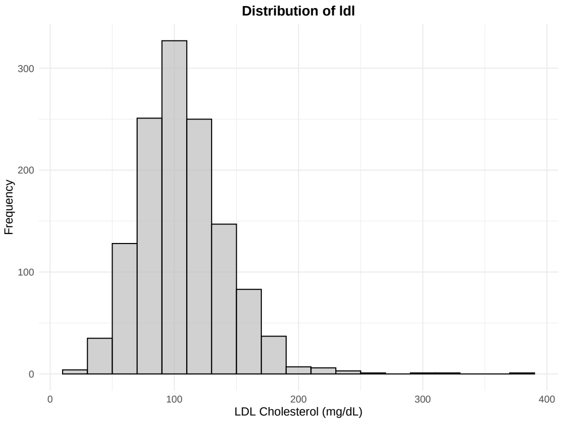

Figure 10: Distribution of LDL cholesterol levels.

3 PCA Analysis

Population stratification refers to differences in allele frequencies between subpopulations due to systematic ancestry differences, potentially confounding GWAS results. Principal Components Analysis (PCA) is widely used to detect and correct for this stratification. By reducing multidimensional genotype data into principal components (PCs), PCA captures major axes of genetic variation. The first few PCs often reflect population structure, clustering individuals from different subpopulations along these axes. Including these PCs as covariates in the GWAS model controls for stratification, reducing false-positive associations from ancestry differences rather than true genetic associations with the trait.

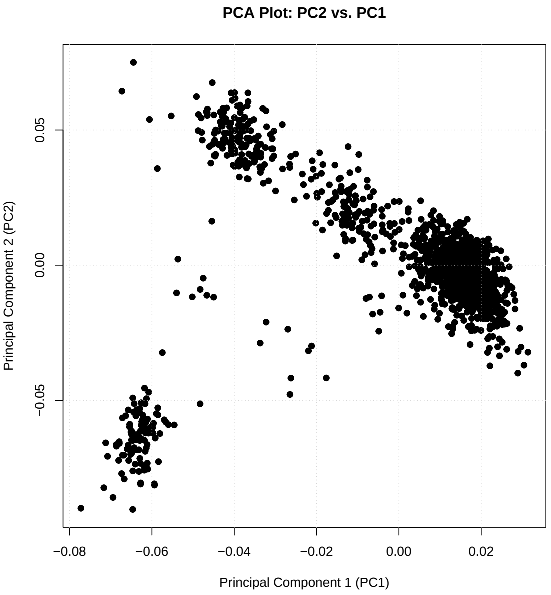

Figure 11: Plot of PC2 vs PC1 showing the population stratification.

The PCA scores plot in Figure 11 provides insights into population structure and genetic variation. It displays the first two principal components (PC1 and PC2), capturing major genetic variation axes. The spread along PC1 and PC2 indicates genetic diversity, and the clustering suggests subpopulations within the dataset. This highlights the need to account for population structure in GWAS to prevent spurious associations. The range of values along PC1 and PC2 suggests these components explain significant genetic variation, highlighting their importance in understanding the study population's genetic architecture. Therefore, including these PCs as covariates in the GWAS model is crucial for ensuring result robustness and validity. Furthermore, the first five PCs show important variance. Hence, in addition to the first two, the 3rd, 4th, and 5th PCs will be included as covariates.

4 Pre-processing and Quality Control (QC) Steps

Pre-processing and quality control (QC) are critical steps in genomic studies to ensure data reliability and validity.

SNP QC involves the following criteria:

- *Missing Data*: SNPs with high missing data rates ($> 10\%$) are imputed with the most common allele combination.
- *Minor Allele Frequency (MAF)*: SNPs with low MAF (< 0.01) are removed to avoid statistical power issues and false positives.
- *Hardy-Weinberg Equilibrium (HWE)*: SNPs significantly deviating from HWE (p-value $<$ 1e-10) are filtered out. HWE indicates that allele and genotype frequencies remain constant without evolutionary influences; deviations may suggest genotyping errors, population stratification, or selection pressure.

Sample QC involves the following criteria:

- *Missing Data*: Samples with excessive missing data ($> 10\%$) are excluded to maintain data integrity.
- *Heterozygosity Rates*: Samples have no big outliers, therefore they don't need imputation for this parameter.
- *Population Structure*: PCA is used to assess population structure. They will be included as covariates.

Overall QC Explanation: QC steps identify and remove unreliable data points, reducing noise and improving downstream analysis accuracy. QC for both SNPs and samples ensures a high-quality dataset free from technical artifacts, suitable for robust genetic association study.

5 Association / Post-association Analysis

Association analysis identifies genetic variants (e.g., SNPs) associated with a trait or disease by comparing variant frequencies in cases (with the trait) and controls (without the trait). The goal is to determine if a specific variant is more common in cases, suggesting a potential link.

This analysis uses a logistic regression model to include multiple predictors and adjust for confounders. The outcome variable Y represents the binary trait status (1 for cases, 0 for controls), with predictor variables including genotype G for the SNP and additional covariates $\mathbf{X}$ to correct for population stratification:

$$\text{logit}(P(Y = 1 \mid G, \mathbf{X})) = \beta_0 + \beta_1 G + \beta_2 X_1 + \beta_3 X_2 + \cdots + \beta_k X_k$$

where $\text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$ is the log-odds of the probability P of the trait given the genotype and covariates. β_0 is the intercept, β_1 the genotype coefficient, and $\beta_2, \beta_3, \ldots, \beta_k$ are the covariate coefficients.

The null hypothesis H_0 states $\beta_1 = 0$ (no association), while the alternative H_A states $\beta_1 \neq 0$ (potential association). Statistical tests (e.g., Wald test, likelihood ratio test, score test) evaluate β_1's significance. A significant result suggests an association between the variant and the trait, offering insights into the genetic basis of the disease.

To control for population stratification, principal components (PCs) from PCA are included as covariates. These PCs capture major genetic variation axes, helping to control for ancestry differences that might confound results. Including PCs ensures observed associations are due to the genetic variant rather than population structure.

In addition, age, sex, HDL cholesterol level, LDL cholesterol level, and triglycerides level are included in the model as they are known risk factors for CAD. These covariates reduce the impact of potential confounders, allowing for the identification of statistically significant genetic associations.

5.1 Mathematical Model

The logistic regression model for a given SNP_i is:

$$\begin{aligned}\log\left(\frac{\hat{P}(\text{CAD} = 1 \mid \text{SNP}_i, \text{Age}, \text{Sex}, \text{TG}, \text{HDL}, \text{LDL}, \text{PC}_{1:5})}{\hat{P}(\text{CAD} = 0 \mid \text{SNP}_i, \text{Age}, \text{Sex}, \text{TG}, \text{HDL}, \text{LDL}, \text{PC}_{1:5})}\right) &= \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{SNP}_i + \hat{\beta}_2 \cdot \text{Age} \\ &+ \hat{\beta}_3 \cdot \text{Sex} + \hat{\beta}_4 \cdot \text{TG} \\ &+ \hat{\beta}_5 \cdot \text{HDL} + \hat{\beta}_6 \cdot \text{LDL} \\ &+ \sum_{j=1}^{5} \hat{\beta}_{6+j} \cdot \text{PC}_j\end{aligned}$$

Where:

- $\hat{P}(CAD = 1)$ is the predicted probability of the CAD trait being present.

- $\hat{P}(CAD = 0)$ is the predicted probability of the CAD trait being absent.
- $\hat{\beta}_0$ is the intercept term.
- $\hat{\beta}_1$ is the estimated coefficient for the SNP of interest.
- $\hat{\beta}_2$ is the estimated coefficient for Age.
- $\hat{\beta}_3$ is the estimated coefficient for Sex.
- $\hat{\beta}_4$ is the estimated coefficient for Triglycerides (TG).
- $\hat{\beta}_5$ is the estimated coefficient for High-Density Lipoprotein (HDL).
- $\hat{\beta}_6$ is the estimated coefficient for Low-Density Lipoprotein (LDL).
- $\hat{\beta}_{6+i}$ (for $i = 1, \ldots, 5$) are the estimated coefficients for the first five principal components (PC$_i$).

5.2 Manhattan Plot

The Manhattan plot in Figure 12 visualizes for each points, the -log$_{10}$ p -values of SNPs across the genome, highlighting regions with significant associations. Two thresholds are highlighted: the green one is the Bonferroni threshold(-log$_{10}(5 \times 10^{-8})$) which is a default significance threshold, and the purple one is a selected threshold of -log$_{10}(5 \times 10^{-6})$ allowing to search for still significant , but less extreme p values associated with SNPs. We can see two SNPs strongly associated with CAD on chromosome 5. Other clusters of p-values above the selected threshold associated with their respective SNPs can be found on chromosomes 4, 5, 9 and 11. These SNPs of interest will be identified later with their specific marker IDs.

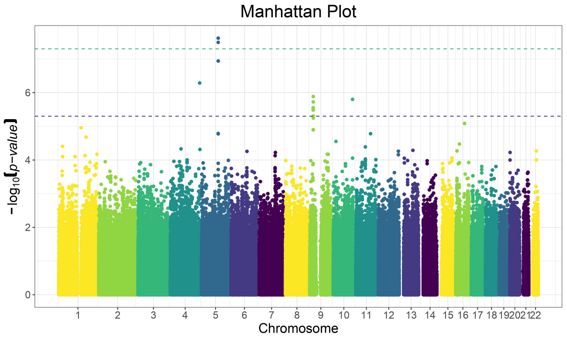

Figure 12: Manhattan plot of the association analysis.

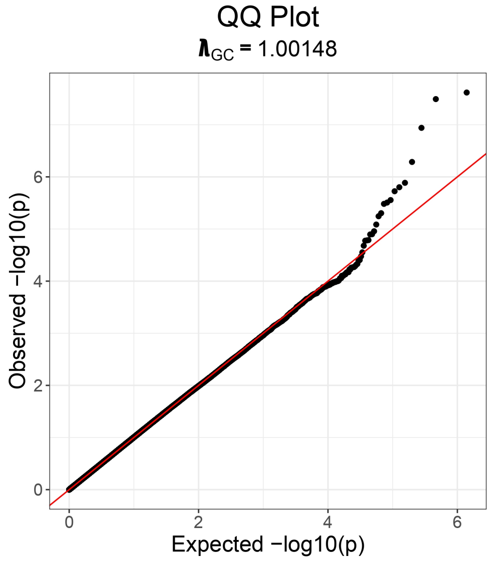

Figure 13: Q-Q plot of the association analysis.

5.3 Lambda Analysis and Q-Q Plot

Lambda (λ) is the genomic inflation factor, indicating the degree of inflation in test statistics due to population stratification or other confounders. A Q-Q plot in Figure 13 compares the observed p-values to the expected p-values under the null hypothesis.

The Q-Q plot provided compares the observed -log10(p-values) from the GWAS to the expected -log10(p-values) under the null hypothesis of no association.

Diagonal red line (Expected Values)

The diagonal line represents the expected distribution of p-values under the null hypothesis. If the observed p-values follow the null hypothesis perfectly, the points will lie on this line.

Inflation Factor (λ_{GC})

The plot includes a genomic inflation factor (λ_{GC}) value of 1.00148.

- $\lambda_{GC} \approx 1$: Indicates little to no inflation, suggesting that population stratification, cryptic relatedness, or other confounding factors are well controlled.
- $\lambda_{GC} > 1.2$: Indicates potential inflation, suggesting possible confounding factors that need to be addressed.

Specific Observations from the Q-Q Plot

- **Points Near the Diagonal**: Most points lie on or near the diagonal line, indicating that the majority of SNPs follow the expected distribution under the null hypothesis.
- **Deviation from the Diagonal**: Some points deviate above the diagonal line, particularly at the tail end (low p-values). This deviation suggests the presence of SNPs that are significantly associated with the trait of interest (CAD).
- **Genomic Inflation Factor (λ_{GC})**: The λ_{GC} value of 1.00148 is very close to 1, indicating minimal inflation. This suggests that the analysis is well-controlled for population stratification and other potential confounders.

The Q-Q plot indicates that the GWAS results are largely consistent with the null hypothesis, with a few SNPs showing significant associations. The minimal genomic inflation factor ($\lambda_{GC} = 1.00677$) suggests that the results are reliable and not significantly affected by population stratification or other confounding factors.

6 Final Estimated Model

Note on the computed values: The function used in the GWAS tutorial was employed to generate the Manhattan plot in Figure 12. However, the computed values in Table 1 were obtained using the glm function. This leads to very slight variations only in the values for SNPs with the smallest p-values. These discrepancies are likely due to the different implementations of the two functions, which approximate values differently. Nonetheless, the differences are minimal and do not alter the overall analysis.

Detailed Table

SNP Marker ID	Estimate	Std Error	Z Value	P Value
rs4957723	0.619	0.112	5.511	3.565e-08
rs17160628	0.612	0.112	5.449	5.060e-08
rs7683009	-0.495	0.098	-5.031	4.877e-07
rs10515379	0.596	0.116	5.148	2.632e-07
rs9632884	-0.482	0.101	-4.796	1.616e-06
rs6475606	-0.459	0.095	-4.850	1.233e-06
rs10757272	-0.437	0.096	-4.566	4.975e-06
rs4977574	-0.451	0.096	-4.703	2.567e-06
rs2891168	-0.446	0.096	-4.659	3.171e-06
rs1333049	-0.460	0.096	-4.807	1.530e-06
rs7088780	0.524	0.112	4.680	2.875e-06

Table 1: Detailed logistic regression results for SNPs of interest.

This table summarizes the information about SNPs that have significant associations with CAD status. The SNPs rs4957723 and rs17160628 correspond to the two points on chromosome 5 in Figure 12 that are above the Bonferroni threshold (p-value $< 5 \times 10^{-8}$). These SNPs have the lowest p-values, indicating the strongest associations with the disease. Their respective estimates show a positive association with CAD.

With another suggestive threshold for p-values $< 5 \times 10^{-6}$, which denotes a less strong but still relevant association, the remaining SNPs of interest, apart from rs10515379 and rs7088780, suggest a negative association with CAD.

Variable	Estimate Range	Std Error Range	Z Value Range	P Value Range
$\hat{\beta}_0$ (Intercept)	[8.781, 9.5]	[0.666, 0.687]	[13.091, 13.837]	[1.002e-42, 9.978e-43]
$\hat{\beta}_2$ (age)	[-0.126, -0.12]	[0.01, 0.01]	[-12.82, -12.368]	[1.272e-37, 7.926e-37]
$\hat{\beta}_3$ (sex)	[-0.711, -0.659]	[0.15, 0.151]	[-4.735, -4.393]	[1.120e-05, 8.461e-06]
$\hat{\beta}_4$ (tg)	[0.004, 0.004]	[0.001, 0.001]	[3.686, 4.015]	[1.096e-04, 9.473e-05]
$\hat{\beta}_5$ (hdl)	[-0.019, -0.016]	[0.006, 0.006]	[-3.131, -2.777]	[1.742e-03, 5.479e-03]
$\hat{\beta}_6$ (ldl)	[-0.001, -0.001]	[0.002, 0.002]	[-0.653, -0.321]	[5.135e-01, 7.485e-01]
$\hat{\beta}_7$ (PC1)	[-0.001, 0.001]	[0.001, 0.001]	[-0.425, 0.993]	[3.208e-01, 8.939e-01]
$\hat{\beta}_8$ (PC2)	[-0.001, 0]	[0.002, 0.002]	[-0.301, -0.022]	[7.633e-01, 9.821e-01]
$\hat{\beta}_9$ (PC3)	[0, 0.001]	[0.002, 0.002]	[0.209, 0.538]	[5.904e-01, 8.341e-01]
$\hat{\beta}_{10}$ (PC4)	[-0.002, -0.001]	[0.002, 0.002]	[-0.941, -0.563]	[3.467e-01, 5.733e-01]
$\hat{\beta}_{11}$ (PC5)	[-0.001, -0.001]	[0.003, 0.003]	[-0.452, -0.256]	[6.511e-01, 7.981e-01]
$\hat{\beta}_1$ (SNP$_i$)	[-0.495, 0.619]	[0.095, 0.116]	[-5.031, 5.511]	[1.233e-06, 5.060e-08]

Table 2: Summary of logistic regression results for SNPs of interest.

Based on these identified SNPs, a logistic regression model was performed on each SNP, resulting in the estimated regression results shown in Table 2. The range of values is consistent among all variable estimates, except for the SNPs of interest, as previously explained. These results provide interesting insights into the association of variables with CAD. At a significance level of p $<$ 0.001, only the intercept, age, sex, TG, and SNPs have statistically significant estimates, meaning the estimates for HDL cholesterol level, LDL cholesterol, and all PCs do not have statistically significant estimates.

7 Conclusion

In this genome-wide association study (GWAS) aimed at identifying genetic variants associated with coronary artery disease (CAD), we conducted an analysis involving several critical steps. QC measures were applied to ensure data integrity, including filtering SNPs based on missing data rates, minor allele frequency (MAF), and Hardy-Weinberg equilibrium (HWE). Similarly, sample QC involved excluding samples with excessive missing data, and assessing population structure using principal components analysis (PCA). The EDA provided insights into the distribution of CAD cases and controls, missing data patterns, MAF distribution, and covariate distributions. These analyses highlighted the need to address imbalances and missing data to prevent biases in downstream analyses. PCA revealed significant population stratification within the dataset, with the first few principal components capturing the major axes of genetic variation. These PCs were included as covariates in the regression models to control for population structure and reduce false-positive associations. Using logistic regression models, we identified several SNPs significantly associated with CAD. The models included genotype data and covariates such as age, sex, HDL cholesterol, LDL cholesterol, triglycerides, and the first five principal components. The Manhattan and Q-Q plots visualized the results, confirming significant associations and minimal genomic inflation, indicating well-controlled confounding factors.

The analysis allowed to determine SNPs such as rs4957723 and rs17160628 with strong positive associations to CAD, exceeding the Bonferroni significance threshold. The low genomic inflation factor (λ_{GC}) indicated that population stratification and other confounders were well controlled. Including covariates like age, sex, and Triglyceride levels in the models significantly improved the robustness of the results, highlighting their importance in CAD risk assessment.

8 References

- Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, 34(28), 3769–3792. doi:10.1002/sim.6605. Available at: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5019244`
- Breheny, P., Reisetter, A., Harris, L., & Peter, T. (2022). *GWAS tutorial: An Introduction*. Available at: `https://pbreheny.github.io/adv-gwas-tutorial/index.html`