Genome-Wide Association Study of Coronary Artery Disease and Cardiovascular Risk Factors

Tao Sun Sciper 279288 tao.sun@epfl.ch

1 Introduction

Genome-wide association (GWA) analysis typically focuses on the associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases. In this project, a GWA analysis on coronary artery disease (CAD) is carried out following the tutorial provided by [1]. The project consists of three parts: data preprocessing, data generation, and GWA analysis, and also includes visualization and discussion.

The aim of this project is to find out which SNPs are statistically significantly associated with CAD and the risk factors for CAD. In this project, high-density lipoprotein (HDL)-cholesterol is chosen as the risk factor.

2 Data Description

This project utilizes the PennCath cohort data which was produced in a previous GWA study of CAD and cardiovascular risk factors at University of Pennsylvania Medical Center [2]. The complete data set has a total of $n = 3850$ individuals that were recruited between July 1998 and March 2003.

In this project, we use information of 1401 individuals with genotype information across 861,473 SNPs. The data are separated in three different files described as following:

.bim: Information on each SNP, including chromosome number, RS number (SNP identifier), genetic distance, the position on the chromosome and the two alleles.

.bed: Every SNP and the genotype at that SNP for each individual. The genotype would be transformed into a enumerate value of 0, 1, 2 representing three possible cases of genotype at one SNP.

.fam: The participant identification information, including family ID, sample ID, paternal ID, maternal ID, sex and phenotype.

Meanwhile, the analysis also uses corresponding clinical data, including CAD status, sex, age, triglyceride (TG), high-density lipoprotein (HDL)-cholesterol and low-density lipoprotein (LDL)-cholesterol. HDL-cholesterol, LDL-cholesterol and TG are all quantitative traits and well-described risk factors for cardiovascular disease.

3 Data Preprocessing

3.1 SNP Level Filtering (Part1)

The original SNP data needs to be cleaned for later analysis since there is a large amount of missing data, low variability and possible genotyping error. The cleaning process begins with filtering out missing data and low variability SNPs, which is followed by the sample level filtering. Finally, the filtering on genotype error is performed. For the first filtering step, some minimum criteria needs to be specified:

Call rate In order to filter missing data, the call rate is used. The call rate of a given SNP is defined as the proportion of individuals in the study for which the corresponding SNP information is note missing. In this study, SNPs with a call rate less than 95% are removed. Also, the NA's need to be removed.

Low variability The variability of SNP can be measured as the minor allele frequency (MAF). At a given SNP, if there is low variability amount participants, then such SNP would not be considered as statistically significantly related with the traits in the study, or can be regarded as irrelevant. Thus, we remove SNPs for which the MAF is less than 1%.

3.2 Sample Level Filtering

At the sample level, individuals who fail the minimum criteria for a valid sample need to be excluded for the analysis. The major concerns in filtering sample data are missing values, sample contamination, correlation, ambiguity or discordance of race, ethnic group and gender.

Call rate The call rate of sample is defined as for one individual, the proportion of genotype data that is not missing. Here, a threshold of 95% is applied to filter. As result, we keep individuals with a certain amount of genotype information.

Heterozygosity A low proportion of heterozygous SNPs within an individual is another cause of poor sample quality, since deficient heterozygosity can indicate inbreeding or other substructure of that person and is also not informative. Thus, we remove those individuals who have more than 10% difference between observed and expected counts of heterozygous SNPs.

Identity by descent The independence across individuals is a basic assumption of association analysis of typed SNPs, which requires that people from the same family cannot be in the data at the same time. Identity by descent (IBD) is commonly used in the measure of duplication between pair of samples, where a kinship coefficient of greater than 0.1 can be regarded as existence of duplicates. The kinship coefficient is a simple measure of relatedness, defined as the probability that a pair of randomly sampled homologous alleles are identical by descent. [3] More simply, it is the probability that an allele selected randomly from an individual, i, and an allele selected at the same autosomal locus from another individual, j, are identical and from the same ancestor.

Before IBD analysis, we firstly reduce the dimension by applying linkage disequilibrium (LD) pruning. LD means a nonrandom association of alleles at two or more loci.

[4] In this project, we utilize the composite measure of LD. In the notation of [5], the population value for for the composite measure at two biallelic loci A and B, with alleles A, a and B, b is

$$D_{AB} = 2 \times P_{AB}^{AB} + 2 \times P_{Ab}^{AB} + 2 \times P_{aB}^{AB} + P_{aB}^{Ab} + P_{ab}^{AB} - 2 \times p_A p_B \quad . \tag{1}$$

In this equation, p_A and p_B are the probabilities of alleles A and B, and the two-locus genotypic probabilities P_{AB}^{AB}, P_{Ab}^{AB}, P_{aB}^{AB}, P_{Ab}^{aB}, P_{AB}^{ab} are written so that the double heterozygotes can be distinguished. [6]

In this step, we apply LD pruning to remove samples with $|D|$ greater than 0.2. Then, pairwise IBD kinship coefficients are calculated. Those with the greatest number of coefficients that are greater than 0.1 are iteratively removed.

Ancestry To identify the problems caused by ancestry, principal components analysis (PCA) is performed to visualize and cluster individuals into ancestry groups based the genotype information. Figure 1 illustrates all observations in a 2-dimension space by the first two principal components. There seem to be 2 groups in PCA plot and some filtering is needed. However, since we know that the population used in the PennCath data is homogenous from European ancestry and the data are pre-filtered, we can say that there are no obvious outliers and just omit this step of filtering.

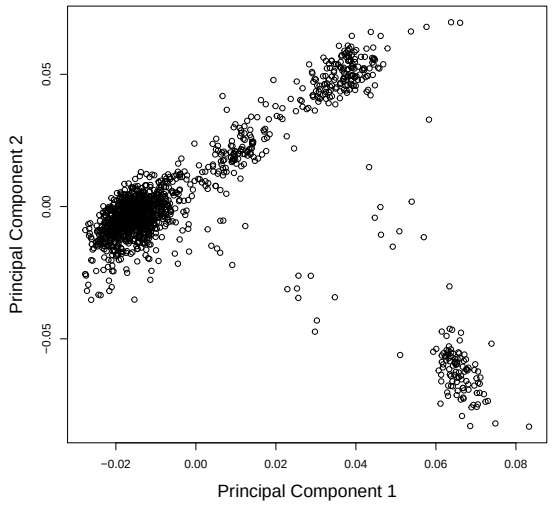


Figure 1: Ancestry plot based on the first two principal components.

3.3 SNP Level Filtering (Part2)

Once we finish filtering the samples, we proceed to filter at the SNP level using Hardy Weinberg Equilibrium (HWE). HWE states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. Thus, violations of HWE can be regarded as the presence of population substructure or the occurrence of a genotyping error. A χ^2 goodness of fit test at a given SNP is performed to test the departures of HWE, the statistic of which is defined as:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad , \tag{2}$$

where O_i is the observed frequency of genotype i and E_i is the expected frequency of genotype i. The null hypothesis behind the test is that there is no difference from HWE between the observed and the expected values. Thus, we filter out those SNPs with a p-value $< 10^{-6}$, since these suggest the rejection of the null hypothesis and a departure from HWE.

3.4 After filtering

We end up with 656,890 SNPs and 1104 individuals for further analysis.

4 Data Generation

For the aim of analysis, two types of data need to be generated. The first are principal components (PCs) for capturing latent population substructure and the second are genotypes of untyped SNPs which may have some functional relationship to the traits in the study.

4.1 Principal Components

The substructure of population refers to the presence of genetic diversity within a homogeneous population, which, on the other hand, can be regarded as subgroup or latent connection between individuals in the population. The principal components are calculated straightforwardly with the help of PCA on the observed genotype capture information. Before PCA, LD pruning as defined in (1) with a threshold of 0.2 is applied. Typically, the first 10 PCs, which together may explain a large amount of variance in the data, are chosen as possible confounders and would be used in the GWA models explained in next section.

4.2 Imputation of Non-Typed SNPs

The analysis of asssociation of genotype of non-typed SNPs with disease outcome is far more important and interesting than that of typed SNPs which vary in at least 1% within the general population. Unmeasured genotype data can be imputed from external resources and we use *1000 Genomes* [7] data as reference here. According to previous study (e.g., [8]), the cholesteryl ester transfer protein (CETP) gene region, which is a well-characterized gene on chromosome 16, has been associated with HDL-cholesterol. Therefore, in this project, we choose chromosome 16 only and impute those SNPs that are in *1000 Genomes* but not in our data.

Imputed genotypes can be reported as the "best guess" genotype or as the posterior probability of each genotype at a given location on the genome [1]. The imputation is based on a recursive regression model fitting. We fit a regression model (3) for genotypes of untyped SNPs associated with near-by typed SNPs:

$$Untyped = \sum_{i \in \mathcal{I}} \alpha_i \times Typed_i + b + \varepsilon \quad , \tag{3}$$

where $\mathcal{I}$ denotes a subset of the nearest 50 typed SNPs around the untyped one, b is the bias and ε is white noise. The regression starts with the nearest SNP and new SNPs are added to the subset in the ascending order of distance until the model is powerful enough.

That is to say, we performe a forward step-wise regression on the 50 nearest SNPs in the present set, stopping each search either when the R^2 for prediction exceeds 0.95, or after including 4 SNPs in the regression, or until R^2 is not improved by at least 0.05. R^2 is a measure of the global fit of the model and is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad , \tag{4}$$

where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares.

After imputing action, those values with a high degree of uncertainty ($R^2 < 0.7$) are filtered out, along with those with low estimated MAF (MAF < 0.01), and the failed imputations.

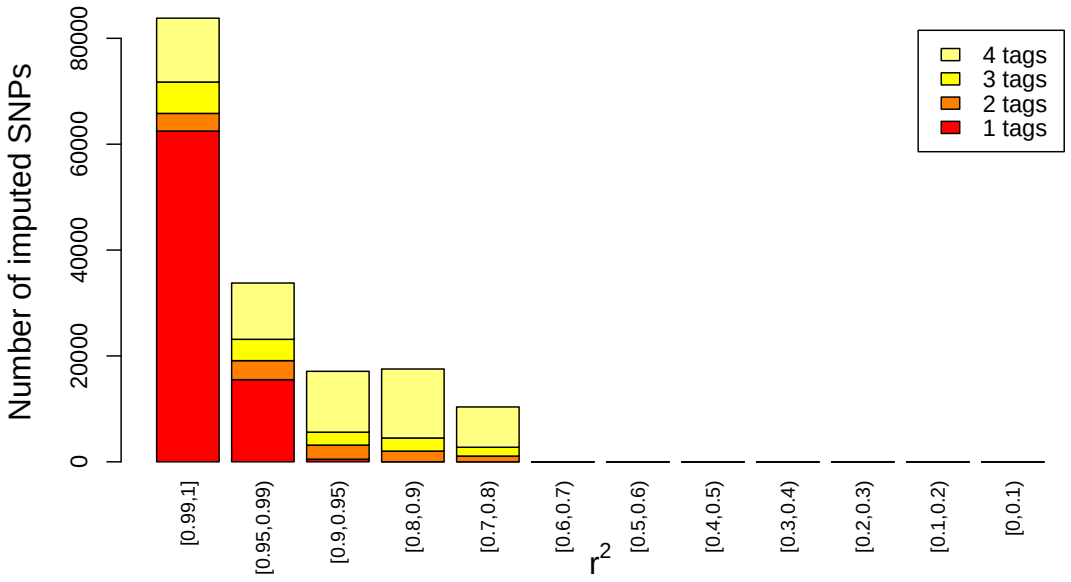


Figure 2: The R^2 distribution for imputation results.

Figure 2 displays the R^2 distribution of imputation results with respect to the number of typed SNPs i.e. tags for each untyped SNP. The majority of imputation takes into account only 1 tag. Generally, the fewer the number of tags involved is, the higher the R^2 of the fitting would be.

5 Genome-Wide Association Analysis

The GWA analysis is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. One basic assumption behind GWA is the traits are normal distributed. In our project, the aim is to identify any statistically significant association between SNPs and CAD. Since HDL-cholesterol level is a typical risk factors of cardiovascular disease, we choose it as the trait in our analysis, or the response. Specifically, the trait is defined as phenotype value which is calculated by a normal rank-transformation of the HDL-cholesterol level. The QQ plots

of HDL-cholesterol level before and after transformation are shown in Figure 3. Before transformation, most of the points lie outside of the boundary; after transformation, most of the points lie on the diagonal, which suggests that we can consider the distribution of transformed HDL-cholesterol level as a normal distribution.

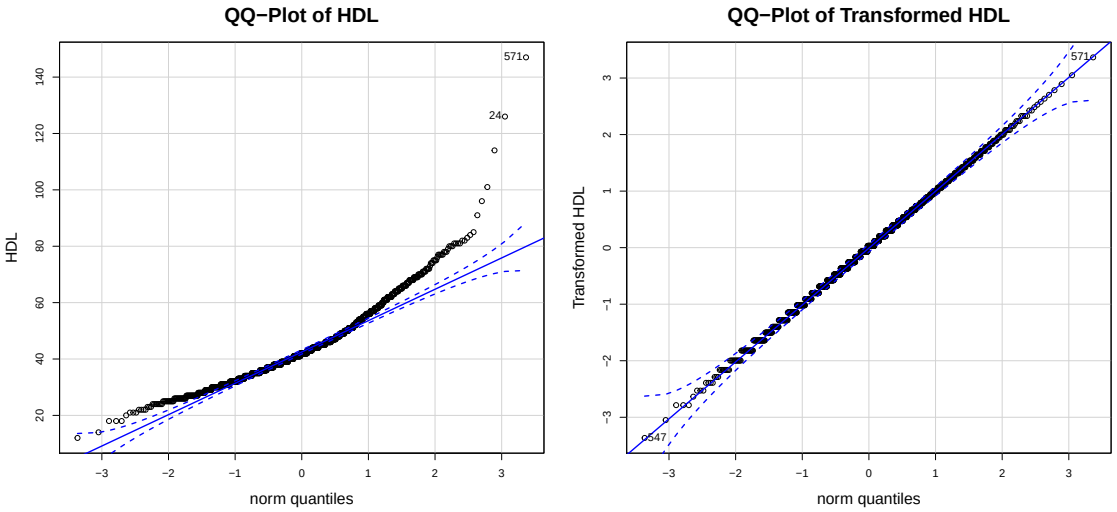


Figure 3: Quantile-quantile plots for HDL before and after normalization

5.1 Association Analysis for Typed Data

The association analysis for typed SNPs is performed using the Gaussian general linear model for SNPs in all 22 chromosomes with the information of samples after filtering. The model is defined in (5). *Phenotype* is the transformed HDL level as described above. Sex is a binary valuable of 0 and 1. *Age* is a simple integer. PC_i is the first 10 principle components that we calculate above. *Genotype* is defined as an enumerate value of 0, 1 and 2, which represents three possible cases of genotype at one SNP. b is the bias and ε is the white noise. The aim is to test for each SNP, whether the genotype is significantly related to the phenotype. Statistically, the null hypothesis is $\mathcal{H}_0 : \theta = 0$ and the alternative hypothesis is $\mathcal{H}_1 : \theta \neq 0$.

$$Phenotype = \alpha \times Sex + \beta \times Age + \sum_{i=1}^{10} \gamma_i \times PC_i + \theta \times Genotype + b + \varepsilon \tag{5}$$

The model is fitted with the help of parallel programming. For each SNP, the statistics for θ is recorded and the 10 most significant typed SNPs are presented in Table 1. The *-logP* column refers to $log_{10}(p.value)$, which is designed for plotting.

5.2 Association Analysis for Imputed Data

The association analysis for imputed SNPs is performed similarly to the general linear model fitting for typed data, where SNP genotypes are treated as independent variables.

The main difference is here the information of the imputation model based on (3) is also required. The statistical significance for one imputed SNP is calculated as a combination effect of significance of those typed ones that are used for imputing it.

SNP	chr	position	Estimate	Std.Error	t.value	p.value	-logP
rs1532625	16	57005301	0.20	0.04	5.39	8.45e-08	7.07
rs247617	16	56990716	0.21	0.04	5.32	1.24e-07	6.91
rs10945761	6	162065367	0.19	0.04	4.54	6.29e-06	5.20
rs3803768	17	80872028	-0.31	0.07	-4.53	6.45e-06	5.19
rs4821708	22	38164106	-0.18	0.04	-4.52	6.83e-06	5.17
rs9647610	6	162066421	0.18	0.04	4.49	7.61e-06	5.12
rs11934535	4	94936015	0.17	0.04	4.44	9.75e-06	5.01
rs2469832	17	38999150	-0.16	0.04	-4.37	1.34e-05	4.87
rs11815950	10	44613217	-0.21	0.05	-4.31	1.75e-05	4.76
rs11682970	2	7477186	0.20	0.05	4.29	1.92e-05	4.72
rs12268604	10	44612433	-0.21	0.05	-4.28	1.98e-05	4.70
rs845966	21	32688703	-0.16	0.04	-4.27	2.06e-05	4.69
rs9591347	13	51257516	0.30	0.07	4.26	2.20e-05	4.66
rs2658621	10	59711428	-0.16	0.04	-4.25	2.30e-05	4.64
rs11831773	12	32310847	-0.26	0.06	-4.24	2.44e-05	4.61

Table 1: Summary of association analysis for 15 most significant typed SNPs.

5.3 Post-Analytic Discussion

5.3.1 Significance

In practice, a Bonferonni-corrected genome-wide significance threshold of 5×10^{-8} is used for control of the family-wise error rate. This threshold is based on research, suggesting approximately one-million independent SNPs across the genome (e.g., [9]), so tends be applied regardless of the actual number of typed or imputed SNPs under investigation. [1]

In our setting, none of the SNPs reach the Bonferroni level of significance. However, two typed SNPs, rs1532625 and rs247617 are suggestive of association ($p < 5 \times 10^{-6}$) with respective pvalues of 8.45×10^{-8} and 1.24×10^{-7}. On the other hand, for untyped SNPs, in total, there are 22 imputed SNPs on chromosome 16 that are significant at a suggestive association threshold of 5×10^{-6}. The information for these 24 SNPs are summarized in Table 2.

The Manhattan plot of GWA significant results is shown in Figure 4. Typed and imputed SNPs are represented by black and blue, respectively. We label the typed SNPs with signals that have surpassed the 5×10^{-6} threshold.

As shown in Table 2 and Figure 4, the 24 SNPs identified as having significant association with the HDL level all locate at chromosome 16 with the imputed ones quite close to the two typed SNPs. The result is reasonable because the two most significant typed SNPs are on chromosome 16 and those untyped SNPs that are imputed by these two are likely to have significance due to the imputation methods. Actually, the SNP with the lowest p-value in both the typed and imputed SNP analysis lies within the boundaries of CETP. Thus, it would be interesting to focus on the single gene, CETP, and explore a little more about it.

	SNP	chr	position	p.value	-logP	type
1	rs1532625	16	57005301	8.45e-08	7.07	typed
2	rs1532624	16	57005479	9.81e-08	7.01	imputed
3	rs7205804	16	57004889	9.81e-08	7.01	imputed
4	rs247617	16	56990716	1.24e-07	6.91	typed
5	rs12446515	16	56987015	1.43e-07	6.84	imputed
6	rs17231506	16	56994528	1.43e-07	6.84	imputed
7	rs173539	16	56988044	1.43e-07	6.84	imputed
8	rs183130	16	56991363	1.43e-07	6.84	imputed
9	rs247616	16	56989590	1.43e-07	6.84	imputed
10	rs3764261	16	56993324	1.43e-07	6.84	imputed
11	rs821840	16	56993886	1.43e-07	6.84	imputed
12	rs56156922	16	56987369	1.43e-07	6.84	imputed
13	rs72786786	16	56985514	7.55e-07	6.12	imputed
14	rs11508026	16	56999328	1.15e-06	5.94	imputed
15	rs12444012	16	57001438	1.15e-06	5.94	imputed
16	rs12720926	16	56998918	1.15e-06	5.94	imputed
17	rs4784741	16	57001216	1.15e-06	5.94	imputed
18	rs34620476	16	56996649	1.16e-06	5.94	imputed
19	rs708272	16	56996288	1.16e-06	5.94	imputed
20	rs711752	16	56996211	1.16e-06	5.94	imputed
21	rs12720922	16	57000885	3.24e-06	5.49	imputed
22	rs8045855	16	57000696	3.24e-06	5.49	imputed
23	rs12149545	16	56993161	3.25e-06	5.49	imputed
24	rs56228609	16	56987765	3.25e-06	5.49	imputed

Table 2: Summary of association analysis for the 24 most significant SNPs.

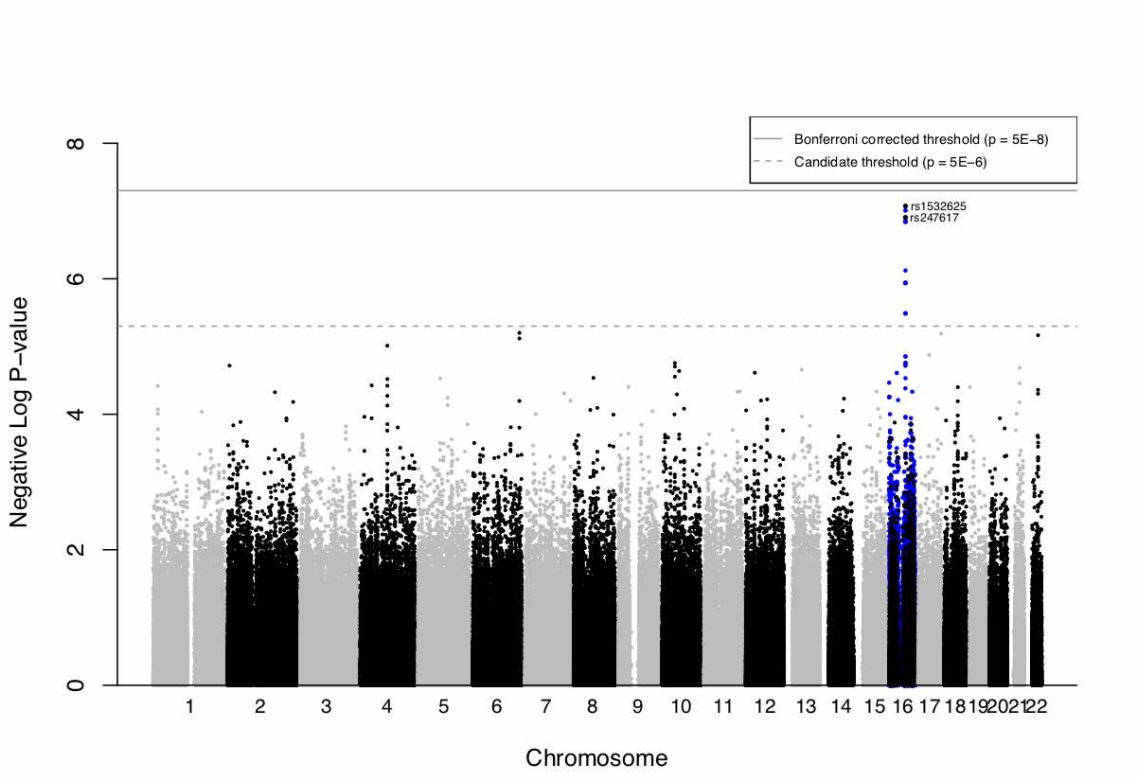


Figure 4: Manhattan plot of genome-wide association analysis results.

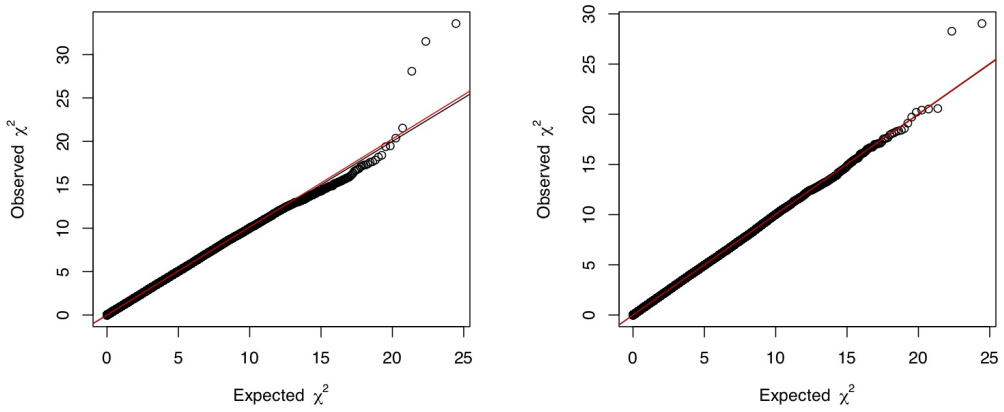
5.3.2 Quality control

Before doing further analysis on CETP, we test whether our results are influenced by confounding. To do so, we fit a new model that does not account for covariates (PCs, age and sex), which we call an unadjusted model opposite to previous adjusted model. Quantile-quantile (Q-Q) plots are used to visualize the relationship between the expected and observed distributions of SNP-level test statistics, as illustrated in Figure 5(a) and (b). Actually, it is the squared t-test statistics which follow the χ^2 distribution that are plotted. We can see from the plot that the tail of the distribution is closer to the $y = x$ line after accounting for potential confounding by population substructure in the model. In Figure 5(b), there is also a slight deviation in the upper right tail which suggests the existence of some kind of dependence.

Meanwhile, the λ-statistic can be used to present the degree of deviation from the $y = x$ line formally. By definition, λ is the the median of the observed chi-squared test statistics divided by the expected median of the chi-squared distribution:

$$\lambda = \frac{Median_{Observed}}{Median_{Expected}} \quad . \tag{6}$$

A value of λ closer to one suggests low degree of deviation. By adding confounders into model, λ is improved from 1.0142 to 1.0032, because of which we can conclude that the fitting results are significantly influenced by confounding.



(a) Unadjusted model; $\lambda = 1.0142$. (b) Adjusted for PCs, age and sex; $\lambda = 1.0032$.

Figure 5: Quantilequantile plots for quality control check.

5.3.3 LD pattern in CETP

In this section, we visualize LD patterns between the most significant SNP from our analysis (rs1532625) and SNPs in nearby regions, CETP, by the heatmap in Figure 6. At the top of the heatmap, one scatter plot shows the negative log p-values of typed SNPs (black) and imputed SNPs (red) with respect to their position at the gene. In the inverted pyramid, the degree of shading indicates the amount of LD so that darker squares indicates higher LD. In this pyramid, LD is measured by R^2, which is transformation of the scalar D, defined as the difference between the joint probability of the two major alleles and the product of the two marginal probabilities, where the adjustment is based on allele frequencies. At two biallelic loci A and B, with alleles A, a and B, b, D and R^2 can be mathematically expressed as following:

$$D = P_{AB} - p_A p_B \quad , \tag{7}$$

$$R^2 = \frac{D}{p_A p_B p_a p_b} \quad . \tag{8}$$

P_{AB} is the frequency of the genotype AB and p_A, p_B, p_a, p_b are the occurrence frequency of A, B, a, b respectively. If A and B are completely independent, D and R^2 equal 0. To be noted, the definition of D is sightly different from (1).

If two biallelic loci have strong LD and thus, are dependent to some degree, then these two loci are more likely to be inherited by descendants together after crossover. Thus, the combined effect of these two related loci are worth further study. In Figure 6, we observe the presence of two distinct LD blocks within the CETP gene region, with high levels of LD between SNPs within each block and lower LD between SNPs across the two blocks. The result suggests that within each block in CETP, gene are more likely to be passed on to the next generation together even with crossing over.

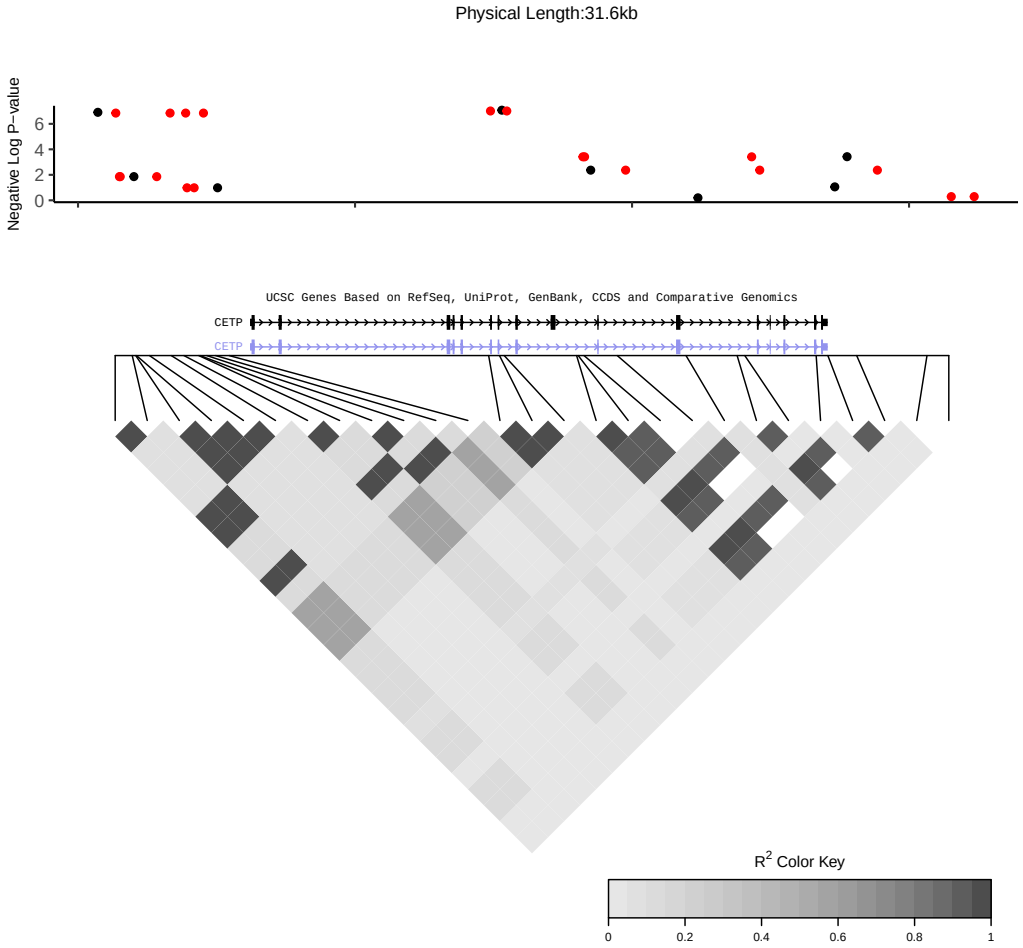


Figure 6: Heatmap of linkage disequilibrium.

6 Conclusion

In this project, we carry out a standard GWA analysis on CAD and cardiovascular risk factors. Data preprocessing is performed step-by-step to remove data with little prediction power. Principal components of typed SNPs are selected out for association analysis. Imputation of untyped SNPs is implemented with *1000 Genomes* dataset and we use typed SNPs nearby to predict one untyped SNP.

The result of GWA on typed and imputed data suggests that typed SNPs, rs1532625 and rs247617, along with other 22 untyped SNPs within the CETP gene region at chromosome 16 are statically significantly associated with the HDL level (p-value$< 5 \times 10^{-6}$), a commonly believed cardiovascular risk factor. The genotypes of these 24 SNPs can be considered as the predictor for CAD and thus, can be used for diagnosing potential patients.

On the other hand, 2 distinct LD blocks within the CETP gene region are detected. The combined effect of gene within these two blocks need further study to have a better understanding of the genetic effect for CAD.

References

[1] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes, "A guide to genome-wide association analysis and post-analytic interrogation," *Statistics in medicine*, vol. 34, no. 28, pp. 3769–3792, 2015.

[2] M. P. Reilly, M. Li, J. He, J. F. Ferguson, I. M. Stylianou, N. N. Mehta, M. S. Burnett, J. M. Devaney, C. W. Knouff, J. R. Thompson *et al.*, "Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies," *The Lancet*, vol. 377, no. 9763, pp. 383–392, 2011.

[3] D. Schaid, "Mathematical and statistical methods for genetic analysis," *Publications of the American Statistical Association*, vol. 100, no. 470, pp. 712–712, 2003.

[4] M. Slatkin, "Linkage disequilibrium–understanding the evolutionary past and mapping the medical future." *Nature Reviews Genetics*, vol. 9, no. 6, pp. 477–485, 2008.

[5] B. S. Weir, *Genetic Data Analysis II*. Sunderland, Massachusetts: Sinauer Associates, Inc., 1996.

[6] D. C. Hamilton, Q. Liu, and D. E. Cole, "Approximate variance for a standardized composite measure of linkage disequilibrium," *Annals of Human Genetics*, vol. 70, no. 4, pp. 535–540, 2006.

[7] G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. Mcvean, "A map of human genome variation from population-scale sequencing." *Nature*, vol. 467, no. 7319, pp. 1061–73, 2010.

[8] A. C. Edmondson, P. S. Braund, I. M. Stylianou, A. V. Khera, C. P. Nelson, M. L. Wolfe, S. L. Derohannessian, B. J. Keating, L. Qu, and J. He, "Dense genotyping of candidate gene loci identifies variants associated with high-density lipoprotein cholesterol," *Circulation Cardiovascular Genetics*, vol. 4, no. 2, p. 145, 2011.

[9] M. X. Li, J. M. Y. Yeung, S. S. Cherny, and P. C. Sham, "Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets," *Human Genetics*, vol. 131, no. 5, pp. 747–756, 2012.