Genome-Wide Analysis of Cardiovascular Risk Factors

Zoé Weissbaum

July 2024

1 Introduction

In the following, we will be performing a genome-wide association study (GWAS), taking after the tutorial [1]. A GWAS is an observational study that aims to detect if any variant is associated with a trait, typically a disease. It is usually done on large sets of data and attempts to look through the entire human genome to uncover potential association.

In this tutorial, we test association for each single nucleotide polymorphism (SNP) independently. We focus on common variants, meaning that they are present in at least 1% of the population.

The data we are using is the PennCATH cohort data on 1401 individuals with genotype information across 861'473 SNPs, arising from a GWAS of coronary artery disease and cardiovascular risk factors based at University of Pennsylvania Medical Center. The data includes sex, age, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, triglycerides and coronary artery disease status. Cholesterol and triglycerides are well-known cardiovascular disease risk factors. Cardiovascular diseases are one of the leading causes of death worldwide, which makes understanding and identifying associated genetic variants particularly valuable. We are thus looking for association between certain variants and those risk factors, especially in the cholesteryl ester transfer protein (CETP) gene region. Such variants in the CETP region could be responsible for coronary artery disease.

The data is already de-identified and filtered.

We will conduct our analysis using R and the packages and methods specific to GWAS. The R files and detailed methods are available in [2]. This is a complement to the tutorial [1] and both have the same structure.

This study can be sectioned into 4 main parts, each divided into smaller steps, according to the structure of the tutorial. We outline the procedure in the overview below.

i) Data pre-processing

1. Reading data into R to create an R object
2. SNP-level filtering (part 1)
3. Sample-level filtering
4. SNP-level filtering (part 2)

ii) New data generation

5. Principal component analysis (PCA)
6. Imputation of non-typed genotypes

iii) Statistical analysis

7. Association analysis of typed SNPs
8. Association analysis of imputed data

iv) Post-analytic interrogation

9. Integration of imputed and typed SNP results
10. Visualization and quality control of association findings

2 Data pre-processing

The first step is to load the data and prepare the files which we will use to save our results. The data contains genotype information on 861'473 SNPs and 1401 individuals.

We can then start the data pre-processing. This step is essential as it allows us to clean the data, so that we are only left with complete, usable data. The first filtering step is to remove SNPs that have a lot of missing data, a low variability or that involve genotyping errors. We first get rid of SNPs for which there is more than 5% missing data, so we filter with a call rate of 95%. The low variability is measured in terms of the minor allele frequency: if the minor allele frequency is less than 1%, the SNP is removed. We do this because low variability implies low statistical power. We thus removed 203'287 SNPs and are left with 658'186 SNPs.

The next step is at the sample-level. We filter out individuals with missing data, poor sample quality or correlation in the population. As for SNPs, we remove individuals with more than 5% missing data on the SNPs.

We then remove individuals with too high or too low heterozygosity, according to the Hardy-Weinberg equilibrium (HWE). HWE requires 5 assumptions: random mating, infinite population size, no mutations, no genetic migration and no natural selection. Then if all these assumptions are satisfied, the alleles are distributed as follows: for a two allele locus with alleles A and a, where $p(A) = p$ and $p(a) = 1-p = q$, the genotypes AA, Aa and aa occur in proportions p^2, $2pq$ and q^2 respectively. If the observed proportions do not match the expectations, one of the assumptions is violated, for example in the case of population stratification or inbreeding. In this instance we have non-random mating and an increase in homozygosity. Heterozygos-

ity is the proportion of genotypes Aa, which is expected to occur in proportion $2pq$. We then compute an inbreeding coefficient $|F| = 1 - O/2pq$, where O is the observed proportion of heterozygous SNPs in a given individual. If $|F| > 0.1$, the individual is discarded.

Next, we check unrelatedness and independence across the population. We compute a pairwise identity by descent kinship coefficient, and if it is larger than 0.1, we remove the individual in the pair with the lowest call rate. We also prune based on linkage disequilibrium with a coefficient of 0.2. Linkage disequilibrium is the non-random association of alleles at two (or more) loci. For two loci A, B, each with two alleles A, a, B, b, and under independence, we must have $p_{AB} = p_A p_B$ where p_{AB} denotes the occurrence of genotype AB (and similarly with other genotypes and all other combinations). If the difference between p_{AB} and $p_A p_B$ is too big, there is linkage disequilibrium.

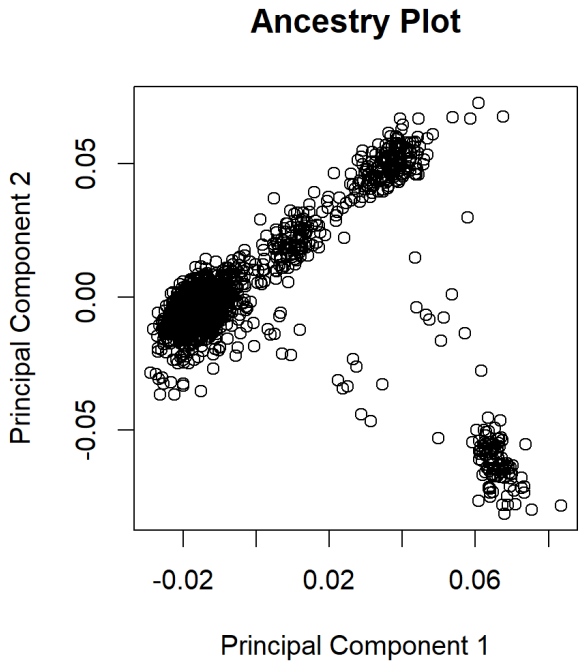

Figure 1: PCA plot of ancestry

We then check for ancestry by determining the first two principal components of the genotype data and inspecting the PCA plot. The principal components allow us to classify individuals into ancestry groups based on their genetic data. We do this as self-reported race and ethnicity might not align perfectly with genetically defined clusters. Additionally, identifying individuals who do not fit within a specific genetic cluster can be indicative of sample-level errors. The plot with the two first principal components can be seen on Figure 1. Since we do not observe any particular

deviations or clear outliers, no samples are removed at this step. This is expected as the data we are using is a pre-filtered, homogeneous sample.

After this, we go back to SNP-level to investigate for genotyping errors. We do this after the sample-level filtering because it is also based on Hardy-Weinberg equilibrium. In a general setting, we could go back and forth between both levels until no additional subjects are removed. In the present case, it is not necessary as none were eliminated at sample-level. This time, we measure deviation from Hardy-Weinberg equilibrium with a χ^2 goodness-of-fit test, and discard any SNP with a p-value less than 10^{-6}. Here, we filter out an additional 1296 SNPs and are hence left with 656'890 SNPs.

All of these quality controls are necessary to ensure that our data follows a certain number of assumptions (such as the assumptions of HWE and independence) which we require in order to accurately perform our association analysis.

3 Principal components and imputed data

We now have clean data and, before actually performing our analysis, we generate new data that we will use together with the genotype data in the association analysis: principal components and imputed SNPs.

We first want to define the principal components that account for population substructure and stratification. It is important to adjust for population substructure as in can be a confounder and can cause spurious association. We prune again based on a linkage disequilibrium coefficient of 0.2 and then compute the first 10 principal components. This is an arbitrary but typical number. An other option would be to define a certain amount of variability that the principal components explain and then keep the number of principal components that account for this variability. We will later use the principal components as covariates in the regression models for the association analysis.

Next, we use external resources and references to impute SNP data that could not be measured. We use the 1000 Genomes data to impute SNPs on chromosome 16, as an example. In a general setting, we would typically impute data on all chromosomes. We note that the imputed SNPs are separated from the measured ones in the association analysis. After imputation, we need to do a quality control and remove SNPs that could not be imputed, those with high uncertainty and those with low minor allele frequency. We end up with 162'565 imputed SNPs that we can use in our association analysis.

4 Association analysis

Now that our data is loaded and filtered, and that we imputed additional SNP genotypes, we can finally perform our genome-wide association analysis.

As mentioned above, we analyse separately the typed and imputed SNPs. We begin with the association analysis of the typed (measured) SNPs. We use an additive model, so the genotype at each SNP is represented as the number of minor alleles (0, 1 or 2). We perform a regression on each single SNP with age, sex (1 for male, 2 for female) and the 10 principal components as the covariates.

We use inverse normally transformed high-density lipoprotein (HDL) cholesterol as the response variable, because we require a normally distributed variable, otherwise our model would not be appropriate. Visual inspection of a histogram of HDL cholesterol revealed some extreme values so we selected an inverse normal transformation. A plot comparing the two distributions can be viewed in Figure 2.

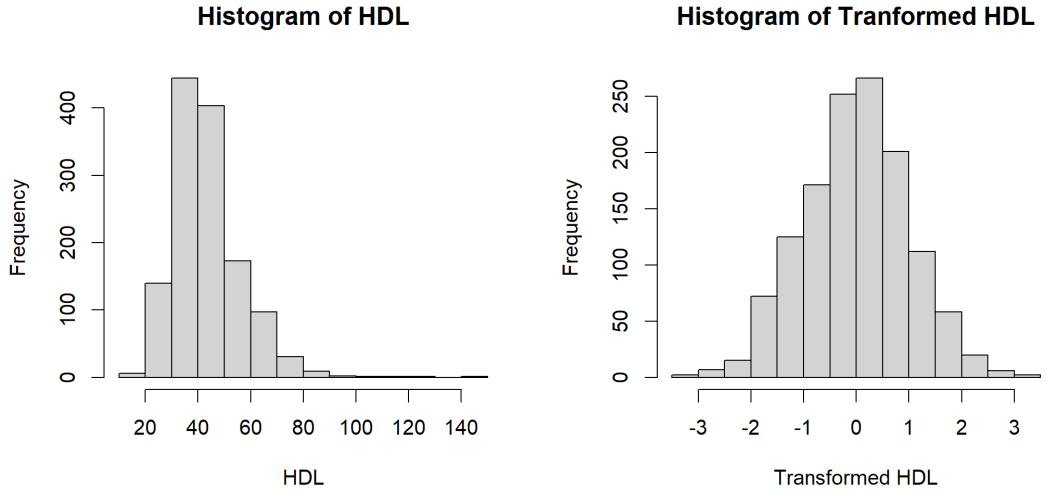

Figure 2: Histograms of HDL cholesterol and inverse normally transformed HDL cholesterol

The final estimated model for a given SNP is:

$$\widehat{transfHDL} = \hat{\beta}_0 + \hat{\beta}_{age} \cdot age + \hat{\beta}_{sex} \cdot sex + \sum_{i=1}^{10} \hat{\beta}_i \cdot PC_i + \hat{\beta}_{SNP} \cdot SNP\# + \epsilon.$$

Association is tested for using a Cochran-Armitage trend test. We run the model fitting in parallel due to the large number of SNPs. As is typical, a Bonferroni correction is applied, with a genome-wide significance threshold of $5 \cdot 10^{-8}$. We find potential association (p-value $< 5 \cdot 10^{-6}$) for two SNPs in the cholesteryl ester transfer protein (CETP) gene region. The SNPs rs1532625 and rs247617 have p-values $8.94 \cdot 10^{-8}$ and $1.52 \cdot 10^{-7}$ respectively.

We then proceed similarly with the imputed SNPs. We use the same model with sex, age and the 10 principal components, and test for association with a Cochran-Armitage trend test. We also map the associated SNPs to the corresponding gene regions. We end up finding 16 SNPs in the CETP region that are significant (p-value $< 5 \cdot 10^{-6}$). A complete table of results for these SNPs can be viewed in the Appendix.

5 Data integration and quality control

After concluding our analysis, we want to present and visualize our results. First, we want to put all our data together so that it is easier to find the relevant information. We link each SNP to its corresponding locus, chromosome and base pair location. We can combine the data from the typed and imputed SNPs to have all the results in one place. It is hence easy to visualize results globally or only on typed or imputed data.

In the last step, we use several methods and plots to visualize the results. These will also help us check and control for quality and consistency in our findings. Mainly, we will use two visual tools: Manhattan plots and Q-Q plots.

Manhattan plots allow to display genome-wide association significance level by chromosome. Each dot represents one SNP and the higher the dot, the more significant the SNP (we plot the -$\log_{10}$ of the p-value, so large values correspond to small p-values). The x-axis corresponds to the chromosomal location of each SNP. The Manhattan plot of our analysis can be seen on Figure 3. Imputed SNPs are represented in blue. We observe that no SNP reached the Bonferroni corrected significance threshold (p-value $< 5 \cdot 10^{-8}$), but two typed and 22 imputed SNPs reached the more flexible suggestive association threshold (p-value $< 5 \cdot 10^{-6}$).

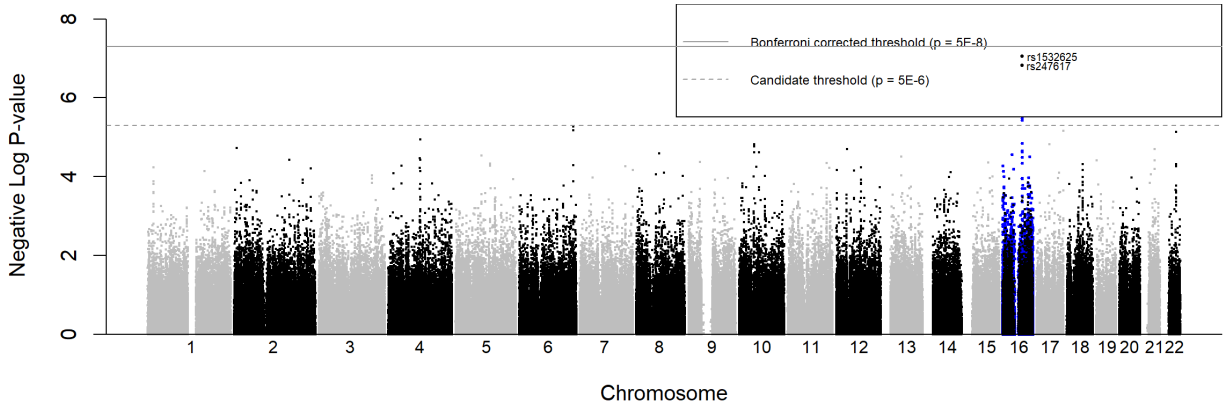

Figure 3: Manhattan plot

We then use quantile-quantile plots to compare expected and observed distributions of SNP-level test statistics. We provide two plots on Figure 4, one with the

unadjusted model and one where the model has been adjusted for the principal components. We observe a slight improvement (the tail of the distribution is closer to the line) after adjusting for confounders and population substructure, but in both cases there is no significant deviation. The few observed statistics that are far from the line are suggestive of association.

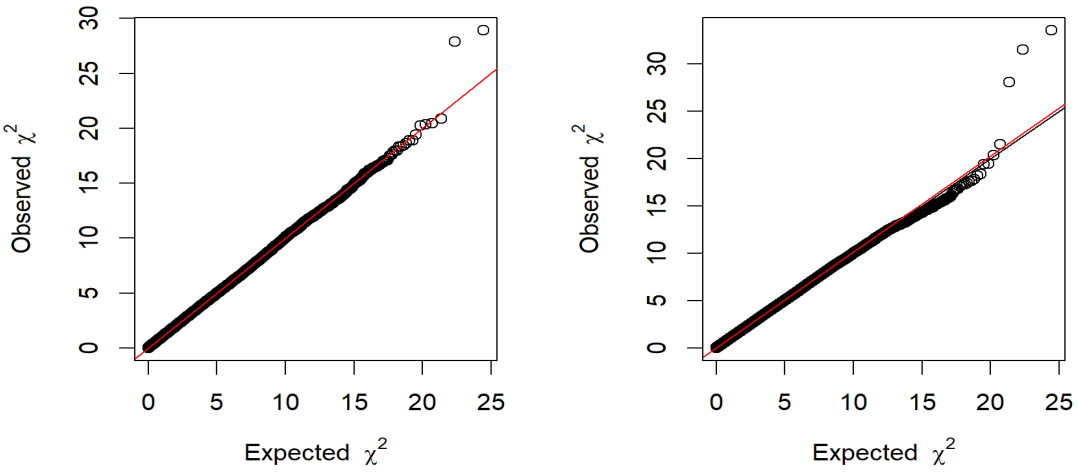

Figure 4: Q-Q plots of adjusted model (left) and unadjusted model (right)

The deviation from the expected distribution is measured with the λ-statistic. It has been established that with population substructure, the distribution of Cochran-Armitage trend tests is inflated by a constant multiplicative factor λ. This factor can be estimated with a χ^2 distribution: $\lambda = \text{median}(\chi^2)/0.456$. A value of $\lambda > 1.2$ suggests population substructure. A value closer to 1 means better adjustment for potential stratification. We obtain the following values:

- Unadjusted λ: 1.0141
- Standardized unadjusted λ: 1.0108
- Adjusted λ: 1.0005
- Standardized adjusted λ: 1.0004

We do get values closer to 1 after adjusting for confounders, but the gain is very limited as the data consists of a rather homogeneous population. We can conclude that there is no clear bias in the data, and that our chosen model is satisfactory.

6 Conclusions

In this comprehensive genome-wide association study, we identified numerous genetic variants significantly associated with various cardiovascular risk factors.

Following the tutorial, [1], we were able to discover and present the steps and methods involved in a genome-wide association study. We prepared and cleaned the data by filtering both at SNP-level and at sample-level, generated principal components to account for population substructure and imputed non-typed SNPs. With the newly computed principal components, we wrote out a linear model for the response variable, the inverse normally transformed high-density lipoprotein cholesterol. We performed a regression for each SNP. We then tested for association using a Cochran-Armitage trend test, once on the measured data and once on the imputed data.

We found suggestive association for two typed SNPs and 16 imputed SNPs in the CETP region, at the significance threshold of $p < 5 \cdot 10^{-6}$. The cholesteryl ester transfer protein is linked with high-density lipoprotein cholesterol, which is a risk factor for cardiovascular diseases such as coronary artery disease.

Finally, we visualized and inspected our results using several tools. The Manhattan plot summarized all results in one graph, displaying significance levels of all SNPs and classified under chromosomal location. The Q-Q plots allowed us to confirm normal distribution of the data and to conclude that our results were consistent with our assumptions.

Throughout this analysis, we experienced no major hiccups as we worked on prefiltered, homogeneous data. However, we still worked through the key steps of a GWAS, and [1] provides some more general information that can be used in a broader context.

7 Bibliography

[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. 769–3792, 2015. DOI: https://doi.org/10.1002/sim.6605.

[2] A. Alhendi. "Genome-wide association study (gwas) tutorial." (), [Online]. Available: https://github.com/AAlhendi1707/GWAS.

A Appendix

Table 1: Table of results for the CETP region (chromosome 16), typed and imputed SNPs with significant p-value (p-value $< 5 \cdot 10^{-6}$)

SNP	**p-value**	**-log(p)**	**chr**	**position**	**type**
rs1532625	$8.94 \cdot 10^{-8}$	7.0484	16	57005301	typed
rs247617	$1.52 \cdot 10^{-7}$	6.8169	16	56990716	typed
rs1532624	$1.04 \cdot 10^{-7}$	6.9845	16	57005479	imputed
rs7205804	$1.04 \cdot 10^{-7}$	6.9845	16	57004889	imputed
rs17231506	$1.75 \cdot 10^{-7}$	6.7579	16	56994528	imputed
rs183130	$1.75 \cdot 10^{-7}$	6.7579	16	56991363	imputed
rs3764261	$1.75 \cdot 10^{-7}$	6.7579	16	56993324	imputed
rs821840	$1.75 \cdot 10^{-7}$	6.7579	16	56993886	imputed
rs11508026	$1.17 \cdot 10^{-6}$	5.9300	16	56999328	imputed
rs12444012	$1.17 \cdot 10^{-6}$	5.9300	16	57001438	imputed
rs12720926	$1.17 \cdot 10^{-6}$	5.9300	16	56998918	imputed
rs4784741	$1.17 \cdot 10^{-6}$	5.9300	16	57001216	imputed
rs34620476	$1.18 \cdot 10^{-6}$	5.9289	16	56996649	imputed
rs708272	$1.18 \cdot 10^{-6}$	5.9289	16	56996288	imputed
rs711752	$1.18 \cdot 10^{-6}$	5.9289	16	56996211	imputed
rs12720922	$3.39 \cdot 10^{-6}$	5.4695	16	57000885	imputed
rs8045855	$3.39 \cdot 10^{-6}$	5.4695	16	57000696	imputed
rs12149545	$3.69 \cdot 10^{-6}$	5.4327	16	56993161	imputed