Genome-wide association study on coronary artery disease

Fahim Beck

1 Introduction and background

Coronary artery disease (CAD) is one of the major causes of death worldwide. It causes a reduction in blood flow in the arteries of the heart through plaque formation (arteriosclerosis). There are many risk factors (smoking, alcohol, high blood cholesterol, obesity, etc). However, the heritability of this disease has been estimated between 40% and 60% (McPherson and Tybjaerg-Hansen, 2016). Hence the interest in carrying out a genome-wide association study.

1.1 Research questions and approaches

The aim of this study is to identify associations among SNPs and the presence of CAD. For a GWAS, the four main steps are: (i) data pre-processing; (ii) new data generation; (iii) statistical analysis; and (iv) post-analytic interrogation.

1.2 Dataset

The dataset comes from a GWA study of CAD (PennCATH) of the University of Pennsylvania Medical Center. It includes 3850 individuals enrolled between July 1998 and March 2003. A part of them of European ancestry were selected for genome-wide genotyping. Here, we consider anonymised data that includes 1401 individuals with genotype information over 861,473 SNPs. The clinical data gives information about the age, sex, HDL and LDL cholesterol, triglycerides and CAD status.

2 GWA analysis

2.1 Data pre-processing / QC steps

Before starting the GWA analysis, it is necessary to proceed to some quality-control steps in order to remove (or ”filter”) poor quality data. Filtering will be done at two levels: at SNP and sample levels. Excluding SNPs that have been found to be of insufficient quality can be justified by several reasons, such as the large quantity of missing data, low variability, or genotyping errors. As for the exclusion of individuals, this may be due to missing data, sample contamination, correlation or other issues related to race, ethnicity or gender.

SNP-level filtering – part 1

Here, the data are filtered according to two criteria: the call rate and the minor allele frequency (MAF).

The call rate is defined for a given SNP as the proportion of individuals for which the corresponding SNP information is not missing. In our case, we filter with a call rate of 95%. All SNPs that have more than 5% missing data will be removed.

As for the second criterion, we remove all SNPs for which the MAF is less than 1%. Indeed, if a large majority of the participants have two copies of the major allele, then there will be little variability and the power of our statistical tests will be reduced, i.e. it will be more difficult to infer a significant relationship between the SNP and the trait studied.

After this step, we are left with 658,186 SNPs. Indeed, 203,287 SNPs were removed.

Sample-level filtering

Similar to what was done at the SNP level on the call rate criterion, we remove the individuals whose percentage of missing SNPs is more than 5% (call rate is 95%).

Then, a second criterion concerns heterozygosity. Under Hardy-Weinberg equilibrium (HWE), the probability of observing both alleles at a given SNP is $2p(1-p)$ where p is the frequency of the dominant allele. We would therefore like to remove individuals with an inbreeding coefficient $|F| = (1 - O/E) > 0.1$, where O and E are respectively the observed and expected numbers of heterozygous SNPs.

Another problem concerns the relationship between individuals. In such studies, it is possible to unintentionally recruit two or more individuals from the same family. One measure to quantify relatedness is the identity by descent (IBD). An IBD kinship coefficient of more than 0.1 can reinforce the hypothesis that the pair in question is related. Before proceeding to this analysis, linkage disequilibrium (LD) pruning is applied with a threshold of 0.2. This will eliminate duplicates and other types of redundancy. It will also result in a lot of computational savings. This filtering reduces considerably the SNPs from 658,186 to 72,812. Then, the IBD analysis indicates that no pair suggests a family relationship.
With the same filtering, we consider a last aspect which is about ancestry. Here, we would like to remove individuals who are not part of an ethnic/racial group. For this, PCA helps us visualising the different ancestry groups according to the genetic information. The 72,812 SNPs are given as input for the PCA. Figure 1 shows a plot of the first principal component against the second one.

Visual inspection of this plot does not require removing more individuals. It should be noted that the PennCATH data has also been pre-filtered.

SNP-level filtering – part 2 (HWE)

A final filtering at the SNP level consists of removing SNPs that violate HWE. These may indicate a genotyping error. To do this, SNPs whose HWE test statistic (obtained by a χ^2 test between the observed and expected genotypes) has a p-value lower than 1×10^{-6} in

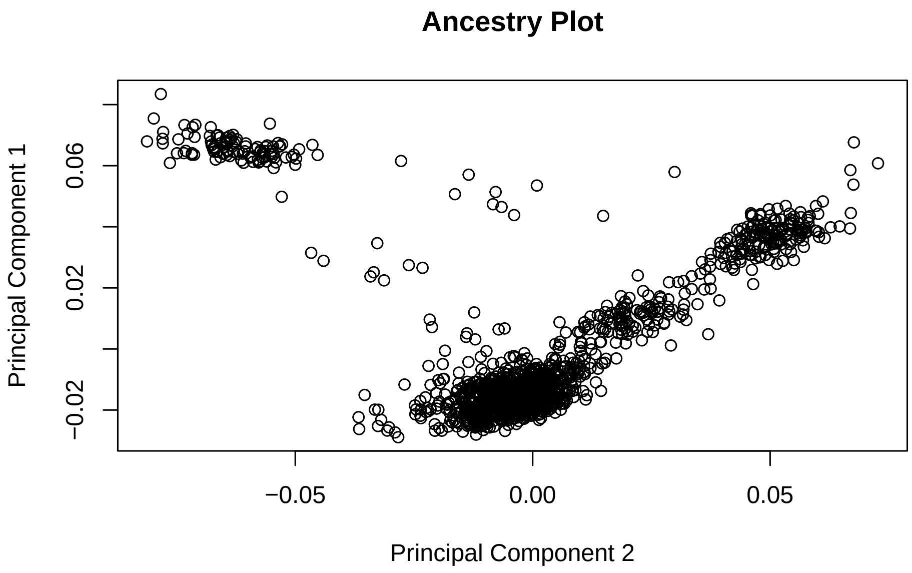

Figure 1: Plot of PC1 against PC2 for representing individuals by ancestry groups based on their genetic composition.

CAD controls are removed. This corresponds to 1,296 SNPs. This leaves 656,890 SNPs for association analysis.

2.2 Association / post-association analysis

3 Conclusion

4 References

[1] R. McPherson and A. Tybjaerg-Hansen. Genetics of coronary artery disease. *Circulation Research*, 118(4):564–578, February 2016. doi: 10.1161/CIRCRESAHA.115.306566.

[2] E. Reed, S. Nunez, D. Kulp, J. Qian, M. Reilly, and A. Foulkes. A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in medicine*, 34, September 2015. doi: 10.1002/sim.6605.