AI Hack 2018: Project Report GWAS Challenge

Team 25: Lean Mean Gene Machine

November 18, 2018

Abstract

Genome Wide Association Studies (GWAS) aim to understand the effect of Single Nucleotide Polymorphisms (SNPs) in causing disease. Associations can be made using statistical techniques such as PCA and regression analysis to draw conclusions about risk exacerbating genes for a specific clinical outcome. In this paper we explore SNPs that play a part in increasing cardiovascular disease risk in addition to other routine biochemical blood tests.

1 Introduction

Genome Wide Association Studies (GWAS) aim to understand the effect of Single Nucleotide Polymorphisms (SNPs) in causing disease. SNPs are characterised by a single base pair change that usually occurs as a result of errors during DNA replication [1]. Although they are small chance errors independently, they can have major additive effects downstream in biological processes. By studying the association between genetic variants and the trait (in this case a clinical outcome), one hopes to better understand the underlying pathological processes to inform areas of future research.

2 Data preprocessing

Data is sourced from a study of coronary artery disease (CAD), containing 1,401 study individuals with genotype information across 861,473 single nucleotide polymorphisms (SNPs). The individuals were surveyed between July 1998 and March 2003, where the case-control study was based on European ancestry severe angiographic CAD status. CSV files of data consisted of SNP matrices for all 22 chromosomes and 1,401 patients; Clinical matrix for all 1,401 patients.

SNPs with more than 5% of their data missing were removed to minimise imbalance and cause a detriment in the validity of subsequent statistical analysis. This was carried out by finding the allele genotype distribution for each SNP, and filtering. Further, SNPs with an allele distribution not obeying the Hardy-Weinberg distribution was tagged for error-analysis. This was done by comparing the proportions of the dominant (p) and recessive (q) alleles and seeing that they obey the simple relation $p + q = 1$. If they didnt fall within a 5% level of significance of 1, it was omitted [2][3].

Data Imputation has been shown to increase power of the study and can be implemented by softwares such as PLINK. Linkage disequilibrium (LD) deals with the non random association of alleles at different sites, because sometimes it is possible to make

predictions based on mapping these relations. To this end, we create a visual heatmap of the pairwise correlation between variables. The goal of GWAS is to use these indirect associations to inform disease susceptibility. Much of genome can be split into haplotype blocks which are closely linked alleles that tend to be inherited over evolutionary time. Hence, by finding groups of of SNPs exhibiting linkage disequilibrium one can infer a risk association.

3 Principal components analysis

Principal component analysis (PCA) is a widely-used tool in statistical genetics used to infer obscure population structure from genome-wide data such as SNPs, and/or to facilitate further analyses such as genome-wide association studies [2016]. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Hence, by using only the first few principal components, PCA maps the high-dimensional data spaces of the SNP matrices to a lower-dimensional picture that contains a large amount of information of the sample variability.

To execute the PCA on the SNP matrix we need a data array in the form of a "correlation matrix". We can't simply use the given chromosome matrices due to the categorical nature of the element values in the SNP matrices 0,1,2. For each chromosome (15, 16, 17), we have a $M \times N$ genotype matrix $\mathbf{X}$, where M is the number fo SNPs and N is the number of individuals in the study. Each matrix entry x_{ij} takes a value 0,1,2 which represents the count of variant alleles for an individual at SNP i. From here, we construct the $M \times N$ normalised genomic matrix $\mathbf{Y}$ where each row $\mathbf{y_i}$ has approximately mean 0 and variance 1 for SNPs in the Hardy-Weinberg equilibrium [4][5].

$$\mu_i = \frac{\sum_{j=1}^{N} x_{ij}}{N_i} \tag{1}$$

$$p_i = \frac{1 + \sum_{j=1}^{N} x_{ij}}{2 + 2N_i} \tag{2}$$

$$y_{ij} = \frac{x_{ij} - \mu_i}{\sqrt{p_i(1 - p_i)}} \tag{3}$$

where p_i is the sample allele frequency for SNP i. Then, the genomic relationship matrix (GRM) ψ is given by:

$$\psi = \mathbf{Y}^T\mathbf{Y}/M \tag{4}$$

The leading 10 principal components of $\mathbf{Y}$ are obtained by performing the eigendecomposition of the GRM ψ. This was implemented in Python with functionalities from the `sklearn.decomposition.PCA` libary.

4 Regression Analysis

In this section we describe how we calculate the p values required to measure the association between SNPs and the given phenotype (coronary artery disease risk). A logistic

regression was used to better account for the categorical nature of the data (being labelled based on the allele frequency (0, 1, or 2). A future consideration would be to use a mixed model regression. We looked at the pairwise correlations of the SNPs to identify SNP interactions and weight the regression.

5 Results and visualisations

Below we present the manhattan plots, describing the association between SNP sites and deviation between control and target cases. We interpret a large $log(p_i)$ value for site i as meaning there is a significant difference in the allele frequency between control and target cases. This means that a particular phenotype can be associated with a particular set of SNPs.

We now present the SNP labels associated with CAD, HDL and LDL with the p values overcoming a candidate threshold value.

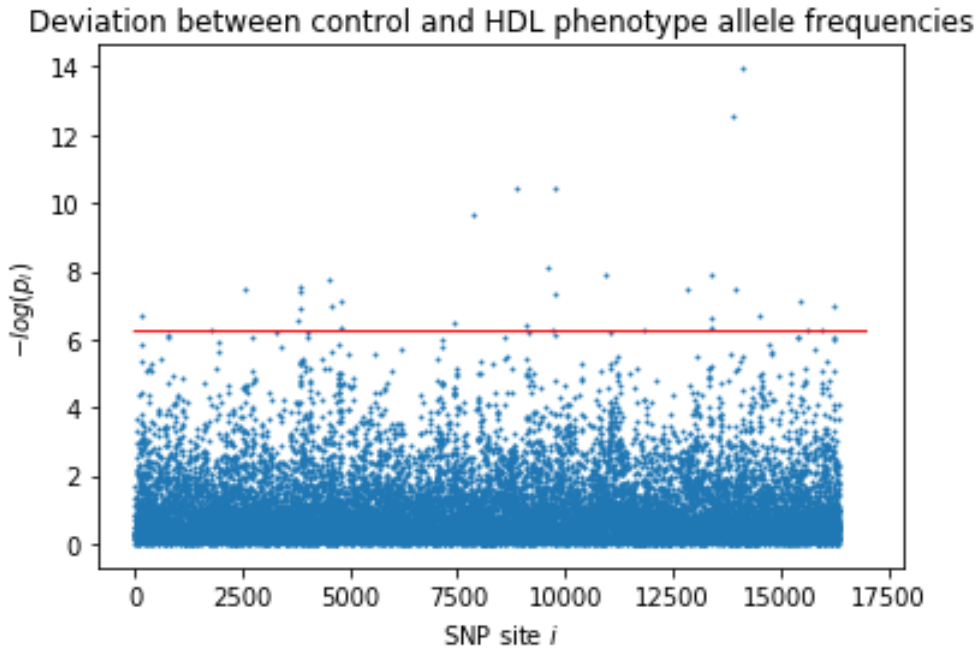

Figure 1: Manhattan Plot of the deviation between the control genotypes and HDL phenotype allele frequencies. Bonferonni correction is put in as a threshold represented by the straight red line. SNPs above this point surpass the candidate threshold and are important for lab biologists.

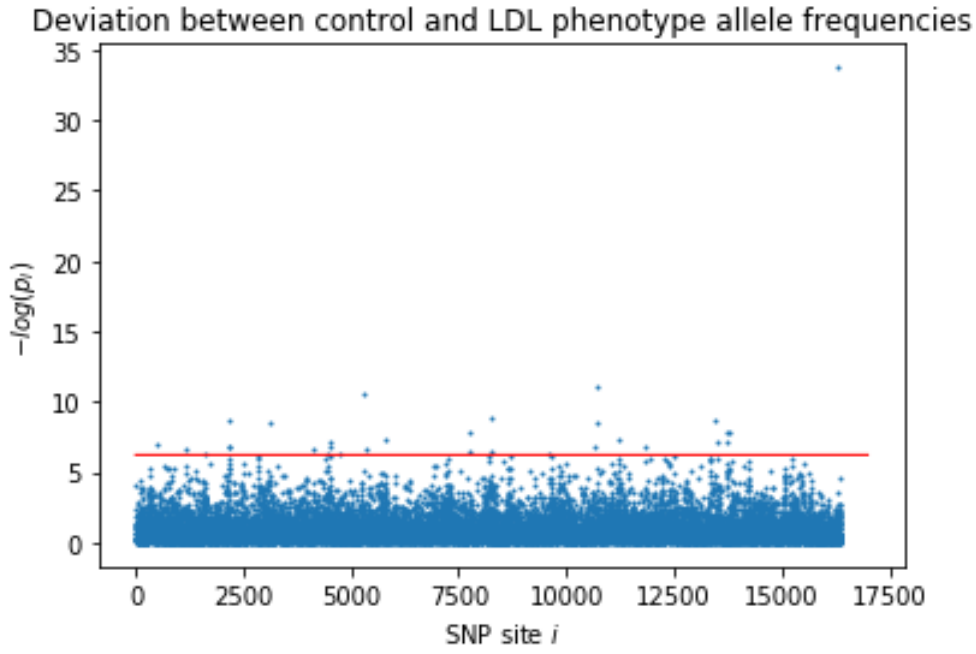

Figure 2: Manhattan Plot of the deviation between the control genotypes and LDL phenotype allele frequencies. Bonferonni correction is put in as a threshold represented by the straight red line. SNPs above this point surpass the candidate threshold and are important for lab biologists.

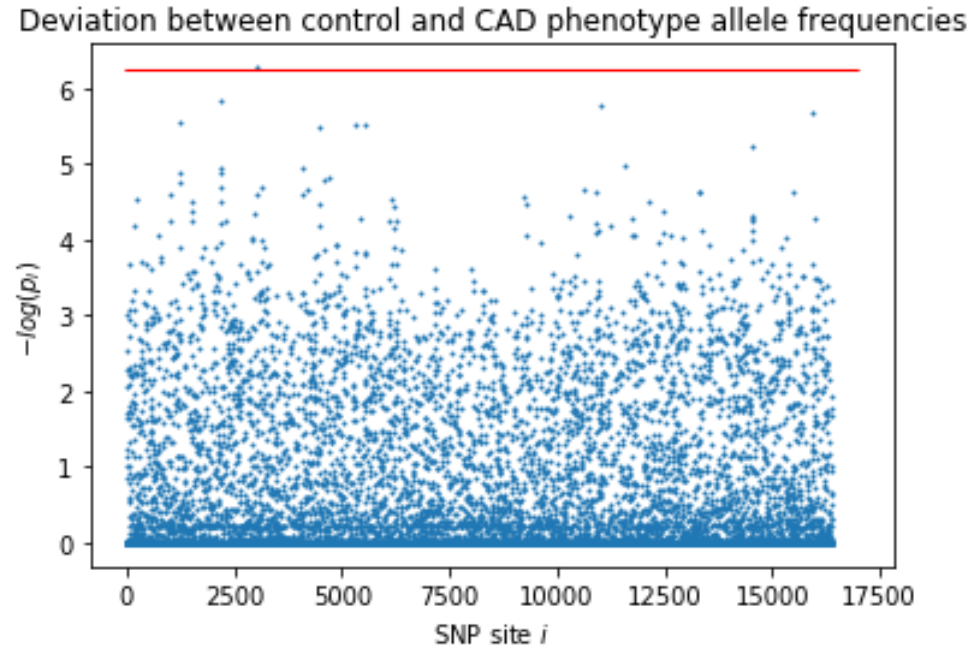

Figure 3: Manhattan Plot of the deviation between the control genotypes and CAD phenotype allele frequencies. Bonferonni correction is put in as a threshold represented by the straight red line. SNPs above this point surpass the candidate threshold and are important for lab biologists.

We set a candidate =. Given that analysis has been performed multiple times on the same data set, we must apply the Bonferroni correction to take into account the fact

that the candidate will have a high probability of giving a type 1 hypothesis testing error across at least one of the models. The correction means that the candidate alpha value must be divided by m = 20 to take into account the repeated models on the same data. This ensures that the of the whole model will be that of the candidate.

6 Conclusion

While performing all of these complex statistical metrics on the genetic data, it is important to keep in mind that many assumptions are in play. As much as we try and reduce bias we are still only making associations, the biological basis for the disease and the pathological mechanisms involved still need to be evaluated. Minimising the statistical biases, we can conclude with some certainty that the above mentioned SNPs should be candidates for future research in the prevalence of SNPs.

References

[1] P. M. Visscher et al., 10 Years of GWAS Discovery: Biology, Function, and Translation, Am J Hum Genet, vol. 101, no. 1, pp. 522, Jul. 2017.

[2] N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher, Pitfalls of predicting complex traits from SNPs, Nature Reviews Genetics, vol. 14, no. 7, pp. 507515, Jul. 2013.

[3] S. Service et al., Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies, Nature Genetics, vol. 38, no. 5, pp. 556560, May 2006.

[4] K. J. Galinsky et al., Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia, Am. J. Hum. Genet., vol. 98, no. 3, pp. 456472, Mar. 2016.

[5] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, Principal components analysis corrects for stratification in genome-wide association studies, Nat. Genet., vol. 38, no. 8, pp. 904909, Aug. 2006.