Genome-wide Association Study (GWAS) on HDL Cholesterol Levels

Arpi Beshlikyan (SID: 404239449)

Professor Jae Hoon Sul

UCLA Fall 2020, CS CM121

Introduction

With the discovery of DNA, and later the invention of sequencing technology, came an important and immediately apparent question -- how do our genes and genetic variants influence our phenotypic traits? As most genetic variants are single nucleotide polymorphisms (SNPs), where a single base varies between two alleles, we conduct association studies between a given SNP and a phenotypic trait, which can either be quantitative or binary. An association study computes the correlation between the presence of a certain allele at a set base location and the sample's phenotypic value; if that correlation is above a certain threshold, then that SNP is said to be associated with the disease in question. In a Genome-wide Association Study (GWAS), the association test is performed on multiple SNPs, and the SNPs that peak in their correlations with the phenotype are said to have a strong association with the phenotype.

To perform a GWAS, one must first collect diverse phenotypic samples of cases and controls, as well as their genotypic (SNP) information, then perform quality control to identify and remove the SNP and individuals with poor quality, and finally conduct the association analysis to identify the SNPs (if any) that are most associated with the phenotype of interest. After the analysis, supplemental analyses such as estimating heritability and calculating polygenic risk scores to predict the likelihood of an individual having a certain phenotype can then be used to understand the genetic basis of a phenotype or disease.

Method

I started the analysis by downloading the provided low-coverage whole-genome sequencing data files that had been preprocessed to remove the rare SNPs (with MAF < 15%), as well as those that are highly correlated to each other. The resulting data is on 2,504 individuals and 828,325 SNPs in binary PLINK format[1].

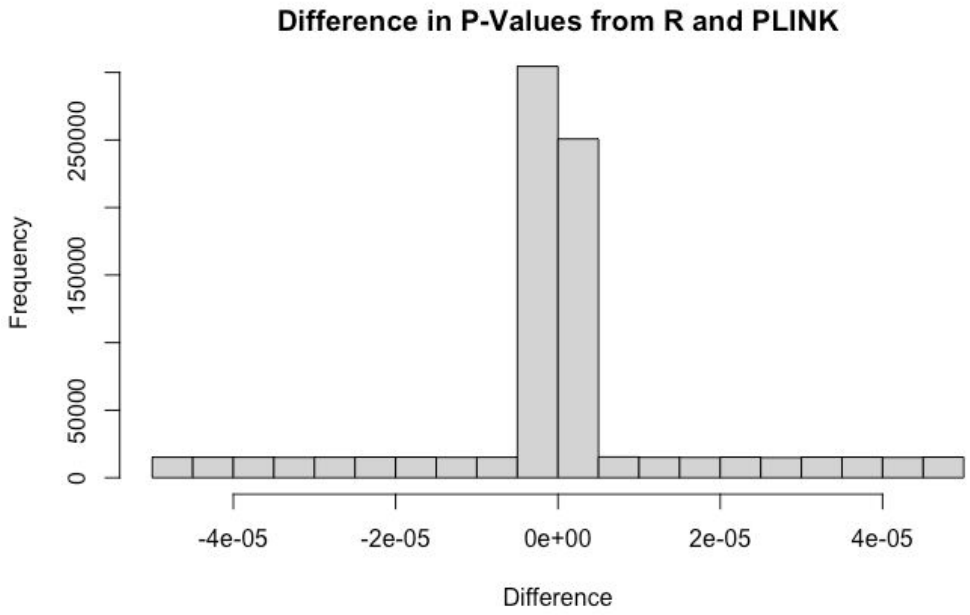
I then performed univariate linear regressions using the SNP data and phenotype data with PLINK, which outputted the results into the file *plink.assoc.linear*. I converted the encoding of the pairs of SNP data to be in the additive encoding, performed the linear regression in R, and plotted the differences between the p-values from the regressions in R and the regressions in PLINK.

[1] More details here: http://zzz.bwh.harvard.edu/plink/data.shtml

I used the provided script to draw the Manhattan plot of the p-values from the PLINK linear regressions, visualizing which SNPs have the strongest association to the level of HDL cholesterol. I found the SNP with the smallest p-value in each chromosome as calculated using the PLINK regression, and compared it to the corresponding SNP's p-value listed in the UK Biobank GWAS summary statistics[2], and summarized my findings in a table. I then looked through the dbSNP database to see if the same SNPs I identified to have a strong association with HDL cholesterol levels appeared in previous studies about the same phenotype.

Results

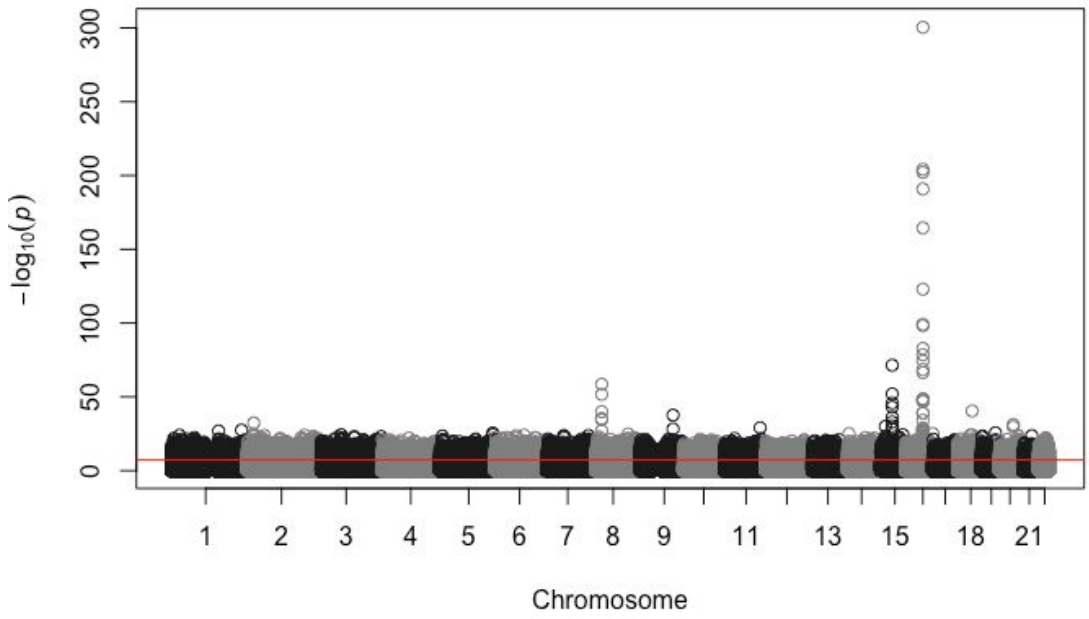
The frequencies of the differences between the p-values I calculated using R and using PLINK are shown below. The median difference is 0.000, while the largest magnitude of difference observed is 5.000e-05. This indicates that there are very minimal differences between the two sets of p-values and that the p-values calculated using R are fairly accurate.



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-5.000e-05	-1.171e-06	0.000e+00	-7.390e-09	1.166e-06	5.000e-05

[2] https://www.dropbox.com/s/65jisgxwbbdrkaw/30760_irnt.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0

The Manhattan plot for the p-values calculated using PLINK is below. There are significant peaks, showing significant association for that SNP, in Chromosomes 8, 9, 15, 16, and 18.



The table on the following page compares the p-values calculated in this association study with the p-values in the UK Biobank GWAS summary for the same study on HDL cholesterol. The three most significantly low p-values from my analysis with PLINK have p-values of 0.00 in the Biobank statistics, as shown in the rows shaded green. The next two lowest p-values from my PLINK analysis are also the next two lowest p-values in the Biobank p-values, in the magnitude of E-178 and E-160, as shown in the rows shaded yellow.

There are a few SNPs that have a high but still seemingly significant PLINK p-value, but an objectively insignificant Biobank p-value. This could be due to the Biobank GWAS having a number of sampled much than 2,504 to conduct the study on, effectively amplifying the significance of the truly associated SNPs and the insignificance of the irrelevant SNPs. This would also explain how the Biobank p-values for the three SNPs with the greatest association found in my PLINK analysis can be 0.00.

Chromosome	SNP ID	PLINK p-value	Biobank p-value
1	rs4846920	3.33E-28	3.55E-108
2	rs6754295	4.80E-33	9.48E-92
3	rs1112403	4.61E-25	9.92E-01
4	rs6532041	7.85E-23	4.76E-01
5	rs1036172	3.01E-26	1.58E-02
6	rs2819974	8.16E-25	3.55E-01
7	rs834793	1.42E-24	5.47E-01
8	rs35617716	2.84E-59	0.00E+00
9	rs2740488	2.47E-38	2.90E-160
10	rs4300315	4.61E-22	3.51E-01
11	rs10750097	8.95E-30	3.40E-113
12	rs10846772	1.70E-21	1.11E-01
13	rs9519977	1.14E-21	3.96E-01
14	rs2332101	8.61E-26	9.90E-01
15	rs1077835	3.58E-72	0.00E+00
16	rs11076175	4.12E-301	0.00E+00
17	rs35772501	5.58E-22	8.86E-02
18	rs4939883	3.50E-41	6.83E-178
19	rs429358	2.31E-26	1.23E-127
20	rs6073958	6.01E-32	1.77E-108
21	rs9983496	1.83E-24	7.82E-02
22	rs732381	4.45E-23	6.56E-01

Discussion

The most significant SNP (`rs110761`), on Chromosome 16, was found to be significant in a 2013 paper on the HDL cholesterol levels in the Latvian population[3], as well as identified as a variant associated with a lipid-related gene in a paper from 2011[4]. However, the second most significant SNP (`rs1077835`), found in my analysis, on Chromosome 15, is mentioned many more times in papers about plasma lipids and HDL, most recently in a 2020 paper[5]. I did not find any papers through the dbSNP that mention the third most significant SNP (`rs35617716`), but the fourth most significant SNP () on Chromosome 18 was mentioned many times in papers on lipid levels and hypercholesterolemia, most recently in a paper from 2019[6].

In my analysis, many of the processes were complicated to code or took a long time to run through R, as the datasets were very large and could not be loaded into memory easily. Thus, a future improvement could be to store the datasets in a database and use SQL to optimize the loading of only the data that is necessary for the current process into memory. I also had trouble with the provided R script to generate the Manhattan plot and had to change line 37 to set *d$pos* to *NaN*, and not to NA.

In conclusion, the GWAS performed on the data provided, both through PLINK and through R, was successful in identifying the SNPs most associated with HDL cholesterol levels as confirmed by the statistics from the UK Biobank GWAS, as well as the studies related to those SNPs in the dbSNP[7] database.

---

[3] Radovica I, Fridmanis D, Vaivade I, Nikitina-Zake L, Klovins J. The association of common SNPs and haplotypes in CETP gene with HDL cholesterol levels in Latvian population. PLoS One. 2013 May 13;8(5):e64191. doi: 10.1371/journal.pone.0064191. PMID: 23675527; PMCID: PMC3652817.

[4] Carlquist JF, McKinney JT, Horne BD, Camp NJ, Cannon-Albright L, Muhlestein JB, Hopkins P, Clarke JL, Mower CP, Park JJ, Nicholas ZP, Huntinghouse JA, Anderson JL. Common Variants in 6 Lipid-Related Genes Discovered by High-Resolution DNA Melting Analysis and Their Association with Plasma Lipids. J Clin Exp Cardiolog. 2011 Jul 10;2(138):2155-9880-2-138. doi: 10.4172/2155-9880.1000138. PMID: 22229114; PMCID: PMC3251308.

[5] Wei W, Hu T, Luo H, Ye Z, Lu F, Wu Y, Ying M. The cross-sectional study of hepatic lipase SNPs and plasma lipid levels. Food Sci Nutr. 2020 Jan 13;8(2):1162-1172. doi: 10.1002/fsn3.1403. PMID: 32341780; PMCID: PMC7180388.

[6] Yang S, Yin RX, Miao L, Zhang QH, Zhou YG, Wu J. Association between the LIPG polymorphisms and serum lipid levels in the Maonan and Han populations. J Gene Med. 2019 Feb;21(2-3):e3071. doi: 10.1002/jgm.3071. Epub 2019 Feb 4. PMID: 30657227; PMCID: PMC6590183.

[7] Searched through here: https://www.ncbi.nlm.nih.gov/snp/?cmd=search