COM SCI CM121 Final Project

HDL Cholesterol Level GWAS

Nuutti Barron, 805 205 422

Introduction:

This project consists of analyzing a dataset of 828,325 preselected SNPs (Single Nucleotide Polymorphisms) over 2,504 individuals in an attempt to uncover associations between SNPs and HDL cholesterol level in an additive model of alleles for phenotype, and represent these associations graphically. We investigate various means of data analysis, including implementing our own linear regression in R, and using the plink whole genome association analysis toolset software. The data used in this project is provided in binary PLINK format, and was generated for the purpose of this project, but we compare the results we find in this project to similar data obtained from the publicly accessible UK Biobank and dbSNP database.

Methods and Results:

1) In this part we use the plink genome association analysis toolset to perform a linear regression for each predetermined SNP and the phenotype data in our dataset. We obtain the genomic data from the "snpdata" files (.bim, .bed, and .map) and the phenotype data from a separate text file. The linear regression outputs a file (plink.assoc.linear) with identifying information for each SNP, and the statistical p-value of association between each individual SNP and HDL cholesterol level as determined by linear regression. Using the tool provided by plink, this took about two hours to complete. Instructions on how to perform this analysis were found on plink website's documentation.

2) First, we use the recode tool offered by plink to convert the binary data file to a text file that can be read into an R script. The file format was transposed in the process so that each row (as opposed to each column) in the produced text file would contain all of the necessary information for a single allele. This helps optimize the linear regression we do later in R, because we can read all the information for a single SNP (over all 2,504 individuals) in one read of the text file. This is important, because the data file is so large (around 8 GB) that we cannot read all of it into an R program at once, so we have to dynamically buffer it in a smaller amount of data at a time. This means that in order to perform a linear regression in R on a SNP and phenotype

data, we need all the SNP data to be encoded into a single row (all of the SNP data is required at the time of performing the regression, and the size of the file prohibits access to all of this data at once if it is contained in a column on the text file), hence why the file is transposed. When converting the binary data to text data, we also encode the major and minor alleles as 1 and 2 (as opposed to A, C, T, or G), so that we can later encode SNPs as 0 (homozygous minor), 1 (for heterozygous), or 2 (homozygous major). This encoding is critical for performing a linear regression on the allele and phenotype data in R. After encoding the allele into this format, the linear regression is performed (with allele encoding as a independent variable, and phenotype as the dependent variable), and the p-value of association between each allele and HDL cholesterol level is determined and compared to the p-value obtained from the plink linear regression tool.

Based on the scatterplot depicting the different between p-values for each SNP as determined by plink and linear regression in R, as well as manual inspection of the p-values both output found that both methods found each SNP to have effectively the same resulting p-value (association with HDL cholesterol level). There are very minute differences in p-values, but these seem to arise because the plink linear regression tool truncates p-values to a certain number of digits, and round to the nearest digit, while the p-values determined by linear regression in R had more precision (more decimal places). While plink may not have had the same level of precision as the p-values determined in R, the linear regression in plink was much faster, as the linear regression in R for every SNP took nearly ten hours, and plink was able to do linear regression in around two hours on the machine I used

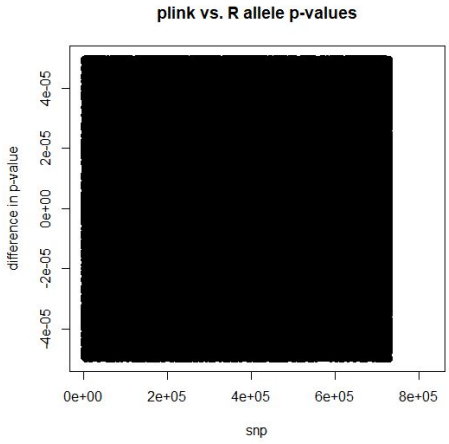

Figure 1: Difference in p-value determined by linear regression in R and plink for each SNP. Note from the y-axis that the variation in p-value is < ~0.0001

3) To create the Manhattan plot for this GWAS, we utilized the provided qqman.r script. The function in this script that creates the graphic requires as input an R dataframe which is formatted to include the chromosome number, base pair location on its chromosome, and statistical p-value for each SNP. To create this formatted dataframe, we first read in all the data from the file containing the plink linear regression data (from problem 1), and keep the columns that contain chromosome, base pair, and p-value date. Then this data frame is passed to the manhattan() function, which creates the Manhattan plot for this study.

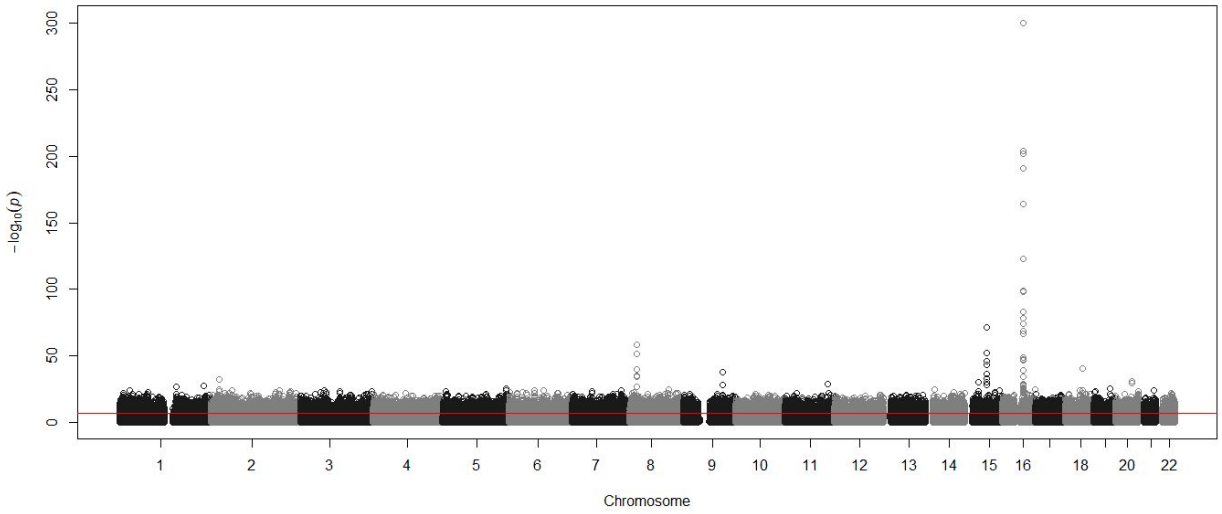

Figure 2: Manhattan Plot for HDL cholesterol in this GWAS

4) To find the SNP with minimum p-value (with the greatest association with HDL cholesterol level) for each chromosome, we first create a container for each chromosome that contains all the SNPs and SNP data we analyze located on that chromosome. Then we simply find the SNP with minimum p-value in each container, and save its information. To get a reference for the accuracy of our results, we compare our results to data obtained from UK Biobank GWAS summary statistics for HDL cholesterol. We download the summary file, and find the p-value the UK Biobank GWAS study found for each of the SNPs we found to have minimum value in our study. Once again, the file is too large to pull all of its content into an R code at once, so we dynamically buffer parts of the file into our R program and search for the SNPs (based on chromosome and base pair information) we identified in our study.

Chr	1	2	3	4	5	6	7	8	9	10	11
SNP	rs4846920	rs6754295	rs1112403	rs6532041	rs1036172	rs2819974	rs834793	rs35617716	rs2740488	rs4300315	rs10750097
-log(p_R)	27.477	32.319	24.336	22.104	25.521	24.088	23.847	58.546	37.608	21.336	29.048
-log(p_U)	107.44	91.023	0.0035	0.3222	1.8011	0.4503	0.2622	NA	159.53	0.4548	112.46

Chr	12	13	14	15	16	17	18	19	20	21	22
SNP	rs10846772	rs9519977	rs2332101	rs1077835	rs11076175	rs35772501	rs4939883	rs429358	rs6073958	rs9983496	rs732381
-log(p_R)	20.770	20.942	25.065	71.445	300.38	21.253	40.455	25.636	31.221	23.737	22.351
-log(p_U)	0.954	0.402	0.004	NA	NA	1.0523	1.6536	126.90	107.75	1.1068	0.1833

Figure 3: SNP with minimum p-value for every chromosome determined by plink (denoted P_R), and the corresponding SNP's p-value in UKBB database (denoted P_U). NA entry: p = 0

5) I looked up each of the SNPs identified in part (4) in the dbSNP database and looked at each one's 'clinical significance', as well as references to these SNPs in Pubmed to determine if previous studies have found an association with these SNPs and HDL cholesterol level in the past. The following SNPs have been found to be associated with HDL cholesterol level in past studies:

- **rs6754295 - Chromosome 2**: 6 PubMed studies published on association with HDL cholesterol level and heart disease
- **rs10750097 - Chromosome 11**: 10 PubMed studies published on this SNP and heart conditions, many of which name association with HDL cholesterol level in title
- **rs1077835 - Chromosome 15**: 7 PubMed studies published on association with this SNP and HDL cholesterol
- **rs11076175 - Chromosome 16**: 2 PubMed studies found an association with this SNP and HDL cholesterol
- **rs4939883 - Chromosome 18**: 26 PubMed citations to this SNP, many of which cite an association to HDL cholesterol

Note: Some of the other SNPs I researched had studies published on their significance (for example, rs429358 on chromosome 19 had over 200 PubMed citations for its association with neurodegenerative disorders such as alzheimers and parkinsons disease), but none of these studies cited an association with HDL cholesterol.

Discussion:

A few things were accomplished during this project. I performed linear regression both in R and using the plink toolset, and found the statistical association between 828,325 common SNPs and HDL cholesterol levels in a sample of 2,504 people. Both methods had nearly identical results, and there are advantages to both. Linear Regression in R was more precise, but using plink linear regression was performed much faster. We were able to compare the results of this study to previous studies as well, by comparing p-value determined in this study to studies found in UK Biobank and dbSNP database, and found that some of these SNPs have been found to be associated with HDL cholesterol level in previous studies.

There were two major challenges in this project. One was writing R code that would execute on such a large set of data in a reasonable amount of time. I initially used a lot of for-loops, which tend to be slow in R, and which I was able to substitute with calls to apply() in R to optimize my code. This made the runtime of my code go from the timescale of days to hours, displaying massive improvement. The second major challenge was reading data from a file larger than what R is able to interpret in a single read. This meant that in my algorithms, all the data read by R from large files had to be buffered bit by bit. An avenue of further optimization for this code would be determining the ideal chunk size to read from the file in a single read. I used 425 lines, as it is a factor of 828,325 (the number of SNPs). Another area of possible optimization would be if using plink we could both transpose and encode SNPs into the 0, 1, 2 encoding used for linear regression in R. plink currently only allows you to do one or the other, and since I needed the data file to be transposed, it meant I had to manually encode the SNPs, which likely is inefficient compared to if the data file was already encoded before it was read in by R script. Lastly, using linear regression assumes an additive mode of major and minor alleles, but there are other biological models which describe phenotype as a different function of alleles. This is an avenue that can be explored in future studies.