HGG621 Final Project Report

Ruijie Yin

Lab date: 10/8/2019. Report date: 12/2/2019.

Abstract

Left atrial enlargement has been proved to be related to many cardiovascular disease, however, genetic factors contributing to the left atrium (LA) size are poorly understood. This project aims to map susceptibility genes for LA size using the data from Northern Manhattan Study (NOMAS), 1350 individuals participated in this study and 942 of them had their LA sizes measured using transthoracic echocardiogram (TTE). Genome-wide association study (GWAS) was performed using the single-nucleotide polymorphism (SNPs) genotyped in this data and evidence for association were found in genes such as PTPRG, SEMA3C and CBFA2T3.

1 Introduction

The left atrium plays a major role in cardiac physiology by collecting blood during systole and modulating left ventricular filling during diastole [1]. Left ventricular diastolic dysfunction or mitral valve disease may lead to left atrial pressure or volume overload which, if chronically maintained, could result in left atrial enlargement and remodeling [2]. As a marker of left ventricular diastolic dysfunction or increased filling volumes, left atrial size may provide important prognostic information [3]. In this regard, left atrial enlargement has been related to higher risk of atrial fibrillation, a condition that by itself can increase the risk of ischemic stroke and death [4].The association between increased risk of ischemic stroke and enlarged left atrial size has been well documented,even in subjects without atrial fibrillation. Multiple acquired conditions have been associated with LA enlargement. Among them are mitral valve disease, arterial hypertension and any condition that increases the left ventricular filling pressures[1]. Understanding the genetic influence and factors on left atrial size would help in identifying subjects at increased risk for developing an enlarged LA especially at an early stage. In addition, this knowledge is essential for understanding cardiac structure and function at the molecular level and identifying therapeutic targets in the management of left atrial size enlargement. A recent study identified several possible susceptibility genes affecting left atrial size. Among them, MYOCD has been shown to serve as a key transducer of hypertrophic signals in cardiomyocytes in vitro. Evidence from linkage and association study, together with the known function, suggests that polymorphisms in MYOCD gene modify left atrial size [5]. This final project aims to find out if there is any genotyped SNPs associated with left atrial size in the data from Northern Manhattan Study (NOMAS) using Genome-wide association study (GWAS).

2 Study Design and Methods

2.1 Data

The dataset used in this project was provided by Drs.Ralph Sacco and Tatjana Rundek of the Northern Manhattan Study (NOMAS). The data consist of 1397 individuals with genotype and phenotype information. Among the 1397 individuals, 943 of them had the phenotype of interest, left atrial size measured. LA size was further dichotomized using the common threshold of log(29) (mm^3/m^2) in clinical practice; Subjects with LA sizes greater size than log(29) are labeled as 'cases' and are labeled as 'controls' if LA sizes are less than log(29). The criterions of selection of the subjects, the baseline evaluation of the subjects, laboratory assessments, annual prospective follow-up scheme and outcome classification, etc. can be found in the relevant literature on this study [6]. The self-reported gender, smoking status, age at time of phenotype measurement, diabetes status, Body Mass Index(BMI), hypertension status and dislipidemia status were also reported and served as the covariates in our association analysis.

2.2 GWAS

Genome-wide association study (GWAS) refers to study in which hundreds of thousands of single-nucleotide polymorphisms (SNPs) are genotyped across the genome and tested for association with the phenotype of interest. In the past few years, numerous genetic susceptibility loci have been identified to be associated with many complex diseases via GWAS, including but not limited to a variety of cancers, bipolar disorder and coronary artery disease[7]. Typically, the free, open-source whole genome association analysis toolset **Plink** (Version 1.9), was used throughout the project; The linkage disequilibrium blocks of the selected candidate SNPs from GWAS were generated using **Locuszoom** and annotations were obtained from UCSC Genome Browser and Ensembl. The null hypothesis for GWAS in this project is "none of the SNP loci genotyped in these data are associated with LA size." and the alternate hypothesis is "at least 1 of the genotyped SNPs is associated with the LA size in these data."

2.3 Logistic Analysis

Denote Y_i as the qualitative phenotypic value of the i^{th} subject (0=control, 1=case), $X_{ij} = \{0, 1, 2\}$ is the genotypic value of the j^{th} SNP of the i^{th} subject, in the additive model assumption this project used, $X_{ij} = 0$ represents a homozygote, $X_{ij} = 1$ stands for a heterozygote and $X_{ij} = 2$ is another homozygote; Denote Z_{ik} as the value (quantitative or qualitative) of the k^{th} covariate of the i^{th} individual, the Logistic Regression model can be stated as following:

$$\Pr(\mathrm{Y}_i = 1|X_{ij}) = \frac{exp(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+...+\beta_j X_{ij}+\gamma_1 Z_{i1}+...+\gamma_k Z_{ik})}{1+exp(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+...+\beta_j X_{ij}+\gamma_1 Z_{i1}+...+\gamma_k Z_{ik})}$$

We test if any of the regression coefficients β_i is different from 0, it is equivalently to testing if the odds ratio for the association between the j^{th} SNP and the phenotype is different from 1. The P-values of odds ratios were reported and those SNPs that have P-values below a predetermined significance level are selected for further analysis.

2.4 Association Analysis with Bonferroni Correction

Bonferroni correction offers a convenient way to control the family-wise error rate (FWER) by dividing α by m provided that the markers are uncorrelated. The FWER is the probability of rejecting at least one null hypothesis when all the null hypothesis are correct [7]. The resulting significance level is $\frac{\alpha}{m}$, an SNP is then considered to be statistically significant if its P-value is less than this adjusted significance level. Since $616,027$ SNPs were tested, the significance level in this project was defined at $0.05/616027 = 8.12 \times 10^{-8}$. For the given sample size in this project, the minimum detectable risk is 0.345 or 1.855 on the other side, the power, however, is only about 35%.

3 Results

3.1 Descriptive Statistics

The demographics statistics of the subjects involved are summarized in Table 1. Participants with missing measurements are excluded.

3.2 Quality Control

Several quality control steps were conducted prior to genetic association analysis, and the results corresponding to each step is shown as below:

3.2.1 Missingness of SNPs and individuals

SNPs and individuals with very high levels of missingness were first filtered out based on a relaxed threshold (0.2; $>$ 20%).Then a filter with a more stringent threshold were applied (0.02) and SNP filtering were performed before individual filtering:

Characteristics	Controls(N(%))	Cases(N(%))
# *of subjects*	75(7.96%)	867(92.04%)
Race		
Hispanic	58(8.52%)	623(91.48%)
Non-Hispanic Black	12(7.89%)	140(92.11%)
Non-Hispanic White	5(5.21%)	91(94.79%)
Other	0(0%)	13(100%)
Sex		
Male	14(3.65%)	369(96.35%)
Female	36(6.44%)	523(93.56%)
Smoking		
Yes	26(5.22%)	472(94.78%)
No	24(5.4%)	420(94.6%)
Diabetes		
Yes	7(3.72%)	181(96.28%)
No	43(5.7%)	711(94.3%)
Hypertension		
Yes	26(3.88%)	644(96.12%)
No	24(8.79%)	249(91.21%)
Dislipidemia		
Yes	20(3.73%)	516(96.27%)
No	31(6.31%)	460(93.69%)
	$Mean \pm SD$	$Mean \pm SD$
Age at 1st Measurement	64.95 ± 8.15	66.01 ± 7.84
BMI	25.35 ± 4.27	28.63 ± 4.92

Table 1: The demographics statistics of the 942 participants.

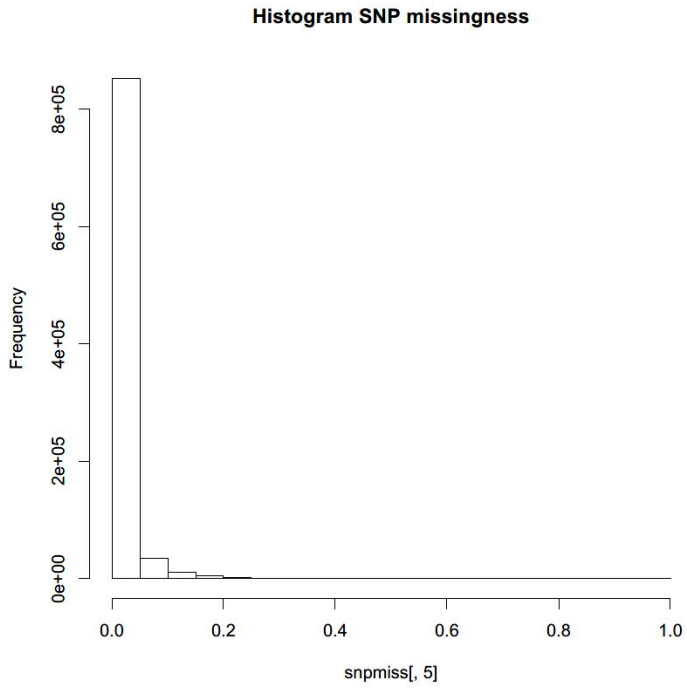

Figure 1: Proportion of missing individuals per SNP.

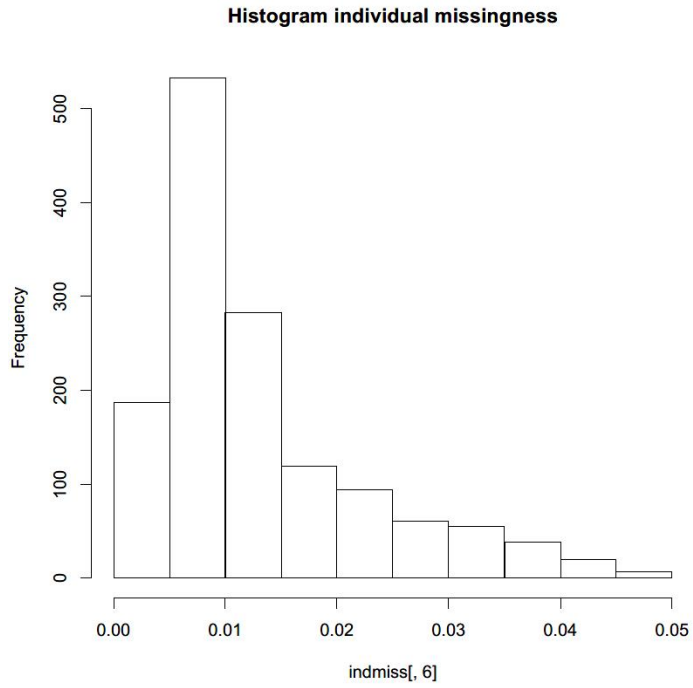

Figure 2: Proportion of missing SNPs per individual.

Results shows that only a very small proportion of SNPs have missing individuals and most of the individuals do not have missing SNPs or have low missingness rate.

3.2.2 Sex Discrepancy

This step checks for discrepancies between self-reported sex of the individuals and their sex based on X chromosome heterozygosity/homozygosity rates. Males should have an X chromosome homozygosity estimate greater than 0.8 and females should be than 0.2. Individuals who do not fulfil these requirements were then deleted.The plots seem fine and 753,566 variants and 1379 people pass filters and QC:

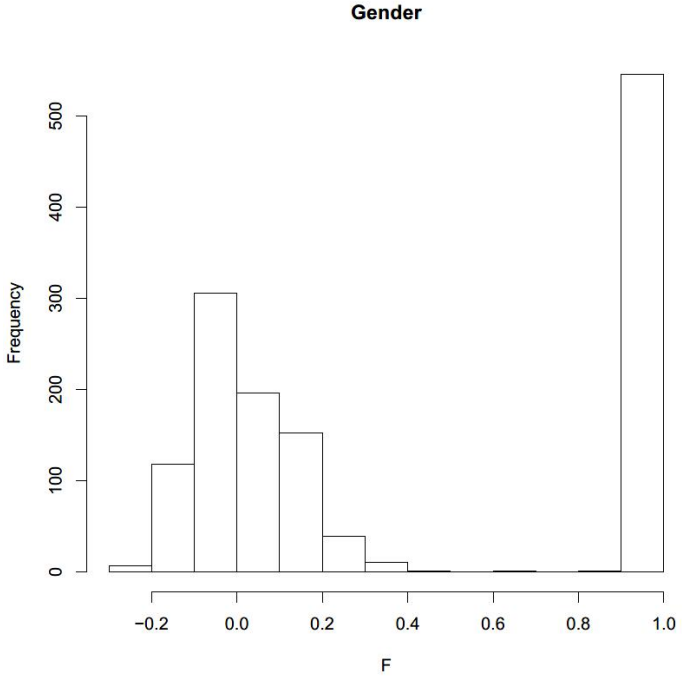

Figure 3: Sex Discrepancy check for all genders.

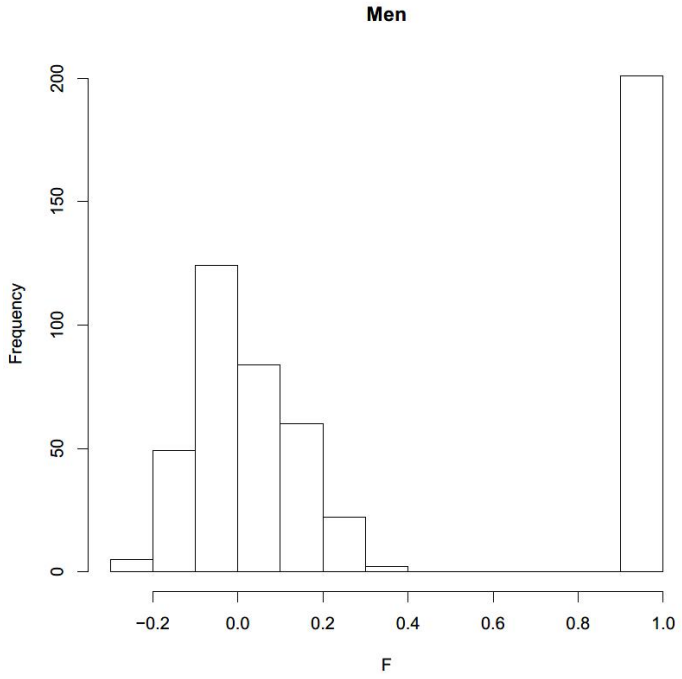

Figure 4: Sex Discrepancy check for men.

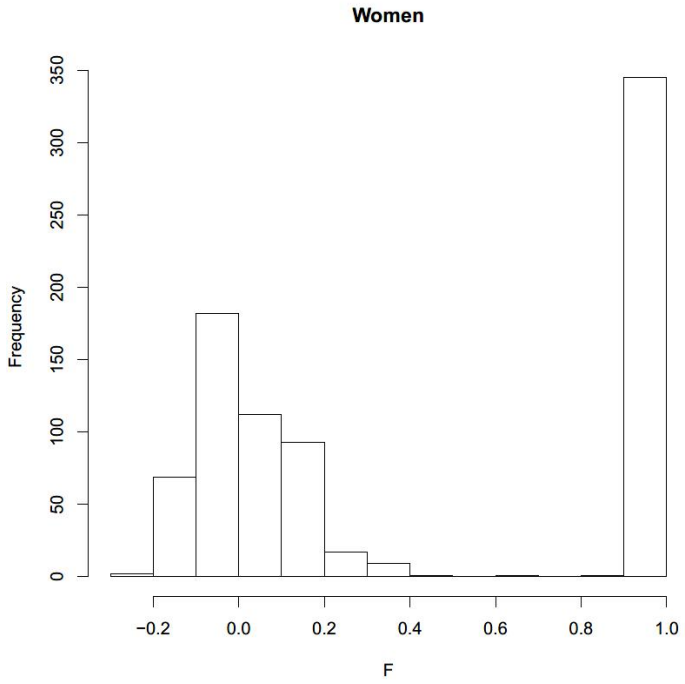

Figure 5: Sex Discrepancy check for women.

3.2.3 Minor allele frequency (MAF)

A conventional MAF threshold for a regular GWAS is between 0.01 or 0.05, a plot of the MAF distribution is shown below and it can been seen from the histogram that roughly 10,000 SNPs have very low MAF frequency and are removed in this step; 626,146 variants and 1379 people pass filters and QC:

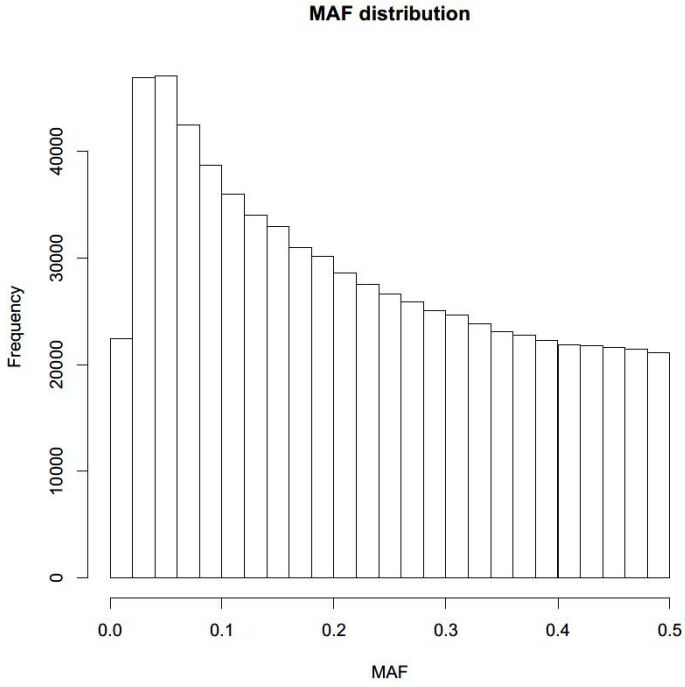

Figure 6: The distribution of MAF

3.2.4 HardyWeinberg equilibrium (HWE)

Although the conventional standard is to exclude SNPs with HWE P-value $< 1 \times 10^{-10}$ in cases and $< 1 \times 10^{-6}$ in controls, given the relatively small sample size in this project, we implemented a less strict case threshold of P-value $< 1 \times 10^{-6}$; The histogram of the HWE p-values are shown in the following two figures, a number of SNPs are deviated and are thus deleted:

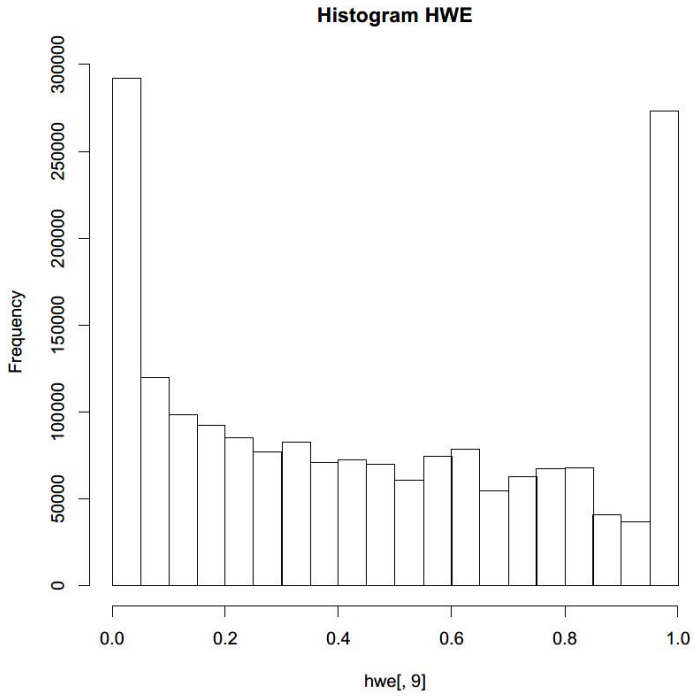

Figure 7: The distribution of HWE p-values of all SNPs.

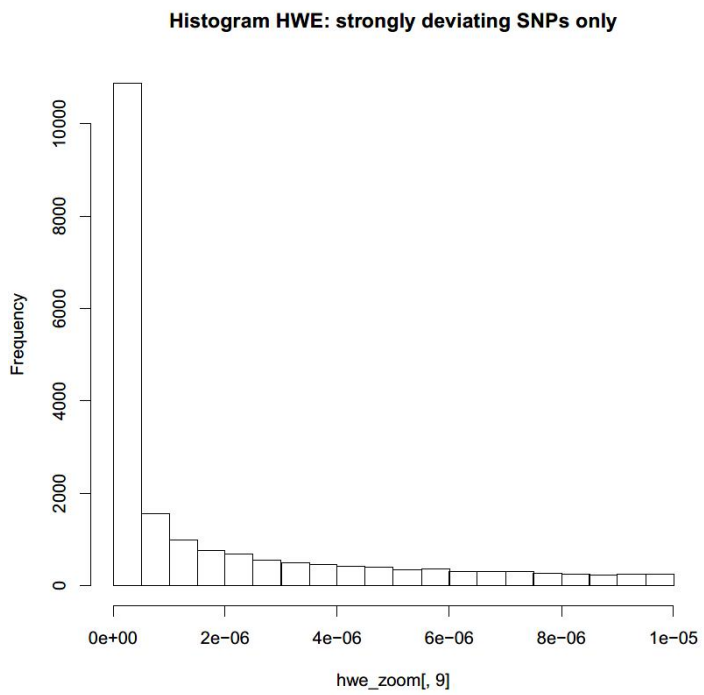

Figure 8: the distribution of HWE p-values of strongly deviating SNPs.

3.2.5 Heterozygosity

In this step, individuals that deviate ±3 SD from the samples' heterozygosity rate mean were removed. A plot of the distribution of the heterozygosity rate of the subjects is shown in Fugure 9:

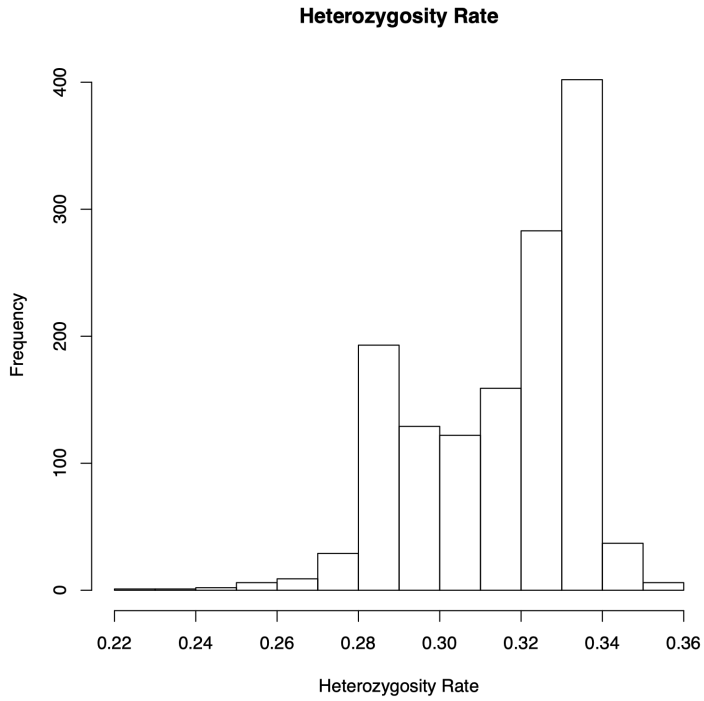

Figure 9: The distribution of the heterozygosity rate of the subjects.

3.2.6 Relatedness

The fact that cryptic relatedness can interfere with the association analysis drives us to detect the subjects that are related. A pi-hat threshold of 0.2 was applied and all individuals above this threshold are excluded. Since our data only have the founders, we expected and did see an empty plot of relatedness of the subjects as below:

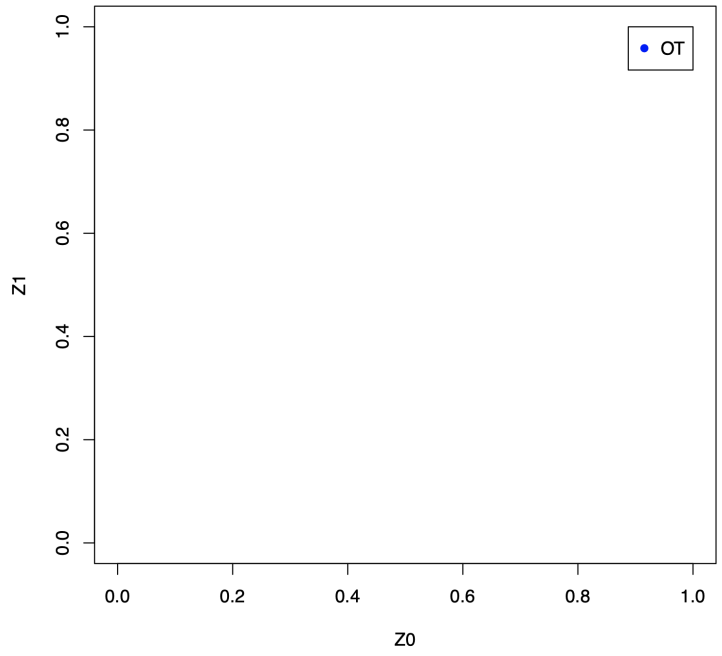

Figure 10: The relatedness of the subjects.

After this step, 616,027 variants and 1372 people have passed filters and QC.

3.2.7 Population Stratification

Testing and controlling for the presence of population stratification is an essential QC step prior to conducting GWAS. The multidimensional scaling (MDS) approach was implemented in this project and the population stratification was checked using data from the 1000 Genomes Project (1KG). Given the ethnic structure in our data, none of the individuals were removed and 10 main components were subsequently used as covariates in the logistic analysis.

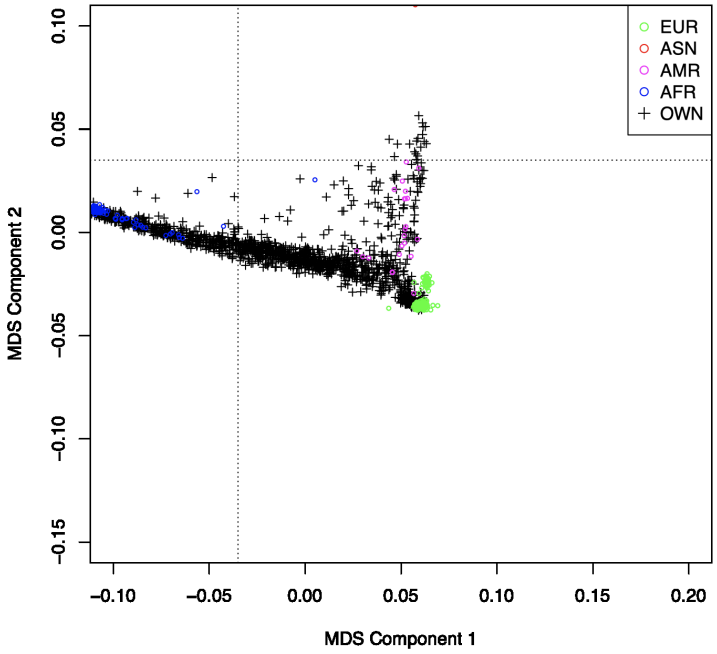

Figure 11: Multidimensional scaling (MDS) plot of 1KG against NOMAS data.

3.3 Association Analysis

3.3.1 Logistic Analysis

In the logistic analysis, 10 MDS principal components we obtained in the population stratification step along with several other variables (such as self-reported gender, smoking status, age at time of phenotype measurement,etc.) were served as our covariates. The corresponding Quantile-Quantile plot and the Manhattan plot are shown in Figure 12 and 13, respectively:

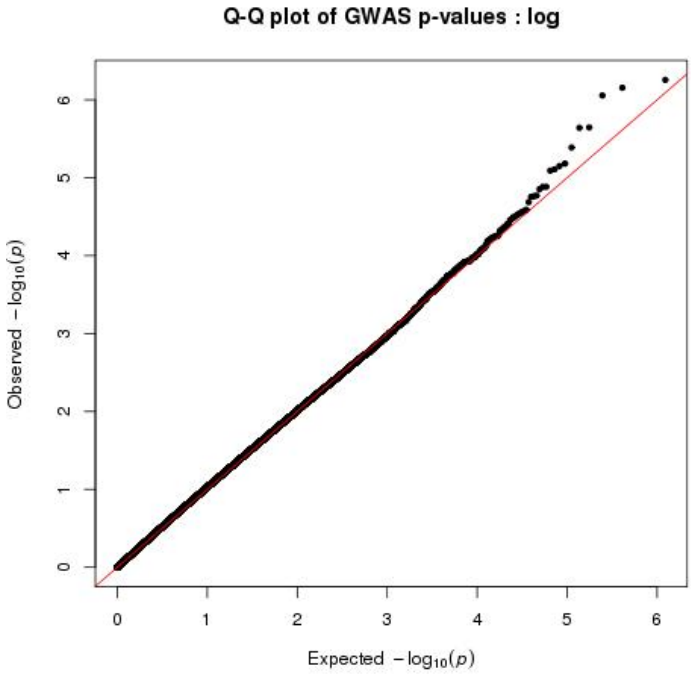

Figure 12: QQ-Plot of logistic analysis

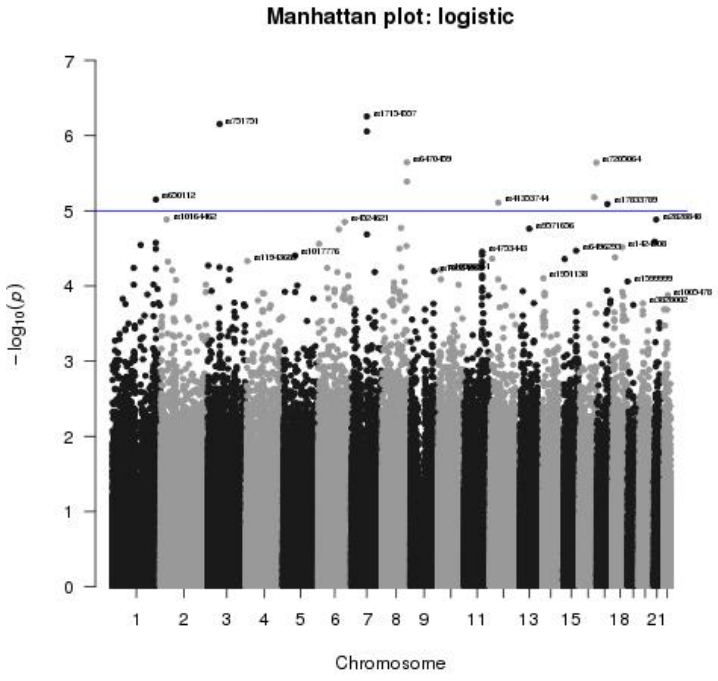

Figure 13: Manhattan plot of logistic analysis

Q-Q plot compares the observed P-vales (Y-axis) with the expected P-values (X-axis, uniformly distributed). In the above Q-Q plot, those SNPs not associated with the LA size should have large observed P-values and they match with the large expected P-values (points fall onto the diagonal solid red line on the left bottom of the plot), which indicates

that our analysis model fits the data well; For those points deviates from the diagonal line on the top right corner, it means these points are possible susceptibility SNPs that are associated with LA size, since their observed P-values are less than the expected ones (or larger than the expected ones in log scale), showing their effects are greater than random effects and suggests that they are probably associated with our phenotype. In the Manhattan plot, even though none of the SNPs appear to have smaller P-values than the significance level, the P-values of some SNPs are small enough to catch our attention and are selected for further investigation. The explanations are similar for the Q-Q plot and the Manhattan plot from the association analysis with Bonferroni correction shown in Figure 14 and 15.

3.3.2 Association Analysis with Bonferroni Correction

The Q-Q plot and Manhattan plot are:

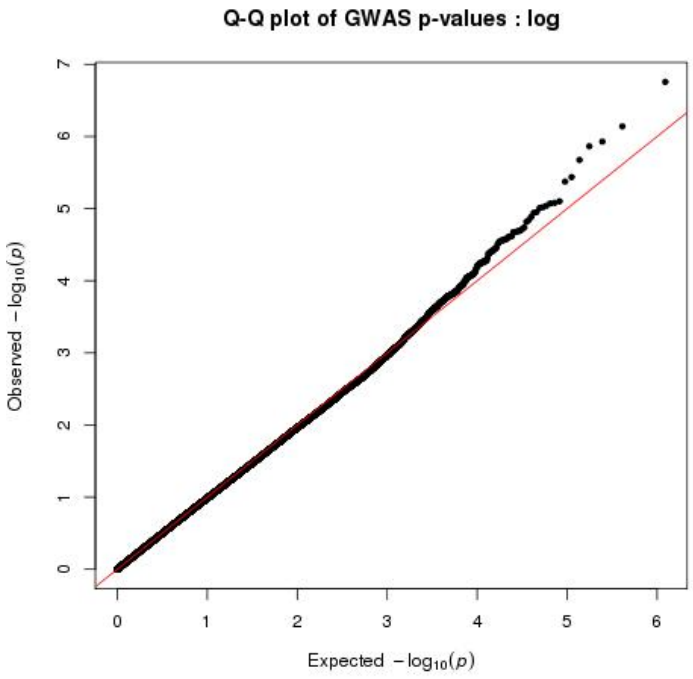

Figure 14: QQ-Plot

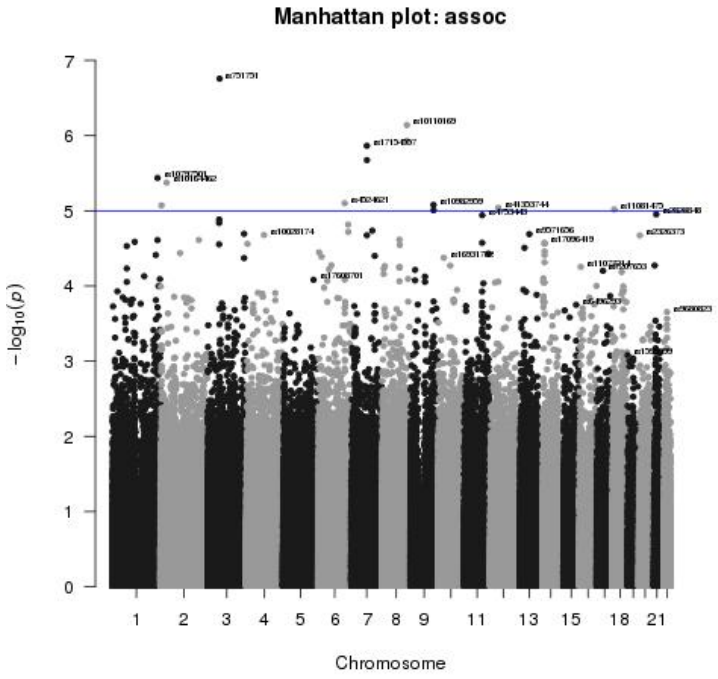

Figure 15: Manhattan plot

3.4 Visualization of Linkage Disequilibrium Blocks

Once significant susceptibility loci were found, the computation of linkage disequilibrium among their neighboring SNPs upstream and downstream and understanding the population haplotype structure were of interest [7]. Direct inference from significantly associated single-nucleotide polymorphisms (SNPs) rarely yields functional variants; More commonly, GWAS hits span a genomic region (GWAS risk loci) that is characterized by multiple correlated SNPs, and may cover multiple closely located genes. Some of these genes could be relevant to the disease, while others are not, yet due to the correlated nature of closely located genetic variants, distinguishing relevant from non-relevant genes is often not possible based on association P-values alone. Identifying the most likely relevant, causal genes and variants requires integrating available information about regional linkage disequilibrium (LD) patterns and functional consequences of correlated SNPs [8].This was done using **LocusZoom** (Haploview also provides such analyses, however, by the time this project was done, its Hapmap database was still not accessible.) Each SNP is color-coded based on the highest r^2 to one of the independent significant SNPs, if that is greater or equal to the threshold. Other SNPs that are below the pre-defined r^2 are colored in grey. The top lead SNPs in genomic risk loci, lead SNPs and independent significant SNPs are circled in black and colored in dark-purple, purple and red, respectively. P-values in $-log(10)$ scale as in Manhattan plot are shown on the left vertical axis, the recombination rates are on the right vertical axis, and the chromosomal positions are on the horizontal axis.

The LD block for rs650112 on chromosome 1 (An intronic variant; DNAH14; a protein-coding gene):

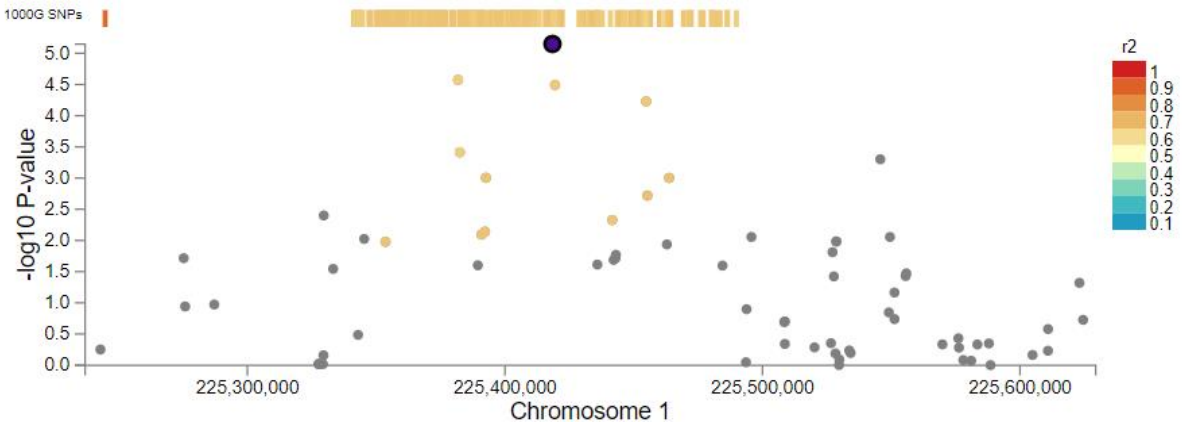

Figure 16: LD block for chr1:225418456

The LD block for rs751751 on chromosome 3 (An intronic variant; PTPRG; a protein-coding gene):

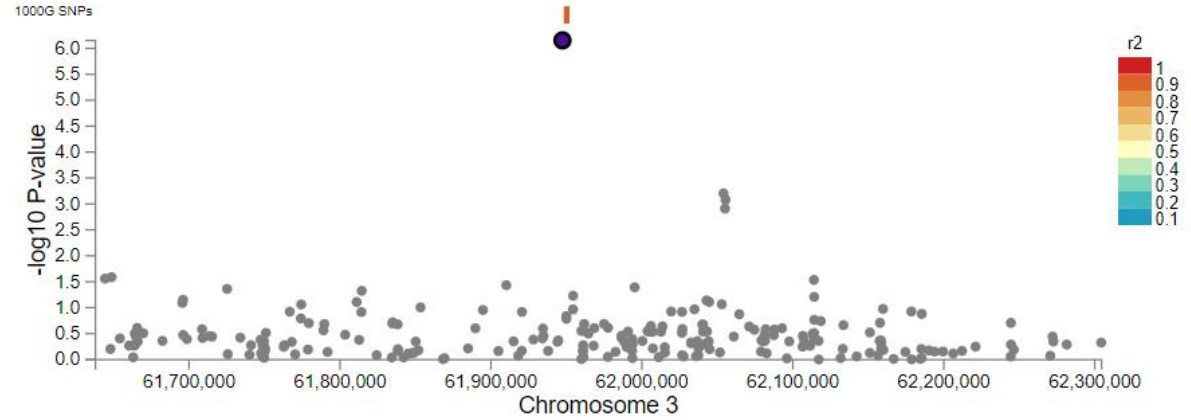

Figure 17: LD block for chr3:61947369

The LD block for rs17154557 on chromosome 7 (intronic variant; SEMA3C; a protein-coding gene):

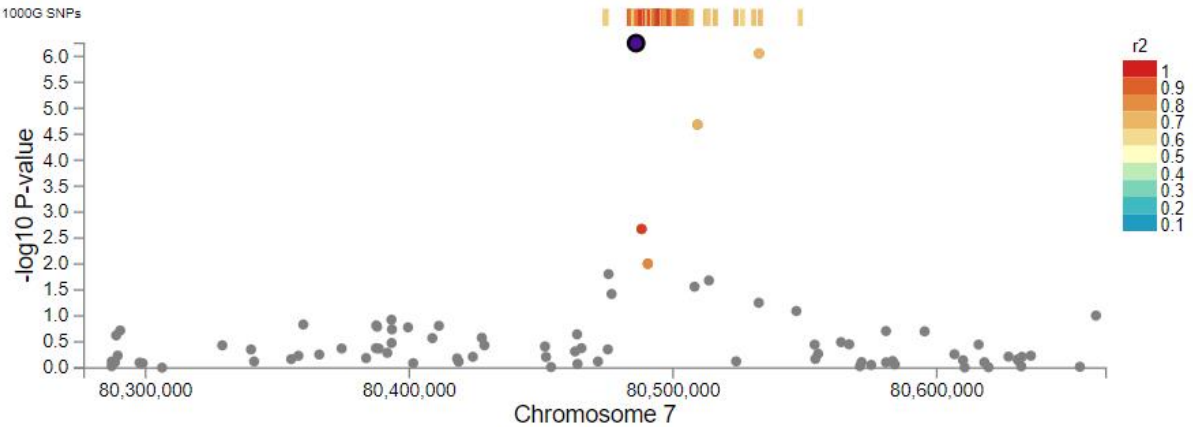

Figure 18: LD block for chr7:80485976

The LD block for rs6470459 on chromosome 8 (intronic variant; PCAT1; a non coding transcript variant):

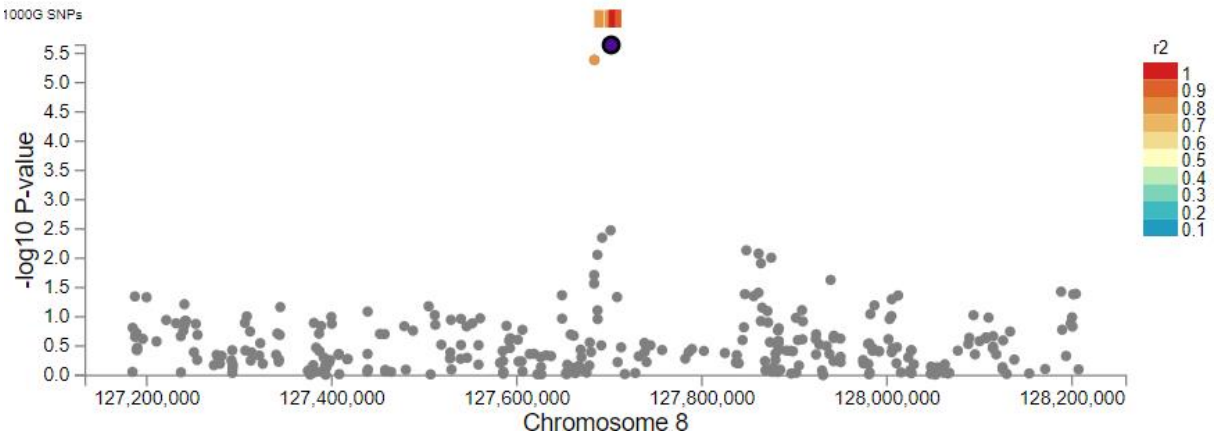

Figure 19: LD block for chr8:127701948

The LD block for rs41353744 on chromosome 12 (intergenic variant):

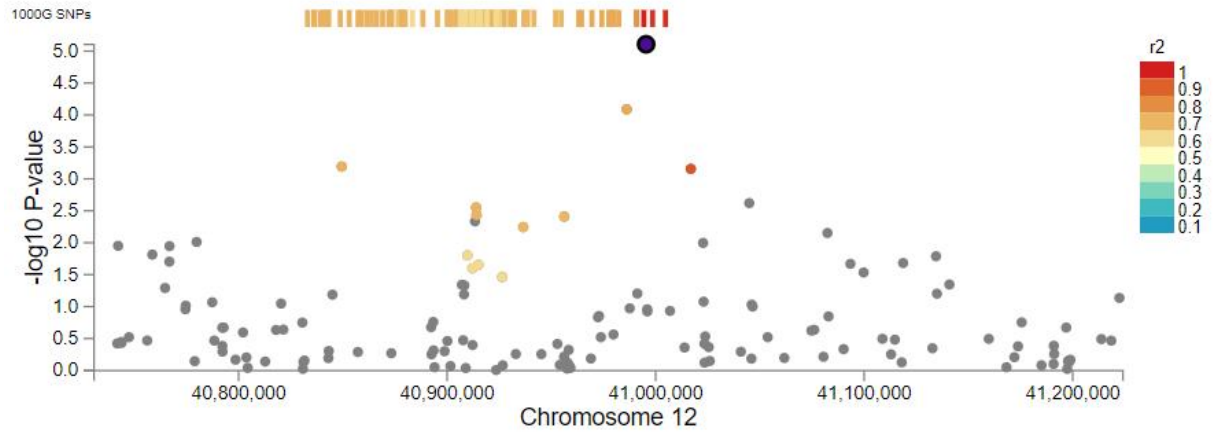

Figure 20: LD block for chr12:40995423

The LD block for rs17771505 on chromosome 16 (intergenic variant):

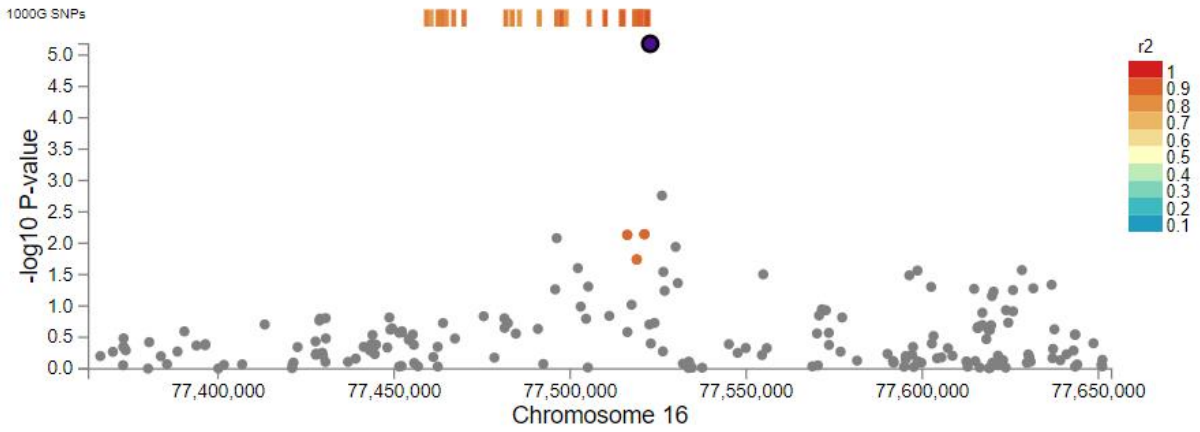

Figure 21: LD block for chr16:77522735

The LD block for rs7205064 on chromosome 16 (intron variant; CBFA2T3; a protein-coding gene):

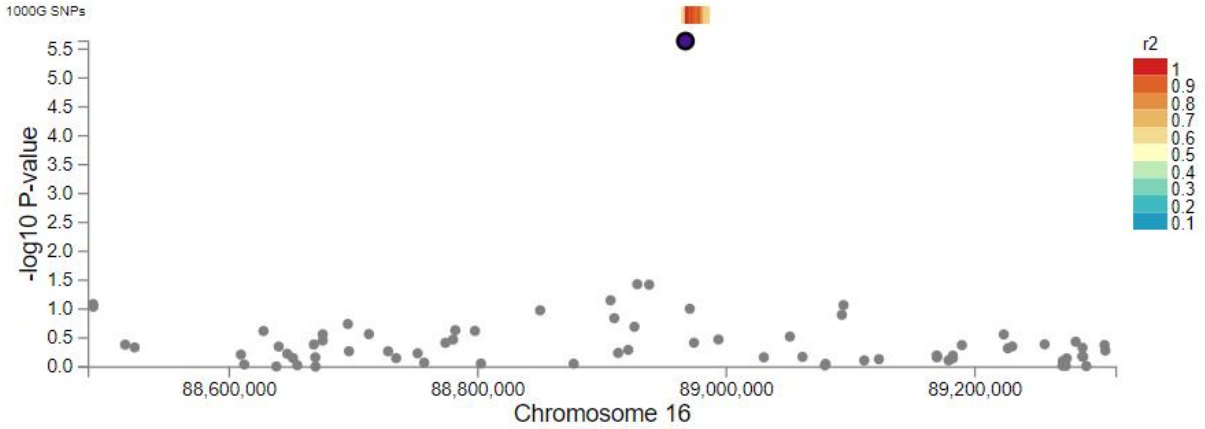

Figure 22: LD block for chr16:88967053

The LD block for rs17833789 on chromosome 17 (no relevant annotation was found for this SNP):

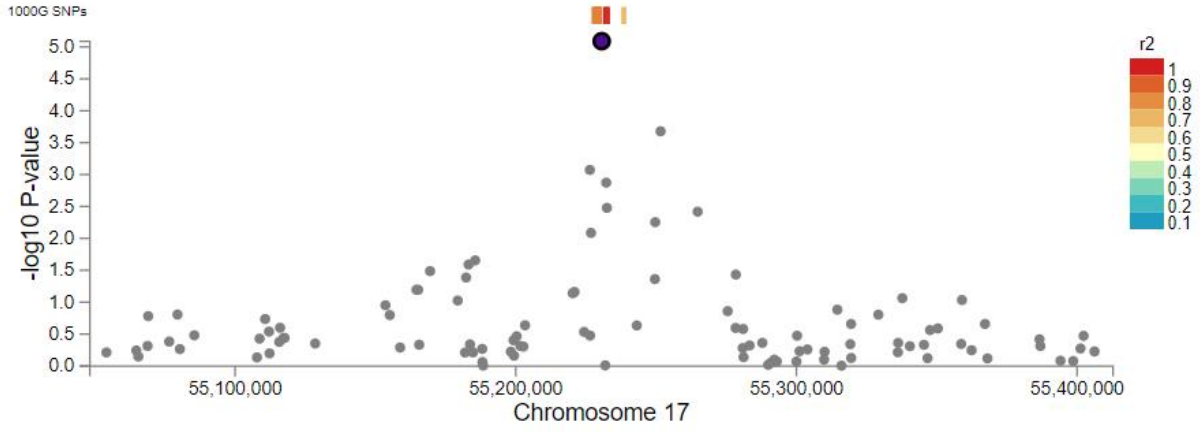

Figure 23: LD block for chr17:55230628

No.	CHR	SNP	A1	Odds Ratio	P-value	Gene
1	1	rs650 112	A	0.4045	7.078×10^{-6}	DNAH14
2	3	rs751 751	G	0.3262	7.009×10^{-7}	PTPRG
3	7	rs17 154 557	A	0.3129	5.557×10^{-7}	SEMA3C
4	8	rs6 470 459	A	0.368	2.267×10^{-6}	PCAT1
5	12	rs41 353 744	A	0.3095	7.801×10^{-6}	NA
6	16	rs17 771 505	G	0.3293	6.597×10^{-6}	NA
7	16	rs7 205 064	A	0.1857	2.29×10^{-6}	CBFA2T3
8	17	rs17 833 789	A	0.3864	8.123×10^{-6}	NA

Table 2: Identified lead SNPs from logistic analysis.

Gene	Summary
DNAH14	Dyneins are microtubule-associated motor protein complexes composed of several heavy, light, and intermediate chains. Two major classes of dyneins, axonemal and cytoplasmic, have been identified. DNAH14 is an axonemal dynein heavy chain (DHC).
PTPRG	The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation.
SEMA3C	Encodes a secreted glycoprotein that belongs to the semaphorin class 3 family of neuronal guidance cues. The encoded protein contains an N-terminal sema domain, integrin and immunoglobulin-like domains, and a C-terminal basic domain.
PCAT1	Produces a long non-coding RNA that promotes cell proliferation and is upregulated in prostate, colorectal, and other cancers.
CBFA2T3	Encodes a member of the myeloid translocation gene family which interact with DNA-bound transcription factors and recruit a range of corepressors to facilitate transcriptional repression.

Table 3: Identified lead SNPs from logistic analysis.

3.5 Susceptibility SNPs

None of the SNPs have passed the adjusted significance level, all of the lead SNPs with relatively low P-values are summarized in Table 2 using results from above, along with their odds-ratios, P-values and mapped genes. A detailed description of the selected intronic genes are given in Table 3.

4 Discussion and Interpretation

Several possible susceptibility genes that could be affecting LA size were identified, such as DNAH14, PTPRG, SEMA3C, PCAT1 and CBFA2T3, none of them coincide with the findings in previous studies; Besides, since we do not have enough power for this study, the results may be more reliable or even different than what we have got if we can instead use a linear association analysis without dichotomizing the phenotype.

5 References

[1]Abhayaratna WP, Seward JB, Appleton CP, et al. Left atrial size: physiologic determinants and clinical applications. J Am Coll Cardiol 2006;47:235763

[2]Leung DY, Boyd A, Ng AA, et al. Echocardiographic evaluation of left atrial size and function: current understanding, pathophysiologic correlates, and prognostic implications. Am Heart J 2008;156:105664

[3]Douglas PS. The left atrium: a biomarker of chronic diastolic dysfunction and cardiovascular disease risk. J Am Coll Cardiol 2003;42:12067

[4]Tsang TS, Barnes ME, Bailey KR, et al. Left atrial volume: important risk marker of incident atrial fibrillation in 1655 older men and women. Mayo Clin Proc 2001;76:46775

[5]Wang L, Di Tullio MR, Beecham A, et al. A comprehensive genetic study on left atrium size in Caribbean Hispanics identifies potential candidate genes in 17p10. Circ Cardiovasc Genet. 2010;3(4):386392. doi:10.1161/CIRCGENETICS.110.938381

[6]Sacco, Ralph Anand, Kishlay Lee, Hye-Seung Boden-Albala, Bernadette Stabler, Sally Allen, Robert Paik, Myunghee. (2004). Homocysteine and the Risk of Ischemic Stroke in a Triethnic Cohort The Northern Manhattan Study. Stroke; a journal of cerebral circulation. 35. 2263-9. 10.1161/01.STR.0000142374.33919.92.

[7]Zeng, Ping Zhao, Yang Qian, Cheng Zhang, Liwei Zhang, Ruyang Gou, Jianwei Liu, Jin Liu, Liya Chen, Feng. (2015). Statistical analysis for genome-wide association study. Journal of biomedical research. 29. 285-97. 10.7555/JBR.29.20140007.

[8]Watanabe, K., Taskesen, E., Bochoven, A. et al. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 8, 1826 (2017) doi:10.1038/s41467-017-01261-5