The Pennsylvania State University

The Graduate School

TRANS-ETHNIC META-ANALYSIS ON LIPID EXOME SEQUENCING DATA

A Thesis in

Bioinformatics and Genomics

by

Yuhuan Cheng

© 2021 Yuhuan Cheng

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2021

The thesis of Yuhuan Cheng was reviewed and approved by the following:

Dajiang Liu
Associate Professor of Public Health Science
Thesis Advisor

Laura Carrel
Associate Professor of Biochemistry and Molecular Biology

Duanping Liao
Professor of Public Health Science

George Perry
Associate Professor of Anthropology and Biology
Chair of the Bioinformatics and Genomics program

ABSTRACT

Genome-wide association studies (GWAS) have provided evidence for associations between genetic variants with lipid levels in human blood that function as risk factors in cardiovascular diseases. Combined with meta-analysis, more connections between genetic variants (including both common variants and rare variants) and complex disease traits can be found in larger sample sizes. In sample sizes of up to 150,000 individuals, we applied and evaluated our proposed approach for performing a meta-analysis of DNA sequence variants association test on their exome sequencing data with six different blood lipid level traits, including high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), non-high-density lipoprotein cholesterol (Non-HDL-C), triglycerides (TG), TG/HDL ratio and total cholesterol (TC). In single variant association tests, we identified 623 variants that met the significance threshold ($P < 5 \times 10^{-8}$), at 10 novel loci and 90 known loci, associated with one or more of four traits (HDL, LDL, TG, TC). We also identified 37 significant loci for Non-HDL trait and 30 significant loci for TG/HDL ratio trait. By comparing to other three methods, including fixed effects model, random effects model and meta-regression model, our results show that our method has an advantage in identifying significant variants, and also rare variants with lower minor allele frequency (MAF<0.01).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor Dr. Dajiang Liu for his relentless support through my master's study; he convincingly guided and encouraged me to do my research even when the road got tough. Without his persistent help, this thesis would not have been completed. I would also like to extend my deepest gratitude to my other two committee members: Dr. Laura Carrel and Dr. Duanping Liao, who gave me constructive advice on this thesis and invaluable insight into the application part of it.

It has been a wonderful journey in Dr. Liu's lab, many thanks to all the lab members, who were always willing to help me out whenever I had questions about research. Especially helpful to me on this thesis was Xingyan Wang, who explained many statistical models to me and provided tons of practical suggestions. I really had the great pleasure of working and learning in Dr. Liu's lab, surrounded by a welcoming and conducive learning environment.

I cannot leave Penn State without mentioning Dr. Cooduvalli Shashikant, who always holds a profound belief in my work and my abilities. I would like to thank him for consistent support and guidance, without whom I would not have made my mind to continue academic in bioinformatics after graduation. Also, I gratefully acknowledge the love and support that I received from my family and friends; special thanks to my boyfriend for staying up late all the nights with me when I was writing this thesis.

Chapter 1

Introduction

GWAS is a useful tool for finding single-nucleotide polymorphisms (SNP) across genome and identifying their associations with traits of particular diseases in the population. The basis of GWAS is linkage disequilibrium (LD), the non-random association of alleles at two or more loci across the human genome in a definite population (*Goode et al., 2011*). Not only some disease risk SNPs can be directly genotyped in the study and found to be statistically associated with the trait, but through indirect association, genotyped SNPs can also be identified as a tag SNP in high LD region with the influential SNP then count for the statistical association with the trait *(Bush et al., 2012).* As a consequence, GWAS has an advantage over linkage analysis for the study of complex diseases, like cardiovascular diseases or age-related macular degeneration (AMD) (*Klein et al., 2005; Fritsche et al., 2016*), because multiple genetic variants across genome contribute to disease risk. Along with the development of large-scale exome sequencing, researchers are able to further focus on rare exonic variants that were hard to be located and studied because of the low allele frequency *(Kiezun et al., 2012; Lee et al., 2014);* they usually have specific functions such as regulating and participating protein synthesis inside the human body, with relatively large effect sizes.

Both clinical and genetic evidence has indicated the causal relationship between blood cholesterol concentrations and cardiovascular disease (*Ference et al., 2017*). In general, cholesterol travel through the bloodstream in lipoprotein particles, and low-density

lipoprotein (LDL) delivers cholesterol to cells inside human body while high-density lipoprotein (HDL) assists in the removal of cholesterol from the bloodstream. As the most common form of lipids in the blood, triglycerides (TG) can attach to HDL then impair its structure. Consequently, higher LDL or TG and lower HDL can cause endothelial dysfunction and cholesterol plaques build-up along with the arterial walls, leading to narrow arteries *(Sobczak et al., 2019)*. Narrow arteries can be damaged with these plaques and more susceptible to blood clots, which increase the risk of myocardial infarction as well as heart attack *(Holmes et al., 2018).*

The majority of common variants at loci associated with lipid levels were uncovered by GWAS and some of them were validated by animal experiments, together supporting that lipid traits including HDL, LDL, TG, Non-HDL and TC are heritable risk factors for cardiovascular diseases *(Global Lipids Genetics Consortium et al., 2013; Kathiresan et al., 2007; Teslovich et al., 2010; The Emerging Risk Factors Collaboration et al., 2009).* Also, an abnormal ratio of triglycerides to HDL-cholesterol (TG/HDL-C) indicates a risk for the development of coronary disease, according to univariate analysis based on clinical data *(da Luz et al., 2008)*; usually high ratio of TG/HDL-C, especially larger than 3, can be seen as a significant risk indicator of heart attack and stroke.

Along with showing the genetic evidence of variants in *PCSK9* gene associated with lower LDL among both African-Americans group and European-Americans group, *Cohen et al.* conducted a 15-year follow-up study reporting that those genetic variants in *PCSK9* conferred protection against coronary heart disease *(Cohen et al., 2006).* Besides, clinical trials showed lowering of low-density lipoprotein cholesterol (LDL-C) by statin therapy can improve cardiovascular outcomes *(Cannon et al., 2015)*. Identification of novel

associations between blood lipids and genetic variants can yield insights on potential pathways or therapeutic targets for cardiovascular diseases; as a reference, PCSK9 inhibitors have already been used as medications to reduce LDL-C especially in selective high-risk patients, who are statin-intolerant or have familial hypercholesterolemia *(Chaudhary et al., 2017).*

Throughout time, GWAS studies for uncovering genetic associations with lipids have been starting with bigger sample sizes from 100,000 to 300,000 and focusing on lower variant frequency from common variants to rare variants, generating a larger number of loci from 95 to 268 (*Teslovich et al., 2010; Klarin et al., 2018*). A previous study based on the exome-chip design (*Liu et al., 2017*) also reported lipoprotein levels altered by genetic loci to have an association with the risk reduction of both coronary heart disease and diabetes. Although different ancestries were included in the previous analysis, most studies primarily relied on fixed effects methods without considering ancestry-specific information while modeling.

Considering that including trans-ethnic information can be helpful to identify genetic variants with large effect size but only occur in specific ancestry, here we perform a meta-analysis on whole-exome sequencing data of up to 150,000 individuals from six different ancestries, which consist of 84,633 Europeans, 10,652 Africans, 4,479 East Asians, 26,033 South Asians, 5,715 Hispanics, and 1,182 Samoans, to identify novel variants associated with six different blood lipid levels on an exome-wide scale. The approach we applied to perform our trans-ethnic meta-analysis combines the strength of all fixed effects model, random effects model and meta-regression model, with the simulation results proved enough power to detect novel associations in heterogeneous populations.

For those variants identified with *p*-values that met genome-wide significance threshold ($P < 5 \times 10^{-8}$), we use one megabase sliding window to define gene locus, which means the one megabase region surrounding the identified variants. We mapped our results to previous study results and recorded the identified novel loci. Next, to evaluate the performance of our model on single variant association study, we compared our results with other three conventional methods, including fixed effects model, random effects model and meta-regression model.

Chapter 2

Methods

Study samples and summary statistics

Fifteen studies from four different cohorts (MIGen, TOPMed, UKBB13k and UKBB50k) contributed association results for whole-exome sequence data and plasma lipid levels. Each contributing cohort analyzed the ancestries within their cohorts separately and studies collected on case/control status analyzed cases separately from the controls (*Liu et al., 2017).* We did multidimensional scaling (MDS) analysis to check the similarity between ancestry of each study with 1000 Genome Project *(The 1000 Genomes Project Consortium., 2015)* as reference. In MDS analysis, we assigned these fifteen studies and five reference ancestries as points to coordinates in 3-dimensional space, following by the calculation of Euclidean distances for all pairs of points to generate the similarity matrix between studies and reference ancestries; then we visualize our results in three graphs (each graph comparing two dimensions) regarding each trait.

For each study, GWAS summary association test statistics were generated from the whole-exome sequencing data using RVTESTS *(Zhan et al., 2016)* by each autosomal chromosome. Both score test statistics files and covariance test statistic file were obtained, and we merged those score test statistics of all 22 autosomal chromosomes within each study as our input score statistics files to perform the meta-analysis. Since our association study focused on autosomal genes, we excluded sex chromosomes data while generating the summary association test statistics.

When looking into the score statistics files, we found that allele frequency value for some variants were missing as NA but with heterozygotes and alternative allele homozygotes recorded. Considering that our model uses allele frequency to calculate the results, we corrected the allele frequency in original score statistics file based on the given information on counts and copies of genotyped alleles by following function:

$$AF = \frac{INFORMATIVE\_ALT\_AC}{2 \times (N_{REF} + N_{HET} + N_{ALT})}$$

where INFORMATIVE_ALT_AC represents the number of alternative alleles in the analyzed samples, and N_{REF} / N_{HET} / N_{ALT} represents the number of samples carrying homozygous reference/ heterozygous/ homozygous alternative alleles.

From the merged score statistics files for each study, we pooled all genotyped variants within each trait into a variant list file that only record the position on chromosome of those variants, in order to run the following meta-analysis in 390 batches for each trait; each batch consists of 50,000 variants.

In addition to the score statistics files, imputation quality files are also required. Considering that imputation files were not given along with the original genotypes study files and higher value of R square quality brings better genotype imputation quality, we imputed all genotyped variants by giving them a simulated R square quality as 1. Both merged score statistics files and imputation quality files were indexed using Tabix *(Li et al., 2011).*

Meta-analysis

We performed meta-analysis using our proposed method, MEMO (Mixed Effect Meta-Regression for Optimal Trans-ethnic Meta-analysis) that is implemented in R software package rareGWAMA *(Liu et al., 2014).* All single variant tests in the meta-analysis for all traits were two-sided. Our full model can be described as follow:

$$b_{jk} = \sum_{l=0}^{L} Z_{lk}\,\gamma_{jl} + e_{jk} + \epsilon_{jk}$$

Where b_{jk} is the genetic effect for jth variant in the kth study. $e_{jk} \sim N(0, \tau^2)$ is the random effect that captures the extra heterogeneity. Z_{lk} is the lth genetic variation (or principle component, PC) for the kth study with $Z_{0k} = 1$. Correspondingly, γ_{jl} captures the effect of the lth axis of genetic variation for the jth variant with γ_{j0} as an intercept in the model. $\epsilon_{jk} \sim N(0, s_{jk}^2)$ is the random error and s_{jk} are standard errors for b_{jk}. In our analysis, we found that the first 3 genetic variations can already separate our study ancestry groups, so we set $L = 3$.

Instead of fitting a single model, we fit a series of models that nested within the full model described above. First, we fitted the model that only contains an intercept. In another word, we excluded all genetic variations as well as the random effect. The hypothesis test is performed to examine if the intercept is 0 or not. Next, we include the first l PCs in the model. Each time we add one more PCs that up to 3. Hypothesis tests are performed for each model to see if all γs are equals to 0. Finally, for the full model, we are interested in testing whether both all γs and the variance of random effect τ^2 equal to 0.

Due to the fact that all these tests are performed on the same data, which implies an unignorable correlation, statistical significance for our final model was calculated using Gaussian copula approach to synthesis all these different models into a single final model.

Besides estimated effect size b_{jk} and its variance estimate s_{jk}^2, we also performed study-specific genomic control (GC) *(Devlin et al., 1999)* based on MAF. To be more precise, we performed GC values control based on both rare variants (MAF < 1%) and common variants (MAF >= 1%) for each participated study.

Additionally, study genetic variation Z_{lk} is calculated using MDS from allele frequency. We defined the genetic distance between 2 studies, i.e. study k and k', with J variants, as:

$$d_{kk'} = \frac{\sum_j \left(f_{jk} - f_{jk'}\right)^2}{J}$$

Where f_{jk} and $f_{jk'}$ are the allele frequency for the jth variant for study k and k', respectively. Both GC values and MDS distances are helpful for population stratification at the meta-analytic level.

To sum up, the input data comprise tabix-indexed score statistics files and imputation quality files, variant list files, also MDS distances and GC values. In terms of summarizing results, we compared our meta-analysis results with commonly used genome-wide p-value threshold of 5×10^{-8}. After identified with p-values that met our statistical significance threshold, variants were sorted in ascending order of their p-value. Then, we summarized results on the level of significant loci to take into consideration of linkage disequilibrium (LD) structure.

Starting from the variant with the lowest p-value, a significant locus is defined as a 1 megabase (MB) region surrounding the variant on human genome. When identified significant variants are close to each other on exome, they likely locate on the same locus or overlapped loci, therefore, we collapsed the overlapped loci into a single locus and recorded the 'sentinel' variant (the variant in the locus with the lowest p-value). All novel loci reported in the result are not within 1 megabase of any previously known *(Klarin et al., 2018)* blood lipid level associated GWAS SNP.

Annotation

For all four lipid traits that have been reported to be associated with previously known GWAS loci, all sentinel variants identified in our method were annotated using ANNOVAR (version 2018Apr16) *(Wang et al., 2010)*. We annotated the variants with Genome Reference Consortium Human Build 38 (GRCh38) to get the gene information (loci), function information (exonic, intronic, UTR5, UTR3, intergenic and ncRNA_exonic) and exonic function information (nonsynonymous or synonymous); we also annotated the variants with avsnp150 database, which is dbSNP150 database with allelic splitting and left-normalization, to obtain the Reference SNP cluster ID (rsID) for each variant. Besides, we annotated the variants with 1000 Genome Project database containing all ancestry (version 1000g2015aug_all) to check the detailed genetic variants information on annotated loci.

Comparison with other methods

To evaluate the performance of our method, we compared it with fixed effects model (FE), random effects model (RE) and our meta-regression model. *P*-values for FE and RE were generated using METASOFT *(Han et al., 2011)*, both FE and conventional RE are based on inverse-variance-weighted effect size. *P*-values for the meta-regression model were shown in our output results.

We plotted the identified significant variants in our method as Manhattan plots and Quantile-quantile plots and annotated the sentinel variants within novel identified loci in other methods' Manhattan plots. After identifying significant variants using same threshold ($P < 5 \times 10^{-8}$), we applied the same definition to count the loci numbers that were identified in FE and RE, the comparison results were shown as a table in the results chapter.

Chapter 3

Results

Calculated allele frequency showed linear regression pattern

Since some missing allele frequency (AF) values were found in the summary statistics files, we calculated the allele frequency for each genotyped variant using the function mentioned in the method chapter. We noticed that the calculated AF didn't match all non-missing AF in the original files exactly, so we generated a scatterplot to check the relationship between non-missing AF in the original files and our calculated AF for 15 studies associated with HDL trait. Figure 3-1 shows the linear regression pattern between original non-missing AF and calculated AF, which means the correction of AF is meaningful and should not bring much deviation to the results. Although it should be noted that a considerable amount of AF values changed to 1.00 after correction, additional variants with retrieved effective AF values were able to participate in our modelling. Consequently, our AF correction reduced the amount of *NaN p*-values found in our results, part of which were caused by missing AF, by 25%.

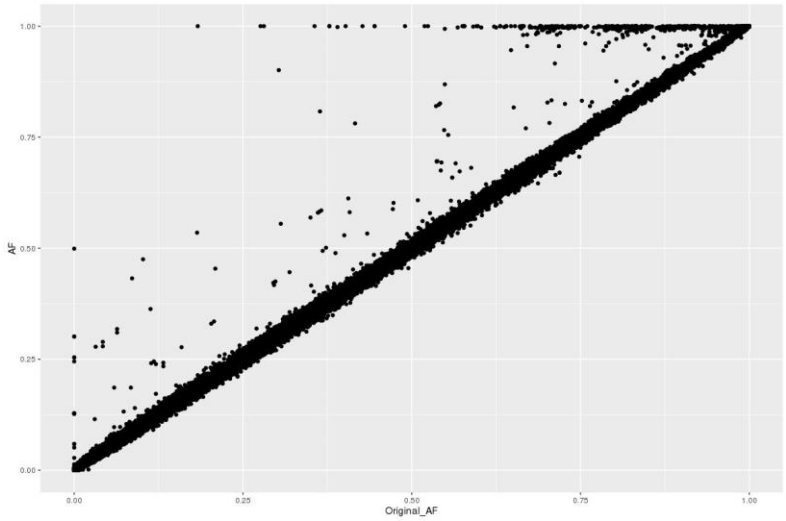
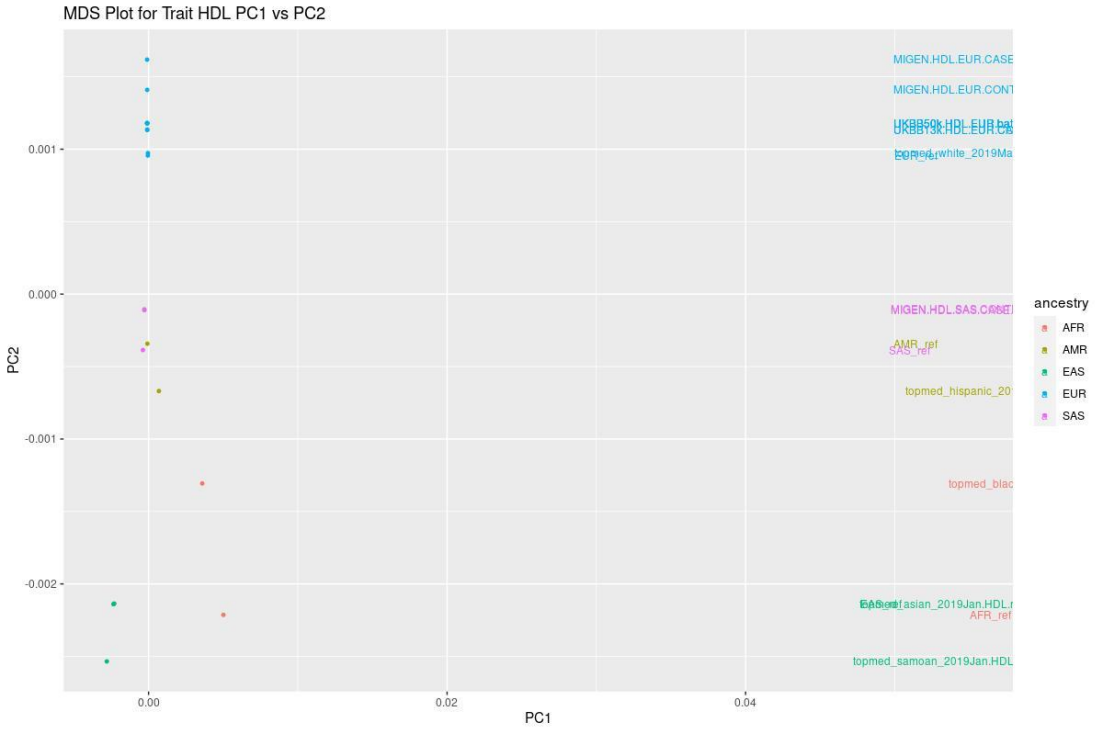
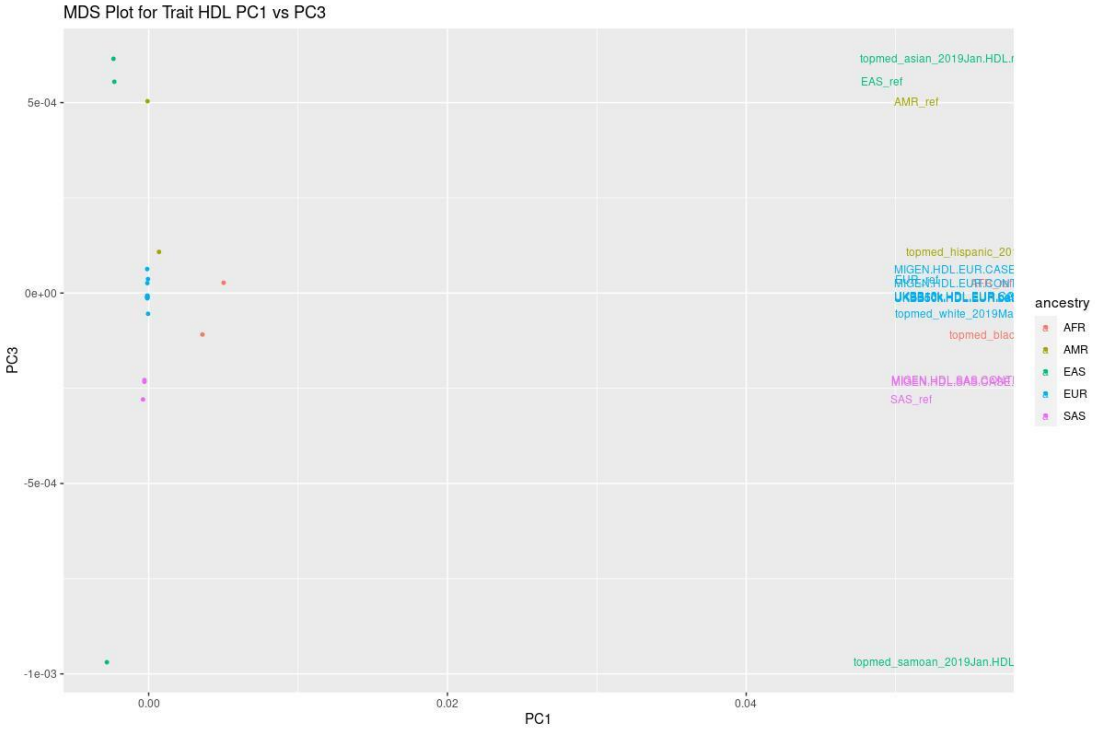


Figure **3-1**: Scatterplot showing relationship between original AF (x-axis) and calculated AF (y-axis) for HDL trait associated studies.

MDS analysis checked the population heterogeneity in study samples

Given the study samples across six different ancestries, we performed MDS analysis within each trait to check the population stratification with reference ancestries in 3-dimensions and visualized our results as stated in the method chapter. Figure 3-2 shows the calculated distance between fifteen study samples associated with HDL trait and five 1000G reference ancestries. In general, our MDS results confirmed the ancestry of each study, although Samoan ancestry has not yet been recorded in 1000G reference ancestry, the study samples from Samoan ancestry only include 1,182 individuals, which is a relatively small sample size. Therefore, we believe the lack of Samoan reference ancestry should not influence our results to a recognizable degree.


MDS Plot for Trait HDL PC1 vs PC2
PC2
PC1
0.001
0.000
-0.001
-0.002
0.00
0.02
0.04
MIGEN.HDL.EUR.CASE
MIGEN.HDL.EUR.CONT
topmed_white_2019Ma
AMR_ref
topmed_hispanic_20
topmed_blac
AFR_ref
topmed_samoan_2019Jan.HDL
ancestry
AFR
AMR
EAS
EUR
SAS


MDS Plot for Trait HDL PC1 vs PC3
PC3
PC1
5e-04
0e+00
-5e-04
-1e-03
0.00
0.02
0.04
EAS_ref
AMR_ref
topmed_hispanic_20
MIGEN.HDL.EUR.CASE
topmed_white_2019Ma
topmed_blac
SAS_ref
topmed_samoan_2019Jan.HDL
ancestry
AFR
AMR
EAS
EUR
SAS

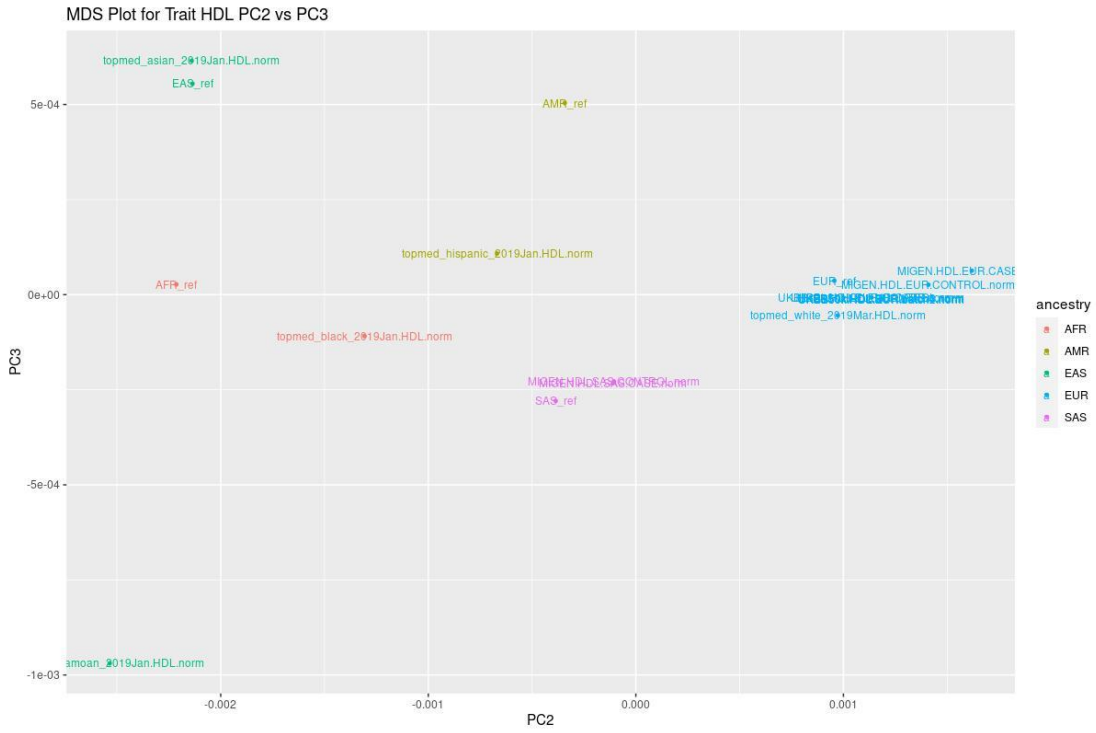


Figure **3-2**: MDS plots for HDL trait.

Each plot compares two PCs from the three-dimensional analysis results, in the order of PC1 vs PC2, PC1 vs PC3 and PC2 vs PC3. Different colors represent different ancestry groups. The distance between each point shows the similarity between studies or reference. Overlapped labels/points represent case/control studies or studies of different batches from same ancestry within same cohort.

Quantile-quantile plots showed enough power to detect association

We generated quantile-quantile (QQ) plots for the *p*-values of our method to check the quantile distribution of our observed p-values versus the quantile distribution of expected p-values (where both have been –log transformed). Figure 3-3 shows the quantile-quantile plots for our method were well behaved, the tail in our QQ plots shows our method has enough power to detect positions of causal polymorphisms for each trait.

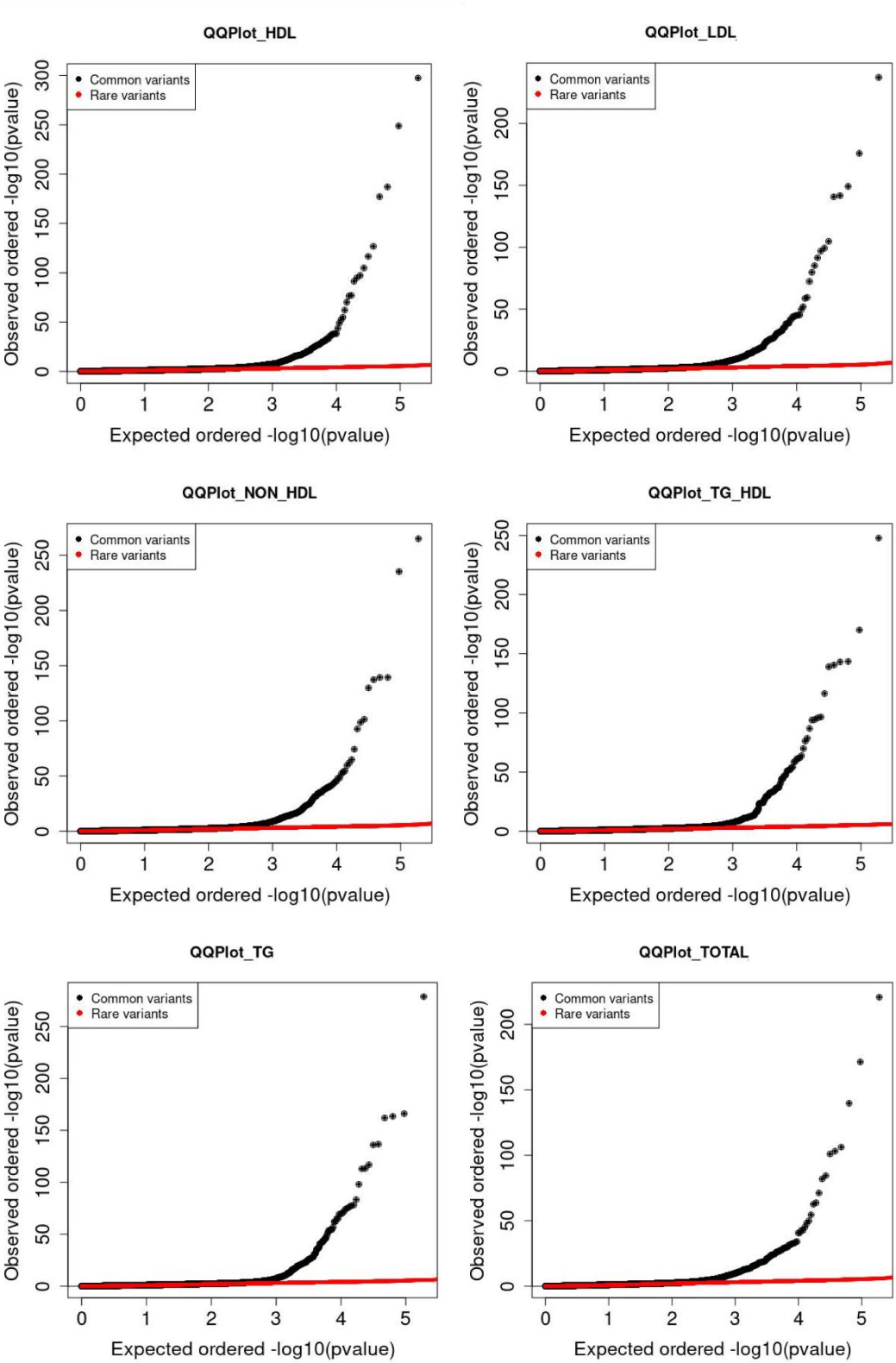
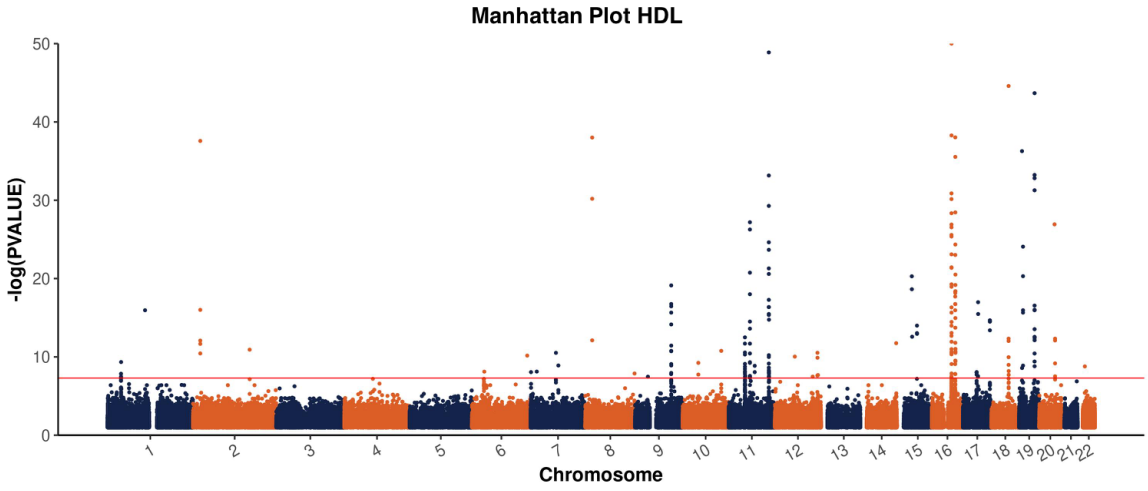
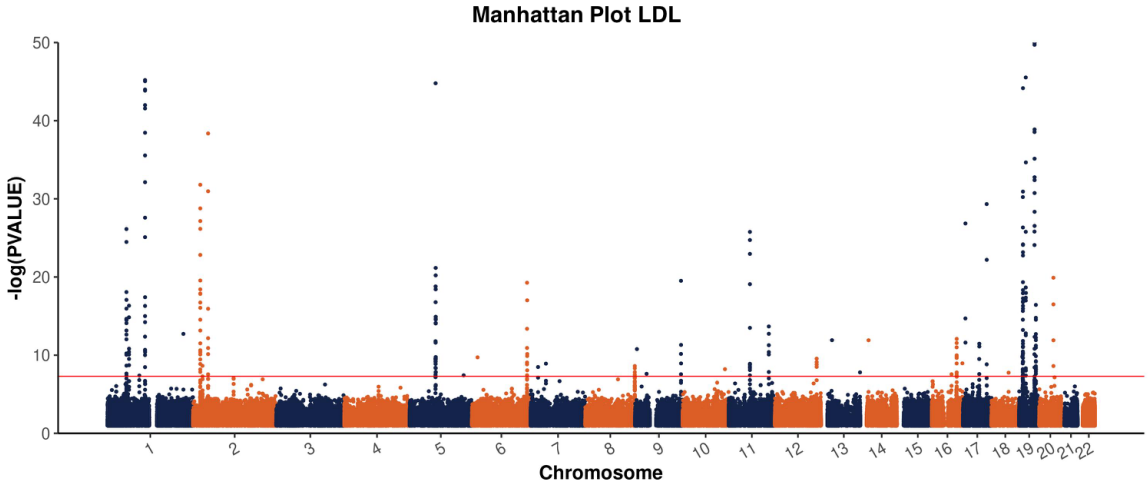

QQPlot_HDL
Common variants
Rare variants
Observed ordered -log10(pvalue)
Expected ordered -log10(pvalue)
QQPlot_LDL
Common variants
Rare variants
Observed ordered -log10(pvalue)
Expected ordered -log10(pvalue)
QQPlot_NON_HDL
Common variants
Rare variants
Observed ordered -log10(pvalue)
Expected ordered -log10(pvalue)
QQPlot_TG_HDL
Common variants
Rare variants
Observed ordered -log10(pvalue)
Expected ordered -log10(pvalue)
QQPlot_TG
Common variants
Rare variants
Observed ordered -log10(pvalue)
Expected ordered -log10(pvalue)
QQPlot_TOTAL
Common variants
Rare variants
Observed ordered -log10(pvalue)
Expected ordered -log10(pvalue)

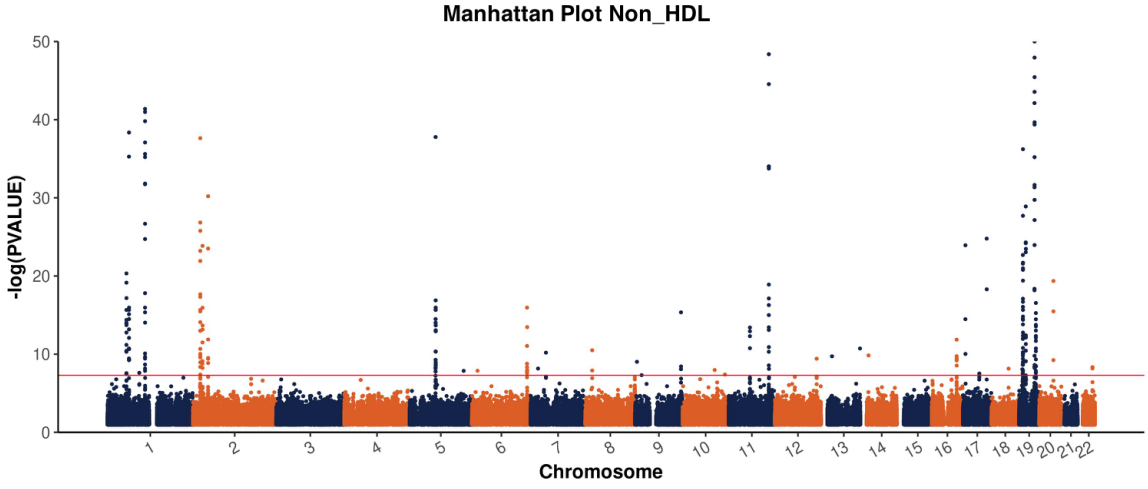
Figure **3-3**: Quantile-quantile plots for *p*-value of our meta-analysis results for six blood lipid traits.
x-axis represents –log transformed expected *p*-value, and y-axis represents –log transformed observed *p*-value. Black line represents the distribution for common variants and red line represents the distribution for rare variants.

Manhattan plots showed significant variants identified in our method

Compared with commonly used genome-wide p-value threshold of 5×10^{-8}, our method identified 220 significant variants associated with HDL trait, 292 significant variants associated with LDL trait, 267 significant variants associated with Non_HDL trait, 209 significant variants associated with TG_HDL trait, 214 significant variants associated with TG trait and 309 significant variants associated with TC trait. Considering the large number of variants genotyped, Manhattan plots were generated for each trait to display significant SNPs across the genome. The higher of genetic variants locate in the Manhattan plots, the more likely of the its association with given trait. Figure 3-4 shows the peak of SNPs identified in our method, and to display the peak more clearly, it only shows part of identified significant variants but omitting those with extremely low p-values. Loci were defined from the pool of our identified significant SNPs.


Manhattan Plot HDL
-log(PVALUE)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


Manhattan Plot LDL
-log(PVALUE)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


Manhattan Plot Non_HDL
-log(PVALUE)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

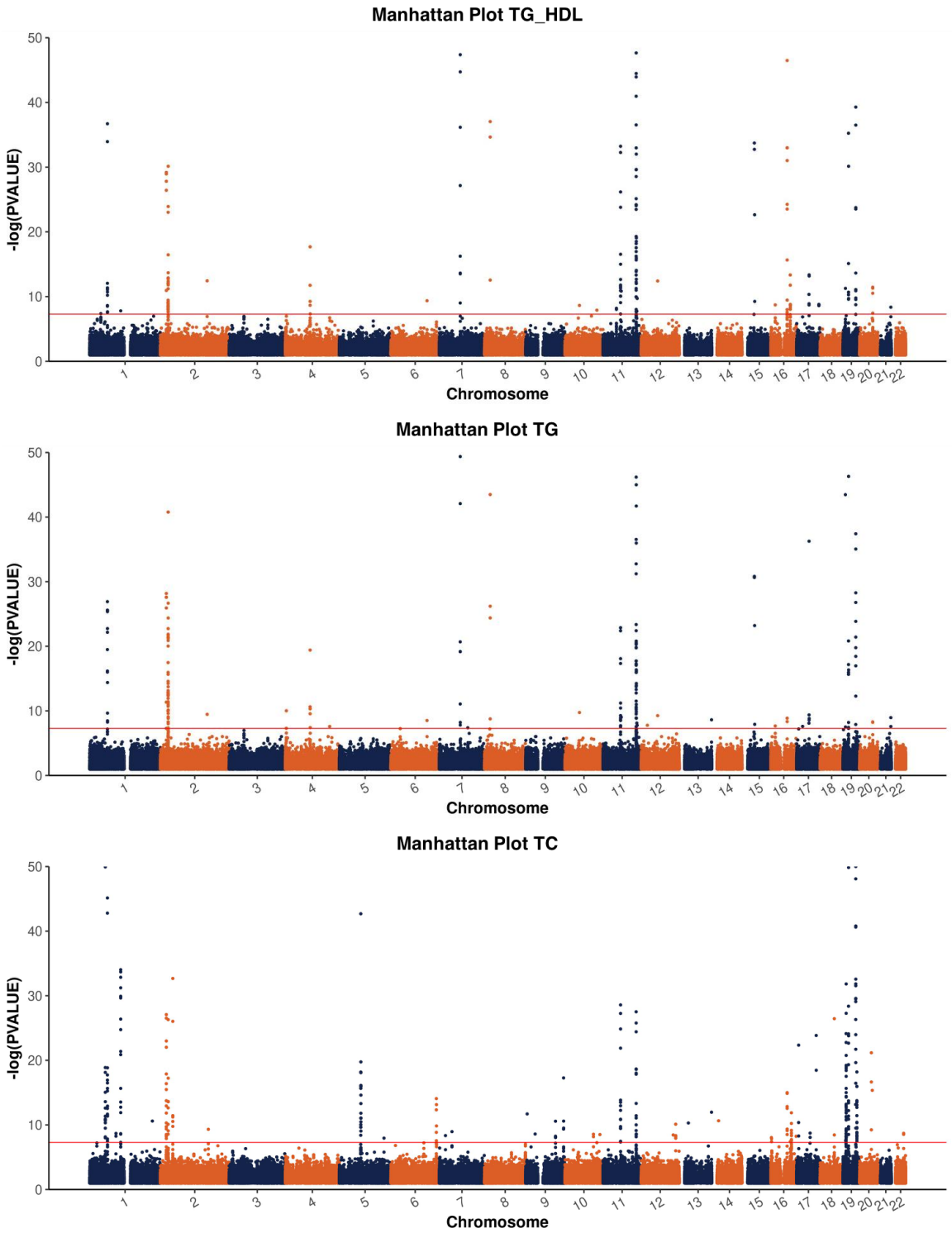


Figure **3-4**: Manhattan plots for *p*-value of our meta-analysis results for six blood lipid traits.

Each point represents a genetic variant genotyped from the study samples. The x axis shows its position on autosomal chromosome, and the y axis represents –log transformed *p*-value of our results, range from 0 to 50. Red line represents our significance p-value threshold of 5×10^{-8}.

Ten Novel loci identified by our method

From the pool of our identified significant SNPs, we defined 106 loci surrounding the sentinel variants using the definition as described in the method chapter. For four traits (HDL, LDL, TG, TC), there were 386 associated loci reported in previous GWAS study (*Klarin et al., 2018*). After annotation with GRch38, we identified 13 potential novel loci associated with one or more of these four traits, and 87 of 386 known loci associated with these four traits, showing the reliability of our results. Considering *Klarin et al., 2018* may not cover the latest associations, we checked with the updated data on GWAS Catalog (*Buniello et al., 2019*). As a result, we found 3 of 13 these loci were reported associated with LDL or TG in previous studies *(Richardson et al., 2020; Ripatti et al., 2020)*. In conclusion, our method identified 10 novel loci in total. Table 3-1 summarized the novel loci with statistics results and gene information; 6 out of 10 sentinel variants were annotated as nonsynonymous variants across exome, resulting in malformed or dysfunctional protein products. It's worth noting that half of these nonsynonymous variants are rare variants (AF < 0.01).

Table **3-1**: Novel significant loci identified for four traits.

HDL						
Chr: Pos	**rsid**	**Ref/Alt**	**AF**	**P-value**	**Gene**	**Annotation**
9:36966703	.	G/A	2.43E-05	3.38E-08	PAX5	nonsynonymous
17:39905964	rs2305479	C/T	0.439	8.74E-09	GSDMB	nonsynonymous
20:45941291	rs6065908	C/T	0.183	4.69E-13	PCIF1	intronic
LDL						
Chr: Pos	**rsid**	**Ref/Alt**	**AF**	**P-value**	**Gene**	**Annotation**
9:33113783	rs551564683	T/C	0.000822	2.43E-08	B4GALT1	nonsynonymous
19:9380097	rs141144143	G/A	3.41E-05	1.60E-10	ZNF177; ZNF559-ZNF177	nonsynonymous
TG						
Chr: Pos	**rsid**	**Ref/Alt**	**AF**	**P-value**	**Gene**	**Annotation**
7:100113754	rs143637758	G/C	4.13E-05	4.15E-08	TAF6	nonsynonymous
11:119019214	.	C/T	9.28E-05	7.74E-09	TRAPPC4	synonymous
11:120327605	rs760568794	G/A	0.000721	1.68E-08	TLCD5	nonsynonymous
17:21214278	rs61121321	G/C	0.0107	2.49E-08	TMEM11	UTR5
20:45947863	rs7679	T/C	0.184	5.16E-09	PCIF1	UTR3
TC						
Chr: Pos	**rsid**	**Ref/Alt**	**AF**	**P-value**	**Gene**	**Annotation**
9:33113783	rs551564683	T/C	0.000803	2.56E-09	B4GALT1	nonsynonymous
16:1587897	rs748463014	C/T	0.000118	9.09E-09	LOC105-371046	ncRNA_intronic
19:9380097	rs141144143	G/A	3.28E-05	2.73E-08	ZNF177; ZNF559-ZNF177	nonsynonymous

Chr: Pos shorts for chromosome: position; Ref/Alt shorts for reference allele/alternative allele; Annotation describes the function of given variant. The dot in rsid column represents NA string.

Among the nonsynonymous variants we identified that associated with lipid levels, *PAX5* was reported to be associated with coronary artery disease (CAD) in type 1 diabetes (T1D) European ancestry cases *(Charmet et al., 2018). Charmet et al., 2018* identified variant rs143723948 at the UTR5 region of *PAX5* with *p*-value 6.01×10^{-7}, showing significance at their *p*-value threshold of 1×10^{-5}. Consistently, our result suggested the association between variant 9:36966703 at *PAX5* annotated as a nonsynonymous exonic variant and HDL, with significant *p*-value 3.38×10^{-8}. Therefore, our result provided a potential explanation for the association between *PAX5* with CAD in T1D patients by demonstrating evidence for a suggestive association between *PAX5* with HDL.

Additionally, another novel locus *TLCD5* associated with TG that showed significance in our result, also known as *TMEM136*, was reported to be associated with CAD by a trans-ancestry meta-analysis on Japanese and European populations *(Koyama et al., 2020). Koyama et al., 2020* identified variant rs1893261 at *TMEM136* as a synonymous variant with significant *p*-value 3.1×10^{-8}, suggesting its association with CAD. The nonsynonymous variant rs760568794 at TLCD5 that was identified associated with TG (*p*-value 1.68×10^{-8}) in our result is consistent with the previous finding. Thus, our result indicated the interrelationship of TG associated variant at locus *TLCD5* with CAD.

Besides, we also identified a novel locus *GSDMB* associated with HDL, which was reported to be associated with T1D in African American cases *(Onengut-Gumuscu et al., 2019).* Previous epidemiological studies have provided evidence that CAD was 4 times more prevalent in T1D patients *(Lind et al., 2014).* Combined with our result suggesting

that the nonsynonymous variant rs2305479 at *GSDMB* is associated with HDL, it can be seen that *GSDMB* may be a risk factor for CAD in T1D.

Comparison with other methods

To evaluate the performance of our method, we looked into the p-value for meta-regression model that nested in our method, also for fixed effects (FE) model and random effects (RE) model, both of which were performed by METASOFT as described in the method chapter. Then we counted the number of identified loci for each method using the same definition. Since our method fitted all these three models, we expected to see our method have a combined strength over each conventional simple model. Table 3-2 summarized the significant loci associated with each trait for each method, and our MEMO method shows advantage on identifying significant loci over all other three methods. Exceptionally, we noticed that FE identified two more loci associated with TC trait than our MEMO method, but after annotation, we found that those two loci were both known loci on *GPAM* (rs2254537) and *FRK* (rs3756772). Overall, our method prevails over other three methods on identifying novel loci.

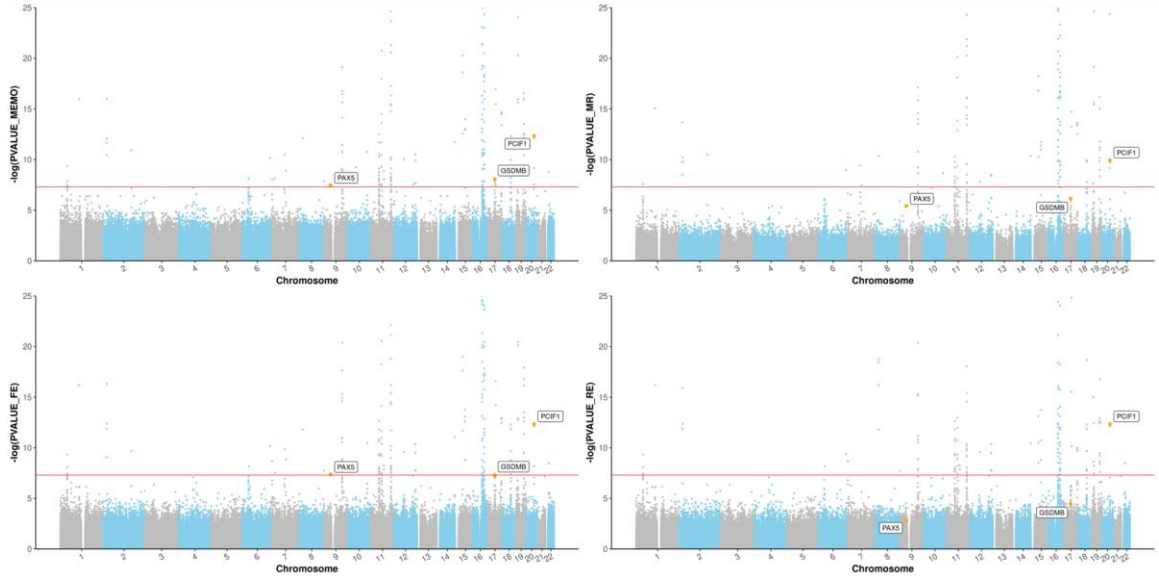
Table **3-2**: The number of significant loci identified by each method.

MEMO is our proposed method, MR_MEGA is meta-regression model. For all four methods, we used the same significance *p*-value threshold of 5×10^{-8}.

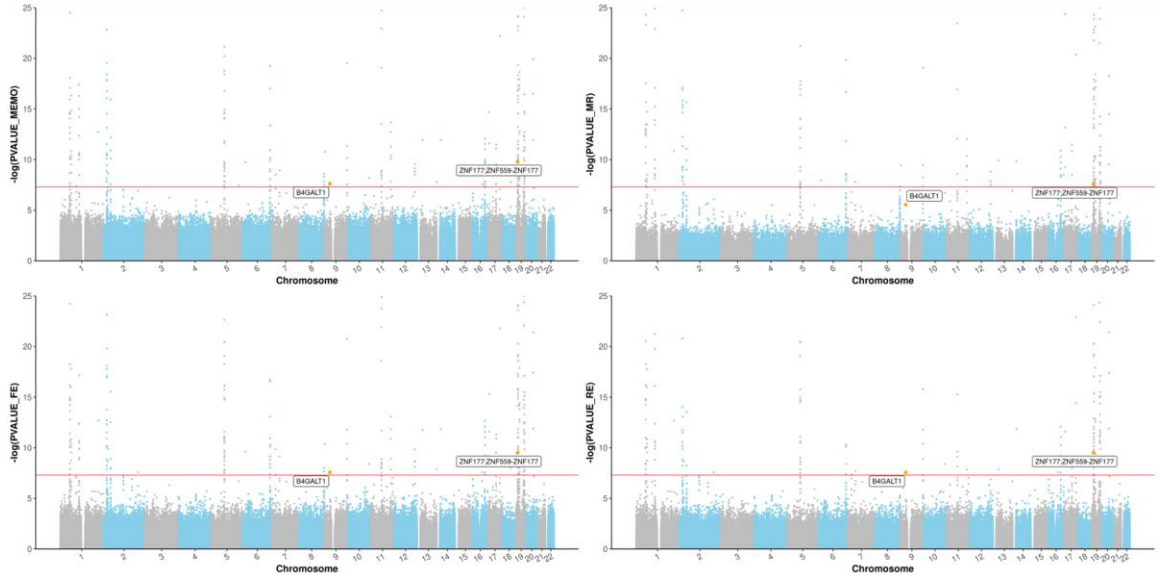
Methods	Loci Number						
	HDL	LDL	Non_HDL	TG_HDL	TG	TC	Total
MEMO	39	38	37	30	30	44	218
MR_MEGA	28	30	27	22	17	34	158
FE	38	37	37	30	28	46	216
RE	30	30	28	28	22	41	179

Focusing on our identified novel loci, Manhattan plots with identified novel sentinel variants annotation were generated for all four methods and displayed side by side in 2× 2 order. Consistent with the difference in counted loci number shown in Table 3-2, bunches of novel loci didn't meet the same significance threshold in both meta-regression model and random effects model. Besides, fixed effects model failed to identify the novel locus *GSDMB* with p-value as 5.37E-08, while our method showed significance of the sentinel variant on *GSDMB* locus, annotated as a nonsynonymous exonic variant associated with HDL trait, with p-value as 8.74E-09. It can be seen from the comparison with other three methods that our method has the advantage to suggest novel associations between genetic variants and blood lipid levels.

Manhattan Plot all methods HDL



Manhattan Plot all methods LDL



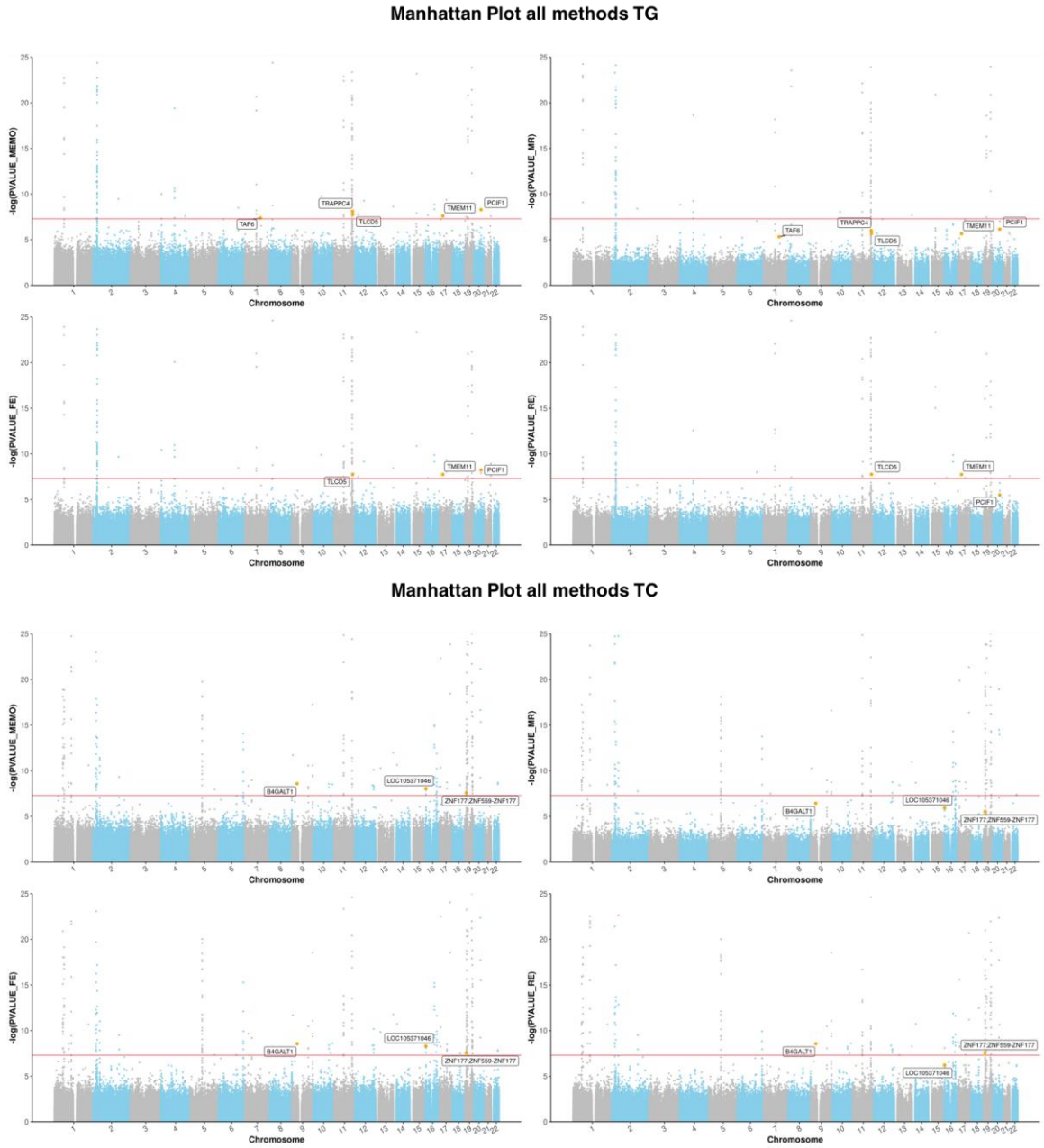


Figure **3-5**: Manhattan plots for all four methods with identified novel loci annotation.

For each trait, top left shows result of our method MEMO, top right shows result of meta-regression model (MR_MEGA), bottom left shows result of fixed effects model (FE), bottom right shows result of random effects model (RE). Orange points represent identified

novel sentinel variants in our method, labeled by the loci name. The x axis shows the position on autosomal chromosome, and the y axis represents –log transformed *p*-value range from 0 to 25. Red line represents same significance p-value threshold of 5×10^{-8}.

Chapter 4

Discussion

We leveraged whole-exome sequencing data of up to 150,000 individuals from multi-ethnic cohorts to investigate genetic variants' associations with blood lipids. Our investigation robustly confirmed 90 previously reported loci associated with one or more of four lipid traits (HDL, LDL, TG, TC), demonstrating our proposed method's reliability. Simultaneously, we uncovered an additional ten novel genome-wide significant loci by our approach, following by comparing the performance with other three statistical models that include fixed effects model, random effects model, and meta-regression model. The result showed that our method has the advantage over all these models on identifying novel associations with statistical significance.

Among the novel loci we identified that have suggestive associations with lipid traits, we found three of them have been reported to be associated with coronary artery diseases or type 1 diabetes. Given the incontrovertible evidence that lipids play a causal role in cardiovascular diseases, our study should provide more reliable insights into the underlying mechanism of cardiovascular diseases by uncovering the novel genetic associations with blood lipids.

Also, it can be seen from the comparison results that our method outperformed three conventional models in single variant association tests. As a powerful statistical tool, our approach also has potential applications in other GWAS studies to identify novel genetic associations on a genome-wide scale with different complex diseases.

An interesting finding in our results is that some of our identified significant sentinel variants were located in introns or UTR regions, even though the original sequence was genotyped with an exome-focused array. The reason why intronic variants are included in the exome sequencing data can be explained as the sequencing byproduct. While exome sequencing primarily targets exons, it may capture some noncoding regions such as introns, intron-exon boundary regions, UTRs, and intergenic regions concurrently *(Guo et al., 2012)*. According to previous research, information concerning these off-target regions obtained from exome sequencing data also contributed to genetic epidemiology studies *(Romasko et al., 2018).*

Our method also identified 37 significant loci for Non-HDL trait and 30 significant loci for TG/HDL ratio trait, showing the same performance as the fixed effects model. Some risk-evaluation and risk-modeling studies *(Brunner et al., 2019; Hajian-Tilaki K et al., 2020)* have provided evidence supporting the associations between phenotypes of Non_HDL and TG/HDL ratio with cardiovascular diseases. But the genetic basis of the associations remains unclear; our findings suggest the feasibility of our method to screen genetic variants with statistical significance.

Previous GWAS studies often focus on common variant associations, given the challenge of identifying rare variants across the human genome. Even though occurring with small allele frequencies in populations, rare coding variants can help understand complex diseases with their large effect size on human genes. Our research shed light on nominating putative targets that may be repurposed for treating cardiovascular diseases by uncovering novel associations between rare variants and lipid traits.

As a result of single variant level association tests, our study can be followed by a series of gene-level association tests. By collapsing multiple rare variants in a gene as gene burden, we can increase the statistical power of regression analysis, and therefore some signals driven by very rare variants (frequency < 0.05%) can be recovered (*Liu et al., 2014*). Another direction of the future work can be fine-mapping, which can specify and prioritize the causal genetic variants identified by our approach associated with lipids traits. The identification of candidate causal variants by statistical fine-mapping will help understand the genetic basis of human complex disease and also allow drug target for precision medicine in the future.

References

- Brunner FJ, et al. Application of non-HDL cholesterol for population-based cardiovascular risk stratification: results from the Multinational Cardiovascular Risk Consortium. The Lancet. 2019; 394(10215): 2173-2183.
- Buniello A, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019; 47(D1):D1005-D1012. [PubMed: 30445434]
- Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822. Epub 2012 Dec 27. [PubMed: 23300413]
- Cannon CP, et al. Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. N Engl J Med. 2015;372(25):2387-97. [PubMed: 26039521]
- Charmet R, et al. Novel risk genes identified in a genome-wide association study for coronary artery disease in patients with type 1 diabetes. Cardiovasc Diabetol. 2018;17(1):61. [PubMed: 29695241]
- Chaudhary R, et al. PCSK9 inhibitors: A new era of lipid lowering therapy. World J Cardiol. 2017; 9(2):76-91. [PubMed: 28289523]
- Cohen JC, et al. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006; 354(12):1264-72. [PubMed: 16554528]

- da Luz PL, et al. High ratio of triglycerides to HDL-cholesterol predicts extensive coronary disease. Clinics (Sao Paulo, Brazil). 2008; 63(4):427-32. [PubMed: 18719750]
- Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997-1004. [PubMed:11315092]
- Fritsche LG, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet. 2016; 48(2):134-143. [PubMed: 26691988]
- Global Lipids Genetics Consortium. et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013; 45:1274–83. [PubMed: 24097068]
- Guo Y, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 2012; 13:194. [PubMed: 22607156]
- Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet. 2011;88(5):586-598. [PubMed: 21565292]
- Hajian-Tilaki K, et al. Triglyceride to high-density lipoprotein cholesterol and low-density lipoprotein cholesterol to high-density lipoprotein cholesterol ratios are predictors of cardiovascular risk in Iranian adults: Evidence from a population-based cross-sectional study. Caspian J Intern Med. 2020; 11(1):53-61. [PubMed: 32042387]

- Holmes MV, et al. Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. J Am Coll Cardiol. 2018; 71(6):620-632. [PubMed: 29420958]
- Koyama S, et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. Nat Genet. 2020; 52(11):1169-1177. [PubMed: 33020668]
- Kathiresan S, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Med Genet. 2007; 8(1):S17. [PubMed: 17903299]
- Kiezun A, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012; 44:623–30. [PubMed: 22641211]
- Klarin D, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018; 50: 1514-1523. [PubMed: 30275531]
- Klein RJ, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308(5720):385-389. [PubMed: 15761122]
- Lee, S, et al. Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 2014; 95:5–23. [PubMed: 24995866]
- Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011; 27(5): 718-719. [PubMed: 21208982]
- Lind M, et al. Glycemic control and excess mortality in type 1 diabetes. *N Engl J Med.* 2014; **371**:1972–1982. [PubMed: 25409370]

- Liu DJ, et al. Exome-wide association study of plasma lipids in >300,000 individuals. Nat Genet. 2017;49(12):1758-1766. [PubMed: 29083408]
- Liu DJ, et al. Meta-analysis of gene-level tests of rare variant association. Nat Genet. 2014; 46: 200-204. [PubMed: 24336170]
- Onengut-Gumuscu S, et al. Type 1 Diabetes Risk in African-Ancestry Participants and Utility of an Ancestry-Specific Genetic Risk Score. Diabetes Care. 2019; 42(3):406-415. [PubMed: 30659077]
- Richardson TG, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLoS Med. 2020;17(3):e1003062. [PubMed: 32203549]
- Ripatti P, et al. Polygenic Hyperlipidemias and Coronary Artery Disease Risk. Circ Genom Precis Med. 2020;13(2):e002725. [PubMed: 32154731]
- Romasko EJ, et al. Utility and limitations of exome sequencing in the molecular diagnosis of pediatric inherited platelet disorders. Am J Hematol. 2018; 93(1):8-16. [PubMed: 28960434]
- Shi J, Lee S. A novel random effect model for GWAS meta-analysis and its application to trans-ethnic meta-analysis. Biometrics. 2016; 72(3):945-54. [PubMed: 26916671]
- Sobczak AIS, Stewart AJ. Coagulatory Defects in Type-1 and Type-2 Diabetes. Int J Mol Sci. 2019; 20(24):6345. [PubMed: 31888259]

- Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–13. [PubMed: 20686565]
- The Emerging Risk Factors Collaboration. et al. Major lipids, apolipoproteins, and risk of vascular disease. J. Am. Med. Assoc. 2009; 302:1993–2000. [PubMed: 19903920]
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526: 68-74. [PubMed: 26432245]
- Wang K, et al. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164. [PubMed: 20601685]
- Zhan X, et al. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016; 32:1423–1426. [PubMed:27153000]

Appendix

Simulation results for our method

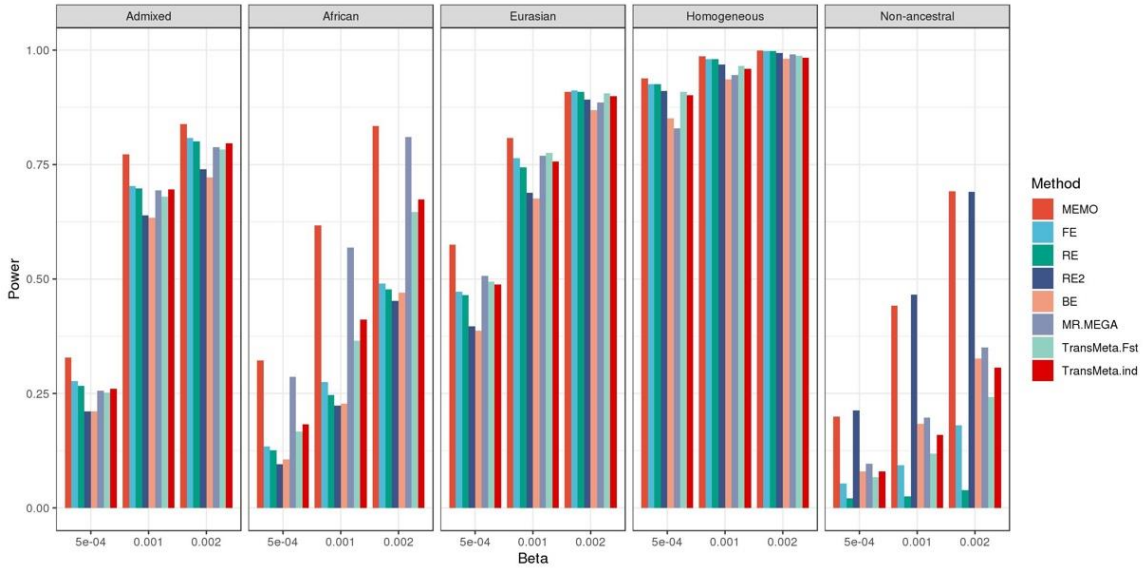
Here we show the simulation results for comparison of our proposed approach MEMO with other 7 methods. Firstly, we estimated type I error rates at three stringent α levels using the proportion of empirical p-values smaller than the given level, and found that empirical type I error rates were well controlled for our method MEMO.

Supplementary Table **1**: Type I error rate estimates at different α levels.

RE2 stands for Han and Eskin's Random Effects model, which is optimized to detect associations under heterogeneity; BE stands for binary effects model, which is optimized to detect associations when some studies have an effect and some studies do not; TransMeta.Fst and TransMeta.ind are F_{st} kernel and the independent kernel in TransMeta *(Shi et al., 2016).*

α	MEMO	FE	RE	RE2	BE	MR-MEGA	TransMeta-Fst	TransMeta-ind
5e-5	3.67E-05	4.36E-05	4.29E-05	2.50E-05	3.02E-05	3.14E-05	3.46E-05	3.02E-05
5e-7	3.96E-07	4.74E-07	4.67E-07	2.93E-07	3.46E-07	3.29E-07	3.74E-07	3.35E-07
5e-9	5.00E-09	5.21E-09	6.15E-09	4.18E-09	5.01E-09	4.23E-09	4.81E-09	4.47E-09

Also, we carried out comparisons of our method and other seven methods under five different scenarios, which cover a wide range of possible scenarios of genetic heterogeneity. Admixed scenario assumes genetic effects exist equally in African and Native American ancestry; African scenario assumes genetic effects exist only in African ancestry; Eurasian scenario assumes genetic effects exist in Native American, East Asian (double effects), European and South Asian ancestries; Homogenous scenario assumes effects are the same in all ancestries while Non-ancestral means effects are different across all ancestries. Supplementary Figure 1 showed the empirical power of all methods under all five scenarios. Our method MEMO yields the highest or near highest power among all eight methods at different estimate beta, except for the cases where beta equals to 0.002 in Eurasian scenario and non-ancestral scenario, but the difference is not significant.



Supplementary Figure **1**: Simulation results for comparing our methods with other 7 methods.

Each box represents each different scenario. X axis represents estimate beta and y axis represents the performance of each method.

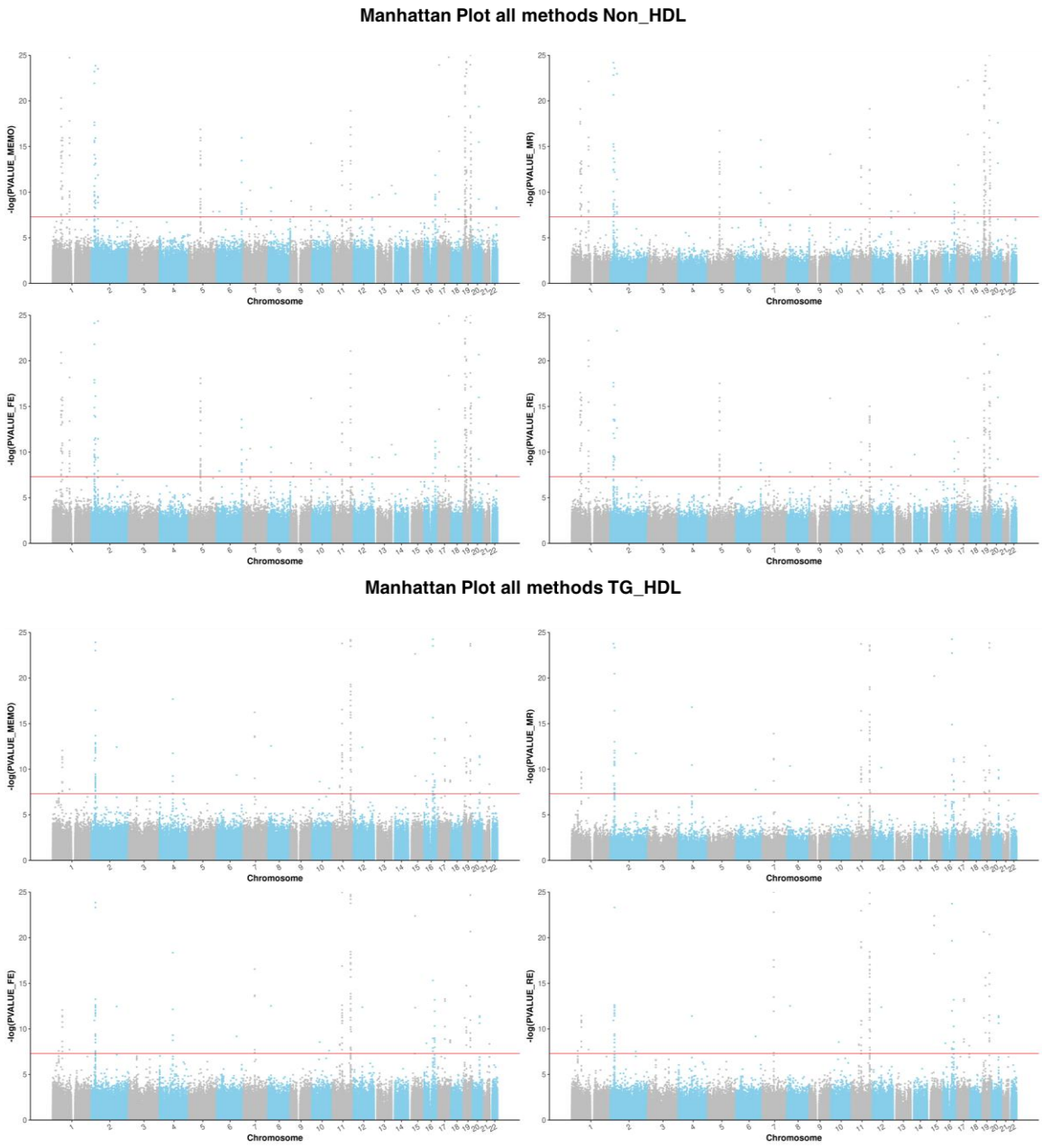
Genomic control values for each study

Supplementary Table **2**: Study-specific genomic control values for each trait (common variants/rare variants).

Traits / Studies	HDL	LDL	Non_HDL	TG/HDL	TG	TC
MIGen.EUR.Case	1.005/0.996	0.992/0.998	0.980/1.001	1.009/0.993	0.980/0.992	0.984/0.985
MIGen.EUR.Control	0.977/0.962	1.000/0.983	0.988/0.982	0.962/0.974	0.964/1.010	0.998/0.991
MIGen.SAS.CASE	0.951/0.978	0.983/0.972	0.982/0.958	0.987/1.002	0.991/1.003	0.972/0.959
MIGen.SAS.CONTROL	0.986/0.987	0.994/0.982	1.000/0.989	0.997/1.014	0.991/1.003	0.991/0.978
TOPMed.ASIAN	0.996/1.026	1.007/0.957	1.000/1.000	1.004/1.018	1.002/0.990	1.008/0.992
TOPMed.BLACK	1.005/1.015	0.996/1.005	1.005/0.997	1.012/0.996	1.015/1.004	1.004/1.010
TOPMed.HISPANIC	0.999/0.989	0.998/0.978	1.007/0.990	0.992/1.000	0.995/0.995	1.010/0.996
TOPMed.SAMOAN	0.977/0.970	0.985/1.021	0.983/1.007	0.992/1.023	0.989/0.969	0.981/0.914
TOPMed.WHITE	0.959/0.999	0.959/0.993	0.955/1.004	0.975/0.974	0.980/0.948	0.956/1.007
UKBB13k.EUR.Case	0.992/1.001	0.980/0.998	0.995/1.005	0.997/0.999	0.987/0.999	0.980/0.994
UKBB13k.EUR.Control	0.984/0.996	0.975/1.001	0.997/0.999	0.992/0.995	0.985/0.991	0.980/0.997
UKBB50k.EUR.Batch1	0.979/0.995	0.980/0.989	0.985/0.987	0.978/1.010	0.984/0.992	0.987/0.006
UKBB50k.EUR.Batch2	0.982/1.014	0.985/0.991	0.988/0.990	0.992/0.999	0.987/1.005	0.980/0.993
UKBB50k.EUR.Batch3	0.977/1.000	0.978/0.996	0.980/0.997	0.990/0.997	0.982/0.996	0.985/1.002
UKBB50k.EUR.Batch4	0.981/1.004	0.967/1.008	0.978/0.996	0.987/1.010	0.987/1.007	0.970/1.005

Studies were labeled according to each cohort and ancestry.

Manhattan plots for variants associated with Non-HDL and TG/HDL



Supplementary Figure **2**: Manhattan plots for variants associated with two traits Non-HDL and TG/HDL.

For each trait, top left shows the result of our method MEMO, top right shows result of MR_MEGA, bottom left shows the result of FE, bottom right shows the result of RE.