UNIVERSITY OF HELSINKI

UNIVARIATE AND MULTIVARIATE STATISTICAL TESTS IN GENETIC ASSOCIATION STUDIES

Sanni Emilia Ruotsalainen

UNIVERSITY OF HELSINKI
Department of Mathematics and Statistics
Statistics

Master's Thesis
May 2017

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty Science	Laitos — Institution — Department Mathematics and statistics

Tekijä — Författare — Author

Sanni Emilia Ruotsalainen

Työn nimi — Arbetets titel — Title

Univariate and multivariate statistical tests in genetic association studies

Oppiaine — Läroämne — Subject

Statistics

Työn laji — Arbetets art — Level Master's thesis	Aika — Datum — Month and year May 2017	Sivumäärä — Sidoantal — Number of pages 49 p.

Tiivistelmä — Referat — Abstract

Genome-wide association studies have identified hundreds of genomic loci associated with a wide range of human conditions and quantitative traits, such as cholesterol level and diabetes. However, most of these studies have focused on analysing single traits, even the studies involving multiple related traits. Growing evidence for pleiotropy, where the same genetic locus is associated with multiple traits, supports the idea that multivariate methods could provide a remarkable boost in statistical power compared to univariate methods.

In this thesis the main research question is to compare the multivariate Wald test to the corresponding univariate test, and to see when multivariate testing is more useful. My second research question is to compare the multivariate Wald test and another multivariate method called Canonical Correlation Analysis (CCA), and to see if they yield the same result.

To examine these topics I performed a simulation study in which I simulated data set with 1,000 genotypes and 1,000 individuals. In addition I simulated bivariate phenotypes that were differently correlated with each other, and the genotypes. I performed the univariate Wald test for each trait against each genotype, and the multivariate Wald test for each trait pair against each genotype. I also performed the corresponding CCA to compare those results with the Wald test.

In addition to the simulation study I performed the similar analyses for real data from The National FINRISK Study. I used three different blood lipid measuerements, HDL-cholesterol, LDL-cholesterol and triglycerides as example traits, and 157 genomic loci previously known to associate with blood lipid levels. These blood lipid levels were approppriate example traits for this study because they are correlated differently with each other, and they are differently associated with the 157 genomic loci used here. Therefore I found many different combinations of correlation between traits, and directions of genetic effects for different traits.

Based on my simulation studies I can say that the multivariate testing is never much worse in terms of power to detect associations than the corresponding univariate tests, and in some cases it is much more powerful. Thus there is no reason not to do the multivariate analysis first in case of studying multiple related traits. Multivariate testing is more powerful in cases where the correlation between the traits is large and the genetic effects for the traits show opposite directions compared to the trait correlation. The least effective multivariate testing is compared to univariate testing when the correlation between the traits is small, and the directions of genetic effects is consistent with the trait correlation. Based on my results multivariate Wald test and CCA yield the same results, with some minor approximation differencies in small sample sizes.

Avainsanat — Nyckelord — Keywords

Multivariate analysis, Lipids, Genome-wide association studies, Cardiovascular disesases
Canonical correlation analysis, Multivariate Wald test

Säilytyspaikka — Förvaringsställe — Where deposited

Kumpula campus library

Muita tietoja — Övriga uppgifter — Additional information

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty Matemaattis-luonnontieteellinen	Laitos — Institution — Department Matematiikan ja tilastotieteen laitos

Tekijä — Författare — Author
Sanni Emilia Ruotsalainen

Työn nimi — Arbetets titel — Title

Univariate and multivariate statistical tests in genetic association studies

Oppiaine — Läroämne — Subject

Tilastotiede

Työn laji — Arbetets art — Level Pro gradu -tutkielma	Aika — Datum — Month and year Toukokuu 2017	Sivumäärä — Sidoantal — Number of pages 49 s.

Tiivistelmä — Referat — Abstract

Perimänlaajuisten assosiaatioanalyysien avulla on löydetty satoja perimän kohtia, jotka ovat yhteydessä useisiin sairauksiin tai ominaisuuksiin kuten kolesterolitasoihin ja diabetekseen. Useimmat näistä tutkimuksista ovat kuitenkin tutkineet ainoastaan yhtä ominaisuutta kerrallaan, vaikka tutkimus käsittelisikin useita toisiinsa liittyviä muuttujia. Kasvava näyttö siitä, että yksi perimän kohta on yhteydessä useisiin ominaisuuksiin (pleiotropia) tukee ajatusta, että monimuuttujamenetelmät voisivat olla tehokkaampia kuin yhden muuttujan menetelmät.

Tutkielmani päätutkimuskysymys on vertailla moniulotteista Waldin testiä vastaavaan yhden muuttujan testiin, ja katsoa millaisissa tilanteissa on tehokkaampaa käyttää monimuuttujatestausta. Toinen tutkimuskysymykseni on vertailla moniulotteista Waldin testiä toiseen monimuuttujamenetelmään, kanoniseen korrelaatioanalysiin (CCA) ja katsoa tuottavatko nämä menetelmät saman tuloksen.

Tutkiakseni näitä asioita tein simulaatiotutkimuksen, jossa simuloin aineiston, jossa on 1,000 genotyyppiä 1,000 henkilölle. Tämän lisäksi simuloin muuttujapareja, jotka ovat eri tavoin yhteydessä toisiinsa, sekä simuloituihin genotyyppeihin. Tein jokaiselle muuttujalle yhden muuttujan Waldin testin jokaista genotyyppiä vastaan, sekä jokaiselle muuttujaparille moniulotteisen Waldin testin jokaista genotyyppiä vastaan. Tein myös vastaavat kanoniset korrelaatioanalyysit jotta voin vertailla näitä tuloksia moniulotteisen Waldin testin tuloksiin.

Simulaatiotutkimuksen lisäksi tein vastaavat analyysit myös aineistolle FINRISKI-tutkimuksesta. Esimerkkimuuttujinani käytin kolmea veren lipidiarvoa, HDL-kolesterolia, LDL-kolesterolia ja triglyserideja, sekä 157 perimän kohtaa, joiden tiedetään olevan yhteydessä veren lipiditasoihin. Lipiditasot olivat hyvä esimerkki tähän tutkimukseen, koska ne ovat erilailla yhteydessä keskenään, sekä näiden 157 perimän kohtien kanssa. Näin ollen löysin useita erilaisia yhdistelmiä lipidien välisistä yhteyksistä sekä perimän vaikutuksista eri lipideille.

Simulaatiotutkimusteni perusteella voimme sanoa, että monimuuttujatestaus on lähes aina vähintään yhtä voimakas havaitsemaan yhteyksiä kuin vastaavat yhden muuttujan testit, ja joissain tapauksissa se on paljon voimakkaampi. Näin ollen ei ole mitään syytä olla suorittamatt monimuuttujatestausta ensin, kun on kyse useista toisiinsa liittyvistä muuttujista. Monimuuttujatestaus on selkeästi voimakkaampi tilanteissa, joissa muuttujien välinen korrelaatio on suurta ja perimän vaikutus näihin muuttujiin on erisuuntaista. Vähiten monimuuttujatestaamisesta on hyötyä yhden muuttujan testiin verrattuna silloin, kun muuttujien välinen korrelaatio on pientä, ja perimän vaikutus muuttujiin on samansuuntaista. Tutkimusteni perusteella voimme myös sanoa, että moniulotteinen Waldin testi ja CCA tuottavat saman tuloksen, joskin pienillä otoskoilla huomataan pieniä approksimaatioeroja.

Avainsanat — Nyckelord — Keywords
Monimuuttujamenetelmät, Perimänlaajuinen assosiaatiotutkimus, Sydän- ja verisuonitaudit, Kanoninen korrelaatioanalyysi, Moniulotteinen Waldin testi

Säilytyspaikka — Förvaringsställe — Where deposited

Kumpulan tiedekirjasto

Muita tietoja — Övriga uppgifter — Additional information

Contents

Chapter 1

Introduction

Years of Genome-wide association studies (GWAS) have successfully identified common variants at more than 1,000 genomic loci robustly associated with a wide range of human conditions and quantitative traits [Price et al., 2015]. For example, around a hundred genetic loci with genome-wide significant association on blood lipid levels have been identified [Surakka, 2014, Willer et al., 2013]. Despite this progress, one big limitation is that almost all GWAS performed have focused on analysing a single trait at a time, even the studies involving multiple related traits, such as blood lipid levels. Growing evidence for pleiotropy, where the same genetic locus is associated with multiple traits, supports the idea that multivariate analysis of multiple related traits can provide a remarkable boost in power for locus discovery, compared to an univariate analysis of a single trait [Inouye et al., 2012].

My aim in this thesis is to demonstrate the nature of multivariate analysis compared to the corresponding univariate analysis by comparing the multivariate Wald test to the corresponding univariate Wald tests. In addition I will compare multivariate Wald test also to another multivariate method, Canonical Correlation Analysis (CCA) to see if they yield the same result. I will first introduce the univariate and multivariate Wald test and CCA and statistical methods behind them in Methods. In Results I will introduce my results first from simulation studies and later a few examples from real data to show how the multivariate methods perform compared to the univariate methods in practice. The results I get from these studies are further discussed in Conclusions.

In my examples I will use blood lipid levels of high-density lipoprotein cholesterol, low-density lipoprotein cholesterol and triglycerides as example traits to demonstrate the multivariate analysis compared to the univariate analysis. These blood lipid levels are appropriate examples for this purpose because in addition to their interesting associations with cardiovascular diseases, and thus to the public health, they are correlated in different ways between each other and also with different genetic loci. Therefore we will find lots of different combinations of correlation between traits, and directions of genetic effects for different traits to show many examples how multivariate tests acts compared to univariate tests in these various cases.

Cardiovascular diseases (disorders of the heart and blood vessels) are a leading cause of death worldwide, and they have been under great interest for epidemiological studies [Webb et al., 2013]. Back in 1971, the Framingham Heart Study offered the first bit of evidence that elevated cholesterol levels are an important risk for heart disease [Kannel et al., 1961], but still decades later the preventative actions for cardiovascular diseases are limited and are mainly concentrated on lifestyle changes. In clinical practice the most often used biomarkers for cardiovascular diseases are the circulating blood lipid levels (cholesterol and triglycerides), which are well-established risk factors for cardiovascular diseases [Webb et al., 2013].

Levels of circulating blood lipids are largely affected by environmental factors such as diet, body composition, smoking and alcohol usage. Therefore the risk for unfavourable lipid levels caused by these factors can be lowered by individual's behavioural and lifestyle changes. However, only a bit less than a half of the variability in lipid levels is attributable to these environmental factors and the remaining proportion is because of the genetic effects [van Dongen et al., 2013]. Therefore, as the genetic factors explain around half of the population lipid variation, it is crucial to understand the genetic mechanisms behind the lipid levels in order to better understand the biological pathways behind cardiovascular diseases [Surakka, 2014].

Chapter 2

Background

2.1 Multivariate analysis

Multivariate analysis can be defined as the application of methods that deal with a reasonably large numbers of measurements made on each object in one or more samples *simultaneously*. The important point is that multivariate analysis deals with the *simultaneous relationships among variables*. Multivariate techniques differ from univariate analysis by directing attention away from the analysis of the mean and variance of a single variable to the analysis of the correlations which reflect the extent of relationship among several variables. [Dillon and Goldstein, 1983] Advantages of using multivariate analysis are that it looks the phenomena in a more general way and it can help control for Type 1 error (incorrect rejection of a true null hypothesis, ”a false positive”) [Rencher and Christensen, 2011].

In practice, multivariate data sets are common, although they are not always analysed as such. The exclusive use of univariate methods with such data is no longer excusable, given the availability of multivariate techniques and inexpensive computing power to carry them out. In the past, the computations were overwhelming even with smaller datasets, and so multivariate analyses were typically avoided, but now this is not a problem any more. [Rencher and Christensen, 2011]

Biological processes, such as metabolism and circulation of blood lipids, are very complex by nature. As such it is rare that a single response variable is sufficient to describe a biological system entirely. Rather, multiple response variables are often measured to gain a biological insight. For example, in the case of circulating blood lipid levels it is common to measure several cholesterol levels. Thus it seems intuitive that the multivariate methods would be better suited to study biological processes, such as circulation of blood lipids.

2.2 Genetics

2.2.1 Human genome

Genome structure

In this section I will briefly introduce the basics of the human genome's structure and function as well as a widely used method for studying genetic associations with different traits and diseases. This section is based on [Klug et al., 2012]. Genetic information of human is encoded in the genome, which is all the DNA in a cell. DNA can be found in almost every cell in the body inside the nucleus as 23 chromosome pairs (total of 46 chromosomes). Each chromosome is a double stranded string built with smaller particles; nucleotides. There are four different types of nucleotides in DNA: adenine (A), guanine (G), cytosine (C) and thymine (T), depending on the base in nucleotide. These nucleotides form a string, and thus DNA can be thought as a ”two complementary string of letters”. The human genome consists of approximately 3 billion base pairs of which only around 1.5% is covered by regions of the genome that codes for proteins, called exons. [Klug et al., 2012]

Genome function

A gene is a piece of DNA that has promoter, exons and introns. Exons are the protein coding parts of the gene, and therefore they define the amino acid structure in the resulting protein. Introns are the parts of genes that do not directly code for proteins, but are integral to gene expression regulation. The exonic DNA has triplets of nucleotides, called codons, each of which has a corresponding amino acid.

In gene transcription the double strand of the DNA is splitted and the transcription (Fig. 2.1) is started. In case of a protein coding-gene, an enzyme called RNA polymerase starts to move along the template strand (non-coding strand) of the DNA and copies it into precursor messenger RNA (pre-mRNA) starting from a start codon and ending at a stop codon. The structure of the resulting single stranded pre-mRNA is similar to that in the coding DNA strand, except that base thymine is replaced with uracil (U).

Once the pre-mRNA is ready, the introns are spliced out from the sequence and the mRNA moves outside of the cell nucleus to the cell cytoplasm where it binds to the ribosome. The ribosome starts to translate the genetic code in the mRNA into amino acid code. Each of these codons, except for stop codons (UAG, UGA and UAA), has one pairing amino acid. In contrary, one amino acid can have multiple corresponding codons which allows a mutation to

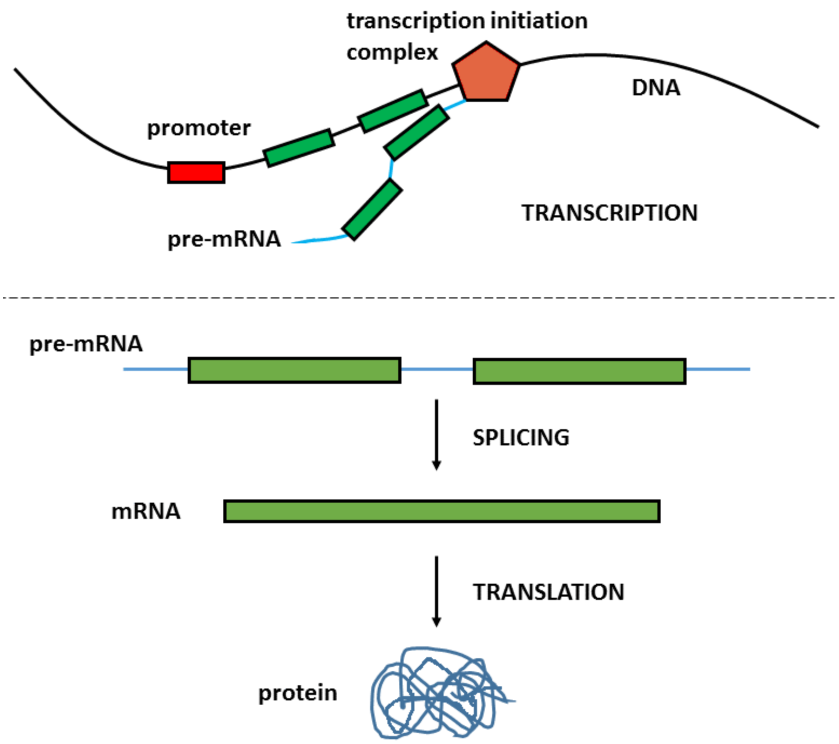


Figure 2.1: From DNA to protein. Demonstrative figure of transcription and translation of a protein coding gene. Figure adapted from [Surakka, 2014].

occur in the genome without changing the protein structure.

Genetic variation

An allele is an alternative form of the DNA sequence in a region of the genome (a genomic region is also called a locus, plural loci). Most multicellular organisms, like humans, are diploid which means that they have two sets of chromosomes. Diploid organism have one copy of each locus (and one allele) on each chromosome, a total of two alleles at each locus. An individual that has a pair of identical alleles at a locus is said to be *homozygous* at that locus and an individual that has two different alleles at a locus is said to be *heterozygous* for that locus.

The process of mutation is the source of new alleles. For a new allele to be recognized by observation of an organism, the allele must cause a change in the phenotype. Some mutations can change the physical appearance, the phenotype, of the organisms, while some others may have no apparent effect on the organism. Mutation can originate as an alteration in DNA sequence occurring during meiosis or because of radiation or mutagens, that has escaped the DNA repair system. Any base-pair change in any part of a DNA molecule can be considered as a mutation. When a non-fatal mutation occurs in the germ line cells, it can be passed on to the next generation and when the frequency of the mutated allele in the population rises up to 1%, the mutation is called a polymorphism.

2.2.2 Genome-wide association studies (GWAS)

This section is based on [Pearson and Manolio, 2008]. Genome-wide association study (or GWAS) is defined by the National Institutes of Health as a study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits. In other words it is an examination of multiple genetic variants in different individuals to see if any of those variants is associated with the trait under study. The idea is to search the whole genome for small variations in the genome, called single nucleotide polymorphisms, or SNPs, that occur more frequently in people with a particular disease than in people without the disease, or in case of quantitative (continuous) trait to see if the trait under study is distributed differently among the genotypes of SNPSs under study.

Briefly, the idea in GWAS is to search for genotype-phenotype associations that happens when one or more genotypes within a population co-occur with a trait under study more often than it would be expected by change. Identifying such associations is very important, because they give us hints of biology behind the diseases and traits and thus GWA studies can also give us hints of targets for therapeutics. Because GWAS examines SNPs across the genome, they represent a promising way to study complex, common diseases and traits, in which many genetic variations contribute to a person's risk for the disease. GWA studies typically perform the first analysis in a discovery cohort, followed by validation of the most significant SNPs in an independent validation cohort.

Family-based linkage studies have been successful in identifying genes of large effect in Mendelian diseases (diseases controlled by a single locus), such as cystic fibrosis, but have had limited success in common, non-Mendelian conditions, such as asthma. Major limitations of linkage studies are relatively low statistical power for complex diseases influenced by multiple genes, and the large size of the chromosomal regions shared among family members, in whom it can be difficult to narrow the linkage signal sufficiently to identify a causative gene. For non-Mendelian conditions, GWA studies represent a valuable advance over family-based linkage studies, in which multiple affected families are arduously assembled and inheritance patterns are related to only a few hundred markers throughout the genome. [Pearson and Manolio, 2008]

The number of SNPs tested in GWA studies depends on the genotyping technology, but is typically one million or more. There are multiple ways to test the significance of the association between genotype and trait, depending on the type of the trait under study. When studying a quantitative traits, such as blood lipid levels, statistical significance of the association can be tested using simple linear regression where the genotype is the explanatory variable and the trait is the response variable. The significance of the regression coefficient in that model can be tested using univariate Wald test, for example. In that case the p-value is calculated from standard normal distribution.

The most frequently used GWA study desing to date has been the case-control design that is used to study a disease. In case-control design, for each of the SNPs it is investigated if the allele frequency is significantly altered between the case and the control groups. Because each individual carries two copies of each autosomal SNP, the frequency of each of the three possible genotypes can be tested. In case-control set-ups, the fundamental unit for reporting effect sizes is the odds ratio. The odds ratio is the ratio of two odds, which in the context of GWA studies are the odds of disease for individuals having a specific allele and the odds of disease for individuals who do not have that same allele. When the allele frequency in the case group is much higher than in the control group, the odds ratio is higher than 1, and vice versa for lower allele frequency. Additionally, a p-value for the significance of the odds ratio is typically calculated using a simple χ^2 -test. Finding odds ratios that are significantly different from 1 is the objective of the GWA study because this shows that a SNP is associated with the disease. [Pearson and Manolio, 2008]

The exact threshold for statistical significance varies by study, but the conventional threshold is 5×10^{-8} to be significant in the face of hundreds of thousands to millions of tested SNPs. One of the biggest problems in the GWA analyses is the multiple testing dilemma; when analyzing hundreds of thousans or even millions of SNPs simultaneously one must account for the fact that probability to detect at least one association by change (type 1 error) rises with each independent test. However, as SNPs in the data are not truly independent because of the linkage disequilibrium [1], a simple Bonferroni correction that corrects for the number of tests is highly conservative. [Surakka, 2014, Sham and Purcel, 2014]

The most powerful way to take the multiple testing challenge into account would be to use permutation procedures; simulate the null distribution of the test statistics in the case of no association. However, as the magnitude of SNPs in the GWA analyses can be in the millions, the computational challenges have made it nearly impossible to use permutations in the large GWA studies. Due to these problems, Bonferroni correction for one million independent tests, p-value $< 0.05/10^6 = 5 \times 10^{-8}$, is commonly used as a significance threshold in GWA studies. This threshold has proven to work well in the published studies as most of the findings

[1] When two genetic loci are positioned close to each other in the genome, they are more likely to be inherited together. Alleles that are not independently inherited are said to be linked with each other, and they are said to be in linkage disequilibrium [Klug et al., 2012].

have been successfully replicated and it has thus become a general genome-wide significance threshold. [Surakka, 2014, Sham and Purcel, 2014]

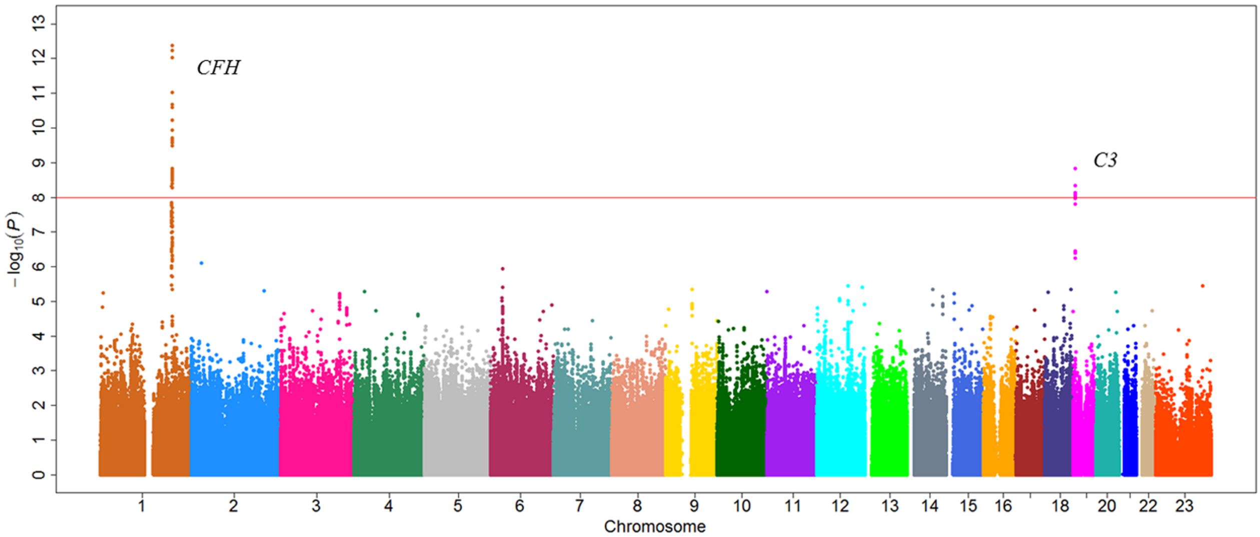


Figure 2.2: Example Manhattan plot of genome-wide association analysis for serum C3 level. X-axis shows chromosomal positions. Y-axis shows $-\log_{10}$ (p-values) from linear regression adjusted for age, smoking, and log(BMI). The horizontal solid line indicates the present threshold of $p = 5 \times 10^{-8}$. Figure from [Yang et al., 2012]

.

After calculation of p-values for all SNPs, a common approach to examine and present the results is to create a Manhattan plot of the results. In the context of GWA studies, this plot shows the negative logarithm of the p-value as a function of genomic location. Thus the SNPs with the most significant association stands out on the plot, usually as stacks of points because of the correlation structure of the genome. Therefore in a good Manhattan plot true signals are supported by many neighbouring SNPs and are not represented by only a single dot that stands out. There is an example of Manhattan plot in the Fig.2.2 [Yang et al., 2012]. In this example the trait under investigation is serum C3 levels. In the x-axis there is the chromosomal position, and each of the chromosomes is drawn with different colours to make it easier to distinguish them. In this example, the strongest associations are seen on chromosomes 1 (CFH locus) and 18 (C3 locus). Both of these signals are supported by the SNPs close by, which supports that these are true associations.

The GWA approach is revolutionary because it enables examination of the entire human genome at levels of resolution previously unattainable, in thousands of unrelated individuals, unconstrained by prior hypotheses regarding genetic associations with disease. However, the GWA approach can also be problematic because the massive number of statistical tests performed presents an unprecedented potential for false-positive results, leading to new stringency in acceptable levels of statistical significance and requirements for replications of findings [Pearson and Manolio, 2008].

2.3 Circulating blood lipids

In this section I will introduce the very basics of blood lipids and their function as a motivation for doing GWA studies on them. This section is based on [Surakka, 2014]. Blood lipids (or blood fats) are lipids in the blood, either free or bound to other molecules. Blood lipids are mainly fatty acids and cholesterol. The density of the lipids and type of protein determines the fate of the particle and its influence on metabolism. The concentration of blood lipids depends on intake and secretion from the intestine, and uptake and secretion from cells.

Cholesterol is made by the liver and it is an essential part of cell walls and nerves. Cholesterol cannot dissolve in the blood. Therefore it must be transported through bloodstream by carriers, called lipoproteins, which got their name because they are made of fat (lipid) and proteins. Lipoproteins are named based on their size and density; the lower the density, the larger the particle. There are total of five major groups of lipoproteins; chylomicrons, very low-density lipoprotein (VLDL), intermediate-density lipoprotein (IDL), low-density lipoprotein (LDL) and high-density lipoprotein (HDL). Low-density lipoprotein delivers cholesterol to cells for membrane production, and high-density lipoprotein scavenges excess cholesterol for return to the liver.

LDL cholesterol, LDL-C, is what's considered "bad" cholesterol, as it leads to a build-up of cholesterol in arteries. LDL contributes to plaque, a thick, hard deposit that can clog arteries and make them less flexible. This condition is known as atherosclerosis. If a clot forms and blocks a narrowed artery, heart attack or stroke can result. Another condition called peripheral artery disease can develop when plaque build-up narrows an artery supplying blood to the legs.

HDL cholesterol, HDL-C, is what's considered "good" cholesterol, as it carries cholesterol from other parts of the body back to the liver, where it is broken down and passed from the body. One-fourth to one-third of blood cholesterol is carried by HDL. A healthy level of HDL cholesterol may protect against heart attack and stroke, and low levels of HDL cholesterol have been shown to increase the risk of heart disease.

Triglycerides (TG) are fats from the food we eat that are carried in the blood, and they are used to store excess energy from our diet. Most of the fats we eat are in triglyceride form. Excess calories, alcohol or sugar in the body turn into triglycerides and are stored in fat cells throughout the body. Triglycerides and cholesterol are both fatty substances, lipids, but triglycerides are fats and cholesterol is not. An elevated triglyceride level is associated with an increase in the risk of heart disease. High levels of triglycerides in the blood are associated with atherosclerosis. Elevated triglycerides can be caused by overweight and obesity, physical inactivity, cigarette smoking, excess alcohol consumption and a diet very high in carbohydrates. Underlying diseases or genetic disorders are sometimes the cause of high triglyceride levels. People with high triglyceride levels often also have a high total cholesterol level, including a high LDL cholesterol level and a low HDL cholesterol level. Many people with heart disease or diabetes also have high triglyceride levels.

Cholesterol metabolism plays a central role in cardiovascular diseases. The functions of HDL and LDL particles explain why LDL-C levels have positive correlation with cardiovascular events and HDL-C levels have negative correlations. The excess amount of LDL and insufficient HDL lipid clearance in the blood stream can cause arterial inflammation leading to an atherosclerotic plaque blocking the artery. This connection between circulating blood lipids and cardiovascular disease risk has made lipids part of the most studied human traits. As the different enzymatically measurable lipid traits, HDL-C, LDL-C, total cholesterol TC (which can be calculated for certain units of measurements using Friedewald's equation TC=LDL+ HDL+TG/5) and triglycerides (TG), also seem to be highly heritable, they have been under great interest in genetic studies.

Chapter 3

Methods

3.1 Multivariate Linear Regression

In this section I will introduce multivariate linear regression, and its parameter estimation and testing. Theory in this section is based on chapter 10 in [Rencher and Christensen, 2011]. Simple and multiple linear models are used to study how a single quantitative variable Y depends on one or more predictor variables X, respectively. Multivariate linear model is their extension and it is used to study how multiple quantitative variables Y depend on one or more predictor variables X. The predictor variables in this model may be quantitative or qualitative. As simple linear regression can be used in parameter testing in GWA studies in case of one trait, the multivariate linear regression can be used to test the significance of the association between SNPs and multiple traits simultaneously.

3.1.1 Multivariate Linear model

In multivariate linear regression multiple Y's are measured corresponding to each set of X's and each $Y_1, Y_2, \ldots, Y_q$ is to be predicted by all of $X_1, X_2, \ldots X_p$. The n observed values of the vector of Y's can be listed as rows in the following matrix:

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} \mathbf{y}'_1 \\ \mathbf{y}'_2 \\ \vdots \\ \mathbf{y}'_n \end{pmatrix}.$$

Thus each row of $\mathbf{Y}$ ($\mathbf{y}_i'$) contains the values of the q dependent variables measured on subject i, $i = 1, 2, \ldots, n$ and each column of $\mathbf{Y}$ consists of the n observations on one of the q variables $Y_1, Y_2, \ldots, Y_q$.

The n values of predictor variables $X_1, X_2, \ldots, X_p$ can be placed in a matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

Since each of the Y's depends on the X's in its own way, each of them will need different regression coefficients (β's). Thus we have a column of β's for each column of $\mathbf{Y}$, and these columns form a matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_q)$:

$$\mathbf{B} = \begin{pmatrix} \beta_{01} & \beta_{02} & \ldots & \beta_{0q} \\ \beta_{11} & \beta_{12} & \ldots & \beta_{1q} \\ \vdots & \vdots & & \vdots \\ \beta_{p1} & \beta_{p2} & \ldots & \beta_{pq} \end{pmatrix}.$$

The multivariate model is therefore:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{\Xi}, \tag{3.1}$$

where $\mathbf{Y}$ is $n \times q$, $\mathbf{X}$ is $n \times (p+1)$, $\mathbf{B}$ is $(p+1) \times q$, and $\boldsymbol{\Xi}$ is the residual error matrix. The model can be written for each column $i = 1, 2, \ldots, q$ of $\mathbf{Y}$ separately as:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{ni} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \\ \vdots \\ \beta_{pi} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{ni} \end{pmatrix}. \tag{3.2}$$

There are some additional assumptions that lead to good estimates. First of these assumptions is that $E(\mathbf{Y}) = \mathbf{XB}$, or $E(\boldsymbol{\Xi}) = \mathbf{0}$. This assumption states that the linear model is correct and that no additional X's are needed to predict the Y's. Second assumption is that $\text{cov}(\mathbf{y}_i) = \Sigma$ for all $i = 1, 2, \ldots, n$, where $\mathbf{y}_i'$ is the ith row of $\mathbf{Y}$. This assumption asserts that each of the n observation vectors (rows in $\mathbf{Y}$) has the same covariance matrix Σ. Third of these assumptions is that $\text{cov}(\mathbf{y}_i,\mathbf{y}_j)=\mathbf{0}$ for all $i \neq j$, which declares that observation vectors $\mathbf{y}_i$ are not correlated with each other. Thus we assume that the Y's within an observation vector are correlated with each other but independent of the Y's in any other observation vector.

The covariance matrix Σ mentioned in the second assumption earlier contains the variances and covariances of $y_{i1}, y_{i2}, \ldots, y_{iq}$ in any $\mathbf{y}_i$:

$$\Sigma = \mathrm{cov}(\mathbf{y}_i) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2q} \\ \vdots & \vdots & & \vdots \\ \sigma_{q1} & \sigma_{p2} & \ldots & \sigma_{qq} \end{pmatrix}$$

The assumption 3 says that the covariances of each $y_{i1}, y_{i2}, \ldots, y_{iq}$ with each of $y_{j1}, y_{j2}, \ldots, y_{jq}, (i \neq j)$ are zero:

$$\begin{pmatrix} \mathrm{cov}(y_{i1}, y_{j1}) & \mathrm{cov}(y_{i1}, y_{j2}) & \ldots & \mathrm{cov}(y_{i1}, y_{jq}) \\ \mathrm{cov}(y_{i2}, y_{j1}) & \mathrm{cov}(y_{i2}, y_{j2}) & \ldots & \mathrm{cov}(y_{i2}, y_{jq}) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(y_{iq}, y_{j1}) & \mathrm{cov}(y_{iq}, y_{j2}) & \ldots & \mathrm{cov}(y_{iq}, y_{jq}) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}.$$

Using vectorization the model in 3.2 can be re-written as

$$\mathrm{vec}\mathbf{Y} = \mathbf{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*, \tag{3.3}$$

where $\mathbf{X}^* = \mathbf{I}_q \otimes \mathbf{X}$ is a $qn \times q(p+1)$ block-diagonal matrix, $\boldsymbol{\beta}^* = \mathrm{vec}\mathbf{B}$ is a vector of length $q(p+1)$, and $\boldsymbol{\varepsilon}^*$ =vec $\boldsymbol{\Xi}$ is a vector of length qn.

3.1.2 Estimation of the parameters

Least Squares Estimation for B

The matrix $\mathbf{B}$ of β's is estimated with

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.4}$$

$\hat{\mathbf{B}}$ is the *least squares estimator* for $\mathbf{B}$ because it minimizes matrix $\mathbf{E} = \hat{\boldsymbol{\Xi}}'\hat{\boldsymbol{\Xi}}$ that is analogous to *error sum-of-squares (SSE)* in univariate case:

$$\mathbf{E} = \hat{\boldsymbol{\Xi}}'\hat{\boldsymbol{\Xi}} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}).$$

The matrix $\hat{\mathbf{B}}$ minimizes $\mathbf{E}$ in the following sense: if we let $\mathbf{B}_0$ be an estimate that may possibly be better than $\hat{\mathbf{B}}$ and add $\mathbf{X}\hat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0$ to $\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}$, we find that this adds a positive definite matrix to $\mathbf{E} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$'$(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$. Thus we cannot improve on $\hat{\mathbf{B}}$ [Rencher and Christensen, 2011].

Obtaining this least squares estimator can be done without imposing the assumptions $\mathrm{E}(\mathbf{Y}) = \mathbf{XB}, \mathrm{cov}(\mathbf{y}_i) = \mathbf{\Sigma}$ and $\mathrm{cov}(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{0}$. However, when these assumptions hold, $\hat{\mathbf{B}}$ has the following properties:

- $\hat{\mathbf{B}}$ is unbiased, that is, $E(\hat{\mathbf{B}}) = \mathbf{B}$.
- All $\hat{\beta}_{ji}$'s in $\hat{\mathbf{B}}$ are correlated with each other, which is due to the correlations among the x's and the y's. Because of the correlations among the columns of $\hat{\mathbf{B}}$, we need multivariate test for hypotheses about $\mathbf{B}$
- The least squares estimators $\hat{\beta}_{ji}$ have minimum variance among all possible linear unbiased estimators, i.e. $\hat{\mathbf{B}}$ is the best linear unbiased estimator (BLUE) for $\mathbf{B}$.

An estimator for $\mathbf{\Sigma}$

An unbiased estimator of $\mathrm{cov}(\mathbf{y}_i) = \mathbf{\Sigma}$ is given by

$$\begin{aligned} \mathbf{S}_e = \frac{\mathbf{E}}{n-q-1} &= \frac{(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})}{n-q-1} \\ &= \frac{\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}}{n-q-1}. \end{aligned} \tag{3.5}$$

3.1.3 Model Corrected for Means

It is often convenient to "center" the X's by subtracting their means, $\bar{X}_1 = \sum_{i=1}^{n} x_{i1}/n, \bar{X}_2 = \sum_{i=1}^{n} x_{i2}/n$, and so on. When the X's are centered by subtracting their means, we get the centered X matrix

$$\mathbf{X}_c = \begin{pmatrix} x_{11} - \bar{X}_1 & x_{12} - \bar{X}_2 & \dots & x_{1p} - \bar{X}_p \\ x_{21} - \bar{X}_1 & x_{22} - \bar{X}_2 & \dots & x_{2p} - \bar{X}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{X}_1 & x_{n2} - \bar{X}_2 & \dots & x_{np} - \bar{X}_p \end{pmatrix}$$

In terms of centered x's, the model for each y_{ij} in (3.2) becomes

$$y_{ij} = \alpha + \beta_{1i}(x_{j1} - \bar{X}_1) + \beta_{2i}(x_{j2} - \bar{X}_2) + \dots + \beta_{pi}(x_{jp} - \bar{X}_p) + \varepsilon_{ji}, \tag{3.6}$$

where $\alpha = \beta_{0i} + \beta_{1i}\bar{X}_1 + \beta_{2i}\bar{X}_2 + \cdots + \beta_{pi}\bar{X}_p$.

The matrix $\mathbf{B}$ can be partitioned as

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_0' \\ \mathbf{B}_1 \end{pmatrix} = \left(\begin{array}{cccc} \beta_{01} & \beta_{02} & \cdots & \beta_{0q} \\ \hline \beta_{11} & \beta_{12} & \cdots & \beta_{1q} \\ \vdots & \vdots & & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pq} \end{array}\right). \tag{3.7}$$

Now using the centered X's in the matrix $\mathbf{X}_c$ the estimates for $\mathbf{B}_1$ and $\boldsymbol{\beta}_\mathbf{0}'$ are

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c})^{-1}(\mathbf{X}_\mathbf{c}'\mathbf{Y}), \tag{3.8}$$

$$\hat{\boldsymbol{\beta}}_\mathbf{0}' = \bar{\mathbf{Y}}' - \bar{\mathbf{X}}'\hat{\mathbf{B}}_1, \tag{3.9}$$

where $\bar{\mathbf{Y}} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_q)$ and $\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)$. These estimates give the same results as $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in (3.4).

The estimate $\hat{\mathbf{B}}_1$ in (3.8) can be expressed in terms of sample covariance matrices $\mathbf{S}_{xx}$ and $\mathbf{S}_{xy}$. If (3.8) is divided and multiplied by $n-1$ we get

$$\begin{aligned} \hat{\mathbf{B}}_1 &= (n-1)(\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c})^{-1}\frac{\mathbf{X}_\mathbf{c}'\mathbf{Y}_\mathbf{c}}{n-1} = \left(\frac{\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}}{n-1}\right)^{-1}\frac{\mathbf{X}_\mathbf{c}'\mathbf{Y}_\mathbf{c}}{n-1} \\ &= \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} \end{aligned} \tag{3.10}$$

where $\mathbf{S}_{xx}$ and $\mathbf{S}_{xy}$ are blocks from the overall sample covariance matrix of the vector $(Y_1, Y_2, \ldots, Y_q, X_1, X_2, \ldots, X_p)'$:

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}. \tag{3.11}$$

3.1.4 Statistical testing

In this section I will introduce univariate and multivariate Wald test which are used for testing significance of regression coefficients. Theory in this section is based on chapter 5.2 in [Rencher and Christensen, 2011]. Univariate Wald test is used for testing significance of a single coefficient, say β, and the multivariate Wald test is used for testing the joint significance of several components of a vector of coefficients $\boldsymbol{\beta}$.

Univariate Wald test

Let us first consider testing the significance of just one particular coefficient, say β in case of a single response variable y and a single predictor variable x. In this case the model is Y=Xβ+ϵ, where X and Y are centered $n \times 1$ vectors and β is a scalar. Under the univariate Wald test, the maximum likelihood estimate (m.l.e.) $\hat{\beta}$ of β is compared with the proposed value β_0 with the assumption that the difference between the two will be approximately normally distributed. For example when testing the significance of a genetic effect, it is tested whether the effect is zero. The null hypothesis in that case is that β is 0:

$$H_0 : \beta = 0.$$

In more generally the null hypothesis is $H_0 : \beta = \beta_0$.

Under $H_0 : \beta = 0$, the m.l.e. $\hat{\beta}$ has a distribution with mean 0 and variance

$$\begin{aligned} \mathrm{var}(\hat{\beta}) &= \mathrm{var}\Big(\frac{X^TY}{X^TX}\Big) = \frac{1}{(X^TX)^2}\mathrm{var}\Big(X^TY\Big) \\ &= \frac{1}{(X^TX)^2}\mathrm{var}\Big(\sum_{i=1}^{n} x_i y_i\Big) \\ &= \frac{\sum x_i^2 \mathrm{var}(y_i)}{(X^TX)^2} = \frac{\mathrm{var}(Y)}{X^TX} \\ &= \sigma^2 (X^TX)^{-1}, \end{aligned} \tag{3.12}$$

where σ^2 is the constant variance of the errors ϵ. This is usually unknown and in practice it is replaced by the unbiased estimate based on the residuals sum of squares.

Thus, we can base our univariate test statistic t on the ratio

$$t = \frac{\hat{\beta}}{\sqrt{\mathrm{var}(\hat{\beta})}}. \tag{3.13}$$

In more generally the test statistic is

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{\mathrm{var}(\hat{\beta})}}. \tag{3.14}$$

Under the assumption of normality of the errors, the ratio of the coefficient to its standard error t has under H_0 a *Student's t* distribution with $n - p$ degrees of freedom when σ^2 is estimated, and a standard normal distribution if σ^2 is known.

Under the weaker second-order assumptions concerning the means, variances and covariances of the observations, the ratio has approximately in large samples a standard normal distribution. This result provides a basis for approximate inference in large samples.

The t test can also be used to construct a confidence interval for a coefficient β. It can be stated with $100(1-\alpha)\%$ confidence that β is between the bounds

$$\hat{\beta} \pm t_{1-\alpha/2,n-p}\sqrt{\mathrm{var}(\hat{\beta})}, \tag{3.15}$$

where $t_{1-\alpha/2,n-p}$ is the two-sided critical value of Student's t distribution with $n-p$ d.f. for a test of size α, n is the sample size and p is the number of predictor variables in $\mathbf{X}$.

Multivariate Wald test

The Wald test can also be used to test the joint significance of several coefficients, for example testing the significance of an effect of a single locus on multiple phenotypes simultaneously. In this case the model is $\mathbf{Y} = \mathbf{XB} + \Xi$, where $\mathbf{X}$ is $n \times 1$ matrix and both $\mathbf{X}$ and $\mathbf{Y}$ are centered. If we have a vector of coefficients, say $\boldsymbol{\beta}$, of length q, then the null hypothesis is:

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \tag{3.16}$$

that is, all the β's in $\boldsymbol{\beta}$ are 0. The multivariate Wald statistic W to test this hypothesis is calculated as follows:

$$W = \hat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_{\beta}^{-1} \hat{\boldsymbol{\beta}} \tag{3.17}$$

where $\hat{\boldsymbol{\beta}}$ is the m.l.e. of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\beta}$ is variance-covariance matrix of $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{\Sigma}_{\beta} = \begin{pmatrix} \mathrm{var}(\hat{\beta}_1) & \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) & \dots & \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_q) \\ \mathrm{cov}(\hat{\beta}_2, \hat{\beta}_1) & \mathrm{var}(\hat{\beta}_2) & \dots & \mathrm{cov}(\hat{\beta}_2, \hat{\beta}_q) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(\hat{\beta}_q, \hat{\beta}_1) & \mathrm{cov}(\hat{\beta}_q, \hat{\beta}_2) & \dots & \mathrm{var}(\hat{\beta}_q) \end{pmatrix}, \tag{3.18}$$

where covariance of $\hat{\beta}_j$ and $\hat{\beta}_k$ is

$$\begin{aligned}
\text{cov}(\hat{\beta}_j, \hat{\beta}_k) &= \text{cov}\Big(\frac{X^T Y_j}{X^T X}, \frac{X^T Y_k}{X^T X}\Big) \\
&= \frac{1}{(X^T X)^2}\text{cov}\Big(X^T Y_j, X^T Y_k\Big) \\
&= \frac{1}{(X^T X)^2}\text{cov}\Big(\sum_{i=1}^{n} x_i y_{ij}, \sum_{i=1}^{n} x_i y_{ik}\Big) \\
&= \frac{1}{(X^T X)^2}\Big(\sum_{i=1}^{n} x_i^2 \text{cov}\Big(Y_j, Y_k\Big)\Big) \\
&= \text{cov}(Y_j, Y_k)(X^T X)^{-1} \\
&= \sigma_{jk}^2 (X^T X)^{-1}
\end{aligned} \tag{3.19}$$

Thus $\Sigma_\beta = (\mathbf{X}^T\mathbf{X})^{-1}\Sigma$, where Σ is the covariance matrix of the error terms. As in univariate case, also in the multivariate case the covariances of the coefficients depends on covariance of the Y_j and Y_k, σ_{jk}^2. These are usually unknown and in practice we substitute the estimate based on the residual sum of squares.

Asymptotic theory tells that under H_0 the large-sample distribution of the m.l.e. $\hat{\boldsymbol{\beta}}$ is multivariate normal with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_\beta$ (which is positive definite matrix) i.e $\hat{\boldsymbol{\beta}} \sim \text{N}_p(0, \boldsymbol{\Sigma})$. If we use the known result[1] concerning the multivariate normal distribution and χ^2- distribution, we get that $(\hat{\boldsymbol{\beta}} - \mathbf{0})'\boldsymbol{\Sigma}_\beta{}^{-1}(\hat{\boldsymbol{\beta}} - \mathbf{0}) \sim \chi_p^2 \Rightarrow \hat{\boldsymbol{\beta}}'\boldsymbol{\Sigma}_\beta{}^{-1}\hat{\boldsymbol{\beta}} \sim \chi_p^2$. This means that the large-sample distribution of the $W = \hat{\boldsymbol{\beta}}'\boldsymbol{\Sigma}_\beta{}^{-1}\hat{\boldsymbol{\beta}}$ is chi-squared with p degrees of freedom. This result holds whether the Σ_β is known or estimated.

Under the assumption of normality there is a stronger result: if Σ_β is known, the distribution of W is exactly χ^2 with p degrees of freedom. In the more general case where Σ_β is estimated using a residual sum of squares based on $n - p$ degrees of freedom, the distribution of W/p is an F with p and $n - p$ degrees of freedom. As n approaches infinity ($n - p$ approaches infinity), the F distribution times p approaches a χ^2 distribution with p degrees of freedom. Thus, in large samples it makes no difference whether one treats W as χ^2 or W/p as an F statistic, and often W is treated as χ^2 as a large sample approximation.

3.2 Canonical Correlation Analysis (CCA)

In this section I will introduce the Canonical correlation analysis (CCA). Theory in this section is based on chapter 9 in [Dillon and Goldstein, 1983]. Canonical correlation analysis

[1] According to this result $Z \sim \text{N}_k(\mu, \Sigma) \Rightarrow (Z - \mu)'\Sigma_\beta^{-1}(Z - \mu) \sim \chi_k^2$ (assuming that Σ_β is positive definite)

is a well-established multivariate technique for detecting linear relationships between two sets of variables, predictor and response variables. It should be used in analysing several predictor and response variables *simultaneously*, and it is particularly appropriate when the response variables are themselves correlated.

In CCA *canonical variates* are computed from both sets of variables. A variate in canonical correlation analysis is analogous to a dimension or factor in a principal components analysis. The difference is that a canonical variate consists of maximally correlated predictor and response parts. A maximum of M variates can be extracted, where M is the number of variables in the smallest set, that is $M = min(p, q)$, where p is the number of predictor variables, and q is the number of response variables. The M variates are extracted such that they are independent of each other. To test the significance of the relationships between canonical variates the data should meet the requirements of multivariate normality and homogeneity of variance.

3.2.1 The Population Model

Let p be the number of predictor variables and q be the number of response variables, and assume that $p \geqslant q$. Denote by $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$ the p dimensional vector of predictor variables, $X's$, and by $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_q)$ the q dimensional vector of response variables, $Y's$. Letting $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ denote the respective mean vectors associated with $\mathbf{X}$ and $\mathbf{Y}$, the population variance-covariance matrices can be defined as

$$\begin{aligned}\boldsymbol{\Sigma}_{xx} &= E\Big\{(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)'\Big\}\\ \boldsymbol{\Sigma}_{yy} &= E\Big\{(\mathbf{Y}-\boldsymbol{\mu}_y)(\mathbf{Y}-\boldsymbol{\mu}_y)'\Big\}\\ \boldsymbol{\Sigma}_{xy} &= E\Big\{(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{Y}-\boldsymbol{\mu}_y)'\Big\}.\end{aligned} \tag{3.20}$$

The objective of CCA is to find the linear combination of X's that is maximally correlated with some linear combination of the Y's. If we denote the respective linear combinations by

$$X^* = \mathbf{a}^{'}\mathbf{x} = a_1x_1 + a_2x_2 + \cdots + a_px_p \tag{3.21}$$

and

$$Y^* = \mathbf{b}^{'}\mathbf{y} = b_1y_1 + b_2y_2 + \cdots + b_qy_q, \tag{3.22}$$

then finding the linear combination of p predictor variables that are maximally correlated with the linear combination of the Y's corresponds to finding vectors $\mathbf{a}$ and $\mathbf{b}$ that maximizes

$$r = \frac{(X\mathbf{a})^{'}(Y\mathbf{b})}{\parallel X\mathbf{a} \parallel\parallel Y\mathbf{b} \parallel} = \frac{\mathbf{a}^{'}\Sigma_{xy}\mathbf{b}}{\sqrt{\mathbf{a}^{'}\mathbf{\Sigma}_{xx}\mathbf{a}}\sqrt{\mathbf{b}^{'}\mathbf{\Sigma}_{yy}\mathbf{b}}}. \tag{3.23}$$

The maximized correlation r is called *canonical correlation* between $\mathbf{X}$ and $\mathbf{Y}$. Since r is invariant under scaling of $\mathbf{a}$ and $\mathbf{b}$, we can make an arbitrary normalization of $\mathbf{a}$ and $\mathbf{b}$. We will require that $\mathbf{a}$ and $\mathbf{b}$ be such that X^* and Y^* have unit variance, that is, $\mathbf{a}^{'}\mathbf{\Sigma_{xx}a} = \mathbf{b'\Sigma_{yy}b} = 1$, and that $E(X^*) = 0$ and similarly $E(Y^*) = 0$. This problem is equivalent to solving the canonical equations:

$$\Big(\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx} - \lambda\mathbf{I}\Big)\mathbf{a} = \mathbf{0} \tag{3.24}$$

and

$$\Big(\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy} - \lambda\mathbf{I}\Big)\mathbf{b} = \mathbf{0} \tag{3.25}$$

where $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{yy}$, $\mathbf{\Sigma}_{yx}$ and $\mathbf{\Sigma}_{xy}$ $(= \mathbf{\Sigma}_{xy}^{T})$ are defined as before in (3.20), $\mathbf{I}$ is the identity matrix, and λ is the largest eigenvalue for the *characteristic equations*

$$|\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx} - \lambda\mathbf{I}| = 0 \tag{3.26}$$

and

$$|\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy} - \lambda\mathbf{I}| = 0. \tag{3.27}$$

The largest eigenvalue of the product matrix $\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}$ or $\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}$ is the squared canonical correlation coefficient r^2. The eigenvectors associated with the eigenvalue λ then become the vector of coefficients $\mathbf{a}$ and $\mathbf{b}$. There are two sets of eigenvectors, one for $\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}$ and one for $\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}$. It can be shown that

$$\mathbf{a} = \frac{\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}\mathbf{b}}{\sqrt{\lambda}} \tag{3.28}$$

and

$$\mathbf{b} = \frac{\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}\mathbf{a}}{\sqrt{\lambda}} \tag{3.29}$$

which means that it is not necessary to solve for both characteristic equations, since the eigenvectors $\mathbf{a}$ and $\mathbf{b}$ are themselves defined already when one of them is known.

3.2.2 Sample-Based Canonical Correlation Analysis

So far I have been considering population variance-covariance matrices $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{xy}$, $\mathbf{\Sigma}_{yx}$ and $\mathbf{\Sigma}_{yy}$. In most applications, however, these matrices are not known. A canonical correlation analysis usually starts with a sample of n measurements on the $(p+q)$ dimensional variable $\mathbf{Z}$=($\mathbf{X}$,$\mathbf{Y}$), that is, with the data matrix

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} & y_{11} & y_{12} & \dots & y_{1q} \\ x_{21} & x_{22} & \dots & x_{2p} & y_{21} & y_{22} & \dots & y_{2q} \\ \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} & y_{n1} & y_{n2} & \dots & y_{nq}. \end{pmatrix}$$

The components of the variance-covariance matrix generated from a data matrix like that shown above are then used to estimate the coefficients of each pair of canonical variates. That is, the two product matrices that are used in the analysis correspond to

$$\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx} \tag{3.30}$$

and

$$\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} \tag{3.31}$$

where $\mathbf{S}_{xx}$, $\mathbf{S}_{xy}$, $\mathbf{S}_{yx}$ and $\mathbf{S}_{yy}$ are, respectively, the sample-based estimates of $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{xy}$, $\mathbf{\Sigma}_{yx}$ and $\mathbf{\Sigma}_{yy}$. Given the necessary inverses, the procedures followed here are precisely the same as those described above for the population model.

Often the measurements collected have different properties, which means that they are not comparable. In such cases the $\mathbf{X}$ and $\mathbf{Y}$ variables are first standardized to have unit variance so that the variance-covariance matrix is a correlation matrix. Following the previous approach, the two product matrices that become the input to the analysis are

$$\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx} \tag{3.32}$$

and

$$\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} \tag{3.33}$$

where $\mathbf{R}_{xx}$ is the correlation matrix formed from considering the $\mathbf{X}$ variables alone, $\mathbf{R}_{yy}$ is the correlation matrix formed from considering the $\mathbf{Y}$ variables alone, and $\mathbf{R}_{xy}$ ($\mathbf{R}_{yx}$) is the correlation matrix obtained from considering both the $\mathbf{X}$ and $\mathbf{Y}$ variables together. The same canonical correlation r^2 will be obtained whether one is using (3.32) and (3.33) or (3.30) and (3.31).

The sample-based estimates of the canonical weights $\mathbf{a}$ and $\mathbf{b}$ will be denoted by $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. When the sample-based estimates of the variance-covariance matrices (3.30 and 3.31) are used, the elements of $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ will be in units proportional to those of the respective responses in each set, and the dimensionality of the respective canonical variables will thus have a meaning. In contrast, canonical variates based on the correlation matrices (3.32 and 3.33) are dimensionless, and thus in computing the correlation-based canonical variates one should use the standardized scores of the original variables.

3.2.3 Statistical Testing

In CCA we want to test the null hypothesis that the q response variables are not linearly associated with the p predictor variables, that is

$$\begin{aligned} H_0 &: \boldsymbol{\Sigma}_{yx} = \mathbf{0} \\ H_1 &: \boldsymbol{\Sigma}_{yx} \neq \mathbf{0}. \end{aligned} \tag{3.34}$$

To do that, we need to define a suitable test statistic called Wilks's lambda Λ as follows:

$$\Lambda = \prod_{j=1}^{M}(1 - \hat{\lambda}_j) = \frac{|\mathbf{S}|}{|\mathbf{S}_{xx}||\mathbf{S}_{yy}|} \tag{3.35}$$

where $M = \min(p, q)$, and λ_j is the jth largest eigenvalue for the characteristic equations (in 3.26 and 3.27). Bartlett's χ^2 approximation for the distribution of Λ is derived for

$$X^2 = -[(n-1) - \frac{1}{2}(p+q+1)]\ln\Lambda, \tag{3.36}$$

which under the H_0 in 3.34 follows χ^2_{pq} distribution.

3.3 The relationship between Canonical correlation analysis and Multivariate linear model

Theory in this section is based on chapter 11.6 in [Rencher and Christensen, 2011]. To reveal the relationship between multivariate linear model and CCA, let us examine the linear model with one response variable X, and two predictor variables Y_1 and Y_2: $X = \alpha_1 Y_1 + \alpha_2 Y_2 + \epsilon$. In this model the proportion of the total variation in the response variable X that can be attributed to regression on the Y's is denoted by R^2:

$$\begin{aligned} R^2 &= \frac{\text{regression sum of squares (SSR)}}{\text{total sum of squares (SST)}} \\ &= \frac{\hat{\alpha}'\mathbf{Y}'\mathbf{X} - n\bar{\mathbf{X}}^2}{\mathbf{X}'\mathbf{X} - n\bar{\mathbf{Y}}^2}. \end{aligned} \tag{3.37}$$

This ratio R^2 is called the *squared multiple correlation*. This can also be expressed in terms of sample variances, covariances and correlations:

$$R^2 = \frac{\mathbf{s'}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{s}_{xy}}{\mathbf{s}_{xx}} = \mathbf{r'}_{xy}\mathbf{R^{-1}}_{yy}\mathbf{r}_{xy}, \tag{3.38}$$

where $\mathbf{S}_{xx}, \mathbf{S}_{xy}$ and $\mathbf{S}_{yy}$ are defined in 3.31 and $\mathbf{r}_{xy}$ and $\mathbf{R}_{yy}$ are from analogous partitioning of the sample correlation matrix of the $y's$ and the x:

$$\mathbf{R} = \left(\begin{array}{c|cc} 1 & r_{x,y_1} & r_{x,y_2} \\ \hline r_{y_1,x} & r_{y_1,y_1} & r_{y_1,y_2} \\ r_{y_2,x} & r_{y_2,y_1} & r_{y_2,y_2} \end{array}\right) = \begin{pmatrix} 1 & \mathbf{r'}_{yx} \\ \mathbf{r}_{yx} & \mathbf{R}_{yy} \end{pmatrix}$$

The F-test for overall regression can be expressed in terms of R^2 as

$$F = \frac{n-q-1}{q}\frac{R^2}{1-R^2}. \tag{3.39}$$

Canonical correlation can be defined as an extension of this multiple correlation R^2. When one of the two sets of variables has only one variable, canonical correlation reduces to multiple correlation. For example, when $p = 1$, $\mathbf{R}_{xx}$ becomes 1, and the single squared canonical correlation reduces to $r^2 = \mathbf{r'}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{r}_{xy}$, which can be recognized as R^2. As the statistical testing in CCA is in practice testing the significance of the canonical correlation coefficient, it is equivalent to F test, which tests the significance of multiple correlation coefficient. It can be shown that F test and Wald tests are asymptotically equivalent, and since the Wald test yields the same results no matter which way you treat the association (whether X is the response variable or the Y's), we can say that the Wald test and CCA are asymptotically equivalent in case of one variable in the other group of variables.

Chapter 4

Results

4.1 Simulations

In this section I will introduce the simulations I did, and the results I got from them. I had two motivations to do these simulations: first I wanted to illustrate the nature of multivariate Wald test compared to the corresponding univariate Wald tests i.e. in what kind of situations the multivariate test is more useful than the univariate test. The second motivations was to examine whether multivariate Wald test yields the same results as canonical correlation analysis. Simulation scenarios are adapted from [Stephens, 2013]. The R code I used for these simulations, and their statistical analyses is in Appendix A.

To illustrate these two things, I made bivariate simulations in which two phenotypes, Y_1 and Y_2 are associated in varying ways with SNP genotypes g and with each other. Each simulation scenario is defined by three parameters, (β_1, β_2, ρ), which denote, respectively, the effects of genotype g on Y_1 and Y_2, and the correlation coefficient of Y_1 and Y_2. I simulated datasets of 1,000 individuals, where for each individual i I simulated 1,000 genotypes from the distribution $g_i \sim \text{Bin}(2, 0.2)$, that is, the minor allele frequency for all of these 1,000 genotypes is 0.2.

After genotype simulations I simulated bivariate phenotypes (Y_1, Y_2) for every SNP from the distribution $(Y_{i1}, Y_{i2})|g_{i,j} \sim N_2(\mu_i, \Sigma)$ where $\mu_i = (\beta_1 g_i, \beta_2 g_i)$ and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. I fixed $\beta_2 = 0.2$ (Y_2 is associated with g) and considered two different levels of correlation $\rho = (0.3, 0.7)$, β_1 had three different values, $(-0.2, 0, 0.2)$ (Y_1 is associated, not-associated and associated, respectively, with g.). Thus, there was total of 12,000 phenotypes for 1,000 individuals (6 bivariate phenotype-pairs for 1,000 individuals). The simulation scenarios were as follow:

1. simulation: $\beta_1 = -0.2$, $\beta_2 = 0.2$ and ρ= 0.3 (opposite directions, small correlation)
2. simulation: $\beta_1 = 0$, $\beta_2 = 0.2$ and ρ= 0.3 (one effect, small correlation)
3. simulation: $\beta_1 = 0.2$, $\beta_2 = 0.2$ and ρ= 0.3 (same direction, small correlation)
4. simulation: $\beta_1 = -0.2$, $\beta_2 = 0.2$ and ρ= 0.7 (opposite directions, large correlation)
5. simulation: $\beta_1 = 0$, $\beta_2 = 0.2$ and ρ= 0.7 (one effect, large correlation)
6. simulation: $\beta_1 = 0.2$, $\beta_2 = 0.2$ and ρ= 0.7 (same direction, large correlation)

The non-genetic variance of every trait is simulated to be 1, thus the simulated standard deviation of the traits are $\sqrt{1} = 1$. If the genetic effect size (0.2 or -0.2) is compared to that, we notice that it is one fifth of the trait's non-genetic standard deviation, which is a quite remarkable effect in GWAS, but it was chosen to demonstrate the methods using a sample 1,000 individuals.

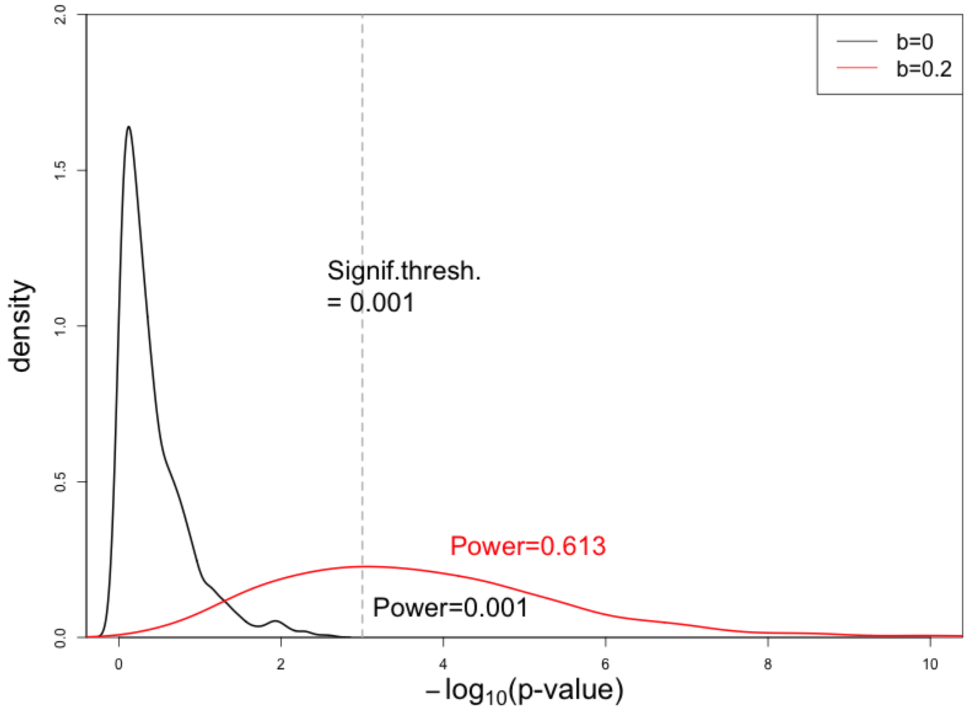


Figure 4.1: Demonstration of the power in the univariate tests with the significance threshold of 0.001, and the effect of 0 and 0.2. The lines represents the density plots for 1,000 simulated tests, x-axis being the -$\log_{10}$(p-value).

The power of the test is the probability that when there is a true effect the test statistic will reach the given threshold. Power depends on the sample size n, allele frequency f and the effect size β. For quantitative traits power increases with $nf(1-f)\beta^2$ [Sham and Purcel, 2014]. The power to detect the effect in these simulation scenarios (0.2 or -0.2) when used significance threshold 0.001 is 0.613, and in the case of no effect ($\beta = 0$) the power is 0.001. That is illustrated in the Fig. 4.1.

4.1.1 Comparing univariate and multivariate Wald tests

I calculated univariate Wald test statistics for every simulated phenotype-SNP-pairs and then I calculated corresponding p-values from standard normal distribution, and -log$_{10}$(p-values). Distributions of -log$_{10}$(p-values) I got from univariate Wald tests are in Fig. 4.2 showing that when there is an effect (0.2 or -0.2) the boxplots are very similar, and when there is no effect the p-values are also very large (-log$_{10}$(p-values) are small), as would be expected. Boxplots with the same colour represents the tests against phenotypes from the same simulation scenarios

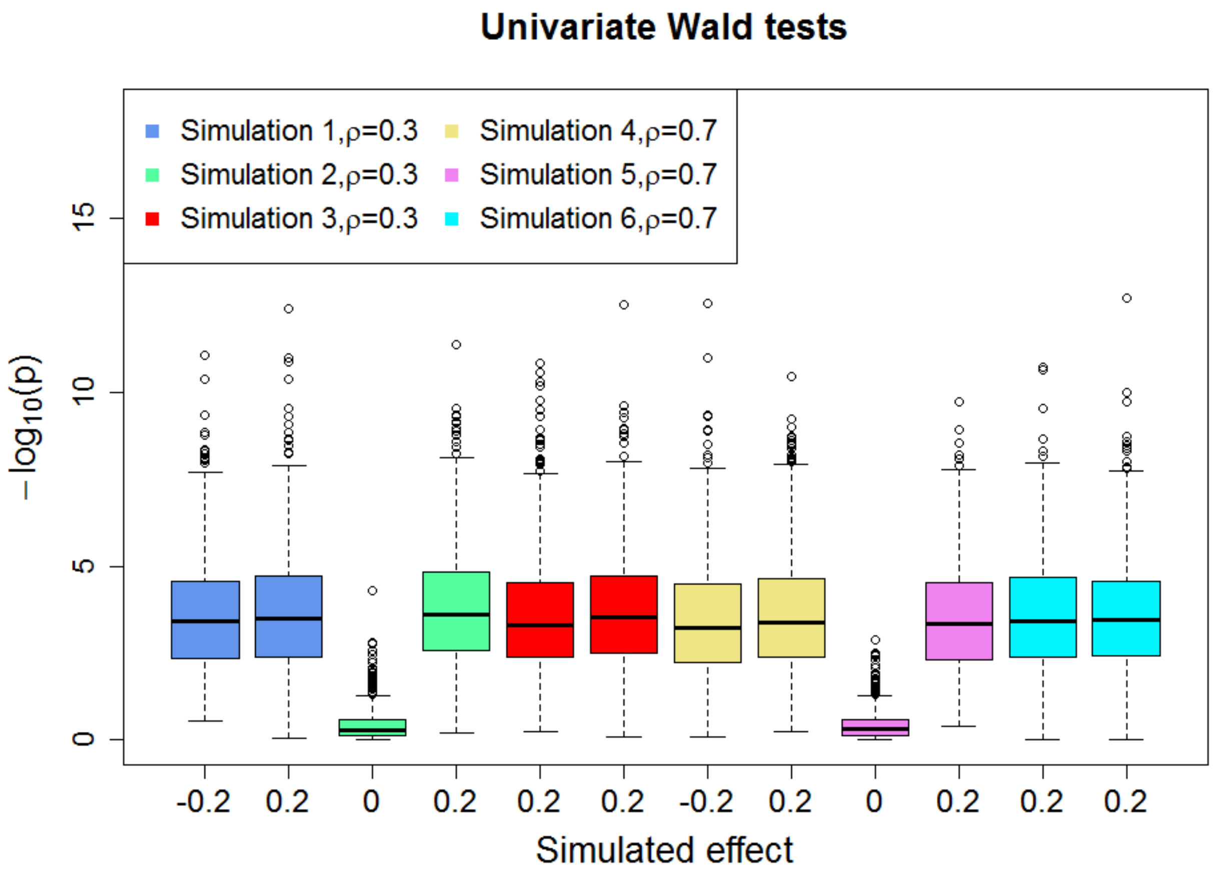


Figure 4.2: Univariate Wald test -log$_{10}$(p- values). Each boxplot represents a -log$_{10}$(p- values) for one genotype-phenotype pair, thus in each boxplot there is the results from 1,000 tests.

After univariate tests I calculated multivariate Wald test statistics for every simulated SNP-bivariate phenotype- pairs, and calculated corresponding p-values from χ^2-distribution with two degrees of freedom, and then -$\log_{10}$(p-values). Distributions of -$\log_{10}$(p-values) I got are in Fig. 4.3, where one boxplot represents one phenotype-genotype pair (two in each simulation scenarios). Thus in each boxplot there are -$\log_{10}$(p-values) for 1,000 tests. Colour of the boxplots represents the relationship of the two effects (opposite direction, same direction or single effect), and the simulated non-genetic correlation in each scenario is on x- axis. It can be seen that the multivariate Wald test is most efficient in cases, where the two effects have opposite directions (-0.2 and 0.2), especially when correlation between two phenotypes is large (0.7, the fourth boxplot).

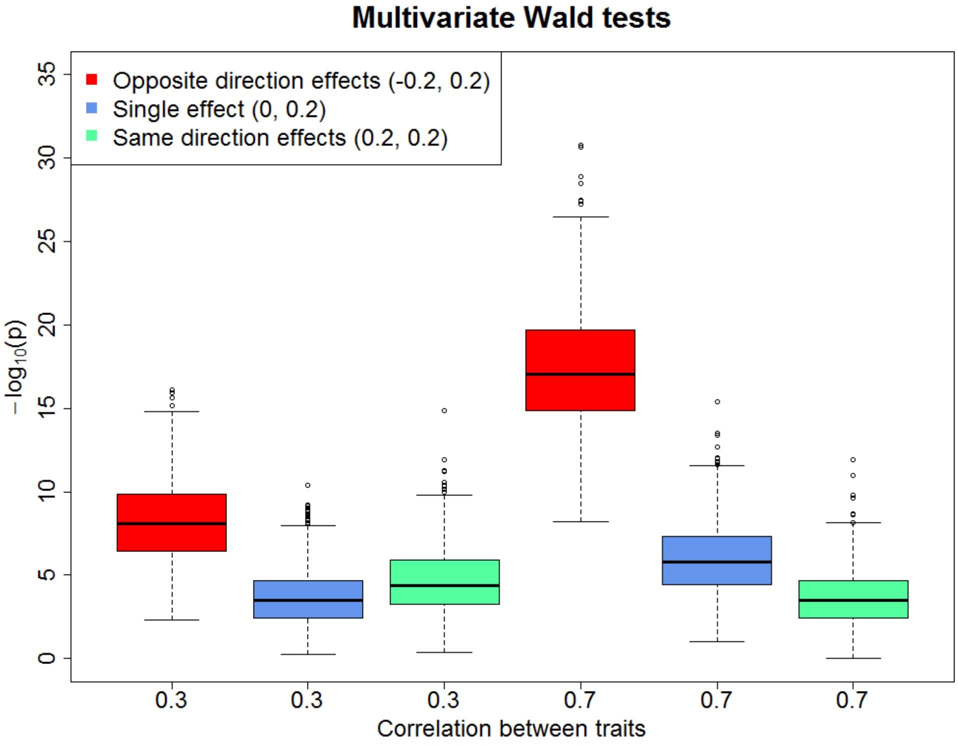


Figure 4.3: Multivarite Wald test -$\log_{10}$(p)- values. Every boxplot represents -$\log_{10}$(p values) of one simulation scenarios, and thus in every boxplot there is 1,000 observations.

I compared the univariate tests, and corresponding multivariate tests in Fig. 4.4 showing that multivariate test (blue boxplots) are most efficient compared to univariate tests (green boxplots) when the two effects have opposite directions. When the effects have same direction, it seems that the multivariate test is most efficient when the correlation between the traits is small, and even then barely more efficient than either of the corresponding univariate tests.

When g has effect only on the other phenotype the multivariate test is most efficient compared to univariate test when the correlation between the traits is large. When the correlation

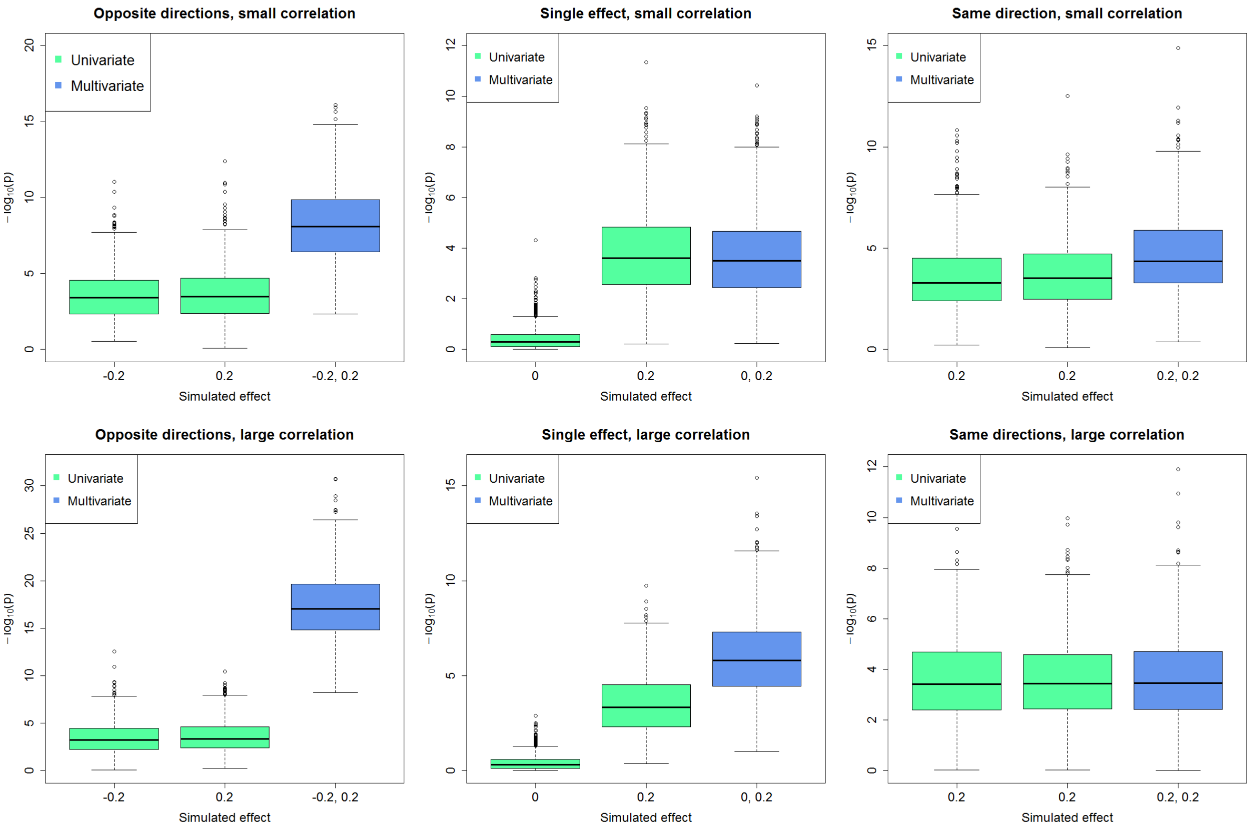


Figure 4.4: Univariate vs. Multivariate Wald test results by simulation scenarios.

between the traits is small, the multivariate test is about as effective as univariate test for the trait that genotype has effect on. From Fig. 4.4 can also be seen that the multivariate test is never much worse in terms of power than corresponding univariate tests.

These same results can be seen from Table 4.1, where I have calculated the percentage of the cases, where multivariate Wald test yields smaller p-value (larger -$\log_{10}$(p-value)) than either of the corresponding univariate Wald tests. When effects have opposite directions, that percentage is 100 in both small and large correlation cases. Thereby it seems that multivariate Wald test is very useful compared to univariate tests when the effects have opposite directions.

Table 4.1: Comparison of univariate and multivariate Wald tests. Percentages reported here tells when the multivariate test yielded smaller p-value than either of the corresponding univariate tests.

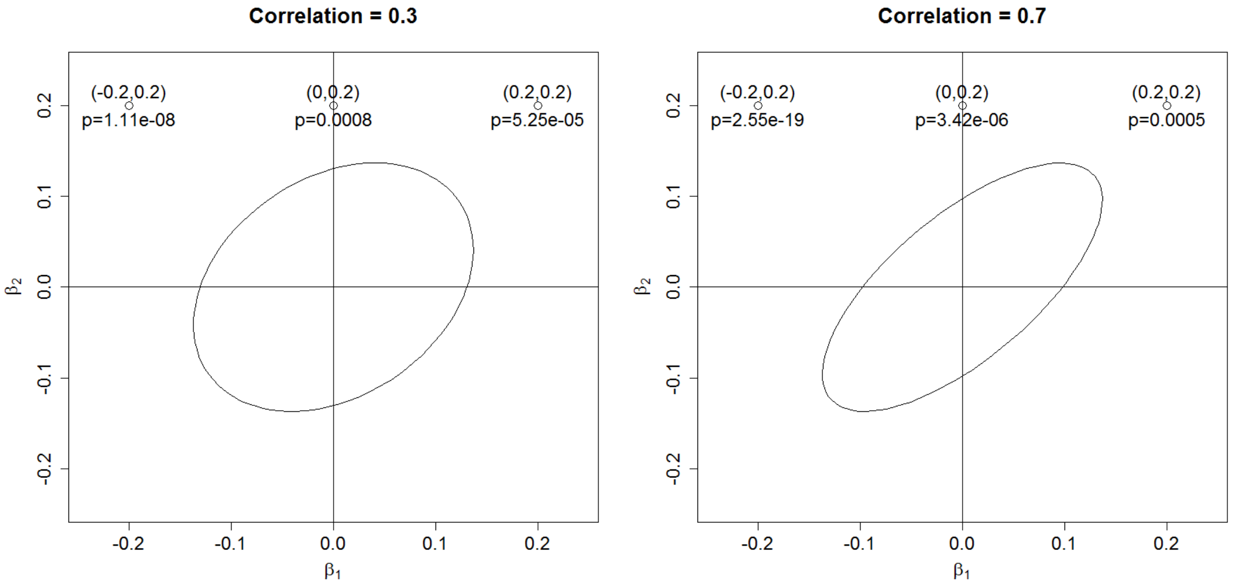
Simulation	%	Effects	Correlation
1	100	Opposite directions	0.3
2	31.2	Single	0.3
3	60.4	Same direction	0.3
4	100	Opposite directions	0.7
5	97.4	Single	0.7
6	2.4	Same direction	0.7

When g has effect only on the other trait, multivariate Wald test yields smaller p-value in 31.2% of cases when correlation between the traits is 0.3, and 97.4% of cases when correlation between the traits is 0.7. When the two effects have same direction, multivariate Wald test yields smaller p-value than either of the corresponding univariate tests in 60.4% of cases when the correlation between the traits is 0.3, and 2.4% when that correlation is 0.7.

The small percentages seen here, such as 2.4 and 31.2 might give the impression that the multivariate test in these cases is less efficient than corresponding univariate tests, but the Fig. 4.4 shows that multivariate test in practice yields almost as small p-value as either of the corresponding univariate tests.

Intuitive explanation for the pattern

The intuitive explanation for this pattern described above can be given by considering the null models for the effects. In Fig. 4.5 are the 95% highest probability regions (areas inside the ellipses) for the null model in case of both correlations in my simulations, 0.3 and 0.7. So in case of no genetic effects (that is both β_1 and $\beta_2 = 0$), 95% of the cases would be expected to be within these ellipses.



(a) Small correlation

(b) Large correlation

Figure 4.5: 95% highest probability regions for the null models. Each three different combinations of the genetic effects (opposite directions, same direction and single effect) are demonstrated in the figure.

Fig. 4.5a shows the 95 % confidence ellipse of the null model for two genetic effects in case the correlation between the traits is 0.3. I also plotted the three dots to demonstrate the effect pairs in my simulations, and the corresponding theoretical p-values for each of the cases to also show the mathematical background for the pattern besides the figure. From this figure it can be seen that the case of opposite directions (-0.2, 0.2) is most deviated from this ellipse (which means it will give the smallest p-value as the idea in the test is to test how well the result fits to the null model), and the single effect case (0, 0.2) is closest to this ellipse. This demonstrates the results described earlier: multivariate test is most efficient in the case of opposite directions, and least effective in the case of a single effect when the correlation between traits is small.

Fig. 4.5b also shows the 95 % confidence ellipse for the null model, as Fig. 4.5a, but in this case the correlation between the traits is 0.7. In this case the opposite direction effect situation is again the most deviated from the ellipse, and the same direction effect situation (0.2, 0.2) is closest to model. This demonstrates the results described earlier: multivariate test was most efficient in the case of opposite directions, and least effective in the same direction effect cases when the correlation between the traits is large.

When the correlation between the traits approaches to zero, the confidence ellipse gets more and more close to circle. In extreme situation where the correlation is zero, and the confidence ellipse is a circle, points (0.2, 0.2) and (-0.2, 0.2) would be equally deviated from the ellipse, and in that case the multivariate test would be equally efficient in these two cases, and least effective in the case of a single effect.

When the correlation between the traits approaches to 1, the confidence ellipse gets very narrow, and eventually when the correlation is 1, it is a line. In that case the point (-0.2, 0.2) would be most deviated from that line, and the point (0.2, 0.2) would be closest to that line. In that case the multivariate test would be most efficient in the case of opposite direction effects, and least effective in the case of same direction effects.

4.1.2 Comparing multivariate Wald test and Canonical correlation analysis

The other motivation for these simulations was to compare multivariate Wald test and Canonical correlation analysis and to see whether they yield the same results, that is, whether they yield the same p-value for the significance of the association between genotype and phenotypes. I executed the CCA for all the bivariate phenotype-genotype pairs as explained in Methods. The -$\log_{10}$(p-values) I got from these calculations are in Fig. 4.6. It seems that those results are similar to those of multivariate Wald test in Fig. 4.3. To check this more

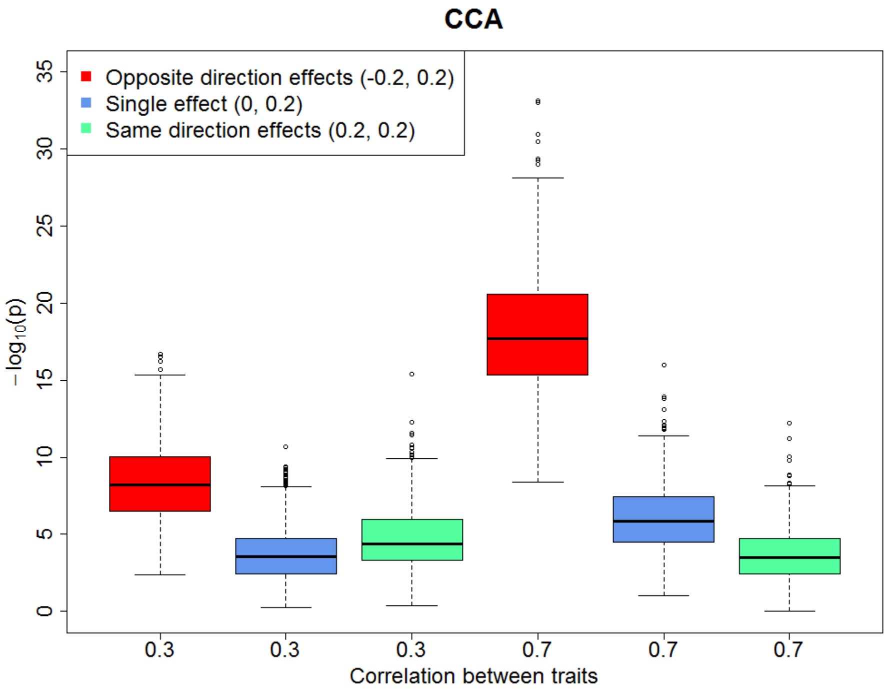


Figure 4.6: Canonical correlation analysis -$\log_{10}$(p-values). Every boxplot represents -$\log_{10}$(p-values) of one simulation scenarios, and thus in every boxplot there is 1,000 observations.

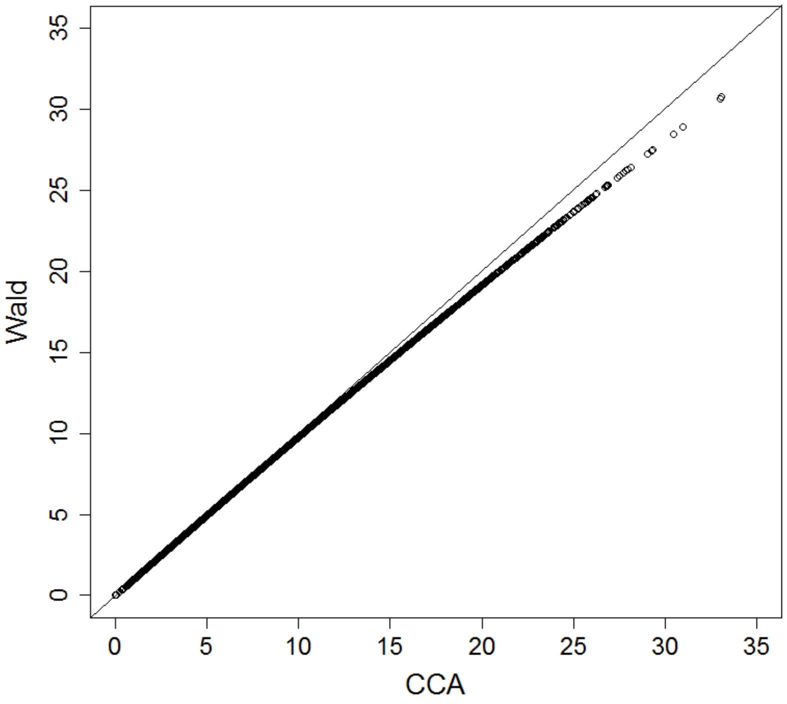


Figure 4.7: CCA vs. Wald Multivariate -$\log_{10}$(p-values).

closely I plotted -$\log_{10}$(p-values) from both of these models against each other in Fig. 4.7. From that figure it can be seen that they are not exactly the same, especially when p-values get very small (-$\log_{10}$(p-values) get large). These differences are most likely due to differences in asymptotics of approximations in the two methods, and have very little practical consequence. In genome-wide association studies, the most commonly used threshold for statistical significance of the effect is 5×10^{-8}, and from the figure we can see that the -$\log_{10}$(p-values) are the same until that threshold . Thus for all practical purposes it does not make a difference if one is using multivariate Wald test or Canonical correlation analysis.

4.2 Data from The National FINRISK Study

In addition to simulation studies, to illustrate the nature of multivariate Wald test compared to univariate Wald test and CCA I did the same analysis for real data as for simulated data described in the previous section. I used data from The National FINRISK Study, which is a large Finnish population survey on risk factors on chronic, non-communicable diseases coordinated by National Institute for Health and Welfare (THL). The survey is carried out every five years using independent, random and representative population samples across Finland [FINRISK homepage].

The National FINRISK Study was earlier known as the North Karelia Project and it was part of the World Health Organization MONICA Project (FINMONICA) in 1982-1992. The research work of the project starting from 1972 is called The National FINRISK Study [FINRISK homepage]. In my study, I had access to The National FINRISK Study 1992–2012 collections, with 20,626 individuals at the beginning. 1,792 individuals were excluded from the analysis because 1,512 of them did not have any lipid measurements, and additional 280 individuals did not have LDL-cholesterol measurements available. Thus the final number of individuals in my analyses was 18,834.

I considered 3 traits into my analysis, HDL-cholesterol, LDL-cholesterol and triglycerides (TG), and I examined 157 SNPs that have been previously reported to be associated with blood lipid levels in [Willer et al., 2013]. The use of genomic data for these 157 genomic regions together with lipid measurements has been approved by the FINRISK Management Group (Applicant: M. Pirinen; Project: #60/2015 ”Fine-mapping genomics regions associated with lipid levels”). The complete list of SNPs I used, and their features is in Table B.1. My traits were inverse rank normalized, so the βs in my analyses are directly comparable with standard deviations. First I examined correlations between the three traits that are listed below:

- cor(HDL, LDL)= -0.0781

- cor(HDL, TG)= -0.4334

- cor(LDL, TG)= 0.2049

For each of the 157 SNPs in the data I executed three univariate Wald tests againts each three traits (HDL, LDL and TG), three bivariate Wald tests (HDL and LDL, HDL and TG and LDL and TG), and one multivariate test against all three phenotypes, and also corresponding CCA. In Figures 4.8-4.10 are the results for the univariate tests. In these figures the size of the bar represents the significance of the SNP ($-\log_{10}$(p-values)), the colour of the bar represents the direction (with respect to minor allele) of the effect (red=negative and green= positive) and the colour of the background represents the minor allele frequency, MAF (frequency of the least common allele): dark grey= MAF $\geqslant$0.05 and light grey= MAF<0.05. These results are also in Table C.1.

In all of these barplot-figures I have cut the -$\log_{10}$(p-values) at 10 for clarity. That means that if some p-value was smaller than 10e^{-10}, its -$\log_{10}$(p-value) is in these figures still 10. These results are sorted by p-value, and the name of the bar is the name of the gene that has been reported for that particular SNP by [Willer et al., 2013].

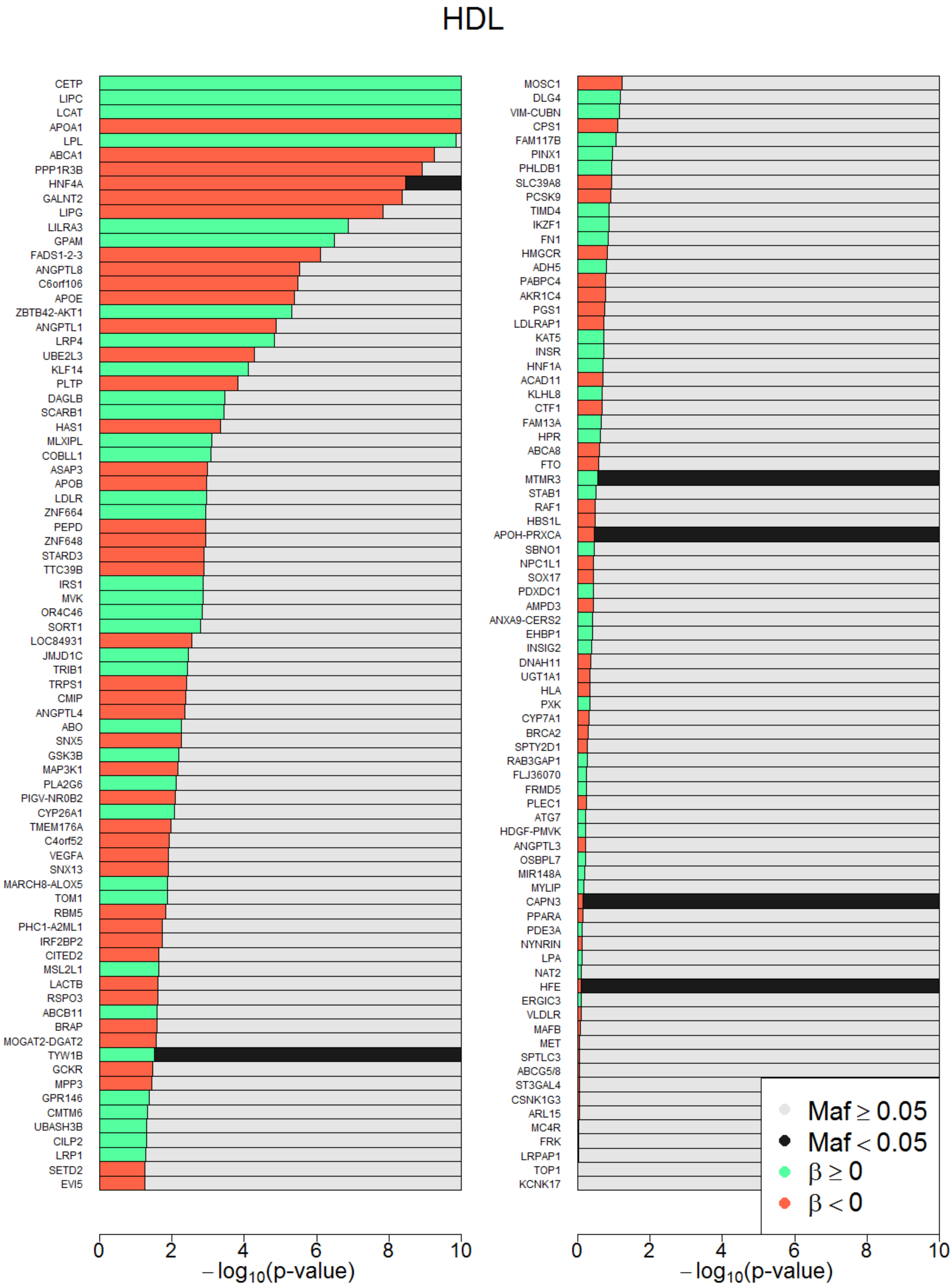


Figure 4.8: Results for univariate tests (HDL)

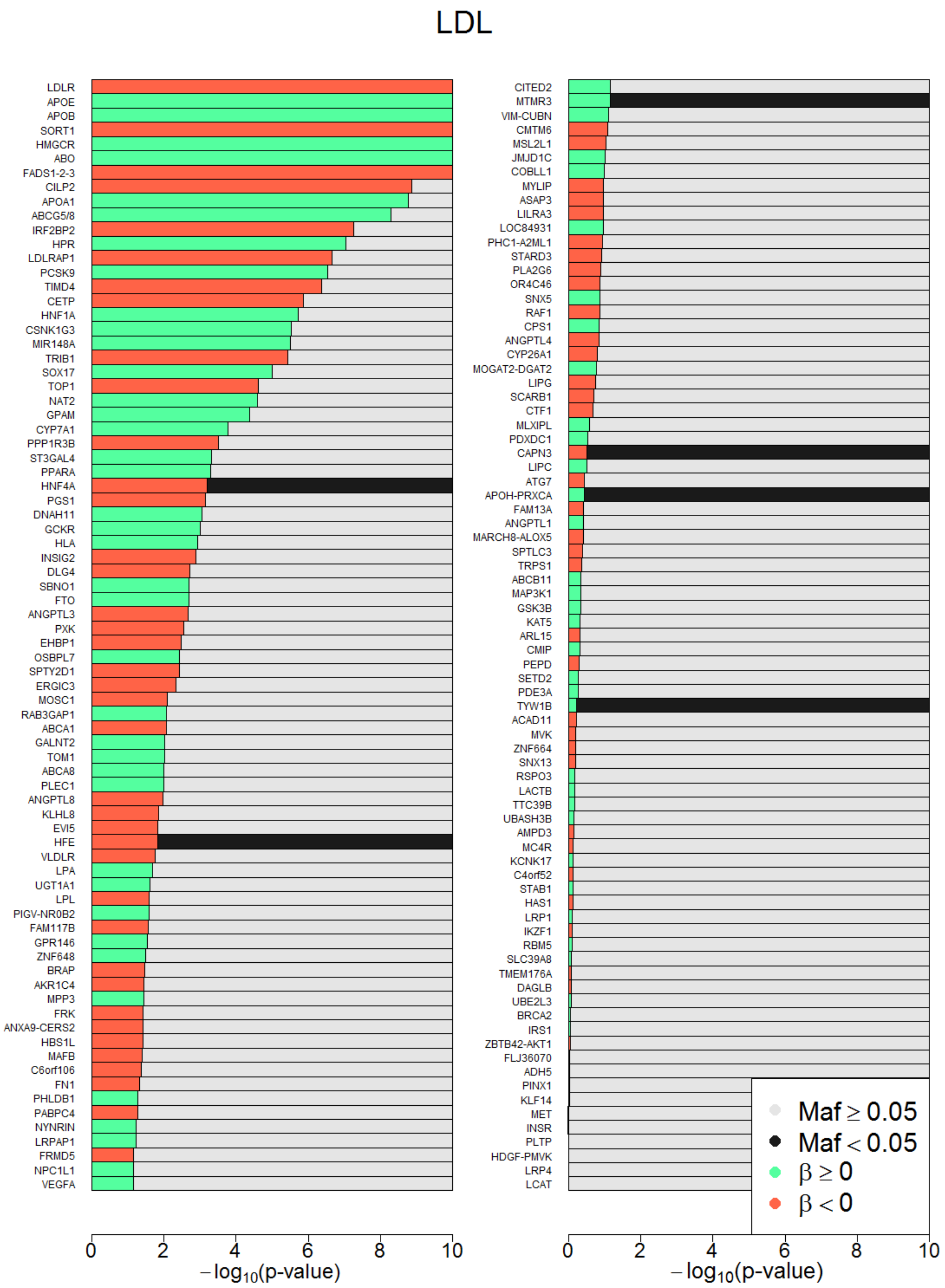


Figure 4.9: Results for univariate tests (LDL)

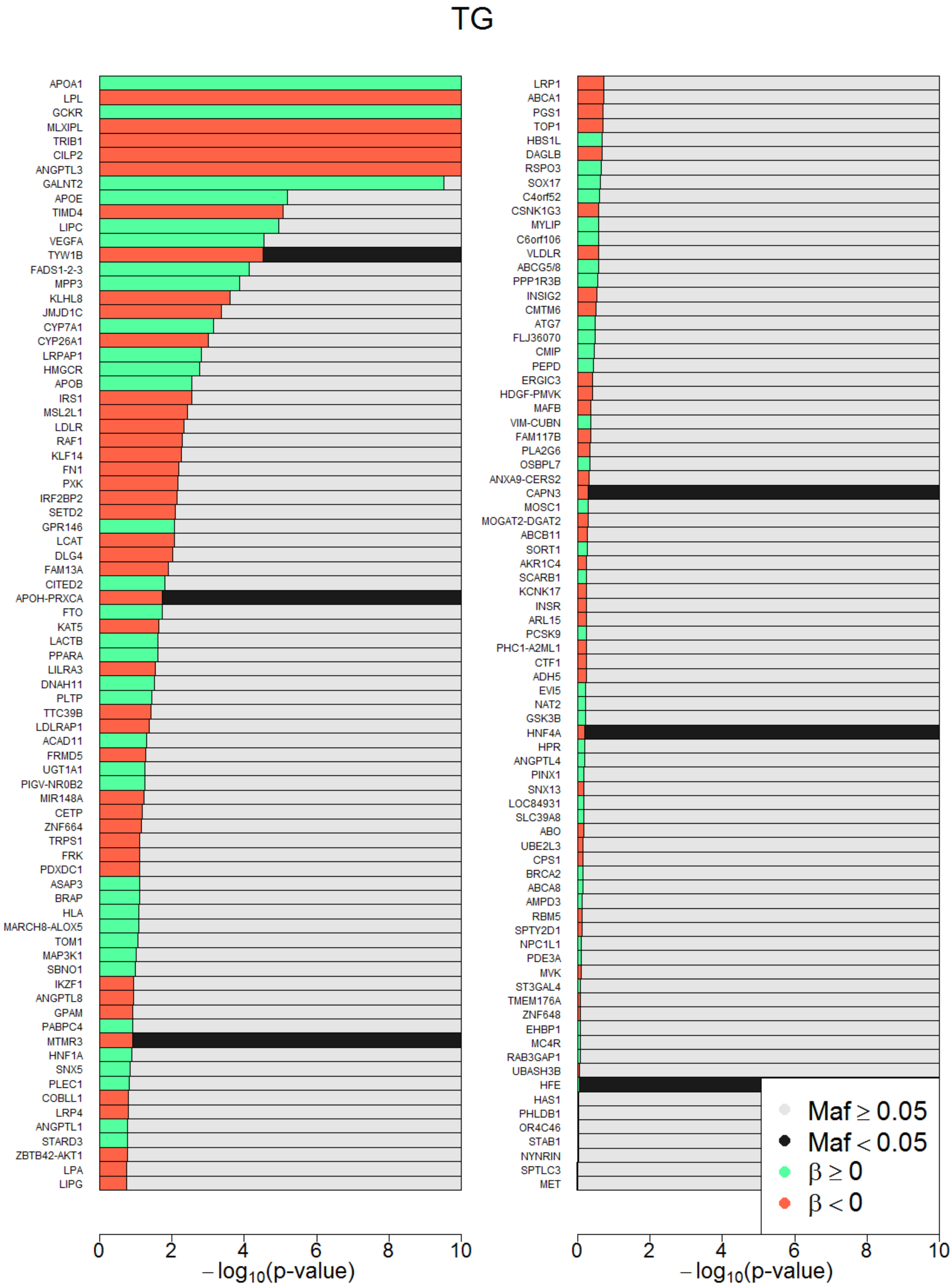


Figure 4.10: Results for univariate tests (TG)

After univariate tests I executed the multivariate tests for every trait-pairs (HDL-LDL, HDL-TG and LDL-TG) and also for all three phenotypes. These results are in Figures 4.11-4.14. In these figures the size of the bar represents, as in the univariate case, the significance of the SNP, and the colour of the bar represents the relationship of the two effects (opposite direction, same direction, single effect or no effect). In this case I have defined the β as an effect, if it's absolute value is larger than 0.02, so the case "no effect" means that both β_1 and β_2 are smaller or equal to 0.02. In the case of 3 traits, there is no colour coding for the bars. In that case again, the size of the bar represents the the significance of the SNP.

From these figures it can be seen that the p-values seems to be smaller in multivariate cases than in univariate cases (the bars are longer), especially in the case of 3 traits. It can also be seen from the results for 2 traits that the most significant results seem to arise from either opposite direction effects cases, or single effect cases, which supports the results from my simulations studies.

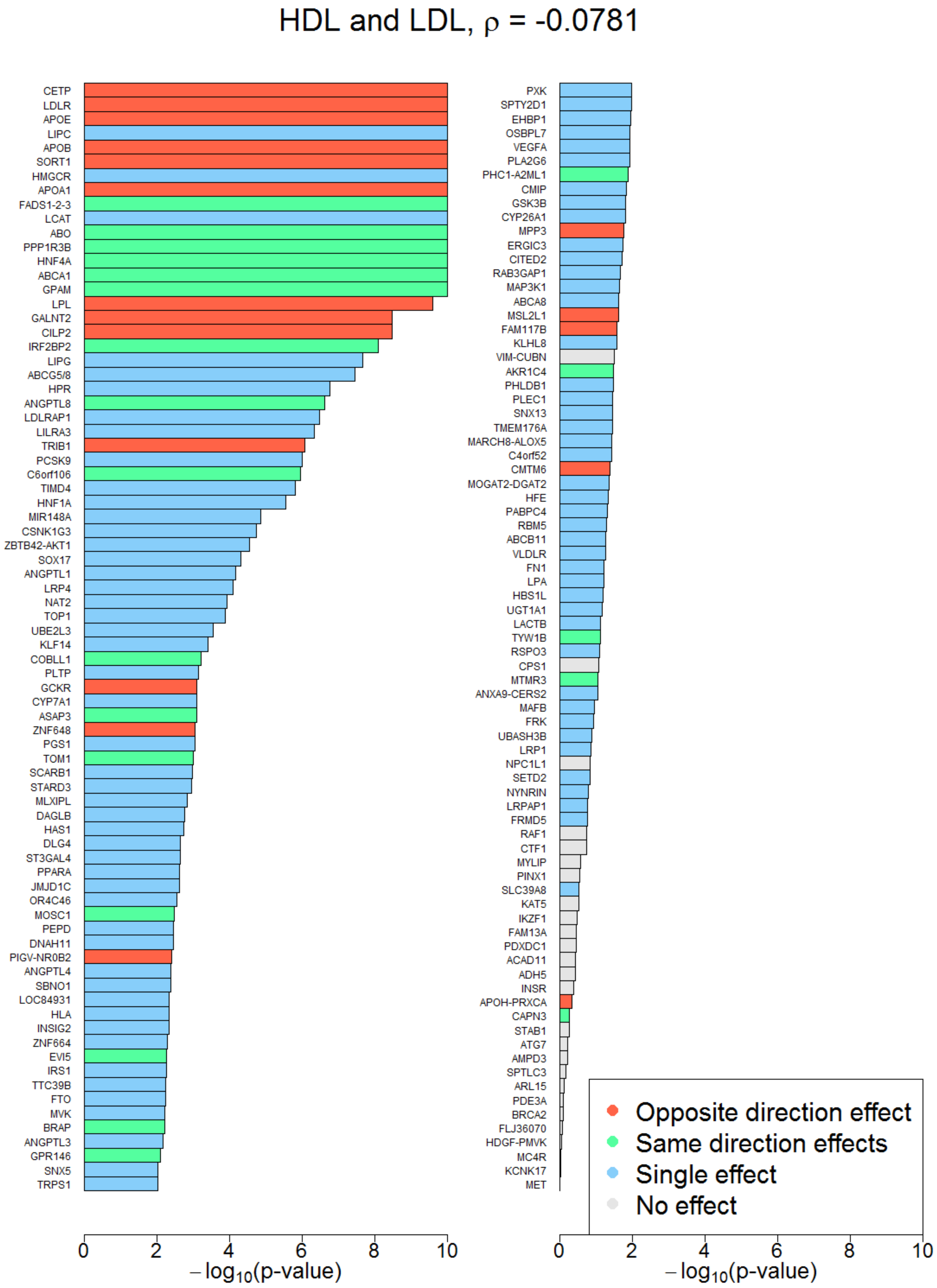


Figure 4.11: Results for multivariate tests for 2 traits (HDL and LDL)

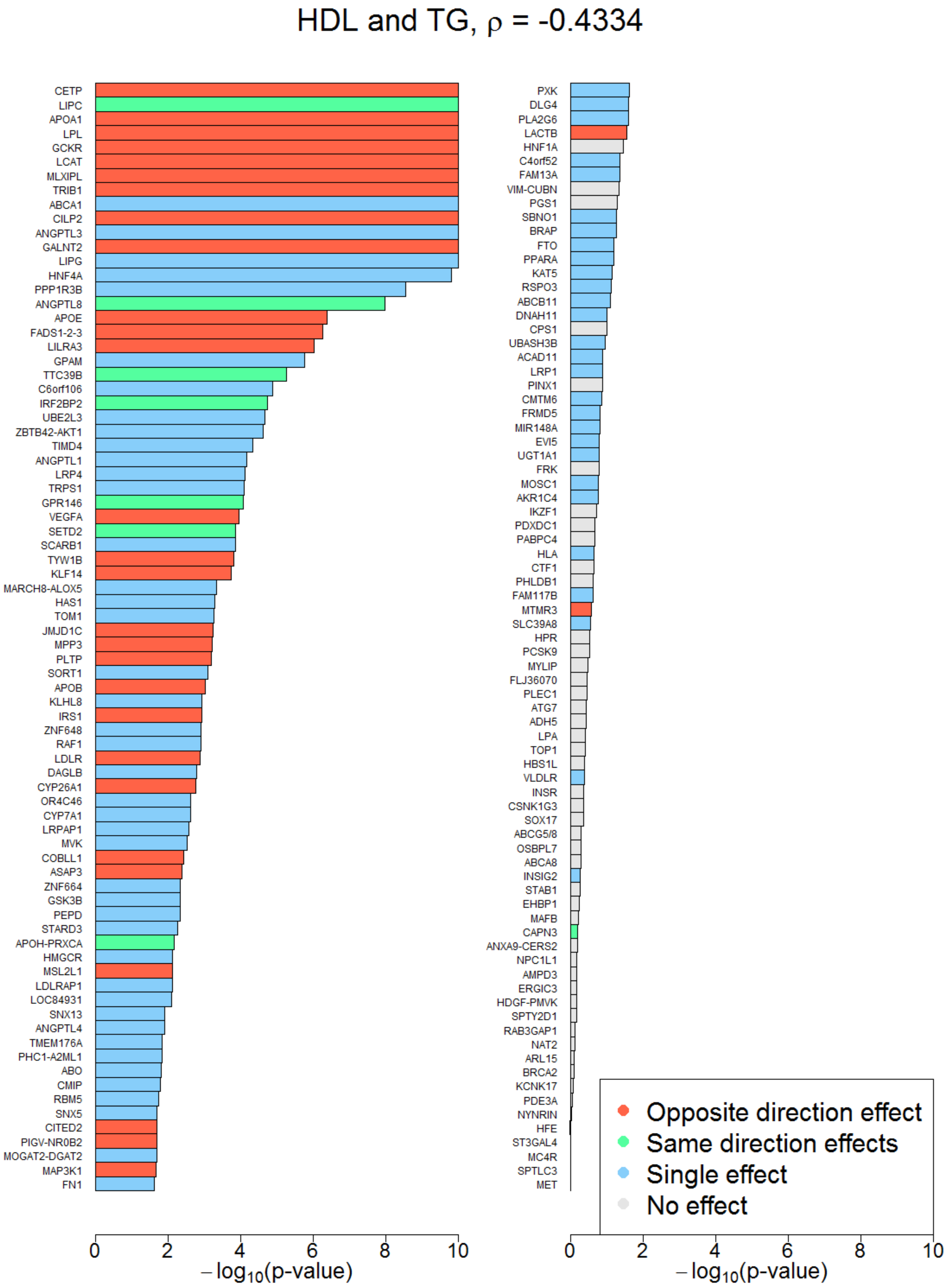


Figure 4.12: Results for multivariate tests for 2 traits (HDL and TG)

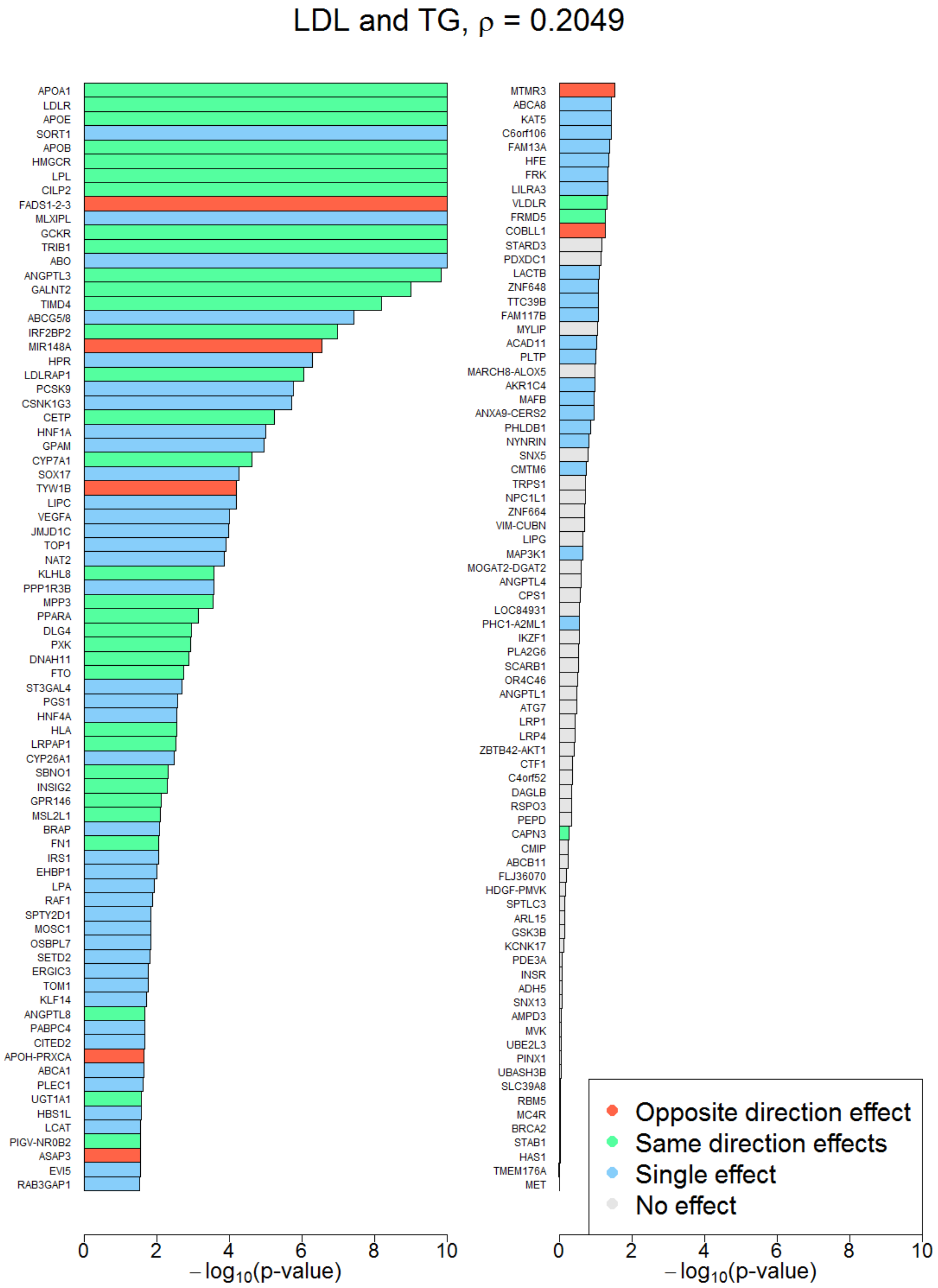


Figure 4.13: Results for multivariate tests for 2 traits (LDL and TG)

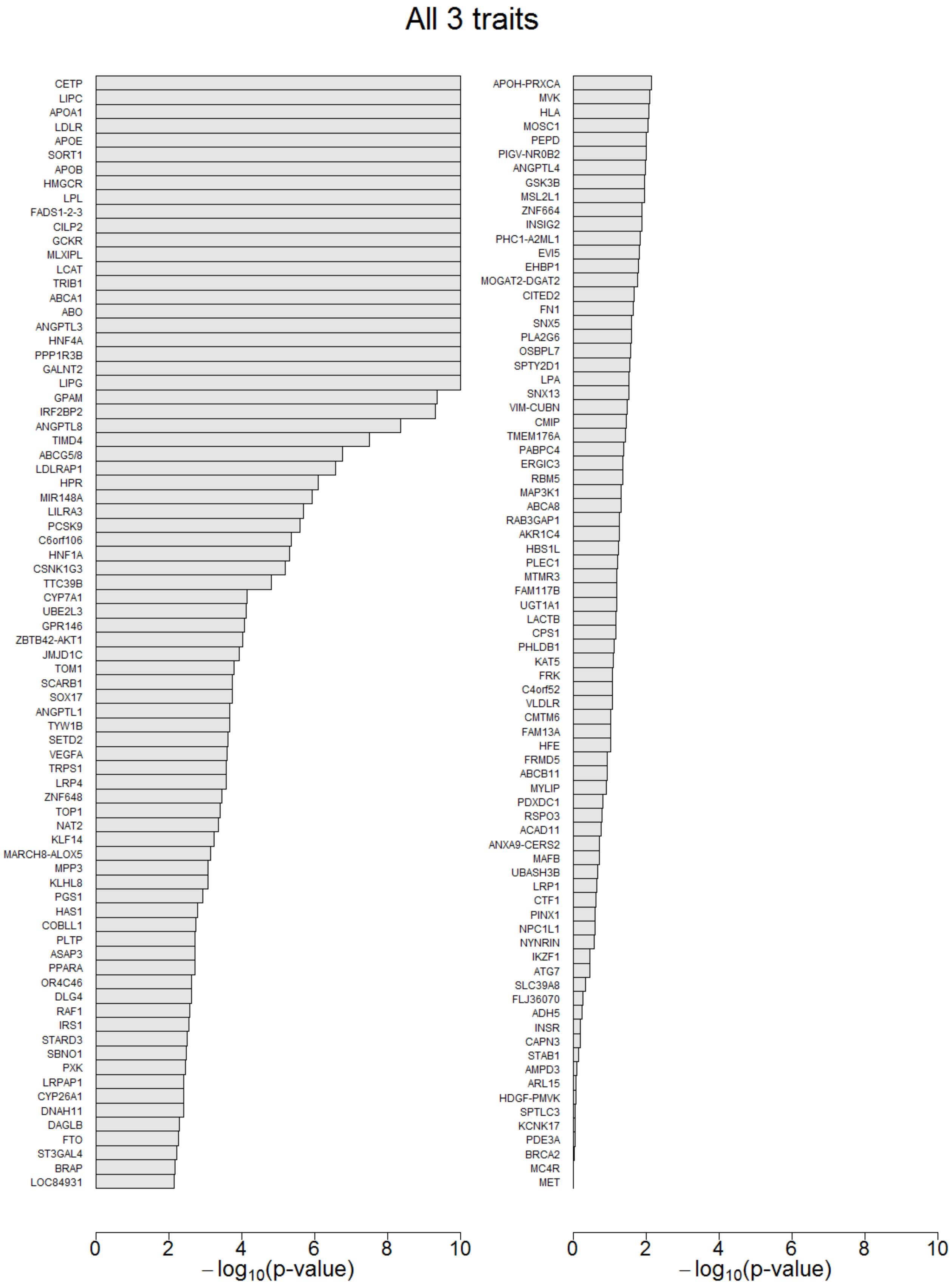
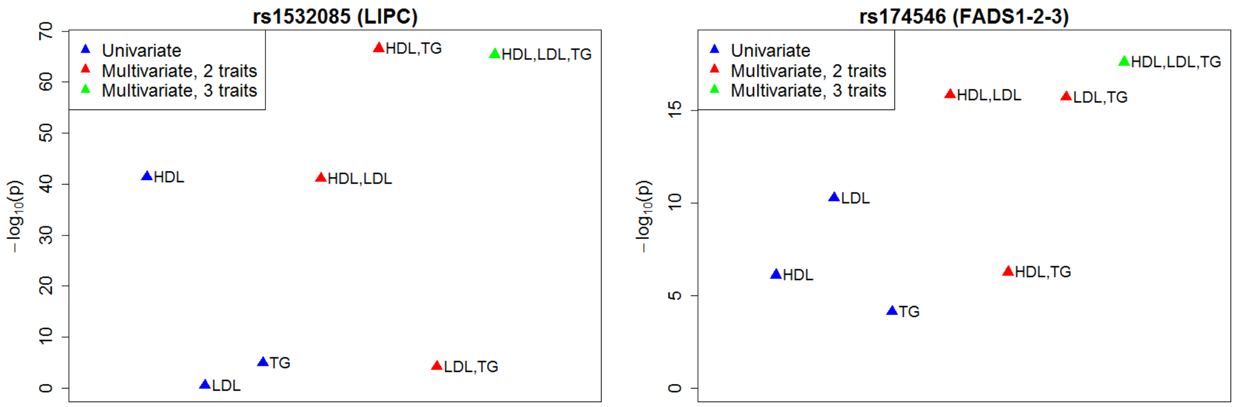


Figure 4.14: Results for multivariate tests for 3 traits (HDL, LDL and TG)

Next I looked the 157 SNPs in the data individually to examine how multivariate tests act compared to univariate tests and chose a couple of them as examples. First example is rs1532085 (in LIPC gene) in Fig. 4.15a, which has the same direction of effect on HDL and TG. The correlation between HDL and TG was -0.4334 (negative and quite large). In this case we can see that the multivariate test against HDL and TG is much more efficient than either of the univariate tests. This case is the same as if there was a positive correlation between traits and the effects would have opposite directions, because in the case of negative correlation the null model ellipse would be a mirror image of the ones in Fig.4.5.



(a) Example for the case of same direction effect on HDL and TG

(b) Example for the case of opposite direction effect on LDL and TG

Figure 4.15: Examples from The National FINRISK Study

The other example is rs174546 (in FADS1-2-3 gene), which has opposite direction effects on LDL and TG (in Fig. 4.15b). The correlation between these traits was 0.2049 (positive and quite small). It can be seen that in this case the multivariate test against LDL and TG is more efficient than either of the corresponding univariate cases. This supports the results already seen in the simulation study: even with quite small positive correlation between traits but in case of opposite direction effects the multivariate test is more efficient than the univariate tests.

To illustrate the nature of multivariate test compared to univariate tests more generally I plotted all the $-\log_{10}$(p-values) from the three dimensional multivariate tests against the minimum of the corresponding univariate tests (that is the one of the three univariate tests that yields the largest p-value), and also for the minimum of the corresponding multivariate tests against two traits. These plots are in Fig. 4.16.

In these figures the x-axis and the y-axis have been cut from value 30. It can be seen that the multivariate test for three traits has never higher p-value than the least effective corresponding univariate test, or least effective multivariate test for two traits.

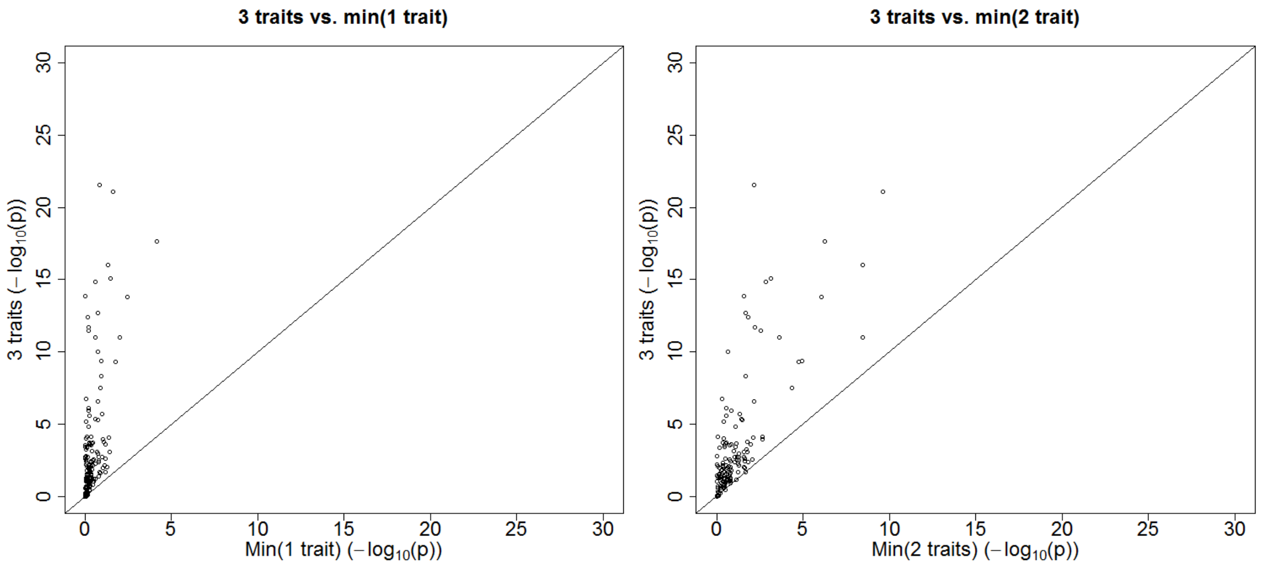


Figure 4.16: Multivariate test for 3 traits vs. minimum of univariate tests and minimum of multivariate tests for 2 traits on -log$_{10}$ scale.

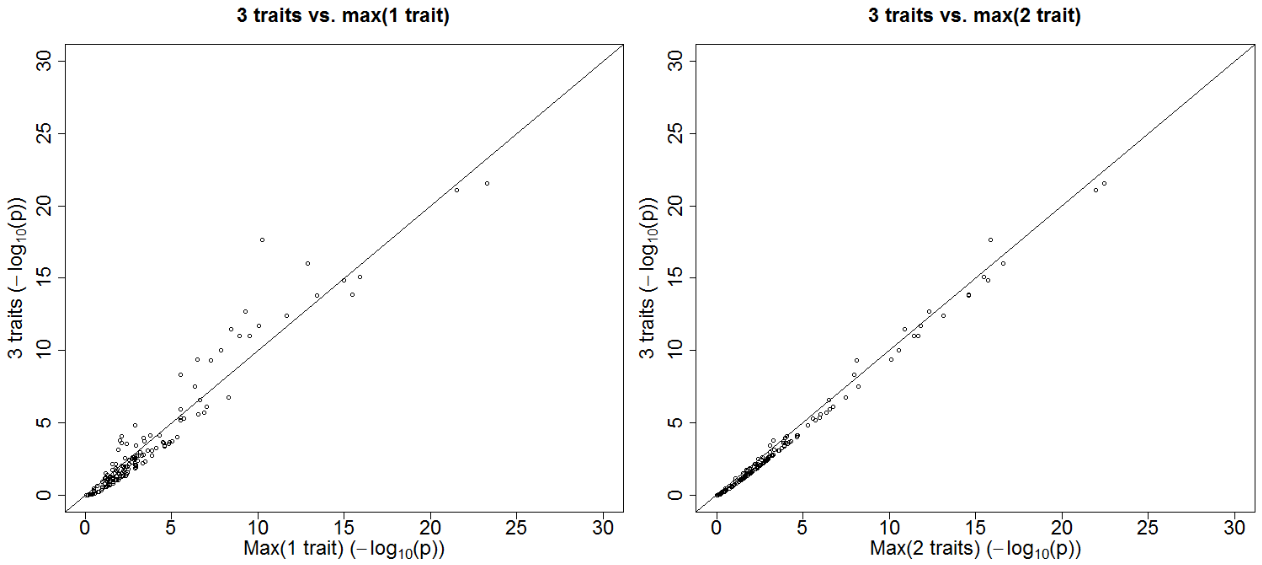


Figure 4.17: Multivariate test for 3 traits vs. maximum of univariate tests and maximum of multivariate tests for 2 traits on -log$_{10}$ scale.

This same thing can be seen from Fig. 4.17. In that figure I have plotted all the -log$_{10}$(p-values) from the multivariate tests for all three traits against the maximum of the corresponding univariate tests, and also for the maximum of the corresponding multivariate tests against two traits. From that figure it can be seen that the multivariate test for 3 traits is not noticeably less efficient than the corresponding univariate test, or multivariate test against two traits that yields the most significant results. And in some cases it can be more efficient than the corresponding univariate test that yields the most significant result. So based on these results we can say that there is no reason not to do the multivariate test first.

I also checked the relationship between multivariate Wald test and CCA. In figure 4.18 are plotted all multivariate Wald tests against two traits and against 3 traits against corresponding CCA results. From that figure it can be seen the same thing as earlier; that those two methods yields the same results. This time they are exactly the same all the way through, which is propably due to the fact that in this data there is approximatley 19 times more individuals (18,834 vs 1,000) and thus the differences in approximations seen in previous section disappears.

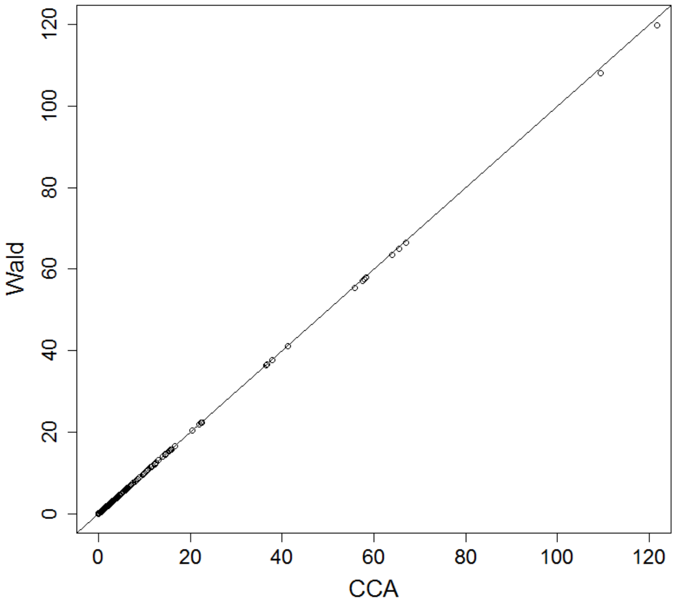


Figure 4.18: CCA and Multivariate Wald test -$\log_{10}$(p-values) comparison from The National FINRISK Study.

Chapter 5

Conclusions

There were two main goals in this thesis: to demonstrate the nature of multivariate test compared to the corresponding univariate tests, and to see if two multivariate methods, multivariate Wald test and CCA yield the same results. Based on my simulation studies and examples from The National FINRISK study we can say that multivariate test is never much worse in terms of power to detect associations than the corresponding univariate tests but in some situations it can be much more efficient. In genetic association studies multivariate analysis of genomic regions and correlated traits can provide new insights to genetic mechanism behind complex, non-Mendelian diseases such as cardiovascular diseases. For example, for correlated lipid traits using multivariate tests we can detect more genomic regions associated with blood lipids, than we would by using only univariate tests. Finding these associations and the genetic mechanism behind blood lipid levels is important, as they are well-established risk factors for cardiovascular diseases, the most common cause of death worldwide.

The multivariate test is (in case of two traits) especially efficient when the correlation between the traits is large and its direction is opposite to the direction of the genetic effects on them. This happens when the correlation between the traits is positive and the effects have opposite directions, or the correlation is negative and the effects have same direction. Based on Figures 4.16 and 4.17 it seems that in general adding traits to analysis does not decrease the power to detect the significant effects. For example, the multivariate tests for three traits in practice always yields at least as small p-value as the one for only 2 traits. Thus, when testing for example the association of a SNP and multiple correlated traits there is no reason not to do the multivariate test first and after that, if one is interested, carry out the corresponding univariate tests.

The other main goal in this thesis was to look the relationship of Mutivariate Wald test and CCA and see if they yield the same results. Based on the simulation studies and examples from The National FINRISK Study we can say that these two methods yield the same results, with some small differences probably due to slightly different approximations.

The differences with smaller sample sizes have very little practical consequence in locus discovery, because when finding statistically significant SNPs for some trait, the most commonly used threshold in GWA studies for significance is 5×10^{-8}, and until that value these two methods yields the same results. Thus for all practical purposes, we can say that in case of testing the statistical significance of the the genetic effects of one genotype on multiple traits, it does not make a difference if one uses multivariate Wald test or CCA. There is, however, a limitation in multivariate Wald test compared to CCA: Wald test cannot be used to test the significance of multiple SNPs and traits, which can be done using CCA.

In this thesis I examined only cases of two or three traits and one genomic loci at a time. I think that it would be really interesting to look how these patterns shown in this thesis could be generalized into cases of tens of traits simultaneously. Because many biological processes consist of much more than two or three separate measurements, examining even more traits would provide even more useful tools to study complex diseases. Based on future results it could be possible to create an algorithm to estimate beforehand if one should use multivariate test instead of multiple univariate tests based on the effect sizes and correlation structure of the traits.

Acknowledgements

First of all I would like to thank my supervisors Matti Pirinen and Samuli Ripatti for providing me with the opportunity to write my thesis at FIMM, and giving me the guidance and support needed. I would also like to thank Sirkka-Liisa Varvio for introducing me to Matti and Samuli, and thus providing me this opportunity. I would also like to thank all my colleagues at FIMM, especially my roommates for providing me such a warm and inspiring work environment. I would also want to thank anonymous participants of The National FINRISK Study and the researchers of The FINRISK Study for a permission to use the cohort in this thesis work. Last, but definitely not least, I would like to thank my family and friends for always being there for me when needed as a listening ears and shoulders to lean on.

Helsinki, 23.5.2017

Sanni Ruotsalainen

Bibliography

[Cichonska et al., 2016] Cichonska, A., Rousu, J., Marttinen, P., Kangas, A. J., Soininen, P., Lehtimäki, T., Raitakari, O. T., Järvelin, M.-R., Salomaa, V., Ala-Korpela, M., Ripatti, S., and Pirinen, M. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics*, 32(13):1981.

[Dillon and Goldstein, 1983] Dillon, W. R. and Goldstein, M. (1983). *Multivariate analysis: methods and applications*. Wiley Series in probapility and mathematical statistics: Applied probability and Statistics. Wiley.

[FINRISK homepage] FINRISK homepage. National Institute for Health and Welfare, THL. `https://www.thl.fi/en/web/thlfi-en/research-and-expertwork/population-studies/the-national-finrisk-study`. Accessed: 2016-009-27.

[Inouye et al., 2012] Inouye, M., Ripatti, S., Kettunen, J., Lyytikäinen, L.-P., Oksala, N., Laurila, P.-P., Kangas, A. J., Soininen, P., Savolainen, M. J., Viikari, J., Kähönen, M., Perola, M., Salomaa, V., Raitakari, O., Lehtimäki, T., Taskinen, M.-R., Järvelin, M.-R., Ala-Korpela, M., Palotie, A., and de Bakker, P. I. W. (2012). Novel Loci for Metabolic Networks and Multi-Tissue Expression Studies Reveal Genes for Atherosclerosis. *PLoS Genetics*, 8(8):e1002907.

[Kannel et al., 1961] Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N., and Stokes, J. (1961). Factors of Risk in the Development of Coronary Heart Disease—Six-Year Follow-up Experience, The Framingham Study. *Annals of Internal Medicine*, 55(1):33–50.

[Klug et al., 2012] Klug, W., Cummings, M., Palladino, M., and Spencer, C. (2012). *Concepts of Genetics (Tenth Edition)*. Pearson Education.

[McCarthy et al., 2008] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.

[Pearson and Manolio, 2008] Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA*, 299(11):1335–1344.

[Price et al., 2015] Price, A. L., Spencer, C. C. A., and Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society of London B: Biological Sciences*, 282:20151684.

[Rencher and Christensen, 2011] Rencher, A. C. and Christensen, W. (2011). *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. Wiley.

[Sham and Purcel, 2014] Sham, P. C. and Purcel, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346.

[Stephens, 2013] Stephens, M. (2013). A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE*, 8(7):e65245.

[Surakka, 2014] Surakka, I. (2014). *Genetics of Circulating Blood Lipids*. PhD thesis, University of Helsinki.

[van Dongen et al., 2013] van Dongen, J., Willemsen, G., Chen, W.-M., de Geus, E., and Boomsma, D. (2013). The heritability of metabolic syndrome traits in a large population-based sample. *Journal of Lipid Research*, 54(10):2914–2923.

[Webb et al., 2013] Webb, J., Gonna, H., and Ray, K. K. (2013). Lipid management: maximising reduction of cardiac risk. *Clinical Medicine*, 13(6):618–620.

[Willer et al., 2013] Willer, C. J., Mohlke, K. L., Ingelsson, E., and Abecasis, G. R. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1274.

[Yang et al., 2012] Yang, X., Sun, J., Gao, Y., Tan, A., Zhang, H., Hu, Y., Feng, J., Qin, X., Tao, S., Chen, Z., Kim, S.-T., Peng, T., Liao, M., Lin, X., Zhang, Z., Tang, M., Li, L., Mo, L., Liang, Z., Shi, D., Huang, Z., Huang, X., Liu, M., Liu, Q., Zhang, S., Trent, J. M., Zheng, S. L., Xu, J., and Mo, Z. (2012). Genome-Wide Association Study for Serum Complement C3 and C4 Levels in Healthy Chinese Subjects. *PLoS Genetics*, 8(9):e1002916.

Appendix A

R script for simulations

```
n<-1000 #number of "individuals"
g<-1000 #number of genotypes
m<-6 #number of bivariate phenotype-pairs

#simulation of genotypes
G <- lapply(1:n, function(x) (rbinom(n, 2, 0.2)))
GT<-lapply(1:g, function(x) rep(G[[x]],each=2))

#simulation of phenotypes
beta1<-c(-0.2,0,0.2,-0.2,0,0.2) #effects on the first phenotype
beta2<-c(rep(0.2, times=m)) #effects on the second phenotype
Beta<-lapply(1:m, function(x) rep(c(beta1[x],beta2[x]),times=1000))

#covariances between the phenotypes
Cov<-c(rep(0.3,times=3),rep(0.7,times=3))

#covariance matrices
Sigma<-lapply(1:m, function(x) matrix(c(1,Cov[x],Cov[x],1),ncol=2))
#XB-matrix
XB<-lapply(1:m, function(x) sapply(1:n, function(y)  GT[[y]]*Beta[[x]]))
library(MASS)
#Epsilon-matrix
epsilon<-lapply(1:m, function(x) mvrnorm(n=1000, c(0,0), Sigma[[x]]))
#Phenotypes
Y1<-lapply(1:m, function(x) sapply(1:n, function(y)  XB[[x]][,y][c(TRUE,FALSE)]
+epsilon[[x]][,1]))
Y2<-lapply(1:m, function(x) sapply(1:n, function(y)  XB[[x]][,y][c(FALSE,TRUE)]
+epsilon[[x]][,2]))
FT<-c(Y1,Y2)

#X and Y matrices, centered for means
X<-lapply(1:n, function(x) matrix(c(G[[x]]-mean(G[[x]])),ncol=1))
Y<-lapply(1:(m*2), function(x) (sapply(1:n, function(y) FT[[x]][,y]-
mean(FT[[x]][,y]))))

#Beta-estimates
betaHat<-lapply(1:(m*2), function(x)(sapply(1:n, function(y) solve
(t(X[[y]])%*%X[[y]]) %*% t(X[[y]]) %*%Y[[x]][,y])))

#Variances of beta-estimates
var_betaHat<-lapply(1:12,function(x) (sapply(1:n, function(y) var(FT[[x]][,y])*
solve(t(X[[y]]) %*% X[[y]]))))

#univariate test statistics
Z<-lapply(1:(m*2), function(x) (sapply(1:n, function(y) betaHat[[x]][y]
/sqrt(var_betaHat[[x]][y]))))

#univariate p-values
pvaluesuv<-lapply(1:(m*2), function(x) (sapply(1:n, function(y)
2*pnorm(-abs(Z[[x]][[y]])))))

#univariate -log10(p)- values
logpvaluesuv<-lapply(1:(m*2), function(x) (sapply(1:n, function(y) (-log10
(pvaluesuv[[x]][y])))))

#Multivariate Wald test

#covariances of the estimates
covariance<-lapply(1:m, function(x)(sapply(1:n, function(y) (cov(FT[[x]][,y],
FT[[x+m]][,y]))*solve(t(X[[y]]) %*% X[[y]]))))

#multivariate test statistics
ZM<-lapply(1:m, function(x)(sapply(1:n, function(y) (c(betaHat[[x]][y],
betaHat[[x+6]][y]))%*%(solve(matrix(c(var_betaHat[[x]][y],covariance[[x]][y],
covariance[[x]][y],
var_betaHat[[x+6]][y]),ncol=2)))%*%(matrix(c(betaHat[[x]][y],
betaHat[[x+6]][y]), ncol=1)))))

#multivariate p-values
pvaluesmv<-lapply(1:m, function(x) (sapply(1:n, function(y) pchisq(ZM[[x]][y],2,
lower.tail=FALSE))))

#multivariate-log10(p)-values
logpvaluesmv<-lapply(1:m, function(x) sapply(1:n, function(y) (-log10(pvaluesmv[[x]]
[y]))))

#CCA
sigma_xx<-lapply(1:n, function(x) matrix(c(var(G[[x]])),1,1))

sigma_yy<-lapply(1:m, function(x)(lapply(1:n, function(y) matrix(c(var(FT[[x]][,y]),
cov(FT[[x]][,y],FT[[x+6]][,y]),cov(FT[[x]][,y],FT[[x+6]][,y]),var(FT[[x+6]][,y])),
2,2))))

sigma_all<-lapply(1:m, function(x) (lapply(1:n, function(y) matrix(c(var(FT[[x]]
[,y]),cov(FT[[x]][,y],FT[[x+6]][,y]),cov(FT[[x]][,y],G[[y]]), cov(FT[[x]][,y],
FT[[x+6]][,y]),var(FT[[x+6]][,y]),cov(FT[[x+6]][,y],G[[y]]),cov(FT[[x]][,y],G[[y]]),
cov(FT[[x+6]][,y],G[[y]]),var(G[[y]])),3,3))))

Lambda<-lapply(1:m, function(x) (sapply(1:n,function(y) det(sigma_all[[x]][[y]])/
((det(sigma_xx[[y]]))%*%(det(sigma_yy[[x]][[y]]))))))

Chi_appro<-lapply(1:m, function(x) (sapply(1:n, function(y) -((n-1)-0.5*(2+1+1))*
log(Lambda[[x]][y]))))

#CCA p-values
pvaluescca<-lapply(1:m, function(x) (sapply(1:n, function(y) pchisq(Chi_appro[[x]][y],
2, lower.tail=F))))

#CCA -log(p)-values
logpvaluescca<-lapply(1:m, function(x) (sapply(1:n, function(y) (-log10(pvaluescca
[[x]][y])))))
```

Appendix B

List of SNPs used in analysis of The National FINRISK Study

Table B.1: List of SNPs used in analysis [Willer et al., 2013]

Gene	Rsid	Chr	Reported associated traits	A1	A2	Position	A1 freq
ASAP3	rs1077514	1	TC	t	c	23766233	0.8707
LDLRAP1	rs12027135	1	TC,LDL	t	a	25775733	0.5343
PIGV-NR0B2	rs12748152	1	HDL,LDL,TG	c	t	27138393	0.92876
PABPC4	rs4660293	1	HDL	a	g	40028180	0.7639
PCSK9	rs2479409	1	LDL,TC	a	g	55504650	0.6675
ANGPTL3	rs2131925	1	TG,LDL,TC	t	g	63025942	0.69
EVI5	rs7515577	1	TC	c	a	93009438	0.1939
SORT1	rs629301	1	LDL,TC	g	t	109818306	0.2124
ANXA9-CERS2	rs267733	1	LDL	g	a	150958836	0.1372
HDGF-PMVK	rs12145743	1	HDL	g	t	156700651	0.3311
ANGPTL1	rs4650994	1	HDL	g	a	178515312	0.5172
ZNF648	rs1689800	1	HDL	a	g	182168885	0.6728
MOSC1	rs2642442	1	TC,LDL	t	c	220973563	0.7282
GALNT2	rs4846914	1	HDL,TG	a	g	230295691	0.5844
IRF2BP2	rs514230	1	TC,LDL	t	a	234858597	0.7
APOB	rs1367117	2	LDL,TC	g	a	21263900	0.7124
GCKR	rs1260326	2	TG,TC	c	t	27730940	0.5871
ABCG5/8	rs4299376	2	LDL,TC	t	g	44072576	0.7032
EHBP1	rs2710642	2	LDL	g	a	63149557	0.3813
INSIG2	rs10490626	2	LDL,TC	a	g	118835841	0.07916
LOC84931	rs2030746	2	LDL,TC	c	t	121309488	0.6016
RAB3GAP1	rs7570971	2	TC	a	c	135837906	0.4908
COBLL1	rs12328675	2	HDL	c	t	165540800	0.1491
ABCB11	rs2287623	2	TC	g	a	169830155	0.405
FAM117B	rs11694172	2	TC	a	g	203532304	0.7836
CPS1	rs1047891	2	HDL	c	a	211540507	0.6979
FN1	rs1250229	2	LDL	t	c	216304384	0.2111
IRS1	rs2972146	2	HDL,TG	g	t	227100698	0.3773
UGT1A1	rs11563251	2	TC,LDL	t	c	234679384	0.1253
ATG7	rs2606736	3	HDL	c	t	11400249	0.3945
RAF1	rs2290159	3	TC	g	c	12628920	0.82
CMTM6	rs7640978	3	LDL,TC	t	c	32533010	0.1055

Table B.1: List of SNPs used in analysis [Willer et al., 2013]

Gene	Rsid	Chr	Reported associated traits	A1	A2	Position	A1 freq
SETD2	rs2290547	3	HDL	g	a	47061183	0.7889
RBM5	rs2013208	3	HDL	t	c	50129399	0.5053
STAB1	rs13326165	3	HDL	a	g	52532118	0.1873
PXK	rs13315871	3	TC	g	a	58381287	0.91953
GSK3B	rs6805251	3	HDL	t	c	119560606	0.3813
ACAD11	rs17404153	3	LDL,HDL	t	g	132163200	0.1438
MSL2L1	rs645040	3	TG	g	t	135926622	0.2309
LRPAP1	rs6831256	4	TG,TC,LDL	a	g	3473139	0.591
C4orf52	rs10019888	4	HDL	a	g	26062990	0.8364
KLHL8	rs442177	4	TG	g	t	88030261	0.4472
FAM13A	rs3822072	4	HDL	g	a	89741269	0.5119
ADH5	rs2602836	4	HDL	a	g	100014805	0.4274
SLC39A8	rs13107325	4	HDL	c	t	103188709	0.92216
ARL15	rs6450176	5	HDL	g	a	53298025	0.7216
MAP3K1	rs9686661	5	TG	c	t	55861786	0.8232
HMGCR	rs12916	5	TC,LDL	c	t	74656539	0.4314
CSNK1G3	rs4530754	5	LDL,TC	a	g	122855416	0.5818
TIMD4	rs6882076	5	TC,TG,LDL	c	t	156390297	0.6662
MYLIP	rs3757354	6	LDL,TC	t	c	16127407	0.2098
HFE	rs1800562	6	LDL,TC	g	a	26093141	0.95383
HLA	rs3177928	6	TC,LDL	a	g	32412435	0.1807
C6orf106	rs2814982	6	TC	c	t	34546560	0.8931
KCNK17	rs2758886	6	TC	g	a	39250837	0.715
VEGFA	rs998584	6	TG,HDL	c	a	43757896	0.4855
FRK	rs9488822	6	TC,LDL	a	t	116312893	0.7
RSPO3	rs1936800	6	HDL, TG	c	t	127436064	0.5277
HBS1L	rs9376090	6	TC	t	c	135411228	0.7282
CITED2	rs605066	6	HDL	t	c	139829666	0.562
LPA	rs1564348	6	LDL,TC	t	c	160578860	0.8549
GPR146	rs1997243	7	TC	g	a	1083777	0.1306
DAGLB	rs702485	7	HDL	g	a	6449272	0.4499
SNX13	rs4142995	7	HDL	g	t	17919258	0.6161
DNAH11	rs12670798	7	TC,LDL	t	c	21607352	0.7757
MIR148A	rs4722551	7	LDL,TG,TC	c	t	25991826	0.1702
NPC1L1	rs2072183	7	TC,LDL	g	c	44579180	0.7625
IKZF1	rs4917014	7	HDL	g	t	50305863	0.3404
TYW1B	rs13238203	7	TG	t	c	72129667	0.03562
MLXIPL	rs17145738	7	TG,HDL	t	c	72982874	0.1174
MET	rs38855	7	TG	g	a	116358044	0.4736
KLF14	rs4731702	7	HDL	t	c	130433384	0.4604
TMEM176A	rs17173637	7	HDL	t	c	150529449	0.90237
PPP1R3B	rs9987289	8	HDL,TC,LDL	g	a	9183358	0.9248
PINX1	rs11776767	8	TG	g	c	10683929	0.5937
NAT2	rs1495741	8	TG,TC	a	g	18272881	0.7493
LPL	rs12678919	8	TG,HDL	g	a	19844222	0.1214
SOX17	rs10102164	8	LDL, TC	g	a	55421614	0.8259
CYP7A1	rs2081687	8	TC,LDL	t	c	59388565	0.3522
TRPS1	rs2293889	8	HDL	g	t	116599199	0.5871
TRIB1	rs2954029	8	TG,TC,LDL,HDL	t	a	126490972	0.4683
PLEC1	rs11136341	8	LDL,TC	g	a	145043543	0.3694
VLDLR	rs3780181	9	TC,LDL	a	g	2640759	0.94723
TTC39B	rs581080	9	HDL,TC	c	g	15305378	0.8206
ABCA1	rs1883025	9	HDL,TC	c	t	107664301	0.7573
ABO	rs635634	9	-	t	c	136155000	0.1873
AKR1C4	rs1832007	10	TG	a	g	5254847	0.8681

Table B.1: List of SNPs used in analysis [Willer et al., 2013]

Gene	Rsid	Chr	Reported associated traits	A1	A2	Position	A1 freq
VIM-CUBN	rs10904908	10	TC	g	a	17260290	0.4538
MARCH8-ALOX5	rs970548	10	HDL, TC	c	a	46013277	0.277
JMJD1C	rs10761731	10	TG	t	a	65027610	0.37
CYP26A1	rs2068888	10	TG	a	g	94839642	0.4908
GPAM	rs2255141	10	TC,LDL	a	g	113933886	0.3193
AMPD3	rs2923084	11	HDL	a	g	10388782	0.847
SPTY2D1	rs10128711	11	TC	c	t	18632984	0.719
LRP4	rs3136441	11	HDL	c	t	46743247	0.1372
OR4C46	rs11246602	11	HDL	c	t	51512090	0.1332
FADS1-2-3	rs174546	11	TG,LDL,TC,HDL	c	t	61569830	0.6425
KAT5	rs12801636	11	HDL	a	g	65391317	0.2243
MOGAT2-DGAT2	rs499974	11	HDL	c	a	75455021	0.8245
APOA1	rs964184	11	TG,TC,HDL,LDL	c	g	116648917	0.78
PHLDB1	rs11603023	11	TC	t	c	118486067	0.4512
UBASH3B	rs112302432	11	-	c	t	122522375	0.3971
ST3GAL4	rs11220462	11	LDL,TC	g	a	126243952	0.8575
PHC1-A2ML1	rs4883201	12	TC	a	g	9082581	0.8865
PDE3A	rs7134375	12	HDL	a	c	20473758	0.4169
LRP1	rs11613352	12	TG,HDL	t	c	57792580	0.1913
MVK	rs7134594	12	HDL	t	c	110000193	0.5554
BRAP	rs11065987	12	TC,LDL	a	g	112072424	0.5778
HNF1A	rs1169288	12	TC,LDL	c	a	121416650	0.3338
SBNO1	rs4759375	12	HDL	t	c	123796238	0.09367
ZNF664	rs4765127	12	HDL,TG	t	g	124460167	0.3628
SCARB1	rs838880	12	HDL	c	t	125261593	0.3259
BRCA2	rs4942486	13	LDL	c	t	32953388	0.5383
NYNRIN	rs8017377	14	LDL	g	a	24883887	0.5409
ZBTB42-AKT1	rs4983559	14	HDL	g	a	105277209	0.3773
CAPN3	rs2412710	15	TG	g	a	42683787	0.97757
FRMD5	rs2929282	15	TG	a	t	44245931	0.85
LIPC	rs1532085	15	HDL,TC,TG	a	g	58683366	0.3668
LACTB	rs2652834	15	HDL	g	a	63396867	0.7652
PDXDC1	rs3198697	16	TG	t	c	15129940	0.3826
CTF1	rs11649653	16	TG	g	c	30918487	0.36
FTO	rs1121980	16	HDL,TG	g	a	53809247	0.5528
CETP	rs3764261	16	HDL,LDL,TC,TG	a	c	56993324	0.2942
LCAT	rs16942887	16	HDL	a	g	67928042	0.1332
HPR	rs2000999	16	TC,LDL	a	g	72108093	0.1847
CMIP	rs2925979	16	HDL	c	t	81534790	0.7045
DLG4	rs314253	17	TC,LDL	c	t	7091650	0.3351
STARD3	rs11869286	17	HDL	c	g	37813856	0.6755
MPP3	rs8077889	17	TG	a	c	41878166	0.7559
OSBPL7	rs7206971	17	LDL,TC	a	g	45425115	0.4723
APOH-PRXCA	rs1801689	17	LDL	a	c	64210580	0.96306
ABCA8	rs4148008	17	HDL	c	g	66875294	0.6913
PGS1	rs4129767	17	HDL	a	g	76403984	0.5237
LIPG	rs7241918	18	HDL,TC	t	g	47160953	0.8232
MC4R	rs12967135	18	HDL	g	a	57849023	0.7691
INSR	rs7248104	19	TG	a	g	7224431	0.4169
ANGPTL4	rs7255436	19	HDL	a	c	8433196	0.5435
LDLR	rs6511720	19	LDL,TC	t	g	11202306	0.09763
ANGPTL8	rs737337	19	HDL	t	c	11347493	0.9314
CILP2	rs10401969	19	TC,TG,LDL	c	t	19407718	0.07124
PEPD	rs731839	19	TG, HDL	a	g	33899065	0.6583
APOE	rs4420638	19	LDL,TC,HDL	a	g	45422946	0.814

Table B.1: List of SNPs used in analysis [Willer et al., 2013]

Gene	Rsid	Chr	Reported associated traits	A1	A2	Position	A1 freq
FLJ36070	rs492602	19	TC	a	g	49206417	0.5699
HAS1	rs17695224	19	HDL	g	a	52324216	0.7612
LILRA3	rs386000	19	HDL	c	g	54792761	0.1992
SPTLC3	rs364585	20	LDL	g	a	12962718	0.6332
SNX5	rs2328223	20	LDL	a	c	17845921	0.7507
ERGIC3	rs2277862	20	TC	c	t	34152782	0.8681
MAFB	rs2902940	20	TC,LDL	a	g	39091487	0.7586
TOP1	rs6029526	20	LDL,TC	t	a	39672618	0.36
HNF4A	rs1800961	20	HDL,TC	c	t	43042364	0.9657
PLTP	rs6065906	20	HDL,TG	t	c	44554015	0.8021
UBE2L3	rs113359481	22	-	c	t	21932068	0.8008
MTMR3	rs5763662	22	LDL	t	c	30378703	0.02507
TOM1	rs138777	22	TC	a	g	35711098	0.3483
PLA2G6	rs5756931	22	TG	c	t	38546033	0.3641
PPARA	rs4253772	22	TC,LDL	c	t	46627603	0.8813

Appendix C

Results from The National FINRISK Study

Table C.1: Results for univariate Wald tests

	β			s.e.(β)			Z			p-value		
Gene	HDL	LDL	TG	HDL	LDL	TG	HDL	LDL	TG	HDL	LDL	TG
ASAP3	-0.0512	-0.0253	0.0266	0.0156	0.0158	0.0151	-3.2812	-1.6065	1.762	0.001	0.1082	0.0781
LDLRAP1	-0.0135	-0.0533	-0.02	0.0102	0.0103	0.0099	-1.325	-5.1812	-2.0338	0.1852	2.2e-07	0.042
PIGV-NR0B2	-0.0503	0.0429	0.0351	0.019	0.0192	0.0184	-2.6444	2.2302	1.9054	0.0082	0.0257	0.0567
PABPC4	-0.0171	-0.0239	0.0184	0.0123	0.0124	0.0119	-1.3881	-1.9251	1.5505	0.1651	0.0542	0.121
PCSK9	-0.0173	0.058	0.0061	0.0112	0.0113	0.0108	-1.5485	5.1282	0.5626	0.1215	2.9e-07	0.5737
ANGPTL3	-0.0059	-0.0361	-0.0729	0.0116	0.0117	0.0112	-0.506	-3.0728	-6.4918	0.6129	0.0021	8.5e-11
EVI5	-0.0233	-0.0302	0.0064	0.0122	0.0124	0.0118	-1.9026	-2.4467	0.5408	0.0571	0.0144	0.5887
SORT1	0.0389	-0.1592	0.0071	0.0123	0.0125	0.0119	3.151	-12.7752	0.5975	0.0016	2.3e-37	0.5502
ANXA9-CERS2	0.0128	-0.0305	-0.0098	0.0146	0.0147	0.0141	0.8808	-2.071	-0.6968	0.3784	0.0384	0.4859
HDGF-PMVK	0.0055	3e-04	-0.009	0.0108	0.0109	0.0104	0.5098	0.0252	-0.8601	0.6102	0.9799	0.3897
ANGPTL1	-0.0447	0.0089	0.0138	0.0103	0.0104	0.0099	-4.354	0.861	1.3892	1.3e-05	0.3892	0.1648
ZNF648	-0.0372	0.0248	-0.0022	0.0115	0.0116	0.0111	-3.242	2.143	-0.1976	0.0012	0.0321	0.8433
MOSC1	-0.0211	-0.0299	0.0071	0.0112	0.0113	0.0108	-1.8879	-2.6424	0.6522	0.059	0.0082	0.5143
GALNT2	-0.0604	0.0269	0.0627	0.0103	0.0104	0.01	-5.8675	2.587	6.2994	4.4e-09	0.0097	3e-10
IRF2BP2	-0.0242	-0.0564	-0.0268	0.0103	0.0104	0.0099	-2.3549	-5.4294	-2.6934	0.0185	5.7e-08	0.0071
APOB	-0.0372	0.149	0.0329	0.0114	0.0115	0.011	-3.2701	12.9799	2.9903	0.0011	1.6e-38	0.0028
GCKR	-0.0226	0.0356	0.0858	0.0107	0.0108	0.0104	-2.1133	3.2918	8.2867	0.0346	0.001	1.2e-16
ABCG5/8	-0.0015	0.0734	0.0133	0.0124	0.0126	0.012	-0.1244	5.8464	1.103	0.901	5e-09	0.27
EHBP1	0.0089	-0.0308	0.0019	0.0104	0.0105	0.0101	0.8571	-2.9307	0.1914	0.3914	0.0034	0.8482
INSIG2	0.017	-0.0658	-0.0204	0.0203	0.0205	0.0196	0.84	-3.2197	-1.0407	0.4009	0.0013	0.298
LOC84931	-0.0318	0.0172	0.0044	0.0107	0.0108	0.0103	-2.98	1.5955	0.4253	0.0029	0.1106	0.6706
RAB3GAP1	0.0064	0.0274	0.0017	0.0103	0.0104	0.01	0.6261	2.6373	0.1719	0.5313	0.0084	0.8635
COBLL1	0.0576	0.0284	-0.0232	0.0172	0.0174	0.0166	3.3481	1.6379	-1.3927	8e-04	0.1014	0.1637
ABCB11	0.0228	0.0078	-0.0059	0.0102	0.0103	0.0099	2.2342	0.7519	-0.6015	0.0255	0.4521	0.5475
FAM117B	0.024	-0.0311	-0.0104	0.014	0.0142	0.0136	1.71	-2.1974	-0.7653	0.0873	0.028	0.4441
CPS1	-0.0191	0.016	-0.0039	0.0108	0.0109	0.0104	-1.7716	1.4721	-0.379	0.0765	0.141	0.7047
FN1	0.0186	-0.0255	-0.0334	0.0127	0.0129	0.0123	1.4576	-1.9807	-2.7163	0.1449	0.0476	0.0066
IRS1	0.0338	0.0016	-0.0305	0.0106	0.0107	0.0102	3.1986	0.1539	-2.9798	0.0014	0.8777	0.0029
UGT1A1	-0.0128	0.0382	0.0311	0.0168	0.017	0.0163	-0.7621	2.2506	1.9135	0.446	0.0244	0.0557
ATG7	0.0055	-0.0094	0.0098	0.0103	0.0104	0.01	0.536	-0.9058	0.9832	0.5919	0.365	0.3255
RAF1	-0.0117	-0.0178	-0.032	0.0118	0.012	0.0114	-0.9844	-1.4928	-2.7975	0.3249	0.1355	0.0052
CMTM6	0.0362	-0.0318	-0.0178	0.0182	0.0184	0.0176	1.9853	-1.7273	-1.0095	0.0471	0.0841	0.3127
SETD2	-0.023	0.0077	-0.0308	0.0121	0.0122	0.0117	-1.9044	0.631	-2.6411	0.0569	0.5281	0.0083
RBM5	-0.0247	0.0025	-0.003	0.0102	0.0103	0.0098	-2.4305	0.2403	-0.3078	0.0151	0.8101	0.7582
STAB1	0.0136	0.0041	-7e-04	0.0133	0.0134	0.0128	1.0212	0.3091	-0.0512	0.3072	0.7572	0.9591
PXK	0.0127	-0.052	-0.045	0.0172	0.0174	0.0167	0.7348	-2.9873	-2.6978	0.4624	0.0028	0.007
GSK3B	0.0286	0.0078	0.0053	0.0105	0.0106	0.0101	2.7313	0.737	0.526	0.0063	0.4611	0.5989
ACAD11	-0.0171	-0.007	0.0254	0.0134	0.0135	0.0129	-1.2781	-0.5152	1.9597	0.2012	0.6064	0.05
MSL2L1	0.0331	-0.0249	-0.0409	0.0146	0.0147	0.0141	2.2705	-1.693	-2.9034	0.0232	0.0905	0.0037
LRPAP1	-7e-04	0.0204	0.0329	0.0108	0.0109	0.0104	-0.0675	1.8807	3.1633	0.9462	0.06	0.0016
C4orf52	-0.0377	-0.0049	0.0166	0.015	0.0151	0.0145	-2.5104	-0.326	1.1472	0.0121	0.7444	0.2513
KLHL8	0.0129	-0.0253	-0.0362	0.0102	0.0103	0.0099	1.2654	-2.4597	-3.67	0.2057	0.0139	2e-04
FAM13A	0.0126	-0.0089	-0.0246	0.0102	0.0103	0.0099	1.2346	-0.867	-2.492	0.217	0.3859	0.0127
ADH5	0.0143	0.001	-0.0054	0.0102	0.0103	0.0099	1.4019	0.0992	-0.5469	0.1609	0.921	0.5845

Table C.1: Results for univariate Wald tests

	β			s.e.(β)			Z			p-value		
Gene	HDL	LDL	TG	HDL	LDL	TG	HDL	LDL	TG	HDL	LDL	TG
SLC39A8	-0.0747	0.0102	0.0194	0.0476	0.0481	0.0461	-1.5678	0.2113	0.4208	0.1169	0.8327	0.6739
ARL15	-0.0013	-0.0086	-0.0066	0.012	0.0121	0.0116	-0.1088	-0.713	-0.5707	0.9134	0.4758	0.5682
MAP3K1	-0.0397	0.0109	0.0236	0.0146	0.0148	0.0141	-2.7114	0.7372	1.6669	0.0067	0.461	0.0955
HMGCR	-0.0148	0.1046	0.031	0.0102	0.0103	0.0099	-1.4462	10.110	3.1301	0.1481	5e-24	0.0017
CSNK1G3	-0.0012	0.048	-0.0112	0.0102	0.0103	0.0098	-0.1135	4.6693	-1.1341	0.9097	3e-06	0.2568
TIMD4	0.0164	-0.0555	-0.0468	0.0109	0.011	0.0105	1.5056	-5.0528	-4.4508	0.1322	4.4e-07	8.6e-06
HFE	-0.007	-0.0656	0.0031	0.0267	0.027	0.0258	-0.2637	-2.4316	0.1215	0.792	0.015	0.9033
HLA	-0.0092	0.0409	0.0211	0.0125	0.0126	0.0121	-0.7359	3.2413	1.7448	0.4618	0.0012	0.081
C6orf106	-0.0638	-0.0281	0.0147	0.0137	0.0138	0.0132	-4.655	-2.0309	1.1118	3.2e-06	0.0423	0.2662
KCNK17	-2e-04	0.0041	-0.0067	0.012	0.0121	0.0116	-0.0177	0.3344	-0.5791	0.9859	0.7381	0.5625
VEGFA	-0.0256	0.0188	0.0413	0.0102	0.0103	0.0099	-2.5047	1.8168	4.1812	0.0123	0.0692	2.9e-05
FRK	0.001	-0.0221	-0.0181	0.0105	0.0107	0.0102	0.0922	-2.0768	-1.77	0.9265	0.0378	0.0767
RSPO3	-0.023	0.0046	0.0121	0.0102	0.0103	0.0099	-2.244	0.4406	1.2258	0.0248	0.6595	0.2203
HBS1L	-0.0103	-0.0225	0.0132	0.0108	0.0109	0.0104	-0.9605	-2.0696	1.2655	0.3368	0.0385	0.2057
CITED2	-0.0232	0.0187	0.0238	0.0102	0.0103	0.0099	-2.2759	1.8135	2.4167	0.0229	0.0698	0.0157
LPA	0.0044	0.0346	-0.0194	0.0147	0.0149	0.0143	0.2967	2.324	-1.3601	0.7667	0.0201	0.1738
GPR146	0.0309	0.0337	0.0388	0.0152	0.0154	0.0147	2.031	2.1926	2.6355	0.0423	0.0283	0.0084
DAGLB	0.0364	-0.0021	-0.0125	0.0102	0.0103	0.0099	3.5727	-0.208	-1.2655	4e-04	0.8352	0.2057
SNX13	-0.028	-0.0052	-0.0046	0.0112	0.0113	0.0108	-2.5017	-0.4614	-0.4278	0.0124	0.6445	0.6688
DNAH11	-0.0092	0.0404	0.0251	0.012	0.0122	0.0116	-0.7672	3.3208	2.1579	0.443	9e-04	0.0309
MIR148A	0.0056	0.0573	-0.0223	0.0122	0.0123	0.0118	0.459	4.659	-1.8968	0.6463	3.2e-06	0.0579
NPC1L1	-0.0095	0.0195	0.0028	0.0106	0.0107	0.0103	-0.8986	1.8195	0.2715	0.3689	0.0688	0.786
IKZF1	0.0166	-0.003	-0.0172	0.0112	0.0113	0.0108	1.4821	-0.2642	-1.5885	0.1383	0.7916	0.1122
TYW1B	0.0963	0.0234	-0.18	0.0447	0.0451	0.0432	2.1575	0.5182	-4.1671	0.031	0.6043	3.1e-05
MLXIPL	0.0512	0.0171	-0.1187	0.0153	0.0154	0.0148	3.3494	1.1099	-8.0244	8e-04	0.267	1e-15
MET	-0.0013	4e-04	-4e-04	0.0102	0.0103	0.0099	-0.1291	0.0369	-0.0371	0.8973	0.9705	0.9704
KLF14	0.0404	-8e-04	-0.0274	0.0102	0.0103	0.0099	3.9574	-0.0822	-2.77	7.6e-05	0.9345	0.0056
TMEM176A	-0.0367	-0.003	-0.0029	0.0144	0.0145	0.0139	-2.5538	-0.2093	-0.2085	0.0107	0.8342	0.8349
PPP1R3B	-0.0874	-0.0523	0.0153	0.0144	0.0145	0.0139	-6.0811	-3.6028	1.1011	1.2e-09	3e-04	0.2709
PINX1	0.0177	-0.0011	0.0047	0.011	0.0111	0.0107	1.6044	-0.0954	0.445	0.1086	0.924	0.6563
NAT2	0.0031	0.0498	0.0061	0.0117	0.0118	0.0113	0.2671	4.2077	0.5364	0.7894	2.6e-05	0.5917
LPL	0.1161	-0.0408	-0.1699	0.0181	0.0183	0.0175	6.4171	-2.2307	-9.706	1.4e-10	0.0257	2.8e-22
SOX17	-0.0108	0.0541	0.0141	0.0121	0.0122	0.0117	-0.8955	4.4223	1.2004	0.3705	9.8e-06	0.23
CYP7A1	-0.0075	0.0399	0.0344	0.0105	0.0106	0.0102	-0.7179	3.7571	3.3896	0.4728	2e-04	7e-04
TRPS1	-0.0323	-0.0088	-0.0193	0.0112	0.0113	0.0108	-2.8786	-0.7753	-1.7749	0.004	0.4382	0.0759
TRIB1	0.0296	-0.0476	-0.0745	0.0102	0.0103	0.0098	2.9067	-4.632	-7.5744	0.0037	3.6e-06	3.6e-14
PLEC1	-0.0058	0.0274	0.0148	0.0106	0.0107	0.0102	-0.5463	2.5643	1.4419	0.5848	0.0103	0.1493
VLDLR	-0.0055	-0.0546	-0.0243	0.0228	0.023	0.022	-0.2431	-2.3754	-1.1043	0.8079	0.0175	0.2695
TTC39B	-0.0466	0.0061	-0.029	0.0145	0.0147	0.014	-3.2126	0.4141	-2.0659	0.0013	0.6788	0.0388
ABCA1	-0.0809	-0.0347	-0.0165	0.0131	0.0132	0.0126	-6.2005	-2.6343	-1.3047	5.6e-10	0.0084	0.192
ABO	0.0353	0.09	-0.005	0.0127	0.0128	0.0123	2.7782	7.018	-0.4093	0.0055	2.3e-12	0.6823
AKR1C4	-0.0209	-0.0323	-0.0086	0.0151	0.0153	0.0146	-1.3831	-2.1086	-0.5857	0.1666	0.035	0.5581
VIM-CUBN	0.019	0.0184	0.0078	0.0104	0.0105	0.0101	1.8246	1.7513	0.7768	0.0681	0.0799	0.4373
MARCH8-ALOX5	0.0282	-0.0098	0.0191	0.0114	0.0115	0.011	2.4827	-0.8504	1.7347	0.013	0.3951	0.0828
JMJD1C	0.0306	0.0175	-0.0358	0.0105	0.0106	0.0102	2.9125	1.6524	-3.5227	0.0036	0.0985	4e-04
CYP26A1	0.0268	-0.0145	-0.0325	0.0102	0.0103	0.0099	2.6225	-1.4043	-3.2921	0.0087	0.1602	0.001
GPAM	0.0554	0.0448	-0.0163	0.0108	0.0109	0.0105	5.1135	4.0979	-1.5595	3.2e-07	4.2e-05	0.1189
AMPD3	-0.0128	-0.0053	0.0044	0.0145	0.0146	0.014	-0.8872	-0.3625	0.3175	0.375	0.717	0.7509
SPTY2D1	-0.0066	-0.0305	-0.0029	0.0104	0.0105	0.0101	-0.6295	-2.8978	-0.2885	0.529	0.0038	0.773
LRP4	0.0519	-2e-04	-0.0161	0.012	0.0121	0.0116	4.3329	-0.0162	-1.3914	1.5e-05	0.9871	0.1641
OR4C46	0.0417	-0.0197	-7e-04	0.0131	0.0132	0.0126	3.1869	-1.4958	-0.0555	0.0014	0.1347	0.9557
FADS1-2-3	-0.0511	-0.0686	0.0397	0.0104	0.0105	0.01	-4.9413	-6.5594	3.9665	7.8e-07	5.4e-11	7.3e-05
KAT5	0.0167	0.0092	-0.0278	0.0127	0.0128	0.0122	1.3209	0.7185	-2.2673	0.1865	0.4725	0.0234
MOGAT2-DGAT2	-0.0272	0.0171	-0.0078	0.0124	0.0125	0.012	-2.201	1.3679	-0.6496	0.0277	0.1713	0.5159
APOA1	-0.0999	0.0893	0.2427	0.0147	0.0148	0.0142	-6.8029	6.0265	17.0978	1e-11	1.7e-09	1.5e-65
PHLDB1	0.0163	0.0202	-7e-04	0.0103	0.0104	0.01	1.5855	1.9459	-0.0704	0.1129	0.0517	0.9438
UBASH3B	0.0207	0.004	-0.0016	0.0105	0.0106	0.0102	1.9642	0.375	-0.1565	0.0495	0.7077	0.8757
ST3GAL4	-0.0015	0.0458	0.0028	0.013	0.0131	0.0126	-0.1144	3.4938	0.225	0.9089	5e-04	0.822
PHC1-A2ML1	-0.0374	-0.0253	-0.0086	0.0159	0.016	0.0153	-2.3592	-1.5797	-0.559	0.0183	0.1142	0.5762
PDE3A	0.0035	0.0064	0.0023	0.0103	0.0104	0.01	0.3343	0.6122	0.2319	0.7381	0.5404	0.8166
LRP1	0.0235	0.0034	-0.0153	0.012	0.0122	0.0116	1.9466	0.2834	-1.3137	0.0516	0.7769	0.1889
MVK	0.0327	-0.005	-0.0023	0.0102	0.0103	0.0099	3.1944	-0.4859	-0.2314	0.0014	0.627	0.817
BRAP	-0.0233	-0.0224	0.0178	0.0105	0.0106	0.0101	-2.2217	-2.1221	1.7604	0.0263	0.0338	0.0783
HNF1A	0.0136	0.0504	0.0153	0.0105	0.0106	0.0101	1.2986	4.7648	1.5137	0.1941	1.9e-06	0.1301
SBNO1	0.0147	0.0489	0.0248	0.0157	0.0158	0.0152	0.9364	3.0922	1.6394	0.3491	0.002	0.1011
ZNF664	0.0368	-0.0054	-0.0197	0.0113	0.0114	0.011	3.2446	-0.4727	-1.8029	0.0012	0.6364	0.0714
SCARB1	0.0365	-0.0133	0.0058	0.0103	0.0104	0.0099	3.5577	-1.2845	0.5838	4e-04	0.199	0.5594
BRCA2	-0.0068	0.0017	0.0037	0.0103	0.0104	0.0099	-0.6598	0.1673	0.3715	0.5094	0.8671	0.7103
NYNRIN	-0.0032	0.0202	-4e-04	0.0105	0.0106	0.0102	-0.3003	1.8992	-0.044	0.7639	0.0575	0.9649
ZBTB42-AKT1	0.0476	-0.0014	-0.0138	0.0104	0.0105	0.0101	4.5714	-0.131	-1.366	4.8e-06	0.8958	0.1719
CAPN3	-0.0372	-0.1023	-0.065	0.1008	0.1018	0.0975	-0.3695	-1.0055	-0.6664	0.7117	0.3146	0.5051
FRMD5	0.0135	-0.0435	-0.0439	0.0236	0.0238	0.0228	0.5721	-1.8252	-1.9272	0.5673	0.068	0.054
LIPC	0.1406	0.0104	0.0439	0.0103	0.0104	0.01	13.6071	0.9974	4.394	3.6e-42	0.3186	1.1e-05
LACTB	-0.028	0.0055	0.0271	0.0124	0.0126	0.012	-2.254	0.4397	2.2576	0.0242	0.6601	0.024
PDXDC1	0.0092	0.0112	-0.0178	0.0104	0.0105	0.01	0.8889	1.0661	-1.7669	0.374	0.2864	0.0773

Table C.1: Results for univariate Wald tests

	β			s.e.(β)			Z			p-value		
Gene	HDL	LDL	TG	HDL	LDL	TG	HDL	LDL	TG	HDL	LDL	TG
CTF1	-0.0129	-0.013	-0.0055	0.0103	0.0104	0.01	-1.2533	-1.2509	-0.5486	0.2101	0.211	0.5833
FTO	-0.0115	0.0322	0.0235	0.0103	0.0104	0.01	-1.1149	3.0867	2.3528	0.2649	0.002	0.0186
CETP	0.2517	-0.0556	-0.0203	0.0114	0.0115	0.011	22.0815	-4.8336	-1.8388	4.8e-108	1.3e-06	0.0659
LCAT	0.1137	-6.6e-05	-0.0354	0.0139	0.0141	0.0135	8.161	-0.0047	-2.6269	3.3e-16	0.9963	0.0086
HPR	0.0152	0.0702	0.006	0.013	0.0131	0.0126	1.1714	5.3431	0.4748	0.2415	9.1e-08	0.635
CMIP	-0.0316	0.0079	0.01	0.011	0.0112	0.0107	-2.8631	0.7121	0.9381	0.0042	0.4764	0.3482
DLG4	0.0192	-0.0326	-0.026	0.0104	0.0105	0.0101	1.8484	-3.1066	-2.5874	0.0645	0.0019	0.0097
STARD3	-0.0356	-0.0173	0.0146	0.0111	0.0112	0.0107	-3.2235	-1.5515	1.3664	0.0013	0.1208	0.1718
MPP3	-0.0264	0.0265	0.0463	0.0126	0.0127	0.0121	-2.1032	2.0895	3.8132	0.0354	0.0367	1e-04
OSBPL7	0.0051	0.0299	0.0073	0.0102	0.0103	0.0099	0.5014	2.9007	0.7379	0.6161	0.0037	0.4606
APOH-PRXCA	-0.0516	0.0495	-0.1235	0.0542	0.0548	0.0525	-0.9504	0.9039	-2.3546	0.3419	0.3661	0.0185
ABCA8	-0.0125	0.0285	0.0037	0.011	0.0111	0.0106	-1.1366	2.5709	0.3447	0.2557	0.0101	0.7303
PGS1	-0.0141	-0.0361	-0.0132	0.0106	0.0107	0.0102	-1.3387	-3.3875	-1.2921	0.1807	7e-04	0.1963
LIPG	-0.0778	-0.0184	-0.0178	0.0137	0.0139	0.0133	-5.6698	-1.3301	-1.3443	1.4e-08	0.1835	0.1788
MC4R	-0.0013	-0.0046	0.0022	0.0133	0.0135	0.0129	-0.097	-0.3408	0.1735	0.9227	0.7332	0.8622
INSR	0.0136	4e-04	-0.0057	0.0104	0.0105	0.01	1.3136	0.0344	-0.5713	0.189	0.9726	0.5678
ANGPTL4	-0.0293	-0.0152	0.0045	0.0103	0.0104	0.0099	-2.8531	-1.4638	0.45	0.0043	0.1433	0.6527
LDLR	0.0534	-0.2676	-0.0448	0.0163	0.0165	0.0158	3.266	-16.2122	-2.8334	0.0011	4.1e-59	0.0046
ANGPTL8	-0.0834	-0.0461	-0.0271	0.0178	0.018	0.0172	-4.6768	-2.5625	-1.5723	2.9e-06	0.0104	0.1159
CILP2	0.0423	-0.132	-0.1544	0.0216	0.0218	0.0209	1.9622	-6.0621	-7.4042	0.0497	1.3e-09	1.3e-13
PEPD	-0.0345	-0.0069	0.0093	0.0106	0.0107	0.0103	-3.2428	-0.6379	0.9014	0.0012	0.5235	0.3674
APOE	-0.0525	0.1835	0.0498	0.0114	0.0115	0.011	-4.6061	15.9347	4.5173	4.1e-06	3.6e-57	6.3e-06
FLJ36070	0.0062	0.0011	0.0099	0.0105	0.0106	0.0102	0.5879	0.1052	0.9709	0.5566	0.9162	0.3316
HAS1	-0.0465	-0.004	-0.0011	0.0133	0.0134	0.0128	-3.5069	-0.2996	-0.0851	5e-04	0.7645	0.9322
LILRA3	0.0579	-0.0177	-0.0232	0.011	0.0111	0.0106	5.2699	-1.5972	-2.1834	1.4e-07	0.1102	0.029
SPTLC3	-0.0014	-0.0092	-4e-04	0.0107	0.0109	0.0104	-0.1285	-0.844	-0.0391	0.8977	0.3987	0.9688
SNX5	-0.0344	0.0187	0.0177	0.0124	0.0125	0.012	-2.7737	1.494	1.4761	0.0055	0.1352	0.1399
ERGIC3	0.0044	-0.0511	-0.0151	0.0178	0.018	0.0173	0.2489	-2.8344	-0.8741	0.8034	0.0046	0.3821
MAFB	-0.0028	-0.0258	-0.0095	0.0124	0.0125	0.012	-0.2284	-2.0627	-0.7945	0.8194	0.0391	0.4269
TOP1	3e-04	-0.0434	-0.0125	0.0102	0.0103	0.0099	0.0312	-4.2184	-1.273	0.9751	2.5e-05	0.203
HNF4A	-0.1494	-0.0874	-0.0117	0.0253	0.0255	0.0245	-5.9046	-3.4206	-0.4799	3.5e-09	6e-04	0.6313
PLTP	-0.0542	-4e-04	0.0289	0.0143	0.0144	0.0138	-3.7968	-0.0308	2.095	1e-04	0.9754	0.0362
UBE2L3	-0.044	0.0023	-0.0041	0.0109	0.011	0.0105	-4.0466	0.2069	-0.3935	5.2e-05	0.8361	0.694
MTMR3	0.032	0.0531	-0.0434	0.0291	0.0294	0.0281	1.0995	1.809	-1.5425	0.2715	0.0705	0.123
TOM1	0.0261	0.0276	0.0174	0.0106	0.0107	0.0102	2.4672	2.5833	1.6983	0.0136	0.0098	0.0894
PLA2G6	0.0278	-0.0159	-0.0075	0.0104	0.0105	0.01	2.6754	-1.5151	-0.7437	0.0075	0.1298	0.4571
PPARA	-0.0068	0.0669	0.0415	0.0191	0.0193	0.0184	-0.3554	3.4712	2.2495	0.7223	5e-04	0.0245
MYLIP	0.005	-0.0189	0.0127	0.0116	0.0117	0.0112	0.4316	-1.6112	1.1312	0.666	0.1071	0.258

Table C.2: Results for multivariate Wald test and CCA

	Z multi				p- value Wald				χ^2 approximation				p-value CCA			
Gene	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3
ASAP3	14.2584	10.9629	7.1461	14.9199	8e-04	0.0042	0.0281	0.0019	14.2623	10.9649	7.1467	14.9239	8e-04	0.0042	0.0281	6e-04
LDLRAP1	29.8557	9.805	27.8312	33.3267	3.3e-07	0.0074	9e-07	2.7e-07	29.8762	9.8065	27.8488	33.3518	3.3e-07	0.0074	9e-07	5.7e-08
PIGV-NR0B2	11.1127	7.7855	7.1637	11.3767	0.0039	0.0204	0.0278	0.0099	11.1148	7.7862	7.1643	11.3786	0.0039	0.0204	0.0278	0.0034
PABPC4	6.0877	3.0759	7.6549	8.2988	0.0477	0.2148	0.0218	0.0402	6.088	3.0758	7.6556	8.2995	0.0476	0.2148	0.0218	0.0158
PCSK9	27.6237	2.4053	26.5472	28.7491	1e-06	0.3004	1.7e-06	2.5e-06	27.6411	2.4052	26.5631	28.7673	9.9e-07	0.3004	1.7e-06	5.7e-07
ANGPTL3	10.0021	54.4569	45.3133	57.533	0.0067	1.5e-12	1.4e-10	2e-12	10.0037	54.53	45.3631	57.6134	0.0067	1.4e-12	1.4e-10	3.1e-13
EVI5	10.3974	3.6935	7.1203	10.4493	0.0055	0.1577	0.0284	0.0151	10.3992	3.6935	7.1208	10.4508	0.0055	0.1578	0.0284	0.0054
SORT1	167.8678	14.2958	173.9975	188.688	3.5e-37	8e-04	1.6e-38	1.2e-40	168.6025	14.2997	174.7877	189.6144	2.4e-37	8e-04	1.1e-38	6.7e-42
ANXA9-CERS2	4.8093	0.909	4.3667	4.8096	0.0903	0.6348	0.1127	0.1863	4.8094	0.9089	4.3667	4.8096	0.0903	0.6348	0.1127	0.0903
HDGF-PMVK	0.2642	0.7683	0.7822	0.8102	0.8763	0.681	0.6763	0.847	0.2642	0.7683	0.7821	0.8101	0.8763	0.681	0.6763	0.6669
ANGPTL1	19.23	19.1636	2.2767	19.5481	6.7e-05	6.9e-05	0.3203	2e-04	19.2377	19.1713	2.2766	19.5556	6.6e-05	6.9e-05	0.3204	5.7e-05
ZNF648	14.103	13.3727	5.0157	18.4731	9e-04	0.0012	0.0814	4e-04	14.1067	13.3761	5.0158	18.4797	9e-04	0.0012	0.0814	9.7e-05
MOSC1	11.3956	3.5845	8.4695	11.5533	0.0034	0.1666	0.0145	0.0091	11.3978	3.5845	8.4705	11.5553	0.0033	0.1666	0.0145	0.0031
GALNT2	38.9856	52.502	41.436	54.3278	3.4e-09	4e-12	1e-09	9.6e-12	39.0219	52.5697	41.4772	54.3991	3.4e-09	3.8e-12	9.8e-10	1.5e-12
IRF2BP2	37.2494	21.7881	32.0868	46.3365	8.2e-09	1.9e-05	1.1e-07	4.8e-10	37.2824	21.7984	32.1107	46.3875	8e-09	1.8e-05	1.1e-07	8.5e-11
APOB	173.5987	13.9253	168.5925	174.0099	2e-38	9e-04	2.5e-37	1.7e-37	174.3852	13.9289	169.3336	174.7955	1.4e-38	9e-04	1.7e-37	1.1e-38
GCKR	14.3021	70.7659	71.3205	73.3817	8e-04	4.3e-16	3.3e-16	8.1e-16	14.306	70.8916	71.4483	73.5152	8e-04	4e-16	3.1e-16	1.1e-16
ABCG5/8	34.292	1.3499	34.1902	34.2938	3.6e-08	0.5092	3.8e-08	1.7e-07	34.3196	1.3498	34.2176	34.3205	3.5e-08	0.5092	3.7e-08	3.5e-08
EHBP1	8.9859	1.0948	9.2437	10.3494	0.0112	0.5785	0.0098	0.0158	8.9871	1.0947	9.2449	10.3509	0.0112	0.5785	0.0098	0.0057
INSIG2	10.715	1.285	10.518	10.7414	0.0047	0.526	0.0052	0.0132	10.7169	1.2849	10.5199	10.743	0.0047	0.526	0.0052	0.0046
LOC84931	10.7486	9.6694	2.5559	12.1166	0.0046	0.0079	0.2786	0.007	10.7505	9.6709	2.5558	12.1189	0.0046	0.0079	0.2786	0.0023
RAB3GAP1	7.652	0.6163	7.097	7.6535	0.0218	0.7348	0.0288	0.0537	7.6527	0.6162	7.0976	7.6541	0.0218	0.7348	0.0288	0.0218
COBLL1	14.8401	11.2098	5.8008	14.9816	6e-04	0.0037	0.055	0.0018	14.8444	11.2119	5.8011	14.9856	6e-04	0.0037	0.055	6e-04
ABCB11	5.855	5.1178	1.1612	5.8881	0.0535	0.0774	0.5596	0.1172	5.8553	5.1179	1.1611	5.8883	0.0535	0.0774	0.5596	0.0526
FAM117B	7.2094	2.9282	4.9319	7.3229	0.0272	0.2313	0.0849	0.0623	7.21	2.9281	4.9321	7.3233	0.0272	0.2313	0.0849	0.0257
CPS1	4.9282	4.6327	2.6507	7.1907	0.0851	0.0986	0.2657	0.0661	4.9283	4.6328	2.6506	7.1911	0.0851	0.0986	0.2657	0.0274
FN1	5.631	7.5119	9.495	9.6369	0.0599	0.0234	0.0087	0.0219	5.6312	7.5126	9.4964	9.6381	0.0599	0.0234	0.0087	0.0081
IRS1	10.3954	13.5384	9.4892	14.1234	0.0055	0.0011	0.0087	0.0027	10.3971	13.5418	9.4906	14.1268	0.0055	0.0011	0.0087	9e-04
UGT1A1	5.411	3.6625	7.2668	7.2671	0.0668	0.1602	0.0264	0.0639	5.4112	3.6625	7.2674	7.2676	0.0668	0.1602	0.0264	0.0264
ATG7	1.0383	2.0404	2.2466	3.3381	0.595	0.3605	0.3252	0.3424	1.0383	2.0403	2.2465	3.3379	0.595	0.3605	0.3252	0.1884
RAF1	3.4483	13.3675	8.7085	14.2171	0.1783	0.0013	0.0129	0.0026	3.4482	13.3708	8.7096	14.2206	0.1783	0.0012	0.0128	8e-04
CMTM6	6.4285	3.984	3.4322	6.4376	0.0402	0.1364	0.1798	0.0922	6.4289	3.984	3.4322	6.4379	0.0402	0.1364	0.1798	0.04
SETD2	3.8608	17.8239	8.4098	19.3183	0.1451	1e-04	0.0149	2e-04	3.8608	17.8304	8.4108	19.3256	0.1451	1e-04	0.0149	6.4e-05
RBM5	5.9098	7.9919	0.1909	8.1015	0.0521	0.0184	0.909	0.044	5.9101	7.9927	0.1908	8.1022	0.0521	0.0184	0.909	0.0174
STAB1	1.1951	1.2095	0.1093	1.3108	0.5502	0.5462	0.9468	0.7266	1.195	1.2095	0.1093	1.3107	0.5502	0.5462	0.9468	0.5193
PXK	9.1771	7.4544	13.4646	13.6256	0.0102	0.0241	0.0012	0.0035	9.1784	7.4551	13.468	13.6287	0.0102	0.0241	0.0012	0.0011
GSK3B	8.3687	10.7735	0.69	11.156	0.0152	0.0046	0.7082	0.0109	8.3697	10.7754	0.6899	11.1578	0.0152	0.0046	0.7082	0.0038
ACAD11	2.0141	4.1034	4.7179	4.9735	0.3653	0.1285	0.0945	0.1737	2.014	4.1034	4.718	4.9735	0.3653	0.1285	0.0945	0.0832
MSL2L1	7.4664	9.8074	9.6885	11.0859	0.0239	0.0074	0.0079	0.0113	7.4671	9.8089	9.69	11.0877	0.0239	0.0074	0.0079	0.0039
LRPAP1	3.5433	11.8694	11.5921	13.4292	0.1701	0.0026	0.003	0.0038	3.5433	11.8719	11.5945	13.4322	0.1701	0.0026	0.003	0.0012
C4orf52	6.5763	6.316	1.6447	6.6256	0.0373	0.0425	0.4394	0.0848	6.5768	6.3164	1.6446	6.6259	0.0373	0.0425	0.4394	0.0364
KLHL8	7.2089	13.5467	16.5125	16.5833	0.0272	0.0011	3e-04	9e-04	7.2095	13.5501	16.518	16.5884	0.0272	0.0011	3e-04	2e-04
FAM13A	2.1216	6.2599	6.3428	6.3936	0.3462	0.0437	0.0419	0.094	2.1214	6.2602	6.3432	6.3939	0.3462	0.0437	0.0419	0.0409
ADH5	2.0092	1.9668	0.3456	2.0092	0.3662	0.374	0.8413	0.5705	2.0091	1.9667	0.3456	2.0091	0.3662	0.3741	0.8413	0.3662
SLC39A8	2.4659	2.5209	0.1934	2.5404	0.2914	0.2835	0.9078	0.468	2.4658	2.5208	0.1934	2.5402	0.2914	0.2835	0.9078	0.2808
ARL15	0.5356	0.4695	0.6966	0.8369	0.7651	0.7908	0.7059	0.8406	0.5356	0.4694	0.6965	0.8368	0.7651	0.7908	0.7059	0.6581
MAP3K1	7.6295	7.7094	2.942	7.8865	0.022	0.0212	0.2297	0.0484	7.6303	7.7101	2.9419	7.8871	0.022	0.0212	0.2297	0.0194
HMGCR	102.6473	9.8249	103.3833	103.4398	5.1e-23	0.0074	3.6e-23	2.8e-22	102.9171	9.8264	103.6571	103.7111	4.5e-23	0.0073	3.1e-23	3e-23
CSNK1G3	21.8663	1.6964	26.3664	26.8264	1.8e-05	0.4282	1.9e-06	6.4e-06	21.8767	1.6963	26.382	26.842	1.8e-05	0.4282	1.9e-06	1.5e-06
TIMD4	26.7721	19.9471	37.7069	37.8216	1.5e-06	4.7e-05	6.5e-09	3.1e-08	26.7883	19.9556	37.7407	37.8546	1.5e-06	4.6e-05	6.4e-09	6e-09

Table C.2: Results for multivariate Wald test and CCA

	Z multi				p- value Wald				χ^2 approximation				p-value CCA			
Gene	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3
MYLIP	2.6901	2.2573	4.8253	5.8313	0.2605	0.3235	0.0896	0.1201	2.69	2.2572	4.8254	5.8314	0.2605	0.3235	0.0896	0.0542
HFE	6.12	0.0697	6.3138	6.3603	0.0469	0.9657	0.0426	0.0953	6.1204	0.0697	6.3142	6.3605	0.0469	0.9657	0.0425	0.0416
HLA	10.7402	3.0444	11.7246	11.726	0.0047	0.2182	0.0028	0.0084	10.7421	3.0443	11.7271	11.7281	0.0046	0.2182	0.0028	0.0028
C6orf106	27.439	22.4725	6.5615	27.6392	1.1e-06	1.3e-05	0.0376	4.3e-06	27.4561	22.4836	6.562	27.6559	1.1e-06	1.3e-05	0.0376	9.9e-07
KCNK17	0.1119	0.4154	0.5496	0.6316	0.9456	0.8125	0.7597	0.8892	0.1119	0.4153	0.5496	0.6315	0.9456	0.8125	0.7597	0.7292
VEGFA	8.9179	18.2047	18.4447	19.1793	0.0116	1e-04	9.9e-05	3e-04	8.9191	18.2115	18.4518	19.1866	0.0116	1e-04	9.8e-05	6.8e-05
FRK	4.3182	3.6285	6.2	6.6771	0.1154	0.163	0.045	0.0829	4.3182	3.6285	6.2004	6.6774	0.1154	0.163	0.045	0.0355
RSPO3	5.1065	5.1419	1.5401	5.1851	0.0778	0.0765	0.463	0.1587	5.1066	5.1421	1.54	5.1851	0.0778	0.0765	0.463	0.0748
HBS1L	5.5504	1.8314	7.2634	7.4764	0.0623	0.4002	0.0265	0.0582	5.5506	1.8313	7.264	7.4769	0.0623	0.4002	0.0265	0.0238
CITED2	7.8716	7.8031	7.6543	9.6457	0.0195	0.0202	0.0218	0.0218	7.8724	7.8039	7.655	9.6469	0.0195	0.0202	0.0218	0.008
LPA	5.631	1.9356	8.9207	9.0186	0.0599	0.3799	0.0116	0.029	5.6312	1.9355	8.9219	9.0195	0.0599	0.3799	0.0116	0.011
GPR146	9.6878	18.7117	9.7963	21.4759	0.0079	8.6e-05	0.0075	8.4e-05	9.6893	18.719	9.7978	21.4853	0.0079	8.6e-05	0.0075	2.2e-05
DAGLB	12.7692	12.8193	1.6042	12.82	0.0017	0.0016	0.4484	0.005	12.7721	12.8223	1.6041	12.8227	0.0017	0.0016	0.4484	0.0016
SNX13	6.6928	8.8443	0.3288	8.9736	0.0352	0.012	0.8484	0.0296	6.6933	8.8454	0.3288	8.9746	0.0352	0.012	0.8484	0.0113
DNAH11	11.287	4.6759	13.3061	13.3197	0.0035	0.0965	0.0013	0.004	11.2892	4.676	13.3094	13.3226	0.0035	0.0965	0.0013	0.0013
MIR148A	22.3874	3.7265	30.1936	30.3515	1.4e-05	0.1552	2.8e-07	1.2e-06	22.3983	3.7265	30.2146	30.372	1.4e-05	0.1552	2.7e-07	2.5e-07
NPC1L1	3.8861	0.8197	3.3212	4.0908	0.1433	0.6638	0.19	0.2518	3.8861	0.8196	3.3211	4.0907	0.1433	0.6638	0.19	0.1293
IKZF1	2.2186	3.3434	2.5272	3.3465	0.3298	0.1879	0.2826	0.3412	2.2185	3.3433	2.5271	3.3464	0.3298	0.1879	0.2826	0.1877
TYW1B	5.1293	17.5898	19.33	19.5453	0.0769	2e-04	6.3e-05	2e-04	5.1294	17.5962	19.3378	19.5529	0.0769	2e-04	6.3e-05	5.7e-05
MLXIPL	13.1114	64.3913	72.3092	72.3093	0.0014	1e-14	2e-16	1.4e-15	13.1145	64.4947	72.4407	72.4388	0.0014	9.9e-15	1.9e-16	1.9e-16
MET	0.0174	0.0266	0.0034	0.0287	0.9913	0.9868	0.9983	0.9987	0.0174	0.0266	0.0034	0.0287	0.9913	0.9868	0.9983	0.9857
KLF14	15.7125	17.2055	7.9188	17.429	4e-04	2e-04	0.0191	6e-04	15.7174	17.2116	7.9196	17.4348	4e-04	2e-04	0.0191	2e-04
TMEM176A	6.6903	8.4563	0.0724	8.4783	0.0353	0.0146	0.9644	0.0371	6.6908	8.4573	0.0724	8.4791	0.0352	0.0146	0.9644	0.0144
PPP1R3B	53.7114	39.4025	16.5115	54.3406	2.2e-12	2.8e-09	3e-04	9.5e-12	53.7824	39.4396	16.517	54.4119	2.1e-12	2.7e-09	3e-04	1.5e-12
PINX1	2.5749	4.0591	0.2343	4.1013	0.276	0.1314	0.8894	0.2507	2.5748	4.0591	0.2343	4.1012	0.276	0.1314	0.8894	0.1287
NAT2	18.0619	0.5765	17.8156	18.0715	1e-04	0.7496	1e-04	4e-04	18.0687	0.5765	17.8221	18.0777	1e-04	0.7496	1e-04	1e-04
LPL	44.1876	101.1466	94.2665	101.2178	2.5e-10	1.1e-22	3.4e-21	8.5e-22	44.2348	101.4084	94.4932	101.4773	2.5e-10	9.5e-23	3e-21	9.2e-23
SOX17	19.8611	1.6325	19.6471	19.8679	4.9e-05	0.4421	5.4e-05	2e-04	19.8695	1.6324	19.6553	19.8758	4.8e-05	0.4421	5.4e-05	4.8e-05
CYP7A1	14.2966	12.0565	21.2789	21.8113	8e-04	0.0024	2.4e-05	7.1e-05	14.3005	12.0591	21.2887	21.8211	8e-04	0.0024	2.4e-05	1.8e-05
TRPS1	9.2932	18.9106	3.3272	19.0632	0.0096	7.8e-05	0.1895	3e-04	9.2945	18.9181	3.3272	19.0703	0.0096	7.8e-05	0.1895	7.2e-05
TRIB1	27.9709	57.4352	67.2725	67.3248	8.4e-07	3.4e-13	2.5e-15	1.6e-14	27.9887	57.5169	67.3858	67.4364	8.4e-07	3.2e-13	2.3e-15	2.3e-15
PLEC1	6.6961	2.0823	7.4524	7.4539	0.0352	0.3531	0.0241	0.0588	6.6966	2.0822	7.4531	7.4544	0.0351	0.3531	0.0241	0.0241
VLDLR	5.8274	1.8115	6.0405	6.6074	0.0543	0.4042	0.0488	0.0855	5.8277	1.8114	6.0408	6.6077	0.0543	0.4043	0.0488	0.0367
TTC39B	10.3474	24.2408	5	25.0317	0.0057	5.4e-06	0.0821	1.5e-05	10.3491	24.2539	5.0001	25.045	0.0057	5.4e-06	0.0821	3.6e-06
ABCA1	48.2334	56.5429	7.55	62.1244	3.4e-11	5.3e-13	0.0229	2.1e-13	48.2901	56.6219	7.5507	62.2188	3.3e-11	5.1e-13	0.0229	3.1e-14
ABO	60.3861	8.381	52.8147	60.7184	7.7e-14	0.0151	3.4e-12	4.1e-13	60.4767	8.382	52.8833	60.8084	7.4e-14	0.0151	3.3e-12	6.2e-14
AKR1C4	6.8566	3.5325	4.4707	7.6059	0.0324	0.171	0.107	0.0549	6.8571	3.5325	4.4707	7.6064	0.0324	0.171	0.107	0.0223
VIM-CUBN	6.9381	6.1629	3.2495	8.7517	0.0311	0.0459	0.197	0.0328	6.9386	6.1632	3.2494	8.7525	0.0311	0.0459	0.197	0.0126
MARCH8-ALOX5	6.5971	15.3754	4.5271	16.9594	0.0369	5e-04	0.104	7e-04	6.5976	15.38	4.5272	16.9648	0.0369	5e-04	0.104	2e-04
JMJD1C	12.0387	14.9608	18.2939	20.7871	0.0024	6e-04	1e-04	1e-04	12.0412	14.9651	18.3009	20.7958	0.0024	6e-04	1e-04	3e-05
CYP26A1	8.3245	12.7518	11.3935	13.3232	0.0156	0.0017	0.0034	0.004	8.3255	12.7548	11.3957	13.3261	0.0156	0.0017	0.0034	0.0013
GPAM	46.4994	26.5238	22.801	46.5617	8e-11	1.7e-06	1.1e-05	4.3e-10	46.5519	26.5397	22.8124	46.6131	7.8e-11	1.7e-06	1.1e-05	7.6e-11
AMPD3	0.9748	0.7902	0.2916	0.9756	0.6142	0.6736	0.8643	0.8072	0.9747	0.7901	0.2916	0.9755	0.6142	0.6736	0.8643	0.614
SPTY2D1	9.1341	0.7602	8.4943	9.1358	0.0104	0.6838	0.0143	0.0275	9.1353	0.7602	8.4953	9.1368	0.0104	0.6838	0.0143	0.0104
LRP4	18.879	18.9701	2.0116	19.0295	8e-05	7.6e-05	0.3658	3e-04	18.8864	18.9777	2.0115	19.0366	7.9e-05	7.6e-05	0.3658	7.3e-05
OR4C46	11.7201	12.0847	2.3031	14.465	0.0029	0.0024	0.3161	0.0023	11.7225	12.0873	2.303	14.4686	0.0028	0.0024	0.3162	7e-04
FADS1-2-3	72.9529	28.8677	72.4643	85.192	1.4e-16	5.4e-07	1.8e-16	2.4e-18	73.0868	28.8868	72.5964	85.374	1.3e-16	5.3e-07	1.7e-16	2.9e-19
KAT5	2.4241	5.3169	6.6017	6.7704	0.2976	0.0701	0.0369	0.0796	2.424	5.3171	6.6022	6.7707	0.2976	0.0701	0.0368	0.0339
MOGAT2-DGAT2	6.2832	7.7851	2.7738	10.2001	0.0432	0.0204	0.2498	0.0169	6.2836	7.7859	2.7737	10.2015	0.0432	0.0204	0.2499	0.0061
APOA1	76.6588	292.4269	298.9778	299.0589	2.3e-17	3.2e-64	1.2e-65	1.6e-64	76.8071	294.6897	301.3444	301.4188	2.1e-17	1e-64	3.7e-66	3.5e-66

Table C.2: Results for multivariate Wald test and CCA

	Z multi				p- value Wald				χ^2 approximation				p-value CCA			
Gene	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3
PHLDB1	6.8244	2.9284	4.0165	6.8887	0.033	0.2313	0.1342	0.0755	6.8249	2.9283	4.0165	6.8891	0.033	0.2313	0.1342	0.0319
UBASH3B	4.1391	4.3789	0.1974	4.539	0.1262	0.112	0.906	0.2088	4.1391	4.3789	0.1974	4.5389	0.1262	0.112	0.906	0.1034
ST3GAL4	12.232	0.0512	12.4583	12.4608	0.0022	0.9747	0.002	0.006	12.2347	0.0512	12.4611	12.4632	0.0022	0.9747	0.002	0.002
PHC1-A2ML1	8.697	8.4129	2.5534	10.5902	0.0129	0.0149	0.279	0.0142	8.6981	8.4139	2.5533	10.5917	0.0129	0.0149	0.279	0.005
PDE3A	0.5218	0.2773	0.3867	0.6061	0.7704	0.8705	0.8242	0.895	0.5217	0.2773	0.3866	0.6061	0.7704	0.8705	0.8242	0.7386
LRP1	3.9802	4.1005	2.0446	4.4063	0.1367	0.1287	0.3598	0.2208	3.9802	4.1005	2.0445	4.4062	0.1367	0.1287	0.3598	0.1105
MVK	10.2605	11.6417	0.2543	11.8661	0.0059	0.003	0.8806	0.0079	10.2622	11.6441	0.2543	11.8682	0.0059	0.003	0.8806	0.0026
BRAP	10.2383	5.7883	9.5335	12.1607	0.006	0.0553	0.0085	0.0069	10.24	5.7886	9.5349	12.1631	0.006	0.0553	0.0085	0.0023
HNF1A	25.5128	6.7653	23.005	27.3355	2.9e-06	0.034	1e-05	5e-06	25.5274	6.7658	23.0167	27.3518	2.9e-06	0.0339	1e-05	1.1e-06
SBNO1	10.9582	5.8363	10.6178	13.6919	0.0042	0.054	0.0049	0.0034	10.9603	5.8366	10.6196	13.695	0.0042	0.054	0.0049	0.0011
ZNF664	10.5758	10.7823	3.2617	10.7982	0.0051	0.0046	0.1958	0.0129	10.5777	10.7842	3.2617	10.7999	0.005	0.0046	0.1958	0.0045
SCARB1	13.6762	17.7632	2.3987	19.9124	0.0011	1e-04	0.3014	2e-04	13.6797	17.7697	2.3986	19.9203	0.0011	1e-04	0.3014	4.7e-05
BRCA2	0.4488	0.447	0.1467	0.4565	0.799	0.7997	0.9293	0.9283	0.4488	0.447	0.1467	0.4564	0.799	0.7997	0.9293	0.796
NYNRIN	3.6303	0.1244	3.803	3.936	0.1628	0.9397	0.1493	0.2685	3.6303	0.1244	3.803	3.9359	0.1628	0.9397	0.1493	0.1397
ZBTB42-AKT1	20.9493	21.2329	1.8891	21.247	2.8e-05	2.5e-05	0.3888	9.4e-05	20.9588	21.2426	1.889	21.2562	2.8e-05	2.4e-05	0.3889	2.4e-05
CAPN3	1.2131	0.9469	1.2323	1.7257	0.5452	0.6228	0.54	0.6312	1.213	0.9469	1.2322	1.7256	0.5452	0.6229	0.54	0.422
FRMD5	3.517	3.7756	5.8494	5.9057	0.1723	0.1514	0.0537	0.1163	3.5169	3.7756	5.8497	5.9058	0.1723	0.1514	0.0537	0.0522
LIPC	189.4257	306.5293	19.3166	306.5302	7.4e-42	2.7e-67	6.4e-05	3.8e-66	190.3646	309.0185	19.3245	309.0111	4.6e-42	7.9e-68	6.4e-05	7.9e-68
LACTB	5.1506	7.1973	5.0973	7.1974	0.0761	0.0274	0.0782	0.0659	5.1507	7.1979	5.0975	7.1979	0.0761	0.0274	0.0782	0.0274
PDXDC1	2.0877	3.1517	5.2509	5.2771	0.3521	0.2068	0.0724	0.1526	2.0876	3.1516	5.2511	5.2771	0.3521	0.2068	0.0724	0.0715
CTF1	3.4014	2.9463	1.654	4.271	0.1826	0.2292	0.4374	0.2336	3.4014	2.9462	1.6539	4.271	0.1826	0.2292	0.4374	0.1182
FTO	10.2955	5.5596	12.6165	12.647	0.0058	0.0621	0.0018	0.0055	10.2973	5.5598	12.6194	12.6496	0.0058	0.062	0.0018	0.0018
CETP	497.3125	551.9894	24.1145	574.3615	1e-108	1.4e-120	5.8e-06	3.6e-124	503.943	560.1809	24.1273	583.2247	3.7e-110	2.3e-122	5.8e-06	2.3e-127
LCAT	67.0045	67.2844	7.1976	67.5213	2.8e-15	2.5e-15	0.0274	1.4e-14	67.1168	67.3977	7.1982	67.6336	2.7e-15	2.3e-15	0.0273	2.1e-15
HPR	31.0882	2.4839	28.9497	31.0886	1.8e-07	0.2888	5.2e-07	8.1e-07	31.1106	2.4838	28.9689	31.1101	1.8e-07	0.2888	5.1e-07	1.8e-07
CMIP	8.4372	8.2712	1.1621	8.5754	0.0147	0.016	0.5593	0.0355	8.4382	8.2721	1.162	8.5762	0.0147	0.016	0.5593	0.0137
DLG4	12.2447	7.4235	13.6234	14.3863	0.0022	0.0244	0.0011	0.0024	12.2474	7.4241	13.6269	14.3899	0.0022	0.0244	0.0011	8e-04
STARD3	13.663	10.392	5.3684	13.8119	0.0011	0.0055	0.0683	0.0032	13.6665	10.3938	5.3686	13.8151	0.0011	0.0055	0.0683	0.001
MPP3	8.1524	14.8721	16.3265	16.67	0.017	6e-04	3e-04	8e-04	8.1533	14.8764	16.3319	16.6751	0.017	6e-04	3e-04	2e-04
OSBPL7	8.9475	1.3305	8.4356	9.1845	0.0114	0.5142	0.0147	0.0269	8.9487	1.3304	8.4366	9.1856	0.0114	0.5142	0.0147	0.0101
APOH-PRXCA	1.5958	10.0184	7.5503	12.0701	0.4503	0.0067	0.0229	0.0071	1.5957	10.02	7.551	12.0724	0.4503	0.0067	0.0229	0.0024
ABCA8	7.4908	1.3111	6.6444	7.8792	0.0236	0.5192	0.0361	0.0486	7.4915	1.311	6.6449	7.8798	0.0236	0.5192	0.0361	0.0195
PGS1	14.0618	5.9068	11.8484	15.9875	9e-04	0.0522	0.0027	0.0011	14.0655	5.9071	11.8508	15.9921	9e-04	0.0522	0.0027	3e-04
LIPG	35.3101	48.5985	2.9682	49.6511	2.2e-08	2.8e-11	0.2267	9.5e-11	35.3395	48.6561	2.9681	49.7101	2.1e-08	2.7e-11	0.2267	1.6e-11
MC4R	0.1315	0.0309	0.178	0.1786	0.9363	0.9847	0.9149	0.981	0.1315	0.0309	0.178	0.1786	0.9364	0.9847	0.9149	0.9146
INSR	1.7445	1.7266	0.3504	1.7478	0.418	0.4218	0.8393	0.6264	1.7444	1.7264	0.3503	1.7477	0.418	0.4218	0.8393	0.4173
ANGPTL4	11.0029	8.786	2.7297	11.2437	0.0041	0.0124	0.2554	0.0105	11.005	8.7872	2.7297	11.2456	0.0041	0.0124	0.2554	0.0036
LDLR	266.8577	13.3142	263.0861	268.94	1.1e-58	0.0013	7.4e-58	5.2e-58	268.7379	13.3175	264.9128	270.8428	4.4e-59	0.0013	3e-58	1.5e-59
ANGPTL8	30.4976	36.7312	7.7109	41.7706	2.4e-07	1.1e-08	0.0212	4.5e-09	30.5191	36.7632	7.7117	41.8114	2.4e-07	1e-08	0.0212	8.3e-10
CILP2	38.9781	56.2914	76.382	77.7685	3.4e-09	6e-13	2.6e-17	9.2e-17	39.0143	56.3697	76.5292	77.9191	3.4e-09	5.7e-13	2.4e-17	1.2e-17
PEPD	11.3153	10.7512	1.5188	11.4182	0.0035	0.0046	0.4679	0.0097	11.3175	10.7531	1.5187	11.4201	0.0035	0.0046	0.468	0.0033
APOE	265.2794	29.4385	255.5503	265.2905	2.5e-58	4.1e-07	3.2e-56	3.2e-57	267.1372	29.4584	257.2727	267.1413	9.8e-59	4e-07	1.4e-56	9.8e-59
FLJ36070	0.3686	2.1253	0.9518	2.1361	0.8317	0.3455	0.6213	0.5446	0.3685	2.1252	0.9518	2.1359	0.8317	0.3456	0.6213	0.3437
HAS1	12.6292	15.1501	0.0904	15.2172	0.0018	5e-04	0.9558	0.0016	12.6321	15.1546	0.0904	15.2213	0.0018	5e-04	0.9558	5e-04
LILRA3	29.1855	27.7717	6.1471	29.2379	4.6e-07	9.3e-07	0.0463	2e-06	29.2051	27.7892	6.1474	29.2568	4.6e-07	9.2e-07	0.0462	4.4e-07
SPTLC3	0.7503	0.0268	0.7309	0.7542	0.6872	0.9867	0.6939	0.8604	0.7502	0.0268	0.7309	0.7541	0.6872	0.9867	0.6939	0.6859
SNX5	9.3346	7.8228	3.6608	9.3487	0.0094	0.02	0.1604	0.025	9.336	7.8236	3.6607	9.3497	0.0094	0.02	0.1604	0.0093
ERGIC3	8.0345	0.7795	8.1235	8.1343	0.018	0.6772	0.0172	0.0433	8.0353	0.7794	8.1244	8.135	0.018	0.6772	0.0172	0.0171
MAFB	4.4075	1.006	4.3991	4.7563	0.1104	0.6047	0.1109	0.1905	4.4075	1.006	4.3991	4.7563	0.1104	0.6047	0.1109	0.0927
TOP1	17.8849	1.9172	17.9695	18.234	1e-04	0.3834	1e-04	4e-04	17.8915	1.9171	17.9761	18.2404	1e-04	0.3834	1e-04	1e-04

Table C.2: Results for multivariate Wald test and CCA

	Z multi				p- value Wald				χ^2 approximation				p-value CCA			
Gene	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3	HDL,LDL	HDL,TG	LDL,TG	All 3
HNF4A	50.027	45.1894	11.7516	56.371	1.4e-11	1.5e-10	0.0028	3.5e-12	50.0882	45.2389	11.754	56.448	1.3e-11	1.5e-10	0.0028	5.5e-13
PLTP	14.5239	14.7459	4.6102	14.9447	7e-04	6e-04	0.0998	0.0019	14.5279	14.7501	4.6102	14.9487	7e-04	6e-04	0.0997	6e-04
UBE2L3	16.3866	21.5412	0.2411	21.6492	3e-04	2.1e-05	0.8864	7.7e-05	16.392	21.5513	0.2411	21.6588	3e-04	2.1e-05	0.8864	2e-05
MTMR3	4.8217	2.6355	7.0931	7.3331	0.0897	0.2677	0.0288	0.062	4.8218	2.6354	7.0937	7.3336	0.0897	0.2678	0.0288	0.0256
TOM1	13.8411	15.0159	8.0998	20.1125	0.001	5e-04	0.0174	2e-04	13.8447	15.0203	8.1006	20.1206	0.001	5e-04	0.0174	4.3e-05
PLA2G6	8.8738	7.3176	2.4914	9.311	0.0118	0.0258	0.2877	0.0254	8.8749	7.3183	2.4913	9.312	0.0118	0.0258	0.2878	0.0095
PPARA	12.0561	5.4606	14.5185	14.8903	0.0024	0.0652	7e-04	0.0019	12.0587	5.4608	14.5226	14.8942	0.0024	0.0652	7e-04	6e-04