UNIVERSITY OF
CAMBRIDGE

Genomics of lipid metabolism

Identification of genetic determinants of lipid metabolites and the effect of perturbations of lipid levels on coronary heart disease risk factors

Eric Leigh Harshfield

St John's College

This dissertation is submitted for the degree of
Doctor of Philosophy (PhD)

Department of Public Health & Primary Care
University of Cambridge

September 2017

Summary

Background Coronary heart disease (CHD) is one of the leading causes of death worldwide, and global mortality rates are expected to continue to rise over the coming decades. In Pakistan in particular, chronic diseases are responsible for 50 % of the total disease burden. Circulating lipids are strongly and linearly associated with risk of CHD; however, despite considerable efforts to demonstrate causality, available evidence is conflicting and insufficient. Study of the underlying metabolic pathways implicated in the association between lipids and CHD would help to disentangle and elucidate these complex relationships.

Objectives The primary objectives of this dissertation were to (1) identify the genetic determinants of lipid metabolites and (2) advance understanding of the effect of perturbations in lipid metabolite levels on CHD and its risk factors.

Methods Direct infusion high-resolution mass spectrometry was performed on 5662 participants from the Pakistan Risk of Myocardial Infarction Study to obtain signals for 444 known lipid metabolites. Correlations and associations of the lipids with smoking, physical activity, circulating biomarkers, and other CHD risk factors were assessed. Genome-wide analyses were conducted to analyse the association of each lipid with over 6.7 million imputed single nucleotide polymorphisms. Functional annotation and Gaussian Graphical Modelling were used to link the variants associated with each lipid to the most likely mediating gene, discern the underlying metabolic pathways, and provide a visual representation of the genetic determinants of human metabolism. Mendelian randomisation was also implemented to examine the causal effect of lipids on risk of CHD.

Results The lipids were highly correlated with each other and with levels of major circulating lipids, and they exhibited significant associations with several CHD risk factors. There were 254 lipids that had significant associations with one or more genetic variants and 355 associations between lipids and variants, with a total of 89 sentinel variants from 23 independent loci. The analyses described in this dissertation resulted in the discovery of four novel loci, identified novel relationships between genetic variants and lipids, and revealed new biological insights into lipid metabolism.

Conclusion Analyses of lipid metabolites in large epidemiological studies can contribute to enhanced understanding of mechanisms for CHD development and identification of novel causal pathways and new therapeutic targets.

Dedication

This dissertation is dedicated to my family.

> “*Knowledge is as wings to man’s life, and a ladder for his ascent. Its acquisition is incumbent upon everyone. The knowledge of such sciences, however, should be acquired as can profit the peoples of the earth, and not those which begin with words and end with words. Great indeed is the claim of scientists and craftsmen on the peoples of the world. . . . In truth, knowledge is a veritable treasure for man, and a source of glory, of bounty, of joy, of exaltation, of cheer and gladness unto him.*”
>
> — Bahá’u’lláh, *Tajallíyát (Effulgences),* in *Tablets of Bahá’u’lláh Revealed after the Kitáb-i-Aqdas*, pp. 51-52

Preface

The central focus of the work presented in this dissertation is the acquisition, processing, and analysis of data on levels of lipid metabolites that were measured using direct infusion high-resolution mass spectrometry. The tools of epidemiology and genetics are employed to aid understanding of the relevant metabolic pathways, the genetic determinants of lipids, and potential causal mechanisms that may lead to chronic disease outcomes. In addition to the work described here, I have conducted other analyses outside the scope of this dissertation during the course of my PhD, such as the association of depressive symptoms with risk of cardiovascular diseases and cause-specific mortality in the Emerging Risk Factors Collaboration (ERFC) and UK Biobank, and the association of hypertension and hyperglycaemia with wealth and educational attainment in Bangladesh. Much of this work has contributed to scientific manuscripts, which are listed in Appendix B.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as indicated in the Acknowledgements or specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed limit of 60 000 words (excluding figures, tables, references, and appendices).

Eric Leigh Harshfield
September 2017

Acknowledgements

Although the work described in this dissertation is my own, it would not have been possible without the help and support of a number of individuals. I would like to thank my primary supervisor, Dr. Angela Wood, and my co-supervisors, Dr. Adam Butterworth and Professor John Danesh, for their advice, guidance, and support throughout the course of my doctoral studies. I am particularly grateful to Dr. Angela Wood for dedicating her time to regularly meet with me and providing helpful comments.

A number of other people have contributed to the work presented in this dissertation and their contributions are gratefully acknowledged. Dr. Albert Koulman and Dr. Julian Griffin have provided expert advice and contributions in many aspects related to processing, analysing, and interpreting the lipidomics data. The staff and laboratory technicians at the Medical Research Council (MRC) Human Nutrition Research (HNR) laboratory in Cambridge (which has now been restructured and renamed as the MRC Elsie Widdowson Laboratory) provided invaluable contributions by processing the serum samples and obtaining the lipid signal data. I would also like to thank the members of the Cardiovascular Epidemiology Unit, especially those who provided useful comments and feedback during the numerous opportunities in which I was able to present my work to colleagues during the course of my research.

During the course of my PhD, I established a PROMIS lipidomics working group to help advance the progress of the analyses and completion of the manuscripts. I would like to thank all of the members of the working group—Albert Koulman, Jules Griffin, Angela Wood, Adam Butterworth, David Stacey, Dirk Paul, Daniel Ziemek, Eric Fauman, and Danish Saleheen—who took part in biweekly teleconferences, whose valuable insights have helped enhance the quality and thoroughness of the research and the manuscripts.

On a personal note, I would like to my family and friends for their continued support over the past four years. In particular, I would like to thank my wife, Amelia, for her immense encouragement and unwavering support throughout the whole process, as well as reading through all of the chapters and providing helpful feedback and suggestions.

I would also like to thank several individuals who contributed specifically to certain aspects of the work presented in this dissertation, the details of which are provided below.

Chapter 1: I conducted extensive research and wrote the text for this chapter. Several of the figures were taken from previous publications or online sources, and have been referenced appropriately. Angela Wood provided helpful comments. Some of the text for this chapter, in particular the literature review of genetics and metabolomics studies, was incorporated into a manuscript on the genetics of lipids, which was recently published in *Cardiovascular Resesearch* (listed in Appendix B). Additionally, parts of the section on Mendelian randomisation that appears in this chapter was originally written for a manuscript that I co-authored with Stephen Burgess on Mendelian randomisation using high-dimensional "omics" platforms, which was published in *Current Opinion in Endocrinology, Diabetes, and Obesity* (listed in Appendix B)

Chapter 2: Danish Saleheen and John Danesh are the co-principal investigators of the Pakistan Risk of Myocardial Infarction Study (PROMIS). Much of the text describing the study is derived from the protocol paper, which has been referenced appropriately in the text. The overall design of the study, collection of questionnaire data, and measurement of circulating biomarkers and major blood lipids were undertaken by PROMIS researchers and technicians, including a number of field workers in Pakistan. Field-work, genotyping, and standard clinical chemistry assays in PROMIS were principally supported by grants awarded to the University of Cambridge from the British Heart Foundation, UK Medical Research Council, Wellcome Trust, European Commission Framework Framework 6–funded Bloodomics Integrated Project, Pfizer, Novartis, and Merck. I adapted and modified SAS scripts, which were originally written by Mat Walker, to perform data management for PROMIS as a whole, as well as for the specific dataset used for analysis in this dissertation. I cleaned the PROMIS data and created a merged dataset containing the phenotype information for the full set of PROMIS participants with identifiers linking the phenotype data to the lipidomics data for the subset of participants who had this information. Robin Young cleaned the genetic data in PROMIS and performed quality control and genetic imputation. Weang Kee Ho identified cryptically-related individuals based on the genetic data and ran a principal component analysis to produce the multi-dimensional scaling matrix, which I used to account for ancestry when I performed the genome-wide association analyses. I drafted the text for this chapter and Angela Wood provided helpful comments. Some of the text for this chapter and the two chapters that follow has also been incorporated into a manuscript describing the lipidomics study and its applications to coronary heart disease and its risk factors, which is currently under review at time of writing (listed in Appendix B). I conducted all of the analyses for the manuscript, and I am the lead author and drafted and edited the text. However, the other co-authors who were part of the PROMIS lipidomics working group provided helpful comments and input on the manuscript, which were taken into consideration when writing this chapter.

Chapter 3: Serum samples from PROMIS participants were sent from Pakistan by Asif Rasheed and Danish Saleheen to a laboratory in the Cardiovascular Epidemiology Unit at the University of Cambridge. Nasir Sheikh and Philip Haycock assisted with the laboratory work by pipetting the serum samples into 1.2 mL Cryovial tubes and arranging the samples onto 96-well plates according to a specified plan. The sample placement was determined by a block randomisation algorithm, which I implemented with helpful advice from Michael Eiden. Luke Marney wrote R and Python scripts, which I adapted, to process the raw lipidomics data and perform a peak-picking algorithm to select peaks corresponding to known lipids of interest. Albert Koulman performed quality control on the cleaned lipidomics dataset by comparing the spectral data of the lipid metabolites with that of the blanks and QC samples. Although a large number of individuals were involved in setting up the PROMIS study as a whole, as well as bringing together the specific pieces of information that were necessary for the analyses described in this dissertation to take place, I was ultimately responsible and led all of the steps from start to finish. I devised the protocols and overall analysis strategy for the entire lipidomics project, including obtaining the samples, performing the block randomisation, processing and cleaning the raw data, and performing the analyses. I produced all of the tables and figures myself, or adapted them from other sources as indicated, which have been referenced appropriately. I wrote the text of this chapter, and Angela Wood provided helpful comments.

Chapter 4: I wrote the statistical analysis plan in consultation with Angela Wood and Albert Koulman, who provided helpful input and suggestions for the analyses that appear in this chapter. I analysed the data myself, produced all tables and figures, and drafted the text.

Chapter 5: I wrote the protocol and analysis plan for conducting the univariate genome-wide association analyses for each lipid metabolite in consultation with Angela Wood, Adam Butterworth, and Jo Howson, who provided helpful advice and guidance. The list of analysed ratios with biological significance was curated by Albert Koulman and Jules

Griffin. I wrote the R scripts and job submission scripts that were used to run the genome-wide association analyses, filter and process the results, run the meta-analyses for each trait, and produce the Manhattan plots and Q-Q plots, though some scripts were based on templates from Robin Young and Praveen Surendran that I adapted and modified. Genetic analyses were performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (http://www.hpc.cam.ac.uk), which is provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England and funding from the Science and Technology Facilities Council. I wrote the text of this chapter, and Angela Wood and Adam Butterworth provided helpful comments.

Chapter 6: Mihir Kamat assisted with annotating the lead variants from the conditional analyses using VEP and ANNOVAR. David Stacey and Dirk Paul assisted with the functional annotation of the variants and identification of probable causal genes. Eric Fauman contributed additional evidence to support the functional annotation pipeline, and Daniel Ziemek and David Stacey assisted with the incorporation of genetic and GGM data into the network diagrams. Although these individuals made important contributions to these specific aspects of the work, I coordinated and planned the analysis strategy and oversaw all aspects of the analyses and interpretation of the results. I wrote the text of this chapter, and Angela Wood and Adam Butterworth provided helpful comments. A manuscript is currently in preparation which describes the GWAS results and interpretation (listed in Appendix B), which is largely based on content from this chapter and the previous chapter. Comments on the manuscript provided by the PROMIS lipidomics working group were also taken into consideration when writing this chapter.

Chapter 7: I wrote the protocol and analysis plan for conducting the Mendelian randomisation analysis, with advice from Angela Wood and Stephen Burgess. Stephen Burgess also provided the R code that I adapted and used to run the Mendelian randomisation analyses on my own data. I used PhenoScanner to identify a list of proxy variants for each of the sentinel variants from the conditional analyses. James Staley assisted by looking up proxy variants for the multi-allelic SNPs, indels (insertion–deletion SNPs), and SNPs that were not able to be found in 1000 Genomes phase 3. I wrote the text of this chapter, and Angela Wood provided helpful comments.

Chapter 8: I conducted all the analyses that were described using the INTERVAL data. I wrote the text of this chapter, and Angela Wood provided helpful comments.

List of Abbreviations

Abbreviation	Description
ALD	Alcohol-related Liver Disease
AMU	Atomic Mass Units
ANNOVAR	ANNOtate VARiation
BHC	Bayesian Hierarchical Clustering
BMI	Body Mass Index
bp	Base Pairs (1 bp = 3.4 Å = 340 pm)
CE	Cholesteryl Ester
Cer	Ceramide
CAD	Coronary Artery Disease
CEU	Cardiovascular Epidemiology Unit
CHD	Coronary Heart Disease
CI	Confidence Interval
CNCD	Center for Non-Communicable Diseases
CSV	Comma-Separated Values
CV	Coefficient of Variation
CVD	Cardiovascular Disease
DBP	Diastolic Blood Pressure
DG	Diacylglycerol/Diglyceride
DIHRMS	Direct Infusion High-Resolution Mass Spectrometry
DNA	Deoxyribonucleic Acid
EA	Effect Allele
EDTA	Ethylenediaminetetraacetic Acid
EFO	Experimental Factor Ontology
EPIC	European Prospective Investigation into Cancer and Nutrition
ERFC	Emerging Risk Factors Collaboration
eQTL	Expression Quantitative Trait Loci
FA	Fatty Acyl
FDR	False Discovery Rate
FPG	Fasting Plasma Glucose
FreeFA	Free Fatty Acid
GC	Gas Chromatography
GCKR	Glucokinase Regulatory Protein
GCTA	Genome-wide Complex Trait Analysis
GGM	Gaussian Graphical Model
GL	Glycerolipid
GLGC	Global Lipids Genetics Consortium
GO	Gene Ontology
GP	Glycerophospholipid
GRCh37	Genome Reference Consortium human genome (build 37)
GTEx	Genotype-Tissue Expression
GWAS	Genome-Wide Association Study
HbA_{1c}	Glycated Haemoglobin (Haemoglobin A_{1c})
HDL-C	High-Density Lipoprotein Cholesterol
HGNC	HUGO Gene Nomenclature Committee
HMDB	Human Metabolome Database
HNR	Human Nutrition Research
HPCS	High Performance Computing Service
HUGO	Human Genome Organisation
HWE	Hardy-Weinberg Equilibrium
IHD	Ischaemic Heart Disease
IPA	Ingenuity Pathway Analysis
IV	Instrumental Variable
Kb	Kilo-base pairs (1000 base pairs)
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	Liquid Chromatography
LCL	Lymphoblastoid Cell Lines

List of Abbreviations (. . . continued)

Abbreviation	Description
LD	Linkage Disequilibrium
LDL-C	Low-Density Lipoprotein Cholesterol
LIPID MAPS	Lipid Metabolites And Pathways Strategy
LMPD	LIPID MAPS Proteome Database
Lp(a)	Lipoprotein(a)
LysoPC	Lysophosphocholine
MAF	Minor Allele Frequency
Mb	Mega-base pairs (1 000 000 base pairs)
MGI	Mouse Genome Informatics
mGWAS	Metabolomics Genome-Wide Association Study
MI	Myocardial Infarction
MR	Mendelian Randomisation
MRC	Medical Research Council
MS	Mass Spectrometry
MTBE	Methyl *Tert*-Butyl Ether
m/z	Mass-to-charge ratio
mzXML	Mass-to-charge ratio eXtensible Markup Language
NAFLD	Non-Alcoholic Fatty Liver Disease
NASH	Non-Alcoholic Steatohepatitis
NEA	Non-Effect Allele
NMR	Nuclear Magnetic Resonance
OMIM	Online Mendelian Inheritance in Man
OR	Odds Ratio
PA	Phosphatic Acid
PBS	Phosphate Buffered Saline
PC	Phosphatidylcholine
PCA	Principal Component Analysis
PE	Phosphatidylethanolamine
PG	Phosphatidylglycerol
PI	Phosphatidylinositol
PLS-DA	Partial Least Squares Discriminant Analysis
PROMIS	Pakistan Risk of Myocardial Infarction Study
PS	Phosphatidylserine
PUFA	Polyunsaturated Fatty Acid
QC	Quality Control
Q-Q	Quantile-Quantile
SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Standard Error
SM	Sphingomyelin
SNP	Single nucleotide polymorphism
SP	Sphingolipid
ST	Sterol lipid
T2D	Type 2 Diabetes Mellitus
TG	Triacylglycerol/Triglyceride
UK	United Kingdom
VEP	Variant Effect Predictor
VLDL-C	Very-Low-Density Lipoprotein Cholesterol
YLL	Years of Life Lost

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction and literature review:

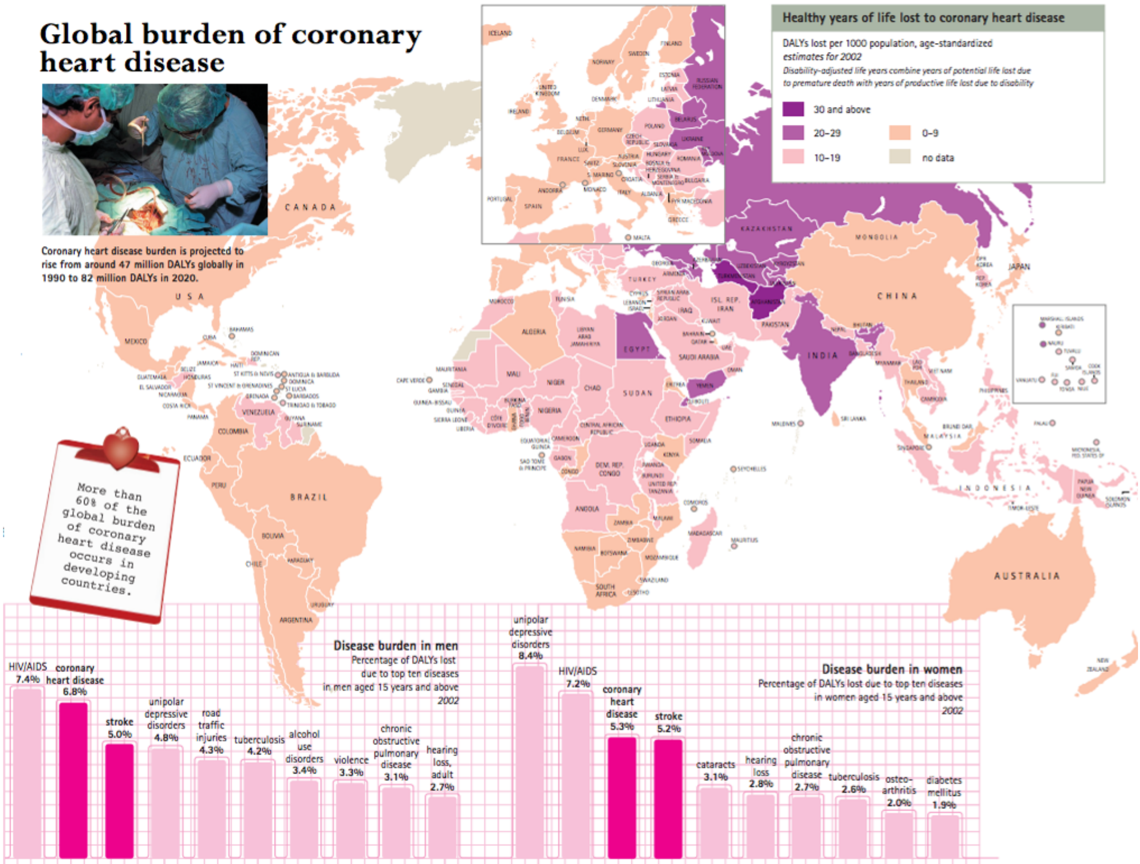
Cardiovascular disease, genetics, and metabolomics

Chapter summary

Cardiovascular disease (CVD) is one of the leading causes of mortality and morbidity worldwide. Although advances in medicine, particularly the use of statins, have reduced mortality rates from CVD in many Western countries during the past half-century, it remains a major cause of death and morbidity, and the burden of coronary heart disease (CHD), which is one of the most common forms of CVD, has been increasing at an alarming rate in South Asian populations. In order to reduce the burden of CVD in our society, a more thorough understanding of the various genetic, lifestyle, and environmental risk factors is needed. This dissertation aimed to measure and analyse lipid profiles in a Pakistani population, assess the association of hundreds of lipids with circulating biomarkers and CHD risk factors, identify the genetic determinants of these lipids, and determine their causal relevance and potential clinical applications.

This chapter provides an introduction and literature review of established risk factors for CHD, with an emphasis on levels of major blood lipids. Introductions to genetics and metabolomics are also provided, and the current state of knowledge is described vis-à-vis the link between genetics, levels of lipid metabolites, and CHD. A literature review is included that summarises all of the currently published metabolomics GWAS studies and their main findings. Additionally, an overview of Mendelian randomisation (MR) is provided which describes some of the recent methodological advances in MR and the application of these approaches to assessing the causal relevance of major lipids and metabolites for risk of CHD. Finally, a conceptual framework of the overall approach used in this dissertation is presented along with a summary of the contents of each chapter.

Figure 1.1: Map of global burden of CHD


Global burden of coronary heart disease
Coronary heart disease burden is projected to rise from around 47 million DALYs globally in 1990 to 82 million DALYs in 2020.
More than 60% of the global burden of coronary heart disease occurs in developing countries.
Healthy years of life lost to coronary heart disease
DALYs lost per 1000 population, age-standardized estimates for 2002
Disability-adjusted life years combine years of potential life lost due to premature death with years of productive life lost due to disability
30 and above
20–29
10–19
0–9
no data
CANADA
USA
MEXICO
BRAZIL
BOLIVIA
CHILE
PARAGUAY
URUGUAY
ARGENTINA
COLOMBIA
VENEZUELA
PERU
ECUADOR
ALGERIA
LIBYAN ARAB JAMAHIRIYA
EGYPT
MALI
NIGER
CHAD
SUDAN
NIGERIA
ETHIOPIA
ANGOLA
ZAMBIA
ZIMBABWE
NAMIBIA
BOTSWANA
MOZAMBIQUE
MADAGASCAR
SOUTH AFRICA
TURKEY
ISL. REP. IRAN
SAUDI ARABIA
KAZAKHSTAN
MONGOLIA
CHINA
INDIA
PAKISTAN
JAPAN
INDONESIA
AUSTRALIA
NEW ZEALAND
Disease burden in men
Percentage of DALYs lost due to top ten diseases in men aged 15 years and above
2002
HIV/AIDS 7.4%
coronary heart disease 6.8%
stroke 5.0%
unipolar depressive disorders 4.8%
road traffic injuries 4.3%
tuberculosis 4.2%
alcohol use disorders 3.4%
violence 3.3%
chronic obstructive pulmonary disease 3.1%
hearing loss, adult 2.7%
Disease burden in women
Percentage of DALYs lost due to top ten diseases in women aged 15 years and above
2002
unipolar depressive disorders 8.4%
HIV/AIDS 7.2%
coronary heart disease 5.3%
stroke 5.2%
cataracts 3.1%
hearing loss 2.8%
chronic obstructive pulmonary disease 2.7%
tuberculosis 2.6%
osteo-arthritis 2.0%
diabetes mellitus 1.9%

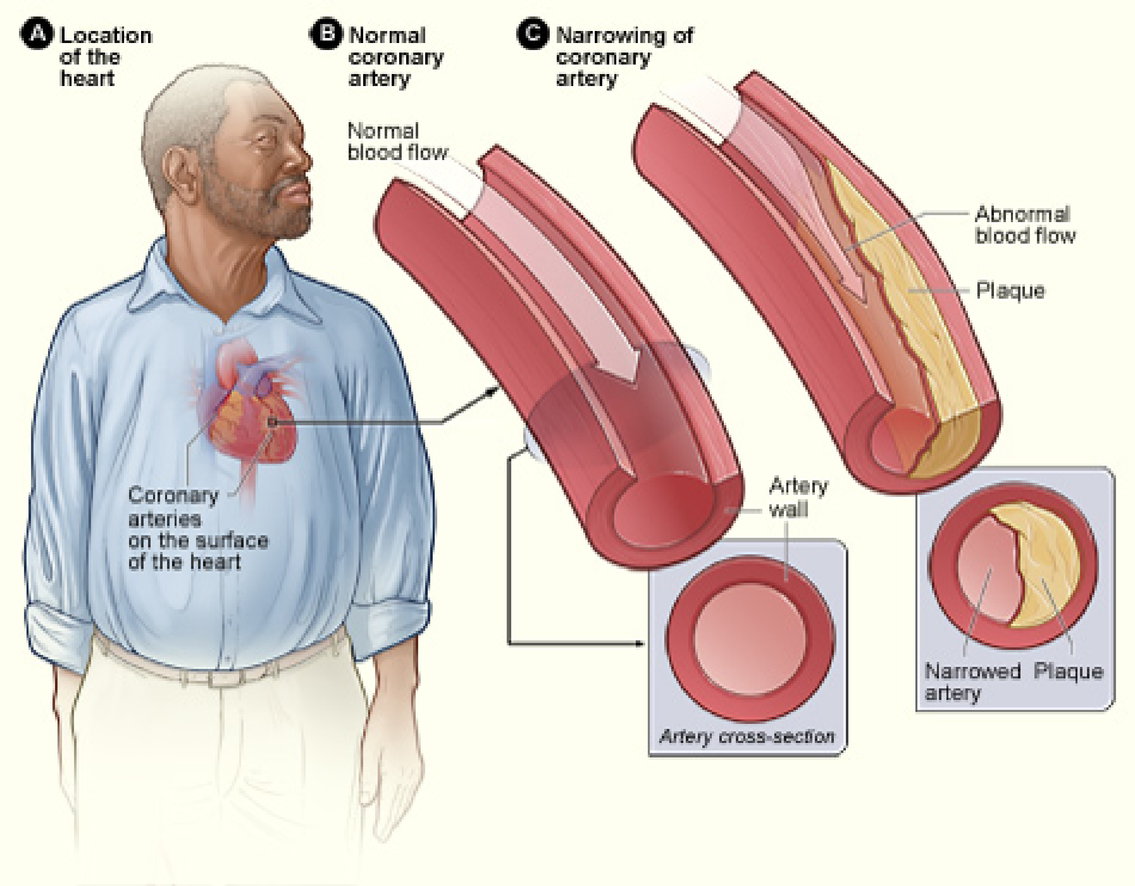
The map shows the disability-adjusted life years (DALYs) lost per 1000 population, with the number of healthy years of life lost due to CHD shown as a spectrum from 0-9 (coloured peach) to 30 and above (dark purple). The bar charts show the top ten diseases resulting in the highest percentage of DALYs lost in men (left) and women (right). Source: World Health Organization, 2004[3].

1.1 Cardiovascular disease

Infectious diseases and malnutrition-related childhood illnesses were once the primary causes of death, but in recent decades, the global burden of disease has been gradually shifting from communicable to non-communicable diseases, and there has been a concomitant rise in the incidence of chronic diseases[1]. Cardiovascular disease (CVD) is one of the leading causes of death worldwide and mortality rates are expected to continue to rise over the next twenty years[2]. As the map in Figure 1.1 illustrates, coronary heart disease (CHD) and stroke, two of the most common forms of CVD, are among the top ten diseases with the highest burden (in terms of healthy years of life lost) in both men and women.

Ischaemic heart disease (IHD), also known as coronary artery disease (CAD), occurs when fatty deposits of plaque build up within the walls of the coronary arteries until blood flow to the heart is restricted. This process is called atherosclerosis and results in CHD. The terms IHD, CAD, and CHD are therefore often used interchangeably. A

Figure 1.2: Build-up of plaque in the coronary arteries through atherosclerosis



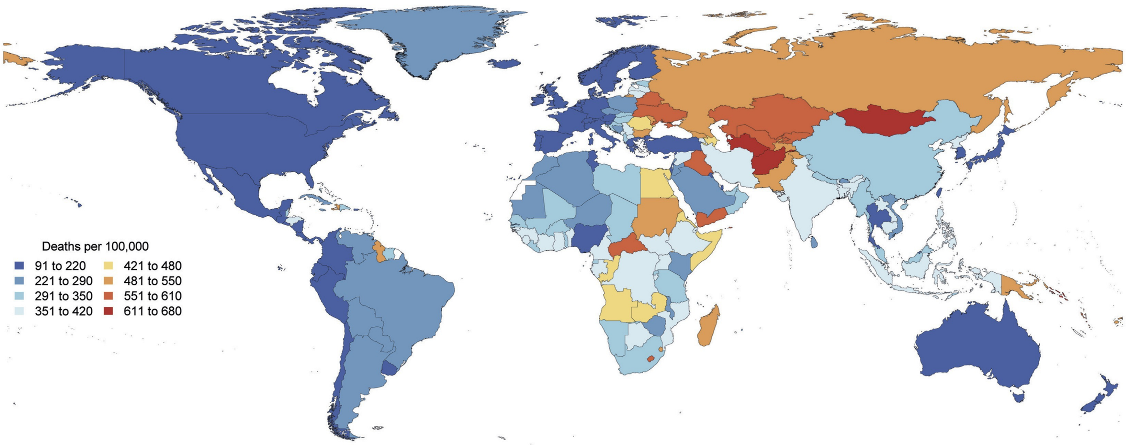
Source: National Heart, Lung, and Blood Institute, 2015[4].

diagram depicting the difference between individuals with healthy arteries and those with atherosclerosis in the coronary arteries is shown in Figure 1.2. In a healthy individual, blood flows throughout the body delivering nutrients and oxygen to all the cells and vital organs. However, when the blood flow has been obstructed due to the build-up of plaque along the walls of the arteries, this can lead to a heart attack or stroke. A stroke is similar to CHD, except instead of blood flow being restricted to the heart, the blood supply is cut-off from reaching the brain.

1.1.1 Burden of CHD in South Asia

Throughout most of South Asia, lower respiratory infections, diarrhoeal diseases, and congenital anomalies are among the leading causes of premature mortality[1]. However, in Pakistan, chronic diseases are responsible for 50 % of the total disease burden[5]. Years of life lost (YLLs) is a measure of premature mortality that provides an estimate of the average number of years a person would have lived if he or she had not died prematurely. IHD and stroke are two of the top ten causes of YLLs in Pakistan, and observed YLLs due to IHD

Figure 1.3: Global map of age-standardised death rate of CVD



Countries with the lowest death rate due to CVD are shown in dark blue (91 to 220 deaths per 100 000 people), while countries with the highest death rate due to CVD are shown in dark red (611 to 680 deaths per 100 000 people) (data as of 2015). Source: Roth GA, et al. *J Am Coll Cardiol.* 2017;70(1):1-25[8].

are nearly twice as high as expected (ratio of 1.81 in 2015) based on socio-demographic index[1]. Pakistan is also among the top ten countries in the world with the highest number of people living with diabetes[6].

The age-standardised mortality rate due to CVD increased in Pakistan between 2000 to 2012 from 250.6 to 274.2 per 100 000 population, with a more pronounced increase in males than in females[7]. Over this time period, the CVD mortality rate decreased in 153 out of 172 countries; Pakistan was one of only 19 countries where the CVD mortality rate actually worsened rather than improving[7]. Additionally, Pakistan was the only country in South Asia where the CVD mortality rate increased[7]. A map of the CVD mortality rate for each country is shown in Figure 1.3.

Studies have shown that South Asians who migrate to other countries are at increased risk for CHD compared with the reference population of their adopted countries[6,9]. Amongst males living in England and Wales, CHD is responsible for 27 % of all deaths in South Asians compared with 18 % overall[9]. For women, the difference is less striking but still noticeable, with CHD responsible for 18 % of all deaths in South Asian females, compared with 13 % for the overall female population[9]. The prevalence of CHD is highest in Pakistani men living in England than in any other ethnic group in this country[9].

Understanding the reasons for the high burden of CHD in South Asians, and Pakistanis in particular, would aid efforts to reduce CHD incidence in this population. The findings could also have important applications and implications for reducing risk of CHD in other parts of the world.

1.1.2 Established risk factors for CHD

In addition to certain non-modifiable risk factors for CHD, such as age, gender, ethnicity, and family history of CVD, a number of modifiable risk factors have been identified, which can be influenced by the lifestyle choices that one makes. A few of the most common modifiable risk factors are hypertension (high blood pressure), smoking or tobacco use, diabetes, high cholesterol levels, lack of physical exercise, an unhealthy diet, and obesity[4,7], although stress and alcohol consumption are also contributing factors. The genes that one inherits at birth can also influence both modifiable and non-modifiable risk factors and prescribe a portion of an individual's overall CHD risk (approximately 28 %[10]), regardless of other environmental factors.

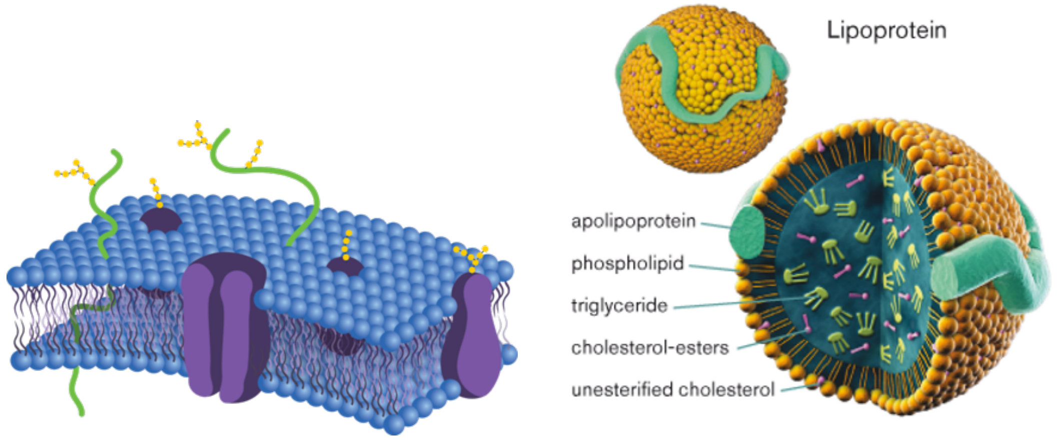
Despite existing knowledge of the major risk factors for CHD and the influence of abnormal levels of major lipids, there is much less known about how levels of the hundreds of individual lipid subfractions and other metabolic markers are involved in the development of atherosclerosis and the onset of CHD[11].

1.1.3 Major blood lipids and risk of CHD

Lipids are essential for life, and have several important biological functions. These include (1) energy storage, since lipids are largely composed of fats; (2) formation of a phospholipid bilayer that is a core part of the cellular membrane and protects the cell (Figure 1.4a); (3) signalling, as lipids are used as messengers and form signalling molecules that influence cellular responses; and (4) transport, since lipids assemble to form lipoproteins that carry vitamins and nutrients throughout the body (Figure 1.4b)[11]. Given the important role that lipids play in the body, the "lipidome", which is the totality of lipid molecules in cells, can therefore reflect underlying metabolic processes that may be influenced by dietary, environmental, and genetic factors[12].

Despite the diversity of lipid species and the wide array of functions that lipids are involved in, most studies of lipids up until recent years have been relatively crude, as they have mainly focused on major circulating lipids that can be easily measured by standard assays, such as total cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides. Several lipid-related targets have been studied as potential pathways that could be modified to reduce the risk of dyslipidaemia, a condition in which abnormal levels of blood lipids contribute to the development of atherosclerosis. Apolipoprotein A-V (ApoA5), apolipoprotein C-III (ApoC3), angiopoietin-like 3 (ANGPTL3), and lipoprotein(a) [Lp(a)] have generated substantial interest in the

Figure 1.4: Structure of phospholipid bilayer and apolipoproteins



(a) Phospholipid bilayer, a core part of the cellular membrane. Source: Wilkin D & Brainard J, 2016[13].

(b) Lipids forming an apolipoprotein. Source: Mabtech, 2013[14].

scientific community[15–19].

For triglycerides in particular, several large-scale prospective studies have demonstrated that levels of triglycerides are strongly and linearly associated with CHD[20,21], suggesting that triglycerides should be considered an independent risk factor for CHD[22]. However, the Emerging Risk Factors Collaboration (ERFC), in a meta-analysis of individual participant data from 68 long-term prospective studies, demonstrated that despite the association of triglycerides with CHD when adjusting for age and sex, the association is attenuated after adjusting for conventional cardiovascular risk factors such as systolic blood pressure, smoking status, and history of diabetes[23], which would suggest that the association of triglycerides with CHD is mediated by these other established risk factors. But in contrast to these findings, a subsequent study, examining a genetic variant that regulates triglyceride concentration, the *APOA5* gene promoter (-1131T>C, rs662799), found evidence that is consistent with a causal association between triglyceride-mediated pathways and risk of CHD[24]. Therefore, further research is of paramount importance to shed clarity on the nature of this association and the underlying mechanisms.

An analysis in the Pakistan Risk of Myocardial Infarction Study (PROMIS)[25] found that levels of major blood lipids—LDL-C, HDL-C, and triglycerides—are each strongly associated with CHD[26]. Several lipid-related genetic variants were found to be common to Pakistanis and Europeans, although they explained only a modest portion of the population variation in lipid concentration; additionally, the study found that allelic frequencies and effect sizes of lipid-related variants can differ between Pakistanis and Europeans[26]. Additional research on lipid metabolism, especially in the Pakistani population, would help elucidate these findings and further explain their significance.

It is now widely recognised that lowering levels of overall triglycerides is important for reducing CHD risk; indeed, the question is no longer *whether* we should lower triglycerides, but *how* we should lower triglycerides to reduce CHD risk. Evidence for the growing recognition of this important question can be discerned by the launch of several phase III clinical trials with the aim of reducing triglycerides in adults with severe hypertriglyceridemia. Two trials of omega-3 fatty acids are nearing completion: the REDUCE-IT trial by Amarin Pharma Inc. for the drug Vascepa, a highly purified ethyl ester of eicosapentaenoic acid (EPA), which will be completed in December 2017 (with results expected to be available before the end of Q3 2018)[27], and the STRENGTH trial by AstraZeneca for the drug Epanova, made up of omega-3 carboxylic acids, which will be completed in November 2019[28]. In addition, the PROMINENT trial by Kowa Pharmaceutical was announced in April 2016 for the drug PemaFibrate, which is a peroxisome proliferator-activated receptor (PPAR) alpha agonist, a key regulator of lipid and glucose metabolism that has been implicated in inflammation, which will be tested for the treatment of dyslipidaemia[29]. The substantial attention (and capital) invested by the pharmaceutical industry supports corresponding evidence in the scientific literature that the reduction of triglycerides is an important (and marketable) health concern.

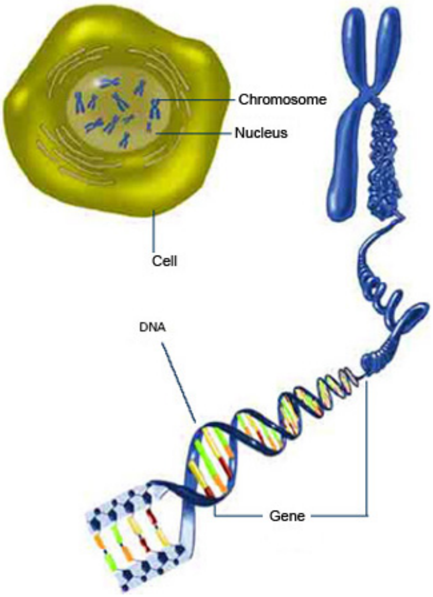
1.2 Genetics

1.2.1 Introduction to genetics

Deoxyribonucleic acid (DNA) consists of two long, twisted chains made up of nucleotides[30]. The bases in DNA nucleotides are adenine (A), thymine (T), cytosine (C) and guanine (G). Long strings of nucleotides form genes, and groups of genes are packaged tightly into structures called chromosomes[30]. A diagram portraying the structure of DNA is shown in Figure 1.5.

A gene is the basic unit of heredity and contains all of the information necessary to synthesise a protein. It is estimated that humans have between 20 000 to 23 000 genes[31]. The term genotype, then, refers to the genetic constitution of an individual, either overall or at a specific locus (i.e. location on a chromosome). All people have two copies of each chromosome, one from each parent. An allele refers to the existence of two or more alternative forms of a gene that are found at the same position on a chromosome, which arise by mutation (for example, A instead of G)[30]. Individuals are homozygous for a

Figure 1.5: Structure of DNA

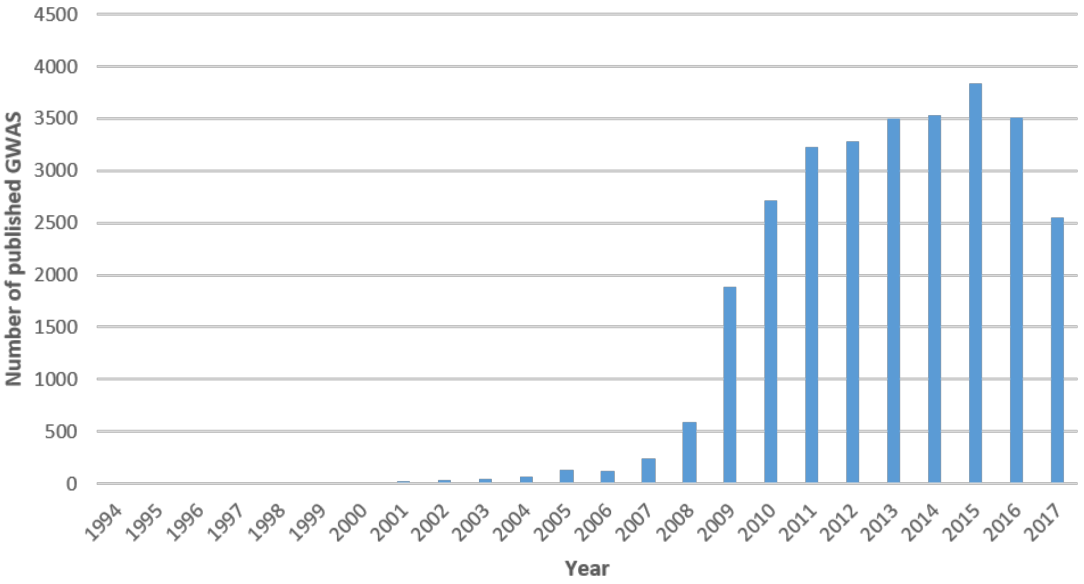


Deoxyribonucleic acid (DNA), which makes up genes, is spooled within chromosomes inside the nucleus of a cell. Source: National Institute of General Medical Sciences. *The New Genetics.* 2010[30].

given locus if they have two identical alleles at that locus, and heterozygous if they have two different alleles at that locus. Furthermore, if the alleles are identical, they can be homozygous for either the major allele or the minor allele, which gives a total of three possible genetic states for a given trait (i.e. AA, Aa, or aa, where A refers to the major allele and a refers to the minor allele).

A phenotype or trait is an observable (i.e. measurable) characteristic of a cell or organism. Phenotypes can be categorical, such as the presence or absence of a given disease, or continuous, such as the measurement of a biomarker (e.g. blood pressure or body mass index). Single nucleotide polymorphisms (SNPs) are small modifications in an individual nucleotide of the genome. Another important concept is linkage disequilibrium (LD), which is an assessment of the degree of correlation between nearby variants. Genetic loci are in LD with each other when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly. The theory behind LD is that every individual has a set of ancestral chromosomes, but eventually a mutation will occur followed by recombination. After many generations through which this mutation is propagated and new mutations occur, there will be a wide degree of variation between variants. However, since they all derive from a common ancestry, they should be mostly correlated. LD thus allows variants that have not been measured to be imputed from these haplotypes. The uncertainty is represented by assigning the probabilities of each of the three genotypes for each individual. Imputation is particularly

Figure 1.6: Plot showing number of genome-wide association studies published per year



A PubMed search was conducted (last updated on 19 September 2017) using the following search term: *("genome-wide association study") OR (GWAS)*. A summary of the number of published studies that were returned in each year was downloaded and used to produce this plot. Although some results were returned for studies published prior to the first GWAS in 2005, an examination of the titles and abstracts for these studies indicates that they only analysed at most a few hundred susceptibility loci that were already known to be associated with the trait or disease of interest. Therefore, these studies were not truly "genome-wide", which typically refers to the analysis of a trait or disease with millions of variants across the entire genome[33].

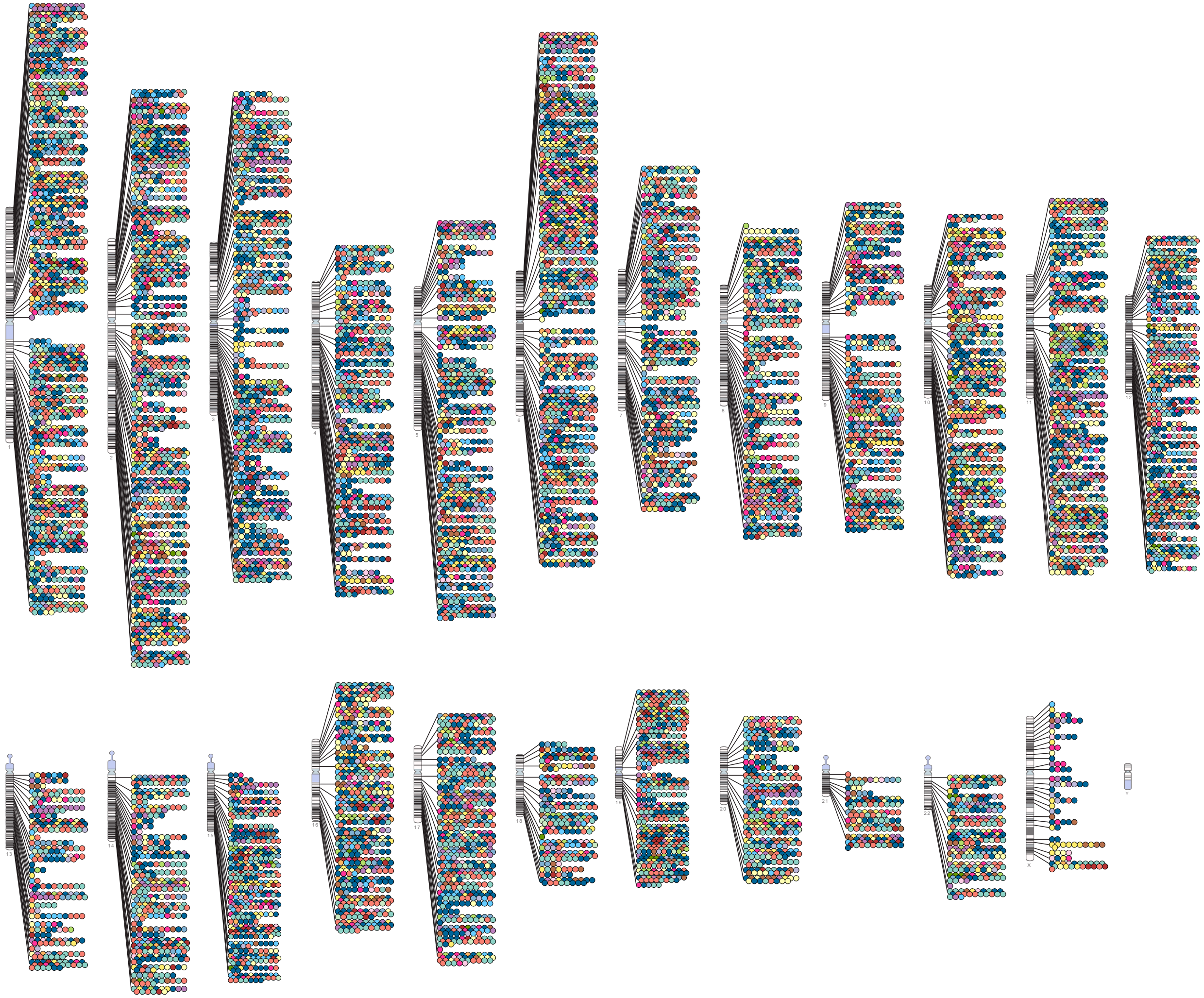
useful for common and low-frequency variants.

Rigorously conducted association study designs typically involve two stages. First, there is a discovery stage, where the goal is to capture as much common genomic variation as possible by LD. Since a large number of statistical tests are conducted, this necessitates using a stringent threshold for genome-wide significance. Second, there is a replication or follow-up stage, which aims to assess many fewer SNPs that represent the top signals from the discovery stage. A meta-analysis of the results from the two stages then leads to a subset of significant variants that can be ultimately reported.

The first genome-wide association study (GWAS) was published in 2005 in the journal *Science*. The study found that a variant in the *CFH* gene is strongly associated with age-related macular degeneration (with an odds ratio of 7.4) and was discovered using only 100 000 SNPs with just 96 cases and 50 controls[32]. There have been incredible advances in the pace of GWAS since then, and new associations are constantly being discovered (nearly 30 000 GWAS publications in the past 10 years. A plot showing the number of GWAS published each year is shown in Figure 1.6.

An overview of all reported SNP–trait associations has been compiled and is continuously updated by the NHGRI-EBI GWAS Catalog[34]. The most recent version at time of writing is shown in Figure 1.7, which includes 49 769 reported unique SNP–trait associations[34].

Figure 1.7: GWAS diagram of all SNP–trait associations



GWAS diagram of all SNP–trait associations, with P-values $\leq 5 \times 10^{-8}$, mapped onto the human genome by chromosomal locations and displayed on the human karyotype (downloaded 01-Sep-2017 from http://www.ebi.ac.uk/gwas/). Source: Welter D, et al. *Nucleic Acids Res.* 2014;42:D1001-6[34].

1.2.2 Association of genetic loci with major lipids

In order to better understand the role that genetics plays in affecting lipid levels, a number of association studies of major lipids have been conducted. One of the first large-scale meta-analyses of circulating lipids was published in 2010, which reported the discovery of 95 genetic loci significantly associated with plasma concentrations of total cholesterol, LDL-C, HDL-C, and triglycerides[19]. Of these loci, 59 were novel at the time. A subsequent large-scale meta-analysis resulted in the discovery of 62 additional novel loci, bringing the total number of loci associated with major lipid traits to 157[35]. At present, association studies[19,35–40] have uncovered 175 genetic loci that affect lipid levels in the population, which are listed in Table 1.1. Most of these variants reside in non-coding portions of the genome, where the precise function is often not well known.

1.2.3 Association of genetic loci with risk of CHD

Genetics is a useful tool to provide information about the heritability of diseases, including CHD. Due to the influence of inherited genes on expression of genes that regulate pathways for CHD risk factors, family history of CHD is a strong risk factor for CHD.

Over time, genetic collaborations have identified an increasing number of loci associated with CAD, starting with nine loci in 2007 with the Wellcome Trust Case Control Consortium (WTCCC)[41], and increasing to 34 distinct loci by 2011[42]. The CARDIoGRAMplusC4D Consortium then discovered additional loci in 2013, bringing the total number of susceptibility loci for CAD to 46, of which 12 showed a significant association with a lipid trait[43]. Today, using the latest information available at time of writing, 58 independent and replicated loci for CAD have been identified in European populations (see Table 1.2), and from joint association analyses the total heritability of CAD is estimated to be 28 %[10]. While these studies indicate that our genotype only explains a portion of CHD risk, they nevertheless help elucidate how our genetic make-up is linked to CHD.

One example of the important role that genetics can play in disease development is with hepatic glucokinase, which binds to glucokinase regulatory protein (GCKR) in the presence of fructose 6-phosphate to regulate glucose storage and disposal in the liver. A study revealed that carriers of the *GCKR*-L446 polymorphism were protected against type 2 diabetes (T2D) despite higher triglyceride levels and risk of dyslipidaemia, suggesting that genetic differences in certain individuals can improve their ability to regulate insulin and triglyceride levels[44].

Table 1.1: List of 175 genetic loci associated with major lipids

Gene	**Trait**	**rsid**	**Chr:Pos (GRCh37)**	**EA/NEA**
PIGV	HDL	rs12748152	chr1:27138393	T/C
PABPC4	HDL	rs4660293	chr1:40028180	G/A
RRNAD1	HDL	rs12145743	chr1:156700651	G/T
C1orf220	HDL	rs4650994	chr1:178515312	A/G
ZNF648	HDL	rs1689800	chr1:182168885	G/A
GALNT2	HDL	rs4846914	chr1:230295691	A/G
COBLL1	HDL	rs12328675	chr2:165540800	C/T
CPS1	HDL	rs1047891	chr2:211540507	A/C
PRKAG3	HDL	rs78058190	chr2:219699999	A/G
LOC646736	HDL	rs2972146	chr2:227100698	T/G
ATG7	HDL	rs2606736	chr3:11400249	T/C
NBEAL2	HDL	rs2290547	chr3:47061183	T/C
MON1A	HDL	rs2013208	chr3:50129399	A/C
STAB1	HDL	rs13326165	chr3:52532118	G/A
GSK3B	HDL	rs6805251	chr3:119560606	T/C
C4orf52	HDL	rs10019888	chr4:26062990	C/G
FAM13A	HDL	rs3822072	chr4:89741269	A/G
LOC100507053	HDL	rs2602836	chr4:100014805	G/A
SLC39A8	HDL	rs13107325	chr4:103188709	T/C
ARL15	HDL	rs6450176	chr5:53298025	A/G
RSPO3	HDL	rs1936800	chr6:127436064	T/C
LOC645434	HDL	rs605066	chr6:139829666	T/C
DAGLB	HDL	rs702485	chr7:6449272	G/A
SNX13	HDL	rs4142995	chr7:17919258	T/C
C7orf72	HDL	rs4917014	chr7:50305863	G/T
CD36	HDL	rs3211938	chr7:80300449	G/T
KLF14	HDL	rs4731702	chr7:130433384	T/C
TMEM176A	HDL	rs17173637	chr7:150529449	C/T
LOC157273	HDL	rs9987289	chr8:9183358	G/A
TRPS1	HDL	rs2293889	chr8:116599199	G/T
TTC39B	HDL	rs581080	chr9:15305378	C/G
ABCA1	HDL	rs1883025	chr9:107664301	T/C
MARCH8-ALOX5	HDL	rs970548	chr10:46013277	C/A
CAND1.11	HDL	rs2923084	chr11:10388782	G/A
F2	HDL	rs3136441	chr11:46743247	C/T
OR4C46	HDL	rs11246602	chr11:51512090	C/T
PCNXL3	HDL	rs12801636	chr11:65391317	A/G
MOGAT2	HDL	rs499974	chr11:75455021	A/C
LOC100506393	HDL	rs7134375	chr12:20473758	A/C
MMAB	HDL	rs7134594	chr12:110000193	T/C
SBNO1	HDL	rs4759375	chr12:123796238	T/C
ZNF664,ZNF664-FAM101A	HDL	rs4765127	chr12:124460167	T/G
SCARB1	HDL	rs838880	chr12:125261593	T/C
ZBTB42	HDL	rs4983559	chr14:105277209	A/G
AQP9	HDL	rs1532085	chr15:58683366	G/A
TPM1	HDL	rs2652834	chr15:63396867	G/A
RMI2	HDL	rs7188861	chr16:11454650	T/C
FTO	HDL	rs1121980	chr16:53809247	A/G
HERPUD1	HDL	rs3764261	chr16:56993324	A/C
PSKH1	HDL	rs16942887	chr16:67928042	A/G
CMIP	HDL	rs2925979	chr16:81534790	C/T
STARD3	HDL	rs11869286	chr17:37813856	C/G
CD300LG	HDL	rs72836561	chr17:41926126	T/C
ABCA8	HDL	rs4148008	chr17:66875294	A/G
PGS1	HDL	rs4129767	chr17:76403984	A/G
LIPG	HDL	rs7241918	chr18:47160953	T/G
PMAIP1	HDL	rs12967135	chr18:57849023	C/T
ANGPTL4	HDL	rs7255436	chr19:8433196	A/C
DOCK6	HDL	rs737337	chr19:11347493	C/T
FPR3	HDL	rs17695224	chr19:52324216	A/G
MIR4752	HDL	rs386000	chr19:54792761	C/G
HNF4A	HDL	rs1800961	chr20:43042364	T/C
PCIF1	HDL	rs6065906	chr20:44554015	C/T
UBE2L3	HDL	rs181362	chr22:21932068	T/C
PCSK9	LDL	rs2479409	chr1:55504650	A/G
CELSR2	LDL	rs629301	chr1:109818306	T/G
ANXA9	LDL	rs267733	chr1:150958836	G/A
APOB	LDL	rs1367117	chr2:21263900	A/G
ABCG8	LDL	rs4299376	chr2:44072576	T/G
EHBP1	LDL	rs2710642	chr2:63149557	G/A
CCDC93	LDL	rs10490626	chr2:118835841	A/G
LINC01101	LDL	rs2030746	chr2:121309488	T/C
FN1	LDL	rs1250229	chr2:216304384	C/T
CMTM6	LDL	rs7640978	chr3:32533010	T/C
DNAJC13	LDL	rs17404153	chr3:132163200	T/G
CSNK1G3	LDL	rs4530754	chr5:122855416	A/G
DTNBP1	LDL	rs3757354	chr6:16127407	T/C
HFE	LDL	rs1800562	chr6:26093141	A/G
SLC22A1	LDL	rs1564348	chr6:160578860	C/T
MIR148A	LDL	rs4722551	chr7:25991826	C/T
SOX17	LDL	rs10102164	chr8:55421614	A/G
PLEC	LDL	rs11136341	chr8:145043543	G/A
ABO	LDL	rs635634	chr9:136155000	T/C
ST3GAL4	LDL	rs11220462	chr11:126243952	A/G
BRCA2	LDL	rs4942486	chr13:32953388	C/T
NYNRIN	LDL	rs8017377	chr14:24883887	A/G
KPNB1	LDL	rs7206971	chr17:45425115	A/C
APOH	LDL	rs1801689	chr17:64210580	C/A
GATA6	LDL	rs79588679	chr18:19907770	T/C
LDLR	LDL	rs6511720	chr19:11202306	T/G
APOC1	LDL	rs4420638	chr19:45422946	G/A
ZNF274	LDL	rs117492019	chr19:58681861	T/G
LOC101929486	LDL	rs364585	chr20:12962718	G/A
BANF2	LDL	rs2328223	chr20:17845921	C/A

Table: List of 175 genetic loci associated with major lipids (. . . continued)

Gene	Trait	rsid	Chr:Pos (GRCh37)	EA/NEA
TOP1	LDL	rs6029526	chr20:39672618	A/T
MTMR3	LDL	rs5763662	chr22:30378703	T/C
ASAP3	TC	rs1077514	chr1:23766233	T/C
TMEM57	TC	rs12027135	chr1:25775733	T/A
EVI5	TC	rs7515577	chr1:93009438	A/C
FCGR2A	TC	rs1801274	chr1:161479745	G/A
MOSC1	TC	rs2642442	chr1:220973563	G/A
IRF2BP2	TC	rs514230	chr1:234858597	T/G
RAB3GAP1	TC	rs7570971	chr2:135837906	A/C
ABCB11	TC	rs2287623	chr2:169830155	A/G
FAM117B	TC	rs11694172	chr2:203532304	G/A
UGT1A1	TC	rs11563251	chr2:234679384	T/C
RAF1	TC	rs2290159	chr3:12628920	C/G
PXK	TC	rs13315871	chr3:58381287	A/G
ADAMTS3	TC	rs117087731	chr4:73696709	T/A
MTHFD2L	TC	rs182616603	chr4:75084732	T/C
HMGCR	TC	rs12916	chr5:74656539	T/C
TIMD4	TC	rs6882076	chr5:156390297	C/T
HLA-DRA	TC	rs3177928	chr6:32412435	A/G
SPDEF	TC	rs2814982	chr6:34546560	T/C
KCNK5	TC	rs2758886	chr6:39250837	A/G
FRK	TC	rs9488822	chr6:116312893	T/A
HBS1L	TC	rs9376090	chr6:135411228	C/T
C7orf50	TC	rs1997243	chr7:1083777	G/A
DNAH11	TC	rs12670798	chr7:21607352	C/T
NPC1L1	TC	rs2072183	chr7:44579180	C/G
UBXN2B	TC	rs2081687	chr8:59388565	C/T
VLDLR	TC	rs3780181	chr9:2640759	G/A
VIM-AS1	TC	rs10904908	chr10:17260290	G/A
ERLIN1	TC	rs11597086	chr10:101953705	C/T
GPAM	TC	rs2255141	chr10:113933886	G/A
SPTY2D1	TC	rs10128711	chr11:18632984	C/T
PHLDB1	TC	rs11603023	chr11:118486067	C/T
MIR100HG	TC	rs7941030	chr11:122522375	C/T
PHC1	TC	rs4883201	chr12:9082581	G/A
ATXN2	TC	rs11065987	chr12:112072424	G/A
HNF1A	TC	rs1169288	chr12:121416650	C/A
HPR	TC	rs2000999	chr16:72108093	A/G
ASGR1	TC	rs314253	chr17:7091650	C/T
SUGP1	TC	rs10401969	chr19:19407718	C/T
FUT2	TC	rs492602	chr19:49206417	G/A
C20orf173	TC	rs2277862	chr20:34152782	C/T
MAFB	TC	rs2902940	chr20:39091487	A/G
TOM1	TC	rs138777	chr22:35711098	G/A
PPARA	TC	rs4253772	chr22:46627603	T/C
DOCK7	TG	rs2131925	chr1:63025942	T/G
PROX1	TG	rs340839	chr1:214161820	C/T
GCKR	TG	rs1260326	chr2:27730940	C/T
CEP68	TG	rs2540948	chr2:65284623	C/T
MSL2	TG	rs645040	chr3:135926622	T/G
DOK7	TG	rs6831256	chr4:3473139	G/A
AFF1	TG	rs442177	chr4:88030261	T/G
LOC101928448	TG	rs9686661	chr5:55861786	T/C
HLA-C	TG	rs2247056	chr6:31265490	C/T
VEGFA	TG	rs998584	chr6:43757896	A/C
TYW1B	TG	rs13238203	chr7:72129667	T/C
TBL2	TG	rs17145738	chr7:72982874	T/C
GPR85	TG	rs2255811	chr7:112722196	G/A
MET	TG	rs38855	chr7:116358044	G/A
PINX1	TG	rs11776767	chr8:10683929	C/G
NAT2	TG	rs1495741	chr8:18272881	A/G
LPL	TG	rs12678919	chr8:19844222	G/A
TRIB1	TG	rs2954029	chr8:126490972	T/A
AKR1C4	TG	rs1832007	chr10:5254847	G/A
JMJD1C	TG	rs10761731	chr10:65027610	T/A
CYP26A1	TG	rs2068888	chr10:94839642	A/G
FADS1	TG	rs174546	chr11:61569830	T/C
ZPR1	TG	rs964184	chr11:116648917	C/G
R3HDM2	TG	rs11613352	chr12:57792580	T/C
CAPN3	TG	rs2412710	chr15:42683787	A/G
FRMD5	TG	rs2929282	chr15:44245931	T/A
PDXDC1	TG	rs3198697	chr16:15129940	T/C
STX4	TG	rs11649653	chr16:30918487	T/C
SERPINF2	TG	rs2070863	chr17:1648502	T/C
TM4SF5	TG	rs193042029	chr17:4667984	G/T
MPP3	TG	rs8077889	chr17:41878166	C/A
INSR	TG	rs7248104	chr19:7224431	A/G
PEPD	TG	rs731839	chr19:33899065	A/G
COL18A1	TG	rs114139997	chr21:46875775	A/G
PLA2G6	TG	rs5756931	chr22:38546033	C/T
PNPLA3	TG	rs738409	chr22:44324727	G/C

List of 175 genetic loci that have been identified by association studies (as of April 2017) that affect lipid levels in the population. **Abbreviations:** **EA** = Effect allele; **GRCh37** = Genome Reference Consortium human genome (build 37); **HDL** = High-density lipoprotein cholesterol; **LDL** = Low-density lipoprotein cholesterol; **NEA** = Non-effect allele; **TC** = Total cholesterol; **TG** = Triglycerides.

Table 1.2: Replicated genome-wide significant loci for CAD

Chr	Closest Gene(s)	Putative Functions of Possible Relevance to CAD	Lead SNP	EAF	OR
1	PPAP2B	Regulation of cell–cell interactions	rs17114036	0.92	1.13
1	PCSK9*	Regulation of LDL receptor recycling	rs11206510	0.85	1.08
1	SORT1	Regulate apoB secretion and LDL catabolism	rs599839	0.78	1.11
1	IL6R	IL-6 receptor, immune response	rs4845625	0.46	1.05
1	MIA3	Collagen secretion	rs17465637	0.73	1.08
2	LINC00954	LncRNA of unknown function	rs16986953	0.07	1.09
2	APOB*	Major apolipoprotein of LDL	rs515135	0.78	1.07
2	ABCG5/G8*	Cholesterol absorption and secretion	rs6544713	0.29	1.05
2	VAMP5/8-GGCX	Intracellular vesicle trafficking	rs1561198	0.47	1.06
2	ZEB2-AC074093.1	ZEB2: Transcriptional repressor	rs2252641	0.44	1.03
2	WDR12	Component of nucleolar protein complex	rs6725887	0.14	1.14
3	MRAS	Cell growth and differentiation	rs9818870	0.15	1.07
4	EDNRA	Receptor for endothelin—vasoconstriction	rs1878406	0.16	1.06
4	GUCY1A3	Nitric oxide signaling	rs7692387	0.80	1.07
4	REST-NOA1	REST maintains VSMCs in a quiescent state	rs17087335	0.21	1.06
5	SLC22A4/A5	Organic cation transporter	rs273909	0.14	1.06
6	ANKS1A	May inhibit PDGF-induced mitogenesis	rs17609940	0.82	1.03
6	PHACTR1	Regulates protein phosphatase 1 activity	rs12526453	0.71	1.10
6	KCNK5*	Potassium channel protein	rs10947789	0.78	1.05
6	TCF21	Transcriptional regulator	rs12190287	0.64	1.06
6	SLC22A3-LPAL2-LPA	Lipoprotein(a)	rs2048327 rs3789220	0.35 0.02	1.06 1.42
6	PLG	Fibrinolysis	rs4252120	0.74	1.03
7	NOS3	Production of nitric oxide	rs3918226	0.06	1.14
7	HDAC9	Represses MEF2 activity/beige adipogenesis	rs2023938	0.10	1.08
7	ZC3HC1	Encodes NIPA, regulator of cell proliferation	rs11556924	0.69	1.08
8	LPL*	Lipolysis of TG-rich lipoproteins	rs264	0.85	1.06
8	TRIB1*	TG, MAPK signaling, SMC proliferation	rs2954069	0.55	1.04
9	CDKN2BAS	Cellular proliferation, platelet function	rs10757274	0.48	1.21
9	ABO*	IL-6, E-selectin, LDL-C levels	rs579459	0.21	1.08
10	KIAA1462	Component of endothelial cell–cell junctions	rs2505083	0.40	1.07
10	CXCL12	Endothelial regeneration; neutrophil migration	rs501120 rs2047009	0.81 0.48	1.08 1.06
10	LIPA	Intracellular hydrolysis of cholesteryl esters	rs1412444 rs11203042	0.37 0.45	1.07 1.04
10	CYP17A1-CNNM2-NT5C2	CYP17A1: Steroidogenic pathway	rs12413409	0.89	1.08
11	PDGFD	Role in SMC proliferation	rs974819	0.33	1.07
11	SWAP70	Leukocyte and VSMC migration and adhesion	rs10840293	0.55	1.06
11	ZNF259 APOA5 APOC3	TG-rich lipoprotein metabolism	rs964184	0.18	1.05
12	SH2B3	Negative regulator of cytokine signaling	rs3184504	0.42	1.07
12	ATP2B1	Intracellular calcium homeostasis	rs7136259	0.43	1.04
12	KSR2	Suppressor of Ras2–cell proliferation; obesity	rs11830157	0.36	1.12†
13	FLT1	VEGFR family; angiogenesis	rs9319428	0.31	1.04
13	COL4A1/A2	Type IV collagen chain of basement membrane	rs4773144 rs9515203	0.43 0.76	1.05 1.07
14	HHIPL1	Unknown	rs2895811	0.41	1.04
15	ADAMTS7	Proliferative response to vascular injury	rs7173743	0.56	1.08
15	SMAD3	Downstream mediator of TGF-β signaling	rs56062135	0.79	1.07
15	MFGE8-ABHD2	MFGE8: Lactadherin–VEGF neovascularization	rs8042271	0.90	1.10
15	FURIN	Endoprotease—TGF-β1 precursor and type I MMP	rs17514846	0.44	1.05
17	BCAS3	Rudhira—EC polarity and angiogenesis	rs7212798	0.15	1.08
17	RAI1-PEMT-RASD1	PEMT encoded protein converts PE to PC	rs12936587	0.61	1.03
17	SMG6	Role in nonsense mediated RNA decay	rs216172	0.35	1.05
17	UBE2Z	Protein ubiquination; apoptosis	rs46522	0.51	1.04
18	PMAIP1-MC4R*	PMAIP1: HIF1A-induced proapoptotic gene; MC4R: Leptin signaling—obesity	rs663129	0.26	1.06
19	LDLR*	LDL clearance	rs1122608	0.77	1.08
19	APOE	LDL and VLDL clearance	rs4420638	0.17	1.10
19	ZNF507	Unknown	rs12976411	0.09	0.67†
21	KCNE2	Maintains cardiac electric stability	rs9982601	0.13	1.12
22	POM121L9P-ADORA2A	Adenosine A2a receptor: infarct-sparing effects	rs180803	0.97	1.20

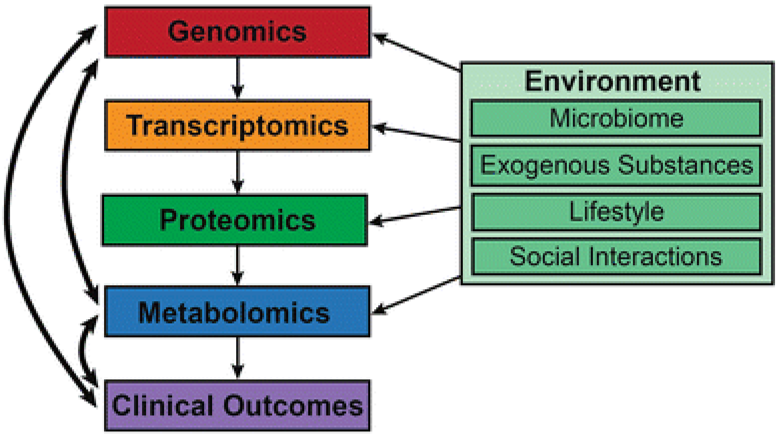
List of independent and replicated loci for CAD based on published studies of CAD (as of April 2017). *Locus is also associated with major lipids (see Table 1.1). †By recessive model. **Abbreviations:** **CAD** = Coronary artery disease; **Chr** = Chromosome; **EAF** = Effect allele frequency; **EC** = Endothelial cells; **HIF1A** = Hypoxia inducible factor 1A; **IL** = Interleukin; **LDL** = Low-density lipoprotein; **LDL-C** = Low-density lipoprotein cholesterol; **LncRNA** = Long noncoding RNA; **MAPK** = Mitogen-activated protein kinase; **MEF2** = Myocyte enhancer factor 2; **MMP** = Matrix metalloproteinase; **NIPA** = Nuclear interacting partner of anaplastic lymphoma kinase; **OR** = Odds ratio; **PC** = Phosphatidylcholine; **PDGF** = Platelet-derived growth factor; **PE** = Phosphatidylethanolamine; **REST** = RE-1 silencing transcription factor; **SMC** = Smooth muscle cell; **SNP** = Single-nucleotide polymorphism; **TG** = Triglyceride; **TGF-β** = Transforming growth factor-β; **VEGFR** = Vascular endothelial growth factor receptor; **VLDL** = Very low-density lipoprotein; **VSMC** = Vascular smooth muscle cells.

1.3 Metabolomics

1.3.1 Introduction to metabolomics

Metabolomics attempts to capture the complexity of metabolic networks by simultaneously studying a range of metabolic markers, called metabolites, within biological fluids, cells, and tissues[45]. Measurement of metabolites can provide a direct reflection of the physiological state, making them an ideal method of tracking changes induced by disease or treatment.

Figure 1.8: Role of genomics, metabolomics, and other types of "–omics" data in informing clinical outcomes



Genomics contributes to variations at the level of the transcriptome, proteome, and metabolome, which influences clinical outcomes. In addition, environmental factors can influence all four "–omics" levels leading to a clinical phenotype. The double-ended arrows indicate the interaction of metabolites with clinical outcome measures and genomic variants, which could reveal novel pathways associated with clinical phenotypes, which can then be validated by functional genomic studies and by investigating the interaction of those genetic variants with clinical outcomes. Source: Neavin D, et al. *Metabolomics*. 2016;12(7):1-6[47].

Metabolomics can also be used to identify specific metabolic phenotypes that are associated with given genetic modifications[45]. Since metabolites are closer in proximity to clinical outcomes than proteins or genes, they contain more information on the health status of individuals compared to other "–omics" technologies[46]. The relationship between genomics, metabolomics, and other types of "–omics" data, and how this information can inform clinical outcomes and the identification of novel metabolic pathways, is shown in Figure 1.8.

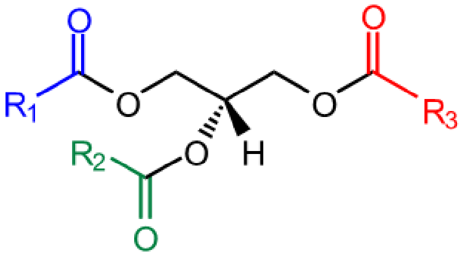
Lipidomics, a subset of metabolomics concerned with the study of lipid profiles, involves profiling a biological sample to yield information on the composition and abundance of lipid subfractions in the body. While lipidomics technically falls under the umbrella of metabolomics, it can also be viewed as a stand-alone field due to the uniqueness of lipids and their specific functions relative to other metabolites[48]. The field has seen recent technological advances, particularly in mass spectrometry[49,50] and data processing[51]. The various types of platforms and processing methods used for lipidomics will be described in Section 3.1. There are currently more than 30 000 unique identifiable lipid species[52]. Lipidomic analysis of human blood has the potential to uncover novel biomarkers and identify the role of specific lipids in diseases, including CVD[53], cancer[54], and neurodegenerative disorders[55]. Lipidomics can also be used to identify dietary patterns and objectively assess adherence to dietary programmes aimed at reducing obesity and dyslipidaemia[56].

Although standard lipid biomarkers remain a fundamental part of everyday clinical practice, lipidomics takes a more global view of lipid metabolism and can provide a detailed picture of abnormalities in lipid levels, in contrast to the measurement of isolated lipoproteins[11]. Lipidomics can be used to describe specific lipids involved in dynamic physiological changes and characterise any abnormalities in lipid metabolism that impact disease aetiology[11]. Studies of the hundreds of different types of lipids that make up these broad overall lipid classes have the ability to explore the underlying lipid pathways and the association with CHD in much greater detail, and can lead to the identification of new therapeutic targets and novel therapeutic agents[11,57–59].

Lipid species share common characteristics as they are all made up of fatty acids attached to a backbone. Lipid classes are defined based on a characteristic head group within the backbone, while the diversity of lipid species within each lipid class derives from the various combinations of fatty acids, which can vary in several characteristics, including the length of each carbon chain, the number of double bonds on each chain, the position of the double bonds, and the configuration (*cis* or *trans*) of the double bonds[48,60].

To return to the specific triglyceride example mentioned earlier, most studies of triglycerides have treated triglycerides as a single compound, whereas in reality triglycerides are composed of three acyl chains esterified to a glycerol backbone and can take on a broad range of molecular weights[61]. The generic structure of a triglyceride molecule is shown in Figure 1.9. Together the diverse range of individual triglyceride metabolites make up the overall lipid class for triglycerides that is measured by a standard clinical chemistry assay; thus, through lipidomics the individual subtypes of triglycerides can be explored in much greater detail. A study of triglyceride metabolites found that triglycerides with a fewer number of acyl chain carbons and fewer acyl chain double bonds were associated with an increased risk of T2D, whereas triglycerides with a higher number of acyl chain carbons and more double bonds were associated with a decreased risk of T2D, even after adjustment for age, sex, body mass index (BMI), fasting glucose, fasting insulin, total triglycerides, and HDL-C[61]. In addition to T2D, the nature of the association of triglycerides with CHD may depend on their structure—it is likely that some subtypes of triglycerides are associated with increased risk of CHD while others are associated with decreased risk, but the metabolic pathways involved are currently not well understood. Therefore, a more thorough characterisation of triglyceride-mediated pathways and the associated risks of various subtypes of triglycerides would contribute to a better understanding of the nature of the association of triglycerides with risk of CHD. In addition to triglycerides, other lipid species may also have different patterns of risk depending on their structure and the metabolic pathways involved.

Figure 1.9: Generic structure of a triglyceride molecule



A triglyceride molecule consists of a glycerol backbone with three fatty acid chains (labelled in this diagram as R_1, R_2, and R_3). The number of carbon atoms and double bonds on each chain can differ. Each unique combination of carbon atoms and double bonds on the three chains results in a different type of triglyceride. Source: Adapted from Quehenberger O, et al. *J Lipid Res*, 2010;51(11):3299-3305[62].

1.3.2 Association of metabolites with risk of CHD

Clinicians currently use a relatively narrow set of blood chemistry analytes to assess health and disease states, such as measuring glucose to monitor diabetes, circulating lipid levels—cholesterol, LDL-C, HDL-C, and triglycerides—to assess cardiovascular health, and markers such as creatinine to diagnose renal disorders. However, particularly in the case of major lipids, these measures represent the sum of a complex mixture of molecular species present in the lipoprotein particles, and although they capture the bulk components of the lipodome and provide modest insight into the relative distributions of lipoproteins, they do not capture the full complexity of the lipidome[12]. A partial map of metabolic pathways between lipid subclasses is shown in Figure 1.10, which clearly illustrates the complexity of lipid pathways. Metabolomics offers the ability to reveal a far more comprehensive metabolic profile for individuals or patients, and lipid metabolites can offer improved CVD risk prediction beyond that of traditional risk factors[53]. Given that many cardiovascular pathologies have an underlying metabolic basis, metabolomics can reasonably be used to estimate the relative risk of patients, understand pathophysiological mechanisms, and monitor treatment progress[45]. Metabolomics and lipidomics are also expected to play an important role in identifying and characterising disease states and in cardiometabolic drug development[52]. The authors of a recent white paper indicated that including metabolomics data in precision medicine initiatives is "vital and necessary"[63], with the rationale that future metabolic signatures will:

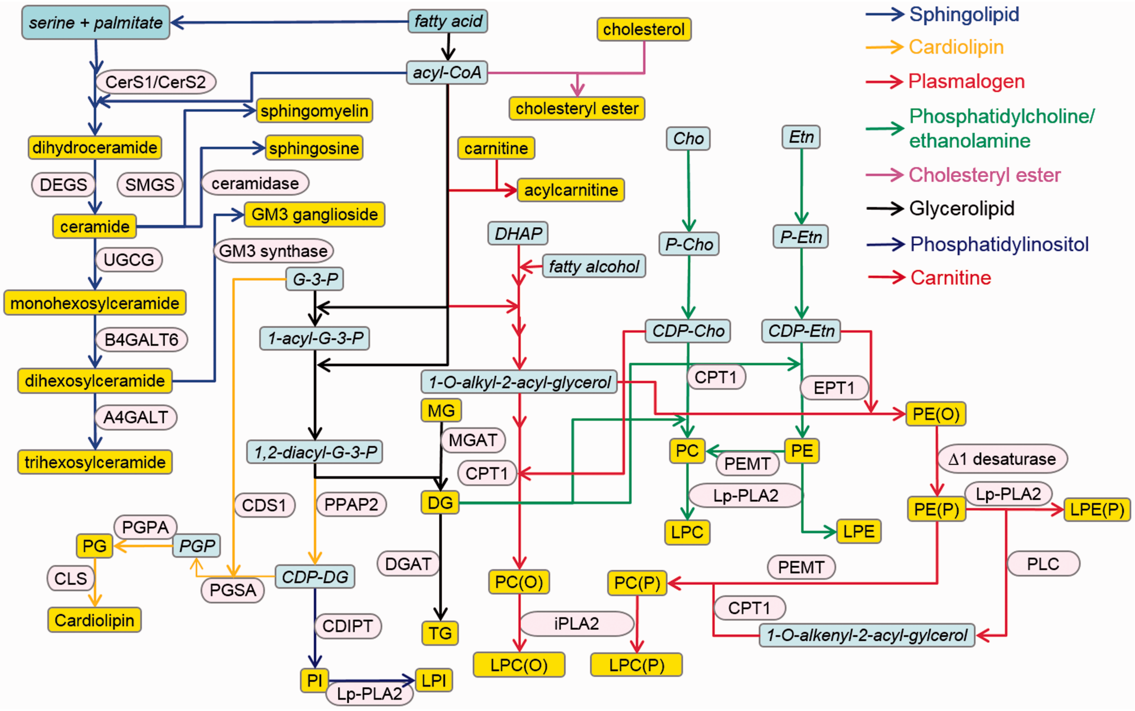
> *(1) provide predictive, prognostic, diagnostic, and surrogate markers of diverse disease states; (2) inform on underlying molecular mechanisms of diseases; (3) allow for sub-classification of diseases, and stratification of patients based on metabolic pathways impacted; (4) reveal biomarkers for drug response phen-*

> *otypes, providing an effective means to predict variation in a subject's response to treatment (pharmacometabolomics); (5) define a metabotype for each specific genotype, offering a functional read-out for genetic variants; (6) provide a means to monitor response and recurrence of diseases, such as cancers; and (7) describe the molecular landscape in human performance applications and extreme environments*[63].

Given the large scope and potential for metabolomics in informing CHD risk and numerous other diseases, a deeper understanding of the metabolic pathways involved and the genetic determinants of these metabolites is needed.

A number of studies have examined the association of metabolites with risk of CHD and related traits, such as diabetes and hypertension. To give several examples, a nested case-control study using samples from two large prospective cohorts found that the metabolite 2-aminoadipic acid was associated with risk of developing diabetes, indicating that this metabolite could be an important biomarker for diabetes risk and a potential modulator of glucose homoeostasis[64]. Another study found significant differences in individuals with impaired fasting glucose and T2D compared with healthy controls in levels of multiple metabolites, including α-hydroxybutyrate, alanine, proline, phenylalanine, glutamine, several branched-chain amino acids, several low-carbon number lipids, pyroglutamic acid, glycerophospholipids, and sphingomyelins[65]. Additionally, a nested case-cohort study within the EPIC study found that odd-chain (C15:0 [pentadecanoic acid] and C17:0 [heptadecanoic acid]) and long-chain (C20:0 [arachidic acid], C22:0 [behenic acid], C23:0 [tricosanoic acid], and C24:0 [lignoceric acid]) saturated fatty acids were negatively associated with incident T2D, while short-chain saturated fatty acids (C14:0 [myristic acid], C16:0 [palmitic acid], and C18:0 [stearic acid]) were positively associated with incident T2D[66]. Another analysis using the same cohort found that among long-chain ω-3 polyunsaturated fatty acids (PUFAs), α-linoleic acid (ALA) was inversely associated with T2D, but eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) were not significantly associated; among ω-6 PUFAs, linoleic acid (LA) and eicosadienoic acid (EDA) were inversely associated with T2D[67]. Other studies have identified increased levels of certain metabolites in individuals with hypertension compared with controls, as described in a review by Nikolic et al[68]. Regarding CHD itself, a lipidomics analysis on a subcohort of the prospective population-based Malmö Diet and Cancer study found that reduced levels of several species of lysophosphatidylcholines and triglycerides and increased levels of a sphingomyelin species were associated with incident CVD[69]. Another analysis using metabolomics data from several

Figure 1.10: Map of lipid subclass metabolic pathways



Partial metabolic map of lipid subclasses commonly measured in lipidomics studies. Source: Mundra PA, Shaw JE & Meikle PJ. *Int J Epidemiol.* 2016;45(5):1329-1338[12].

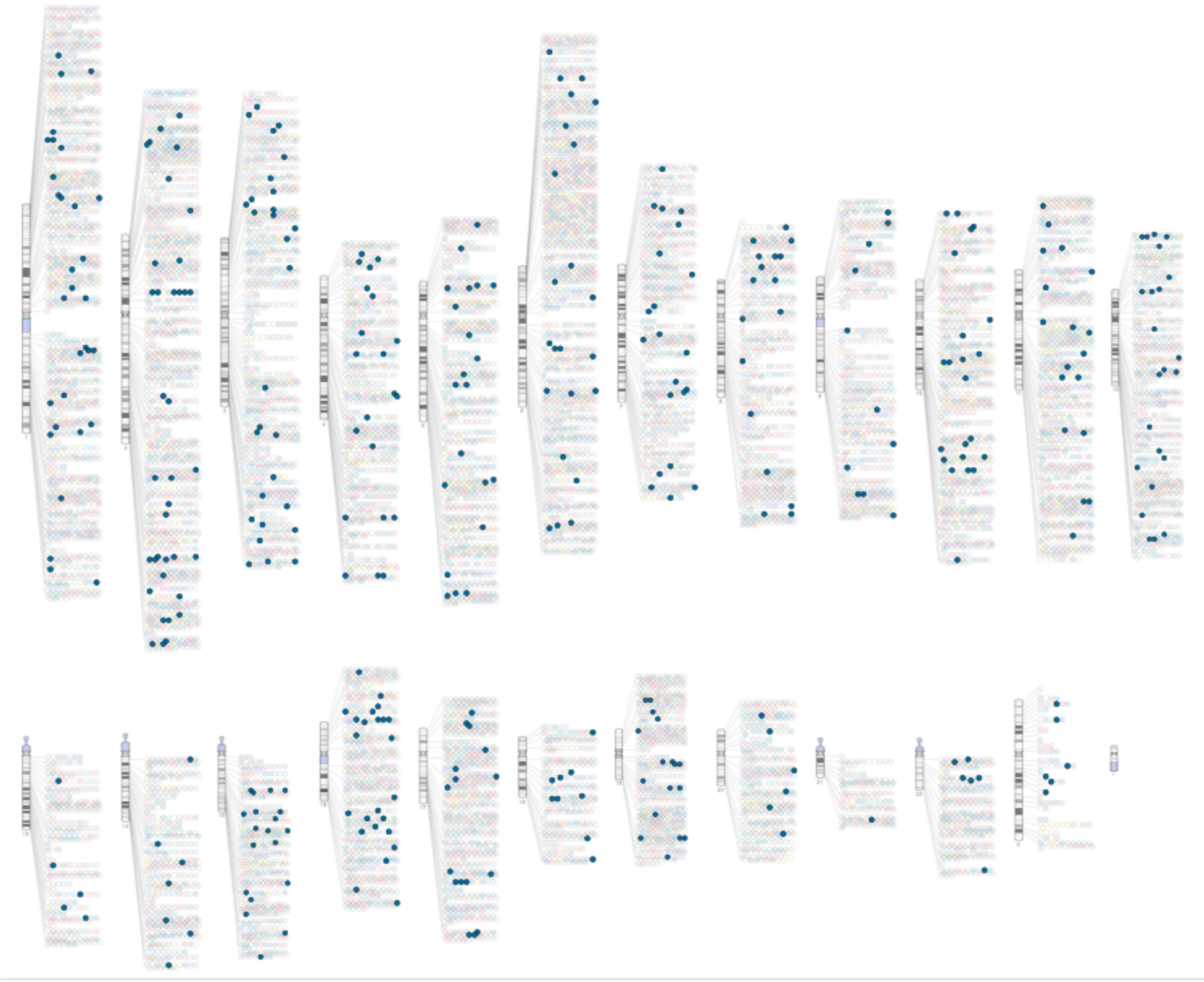
population-based cohorts found that certain species of lysophosphatidylcholines and sphingomyelins were negatively associated with incident CVD, while a monoacylglycerol species was positively associated with incident CVD[70]. Additionally, a prospective population-based lipidomics study found that species of cholesteryl esters, lysophosphatidylcholines, phosphatidylcholines, phosphatidylethanolamines, sphingomyelins, and triglycerides were associated with risk of CVD[53].

1.4 Genetic determinants of metabolites

1.4.1 Published genetic associations with metabolites

Out of the 49 769 reported unique SNP–trait associations from the GWAS diagram shown previously, 390 of these associations involve metabolite measurements as traits. Figure 1.11 shows associations with metabolites highlighted in blue, with all other traits greyed out. The figure clearly demonstrates that various metabolites are associated with a wide range of SNPs across the entire genome. This is important because it shows that not only is there an enormous diversity of metabolites themselves, but they are also influenced by a wide range of genetic determinants and they affect many different metabolic pathways.

Figure 1.11: GWAS diagram of all SNP–trait associations with metabolite measurements highlighted



GWAS diagram of all SNP–trait associations, with P-values $\leq 5 \times 10^{-8}$, mapped onto the human genome by chromosomal locations and displayed on the human karyotype, with 390 SNP–trait associations for "metabolite measurements" highlighted (downloaded 01-Sep-2017 from http://www.ebi.ac.uk/gwas/). Source: Welter D, et al. *Nucleic Acids Res.* 2014;42:D1001-6[34].

1.4.2 Literature review of published mGWAS

A metabolite-based genome-wide association study (mGWAS) is defined as a GWAS where metabolic traits are used as the phenotypic traits[71]. This is in contrast to a metabolome-wide association study (MWAS), which investigates associations between metabolic phenotypes and disease risk so it can be used to identify disease-related biomarkers[71]. The first-ever mGWAS investigating associations between genetic variants and metabolite profiles, published in 2008, involved quantitative measurement of 363 metabolites in serum samples from 284 male participants in the KORA study[72]. Although four loci associated with metabolites were discovered, the sample size was quite limited and there was no replication data. However, a follow-up study was published in 2010 using a much larger sample from the same population (1809 participants) along with a replication cohort of 422 participants from TwinsUK, which resulted in the discovery of eight replicated loci associated with metabolites[73]. Building on this work, a study published in 2011 involving 1768 participants from KORA and 1052 participants from TwinsUK found significant associations of metabolites with 37 loci. Meanwhile, several other studies were published over this time period based on different populations, and lately the number of published mGWAS has been continuing to steadily rise.

A recent review of mGWAS published in 2015 identified 21 such publications[74]. However, this review missed several important studies and numerous additional studies have since been published[75–84]. In theory, one could say that any study involving metabolic phenotyping using mass spectrometry (MS) or nuclear magnetic resonance (NMR) should be included as an mGWAS. Indeed, a search for "metabolite measurement" using Experimental Factor Ontology (EFO) (http://www.ebi.ac.uk/efo/) search terms identified 168 studies. However, this would end up including studies that only involve a single trait measured using MS (e.g. vitamin D or Bisphenol A [BPA]), or that measured substances not typically assayed by large-scale metabolomics platforms such as heavy metals or trace minerals (e.g. copper, selenium, zinc, arsenic, or lead). While it is difficult to provide a precise definition of mGWAS, the literature review used in this dissertation restricted the definition of mGWAS so that it only included studies of high-dimensional metabolomics, i.e. studies that measured a wide-variety of metabolic traits that are involved in human metabolism. Therefore, some studies that measured traits such as heavy metals and trace minerals were excluded if they only measured a handful of metabolites, while metabolites classed as xenobiotics (foreign chemical substances in the body such as drugs) were included if they were measured by a high-dimensional metabolomics platform that also assayed hundreds of other metabolites. Despite this inconsistency, this definition narrowed the focus of the literature review to specific types of high-dimensional metabolomics studies, rather than including studies that measured any possible trait using MS or NMR. The metabolites that were measured by these mGWAS studies, therefore, included all lipid-related traits, apolipoproteins, amino acids, ketone bodies, glycolysis-related metabolites, carbohydrates, cofactors and vitamins, energy-related metabolites, nucleotides, peptides, and xenobiotics. The review paper was used to develop an initial list of references, and additional studies[75–84] were identified through further scanning of the literature. April 2017 was used as the cut-off date, so any mGWAS studies published on or before this date were included. Although this approach was not technically a systematic review, it was still quite thorough and comprehensive. Supplementary data from each of the published studies were downloaded (except for a few studies where the data were unavailable), and the files were compiled into a single database in a consistent format that contained information such as the name of the study, trait, SNP, chromosome, position, effect allele, non-effect allele, effect size, standard error, and P-value. In total, 31 published mGWAS were identified through the literature review, which are listed in Table 1.3. A plot of the cumulative number of mGWAS published per year is shown in Figure 1.12.

While all of the mGWAS studies used either MS or NMR (58 % used MS, 35 % used NMR, and 6 % used both), a variety of different metabolomics platforms were employed to obtain the metabolite measurements. Biocrates (typically the AbsoluteIDQ) was used by five studies (16 %), Bruker (usually the Biospin Avance) was used by eight studies (26 %), Metabolon (usually GC-MS, UHPLC-MS/MS^2, and/or LC-MS) was used by six studies (19 %), and four studies (13 %) used multiple metabolomics platforms. The mGWAS studies also measured the samples in a range of different fluids: 17 (55 %) used serum, six (19 %) used plasma, four (13 %) used urine samples, and the remaining four studies used either whole blood or a combination of different biofluids. Each study measured a median of 217 metabolites, and out of the 13 studies that also analysed ratios of metabolites, a median of 15 180 ratios were used. The metabolites measured included amino acids, acylcarnitines, phospholipids, sphingolipids, fatty acids, and NMR peaks. The median number of participants in the discovery samples was 1960, and out of the nine studies that included a replication cohort, there was a median of 923 participants in the replication sample. Twenty-three (74 %) of the studies involved European participants and six (19 %) involved study participants from North America.

Table 1.3: Summary of mGWAS studies

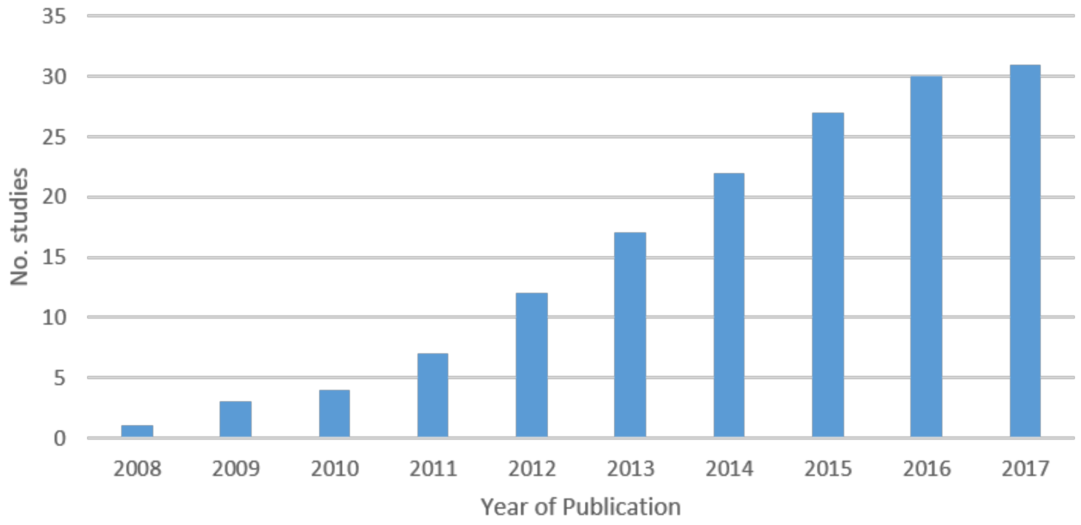
First author	Year	Journal	PMID	Cohort(s)	Metabolomics Method(s)	Metabolomics Platform(s)	Biofluid(s)	Metabolic trait(s)	No. metabolites	Study population(s)	No. participants	No. loci	No. SNPs	No. unique SNPs
Burkhardt	2015	*PLoS Genet*	26401656	LIFE Leipzig Heart Study, Sorb Study	MS	Applied Biosystems	Blood	Amino acids, acylcarnitines	62 + 34 ratios	German	2 107 + 923	16	4 522	876
Chasman	2009	*PLoS Genet*	19936222	Women's Genome Health Study (WGHS)	NMR	LipoProtein-II	Plasma	NMR-based lipoprotein fractions	17	North American (European ancestry)	17 296	43	668	71
Demirkan	2012	*PLoS Genet*	22359512	ERF, MICROS, NSPHS, ORCADES, VIS	MS	Micromass	Plasma	Phospholipids & sphingolipids	153 + ratios	European	4 034	35	1 122	35
Demirkan	2015	*PLoS Genet*	25569235	Erasmus Rucphen Family (ERF)	NMR	Bruker BioSpin Avance II	Serum	NMR-derived metabolites	42	European	2 118	8	241	241
Draisma	2015	*Nat Commun*	26068415	TwinsUK, KORA, EGCUT, LLS, QIMR, ERF, NTR	MS	Biocrates AbsoluteIDQ p150	Serum	Mainly phospholipids	129	European	7 478 + 1 182	31	123	59
Gieger	2008	*PLoS Genet*	19043545	KORA	MS	Biocrates	Serum	Targeted MS	363 + 131 769 ratios	German	284	4	30 432	2 906
Hartiala	2016	*Nat Commun*	26822151	Cleveland Clinic GeneBank Study	MS	Applied Biosystems	Plasma	Betaine	1	American	8 668	2	6	6
Hicks	2009	*PLoS Genet*	19798445	ERF, MICROS, NSPHS, ORCADES, VIS	MS	Micromass	Plasma & serum	Sphingolipids	33 + 43 ratios	European	4 400	5	483	53
Hong	2013	*Hum Mutat*	23281178	CAPS	MS	Proteomics & Metabolomics Facility at Colorado State University	Serum	MS peaks	6 138	Swedish	402 + 489	7	15 825	3 644
Illig	2010	*Nat Genet*	20037589	KORA, TwinsUK	MS	Biocrates AbsoluteIDQ	Serum	Mainly phospholipids	163 + 26 406 ratios	German, British	1 809 + 422	9	9 358	4 767
Inouye	2012	*PLoS Genet*	22916037	Young Finns Study (YFS), Northern Finland Birth Court 1966 (NFBC66)	NMR	Bruker Biospin Avance III	Serum	Mainly lipid traits & low-weight metabolites	130	Finnish, British	1 905 + 4 703	34	34	34
Kettunen	2012	*Nat Genet*	22286219	YFS, NFBC1966, HBCS, GenMets, DILGOM, Twins	NMR	Bruker Biospin Avance III	Serum	Mainly lipid traits	117 + 99 ratios	Finnish	8 330	31	14 173	1 034
Kettunen	2016	*Nat Commun*	27005778	14 cohorts from Europe	NMR	Bruker Biospin Avance III	Blood	Lipoprotein lipids & subclasses, fatty acids, amino acids, glycolysis precursors	123	European	24 925	62	8	8
Kraus	2015	*PLoS Genet*	26540294	CATHGEN	MS	Waters Corporation	Plasma	FIA-MS derived, mainly acylcarnitines and amino acids	63	American	1 490 + 2 022	6	50	32
Krumsiek	2012	*PLoS Genet*	23093944	KORA	MS	Metabolon UHPLC-MS-MS2 & GC-MS	Serum	Non-targeted MS (unknowns)	517	German	1 768	34	948	474
Long	2017	*Nat Genet*	28263315	TwinsUK	MS	Metabolon UHPLC-MS-MS2	Serum	Non-targeted MS	644	British	1 960	101	198 279	128 007
Montoliu	2013	*Genes Nutr*	23065485	São Paulo Brazil general population	NMR	Biocrates AbsoluteIDQ	Urine	NMR peaks	2 425	Brazilian	265	2	0	0

Table: Summary of mGWAS studies (. . . continued)

First author	Year	Journal	PMID	Cohort(s)	Metabolomics Method(s)	Metabolomics Platform(s)	Bio-fluid(s)	Metabolic trait(s)	No. metabolites	Study population(s)	No. participants	No. loci	No. SNPs	No. unique SNPs
Nicholson	2011	*PLoS Genet*	21931564	MolTWIN, MolObb	NMR & MS	Biocrates AbsoluteIDQ & Bruker BioSpin	Urine & plasma	Urine: NMR peaks; Plasma: mainly phospholipids	Urine: 512, plasma: 163 + ratios	British	211	3	1 807	1 297
Petersen	2014	*Hum Mol Genet*	24014485	KORA	NMR & MS	Biocrates, Metabolon GC-MS & LC-MS-MS, & Lipofit	Serum	Non-targeted MS (knowns & unknowns); NMR-derived lipid-related traits	649 (151 + 483 + 15)	German	1 814	20	2 909	1 600
Raffler	2013	*Genome Med*	23414815	KORA	NMR	Biocrates AbsoluteIDQ p150	Plasma	NMR peaks	8 600 + 124 750 ratios	German	1 757	7	7	7
Raffler	2015	*PLoS Genet*	26352407	SHIP, KORA	NMR	Bruker BioSpin GmbH	Urine	NMR-derived metabolites & NMR peaks	15 379 features (incl. ratios)	European	3 861 + 1 691	26	35	29
Rhee	2013	*Cell Metab*	23823483	Framingham Heart Study (FHS)	MS	Applied Biosystems	Plasma	Amino acids, amines, polar metabolites, lipids	217	American (European ancestry)	2 076	31	6 730	31
Ried	2014	*Hum Mol Genet*	24927737	KORA, TwinsUK	MS	Biocrates AbsoluteIDQ p150 & Metabolon	Serum	Targeted MS: mainly phospholipids; Non-targeted MS: knowns	344 (151 + 193)	European	1 809 + 843	12	0	0
Rueedi	2014	*PLoS Genet*	24586186	Cohorte Lausannoise (CoLaus), TasteSensomics	NMR	Bruker Biospin Avance III	Urine	NMR peaks	1276	European, Brazilian	835 + 601	11	76	34
Shin	2014	*Nat Genet*	24816252	KORA, TwinsUK	MS	Metabolon UHPLC-MS-MS2 & GC-MS	Serum	Non-targeted MS (knowns & unknowns)	486 + 98 346 ratios	European	7 824	145	144	144
Suhre	2011	*Nat Genet*	21572414	SHIP, KORA	NMR	Bruker BioSpin DRX-400	Serum	NMR-derived metabolites	59 + 1 661 ratios	German	862 + 992	5	575 628	544 540
Suhre	2011	*Nature*	21886157	KORA, TwinsUK	MS	Metabolon UHPLC-MS-MS2 & GC-MS	Urine	Non-targeted MS (knowns)	276 + 37 179 ratios	German, British	1 768 + 1 052	37	15 475	12 874
Tukiainen	2012	*Hum Mol Genet*	22156771	YFS, NFBC1966, HBCS, GenMets, DILGOM	NMR	Bruker Biospin Avance III	Serum	Mainly lipid traits	117 + 99 ratios	Finnish	8 330	30	62 341	3 008
Yet	2016	*PLoS Genet*	27073872	TwinsUK	MS	Biocrates AbsoluteIDQ & Metabolon	Serum	Targeted & non-targeted MS	605 unique (160 + 488)	British	1 001	26	62	40
Yu	2013	*Genet Epidemiol*	23934736	ARIC	MS	Metabolon GC-MS & LC-MS	Serum	Non-targeted MS (knowns & unknowns)	3	African American	1 260	2	76	76
Yu	2014	*PLoS Genet*	24625756	ARIC	MS	Metabolon GC-MS & LC-MS	Serum	Non-targeted MS (knowns & unknowns)	308	African American	1 260	19	19	15

The list of metabolomics GWAS studies included in this table is based on a literature search conducted in April 2017, so only studies published on or prior to this date are included.

Figure 1.12: Plot showing number of mGWAS published per year



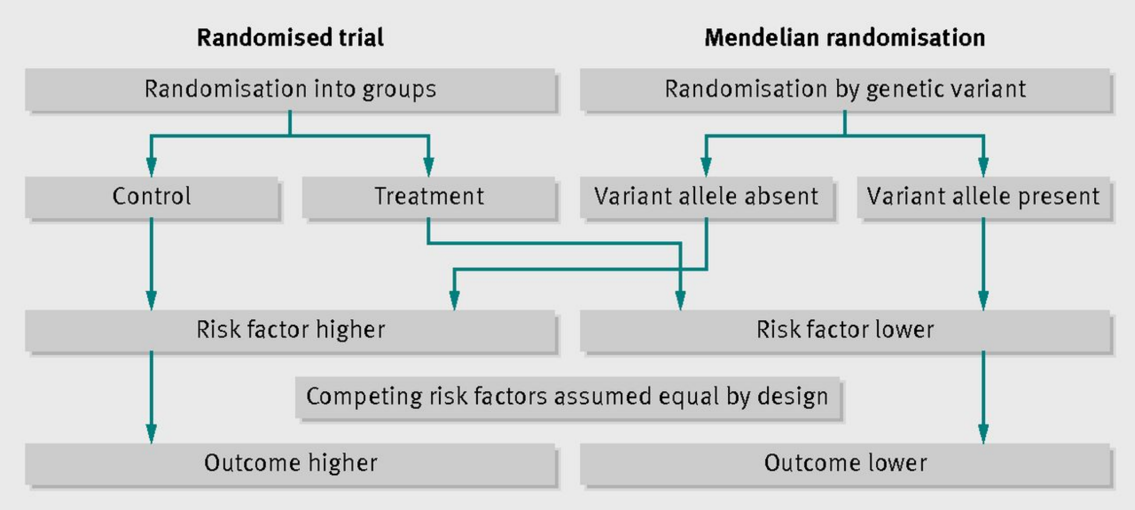
A literature review was conducted to determine the number of metabolite-based genome-wide association studies (mGWAS) published per year.

1.5 Overview of Mendelian randomisation

Mendelian randomisation (MR) is the use of genetic variants as proxies for increased or decreased exposure to a modifiable phenotype (i.e. risk factor) to help judge whether clinical or pharmaceutical interventions on the risk factor are likely to lead to changes in a disease outcome[85,86]. MR is sometimes described as "nature's randomised trial" because genetic variants can be considered to be randomly distributed in a population with respect to environmental and social factors which may be important confounders (Figure 1.13)[87]. Since genetic polymorphisms are allocated approximately randomly at the time of conception, inheriting an allele associated with lower levels of low-density lipoprotein cholesterol (LDL-C), for example, is analogous to being randomly allocated to an LDL-C-lowering therapy at birth, while inheriting the other allele is analogous to being randomly allocated usual care[88]. Provided that the polymorphism is only associated with LDL-C but not with any other traits, then the only difference between groups being compared should be their LDL-C levels. Therefore, comparing the risk of CHD among individuals with and without an LDL-C-lowering allele should provide an estimate of the causal effect of LDL-C on risk of CHD in a manner analogous to a long-term randomised trial[88].

In addition to the analogy with randomised trials, MR is also akin to instrumental variable (IV) approaches that are frequently utilised in econometrics, in which the instrument is a variable that is related to the outcome only through its association with the modifiable

Figure 1.13: Analogy between a Mendelian randomisation study and a randomised trial



Source: Burgess S, et al. *BMJ.* 2012;345:e7325[87].

exposure of interest[89]. There are three basic assumptions of IV analysis that are required to be met in order for a genetic variant to be used to estimate a causal estimate for a risk factor[86]. These assumptions are that:

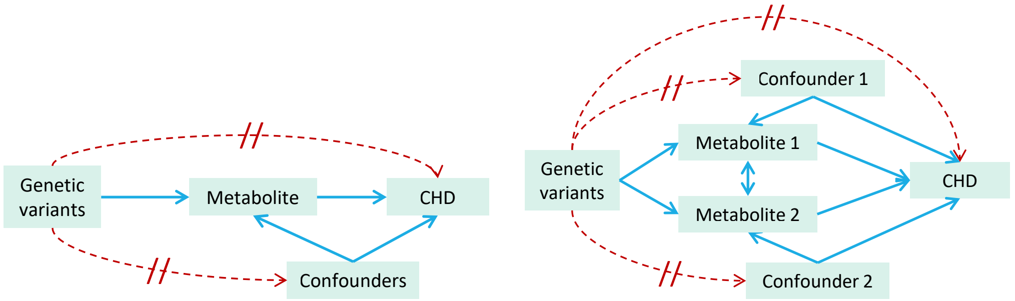
1. The variant is associated with the risk factor,
2. The variant is not associated with any confounder of the risk factor–outcome association, and
3. The variant does not affect the outcome, except possibly via its association with the risk factor.

The first assumption ensures that the genetic variant is not a weak instrument for evaluating the causal effect. Since the variant is used to define subgroups of individuals who either have or do not have the effect allele, the second assumption ensures that all other variables are distributed equally between the subgroups except for the risk factor of interest. The third assumption ensures that the only causal pathway(s) from the genetic variant to the outcome are via the risk factor, which means that the variant is not directly associated with the outcome.

If G is the genetic variant, X is the exposure or risk factor of interest, Y is the outcome, and U is any unmeasured confounder, then the IV assumptions can be stated as[86]:

1. G is not independent of X,
2. G is independent of U, and
3. G is independent of Y conditional on X and U.

Figure 1.14: Directed acyclic graph of instrumental variable assumptions for univariable and multivariable MR



(a) Univariable Mendelian randomisation

(b) Multivariable Mendelian randomisation

A blue arrow indicates that there should be an association between the two boxes under instrumental variable (IV) assumptions; a red dashed arrow indicates that an association between the boxes would violate IV assumptions.

MR can be performed using a single trait as the risk factor of interest or using multiple traits. Since lipid metabolites are highly correlated, univariable MR for a single lipid may not fully take into account the influence of modifications in lipid levels on disease outcomes. In univariable MR of lipid metabolites, one or more genetic variants are used as instrumental variables to help determine whether a single lipid has a causal effect on a disease outcome, whereas in multivariable MR, multiple lipids are assessed for a causal effect on a disease. A directed acyclic graph depicting the MR assumptions as random variables for both the univariable and multivariable scenarios is shown in Figure 1.14. In both situations, the assumptions of MR would be violated if any of the genetic variants used as instrumental variables are directly associated with the outcome, or if they are associated with any confounders of the association between the risk factor(s) and the outcome. If any of the MR assumptions are violated, the genetic variants would have a causal path to the outcome using a different route other than via the risk factor(s) of interest, so they would not be suitable instrumental variables.

The simplest way of estimating the causal effect of a risk factor on a outcome is using the ratio of coefficients method, or the Wald method[86]. To use this approach, the IV can be thought of as a SNP where two of the three subgroups are merged together, which would reflect either a dominant or recessive genetic model (e.g. for a dominant model, the two subgroups would be AA [major homozygote] and Aa/aa [heterozygote/minor homozygote])[86]. Under the assumption of linearity, the ratio estimate can be calculated simply by dividing the average difference in the outcome between the two subgroups by the average difference in the risk factor between the two subgroups ($\frac{\Delta Y}{\Delta X}$)[86]. The equation for estimating the causal effect using the ratio method can therefore be expressed as:

$$\text{Ratio method estimate (continuous IV)} = \frac{\hat{\beta}_{Y|G}}{\hat{\beta}_{X|G}} \tag{1.1}$$

where $\hat{\beta}_{X|G}$ is the coefficient of the gene (G) in the regression of the risk factor (X) on G, representing the change in X for a unit change in G, and $\hat{\beta}_{Y|G}$ is the coefficient of G in the regression of Y on G.

The standard error of the ratio estimate can then be approximated as:

$$\text{Standard error of ratio estimate} \simeq \sqrt{\frac{\text{se}(\hat{\beta}_{Y|G})^2}{\hat{\beta}^2_{X|G}} + \frac{\hat{\beta}^2_{Y|G}\,\text{se}(\hat{\beta}_{X|G})^2}{\hat{\beta}^4_{X|G}}} \tag{1.2}$$

The ratio method only works for a single genetic variant, so more sophisticated approaches are needed when there are multiple IVs. One approach that has been developed for combining summary-level data on multiple genetic variants that are not in LD is to combine the ratio estimates from each variant in an inverse-variance weighted (IVW) meta-analysis[86]. The combined IVW estimate (where k refers to each of the genetic variants used as IVs) is:

$$\hat{\beta}_{IVW} = \frac{\sum_k \hat{\beta}_{Xk}\hat{\beta}_{Yk}\sigma_{Yk}^{-2}}{\sum_k \hat{\beta}^2_{Xk}\sigma_{Yk}^{-2}} \tag{1.3}$$

The approximate standard error of the IVW estimate is:

$$\text{se}(\hat{\beta}_{IVW}) \simeq \sqrt{\frac{1}{\sum_k \hat{\beta}^2_{Xk}\sigma_{Yk}^{-2}}} \tag{1.4}$$

The most straightforward application of MR involves taking a single genetic variant that is associated with the risk factor, but not associated with other risk factors that may represent confounders or alternative causal pathways to the outcome[90]. Such a genetic variant may be hard to find, but for protein biomarkers, such as fibrinogen or C-reactive protein, genetic variants in or near the relevant coding region (in these cases, the *FGB* and *CRP* gene regions, respectively[91,92]) have been shown to have good specificity of association with the risk factor in a number of cases. An association between such a genetic variant and the outcome under these circumstances is indicative of a causal effect of the risk factor on the outcome[93]. In other cases, such as for interleukin-1[94], genetic variants may be associated with alternative risk factors, but so long as these risk factors represent a single causal pathway—that is, they are upstream or downstream of the target risk factor and there is no alternative causal pathway from the genetic variants to the outcome that does not go via the target risk factor—the assumptions necessary for MR would not be violated[95].

Beyond simply determining whether or not an association is causal, an estimate of the causal effect of the risk factor on the outcome can be obtained under further parametric assumptions including linearity[96]. Although the causal estimate is likely to differ from the impact of intervening on the risk factor in practice, and so the magnitude of the causal estimate should not be taken too literally[87], the causal estimate is a valid test statistic for testing the causal hypothesis of whether there is a causal effect of the exposure on the outcome. This enables information on multiple genetic variants to be combined into a single causal estimate, which has greater power to detect a causal effect than a test of the association of any of the individual genetic variants with the outcome[97].

One recent innovation in MR is the use of summarised data on genetic associations with the risk factor and the outcome to obtain a causal estimate[98,99]. These associations can come from a single dataset (one-sample setting), or from separate datasets (two-sample setting)[100]. A practical advantage of the use of summarised data is the ability to analyse publicly-available data from large consortia[101]—such as the Global Lipids Genetics Consortium (GLGC)[35], which has made associations of genetic variants with LDL-C, HDL-C, and triglycerides in over 188 000 individuals available (http://www.sph.umich.edu/csg/abecasis/public/lipids2013/), and CARDIoGRAMplusC4D[43], which has made associations with CHD risk available in over 60 000 cases and 125 000 controls (http://www.cardiogramplusc4d.org). These methods and data resources—in particular, the large sample sizes of consortium data and the ease of obtaining genetic association estimates—have revolutionised the practice and power of MR investigations[102].

1.5.1 MR analyses of major lipids on CHD risk

MR provides an invaluable tool to assess the causal effect of major circulating lipids on risk of CHD. However, in fact, the genetic evidence for a link between hypercholesterolaemia and CHD risk has a long history[103] that precedes the popularisation of MR. Links between LDL-C and CHD risk are well established for both common and rare genetic variants[104], and formal approaches for MR have clearly shown a deleterious causal effect of increased

LDL-C on CHD risk[105,106]. In many ways, despite not being a protein biomarker, LDL-C is an ideal risk factor for use in MR. Several genetic variants associated with LDL-C are located in gene regions that also have corresponding pharmaceutical interventions, such as the *HMGCR* gene region for statins[107], and the *PCSK9* gene region for PCSK9-inhibitors[108,109]. Indeed, an MR analysis using variants in the *NPC1L1* gene region[110] was published in advance of a large trial of ezetimibe (an NPC1L1-inhibitor)[111], and correctly predicted its result. The possible benefit of combination therapy by statin and ezetimibe has been considered in a factorial MR analysis, comparing individuals with genetically-lowered LDL-C due to (1) *HMGCR* variants alone, (2) *NPC1L1* variants alone, and (3) the presence of variants in both gene regions[112]. Genetic variants in different gene regions, as well as genetic variants with varying strengths of association with LDL-C concentrations (including rare gain-of-function and loss-of-function mutations with large effects on LDL-C) have been shown to have associations with CHD risk that are proportional to their association with LDL-C[113], both strengthening the argument that LDL-C is the relevant causal risk factor, and suggesting that all mechanisms of LDL-C-lowering seem to have similar effects on CHD risk. However, the magnitude of the genetically-predicted causal effect of LDL-C on CHD risk is much larger than the observed reduction in CHD risk from taking statins; the MR estimate is 3.5 times larger than the estimate from trials[87] (based on taking statins for 5+ years in primary prevention[114]). One explanation for this is that genetically-predicted variation in LDL-C concentrations is lifelong, and so the MR estimate represents the effect of long-term reduction in LDL-C. Genetic studies have corroborated the slight increases in type 2 diabetes (T2D) risk that are observed in statin trials[115] with several LDL-C-lowering variants showing suggestive associations with increased T2D risk[116]. This suggests that the increase in T2D risk is likely to be an on-target effect of statin drugs, rather than an off-target effect; also that it may be a consequence of LDL-C-lowering more widely rather than a specific effect of intervention on the *HMGCR* pathway.

A similar story can be told for lipoprotein(a) [Lp(a)]. The kringle IV type 2 size polymorphism (a copy number variant) is highly predictive of Lp(a) concentrations, explaining 21 % of variation in Lp(a)[17]—in contrast, no genetic variant for LDL-C, HDL-C, or triglycerides explains more than 1 % of variation[35]. This polymorphism (and also variants in the *LPA* region[16]) are also associated with CHD risk, suggesting a deleterious causal effect of increased Lp(a) on CHD risk[17]. Similarly to that for LDL-C, the effect estimate from the MR analysis is 2.5 times greater than that from a standard observational analysis[17].

For triglycerides, the story is less clear due to a lack of genetic variants associated with triglycerides that do not also associate with LDL-C and/or HDL-C. A methodological development to address this is multivariable MR, in which the causal effects of multiple risk factors can be estimated simultaneously[117]. This requires genetic variants to be associated with one or more of the risk factors, but not associated with other risk factors that may represent confounders of any risk factor–outcome association or alternative causal pathways to the outcome that are not via one of the target risk factors. Multivariable MR analyses have suggested a deleterious causal effect of increased triglycerides on CHD risk[118,119].

While there are genetic variants that have specific associations with HDL-C, these variants are not associated with CHD risk[120]. However, an allele score based on all the genetic variants known to be associated with HDL-C at a genome-wide level of significance is associated with CHD risk, suggesting a protective causal effect of HDL-C provided that the MR assumptions are satisfied[121]. Holmes et al. demonstrated an inverse association with CHD risk for an unrestricted score that explained 3.8 % of the variance in HDL-C, but no association for a restricted score omitting variants additionally associated with LDL-C or triglycerides that explained 0.3 % of the variance in HDL-C[121]. One explanation for the null finding with the restricted score is that the analysis lacked the power to detect a causal effect. Multivariable MR is a useful tool in this case, as a multivariable analysis focusing on HDL-C can include genetic variants that have pleiotropic associations with either LDL-C or triglycerides. This provides robustness to pleiotropy but still reasonable power to detect a causal effect. A multivariable MR analysis using a limited number of genetic variants did not reveal a causal effect of HDL-C[117], and neither did an initial analysis including all genome-wide significant variants[118]. Although a more principled multivariable MR analysis taking into account the relative weights of the genetic variants did suggest a protective effect of HDL-C[119], the magnitude of the effect was too small to be clinically relevant; there is also the potential of some residual bias due to pleiotropic associations of the 185 genetic variants.

1.5.2 Methodological advances in MR and relation to major lipids

There are two other pertinent methodological advances for using MR to assess the causal relevance of major lipids. These are: (1) MR-Egger[122] and (2) a weighted median method[123]. MR-Egger is a method adapted from the meta-analysis literature on publication bias[124]. In an MR setting, each genetic variant contributes an estimate of the causal effect, and a pooled estimate is calculated based on all the genetic variants, which are treated similarly to studies in a meta-analysis. However, if even one of the genetic variants violates the MR assumptions, then the causal estimate from that variant will be biased, and the usual pooled estimate (known as the inverse-variance weighted estimate[99]) will be biased and have an inflated Type 1 error rate. This may lead to false positive findings when genetic variants are pleiotropic[125]. Rather than the standard approach, which assesses whether genetic variants associated with the risk factor are also associated with the outcome, MR-Egger assesses whether there is a dose–response relationship in the genetic associations with the risk factor and with the outcome. This is a higher standard of proof than demanded in a standard MR analysis, and so MR-Egger has reduced Type 1 error rates[122]. MR-Egger also enables a test of "directional pleiotropy", whether pleiotropic associations of genetic variants are likely to bias causal estimates in one particular direction. Additionally, under the assumption that genetic variants may have pleiotropic effects on the outcome, but that these pleiotropic effects are uncorrelated with instrument strength[126], MR-Egger provides a consistent estimate of the causal effect[122].

The second methodological advance, the weighted median method, is a simple idea: rather than calculating a pooled estimate using a weighted mean of the causal estimates based on each genetic variant individually, one can instead report a pooled estimate that is calculated using a weighted median[127]. Unlike the mean, the median is not affected by outlying results, so the weighted median estimate is not sensitive to a handful of pleiotropic genetic variants. Formally, it is a consistent estimate of the causal effect if at least half of the genetic variants (by weight) are valid instruments[123].

Both the MR-Egger and weighted median approaches are worthwhile sensitivity analyses for MR when some genetic variants are suspected to be pleiotropic. The MR-Egger estimate has the advantage that it allows all genetic variants to be pleiotropic, although it makes an assumption on the distribution of these pleiotropic effects. However, the MR-Egger estimate may be imprecise, and it is highly influenced if there are one or two strong variants. The weighted median estimate is more precise and more stable, but relies on the majority of the evidence in the analysis being reliable.

The application of these methods to major lipids is very revealing: using all genome-wide significant variants, standard MR, MR-Egger, and weighted median analyses all suggest causal effects of LDL-C and triglycerides on CHD risk, with no evidence of directional pleiotropy. However, while the standard MR analysis using all genome-wide significant variants for HDL-C suggests a protective effect of HDL-C on CHD risk, the MR-Egger and weighted median analyses suggest a null effect, with evidence of directional pleiotropy in the MR-Egger analysis[123]. This null finding is supported by trial evidence on CETP inhibitors, which raise HDL-C levels and lower LDL-C levels, but do not lower CHD incidence[128].

The above examples demonstrate that MR analyses can be simple or not, depending on the available genetic variants and their specificity of association with the risk factor under analysis. A naïve MR analysis, particularly one using a large number of genetic variants, can be misleading. However, the development of new methods can help either to add confidence in the finding from an MR analysis, or to call it into question[129]. The diversity of methods available for conducting MR are especially useful when evaluating the causal effect of lipid metabolites on CHD. The high degree of correlations between lipid metabolites and extensive pleiotropy require careful application of various MR methods to instil further confidence in the results.

1.5.3 Application of MR to metabolomics

As described in Chapter 1, the widespread measurement of high-dimensional phenotypic traits brings novel opportunities to perform GWAS that can examine the associations of millions of genetic variants with hundreds of metabolites or thousands of proteins[130]. While numerous metabolomics GWAS have been performed in recent years as the literature review showed Table 1.3, very few metabolomics studies have used an MR approach to assess whether the associated phenotypic traits that they identified could have causal effects on diseases or risk factors.

The studies that have employed MR on high-throughput data have taken either of two approaches: (1) to determine the causal role of conventional risk factors on levels of metabolites, or (2) to determine whether metabolites have a causal effect on diseases or traits. As an example of the first approach, a meta-analysis of four Finnish population cohorts obtained levels of 82 different metabolites and metabolic measures using nuclear magnetic resonance, including lipoprotein lipids, fatty acids, and amino acids[131]. The authors found evidence that strongly supports causal effects of adiposity on 24 metabolites that are potential cardiometabolic risk factors[131]. Another study using mass spectrometry in a British population determined that gene expression levels derived from expression quantitative trait loci (eQTLs) in fat, skin, and lymphoblastoid cell lines could play a causal role on levels of a wide range of metabolites[132]. The authors identified two loci (*THEM4* and *CYP3A5*) where the allele associated with increased metabolite levels was significantly associated with decreased gene expression in one or more tissues, supporting the notion that the underlying causal variants for these two loci could have regulatory consequences[132].

To illustrate the second approach, a study using the same platform could not find any evidence that the levels of three metabolites—urate, mannose, and unnamed metabolite X12063—have a causal effect on relative appendicular lean mass[133]. However, a prospective cohort study that conducted mass spectrometry used summarised CHD association results from CARDIoGRAMplusC4D[43] to find four lipid-related metabolites (lysophosphatidylcholines 18:1 and 18:2, monoglyceride 18:2, and sphingomyelin 28:1) with evidence for a causal role in CHD development[70]. An interesting approach used by another study did not make any *a priori* assumptions on whether metabolites should be considered as the risk factor or the outcome; they performed bidirectional analyses but could not detect any causal associations in either direction between mRNAs and metabolites[134].

Metabolomics and proteomics particularly stand to benefit from the availability of summarised data for MR and a two-sample setting, where the associations of high-dimensional phenotypic traits with genetic variants are measured in one population (usually a small cross-sectional study of healthy individuals) and the associations of those variants with diseases and risk factors are measured in another population such as large consortia (for disease outcomes, usually a consortium of case–control studies)[101]. Furthermore, the multivariable MR approach will be particularly relevant to high-dimensional platforms, and lipidomics in particular, as it may be difficult to find genetic variants having a specific association with a single variable when the traits are highly correlated with each other[135]. The greatest challenge of MR with lipidomics lies in identifying a suitable set of genetic variants for a particular lipid metabolite (or several lipid metabolites for multivariable MR) that will not violate the IV assumptions. Thus, the MR-Egger and weighted median methods could be especially important to provide some robustness against pleiotropic variants.

There is tremendous scope and untapped potential to apply MR in investigating plausible novel causal pathways of high-dimensional phenotypic traits with diseases and risk factors. MR is a tool that can provide additional evidence to prioritise further research and clinical applications, or just as importantly, to discourage additional resource allocation towards a specific pathway. Over the coming years, MR is likely to be applied with increasing regularity to high-dimensional phenotypic data where concomitant genetic information is available, and in lipidomics in particular.

1.6 Dissertation outline

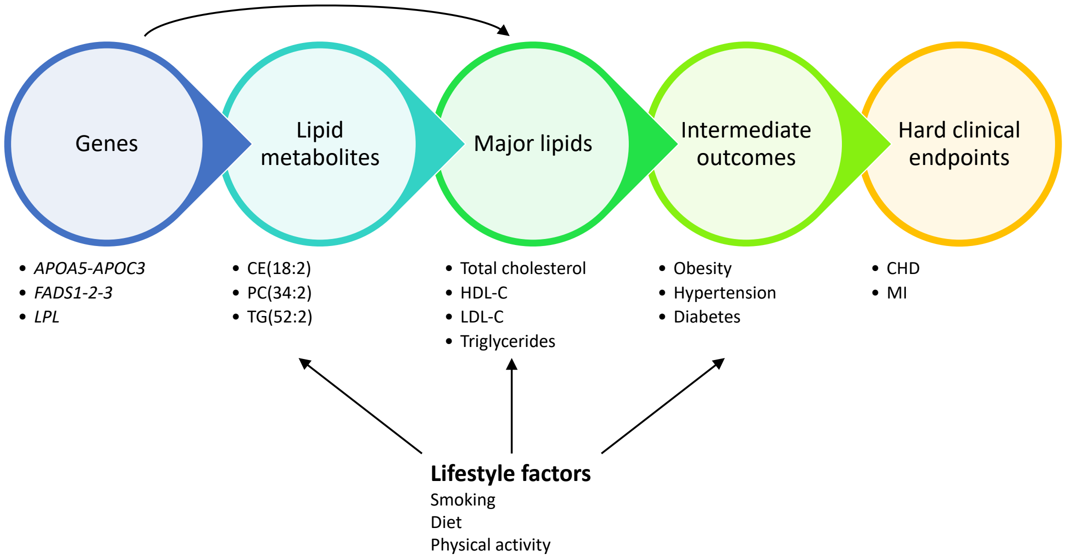
The primary objectives of this dissertation were (1) to identify the genetic determinants of lipid metabolites, and (2) to advance understanding of the effect of perturbations in lipid metabolite levels on CHD and its risk factors. Through advancement of the knowledge base in this important field, it is intended that the findings of this dissertation could lead to further studies that would help advance mechanistic understanding and prioritise novel therapeutic targets for drug development and personalised medicine. Each of the following chapters of this dissertation seeks to address various facets of this overall goal and approach it from different angles, such as the association of lipid metabolites with major lipids and CHD risk factors (Chapter 4), the genetic determinants of lipid metabolites (Chapter 5 and Chapter 6), and the causal relevance of lipid metabolites with risk of CHD (Chapter 7). The influence of various lifestyle factors such as smoking, diet, and physical activity on all of these associations is also considered. A conceptual framework that portrays the connection between all of these various aspects is shown in Figure 1.15. For clarity, whenever the word “lipids” is used in this dissertation, it refers to lipid metabolites—it only refers to major circulating lipids such as HDL-C and LDL-C if appropriately qualified as such.

The primary dataset that was used for analysis in this dissertation is the Pakistan Risk of Myocardial Infarction Study (PROMIS). Chapter 2 provides a background of PROMIS, presents descriptive statistics for the subset of PROMIS participants for whom lipidomics measurements were taken, and describes the data management steps involved in processing and cleaning the PROMIS phenotypic, biomarker, and genetic data. To provide a wider context, demographic and clinical characteristics of the controls in PROMIS with lipidomics measurements are also compared with the wider set of controls in PROMIS and in Pakistan as a whole, and differences between the full set of cases and controls in PROMIS are also compared.

The lipidomics platform is described in Chapter 3, including the methods used to process the lipidomics data and the quality control steps that were performed.

Chapter 4 presents results of descriptive analyses of the lipidomics data, including heat maps of cross-correlations of the lipid metabolites, correlations of the lipid metabolites with major lipids and other circulating biomarkers, principal component analysis, partial least squares discriminant analysis, and Gaussian graphical modelling on the lipidomics data, and the association of the principal components of the lipid metabolites with CHD risk factors.

Figure 1.15: Conceptual framework of analysis approach in this dissertation



This figure presents a conceptual framework of the analysis approach used in this dissertation. Genes can influence levels of lipid metabolites, which in turn make up major lipids. Changes in levels of these lipids can affect intermediate CHD outcomes such as obesity, hypertension, and diabetes, which can eventually lead to an MI or CHD. Lifestyle factors can also play an important role in influencing levels of individual lipid metabolites, major lipids, and the development of CHD outcomes. Examples of the various aspect of the conceptual framework are given below each heading. Chapter 4 explores cross-correlations of lipid metabolites, the association of lipid metabolites with major lipids, the association of lifestyle and environmental factors (e.g. diet, physical activity, and smoking) with lipid metabolites, and the association of lipid metabolites with intermediate CHD outcomes. Chapter 5 and Chapter 6 explore the association of genetic factors with lipid metabolites, and Chapter 7 describes a Mendelian randomisation study that was used to assess whether lipid metabolites have a causal association with risk of CHD.

Chapter 5 describes the steps involved in conducting univariate genome-wide association analyses on all 444 lipids that were measured using the lipidomics platform, and presents the overall findings from the GWAS results.

Chapter 6 provides an analysis and interpretation of the GWAS results, including variant annotation and incorporation of information from pharmacological and functional databases to aid biological understanding of mechanisms through which genetic variants influence metabolic pathways. Also included in this chapter is a description of the conditional analyses and the steps that were involved to determine the number of independent loci and how many of these loci are novel.

Chapter 7 provides an overview of MR and its use for determining the causal effect of perturbations in levels of a risk factor on a disease outcome, particularly in the area of circulating lipids and lipid metabolism. It then describes the application of MR using several different methods to assess the causal relevance of the lipid metabolites in this study for risk of CHD.

Chapter 8 provides a summary of the key findings from this dissertation and their potential implications. It also summarises the strengths and limitations of the dataset used for this dissertation and fruitful avenues for future research. While the genetic analyses described in this dissertation primarily focus on a single discovery stage from one study, preliminary results from the replication of these findings in a different study population with nearly a threefold increase in the number of participants, using the exact same lipidomics platform, will be described in this final chapter.

CHAPTER 2

Description of the Pakistan Risk of Myocardial Infarction Study

Chapter summary

This chapter provides a background and description of the main study that was used for analysis in this dissertation, the Pakistan Risk of Myocardial Infarction Study (PROMIS), which is a case-control study of myocardial infarction (MI) with over 35 000 participants recruited from nine hospitals in urban Pakistan. This chapter also includes descriptive statistics of the PROMIS data, explains the data management steps that were performed in order to clean, process, and harmonise the data, and describes the steps that were involved in quality control of the genetic information.

Lipidomics measurements were only obtained for controls who were free from MI at recruitment, but this population was still at risk of chronic diseases since 56 % of the participants were overweight, 17 % were obese, 18 % had hypertension, and 38 % had diabetes. Genotyped data were imputed and cleaned to yield genetic information on over 6.7 million variants. The combination of lipid and other biomarker data, information on CHD risk factors, and genetic data makes PROMIS a rich resource for investigating and identifying emerging and established risk factors in Pakistan.

2.1 Introduction

The burden of CVD has been increasing in South Asia at a greater rate than in any other region of the world. While CVD mortality has decreased in most high-income countries from 1990 to 2015, there has been no such decline in South Asia[8]. In many South Asian countries, CHD manifests about 10 years earlier on average compared with the rest of the world. Whereas CVD in Western countries is often considered a disease of late middle-age and the elderly (i.e. about 77 % of CVD deaths occur above the age of 70), the vast majority of CVD deaths in South Asian countries occur amongst people under 70 years of age, resulting in substantial loss of productive working years due to premature CVD morbidity and mortality[1].

Large-scale population studies in South Asia provide a useful resource for the investigation of locally relevant chronic disease risk factors in order to understand the reasons for the high burden of chronic diseases in this setting and what steps can be taken to address the problem.

2.2 Overview of PROMIS

The Pakistan Risk of Myocardial Infarction Study (PROMIS) is a case-control study in Pakistan. The primary outcome is first-ever acute myocardial infraction (MI). An overview of PROMIS and a summary of the key methodological features is provided in this chapter; further details and specifics about the study can be found in the protocol paper[25].

2.2.1 Recruitment of participants

PROMIS participants were recruited from participating hospitals in nine urban centres in Pakistan. A map indicating the name of each recruitment centre and their geographic location is shown in Figure 2.1; the number of participants recruited from hospitals in each city is shown in Figure 2.2. Four hospitals in Karachi, which is the largest city in Pakistan and the sixth largest city in the world[136], collectively recruited over half (54 %) of the PROMIS participants.

Cases were selected from hospital patients at each centre who had recently had an MI. Patients were eligible for selection if they were between the ages of 30 to 80 years, they had been admitted to the emergency room for an MI with sustained clinical symptoms lasting at least 20 minutes within the 24 hours prior to hospitalisation, and they had electrocardiogram changes typical of an MI and a positive troponin-T test. Controls were concurrently identified and recruited in the sample hospital as cases according to the following order of priority: (1) visitors of patients attending the outpatient department; (2) patients attending outpatient clinics for non-cardiac-related symptoms (e.g. routine health check-ups; refraction and cataracts patients; minor ear, nose, and throat patients; or individuals undergoing minor elective surgery), and (3) non-first-degree relative visitors of MI cases. The controls were frequency-matched to cases based on sex and age in five-year bands. The exclusion criteria included if participants had (1) any previous history of CVD; (2) an onset of chest symptoms and hospitalisation for MI lasting longer than 24 hours; (3) a history of viral or bacterial infection in the past two weeks; (4) presence of chronic conditions (e.g. tuberculosis, malaria, hepatitis, or renal failure); (5) pregnancy; or (6) failure to give informed consent.

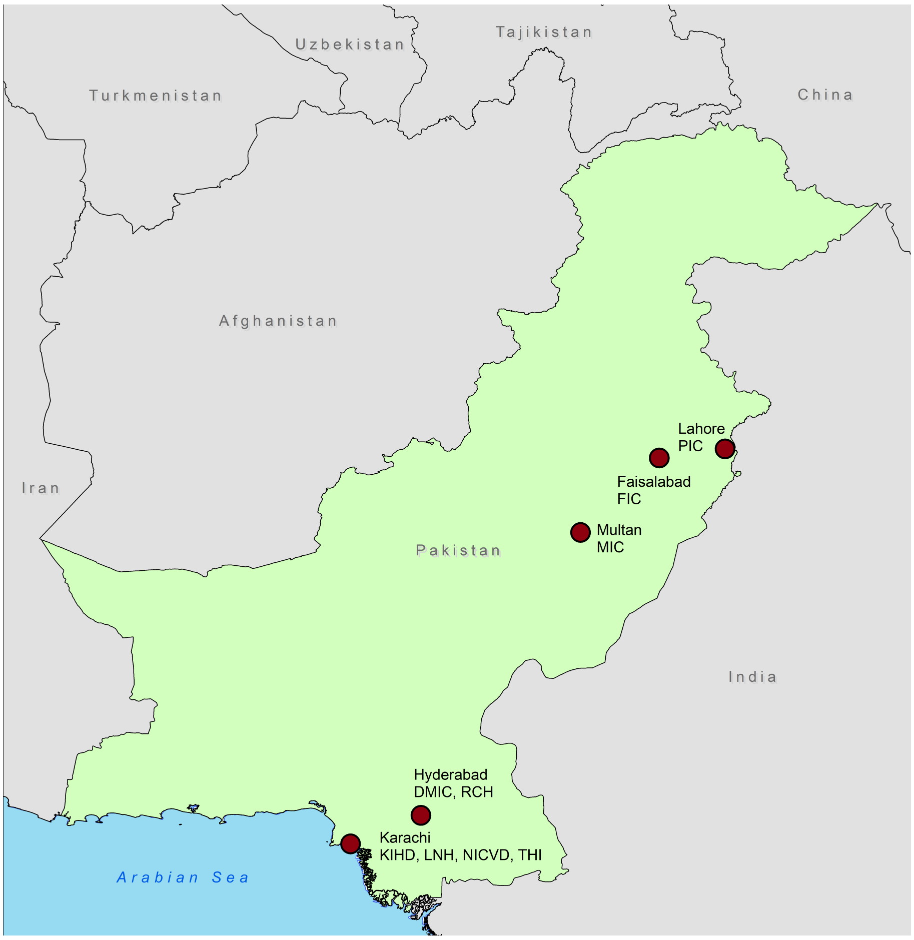
Recruitment of PROMIS participants took place from 2005 to 2011, and the date that each survey was completed is shown in Figure 2.3. The study has now completed recruitment, and the final sample size is approximately 16 700 cases and 18 600 controls.

Ethical approval was obtained from the ethics committees responsible for each of the PROMIS recruitment centres, as well as from the Center for Non-Communicable Diseases (CNCD) in Pakistan.

2.2.2 Questionnaire

All participants who were enrolled in the study received a questionnaire that was specifically designed for the study population. The questionnaire consisted of approximately 150 questions and took about three hours to complete on average. However, response time could vary since some of the questions pertained only to cases (i.e. in relation to their hospitalisation for MI) or only to women (i.e. in relation to reproductive health). The questionnaire assessed information about each participant concerning a range of different factors, including demographics, socioeconomic status, tobacco and alcohol consumption, and physical activity. There was also a dietary section, which obtained detailed information about the frequency in which participants consumed various locally-relevant foods. All of the questions, especially the ones related to smoking and diet, were tailored to the specific setting in Pakistan to ensure that the questionnaire was culturally relevant. For instance, participants were asked about their use of cigarettes, beedies, huqqa/chilum, paan, naswar, gutka, and supari, which are all various types of tobacco products that are commonly used in Pakistan. In the food frequency section, participants were asked about their consumption of approximately 160 different types of locally relevant foods,

Figure 2.1: Map of recruitment centres in PROMIS



Map was created using ArcGIS® software (ArcMap 10.4.1) by Esri[137]. Country level administrative boundary shapefiles were obtained from GADM version 2.8[138]. Red circles are shown and labelled with the city where each PROMIS recruitment centre is located. **Abbreviations: DMIC** = Deewan Mushtaq Institute of Cardiology; **FIC** = Faisalabad Institute of Cardiology; **KIHD** = Karachi Institute of Heart Diseases; **LNH** = Liaquat National Hospital & Medical College; **MIC** = Multan Institute of Cardiology; **NICVD** = National Institute of Cardiovascular Diseases; **RCH** = Red Crescent Hospital; **PIC** = Punjab Institute of Cardiology; **THI** = Tabba Heart Institute.

Figure 2.2: Number of PROMIS participants recruited from hospitals in each city

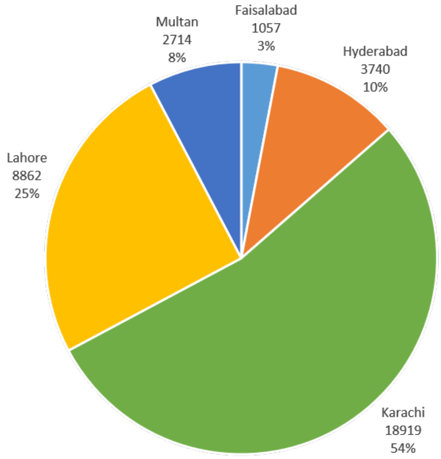
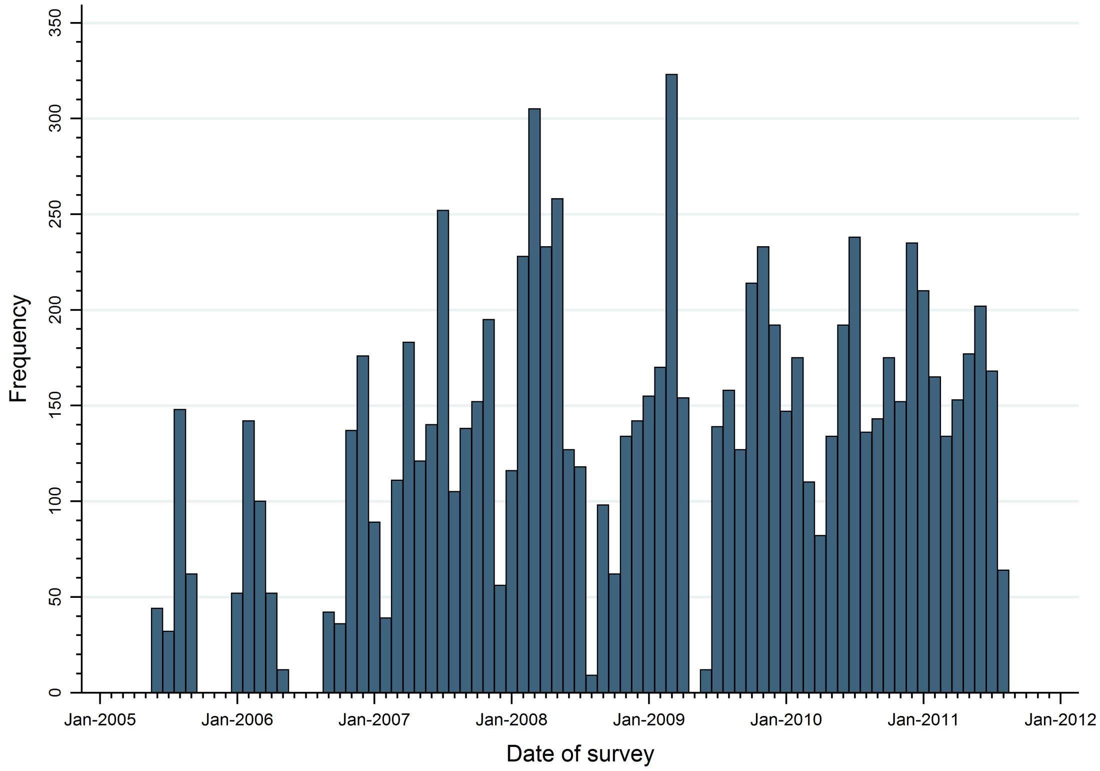


Figure 2.3: Distribution of survey dates in PROMIS participants with lipidomics measurements



including paratha, roti, daliya, daal, pakoray, kheer, and halwa, which were grouped into 54 categories by a local nutritionist.

In addition to the self-reported questions, other measurements were also taken that are not susceptible to recall or social desirability bias. For example, anthropometry measurements were taken to record the participants' systolic and diastolic blood pressure (SBP and DBP, respectively), resting heart rate, height, weight, and waist and hip circumference. These measurements were all taken using standardised procedures and equipment as described in the protocol paper[25].

A full-length copy of the entire PROMIS questionnaire can be found in Appendix C.

All of the information extracted from the questionnaire, as well as measurement data, was entered in duplicate into a central database in Pakistan, then sent to the data management team at the University of Cambridge. The data were then extracted, cleaned, and harmonised as described in subsection 2.2.6.

2.2.3 Biomarker data

Non-fasting blood samples (with the time since last meal recorded) were drawn from each participant and centrifuged within 45 minutes of venepuncture. Serum samples were stored at −80 °C until use. The samples were assayed at the CNCD in Pakistan to measure values of standard biomarkers, including total cholesterol, HDL-C, LDL-C, triglycerides, glucose, creatinine, and HbA_{1c}. Serum and ethylenediaminetetraacetic acid (EDTA) plasma samples were also sent to various laboratories, including the University of Cambridge, the University of Pennsylvania, Tufts University, and the University of Washington, for analysis of over 150 additional biomarkers, including interleukin-1 receptor antagonist, interleukin-6, interleukin-18, adiponectin, C-peptide, E-selectin, matrix metalloproteinase-9, serum amyloid, intercellular adhesion molecule 1, alkaline phosphatase, alanine transaminase, aspartate aminotransferase, and many others.

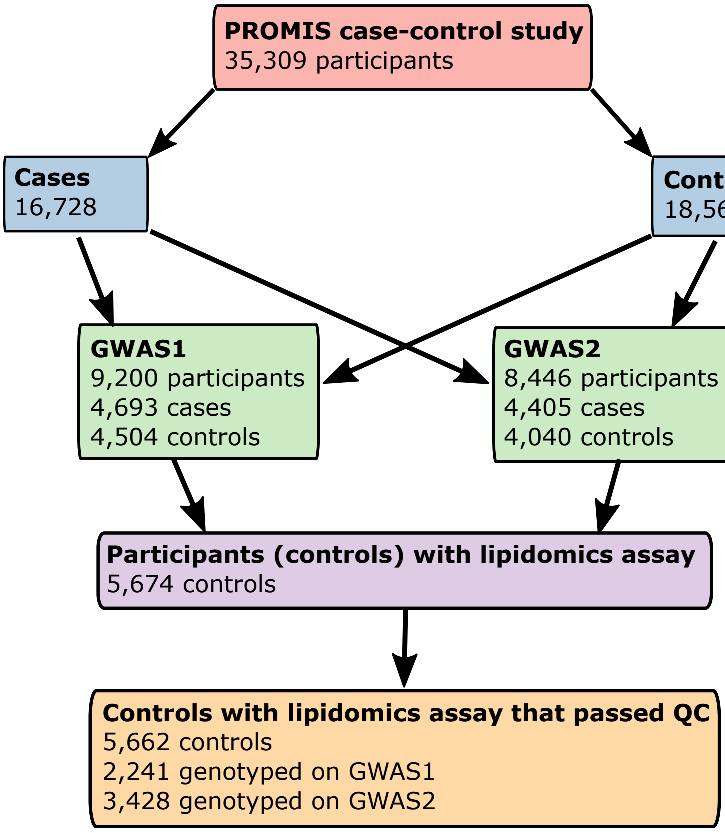
2.2.4 Lipidomics data

It was decided to prioritise the selection of participants for the lipidomics assay based on several criteria. Participants had to be a control free from MI at baseline since the analysis of lipid levels from blood samples taken immediately following an MI would not be an accurate portrayal of normal lipid levels. In addition, participants needed to have available genetic data from a GWAS platform to facilitate genetic analyses on the lipidomics data. Furthermore, the participants needed to have complete information on age, sex, ethnicity, centre from which they were recruited, and the date that the survey was completed. This resulted in 5674 PROMIS participants for whom lipidomics measurements were taken using direct infusion high-resolution mass spectrometry, of whom 5662 passed QC. An overview of the process of selecting participants for this study is shown in Figure 2.4, which includes details on the overall number of cases and controls in PROMIS, the number of participants genotyped on each GWAS platform, and the final number of controls with lipidomics data that were included in the analysis. A detailed description of the lipidomics platform will follow in Chapter 3.

2.2.5 Genetics data

An overview of the quality control (QC) steps that were performed on the genetic data are shown in Figure 2.5. DNA samples were extracted from leukocytes in Pakistan and genotyped at the Wellcome Trust Sanger Institute in Cambridge, UK. PROMIS participants were genotyped on either of two GWAS platforms: (1) the Illumina 660-Quad GWAS

Figure 2.4: Flowchart of participant selection for lipidomics study



GWAS1 refers to the Illumina 660-Quad GWAS platform; GWAS2 refers to the Illumina HumanOmniExpress GWAS platform.

platform (referred to in this dissertation as GWAS1), which consisted of 527 925 genotyped variants after QC steps were performed, or (2) the Illumina HumanOmniExpress GWAS platform (referred to in this dissertation as GWAS2), which consisted of 643 333 genotyped variants after QC. Overlapping samples were removed so that there was a distinct set of individuals on each genetic platform: out of the controls with lipidomics measurements, 2241 participants remained that were genotyped on GWAS1, and 3428 participants remained that were genotyped on GWAS2. Genetic samples were removed if they were heterozygosity outliers (heterozygosity > mean ± 3-SD), if the call rate was less than 97 %, if there was discordant sex between genotype and phenotype, or if they were duplicate or related pairs (kinship coefficient > 0.375). Single nucleotide polymorphisms (SNPs) were excluded if the call rate was less than 97 %, there was evidence of departure from Hardy-Weinberg Equilibrium (HWE) at a P-value of less than 1×10^{-7}, or the minor allele frequency (MAF) was less than 1 %.

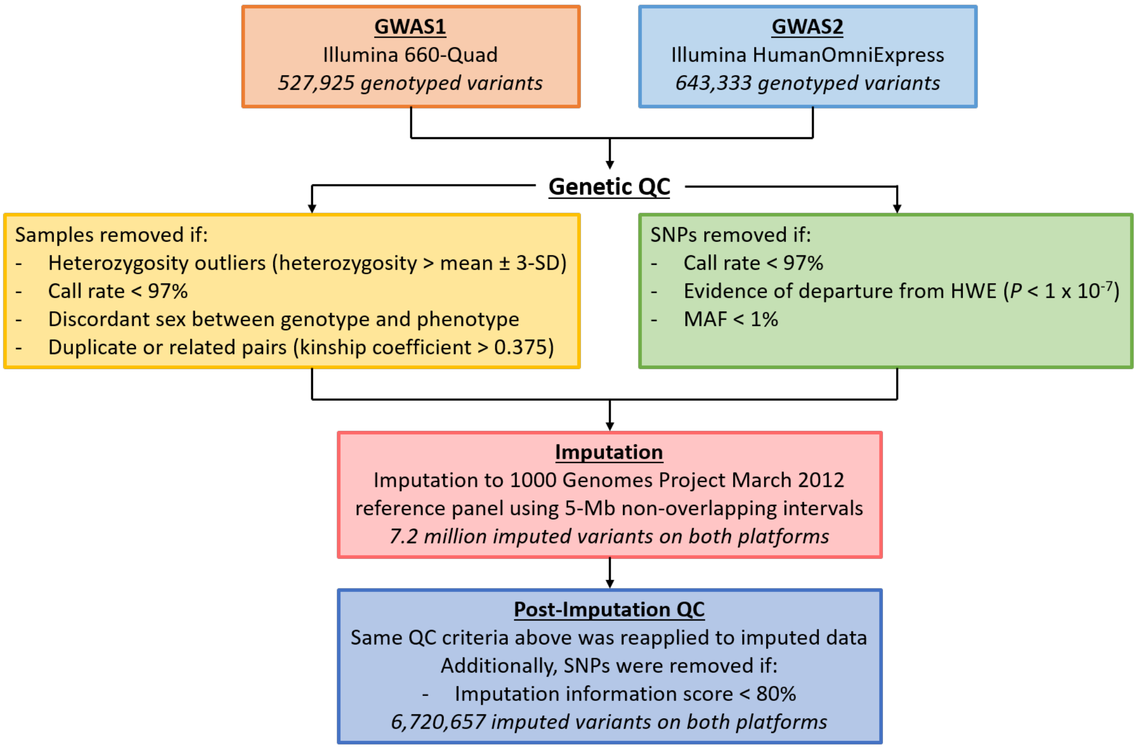
The kinship coefficient mentioned above (0.375) was selected as the cut-off in this study due to the high degree of consanguinity in marriages in Pakistan. More than half of all marriages in Pakistan (56 %) are between first and second cousins[139]. A kinship coefficient of 1 indicates duplicate individuals or monozygotic twins, a value of 0.5 ($\frac{1}{2}$) indicates first-degree relatives, a value of 0.25 ($\frac{1}{4}$) indicates second-degree relatives, and a value of 0.125 ($\frac{1}{8}$) indicates third-degree relatives. In most GWAS studies, it is typical to remove one individual from each pair with a kinship coefficient > 0.1875, which is halfway between the expected values for second- and third-degree relatives[140]. However, due to the extent of intermarriage in the population, a less stringent threshold was necessary. Therefore, a kinship coefficient > 0.375 ($\frac{3}{8}$) was chosen as the cut-off, which is halfway between the expected values for first- and second-degree relatives.

Imputation was applied to both of the cleaned PROMIS genotyped datasets using the 1000 Genomes Project[141] March 2012 (v3) release as the reference panel. Imputation was conducted using IMPUTE v2.1.0[142] using 5-Mb non-overlapping intervals for the whole genome. Once imputation had been performed on GWAS1 and GWAS2 separately, there were over 7.2 million directly genotyped or imputed SNPs available for analyses in either dataset before further QC.

In total, 5662 individuals had concomitant information on lipidomics data and directly genotyped or imputed SNPs. QC filters for SNPs were applied during the analysis stage using SNPTEST v2.4.1[143] to remove SNPs that were poorly imputed. The imputation information score is a metric between 0 and 1 that provides an assessment of the level of accuracy of imputation. A value of 1 indicates that there is no uncertainty in the imputed genotypes whereas a value of 0 means that there is complete uncertainty about the genotypes. SNPs were removed if they had an imputation information score < 0.80. The results were then extracted from the output files and the original QC filters were reapplied to the combined genotyped and imputed data (i.e. MAF, HWE P-value, or call rate below specified cut-offs). After the final QC filters were applied, there were 6 720 657 SNPs remaining. Additional QC steps that were performed on the genetic data as part of the GWAS analyses will be described in Chapter 5.

The two GWAS platforms that were described above (GWAS1 and GWAS2) are the only sources of genetic data that were analysed for the purposes of this dissertation; however, PROMIS participants have also been genotyped on several other platforms, including the Metabochip array, a customised Metabochip array (known as Metabochip+), and a customised Exome array (known as Exome+). PROMIS participants also underwent whole genome sequencing (WGS) and whole Exome sequencing (WES). A significant proportion of the PROMIS participants with lipidomics data were also assayed on these other genetic platforms and the association of each lipid metabolite with SNPs on these platforms have

Figure 2.5: Quality control steps performed on genetic data


GWAS1
Illumina 660-Quad
527,925 genotyped variants
GWAS2
Illumina HumanOmniExpress
643,333 genotyped variants
Genetic QC
Samples removed if:
- Heterozygosity outliers (heterozygosity > mean ± 3-SD)
- Call rate < 97%
- Discordant sex between genotype and phenotype
- Duplicate or related pairs (kinship coefficient > 0.375)
SNPs removed if:
- Call rate < 97%
- Evidence of departure from HWE ($P < 1 \times 10^{-7}$)
- MAF < 1%
Imputation
Imputation to 1000 Genomes Project March 2012
reference panel using 5-Mb non-overlapping intervals
7.2 million imputed variants on both platforms
Post-Imputation QC
Same QC criteria above was reapplied to imputed data
Additionally, SNPs were removed if:
- Imputation information score < 80%
6,720,657 imputed variants on both platforms

also been analysed, but as that work is beyond the scope of this dissertation, only the results pertaining to the two mentioned GWAS platforms will be described.

2.2.6 Data processing, cleaning, and harmonisation

Alongside doctoral research, the author of this dissertation also carried out data management for PROMIS within the Cardiovascular Epidemiology Unit (CEU). The responsibilities for this role included collating, cleaning, harmonising, and managing a database of questionnaire and laboratory data for over 35 000 PROMIS participants, including the socio-demographic, anthropometric, medical history, lifestyle, dietary, and biochemical information described above. It also involved maintaining and continually updating a reference file that enabled the phenotype data to be linked to the genetic data by mapping each participant ID with the corresponding genetic sample(s) on each genetic platform.

Whenever new PROMIS data arrived, it was merged into the overall PROMIS database by matching on both the variable names (columns) and the list of participants (rows). If the variable already existed, then the new values were used to update the existing values for each participant. However, if new variables were provided, then these were added to the dataset. The values of each variable were also converted from the original units (e.g. ng/mL for C-peptide) to a standard set of units used internally by the CEU (i.e. nmol/L for this example). Additionally, any duplicate participants or variables were removed before merging in the data. Whenever there were any discrepancies in the new dataset, queries were sent back to the provider of the data and followed up to ensure that the issues were resolved.

Another important aspect of this role involved producing datasets containing specified variables of interest on subsets of participants, and answering any queries that arose. Over 100 datasets were produced for colleagues in the CEU as part of this data management role, each customised to meet the requirements of the specific project that the analysts were working on, and updated datasets were created whenever new information became available.

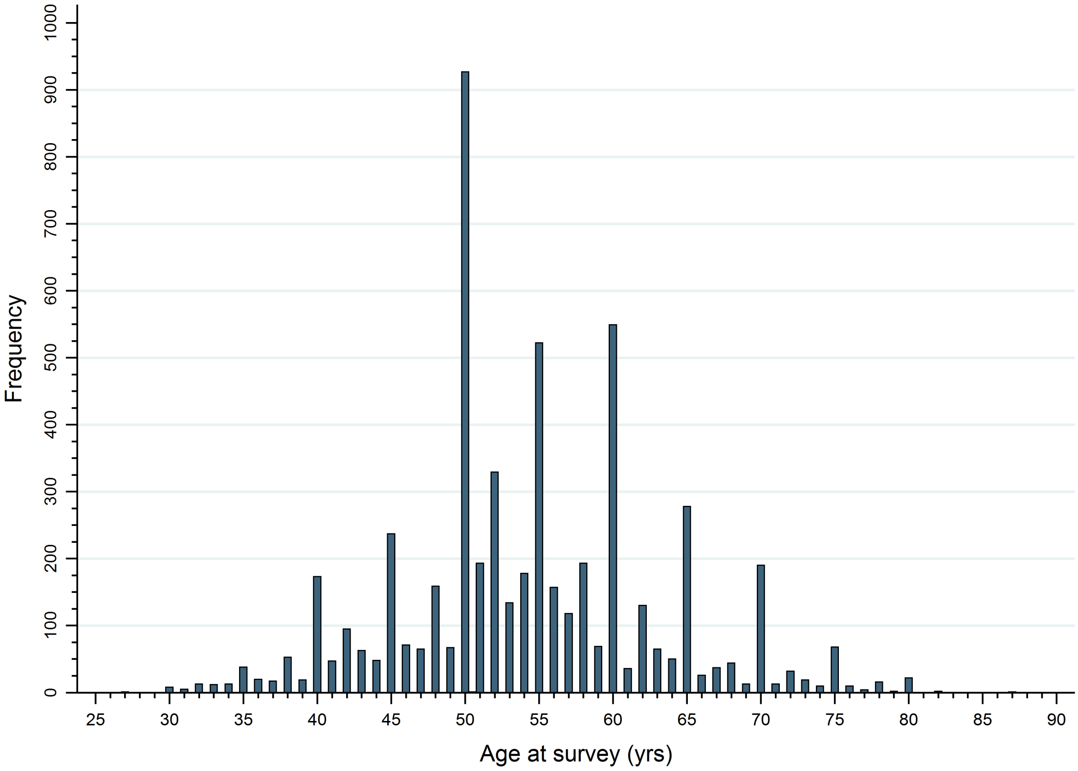
The statistical software SAS[144] v9.3 was used for all data management tasks. A number of SAS macro scripts, some specific to PROMIS and some applicable to all projects in the CEU, were used to perform functions such as merging in new datasets, appending datasets, checking for duplicate observations, comparing values between datasets, cleaning variables, converting variables to consistent units, and producing cleaned datasets. These scripts were instrumental in performing data management responsibilities.

2.3 Demographic and clinical characteristics of PROMIS participants

Demographic and clinical characteristics of the 5662 PROMIS participants with lipidomics information are shown in Table 2.1. The median age was 54 years, with a range from 27 to 87. Age and/or date of birth were self-reported on the questionnaire, and it appears that a majority of participants rounded their age to the nearest five years since the top ten most commonly reported ages of participants were 50, 60, 55, 52, 65, 45, 51, 58, 70, and 54; fully 53 % of the participants had ages that were divisible by five. There is no other reasonable explanation for the fact that 928 participants were age 50, but only 67 participants were 49; likewise, 190 participants were 70, but only thirteen participants were 69 and thirteen participants were 71. A histogram depicting this phenomenon is shown in Figure 2.6.

The majority of participants (79 %) were male. The most common ethnicities were Urdu and Punjabi (37 % and 36 % of participants, respectively). The average BMI was 26 kg/m^2, which is rather high since any BMI over 25 is considered overweight. In fact, although all participants in this analysis were controls without myocardial infarction at baseline, they represent a population at increased risk of MI, since 56 % of the participants were

Figure 2.6: Distribution of ages of PROMIS participants with lipidomics measurements



overweight, 17 % were obese, 18 % had hypertension, and the proportion of participants with diabetes was 38 % according to elevated HbA_{1c} levels, or 31 % based on elevated fasting plasma glucose (FPG) levels. Hypertension was defined as SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg; overweight and obese were defined as body mass index (BMI) $\geq$ 25 kg/m^2 or BMI $\geq$ 30 kg/m^2, respectively; and diabetes was defined as either FPG $\geq$ 126 mg/dL or $HbA_{1c} \geq 6.5\,\%$.

There were significant differences in levels of CHD risk factors according to ethnicity. Participants from Urdu and Punjabi ethnic backgrounds had significantly higher levels of being overweight and having obesity, hypertension, and diabetes compared with participants from other ethnic backgrounds (which includes Pathan, Balochi, Sindhi, Memon, and Gujrati) (chi-square test $P < 0.001$ for all comparisons).

In order to validate the extent to which controls in PROMIS were indeed at elevated risk of MI, the levels of demographic and clinical characteristics were also compared between cases and the full set of controls in PROMIS (Table 2.2). There were statistically significant differences ($P < 0.0001$) between cases and controls in all anthropometric markers, all circulating lipid biomarkers, all categorical variables, and almost all CHD risk factors. Compared with all MI cases, controls in the overall PROMIS study were slightly older and had slightly higher BMI, SBP, DBP, and levels of triglycerides. Also a significantly higher proportion of controls were Punjabi ($P < 0.0001$), though the absolute difference was marginal (40 % of controls versus 36 % of controls). On the other hand, controls had slightly higher levels of HDL-C (the "good" cholesterol) and slightly lower levels of LDL-C (the "bad" cholesterol), and were less likely to smoke, take diabetic or hypertensive drugs, or have diabetes. Despite the statistically significant differences in many relevant characteristics between cases and controls, levels of relevant CHD risk factors were still elevated beyond the normal healthy range in a significant proportion of the controls, putting them at increased risk of an MI.

To examine the representativeness of the sample with lipidomics measurements, these results were compared to the broader population of Pakistan. The subset of PROMIS controls selected for the lipidomics assay (n = 5662) were comparable to all controls in the entire PROMIS study (n = 18 564) on the basis of the demographic and clinical characteristics listed in Table 2.1. The PROMIS participants were also compared with the general population of Pakistan (n = 13 558) using the latest available data obtained from the Demographic and Health Survey for Pakistan[139]. This analysis showed that PROMIS participants were older on average, and a higher proportion consumed tobacco and were overweight, compared with the head of household in the general Pakistani population. However, these differences are perhaps to be expected because the controls in PROMIS were visitors and blood relatives of patients in the hospital who had recently been admitted due to a heart attack, and the controls were matched to the cases according to sex and age bracket. Therefore, since individuals are more likely to have a heart attack at an older age, and are at higher risk for a heart attack if they are overweight and consume tobacco, it makes sense that the controls in PROMIS had similar characteristics to the MI patients themselves but different from the general population of Pakistan. One further difference between the PROMIS and DHS datasets is that 22 % of the general Pakistan population is Pashto, but no individuals with this ethnicity were recruited to PROMIS. Besides Urdu and Punjabi, the other ethnicities of PROMIS participants included Pathan, Balochi, Sindhi, Memon, and Gujrati, whereas for the overall Pakistan population the most common other ethnicities are Sindhi, Pushto, Balochi, Barauhi, Siraiki, Hindko, Shina, and Balti. Thus, the ethnicities of PROMIS participants, along with many other demographic and clinical characteristics, were not especially representative of the country as a whole.

Table 2.1: Demographic and clinical characteristics and coronary heart disease risk factors of individuals assayed by DIHRMS in PROMIS

Variable	**PROMIS controls assayed by DIHRMS** (n=5662)		**All PROMIS controls** (n=18 564)		**DHS Pakistan** (n=13 558)	
	No. of subjects	**Mean (SD) or %**	**No. of subjects**	**Mean (SD) or %**	**No. of subjects**	**Mean (SD) or %**
Anthropometric markers						
Age at survey (yrs)	5662	54 (9)	18 564	56 (9)	13 558	33 (9)
Body-mass index (kg/m^2)	5562	26 (5)	18 290	26 (5)	4698	25 (6)
Waist-to-hip ratio	5590	0.96 (0.13)	18 344	0.95 (0.06)	–	–
Systolic blood pressure (mmHg)	5587	128 (17)	18 255	128 (17)	–	–
Diastolic blood pressure (mmHg)	5584	81 (9)	18 247	81 (10)	–	–
Circulating lipid biomarkers						
Total cholesterol (mmol/L)	5542	4.63 (1.33)	17 935	4.68 (1.30)	–	–
HDL cholesterol (mmol/L)	5530	0.89 (0.27)	17 881	0.93 (0.28)	–	–
LDL cholesterol (mmol/L)	5459	2.77 (1.03)	17 491	2.81 (1.01)	–	–
Non-HDL cholesterol (mmol/L)	5530	3.75 (1.31)	17 884	3.75 (1.27)	–	–
Log$_e$ triglycerides (mmol/L)	5537	0.74 (0.53)	17 920	0.69 (0.53)	–	–
Categorical variables						
Sex	5662		18 564		13 558	
Male	4466	79 %	14 049	76 %	12 409	92 %
Female	1196	21 %	4515	24 %	1149	8 %
Ethnicity	5662		18 495		13 553	
Urdu	2113	37 %	6160	33 %	1286	9 %
Punjabi	2039	36 %	7404	40 %	3062	23 %
Other*	1510	27 %	4931	27 %	9205	68 %
Tobacco consumption status	5651		18 512		13 542	
Not current	3924	69 %	13 218	71 %	12 256	92 %
Current	1727	31 %	5294	29 %	1016	8 %
History of diabetes	5651		18 516		–	
No	4871	86 %	16 081	87 %	–	–
Yes	780	14 %	2435	13 %	–	–
Diabetic drug use status	5654		18 540		–	
No	5093	90 %	16 693	90 %	–	–
Yes	561	10 %	1847	10 %	–	–
Hypertensive drug use status	5655		18 539		–	
No	4746	84 %	15 231	82 %	–	–
Yes	909	16 %	3308	18 %	–	–
CHD risk factors						
Overweight	5562		18 290		4698	
No	2446	44 %	7830	43 %	2807	60 %
Yes	3116	56 %	10 460	57 %	1891	40 %
Obese	5562		18 290		4698	
No	4636	83 %	15 339	84 %	4031	86 %
Yes	926	17 %	2951	16 %	667	14 %
Hypertension	5587		18 257		–	
No	4600	82 %	15 017	82 %	–	–
Yes	987	18 %	3240	18 %	–	–
Diabetes (HbA$_{1c}$ definition)	4212		8503		–	
No	2600	62 %	5500	65 %	–	–
Yes	1612	38 %	3003	35 %	–	–
Diabetes (FPG definition)	5533		17 782		–	
No	3828	69 %	12 337	69 %	–	–
Yes	1705	31 %	5445	31 %	–	–

Abbreviations: BMI = Body mass index; **CHD** = Coronary heart disease; **DBP** = Diastolic blood pressure; **DHS** = Demographic & Health Surveys; **FPG** = Fasting plasma glucose; **SBP** = Systolic blood pressure; **SD** = Standard deviation.

Definitions: Diabetes (FPG) = FPG $\geq$ 126 mg/dL; **Diabetes (HbA$_{1c}$)** = HbA$_{1c}$ $\geq$ 6.5 %; **Hypertension** = SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg; **Obese** = BMI $\geq$ 30 kg/m^2; **Overweight** = BMI $\geq$ 25 kg/m^2.

Data for the overall Pakistani population were obtained from the DHS[139]. A dash (–) indicates that data were not available.

***** "Other" ethnicity for PROMIS participants includes Pathan, Balochi, Sindhi, Memon, and Gujrati. For DHS Pakistan, "Other" ethnicity category includes Sindhi, Pushto, Balochi, Barauhi, Siraiki, Hindko, Shina, and Balti.

Table 2.2: Comparison of demographic and clinical characteristics and coronary heart disease risk factors between cases and controls in PROMIS

Variable	**Controls** (n=18 564)		**Cases** (n=16 728)		***P*-value**
	No. of subjects	**Mean (SD) or %**	**No. of subjects**	**Mean (SD) or %**	
Anthropometric markers					
Age at survey (yrs)	18 564	56 (9)	16 727	54 (10)	<0.0001
Body-mass index (kg/m^2)	18 290	26 (5)	15 162	26 (4)	0.0009
Waist-to-hip ratio	18 344	0.95 (0.06)	15 420	0.97 (0.06)	<0.0001
Systolic blood pressure (mmHg)	18 255	129 (17)	16 099	126 (20)	<0.0001
Diastolic blood pressure (mmHg)	18 247	81 (10)	16 082	80 (11)	<0.0001
Circulating lipid biomarkers					
Total cholesterol (mmol/L)	17 935	4.68 (1.30)	15 405	5.04 (1.31)	<0.0001
HDL cholesterol (mmol/L)	17 881	0.93 (0.28)	15 372	0.91 (0.26)	<0.0001
LDL cholesterol (mmol/L)	17 491	2.81 (1.01)	15 025	3.24 (1.10)	<0.0001
Non-HDL cholesterol (mmol/L)	17 884	3.75 (1.27)	15 367	4.14 (1.26)	<0.0001
Log$_e$ triglycerides (mmol/L)	17 920	0.69 (0.53)	15 404	0.63 (0.55)	<0.0001
Categorical variables					
Sex	18 564		16 728		<0.0001
Male	14 049	76 %	14 060	84 %	
Female	4515	24 %	2668	16 %	
Ethnicity	18 495		16 636		<0.0001
Urdu	6160	33 %	5905	36 %	
Punjabi	7404	40 %	6007	36 %	
Other*	4931	27 %	4724	28 %	
Tobacco consumption status	18 512		16 355		<0.0001
Not current	13 218	71 %	8534	52 %	
Current	5294	29 %	7821	48 %	
History of diabetes	18 516		16 408		<0.0001
No	16 081	87 %	13 399	82 %	
Yes	2435	13 %	3009	18 %	
Diabetic drug use status	18 540		16 454		<0.0001
No	16 693	90 %	14 079	86 %	
Yes	1847	10 %	2375	14 %	
Hypertensive drug use status	18 539		16 455		<0.0001
No	15 231	82 %	10 063	61 %	
Yes	3308	18 %	6392	39 %	
CHD risk factors					
Overweight	18 290		15 162		0.1592
No	7830	43 %	6607	44 %	
Yes	10 460	57 %	8555	56 %	
Obese	18 290		15 162		<0.0001
No	15 339	84 %	13 100	86 %	
Yes	2951	16 %	2062	14 %	
Hypertension	18 257		16 100		0.0010
No	15 017	82 %	13 020	81 %	
Yes	3240	18 %	3080	19 %	
Diabetes (HbA$_{1c}$ definition)	8503		9503		<0.0001
No	5500	65 %	463	43 %	
Yes	3003	35 %	5440	57 %	
Diabetes (FPG definition)	17 782		15 343		<0.0001
No	12 337	69 %	5013	33 %	
Yes	5445	31 %	10 330	67 %	

Abbreviations: BMI = Body mass index; **CHD** = Coronary heart disease; **DBP** = Diastolic blood pressure; **DHS** = Demographic & Health Surveys; **FPG** = Fasting plasma glucose; **SBP** = Systolic blood pressure; **SD** = Standard deviation.

Definitions: Diabetes (FPG) = FPG $\geq$ 126 mg/dL; **Diabetes (HbA$_{1c}$)** = HbA$_{1c}$ $\geq$ 6.5 %; **Hypertension** = SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg; **Obese** = BMI $\geq$ 30 kg/m^2; **Overweight** = BMI $\geq$ 25 kg/m^2.

***** "Other" ethnicity includes Pathan, Balochi, Sindhi, Memon, and Gujrati.

2.4 Discussion

PROMIS has been established as a resource for the investigation and identification of emerging and established risk factors in Pakistan. Over 35 000 participants have been recruited, and information on a wide range of anthropometric measurements, demographic characteristics, lifestyle factors, lipids, clinical chemistry biomarkers, genetics, and other data are available.

Given that PROMIS used a case-control study design, it has certain inherent limitations. For instance, most of the questions on the questionnaire were self-reported, which has a potential for recall bias. Participants may lie about (or genuinely not remember) information, as was clearly evidence with self-reported age (see Figure 2.6). Participants may have also provided incorrect information about smoking habits, physical activity, the types of food that they eat and the frequency that they consume them, or other factors, though unlike with the age example, for lifestyle factors it is very difficult to detect whether there is bias or the reported lifestyle habits are genuine. The consequence of these biases is that most of the information on the questionnaire itself has to be taken with a healthy dose of scepticism. However, the laboratory measurements and genetic data are not subject to recall bias, and are therefore likely to be more reliable and accurate.

There is also a possibility of selection bias, which occurs when the outcome is influenced by differences between groups. However, in PROMIS the controls were recruited from hospital visitors and people admitted to outpatient clinics. They were likely to be fairly similar to cases, who were admitted to the hospital for an MI, so the likelihood of this type of bias is minimal. In the analyses for this dissertation, only controls were examined since lipidomics measurements were not obtained in cases. However, because only controls were analysed that had completion information on all variables used in the model, there is a potential for selection bias caused by missing data. A related concern is that although the cases and controls in PROMIS were relatively similar to each other, they were not very representative of the Pakistani population as a whole. This potentially limits the generalisations that can be made to extend the findings from this dissertation to the wider population of Pakistan and South Asia.

Another limitation that emerged due to the study design is the lack of follow-up data. Since PROMIS is a case-control study, the participants were only surveyed at baseline without any follow-up. It would be very useful to know if controls with increased levels of certain lipids, but who lacked symptoms of MI at the time of recruitment, ended up developing CHD in the period thereafter. Without follow-up data, it is not possible to determine this information.

Despite these limitations, PROMIS is a large study in a population that has not been widely studied, and has an extensive range of biological measurements available, including an exceptional quantity of high-quality genetics and lipidomics data. The use of such a befitting bioresource for this dissertation provides a unique opportunity to investigate and identify novel therapeutic targets through the study of high-dimensional phenotypic traits. The lipidomics platform that was analysed in this dissertation will be described in the next chapter.

CHAPTER 3

Processing of lipidomics data

Chapter summary

Intact lipid profiling by direct infusion high-resolution mass spectrometry (DIHRMS) was performed on 5662 serum samples from healthy PROMIS participants. A novel peak-picking algorithm was developed to identify and record signals at mass-to-charge ratios corresponding to 444 known lipids in positive and negative ionisation modes with an average coefficient of variation of 13.44 %. The lipids belonged to five overall categories: fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, and sterol lipids. This chapter describes the steps involved in assaying the serum samples on the lipidomics platform, extracting the signal data using the peak-picking algorithm, performing quality control, and producing a clean dataset for analysis. This newly developed open-profiling method is highly suitable to provide detailed lipid profiles in large-scale epidemiological studies, with a wider coverage of lipids than most other high-throughput profiling methods. The lipidomics data generated by this platform can be utilised to provide many novel insights into the effect of physiology and diet on lipid metabolism, the genetic determinants of lipids, and the relationship between lipid subfractions and CHD.

3.1 Introduction

3.1.1 Comparison of metabolomics platforms

There are important prerequisites for any method that uses profiles of serum lipids to study lipid metabolism. Since by definition, an open-profiling method does not target particular lipid classes or species within each class, it must have the ability to measure and discriminate a wide range of lipids with minimal bias across different lipid species and classes[145]. If such approaches are to be used on a large scale, such as in epidemiological studies with thousands of samples that consist of multiple batches and may run over several months, then the method needs to be high-throughput and reproducible. It is essential that the profiling method is robust so that it can deal with slight variations and fluctuations in the ionisation efficiency and ion transfer over time, and also overcome batch effects that could be introduced as a result of regular cleaning of the instrument[145].

Mass spectrometry

A variety of sample preparation and analytical techniques can be used to generate mass spectra of lipid profiles. There are two commonly used metabolomics technologies for obtaining mass spectral data. The first, mass spectrometry (MS), is a relatively inexpensive and highly sensitive technique that measures the mass-to-charge ratios (m/z) of ions formed from molecules[46]. MS is often performed following liquid or gas chromatographic separation, which reduces the complexity of the data by separating the metabolites in a time dimension (recorded as the retention time for each metabolite)[45,146,147]. MS can detect over 300 distinct metabolites using gas chromatography and over 1000 metabolites using liquid chromatography; however, since each metabolite has to be measured individually, MS is slower than other metabolomics approaches, especially when including the prior chromatographic separation step[45,146]. Furthermore, MS breaks up the samples and destroys them so that they are not usable for further analyses[45,146].

Nuclear magnetic resonance

The second approach, nuclear magnetic resonance (NMR) spectroscopy, works by applying a magnetic field to small molecules, taking advantage of the magnetic spin of certain nuclei (e.g. ^{1}H, ^{13}C, and ^{31}P) to record their energy levels using radio-frequency waves[46]. It identifies molecules by the specific pattern in the chemical shift of specific atoms[148]. NMR has a higher throughput than MS because it can measure all of the metabolites simultaneously, it is very accurate and replicable, and it does not destroy the samples during the process[45,146]. However, NMR is much less sensitive than MS because it can only detect around 30 to 100 metabolites; it is also much more expensive, and the measurements are less precise[45,146]. Although ^{1}H-NMR spectroscopy has been applied to a wide range of studies due to its reproducibility and rapid analysis speed, it cannot measure metabolites that are present in low concentrations, which require more sensitive techniques such as MS to be identified and quantified[46].

Direct infusion high-resolution mass spectrometry

Direct infusion coupled to high-resolution mass spectrometry (DIHRMS) is an alternative metabolomics approach where the sample is introduced directly into the mass spectrometer[58]. DIHRMS benefits from a rapid analysis time, has high technical reproducibility, and can detect a wide range of metabolites at a reasonable cost[45,146,147,149]. The measurements are comparable to other mass spectrometry approaches and only minimal sample biomass is needed[45,149]. It does not require a prior separation step using gas or liquid chromatography, which is a significant advantage given that it speeds up the analysis time substantially and does not destroy the biological sample if needed for further analysis[149].

Despite these significant advantages, DIHRMS does have several limitations. First, signals can be suppressed or enhanced when all the samples are introduced simultaneously into the ionisation source[149]. This results in the ionisation capability of one metabolite being modified because the charge is associated with another metabolite[45,146].

Second, DIHRMS produces complex spectra that yield *m/z* and intensity values only, without a retention time[45,149]. This makes the precise identification of lipid species challenging because chemical isomers cannot be distinguished, which would require chromatographic separation[45,146]. DIHRMS uses a fingerprinting approach which means that the lipids to be measured are not identified *a priori*, so they have to be identified on the basis of their monoisotopic mass[150]. Using DIHRMS, lipids can be identified to a molecular formula level but not beyond as it cannot separate isobaric species[149]. Therefore, for some lipids it can be difficult even to determine their subclass, for instance whether they are a phosphocholine (PC) or phosphoethanolamine (PE).

Despite the above limitation, identification at the chemical formula level using DIHRMS still allows the modelling of changes within lipids according to biological processes such as chain elongation and desaturation, while further characterisation of *m/z* peaks can be performed using LC-MS/MS or GC-MS/MS[45]. Therefore, when DIHRMS is performed with consideration to these points and selection of lipids that can be identified more readily, these limitations should not be an overwhelming cause for concern.

Regardless of the method used for obtaining mass spectral data, most metabolomics equipment is subject to non-biological artefacts in measurements, which produces a batch effect over time as each new plate is run (also known as “machine drift”)[151]. To compensate, it is necessary to include blanks and quality control (QC) samples on each plate and add internal standards to each sample to facilitate validation, cleaning, and normalisation of the results[151]. It is also important to use a balanced randomisation which helps to avoid batch effects so that any differences between lipid profiles of the various samples are real rather than artefacts[152].

3.1.2 Peak-picking to identify lipid metabolites

When using an open-profiling method, the complexity and large number of lipids that can be recorded requires that a suitable data processing approach is used to obtain relevant and specific information on the lipids in the samples. A single sample analysed using a mass spectrometer often produces tens of thousands of data points, but the actual number of lipids in the sample is typically only in the couple thousands. Therefore the majority of data points are actually extraneous, the result of instrument noise, process artefacts, and redundant ion features. This makes identifying true signals a laborious and time-consuming process with a high risk of false positives[150,151,153].

Most feature-selection algorithms are aimed at chromatographic approaches and are not suitable for direct infusion. However, a type of feature selection that is effective for this type of data is peak-picking, which is the algorithmic process of identifying a particular distribution of a signal that is peak-shaped (i.e. Gaussian). Software algorithms have been developed that aim to mimic the human process of identifying baseline signals, examining the height and shape of a peak, recording a measurement of the peak width and height (known as peak integration), and assigning the mean m/z value for that peak. Various advanced methods of peak-picking apply algorithmic filters, such as the frequently employed Gaussian filter, but signal processing methods such as wavelet analysis have also been developed, which aim to extract all possible signals in a mass spectrum[150,151]. These popular advanced methods can be prohibitively computationally intensive when applied to thousands of mass spectra, with thousands of signals to analyse per spectra, and their use is discouraged particularly when it is not entirely known what compounds are in a particular sample. However, for large-scale studies where the composition of the samples is reasonably similar (such as human serum), the vast majority of the detected signals will be similar. Therefore, a more targeted approach to signal processing can be used, focusing on lipids that can be expected in human serum, and can be readily applied to studies with large numbers of samples[58].

Peak-picking is the approach that was used in this study. A comprehensive lipid list was compiled using a well-established approach based on all lipids that could theoretically be present in human serum[48,154]. The mass spectrometer measured all compounds within a specified *m/z* window and a novel peak-picking algorithm developed for this analysis was used to extract the lipid signals corresponding to all lipids on this list. The list was then further refined based on a specific lipids that were actually detectable in this population.

3.2 Methods

3.2.1 Sample selection

Serum samples from 5674 PROMIS participants were pipetted into 1.2 mL Cryovial tubes and manually arranged into 72 boxes according to a randomised block design so that each box had no more than 80 samples. Each box corresponded to a 96-well plate, and each tube within the box was assigned a corresponding well on that plate. Block randomisation was performed using the "blockTools" package[155] in R v3.1.2[156], which assigned each of the samples to experimental blocks, with one sample per factor, by creating a measure of multivariate distance between all possible pairs of units. The factors used were participant sex, age, ethnicity, recruitment centre, and time in years since date of survey. Therefore, instead of a simple randomisation, which reorders participants in an arbitrary manner and could end up with a clustering of participants with similar values on the same plate by chance, block randomisation ensured that the values of each of the specified variables were evenly distributed both within the 80 samples on each plate as well as across the 72 plates. Thus, block randomisation is not actually random in the true sense of the word; rather, this approach ensured that all of the relevant factors were distributed in consistent proportions across all of the plates as well as within each plate[152].

A QC sample was created by pooling 100 µL of 200 randomly selected samples, which was mixed and aliquoted to be used on each plate. A subset of the pooled sample was diluted with phosphate buffered saline (PBS) solution to two different concentrations giving three different QCs (QC1 was undiluted, QC2 was 1:1 diluted, and QC3 was 1:3 diluted). Samples were sent to the Medical Research Council (MRC) Human Nutrition

Research (HNR) laboratory in Cambridge, UK, for lipid extraction and data acquisition using DIHRMS.

3.2.2 Lipid extraction

An automated method for the extraction of lipids was developed using an Anachem Flexus automated liquid handler (Anachem, Milton Keynes, UK). A total of 80 samples, four blanks, and twelve QC samples in 1.2 mL Cryovials were placed on the Flexus, then 100 µL of MilliQ H_2O was added to each of the wells and mixed, and then 100 µL of the mixture was transferred to a glass-coated 2.4 mL deep well plate (Plate+^{TM}, Esslab, Hadleigh, UK). Next, 250 µL of MeOH was added containing six internal standards (0.6 µM 1,2-di-*O*-octadecyl-*sn*-glycero-3-phosphocholine, 1.2 µM 1,2-di-*O*-phytanyl-*sn*-glycero-3-phosphoethanolamine, 0.6 µM C8-ceramide, 0.6 µM *N*-heptadecanoyl-*D*-*erythro*-sphingosylphosphorylcholine, 6.2 µM undecanoic acid, 0.6 µM trilaurin), followed by 500 µL of methyl *tert*-butyl ether (MTBE). The plates were then sealed using Corning aluminium micro-plate sealing tape (Sigma Aldrich Company, UK) and shaken for 10 min at 600 rpm, after which the plate was transferred to a centrifuge and spun for 10 min at 6000 rpm. Each well in the resulting plate had two layers, with an aqueous layer at the bottom and an organic layer on top. A 96-head micro-dispenser (Hydra Matrix, Thermo Fisher Ltd, Hemel Hampstead, UK) was used to transfer 25 µL of the organic layer to a glass coated 240 µL low well plate (Plate+^{TM}, Esslab, Hadleigh, UK), and 90 µL of MS-mix (7.5 µM NH_4Ac IPA:MeOH [2:1]) was added using a Hydra Matrix, after which the plate was sealed and stored at −20 °C until analysis.

3.2.3 Data acquisition

All samples were infused into an Exactive Orbitrap (Thermo, Hemel Hampstead, UK) using a Triversa Nanomate (Advion, Ithaca, US). The Nanomate infusion mandrel was used to pierce the seal of each well before analysis, after which, with a fresh tip, 5 µL of sample was aspirated, followed by an air gap (1.5 µL). The tip was pressed against a fresh nozzle and the sample was dispensed using 0.2 psi nitrogen pressure. Ionisation was achieved with a 1.2 kV voltage. The Exactive started acquiring data 20 seconds after sample aspiration began. After 72 seconds of acquisition in positive mode, the Nanomate and the Exactive switched over to negative mode, decreasing the voltage to −1.5 kV. The spray was maintained for another 66 seconds, after which the analysis was stopped and the tip discarded, before analysis of the next sample began. Throughout the analysis the sample plate was kept at 15 °C. Samples were run in row order and repeated multiple times if necessary to ensure accuracy. The mass spectrometer had a resolution of 65 000 at 400 *m/z* with an average mass accuracy error of 0.85 ppm in the measurement of the *m/z* across all intact lipids.

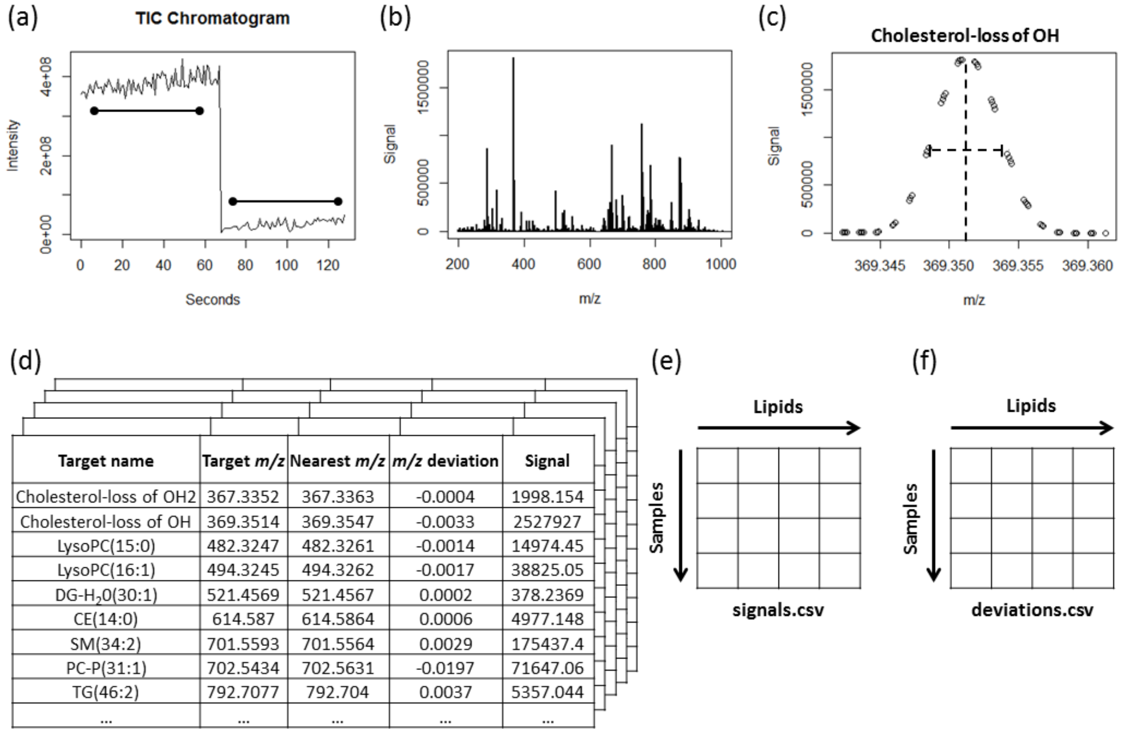
3.2.4 Data processing and peak-picking

The Exactive provided spectral data in a compressed proprietary raw format. It was necessary to repeat the assay for 259 samples where the first run produced poor quality spectra. For these 259 duplicate samples, the sample with the highest quality mass spectral data was retained, which was determined from visual inspection of the raw spectra. In the event that the spectra appeared similar so that it was not possible to determine which one to keep, the run with the latest time-stamp was retained. Once a clean list of 96 raw files was obtained for each of the 72 plates, including blanks and QC samples, the files were decompressed and converted to an open-source spectral format (mzXML) using the "msconvert" tool in ProteoWizard[157,158]. This tool was called from a Python[159] script to take advantage of multi-threading, which allowed multiple samples to be processed in parallel on a Linux server. The computing cluster that was used for these analyses was hosted by the High Performance Computing Service (HPCS) at the University of Cambridge. The compressed raw data were approximately 180 GB, which expanded to almost one terabyte when uncompressed and converted.

For each infusion an average spectrum was calculated from a user-defined retention time window, which was set at 20 to 70 seconds for positive ionisation mode and 95 to 145 seconds for negative ionisation mode. The *m/z* window was set at 185 to 1000 for both ionisation modes. The R[156] package "xcms"[51] was used to average fifty spectra per mode, which was also called from a Python script on the same Linux server so that multi-threading could be used. XCMS software[51] has become a popular tool for peak alignment, peak detection, and performing other aspects of mass spectrometry data processing, and has been used in many non-targeted metabolomics studies[70,160–165].

The list of *m/z* values, based on expected and possible lipids in serum, was extracted to a file containing comma-separated values (CSV) and used to extract small windows of data around the target *m/z* in the average spectrum. The peak maximum was recorded and the two closest points to the half-height of the peak on either side were identified, yielding a total of four points. The points with which a horizontal line at half-height intersected a line connecting the two points on either side of the peak (one above the half-height of the

Figure 3.1: Schematic of the peak-picking process



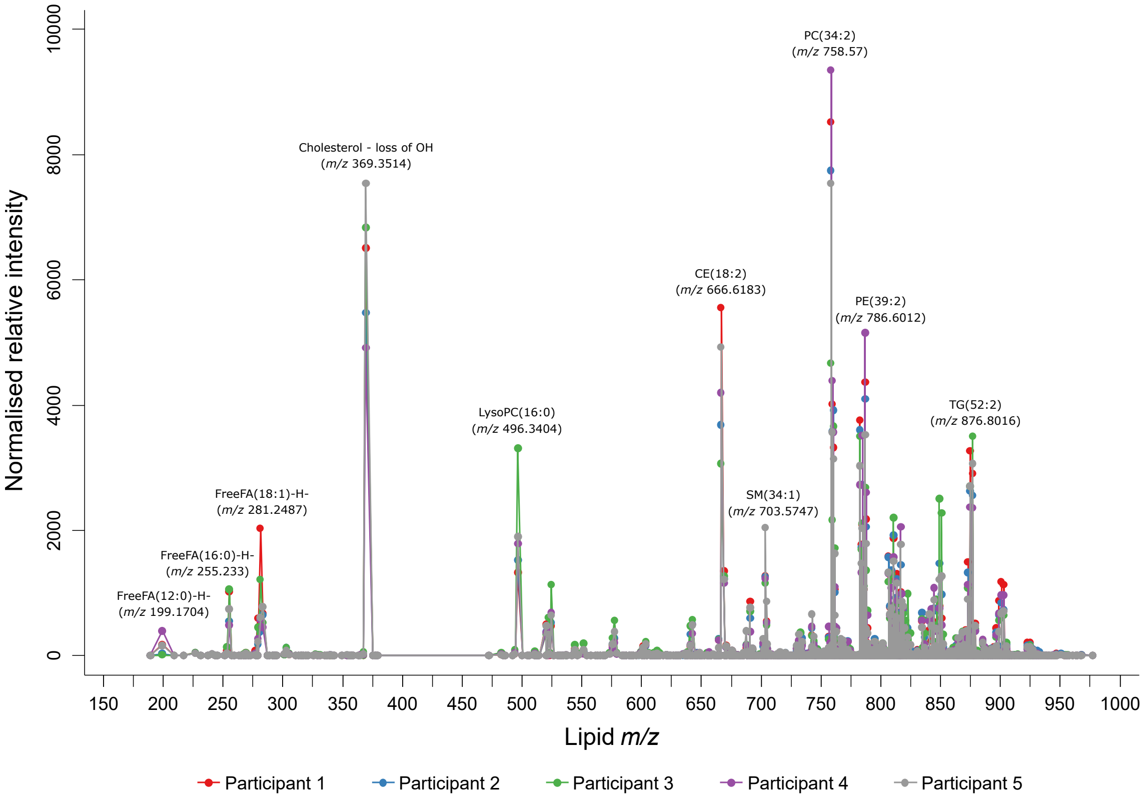
Target name	Target *m/z*	Nearest *m/z*	*m/z* deviation	Signal
Cholesterol-loss of OH2	367.3352	367.3363	-0.0004	1998.154
Cholesterol-loss of OH	369.3514	369.3547	-0.0033	2527927
LysoPC(15:0)	482.3247	482.3261	-0.0014	14974.45
LysoPC(16:1)	494.3245	494.3262	-0.0017	38825.05
DG-H$_2$0(30:1)	521.4569	521.4567	0.0002	378.2369
CE(14:0)	614.587	614.5864	0.0006	4977.148
SM(34:2)	701.5593	701.5564	0.0029	175437.4
PC-P(31:1)	702.5434	702.5631	-0.0197	71647.06
TG(46:2)	792.7077	792.704	0.0037	5357.044
...	...	...	...	...

(a) XCMS was used to average fifty spectra in positive and negative ionisation modes, yielding (b) the average mass spectrum for that particular polarity, for which signals were obtained using a peak-picking algorithm that determined the (c) peak signal at the midpoint of a line drawn at half-height for peaks near signals that correspond to known lipids. Signals and deviations that represented known lipids were then (d) combined in a database, and split into separate files for (e) signals and (f) deviations for each lipid. Source: Figure produced by Luke Marney for manuscript by Harshfield E, et al. (Manuscript under review; see Appendix B).

peak and one below) was used to obtain a peak width calculation (distance of the line) and a more accurate m/z value for the peak maximum (midpoint of the line). For all the m/z values, the maximum intensity was recorded as well as the deviation of the peaks' accurate m/z. This was performed independently for each sample and run in parallel. The final step was the combination of all the signals into one CSV file and the deviations in a second CSV file. The technical set-up yielded average deviations of less than 4 ppm ($1\,000\,000 \times m/z$ deviation / m/z target). An overview schematic of the peak-picking process is shown in Figure 3.1.

As an illustrative example, lipid profiles for five PROMIS participants are shown in Figure 3.2. These five individuals were chosen to be representative of the whole set of participants, while allowing the differences in lipid profiles between individuals to be more readily visible. The intensities are shown in a different colour for each participant. The intensities were measured on a continuous scale across the mass spectrum, but peaks corresponding to the various lipids can easily be seen. The intensity values at each of these

Figure 3.2: Illustrative example showing relative intensities of mass spectra for five PROMIS participants



The normalised relative intensities for each lipid (calculated for each participant by dividing the signal for each lipid by the total signal of all the lipids) are shown for five randomly selected participants. Several of the lipids with the highest peaks for one or more participants are labelled according to their name and m/z value.

peaks were then extracted according to the peak-picking process as described above.

The signals recorded for each sample are relative intensities, which represent the relevant abundance of the ions that correspond to each lipid. Since most of the ions have a single charge, the m/z value is usually equivalent to the mass of the ion itself. By subtracting the mass of the adduct ion (e.g. $[M+H]^+$, $[M+NH_4]^+$, $[M-H]^-$, or $[M+OAc]^-$) from the observed peak, one can determine the exact mass of the lipid and thereby identify its molecular structure and the number of carbon atoms and double bonds that it contains.

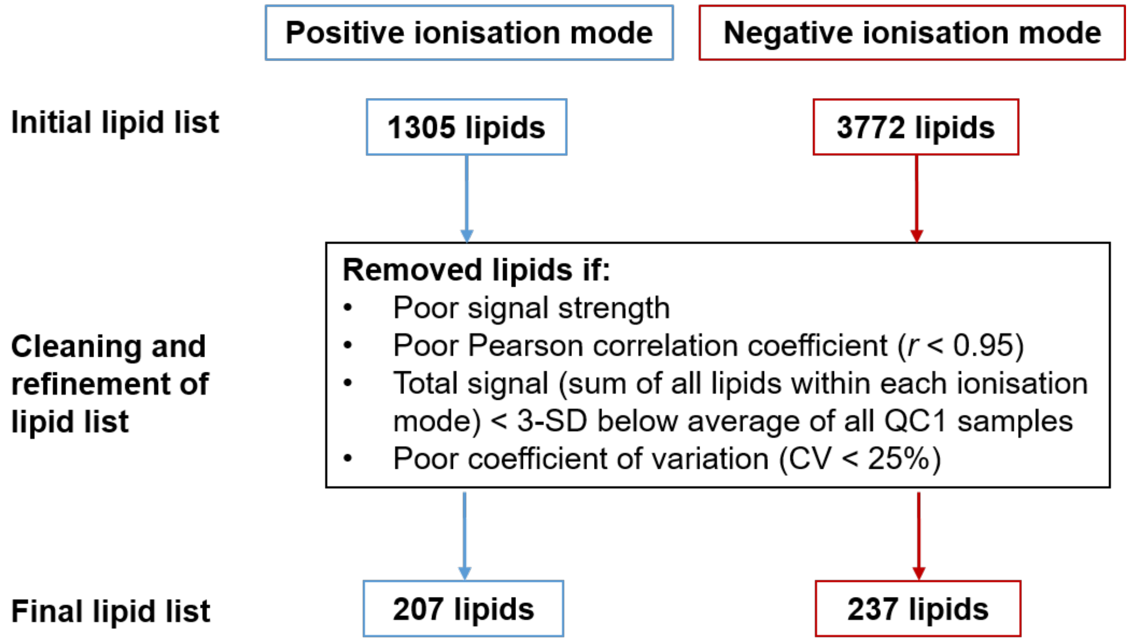
Because the open-profiling approach used in DIHRMS does not predetermine which lipid species will be detected, data are provided on all ionisable molecules and the assay is therefore very sensitive to contamination, especially of compounds with high ionisation efficiencies. All plates contained a presence of adipates (m/z 371.316) and organophosphates such as Tris(2,4-di-*tert*-butylphenyl) phosphite (m/z 647.459) and its oxidation products (m/z 663.454), which leach from plastics into organic solvents. However, the use of glass-coated well-plates minimised the contact time of the samples with the plastics, and using blanks and three QC levels, the contamination ions (approximately half of all the signals) were able to be excluded from the final data set. The use of glass-coated well-plates was therefore essential to obtain both precise and reliable data. Furthermore, as the method relies on nano-flow, the contaminations had minimal impact on the ionisation efficiency.

The peak-picking algorithm described above enabled the extraction of signals and deviations from 1305 lipids in positive ionisation mode and 3772 lipids in negative ionisation mode for the 5674 participants plus blanks and QC samples, resulting in more than 69 million retrieved data values and nearly two terabytes of uncompressed data. Since the method processed each file independently and could run the analyses in parallel, there was no requirement to load all the files jointly into memory to perform the alignment, which greatly increased processing speeds. In contrast, most other metabolomics processing methods require loading the entire dataset for all metabolites and all samples into memory, which requires an extensive amount of processing power and takes a long time to perform the calculations. However, this peak-picking approach is only suitable to compare samples that are very similar and where the same lipids are expected for each individual, since it requires an input file with the expected m/z (target). The results also required manual QC as in certain cases the m/z target could be too close to an isotope or adduct of another lipid. The identities of all the ions that passed QC were therefore confirmed.

3.2.5 Refinement of lipid list

Although the original list of all known lipids consisted of 1305 lipids in positive ionisation mode and 3772 lipids in negative ionisation mode, not all of these lipids were actually detectable in the PROMIS serum samples using the lipidomics platform. Quality control was therefore performed to remove lipids that were undetectable or did not align with internal standards. All QC samples with poor signal strength, all contaminated blanks, lipids with fewer than 700 detectable peaks, lipids that had a poor correlation with the QC samples (Pearson correlation $r < 0.95$), and lipids where the total signal was less than three standard deviations below the average of all the QC1 samples were removed. The coefficient of variation (CV) for each QC sample was then determined for each lipid, and the lowest of the three QC samples at each dilution was kept. Any lipids with a CV of less than 25 % were omitted from further analysis. An overview of the QC steps that were performed is shown in Figure 3.3.

Figure 3.3: QC steps performed to obtain final list of 444 lipid metabolites


Positive ionisation mode
Negative ionisation mode
Initial lipid list
1305 lipids
3772 lipids
Cleaning and refinement of lipid list
Removed lipids if:
• Poor signal strength
• Poor Pearson correlation coefficient (r < 0.95)
• Total signal (sum of all lipids within each ionisation mode) < 3-SD below average of all QC1 samples
• Poor coefficient of variation (CV < 25%)
Final lipid list
207 lipids
237 lipids

3.2.6 Data cleaning and quality control

The total signal was calculated separately for the lipid metabolites in each ionisation mode as the sum of the signals of all lipids in that mode. Participants were excluded from analysis in a particular ionisation mode if the total signal for the lipids in that mode was less than 5 000 000 (relative units), which indicated a poor infusion. This cut-off was based on the lower QC value so that only samples with proven robustness were included. Each lipid was normalised by expressing the signal as a proportion of the total signal for each participant. This was applied to reduce the batch effect and facilitate meaningful comparisons of relative intensities of lipids across different plates. Since the distributions of the normalised signals for most of the lipids showed approximate log-normality, natural log-transformation was applied to each lipid. Lipid signals for individual participants were considered outliers and thus excluded if the normalised, log-transformed signal was more than 10 standard deviations (SD) from the mean for that lipid across all the participants.

An assumption of the models used to run the GWAS is that the data are approximately normally distributed, so it is common practice in most metabolomics analyses to deal with outliers. Many metabolomics studies treat any values more than 3 to 5 SD from the mean as outliers[75,80,83,166–169]. However, since the majority of PROMIS participants were not fasting at time of blood draw, their lipid levels could have spiked if they had recently consumed a high-fat meal, so to avoid being overly conservative, a cut-off of 10-SD from the mean was used. It is unlikely that lipids would actually have true values more than

10-SD from the mean, so these measurements were either below the lower limit of detection (which would be expected when the distribution of the measurements for that particular lipid showed that it could be expected to overlap with the limit of detection) or the measurement was false, perhaps due to a contaminant or another unknown reason. For the GWAS analyses (described in Chapter 5) it was reasonable to set these outliers to missing to ensure that the assumptions of approximate log-normality held, but for the Gaussian Graphical Modelling (GGM) analyses (described in Chapter 4 and Chapter 6), which required complete cases, all missing values were imputed with the median value for that lipid.

3.2.7 Scatter plot of lipid signals

In order to observe the distribution of each lipid across the full set of participants, the normalised relative intensities of each lipid for each participant were plotted across the *m/z* spectrum. These were displayed as a scatter plot with the *m/z* of each lipid on the x-axis, and the normalised relative intensity of the signal of each lipid on the y-axis.

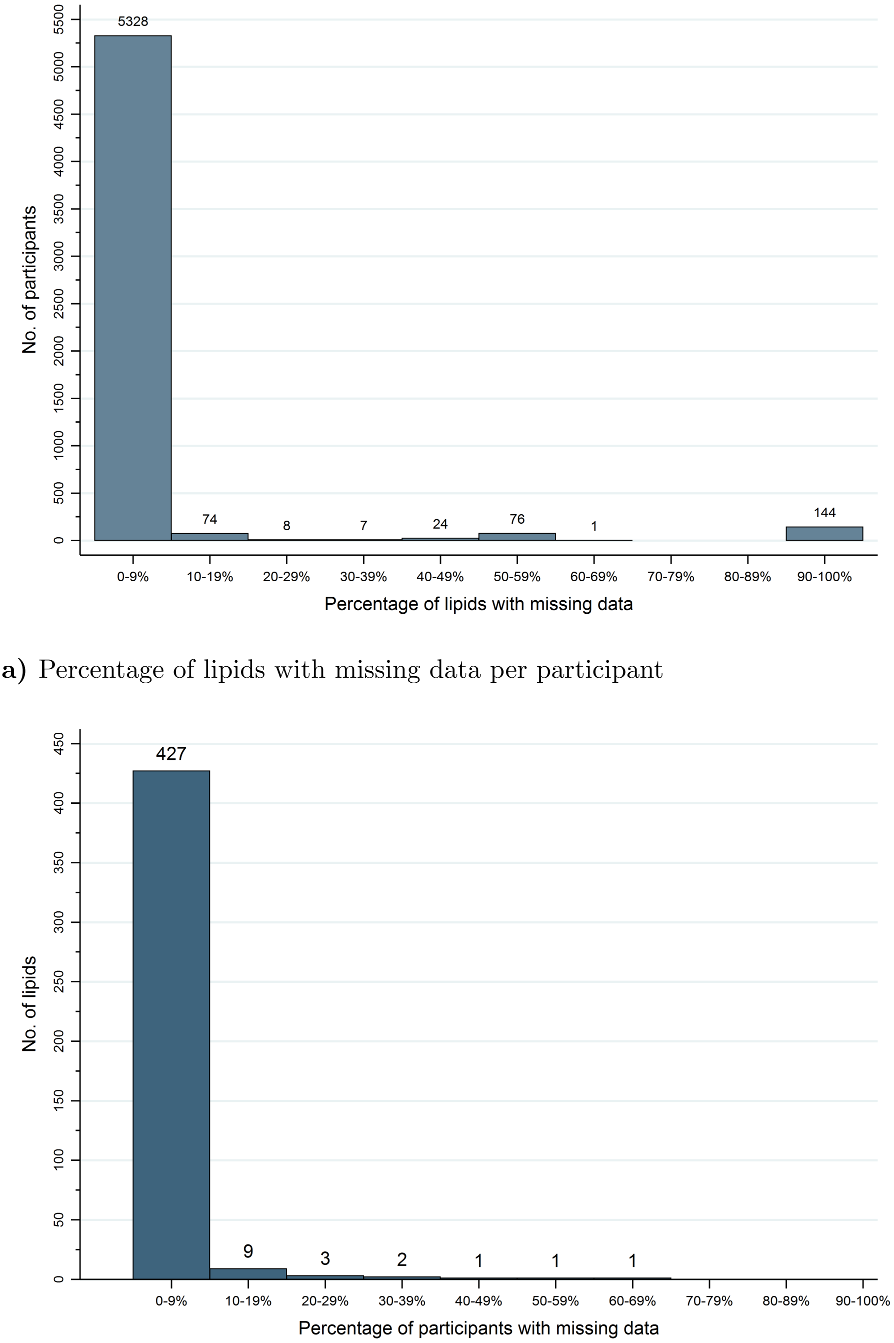
3.3 Results of lipidomics processing

Lipid profiles obtained using DIHRMS were available for 5662 PROMIS participants following data processing, cleaning, and QC; 123 out of the original 5674 samples (2 %) were excluded during the QC stage. The extent of missing data is shown in Figure 3.4. As shown in Figure 3.4a, 5328 participants (94 %) were missing data from less than 10 % of the 444 lipids, and as Figure 3.4b shows, 427 lipids (96 %) were missing data in less than 10 % of the participants. There were only 17 lipids that were missing data in more than 10 % of the participants.

A scatter plot showing the normalised relative intensities of all the lipids according to their *m/z*, grouped by participant, is shown in Figure 3.5. The wide distribution of the signals across certain lipids, such as cholesterol with loss of −OH, CE(18:2), PC(34:2), and TG(52:2) (see Figure 3.2), suggests that the levels of these lipids vary significantly across individuals, whereas the levels of other lipids are more consistent.

A distinct advantage of DIHRMS is its low cost (approximately £10 per sample) and rapid data collection. The high throughput of the method means that with an analysis time of just over two minutes per sample, it is possible to run a full plate within four hours. The automated sample preparation of one plate is possible in 1.5 hours, which makes this approach especially useful for large-scale lipid profiling.

Figure 3.4: Extent of missing data according to the number of lipids and participants

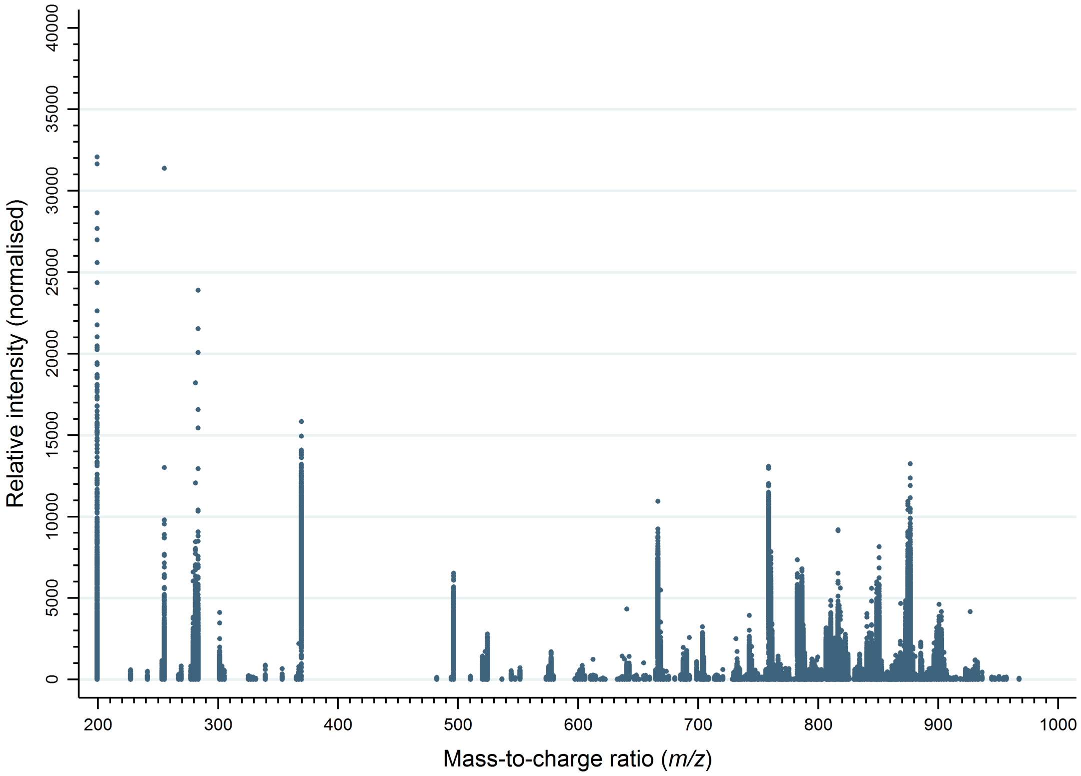


(a) Percentage of lipids with missing data per participant

(b) Percentage of participants with missing data per lipid

Figure 3.4a shows the number of participants with lipidomics data according to the proportion of lipids with missing data (e.g. 5328 participants had between 0 % to 9 % missing data). Figure 3.4b shows the number of lipids that were measured according to the proportion of participants with missing data (e.g. 427 lipids had complete signal data measured in all except 0 % to 9 % of participants).

Figure 3.5: Normalised relative intensities of lipid metabolites



3.3.1 Lipid classification

This DIHRMS method for lipidomics covers a wide range of lipids, including fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, and sterol lipids (Figure 3.6), and does not require a prior selection of specific lipids or lipid classes, in contrast to fragmentation-based approaches using tandem mass spectrometry.

The DIHRMS method included measurement of neutral lipids such as triglycerides and cholesteryl esters, which are not covered by the commercial platforms that are currently most widely used in large-scale metabolite phenotyping for genome-wide association studies[74]. This study included analysis of approximately 125 metabolic features that have not yet been assessed in any of the major genome-wide association studies of human metabolism[74]. Additionally, the platform included measurement of lipids that contain odd-chain fatty acids, which have generally been ignored in previous metabolic profiling efforts[170], while this analysis shows that these lipids are important to human metabolism.

The classification scheme that was employed for the lipids measured by this platform is based on the taxonomy used by a number of lipid databases, including Lipid Metabolites and Pathways Strategy (LIPID MAPS, http://www.lipidmaps.org)[62,154], Swiss Lipids (http://www.swisslipids.org)[171], and the Human Metabolome Database

Figure 3.6: Classification of lipid metabolites according to overall lipid category, main class, and subclass

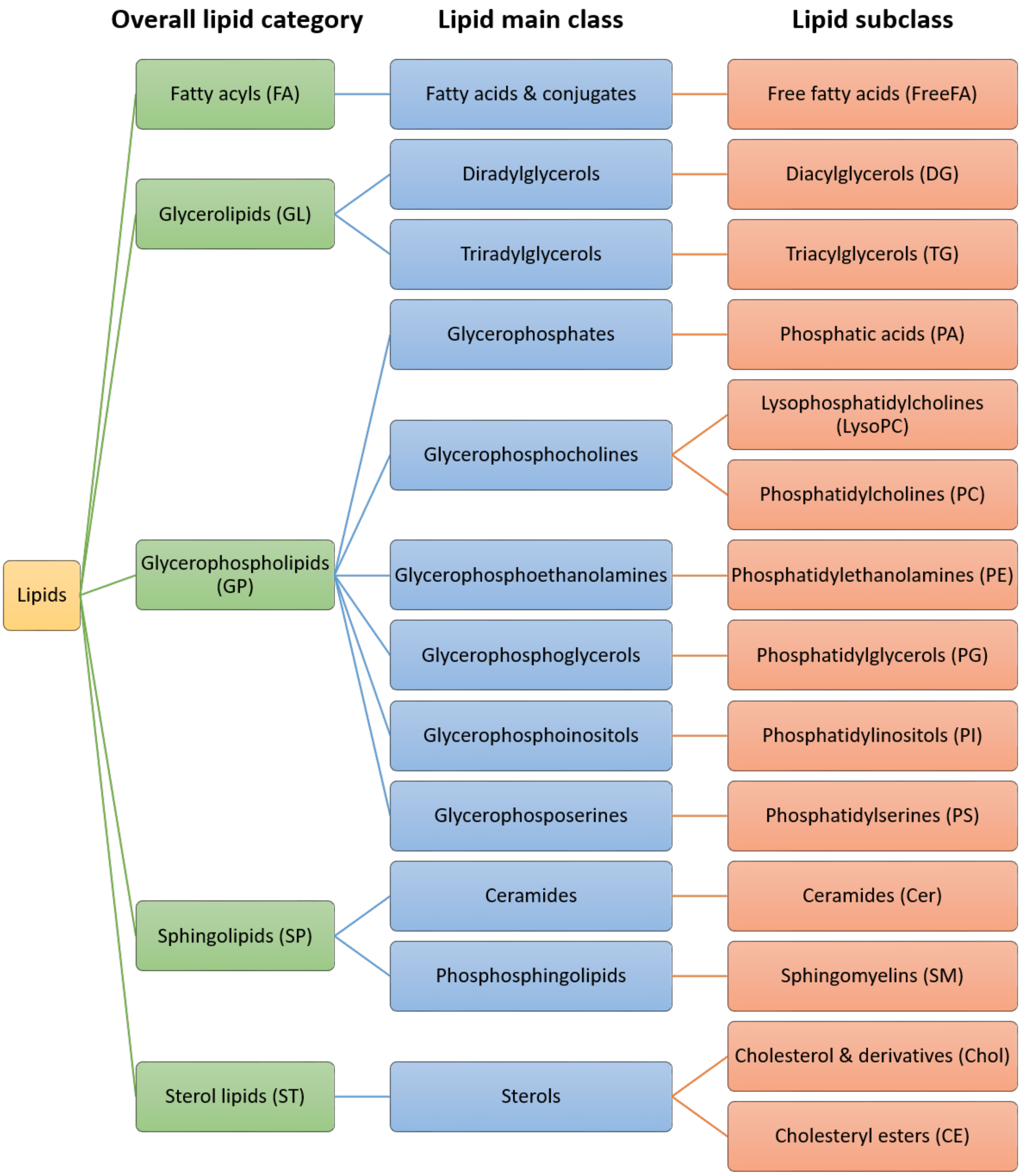


Table 3.1: Categorisation of lipid metabolites in positive and negative ionisation mode measured by DIHRMS in PROMIS

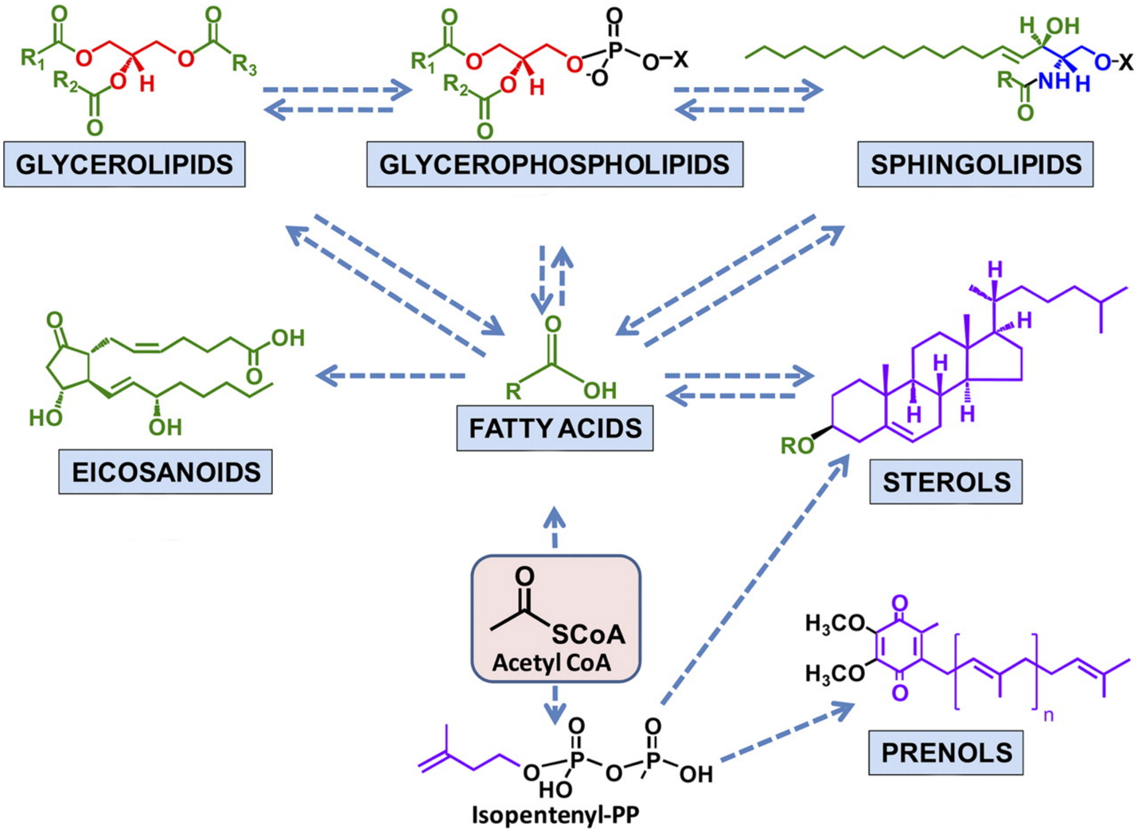
Overall lipid category	**Lipid subclass**	**No. (%) of lipid metabolites**
Fatty acyls (FA)	Free fatty acids (FreeFA) (−)	22 (5.0)
Glycerolipids (GL)	Diacylglycerols (DG) (+)	19 (4.3)
	Triacylglycerols (TG) (+)	56 (12.6)
Glycerophospholipids (GP)	Lysophosphatidylcholines (LysoPC) (+)	8 (1.8)
	Phosphatic acids (PA) (−)	20 (4.5)
	Phosphatic acids (PA) (+)	13 (2.9)
	Phosphatidylcholines (PC) (−)	52 (11.7)
	Phosphatidylcholines (PC) (+)	54 (12.2)
	Phosphatidylethanolamines (PE) (−)	24 (5.4)
	Phosphatidylethanolamines (PE) (+)	16 (3.6)
	Phosphatidylglycerols (PG) (−)	5 (1.1)
	Phosphatidylinositols (PI) (−)	25 (5.6)
	Phosphatidylserines (PS) (−)	22 (5.0)
Sphingolipids (SP)	Ceramides (Cer) (−)	16 (3.6)
	Sphingomyelins (SM) (−)	51 (11.5)
	Sphingomyelins (SM) (+)	27 (6.1)
Sterol lipids (ST)	Cholesterol & derivatives (Chol) (+)	2 (0.5)
	Cholesteryl esters (CE) (+)	12 (2.7)
Total lipid metabolites		**444 (100%)**

(+) denotes lipid metabolite measured in positive ionisation mode; (−) denotes lipid metabolite measured in negative ionisation mode.

(HMDB, http://www.hmdb.ca). While there are slight differences between these classification schemes, they all include the five overall lipid categories that were measured using DIHRMS, which are then subdivided into various lipid classes and subclasses (Figure 3.6). For instance, two of the lipid subclasses that were measured using the platform, lysophosphatidylcholines and phosphatidylcholines, are both a type of glycerophosphocholine, which in turn is one of the six types of glycerophospholipids that were measured using the platform. The number of lipids that were measured within each lipid subclass, in both positive and negative ionisation modes, are listed in Table 3.1. It is worth noting that beyond the five overall categories of lipids listed here, there are also several broad lipid categories that were either not measured by this platform or were not detectable in this population—these include prenol lipids (such as Vitamin E and Vitamin K), saccharolipids (such as disaccharides), and polyketides (such as erythromycins and tetracyclines).

As a brief introduction to lipid biology and structure, the five overall lipid categories and their constituent lipid classes and subclasses will be described. The structures of these classes are shown in Figure 3.7. Lipids in the first category, *fatty acids*, consist of a hydrocarbon chain that terminates with a carboxylic acid group. Fatty acids are commonly used as building blocks of more structurally complex lipids. A few well-known examples are linoleic acid, alpha-linoleic acid, eicosapentaenoic acid (EPA), and docosahexaenoic

Figure 3.7: Chemical structure of overall lipid categories



Relationships among the major lipid categories are shown starting with the 2-carbon precursor acetyl CoA, which is a building block in the biosynthesis of fatty acids. Fatty acyl substituents in turn may be transferred to form complex lipids, namely sphingolipids, glycerolipids, glycerophospholipids, and sterols (as steryl esters). Some fatty acids are converted to eicosanoids. A second major biosynthetic route from acetyl CoA generates the 5-carbon isoprene precursor isopentenyl pyrophosphate, which provides the building blocks for the prenol and sterol lipids. Fatty acyl-derived substituents are coloured green; isoprene-derived atoms are coloured purple; glycerol and serine-derived groups are coloured red and blue, respectively. Arrows denote multi-step transformations among the major lipid categories starting with acetyl CoA. Source: Figure and caption are adapted from Quehenberger O, et al. *J Lipid Res.* 2010;51(11):3299-3305[62].

acid (DHA).

Diacylglycerols and triacylglycerols consist of two or three fatty acid chains, respectively, bound to a glycerol backbone. These fall into the second category, *glycerolipids*, which are especially important for storing fat.

The third category, *glycerophospholipids*, are phospholipids with a glycerol backbone, and are major constituents of membrane bilayers. A phospholipid consists of two hydrophobic fatty acid tails and a hydrophilic head, which is a phosphate group. The phosphate groups can be modified with simple organic molecules such as choline or ethanolamine. For instance, phosphatidylcholines are simply glycerophospholipids with choline head groups, while phosphatidylethanolamines are glycerophospholipids with ethanolamine head groups.

The fourth category, *sphingolipids*, are a class of lipids that contain a backbone of sphingoid bases. A ceramide is a specific type of sphingolipid where the head group consists of only a hydrogen atom. Finally, *sterol lipids* are another important component of cellular membranes. While all steroids are derived from the same fused four-ring core structure, which consists of three cyclohexane rings and one cyclopentane ring, they can have different biological roles such as hormones and signalling.

Each lipid metabolite can be classified according to the total number of carbon atoms and double bonds that it contains. This is denoted by a:b, where a is the number of carbons in the side chain(s) and b is the number of double bonds. As an example, CE(14:0) (m/z 614.587) is a cholesteryl ester that contains 14 carbon atoms and zero double bonds, so it is fully saturated. It only has one side chain which consists of a single fatty acid, FA(14:0), also known as myristic acid. Another example is PC(34:2) (m/z 758.57), which consists of 34 carbon atoms with two double bonds so it is polyunsaturated. While the most likely combination of fatty acid chains that makes up this phosphocholine is [FA(16:0/18:2)] (palmitic acid and linoleic acid, respectively), it is possible that other combinations of fatty acids could exist for a lipid with this m/z. A more complex example is TG(54:4) (m/z 900.8015), which contains three fatty acid chains. The most likely combination of fatty acids for a triglyceride with this m/z is [FA(18:0/18:2/18:2)] (stearic acid and two linoleic acids); however, it is also quite possible that this triglyceride could consist of a different combination of fatty acids, such as [FA(16:0/18:0/20:4)] (palmitic acid, stearic acid, and arachidonic acid).

As the above examples illustrate, a drawback of the DIHRMS assay is that the precise fatty acids that make up each lipid cannot be determined, nor can the location of the double bonds on each fatty acid chain be identified. However, despite this limitation, the lipids can be grouped according to the total number of carbon atoms and double bonds on the entire lipid. Figure 3.8 shows each lipid plotted according to the number of carbon atoms and double bonds that make up that lipid. Lipids measured in positive ionisation mode are shown in blue, while lipids measured in negative ionisation mode are shown in red. The lipids are plotted semi-transparently, so that a darker shade of blue indicates that multiple lipids with the same number of carbon atoms and double bonds in positive ionisation mode exist at a single location, while a darker shade of red indicates a large number of lipids in negative ionisation mode with the same number of carbon atoms and double bonds. In contrast, a purple colour indicates that lipids in both positive and negative ionisation mode with the same number of carbon atoms and double bonds were measured, so these are plotted on top of each other. The lipids have been grouped into the five overall lipid categories.

The majority of fatty acyls were fully saturated (i.e. zero double bonds), while the majority of sphingolipids were either fully saturated, monounsaturated (with one double bond), or polyunsaturated with two double bonds. For all of the lipid classes, but especially for the glycerolipids and glycerophospholipids, a clear pattern emerged, which is that as the number of carbon atoms increased, the number of double bonds also increased. Figure 3.8 confirms the information listed in Table 3.1, namely that all of the glycerolipids and sterol lipids were detected only in positive ionisation mode, while all of the fatty acyls were detected only in negative ionisation mode. The glycerophospholipids and sphingolipids, meanwhile, consisted of a range of lipids in both positive and negative ionisation modes.

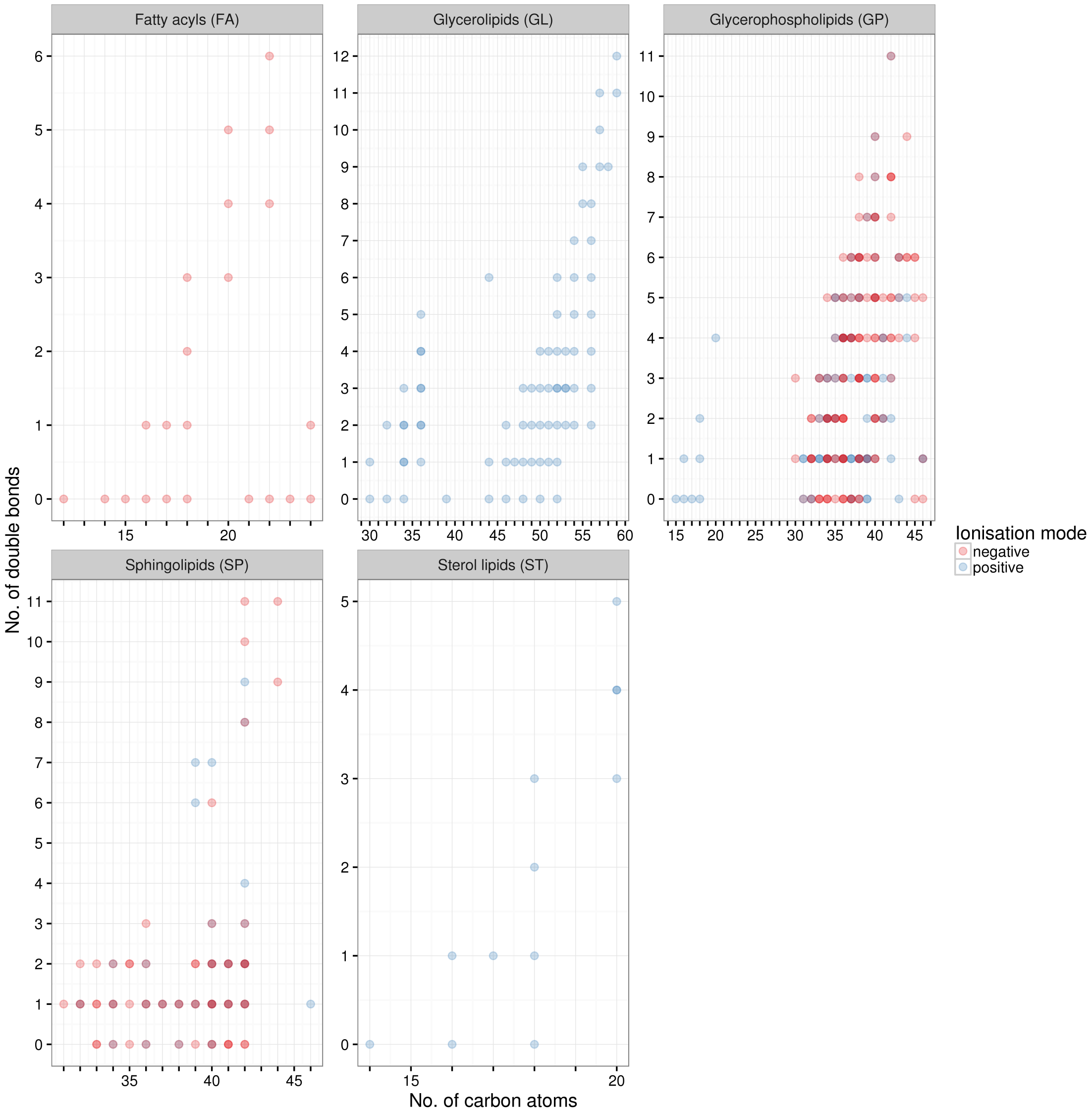
3.3.2 Correction for batch effect

There was a significant observable batch effect (machine drift) in the pre-normalised (raw) data across the 72 plates that were measured using DIHRMS. The batch effect was relatively minor for the first few plates, but became more pronounced after plate 14 and especially apparent from plate 23 upwards, as seen in the box plots (Figure 3.9) and scatter plots (Figure 3.10) of the raw (pre-normalised) relative intensities for each plate. This can also be seen in the plot of the first two principal components for each plate (Figure 3.11), which together explained 69 % of the variance in the relative intensities (PCA analysis will be described in Chapter 4). In the scatter plot and PCA plot, the QC samples and blanks are represented by black tick marks, while the samples on each plate are each portrayed in different colours. As each successive plate was run, the variance in the relative intensities became more prominent. It is likely that a significant contributor to the observed batch effect, particularly the extreme shift in the average relative intensity between plates 22 and 23, was the cleaning of the instrument heads. However, after normalising the lipids by dividing the signal of each lipid for each participant by the total signal (sum of all the lipids) for that participant, the batch effect was no longer apparent (Figure 3.12) since the average relative intensity was consistent across each plate.

3.3.3 Quality control results

Quality control of the lipid metabolites resulted in a list of 207 lipids in positive ionisation mode and 237 lipids in negative ionisation mode, all with unique mass-to-charge ratios and identifiers. However, since no chromatography step was used in the method, each m/z could theoretically represent a number of individual isobaric lipid species. The CVs for each lipid that were retained in each ionisation mode are shown in Figure 3.13. The precision was higher in positive mode (average CV 13.01 %, median CV 11.48 %) than in

Figure 3.8: Number of carbon atoms and double bonds for each lipid by ionisation mode within each lipid class



The lipids are plotted semi-transparently so that circles with darker shades of colour indicate that multiple lipids in either positive (blue) or negative (red) ionisation mode were measured with the same number of carbon atoms and double bonds. A purple colour indicates that lipids with the same number of carbon atoms and double bonds were measured in both positive and negative ionisation mode, which are plotted on top of each other.

Figure 3.9: Box plots of raw (pre-normalised) relative intensities of lipids for each participant across each plate

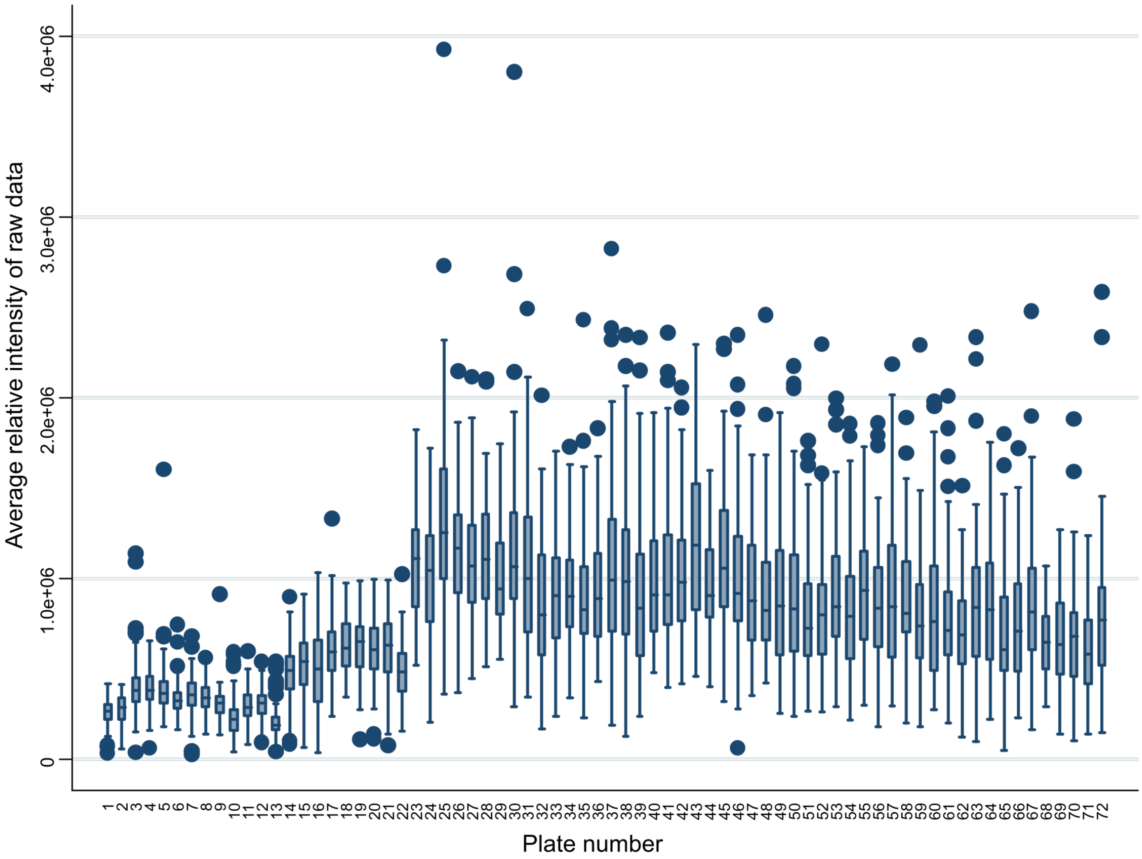
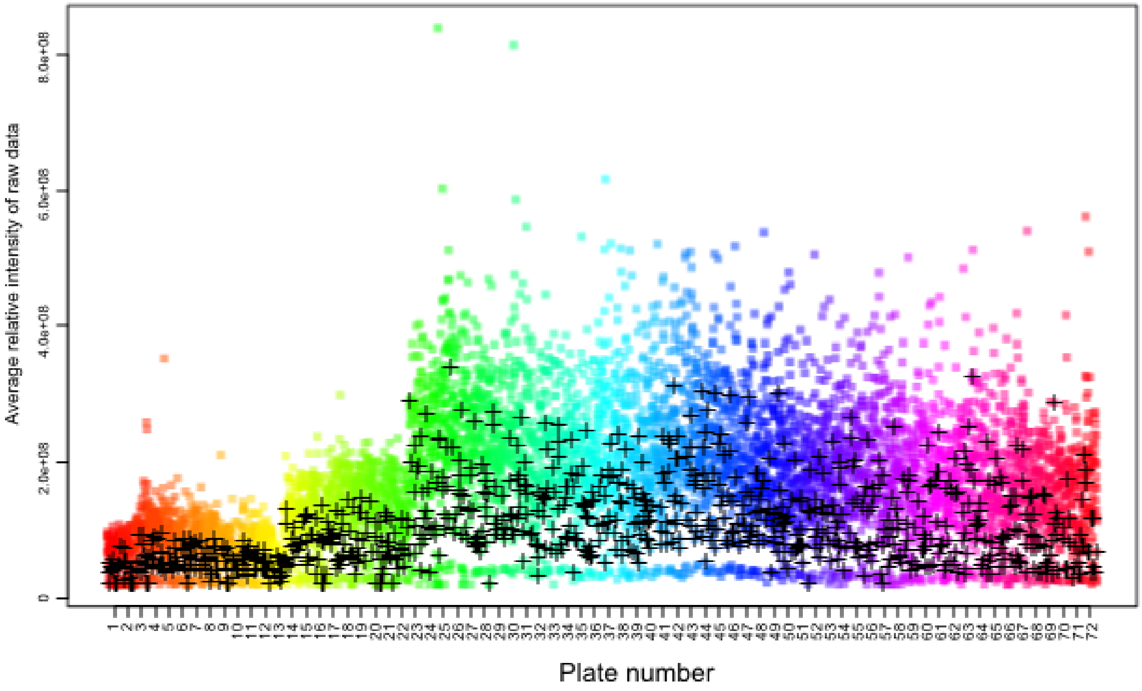
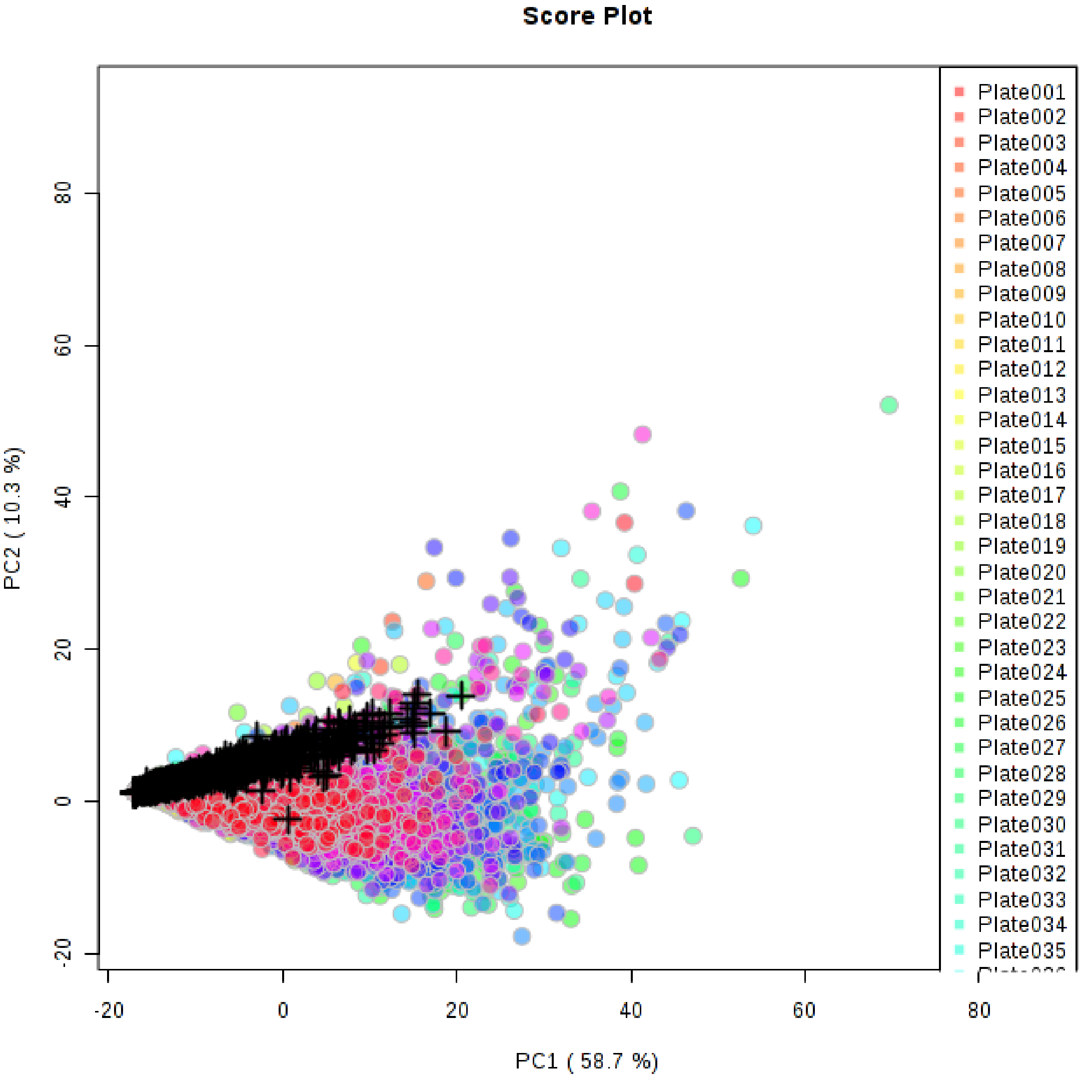


Figure 3.10: Scatter plot of raw (pre-normalised) relative intensities of lipids for each participant across each plate



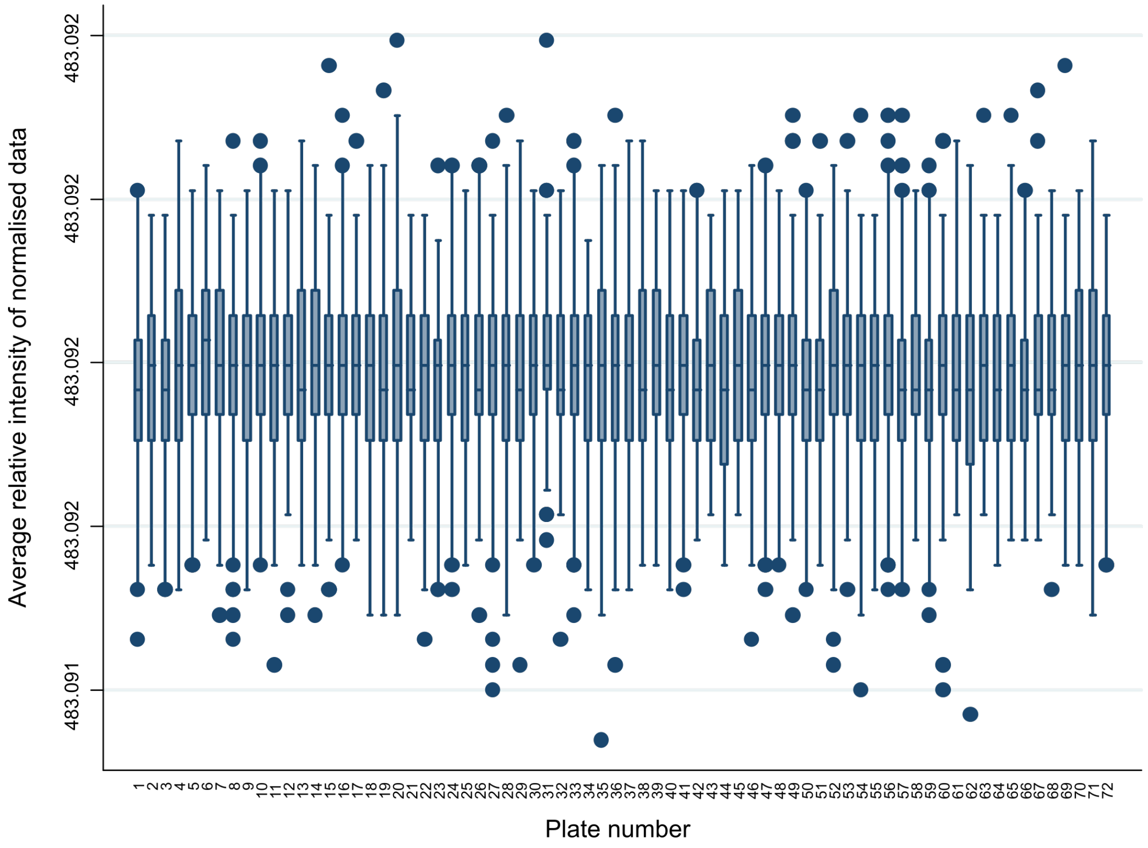
This figure was produced using MetaboAnalyst 2.0 (http://www.metaboanalyst.ca)[172].

Figure 3.11: Score plot of first and second principal components of raw (pre-normalised) relative intensities of lipids for each participant for each plate



This figure was produced using MetaboAnalyst 2.0 (http://www.metaboanalyst.ca)[172].

Figure 3.12: Box plots of normalised relative intensities of lipids for each participant across each plate



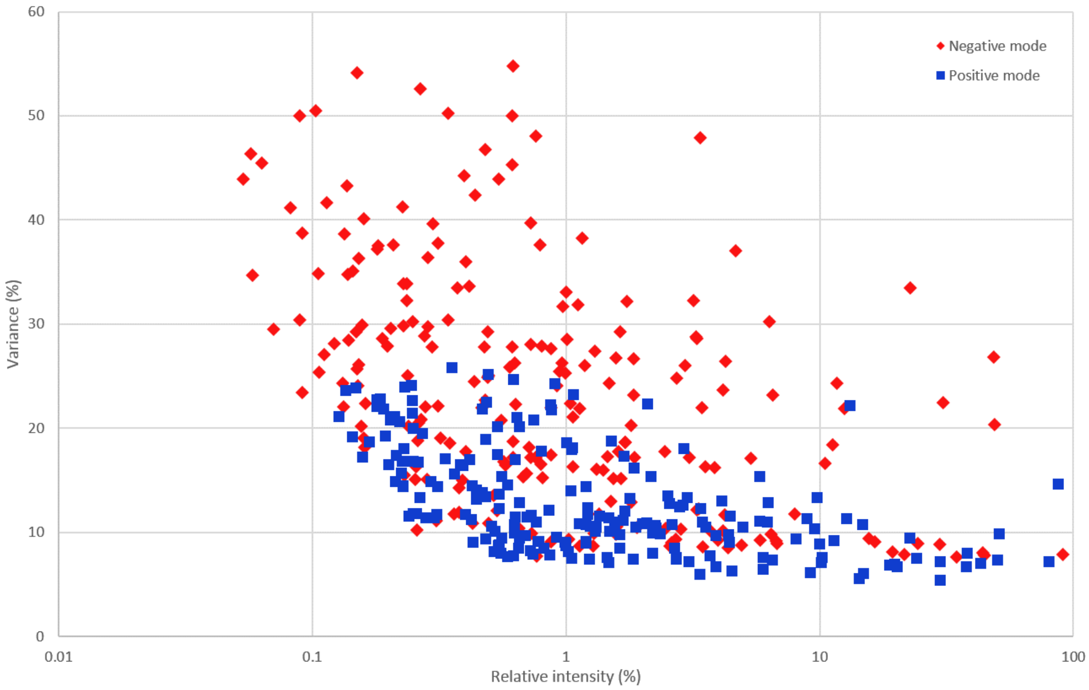
negative mode (average CV 23.67 %, median CV 22.34 %). However, the CVs demonstrate that the normalisation gave reproducible data on par with other high-throughput metabolic profiling methods[57,73].

3.4 Discussion

DIHRMS was used to record signals across a broad spectrum of mass-to-charge ratios in 5662 individuals. Raw signal data were converted and signals corresponding to known lipids were extracted using a specialised peak-picking algorithm. The lipid signals were normalised to reduce the batch effect and facilitate meaningful comparisons of relative intensities of lipids across different plates. After applying the normalisation and additional QC steps, the resulting cleaned signal data were available for 444 lipids.

The newly developed method described in this chapter used DIHRMS to yield the intensities of several thousand features. Although the platform itself is non-targeted and measures all signals within a specified m/z window, a customised peak-picking approach was used, which is scalable to studies with large-scale epidemiological data, to obtain signals for a targeted set of lipids that can reasonably be expected from human serum.

Figure 3.13: Coefficient of variation for each lipid in positive and negative ionisation mode



This newly developed open-profiling method is highly suitable to provide detailed lipid profiles in large-scale epidemiological studies, with a wider coverage of lipids than most other high-throughput profiling methods.

One of the apparent limitations of the method described in this chapter is that a large number of measured lipids were excluded from the analysis. The lipidomics platform was untargeted in the sense that it measured all signals within a mass-charge ratio window from 185 to 1000, but it would be impractical to include all signals within that range since the size of the dataset would essentially be infinite. The peak-picking algorithm was therefore implemented to extract all signals corresponding to known lipids that were theoretically detectable in human serum (1305 lipids in positive ionisation mode and 3772 lipids in negative ionisation mode). However, in practice, not all of these lipids were detectable in this specific population, and some of the detected lipids were excluded if they failed QC, so this resulted in a final dataset of 207 lipids in positive ionisation mode and 237 lipids in negative ionisation mode. Although this can be perceived as a major reduction in the overall number of lipids compared to the total number of signals that were extracted, in reality, the fact that 444 lipids were detectable and passed QC with low CVs is reasonable and to be expected for this lipidomics platform.

Another potential limitation is the method that was used to address the batch effect, which was readily apparent from the raw signal data. The strategy that was selected for batch correction was chosen because it was straightforward to implement and reasonably effective at reducing the batch effect across the 72 plates. The approach involved expressing the signal for each lipid as a proportion of the total signal (the sum of the signals for all the lipids within that ionisation mode). MetaboAnalyst 2.0 software[172] was used to evaluate the extent to which the batch effect was a concern and how well it was able to be addressed. An alternative strategy for batch correction that could have been used is Quality Control-Robust Spline Correction (QC-RSC), which is based on an adaptive cubic smoothing spline algorithm[149], but since the approach that was used sufficiently addressed the batch effect it was not necessary to implement a more complex approach. The genetic analyses also adjusted for plate to further account for any remaining batch effect.

Signals were annotated to a molecular formula on the basis of the accurate mass. The resolution of 65 000 at 400 *m/z*, as used in this study, allowed for the baseline separation of, for instance, molecular formulae $C_{41}H_{78}NPO_8{+}H^+$ (*m/z* 744.554) and $C_{42}H_{82}NPO_7{+}H^+$ (*m/z* 744.590), but was unable to determine if the former was PC(33:2) or PE(36:2) as the species are isobaric. The average mass accuracy error in the measurement of the *m/z* was 0.85 ppm across all intact lipids, with the highest difference for detected lipids of 2.69 ppm for PC ae (37:4). However, this did not mean that only one lipid species contributed to a specific *m/z*. For example, the ion which was identified as TG(52:2) with *m/z* 876.802 could be many different triglyceride lipids [e.g. TG(16:0/18:2/18:0), TG(14:0/16:0/22:2) or TG(16:0/18:1/18:1)], which all have the same molecular weight. Interpretation of species at the level of chemical formulas allows one to model changes within lipid pools according to biological processes such as chain elongation and desaturation. One can assume that a given signal peak was likely to be a combination of several lipid species. The annotation was further based on fragmentation data for the most predominant ion through additional LC-MS/MS analyses.

Because the open-profiling approach did not predetermine which lipid species would be detected, as the name suggests, it provided data on all ionisable molecules and was therefore very sensitive to contamination, especially of compounds with high ionization efficiencies. In all analyses, there was evidence of adipates (*m/z* 371.316) and organophosphates, such as Tris(ditert-butylphenyl)phosphite (*m/z* 647.459) and its oxidation products (*m/z* 663.454), which leached from plastics into the organic solvents. However, using glass-coated well-plates minimised the contact time of the samples with the plastics, and by using blanks and QCs at three different concentrations, the contaminating ions (approximately half of all the signals) were excluded from the final data set. The use of glass-coated well plates was essential to obtain both precise and reliable data. Furthermore, as the method relied on nanoflow, contaminants had minimal impact on the ionisation efficiency.

The developed peak-picking algorithm enabled the processing of almost seven thousand data files in about two minutes per sample. The analysis time was greatly sped up by processing each file independently and performing analyses in parallel, since unlike other metabolomics processing techniques, there was no requirement to load all files jointly into memory to perform the alignment. This approach is only suitable to compare similar samples where the same lipids are expected, as it requires known lipids with their target *m/z*. The results required manual curation as in certain cases the target *m/z* was too close to an isotope or adduct of another lipid. Therefore, the identity of all the ions that passed the QC filter was confirmed, and selected samples were analysed by high-resolution LC-MS/MS to confirm lipid annotations.

The approach that was used has three practical advantages over most published approaches: (1) The method is extremely fast, so that with an analysis time of just over two minutes per sample it is possible to run several hundred samples per day; (2) as a consequence of the high throughput, the cost per sample is greatly reduced; and (3) by virtue of the simplicity of the method it is also robust, with low CVs achievable for many of the lipid species detected. This opens up the possibility of applying the method to much larger studies or exploring possible applications in routine assessment of patients and health screenings.

Open-profiling lipidomics using DIHRMS, combined with a novel peak-picking algorithm, proved highly effective for obtaining concentrations of 444 lipid metabolites in 5662 individuals from PROMIS. The following chapters describe the analyses of the lipidomics signal data.

CHAPTER 4

Descriptive analyses of lipidomics data

Chapter summary

The analyses in this chapter focus on characterising the lipid metabolites and their association with CHD risk factors such as smoking, physical activity, levels of major circulating lipids, diabetes, hypertension, and obesity. While dimension reduction techniques were employed to facilitate analysis of the overall patterns in the signal data across lipids, the lipids were also analysed on an individual basis to examine their characteristics and associations more comprehensively. However, rather than displaying results for the association of hundreds of individual lipids with dozens of different traits, which would produce an overload of information that is not particularly informative, this chapter presents a selected set of descriptive results that best illustrates the diversity of the lipids and the utility of this platform to facilitate a deeper understanding of lipid metabolism.

Descriptive analyses were conducted on the cleaned lipidomics dataset, which consisted of signal data for 444 lipid metabolites in 5662 individuals, to examine cross-correlations of each pairwise combination of lipids. The lipids were analysed for their association with lifestyle and environmental factors such as smoking and physical activity, and their association with major lipids such as HDL-C, LDL-C, and triglycerides. Principal component analysis was performed as a dimension reduction technique to reduce the complexity of the data and identify patterns, and then principal components of the lipids were analysed for their association with CHD risk factors such as obesity, hypertension, and diabetes. Partial least squares discriminant analysis was implemented to determine whether levels of lipid metabolites could be used to classify individuals according to their body mass index and Framingham 10-year risk of cardiovascular disease. Gaussian graphical modelling (GGM) was also employed to obtain partial correlations of the lipid metabolites. A heat map and network map were produced of the observed GGM edges compared with the number of GGM edges that would be expected due to chance. Specific results are also shown for a subset of the triglyceride metabolites and for lipids associated with a variant in the *APOA5* locus that is known to be associated with major lipids and CAD. Overall, these analyses showed that the lipid metabolites are highly correlated with each other and display associations with major lipids and several CHD risk factors.

4.1 Introduction

Lipidomics can provide a wealth of useful information on lipid profiles in the body and the metabolic pathways that are involved in lipid metabolism, which are closely linked to the onset of many chronic diseases as described in Section 1.3. Analyses of the epidemiological associations between lipids and CHD risk factors can yield valuable insights into lipid metabolism and the onset of disease.

Lipidomics and other high-throughput analytical platforms lead to the production of extensive amounts of data[149]. In addition, metabolic pathways are enormously complex, so understanding and interpreting the relative composition and abundance of metabolites across a set of samples can be challenging, especially when aiming to translate lipid metabolites into meaningful biomarkers with useful clinical applications[173,174]. A number of statistical methods, such as data reduction techniques to reduce the complexity and dimensionality of the data, are needed to analyse these types of datasets in a thorough and consistent manner.

One of the most commonly used methods to analyse biological high-throughput data is principal component analysis (PCA)[58,150,175]. PCA is an approach to capture the structure of multidimensional data by reducing it into a series of artificial variables (i.e. "principal components") that explain a moderate proportion of the variation in the data, which may then be used as predictors or covariates in subsequent analyses[176]. PCA is performed on a matrix of correlation coefficients and requires that the variables are normally distributed and that pairs of variables display a bivariate normal distribution[176]. In order to meet these assumptions, mass spectral data need to undergo log transformation, data centring, and statistical normalisation techniques such as Pareto Scaling, in which each variable is divided by the square root of its standard deviation[150,177]. Depending on the type of centring, scaling, and transformation techniques used, researchers can obtain vastly different results that highlight completely different findings[177]; thus, these data preparation steps mask the essential structure of the data and the associated quality control information that is critical for high-throughput analyses. Additionally, the projection of the original spectral data onto a multivariate space of lower dimensionality, albeit allowing for spectral clustering, produces a loading matrix that displays correlations but whose values have no real-world interpretation[150,176]. Therefore, while PCA is useful for verifying the accuracy of quality control samples, detecting outliers, and initially examining clusters and patterns in the data, this analytical technique is otherwise limited in its applicability[178]. For this reason, PCA can be helpfully employed to initially identify patterns in the lipidomics data, and the associations of these principal components with CHD risk factors can also be assessed, but further analyses are also needed that delve deeper into individual lipids with interesting associations, which can provide detailed information that is more straightforward to interpret[175].

Another approach, Gaussian graphical modelling (GGM), is an established technique that can lead to insights into the dependency structure between lipids, such as (1) which lipids might influence each other, and (2) which lipids might be under shared regulatory mechanisms[132,179,180]. GGMs are also known as partial correlation networks, for which the partial correlations are obtained by determining the correlation between two metabolites while holding all other metabolites constant, and repeating this for each pair of metabolites[179]. The partial correlation between metabolites A and B while holding C constant is determined by performing normal linear least-squares regression with A as the target and C as the predictor, then performing a normal linear least-squares regression with B as the target and C as the predictor, then calculating the residuals for each regression and the correlation coefficient between the two sets of residuals[179].

PCA and GGM are "unsupervised" statistical techniques for pattern recognition because they are not based on any prior knowledge about the samples, such as phenotype or outcome information, so the results depend solely on the structures inherent in the underlying data[58]. In contrast, when using "supervised" methods, additional external details are included that help arrange or group the data according to the provided characteristics[58]. A variety of supervised methods are useful for analysis of high-throughput metabolomics data. One such technique is partial least squares discriminant analysis (PLS-DA), which is a regression extension of PCA. PLS-DA involves selecting a set of coefficients that define the single linear combination of factors (e.g. phenotype data and lipid levels) that best differentiates (i.e. discriminates) between groups based on a chosen outcome of interest[181]. PLS-DA has been widely used in metabolomics studies to sharpen the distinction between groups according to shared characteristics[175]. For example, PLS-DA has been applied in order to identify differences in levels of metabolites according to chronic kidney disease and hyperlipidaemia status in rats[182,183], to distinguish individuals with dietary patterns at increased risk of chronic diseases based on levels of metabolites[56], and to characterise differences in levels of metabolites according to the flesh colour of a potato[184]. The use of supervised techniques to identify a specific set of lipid metabolites that distinguishes between individuals with various CHD risk factors (e.g. obese individuals compared to those with normal adiposity) has important clinical applications for personalised medicine and prevention of CHD and could be used in risk prediction models.

4.2 Methods

4.2.1 Descriptive analyses and data reduction

Histograms and quantile-quantile plots of lipid metabolites

Histograms and quantile-quantile (Q-Q) plots were constructed to examine the distributions of each lipid metabolite. The lipid signals on the raw scale were compared with the normalised, log-transformed scale in order to confirm approximate log-normality.

Principal component analysis

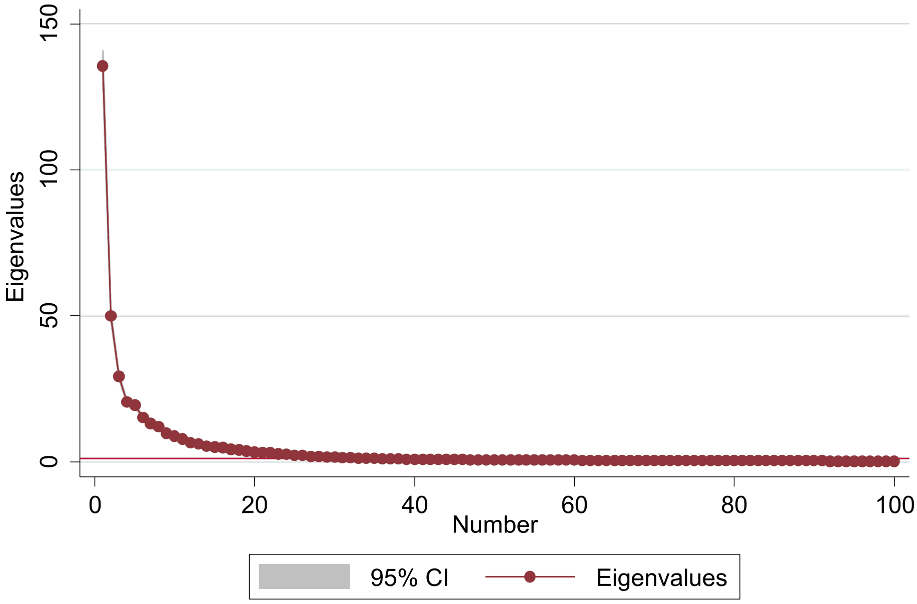
PCA[176] was conducted on the normalised relative intensities of the lipids. The matrix loadings were orthogonally rotated and the first four principal components were retained based on examination of the scree plot of the eigenvalues (Figure 4.1). These four principal components explained 55.1 % of the variance in the relative intensities of the lipids, while altogether, 56 principal components explained 95 % of the variance. Since PCA is more effective when the input data matrix has complete information on all variables included in the calculation of principal components, the 17 lipids (3.8 %) that had more than 10 % missing data (as described in Chapter 3) were excluded from the analysis. Scatter plots were produced comparing the first versus second principal components and third versus fourth principal components.

Partial least squares discriminant analysis

PLS-DA was also conducted on the normalised relative intensities of the lipids in order to separate individuals according to CHD risk factors using a score comprised of all lipids. The PLS-DA analysis was performed in R using the “mixOmics” package[185,186]. Individuals were categorised according to whether their body mass index (BMI) was underweight ($< 18.5\,\mathrm{kg/m^2}$), normal ($18.5\,\mathrm{kg/m^2}$ to $24.9\,\mathrm{kg/m^2}$), overweight ($25\,\mathrm{kg/m^2}$ to $29.9\,\mathrm{kg/m^2}$), or obese ($\geq 30\,\mathrm{kg/m^2}$). PLS-DA was then applied to separate individuals based on their lipid levels into these BMI categories.

PLS-DA was also applied to separate individuals according to their Framingham 10-year relative risk for cardiovascular disease (CVD)[187]. Using the standard Framingham CVD algorithm[187], each individual was assigned points based on their age, sex, high-density

Figure 4.1: Scree plot of eigenvalues from PCA



lipoprotein cholesterol (HDL-C) levels, total cholesterol levels, systolic blood pressure (SBP), smoking status, and diabetes status, which were then converted to a 10-year CVD risk score and classified as low ($< 10\,\%$), moderate ($10\,\%$ to $20\,\%$), or high ($\geq 30\,\%$) risk.

4.2.2 Assessment of correlations between metabolites

Cross-correlations of lipid metabolites

Initially, Pearson correlation coefficients were calculated to determine cross-correlations of each of the lipid metabolites within each ionisation mode. Ward's minimum variance agglomerative method[188], which minimises the sum of squared distance of pairs of objects, was applied iteratively to determine the hierarchy of the entire set of lipids. A heat map was produced to display the correlation coefficient for each pairwise combination of lipids. However, Pearson correlation coefficients have limitations, as they are generally unable to distinguish between direct and indirect metabolic interactions[179], so further techniques were also employed.

Gaussian graphical modelling

In order to better resolve lipid cross-correlations, a GGM was estimated on the normalized relative intensities of the lipids. This was performed between each pair of lipid metabolites while holding all other metabolites constant. The GGM resulted in a set of edges in which each edge connected two detected lipids if their cross-correlation conditioned on all other lipids was significantly different from zero. Participants with more than 10 % missing lipids, as well as lipids with more than 20 % missing participants, were removed from the analysis. The "genenet" R package was used to infer the GGM[189]. A similar approach for metabolomics data has been suggested previously[179]. To focus on strong effects, edges in the model were only retained if they met a false discovery rate (FDR) cutoff of 0.05 and had a partial correlation coefficient greater than 0.2. This resulted in a network of the partial correlations between all metabolites. The results were also summarised and combined within each lipid class to produce a heat map of the partial correlations between each of the lipid classes.

The heat map was constructed by first calculating the partial correlation between each pairwise combination of lipids, then counting the number of GGM edges between lipids that belong to the same lipid subclass. Next, the connections between lipids and subclasses were randomly shuffled and the number of GGM edges was then counted; this process was repeated for 1000 permutations so that an average number of expected GGM edges was obtained. The actual (observed) number of GGM edges was then compared to the expected number of GGM edges if this assignment was random. These were then summarised within lipid subclasses. Each cell of the heat map showed the ratio of the observed number of GGM edges for all lipids within that subclass to the total number of connections between lipids within that subclass. The cells were coloured red or blue according to whether the observed number of GGM edges was more or less than expected due to chance alone.

Fatty acid chain enrichment analysis

The detected lipids were also manually annotated with respect to their constituent fatty acid chains. To test whether edges from the GGM were enriched for any combination of fatty acid chains, the annotations were permuted 1000 times, keeping the number of annotations per lipid and fatty acid chain constant using the R package "BiRewire"[190]. For each combination of fatty acid chains, the number of GGM edges connecting lipids with that specific combination were counted using the true annotation as well as the permuted versions. These counts were then used to directly estimate P-values of enrichment and depletion.

4.2.3 Association of lipid metabolites with CHD risk factors

There are a range of established CHD risk factors, which include hypertension, smoking, diabetes, LDL-C, HDL-C, physical activity, diet, and obesity. As set out in Figure 1.15, one of the goals of this dissertation was to determine whether these factors influence levels of lipid metabolites, or whether levels of lipid metabolites can result in increased risk of CHD due to undesirable modifications in these risk factors. The association of lipid metabolites with these CHD risk factors was analysed in three groups: (1) Lifestyle and behavioural factors such as diet, physical activity, and smoking; (2) Circulating biomarkers such as LDL-C and HDL-C; and (3) Intermediate outcomes such as hypertension, diabetes, and obesity.

It is often challenging to portray the results from the association of 444 lipid metabolites with a range of different risk factors as there are simply too many data points to process and present in an understandable and interpretable format. Examination of a few traits across all of the lipids can easily produce thousands of association statistics. Therefore, a number of the following analyses in this and subsequent chapters have focused on a subset of lipid metabolites from a particular subclass, such as triglycerides, or on a subset of lipids that are associated with a particular genetic region of interest that is known to be associated with risk of CHD. As will be described in Chapter 5 and Chapter 6, the GWAS analyses resulted in the discovery of a number of lipid metabolites that were significantly associated with known genetic regions. Some of the analyses have focused on the lipid metabolites that are associated with rs662799, a common polymorphism in the *APOA5* region that is known to be associated with major lipids and CAD (Table 1.2). While the actual genetic analyses will be described in the following chapters, this chapter describes the epidemiological (non-genetic) associations of the lipids that are significantly associated with this variant.

Association with lifestyle factors

It is already well established that lifestyle factors such as diet, smoking, and physical activity can influence levels of major circulating lipids. Therefore, this analysis aimed to examine the association of these behavioural factors on concentrations of lipid metabolites, for which less is known. Logistic regression models adjusted for age and sex were constructed examining the association of smoking and physical activity with a wide range of lipid metabolites. The variables for smoking status and physical activity were reclassified as binary variables (i.e. smoker versus non-smoker and physically active at any level versus not physically active).

Although a detailed food frequency questionnaire was obtained from PROMIS participants as described in Chapter 2, the information was self-reported and subject to recall bias. Furthermore, although the survey was comprehensive and asked about consumption of approximately 160 different foods, the nutritional content and make-up of each type of food on the questionnaire has not been determined, so it is difficult to identify overall dietary patterns that can be linked to lipid levels without the use of PCA, which is again subject to limitations. For these reasons, dietary patterns were not examined and the analysis of lifestyle factors focused on smoking and physical activity.

Association with circulating biomarkers

PROMIS is a rather unique bioresource in that not only does it measure standard major lipids such as total cholesterol, HDL-C, LDL-C, and triglycerides, and it has detailed information on hundreds of lipids using a novel lipidomics platform, additionally, it has measurements on hundreds of clinical chemistry biomarkers, as described in Chapter 2. These include glucose, creatinine, C-reactive protein, $\mathrm{HbA_{1c}}$, C-peptide, ferritin, apolipoprotein B, interleukin-6, alkaline phosphatase, alanine transaminase, aspartate aminotransferase, and many others.

The lipid metabolites that were most strongly associated with *APOA5* within each lipid category were examined to determine their cross-correlations with a range of circulating biomarkers. Partial correlation coefficients were calculated with adjustment for age and sex.

Association with intermediate outcomes

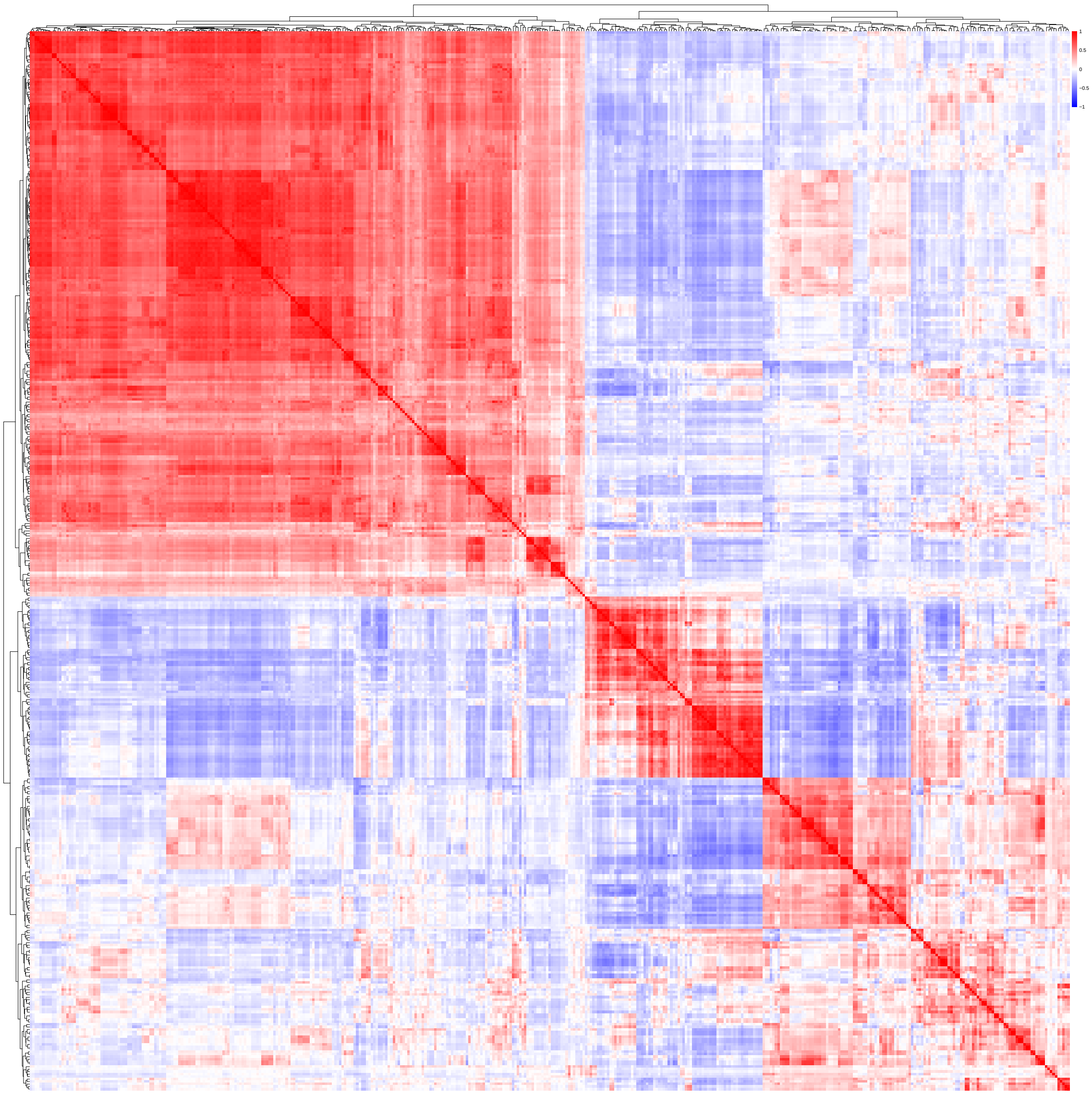
The association of lipid metabolites with several intermediate outcomes, which are soft clinical endpoints that can eventually lead to MI or CHD, were also examined. Unconditional logistic regression models adjusting for age and sex were used to assess the association of the second, third, and fourth principal components with having hypertension, being overweight or obese, and having diabetes. Hypertension was defined as SBP $\geq 140\,\mathrm{mmHg}$ or DBP $\geq 90\,\mathrm{mmHg}$; overweight and obese were defined as body mass index (BMI) $\geq 25\,\mathrm{kg/m^2}$ or BMI $\geq 30\,\mathrm{kg/m^2}$, respectively; and diabetes was defined as $\mathrm{HbA_{1c}} \geq 6.5\,\%$.

4.3 Results

4.3.1 Cross-correlations of lipid metabolites

A heat map with hierarchical clustering of Pearson correlation coefficients showed significant clustering of lipid metabolites in both positive and negative ionisation modes (Figure 4.2). Although the lipids were clustered broadly into three overall groups, they were correlated

Figure 4.2: Heat map of Pearson correlations of lipid metabolites with hierarchical clustering

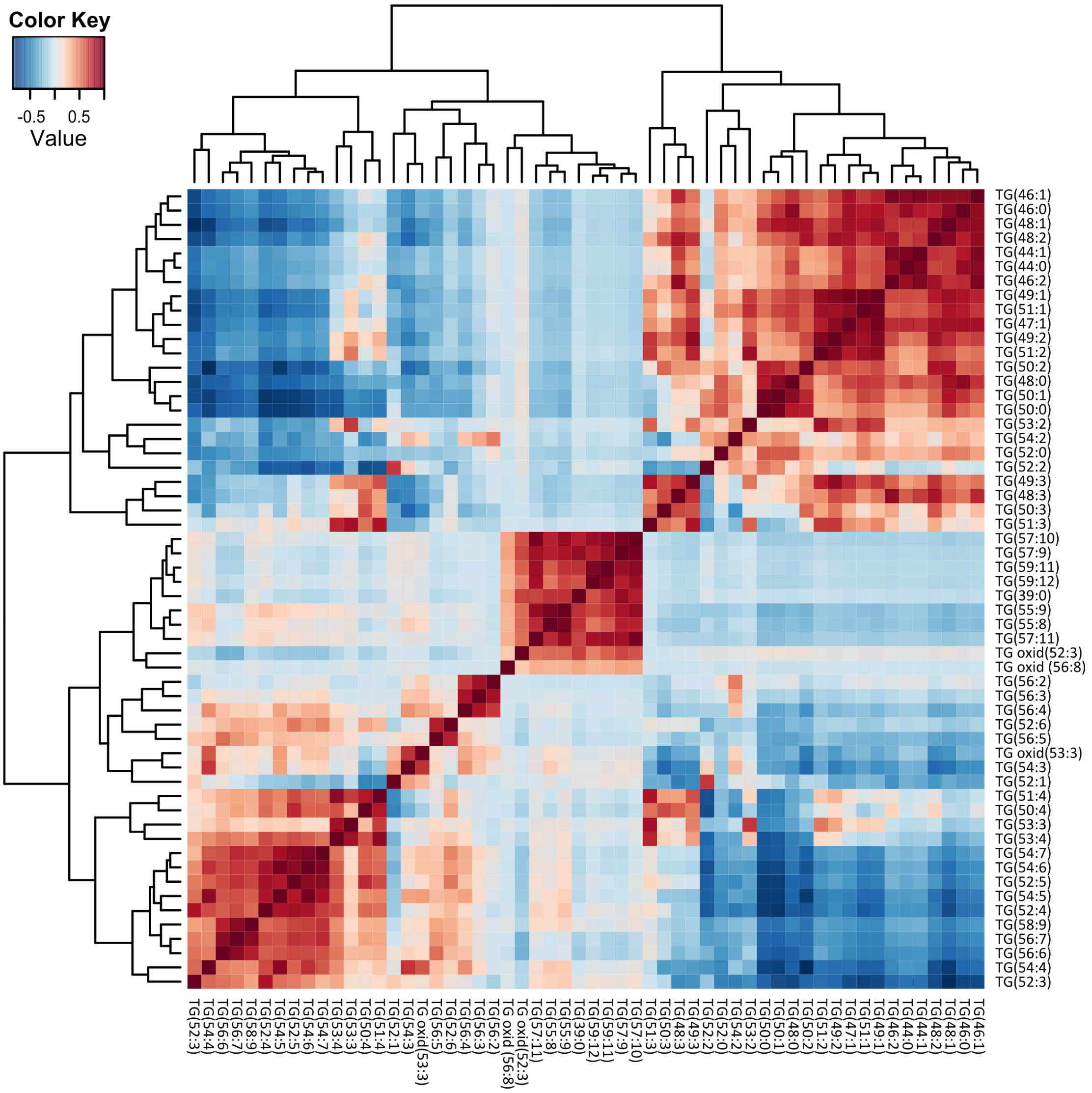


Both axes show the full list of 444 lipids, which have been ordered using hierarchical clustering.

with other lipids both within and across lipid classes and subclasses.

The correlations between lipid metabolites can be seen more clearly in a separate heat map on a subset of just the triglyceride-related metabolites (Figure 4.3). As one would expect, triglycerides with the same carbon number but differing numbers of double bonds were strongly positively correlated with each other. Interestingly, while most of the triglycerides with lower carbon numbers and fewer double bonds were strongly positively correlated with each other, these triglycerides were strongly negatively correlated with the triglycerides with higher carbon numbers and more double bonds. Additionally, peroxidised triglycerides [e.g. TG oxid(52:3) and TG oxid(53:3)] were negatively correlated with their regular triglycerides counterparts [TG(52:4) and TG(53:4), respectively].

Figure 4.3: Heat map of Pearson correlations of triglyceride-related metabolites with hierarchical clustering



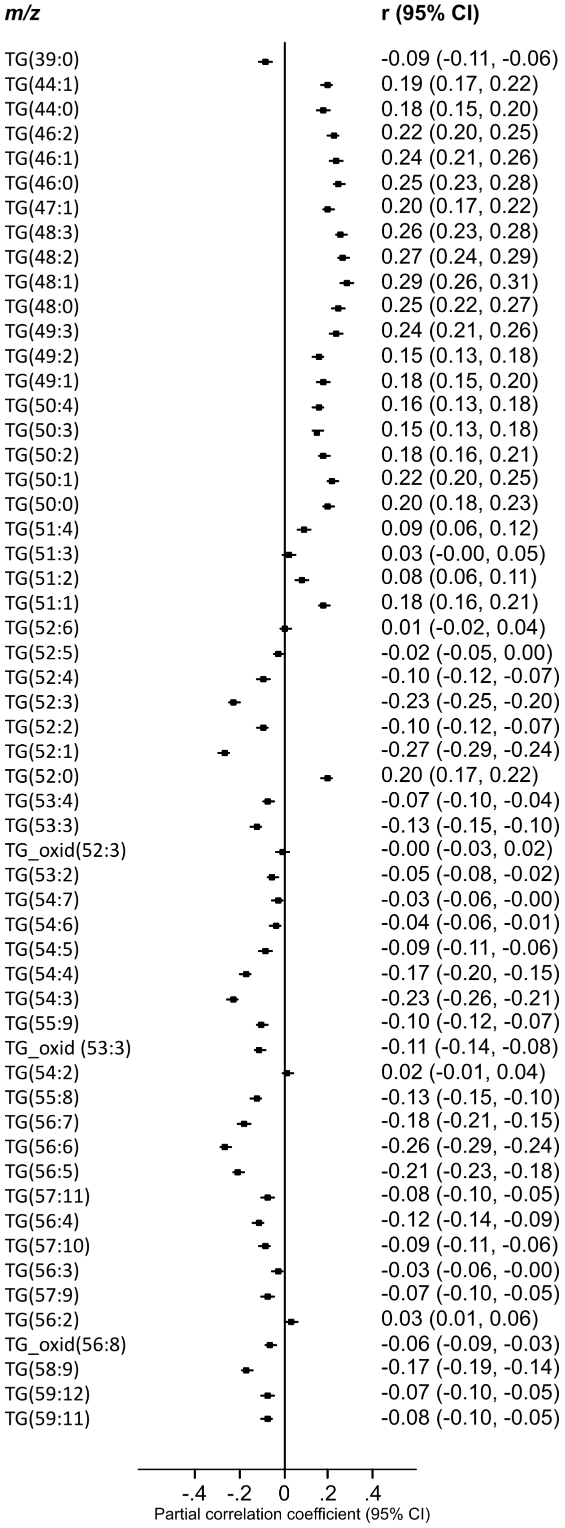
While most lipid metabolites were correlated with levels of major lipids, confirming the validity of the platform, the direction of the correlation varied according to the structure of individual metabolites. A graph of the partial correlation coefficients and 95 % CIs for each of the triglyceride-related metabolites with circulating triglyceride levels is shown in Figure 4.4. While most of the triglyceride-related metabolites with fewer carbon atoms and fewer double bonds in their fatty acid side-chains were positively correlated with circulating triglycerides, many of the triglyceride-related metabolites with comparatively more carbon atoms and double bonds in their fatty acid side-chains were inversely correlated with circulating triglycerides.

For many of the triglycerides, as the number of double bonds increased or the level of saturation decreased [e.g. from TG(48:1) to TG(48:2) to TG(48:3), or from TG(50:1) to TG(50:2) to TG(50:3)], the correlation with circulating triglycerides decreased, indicating an inverse correlation between the number of double bonds and circulating triglycerides. Likewise, for triglycerides that had an inverse correlation with circulating triglycerides, the same trend was observed: as the number of double bonds increased [e.g. from TG(54:3) to TG(54:4) to TG(54:5), or from TG(56:6) to TG(56:7) to TG(56:8)], the correlation with circulating triglycerides decreased. The explanation for this pattern is unclear, but the correlation of triglycerides with circulating triglycerides provides validation of the platform and demonstrates the diversity of lipids with varying characteristics.

4.3.2 Gaussian graphical modelling

The GGM approach employed partial correlations to determine if specific lipids were still strongly correlated after adjusting for all other lipids, and if lipid signals that were assigned to specific lipids were in fact dominated by isotopologues of other lipids or signals present as artefacts. The relationships between the lipid subclasses are shown in a heat map (Figure 4.5). From the 314 GGMs, there were ten correlations that were purely M+1 isotopes (the same lipid that contained one ^{13}C isotope) of other lipid signals and four correlations that were purely M+2 isotopes (the same lipid that contained two ^{13}C isotopes), based on very high correlations ($r > 0.997$) and correct isotope ratios. There were 26 correlations where the M+1 isotope contributed considerably to the signal and four where the M+2 isotope contributed predominantly to the signal. However, those signals also showed contributions of different lipid signals, for which the correlations were not as high ($r < 0.997$) or the isotope ratio was incorrect. There were 36 correlations where the signals came from the same lipid in both positive and negative ionization modes,

Figure 4.4: Partial correlation coefficients of triglyceride-related metabolites with circulating triglyceride levels



and two sets of lipids for which the signals overlapped and the peak-picking algorithm was unable to distinguish the signals. The remaining 222 significant correlations were not caused by any technical artefacts and were therefore most likely driven by biology, such as correlations between lipids containing the same fatty acids (Figure 4.6).

4.3.3 PCA of lipid metabolites

PCA of overall lipids

The overall differences in lipid metabolism were assessed in the cohort using PCA. Scatter plots were produced of the matrix loadings of the first versus second (Figure 4.7a) and third versus fourth (Figure 4.7b) principal components. The individual lipids are distinguished by colour on the figures according to the overall lipid category to which they belong.

The first principal component (which explained 31.8 % of the variance in the lipid levels) correlated with differences between the positive and negative ionisation modes. The dynamic range of the negative mode data was more limited than the positive mode data, and due to the lower ionisation efficiency, the data were more prone to ion suppression. These differences between the ionisation modes were amplified when the data were expressed relative to total signal intensity. The first principal component was therefore excluded from further data analysis.

The second component (which explained 11.7 % of the variance) was dominated by triglycerides containing shorter and more saturated fatty acids, which had the highest positive loadings, versus fatty acids (e.g. free linoleic acid) and cholesterol esterified with polyunsaturated fatty acids [e.g. CE(18:2)], which had the strongest negative loadings Figure 4.7a).

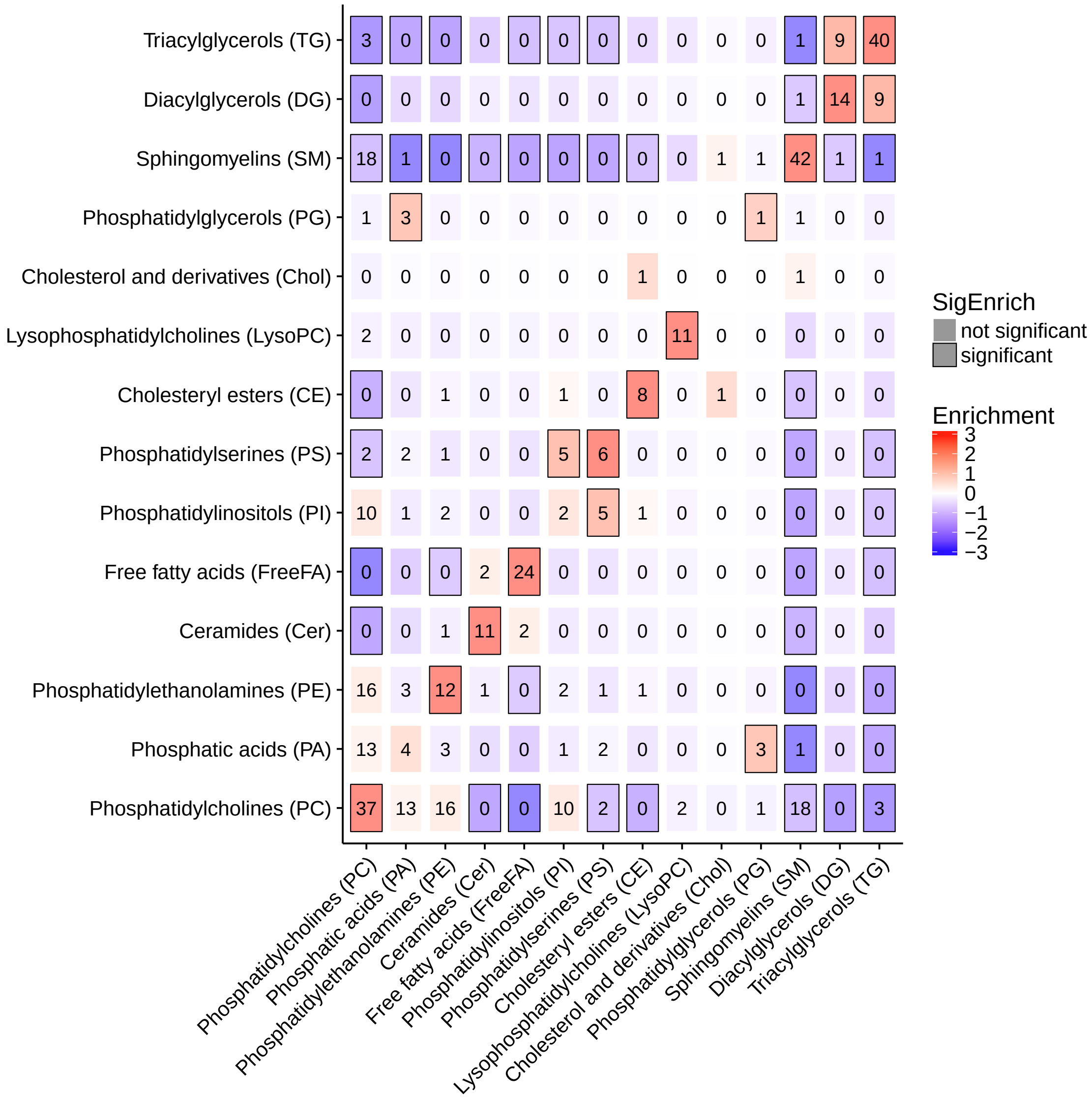
The third component (which explained 6.9 % of the variance) differentiated saturated phosphatidylcholines [e.g. PC(32:0), PC(34:0), PC(32:0)] from triglycerides containing longer, unsaturated fatty acids [e.g. TG(54:5), TG(54:7), TG(56:7)] (Figure 4.7b).

The fourth component (which explained 4.8 % of the variance) differentiated the odd-chain fatty acid–containing sphingomyelins with the highest positive loadings [e.g. SM(39:1), SM(41:1), SM(37:1)] versus saturated free fatty acids and triglycerides with the strongest negative loadings [e.g. TG(52:2), TG(54:2)] (Figure 4.7b).

PCA of triglycerides

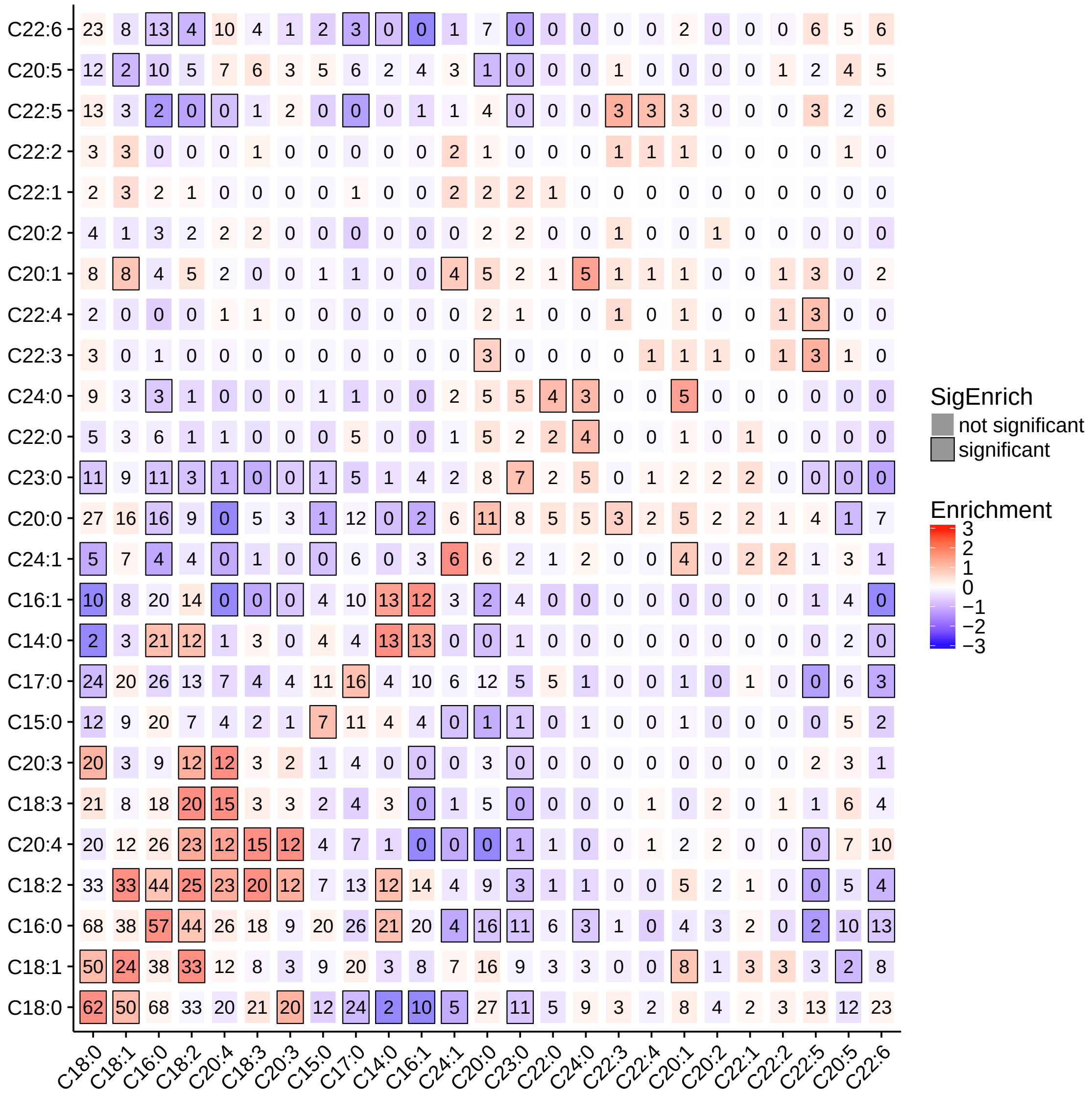
In addition to the PCA analysis on the full set of lipids, PCA was also performed on a subset of lipids consisting of only the triglycerides. The corresponding scatter plots of

Figure 4.5: Heat map showing relationships between lipid subclasses based on partial correlations derived using Gaussian Graphical Modelling



This heat map shows the relationships between lipid subclasses based on the inferred Graphical Gaussian Model (GGM). The number in each cell shows the observed number of GGM edges connecting two lipids in subclasses specified on the x- and y-axes. The cells are coloured red or blue according to whether the observed number of GGM edges is more or less than expected due to chance alone, and a box is drawn around the cell if there is a significant difference between the number of observed versus expected GGM edges.

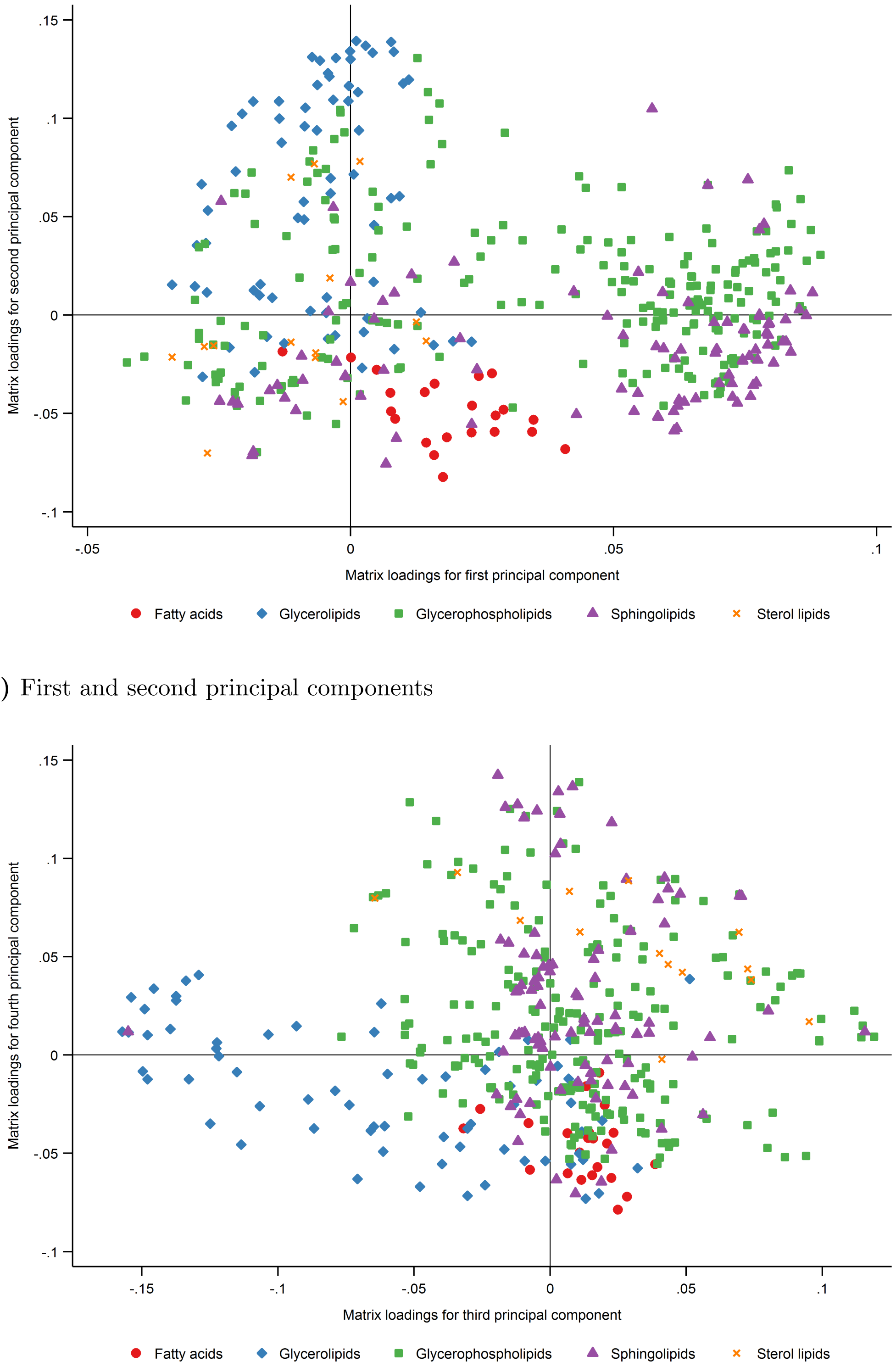
Figure 4.6: Heat map showing partial correlations between constituent fatty acid chains of lipids based on partial correlations derived using Gaussian Graphical Modelling



	C18:0	C18:1	C16:0	C18:2	C20:4	C18:3	C20:3	C15:0	C17:0	C14:0	C16:1	C24:1	C20:0	C23:0	C22:0	C24:0	C22:3	C22:4	C20:1	C20:2	C22:1	C22:2	C22:5	C20:5	C22:6
C22:6	23	8	13	4	10	4	1	2	3	0	0	1	7	0	0	0	0	0	2	0	0	0	6	5	6
C20:5	12	2	10	5	7	6	3	5	6	2	4	3	1	0	0	0	1	0	0	0	0	1	2	4	5
C22:5	13	3	2	0	0	1	2	0	0	0	1	1	4	0	0	0	3	3	3	0	0	0	3	2	6
C22:2	3	3	0	0	0	1	0	0	0	0	0	2	1	0	0	0	1	1	1	0	0	0	0	1	0
C22:1	2	3	2	1	0	0	0	0	1	0	0	2	2	2	1	0	0	0	0	0	0	0	0	0	0
C20:2	4	1	3	2	2	2	0	0	0	0	0	0	2	2	0	0	1	0	0	1	0	0	0	0	0
C20:1	8	8	4	5	2	0	0	1	1	0	0	4	5	2	1	5	1	1	1	0	0	1	3	0	2
C22:4	2	0	0	0	1	1	0	0	0	0	0	0	2	1	0	0	1	0	1	0	0	1	3	0	0
C22:3	3	0	1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	1	1	0	1	3	1	0
C24:0	9	3	3	1	0	0	0	1	1	0	0	2	5	5	4	3	0	0	5	0	0	0	0	0	0
C22:0	5	3	6	1	1	0	0	0	5	0	0	1	5	2	2	4	0	0	1	0	1	0	0	0	0
C23:0	11	9	11	3	1	0	0	1	5	1	4	2	8	7	2	5	0	1	2	2	2	0	0	0	0
C20:0	27	16	16	9	0	5	3	1	12	0	2	6	11	8	5	5	3	2	5	2	2	1	4	1	7
C24:1	5	7	4	4	0	1	0	0	6	0	3	6	6	2	1	2	0	0	4	0	2	2	1	3	1
C16:1	10	8	20	14	0	0	0	4	10	13	12	3	2	4	0	0	0	0	0	0	0	0	1	4	0
C14:0	2	3	21	12	1	3	0	4	4	13	13	0	0	1	0	0	0	0	0	0	0	0	0	2	0
C17:0	24	20	26	13	7	4	4	11	16	4	10	6	12	5	5	1	0	0	1	0	1	0	0	6	3
C15:0	12	9	20	7	4	2	1	7	11	4	4	0	1	1	0	1	0	0	1	0	0	0	0	5	2
C20:3	20	3	9	12	12	3	2	1	4	0	0	0	3	0	0	0	0	0	0	0	0	0	2	3	1
C18:3	21	8	18	20	15	3	3	2	4	3	0	1	5	0	0	0	0	1	0	2	0	1	1	6	4
C20:4	20	12	26	23	12	15	12	4	7	1	0	0	0	1	1	0	0	1	2	2	0	0	0	7	10
C18:2	33	33	44	25	23	20	12	7	13	12	14	4	9	3	1	1	0	0	5	2	1	0	0	5	4
C16:0	68	38	57	44	26	18	9	20	26	21	20	4	16	11	6	3	1	0	4	3	2	0	2	10	13
C18:1	50	24	38	33	12	8	3	9	20	3	8	7	16	9	3	3	0	0	8	1	3	3	3	2	8
C18:0	62	50	68	33	20	21	20	12	24	2	10	5	27	11	5	9	3	2	8	4	2	3	13	12	23

This heat map shows the relationships between the constituent fatty acid chains based on the inferred Graphical Gaussian Model (GGM). The number in each cell shows the observed number of GGM edges connecting two lipids with the constituent fatty acid chains specified on the x- and y-axes. The cells are coloured red or blue according to whether the observed number of GGM edges is more or less than expected due to chance alone, and a box is drawn around the cell if there is a significant difference between the number of observed versus expected GGM edges.

Figure 4.7: Scatter plot showing matrix loadings of normalised relative intensities of lipids from PCA, coloured by overall lipid category



(a) First and second principal components

(b) Third and fourth principal components these principal components are shown in Figure 4.8. There is much that can be learned by examining individual lipid subclasses. The first principal component of the triglycerides appeared to distinguish triglycerides with an odd number of carbon atoms [e.g. TG(49:3), TG(51:4), TG(55:9)], shown in the blue oval on the right side of Figure 4.8a, from those with an even number of carbon atoms [e.g. TG(50:0), TG(52:1), TG(54:3)], shown in the green oval on the left side. Odd-chain fatty acids primarily derive from dairy consumption, while even-chain fatty acids are predominantly synthesised in the liver through *de novo* lipogenesis.

The second principal component appeared to differentiate triglycerides with saturated and monounsaturated fatty acid chains [e.g. TG(44:1), TG(47:1), TG(49:2)], shown in the green and blue ovals at the top of the figure, from triglycerides containing ω-3 and ω-6 polyunsaturated fatty acids [e.g. TG(54:4), TG(58:9), TG(57:10)], shown in the orange oval at the bottom of the figure. These polyunsaturated fatty acids come primarily from fish consumption. There are also two pink ovals in Figure 4.8a, which represent a distinct category (or possibly multiple categories) from dairy consumption, fish consumption, and hepatic synthesis. However, further research is needed to determine the biological significance of the triglycerides in this grouping.

The third principal component appeared primarily to distinguish triglycerides containing polyunsaturated fatty acids from triglycerides containing even-chain fatty acids (Figure 4.8b), while the interpretation of the fourth principal component of the triglycerides was not readily apparent.

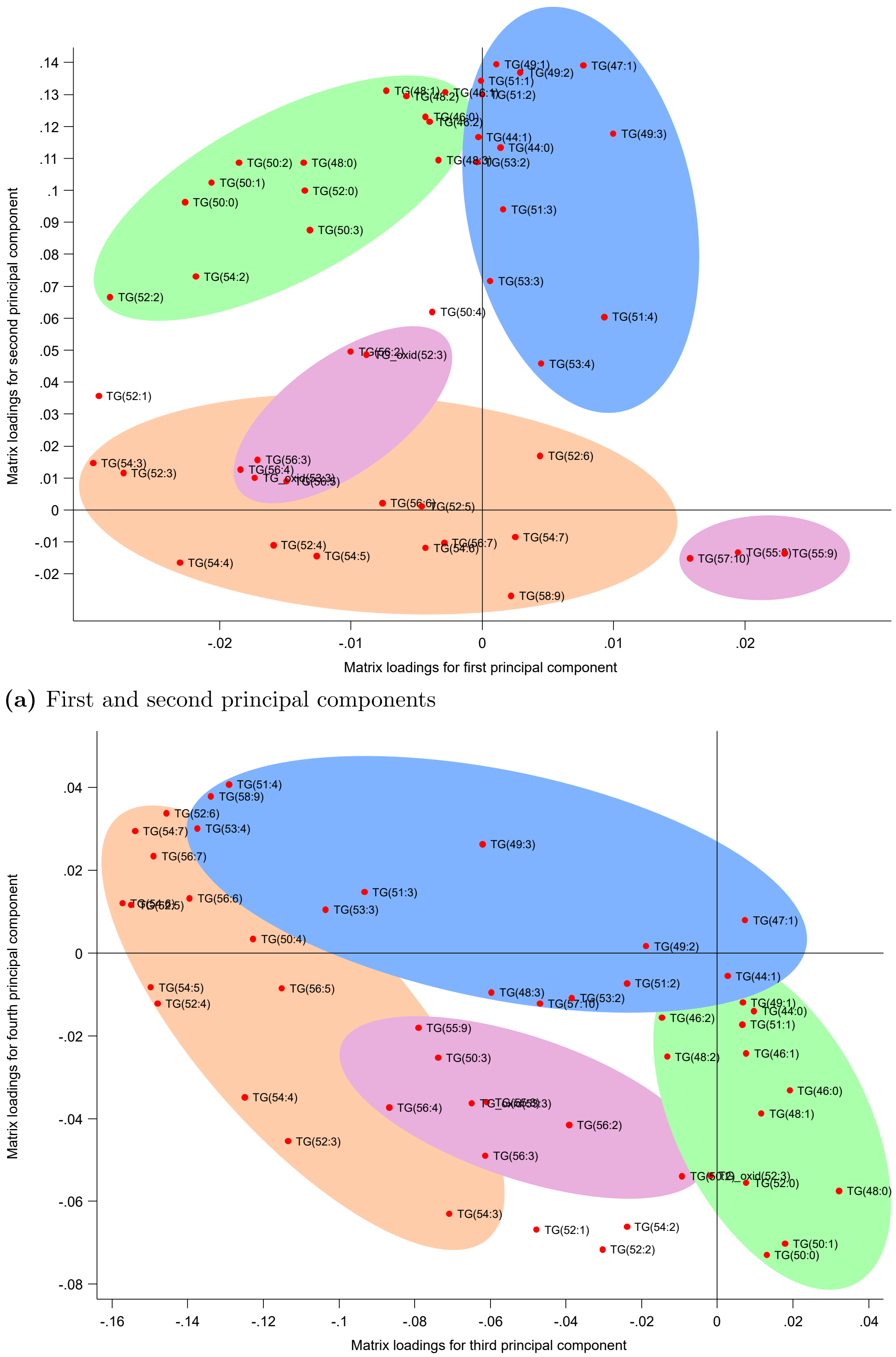
PCA of lipids associated with *APOA5*

PCA was performed on a subset of the lipids that were significantly associated with the rs662799 (chr11:116663707) variant in the *APOA5* gene at $P < 8.9 \times 10^{-10}$. The scatter plots of these principal components are shown in Figure 4.9. The first principal component is difficult to interpret, but it is clear that bar a few exceptions, the second, third, and fourth principal components were very effective at differentiating glycerolipids (i.e. diglycerides and triglycerides) from the other lipid classes.

4.3.4 Association of lipids with lifestyle factors

As shown in Figure 4.10, out of the top twenty lipids that are most strongly associated with *APOA5*, some lipids had significantly positive associations with smoking and physical activity, while other lipids had significant inverse associations. For instance, smoking was

Figure 4.8: Scatter plot showing matrix loadings of normalised relative intensities of triglycerides from PCA

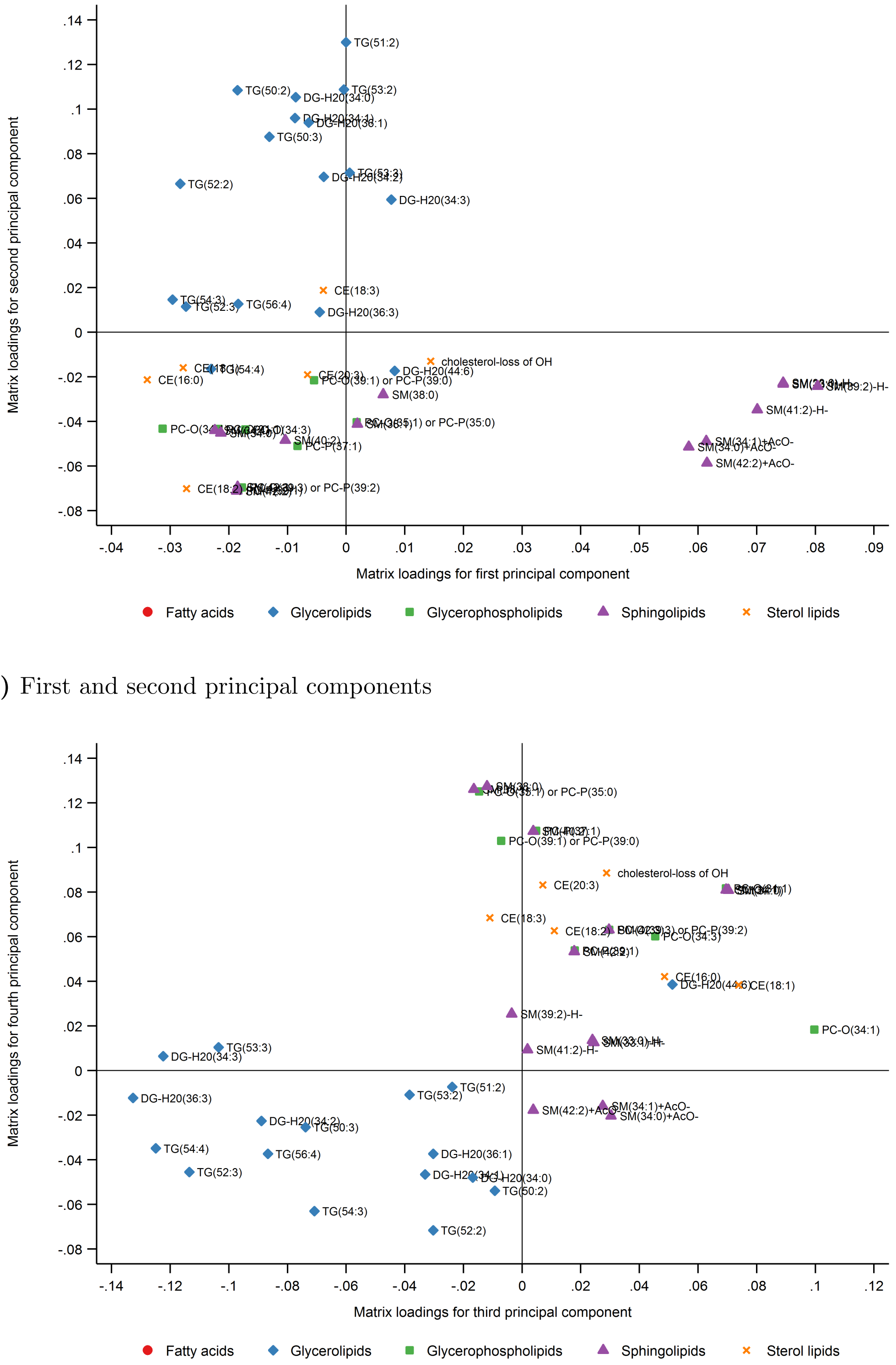


(a) First and second principal components

(b) Third and fourth principal components

Blue oval indicates triglycerides containing odd-chain fatty acids from dairy consumption; orange oval indicates triglycerides containing ω-3 and ω-6 polyunsaturated fatty acids from fish consumption; green oval indicates triglycerides containing even-chain fatty acids formed through *de novo* lipogenesis.

Figure 4.9: Scatter plots showing matrix loadings of normalised relative intensities of lipids from PCA that are significantly associated with rs662799 in the *APOA5* locus



(a) First and second principal components

(b) Third and fourth principal components associated with decreased levels of TG(52:4) but increased levels of TG(52:2) (Figure 4.10a). Since both of these triglycerides have the same number of carbon atoms on their fatty acid chains, the only difference is the level of saturation, since TG(52:4) has two additional double bonds on its three fatty acid chains. Thus, TG(52:2) has either two saturated fatty acids (zero double bonds) and a polyunsaturated fatty acid (two double bonds), or two monounsaturated fatty acids (one double bond each) and a saturated fatty acid (zero double bonds). In contrast, TG(52:4) most likely has a mixture of monounsaturated and polyunsaturated fatty acids but probably does not have any saturated fatty acids, which could explain why it was more likely to be present in the blood samples of non-smokers. Although this does not imply that smoking causes changes in lipid profiles, the adverse behaviour of smoking quite likely has a mediating effect.

Similar patterns in lipids with differing directions of effect can be observed for the association with physical activity, although the majority of the associations were not statistically significant (Figure 4.10b). However, one aspect to keep in mind is that although all of these lipids are significantly associated with *APOA5*, not all of them are influenced by lifestyle choices such as smoking or being physically active. Other factors are likely to play an important role as well and are more likely to be the main causal drivers.

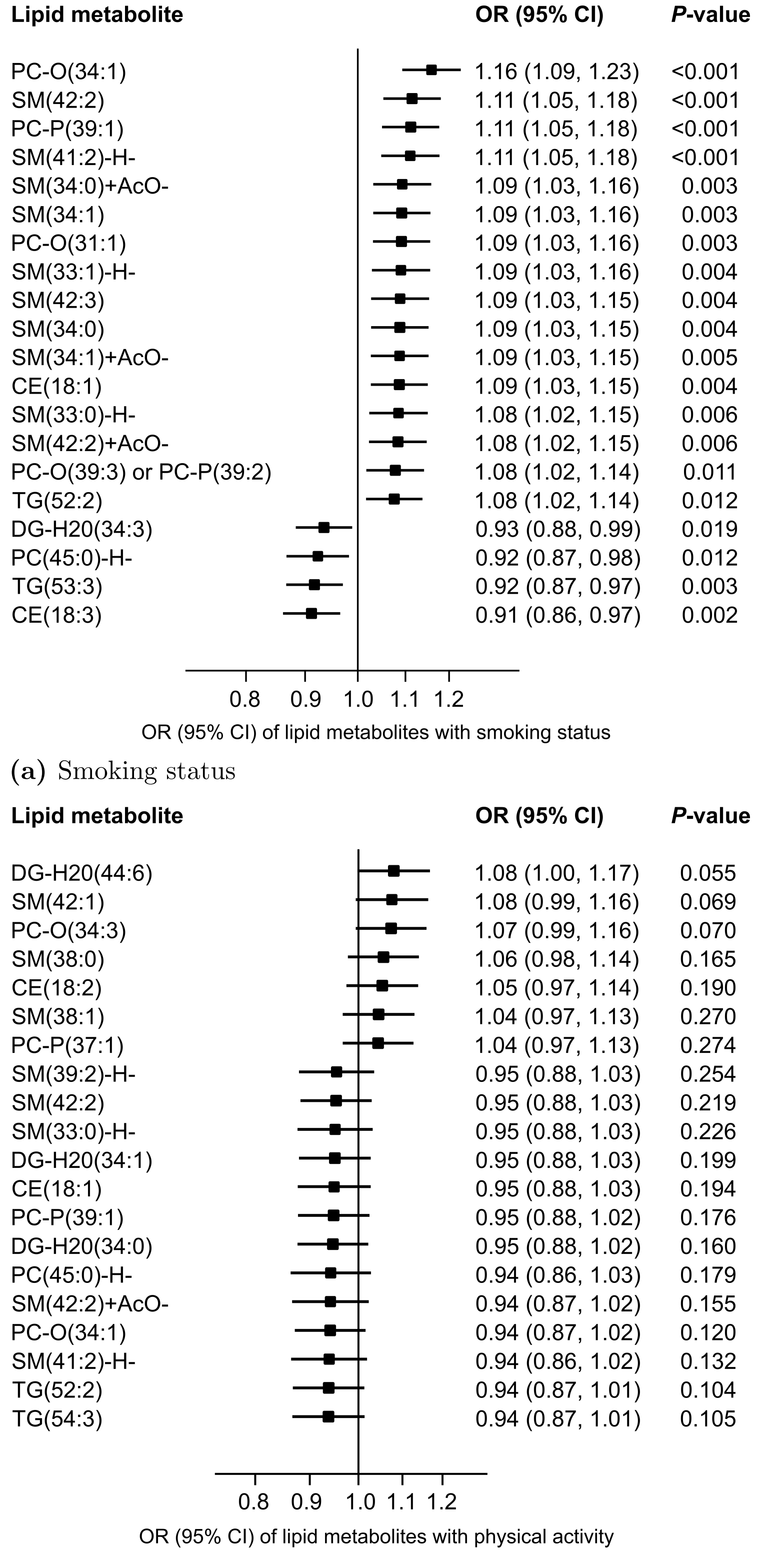
4.3.5 PLS-DA of lipid metabolites

A PLS-DA model was used to discriminate individuals according to their BMI category. However, lipid metabolites were unable to distinguish individuals who were underweight, normal weight, overweight, or obese due to the significant overlap between BMI categories. A plot of the first and second components are shown in Figure 4.11a. A PLS-DA model was also applied to distinguish individuals according to their Framingham 10-year risk for CVD. However, again, lipid metabolites were unable to effectively discriminate individuals who were at low, moderate, or high risk of CVD, as shown in the plot of the first and second components in Figure 4.11b.

4.3.6 Association of lipid metabolites with circulating biomarkers

For the lipids within each overall lipid category that were most strongly associated with rs662799 (chr11:116663707) in the *APOA5* region [i.e. TG(53:3) (m/z 888.8016) for glycerolipids, PC-O(39:3) or PC-P(39:2) (m/z 812.6532) for glycerophospholipids, SM(42:3) (m/z 811.6688) for sphingolipids, and CE(20:3) (m/z 692.6339) for sterol lipids], the cross-correlation of each lipid metabolite with a wide range of major lipids and other circulating

Figure 4.10: Association of lipid metabolites with smoking status and physical activity

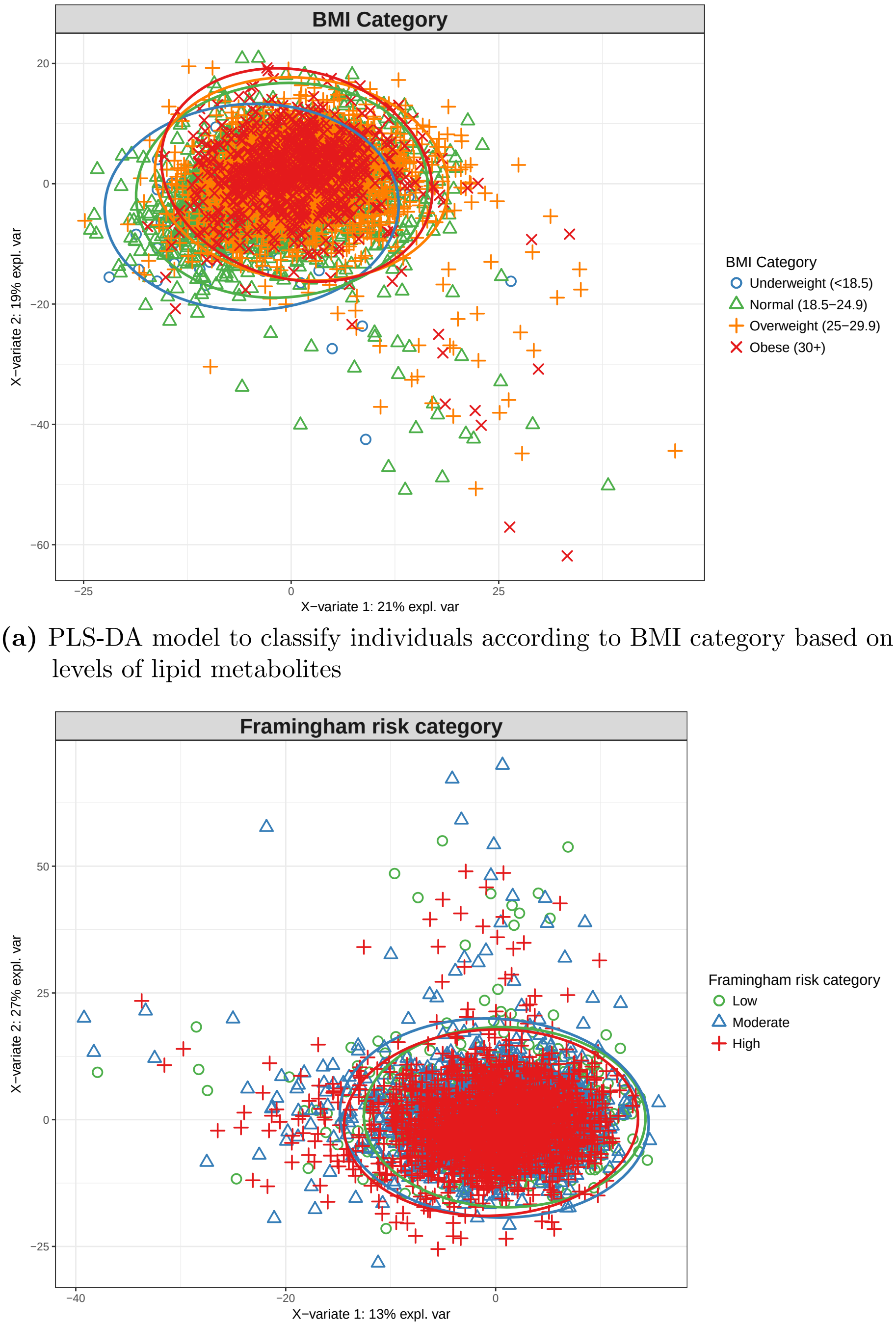


(a) Smoking status

(b) Physical activity

All analyses were adjusted for age and sex. Out of the lipids that were associated with rs662799 in the *APOA5* locus, results are shown for the top twenty lipids that were most significantly associated with smoking status and the top twenty lipids that were most significantly associated with physical activity.

Figure 4.11: Separation of individuals according to body mass index category and Framingham 10-year CVD risk based on a lipid score using partial least-squares discriminant analysis



(a) PLS-DA model to classify individuals according to BMI category based on levels of lipid metabolites

(b) PLS-DA model to classify individuals according to Framingham 10-year CVD risk category based on levels of lipid metabolites

Body mass index (BMI) was classified as underweight ($< 18.5\,\mathrm{kg/m^2}$), normal ($18.5\,\mathrm{kg/m^2}$ to $24.9\,\mathrm{kg/m^2}$), overweight ($25\,\mathrm{kg/m^2}$ to $29.9\,\mathrm{kg/m^2}$), or obese ($\geq 30\,\mathrm{kg/m^2}$). Framingham 10-year relative risk for cardiovascular disease (CVD) was calculated using an algorithm based on age, sex, high-density lipoprotein cholesterol (HDL-C) levels, total cholesterol levels, systolic blood pressure (SBP), smoking status, and diabetes status[187]. The CVD risk score was then classified as low ($< 10\,\%$), moderate ($10\,\%$ to $20\,\%$), or high ($\geq 30\,\%$) risk.

biomarkers was determined. The biomarkers that were examined included total cholesterol, LDL-C, HDL-C, $\mathrm{HbA_{1c}}$, apolipoprotein B, and C-reactive protein (Figure 4.12).

The free fatty acid was inversely associated with total cholesterol, non-HDL cholesterol, LDL-C, triglycerides, ApoB, ApoC3, ApoE, and several other biomarkers (Figure 4.12a).

Triglyceride TG(53:3) (Figure 4.12b) showed a significant positive correlation with levels of major circulating triglycerides, as would be expected, but also with ApoB, ApoC3 and ApoE, total cholesterol, non-HDL cholesterol, and several other biomarkers. This triglyceride also exhibited a significant negative correlation with HDL-C and ApoA1.

For both sphingolipids (Figure 4.12d) and sterol lipids (Figure 4.12e), the strongest inverse associations were found with major circulating triglycerides, ApoC3, and ApoE. Sphingolipids had the strongest correlation with HDL-C, while sterol lipids esters had the strongest association with LDL-C.

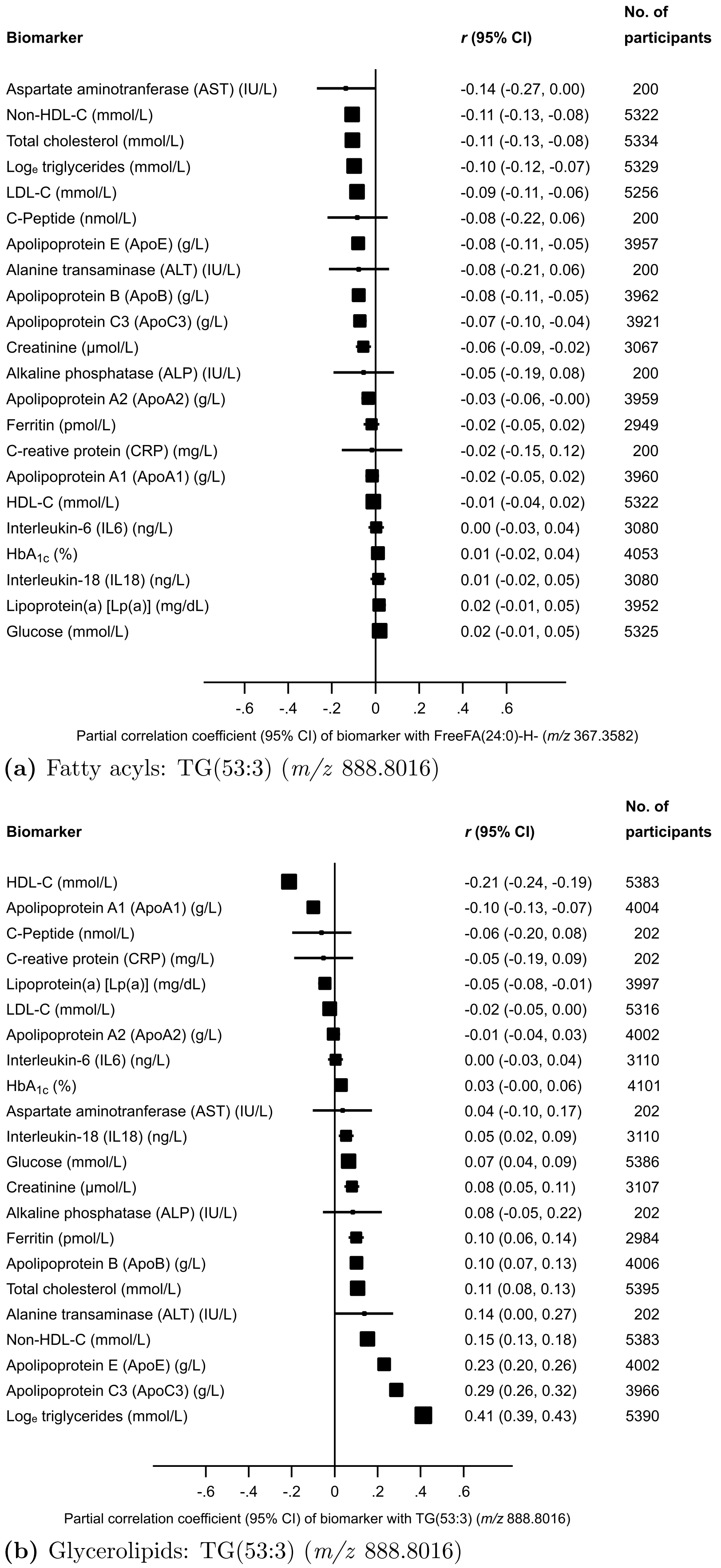
4.3.7 Association of lipid metabolites with intermediate outcomes

A plot displaying graphical representations of the ORs and associated 95 % CIs for associations of the second, third, and fourth principal components with several intermediate outcomes that are risk factors for CHD (i.e. overweight, obese, hypertension, and diabetes) is shown in Figure 4.13.

Individuals who were overweight or diabetic—i.e. with high levels of CHD risk factors—were more likely to have lipid profiles similar to those corresponding to the second principal component, whereas individuals who were not overweight or diabetic and did not have hypertension—i.e. at reduced risk of CHD—were more likely to have lipid profiles matching the third or fourth principal components. For example, a 1-SD increase in the loading scores of the lipids that made up the third principal component resulted in a 20 % reduction in the risk of being overweight (OR = 0.80, 95 % CI 0.76–0.84) and a 33 % reduced risk of having diabetes according to levels of $\mathrm{HbA_{1c}}$ (OR = 0.77, 95 % CI 0.72–0.81), which is a reflection of long-term blood glucose levels rather than short-term fluctuations that are measured by fasting plasma glucose levels.

Individuals with a pattern of lipid metabolite levels that could primarily be explained by the third principal component had a more than 20 % reduced risk of being overweight (OR = 0.80, 95 % CI 0.76–0.84). According to $\mathrm{HbA_{1c}}$ levels, the third and fourth principal components were both associated with a 33 % reduced risk of having diabetes (OR = 0.77, 95 % CI 0.72–0.81 and 0.73–0.82, respectively).

Figure 4.12: Cross-correlations of circulating biomarkers with the lipids within each overall lipid category most strongly associated with rs662799 in the *APOA5* locus



Biomarker	*r* (95% CI)	No. of participants
Aspartate aminotranferase (AST) (IU/L)	-0.14 (-0.27, 0.00)	200
Non-HDL-C (mmol/L)	-0.11 (-0.13, -0.08)	5322
Total cholesterol (mmol/L)	-0.11 (-0.13, -0.08)	5334
Log$_e$ triglycerides (mmol/L)	-0.10 (-0.12, -0.07)	5329
LDL-C (mmol/L)	-0.09 (-0.11, -0.06)	5256
C-Peptide (nmol/L)	-0.08 (-0.22, 0.06)	200
Apolipoprotein E (ApoE) (g/L)	-0.08 (-0.11, -0.05)	3957
Alanine transaminase (ALT) (IU/L)	-0.08 (-0.21, 0.06)	200
Apolipoprotein B (ApoB) (g/L)	-0.08 (-0.11, -0.05)	3962
Apolipoprotein C3 (ApoC3) (g/L)	-0.07 (-0.10, -0.04)	3921
Creatinine (µmol/L)	-0.06 (-0.09, -0.02)	3067
Alkaline phosphatase (ALP) (IU/L)	-0.05 (-0.19, 0.08)	200
Apolipoprotein A2 (ApoA2) (g/L)	-0.03 (-0.06, -0.00)	3959
Ferritin (pmol/L)	-0.02 (-0.05, 0.02)	2949
C-reative protein (CRP) (mg/L)	-0.02 (-0.15, 0.12)	200
Apolipoprotein A1 (ApoA1) (g/L)	-0.02 (-0.05, 0.02)	3960
HDL-C (mmol/L)	-0.01 (-0.04, 0.02)	5322
Interleukin-6 (IL6) (ng/L)	0.00 (-0.03, 0.04)	3080
HbA$_{1c}$ (%)	0.01 (-0.02, 0.04)	4053
Interleukin-18 (IL18) (ng/L)	0.01 (-0.02, 0.05)	3080
Lipoprotein(a) [Lp(a)] (mg/dL)	0.02 (-0.01, 0.05)	3952
Glucose (mmol/L)	0.02 (-0.01, 0.05)	5325

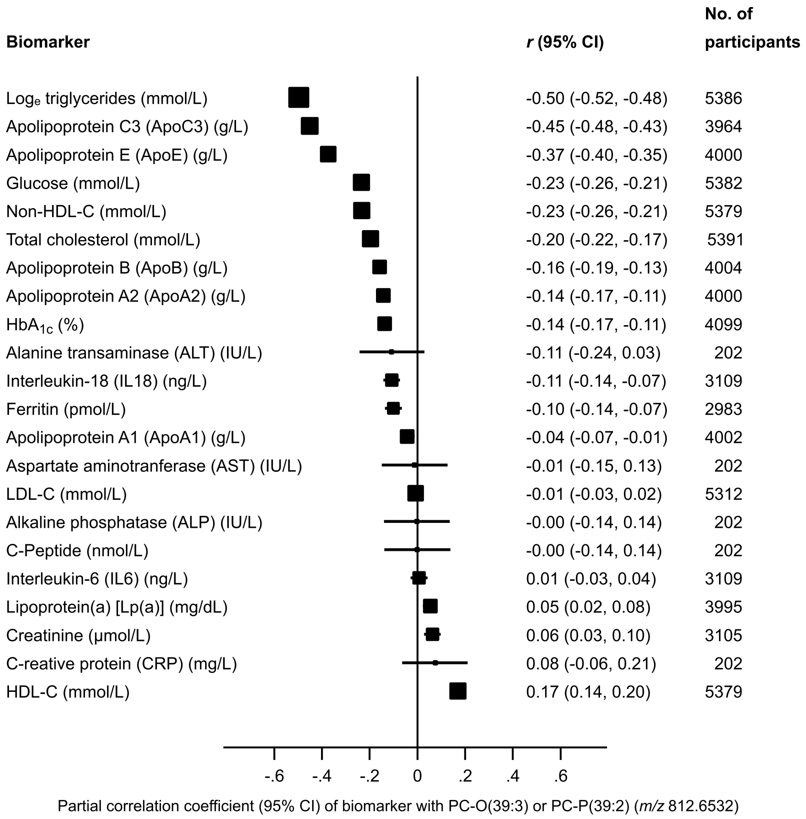
-.6
-.4
-.2
0
.2
.4
.6
Partial correlation coefficient (95% CI) of biomarker with FreeFA(24:0)-H- (*m/z* 367.3582)

(a) Fatty acyls: TG(53:3) (*m/z* 888.8016)

Biomarker	*r* (95% CI)	No. of participants
HDL-C (mmol/L)	-0.21 (-0.24, -0.19)	5383
Apolipoprotein A1 (ApoA1) (g/L)	-0.10 (-0.13, -0.07)	4004
C-Peptide (nmol/L)	-0.06 (-0.20, 0.08)	202
C-reative protein (CRP) (mg/L)	-0.05 (-0.19, 0.09)	202
Lipoprotein(a) [Lp(a)] (mg/dL)	-0.05 (-0.08, -0.01)	3997
LDL-C (mmol/L)	-0.02 (-0.05, 0.00)	5316
Apolipoprotein A2 (ApoA2) (g/L)	-0.01 (-0.04, 0.03)	4002
Interleukin-6 (IL6) (ng/L)	0.00 (-0.03, 0.04)	3110
HbA$_{1c}$ (%)	0.03 (-0.00, 0.06)	4101
Aspartate aminotranferase (AST) (IU/L)	0.04 (-0.10, 0.17)	202
Interleukin-18 (IL18) (ng/L)	0.05 (0.02, 0.09)	3110
Glucose (mmol/L)	0.07 (0.04, 0.09)	5386
Creatinine (µmol/L)	0.08 (0.05, 0.11)	3107
Alkaline phosphatase (ALP) (IU/L)	0.08 (-0.05, 0.22)	202
Ferritin (pmol/L)	0.10 (0.06, 0.14)	2984
Apolipoprotein B (ApoB) (g/L)	0.10 (0.07, 0.13)	4006
Total cholesterol (mmol/L)	0.11 (0.08, 0.13)	5395
Alanine transaminase (ALT) (IU/L)	0.14 (0.00, 0.27)	202
Non-HDL-C (mmol/L)	0.15 (0.13, 0.18)	5383
Apolipoprotein E (ApoE) (g/L)	0.23 (0.20, 0.26)	4002
Apolipoprotein C3 (ApoC3) (g/L)	0.29 (0.26, 0.32)	3966
Log$_e$ triglycerides (mmol/L)	0.41 (0.39, 0.43)	5390

-.6
-.4
-.2
0
.2
.4
.6
Partial correlation coefficient (95% CI) of biomarker with TG(53:3) (*m/z* 888.8016)

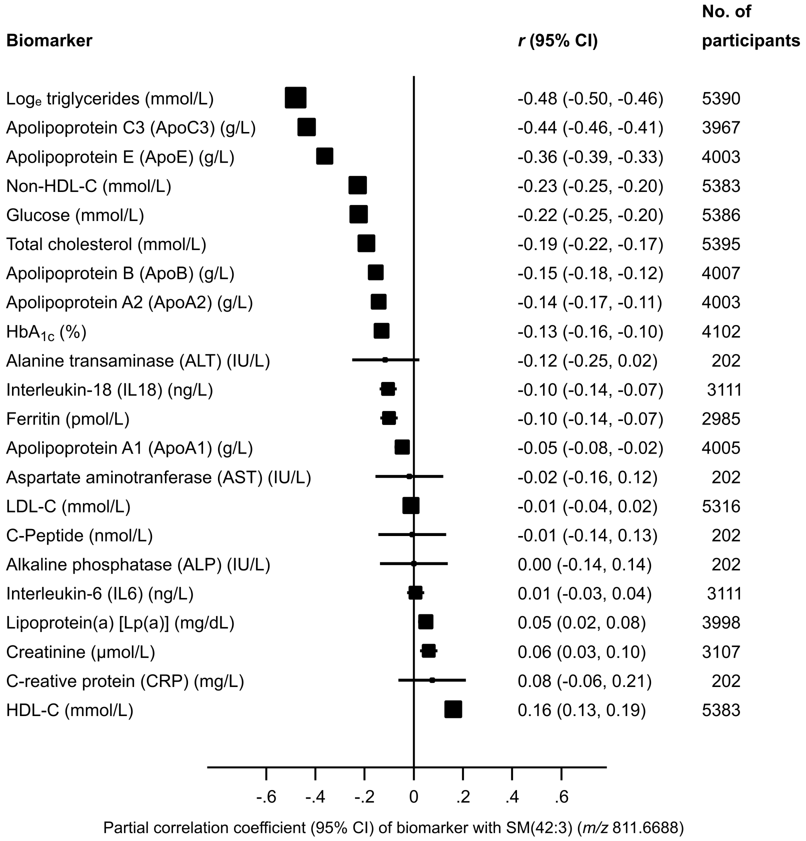
(b) Glycerolipids: TG(53:3) (*m/z* 888.8016)



Biomarker	*r* (95% CI)	No. of participants
Log$_e$ triglycerides (mmol/L)	-0.50 (-0.52, -0.48)	5386
Apolipoprotein C3 (ApoC3) (g/L)	-0.45 (-0.48, -0.43)	3964
Apolipoprotein E (ApoE) (g/L)	-0.37 (-0.40, -0.35)	4000
Glucose (mmol/L)	-0.23 (-0.26, -0.21)	5382
Non-HDL-C (mmol/L)	-0.23 (-0.26, -0.21)	5379
Total cholesterol (mmol/L)	-0.20 (-0.22, -0.17)	5391
Apolipoprotein B (ApoB) (g/L)	-0.16 (-0.19, -0.13)	4004
Apolipoprotein A2 (ApoA2) (g/L)	-0.14 (-0.17, -0.11)	4000
HbA$_{1c}$ (%)	-0.14 (-0.17, -0.11)	4099
Alanine transaminase (ALT) (IU/L)	-0.11 (-0.24, 0.03)	202
Interleukin-18 (IL18) (ng/L)	-0.11 (-0.14, -0.07)	3109
Ferritin (pmol/L)	-0.10 (-0.14, -0.07)	2983
Apolipoprotein A1 (ApoA1) (g/L)	-0.04 (-0.07, -0.01)	4002
Aspartate aminotranferase (AST) (IU/L)	-0.01 (-0.15, 0.13)	202
LDL-C (mmol/L)	-0.01 (-0.03, 0.02)	5312
Alkaline phosphatase (ALP) (IU/L)	-0.00 (-0.14, 0.14)	202
C-Peptide (nmol/L)	-0.00 (-0.14, 0.14)	202
Interleukin-6 (IL6) (ng/L)	0.01 (-0.03, 0.04)	3109
Lipoprotein(a) [Lp(a)] (mg/dL)	0.05 (0.02, 0.08)	3995
Creatinine (μmol/L)	0.06 (0.03, 0.10)	3105
C-reative protein (CRP) (mg/L)	0.08 (-0.06, 0.21)	202
HDL-C (mmol/L)	0.17 (0.14, 0.20)	5379

-.6
-.4
-.2
0
.2
.4
.6
Partial correlation coefficient (95% CI) of biomarker with PC-O(39:3) or PC-P(39:2) (*m/z* 812.6532)

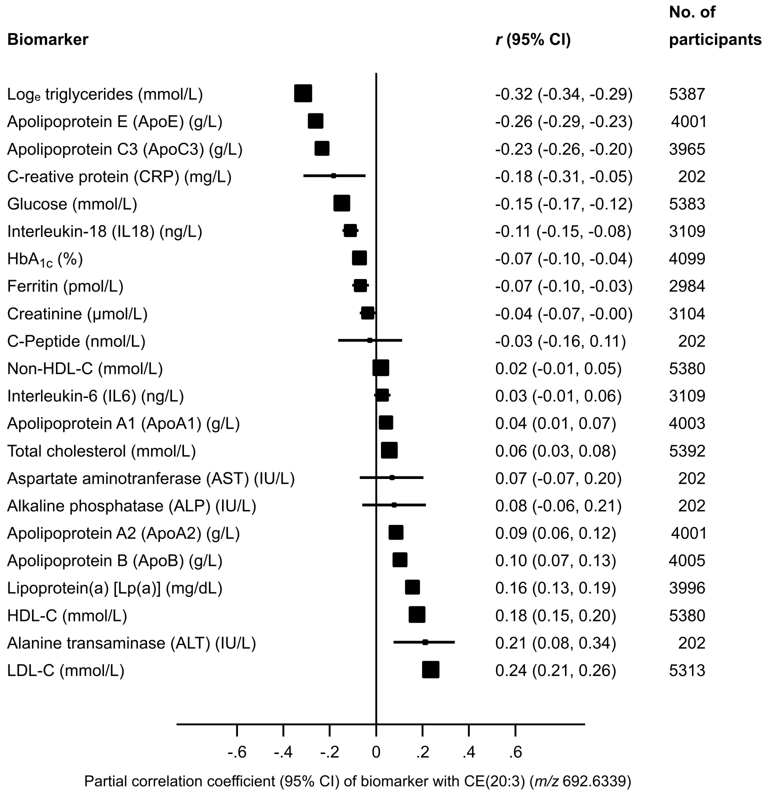
(c) Glycerophospholipids: PC-O(39:3) or PC-P(39:2) (*m/z* 812.6532)



Biomarker	*r* (95% CI)	No. of participants
Log$_e$ triglycerides (mmol/L)	-0.48 (-0.50, -0.46)	5390
Apolipoprotein C3 (ApoC3) (g/L)	-0.44 (-0.46, -0.41)	3967
Apolipoprotein E (ApoE) (g/L)	-0.36 (-0.39, -0.33)	4003
Non-HDL-C (mmol/L)	-0.23 (-0.25, -0.20)	5383
Glucose (mmol/L)	-0.22 (-0.25, -0.20)	5386
Total cholesterol (mmol/L)	-0.19 (-0.22, -0.17)	5395
Apolipoprotein B (ApoB) (g/L)	-0.15 (-0.18, -0.12)	4007
Apolipoprotein A2 (ApoA2) (g/L)	-0.14 (-0.17, -0.11)	4003
HbA$_{1c}$ (%)	-0.13 (-0.16, -0.10)	4102
Alanine transaminase (ALT) (IU/L)	-0.12 (-0.25, 0.02)	202
Interleukin-18 (IL18) (ng/L)	-0.10 (-0.14, -0.07)	3111
Ferritin (pmol/L)	-0.10 (-0.14, -0.07)	2985
Apolipoprotein A1 (ApoA1) (g/L)	-0.05 (-0.08, -0.02)	4005
Aspartate aminotranferase (AST) (IU/L)	-0.02 (-0.16, 0.12)	202
LDL-C (mmol/L)	-0.01 (-0.04, 0.02)	5316
C-Peptide (nmol/L)	-0.01 (-0.14, 0.13)	202
Alkaline phosphatase (ALP) (IU/L)	0.00 (-0.14, 0.14)	202
Interleukin-6 (IL6) (ng/L)	0.01 (-0.03, 0.04)	3111
Lipoprotein(a) [Lp(a)] (mg/dL)	0.05 (0.02, 0.08)	3998
Creatinine (μmol/L)	0.06 (0.03, 0.10)	3107
C-reative protein (CRP) (mg/L)	0.08 (-0.06, 0.21)	202
HDL-C (mmol/L)	0.16 (0.13, 0.19)	5383

-.6
-.4
-.2
0
.2
.4
.6
Partial correlation coefficient (95% CI) of biomarker with SM(42:3) (*m/z* 811.6688)

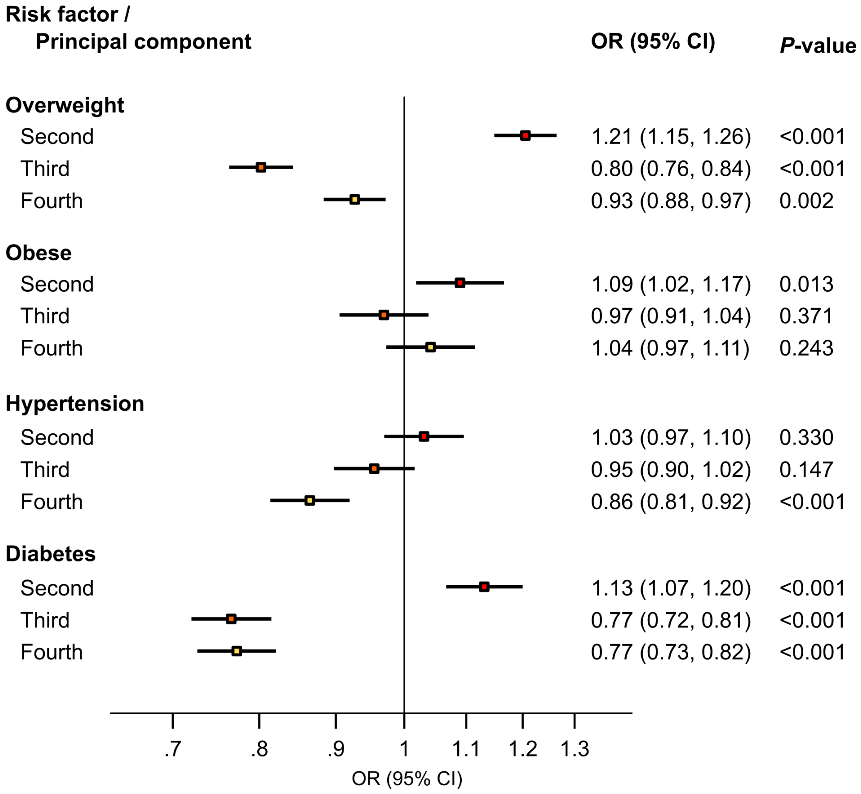
(d) Sphingolipids: SM(42:3) (*m/z* 811.6688)



(e) Sterol lipids: CE(20:3) (m/z 692.6339)

For the lipids within each overall lipid category that were most strongly associated with rs662799 (chr11:116663707) in the *APOA5* region, the correlations of these lipids with a range of circulating biomarkers are shown. Analyses were adjusted for age and sex.

Figure 4.13: Association of established coronary heart disease risk factors with principal components of lipid levels



Risk factor / Principal component	OR (95% CI)	*P*-value
Overweight		
Second	1.21 (1.15, 1.26)	<0.001
Third	0.80 (0.76, 0.84)	<0.001
Fourth	0.93 (0.88, 0.97)	0.002
Obese		
Second	1.09 (1.02, 1.17)	0.013
Third	0.97 (0.91, 1.04)	0.371
Fourth	1.04 (0.97, 1.11)	0.243
Hypertension		
Second	1.03 (0.97, 1.10)	0.330
Third	0.95 (0.90, 1.02)	0.147
Fourth	0.86 (0.81, 0.92)	<0.001
Diabetes		
Second	1.13 (1.07, 1.20)	<0.001
Third	0.77 (0.72, 0.81)	<0.001
Fourth	0.77 (0.73, 0.82)	<0.001

All analyses were adjusted for age and sex. Odds ratios (OR) and 95 % confidence intervals (CI) for each principal component are expressed per 1-SD increase in the loading scores of the lipids that make up that component. **Abbreviations: BMI** = body mass index; **CHD** = coronary heart disease; **DBP** = diastolic blood pressure; **SBP** = systolic blood pressure. **Definitions: Diabetes** = $\text{HbA}_{1c} \geq 6.5\,\%$; **Hypertension** = SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg; **Obese** = BMI $\geq 30\,\text{kg/m}^2$; **Overweight** = BMI $\geq 25\,\text{kg/m}^2$.

4.4 Discussion

The analyses conducted in this chapter showed that lipid metabolites were highly correlated with each other and with levels of major circulating lipids. PCA was used to identify the principal components that explained the majority of the variance in the levels of the metabolites. These principal components were associated with increased levels of several CHD risk factors.

The second principal component revealed a contrast between free fatty acid levels versus small, saturated triglycerides. Several factors could have contributed to this differentiation. Volunteers were recruited at different hospitals and blood samples were taken directly after consent. This means that there was significant variation in the time since participants had eaten their last meal, which would have strongly affected both the free fatty acid and triglyceride pools. The lipid species that contributed most to the second principal component are also affected by obesity and insulin secretion/sensitivity. At the same time, adiposity and insulin secretion will have an effect on free fatty acids levels. The second principal component also demonstrated a significant positive association with the relative likelihood for being overweight and having diabetes.

The third principal component was most closely characterized by unsaturated triglycerides. The loadings of the fourth principal component showed that linoleic acid, whether as a free fatty acid or as one of the fatty acid chains that made up a triglyceride, had negative loading scores, while sphingomyelins containing odd-chain fatty acids and desaturated phospholipids had positive loading scores. Both the third and fourth principal components showed a negative association with the relative risk for being overweight and having diabetes, while only the fourth principal component also showed a negative association with the relative risk for hypertension.

Despite the findings that were able to be drawn from the PCA analysis, PCA does have some inherent limitations. The data are required to be normally distributed, which despite normalisation and log transformation may have not been fully addressed. Furthermore, the principal components can be difficult to interpret and they reflect overall trends and general patterns in the distribution of the variance between the levels of the metabolites, rather than the actual lipid levels, so they have no direct real-world interpretation. Nevertheless, PCA provided a useful tool to examine clusters and patterns in the data and detect interesting associations that can be explored further.

In order to further investigate patterns in the levels of lipid metabolites, PLS-DA was conducted to study the ability of lipid levels in discriminating individuals according to their levels of CHD risk factors and risk of CVD. PLS-DA has been successfully applied to other metabolomics studies[56,182–184,191], so it was anticipated that this approach might expand the knowledge base of the role of lipids in risk of CHD. However, the PLS-DA analysis in this study did not yield any meaningful insights since the risk categories overlapped very closely, even for the Framingham risk score which was partially calculated from levels of major circulating lipids (HDL-C and total cholesterol). Although numerous studies have shown that there are important metabolic differences between individuals with different vascular risk profiles[56,65–67,131,192], in this study it was not possible to disentangle the role that levels of lipid metabolites play in this process, although lipid metabolites are undoubtedly still either directly involved or implicated in atherosclerosis and the onset of cardiovascular diseases. Therefore, the null findings from the PLS-DA indicate that although the principal components obtained using PCA were significantly associated with CHD risk factors, these associations are not sufficient to infer biological or clinical relevance[193] and further research is needed to elucidate the role that lipids play.

In addition to the dimension reduction analysis approach, analyses were also conducted on individual lipids. However, it was beyond the scope of this dissertation to examine the association of each of the 444 lipids with a range of risk factors, as this would have resulted in too many data points that would not have been possible to interpret and present in a concise and coherent manner. This is one of the challenges of working with high-dimensional phenotypic data. Therefore, analyses of individual lipids were limited to subsets that were most likely to be of scientific interest, such as triglycerides and lipids that were significantly associated with a common polymorphism in the *APOA5* cluster known to be associated with major lipid loci and CAD.

The lipids exhibited a range of different associations with circulating biomarkers. Additionally, the analyses of lifestyle factors revealed that that some lipids were positively associated with smoking status and physical activity, while other lipids exhibited an inverse association. For triglycerides, the overall trend was that triglycerides with fewer numbers of double bonds were associated with increased levels of smoking, while triglycerides with a higher number of double bonds were associated with decreased levels of smoking. This trend was also observed somewhat for physical activity although most of the associations were not significant.

The focused analyses that were performed for a subset of the triglycerides and a subset of the lipids significantly associated with *APOA5* could have also been performed for other lipid subclasses and gene regions, so the analyses in this chapter were not by any means comprehensive. However, it provides a detailed overview of the characteristics of the lipid metabolites, their cross-correlations, their correlation with major circulating lipids and with CHD risk factors, and the association of principal components of the lipids with CHD risk factors, as well as insights into a few specific subsets of the lipids.

The GGM analyses that were performed for the lipid subclasses and constituent fatty acid chains of the lipids could have also been performed for each individual lipid or subsets of lipids from various subclasses, which would provide much greater detail and insights into the partial correlations between lipids. However, such analyses were beyond the scope of the epidemiological investigations conducted in this chapter.

This lipidomics platform provides many novel insights into the effect of physiology and behavioural choices on lipid metabolism, and detailed information about the relationship between lipid subfractions and CHD risk factors. This chapter has described the epidemiology of lipid metabolites. In the following chapters, the genetic associations with the lipid metabolites will be examined in detail.

CHAPTER 5

Genome-wide association study of lipid metabolites

Chapter summary

After presenting the epidemiology of lipid metabolites in the last chapter, the current and subsequent chapters focus on the genetic epidemiology of these lipids. This chapter presents results from genome-wide association analyses of the lipids, Chapter 6 describes the annotation of the genetic associations and interpretation of the biological insights that emerged, and Chapter 7 describes the causal relevance of lipids for coronary heart disease (CHD) using genetic variants associated with both lipids and CHD as instrumental variables.

Before conducting the GWAS analyses, the distributions of the log-transformed lipids were examined to ensure approximate log-normality, an appropriate set of adjustment variables was determined, and the covariates were regressed out by calculating the residuals for each model. GWAS analyses were run on the residuals of the regression models for the association of each of the 444 lipids with over 6.7 million genetic variants in 5662 individuals. Manhattan plots and quantile–quantile plots were produced to summarise the overall GWAS results for each lipid, and regional association plots were produced for all gene regions containing one or more genome-wide significant associations. Conditional analyses were then conducted to identify the total number of conditionally independent loci, and finally stringent post-analysis quality control filters were applied.

Out of the 444 lipids that were analysed, 254 lipids (57 %) were significantly associated with one or more genetic variant(s). From the final results of the conditional analyses, there were 355 associations between SNPs and lipids, with a total of 89 sentinel variants from 23 independent loci. For each lipid, the sentinel variant within each locus was defined as the variant with the strongest P-value of association. Heat maps were produced summarising the association of the most strongly associated lipid within each lipid subclass with each genetic locus as a visual representation of the overall genetic findings. The association of lipid metabolites with the 175 genetic loci that are known to be associated with major circulating lipids is also presented in the form of a heat map. Only 13 out of the 175 loci associated with major lipids were also associated with lipid metabolites.

In addition to the primary association analyses of the individual lipids, GWAS were also performed on principal components of the lipids and selected ratios of lipid metabolites with known biological significance. The second, third, and fourth principal components were associated with variants in the *FADS1-2-3* and *APOA5-APOC3* loci, which had already been detected in the univariate GWAS. However, seventeen ratios were significantly associated with one or more variants, and four additional independent loci were identified from analysis of the ratios that were not associated with the individual lipids that made up that ratio.

5.1 Introduction

As described in Section 1.2, the first genome-wide association study (GWAS) was conducted in 2005, and ever since then the number of GWAS studies conducted per year and the number of identified associated traits has been rapidly increasing. At a most basic level, the output from a GWAS is a list of association results between SNPs and one or more traits. This information can then be further analysed and interpreted to lead to more insightful conclusions about the significance of the results and how they can be applied to clinical and pharmacological settings.

GWAS have collectively identified 49 769 unique SNP–trait associations[34]. Out of these, 390 associations involve metabolite measurements as traits. The GWAS analyses described in this chapter for 444 lipids, using a novel lipidomics platform for which results have never before been published, may help contribute towards increasing the number of known associations between lipid metabolites and genetic variants.

5.2 Methods

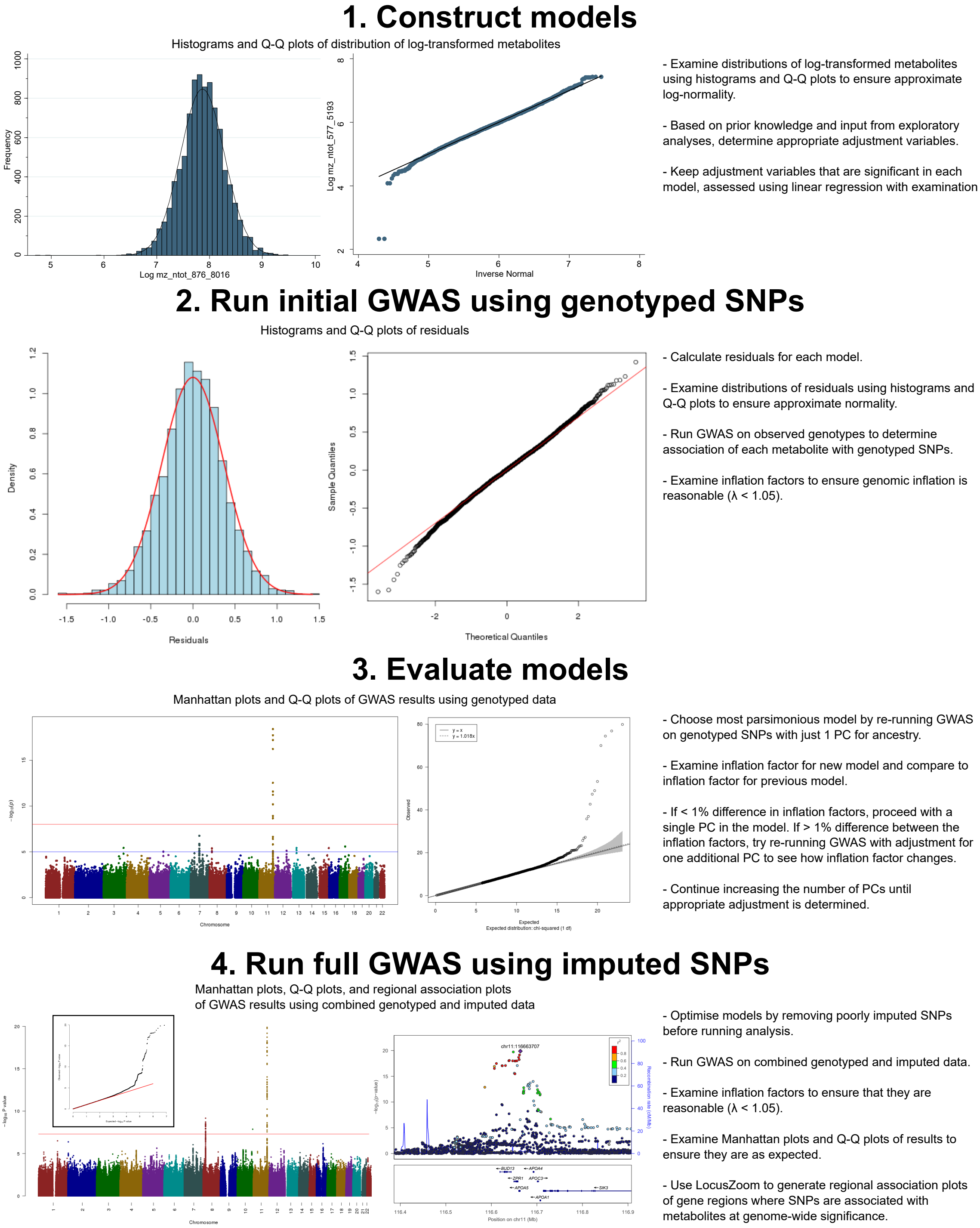
Genome-wide association analyses were run on the residuals of each lipid using linear regression. An analysis plan was developed in advance and followed to ensure that the process was rigorous and not prone to researcher bias. An overview of the procedure used to perform each stage of the GWAS analyses is provided in Figure 5.1 and described in more detail below. For each step of the flowchart, the procedure that was followed for one lipid, TG(52:2) (*m/z* 876.8016), is shown as an example.

5.2.1 Regression models used to perform GWAS

Histograms and Q-Q plots were produced to examine the distributions of the lipids. Since some of the lipids were not normally distributed, natural log transformation was applied to all of the lipids to achieve approximately normal distributions.

An appropriate level of adjustment for all of the lipids was determined based on both prior knowledge and input from exploratory analyses. In addition to consideration of factors that could lead to differences in lipid profiles between individuals, such as age, sex, date of blood sample collection, and fasting status, other technical factors were also considered such as plate number, date and time of assay, and the technician who ran the assay. Linear regression models were constructed with progressive adjustment for each of the potential variables to determine which factors were significant. An additive linear

Figure 5.1: Flow chart of procedure used to perform GWAS



1. Construct models

Histograms and Q-Q plots of distribution of log-transformed metabolites

- Examine distributions of log-transformed metabolites using histograms and Q-Q plots to ensure approximate log-normality.
- Based on prior knowledge and input from exploratory analyses, determine appropriate adjustment variables.
- Keep adjustment variables that are significant in each model, assessed using linear regression with examination

2. Run initial GWAS using genotyped SNPs

Histograms and Q-Q plots of residuals

- Calculate residuals for each model.
- Examine distributions of residuals using histograms and Q-Q plots to ensure approximate normality.
- Run GWAS on observed genotypes to determine association of each metabolite with genotyped SNPs.
- Examine inflation factors to ensure genomic inflation is reasonable (λ < 1.05).

3. Evaluate models

Manhattan plots and Q-Q plots of GWAS results using genotyped data

- Choose most parsimonious model by re-running GWAS on genotyped SNPs with just 1 PC for ancestry.
- Examine inflation factor for new model and compare to inflation factor for previous model.
- If < 1% difference in inflation factors, proceed with a single PC in the model. If > 1% difference between the inflation factors, try re-running GWAS with adjustment for one additional PC to see how inflation factor changes.
- Continue increasing the number of PCs until appropriate adjustment is determined.

4. Run full GWAS using imputed SNPs

Manhattan plots, Q-Q plots, and regional association plots of GWAS results using combined genotyped and imputed data

- Optimise models by removing poorly imputed SNPs before running analysis.
- Run GWAS on combined genotyped and imputed data.
- Examine inflation factors to ensure that they are reasonable (λ < 1.05).
- Examine Manhattan plots and Q-Q plots of results to ensure they are as expected.
- Use LocusZoom to generate regional association plots of gene regions where SNPs are associated with metabolites at genome-wide significance.

Using a single lipid, TG(52:2) (m/z 876.8016), as an example, a flow chart is shown of the steps used to perform the GWAS: (1) The distributions of the lipids were examined and appropriate adjustment variables were determined to construct the models used in the GWAS; (2) The initial GWAS was run using the residuals of the adjusted model on the genotyped data; (3) The model was evaluated based on the inflation factor, Manhattan plot, and Q-Q plot of the GWAS results; (4) Once the model was optimised, the full GWAS was run on the combined genotyped and imputed data, and then Manhattan plots, Q-Q plots (inset), and regional association plots were created.

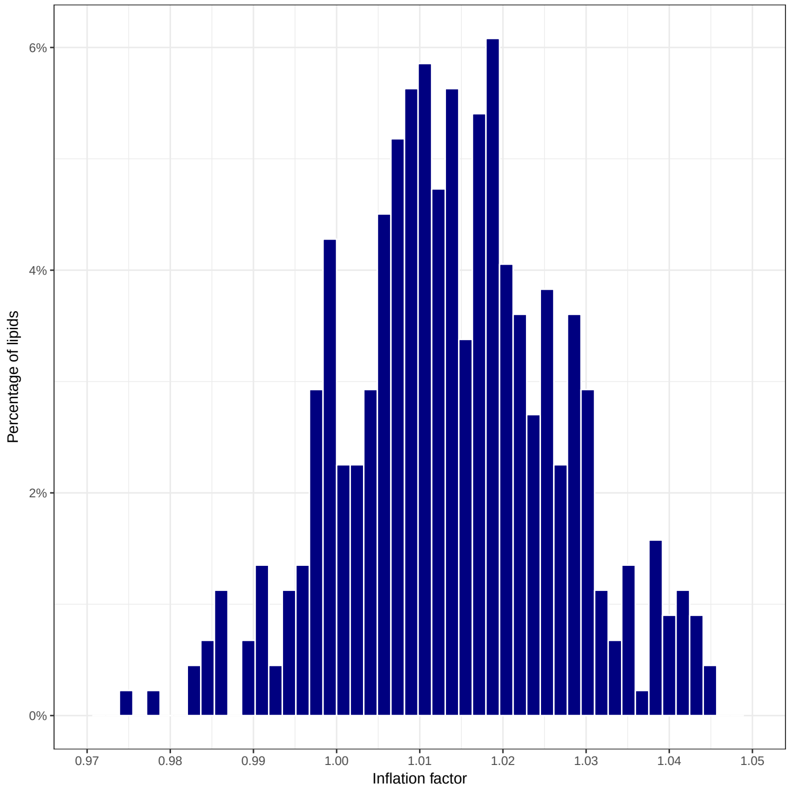
model was used, which assumes a linear additive effect of the minor alleles on the lipid concentrations[148]. For consistency, comparability, and practicality of implementation, the models used for each lipid were adjusted for the same set of covariates.

Since the date of blood sample collection had missing data for some of the participants (12 %), the date of completion of the questionnaire was used instead, which had no missing data. Blood samples were nearly always taken on the same day that the survey was completed, so using the date of survey instead was a suitable approximation. The reason that the date of sample collection was considered to be important is because while the lipidomics assay was conducted in 2013, some of the samples had been collected several years prior—the range of survey dates was from 2005 to 2011 (see Figure 2.3). A concern, especially if the blood samples were subjected to repeated thawing and refreezing in order to conduct various assays and biomarker measurements, was that the length of storage time could have had an important impact on the lipid profiles since levels of lipids in human serum can deteriorate over time. Thawing and refreezing was kept to a minimum over the duration of the study (although the number of times this occurred was not recorded), but adjusting for date of sample collection attempted to take this into account. Otherwise, if levels of some lipids deteriorated more rapidly than others, or certain samples were thawed and refrozen more frequently than other samples, this could have introduced bias or confounding of the genetic associations with each lipid.

Fasting status was an important source of variability to consider because individuals who have recently eaten a high-fat meal would have much higher levels of triglycerides and other lipids, known as postprandial lipaemia[194,195]. Fasting status was categorised as fasting for less than eight hours, fasting for eight or more hours, or fasting for unknown duration (i.e. on the questionnaire they responded that they had fasted but did not provide a length of time since they had eaten their last meal). There were significant differences in fasting status between the two GWAS platforms when including all individuals (chi-square test $P < 0.001$), but when excluding the 266 (5 %) individuals who fasted for an unknown duration, there was no significant difference between fasting status (chi-square test $P = 0.916$) or fasting duration recorded in hours since last meal (t-test $P = 0.744$). Nevertheless, fasting status was included as a variable in the model because of the direct impact that fasting has on lipid levels.

The final set of adjustment variables for each log-transformed lipid was age, sex, date of survey, fasting status, and plate number (batch). Continuous adjustment variables were converted to categorical: age was categorised as <40, 40–49, 50–59, 60–69, or ≥70; date of

Figure 5.2: Histogram of inflation factors across all lipids



survey was categorised by quarter-years (e.g. 2005q1, 2005q2, 2005q3, . . . , 2011q3); and fasting status was categorised as fasting for eight or more hours, less than eight hours, or unknown duration. To account for population stratification and genetic substructure in the data, principal component analysis (PCA) was conducted on the multi-dimensional scaling matrix created from autosomal SNPs (chromosomes 1 to 22, excluding the X and Y sex chromosomes) as implemented in PLINK[196]; the first six principal components were subsequently added to each model. The number of principal components to use was determined by adding each principal component to the model, running the linear regression model to assess whether the factors were significant ($P < 0.05$), and conducting the GWAS to examine the inflation factor. The first six principal components were significant in the majority of the lipids examined and all of the genomic inflation factors (λ) were less than 1.05. Inflation factors were calculated using the "GenABEL" package v1.8-0[197] in the statistical programme R[156]. A histogram of the inflation factors across all lipid metabolites is shown in Figure 5.2, which naturally form a Gaussian distribution centred close to 1.0, as expected. The mean (SD) of the inflation factors was 1.014 (0.013), and all values were between 0.974 and 1.046.

Residuals were calculated from the adjusted models for each lipid, since regressing out the covariates from the models enhanced performance speed and efficiency. The distributions of the residuals were examined using histograms and Q-Q plots in order to ensure that the residuals were approximately normal. A Q-Q plot is a graphical method of comparing probability distributions by plotting the negative logarithm of the observed versus expected P-value for each SNP, which can be used to determine whether there are deviations from the expected P-value distribution. An over-representation of highly significant P-values in the tail area indicates possible true positive associations and that population stratification was adequately controlled[148].

5.2.2 Univariate GWAS

As shown in Figure 5.1, once the models were constructed and residuals were calculated, the GWAS was performed first on the genotyped variants and then on the full set of 6.7 million directly genotyped and imputed variants. This provided the opportunity to confirm that the results of the genotyped SNPs were as expected, double-check the inflation factors, and ensure that the most parsimonious model was being used before scaling up to the full analysis.

Initial GWAS using genotyped SNPs

Linear regression was used to determine the association of each lipid with genotyped SNPs using SNPStats v1.12.0[198], which was performed separately for the samples genotyped on each of the two platforms, GWAS1 and GWAS2. As described in Chapter 2, there were 527 925 genotyped variants on GWAS1 and 643 333 genotyped variants on GWAS2. Q-Q plots of the results were constructed for each lipid and the inflation factors were checked to ensure that the values were not too extreme ($\lambda < 1.05$). Manhattan plots were also constructed to examine the associations of genotyped SNPs with each lipid across the entire genome.

It was important to use the most parsimonious model as possible since over-adjustment with too many inflation factors could lead to reduced statistical power to detect true associations. Therefore, the models that were used for the genotyped SNPs were rerun with adjustment for just a single principal components for ancestry, and the inflation factors between the simple model (one principal component) and the most comprehensively adjusted model (six principal components) were compared. If there was a difference of less than 1 % between the two inflation factors, that would have suggested that adjustment for just a single principal component would have been sufficient. However, since the difference was greater than 1 %, the GWAS on the genotyped SNPs was rerun with adjustment for two principal components to determine how the inflation factors changed. This process was continued until the appropriate number of principal components to adjust for was determined. The end result was that six principal components were retained in the models.

Full GWAS using genotyped and imputed SNPs

After confirming that the results from the genotyped SNPs were sensible and as expected, datasets were generated with the participant IDs listed in the same order as the samples imputed on GWAS1 and GWAS2. SNPTEST v2.4.1[143] was then used to conduct linear regression assessing the association of each lipid with over 6.7 million variants, which consisted of the combined directly genotyped and imputed SNPs. The models that were used for analysis of the genotyped SNPs were also used when analysing the combined genotyped and imputed SNPs, with adjustment for the same set of covariates including six principal components for ancestry.

A missing data likelihood score test was used when testing for association at imputed SNPs to account for genotype uncertainty. This involved calculating an observed data likelihood in which the contribution of each possible genotype was weighted by its imputation probability[143]. The score test then attempted to maximise the likelihood by evaluating the first and second derivatives of the likelihood under the null hypothesis that there is no association, which works well when the log-likelihood is close to a quadratic[143]. If the score test did not produce a sensible result then an EM algorithm was used instead[143].

The imputed data files on each GWAS platform had been split into 566 chunks across all 22 autosomal chromosomes, each chunk consisting of a 5-Mb region of the genome. The analyses were conducted separately and in parallel for each of the 566 chunks on each platform and then the SNPTEST output results were combined into a single results file for GWAS1 and GWAS2. This file was also formatted by extracting relevant statistics and filtered based on cut-offs that were established as part of the analysis plan. The following information was obtained: chromosome, position, SNP name, call rate, Hardy–Weinberg equilibrium (HWE) P-value, imputation information score, effect allele, non-effect allele, effect allele frequency, non-effect allele frequency, MAF, beta (β), standard error, and P-value. Variants were excluded if the HWE P-value was $< 1 \times 10^{-7}$, the call rate was < 0.97, the minor allele frequency (MAF) was < 0.01, or the information (imputation) score was < 0.80.

Since by definition a GWAS is conducted on a genome-wide basis, it implies the hypothesis-free testing of millions or tens of millions of associations of SNPs with a single trait. The now-standard P-value threshold for assessing whether an association result has reached genome-wide significance is 5×10^{-8}, which corresponds to $P < 0.05$ after adjusting for the number of independent variants among the HapMap phase II genotyped SNPs[199]. However, in the advent of high-dimensional phenotyping platforms, it has now become common to conduct multiple GWAS of related traits, thereby requiring a more conservative threshold for genome-wide significance. A straightforward means of adjusting the threshold for genome-wide significance is to apply a Bonferroni correction by dividing the standard genome-wide significance threshold by the number of traits being analysed, which controls the family-wise error rate to reduce the probability of making a type I error (false positives)[12]. However, in the case of metabolomics, where the traits are all highly correlated, this may result in an overly stringent significance threshold, which would yield false negatives since it is likely that many true associations between SNPs and metabolites would be discounted. An approach that was implemented in a recent metabolomics study[78], which has been validated using a permutation test[200], is to conduct PCA on the metabolites and determine the number of principal components that explain at least 95 % of the variance in the levels of the metabolites, and to use this as the correction factor. Although this only serves as an approximation of the effective number of independent tests, it provides a highly accurate estimation of the threshold for multiple-testing correction that would be achieved using the permutation approach, while greatly reducing the computing-time and resources required[200]. To correct for multiple testing in this analysis, a Bonferroni correction was applied using a cut-off for statistical genome-wide significance of $P < 8.929 \times 10^{-10}$, which was calculated by dividing the standard genome-wide significance level (5×10^{-8}) by the number of principal components (56) that explained over 95 % of the variance in the levels of the lipids, as described in Chapter 4.

Manhattan plots are an approach that is commonly employed to graphically display the results of a GWAS. Typically, the position of each analysed SNP in base pairs along each chromosome is displayed on the horizontal axis, while the $-\log_{10}$ P-value for the association of the SNP with the trait of interest is displayed on the vertical axis. The "towers" that may emerge on the figure when there are a large number of highly significant SNPs clustered within particular loci are said to resemble the Manhattan skyline, which is how this style of figure gets its name.

Regional association plots also show $-\log_{10}$ P-values versus chromosomal positions, but they focus on specific genetic regions of interest. Regional association plots are often called "LocusZoom" plots because this is the name of the software package that is most commonly used to produce the plots. They show the lead SNP or "sentinel SNP" (i.e. the SNP with the strongest P-value) along with other SNPs in the same gene region. The location of the genes is annotated, allowing one to examine nearby genes and identify potential causal candidate(s). The other SNPs in the region are coloured according to their degree of linkage disequilibrium (LD) with the lead SNP. A high LD r indicates a non-random association between neighbouring genetic variants, which is used to describe a region of high correlation between SNPs[148].

For all lipids, Manhattan plots and Q-Q plots were constructed and inflation factors were verified. Regional association plots were produced in LocusZoom v1.3[201] using the meta-analysed results for 200-Kb regions upstream and downstream of any SNPs with Bonferroni-corrected genome-wide significant P-values, and summary results were compiled for all significant SNPs.

Computing resources used for genetic analyses

Conducting the genetic association analyses described above required a substantial amount of computing resources. The imputed genetic files for GWAS1 and GWAS2 were each over 100 GB, and these files had to be read into memory for each of the 444 traits analysed. Over 6.7 million association analyses were conducted for each of the 444 traits. On a normal desktop computer, this would have taken several years of computing time to complete the analyses; however, through the use of parallel jobs that were submitted to a high-performance computing cluster, the full set of GWAS results for each lipid was obtained in less than four hours, and analysis of the full set of lipids only took a few days.

For the analyses using only the genotyped data, which had smaller file sizes with fewer variants and thus ran much more quickly, "cardio" was used, which is a computing cluster at the Cardiovascular Epidemiology Unit (CEU) hosted by the High Performance Computing Service (HPCS) at the University of Cambridge. Cardio has six compute nodes connected by 10 Gbit ethernet: one high memory Dell PowerEdge R820 node (1 TB RAM, 4x8 cores, 2.40 GHz Intel Xeon E5-4640), two Dell PowerEdge R720 nodes (256 GB RAM, 2x12 cores, 2.40 GHz Intel Xeon E5-4640) and three Dell PowerEdge R630 nodes (256 GB RAM, 2x10 cores, 2.60 GHz Intel Xeon E5-2660). The cluster is linked to a Lustre file system with 291 TB of usable storage space on Dell MD disk vaults.

For the analyses using the combined genotyped and imputed data, which had significantly larger file sizes with many more variants and thus required increased computing power, the Darwin computing cluster hosted by HPCS was used. 9600 Sandy Bridge cores are provided by a 600 quad server Dell C6220 chassis. Each node consists of two 2.60 GHz eight core Intel Sandy Bridge E5-2670 processors, giving sixteen cores in total, forming a single NUMA (Non-Uniform Memory Architecture) server with 64 GB of RAM (4 GB per core), 376 GB of local storage, and Mellanox FDR ConnectX3 interconnect. Thus, rather than being limited to six computing nodes as was the case on Cardio, the Darwin computing cluster enabled access to up to 600 nodes—albeit shared with many other users across the University of Cambridge—which dramatically sped up the analysis time.

Meta-analysis of univariate GWAS results

Beta estimates and standard errors from the association results of the two genotyping platforms were combined in a fixed-effect inverse-variance weighted meta-analysis using the latest version of METAL (2011-03-25)[202]. This approach took into account the sample sizes and direction of effect of the results obtained using each genotyping platform. For each variant, a reference allele was selected and a z-statistic was calculated that summarised the magnitude and direction of effect relative to the reference allele. The results for both genotyping platforms were aligned to the same reference allele. An overall z-statistic and P-value were calculated from the weighted sum of the individual statistics, with the weights proportional to the square-root of the number of participants in each sample and selected such that the sum of the squares of the weights equalled 1.0.

Although the effect allele and non-effect allele were specified when running METAL, the program has been designed to ignore this information and meta-analyse the results from each study according to its own criteria of which allele should be the effect allele. Fortunately, the programme does align the results to keep the alleles consistent across all the studies that are included in the meta-analysis, but the downside of this approach is that the effect allele that was output from the meta-analysis was sometimes different from the effect allele from the individual studies. For example, if in the input dataset for a given variant, G is the effect allele and A is the reference allele, METAL flips it around so that the β estimate is with respect to A. Therefore, after running the meta-analysis, if the effect allele from the meta-analysis output did not match the effect allele from the two individual studies, then the reference allele and effect allele in the meta-analysis results were swapped and the sign of the β estimate was reversed (i.e. from negative to positive, or vice-versa), while the magnitude was left the same. This ensured that the effect alleles and β coefficients in the combined output results were consistent with the SNPTEST results from the two GWAS platforms.

Compilation of summary GWAS results

Similar procedures that were performed for each of the individual GWAS platforms were also followed for the combined results. Using the meta-analysis results for each lipid, Manhattan plots and Q-Q plots were constructed and inflation factors were verified to ensure that they were reasonable. Regional association plots were produced in LocusZoom v1.3[201] using the meta-analysed results for any SNPs with significant $-\log_{10}$ P-values less than the genome-wide significance level of 5×10^{-8} and any SNPs within 200-Kb of the lead SNP. A Bonferroni correction was applied using a cut-off for statistical genome-wide significance of $P < 8.929 \times 10^{-10}$. Summary results were compiled for any SNPs with $-\log_{10}$ P-values less than the corresponding Bonferroni-corrected P-values.

Conditional analyses of univariate GWAS results

Conditional analyses provide a means to examine GWAS results in further detail by determining if the significant variants within a particular genetic region (locus) are harbouring any secondary, tertiary, or higher-level signals in that same locus or nearby. These further signals can be detected when the original model is conditioned on the sentinel variants. Any variants that are highly correlated with the sentinel variants will drop away (i.e. lose statistical significance), while variants that are not correlated with the sentinel variants, which represent a conditionally independent signal, will still be significant. In order to perform conditional analyses, first genetic association analyses are performed as normal, and then the same models are rerun with adjustment for the same set of covariates along with conditional adjustment for each of the sentinel SNPs, which is used to determine if any secondary signals can be detected. If so, one can adjust for both the primary and secondary signals to determine if there are any tertiary signals, and this process can continue to be repeated until there are no more significant variants.

In this dissertation, the conditional analysis approach described above was applied by manually re-running SNPTEST for each lipid metabolite that had a significant SNP, with conditioning on the SNP with the most significant P-value. Several programmes exist which can perform this process in a semi-automated fashion. One such programme is GCTA (Genome-wide Complex Trait Analysis, http://cnsgenomics.com/software/gcta/). However, this programme requires as input the summary-level statistics from the meta-analysis results since it conducts the conditional analyses genome-wide. Since the imputed genetic files from PROMIS were separated into 5-Mb chunks within each chromosome for convenience and analysis efficiency, it would have been necessary to combine each of the separate chunks into one enormous file for each chromosome in order to feed this into the programme. Instead, it was decided for this dissertation to manually conduct the conditional analyses within each 5-Mb chunk on each chromosome for each lipid in order to keep the workflow more manageable. These regions were wide enough that all of the significant SNPs in each gene region were contained within a single chunk. A limitation of this approach is that it was not possible to test for SNPs in long-range LD that fell outside the 5-Mb chunk, which could theoretically become significant once the sentinel SNPs were conditioned on. However, apart from this, there were not any situations where significant SNPs in the same gene region as the sentinel SNP were excluded because they were on a separate chunk, so the decision to conduct the conditional analyses within 5-Mb chunks was sufficient and fairly robust.

For each lipid, conditional analyses were run based on the results of the meta-analysis following the GWAS. The first round of conditional analyses was completed as follows. From the meta-analysis GWAS results, all SNPs were selected where $P < 8.9 \times 10^{-10}$, the 5-Mb chunks were identified where each of these SNPs were located, and the lead SNPs were selected within each chunk that had the strongest P-value. On an individual lipid basis, for each 5-Mb chunk that was identified, SNPTEST was run on the genotyped and imputed data for each GWAS platform using the same model as before, except also conditioning on the sentinel SNP in the identified chunk. The results from the samples analysed on each genotyping platform were combined in a meta-analysis using METAL as described in subsection 5.2.2, and all SNPs were identified where $P < 8.9 \times 10^{-10}$.

In order to complete the second round of conditional analyses, the sentinel SNP from the meta-analysed results of the first conditional analysis was identified, and the above process was repeated for each chunk. In other words, the original model that was used in the univariate GWAS was implemented for each lipid, with adjustment for both the first sentinel SNP (the most significant variant from the univariate GWAS) and the second sentinel SNP (the most significant variant after conditioning on the first sentinel SNP).

The script was written in such a manner that if there were still significant SNPs remaining, further rounds of conditional analyses would be conducted in the same manner, by continuing to add additional sentinel SNPs to the model for each lipid until there were no more significant SNPs remaining. The final list of SNPs that were conditionally independent for each lipid was identified and combined into a single list across all lipids.

Post-analysis quality control of univariate GWAS results

To verify the robustness and validity of the GWAS results, a number of post-analysis quality control (QC) steps were performed. Essentially, this involved checking the results for the individual GWAS results on each platform and comparing them to the meta-analysis results.

It was important to perform post-analysis QC because the association results contradicted each other for some of the variants on the GWAS1 and GWAS2 platforms. For example, regarding the association of m/z 947.5866 with rs7234716 (chr18:68795397) and the association of m/z 950.6128 with rs66801830 (chr22:29350750), the SNP measured on the GWAS1 array had a strongly significant inverse association with the lipid, while the same SNP measured on the GWAS2 array had a non-significant positive association with the lipid (see Figure 5.3). This suggests that the significant GWAS1 result for this variant may have been a spurious finding (i.e. false positive).

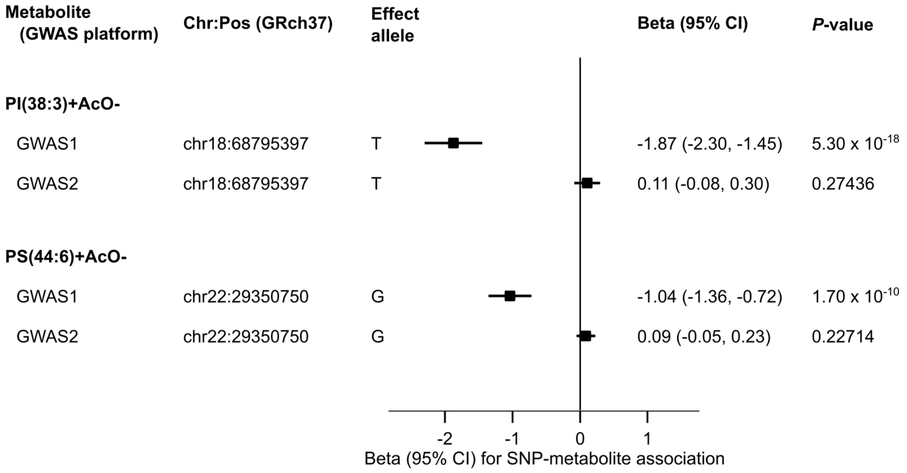
A number of statistics were extracted from the raw SNPTEST results output from both GWAS platforms ("batches"), including the beta coefficient (β), standard error, P-value, effect allele, non-effect allele, MAF, HWE P-value, imputation information score, and call rate. These statistics were compared across the two batches for consistency, and the regional association plots were also checked for each locus to ensure that they had sensible peaks. The Q-Q plots, Manhattan plots, and inflation factors were also re-verified.

Several QC filters had already been applied to the GWAS data from each platform (i.e. HWE $P < 1 \times 10^{-7}$, call rate < 0.97, MAF < 0.01, and imputation information score < 0.80), but for the post-analysis QC, additional filters were applied. The lead SNPs from the meta-analysis were only retained if they (1) had β estimates in the same direction on both platforms (i.e. β estimates were both negative or both positive); and (2) had $P < 0.01$ on both platforms, with $P < 8.9 \times 10^{-10}$ in the meta-analysis. All of the genetics results presented in this dissertation are based on the associations that passed QC.

Grouping independent SNPs into loci

Once the SNPs were identified for both the univariate GWAS results and the conditional analysis results, each set of variants were grouped into loci using ±500-Kb "rolling windows". To define the loci, the variants were sorted in order of position within each chromosome; SNPs on the same chromosome were assigned to the same locus if the 500-Kb windows on either side of two adjacent variants overlapped (in other words, if the variants were within 1-Mb of each other). Additional variants were added to each locus if

Figure 5.3: Examples of spurious SNP associations with lipids



Examples of associations of lipids with SNPs where the results on GWAS1 and GWAS2 differed significantly. Shown in this figure are the association of PI(38:3)+AcO^{-} (*m/z* 947.5866) with rs7234716 (chr18:68795397) (*top*), and the association of PS(44:6)+AcO^{-} (*m/z* 950.6128) with rs66801830 (chr22:29350750) (*bottom*). In both cases, the SNP measured on the GWAS1 array had a highly significant inverse association with the lipid, but the same SNP measured on the GWAS2 array had a non-significant positive association with the lipid. This calls into question whether the association for GWAS1 was legitimate or a false positive; therefore, these associations were excluded from the analysis.

the next variant on the list (in order of position on each chromosome) was within 1-Mb of the previous variant, so that the size of the locus kept expanding (hence the name "rolling windows") until there were no more variants with windows that overlapped the locus. Initially, the name of each locus was determined by concatenating the names of the nearest genes of all the significant variants within each locus (e.g. if the nearest genes to the conditionally independent variants within the locus were *APOA5*, *APOA4*, *APOC3*, and *APOA1*, then the gene was named as *APOA5-APOA4-APOC3-APOA1*). However, after utilising a functional annotation pipeline to identify the most likely causal gene(s) for each variant (which will be described in Chapter 6), the names of each locus were renamed accordingly. For this example, the locus was named *APOA5-APOC3*, which is quite similar and can be understood intuitively, but in the case of rs71661463, although the nearest gene was *ZNRF3*, the predicted causal gene was *XBP1*. The most likely causal genes rather than the nearest genes are presented throughout the remainder of this dissertation.

5.2.3 Association of lipid metabolites with major lipid loci

Table 1.1 lists 175 genetic loci that are known to be associated with major circulating lipids. In order to examine these loci in detail, the GWAS results for all lipid metabolites that were significantly associated ($P < 8.9 \times 10^{-10}$) with one or more of the 175 variants in these loci were extracted. These results were plotted in the form of a heat map.

5.2.4 GWAS of lipid principal components

Secondary discovery analyses were conducted to analyse the association of the first four principal components of the lipid metabolites with genetic variants. Although the traits themselves were different, the adjustment factors used in the models were exactly the same as for the univariate GWAS analyses of the lipids. A meta-analysis was performed to combine results from the two genotyping platforms using a fixed-effect inverse-variance weighted meta-analysis. Since there were fewer statistical tests than for the univariate lipids, the combined results file for each principal component was filtered using the standard threshold for genome-wide significance of $P < 5 \times 10^{-8}$. Conditional analyses of the principal components were not performed.

5.2.5 GWAS of ratios of lipid metabolites

A third discovery step was carried out by testing genome-wide associations on 26 pairwise ratios of lipid concentrations, which are listed in Table 5.1. Ratios were identified through expert curation based on those that had strong biological rationales and that acted through thoroughly understood metabolic pathways. It has been proposed that all pairwise combinations of metabolite ratios can be analysed to gain further biochemical insights into metabolic pathways[72,130,168,203,204], but this requires a much more stringent threshold for genome-wide significance due to the extensive number of statistical tests performed. Although the use of ratios decreases the variance and can therefore lead to decreased P-values[72,73,130], previous mGWAS that have examined all possible combinations of ratios have not found very many additional significant traits since often the metabolites in the numerator and denominator were already identified from the univariate GWAS[132]. Therefore, a subset of ratios with known biological importance was analysed.

Following the same procedures that were implemented for the univariate GWAS of each lipid and the principal components of lipids, likewise for the ratios a meta-analysis was performed and the results that reached the standard threshold for genome-wide significance were analysed. Conditional analyses of the ratios were not performed.

Table 5.1: Ratios of lipid metabolites that were analysed and their biological significance

Numerator		**Denominator**		**Classification**
Name	***m/z***	**Name**	***m/z***	
FreeFA(16:1)	253.2174	FreeFA(18:2)	279.233	Adipose tissue activity
FreeFA(18:3)	277.2174	FreeFA(18:2)	279.233	Desaturase activity
FreeFA(20:3)	305.2487	FreeFa(18:3)	277.2174	Elongase activity
LysoPC(16:0)	496.3404	PE(37:0)	762.6012	Cardiovascular disease risk
LysoPC(17:0)	510.356	LysoPC(18:2)	520.3404	Alpha oxidation
LysoPC(18:2)	520.3404	PC(34:2)	758.57	Lipase activity
PC(32:0)	734.5699	TG(52:3)	874.7859	Insulin production
PC(32:1)	732.5541	CE(18:2)	666.6183	Insulin production
PC(32:1)	732.5541	PC(34:2)	758.57	Palmitolate production
PC(32:1)	732.5541	TG(52:3)	874.7859	Insulin production
PC(34:3)	756.5541	PC(34:2)	758.57	Desaturase
PC(34:3)	756.5541	PE(41:4)	810.6012	Cardiovascular disease risk
PC(36:4)	782.5699	LysoPC(16:0)	496.3404	Inflammation
PC(36:4)	782.5699	PC(34:2)	758.57	ω-6 production
PC(38:5)	808.5851	PC(34:2)	758.57	Eicosapentaenoic acid (EPA) levels
PC(38:6)	806.5694	PC(34:2)	758.57	Docosahexaenoic acid (DHA) levels
PE(34:2)	716.523	CE(20:4)	690.6183	Glucose control
PE(39:2)	786.6012	PC(34:2)	758.57	Elongase
SM(32:1)	675.5434	SM(39:1)	773.6531	Dairy fat intake
SM(42:1)	815.7001	TG(48:2)	820.739	Insulin production
TG(46:0)	796.7393	TG(54:4)	900.8015	*De novo* lipogenesis
TG(48:1)	822.7546	TG(54:5)	898.7856	*De novo* lipogenesis
TG(52:5)	870.7544	TG(52:4)	872.7702	Desaturase
TG(53:2)	890.8172	TG(52:2)	876.8016	Alpha oxidation
TG(54:3)	902.8175	TG(52:3)	874.7859	Elongase
TG(54:5)	898.7856	CE(18:2)	666.6183	C-peptide level

5.3 Results

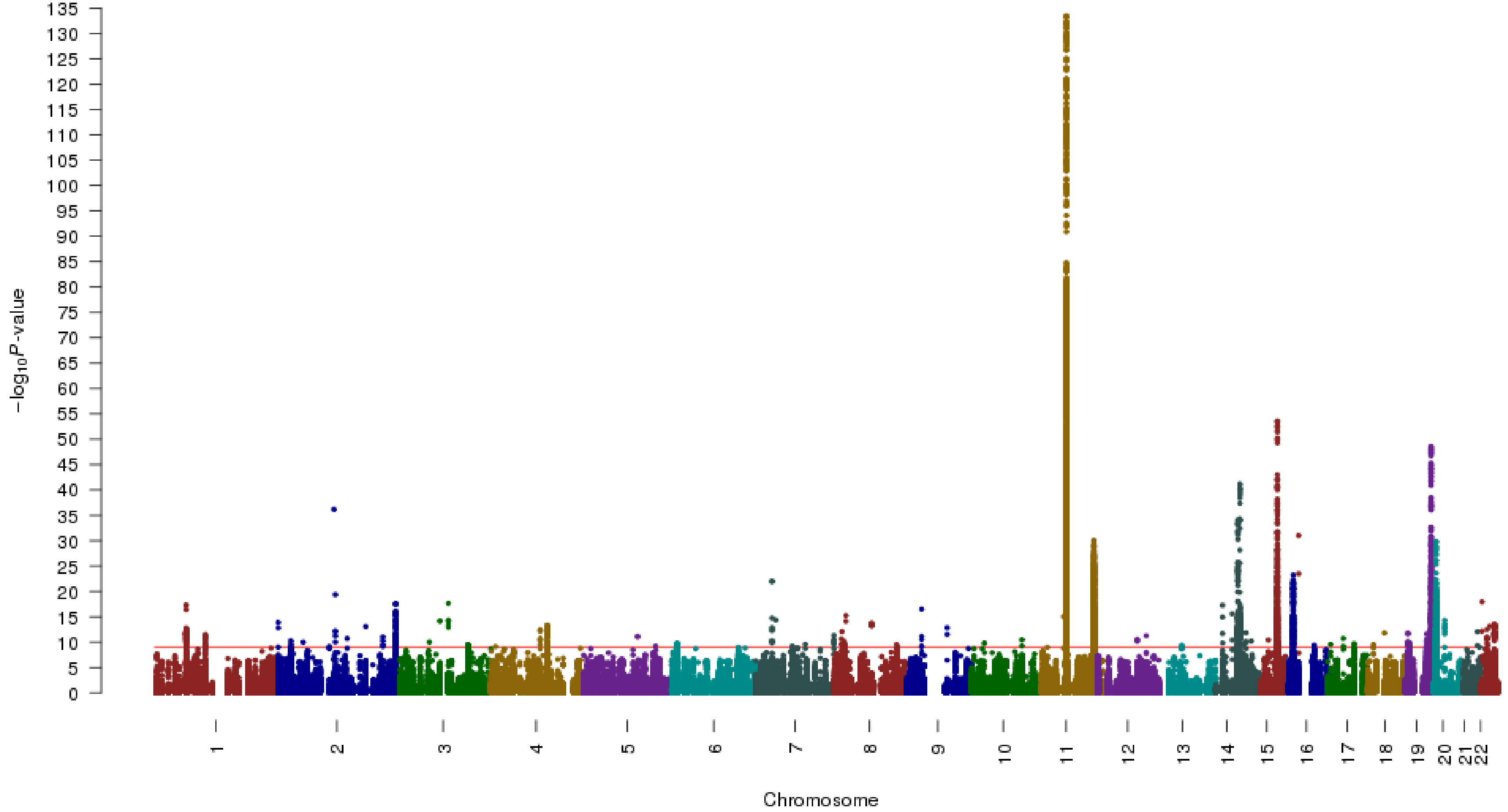
5.3.1 Univariate GWAS results

When the meta-analysis results of the univariate GWAS were combined, there were 14 423 significant associations between SNPs and lipids at the Bonferroni-corrected significance threshold. There were 1727 significant SNPs, and 254 lipids that had at least one significant association.

A global Manhattan plot depicting the combined association of each of the lipids with each SNP is shown in Figure 5.4. Unlike an ordinary Manhattan plot, which shows genome-wide association results for a single trait, in this “global” Manhattan plot the association results were first combined for all 444 traits, then they were plotted together according to P-value and chromosomal position in a single figure. The name of some of the most significant genes are labelled on the diagram. A few of the loci containing SNPs with the strongest association with one or more lipids were *FADS1-2-3*, *APOA5-APOC3*, *PIGH-TMEM229B*, *LIPC*, *MBOAT7*, and *SPTLC3*.

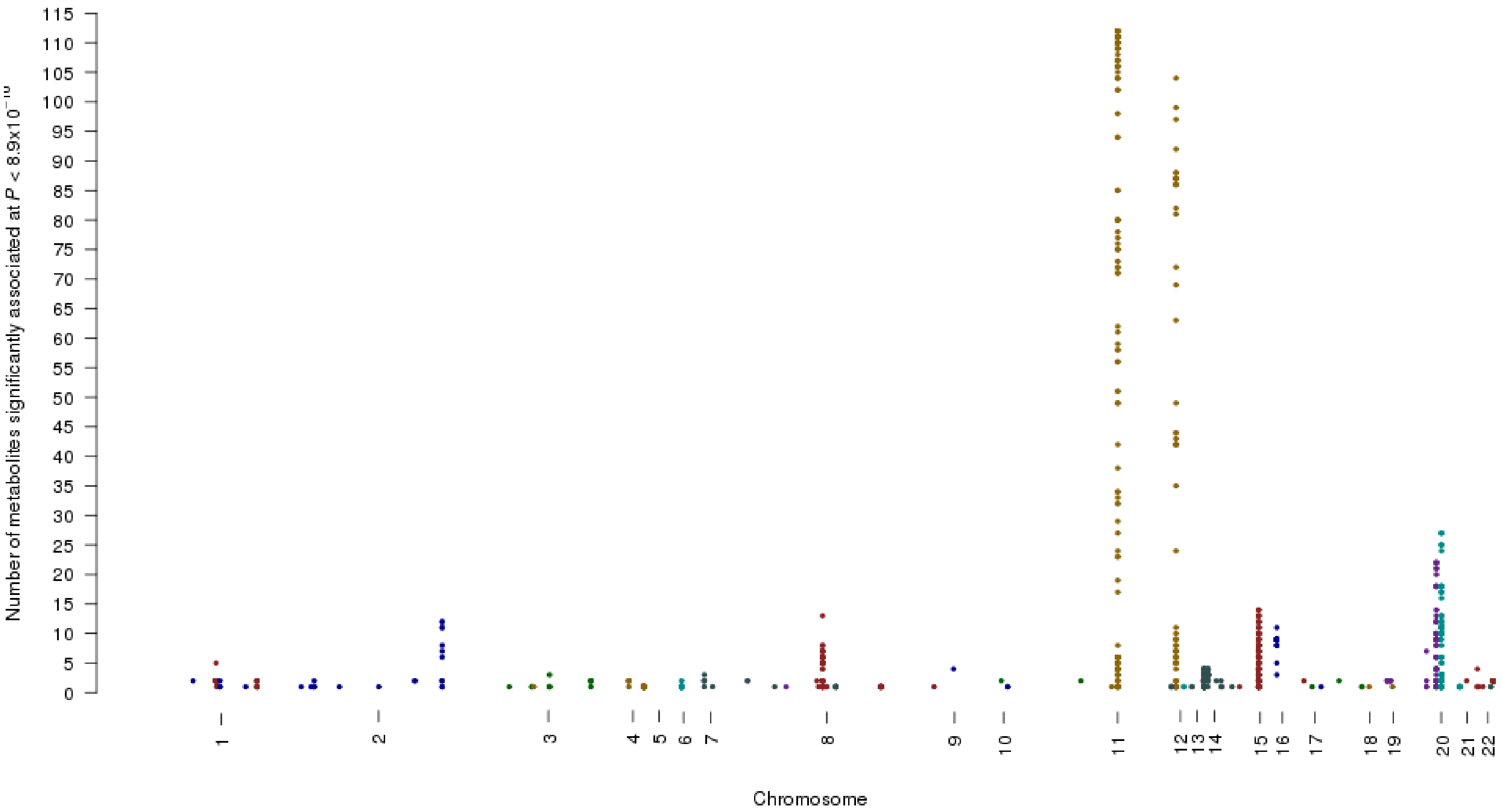
A modified version of this global Manhattan plot, displaying instead on the y-axis the number of lipids associated with each SNP at the Bonferroni-corrected significance threshold, is shown in Figure 5.5. One can clearly see that SNPs in the *FADS1-2-3* and *APOA5-APOC3* regions were associated with the most number of lipids (over 100), which is often referred to as pleiotropy.

Figure 5.4: Global Manhattan plot showing association of 444 lipid metabolites with 6.7 million variants



Manhattan plot of combined results from GWAS analysis for all lipids. P-values are shown for association of each SNP with each lipid. Red line indicates Bonferroni-corrected P-value for genome-wide significance (8.9×10^{-10}).

Figure 5.5: Number of lipids associated with each SNP



As an alternative to a traditional Manhattan plot, this figure, using the combined results from the GWAS analysis for all lipids, shows on the y-axis the number of lipids significantly associated with each SNP at the Bonferroni-corrected significance threshold ($P < 8.9 \times 10^{-10}$).

5.3.2 Conditional analysis results

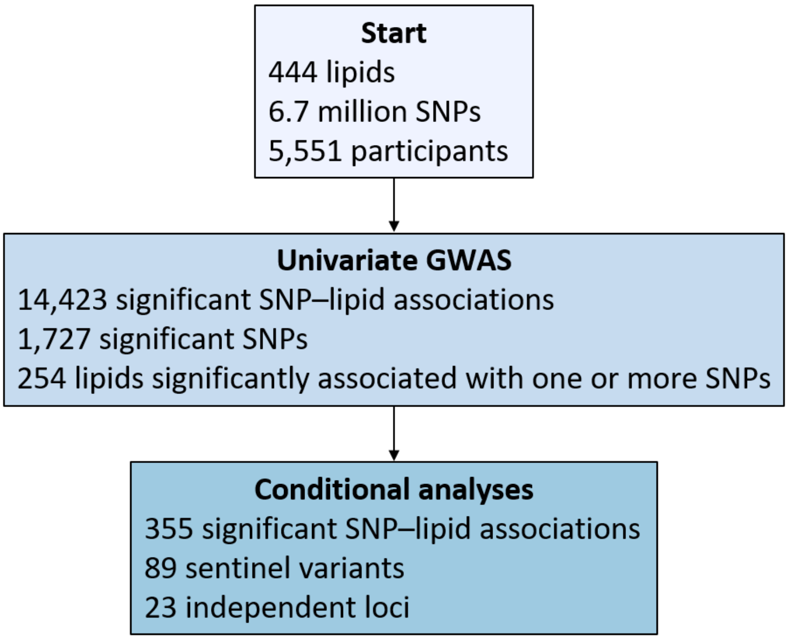
The conditional analyses performed for each lipid, when combined across lipids, identified 355 genome-wide significant SNP–lipid associations at 89 statistically independent SNPs, which corresponded to 23 loci using a distance-based measure of ±500-Kb. A summary of the overall findings from the univariate GWAS and conditional analyses is shown in Figure 5.6.

Most of the genetic loci did not reveal any secondary signals. However, one locus (*LIPC*) harboured secondary conditionally independent associations with 13 lipids and even a tertiary conditionally independent association with one lipid [PE(36:4)-H$^-$ (*m/z* 762.5079)]. Regional association plots from the conditional analyses of this lipid with variants in the *LIPC* region are shown in Figure 5.7 as an example. From the univariate GWAS, the lead SNP was rs1077835 (chr15:58723426), which had a *P*-value of 6.036×10^{-51} (Figure 5.7a). After adjusting for this SNP in the first round of conditional analyses, the new lead SNP was rs2043085 (chr15:58680954), which had a *P*-value of 2.45×10^{-41} (Figure 5.7b). Then after adjusting for both the first and second sentinel variants in the second round of conditional analyses, the new lead SNP was rs11071371 (chr15:58576226), which had a *P*-value of 7.083×10^{-14} (Figure 5.7c). Finally, after adjusting for the first, second, and third sentinel variants in the third round of conditional analyses, the lead variant was rs79341002 (chr15:58707990), which had a *P*-value of 2.012×10^{-8} (Figure 5.7d). Since there were no longer any SNPs remaining that reached genome-wide significance for association with PE(36:4)-H$^-$, the conditional analyses for this lipid stopped after the third round.

The number of significant variants from the conditional analyses that were associated with each lipid subclass is shown in Table 5.2. There were 42 variants from 13 independent loci that were significantly associated with phosphatidylcholines, and 20 variants from 11 independent loci that were significantly associated with sphingomyelins.

A heat map of the effect size and *P*-value for the association of the most strongly associated lipid within each subclass with the sentinel SNPs within each locus is shown in Figure 5.8. Colours indicate the magnitude of the *P*-value from light to dark blue, where genome-wide significant associations ($P < 8.9 \times 10^{-10}$) are shown in the darkest colour. It is evident from the heat map that for certain loci such as *FADS1-2-3*, independent SNPs in this region were associated with lipids in nearly all of the subclasses. Lead SNPs in the *APOA5-APOC3* region were also associated with a wide range of lipid subclasses. In contrast, SNPs in the *PNPLA3* region were only associated with two triglycerides, but not with lipids from any other subclass. The heat map also shows that phosphatidylcholines

Figure 5.6: Results from univariate GWAS and conditional analyses



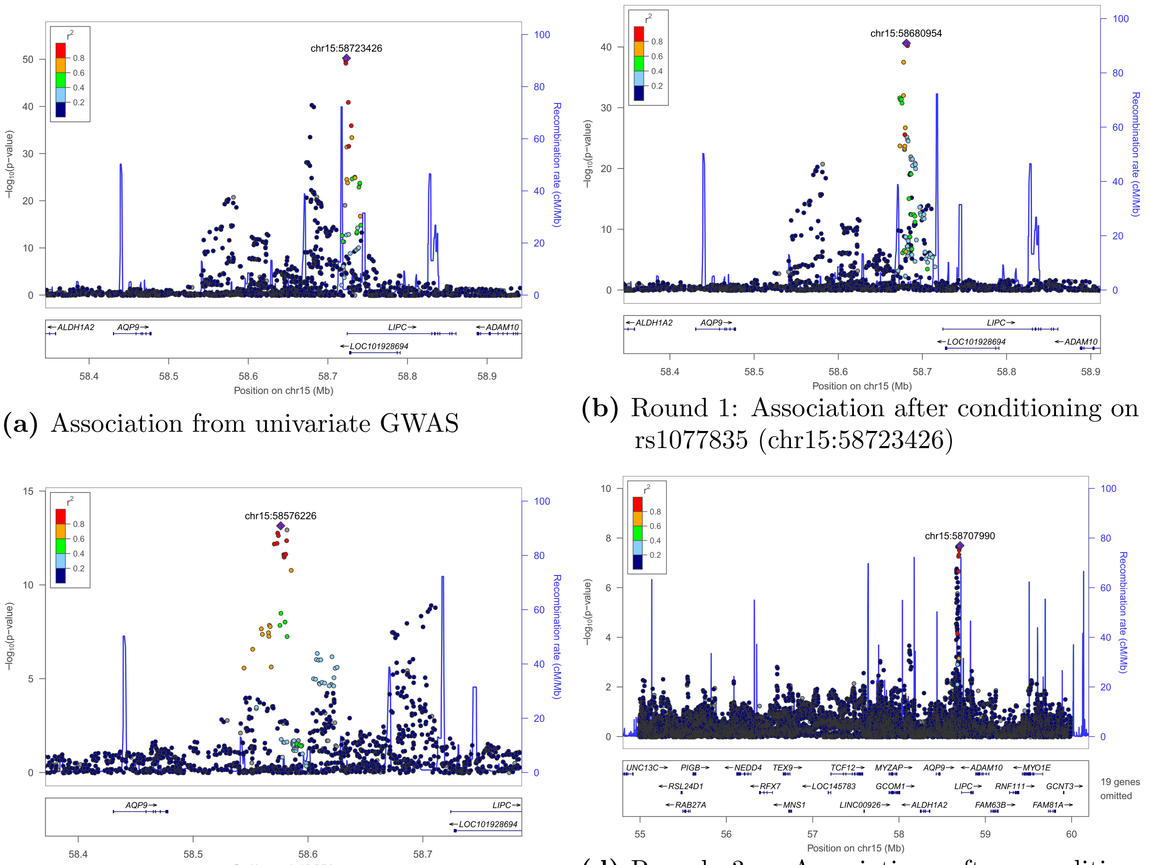
were associated with one or more lead SNPs in the majority of the gene regions, while lipids in other subclasses had greater specificity, as they were only associated with lead SNPs in a single locus. For example, ceramides were only significantly associated with SNPs in the *SPTLC* locus, and cholesterol and derivatives were only significantly associated with SNPs in the *APOA5-APOC3* locus.

5.3.3 Lipidome scan of conditional analysis loci

In Chapter 4, cross-correlations of lipid metabolites with circulating biomarkers were shown for the lipids within each subclass that were most strongly associated with a variant in the *APOA5* locus, and other results were shown for a subset of lipids that were significantly associated with *APOA5*. Following on from those analyses, in this chapter the genetic associations of lipids with all of the significant loci from the conditional analyses were examined in detail. The plots for each locus are shown in Figure 5.9.

For five of the loci (*ANGPTL3*, *APOE-C1-C2-C4*, *CERS4*, *GCKR*, *MLXIPL*), there were consistent associations in the same direction with a similar magnitude for nearly all of the top twenty lipids that were associated with that locus. However, for the remaining 18 loci (*APOA5-APOC3*, *CETP*, *ELOVL2*, *FADS1-2-3*, *GAL3ST1*, *LIPC*, *LPL*, *MBOAT7*, *PAQR9*, *PCTP*, *PIGH-TMEM229B*, *PLA2G10-NTAN1-NPIPA5*, *PNPLA3*, *SCD*, *SGPP1*, *SPTLC3*, *UGT8*, and *XBP1*), the direction and magnitude of the associations varied.

Figure 5.7: Regional association plots showing conditional analyses of PE(36:4)-H⁻ in *LIPC* locus



(a) Association from univariate GWAS

(b) Round 1: Association after conditioning on rs1077835 (chr15:58723426)

(c) Round 2: Association after conditioning on rs1077835 (chr15:58723426) and rs2043085 (chr15:58680954)

(d) Round 3: Association after conditioning on rs1077835 (chr15:58723426), rs2043085 (chr15:58680954), and rs11071371 (chr15:58576226)

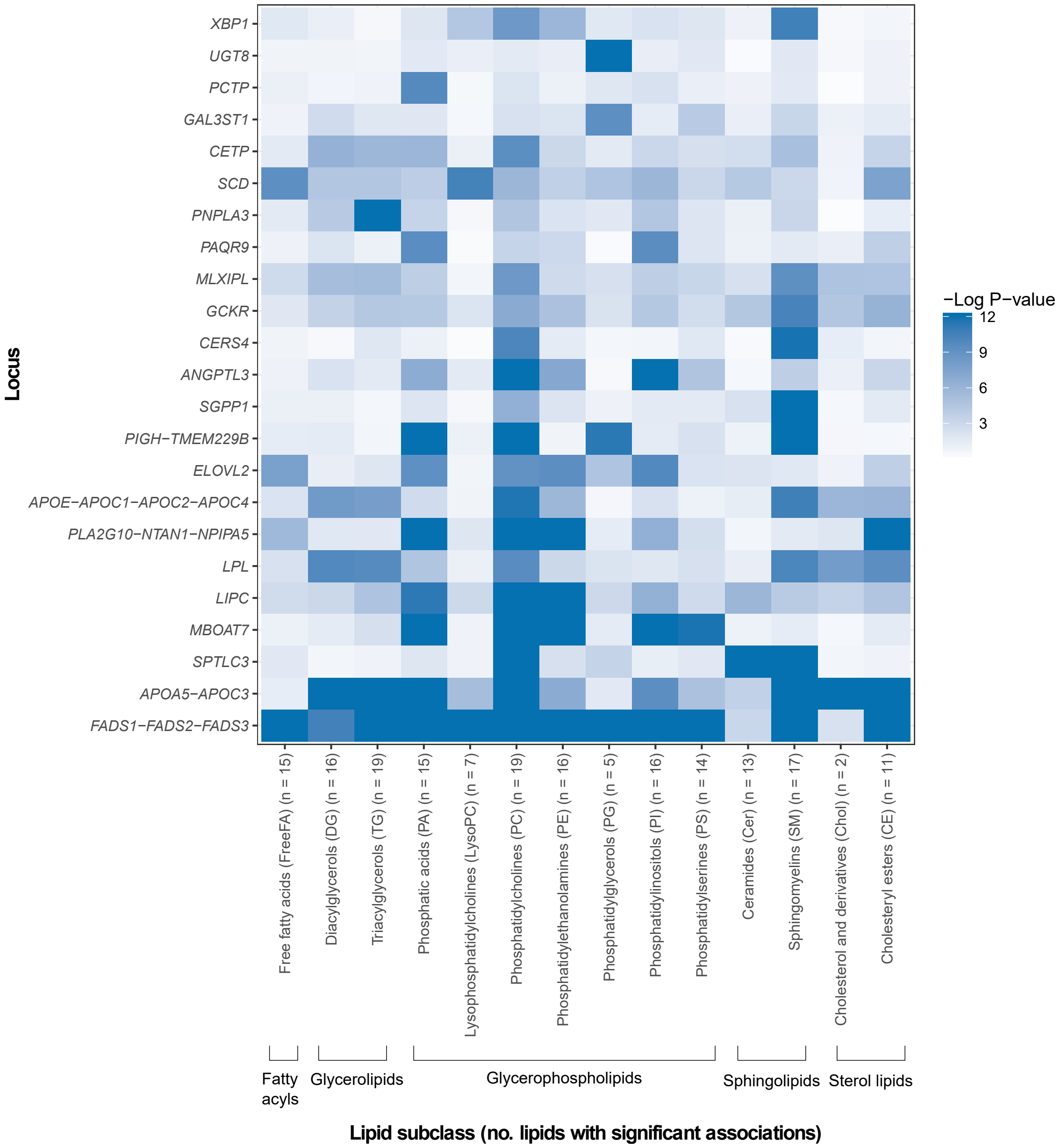
Regional association plots show the association of PE(36:4)-H⁻ (*m/z* 762.5079) with variants in the *LIPC* locus from the univariate GWAS and after conditioning on the sentinel SNP from each round of the conditional analyses. There were no longer any significant variants associated with PE(36:4)-H⁻ after the third round of conditional analyses.

Table 5.2: Number of variants and loci significantly associated with each lipid subclass

Lipid subclass	**No. lipids**	**No. lipids significantly associated with one or more variants**	**No. significant variants**	**No. loci**
Free fatty acids (FreeFA)	22	5 (23 %)	5	2
Diacylglycerols (DG)	19	14 (74 %)	5	2
Triacylglycerols (TG)	56	25 (45 %)	11	4
Phosphatic acids (PA)	33	19 (58 %)	21	9
Lysophosphatidylcholines (LysoPC)	8	2 (25 %)	2	2
Phosphatidylcholines (PC)	106	68 (64 %)	42	13
Phosphatidylethanolamines (PE)	40	22 (55 %)	23	5
Phosphatidylglycerols (PG)	5	3 (60 %)	5	4
Phosphatidylinositols (PI)	25	18 (72 %)	17	5
Phosphatidylserines (PS)	22	5 (23 %)	5	2
Ceramides (Cer)	16	9 (56 %)	4	1
Sphingomyelins (SM)	78	52 (67 %)	20	11
Cholesterol & derivatives (Chol)	2	1 (50 %)	1	1
Cholesteryl esters (CE)	12	11 (92 %)	8	4
Total unique	**444**	**254**	**89**	**23**

Results are shown for the conditional analyses.

Figure 5.8: Heat map showing associations between lipid subclasses and significant loci from conditional analyses



The effect estimates of the associations between significant variants and lipid subclasses are plotted as a heat map for the conditional analysis results. Results are shown for the association of the most strongly associated (smallest P-value) lipid within each lipid subclass with the lead variant within each locus. The rows show the 23 loci while the columns show the 14 lipid subclasses. The magnitude of the P-values of association are indicated by the colour scale from light to dark blue, where genome-wide significant associations ($P < 8.9 \times 10^{-10}$) are shown in the darkest colour. The P-values are capped at a maximum of $P < 1 \times 10^{-12}$ so as to make the differences between significant and non-significant associations more readily apparent. The lipid subclasses are arranged according to their classification by overall category, main class, and then alphabetically by subclass. The loci are arranged in order of the number of lipids significantly associated with each locus, with ties ordered alphabetically.

Some of the lipids were positively associated with the lead variant in that locus and other lipids had an inverse association with the lead variant in that locus.

The association of each lipid with a gain-of-function variant in the *LPL* gene (rs328) (Figure 5.9k) revealed that the magnitude and direction of the association of different lipids with these variants varied significantly. The effect allele for this variant was associated with genome-wide significant decreases ($P < 8.9 \times 10^{-10}$) in levels of multiple diglyceride and triglyceride species and increases in the levels of several different cholesterol esters, sphingomyelins, and phosphocholines (Figure 5.9k). Within the triglycerides, those containing monounsaturated fatty acids within the fatty acid side chains had the greatest magnitude of effect and significance.

The association of each lipid with a common polymorphism (rs662799) in the *APOA5-APOC3* cluster also revealed differences in the magnitude and direction of the association according to overall lipid category (Figure 5.9b). The four glycerolipids (diglycerides and triglycerides) had inverse associations with the effect allele for this variant, while the cholesterol esters, phosphocholines, sphingomyelins, and cholesterol had positive associations.

Meanwhile, in the *APOE-C1-C2-C4* region, the two significantly associated diglycerides had positive associations with the lead variant (rs483082), while the remaining phosphocholines and sphingomyelins had inverse associations (Figure 5.9c).

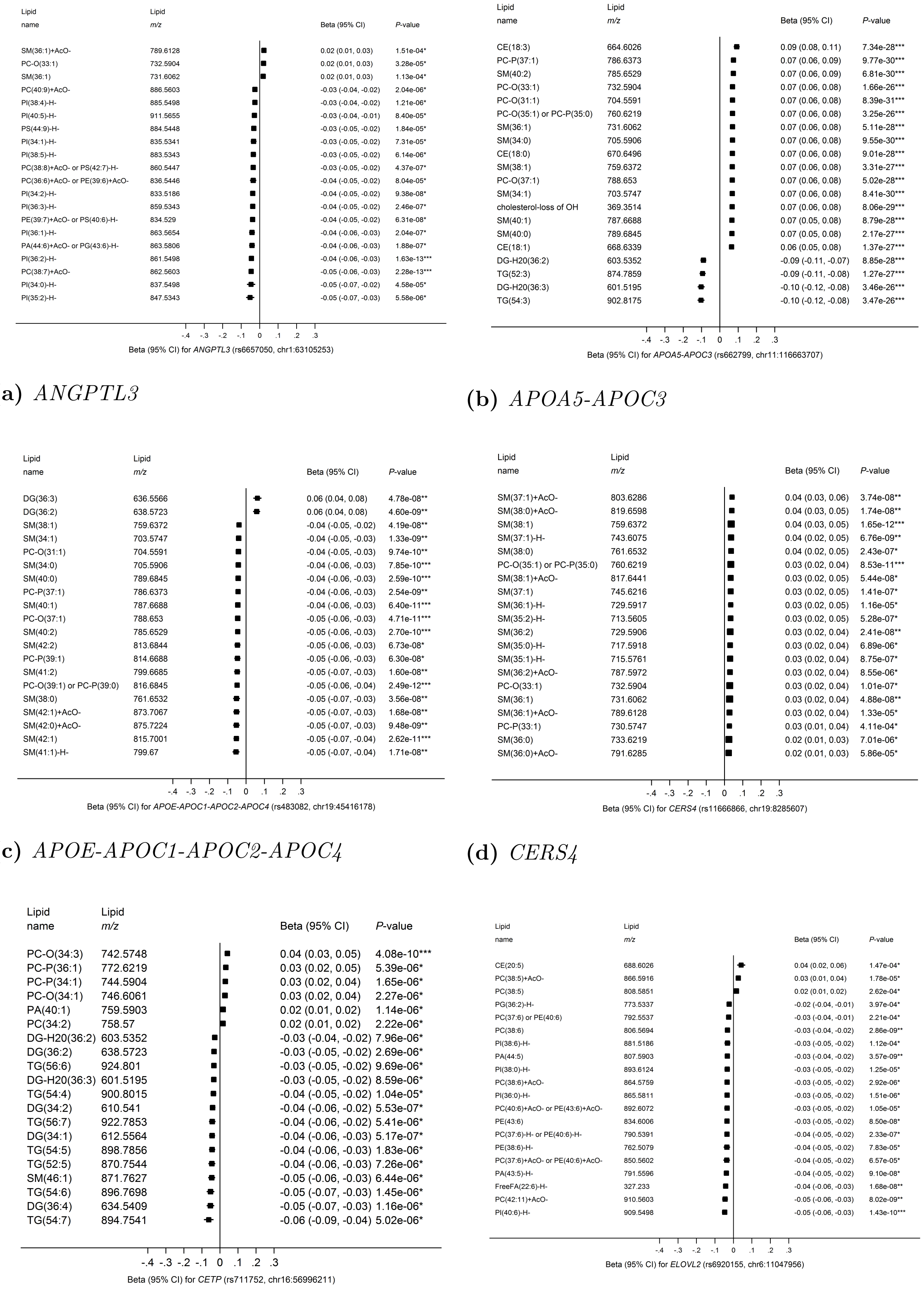
In the *CETP* region, there was a similar division. Phosphoholines and a phosphatic acid had significant positive associations with the lead variant (rs711752), while diglycerides and triglycerides had significant inverse associations (Figure 5.9e).

5.3.4 Association of lipid metabolites with major lipid loci

As related in Chapter 1, at present GWAS have identified 175 genetic loci that are associated with major circulating lipids[19,35–40] (listed in Table 1.1). The majority of these variants reside in non-coding portions of the genome, where the precise function is often not well known. By examining the association of lipid metabolites with these major lipid loci, further insights into lipid metabolism can be identified, which can aid efforts to determine gene functions.

Figure 5.10 shows a heat map of the GWAS results for all lipid metabolites significantly associated with one or more of the 175 variants in these loci. Only 13 of the 175 loci (*FADS1-2-3*, *APOA5-APOC3*, *LOC101929486*, *LPL*, *LIPC*, *PLA2G10*, *GCKR*, *ANGPTL3*, *DOCK6*, *PNPLA3*, *TBL2*, *TRIB1*, and *HERPUD1*) were significantly associated with one or more lipids.

Figure 5.9: Association of the top twenty most significantly associated lipids with the lead variant in each significant locus from the conditional analyses



Lipid name	Lipid *m/z*	Beta (95% CI)	*P*-value
SM(36:1)+AcO-	789.6128	0.02 (0.01, 0.03)	1.51e-04*
PC-O(33:1)	732.5904	0.02 (0.01, 0.03)	3.28e-05*
SM(36:1)	731.6062	0.02 (0.01, 0.03)	1.13e-04*
PC(40:9)+AcO-	886.5603	-0.03 (-0.04, -0.02)	2.04e-06*
PI(38:4)-H-	885.5498	-0.03 (-0.04, -0.02)	1.21e-06*
PI(40:5)-H-	911.5655	-0.03 (-0.04, -0.01)	8.40e-05*
PS(44:9)-H-	884.5448	-0.03 (-0.05, -0.02)	1.84e-05*
PI(34:1)-H-	835.5341	-0.03 (-0.05, -0.02)	7.31e-05*
PI(38:5)-H-	883.5343	-0.03 (-0.05, -0.02)	6.14e-06*
PC(38:8)+AcO- or PS(42:7)-H-	860.5447	-0.03 (-0.05, -0.02)	4.37e-07*
PC(36:6)+AcO- or PE(39:6)+AcO-	836.5446	-0.04 (-0.05, -0.02)	8.04e-05*
PI(34:2)-H-	833.5186	-0.04 (-0.05, -0.02)	9.38e-08*
PI(36:3)-H-	859.5343	-0.04 (-0.05, -0.02)	2.46e-07*
PE(39:7)+AcO- or PS(40:6)-H-	834.529	-0.04 (-0.05, -0.02)	6.31e-08*
PI(36:1)-H-	863.5654	-0.04 (-0.06, -0.03)	2.04e-07*
PA(44:6)+AcO- or PG(43:6)-H-	863.5806	-0.04 (-0.06, -0.03)	1.88e-07*
PI(36:2)-H-	861.5498	-0.04 (-0.06, -0.03)	1.63e-13***
PC(38:7)+AcO-	862.5603	-0.05 (-0.06, -0.03)	2.28e-13***
PI(34:0)-H-	837.5498	-0.05 (-0.07, -0.02)	4.58e-05*
PI(35:2)-H-	847.5343	-0.05 (-0.07, -0.03)	5.58e-06*

(a) *ANGPTL3*

Lipid name	Lipid *m/z*	Beta (95% CI)	*P*-value
CE(18:3)	664.6026	0.09 (0.08, 0.11)	7.34e-28***
PC-P(37:1)	786.6373	0.07 (0.06, 0.09)	9.77e-30***
SM(40:2)	785.6529	0.07 (0.06, 0.09)	6.81e-30***
PC-O(33:1)	732.5904	0.07 (0.06, 0.08)	1.66e-26***
PC-O(31:1)	704.5591	0.07 (0.06, 0.08)	8.39e-31***
PC-O(35:1) or PC-P(35:0)	760.6219	0.07 (0.06, 0.08)	3.25e-26***
SM(36:1)	731.6062	0.07 (0.06, 0.08)	5.11e-28***
SM(34:0)	705.5906	0.07 (0.06, 0.08)	9.55e-30***
CE(18:0)	670.6496	0.07 (0.06, 0.08)	9.01e-28***
SM(38:1)	759.6372	0.07 (0.06, 0.08)	3.31e-27***
PC-O(37:1)	788.653	0.07 (0.06, 0.08)	5.02e-28***
SM(34:1)	703.5747	0.07 (0.06, 0.08)	8.41e-30***
cholesterol-loss of OH	369.3514	0.07 (0.06, 0.08)	8.06e-29***
SM(40:1)	787.6688	0.07 (0.05, 0.08)	8.79e-28***
SM(40:0)	789.6845	0.07 (0.05, 0.08)	2.17e-27***
CE(18:1)	668.6339	0.06 (0.05, 0.08)	1.37e-27***
DG-H20(36:2)	603.5352	-0.09 (-0.11, -0.07)	8.85e-28***
TG(52:3)	874.7859	-0.09 (-0.11, -0.08)	1.27e-27***
DG-H20(36:3)	601.5195	-0.10 (-0.12, -0.08)	3.46e-26***
TG(54:3)	902.8175	-0.10 (-0.12, -0.08)	3.47e-26***

(b) *APOA5-APOC3*

Lipid name	Lipid *m/z*	Beta (95% CI)	*P*-value
DG(36:3)	636.5566	0.06 (0.04, 0.08)	4.78e-08**
DG(36:2)	638.5723	0.06 (0.04, 0.08)	4.60e-09**
SM(38:1)	759.6372	-0.04 (-0.05, -0.02)	4.19e-08**
SM(34:1)	703.5747	-0.04 (-0.05, -0.03)	1.33e-09**
PC-O(31:1)	704.5591	-0.04 (-0.05, -0.03)	9.74e-10**
SM(34:0)	705.5906	-0.04 (-0.06, -0.03)	7.85e-10***
SM(40:0)	789.6845	-0.04 (-0.06, -0.03)	2.59e-10***
PC-P(37:1)	786.6373	-0.04 (-0.06, -0.03)	2.54e-09**
SM(40:1)	787.6688	-0.04 (-0.06, -0.03)	6.40e-11***
PC-O(37:1)	788.653	-0.05 (-0.06, -0.03)	4.71e-11***
SM(40:2)	785.6529	-0.05 (-0.06, -0.03)	2.70e-10***
SM(42:2)	813.6844	-0.05 (-0.06, -0.03)	6.73e-08*
PC-P(39:1)	814.6688	-0.05 (-0.06, -0.03)	6.30e-08*
SM(41:2)	799.6685	-0.05 (-0.07, -0.03)	1.60e-08**
PC-O(39:1) or PC-P(39:0)	816.6845	-0.05 (-0.06, -0.04)	2.49e-12***
SM(38:0)	761.6532	-0.05 (-0.07, -0.03)	3.56e-08**
SM(42:1)+AcO-	873.7067	-0.05 (-0.07, -0.03)	1.68e-08**
SM(42:0)+AcO-	875.7224	-0.05 (-0.07, -0.03)	9.48e-09**
SM(42:1)	815.7001	-0.05 (-0.07, -0.04)	2.62e-11***
SM(41:1)-H-	799.67	-0.05 (-0.07, -0.04)	1.71e-08**

-.4 -.3 -.2 -.1 0 .1 .2 .3
Beta (95% CI) for *APOE-APOC1-APOC2-APOC4* (rs483082, chr19:45416178)

(c) *APOE-APOC1-APOC2-APOC4*

Lipid name	Lipid *m/z*	Beta (95% CI)	*P*-value
SM(37:1)+AcO-	803.6286	0.04 (0.03, 0.06)	3.74e-08**
SM(38:0)+AcO-	819.6598	0.04 (0.03, 0.05)	1.74e-08**
SM(38:1)	759.6372	0.04 (0.03, 0.05)	1.65e-12***
SM(37:1)-H-	743.6075	0.04 (0.02, 0.05)	6.76e-09**
SM(38:0)	761.6532	0.04 (0.02, 0.05)	2.43e-07*
PC-O(35:1) or PC-P(35:0)	760.6219	0.03 (0.02, 0.04)	8.53e-11***
SM(38:1)+AcO-	817.6441	0.03 (0.02, 0.05)	5.44e-08*
SM(37:1)	745.6216	0.03 (0.02, 0.05)	1.41e-07*
SM(36:1)-H-	729.5917	0.03 (0.02, 0.05)	1.16e-05*
SM(35:2)-H-	713.5605	0.03 (0.02, 0.05)	5.28e-07*
SM(36:2)	729.5906	0.03 (0.02, 0.04)	2.41e-08**
SM(35:0)-H-	717.5918	0.03 (0.02, 0.04)	6.89e-06*
SM(35:1)-H-	715.5761	0.03 (0.02, 0.04)	8.75e-07*
SM(36:2)+AcO-	787.5972	0.03 (0.02, 0.04)	8.55e-06*
PC-O(33:1)	732.5904	0.03 (0.02, 0.04)	1.01e-07*
SM(36:1)	731.6062	0.03 (0.02, 0.04)	4.88e-08**
SM(36:1)+AcO-	789.6128	0.03 (0.02, 0.04)	1.33e-05*
PC-P(33:1)	730.5747	0.03 (0.01, 0.04)	4.11e-04*
SM(36:0)	733.6219	0.02 (0.01, 0.03)	7.01e-06*
SM(36:0)+AcO-	791.6285	0.02 (0.01, 0.03)	5.86e-05*

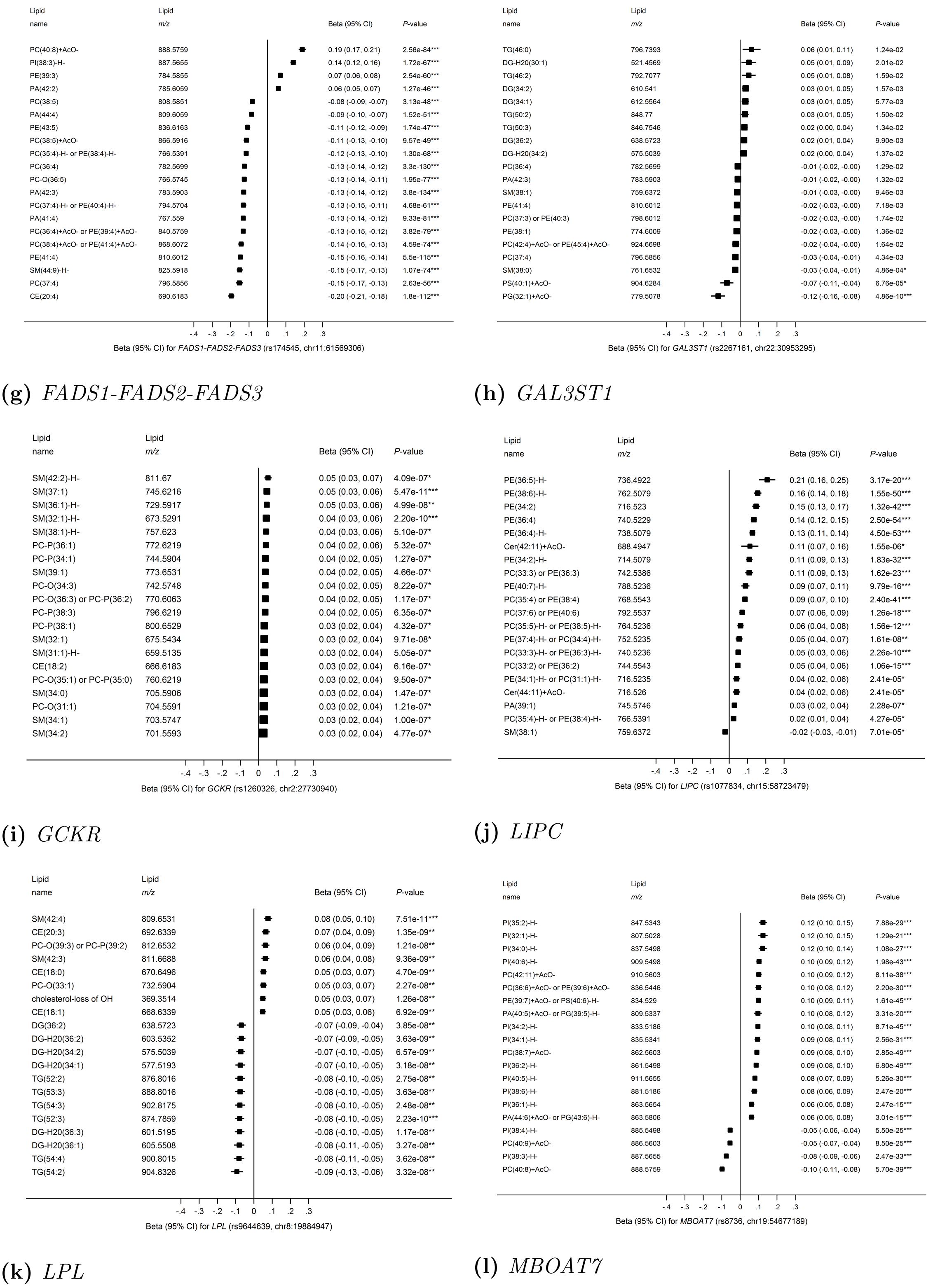
(d) *CERS4*

Lipid name	Lipid *m/z*	Beta (95% CI)	*P*-value
PC-O(34:3)	742.5748	0.04 (0.03, 0.05)	4.08e-10***
PC-P(36:1)	772.6219	0.03 (0.02, 0.05)	5.39e-06*
PC-P(34:1)	744.5904	0.03 (0.02, 0.04)	1.65e-06*
PC-O(34:1)	746.6061	0.03 (0.02, 0.04)	2.27e-06*
PA(40:1)	759.5903	0.02 (0.01, 0.02)	1.14e-06*
PC(34:2)	758.57	0.02 (0.01, 0.02)	2.22e-06*
DG-H20(36:2)	603.5352	-0.03 (-0.04, -0.02)	7.96e-06*
DG(36:2)	638.5723	-0.03 (-0.05, -0.02)	2.69e-06*
TG(56:6)	924.801	-0.03 (-0.05, -0.02)	9.69e-06*
DG-H20(36:3)	601.5195	-0.03 (-0.05, -0.02)	8.59e-06*
TG(54:4)	900.8015	-0.04 (-0.05, -0.02)	1.04e-05*
DG(34:2)	610.541	-0.04 (-0.06, -0.02)	5.53e-07*
TG(56:7)	922.7853	-0.04 (-0.06, -0.02)	5.41e-06*
DG(34:1)	612.5564	-0.04 (-0.06, -0.03)	5.17e-07*
TG(54:5)	898.7856	-0.04 (-0.06, -0.03)	1.83e-06*
TG(52:5)	870.7544	-0.04 (-0.06, -0.03)	7.26e-06*
SM(46:1)	871.7627	-0.05 (-0.06, -0.03)	6.44e-06*
TG(54:6)	896.7698	-0.05 (-0.07, -0.03)	1.45e-06*
DG(36:4)	634.5409	-0.05 (-0.07, -0.03)	1.16e-06*
TG(54:7)	894.7541	-0.06 (-0.09, -0.04)	5.02e-06*

(e) *CETP*

Lipid name	Lipid *m/z*	Beta (95% CI)	*P*-value
CE(20:5)	688.6026	0.04 (0.02, 0.06)	1.47e-04*
PC(38:5)+AcO-	866.5916	0.03 (0.01, 0.04)	1.78e-05*
PC(38:5)	808.5851	0.02 (0.01, 0.02)	2.62e-04*
PG(36:2)-H-	773.5337	-0.02 (-0.04, -0.01)	3.97e-04*
PC(37:6) or PE(40:6)	792.5537	-0.03 (-0.04, -0.01)	2.21e-04*
PC(38:6)	806.5694	-0.03 (-0.04, -0.02)	2.86e-09**
PI(38:6)-H-	881.5186	-0.03 (-0.05, -0.02)	1.12e-04*
PA(44:5)	807.5903	-0.03 (-0.04, -0.02)	3.57e-09**
PI(38:0)-H-	893.6124	-0.03 (-0.05, -0.02)	1.25e-05*
PC(38:6)+AcO-	864.5759	-0.03 (-0.05, -0.02)	2.92e-06*
PI(36:0)-H-	865.5811	-0.03 (-0.05, -0.02)	1.51e-06*
PC(40:6)+AcO- or PE(43:6)+AcO-	892.6072	-0.03 (-0.05, -0.02)	1.05e-05*
PE(43:6)	834.6006	-0.03 (-0.05, -0.02)	8.50e-08*
PC(37:6)-H- or PE(40:6)-H-	790.5391	-0.04 (-0.05, -0.02)	2.33e-07*
PE(38:6)-H-	762.5079	-0.04 (-0.05, -0.02)	7.83e-05*
PC(37:6)+AcO- or PE(40:6)+AcO-	850.5602	-0.04 (-0.05, -0.02)	6.57e-05*
PA(43:5)-H-	791.5596	-0.04 (-0.05, -0.02)	9.10e-08*
FreeFA(22:6)-H-	327.233	-0.04 (-0.06, -0.03)	1.68e-08**
PC(42:11)+AcO-	910.5603	-0.05 (-0.06, -0.03)	8.02e-09**
PI(40:6)-H-	909.5498	-0.05 (-0.06, -0.03)	1.43e-10***

(f) *ELOVL2*



Lipid name	Lipid m/z	Beta (95% CI)	P-value
PC(40:8)+AcO-	888.5759	0.19 (0.17, 0.21)	2.56e-84***
PI(38:3)-H-	887.5655	0.14 (0.12, 0.16)	1.72e-67***
PE(39:3)	784.5855	0.07 (0.06, 0.08)	2.54e-60***
PA(42:2)	785.6059	0.06 (0.05, 0.07)	1.27e-46***
PC(38:5)	808.5851	-0.08 (-0.09, -0.07)	3.13e-48***
PA(44:4)	809.6059	-0.09 (-0.10, -0.07)	1.52e-51***
PE(43:5)	836.6163	-0.11 (-0.12, -0.09)	1.74e-47***
PC(38:5)+AcO-	866.5916	-0.11 (-0.13, -0.10)	9.57e-49***
PC(35:4)-H- or PE(38:4)-H-	766.5391	-0.12 (-0.13, -0.10)	1.30e-68***
PC(36:4)	782.5699	-0.13 (-0.14, -0.12)	3.3e-130***
PC-O(36:5)	766.5745	-0.13 (-0.14, -0.11)	1.95e-77***
PA(42:3)	783.5903	-0.13 (-0.14, -0.12)	3.8e-134***
PC(37:4)-H- or PE(40:4)-H-	794.5704	-0.13 (-0.15, -0.11)	4.68e-61***
PA(41:4)	767.559	-0.13 (-0.14, -0.12)	9.33e-81***
PC(36:4)+AcO- or PE(39:4)+AcO-	840.5759	-0.13 (-0.15, -0.12)	3.82e-79***
PC(38:4)+AcO- or PE(41:4)+AcO-	868.6072	-0.14 (-0.16, -0.13)	4.59e-74***
PE(41:4)	810.6012	-0.15 (-0.16, -0.14)	5.5e-115***
SM(44:9)-H-	825.5918	-0.15 (-0.17, -0.13)	1.07e-74***
PC(37:4)	796.5856	-0.15 (-0.17, -0.13)	2.63e-56***
CE(20:4)	690.6183	-0.20 (-0.21, -0.18)	1.8e-112***

-.4 -.3 -.2 -.1 0 .1 .2 .3
Beta (95% CI) for FADS1-FADS2-FADS3 (rs174545, chr11:61569306)

(g) *FADS1-FADS2-FADS3*

Lipid name	Lipid m/z	Beta (95% CI)	P-value
TG(46:0)	796.7393	0.06 (0.01, 0.11)	1.24e-02
DG-H20(30:1)	521.4569	0.05 (0.01, 0.09)	2.01e-02
TG(46:2)	792.7077	0.05 (0.01, 0.08)	1.59e-02
DG(34:2)	610.541	0.03 (0.01, 0.05)	1.57e-03
DG(34:1)	612.5564	0.03 (0.01, 0.05)	5.77e-03
TG(50:2)	848.77	0.03 (0.01, 0.05)	1.50e-02
TG(50:3)	846.7546	0.02 (0.00, 0.04)	1.34e-02
DG(36:2)	638.5723	0.02 (0.01, 0.04)	9.90e-03
DG-H20(34:2)	575.5039	0.02 (0.00, 0.04)	1.37e-02
PC(36:4)	782.5699	-0.01 (-0.02, -0.00)	1.29e-02
PA(42:3)	783.5903	-0.01 (-0.02, -0.00)	1.32e-02
SM(38:1)	759.6372	-0.01 (-0.03, -0.00)	9.46e-03
PE(41:4)	810.6012	-0.02 (-0.03, -0.00)	7.18e-03
PC(37:3) or PE(40:3)	798.6012	-0.02 (-0.03, -0.00)	1.74e-02
PE(38:1)	774.6009	-0.02 (-0.03, -0.00)	1.36e-02
PC(42:4)+AcO- or PE(45:4)+AcO-	924.6698	-0.02 (-0.04, -0.00)	1.64e-02
PC(37:4)	796.5856	-0.03 (-0.04, -0.01)	4.34e-03
SM(38:0)	761.6532	-0.03 (-0.04, -0.01)	4.86e-04*
PS(40:1)+AcO-	904.6284	-0.07 (-0.11, -0.04)	6.76e-05*
PG(32:1)+AcO-	779.5078	-0.12 (-0.16, -0.08)	4.86e-10***

(h) *GAL3ST1*

Lipid name	Lipid m/z	Beta (95% CI)	P-value
SM(42:2)-H-	811.67	0.05 (0.03, 0.07)	4.09e-07*
SM(37:1)	745.6216	0.05 (0.03, 0.06)	5.47e-11***
SM(36:1)-H-	729.5917	0.05 (0.03, 0.06)	4.99e-08**
SM(32:1)-H-	673.5291	0.04 (0.03, 0.06)	2.20e-10***
SM(38:1)-H-	757.623	0.04 (0.03, 0.06)	5.10e-07*
PC-P(36:1)	772.6219	0.04 (0.02, 0.06)	5.32e-07*
PC-P(34:1)	744.5904	0.04 (0.02, 0.05)	1.27e-07*
SM(39:1)	773.6531	0.04 (0.02, 0.05)	4.66e-07*
PC-O(34:3)	742.5748	0.04 (0.02, 0.05)	8.22e-07*
PC-O(36:3) or PC-P(36:2)	770.6063	0.04 (0.02, 0.05)	1.17e-07*
PC-P(38:3)	796.6219	0.04 (0.02, 0.05)	6.35e-07*
PC-P(38:1)	800.6529	0.03 (0.02, 0.04)	4.32e-07*
SM(32:1)	675.5434	0.03 (0.02, 0.04)	9.71e-08*
SM(31:1)-H-	659.5135	0.03 (0.02, 0.04)	5.05e-07*
CE(18:2)	666.6183	0.03 (0.02, 0.04)	6.16e-07*
PC-O(35:1) or PC-P(35:0)	760.6219	0.03 (0.02, 0.04)	9.50e-07*
SM(34:0)	705.5906	0.03 (0.02, 0.04)	1.47e-07*
PC-O(31:1)	704.5591	0.03 (0.02, 0.04)	1.21e-07*
SM(34:1)	703.5747	0.03 (0.02, 0.04)	1.00e-07*
SM(34:2)	701.5593	0.03 (0.02, 0.04)	4.77e-07*

(i) *GCKR*

Lipid name	Lipid m/z	Beta (95% CI)	P-value
PE(36:5)-H-	736.4922	0.21 (0.16, 0.25)	3.17e-20***
PE(38:6)-H-	762.5079	0.16 (0.14, 0.18)	1.55e-50***
PE(34:2)	716.523	0.15 (0.13, 0.17)	1.32e-42***
PE(36:4)	740.5229	0.14 (0.12, 0.15)	2.50e-54***
PE(36:4)-H-	738.5079	0.13 (0.11, 0.14)	4.50e-53***
Cer(42:11)+AcO-	688.4947	0.11 (0.07, 0.16)	1.55e-06*
PE(34:2)-H-	714.5079	0.11 (0.09, 0.13)	1.83e-32***
PC(33:3) or PE(36:3)	742.5386	0.11 (0.09, 0.13)	1.62e-23***
PE(40:7)-H-	788.5236	0.09 (0.07, 0.11)	9.79e-16***
PC(35:4) or PE(38:4)	768.5543	0.09 (0.07, 0.10)	2.40e-41***
PC(37:6) or PE(40:6)	792.5537	0.07 (0.06, 0.09)	1.26e-18***
PC(35:5)-H- or PE(38:5)-H-	764.5236	0.06 (0.04, 0.08)	1.56e-12***
PE(37:4)-H- or PC(34:4)-H-	752.5235	0.05 (0.04, 0.07)	1.61e-08**
PC(33:3)-H- or PE(36:3)-H-	740.5236	0.05 (0.03, 0.06)	2.26e-10***
PC(33:2) or PE(36:2)	744.5543	0.05 (0.04, 0.06)	1.06e-15***
PE(34:1)-H- or PC(31:1)-H-	716.5235	0.04 (0.02, 0.06)	2.41e-05*
Cer(44:11)+AcO-	716.526	0.04 (0.02, 0.06)	2.41e-05*
PA(39:1)	745.5746	0.03 (0.02, 0.04)	2.28e-07*
PC(35:4)-H- or PE(38:4)-H-	766.5391	0.02 (0.01, 0.04)	4.27e-05*
SM(38:1)	759.6372	-0.02 (-0.03, -0.01)	7.01e-05*

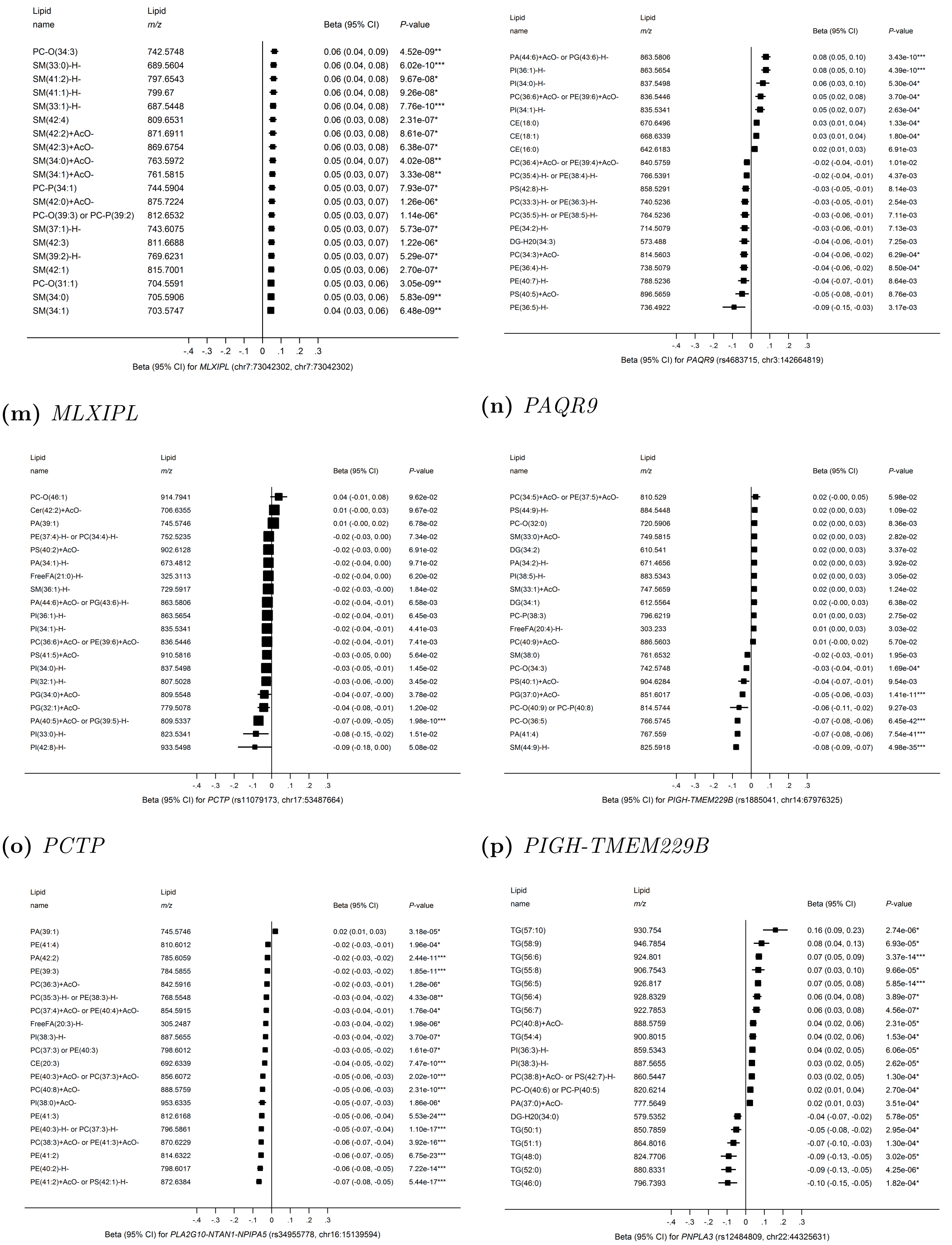
(j) *LIPC*

Lipid name	Lipid m/z	Beta (95% CI)	P-value
SM(42:4)	809.6531	0.08 (0.05, 0.10)	7.51e-11***
CE(20:3)	692.6339	0.07 (0.04, 0.09)	1.35e-09**
PC-O(39:3) or PC-P(39:2)	812.6532	0.06 (0.04, 0.09)	1.21e-08**
SM(42:3)	811.6688	0.06 (0.04, 0.08)	9.36e-09**
CE(18:0)	670.6496	0.05 (0.03, 0.07)	4.70e-09**
PC-O(33:1)	732.5904	0.05 (0.03, 0.07)	2.27e-08**
cholesterol-loss of OH	369.3514	0.05 (0.03, 0.07)	1.26e-08**
CE(18:1)	668.6339	0.05 (0.03, 0.06)	6.92e-09**
DG(36:2)	638.5723	-0.07 (-0.09, -0.04)	3.85e-08**
DG-H20(36:2)	603.5352	-0.07 (-0.09, -0.05)	3.63e-09**
DG-H20(34:2)	575.5039	-0.07 (-0.10, -0.05)	6.57e-09**
DG-H20(34:1)	577.5193	-0.07 (-0.10, -0.05)	3.18e-08**
TG(52:2)	876.8016	-0.08 (-0.10, -0.05)	2.75e-08**
TG(53:3)	888.8016	-0.08 (-0.10, -0.05)	3.63e-08**
TG(54:3)	902.8175	-0.08 (-0.10, -0.05)	2.48e-08**
TG(52:3)	874.7859	-0.08 (-0.10, -0.05)	2.23e-10***
DG-H20(36:3)	601.5195	-0.08 (-0.10, -0.05)	1.17e-08**
DG-H20(36:1)	605.5508	-0.08 (-0.11, -0.05)	3.27e-08**
TG(54:4)	900.8015	-0.08 (-0.11, -0.05)	3.62e-08**
TG(54:2)	904.8326	-0.09 (-0.13, -0.06)	3.32e-08**

(k) *LPL*

Lipid name	Lipid m/z	Beta (95% CI)	P-value
PI(35:2)-H-	847.5343	0.12 (0.10, 0.15)	7.88e-29***
PI(32:1)-H-	807.5028	0.12 (0.10, 0.15)	1.29e-21***
PI(34:0)-H-	837.5498	0.12 (0.10, 0.14)	1.08e-27***
PI(40:6)-H-	909.5498	0.10 (0.09, 0.12)	1.98e-43***
PC(42:11)+AcO-	910.5603	0.10 (0.09, 0.12)	8.11e-38***
PC(36:6)+AcO- or PE(39:6)+AcO-	836.5446	0.10 (0.08, 0.12)	2.20e-30***
PE(39:7)+AcO- or PS(40:6)-H-	834.529	0.10 (0.09, 0.11)	1.61e-45***
PA(40:5)+AcO- or PG(39:5)-H-	809.5337	0.10 (0.08, 0.12)	3.31e-20***
PI(34:2)-H-	833.5186	0.10 (0.08, 0.11)	8.71e-45***
PI(34:1)-H-	835.5341	0.09 (0.08, 0.11)	2.56e-31***
PC(38:7)+AcO-	862.5603	0.09 (0.08, 0.10)	2.85e-49***
PI(36:2)-H-	861.5498	0.09 (0.08, 0.10)	6.80e-49***
PI(40:5)-H-	911.5655	0.08 (0.07, 0.09)	5.26e-30***
PI(38:6)-H-	881.5186	0.08 (0.06, 0.09)	2.47e-20***
PI(36:1)-H-	863.5654	0.06 (0.05, 0.08)	2.47e-15***
PA(44:6)+AcO- or PG(43:6)-H-	863.5806	0.06 (0.05, 0.08)	3.01e-15***
PI(38:4)-H-	885.5498	-0.05 (-0.06, -0.04)	5.50e-25***
PC(40:9)+AcO-	886.5603	-0.05 (-0.07, -0.04)	8.50e-25***
PI(38:3)-H-	887.5655	-0.08 (-0.09, -0.06)	2.47e-33***
PC(40:8)+AcO-	888.5759	-0.10 (-0.11, -0.08)	5.70e-39***

(l) *MBOAT7*



(m) *MLXIPL*

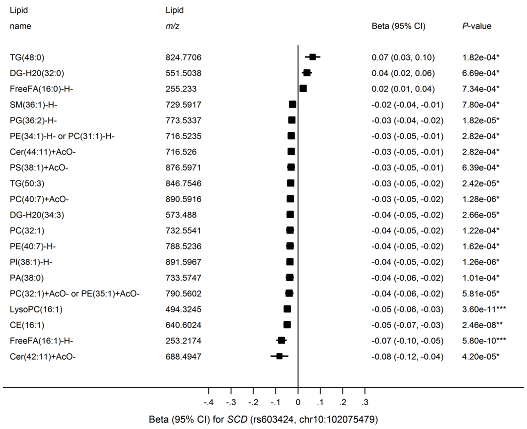
(n) *PAQR9*

(o) *PCTP*

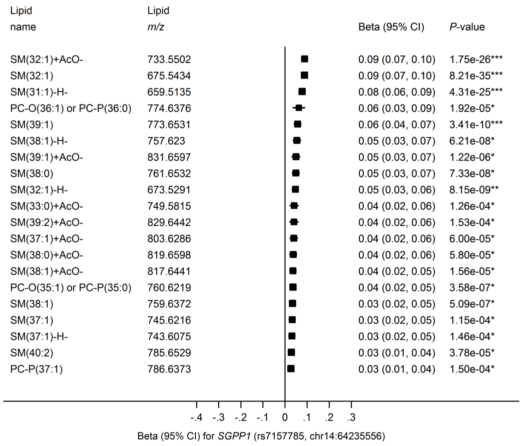
(p) *PIGH-TMEM229B*

(q) *PLA2G10-NTAN1-NPIPA5*

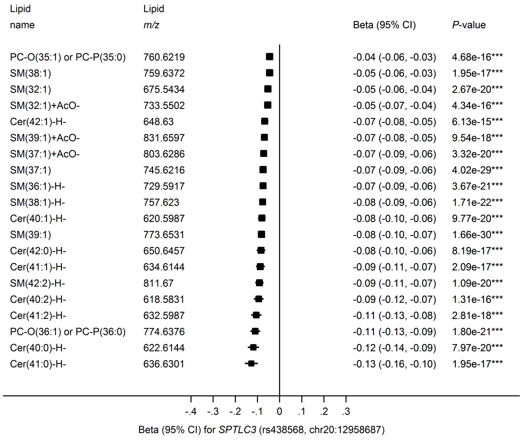
(r) *PNPLA3*



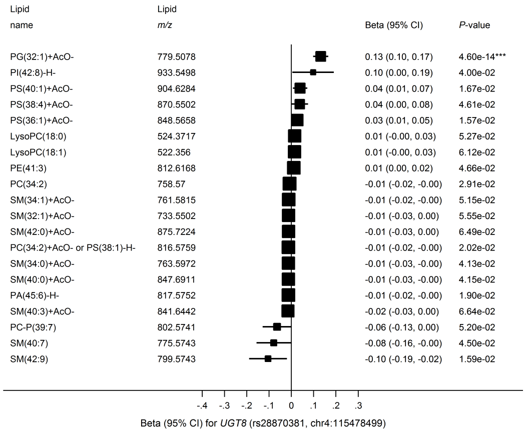
(s) *SCD*



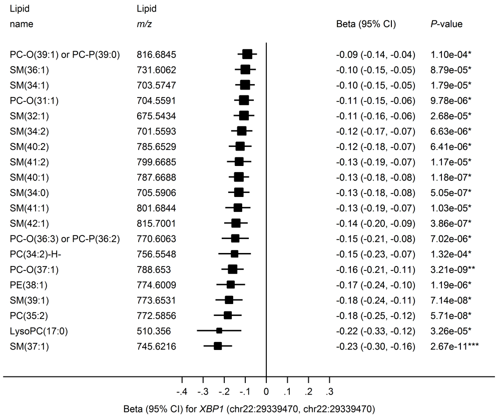
(t) *SGPP1*



(u) *SPTLC3*



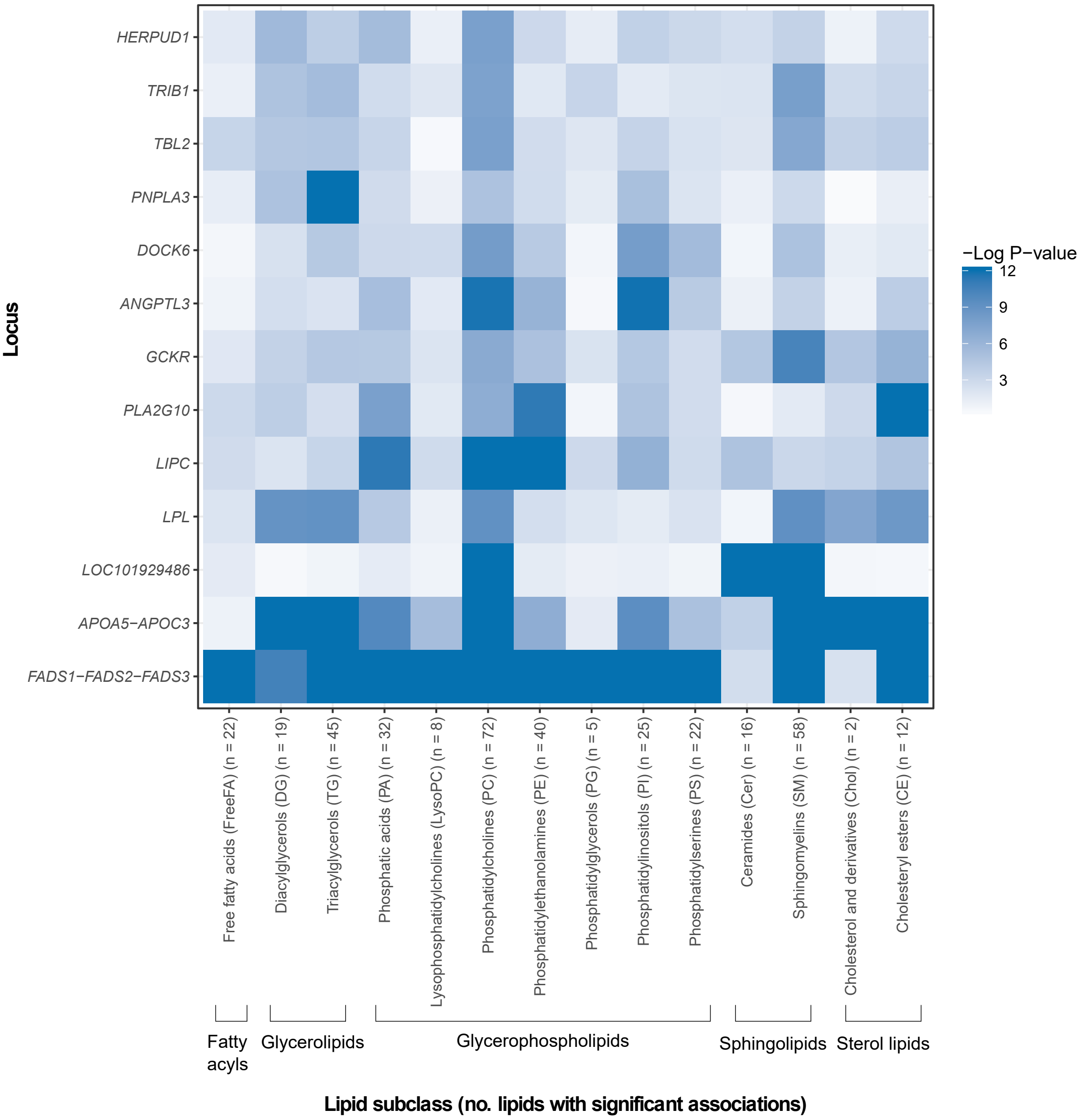
(v) *UGT8*



(w) *XBP1*

Note: * = $P < 0.001$; ** = $P < 5 \times 10^{-8}$; *** = $P < 8.9 \times 10^{-10}$.

Figure 5.10: Heat map showing associations between lipid subclasses and significant loci from the 175 major lipid loci



The effect estimates of the associations between lipid subclasses and significant variants from the 175 major lipid loci are plotted as a heat map. Results are shown for the association of the most strongly associated (smallest P-value) lipid within each lipid subclass with each locus. The rows show the subset of the 175 loci with at least one significantly associated variant while the columns show the 14 lipid subclasses. The magnitude of the P-values of association are indicated by the colour scale from light to dark blue, where genome-wide significant associations ($P < 8.9 \times 10^{-10}$) are shown in the darkest colour. The P-values are capped at a maximum of $P < 1 \times 10^{-12}$ so as to make the differences between significant and non-significant associations more readily apparent. The lipid subclasses are arranged according to their classification by overall category, main class, and then alphabetically by subclass. The loci are arranged in order of the number of lipids significantly associated with each locus, with ties ordered alphabetically.

5.3.5 Results of GWAS of principal components

Following on from the PCA of the lipid metabolites described in Chapter 4 and the association of these principal components with CHD risk factors, a GWAS was conducted on the second, third, and fourth principal components. The Manhattan plots summarising the genetic associations for these principal components are shown in Figure 5.11, and the regional association plots for the variants that reached genome-wide significance are shown in Figure 5.12. Table 5.3 presents a summary of the associations for the variants that reached genome-wide significance.

There was only one variant (rs662799, chr11:116663707) in the *APOA5-APOC3* genetic locus that showed a genome-wide significant association with the second principal component (Figure 5.11a and Figure 5.12a), but there were 74 variants that were associated with the third and fourth principal components (Table 5.3).

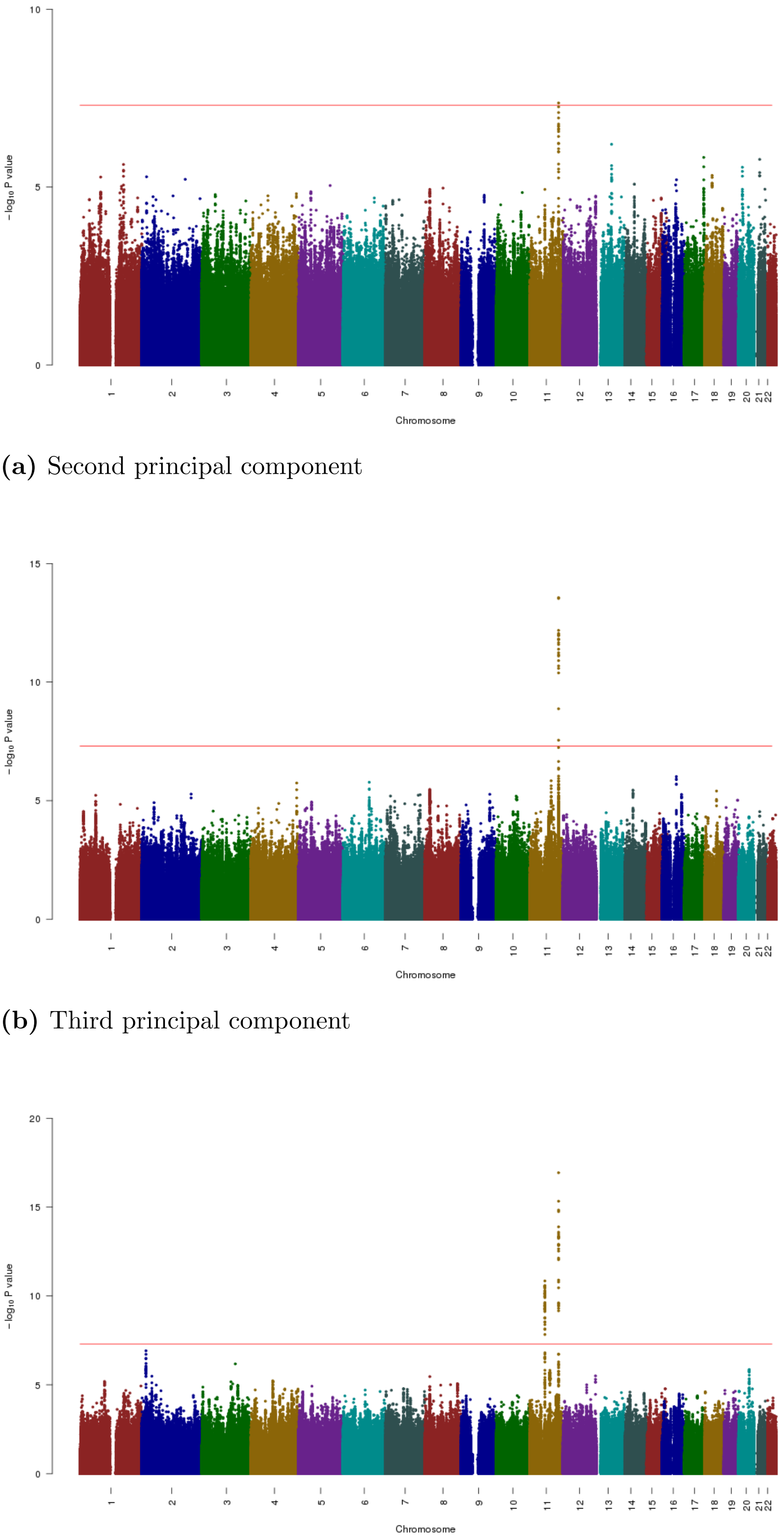
Although the third principal component was most closely characterised by unsaturated triglycerides, it did not show any significant associations with variants in the *FADS1-2-3* locus, and was only associated with variants in the *APOA5-APOC3* region (Figure 5.11b and Figure 5.12b). In contrast, variants in both the *FADS1-2-3* and *APOA5-APOC3* regions were significantly associated with the fourth principal component (Figure 5.11c, Figure 5.12c and Figure 5.12d).

Conditional analyses were not performed on the GWAS of the principal components because the sentinel variants in the *FADS1-2-3* and *APOA5-APOC3* regions had already been identified from the conditional analyses of the univariate GWAS. Therefore, conditional analyses of the principal components would not have yielded much additional information.

5.3.6 Results of GWAS of ratios of lipid metabolites

Twenty-six ratios of lipids were identified and selected that had strong biological rationales and acted through thoroughly understood metabolic pathways. Seventeen of the ratios had one or more genome-wide significant associations, for which the Manhattan plots are shown in Figure 5.13. A summary of the associations for the most strongly associated variant within each locus that was associated with each ratio is shown in Table 5.4. The analysis of these ratios resulted in the identification of four additional independent loci that were not detected in the univariate GWAS. First, the ratio LysoPC(18:2) / PC(34:2) (m/z 520.3404 / 758.57), which is indicative of lipase activity, was significantly associated with variants in the *MYCL1-MFSD2A* locus (Figure 5.13i). Major facilitator

Figure 5.11: Manhattan plots of principal components of lipid levels

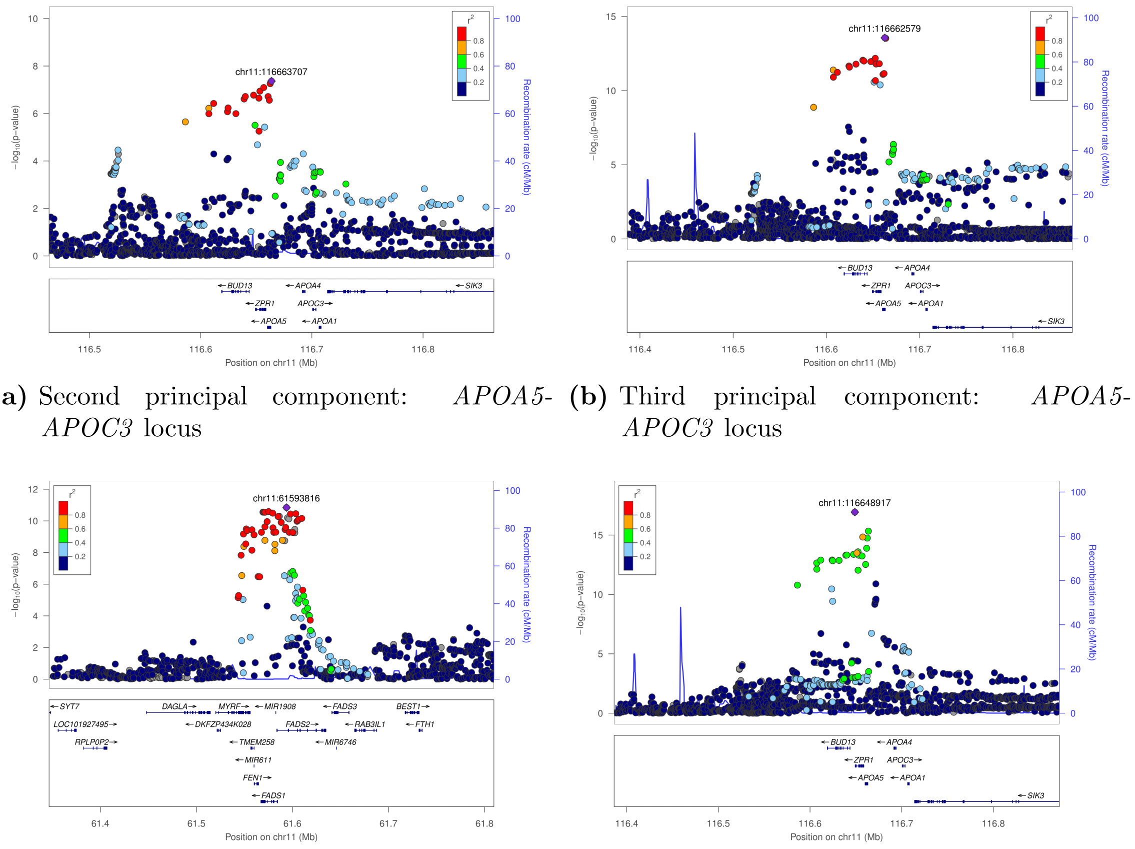


(a) Second principal component

(b) Third principal component

(c) Fourth principal component

Figure 5.12: Regional association plots of principal components of lipid levels



(a) Second principal component: *APOA5-APOC3* locus

(b) Third principal component: *APOA5-APOC3* locus

(c) Fourth principal component: *FADS1-2-3* locus

(d) Fourth principal component: *APOA5-APOC3* locus

Table 5.3: Summary of genetic associations of principal components of lipids

Principal component	rsID	Chr:Pos (GRCh37)	EA	NEA	β	SE	P-value	Locus
Second	rs662799	chr11:116663707	A	G	−0.9963	0.1819	4.3×10^{-8}	*APOA5-APOC3*
Third	rs651821	chr11:116662579	T	C	1.2532	0.1647	2.7×10^{-14}	*APOA5-APOC3*
Third	rs662799	chr11:116663707	A	G	1.2561	0.1653	3.0×10^{-14}	*APOA5-APOC3*
Third	rs6589566	chr11:116652423	A	G	1.1600	0.1614	6.6×10^{-13}	*APOA5-APOC3*
Third	rs10790162	chr11:116639104	A	G	−1.1522	0.1612	8.9×10^{-13}	*APOA5-APOC3*
Third	rs6589565	chr11:116640237	A	G	−1.1488	0.1611	1.0×10^{-12}	*APOA5-APOC3*
Third	rs964184	chr11:116648917	C	G	1.0656	0.1496	1.1×10^{-12}	*APOA5-APOC3*
Third	rs2160669	chr11:116647607	T	C	1.1438	0.1607	1.1×10^{-12}	*APOA5-APOC3*
Third	rs10750096	chr11:116656788	A	C	1.1443	0.1618	1.5×10^{-12}	*APOA5-APOC3*
Third	rs2075290	chr11:116653296	T	C	1.1359	0.1608	1.6×10^{-12}	*APOA5-APOC3*
Third	rs3825041	chr11:116631707	T	C	−1.1263	0.1595	1.6×10^{-12}	*APOA5-APOC3*
Third	rs6589564	chr11:116624153	C	G	−1.1207	0.1596	2.2×10^{-12}	*APOA5-APOC3*
Third	rs7930786	chr11:116624727	C	G	−1.1135	0.1591	2.6×10^{-12}	*APOA5-APOC3*
Third	rs1558860	chr11:116607368	A	C	−1.1160	0.1610	4.1×10^{-12}	*APOA5-APOC3*
Third	rs9326246	chr11:116611733	C	G	−1.1046	0.1604	5.8×10^{-12}	*APOA5-APOC3*
Third	rs2072560	chr11:116661826	T	C	−1.1511	0.1679	7.1×10^{-12}	*APOA5-APOC3*
Third	rs2266788	chr11:116660686	A	G	1.0834	0.1584	8.0×10^{-12}	*APOA5-APOC3*
Third	rs1558861	chr11:116607437	T	C	1.0867	0.1604	1.3×10^{-11}	*APOA5-APOC3*
Third	rs7483863	chr11:116652491	A	G	−1.0649	0.1589	2.1×10^{-11}	*APOA5-APOC3*
Third	rs11604424	chr11:116651115	T	C	0.8792	0.1319	2.6×10^{-11}	*APOA5-APOC3*
Third	rs3741298	chr11:116657561	T	C	0.8822	0.1337	4.1×10^{-11}	*APOA5-APOC3*
Third	rs7350481	chr11:116586283	T	C	−0.8967	0.1479	1.3×10^{-9}	*APOA5-APOC3*
Third	rs180327	chr11:116623659	T	C	0.7048	0.1269	2.8×10^{-8}	*APOA5-APOC3*
Fourth	rs964184	chr11:116648917	C	G	1.2218	0.1428	1.2×10^{-17}	*APOA5-APOC3*
Fourth	rs662799	chr11:116663707	A	G	1.2808	0.1577	4.7×10^{-16}	*APOA5-APOC3*
Fourth	rs3741298	chr11:116657561	T	C	1.0166	0.1274	1.5×10^{-15}	*APOA5-APOC3*
Fourth	rs651821	chr11:116662579	T	C	1.2499	0.1571	1.8×10^{-15}	*APOA5-APOC3*
Fourth	rs2072560	chr11:116661826	T	C	−1.2342	0.1602	1.3×10^{-14}	*APOA5-APOC3*
Fourth	rs6589566	chr11:116652423	A	G	1.1734	0.1540	2.6×10^{-14}	*APOA5-APOC3*
Fourth	rs11604424	chr11:116651115	T	C	0.9542	0.1258	3.3×10^{-14}	*APOA5-APOC3*
Fourth	rs2160669	chr11:116647607	T	C	1.1589	0.1534	4.2×10^{-14}	*APOA5-APOC3*
Fourth	rs6589565	chr11:116640237	A	G	−1.1592	0.1538	4.8×10^{-14}	*APOA5-APOC3*
Fourth	rs2075290	chr11:116653296	T	C	1.1567	0.1536	5.0×10^{-14}	*APOA5-APOC3*
Fourth	rs10790162	chr11:116639104	A	G	−1.1591	0.1539	5.0×10^{-14}	*APOA5-APOC3*
Fourth	rs10750096	chr11:116656788	A	C	1.1604	0.1544	5.8×10^{-14}	*APOA5-APOC3*
Fourth	rs6589564	chr11:116624153	C	G	−1.1287	0.1523	1.3×10^{-13}	*APOA5-APOC3*
Fourth	rs9326246	chr11:116611733	C	G	−1.1344	0.1531	1.3×10^{-13}	*APOA5-APOC3*
Fourth	rs3825041	chr11:116631707	T	C	−1.1264	0.1522	1.4×10^{-13}	*APOA5-APOC3*
Fourth	rs7930786	chr11:116624727	C	G	−1.1225	0.1519	1.5×10^{-13}	*APOA5-APOC3*
Fourth	rs1558860	chr11:116607368	A	C	−1.1264	0.1536	2.3×10^{-13}	*APOA5-APOC3*
Fourth	rs2266788	chr11:116660686	A	G	1.1044	0.1514	3.0×10^{-13}	*APOA5-APOC3*
Fourth	rs1558861	chr11:116607437	T	C	1.0976	0.1531	7.5×10^{-13}	*APOA5-APOC3*
Fourth	rs7483863	chr11:116652491	A	G	−1.0820	0.1515	9.1×10^{-13}	*APOA5-APOC3*
Fourth	rs71462009	chr11:116671824	T	C	−0.9782	0.1445	1.3×10^{-11}	*APOA5-APOC3*
Fourth	rs7350481	chr11:116586283	T	C	−0.9501	0.1411	1.6×10^{-11}	*APOA5-APOC3*
Fourth	rs180327	chr11:116623659	T	C	0.8018	0.1210	3.5×10^{-11}	*APOA5-APOC3*
Fourth	rs11216140	chr11:116672013	T	C	−0.9199	0.1453	2.5×10^{-10}	*APOA5-APOC3*
Fourth	rs6589569	chr11:116671476	T	C	−0.9135	0.1447	2.7×10^{-10}	*APOA5-APOC3*
Fourth	rs9667814	chr11:116671823	C	G	−0.9108	0.1449	3.3×10^{-10}	*APOA5-APOC3*
Fourth	rs180326	chr11:116624703	T	G	0.7828	0.1250	3.8×10^{-10}	*APOA5-APOC3*
Fourth	rs4938313	chr11:116671005	A	G	−0.8937	0.1437	5.0×10^{-10}	*APOA5-APOC3*
Fourth	rs6589567	chr11:116670676	A	C	−0.8835	0.1432	6.9×10^{-10}	*APOA5-APOC3*
Fourth	rs174568	chr11:61593816	T	C	−1.0686	0.1582	1.4×10^{-11}	*FADS1-2-3*
Fourth	rs174553	chr11:61575158	A	G	1.0427	0.1564	2.6×10^{-11}	*FADS1-2-3*
Fourth	rs174545	chr11:61569306	C	G	1.0406	0.1563	2.8×10^{-11}	*FADS1-2-3*
Fourth	rs174546	chr11:61569830	T	C	−1.0395	0.1562	2.8×10^{-11}	*FADS1-2-3*
Fourth	rs174547	chr11:61570783	T	C	1.0395	0.1562	2.8×10^{-11}	*FADS1-2-3*
Fourth	rs174550	chr11:61571478	T	C	1.0395	0.1562	2.8×10^{-11}	*FADS1-2-3*
Fourth	rs174554	chr11:61579463	A	G	1.0382	0.1565	3.3×10^{-11}	*FADS1-2-3*
Fourth	rs174576	chr11:61603510	A	C	−1.0250	0.1548	3.5×10^{-11}	*FADS1-2-3*
Fourth	rs1535	chr11:61597972	A	G	1.0334	0.1562	3.7×10^{-11}	*FADS1-2-3*
Fourth	rs35473591	chr11:61586328	CT	C	−1.0295	0.1568	5.1×10^{-11}	*FADS1-2-3*
Fourth	rs174562	chr11:61585144	A	G	1.0283	0.1567	5.3×10^{-11}	*FADS1-2-3*
Fourth	chr11:61594920	chr11:61594920	CT	C	1.0243	0.1568	6.4×10^{-11}	*FADS1-2-3*
Fourth	rs174583	chr11:61609750	T	C	−0.9909	0.1520	7.0×10^{-11}	*FADS1-2-3*
Fourth	chr11:61596322	chr11:61596322	CA	C	1.0201	0.1567	7.5×10^{-11}	*FADS1-2-3*

Table: Summary of genetic associations of principal components of lipids (...continued)

Principal component	rsID	Chr:Pos (GRCh37)	EA	NEA	β	SE	P-value	Locus
Fourth	rs174580	chr11:61606642	A	G	1.0100	0.1553	7.9×10^{-11}	*FADS1-2-3*
Fourth	rs174581	chr11:61606683	A	G	-1.0102	0.1553	7.9×10^{-11}	*FADS1-2-3*
Fourth	rs174578	chr11:61605499	A	T	-1.0045	0.1552	9.6×10^{-11}	*FADS1-2-3*
Fourth	rs174577	chr11:61604814	A	C	-1.0013	0.1550	1.1×10^{-10}	*FADS1-2-3*
Fourth	rs174551	chr11:61573684	T	C	1.0622	0.1646	1.1×10^{-10}	*FADS1-2-3*
Fourth	rs174564	chr11:61588305	A	G	1.0204	0.1586	1.3×10^{-10}	*FADS1-2-3*
Fourth	rs174549	chr11:61571382	A	G	-1.0376	0.1637	2.4×10^{-10}	*FADS1-2-3*
Fourth	rs174567	chr11:61593005	A	G	0.9949	0.1574	2.6×10^{-10}	*FADS1-2-3*
Fourth	rs174555	chr11:61579760	T	C	1.0359	0.1641	2.8×10^{-10}	*FADS1-2-3*
Fourth	rs174536	chr11:61551927	A	C	0.9513	0.1513	3.3×10^{-10}	*FADS1-2-3*
Fourth	chr11:61602460	chr11:61602460	CA	C	1.0333	0.1647	3.5×10^{-10}	*FADS1-2-3*
Fourth	rs174566	chr11:61592362	A	G	0.9810	0.1564	3.5×10^{-10}	*FADS1-2-3*
Fourth	rs102274	chr11:61557826	T	C	0.9506	0.1517	3.7×10^{-10}	*FADS1-2-3*
Fourth	rs174537	chr11:61552680	T	G	-0.9432	0.1509	4.1×10^{-10}	*FADS1-2-3*
Fourth	rs174561	chr11:61582708	T	C	1.0210	0.1643	5.2×10^{-10}	*FADS1-2-3*
Fourth	rs174544	chr11:61567753	A	C	-1.0301	0.1658	5.3×10^{-10}	*FADS1-2-3*
Fourth	rs99780	chr11:61596633	T	C	-0.9723	0.1565	5.3×10^{-10}	*FADS1-2-3*
Fourth	rs174556	chr11:61580635	T	C	-1.0183	0.1640	5.3×10^{-10}	*FADS1-2-3*
Fourth	rs174574	chr11:61600342	A	C	-0.9653	0.1555	5.4×10^{-10}	*FADS1-2-3*
Fourth	chr11:61602459	chr11:61602459	CCA	C	1.0396	0.1676	5.6×10^{-10}	*FADS1-2-3*
Fourth	rs174533	chr11:61549025	A	G	-0.9309	0.1508	6.8×10^{-10}	*FADS1-2-3*
Fourth	rs174538	chr11:61560081	A	G	-0.9934	0.1615	7.6×10^{-10}	*FADS1-2-3*
Fourth	rs28456	chr11:61589481	A	G	1.0099	0.1676	1.7×10^{-9}	*FADS1-2-3*
Fourth	rs174548	chr11:61571348	C	G	0.9823	0.1631	1.7×10^{-9}	*FADS1-2-3*
Fourth	chr11:61591995	chr11:61591995	G	GAA	-0.9896	0.1645	1.8×10^{-9}	*FADS1-2-3*
Fourth	rs174535	chr11:61551356	T	C	0.8947	0.1507	2.9×10^{-9}	*FADS1-2-3*
Fourth	rs174560	chr11:61581764	T	C	0.9703	0.1636	3.0×10^{-9}	*FADS1-2-3*
Fourth	rs174534	chr11:61549458	A	G	0.9292	0.1581	4.1×10^{-9}	*FADS1-2-3*
Fourth	rs102275	chr11:61557803	T	C	0.8650	0.1495	7.1×10^{-9}	*FADS1-2-3*
Fourth	rs174559	chr11:61581656	A	G	-0.9890	0.1712	7.7×10^{-9}	*FADS1-2-3*
Fourth	rs174530	chr11:61546592	A	G	0.8397	0.1483	1.5×10^{-8}	*FADS1-2-3*

The association of the second, third, and fourth principal components of lipids with genetic variants is shown in order of P-value of significance within each principal component and genetic locus. The sentinel variant (SNP most strongly associated with each principal component) within each locus is highlighted. **Abbreviations: EA** = Effect allele; **GRCh37** = Genome Reference Consortium human genome build 37; **NEA** = Non-effect allele; **SE** = Standard error; **SNP** = Single nucleotide polymorphism.

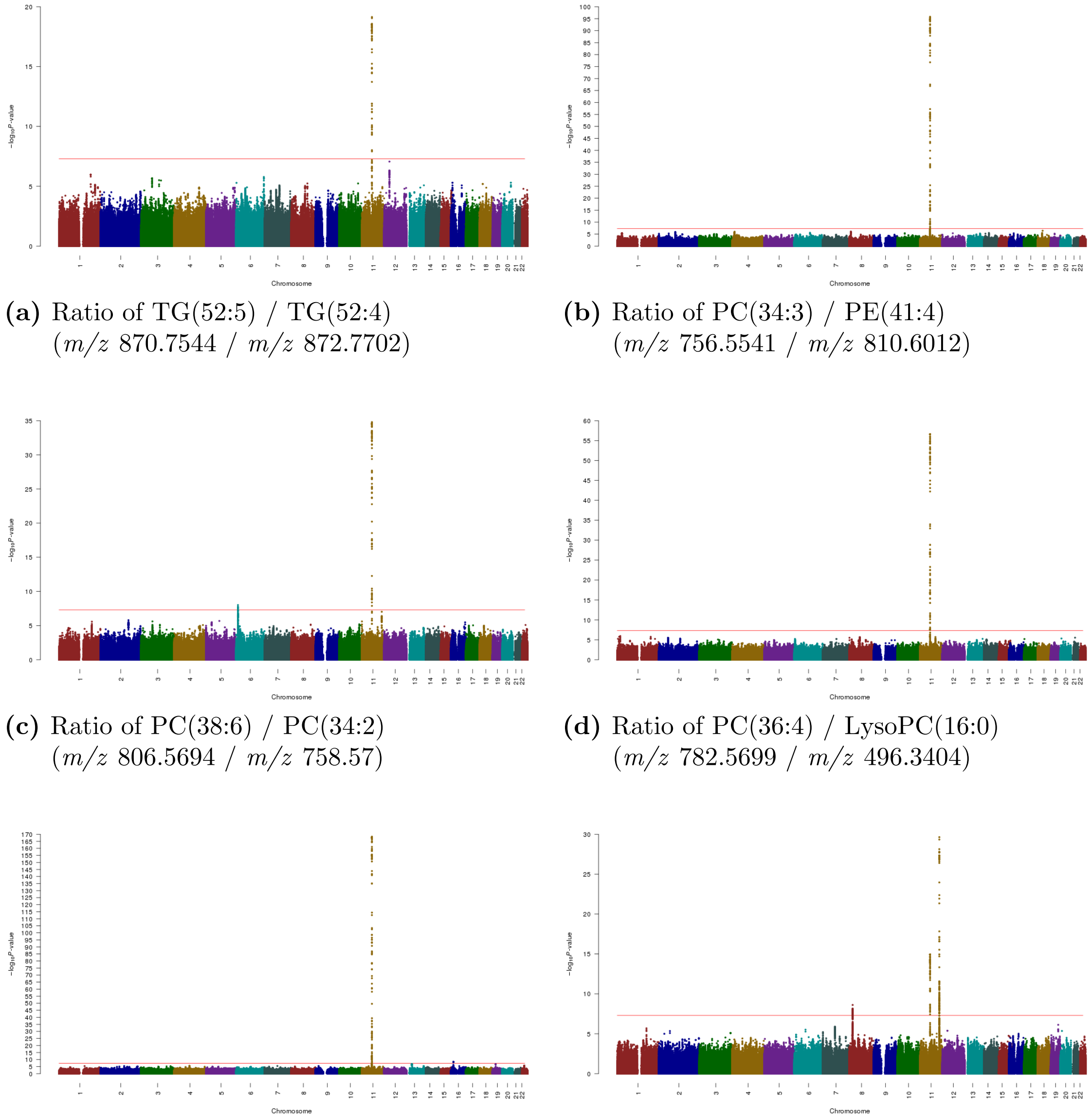
superfamily domain-containing 2A (*MFSD2A*) is a transmembrane protein and lysophosphatidylcholine transporter, so the association with a ratio containing LysoPC(18:2) is not surprising. The significant variants in this region are known to be associated with HDL cholesterol[205] and several lipid metabolites (1-oleoylglycerophosphoethanolamine, ratio of 1-eicosatrienoylglycerophosphocholine / 2-oleoylglycerophosphocholine, and ratio of phenyllactate / pyroglutamylglycine)[57].

Second, the ratio PE(39:2) / PC(34:2) (m/z 786.6012 / 758.57), which is indicative of an elongase enzyme, was significantly associated with variants in the *LPGAT1* locus (Figure 5.13m). Lysophosphatidylglycerol acyltransferase 1 (*LPGAT1*) encodes a member of the lysophospholipid acyltransferase family and catalyses the reacylation of lysophosphatidylglycerols to phosphatidylglycerols. The significant variants in this locus are known to be associated with stearic acid[206] and other metabolites (ratio of arabinose / propionylcarnitine and ratio of catechol sulfate / hippurate)[57].

Third, the ratio TG(48:1) / TG(54:5) (*m/z* 822.7546 / 898.7856), which is indicative of *de novo* lipogenesis, was significantly associated with rs71487172 (chr10:86891371) in the *LOC100507470* locus (Figure 5.13p). This variant has been reported for association with two metabolites (erthronate and glutamate)[132], but has not previously been reported for association with triglycerides.

Finally, the ratio TG(54:3) / TG(52:3) (*m/z* 902.8175 / 874.7859), which is also indicative of an elongase enzyme, was significantly associated with variants in the *HAPLN4-TM6SF2-PBX4* region (Figure 5.13l). Transmembrane 6 superfamily member 2 (*TM6SF2*) is a regulator of liver fat metabolism influencing triglyceride secretion, so the association with a ratio of two triglycerides is expected. Additionally, the significant variants in this region have been previously reported for association with several major circulating lipids, specifically total cholesterol, LDL cholesterol, and triglycerides[35].

Figure 5.13: Manhattan plots of ratios of lipid metabolites that had significant associations with one or more variants



(a) Ratio of TG(52:5) / TG(52:4) (m/z 870.7544 / m/z 872.7702)

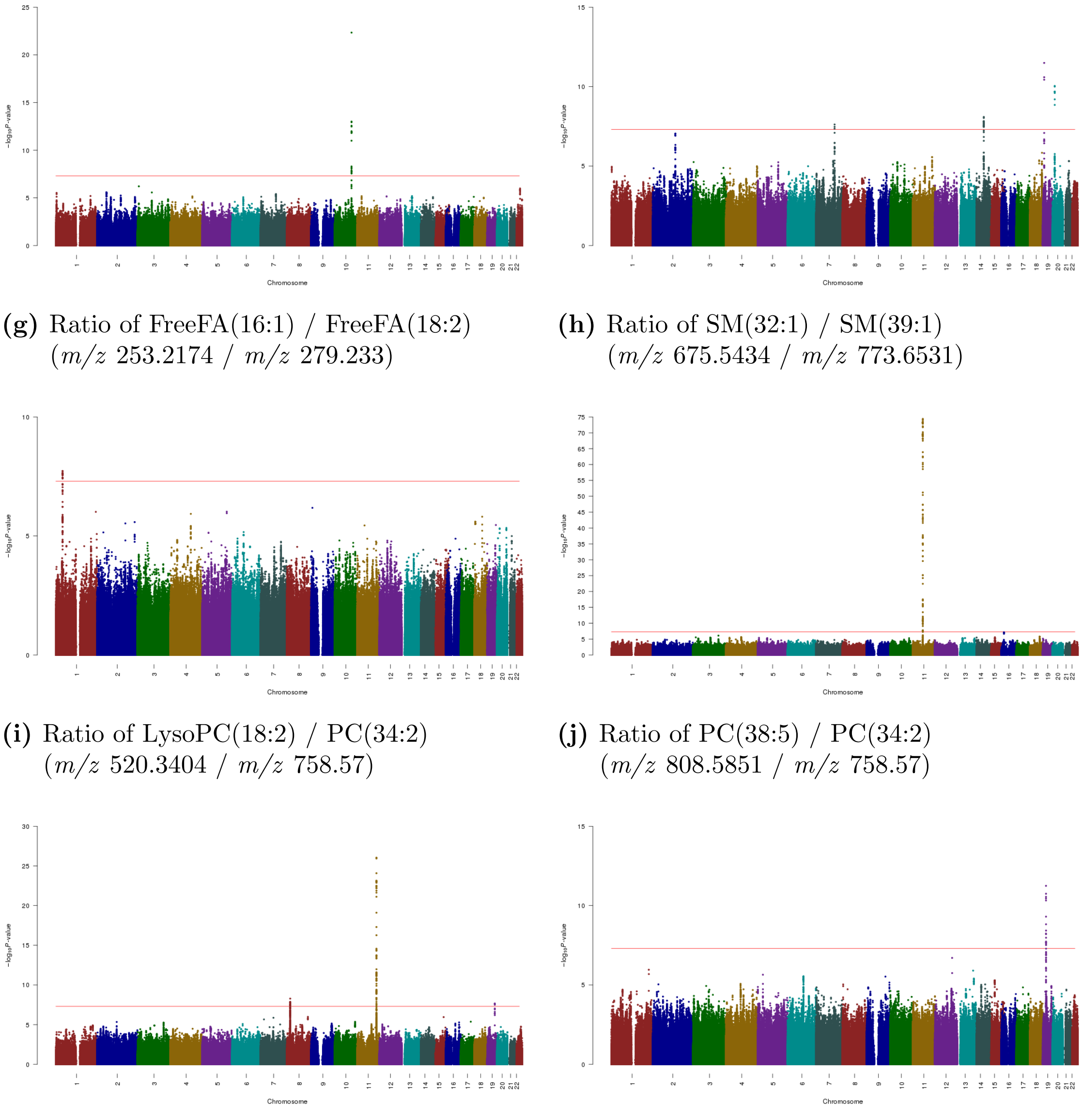
(b) Ratio of PC(34:3) / PE(41:4) (m/z 756.5541 / m/z 810.6012)

(c) Ratio of PC(38:6) / PC(34:2) (m/z 806.5694 / m/z 758.57)

(d) Ratio of PC(36:4) / LysoPC(16:0) (m/z 782.5699 / m/z 496.3404)

(e) Ratio of PC(36:4) / PC(34:2) (m/z 782.5699 / m/z 758.57)

(f) Ratio of PC(32:0) / TG(52:3) (m/z 734.5699 / m/z 874.7859)



(g) Ratio of FreeFA(16:1) / FreeFA(18:2) (m/z 253.2174 / m/z 279.233)

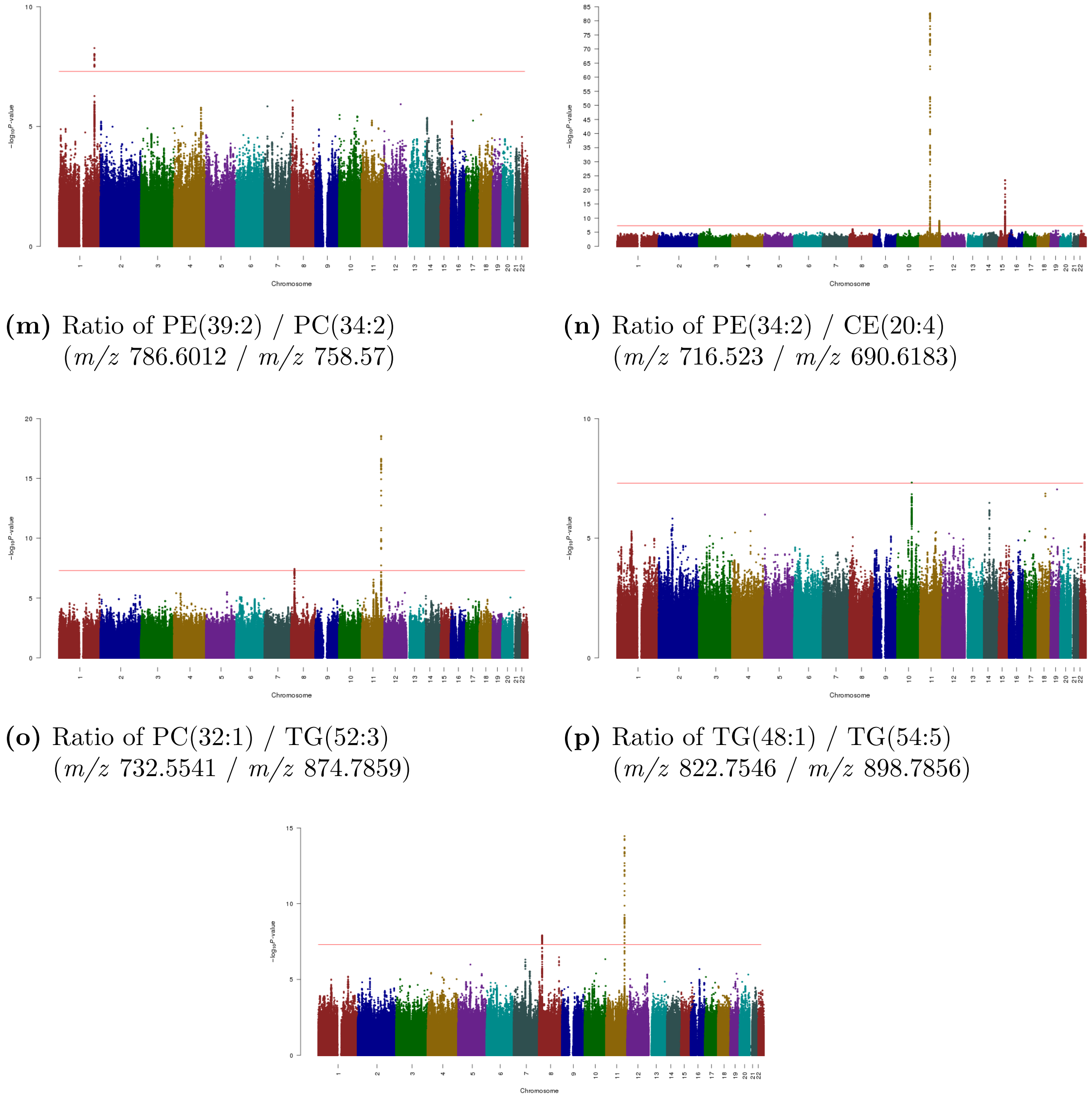
(h) Ratio of SM(32:1) / SM(39:1) (m/z 675.5434 / m/z 773.6531)

(i) Ratio of LysoPC(18:2) / PC(34:2) (m/z 520.3404 / m/z 758.57)

(j) Ratio of PC(38:5) / PC(34:2) (m/z 808.5851 / m/z 758.57)

(k) Ratio of TG(54:5) / CE(18:2) (m/z 898.7856 / m/z 666.6183)

(l) Ratio of TG(54:3) / TG(52:3) (m/z 902.8175 / m/z 874.7859)



(m) Ratio of PE(39:2) / PC(34:2) (m/z 786.6012 / m/z 758.57)

(n) Ratio of PE(34:2) / CE(20:4) (m/z 716.523 / m/z 690.6183)

(o) Ratio of PC(32:1) / TG(52:3) (m/z 732.5541 / m/z 874.7859)

(p) Ratio of TG(48:1) / TG(54:5) (m/z 822.7546 / m/z 898.7856)

(q) Ratio of SM(42:1) / TG(48:2) (m/z 815.7001 / m/z 820.739)

Table 5.4: Summary of genetic associations of ratios of lipid metabolites

Lipid ratio			**rsID**	**Chr:Pos (GRCh37)**	**EA / NEA**	$\boldsymbol{\beta}$	**SE**	***P*-value**	**Locus**
Name	***m/z***	**Classification**							
FreeFA(16:1) / FreeFA(18:2)	253.2174 / 279.233	Adipose tissue activity	rs603424	chr10:102075479	A / G	−0.1653	0.0167	4.65×10^{-23}	*PKD2L1*
LysoPC(18:2) / PC(34:2)	520.3404 / 758.57	Lipase activity	rs4381172	chr1:40400397	T / C	0.1194	0.0212	1.86×10^{-8}	*MYCL1-MFSD2A*
SM(32:1) / SM(39:1)	675.5434 / 773.6531	Dairy fat intake	rs62126382	chr19:8272916	T / C	−0.1354	0.0194	3.23×10^{-12}	*CERS4*
SM(32:1) / SM(39:1)	675.5434 / 773.6531	Dairy fat intake	rs12532251	chr7:111821556	T / G	−0.1744	0.0313	2.42×10^{-8}	*DOCK4*
SM(32:1) / SM(39:1)	675.5434 / 773.6531	Dairy fat intake	rs12880341	chr14:64236191	T / C	−0.1567	0.0272	8.12×10^{-9}	*SGPP1*
SM(32:1) / SM(39:1)	675.5434 / 773.6531	Dairy fat intake	rs364585	chr20:12962718	A / G	−0.1299	0.0200	9.07×10^{-11}	*SPTLC3*
PE(34:2) / CE(20:4)	716.523 / 690.6183	Glucose control	rs662799	chr11:116663707	A / G	−0.1521	0.0248	8.63×10^{-10}	*APOA5-APOC3*
PE(34:2) / CE(20:4)	716.523 / 690.6183	Glucose control	rs1535	chr11:61597972	A / G	−0.4726	0.0244	2.19×10^{-83}	*FADS1-2-3*
PE(34:2) / CE(20:4)	716.523 / 690.6183	Glucose control	rs1077835	chr15:58723426	A / G	−0.2258	0.0222	3.29×10^{-24}	*LIPC*
PC(32:1) / TG(52:3)	732.5541 / 874.7859	Insulin production	rs964184	chr11:116648917	C / G	0.1986	0.0221	2.84×10^{-19}	*APOA5-APOC3*
PC(32:1) / TG(52:3)	732.5541 / 874.7859	Insulin production	rs9644639	chr8:19884947	C / G	−0.1904	0.0346	3.72×10^{-8}	*LPL*
PC(32:0) / TG(52:3)	734.5699 / 874.7859	Insulin production	rs662799	chr11:116663707	A / G	0.2831	0.0247	2.32×10^{-30}	*APOA5-APOC3*
PC(32:0) / TG(52:3)	734.5699 / 874.7859	Insulin production	rs174546	chr11:61569830	T / C	−0.1952	0.0244	1.14×10^{-15}	*FADS1-2-3*
PC(32:0) / TG(52:3)	734.5699 / 874.7859	Insulin production	rs9644639	chr8:19884947	C / G	−0.2095	0.0351	2.47×10^{-9}	*LPL*
PC(34:3) / PE(41:4)	756.5541 / 810.6012	Cardiovascular disease risk	rs174566	chr11:61592362	A / G	−0.4993	0.0240	1.57×10^{-96}	*FADS1-2-3*
PC(36:4) / LysoPC(16:0)	782.5699 / 496.3404	Inflammation	rs174550	chr11:61571478	T / C	0.3616	0.0227	2.43×10^{-57}	*FADS1-2-3*
PC(36:4) / PC(34:2)	782.5699 / 758.57	ω-6 production	rs174564	chr11:61588305	A / G	0.6929	0.0250	5.47×10^{-169}	*FADS1-2-3*
PC(36:4) / PC(34:2)	782.5699 / 758.57	ω-6 production	rs4122352	chr16:15174571	A / G	0.1543	0.0261	3.34×10^{-9}	*NTAN1-RRN3*
PE(39:2) / PC(34:2)	786.6012 / 758.57	Elongase	rs72747041	chr1:211914153	A / G	0.1452	0.0249	5.34×10^{-9}	*LPGAT1*
PC(38:6) / PC(34:2)	806.5694 / 758.57	DHA level	rs174566	chr11:61592362	A / G	0.3027	0.0243	1.61×10^{-35}	*FADS1-2-3*
PC(38:6) / PC(34:2)	806.5694 / 758.57	DHA level	rs6920155	chr6:11047956	A / C	−0.1089	0.0190	9.83×10^{-9}	*ELOVL2*
PC(38:5) / PC(34:2)	808.5851 / 758.57	EPA level	rs174567	chr11:61593005	A / G	0.4339	0.0237	4.04×10^{-75}	*FADS1-2-3*
SM(42:1) / TG(48:2)	815.7001 / 820.739	Insulin production	rs662799	chr11:116663707	A / G	0.1940	0.0246	3.46×10^{-15}	*APOA5-APOC3*
SM(42:1) / TG(48:2)	815.7001 / 820.739	Insulin production	rs115129770	chr8:19844439	C / G	−0.1856	0.0326	1.25×10^{-8}	*LPL*
TG(48:1) / TG(54:5)	822.7546 / 898.7856	*De novo* lipogenesis	rs71487172	chr10:86891371	T / C	0.1735	0.0318	4.73×10^{-8}	*LOC100507470*
TG(52:5) / TG(52:4)	870.7544 / 872.7702	Desaturase	rs174567	chr11:61593005	A / G	0.2203	0.0241	7.29×10^{-20}	*FADS1-2-3*
TG(54:5) / CE(18:2)	898.7856 / 666.6183	C-peptide level	rs79626409	chr19:45417638	CTTCG / C	0.1975	0.0353	2.21×10^{-8}	*APOC1*
TG(54:5) / CE(18:2)	898.7856 / 666.6183	C-peptide level	rs662799	chr11:116663707	A / G	−0.2635	0.0246	9.02×10^{-27}	*APOA5-APOC3*
TG(54:5) / CE(18:2)	898.7856 / 666.6183	C-peptide level	rs9644639	chr8:19884947	C / G	0.2035	0.0349	5.29×10^{-9}	*LPL*
TG(54:3) / TG(52:3)	902.8175 / 874.7859	Elongase	rs8107974	chr19:19388500	A / T	−0.2159	0.0313	5.68×10^{-12}	*HAPLN4-TM6SF2-PBX4*

This table shows only the most strongly associated variant within each locus that was associated with each ratio. **Abbreviations: EA** = Effect allele; **GRCh37** = Genome Reference Consortium human genome build 37; **NEA** = Non-effect allele; **SE** = Standard error.

5.4 Discussion

The GWAS conducted on 444 lipid metabolites in 5662 individuals revealed a considerable number of associations between SNPs and lipid metabolites, particularly in the *FADS1-2-3* and *APOA5-APOC3* regions. The *FADS1-2-3* locus impacts fatty acid desaturase and has previously been shown to be associated with lipid metabolites[130,132]. The *APOA5-APOC3* locus is also a well-known CHD gene region. These and other findings helped confirm that the GWAS results were as expected and validated existing findings.

Overall, the conditional analyses identified 254 lipids that were significantly associated with one or more genetic variant(s) and 355 associations between SNPs and lipids, with a total of 89 lead variants from 23 loci.

A potential limitation of this analysis is that the sentinel variants identified within each locus were lipid-specific, so this means that while the list of significant variants for each lipid were independent, the combined list of sentinel variants across all lipids included variants that were in LD with each other. This could have been avoided by selecting one sentinel variant for each locus across all lipids, but this would have meant that for many lipids the sentinel variant would not be the variant with the strongest P-value.

For the vast majority of the significant loci from the conditional analyses, there were differences in the magnitude and direction of the associations of lipids from different subclasses with the lead variant in each locus. For *LPL* in particular, there were 13 lipids that had significant associations with variants in the *LPL* region, which is also a well-characterised CHD locus. This lipidomics platform showed that individuals with genetically lower *LPL* activity have increased levels of triglycerides—especially those containing monounsaturated fatty acids—and decreased levels of sphingomyelins and cholesterol esters. This is the first analysis to show a link between *LPL* activity and sphingomyelin levels.

Out of the 175 loci associated with major circulating lipids, it was observed that lipid metabolites were only significantly associated with 13 of these loci. However, this may be due to the fact that the 175 major lipid loci were identified from large consortia such as the GLGC[35], which included 188 577 individuals in its analysis. The analysis using this lipidomics platform may not have had sufficient power to detect these associations.

The analyses of principal components and ratios of lipids also provided new biological insights into lipid metabolism. As described in Chapter 4, the PCA of the lipid metabolites showed that the second principal component revealed a contrast between free fatty acid levels versus small, saturated triglycerides. The significant variation in length of fasting status prior to blood draw would have strongly affected the levels of both free fatty acids and triglycerides, but may have also obscured the genetic associations. Indeed, in the GWAS of the first four principal components there was only one variant (rs662799, chr11:116663707) in the *APOA5-APOC3* genetic locus that showed a genome-wide significant association with the second principal component (Figure 5.11a and Figure 5.12a). Variants in the *APOA5-APOC3* region have been associated with type 2 diabetes[15], fatty liver disease[207], hypertrigylceridemia[208,209], and dyslipidaemia[210]. This gene region has also been associated with metabolites such as 1-linoleoylglycerol in previous metabolomics studies[132]. However, as only one variant was significantly associated with the second principal component compared to 74 variants that were associated with the third and fourth principal components (Table 5.3), this suggests that the second component was largely driven by dietary patterns rather than genetic differences. As described earlier, the lipid species that contributed most to the second principal component are also affected by obesity and insulin secretion/sensitivity, and this principal component is significantly associated with being overweight and having diabetes.

Although the third principal component was most closely characterised by unsaturated triglycerides, it did not show any significant genetic associations with variants in the *FADS1-2-3* locus, and was only associated with variants in the *APOA5-APOC3* region (Figure 5.11b and Figure 5.12b). In contrast, variants in both the *FADS1-2-3* and *APOA5-APOC3* regions were significantly associated with the fourth principal component (Figure 5.11c, Figure 5.12c and Figure 5.12d). The *APOA5-APOC3* locus has been previously highlighted in other lipidomics studies[70] and previous GWAS[18,211,212]. As described in the previous chapter, the loadings of the fourth principal component showed that linoleic acid–containing lipids had negative loading scores, while sphingomyelins containing odd-chain fatty acids and desaturated phospholipids had positive loading scores. The association between SNPs in the *FADS1-2-3* region with sphingomyelins has been described previously[213], although not explained, and has not previously been described for odd-chain fatty acid–containing sphingomyelins. The impact on triglycerides also explained the association of SNPs in the *APOA5-APOC3* locus with the fourth principal component. Both the third and fourth principal components showed negative associations with the relative risk for being overweight and having diabetes, while only the fourth principal component showed a negative association with the relative risk for hypertension. When considering the additional evidence that the genetic findings bring, this last observation is striking since sphingomyelins have thus far been implicated in hypertension as precursors to ceramide production, but odd-chain fatty acid–containing sphingomyelins have mostly been unexplored.

Previous mGWAS have shown that genetic analyses of ratios of metabolites can yield additional insights that are not detected from analysis of the individual metabolites that make up the ratio[57,72,73,82,132,168]. For instance, one study identified eight additional loci from metabolite ratios that were not detected from the GWAS of the individual metabolites[132]. In several cases, the metabolite ratio appeared to reflect flux through a particular metabolic reaction that was influenced by the associated variant, while in other cases, the metabolites in the ratio were linked to either a substrate or a product, so the genetic variant may have caused one molecule to be consumed or acted on faster than the other[132]. Ultimately, despite testing 98 346 pairwise metabolite ratios, the study only detected eight additional loci[132], which is a limited gain for substantial effort. The analysis of such a large number of ratios also required a more stringent penalty for the Bonferroni-corrected P-value, which may have resulted in discarding true signals that could been identified with a more focused approach. For these reasons, in the analyses for this dissertation it was decided to limit the analysis to 26 pairwise ratios, identified through expert curation, that had strong biological rationales and acted through thoroughly understood metabolic pathways. Testing all possible pairwise ratios between lipid metabolites may have led to the identification of additional loci associated with lipids and their ratios, but such intensive analyses were beyond the scope of the genetic analyses presented in this chapter.

This platform can provide many novel insights into the genetic determinants of lipids and detailed information about lipid metabolism. The genetic findings that were described in this chapter will be explored further in Chapter 6. Additionally, the significantly associated variants will be annotated to identify the most likely causal gene(s), the biological insights will be interpreted, and novel lipids will be identified which have not previously been reported in GWAS of major circulating lipids or metabolomics.

CHAPTER 6

Identification of novel loci and interpretation of genetic findings

Chapter summary

This chapter provides an in-depth analysis and interpretation of the genome-wide association study (GWAS) results of the lipid metabolites. The pipeline that was used to annotate the significant variants is described, including the incorporation of information from pharmacological and functional databases to aid biological understanding of mechanisms through which variants influence metabolic pathways. The GWAS findings were also compared to previously published studies of metabolomics and circulating major lipids to identify novel loci, which had not previously been reported for association with major lipids or metabolites. This resulted in the discovery of four novel loci: *UGT8*, *XBP1*, *GAL3ST1*, and *PNPLA3*. Novel relationships between genetic variants and lipids were also identified, and new biological insights into lipid metabolism for existing loci were revealed.

6.1 Introduction

Metabolomics GWAS (mGWAS), which were described in Chapter 1, are genome-wide association studies using multiple metabolites measured on high-dimensional phenotyping platforms as phenotypic traits[71]. The literature review of published mGWAS identified 31 studies, which are listed in Table 1.3. The supplementary data from each of these studies were downloaded and combined into a single dataset, which provides a valuable resource to compare against when identifying novel loci from a newly conducted GWAS.

The translation of genetic findings from isolated variants to gene function and impact on diseases is challenging. Genetic association studies essentially do little more than identify a genomic location related to a trait or disease, but provide meagre evidence of gene function unless the single nucleotide polymorphisms (SNPs) have been found to exhibit predictable effects on gene expression[33]. Efforts to discern the most likely causal gene and its associated function have included examining of known variants in high linkage disequilibrium (LD) with the associated SNP to identify variants with plausible biological effects, studying gene expression from tissue samples or cell lines, and conducting other functional studies such as knockouts in cell or animal models[33]. A wealth of functional information is now available in large databases, which can be utilised effectively to annotate significant variants identified from a GWAS and determine the most likely causal genes and relevant biological pathways that may be implicated in disease outcomes.

6.2 Methods

6.2.1 Functional annotation pipeline

In order to annotate SNPs with potential causal genes, a two-pronged strategy was implemented consisting of a bottom-up and top-down approach. The approach described here was customised specifically for the analysis presented in this dissertation, but a more generalised version of the functional annotation pipeline has been published[214] and is publicly available (https://github.com/ds763/ProGeM) to facilitate implementation in other projects. An overview of the functional annotation pipeline is shown in Figure 6.1.

In the bottom-up approach, SNPs were annotated according to their putative effects on proximal gene function by examining whether these SNPs influence protein sequencing, gene splicing, and/or mRNA levels of a local gene. Conversely, in the top-down approach, SNPs were annotated according to previous knowledge concerning local gene function by examining whether proximal genes have been previously implicated in lipid metabolism. In cases where (1) SNPs were purported to exert effects on more than one local gene, and/or (2) more than one local gene was previously implicated in lipid metabolism, SNPs were assigned to multiple genes rather than force-assigning each to a single gene. In cases where it was not possible to annotate SNPs using either the bottom-up or top-down approach, the SNPs were assigned to their nearest gene.

For the bottom-up approach, SNPs were annotated based on their putative effects on proximal gene function if any of the following conditions were met: (1) the SNP resided within an exonic sequence of a gene; (2) the SNP resided within a splice-site (±2-bp from an intron-exon boundary); (3) the SNP was in high LD ($r^2 \geq 0.8$) with a non-synonymous SNP; and/or (4) the SNP was a *cis*-eQTL for a local gene. To identify any exonic and splice site SNPs within the SNP list, Variant Effect Predictor (VEP)[215] (http://www.ensembl.org/common/Tools/VEP/) was run on the list of variants with the "pick" option (which outputs one block of annotation per variant), and Ensembl transcripts used as the reference for determining consequences. SNPs in high LD with the list of associated SNPs were identified within the imputed dataset and run through VEP to select only non-synonymous SNPs. *Cis*-eQTLs within the list of associated SNPs were identified using eQTL data provided by the Genotype-Tissue Expression (GTEx) project[216] (http://www.gtexportal.org), keeping significant SNP–gene associations only (filename: ⟨GTEx_Analysis_V6_eQTLs.tar.gz⟩). SNPs were only annotated if they were significant eQTLs in at least one of the following tissues that were deemed most relevant for lipid-related phenotypes: subcutaneous adipose tissue, visceral adipose tissue, liver, and/or whole blood.

In the top-down approach, for each of the associated SNPs, a list of all proximal genes located ≤ 500-Kb upstream or downstream was first identified using the ANNOVAR tool (http://annovar.openbioinformatics.org). All genes previously associated with a lipid-related biological process or function were then identified from the following databases: (1) LIPID MAPS Proteome Database (LMPD) (http://www.lipidmaps.org/data/proteome/LMPD.php); (2) Gene Ontology (GO) (http://geneontology.org); (3) Online Mendelian Inheritance in Man (OMIM) (http://www.omim.org); (4) Mouse Genome Informatics (MGI) (http://www.informatics.jax.org); (5) Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/); and/or (6) Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com/products/ipa/).

LMPD is an object-relational database of lipid-associated genes and proteins across multiple species including human, mouse, and fruit fly; all human genes (1116 in total) were simply extracted from this database (accessed 16-Mar-2016). For GO and OMIM, all terms or Mendelian diseases containing one or more lipid-related keyword(s) using HumanMine (http://www.humanmine.org) were identified, then all human genes associated with one or more of these terms were extracted (accessed 01-Apr-2016 and 07-Apr-2016, respectively). Similarly, for MGI all mouse genes were extracted using MouseMine (http://www.mousemine.org/mousemine/) (accessed 31-Mar-2016) that were associated with the following manually-selected lipid-related terms and their children: (1) abnormal lipid homeostasis (MP:0002118); (2) abnormal lipoprotein level (MP:0010329); (3) abnormal lipid metabolism (MP:0013245); and (4) adipose tissue phenotype (MP:0005375). From the KEGG database all lipid compounds (with "C" number IDs) with biological roles were first extracted in order to then identify all genes associated with reactions (with "R" number IDs) involved in lipid biology using MitoMiner (http://mitominer.mrc-mbu.cam.ac.uk/release-3.1/) (accessed 31-Mar-2016). Finally, from IPA the interaction networks were downloaded for all fourteen of the lipid subclasses in order to extract all genes in a compound-specific manner (accessed 13-Apr-2016).

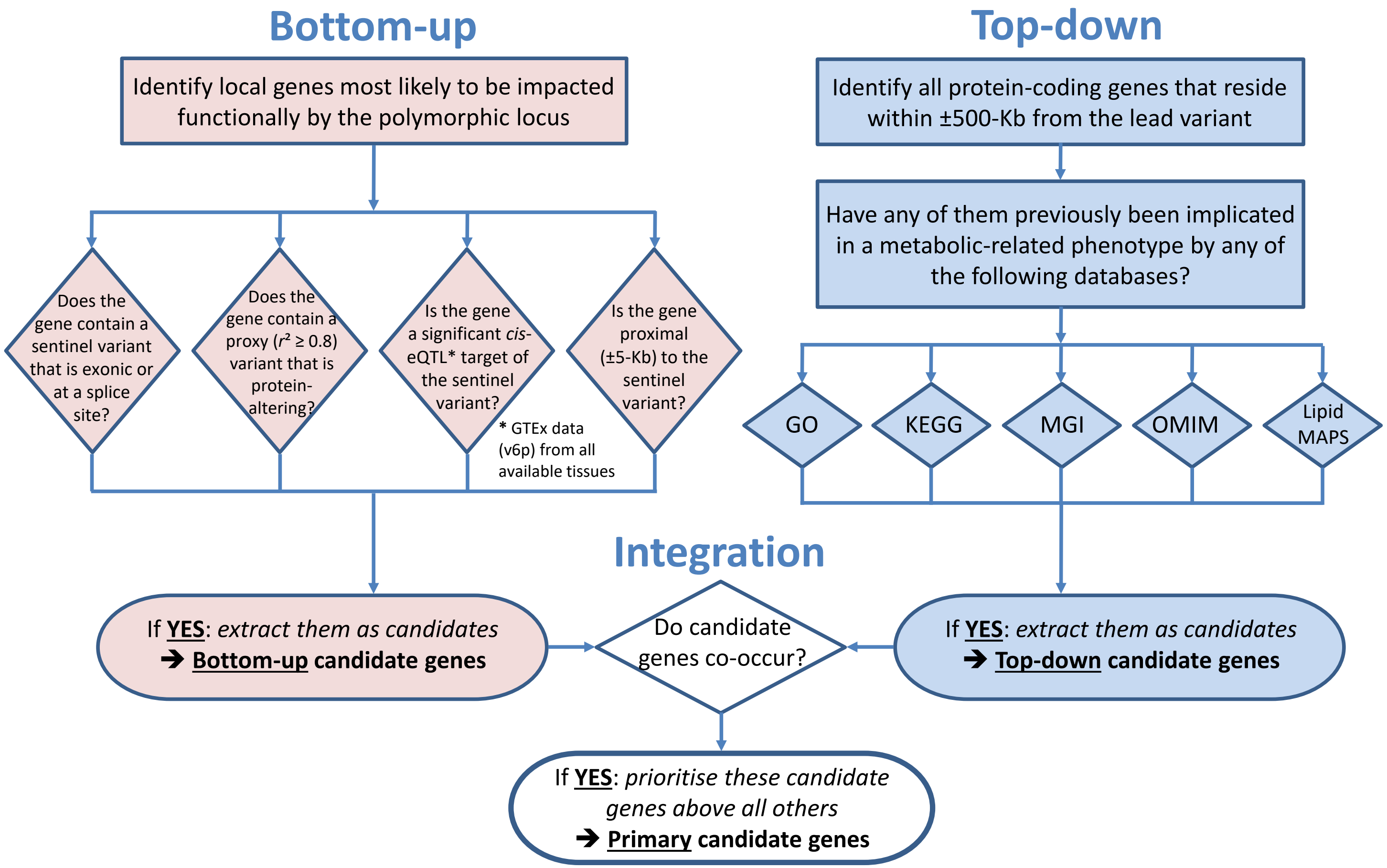
Once the lists of lipid-related genes were obtained from the aforementioned databases, HUGO Gene Nomenclature Committee (HGNC) symbols were used to search for overlap with the list of proximal ($\leq$ 500-Kb) genes, thereby annotating SNPs with proximal genes where there was evidence that at least one might be involved in lipid-related biology. For each lead SNP, it was first recorded whether there was any compound-specific evidence from IPA for a SNP–gene assignment whereby both the SNP (from this study) and the gene (from IPA) were associated with the same lipid subclass. Then from the five remaining (compound non-specific) databases, the overlapping genes were categorised as either (1) recurrent candidates, in that they were highlighted in at least two different databases, or (2) single candidates. Further, the recurrent candidates were assigned a score out of five for prioritisation purposes, with one point awarded for each database highlighting them as being lipid-related.

After performing comprehensive annotation of SNPs as per the bottom-up and top-down procedures described above, this information was then integrated to predict the most likely causal gene(s) using a hierarchical approach as follows: (1) For those lead SNPs where the same gene was highlighted by both the bottom-up and top-down approach, this gene was selected as the putative causal gene; (2) If both the SNP (from this study) and the proximal gene (from IPA) were associated with the same lipid subclass, further

SNP–gene assignments were made accordingly; (3) Finally, for each of the remaining lead SNPs, the highest scoring top-down gene and any bottom-up genes were assigned as the likely causal gene(s).

Separately, an expert-curated causal gene was assigned to each variant and compared to the predicted causal genes identified by the functional annotation pipeline to assess concordance and validate the pipeline.

Figure 6.1: Flow diagram outlining strategy for causal gene prioritisation


Bottom-up
Identify local genes most likely to be impacted functionally by the polymorphic locus
Does the gene contain a sentinel variant that is exonic or at a splice site?
Does the gene contain a proxy ($r^2 \geq 0.8$) variant that is protein-altering?
Is the gene a significant *cis*-eQTL* target of the sentinel variant?
* GTEx data (v6p) from all available tissues
Is the gene proximal (±5-Kb) to the sentinel variant?
Top-down
Identify all protein-coding genes that reside within ±500-Kb from the lead variant
Have any of them previously been implicated in a metabolic-related phenotype by any of the following databases?
GO
KEGG
MGI
OMIM
Lipid MAPS
Integration
If **YES**: *extract them as candidates*
➔ **Bottom-up candidate genes**
Do candidate genes co-occur?
If **YES**: *extract them as candidates*
➔ **Top-down candidate genes**
If **YES**: *prioritise these candidate genes above all others*
➔ **Primary candidate genes**

A proxy is defined as those variants with $r^2 \geq 0.8$ with the lead variant (EUR population, 1000 Genomes). GTEx data (v6p) from all available tissues were used as a source for identifying *cis*-eQTLs. **Abbreviations: eQTL** = Expression Quantitative Trait Locus; **GO** = Gene Ontology; **GTEx** = Genotype–Tissue Expression; **KEGG** = Kyoto Encyclopedia of Genes and Genomes; **Lipid MAPS** = Lipid Metabolites and Pathways Strategy; **MGI** = Mouse Genome Informatics; **OMIM** = Online Inheritance in Man. Source: Adapted from Stacey D, et al. *BioRxiv*, 2017;230094[214].

6.2.2 Identification of novel loci

In order to determine the number of novel loci from the univariate GWAS and conditional analyses, the raw results files from a GWAS of circulating lipids published by the Global Lipids Genetics Consortium (GLGC)[35] were downloaded and combined with the published results from the metabolomics GWAS papers. The published results database was filtered to only include results significant at $P < 5 \times 10^{-8}$, the standard level for genome-wide significance.

In order to be thorough and comprehensive, several different approaches were devised and compared to identify novel loci. First, all variants from the combined univariate GWAS of each lipid and the conditional analyses that were significant at $P < 8.929 \times 10^{-10}$ were compared to the published SNPs to identify novel SNPs that were more than 500-Kb from any published SNPs—in other words, SNPs that had not previously been reported in GWAS of either metabolomics or circulating lipids.

A second approach that was used to identify novel loci was based on simply determining if the variants matched any previously published variants or any proxies of those variants. Proxy SNPs in linkage disequilibrium (LD) with the published SNPs were identified using PhenoScanner[217], an in-house resource developed by the Cardiovascular Epidemiology Unit, which has also been made publicly accessible online (http://www.phenoscanner.medschl.cam.ac.uk). PhenoScanner was used to search for proxies of published variants found in dbSNP version 138[218] and 1000 Genomes phase 3[141] with a maximum distance of ±500-Kb and an LD $r^2 \geq 0.6$. SNPs with a MAF $\geq 0.5\,\%$, that were biallelic, or that had small insertions or deletions (indels) (< 5 bases) were also included. Using this approach, any variants that did not match previously published variants or proxies of those variants were considered to be novel.

A third approach was also used to identify novel loci that combined information about both distance and LD. All published results from the GWAS of circulating lipids (GLGC) and metabolomics GWAS were selected and filtered to only include variants reached genome-wide significance ($P < 5 \times 10^{-8}$). A window was then established around each published variant, with the size of the window determined by the distance of the farthest proxy SNP upstream and downstream of the published SNP with an LD $r^2 \geq 0.6$. Again, the proxies were searched using PhenoScanner with the same criteria as described in the previous approach. If no proxies were found for the published SNP, a default distance of ±500-Kb was used for the size of the window around the published SNP. Finally, all PROMIS variants were selected ($P < 8.929 \times 10^{-10}$) that fell outside those windows, and these were determined to be novel.

6.2.3 Network of genetic and metabolic associations

Cytoscape v3.4.0[219] was used to generate a network of associations between genes and lipids. Using the approach described in Chapter 4, a Gaussian graphical model (GGM) was constructed to connect lipids with other lipids based on partial correlation coefficients[130,179,180]. Metabolites were also connected with genetic loci using the univariate GWAS results, with one link for each genome-wide significant association. The full network facilitates visualisation of the genetic determinants of human metabolism and the relationships between genetic loci and lipids.

The network diagrams were created by combining two parts to integrate different sources of information. The first part was created by loading the reported associations between lipids and genes into Cytoscape. Lipid species were clustered according to the lipid subclass they belong to, resulting in fourteen distinct lipid nodes in the network. The 89 identified lead SNPs from the conditional analyses were clustered according to their corresponding predicted causal gene(s).

For the second part, a functional interaction network consisting solely of the list of predicted causal genes/loci was created in Cytoscape using interaction network data downloaded from Ingenuity Pathway Analysis (IPA) that had been merged using in-house R[156] scripts to create a `.sif` file. For loci with multiple potential causal genes, interaction networks for all genes were extracted from IPA and an edge was drawn if at least one gene at that locus functionally interacted with another of the lipid-associated genes according to IPA. Finally, these two parts were merged together by node names (i.e. gene symbols). No enrichment statistics (e.g. KEGG pathways or GO terms) or other statistical information was used to produce the network, since this information was already incorporated to inform the predictions of the most likely causal genes, and would therefore invalidate the conclusions if also used to inform the network.

In order to examine the associations between genes and lipids in the network more closely, a second network diagram was created showing the individual lipid species rather than just the lipid subclasses as a whole. However, since a network displaying hundreds of lipids and their associated genetic loci would have been overly complex and too difficult to read or interpret, only a portion of the network was drawn focusing on just the triglycerides and their genetic associations. Thus, this network diagram portrayed the partial correlations of the triglycerides with each other and the association of each triglyceride with genetic loci.

6.3 Results

6.3.1 Annotation of significant variants from conditional analyses

The functional annotation pipeline was used to annotate the 89 sentinel variants (i.e. those with the strongest P-values) from the conditional analyses. Table 6.1 shows the lipid subclasses associated with each variant, summarises the consequence of each variant on the protein sequence, and indicates the predicted causal locus based on integration of information from the bottom-up and top-down SNP annotation approaches together with expert curation. The concordance between the predicted causal genes identified by the functional annotation pipeline and the expert-curated causal genes was 85 %, demonstrating that the pipeline exhibited high sensitivity.

Once a predicted causal locus had been identified for each sentinel variant, further steps were taken to annotate the loci and summarise known associations with the variants in each locus. A list of the 23 significant loci from the conditional analyses is provided in Table 6.2. The names and putative functions of the genes in each locus are described, along with a list of the lipid-related biomarkers, diseases, and tissues from gene expression data that are associated with the significant variants and their proxies in each locus, which were looked up from PhenoScanner[217] and the mGWAS literature review of published metabolomics GWAS, and supplemented with a further literature search. Data on CAD were contributed by CARDIoGRAMplusC4D investigators (downloaded from `http://www.cardiogramplusc4d.org`)[43]; data on other outcomes were obtained from the NHGRI-EBI GWAS Catalog[34]; gene expression tissue data were obtained from the GTEx Consortium[216]. An exhaustive list of all the associations can be downloaded from PhenoScanner, but Table 6.2 provides an overview of the principle associations that are reported for each locus. Additionally, a few summary results from PROMIS are shown to facilitate comparison of the lipidomics analysis from this dissertation with the published data. The last two columns provide a list of lipid subclasses and the number of lipid metabolites from PROMIS that are significantly associated with the sentinel variants in each locus.

One of the findings from this table is that the *SCD* locus, which has been shown to be associated with phosphocholines and sphingomyelins, was only associated with free fatty acids and lysophosphocholines in this study. The putative gene function for *SCD* is fatty acid biosynthesis, so the association with free fatty acids is not surprising, but the lack of

Table 6.1: Functional annotation of significant variants from conditional analyses

Locus	rsID	Chr:Pos (GRCh37)	EA/NEA	Associated lipid subclasses	SNP consequence	Predicted causal locus
1	rs6657050	chr1:63105253	G/A	PC, PI	Intronic	*ANGPTL3* §#
2	rs1260326	chr2:27730940	C/T	SM	Missense, splice site	*GCKR* *†#
3	rs9821138	chr3:142659837	G/A	PI	Intronic	*PAQR9*
	rs4683715	chr3:142664819	G/A	PA	Downstream	*PAQR9*
4	rs28870381	chr4:115478499	T/G	PG	Intergenic	*UGT8*
5	rs9468308	chr6:11044068	G/T	PE	Intronic	*ELOVL2*
	rs6920155	chr6:11047956	A/C	PI	Upstream	*ELOVL2*
	rs9393915	chr6:11072322	T/C	PA, PC	Intronic	*ELOVL2*
6	chr7:73042302	chr7:73042302	G/GCTTT	SM	Upstream	*MLXIPL* ‡#
7	rs117199990	chr8:19820916	T/C	DG, SM, TG	Intronic	*LPL* ‡¶#
	rs17482753	chr8:19832646	T/G	CE	Intergenic	*LPL* ‡#
	rs115129770	chr8:19844439	G/C	PC	Intergenic	*LPL* ‡#
	rs77237194	chr8:19865455	T/A	DG	Regulatory region	*LPL* ‡#
	rs9644639	chr8:19884947	G/C	SM, TG	Intergenic	*LPL* ‡¶#
8	rs603424	chr10:102075479	A/G	FreeFA, LysoPC	Intronic	*SCD* §¶#
9	rs174530	chr11:61546592	G/A	PA, PC	Intronic	*FADS1-2-3* §#
	rs174533	chr11:61549025	A/G	PC, PE, PI	Intronic	*FADS1-2-3* §#
	rs102275	chr11:61557803	C/T	PC	Intronic	*FADS1-2-3* §#
	rs174544	chr11:61567753	A/C	PC, PE, PI	3' UTR	*FADS1-2-3* *§#
	rs174545	chr11:61569306	G/C	PA, PC, PS	3' UTR	*FADS1-2-3* *§#
	rs174546	chr11:61569830	T/C	CE, PI	3' UTR	*FADS1-2-3* *§#
	rs174548	chr11:61571348	G/C	PE, PI	Intronic	*FADS1-2-3* §#
	rs174549	chr11:61571382	A/G	PA, TG	Intronic	*FADS1-2-3* §¶#
	rs174551	chr11:61573684	C/T	PC	Intronic	*FADS1-2-3* §#
	rs174553	chr11:61575158	G/A	CE, PC, PE, PI	Intronic	*FADS1-2-3* §#
	rs174554	chr11:61579463	G/A	PA	Intronic	*FADS1-2-3* §#
	rs174560	chr11:61581764	C/T	PA, PC, PE, PI, PS	Intronic	*FADS1-2-3* §#
	rs174561	chr11:61582708	C/T	FreeFA, PC	Intronic	*FADS1-2-3* §¶#
	rs28456	chr11:61589481	G/A	PA, TG	Upstream	*FADS1-2-3* §¶#
	rs174565	chr11:61591636	G/C	PA, PE	Upstream	*FADS1-2-3*
	rs174566	chr11:61592362	G/A	PC, PE, PG, SM	Upstream	*FADS1-2-3* §#
	rs174567	chr11:61593005	G/A	PA, PC, PE, PS	Upstream	*FADS1-2-3* §#
	rs174568	chr11:61593816	T/C	PC	Upstream	*FADS1-2-3* §#
	chr11:61594920	chr11:61594920	C/CT	PS	Upstream	*FADS1-2-3* ‡
	rs968567	chr11:61595564	T/C	PE	5' UTR	*FADS1-2-3* *§#
	chr11:61596322	chr11:61596322	C/CA	PA, PC	Intronic	*FADS1-2-3* ‡
	rs99780	chr11:61596633	T/C	FreeFA	Intronic	*FADS1-2-3* §¶#
	rs1535	chr11:61597972	G/A	PC, PG	Intronic	*FADS1-2-3* §#
	rs61897793	chr11:61599347	A/G	FreeFA	Intronic	*FADS1-2-3* §¶#
	rs174574	chr11:61600342	C/A	FreeFA, LysoPC	Intronic	*FADS1-2-3* §¶#
	chr11:61602460	chr11:61602460	C/CA	PA, PC, PI	Intronic	*FADS1-2-3*
	rs174576	chr11:61603510	A/C	PA, PC, PI	Intronic	*FADS1-2-3* §#
	rs174578	chr11:61605499	A/T	PC	Intronic	*FADS1-2-3* §#
	rs174580	chr11:61606642	G/A	PA, PC, PE	Intronic	*FADS1-2-3* §#
	rs174581	chr11:61606683	A/G	PC	Intronic	*FADS1-2-3* §#
	rs174582	chr11:61607168	G/A	PE	Intronic	*FADS1-2-3* §#
	rs61897795	chr11:61618169	G/A	PE	Intronic	*FADS1-2-3* §#
10	rs964184	chr11:116648917	C/G	CE, DG, PA, PC, SM, TG	3' UTR	*APOA5-APOC3* *¶#
	rs11604424	chr11:116651115	T/C	CE	Intronic	*APOA5-APOC3* ¶
	rs6589566	chr11:116652423	A/G	SM, TG	Intronic	*APOA5-APOC3* ¶
	rs2075290	chr11:116653296	T/C	TG	Intronic	*APOA5-APOC3* ¶
	rs3741298	chr11:116657561	T/C	TG	Intronic	*APOA5-APOC3* ¶
	rs651821	chr11:116662579	T/C	CE, DG, PC, SM, TG	5' UTR	*APOA5-APOC3* *¶#
	rs662799	chr11:116663707	A/G	CE, Chol, DG, PC, SM, TG	Upstream	*APOA5-APOC3* ¶§
11	rs7160525	chr14:64232220	A/G	SM	Downstream	*SGPP1*
	rs7157785	chr14:64235556	T/G	SM	Regulatory region	*SGPP1*
12	rs11158671	chr14:67965452	A/G	PG	Intronic	*PIGH-TMEM229B* §
	rs1885041	chr14:67976325	T/C	PA, PC, SM	Intronic	*PIGH-TMEM229B* §
13	rs11071371	chr15:58576226	T/C	PE	Intronic	*LIPC* ‡#
	rs2043085	chr15:58680954	C/T	PC, PE	Intronic	*LIPC*
	rs1532085	chr15:58683366	G/A	PA, PC, PE	Intronic	*LIPC*
	rs1077835	chr15:58723426	G/A	PC, PE	Intronic	*LIPC*
	rs1077834	chr15:58723479	C/T	PE	Intronic	*LIPC*
	chr15:58723675	chr15:58723675	T/C	PC, PE	Intronic	*LIPC*
	rs2070895	chr15:58723939	A/G	PC, PE	Intronic	*LIPC*
14	rs4985124	chr16:15125441	G/T	PA	Intronic	*PLA2G10-NTAN1-NPIPA5* ‡§
	rs1135999	chr16:15131962	G/A	PE	Missense	*PLA2G10-NTAN1-NPIPA5* *§
	rs34955778	chr16:15139594	C/T	PC, PE	Intronic	*PLA2G10-NTAN1-NPIPA5* ¶§
	rs11644601	chr16:15172118	C/T	CE	Intronic	*PLA2G10-NTAN1-NPIPA5* §
15	rs711752	chr16:56996211	A/G	PC	Intronic	*CETP* ¶
16	rs11079173	chr17:53487664	C/A	PA	Intronic	*PCTP*
17	rs11666866	chr19:8285607	A/G	PC, SM	Intronic	*CERS4* §#
18	rs75627662	chr19:45413576	T/C	SM	Upstream	*APOE-C1-C2-C4* ¶
	rs483082	chr19:45416178	T/G	PC, SM	Upstream	*APOE-C1-C2-C4* ¶
19	rs4806498	chr19:54674742	T/C	PA, PE, PI	Intronic	*MBOAT7* ¶§
	rs641738	chr19:54676763	T/C	PA, PC, PI	Missense	*MBOAT7* *¶§
	rs34564463	chr19:54676814	GC/G	PC, PI	5' UTR	*MBOAT7* *¶
	rs626283	chr19:54677001	C/G	PC, PI	Upstream	*MBOAT7* ¶§
	rs8736	chr19:54677189	T/C	PC, PI	3' UTR	*MBOAT7* *¶#
	rs10416555	chr19:54677897	A/G	PI, PS	Synonymous	*MBOAT7* *¶
20	rs438568	chr20:12958687	G/A	Cer, PC, SM	Intergenic	*SPTLC3* §#
	rs4814175	chr20:12959094	T/A	Cer	Intergenic	*SPTLC3* §#
	rs4814176	chr20:12959398	C/T	PC, SM	Intergenic	*SPTLC3* §#
	rs364585	chr20:12962718	G/A	Cer, SM	Intergenic	*SPTLC3* §#
	rs168622	chr20:12966089	G/T	Cer	Intergenic	*SPTLC3* §#
	rs686548	chr20:12973521	T/A	PC, SM	Intergenic	*SPTLC3* §#
21	chr22:29339470	chr22:29339470	T/TTCTC	SM	Intronic	*XBP1*
22	rs2267161	chr22:30953295	T/C	PG	Missense	*GAL3ST1* *#
23	rs12484809	chr22:44325631	T/C	TG	Intronic	*PNPLA3* ¶§

This table shows annotation of 89 significant variants from conditional analyses, including the lipid subclasses significantly associated with each variant, the consequence of each variant on the protein sequence, and the predicted causal locus, which is based on integration of information from the bottom-up and top-down SNP annotation approaches (see subsection 6.2.1) together with expert curation. Chromosomal loci (defined using ±500-Kb rolling windows around each SNP) are indicated in the table by alternating white background and grey shading. *SNP is exonic. †SNP is located at a splice site. ‡SNP is in high LD with a non-synonymous SNP. §SNP is in high LD with a significant *cis*-eQTL. ¶There is lipid compound-specific evidence from IPA supporting the SNP-gene annotation. #Gene symbol was highlighted by both the bottom-up and top-down approaches. **Abbreviations: ANNOVAR** = ANNOtate VARiation; **CE** = Cholesterol ester; **Cer** = Ceramide; **DG** = Diacylglycerol; **eQTL** = Expression Quantitative Trait Loci; **FA** = Fatty acids; **GL** = Glycerolipid; **GLGC** = Global Lipids Genetics Consortium; **GP** = Glycerophospholipid; **GRCh37** = Genome Reference Consortium human genome build 37; **IPA** = Ingenuity Pathway Analysis; **LD** = Linkage Disequilibrium; **LPC** = Lysophosphatidylcholine; **PA** = Phosphatidate; **PC** = Phosphatidylcholine; **PE** = Phosphatidylethanolamine; **PG** = Phosphatidylglycerol; **PI** = Phosphatidylinositol; **PS** = Phosphatidylserine; **SL** = Sphingolipid; **SM** = Sphingomyelin; **SNP** = Single Nucleotide Polymorphism; **ST** = Steroids and derivatives; **TG** = Triacylglycerol; **VEP** = Variant Effect Predictor.

associations with other lipid subclasses that have been reported in previous studies could be due to lack of statistical power or the fact that this study was conducted in a Pakistani population, whereas all of the published mGWAS studies are based in European, American, or Brazilian populations. Although the GLGC included fine-mapping in a South Asian population, *SCD* was not one of the 157 significant loci identified in this paper[35].

Additionally, the *CERS4* has previously been reported to be associated with ceramides and sphingomyelins, whereas in this analysis it was associated with phosphocholines and sphingomyelins. Another difference is that the *PIGH-TMEM229B* locus has only been reported to be associated with phosphocholines, whereas this analysis extends further what is known about this gene region, as associations were also found with phosphatic acids, phosphoglycerols, and sphingomyelins.

The *APOA5-APOC3* locus is the only gene region from the conditional analyses that was significantly associated with CHD. Together with *FADS1-2-3*, these two loci have been widely studied and are associated with a large number of different lipid-related traits in the published literature, including apolipoproteins, cholesterol esters, free cholesterol, HDL-C, LDL-C, linoleic acid, triglycerides, and omega-3 and omega-6 fatty acids. Accordingly, they were each associated with over 100 lipids in PROMIS from a wide diversity of lipid subclasses.

Table 6.2: Summary of known and novel information about conditional analysis loci

Genetic locus				Published associations with locus			Associations in PROMIS	
Cytogenetic band	**Locus name**	**Full gene name(s)**	**Gene function(s)**	**Biomarkers**	**Diseases**	**Gene expression tissues**	**Lipid subclasses**	**No. lipid species**
1p31.3	*ANGPTL3*	Angiopoietin-like 3	Angiogenesis	Cholesterol esters, free cholesterol, LA, ω-6 fatty acids, phosphoglycerides, total cholesterol, triglycerides, VLDL	**CAD**	Lymphoblastoid cell lines, skin, transformed fibroblast cells	PC, PI	2
2p23.3	*GCKR*	Glucokinase regulator	Glucokinase inhibition	Alanine, albumin, ApoA1, ApoB, ApoC3, blood proteins, cholesterol, CRP, fasting glucose, HDL, LDL, LA, metabolite levels, ω-3 fatty acids, ω-6 fatty acids, polyunsaturated fatty acids, phospholipids, total cholesterol, total fatty acids, triglycerides, VLDL	CKD, gout, hypertriglyceridemia, T2D, NAFLD, Crohn's disease	Liver, lymphoblastoid cell lines, peripheral blood, skeletal muscle, transformed fibroblast cells, thyroid, whole blood	SM	2
3q23	*PAQR9*	Progestin and adipoQ receptor family member 9	Receptor activity	Phospholipids, phosphoethanolamines	N/A	N/A	PA, PI	2
4q26	*UGT8*	Uracil-diphosphate glucose (UDP) glycosyltransferase 8	Galactose to ceramide catalysis	N/A	N/A	Transformed fibroblast cells	PG	1
6p24.2	*ELOVL2*	Elongation of very long chain (ELOVL) fatty acid elongase 2	Fatty acid elongation	Blood metabolite ratios, DHA	N/A	Testis	PA, PC, PE, PI	4
7q11.23	*MLXIPL*	Max-like protein X (MLX) interacting protein like	Transcriptional repressor	Cholesterol esters, free cholesterol, phospholipids, total cholesterol, total lipids, triglycerides, VLDL	**CAD**	Esophagus mucosa, pancreas, skin, subcutaneous adipose, thyroid, transformed fibroblast cells	SM	2
8p21.3	*LPL*	Lipoprotein lipase	Triglyceride hydrolysis, ligand/bridging factor for receptor-mediated lipoprotein uptake	ApoB, cholesterol esters, free cholesterol, HDL, phospholipids, total cholesterol, triglycerides, VLDL	**CAD**, hypertriglyceridemia, metabolic syndrome	Peripheral blood, peripheral blood monocytes	CE, DG, PC, SM ,TG	13
10q24.31	SCD	Stearoyl-CoA desaturase	Fatty acid biosynthesis	Blood metabolite levels, lysophosphocholines, palmitate, palmitoleic acid, phosphocholines, sphingomyelins	N/A	Peripheral blood	FreeFA, LysoPC	2
11q12.2	*FADS1-FADS2-FADS3*	Fatty acid desaturase 1-2-3	Fatty acid desaturation, fatty acid biosynthesis	ApoA1, blood metabolite levels, cholesterol esters, free cholesterol, DPA, EPA, fasting glucose, HDL, LDL, LA, ω-3-fatty acids, phosphocholines, plasma oleic acid, red blood cell fatty acids, total cholesterol, triglycerides	Crohn's disease, inflammatory bowel disease, laryngeal squamous cell carcinoma, lung cancer, rheumatoid arthritis	Adipose omental, brain cerebellum, esophagus mucosa, heart left ventricle, liver, lymphoblastoid cell lines, pancreas, peripheral leukocytes, skeletal muscle, spleen, thyroid, tibial nerve, transformed fibroblast cells, transverse colon, whole blood	CE, FreeFA, LysoPC, PA, PC, PE, PG, PI, PS, SM, TG	110

Table: Summary of known and novel information about conditional analysis loci (. . . continued)

Genetic locus			**Published associations with locus**				**Associations in PROMIS**	
Cytogenetic band	**Locus name**	**Full gene name(s)**	**Gene function(s)**	**Biomarkers**	**Diseases**	**Gene expression tissues**	**Lipid subclasses**	**No. lipid species**
11q23.3	*APOA5-APOC3*	Apolipoprotein A5-C3	Plasma triglyceride regulation, lipoprotein and hepatic lipase inhibition	ApoB, blood metabolite levels, cholesterol esters, free cholesterol, HDL, IDL, LDL, LA, LpPLA$_2$ activity, mono unsaturated fatty acids, ω-3 fatty acids, ω-6 fatty acids, phosphocholines, phosphoglycerides, sphingomyelins, total cholesterol, triglycerides, vitamin E, VLDL	**CAD**, **CHD**, hypertriglyceridemia, metabolic syndrome	Peripheral blood, peripheral blood monocytes	Chol, CE, DG, PA, PC, SM, TG	105
14q23.2	*SGPP1*	Sphingosine-1-phosphate phosphatase 1	Intracellular and extracellular sphingosine-1-phosphate (S1P) regulation	Blood metabolite levels, sphingomyelins	N/A	N/A	SM	4
14q24.1	*PIGH-TMEM229B*	Phosphatidylinositol glycan anchor biosynthesis class H, transmembrane protein 229B	Glycosylphosphatidylinositol anchor biosynthesis	Phosphocholines	Breast cancer, Parkinson's disease	N/A	PA, PC, PG, SM	4
15q21.3	*LIPC*	Lipase C, hepatic type	Triglyceride hydrolysis, ligand/bridging factor for receptor-mediated lipoprotein uptake	ApoA1, blood metabolite levels, cholesterol esters, free cholesterol, DHA, HDL, IDL, LA, ω-3 fatty acid, ω-6 fatty acids, phosphoethanolamines, total cholesterol, triglycerides, VLDL	Advanced age-related macular degeneration, Alzheimer's disease, amyotrophic lateral sclerosis (ALS), schizophrenia, gout, ischaemic stroke, metabolic syndrome	N/A	PA, PC, PE	14
16p13.11, 16p13.12	*PLA2G10-NTAN1-NPIPA5*	Phospholipase A2 group X, N-terminal asparagine amidase, nuclear pore complex interacting protein family member A5	Glycerophospholipid hydrolysis	Dihomo linolenate, double bonds in fatty acids, LA, PUFA	N/A	Blood, lymphoblastoid cell lines, skin, transformed fibroblast cells	CE, PA, PC, PE	11
16q13	*CETP*	Cholesteryl ester transfer protein	Cholesteryl ester transfer	ApoA1, cholesterol esters, free cholesterol, HDL, LDL, phosphocholines, total cholesterol, triglycerides, VLDL	Advanced age-related macular degeneration, **CAD**	Peripheral blood, transformed fibroblast cells	PC	1
17q22	*PCTP*	Phosphatidylcholine transfer protein	Lipid binding, phosphatidylcholine transfer	N/A	N/A	Lymphoblastoid cell lines, peripheral blood	PA	1
19p13.2	*CERS4*	Ceramide synthase 4	Sphingosine *N*-acyltransferase activity	Ceramides, sphingomyelins	N/A	Adipose subcutaneous, esophagus muscularis, skeletal muscle, skin, tibial artery, transformed fibroblast cells	PC, SM	2

Table: Summary of known and novel information about conditional analysis loci (. . . continued)

Genetic locus				**Published associations with locus**			**Associations in PROMIS**	
Cytogenetic band	**Locus name**	**Full gene name(s)**	**Gene function(s)**	**Biomarkers**	**Diseases**	**Gene expression tissues**	**Lipid subclasses**	**No. lipid species**
19q13.32	*APOE-APOC1-APOC2-APOC4*	Apolipoprotein E-C1-C2-C4	Protein homodimerization activity, receptor binding, cholesteryl ester transfer protein inhibition, phosphatidylcholine binding, phospholipase inhibitor activity, lipoprotein lipase activation, lipid transporter activity	Cholesterol esters, free cholesterol, LDL, phospholipids, total cholesterol, triglycerides, VLDL	Advanced age-related macular degeneration, Alzheimer's disease, **CAD**, carotid intima media thickness, cognitive decline, T2D	Lymphoblastoid cell lines	PC, SM	8
19q13.42	*MBOAT7*	Membrane bound *O*-acyltransferase domain containing 7	Lysophospholipid acyltransferase activity	Blood metabolite levels, phosphoinositols	Alcohol-related cirrhosis	Adipose subcutaneous, adipose visceral omentum, adrenal gland, aorta artery, breast mammary tissue, coronary artery, esophagus mucosa, esophagus muscularis, heart atrial appendage, heart left ventricle, sigmoid colon, skeletal muscle, spleen, testis, thyroid, transformed fibroblast cells, tibial artery, tibial nerve, whole blood	PA, PC, PE, PI, PS	22
20p12.1	*SPTLC3*	Serine palmitoyltransferase long chain base subunit 3	Pyridoxal phosphate binding, transaminase activity	Blood metabolite levels, LDL, phosphocholines, sphingomyelins	N/A	N/A	Cer, PC, SM	27
22q12.1	*XBP1*	X-box binding protein 1	Transcription factor activity, sequence-specific DNA binding	N/A	Esophageal squamous cell carcinoma, pancreatic cancer	N/A	SM	1
22q12.2	*GAL3ST1*	Galactose-3-*O*-sulfotransferase 1	Sulfotransferase activity, galactosylceramide sulfotransferase activity	N/A	N/A	Peripheral blood monocytes	PG	1
22q13.31	*PNPLA3*	Patatin-like phospholipase domain containing 3	Phospholipase A_2 activity, mono-olein transacylation activity	N/A	Alcohol-related cirrhosis, NAFLD, metabolic disease, obesity	N/A	TG	2

The list of associations of biomarkers, diseases, and tissues from gene expression data in this table is based on lookups performed in PhenoScanner[217] and the mGWAS database that was compiled from a literature review of published metabolomics GWAS studies, but has also been supplemented by further literature searches for each locus. Nevertheless, the associations listed in this table are undoubtedly still incomplete and are meant to provide a representative overview of the type of associations that exist for the significant variants in each locus from the conditional analyses, rather than providing a comprehensive list of all known associations with all variants in these loci (which can be downloaded from PhenoScanner itself). In particular, certain traits that are not directly lipid-related, such as body mass index, height, and waist circumference, for instance, have been excluded to conserve space in the table. Associations with CAD and CHD are highlighted in bold. **Abbreviations: CAD** = Coronary artery disease; **CHD** = Coronary heart disease; **CRP** = C-reactive protein; **DHA** = Docosahexaenoic acid; **EPA** = Eicosapentaenoic acid; **HDL** = High-density lipoprotein cholesterol; **IDL** = Intermediate-density lipoprotein cholesterol; **LA** = Linoleic acid; **LDL** = Low-density lipoprotein cholesterol; **NAFLD** = Non-alcoholic fatty liver disease; **PUFA** = Polyunsaturated fatty acid; **T2D** = Type 2 diabetes; **VLDL** = Very low density lipoprotein cholesterol.

6.3.2 Novel loci

As described in Chapter 5, there were 89 sentinel variants that belonged to 23 independent loci. The next step was to identify novel loci that had not previously been reported in the GLGC for circulating lipids or published in previous metabolomics GWAS papers at time of writing (Table 1.3). When using a simple distance-based measure of ±500-Kb to define novel loci—that is, the variant had to be more than 500-Kb from a previously published variant to be considered novel—two novel loci were identified: *XBP1* and *GAL3ST1*. However, this approach was probably overly conservative since 1-Mb around each published variant is a rather wide region, so there were likely to be other novel variants that were not picked up by this method.

The second approach that was used to identify novel loci was based on simply determining if the variants matched any previously published variants or any proxies of those variants. PhenoScanner[217] was used to search for proxies of published variants. Using this approach, six novel loci were identified: *UGT8*, *PCTP*, *MBOAT7*, *XBP1*, *GAL3ST1*, and *PNPLA3*. However, this approach was likely too relaxed since variants could have been within the same gene region and only a few base-pairs away from a previously published variant or proxy, but still be considered novel if they were not in LD.

Therefore, the third approach tried to strike a balance between the two methods described above by incorporating information about both LD and distance to determine the size of the window around each published SNP. Again, the variants from the conditional analyses were compared to the variable windows around each published SNP to identify novel SNPs. Since PhenoScanner does not have any LD information on proxy SNPs more than 500-Kb in either direction from the specified variant or proxies with an $R^2 < 0.6$, this meant that all of the resulting windows upstream and downstream from the specified variants were now less than or equal to 500-Kb. However, as mentioned above, the distance-based approach may have been overly conservative and the proxy method was too relaxed, so defining the windows using the LD-based distance method was likely to be the most rigorous and accurate. Using this approach, there were four novel loci: *UGT8*, *XBP1*, *GAL3ST1*, and *PNPLA3*. While results from all three approaches have been presented here for comparability, all of the figures and tables in this chapter that display the number of novel variants are based on the latter approach.

As noted in subsection 5.3.4, rs738409, a missense variant in the *PNPLA3* locus which was already known to be associated with non-alcoholic fatty liver disease, was found in an exome-wide association analysis to also be significantly associated with circulating levels of triglycerides[40], and it was therefore included as one of the 175 major lipid loci in Figure 5.10. However, despite this published association with triglycerides, *PNPLA3* was considered to be a novel locus in this dissertation because it was not significantly associated with either major lipids or metabolites in the GWAS databases that were searched, namely the GLGC and literature review of mGWAS studies. The lead variant in PROMIS was rs12484809, which although it is different, has an LD r^2 of 0.738 with the previously published variant.

Table 6.2 correctly identifies *UGT8*, *XBP1*, *GAL3ST1*, and *PNPLA3* as novel loci since they were not associated with any lipid-related biomarkers. However, the table also indicates that *PCTP* is not known to be associated with any lipid-related biomarkers. In fact, this is correct, but the reason that the significant variant from the conditional analyses in the *PCTP* locus was not considered to be novel is because this variant (rs1107917, chr17:53487664) is just 23-Kb from a published SNP (chr17:533511321) from an mGWAS study[81] that is significantly associated with tryptophan. However, given that there was no LD information available for the published variant at time of writing, and tryptophan is not closely related to lipid metabolism, it is possible that *PCTP* should also be considered among the novel loci discovered in this dissertation. Interestingly, although the putative function of *PCTP* is catalysing the transfer of phosphatidylcholines between membranes, in PROMIS a phosphatic acid, rather than a phosphatidylcholine, was significantly associated with a variant in this locus.

A summary of the novel variants in these regions is shown in Table 6.3. The association of each lipid with the variants within each novel locus are shown in regional association plots that were produced using LocusZoom[201] (Figure 6.2). The plots have been annotated to distinguish between coding and non-coding variants, and the LD information that is shown between the sentinel variant and nearby variants within the region was calculated based on the PROMIS data itself rather than importing LD information from an external database such as 1000 Genomes[141].

Table 6.3: Association of lipids with loci not previously known to be associated with major lipids or metabolites

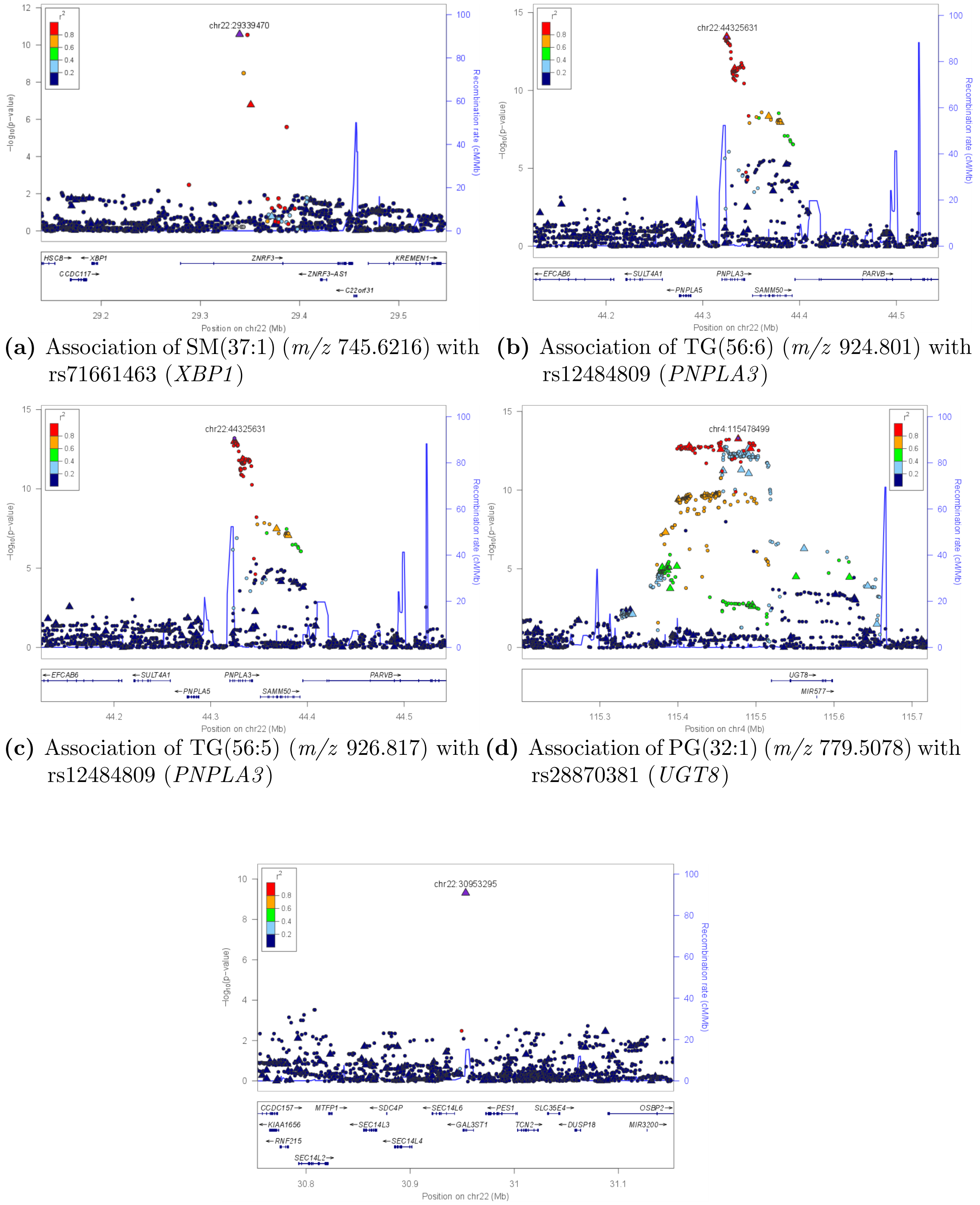
Lipid name (*m/z*)	**rsid (GRCh37 Chr:Pos)**	**EA / NEA**	**MAF (GWAS1 & GWAS2)**	**Info score (GWAS1 & GWAS2)**	***β***	**SE**	***P*-value**	**SNP consequence**	**Predicted causal gene**
SM(37:1) (745.6216)	rs71661463 (chr22:29339470)	T / TTCTC	0.01580 0.01272	0.8786 0.8646	−0.2298	0.0345	2.67×10^{-11}	Intronic	*XBP1*
PG(32:1)+AcO$^-$ (779.5078)	rs28870381 (chr4:115478499)	T / G	0.28302 0.28946	0.9749 0.9974	0.1316	0.0174	4.60×10^{-14}	Intergenic	*UGT8*
PG(32:1)+AcO$^-$ (779.5078)	rs2267161 (chr22:30953295)	T / C	0.21539 0.22002	1.0000 1.0000	−0.1197	0.0192	4.86×10^{-10}	Missense	*GAL3ST1*
TG(56:6) (924.801)	rs12484809 (chr22:44325631)	T / C	0.19835 0.20544	0.9854 0.9956	0.0700	0.0092	3.37×10^{-14}	Intronic	*PNPLA3*
TG(56:5) (926.817)	rs12484809 (chr22:44325631)	T / C	0.19835 0.20544	0.9854 0.9956	0.0657	0.0087	5.85×10^{-14}	Intronic	*PNPLA3*

For a description of how predicted causal genes were determined, see subsection 6.2.1. **Abbreviations: EA** = Effect allele; **GRCh37** = Genome Reference Consortium human genome build 37; **MAF** = Minor allele frequency; ***m/z*** = Mass-to-charge ratio; **NEA** = Non-effect allele; **SE** = Standard error; **SNP** = Single nucleotide polymorphism.

The four novel loci identified from this study will be described. First, patatin-like phospholipase domain containing protein 3 (*PNPLA3*) is a multifunctional enzyme that encodes a triacylglycerol lipase, which mediates triacylglycerol hydrolysis in adipocytes, and also has acylglycerol *O*-acyltransferase activity. A recent meta-analysis provided strong and unequivocal evidence for a significant role of rs738409, a nonsynonymous variant (p.Ile148Met) in the *PNPLA3* gene, in the progression of alcohol-related liver disease (ALD)[220]. Isoleucine to methionine substitution at position 148 in the *PNPLA3* gene (I148M) is a loss-of-function allele that has been shown to impair triglyceride hydrolysis in the liver and secretion of triglyceride-rich very low density lipoproteins, leading to altered fatty acid composition of liver triglycerides. However, the mechanisms by which this *PNPLA3* variant and other variants in the *TM6SF2* and *MBOAT7* genes confer risk for ALD, and the nature of the functional interplay between them, has not yet been determined[221]. Paradoxically, although this *PNPLA3* variant is associated with increased risk of liver disease, it is also associated with reduced risk of CHD. This suggests that targeting hepatic pathways to reduce cardiovascular risk may be complex, despite the clustering of cardiovascular and hepatic diseases in people with metabolic syndrome. The results from the conditional analysis in this dissertation found associations of two triglycerides—TG(56:6) (m/z 924.801) and TG(56:5) (m/z 926.817)—with rs12484809 in the *PNPLA3* locus, a variant in LD with rs738409 ($r^2 = 0.738$) that has not been previously reported. Therefore, this newly discovered association with a variant in *PNPLA3* could play an important role in ALD, non-alcoholic steatohepatitis (NASH), or non-alcoholic fatty liver disease (NAFLD), but a more thorough understanding of the pathways involved are needed.

Further investigation of the rs738409 (I148M) variant was also undertaken to study the associations of lipids with *PNPLA3* in greater detail. The loss-of-function variant was focused on since more is known about its role, and therefore it is more likely to have clinical applications than the novel variant that was identified. As shown in Figure 6.3, the *PNPLA3* I148M allele was associated with increased levels of lipids of higher carbon number and double-bond content, and consistently, with decreased levels of lipids of lower carbon number and double-bond content. This allele is associated with reduced risk of CHD but increased risk of non-alcoholic fatty liver disease (NAFLD). There were also significant differences between the mean level of lipids between individuals according to whether they were homozygous for the major allele, heterozygous, or homozygous for the minor allele (Figure 6.4).

Figure 6.2: Regional association plots for association of each lipid with novel loci



(a) Association of SM(37:1) (m/z 745.6216) with rs71661463 (*XBP1*)

(b) Association of TG(56:6) (m/z 924.801) with rs12484809 (*PNPLA3*)

(c) Association of TG(56:5) (m/z 926.817) with rs12484809 (*PNPLA3*)

(d) Association of PG(32:1) (m/z 779.5078) with rs28870381 (*UGT8*)

(e) Association of PG(32:1) (m/z 779.5078) with rs2267161 (*GAL3ST1*)

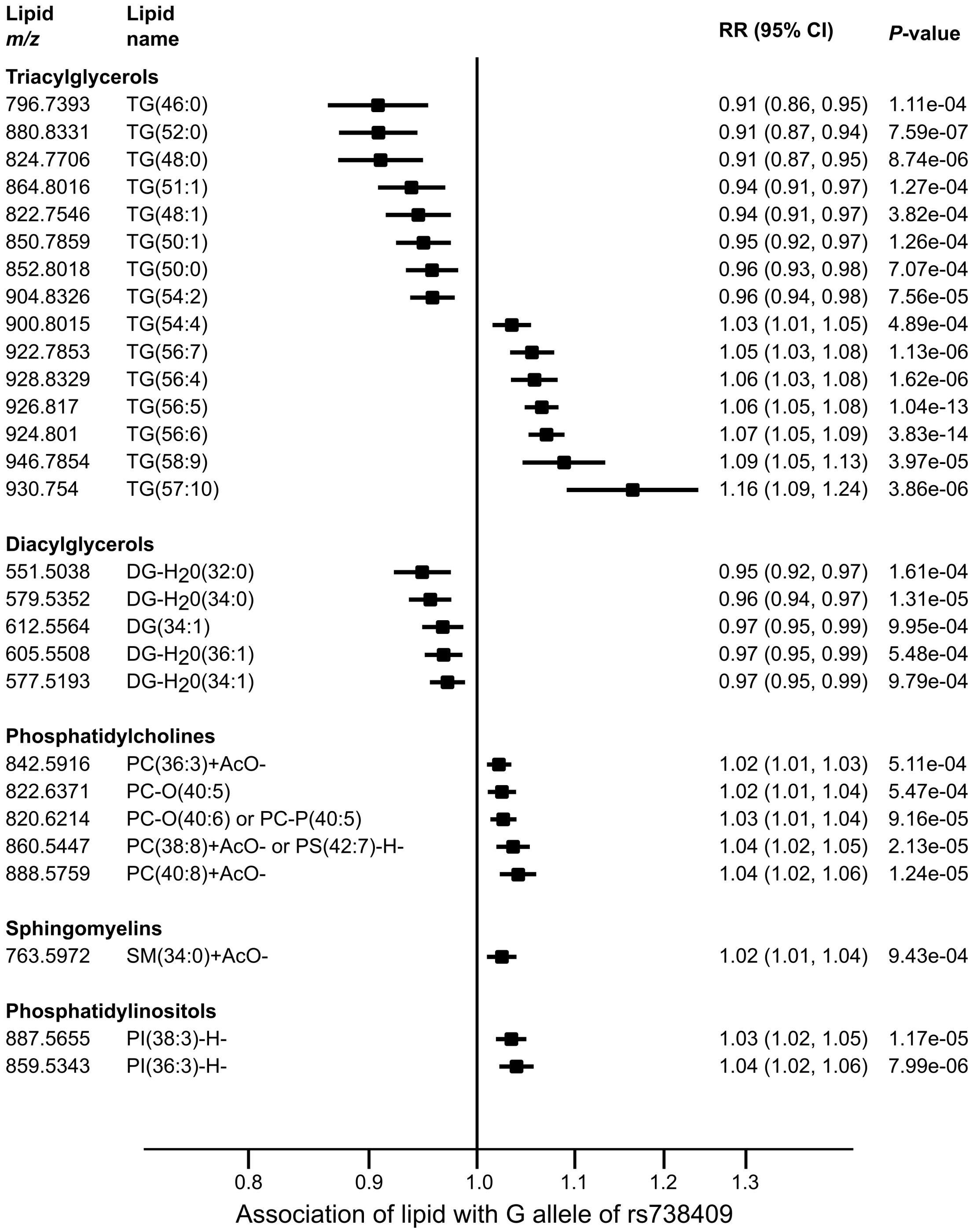
The association of each lipid with the variants within the loci not previously reported to be associated with major lipids or metabolites is shown in a regional association plot produced using LocusZoom[201]. Coding variants are indicated with a filled triangle, while non-coding variants are indicated with a filled circle. The coding variant rs2267161 in the *GAL3ST1* region is the only sentinel variant that was directly genotyped; the other sentinel variants were imputed. Linkage disequilibrium (LD) was calculated based on the correlation (r^2) of each SNP, and is indicated by the colour of the fill of each plotted variant. The lead SNP within each locus is shown in purple and labelled with the chromosomal position.

The second novel locus, galactose-3-*O*-sulfotransferase 1 (*GAL3ST1*), catalyses the sulfation of membrane glycolipids and the synthesis of galactosylceramide sulfate (sulfatide), a major lipid component of the myelin sheath. It also acts on lactosylceramide, galactosyl 1-alkyl-2-*sn*-glycerol, and galactosyl diacylglycerol. The conditional analysis results found that rs2267161, a missense variant in the *GAL3ST1* locus, is associated with PG(32:1) (*m/z* 779.5078). Variations for this same SNP have been linked to reduced insulin resistance and increased risk of T2D[222], but it has not been previously reported in any associations with circulating lipids or mGWAS, so it is an interesting finding from this study that the variant is also associated with a phosphatidylglycerol. Sulfonation is an important step in the metabolism of many drugs, so it is possible that this novel variant could be a useful pathway for new drugs targeting insulin resistance through this metabolic pathway, although further research in this area is needed.

The third locus, uracil-diphosphate glucose (UDP) glycosyltransferase 8 (*UGT8*) catalyses the transfer of galactose to ceramide, a key enzymatic step in the biosynthesis of galactocerebrosides, which are abundant sphingolipids of the myelin membrane of the central nervous system and peripheral nervous system. The conditional analysis results found that rs28870381, an intergenic variant in the *UGT8* locus, was associated with PG(32:1) (*m/z* 779.5078). Again, there has not been much research on *UGT8*, but it is surprising to find an association with a phosphatidylglycerol since one would expect that variants in *UGT8* would be associated with ceramides. This is another area that would be useful for follow-up research.

Finally, X-box binding protein 1 (*XBP1*) functions as a transcription factor during endoplasmic reticulum stress by regulating the unfolded protein response. It also functions as a major regulator of the unfolded protein response in obesity-induced insulin resistance and type 2 diabetes for the management of obesity and diabetes prevention, and recent studies have shown that compounds targeting the *XBP1* pathway are a potential approach for the treatment of metabolic diseases[223]. In addition, *XBP1* protein expression, which is induced in the liver by a high carbohydrate diet, is directly involved in fatty acid synthesis through *de novo* lipogenesis, so compounds that inhibit *XBP1* activation may also be useful for treatment of NAFLD[224]. The conditional analysis results found that rs71661463, an intronic variant in the *XBP1* locus, was associated with SM(37:1) (*m/z* 745.6216).

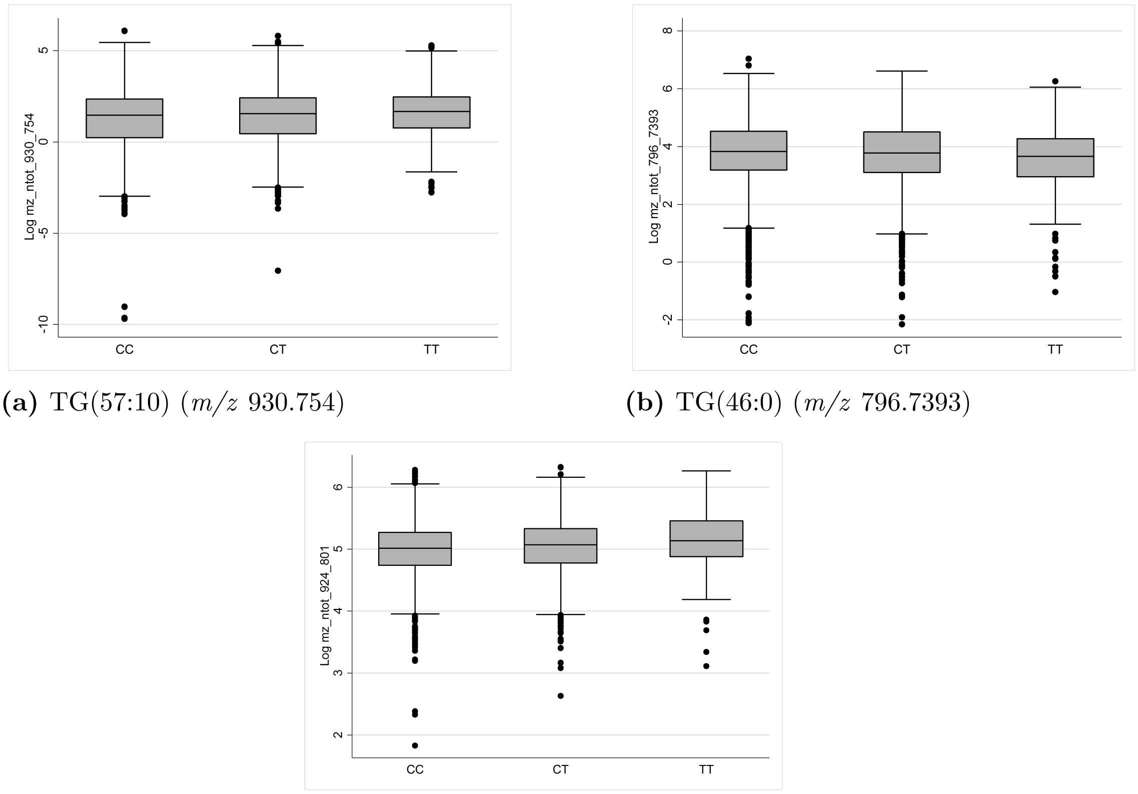
Figure 6.3: Association of lipids with rs738409 variant in *PNPLA3*



Lipid *m/z*	Lipid name	RR (95% CI)	*P*-value
Triacylglycerols			
796.7393	TG(46:0)	0.91 (0.86, 0.95)	1.11e-04
880.8331	TG(52:0)	0.91 (0.87, 0.94)	7.59e-07
824.7706	TG(48:0)	0.91 (0.87, 0.95)	8.74e-06
864.8016	TG(51:1)	0.94 (0.91, 0.97)	1.27e-04
822.7546	TG(48:1)	0.94 (0.91, 0.97)	3.82e-04
850.7859	TG(50:1)	0.95 (0.92, 0.97)	1.26e-04
852.8018	TG(50:0)	0.96 (0.93, 0.98)	7.07e-04
904.8326	TG(54:2)	0.96 (0.94, 0.98)	7.56e-05
900.8015	TG(54:4)	1.03 (1.01, 1.05)	4.89e-04
922.7853	TG(56:7)	1.05 (1.03, 1.08)	1.13e-06
928.8329	TG(56:4)	1.06 (1.03, 1.08)	1.62e-06
926.817	TG(56:5)	1.06 (1.05, 1.08)	1.04e-13
924.801	TG(56:6)	1.07 (1.05, 1.09)	3.83e-14
946.7854	TG(58:9)	1.09 (1.05, 1.13)	3.97e-05
930.754	TG(57:10)	1.16 (1.09, 1.24)	3.86e-06
Diacylglycerols			
551.5038	DG-H_20(32:0)	0.95 (0.92, 0.97)	1.61e-04
579.5352	DG-H_20(34:0)	0.96 (0.94, 0.97)	1.31e-05
612.5564	DG(34:1)	0.97 (0.95, 0.99)	9.95e-04
605.5508	DG-H_20(36:1)	0.97 (0.95, 0.99)	5.48e-04
577.5193	DG-H_20(34:1)	0.97 (0.95, 0.99)	9.79e-04
Phosphatidylcholines			
842.5916	PC(36:3)+AcO-	1.02 (1.01, 1.03)	5.11e-04
822.6371	PC-O(40:5)	1.02 (1.01, 1.04)	5.47e-04
820.6214	PC-O(40:6) or PC-P(40:5)	1.03 (1.01, 1.04)	9.16e-05
860.5447	PC(38:8)+AcO- or PS(42:7)-H-	1.04 (1.02, 1.05)	2.13e-05
888.5759	PC(40:8)+AcO-	1.04 (1.02, 1.06)	1.24e-05
Sphingomyelins			
763.5972	SM(34:0)+AcO-	1.02 (1.01, 1.04)	9.43e-04
Phosphatidylinositols			
887.5655	PI(38:3)-H-	1.03 (1.02, 1.05)	1.17e-05
859.5343	PI(36:3)-H-	1.04 (1.02, 1.06)	7.99e-06

Association of *G* allele of rs738409 in *PNPLA3* locus with levels of various lipids. The black lines denote 95 % confidence intervals.

Figure 6.4: Box plots displaying levels of three triglycerides in individuals grouped by their genotypes at the variant rs738409 in *PNPLA3*



(a) TG(57:10) (*m/z* 930.754)

(b) TG(46:0) (*m/z* 796.7393)

(c) TG(56:6) (*m/z* 924.801)

ANOVA test of difference in mean levels of triglycerides by genotype: (a) TG(57:10): P = 0.013; (b) TG(46:0): P < 0.001; and (c) TG(56:6): P < 0.001.

6.3.3 Novel relationships and new biological insights into lipid metabolism

This analysis replicated and confirmed known associations between lipids and genetic loci while also further extending what is known about these loci, thus providing new biological insights into lipid metabolism for known loci. For instance, the associations of lipids with apolipoprotein A5 (*APOA5*) were described in Chapter 5, and the associations of those lipids with CHD risk factors were presented in Chapter 4. Additionally, membrane bound *O*-acyltransferase domain containing 7 (*MBOAT7*), which contributes to the regulation of free arachidonic acid in the cell through the remodelling of phospholipids, was reported as being associated with the metabolite 1-arachidonoylglycerophosphoinositol [i.e. PI(36:4)] in a previous mGWAS[132], but this analysis found that the lead SNP in this locus, rs8736 (chr19:54677189), was associated with a wide range of phosphatic acids [e.g. PA(40:5) and PA(44:6)], phosphocholines [e.g. PC(36:6) and PC(42:11)], phosphoethanolamines [e.g. PE(39:7)], and phosphoinositols [e.g. PI(34:1) and PI(36:1)] (see Figure 5.9l).

Another significant example where this analysis extends the knowledge base of lipid metabolism is hepatic type lipase C (*LIPC*). *LIPC* is an important enzyme in HDL metabolism that has the capacity to catalyse hydrolysis of phospholipids, mono-, di-, and triglycerides, and acyl-CoA thioesters. *LIPC* has been reported as being associated with several different metabolites, including linoleic acid, docosahexaenoic acid (DHA), 1-linoleoylglycerophosphoethanolamine, and others[78,132]; however, this analysis reports additional associations with an even wider range of metabolites, including PA(39:1), PC(35:4), PE(36:4), PE(36:5), and PE(38:6) (see Figure 5.9j).

Apolipoprotein E-C1-C2-C4 (*APOE-C1-C2-C4*) is a well-studied locus that has been reported for association with cholesterol esters, free cholesterol, LDL-C, phospholipids, total cholesterol, triglycerides, and VLDL-C[35,78]. However, the previous studies that looked at this locus were based on an NMR metabolomics platform, which differentiates lipids based on particle size[78], and the association with major circulating lipids[35], which does not examine individual species of lipid metabolites. In comparison, this analysis found associations with six specific sphingomyelins [SM(34:0), SM(40:0), SM(40:1), SM(40:2), SM(42:0)+AcO$^-$, and SM(42:1)] and two phosphocholines [PC-O(37:1) and PC-O(39:1)] (see Figure 5.9c), which are defined based on the number of carbon atoms and double bonds rather than the particle size. Thus, these results contribute valuable information to the body of existing knowledge regarding this locus.

Sphingosine-1-phosphate phosphatase 1 (*SGPP1*) has been previously been reported for association with overall blood metabolite levels[132] and sphinomyelins in particular[225]. The analyses in this dissertation found associations of *SGPP1* with four additional sphingomyelins [SM(31:1)-H$^-$, SM(32:1), SM(32:1)+AcO$^-$, and SM(39:1)] that were not discovered in those studies (see Figure 5.9t).

Likewise, serine palmitoyltransferase long chain base subunit 3 (*SPTLC3*) has previously been reported for association with blood metabolite levels[132], LDL-C[35], phosphocholines[73], and sphingomyelins[73]. This analysis also found associations with three phosphocholines and fifteen sphingomyelins, but different ones that have not been previously reported [e.g. PC-P(38:1), SM(31:1)-H$^-$, SM(32:1), and SM(37:1)]. Additionally, this analysis found significant associations with nine ceramides [Cer(40:0)-H$^-$, Cer(40:1)-H$^-$, Cer(40:2)-H$^-$, Cer(41:0)-H$^-$, Cer(41:1)-H$^-$, Cer(41:2)-H$^-$, Cer(42:0)-H$^-$, Cer(42:1)-H$^-$, and Cer(42:2)-H$^-$] which have not ever been reported in relation to this locus (see Figure 5.9u).

6.3.4 Network diagrams

Genetic associations with lipid metabolite concentrations were summarised within each lipid subclass and combined with partial correlations from the GGM to produce a network of associations between genes and lipid subclasses, which is shown in Figure 6.5. The network diagram facilitates visualisation of the genetic determinants of human metabolism and the relationships between genetic loci and lipid subclasses.

The network shows the connections between the various lipid subclasses and their association with genetic loci. For example, it is evident that diglycerides and triglycerides had strong over-representation in the GGM, which means that there were more connections between diglycerides and triglycerides than would have been expected due to chance alone, whereas sphingomyelins and triglycerides had strong under-representation in the GGM, which means that there were fewer connections between lipid species in these subclasses than would have been expected due to chance alone.

Another observation that can be drawn from the network diagram is that sphingomyelins are associated with two loci that are also associated with a range of other lipid subclasses, namely *SPTLC3* and *FADS1-2-3*, but they are also associated with exclusively with four loci that are not associated with any other lipid subclasses: *GCKR*, *SGPP1*, *MLXIPL*, and *XBP1*. Sphingomyelins have previously been shown to be associated with *SGPP1*, but the associations of sphingomeylins with the other three loci have not ever been reported prior to this analysis.

A second network diagram was also generated for a subset of the triglyceride species

(Figure 6.6) showing the partial correlations of individual triglycerides and the detailed associations between triglycerides with genetic loci. A number of observations can also be drawn from this second diagram. First, the network shows that variants in the *APOA5-APOC3* locus, which is in the centre of the figure, are associated with a wide range of triglycerides. Apolipoprotein A-V (ApoA5) is an apolipoprotein encoded by the *APOA5* gene that regulates levels of circulating triglycerides, but the mechanisms as to how it does this are unclear. It is thought that either ApoA5 regulates the catabolism of triglyceride-rich lipoprotein particles by *LPL*, or that ApoA5 plays a role in the assembly of VLDL particles, or perhaps both play a role[205,210,226–228]. Figure 6.6 mainly shows links with triglyerides containing polyunsaturated fatty acids but no direct links with completely saturated triglycerides, suggesting that variants in the *APOA5-APOC3* locus mainly affect the catabolism of triglyceride-rich lipoproteins and not so much the assembly of VLDL particles in the liver, where we would expect an association with saturated triglycerides that are produced in the liver.

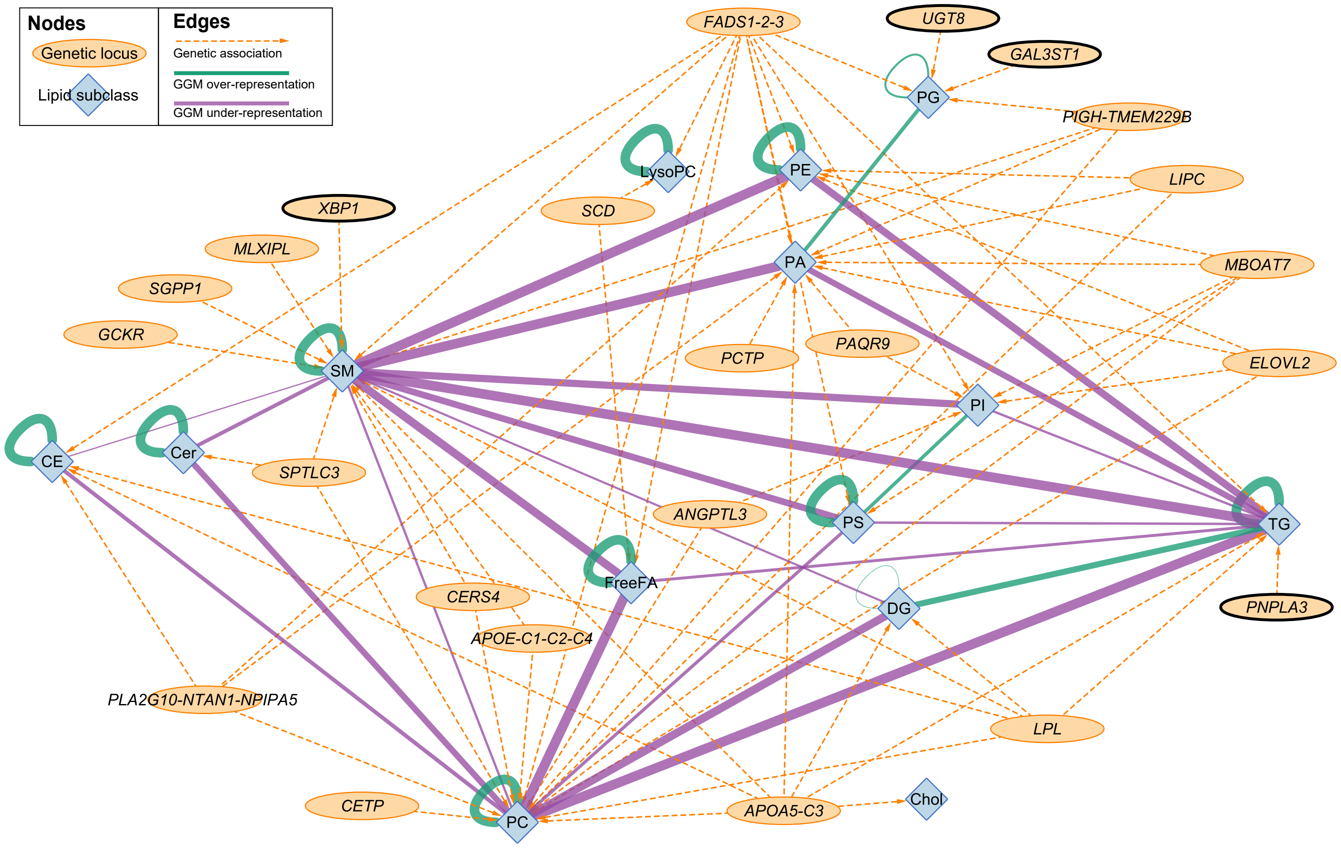
Second, fatty acid desaturase is key in the production of polyunsaturated fatty acids, so differences in the activity of *FADS1-2-3* will most clearly be seen in triglycerides with a large number of double bonds that also have a large number of carbon atoms. That is why this diagram shows that *FADS1-2-3* is only linked to TG(56:6), TG(56:7), and TG(58:9).

Third, it is unclear why variants in the *PNPLA3* locus also have the strongest associations with some of the same largest triglycerides, namely TG(56:5) and TG(56:6). Presumably the variants in the *PNPLA3* locus that have significant effects are changing the substrate specificity so that there is a shift in the relative amounts of triglycerides that are exported from the liver.

Fourth, as described in Chapter 5, *LPL* is mainly active on monounsaturated fatty acids in triglyceride species, which is what can be seen in Figure 6.6. *LPL* links to the triglycerides TG(52:2), TG(52:3), TG(53:2), and TG(53:3), which have a high probability of containing one or more mono-unsaturated fatty acids within their fatty acid side chains.

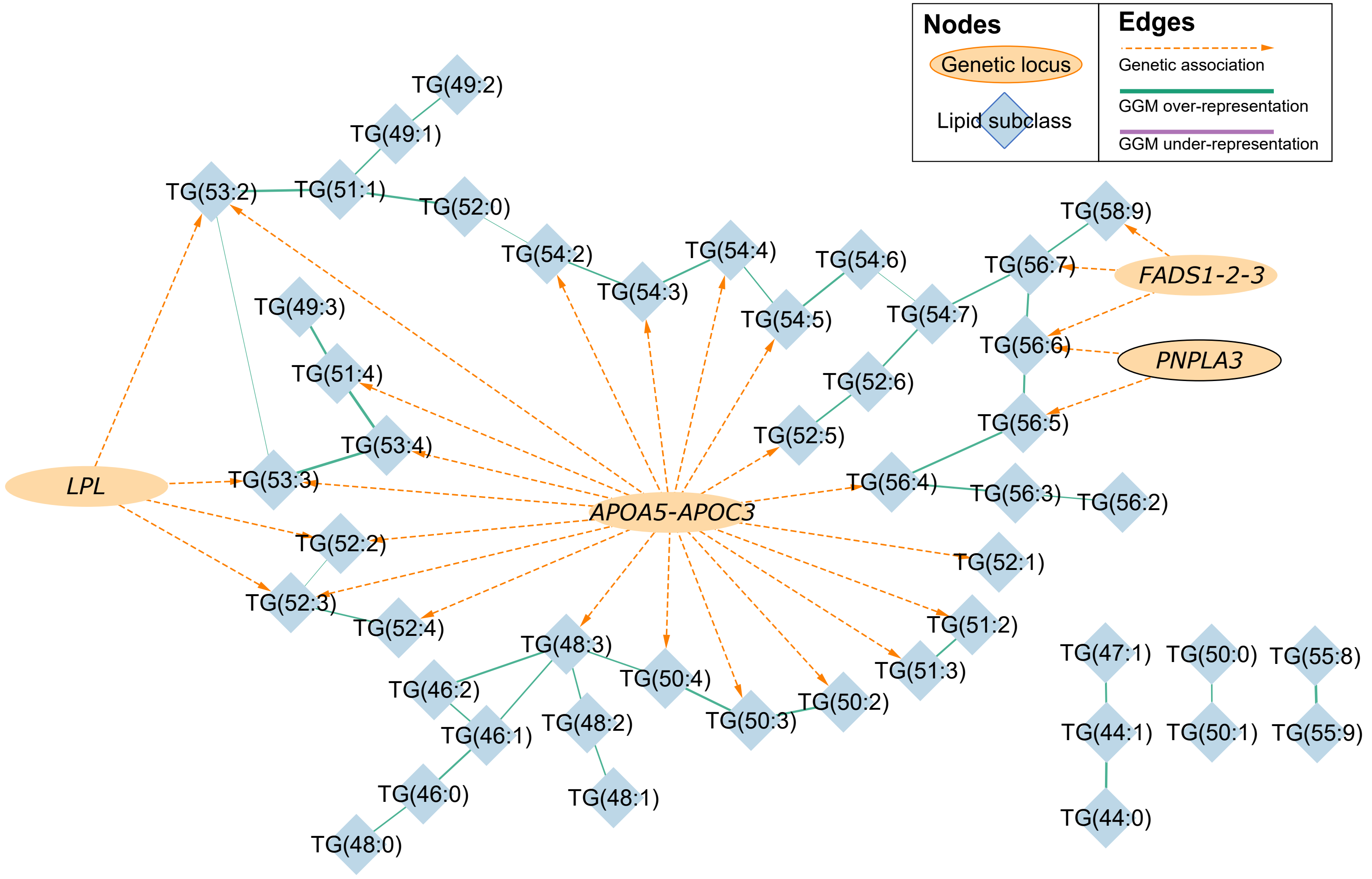
A final observation is that completely saturated triglycerides such as TG(44:0), TG(46:0), and TG(50:0) do not show any evidence of direct genetic associations, which suggests that the main driver of these lipids is *de novo* lipogenesis and that this process is independent of these genes.

Figure 6.5: Combined network view of genetic and lipid metabolite associations



Nodes representing genetic loci are labelled with the most likely causal gene at that locus according to the functional annotation approach (see subsection 6.2.1). In order for an edge to be drawn between a genetic locus and a lipid subclass, there must have been a minimum of one variant at that locus significantly ($P < 8.929 \times 10^{-10}$) associated with a minimum of one lipid species belonging to the lipid subclass. Edges between lipid subclasses indicate whether there was either a significant over-representation (green edges) or under-representation (red edges) of GGM connections between lipid species belonging to different lipid subclasses (see subsection 6.2.3), with the magnitude indicated by the thickness of the edges. Likewise, “self-loops” indicate either under- or over-representation of connections between lipid species belonging to a single lipid subclass. Loci not previously known to be associated with major lipids or metabolites are indicated with a black border around the oval. **Abbreviations: CE** = Cholesteryl Esters; **Cer** = Ceramides; **Chol** = Cholesterol and derivatives; **DG** = Diacylglycerols; **FreeFA** = Free Fatty Acids; **LysoPC** = Lysophosphatidylcholines; **PA** = Phosphatic Acids; **PC** = Phosphatidlycholines; **PE** = Phosphatidylethanolamines; **PG** = Phosphatidylglycerols; **PI** = Phosphatidylinositols; **PS** = Phosphatidylserines; **SM** = Sphingomyelins; **TG** = Triacylglycerols.

Figure 6.6: Combined network view of genetic and lipid metabolite associations for individual triglycerides



Nodes representing genetic loci are labelled with the most likely causal gene at that locus according to the functional annotation approach (see subsection 6.2.1). In order for an edge to be drawn between a genetic locus and a triglyceride, there must have been a minimum of one variant at that locus significantly ($P < 8.929 \times 10^{-10}$) associated with a minimum of one lipid species belonging to the lipid subclass. Edges between triglycerides indicate whether there was either a significant over-representation (green edges) or under-representation (purple edges) of GGM connections between lipid species belonging to different lipid subclasses (see subsection 6.2.3), with the magnitude indicated by the thickness of the edges. Loci not previously known to be associated with major lipids or metabolites are indicated with a black border around the oval.

6.4 Discussion

In this chapter, a functional annotation pipeline was used to annotate the 89 significant variants from the conditional analyses and identify the probable causal genes for each of the 23 significant loci. Detailed information about the consequence of each variant on the protein sequence and the predicted causal gene were determined by integrating information from the bottom-up and top-down approaches together with expert curation.

As described in subsection 6.2.2, three different approaches were utilised to identify novel variants. The first only considered variants more than 500-Kb from published variants to be novel, the second approach only considered published variants and their proxies to be novel, and the third method, which was the preferred approach for this dissertation, defined the size of the region around each published variant based on available LD information, therefore considering variants outside this variable window size to be novel. As would be expected, the significant variant from the conditional analyses in the *PCTP* locus was not considered to be novel using the third method since it was only 23-Kb from a published variant. However, no LD information was available for the published variant, and the trait that it is associated with, tryptophan, is an amino acid that is used in biosynthesis of proteins, which is not very closely related to lipid metabolites or circulating major lipids, despite being a metabolite that was published in an mGWAS[81]. Therefore, although it was thought that the first approach was too conservative in identifying novel variants and the second approach was too liberal, while the third struck a good balance, it turns out that the second approach was actually more reliable in this instance. This demonstrates how challenging it can be to accurately determine what is truly novel. The definition of novelty can vary tremendously depending on which criteria are used.

Another limitation of the approach used to identify novel loci is that the only database of major circulating lipids that was used to compare against the PROMIS results was the GLGC. While 157 loci were identified by the GLGC[35], subsequent studies have identified 175 major lipid loci at the time of writing (listed in Table 1.1), so including additional studies would have been more comprehensive. This is why *PNPLA3* was reported as a novel locus in this dissertation, even though it had been reported for association with levels of circulating triglycerides in another study of major lipids that was not included in the comparison[40]. Likewise, while 31 published mGWAS were identified at time of writing based on a previous review[74], with additional studies identified through further scanning of the literature (listed in Table 1.3), a comprehensive systematic review may have been able to identify additional studies.

A genome-wide threshold ($P < 5 \times 10^{-8}$) was used to identify previously published variants and determine whether these variants were successfully replicated in this study. Use of a lower threshold would enable identification of additional variants that variants at known loci that successfully replicated, but on the other hand, since additional published results would be included that did not reach genome-wide significance, this would result in potentially failing to identify loci in this study that were truly novel.

Although several limitations have been noted, the overall approach to identify novel loci was rigorous and robust. Based on the literature review of published GWAS of metabolomics and circulating lipids, four novel loci not previously reported for association with major lipids or metabolites were discovered (*XBP1*, *PNPLA3*, *UGT8*, and *GAL3ST1*). However, based on the above findings, while not consistent with the analysis plan that was followed and hence why this change was not formally adopted, an argument could be made for including *PCTP* amongst the novel loci identified by this dissertation while discarding the novel discovery in the *PNPLA3* locus.

The in-depth analyses of the GWAS results presented in this chapter help enhance understanding of lipid biology and genetic associations. This study found that a wide range of glycerophospholipids are associated with variants in the *MBOAT7* locus. This is the first analysis to report associations of these lipids with this locus. Additionally, while *LIPC* is known to be associated with several metabolites, this analysis found additional associations with other lipids (phosphatic acids, phosphocholines, and phosphoethanolamines) that have not been previously reported. Novel associations were also identified for sphingomyelins and phosphocholines with variants in the *APOE-C1-C2-C4* region, and four additional sphingoymelins were identified that are associated with variants in the *SGPP1* locus, which has not been previously reported. Furthermore, this analysis identified phosphocholines, sphingomyelins, and ceramides that have not been previously reported for association with variants in the *SPTLC3* locus. Finally, as noted in the previous chapter, this is the first analysis to show a link between *LPL* activity and sphingomyelins.

The network diagrams shown in this chapter incorporated information from a GGM of lipid metabolites and genetic associations with lipids to provide visual representations of the genetic determinants of human metabolism. This resulted in a number of interesting findings. Diglycerides and triglycerides had strong over-representation in the GGM, indicating that there were more connections between diglycerides and triglycerides than would be expected due to chance alone, whereas sphingomyelins and triglycerides had strong under-representation in the GGM, indicating that there were fewer connections between lipid species in these subclasses than would be expected due to chance alone. The network analysis also showed links with triglyerides containing polyunsaturated fatty acids but no direct links with completely saturated triglycerides, suggesting that variants in the *APOA5-APOC3* locus mainly affect the catabolism of triglyceride-rich lipoproteins but have less of an impact on the assembly of VLDL particles in the liver. Another finding from the network analysis is that variants in the *PNPLA3* locus had the strongest associations with some of the same largest triglycerides [i.e. TG(56:5) and TG(56:6)]. This may have been due to the effects of these variants on changing the substrate specificity so that there is a shift in the relative amounts of triglycerides that are exported from the liver.

Follow-up efforts are currently under way to investigate the association of lipid metabolites with the *PNPLA3* locus in greater detail. A proposal for a recall-by-genotype study has been submitted and successfully approved to recall 60 (initially) healthy volunteers registered in the NIHR Cambridge BioResource based on their *PNPLA3* genotype (i.e. 30 volunteers per homozygous group) matched for sex (50 % men), age (30–40 years), and BMI (25–32 kg/m^2). Recruitment for this study is expected to start soon.

As part of the recall-by-genotype study, the BMI, blood pressure, heart rate, height, weight, medical history, lifestyle factors, and demographics of each individual will be assessed. A baseline blood sample will be taken on the first day, the participants will consume an energy-balanced dinner of standard macronutrient composition at 7pm, and drink deuterium-labelled water (3 g/kg body water) at 8pm and 10pm (i.e. 2-hour gap between the loading doses) divided into two portions of equal sizes. The participants will be told that they should not exercise, have any further meals, or consume alcohol after dinner.

The participants will have an overnight fast, and then a fasting blood sample will be taken at 8am (i.e. 12 hours after the first loading dose of deuterated water) to measure *de novo* lipogenesis by gas chromatography mass spectrometry (GC-MS), lipid profiles using direct infusion high-resolution mass spectrometry (DIHRMS) and size exclusion chromatography coupled to electrospray ionisation mass spectrometry (SEC-ESI-MS), and other biochemical parameters. The participants will then take a maintenance dose of deuterium-labelled water (0.04 g/kg body water) at 8:15am and eat breakfast at 8:30am (approximately 70 % fructose co-ingested with glucose [CHO], 10 % fat; 30 % overfeeding). A postprandial blood sample will then be taken after 3.5 hours (at 12:00pm) to again measure *de novo* lipogenesis by GC-MS, lipid profiles using DIHRMS and SEC-ESI-MS, and other biochemical parameters.

The use of deuterium incorporation from enriched drinking water allows for the quantification of palmitate, synthesised via *de novo* lipogenesis within blood plasma triglycerides during high-carbohydrate feeding. Thus, the data would implicate triglycerides involved in *de novo* lipogenesis most likely produced in the liver (rather than adipose tissue). Without *in vitro* or *in vivo* experiments, it would be difficult to establish the exact molecular mechanism or function of *PNPLA3*. However, the newly generated DIRHMS data that this study will measure would confirm and improve the qualitative lipid measurements in the PROMIS cohort.

Overall, the genetic analyses in this chapter identified several novel relationships between genetic variants and lipids and also revealed numerous new biological insights into lipid metabolism for existing loci. These findings strengthen and expand the knowledge base for understanding the genetic determinants of lipid metabolites and their association with metabolic disease-related loci, and highlight useful areas for follow-up studies to identify possible therapeutic targets. In the next chapter, the causal relevance of these lipids for risk of coronary heart disease will be investigated.

CHAPTER 7

Mendelian randomisation study:

Causal effect of lipid metabolites on risk of coronary heart disease

Chapter summary

This chapter provides an overview of Mendelian randomisation (MR) and its use for determining the causal effect of perturbations in levels of a risk factor on a disease outcome. First, a literature review is presented summarising the available evidence for a causal association between modifications in levels of major lipids and risk of coronary heart disease (CHD). Next, although the application of MR to metabolomics is still an emerging field, the few studies that have been published thus far assessing the causal relevance of metabolites for risk of CHD are described. Finally, as a natural succession to previous chapters that described the association of lipid metabolites with CHD risk factors and genetic variants, MR is employed in this chapter to investigate the causal paths of lipid metabolites in relation to risk of CHD. A comprehensive MR analysis strategy was developed and tested to assess several different research questions, namely the causal effect on CHD risk of levels of (1) individual lipids, (2) lipids grouped according to their corresponding subclasses, and (3) weighted linear combinations of lipids. To address pleiotropy, which occurs when a genetic variant is associated with multiple risk factors and could lead to a violation of the instrumental variable assumptions, the most highly pleiotropic variants were pruned from the set of variants used as instrumental variables and the causal estimates were

compared before and after pruning. Additionally, multiple MR approaches were employed, in particular the inverse-variance weighted, MR-Egger, and weighted median methods, and the results were compared. The use of several different methods to address each research question helped to provide robustness of the causal estimates.

There were eighteen lipids with evidence of a causal effect. However, there was extensive pleiotropy for which the methods employed were unable to fully account, and the majority of the lipids did not have a sufficiently large number of signficantly associated variants that could be used as instrumental variables. Therefore, the precise individual lipid species that have a causal effect on CHD could not be determined, although broad generalisations about the nature of the causal pathways could be observed from the association of lipids within particular subclasses. Fatty acids, sphingolipids, and sterol lipids appeared to have a protective effect on risk of CHD, while increased levels of glycerolipids were associated with increased risk of CHD. Elevated levels of glycerophospholipids with fewer double bonds predominantly increased the risk of CHD, while higher levels of glycerophospholipids with more double bonds had a protective effect on CHD.

7.1 Introduction

Mendelian randomisation (MR), as described in Chapter 1, is the use of genetic variants as instrumental variables to determine whether increased or decreased exposure to a risk factor is causally associated with a disease outcome[85,86]. While MR has been widely used in assessing whether major lipids have a causal effect on CHD, the use of MR in metabolomics is much less common. However, since mGWAS have identified a large number of loci that are associated with numerous metabolites, this presents an ideal opportunity to examine whether any of these metabolites have causal effects on CHD or its risk factors.

Good candidates for MR analyses are metabolites where variations in their concentrations are associated with loci in or near genes encoding metabolic enzymes or carriers, which are also linked to CHD and related metabolic diseases. A few examples of loci that meet these criteria are *APOA5-APOC3*, *FADS1-2-3*, *GCKR*, *LPL*, and *LIPC*. For example, SNP rs1260326 in the *GCKR* loci is known to lower fasting glucose and triglyceride levels and reduce the risk of T2D, and this loci is also associated with phosphatidylcholine ratios[73]. Therefore, an MR study could be conducted to determine whether variations in levels of phosphatidylcholine ratios have a causal effect on T2D. Additionally, SNP rs964184 in the apolipoprotein cluster *APOA5-APOC3* is strongly associated with blood triglyceride levels and is a well-known CHD loci, but is also associated with ratios between different phosphatidylcholines, which are biochemically related to triglycerides by the intermediary of only a few enzymatic reaction steps[73].

7.2 Methods

A two-sample MR analysis was conducted using summarised statistics from genetic associations with CHD risk in over 60 000 cases and 125 000 controls from the CARDIoGRAMplusC4D consortium[43]. MR-Egger and the weighted median method were used to provide some robustness against pleiotropic variants.

The overall objective of the MR analysis was to address three principal research questions concerning the causal effect of lipid metabolites (considered either individually or in combination) with CHD risk. An analysis plan was developed prior to conducting the analyses that would seek to address these research questions, which is summarised in Figure 7.1. The first question was whether levels of individual lipids have a causal effect on CHD risk. This was assessed using a crude approach, which included all significantly associated variants as instrumental variables, and a pruned approach, which omitted the most highly pleiotropic variants. The pruned approach was performed using three different methodological approaches (which were described in more detail in Section 1.5): (1) the inverse-variance weighted (IVW) method, which is a standard regression with inverse-variance weights and the intercept term set to zero; (2) the MR-Egger method, which is a standard regression with inverse-variance weights and the intercept term estimated; and (3) the weighted median method, which determines the median of the causal estimates based on the individual candidate instruments, using inverse-variance weights so that more precise estimates receive more weight in the analysis[229]. As part of the analysis plan for the first research question, it was also decided that a multivariable approach would be used, which would still consider each lipid individually but allow for variants to be associated with multiple lipids.

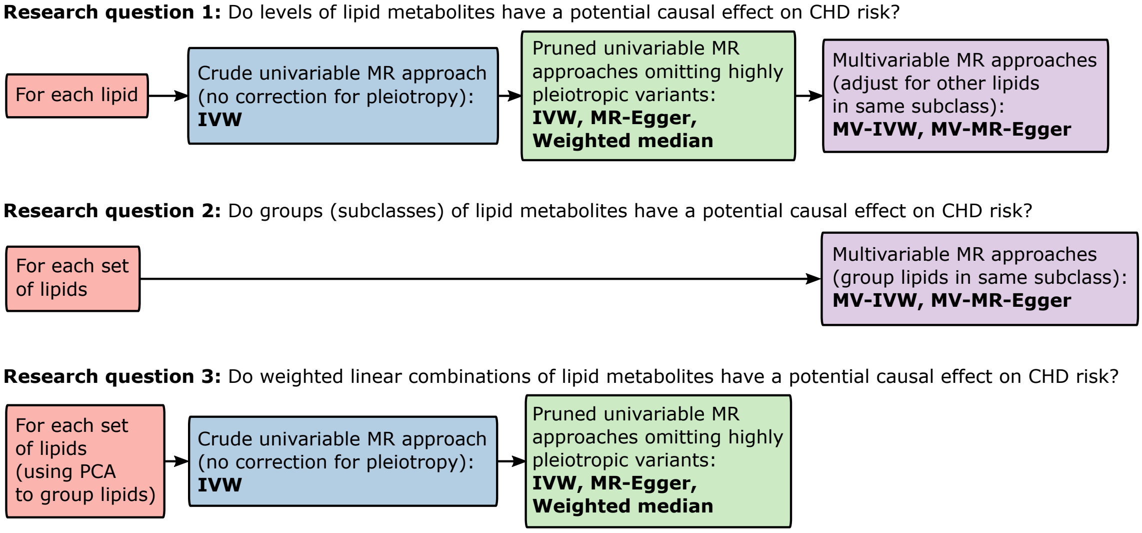
The second research question was whether subclasses of lipids have a causal effect on risk of CHD. The analysis plan for this question involved combining the lipid subclasses and using a multivariable MR approach to obtain effect estimates for each lipid within each of the subclasses.

Finally, the third research question assessed whether weighted linear combinations of correlated lipids, which accounted for most of the variance in the levels of the lipids, have a causal effect on risk of CHD. In order to address this question, it was decided that principal component analysis (PCA) would be used to reduce the number of lipids to a smaller number of principal components that account for the majority of the variance in the levels of the lipids, and determine whether these principal components are causally related to risk of CHD.

7.2.1 Causal effect of lipids on risk of CHD

The first research question examined whether levels of individual lipids have a causal effect on CHD risk. A crude univariable approach was initially applied to provide a “first pass” in determining whether each lipid, when considered on an individual basis, was causally related to risk of CHD while ignoring correlations between lipids and not accounting for pleiotropy. This approach was conducted for each lipid by selecting all lead variants from the conditional analyses that were associated with the lipid at the Bonferroni-corrected P-value of 8.929×10^{-10}, then calculating the MR estimate using the IVW method. This method assumes all SNPs are valid instrumental variables, which is equivalent to a gene score when variants are uncorrelated. An adaptation of this method was used to correct for correlations between variants. The association of each variant with CHD was looked up

Figure 7.1: Flow chart of planned MR analysis approach


Research question 1: Do levels of lipid metabolites have a potential causal effect on CHD risk?
For each lipid
Crude univariable MR approach (no correction for pleiotropy): IVW
Pruned univariable MR approaches omitting highly pleiotropic variants: IVW, MR-Egger, Weighted median
Multivariable MR approaches (adjust for other lipids in same subclass): MV-IVW, MV-MR-Egger
Research question 2: Do groups (subclasses) of lipid metabolites have a potential causal effect on CHD risk?
For each set of lipids
Multivariable MR approaches (group lipids in same subclass): MV-IVW, MV-MR-Egger
Research question 3: Do weighted linear combinations of lipid metabolites have a potential causal effect on CHD risk?
For each set of lipids (using PCA to group lipids)
Crude univariable MR approach (no correction for pleiotropy): IVW
Pruned univariable MR approaches omitting highly pleiotropic variants: IVW, MR-Egger, Weighted median

Abbreviations: CHD = Coronary heart disease; **IVW** = Inverse-variance weighted; **MR** = Mendelian randomisation; **MV** = Multivariable; **PCA** = Principal component analysis.

from the 2015 version of CARDIoGRAMplusC4D, using proxy variants (correlated with the lead variant with an LD $r^2 > 0.6$) if the lead variant was not available. Lookups of genetic associations with the outcome and identification of proxy variants were performed using PhenoScanner[217].

A pruned univariable MR approach was then used in order to progress the lipids that were significant in the crude approach, accounting for pleiotropy by omitting highly pleiotropic variants. Multiple SNPs in the *APOA5-APOC3* and *FADS1-2-3* regions were associated with more than 100 lipids, whereas SNPs in all other regions were associated with fewer than 30 lipids. Since these two regions were not likely to be informative for MR of specific lipids due to extensive pleiotropy, and were likely to drive significant results for a large number of lipids, they were omitted from the analysis to prioritise the discovery of novel associations with disease risk. For this approach, the IVW method was used for all lipids, and the MR-Egger and weighted median method were also used for lipids associated with more than two variants in the conditional analyses.

Given that lipids within the same subclass have similar biological structures and may affect similar biological pathways, it was planned that a multivariable MR approach would be used to determine whether each lipid, when taking into account the effect of other lipids within the same subclass, is causally related to risk of CHD. For each lipid subclass, all lead variants from the conditional analysis would be selected that were associated with any of the lipids within the subclass at the Bonferroni-corrected P-value of 8.929×10^{-10}, and MR would be run using a multivariable analogue of the IVW method.

7.2.2 Causal effect of lipid subclasses on risk of CHD

A separate set of MR analyses were also developed as part of the analysis plan to address the second research question, whether subclasses of lipids have a causal effect on risk of CHD. It was planned that a univariable MR approach would be used to consider whether the entire lipid subclass as a whole, rather than the individual lipids assessed in a univariable or multivariable analysis, have a causal effect on risk of CHD.

In the first instance, a GWAS was run on each of the five overall lipid categories before assessing the fourteen lipid subclasses. The signals for each of the individual lipids that make up each overall lipid category were aggregated by adding them together, then a rank-based inverse normal transformation was performed. Histograms and Q-Q plots of each lipid category were examined to ensure approximate normality. A univariate GWAS was then performed on each lipid category with adjustment for the same set of variables as was used in the GWAS of the individual lipids (see Chapter 5). MR could then be conducted using the same approach as was used for the individual lipids.

7.2.3 Causal effect of linear combinations of lipids on risk of CHD

The third research question assessed whether linear combinations of correlated lipids, which account for most of the variance in levels of the lipids, have a causal effect on risk of CHD. Given that most of the lipids are highly correlated, this analysis aimed to reduce the number of lipids to a smaller number of principal components that account for most of the variance in the levels of the lipids, and to determine whether these principal components are causally related to risk of CHD. The rationale for this approach is that for highly correlated lipids, the causal effect estimates may be imprecise for both lipids if both are included, so dimension reduction through PCA is an effective way to address this. PCA was performed on all lipids and the number of components were selected that explain at least 95 % of the variance. A GWAS was run on each principal component (as described in Chapter 5), and it was planned that MR would be run using a multivariable analogue of the IVW method, where the risk factors would be each of the principal components, and the variants identified from the GWAS that were associated with any of the principal components would be used as instrumental variables.

7.3 Results

The MR analyses described in this chapter were conducted to assess whether changes in lipid concentrations have a causal effect on risk of CHD. The causal effect of lipids was assessed for each lipid individually, for subgroups of lipids by subclass, and for principal components of lipids.

7.3.1 Causal effect of lipids on risk of CHD

The univariable MR analysis identified eighteen lipids with evidence of a causal effect on CHD. A summary of the most significantly associated lipids is shown in Table 7.1. The results for both the crude and pruned approach are presented in the table, but only the causal effect estimates that were obtained using the ratio (if only one variant was used as an instrumental variable) or IVW method (if two or more variants were used) are shown. The more sophisticated approaches, namely the MR-Egger and weighted median methods, were also used to obtain causal estimates, but these methods could only be employed when there were three or more variants used as instrumental variables. The majority of the lipids were not eligible for these approaches because they were only significantly associated with one or two independent genetic variants that were also associated with CHD.

Likewise, the multivariable MR approach was not able to be conducted due to the lack of sufficient variants to use as instrumental variables. Multivariable MR requires at least one more variant used as instrumental variables than there are risk factors (i.e. if n is the number of lipids, then the number of variants used as instrumental variables must be at least $n + 1$). Since most lipid metabolites were only associated with a few genetic variants and there was significant overlap of associations due to pleiotropy, the lipidomics dataset did not meet the minimum requirements needed to run multivariable MR.

Figure 7.2 shows representative scatter plots of the β estimates for each SNP that was used as an instrumental variable for two of the lipids with significant causal effect estimates, PE(36:4)-H$^-$ (m/z 738.5079) and PE(38:6)-H$^-$ (m/z 762.5079). The figures show the β estimate for the effect of the SNP on the lipid on the x-axis, and the β estimate for the effect of the SNP on CHD on the y-axis. Confidence intervals for the β estimates are shown around each plotted SNP in both the x and y directions. Lines are shown based on the intercept and slope for each MR method that was used, namely the crude and pruned approach (which are the same for these two lipids because no SNPs were removed from the pruned approach), the MR-Egger method, and the weighted median approach.

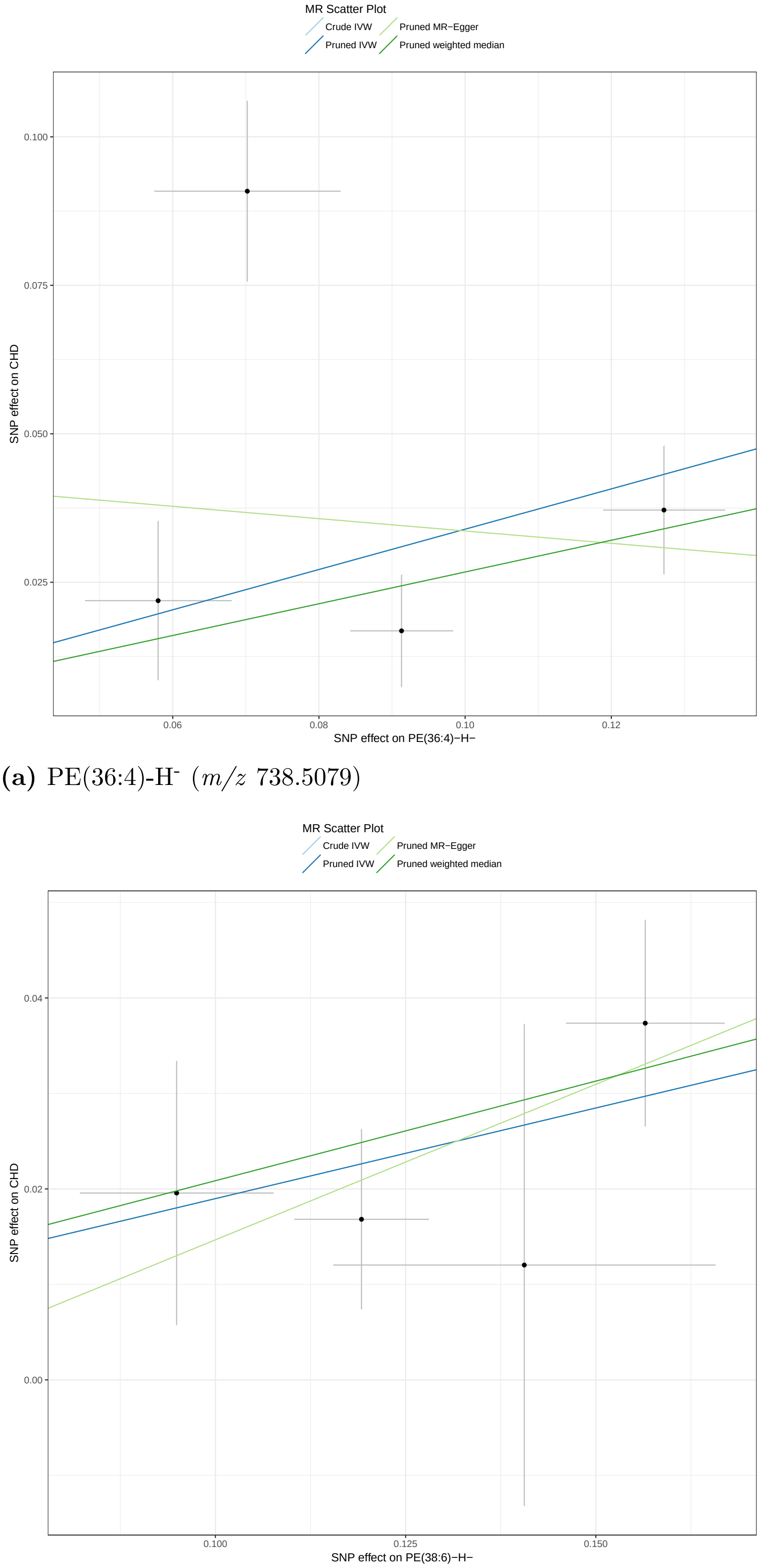
The causal effect of the lipids on CHD, grouped according to the five overall lipid categories, is shown in Figure 7.3. Results using both the crude and pruned approaches are presented. Loci are underlined if the SNPs in that locus were used as instrumental variables in both the crude and pruned approaches, whereas loci that are not underlined were only used in the crude approach. It is readily apparent that a large number of lipids have a causal effect on CHD through the same pathways. For instance, nearly all of the sphingomyelins have a protective effect for CHD through the *LPL* gene (Figure 7.3a), while diglycerides and triglycerides have a significant positive effect on CHD through the *LPL* and *APOA5-APOC3* genes (Figure 7.3b). Phosphocholines and phosphoethanolamines, on the other hand, have a significant positive causal effect on CHD through the *LIPC* pathway (Figure 7.3c). Unfortunately, due to extensive pleiotropy, it was not possible to identify which lipids are actually causal for CHD, since it could be any of the lipids with a significant causal effect, or perhaps multiple lipids working together to influence the pathway. Therefore, the MR analysis did not yield anything particularly informative as to the causal effect of individual lipid metabolites on risk of CHD.

A notable finding that emerged from this analysis is that TG(56:5), which is one of the lipids that was significantly associated with the novel variant in the *PNPLA3* gene (as described in subsection 6.3.2), had a significant protective causal effect for CHD using the pruned approach (Figure 7.3b), which would suggest that individuals with elevated levels of this triglyceride are at reduced risk of CHD. However, since this association was only just barely significant, this finding should not be overemphasised.

7.3.2 Causal effect of lipid subclasses on risk of CHD

The next stage of the analysis plan was to address the second research question, which aimed to determine whether lipid subclasses have a causal effect on risk of CHD. The GWAS results of the overall lipid categories revealed that fatty acyls and glycerophospholipids did not have any associations that reached genome-wide significance (Figure 7.4). Glycerolipids, sphingolipids, and sterol lipids were associated with variants in the *LPL* and *APOA5-APOC3* loci. Additionally, sphingolipids were also associated with variants in the *ANGPTL3*, *MLXIPL*, and *APOE-C1-C2-C4* regions, and sterol lipids were also associated with variants in the *APOE-C1-C2-C4* locus. Regrettably, these associations for the overall lipid categories were also found for the individual lipids. Due to the extensive pleiotropy that was found for many of the metabolites (particularly with the *APOA5-APOC3* and *LPL* regions, which were associated with most of the lipid classes), there was little ad-

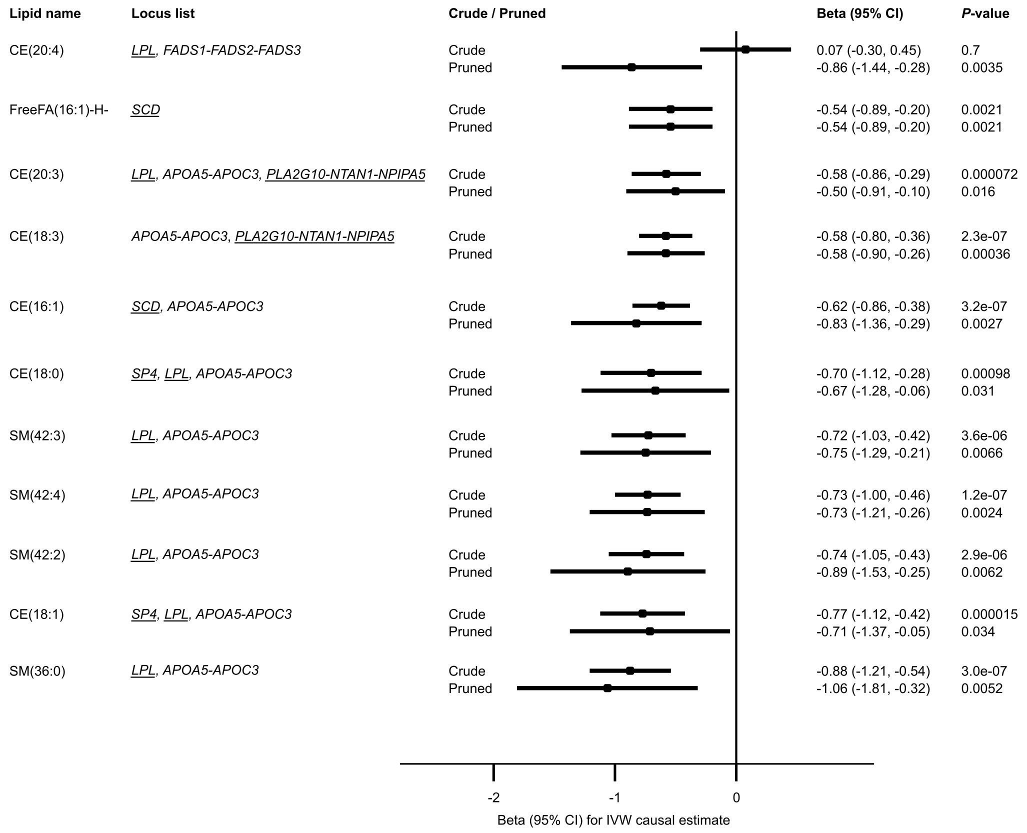
Figure 7.2: Scatter plot of MR results for the association of two lipids with CHD



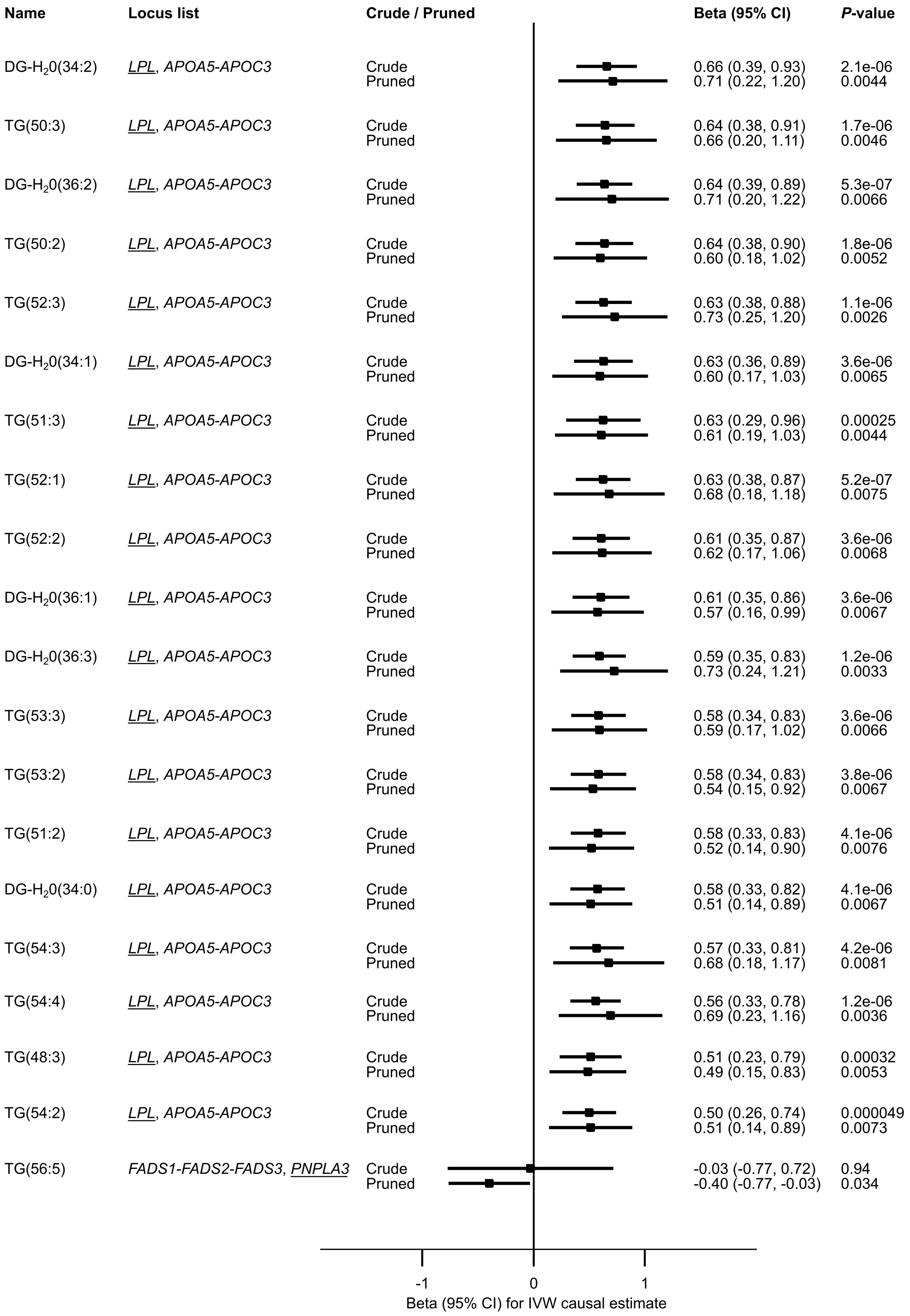
(a) PE(36:4)-H$^-$ (m/z 738.5079)

(b) PE(38:6)-H$^-$ (m/z 762.5079)

Figure 7.3: Results of MR analysis to assess causal effect of lipids on risk of CHD

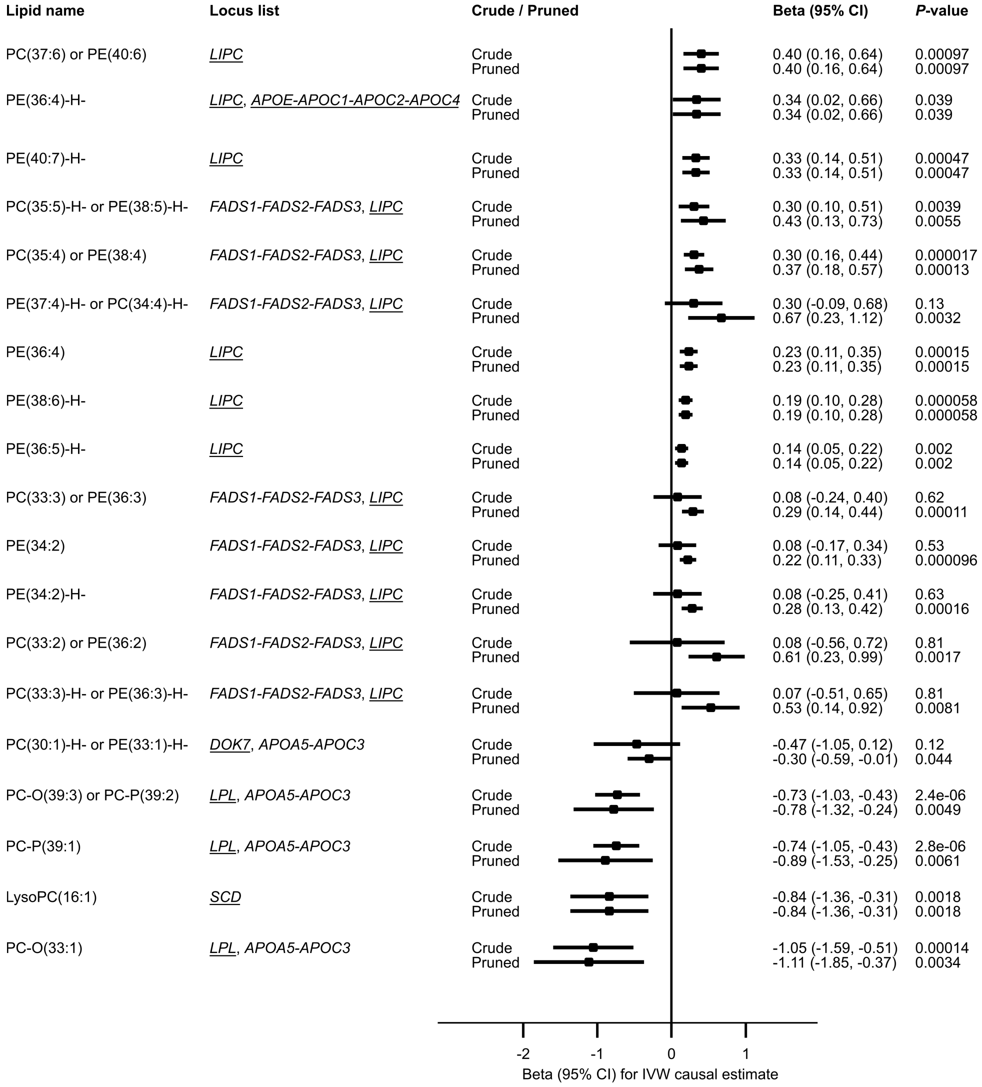


(a) Fatty acids, sphingolipids, and sterol lipids



Name	Locus list	Crude / Pruned	Beta (95% CI)	*P*-value
DG-H_20(34:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.66 (0.39, 0.93) 0.71 (0.22, 1.20)	2.1e-06 0.0044
TG(50:3)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.64 (0.38, 0.91) 0.66 (0.20, 1.11)	1.7e-06 0.0046
DG-H_20(36:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.64 (0.39, 0.89) 0.71 (0.20, 1.22)	5.3e-07 0.0066
TG(50:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.64 (0.38, 0.90) 0.60 (0.18, 1.02)	1.8e-06 0.0052
TG(52:3)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.63 (0.38, 0.88) 0.73 (0.25, 1.20)	1.1e-06 0.0026
DG-H_20(34:1)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.63 (0.36, 0.89) 0.60 (0.17, 1.03)	3.6e-06 0.0065
TG(51:3)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.63 (0.29, 0.96) 0.61 (0.19, 1.03)	0.00025 0.0044
TG(52:1)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.63 (0.38, 0.87) 0.68 (0.18, 1.18)	5.2e-07 0.0075
TG(52:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.61 (0.35, 0.87) 0.62 (0.17, 1.06)	3.6e-06 0.0068
DG-H_20(36:1)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.61 (0.35, 0.86) 0.57 (0.16, 0.99)	3.6e-06 0.0067
DG-H_20(36:3)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.59 (0.35, 0.83) 0.73 (0.24, 1.21)	1.2e-06 0.0033
TG(53:3)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.58 (0.34, 0.83) 0.59 (0.17, 1.02)	3.6e-06 0.0066
TG(53:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.58 (0.34, 0.83) 0.54 (0.15, 0.92)	3.8e-06 0.0067
TG(51:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.58 (0.33, 0.83) 0.52 (0.14, 0.90)	4.1e-06 0.0076
DG-H_20(34:0)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.58 (0.33, 0.82) 0.51 (0.14, 0.89)	4.1e-06 0.0067
TG(54:3)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.57 (0.33, 0.81) 0.68 (0.18, 1.17)	4.2e-06 0.0081
TG(54:4)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.56 (0.33, 0.78) 0.69 (0.23, 1.16)	1.2e-06 0.0036
TG(48:3)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.51 (0.23, 0.79) 0.49 (0.15, 0.83)	0.00032 0.0053
TG(54:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	0.50 (0.26, 0.74) 0.51 (0.14, 0.89)	0.000049 0.0073
TG(56:5)	*FADS1-FADS2-FADS3*, *PNPLA3*	Crude Pruned	-0.03 (-0.77, 0.72) -0.40 (-0.77, -0.03)	0.94 0.034

(b) Glycerolipids



Lipid name	Locus list	Crude / Pruned	Beta (95% CI)	*P*-value
PC(37:6) or PE(40:6)	*LIPC*	Crude Pruned	0.40 (0.16, 0.64) 0.40 (0.16, 0.64)	0.00097 0.00097
PE(36:4)-H-	*LIPC*, *APOE-APOC1-APOC2-APOC4*	Crude Pruned	0.34 (0.02, 0.66) 0.34 (0.02, 0.66)	0.039 0.039
PE(40:7)-H-	*LIPC*	Crude Pruned	0.33 (0.14, 0.51) 0.33 (0.14, 0.51)	0.00047 0.00047
PC(35:5)-H- or PE(38:5)-H-	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.30 (0.10, 0.51) 0.43 (0.13, 0.73)	0.0039 0.0055
PC(35:4) or PE(38:4)	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.30 (0.16, 0.44) 0.37 (0.18, 0.57)	0.000017 0.00013
PE(37:4)-H- or PC(34:4)-H-	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.30 (-0.09, 0.68) 0.67 (0.23, 1.12)	0.13 0.0032
PE(36:4)	*LIPC*	Crude Pruned	0.23 (0.11, 0.35) 0.23 (0.11, 0.35)	0.00015 0.00015
PE(38:6)-H-	*LIPC*	Crude Pruned	0.19 (0.10, 0.28) 0.19 (0.10, 0.28)	0.000058 0.000058
PE(36:5)-H-	*LIPC*	Crude Pruned	0.14 (0.05, 0.22) 0.14 (0.05, 0.22)	0.002 0.002
PC(33:3) or PE(36:3)	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.08 (-0.24, 0.40) 0.29 (0.14, 0.44)	0.62 0.00011
PE(34:2)	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.08 (-0.17, 0.34) 0.22 (0.11, 0.33)	0.53 0.000096
PE(34:2)-H-	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.08 (-0.25, 0.41) 0.28 (0.13, 0.42)	0.63 0.00016
PC(33:2) or PE(36:2)	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.08 (-0.56, 0.72) 0.61 (0.23, 0.99)	0.81 0.0017
PC(33:3)-H- or PE(36:3)-H-	*FADS1-FADS2-FADS3*, *LIPC*	Crude Pruned	0.07 (-0.51, 0.65) 0.53 (0.14, 0.92)	0.81 0.0081
PC(30:1)-H- or PE(33:1)-H-	*DOK7*, *APOA5-APOC3*	Crude Pruned	-0.47 (-1.05, 0.12) -0.30 (-0.59, -0.01)	0.12 0.044
PC-O(39:3) or PC-P(39:2)	*LPL*, *APOA5-APOC3*	Crude Pruned	-0.73 (-1.03, -0.43) -0.78 (-1.32, -0.24)	2.4e-06 0.0049
PC-P(39:1)	*LPL*, *APOA5-APOC3*	Crude Pruned	-0.74 (-1.05, -0.43) -0.89 (-1.53, -0.25)	2.8e-06 0.0061
LysoPC(16:1)	*SCD*	Crude Pruned	-0.84 (-1.36, -0.31) -0.84 (-1.36, -0.31)	0.0018 0.0018
PC-O(33:1)	*LPL*, *APOA5-APOC3*	Crude Pruned	-1.05 (-1.59, -0.51) -1.11 (-1.85, -0.37)	0.00014 0.0034

(c) Glycerophospholipids

The inverse-variance weighted (IVW) method was used to assess the causal effect of each lipid on coronary heart disease (CHD). The predicted causal locus is listed for each SNP used as an instrumental variable in the MR analysis. The SNPs were selected from the conditional analysis results following QC for which $P < 0.05$ for either the crude or the pruned approach using the IVW method. The crude approach included all variants, while the pruned approach excluded highly pleiotropic variants associated with 100 or more lipids (i.e. *FADS1-2-3* and *APOA5-APOC3*). SNPs belonging to loci with underlined names were used as instrumental variables in both the crude and pruned approaches to estimate the causal effect.

Table 7.1: Results of MR analysis to assess causal effect of lipid metabolites on risk of coronary heart disease

Lipid name	**Lipid** *m/z*	**Crude approach**				**Pruned approach**			
		Ratio/IVW results				**Ratio/IVW results**			
		No. variants	β	**SE**	***P*-value**	**No. variants**	β	**SE**	***P*-value**
FreeFA(16:1)-H$^-$	253.2174	1	-0.54	0.18	2.1×10^{-3}	1	-0.54	0.18	2.1×10^{-3}
LysoPC(16:1)	494.3245	1	-0.84	0.27	1.8×10^{-3}	1	-0.84	0.27	1.8×10^{-3}
DG-H20(34:2)	575.5039	2	0.66	0.14	2.1×10^{-6}	1	0.71	0.25	4.4×10^{-3}
DG-H20(34:1)	577.5193	2	0.63	0.14	3.6×10^{-6}	1	0.60	0.22	6.5×10^{-3}
DG-H20(34:0)	579.5352	2	0.58	0.13	4.1×10^{-6}	1	0.51	0.19	6.7×10^{-3}
DG-H20(36:3)	601.5195	2	0.59	0.12	1.2×10^{-6}	1	0.73	0.25	3.3×10^{-3}
DG-H20(36:2)	603.5352	2	0.64	0.13	5.3×10^{-7}	1	0.71	0.26	6.6×10^{-3}
DG-H20(36:1)	605.5508	2	0.61	0.13	3.6×10^{-6}	1	0.57	0.21	6.7×10^{-3}
CE(16:1)	640.6024	2	-0.62	0.12	3.2×10^{-7}	1	-0.83	0.27	2.7×10^{-3}
CE(18:3)	664.6026	3	-0.58	0.11	2.3×10^{-7}	2	-0.58	0.16	3.6×10^{-4}
CE(18:1)	668.6339	3	-0.77	0.18	1.5×10^{-5}	2	-0.71	0.34	3.4×10^{-2}
CE(18:0)	670.6496	3	-0.70	0.21	9.8×10^{-4}	2	-0.67	0.31	3.1×10^{-2}
CE(20:4)	690.6183	2	0.07	0.19	7.0×10^{-1}	1	-0.86	0.30	3.5×10^{-3}
CE(20:3)	692.6339	3	-0.58	0.15	7.2×10^{-5}	2	-0.50	0.21	1.6×10^{-2}
PC(30:1)-H$^-$ or PE(33:1)-H$^-$	702.5079	2	-0.47	0.30	1.2×10^{-1}	1	-0.30	0.15	4.4×10^{-2}
PE(34:2)-H$^-$	714.5079	3	0.08	0.17	6.3×10^{-1}	2	0.28	0.07	1.6×10^{-4}
Cer(44:11)+AcO$^-$	716.523	3	0.08	0.13	5.3×10^{-1}	2	0.22	0.06	9.6×10^{-5}
PC-O(33:1)	732.5904	2	-1.05	0.28	1.4×10^{-4}	1	-1.11	0.38	3.4×10^{-3}
SM(36:0)	733.6219	2	-0.88	0.17	3.0×10^{-7}	1	-1.06	0.38	5.2×10^{-3}
PE(36:5)-H$^-$	736.4922	2	0.14	0.04	2.0×10^{-3}	2	0.14	0.04	2.0×10^{-3}
PE(36:4)-H$^-$	738.5079	4	0.34	0.16	3.9×10^{-2}	4	0.34	0.16	3.9×10^{-2}
PE(36:4)	740.5229	2	0.23	0.06	1.5×10^{-4}	2	0.23	0.06	1.5×10^{-4}
PC(33:3)-H$^-$ or PE(36:3)-H$^-$	740.5236	3	0.07	0.29	8.1×10^{-1}	2	0.53	0.20	8.1×10^{-3}
PC(33:3) or PE(36:3)	742.5386	3	0.08	0.16	6.2×10^{-1}	2	0.29	0.07	1.1×10^{-4}
PC(33:2) or PE(36:2)	744.5543	3	0.08	0.33	8.1×10^{-1}	2	0.61	0.19	1.7×10^{-3}
PE(37:4)-H$^-$ or PC(34:4)-H$^-$	752.5235	2	0.30	0.20	1.3×10^{-1}	1	0.67	0.23	3.2×10^{-3}
PE(38:6)-H$^-$	762.5079	4	0.19	0.05	5.8×10^{-5}	4	0.19	0.05	5.8×10^{-5}
PC(35:5)-H$^-$ or PE(38:5)-H$^-$	764.5236	3	0.30	0.11	3.9×10^{-3}	2	0.43	0.15	5.5×10^{-3}
PC(35:4) or PE(38:4)	768.5543	3	0.30	0.07	1.7×10^{-5}	2	0.37	0.10	1.3×10^{-4}
PE(40:7)-H$^-$	788.5236	2	0.33	0.09	4.7×10^{-4}	2	0.33	0.09	4.7×10^{-4}
PC(37:6) or PE(40:6)	792.5537	2	0.40	0.12	9.7×10^{-4}	2	0.40	0.12	9.7×10^{-4}
SM(42:4)	809.6531	2	-0.73	0.14	1.2×10^{-7}	1	-0.73	0.24	2.4×10^{-3}
SM(42:3)	811.6688	2	-0.72	0.16	3.6×10^{-6}	1	-0.75	0.27	6.6×10^{-3}
PC-O(39:3) or PC-P(39:2)	812.6532	2	-0.73	0.15	2.4×10^{-6}	1	-0.78	0.28	4.9×10^{-3}
SM(42:2)	813.6844	2	-0.74	0.16	2.9×10^{-6}	1	-0.89	0.33	6.2×10^{-3}
PC-P(39:1)	814.6688	2	-0.74	0.16	2.8×10^{-6}	1	-0.89	0.32	6.1×10^{-3}
TG(48:3)	818.7236	2	0.51	0.14	3.2×10^{-4}	1	0.49	0.18	5.3×10^{-3}
TG(50:3)	846.7546	2	0.64	0.13	1.7×10^{-6}	1	0.66	0.23	4.6×10^{-3}
TG(50:2)	848.77	2	0.64	0.13	1.8×10^{-6}	1	0.60	0.22	5.2×10^{-3}
TG(51:3)	860.7703	2	0.63	0.17	2.5×10^{-4}	1	0.61	0.21	4.4×10^{-3}
TG(51:2)	862.7857	2	0.58	0.13	4.1×10^{-6}	1	0.52	0.20	7.6×10^{-3}
TG(52:3)	874.7859	2	0.63	0.13	1.1×10^{-6}	1	0.73	0.24	2.6×10^{-3}
TG(52:2)	876.8016	2	0.61	0.13	3.6×10^{-6}	1	0.62	0.23	6.8×10^{-3}
TG(52:1)	878.8172	2	0.63	0.12	5.2×10^{-7}	1	0.68	0.25	7.5×10^{-3}
TG(53:3)	888.8016	2	0.58	0.13	3.6×10^{-6}	1	0.59	0.22	6.6×10^{-3}
TG(53:2)	890.8172	2	0.58	0.13	3.8×10^{-6}	1	0.54	0.20	6.7×10^{-3}
TG(54:4)	900.8015	2	0.56	0.12	1.2×10^{-6}	1	0.69	0.24	3.6×10^{-3}
TG(54:3)	902.8175	2	0.57	0.12	4.2×10^{-6}	1	0.68	0.25	8.1×10^{-3}
TG(54:2)	904.8326	2	0.50	0.12	4.9×10^{-5}	1	0.51	0.19	7.3×10^{-3}
TG(56:5)	926.817	2	-0.03	0.38	9.4×10^{-1}	1	-0.40	0.19	3.4×10^{-2}

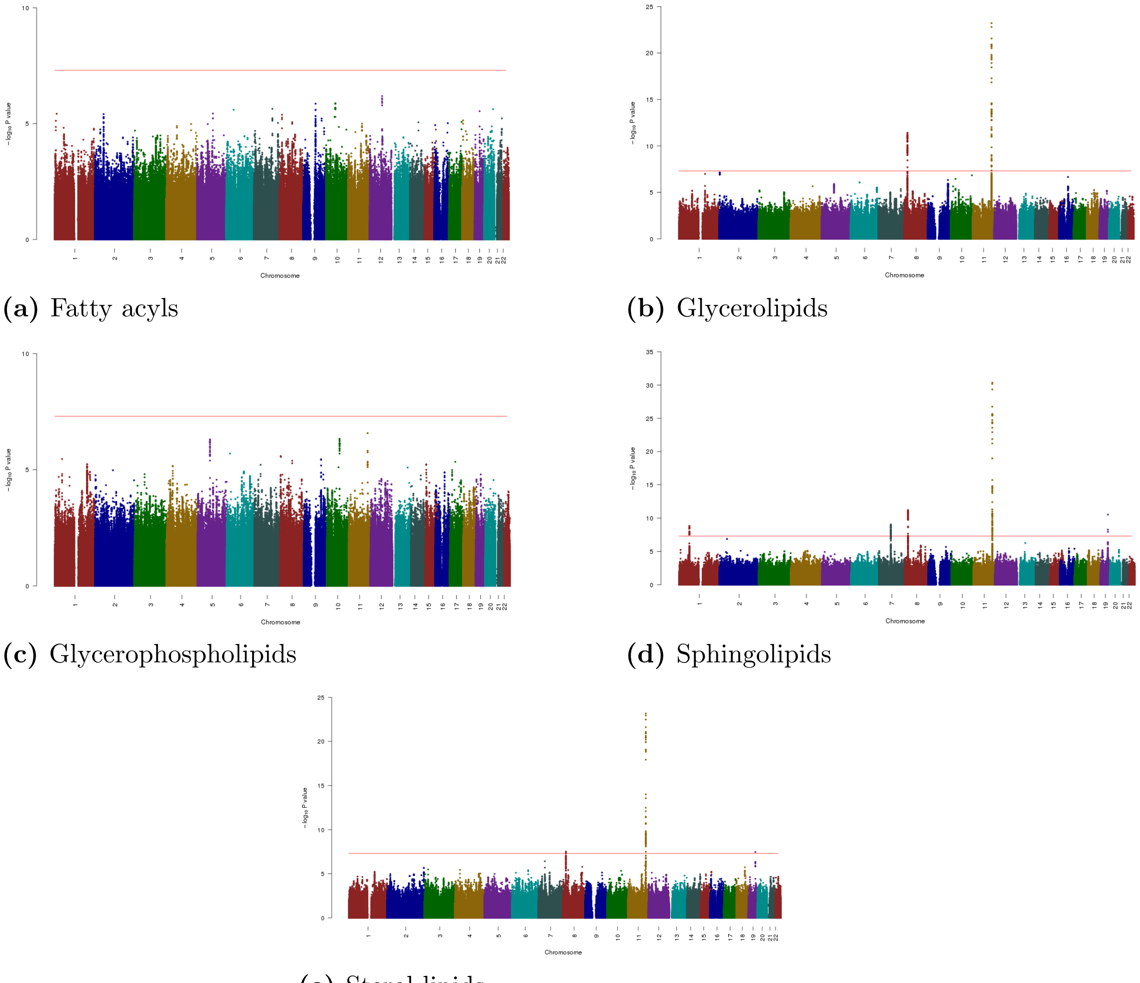
Abbreviations: IVW = Inverse-variance weighted; ***m/z*** = Mass-to-charge ratio; **SE** = Standard error.

ditional information that could be gleaned from performing an MR of the overall lipid categories or fourteen lipid subclasses.

7.3.3 Causal effect of linear combinations of lipids on risk of CHD

Finally, the third research question aimed to address whether weighted linear combinations of lipids, which were derived using PCA, have a causal effect on risk of CHD. The results of the GWAS of the principal components of the lipids were presented in subsection 5.3.5 and the Manhattan plots were shown in Figure 5.11.

Figure 7.4: Manhattan plots from GWAS of overall lipid categories



(a) Fatty acyls

(b) Glycerolipids

(c) Glycerophospholipids

(d) Sphingolipids

(e) Sterol lipids

The second principal component was significantly associated with only one variant in the *APOA5-APOC3* locus. The third principal component was also only associated with variants in the *APOA5-APOC3* locus, and the fourth principal component was associated with variants in both the *FADS1-2-3* and *APOA5-APOC3* loci. However, as discussed, these two loci were associated with over 100 lipid metabolites, and were excluded from the pruned approach that was used for MR of the individual lipids. Thus, even more so than for the lipid subclasses, extensive pleiotropy made it impossible to conduct an MR analysis on the principal components of the lipids.

7.4 Discussion

An MR analysis was conducted to assess the causal relevance of lipid metabolites for risk of CHD. It was intended that this would be performed by investigating lipids when considered on an individual basis, for each lipid subclass as a whole, and for linear combinations of lipids. For each of these distinct research questions, a detailed analysis plan was devised that employed several different MR methods, namely IVW, MR-Egger, and the weighted median approach.

The MR analyses identified eighteen lipids with evidence of a causal effect. However, there was extensive pleiotropy since the variants that were used as instrumental variables were associated with numerous lipids, which violates the assumptions of MR since there was not a single causal pathway from the genetic variants to the outcome that occurred only via the risk factor(s) under investigation. Therefore, it was not possible to determine which lipids are actually causally associated with CHD. Instead, only broad generalisations about the general direction of the association of lipids within particular subclasses can be made.

It appears that free fatty acids, cholesterol esters, and sphingomyelins have a protective effect on risk of CHD, while diglycerides and triglycerides have a positive association with risk of CHD. All of these associations are predominantly driven by variants in the *LPL* locus. Meanwhile, the association of glycerophospholipids with risk of CHD appears to vary depending on the lipid structure, which is predominantly driven by variants in the *LIPC* locus. Phosphocholines and phosphoethanolamines with a larger number of double bonds [e.g. PE(37:6), PE(40:7), PC(35:5)-H$^-$, and PE(37:4)-H$^-$] generally have a positive effect on CHD, while monounsaturated phosphocholines and phosphoethanolamines (i.e. with only one double bond) [e.g. PC-P(39:1), LysoPC(16:1), and PC-O(33:1)] appear to have a protective effect on CHD.

It was interesting that one of the triglycerides that was significantly associated with the newly discovered novel variant in the *PNPLA3* locus had a significant protective causal effect for CHD using the pruned approach, suggesting that individuals with elevated levels of this triglyceride are at reduced risk of CHD. However, this finding should not be overemphasised until it is explored further. The planned recall-by-genotype study described in Chapter 6 may be able to shed additional light on this finding.

Only the lipids that exhibited evidence of a significant causal effect on risk of CHD at $P < 0.05$ are shown in Figure 7.3; therefore, one can discern that predominantly diglycerides and triglycerides with zero, one, two, or three double bonds had significant causal effects. The same also holds for cholesterol esters and sphingomyelins. These lipids mostly contain saturated or monounsaturated fatty acids in their constituent side chains. Therefore, the overall takeaway from this MR analysis, although extensive pleiotropy prevents making any definite conclusions regarding causality, is that saturated and monounsaturated fatty acids are associated with reduced risk of CHD when they are found in the bloodstream either as free fatty acids or as part of cholesterol esters, sphingomyelins, phosphocholines, or lysophosphocholines, but they are associated with increased risk of CHD as constituents of diglycerides and triglycerides.

MR has been successfully utilised in other metabolomics studies; however, it was not possible to sufficiently address pleiotropy concerns in this analysis, even when considering alternative approaches such as examining the lipid subclasses as a whole, using principal component analysis, and using more sophisticated techniques that are more effective at addressing pleiotropy such as MR-Egger and the weighted median approach. It is likely that the high degree of pleiotropy in PROMIS occurred because of the particular lipidomics platform that was used. Other metabolomics platforms that use nuclear magnetic resonance (NMR) or mass spectrometry measure traits that are less closely correlated and therefore encounter less pleiotropy. The DIHRMS platform only measured lipids that belong to five overall categories, so it is unsurprising that the traits are highly correlated.

In addition to the lipid traits being highly correlated, there were an extremely limited number of genetic variants that were significantly associated with each lipid that could be used as instrumental variables in the MR analysis. Many of these variants included established CHD loci so the finding that these lipids have a significant causal effect for CHD is to be expected, but provides little added value and does not facilitate drawing any specific inference about the causal relevance of individual lipids due to the extensive pleiotropy. The nature of the lipidomics platform that was used and the modest number of significant genetic associations with these lipids meant that there was very little that could be done to overcome this limitation.

Another potential limitation of this two-sample MR analysis is that the GWAS association results came from very different populations, which could make them less suitable to be combined. The PROMIS GWAS results came from a Pakistani population, whereas the CHD results came from the CARDIoGRAMplusC4D consortium, which involves GWAS summary statistics from a meta-analysis of cases and controls of European descent. If the association results for the exposure and outcome come from different underlying populations then the inferences could be misleading if the variant is not a valid instrumental variable in both samples[101]. However, the genotyped samples in PROMIS were imputed to the 1000 Genomes reference panel, which is representative of a wide-range of ethnicities including South Asian and European, so this should help to overcome some of the differences in ethnicities between the two populations used in the MR analysis.

Despite the lack of evidence to identify a causal effect, the significant associations of lipid metabolites with CHD risk factors and genetic variants that were discussed in the previous chapters still hold. Even though specific individual lipids can not be pinpointed that are the key to causality, the overall findings show that specific lipid subclasses do have a causal effect on CHD. Future applications of this work could involve efforts to develop drugs that target the modification of levels of lipids that belong to specific subclasses.

CHAPTER 8

Summary of findings and potential implications

Chapter summary

This chapter summarises the aims of this dissertation, the methods that were followed, and the analyses that were conducted, and provides an overview of the key findings and their potential implications. Additionally, this chapter summarises the strengths of this analysis and a few of the limitations that were not able to be addressed. Furthermore, opportunities to further extend the analyses described in this dissertation in ongoing and future studies is discussed.

8.1 Summary of key findings

The aim of this dissertation was the identification of novel therapeutic targets through the study of high-dimensional phenotypic traits from blood lipids. Direct infusion high-resolution mass spectrometry yielded signal data for 444 lipid metabolites in 5662 individuals from the Pakistan Risk of Myocardial Infarction Study (PROMIS) following sample processing, data cleaning, and quality control (QC) filtering. Analyses of the Pearson and partial correlations between lipids showed that the lipids were highly correlated with each other and with levels of major circulating lipids and other biomarkers. Principal components of the lipids were also associated with increased levels of several coronary heart disease (CHD) risk factors. The lipids also exhibited associations with smoking status and physical activity. Triglycerides with fewer numbers of double bonds were associated with increased levels of smoking, while triglycerides with a higher number of double bonds were associated with decreased levels of smoking.

Analyses of the genetic determinants of the lipids led to the identification of 254 lipids that were significantly associated with one or more genetic variant(s) and 355 associations between single nucleotide polymorphisms (SNPs) and lipids. In total, these analyses identified 89 sentinel variants from 23 independent loci that were significantly associated with lipid metabolites. Four of these loci were considered novel as they had not previously been reported for association with major circulating lipids or metabolites. Genetic analyses of principal components and ratios of the lipids also led to the discovery of further associations.

The analyses conducted in this dissertation yielded a number of new biological insights into lipid metabolism. In addition to replicating and confirming known associations between lipids and genetic loci, this study also further extended what is known about these loci by identifying new associations. Several of the key novel genetic findings identified by this study are that: (1) decreased levels of sphingomyelins are associated with genetically lower *LPL* activity; (2) a wide range of glycerophospholipids are associated with variants in the *MBOAT7* locus; (3) several newly identified phosphatic acids, phosphocholines, and phosphoethanolamines are associated with variants in the *LIPC* region; (4) several novel associations were identified for sphingomyelins and phosphocholines with variants in the *APOE-C1-C2-C4* cluster; (5) four newly identified sphingomyelins are associated with variants in the *SGPP1* locus; and (6) several previously unreported phosphocholines, sphingomyelins, and ceramides are associated with variants in the *SPTLC3* locus. This is just a sample of the many novel associations that were identified as part of the analyses for this dissertation.

It was not possible to identify individual lipids with a causal effect on CHD, but general trends in the direction of association of lipid subclasses were noted. It was observed that saturated and monounsaturated fatty acids are associated with reduced risk of CHD when they are found in the bloodstream either as free fatty acids or as part of cholesterol esters, sphingomyelins, phosphocholines, or lysophosphocholines, but they are associated with increased risk of CHD as constituents of diglycerides and triglycerides.

As will be described in subsection 8.3.2, the preliminary results from a GWAS of INTERVAL participants using the same lipidomics platform helps to validate the PROMIS results as well as identify additional loci associated with these lipids. Further research is needed to extend the findings from these two studies and apply them to clinical and pharmacological settings.

A question that researchers are currently trying to answer is *how* should we go about lowering triglycerides to reduce risk of CHD? As discussed in Chapter 1, several phase III clinical trials are currently ongoing to investigate the reduction of triglyceride levels in adults with severe hypertriglyceridemia. Through the research described in this dissertation, a distinct pattern was identified in the association of triglyceride metabolites with major circulating triglycerides, and novel associations between triglycerides and genetic loci were also discovered. In particular, the detailed information about how associations vary according to the diversity of triglycerides and their constituent fatty acids sheds further clarity on this important topic. These insights could help inform further studies of triglyceride-lowering drugs.

8.2 Strengths and limitations of this dissertation

8.2.1 Strengths

This dissertation differs from previous studies of the genetic determinants of lipids in several important ways that enhance its scientific merit. First, the research involved participants from a novel population in Pakistan, thereby enhancing scientific understanding of lipid levels in this relatively understudied population. Second, the analysis was based on a relatively large dataset of 5662 participants, thereby increasing statistical power to detect associations. Third, the analyses were performed on individuals free from known myocardial infarction (MI) at baseline, which avoids the impact of having recently had an MI on the lipidome, the effects of which are largely unknown. Fourth, a newly developed open-profiling lipidomics platform was utilised to provide detailed lipid profiles, with a wider coverage of lipids than most other high-throughput profiling methods. Fifth, the genetic analyses resulted in the identification of novel loci and new biological insights into lipid metabolism. Sixth, a novel Mendelian randomisation (MR) approach was developed for analysing high-dimensional data, which has not ever been employed by other metabolomics studies. Finally, the attempt to implement this MR analysis plan led to the realisation that MR with lipidomics in particular is difficult to implement successfully due to a high degree of correlation between lipids and the resulting pleiotropy, although other high-dimensional –omics platforms such as metabolomics or proteomics may have more success with implementing the methods that were developed as part of this dissertation.

8.2.2 Limitations

The limitations of this dissertation also merit consideration. First, since PROMIS is a case-control study of MI, the usual limitations of case-control data apply, namely recall bias and selection bias. Recall bias could easily arise since the majority of the questionnaire was self-reported information. However, while recall bias did likely impact the questionnaire data to some extent, it could not have had an impact on the blood samples that were used to obtain the lipidomics measurements and genetic data. Furthermore, since controls were recruited from visitors and patients of people attending out-patient clinics and unrelated visitors of cardiac patients, this minimises the risk of selection bias amongst the controls who were used for this study. Another inherent limitation of the case-control study design is that participants were surveyed at baseline but there was not any available follow-up data. It would have been useful to know if some of the controls, even though they were free from cardiovascular disease (CVD) at baseline, actually ended up developing CVD later on, which could have been in part due to their lipid levels. However, it was not possible to obtain this information since the participants were not followed up.

Second, the serum samples were stored in freezers for several years before aliquots were taken for the lipidomics measurements. Although attempts were made to account for this by adjusting the analyses by the number of years that the samples had been stored, residual confounding may still be an issue, but even more importantly, the lipid profiles may have deteriorated, making an accurate assessment of the lipid signals difficult. However, since the samples were stored at $-80\,^{\circ}\mathrm{C}$ until use, this should not be an overwhelming concern.

A third limitation is that a majority of the participants were not fasting at time of blood draw (76 %), and there was also a small proportion of participants that reported having fasted but the duration was unknown (7 %). Recent food consumption may have had a significant impact on lipid levels and influenced the results. Although the analyses were adjusted for fasting status, a more strongly-powered study would be able to stratify by participants who were fasting versus non-fasting, thereby facilitating a comparison of the lipid profiles between these sets of individuals and a determination of whether there was a significant difference in lipid levels. Unfortunately, there was not enough statistical power to do so.

Fourth, the data consisted of participants from multiple centres in urban Pakistan, but it is unclear whether the findings from this study would be applicable to individuals living in rural villages and other parts of Pakistan, or in other South Asian countries—or more broadly, to the rest of the world. Since many characteristics of the PROMIS participants were markedly different from the overall population of Pakistan, the sample analysed in this dissertation many not be very representative. However, since many of the same lipids from this study were associated with known genetic regions such as *APOA5-APOC3* and *FADS1-2-3*, which have already been shown to be associated with multiple lipids in Western populations, this helps validate the legitimacy of the findings from this study.

Fifth, since most of the lipids were only associated with a small number of genetic regions, and there was significant pleiotropy in these regions, it was not possible to obtain causal estimates of individual lipid metabolites for CHD, which was one of the primary aims of this dissertation. However, the identification of novel loci using this platform, and the observation of consistent trends for the association with CHD amongst lipid subclasses, provides an impetus for further work.

Finally, the genetic analyses did not include a replication study to validate the findings. Ideally, the analyses would have been replicated in another cohort to ensure that the signals were legitimate and not false positives. Although especially stringent procedures were followed, highly conservative cut-offs were used to determine statistical significance, and rigorous pre-analysis and post-analysis QC was performed, there is still a possibility that some of the findings were false positives that arose due to artefacts rather than being true signals.

8.3 Ongoing and future studies

The present study has provided a comprehensive assessment of the genetic determinants of lipid metabolites and the association of lipid metabolites with CHD risk factors, and attempted to identify the causal relevance of these lipids for risk of CHD. Nevertheless, as outlined above and in the discussion sections of the individual chapters, there are opportunities to address some of the limitations of the present study and extend the analyses further.

8.3.1 Further analyses of lipidomics data in PROMIS

Many of the analyses that were conducted in this dissertation could have been extended further, which would be excellent for follow-up work.

First, while principal component analysis was used as a dimension reduction technique to facilitate analyses of the lipidomics data, other more sophisticated techniques such as Bayesian Hierarchical Clustering, Random Forests, or Support Vector Machines could be employed in follow-up analyses. These approaches could lead to additional insights into the correlations between lipids and the patterns that emerge in the data.

Second, while genetic analyses were performed using GWAS data, which can only detect variants with a minor allele frequency (MAF) of greater than 1 % (i.e. between 0.01 and 0.50), fine-mapping could be performed using Exome+ data, which can detect much rarer variants with a MAF of much less than 1 % (i.e. between 0.0001 and 0.01). This is important because coding variants with deleterious effects on corresponding proteins can help identify causal genes, but such variants are usually rare and the power to robustly detect the effect of such variants is small, even in a study with a moderate sample size. Gene burden tests on coding variants, which group rare variants together, can improve the analytical power and facilitate the identification of causal genes. By zeroing in on specific genetic regions of interest and exploring the associations with extremely rare variants using Exome+ genotypes, novel insights into the genetic determinants of lipid metabolism can be uncovered and potentially causal genes associated with various lipid metabolites can be more readily identified.

Third, recall-by-genotype studies can be employed to compare lipid levels in individuals with and without genetic mutations. By recruiting participants to take an oral fat challenge test or glucose challenge test and measuring their lipid metabolite levels both before and afterwards, their metabolic response to fat and glucose can be determined and the role that genetic mutations play in lipid metabolism can be studied in greater detail. Studies of knockout mice can also be employed effectively. As discussed in Chapter 6, a study to recall healthy volunteers registered in the NIHR Cambridge BioResource based on their *PNPLA3* genotype has been approved and will soon begin recruitment. These types of studies can be used to shed further light on the genetic determinants of lipid metabolism.

Finally, an approach known as statistical co-localisation can be used in situations where specific genetic variants may be causative of multiple phenotypes (i.e. pleiotropy). This approach involves integrating information on gene regulation, gene expression, metabolic pathways, and complex diseases under the assumption that genetic association signals shared between phenotypes are supportive of a causal genomic region[230]. Co-localisation has been successfully applied to test for overlapping genetic association signals across the entire genome with gene expression, CVD risk factors, and coronary artery disease (CAD) simultaneously[230]. Given the extensive pleiotropy amongst the lipidomics data, which made it difficult to assess causality using MR (see Chapter 7), this lipidomics platform may present an ideal opportunity to implement statistical co-localisation to provide further information about candidate causal genes for lipid metabolites and help prioritise metabolic pathways for functional follow-up.

8.3.2 Lipidomics analyses in INTERVAL

As mentioned, ideally a replication study would have been performed using the same lipidomics platform in a different but similar population, keeping only the genetic variants that were found in both the discovery dataset and the replication dataset. Although this was not possible for the present study, an assay is currently in progress to conduct DIHRMS in all 50 000 participants from the INTERVAL study. Data collection is still taking place in INTERVAL, but a few preliminary analyses have already been conducted, which are described here.

Background

INTERVAL was established as a randomised trial of whole blood donors enrolled from 25 centres of the National Health Service (NHS) Blood and Transplant in England[231]. The original purpose of the study was to determine whether the recommended length of time between blood donations can be "safely and acceptably decreased to optimise blood supply whilst maintaining the health of donors"[231]. However, the study was re-purposed by taking advantage of the bioresource of blood samples that had been collected from the study participants and conducting multiple types of assays from these samples, including genomics, proteomics, metabolomics, and lipidomics. Since the INTERVAL study recruited healthy blood donors from England, it is a considerably different population from urban Pakistan, but the use of the same lipidomics platform that was used in PROMIS will facilitate a proper replication study to take place. At the time of this writing, lipidomics data have only been measured in about 15 000 participants in INTERVAL, but this is already nearly triple the sample size compared to the number of PROMIS participants with available lipidomics data.

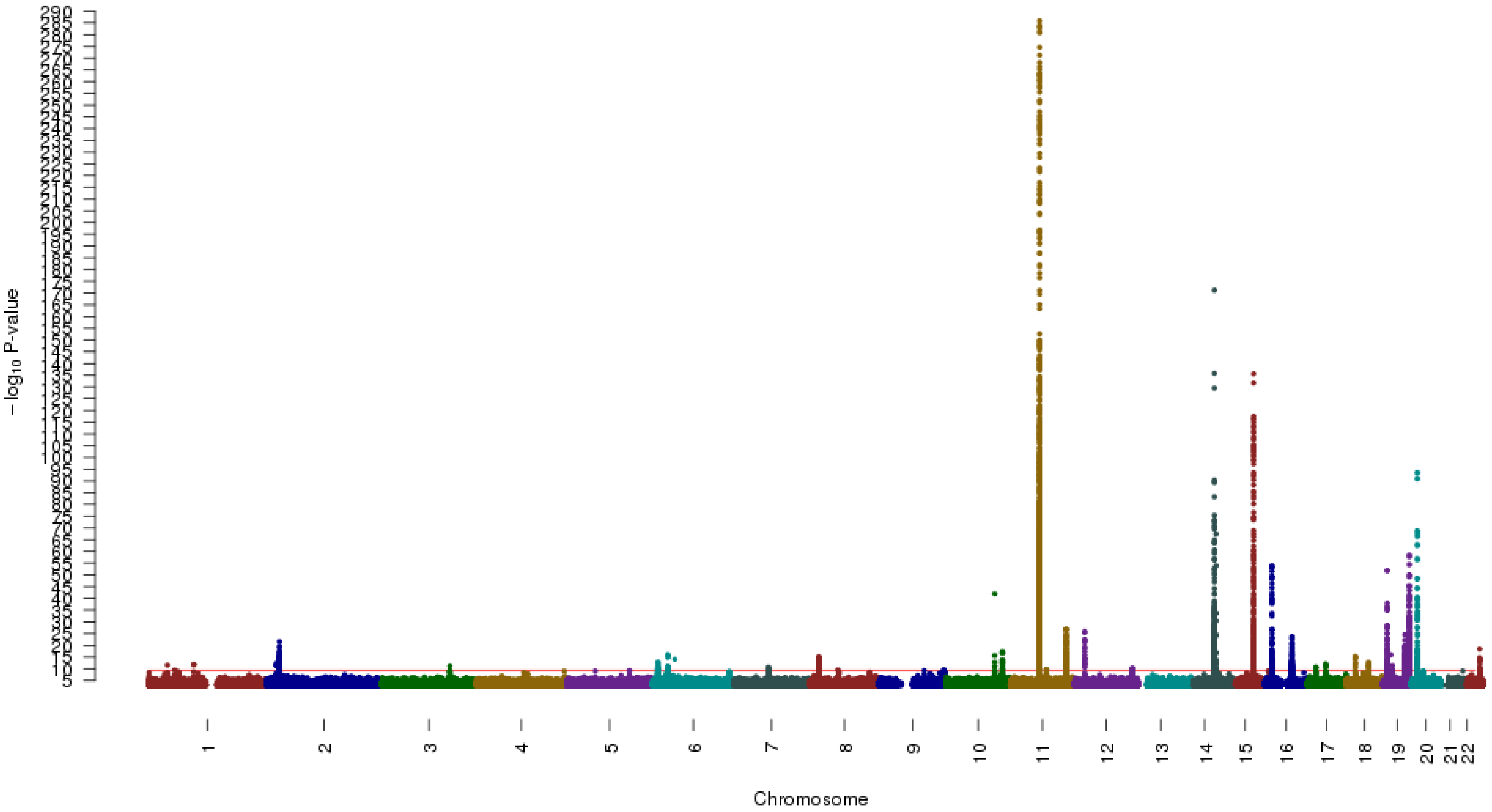
Methods

A univariate GWAS was conducted for each lipid metabolite measured in INTERVAL. One major difference in how the genetic analyses were conducted in PROMIS and INTERVAL is that to analyse the INTERVAL data BOLT-LMM was used instead of SNPTEST. The BOLT-LMM algorithm uses a linear mixed-model (LMM) association method that has been specifically designed to improve efficiency and increase the power to detect associations, particularly for GWAS conducted in large cohorts[232]. SNPTEST would have struggled to conduct this analysis in 15 000 individuals, and would not be able to handle the full 50 000 participants from INTERVAL when the complete lipidomics data becomes available, so BOLT-LMM is the best option. In contrast to SNPTEST, the algorithms used in BOLT-LMM rely on approximations that hold only at large sample sizes, but the programme does not work reliably with sample sizes of fewer than 5000 samples, so it would not have been suitable for analysis with the PROMIS data, which was only just over 5000 participants. Although the analyses from the two studies were conducted using different software packages, previous analyses comparing the output from the two programmes shows that the results are nearly identical.

Results

A Manhattan plot (Figure 8.1) was produced showing the combined genome-wide associations from these 558 lipids. So far the analyses have only been conducted using the genotyped data, but already the association of multiple lipids with variants in the *FADS1-2-3* locus is $P < 10^{-250}$. The Bonferroni-corrected P-value for genome-wide significance is 4.5×10^{-10} for INTERVAL, so these P-values in the *FADS1-2-3* locus are extremely highly significant. It is expected that a significant number of additional loci will reach genome-wide significance when these lipids are analysed using the imputed data. There were 37 loci in INTERVAL that reached genome-wide significance using the genotyped data, of which four of these were novel. Eighteen of these loci were the same ones that had been detected in PROMIS, and can therefore be considered to have replicated, which helps to validate the PROMIS results. The remaining 19 loci were only found in INTERVAL but not in PROMIS, which can be attributed not only to the fact that INTERVAL was conducted in a different study population but also that it has a larger sample size.

Figure 8.1: Global Manhattan plot using INTERVAL data showing association of 558 lipid metabolites with genotyped variants



Manhattan plot of combined results from genome-wide association study analysis using INTERVAL data for all lipids. *P*-values are shown for association of each SNP with each lipids. Red line indicates Bonferroni-corrected *P*-value for genome-wide significance (4.5×10^{-10}).

Comparison with PROMIS

More detailed analyses of the lipidomics data in INTERVAL awaits completion of data collection. Although lipidomics data on 15 000 participants has already been assayed, cleaned, and analysed, lipidomics measurements for the remaining 35 000 participants still needs to take place.

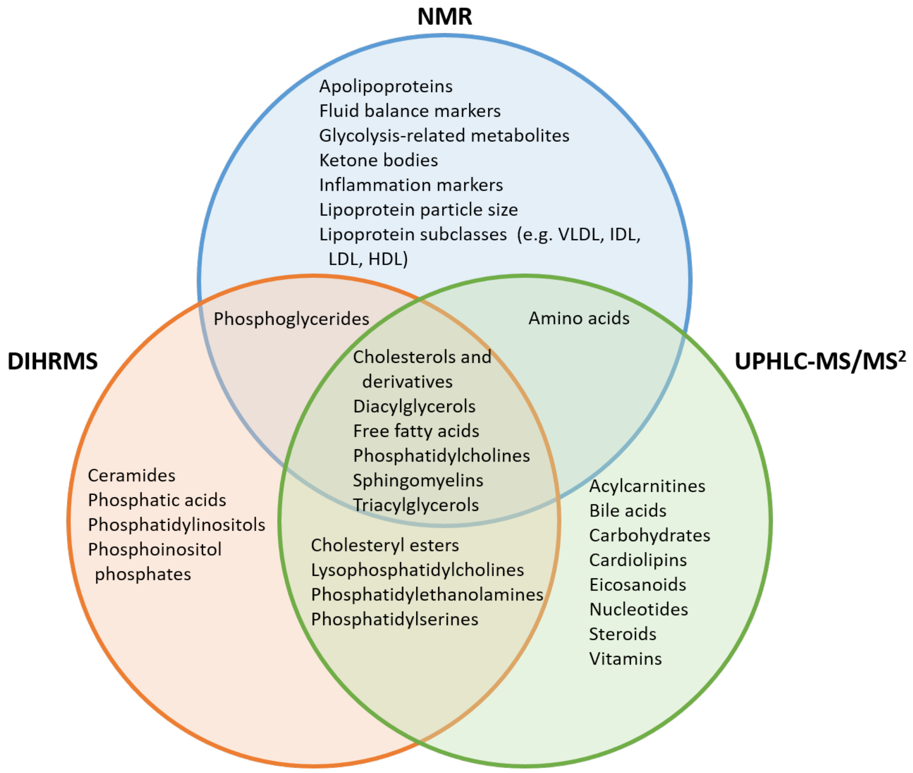
In addition to the substantially larger sample size, another significant advantage of INTERVAL is that multiple –omics assays were conducted in the same set of participants, facilitating the overlay of different sources of information. Demonstrating that high quality and reliable data can be obtained using different metabolomics approaches and assessing the differences between them would help to (1) enable cross-validation of findings resulting from analyses of data generated from multiple metabolomics assays, (2) generate hypotheses of biological mechanisms and metabolic pathways underlying CHD association signals, and (3) demonstrate robustness of measurements and quality of results. Proteomics data (3283 analytes derived from 2995 unique proteins) were obtained from 3301 INTERVAL participants using the Sysmex array, metabolomics data were collected using both ultra-high performance liquid chromatography–tandem mass spectrometry (UHPLC-MS/MS^2) (995 metabolites in 8536 participants) and nuclear magnetic resonance (NMR) spectroscopy (225 metabolites in 46 190 INTERVAL participants), and lipidomics data (558 lipid metabolites in 13 992 participants) were collected using DIHRMS. Furthermore, in terms of genetics data in INTERVAL there is GWAS, exome, whole genome sequencing (WGS) data, and whole exome sequencing (WES) data. This enables rich comparisons to take place, especially since some of the same lipids and other metabolites were measured using DIHRMS, UHPLC-MS/MS^2, and NMR. The overlap between metabolites measured on the three lipidomics/metabolomics platforms is shown in Figure 8.2.

By measuring the same metabolites on multiple platforms, the between-platform correlation in the levels of these overlapping metabolites can be assessed. Additionally, the strength of the association of these overlapping metabolites with previously identified genetic loci can be compared across platforms to assess the potential of each platform for large-scale discovery.

8.4 Conclusion

The primary objectives of this dissertation were (1) to identify the genetic determinants of lipid metabolites, and (2) to advance understanding of the effect of perturbations

Figure 8.2: Overlap of metabolites assayed by different metabolomics platforms



This Venn diagram shows types of metabolites that are measured by different metabolomics platforms. **Abbreviations: DIHRMS** = Direct Infusion High-Resolution Mass Spectrometry; **NMR** = Nuclear Magnetic Resonance spectroscopy; **UHPLC-MS/MS^2** = Ultra-High Performance Liquid Chromatography–Tandem Mass Spectrometry.

in lipid metabolite levels on CHD and its risk factors. The analyses conducted in this dissertation involved research in a novel population in Pakistan using a newly developed high-throughput open-profiling lipidomics platform. The diversity of lipids were analysed and their cross-correlations and associations with circulating biomarkers, lifestyle factors, and CHD risk factors was assessed. The analyses yielded many new biological insights into lipid metabolism, replicated and confirmed known associations between lipids and genetic loci, and further extended what is known about these loci by identifying new genetic associations. The findings in this dissertation have made important contributions to the advancement of the knowledge base in genetic epidemiology, lipid metabolism, and understanding the onset and development of coronary heart disease. It is hoped that further studies will follow on from this work to help advance mechanistic understanding of the genetic determinants of lipid metabolism and prioritise novel therapeutic targets for drug development and personalised medicine.

APPENDIX A

References

[1] GBD Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388(10053):1459–1544. doi:10.1016/S0140-6736(16)31012-1.

[2] Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med*. 2006;3(11):e442. doi:10.1371/journal.pmed.0030442.

[3] World Health Organization. *The atlas of heart disease and stroke*; 2004. Available from: http://www.who.int/cardiovascular_diseases/resources/atlas/.

[4] National Heart, Lung, and Blood Institute. *What is coronary heart disease?*; 2015. Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/cad.

[5] Abegunde DO, Mathers CD, Adam T, Ortegon M, Strong K. The burden and costs of chronic diseases in low-income and middle-income countries. *Lancet*. 2007;370(9603):1929–38. doi:10.1016/S0140-6736(07)61696-1.

[6] Ghaffar A, Reddy KS, Singhi M. Burden of non-communicable diseases in South Asia. *BMJ*. 2004;328(7443):807–10. doi:10.1136/bmj.328.7443.807.

[7] World Health Organization. *Cardiovascular diseases mortality: age-standardized death rate per 100 000 population, 2000-2012*; 2014. Available from: http://www.who.int/gho/en/.

[8] Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am Coll Cardiol*. 2017;70(1):1–25. doi:10.1016/j.jacc.2017.04.052.

[9] British Heart Foundation Health Promotion Research Group. *Ethnic differences in cardiovascular disease: 2010 edition*. British Heart Foundation; 2010. Available from: https://www.bhf.org.uk/publications/statistics/ethnic-differences-in-cardiovascular-disease-2010.

[10] McPherson R, Tybjærg-Hansen A. Genetics of coronary artery disease. *Circ Res*. 2016;118(4):564–78. doi:10.1161/CIRCRESAHA.115.306566.

[11] Kolovou G, Kolovou V, Mavrogeni S. Lipidomics in vascular health: current perspectives. *Vasc Health Risk Manag*. 2015;11:333–42. doi:10.2147/VHRM.S54874.

[12] Mundra PA, Shaw JE, Meikle PJ. Lipidomic analyses in epidemiology. *Int J Epidemiol*. 2016;45(5):1329–1338. doi:10.1093/ije/dyw112.

[13] Wilkin D, Brainard J. *Phospholipid bilayers*; 2016. Available from: `http://www.ck12.org/biology/Phospholipid-Bilayers/lesson/Phospholipid-Bilayers-BIO/`.

[14] Mabtech. *Apolipoproteins*; 2013. Available from: `https://www.mabtech.com/knowledge-center/applied-research/apolipoproteins`.

[15] Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*. 2008;40(2):161–9. doi:10.1038/ng.76.

[16] Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, et al. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med*. 2009;361(26):2518–28. doi:10.1056/NEJMoa0902604.

[17] Kamstrup PR, Tybjærg-Hansen A, Steffensen R, Nordestgaard BG. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *JAMA*. 2009;301(22):2331–9. doi:10.1001/jama.2009.801.

[18] Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet*. 2009;41(1):56–65. doi:10.1038/ng.291.

[19] Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010;466(7307):707–13. doi:10.1038/nature09270.

[20] Sarwar N, Danesh J, Eiriksdottir G, Sigurdsson G, Wareham N, Bingham S, et al. Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies. *Circulation*. 2007;115(4):450–8. doi:10.1161/CIRCULATIONAHA.106.637793.

[21] Sharrett AR, Ballantyne CM, Coady SA, Heiss G, Sorlie PD, Catellier D, et al. Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein(a), apolipoproteins A-I and B, and HDL density subfractions: the Atherosclerosis Risk in Communities (ARIC) Study. *Circulation*. 2001;104(10):1108–13. doi:10.1161/hc3501.095214.

[22] McBride P. Triglycerides and risk for coronary artery disease. *Curr Atheroscler Rep*. 2008;10(5):386–90. doi:10.1007/s11883-008-0060-9.

[23] Emerging Risk Factors Collaboration. Major lipids, apolipoproteins, and risk of vascular disease. *JAMA*. 2009;302(18):1993–2000. doi:10.1001/jama.2009.1619.

[24] Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration, Sarwar N, Sandhu MS, Ricketts SL, Butterworth AS, Di Angelantonio E, et al. Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *Lancet*. 2010;375(9726):1634–9. doi:10.1016/S0140-6736(10)60545-4.

[25] Saleheen D, Zaidi M, Rasheed A, Ahmad U, Hakeem A, Murtaza M, et al. The Pakistan Risk of Myocardial Infarction Study: a resource for the study of genetic, lifestyle and other determinants of myocardial infarction in South Asia. *Eur J Epidemiol*. 2009;24(6):329–38. doi:10.1007/s10654-009-9334-y.

[26] Saleheen D, Soranzo N, Rasheed A, Scharnagl H, Gwilliam R, Alexander M, et al. Genetic determinants of major blood lipids in Pakistanis compared with Europeans. *Circ Cardiovasc Genet*. 2010;3(4):348–57. doi:10.1161/CIRCGENETICS.109.906180.

[27] United States National Institutes of Health. *A study of AMR101 to evaluate its ability to reduce cardiovascular events in high risk patients with hypertriglyceridemia and on statin. The primary objective is to evaluate the effect of 4 g/day AMR101 for preventing the occurrence of a first major cardiovascular event. (REDUCE-IT)*; 2016. Available from: https://clinicaltrials.gov/ct2/show/NCT01492361.

[28] United States National Institutes of Health. *Outcomes study to assess statin residual risk reduction with EpaNova in high CV risk patients with hypertriglyceridemia (STRENGTH)*; 2017. Available from: https://clinicaltrials.gov/ct2/show/NCT02104817.

[29] Kowa Research Institute, Inc. PR Newswire, editor. *Landmark trial entitled "PROMINENT" to explore the prevention of heart disease in diabetic patients with high triglycerides and low HDL-C*; 2016. Available from: http://www.prnewswire.com/news-releases/landmark-trial-entitled-prominent-to-explore-the-prevention-of-heart-disease-in-diabetic-patients-with-high-triglycerides-and-low-hdl-c-300201581.html.

[30] National Institute of General Medical Sciences. National Institutes of Health, editor. *The new genetics*. U.S. Department of Health and Human Services; 2010. Available from: https://publications.nigms.nih.gov/thenewgenetics.

[31] Pertea M, Salzberg SL. Between a chicken and a grape: estimating the number of human genes. *Genome Biol*. 2010;11(5):206. doi:10.1186/gb-2010-11-5-206.

[32] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308(5720):385–9. doi:10.1126/science.1109557.

[33] Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA*. 2008;299(11):1335–44. doi:10.1001/jama.299.11.1335.

[34] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res*. 2014;42(Database issue):D1001–6. doi:10.1093/nar/gkt1229.

[35] Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45(11):1274–83. doi:10.1038/ng.2797.

[36] Asselbergs FW, Guo Y, van Iperen EP, Sivapalaratnam S, Tragante V, Lanktree MB, et al. Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am J Hum Genet*. 2012;91(5):823–38. doi:10.1016/j.ajhg.2012.08.032.

[37] Albrechtsen A, Grarup N, Li Y, Sparso T, Tian G, Cao H, et al. Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia*. 2013;56(2):298–310. doi:10.1007/s00125-012-2756-1.

[38] Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet*. 2014;94(2):223–32. doi:10.1016/j.ajhg.2014.01.009.

[39] Surakka I, Horikoshi M, Mägi R, Sarin AP, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. *Nat Genet.* 2015;47(6):589–97. doi:10.1038/ng.3300.

[40] Tang CS, Zhang H, Cheung CYY, Xu M, Ho JCY, Zhou W, et al. Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese. *Nat Commun.* 2015;6:10206. doi:10.1038/ncomms10206.

[41] Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. *N Engl J Med.* 2007;357(5):443–53. doi:10.1056/NEJMoa072366.

[42] Peden JF, Farrall M. Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum Mol Genet.* 2011;20(R2):R198–205. doi:10.1093/hmg/ddr384.

[43] CARDIoGRAMplusC4D Consortium. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet.* 2013;45(1):25–33. doi:10.1038/ng.2480.

[44] Vaxillaire M, Cavalcanti-Proenca C, Dechaume A, Tichet J, Marre M, Balkau B, et al. The common P446L polymorphism in *GCKR* inversely modulates fasting glucose and triglyceride levels and reduces type 2 diabetes risk in the DESIR prospective general French population. *Diabetes.* 2008;57(8):2253–7. doi:10.2337/db07-1807.

[45] Griffin JL, Atherton H, Shockcor J, Atzori L. Metabolomics as a tool for cardiac research. *Nat Rev Cardiol.* 2011;8(11):630–643. doi:10.1038/nrcardio.2011.138.

[46] Dehghan A. Mass spectrometry in epidemiological studies: What are the key considerations? *Eur J Epidemiol.* 2016;31(8):715–6. doi:10.1007/s10654-016-0195-x.

[47] Neavin D, Kaddurah-Daouk R, Weinshilboum R. Pharmacometabolomics informs pharmacogenomics. *Metabolomics.* 2016;12(7):1–6. doi:10.1007/s11306-016-1066-x.

[48] Han X, Yang K, Gross RW. Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. *Mass Spectrom Rev.* 2012;31(1):134–78. doi:10.1002/mas.20342.

[49] Han X, Gross RW. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. *J Lipid Res.* 2003;44(6):1071–9. doi:10.1194/jlr.R300004-JLR200.

[50] Graessler J, Schwudke D, Schwarz PE, Herzog R, Shevchenko A, Bornstein SR. Top-down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients. *PLoS One.* 2009;4(7):e6261. doi:10.1371/journal.pone.0006261.

[51] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem.* 2006;78(3):779–87. doi:10.1021/ac051437y.

[52] Puri R, Duong M, Uno K, Kataoka Y, Nicholls SJ. The emerging role of plasma lipidomics in cardiovascular drug discovery. *Expert Opin Drug Discov.* 2012;7(1):63–72. doi:10.1517/17460441.2012.644041.

[53] Stegemann C, Pechlaner R, Willeit P, Langley SR, Mangino M, Mayr U, et al. Lipidomics profiling and risk of cardiovascular disease in the prospective population-based Bruneck Study. *Circulation.* 2014;129(18):1821–1831. doi:10.1161/circulationaha.113.002500.

[54] Ravipati S, Baldwin DR, Barr HL, Fogarty AW, Barrett DA. Plasma lipid biomarker signatures in squamous carcinoma and adenocarcinoma lung cancer patients. *Metabolomics*. 2015;11(6):1600–1611. doi:10.1007/s11306-015-0811-x.

[55] Mapstone M, Cheema AK, Fiandaca MS, Zhong X, Mhyre TR, MacArthur LH, et al. Plasma phospholipids identify antecedent memory impairment in older adults. *Nat Med*. 2014;20(4):415–8. doi:10.1038/nm.3466.

[56] Garcia-Perez I, Posma JM, Gibson R, Chambers ES, Hansen TH, Vestergaard H, et al. Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. *Lancet Diabetes Endocrinol*. 2017;5(3):184–195. doi:10.1016/S2213-8587(16)30419-3.

[57] Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wägele B, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. 2011;477(7362):54–60. doi:10.1038/nature10354.

[58] Heather LC, Wang X, West JA, Griffin JL. A practical guide to metabolomic profiling as a discovery tool for human heart disease. *J Mol Cell Cardiol*. 2013;55:2–11. doi:10.1016/J.YJMCC.2012.12.001.

[59] Dehairs J, Derua R, Rueda-Rincon N, Swinnen JV. Lipidomics in drug development. *Drug Discov Today Technol*. 2015;13:33–8. doi:10.1016/j.ddtec.2015.03.002.

[60] Pechlaner R, Kiechl S, Mayr M. Potential and caveats of lipidomics for cardiovascular disease. *Circulation*. 2016;134(21):1651–1654. doi:10.1161/circulationaha.116.025092.

[61] Rhee EP, Cheng S, Larson MG, Walford GA, Lewis GD, McCabe E, et al. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. *J Clin Invest*. 2011;121(4):1402–11. doi:10.1172/JCI44442.

[62] Quehenberger O, Armando AM, Brown AH, Milne SB, Myers DS, Merrill AH, et al. Lipidomics reveals a remarkable diversity of lipids in human plasma. *J Lipid Res*. 2010;51(11):3299–305. doi:10.1194/jlr.M009449.

[63] Beger RD, Dunn W, Schmidt MA, Gross SS, Kirwan JA, Cascante M, et al. Metabolomics enables precision medicine: “A White Paper, Community Perspective”. *Metabolomics*. 2016;12(10):1–15. doi:10.1007/s11306-016-1094-6.

[64] Wang TJ, Ngo D, Psychogios N, Dejam A, Larson MG, Vasan RS, et al. 2-Aminoadipic acid is a biomarker for diabetes risk. *J Clin Invest*. 2013;123(10):4309–17. doi:10.1172/JCI64801.

[65] Xu F, Tavintharan S, Sum CF, Woon K, Lim SC, Ong CN. Metabolic signature shift in type 2 diabetes mellitus revealed by mass spectrometry-based metabolomics. *J Clin Endocr Metab*. 2013;98(6):E1060–5. doi:10.1210/jc.2012-4132.

[66] Forouhi NG, Koulman A, Sharp SJ, Imamura F, Kroger J, Schulze MB, et al. Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: the EPIC-InterAct case-cohort study. *Lancet Diabetes Endocrinol*. 2014;2(10):810–818. doi:10.1016/S2213-8587(14)70146-9.

[67] Forouhi NG, Imamura F, Sharp SJ, Koulman A, Schulze MB, Zheng J, et al. Association of plasma phospholipid n-3 and n-6 polyunsaturated fatty acids with type 2 diabetes: the EPIC-InterAct case-cohort study. *PLoS Med*. 2016;13(7):e1002094. doi:10.1371/journal.pmed.1002094.

[68] Nikolic SB, Sharman JE, Adams MJ, Edwards LM. Metabolomics in hypertension. *J Hypertens.* 2014;32(6):1159–69. doi:10.1097/HJH.0000000000000168.

[69] Fernandez C, Sandin M, Sampaio JL, Almgren P, Narkiewicz K, Hoffmann M, et al. Plasma lipid composition and risk of developing cardiovascular disease. *PLoS One.* 2013;8(8):e71846. doi:10.1371/journal.pone.0071846.

[70] Ganna A, Salihovic S, Sundstrom J, Broeckling CD, Hedman AK, Magnusson PK, et al. Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. *PLoS Genet.* 2014;10(12):e1004801. doi:10.1371/journal.pgen.1004801.

[71] Pirih N, Kunej T. Toward a taxonomy for multi-omics science? Terminology development for whole genome study approaches by omics technology and hierarchy. *OMICS: A Journal of Integrative Biology.* 2017;21(1):1–16. doi:10.1089/omi.2016.0144.

[72] Gieger C, Geistlinger L, Altmaier E, Hrabě de Angelis M, Kronenberg F, Meitinger T, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* 2008;4(11):e1000282. doi:10.1371/journal.pgen.1000282.

[73] Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, et al. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet.* 2010;42(2):137–41. doi:10.1038/ng.507.

[74] Kastenmüller G, Raffler J, Gieger C, Suhre K. Genetics of human metabolism: an update. *Hum Mol Genet.* 2015;24(R1):R93–R101. doi:10.1093/hmg/ddv263.

[75] Burkhardt R, Kirsten H, Beutner F, Holdt LM, Gross A, Teren A, et al. Integration of genome-wide SNP data and gene-expression profiles reveals six novel loci and regulatory mechanisms for amino acids and acylcarnitines in whole blood. *PLoS Genet.* 2015;11(9):e1005510. doi:10.1371/journal.pgen.1005510.

[76] Chasman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, et al. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet.* 2009;5(11):e1000730. doi:10.1371/journal.pgen.1000730.

[77] Hartiala JA, Tang WH, Wang Z, Crow AL, Stewart AF, Roberts R, et al. Genome-wide association study and targeted metabolomics identifies sex-specific association of *CPS1* with coronary artery disease. *Nat Commun.* 2016;7:10558. doi:10.1038/ncomms10558.

[78] Kettunen J, Demirkan A, Würtz P, Draisma HHM, Haller T, Rawal R, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of *LPA*. *Nat Commun.* 2016;7:11122. doi:10.1038/ncomms11122.

[79] Kraus WE, Muoio DM, Stevens R, Craig D, Bain JR, Grass E, et al. Metabolomic quantitative trait loci (mQTL) mapping implicates the ubiquitin proteasome system in cardiovascular disease pathogenesis. *PLoS Genet.* 2015;11(11):e1005553. doi:10.1371/journal.pgen.1005553.

[80] Long T, Hicks M, Yu HC, Biggs WH, Kirkness EF, Menni C, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet.* 2017;49(4):568–578. doi:10.1038/ng.3809.

[81] Petersen AK, Zeilinger S, Kastenmüller G, Römisch-Margl W, Brugger M, Peters A, et al. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Hum Mol Genet.* 2014;23(2):534–45. doi:10.1093/hmg/ddt430.

[82] Raffler J, Friedrich N, Arnold M, Kacprowski T, Rueedi R, Altmaier E, et al. Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genet.* 2015;11(9):e1005487. doi:10.1371/journal.pgen.1005487.

[83] Yet I, Menni C, Shin SY, Mangino M, Soranzo N, Adamski J, et al. Genetic influences on metabolite levels: a comparison across metabolomic platforms. *PLoS One.* 2016;11(4):e0153672. doi:10.1371/journal.pone.0153672.

[84] Yu B, Zheng Y, Alexander D, Manolio TA, Alonso A, Nettleton JA, et al. Genome-wide association study of a heart failure related metabolomic profile among African Americans in the Atherosclerosis Risk in Communities (ARIC) study. *Genet Epidemiol.* 2013;37(8):840–5. doi:10.1002/gepi.21752.

[85] Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol.* 2003;32(1):1–22. doi:10.1093/ije/dyg070.

[86] Burgess S, Thompson SG. *Mendelian randomization: methods for using genetic variants in causal estimation.* Chapman & Hall; 2015.

[87] Burgess S, Butterworth A, Malarstig A, Thompson S. Use of Mendelian randomisation to assess potential benefit of clinical intervention. *BMJ.* 2012;345:e7325. doi:10.1136/bmj.e7325.

[88] Ference BA. *Mendelian randomization studies: nature's randomized trials.* American College of Cardiology; 2015. Available from: http://www.acc.org/latest-in-cardiology/articles/2015/06/11/13/17/Mendelian-randomization-studies.

[89] Davey Smith G, Ebrahim S. Mendelian randomization: genetic variants as instruments for strengthening causal inference in observational studies. In: Weinstein M, Vaupel JW, Wachter KW, editors. Biosocial surveys. The National Academies Press; 2008. p. 336–366. Available from: https://www.nap.edu/catalog/11939/biosocial-surveys. doi:10.17226/11939.

[90] Hingorani A, Humphries S. Nature's randomised trials. *Lancet.* 2005;366(9501):1906–1908. doi:10.1016/s0140-6736(05)67767-7.

[91] Keavney B, Danesh J, Parish S, Palmer A, Clark S, Youngman L, et al. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. *Int J Epidemiol.* 2006;35(4):935–943. doi:10.1093/ije/dyl114.

[92] C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC), Wensley F, Gao P, Burgess S, Kaptoge S, Di Angelantonio E, et al. Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *BMJ.* 2011;342:d548. doi:10.1136/bmj.d548.

[93] VanderWeele T, Tchetgen Tchetgen E, Cornelis M, Kraft P. Methodological challenges in Mendelian randomization. *Epidemiology.* 2014;25(3):427–435. doi:10.1097/ede.0000000000000081.

[94] Interleukin 1 Genetics Consortium. Cardiometabolic consequences of genetic upregulation of the interleukin 1 receptor antagonist: a Mendelian randomisation analysis. *Lancet Diabetes Endocrinol.* 2015;3(4):243–253. doi:10.1016/S2213-8587(15)00034-0.

[95] Burgess S, Butterworth AS, Thompson JR. Beyond Mendelian randomization: how to interpret evidence of shared genetic predictors. *J Clin Epidemiol.* 2015;69:208–216. doi:10.1016/j.jclinepi.2015.08.001.

[96] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res.* 2007;16(4):309–330. doi:10.1177/0962280206077743.

[97] Pierce BL, Ahsan H, VanderWeele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol.* 2011;40(3):740–752. doi:10.1093/ije/dyq151.

[98] Johnson T. *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits.* Queen Mary University of London; 2011. Available from: `http://webspace.qmul.ac.uk/tjohnson/gtx/outline2.pdf`.

[99] Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol.* 2013;37(7):658–665. doi:10.1002/gepi.21758.

[100] Pierce B, Burgess S. Efficient design for Mendelian randomization studies: subsample and two-sample instrumental variable estimators. *Am J Epidemiol.* 2013;178(7):1177–1184. doi:10.1093/aje/kwt084.

[101] Burgess S, Scott R, Timpson N, Davey Smith G, Thompson S, EPIC-InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol.* 2015;30(7):543–552. doi:10.1007/s10654-015-0011-z.

[102] Burgess S, Timpson NJ, Ebrahim S, Smith GD. Mendelian randomization: where are we now and where are we going? *Int J Epidemiol.* 2015;44(2):379–388. doi:10.1093/ije/dyv108.

[103] Austin MA, Hutter CM, Zimmern RL, Humphries SE. Familial hypercholesterolemia and coronary heart disease: a HuGE association review. *Am J Epidemiol.* 2004;160(5):421–429. doi:10.1093/aje/kwh237.

[104] Strong A, Rader DJ. Clinical implications of lipid genetics for cardiovascular disease. *Curr Cardiovasc Risk Rep.* 2010;4(6):461–468. doi:10.1007/s12170-010-0131-7.

[105] Linsel-Nitschke P, Götz A, Erdmann J, Braenne I, Braund P, Hengstenberg C, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease—a Mendelian randomisation study. *PLoS One.* 2008;3(8):e2986. doi:10.1371/journal.pone.0002986.

[106] Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, Ripatti S, et al. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol.* 2010;30(11):2264–76. doi:10.1161/ATVBAHA.109.201020.

[107] Pedersen T, Kjekshus J, Berg K, Haghfelt T, Færgeman O, Thorgeirsson G, et al. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease:

the Scandinavian Simvastatin Survival Study (4S). *Lancet.* 1994;344(8934):1383–1389. doi:10.1016/S0140-6736(94)90566-5.

[108] Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N Engl J Med.* 2006;354(12):1264–1272. doi:10.1056/nejmoa054013.

[109] Fitzgerald K, Frank-Kamenetsky M, Shulga-Morskaya S, Liebow A, Bettencourt BR, Sutherland JE, et al. Effect of an RNA interference drug on the synthesis of proprotein convertase subtilisin/kexin type 9 (PCSK9) and the concentration of serum LDL cholesterol in healthy volunteers: a randomised, single-blind, placebo-controlled, phase 1 trial. *Lancet.* 2014;383(9911):60–68. doi:10.1016/s0140-6736(13)61914-5.

[110] Stitziel NO, Won HH, Morrison AC, Peloso GM, Do R, Lange LA, et al. Inactivating mutations in *NPC1L1* and protection from coronary heart disease. *N Engl J Med.* 2014;371(22):2072–2082. doi:10.1056/NEJMoa1405386.

[111] Cannon CP, Blazing MA, Giugliano RP, McCagg A, White JA, Theroux P, et al. Ezetimibe added to statin therapy after acute coronary syndromes. *N Engl J Med.* 2015;372(25):2387–2397. doi:10.1056/NEJMoa1410489.

[112] Ference BA, Majeed F, Penumetcha R, Flack JM, Brook RD. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in *NPC1L1*, *HMGCR*, or both: a 2 × 2 factorial Mendelian randomization study. *J Am Coll Cardiol.* 2015;65(15):1552–61. doi:10.1016/j.jacc.2015.02.020.

[113] Ference BA, Yoo W, Alesh I, Mahajan N, Mirowska KK, Mewada A, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. *J Am Coll Cardiol.* 2012;60(25):2631–2639. doi:10.1016/j.jacc.2012.09.017.

[114] Taylor F, Huffman MD, Macedo AF, Moore TH, Burke M, Davey Smith G, et al. Statins for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev.* 2013;2013(1):CD004816. doi:10.1002/14651858.CD004816.pub5.

[115] Sattar N, Preiss D, Murray HM, Welsh P, Buckley BM, de Craen AJ, et al. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. *Lancet.* 2010;375(9716):735–742. doi:10.1016/S0140-6736(09)61965-6.

[116] Fall T, Xie W, Poon W, Yaghootkar H, Mägi R, GENESIS Consortium, et al. Using genetic variants to assess the relationship between circulating lipids and type 2 diabetes. *Diabetes.* 2015;64(7):2676–84. doi:10.2337/db14-1710.

[117] Burgess S, Thompson S. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol.* 2015;181(4):251–260. doi:10.1093/aje/kwu283.

[118] Do R, Willer CJ, Schmidt EM, Sengupta S, Gao C, Peloso GM, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet.* 2013;45(11):1345–52. doi:10.1038/ng.2795.

[119] Burgess S, Freitag DF, Khan H, Gorman DN, Thompson SG. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS One.* 2014;9(10):e108891. doi:10.1371/journal.pone.0108891.

[120] Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet*. 2012;380(9841):572–80. doi:10.1016/S0140-6736(12)60312-2.

[121] Holmes MV, Asselbergs FW, Palmer TM, Drenos F, Lanktree MB, Nelson CP, et al. Mendelian randomization of blood lipids for coronary heart disease. *Eur Heart J*. 2015;36(9):539–50. doi:10.1093/eurheartj/eht571.

[122] Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015;44(2):512–525. doi:10.1093/ije/dyv080.

[123] Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*. 2016;40(4):304–14. doi:10.1002/gepi.21965.

[124] Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629–634. doi:10.1136/bmj.315.7109.629.

[125] Pickrell J. Fulfilling the promise of Mendelian randomization. *bioRxiv*. 2015;018150. doi:10.1101/018150.

[126] Kolesár M, Chetty R, Friedman J, Glaeser E, Imbens G. Identification and inference with many invalid instruments. *J Bus Econ Stat*. 2014;33(4):474–484. doi:10.1080/07350015.2014.978175.

[127] Han C. Detecting invalid instruments using L_1-GMM. *Econ Lett*. 2008;101(3):285–287. doi:10.1016/j.econlet.2008.09.004.

[128] Schwartz GG, Olsson AG, Abt M, Ballantyne CM, Barter PJ, Brumm J, et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. *N Engl J Med*. 2012;367(22):2089–2099. doi:10.1056/nejmoa1206797.

[129] Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology*. 2017;28(1):30–42. doi:10.1097/EDE.0000000000000559.

[130] Suhre K, Raffler J, Kastenmüller G. Biochemical insights from population studies with genetics and metabolomics. *Arch Biochem Biophys*. 2016;589:168–76. doi:10.1016/j.abb.2015.09.023.

[131] Würtz P, Wang Q, Kangas AJ, Richmond RC, Skarp J, Tiainen M, et al. Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Med*. 2014;11(12):e1001765. doi:10.1371/journal.pmed.1001765.

[132] Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014;46(6):543–50. doi:10.1038/ng.2982.

[133] Korostishevsky M, Steves CJ, Malkin I, Spector T, Williams FM, Livshits G. Genomics and metabolomics of muscular mass in a community-based sample of UK females. *Eur J Hum Genet*. 2016;24(2):277–83. doi:10.1038/ejhg.2015.85.

[134] Bartel J, Krumsiek J, Schramm K, Adamski J, Gieger C, Herder C, et al. The human blood metabolome–transcriptome interface. *PLoS Genet*. 2015;11(6):e1005274. doi:10.1371/journal.pgen.1005274.

[135] Würtz P, Kangas AJ, Soininen P, Lehtimäki T, Kähönen M, Viikari JS, et al. Lipoprotein subclass profiling reveals pleiotropy in the genetic variants of lipid risk factors for coronary heart disease: a note on Mendelian randomization studies. *J Am Coll Cardiol*. 2013;62(20):1906–1908. doi:10.1016/j.jacc.2013.07.085.

[136] Demographia. *Demographia world urban areas (13th ed.)*. Demographia; 2017. Available from: http://www.demographia.com/db-worldua.pdf.

[137] ESRI. *ArcGIS for Desktop: Release 10.4.1* [Computer Program]. Environmental Systems Research Institute (ESRI); 2016. Available from: http://desktop.arcgis.com.

[138] Hijmans R, Kapoor J, Wieczorek J, Garcia N, Maunahan A, Rala A, et al.. *GADM database of Global Administrative Areas* [Database]. GADM; 2015. Available from: http://gadm.org.

[139] National Institute of Population Studies (NIPS) Pakistan, ICF International. *Pakistan Demographic and Health Survey 2012-13*. NIPS/Pakistan and ICF International; 2013. Available from: http://dhsprogram.com/pubs/pdf/FR290/FR290.pdf.

[140] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc*. 2010;5(9):1564–1573. doi:10.1038/nprot.2010.116.

[141] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65. doi:10.1038/nature11632.

[142] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5(6):e1000529. doi:10.1371/journal.pgen.1000529.

[143] Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*. 2010;11(7):499–511. doi:10.1038/nrg2796.

[144] SAS Institute Inc. *SAS* [Computer Program]. SAS Institute Inc; 2010. Available from: https://www.sas.com.

[145] Dejaegher B, Heyden YV. Ruggedness and robustness testing. *J Chromatogr A*. 2007;1158(1-2):138–57. doi:10.1016/j.chroma.2007.02.086.

[146] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. *Mass Spectrom Rev*. 2007;26(1):51–78. doi:10.1002/mas.20108.

[147] Li M, Yang L, Bai Y, Liu H. Analytical methods in lipidomics and their applications. *Anal Chem*. 2014;86(1):161–175. doi:10.1021/ac403554h.

[148] Suhre K, Gieger C. Genetic variation in metabolic phenotypes: study designs and applications. *Nat Rev Genet*. 2012;13(11):759–69. doi:10.1038/nrg3314.

[149] Kirwan JA, Weber RJ, Broadhurst DI, Viant MR. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Sci Data*. 2014;1:140012. doi:10.1038/sdata.2014.12.

[150] Courant F, Antignac JP, Dervilly-Pinel G, Le Bizec B. Basics of mass spectrometry based metabolomics. *Proteomics*. 2014;14(21-22):2369–2388. doi:10.1002/pmic.201400255.

[151] Hendriks MMWB, Eeuwijk FAv, Jellema RH, Westerhuis JA, Reijmers TH, Hoefsloot HCJ, et al. Data-processing strategies for metabolomics studies. *Trends Anal Chem (TrAC)*. 2011 11;30(10):1685–1698. doi:10.1016/J.TRAC.2011.04.019.

[152] Qin LX, Zhou Q, Bogomolniy F, Villafania L, Olvera N, Cavatore M, et al. Blocking and randomization to improve molecular biomarker discovery. *Clin Cancer Res*. 2014;20(13):3371–8. doi:10.1158/1078-0432.CCR-13-3155.

[153] Evans A. Precision metabolomics: a single technology for compliance testing, toxicology and personalized medicine. In: HT-ADME 2017 Conference. The Boston Society; 2017. Available from: https://www.bostonsociety.org/HT-ADME/images/HT-ADME/pdfs/2017/HTADME2017program.pdf.

[154] Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CR, Shimizu T, et al. Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res*. 2009;50 Suppl:S9–14. doi:10.1194/jlr.R800095-JLR200.

[155] Moore RT. *blockTools: Blocking, assignment, and diagnosing interference in randomized experiments* [Computer Program]; 2014. R package. Available from: https://cran.r-project.org/package=blockTools.

[156] R Core Team. *R: a language and environment for statistical computing* [Computer Program]; 2013. Available from: http://www.r-project.org.

[157] Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnol*. 2012;30(10):918–20. doi:10.1038/nbt.2377.

[158] Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008;24(21):2534–6. doi:10.1093/bioinformatics/btn323.

[159] Python Software Foundation. *Python* [Computer Program]; 2013. Available from: https://www.python.org.

[160] Masoodi M, Eiden M, Koulman A, Spaner D, Volmer DA. Comprehensive lipidomics analysis of bioactive lipids in complex regulatory networks. *Anal Chem*. 2010;82(19):8176–85. doi:10.1021/ac1015563.

[161] Hu C, Kong H, Qu F, Li Y, Yu Z, Gao P, et al. Application of plasma lipidomics in studying the response of patients with essential hypertension to antihypertensive drug therapy. *Mol Biosyst*. 2011;7(12):3271–9. doi:10.1039/c1mb05342f.

[162] Brugnara L, Vinaixa M, Murillo S, Samino S, Rodriguez MA, Beltran A, et al. Metabolomics approach for analyzing the effects of exercise in subjects with type 1 diabetes mellitus. *PLoS One*. 2012;7(7):e40600. doi:10.1371/journal.pone.0040600.

[163] Hong MG, Karlsson R, Magnusson PK, Lewis MR, Isaacs W, Zheng LS, et al. A genome-wide assessment of variability in human serum metabolism. *Hum Mutat*. 2013;34(3):515–24. doi:10.1002/humu.22267.

[164] Shen M, Broeckling CD, Chu EY, Ziegler G, Baxter IR, Prenni JE, et al. Leveraging non-targeted metabolite profiling via statistical genomics. *PLoS One*. 2013;8(2):e57667. doi:10.1371/journal.pone.0057667.

[165] Koulman A, Prentice P, Wong MCY, Matthews L, Bond NJ, Eiden M, et al. The development and validation of a fast and robust dried blood spot based lipid profiling method to study infant metabolism. *Metabolomics*. 2014;10(5):1018–1025. doi:10.1007/s11306-014-0628-z.

[166] Demirkan A, Henneman P, Verhoeven A, Dharuri H, Amin N, van Klinken JB, et al. Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet*. 2015;11(1):e1004835. doi:10.1371/journal.pgen.1004835.

[167] Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikainen LP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet*. 2012;44(3):269–76. doi:10.1038/ng.1073.

[168] Raffler J, Römisch-Margl W, Petersen AK, Pagel P, Blochl F, Hengstenberg C, et al. Identification and MS-assisted interpretation of genetically influenced NMR signals in human plasma. *Genome Med*. 2013;5(2):13. doi:10.1186/gm417.

[169] Ried JS, Shin SY, Krumsiek J, Illig T, Theis FJ, Spector TD, et al. Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses. *Hum Mol Genet*. 2014;23(21):5847–57. doi:10.1093/hmg/ddu301.

[170] Jenkins BJ, Seyssel K, Chiu S, Pan PH, Lin SY, Stanley E, et al. Odd chain fatty acids; new insights of the relationship between the gut microbiota, dietary intake, biosynthesis and glucose intolerance. *Sci Rep*. 2017;7:44845. doi:10.1038/srep44845.

[171] Aimo L, Liechti R, Hyka-Nouspikel N, Niknejad A, Gleizes A, Götz L, et al. The SwissLipids knowledgebase for lipid biology. *Bioinformatics*. 2015;31(17):2860–6. doi:10.1093/bioinformatics/btv285.

[172] Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0–a comprehensive server for metabolomic data analysis. *Nucleic Acids Res*. 2012;40(Web Server issue):W127–33. doi:10.1093/nar/gks374.

[173] Koulman A, Lane GA, Harrison SJ, Volmer DA. From differentiating metabolites to biomarkers. *Anal Bioanal Chem*. 2009;394(3):663–70. doi:10.1007/s00216-009-2690-3.

[174] Watson AD. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Lipidomics: a global approach to lipid analysis in biological systems. *J Lipid Res*. 2006;47(10):2101–11. doi:10.1194/jlr.R600022-JLR200.

[175] Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr Bioinform*. 2012;7(1):96–108. doi:10.2174/157489312799304431.

[176] Joliffe IT, Morgan BJ. Principal component analysis and exploratory factor analysis. *Stat Methods Med Res*. 1992;1(1):69–95. doi:10.1177/096228029200100105.

[177] van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*. 2006;7:142. doi:10.1186/1471-2164-7-142.

[178] Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *J Proteome Res*. 2012;11(8):4120–31. doi:10.1021/pr300231n.

[179] Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol.* 2011;5:21. doi:10.1186/1752-0509-5-21.

[180] Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohney RP, Milburn MV, et al. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.* 2012;8(10):e1003005. doi:10.1371/journal.pgen.1003005.

[181] Barker M, Rayens W. Partial least squares for discrimination. *J Chemometr.* 2003;17(3):166–173. doi:Doi 10.1002/Cem.785.

[182] Kim JA, Choi HJ, Kwon YK, Ryu DH, Kwon TH, Hwang GS. ^{1}H NMR-based metabolite profiling of plasma in a rat model of chronic kidney disease. *PLoS One.* 2014;9(1):e85445. doi:10.1371/journal.pone.0085445.

[183] Miao H, Chen H, Zhang X, Yin L, Chen DQ, Cheng XL, et al. Urinary metabolomics on the biochemical profiles in a rat hyperlipidemia model using ultra-performance liquid-chromatography coupled with quadrupole time-of-flight synapt high-definition mass spectrometry. *J Anal Methods Chem.* 2014;184162:1–9. doi:10.1155/2014/184162.

[184] Acharjee A, Finkers R, Visser RG, Maliepaard C. Comparison of regularized regression methods for omics data. *Metabolomics: Open Access.* 2013;3(3):126. doi:10.4172/2153-0769.1000126.

[185] Lê Cao KA, Rohart F, González I, Déjean S, Gautier B, Bartolo F, et al.. *mixOmics: Omics Data Integration Project*; 2017. R package version 6.3.1. Available from: `https://CRAN.R-project.org/package=mixOmics`.

[186] Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Comput Biol.* 2017;13(11):e1005752. doi:10.1371/journal.pcbi.1005752.

[187] D’Agostino S R B, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation.* 2008;117(6):743–53. doi:10.1161/CIRCULATIONAHA.107.699579.

[188] Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58(301):236–244. doi:10.1080/01621459.1963.10500845.

[189] Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol.* 2007;1:37. doi:10.1186/1752-0509-1-37.

[190] Gobbi A, Iorio F, Dawson KJ, Wedge DC, Tamborero D, Alexandrov LB, et al. Fast randomization of large genomic datasets while preserving alteration counts. *Bioinformatics.* 2014;30(17):i617–23. doi:10.1093/bioinformatics/btu474.

[191] González I, Lê Cao KA, Davis MJ, Déjean S. Visualising associations between paired ‘omics’ data sets. *BioData Min.* 2012;5(1):19. doi:10.1186/1756-0381-5-19.

[192] Imamura F, Sharp SJ, Koulman A, Schulze MB, Kroger J, Griffin JL, et al. A combination of plasma phospholipid fatty acids and its association with incidence of type 2 diabetes: the EPIC-InterAct case-cohort study. *PLoS Med.* 2017;14(10):e1002409. doi:10.1371/journal.pmed.1002409.

[193] Worley B, Powers R. Multivariate analysis in metabolomics. *Curr Metabolomics*. 2013;1(1):92–107. doi:10.2174/2213235X11301010092.

[194] Wojczynski MK, Glasser SP, Oberman A, Kabagambe EK, Hopkins PN, Tsai MY, et al. High-fat meal effect on LDL, HDL, and VLDL particle size and number in the Genetics of Lipid-Lowering drugs and diet network (GOLDN): an interventional study. *Lipids Health Dis*. 2011;10(1):181. doi:10.1186/1476-511X-10-181.

[195] Wojczynski MK, Parnell LD, Pollin TI, Lai CQ, Feitosa MF, O'Connell JR, et al. Genome-wide association study of triglyceride response to a high-fat meal among participants of the NHLBI Genetics of Lipid Lowering Drugs and Diet Network (GOLDN). *Metabolism*. 2015;64(10):1359–71. doi:10.1016/j.metabol.2015.07.001.

[196] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75. doi:10.1086/519795.

[197] Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*. 2007;23(10):1294–6. doi:10.1093/bioinformatics/btm108.

[198] Clayton D. *snpStats: SnpMatrix and XSnpMatrix classes and methods* [Computer Program]; 2013. Available from: `https://bioconductor.org/packages/snpStats/`.

[199] Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851–61. doi:10.1038/nature06258.

[200] Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet Epidemiol*. 2010;34(1):100–5. doi:10.1002/gepi.20430.

[201] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336–7. doi:10.1093/bioinformatics/btq419.

[202] Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190–1. doi:10.1093/bioinformatics/btq340.

[203] Altmaier E, Ramsay SL, Graber A, Mewes HW, Weinberger KM, Suhre K. Bioinformatics analysis of targeted metabolomics–uncovering old and new tales of diabetic mice under medication. *Endocrinology*. 2008;149(7):3478–89. doi:10.1210/en.2007-1747.

[204] Petersen AK, Krumsiek J, Wägele B, Theis FJ, Wichmann HE, Gieger C, et al. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics*. 2012;13:120. doi:10.1186/1471-2105-13-120.

[205] Weissglas-Volkov D, Aguilar-Salinas CA, Nikkola E, Deere KA, Cruz-Bautista I, Arellano-Campos O, et al. Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. *J Med Genet*. 2013;50(5):298–308. doi:10.1136/jmedgenet-2012-101461.

[206] Wu JH, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. *Circ Cardiovasc Genet*. 2013;6(2):171–83. doi:10.1161/CIRCGENETICS.112.964619.

[207] Feng Q, Baker SS, Liu W, Arbizu RA, Aljomah G, Khatib M, et al. Increased apolipoprotein A5 expression in human and rat non-alcoholic fatty livers. *Pathology*. 2015;47(4):341–8. doi:10.1097/PAT.0000000000000251.

[208] Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*. 2001;294(5540):169–73. doi:10.1126/science.1064852.

[209] Talmud PJ, Hawe E, Martin S, Olivier M, Miller GJ, Rubin EM, et al. Relative contribution of variation within the *APOC3/A4/A5* gene cluster in determining plasma triglycerides. *Hum Mol Genet*. 2002;11(24):3039–46. doi:10.1093/hmg/11.24.3039.

[210] Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet*. 2010;42(8):684–7. doi:10.1038/ng.628.

[211] Talmud PJ, Drenos F, Shah S, Shah T, Palmen J, Verzilli C, et al. Gene-centric association signals for lipids and apolipoproteins identified via the HumanCVD BeadChip. *Am J Hum Genet*. 2009;85(5):628–42. doi:10.1016/j.ajhg.2009.10.014.

[212] Plaisier CL, Horvath S, Huertas-Vazquez A, Cruz-Bautista I, Herrera MF, Tusie-Luna T, et al. A systems genetics approach implicates *USF1*, *FADS3*, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet*. 2009;5(9):e1000642. doi:10.1371/journal.pgen.1000642.

[213] Draisma HH, Pool R, Kobl M, Jansen R, Petersen AK, Vaarhorst AAM, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun*. 2015;6:7208. doi:10.1038/ncomms8208.

[214] Stacey D, Fauman E, Ziemek D, Sun B, Harshfield E, Wood A, et al. ProGeM: a framework for the prioritisation of candidate causal genes at molecular quantitative trait loci. *bioRxiv*. 2017;230094. doi:10.1101/230094.

[215] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biology*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4.

[216] GTEx Consortium, Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648–60. doi:10.1126/science.1262110.

[217] Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*. 2016;32(20):3207–3209. doi:10.1093/bioinformatics/btw373.

[218] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308–11. doi:10.1093/nar/29.1.308.

[219] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc.* 2007;2(10):2366–82. doi:10.1038/nprot.2007.324.

[220] Chamorro AJ, Torres JL, Miron-Canelo JA, Gonzalez-Sarmiento R, Laso FJ, Marcos M. Systematic review with meta-analysis: the I148M variant of patatin-like phospholipase domain-containing 3 gene (*PNPLA3*) is significantly associated with alcoholic liver cirrhosis. *Aliment Pharmacol Ther.* 2014;40(6):571–81. doi:10.1111/apt.12890.

[221] Stickel F, Moreno C, Hampe J, Morgan MY. The genetics of alcohol dependence and alcohol-related liver disease. *J Hepatol.* 2017;66(1):195–211. doi:10.1016/j.jhep.2016.08.011.

[222] Roeske-Nielsen A, Buschard K, Månson JE, Rastam L, Lindblad U. A variation in the cerebroside sulfotransferase gene is linked to exercise-modified insulin resistance and to type 2 diabetes. *Exp Diabetes Res.* 2009;2009:429593. doi:10.1155/2009/429593.

[223] Piperi C, Adamopoulos C, Papavassiliou AG. *XBP1*: a pivotal transcriptional regulator of glucose and lipid metabolism. *Trends Endocrinol Metab.* 2016;27(3):119–22. doi:10.1016/j.tem.2016.01.001.

[224] Glimcher LH, Lee AH. From sugar to fat: how the transcription factor *XBP1* regulates hepatic lipogenesis. *Ann N Y Acad Sci.* 2009;1173 Suppl 1:E2–9. doi:10.1111/j.1749-6632.2009.04956.x.

[225] Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, Ugocsai P, et al. Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genet.* 2009;5(10):e1000672. doi:10.1371/journal.pgen.1000672.

[226] Schaap FG, Rensen PC, Voshol PJ, Vrins C, van der Vliet HN, Chamuleau RA, et al. ApoAV reduces plasma triglycerides by inhibiting very low density lipoprotein-triglyceride (VLDL-TG) production and stimulating lipoprotein lipase-mediated VLDL-TG hydrolysis. *J Biol Chem.* 2004;279(27):27941–7. doi:10.1074/jbc.M403240200.

[227] Ariza MJ, Sánchez-Chaparro MA, Barón FJ, Hornos AM, Calvo-Bonacho E, Rioja J, et al. Additive effects of *LPL*, *APOA5* and *APOE* variant combinations on triglyceride levels and hypertriglyceridemia: results of the ICARIA genetic sub-study. *BMC Med Genet.* 2010;11:66. doi:10.1186/1471-2350-11-66.

[228] De Castro-Orós I, Cenarro A, Tejedor MT, Baila-Rueda L, Mateo-Gallego R, Lamiquiz-Moneo I, et al. Common genetic variants contribute to primary hypertriglyceridemia without differences between familial combined hyperlipidemia and isolated hypertriglyceridemia. *Circ Cardiovasc Genet.* 2014;7(6):814–21. doi:10.1161/CIRCGENETICS.114.000522.

[229] Burgess S, Bowden J, Dudbridge F, Thompson SG. *Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization.* arXiv; 2016. arXiv:1606.03729 [stat.ME]. Available from: `http://arxiv.org/abs/1606.03729`.

[230] Foley CN, Staley JR, Dudbridge F, Howson JMM. Test to identify co-localization of genetic association signals across multiple traits using summary statistics. *Genet Epidemiol.* 2016;40(7):613–614. doi:10.1002/gepi.22001.

[231] Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials*. 2014;15:363. doi:10.1186/1745-6215-15-363.

[232] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*. 2015;47(3):284–90. doi:10.1038/ng.3190.

APPENDIX B

Relevant publications and conferences

B.1 List of publications authored and co-authored during PhD

1. **Harshfield EL**, Stacey D, Paul DS, Fauman EB, Ziemek D, Koulman A, Griffin JL, Wood AM, Butterworth AS, Danesh J & Saleheen D. (2018). The PROMIS of lipid biomarkers in human cardiovascular disease: analysis of genetic determinants of lipids identifies novel associations between lipids and metabolic disease-related loci. *Manuscript in preparation.*

2. **Harshfield EL**, Davidson KW, Shaffer JA, Pennells L, Kaptoge S, [29 other co-authors], Danesh J & Di Angelantonio E, for the Emerging Risk Factors Collaboration/UK Biobank Depression Study Group. (2018). Association of depressive symptoms with risk of cardiovascular diseases and cause-specific mortality: individual participant meta-analysis of 24 prospective studies. *Manuscript in preparation.*

3. **Harshfield EL**, Koulman A, Ziemek D, Fauman EB, Marney L, Paul DS, Stacey D, Rasheed A, Lee J-J, Shah N, Jabeen S, Imran A, Abbas S, Majeed F, Qamar N, Mallick N, Yaqoob Z, Saghir T, Rizvi SNH, Memon F-u-R, Qureshi IH, Ishaq M, Frossard P, Danesh J, Saleheen D, Butterworth AS, Wood AM & Griffin JL. (2018). An unbiased lipid phenotyping approach to study the genetic determinants of lipids and their association with coronary heart disease risk factors. *Manuscript under review.*

4. van der Laan SW, **Harshfield EL**, Hemerich D, Stacey D, Wood AM, Asselbergs FW. (2018). From lipid locus to drug target through human genomics. *Cardiovasc Res.* Published online 23 May 2018 [Epub ahead of print]. doi:10.1093/cvr/cvy120.

5. Stacey D, Fauman EB, Ziemek D, Sun BB, **Harshfield EL**, Wood AM, Butterworth AS, Suhre K, Paul DS. (2017). ProGeM: a framework for the prioritisation of candidate causal genes at molecular quantitative trait loci. *bioRxiv* 230094. doi:10.1101/230094.

6. **Harshfield EL**, Stacey D, Paul DS, Koulman A, Wood AM, Butterworth AS, Fauman E, Griffin JL, Danesh J & Saleheen D. (2016). Genomics of lipid metabolism: identifying novel causal pathways and new therapeutic targets for redu-

cing risk of coronary heart disease [Abstract]. *Genet Epidemiol*, *40*(7), 614-615. doi:10.1002/gepi.22001.

7. Burgess S & **Harshfield E** (2016). Mendelian randomization to assess causal effects of blood lipids on coronary heart disease: lessons from the past and applications to the future. *Curr Opin Endocrinol Diabetes Obes*, *23*(2): 124-130. doi:10.1097/MED.0000000000000230.

8. **Harshfield E**, Chowdhury R, Harhay MN, Bergquist H & Harhay MO. (2015). Association of hypertension and hyperglycaemia with socioeconomic contexts in resource-poor settings: the Bangladesh Demographic and Health Survey. *Int J Epidemiol*, *44*(5): 1625-1636. doi:10.1093/ije/dyv087.

9. White IR, Rapsomaniki E, Emerging Risk Factors Collaboration [**Harshfield E** was one of 207 collaborators]. (2015). Covariate-adjusted measures of discrimination for survival data. *Biometrical J*, *57*(4):592-613. doi:10.1002/bimj.201400061.

10. Chowdhury R, [19 other co-authors], **Harshfield E**, [8 other co-authors], Danesh J & Di Angelantonio E. (2015). The Bangladesh Risk of Acute Vascular Events (BRAVE) Study: objectives and design. *Eur J Epidemiol*, *30*(7): 577-587. doi:10.1007/s10654-015-0037-2.

11. Freitag DF, [12 other co-authors], **Harshfield E**, [152 other co-authors] & Danesh J. (2015). Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: a Mendelian randomisation analysis. *Lancet Diabetes Endocrinol*, *3*(4):243-253. doi:10.1016/S2213-8587(15)00034-0.

12. Chowdhury R, **Harshfield E**, Roy S, Flora MS, Akram K, Bhuiya A, & Ahsan H. (2014). Life, health, and safety of industrial workers in Bangladesh: should they be driven by economic rationale or moral imperative? *J Occup Environ Med*, *56*(4):e12-e13. doi:10.1097/JOM.0000000000000126.

13. Emerging Risk Factors Collaboration [**Harshfield E** was one of 189 collaborators], [73 other co-authors] & Danesh J. (2014). Glycated hemoglobin measurement and prediction of cardiovascular disease. *JAMA*, *311*(12):1225-1233. doi:10.1001/jama.2014.1873.

B.2 Conferences presented research at during PhD

1. Stacey D, Fauman EB, Ziemek D, Sun BB, **Harshfield EL**, Wood AM, Butterworth AS, Suhre K, Paul DS. (6-9 Sep 2017). ProGeM: a general framework and tool for the prioritisation of candidate causal genes at molecular QTLs [poster presentation by Stacey D]. *The Genomics of Common Diseases.* Wellcome Genome Campus, Hinxton, Cambridge, UK.

2. **Harshfield E**. (04 Jul 2017). Genomics of a detailed profile of soluble lipids [poster presentation]. *Big Data in Medicine: Tools, Transformation and Translation.* University of Cambridge. Cambridge, UK.

3. **Harshfield E**. (22 Nov 2016). Genomics of a detailed profile of soluble lipids [poster presentation]. *British Heart Foundation Site Visit for Quinquennial Review.* University of Cambridge. Cambridge, UK.

4. **Harshfield E**. (25 Oct 2016). Genomics of lipid metabolism: Identifying novel causal pathways and new therapeutic targets for reducing risk of coronary heart disease [oral presentation]. *International Genetic Epidemiology Society 2016 Annual Meeting.* Toronto, Canada.

5. **Harshfield E**. (29 Jun 2016). Genomics of lipid metabolism: Identifying novel causal pathways and new therapeutic targets for reducing risk of coronary heart disease [poster presentation]. *Wellcome Trust Site Visit for Proposed Translational Centre for Global Ageing.* University of Cambridge. Cambridge, UK.

6. **Harshfield E**. (28 Nov 2014). Lipid metabolites and risk of coronary heart disease [oral presentation]. *UK MEG 2014 Workshop: Metabonomics in Molecular Epidemiology.* Imperial College London. London, UK.

7. Marney L, Richardson L, Koulman A, **Harshfield E**, Forouhi N & Griffin J. (2-5 Sept 2014). High-throughput metabolomics in the epidemiological study of metabolic disease [poster presentation by Marney L]. *The Association for Mass Spectrometry: Applications to the Clinical Lab (MSACL) 2014 EU.* Salzburg, Austria.

8. Topi G, Chowdhury R, **Harshfield E**, Darweesh SKL, Bautista PK, Voortman T, Moreira EM, Bramer WM, Van Den Hooven EH, Franco OH. (8-10 May 2014). Associations of air pollution and ambient temperature with the risk of stroke: a systematic review and meta-analysis [poster presentation by Topi G]. *EuroPRevent 2014.* Amsterdam, The Netherlands.

APPENDIX C

PROMIS questionnaire

Pakistan Risk Of Myocardial Infarction Study (PROMIS)

Inclusion criteria for Cases

- First time ever acute MI □
- Age range (30-80 years) □
- 1 mm or more ST elevation in any two or more contiguous limb leads with Trop.Positive □
- Non ST elevation MI with Trop. Positive □
- Troponin levels □
- New onset LBB with Trop. positive (previous ECG req) □

Exclusion criteria for Cases

- Onset of chest symptoms and hospitalization for MI more than 24 hours □
- Presence of cardiogenic shock □
- Presence of chronic conditions (malignancy, , infection such as TB, Malaria, hepatitis, renal failure, leprosy and etc.) □
- Viral or bacterial infection in past 2 weeks □
- Pregnancy □
- Any prior cardiac event □
- Failure to give Informed consent □

Source of Controls

1. Attendants and visitors/relatives of patients presenting with non-cardiac reasons □
2. unrelated (not first-degree relative) visitors of a cardiac patient □
3. individuals undergoing routine health checkup □
4. Refraction and Cataracts patients □
5. Minor ear, nose and throat patients □
6. Individuals undergoing elective minor surgery (skin disorders, orthopaedic surgery, haemorrhoids and hernia) □

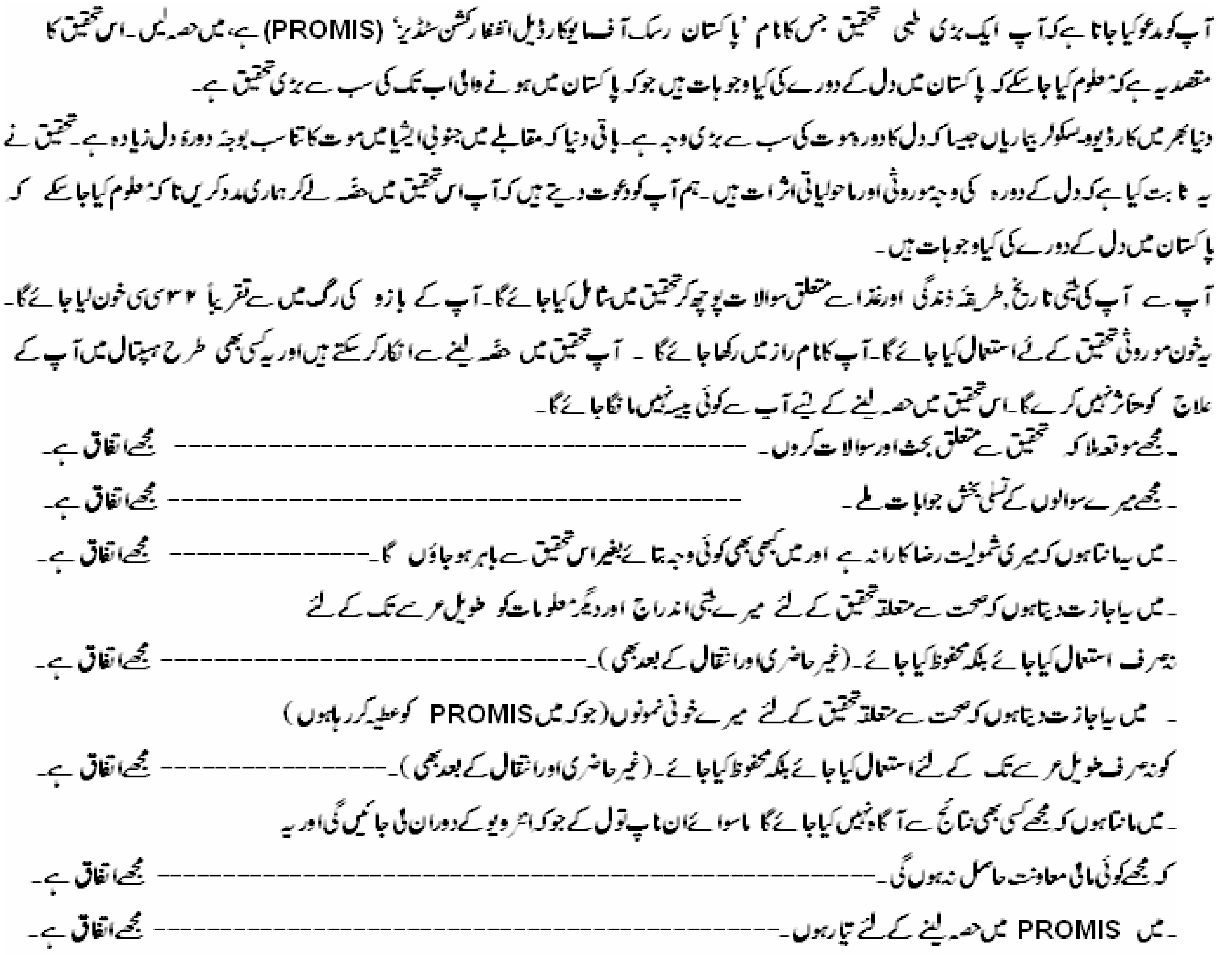
Exclusion criteria for Controls

- Any prior cardiac event □
- Pregnancy □
- Presence of chronic conditions (malignancy, infection such as TB, Malaria, hepatitis, renal failure, leprosy, TIA, Stroke and etc.) □
- Viral or bacterial infection in past 2 weeks □
- Onset of chest symptoms and hospitalization for MI more than 24 hours □
- Failure to give informed consent □

Check every option appropriately

FOR DATA ENTRY ONLY:

	Date	**Name and Signature**
Entry of selected variables		
Complete entry (1st time)		
Complete entry (2nd time)		
Data rechecked		



آپ کو مدعو کیا جاتا ہے کہ آپ ایک بڑی طبی تحقیق جس کا نام 'پاکستان رسک آف مایوکارڈیل انفارکشن سٹڈیز' (PROMIS) ہے، میں حصہ لیں۔ اس تحقیق کا مقصد یہ ہے کہ معلوم کیا جا سکے کہ پاکستان میں دل کے دورے کی کیا وجوہات ہیں جو کہ پاکستان میں ہونے والی اب تک کی سب سے بڑی تحقیق ہے۔

دنیا بھر میں کارڈیوویسکولر بیماریاں جیسا کہ دل کا دورہ موت کی سب سے بڑی وجہ ہے۔ باقی دنیا کہ مقابلے میں جنوبی ایشیا میں موت کا تناسب بوجہ دورۂ دل زیادہ ہے۔ تحقیق نے یہ ثابت کیا ہے کہ دل کے دورہ کی وجہ موروثی اور ماحولیاتی اثرات ہیں۔ ہم آپ کو دعوت دیتے ہیں کہ آپ اس تحقیق میں حصہ لے کر ہماری مدد کریں تاکہ معلوم کیا جا سکے کہ پاکستان میں دل کے دورے کی کیا وجوہات ہیں۔

آپ سے آپ کی طبی تاریخ، طریقۂ زندگی اور غذا سے متعلق سوالات پوچھ کر تحقیق میں شامل کیا جائے گا۔ آپ کے بازو کی رگ میں سے تقریباً ۳۲ سی سی خون لیا جائے گا۔ یہ خون موروثی تحقیق کے لئے استعمال کیا جائے گا۔ آپ کا نام راز میں رکھا جائے گا ۔ آپ تحقیق میں حصہ لینے سے انکار کر سکتے ہیں اور یہ کسی بھی طرح ہسپتال میں آپ کے علاج کو متاثر نہیں کرے گا۔ اس تحقیق میں حصہ لینے کے لیے آپ سے کوئی پیسہ نہیں مانگا جائے گا۔

۔ مجھے موقع ملا کہ تحقیق سے متعلق بحث اور سوالات کروں۔ -- مجھے اتفاق ہے۔

۔ مجھے میرے سوالوں کے تسلی بخش جوابات ملے۔ -- مجھے اتفاق ہے۔

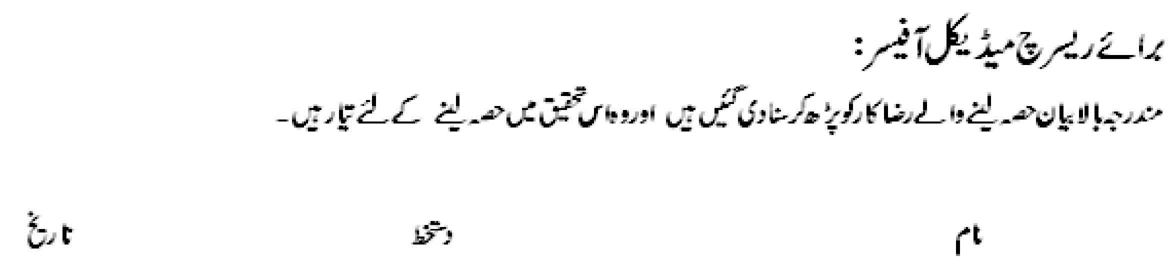
۔ میں یہ مانتا ہوں کہ میری شمولیت رضاکارانہ ہے اور میں کبھی بھی کوئی وجہ بتائے بغیر اس تحقیق سے باہر ہو جاؤں گا۔--------------- مجھے اتفاق ہے۔

۔ میں یہ اجازت دیتا ہوں کہ صحت سے متعلق تحقیق کے لئے میرے طبی اندراج اور دیگر معلومات کو طویل عرصے تک کے لئے
نہ صرف استعمال کیا جائے بلکہ محفوظ کیا جائے۔ (غیر حاضری ورانتقال کے بعد بھی)۔---------------------------------- مجھے اتفاق ہے۔

۔ میں یہ اجازت دیتا ہوں کہ صحت سے متعلق تحقیق کے لئے میرے خونی نمونوں (جو کہ میں PROMIS کو عطیہ کر رہا ہوں)
کو نہ صرف طویل عرصے تک کے لئے استعمال کیا جائے بلکہ محفوظ کیا جائے۔ (غیر حاضری اور انتقال کے بعد بھی)۔---------------- مجھے اتفاق ہے۔

۔ میں مانتا ہوں کہ مجھے کسی بھی نتائج سے آگاہ نہیں کیا جائے گا ماسوائے ان ناپ تول کے جو کہ انٹرویو کے دوران لی جائیں گی اور یہ
کہ مجھے کوئی مالی معاونت حاصل نہ ہوں گی۔-- مجھے اتفاق ہے۔

۔ میں PROMIS میں حصہ لینے کے لئے تیار ہوں۔--- مجھے اتفاق ہے۔



برائے ریسرچ میڈیکل آفیسر:

مندرجہ بالا بیان حصہ لینے والے رضاکار کو پڑھ کر سنادی گئیں ہیں اور وہ اس تحقیق میں حصہ لینے کے لئے تیار ہیں۔

نام　　　　　　　　　　دستخط　　　　　　　　　　تاریخ

English translation:

You are being invited to take part in a major medical research project called the “Pakistan Risk of Myocardial Infarction Study” (PROMIS). The aim of this study is to help better understand the causes of heart attacks in Pakistan, and it is the biggest such study ever done in Pakistan.

Cardiovascular diseases such as heart attacks are the leading cause of deaths worldwide. Deaths due to heart attacks are higher in South Asia than in any other part of the world. Research has shown that heart attacks are caused by a combination of genetic and environmental factors. We invite you to participate in this study to help us better understand the causes of heart attacks in the Pakistani population.

Your participation in this research will involve you being asked some questions about your medical history, lifestyle and your diet. About 32cc of blood will be collected from a vein in your arm. This blood will be used for conducting genetic research and measuring markers in your blood. Your name will be kept confidential. You can refuse to participate in this research and this will not by any means affect your treatment at the hospital. You will not be asked to pay any money if you agree to participate in our research.

I have had the chance to discuss and ask any questions about the research	**I agree**
I have received satisfactory answers to my questions	**I agree**
I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason	**I agree**
I give permission for access to my medical records, and for long-term storage and use of this and other information about me for purposes of health-related research (even after my incapacity or death)	**I agree**
I give permission for long-term storage and use of my blood samples for health-related research (even after my incapacity or death), and I relinquish all rights to these samples which I am donating to PROMIS	**I agree**
I understand that, except for some measurements taken during this interview, none of my results will be given to me and that I will not benefit financially in anyway from taking part	**I agree**
I agree to take part in PROMIS	**I agree**

For the Research Medical Officer:
The above statement has been read to the participant. He/she has agreed to each of the above statements and has agreed to participate in this research

Name: Signature: Date:

Date of Interview □□ □□ □□ **Interviewer's Name:** ______________
day *month* *year*

1. Study ID# □ □□□□ **2. Status:** □ **Control** □ **Case**

3. Centre _____________ **Hospital ID.** ______________

Shifted to Ward: _____________ **Bed Number**______________

4. Name____

5.Address____

____

6. Patient's Mobile #: _______________________________**Patient's Land Line #:** _________________________

Pt.'s Relative Mobile #:______________________________ **Pt.'s Relative Land Line #:**_______________________

7. Gender: Male (1) □ Female (2) □ **DOB:** □□ □□ □□ or **8.Age** _______

9. Onset of Symptoms Date □□ □□ □□ Time **:** □□ □□
day *month* *year* *hour* *mm*

10. Arrived at hospital Date □□ □□ □□ Time **:** □□ □□

day *month* *year* *hour* *mm*

11. Last meal Date □□ □□ □□ Time : □□ □□

day *month* *year* *hour* *mm*

12. Sampling Date □□ □□ □□ Time : □□ □□

13. ER- Thrombolysis time □□ □□ **14. Thrombolytic infusion:** Yes □ No □ NA □
hour *mm*

15. Sample taken within 24 hours of symptoms Yes □ No □

16. Source of Information: Patient □ Attendant □

	No	**Yes**
17. Aspirin Intake	□	□
18. Nitroglycerine Intake	□	□
19. Did the patients symptoms resolve on aspirin intake?	□	□
20. Did the patients symptoms resolve on nitroglycerine intake?	□	□

ECG Changes

	No	**Yes**
21) ST segment elevation in two or more contiguous leads with reciprocal changes > 1mm.	□	□
22) New onset left bundle branch block	□	□
23) New pathological Q waves	□	□
24) Any other ECG changes______________________________		

25) S in lead V1+ R in V5/ V6 = > 35 mm Y □ **N** □

26)SV3 + R avl > 28 mm in men Y □ **N** □

27) SV3 + R avl > 20 mm in women Y □ **N** □

28) R avl > 11mm **Y** □ **N** □ **R V4-6 > 25mm** **Y** □ **N** □

29) S V 1-3 > 25 mm **Y** □ **N** □ **R I + S III > 25 mm** **Y** □ **N** □

30) R V5 or V6 > 35 mm **Y** □ **N** □ **S V1 or V2** **Y** □ **N** □

31) Type of MI (1) Anterior □ (2) Antero-Septal □ (3) Inferior □ (4) Lateral □ (5) Posterior □
(6) Right ventricle □ (7) NSTEMI □

32) Reversal of symptoms on streptokinase infusion Y □ N □

33) Reversal of ST- segment elevation by > 50% on streptokinase infusion Y □ N □

34) Troponin Y □ N □ NA □

35) Medications (for non-hospital controls, complete the Pre-Admission section only)

	Pre Admission		In Hospital		Duration Pre-admission				Pre Admission		In Hospital		Duration Pre-admission	
	No	Yes	No	Yes	days	months			No	Yes	No	Yes	days	months
a) ACE-i	□	□	□	□	□	□	i) Heparin	→ LDH			□	□	□	□
								→ FDH			□	□	□	□
b) ASA	□	□	□	□	□	□	j) Hormones		□	□	□	□	□	□
c) Beta Blocker	□	□	□	□	□	□	k) Insulin		□	□	□	□	□	□
d) Ca Channel Blocker	□	□	□	□	□	□	l) Nitrates		□	□	□	□	□	□
e) Cholesterol ↓	□	□	□	□	□	□	m) Oral Hypoglycemics		□	□	□	□	□	□
f) Digoxin	□	□	□	□	□	□	n) Thrombolysis				□	□	□	□
g) Thiazides	□	□	□	□	□	□	o) Warfarin		□	□	□	□	□	□
h) Diuretics	□	□	□	□	□	□	p) Platelet antagonist		□	□	□	□	□	□
Homeopathic	□	□	□	□	□	□	q) Angiotensin Receptor Blocker		□	□	□	□	□	□

[redacted]

36) Past medical history ____

____

	HTN	Age of diag.	DM	Currently on Ins? (Y/N)	Approx. age of diag.	Angina	MI	Approx. age of diag.	Stroke	Approx. age of diag.	CLD	CA	Site	Sudden Death
Subject	□	□	□	□	□□	□	□	□□	□	□□	□	□	□	
Mother	□	□	□	□	□□	□	□	□□	□	□□	□	□	□	□
Father	□	□	□	□	□□	□	□	□□	□	□□	□	□	□	□
Sister(s)	□	□	□	□	□□	□	□	□□	□	□□	□	□	□	□
Brother(s)	□	□	□	□	□□	□	□	□□	□	□□	□	□	□	□
Son(s)	□	□	□	□	□□	□	□	□□	□	□□	□	□	□	□
Daughter(s)	□	□	□	□	□□	□	□	□□	□	□□	□	□	□	□

37) Do you feel pain in your lower extremities while walking? No □ Yes □
(only Calf muscles).

38) Does this pain go away at rest immediately? No □ Yes □

38b) If yes how long does it take for the pain to get relieved ________ minutes

39) Does this pain occur every time when you walk? No □ Yes □
(0) (1)

Does your heart beat increase or become irregular quite often? No □ Yes □

Have you ever been diagnosed of atrial fibrillation? No □ Yes □

40) Past Surgical history (most recent)

Procedure Performed______________________ Year of Surgery ________________

TOBACCO USAGE

41) Which best describes subject's history of tobacco use?

□ I) Ever used tobacco Product 2)Currently uses tobacco products □ 3) Never used tobacco products □

42) Does/did subject use any of the following tobacco products? (check all that apply)

Cigarettes (1) □ Beedies (2) □ Huqqa/ Chilum □ (3) Paan (4) □ Naswar (5) □ Gutka □ (6) Supari (7) □

(check pann, gutka and Supari only if they are used in a tobacco form)

43) If yes to Cigarettes and Beedies ⟶ **44)** How many per day? □□□

44) Type of cigarettes? (*Check one only*) 1)Filter □ 2)Non-Filter □ 3)Both □

45) What brand of cigarettes does/did subject most commonly smoke? ______________________________

46) How many years patient has smoked **47) At what age did subject quit?** □□ **Yrs N/A** □

48) Has subject smoked > 100 cigarettes/ beddies in your life? No □ **Yes** □

49) If yes to Huqqa/Chilum, how many years patient has smoked □□

50) At what age did subject quit Huqqa/Chilum □□ **yrs N/A** □

51) If yes to Paan ⟶ 1) Tambako □ 2) Non-Tambako □

52) How many per day? □□□

53) At what age did subject start? □□ **Yrs** **54) At what age did subject quit?** □□ **Yrs N/A** □

55) If yes to Gutka ⟶ Tambako □ Non-Tambako □

56) How many per day? □□□

57) At what age did subject start? □□ **Yrs** **58) At what age did subject quit?** □□ **Yrs N/A** □

59) If yes to Supari such as *"city"* → Tambako □ Non-Tambako □

60) How many per day? □□□

61) At what age did subject start? □□ **Yrs** **62) At what age did subject quit?** □□ **Yrs** **N/A** □

63) If yes to Naswar → Black □ Green □ Kashmiri □

64) How many packets per day? □□□

65) At what age did subject start? □□ **Yrs** **66) At what age did subject quit?** □□ **Yrs** **N/A** □

66) Are you exposed to other people's smoke?
1. Never □ 2. Work Place □ 3. Home □ 4. With Friends □

67) Has subject had an acute febrile illness within the previous six months?

1. >4 days □ 2. Required antibiotics regularly □ 3. Required hospitalization □ 4.None. □

68) How many jobs does subject currently have? (1) **1** □ (2) **2+** □ (3) Unemployed □ (4) Retired □ (5) N/A □

69) Which category below best describes subject's main occupation? (Select one only)

(1) Professional □ (2) Skilled Labour □ (3) General Labour □ (4) Housewife □ (5) Farmer □

(6) Business □ (7) Clerical □ (8) Self employed □ (9) Other *(specify)* ___ ____________

PHYSICAL ACTIVITY

70. At work: (1) Light □ **(2) Moderate** □ **(3) Active** □ **(4) N/A** □

Key: **light (**physically very easy, sitting office work or sitting shop work or secretary)
Moderate (Standing and walking, eg. Store assistant, light industrial worker)
Active (walking and lifting, heavy manual labor eg. mazdoor, high weight industrial worker, Farmer)

71. Daily Commuting:
a) Using motorized transport: (1) Personal Car □ (2) Motor Bike □ (3) Bus □ (4) Reksha/Taxi □ (5) or no walk or cycling □
b) Walking or bicycling 1 to 29 minutes □ (6)
c) Walking or bicycling > 30 minutes □ (7) **d)** N/A □ (8)

72. Leisure Time: (1) **Low** □ (2) **Moderate** □ (3) **Heavy** □
Key: **Low** (almost completely inactive eg reading, watching TV, doing some minor physical activity)
Moderate (some physical activity for > 4 h/week eg. Walking, cycling, light gardening but excluding travel to work)
High (vigorous physical activity for > 3 h/week, eg. Running, jogging, swimming, weight lifting)

73. In case of exercise → 1. Aerobic □ 2. Anaerobic □ (weight lifting or sprinting)
74. If aerobic:
Type I – 15-20 min. 3x week (hiking, jogging, running, treadmill, stationary cycling) □

Type II -- 30 min. 3x week (bicycling, swimming, tennis, walking briskly) □

Type III cricket, football, badminton, volley ball. □

1. Sindh, 2. Punjab, 3. Serhad, 4. Balochistan
5. Other

75. Place of birth: ______________ ______________
Province *Country*

Ethnicity **76)** Subject# □ **77)** Mother# □ **78)** Father# □

1. Urdu 5. Sindhi
2. Punjabi 6. memon
3. Pathan 7. Gujrati
4. Balochi 8. Others

79. Were Subject's father and mother first degree relatives? No □ Yes □

80. Is subject's spouse his/her first degree relative? No □ Yes □

81) Religion: 1. Islam □ 2. Hinduism □ 3. Christianity □ 4. Sikhism □ 5. Athiest □
6. Others □

82) Do you practice religion regularly? (Incase of Islam, do you pray namaz regularly) No □ Yes □

The number of years of formal education completed (check highest level only):

83) Subject None □ Number of years of formal education □ Number of years of Madarsah education □

84) Mother None □ Number of years of formal education □ Number of years of Madarsah education □

85) Father None □ Number of years of formal education □ Number of years of Madarsah education. □

86) Marital Status

1. Single □ 2. Married □ 3. Divorced □ 4. Separated □ 5. Widow □

87) How many children do you have __________________

88) Has the subject experienced any of the following in the past year?

1. Marital separation/Divorce □ No □ Yes 4.Major personal injury or illness □ No □ Yes

2. Loss of job/retirement □ No □ Yes 5.Death/major illness of a close family member □ No □ Yes

3. Loss of crop □ No □ Yes 6. Death of a spouse □ No □ Yes

88) Monthly income of the Family:

89) How many household members (including children) are there:

90) Note which of the following objects are <u>owned</u> by household members: (check all that apply)

(1) Home □ (2) Car/Auto □ (3) Motorcycle/Scooter □ (4) Bicycle □ (5) Radio/Stereo □ (6) Television □
(7) Other Land/Property □ (8) Computer □ (9) Livestock/Cattle □ (10) Mobile □ (11) AC □ (12) Number of servants ___

91) What level of stress does the subject feel:

			(1) None/Mild	**(2)** Moderate	**(3)** High/Severe	**(4)** N/A
	a) <u>At work:</u>	i) Mental/Emotional:	□	□	□	□
		ii) Physical:	□	□	□	□
	b) <u>At home:</u>	i) Mental/Emotional:	□	□	□	
		ii) Physical	□	□	□	
	c) <u>Financial:</u>		□	□	□	

For Women Only (Questions 92-99)

92. Has subject ever used birth control pills or had birth control depot injections?

□ Yes □ No → **93.** Age when began □□ years old **95.** Brand: ________________

→ **94.** For how many years? □□ Years

96. Has the subject had a menstrual period in the last 12 months?

□ Yes □ No → **97.** At what age did they stop? □□ Years

→ **98.** Why did they stop? Menopause □ (1) Hysterectomy □ (2) Radiation □ (3)

(check only one) □ 4.Other (specify)____________________________

99. Has subject ever used female hormone replacements?

□ Yes □ No → **100.** How long? □□ Years → **101.** Are you still taking them? yes □ no □

→ **102.** Which type of replacement have you used the most?

□ Estrogen alone □ Estrogen + Progesterone

103. Type of oil/fat used most often in cooking.

□ Desi Ghee (1) □ Ghee (2) □ Oil (3)

□ Butter (4) □ margarine (5)

104. Note down the brand of ghee/oil _____________ (note down separately)

105a. How much ghee is bought per month for cooking purposes ____________

105b. How much oil is bought per month for cooking purposes ______________

106. For How many people (aged 6 years and above) it is used for preparing food _____________.

107. During the past 12 months, has subject had a drink of beer, wine, liquor or any other alcoholic beverage?

□ No □ Yes ***How often*** → **108.** □ once a year **(1)** □ once a month **(2)** □ 2-3 times a month **(3)** □ once a week **(4)**

(mark one box only) □ 2-3 times a **(5)** □ 4-6 times a **(6)** □ everyday **(7)**

Food Item	**1) > 4 per day**	**2) 2-4 per day**	**3) 1/ day**	**4) > 3 times/ week**	**5) 2-3 times/ week**	**6) 1/ week**	**7)< a week but once or more /month**	**8) 1/ mont h**	**9) <1/ month or occasio -nal**	**10) In ramaz-an only**	**11) none**
1. Tea											
2. Coffee											
3. Qahwa											
4. Herbal tee											

10. Do you add extra salt to tea/Qahwa/lassi? No □ Yes □

111. Do you add extra salt to food? No □ Yes □

Food Item	1) >4/ day	2) 2-4 per day	3) 1/ day	4) > 3 times/ week	5) 2-3 times/ week	6) 1/ week	7)< a week but once or more /month	8) 1/ month	9) <1/ month or occasi-onal	10) In ramaz -an only	11) none
1. Khameri or Nan											
2. Paratha/Puri											
3. Roti (Safed Aaata)											
4. Roti (Laal Aaata/chakki)											
5. Roti (mixed aata)											
6. Daliya											
7. Eggs											
8. Dark green leafy vegetables and yellow Vegetables **(COOKED)**											
9. Cruciferous vegetables (Gobi, phool gobi, band gobi, sursoon, others) **(COOKED)**											
10. Other vegetables excluding potatoes **(COOKED)**											
11. Dark green leafy vegetables and yellow Vegetables **(SALAD)**											
12. Cruciferous vegetables (Gobi, phool gobi, band gobi, sursoon, others) **(SALAD)**											
13. Other vegetables excluding potatoes **(SALAD)**											
14. Cooked Potatoes											
15. Fried potatoes, French fries											

Food Item	1) >4/ day	2) 2-4 per day	3) 1/ day	4) > 3 times/ week	5) 2-3 times/ week	6) 1/ week	7)< a week but once or more /month	8) 1/ month	9) <1/ month or occasi o-nal	10) In ramaz -an only	11) none
16. Tomatoes as Salad											
17. Onion raw											
18. Rice											
19. Chicken Biryani											
20. Beef or Mutton Biryani											
21. Daal, lobia, channa, cholay											
22. Fruits											
23. Fruit juice											
24. Kata kut or organ meet											
25. Beef Salan											
26. Mutton Salan											
27. Chicken Salan											
28. Fish salan											
29.Beef boti, tikka, kabab, Beef shawarma and others											
30.Chicken boti, tikka, kabab, chicken roll, chicken shawarma and others											
31. Chicken Fried, Broast											
32. Fried fish											
33. Paya											
34.Nehari											
35. Pakoray, other basen products											
36. Samosay											
37. Burger, Bun kebab											
38. Wisky											

Food Item	**1) >4/ day**	**2) 2-4 per day**	**3) 1/ day**	**4) > 3 times/ week**	**5) 2-3 times/ week**	**6) 1/ week**	**7)< a week but once or more /month**	**8) 1/ month**	**9) <1/ month or occasi-onal**	**10) In ramaz -an only**	**11) none**
39. Beer											
40. Bhang											
41. Tharra (desi sharab)											
42. Milk											
43.Makkhan, Margarine.											
44. Other dairy products (Lassi, Curd)											
45. Sugars (Gurh, Shakkar, honey, Jam, marmalade)											
47. Carobonated beverages.pepsi											
48. NonCarbonated beverages (rooh afza)											
49. Pickels.											
50. Bakery Items (Bagarkhani, Papay, cakes, biscuits others)											
51. Kheer, Custard, Milk shakes and other milk based sweet dishes, ice cream											
52. Halwa, Mithai, Jalabe chocolate											
53. Nimko and other fried items											
54. Nuts/Seeds											

112) Do you have any addictions? □No Yes □

113) If yes please describe________________

For how long have you been taking? ____________

[redacted]

Physical Measurements		**First Reading**	**Second Reading**
Blood Pressure	**114)** systolic	□□□	□□□
	115) Diastolic	□□□	□□□

	116) Waist (cm)		**117) Hip (cm)**		
#1	□□□	**#1**	□□□	**118) Weight**	□□□ **.** □ **kg**
#2	□□□	**#2**	□□□	**119) Height**	□□□**.** □**cm**

120) Heart Rate (beats/minute) □□□

121) QT Interval __________

Sample Information:

Sample ID#: □ **-** □□□□

Blood Sample collected: **YES** □ **NO** □

No. of EDTA tubes □

No. of Serum tubes □

Shifting date of samples to AKUH ____________________

Time of Departure at the collection center ________________

Time of delivery at AKUH __________________