# Discovery and Application of Genetic Determinants of Cardiovascular Disease Risk Factors

SONIA HARKCHAND SHAH

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF GENETICS, EVOLUTION AND ENVIRONMENT
UNIVERSITY COLLEGE LONDON
December 2013

# Declaration

I, Sonia Harkchand Shah, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been clearly indicated.

Sonia Harkchand Shah

# Abstract

The focus of my PhD has been two-fold:

First, to improve the understanding of the biology behind a well-known cardiovascular disease (CVD) risk factor - left ventricular mass, by identifying novel genetic loci associated with this risk factor. A large-scale association meta-analysis in over 10,000 individuals identified four novel loci associated with electrocardiographically-determined left ventricular mass.

Second, to explore the application of known genetic determinants of the main blood lipid fractions, the latter being well-known CVD risk factors and therapeutic targets. I assess the use of genetic variants associated with total cholesterol, low-density lipoprotein-cholesterol (LDL-C), high-density lipoprotein-cholesterol (HDL-C) and triglycerides for discriminating healthy individuals from those that have a high absolute risk of CVD, those that require lipid-lowering medication, and those that have a coronary event. The lipid genetic variants showed poor discriminatory ability for all three outcomes and provided no improvement over the widely-used, non-genetic Framingham 10 year CVD risk score. Lipid-associated genetic variants were also used to generate genetic risk score instruments for LDL-C, HDL-C and triglycerides, which were applied in a Mendelian randomisation analysis to determine their causal relationship with carotid-intima media thickness (CIMT). CIMT has been a widely used surrogate outcome measure in clinical trials of CVD drugs. LDL-C-lowering drugs have shown to reduce CIMT progression and CHD risk in clinical trials. However, the extent of any causal association between HDL-C or triglycerides and CIMT is unclear. The results from this MR analysis support a casual relationship with LDL-C, but not with HDL-C and triglycerides, which may indicate that CIMT is a less useful surrogate end point in clinical trials of primarily HDL-C or triglyceride modifying therapies.

# Statement of Collaboration

The work in chapter 2 was carried out in collaboration with several cohorts. Analysis in the British Women's Health and Heart Study (BWHHS) and Genetic Regulation of Arterial Pressure of Humans in the Community study (GRAPHIC) was done by Dr. Tom Gaunt and Dr. Chris Nelson, respectively. Analysis in each of the replication cohorts was carried out by the study-specific analyst. All other statistical analysis was carried out independently and in parallel by me and Dr. Nelson. This work has been published in: Shah, S., Nelson, C. P., Gaunt, T. R., Van der Harst, P., Barnes, T., Braund, P. S., Lawlor, D. A., et al. (2011). Four genetic loci influencing electrocardiographic indices of left ventricular hypertrophy. *Circulation Cardiovascular Genetics, 4*(6), 626–635.

In chapter 3, analysis of BWHHS data was carried out by Dr. Tom Gaunt using R and STATA scripts provided by me. Calculation of the Framingham 10 year CVD risk score in both studies was carried out using a STATA script written by Dr. Jackie Cooper. This work has been published in: Shah, S., Casas, J. P., Gaunt, T. R., Cooper, J., Drenos, F., Zabaneh, D., Swerdlow, D. I., et al. (2013). Influence of common genetic variation on blood lipid levels, cardiovascular risk, and coronary events in two British prospective cohort studies. *European Heart Journal, 34*(13), 972-981.

In chapter 5, quality control analysis of IMPROVE genotype and phenotype data was performed by Dr. Delilah Zabaneh and Dr. Karl Gertow. This work has been published in: Shah, S., Casas, J.-P., Drenos, F., Whittaker, J., Deanfield, J., Swerdlow, D. I., Holmes, M. V, et al. (2012). Causal relevance of blood lipid fractions in the development of carotid atherosclerosis: Mendelian randomisation analysis. *Circulation Cardiovascular Genetics*, *6*(1), 63-72.

# Acknowledgements

First and foremost, I would like to thank Aroon D. Hingorani and Steve E. Humphries for giving up their time to be my supervisors and mentors at University College London. They have provided me with great insight into the field of Cardiovascular Genetics, and without their support, guidance and constructive critiques and recommendations, this thesis would not be the same.

Second, I would like to thank everyone who has been part of the WHII-50K group, UCLEB, and the UCL Genetics Institute, especially Delilah Zabaneh, Jon White, Fotios Drenos and John Whittaker. They have been a fantastic support network that has allowed me to seek unlimited advice.

Last but not least, I would like to thank my family and friends for all their support over the last few years. Gordon, thank you for all your patience and support. This PhD would have taken a lot longer if it wasn’t for you.

# Contents

# Abbreviations

| | |
|---|---|
| 2SLS | Two-Stage Least Squares |
| ACE | Angiotensin Converting Enzyme |
| AGT | Angiotensinogen |
| AGTR1 | Angiotensin II Receptor, Type 1 |
| APOB | Apolipoprotein B |
| APOE | Apolipoprotein E |
| ANRIL | Antisense Non-coding RNA in the INK4 Locus |
| BMI | Body Mass Index |
| BNF | British National Formulary |
| BRHS | British Regional Heart Study |
| BRIGHT | British Genetics Of Hypertension |
| BWHHS | British Women’s Heart and Health Study |
| CAD | Coronary Artery Disease |
| CARDIoGRAM | Coronary Artery Disease Genome-Wide Replication And Meta-Analysis |
| CD-CV | Common Disease – Common Variant |
| CHD | Coronary Heart Disease |
| ChIP-Seq | Chromatin Immuno-Precipitation With Sequencing |
| CI | Confidence Intervals |
| CIMT | Carotid Intima-Media Thickness |
| CNV | Copy Number Variation |
| CRP | C-Reactive Protein |
| CVD | Cardiovascular Disease |
| CYP11B2 | Cytochrome P450, Family 11, Subfamily B, Polypeptide 2 |
| DNase-Seq | DNase I Hypersensitive Sites Sequencing |
| ECG | Electrocardiogram |
| ECG-LV mass | Electrocardiographically-Derived Left Ventricular Mass |
| Echo-LV mass | Echocardiographically-Derived Left Ventricular Mass |
| ENCODE | Encyclopaedia Of DNA Elements |
| FAIRE-Seq | Formaldehyde-Assisted Isolation Of Regulatory Elements With Sequencing |
| FH | Familial Hypercholesterolemia |
| GLGC | Global Lipids Genetics Consortium |
| GRAPHIC | Genetic Regulation Of Arterial Pressure Of Humans In The Community |
| GWA | Genome-Wide Association |
| GWAS | Genome-Wide Association Studies |
| HCM | Hypertrophic Cardiomyopathy |
| HDL | High-Density Lipoprotein |

| | |
|---|---|
| HDL-C | High-Density Lipoprotein-Cholesterol |
| HWE | Hardy-Weinberg Equilibrium |
| IBD | Identity-By-Descent |
| IBS | Identity-By-State |
| I/D | Insertion/Deletion |
| IDL | Intermediate-Density Lipoprotein |
| IMPROVE | IMT Progression As Predictors Of Vascular Events In A High Risk European Population |
| KCNA5 | Potassium Voltage-Gated Channel, Shaker-Related Subfamily, Member 5 |
| LD | Linkage Disequilibrium |
| LDL | Low-Density Lipoprotein |
| LDL-C | Low-Density Lipoprotein-Cholesterol |
| LDLR | Low-Density Lipoprotein Receptor |
| LPL | Lipoprotein Lipase |
| LV | Left Ventricular |
| LVH | Left Ventricular Hypertrophy |
| MAF | Minor Allele Frequency |
| MDS | Multi-Dimensional Scaling |
| MI | Myocardial Infarction |
| MR | Mendelian Randomisation |
| MRI | Magnetic Resonance Imaging |
| MYOZ2 | Myozenin 2 |
| NHGRI | National Human Genome Research Institute |
| NHS | National Health Service |
| NICE | National Institute Of Health And Clinical Excellence |
| NIH | National Institute Of Health |
| OLS | Ordinary Least Squares |
| OMIM | Online Mendelian Inheritance In Man |
| PCSK9 | Proprotein Convertase Subtilisin/Kexin Type 9 |
| PREVEND | Prevention Of Renal And Vascular End Stage Disease |
| QC | Quality Control |
| QQ | Quantile-Quantile |
| SCD | Sudden Cardiac Death |
| SCN5A | Sodium Channel, Voltage-Gated, Type V, Alpha Subunit |
| SD | Standard Deviation |
| SLOCB1 | Solute Carrier Organic Anion Transport Family Member 1b1 |
| SNP | Single Nucleotide Polymorphism |
| TF | Transcription Factor |
| TTN | Titin |
| UTR | Untranslated Region |
| VIF | Variance Inflation Factor |

| | |
|---|---|
| VLDL | Very Low-Density Lipoprotein |
| WHII | Whitehall II |
| WHO | World Health Organisation |

# 1 Introduction

## 1.1 Background

Cardiovascular disease (CVD) is the main cause of death in the UK. The World Health Organisation (WHO) has predicted that 23 million people per year will die from CVD globally by 2030 (World Health Organisation 2012). CVD encompasses a wide range of Mendelian and complex disorders, including diseases of the vasculature, diseases of the myocardium, diseases of the heart's electrical circuit, and congenital heart disease (Table 1.1)

**Table 1.1 Cardiovascular disorders.**

| CVD Group | Description |
|---|---|
| Coronary heart disease (CHD) or coronary artery disease (CAD) | Disease of the blood vessels supplying the heart muscle; mainly caused by a build-up of fatty deposits on the inner walls of the blood vessels that reduce or prevent blood flow to the heart, which can result in angina (chest pain) or myocardial infarction (heart attack). |
| Stroke or cerebrovascular disease | Disease of the blood vessels supplying the brain; the main causes of stroke include restriction of blood flow to the brain either due to fatty-acid build up or formation of blood clots due to bleeding from a blood vessel in the brain. |
| Peripheral arterial disease | Disease of blood vessels supplying the arms and legs. |
| Cardiomyopathy | Disease of the heart muscle where the muscle becomes enlarged, thick, or rigid, preventing efficient pumping of blood through the body. |
| Cardiac dysrhythmia | Disorders of the heart rhythm due to abnormal electrical activity in the heart. |
| Rheumatic heart disease | Damage to the heart muscle and heart valves from rheumatic fever, caused by streptococcal bacteria. |
| Congenital heart disease | Malformations of heart structure existing at birth. |
| Deep vein thrombosis and Pulmonary embolism | Blood clots in the leg veins which can dislodge and move to the heart and lungs. |

Modified from WHO Cardiovascular diseases Fact sheet N°317 (September 2012) (World Health Organisation 2012).

Family and twin studies of common CVDs provide evidence for considerable genetic contribution (Marenberg et al. 1994; Wienke et al. 2005). Hence, genetic research has long been pursued to enhance our understanding of the contributing molecular mechanisms, with the potential of identifying new therapeutic targets and improving current disease risk prediction models.

There are two main approaches to the discovery of genes for CVD and its risk factors in humans – linkage analysis and genetic association. For Mendelian disorders, where a simple pattern of inheritance suggests a single casual gene with large effect on phenotype (Kathiresan & Srivastava 2012), linkage analysis has largely been successful in identifying the causal mutation. For example, in 1989 Jarcho et al. localised the chromosomal position of a causal gene for familial hypertrophic cardiomyopathy (Jarcho et al. 1989), and a year later causal mutations in the beta cardiac myosin heavy chain were identified within this region (Geisterfer-Lowrance et al. 1990). On the other hand, complex disorders follow complex inheritance patterns that are suggestive of interaction between multiple loci and non-genetic factors. Early genetic research into common CVD using a candidate-gene approach revealed few replicated positive findings. For example, candidate-gene sequencing identified mutations in the *MEF2A* gene, a member of the myocyte enhancer family of transcription factors. This was the first locus to be implicated in the autosomal dominant form of CAD  (Wang et al. 2003), but the casual role of this gene was soon disputed due to lack of replication (Weng et al. 2005).

The sequencing of the human genome (Lander et al. 2001; Venter et al. 2001) and the subsequent development of affordable high-throughput genotyping technology have made it possible to identify associations of common genetic variants with disease events or with risk factors at the genome-wide level – genome-wide association studies (GWAS) (described later in section 1.3.3). As a result, the last five years have seen the identification of numerous replicated novel loci for some of the most important CVD traits, including CAD (McPherson et al. 2007; Samani et al. 2007; Schunkert et al. 2011; Davies et al. 2012) and myocardial infarction (MI) (Helgadottir et al. 2007; Kathiresan et al. 2009).

Efforts have also been focused on large-scale identification of genetic determinants of well-known modifiable CVD risk factors, including serum lipids (e.g. low- and high-density lipoprotein-cholesterol (LDL-C and HDL-C), and triglycerides) (Talmud et al. 2009; Teslovich et al. 2010; Asselbergs et al. 2012), body mass index (BMI) (Speliotes et al. 2010), electrocardiogram (ECG) measures (Chambers et al. 2010; Pfeufer et al. 2010; Sotoodehnia et al. 2010), left ventricular (LV) mass (Mayosi et al. 2008; Arnett et al. 2009), carotid-intima media thickness (CIMT) (Baldassarre et al. 2010; Bis et al. 2011) and blood pressure (Levy et al. 2009; Ganesh et al. 2013). Some of these risk factors are discussed in more detail in section 1.5.

The proportion of the total phenotypic variance explained by all genetic contributions is known as broad-sense heritability, while that attributed only to the additive genetic contribution is referred to as narrow-sense heritability. For all the mentioned CVD and risk factor traits, despite the discovery of numerous associated loci, their combined effects explain only a modest fraction of the estimated (narrow-sense) heritability of the trait (Manolio et al. 2009), indicating that there may be other genetic factors that are yet to be identified. This may explain why in most analyses the current known genetic risk factors have failed to provide substantial improvement over traditional non-genetic risk factors in disease risk prediction (Kathiresan et al. 2008; Talmud et al. 2010), leaving their utility for this purpose unclear. Another application of genetic variants has been in addressing the question of whether an epidemiologically observed relationship between risk factor and outcome is due to the causal effect of the former on the latter, termed Mendelian randomisation (MR) (Smith & Ebrahim 2003). Such studies have confirmed a causal relationship between LDL-C and coronary disease (Ference et al. 2012; Linsel-Nitschke et al. 2008).

In the rest of this chapter, I describe the biology and genetics behind CVD and some of the above mentioned risk factors, as well as introduce methodological concepts, both of which aim to provide the background to the work presented in this thesis.

## 1.2 Genetic Variation and Disease

Disease causing variants range from exceedingly rare mutations to very common genetic variations, and their effects from large to negligible (Marian & Belmont 2011). Single nucleotide polymorphisms (SNPs), which consist of differences between individuals at a single nucleotide position, are the most common type of variation, with approximately 15 million SNPs identified in European, African and Asian populations (The 1000 Genomes Project Consortium 2010). Other types of variation include deletions, duplications, and copy number variations (CNVs). The vast majority of known SNPs are thought to be neutral as they are located in apparently non-functional regions of the genome (Frazer et al. 2009). However, when they occur within coding or regulatory regions they may affect protein sequence or expression, potentially resulting in an altered phenotype.

Mendelian disorders are usually caused by mutations in a single gene that often have large, deleterious effects. Though there are exceptions (e.g. cystic fibrosis transmembrane conductance regulator (*CTFR*) mutations that cause cystic fibrosis (Mateu et al. 2002)), such mutations tend to have more recent origins and are rare due to negative selection. Though important for conferring disease risk in the individual carrying the mutation and in relatives, the impact of rare mutations at the population level is usually low. Complex diseases, on the other hand, are polygenic in nature with significant genetic and environmental contribution (Schork 1997). The underlying architecture of common complex diseases has been attributed to four possible models:

1. Common disease – common variant (CD-CV) model: This assumes an additive contribution of multiple common polymorphisms (defined as having a minor allele frequency (MAF) >5%) in multiple loci (Bodmer & Bonilla 2008). However, almost without exception, the combined effect of common genetic variants accounts for only a small proportion of the trait variance – referred to as the 'missing heritability' problem (Manolio et al. 2009). Despite this, the attributable risk of common alleles in a population may be considerable due to their frequency in the population (Marian & Belmont 2011).

2. Infinitesimal model: Every gene contributes to every trait, but with effect sizes that are so small that samples greater than the population size of the species would be needed to detect them (Gibson 2012). The GWAS on height in over >180,000 individuals, identified hundreds of variants in at least 180 loci that passed the pre-defined significance threshold, but these only explained 10% of the phenotypic variation in height (Lango et al. 2010). Rather than selecting variants passing a pre-defined association significance threshold, a recently developed method considering all SNPs simultaneously, estimated that 45% of the phenotypic variance in height was explained by all SNPs on the genotyping platform used in the height GWAS (Yang et al. 2010). This suggests the idea that heritability is not so much missing as it is hidden below the stringent significance thresholds used in GWAS (Gibson 2010).

3. Rare allele model: This model is the opposite of the CD-CV model, assuming many rare alleles with large effects are the main genetic contributors. (Bodmer & Bonilla 2008).

4. Broad sense heritability model: Where a combination of genotypic, environmental and epigenetic interactions contribute to the phenotype (Gibson 2012).

Currently there is insufficient empirical evidence to support any single model, and no single technological platform to assess all models simultaneously. However, it is reasonable to assume that allelic architecture (number, type, effect size and frequency of genetic variants) may differ across traits (Manolio et al. 2009), and that all models may contribute, at varying levels, to different diseases or traits. For example, GWAS analysis identified common variants in seven loci associated with triglyceride levels, while re-sequencing of selected genes identified an excess of rare, non-synonymous variants across four genes when comparing individuals in extremes of the plasma triglyceride distribution (Johansen et al. 2010). In addition, the large overlap between genes identified by GWAS and those identified earlier through Mendelian families also provides support for  genetic variants across the spectrum of allele frequencies contributing to complex common diseases (Kathiresan & Srivastava 2012).

## 1.3 Investigating the Genetic Basis of Disease

### 1.3.1 Linkage Disequilibrium

The International HapMap project (HapMap Consortium 2003) was launched in 2002 with the aim of providing a public resource to aid medical genetic research. The goal was to characterise common genetic variants (MAF >5%) in humans and determine their frequency and correlation within different ethnic populations. Two hundred and seventy individuals with African, European and Asian ancestry were sequenced, characterising over 3 million SNPs. The co-inheritance of SNP alleles on a chromosome leads to associations between these alleles in the population (known as linkage disequilibrium (LD)). Recombination is more likely to occur as the distance between two SNPs increases, and the correlation between two SNPs is therefore likely to decline with physical distance on the chromosome.

The presence of LD between SNPs in a region means that genotyping only a few, carefully chosen ‘tag’ SNPs in the region will provide enough information to predict much of the information about the remainder of the common SNPs in that region (HapMap Consortium 2003). The genomic distance at which LD decays determines how many genetic markers are needed to tag a haplotype (Visscher et al. 2012). On the basis of empirical studies, it has been estimated that genotyping around 500,000 common SNPs, combined with the knowledge of LD structure, is sufficient to allow the vast majority of common variants to be tested for association with phenotypes in non-African populations (The International HapMap Consortium 2005). This serves the basis for commercial genome-wide association SNP arrays, which due to the limitation of the physical size of the array, use a tag SNP approach to capture a large proportion of the variation in the genome using a substantially smaller number of SNPs on the array (de Bakker et al. 2005).

With recent advances in DNA sequencing technologies (next-generation sequencing), the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010) was launched in 2008. The aim was to discover and provide accurate haplotype information for most genetic variants that have frequencies of at least 1% in about 2500 individuals from 27 different populations with ancestry from Europe, East Asia, South Asia, West Africa and the Americas (The 1000 Genomes Project Consortium 2010). The data released thus far provides a much deeper, more uniform picture of human genetic variation than was previously available through HapMap (The 1000 Genomes Project Consortium 2010). The development of computational and statistical methods make it possible to use the haplotypic information from HapMap and 1000 Genomes to computationally impute the genotypes in samples for millions of additional variants beyond those that are genotyped using commercial SNP arrays, with the added benefit of no additional genotyping cost. In all approaches to gene mapping, the underlying assumption is that a disease-predisposing allele will pass from generation to generation together with other variants in high LD (Balding 2006).

### 1.3.2 Linkage Analysis

Linkage analysis tests whether a genetic marker tagging a region along the genome is co-transmitted with disease more often than expected by chance within large pedigrees, followed by fine-mapping of these regions to identify the casual gene and variant. The success of linkage studies depends on the availability of phenotypically well-characterised families that include a sufficiently large number of informative affected individuals (Cambien & Tiret 2007) and where the causal variants have strong heritability. An important advantage of linkage methods is that the combined information from several affected families has the potential to identify a causal locus even when different rare variants within the same locus are responsible for the disease in different families, making it appropriate when many rare variants at a locus each contribute to disease risk (Balding 2006). However, even in monogenic disorders the relationship between genotype and phenotype can be complex due to three genetic phenomena:

1. Penetrance – not all individuals with a given genotype will exhibit the phenotype associated with the genotype, referred to as incomplete penetrance. This may be due to modifier gene or environmental interactions (Kathiresan & Srivastava 2012).

2. Expressitivity – individuals with the same genotype can show varying degrees of the same phenotype.

3. Pleiotropy – mutations in a single gene can influence multiple phenotypic traits.

Some non-hereditary phenotypes, induced by certain environmental conditions, can also resemble a phenotype with a known genetic cause (phenocopy). Together these make gene discovery more difficult, since genotype may not segregate perfectly with phenotype (Kathiresan & Srivastava 2012). Linkage analysis generally lacks statistical power for identifying causal variants with low penetrance and small effect, which are more characteristic of complex disease traits (Balding 2006).

### 1.3.3 Genetic Association Analysis

Genetic association studies compare the frequency of genetic variants in control and disease groups of unrelated individuals, or test for association with a continuous trait. If the proportion of cases to controls, or the mean of the continuous trait significantly differs between the genotype groups for a particular genetic variant, then this provides evidence for association (Balding 2006). Association studies in unrelated individuals using genome-wide SNP arrays have been widely used to identify genetic variants associated with complex diseases without *a priori* knowledge of candidate genes and pathways, with the assumption that the associated variant is either casual or is tagging the causal variant. The power to detect associations using genome-wide platforms depends directly on sample size, MAF, strength of LD between the genotyped variant and causal variant, and the effect size (Spencer et al. 2009). Since 2005, the number of published GWAS has dramatically increased. A catalogue of published GWAS (Hindorff et al. 2009) is maintained by the National Human Genome Research Institute (NHGRI) at the National Institutes of Health (NIH). By mid-2012 it contained results from over 1300 published GWAS on over 200 traits.

Despite the value of the genome-wide approach, the technology still has considerable cost and relatively low power to detect subtle, but potentially important effects, in studies of typical sample sizes (a few thousand individuals) (Keating et al. 2008). In addition, given the large number of statistical tests performed when using such platforms, very stringent statistical significance thresholds have been adopted to reduce false discoveries, at the cost of discarding true associations. This is evident in the fact that significant associations reported in publications explain a relatively small proportion of the total phenotypic variation, despite studies showing a much higher heritability estimate when all SNPs on the genotyping platform are considered (Yang et al. 2010). Due to the tag-SNP approach in the design of genome-wide arrays, and depending on the size of the LD block tagged by the genotyped SNPs, fine-mapping of the identified regions is often needed in order to identify the causal variant. In addition, rare genetic variants are difficult to tag with common markers and are therefore under-represented on genome-wide platforms.

Array-based genotyping technologies that have enabled GWAS also permit flexibility in choosing the scope and density of SNPs for candidate gene studies. This has led to the development of several custom gene-centric chips such as the Cardiochip (Keating et al. 2008) designed for studying cardiovascular disease, the Metabochip for cardiovascular and metabolic traits (Voight et al. 2012), and the Immunochip (Cortes & Brown 2011) for auto-immune and inflammatory traits. Compared to genome-wide arrays, though these have fewer SNPs, they provide much denser coverage of candidate genetic loci enabling cost-effective fine-mapping of loci for both rare and common variants, and reduce the multiple-testing problem that is a major issue in genome-wide studies.

### 1.3.4 Association with Disease Outcome versus Continuous Risk Factors

Association studies with disease outcomes have the potential to uncover novel disease pathways. However, there are two major limitations of such an approach to genetic discovery. Firstly, CVD encompasses a range of conditions, and for the purpose of GWAS, phenotypes are often classified on the basis of observed disease outcomes such as CAD, MI or stroke. However, even within these broad categories individuals may have different disease aetiologies as well as different genetic and biological risk factors (Arking & Chakravarti 2009). Secondly, the success of such studies for identifying common variants with small effect relies heavily on power and sample size. This realisation has led to the formation of large, collaborative consortia which enable combining of results from multiple independent studies (meta-analysis) in order to obtain sufficient numbers of case samples. One such effort is the Coronary Artery Disease Genome-wide Replication And Meta-Analysis (CARDIoGRAM) consortium which combines 14 CAD GWAS, accruing data for 22,233 CAD cases and 64,742 controls (Schunkert et al. 2011).

Given these limitations, many groups have diverted efforts to studying intermediate CVD risk factors as they tend to be more homogenous and easily obtainable in existing population cohorts. Genetic determinants of intermediate phenotypes could potentially be linked to disease risk using a step-wise process of association in circumstances where the overall association between genotype and disease would be too small for direct detection in a case-control GWAS design (Carvajal-Carmona 2010). More recently, interest in the discovery of genetic determinants of risk factors has also been growing given their potential application in MR analysis to determine the causal relationship between risk factor and disease (section 1.6.2).

## 1.4 Genetics of Cardiovascular Disease

### 1.4.1 Mendelian Disorders

#### 1.4.1.1 Cardiomyopathy

Cardiomyopathy is the functional deterioration of the cardiac muscle. There are several types of cardiomyopathies, including hypertrophic, dilated and restrictive. The most common type is hypertrophic cardiomyopathy (HCM) which affects up to 1 in 500 individuals (Ramaraj 2008), and is the most common cause of sudden cardiac death (SCD) in young people. HCM is characterised by the presence of LV hypertrophy (increase in cell size), disorganised cardiac myocyte (muscle cell) architecture and widespread myocardial fibrosis (formation of excess fibrous tissue) (Maron et al. 1995). Several hundred mutations in over 20 different HCM susceptibility genes have been identified, most commonly in sarcomeric protein-coding genes (Fokstuen et al. 2011). Other HCM genes, such as Titin (*TTN*) (Satoh et al. 1999) and Myozenin 2 (*MYOZ2*) (Osio et al. 2007), code for sarcomere-interacting Z-disc proteins, which provide mechanic stability and act as nodal points for signalling (Knöll et al. 2011). Though there are several known genetic causes for cardiomyopathy, in around 25-35% of HCM patients the mutation remains unknown (Seidman et al. 2011). There is also marked variation in expressitivity and penetrance, with some patients remaining asymptomatic throughout their lifetime (Charitakis & Basson 2010), making genotype-phenotype correlation complex.

#### 1.4.1.2 Arrhythmogenic Disease

Cardiac arrhythmia is characterised by an abnormal heart rhythm, whereby the heart beats too fast (tachycardia), too slow (bradycardia) or irregularly (fibrillation). Arrhythmogenic diseases can be caused by mutations in ion channels and ion channel-controlling genes (e.g. *SCN5A*, sodium channel, voltage-gated, type V, alpha subunit; *KCNA5*, potassium voltage-gated channel, shaker-related subfamily, member 5), as well as calcium regulatory proteins, all of which play an important role in the propagation of the electrical signal in the heart, resulting in the synchronised contraction and relaxation of the atrial and ventricular chambers. Arrhythmogenic diseases include long-QT syndrome, short-QT syndrome, Brugada syndrome and catecholaminergic polymorphic ventricular tachycardia. The vast proportion of SCD cases are due to cardiomyopathies and arrhythmogenic disease (Pazoki et al. 2010).

#### 1.4.1.3 Lipid Disorders

Familial hypercholesterolaemia (FH) is an autosomal dominant disorder with an estimated prevalence of 1 in 500 in the UK (Marks et al. 2003). It is characterised by exceptionally high total cholesterol and LDL-C levels and is associated with a greatly elevated risk of CHD and death (Simon Broome Register Group 1991). Mutations in three genes are known to cause FH: LDL receptor (*LDLR*), apolipoprotein B (*APOB*), and proprotein convertase subtilisin/kexin type 9 (*PCSK9*) (the functions of the encoded proteins are discussed in section 1.5.2.1). However, in about 60% of clinically diagnosed FH patients no mutations are detected in these three genes (Taylor et al. 2010; Talmud et al. 2013).

### 1.4.2 Complex Disorders – Atherosclerosis-Related Disorders

The underlying disease process in coronary heart and cerebrovascular disease is atherosclerosis. Atherosclerosis is a multi-factorial process resulting in the thickening and hardening of arteries due to the build up of plaque (a combination of white blood cells, fatty acids, fibrous tissue, cholesterol and calcium deposits) within the artery walls. This results in the narrowing of the artery and restricted blood flow. The plaque can rupture leading to the formation of a blood clot which may completely block the artery. If this process occurs within the coronary artery, which supplies the heart, it results in MI, while complete blockage of the carotid artery supplying the brain results in stroke.

As mentioned previously, early candidate gene studies failed to identify replicated associations with CAD. However, in the last 5 years the widespread use of genome-wide association analysis has led to a huge increase in the discovery of replicated genetic variants associated with cardiovascular phenotypes including CAD, MI, heart failure, and stroke. In 2007, three studies (Helgadottir et al. 2007; McPherson et al. 2007; Samani et al. 2007) simultaneously reported genetic variants within a region on chromosome 9p21.3 associated with CAD and MI, and since then many other studies have confirmed this association. The region does not contain any annotated genes and research into the functional relevance of this region is ongoing. At the end of 2011, 26 CAD-risk loci had been identified through large GWASs (Zeller et al. 2011), with all loci exhibiting small to modest effect sizes. Several loci included well-known lipid genes, such as *LDLR* and *PCSK9*, supporting the importance of LDL-C in disease development (Lusis 2012). A few others showed evidence of association with hypertension, however, for most there was no established functional link to CVD pathways.

Despite the identification of multiple loci associated with disease, their combined effect explains a relatively small proportion of the total phenotypic variance. Heritability of CAD has been estimated at around 40% (Marenberg et al. 1994), and the SNPs identified by the CARDIoGRAM study together with previously known loci explain approximately 10% of the additive genetic variance of CAD (Schunkert et al. 2011).

## 1.5 Cardiovascular Risk Factors

### 1.5.1 Left Ventricular Mass

#### 1.5.1.1 Background

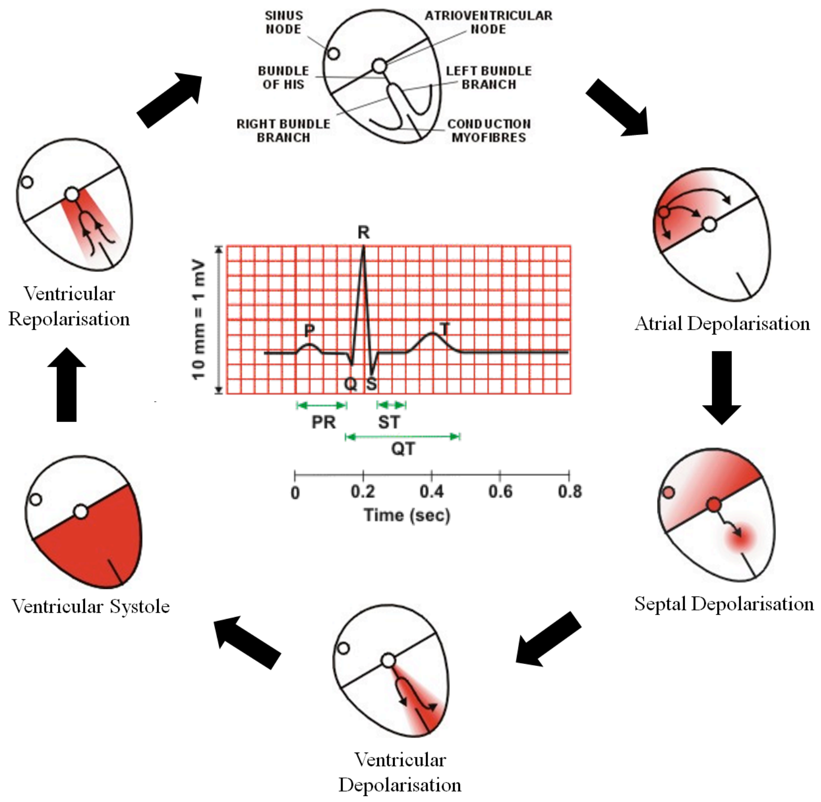
An increase in LV mass is associated with a higher incidence of cardiovascular events (Levy et al. 1990). Gender (independent of body size), age, blood pressure, ethnicity and BMI are all important determinants of this trait. LV mass measures are used in the diagnosis of left ventricular hypertrophy (LVH), the abnormal enlargement of the LV muscle tissue, which is a major risk factor for CVD (Kannel et al. 1987). LVH is a major cause of morbidity and mortality in hypertensive individuals, and historically it was considered to be an adaptive response to increased dynamic load caused by high blood pressure. However, the presence and magnitude of LVH varies substantially among individuals with similar blood pressure, and the relation between LV mass and cardiovascular risk has shown to be continuous in hypertensive individuals (Schillaci et al. 2000). Furthermore, studies on pathways resulting in LVH have shown that LVH may also occur in the absence of clear-cut recognisable changes in cardiac loading conditions (De Simone et al. 2001). Therefore, the discovery of genetic factors that contribute to even small increases in LV mass will likely have clinical importance.

#### 1.5.1.2 Measurement of LV Mass

LV mass can be measured by echocardiography, cardiac magnetic resonance imaging (MRI) or ECG. Echocardiography uses ultrasound techniques to obtain ventricular dimensions. Mathematical formulas are then used to estimate cardiac volume by fitting ventricular shape to geometric figures such as ellipse, cylinder, cone, and truncated polyhedrons (Foppa et al. 2005). MRI calculates volume from a three-dimensional set of images, without requirement for any assumptions about ventricular geometry. ECG is a graphical representation of the electrical activity of the heart over time (Figure 1.1) as measured by electrodes placed on the surface of the skin (Figure 1.2). The different ECG components are used to calculate ECG indices of LV mass (Figure 1.2). Increased LV mass is known to increase the height and depth of the QRS complex (Figure 1.1) and the length of the QRS duration (Figure 1.1).

Many different criteria for electrocardiographic LV mass (ECG-LV mass) measures have been proposed over the years, the most commonly used including Cornell Product, Sokolow-Lyon Index, QRS Voltage Product and QRS Voltage Sum (Figure 1.2). These four indices incorporate different components of the ECG, show differential pairwise correlation (shown in Section 2.3.2) and have a range of reported heritability estimates (mentioned below). Each measure, therefore, may provide independent information. Though cardiac MRI is currently considered the

**Figure 1.1 Description of the main ECG components.**



**P wave:** Represents atrial depolarisation that results in the contraction of the atria and the expulsion of blood into the ventricles.

**QRS complex:** Corresponds to ventricular depolarisation. The first downward deflection in the QRS complex is the **Q wave**, which represents septal depolarisation. The first upward deflection of the QRS is called the **R wave**. Most of the ventricle is activated during the R wave. The R wave may be unusually tall in the presence of left ventricular hypertrophy. The rim of the ventricular muscle is the last to contract and this late depolarisation is represented by the **S wave**, shown as the downward deflection following the R wave. An abnormally large S wave may also indicate hypertrophy. If a second upward deflection is recorded, this is indicative of a problem in the ventricular conduction system, including conduction blocks in the branches of the bundle of His. These are referred to as the R-prime and S-prime waves (R’ and S’).

**QRS Duration:** Duration of the QRS complex.

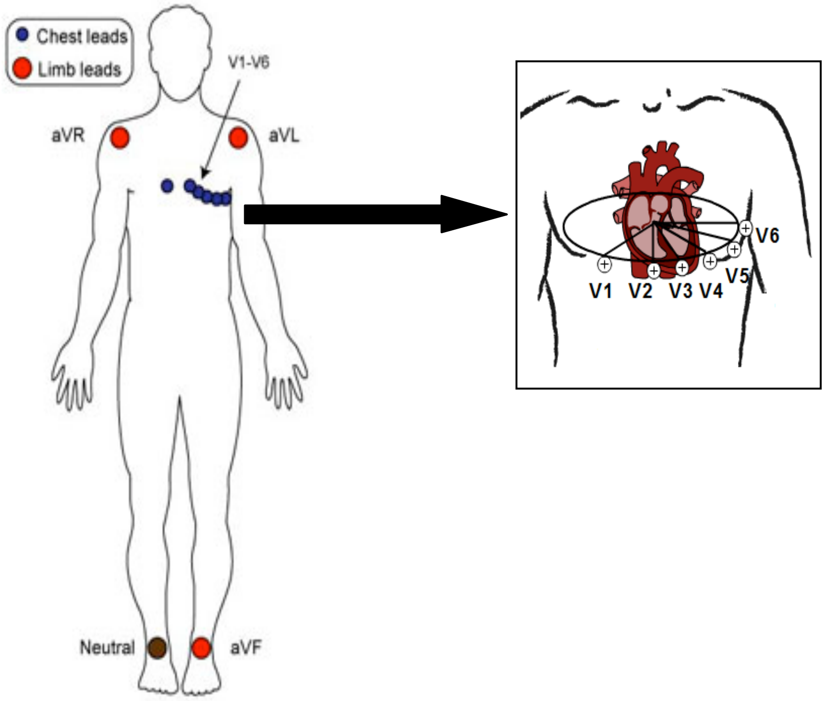
**ST segment:** The isoelectric period when the entire ventricle is depolarized and roughly corresponds to the plateau phase of the ventricular action potential.

**T wave:** Represents repolarisation and relaxation of the ventricles.

**QT interval:** Measured from the beginning of the QRS complex to the end of the T wave, the QT interval represents the time for both ventricular depolarisation and repolarisation to occur. It is dependent on the heart rate (the faster the heart rate, the shorter the QT interval) and is therefore usually reported after correcting for heart rate.

(Images modified from www.cvphysiology.com and www.cardionetics.com)

**Figure 1.2 Left ventricular mass indices from 12-lead ECG.**



The figure shows the placement of 10 electrodes (6 chest and 4 limb) for the measurement of 12-lead ECG. ECG-LV mass indices are calculated as follows:

- **Sokolow-Lyon Index** (µV) = SV1 + max (RV5, RV6) i.e. (Sokolow & Lyon 1949)
- **Cornell Product** (µV.s) = Cornell voltage x QRS Duration (where Cornell voltage = RaVL + SV3 (600µV added for females) (Casale et al. 1987)
- **QRS Voltage Sum** (µV) = the sum of |Q| + R + |S| + R' + |S'| amplitudes in all 12 leads (Molloy et al. 1992; Okin et al. 1995)
- **QRS Voltage Product** (µV.s) = QRS Voltage Sum x QRS duration (Molloy et al. 1992; Okin et al. 1995)

where,
SV1 and SV3 – amplitude of the S wave as measured by lead V1 and lead V3
RV5 and RV6 – amplitude of the R wave as measured by leads V5 and V6
RaVL – amplitude of the R wave as measured by the aVL lead
|Q| and |S| – absolute amplitude of Q wave and S wave
|R'| and |S'| – absolute amplitude of R' and S' wave
R – amplitude of the R wave

gold standard for estimating LV mass, its utility is restricted due to high cost and limited availability. There is evidence that echocardiographically-derived LV mass (echo-LV mass) has greater sensitivity than ECG-LV mass for LVH diagnosis. However, a study in 475 elderly men comparing echo-LV mass with ECG-LV mass (based on Cornell Product) for the diagnosis of LVH concluded that the two predict mortality independently of each other and other CVD risk factors, suggesting that they capture somewhat different information on cardiac status (Sundström et al. 2001). The cost and operational considerations tend to limit the use of echocardiography in large-scale population studies and clinical trials. The ECG, on the other hand, is inexpensive and data easily obtainable or already available for participants of existing epidemiological studies.

### 1.5.1.3 Genetics of Left Ventricular Mass and Hypertrophy

Echo-LV mass has significant heritability, with reported estimates between 24-50% (Swan et al. 2003; Post & Levy 1994). ECG-LV mass indices have also been shown to have significant heritability (Sokolow-Lyon Index ~40%, Cornell Product ~23%) (Mayosi 2002). Until recently, discovery of genes associated with LVH in humans has mostly been restricted to severe familial forms of hypertrophy, such as HCM, with causal mutations being identified in several sarcomeric protein-coding genes. However, it is unclear whether genetic variation in these genes influences less severe forms of LVH. Genes involved in haemodynamic load, calcium homeostasis and cell growth have also been suggested to play a role in LVH development (Arnett et al. 2004). For example, polymorphisms in the angiotensin converting enzyme (ACE), which plays an important role in arterial vasoconstriction, have been associated with LVH in some studies (Gharavi et al. 1996; Perticone et al. 1997), but have failed to replicate in others (Kauma et al. 1998; Gomez-Angelats et al. 2000).

A previous genome-wide linkage analysis of ECG- and echo-LV mass in hypertensive families found suggestive evidence for loci on chromosomes 10q23.1 for Sokolow-Lyon Index, on 17p13.3 for Cornell Product and on 5p14.1 for echo-LVH (Mayosi et al. 2008). The identified regions, however, were large and spanned several genes and a causal mutation has yet to be identified in these regions. However, a recent study did show that one of the SNPs in the 17p13.3 region lies within the 3'UTR of a gene with unknown function, *TLCD2*, which is also immediately downstream of a microRNA (*miR-22*) (Harper et al. 2013). MicroRNAs are a group of small non-coding RNA molecules involved in posttranscriptional gene regulation, and there are now several studies supporting the role of *miR-22* as a pro-hypertrophic modulating miRNA (Jentzsch et al. 2012; Gurha et al. 2012; Xu et al. 2012). Therefore, *miR-22* and *TLCD2* may be strong candidates to account for this observation (Harper et al. 2013). The few GWAS studies that have been published have not been very successful in identifying many common variants. Two studies on echo-LV mass (including the largest GWAS to date for this trait, with discovery in 12,612 individuals and replication in 4,094 individuals) reported no definite associations with LV mass (Vasan et al. 2009; Arnett et al. 2011). Therefore, the pathophysiological mechanisms that underlie LVH remain incompletely characterised and genetic studies may help expose mechanisms not previously recognised to play a role in the development of LVH.

## 1.5.2 Lipids

### 1.5.2.1 Background

Lipids have been known CVD risk factors for over half a century. Cholesterol and triglycerides are two types of lipids that circulate within the blood. Triglyceride is the form in which dietary intake of fat or excess carbohydrate is stored. Cholesterol is an essential steroid found in the plasma membrane of all cells and is the precursor to all steroid hormones. Cholesterol is present in tissues and in plasma either as free cholesterol or as a storage form, combined with a long-chain fatty acid (cholesteryl ester) (Mayes & Botham 2012). The cholesterol in the body is synthesised and also provided by the average diet. When cholesterol levels in the blood are excessive, the liver secretes it into bile for excretion from the body. Lipids, being insoluble in aqueous solution, are transported in plasma by apolipoproteins, forming lipid-protein complexes known as lipoproteins (Feher & Richmond 2006). LDL (apolipoprotein B being the main protein component) and HDL (apolipoprotein A1 being the main protein component) are two classes of lipoproteins. Plasma lipoproteins are characterised by the proportion of protein in the lipoprotein complex, which determines their density. There are five major density classes: chylomicrons, very low-density lipoprotein (VLDL), intermediate-density lipoprotein (IDL), LDL and HDL (Feher & Richmond 2006). Plasma LDL is the vehicle of uptake of cholesterol and cholesteryl ester by many tissues while HDL transports excess cholesterol to the liver for elimination (Mayes & Botham 2012). There are three main pathways involved in the synthesis and transport of lipids within the body:

1. Exogenous (Dietary) Lipid Pathway: After digestion and absorption of dietary fat, triglycerides and cholesterol are packaged into chylomicrons in the intestine. Chylomicrons consist mainly of triglycerides. These are secreted into the lymphatic system and eventually join the blood circulation. When the chylomicrons reach muscle and adipose tissue, the triglycerides are hydrolysed by lipoprotein lipase (LPL) to release fatty acids that are taken up by the cells for energy or storage. The remaining components are known as chylomicron remnants. Cholesterol that is not used by the cells remains in the chylomicron remnants and these are eventually taken up by the liver (Figure 1.3). Apolipoprotein E (ApoE), the main protein component of chylomicron remnants, acts as a binding ligand for receptors located on the liver.

2. Endogenous Pathway: This involves the synthesis of cholesterol and triglycerides by the liver. These are transported in the blood stream to muscle and adipose tissue by VLDLs, where triglycerides are processed by LPL. Some of the VLDL remnant particles are removed from circulation by the liver via LDL receptors, while others are hydrolysed to form smaller, denser LDL particles, which are cholesterol-rich particles (Figure 1.3). Most of the LDL particles are also taken up by LDL receptors on hepatic cells, releasing free cholesterol which accumulates within the liver cells. The number of LDL receptors on the surface of the liver determines how quickly LDL particles are cleared from the bloodstream. When cells have abundant cholesterol, LDL receptor synthesis at the level of transcription, is blocked to prevent further cholesterol uptake. Conversely, more LDL receptors are made when the cell is deficient in cholesterol. Proprotein convertase subtilisin/kexin type 9 (encoded by the *PCSK9* gene) is believed to induce degradation of LDL receptors in the liver, resulting in reduced clearance of LDL particles from the blood.

3. Reverse Cholesterol Transport: This is the process by which excess cholesterol is removed from the tissues and returned to the liver by HDL, where it is metabolised to bile acids and salts that are eliminated from the body.

Excess circulating LDL-C can penetrate endothelial walls where it is oxidised by free radicals and becomes toxic to cells. The damage caused to the artery wall by oxidised LDL triggers an immune response, resulting in the recruitment of macrophages to the site of damage. The oxidised LDL molecules are taken up by macrophages which become engorged and form foam cells (cholesterol-loaded cells) (Libby et al. 2011). These foam cells may rupture, depositing a greater amount of oxidised cholesterol into the artery wall. This deposition of necrotic debris provokes further inflammation, continuing the cycle (Lusis 2012). The surrounding

**Figure 1.3 Synthesis of lipoprotein complexes in the small intestine, liver, and blood plasma and their delivery to peripheral tissues of the body.**

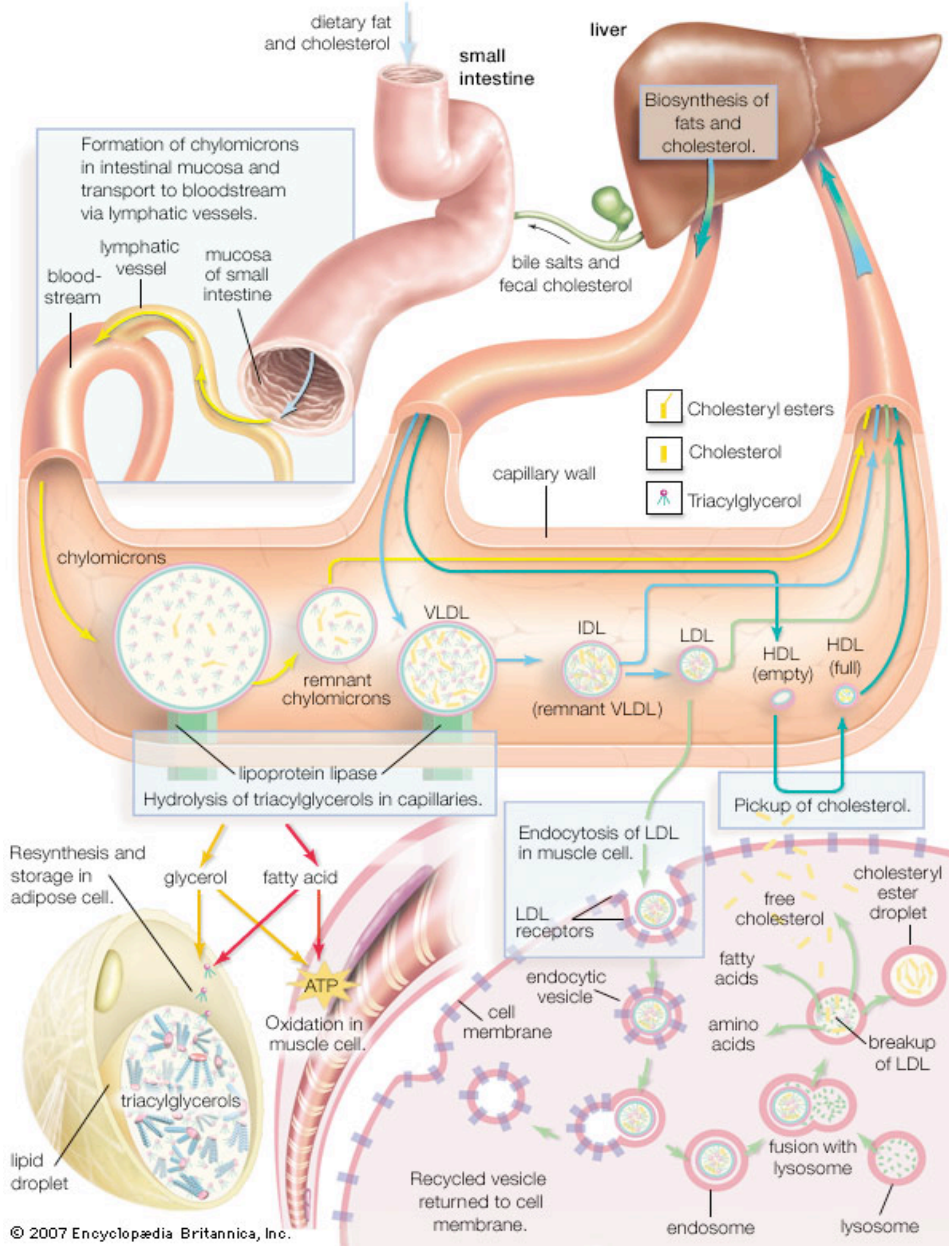

dietary fat and cholesterol
small intestine
liver
Biosynthesis of fats and cholesterol.
Formation of chylomicrons in intestinal mucosa and transport to bloodstream via lymphatic vessels.
lymphatic vessel
blood-stream
mucosa of small intestine
bile salts and fecal cholesterol
Cholesteryl esters
Cholesterol
Triacylglycerol
capillary wall
chylomicrons
remnant chylomicrons
VLDL
IDL
(remnant VLDL)
LDL
HDL (empty)
HDL (full)
lipoprotein lipase
Hydrolysis of triacylglycerols in capillaries.
Pickup of cholesterol.
Endocytosis of LDL in muscle cell.
LDL receptors
Resynthesis and storage in adipose cell.
glycerol
fatty acid
ATP
Oxidation in muscle cell.
triacylglycerols
lipid droplet
endocytic vesicle
cell membrane
free cholesterol
cholesteryl ester droplet
fatty acids
amino acids
breakup of LDL
fusion with lysosome
Recycled vesicle returned to cell membrane.
endosome
lysosome
© 2007 Encyclopædia Britannica, Inc.

Image taken from Encyclopaedia Britannica (http://www.britannica.com/bps/media-view/92255/0/0/0) muscle cells secrete a collagen-rich extracellular matrix to cover the lesion with a fibrous cap that separates the plaque from the blood (Figure 1.4). The formation of plaque causes the artery to narrow and restricts blood flow. Rupture of the plaque leads to the formation of a blood clot which may completely block the artery. High levels of circulating LDL-C are known to contribute to the development of atherosclerosis, while high levels of circulating HDL-C are thought to have a protective effect from atherosclerosis, as it removes excess cholesterol deposited in the blood vessels and transports it back to the liver.

**Figure 1.4 Formation of atherosclerotic plaque.**

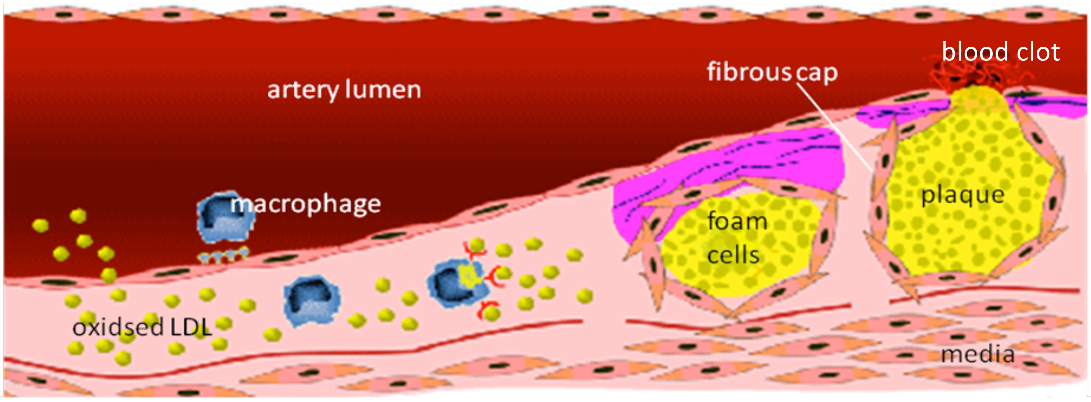


Image taken from http://users-phys.au.dk/jvn/Research-CABRA.htm

### 1.5.2.2 Measurement of Lipid Levels

Serum total cholesterol and triglycerides can be measured using a centrifugal analyser. HDL-C can be measured by first precipitating apoB-containing lipoproteins (non-HDL-C components) with dextran sulphate-magnesium chloride, followed by centrifugation. The HDL-C in the supernatant can then be measured by an enzymatic procedure. Direct measurement of LDL-C is time-consuming and requires expensive instrumentation that is not available in routine laboratories (Branchi et al. 1998). Therefore, LDL-C is commonly calculated from HDL-C and triglyceride measurements using an empirical equation - the Friedewald formula (Friedewald et al. 1972). However, the accuracy of LDL-C estimation can be affected by serum triglyceride levels and errors inherent to the methods used to obtain HDL-C and triglyceride measures.

### 1.5.2.3 Lipid Genetics

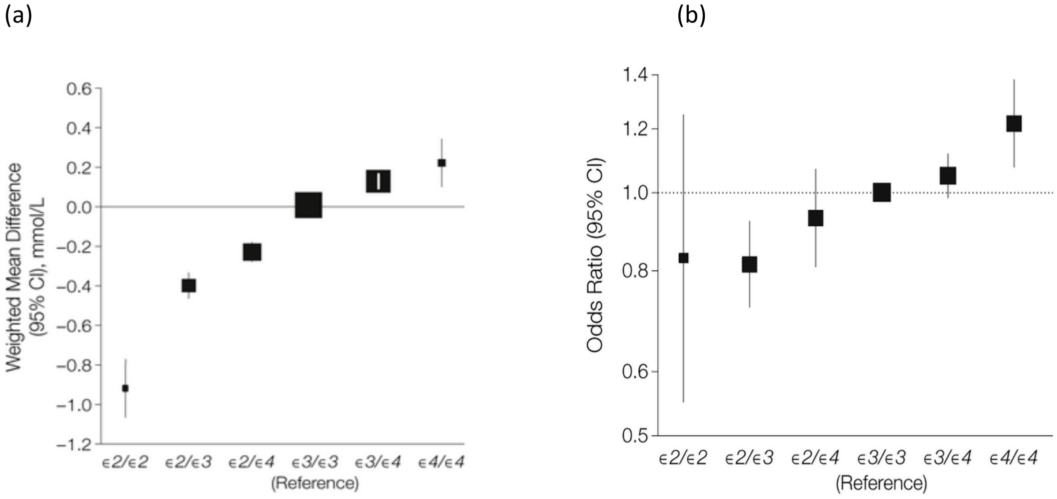
Plasma-lipid levels are highly heritable traits with estimates ranging from 40% – 70% (Krauss 2008; Weiss et al. 2006). A major contributor to LDL-C and total cholesterol levels is the *APOE* gene which codes for apolipoprotein E. SNPs at two sites (rs7412 and rs429358) result in three major protein isoforms with either a cysteine (cys) or arginine (arg) residue at positions 112 and 158: ApoE2 (cys112, cys158), ApoE3 (cys112, arg158; the most common form), and ApoE4 (arg112, arg158) (Utermann et al. 1980; Weisgraber et al. 1981) (Table 1.2)

**Table 1.2 Combinations of rs429358 and rs7412 responsible for the observed *APOE* alleles.**

| *APOE* alleles | rs429358 | rs7412 |
|---|---|---|
| ε2 | T | T |
| ε3 | T | C |
| ε4 | C | C |
| Not observed | C | T |

These differences in the amino acid composition affect its binding to hepatic lipoprotein receptors. The three common alleles: ε2, ε3 and ε4, result in six possible genotypes: ε2ε2, ε2ε3, ε2ε4, ε3ε3, ε3ε4 and ε4ε4. A large meta-analysis in 61,463 healthy participants showed an approximately linear relationship of *APOE* genotypes with LDL-C levels, and with coronary risk in 21,331 cases and 47,467 controls (Figure 1.5). However, the presence of the Apoe E2/E2 genotype is also known to cause hyperlipoproteinemia type III, characterised by the accumulation of remnant lipoproteins in the plasma and development of premature atherosclerosis, which complicates the relationship with CVD.

**Figure 1.5 Association of *APOE* genotypes with lipid levels and coronary risk.** (a) Differences in LDL-C levels and (b) odds ratios for coronary disease by *APOE* genotypes using individuals with the ε3ε3 genotype as reference group. Size of data markers is proportional to the inverse of the variance of the weighted mean difference or odds ratios (ε3ε3 is represented by a square with arbitrarily fixed size) and vertical lines represent 95% confidence intervals (CIs).



Images taken from Bennet et al (2007).

Several large-scale association analyses have identified large numbers of genetic variants associated with lipid traits (Talmud et al. 2009; Teslovich et al. 2010; Asselbergs et al. 2012), some of which have also been shown to be associated with CAD (Teslovich et al. 2010). These loci explain around 10-15% of the total phenotypic variance of the lipid traits (Teslovich et al. 2010). Many of these loci show strong overlap with those responsible for Mendelian disorders - of the 19 genes identified as monogenic causes of extremely low or high levels of LDL-C, HDL-C and triglycerides, common variants in or near 16 of these genes have been identified through GWAS (Kathiresan & Srivastava 2012) (Figure 1.6). Around one-third of the identified genes are known to play a role in lipid metabolism, including those that are already targets of lipid-modifying therapies. The rest offer novel insight into lipid biology, and for a few of these, mouse models have subsequently confirmed their role in lipoprotein regulation (Musunuru et al. 2010; Varbo et al. 2011). Therefore, despite each independent SNP explaining a very small proportion of the phenotypic variance, the biological and therapeutic value of the gene mapped by the variant may be very high.

**Figure 1.6 Lipid Genetic Loci.** Of 19 genes previously implicated in Mendelian lipid disorders, 16 lie within 100 kilobases of one of the lead SNPs mapped by GWAS, including nine that lie within 10 kilobases of the nearest lead SNP.

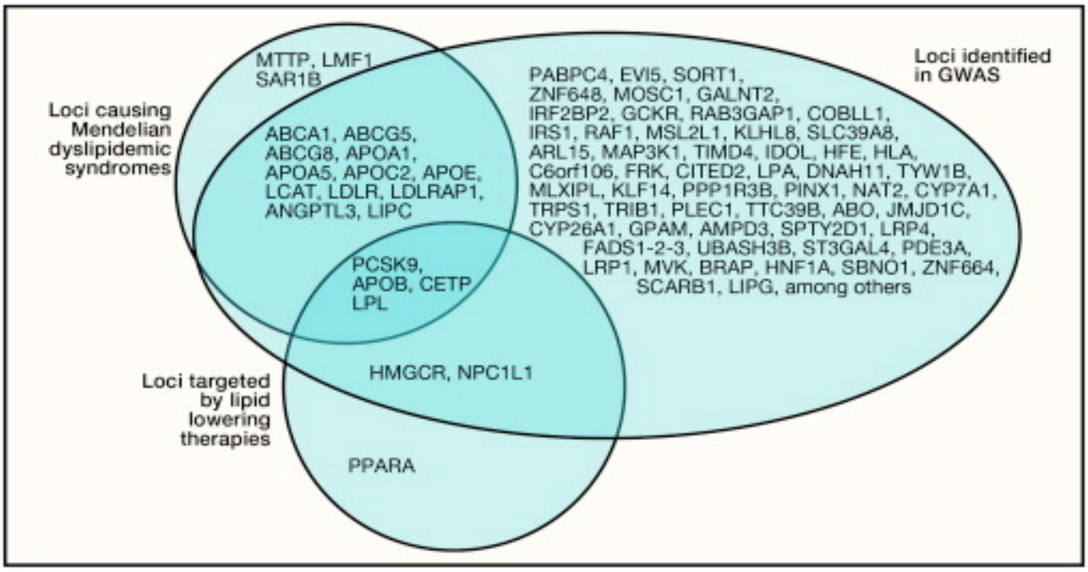

Loci causing Mendelian dyslipidemic syndromes
Loci identified in GWAS
Loci targeted by lipid lowering therapies
MTTP, LMF1 SAR1B
ABCA1, ABCG5, ABCG8, APOA1, APOA5, APOC2, APOE, LCAT, LDLR, LDLRAP1, ANGPTL3, LIPC
PCSK9, APOB, CETP LPL
PABPC4, EVI5, SORT1, ZNF648, MOSC1, GALNT2, IRF2BP2, GCKR, RAB3GAP1, COBLL1, IRS1, RAF1, MSL2L1, KLHL8, SLC39A8, ARL15, MAP3K1, TIMD4, IDOL, HFE, HLA, C6orf106, FRK, CITED2, LPA, DNAH11, TYW1B, MLXIPL, KLF14, PPP1R3B, PINX1, NAT2, CYP7A1, TRPS1, TRIB1, PLEC1, TTC39B, ABO, JMJD1C, CYP26A1, GPAM, AMPD3, SPTY2D1, LRP4, FADS1-2-3, UBASH3B, ST3GAL4, PDE3A, LRP1, MVK, BRAP, HNF1A, SBNO1, ZNF664, SCARB1, LIPG, among others
HMGCR, NPC1L1
PPARA

Image taken from Kathiresan & Srivastava (2012).

## 1.5.3 Carotid Intima-Media Thickness

### 1.5.3.1 Background

CIMT is a measure of the thickness of the two inner-most layers, the tunica intima and tunica media, of the carotid artery wall. The right and left common carotid arteries extend up the right and left side of the neck, respectively, to supply the head, neck and brain with oxygenated blood. The common carotid arteries branch at the bifurcation point into the internal (supplying the brain) and external (supplying the face and neck) carotid arteries (Figure 1.7).

**Figure 1.7 Carotid arteries.**

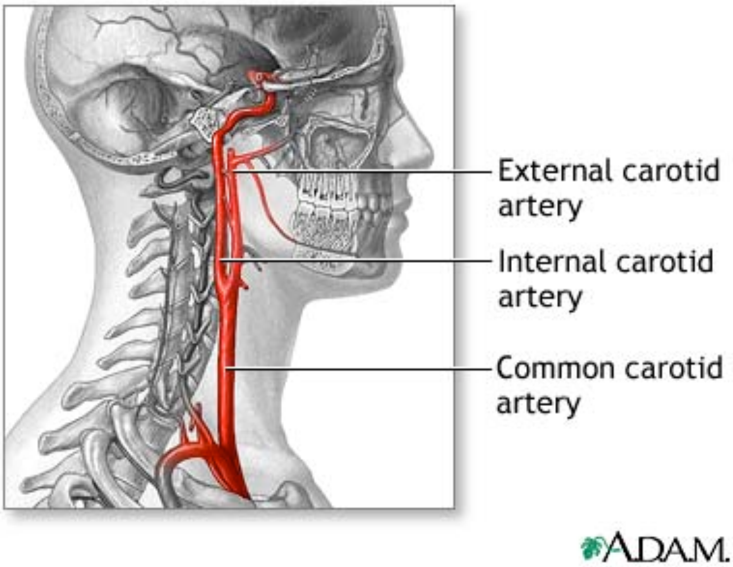


Image taken from http://www.umm.edu/graphics/images/en/13939.jpg

Increased CIMT is strongly associated with atherosclerosis and cardiovascular events (Bots et al. 1997; Chambless et al. 1997, 2000; O'Leary et al. 1999). As a result, it is commonly used as a surrogate endpoint for cardiovascular events in intervention trials. The use of a surrogate marker in clinical drug trials is much more cost-effective, as the sample size requirements are not as large, and the follow-up time not as long compared to trials with disease morbidity or mortality end-points. Trials using CIMT as a surrogate end-point determine drug efficacy by its ability to regress or slow progression of CIMT, with the assumption that this translates into a reduction in cardiovascular risk.

### 1.5.3.2 Measurement of CIMT

CIMT can be measured using non-invasive ultrasound imaging as shown in Figure 1.8**.** Measurements can be made from different combinations of segments (common carotid, carotid bifurcation and/or internal carotid artery), walls (far wall and/or near wall), and angles (single angle or a combination of angles) (Dogan et al. 2010), all of which are associated with differences in reproducibility, magnitude, and precision of CIMT measurement (Dogan et al. 2010). Lack of standardised protocol for CIMT measurement usually means that there is significant diversity in the protocols used to measure CIMT in different studies.

**Figure 1.8 Ultrasound scan of the carotid artery.** The arrows mark the inside border of the carotid tunica intima (innermost) layer, which is in direct contact with the blood, and the outside border of the tunica media, which consists of smooth muscle and elastic tissue.

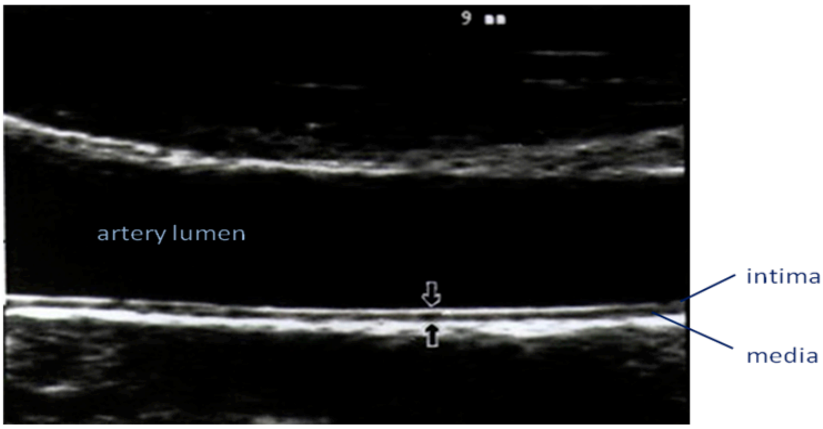


Image taken from http://londoncardiovascularclinic.co.uk

### 1.5.3.3 Genetics of CIMT

CIMT constitutes an attractive quantitative intermediate disease phenotype for the study of atherosclerosis-related CVD (Gertow et al. 2012). Although family studies have shown consistent evidence for moderate heritability for CIMT (Zhao et al. 2008; Sacco et al. 2009), candidate gene studies have not found consistent genetic associations with CIMT (Bis et al. 2011). Two large-scale association studies have reported significant, but different, associations with common CIMT. The first is a GWAS meta-analysis identifying three independent loci, including the *APOC1* gene (codes for apolipoprotein C1) on 19q13, a region that also includes *APOE*, *APOC2*, and *APOC4* genes (Bis et al. 2011), and the second a gene-centric analysis identifying the *BCAR1-CFDP1-TMEM170A* locus on chromosome 16, which was also shown to be associated with CAD (Gertow et al. 2012).

## 1.6 Applications of Genetic Variants

### 1.6.1 Disease Risk Prediction

In clinical practice, risk prediction algorithms have been used to identify individuals at high risk of developing CVD in the short term. These individuals could then be selected to receive therapeutic or lifestyle interventions that would reduce their risk and prevent or postpone the occurrence of disease. With the development of LDL-C-lowering drugs, as well as several randomised-controlled treatment trials confirming their casual role in the development of CHD (LaRosa et al. 2000; The Lipid Research Clinics Coronary Primary Prevention Trial 1984), a more targeted approach has been adopted world-wide in order to balance the inconvenience, risks and costs of intervention with the potential benefits of risk reduction (Dent 2009). Prescription of LDL-C-lowering statin drugs for primary prevention of CHD in the general population was initially informed by lipid level thresholds. However, cholesterol levels identify patients at risk of future coronary events only moderately well (Law & Wald 2002). Many individuals have a cholesterol concentration sufficient to raise the risk of coronary events, but the strength of the association of LDL-C with coronary events is only modest, with about a three-fold relative difference in the risk of coronary events among those at the extremes of the population LDL-C distribution (Di Angelantonio et al. 2009). There is a lot of supporting evidence that major CVD risk factors like blood pressure or blood lipid levels are individually poor predictors of a patient's CVD risk when compared with multifactor CVD risk prediction estimates (Jackson 2008). As a result, current UK, European, and Australasian guidelines recommend prescription of statins on the basis of absolute CVD risk rather than solely on LDL-C thresholds.

Guidelines from the National Institute for Health and Clinical Excellence (NICE) and other associations recommend commencing statin therapy in individuals estimated to have a 10 year absolute risk of CVD greater than 20% (NICE, 2010). The recommended methods for evaluating absolute CVD risk incorporate multiple, established risk factors. The most widely-used model is the Framingham 10 year CVD risk score (Anderson et al. 1991). This method uses a multivariable regression equation derived from a population sample of the Framingham Heart Study and the Framingham Offspring Study (Anderson et al. 1991). The Framingham Heart Study has been operational for more than 40 years and has identified a number of risk factors that have a cumulative impact on CVD (Anderson et al. 1991). The initial study in around 5,500 residents of Framingham, a town in Massachusetts, USA, included parents and offspring between the age of 30 and 74 who were initially free of cardiovascular disease and presented prediction equations for several CVD endpoints based on measurements of several known risk factors, including lipids, age, sex, blood pressure, smoking habit, and diabetes status (Anderson et al. 1991). Other risk scores used in Europe include QRISK (Hippisley-Cox et al. 2007), EuroSCORE (Nashef et al. 1999)and PROCAM (Assmann et al. 2002), with the latter incorporating information on family history, a surrogate for genetic effects.

Unfortunately, all the available risk models are far from perfect. This rests on the fact that our knowledge of the disease's aetiology is incomplete, both in terms of which risk factors are independently important and how they should each be weighted. Many of the important risk factors, such as blood pressure and serum cholesterol level, also show considerable intra-individual variation and cannot be measured with sufficient accuracy to support risk assessment with the required degree of certainty (Dent 2009). Given the number of genetic variants associated with CVD risk factors there is great interest in their utility for improving risk prediction, since genotype is fixed from conception and should therefore capture long-term differences in risk factor values without the biological variation that affects measurement.

## 1.6.2 Mendelian Randomisation Analysis

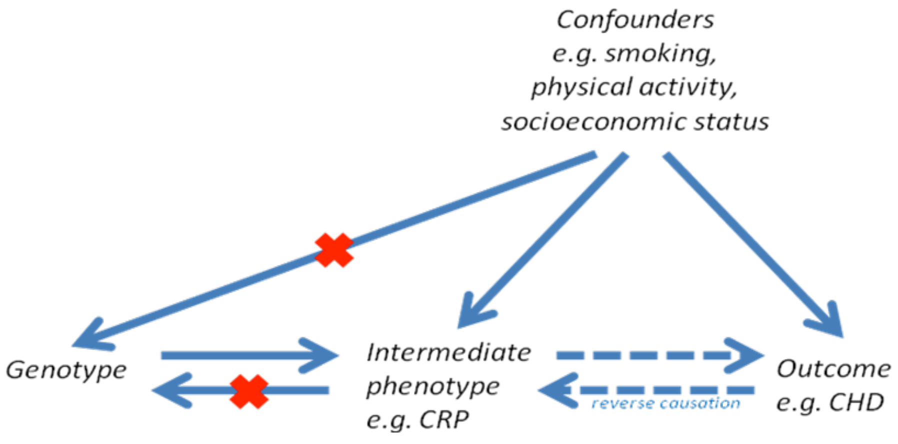
### 1.6.2.1 Background

Many known risk factors for CVD have been identified through epidemiological studies that have examined the direct association of the observed risk factor with disease outcome. Identification of modifiable risk factors can inform lifestyle or therapeutic interventions to reduce disease risk. However, inferring causality from observational data is problematic as it is not always clear which of the two associated variables is the cause and which the effect, or whether both are common effects of a third unobserved variable, or confounder (Sheehan et al. 2008). Confounding factors such as social, behavioural and environmental factors, are often more difficult to measure and control for, and it may not be possible to identify and account for all the relevant confounders.

Randomised controlled trials provide the most robust estimate of causal effect, however, they are not always feasible and have cost and ethical implications (Lawlor et al. 2004). One approach which circumvents the issues faced in observational studies, and that is increasingly being used to determine causal relationships, is MR analysis. MR is a relatively recent development in genetic epidemiology where genetic variants are used as a proxy for modifiable risk factors that are associated with disease (Thomas & Conti 2004). Since heritable units are randomly assigned at conception, genotypes should not be associated with confounding factors, such as smoking and socioeconomic circumstances, nor will the genotype be affected by disease processes that influence the intermediate risk factor (reverse-causation) (Smith & Ebrahim 2003).

Observational studies, for example, have shown C-reactive protein (CRP), a nonspecific marker of acute phase inflammatory response, to be a strong marker for CHD risk (Chambless et al. 1997). Inflammation plays a key role in the underlying disease process for the development of CHD, and CRP is currently the most widely used biomarker of inflammation (Ridker et al. 2000). However, it is unclear whether CRP is a causal factor. If a casual relationship is established, it would warrant the development of drugs specifically targeted at reducing CRP activity. However, if CRP levels are simply a marker of inflammation, targeting CRP is unlikely to be an effective means of reducing cardiovascular disease burden. Epidemiological studies may identify spurious associations due to confounding factors related to both exposure and disease outcome. In the case of CRP, higher levels are associated with smoking, which is also a risk factor for CHD, and this three-way relationship (Figure 1.9) might confound the purported causal link between CRP and CHD (Lawlor et al. 2008).

**Figure 1.9 The Mendelian randomisation paradigm using CRP and CHD as an example.**



A recent meta-analysis used MR analysis to look at the association of known CRP-associated genetic variants with CHD in 47 studies (total N = 194,418, with 46,557 CHD cases) (Wensley et al. 2011). Their findings indicate that CRP concentration itself is unlikely to be even a modest causal factor in CHD.

#### 1.6.2.2 Principles of Instrumental Variable Analysis

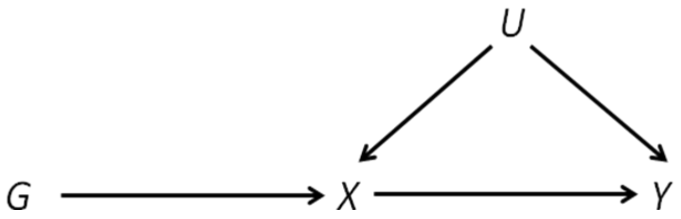
The method of instrumental variables (IVs) is an established method in econometrics used to estimate causal relations using observational data (Angrist et al. 1996). Standard regression estimates of the relation of interest may be biased because of the presence of unmeasured confounding factors, reverse causality, selection bias, or measurement error (Stock 2001). In such cases a third, `instrumental' variable can be used to extract variation in the variable of interest that is unrelated to these problems. This variation can then be used to estimate its causal effect on an outcome measure (Stock 2001). IV analysis where genotype is used as an instrument is known as MR. A valid instrument is defined as a variable that satisfies the following three assumptions:

1. The instrument (*G*), which in MR analysis is genotype, is strongly associated with the modifiable risk factor of interest (*X*) (Figure 1.10).

2. The instrument (*G*) is independent of any unmeasured confounding factors, (*U*) (Figure 1.10).

3. The instrument (*G*) is related to the outcome (*Y*) only via the risk factor of interest (*X*).

**Figure 1.10 The Mendelian randomisation model.** *G*, a genetic instrument with a specific effect on an intermediate phenotype, X; Y, an outcome; U, unobserved confounders of the suggested X-Y relationship.



### 1.6.2.3 Estimating Causal Effects using IV Methods

In the case of linear associations and a continuous outcome, the IV estimate of the causal effect of the exposure *X* on *Y* has been shown to be:

**Equation 1.1**

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{GY}}{\hat{\beta}_{GX}}$$

where $\hat{\beta}_{GY}$ is the coefficient for the regression of outcome *Y* on the genetic instrument *G*, and $\hat{\beta}_{GX}$ is the coefficient for the regression of the exposure *X* on the genetic instrument *G* (Thomas & Conti 2004; Lawlor et al. 2008). The above estimator $\hat{\beta}_{IV}$ only applies when there is a single IV (Lawlor et al. 2008) and this approach is known as the ratio of coefficients method. Where there is more than one IV, the simplest and most commonly used technique is the two-stage least squares (2SLS) method (Basmann 1957). The IV estimate is derived, as the name suggests, by two regression steps:

1. Performing a least-squares regression of the intermediate phenotype *X* on the instrumental variable *G*.

2. A second least-squares regression of the outcome *Y* on the predicted values of *X* obtained from the first regression. 2SLS assumes linear relationship between *G*, *X* and *Y*. The causal estimate is derived from this second regression.

Both 2SLS and the ratio method are applicable with a single instrument, in which case the causal estimates are identical, but 2SLS can also be used with multiple instruments.

#### 1.6.2.4 Instrument Strength and Weak Instrument Bias

The power to detect a casual relationship depends on the sample size and the strength of the instrument, which in MR studies depends on the proportion of variance explained by the known genetic factors ($R^2$) (Pierce et al. 2010). Although genetic variants are independent of confounders, confounders will not be perfectly balanced between genotypic sub-groups in finite samples (Burgess & Thompson 2011). If the instrument is weak and does not explain much of the intermediate phenotypic variance, the chance difference in confounders may explain more of the phenotypic difference between sub-groups than the instrument (Burgess & Thompson 2011). In finite samples, IV estimates are biased in the same direction as ordinary least squares (OLS) (a method for estimating the unknown parameters in a linear regression model) estimates between the observed intermediate phenotype and outcome. The magnitude of the bias of IV estimates approaches that of OLS estimates as $R^2$ approaches zero (Bound et al. 1995). However, it is important to note that the sample sizes typical of genetic studies nowadays are usually large enough to avoid weak instrument bias.

#### 1.6.2.5 Population Stratification

Since independent heritable units are randomly assigned from parent to offspring during gamete formation, they are not expected to be affected by any confounding factors other than ancestry (Pierce et al. 2010). Presence of population stratification may therefore violate the second assumption of IV analysis. However, restricting analysis within ethnically homogenous populations or incorporating population structure into the analysis should overcome this problem.

#### 1.6.2.6 Pleiotropy

SNPs identified from association studies may be within pleiotropic genes (genes that affect multiple phenotypic traits) that influence outcome directly or indirectly through risk factors other than the intermediate trait of interest (Palmer et al. 2011). Unless it is known that these other risk factors are in the same pathway downstream of the intermediate trait of interest, these may not be valid instruments. In the case of protein traits, such as CRP levels where a cis-acting genetic variant is used as an instrument, it is known that any association of this variant with other risk factors is via the downstream effect on CRP and assumptions for MR are not violated. However, in the case of non-protein traits, such as lipids, several SNPs in different genes have been reported to be associated with more than one of the three lipid fractions and also with other CVD risk factors. Such SNPs may violate the assumptions of IV and would need to be carefully considered when used in MR analysis.

#### 1.6.2.7 Linkage Disequilibrium

Genetic association studies rely on the LD between tag SNPs and functional variants. IV assumptions are not violated when tag SNPs are used as IVs, unless they are also in LD with a functional variant that affects the outcome through a pathway that does not include the risk factor of interest (Palmer et al. 2011).

### 1.6.2.8 Single SNP, Genetic Risk Score and Multiple SNP IVs

When multiple SNPs are associated with an exposure of interest, either a single SNP may be chosen as the IV, the SNPs can be used as multiple IVs , or multiple SNPs can be combined into a single genetic risk score IV (Figure 1.11).

**Figure 1.11 Causal diagram for a Mendelian randomisation study.** (a) a single SNP instrumental variable, (b) multiple, independent instrumental variables and (c) combining SNPs into a composite genetic score instrumental variable. The effect size for each relationship is denoted by β.

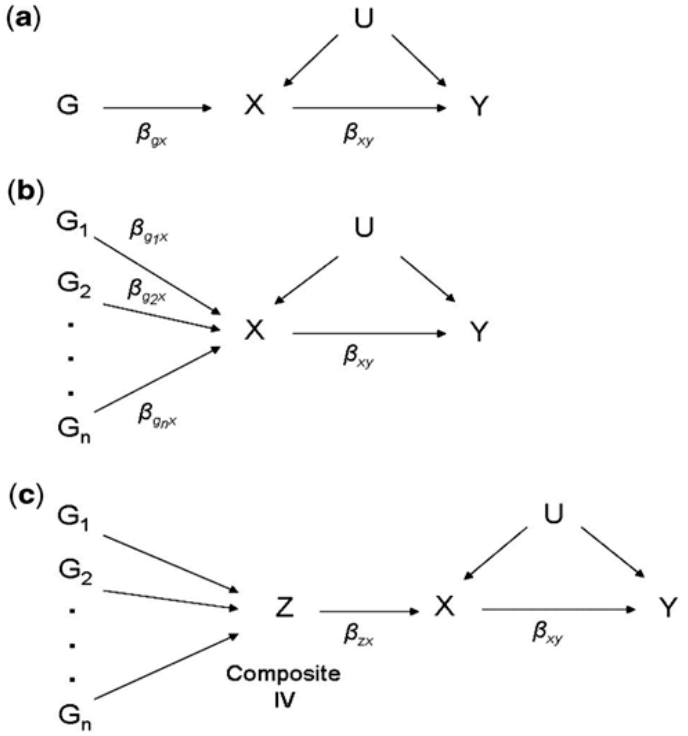


Image taken from Pierce et al (Pierce et al. 2010)

Previous lipid MR analyses have either used the single most significantly associated SNP as an instrument (Sarwar et al. 2010) or a combined genetic risk score where the lead SNP from each locus has been selected (Kathiresan et al. 2008; Levy et al.

2009). Methodological studies have shown that a combined genetic score or multiple instruments approach are more appropriate when multiple SNPs are associated with the exposure (Pierce et al. 2010; Burgess & Thompson 2011).

## 1.7 Thesis Outline

As the title of my thesis suggests, the focus of my PhD has been two-fold: Firstly, to improve the understanding of the biology behind well-known cardiovascular risk factors, and secondly to explore the application of known genetic variants associated with CVD risk factors in disease risk prediction, and in determining causality between CVD risk factors and clinically relevant outcomes.

In Chapter 2 I report the large-scale discovery of genetic variants associated with LV mass. As mentioned in section 1.5.1, the pathophysiological mechanisms of LVH remain incompletely characterised, and novel loci associated with LV mass may provide insight into the pathways involved in the development of LVH. Initially, digital ECG measures were available for around 5000 individuals from 20 London-based Civil service departments (Whitehall II (WHII) Cohort). These individuals had also been genotyped using the Cardiochip SNP array. To increase the sample size, and hence power for detection of genetic associations, collaboration with two additional UK studies: the British Women's Health and Heart Study (BWHHS) and the Genetic Regulation of Arterial Pressure of Humans in the Community (GRAPHIC) study doubled the sample size for discovery through meta-analysis of summary-level data. Both additional studies had digital ECG data, and study participants had already been genotyped using the Cardiochip. Significant associations were validated in three additional replication cohorts.

In Chapter 3, I explore the potential of known lipid-associated genetic variants in risk prediction for clinically relevant outcomes, including high CVD risk status, need for lipid therapeutic intervention, and CHD outcome. The ability of genetic data to discriminate individuals in each outcome category was compared with the commonly used non-genetic CVD risk score – the Framingham 10 yr CVD risk score. This work was based on lipid genetic variants identified by a large-scale association analysis in the WHII study published in 2009 (Talmud et al. 2009). Analysis was carried out in two British cohorts – WHII and BWHHS.

Finally, I use known genetic variants associated with lipids to determine their causal relationship with common CIMT. Clinical trials of drugs targeting HDL-C and triglycerides that use progression of CIMT as a marker for drug efficacy have provided contradictory results, leaving their causal role in atherosclerosis and CHD uncertain. Since many lipid-associated variants have been identified, I assessed the suitability of different approaches for instrument development for LDL-C, HDL-C and triglycerides in Chapter 3. The final instruments, together with a second set of independently derived instruments based on SNPs reported by the GLGC, were applied in an MR analysis in WHII, in around 3000 individuals, and in the IMT Progression as Predictors of Vascular Events in a High Risk European Population (IMPROVE) study, in around 3400 individuals, to determine the casual relationship between the three lipid fractions and common CIMT.

# 2 Discovery of Genetic Determinants of Left Ventricular Mass

## 2.1 Introduction

Measures of LV mass are used in the diagnosis of LVH, the abnormal enlargement of the LV muscle tissue, which is a major cause of morbidity and mortality. The relation between LV mass and cardiovascular risk has shown to be continuous (Levy et al. 1989; Schillaci et al. 2000), therefore the discovery of genetic factors that contribute to even small increases in LV mass will likely have clinical importance. Though echocardiography is more sensitive than ECG for detecting LVH, the cost and operational considerations limit its use in large-scale population studies. ECG data is more widely available in existing cohorts and several methods exist to calculate indices of LV mass from this data (refer to section 1.5.1.2).

Though both echocardiographic and ECG LV mass measures have shown to have significant heritability (Mayosi 2002), to date very few loci have been robustly associated with these traits. Two studies on echo-LV mass (including the largest GWAS to date for this trait, with discovery in 12,612 individuals  and replication in 4,094 individuals – the EchoGen consortium) reported no definite associations with LV mass (Vasan et al. 2009; Arnett et al. 2011). A study in a total of 202 individuals from the extreme tails of the LV mass distribution and replication in 704 Caucasian individuals reported associations in two regions (5p13.2 and 12q14.3) (Arnett et al. 2009), and a genome-wide linkage analysis found suggestive evidence for loci on chromosomes 10q23.1 for the Sokolow-Lyon index, and on 17p13.3 for the Cornell product (Mayosi et al. 2008). It is unclear whether genetic variation in genes causal of Mendelian forms (e.g. sarcomeric genes) influences less severe forms of LVH as these have not been identified in previous studies.

Given the lack of loci reported for LV mass, and the increasing interest of linking genetic variants affecting genes involved in cardiovascular disease pathways and the more common forms of LVH, a large-scale cardiovascular gene-centric analysis of four ECG-derived indices of LV mass (Sokolow-Lyon Index, Cornell Product, 12-lead QRS Voltage Sum and 12-lead QRS Voltage Product) was carried out in three population-based cohorts: British Women's Health and Heart Study (BWHHS), Genetic Regulation of Arterial Pressure of Humans in the Community (GRAPHIC) study and Whitehall II (WHII), with a total sample size over 10,000 individuals. All studies had previously collected digital ECG and biometric data, and genotyped individuals using the Illumina cardiovascular gene-centric 50K SNP array (Cardiochip) (Keating et al. 2008). Promising signals were replicated in three further cohorts, with a total sample size of 11,777 individuals.

## 2.2 Materials & Methods

### 2.2.1 Discovery Study Cohorts

#### 2.2.1.1 British Women's Heart and Health Study (BWHHS)

The BWHHS is a prospective cohort study of 4286 British women who were between the ages of 60 and 79 at baseline (1999 - 2001) (Ebrahim et al. 2008). Participants were randomly selected from general practice registers in 23 British towns (Lawlor et al. 2003). Baseline demographic, anthropometric, 12-lead ECG and biological data were collected between 1999 and 2001 and used in this analysis. During this time face-to-face interviews were also conducted for the completion of medical questionnaires, and a DNA repository was made. Ethical committee approval was obtained for the study.

#### 2.2.1.2 Genetic Regulation of Arterial Pressure of Humans in the Community (GRAPHIC) Study

The GRAPHIC study selected 2024 individuals from 520 nuclear families recruited from the general population in Leicestershire, UK, between 2003 - 2005 for the purpose of investigating the genetic determinants of blood pressure and related cardiovascular traits (Tobin et al. 2008). Families were included if both parents were aged between 40 and 60 years, and two offspring were 18 years or older and wished to participate. A detailed medical history was obtained from study subjects by standardised questionnaires, and clinical examinations were performed by research nurses following standard procedures. Blood samples and other measurements such as height, weight, waist-hip ratio, clinic and ambulatory blood pressure, and 12-lead ECG were obtained from participants. Ethical committee approval was obtained for the study.

#### 2.2.1.3 Whitehall II (WHII)

The WHII study recruited 10,308 participants (70% men) between 1985 and 1989 from 20 London-based Civil service departments (Marmot & Brunner 2005). The study was initially set up as a longitudinal study of cardiorespiratory disease and diabetes. Clinical measurements are taken every 5 years and postal questionnaires are conducted in between clinical phases. Clinical data were available from four phases: 1985 - 1988, 1991 - 1993, 1995 - 1999 and 2003 - 2004. Clinical and questionnaire data collected between 1991 and 1993 provided the first comprehensive phenotypic dataset for WHII and is considered the baseline phase. By 2003, only 6914 of the original 10,308 participants attended the clinic. Blood samples for DNA were collected between 2002 and 2004. For the purpose of this study, data collection from 2003 - 2004 provided the most comprehensive ECG data, and these were used in the calculation of the ECG-LV mass indices investigated in this analysis. Participant age during this period ranged between 50 and 75. Ethical committee approval was obtained for the study.

## 2.2.2 Replication Study Cohorts

### 2.2.2.1 British Regional Heart Study (BRHS)

The initial focus of the BRHS was on the prevalence and incidence of CVD and their relations to established behavioural and biological risk factors (Walker et al. 2004). The study comprises of 7735 men aged between 40 and 59 years recruited from 24 medium sized British towns between 1978 and 1980. Clinical measurements were made at baseline. Twenty years later (1998 - 2000) participants were re-measured, including the application of 12-lead ECG, and whole blood samples taken for DNA analysis. Phenotypic measures from the follow-up phase (1998 - 2000) were used in this analysis. Ethical committee approval was obtained for the study.

### 2.2.2.2 British Genetics of Hypertension (BRIGHT) Study

The BRIGHT study comprises of hypertensive families recruited between 1996 and 2002. Cases were defined as having blood pressure readings ≥150/100 mmHg (if based on one reading), or ≥145/95 mmHg (if based on the mean of three readings). Each family contained at least two affected siblings, in whom onset of hypertension was diagnosed before the age of 60 years. Hypertensive individuals who self-reportedly consumed more than 21 units of alcohol per week; had diabetes; had intrinsic renal disease; had a self-reported history of secondary hypertension that was confirmed by the family physician; or had coexisting illness, were excluded. Recruitment was aimed for hypertensive individuals with BMI less than 30 kg/m$^2$. ECGs were obtained at the time of recruitment. Only single individuals from each family were genotyped (Caulfield et al. 2003). Ethical committee approval was obtained for the study.

#### 2.2.2.3 Prevention of Renal and Vascular End Stage Disease (PREVEND) Study

The PREVEND study is a prospective investigation of the natural course of albuminuria (elevated levels of albumin in the urine), and its relationship to renal and cardiovascular disease. The patients of the PREVEND cohort were selected in 1997 from 40,856 individuals from the general population in the Netherlands (Smilde et al. 2005). In total, 8592 subjects were included in the PREVEND baseline cohort. At baseline (1997-1998) biometric measurements were taken; participants completed a questionnaire on demographics, CVD history, renal disease history, and use of hypertensive medication; blood pressure was measured; a fasting blood sample was drawn and standard 12-lead ECGs were recorded. Ethical committee approval was obtained for the study.

### 2.2.3 Calculation of ECG-LV Mass Indices

For all studies the standard 12-lead ECG digital data was transferred to the University of Glasgow ECG Core Lab based at the Glasgow Royal Infirmary, where they were reviewed manually and checked for technical problems which would have interfered with analysis. Technically unsatisfactory ECGs (which may include ECGs of individuals with existing CHD) were excluded. The reviewed ECGs were then analysed by the University of Glasgow ECG analysis program (Macfarlane et al. 2005) and four LV mass indices (Sokolow-Lyon Index, Cornell Product, 12-lead QRS Voltage Sum and 12-lead QRS Voltage Product) generated (refer to Figure 1.2 for calculation of these measures). This software meets all of the required specifications in terms of measurement accuracy and is used widely in various commercial products and clinical trials. For each ECG-LV mass index, outliers more than 3 standard deviations (SD) away from the mean were excluded from the analysis. Based on the observed distribution of the phenotypic measures, analysis was done on untransformed data.

### 2.2.4 Genotyping

The discovery cohorts were all genotyped using the ITMAT Broad-CARe (Cardiochip) (Keating et al. 2008). The Cardiochip is a gene-centric SNP array containing ~50,000 SNPs covering ~2000 loci that are known to be involved in cardiovascular pathways, as determined from GWAS of vascular and inflammatory disease and comprehensive literature searching. During array design, gene loci were prioritised into 3 density groups:

- **Group 1** (n = 435 loci): Genes and regions with a high likelihood of functional significance, including established mediators of vascular disease, loci derived from GWAS and those shown to be associated with cardiovascular phenotypes of interest. SNPs were inclusive of the intronic, exonic, untranslated regions (UTRs) and 5 kilobases of the proximal promoter regions. The average number of SNPs in the Group 1 loci is ~35.6 (Keating et al. 2008)

- **Group 2** (n = 1,349 loci): Candidate loci that are potentially involved in phenotypes of interest. SNPs were inclusive of intronic, exonic and flanking UTRs. The average number of SNPs in the Group 2 loci is ~16.3 (Keating et al. 2008)

- **Group 3** (n = 232 loci): Comprised mainly of the larger genes (>100 kb) which were of lower interest *a priori*. Only non-synonymous SNPs (nsSNPs) and known functional variants of MAF>0.01 were captured for these loci (Keating et al. 2008)

Genotyping in each study was done at different genotyping centres. However, for all studies that used the Cardiochip, genotypes were called using the Illumina BeadStudio (version 3) Genotyping Module using the default GenCall software application (Illumina 2005) to automatically cluster and call genotypes from the intensity data. The same quality control (QC) criteria were applied, in the order specified below, to each discovery study to ensure data integrity.

1. Exclusion of low-performing SNPs, where the percentage of missing calls per SNP was > 10%.

2. Exclusion of low-quality samples with percentage missing calls > 5%.

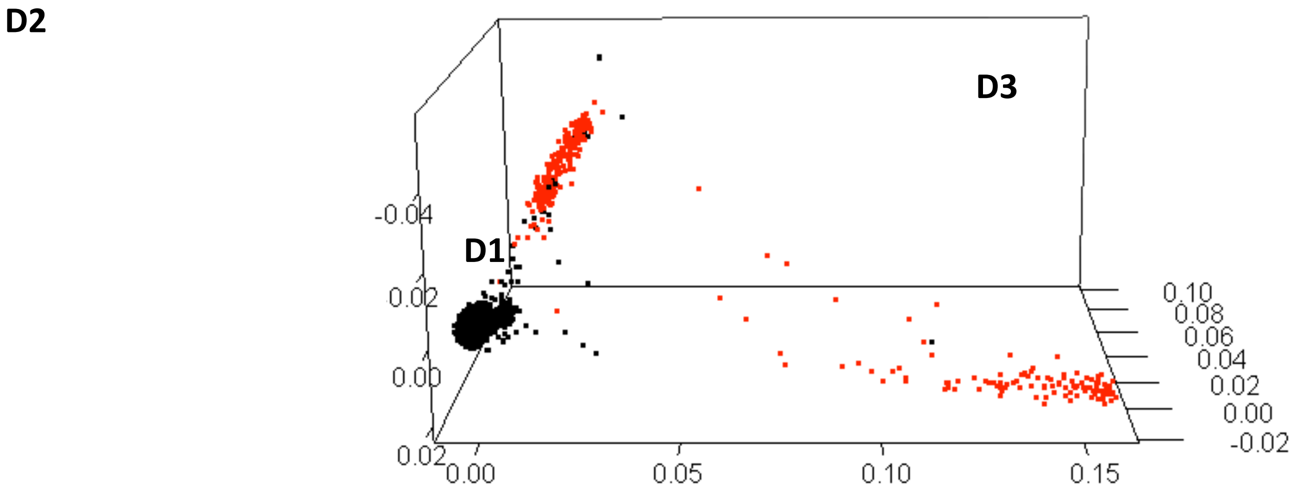
3. Identity-by-descent (IBD) was used to estimate relatedness between every pair of individuals in the dataset. IBD is a measure of how many alleles at any marker in each of the two samples came from the same ancestral chromosomes. IBD was calculated using PLINK software (Purcell et al. 2007) which uses a Hidden Markov Model in which the hidden IBD state is estimated given the observed identity-by-state (IBS - a measure of how many alleles at any marker in each of the two samples happen to be the same). IBD should identify known duplicates that have been genotyped multiple times for quality control purposes. IBD will also identify individuals in the study sample that have unknown residual, non-trivial degrees of relatedness, which can violate the independence assumptions of standard statistical techniques (McCarthy et al. 2008). Unexpected duplicate pairs may indicate sample mix-up or sample contamination. Given the gene-centric design of the array with dense coverage of loci, there are many clusters of highly correlated SNPs. To avoid biases from groups of correlated markers, pairwise IBD was performed on a pruned set of 'independent' SNPs selected based on a SNP pairwise correlation threshold ($R^2$<0.5). One sample from each known duplicate or related pair was removed from further analysis. Unknown duplicate samples are likely to arise from contamination

or mix-up, and these were also excluded. Since GRAPHIC is a family-based study, related individuals were included in the data and family structure was taken into account in downstream analysis.

4. Multidimensional scaling (MDS) was used to investigate any structure in the data which may be due to population structure, family relatedness, long-range LD or genotyping assay artefacts. MDS is a method that represents measurements of similarity (or dissimilarity) among pairs of objects as distances between points in a low-dimensional space (Borg & Groenen 2005). MDS can be applied to either IBD or IBS to identify population structure and outliers. In this case, MDS was applied to the pairwise IBS distances (calculated using the same set of independent SNPs as for the IBD calculation) to cluster groups of individuals with similar genotypes. MDS allows the representation of the genotype data in a lower-dimensional space, in this case only 3 dimensions were used, which enables the visualisation of any significant structure present in the data. Together with self-reported ethnicity, arbitrary cut-offs were used to exclude non-Caucasian samples and any additional outliers based on the plot of the first 3 dimensions from the MDS analysis. This is shown for WHII data in Figure 2.1.

5. Samples where the genotype-inferred sex did not match the reported sex were excluded. This was determined by looking at the call rates of the Y chromosome SNPs and also by calculating the homozygosity rate of X chromosome SNPs. Discordance between reported and genetically-estimated gender is most likely to occur due to sample mix-up or contamination.

**Figure 2.1 Population structure in WHII based on multidimensional scaling**. The first 3 dimensions are plotted below. Individuals in black have self-reported Caucasian ethnicity and those in red have self-reported non-Caucasian ethnicity. The three clusters correspond to Caucasian, Asian and African ancestry. Individuals that clustered away from the Caucasian individuals were assumed to be ethnic outliers and excluded from further analysis.



6. The Hardy-Weinberg equilibrium (HWE) test determines whether the observed genotype frequencies for a SNP significantly differ from the expected frequencies, and is used to flag poorly performing assays and anomalous genotype clustering. The HWE test was performed after removal of outliers and in founders only. SNPs with HWE p-value < 1 x $10^{-04}$ are considered to be of poor quality (Laurie et al. 2010) and were excluded from further analysis.

7. Rare SNPs are more prone to error, as fewer samples would be within a genotype cluster and most clustering-based calling algorithms do not perform well with rare alleles (Neale and Purcell 2008; Teo 2008). The power to detect association is also much lower for rare SNPs. Therefore, SNPs with minor allele frequency (MAF) < 1% were excluded from analysis.

SNPs and samples that passed the above QC filters were used in the discovery association analysis. For replication of the SNPs taken forward from the discovery phase, genotypes were generated *de novo* for the PREVEND Study and BRHS using

KASPAR assays (KBioSciences) and extracted *in silico* from Cardiochip array data for the BRIGHT Study. Replication studies with existing Cardiochip data used their own QC steps and thresholds, which were similar to those used for the discovery studies. For all studies analysis was restricted to Caucasian samples.

### 2.2.5 Statistical Analysis

#### 2.2.5.1 Within-Study Association Analysis

In each of the discovery cohorts, linear regression analyses were performed for each SNP with each ECG-LV mass index assuming a per-allele additive genetic model. Age, sex, BMI, and systolic blood pressure (SBP) were added as covariates in the model. For individuals on blood pressure lowering medication, SBP was adjusted by adding a constant of 15mmHg based on a study comparing methods for adjusting blood pressure for treatment effects (Tobin et al. 2005). Covariates were selected based on prior knowledge of non-genetic risk factors of LVH. In GRAPHIC, additional adjustments for age$^2$ (due to the older age of parents compared to children in the cohort) and familial correlation were taken into account using generalised estimating equations with an exchangeable correlation structure (Liang & Zeger 1986).

#### 2.2.5.2 Between-Study Meta-Analysis

Meta-analysis is a statistical method for combining results (in this case the per-allele beta-coefficients) from multiple independent studies to estimate the combined effect. This has more power to detect an effect than any of the studies individually. Since some studies may have more precise estimates than others, rather than calculating a simple mean of the effect sizes, a weighted mean is calculated where the study-specific effects are weighted by the inverse of the study variance i.e. studies with large variance will be down-weighted and contribute less to the overall effect estimate. Meta-analysis can be performed using a fixed or random effects model. Under the fixed effect model it is assumed that there is one true effect size which is shared by all the included studies. By contrast, the random effects model allows the true effect to vary from study to study. This between study heterogeneity is incorporated into the meta-analysis.

The relative strengths of fixed and random effects analyses remain controversial (Thompson et al. 2011) and both have been applied in published meta-analyses. Heterogeneity was measured using $I^2$ (Higgins et al. 2003), which describes the percentage of variation across studies attributed to heterogeneity rather than chance. An $I^2$ value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity. Because ~30% of SNPs passing the discovery threshold (specified below) showed moderate to high heterogeneity ($I^2$ between 40 and 70), a random-effects model was applied using the commonly used DerSimonian and Laird procedure for a random effects meta-analysis (DerSimonian & Laird 1986), whereby the between-study heterogeneity estimates are used to adjust the standard errors of each study-specific estimate. The fixed effect meta-analysis did not identify SNPs in any additional genes to the ones reported by the random effects model. Therefore only results from the random effects model are reported here.

If a Bonferroni correction for multiple testing of 33,950 SNPs that passed QC in all three discovery cohorts was applied, the p-value threshold would have been $1.47\text{x}10^{-6}$. However, given the lower number of independent SNPs, the higher prior odds of association due to the informed selection of loci covered by the array and the opportunity to eliminate false positives at the replication stage, the discovery threshold to take SNPs forward for replication was relaxed to $1\text{x}10^{-4}$.

### 2.2.5.3 Conditional Analysis

Typically many SNPs in a region harbouring one or more causal variant(s) demonstrate univariate associations with the traits of interest but the majority of these associations are indirect and operate through LD with the causal site(s). When multiple SNPs within the same locus reached significance, conditional association analyses were performed, whereby the lead (most significant) SNP was added to the linear regression model as a covariate. Any SNPs that remained significant at the discovery threshold in conditional analysis were considered to be independent association signals from the lead SNP. If the most significant SNP in the conditional analysis passed the discovery p-value threshold it was also added as a covariate to the regression model in addition to the original lead SNP. This process was repeated until no more SNPs showed an association p-value less the discovery threshold. All independent signals were taken forward for replication.

### 2.2.5.4 Genomic Inflation

When association analysis is carried out on a large number of SNPs it is important to test the distribution of the test statistic in comparison with the expected null distribution. Any deviation of the observed test statistic distribution from the null distribution may suggest systematic bias (from unrecognized population structure, analytical approach, genotyping artefacts, array design etc). This deviation is quantified by calculating the genomic inflation factor and can be visualised by quantile–quantile (QQ) plots. The genomic inflation factor, lambda, is defined as the ratio of the median of the observed distribution of the test statistic to the expected median, thus quantifying the extent of the bulk inflation (Devlin & Roeder 1999).

### 2.2.5.5 Replication

Linear regression analysis was carried out in each of the three replication cohorts for the selected SNPs with the ECG-LV mass trait with which they were significantly associated in the discovery stage. As before, age, sex, BMI and corrected SBP were added as covariates in the regression model. A meta-analysis of the results from the replication studies was done using the same method as in the discovery phase. A Bonferroni correction for the number of independent signals taken forward for replication (12 SNPs) was applied to the standard 0.05 significance threshold, giving a replication threshold of $p=4.17x10^{-3}$.

### 2.2.5.6 Calculation of Variance Explained

A univariate linear regression was used to determine the proportion of trait variation ($R^2$) explained by each of the four replicated SNPs within each study. In GRAPHIC only the parental generation was used to calculate the $R^2$ value.

### 2.2.5.7 Top Decile Analysis

To examine the effect of each of the four replicated SNPs on the odds of being in the top decile of their associated trait distribution i.e. QRS Voltage Sum (rs2290893, rs2292462, rs4966014) and Cornell Product (rs6797133), logistic regression was performed in each discovery study (adjusting for the same covariates as previously), and the results meta-analysed using random effects. In addition, to look at the combined effect of the three QRS Voltage Sum-associated SNPs, a genetic risk score was generated for each individual which was the sum of the number of risk alleles (trait-raising alleles) across the three SNPs (for a single SNP, an individual can have 0, 1 or 2 risk alleles), assuming a per-allele additive effect. Individuals carrying 0-2 risk alleles were used as the reference group, and the odds of being in the top decile of QRS Voltage Sum were calculated for subjects carrying 3, 4, 5 or 6 risk alleles.

### 2.2.6 Functional Analysis

Bioinformatics resources were used to assess if the replicated SNPs, or SNPs in strong LD with them, could have a functional impact on nearby genes. For SNPs found in coding regions, their likelihood of affecting protein structure and function was assessed using Polyphen (http://genetics.bwh.harvard.edu/pph/), which uses structural knowledge, evolutionary conservation and knowledge of functional sites to assess how likely an amino acid substitution will alter protein structure and function (Adzhubei et al. 2010). For SNPs found outside of coding regions the RegulomeDB bioinformatics server (http://regulome.stanford.edu/) was used to assess whether the SNPs were found within any regulatory regions, which may suggest a regulatory role on gene expression. The RegulomeDB server uses publicly available data to annotate SNPs within known and predicted regulatory elements. One such available resource is the Encyclopaedia of DNA Elements (ENCODE) (Birney et al. 2007). The ENCODE project used genome-wide experimental techniques to identify the following regulatory regions in multiple cell types:

1. Transcription factor (TF) binding sites identified by chromatin immuno-precipitation with sequencing (ChIP-Seq) (Landt et al. 2012). TF-bound DNA from nuclear extract can be isolated using a TF-specific antibody (by chromatin immuno-precipitation) and then sequenced. Mapping the sequenced DNA fragments to the genome locates the regions where the TF was bound.

2. Open chromatin regions – Active regulatory regions (including enhancers, silencers and promoters) tend to have an open chromatin structure which allows access to DNA-binding regulatory proteins. The open nature of the

chromatin structure also increases sensitivity of these regions to digestion by nucleases such as DNase I. DNase I hypersensitive sites sequencing (DNase-Seq) (Song & Crawford 2010) was used to map DNase I hypersensitivity sites across the genome. Open chromatin regions can also be identified by Formaldehyde-Assisted Isolation of Regulatory Elements with sequencing (FAIRE-Seq) which separates nucleosomes from open chromatin. Sequencing of these open chromatin regions followed by mapping onto the genome provides the location of these regions.

3. Histone-methylation - chemical modifications (e.g. methylation and acylation) to the histone proteins present in chromatin influence gene expression by changing how accessible the chromatin is to transcription. Specific histone modifications can differentiate between promoter and enhancer-sites e.g. H3K4Me3 is Histone H3 with the addition of 3 methyl groups to Lysine reside 4, which is associated with active promoters. The same ChIP-Seq method using antibodies specific for each type of histone modification can be applied to identify regions where specific histone modification has occurred.

In addition, the RegulomeDB server queries TF binding site databases which contain information on both predicted and experimentally validated TF binding sites. Another source of functional information is expression quantitative trait loci (eQTL) data. These are genomic loci that regulate expression levels of mRNAs. RegulomeDB queries eQTL data from lymphoblastoid cells (Montgomery et al. 2010). The Genevar v3.2.0 software (Yang et al. 2010) was also used to query eQTL datasets from lymphoblastoid cells, skin cells, adipose tissue, T-cells and fibroblasts (Grundberg et al. 2012; Stranger et al. 2012; Nica et al. 2011; Dimas et al. 2009).

### 2.2.7 Comparison of ECG- and Echo-LV Mass-Associated Variants

To assess if any of the replicated SNPs were also associated with echo-LV mass, a look-up of their association in the EchoGen Consortium (Vasan et al. 2009), who were responsible for the largest published GWA meta-analysis on echo-LV mass to date, was carried out. Briefly, EchoGen had performed meta-analysis of GWA data from 5 population-based cohort studies (Cardiovascular Health Study, Framingham Heart Study, Rotterdam Study, Multinational Monitoring of Trends and Determinants in Cardiovascular Disease study (MONICA-KORA), and Gutenberg Heart Study) with a total sample size of 12,612. Within each study summary estimates for each SNP were obtained using linear regression assuming an additive model with age, sex, height and weight as covariates, and a meta-analysis carried out using an inverse variance fixed effect model.

### 2.2.8 Association with Candidate Genes

Previous studies have suggested association of variants in genes of the renin-angiotensin system cascade with LVH, notably the A1166C variant in angiotensin II receptor, type 1 (*AGTR1*), the M235T polymorphism in angiotensinogen (*AGT*), the insertion/deletion (I/D) polymorphism in the *ACE* gene and the -344 C/T polymorphism in cytochrome P450, family 11, subfamily B, polypeptide 2 (*CYP11B2*). Genes involved in haemodynamic load, calcium homeostasis have also been suggested to play a role in LVH development and causal mutations in sarcomeric-protein coding genes are known to be responsible for Mendelian forms of cardiac hypertrophy. Association of such genes implicated in LVH (identified through literature search) were also reported.

## 2.3 Results

### 2.3.1 Cohort Characteristics

After QC, the discovery sample with genotype data comprised of 10,497 individuals: 3414 from BWHHS, 2024 from GRAPHIC and 5059 from the WHII Study. Replication was undertaken in 3 additional cohorts (PREVEND, BRHS and BRIGHT) totalling 11,777 individuals. Discovery and replication cohort characteristics are shown in Table 2.1.

### 2.3.2 Phenotype Distribution

The distributions of the ECG-LV mass measures were very similar in each cohort (Figure 2.2). There was strong intra-individual correlation between 12-lead QRS Voltage Sum and 12-lead QRS Voltage Product, moderate correlations between either Cornell Product or Sokolow-Lyon index and the 12-lead QRS indices and no correlation between Cornell Product and Sokolow-Lyon index (Table 2.2).

### 2.3.3 SNP Quality Control Analysis

After genotyping QC, over 34,000 SNPs in each study remained for analysis. A large number of these were excluded due to low MAF (<1%), since rare SNPs are more prone to genotyping error. A total of 33,950 SNPs were present in all three studies after applying QC thresholds.

**Table 2.1 Demographic and clinical characteristics for discovery and replication studies.**

| Discovery Cohorts | Whitehall II | | BWHHS | | GRAPHIC | |
|---|---|---|---|---|---|---|
| | Males<br>(N=3721) | Females<br>(N=1338) | Males<br>(N=0) | Females<br>(N=3414) | Males<br>(N=1021) | Females<br>(N=1003) |
| **Age(yrs)** | 60.8(5.9) | 61.2(6.1) | | 68.9(5.5) | 39.4(15.1) | 39.2(13.9) |
| **BMI(kg/m$^2$)** | 26.6(3.8) | 27.0(5.5) | | 27.8(4.9) | 26.4(4.3) | 25.8(4.9) |
| **DBP(mmHg)** | 75.1(10.4) | 73.2(10.7) | | 79.4(11.9) | 73.5(8.1) | 69.9(6.8) |
| **SBP(mmHg)** | 128.7(16.1) | 126.5(18.2) | | 147.0(25.1) | 122.5(10.1) | 114.9(9.8) |
| **% On BP-lowering drugs** | 23.1 | 22.1 | | 30 | 7.9 | 5.4 |
| **Corrected SBP(mmHg)** | 132.2(18.1) | 129.7(20.6) | | 151.6(27.3) | 123.4(10.9) | 115.5(10.6) |
| **Cornell Product(μV.s)** | 147.5(58.8) | 163.8(54.1) | | 183.9(53.3) | 137.0(48.6) | 143.5(40.0) |
| **QRS Voltage Product(μV.s)** | 1437.5(398.9) | 1127.7(312.6) | | 1197.6(340.4) | 1570.3(433.2) | 1130.4(265.8) |
| **QRS Voltage Sum(μV)** | 14740.4(3056.7) | 12488.1(2629.2) | | 13158.9(2812.2) | 16351.1(3830.4) | 12923.7(2636.3) |
| **Sokolow-Lyon Index(μV)** | 2156.2(718.1) | 1924.6(587.9) | | 1850.3(541.0) | 2395.7(643.6) | 2125.1(539.2) |
| **Replication Cohorts** | **BRHS** | | **BRIGHT** | | **PREVEND** | |
| | Males<br>(N=3519) | Females<br>(N=0) | Males<br>(N=485) | Females<br>(N=713) | Males<br>(N=3298) | Females<br>(N=3762) |
| **Age(yrs)** | 68.8(5.5) | | 56.7(10.7) | 57.7(10.1) | 49.6(12.8) | 47.9(12.1) |
| **BMI(kg/m$^2$)** | 26.8(3.7) | | 27.7(3.8) | 27.1(4.8) | 26.2(3.6) | 25.8(4.6) |
| **DBP(mmHg)** | 85.2(11.1) | | 96.8(11.0) | 91.7(12.0) | 76.7(9.6) | 70.9(8.9) |
| **SBP(mmHg)** | 149.2(24.2) | | 158.0(20.9) | 151.2(21.8) | 133.3(18.2) | 123.8(20.3) |
| **% On BP lowering drugs** | 28.0 | | 100 | 100 | 14.6 | 13.5 |
| **Corrected SBP(mmHg)** | 153.4(25.8) | | 172.9(20.9) | 166.2(21.9) | 135.4(20.2) | 125.8(22.5) |
| **Cornell Product(μV.s)** | 143.4 (65.0) | | 174.3(94.3) | 188.3(73.3) | 126.7(55.7) | 146.4(46.4) |
| **QRS Voltage Product(μV.s)** | 1384.6(405.2) | | 1471.5(413.6) | 1192.4(335.7) | 1576.6(365.8) | 1268.5(302.5) |
| **QRS Voltage Sum(μV)** | 13807.9(2887.8) | | 14587(3407) | 12563(2990) | 15700.1(3349.6) | 13761.8(3020.5) |
| **Sokolow-Lyon Index(μV)** | 2069.5(677.3) | | 2326(743) | 2180(650) | 2522.5(695.6) | 2169.0(591.7) |

BMI=Body Mass Index; DBP=Diastolic Blood Pressure; SBP=Systolic Blood Pressure; μV=microvolts; μV.s=microvolt seconds

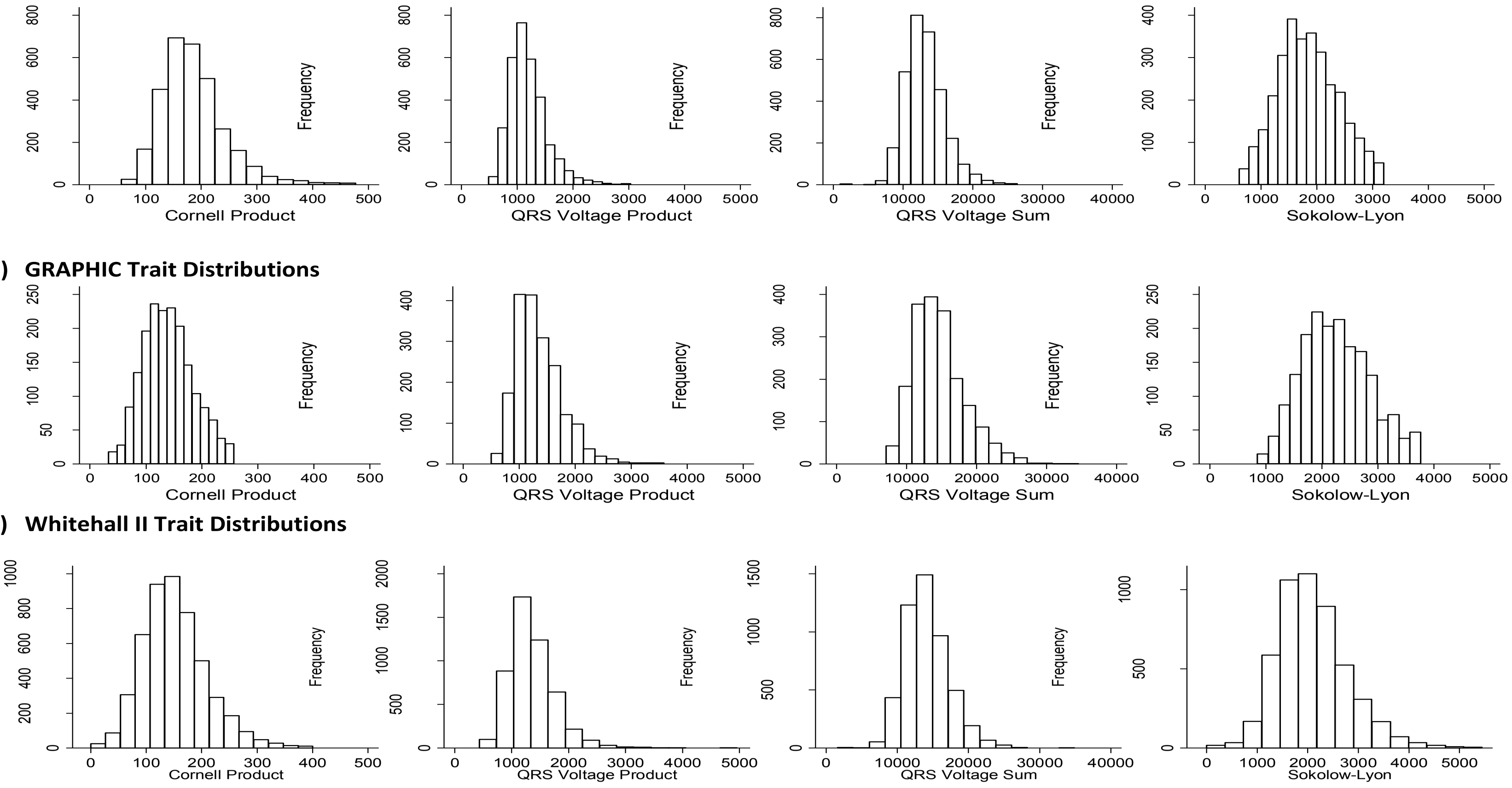
**Table 2.2 Pearson correlation coefficients for traits and covariates.**

| | | **Cornell Product** | **Sokolow-Lyon** | **QRS Voltage Sum** |
|---|---|---|---|---|
| **Sokolow-Lyon** | BWHHS | -0.0139 | | |
| | GRAPHIC | -0.0588 | | |
| | WHII | -0.0265 | *1.000* | |
| **QRS Voltage Sum** | BWHHS | 0.3917 | 0.614 | |
| | GRAPHIC | 0.1199 | 0.7499 | |
| | WHII | 0.330 | 0.684 | *1.000* |
| **QRS Voltage Product** | BWHHS | 0.6138 | 0.429 | 0.8434 |
| | GRAPHIC | 0.2121 | 0.6425 | 0.9269 |
| | WHII | 0.444 | 0.500 | 0.880 |

### 2.3.4 Association Analysis

A total of 47 SNPs in 12 loci passed the discovery meta-analysis p-value threshold of 1x10$^{-4}$. The QQ plots for each cohort show that the distribution of the observed p-values did not deviate much from the expected distribution (Figure 2.3), with genomic inflation factors ranging between 1.000 and 1.09. Conditional analysis of the lead SNPs in each locus did not identify any additional independent effects and therefore only the lead SNPs were selected for replication. These comprised six SNPs selected on the basis of an association with 12-lead QRS Voltage Sum, one with 12-lead QRS Voltage Product, four with Cornell Product and one with Sokolow-Lyon Index (Table 2.3).

**Figure 2.2 ECG-LV Mass distributions in the discovery cohorts.**



**A) BWHHS Trait Distributions**

**B) GRAPHIC Trait Distributions**

**C) Whitehall II Trait Distributions**

### 2.3.5 Replication

Of the 12 SNPs taken forward for replication, four showed evidence of significant association for their specific trait in the replication studies after allowing for multiple testing (p-value < 4.17x10$^{-3}$) (Table 2.3). These were variants in the *PTGES3* (12q13.3), *NMB* (15q25.2) and *IGF1R* (15q26.3) genes for 12-lead QRS Voltage Sum and the *SCN5A* (3p22.2) gene for Cornell Product. Meta-analysis of the combined discovery and replication data gave p-values of 3.74x10$^{-8}$, 3.23x10$^{-9}$, 1.26x10$^{-7}$ and 1.22x10$^{-7}$ for the lead SNPs in the *PTGES3 (*rs2290893), *NMB* (rs2292462), *IGF1R* (rs4966014) and *SCN5A* (rs6797133) loci, respectively. For each of the replicated loci the study-specific estimates for each SNP and the pattern of association across the four ECG-LV mass traits in the regions are shown in Figure 2.4 – Figure 2.11. Summary findings in the replication samples for the 8 SNPs that did not pass the replication p-value threshold are also shown in Figure 2.12 A-H.

**Figure 2.3 Quantile-quantile plots for each trait in each study.** Genomic inflation factors (lambda) are also shown.

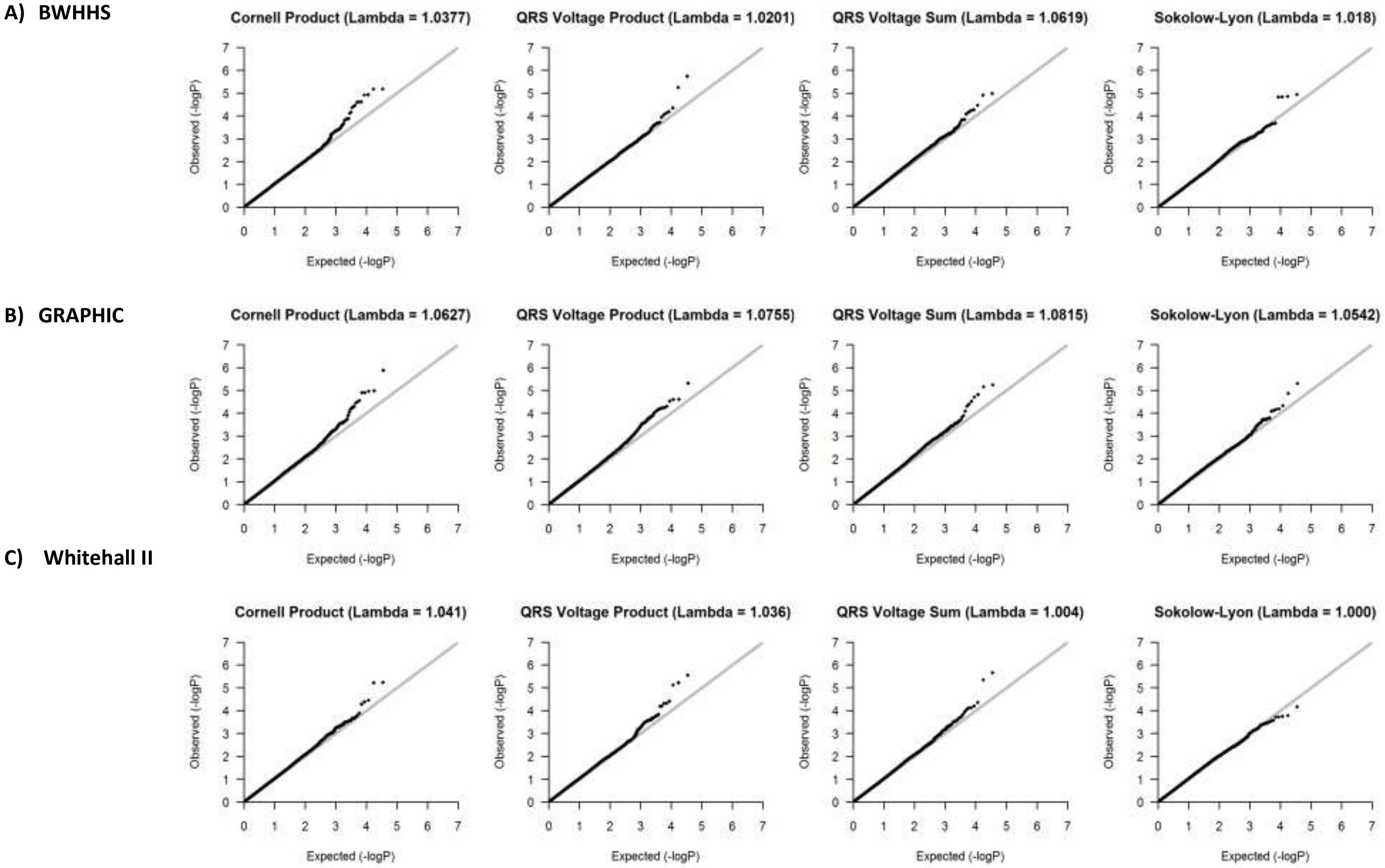


**Table 2.3 Meta-analysis results for the 12 SNPs associated with ECG-LV mass indices.** The table shows the discovery statistics for the 12 SNPs taken forward for replication. Results from the meta-analysis of the replication samples are shown in the penultimate two columns. Loci that were replicated ($p<4.17x10^{-3}$) are shown in bold. CHR=chromosome; N=total number of subjects analysed in the discovery cohorts for the specific SNP; Mean MAF=Mean of minor allele frequency.

| Phenotype | CHR | Gene Locus | SNP | Coded allele | Non-coded allele | N | Mean MAF | Discovery Beta (95% CI) | Discovery p-value | Replication Beta (95% CI) | Replication p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cornell Product | **3p22.2** | ***SCN5A*** | **rs6797133** | **A** | **G** | **9367** | **0.40** | **-3.7(-5.3,-2.2)** | **$3.01x10^{-6}$** | **-2.1(-3.6,-0.7)** | **$3.87x10^{-3}$** |
| | 4q26 | *MYOZ2* | rs9993110 | T | C | 9362 | 0.45 | 3.3(1.8,4.9) | $1.88x10^{-5}$ | 1.1(-0.3,2.6) | 0.11 |
| | 11q12.1 | *SERPING1* | rs11603020 | C | T | 9375 | 0.27 | 3.4(1.7,5.1) | $7.91x10^{-5}$ | 0.6(-1.7,2.9) | 0.59 |
| | 12q12 | *TMEM117* | rs860867 | T | C | 9387 | 0.10 | -5.1(-7.7,-2.6) | $5.94x10^{-5}$ | 1.3(-1.0,3.7) | 0.27 |
| QRS Voltage Product | 6p21.31 | *PPARD* | rs4713854 | C | A | 9410 | 0.11 | -33.3(-49.5,-17.1) | $5.84x10^{-5}$ | -0.4(-15.3,14.5) | 0.68 |
| QRS Voltage Sum | 4q26 | *FABP2* | rs11724758 | A | G | 9413 | 0.47 | 176.5(95.2,257.8) | $2.09x10^{-5}$ | 83.6(5.52,161.7) | 0.036 |
| | 5p15.31 | *MTRR* | rs3815743 | G | A | 9386 | 0.16 | -302.8(-412.7,-192.8) | $6.76x10^{-8}$ | -22.6(-125.8,80.5) | 0.51 |
| | 11q12.1 | *TCN1* | rs17597065 | G | A | 9427 | 0.04 | -396.6(-592.6,-200.6) | $7.32x10^{-5}$ | 110.0(-175.8,395.8) | 0.45 |
| | **12q13.3** | ***PTGES3*** | **rs2290893** | **A** | **G** | **9427** | **0.35** | **-201.4(-285.6,-117.1)** | **$2.84x10^{-6}$** | **-126.7(-205.4,-48.1)** | **$1.59x10^{-3}$** |
| | **15q25.3** | ***NMB*** | **rs2292462** | **G** | **T** | **9427** | **0.45** | **-218.6(-315.7,-121.5)** | **$1.02x10^{-5}$** | **-164.3(-249.8,-78.7)** | **$1.68x10^{-4}$** |
| | **15q26.3** | ***IGF1R*** | **rs4966014** | **C** | **T** | **9437** | **0.30** | **-181.8(-269.3,-94.4)** | **$4.65x10^{-5}$** | **-143.5(-225.7,-61.3)** | **$6.23x10^{-4}$** |
| Sokolow-Lyon | 2q31.2 | *TTN/PRKRA* | rs10497520 | T | C | 9321 | 0.13 | -68.9(-94.6,-43.1) | $1.58 x10^{-7}$ | -24.3(-48.0,-0.55) | 0.045 |

**Figure 2.4 Genetic association of rs6797133 (*SCN5A*) with Cornell Product.** Associations are reported by study separately for the discovery and replication cohorts, together with pooled discovery, replication and overall estimates. Beta-coefficients (with 95% confidence intervals) describe per allele effect of the minor allele of the SNP for the trait shown. A negative beta indicates that the trait had a lower value in those carrying the minor allele. The heterogeneity index ($I^2$) value is shown in the bottom left hand corner.

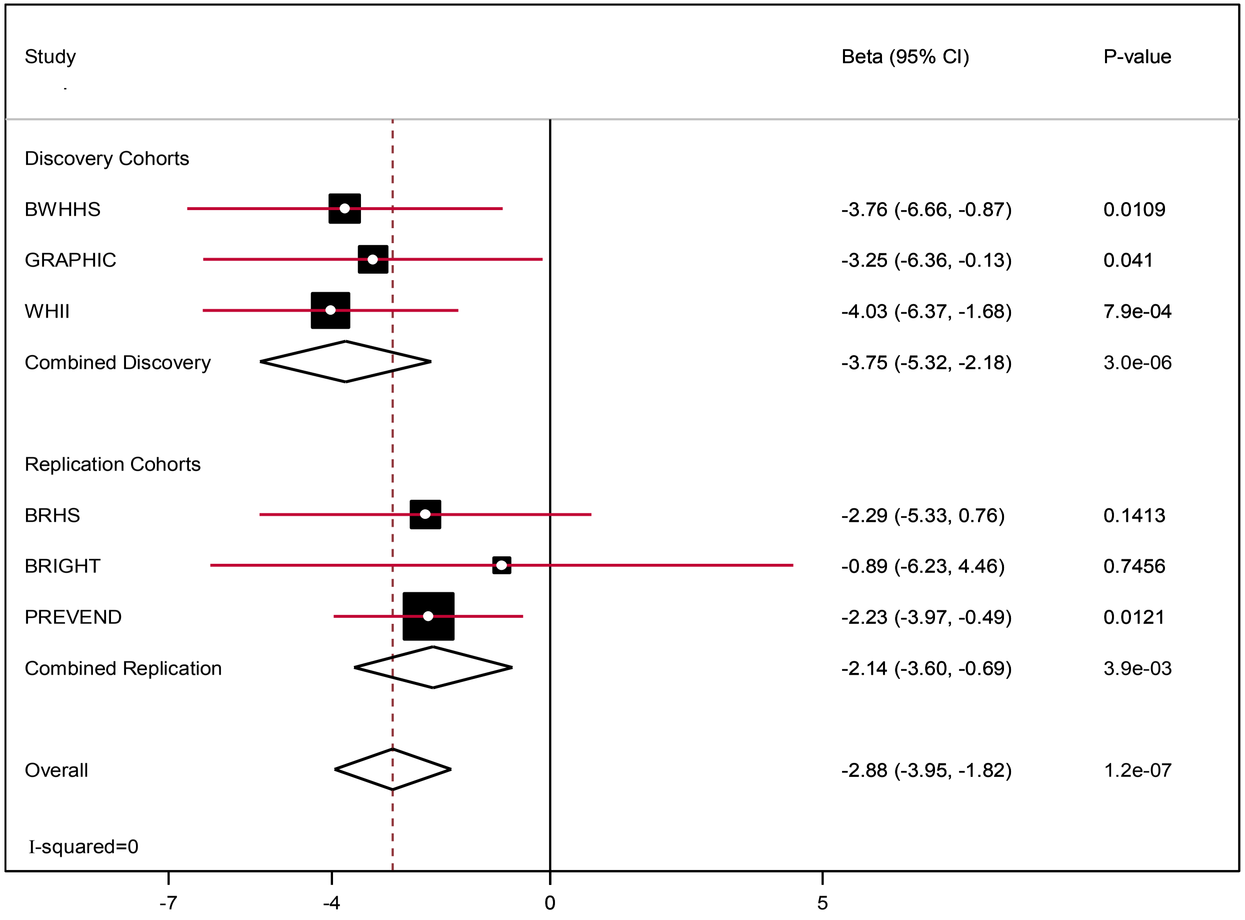


**Figure 2.5 Association signals around the *SCN5A* gene with ECG-LV mass indices.** A snapshot from the UCSC genome browser (http://genome.ucsc.edu/) showing bar plots of the $-\log_{10}$(p-values) from the discovery meta-analysis of all SNPs on the array in and around the *SCN5A* gene with all four ECG-LV mass indices. The horizontal lines at $-\log_{10}$(p-values) = 4 represent the discovery p-value threshold. Genes (exons as boxes and introns as horizontal lines) within the regions are displayed at the bottom.

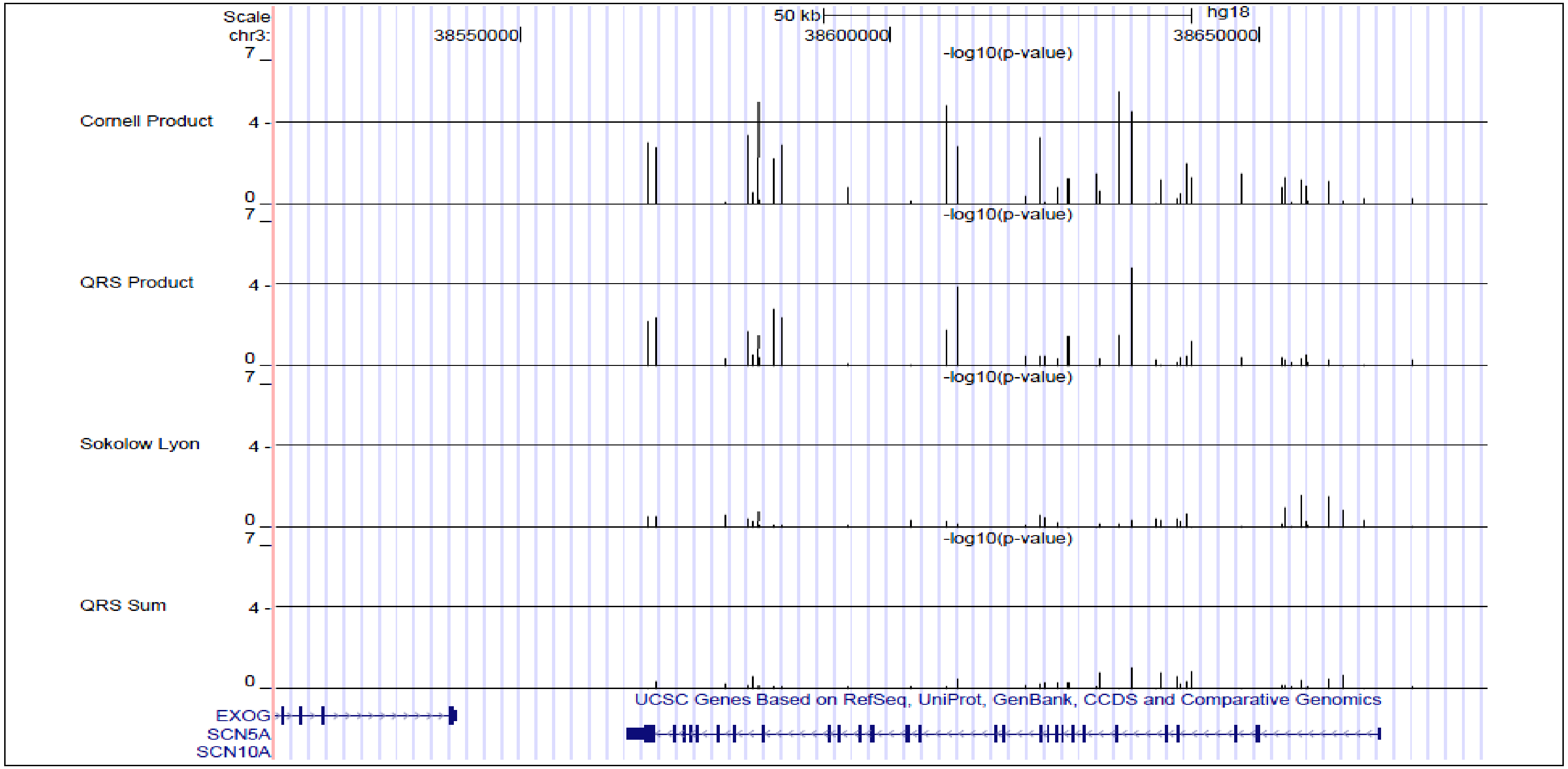


**Figure 2.6 Genetic association of rs2290893 (*PTGES3*) with QRS Voltage Sum.** Associations are reported by study separately for the discovery and replication cohorts, together with pooled discovery, replication and overall estimates. Beta-coefficients (with 95% confidence intervals) describe per allele effect of the minor allele of the SNP for the trait shown. A negative beta indicates that the trait had a lower value in those carrying the minor allele. The heterogeneity index ($I^2$) value is shown in the bottom left hand corner.

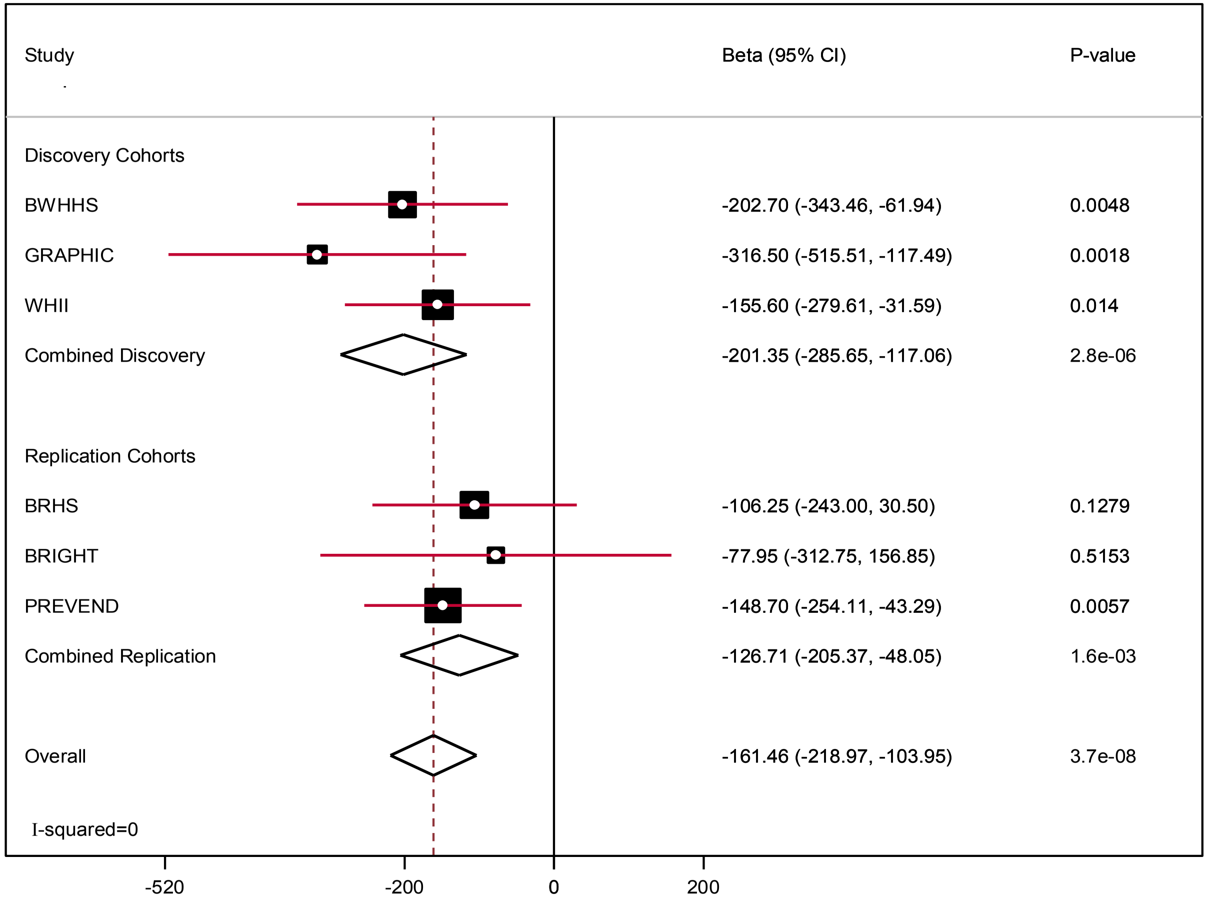


**Figure 2.7 Association signals around the *PTGES3* gene with ECG-LV mass indices.** A snapshot from the UCSC genome browser ([http://genome.ucsc.edu/](http://genome.ucsc.edu/)) showing bar plots of the –log$_{10}$(p-values) from the discovery meta-analysis of all SNPs on the array in and around the *PTGES3* gene with all four ECG-LV mass indices. The horizontal lines at –log$_{10}$(p-values) = 4 represent the discovery p-value threshold. Genes (exons as boxes and introns as horizontal lines) within the regions are displayed at the bottom.

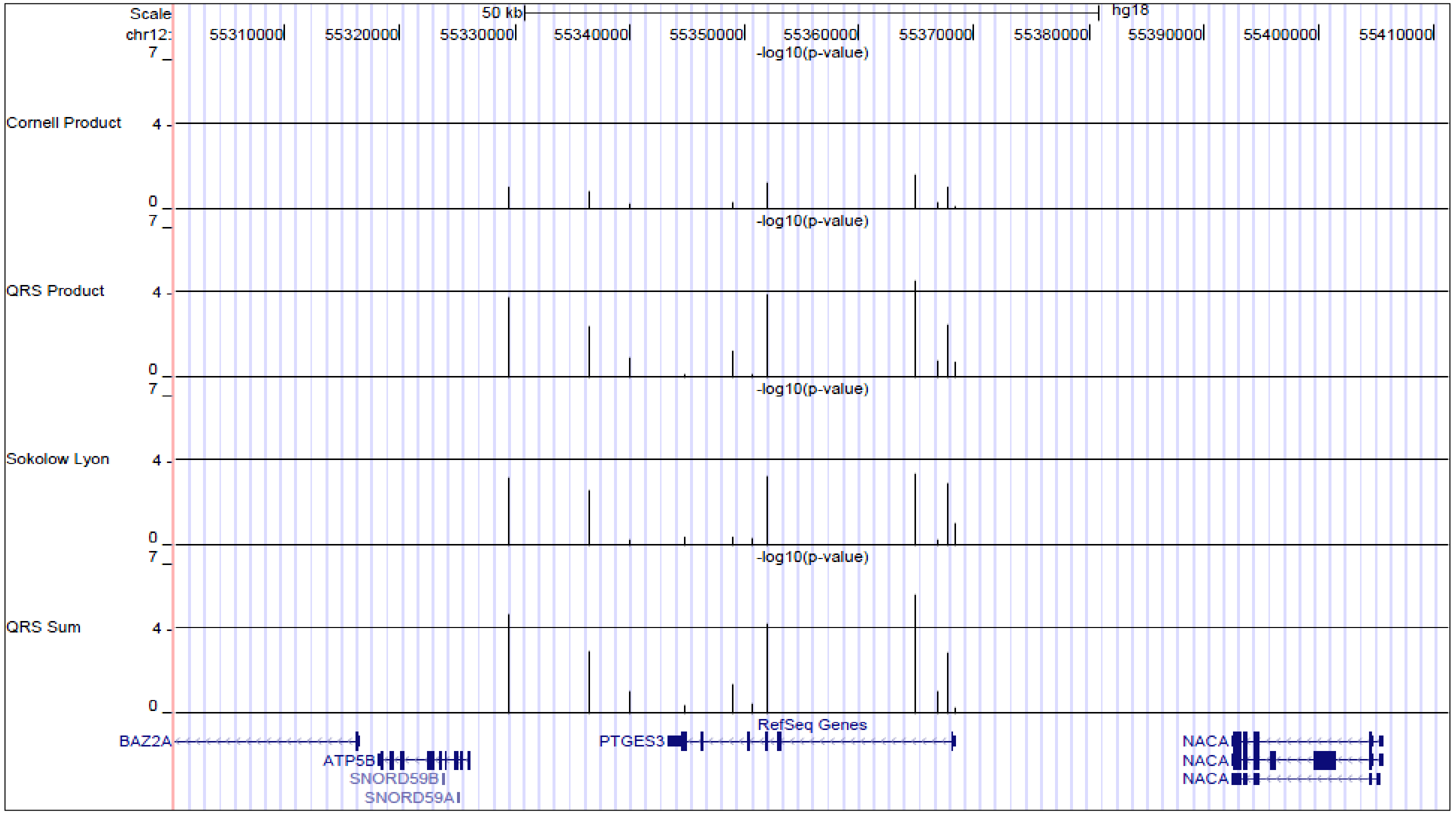
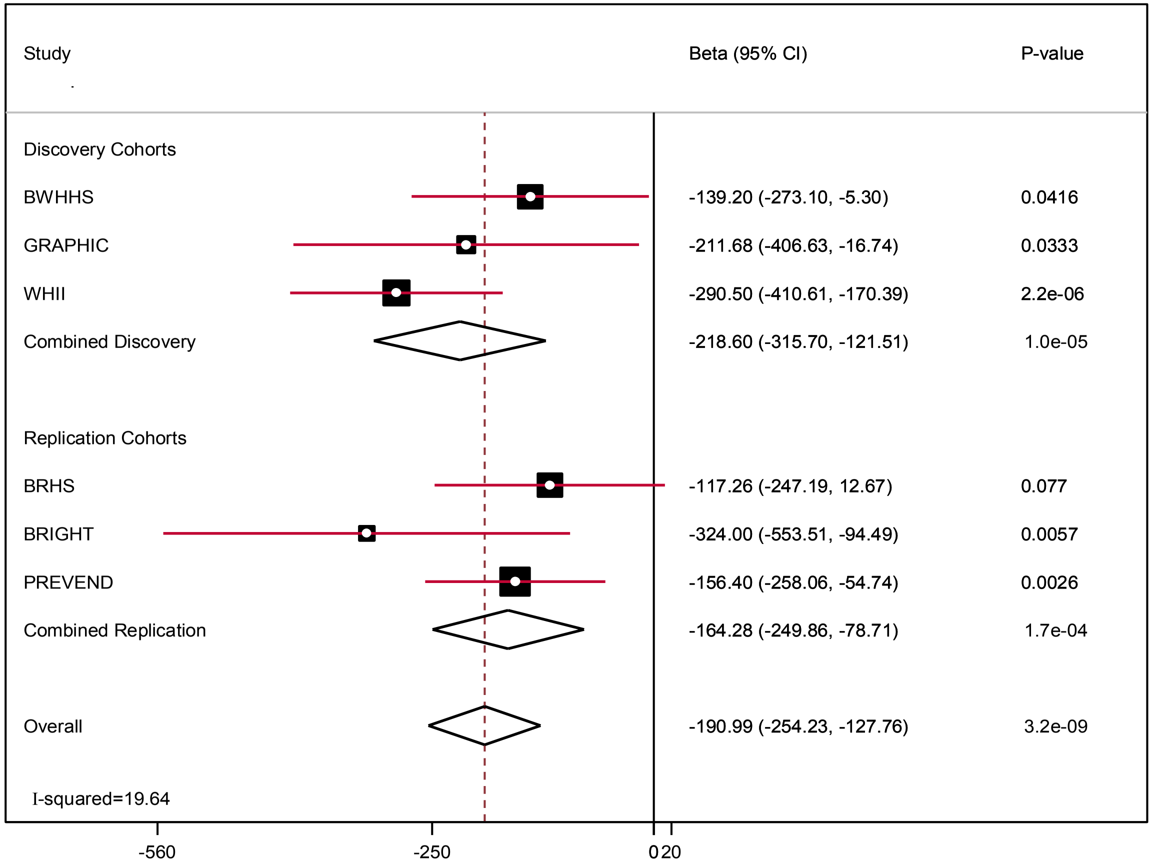


**Figure 2.8 Genetic association of rs2292462 (*NMB*) with QRS Voltage Sum.** Associations are reported by study separately for the discovery and replication cohorts, together with pooled discovery, replication and overall estimates. Beta-coefficients (with 95% confidence intervals) describe per allele effect of the minor allele of the SNP for the trait shown. A negative beta indicates that the trait had a lower value in those carrying the minor allele. The heterogeneity index ($I^2$) value is shown in the bottom left hand corner.



| Study | Beta (95% CI) | P-value |
|---|---|---|
| Discovery Cohorts | | |
| BWHHS | -139.20 (-273.10, -5.30) | 0.0416 |
| GRAPHIC | -211.68 (-406.63, -16.74) | 0.0333 |
| WHII | -290.50 (-410.61, -170.39) | 2.2e-06 |
| Combined Discovery | -218.60 (-315.70, -121.51) | 1.0e-05 |
| Replication Cohorts | | |
| BRHS | -117.26 (-247.19, 12.67) | 0.077 |
| BRIGHT | -324.00 (-553.51, -94.49) | 0.0057 |
| PREVEND | -156.40 (-258.06, -54.74) | 0.0026 |
| Combined Replication | -164.28 (-249.86, -78.71) | 1.7e-04 |
| Overall | -190.99 (-254.23, -127.76) | 3.2e-09 |

I-squared=19.64

.

**Figure 2.9 Association signals around the *NBM* gene with ECG-LV mass indices.** A snapshot from the UCSC genome browser (http://genome.ucsc.edu/) showing bar plots of the –log$_{10}$(p-values) from the discovery meta-analysis of all SNPs on the array in and around the *NMB* gene with all four ECG-LV mass indices. The horizontal lines at –log$_{10}$(p-values) = 4 represent the discovery p-value threshold. Genes (exons as boxes and introns as horizontal lines) within the regions are displayed at the bottom.

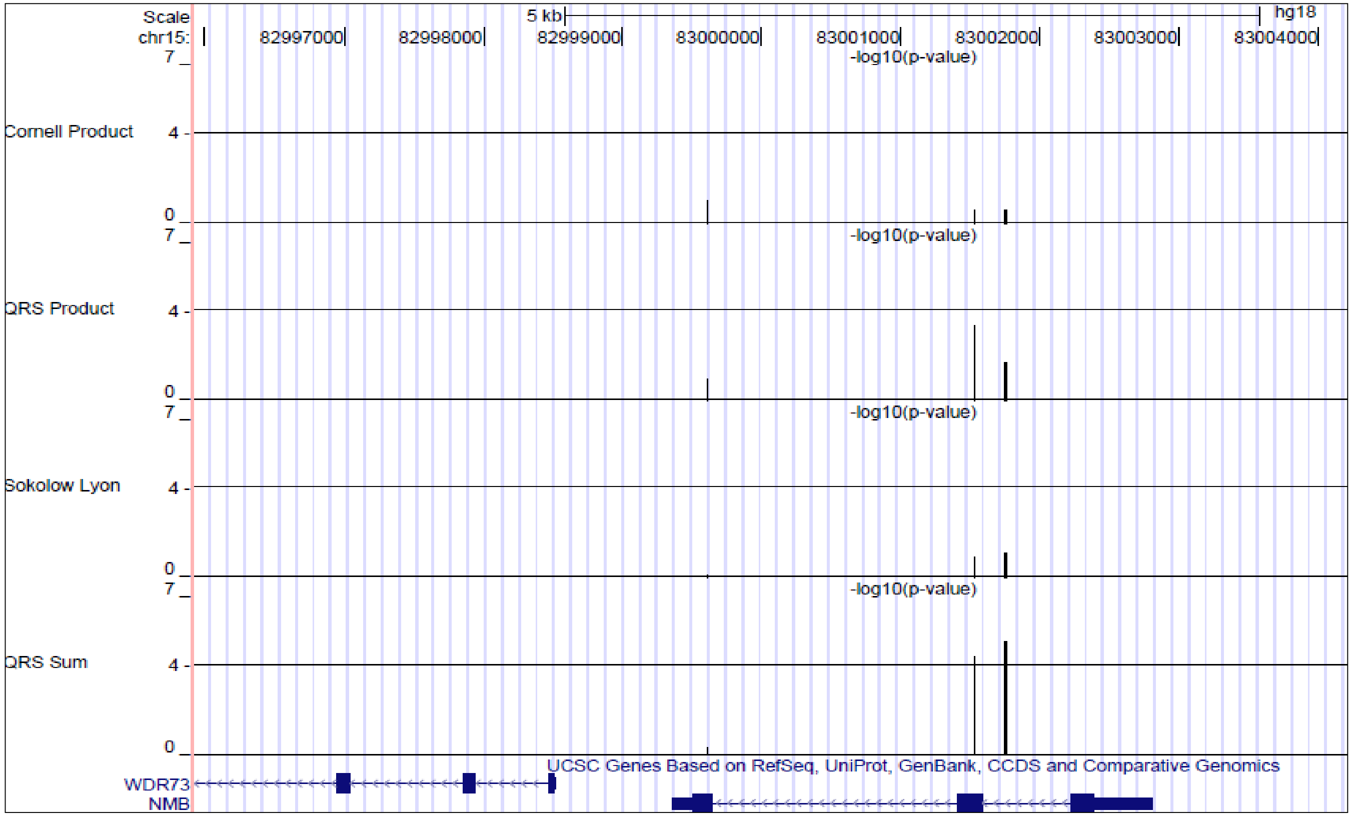


**Figure 2.10 Genetic association of rs4966014 (*IGF1R*) with QRS Voltage Sum.** Associations are reported by study separately for the discovery and replication cohorts, together with pooled discovery, replication and overall estimates. Beta-coefficients (with 95% confidence intervals) describe per allele effect of the minor allele of the SNP for the trait shown. A negative beta indicates that the trait had a lower value in those carrying the minor allele. The heterogeneity index ($I^2$) value is shown in the bottom left hand corner.

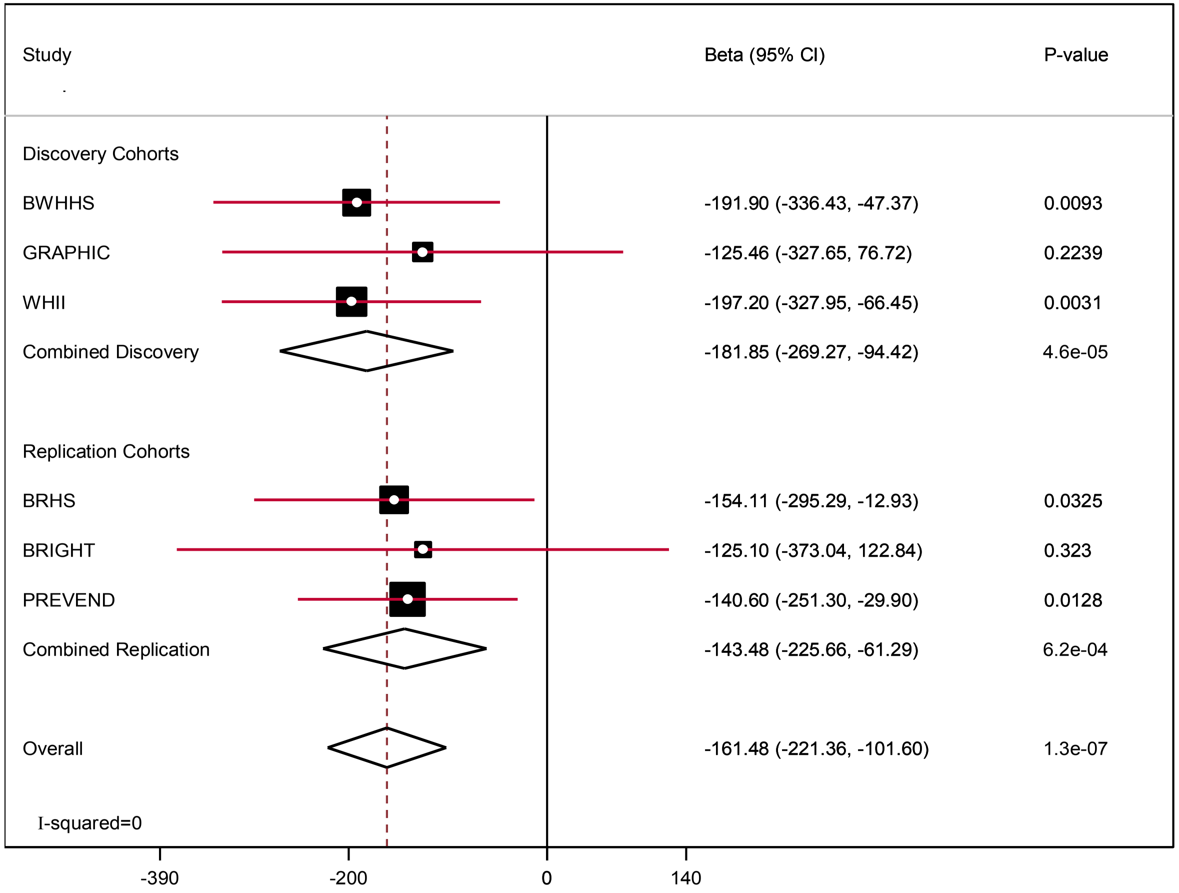


**Figure 2.11 Association signals around the *IGF1R* gene with ECG-LV mass indices.** A snapshot from the UCSC genome browser (http://genome.ucsc.edu/) showing bar plots of the –log$_{10}$(p-values) from the discovery meta-analysis of all SNPs on the array in and around the *IGF1R* gene with all four ECG-LV mass indices. The horizontal lines at –log$_{10}$(p-values) = 4 represent the discovery p-value threshold. Genes (exons as boxes and introns as horizontal lines) within the regions are displayed at the bottom.

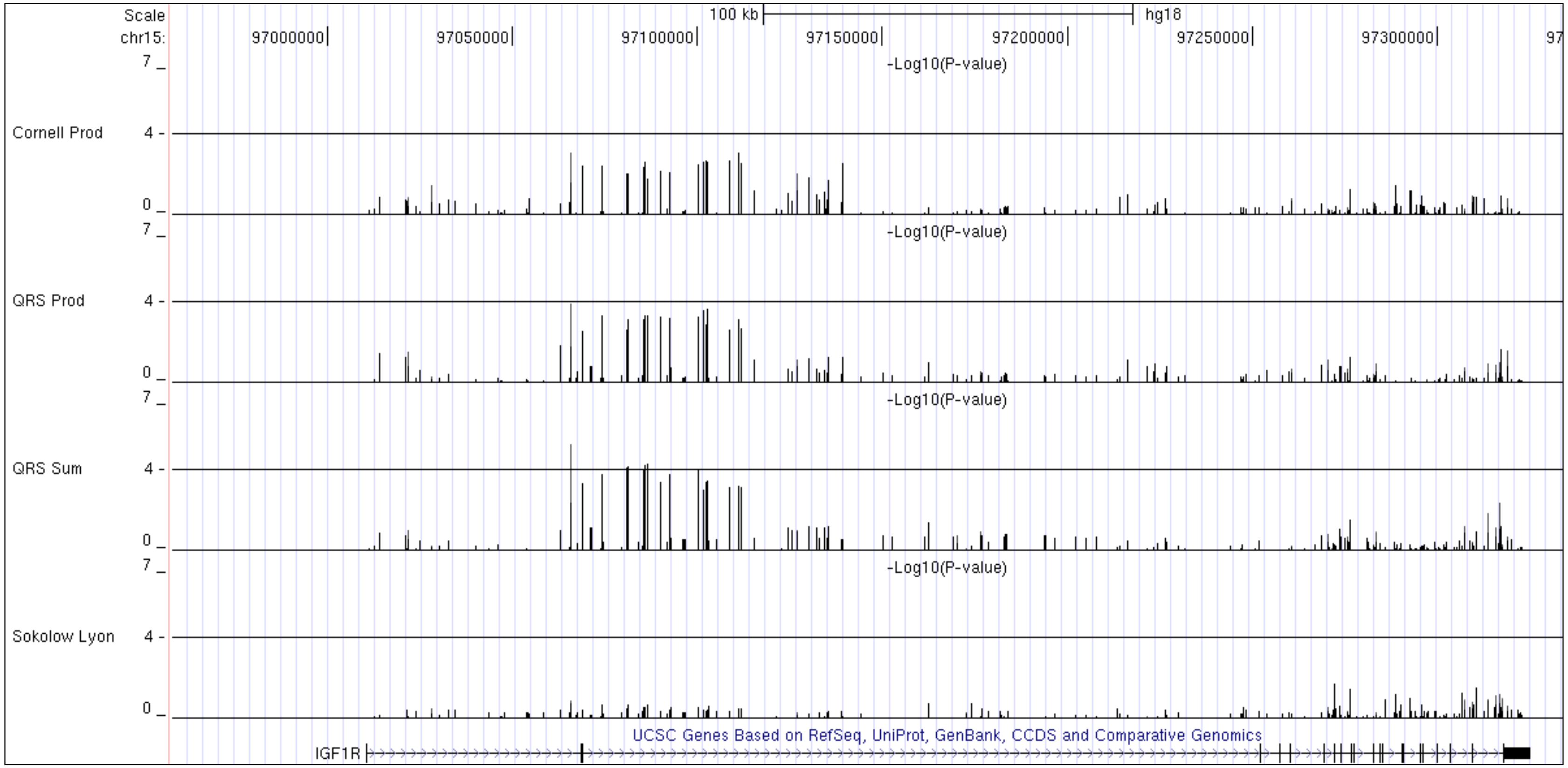
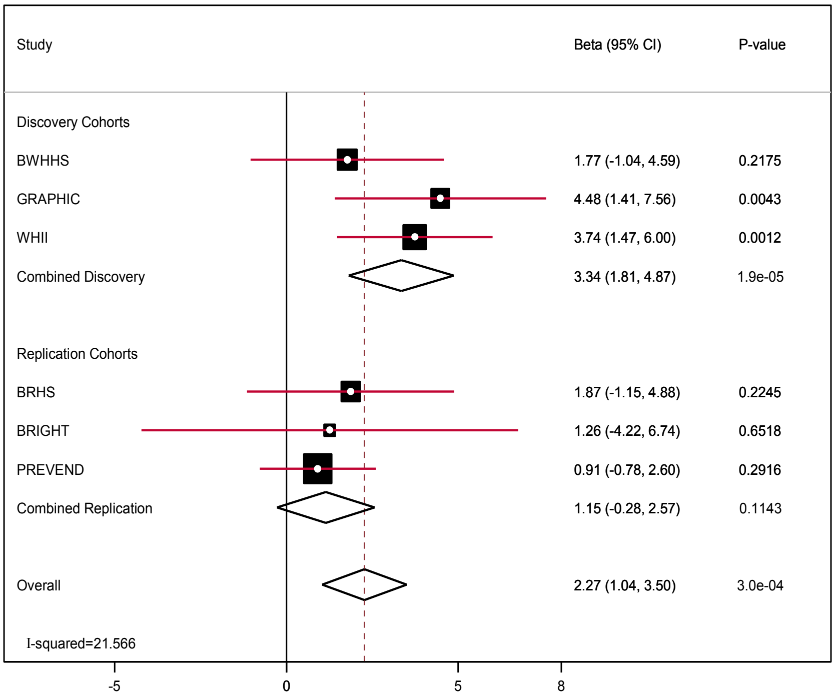
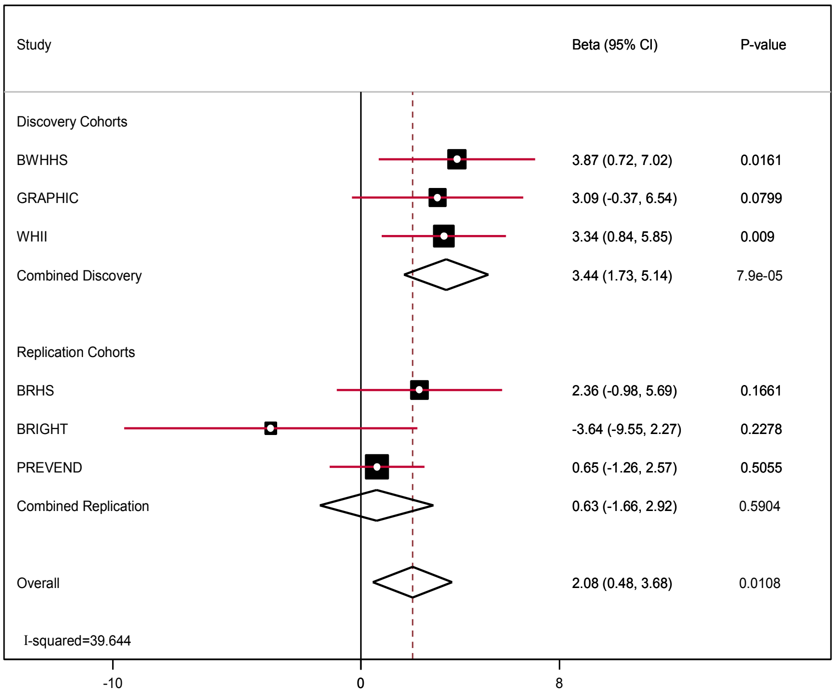


**Figure 2.12 A-H Forest plots of unreplicated SNPs.** Associations are reported by study separately for the discovery and replication cohorts, together with pooled discovery, replication and overall estimates. Beta-coefficients (with 95% confidence intervals) describe per allele effect of the minor allele of the SNP for the trait shown. The heterogeneity index ($I^2$) value is shown in the bottom left hand corner.

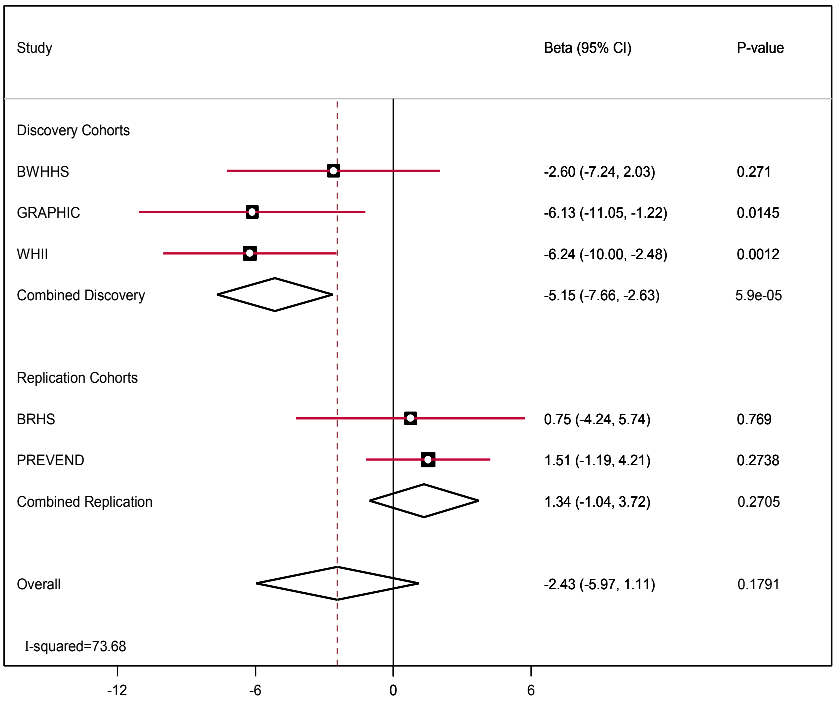
**A.** rs999310 (*MYOZ2*) with Cornell Product



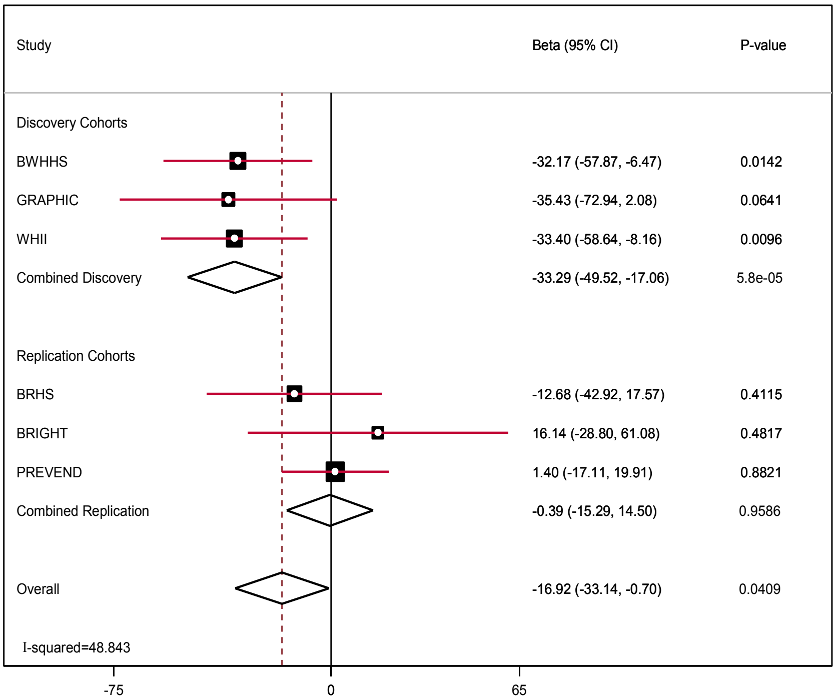
**B.** rs11603020 (*SERPING*) with Cornell Product



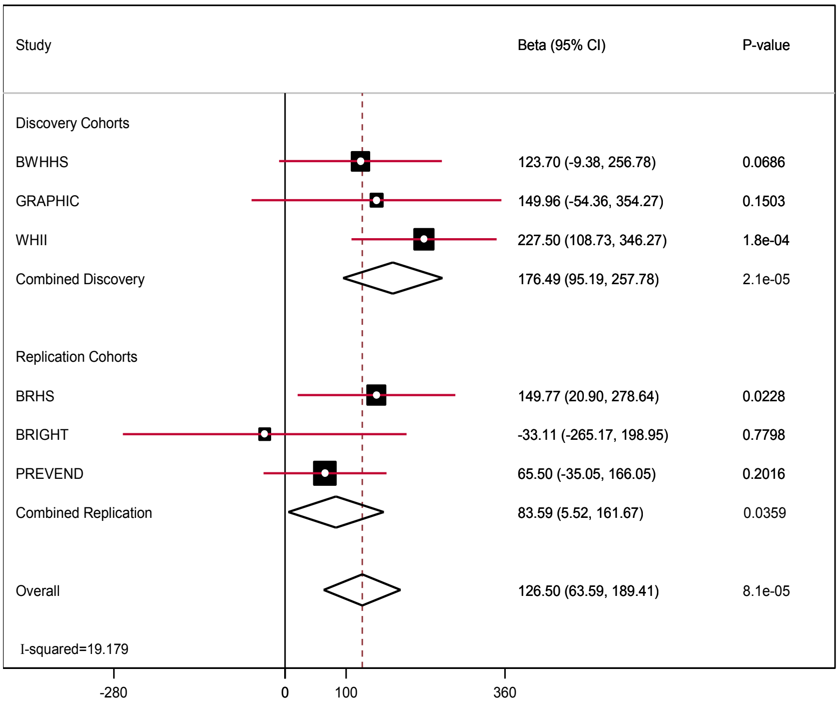
**C.** rs860867 (*TMEM117*) with Cornell Product



**D.** rs4713845 (*PPARD*) with QRS Voltage Product

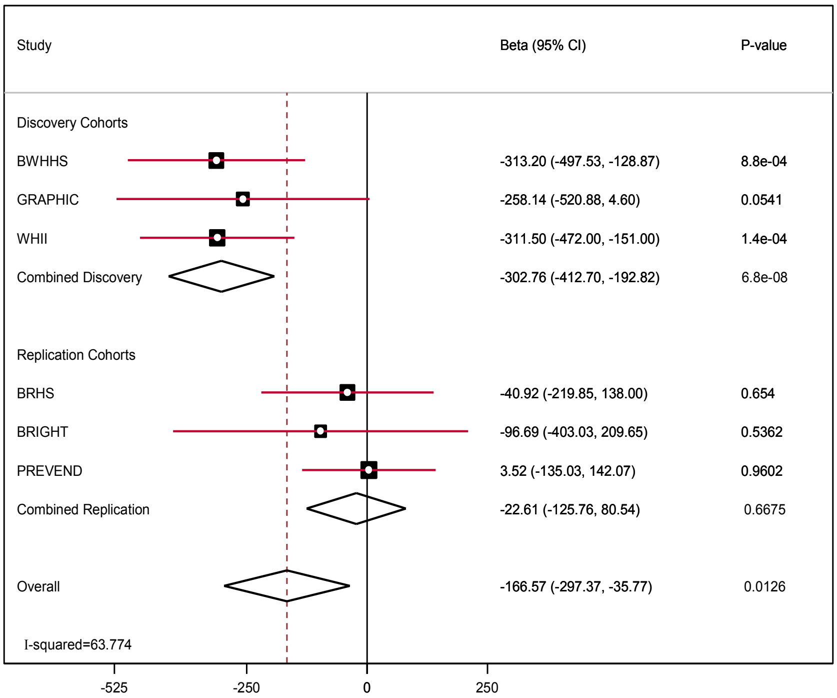


**E.** rs11724758 (*FABP2)* with QRS Voltage Sum



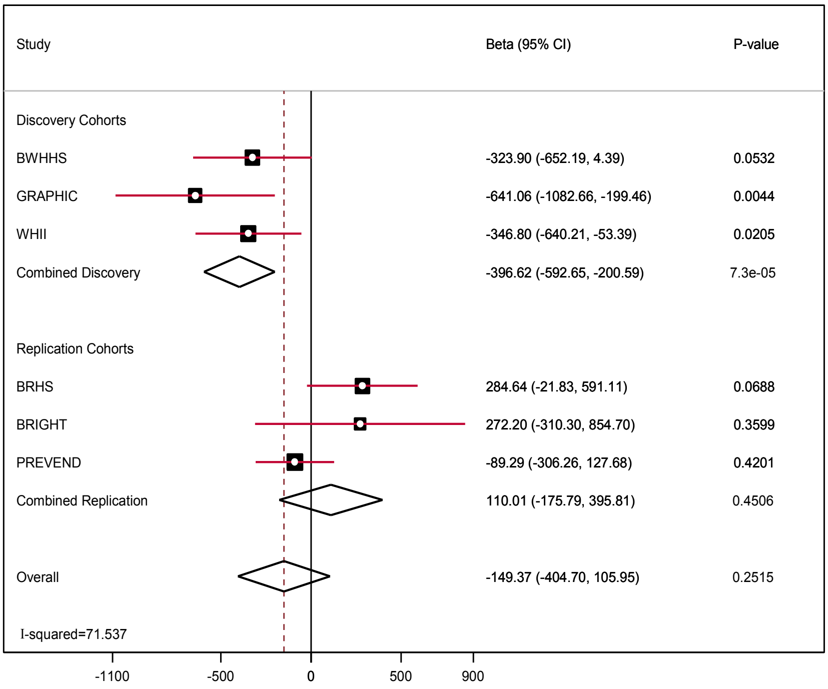
| Study | Beta (95% CI) | P-value |
|---|---|---|
| Discovery Cohorts | | |
| BWHHS | 123.70 (-9.38, 256.78) | 0.0686 |
| GRAPHIC | 149.96 (-54.36, 354.27) | 0.1503 |
| WHII | 227.50 (108.73, 346.27) | 1.8e-04 |
| Combined Discovery | 176.49 (95.19, 257.78) | 2.1e-05 |
| Replication Cohorts | | |
| BRHS | 149.77 (20.90, 278.64) | 0.0228 |
| BRIGHT | -33.11 (-265.17, 198.95) | 0.7798 |
| PREVEND | 65.50 (-35.05, 166.05) | 0.2016 |
| Combined Replication | 83.59 (5.52, 161.67) | 0.0359 |
| Overall | 126.50 (63.59, 189.41) | 8.1e-05 |

**F.** rs3815743 (*MTRR*) QRS Voltage Sum



| Study | Beta (95% CI) | P-value |
|---|---|---|
| Discovery Cohorts | | |
| BWHHS | -313.20 (-497.53, -128.87) | 8.8e-04 |
| GRAPHIC | -258.14 (-520.88, 4.60) | 0.0541 |
| WHII | -311.50 (-472.00, -151.00) | 1.4e-04 |
| Combined Discovery | -302.76 (-412.70, -192.82) | 6.8e-08 |
| Replication Cohorts | | |
| BRHS | -40.92 (-219.85, 138.00) | 0.654 |
| BRIGHT | -96.69 (-403.03, 209.65) | 0.5362 |
| PREVEND | 3.52 (-135.03, 142.07) | 0.9602 |
| Combined Replication | -22.61 (-125.76, 80.54) | 0.6675 |
| Overall | -166.57 (-297.37, -35.77) | 0.0126 |

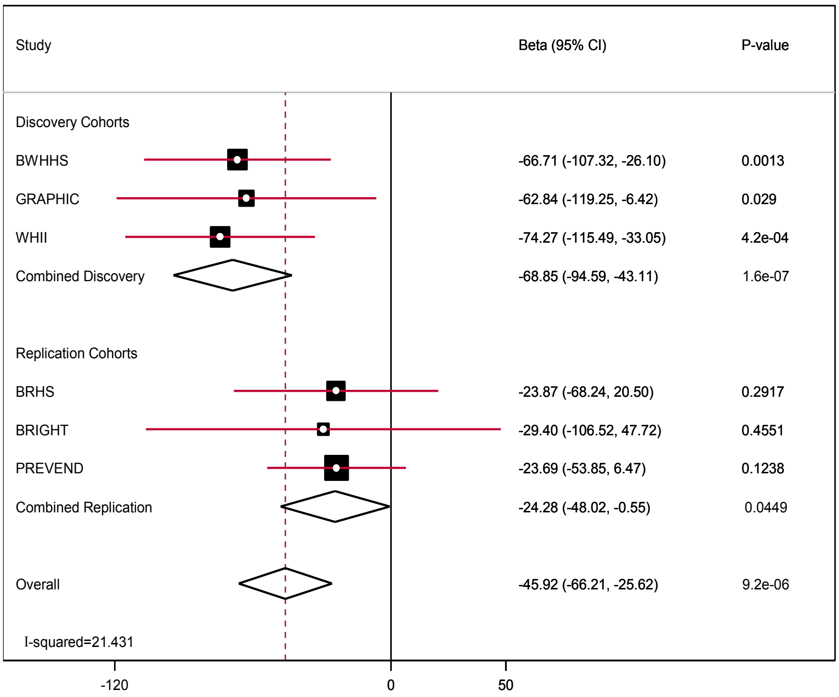
**G.** rs17597065 (*TCN1*) with QRS Voltage Sum



| Study | Beta (95% CI) | P-value |
|---|---|---|
| Discovery Cohorts | | |
| BWHHS | -323.90 (-652.19, 4.39) | 0.0532 |
| GRAPHIC | -641.06 (-1082.66, -199.46) | 0.0044 |
| WHII | -346.80 (-640.21, -53.39) | 0.0205 |
| Combined Discovery | -396.62 (-592.65, -200.59) | 7.3e-05 |
| Replication Cohorts | | |
| BRHS | 284.64 (-21.83, 591.11) | 0.0688 |
| BRIGHT | 272.20 (-310.30, 854.70) | 0.3599 |
| PREVEND | -89.29 (-306.26, 127.68) | 0.4201 |
| Combined Replication | 110.01 (-175.79, 395.81) | 0.4506 |
| Overall | -149.37 (-404.70, 105.95) | 0.2515 |

I-squared=71.537

**H.** rs10497520 (*TTN*) with Sokolow-Lyon



| Study | Beta (95% CI) | P-value |
|---|---|---|
| Discovery Cohorts | | |
| BWHHS | -66.71 (-107.32, -26.10) | 0.0013 |
| GRAPHIC | -62.84 (-119.25, -6.42) | 0.029 |
| WHII | -74.27 (-115.49, -33.05) | 4.2e-04 |
| Combined Discovery | -68.85 (-94.59, -43.11) | 1.6e-07 |
| Replication Cohorts | | |
| BRHS | -23.87 (-68.24, 20.50) | 0.2917 |
| BRIGHT | -29.40 (-106.52, 47.72) | 0.4551 |
| PREVEND | -23.69 (-53.85, 6.47) | 0.1238 |
| Combined Replication | -24.28 (-48.02, -0.55) | 0.0449 |
| Overall | -45.92 (-66.21, -25.62) | 9.2e-06 |

I-squared=21.431

### 2.3.6 Variance Explained

The percentage of trait variance explained by each of the four novel loci was less than 0. 5% in each study (Table 2.4).

**Table 2.4 Percentage variance explained.**

| | | | % Variance Explained | | |
|---|---|---|---|---|---|
| SNP | Gene Locus | Trait | BWHHS | GRAPHIC | WHII |
| rs6797133 | *SCN5A* | Cornell Product | 0.12 | 0.08 | 0.25 |
| rs2290893 | *PTGES3* | QRS Voltage Sum | 0.26 | 0.46 | 0.06 |
| rs2292462 | *NMB* | QRS Voltage Sum | 0.11 | 0.17 | 0.34 |
| rs4966014 | *IGF1R* | QRS Voltage Sum | 0.28 | 0.01 | 0.11 |

### 2.3.7 Risk Allele Count

To investigate the potential clinical relevance of these findings, the extent to which carriage of a trait-raising allele increased the chance of being in the top decile of the trait distribution was determined. Carriage of the rs6797133 (*SCN5A*) risk-allele increased the chances of being in the top decile of the Cornell Product distribution by 8%, while carriage of the rs2292462 (*NMB*) risk-allele increased the chances of being in the top decile of the QRS Voltage Sum distribution by 19% (Figure 2.13). To assess the combined effects of the three loci (*PTGES3*, *NMB* and *IGF1R*) affecting QRS Voltage Sum, the odds ratio of being in the top decile of the trait for those carrying 3 or more trait-raising alleles versus those carrying 0-2 alleles was calculated. Individuals carrying 6 risk alleles had a 1.60-fold (95% CI = 1.20 – 2.29) increased likelihood of being in the top decile of the QRS Voltage Sum distribution (Figure 2.14).

**Figure 2.13 Odds Ratios for being in the top decile for the associated ECG-LV mass trait per trait-raising allele for each of the four replicated SNPs.** Data from meta-analysis of all three discovery cohorts are shown and represented as odds ratio (95% CI) and association p-values. For rs2290893, rs2292462 and rs4966014 the odds ratios for being in the top decile of the QRS Voltage Sum distribution are shown, while for rs6797133 the odds ratios for being in the top decile of the Cornell Product distribution is shown,

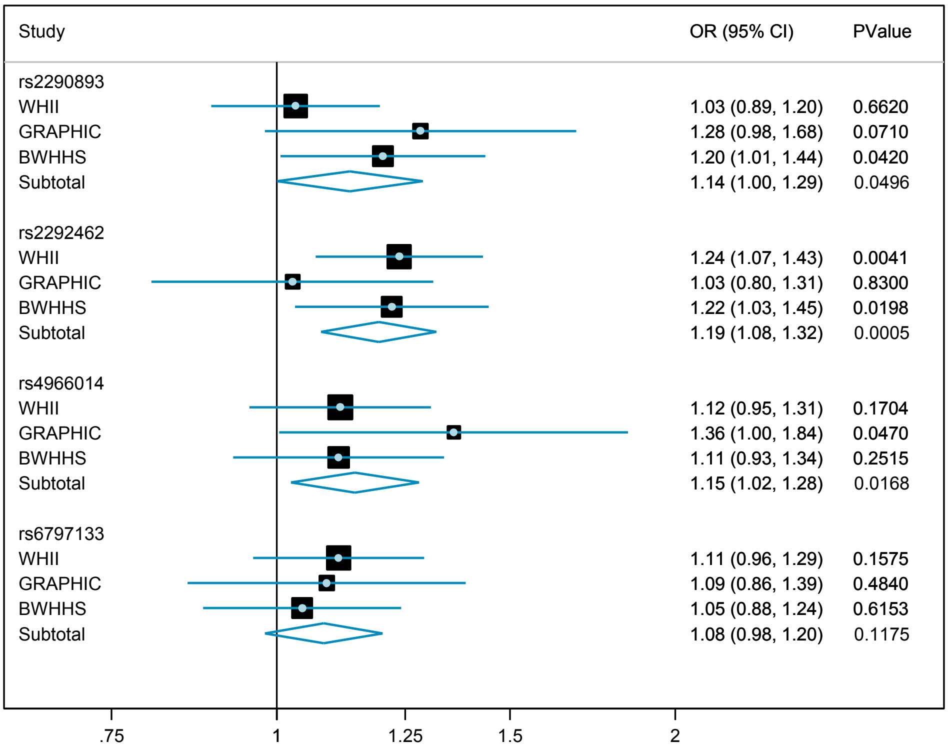
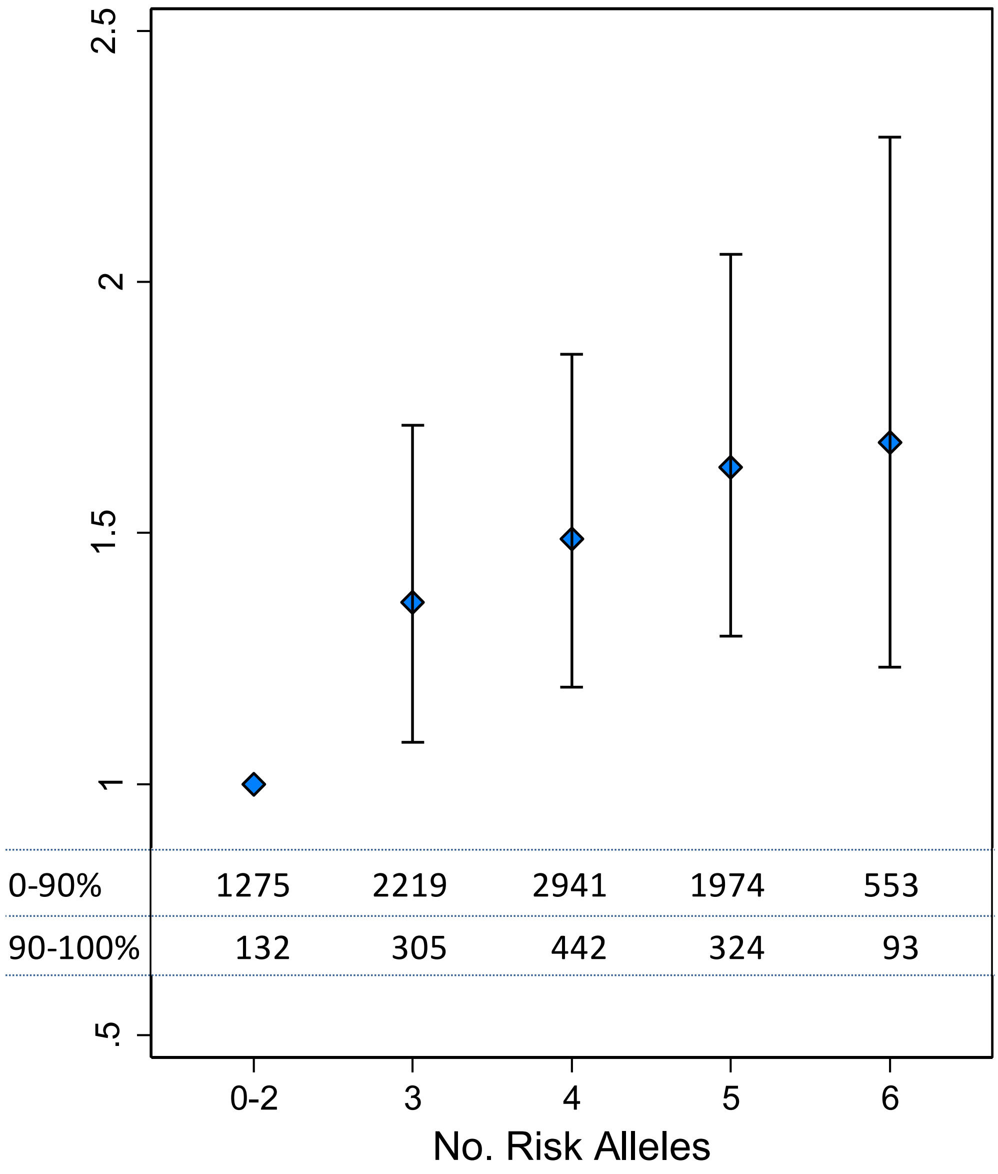


**Figure 2.14 Meta analysis of the odds ratio for being in the top decile for QRS Voltage Sum decile** The group carrying 0-2 alleles was used as the reference group. Data shows the odds ratio and 95% CI of each risk-allele score derived from the three SNPs associated with this trait (rs2290893, rs2292462, rs4966014). The sample size in each risk allele and decile group are shown at the bottom of the figure.



| | 0-2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 0-90% | 1275 | 2219 | 2941 | 1974 | 553 |
| 90-100% | 132 | 305 | 442 | 324 | 93 |

### 2.3.8 Functional Analysis

SNP rs2290893 is located in the first intron of the *PTGES3* gene. Based on analysis using the RegulomeDB server there was evidence that this SNP could have an effect on protein binding and expression: it was found within a region enriched for histone-modifications characteristic of both promoters and enhancers, as well as within a DNase hypersensitive region, characteristic of active promoters. The SNP also lies within an experimentally-determined binding site for the transcription factor Cdx-2. It was also significantly associated with higher *RBMS2* (a gene upstream of *PTGES3*) expression in adipose and skin tissue (p-value < 1x10$^{-06}$). At the *NMB* locus, the lead SNP rs2292462 was associated with expression of *NMB* in adipose tissue (p-value < 1x10$^{-06}$). The SNP was also in LD ($r^2$=0.48, D'=1.0) with a non-synonymous SNP (rs1051168, Proline to Threonine), 234 base pairs away that had an association signal (p-value = 4.3x10$^{-5}$) with QRS Voltage Sum in the discovery meta-analysis, comparable with that of the lead SNP. However, based on Polyphen prediction this SNP is unlikely to affect protein structure or function. There was no evidence to support any functional impact of the other two SNPs.

### 2.3.9 Variants Associated with Echo-LV Mass

There was no evidence for association of any of the replicated variants with echo-LV mass in the EchoGen data (Table 2.5).

**Table 2.5 Association of ECG-LV mass signals with echo-LV mass.**

| SNP | Gene | Chr | Position | Coded Allele | Non-Coded Allele | Meta Beta | Meta SE | Meta p-value | Mean MAF |
|---|---|---|---|---|---|---|---|---|---|
| rs2290893 | PTGES3 | 12 | 55364887 | A | G | -0.08 | 0.46 | 0.87 | 0.36 |
| rs2292462 | NMB | 15 | 83001758 | G | T | -0.33 | 0.45 | 0.47 | 0.46 |
| rs4966014 | IGF1R | 15 | 97065541 | C | T | 0.18 | 0.54 | 0.74 | 0.32 |
| rs6797133 | SCN5A | 3 | 38631037 | A | G | 0.17 | 0.45 | 0.71 | 0.39 |

### 2.3.10 Association with Candidate Genes

Renin-angiotensinogen system polymorphisms, previously suggested to be associated with LVH, were not associated with ECG-LV mass indices (Table 2.6). Upon examination of variants in other candidate genes and pathways linked to development of LVH, nominal associations with variants in several genes were found (Table 2.7). Although the level of significance achieved for these variants in the context of the large number of SNPs examined cannot exclude the possibility that many of these associations are false positives, their location within genes known to be involved in LVH suggests the need for further analysis in much larger sample sizes.

## 2.4 Discussion

### 2.4.1 Summary of Results

Large-scale association meta-analysis identified four genetic variants robustly associated with some of the ECG-derived indices of LV mass, providing novel insights into the genetic determinants of this widely assessed cardiovascular trait. The SNPs demonstrated association with Cornell Product at 3p22.2 in *SCN5A* and with QRS Voltage Sum at 12q13.3 in *PTGES3*, 15q25.2 in *NMB* and 15q26.3 in *IGF1R*. These variants do not appear to be associated with echo-LV mass suggesting that these phenotypes may measure somewhat distinct aspects of cardiac biology.

### 2.4.2 Insulin Growth Factor Pathway in Cardiac Biology

Cardiac hypertrophy is characterised by an increase in cardiomyocyte size, disarray of myofibrils, fibrosis in the extracellular matrix, re-activation of fetal transcriptional

**Table 2.6 Association analysis of candidate renin-angiotensinogen system polymorphisms with ECG-LVH traits.**

| SNP | Coded Allele | Non-coded Allele | Gene (Polymorphism) | Index of Left Ventricular Hypertrophy | Discovery Beta (SE) | Discovery P-value |
|---|---|---|---|---|---|---|
| rs5186 | C | A | *AGTR1* (A1166C) | Cornell Product | -1.64 (0.86) | 0.058 |
| | | | | QRS Voltage Sum | -13.6 (46.2) | 0.77 |
| | | | | QRS Voltage Product | -3.58 (5.73) | 0.53 |
| | | | | Sokolow-Lyon | -4.81 (9.56) | 0.61 |
| rs699 | C | G | *AGT* (M235T) | Cornell Product | 0.25 (0.79) | 0.75 |
| | | | | QRS Voltage Sum | 46.2 (43.5) | 0.29 |
| | | | | QRS Voltage Product | 13.0 (5.36) | 0.016 |
| | | | | Sokolow-Lyon | 7.89 (8.99) | 0.38 |
| rs1799998 | C | T | *CYP11B2* (-344 C/T) | Cornell Product | -0.0052 (0.81) | 0.99 |
| | | | | QRS Voltage Sum | 48.1 (82.6) | 0.56 |
| | | | | QRS Voltage Product | 4.16 (10.8) | 0.70 |
| | | | | Sokolow-Lyon | -5.92 (12.4) | 0.63 |
| rs4343 | A | G | *ACE* (G2350A, I/D) | Cornell Product | -0.27 (0.80) | 0.74 |
| | | | | QRS Voltage Sum | 34.7 (43.1) | 0.42 |
| | | | | QRS Voltage Product | 1.83 (5.30) | 0.36 |
| | | | | Sokolow-Lyon | 6.68 (8.86) | 0.45 |

**Table 2.7 Association of variants in LVH candidate genes with ECG-LV mass traits.** The table shows the discovery meta-analysis summary statistics for SNPs in candidate genes. Only the most significant SNP association with any one of the ECG-LV mass indices are shown for each gene.

| Biological Process | SNP | CHR | BP | Coded allele | Non-coded allele | Gene | Index of Left Ventricular Hypertrophy | Discovery Beta (95% CI) | Discovery P-value |
|---|---|---|---|---|---|---|---|---|---|
| Calcium Homeostasis | rs2746073 | 1 | 191045850 | A | T | *RGS2* | Cornell Product | -2.03 ( -3.78 , -0.28 ) | 0.02 |
| | rs7554607 | 1 | 235333226 | G | A | *RYR2* | QRS Voltage Sum | 137.3 ( 57.03 , 217.5 ) | 0.0008 |
| | rs3752581 | 6 | 118976423 | G | A | *PLN* | Sokolow-Lyon | -15.19 ( -32.3 , 1.92 ) | 0.08 |
| | rs10849860 | 12 | 120152637 | G | A | *P2RX4* | Cornell Product | -1.93 ( -3.97 , 0.11 ) | 0.06 |
| Hemodynamic load | rs2493129 | 1 | 228907565 | A | G | *AGT* | Cornell Product | -5.62 ( -10.37 , -0.86 ) | 0.02 |
| | rs12721272 | 3 | 149929654 | A | G | *AGTR1* | QRS Voltage Sum | 293.5 ( 31.68 , 555.3 ) | 0.03 |
| | rs4714384 | 6 | 12405839 | G | A | *EDN1* | QRS Voltage Product | -11.93 ( -22.6 , -1.25 ) | 0.03 |
| | rs2853796 | 7 | 150334848 | C | A | *NOS3* | QRS Voltage Product | -13.97 ( -24.05 , -3.9 ) | 0.007 |
| | rs4917675 | 10 | 115789467 | G | A | *ADRB1* | QRS Voltage Product | 12.13 ( 0.65 , 23.61 ) | 0.04 |
| | rs4354 | 17 | 58925184 | A | G | *ACE* | QRS Voltage Product | -39.16 ( -72.32 , -6.01 ) | 0.02 |
| | rs4646124 | 23 | 15526717 | A | G | *ACE2* | Cornell Product | -2.27 ( -4.25 , -0.29 ) | 0.02 |
| Structural | rs868407 | 1 | 199607964 | G | A | *TNNT2* | Sokolow-Lyon | 17.25 ( -1.48 , 35.99 ) | 0.07 |
| | rs936175 | 3 | 46879712 | A | C | *MYL3* | QRS Voltage Product | 11.54 ( -3.89 , 26.98 ) | 0.1 |
| | rs10865971 | 3 | 52456446 | G | A | *TNNC1* | QRS Voltage Product | 13.91 ( -0.66 , 28.47 ) | 0.06 |
| | rs3729989 | 11 | 47326617 | G | A | *MYBPC3* | Cornell Product | 3.6 ( 1.32 , 5.88 ) | 0.002 |
| | rs3729823 | 14 | 22956249 | G | C | *MYH7* | QRS Voltage Product | -7.88 ( -69.15 , 53.38 ) | 0.8 |
| | rs893132 | 15 | 32877352 | G | A | *ACTC1* | QRS Voltage Product | -16.11 ( -26.19 , -6.03 ) | 0.002 |
| | rs17752921 | 15 | 61130684 | G | A | *TPM1* | Cornell Product | -2.99 ( -5.51 , -0.47 ) | 0.02 |
| | rs3729709 | 19 | 60359618 | G | A | *TNNI3* | Cornell Product | -4.27 ( -7.16 , -1.38 ) | 0.004 |

programs, and decreased cardiac function (Sun et al. 2009). Cardiac myocytes undergo rapid proliferation during fetal life, but in the perinatal period proliferation ceases. Adult cardiac myocytes generally do not re-enter the cell cycle when exposed to growth signals, and further increases in cardiac mass are partly achieved through an increase in cell size (hypertrophy) (Ahuja et al. 2007). The association of variants in the insulin growth factor 1 receptor (*IGF1R*) gene with ECG-LV mass is plausible given the current knowledge of the role of the IGF pathway in the heart. IGF1 binds to IGF1R triggering a signalling cascade that plays an essential regulatory role in cardiac biology. During the post-natal period, the switch in cardiac metabolism and the cardiomyocyte's withdrawal from the cell cycle is characterized by a marked decrease in the expression of IGF1 and IGF1R (Knezevic et al. 2012). There have been several mouse studies that have demonstrated the importance of the IGF-1 signalling pathway in cardiac remodelling. Transgenic mice over-expressing IGF1R in the heart displayed cardiac hypertrophy, which was the result of an increase in myocyte size (McMullen et al. 2004). A recent study in mice also demonstrated that postnatal repression of cardiac IGF1R is associated with the up-regulation of a micro-RNA (miR-378) promoting cardiomyocyte apoptosis (Knezevic et al. 2012). IGF-1-injected hearts of infarcted mice showed improved ventricular function and cardiomyocyte survival (Urbanek et al. 2005; Welch et al. 2002). Interestingly, SNPs in the *IGF1* gene did not show association with ECG-LV mass traits in this analysis.

### 2.4.3 *SCN5A*

Mutations in *SCN5A*, which encodes the sodium channel, voltage-gated, type V, alpha sub-unit, cause long QT syndrome, a Mendelian arrhythmogenic disease characterised by a prolonged QT interval (represents the ventricular contraction on the ECG (refer to Figure 1.1). SNPs in this gene have been associated with two other ECG parameters, PR Interval and QRS duration (Chambers et al. 2010; Holm et al.

2010). Increased LV mass is known to increase the height and depth of the QRS complex and the length of the QRS duration. Association of *SCN5A* variants were only observed with the two ECG-LV mass indices incorporating QRS duration — Cornell Product and QRS Voltage Product. Given the high correlation of the QRS Voltage Product with the QRS Voltage Sum and the lack of association with the latter index which incorporates the amplitude of the Q, R and S waves but not the duration, it is unclear whether the observed association of *SCN5A* variants with ECG-LV mass reflects changes in the function of the sodium channel that simply affect the propagation of the electrical signal, or whether variants in *SCN5A* actually affect myocyte size.

### 2.4.4 *PTGES3* and *NMB*

Neither *PTGES3* nor *NMB* are *a priori* biological candidates likely to influence ECG-LV mass. *PTGES3* codes for prostaglandin E synthase 3 (also known as p23 or TERT binding protein). *NMB* codes for neuromedin B, the mammalian homologue of bombesin-like peptide. Fine-mapping of these regions may help identify the causal gene in these regions.

### 2.4.5 Sarcomeric Protein-Coding Genes

Though SNPs in *MYOZ2* (Myozenin 2) and *TTN* (Titin), both playing an important role in sarcomere structure and function, did not pass the replication threshold, they showed suggestive evidence for association in the replication studies. Given that both genes are associated with hypertrophic cardiomyopathy (Osio et al. 2007; Satoh et al. 1999), they may be plausible candidates for LV mass variation.

### 2.4.6 Comparison of Echo-LV Mass and ECG-LV Mass

Interestingly, the loci associated with ECG indices of LV mass did not show evidence of association with echo-derived LV mass in the large meta-analysis by the EchoGen Consortium (Vasan et al. 2009). However, several observations indicate that the two may reflect different biological processes. The electrocardiogram measures the algebraic sum of the action potentials of myocardial fibres. Therefore, the ECG changes in cardiac hypertrophy reflect the electrical remodelling of the action potential of the cardiac myofibres, which is measured in voltage and time (Hill 2003). By contrast, the echocardiogram captures anatomical remodelling of the myofibres, fibroblasts, other interstitial changes (such as inflammation) and the cardiac chambers of the heart. This is reflected in the poor correlation between ECG- and echo-LV mass measures in several clinical contexts (Casale et al. 1987; Epstein et al. 1990; Rosenzweig et al. 1991). More direct evidence that they may be genetically different comes from assessment of ECG-LVH indices and echo-LV mass in the same families showing greater heritability for the ECG indices (Mayosi 2002). This observation underscores the importance and relevance of identifying genetic determinants of both traits.

### 2.4.7 Combined Effect of Four Loci

Consistent with variants identified for other complex quantitative traits, the amount of trait variance explained by each SNP individually was low (<1%). Carriage of the trait-raising allele at each of the locus was associated with an 8 - 19% higher probability of lying in the top 10% of the population distribution for that trait. The effect of the three loci affecting QRS Voltage Sum was additive. Individuals carrying all six trait raising alleles for these loci (~ 6% of the population) had a 1.60 (95% CI = 1.23 - 2.29) fold increased probability of lying in the top decile for QRS Voltage Sum compared with those carrying 0-2 alleles. Whether these differences impact on the cardiovascular risk associated with ECG-LVH will require further evaluation in large-scale population samples.

### 2.4.8 Limitations

The Cardiochip array contains only about 10% of all genes in the human genome with a known or suspected cardiovascular function (Keating et al. 2008). While providing a cost-effective analysis of variants in these genes, a significant limitation is that it does not provide full genome coverage and there are likely to be many missed variants with similar or indeed greater effects than those identified that lie outside of known cardiovascular pathways. Even within the genes studied additional variants may not have been identified, given the threshold used for taking variants forward for replication. Larger and more comprehensive studies will be required to identify more loci associated with ECG-LV mass traits and explain a larger proportion of their variances. Analysis in the extremes of the LV mass distribution may be an additional strategy for genetic discovery if larger studies are available.

As with any study of this type, the association findings provide a first step towards identifying and studying the functionality of candidate genes in the associated region. For some of the identified regions, where no functional link to LV mass could be identified, the causal gene is unclear. In such cases fine-mapping via imputation of additional SNPs or sequencing in these regions may help refine the location of the causal gene and variant.

Though there is evidence for some of these SNPs being associated with gene expression of nearby genes, to establish a more confirmatory role in relation to the trait of interest, eQTL analysis of heart tissue would need to be carried out. Though the data is not yet available for all tissues, the Genotype-Tissue Expression project (https://commonfund.nih.gov/GTEx/), an initiative to understand how genetic variation may control gene activity and its relationship to disease, is currently underway, which will eventually provide eQTL data in heart tissue. SNPs that fall within regulatory regions may alter transcription factor binding, histone methylation signatures and chromatin accessibility. Experimental techniques can be used to measure allele-specific differences in protein binding (electrophoretic mobility shift assay), gene expression (luciferase reporter assay) and open chromatin (FAIRE), to confirm the functional role of SNPs. However, this information would still need to be related to changes in phenotype.

Though these findings provide new insights into the genetic influences on a routinely recorded clinically-relevant cardiovascular trait, the biologic meaning of the findings requires consideration of the specific traits, variants and genes at the loci, and annotations of their potential functions before their possible clinical relevance will be understood.

# 3 Lipid-Associated Genetic Variants for Risk Prediction

## 3.1 Introduction

Ideally, a predictive model would be able to categorise people dichotomously into those who would develop CHD and those who would not (Dent 2009). Those predicted to have a CHD event could then be targeted to receive therapeutic or lifestyle interventions that would reduce their risk and prevent or postpone the occurrence of the disease. Unfortunately, all available risk models are far from perfect and there is ongoing research into whether additional risk factors can improve current risk prediction models.

Blood lipid levels have been known CVD risk factors for over half a century, and therapeutic intervention for primary prevention of CVD in the general population was initially informed by lipid level thresholds. However, lipid levels identify patients at risk of future coronary events only moderately well. NICE and other organisations recommend commencing LDL-C-lowering statin therapy for primary prevention in individuals estimated to have a 10 year absolute risk of CVD >20%, where the recommended methods for evaluating absolute CVD risk are based on multiple, established risk factors. The most widely-used model is the Framingham 10 year CVD risk score (Anderson et al. 1991), based on analysis carried out in the Framingham Heart Study (see section 1.6.1). Despite these guidelines, doctors may still be persuaded in their therapeutic decisions by high absolute values of total cholesterol or LDL-C.

All the principal blood lipid fractions: total cholesterol, LDL-C, HDL-C, and triglycerides, have both environmental and genetic determinants, with a reported heritability of 40 - 70% (Krauss 2008). Recently, association studies using whole genome and dense gene-centric arrays have identified numerous common SNPs associated with these four lipid fractions. Each SNP has a small average effect but any individual may carry numerous variants which, collectively, have a more substantial influence on blood lipid levels (Talmud et al. 2009). Genetic variants have some potential advantages over non-genetic risk factors as predictors of disease risk. Genotype is fixed from conception and so should represent long-term differences in blood lipid values without the biological variation that affects assays of blood lipids themselves. Genotyping assays have very high fidelity and low cost. This has led to an interest in the potential use of genetic information for evaluation of cardiovascular risk.  Despite lipids being important CHD risk factors, there is little information on the population effect of multiple lipid-associated SNPs on clinically relevant healthcare outcomes such as estimates of cardiovascular risk, prescription of lipid-lowering drug therapies, and subsequent clinical events.

The aim of the work in this chapter was to evaluate the influence of common SNPs associated with total cholesterol, LDL-C, HDL-C and triglyceride levels on the following outcomes: (1) being identified as a 'high-risk' individual as determined by a Framingham 10 year CVD risk greater than 20%, which is the qualifying threshold used to identify such individuals in Britain (and many other countries), and is a reference against which many other methods of risk prediction are routinely assessed; (2) receiving lipid lowering treatment, since guidelines encourage primary therapeutic intervention for these high-risk individuals; and (3) coronary disease events. Analysis was carried out in two British cohorts (WHII and BWHHS), in which prescribing decisions were made without knowledge of participants' genotype. For comparison, the association of the Framingham 10 year CVD risk score (Anderson et al. 1991), which is based on phenotypic rather than genetic measurements, with the odds of receiving lipid medication and CHD outcome was also assessed.

## 3.2 Materials and Methods

### 3.2.1 Study Populations

Data from the WHII and BWHHS cohorts were used for this analysis, both of which were described previously in section 2.2.1. Baseline lipid and CVD risk factor measurements (1991-1993 in WHII, and 1999-2001 in BWHHS) were used, while information on CHD events and lipid medication use was obtained from the 2003-2004 follow-up phase (10 years from baseline) in WHII and the 2007 follow-up phase (8 years from baseline) in BWHHS.

### 3.2.2 Lipid Measurements

In WHII venous blood samples at each examination were taken after at least 5 hours of fasting. Serum obtained after centrifugation was refrigerated at -4°C and assayed within 72 hours of the blood draw. Total cholesterol and triglycerides were measured using a centrifugal analyser. HDL-C levels were determined by measuring cholesterol in the supernatant fluid obtained after precipitating non-HDL-C with dextran sulfate-magnesium chloride with the use of a centrifuge (Kivimäki et al. 2008). In BWHHS, total cholesterol, HDL-C, and triglycerides were measured on frozen serum sample using the same methods used in WHII. In both studies, LDL-C concentration was estimated based on the Friedewald formula (Friedewald et al. 1972). Because isolation of the LDL fraction requires ultracentrifugation, a technique not generally available in service laboratories, the concentration of LDL-C was calculated by this formula. Individuals with triglyceride levels >4.5 mmol/L did not have their LDL-C levels calculated and were set as missing, since calculated LDL-C cannot be accurately estimated when triglyceride levels exceed this threshold (Friedewald et al. 1972).

### 3.2.3 Coronary Events Data Collection

In WHII, questionnaires were sent at each phase of data collection to gather information on self-reported non-fatal coronary events (MI or angina), and this was supplemented by information on coronary events identified by research clinic ECGs, and through verification of primary care and hospital records (Marmot & Brunner 2005). At baseline in BWHHS, women were recorded to have CHD if they had a medical record of an MI or angina, or if they self-report that a doctor had ever diagnosed a heart attack or angina. At subsequent phases in BWHHS, incident (new) cases of CHD were collected through detailed medical record reviews and participant questionnaires and by linkage to the National Health Service (NHS) Central Register for information on date and cause(s) of all deaths during follow-up.

### 3.2.4 Lipid Medication Use Data Collection

In WHII, participants were asked to name any medication taken in the 14 days prior to the survey at each phase. The medication list was recoded using the British National Formulary (BNF) codes (http://bnf.org/bnf/index.htm) and participants were categorised as users of lipid lowering drug therapy if they used statins or other lipid lowering drugs such as fibrates, nicotinic acid and its derivatives, cholesterol absorption inhibitors, or omega-3 fatty acid compounds. Baseline lipid-lowering drug use in BWHHS was determined by face to face interview. Participants were asked to bring to the assessment their repeat medication slips or their actual medications. Data on all medications, including their dosage, were entered onto a questionnaire sheet by the interviewer. For women who forgot their repeat prescription document they were asked about any medications, including dosage and the reason for which they were prescribed the medication. Medications were catalogued using codes from the BNF. For subsequent phases in BWHHS, information on medication was obtained from self-administered postal questionnaire, where participants were encouraged to write medication details direct from their repeat prescription sheet and/or mail a copy of the prescription sheet back with the questionnaire.

### 3.2.5 Genotyping

As described in section 2.2.4 both WHII and BWHHS individuals were genotyped using the Cardiochip (Keating et al. 2008), which contains ~50,000 SNPs covering ~2000 loci that are known to be involved in cardiovascular pathways. After quality control filters (previously described in section 2.2.4), 5059 WHII and 3414 BWHHS individuals with genotype data remained for analysis.

### 3.2.6 Selection of Lipid-Associated SNPs

A previously published large-scale genetic association analysis of SNPs on the Cardiochip genotyping platform with baseline measurements of blood lipids in the 5059 individuals from WHII (Talmud et al. 2009) reported 60 SNPs in 12 genes to be associated with LDL-C, 73 SNPs in 10 genes associated with triglycerides, and 71 SNPs in 5 genes associated with HDL-C (Talmud et al. 2009), all passing the significance threshold of p-value $<1x10^{-05}$. Associations with total cholesterol (53 SNPs) were also identified, though not published.

Since the genotyping platform used is a gene-centric array and was not designed using a tag-SNP approach, typically many SNPs in a region harbouring one or more causal variant(s) demonstrate univariate associations with the traits of interest, but the majority of these associations are indirect and operate through LD with the causal site(s). In the study by Talmud et al (2009), the lipid-associated SNPs were therefore passed through a stepwise variable selection scheme with the Akaike's Information Criterion (AIC) (Akaike 1974), with the aim of removing redundant associations and retaining the best predictors for each lipid trait (Talmud et al. 2009). Information criteria are used to select the best regression model from a set of possible models given the data. AIC is the most commonly used criterion. The AIC value reflects the goodness of fit of the model, but also includes a penalty that increases as the number of estimated parameters increases, which discourages over-fitting. The preferred model is the one with the minimum AIC value. SNP selection was carried out separately for each chromosome (since independence between SNPs on different chromosomes is expected) whereby the genetic model assumed an additive effect. Age and gender were included in the base model for the variable selection stage for all three lipid traits (Talmud et al. 2009).

For the selection of total cholesterol and LDL-associated SNPs, the *APOE* genotype was also included in the base model as it is the major determinant of total cholesterol and LDL-C levels (see section 1.5.2.3). Of the two SNPs that determine the major APOE isoforms, only rs7412 is represented on the Cardiochip genotyping platform. However, in both studies the two SNPs had been separately genotyped and the *APOE* genotype determined (Abdollahi et al. 2006; Sabia et al. 2010), and this data was used for this analysis. The SNPs that were retained after variable selection included 21 SNPs (including the 2 *APOE* SNPs) for total cholesterol, 23 SNPs (including the 2 *APOE* SNPs) for LDL-C, 12 SNPs for HDL-C, and 16 SNPs for triglycerides (Table 3.1 - Table 3.4).

### 3.2.7 Calculation of Lipid Genetic Risk Scores

In order to evaluate the combined effects of the common lipid-associated SNPs, a genetic risk score was calculated for each lipid fraction in each individual, which was a simple count of the number of risk alleles (for HDL-C-associated SNPs these are the number of HDL-C-lowering alleles) present in each individual. The score represents a summary of the genetic risk from the different variants that predispose an individual to increased lipid levels (in the case of HDL-C, decreased levels) and therefore increased CHD risk. For each lipid trait, genetic scores for each participant were calculated by summing the number of risk alleles (0, 1 or 2 for each SNP). Based on prior knowledge of the relationship between the *APOE* genotypes and LDL-C levels from a large meta-analysis in 86,067 healthy participants (Bennet et al. 2007) (refer to section 1.5.2.3), and for simplicity, the *APOE* risk count was coded as follows: ε2 carriers (ε2ε2, ε2ε3, ε2ε4) = 0, ε3ε3 = 1 and ε4 carriers (ε3ε4, ε4ε4) = 2, and included in the LDL-C and total cholesterol score calculation. The genetic scores were calculated in the same manner for BWHHS participants. Individuals with missing genotypes were excluded.

Where per-allele effects for each SNP are similar, a simple risk allele count is an easy and appropriate method to generate genetic risk scores since it assumes that each risk allele contributes equally to the phenotype. However, if the effect sizes are known to be different across risk alleles, then a weighted score is more appropriate. Weighted genetic risk scores can be calculated as follows:

**Equation 3.1**

$$GS_i = \sum_{j=1}^{m} \beta_j X_{ij}$$

where, $GS_i$ is the genetic score for the *i*th individual, *m* is the number of SNPs used in the score calculation, $X_{ij}$ is the risk allele count for the *j*-th SNP in individual *i*, and $\beta_j$ is the per-risk allele effect of SNP *j* on the trait of interest. Though a weighted score was developed in WHII using the per risk allele beta-coefficients from the regression of lipid fractions on SNPs in WHII, it provided very similar results to that of the unweighted score. The unweighted score is more likely to have clinical application because of its simplicity. Therefore, for ease of interpretation, only the results from the unweighted scores are presented here.

**Table 3.1 SNPs on the Cardiochip associated with total cholesterol in WHII, selected by stepwise regression using the Akaike Information Criterion for model selection.**

| SNP | CHR | BP (NCBI36) | GENE | Alleles | | | MAF | Common homozygotes mean total cholesterol mmol/L | Heterozygotes mean total cholesterol mmol/L | Rare homozygotes mean total cholesterol mmol/L | Risk-allele beta from regression with total Cholesterol | % total cholesterol variance explained |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Risk | Minor | Common | | | | | | |
| rs11591147 | 1 | 55278235 | *PCSK9* | G | T | G | 0.02 | 6.46 | 5.93 | 5.48 | 0.53 | 0.7179 |
| rs4970834 | 1 | 10961640 | *CELSR2* | G | A | G | 0.18 | 6.50 | 6.35 | 6.11 | 0.16 | 0.5824 |
| rs12740374 | 1 | 10961911 | *CELSR2* | G | T | G | 0.21 | 6.51 | 6.35 | 6.20 | 0.16 | 0.6349 |
| rs629301 | 1 | 10961982 | *CELSR2* | T | G | T | 0.21 | 6.51 | 6.36 | 6.20 | 0.15 | 0.5938 |
| rs934197 | 2 | 21120966 | *APOB* | A | A | G | 0.33 | 6.38 | 6.47 | 6.62 | 0.12 | 0.3987 |
| rs562338 | 2 | 21141826 | *APOB* | G | A | G | 0.18 | 6.50 | 6.33 | 6.20 | 0.17 | 0.6164 |
| rs4299376 | 2 | 43926080 | *ABCG8* | G | G | T | 0.32 | 6.35 | 6.50 | 6.63 | 0.14 | 0.7349 |
| rs12916 | 5 | 74692295 | *HMGCR* | C | C | T | 0.40 | 6.35 | 6.47 | 6.56 | 0.11 | 0.4695 |
| rs2575876 | 9 | 10670556 | *ABCA1* | G | A | G | 0.25 | 6.51 | 6.38 | 6.31 | 0.12 | 0.4098 |
| rs2072560 | 11 | 11616703 | *APOA5* | T | T | C | 0.06 | 6.41 | 6.67 | 7.51 | 0.31 | 0.8122 |
| rs9804646 | 11 | 11617028 | *APOA5* | G | A | G | 0.09 | 6.47 | 6.29 | 6.21 | 0.18 | 0.3832 |
| rs17248720 | 19 | 11059187 | *LDLR* | C | T | C | 0.13 | 6.53 | 6.21 | 5.9 | 0.31 | 1.79 |
| rs8102912 | 19 | 11066975 | *LDLR* | G | A | G | 0.23 | 6.53 | 6.33 | 6.27 | 0.17 | 0.7806 |
| rs2228671 | 19 | 11071912 | *LDLR* | C | T | C | 0.13 | 6.50 | 6.28 | 6.29 | 0.19 | 0.6905 |
| rs2304128 | 19 | 19607151 | *GMIP* | C | A | C | 0.09 | 6.47 | 6.33 | 5.67 | 0.18 | 0.4449 |
| rs10402271 | 19 | 50021054 | *BCAM* | G | G | T | 0.32 | 6.34 | 6.51 | 6.61 | 0.15 | 0.7496 |
| rs519113 | 19 | 50068124 | *PVRL2* | C | G | C | 0.24 | 6.50 | 6.39 | 6.26 | 0.12 | 0.403 |
| rs6859 | 19 | 50073874 | *PVRL2* | A | A | G | 0.41 | 6.35 | 6.46 | 6.60 | 0.12 | 0.5741 |
| rs12721109 | 19 | 50139061 | *APOC4* | G | A | G | 0.02 | 6.47 | 5.92 | 6.34 | 0.53 | 0.9296 |

**Table 3.2 SNPs on the Cardiochip associated with LDL-C in WHII, selected by stepwise regression using the Akaike Information Criterion for model selection.**

| SNP | CHR | BP (NCBI36) | GENE | Alleles |  |  | MAF | Common homozygotes mean LDL-C mmol/L | Heterozygotes mean LDL-C mmol/L | Rare homozygotes mean LDL-C mmol/L | Risk-allele beta from regression with LDL-C levels adjusted for sex and age | % LDL-C variance explained |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  | Risk | Minor | Common |  |  |  |  |  |  |
| rs11591147 | 1 | 55278235 | *PCSK9* | G | T | G | 0.02 | 4.38 | 3.84 | 2.98 | 0.55 | 0.95 |
| rs4970834 | 1 | 109616403 | *CELSR2* | G | A | G | 0.18 | 4.42 | 4.27 | 4.08 | 0.16 | 0.67 |
| rs12740374 | 1 | 109619113 | *CELSR2* | G | T | G | 0.21 | 4.43 | 4.28 | 4.12 | 0.15 | 0.75 |
| rs629301 | 1 | 109619829 | *CELSR2* | T | G | T | 0.21 | 4.43 | 4.28 | 4.14 | 0.15 | 0.7 |
| rs693 | 2 | 21085700 | *APOB* | A | G | A | 0.48 | 4.48 | 4.34 | 4.27 | 0.11 | 0.52 |
| rs934197 | 2 | 21120966 | *APOB* | A | A | G | 0.33 | 4.30 | 4.40 | 4.53 | 0.12 | 0.53 |
| rs562338 | 2 | 21141826 | *APOB* | G | A | G | 0.18 | 4.43 | 4.24 | 4.16 | 0.18 | 0.84 |
| rs4299376 | 2 | 43926080 | *ABCG8* | G | G | T | 0.32 | 4.28 | 4.42 | 4.53 | 0.13 | 0.78 |
| rs12916 | 5 | 74692295 | *HMGCR* | C | C | T | 0.41 | 4.26 | 4.39 | 4.51 | 0.12 | 0.72 |
| rs3804231 | 5 | 74732535 | *COL4A3BP* | A | A | G | 0.13 | 4.33 | 4.46 | 4.62 | 0.14 | 0.4 |
| rs2072560 | 11 | 116167036 | *APOA5* | T | T | C | 0.06 | 4.34 | 4.51 | 5.33 | 0.22 | 0.49 |
| rs17231506 | 16 | 55552029 | *CETP* | C | T | C | 0.32 | 4.44 | 4.33 | 4.22 | 0.11 | 0.51 |
| rs1529729 | 19 | 11024562 | *SMARCA4* | G | G | A | 0.45 | 4.28 | 4.39 | 4.44 | 0.09 | 0.36 |
| rs17248720 | 19 | 11059187 | *LDLR* | C | T | C | 0.13 | 4.45 | 4.14 | 3.81 | 0.31 | 0.22 |
| rs8102912 | 19 | 11066975 | *LDLR* | G | A | G | 0.23 | 4.44 | 4.26 | 4.19 | 0.16 | 0.84 |
| rs2228671 | 19 | 11071912 | *LDLR* | C | T | C | 0.13 | 4.42 | 4.21 | 4.18 | 0.18 | 0.81 |
| rs10402271 | 19 | 50021054 | *PVRL2* | G | G | T | 0.33 | 4.26 | 4.44 | 4.53 | 0.15 | 0.97 |
| rs519113 | 19 | 50068124 | *PVRL2* | C | G | C | 0.24 | 4.42 | 4.31 | 4.18 | 0.12 | 0.5 |
| rs6859 | 19 | 50073874 | *PVRL2* | A | A | G | 0.41 | 4.27 | 4.38 | 4.52 | 0.12 | 0.69 |
| rs283813 | 19 | 50081014 | *PVRL2* | T | A | T | 0.07 | 4.39 | 4.23 | 3.70 | 0.19 | 0.4 |
| rs12721109 | 19 | 50139061 | *APOC2* | G | A | G | 0.02 | 4.39 | 3.82 | 4.32 | 0.54 | 0.12 |

**Table 3.3 SNPs on the Cardiochip associated with HDL-C in WHII, selected by stepwise regression using the Akaike Information Criterion for model selection.**

| | | | | Alleles | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Chr** | **BP (NCBI 36)** | **Gene** | **Risk** | **Minor** | **Common** | **MAF** | **Common homozygotes mean HDL-C mmol/l** | **Heterozygotes mean HDL-C mmol/l** | **Rare homozygotes mean HDL-C mmol/l** | **Risk-allele beta from regression with $\log_e$(HDL-C) adjusted for sex and age** | **% HDL-C variance explained** |
| rs301 | 8 | 19861 | *LPL* | T | C | T | 0.25 | 1.41 | 1.46 | 1.52 | -0.04 | 0.69 |
| rs17410962 | 8 | 19892 | *SLC18A* | G | A | G | 0.13 | 1.42 | 1.48 | 1.52 | -0.04 | 0.55 |
| rs662799 | 11 | 11616 | *APOA4* | G | G | A | 0.06 | 1.44 | 1.37 | 1.28 | -0.06 | 0.47 |
| rs11820589 | 11 | 11613 | *BUD13* | A | A | G | 0.06 | 1.44 | 1.39 | 1.14 | -0.05 | 0.42 |
| rs4775041 | 15 | 56461 | *LIPC* | G | C | G | 0.30 | 1.41 | 1.43 | 1.53 | -0.03 | 0.45 |
| rs261342 | 15 | 56518 | *LIPC* | C | G | C | 0.22 | 1.41 | 1.45 | 1.54 | -0.03 | 0.55 |
| rs9989419 | 16 | 55542 | *CETP* | A | A | G | 0.40 | 1.49 | 1.42 | 1.34 | -0.05 | 1.58 |
| rs12708967 | 16 | 55550 | *CETP* | C | C | T | 0.19 | 1.47 | 1.38 | 1.29 | -0.06 | 1.71 |
| rs17231506 | 16 | 55552 | *CETP* | C | T | C | 0.32 | 1.37 | 1.47 | 1.56 | -0.07 | 2.62 |
| rs711752 | 16 | 55553 | *CETP* | G | A | G | 0.43 | 1.35 | 1.45 | 1.52 | -0.06 | 2.36 |
| rs5883 | 16 | 55564 | *CETP* | C | T | C | 0.06 | 1.42 | 1.51 | 1.54 | -0.05 | 0.43 |
| rs5880 | 16 | 55572 | *CETP* | C | C | G | 0.05 | 1.44 | 1.34 | 1.20 | -0.08 | 0.7 |

**Table 3.4 SNPs on the Cardiochip associated with triglycerides in WHII, selected by stepwise regression using the Akaike Information Criterion for model selection.**

| | | | | Alleles | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SNP | Chr | BP (NCBI36) | Gene | Risk | Minor | Common | MAF | Common homozygotes mean Triglyceride mmol/l | Heterozygotes mean Triglyceride mmol/l | Rare homozygotes mean Triglyceride mmol/l | Risk-allele beta from regression with log$_e$(Triglycerides) adjusted for sex and age | % Triglyceride variance explained |
| rs1260326 | 2 | 27584444 | *GCKR* | T | T | C | 0.40 | 1.38 | 1.43 | 1.63 | 0.06 | 0.57 |
| rs1714571 | 7 | 72542746 | *BAZ1B* | C | T | C | 0.20 | 1.50 | 1.36 | 1.25 | 0.09 | 0.71 |
| rs285 | 8 | 19859469 | *LPL* | C | T | C | 0.47 | 1.54 | 1.45 | 1.31 | 0.07 | 0.77 |
| rs331 | 8 | 19864685 | *LPL* | G | A | G | 0.27 | 1.51 | 1.39 | 1.24 | 0.08 | 0.81 |
| rs3289 | 8 | 19867472 | *LPL* | C | C | T | 0.03 | 1.43 | 1.70 | 1.04 | 0.16 | 0.38 |
| rs3916027 | 8 | 19869148 | *SLC18A* | G | A | G | 0.27 | 1.51 | 1.40 | 1.24 | 0.08 | 0.74 |
| rs1050366 | 8 | 19891970 | *SLC18A* | C | A | C | 0.11 | 1.48 | 1.31 | 1.08 | 0.11 | 0.73 |
| rs1732151 | 8 | 12655559 | *TRIB1* | A | G | A | 0.47 | 1.53 | 1.43 | 1.37 | 0.05 | 0.43 |
| rs1710899 | 10 | 95354023 | *RBP4* | G | G | C | 0.03 | 1.42 | 1.72 | 1.53 | 0.14 | 0.42 |
| rs6589565 | 11 | 11614544 | *BUD13* | A | A | G | 0.07 | 1.39 | 1.81 | 1.92 | 0.19 | 1.63 |
| rs1228603 | 11 | 11615741 | *ZNF259* | T | T | C | 0.06 | 1.42 | 1.64 | 1.83 | 0.13 | 0.80 |
| rs651821 | 11 | 11616778 | *APOA5* | C | C | T | 0.06 | 1.39 | 1.81 | 2.04 | 0.21 | 1.64 |
| rs1075009 | 11 | 11616924 | *APOA4* | G | G | A | 0.21 | 1.38 | 1.53 | 1.70 | 0.09 | 0.81 |
| rs3398910 | 11 | 11620535 | *APOC3* | T | T | C | 0.25 | 1.40 | 1.47 | 1.67 | 0.07 | 0.54 |
| rs5072 | 11 | 11621279 | *APOA1* | A | A | G | 0.08 | 1.41 | 1.60 | 1.84 | 0.11 | 0.55 |
| rs2304128 | 19 | 19607151 | *GMIP* | G | T | G | 0.09 | 1.47 | 1.31 | 0.99 | 0.10 | 0.53 |

### 3.2.8 Estimation of the 10-year Absolute Risk of Cardiovascular Disease

To calculate the Framingham 10 year risk of CVD, the equation presented by Anderson et al. (1991) was used, which uses linear model coefficients estimated in the Framingham Heart Study to weight each risk factor contribution to the overall risk. The equation incorporates information on gender, age, diabetes mellitus status, smoking status (current), systolic blood pressure, total cholesterol and HDL-C. The text and equations below, extracted from Anderson et al. (1991), describe how the risk of a CVD event within a given time frame can be calculated.

A parametric statistical model (Anderson 1991) was used to provide predicted probabilities for CVD outcome. This modelling is based on risk factor levels and times until events. Let *T* denote the time until the event of interest, and $x_1,\ x_2 \ldots\ x_k$ represent the risk factor measurements for an individual. The coefficients $\beta_0, \beta_1, \beta_2 \ldots \beta_k$, as well as θ and $\theta_1$, are the parameters estimated from the Framingham Heart Study and are extracted from Table 1 of the published manuscript by Anderson et al. (1991). The model assumes that *T* follows a Weibull distribution (Weibull 1951), and θ and $\theta_1$ are the scale and shape parameters for this distribution.

**Equation 3.2**

$$\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

where, $\mu$ is assumed to be a linear function of the risk factors and $\log(\sigma) = \theta + \theta_1\mu$ is considered to be a linear function of $\mu$. To compute the probability that time until event is less than some arbitrary time *t* for given values of $\mu$ and $\sigma$, let

**Equation 3.3**

$$u = \frac{\log(t) - \mu}{\sigma}$$

Assume,

$$P(T > t) = P\left\{\frac{\log(T) - \mu}{\sigma} > u\right\}$$

then,

**Equation 3.4**

$$\mathrm{P}(T < t) = 1 - \exp(-\exp(u))$$

where, *P(T<t)* is the predicted probability of an event by time *t.* The weights for the CVD risk factors and the values for Θ, $\Theta_1$ and $\beta_0$ are given below (extracted from Anderson et al. (1991)):

| | |
|---|---|
| Θ | 0.6536 |
| $\Theta_1$ | -0.2402 |
| $\beta_0$ | 18.8144 |
| Female | -1.2146 |
| $\log_e$(age) | -1.8443 |
| $\log_e$(age) x female | 0.3668 |
| $\log_e$(systolic BP) | -1.4032 |
| Smoking status | -0.3899 |
| $\log_e$(total cholesterol/ HDL-C) | -0.5390 |
| Diabetes status | -0.3036 |
| Diabetes status x female | -0.1697 |

Baseline measures in individuals with complete phenotype data in both studies (1991-1993 in WHII, and 1999-2001 in BWHHS) were used for calculating the Framingham 10 year CVD risk score. Since the equation incorporates total cholesterol levels and is designed for estimating risk in individuals without heart disease, the risk was not estimated for participants on lipid lowering medication or with CHD at baseline.

### 3.2.9 Association of Genetic Scores with Lipid Levels

The effects of the genetic scores on baseline lipid concentration were expressed per additional risk allele (equivalent to a unit change in the score), and also as the difference in lipid value between participants in the highest and lowest quintile of the genetic score distribution. Individuals on lipid medication at baseline were excluded from this analysis. Linear regression analysis was performed unadjusted and adjusted for gender (only in WHII) and age. To obtain a normal distribution, the HDL-C and triglyceride variables were $\log_e$-transformed prior to analysis.

### 3.2.10 Association of Genetic Scores with 'High-Risk' Status

An individual with Framingham 10 year risk of CVD greater than 20% was considered as 'high-risk', since this is the cut-off that has been used for therapeutic intervention in the UK. Using logistic regression, the odds ratio for having a baseline 10 year CVD risk > 20% was calculated for individuals in the top quintile of each lipid genetic score distribution with reference to those in the lowest quintile, unadjusted and adjusted for the respective lipid fraction.

### 3.2.11 Association of Genetic Scores with Lipid Medication Use

Using logistic regression and lipid medication data from the follow-up phases in both studies, the odds ratios for lipid medication use for primary prevention were calculated for individuals in the top quintile of the genetic score distribution with reference to individuals in the lowest quintile, unadjusted and adjusted for the respective lipid fraction. Since genotype precedes outcome, both incident and prevalent lipid drug users in the follow-up phases were considered. To ensure that analysis was restricted to subjects receiving lipid lowering treatment for primary rather than secondary prevention, individuals who had a CHD event prior to receiving lipid medication were excluded from this analysis. For comparison, the odds ratio for lipid medication use was also calculated for individuals with a Framingham 10 year CVD risk greater than 20% compared to those with lower risk.

### 3.2.12 Association of Genetic Scores with Coronary Events

Using CHD event data at the follow-up phases in both studies, the odds ratios for having a CHD event for those in the top quintile of the lipid score distribution compared to those in the bottom quintile, both unadjusted and adjusted for the relevant lipid level, were calculated using logistic regression. Since genotype precedes outcome, all individuals with a CHD event by the follow-up phase (both incident and prevalent) were included in the analysis. For comparison, the unadjusted odds ratio of developing CHD for individuals with high baseline Framingham 10 year CVD risk (> 20%) compared to those with lower risk was also calculated.

### 3.2.13 Discriminative Ability of Genetic Risk Scores

To evaluate the potential value of the lipid genetic scores for discrimination, the area under the receiver operating characteristic curve (AUROC) was calculated for each lipid genetic score for distinguishing ‘high-risk’ individuals, lipid medication usage, and CHD outcome. The receiver operating characteristic curve illustrates the performance of a binary classifier system, as its discrimination threshold is varied, by plotting the proportion of true positives (sensitivity) versus the proportion of false positives (specificity), at various threshold settings. For comparison, the AUROC using the individual lipid levels as predictor of all three outcomes were also calculated. The AUROC for the Framingham 10 year CVD risk score as a predictor for lipid drug use and CHD outcome were also calculated.

### 3.2.14 Improvement over Framingham 10 year CVD Risk

To determine if the lipid genetic scores improved discrimination of lipid drug users and CHD events above the Framingham 10 year CVD risk score, the predicted risk for each outcome was first calculated using a logistic regression model with only the Framingham risk score as a predictor (the baseline model), followed by both the Framingham risk score and each lipid genetic score in turn as predictors (enhanced model). The AUROC for both the baseline and enhanced models for each lipid genetic score were calculated. Analysis was done using the PredictABEL package v1.2.1 (Kundu et al. 2011) in R CRAN (R Development Core Team 2012).

### 3.2.15 Net Reclassification Improvement for in WHII

To quantify any improvement in classification of individuals with and without CHD when the lipid genetic scores were used in addition to the Framingham 10 year CVD risk score, the commonly used net reclassification improvement (NRI) (Pencina et al. 2008) was calculated in WHII. When calculating NRI, each subject in the data set has two risk values calculated, one according to the baseline model and one according to the enhanced model, and are then classified into pre-defined risk categories. The proportion of individuals that are reclassified into a different risk category by the enhanced model as compared with the baseline model are determined. The percentage that are correctly reclassified are CHD cases that are classified into a higher risk category, or non-cases that are reclassified into a lower risk category, in the enhanced model compared to the baseline model. The NRI summarises this information into a single value and represents the difference in proportions moving up and down risk categories among cases versus controls:

**Equation 3.5**

$$\mathrm{NRI} = [\mathrm{P}(\mathrm{up} \mid \mathrm{D} = 1) - \mathrm{P}(\mathrm{down} \mid \mathrm{D} = 1)] - [\mathrm{P}(\mathrm{up} \mid \mathrm{D} = 0) - \mathrm{P}(\mathrm{down} \mid \mathrm{D} = 0)]$$

where, upward movement (up) is the change into a higher category based on the new model and downward movement (down) is the change in the opposite direction, and *D* denotes the event indicator (cases=1, controls=0). An NRI of 0.1 means that 10% more cases were appropriately moved up a risk category than down compared with controls. NRI has the advantage over the ROC curve in that the categories can be formed based on clinically important risk estimates. The risk of a CHD event using either the baseline model (Framingham risk as the predictor) or using the enhanced model (both Framingham risk and the lipid genetic score as predictors) was calculated and individuals classified into two risk categories (risk <20% and risk >=20%), and the NRI calculated. Analysis was done using the PredictABEL package v1.2.1 (Kundu et al., 2011) in R CRAN (R Development Core Team 2012).

### 3.2.16 Summary of Phenotypic Data Used

For clarification, a summary of whether baseline or follow-up variables were used in each of the above described analyses is shown below.

**Table 3.5 A summary of the phenotypic data used in each analysis.**

| **Analysis** | **WHII** | **BWHHS** |
|---|---|---|
| Baseline measurement | 1991-1993 | 1999-2001 |
| Follow-up measurement | 2003-2004 | 2007 |
| Collection of biological samples for DNA extraction | Follow-up | Baseline |
| Calculation of Framingham 10 year CVD risk | Baseline | Baseline |
| Association of gene scores with lipid levels | Baseline | Baseline |
| Association of gene score with lipid medication use | Follow-up | Follow up |
| Association of gene score with CHD | Follow up | Follow up |

## 3.3 Results

### 3.3.1 Participant Characteristics

The baseline characteristics of the participants from the two studies are shown in Table 3.6. WHII individuals are younger and have much lower baseline CVD risk compared to BWHHS individuals. In WHII, of those individuals that did not have CHD and were not on lipid medication at baseline, 8% had an estimated 10-year CVD risk > 20%. On follow-up (~10yrs later) 32% of these 'high-risk' individuals were on lipid medication, while only 7% of low-risk individuals (baseline CVD risk ≤20%) were on lipid medication at follow-up. In BWHHS, 49% had CVD risk > 20% at baseline, of which 34% were on medication at follow-up (~8 years later), while only 8% of BWHHS individuals with baseline CVD risk ≤20% were on lipid medication at follow-up.

**Table 3.6 Cohort characteristics for WHII and BWHHS.**

| | Whitehall II | | BWHHS |
|---|---|---|---|
| | **Men (N=3721)** | **Women (N=1338)** | **Women (N=3414)** |
| **Baseline** | | | |
| Age (yrs) | 49.1 (5.9) | 49.6(6.1) | 68.8(5.5) |
| BMI (kg/m$^2$) | 25.0 (3.1) | 25.3(4.7) | 27.6(4.9) |
| % Smokers (current) | 11 | 15 | 11 |
| % Smokers (ex/current ) | 51 | 46 | 44 |
| Diastolic BP (mmHg) | 80.7 (8.9) | 76.1(9.3) | 79.4 (11.7) |
| Systolic BP (mmHg) | 121.5 (12.8) | 116.6 (13.7) | 146.9 (25.3) |
| Framingham 10yr risk (%) | 10.6 (6.9) | 5.6 (4.9) | 22.1 (11.7) |
| Total Cholesterol (mmol/l) | 6.5 (1.1) | 6.4 (1.2) | 6.6 (1.2) |
| LDL-C (mmol/l) | 4.4 (1.0) | 4.2 (1.1) | 4.2 (1.1) |
| HDL-C (mmol/l) | 1.3 (0.4) | 1.7 (0.4) | 1.7 (0.5) |
| Triglyceride (mmol/l) | 1.6 (1.2) | 1.1 (0.7) | 1.9 (1.2) |
| | | | |
| **Baseline** | | | |
| Lipid drug users | 33 (0.9%) | 10 (0.7%) | 204 (5.9%) |
| CHD cases | 96 (2.6%) | 25 (1.9%) | 460 (13.4%) |
| **Follow-up phase** | | | |
| Duration from baseline | ~10yrs | | ~8yrs |
| Lipid drug users | 426 (11.4%) | 121 (9.0%) | 692 (20.1%) |
| CHD cases | 334 (9.0%) | 87 (6.5) | 802 (23.3%) |

### 3.3.2 Association of Genetic Risk Scores with Blood Lipids

Though per allele effects on lipid values are small (Table 3.7) there is substantial difference in mean lipid levels between individuals in the highest and lowest quintile of the genetic score distribution. Individuals in the top quintile of the cholesterol genetic score distribution have 0.96 (0.85 – 1.07) mmol/L and 0.62 (0.46 – 0.78) mmol/L higher total cholesterol than those in the bottom quintile in WHII and BWHHS, respectively (Table 3.7). Those in the top quintile of the LDL genetic score distribution had 0.85 (0.76 – 0.94) and 0.63 (0.50 – 0.76) mmol/L higher LDL-C than those in the bottom quintile in WHII and BWHHS, respectively. Similarly, the mean HDL-C levels were substantially lower (20% and 15%) and mean triglyceride levels higher (38% and 26%) in the top quintile compared to the bottom quintile in WHII and BWHHS, respectively (Table 3.7).

### 3.3.3 Association of Genetic Risk Scores with 'High-Risk' Status

Individuals in the top quintile of the distributions of each of the four lipid genetic scores tended to have a higher odds of being identified as 'high-risk', as determined by the Framingham 10 year CVD risk >20% (Table 3.8). The triglyceride genetic score showed the strongest association, with individuals in the top quintile of the triglyceride score distribution having a 1.99 (1.39 – 2.85) and 1.56 (1.22 – 2.00) fold higher odds of having 10 year CVD risk >20% compared to those in the bottom quintile in WHII and BWHHS, respectively. Adjusting for the respective baseline lipid levels completely attenuated the association of all genetics scores (Table 3.8). None of the lipid genetic scores were associated with risk factors incorporated in the Framingham risk equation other than blood lipids (shown for WHII in Table 3.9, indicating the association of the genetic scores with Framingham risk is driven simply by the effect of genotype on lipid levels.

**Table 3.7 Association of genetic scores with lipid levels.**

| Genetic score | Outcome | Study | Per allele effect | | | Top vs. bottom quintile | | |
|---|---|---|---|---|---|---|---|---|
| | | | Beta (95%CI)* | P-value | N | Beta (95%CI)* | P-value | N |
| Total cholesterol | Total cholesterol | WHII | 0.10 (0.09, 0.11) | $<1.0x10^{-50}$ | 4677 | 0.96 (0.85, 1.07) | $<1.0x10^{-50}$ | 1517 |
| | | BWHHS | 0.08 (0.07, 0.09) | $1.2x10^{-26}$ | 2560 | 0.62 (0.46, 0.78) | $8.7x10^{-14}$ | 810 |
| LDL | LDL-C | WHII | 0.09 (0.08, 0.10) | $<1.0x10^{-50}$ | 4596 | 0.85 (0.76, 0.94) | $<1.0x10^{-50}$ | 1786 |
| | | BWHHS | 0.07 (0.06, 0.08) | $6.7x10^{-32}$ | 2502 | 0.63 (0.50, 0.76) | $3.5x10^{-21}$ | 1001 |
| HDL | log(HDL-C) | WHII | -0.03 (-0.03, -0.03) | $<1.0x10^{-50}$ | 4534 | -0.20 (-0.23, -0.18) | $<1.0x10^{-50}$ | 2011 |
| | | BWHHS | -0.02 (-0.02, -0.01) | $1.9x10^{-16}$ | 2533 | -0.15 (-0.19, 0.12) | $8.610^{-16}$ | 866 |
| Triglyceride | log(triglyceride) | WHII | 0.04 (0.04, 0.05) | $<1.0x10^{-50}$ | 4549 | 0.38 (0.33, 0.43) | $<1.0x10^{-50}$ | 1725 |
| | | BWHHS | 0.04 (0.04, 0.05) | $2.1x10^{-27}$ | 2553 | 0.26 (0.21, 0.31) | $1.5x10^{-21}$ | 1284 |

*For untransformed outcomes the beta coefficient represents the mmol/l change in lipid levels. For log-transformed outcomes the beta coefficient represents the percentage change in lipid levels. In each case the beta-coefficients are adjusted for gender (only in WHII) and age.

**Table 3.8 Association of genetic scores with 'high-risk' status (Framingham 10yr CVD risk >20%).**

| Genetic Score | Study | Top vs. Bottom Quintile Unadjusted Analysis | | Top vs. Bottom Quintile Adjusted for Lipid Fraction | | N in bottom quintile | | N in top quintile | |
|---|---|---|---|---|---|---|---|---|---|
| | | OR (95%CI) | P-value | OR (95%CI) | P-value | CVD risk >20 | CVD risk <20 | CVD risk >20 | CVD risk <20 |
| Total cholesterol | WHII | 1.53 (1.03, 2.26) | 0.034 | 0.90 (0.58, 1.39) | 0.63 | 62 | 850 | 49 | 440 |
| | BWHHS | 1.30 (0.95, 1.77) | 0.097 | 1.05 (0.76, 1.45) | 0.78 | 206 | 249 | 128 | 119 |
| LDL | WHII | 1.36 (0.93, 1.98) | 0.11 | 0.80 (0.53, 1.20) | 0.28 | 59 | 885 | 59 | 651 |
| | BWHHS | 1.49 (1.14, 1.94) | 0.003 | 1.04 (0.78, 1.39) | 0.79 | 200 | 258 | 229 | 198 |
| HDL | WHII | 1.94 (1.37, 2.74) | 0.00017 | 0.93 (0.64, 1.37) | 0.72 | 63 | 1029 | 79 | 666 |
| | BWHHS | 1.23 (0.91, 1.66) | 0.17 | 0.86 (0.62, 1.19) | 0.35 | 225 | 255 | 139 | 128 |
| Triglyceride | WHII | 1.99 (1.39, 2.85) | 0.00017 | 0.95 (0.63, 1.44) | 0.81 | 59 | 891 | 73 | 553 |
| | BWHHS | 1.56 (1.22, 2.00) | 0.00037 | 1.03 (0.79, 1.35) | 0.81 | 325 | 390 | 228 | 175 |

**Table 3.9 Association in WHII of lipid genetic scores with cardiovascular risk factors, other than lipids, used in the calculation of Framingham 10 year CVD risk.**

| | Cholesterol Score<br>beta (95% CI)*; P-value | LDL Score<br>beta (95% CI)*; P-value | HDL Score<br>beta (95% CI)*; P-value | Triglyceride Score<br>beta (95% CI)*; P-value |
|---|---|---|---|---|
| Gender | 0.01 (-0.03,0.06); 0.58 | 0.03 (-0.007,0.076); 0.10 | 0.007 (-0.03,0.05); 0.71 | 0.009 (-0.03,0.05); 0.67 |
| Age | -0.15 (-0.77, 0.47); 0.64 | -0.18 (-0.74,0.38); 0.53 | 0.19 (-0.34,0.73); 0.48 | -0.08 (-0.64, 0.49); 0.79 |
| SBP | -0.80 (-2.53,0.93); 0.36 | -0.47 (-2.0,1.06); 0.55 | 0.13 (-1.31,1.57); 0.86 | 0.50 (-1.09,2.08); 0.54 |
| DBP | -0.17 (-1.14,0.81); 0.74 | -0.38 (-1.23,0.46); 0.34 | 0.14 (-0.68,0.96); 0.74 | 0.21 (-0.67,1.10); 0.64 |
| Smoking (current) | -0.003 (-0.04,0.03); 0.86 | -0.01 (-0.04,0.02); 0.38 | 0.02 (-0.014,0.05); 0.30 | 0.03 (-0.006,0.06); 0.12 |
| Diabetes | -0.01 (-0.04, 0.02); 0.47 | -0.02 (-0.04,0.008); 0.17 | 0.01 (-0.01,0.04); 0.31 | -0.004 (-0.03,0.02); 0.77 |

*Beta-coefficients shown for the comparison between the top and bottom quintiles of each lipid genetic score

### 3.3.4 Association of Genetic Risk Scores with Lipid Medication Use

Individuals in the top quintile of the LDL-C genetic score had a 2.38 (1.57 - 3.59) and 2.24 (1.52 - 3.29) fold higher odds of receiving lipid medication than those in the lowest quintile (Figure 3.1A) in WHII and BWHHS, respectively. However, adjustment for LDL-C concentration completely attenuated this association in WHII (Figure 3.1B). In BWHHS, though the association was substantially reduced, it remained significant (Figure 3.1B). Individuals in the top quintile of the total cholesterol and triglyceride genetic scores were more likely to use lipid medication and these associations were attenuated to the null after adjusting for total cholesterol and triglyceride levels (Figure 3.1). The HDL-C genetic score was not significantly associated with lipid medication use (Figure 3.1). As expected, individuals with an estimated CVD risk >20% had a higher likelihood (WHII OR = 4.15 (3.04 – 5.67); BWHHS OR = 2.98 (2.32-3.83)) of receiving lipid medication compared to those with lower risk.

### 3.3.5 Association of Genetic Risk Scores with CHD Events

Individuals in the top quintile (compared to bottom quintile) of the LDL-C genetic score distribution had a higher risk of CHD (WHII OR = 1.43 (1.02 – 2.00) and BWHHS OR = 1.31 (0.99 - 1.72)) (Figure 3.2A). After adjusting for LDL-C levels, this association was completely attenuated in WHII but not in BWHHS (Figure 3.2B). Similar associations were seen in both studies for the total cholesterol genetic score (Figure 3.2). The triglyceride score showed association with higher risk of CHD in WHII but not in BWHHS (Figure 3.2). The HDL-C genetic score was not associated with CHD outcome. By comparison, individuals with a Framingham 10 year CVD risk >20% had a 4.21 (3.08 – 5.75) and 2.49 (1.80 – 3.44) fold higher odds of CHD in WHII and BWHHS, respectively.

**Figure 3.1 Association of lipid genetic scores with lipid medication use**. Odds ratio (with 95% CI and p-value) of using lipid-modifying drugs in top vs. bottom quintiles of each genetic score distribution for (A) unadjusted analyses and (B) adjusted for the respective lipid fraction.

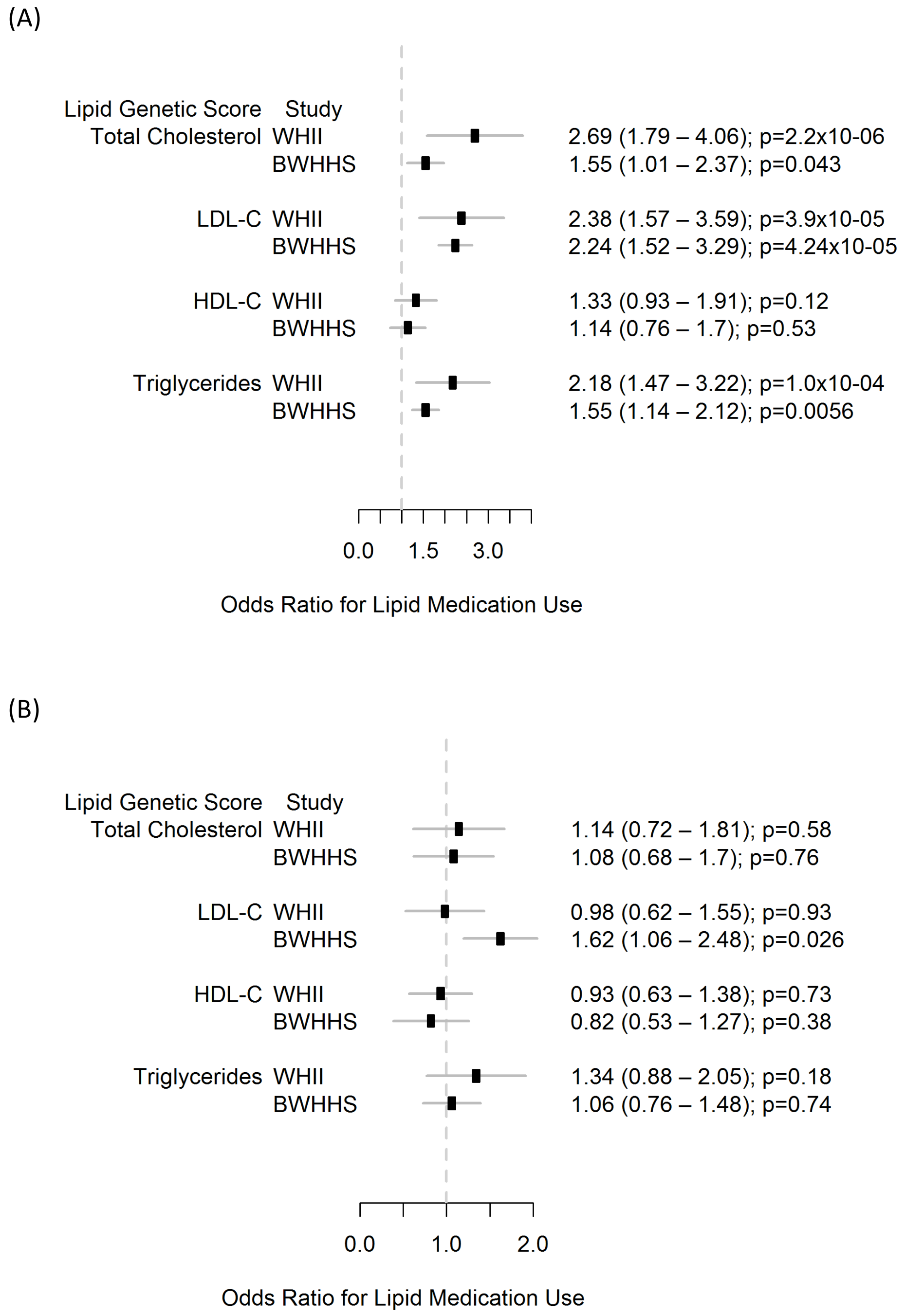
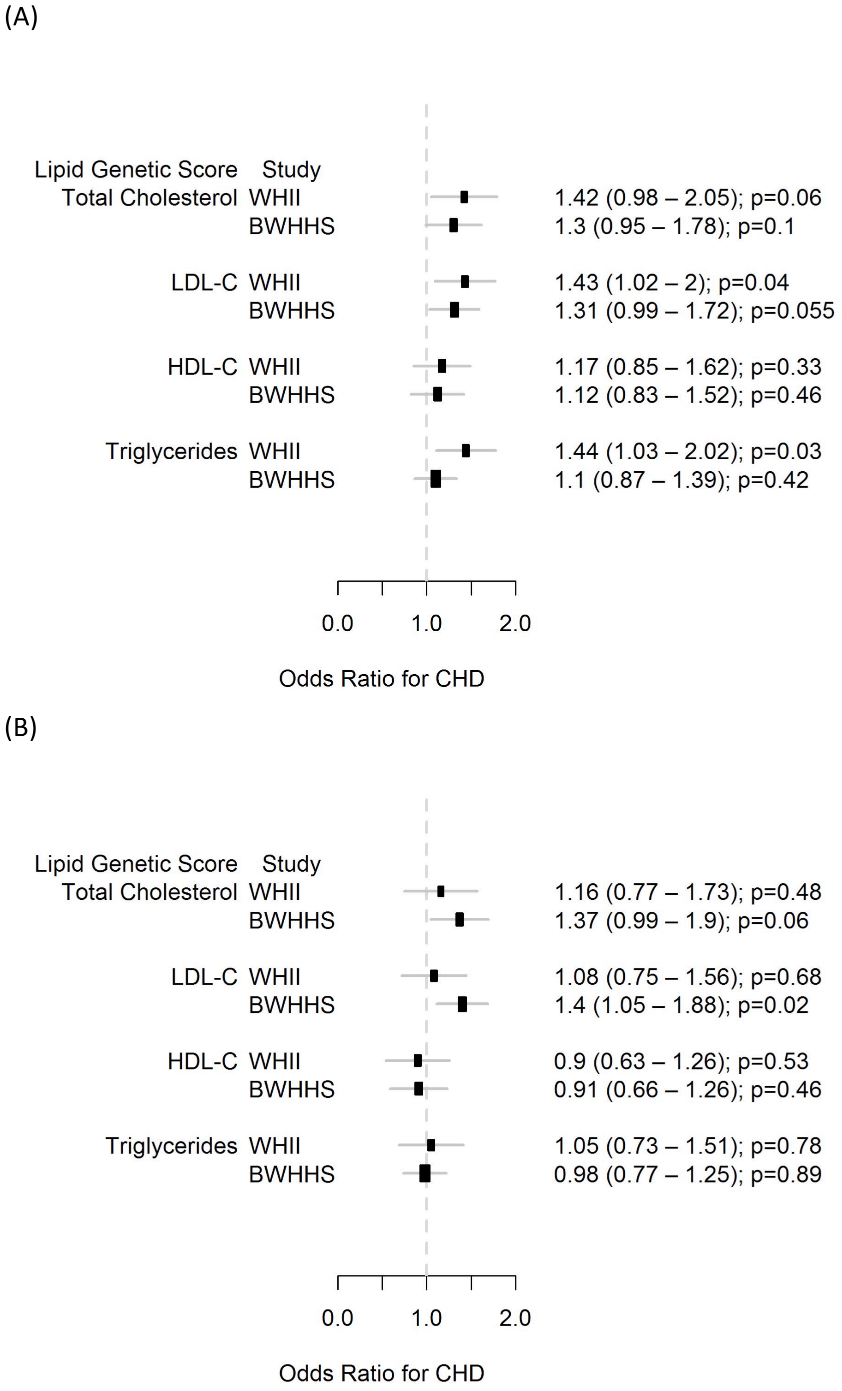


**Figure 3.2 Association of lipid genetic scores with CHD.** Odds ratio (with 95% CI and p-value) of CHD outcome for individuals in the top quintile of each lipid genetic score distribution compared to individuals in the bottom quintile. Odds ratios and p-values are shown for (A) unadjusted analyses and (B) adjusted for the respective lipid fraction.



### 3.3.6 Comparison of Genotype-based and Phenotype-based Discrimination

Blood lipid measurements performed better that the respective genetic scores for discriminating high-risk individuals, lipid medication use and CHD outcome, with all AUROC in WHII above 0.6, while those for the respective genetic scores were all below 0.6 (shown for WHII in Table 3.10). Total cholesterol and LDL-C levels performed the best for discriminating lipid medication, exhibiting an AUROC of 0.79 (0.76 – 0.81) and 0.78 (0.75 – 0.80), while HDL-C and triglyceride levels performed best for discriminating high-risk individuals. The latter is not surprising, since both of these measures are incorporated into the Framingham risk calculation. The Framingham 10 year CVD risk score performed the best for CHD discrimination, and the inclusion of the lipid genetic scores in addition to the Framingham risk score in the model did not improve discrimination (Table 3.11). The performance of each of the lipid genetic scores in comparison to the Framingham 10 year CVD risk score for discriminating high-risk individuals, lipid medication use and CHD in both studies is shown in Figure 3.3.

### 3.3.7 Net Reclassification for Coronary Disease Events in WHII

There was no significant improvement in classification over the Framingham 10 year CVD risk when any of the genetic scores were added to the risk prediction model, with less than 1% of individuals being correctly reclassified in the enhanced model (p-value > 0.13) (Table 3.12). Addition of all four genetic scores to the risk prediction model also did not improve classification (0.03% correctly reclassified, p-value = 0.97).

**Table 3.10 Area under the receiver operating curve (AUROC) for lipid levels and lipid genetic risk scores in WHII**

| | AUROC | | |
|---|---|---|---|
| | High-risk | Lipid drug use | CHD |
| Total Cholesterol | 0.70 (0.67 - 0.73) | 0.79 (0.76 - 0.81) | 0.61 (0.58 - 0.64) |
| Cholesterol genetic score | 0.54 (0.51 - 0.57) | 0.60 (0.57 - 0.64) | 0.54 (0.51 - 0.57) |
| | | | |
| LDL-C | 0.71 (0.68 - 0.74) | 0.78 (0.75 - 0.80) | 0.62 (0.59 - 0.64) |
| LDL-C genetic score | 0.54 (0.50 - 0.57) | 0.59 (0.56 - 0.62) | 0.53 (0.50 - 0.56) |
| | | | |
| HDL-C | 0.76 (0.74 - 0.79) | 0.59 (0.56 - 0.62) | 0.60 (0.57 - 0.63) |
| HDL-C genetic score | 0.58 (0.54 - 0.61) | 0.53 (0.50 - 0.57) | 0.52 (0.49 - 0.55) |
| | | | |
| Triglycerides | 0.78 (0.75 - 0.80) | 0.71 (0.68 - 0.74) | 0.62 (0.59 - 0.65) |
| Triglyceride genetic score | 0.56 (0.53 - 0.59) | 0.56 (0.53 - 0.60) | 0.53 (0.50 - 0.56) |

**Table 3.11 Area under the receiver operating curve (AUROC) for combined Framingham 10yr CVD risk score (FRS) and lipid genetic scores**

| **Exposure** | **Study** | **AUROC for *actual* drug use** | **AUROC for CHD** |
|---|---|---|---|
| FRS | WHII | 0.73 (0.70 – 0.76) | 0.71 (0.67 – 0.74) |
| | BWHHS | 0.67 (0.64 – 0.70) | 0.65 (0.61 – 0.69) |
| FRS + cholesterol genetic score | WHII | 0.74 (0.71 – 0.77) | 0.68 (0.65 – 0.72) |
| | BWHHS | 0.68 (0.65 – 0.71) | 0.65 (0.61 – 0.69) |
| FRS +LDL genetic score | WHII | 0.74 (0.71 – 0.77) | 0.68 (0.64 – 0.71) |
| | BWHHS | 0.68 (0.65 – 0.71) | 0.65 (0.61 – 0.69) |
| FRS +HDL genetic score | WHII | 0.71 (0.67 – 0.74) | 0.70 (0.66 – 0.73) |
| | BWHHS | 0.67 (0.64 – 0.70) | 0.64 (0.60 – 0.68) |
| FRS +triglyceride genetic score | WHII | 0.72 (0.69 – 0.76) | 0.70 (0.66 – 0.73) |
| | BWHHS | 0.67 (0.64 – 0.70) | 0.64 (0.60 – 0.68) |

**Figure 3.3 Discrimination of high-risk individuals, lipid drug usage and coronary heart disease events using lipid genetic scores.** AUROC for discriminating between  high-risk individuals based on a Framingham risk >20% in (A) WHII and (B) BWHHS, actual use of lipid medication in (C) WHII and (D) BWHHS, and coronary heart disease in (E) WHII and (F) BWHHS. TC=total cholesterol, TG=triglycerides.

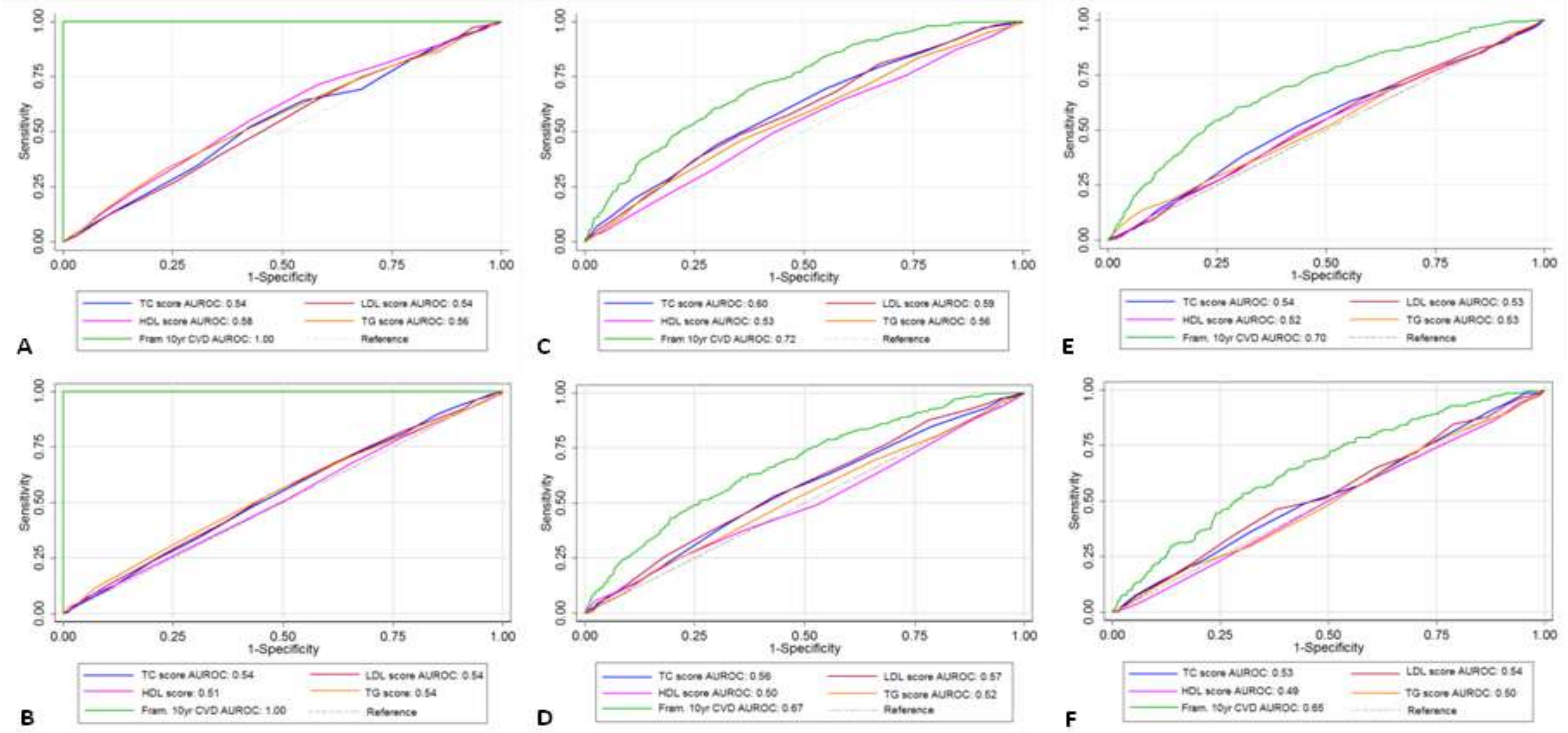


**Table 3.12 Reclassification in WHII based on Framingham 10yr CVD risk and lipid genetic score risk model.** Green signifies correctly reclassified, yellow no reclassification and red incorrectly reclassified. FRS=Framingham 10-year CVD risk score

| | | | FRS + Genetic Score | | | | |
|---|---|---|---|---|---|---|---|
| | | | <20% | ≥20% | % Reclassified | Correctly Reclassified | NRI (p-value) |
| **Cholesterol Genetic score** | | | | | | | |
| **Controls (N=3945)** | **FRS** | **<20%** | 3868 | 2 | 0 | 0.05% | |
| | | **≥20%** | 4 | 71 | 5 | | |
| **Cases (N=252)** | | **<20%** | 235 | 2 | 1 | 0.79% | |
| | | **≥20%** | 0 | 15 | 0 | | 0.0084 (0.13) |
| **LDL Genetic Score** | | | | | | | |
| **Controls (N=3558)** | **FRS** | **<20%** | 3480 | 3 | 0 | 0.06% | |
| | | **≥20%** | 5 | 70 | 7 | | |
| **Cases (N=249)** | | **<20%** | 233 | 2 | 1 | 0.40% | |
| | | **≥20%** | 1 | 13 | 7 | | 0.0045 (0.52) |
| **HDL Genetic Score** | | | | | | | |
| **Controls (N=3984)** | **FRS** | **<20%** | 3907 | 1 | 0 | 0.05% | |
| | | **≥20%** | 3 | 73 | 4 | | |
| **Cases (N=254)** | | **<20%** | 239 | 0 | 0 | 0.00% | |
| | | **≥20%** | 0 | 15 | 0 | | 0.0005 (0.32) |
| **Triglyceride Genetic Score** | | | | | | | |
| **Controls (N=3947)** | **FRS** | **<20%** | 3871 | 1 | 0 | -0.03% | |
| | | **≥20%** | 0 | 75 | 0 | | |
| **Cases (N=252)** | | **<20%** | 236 | 1 | 0 | 0.40% | |
| | | **≥20%** | 0 | 15 | 0 | | 0.0037 (0.35) |

## 3.4 Discussion

### 3.4.1 Summary of Results

Individuals in the top quintile of the LDL-C and total cholesterol genetic score distributions, calculated using 23 LDL-C-associated and 21 total cholesterol-associated genetic variants, respectively, tended to have greater odds of having high CVD risk status, receiving lipid-lowering medication and having a CHD event than individuals in the bottom quintile, in two UK studies of middle-aged men and women. Despite predisposing to lifelong differences in levels of blood lipids, the strength of the genetic associations was insufficiently large to usefully discriminate individuals likely to require lipid-lowering treatment or develop CHD. The Framingham 10 year CVD risk score which incorporates a single mid-life measurement of total cholesterol and HDL-C as well as other non-genetic risk factors, performed better than genetic scores for CHD discrimination, and addition of the genetic risk scores to the Framingham 10 year CVD risk did not improve discrimination or reclassification.

### 3.4.2 Comparison with Previous Studies

Murray et al (2009) found that LDL-C and triglyceride genetic scores based on 7 and 11 SNPs (identified by previous GWAS), respectively, were associated with the likelihood of exceeding the lipid thresholds for intervention, as advocated by US guidelines (by the National Institutes of Health Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults), in an Italian sample of 1155 individuals over 65 years, but that a score based on 9 HDL-C- associated variants was not (Murray et al. 2009). The study also showed that the HDL-C and triglyceride risk allele counts were associated with MI. This study did not examine associations with estimates of absolute CVD risk or the number of individuals actually treated with lipid-modifying drugs. Any improvement in discrimination or reclassification over commonly used non-genetic risk scores was also not assessed. For the analysis in WHII and BWHHS, the genetic scores were based on SNPs identified in WHII using the Cardiochip (Keating et al. 2008), which has denser SNP coverage of many of the loci associated with blood lipid fractions, but lesser genome coverage than the arrays used in GWAS. SNPs studied by Murray et al were either present or had proxy SNPs (based on HapMap CEU LD estimates; $R^2$> 0.8) present on the Cardiochip, which allowed the use of variable selection methods to identify the best genetic predictors of lipid levels from the much larger number of significant associations observed in each region.

A study by Kathiresan et al. (2008) examined the utility of LDL-C and HDL-C genetic scores for the discrimination of CVD events. They generated a single score based on a smaller subset of 11 SNPs in 9 genes associated with either LDL-C or HDL-C from published studies. The genetic score was associated with incident CVD events even after adjustment for lipid levels. They found that a model that incorporated the genetic score did not improve the discrimination of CVD events but did modestly improve risk classification in 193 CVD cases (MI, ischemic stroke, and death from CHD) and 4039 controls. However, the three risk categories used were 0-10%, >10-20%, and >20%. In the context of UK guidelines for primary CVD prevention, there would be no alteration in therapeutic intervention decisions for those reclassified between the lowest two risk categories.

### 3.4.3 CVD-Associated SNPs and Prediction

Since this analysis, new variants associated with the principal lipid fractions have been identified by large-scale association analysis (Teslovich et al. 2010; Asselbergs et al. 2012). In addition, GWAS have also identified multiple variants associated with CVD outcome (McPherson et al. 2007; Samani et al. 2007; Aulchenko et al. 2009). Whether these additional genetic variants can provide sufficiently accurate predictions to enable genetically-informed intervention decisions remains unclear. Several studies have explored whether genetic markers can improve risk prediction using genetic risk scores, but the results from these studies have been conflicting or modest. Although some have shown that genetic variants are associated with CVD outcome independent of conventional risk factors, only one study was able to demonstrate a clinically significant improvement in predictive ability using both measures of risk reclassification and discrimination, which are more informative for the purpose of risk prediction (Table 3.13) (Di Angelantonio & Butterworth 2012). However, these findings need to be considered in the light of several limitations of the studies. Firstly, the number of events considered in these studies is relatively small and therefore they are likely to be underpowered to detect significant improvement in risk prediction. Second, only a few (typically around a dozen) selected genetic variants, are included in the genetic risk score (Di Angelantonio & Butterworth 2012).

New approaches have shown that the proportion of variance explained when all SNPs on the array are considered is much larger than that explained by SNPs passing a preset significance threshold (Yang et al. 2010). Alternative approaches to risk prediction that incorporate all variants nominally associated with CVD risk may provide more power but may also be prone to bias and non-transferability (Di Angelantonio & Butterworth 2012). The selection and combination of genetic variants is essential to maximise the potential improvement in risk prediction over and above risk factors currently used in risk prediction. Recent work on the power and predictive accuracy of polygenic scores has shown that very large sample sizes, up to an order of magnitude greater than currently available, would be needed for estimating predictors to a level which is useful for prediction (Dudbridge 2013). Therefore, as sample sizes begin to grow, prediction using polygenic scores may become more feasible.

**Table 3.13 Prospective studies assessing CVD risk prediction using multiple genetic markers and risk-prediction metrics**

| | | Outcomes Assessed | | Selection of SNPs | | | Variables included in model | | Improvement in risk-prediction metrics | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Association with: | | | | | |
| Authors, Year | Population Source | No. | Type | No. | CVD/CHD | Risk Factors | Established risk factors | Family History | Discrimination | Reclassification |
| Drenos et al, 2007 | General population | 183 | CHD | 12 | No | Yes | Yes | No | Yes | NA |
| Morrison et al, 2007 | General population | 1452 | CHD | 11 | Yes | No | Yes | No | Yes | NA |
| Kathiresan et al, 2008 | General population | 238 | CVD | 11 | No | Yes | Yes | No | No | Yes |
| Paynter et al, 2010 | General population | 777 | CVD | 101 | Yes | Yes | Yes | Yes | No | No |
| | | | | 12 | Yes | No | Yes | Yes | No | No |
| Ripatti et al, 2010 | General population | 1015 | CVD | 13 | Yes | No | Yes | No | No | No |
| Thanassoulis et al, 2012 | General population | 539 | CVD | 102 | Yes | Yes | Yes | Yes | No | No |
| | | | | 29 | Yes | No | Yes | Yes | No | No |
| Vaarhorst et al, 2012 | General population | 642 | CHD | 179 | Yes | Yes | Yes | Yes | No | No |
| | | | | 153 | No | Yes | Yes | Yes | No | No |
| | | | | 29 | Yes | No | Yes | Yes | No | Yes |
| Lluis-Ganella et al, 2012 | General population | 536 | CHD | 8 | Yes | No | Yes | No | Yes | No |
| Andreassi et al, 2012 | Hospital (pre-existing CHD) | 119 | CVD | 48 | Yes | No | Yes | No | No | NA |
| Hughes et al, 2012 | General population | 632 | CHD | 13 | Yes | No | Yes | Yes | No | Yes |
| | | | | 15 | Yes | No | Yes | Yes | Yes | Yes |
| | | | | 8 | Yes | No | Yes | Yes | Yes | Yes |

Table modified from Di Angelantonio & Butterworth (2012)

### 3.4.4 Clinical Implications

Consistent evidence from this and other studies now indicates that prediction based on phenotype outdoes prediction based on common genotypic variation. The current American Heart Association policy on the use of common genetic variants for risk prediction states that though there is robust evidence linking common variants to complex CVD, the minor improvement in discrimination creates scepticism about the clinical utility for risk prediction (Ashley et al. 2012). It is currently uncertain how the genotype results can or should influence clinical management. More importantly, interpretation of results can be challenging and time consuming for clinicians and patients, which is likely to steer them away from the use of genotyping for CVD risk prediction (Ashley et al. 2012). There is a lack of studies informing the clinical benefit of providing such genetic information to patients and funding for such clinical studies is essential to build an evidence base for the field (Ashley et al. 2012).

### 3.4.5 Limitations

Though a large number of SNPs associated with the major blood lipid fractions were studied, these variants collectively explain only a small proportion of the variance in blood lipid levels and only a fraction of the heritability (Talmud et al. 2009). Given the gene-centric design of the array, loci outside known cardiovascular pathways would have been missed. Since this work was completed, the Global Lipids Genetic Consortium (GLGC) conducted a meta-analysis of GWAS involving more than 100,000 participants, and together with the recent Cardiochip-based discovery meta-analysis in over 60,000 individuals, the list of loci influencing the major blood lipid fractions has increased to almost 100 (Teslovich et al. 2010; Asselbergs et al. 2012). Scores based on a larger number of lipid related SNPs will likely explain a larger proportion of the variance in blood lipids and have larger average differences in lipid concentrations in individuals at opposite extremes of the score distribution.

However, the ability of genetic scores incorporating these additional SNPs to identify individuals with ‘high-risk’ status or CHD events may not be correspondingly large because the effect sizes of additional loci identified in very large meta-analysis tend to be extremely small. Additionally, new SNPs are also distributed across different chromosomes and inherited independently, so that only a small proportion of the population carries a large burden of lipid raising alleles. Further analysis based on all currently known lipid related loci will be needed to determine if the interpretations of these findings, based on the Cardiochip array-derived lipid genetic risk scores, on the utility of lipid related SNPs for predicting important healthcare outcomes will substantially alter. It is important to note that despite individuals with Apoe e2e2 having lower total-cholesterol and LDL-C levels based on the Bennet et al meta analysis, the relationship with CHD is complex due to its causal role in hyperlipoproteinemia. However, since this genotype is rare, any impact on the results would be insignificant.

Ongoing efforts to fine map causal variants at the known loci may increase the number of eligible SNPs and improve the performance of lipid related genetic scores. The effects of gene-gene and gene-environment interactions were not modelled and may also contribute to the missing phenotype variance explained. Efforts to deeply re-sequence for rare variants at the relevant genomic regions may also identify highly penetrant (albeit rare) alleles with a larger effect on blood lipid levels than those studied here, and incorporating these into genetic risk score calculations may improve performance.

The associations observed in WHII are likely to be overestimated since the same data was used both for SNP discovery and evaluation of the performance of the allele scores. However, associations and performance estimates were broadly similar in BWHHS. Given the strong association of some lipid genetic scores with lipid drug usage, individuals with a higher number of risk alleles are more likely to be put on lipid medication, thus reducing their risk of an event. Therefore the association of lipid genetic scores with CHD outcome may be underestimated. This also applies to the Framingham 10 year CVD risk score, whereby those with a baseline risk >20% were more likely to be put on lipid medication and the exclusion of higher-risk CHD patients from the latter analysis may have blunted the true association of Framingham risk with CHD outcome.

The Framingham risk equations were developed based on data from a sample population in Framingham, Massachusetts, and there have been studies showing that this method over-estimates risk in other populations (Hense et al. 2003; Kent 2002; Brindle et al. 2003). Despite this, they have been used widely both within the UK and elsewhere. The QRISK cardiovascular disease risk algorithm (QRISK2) (Hippisley-Cox et al. 2007) has been developed to provide accurate estimates of cardiovascular risk in patients from different ethnic groups in England and Wales and would be a more appropriate estimator for the cohorts used. However, the QRISK requires information on participant post codes, which is not available in WHII, and also the equations underpinning the calculation are not freely available since QRISK is licensed for commercial or healthcare use.

# 4 Developing Genetic Instruments for Lipids

## 4.1 Introduction

For prognostic research, all factors associated with an outcome, whether causal or not, are of interest (Sheehan et al. 2008). Causality on the other hand, is relevant for informing health interventions and in drug discovery. MR analysis uses genetic variants as unbiased proxies for modifiable risk factors in order to determine whether their relationship with an outcome is causal (refer to section 1.6.2). Due to the random assortment of alleles at the time of gamete formation, the population distributions of genetic variants are generally independent of behavioural and environmental factors that typically confound epidemiological associations between putative risk factors and outcome (Smith & Ebrahim 2004). The unidirectional flow of biological information from gene through to risk factor and then to disease outcome avoids reverse causation, since the disease or outcome cannot change the inherited genetic variants that are associated with the risk factor.

Higher LDL-C concentration is associated with a higher risk of CHD, and the relationship is considered causal because randomised trials using LDL-C-lowering interventions such as HMG-CoA reductase inhibitors (statins) have shown to reduce CHD risk in proportion to the LDL-C reduction (Baigent et al. 2005; Baigent et al. 2010). Epidemiological studies have shown that increased triglyceride levels and decreased HDL-C levels are both associated with CHD. However, randomised trials of drugs directed at HDL-C and triglycerides have not shown consistent results and have therefore been unable to confirm or refute whether these associations are causal (Cannon et al. 2010; Ginsberg et al. 2010; Jun et al. 2010; NHLBI Communications 2011). There is therefore a lot of interest in assessing the causal relationship of these lipid fractions with various clinically relevant outcomes.

Confirmation of causality would determine whether development of drugs designed to raise HDL-C or lower triglyceride levels are worth pursuing for cardiovascular disease risk management. Given the large number of genetic variants reported to be associated with lipid levels, the aim of this work is to explore different approaches for the development of suitable genetic instruments for LDL-C, HDL-C and triglycerides to use in MR analyses. Genotypic and phenotypic data from 5059 Caucasian individuals from the WHII cohort were used for lipid instrument development.

## 4.2 Materials & Methods

### 4.2.1 Genotypic and Phenotypic Data

The development of genetic instruments utilised genotype data for 5059 Caucasian individuals from the WHII cohort (see section 2.2.1.3 for cohort description) that had been genotyped using the Illumina Cardiochip (as described in section 2.2.4), and baseline (1991-1993) lipid measurements (as described in section 3.2.2).

### 4.2.2 Measuring the Strength of a Genetic Instrument

As mentioned in section 1.6.2.4, the strength of an instrument can be measured by the proportion of the total phenotypic variance explained by the instrument, R$^2$, also known as the coefficient of determination. This can be calculated from the simple linear regression of the genetic instrument with the exposure of interest as follows:

**Equation 4.1**

$$R^2 = 1 - \left( \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \right)$$

where, $y_i$ is the observed risk factor value for the *i*th individual, $\hat{y}_i$ is the fitted value for the *i*th individual and $\bar{y}$ is the mean of the observed data. The numerator is the residual sum of errors (the unexplained variance) and the denominator is the total sum of squares (the total variance in the observed data). The F-statistic is another measure of instrument strength and is related to R$^2$, sample size (*n*) and number of instruments (*k*) used, and is calculated as follows:

**Equation 4.2**

$$F = \frac{R^2(n-k-1)}{(1-R^2)k}$$

As mentioned in section 1.6.2.3, the causal effect can be estimated using the 2SLS method. In the presence of confounding between the exposure and outcome, the casual estimate from a 2SLS regression will be biased towards this confounded association, and the extent of bias is inversely related to the F-statistic (Staiger and Stock 1997). Bias occurs when genetic variants explain not only systematic variation in the risk factor of interest, but also chance variation in the confounders (Burgess & Thompson 2013). As R$^2$ increases, the F-statistic increases and the bias decreases. However, including additional instruments that do not increase the first stage R$^2$ results in a decrease in the F-statistic (Equation 4.2), and hence increases the bias (Palmer et al. 2011). Therefore, weak instrument bias may arise in situations where the number of instruments is large and sample size small. By rule of thumb, an F-statistic >10 is usually an indicator of a strong instrument, as the bias of the IV estimator is 10% of the bias of the observational estimator (Staiger & Stock 1997; Lawlor et al. 2008). However, weak instrument bias is a continuous rather than binary phenomenon, and such application of F-statistic thresholds for assessing weak instruments are not considered useful by some (Burgess & Thompson 2013). However, with large enough sample sizes, estimates using a weak instrument will be consistent for the causal effect. The R$^2$ is usually preferred over the F-statistic when comparing strength of different IVs, since the latter is dependent on the number of instruments used as well as sample size.

### 4.2.3 Unweighted Genetic Risk Score

When multiple, independent SNPs are associated with the exposure of interest, a simple approach for MR analysis is to combine them into a single genetic risk score instrument (refer to section 1.6.2.8). This approach avoids the problem of weak instrument bias that may potentially arise in the presence of a large number of possible instruments. Use of such scores assumes an approximately additive effect on phenotype. In the large-scale lipid association analysis by Talmud et al (2009), SNPs that passed the initial discovery significance threshold (p-value < $1x10^{-05}$; with age and sex as covariates in the regression model) were entered into a stepwise regression step using the AIC (Akaike 1974) for model selection, in an attempt to select a model with the best, non-redundant genetic predictors for each lipid fraction (described in section 3.2.6). The set of SNPs retained in the model included 23 SNPs (including the 2 SNPs making up the *APOE* genotype) for LDL-C, 12 SNPs for HDL-C, and 16 SNPs for triglycerides (refer to Table 3.2 - Table 3.4). The set of SNPs selected for each lipid fraction were combined into a genetic risk score, where for each individual this was the sum of the risk allele counts across the SNPs (previously described in section 3.2.7). To explore how the addition of each SNP to the genetic score affected the strength of the instrument, the following steps, using LDL-C as an example, were carried out:

- Step 1: Genetic variants were ranked in order of decreasing $R^2$ (obtained from the univariate linear regression of LDL-C on each genetic variant).
- Step 2: The unweighted score was first calculated for the single genetic variant with the highest $R^2$. In this instance the genetic score for all

individuals was either 0 (no risk alleles), 1 (heterozygous) or 2 (homozygous for the risk allele).

- Step 3: To assess instrument strength, the $R^2$ and F-statistic for the unweighted score were obtained from the linear regression with LDL-C.
- Step 4: To assess specificity, the $R^2$ values from the linear regression of the LDL-C unweighted score with the other (non-specific) lipid fractions and CVD risk factors (systolic blood pressure (SBP), diastolic blood pressure (DBP), CRP and BMI) were also obtained.
- Step 5: The next ranking SNP was added to the genetic score calculation and steps 3 and 4 repeated with this new score.
- Step 6: Step 5 was repeated until all SNPs had been incorporated into the genetic score calculation.

For simplicity, as was done in the previous chapter (section 3.2.7), the *APOE* genotype was coded as follows: ε2 carriers (ε2ε2/ ε2ε3/ ε2ε4) = 0, ε3ε3 = 1 and ε4 carriers (ε3ε4/ε4ε4) = 2. The above steps were repeated for the HDL-C and triglyceride SNPs, with the $R^2$ and F-statistic for assessing instrument strength derived from the regression with $\log_e$-transformed HDL-C and $\log_e$-transformed triglycerides, respectively.

### 4.2.4 Multiple Instruments Approach

Multiple SNPs associated with the exposure of interest can also be used simultaneously as individual IVs in a multiple instruments approach (refer to section 1.6.2.8). The multiple instruments approach maximises power, while making no assumptions regarding the effect sizes of each SNP (Pierce et al. 2010). However, if a large number of SNPs are used as individual instruments, there is potential for introducing weak instrument bias. As was done with the unweighted score, SNPs were ranked by decreasing $R^2$ values, and for each successive SNP addition, the $R^2$ and F-statistic values were extracted from a multiple regression of the SNPs in question with the respective lipid fraction. To examine specificity, the $R^2$ values were also extracted from the multiple regression of the SNPs with the other non-specific lipid fractions and CVD risk factors.

### 4.2.5 Weighted Genetic Score using Univariate Beta-Coefficients

As discussed in section 3.2.7, genetic risk scores can also be weighted by the effect size of each risk allele, which is more appropriate when effects sizes of SNPs are different. The beta-coefficients from the univariate regression of the lipid fraction on each SNP in WHII were used as weights. Though it was possible to generate weights for the *APOE* genotypes from the WHII data, a previously published meta-analysis with a sample size more than 10 times that of WHII had estimated the effect (in mmol/L) of the *APOE* genotype on LDL-C levels (Bennet et al. 2007) (section 1.5.2.3). Based on this study, where ε3/ε3 individuals were used as the reference group, the *APOE* genotypes were weighted as follows: ε2ε2 = -0.9, ε2ε3 = -0.4, ε2ε4 = -0.2, ε3/ε3 = 0, ε3ε4 = 0.1 and ε4ε4 = 0.2.

### 4.2.6 SNP Multicollinearity

The problem of multicollinearity occurs when two or more predictor variables are highly correlated (Montgomery et al. 2012). Presence of correlated SNPs can lead to large changes in individual effect estimates when other predictors are added or removed from the model, and may also result in an insignificant coefficient of a predictor variable in a multiple regression analysis, despite the simple linear regression showing the coefficient to be significantly different from zero (Montgomery et al. 2012).

To determine the presence and quantify the degree of collinearity between the AIC-selected SNPs, the beta-coefficients from the univariate regression of each SNP with the respective lipid fraction were compared to the beta-coefficients obtained from a multiple regression where all SNPs were used as predictors simultaneously. The variance inflation factor (VIF) was also calculated for each SNP in each lipid genetic score using the *vif()* function from the car package in R CRAN (R Development Core Team 2012). The VIF is a measure of how much the variance of an estimated regression coefficient is increased due to collinearity and is calculated using Equation 4.3.

**Equation 4.3**

$$VIF = \frac{1}{1 - R_j^2}$$

where, $R_j^2$ is the coefficient of determination of a regression of predictor *j* obtained when *j* is regressed on all the other predictors. When the predictor *j* is uncorrelated $R_j^2$ will be small and VIF close to 1, while if *j* is nearly linearly dependent on some of the subset of the remaining predictors, $R_j^2$ will be close to 1 and VIF becomes very large. The square root of the VIF is an indication of how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model (Belsey et al. 1980) e.g if the VIF of a predictor variable is 5, the standard error for the coefficient of that predictor variable is √5 = 2.2 times as large as it would be if that predictor variable was uncorrelated with the other predictor variables. VIFs exceeding 5 are commonly considered high and are an indication of a possible multicollinearity problem, while VIFs exceeding 10 indicate a definite multicollinearity problem.

### 4.2.7 Weighted Genetic Score using Bayesian Information Criterion for SNP Selection followed by Ridge Regression

A straightforward solution to avoid multicollinearity would be to simply use the most significantly associated SNP from each gene region in the score calculation. However, with this approach multiple independent signals in a single gene would be missed, there is ambiguity in defining gene boundaries, and SNPs in different (neighbouring) genes may not necessarily be independent since LD blocks can span several genes. To automate SNP selection and allow multiple SNPs in a gene to be selected, but reduce the multicollinearity problem between selected SNPs, two measures were taken:

1. Firstly, the SNPs associated with baseline LDL-C, HDL-C or triglycerides in WHII (p-value < $1x10^{-05}$) were included in a stepwise variable selection scheme with a more stringent information criterion - the Bayesian Information Criterion (BIC) (Schwarz 1978), using the *step()* function from the stats package in R CRAN (R Development Core Team 2012). The BIC imposes a larger penalty than AIC as the number of predictors in the model increases, and therefore tends to choose a model with fewer SNPs than AIC. Sex and age were included in the baseline model, and for selection of LDL-C SNPs, the *APOE* genotype was also included in the baseline model.

2. Secondly, the beta-coefficients used to weight the SNP risk allele counts were obtained from a Ridge regression (Brown 1994). The Ridge regression is a variant of ordinary multiple linear regression that 'shrinks' the beta-coefficients of redundant SNPs, thereby circumventing issues that may arise if highly-correlated SNPs are present in the model. Though OLS estimates are unbiased, the presence of multicollinearity results in large variances for the estimated regression coefficients. Ridge regression trades a small amount of bias in the coefficient estimates for a substantial reduction in coefficient sampling variance, producing a smaller mean-squared error of

estimation of the coefficients. Ridge regression requires the specification of a Ridge constant, which controls the extent to which Ridge estimates differ from the least-squares estimates. The larger the Ridge constant, the greater the bias and the smaller the variance of the Ridge estimator. The Ridge regression using the Lawless and Wang estimate of the Ridge constant (Lawless & Wang 1976) was carried out on the BIC-selected SNPs using the *lm.ridge()* function from the MASS package in R CRAN (R Development Core Team 2012), with sex and age included in the base model. For the LDL-C SNPs, the *APOE* genotype was also included in the base model. The final weights for each SNP were the beta-coefficients from the Ridge regression. For the *APOE* genotype, weights were derived from the Bennet et al (2007) paper, as specified in section 4.2.5.

## 4.3 Results

### 4.3.1 Comparison of Instrument Strength

The single strongest SNP instruments explained 5%, 3% and 2% of the total variation in LDL-C, HDL-C and triglycerides, respectively (Figure 4.1). As more SNPs were added, either to the genetic score calculation or used as multiple instruments, the $R^2$ value tended to increase. This is shown for the AIC-selected SNPs using both the unweighted genetic score in Figure 4.1, and the multiple instruments approach in Figure 4.2. When all SNPs (whether selected using AIC or BIC) were combined into a single genetic risk score or used as multiple instruments, the final $R^2$ values were twice as large as that of the single, strongest SNP instrument (Table 4.1). For each lipid fraction, three approaches provided the best instruments with equivalent strength: the multiple instruments approach using AIC-selected SNPs, multiple instruments approach using BIC-selected SNPs, and the weighted genetic score based on BIC-selected SNPs and beta-coefficients from a Ridge regression as weights, with all three explaining 13%, 7% and 7% of the total variation in LDL-C, HDL-C and triglycerides, respectively (Table 4.1). For all approaches, the final F-statistics were much larger than 10 (Table 4.1). As the number of SNP instruments increased, the F-statistic for the genetic score also increased, while that for the multiple instruments approach decreased. As SNPs with very small effect are added, the $R^2$ does not increase by much, and since the F-statistic depends on the $R^2$ and the number of instruments used, as more instruments with small effect are added, the F-statistic will decrease. With the genetic risk score the number of instruments is always one, regardless of the number of SNPs used.

### 4.3.2 Instrument Specificity

To evaluate specificity of the instruments, the $R^2$ values from the regression of each instrument with the other, non-specific lipid fractions and CVD risk factors were extracted. Table 4.2 shows the $R^2$ values for each instrument from the regression with the lipid fraction of interest, as well as the largest $R^2$ from the association of the instrument with a non-specific risk factor. As well as explaining 5% of the variation in LDL-C, the *APOE* genotype explained 1% of the total variation in CRP levels. As more SNPs were added to the genetic score, the strength of the association with CRP decreased, with the final AIC-selected unweighted genetic score explaining only 0.2% of the variation in CRP. Overall, the LDL-C genetic scores appeared to be specific for LDL-C, explaining less than 0.7% of the variation in any of the other non-specific CVD risk factors. The single strongest SNP instruments for HDL-C and triglycerides explained less than 0.6% of the variation in any of the other CVD risk factors. However, the HDL-C genetic scores explained 1-2% of the variation in triglycerides, while the triglyceride genetic scores also explained between 1-2% of the variation in HDL-C (Table 4.1). For the multiple instruments approach, whether selected using AIC or BIC, the LDL-C SNPs also explained 3% of the variation in HDL-C, the HDL-C SNPs explained 4% of the variation in

**Figure 4.1 Strength and specificity of the unweighted LDL-C, HDL-C and triglyceride genetic scores.** The plot shows the $R^2$ value of the unweighted genetic score as each AIC-selected SNP is added to the genetic score calculation. SNPs were added in the order of decreasing $R^2$ (strongest, single SNP instrument added first). For each addition of a SNP to the score calculation, the $R^2$ derived from the regression of the unweighted genetic score with its respective lipid fraction and from the regression of the score with the other lipid fractions and CVD risk factors (as a measure of specificity) are shown.

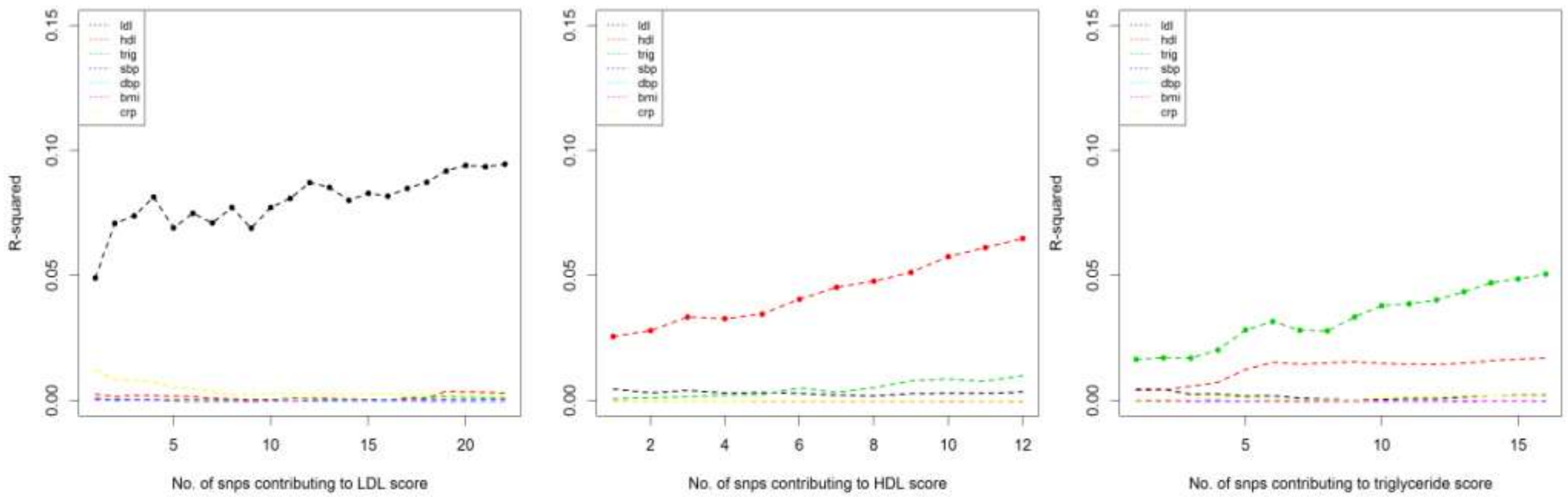


**Figure 4.2 Strength and specificity of the multiple instruments approach.** The plot shows the $R^2$ value of the multiple instruments approach as more AIC-selected SNPs are added as instruments. SNPs were added to the multiple regression model in order of decreasing $R^2$ (strongest SNP instruments added first). For each addition of a SNP to the multiple regression model, the $R^2$ derived from the regression of the SNPs with the respective lipid fraction and from the regression of the SNPs with the other lipid fractions and CVD risk factors (as a measure of specificity) are shown.

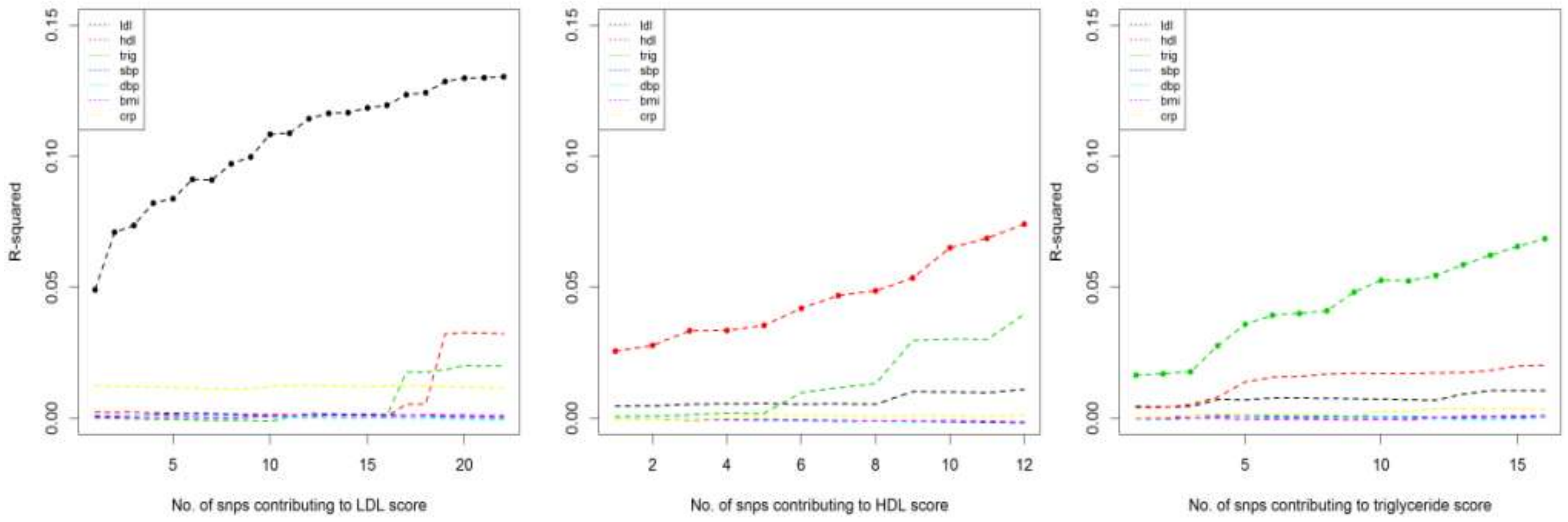


**Table 4.1 Instrument strength**. The final $R^2$ and F-statistic derived from the regression of each lipid genetic instrument with its respective lipid fraction are shown as a measure of instrument strength. The different instruments assessed are: single, strongest SNP instrument; unweighted genetic score using AIC-selected SNPs; weighted genetic score using AIC-selected SNPs and beta-coefficients from the univariate regression of each SNP with the respective lipid fraction as weights; multiple instruments approach using AIC-selected SNPs; weighted genetic score using BIC-selected SNPs and beta-coefficients from a Ridge regression as weights; multiple instruments approach using BIC-selected SNPs.

| | LDL-C | | HDL-C | | Triglycerides | |
|---|---|---|---|---|---|---|
| Instrument | $R^2$ | F | $R^2$ | F | $R^2$ | F |
| Single (strongest) genetic variant | 0.05 | 279 | 0.03 | 127 | 0.02 | 79 |
| Unweighted genetic risk score | 0.09 | 581 | 0.06 | 308 | 0.05 | 219 |
| AIC-selected, univariate coefficient- | 0.11 | 582 | 0.06 | 308 | 0.04 | 219 |
| Multiple instruments (AIC-selected SNPs) | 0.13 | 33 | 0.07 | 33 | 0.07 | 23 |
| BIC-selected, Ridge coefficient-weighted | 0.13 | 697 | 0.07 | 371 | 0.07 | 259 |
| Multiple instruments (BIC-selected SNPs) | 0.13 | 45 | 0.07 | 33 | 0.07 | 27 |

.

**Table 4.2 Instrument specificity.** The final $R^2$ derived from the regression of each lipid genetic instrument with its respective lipid fraction and the largest $R^2$ from the regression with other CVD factors (shown in brackets). The different instruments assessed are: single, strongest SNP instrument; unweighted genetic score using AIC-selected SNPs; weighted genetic score using AIC-selected SNPs and beta-coefficients from the univariate regression of each SNP with the respective lipid fraction as weights; multiple instruments approach using AIC-selected SNPs; weighted genetic score using BIC-selected SNPs and beta-coefficients from a Ridge regression as weights; multiple instruments approach using BIC-selected SNPs.

| Instrument | Genetic instrument for LDL-C | | Genetic instrument for HDL-C | | Genetic instrument for Triglycerides | |
|---|---|---|---|---|---|---|
| | $R^2$ from regression with LDL-C | Largest $R^2$ from regression with a non-specific risk factor (named in | $R^2$ from regression with HDL-C | Largest $R^2$ for non-specific risk factor (named in brackets) | $R^2$ from regression with triglycerides | Largest $R^2$ for non-specific risk factor (named in brackets) |
| Single (strongest) genetic variant | 0.05 | 0.01 ( CRP) | 0.03 | 0.005 (LDL-C) | 0.02 | 0.004 (LDL-C & HDL- |
| Unweighted genetic risk score | 0.09 | 0.003 (HDL-C) | 0.06 | 0.01 (triglycerides) | 0.05 | 0.02 (HDL-C) |
| AIC-selected, univariate coefficient-weighted genetic score | 0.11 | 0.003 (HDL-C) | 0.06 | 0.01 (triglycerides) | 0.04 | 0.01 (HDL-C) |
| Multiple instruments (AIC-selected SNPs) | 0.13 | 0.03 (HDL-C) | 0.07 | 0.04 (triglycerides) | 0.07 | 0.02 (HDL-C) |
| BIC-selected, Ridge coefficient-weighted genetic score | 0.13 | 0.006 (HDL-C) | 0.07 | 0.02 (triglycerides) | 0.07 | 0.02 (HDL-C) |
| Multiple instruments (BIC-selected SNPs) | 0.13 | 0.03 (HDL-C) | 0.07 | 0.04 (triglycerides) | 0.07 | 0.02 (HDL-C) |

triglycerides, and the triglyceride SNPs explained 2% of the variation in HDL-C (Figure 4.2).

### 4.3.3 SNP Collinearity

For the AIC-selected SNPs, rather than a continuous increase in $R^2$, addition of certain SNPs to the genetic score resulted in a drop in the $R^2$ value (Figure 4.1). This was due to the presence of multicollinearity between some SNPs. Multicollinearity was not an issue in the multiple instruments approach since no assumptions are required regarding the effect size for each SNP (Figure 4.2). As mentioned earlier, presence of multicollinearity can result in unstable beta-coefficient estimation. Table 4.3 compares the univariate beta-coefficients with those obtained from a multiple regression of the AIC-selected SNPs with LDL-C. For some SNPs, the effect estimate from the multiple regression analysis was in the opposite direction to that of the univariate beta-coefficient, in others the estimate was inflated, and for some the association with LDL-C was no longer significant. Two SNPs (rs629301 and rs12740374), both in the *CELSR2* gene, had inflated beta-coefficients, and both had very large VIFs (> 100). For rs629301 the effect size was also in the opposite direction (Table 4.3). These two SNPs were found to be in perfect LD according to HapMap and 1000 Genomes data. For the AIC-selected HDL-C SNPs, though the multiple regression p-values were less significant than the univariate p-values, there were no large differences in the beta-coefficients and all VIFs were less than 5 (Table 4.4). Two of the triglyceride-associated SNPs, rs331 and rs3916027, had VIFs >100 (Table 4.5). These two SNPs were also in perfect LD based on HapMap and 1000 Genomes data. In addition, the two SNPs with VIF ~6 were also in LD ($r^2$=0.54; *D′*=1). For three of the triglyceride-associated SNPs, the beta-coefficients from the multiple regression were in the opposite direction to the univariate beta-coefficients (Table 4.5).

**Table 4.3 Effects of multicollinearity on regression coefficients for the LDL-C genetic score SNPs.** Regression coefficients for the AIC-selected SNPs from a univariate regression (single SNP as predictor) and multiple regression (all SNPs as predictors) with LDL-C levels. No adjustment for covariates was made. SNPs where multicollinearity results in either beta-coefficients in the two regression analyses to be in the opposite direction, inflated beta-coefficients in the multiple regression analysis, large variance inflation factors (>5), or where the association p-value in the multiple regression is less than 0.05, are highlighted in red.

| | | | | Univariate analysis | | Multivariate analysis | | VIF |
|---|---|---|---|---|---|---|---|---|
| SNP | Chromosome | Position (NCBI36) | Closest Gene | Beta | P-Value | Beta | P-Value | |
| rs11591147 | 1 | 55278235 | *PCSK9* | 0.55 | 3.46E-11 | 0.53 | 1.36E-11 | 1.0 |
| rs4970834 | 1 | 109616403 | *CELSR2* | 0.15 | 2.94E-08 | 0.06 | **1.62E-01** | 3.0 |
| rs12740374 | 1 | 109619113 | *CELSR2* | 0.15 | 4.98E-09 | **1.21** | 3.42E-04 | **184.7** |
| rs629301 | 1 | 109619829 | *CELSR2* | 0.15 | 2.19E-08 | **-1.09** | 1.12E-03 | **183.6** |
| rs693 | 2 | 21085700 | *APOB* | 0.11 | 4.14E-07 | 0.04 | **1.15E-01** | 1.6 |
| rs934197 | 2 | 21120966 | *APOB* | 0.11 | 1.03E-06 | 0.06 | **2.47E-02** | 1.5 |
| rs562338 | 2 | 21141826 | *APOB* | 0.18 | 2.17E-10 | 0.13 | 3.15E-06 | 1.1 |
| rs4299376 | 2 | 43926080 | *ABCG8* | 0.13 | 2.69E-09 | 0.12 | 1.50E-08 | 1.0 |
| rs12916 | 5 | 74692295 | *HMGCR* | 0.12 | 9.79E-09 | 0.10 | 2.03E-05 | 1.3 |
| rs3804231 | 5 | 74732535 | *COL4A3BP* | 0.13 | 5.61E-05 | 0.05 | **1.40E-01** | 1.3 |
| rs2072560 | 11 | 116167036 | *APOA5* | 0.22 | 1.17E-06 | 0.20 | 2.66E-06 | 1.0 |
| rs17231506 | 16 | 55552029 | *CETP* | 0.11 | 1.91E-06 | 0.10 | 1.25E-06 | 1.0 |
| rs1529729 | 19 | 11024562 | *SMARCA4* | 0.09 | 5.39E-05 | 0.03 | **1.78E-01** | 1.1 |
| rs17248720 | 19 | 11059187 | *LDLR* | 0.31 | 4.73E-23 | 0.43 | 4.95E-15 | 3.4 |
| rs8102912 | 19 | 11066975 | *LDLR* | 0.15 | 1.15E-09 | 0.04 | **2.07E-01** | 1.9 |
| rs2228671 | 19 | 11071912 | *LDLR* | 0.18 | 3.26E-09 | **-0.20** | 1.99E-04 | 3.5 |
| Apoe | 19 | 45411829 | *APOE* | 0.37 | 3.03E-52 | 0.29 | 6.22E-30 | 1.4 |
| rs10402271 | 19 | 50021054 | *PVRL2* | 0.15 | 1.06E-10 | 0.05 | 1.89E-02 | 1.1 |
| rs519113 | 19 | 50068124 | *PVRL2* | 0.12 | 1.90E-06 | 0.05 | 3.87E-02 | 1.1 |
| rs6859 | 19 | 50073874 | *PVRL2* | 0.13 | 7.26E-09 | 0.04 | **8.12E-02** | 1.1 |
| rs283813 | 19 | 50081014 | *PVRL2* | 0.17 | 4.48E-05 | 0.11 | 7.26E-03 | 1.1 |
| rs12721109 | 19 | 50139061 | *APOC2* | 0.54 | 2.30E-13 | 0.23 | 1.51E-03 | 1.2 |

**Table 4.4 Effects of multicollinearity on regression coefficients for the HDL-C genetic score SNPs.** Regression coefficients for the AIC-selected SNPs from a univariate regression (single SNP as predictor) and multiple regression (all SNPs as predictors) with $\log_e$-transformed HDL-C levels. No adjustment for covariates was made. SNPs where multicollinearity results in either beta-coefficients in the two regression analyses to be in the opposite direction, inflated beta-coefficients in the multiple regression analysis, large variance inflation factors (>5), or where the association p-value in the multiple regression is less than 0.05, are highlighted in red.

| | | Position | | Univariate Regression | | Multiple Regression | | |
|---|---|---|---|---|---|---|---|---|
| SNP | Chromosome | (NCBI36) | Closest Gene | Beta | P-value | Beta | P-value | VIF |
| rs301 | 8 | 19861214 | *LPL* | 0.04 | 1.15E-08 | 0.02 | 1.43E-03 | 1.3 |
| rs17410962 | 8 | 19892360 | *SLC18A1* | 0.04 | 6.21E-07 | 0.03 | 1.31E-03 | 1.3 |
| rs11820589 | 11 | 116139072 | *BUD13* | 0.05 | 7.57E-06 | 0.06 | 1.22E-07 | 1.0 |
| rs662799 | 11 | 116168917 | *APOA4* | 0.06 | 1.84E-06 | 0.06 | 5.04E-08 | 1.0 |
| rs4775041 | 15 | 56461987 | *LIPC* | 0.03 | 6.94E-06 | 0.03 | 9.53E-06 | 1.0 |
| rs261342 | 15 | 56518445 | *LIPC* | 0.04 | 2.93E-07 | 0.04 | 2.31E-08 | 1.0 |
| rs9989419 | 16 | 55542640 | *CETP* | 0.05 | 1.00E-17 | 0.02 | 2.14E-02 | 1.6 |
| rs12708967 | 16 | 55550712 | *CETP* | 0.07 | 5.50E-20 | 0.03 | 4.02E-05 | 1.4 |
| rs17231506 | 16 | 55552029 | *CETP* | 0.07 | 9.26E-29 | 0.04 | 5.33E-05 | 2.4 |
| rs711752 | 16 | 55553712 | *CETP* | 0.06 | 2.07E-26 | 0.03 | 6.07E-03 | 2.6 |
| rs5883 | 16 | 55564854 | *CETP* | 0.06 | 3.86E-06 | 0.10 | 3.22E-15 | 1.1 |
| rs5880 | 16 | 55572592 | *CETP* | 0.08 | 5.76E-09 | 0.03 | 1.54E-02 | 1.2 |

.

**Table 4.5 Effects of multicollinearity on regression coefficients for the triglyceride genetic score SNPs.** Regression coefficients for the AIC-selected SNPs from a univariate regression (single SNP as predictor) and multiple regression (all SNPs as predictors) with $\log_e$-transformed triglyceride levels. No adjustment for covariates was made. SNPs where multicollinearity results in either beta-coefficients in the two regression analyses to be in the opposite direction, inflated beta-coefficients in the multiple regression analysis, large variance inflation factors (>5), or where the association p-value in the multiple regression is less than 0.05, are highlighted in red.

| | | | | Univariate Regression | | Multiple Regression | | |
|---|---|---|---|---|---|---|---|---|
| SNP | Chromosome | Position (NCBI36) | Closest Gene | Beta | P-value | Beta | P-value | VIF |
| rs1260326 | 2 | 27584444 | *GCKR* | 0.06 | 1.34E-07 | 0.06 | 1.02E-06 | 1.0 |
| rs17145713 | 7 | 72542746 | *BAZ1B* | 0.08 | 3.57E-08 | 0.08 | 1.33E-08 | 1.0 |
| rs285 | 8 | 19859469 | *LPL* | 0.07 | 1.59E-09 | 0.06 | 9.20E-06 | 1.3 |
| rs331 | 8 | 19864685 | *LPL* | 0.08 | 6.62E-10 | 0.36 | 1.60E-02 | **143.6** |
| rs3289 | 8 | 19867472 | *LPL* | 0.15 | 3.46E-05 | 0.15 | 2.92E-05 | 1.1 |
| rs3916027 | 8 | 19869148 | *SLC18A1* | 0.08 | 2.58E-09 | **-0.33** | 2.54E-02 | **143.7** |
| rs10503669 | 8 | 19891970 | *SLC18A1* | 0.11 | 2.67E-09 | 0.05 | 1.38E-02 | 1.5 |
| rs17321515 | 8 | 126555590 | *TRIB1* | 0.05 | 2.22E-05 | 0.05 | 1.97E-05 | 1.0 |
| rs17108993 | 10 | 95354023 | *RBP4* | 0.13 | 5.55E-05 | 0.11 | 4.45E-04 | 1.0 |
| rs6589565 | 11 | 116145446 | *BUD13* | 0.20 | 2.32E-19 | 0.11 | 3.56E-02 | **6.1** |
| rs12286037 | 11 | 116157416 | *ZNF259* | 0.15 | 1.04E-10 | 0.21 | 1.75E-13 | 1.7 |
| rs651821 | 11 | 116167788 | *APOA5* | 0.21 | 2.00E-19 | 0.15 | 1.34E-02 | **6.9** |
| rs10750097 | 11 | 116169249 | *APOA4* | 0.09 | 3.68E-11 | **-0.05** | 1.23E-02 | 2.2 |
| rs33989105 | 11 | 116205351 | *APOC3* | 0.07 | 4.98E-07 | 0.04 | 2.09E-03 | 1.3 |
| rs5072 | 11 | 116212792 | *APOA1* | 0.11 | 2.15E-07 | **-0.03** | **2.09E-01** | 1.8 |
| rs2304128 | 19 | 19607151 | *GMIP* | 0.10 | 8.54E-07 | 0.09 | 8.56E-06 | 1.0 |

### 4.3.4 BIC-SNP Selection and Ridge Regression

Variable selection using AIC failed to remove all highly correlated SNPs. Therefore the more stringent BIC was used for model selection. For both LDL-C and triglycerides, stepwise regression using BIC retained a smaller number of genetic predictors in the model than with AIC (Table 4.6). None of the triglyceride-associated SNPs selected by the BIC had large VIFs. Though the 2 AIC-selected, LDL-C-associated SNPs with very large VIFs were retained by the model selection using BIC, applying a Ridge regression shrunk the coefficients for these 2 SNPs from 1.21 and -1.09 (from the multiple regression) to 0.23 and -0.068, respectively. For HDL-C, using the BIC selected the same number of SNPs (Table 4.6) (but not necessarily the same SNPs) as AIC.

**Table 4.6 Comparison of SNP selection by AIC and BIC.** The table compares the number of SNPs retained after variable selection using AIC or BIC. BIC selects a smaller number of LDL-C- and triglyceride-associated SNPs. Presence of SNPs with high variance inflation factors (VIF >5) indicates presence of multicollinearity. With AIC, SNP multicollinearity is an issue with the selected set of LDL (2 SNPs with VIF>5) and triglyceride SNPs (4 SNPs with VIF>5). Using BIC overcomes this problem for the triglyceride SNPs, but multicollinearity remains even after BIC-selection of LDL-C associated SNPs.

| | Number of genetic predictors selected by model | | Number of genetic predictors with VIF>5 | |
|---|---|---|---|---|
| Exposure | AIC | BIC | AIC | BIC |
| LDL | 22 | 16 | 2 | 2 |
| HDL | 12 | 12 | 0 | 0 |
| Triglyceride | 16 | 13 | 4 | 0 |

## 4.4 Discussion

### 4.4.1 Summary of Results

Two important considerations when performing MR analysis are the strength and specificity of the instrument. Based on the SNPs associated with lipid levels in WHII, assessment of the different approaches to instrument development shows that the weighted genetic score consisting of SNPs selected by a stepwise regression approach using the BIC for model selection and weighted by coefficients derived from a Ridge regression, provide the best instruments in terms of both strength and specificity for the exposure of interest. Though the multiple instruments approach was equivalent in strength to the genetic score, a genetic score instrument that combines multiple SNPs into a single IV is less likely to suffer from weak-instrument bias when the number of SNPs used is large and/or sample size is small.

### 4.4.2 Multicollinearity

Multicollinearity is often caused by the choice of model, such as when two highly correlated predictors are used in the regression model (Montgomery et al. 2012). Stepwise regression using AIC for model selection was initially adopted to select the best genetic predictors of each lipid fraction, with the assumption that redundant SNPs would be removed. However, several correlated SNPs, some of which are in perfect LD, were retained by the model. Re-specifying the model by manually removing the correlated SNPs can overcome the problem of multicollinearity. However, this would require several iterations of model specification and may not be a satisfactory solution if the SNPs dropped from the model have significant explanatory power (Montgomery et al. 2012). Presence of multicollinearity may result in poor estimates of the regression coefficients, and subsequently, inaccurate weights for the genetic score. The above issues were overcome by using a more stringent model selection criterion which excluded some of the correlated SNPs, followed by the use of a Ridge regression to obtain the weights for each SNP. The latter has the effect of shrinking the beta-coefficients and thus the contribution of redundant SNPs towards zero. An alternative to this multi-stage approach to SNP selection and weight estimation is to use a method such as lasso, which is a penalised regression technique that carries out variable selection and estimates coefficients of the selected variables, shrinking estimates of redundant variables towards zero. The lasso method eliminates the need to select an initial subset of SNPs using a pre-defined association p-value threshold, and available software that implement this method enable the efficient analysis of a very large number of SNPs characteristic of large-scale association studies.

### 4.4.3 Weak Instrument Bias

The F-statistic can be used as an indication of the degree of bias of the 2SLS estimator to the observed association between exposure and outcome. Based on previous studies, where the F-statistic is greater than 10, the bias becomes typically negligible (Pierce et al. 2010). In the case of the lipid instruments considered here, whether using genetic scores or multiple instruments, all F-statistic*s* were much greater than 10. However, if the number of SNPs used in a multiple instruments approach increases with little increase in $R^2$, this may create weak instrument bias, since the F-statistic is dependent on $R^2$ and the number of instruments. Current genetic studies usually have large enough sample sizes to avoid weak instrument bias, however, there will be a limit to the number of lipid-associated SNPs that can be used with the multiple instruments approach. On the other hand, weak instrument bias is unlikely to be an issue when a large number of SNPs are combined into a single genetic score instrument.

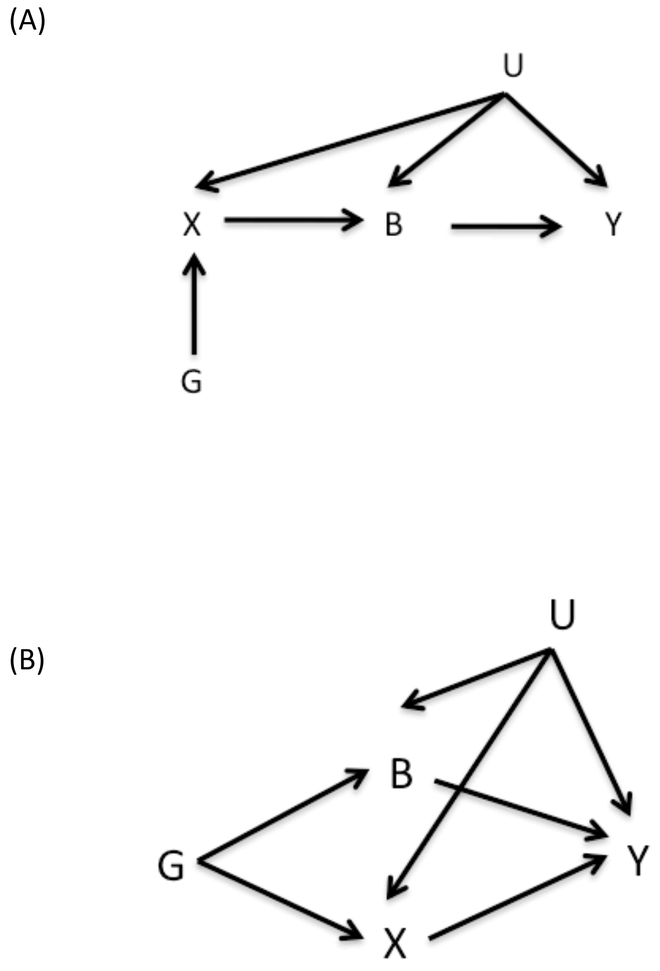
### 4.4.4 Specificity

At the individual SNP level, some SNPs may be associated with more than one lipid fraction or with other CVD risk factors. If these risk factors are downstream in the same biological pathway (Figure 4.3A) then the assumption that the instrument only affects outcome via the exposure of interest is not violated. However, this is not the case when a SNP exerts an effect on outcome via an independent pathway to the exposure of interest (Figure 4.3B). Unfortunately, the path of association between instrument and exposure of interest is not always known, making it difficult to interpret results when an instrument is associated with other risk factors. A number of SNPs were associated with both triglycerides and HDL-C. Further selection of SNPs associated only with the lipid fraction of interest would provide a more specific instrument and may help better dissect the causal pathway. Using a multiple instruments approach appears to be less specific than using a weighted genetic score derived from the same set of SNPs, which may suggest that when combining multiple SNPs into a single score, some of the pleiotropic effects may balance out.

### 4.4.5 Limitations

The major limitation of this analysis is that instrument strength was estimated in the same sample used to identify the SNPs and estimate effect sizes. Instrument strength reported here will therefore be inflated due to discovery bias. If MR analysis is also carried out in the same dataset, there is likely to be bias due to “winner’s curse”. Recent work by Burgess and Thompson (2013) showed that the use of variants and weights chosen based on the strength of their association with the risk factor in the data under analysis gives biased causal estimates in the direction of the confounded association. The genetic instruments developed in one dataset should therefore be applied in an MR analysis in an independent dataset.

**Figure 4.3 Association of a variant in a causal pathway.** *X* denotes the exposure of interest, *B* an additional risk factor, *Y* the outcome of interest and *G* the genetic instrument. (A) Though a genetic variant determining the exposure of interest may also be associated with other downstream risk factors, it is still a valid instrument. (B) Where the instrument is associated independently with both the exposure of interest and other risk factors it becomes invalid.



When the only source of information on weights is the data under analysis, an alternative approach recommended by Burgess and Thompson is to use a ten-foldcross-validation approach whereby weights are calculated based on 90% of the data and the ten sets of weights are applied in the (10%) validation data (Burgess & Thompson 2013). In this way the correlation between the weights and the data under analysis is removed. However, large enough sample sizes would be required to allow sufficient numbers in the validation data.

The analysis carried out in this chapter assumes only additive effects of alleles for each SNP and no gene-gene interactions. Though not explored in this work, if these assumptions did not hold, it would be possible to incorporate such knowledge into the model. However, analysis by Burgess & Thompson (2013) showed that an unweighted allele score is robust to model mis-specification.

Since this analysis, the GWAS meta-analysis by the GLGC in over 100,000 individuals (Teslovich et al. 2010) and the more recent gene-centric meta-analysis (using the Illumina Cardiochip) in around 66,000 individuals (Asselbergs et al. 2012) have both identified a large number of novel loci associated with each of the three lipid fractions. The SNPs and regression coefficients reported by these studies could be used to generate unbiased weighted lipid genetic scores for MR analysis. Using more than one genetic score generated from SNPs and weights identified and estimated from independent datasets would provide evidence for the robustness of the IV estimate, as independently derived genetic scores are unlikely to be influenced by the same pleiotropy or linkage disequilibrium-induced confounding.

Since the study by Yang et al (2010), which showed that when all SNPs on the GWA platform are considered they explain a much larger proportion of the phenotypic variation than those chosen by a p-value threshold, there has been a lot of interest in the use of all SNPs on the genotyping platform for MR instrument development. However, the issue of specificity and weak instrument bias would need to be carefully considered when exploring such an approach for MR analysis.

# 5 Determining the Causal Relationship between Blood Lipids and Carotid-Intima Media Thickness: a Mendelian Randomisation Analysis

## 5.1 Introduction

Higher LDL-C concentration is associated with a higher risk of CHD, and the relationship is considered causal because randomised trials using LDL-C-lowering interventions such as statins have shown to reduce CHD risk in proportion to the degree of LDL-C reduction (Baigent et al. 2005; Baigent et al. 2010). Interventions to elevate HDL-C or reduce triglycerides might also confer incremental protection against CHD, but thus far randomised trials of drugs directed at these two lipid fractions have been unable to confirm or refute such effects (Cannon et al. 2010; Forrest et al. 2008; Ginsberg et al. 2010; Jun et al. 2010; NHLBI Communications 2011).

Conclusive demonstration of the benefit and safety of new lipid-modifying interventions requires evaluation in large, expensive randomised trials with hard clinical end points in people already receiving effective drugs for CHD prevention. Approaches that help validate treatment targets ahead of such trials may help reduce the risk of late-stage failures in drug development. One approach has been to use a non-invasive measure of atherosclerosis, CIMT, as a surrogate end-point. CIMT is considered to be a subclinical measure of atherosclerosis, which is strongly associated with risk of CHD (O'Leary et al. 1999; Chambless et al. 1997). Statin drugs that are effective in reducing CHD also reduced progression of CIMT in proportion to the degree of LDL-C-lowering (Bedi et al. 2010; Espeland et al. 2005; Kastelein et al. 2003). However, interventions developed so far that reduce triglycerides or raise HDL-C have shown inconsistent effects on CIMT (Bots et al. 2007; Hiukka et al. 2008; Taylor et al. 2009), making it uncertain whether the specific agents are ineffective for CHD prevention, whether these two lipid fractions in general are not causally related to CHD and therefore invalid targets, or whether CIMT is an inadequate marker of HDL-C or triglyceride-mediated effects on CHD risk. MR, described in section 1.6.2, provides a means of evaluating and quantifying the extent to which associations between a putative risk factor (e.g. HDL-C or triglycerides) and an outcome, such as CHD or CIMT, are causal (Ebrahim & Smith, 2008).

Using large-scale genotyping arrays, many SNPs influencing LDL-C, HDL-C and triglycerides have recently been identified (Talmud et al. 2009; Teslovich et al. 2010; Asselbergs et al. 2012), and these provide potential instruments for MR analyses. The aim of the work in this chapter was to determine the casual relationship between these three lipid fractions and common CIMT using MR analysis. The LDL-C, HDL-C and triglyceride weighted genetic scores developed in the previous chapter, based on lipid-associated SNPs on the Cardiochip and weights estimated in WHII, were used as instruments for the three lipid fractions to estimate their causal association with common CIMT in around 3000 participants from the WHIII study (Marmot & Brunner 2005) and also in around 3400 individuals from the IMT Progression as Predictors of Vascular Events in a High Risk European Population (IMPROVE) study (Baldassarre et al. 2010) using the 2SLS approach for instrumental variable analysis (described in section 1.6.2.3). In addition, the causal association was also determined in both studies using weighted genetic scores based on the independently identified lipid-associated SNPs and effect sizes reported by the GLGC GWAS meta-analysis in over 100,000 individuals (Teslovich et al. 2010).

## 5.2 Materials & Methods

### 5.2.1 Study Cohorts

The WHII cohort has been described in section 2.2.1.3. The IMPROVE longitudinal study (Baldassarre et al. 2010) recruited a total of 3711 individuals (48% men) between March 2004 and April 2005 from 7 centres in 5 European countries, with a median age of 64.4 years. Eligibility criteria included age between 55 to 79 years, presence of at least three vascular risk factors, and absence of symptoms of cardiovascular diseases and any conditions that might limit longevity or visualisation of the carotid intima. The study was designed in accordance with the rules of Good Clinical Practice, and with the ethical principles established in the Declaration of Helsinki. Informed consent was obtained from all participants. Baseline data from the IMPROVE data was used in analysis.

### 5.2.2 CIMT Measurement

In WHII, ultrasound vascular measurements were taken in the 2003-2004 follow-up phase. Participants were examined in a supine position, with the head turned to a 45 degree angle away from the side to be scanned. The far walls of the left and right common carotid artery were visualised in the lateral projection. The common CIMT was measured at its thickest part, 1 cm proximal to the bifurcation. A measurement was taken between the leading edge of the intima and the media adventitia on three separate images on each side using electronic callipers, and the mean of the six measures was used for analysis. The overall coefficient of variation (defined as the ratio of the standard deviation to the mean) for repeated measures of CIMT was 4.7% (N = 89), indicating high reproducibility (Kivimäki et al. 2008). For all analyses, the CIMT variable was $\log_e$-transformed.

In IMPROVE, the far walls of the left and right common carotid artery were visualised in the lateral projection and recorded on sVHS videotapes. Measurements were taken at the thickest part of common carotids, 1 cm proximal to the bifurcation. The far walls of the common carotids in their entire length were measured in at least three different images on each side using dedicated software able to automatically recognise the leading edge of the intima and the media adventitia. For each segment the mean of the six measures was used for the analysis. The overall coefficient of variation for repeated measures of CIMT was 3.9% (N = 121) (Baldassarre et al. 2010). In both studies, to obtain a normal distribution, the CIMT variable was $\log_e$-transformed.

### 5.2.3 Lipid Measurements

For WHII, baseline lipid measurements were used for the analysis, since very few individuals were on lipid-lowering therapy at baseline (previously described in section 3.2.2)**.** In the IMPROVE study, blood sampling for laboratory tests was performed after an overnight fast. Serum was frozen at –80°C prior to shipment for centralised biochemical analyses and biobanking in Stockholm (Karolinska Institute Stockholm, Sweden). Serum concentrations of total cholesterol, HDL-C and triglycerides were analysed in a centralised laboratory. LDL-C concentration was calculated using the Friedewald (Friedewald et al. 1972). HDL-C and triglyceride variables were $\log_e$-transformed for all analyses.

### 5.2.4 Genotyping

As described in section 2.2.4, DNA was extracted from whole blood samples from WHII participants between 2003 and 2004, and genotyping on a total of 5592 samples using the Illumina Cardiochip (Keating et al. 2008) was completed in 2008.

At the beginning of 2012, genotyping of 3413 WHII samples was also completed using the Illumina Metabochip platform (Voight et al. 2012). In the IMPROVE study, 3695 samples were genotyped using the Metabochip. The Metabochip is a custom genotyping array that provides cost-effective genotyping of nearly 200,000 SNPs chosen based on GWAS results from meta-analyses of 23 traits that include cardiovascular disease outcomes (CAD, type 2 diabetes (T2D) and MI), and risk factors (fasting glucose, fasting insulin, 2-hour glucose, glycated haemoglobin (HbA1c), T2D age of diagnosis, LDL-C, HDL-C triglycerides, total cholesterol, systolic and diastolic blood pressure, QT interval, BMI, waist-to-hip ratio adjusted for BMI, waist circumference adjusted for BMI, height, body fat percentage, platelet count, mean platelet volume, and white blood cell count).

After applying quality control steps (refer to section 2.2.4) (filtering for duplicates, cryptic relatedness, ambiguous gender, self-reported non-Caucasians, outliers based on the genome-wide identity-by-state analysis implemented in PLINK, sample call rate <80% and SNP call rate <98%), 5059 Cardiochip genotyped samples and 3126 Metabochip genotyped samples from WHII were available for the analysis. In the IMPROVE study, after quality control (SNP and sample call rate <95%, removing individuals for relatedness (confirmed or cryptic), reported non-European descent, outliers identified by multi-dimensional scaling, and mismatch between recorded and genotype-determined sex), 3430 individuals remained for analysis.

### 5.2.5 Generating Lipid Genetic Score Instruments

Two lipid genetic scores were derived: one based on an internal analysis in the WHII study using SNPs present on the Cardiochip and one based on lipid-associated variants reported by the GLGC (Teslovich et al. 2010), as described below.

#### 5.2.5.1 Cardiochip Genetic Scores

The SNPs and weights used for calculating the Cardiochip genetic scores for LDL-C, HDL-C and triglycerides in both WHII and IMPROVE were derived using variable selection with BIC and Ridge regression in WHII (Table 5.1 - Table 5.3), as described in the previous chapter (section 4.3.7). SNPs not present in the IMPROVE data (because they were not represented on the Metabochip genotyping platform or failed quality control) were excluded from the genetic score calculations. In both studies, individuals with any missing genotypes for the SNPs used in the score calculation were excluded from the analysis.

#### 5.2.5.2 GLGC Genetic Scores

In the published GLGC meta-analysis, an association p-value $<5\times10^{-8}$ was used to denote significant association between SNPs and lipid traits (Teslovich et al. 2010). For the purpose of the genetic score calculation only the lead SNP from each locus was selected, and if a SNP was associated with more than one lipid fraction it was only used in the genetic score calculation for the trait with which it had the most significant association p-value (primary trait association). These are the SNPs listed in the primary tables of the GLGC publication (Teslovich et al. 2010). Since the design of the Metabochip was based on GWAS results, a large number of the lipid-associated SNPs reported by the GLGC meta-analysis were present on the Metabochip platform. Risk allele counts were therefore calculated in WHII and IMPROVE using the Metabochip genotype data. Risk allele counts were weighted using the univariate beta-coefficients reported by the GLGC discovery meta-analysis (Table 5.1 - Table 5.3). The GLGC reports a strong association of rs4420638 on chromosome 19 with LDL-C levels, but this SNP is not independent of the *APOE* SNP rs429358. Therefore, as was done for the Cardiochip scores, the weighted *APOE* genotypes were used in the calculation of the LDL-C genetic score instead of the

GLGC-reported SNP. Since the discovery of the GLGC SNPs was carried out in an independent dataset and only a single SNP was selected at each locus, the issues of discovery bias and multicollinearity due to LD were minimised. SNPs not present in the data (because they were not represented on the genotyping platform or failed quality control) were excluded from the genetic score calculations. Individuals with missing data for SNPs used in the score calculation were excluded.

**Table 5.1 SNPs contributing to the LDL-C genetic scores.**

| **SNP** | **Gene** | **In Cardiochip Score** | **In GLGC Score** | **Risk Allele** | **Non-risk Allele** | **Cardiochip Weight** | **GLGC weight** |
|---|---|---|---|---|---|---|---|
| rs4299376 | *ABCG8* | **yes** | **yes** | G | T | 0.12 | 0.071 |
| rs2072560 | *APOA5* | **yes** | no | T | C | 0.21 | - |
| rs1367117 | *APOB* | no | **yes** | A | G | - | 0.1 |
| rs562338 | *APOB* | **yes** | no | G | A | 0.14 | - |
| rs934197 | *APOB* | **yes** | no | A | G | 0.094 | - |
| rs12721109 | *APOC4* | **yes** | no | G | A | 0.26 | - |
| rs10402271 | *BCAM/PVRL2* | **yes** | no | G | T | 0.062 | - |
| rs12740374 | *CELSR2* | **yes** | no | G | T | 0.23 | - |
| rs629301 | *CELSR2* | **yes** | **yes** | T | G | -0.068 | 0.15 |
| rs17231506 | *CETP* | **yes** | no | C | T | 0.1 | - |
| rs1800562 | *HFE* | no | **yes** | G | A | - | 0.057 |
| rs12916 | *HMGCR* | **yes** | no | C | T | 0.11 | - |
| rs8017377 | *KIAA1305* | no | **yes** | A | G | - | 0.029 |
| rs6511720 | *LDLR* | no | **yes** | G | T | - | 0.18 |
| rs17248720 | *LDLR* | **yes** | no | C | T | 0.43 | - |
| rs2228671 | *LDLR* | **yes** | no | C | T | -0.21 | - |
| rs8110695 | *LDLR* | **yes** | no | T | A | 0.066 | - |
| rs3757354 | *MYLIP* | no | **yes** | C | T | - | 0.037 |
| rs2479409 | *PCSK9* | no | **yes** | G | A | - | 0.052 |
| rs11591147 | *PCSK9* | **yes** | no | G | T | 0.52 | - |
| rs283813 | *PVRL2* | **yes** | no | T | A | 0.14 | - |
| rs1564348 | *SLC22A1* | no | **yes** | T | C | - | 0.014 |
| rs11220462 | *ST3GAL4* | no | **yes** | A | G | - | 0.05 |
| Total | | 15 | 10 | | | | |

**Table 5.2 SNPs contributing to HDL-C genetic scores**

| SNP | Gene | In Cardiochip Score | In GLGC Score | Risk Allele | Non-risk Allele | Cardiochip Weight | GLGC weight |
|---|---|---|---|---|---|---|---|
| rs1883025 | *ABCA1* | no | **yes** | T | C | - | 0.024 |
| rs4148008 | *ABCA8* | no | **yes** | G | C | - | 0.011 |
| rs2923084 | *AMPD3* | no | **yes** | G | A | - | 0.011 |
| rs2072560 | *APOA5* | **yes** | no | T | C | 0.064 | - |
| rs6450176 | *ARL15* | no | **yes** | A | G | - | 0.013 |
| rs11820589 | *BUD13* | **yes** | no | A | G | 0.054 | - |
| rs2814944 | *C6orf106* | no | **yes** | A | G | - | 0.013 |
| rs581080 | *C9orf52* | no | **yes** | G | T | - | 0.017 |
| rs3764261 | *CETP* | no | **yes** | C | A | - | 0.088 |
| rs12708967 | *CETP* | **yes** | no | C | T | 0.031 | - |
| rs17231506 | *CETP* | **yes** | no | C | T | 0.041 | - |
| rs5880 | *CETP* | **yes** | no | C | G | 0.04 | - |
| rs5883 | *CETP* | **yes** | no | C | T | 0.091 | - |
| rs711752 | *CETP* | **yes** | no | G | A | 0.023 | - |
| rs9989419 | *CETP* | **yes** | no | A | G | 0.016 | - |
| rs2925979 | *CMIP* | no | **yes** | T | C | - | 0.012 |
| rs737337 | *DOCK6* | no | **yes** | C | T | - | 0.017 |
| rs3136441 | *F2* | no | **yes** | T | C | - | 0.02 |
| rs4846914 | *GALNT2* | no | **yes** | G | A | - | 0.016 |
| rs1800961 | *HNF4A* | no | **yes** | T | C | - | 0.049 |
| rs4731702 | *KLF14* | no | **yes** | C | T | - | 0.015 |
| rs2652834 | *LACTB* | no | **yes** | A | G | - | 0.01 |
| rs386000 | *LILRA3* | no | **yes** | G | C | - | 0.021 |
| rs1532085 | *LIPC* | no | **yes** | G | A | - | 0.037 |
| rs261342 | *LIPC* | **yes** | no | C | G | 0.036 | - |
| rs4775041 | *LIPC* | **yes** | no | G | C | 0.028 | - |
| rs17410962 | *LPL* | **yes** | no | G | A | 0.027 | - |
| rs301 | *LPL* | **yes** | no | T | C | 0.027 | - |
| rs12967135 | *MC4R* | no | **yes** | A | G | - | 0.011 |
| rs4660293 | *PABPC4* | no | **yes** | G | A | - | 0.012 |
| rs7134375 | *PDE3A* | no | **yes** | C | A | - | 0.01 |
| rs4129767 | *PGS1* | no | **yes** | G | T | - | 0.01 |
| rs6065906 | *PLTP* | no | **yes** | C | T | - | 0.024 |
| rs9987289 | *PPP1R3B* | no | **yes** | A | G | - | 0.031 |
| rs16942887 | *PSKH1* | no | **yes** | G | A | - | 0.033 |
| rs838880 | *SCARB1* | no | **yes** | T | C | - | 0.016 |
| rs13107325 | *SLC39A8* | no | **yes** | T | C | - | 0.022 |
| rs11869286 | *STARD3* | no | **yes** | G | C | - | 0.012 |
| rs2293889 | *TRPS1* | no | **yes** | T | G | - | 0.011 |
| rs181362 | *UBE2L3* | no | **yes** | T | C | - | 0.012 |
| rs1689800 | *ZNF648* | no | **yes** | G | A | - | 0.012 |
| Total |  | 12 | 29 |  |  |  |  |

**Table 5.3 SNPs contributing to Triglyceride genetic scores**

| SNP | Gene | Present in Cardiochip Score | Present in GLGC Score | Risk Allele | Non-risk Allele | Cardiochip Weight | GLGC weight |
|---|---|---|---|---|---|---|---|
| rs442177 | *AFF1* | no | **yes** | T | G | - | 0.025 |
| rs10750097 | *APOA5* | **yes** | no | G | A | 0.042 | - |
| rs651821 | *APOA5* | **yes** | no | C | T | 0.22 | - |
| rs33989105 | *APOC3* | **yes** | no | T | C | 0.038 | - |
| rs17145713 | *BAZ1B* | **yes** | no | C | T | 0.085 | - |
| rs10195252 | *COBLL1* | no | **yes** | T | C |  | 0.023 |
| rs2068888 | *CyP26A1* | no | **yes** | G | A | - | 0.026 |
| rs2131925 | *DOCK7* | no | **yes** | T | G | - | 0.058 |
| rs174546 | *FADS1* | no | **yes** | T | C | - | 0.043 |
| rs2412710 | *GANC/CAPN3* | no | **yes** | A | G | - | 0.079 |
| rs1260326 | *GCKR* | **yes** | **yes** | T | C | 0.05 | 0.099 |
| rs2304128 | *GMIP* | **yes** | no | G | T | 0.086 | - |
| rs17108993 | *GPR120* | **yes** | no | G | C | 0.11 | - |
| rs12678919 | *LPL* | no | **yes** | A | G | - | 0.15 |
| rs10503669 | *LPL* | **yes** | no | C | A | 0.047 | - |
| rs285 | *LPL* | **yes** | no | C | T | 0.051 | - |
| rs3289 | *LPL* | **Yes** | no | C | T | 0.15 | - |
| rs331 | *LPL* | **Yes** | no | G | A | 0.032 | - |
| rs9686661 | *MAP3K1* | No | **yes** | T | C | - | 0.029 |
| rs645040 | *MSL2L1* | No | **yes** | T | G | - | 0.025 |
| rs11776767 | *PINX1* | No | **yes** | C | G | - | 0.023 |
| rs5756931 | *PLA2G6* | No | **yes** | T | C | - | 0.017 |
| rs11613352 | *R3HDM2* | No | **yes** | C | T | - | 0.03 |
| rs17145738 | *TBL2* | No | **yes** | C | T | - | 0.11 |
| rs2954029 | *TRIB1* | No | **yes** | A | T | - | 0.064 |
| rs17321515 | *TRIB1* | **Yes** | no | A | G | 0.048 | - |
| rs13238203 | *TYW1B* | No | **yes** | C | T | - | 0.089 |
| rs12286037 | *ZNF259* | **Yes** | no | T | C | 0.18 | - |
| Total |  | 13 | 16 |  |  |  |  |

## 5.2.6 Association of Genetic Scores with Lipid Levels

Linear regression was used to evaluate the association of lipid levels with their respective genetic scores, without any adjustment for covariates. For comparison of effect sizes across the different lipid traits, regression analysis was also performed on standardised variables (Z-scores). The proportion of variance explained ($R^2$) and the F-statistic derived from the regression were reported as measures of the strength of each genetic score as an instrument. The $R^2$ values from the regression of each genetic score with the non-indexed lipid fractions were also reported as an indication of instrument specificity.

### 5.2.7 Observed Association between Lipids and CIMT

Association of CIMT with lipid levels was determined using linear regression, with and without adjustment for sex, age, smoking (current status), diabetes status and statin use. For comparison of effect sizes across traits, regression analysis was also performed on standardised variables.

### 5.2.8 Direct Association between Lipid Genetic Scores and CIMT

Direct association between the genetic scores and CIMT using linear regression was carried out unadjusted for any covariates, and adjusted for the first three dimensions from multidimensional scaling (refer to section 2.2.4). To ensure that any association with CIMT was only through the effect of the genetic scores on the relevant lipid fraction, analysis was also repeated adjusted for the other two lipid fractions.

### 5.2.9 Causal Effect Estimate using 2SLS

To estimate the causal effect of each lipid fraction with CIMT, instrumental variable analysis using 2SLS was carried out without any adjustment for covariates, using the *ivreg()* command from the AER package (Kleiber & Zeileis 2010) in R CRAN (R Development Core Team 2012). A meta-analysis of the effect estimates from the two studies was also carried out using a fixed effect model, where the summary estimate was the weighted (by the inverse of the study variance) mean of the study-specific effects. The analysis was then repeated using lipid levels corrected for statin use by applying multiplicative correction factors derived from an analysis of repeatedly measured lipid levels in WHII, including levels measured before and after lipid-lowering treatment. This method was reported in a recent large-scale genetic meta-analysis (Asselbergs et al. 2012). For statin users, the recorded lipid values were multiplied by a constant: LDL-C by 1.352; HDL-C by 0.949, and triglycerides by 1.210 (Asselbergs et al. 2012). For comparison of effect sizes across lipid traits, regression analysis was also performed on standardised variables.

## 5.3 Results

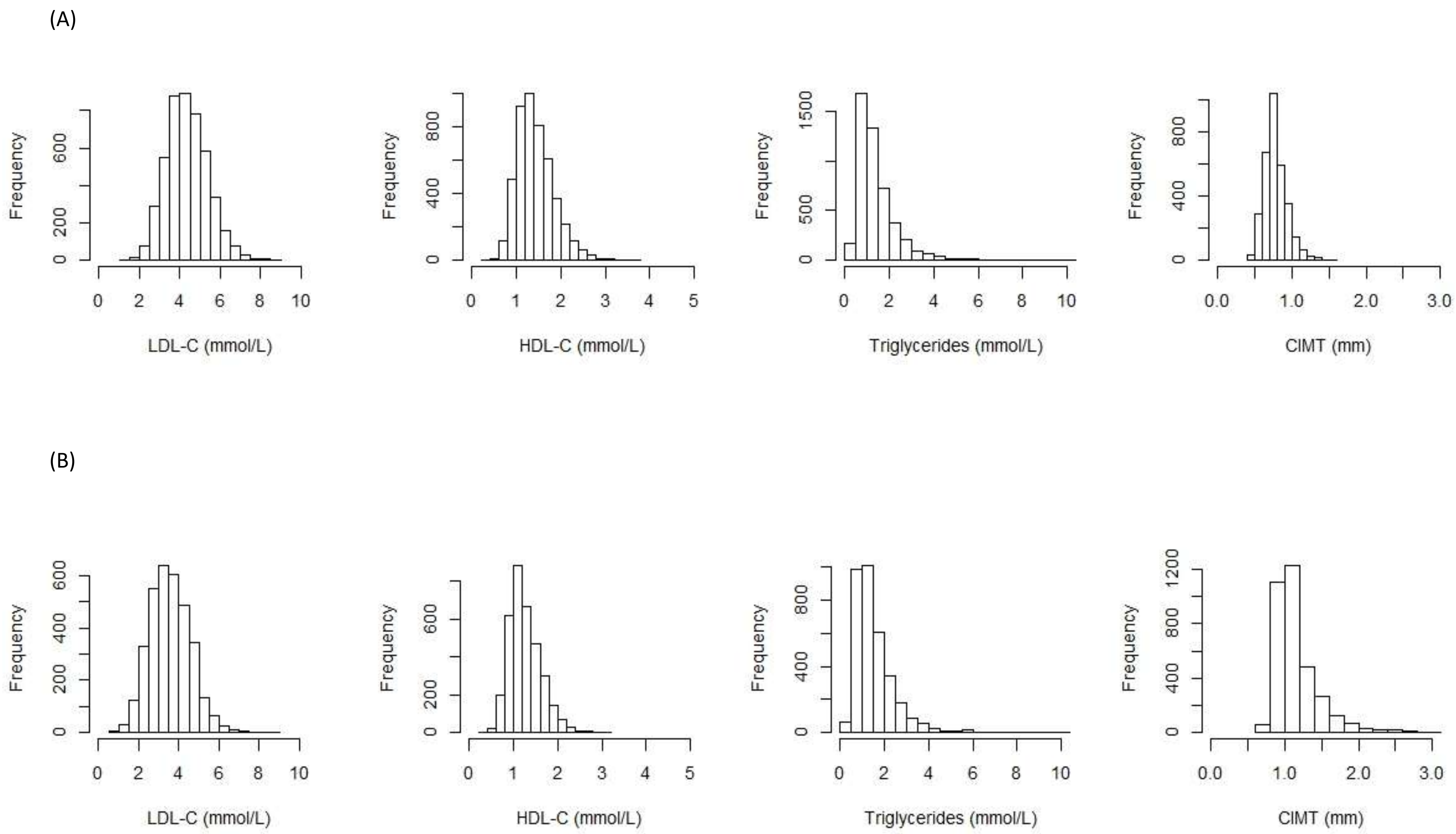
### 5.3.1 Study Characteristics

Population characteristics and sample sizes with both genotype and phenotype data are shown in Table 5.4. The mean age of WHII participants at the follow-up phase when CIMT measurements were taken was 60.9 (SD = 6.0) years, similar to the mean age of 64.2 (SD = 5.4) years for the IMPROVE participants. Mean CIMT in WHII and IMPROVE was 0.8 mm (SD = 0.2) and 1.2 mm (SD = 0.3), respectively. The mean baseline LDL-C level in WHII was 4.4 mmol/L (SD = 1.0). The lower mean LDL-C level in IMPROVE (3.6 mmol/L; SD=1.0) may partly be explained by the larger proportion of participants on statin medication at the time of lipid measurement (0.9% in WHII versus 40% in IMPROVE). The distributions and range of lipid values in the two studies were comparable (Figure 5.1). The range of CIMT values in IMPROVE was larger than in WHII, with 101 individuals having CIMT > 2mm in IMPROVE but none in WHII (Figure 5.1).

**Table 5.4 Cohort characteristics.**

| | **Whitehall II** | **IMPROVE** |
|---|---|---|
| N | 5059 | 3430 |
| % Men | 74 | 48 |
| Age | | |
| Baseline | 49.1 (5.9) | 64.2 (5.4) |
| Follow-up | 60.9 (6.0) | - |
| Mean CIMT (SD) mm | 0.8 (0.2) | 1.2 (0.3) |
| Baseline Mean LDL-C (SD) mmol/L | 4.4 (1.0) | 3.6 (1.0) |
| Baseline Mean HDL-C (SD) mmol/L | 1.4 (0.4) | 1.3 (0.4) |
| Baseline Mean Triglyceride (SD) mmol/L | 1.4 (1.1) | 1.6 (1.2) |
| Baseline % on Statins | 0.9 | 40.3 |
| | | |
| **Number of participants with:** | | |
| CIMT measurement | 3617 | 3430 |
| Cardiochip data | 5059 | 0 |
| Cardiochip data and CIMT | 3256 | 0 |
| Metabochip data | 3126 | 3430 |
| Metabochip data and CIMT | 2138 | 3430 |

**Figure 5.1 Distributions of lipids and CIMT** in (A) WHII and (B) IMPROVE.



### 5.3.2 Cardiochip Lipid Genetic Scores

Seventeen SNPs (including the 2 *APOE* SNPs) contributed to the Cardiochip LDL-C genetic score (Table 5.1), and 12 and 13 SNPs, respectively, to the HDL-C (Table 5.2) and triglyceride (Table 5.3) genetic scores. After applying quality control filters, all SNPs for each lipid score were available in the WHII dataset. In the IMPROVE dataset 13 of 17 LDL-C SNPs (including 2 *APOE* SNPs), 11 of 12 HDL-C SNPs, and 9 of 13 triglyceride SNPs were available for the score calculation.

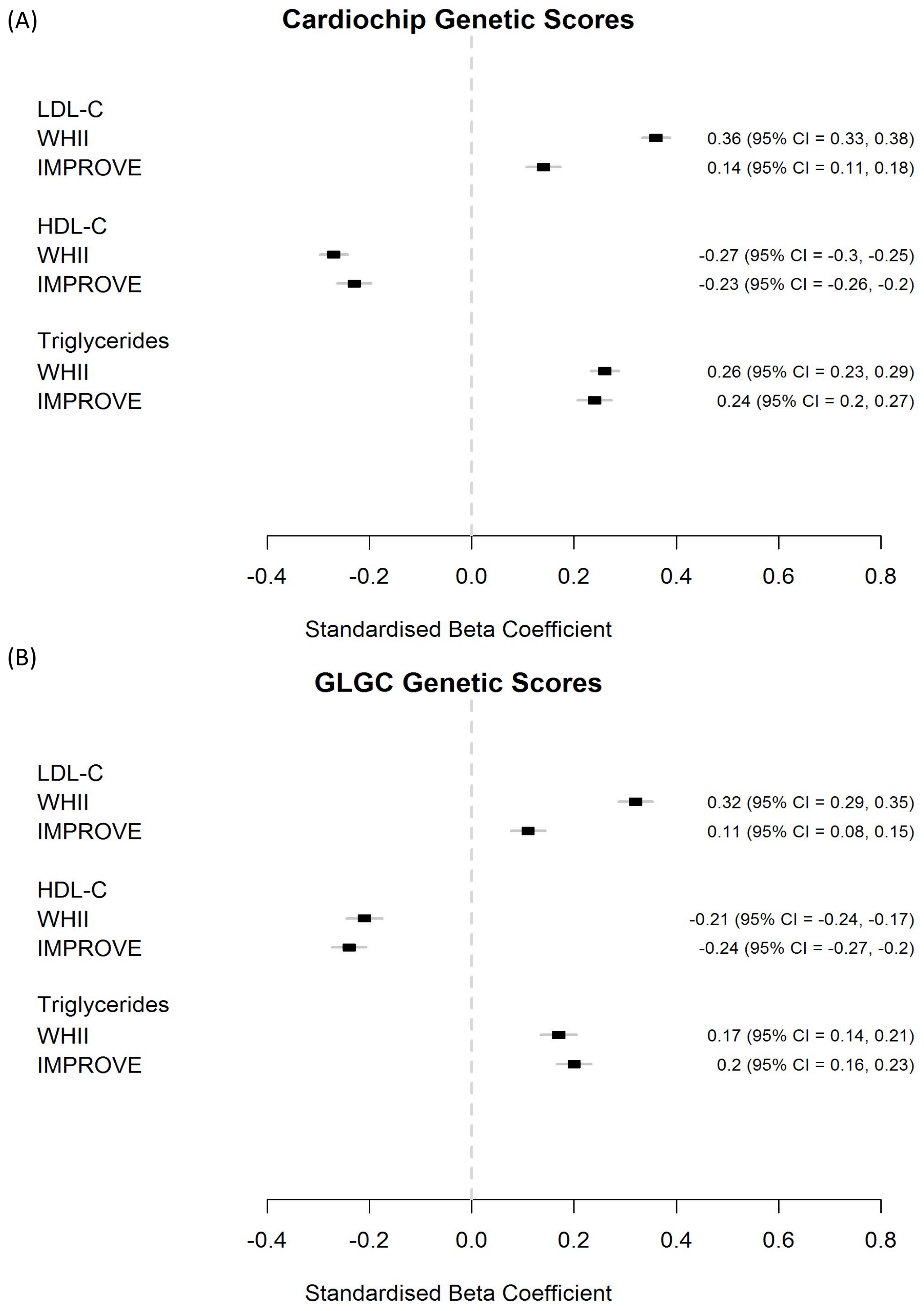
### 5.3.3 GLGC Lipid Genetic Scores

Of the lead SNPs reported by the GLGC meta-analysis, 12 (including the 2 *APOE* SNPs), 29 and 16 SNPs associated with LDL-C, HDL-C and triglycerides, respectively, were present on the Metabochip (Table 5.1 - Table 5.3) and contributed to the genetic scores. In WHII, all LDL-C SNPs, 28 of 29 HDL-C SNPs and all triglyceride SNPs were present. In IMPROVE, 10 of 12 LDL-C SNPs (including the 2 *APOE* SNPs), 28 of 29 HDL-C SNPs, and all triglyceride SNPs were available for score calculation.

### 5.3.4 Association of Lipid Levels with Lipid Genetic Scores

A 1 SD higher Cardiochip LDL-C genetic score was associated with 0.37 mmol/L (95% CI = 0.34 – 0.39) and 0.16 mmol/L (95% CI = 0.13 – 0.20) higher LDL-C in WHII and IMPROVE, respectively. A 1 SD higher HDL-C genetic score was associated with 8% (beta (95% CI) = -0.08 (-0.09, -0.07)) and 6% (beta (95% CI) = -0.06 (-0.07, -0.05)) lower HDL-C in WHII and IMPROVE, respectively. A 1 SD higher triglyceride genetic score was associated with 14% (beta (95% CI) = 0.14 (0.13, 0.16)) and 13% (beta (95% CI) = 0.13 (0.11, 0.14)) higher triglycerides in WHII and IMPROVE, respectively. For comparison of effect sizes, Figure 5.2 shows the standardised beta-coefficients which represent the SD change in lipids per 1 SD change in

**Figure 5.2 Association of lipid genetic scores with lipid levels in WHII and IMPROVE.** Effect sizes are shown for standardised variables: standard deviation change in LDL-C, $\log_e$-transformed HDL-C and $\log_e$-transformed triglycerides per 1 SD change in the respective (A) Cardiochip genetic score and (B) GLGC genetic score.



genetic score. Differences in lipid levels associated with the GLGC genetic scores were in the same direction but were slightly lower in magnitude in both studies (Figure 5.2B)

### 5.3.5 Lipid Genetic Score Instrument Strength

The Cardiochip genetic scores explained 13% and 3% of the total variance in LDL-C, 7% and 5% of the variance in HDL-C, and 7% and 4% of the variance in triglycerides in WHII and IMPROVE, respectively. The GLGC genetic scores explained 11% and 2% of the total variance in LDL-C, 4% and 5% of the variation in $\log_e$(HDL-C), and 2% and 2% of the variation in $\log_e$(triglycerides) in WHII and IMPROVE, respectively (Table 5.5). All genetic scores had very large F-statistics (F > 70) (Table 5.5).

**Table 5.5 Strength of genetic instruments.** $R^2$ and F-statistic obtained from the first stage regression between lipid levels and the respective genetic scores

| | $R^2$ | | F-statistic | | Sample Size | |
|---|---|---|---|---|---|---|
| Genetic Scores | WHII | IMPROVE | WHII | IMPROVE | WHII | IMPROVE |
| Cardiochip | | | | | | |
| LDL-C | 0.13 | 0.03 | 697 | 90 | 4635 | 3354 |
| HDL-C | 0.07 | 0.05 | 371 | 181 | 4745 | 3410 |
| Triglycerides | 0.07 | 0.04 | 259 | 137 | 4760 | 3414 |
| | | | | | | |
| GLGC | | | | | | |
| LDL-C | 0.11 | 0.02 | 366 | 75 | 3005 | 3352 |
| HDL-C | 0.04 | 0.05 | 143 | 194 | 3052 | 3342 |
| Triglycerides | 0.02 | 0.02 | 76 | 78 | 3062 | 3410 |

In WHII, $R^2$ and F-statistic*s* for the Cardiochip scores were much higher than those in IMPROVE due to discovery bias and larger sample size in WHII. The considerably lower $R^2$ values for the LDL-C genetic score in IMPROVE also reflects the large number of individuals on statins. For individuals not on statin medication, the Cardiochip and GLGC LDL-C genetic explained 7.5% and 6%, respectively, of the total variation in LDL-C in IMPROVE. The $R^2$ values from the association of each genetic score with all three lipid fractions are shown in Figure 5.3 and Figure 5.4. Though there is association between the triglyceride score and HDL-C, and the HDL-C score and triglyceride levels, the scores are much stronger instruments for the lipid fraction in question.

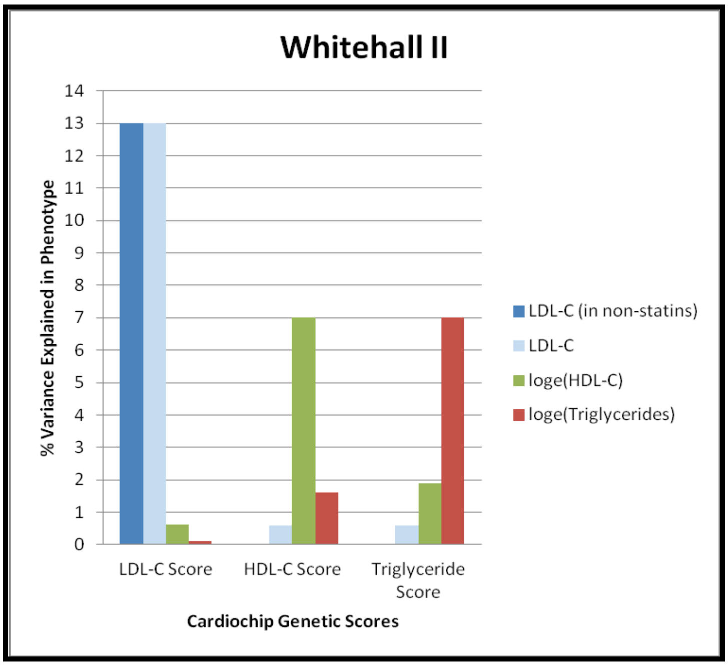
### 5.3.6 Association of Lipid Fractions and CIMT

After adjustment for age, sex, smoking, diabetes mellitus status and statin use, only LDL-C and HDL-C were associated with CIMT in both studies. A 1 mmol/L higher LDL-C was associated with 0.01 mm (95% CI = 0.006 – 0.02) and 0.02 mm (95% CI = 0.005 – 0.03) higher CIMT in WHII and IMPROVE, respectively. Associations of standardised measures, unadjusted and adjusted for covariates are shown in Table 5.6.

### 5.3.7 Direct Association of Lipid Genetic Scores and CIMT

Only the Cardiochip and GLGC LDL-C genetic scores were significantly associated with CIMT in both studies, and this remained the case after adjustment for other lipid fractions and the first three dimensions from MDS (Figure 5.5).

**Figure 5.3 Specificity of the Cardiochip genetic scores for the respective lipid fractions.** For LDL-C, regression was carried out in all individuals and separatly in indviduals not on statin medication (N = 4884 and 2049 in WHII and IMPROVE, respectively).



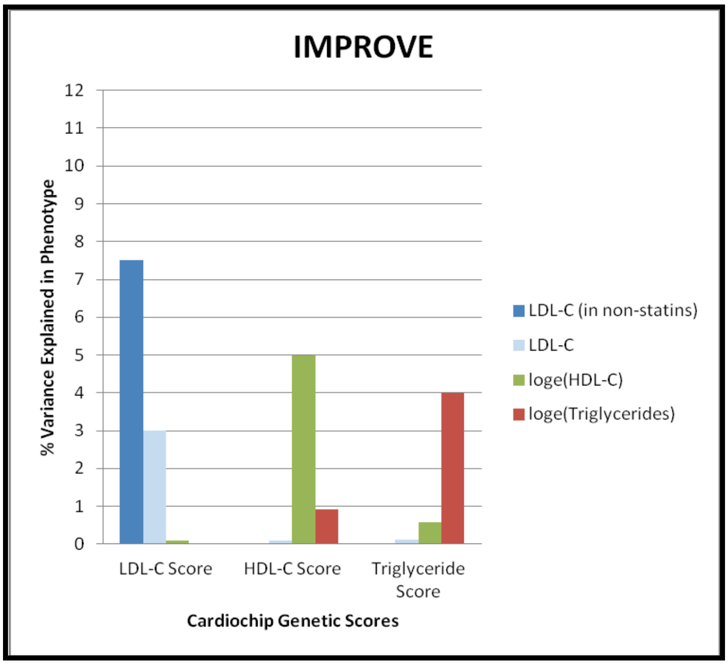


**Figure 5.4 Specificity of the GLGC genetic scores for the respective lipid fractions.** For LDL-C, regression was carried out in all individuals and separatly in indviduals not on statin medication (N = 3047 and 2049 in WHII and IMPROVE, respectively).

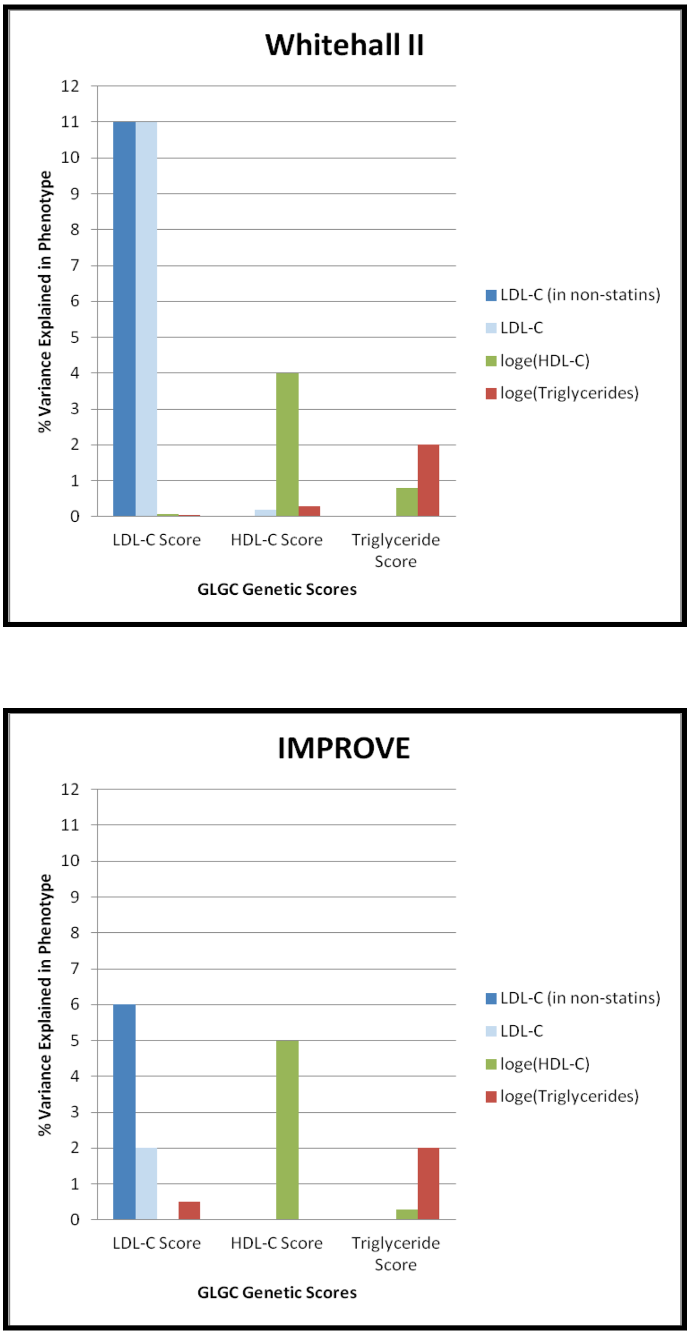


**Table 5.6 Associations of the lipid fractions with CIMT in the WHII and IMPROVE studies.** Effect sizes are shown as SD change in $\log_e$-transformed CIMT per 1 SD change in LDL-C, $\log_e$-transformed HDL-C and $\log_e$-transformed triglycerides. Association is shown for the unadjusted analysis and adjusted for sex, age, smoking, diabetes status and statin use.

| | | **Unadjusted** | | **Adjusted for sex, age, smoking, diabetes status and statin use** | |
|---|---|---|---|---|---|
| **Lipid Phenotype** | **Study** | **Beta (95% CI)** | **P** | **Beta (95% CI)** | **P** |
| Standardised LDL-C | WHII | 0.14 (0.10 – 0.17) | $9\text{x}10^{-14}$ | 0.07 (0.03, 0.10) | 0.0002 |
| | IMPROVE | 0.01 (-0.02, 0.04) | 0.54 | 0.06 (0.02, 0.09) | 0.002 |
| Standardised $\log_e$ (HDL-C) | WHII | -0.07 (-0.10, -0.03) | 0.0003 | -0.06 (-0.10, -0.03) | 0.001 |
| | IMPROVE | -0.09 (-0.12, -0.06)) | $5.2\text{x}10^{-08}$ | -0.06 (-0.09, -0.02) | 0.002 |
| Standardised $\log_e$(triglycerides) | WHII | 0.09 (0.06, 0.13) | $1.4\text{x}10^{-07}$ | 0.05 (0.01, 0.08) | 0.007 |
| | IMPROVE | -0.02 (-0.05, 0.02) | 0.36 | -0.02 (-0.05, 0.01) | 0.24 |

### 5.3.8 Causal Effect Estimation using 2SLS

Based on the meta-analysis of the estimates derived from the 2SLS instrumental variable analysis, a 1 mmol/L higher LDL-C was associated with a 3% (beta (95% CI) = 0.03 (0.02 – 0.05)) and 4% (beta (95% CI) = 0.04 (0.02 – 0.06)) higher CIMT, when using the Cardiochip and GLGC LDL-C genetic scores, respectively, as instruments. Taking the mean CIMT in the two studies (0.8 mm and 1.2 mm), this would translate into a 0.02 – 0.05 mm difference in CIMT per mmol/L difference in LDL-C. HDL-C and triglycerides were not found to be associated with CIMT using instrumental variable analysis. For comparison, results using standardised variables are shown in Figure 5.6. There was no change in the overall IV estimate when using lipid levels corrected for statin use (Figure 5.7).

**Figure 5.5 Association of LDL-C, HDL-C and triglyceride genetic scores with CIMT.** (A) Unadjusted for covariates (B) adjusted for first three principal components (C) adjusted for the first 3 dimensions from MDS and non-index lipid fractions. Beta-coefficients are shown as SD change in log$_e$-transformed CIMT per 1SD change in genetic score.

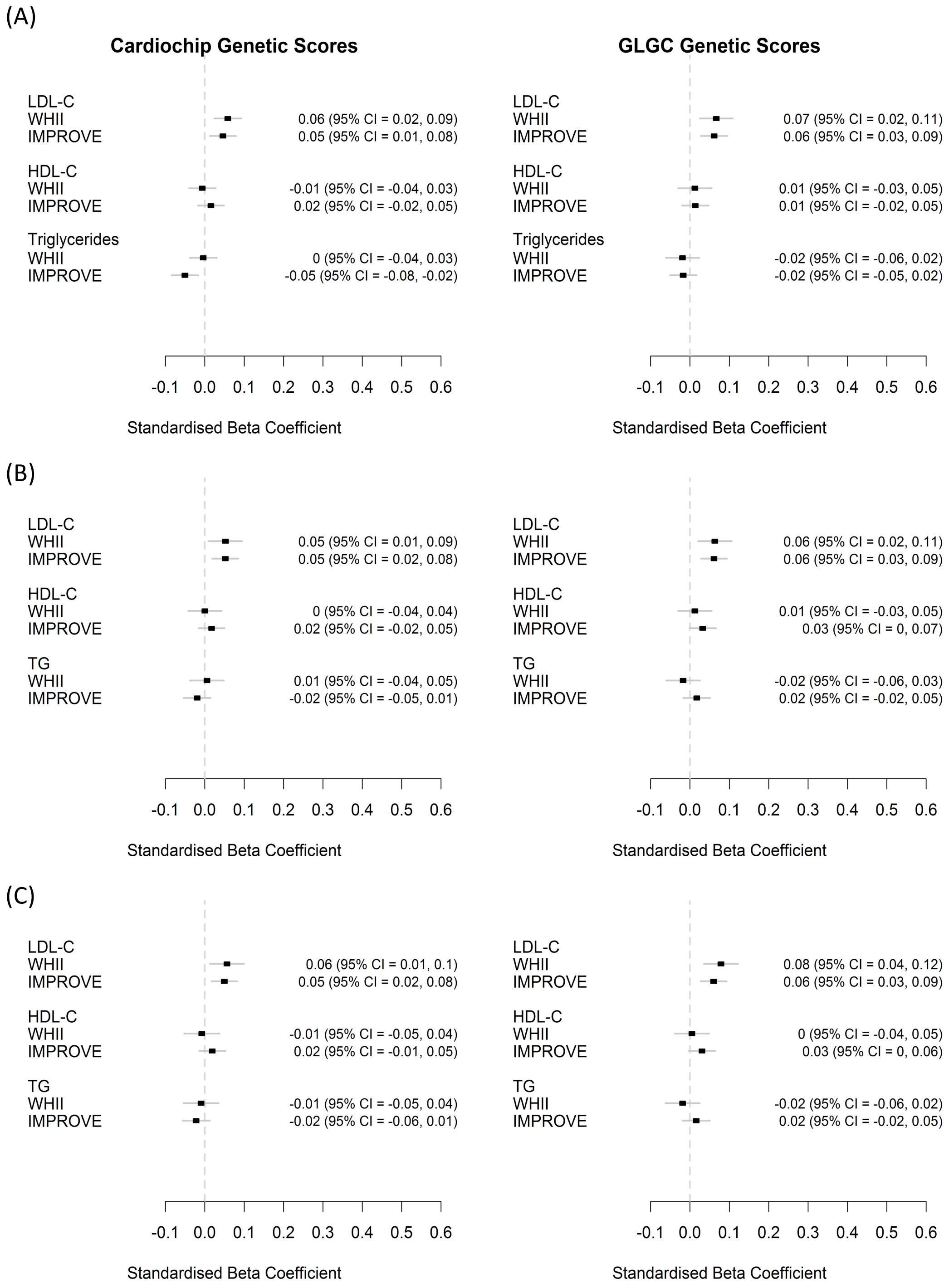


**Figure 5.6 Two-stage least squares regression analysis.** Association of lipid fractions with CIMT obtained from the instrumental variable analysis in which lipid genetic scores act as instruments for the non-confounded effect of each lipid fraction. Effect sizes and 95% confidence intervals in each study and summary estimates from a fixed-effect model are shown as SD change in $\log_e$-transformed CIMT per 1 SD change in LDL-C, $\log_e$-transformed HDL-C and $\log_e$-transformed triglycerides.

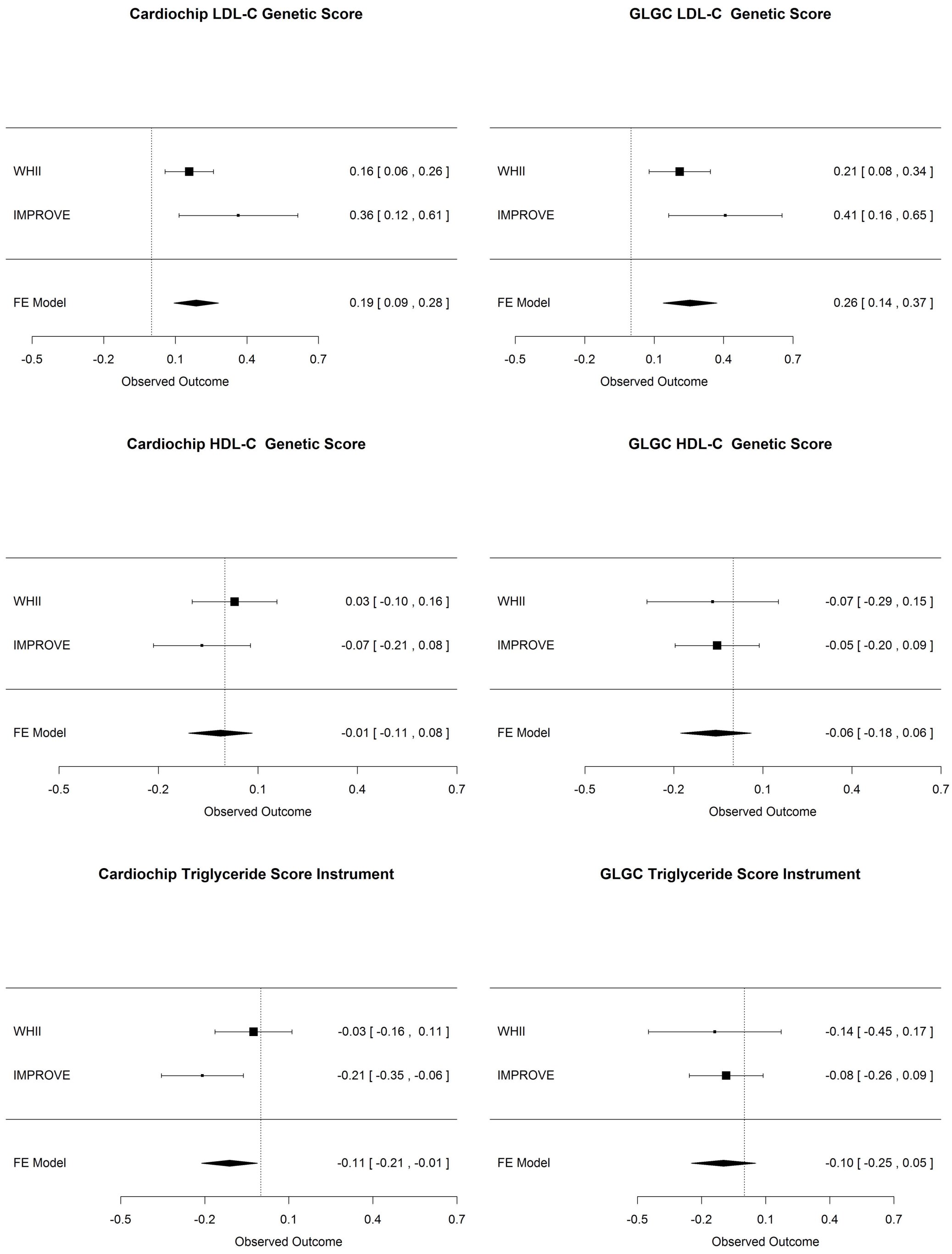
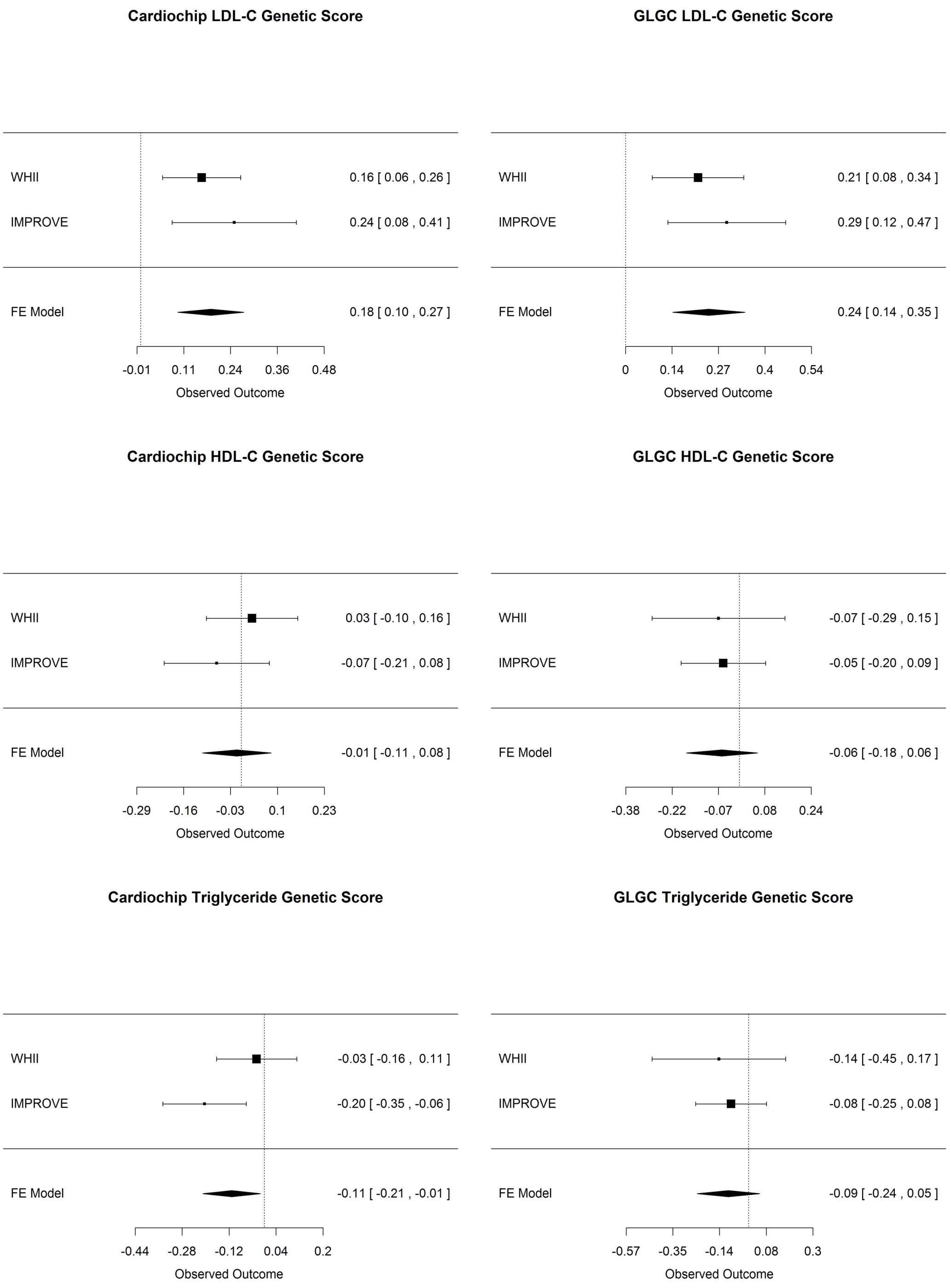


**Figure 5.7 Instrumental variable analysis using lipid levels corrected for statin use.** Effect sizes and 95% confidence intervals in each study and summary estimates from a fixed-effect model are shown as SD change in $\log_e$-transformed CIMT per 1 standard deviation change in LDL-C, $\log_e$-transformed HDL-C and $\log_e$-transformed triglycerides.



## 5.4 Discussion

### 5.4.1 Summary of Results

LDL-C, HDL-C and triglyceride genetic scores were used in an MR analysis to assess the causal relationship between each lipid fraction and CIMT. Though there was a positive association between directly measured LDL-C and common CIMT and a negative association between directly measured HDL-C and CIMT, the results from the MR analysis support a causal association with LDL-C only. Despite differences in cohort characteristics (i.e. healthier individuals and much smaller proportion on lipid-lowering medication in WHII compared to IMPROVE), the different effect of genetic score on lipid levels (smaller in the IMPROVE study) and the different SNPS used in the two genetic scores (Cardiochip versus GLGC), the causal association between LDL-C and CIMT was found to be consistent in both studies for both the Cardiochip and GLGC genetic score instruments. Although the HDL-C and triglyceride genetic scores explained a similar proportion of the variation in the index trait to the LDL-C score (not taking into account the $R^2$ for the Cardiochip genetic score in WHII which was inflated due to discovery bias), there was a lack of significant effect on CIMT for these two lipid fractions. The findings suggest that CIMT is likely to be a reliable surrogate outcome measure in randomised trials of LDL-lowering therapy. However, due to the lack of evidence in the genetic association between HDL-C, triglycerides and CIMT, the findings from the current analysis cast doubt on the use of CIMT as a surrogate outcome measure in trials of HDL-C and triglyceride-modifying therapies.

One criterion for causality is the magnitude of effect. Taking the mean CIMT in the two studies (0.8 mm and 1.2 mm), the IV beta-coefficients estimated from the two genetic scores would translate into approximately a 0.02 – 0.05 mm difference in CIMT per mmol/L difference in LDL-C. To contextualise these findings, a meta-analysis of 11 statin trials (Bedi et al. 2010) found that after treatment with statins (mean treatment duration of 25.6 months), there was a significant reduction in the mean LDL-C (pre-treatment: 4.36 ± 0.85 mmol/L, post-treatment: 2.64 ± 0.72 mmol/L, $P<0.05$, N=2132) and also a 0.04 mm (95% CI=0.028-0.052) difference in mean CIMT between statin therapy arm and placebo arm (Bedi et al., 2010). This is roughly equivalent to a 0.02mm decrease in CIMT per mmol/L decrease in LDL-C, and therefore reasonably concordant with the genetically-inferred casual association from the current analysis.

### 5.4.2 CIMT as a Surrogate Marker

Randomised controlled drug trials with hard clinical end points require a very large number of participants and follow-up over a long period, making them technically and financially challenging. As a result, surrogate endpoints are frequently employed in earlier phase studies to inform the decision to undertake a larger outcome trial. CIMT has been used as a surrogate end-point in many trials of lipid-modifying drugs (Crouse et al. 2007; MacMahon et al. 1998; Smilde et al. 2001; Taylor et al. 2004). However, the suitability of CIMT as a surrogate marker in cardiovascular drug trials is controversial (Lorenz et al. 2012). The underlying assumption in trials using CIMT as a surrogate marker is that the rate of change in CIMT over time in response to drug therapies reflects the change in the risk of cardiovascular outcomes. The majority of CIMT trials, however, have short follow-up periods and modest sample sizes and therefore lack power to identify associations with cardiovascular outcomes. Rather, they are designed to provide inferences on cardiovascular outcomes based on a presumed inverse relationship between atherosclerosis progression and cardiovascular benefit (Taylor et al., 2011).

A recent large-scale meta-analysis of 41 randomised trials assessing CIMT at baseline and follow-up after treatment (Costanzo et al. 2010), including a total of

18,307 participants, concluded that regression of CIMT induced by cardiovascular drug therapies was not associated with a reduction in cardiovascular events. Though the meta-analysis was technically sound, the heterogeneity in the interventions evaluated, the methods used for CIMT measurement, the outcome definition, study design, population characteristics and follow-up time between the 41 trials may have reduced the ability to detect an association between carotid IMT and cardiovascular event reductions in such trials. On the other hand, the ACCORD Lipid trial (Ginsberg et al. 2010) aimed to test whether treatment of T2D patients with fenofibrate, to increase plasma HDL-C and reduce triglyceride levels, would result in additional cardiovascular benefit compared with simvastatin (LDL-lowering) therapy alone. Though the addition of fenofibrates to statin treatment did not show any significant reduction in clinical events in the placebo versus treatment groups, sub-group (defined by baseline lipid levels) analyses suggested benefits of fenofibrate therapy in mixed dyslipidemia individuals (individuals with triglyceride levels in the upper tertile and HDL-C levels in the lower tertile at baseline). Therefore, a MR analysis to determine causality between HDL-C, triglycerides and CIMT in a sufficiently large mixed dyslipidemia sample may be worthwhile.

### 5.4.3 Previous Mendelian Randomisation Studies of Lipids and CIMT

To date there have been very few Mendelian randomisation studies addressing association of lipids with CIMT or CHD. A study by Aulchenko et al (Aulchenko et al. 2009) generated genetic scores for total cholesterol (11 SNPS), LDL-C (8 SNPs), HDL-C (8 SNPs) and triglycerides (11 SNPs) based on SNPs identified in a meta-analysis in around 20,000 European individuals. They looked at the direct association of the genetic scores with CIMT in the Rotterdam study (~5700 individuals over the age of 55) and found only the total cholesterol genetic score and a combined score with all lipid SNPs to be associated with CIMT. The scores explained less than 5% of the variation in each of the respective lipid fractions. However, they did not use an instrumental variable approach to quantify the causal effect.

### 5.4.4 Previous Mendelian Randomisation Studies of Lipids and CVD Events

Four SNPs that were associated with either HDL-C or triglycerides, but not with LDL-C in the GLGC meta-analysis, were also associated with CAD (Teslovich et al. 2010). However, one of the SNPs in the *IRS1* locus was also associated with increased risk of type 2 diabetes mellitus, insulin resistance and hyperinsulinemia. Therefore, the genes in these loci may have pleiotropic effects on non-lipid parameters that are causal for CAD risk reduction (Teslovich et al. 2010). The Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factor Collaboration compared the risk of genetically elevated triglycerides levels based on a single SNP (the *APOA5* SNP rs662799) among over 20,000 CHD cases and 35,000 controls (Sarwar et al. 2010). They concluded that there was a causal role for triglycerides-mediated pathway(s) in CHD. However, the *APOA5* variant was also associated with HDL-C levels. The association of genetically determined triglycerides levels with CHD was also attenuated to the null after adjusting not only for HDL, but also for non-HDL cholesterol and other variables. Since the effect of the rs662799 SNP is not exclusive to triglycerides, this compromises one key assumption for a valid MR analysis and complicates the inference on the potential causal role of triglycerides in CHD. A study by Voight et al (Voight et al. 2012b) used a genetic score consisting of 14 common SNPs selected for a predominant effect on HDL-C and tested this score in up to 12,482 cases of myocardial infarction and 41,331 controls. As a positive control, they also tested a genetic score of 13 common SNPs exclusively associated with LDL-C. They found no casual association of HDL-C with MI but were able to confirm an association of LDL-C with MI. A study currently under review (Do et al. 2013) found that triglycerides may causally effect risk of CAD even after taking into account any pleiotropic effects of triglyceride-associated SNPs with other lipid fractions, and therefore novel therapeutic approaches to triglyceride-rich lipoproteins might be expected to reduce CAD risk. This would suggest that previous failed trials of drugs targeting triglycerides could have been related to the specific drug or drug target, or the use of an unsuitable surrogate marker (i.e. CIMT) for CVD risk.

## 5.4.5 Limitations

### 5.4.5.1 Validity of Instruments

Validity of any MR analysis may be compromised by a) population stratification, where allele frequencies and disease rates differ between population subgroups; b) pleiotropy, where genetic instruments affect the outcome through more than one intermediate risk factor; and c) linkage disequilibrium, where another polymorphism in close proximity (and in linkage disequilibrium) to the variant of interest, is causing disease through another pathway; and d) weak instrument bias.

Analysis in the WHII was restricted to Caucasians and MDS revealed no substantial population stratification after quality control analysis. In the IMPROVE study, though all individuals were Caucasians there was population stratification that reflected the geographical location from which the samples were obtained (Baldassarre et al. 2010). However, the SNPs used to generate the GLGC genetic scores were also discovered in individuals of European descent from the United States, Europe or Australia. Therefore, the scores should be applicable to the general European population and stratification is less likely to be an issue in this MR analysis. Also, adjusting for population stratification did not alter the overall conclusions made from the analysis.

Often, genes act on multiple pathways and may therefore be associated with multiple intermediate phenotypes, especially those that act as transcription factors for other genes. Some SNPs included in the score may be independently associated with other cardiovascular risk factors and so individually they would not be valid instruments. By combining these multiple SNPs into one score the issue of pleiotropy can be addressed. This was demonstrated in chapter 4, where the difference in the proportion of variance explained by the genetic score in the index and non-index traits was much larger than that for a single SNP, suggesting that when a large number of genetic variants are combined into a single genetic score, pleiotropic effects may be expected to balance out. An alternative to overcoming non-specificity would be to only use SNPs associated with the exposure of interest for generating the genetic risk score. This approach has been used by recent work currently under review (Do et al. 2013), whereby a genetic risk score for triglycerides was generated using SNPs that showed large effect on triglyceride levels but minimal effect on LDL-C. Another approach used in this study to overcome pleiotropic effects of SNPs was to use the residuals from the regression of the exposure of interest on the non-specific risk factors, so that any SNP effects on the non-specific risk factors were accounted for. Therefore, several approaches can be adopted to ensure that the problem of specificity of the instrument is appropriately addressed.

Association of an outcome with one polymorphism could have arisen by chance or confounding due to LD, but associations with more than one polymorphism in different genes marking the same exposure are unlikely unless the exposure is causal (Lewis 2010). Given the large number of lipid genetic variants that have been identified by different studies, it is possible to generate many independent combinations of such variants, and from these many independent instrumental variable estimates of the causal effect of exposure of interest on outcome. Both the Cardiochip and GLGC genetic scores, which used only partially overlapping SNP sets, supported the causal association of LDL-C with CIMT in each study. Using two different scores containing only partially overlapping SNPs provides confidence that the results are not biased by the SNP set used, as the two instruments are unlikely to be influenced by the same pleiotropy or LD-induced confounding.

Though the instrument strength of the Cardiochip genetic scores in WHII are inflated due to discovery bias, it is important to note that all genetic scores had comparable instrument strength in the IMPROVE study, and despite the HDL-C genetic scores being the strongest instruments in this cohort, causality was only observed for LDL-C. MR analysis in a much larger sample size sizes would be needed to confirm or refute the findings for HDL-C and triglycerides from this study.

Our method for generating genetic scores makes several assumptions: additive effects of alleles, no gene-gene interactions and a linear effect of lipids on CIMT. Though not explored in this work, if these assumptions did not hold, it would be possible to incorporate such knowledge into the model. However, previous work has shown that genetic scores are robust even in the presence of a different underlying genetic model.

#### 5.4.5.2 Refinement of Instruments

It remains to be seen whether the addition of more SNPs that increase the HDL-C and triglyceride instrument strength will alter the conclusions based on this analysis. However, the effect sizes of additional loci identified in very large meta-analysis tend to be extremely small, therefore addition of these may not significantly improve instrument strength.

Conventional laboratory measures of LDL-C, HDL-C, total cholesterol and triglycerides sum up the lipids carried in lipoprotein particles of various sizes and composition (Tukiainen et al. 2012). Recent developments in high-throughput analytical technologies such as nuclear magnetic resonance (NMR) and mass spectrometry allows more refined metabolic profiling, including measurement of a broader range of lipoprotein subclasses. Large-scale metabolomics studies have provided examples of SNPs associated with HDL-C showing associations in the opposite direction with larger and smaller HDL particles for the same allele (Tukiainen et al. 2012). Therefore, the observed heterogeneity in the biological effects of HDL-C and triglycerides may mask the detection of any true association with CIMT to CHD. Metabolomic studies have also shown that SNPs show stronger association with, and explain a greater proportion of the variance of lipoprotein subclasses compared to enzymatic lipid measures (Tukiainen et al. 2012). Several studies have shown that HDL subclasses differ in their relationship to CHD, with larger HDL particles thought to be more anti-atherogenic than smaller ones (Ala-Korpela 2008; Krauss 2010; Morgan et al. 2004). MR studies using genetic scores representative of the specific lipoprotein subclasses would therefore be a logical future direction.

# 6 Discussion

The key aim of cardiovascular disease genetics has been to correlate genotype with phenotype in order to identify the genes and sequence changes contributing to trait variation and disease susceptibility in humans and, in doing so, provide insight into the biological mechanisms involved in disease development (Kathiresan & Srivastava 2012). Large-scale association studies have identified numerous common genetic variants associated with CVD traits and risk factors. Some of the better studied traits include lipids, blood pressure, CAD, MI and stroke. However, there are still many important CVD risk factors that have not been extensively studied, and large-scale genetic discovery has the potential to provide new insight into the biological pathways responsible for variation in these traits. For risk factors where a large number of genetic variants have already been identified, there is great interest in exploring how these can be applied, not only for disease risk prediction but also for assessing causal disease pathways. The aims of the work in this thesis were therefore two-fold: first, discovery of genetic variants influencing variation in left ventricular mass, an important CVD risk factor for which the few large-scale association studies that have previously been carried out have not been very successful, and second, investigating the application of prior knowledge of the large number of genetic variants associated with well-studied lipid traits for risk prediction and Mendelian randomisation analysis.

In chapter 2, I presented the results from an association analysis with ECG-derived LV mass, a convenient clinical measure of LVH. Genetic variants in four genes (*SCN5A, IGF1R*, *PTGES3* and *NMB*) were robustly associated with one or more of the ECG LV mass indices. Other plausible loci, such as SNPs in sarcomeric genes, showed suggestive association, and larger studies would be needed to confirm or refute these associations. There is some evidence based on ENCODE data that some of the identified variants are within regulatory regions. However, a variant can have a direct effect on gene expression in human tissues or be functional in another way, without necessarily having a causal effect on the trait (Kathiresan & Srivastava 2012).

A combination of fine-mapping of the identified loci, either through imputation or sequencing, and functional experiments would be required to identify (1) the causal gene (2) the causal variant (3) the mechanism by which the variant affects the gene and (4) the mechanism by which the gene affects phenotype, making the road from genotype to phenotype a potentially long and arduous one. The chromosome 9p21.3 locus which was first associated with CAD and MI in 2007 (Samani et al. 2007; McPherson et al. 2007; Helgadottir et al. 2007) provides an extreme example of the difficulties that may be encountered partly perhaps because the most strongly associated SNPs are non-coding and over 100 kilobases downstream of the nearest protein-coding genes *CDKN2A* and *CDKN2B* (both cyclin-dependent kinase inhibitors with a role in cell-cycle regulation). Re-sequencing and fine-mapping studies in this region were unable to identify a causal variant (Shea et al. 2011). Alteration of a non-coding RNA (antisense non-coding RNA in the *INK4* locus (*ANRIL*)), and disruption of binding of the STAT1 transcription factor binding (Harismendy et al. 2011; Holdt & Teupser 2012) have both been explored as potential mechanisms by which the non-coding variants may alter susceptibility to CAD, but no definitive answers have yet emerged. Variants in this region have shown to be independently associated with expression of *CDKN2A, CDKN2B*, and *ANRIL,* with individual SNPs influencing *ANRIL* and *CDKN2B* expression in opposite directions, suggesting that modulation of *ANRIL* expression may mediate susceptibility to disease (Cunnington et al. 2010). A recent eQTL study found the 9p21 locus to be associated with the expression of multiple genes enriched for biomarkers of myocardial infarction, response to wounding and inflammatory processes. However, none of the genes identified as having altered expression in association with the 9p21.3 risk allele remained significant after correction for multiple comparisons (Pilbrow et al. 2012). An independent study showed that the locus had no effect on a wide range of known and putative CVD biomarkers (Angelakopoulou et al. 2012). Six years on, the functional relevance of this region for CAD is therefore yet to be determined.

Genetic association studies have identified over 100 loci associated with one or more of the major lipid fractions, providing substantial biological insights into lipid biology. However, as was shown in chapter 3, their performance in discriminating individuals with high absolute risk of CVD, those that require lipid medication to manage CVD risk and those that develop CHD is poor, offering no improvement over non-genetic factors. This may reflect the fact that the proportion of overall phenotypic variance explained by the combined effects of the variants used is relatively small.  In addition, the distribution of CVD risk alleles in the general population is normal, therefore, more disease events are observed among the large majority who have intermediate numbers of risk alleles than the minority who have a large number of risk alleles and are therefore considered to be at high risk of disease — the prevention paradox (Rose 1985). More recent methodologies that analyse all SNPs simultaneously have shown that the SNPs on the genotyping platform explain a much larger proportion of the phenotypic variance (Yang et al. 2010) than is estimated from SNPs selected based on some significance criteria. Extending such methodology to risk prediction may improve predictive performance. A recent study assessing the performance of risk prediction by polygenic models showed that the predictive ability depends on both the total heritability of phenotype as well as the underlying effect-size distributions (Chatterjee et al. 2013). They show that under the most likely effect-size distributions, the optimal significance threshold for selecting SNPs for prediction models in large GWAS can be much less stringent than the standard (p-value < $5x10^{-08}$) used in discovery association analyses. They also concluded that the effect-size distributions of genetic variants from large GWAS suggest that though increase in total sample size of the training dataset will improve risk prediction models, the improvement will be slow and modest even when common SNPs account for large proportion of heritability of the underlying traits (Chatterjee et al. 2013). Though such polygenic models may not achieve high discriminatory power, it is worth noting that even models with modest discriminatory power may provide important stratification for absolute risk (Chatterjee et al. 2013). Development of robust prediction models in the future will need to consider integrating both common and rare alleles as well as other non-genetic CVD risk factors.

More recently, there has been a lot of interest in identifying suitable instruments that can be used to distinguish causal from non-causal biomarkers of disease. Though both types are useful for predicting disease risk, only causal biomarkers are a suitable as therapeutic targets. In chapter 4, I explored different approaches to instrument development for LDL-C, HDL-C and triglycerides, all of which are polygenic traits with multiple associated SNPs. A weighted genetic score based on SNPs selected using variable selection with the Bayesian Information Criterion for model selection and Ridge regression for shrinkage of beta coefficients provided the best instrument in terms of strength and specificity. These lipid instruments were then applied in an MR analysis to assess the causal relationship between the three lipid fractions and CIMT in two cohorts (chapter 5). Completely independent weighted genetic scores based on SNPs and beta-coefficients identified by the GLGC meta-analysis were also used as instruments. LDL-C was found to be a casual factor in both studies regardless of which genetic score instrument was used, but no robust causal association was found with HDL-C and triglycerides. This would suggest that for trials of therapeutic interventions targeting HDL-C and triglycerides levels, CIMT may not be a suitable surrogate marker for assessing the efficacy of such drugs in reducing CHD risk. However, larger studies are needed to confirm these findings.

It still remains to be confirmed whether these two lipid fractions are causal for CHD. Two clinical trials involving therapeutic elevations of HDL-C on patients on statin therapy were prematurely terminated based on the failure to improve cardiovascular outcomes, including MI and stroke. The AIM-HIGH study (Boden et al. 2011) was conducted with niacin, the most effective HDL-C-raising drug currently on the market; the dal-OUTCOMES study (Schwartz et al. 2009) involved dalcetrapib, a drug in development that partially inhibits CETP, which transfers cholesterol from HDL to VLDL or LDL. A recent MR analysis used a single SNP as well as a 14-SNP genetic score as an instrument for HDL-C and found no casual association with risk of MI in around 12,500 cases and over 41,000 controls, challenging the concept that therapeutic interventions that specifically raise plasma HDL-C will translate into reductions in risk of MI (Voight et al. 2012). These would suggest that HDL-C may simply be a risk marker and not a causal risk mediator (Rader & Tall 2012). However, a small study recently published results on the effects of niacin on HDL function. They assessed functional properties using two tests - cholesterol efflux capacity (a measure of how well HDL removes cholesterol from lipid-loaded cells) and the HDL inflammatory index (which quantifies the antioxidant properties as it relates to preventing the oxidation of LDL). Though treatment of patients on statins with niacin increased HDL-C levels by 29% compared to only a 2% increase in those without niacin, they saw no significant changes in HDL function. Another MR analysis found non-fasting remnant cholesterol to be causally associated with ischemic heart disease (Varbo et al. 2011). Remnant cholesterol is the cholesterol content of triglyceride-rich lipoproteins, composed of very low-density lipoproteins and chylomicron remnants in the nonfasting state. This study suggests that the elevated cholesterol content of triglyceride-rich lipoprotein particles causes ischemic heart disease. However, given the pleiotropic effects of the genetic variants studied, these findings would need to be confirmed using additional genetic variants and/or randomized intervention trials (Varbo et al. 2013). In light of these new studies, it has become apparent that shifting focus from HDL-C levels to HDL function or other lipid and lipoprotein subclasses will be imminent in future research on lipids in relation to CVD. The mechanism by which HDL-C is elevated may be equally important and drugs that raise HDL-C through alternative mechanisms may still be useful. With enhances in high-throughput analytical technologies such as mass spectrometry, it has become feasible to measure a much broader range of  lipoprotein subclasses in large cohorts (Tukiainen et al. 2012). Already, several UK population cohorts (including WHII and BWHHS) have begun metabolic profiling of individuals, with the potential to further dissect the lipid metabolic pathway in relation to CVD outcome and drug development strategies.

In the future, larger collaborations; newer analytical methods; analysis of other types of genetic variants (rare variants and CNVs); integration of genetic, genomic and metabolomic data; functional experiments; and the availability of more affordable next-generation sequencing technology will all ensure continued progression in biological discovery and research into the utility of genetic risk factors for complex CVD traits.

# Bibliography

Abdollahi, M. R., Guthrie, P. A. I., Smith, G. D., Lawlor, D. A., Ebrahim, S., & Day, I. N. M. (2006). Integrated single-label liquid-phase assay of APOE codons 112 and 158 and a lipoprotein study in British women. *Clinical Chemistry*, *52*(7), 1420–1423.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

Ahuja, P., Sdek, P., & MacLellan, W. R. (2007). Cardiac Myocyte Cell Cycle Control in Development, Disease, and Regeneration. *Physiological Reviews*, *87*(2), 521–544.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Ala-Korpela, M. (2008). Critical evaluation of 1H NMR metabonomics of serum as a methodology for disease risk assessment and diagnostics. *Clinical Chemistry and Laboratory Medicine*, *46*(1), 27–42.

Anderson, K. M. (1991). A nonproportional hazards Weibull accelerated failure time regression model. *Biometrics*, *47*(1), 281–288.

Anderson, K. M., Odell, P. M., Wilson, P. W., & Kannel, W. B. (1991). Cardiovascular disease risk profiles. *American Heart Journal*, *121*(1), 293–298.

Angelakopoulou, A., Shah, T., Sofat, R., Shah, S., Berry, D. J., Cooper, J., Palmen, J., et al. (2012). Comparative analysis of genome-wide association studies signals for lipids, diabetes, and coronary heart disease: Cardiovascular Biomarker Genetics Collaboration. *European Heart Journal*, *33*(3), 393–407.

Angrist, J., Imbens, G. W., & Rubin, D. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–472.

Arking, D. E., & Chakravarti, A. (2009). Understanding cardiovascular disease through the lens of genome-wide association studies. *Trends in Genetics*, *25*(9), 387–394.

Arnett, D. K., De las Fuentes, L., & Broeckel, U. (2004). Genes for left ventricular hypertrophy. *Current Hypertension Reports*, *6*(1), 36–41.

Arnett, D. K., Li, N., Tang, W., Rao, D. C., Devereux, R. B., Claas, S. A., Kraemer, R., et al. (2009). Genome-wide association study identifies single-nucleotide polymorphism in KCNB1 associated with left ventricular mass in humans: the HyperGEN Study. *BMC Medical Genetics*, *10*, 43.

Arnett, D. K., Meyers, K. J., Devereux, R. B., Tiwari, H. K., Gu, C. C., Vaughan, L. K., Perry, R. T., et al. (2011). Genetic variation in NCAM1 contributes to left ventricular wall thickness in hypertensive families. *Circulation Research*, *108*(3), 279–283.

Ashley, E. A., Hershberger, R. E., Caleshu, C., Ellinor, P. T., Garcia, J. G. N., Herrington, D. M., Ho, C. Y., et al. (2012). Genetics and cardiovascular disease: a policy statement from the American Heart Association. *Circulation*, *126*(1), 142–157.

Asselbergs, F. W., Guo, Y., van Iperen, E. P., Sivapalaratnam, S., Tragante, V., Lanktree, M. B., Lange, L. A., et al. (2012). Large-Scale Gene-Centric Meta-analysis across 32 Studies Identifies Multiple Lipid Loci. *The American Journal of Human Genetics*, *91*(5), 823–838.

Assmann, G., Cullen, P., & Schulte, H. (2002). Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. *Circulation*, *105*(3), 310–315.

Aulchenko, Y. S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I. M., Pramstaller, P. P., Penninx, B. W. J. H., et al. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature Genetics*, *41*(1), 47–55.

Baigent, C., Blackwell, L., Emberson, J., Holland, L., Reith, C., Bhala, N., Peto, R., et al. (2010). Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170 000 participants in 26 randomised trials. *Lancet*, *376*(9753), 1670–1681.

Baigent, C., Keech, A., Kearney, P. M., Blackwell, L., Buck, G., Pollicino, C., Kirby, A., et al. (2005). Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet*, *366*(9493), 1267–1278.

Baldassarre, D., Nyyssönen, K., Rauramaa, R., De Faire, U., Hamsten, A., Smit, A. J., Mannarino, E., et al. (2010). Cross-sectional analysis of baseline data to identify the major determinants of carotid intima-media thickness in a European population: the IMPROVE study. *European Heart Journal*, *31*(5), 614–622.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*(10), 781–791.

Basmann, R.L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica 25* (1), 77–83.

Bedi, U. S., Singh, M., Singh, P. P., Bhuriya, R., Bahekar, A., Molnar, J., Khosla, S., et al. (2010). Effects of statins on progression of carotid atherosclerosis as measured by carotid intimal-medial thickness: a meta-analysis of randomized controlled trials. *Journal of Cardiovascular Pharmacology and Therapeutics*, *15*(3), 268–273.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (1st ed.). Wiley-Interscience.

Bennet, A. M., Di Angelantonio, E., Ye, Z., Wensley, F., Dahlin, A., Ahlbom, A., Keavney, B., et al. (2007). Association of apolipoprotein E genotypes with lipid levels and coronary risk. *Journal of the American Medical Association*, *298*(11), 1300–1311.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816.

Bis, J. C., Kavousi, M., Franceschini, N., Isaacs, A., Abecasis, G. R., Schminke, U., Post, W. S., et al. (2011). Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. *Nature Genetics*, *43*(10), 940–947.

Boden, W. E., Probstfield, J. L., Anderson, T., Chaitman, B. R., Desvignes-Nickens, P., Koprowicz, K., McBride, R., et al. (2011). Niacin in patients with low HDL cholesterol levels receiving intensive statin therapy. *The New England Journal of Medicine*, *365*(24), 2255–2267.

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, *40*(6), 695–701.

Borg, I., & Groenen, P. (2005). The four purposes of multidimensional scaling. *Modern Multidimensional Scaling: theory and applications.* (2nd ed., pp. 3). New York: Springer.

Bots, M L, Hoes, A. W., Koudstaal, P. J., Hofman, A., & Grobbee, D. E. (1997). Common carotid intima-media thickness and risk of stroke and myocardial infarction: the Rotterdam Study. *Circulation*, *96*(5), 1432–1437.

Bots, Michiel L, Visseren, F. L., Evans, G. W., Riley, W. A., Revkin, J. H., Tegeler, C. H., Shear, C. L., et al. (2007). Torcetrapib and carotid intima-media thickness in mixed dyslipidaemia (RADIANCE 2 study): a randomised, double-blind trial. *Lancet*, *370*(9582), 153–160.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Branchi, A., Rovellini, A., Torri, A., & Sommariva, D. (1998). Accuracy of calculated serum low-density lipoprotein cholesterol for the assessment of coronary heart disease risk in NIDDM patients. *Diabetes Care*, *21*(9), 1397–1402.

Brindle, P., Emberson, J., Lampe, F., Walker, M., Whincup, P., Fahey, T., & Ebrahim, S. (2003). Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. *British Medical Journal*, *327*(7426), 1267.

Brown, P. J. (1993). *Measurement, Regression and Calibration*. Oxford: Clarendon Press.

Burgess, S., & Thompson, S. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, *40*(3), 755–764.

Burgess, S., & Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization (under review).

Cambien, F., & Tiret, L. (2007). Genetics of cardiovascular diseases: from single mutations to the whole genome. *Circulation*, *116*(15), 1714–1724.

Cannon, C. P., Shah, S., Dansky, H. M., Davidson, M., Brinton, E. A., Gotto, A. M., Stepanavage, M., et al. (2010). Safety of anacetrapib in patients with or at high risk for coronary heart disease. *The New England Journal of Medicine*, *363*(25), 2406–2415.

Carvajal-Carmona, L. G. (2010). Genetic dissection of intermediate phenotypes as a way to discover novel cancer susceptibility alleles. *Current Opinion in Genetics & Development*, *20*(3), 308–314.

Casale, P. N., Devereux, R. B., Alonso, D. R., Campo, E., & Kligfield, P. (1987). Improved sex-specific criteria of left ventricular hypertrophy for clinical and computer interpretation of electrocardiograms: validation with autopsy findings. *Circulation*, *75*(3), 565–572.

Caulfield, M., Munroe, P., Pembroke, J., Samani, N., Dominiczak, A., Brown, M., Benjamin, N., et al. (2003). Genome-wide mapping of human loci for essential hypertension. *Lancet*, *361*(9375), 2118–2123.

Chambers, J. C., Zhao, J., Terracciano, C. M. N., Bezzina, C. R., Zhang, W., Kaba, R., Navaratnarajah, M., et al. (2010). Genetic variation in SCN10A influences cardiac conduction. *Nature Genetics*, *42*(2), 149–152.

Chambless, L. E., Folsom, A. R., Clegg, L. X., Sharrett, A. R., Shahar, E., Nieto, F. J., Rosamond, W. D., et al. (2000). Carotid wall thickness is predictive of incident clinical stroke: the Atherosclerosis Risk in Communities (ARIC) study. *American Journal of Epidemiology*, *151*(5), 478–487.

Chambless, L. E., Heiss, G., Folsom, A. R., Rosamond, W., Szklo, M., Sharrett, A. R., & Clegg, L. X. (1997). Association of coronary heart disease incidence with carotid arterial wall thickness and major risk factors: the Atherosclerosis Risk in Communities (ARIC) Study, 1987-1993. *American Journal of Epidemiology*, *146*(6), 483–494.

Charitakis, K., & Basson, C. T. (2010). Can genetic testing improve our aim in hypertrophic cardiomyopathy? *Circulation Research*, *106*(9), 1446–1448.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., & Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics*, *45*(4), 400-405

Cortes, A., & Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy*, *13*(1), 101.

Costanzo, P., Perrone-Filardi, P., Vassallo, E., Paolillo, S., Cesarano, P., Brevetti, G., & Chiariello, M. (2010). Does carotid intima-media thickness regression predict reduction of cardiovascular events? A meta-analysis of 41 randomized trials. *Journal of the American College of Cardiology*, *56*(24), 2006–2020.

Crouse, J. R., Raichlen, J. S., Riley, W. A., Evans, G. W., Palmer, M. K., O'Leary, D. H., Grobbee, D. E., et al. (2007). Effect of rosuvastatin on progression of carotid intima-media thickness in low-risk individuals with subclinical atherosclerosis: the METEOR Trial. *Journal of the American Medical Association*, *297*(12), 1344–1353.

Cunnington, M. S., Santibanez Koref, M., Mayosi, B. M., Burn, J., & Keavney, B. (2010). Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. *PLoS Genetics*, *6*(4), e1000899.

Davey Smith, G., & Ebrahim, S. (2003). "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*(1), 1–22.

Davies, R. W., Wells, G. A., Stewart, A. F. R., Erdmann, J., Shah, S. H., Ferguson, J. F., Hall, A. S., et al. (2012). A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. *Circulation Cardiovascular Genetics*, *5*(2), 217–225.

De Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., & Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, *37*(11), 1217–1223.

De Simone, G., Pasanisi, F., & Contaldo, F. (2001). Link of nonhemodynamic factors to hemodynamic determinants of left ventricular hypertrophy. *Hypertension*, *38*(1), 13–18.

Dent, T. (2009). Predicting the risk of coronary heart disease. Available from www.phgfoundation.org/file/5159/. [Accessed 24 October 2012].

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004.

Di Angelantonio, E., & Butterworth, A. S. (2012). Clinical utility of genetic variants for cardiovascular risk prediction: a futile exercise or insufficient data? *Circulation Cardiovascular Genetics*, *5*(4), 387–390.

Di Angelantonio, E., Sarwar, N., Perry, P., Kaptoge, S., Ray, K. K., Thompson, A., Wood, A. M., et al. (2009). Major lipids, apolipoproteins, and risk of vascular disease. *Journal of the American Medical Association*, *302*(18), 1993–2000.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, *325*(5945), 1246–1250.

Do, R., Willer, C.J., Schmidt, E.M., Sengupta, S., Gao, C., Peloso, G.M., Gustafsson, S., et al. (2013). Polymorphisms associated with plasma triglycerides and risk for coronary artery disease (under review).

Dogan, S., Duivenvoorden, R., Grobbee, D. E., Kastelein, J. J. P., Shear, C. L., Evans, G. W., Visseren, F. L., et al. (2010). Ultrasound protocols to measure carotid intima-media thickness in trials; comparison of reproducibility, rate of progression, and effect of intervention in subjects with familial hypercholesterolemia and subjects with mixed dyslipidemia. *Annals of Medicine*, *42*(6), 447–464.

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PloS Genetics, 9*(3): e1003348.

Ebrahim, S., & Davey Smith, G. (2008). Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics*, *123*(1), 15–33.

Ebrahim, S., Lawlor, D. A., Shlomo, Y. Ben, Timpson, N., Harbord, R., Christensen, M., Baban, J., et al. (2008). Alcohol dehydrogenase type 1C (ADH1C) variants, alcohol consumption traits, HDL-cholesterol and risk of coronary heart disease in women and men: British Women's Heart and Health Study and Caerphilly cohorts. *Atherosclerosis*, *196*(2), 871–878.

Epstein, N. D., Lin, H. J., & Fananapazir, L. (1990). Genetic evidence of dissociation (generational skips) of electrical from morphologic forms of hypertrophic cardiomyopathy. *The American Journal of Cardiology*, *66*(5), 627–631.

Espeland, M. A., O'leary, D. H., Terry, J. G., Morgan, T., Evans, G., & Mudra, H. (2005). Carotid intimal-media thickness as a surrogate for cardiovascular disease events in trials of HMG-CoA reductase inhibitors. *Current Controlled Trials in Cardiovascular Medicine*, *6*(1), 3.

Feher, M. D., & Richmond, W. (2006). *Lipids and Lipid Disorders*. (A. Taylor, Ed.) (3rd ed., pp. 1–31). London: Excerpta Medica Publications.

Ference, B. A., Yoo, W., Alesh, I., Mahajan, N., Mirowska, K. K., Mewada, A., Kahn, J., et al. (2012). 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. *Journal of the American College of Cardiology*, *60*(25), 2631–2639.

Fokstuen, S., Munoz, A., Melacini, P., Iliceto, S., Perrot, A., Ozcelik, C., Jeanrenaud, X., et al. (2011). Rapid detection of genetic variants in hypertrophic cardiomyopathy by custom DNA resequencing array in clinical practice. *Journal of Medical Genetics*, *48*(8), 572–576.

Foppa, M., Duncan, B. B., & Rohde, L. E. P. (2005). Echocardiography-based left ventricular mass estimation. How should we define hypertrophy? *Cardiovascular Ultrasound*, *3*, 17.

Forrest, M. J., Bloomfield, D., Briscoe, R. J., Brown, P. N., Cumiskey, A.-M., Ehrhart, J., Hershey, J. C., et al. (2008). Torcetrapib-induced blood pressure elevation is independent of CETP inhibition and is accompanied by increased circulating levels of aldosterone. *British Journal of Pharmacology*, *154*(7), 1465–1473.

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, *10*(4), 241–251.

Friedewald, W. T., Levy, R. I., & Fredrickson, D. S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*, *18*(6), 499–502.

Ganesh, S. K., Tragante, V., Guo, W., Guo, Y., Lanktree, M. B., Smith, E. N., Johnson, T., et al. (2013). Loci influencing blood pressure identified using a cardiovascular gene-centric array. *Human Molecular Genetics*, *22*(8), 1663-1678.

Geisterfer-Lowrance, A. A. T., Kass, S., Tanigawa, G., Vosberg, H.-P., McKenna, W., Seidman, C. E., & Seidman, J. G. (1990). A molecular basis for familial hypertrophic cardiomyopathy: A β cardiac myosin heavy chain gene missense mutation. *Cell*, *62*(5), 999–1006.

Gertow, K., Sennblad, B., Strawbridge, R. J., Ohrvik, J., Zabaneh, D., Shah, S., Veglia, F., et al. (2012). Identification of the BCAR1-CFDP1-TMEM170A locus as a determinant of carotid intima-media thickness and coronary artery disease risk. *Circulation Cardiovascular Genetics*, *5*(6), 656–665.

Gharavi, A. G., Lipkowitz, M. S., Diamond, J. A., Jhang, J. S., & Phillips, R. A. (1996). Deletion polymorphism of the angiotensin-converting enzyme gene is independently associated with left ventricular mass and geometric remodeling in systemic hypertension. *The American Journal of Cardiology*, *77*(15), 1315–1319.

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nature Genetics*, *42*(7), 558–560.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, *13*(2), 135–145.

Ginsberg, H. N., Elam, M. B., Lovato, L. C., Crouse, J. R., Leiter, L. A., Linz, P., Friedewald, W. T., et al. (2010). Effects of combination lipid therapy in type 2 diabetes mellitus. *The New England Journal of Medicine*, *362*(17), 1563–1574.

Gomez-Angelats, E., De la Sierra, A., Enjuto, M., Sierra, C., Oriola, J., Francino, A., Paré, J. C., et al. (2000). Lack of association between ACE gene polymorphism and left ventricular hypertrophy in essential hypertension. *Journal of Human Hypertension*, *14*(1), 47–49.

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, *44*(10), 1084–1089.

Gurha, P., Abreu-Goodger, C., Wang, T., Ramirez, M. O., Drumond, A. L., Van Dongen, S., Chen, Y., et al. (2012). Targeted deletion of microRNA-22 promotes stress-induced cardiac dilation and contractile dysfunction. *Circulation*, *125*(22), 2751–2761.

HapMap Consortium, I. (2003). The International HapMap Project. *Nature*, *426*(6968), 789–796.

Harismendy, O., Notani, D., Song, X., Rahim, N. G., Tanasa, B., Heintzman, N., Ren, B., et al. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature*, *470*(7333), 264–268.

Harper, A. R., Mayosi, B. M., Rodriguez, A., Rahman, T., Hall, D., Mamasoula, C., Avery, P. J., et al. (2013). Common variation neighbouring micro-RNA 22 is associated with increased left ventricular mass. *PloS One*, *8*(1), e55061.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A., et al. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*, *316*(5830), 1491–1493.

Hense, H.-W., Schulte, H., Löwel, H., Assmann, G., & Keil, U. (2003). Framingham risk function overestimates risk of coronary heart disease in men and women from Germany-results from the MONICA Augsburg and the PROCAM cohorts. *European Heart Journal*, *24*(10), 937–945.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Britsih Medical Journal*, *327*(7414), 557–560.

Hill, J. A. (2003). Electrical remodeling in cardiac hypertrophy. *Trends in Cardiovascular Medicine*, *13*(8), 316–322.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *British Medical Journal*, *335*(7611), 136-141.

Hiukka, A., Westerbacka, J., Leinonen, E. S., Watanabe, H., Wiklund, O., Hulten, L. M., Salonen, J. T., et al. (2008). Long-term effects of fenofibrate on carotid intima-media thickness and augmentation index in subjects with type 2 diabetes mellitus. *Journal of the American College of Cardiology*, *52*(25), 2190–2197.

Holdt, L. M., & Teupser, D. (2012). Recent studies of the human chromosome 9p21 locus, which is associated with atherosclerosis in human populations. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *32*(2), 196–206.

Holm, H., Gudbjartsson, D. F., Arnar, D. O., Thorleifsson, G., Thorgeirsson, G., Stefansdottir, H., Gudjonsson, S. A., et al. (2010). Several common variants modulate heart rate, PR interval and QRS duration. *Nature Genetics*, *42*(2), 117–122.

Illumina. (2005). GenCall. Available from http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf. [Accessed 8 June 2011].

Jackson, R. (2008). Cardiovascular risk prediction: are we there yet? *Heart*, *94*(1), 1–3.

Jarcho, J. A., McKenna, W., Pare, J. A., Solomon, S. D., Holcombe, R. F., Dickie, S., Levi, T., et al. (1989). Mapping a gene for familial hypertrophic cardiomyopathy to chromosome 14q1. *The New England Journal of Medicine*, *321*(20), 1372–1378.

Jentzsch, C., Leierseder, S., Loyer, X., Flohrschütz, I., Sassi, Y., Hartmann, D., Thum, T., et al. (2012). A phenotypic screen to identify hypertrophy-modulating microRNAs in primary cardiomyocytes. *Journal of Molecular and Cellular Cardiology*, *52*(1), 13–20.

Johansen, C. T., Wang, J., Lanktree, M. B., Cao, H., McIntyre, A. D., Ban, M. R., Martins, R. A., et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics*, *42*(8), 684–687.

Jun, M., Foote, C., Lv, J., Neal, B., Patel, A., Nicholls, S. J., Grobbee, D. E., et al. (2010). Effects of fibrates on cardiovascular outcomes: a systematic review and meta-analysis. *Lancet*, *375*(9729), 1875–1884.

Kannel, W. B., Dannenberg, A. L., & Levy, D. (1987). Population implications of electrocardiographic left ventricular hypertrophy. *The American Journal of Cardiology*, *60*(17), 85I–93I.

Kastelein, J. J. P., Wiegman, A., & De Groot, E. (2003). Surrogate markers of atherosclerosis: impact of statins. *Atherosclerosis, 4*(1), 31–36.

Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N. P., Roos, C., Hirschhorn, J. N., et al. (2008). Polymorphisms associated with cholesterol and risk of cardiovascular events. *The New England Journal of Medicine*, *358*(12), 1240–1249.

Kathiresan, S., & Srivastava, D. (2012). Genetics of Human Cardiovascular Disease. *Cell*, *148*(6), 1242–1257.

Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., et al. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics*, *41*(3), 334–341.

Kauma, H., Ikäheimo, M., Savolainen, M. J., Kiema, T. R., Rantala, A. O., Lilja, M., Reunanen, A., et al. (1998). Variants of renin-angiotensin system genes and echocardiographic left ventricular mass. *European Heart Journal*, *19*(7), 1109–1117.

KBioSciences. (2007). KASPar® SNP genotyping on demand Assay design, manufacture, testing and optimisation service. Available from: http://www.kbioscience.co.uk/reagents/KB_PP_KASPar_ondemand.pdf. [Accessed 20 May 2013].

Keating, B. J., Tischfield, S., Murray, S. S., Bhangale, T., Price, T. S., Glessner, J. T., Galver, L., et al. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PloS One*, *3*(10), e3583.

Kent, D. M. (2002). The Framingham scores overestimated the risk for coronary heart disease in Japanese, Hispanic, and native American cohorts. *Evidence-Based Medicine*, *7*(1), 31.

Kivimäki, M., Lawlor, D. A., Davey Smith, G., Kumari, M., Donald, A., Britton, A., Casas, J. P., et al. (2008). Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II Study. *PloS One*, *3*(8), e3013.

Kleiber, C., & Zeileis, A. (2010). AER: Applied Econometrics with R. Available from http://cran.r-project.org/package=AER. [Accessed 17 April 2012]

Knezevic, I., Patel, A., Sundaresan, N. R., Gupta, M. P., Solaro, R. J., Nagalingam, R. S., & Gupta, M. (2012). A novel cardiomyocyte enriched microRNA, miR-378, targets IGF1R: implications in post natal cardiac remodeling and cell survival. *The Journal of Biological Chemistry*, *287*(16), 12913-12926.

Knöll, R., Buyandelger, B., & Lab, M. (2011). The sarcomeric Z-disc and Z-discopathies. *Journal of Biomedicine & Biotechnology*, *2011*, 569628.

Krauss, R. M. (2008). What can the genome tell us about LDL cholesterol? *Lancet*, *371*(9611), 450–452.

Krauss, R. M. (2010). Lipoprotein subfractions and cardiovascular disease risk. *Current Opinion in Lipidology*, *21*(4), 305–311.

Kundu, S., Aulchenko, Y. S., Van Duijn, C. M., & Janssens, A. C. J. W. (2011). PredictABEL: an R package for the assessment of risk prediction models. *European Journal of Epidemiology*, *26*(4), 261–264.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–1831.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838.

LaRosa, J. C., He, J., & Vupputuri, S. (1999). Effect of statins on risk of coronary disease: a meta-analysis of randomized controlled trials. *Journal of the American Medical Association*, *282*(24), 2340–2346.

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602.

Law, M. R., & Wald, N. J. (2002). Risk factor thresholds: their existence under scrutiny. *British Medical Journal*, *324*(7353), 1570–1576.

Lawless, J. F., & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics*, *5*, 307–323.

Lawlor, D A, Bedford, C., Taylor, M., & Ebrahim, S. (2003). Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. *Journal of Epidemiology & Community Health*, *57*(2), 134–140.

Lawlor, D., Harbord, R., Sterne, J., Timpson, N., & Davey-Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 1133–1163.

Lawlor, Debbie A, Davey Smith, G., Kundu, D., Bruckdorfer, K. R., & Ebrahim, S. (2004). Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *Lancet*, *363*(9422), 1724–1727.

Lawlor, Debbie A, Harbord, R. M., Timpson, N. J., Lowe, G. D. O., Rumley, A., Gaunt, T. R., Baker, I., et al. (2008). The association of C-reactive protein and CRP genotype with coronary heart disease: findings from five studies with 4,610 cases amongst 18,637 participants. *PloS One*, *3*(8), e3011.

Levy, D, Garrison, R. J., Savage, D. D., Kannel, W. B., & Castelli, W. P. (1989). Left ventricular mass and incidence of coronary heart disease in an elderly cohort. The Framingham Heart Study. *Annals of Internal Medicine*, *110*(2), 101–107.

Levy, D, Garrison, R. J., Savage, D. D., Kannel, W. B., & Castelli, W. P. (1990). Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. *The New England Journal of Medicine*, *322*(22), 1561–1566.

Levy, Daniel, Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., Glazer, N. L., et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics*, *41*(6), 677–687.

Lewis, S. J. (2010). Mendelian randomization as applied to coronary heart disease, including recent advances incorporating new technology. *Circulation Cardiovascular Genetics*, *3*(1), 109–117.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Libby, P., Ridker, P. M., & Hansson, G. K. (2011). Progress and challenges in translating the biology of atherosclerosis. *Nature*, *473*(7347), 317–325.

Linsel-Nitschke, P., Götz, A., Erdmann, J., Braenne, I., Braund, P., Hengstenberg, C., Stark, K., et al. (2008). 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*, *3*(8), e2986.

Lloyd-Jones, D. M. (2010). Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation*, *121*(15), 1768–1777.

Lorenz, M. W., Polak, J. F., Kavousi, M., Mathiesen, E. B., Völzke, H., Tuomainen, T.-P., Sander, D., et al. (2012). Carotid intima-media thickness progression to predict cardiovascular events in the general population (the PROG-IMT collaborative project): a meta-analysis of individual participant data. *Lancet*, *379*(9831), 2053–2062.

Lusis, A. J. (2012). Genetics of atherosclerosis. *Trends in Genetics, 28*(6), 267-275.

Macfarlane, P., Devine, B., & Clark, E. (2005). The University of Glasgow (Uni-G) ECG Analysis Program. *Computers in Cardiology*, *32*, 451–454.

MacMahon, S., Sharpe, N., Gamble, G., Hart, H., Scott, J., Simes, J., & White, H. (1998). Effects of lowering average of below-average cholesterol levels on the progression of carotid atherosclerosis: results of the LIPID Atherosclerosis Substudy. LIPID Trial Research Group. *Circulation*, *97*(18), 1784–1790.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

Marenberg, M. E., Risch, N., Berkman, L. F., Floderus, B., & De Faire, U. (1994). Genetic susceptibility to death from coronary heart disease in a study of twins. *The New England Journal of Medicine*, *330*(15), 1041–1046.

Marian, A. J., & Belmont, J. (2011). Strategic Approaches to Unraveling Genetic Causes of Cardiovascular Diseases. *Circulation Research*, *108*, 1252–1269.

Marks, D., Thorogood, M., Neil, H. A. W., & Humphries, S. E. (2003). A review on the diagnosis, natural history, and treatment of familial hypercholesterolaemia. *Atherosclerosis*, *168*(1), 1–14.

Marmot, M., & Brunner, E. (2005). Cohort Profile: the Whitehall II study. *International Journal of Epidemiology*, *34*(2), 251–256.

Maron, B. J., Gardin, J. M., Flack, J. M., Gidding, S. S., Kurosaki, T. T., & Bild, D. E. (1995). Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. *Circulation*, *92*(4), 785–789.

Mateu, E., Calafell, F., Ramos, M. D., Casals, T., & Bertranpetit, J. (2002). Can a place of origin of the main cystic fibrosis mutations be identified? *American Journal of Human Genetics*, *70*(1), 257–264.

Mayes, P. A., & Botham, K. M. (2012). Cholesterol Synthesis,Transport, & Excretion. *Harper's Illustrated Biochemistry* (29th ed., pp. 219–230). McGraw-Hill Medical.

Mayosi, B. (2002). Electrocardiographic measures of left ventricular hypertrophy show greater heritability than echocardiographic left ventricular mass. *European Heart Journal*, *23*(24), 1963–1971.

Mayosi, B. M., Avery, P. J., Farrall, M., Keavney, B., & Watkins, H. (2008). Genome-wide linkage analysis of electrocardiographic and echocardiographic left ventricular hypertrophy in families with hypertension. *European Heart Journal*, *29*(4), 525–530.

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

McMullen, J. R., Shioi, T., Huang, W.-Y., Zhang, L., Tarnavski, O., Bisping, E., Schinke, M., et al. (2004). The insulin-like growth factor 1 receptor induces physiological heart growth via the phosphoinositide 3-kinase(p110alpha) pathway. *The Journal of Biological Chemistry*, *279*(6), 4782–4793.

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., et al. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science*, *316*(5830), 1488–1491.

Molloy, T. J., Okin, P. M., Devereux, R. B., & Kligfield, P. (1992). Electrocardiographic detection of left ventricular hypertrophy by the simple QRS voltage-duration product. *Journal of the American College of Cardiology*, *20*(5), 1180–1186.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). MultiCollinearity. In D. J. Balding, N. A. C. Cressie, G. M. Fitzmaurice, H. Goldstein, I. M. Johnstone, G. Molenberghs, D. W. Scott, et al. (Eds.), *Introduction to Linear Regression Analysis* (Fifth., pp. 285–320). Wiley.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, *464*(7289), 773–777.

Morgan, J., Carey, C., Lincoff, A., & Capuzzi, D. (2004). High-density lipoprotein subfractions and risk of coronary artery disease. *Current Atherosclerosis Reports*, *6*(5), 359–365.

Murray, A., Cluett, C., Bandinelli, S., Corsi, A. M., Ferrucci, L., Guralnik, J., Singleton, A., et al. (2009). Common lipid-altering gene variants are associated with therapeutic intervention thresholds of lipid levels in older people. *European Heart Journal*, *30*(14), 1711–1719.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V, Li, X., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, *466*(7307), 714–719.

Nashef, S. A., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., & Salamon, R. (1999). European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery*, *16*(1), 9–13.

National Institute for Health and Clinical Exellence. (2010). Prevention of cardiovascular disease at population level (PH25). Available from http://guidance.nice.org.uk/PH25. [Accessed 18 October 2012].

NHLBI Communications. (2011). *NIH stops clinical trial on combination cholesterol treatment*. Available from http://www.nih.gov/news/health/may2011/nhlbi-26.htm. [Accessed 16 November 2011].

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics*, *7*(2), e1002003.

O'Leary, D. H., Polak, J. F., Kronmal, R. A., Manolio, T. A., Burke, G. L., & Wolfson, S. K. (1999). Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults. Cardiovascular Health Study Collaborative Research Group. *The New England Journal of Medicine*, *340*(1), 14–22.

Okin, P. M., Roman, M. J., Devereux, R. B., & Kligfield, P. (1995). Electrocardiographic identification of increased left ventricular mass by simple voltage-duration products. *Journal of the American College of Cardiology*, *25*(2), 417–423.

Osio, A., Tan, L., Chen, S. N., Lombardi, R., Nagueh, S. F., Shete, S., Roberts, R., et al. (2007). Myozenin 2 is a novel gene for human hypertrophic cardiomyopathy. *Circulation Research*, *100*(6), 766–768.

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., Smith, G. D., et al. (2011). Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research, 21*(3), 223-242.

Pazoki, R., Wilde, A. A. M., & Bezzina, C. R. (2010). Genetic Basis of Ventricular Arrhythmias. *Current Cardiovascular Risk Reports*, *4*(6), 454–460.

Pencina, M. J., D'Agostino, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, *27*(2), 157–172.

Perticone, F., Ceravolo, R., Cosco, C., Trapasso, M., Zingone, A., Malatesta, P., Perrotti, N., et al. (1997). Deletion polymorphism of angiotensin-converting enzyme gene and left ventricular hypertrophy in southern Italian patients. *Journal of the American College of Cardiology*, *29*(2), 365–369.

Pfeufer, A., Van Noord, C., Marciante, K. D., Arking, D. E., Larson, M. G., Smith, A. V., Tarasov, K. V, et al. (2010). Genome-wide association study of PR interval. *Nature Genetics*, *42*(2), 153–159.

Pierce, B. L., Ahsan, H., & Vanderweele, T. J. (2010). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology*, *40*(3), 740–752.

Pignone, M., Phillips, C., & Mulrow, C. (2000). Use of lipid lowering drugs for primary prevention of coronary heart disease: meta-analysis of randomised trials. *British Medical Journal*, *321*(7267), 983–986.

Pilbrow, A. P., Folkersen, L., Pearson, J. F., Brown, C. M., McNoe, L., Wang, N. M., Sweet, W. E., et al. (2012). The chromosome 9p21.3 coronary heart disease risk allele is associated with altered gene expression in normal heart and vascular tissues. *PloS One*, *7*(6), e39574.

Post, W. S., & Levy, D. (1994). New developments in the epidemiology of left ventricular hypertrophy. *Current Opinion in Cardiology*, *9*(5), 534–541.

R Development Core Team. (2012). R: A Language and Environment for Statistical Computing. Available from http://www.r-project.org/. [Accessed 17 April 2012].

Rader, D. J., & Tall, A. R. (2012). The not-so-simple HDL story: Is it time to revise the HDL cholesterol hypothesis? *Nature Medicine*, *18*(9), 1344–1346.

Ramaraj, R. (2008). Hypertrophic cardiomyopathy: etiology, diagnosis, and treatment. *Cardiology in Review*, *16*(4), 172–180.

Ridker, P. M., Hennekens, C. H., Buring, J. E., & Rifai, N. (2000). C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. *The New England Journal of Medicine*, *342*(12), 836–843.

Rose, G. (1985). Sick individuals and sick populations. *International Journal of Epidemiology*, *14*(1), 32–38.

Rosenzweig, A., Watkins, H., Hwang, D. S., Miri, M., McKenna, W., Traill, T. A., Seidman, J. G., et al. (1991). Preclinical diagnosis of familial hypertrophic cardiomyopathy by genetic analysis of blood lymphocytes. *The New England Journal of Medicine*, *325*(25), 1753–1760.

Sabia, S., Kivimaki, M., Kumari, M., Shipley, M. J., & Singh-Manoux, A. (2010). Effect of Apolipoprotein E epsilon4 on the association between health behaviors and cognitive function in late midlife. *Molecular Neurodegeneration*, *5*, 23.

Sacco, R. L., Blanton, S. H., Slifer, S., Beecham, A., Glover, K., Gardener, H., Wang, L., et al. (2009). Heritability and linkage analysis for carotid intima-media thickness: the family study of stroke risk and carotid atherosclerosis. *Stroke, 40*(7), 2307–2312.

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R. J., et al. (2007). Genomewide association analysis of coronary artery disease. *The New England Journal of Medicine*, *357*(5), 443–453.

Sarwar, N., Sandhu, M. S., Ricketts, S. L., Butterworth, A. S., Di Angelantonio, E., Boekholdt, S. M., Ouwehand, W., et al. (2010). Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *Lancet*, *375*(9726), 1634–1639.

Satoh, M., Takahashi, M., Sakamoto, T., Hiroe, M., Marumo, F., & Kimura, A. (1999). Structural analysis of the titin gene in hypertrophic cardiomyopathy: identification of a novel disease gene. *Biochemical and Biophysical Research Communications*, *262*(2), 411–417.

Schillaci, G., Verdecchia, P., Porcellati, C., Cuccurullo, O., Cosco, C., & Perticone, F. (2000). Continuous relation between left ventricular mass and cardiovascular risk in essential hypertension. *Hypertension*, *35*(2), 580–586.

Schork, N. J. (1997). Genetics of complex disease: approaches, problems, and solutions. *American Journal of Respiratory and Critical Care Medicine*, *156*(4), S103–S109.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, *43*, 333–338.

Schwartz, G. G., Olsson, A. G., Ballantyne, C. M., Barter, P. J., Holme, I. M., Kallend, D., Leiter, L. A., et al. (2009). Rationale and design of the dal-OUTCOMES trial: efficacy and safety of dalcetrapib in patients with recent acute coronary syndrome. *American Heart Journal*, *158*(6), 896–901.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Seidman, C. E., Seidman, J., Robbins, J., & Watkins, H. (2011). Identifying Sarcomere Gene Mutations in Hypertrophic Cardiomyopathy. *Circulation Research*, *108*, 743–750.

Shea, J., Agarwala, V., Philippakis, A. A., Maguire, J., Banks, E., Depristo, M., Thomson, B., et al. (2011). Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nature Genetics*, *43*(8), 801–805.

Sheehan, N. A., Didelez, V., Burton, P. R., & Tobin, M. D. (2008). Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine*, *5*(8), e177.

Simon Broome Register Group. (1991). Risk of fatal coronary heart disease in familial hypercholesterolaemia. Scientific Steering Committee on behalf of the Simon Broome Register Group. *British Medical Journal*, *303*(6807), 893–896.

Smilde, T. D. J., Asselbergs, F. W., Hillege, H. L., Voors, A. A., Kors, J. A., Gansevoort, R. T., Van Gilst, W. H., et al. (2005). Mild renal dysfunction is associated with electrocardiographic left ventricular hypertrophy. *American Journal of Hypertension*, *18*(3), 342–347.

Smilde, T., Van Wissen, S., Awollersheim, H., Trip, M., Kastelein, J., & Stalenhoef, A. (2001). Effect of aggressive versus conventional lipid lowering on atherosclerosis progression in familial hypercholesterolemia (ASAP): a prospective, randomised, double-blind trial. *Lancet*, *357*(9256), 577–581.

Smith, G. D., & Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, *33*(1), 30–42.

Sokolow, M., & Lyon, T. P. (1949). The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads. *American Heart Journal*, *37*(2), 161–186.

Song, L., & Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb.prot5384.

Sotoodehnia, N., Isaacs, A., De Bakker, P. I. W., Dörr, M., Newton-Cheh, C., Nolte, I. M., Van der Harst, P., et al. (2010). Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nature Genetics*, *42*(12), 1068–1076.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948.

Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5), e1000477.

Staiger D., & Stock J. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65,* 557-86.

Stock, J. H. (2001). Instrumental Variables in Statistics and Econometrics. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 7577–7582). Elsevier Ltd.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genetics*, *8*(4), e1002639.

Sun, X., Hoage, T., Bai, P., Ding, Y., Chen, Z., Zhang, R., Huang, W., et al. (2009). Cardiac hypertrophy involves both myocyte hypertrophy and hyperplasia in anemic zebrafish. *PloS One*, *4*(8), e6596.

Sundström, J., Lind, L., Arnlöv, J., Zethelius, B., Andrén, B., & Lithell, H. O. (2001). Echocardiographic and electrocardiographic diagnoses of left ventricular hypertrophy predict mortality independently of each other in a population of elderly men. *Circulation*, *103*(19), 2346–2351.

Swan, L., Birnie, D. H., Padmanabhan, S., Inglis, G., Connell, J. M. C., & Hillis, W. S. (2003). The genetic determination of left ventricular mass in healthy adults. *European Heart Journal*, *24*(6), 577–582.

Talmud, P. J., Cooper, J. A., Palmen, J., Lovering, R., Drenos, F., Hingorani, A. D., & Humphries, S. E. (2008). Chromosome 9p21.3 coronary heart disease locus genotype and prospective risk of CHD in healthy middle-aged men. *Clinical Chemistry*, *54*(3), 467–474.

Talmud, P. J., Drenos, F., Shah, S., Shah, T., Palmen, J., Verzilli, C., Gaunt, T. R., et al. (2009). Gene-centric association signals for lipids and apolipoproteins identified via the HumanCVD BeadChip. *American Journal of Human Genetics*, *85*(5), 628–642.

Talmud, P. J., Hingorani, A. D., Cooper, J. A., Marmot, M. G., Brunner, E. J., Kumari, M., Kivimäki, M., et al. (2010). Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *British Medical Journal*, *340*, b4838.

Talmud, P. J., Shah, S., Whittall, R., Futema, M., Howard, P., Cooper, J. A., Harrison, S. C., et al. (2013). Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study. *Lancet*, *381*(9874), 1293–1301.

Taylor, A. J., Bots, M. L., & Kastelein, J. J. P. (2011). Vascular disease: meta-regression of CIMT trials-data in, garbage out. *Nature Reviews Cardiology*, *8*(3), 128–130.

Taylor, A. J., Sullenberger, L. E., Lee, H. J., Lee, J. K., & Grace, K. A. (2004). Arterial Biology for the Investigation of the Treatment Effects of Reducing Cholesterol (ARBITER) 2: a double-blind, placebo-controlled study of extended-release niacin on atherosclerosis progression in secondary prevention patients treated with statins. *Circulation*, *110*(23), 3512–3517.

Taylor, A. J., Villines, T. C., Stanek, E. J., Devine, P. J., Griffen, L., Miller, M., Weissman, N. J., et al. (2009). Extended-release niacin or ezetimibe and carotid intima-media thickness. *The New England Journal of Medicine*, *361*(22), 2113–2122.

Taylor, A., Wang, D., Patel, K., Whittall, R., Wood, G., Farrer, M., Neely, R. D. G., et al. (2010). Mutation detection rate and spectrum in familial hypercholesterolaemia patients in the UK pilot cascade project. *Clinical Genetics*, *77*(6), 572–580.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, *466*(7307), 707–713.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073.

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320.

The Lipid Research Clinics Coronary Primary Prevention Trial. (1984). The Lipid Research Clinics Coronary Primary Prevention Trial results. I. Reduction in incidence of coronary heart disease. *Journal of the American Medical Association*, *251*(3), 351–364.

Thomas, D. C., & Conti, D. V. (2004). Commentary: the concept of “Mendelian Randomization”. *International Journal of Epidemiology*, *33*(1), 21–25.

Thompson, J. R., Attia, J., & Minelli, C. (2011). The meta-analysis of genome-wide association studies. *Briefings in Bioinformatics*, *12*(3), 259–69.

Tobin, M. D., Sheehan, N. A., Scurrah, K. J., & Burton, P. R. (2005). Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine*, *24*(19), 2911–2935.

Tobin, M. D., Tomaszewski, M., Braund, P. S., Hajat, C., Raleigh, S. M., Palmer, T. M., Caulfield, M., et al. (2008). Common variants in genes underlying monogenic hypertension and hypotension and blood pressure in the general population. *Hypertension*, *51*(6), 1658–1664.

Tukiainen, T., Kettunen, J., Kangas, A. J., Lyytikäinen, L.-P., Soininen, P., Sarin, A.-P., Tikkanen, E., et al. (2012). Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Human Molecular Genetics*, *21*(6), 1444–1455.

Urbanek, K., Rota, M., Cascapera, S., Bearzi, C., Nascimbene, A., De Angelis, A., Hosoda, T., et al. (2005). Cardiac stem cells possess growth factor-receptor systems that after activation regenerate the infarcted myocardium, improving ventricular function and long-term survival. *Circulation Research*, *97*(7), 663–673.

Utermann, G., Langenbeck, U., Beisiegel, U., & Weber, W. (1980). Genetics of the apolipoprotein E system in man. *American Journal of Human Genetics*, *32*(3), 339–347.

Varbo, A., Benn, M., Tybjærg-Hansen, A., Grande, P., & Nordestgaard, B. G. (2011). TRIB1 and GCKR polymorphisms, lipid levels, and risk of ischemic heart disease in the general population. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *31*(2), 451–457.

Varbo, A., Benn, M., Tybjærg-Hansen, A., Jørgensen, A. B., Frikke-Schmidt, R., & Nordestgaard, B. G. (2013). Remnant cholesterol as a causal risk factor for ischemic heart disease. *Journal of the American College of Cardiology*, *61*(4), 427–436.

Vasan, R. S., Glazer, N. L., Felix, J. F., Lieb, W., Wild, P. S., Felix, S. B., Watzinger, N., et al. (2009). Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. *Journal of the American Medical Association*, *302*(2), 168–178.

Vasan, R. S., Larson, M. G., Aragam, J., Wang, T. J., Mitchell, G. F., Kathiresan, S., Newton-Cheh, C., et al. (2007). Genome-wide association of echocardiographic dimensions, brachial artery endothelial function and treadmill exercise responses in the Framingham Heart Study. *BMC Medical Genetics*, 8*(Suppl 1)*, S2.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., et al. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genetics*, *8*(8), e1002793.

Voight, B. F., Peloso, G. M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M. K., Hindy, G., et al. (2012b). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet*, *380*(9841), 572-580.

Walker, M., Whincup, P. H., & Shaper, A. G. (2004). The British Regional Heart Study 1975-2004. *International Journal of Epidemiology*, *33*(6), 1185–1192.

Wang, L., Fan, C., Topol, S. E., Topol, E. J., & Wang, Q. (2003). Mutation of MEF2A in an inherited disorder with features of coronary artery disease. *Science*, *302*(5650), 1578–1581.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, *18*, 293–297.

Weisgraber, K. H., Rall, S. C., & Mahley, R. W. (1981). Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *The Journal of Biological Chemistry*, *256*(17), 9077–9083.

Weiss, L. A., Pan, L., Abney, M., & Ober, C. (2006). The sex-specific genetic architecture of quantitative traits in humans. *Nature Genetics*, *38*(2), 218–222.

Welch, S., Plank, D., Witt, S., Glascock, B., Schaefer, E., Chimenti, S., Andreoli, A. M., et al. (2002). Cardiac-specific IGF-1 expression attenuates dilated cardiomyopathy in tropomodulin-overexpressing transgenic mice. *Circulation Research*, *90*(6), 641–648.

Weng, L., Kavaslar, N., Ustaszewska, A., Doelle, H., Schackwitz, W., Hébert, S., Cohen, J. C., et al. (2005). Lack of MEF2A mutations in coronary artery disease. *The Journal of Clinical Investigation*, *115*(4), 1016–1020.

Wensley, F., Gao, P., Burgess, S., Kaptoge, S., Di Angelantonio, E., Shah, T., Engert, J., et al. (2011). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *British Medical Journal*, *342*(2), d548.

Wienke, A., Herskind, A. M., Christensen, K., Skytthe, A., & Yashin, A. I. (2005). The heritability of CHD mortality in danish twins after controlling for smoking and BMI. *Twin Research and Human Genetics*, *8*(1), 53–59.

World Health Organisation. (2012). Cardiovascular diseases Fact sheet N°317. Available from http://www.who.int/mediacentre/factsheets/fs317/en/index.html [Accessed 12 September 2012] .

Xu, X.-D., Song, X.-W., Li, Q., Wang, G.-K., Jing, Q., & Qin, Y.-W. (2012). Attenuation of microRNA-22 derepressed PTEN to effectively protect rat cardiomyocytes from hypertrophy. *Journal of Cellular Physiology*, *227*(4), 1391–1398.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569.

Yang, T., Beazley, C., Montgomery, S. B., Dimas, A. S., Gutierrez-Arcelus, M., Stranger, B. E., Deloukas, P., et al. (2010). Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, *26*(19), 2474–2476.

Zeller, T., Blankenberg, S., & Diemert, P. (2011). Genome-Wide Association Studies in Cardiovascular Disease-An Update 2011. *Clinical Chemistry*, *58*(1), 92–103.

Zhao, J., Cheema, F. A., Bremner, J. D., Goldberg, J., Su, S., Snieder, H., Maisano, C., et al. (2008). Heritability of carotid intima-media thickness: a twin study. *Atherosclerosis*, *197*(2), 814–820.