The Genetics of Coronary Heart Disease across Ethnicities and in those with Type 2 Diabetes

Katherine Ellen Beaney

University College London

A thesis submitted in accordance with the regulations of the University College London for the degree of Doctor of Philosophy

Centre for Cardiovascular Genetics

UCL Institute of Cardiovascular Science

For my Mum

Eileen A. Matthew
1956-2014

“Seek and you shall find.” Mt 7:7

Declaration:

I, Katherine Ellen Beaney confirm that the work presented in this thesis is my own. I have indicated in the thesis where information has been derived from other sources. In particular:

In Chapter 3 - The original 19 SNP gene score was constructed by Prof Steve Humphries and Dr Fotios Drenos. Jackie Cooper calculated the original the QRISK2 CHD risk score for NPHSII participants. Genotyping of the 19 SNPs had been carried out by various members of the CVG lab prior to the start of the project. Saleem Shahid performed genotyping and analysis of the Lahore cohort (with some help from Jackie Cooper). Laurent Larifla genotyped and performed the analysis of the Afro-Caribbean cohort. In both cases this was under my guidance. I performed the literature search. I also calculated the updated gene scores and performed the required analysis for the four studies.

In Chapter 4: I performed the literature search to identify variants for inclusion in the CHD in T2D gene score. The "raw" phenotype (including outcome data) and genotype/imputation data was collected by the UCLEB consortium. The QRISK2 scores for the UCLEB participants were calculated by Jackie Cooper. I calculated that gene scores combined it with QRISK2 and performed the analysis. Jackie Cooper also analysed the association of rs10911021 with CHD and T2D-CHD risk factors in UCLEB and compared risk factors in those in UCLEB with and without T2D. I performed the meta-analysis concerning CHD and rs10911021 with the previously published and UCLEB data. The metabolomic measurements in the UCLEB samples were performed at the NMR metabolomics Laboratory at the University of Eastern Finland by the group headed by Prof Mika Ala-Korpela. Analysis was performed by me except for ET2DS were it was performed by Stela McLachlan with the assistance of Anna Price.

In Chapter 5: The CoRDia trial was devised by Prof Stan Newman and Prof Steve Humphries. The study protocol was developed by the CoRDia Steering Group. Risk factor data was collected by the recruiting primary care providers. The baseline data was entered by me and Dauda Bappa. The majority of the genotyping of the CorDia samples was performed by KaWah Li, with some genotyping performed by me and Phil Howard. I

analysed the baseline CoRDia data and compared it to other relevant data sets (NPHSII, UCLEB and UDACS).

In Chapter 6: The phenome scan in UCLEB was performed by Dr Delilah Zabaneh. Dr Lasse Folkersen performed the eQTL analysis in ASAP. I performed the *in vitro* functional assays (some of the luciferase assays were performed with the assistance of Jutta Palmen) and the other bioinformatics analysis.

Acknowledgements:

Firstly, I would like to acknowledge all of the participants of the studies involved in this project, recruited across from across the world. Without you, none of the work presented in this thesis would have been possible.

I would like to thank my supervisors, Professor Steve Humphries and Dr Fotios Drenos, for their unwavering support throughout the progress of this thesis. You generously shared your time and expertise. I would also like to thank my colleagues throughout my time at CVG – Jay, Jutta, KaWah, Dauda, Jacquie, Andrew, Roaa, Phil, Sam, Philippa, Marta, Ann, Jackie and Ruth – for all for all the many different ways you have helped this PhD. I would especially like to thank Freya, who was a tremendous support to me when my Mum was ill. Thanks also to the team at City University with whom I collaborated as part of the CoRDia trial. A very challenging experience but working with you was a pleasure. Heartfelt thanks are also extended to my funders the MRC and Randox Laboratories (and to the British Heart Foundation who funded much of CVG) for making this project possible. Thanks are also extended to my collaborators and former colleagues at Randox Laboratories – particularly Martin Crockard and Claire Ward.

On a more personal note, thank you to my family, particularly my Auntie Alison and Uncle Alistair who so kindly opened their home to me (for five months – when I said it would be six weeks maximum!). It made the period of transition at the start of this PhD much easier than it otherwise would have been. I must also extend my most sincere gratitude to my parents and my sister. My father and late mother have always given me tremendous support - emotional, spiritual and not least financial.

Finally, to my husband Matthew who was supportive from the outset. Your patience, unfailing enthusiasm and consolation have been an inspiration. “Thank you” seems far too small.

Abstract

Coronary heart disease (CHD) is the most common cause of the death worldwide, presenting a considerable burden to both individual and public health. The genetics of CHD was investigated in two contexts in this thesis - risk prediction and the identification of functional mechanisms through which associated loci affect CHD pathophysiology.

The use of a 19 single nucleotide polymorphism (SNP) CHD gene score (GS) was assessed in three ethnic groups (European, South Asian and Afro-Caribbean), but there was no strong evidence of clinical utility. A systematic literature search identified all variants robustly associated with CHD. Most of these variants were from the meta-analysis performed by the CARDIoGRAMplusC4D consortium. The GS was updated using effect sizes from this meta-analysis, resulting in improved performance. Overall, there was evidence of potential clinical utility in the European and Afro-Caribbean groups and in those with type 2 diabetes (T2D) (all $p<0.05$). However, results from the Pakistani cohorts were inconsistent. T2D-specific GSs were also assessed and were associated with CHD in the T2D group only ($p<0.05$).

Functional analysis of two risk loci was performed. Firstly, rs10911021, previously associated with CHD in T2D and this result was supported by the findings of this thesis. Counterintuitively, the CHD "protective" allele was associated with lower high density lipoprotein (HDL) cholesterol ($p=5x10^{-4}$) and lower large HDL traits (false discovery rate adjusted p-values $p<0.05$) in T2D only, indicating a complex relationship between CHD, T2D and HDL. Secondly, the CHD risk locus on chromosome 21q22 (lead SNP rs9982601) was not associated with any CHD risk factors. Using bioinformatics tools and *in vitro* functional assays, a candidate functional SNP - rs28451064 - was identified (which showed allele-specific protein binding and the minor allele had 12 % higher expression $p=4.82x10^{-3}$). Further investigation is required to define the underlying molecular pathways.

Contents

List of Figures

List of Tables

Publications resulting from this thesis:

Beaney, K.E., Cooper, J.A., Ullah Shahid, S, Ahmed, W., Qamar, R., Drenos, F., Crockard, M. A. and Humphries, S. E. (2015) "Clinical Utility of a Coronary Heart Disease Risk Prediction Gene Score in UK healthy middle aged men and in the Pakistani Population" PLoS One. 2015;10(7):e0130754

Larifla, L., Beaney, K.E., Foucan, L., Bangou, J., Michel, C. T., Martino, J., Velayoudom-Cephise, F. L., Cooper, J. A. and Humphries, S. E. (2016) "Influence of genetic risk factors on coronary heart disease occurrence in Afro-Caribbeans" Canadian Journal of Cardiology. 2016 Jan 14.

Davies, A.K.,McGale, N., Humphries, S.E., Hirani, S.P., Beaney K.E. Bappa, D. A. S., McCabe, J. G. and Newman, S. P. (2015) "Effectiveness of a self- management intervention with personalised genetic and lifestyle-related risk information on coronary heart disease and diabetes-related risk in type 2 diabetes (CoRDia): Study protocol for a randomised controlled trial". Trials 2015 Dec 2 16:547

Submitted – Under Review

Beaney, K.E., Cooper, J.A., McLachlan, S., Wannamethee, S.G., Jefferis, B.J., Whincup, P., Ben-Shlomo, Y., Price, J.F., Kumari, M., Wong, A., Ong, K., Hardy, R., Kuh, D., Kivimaki, M., Kangas, A.J., Soininen, P., Ala-Korpela, M., Drenos, F. and Humphries, S.E. on behalf of the UCLEB consortium. Variant rs10911021 that associates with coronary heart disease in type 2 diabetes, is associated with lower concentrations of circulating HDL cholesterol and large HDL particles but not with amino acids. Submitted June 2016.

In Preparation:

Beaney K.E, Ward C.E., Bappa D.A.S, McGale, N., Davies, A.K., Hirani, S.P., Li, K., Howard, P., Vance, D.V., Crockard, M. A., Lamont, J.V., Newman, S. P. and Humphries, S.E. A 19-SNP Coronary Heart Disease Gene Score profile in Subjects with Type 2 Diabetes : The Coronary heart disease Risk in type 2 DiAbetes (CoRDia study) study baseline characteristics.

Beaney, K.E., Smith, A.J.P., Palmen, J., Zabaneh, D., J.A., Wannamethee, S.G., Jefferis, B.J., Whincup, P., Morris, R., Gaunt, T., Lawlor, D.A., Casas, J.P., Kumari, M., Kivimaki, M, Langenberg, C. and Humphries, S.E. on behalf of the UCLEB consortium. Functional analysis of the CHD risk locus on choromsome 21q22.

List of Acronyms and Abbreviations

3'-UTR	3'-untranslated region
3C	Chromosome confirmation capture
4C	Chromosome confirmation capture on-chip
APS	Ammonium persulphate
ASAP	Advanced study of aortic pathology
AUROC	Area under the receiver operator characteristic
BMI	Body mass index
BRHS	British Regional Heart Study
BWHHS	British Women's Heart and Health Study
CAPS	Caerphilly Prospective Study
CEU	CEPH (Utah residents with ancestry from northern and western Europe)
CHD	Coronary heart disease
CI	Confidence interval
CoRDia	Coronary heart disease and diabetes-related risk
CRF	Conventional risk factor
CS	Catanzaro Study
CVD	Cardiovascular Disease
dIdC	poly(deoxyinosinic-deoxycytidylic) acid sodium salt
DNA	Deoxyribonucleic acid
DMEM	Dulbecco's modified eagle medium
EAS	Edinburgh Artery Study
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
EDTA	Ethylenediaminetetraacetic acid
ELSA	English Longitudinal Study of Aging
EMSA	Electrophoretic mobility shift assay
eQTL	Expression quantitative trait loci
ET2DS	Edinburgh Type 2 Diabetes Study

EUR	European cohort of the 1000 genomes project
FBS	Foetal bovine serum
FDR	False discovery rate
FE	Fixed effects
FH	Familial hypercholesterolaemia
FOXA2	Forkhead Box A2
FREAC4	Forkhead-related transcription factor 4
FRET	Fluorescence resonance energy transfer
GS	Gene score
GWAS	Genome-wide association study
FBS	Foetal bovine serum
GHS	Gargano Heart Study
Hba1c	Glycated haemoglobin
HDL	High density lipoprotein
HMX3	Homeodomain protein H6 family member 3
HPFS	Health Professionals Follow-up Study
HUVEC	Human umbilical vein endothelial cells
HWE	Hardy-Weinberg equilibrium
JHS	Joslin Heart Study
KASP	Kompetitive Allele Specific PCR
KCNE2	Potassium Channel Voltage Gated Subfamily E Regulatory Beta Subunit 2
LB	Lysogeny broth
LD	Linkage disequilibrium
LDL	Low density lipoprotein
MDC	Malmo Diet and Cancer Study
MI	Myocardial infarction
MRC1946	Medical Research Council birth cohort 1946
MRPS6	Mitochondrial ribosomal protein S6
NHGRI	National Human Genome Research Institute
NHS	Nurses' Health Study
NICE	National Institute for Health and Clinical Excellence

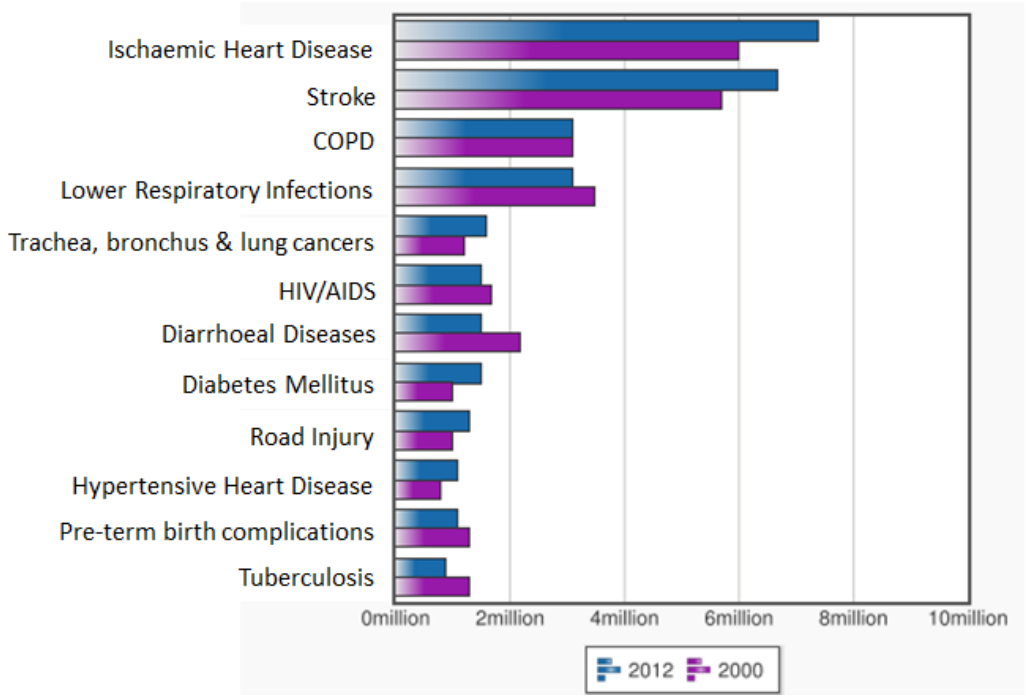
NMR	Nuclear magnetic resonance
NPHSII	Northwick Park Heart Study II
NRI	Net reclassification index
NTC	No template control
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
OR	Odds ratio
RAF	Risk allele frequency
RCT	Randomised controlled trial
RE	Random effects
RNA	Ribonucleic acid
ROC	Receiver operator characteristic
ROS	Reactive oxygen species
RXR	Retinoid X receptor
SD	Standard deviation
SE	Standard error
SLC5A3	Sodium-myoinositol co-transporter
SMI	Self-management intervention
SNP	Single nucleotide polymorphism
SP1	Specificity protein 1
T2D	Type 2 diabetes
TBE	Tris borate EDTA
TEMED	Tetramethylethylenediamine
UCLEB	University College, London School of Hygiene and Tropical Medicine, Edinburgh and Bristol Consortium
UCLH	University College London Hospitals
UCSC	University of California, Santa Cruz
UDACS	University College Diabetes and Cardiovascular Study
UKPDS	UK Prospective Diabetes Study
VDR	Vitamin D receptor
WHII	Whitehall II Study
WHO	World Health Organisation

1 General Introduction

1.1 Cardiovascular disease – a global problem

Cardiovascular disease (CVD) encompasses a number of different conditions including coronary heart disease (CHD) and stroke. It was the most common cause of death in the world between 2000 and 2012 (Figure 1). World Health Organisation (WHO) data for 2012 (the most recent year where data is available) showed that approximately 30% of all deaths were caused by CVD and the highest proportion of these was due to CHD (http://www.who.int/mediacentre/factsheets/fs310/en/). This is despite significant improvements in the treatments available (Nabel and Braunwald 2012). Reducing the number of individuals affected by CVD, and CHD in particular, is therefore an important healthcare issue worldwide.

Figure 1: Comparison of the leading causes of death worldwide between 2000 and 2012

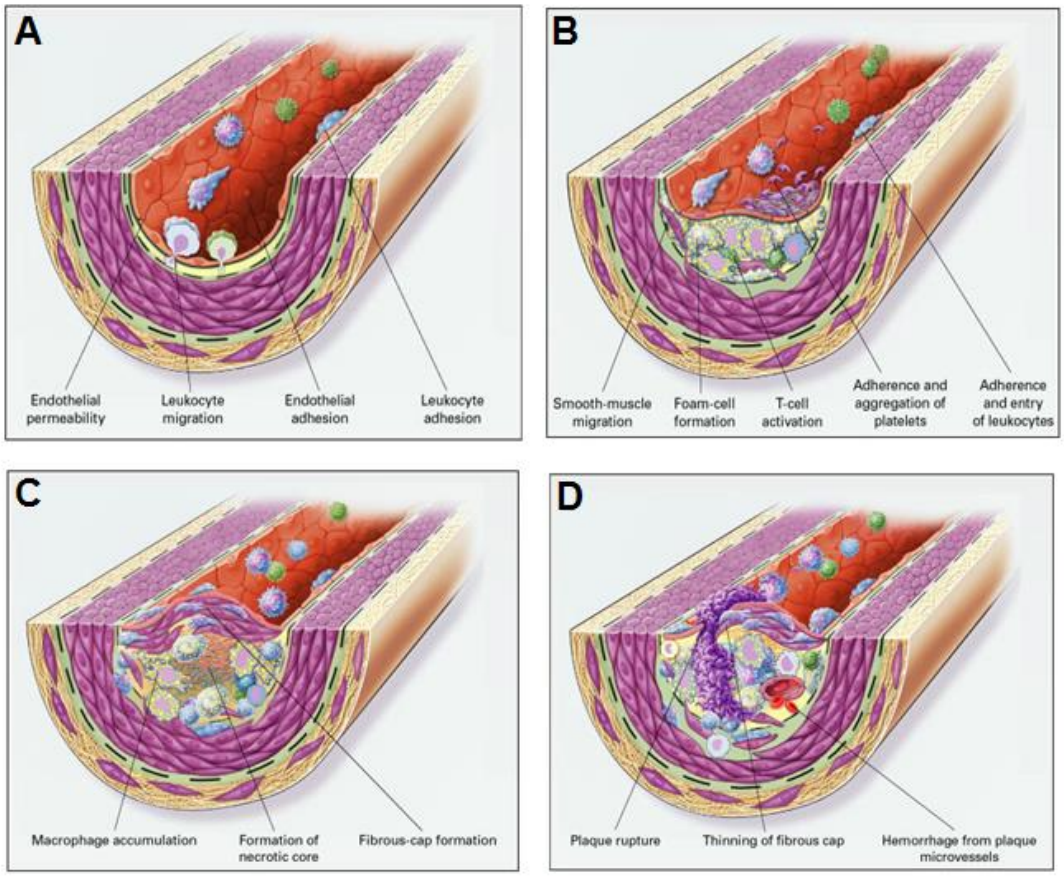


Data collected by and image adapted from the World Health Organisation (http://www.who.int/mediacentre/factsheets/fs310/en/). COPD = Chronic Obstructive Pulmonary Disease. HIV= Human Immunodeficiency Virus. AIDS=Acquired Immune Deficiency Syndrome.

1.2 Pathophysiology of CHD

CHD is characterised by the narrowing of the blood vessels which supply the heart, usually caused by the chronic inflammatory process, atherosclerosis (Figure 2). Atherosclerosis is initiated by endothelial dysfunction where the endothelial cells lining the arterial wall irreversibly lose their homeostatic capabilities (Mudau, Genis et al. 2012). This does not occur uniformly throughout the vasculature, but preferentially at regions with disturbed blood flow such as branch points (VanderLaan, Reardon et al. 2004). The vessel wall becomes more permeable at these sites, allowing low-density lipoprotein (LDL) to accumulate in the arterial wall (Weber and Noels 2011). Oxidisation of lipids in the vessel wall stimulates the dysfunctional endothelial cells to express chemokines, leukocyte adhesion molecules and endothelial adhesion molecules. This promotes the attachment of leukocytes and their transmigration into the arterial wall (Libby and Theroux 2005; Weber and Noels 2011). Monocytes entering the vessel wall can then differentiate into macrophages. Oxidised LDL can interact with these macrophages, causing the release of cytokines, further contributing to the inflammatory process (Businaro, Tagliani et al. 2012). Moreover, the macrophages can become lipid-rich foam cells though phagocytosis of the lipids present in the arterial wall (Nakashima, Fujii et al. 2007). The build-up of monocytes, foam cells and T-lymphocytes forming “fatty streaks” on the vessel wall is usually the first visible manifestation of atherosclerosis. Subsequent accumulation of immune cells, lipids, debris and apoptotic cells results in the development of an atherosclerotic plaque (Ross 1999). Such plaques are covered by a fibrous cap and may have a necrotic core (Weber and Noels 2011). Expansion of the plaque does not necessarily result in a reduction in the diameter of the vessel lumen. The vessel walls can “remodel” to compensate for the presence of the plaque, maintaining the blood carrying capacity of the artery (Glagov, Weisenberg et al. 1987). Most plaques will remain subclinical and the difference between plaques that do and those that do not cause symptoms has been the focus of much research (Falk, Nakano et al. 2013). It is well established that plaques with uniformly dense caps tend to be much more stable. However, thinning of the fibrous cap, usually due to infiltration of immune cells at the “shoulder” of the plaque, can result in rupture. The resultant release of the plaque contents into the coronary artery can occlude the vessel, causing myocardial infarction (MI) (Ross 1999; Weber and Noels 2011).

Figure 2: Schematic representation of the progression of an atherosclerotic plaque.



Reproduced with permission from (Ross 1999). Copyright Massachusetts Medical Society. A) Endothelial dysfunction characterised by increased permeability results in the uptake of plasma constituents (most notably LDL particles) into the intima, promoting a pro-inflammatory response. This leads to adhesion of blood leukocytes and ultimately their migration into the intima. B) The "fatty streak" progresses due to migration of smooth muscle cells (SMCs) into the intima along with T-cell activation, platelet aggregation and formation of foam cells through the uptake of oxidised LDL particles by macrophages. C) With the continued inflammatory response and accumulation of cells and debris, a complicated atherosclerotic lesion can develop, covered by a fibrous cap. Within the lesion, the build-up of extracellular lipids derived from apoptotic or necrotic macrophages and SMCs can result in the formation of a necrotic core. D) Thinning of the lesion's fibrous cap can ultimately lead to its rupture and thrombus formation with complete occlusion of the blood vessel (Ross 1999; Libby, Ridker et al. 2011).

1.3 Risk factors for CHD

1.3.1 Identification of CHD risk factors

CHD and atherosclerosis are commonly thought of as issues of later life. However, it has become clear that atherogenesis can begin early in life. “Fatty streaks” - the precursors of atherosclerotic plaques - have been found in the arteries of teenagers (Strong, Malcom et al. 1999). As the atherosclerotic process can begin many decades before clinical manifestation, this provides a window of time where preventative measures can be employed to avoid its escalation (Falk, Nakano et al. 2013). In order to do this, it is vital to understand the factors that predispose individuals to the development of CHD. From the middle of the last century onwards, numerous prospective studies have been performed to investigate this (and indeed are on-going). One of the most well-known of these is the Framingham Heart Study which started in 1948, following over 5000 residents of Framingham, Massachusetts (Dawber, Meadors et al. 1951). An early publication from the Framingham study reported an association between increased blood pressure and blood cholesterol levels and CHD and MI incidence (Kannel, Kagan et al. 1961). Since then a number of important so-called “conventional risk factors” (CRFs) for CHD have been identified and are shown in Table 1. CHD risk factors can be divided into two categories, those that cannot be modified such as age, family history of early CHD and sex, and those that can be, such as smoking status, LDL-cholesterol levels and blood pressure. It is these modifiable risk factors that clinicians seek to target in order to reduce the incidence of CHD. This strategy has been shown to be effective. From the 1980s onwards CHD mortality fell has fallen in the USA and Western Europe and more than half of this reduction has been due to preventative measures, particularly smoking cessation and reduction of saturated fat in the diet (Capewell and O'Flaherty 2011; Perk, De Backer et al. 2012).

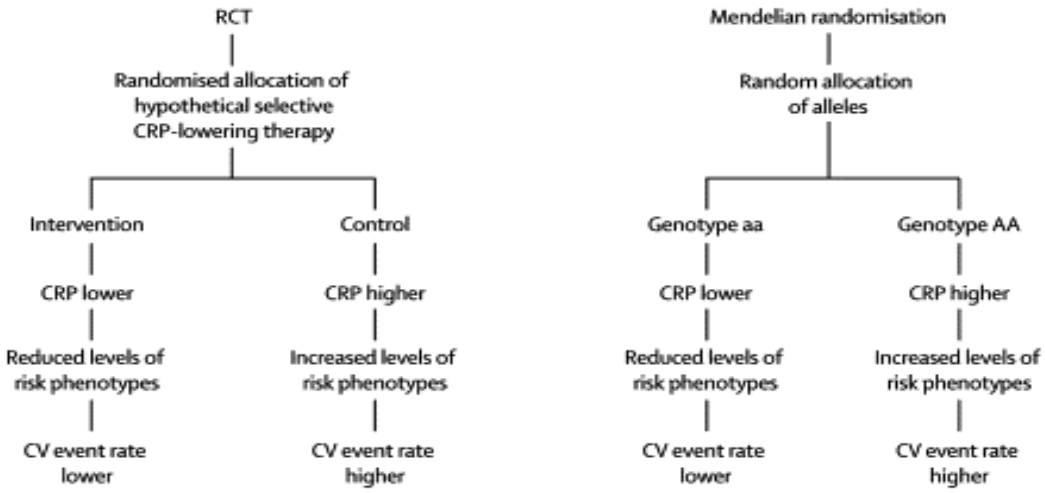
Table 1: Conventional risk factors for CHD

Risk Factor	Epidemiological Evidence
Age	(Jousilahti, Vartiainen et al. 1999)
Sex	(Lloyd-Jones, Larson et al. 1999)
Smoking	(Nyboe, Jensen et al. 1991)
Diabetes	(Huxley, Barzi et al. 2006)
Blood Pressure/Hypertension	(Lloyd-Jones, Evans et al. 2005)
Cholesterol	(Di Angelantonio, Sarwar et al. 2009)
Triglycerides	(Sarwar, Danesh et al. 2007)
Family History of CHD	(Pohjola-Sintonen, Rissanen et al. 1998)

1.3.2 Causality of risk factors

Dozens of risk factors have been identified for CHD, often from case-control or purely observational data. These study designs can be subject to bias from reverse causation or confounding. Therefore, is not possible to determine if the risk factor has a *causal* relationship with CHD. However, there are methods that can be used to assess this which are particularly useful for biomarkers. Firstly, if there is a therapeutic agent that can target the risk factor, it can be determined in a randomised controlled trial (RCT) whether raising or lowering the risk factor (depending on the proposed relationship) reduces the incidence of CHD. This has long been considered the gold-standard for establishing causality. The most pertinent example for CHD is the use of statins, which have been found to lower LDL-cholesterol and also reduce the incidence of MI and CHD in multiple RCTs (Taylor, Huffman et al. 2013). However, RCTs are time-consuming, expensive and require a selective agent however advances made over the past two decades allow genetic studies to be used to assess causality. This study design, referred to as "Mendelian randomisation" (Lawlor, Harbord et al. 2008), requires the identification of a genetic variant associated with the potential risk factor under investigation (but not with any potential confounders). In the study population individuals are then separated by genotype, firstly to confirm there is a relationship between the variant and the potential risk factor. This being so, if there is a causal relationship between the potential risk factor and the disease outcome, the genetic variant should also be associated with the disease outcome (Figure 3). The allocation of alleles at birth is considered to be analogous to the randomisation procedure in an RCT, thus minimising the impact of potential confounders and reverse causation (Lawlor, Harbord et al. 2008). Applicability of this approach is limited to traits for which there is a suitable genetic instrument. Furthermore, suitable genetic variants generally elicit a relatively small effect on the trait in question. Therefore, a very large number of participants are required to perform the study, which can prove difficult to obtain. Mendelian randomisation has been used to assess the causality for a number of potential CHD risk factors (Table 2).

Figure 3: Workflow of A) randomised clinical trial and B) Mendelian randomisation study using the example of C-reactive protein



Only homozygote genotypes are shown for clarity. RCT=randomised clinical trial, CV=cardiovascular, CRP=C-reactive protein. Image reprinted from (Hingorani and Humphries 2005).

Table 2: A selection of potential coronary heart disease risk factors assessed for causality

Risk Factor	Confirmed by Mendelian Randomisation	Reference
LDL-cholesterol	Yes	(Holmes, Asselbergs et al. 2015)
HDL-cholesterol	No	(Holmes, Asselbergs et al. 2015)
Triglycerides	Yes	(Holmes, Asselbergs et al. 2015)
Lipoprotein(a)	Yes	(Kamstrup, Tybjaerg-Hansen et al. 2009)
C-reactive protein	No	(Wensley, Gao et al. 2011)
Interleukin-6	Yes	(Niu, Liu et al. 2012)
Fibrinogen	No	(Keavney, Danesh et al. 2006)

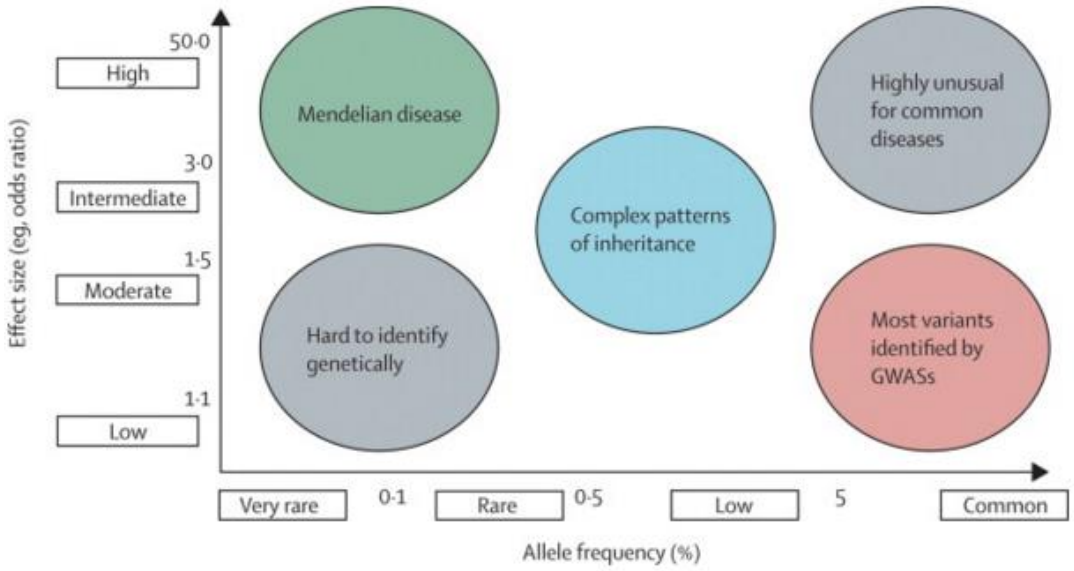
LDL=low density lipoprotein. HDL=high density lipoprotein.

1.4 Genetic risk of CHD

1.4.1 Genetic variation and disease

Family history has long been recognised as a risk factor CHD (Snowden, McNamara et al. 1982; Schildkraut, Myers et al. 1989), highlighting the genetic contribution to this disease. There is a common relationship between the penetrance, frequency and disease susceptibility for risk variants present in the human genome (Figure 4). Evolutionary pressures limit high penetrance variants with a large effect to a low frequency within the population (Blekhman, Man et al. 2008). For example, there is are a multitude of very rare variants that result in genetically strongly raised LDL-cholesterol levels, causing the disease known as familial hypercholesterolaemia (FH) (Futema, Whittall et al. 2013). Lifelong raised LDL-cholesterol in FH patients confers an extremely high risk of CHD, with 50% of men and 30% of women developing the disease before the age of 60 if left untreated. Even the most common FH-causing variant in the UK is present in less than 10% of FH patients (Humphries, Whittall et al. 2006). It follows therefore that many risk alleles for common diseases such as CHD will be present at much higher frequencies in the population but will have a relatively small impact on disease risk. Thus, unlike the disease-causing mutations of rare conditions, to be adequately powered to identify such variants, studies require a large numbers of participants.

Figure 4: The relationship between frequency and susceptibility for genetic variants associated with disease



GWAS = genome-wide association study. Image reprinted from (Speicher, Geigl et al. 2010).

1.4.2 Candidate gene studies

Prior to the sequencing of the human genome, genetic studies of CHD were limited to so-called “candidate gene” studies (Lusis 2012). As atherosclerosis is a multi-factorial process, this provided a number of different pathways to study (such as lipid metabolism and inflammation (Steinberg 2002)) leading to the identification of a number of genes confirmed to be involved in the disease pathway. However, inconsistent findings were common using this approach, making non-replication a major issue (Tabor, Risch et al. 2002). This could result from the small effect size of the risk variant making the association difficult to detect consistently in the relatively small sample sizes available (Ioannidis, Ntzani et al. 2001) or possibly a false association, identified by chance in the initial study (Colhoun, McKeigue et al. 2003). This issue can be partially overcome by systematically reviewing the literature and performing a meta-analysis although this is limited by the quality of the published data which can be reduced by sampling, publication and time-lag bias (Ioannidis, Ntzani et al. 2001). Another limitation of the candidate gene approach is that it was limited to a small number of genes and variants based on the current knowledge of the pathophysiology of CHD and of human genetics (Tabor, Risch et al. 2002). Nevertheless, a number of plausible candidate genes and variants were identified using the candidate gene approach (Table 3).

Table 3: Genetic variants associated with CHD in meta-analyses of candidate gene studies

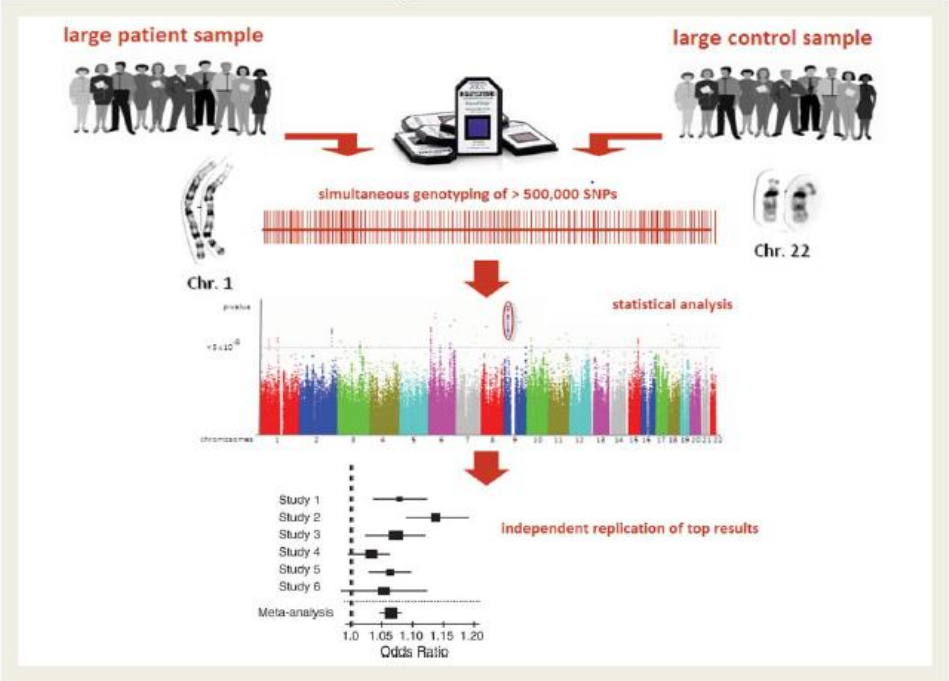
Gene	Chromosome	Variant(s)	Reference
AGT	1	rs699 (M325T)	(Zafarmand, van der Schouw et al. 2008)
MTHFR	1	rs1801133 (C677T)	(Xuan, Bai et al. 2011)
APOB	2	rs1032041 (E4181K) & Signal peptide insertion/deletion	(Chiodini, Barlera et al. 2003)
NOS3	7	rs1799983 (E298D)	(Casas, Cavalleri et al. 2006)
PON1	7	rs662 (Q192)	(Wheeler, Keavney et al. 2004)
SERPINE1	7	rs1799889 (5G/4G)	(Boekholdt, Bijsterveld et al. 2001)
LPL	8	rs328 (S447X)	(Wittrup, Tybjaerg-Hansen et al. 1999)
CETP	16	rs708272 (TaqIB)	(Boekholdt, Sacks et al. 2005)
ACE	17	Insertion/deletion	(Morgan, Coffey et al. 2003)
ITGB3	17	rs5918 (GPIIb-IIIa)	(Morgan, Coffey et al. 2003)
APOE	19	rs7412/rs429358 (e2/e3/e4 polymorphism)	(Bennet, Di Angelantonio et al. 2007)

Data in the table is based on that presented in (Casas, Cooper et al. 2006).

1.4.3 Genome-wide association studies

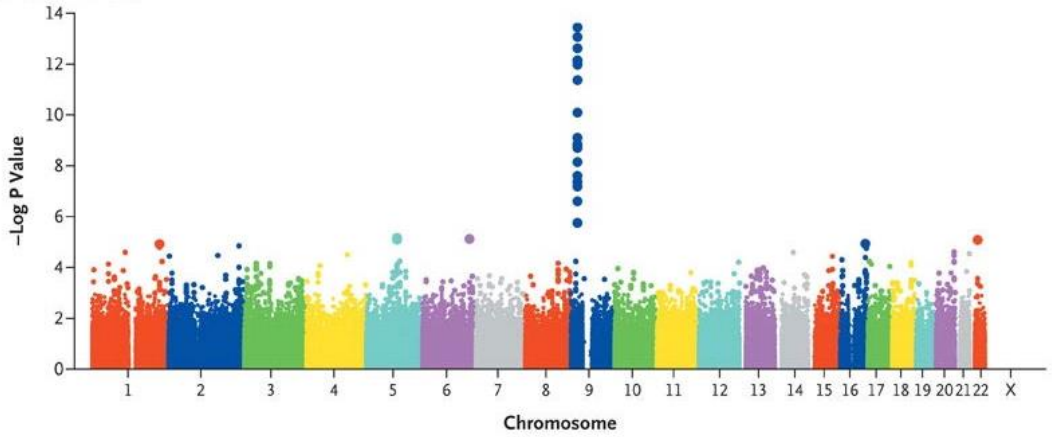
The sequencing of the human genome and advances in genotyping technology gave rise to the ability to perform genome-wide association studies (GWASs) (McCarthy, Abecasis et al. 2008). Here a large number of individuals with and without the trait in question are genotyped using genotyping arrays designed to cover as much of the genome as possible, and association analysis performed. Unlike candidate gene studies, no prior hypothesis as to which genes are important in pathogenesis is required. The GWAS workflow is shown in Figure 5. The p-value threshold for genome-wide significance is usually set in frequentist manner, where it is corrected for multiple testing (often ~$5x10^{-8}$ after adjustment for one-two million independent tests). The results of a GWAS are commonly displayed graphically by chromosome, with p-value on the y axis, referred to as a Manhattan plot (Figure 6). To exploit the potential of GWASs, large consortia have been set up such as CARDIoGRAMplusC4D (for CHD (Deloukas, Kanoni et al. 2013), DIAGRAM (for diabetes (Zeggini, Scott et al. 2008)) and the Global Lipid Genetics Consortium (for lipid traits (Willer, Schmidt et al. 2013)). A number of the risk loci for CHD identified in GWASs (or by fine-mapping GWAS results as in the CARDIoGRAMplusC4D meta-analysis (Deloukas, Kanoni et al. 2013)) lie in genes where risk variants for CHD had been identified in candidate gene studies (Table 4).

Figure 5: Principle of genome-wide association studies



A large number of participants with and without the disease are recruited and genotyped using genome-wide arrays and risk variants identified. Results from different studies in independent cohorts can then be meta-analysed to confirm associations. Image reprinted under licence from the Oxford University Press (Schunkert, Erdmann et al. 2010).

Figure 6: Example Manhattan plot depicting results from a genome-wide association study



Manhattan plot for a GWAS performed for coronary heart disease by the Welcome Trust Case Control Consortium (WTCCC) (Samani, Erdmann et al. 2007). Each dot represents a variant genotyped as part of the GWAS, with chromosome plotted on the x axis and –log p-value on the y axis. Therefore, genomic locations with a strong association will be visually striking. In this example a strong signal has been found on chromosome 9, representing a locus at position p21. GWAS=genome-wide association study. Reproduced with permission from (Samani, Erdmann et al. 2007). Copyright Massachusetts Medical Society.

Table 4: Genes found to be involved in CHD in both candidate gene studies and GWAS-based studies

Gene	Chromosome	Evidence of Association in Candidate Gene Studies	Evidence of Association in GWAS-based Studies
PCSK9	1	(Benn, Nordestgaard et al. 2010)	(Kathiresan, Altschuler et al. 2009)
APOB	2	(Chiodini, Barlera et al. 2003)	(Deloukas, Kanoni et al. 2013)
LPL	8	(Sagoo, Tatt et al. 2008)	(Deloukas, Kanoni et al. 2013)
ABO	9	(Wu, Bayoumi et al. 2008)	(Schunkert, Konig et al. 2011)
APOA5	11	(Sarwar, Sandhu et al. 2010)	(Schunkert, Konig et al. 2011)
APOE	19	(Bennet, Di Angelantonio et al. 2007)	(Deloukas, Kanoni et al. 2013)
LDLR	19	(Linsel-Nitschke, Gotz et al. 2008)	(Kathiresan, Altschuler et al. 2009)

GWAS=genome-wide association study.

1.4.3.1 The impact of the GWAS design

Results from GWASs of common complex diseases have a number of common features (Dandona, Stewart et al. 2010; Imamura and Maeda 2011). Firstly (and not surprisingly as discussed in Chapter 1.4.1) the effect size pertaining to the variants identified is relatively small. Secondly, the majority of the identified risk loci fall outside the exome – pointing to molecular mechanisms that impact on the regulation of gene expression rather than the protein-coding sequence itself(Hindorff, Sethupathy et al. 2009). Furthermore, in general GWASs can only identify a risk locus, not the functional single nucleotide polymorphism (SNP). Linkage disequilibrium (LD) between SNPs allows genotyping chips to cover the genome using a fraction of the variants present. However, as a consequence the lead SNP identified is not necessarily the functional variant. This could be one of many SNPs in strong LD with the lead SNP (that were not genotyped as part of the study). It has also been hypothesised that so-called "synthetic associations" may arise whereby the association of a common genetic variants with a trait results from multiple unobserved low frequency causal variants also present at the locus (Dickson, Wang et al. 2010). Finally, only a minority of the risk variants identified act through known pathogenic mechanisms. For example, of the 53 loci robustly associated with CHD in the meta-analysis published by the CARDIoGRAMplusC4D consortium (Deloukas, Kanoni et al. 2013), only 16 of the 51 were associated with known risk factors. While challenging, this presents an unprecedented opportunity to identify novel pathways which contribute to CHD.

The situation is typified by the example of the CHD risk locus on chromosome 9p21. This was the first CHD risk locus to be identified by a GWAS (Burton, Clayton et al. 2007; McPherson, Pertsemlidis et al. 2007). Of all the loci identified by GWASs, it has the largest effect size for a common polymorphism (minor allele frequency (MAF)>0.05). However, its mechanism of action remains obscure. The locus is not associated with any CRFs for CHD (Deloukas, Kanoni et al. 2013) although an adjacent 11 kb LD block is a GWAS hit for type 2 diabetes (T2D) (Zeggini, Weedon et al. 2007; Zeggini and Ioannidis 2009). The closest protein-coding genes are two cyclin kinase dependent inhibitors involved in cell cycle regulation (*CDKN2A* and *CDKN2B*) located approximately 100 kb upstream (McPherson, Pertsemlidis et al. 2007; Hannou, Wouters et al. 2015). The risk locus overlaps with the sequence of *ANRIL* (also called *CDKN2BAS*), a non-coding ribonucleic acid (RNA) (Pasmant, Laurendeau et al. 2007). Many of the functional studies conducted at the locus have focussed on these genes. Genotype at the 9p21 risk locus has been found to be associated

with expression of *ANRIL* - in whole blood, the risk alleles of SNPs at 9p21 were associated with reduced expression of *ANRIL* (Cunnington, Santibanez Koref et al. 2010) - but not *CDKN2A* and *CDKN2B* (Holdt, Beutner et al. 2010). Moreover, expression of certain *ANRIL* transcripts is associated with plaque burden. This indicates that the effect of the 9p21 locus on CHD is at least partially mediated by influencing the regulation of *ANRIL* expression. There is evidence to suggest that expression of *ANRIL* influences expression of *CDKN2A* and *CDKN2B* (Congrains, Kamide et al. 2012) which could modulate cell cycle regulation (e.g. by influencing macrophage production) which is known to have a key role in progression of atherosclerosis (Braun-Dullaeus, Mann et al. 1998). Chromatin capture techniques have identified a number of enhancers in the 9p21 CHD risk locus and the results implicated disruption of the interferon-ɣ signalling pathway as a possible mechanism (Harismendy, Notani et al. 2011). While subsequent work has not confirmed this relationship (Almontashiri, Fan et al. 2013; Erridge, Gracey et al. 2013), this highlights the possibility that modulation of *ANRIL* expression may affect CHD risk through mechanisms independent of *CDKN2A* and *CDKN2B*.

1.4.4 Publically available data sources used in genetic research

Advances stemming from the sequencing of the human genome have enabled the creation of large data sets concerning different aspects of the genetic contribution to disease and to biology in general. Many projects have been able to create online databases which are publically available. Thus such data can then be used in the functional analysis of GWAS-identified variants.

1.4.4.1 GWAS Catalog

Tens of thousands of associations between SNPs and traits have been identified using GWASs and these are recorded in the GWAS Catalog (Welter, MacArthur et al. 2014) provided by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (part of the European Molecular Biology Laboratory - EMBL-EBI). The details of all studies with a GWAS design meeting the eligibility criteria (assaying at least 100,000 SNPs and using a p-value threshold for association of less than $1x10^{-5}$) are extracted from the literature and added to the catalogue.

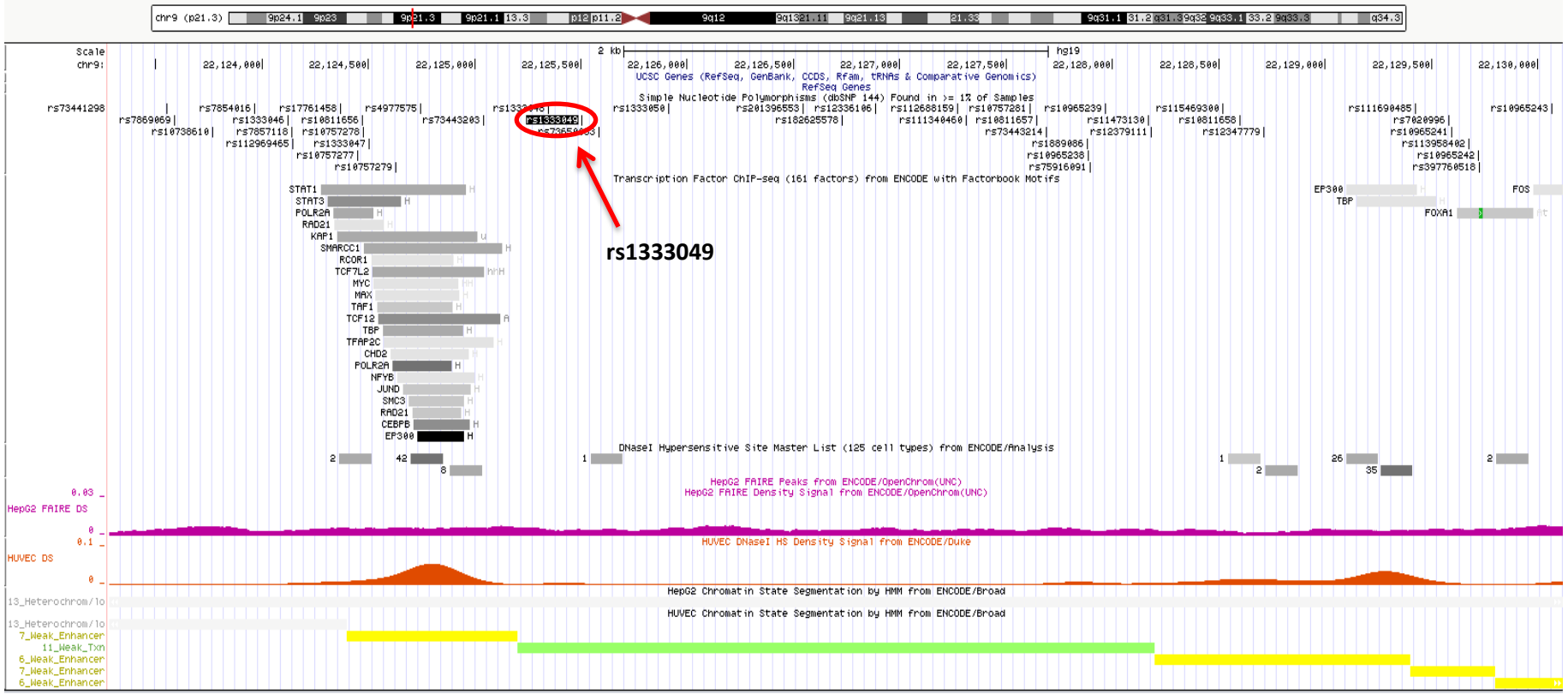
1.4.4.2 HapMap and 1000 Genomes Projects

The International HapMap Project was set up following the sequencing of the human genome to study the common patterns of genetic variation (2003). In total 270 individuals from four populations were genotyped for over three million SNPs (Frazer, Ballinger et al. 2007) and the data is publically available (including MAF and LD data) (Smith 2008). Following on from this the 1000 Genomes Project was created to identify all genetic variants present in at least 1% of individuals in five major populations (European, East Asian, South Asian, West African and American)(Abecasis, Altshuler et al. 2010). In total 2,604 individuals (some in common with the HapMap Project) were genotyped and 88 million variants were identified (Auton, Brooks et al. 2015), providing an invaluable resource to use in the study of genetic variation.

1.4.4.3 Encyclopaedia of DNA Elements Project

The Encyclopaedia of deoxyribonucleic acid (DNA) Elements (ENCODE) project seeks to investigate the functionally of the genome. The first publications from the ENCODE project detailed the findings from 1,640 data sets from 147 (mostly transformed) human cell lines (2012). This included analysis of genome-wide binding of transcription factor binding in 119 cell types (Gerstein, Kundaje et al. 2012), DNase I footprinting in 41 cell and tissue types (Neph, Vierstra et al. 2012) and DNase I hypersensitivity in 125 cell and tissue types (Thurman, Rynes et al. 2012). The project has been expanded to other species, cell types and regulatory features. At the beginning of 2016 the ENCODE repository contained the results of 5000 experiments in seven areas (chromatin structure, 3D genome interactions, DNA-protein interaction, RNA-protein interactions, DNA methylation, transcription and expression) (Diehl and Boyle 2016). Therefore, the ENCODE data provides an invaluable resource to investigate in genomic loci in different cellular environments from direct laboratory analysis (Smith, Humphries et al. 2015). The ENCODE data can be viewed in its genomic context using the UCSC (University of California Santa Cruz) genome browser (Kent, Sugnet et al. 2002). “Tracks” displaying particular data from a cell line of interest can be selected and displayed alongside the genome sequence (see Figure 7 for an example).

Figure 7: Screenshot of ENCODE data for the 9p21 CHD risk locus, displayed using the UCSC Genome Browser (http://genome.ucsc.edu; GR37/hg19)



The lead SNP at the risk locus (rs1333049) is circled in red. The gene "track" is on but there are no markings, indicating that no known genes are present. The other SNPs present at the locus are displayed. To the left of the lead SNP lies a region found to be a DNase I hypersensitivity site in 42 cell types and where multiple transcription factors have been found to bind (also depicted by grey boxes), suggesting this locus is involved in gene regulation. Data from individual cell lines can also be displayed, which is particularly useful if the relevant cells types are known. As an example, tracks showing data from a liver cell-line (HepG2, in pink) and a primary endothelial cell line (human umbilical vein endothelial cells (HUVECs), in orange) are also shown. The peaks represent regions where the feature of interest (e.g. DNAse I hypersensitivity sites) are present. The bottom two tracks represent the chromatin state assigned using the ENCODE data. None were assigned in HepG2 cells but regions were assigned as weak enhancers (yellow) and weakly transcribed (green) in HUVEC cells. ENCODE=encyclopaedia of DNA elements. UCSC=University of California Santa Cruz.

1.4.4.4 Roadmap Epigenomics Consortium

The goal of the Roadmap Epigenomics Consortium is to investigate how epigenomics affects human biology and disease. To create an epigenome, data on DNA accessibility, histone modifications, DNA methylation and RNA expression are used to annotate the genome. Recently data from 111 reference epigenomes from the Roadmap consortium and 16 from the ENCODE project from a variety of different tissues (including adult cells, foetal cells and stem cells) was published (Kundaje, Meuleman et al. 2015). From this data, a 15-state chromatin model has been developed featuring seven repressive and eight active states, which can be highly tissue specific. As such, data from the Roadmap Epigenomics Consortium provides an important insight to functionality and guidance for lab-based functional work (Smith, Humphries et al. 2015). Chromatin state information can be accessed using the HaploReg tool (Ward and Kellis 2012).

1.4.4.5 Genotype-Tissue Expression Consortium

The genotype-tissue expression (GTEX) consortium seeks to provide a resource to study the relationship between genetic variation and gene expression by collecting multiple samples from densely genotypes donors (2013). Previous expression quantitative trait loci (eQTL) studies have been hampered by limited tissue availability and so the GTEX project has collected samples post-mortem. The pilot study included data on 43 tissues taken from 173 individuals (2015) and the project is now being scaled up. The data is publically available on the GTEX browser (http://gtexportal.org/home/). The GTEX database provides a powerful tool to investigate the impact of genetic variation in different tissues. It can be used to search for eQTLs for particular variants or genes and also to compare levels of expression between tissues.

1.5 Epidemiology of CHD

Both CHD incidence and mortality have decreased in high-income nations in recent years (Smolina, Wright et al. 2012; Ford, Roger et al. 2014; Nichols, Townsend et al. 2014). Indeed, CVD is now the second most common cause of the death in a number of European countries including the UK as a result (Nichols, Townsend et al. 2014; 2015). However, this is not reflected in other regions across the world or in some minority ethnic groups within these countries.

1.5.1 CHD in South Asia

The region of South Asia, comprising the countries of India, Pakistan, Bangladesh, Sri Lanka and Nepal contains approximately a quarter of the world's population. As such the region's inhabitants are a heterogeneous group with a variety of different cultures and customs. Nevertheless, South Asian migrants are at a greater risk of developing CHD compared to those of European ethnicity (Zaman, Philipson et al. 2013). There is debate in the literature as to whether the increased burden of CHD in South Asians is accounted for by higher levels of CRFs such as dyslipidaemia and T2D. A large case-control study (the INTERHEART study) which included approximately 30,000 participants from 52 countries found that the excess burden of CHD in South Asians could be attributed to higher levels of CRFs at a younger age (Joshi, Islam et al. 2007). Furthermore, a systematic review of data available for South Asians living in Canada found this group to have a greater prevalence of a number of CRFs including T2D and hypertension (Rana, de Souza et al. 2014). However, overall the findings of the review were consistent with a different cardiovascular risk profile in South Asians compared to those of European ethnicity (e.g. a different relationship between body mass index (BMI) and body fat) indicating that there may be as yet unknown factors - perhaps genetic or epigenetic - contributing to this increased risk.

High quality data concerning CHD in the native South Asian population is relatively scare (Ahmad and Bhopal 2005), however there is evidence that the prevalence has increased greatly over the forty years (Mohan, Deepa et al. 2001; Krishnan, Zachariah et al. 2016), likely as a consequence of increasing urbanisation and the adoption of the more atherogenic "Western" lifestyle.

1.5.2 CHD in populations of African descent

In contrast to South Asian migrants, those of Afro-Caribbean descent living in Western Europe have been found to be at a much lower risk of developing CHD (Wild and McKeigue 1997; Tillin, Hughes et al. 2013). There appears to be a complicated relationship between CRFs and CHD in this group as they have been found to have a protective lipid profile but higher rates of T2D and hypertension (Agyemang, Addo et al. 2009). Curiously, a similarly protective effect was observed in African Americans in the earlier part of the 20th century (Hames and Greenlund 1996), despite a generally poorer risk profile. However, CHD now disproportionately affects African Americans (Crook, Clark et al. 2003) implicating socio-economic and other environmental factors in the aetiology of CHD in this group. For the native African population, data concerning CHD, particularly in Sub-Saharan African is very limited, although it is expected that this will rise as levels of CRFs rise in this region (Onen 2013). There has been a perception that the African population is relatively protected from CHD but it is difficult to assess the evidence as the burden shifts from communicable to non-communicable diseases. A comparison of mortality from CVD in the different ethnic groups in South Africa found that those of African ethnicity had much lower rates of CHD compared to the other ethnic groups - though higher mortality rate from stroke (Bradshaw, Groenewald et al. 2003) – and this was partly attributed to the different ethnic groups being in different stages of the epidemiological transition from a high burden of communicable disease to that of chronic disease (Onen 2013). However, large-scale prospective studies are required to confirm this and to fully investigate the role of CRFs, both traditional and novel.

1.5.3 CHD in East Asia

Unlike the ethnic groups considered thus far, there is no large-scale data on migrant populations from the East Asia. Furthermore, the situation varies greatly between countries within the region. Death rates from CHD in China are increasing, as the burden of infectious disease falls and of non-communicable disease rises (Yang, Kong et al. 2008). There is also an increasing prevalence of CRFs including hypertension (Ma, Mei et al. 2013; Chen, Li et al. 2014) T2D (Zuo, Shi et al. 2014) and dyslipidaemia (Huang, Gao et al. 2014) but not unexpectedly it appears that ethnic groups within China show different CRF clustering (Li, Wang et al. 2012). Whereas, in Japan as CHD mortality declined between the 1960s and the 2000s (Ueshima, Tatara et al. 1987; Hatano 1989). This has been attributed to the reduction in the prevalence of hypertension and smoking, despite the increase in dyslipidaemia although there is evidence to suggest that incidence of CHD and MI is increasing particularly in men (Kitamura, Sato et al. 2008; Rumana, Kita et al. 2008).

1.6 Primary Prevention of CHD

1.6.1 Risk prediction scores

A large proportion of CHD events are preventable (Stamler, Dyer et al. 1993; Stampfer, Hu et al. 2000). Therefore, predicting those at highest risk of developing the disease is an important public health consideration. To take advantage of the combined knowledge of how CRFs affect CHD risk, risk scores have been developed. Here a linear function based on mean values for the risk factors included is calculated and then corrected for the individual's own details (or this is incorporated directly into the original calculation). This value is then exponentiated and incorporated into a survival function to give the CHD risk in a defined period of time, usually ten years. The first risk score for CHD that gained widespread use was developed from the Framingham Heart Study and thus is referred to as the Framingham score (Wilson, D'Agostino et al. 1998). Included in it were age, total cholesterol, HDL-cholesterol, systolic blood pressure, diabetes and smoking (with separate equations for men and women). The score showed good predictively ability in some cohorts similar to that from which it was derived (D'Agostino, Grundy et al. 2001; Simons, Simons et al. 2003). However, it was found to overestimate risk in other ethnic groups (Barzi, Patel et al. 2007) and in other populations of European ethnicity where there was a lower incidence of CHD (Brindle, Emberson et al. 2003; Hense, Schulte et al. 2003). In response to this, region-specific scores have been developed such as SCORE which was derived using data from 12 prospective European cohorts (Conroy, Pyorala et al. 2003). The development of large primary care electronic records has enabled risk scores to be derived from large population cohorts. In England the QRISK score was derived from the QRESEARCH database, (which contains 1.2 million individuals) to estimate risk of CVD (rather than CHD) (Hippisley-Cox, Coupland et al. 2007)) to. This score updated (QRISK2) to include a number of other risk factors, most notably self-reported ethnicity (Hippisley-Cox, Coupland et al. 2008). A similar score (ASSIGN) was developed in Scotland using a nationally representative database – the Scottish Heart Health Extended Cohort (Woodward, Brindle et al. 2007). Both ASSIGN and QRISK2 include measures of social depravation. This was prompted by the observation that the Framingham score underestimated risk in socially deprived individuals and thus could re-enforce social gradients in disease (Brindle, McConnachie et al. 2005; Tunstall-Pedoe and Woodward 2006). The QRISK2 model is updated annually (http://www.qrisk.org/).

1.6.2 Primary prevention strategies

Many of the CRFs for CHD are modifiable and thus lifestyle interventions such as use of smoking cessation services and dietary review, form an important part of the strategy to reduce CHD risk (2014). Prescription of lipid-lowering therapies, primarily statins, has also been used to compliment this. Statin use has been found to reduce risk of CVD events by approximately one fifth per 1mmol/l of LDL-cholesterol reduction in a wide range of individuals (Baigent, Blackwell et al. 2010). A benefit has also been found in those with low CVD risk (Mihaylova, Emberson et al. 2012). However, a Cochrane review of the data concerning the benefit of statin use in primary prevention of CVD found shortcomings in many of the trials identified, with evidence of selective reporting and inclusion of individuals with CVD in many of the trials used in reviews of data on statin use in primary prevention (Taylor, Huffman et al. 2013). Moreover, statin therapies are not without their limitations. Data from both RCTs and Mendelian randomisation has found that statin therapy increases the risk of developing T2D (Sattar, Preiss et al. 2010; Swerdlow, Preiss et al. 2015) in an apparently dose dependent manner (Preiss, Seshasai et al. 2011). Mendelian randomisation also found that LDL-lowering alleles of variants in *HMGCR* (which encodes the protein targeted by statin therapies) were also associated with increased bodyweight, a known causal risk factor for T2D, which may partially explain the relationship between statin use and T2D (Swerdlow, Preiss et al. 2015). Furthermore, statin use has been found to be associated with an increase in glycated haemoglobin levels (a measure of glycaemic control) in those with diabetes (Erqou, Lee et al. 2014). However, both of these effects are relatively modest and outweighed by the protective benefits also found in those with T2D (Kearney, Blackwell et al. 2008). Statin use has also been associated with a greater risk of myopathy in a number of large-scale observational studies (Bruckert, Hayem et al. 2005; Nichols and Koro 2007; Hippisley-Cox and Coupland 2010). However, a meta-analysis of statin RCTs reporting adverse effects found that only a small minority of these were due to statin use (Finegold, Manisty et al. 2014) but the authors stated that the study was probably limited due to the poor reporting of side effects in clinical trials in academic journals (Goldacre 2014).

1.6.3 Current clinical guidance

When risk scores were first introduced into clinical practice, the Framingham risk score was the recommended for use in both the USA and the UK and the high risk group was defined as those having a ten-year risk of CHD ≥20% (Cooper and O'Flynn 2008). Those who fell into that category were then recommended for intensive lifestyle and prescription of lipid-lowering medications (usually statins). However, the joint guidelines issued by the American College of Cardiology (ACC) and the American Heart Association (AHA) developed new risk equations and lowered the high-risk cut-off to ≥7.5% (Goff, Lloyd-Jones et al. 2013). Similarly, in the UK the National Institute of Health and Clinical Excellence (NICE) updated their guidelines to recommend use of QRISK2 and lowered the high-risk threshold to ≥10% (2014). However, given the shortcomings in the available data for statin use there have been concerns particularly regarding the "medicalisation of healthy individuals" and the numbers of adverse events observed in certain groups (Goldacre and Smeeth 2014).There is also evidence that uptake of statins in the those classified in the 10-20 % risk groups is much lower than estimated by NICE (Usher-Smith, Pritchard et al. 2015), although larger studies are required to confirm this.

1.6.4 The "prevention paradox"

The majority of cases of CHD/CVD come from individuals classified with average risk using the CRF risk scores – the so-called prevention paradox (Rose 1981). For example, when use of QRISK2 (2010 version) was validated with data from the health improvement network (THIN), (using a 20% high-risk cut-off), 14% of men and 6% of women were identified as being at high risk. This captured 40% of the cardiovascular events in men and 26% of the cardiovascular events in women (Collins and Altman 2010). This leaves scope for refinement of the risk score to discriminate between those who do and do not go on to develop CVD. In addition to those CRFs already included in risk prediction a many others have been proposed including inflammatory markers (Madjid and Willerson 2011), lipoprotein(a) (Kamstrup, Tybjaerg-Hansen et al. 2013) and genetic information (as discussed in Section 1.6.5).

1.6.5 Use of genetics in risk prediction

The identification of robustly associated CHD risk loci has prompted increasing interest in the inclusion of genetic information in risk prediction, particularly as being fixed at conception, it has a lifelong impact and need only be determined once (Di Angelantonio and Butterworth 2012). Given the relatively small effect sizes pertaining to robustly associated risk loci (OR<1.3, (Deloukas, Kanoni et al. 2013)), it is unsurprising that the addition of one variant into a CRF risk score has not resulted in improved predictive ability (Talmud, Cooper et al. 2008; Brautbar, Ballantyne et al. 2009; Paynter, Chasman et al. 2009). This has led to the development of so-called "gene scores" (GS) where SNPs at independent loci are combined. A GS can be unweighted, where the number of risk alleles at each locus is combined. Alternatively, the individual SNPs can be weighted using the effect size pertaining to their association with CHD prior to their being added together. A number of different combinations of GS and CRF score have been assessed to determine if inclusion of an estimate of genetic risk can improve predictive ability over-and-above the CRF score alone, with mixed results (Paynter, Chasman et al. 2010; Ripatti, Tikkanen et al. 2010; Vaarhorst, Lu et al. 2012; Ganna, Magnusson et al. 2013). It remains unclear whether under the current clinical guidelines it is beneficial to include a CHD risk GS in CHD risk prediction.

1.7 Aims

The aims of this thesis were:

1. To investigate the use of a CHD risk GS in the UK population.

2. To investigate the use of a CHD risk GS in the South Asian and Afro-Caribbean populations.

3. To perform a systematic literature search to identify variants suitable for inclusion in an updated CHD risk GS.

4. To perform a systematic literature search to identify variants suitable for inclusion in a CHD in T2D risk GS.

5. To assess how attendance at a self-management intervention, with and without provision of personalised CHD risk information impacts on behavioural and clinical outcome in those with T2D - as part of the coronary heart disease and diabetes-related risk (CoRDia) study.

6. To perform functional analysis of the CHD risk locus on chromosome 21q22.

2 Methods

2.1 Studies Included

2.1.1 The Second Northwick Park Heart Study

The second Northwick Park heart study (NPHSII) is a prospective study of 3052 men recruited from nine general practices in the UK (Miller, Bauer et al. 1995). All recruits were aged between 50 and 64 and of European ethnicity. Men who were not free of CVD (defined as unstable angina, MI, electrocardiogram evidence of a silent MI, anti-coagulant or aspirin therapy, coronary surgery or other cerebrovascular disease), or who had malignancy or anything that would prevent the giving of informed consent were excluded. A non-fasting sample was taken at baseline, to perform biochemical analysis. CRF risk scores, Framingham and QRISK2 (2012 version) were calculated from baseline data to assess ten-year CHD and CVD risk respectively. These were assessed using ten-year follow-up data. In follow-up CHD was defined as acute MI, silent MI or undergoing coronary surgery. CVD was defined as CHD (as defined above), a new major Q wave on the ECG after five years of follow-up, surgery for angina pectoris with CHD angiographically demonstrated, stroke congestive heart failure or peripheral vascular disease. All subjects gave written informed consent. The study had ethical approval from the institutional ethics committee and was performed in accordance with the Declaration of Helsinki. DNA extracted from blood was used for genotyping. Unless otherwise stated NPHSII had been pre-genotyped prior to the commencement of this project using Taqman assays (Chapter 2.3.1.2) or restriction length fragment polymorphism based methods.

2.1.2 University College, London School of Hygiene and Tropical Medicine, Edinburgh and Bristol Consortium

The University College, London School of Hygiene and Tropical Medicine, Edinburgh and Bristol (UCLEB) Consortium comprises 12 prospective studies, almost all participants of which are of European ethnicity (Shah, Engmann et al. 2013). These include the British Regional Heart Study (BRHS), British Women's Heart and Health Study (BWHHS), Caerphilly Prospective Study (CAPS), Edinburgh Artery Study (EAS), English Longitudinal Study of Aging (ELSA), Edinburgh Type 2 Diabetes Study (ET2DS), Medical Research Council 1946 birth cohort (MRC1946) and the Whilehall II study (WHII) The data collected differs between studies but all have recorded general CHD risk factors (such as those used to calculate the QRISK2 score (Hippisley-Cox, Coupland et al. 2008)). CHD was defined as the occurrence of fatal CHD, non-fatal MI or undergoing coronary artery bypass or angioplasty. CVD was

defined as CHD or stroke. Stroke included all non-fatal stroke (ischaemic and haemorrhagic combined, but excluding transient ischaemic attacks) and fatal stroke. Median follow-up was ten years. The Metabochip platform (Voight, Kang et al. 2012) was used to genotype approximately 21,000 participants included in the UCLEB studies. This platform has approximately 200,000 SNPs, designed to cover regions associated with cardio-metabolic disease. Imputation based on data from the 1000 Genomes European ancestry sample extended the SNP coverage to approximately one million SNPs ($R^2 \geq 0.8$). QRISK2 (2014 version) was used to determine ten-year CVD risk. All subjects included gave informed consent. Individual research ethics committees gave written consent to use anonymised individual level data which had been obtained by each participating study. Analysis concerning the association of rs10911021 with CHD and with CHD risk factors in those with and without T2D was performed using STATA (StataCorp 2013).

2.1.3 Islamabad MI case-control study

The case group was recruited from the Rawalpindi Institute of Cardiology, Pakistan. All cases (n=321) had had an MI as defined by a positive test for troponin T, ST segment changes on electrocardiogram and typical chest pain radiating in the chest that was not relieved at rest. Control subjects (n=228) were recruited from the general population and did not have a history of CVD. Blood samples were taken from all participants and DNA was extracted for genotyping. Diabetes was defined as prescription of diabetes medication and hypertension as blood pressure greater than 140/100 mmHg or prescription of anti-hypertensive medication. The study had approval from the Institutional Review Board and Ethics Committee of Shifa College of Medicine, Shifa International Hospital, Islamabad and all subjects gave written informed consent.

2.1.4 Lahore CHD case-control study

Cases were collected from hospitals in Lahore, Pakistan which covers the whole of the Punjab region. All cases (n=404) had CHD as defined by echocardiogram, angiography and/or biochemical markers. Participants in the control group (n=219) were age and sex matched to the cases, recruited from the general population and did not have a history of CVD. Blood samples were collected for biochemical analysis and DNA extracted for genotyping. Diabetes was defined as fasting blood glucose above 110 mg/dl or non-fasting blood glucose above 140 mg/dl. Hypertension was defined as blood pressure over 150/90 mmHg for participants aged 60 or over, blood pressure over 140/90 mmHg for participants

aged 30-59 and blood pressure over 120/80 mmHg for participants aged under 30. All participants gave written informed consent and the study had ethical approval from the institutional ethical committee, University of the Punjab, Lahore.

2.1.5 Guadeloupe CHD case-control study

Cases were collected from the department of cardiology of the university hospital of Pointe-à-Pitre, Guadeloupe. Cases (n=178) were defined by history of coronary angioplasty and/or coronary bypass surgery or acute MI. Control participants (n=359) were recruited from a public health centre on Guadeloupe and had no history or suspicion of CVD. The study had ethical approval from the inter-regional ethics committee (Sud-Ouest/Outre-Mer III, France) and all participants gave written informed consent. Diabetes was defined as history of diabetes or prescription of hypoglycaemic agents (including insulin). Hypercholesterolaemia was defined as a history of hypercholesterolaemia or prescription of lipid-lowering therapies. Hypertension was defined as a history of hypertesion or prescription of anti-hypertensive therapies.

2.1.6 University College Diabetes and Cardiovascular Study

The University College diabetes and cardiovascular study (UDACS) is a cross-sectional study comprising 1020 participants of mixed ethnicity who were recruited from the University College London Hospitals (UCLH) NHS trust diabetes clinic (Stephens, Hurel et al. 2004). All recruits had diabetes as defined by the WHO criteria (Alberti and Zimmet 1998) but did not require renal dialysis. A number of CRF measures were obtained at recruitment. Approval was obtained from the UCL/UCLH ethics committee.

2.1.7 Advanced Study of Aortic Pathology

Patients undergoing aortic valve surgery at the Karolinksa University Hospital, Stockholm, Sweden were recruited (n=213) into the advanced study of aortic pathology (ASAP) (Folkersen, van't Hooft et al. 2010). Tissue biopsies were taken from the mammary artery, aortic adventitia, aortic intima media heart and liver. Messenger RNA was extracted and measured using the Affymetrix Gene Chip Human Exon 1.0 ST expression array (Santa Clara, CA, USA). Genotyping was performed using the Illumina Human 610W Quad Beadarrays (San Diego, CA, USA). All participants gave informed consent and the study had approval from the ethical committee of the Karolinska Institute.

2.2 Systematic Literature Search

To identify all variants found to be associated with CHD, a systematic literature search was performed.

2.2.1 Search strategy

A computerised literature search of "Web of Science" (Thomson Reuters, New York City, New York, USA) was performed for studies published in English between the inception of the database (1900) and February 2013. The following search terms were used, "coronary artery disease" *or* "coronary heart disease" *or* "acute myocardial infarction" AND "genetics" *or* "risk variants" *or* "single nucleotide polymorphisms". In addition, the GWAS catalog (Welter, MacArthur et al. 2014) was searched using the heading "coronary heart disease".

2.2.2 Study selection

All retrieved articles were assessed for relevance using the following inclusion criteria: 1) the articles reported an original peer-reviewed study; 2) the study was performed in a population of European ethnicity and 3) the study reported an association between a single genetic variant and CHD and provided a quantitative risk estimate. Four different subgroups were considered: variants associated with CHD, variants associated with premature CHD, variants associated with CHD in T2D and variants associated with secondary CHD events. The definitions of CHD accepted were: positive for angiography or angioplasty, bypass surgery, MI, symptomatic or treated angina, CHD death, coronary revascularisation and abnormal electrocardiogram. Studies which included a broader definition of cardiovascular disease or which used an intermediate phenotype (e.g. atherosclerosis) were excluded. Studies conducted in populations suffering from other conditions (except T2D) were also excluded.

2.2.3 Data extraction

The following information was extracted from each study included: effect size of the association between the genetic variant and the phenotype, gene (or locus) the variant is located in, authors, publication date, study design and gender and age of the study participants. For each variant identified, only the most recent meta-analysis was included. Where multiple variants in LD (r^2>0.1 based on the values calculated in the EUR 1000

Genomes pilot 1 data) were identified only articles concerning the most commonly studied variant were further considered.

2.3 Genotyping

2.3.1 Fluorescence based methods

2.3.1.1 DNA preparation

Prior to performing genotyping using fluorescence based methods, DNA concentrations were standardised (e.g. to 15 ng/µl) and diluted to 1.25 ng/µl. Where necessary, the DNA concentration was measured using the Nanodrop 8000 (Thermo Fisher, Scientific Waltham, MA, USA). A robotic liquid handling system (Biomek FX, Beckman Coulter, High Wycombe, UK) was used to transfer 4 µl of DNA (total 5 ng) in to 384-well DNA array plates to be used in genotyping.

2.3.1.2 Taqman genotyping

Taqman genotyping is based on the principle of allele-specific fluorescence emission during a polymerase chain reaction (PCR) (Shen, Abdullah et al. 2009). In each Taqman genotyping assay, there is a forward primer and a reverse primer, which bind either side of the SNP to be genotyped, allowing amplification of that region during PCR. Each assay also contains two allele-specific probes labelled at the 5' end with a different fluorophore (FAM or VIC). A quencher which prevents the fluorophore from fluorescing is bound at the 3' end. During the denaturation step of PCR, these probes can bind to their complementary sequence. During the DNA replication step, the exonuclease activity of the DNA polymerase degrades the bound probe, liberating the fluorophore from the quencher and fluorescence will be emitted. Therefore, if both fluorophores fluoresce the sample is heterozygous. Whereas should there only be a signal from one of the fluorophores, the sample is homozygous for the corresponding allele.

The make-up of the reaction mix depended upon the buffer available. The reaction volumes used with the different buffers are shown below (KAPA buffer Table 5; Taqman genotyping mastermix Table 6). In both cases, 1.8 µl of the reaction mix was added per well. The plate was then centrifuged and the PCR performed using the conditions shown in Table 7.

Table 5: Per-plate master-mix composition for Taqman assays performed KAPA buffer

Reagent	Volume (µl)
KAPA buffer (Kapa Biosystems)	410
Nuclease free water (Sigma- Aldrich)	377.1 / 387.4
40x/80x SNP specific assay (Life Technologies)	20.5 / 10.3
High rox (Kapa Biosystems)	16.4

The volume of SNP specific assay (which contains the primers and probes) depends on whether it is supplied as 40x or 80x. Therefore, the volume of water used also varies accordingly. Sigma-Aldrich-St Louis, MO, USA; Applied Biosystems, Life Technologies - Carlsbad California, USA. Kapa Biosystems – Cape Town, South Africa.

Table 6: Per-plate master-mix composition for Taqman assays performed with Taqman genotyping buffer

Reagent	Volume (µl)
Taqman genotyping buffer (Applied Biosystems)	410
Nuclease free water (Sigma-Aldrich)	389.5 / 399.8
40x/80x SNP specific assay (Applied Biosystems)	20.5 / 10.3

The volume of SNP specific assay (which contains the primers and probes) depends on whether it is supplied as 40x or 80x. Therefore, the volume of water used also varies accordingly. Sigma-Aldrich-St Louis, MO, USA; Applied Biosystems, Life Technologies - Carlsbad California, USA.

Table 7: PCR conditions for Taqman genotyping assays

Temperature	Time	
50°C	2 mins	
95°C	10 mins	
95°C	15 s	40 cycles
60°C	1 mins	

2.3.1.3 KASPar genotyping

KASPar genotyping utilises Kompetitive Allele Specific PCR (KASP) technology (Cuppen 2007). Two components are required for this, the SNP-specific KASP assay mix and the general KASP Master Mix used for all KASPar assays. The KASP assay mix contains two allele specific forward primers, each with a differing 5' tail sequence and a common reverse primer. The KASP Master Mix contains two fluorescence resonance energy transfer (FRET) reporter cassettes. One strand of each FRET reporter cassette has a sequence identical to one of the 5' tails on the allele-specific primers, with either fluorophore FAM or HEX bound at the 5' end. The complementary sequence has a quencher bound at the 3' end and therefore the fluorophore cannot fluoresce. During the PCR reaction, the allele-specific forward primers bind if the corresponding allele is present in the template DNA, resulting in amplification of the SNP-containing region. In subsequent rounds of PCR, the 5' tail sequence of the forward primer is amplified and complementary sequence generated. The fluorophore-labelled component of the FRET reporter cassette can now bind its

complementary sequence, releasing it from the quencher, allowing fluorescence to be emitted.

The reagent volumes used per-plate are stated below (Table 8) with 1.8 µl of reaction mix added per well. PCR was performed with the conditions shown in Table 9. Where distinct genotyping clusters were not observed, three further PCR cycles were performed (conditions shown in Table 10). Where using 1.8 µl of reaction mix per-well of reaction mix did not give acceptable results, the assay was repeated using 3.6 µl of reaction mix.

Table 8: Per-plate master-mix composition for KASPar assays

Reagent	Volume 1.8 µl per-well/ 3.6 µl per-well (µl)
KASP master mix	422.4 / 844.8
Nuclease free water (Sigma-Aldrich)	422.4 / 844.8
KASP assay mix	11.7 / 23.4

Sigma-Aldrich - St Louis, MO, USA.

Table 9: PCR conditions for KASPar genotyping assays

Temperature	Time
94°C	15 mins
94°C	20 s
65°C-57°C	60 s } 10 cycles
94°C	20 s } 26 cycles
57°C	60 s

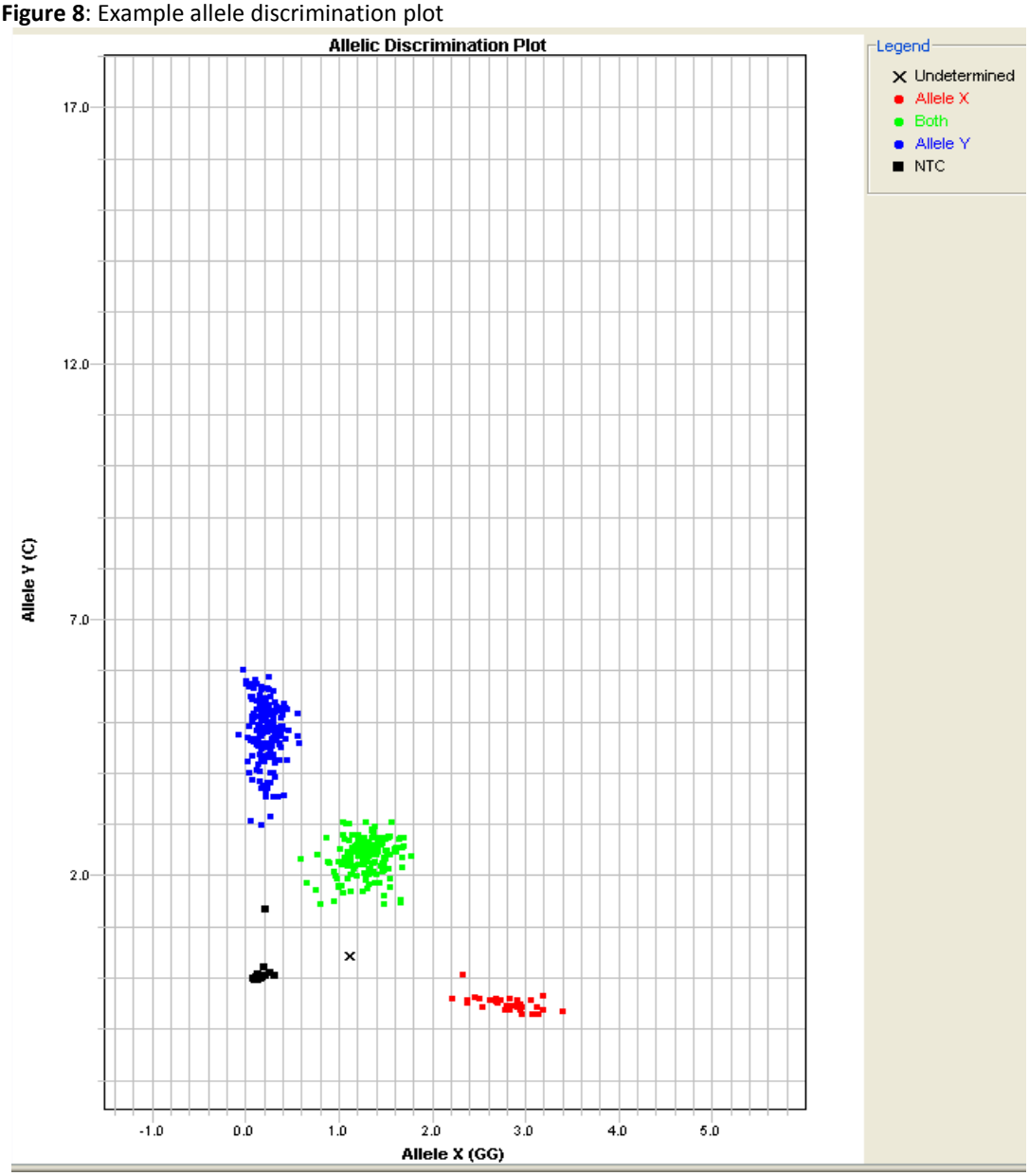
Table 10: PCR conditions for when further cycles are required for KASPar assays.

Temperature	Time
94°C	20 s } 3 cycles
57°C	60 s

2.3.1.4 Signal detection during Taqman and KASPar genotyping

Both Taqman and KASPar genotyping assays are "end-read" methods where the fluorescent signal is detected following completion of the PCR. This was performed using the 7900HT Fast Real-Time PCR System (Applied Biosystems, Life Technologies, Carlsbad California, USA) with the Sequence Detection System (SDS v2.1) software (Applied Biosystems, Life Technologies, Carlsbad California, USA). The fluorescent dyes used in the genotyping assays have different excitation and emission wavelengths allowing them to be differentially detected. Genotypes were assigned by the software based on the fluorescent signal and these were checked manually. An example allele discrimination plot is shown in Figure 8. Each point on the scatterplot represents a well in the 384-well plate (DNA sample or no-template control (NTC)).

Figure 8: Example allele discrimination plot



The plot was generated when genotyping one NPHSII plate for rs12526453 using a KASPar assay. Signal from the VIC dye has been plotted on the x axis and from the HEX dye on the y axis. Blue points represent samples with a strong HEX signal only and are assigned as homozygous for the corresponding allele (in this case C). Similarly, red points represent samples with a strong VIC signal only and are assigned homozygous for the other allele (in this case G). Samples with both a strong VIC and HEX signal are heterozygous and are represented by green dots. Black squares represent no template control (NTC) wells and black "X"s represent wells were the signal did not meet the significance threshold for any genotype and thus are unassigned.

2.3.2 Sanger sequencing

On occasion it was necessary to confirm the genotypes determined using Sanger sequencing (for example rare homozygotes). To do this the region surrounding the SNP was amplified by PCR. The reaction conditions are shown in Table 11. To amplify the sequence surrounding two SNPs in *APOE* (rs7412 and rs429358), presence of ethylene glycol in the reaction was required. The required reagent volumes for this are shown in Table 12. The PCR conditions used in the PCRs are shown in Table 13. Following PCR, the products were purified using the GFX PCR and Gel Band Purification kit (GE Healthcare, Thermo Fisher Scientific Waltham, MA, USA), according to the manufacturer's instructions. Purified PCR products were sent for Sanger sequencing along with the required primers to either Eurofins (Ebersberg, Germany) or Source Bioscience (Cambridge, UK).

Table 11: Sequencing pre-amplification reaction mix (for SNPs not in *APOE*)

Reagent	Volume (µl)
Multiplex PCR mastermix (Qiagen)	10
Nuclease free eater (Sigma-Aldrich)	6
Forward sequencing primer	1
Reverse sequencing primer	1
Template DNA (5 ng/µl)	2
	20

Sigma - Sigma-Aldrich, St Louis, MO, USA; Qiagen - Hilden, Germany.

Table 12: Sequencing pre-amplification reaction mix for SNPs in *APOE*

Reagent	Volume (µl)
Multiplex PCR mastermix (Qiagen)	10
Nuclease free water (Sigma Aldrich)	4
Forward sequencing primer	1
Reverse sequencing primer	1
Ethylene glycol (Sigma-Aldrich)	2
Template DNA (5 ng/µl)	2
	20

Sigma - Sigma-Aldrich, St Louis, MO, USA; Qiagen - Hilden, Germany.

Table 13: PCR cycling conditions for sequencing pre-amplification reactions

Temperature	Time	
95°C	15 mins	
94°C	1 mins	35 cycles
65.5°C	1 mins	
72°C	1 mins	
72°C	5 mins	

2.3.3 Cardiac Risk Prediction Array

The cardiac risk prediction array was developed by Randox Laboratories Ltd (Crumlin, Co Antrim, UK). This was as part of a collaboration between the Centre for Cardiovascular Genetics at UCL, Storegene (London, UK) and Randox Laboratories Ltd which aims to provide an estimate of CHD risk based on a combined CRF and genetic risk score. The array itself simultaneously genotypes 19 CHD risk SNPs (Chapter 3.1, Table 23) and is based on Randox's Biochip Array Technology. The procedure involves amplifying the region surrounding the target SNPs in an allele-specific manner in a multiplex PCR. The PCR products are detected by hybridisation to spatially tethered probes on the biochip array surface. Each position on the biochip array corresponds to a specific allele and genotypes are determined using the Evidence Investigator Analyser (Crumlin, Co Antrim, UK). The array was used to genotype participants in the self-management intervention (SMI) plus risk arm of the CoRDia study (Chapter 5). The protocol was performed according to the manufacturer's instructions.

2.4 *In vitro* functional techniques

2.4.1 Cell culture

Cell culture was performed under sterile conditions. Hepatocellular carcinoma cell lines (Huh-7 and HepG2) were grown in T175 flasks. Both require Dulbecco's Modified Eagle Medium (DMEM) with serum. This was prepared by adding 50 ml of heat inactivated foetal bovine serum (FBS) to 450 ml DMEM (i.e. to give overall 10% FBS content). Both cell lines are strongly adherent and where necessary were removed by trypsin treatment. To do this, the cells were washed in 1x phosphate buffered saline (PBS) and removed from the flask surface by treating with approximately 4 ml of 0.25% trypsin-Ethylenediaminetetraacetic acid (EDTA) solution for 3-5 minutes. The trypsin was deactivated by the addition of approximately 10 ml of DMEM medium.

2.4.2 Electrophoretic mobility shift assay

The electrophoretic mobility shift assay (EMSA) is used to study DNA-protein interactions *in vitro* (Hellman and Fried 2007). Labelled probes corresponding to a genomic region are incubated with nuclear extract from a relevant cell line. This mix in then run on a polyacrylamide gel. The underlying principle is that unbound probe will move more quickly through the gel than DNA-protein complexes. Such complexes will form discrete bands on the gel when it is visualised (e.g. using chemiluminescence for biotin labelled probes). In order to study allele-specific binding, two probe sets (one corresponding to each allele) are designed, the assay performed for both and the band pattern produced compared. Competitor EMSAs can be used investigate proteins that might be involved in the DNA-protein complex. This is carried out by adding a much greater concentration of unlabelled probe which has the protein of interest's consensus binding sequence. Should the EMSA bands observed in a competitor EMSA be much weaker or absent (i.e. are "competed out"), this indicates that the protein under investigation is binding to the genomic sequence of interest. This technique was used to study SNPs located at the CHD risk locus on chromosome 21q22.

2.4.2.1 Extraction of nuclear proteins

Cells were grown to near confluence in T175 flasks and trypsinised as previously described (section 2.4.1). The cells were spun at 1500 rpm for 4 minutes at 4^{o}C and the pellet re-suspended in5 ml ice-cold Buffer A (components of this shown in Table 14 with 50 µl of protease inhibitor 100X added). The mix was then left on ice for ten minutes and thereafter

was spun at 1500 rpm for 4 minutes at 4°C. The pellet was re-suspended in 2 ml Buffer A (with 20 µl of protease inhibitor 100X added) and vortexed for 30 seconds. This mix was spun at 13,000 rpm for 2 minutes at 4°C. The pellet was suspended in 800 µl Buffer C (components shown in Table 15) and 16 µl of protease inhibitor (100X). The mix was vortexed for 1 minute then left on ice for 10 minutes and this was repeated three times. The mix was then spun at 13,000 rpm for 50 minutes at 4°C. The supernatant containing the nuclear proteins was then divided into 50 µl aliquots which were stored at -80°C until required.

Table 14: Components of Buffer A

Reagent	Volume
10 mM HEPES (pH 7.9 4°C)	1ml
1.5 mM $MgCl_2$	150 µl
10 mM KCl	500 µl
dH_20	to 100 ml

HEPES= 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid. dH_20=distilled water.

Table 15: Components of Buffer C

Reagent	Volume
20 mM HEPES pH 7.9	2 ml
25% v/v glycerol	50 ml
0.42 M NaCl	10.5 ml
1.5 mM $MgCl_2$	150 µl
0.2 mM EDTA	40 µl
dH_20	to 100 ml

HEPES= 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid. dH_20=distilled water.
EDTA= Ethylenediaminetetraacetic acid

2.4.2.2 Probe preparation

The probes used for the EMSAs described herein were 25 bases in length, with the SNP position in the middle. Two probes, one corresponding to each allele, were designed for each SNP (and ordered from Eurofins, Ebersberg, Germany). The probe sequences designed for SNPs at the 21q22 CHD risk locus are given in Table 16. The probes were reconstituted in nuclease-free water (Sigman-Aldrich, St Louis MO, USA) according to the manufacturer's instructions.

Table 16: EMSA probe sequences for SNPs at the 21q22 CHD risk locus

SNP - Allele	Sequence
rs9982601 – C	CACAGGGCTGCTCCATGGCCTTGGA
rs9982601 – C	TCCAAGGCCATGGAGCAGCCCTGTG
rs9982601 – T	CACAGGGCTGCTTCATGGCCTTGGA
rs9982601 – T	TCCAAGGCCATGAAGCAGCCCTGTG
rs28451064 – G	CCAGGCCAAAGTGGACACCAAATAC
rs28451064 – G	GTATTTGGTGTCCACTTTGGCCTGG
rs28451064 – A	CCAGGCCAAAGTAGACACCAAATAC
rs28451064 – A	GTATTTGGTGTCTACTTTGGCCTGG
rs9980618 – C	AGGGTGTCTGCTCCAGCACACCATG
rs9980618 – C	CATGGTGTGCTGGAGCAGACACCCT
rs9980618 – T	AGGGTGTCTGCTTCAGCACACCATG
rs9980618 – T	CATGGTGTGCTGAAGCAGACACCCT
rs60687299 – T	CACTGTATTGAATACTGGAGGCAAC
rs60687299 – T	GTTGCCTCCAGTATTCAATACAGTG
rs60687299 – C	CACTGTATTGAACACTGGAGGCAAC
rs60687299 – C	GTTGCCTCCAGTGTTCAATACAGTG
rs9977419 – T	TGTGATAGTGAGTGAGTTCTTACGA
rs9977419 – T	TCGTAAGAACTCACTCACTATCACA
rs9977419 – A	TGTGATAGTGAGAGAGTTCTTACGA
rs9977419 – A	TCGTAAGAACTCTCTCACTATCACA
rs9977093 – G	CCATGCAGAACTGTGAATCAATTAA
rs9977093 – G	TTAATTGATTCACAGTTCTGCATGG
rs9977093 – A	CCATGCAGAACTATGAATCAATTAA
rs9977093 – A	TTAATTGATTCATAGTTCTGCATGG

The SNP position is underlined. EMSA=electrophoretic mobility shift assay.

To enable detection of DNA-protein complexes, a biotin label was attached to the 3' end of all probes. The reaction mixture used for this is shown in Table 17. The reactions were incubated at 37^{o}C for 90 minutes. Thereafter the TdT enzyme was removed adding an equal volume of chroloform : isoamyl alcohol (24:1), vortexing the mix briefly and spinning for 2 minutes at 13,000 rpm. The aqueous DNA containing phase was removed and stored.

Table 17: Biotinylation reaction mix per 30 µl reaction

Reagent	Volume (µl)
5X TdT buffer (Thermo Fisher Scientific)	6 µl
Biotin-11-dUTP (Thermo Fisher Scientific)	3 µl
TdT enzyme (Thermo Fisher Scientific)	0.3
H_20 (Sigma-Aldrich)	17.7
DNA probe	3
	30

Thermo Fisher Scientific - Waltham, MA, USA; Sigma- Aldrich - Sigma-Aldrich, St Louis, MO, USA.

The biotinylated probes were then annealed to form double-stranded probes. To do this equal volumes of the complementary probes were added to 1/10th volume of 10X annealing buffering (made up of 200 mM Tris pH 7.6, 100 mM $MgCl_2$ and 100 mM NaCl) and vortexed gently. An annealing reaction was then performed using the conditions shown in Table 18.

Table 18: Annealing reaction conditions

Temperature	Time
95°C	5 mins
95°C-75°C	25 cycles
4°C	HOLD

2.4.2.3 EMSA binding reactions

Double-stranded biotinylated probes were incubated with nuclear extract. The reaction volumes used are shown in Table 19. Two components of the reaction, poly(deoxyinosinic-deoxycytidylic) acid sodium salt (dIdC) and random sequence oligonucleotides are included as substrates for non-specific DNA binding by proteins in the nuclear extract. All components were combined except for the biotinylated probe and this mix incubated at 4°C for 30 minutes. Thereafter the biotinylated probes were added and the reaction incubated at 25°C for 50 minutes.

Table 19: EMSA binding reaction composition

Reagent	Non-competitor EMSA Volume (µl)	Competitor EMSA Volume (µl)
10X Binding buffer	2	2
dIdC	1	1
50 mM $MgCl_2$ (Thermo Fisher Scientific)	0.5	0.5
Random sequence	1.5	1.5
dH_20	10	6
Nuclear extract	3	3
Competitor probe	-	4
Biotinylated probe	2	2
	20	20

10X binding buffer is 100mM Tris and 500mM KCl, pH 7.5. didC= poly(deoxyinosinic-deoxycytidylic) acid sodium salt. dH_20=distilled water. Thermo Fisher Scientific - Waltham, MA, USA.

2.4.2.4 EMSA Polyacrylamide gel electrophoresis

To separate DNA-bound probes from free probes, the EMSA binding reaction was run on a 6% polyacrylamide gel (1 x1.5 mm). The volume of regents used to prepare the gel is shown in Table 20. Gels were left to set overnight. Prior to running EMSA binding reactions, 15 µl of loading buffer (50% v/v 10X xylene cyanol CFF loading buffer/10X bromophenol blue loading buffer) was run at 120 V and 4°C for approximately 90 minutes in 0.5X TBE (Tris-Borate-EDTA) buffer.

EMSA binding reactions were mixed with 5 µl loading buffer and 16-20 µl of this was loaded on the gel, ensuring an equal volume was loaded into each well. The gel was run for approximately 4 hours at 120 V and 4^{o}C.

Table 20: EMSA polyacrylamide gel reagent volumes

Reagent	Volume
37.5: Acrylamide (Severn Biotech)	12 ml
10X TBE	6 ml
TEMED (Sigma-Aldrich)	50 µl
10 % APS	500 µl
dH_20	47 ml

TBE=Tris Borate EDTA, APS= Ammonium Persulphate, TEMED=Tetramethylethylenediamine. dH_20=distilled water. Severn Biotech – Kidderminster, UK. Sigma-Aldrich- St Louis, MO, USA. EMSA=electrophoretic mobility shift assay.

2.4.2.5 EMSA blotting and detection

The contents of the gel were transferred on to a hybond-N+ membrane using Southern transfer (Southern 1975). To fix the DNA-protein complexes to the membrane, cross-linking was performed at ~254 nm for 3 minutes using the UV Stratalinker2400 (Stratagene, La Jolla, CA, USA). Detection was carried out using chemiluminescence with the LightShift Chemiluminscent EMSA kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions. The membrane was then exposed to X-ray for a minimum of one minute. The film was developed according to the manufacturer's instructions using the SRX-101A film processor (Konica Minolta Medical Imaging, Wayne, NJ, USA).

2.4.3 Luciferase assay

The luciferase dual-reporter assay system (Promega) can be used to quantitatively analyse the impact of genetic variation on mammalian gene expression (Sherf, Navarro et al. 1996; Smith, D'Aiuto et al. 2008; Khamis, Palmen et al. 2015). The system involves transfecting a relevant cell line with two vectors, each containing the sequence corresponding to one of the reporter genes (firefly luciferase gene (*luc+*) and *Renilla* luciferase gene). Gene expression is quantified by measuring the luminescent signal generated by the reporter enzymes in the presence of their substrate. The ratio of *luc+* expression to *Renilla* luciferase expression is then calculated and compared between experimental conditions (e.g. comparing two vectors with different alleles for single SNP). This technique was used to assess the impact of one SNP, rs28451064 on gene expression.

2.4.3.1 Cloning

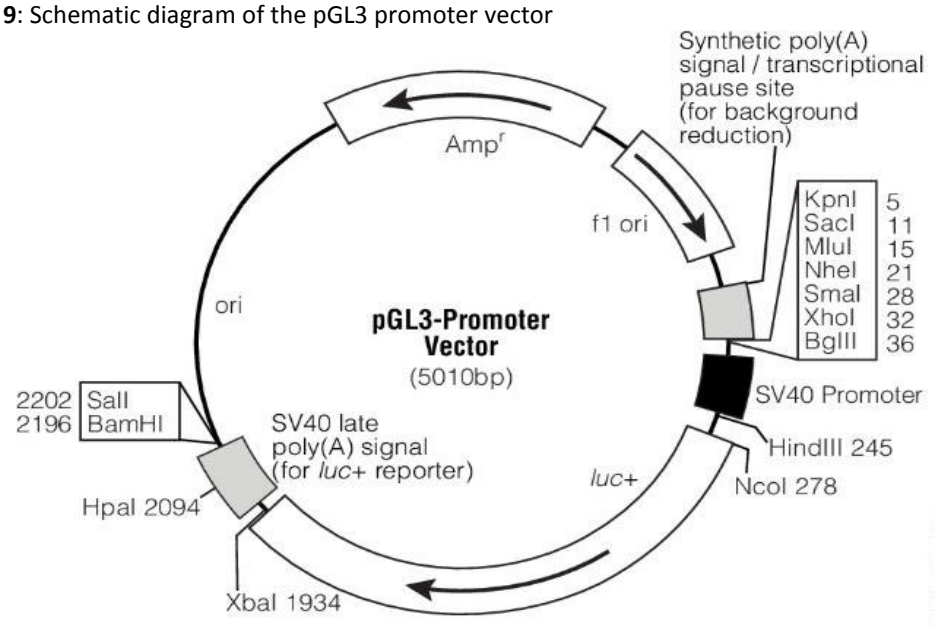
To generate the required vectors, the sequence surrounding rs28451064 was cloned into the pGL3-promoter vector (Figure 9), downstream of the *luc+* gene between the SalI and BamHI restriction sites. This was carried out using the InFusion kit (Clonetech Laboratories, Mountain View, CA, USA). Here, the primers used to amplify the region surrounding the SNP have sequences corresponding to the restriction enzyme sites (SalI and BamHI) in the vector at the 5' end. Thus amplified product containing the SNP sequence can be incorporated into the pre-digested vector by homologous recombination. The primer sequences surrounding the SNP were designed using primer3 (http://primer3.ut.ee/) and the restriction fragment sequence added to the 5' end were designed using the Clontech tool (http://bioinfo.clontech.com/infusion/convertPcrPrimersInit.do) (Table 22). The genomic fragment was amplified using the Phusion High Fidelity PCR kit (New England Biolabs, Ipswich, MA, USA), according to the manufacturer's instructions. The pGL3-promoter vector was digested using the conditions listed in Table 21 and was gel purified using the GFX PCR and Gel Band Purification kit (GE Healthcare, Thermo Fisher Scientific Waltham, MA, USA) according to the manufacturer's instructions. The cloning ligation reaction was then performed with the InFusion kit according to the manufacturer's instructions.

Table 21: Reaction components for plasmid digression with restriction enzymes

Component	Volume (µl)
Plasmid (400 ng/µl)	2.5
SalI	2.5
BamHI	2.5
10x Restriction Enzyme Buffer 3.1	5
dH_20	37.5
	50

Enzymes and buffers supplied by New England Biolabs (Ipswich, MA, USA). dH_20=distilled water.

Figure 9: Schematic diagram of the pGL3 promoter vector



Reproduced with permission from Promega Corporation.

The cloned vector was then transformed into *E.coli* DH5α cells (New England Biolabs, Ipswich, MA, USA) by mixing a 50 µl aliquot of cells with the InFusion reaction mix and leaving this on ice for 20-30 minutes. The mix was then incubated at 42°C for one minute and put back on ice for two minutes. The mix was then added to 500 µl of pre-warmed Lysogeny broth (LB) and incubated for one hour, shaking at 37°C. Thereafter, this mix was spread on to an LB agar plate. Plates were incubated overnight at 37°C. The plasmid DNA was then purified from a single colony using the QIAprep Spin MiniPrep kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions. The enhancer region of the purified plasmid was sequenced using Sanger sequencing (as previously described in section 2.3.2, using a primer with a sequence corresponding to the region of the plasmid the insert was cloned into) to check that the insert was present and to determine which allele was present. To generate a plasmid stock with a high concentration a maxiprep was then performed using the GenElute™ HP Plasmid Maxiprep kit (Sigma-Aldrich, St Louis,

MO, USA), according to the manufacturer's instructions. In order to generate a plasmid with the other allele at the rs28451064 position, the QuickChange Lightning Site-Directed Mutagenesis kit (Agilent Technologies, Santa Clara, CA, USA) was used, according to the manufacturer's instructions. The primers for this were designed using the online tool (www.genomics.agilent.com) (Table 22). The plasmid DNA was purified and sequenced to check the required base change had been made as previously described (section 2.3.2). Thereafter, this fragment was re-cloned into a fresh pGL3-promoter vector using the InFusion kit and the resulting plasmid DNA was purified and concentrated as described in the preceding paragraph.

Table 22: Primers used to generate pGL3promoter vectors with the rs28451064 insert

Primer	Sequence
rs28451064 Cloning forward primer	AAATCGATAAGGATCCCCAGGCACCAGGTAGACTTA
rs28451064 Cloning reverse primer	AAGGGCATCGGTCGACTCTCAGAACTTTACAGAACGCG
SDM Forward primer	CTGGGAGTATTTGGTGTCTACTTTGGCCTGGTAAATT
SDM Reverse Primer	AATTTACCAGGCCAAAGTAGACACCAAATACTCCCAG

SDM=site directed mutagenesis. The sequence of the cloning primers containing the BamHI and SalI restriction sites in the pGL3promoter vector is shown in red.

2.4.3.2 Transfection and luciferase assay

Cells were grown to >90% confluence, trypsinised as previously described (section 2.4.1) and counted using the Advanced Detection and Accurate Measurement (ADAM) cell counter and the Accuchip kit (Digital Bio Pharm, Cambridge UK). Cells were then plated into a 96 well plate at different concentrations ($1x10^5$/ml-$5x10^5$/ml) and left for approximately 24 hours. The plate with cells that were ~70-90% confluent were selected to perform the assay with.

Transfection was performed using either Lipofectamine 2000 or Lipofectamine 3000 (Invitrogen, Life Technologies, Carlsbad CA, USA) according to the manufacturer's instructions. Four control vectors were used: puc19, which does not contain the *luc+* gene, the pGL3control plasmid (containing the *luc+* gene, SV40 bacterial promoter and enhancer), pGL3 promoter plasmid (containing the *luc+* gene, SV40 bacterial promoter but not the enhancer) and the pGL3-basic plasmid (which does not contain the SV40 bacterial promoter or enhancer but does have the *luc+* gene). Each vector was added to twelve wells. The plate was then left for approximately 48 hours at 37°C. The pRL-TK plasmid which contains the *Renilla* luciferase gene was co-transfected into all wells except those with puc19.

To determine levels of luciferase activity, luminescence was detected using the Tropix TR717 Microplate Luminometer with the WinGlow software (Applied Biosystems, Life Technologies, Carlsbad, CA, USA) and the Dual Luciferase assay kit (Promega, Madison, WI, USA) according to the manufacturer's instructions. The ratio of firefly luciferase readings : *Renilla* luciferase readings was then compared between the different constructs using paired t-tests.

2.5 Statistical analysis

2.5.1 General statistical analysis

All statistical analyses were performed using R (R Core Team 2015), unless otherwise stated. Meta-analyses were performed using the R package "metafor" using either a fixed effects (FE) or random effects (RE) (DerSimonian Laird) model (Viechtbauer 2010). Where specific packages have been used to generate figures, this has been indicated. Power calculations for genetic association studies were performed using QUANTO software (Gauderman and Morrison 2001).

2.5.2 Calculation of CHD risk prediction scores

2.5.2.1 CHD GSs

Unweighted GSs were calculated by adding the number of risk alleles present together. Weighted GSs were calculated by multiplying the number of risk alleles at each risk locus by the natural log of the odds ratio (OR) associated with each risk allele (i.e. an additive model). If a recessive model was used, homozygotes for the protective allele and heterozygotes were assigned zero and homozygotes for the risk allele were assigned the natural log of the OR associated with that genotype. These individual "SNP scores" were then added to together to give an overall GS.

2.5.2.2 Framingham risk score

The Framingham risk score was calculated using the equations given in (Wilson, D'Agostino et al. 1998). To combine the Framingham CRF risk with genetic risk, the population mean adjusted Framingham score was combined with the population mean adjusted GS, which is calculated by subtracting the mean population GS based on population risk allele frequencies (RAFs) from the individual's GS. This value was then exponentiatied to give the relative odds ratio (OR) for CHD. Combined ten-year CHD risk was then calculated by incorporating this into the Framingham score survival function (Wilson, D'Agostino et al. 1998).

2.5.2.3 UK Prospective Diabetes Study risk score

The UK Prospective Diabetes Study (UKPDS) risk score was calculated using the equations given in (Stevens, Kothari et al. 2001) to calculate the value "q" from the CRFs values present. To combine the GS with the UKPDS score, the relative OR for CHD was converted to relative risk (incidence of CHD in this group was set at 0.3 – calculated using data

published by the British Heart Foundation (Scarborough, Wickramasinghe et al. 2011)) and included as a term in the calculation of "q". Combined ten-year CHD risk could then be calculated by exponentiating the product by (1-duration of T2D^10)/1-duration of T2D) and subtracting it from one.

2.5.2.4 QRISK2 risk score

QRISK2 was calculated under licence. To combine the QRISK2 score with the GS, the QRISK2 risk was converted to the natural log scale and added to the centred GS (using the population GS) to give value a. This value was then converted back to ten-year CHD probability using the equation: 1/1+e^(-a).

2.5.3 Metrics used to assess risk prediction

2.5.3.1 Calibration

"Calibration" refers to how well the predicted event rate determined using a risk score corresponds to the observed event rate. This was assessed with the Hosmer-Lemeshow test, which is used to determine the "goodness of fit" of a logistic regression model (Hosmer and Lemeshow 1980), such as those used in risk prediction. The sample is divided into deciles according to the risk score values. The expected number of events and observed number of events in the cohort is then determined and a chi-squared test performed using these values. Ten degrees of freedom were used rather than eight as the risk score values and the disease incidence data were obtained independently. A large p-value indicates that the predicted and observed rates are similar and thus that the model used is well calibrated. Hosmer-Lemeshow tests were performed in R using the hoslem.test() function which is part of the "ResourceSelection" package (Lele, Keim et al. 2014). The p-values were calculated from the Hosmer-Lemeshow chi-squared value using the code 1-pchisq(q,df), where q=chi-squared statistic determined in the Hosmer-Lemeshow test and df=the number of degrees of freedom, which was set at ten.

2.5.3.2 Discrimination

The ability of a risk score to discriminate between those who did and did not have an event was assessed using a receiver operator characteristic (ROC) curve. This is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity)(Fawcett 2006). Each participant in a data set has a risk score value and a known outcome. As the outcome is binary, whether the risk score value predicts an event or not depends on the assigned

threshold. Therefore, at a particular threshold the true positive and false positive rates can be calculated. The ROC curve is a graphical summary of the true and false positive rates at all possible thresholds. The area under the ROC curve (AUROC) can then be calculated and this is a measure of how well the risk model discriminates between individuals. An AUROC of 0.5 indicates that the model discriminates poorly, as this value would be expected for a binary outcome by chance whereas an AUROC of 1 indicates perfect discrimination. However, being a rank-based measure the AUROC it can be relatively insensitive to the addition of robustly associated risk factor, with only a small increase in the AUROC observed (ref-Cook). ROC curves were made and AUROC calculated using the "pROC" package in R (Robin, Turck et al. 2011). AUROCs were compared using De Long's test.

2.5.3.3 Reclassification

As discussed in Chapter 1.6.5, CHD risk prediction scores are used to categorise individuals into low or high risk groups. Whether addition of a novel marker to an established risk score improves the classification of individuals can be assessed using the net reclassification index (NRI) (Pencina, D'Agostino et al. 2008). In order to improve risk classification in those who have had an event, risk scores should increase such that those who were classified in the low risk group previously now fall into the high risk group. Conversely to improve classification in the non-event group, risk scores should be lower leading to individuals who were classified as high risk moving into the low risk category. The NRI is the sum of the "event NRI" and the "non-event NRI". The event NRI is calculated by dividing the number of individuals (who had events) who moved up a risk category by the total number of events and subtracting the number of individuals (who had events) who moved down a risk category divided by the number of events. The non-event NRI is calculated in the same manner for those who did not have events. The NRI itself is then calculated by combining these values and as such is not itself a proportion but a unit-less statistic (Leening, Vedder et al. 2014). NRI is sensitive to calibration and give misleading results for poorly calibrated models(Hilden and Gerds 2014). NRI was calculated using the reclassification() function of the R package "PredictABEL"(Kundu, Aulchenko et al. 2011).

2.5.4 Analysis of metabolomics data

Metabolomic traits were determined using a nuclear magnetic resonance (NMR) based platform. All metabolomic measures were adjusted for age, age^2 and sex and an inverse rank transformation was used prior to association analysis (Blom 1958). This was performed using a linear model, adjusted for lipid lowering medication use, in each cohort individually. Separate analysis was performed for those with and without prevalent T2D. The results from the different studies were combined in a FE meta-analysis weighted by sample size. To account for multiple testing and the correlation between the metabolomic traits, p-values were adjusted using the false discovery rate (FDR) from Benjamini-Hochberg-Yekutieli (Benjamini and Yekutieli 2001). An FDR adjusted p-value $p<0.05$ was considered to be statistically significant.

3 Assessment of a CHD GS in the UK and other populations

3.1 Introduction

Data from twin studies has estimated the heritability of CHD mortality to be 40-60% (Zdravkovic, Wienke et al. 2002; Wienke, Herskind et al. 2005) and dozens of loci have been robustly associated with the disease (Casas, Cooper et al. 2006; Deloukas, Kanoni et al. 2013). This has led to development of GSs where a number of risk variants are combined to give an estimate of genetic CHD risk. In order to take account of the different impact that individual risk loci have on CHD risk, each variant can be weighted using its effect size rather than simply constructing a score through allele counting (i.e. an unweighted GS). In addition to the scientific value of developing a CHD GS, they can also provide a tool through which genetic CHD risk can be incorporated into risk prediction in a clinical setting. An individual's GS can be adjusted for the population GS (based on the RAFs present in the population) and then be combined with a CRF score such as the Framingham score or QRISK2 (Hippisley-Cox, Coupland et al. 2008) to give an overall CHD risk estimate (Chapters 2.5.2.2and 2.5.2.4).

Nineteen SNPs from candidate gene studies and GWASs were selected from the literature for inclusion in a CHD GS. The details are given in Table 23. Eight of the SNPs are non-synonymous, seven are located in introns, one is located in a promoter, another in a 3'-untranslated region (3'-UTR) and two are intergenic. The list was finalised in 2010. Since then the field has developed rapidly, most notably with the publication of the CARDIoGRAM GWAS results (Schunkert, Konig et al. 2011) and the subsequent CARDIoGRAMplusC4D meta-analysis (Deloukas, Kanoni et al. 2013) as is discussed in section 3.2.3.1 of this Chapter. Thirteen of the 19 GS SNPs are found in CHD risk loci/genes identified in the CARDIoGRAMplusC4D meta-analysis, although not all of these SNPs are in LD with the corresponding lead SNP. All SNPs were treated additively in the GS, except for rs1799983 which was treated in a recessive manner, due to the association found in the source publication. A kit to genotype all 19 SNPs simultaneously was developed (the Randox Cardiac Risk Prediction array, Chapter 2.3.3) to enable the GS to be used in a clinical context.

Table 23: SNPs included in the CHD risk GSs.

Gene	SNP	Location	Risk Allele	OR	Reference
*APOE**	rs7412	C158R	T[++]	0.80	Bennet, Di Angelantonio et al. (2007)
*APOE**	rs429358	C112R	C	1.06	Bennet, Di Angelantonio et al. (2007)
*MIA3**	rs17465637	Intronic	C	1.14	Samani, Erdmann et al. (2007)
*MRAS**	rs9818870	3'-UTR	T	1.15	Erdmann, Grosshennig et al. (2009)
DAB2IP	rs7025486	Intronic	A	1.16	Harrison, Cooper et al. (2012)
*CXCL12**	rs1746048	Intergenic	C	1.17	Samani, Erdmann et al. (2007)
*APOA5**	rs662799	Promoter Variant	G	1.19	Sarwar, Sandhu et al. (2010)
*SORT1**[+]	rs646776[+]	Intergenic	A	1.19	Kathiresan et al. (Kathiresan, Altschuler et al. 2009)
SMAD3	rs17228212	Intronic	C	1.21	Samani, Erdmann et al. (2007)
ACE	rs4341	Intronic	G	1.22	Casas, Cooper et al. (2006)
*LPL**	rs328	S447X	C	1.25	Casas, Cooper et al. (2006)
CETP	rs708272	Intronic	C	1.28	Casas, Cooper et al. (2006)
CDKN2A /9p21*	rs10757274	Intronic	G	1.29	Samani, Erdmann McPherson, Pertsemlidis et al. (2007); (Samani, Erdmann et al. 2007)
NOS3	rs1799983	E298D	T	1.31	Casas, Baustista (Casas, Bautista et al. 2004)
LPL	rs1801177	D9N	A	1.33	Sagoo, Tatt et al. (2008)
*PCSK9**	rs11591147	R46L	G	1.43	Benn, Nordestgaard et al. (2010)
*LPA**	rs10455872	Intronic	G	1.70	Clarke, Peden et al. (2009)
*APOB**	rs1042031	E4181K	A	1.73	Casas, Cooper et al. (2006)
*LPA**	rs3798220	I1891M	C	1.92	Clarke, Peden et al. (2009)

SNPs marked with an asterisk (*) are included in both the 19 and 13 SNP GS. [+]rs599839 was genotyped in the studies herein instead of rs646776, r^2=0.95 in Europeans. [++]For rs7412, the protective SNP is included in the GS. LD taken from the, 1000 Genomes phase 1 EUR data. OR=odds ratio. GS= gene score. 3'-UTR= 3'-untranslated region. LD=linkage disequilibrium.

The first aim of this study was to investigate the use of the 19 SNP CHD GS in CHD risk prediction in the UK population using the prospective study, NPHSII. Use of the 19 SNP GS was also assessed in both the South Asian and Afro-Caribbean populations using case-control studies from Pakistan and Guadeloupe. A GS comprising only the 13 SNPs located in confirmed CHD risk loci/genes was also assessed in the three ethnic groups to compare its performance to that of the 19 SNP GS. A further aim of the study was to perform a literature search to identify variants associated with CHD that would be suitable for future GSs. Following on from this the final aim was to update the CHD GS using the results of the literature search and to assess whether this improved its performance.

3.2 Results

3.2.1 Assessment of GS in the UK population

3.2.1.1 Baseline characteristics of NPHSII participants

The baseline characteristics of the participants of NPHSII are presented in Table 24. As expected, the men who went on to develop CHD were older, had higher BMI, higher systolic blood pressure, higher total cholesterol, LDL cholesterol, and a higher proportion were smokers and had diabetes, at baseline. Furthermore, those who subsequently developed CHD had a higher ten-year CHD risk as calculated using the Framingham risk score and those who subsequently developed CVD had a higher ten-year CVD risk as calculated using the QRISK2 score.

Table 24: Baseline characteristics in NPHSII for those who did and did not go on to develop CHD during ten-year follow-up

Trait	NPHSII No CHD (n=2491)	NPHSII CHD (n=284)	p-value
Age (years)	55.91 (3.42)	56.64 (3.60)	4.12×10^{-3}
Sex (% Male)	100 %	100 %	-
Smoking	25 %	39 %	2.14×10^{-5}
BMI (kg/m^2)	26.38 (3.42)	27.19 (3.44)	9.61×10^{-4}
Systolic Blood Pressure (mmHg)	137.00 (18.59)	144.09 (20.10)	9.68×10^{-7}
Total Cholesterol (mmol/l)	5.71 (1.01)	6.13 (1.05)	4.79×10^{-8}
LDL-cholesterol (mmol/l)	3.07 (1.00)	3.48 (0.97)	2.66×10^{-7}
HDL-cholesterol (mmol/l)	1.72 (0.59)	1.57 (0.53)	2.60×10^{-4}
Diabetes	2 %	7 %	1.33×10^{-11}
Framingham ten-year CHD risk	0.12 (0.07-0.15)	0.17 (0.09-0.21)	4.33×1011^{-4}
QRISK2 ten-year CVD risk*	0.09 (0.07-0.13)	0.13 (0.09-0.17)	1.93×10^{-14}

All variables are presented as the mean plus standard deviation, unless otherwise stated. Categorical variables were compared using chi-squared tests and continuous variables were compared using Welch's t-tests, apart from the Framingham and QRISK2 risk scores which were compared using Mann Whitney tests (the median and interquartile range are given). *QRISK2 values shown are for those who did and did not go on to develop CVD.

3.2.1.2 GSs in NPHSII

The genotype distribution and RAF of each of the 19 SNPs in NPHSII is shown in Table 25. All SNPs except rs1042031 in *APOB* were in Hardy-Weinberg equilibrium (HWE). A comparison of the RAF in those who did not go to develop CHD and those who did is presented in Table 26. The RAF was higher in those who did develop CHD for rs10757274 at the 9p21 locus (0.48 v 0.64 $p=6 \times 10^{-3}$) and rs1746048 which is located close to the gene *CXCL12* (0.86 v 0.89, p=0.03). Both weighted GSs were higher in those who developed CHD in the ten-year follow-up period (19 SNP GS $p=4.54 \times 10^{-3}$, 13 SNP GS $p=2.63 \times 10^{-3}$). Both the 19 SNP and 13 SNP GSs were associated with CHD (Table

27 and Table 28). The unweighted 19 SNP score was associated with total cholesterol, LDL-cholesterol and HDL-cholesterol, however, only the association with HDL-cholesterol remained with the weighted score (Table 29). Both the weighted and unweighted 19 SNP GSs were associated with the Framingham risk score while only the unweighted score was associated with QRISK2.

Table 25: Genotype distribution and risk allele frequency for each SNP in all NPHSII participants

Gene/Locus	SNP	Genotype Distribution			NPHSII RAF (95% CI)	HWE p-value
APOE	rs429358	**TT**	**TC**	**CC**	0.17 (0.16-0.18)	0.46
		1672	675	61		
APOE	rs429358	**TT**	**TC**	**CC**	0.17 (0.16-0.18)	0.46
		1672	675	61		
MIA3	rs17465367	**CC**	**CA**	**AA**	0.71 (0.69-0.72)	0.97
		1360	1135	236		
MRAS	rs9818870	**CC**	**CT**	**TT**	0.16 (0.15-0.17)	0.86
		1924	709	67		
DAB2IP	rs7025486	**GG**	**GA**	**AA**	0.26 (0.17-0.24)	0.28
		1498	997	185		
CXCL12	rs1746048	**TT**	**TC**	**CC**	0.86 (0.85-0.87)	0.85
		52	641	2035		
APOA5	rs662799	**AA**	**AG**	**GG**	0.06 (0.05-0.07)	0.15
		1793	285	14		
SORT1	rs646776	**AA**	**AG**	**GG**	0.78 (0.77-0.79)	0.18
		1685	902	140		
SMAD3	rs17228212	**TT**	**TC**	**CC**	0.31 (0.30-0.32)	0.09
		1325	1123	277		
ACE	rs4341	**CC**	**CG**	**GG**	0.52 (0.50-0.53)	0.32
		643	1328	740		
LPL	rs328	**CC**	**CG**	**GG**	0.90 (0.89-0.94)	0.77
		2187	497	30		
CETP	rs708272	**CC**	**CT**	**TT**	0.56 (0.55-0.58)	0.07
		798	1320	471		
CDKN2A/ 9p21	rs10757274	**AA**	**AG**	**GG**	0.48 (0.47-0.50)	0.41
		733	1324	637		
NOS3	rs1799983	**GG**	**GT**	**TT**	0.33 (0.32-0.35)	0.91
		1083	1080	272		
LPL	rs1801177	**GG**	**GA**	**AA**	0.01 (0.01-0.02)	0.08
		2413	64	2		
PCSK9	rs11591147	**GG**	**GT**	**TT**	0.99 (0.99-0.99)	0.18
		2401	42	1		
LPA	rs10455872	**AA**	**AG**	**GG**	0.07 (0.07-0.08)	0.42
		2266	373	12		
APOB	rs1042031	**GG**	**GA**	**AA**	0.18 (0.17-0.19)	0.04
		1763	790	66		
LPA	rs3789220	**TT**	**TC**	**CC**	0.02 (0.01-0.02)	1.00
		2636	90	0		

RAF=risk allele frequency. HWE=Hardy-Weinberg equilibrium, CI=confidence interval.

Table 26: Comparison of allele frequencies in those who did and did not go on to develop CHD during ten-year follow-up of NPHSII

Gene/Locus	SNP	NPHSII No CHD RAF (95% CI)	NPHSII CHD RAF (95% CI)	p-value
APOE	rs429358	0.16 (0.15-0.18)	0.18 (0.15-0.22)	0.40
APOE	rs7412	0.91 (0.90-0.91)	0.93 (0.91-0.95)	0.06
MIA3	rs17465367	0.71 (0.69-0.728)	0.71 (0.67-0.75)	0.70
MRAS	rs9818870	0.15 (0.14-0.16)	0.18 (0.15-0.21)	0.16
DAB2IP	rs7025486	0.25 (0.24-0.27)	0.28 (0.24-0.32)	0.16
CXCL12	rs1746048	0.86 (0.85-0.87)	0.89 (0.87-0.92)	0.03
APOA5	rs662799	0.06 (0.05-0.07)	0.06 (0.04-0.09)	0.79
SORT1	rs646776	0.78 (0.77-0.79)	0.79 (0.76-0.83)	0.61
SMAD3	rs17228212	0.31 (0.30-0.32)	0.29 (0.26-0.33)	0.42
ACE	rs4341	0.52 (0.50-0.53)	0.52 (0.47-0.56)	0.96
LPL	rs328	0.90 (0.89-0.90)	0.91 (0.88-0.93)	0.26
CETP	rs708272	0.57 (0.55-0.58)	0.54 (0.50-0.59)	0.31
CDKN2A/ 9p21	rs10757274	0.48 (0.462-0.490)	0.54 (0.50-0.58)	$6x10^{-3}$
NOS3	rs1799983	0.34 (0.32-0.35)	0.31 (0.27-0.35)	0.25
LPL	rs1801177	0.01 (0.01-0.02)	0.02 (0.01-0.04)	0.24
PCSK9	rs11591147	0.99 (0.99-0.99)	0.996 (0.99-1.000)	0.32
LPA	rs10455872	0.07 (0.07-0.08)	0.09 (0.07-0.12)	0.11
APOB	rs1042301	0.18 (0.16-0.19)	0.19 (0.15-0.22)	0.71
LPA	rs3789220	0.02 (0.01-0.02)	0.02 (0.01-0.04)	0.18

Allele frequencies compared using tests of proportion. CI=confidence interval. RAF=risk allele frequency. CHD=coronary heart disease.

Table 27: Comparison of mean GSs in those who did and did not go on to develop CHD during ten-year follow-up of NPHSII

Trait	NPHSII No CHD	NPHSII CHD	p-value
19 SNP unweighted GS	16.07 (2.09) n=1090	16.74 (1.92) n=110	8.26×10^{-4}
19 SNP weighted GS	3.16 (0.53) n=1090	3.31 (0.51) n=110	4.54×10^{-3}
13 SNP unweighted GS	12.63 (1.69) n=1374	13.18 (1.51) n=133	1.21×10^{-4}
13 SNP weighted GS	2.44 (0.49) n=1374	2.56 (0.44) n=133	2.63×10^{-3}

Mean GS plus standard deviation is show in all cases GSs were compared using Welch's t-test. GS=gene score. CHD=coronary heart disease.

Table 28: Association between the GSs and CHD in NPHSII.

Score	Effect Size (95% CI)	p-value
19 SNP GS -unweighted	1.38 (1.13-1.69)	1.55×10^{-3}
19 SNP GS- weighted	1.32 (1.05-1.47)	5.30×10^{-3}
13 SNP GS - unweighted	1.39 (1.16-1.67)	3.77×10^{-4}
13 SNP GS - weighted	1.29 (1.08-1.53)	4.24×10^{-3}

Effect sizes relate to one standard deviation of the variable and were determined by logistic regression, adjusted for age. CI=confidence interval. GS=gene score. CHD=coronary heart disease.

Table 29: Association of 19 SNP GS with CHD risk factors and CRF scores

Trait	19 SNP unweighted GS p-value	19 SNP weighted GS p-value
Smoking	0.15	0.28
BMI	0.08	0.11
Hypertension	0.42	0.38
Cholesterol	0.02	0.18
LDL-cholesterol	9.74×10^{-3}	0.08
HDL-cholesterol	9.69×10^{-3}	0.04
Diabetes	0.78	0.90
Family History of CHD	0.16	0.22
Framingham	7.28×10^{-4}	4.93×10^{-3}
QRISK2	0.04	0.11

The p-values were determined using linear or logistic regression (unadjusted), as appropriate. SNP=single nucleotide polymorphism. GS=gene score. CHD=coronary heart disease. CRF=conventional risk factor.

3.2.1.3 Addition of GSs to CRF scores

As shown in Table 30, complete data (ten-year CHD outcome, genotyping and CRF score) was available for 1022 individuals for the Framingham score plus 19 SNP GS and 1272 individuals for the Framingham score plus 13 SNP GS. Complete data (ten-year CVD outcome, genotyping and CRF score) was available for 1213 NPHSII participants for QRISK2 plus 19 SNP GS and for 1522 participants for QRISK2 plus 13 SNP GS. To combine the CRF with the Framingham score (and indeed QRISK2) the population GS was calculated using the effect sizes and allele frequencies from the source publications (Table 23). To assess the CRF score alone and CRF plus GS scores the number of predicted events was compared to those observed (i.e. calibration of the score) using a Hosmer-Lemeshow goodness-of-fit test. As shown in Figure 10, overall the Framingham score showed poor calibration in NPHSII ($p=2.94x10^{-6}$) and this was worse after addition of the both GSs (both $p<2.94x10^{-6}$). Whereas, QRISK2 showed good calibration in NPHSII (p=0.35) but again calibration was worse after addition of both GSs (19 SNP $p=1.12x10^{-3}$ and 13 SNP $p=4.30x10^{-4}$, Figure 11).

Next the predictive ability (as assessed using the AUROC) of the CRF scores was compared to the combined CRF plus GS risk scores (Figure 12). No statistically significant increase in the AUROC was observed with the addition of either of the GSs to the CRF scores (all p>0.05). Addition of the GSs to the CRF scores did not lead to a significant improvement in risk classification as shown in the reclassification table, which includes the NRI values (

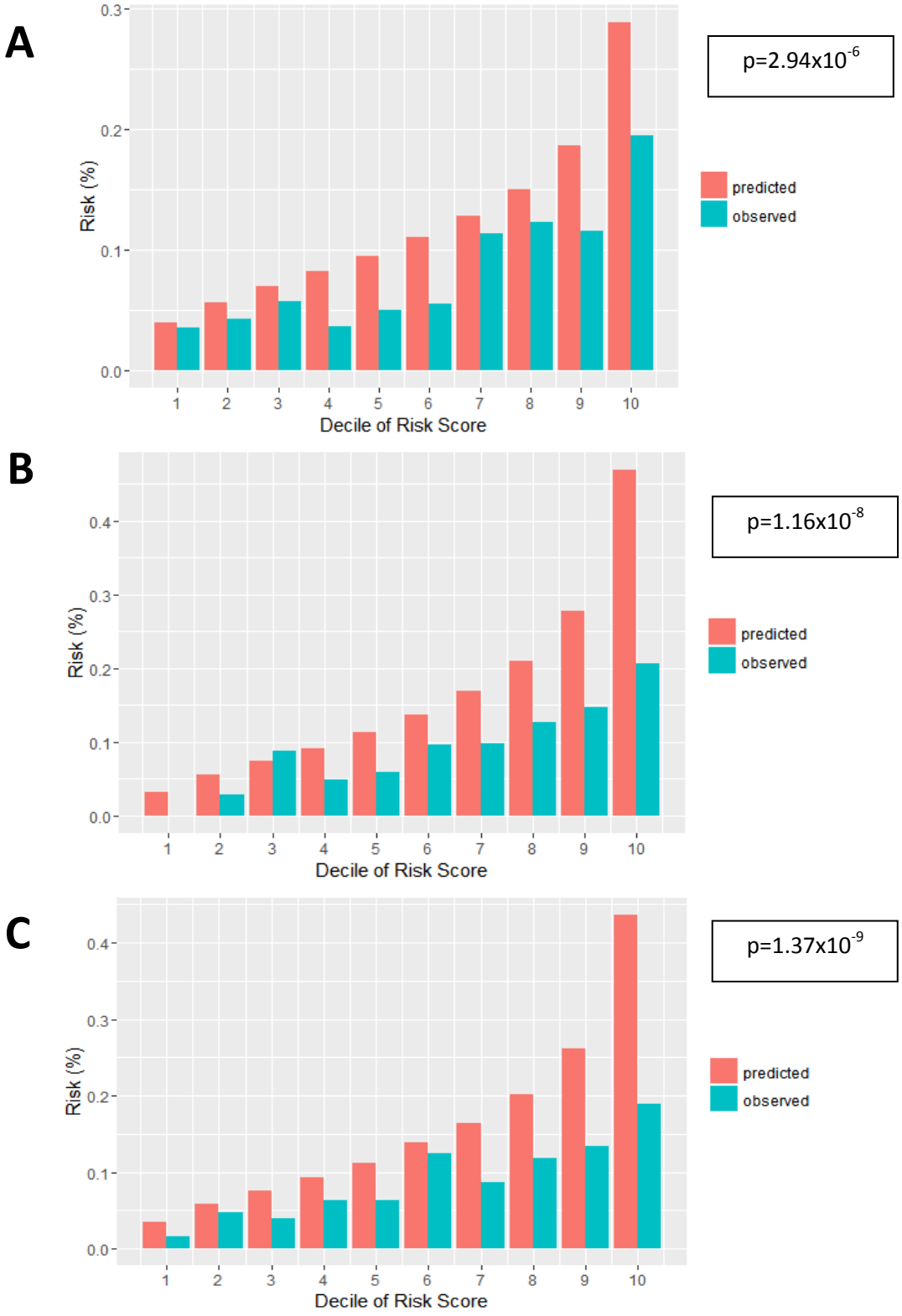
Table 32).

Table 30: Number of NPHSII participants with complete data for each risk score plus GS combination after follow-up of ten years

Score	No CHD n	CHD n	Total N
FRAM + 19 SNP GS	930	92	1022
FRAM + 13 SNP GS	1160	112	1272
Score	No CVD n	CVD n	
QRISK2 + 19 SNP GS	1080	133	1213
QRISK2 + 13 SNP GS	1356	166	1522

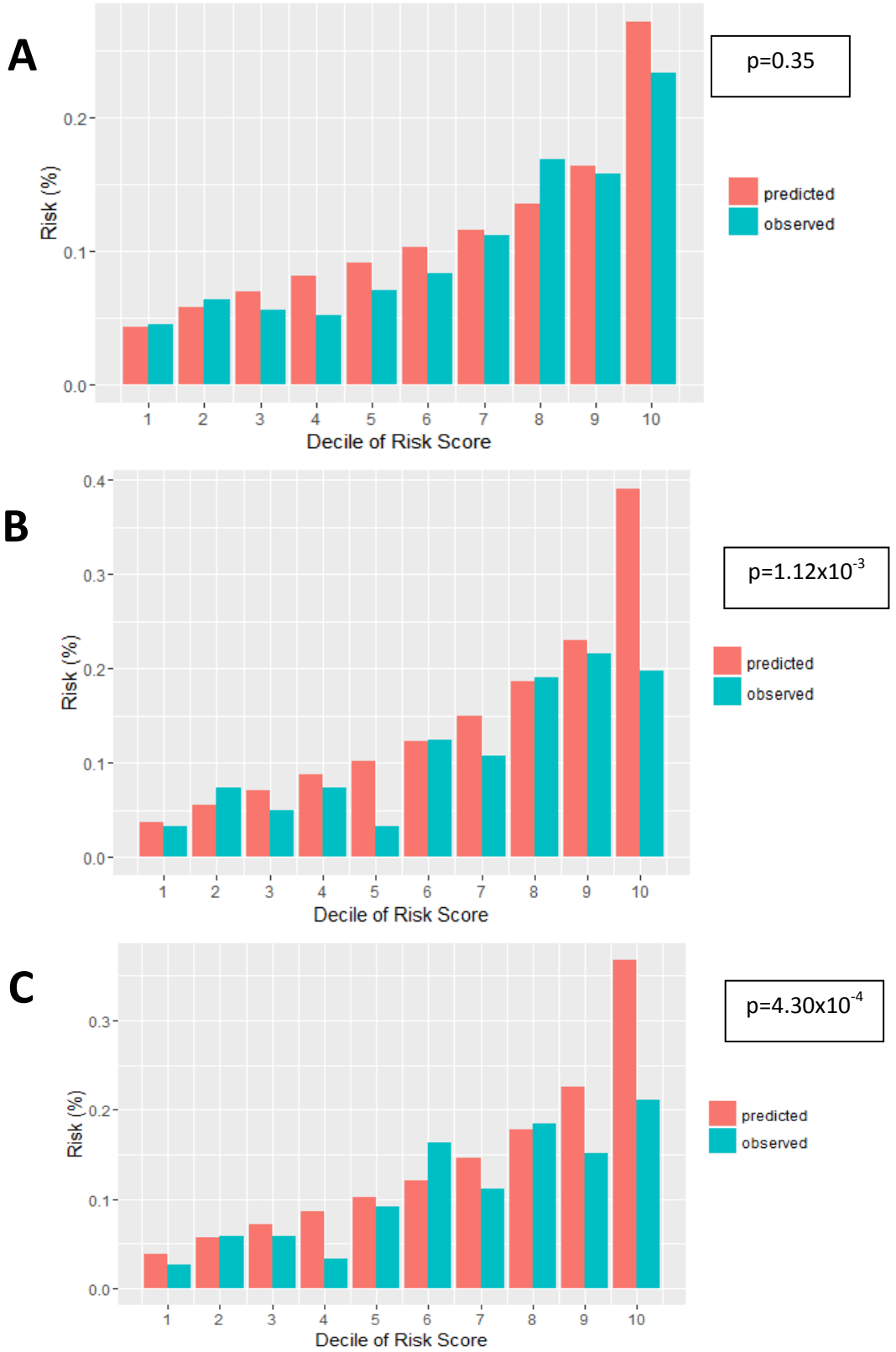
FRAM=Framingham risk score. GS=gene score. CHD=coronary heart disease. SNP=single nucleotide polymorphism.

Figure 10: Observed CHD event rate in NPHSII compared to the predicted event rate determined by A) Framingham score alone; B) Framingham plus 19 SNP GS; C) Framingham plus 13 SNP GS, presented by decile of risk score



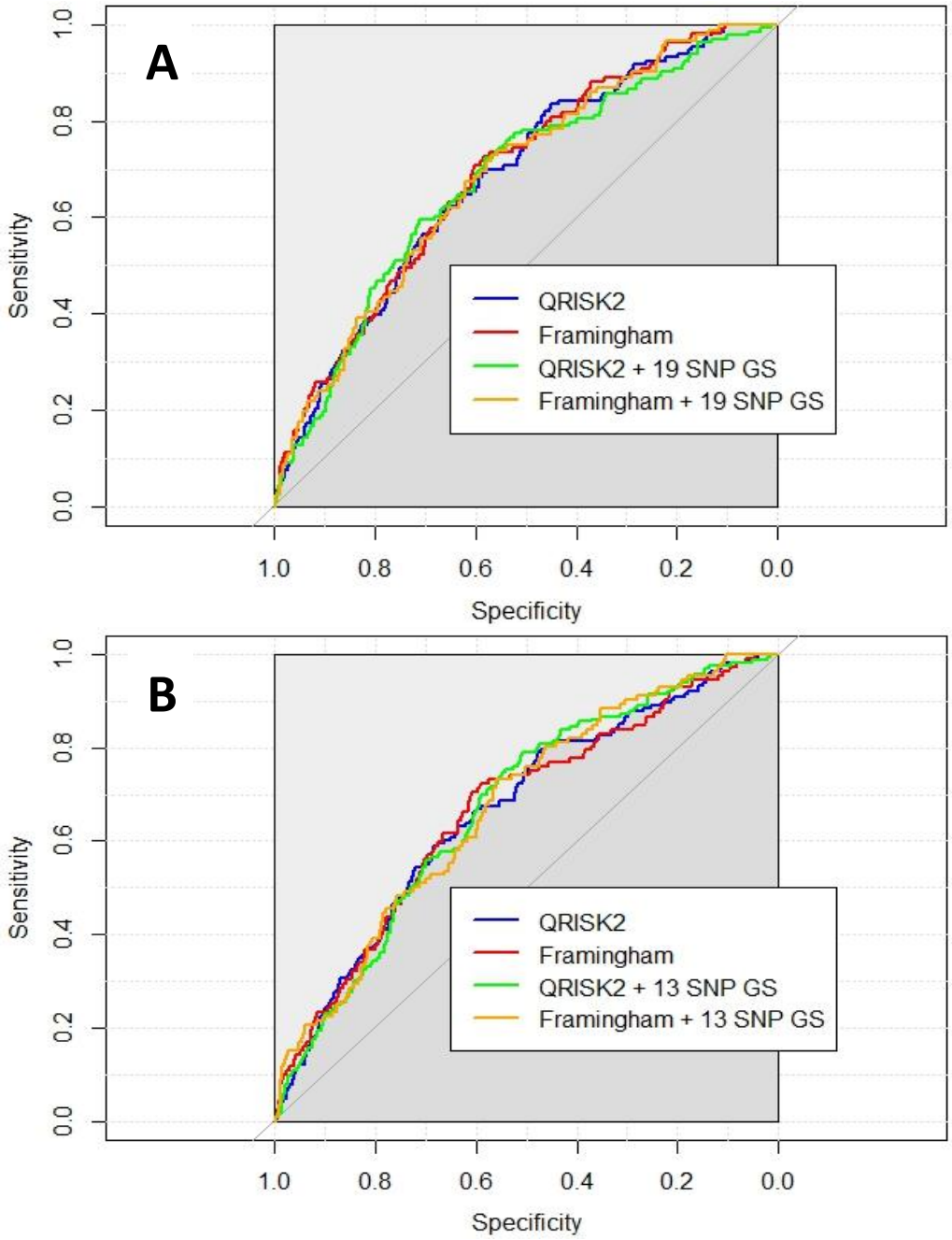
Rates were compared using the Hosmer-Lemeshow test. R packages "ggplot2"(Wickham 2009)," PredictABEL"(Kundu, Aulchenko et al. 2011; Kundu, Aulchenko et al. 2014) and "ResourceSelection" (Lele, Keim et al. 2014) were used to perform the analysis and produce the plots. However, p-values were calculated separately using ten degrees of freedom, rather than with the eight calculated with the R packages.

Figure 11: Observed CHD event rate in NPHSII compared to the predicted event rate determined by A) QRISK2 score alone; B) QRISK2 plus 19 SNP GS; C) QRISK2 plus 13 SNP GS, presented by decile of risk score



Rates were compared using the Hosmer-Lemeshow test. R packages "ggplot2"(Wickham 2009)," PredictABEL"(Kundu, Aulchenko et al. 2014) and "ResourceSelection"(Lele, Keim et al. 2014) were used to perform the analysis and produce the plots. However, p-values were calculated separately using ten degrees of freedom, rather than with the eight calculated with the R packages.

Figure 12: ROC curves for different risk scores – CRF score alone and with the addition of the one the GSs A) CRF scores and the 19 SNP GS and B) CRF score and the 13 SNP GS



Plots were created using the R package "pROC" (Robin, Turck et al. 2011).

Table 31: AUROC for combined CRF plus GS risk scores

Combined Score	AUROC (95% CI)	CRF score	AUROC (95% CI)	p-value
FRAM+19 SNP GS	0.69 (0.64-0.74)	FRAM	0.69 (0.63-0.74)	0.90
FRAM +13 SNP GS	0.67 (0.62-0.72)	FRAM	0.67 (0.61-0.72)	0.75
QRISK2+19 SNP GS	0.68 (0.63-0.73)	QRISK2	0.69 (0.64-0.73)	0.85
QRISK2+13 SNP GS	0.67 (0.63-0.71)	QRISK2	0.67 (0.62-0.71)	0.72

AUROC for different scores were compared using DeLong's test, part of the R package "pROC" (Robin, Turck et al. 2011). AUROC= area under the ROC curve. CI=confidence interval. CRF=conventional risk factor. GS=gene score. FRAM=Framingham risk score.

Table 32: Reclassification of NPHII participants with the addition of the GSs to the CRF scores

Risk Score	Reclassified at lower risk	No change in risk classification	Reclassified at higher risk	NRI (95% CIs)	p-value
FRAM + 19 SNP GS					
No CHD	80	745	105	-0.01 (-0.10-0.09)	0.92
CHD	8	74	10		
Event rate	9.10 %	9.06 %	8.78 %		
FRAM + 13 SNP GS					
No CHD	84	933	143	0.01 (-0.06-0.09)	0.76
CHD	5	95	12		
Event rate	5.62 %	9.24 %	7.69 %		
QRISK2 + 19 SNP GS					
No CHD	101	812	167	0.01 (-0.07-0.10)	0.75
CHD	11	101	21		
Event rate	9.82 %	11.06 %	11.17 %		
QRISK2 + 13 SNP GS					
No CHD	111	1041	204	0.05 (-0.02-0.12)	0.16
CHD	8	130	28		
Event rate	6.61 %	11.10 %	12.07 %		

10 % was used as the high risk cut-off. NRI=net reclassification index. FRAM=Framingham risk score. CI=confidence interval. GS=gene score. CRF=conventional risk factor. GS=gene score. CRF=conventional risk factor. CHD=coronary heart disease.

3.2.2 CHD risk GS in the South Asian and Afro-Caribbean populations

The risk variants included in the risk score were all identified in studies with individuals of European ethnicity. Thus it is unknown how applicable this work is in other ethnic groups, particularly given the differing LD patterns between ethnic groups. Therefore, the performance of the 19 SNP and 13 SNP GSs were assessed in cohorts of South Asian and Afro-Caribbean origin.

3.2.2.1 Basic characteristics of the Islamabad, Lahore and Guadeloupe cohorts

The basic characteristics of the participants from the two Pakistani case-control groups studied are presented in Table 33. In the Islamabad study, data was not collected for up to 60% of participants for all variables except age and sex. As expected, in both studies the case group was older and had a higher proportion of smokers and those with diabetes and hypertension (all $p<0.05$). There was no difference in the proportion of males between cases and controls in either study. While in the Lahore study LDL cholesterol was higher in the case group, surprisingly, in the Islamabad study, total cholesterol and LDL cholesterol did not differ between cases and controls. As all of those in the case group are post-MI, this can be attributed to treatment with lipid-lowering therapies. Data on BMI, triglycerides, family history and HDL-cholesterol were available for some of the participants of the Islamabad study. Only for HDL cholesterol was the difference between cases and controls statistically significant, being lower in the cases.

The basic characteristics of the participants of the Afro-Caribbean case-control study are presented in Table 34. The cases were older and a greater proportion were male and had hypertension, diabetes, hypercholesterolaemia and were smokers. However, there was no difference in the proportion of obese participants or in BMI between cases and controls.

Table 33: Basic characteristics of the participants in the case-control studies of South Asian individuals from Pakistan

Trait	Islamabad			Lahore		
	Controls n=228	Cases n=321	p-value	Controls n=219	Cases n=404	p-value
Age (mean)	38 (11.83)	53 (11.80)	$<2.2x10^{-16}$	56 (10.50)	59 (12.60)	$2x10^{-3}$
Sex (% Male)	66 %	69 %	0.52	54 %	59 %	0.27
BMI (kg/m^2)	24.2 (3.95)	24.3 (4.08)	0.85	-	-	-
Smoking	25 %	46 %	$1.25x10^{-3}$	11 %	30 %	$4.30x10^{-8}$
Hypertension	15 %	46 %	$4.9x10^{-7}$	16 %	62 %	$9.00x10^{-28}$
TC (mmol/l)	4.52 (1.38)	4.71 (3.77)	0.56	-	-	-
LDL-cholesterol (mmol/l)	2.66 (0.78)	2.55 (0.99)	0.33	2.19 (0.44)	2.74 (0.75)	$6.50x10^{-22}$
HDL-cholesterol (mmol/l)	1.35 (1.01)	0.96 (0.22)	$1.01x10^{-4}$	-	-	-
Diabetes	1 %	32 %	$6.6x10^{-9}$	14 %	65 %	$5.10x10^{-34}$

Where appropriate mean and standard deviation (sd) are shown. Categorical variables were compared using a χ^2 test while Welch's t-test was used to compare continuous variables. *Log transformed data. Geometric mean and approximate sd are given. BMI=body mass index. TC=total cholesterol.

Table 34: Basic characteristics of the participants of the case-control study of Afro-Caribbean individuals from Guadeloupe

Trait	Controls n=359	Cases n=178	p-value
Age	51.66 (13.54)	63.20 (10.50)	$<1x10^{-5}$
Sex (% Male)	44 %	64 %	$<1x10^{-5}$
Smoking	13 %	28 %	$<1x10^{-5}$
BMI (kg/m^2)	27.15 (5.62)	27.41 (4.86)	0.54
Obesity (BMI >30 kg/ $m^{2)}$	29 %	24 %	0.22
Hypertension	30 %	79 %	$<1x10^{-5}$
Hypercholesterolemia	15 %	53 %	$<1x10^{-5}$
Diabetes	15 %	54 %	$<1x10^{-5}$

Where appropriate mean and standard deviation (sd) are shown. Categorical variables were compared using a χ^2 test. Student t test was used for comparison between continuous variables. BMI=body mass index.

3.2.2.2 GSs in the Islamabad, Lahore and Guadeloupe cohorts

The 19 SNPs were genotyped in the South Asian and Afro-Caribbean cohorts and the results are presented in Table 35-Table 37. For the Islamabad study, five SNPs were not in HWE - *MIA3* rs17465637, *CXCL12* rs1746048, *MRAS* rs9818870, *LPL* rs1801177 and *APOE* rs7412 - with an excess of homozygotes present in each case. To confirm this, a selection of genotypes were checked by Sanger sequencing and the genotype frequencies remained out of HWE. In the Lahore study, only one SNP – *LPA* rs10455872- was not in HWE. Similarly, in the Guadeloupe study only one SNP, rs646776 (rs599839 was genotyped) close to the gene cluster containing *CELSR2-PSRC1-SORT1*, was not in HWE. The data from the Pakistani control groups was combined and compared to that from the NPHSII (Table 36). The RAF was lower in the Pakistani group for 13 SNPs and higher for three SNPs compared to NPHSII. In the Guadeloupe cohort the RAF was higher for six SNPs and lower for eight SNPs compared to the NPHSII participants (Table 37).

The 19 and 13 SNP GSs were calculated for all cohorts and the results are shown in Table 38. For the 19 SNP GS, full genotyping was available for 294 samples (119 controls/175 cases) in the Islamabad study, 443 samples (134 controls/309 cases) in the Lahore study and 537 samples (359 controls/178 cases) for the Guadeloupe study. There was no difference in the weighted GS between the cases and control in either Pakistani group (Islamabad p=0.35, Lahore p=0.41). However, weighted mean GS was higher in cases compared to controls in the Guadeloupe study (p=0.02). For the 13 SNP GS, full genotyping was available for 317 samples (123 controls/194 cases) in the Islamabad study, 488 samples (145 controls/343 cases) in the Lahore study and 537 samples (359 controls/178 cases) for the Guadeloupe study. In the Islamabad sample, mean weighted 13 SNP GS was found to be higher in case compared to controls (p=0.04) but not in the Lahore sample (p=0.41). The weighted 13 SNP GS was also found to be higher in the Guadeloupe case group compared to controls (p=0.001).

Both GSs (unweighted and weighted) were associated with CHD after adjustment for age and sex in the Guadeloupe cohort but not in either Pakistani study (Table 39, Figure 13). The GSs were not associated with any of the CRFs assessed in the three studies (Table 40, Table 41 and Table 42).

Table 35: Hardy-Weinberg equilibrium results from the Pakistani and Guadeloupe cohorts

Gene/Locus	SNP	Islamabad HWE p-value	Lahore HWE p-value	Guadeloupe HWE p-value
APOE	rs429358	0.03	0.28	0.60
APOE	rs7412	0.27	0.33	0.10
MIA3	rs17465367	0.01	0.53	0.92
MRAS	rs9818870	0.03	0.86	0.44
DAB2IP	rs7025486	0.24	0.80	0.86
CXCL12	rs1746048	$1.4x10^{-3}$	0.94	0.2
APOA5	rs662799	0.37	0.99	0.78
SORT1	rs646776	0.76	0.18	0.02
SMAD3	rs17228212	0.05	0.85	0.48
ACE	rs4341	0.13	0.16	0.63
LPL	rs328	0.65	0.64	0.48
CETP	rs708272	0.70	0.36	0.3
CDKN2A/9p21	rs10757274	0.09	0.40	0.93
NOS3	rs1799983	0.06	0.11	0.63
LPL	rs1801177	$<1x10^{-4}$	-	0.98
PCSK9	rs11591147	-	0.95	-
LPA	rs10455872	0.82	0.01	0.90
APOB	rs1042031	0.14	0.85	0.81
LPA	rs3789220	0.87	0.93	0.74

HWE=Hardy-Weinberg equilibrium.

Table 36: Risk allele frequency in control groups from the Islamabad and Lahore cohorts

Gene/ Locus	SNP	RAF Islamabad Controls (95% CI)	RAF Lahore Controls (95% CI)	p-value	RAF NPSHII (95% CI)	p-value
APOE	rs429358	0.09 (0.06-0.11)	0.11 (0.08-0.14)	0.52	0.17 (0.16-0.18)	$1.60x10^{-9}$
APOE	rs7412	0.96 (0.94-0.98)	0.96 (0.94-0.98)	1	0.91 (0.90-0.92)	$2.69x10^{-6}$
MIA3	rs17465367	0.64 (0.59-0.69)	0.63 (0.58-0.67)	0.68	0.71 (0.69-0.72)	$3.42x10^{-5}$
MRAS	rs9818870	0.10 (0.07-0.13)	0.09 (0.06-0.12)	0.75	0.16 (0.15-0.17)	$2.61x10^{-6}$
DAB2IP	rs7025486	0.31 (0.26-0.35)	0.32 (0.27-0.36)	0.84	0.26 (0.17-0.24)	$4.70x10^{-4}$
CXCL12	rs1746048	0.65 (0.61-0.70)	0.64 (0.59-0.68)	0.65	0.86 (0.85-0.87)	$<2.20x10^{-16}$
APOA5	rs662799	0.15 (0.12-0.18)	0.17 (0.13-0.20)	0.60	0.06 (0.05-0.07)	$<2.20x10^{-16}$
SORT1	rs599839	0.72 (0.68-0.77)	0.74 (0.70-0.79)	0.55	0.78 (0.77-0.79)	$2.03x10^{-3}$
SMAD3	rs17228212	0.19 (0.15-0.22)	0.18 (0.14-0.21)	0.76	0.31 (0.30-0.32)	$4.88x10^{-14}$
ACE	rs4341	0.41 (0.36-0.45)	0.47 (0.43-0.52)	0.05	0.52 (0.50-0.53)	$3.09x10^{-5}$
LPL	rs328	0.92 (0.89-0.94)	0.91 (0.89-0.94)	0.98	0.90 (0.89-0.91)	0.12
CETP	rs708272	0.55 (0.50-0.60)	0.56 (0.51-0.61)	0.83	0.56 (0.55-0.58)	0.62
CDKN2A/ 9p21	rs10757274	0.45 (0.40-0.49)	0.54 (0.49-0.58)	0.01	0.48 (0.47-0.50)	0.68
NOS3	rs1799983	0.16 (0.13-0.20)	0.18 (0.15-0.22)	0.51	0.33 (0.32-0.35)	$<2.20x10^{-16}$
LPL	rs1801177	0.01 (0-0.02)	0 -	NA	0.01 (0.01-0.02)	NA
PCSK9	rs11591147	1.00 -	0.995 (0.00-0.01)	NA	0.99 (0.99-0.99)	NA
LPA	rs10455872	0.01 (0-0.03)	0.01 (0.00-0.03)	1	0.07 (0.07-0.08)	$1.06x10^{-10}$
APOB	rs1042031	0.15 (0.12-0.19)	0.13 (0.09-0.16)	0.24	0.18 (0.17-0.19)	$9.75x10^{-3}$
LPA	rs3789220	0.01 (0-0.02)	0.003 (0-0.01)	0.81	0.02 (0.01-0.02)	0.02

The RAF from the two Pakistani studies were compared to each other, then combined and to NPHSII using tests of proportion. CI=confidence interval. RAF=risk allele frequency.

Table 37: Risk allele frequency in the control group of the Guadeloupe cohort

Gene/Locus	SNP	RAF Guadeloupe Controls (95% CI)	RAF NPSHII (95% CI)	p-value
APOE	rs429358	0.23 (0.19-0.28)	0.17 (0.16-0.18)	0.01
APOE	rs7412	0.93 (0.90-0.96)	0.91 (0.90-0.92)	0.23
MIA3	rs17465367	0.24 (0.80-0.28)	0.71 (0.69-0.72)	$2x10^{-57}$
MRAS	rs9818870	0.08 (0.05-0.11)	0.16 (0.15-0.17)	$1x10^{-4}$
DAB2IP	rs7025486	0.32 (0.27-0.37)	0.26 (0.24-0.28)	0.03
CXCL12	rs1746048	0.53 (0.48-0.58)	0.86 (0.84-0.88)	$1x10^{-40}$
APOA5	rs662799	0.13 (0.10-0.17)	0.06 (0.05-0.07)	$1x10^{-5}$
SORT1	rs599839	0.25 (0.20-0.30)	0.78 (0.76-0.80)	$3x10^{-77}$
SMAD3	rs17228212	0.12 (0.09-0.15)	0.31 (0.30-0.32)	$9x10^{-13}$
ACE	rs4341	0.60 (0.55-0.65)	0.52 (0.50-0.53)	$8x10^{-3}$
LPL	rs328	0.94 (0.91-0.96)	0.90 (0.89-0.91)	0.02
CETP	rs708272	0.76 (0.72-0.80)	0.56 (0.55-0.59)	$11x10^{-11}$
CDKN2A/ 9p21	rs10757274	0.21 (0.17-0.25)	0.48 (0.47-0.50)	$8.19x10^{-20}$
NOS3	rs1799983	0.12 (0.09-0.15)	0.33 (0.32-0.35)	$8x10^{-15}$
LPL	rs1801177	0.001 (0-0.003)	0.01 (0.01-0.02)	NA
PCSK9	rs11591147	1.00 -	0.99 (0.99-1.00)	NA
LPA	rs10455872	0.01 (0-0.03)	0.07 (0.07-0.08)	$1x10^{-5}$
APOB	rs1042031	0.15 (0.12-0.19)	0.18 (0.17-0.19)	0.19
LPA	rs3789220	0.01 (0-0.03)	0.02 (0.01-0.02)	0.21

The RAF from the Guadeloupe study was compared to that of NPHSII participants using tests of proportion. CI=confidence interval. RAF=risk allele frequency.

Table 38: GS values in the Pakistani and Afro-Caribbean cohorts

Study		19 SNP GS			13 SNP GS		
	Score	Controls	Cases	p-value	Controls	Cases	p-value
Islamabad	Unweighted	14.97 (2.12)	15.37 (2.22)	0.11	11.84 (1.63)	12.40 (1.82)	0.004
	Weighted	2.89 (0.50)	2.94 (0.50)	0.35	2.24 (0.42)	2.34 (0.42)	0.04
Lahore	Unweighted	15.28 (2.23)	15.24 (2.20)	0.86	11.85 (1.80)	11.72 (1.72)	0.91
	Weighted	2.95 (0.50)	2.91 (0.46)	0.41	2.25 (0.42)	2.22 (0.36)	0.41
Guadeloupe	Unweighted	13.17 (2.07)	13.90 (2.10)	1.60×10^{-4}	9.35 (1.73)	10.12 (1.68)	<0.001
	Weighted	2.57 (0.47)	2.67 (0.49)	0.02	1.80 (0.42)	1.94 (0.44)	0.001

Mean GS and standard deviation are shown. Welch's t-test was used to compare the GSs between cases and controls. GS=gene score.

Table 39: Association between GS and CHD outcome

Study		19 SNP GS		13 SNP GS	
	Score	OR (95% CI)	p-value	OR (95% CI)	p-value
Islamabad	Unweighted	0.88 (0.64-1.19)	0.41	1.22 (0.90-1.65)	0.20
	Weighted	0.89 (0.65-1.22)	0.48	1.21 (0.90-1.65)	0.21
Lahore	Unweighted	0.98 (0.80-1.20)	0.85	1.00 (0.82-1.22)	0.98
	Weighted	0.92 (0.75-1.13)	0.41	0.91 (0.75-1.11)	0.36
Guadeloupe	Unweighted	1.44 (1.17-1.78)	4.79×10^{-4}	1.57 (1.28-1.95)	2.85×10^{-5}
	Weighted	1.35 (1.10-1.66)	4.93×10^{-3}	1.42 (1.15-1.75)	1.01×10^{-3}

Effect sizes relate to one standard deviation of the variable and were determined by logistic regression, adjusted for age and sex. OR= odds ratio. CI= confidence interval.

Table 40: Association between CHD risk factors and the 19 SNP GS in the Islamabad cohort

Trait	19 SNP unweighted GS p-value	19 SNP weighted GS p-value
Smoking	0.34	0.21
Hypertension	0.47	0.48
TC	0.82	0.89
LDL-cholesterol	0.43	0.68
HDL-cholesterol	0.07	0.28
Diabetes	0.23	0.18

The p-values were determined by linear or logistic regression as appropriate. TC = total cholesterol. GS=gene score.

Table 41: Association between CHD risk factors and the 19 SNP GS in the Lahore cohort

Trait	19 SNP unweighted GS p-value	19 SNP weighted GS p-value
Smoking	0.32	0.43
Hypertension	0.58	0.35
TC	0.59	0.86
LDL-cholesterol	0.15	0.63
HDL-cholesterol	0.09	0.54
Diabetes	0.41	0.83

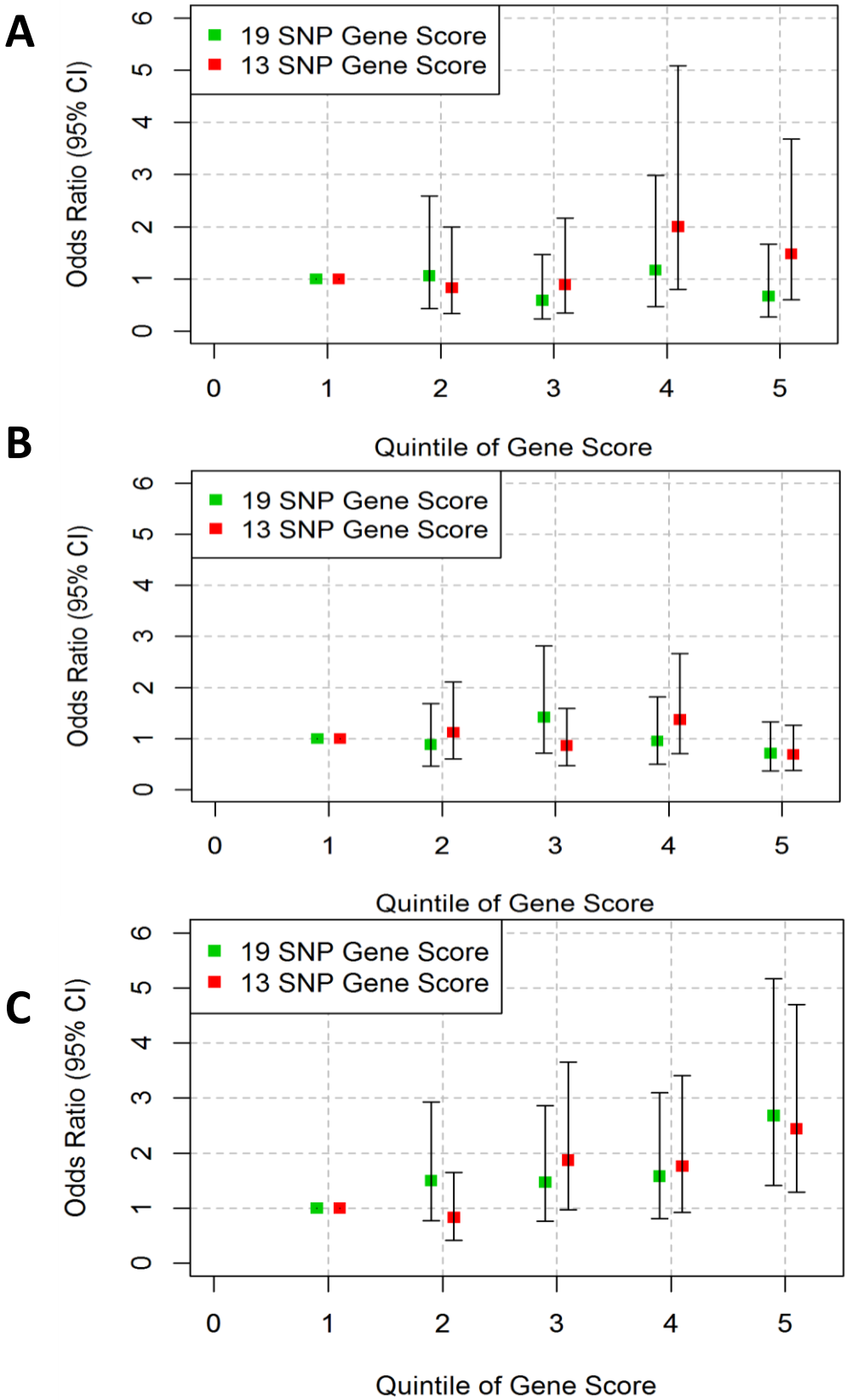
The p-values were determined by linear or logistic regression as appropriate. TC =total cholesterol. GS=gene score.

Table 42: Association between CHD risk factors and the 19 SNP GS in the Guadeloupe cohort

Trait	19 SNP unweighted GS p-value	19 SNP weighted GS p-value
Smoking	0.92	0.83
Hypertension	0.32	0.21
Hypercholesterolemia	0.42	0.80
Diabetes	0.66	0.94

The p-values were determined using t-tests. GS=gene score.

Figure 13: Association between quintile of weighted GS and outcome in the A) Islamabad study B) Lahore study and C) Guadeloupe study



Logistic regression was performed, adjusted for age and sex. Error bars represent 95% CI. CI=confidence intervals. GS=gene score.

3.2.3 Updating the Gene Score

3.2.3.1 Literature search for variants associated with CHD

In order to ascertain the current knowledge regarding the genetics of CHD and to identify candidate SNPs for future CHD GSs, a systematic literature search (up to February 2013) was carried out. Nine search terms relating to the genetics of CHD were used (see Chapter 2.2). An overview of the search strategy and the number of papers identified is shown in Figure 14. Four phenotypes were considered: CHD, premature CHD, CHD in those with T2D and secondary CHD events. Studies which did not meet the inclusion criteria were excluded and only the most recent meta-analysis for a particular variant was retained. For the CHD phenotype, 32 meta-analyses were identified. This included the meta-analysis conducted by the CARDIoGRAMplusC4D consortium (Deloukas, Kanoni et al. 2013). Over 60,000 cases and 130,000 controls were included in the study, by far the largest analysis of CHD genetics published when the search was performed. More than 50 SNPs from 46 loci were robustly associated with CHD and thus these are best candidates to be used in determining genetic risk of CHD (Table 43).

From the remaining 31 meta-analyses, there were 34 variants which had been associated with CHD but were neither among nor in LD with those identified in the CARDIoGRAMplusC4D meta-analysis (Table 44). The data from the CARDIoGRAMplusC4D consortium has been made publically available and was searched to determine if these 34 SNPs had been included in the analysis. If a SNP had been included but had not met the significance threshold, it can be concluded it is not a good candidate to be included in a future CHD risk GS. This cannot be established for SNPs that were not genotyped or imputed in that study. Data on coronary artery disease / myocardial infarction was contributed by CARDIoGRAMplusC4D investigators and was downloaded from www.CARDIOGRAMPLUSC4D.ORG. Of the 34 SNPs, all 29 had been genotyped or imputed in at least one stage of the CARDIoGRAMplusC4D analysis but had not met the significance threshold for association with CHD. Eleven of these SNPs showed a suggestive association with CHD ($p<0.05$) indicating that these loci may be influencing CHD risk but this has not been confirmed. Therefore, these variants can be discounted in favour of those that have been robustly associated with CHD.

Of the remaining five variants the *LPA* SNP, rs10455872, met the criteria for replication in the first stage of the analysis (the CARDIoGRAM GWAS (Schunkert, Konig et al. 2011)) but it was not found in the CARDIoGRAMplusC4D data. It is assumed this is because it is located in the gene *LPA*, where another CHD risk SNP rs3798220, is also located and thus was not included in follow-up. For the remaining four SNPs that were not genotyped or imputed in the CARDIoGRAMplusC4D analysis, the most likely reason is that they are not included on the genotyping arrays used. For one of these SNPs, rs41360247 (*ABCG8*) a proxy SNP, rs4953023, (r^2=1, as determined in the 1000 Genomes phase 1 EUR data) had been genotyped/imputed in stage 1 of the analysis but had not met the threshold for replication (Stage 1 $p=2.15x10^{-6}$, validation at $p<1x10^{-6}$). Of the remaining three SNPs, one is located in the promoter *APOB* (Chiodini, Barlera et al. 2003), which is a CHD risk loci, one is a SNP tagging an insertion/deletion polymorphism in the NFκB subunit encoding gene *NFKB1* (Vogel, Jensen et al. 2011) and the third a missense variant in *TGF-B* (Morris, Moxon et al. 2012).

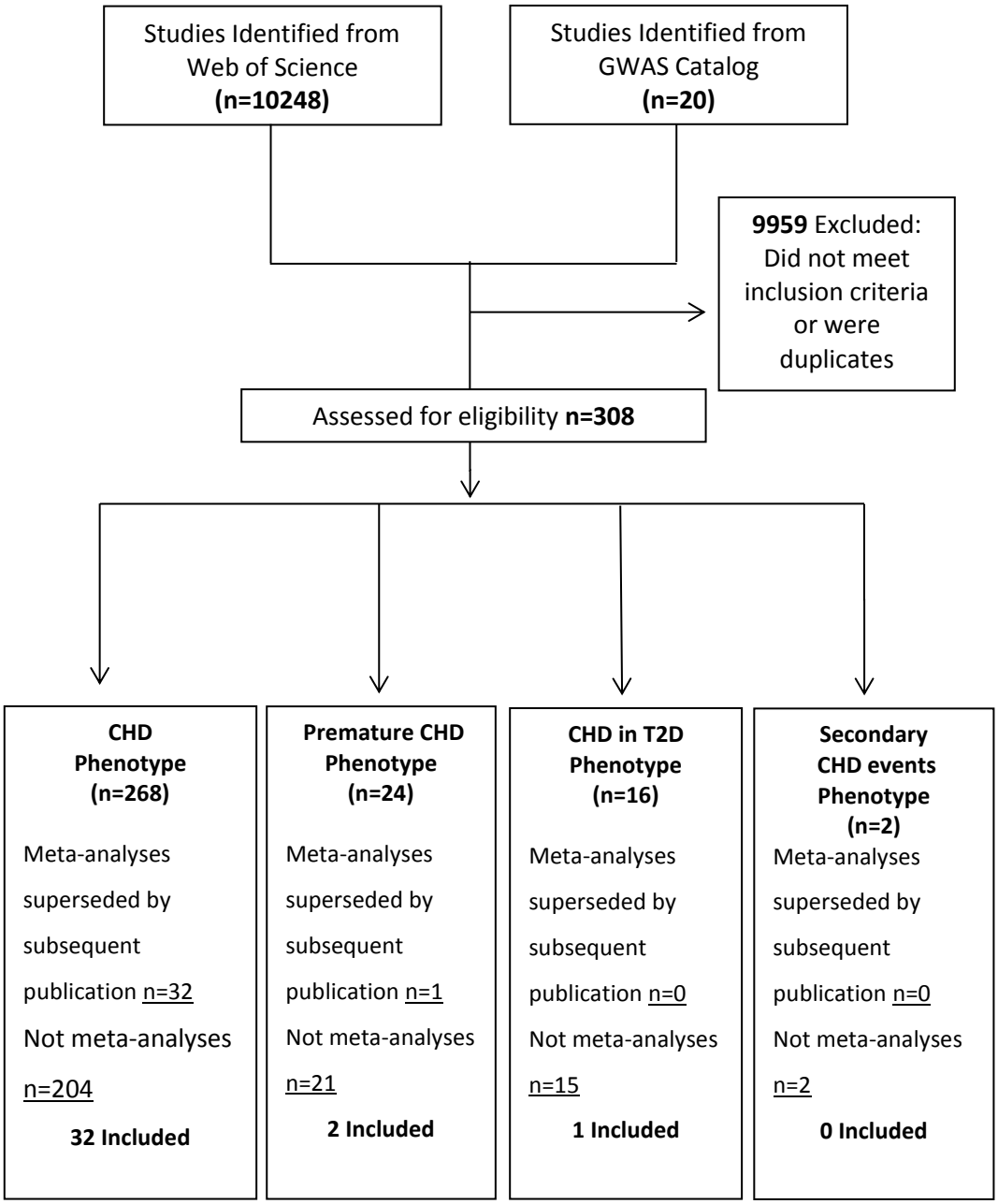
Two meta-analyses investigating variants associated with premature CHD phenotypes were identified (Kathiresan, Altschuler et al. 2009; Xuan, Bai et al. 2011). All of the loci identified by Kathiresan, Altschuler et al. were also identified in the CARDIoGRAMplusC4D meta-analysis. The variant rs1801133 in *MTHFR* investigated by Xuan, Bai et al., was not found to be associated with CHD in the CARDIoGRAMplusC4D meta-analysis (p=0.36). No meta-analyses concerning secondary CHD events were identified. One meta-analysis considering variants associated with CHD in T2D was identified and is discussed in Chapter 4.2.2.

Table 43: Summary of CHD risk loci identified in the CARDIoGRAMplusC4D meta-analysis

Chromosome	Lead SNP	Gene/Locus	OR	RAF
1	rs602633	*SORT1*	1.12	0.77
1	rs11206510	*PCSK9*	1.06	0.84
1	rs4846525	*IL6R*	1.09	0.47
1	rs17114036	*PPAP2B*	1.11	0.91
1	rs17464857	*MIA3*	1.05	0.87
2	rs6725887	*WDR12*	1.12	0.11
2	rs515135	*APOB*	1.03	0.83
2	rs2252641	*ZEB2-AC74093.1*	1.00	0.46
2	rs1561198	*VAMP5-VAMP8-GGCX*	1.07	0.45
3	rs99818870	*MRAS*	1.07	0.14
4	rs7692387	*GUCY1A3*	1.13	0.81
4	rs1878406	*EDNRA*	1.09	0.15
5	rs273909	*SLC22A4-SLC22A5*	1.11	0.14
6	rs3798220	*LPA*	1.28	0.01
6	rs2048327	*LPA*	1.06	0.35
6	rs10947789	*KCNK5*	1.01	0.76
6	rs4252120	*PLG*	1.07	0.73
6	rs12205331	*ANKS1A*	1.04	0.81
6	rs9369640	*PHACTR1*	1.09	0.65
7	rs11556924	*ZC3HC1*	1.09	0.65
7	rs2023938	*HDAC9*	1.13	0.10
7	rs12539895	7q22	1.08	0.19
8	rs264	*LPL*	1.06	0.86
8	rs2954029	*TRIB1*	1.05	0.55
9	rs1333049	9p21	1.23	0.47
9	rs3217992	9p21	1.16	0.38
9	rs579459	*ABO*	1.07	0.21
10	rs12413409	*CYP17A1-CNNM2-NT5C2*	1.10	0.89
10	rs2505083	*KIAA1462*	1.06	0.42
10	rs501120	*CXCL12*	1.07	0.83
10	rs2047009	*CXCL12*	1.05	0.48
10	rs2246833	*LIPA*	1.06	0.38
10	rs11203042	*LIPA*	1.04	0.44
11	rs974819	*PDGRD*	1.07	0.29
11	rs9326246	*ZNF259-APOA5-APOA1*	1.09	0.10
12	rs3184504	*SH2B3*	1.07	0.40
13	rs4773144	*COL4A1-COL4A2*	1.07	0.74
13	rs9515203	*COL4A1-COL4A2*	1.08	0.74
13	rs9319428	*FLT1*	1.10	0.32
14	rs2895811	*HHIPL1*	1.06	0.43
15	rs7173743	*ADAMTS7*	1.07	0.58
15	rs17514846	*FURIN-FES*	1.04	0.44
17	rs12936587	*RAI1-PEMT-RASD1*	1.06	0.59
17	rs15563	*UBE2Z*	1.04	0.52
17	rs2281727	*SMG6*	1.05	0.36
19	rs1122608	*LDLR*	1.10	0.76
19	rs2075650	*ApoE-ApoC1*	1.11	0.14
19	rs445925	*ApoE-ApoC1*	1.13	0.90
21	rs9982601	21q22	1.13	0.13

OR=odds ratio. RAF=risk allele frequency. CHD=coronary heart disease.

Figure 14: Literature search protocol



Two publications covered two of the phenotypes **(**Xuan, Bai et al. 2011**)** for CHD and premature CHD and (Buysschaert, Carruthers et al. 2010) for CHD and secondary CHD events. Therefore, when the total number of studies for each phenotype is added together, the total is 310 rather than 308. T2D=type 2 diabetes. GWAS=genome-wide association study. CHD=coronary heart disease.

Table 44: SNPs identified in the literature search that were not confirmed in the CARDIoGRAMplusC4D meta-analysis

SNP	Gene/ Locus	Study	Literature Search OR	CARDIoGRAM Source	CARDIoGRAM p-value
rs10455872	*LPA*	(Li, Luke et al. 2011)	1.42	CG GWAS	3.08×10^{-13}
rs10846744	*SCARB1*	(Grallert, Dupuis et al. 2012)	1.01	Cplus4DGWAS	3.61×10^{-6}
rs2943634	2p36.3	(Angelakopoulou, Shah et al. 2012)	1.08	Cplus4DGWAS	3.29×10^{-5}
rs4420638	*APOE-C1-C4-C2*	(Grallert, Dupuis et al. 2012)	1.11	CG GWAS	2.14×10^{-4}
rs6922269	*MTHFD1L*	(Angelakopoulou, Shah et al. 2012)	1.10	Cplus4DGWAS	3.58×10^{-4}
rs1801177	*LPL*	(Sagoo, Tatt et al. 2008)	1.33	Cplus4DGWAS	4.04×10^{-4}
rs3869109	6p21.3	(Davies, Wells et al. 2012)	1.10	CG GWAS	1.49×10^{-3}
rs7025486	*DAB2IP*	(Harrison, Cooper et al. 2012)	1.10	Cplus4DGWAS	2.14×10^{-3}
rs2706399	*IL-5*	(Butterworth, Braund et al. 2011)	1.05	CG GWAS	0.01
rs699	*AGT*	(Zafarmand, van der Schouw et al. 2008)	1.08	CG GWAS	0.01
rs266729	*ADIPOQ*	(Zhou, Xi et al. 2012)	1.12	CG GWAS	0.02
rs383830	*APC*	(Angelakopoulou, Shah et al. 2012)	1.10	CG GWAS	0.10
rs2234693	*ESR1*	(Shearman, Cooper et al. 2006)	1.44	CG GWAS	0.15
rs12042319	*ANGPTL3*	(Angelakopoulou, Shah et al. 2012)	1.11	CG GWAS	0.16
rs20455	*KIF6*	(Peng, Lian et al. 2012)	1.27	CG GWAS	0.16
rs1800469	*TGF-B*	(Morris, Moxon et al. 2012)	1.13	CG GWAS	0.22
rs1800896	*IL-10*	(Wang, Zheng et al. 2012)	1.12	CG GWAS	0.24
rs1801133	*MTHFR*	(Xuan, Bai et al. 2011)	1.14	CG GWAS	0.36
rs4343 (I/D)	*ACE*	(Zintzaras, Raman et al. 2008)	1.21	CG GWAS	0.44
rs662	*PON1*	(Wheeler, Keavney et al. 2004)		CG GWAS	0.60
rs5985	*FXIII*	(Voko, Bereczky et al. 2007)	1.23	CG GWAS	0.62
rs1800629	*TNFA*	(Zhang, Xie et al. 2011)	1.50	CG GWAS	0.64
rs5186	*AGT1R*	(Xu, Sham et al. 2010)	1.09	C4D GWAS	0.71
rs1801282	*PPARG*	(Wu, Lou et al. 2012)	1.45	Cplus4DGWAS	0.73
rs2995300	*ADAM8*	(Raitoharju, Seppala et al. 2011)	1.39	CG GWAS	0.79
rs1799983	*NOS3*	(Casas, Cavalleri et al. 2006)	1.13	CG GWAS	0.90
rs708272	*CETP*	(Thompson, Di Angelantonio et al. 2008)		C4D GWAS	0.91
rs17228212	*SMAD3*	(Angelakopoulou, Shah et al. 2012)	1.11	Cplus4DGWAS	0.95
rs1024611	*MCP-1*	(Wang, Zhang et al. 2011)	1.42	CG GWAS	0.99
rs11279109	*APOB*	(Chiodini, Barlera et al. 2003)	1.15	-	-
rs41360247	*ABCG8*	(Teupser, Baber et al. 2010)	1.20	-	-
rs1800471	*TGF-B*	Morris, Moxon et al. 2012	1.21	-	-
rs28362491	*NFKB1*	(Vogel, Jensen et al. 2011)	1.22	-	-
rs116843064	*ANGPTL4*	(Talmud, Smart et al. 2008)	1.48	-	-

Where SNPs have been genotyped or imputed in the CARDIoGRAM GWAS data set, the p-value for association with CHD is included. Data on coronary artery disease / myocardial infarction was contributed by CARDIoGRAMplusC4D investigators and was downloaded from www.CARDIOGRAMPLUSC4D.ORG.Cplus4D = CARDIoGRAMplusC4D meta-analysis. CG GWAS= CARDIoGRAM GWAS. OR=odds ratio.

3.2.3.2 GS SNPs in the CARDIoGRAMplusC4D analysis

The 19 SNPs included in the GS presented in section 3.1, were then considered in the context of the literature search findings. The results of the CARDIoGRAMplusC4D meta-analysis were checked to determine how many of the 19 SNPs were among the robustly associated CHD risk loci. The full data sets were then checked to assess whether the remaining SNPs showed a suggestive association with CHD (at least under the additive model used in that analysis).

3.2.3.2.1 GS SNPs among the 46 confirmed CHD loci

Sixteen of the 19 GS SNPs were genotyped as part of the CARDIoGRAM GWAS or CARDIoGRAMplusC4D meta-analysis. Four of the SNPs were lead SNPs at one of the 46 CHD risk loci identified (*SORT1*, *CXCL12*, *MRAS* and *LPA* rs3798220). The 9p21 SNP used in the 19 SNP GS (rs10757274) and the lead SNP identified in the CARDIoGRAMplusC4D analysis are in strong LD (r^2=0.88 as determined in the 1000 Genomes phase 1 EUR data). Furthermore, rs10757274 is in almost complete LD with the lead 9p21 SNP from the CARDIoGRAM GWAS (r^2=0.99 in the 1000 Genomes phase 1 EUR data). Therefore, it can be concluded that the SNP in the GS covers the locus identified in the CARDIoGRAMplusC4D analysis.

Two of the 19 SNP GS SNPs were moderate LD with the CARDIoGRAMplusC4D lead SNPs and were also genotyped/imputed in the analysis. The *APOA5* promoter SNP, rs662799, is in moderate LD with the lead SNP at the *ZNF259-APOA5-APOA1* locus (r^2=0.73 with rs9326246 as calculated from the 1000 Genomes phase 1 EUR data) while the nonsense *LPL* variant rs328 is in moderate LD with the lead *LPL* SNP identified (rs264, r^2=0.30 with rs328 in the 1000 Genomes phase 1 EUR data). Three of the 19 SNP GS SNPs - rs1801177 in *LPL* (not in LD with rs264), rs7025486 in *DAB2IP* and rs708272 in *CETP* - showed a suggestive association with CHD ($p < 0.05$). Two more of the 19 SNP GS SNPs (rs10455872 in *LPA* and rs17465637 in *MIA3*) were found to be associated with CHD in the CARDIoGRAM GWAS but were not assessed in the CARDIoGRAMplusC4D meta-analysis. The *MIA3* SNP, rs17465637, was not tested for in the CARDIoGRAMplusC4D meta-analysis as neither it nor a good proxy is included on the Metabochip array used (Voight, Kang et al. 2012). As discussed above (section 3.2.3.1), *LPA* rs10455872 was likely not included in the replication step performed in the CARDIoGRAMplusC4D meta-analysis due to its proximity to rs3798220 (also in *LPA*) despite meeting the significance threshold for replication.

Four of the 19 SNP GS SNPs were not found to be associated with CHD in the CARDIoGRAMplusC4D meta-analysis. The *APOB* SNP, rs1042031, is in weak LD with the CARDIoGRAMplusC4D *APOB* SNP rs515135 (r^2=0.28, LD data from the 1000 Genomes pilot CEU data). However, while the minor allele of rs1042031 is the CHD "risk" allele, the common allele is the CHD risk allele for rs515135. A systematic review (Boekholdt, Peters et al. 2003) found that rare allele was associated with lower LDL-cholesterol levels and there was no association between the SNP and CHD. This suggests that the data used in the original meta-analysis (Chiodini, Barlera et al. 2003) may have been subject to bias. The lack of an association between this SNP and CHD is confirmed in the CARDIoGRAMplusC4D meta-analysis (Table 45). No association was found between three other SNPs (*NOS3* rs1799983, *ACE* rs4341, and *SMAD3* rs17228212, all p>0.05) and CHD.

Three of the SNPs in the 19 SNP GS were not genotyped by the CARDIoGRAMplusC4D consortium. However, the *APOE* SNPs rs7412 and rs429358 are in weak LD with the lead SNP identified at the *APOE* locus (rs445925 r^2=0.68 with rs7412; rs2075650 r^2=0.48 with rs429358, LD data taken from 1000 Genomes phase 1 EUR data). For the missense variant rs11591147 in *PCSK9*, while one of the CARDIoGRAMplusC4D lead SNPs (rs11206510) is located close to *PCSK9*, the two SNPs are not in strong LD. However, a study of *PCSK9* variants in an Italian cohort found r^2 = 0.02 and D'=0.66 between the two SNPs and the authors concluded that the SNPs are not independent of each other (Guella, Asselta et al. 2010). Therefore, for these three SNPs it can be concluded that the SNPs included in the 19 SNP GS may at least partially tag the risk loci identified in the CARDIoGRAMplusC4D meta-analysis.

3.2.3.2.2 Updating the GS weightings

The effect sizes used to weight the SNPs in the GS are taken mostly from early GWASs or meta-analysis of candidate gene studies. Given that these are subject to inflation through sources of bias such as the "winner's curse" (Ioannidis 2008) it was decided to update the score using the effect sizes determined in the CARDIoGRAMplusC4D analysis (Deloukas, Kanoni et al. 2013). The updated weightings are shown in Table 45. Where a SNP was not present in the CARDIoGRAMplusC4D data, the effect size from the CARDIoGRAM GWAS (Schunkert, Konig et al. 2011) was used. As discussed above, neither the *APOE* SNPs nor rs11591147 in *PCSK9* were genotyped in CARDIoGRAMplusC4D. Therefore, the most recent meta-analysis identified in the literature search for these SNPs were used for a weighting, which were also the sources used for the weighting in the original score (Table

45). When combining the GS data with the CRF scores, the frequency from the CARDIoGRAM GWAS/CARDIOGRAMplusC4D data (or source publication) was used. All SNPs were treated additively.

Table 45: SNP weightings for updated GS

Gene/Locus	SNP	Risk Allele	OR	OR in original score	Frequency	p-value*	Source
APOE	rs7412	C	1.25	0.80**	0.87	-	- (Bennet, Di Angelantonio et al. 2007)
APOE	rs429358	C	1.06	1.06	0.26	-	(Bennet, Di Angelantonio et al. 2007)
MIA3	rs17465637	C	1.14	1.14	0.74	$1.36x10^{-8}$	CG GWAS
MRAS	rs9818870	T	1.07	1.15	0.14	$2.62x10^{-9}$	Cplus4D
DAB2IP	rs7025486	A	1.04	1.16	0.29	$2.14x10^{-3}$	Cplus4D
CXCL12	rs1746048[b]	C	1.07	1.17	0.83	$1.79x10^{-8}$	Cplus4D
APOA5	rs662799	G	1.05	1.19	0.06	0.01	Cplus4D
SORT1	rs599839	A	1.11	1.19	0.77	$3.8x10^{-15}$	Cplus4D
SMAD3	rs17228212	C	1.01	1.21	0.31	0.94	Cplus4D
ACE	rs4341[c]	G	1.01	1.22	0.52	0.43	CG GWAS
LPL	rs328	C	1.09	1.25	0.91	$2.34x10^{-4}$	CG GWAS
CETP	rs708272[d]	C	1.04	1.28	0.56	0.04	CG GWAS
CDKN2A/9p21	rs10757274[a]	G	1.23	1.29	0.47	$1.39x10^{-52}$	Cplus4D
NOS3	rs1799983	G	1.00	1.31	0.67	0.90	CG GWAS
LPL	rs1801177[e]	A	1.10	1.33	0.06	$4.04x10^{-4}$	Cplus4D
PCSK9	rs11591147	G	1.39	1.43	0.99	-	(Benn, Nordestgaard et al. 2010)
LPA	rs10455872	G	1.32	1.70	0.06	$3.80x10^{-13}$	CG GWAS
APOB	rs1042031	A	1.01	1.73	0.18	0.80	Cplus4D
LPA	rs3798220	C	1.28	1.92	0.01	$4.90x10^{-5}$	Cplus4D

[a]Weighting for rs1333049 (r^2=0.88). [b]Weighting for rs501120 used (r^2=0.97). [c]Weighting for rs4343 (r^2=0.96). [d]Weighting for rs711752 (r^2=1). [e]Weighting for rs7016529 (r^2=1). All r^2 values calculated from 1000 Genomes phase 1 EUR data. Data on coronary artery disease / myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.CARDIOGRAMPLUSC4D.ORG. **In the original score, the protective allele was included rather than the risk allele. OR=odds ratio. Cplus4D = CARDIoGRAMplusC4D meta-analysis. CG GWAS=CARDIoGRAM GWAS. GS=gene score.

3.2.3.2.3 Assessing the updated GS in the UK population

The performance of the updated 19 SNP GS was investigated in NPHSII. A GS removing the SNPs with weak evidence of an association with CHD in the CARDIoGRAMplusC4D was also constructed and its performance compared to that of the 19 SNP GS. Removing SNPs that are not associated with CHD, the performance of the GS should improve (or least show a similar performance) and this will reduce the number of SNPs it is necessary to genotype. Therefore, a 14 SNP score discounting SNPs with a p>0.01 from the CARDIoGRAM GWAS/CARDIoGRAMplusC4D analysis was also tested. Number of NPHSII participants were as shown in Table 30 for the updated 19 GS and 1214 participants (1101 who did not develop CHD/113 who did develop CHD) for the Framingham score plus the 14 SNP GS and 1440 participants (1277 who did not develop CVD/163 who did develop CVD) for QRISK2 plus the 14 SNP GS. For both scores (weighted and un-weighted) the mean GS in the CHD group was higher than in the non-CHD group (Table 46). Both scores were also associated with CHD after adjustment for age and sex, with the weighted scores showing a stronger association (Table 47). As in section 3.2.1.3, the GSs were combined with the Framingham and QRISK2 CRF risk scores (the population GSs were calculated using the effect sizes and allele frequencies shown in Table 45). Ten-year follow-up data was used with the endpoint of CHD for the Framingham score and CVD for the QRISK2 score. Addition of the updated GSs improved calibration in comparison to the Framingham score alone but overall it remained poor (19 SNP GS, $p=4.22x10^{-4}$; 14 SNP GS $p=1.74x10^{-3}$, Figure 15). Whereas, calibration remained good with the addition of both GSs to QRISK2 (19 SNP GS, p=0.17; 14 SNP GS p=0.20, Figure 16).

An improvement in discrimination was observed when the 19 SNP GS was combined with QRISK2 and compared to QRISK2 alone (AUROC 0.68 v 0.70 p=0.02). Addition of the updated 14 SNP GS to the CRF scores also showed improved discrimination compared to both QRISK2 alone (AUROC 0.66 v 0.69 $p=4.69x10^{-4}$) and the Framingham score alone (AUROC 0.67 v 0.69 $p=7.52x10^{-3}$) (Figure 16). There was no difference in AUROC observed with the addition of the 19 SNP GS to the Framingham score compared to the Framingham score alone (p=0.78). Combining the updated GSs with the QRISK2 risk score resulted in improved risk classification in the group who developed CHD giving a positive NRI (19 SNP NRI=0.07, p=0.04; 14 SNP NRI=0.06, p=0.03, Table 49). This was also the case for the addition of the GSs to the Framingham risk score compared to the Framingham score alone

(19 SNP NRI=0.06, p0.03; 14 SNP NRI=0.06, p=0.02, Table 49), but as the combined Framingham plus GS models are poorly calibrated, this may be misleading.

Table 46: Updated GSs in NPHSII

Score	No CHD	CHD	p-value
Updated 19 SNP GS -Unweighted	16.61 (2.18) n=1090	17.29 (2.09) n=110	$1.45x10^{-3}$
Updated 19 SNP GS- Weighted	2.08 (0.24) n=1090	2.17 (0.19) n=110	$1.37x10^{-5}$
Updated 14 SNP GS - Unweighted	12.79 (1.70) n=1294	13.47 (1.53) n=133	$3.61x10^{-6}$
Updated 14 SNP GS - Weighted	2.01 (0.24) n=1294	2.10 (0.21) n=133	$1.01x10^{-6}$

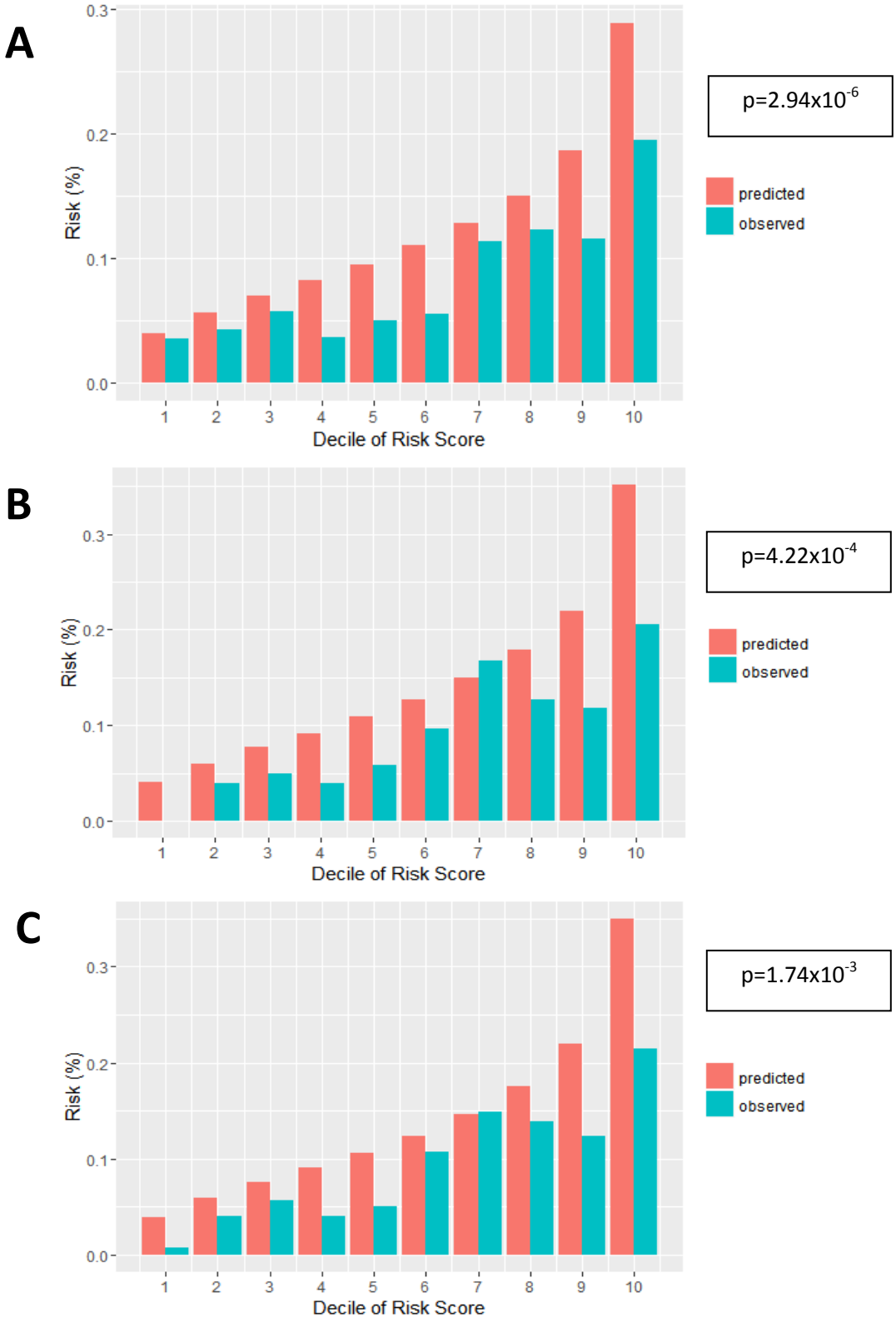
The mean (sd) are presented. The means were compared using Welch's t-test. CI=confidence interval. GS=gene score. CHD=coronary heart disease.

Table 47: Association between updated GSs and CHD in NPHSII

Score	Odds Ratio (95% CI)	p-value
Updated 19 SNP GS - Unweighted	1.38 (1.13-1.68)	$1.83 x10^{-3}$
Updated 19 SNP GS -Weighted	1.47 (1.20-1.80)	$1.99 x10^{-4}$
Updated 14 SNP GS - Unweighted	1.50 (1.25-1.80)	$1.56x10^{-5}$
Updated 14 SNP GS - Weighted	1.51 (1.26-1.82)	$8.90x10^{-6}$

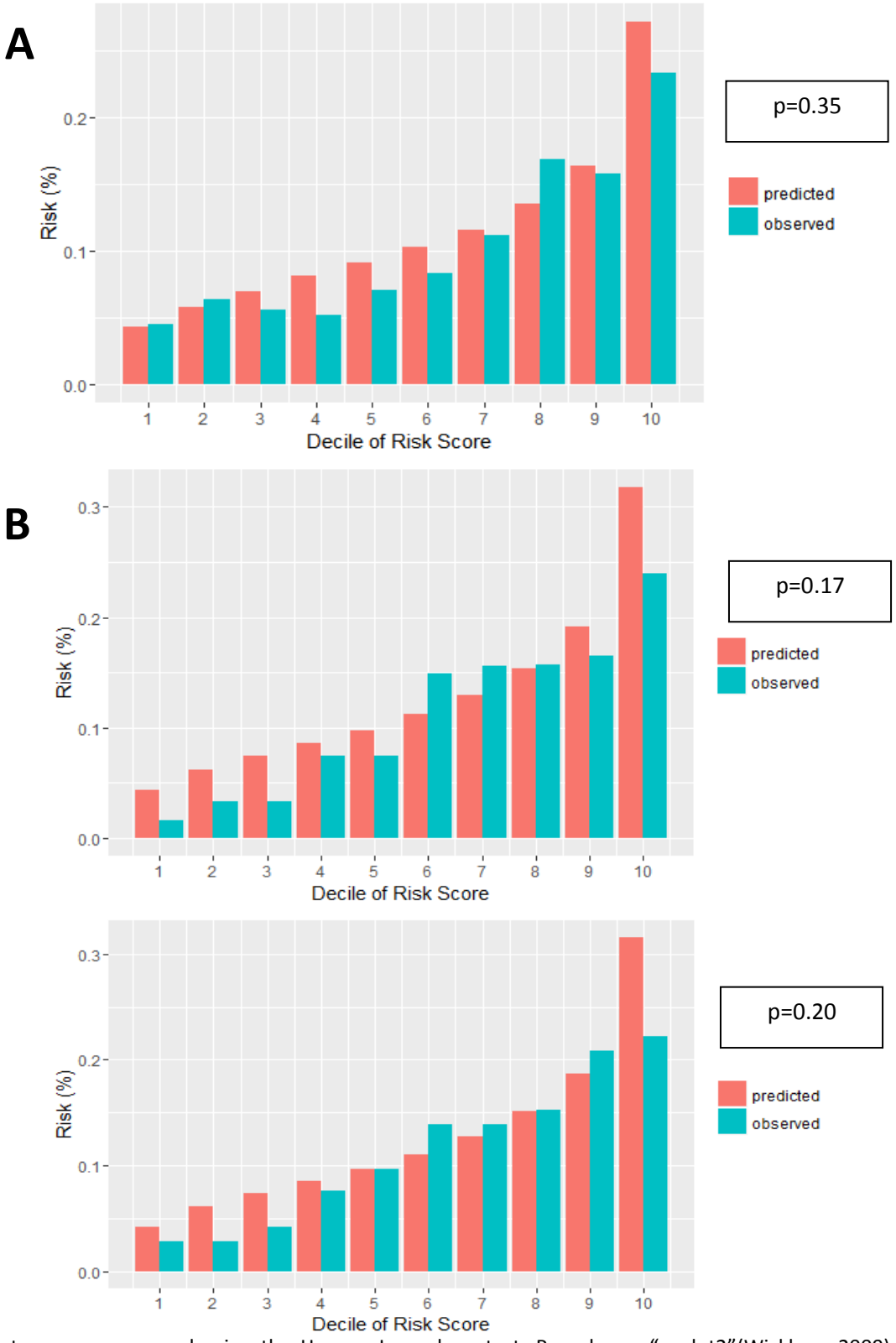
Effect sizes relate to one standard deviation of the variable and were determined by logistic regression, adjusted for age. OR=odds ratio. CI=confidence interval. CHD=coronary heart disease.

Figure 15: Observed CHD event rate in NPHSII compared to the predicted event rate determined by A) Framingham score alone; B) Framingham plus updated 19 SNP GS; C) Framingham plus 14 SNP GS, presented by decile of risk score.



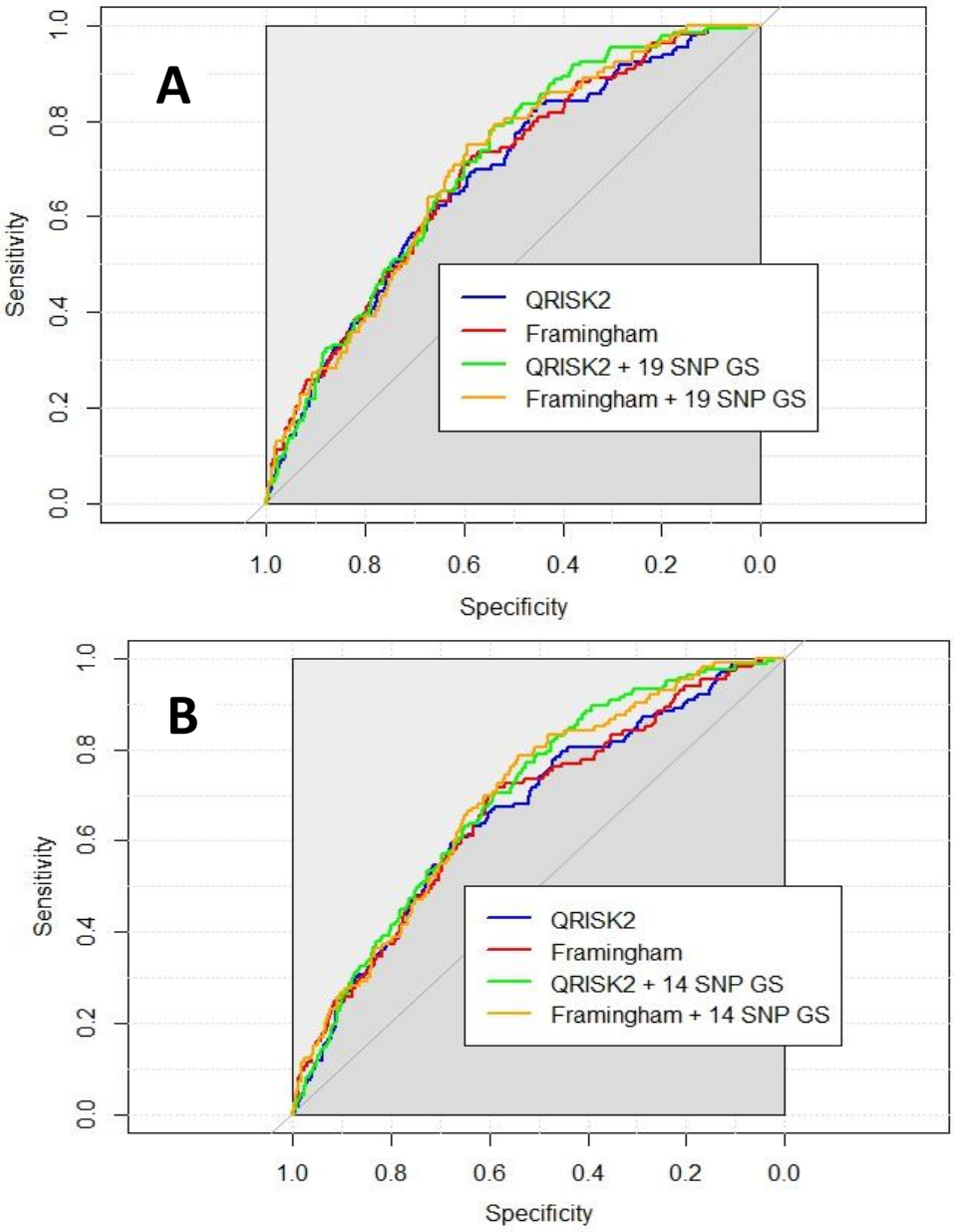
Rates were compared using the Hosmer-Lemeshow test. R packages "ggplot2"(Wickham 2009)," PredictABEL"(Kundu, Aulchenko et al. 2011; Kundu, Aulchenko et al. 2014) and "ResourceSelection" (Lele, Keim et al. 2014) were used to perform the analysis and produce the plots. However, p-values were calculated separately using ten degrees of freedom, rather than with the eight calculated with the R packages.

Figure 16: Observed CHD event rate in NPHSII compared to the predicated event rate determined by A) QRISK2 score alone; B) QRISK2 plus updated 19 SNP GS; C) QRISK2 plus 14 SNP GS, presented by decile of risk score



Rates were compared using the Hosmer-Lemeshow test. R packages "ggplot2"(Wickham 2009)," PredictABEL"(Kundu, Aulchenko et al. 2014) and "ResourceSelection"(Lele, Keim et al. 2014) were used to perform the analysis and produce the plots. However, p-values were calculated separately using ten degrees of freedom, rather than with the eight calculated with the R packages.

Figure 17: ROC curves for different risk scores – CRF score alone with the addition of the one the updated GSs. A) CRF scores and the updated 19 SNP GS and B) CRF score and the 14 SNP GS



Plots were created using the R package "pROC" (Robin, Turck et al. 2011).

Table 48: AUROC for combined CRF plus updated GSs

Combined Score	AUROC (95% CI)	CRF score	AUROC (95% CI)	p-value
FRAM+19 SNP Updated GS	0.70 (0.65-0.75)	FRAM	0.69 (0.64-0.74)	0.78
FRAM+14 SNP Updated GS	0.69 (0.64-0.74)	FRAM	0.67 (0.61-0.72)	7.52×10^{-3}
QRISK2+19 SNP Updated GS	0.70 (0.66-0.75)	QRISK2	0.68 (0.64-0.73)	0.02
QRISK2+14 SNP Updated GS	0.69 (0.65-0.73)	QRISK2	0.66 (0.62-0.70)	4.69×10^{-4}

AUROC were compared using DeLong's test, part of the R package pROC (Robin, Turck et al. 2011)
CI=confidence interval. AUROC=area under the ROC curve. CRF=conventional risk factor. GS=gene score. FRAM=Framingham risk score.

Table 49: Reclassification of NPHII participants with the addition of the updated GSs to the CRF scores

Risk Score	Reclassified at lower risk	No change in risk classification	Reclassified at higher risk	NRI (95 % CIs)	p-value
FRAM + 19 SNP GS					
No CHD	49	828	53	0.06 (0.01-0.12)	0.03
CHD	0	86	6		
Event rate	0 %	9.40 %	10.17 %		
FRAM + 14 SNP GS					
No CHD	55	982	64	0.06 (0.01-0.11)	0.02
CHD	0	105	8		
Event rate	0 %	9.66 %	10.26%		
QRISK2 + 19 SNP GS					
No CHD	51	945	84	0.07 (0.002-0.13)	0.04
CHD	3	114	16		
Event rate	5.56 %	10.76 %	16.00 %		
QRISK2 + 14 SNP GS					
No CHD	63	1115	99	0.06 (0.01-0.12)	0.03
CHD	3	142	18		
Event rate	4.54 %	11.30 %	15.38 %		

10 % was used as the high risk cut-off. FRAM=Framingham risk score. GS=gene score. CRF=conventional risk factor. NRI=net reclassification index. CHD=coronary heart disease. CI=confidence interval.

3.2.3.2.4 Assessing the updated GS in the South Asian and Afro-Caribbean populations

The updated GSs were also assessed in the cohorts from Islamabad, Lahore and Guadeloupe. Both weighted GSs were higher in the cases compared to the controls in the Islamabad (19 SNP GS p=0.01, 14 SNP GS p=0.01, Table 50) and Guadeloupe (19 SNP GS p=0.01, 14 SNP GS $p=9x10^{-4}$;Table 50) cohorts but not in the Lahore cohort (19 SNP GS p=0.83, 14 SNP GS p=0.88; Table 50). After adjustment for age and sex the GSs were not associated with CHD in either Pakistani group (all p>0.05; Table 51) but were associated with CHD in the Guadeloupe cohort (both weighted and unweighted, Table 51; Figure 18). When the association between the GSs and CHD was adjusted for CRFs (age, sex, hypertension, diabetes, hypercholesterolemia and smoking) the association remained for the unweighted score only (19 SNP GS $p=8x10^{-3}$, 14 SNP GS $p=4x10^{-3}$).

Table 50: Updated GS values in the Pakistani and Afro-Caribbean cohorts

		19 SNP GS			14 SNP GS		
Study	Score	Controls	Cases	p-value	Controls	Cases	p-value
Islamabad	Unweighted	14.90 (2.33)	15.42 (2.10)	0.05	12.23 (1.55)	12.73 (1.90)	0.01
	Weighted	1.99 (0.19)	2.06 (0.22)	0.01	1.94 (0.19)	2.00 (0.22)	0.01
Lahore	Unweighted	13.65 (2.36)	13.58 (2.31)	0.78	10.24 (1.88)	10.35 (1.89)	0.55
	Weighted	1.36 (0.22)	1.36 (0.23)	0.83	1.29 (0.22)	1.29 (0.23)	0.88
Guadeloupe	Unweighted	13.17 (2.07)	13.90 (2.10)	$1.60x10^{-4}$	9.69 (1.70)	10.28 (1.68)	$1.60x10^{-4}$
	Weighted	1.68 (0.20)	1.74 (0.24)	0.01	1.61 (0.21)	1.66 (0.23)	$9.00x10^{-4}$

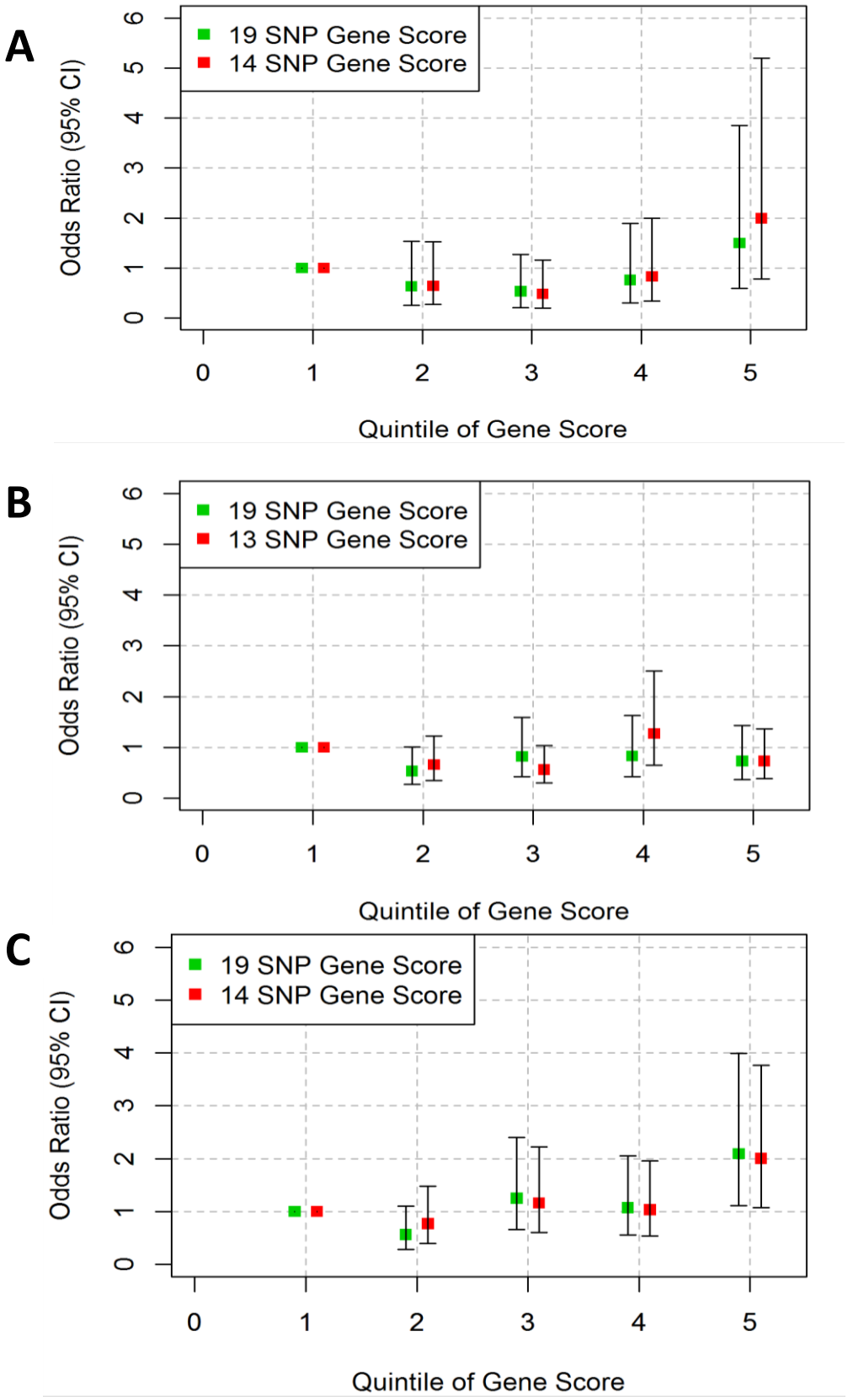
Mean GS and standard deviation are shown. Welch's t-test was used to compare the GSs between cases and controls. GS=gene score.

Table 51: Association between updated GSs and outcome in Pakistani and Afro-Caribbean cohorts

		19 SNP GS		13 SNP GS	
Study	Score	OR (95% CI)	p-value	OR (95% CI)	p-value
Islamabad	Unweighted	0.93 (0.68-1.27)	0.65	1.12 (0.84-1.52)	0.44
	Weighted	1.11 (0.82-1.50)	0.50	1.15 (0.86-1.55)	0.34
Lahore	Unweighted	0.97 (0.79-1.19)	0.76	1.06 (0.87-1.29)	0.58
	Weighted	0.97 (0.79-1.19)	0.77	0.98 (0.80-1.19)	0.82
Guadeloupe	Unweighted	1.50 (1.22-1.86)	$9.29x10^{-3}$	1.41 (1.15-1.73)	$1.16x10^{-3}$
	Weighted	1.32 (1.08-1.64)	$7.37x10^{-3}$	1.32 (1.07-1.62)	$9.29x10^{-3}$

Effect sizes relate to one standard deviation of the GS Logistic regression, adjusted for age and sex, was performed in each case. GS=gene score. OR=odds ratio.CI=confidence interval.

Figure 18: Association between updated weighted GS and outcome in A) the Islamabad study, B) the Lahore study and C) the Guadeloupe study



Logistic regression was performed in each case, adjusted for age and sex. Error bars represent 95 % CI. CI=confidence intervals. GS = gene score.

3.3 Discussion

In this study a CHD GS comprising 19 SNPs selected from candidate gene studies and early GWASs was investigated using data from a prospective study of healthy UK middle-aged men (NPHSII). A smaller 13 SNP GS derived from the 19 SNPs was also assessed. The GSs were originally weighted using the effect sizes determined in meta-analyses of candidate gene studies or GWASs. Both the original 19 SNP and 13 SNP GSs were associated with CHD. However, addition of the GSs to CRF scores (Framingham and QRISK2) to give a combined CHD risk score did not show an improved performance compared to the CRF score alone. Updating the weightings to those determined in the CARDIoGRAM GWAS or CARDIoGRAMplusC4D analysis was found to improve the performance of the 19 SNP GSs. A 14 SNP GS based with five of the 19 SNPs removed (those that did not show an association with CHD in the CARDIoGRAMplusC4D meta-analysis) was also assessed. The combined QRISK2 plus GS risk score showed improved discrimination and reclassification while maintaining good calibration compared to QRISK2 alone. Therefore, our results indicate that including the updated GSs along with QRISK2 may have clinical utility in the UK population. Furthermore, as the updated 19 SNP and 14 SNP GSs performed equally well, only these 14 SNPs of the original 19 require to be used. The results also show that QRISK2 was better at predicating cardiovascular outcome in NPHSII compared to the Framingham score, with the Framingham score overestimating risk in NPHSII. This is consistent with the literature where even the NICE-adjusted Framingham risk equations have been found to overestimate ten-year CHD risk in the UK population, particularly in men (Collins and Altman 2010). The superior performance of QRISK2 compared to the Framingham score is unsurprising given that QRISK2 was derived from a very large British cohort while Framingham was developed from the Framingham study based in Massachusetts, USA (Wilson, D'Agostino et al. 1998; Hippisley-Cox, Coupland et al. 2008; Collins and Altman 2010).

The updated GSs were weighted using the results from the CARDIoGRAMplusC4D meta-analysis (Deloukas, Kanoni et al. 2013). This work has since been expanded and a GWAS meta-analysis investigating over 9 million SNPs using haplotype data from the 1000 genomes project in approximately 185,000 individuals was recently published (Nikpay, Goel et al. 2015). Ten new CHD risk loci were identified (eight from an additive model and two from a recessive model). Association analysis for all the SNPs in the 19 SNP GS with CHD was performed. A comparison of the weightings used in the updated GS and the effect sizes

determined in this meta-analysis is shown in Table 52. For the most part, the effect sizes are very similar and thus it was felt unnecessary to update the weightings in the GS.

Table 52: Comparison of the effect size for the 19 CHD GS SNPs between two CARDIoGRAMplusC4D meta-analyses

Gene/Locus	SNP	Risk Allele	OR in Cardiogram-plusC4D meta-analysis (Deloukas, Kanoni et al. 2013)	OR in most recent meta-analysis (Nikpay, Goel et al. 2015)	p-value in most recent meta-analysis (Nikpay, Goel et al. 2015)
APOE	rs7412	C	1.25	1.14	$8.17x10^{-11}$
APOE	rs429358	C	1.06	1.10	$2.17x10^{-9}$
MIA3	rs17465637	C	1.14	1.08	$3.52x10^{-12}$
MRAS	rs9818870	T	1.07	1.07	$2.21x10^{-6}$
DAB2IP	rs7025486	A	1.04	1.05	$3.26x10^{-6}$
CXCL12	rs1746048	C	1.07[b]	1.08	$6.49x10^{-11}$
APOA5	rs662799	G	1.05	1.06	$4.23x10^{-4}$
SORT1	rs599839	A	1.11	1.10	$9.01x10^{-19}$
SMAD3	rs17228212	C	1.01	1.03	$9.30x10^{-3}$
ACE	rs4341	G	1.01[c]	1.01	0.24
LPL	rs328	C	1.09	1.05	$1.81x10^{-3}$
CETP	rs708272	C	1.04[d]	1.02	0.01
CDKN2A/9p21	rs10757274	G	1.23[a]	1.22	$2.49x10^{-38}$
NOS3	rs1799983	G	1.00	1.03	$4.28x103^{1}$
LPL	rs1801177	A	1.10[e]	1.13	$2.21x10^{-3}$
PCSK9	rs11591147	G	1.39	1.29	$7.47x10^{-6}$
LPA	rs10455872	G	1.32	1.38	$5.73x10^{-39}$
APOB	rs1042031	A	1.01	1.00	0.71
LPA	rs3798220	C	1.28	1.42	$4.66x10^{-9}$

[a]Weighting for rs1333049 (r^2=0.88). [b]Weighting for rs501120 used (r^2=0.97). [c]Weighting for rs4343 (r^2=0.96). [d]Weighting for rs711752 (r^2=1). [e]Weighting for rs7016529 (r^2=1). All r^2 values calculated from 1000 Genomes phase 1 EUR data. Data on coronary artery disease / myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.CARDIOGRAMPLUSC4D.ORG. OR= Odds Ratio. Cplus4D = CARDIoGRAMplusC4D meta-analysis. CG GWAS=CARDIoGRAM GWAS.

In comparison to the genetics of CHD in those of Europeans ethnicity, very little is known about the genetics of CHD in either the South Asian or Afro-Caribbean populations. Here, the GSs were assessed in two case-control cohorts from Pakistan and one from Guadeloupe in the Caribbean. Both the original and updated GSs were associated with CHD (after adjustment for age and sex) in the Afro-Caribbean cohort. The GSs (apart from the 19 SNP GS with the original weightings) were higher in cases compared to controls for both updated GSs in the Islamabad sample, although neither score was associated with CHD after adjustment for age and sex. There was no difference for any of the GSs between cases and controls in the Lahore group. The lack of an association observed for the updated GSs in the Lahore study is unlikely to be due to low power. To have 80% power (at the p=0.05

significance threshold) to detect a similar difference in updated mean GS (for either 19 SNP or 14 SNP updated GS) as was seen in NPHSII, 91 cases and 91 controls are required. The number of cases and controls exceeded this. Rather, the poor performance of the GSs in the Lahore group can be at least partly attributed to the much broader definition of CHD used in recruitment the case group compared to the Islamabad study which used an MI phenotype (more like the "hard" endpoints used in the prospective NPHSII and the Guadeloupe study). Overall, the results indicate that the GSs provide a useful estimate of genetic CHD risk in Afro-Caribbeans from Guadeloupe at least, but firm conclusions cannot be drawn from the Pakistani data.

The improved performance of the GSs with the updated weightings demonstrates that the effect sizes derived from the CARDIoGRAMplusC4D analysis more accurately reflect the impact of the SNP CHD risk. All of the updated weightings were lower, indicating that the original effect sizes were inflated. This is a common problem in genetic studies (Kraft 2008). However, in the Afro-Caribbean group, the updated unweighted GSs remained associated with CHD after adjustment for multiple risk factors while the weighted score did not (Larifla, Beaney et al. 2016). It would be expected that weighting SNPs for their individual impact on CHD risk would improve performance, rather than assuming that all SNPs have the same magnitude of effect. The findings of this study can be partly explained by differing LD patterns between Afro-Caribbeans and those of European ethnicity. A number of the SNPs included in the score are GWAS hits where the lead SNP (i.e. that included in the score) is unlikely to be the functional SNP at that risk locus. LD between the lead and functional SNPs may differ between ethnicities, meaning that some SNPs are better proxies than others. This will reduce the ability of the weighted GS to accurately reflect CHD risk. This problem would be removed by the identification of the functional SNP (or possibly SNPs) at each risk locus.

The results of the systematic literature search demonstrated that SNP selection, which was performed in the early days of GWASs, was suboptimal in light of recent meta-analyses. In order to construct a CHD risk GS now a reasonable strategy would be to use the results of the CARDIoGRAMplusC4D meta-analysis (Deloukas, Kanoni et al. 2013). GSs based on this work have been assessed by others. In six prospective studies with over 10,000 participants of European ethnicity a CRF-CARDIoGRAMplusC4D GS score improved discrimination and reclassification over above the CRF score alone while maintaining good calibration (Ganna,

Magnusson et al. 2013). In the Rotterdam Study the CARDIoGRAMplusC4D GS showed no improvement over CRFs alone (de Vries, Kavousi et al. 2015). In the UCLEB consortium participants a combined QRISK2 plus CARDIoGRAMplusC4D GS score was found to have potential clinical utility for those previously classified as being at intermediate risk although overall the combined score did not show additional benefit over and above QRISK2 alone (Morris, Cooper et al. 2016). It would be interesting to compare the performance of such a score with the 19 SNP/14 SNP GSs developed here. However, the biochip technology utilised by the Randox Cardiac Risk Prediction array (Chapter 2.3.3) developed to genotype the SNPs in a clinical setting is limited to 23 SNPs. Therefore, it would be more pertinent to compare a GS of the top 23 ranked CARDIoGRAMplusC4D SNPs (by RAF multiplied by effect size, Table 53) with the GSs assessed herein. Constructing large-scale GSs using tens of thousands of SNPs with CHD risk estimates (from GWASs) has also been suggested (Dudbridge 2013). While such GSs may ultimately out-perform those constructed with only robustly associated SNPs, such an approach is not practical for a clinical setting, at least in the short-to-medium term.

Table 53: Top 25 CARDIoGRAMplusC4D CHD risk loci ranked by ln(OR) multiplied by RAF

Chromosome	Lead SNP	Gene/Locus	OR	RAF	ln(OR) x RAF
19	rs445925	*ApoE-ApoC1*	1.13	0.9	0.110
4	rs7692387	*GUCY1A3*	1.13	0.81	0.099
9	rs1333049	9p21	1.23	0.47	0.097
1	rs17114036	*PPAP2B*	1.11	0.91	0.095
1	rs602633	*SORT1*	1.12	0.77	0.087
10	rs12413409	*CYP17A1-CNNM2-NT5C2*	1.1	0.89	0.085
19	rs1122608	*LDLR*	1.1	0.76	0.072
13	rs9515203	*COL4A1-COL4A2*	1.08	0.74	0.057
9	rs3217992	9p21	1.16	0.38	0.056
10	rs501120	*CXCL12*	1.07	0.83	0.056
6	rs9369640	*PHACTR1*	1.09	0.65	0.056
7	rs11556924	*ZC3HC1*	1.09	0.65	0.056
8	rs264	*LPL*	1.06	0.86	0.050
13	rs4773144	*COL4A1-COL4A2*	1.07	0.74	0.050
6	rs4252120	*PLG*	1.07	0.73	0.049
1	rs11206510	*PCSK9*	1.06	0.84	0.049
1	rs17464857	*MIA3*	1.05	0.87	0.042
1	rs4846525	*IL6R*	1.09	0.47	0.041
15	rs7173743	*ADAMTS7*	1.07	0.58	0.039
17	rs12936587	*RAI1-PEMT-RASD1*	1.06	0.59	0.034
6	rs12205331	*ANKS1A*	1.04	0.81	0.032
13	rs9319428	*FLT1*	1.1	0.32	0.030
2	rs1561198	*VAMP5-VAMP8-GGCX*	1.07	0.45	0.030
12	rs3184504	*SH2B3*	1.07	0.4	0.027
8	rs2954029	*TRIB1*	1.05	0.55	0.027

OR=odds ratio. RAF=risk allele frequency.

The literature search also found three CHD risk loci which had not been identified in the CARDIoGRAMplusC4D analysis. One variant tags a three codon deletion in the signal peptide region of the *APOB* gene (Boerwinkle and Chan 1989). The second risk locus tags an insertion/deletion variant in *NFKB1* (Vogel, Jensen et al. 2011). The SNP used in genotyping is not available on large scale genotyping chips and no suitable proxies have, as yet, been identified. Another meta-analysis was identified which found an association between variants in *TGF-B* and CHD under a dominant model (Morris, Moxon et al. 2012) (although it is noteworthy that one *TGF-B* variant was included in CARDIoGRAMplusC4D and did not show even a suggestive association p=0.22). Typically, the additive model is used in the GWAS analysis (although this was not the case with the most publication from the CARDIoGAMplusC4D consortium (Nikpay, Goel et al. 2015), so if there is a different relationship between risk allele and CHD, this can easily be missed. At the present the evidence for a robust association between these SNPs and CHD is not as strong as for the CARDIoGRAMplusC4D confirmed loci but a true association cannot be discounted. This serves as a reminder that while the GWAS remains a crucial tool in risk loci discovery and replication, it does not give a complete account of the genetics of a particular phenotype. In addition, one must be mindful of the criteria use to select variants for replication in such large-scale analysis. Variants may not be included due to proximity to another risk locus despite there being strong evidence the effects are independent as was the case with *LPA* in CARDIoGRAMplusC4D.

This study has a number of limitations. All of the participants of NPHSII are male. While there is no evidence to suggest the SNP effect sizes differ between men and women, it is known that the pathogenesis of atherosclerosis is different between the sexes (Appelman, van Rijn et al. 2015). Therefore, it would be ideal to test the GSs in a mixed sample set. Data to calculate either the Framingham score or QRISK2 was not available in either the Pakistani or Afro-Caribbean studies and so the performance of a combined CRF plus genetics CHD risk score could not be assessed in these groups (and the case-control study design precludes assessment of predictive ability). In any case, these CRF scores may not be appropriate to use in these populations. An assessment of CRF scores in the multi-ethnic SABRE cohort in the UK found the performance of both the Framingham score and QRISK2 to vary between ethnicities, being poorer in South Asian women and Afro-Caribbean individuals of both sexes (Tillin, Hughes et al. 2014). Therefore, it may be that specific CHD risk calculators for these populations are required. It might also be necessary to tailor the

GS to particular ethnic groups as all of the variants used in the GSs assessed here were identified in those of European ethnicity. Some of the rarer risk alleles may not be present in all ethnicities - for example none of the participants of the Guadeloupe study carried the rare allele of the *PCSK9* SNP rs11591147. This indicates that in an Afro-Caribbean specific GS, this SNP should be removed. This is only proposed for completely monomorphic variants, as rare variants usually carry relatively large effect sizes and it is important to identify carriers of such variants (e.g. rs3798820 in *LPA*). Moreover, future large-scale analysis in different ethnic groups may identify CHD risk variants which are not present in populations of European ethnicity would merit inclusion in a CHD risk GS used in other populations. It is noteworthy however, that studies performed in African Americans (Lettre, Palmer et al. 2011; Franceschini, Hu et al. 2014) have so far failed to find any such risk loci with large effect sizes, although sample size is relatively small compared to studies performed with individuals of European ethnicity. For the time-being GSs based on data derived from the population remains the best way to estimate genetic CHD risk.

A further limitation is that it is unclear how representative the results in cohorts studied here are of the general South Asian and Afro-Caribbean populations. A small number of Punjabi participants from Lahore (n=96) were genotyped as part of the 1000 Genomes project (phase 3) (Abecasis, Altshuler et al. 2010). The risk allele frequencies for all SNPs except rs10757274 did not differ between this group and the Pakistani control subjects presented here. Given that the frequencies of 19 SNPs were compared between the groups using a p=0.05 significance threshold, it is not unexpected to find one with a significantly different frequency between the two groups by chance. Overall the results suggest that the groups studied here are at least somewhat representative of the Punjabi population in Pakistan. However, five of the 19 SNPs were out of HWE in the Islamabad group. In each case this was due to an excess of homozygotes, indicating the presence of population sub-structure within the cohort. A small number (n=96) of Afro-Caribbean participants from Barbados were also included in the 1000 Genomes phase 3. The RAF differed between this group and the Guadeloupe controls for four SNPs. While this indicates that the results presented here will have some relevance to the wider Afro-Caribbean population, it also highlights the genetic diversity within this ethnic group. Evidently, both genetic and CRF data derived from much larger studies reflecting the different regions of South Asia and the Caribbean is required.

3.4 Conclusion to chapter

The use of the CHD risk GS was optimised by updating the weightings to the effect sizes determined in the CARDIoGRAMplustC4D analysis. Addition of the GS to QRISK2 was found to have potential clinical utility in the UK population. The results from an Afro-Caribbean cohort suggested the GS may also have clinical utility in this group but the results from the Pakistani cohorts were inconclusive.

4 The genetics of CHD in T2D

4.1 Introduction

Data from epidemiological studies has shown that those with T2D have an approximately two-fold greater risk of developing CHD (Woodward, Zhang et al. 2003; Sarwar, Gao et al. 2010). However, given that the two diseases share many common risk factors, it has been difficult to ascertain whether the diabetic state is itself contributing to the pathogenesis of CHD. Recent Mendelian randomisation studies have found that T2D risk variants and variants associated with fasting glucose levels are also associated with CHD, providing evidence for a causal relationship between T2D and CHD (Ahmad, Morris et al. 2015; Jansen, Loley et al. 2015). This is supported by a number of meta-analyses of RCTs that have found improved glycaemic control reduces risk of cardiovascular events (Stettler, Allemann et al. 2006; Mannucci, Monami et al. 2009; Ray, Seshasai et al. 2009). Work is ongoing to identify the mechanism(s) through which diabetes impacts upon CHD risk.

A number of risk factors for CHD in T2D have been identified, including duration of diabetes (Fox, Sullivan et al. 2004; Wannamethee, Shaper et al. 2011) and elevated glycated haemoglobin (Selvin, Marinopoulos et al. 2004). Given this and that those with T2D are already a high risk group specific CHD risk scores for T2D have been developed (Chamnan, Simmons et al. 2009). Data from the UKPDS study was used to construct a T2D-specific CHD risk score (Stevens, Kothari et al. 2001). The 2008 NICE guidelines recommended its use for CHD risk assessment in those with T2D (2008). However, external validation of the UKPDS score with approximately 80,000 newly diagnosed T2D cases taken from the Clinical Practice Research Database found it was poorly calibrated, greatly overestimating risk (Bannister, Poole et al. 2014). The authors concluded that revised risk scores are required. NICE updated its guidance, recommending the use of QRISK2 for those with T2D (2014). Despite the lack of external validation, the Guideline Development Group felt that as the QRISK2 development cohort contains 40,000 individuals with T2D and that QRISK2 is updated regularly, it is the most suitable tool available. Whether updated risk scores for CHD in T2D, developed from a T2D-only cohort will outperform QRISK2 remains to be seen.

The potential of genetics to improve CHD risk prediction was demonstrated in Chapter 3 but whether a general CHD GS is suitable for use in T2D is unknown. It may be more appropriate to use a CHD in T2D specific GS to provide an estimate of genetic risk in this population. Therefore, the aim of this study was to identify risk loci for CHD in T2D from the literature, to construct a GS and assess its performance with data from the UCLEB

consortium. In addition, the general 19 SNP CHD GS was also calculated to assess its performance in those with T2D. Furthermore, by studying the mechanism(s) through which risk variants mediate their effect, genetics can also play a role in elucidating how the diabetic state pre-disposes individuals to CHD. As such, the final aim of this study was to investigate the relationship between the risk variant rs10911021 and CHD in T2D, particularly the relationship between the SNP and T2D-CHD risk factors.

4.2 Results

4.2.1 Systematic literature search results

As when identifying variants associated with the general CHD phenotype, variants associated with CHD in those with T2D were identified in a systematic literature search. The methodology used for the literature search was described in Chapter 2.2 and the workflow is depicted in Figure 14. One meta-analysis which studied variants associated with CHD in T2D was found (Qi, Parast et al. 2011). In this study, twelve GWAS hits for CHD in the general population were genotyped in three cohorts with T2D. Five of the variants were found to be associated with CHD in T2D (Table 54). The authors constructed an unweighted GS using these five SNPs and found it to be associated with CHD in the same three T2D cohorts that were used for the meta-analysis (approximately 1000 CHD cases and 1400 CHD controls), although statistical adjustments were made to account for this. Furthermore, they found that addition of the GS to a CRF model, including glycated haemoglobin, improved discrimination and reclassification compared to the CRF model alone.

Following the completion of the literature search, another locus associated with CHD in T2D was identified by the same group (published September 2013). This locus on chromosome 1 (lead SNP rs10911021) was the first to be associated with CHD in T2D that had not been associated with CHD in the general population (Qi, Qi et al. 2013) and details of the association are shown in Table 54. The SNP was genotyped in the CARDIoGRAM GWAS and found to have a suggestive association with CHD (OR=1.04, p=0.01). The authors state that this effect size was in accordance with what would be expected based on the effect sizes they observed in the T2D and no T2D groups and the prevalence of T2D in the CARDIoGRAM participants. This SNP is studied in section 4.2.4 of this chapter.

Table 54: Variants found to be associated with CHD in T2D.

SNP	Gene/Locus	Risk Allele	OR	Study
rs4977574	9p21	G	1.21	(Qi, Parast et al. 2011)
rs12526453	*PHACTR1*	C	1.25	(Qi, Parast et al. 2011)
rs646776	*SORT1*	T	1.17	(Qi, Parast et al. 2011)
rs2259816	*HNF1A*	T	1.17	(Qi, Parast et al. 2011)
rs11206510	*PCSK9*	T	1.26	(Qi, Parast et al. 2011)
rs10911021	*GLUL*	C	1.36	(Qi, Qi et al. 2013)

The data is taken from either a meta-analysis of 3 independent T2D cohorts (Qi, Parast et al. 2011) or 5 independent T2D cohorts (Qi, Qi et al. 2013). OR=odds ratio.

4.2.2 CHD in T2D GSs

4.2.2.1 Association of CHD in T2D GSs with CHD in T2D

Two CHD in T2D GSs were assessed using data from the UCLEB consortium. One included the five SNPs identified in Qi, Parast et al. 2011 (5 SNP GS) while the other included these plus the chromosome 1 variant, rs10911021 (6 SNP GS). All SNPs were treated additively. The SNPs were weighted using the effect sizes determined in the source publications as listed in Table 54. Of the six SNPs found to be associated with CHD in T2D, two had been genotyped in UCLEB (rs11206510 close to *PCSK9* and rs646776 at the *SORT1* locus). The lead SNP identified by Qi, Parast et al. at the 9p21 locus (rs4977547) had not been genotyped in UCLEB but genotyping data for rs10757274 (r^2=0.99 with rs4977547 as determined in CEU 1000 Genomes phase 1 data) was available. The three other SNPs included in the GS had been imputed. Genotype, diabetic status and follow-up data were available for eight of the UCLEB cohorts (Table 55). Only those with prevalent diabetes at baseline were included to ensure that all CHD events occurred following a diagnosis of T2D.

In total there was complete data for 1535 participants with T2D. There were 160 CHD events during follow-up. The mean weighted and unweighted GS were higher in those who went on to develop CHD for both the 5 SNP and 6 SNP GSs (Table 56 and Table 57). The association between the GSs and CHD was analysed using logistic regression for each study and the results were combined by meta-analysis. Under a FE model both unweighted scores were found to be associated with CHD (6 SNP GS p=2.39x10^{-3} and 5 SNP GS p=0.02; Figure 19). For the 5 SNP GS, each additional risk allele was associated with an increased risk of CHD OR=1.17 (95 % CIs 1.02-1.33), similar to the effect size observed in the original publication (OR=1.19 (95 % CIs 1.13-1.26)) (Qi, Parast et al. 2011). The weighted GSs were also associated with CHD in T2D under a FE model (6 SNP GS p=2.16x10^{-3} and 5 SNP GS p=0.02). However, while there was no evidence of heterogeneity between the studies for the 5 SNP GS (unweighted I^2=0 %, p=0.43; weighted I^2=7 %, p=0.38), there appeared to be at least moderate heterogeneity for the 6 SNP GS, particularly the unweighted score (unweighted I^2=31%, p=0.14, weighted I^2=47%, p=0.07). Under a RE model (using the DerSimonian Laird method), the association was no longer statistically significant for either the unweighted (p=0.14) or weighted score (p=0.06).

Table 55: Number of participants in UCLEB cohorts with genotype, baseline T2D and CHD follow-up data

	BRHS	**BWHHS**	**CAPS**	**EAS**	**ELSA**	**ET2DS**	**MRC1946**	**WHII**	**Total**
Participants with complete genotype and CHD data (n)	2182	1819	1319	744	1705	823	2406	3041	14039
Participants with prevalent T2D (% of total)	261 (12 %)	107 (6 %)	36 (3 %)	59 (8 %)	167 (10 %)	823 (100 %)	50 (2 %)	32 (1 %)	1535 (11 %)
Incident CHD cases (n) (% of T2D total)	72 (28 %)	13 (12 %)	16 (44 %)	13 (22 %)	7 (4 %)	31 (4 %)	5 (10 %)	3 (9 %)	160 (11 %)
Approximate length of follow-up (years)	15	7	9	20	4	4	12	5	-

The median number of follow-up years is shown.

Table 56: Mean weighted and unweighted 6 SNP GS in UCLEB T2D participants who did and did not go on to develop CHD.

Study	Score	No CHD	CHD	p-value
BRHS	Weighted GS (sd)	1.49 (0.32)	1.65 (0.30)	4.05×10^{-4}
	Unweighted GS (sd)	6.94 (1.49)	7.56 (1.45)	2.74×10^{-3}
BWHHS	Weighted GS (sd)	1.42 (0.40)	1.67 (0.34)	0.03
	Unweighted GS (sd)	6.57 (1.87)	7.69 (1.65)	0.04
CAPS	Weighted GS (sd)	1.46 (0.31)	1.44 (0.37)	0.84
	Unweighted GS (sd)	6.75 (1.41)	6.75 (1.57)	1
EAS	Weighted GS (sd)	1.58 (0.34)	1.63 (0.27)	0.61
	Unweighted GS (sd)	7.21 (1.56)	7.62 (1.04)	0.29
ELSA	Weighted GS (sd)	1.48 (0.35)	1.42 (0.22)	0.58
	Unweighted GS (sd)	6.89 (1.59)	6.57 (0.98)	0.43
ET2DS	Weighted GS (sd)	1.50 (0.34)	1.45 (0.27)	0.23
	Unweighted GS (sd)	6.96 (1.56)	6.77 (1.28)	0.47
MRC1946	Weighted GS (sd)	1.52 (0.36)	1.79 (0.19)	0.03
	Unweighted GS (sd)	7.00 (1.68)	8.40 (0.89)	0.02
WHII	Weighted GS (sd)	1.51 (0.35)	1.55 (0.33)	0.86
	Unweighted GS (sd)	7.00 (1.58)	7.33 (1.53)	0.74
Combined	Weighted GS (sd)	-		2.39×10^{-3}
	Unweighted GS (sd)			2.16×10^{-3}

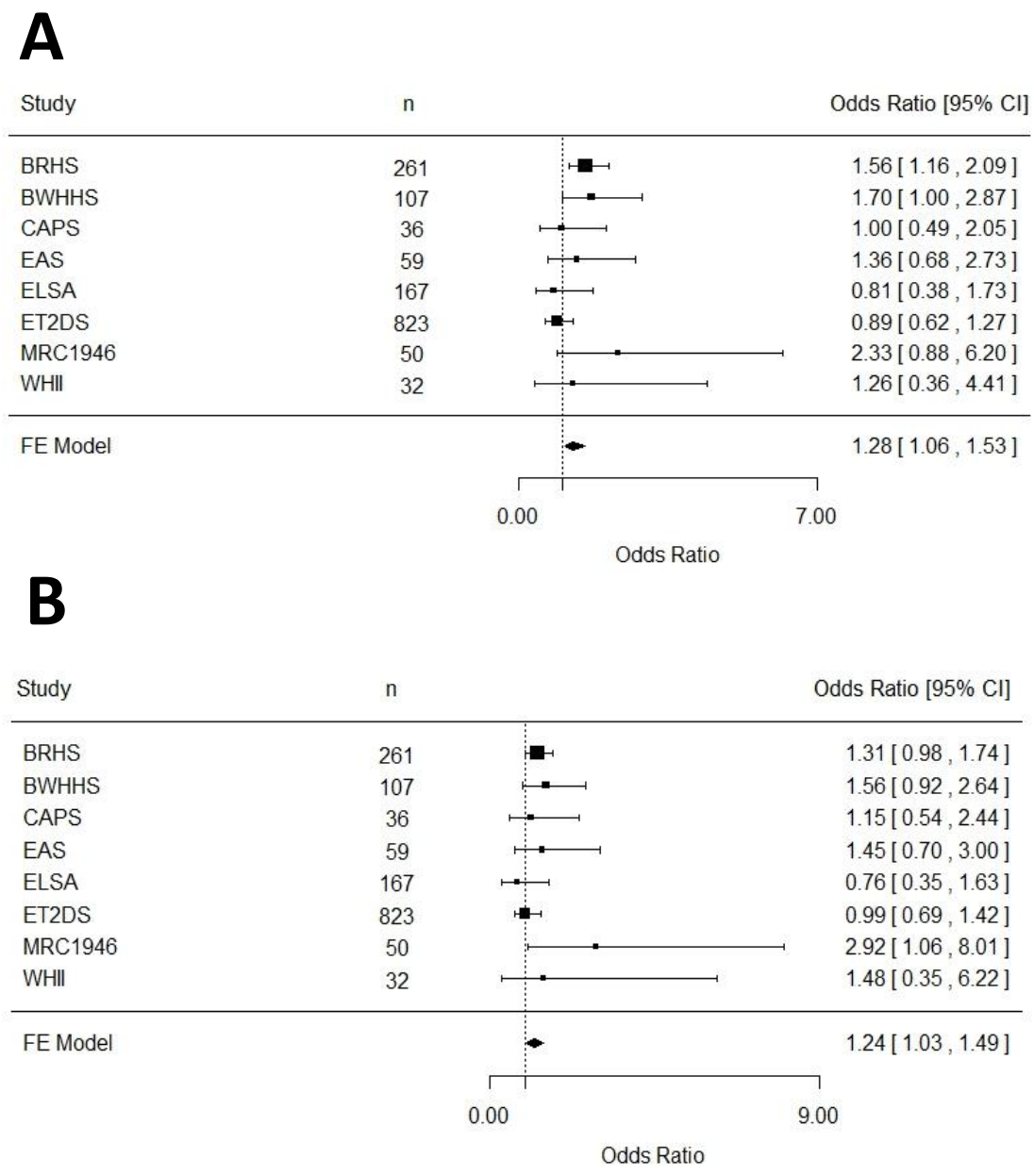
The mean (standard deviation) for each study is shown individually. Mean gene score between the CHD and no CHD groups were compared using Welch's t-tests in individual studies and by ANOVA (with study as a factor) for the combined analysis. GS=gene score.

Table 57: Mean weighted and unweighted 5 SNP GS in UCLEB T2D participants who did and did not go on to develop CHD.

Study	Score	No CHD	CHD	p-value
BRHS	Weighted GS (sd)	1.07 (0.26)	0.14 (0.26)	0.05
	Unweighted GS (sd)	5.56 (1.35)	5.90 (1.37)	0.07
BWHHS	Weighted GS (sd)	1.02 (0.32)	1.17 (0.29)	0.08
	Unweighted GS (sd)	5.26 (1.67)	6.08 (1.50)	0.09
CAPS	Weighted GS (sd)	1.03 (0.25)	1.06 (0.24)	0.78
	Unweighted GS (sd)	5.35 (1.31)	5.50 (1.21)	0.73
EAS	Weighted GS (sd)	1.10 (0.26)	1.15 (0.15)	0.41
	Unweighted GS (sd)	5.67 (1.38)	6.08 (0.76)	0.18
ELSA	Weighted GS (sd)	1.06 (0.29)	0.99 (0.17)	0.32
	Unweighted GS (sd)	5.54 (1.46)	5.14 (0.90)	0.31
ET2DS	Weighted GS (sd)	1.07 (0.27)	1.06 (0.25)	0.85
	Unweighted GS (sd)	5.56 (1.40)	5.55 (1.31)	0.95
MRC1946	Weighted GS (sd)	1.08 (0.30)	1.43 (0.14)	1.63×10^{-3}
	Unweighted GS (sd)	5.56 (1.55)	7.20 (0.84)	6.37×10^{-3}
WHII	Weighted GS (sd)	1.09 (0.23)	1.14 (0.17)	0.66
	Unweighted GS (sd)	5.62 (1.21)	6.00 (1.00)	0.59
Combined	Weighted GS (sd)			0.02
	Unweighted GS (sd)			0.02

The mean GS (standard deviation) for each study is shown individually. Mean GS between the CHD and no CHD groups were compared using Welch's t-tests in individual studies and by ANOVA (with study as a factor) for the combined analysis. GS=gene score.

Figure 19: Association between CHD and A) 6 SNP unweighted GS and B) 5 SNP unweighted GS (unadjusted).



The effect size (95% CI) determined per standard deviation of GS in each study is shown along with the combined effect size. A FE meta-analysis was performed using the R package "metafor" in both cases (Viechtbauer 2010).GS=gene score. FE=fixed effects. CI=confidence interval.

4.2.3 Association of CHD in T2D GSs with T2D-CHD risk factors

To further assess the relationship between the CHD in T2D GSs and CHD, whether the GSs were associated with conventional T2D-CHD risk factors was investigated. For each trait, linear regression with the GS was performed and the results were meta-analysed. An FE model was used unless there was evidence of heterogeneity between the studies (defined as I^2>30%, p<0.05). As shown in Table 58, only LDL-cholesterol showed an association with the unweighted 5 SNP GS (p=0.04) and a suggestive association with the weighted 5 SNP GS (p=0.06). Neither the weighted nor unweighted 6 SNP GS was associated with any of the T2D-CHD risk factors assessed (Table 59).

Table 58: Association between T2D-CHD risk factors and 5 SNP GSs in T2D participants of UCLEB

Trait	5 SNP GS - Unweighted		5 SNP GS - Weighted	
	Beta-coefficient (se)	p-value	Beta-coefficient (se)	p-value
BMI (kg/m^2)	-0.12 (0.09)	0.17	-0.63 (0.44)	0.15
Triglycerides* (mmol/l)	0.01 (0.01)	0.35	0.06 (0.08)	0.39
TC (mmol/l)	0.01 (0.02)	0.46	0.05 (0.09)	0.53
HDL-cholesterol (mmol/l)	0.002 (0.006)	0.79	-0.004 (0.03)	0.90
LDL-cholesterol (mmol/l)	0.05 (0.03)	0.04	0.24 (0.13)	0.06
Systolic blood pressure (mmHg)	0.12 (0.31)	0.38	0.95 (1.61)	0.56
Diastolic blood pressure (mmHg)	-0.13 (0.17)	0.45	-0.71 (0.89)	0.42
Fasting glucose* (mmol/l)	0.007 (0.005)	0.14	0.04 (0.03)	0.12
Insulin* (µIU/ml)	-0.07 (0.16)	0.66	0.05 (0.48)	0.91

For each variable linear regression was performed in individual UCLEB studies and the results meta-analysed using a fixed-effects model. Beta coefficient per unit increase and standard error are shown for each trait. *Variable log transformed. TC=total cholesterol. GS=gene score.

Table 59: Association between T2D-CHD risk factors and 6 SNP GSs in T2D participants of UCLEB

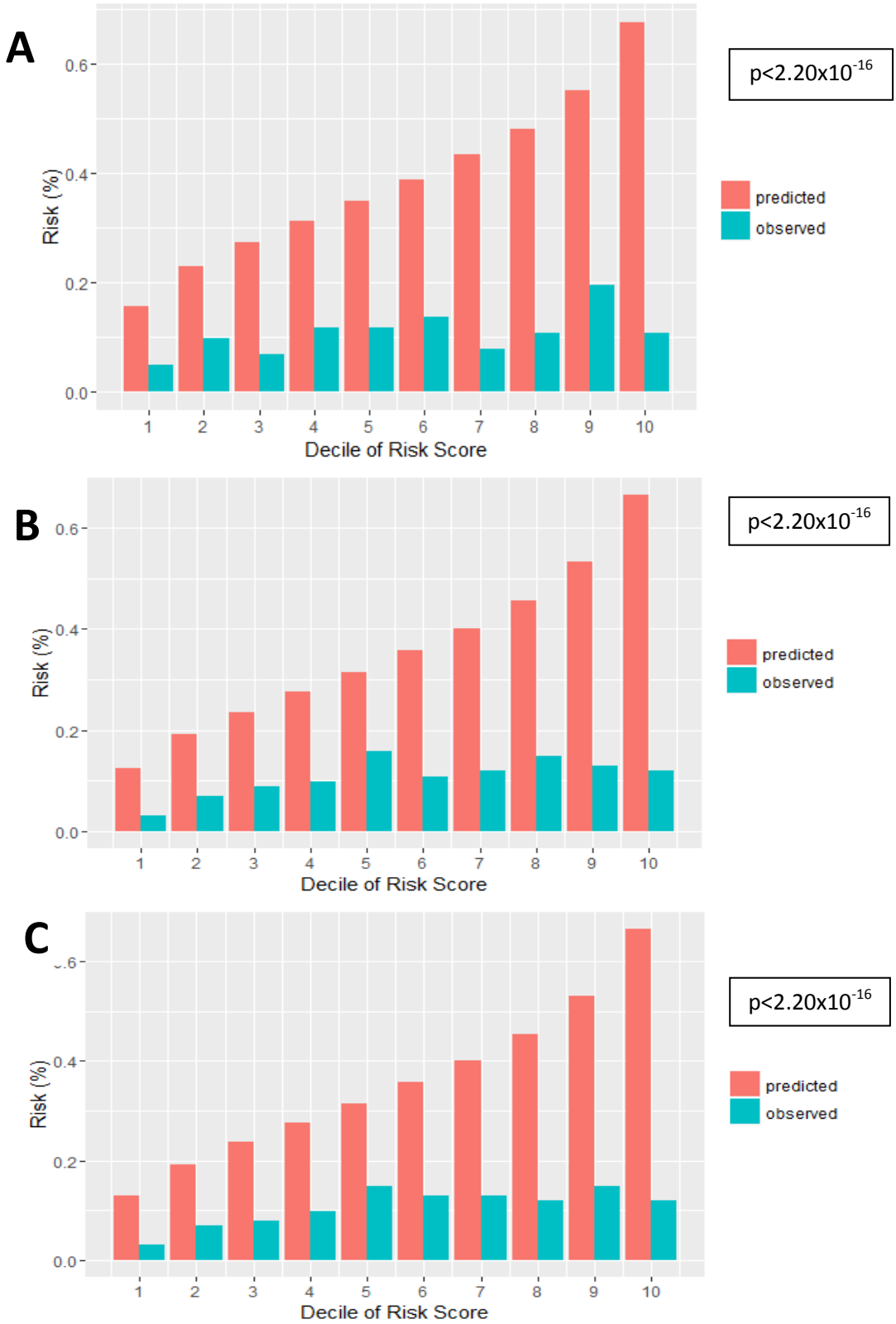
Trait	6 SNP GS - Unweighted		6 SNP GS - Weighted	
	Beta-coefficient (se)	p-value	Beta-coefficient (se)	p-value
BMI (kg/m^2)	-0.10 (0.08)	0.18	-0.48 (0.35)	0.20
Triglycerides* (mmol/l)	0.01 (0.01)	0.42	0.04 (0.06)	0.53
TC** (mmol/l)	0.03 (0.03)	0.29	0.06 (0.07)	0.37
HDL-cholesterol (mmol/l)	0.005 (0.005)	0.32	0.04 (0.02)	0.10
LDL-cholesterol** (mmol/l)	0.02 (0.04)	0.60	-0.01 (0.20)	0.96
Systolic blood pressure (mmHg)	0.07 (0.28)	0.81	0.48 (1.29)	0.75
Diastolic blood pressure (mmHg)	-0.04 (0.15)	0.82	-0.006 (0.71)	0.99
Fasting glucose* (mmol/l)	0.007 (0.005)	0.13	0.03 (0.02)	0.13
Insulin* (µIU/ml)	-0.02 (0.03)	0.52	-0.05 (0.30)	0.72

For each variable linear regression was performed in individual UCLEB studies and the results meta-analysed using a fixed-effects model unless otherwise stated. Beta-coefficient per unit increase and standard error are shown for each trait. *Variable log transformed. **Random effects (DerSimonian Laird method) used in this meta-analysis. TC=total cholesterol. GS=gene score.

4.2.3.1 Addition of CHD in T2D GS to CRF score

Whether addition of the CHD in T2D GSs to a CRF score provides any improvement compared to the performance of a CRF risk score alone was also assessed. Complete QRISK2, GS and ten-year CVD follow-up data was available for 1009 UCLEB participants with T2D (908 no CVD/101 CVD) taken from seven of the UCLEB studies (BWHHS, CAPS, EAS, ELSA, ET2DS, MRC1946 and WHII). To calculate the population weighted 5 SNP GS and 6 SNP GS, the effect sizes were taken from the source publications (Table 54) and the allele frequencies from the EUR group of the 1000 genomes project phase 1. As shown in Figure 20, QRISK2 was very poorly calibrated in the UCLEB participants with T2D ($p<2.20x10^{-16}$). The number of CVD events was similar in each decile of QRISK2 score, a very poor predictive performance. Unsurprisingly addition of the CHD in T2D GSs to QRISK2 did not improve calibration (both 6 SNP and 5 SNP GSs $p<2.20x10^{-16}$). There was no difference in discrimination between QRISK2 and the combined QRISK2 plus weighted GS score for either CHD in T2D GS (6 SNP GS AUROC 0.57 v 0.58 p=0.30, 5 SNP GS AUROC 0.57 v 0.59 p=0.12).

Figure 20: Observed CHD event rate in UCLEB T2D participants compared to the predicted event rate determined by A) QRISK score alone; B) QRISK plus 6 SNP GS C) QRISK plus 5 SNP GS, presented by decile of risk score.



Rates were compared using the Hosmer-Lemeshow test. R packages "ggplot2" (Wickham 2009), "PredictABEL"(Kundu, Aulchenko et al. 2011; Kundu, Aulchenko et al. 2014) and "ResourceSelection" (Lele, Keim et al. 2014). However, p-values were calculated separately using ten degrees of freedom, rather than with the eight calculated with the R packages.

4.2.3.2 CHD in T2D gene scores in those without T2D

The performance of the CHD in T2D GSs was assessed in the non-T2D population, using the NPHSII data set. Complete genotyping and CHD follow-up data was available for 2032 participants (1855 no CHD/177 CHD) for the 6 SNP GS and for 2260 participants (2072 no CHD/188 CHD) for the 5 SNP GS. However, as shown in Table **60**, there was no difference in either gene score (both weighted and unweighted) between those who did and those who did not go on to develop CHD over the ten-year follow-up period. Neither score was found to be associated with CHD (all p>0.05).

Table 60: CHD in T2D GSs in NPHSII

Score	No CHD	CHD	p-value
CHD in T2D 6 SNP GS - Unweighted	7.64 (5.18) n=1855	7.54 (1.37) n=177	0.39
CHD in T2D 6 SNP GS - Weighted	1.65 (0.34) n=1855	2.17 (0.19) n=177	0.27
CHD in T2D 5 SNP GS - Unweighted	12.80 (1.68) n=2072	13.47 (1.49) n=188	0.40
CHD in T2D 5 SNP GS - Weighted	1.23 (0.24) n=2072	1.21 (0.19) n=188	0.34

The GSs were compared using Welch's t-test. GS=gene score.

4.2.3.3 Updated 19 SNP GS in those with T2D

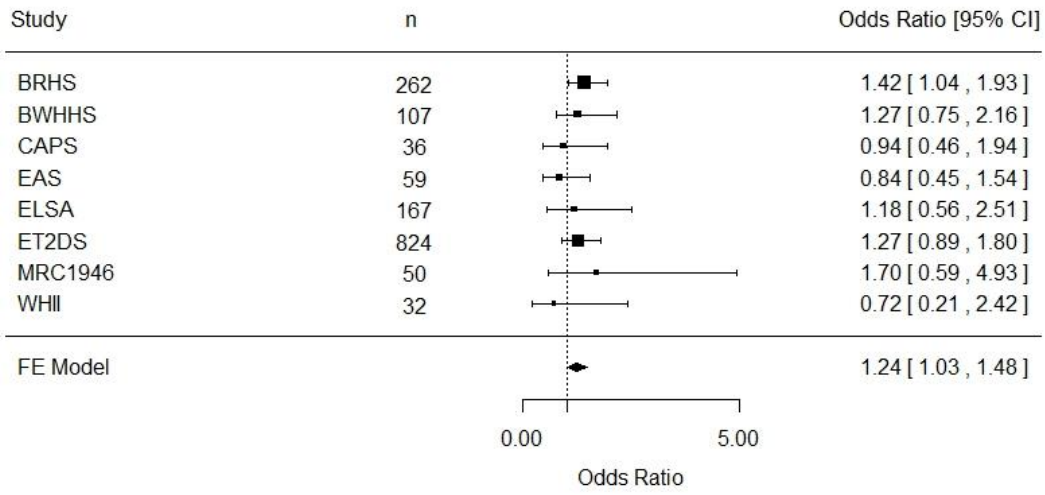
The performance of the updated 19 SNP GS developed in Chapter 3.2.1 was also assesed in the UCLEB participants with T2D. As in the non-T2D population, the GS was associated with CHD (Table 61; Figure 21). Higher unweighted GS values were associated with an increased risk of CHD, for each risk allele OR=1.10 (95% CIs 1.01-1.19). The weighted GS was also associated with CHD in T2D with a similar effect size to the unweighted score (per SD of GS unweighted OR=1.24 (95% CIs 1.03-1.48); weighted OR=1.22 (95% CIs 1.02-1.46, both unadjusted)). This is slightly lower than the effect size observed between the updated 19 SNP GSs in NPSHII (unweighted OR=1.38 (95% CIs 1.13-1.68); weighted NPHSII OR=1.47 (95% CIs 1.20-1.80), both adjusted for age). There was no difference in the discrimiantory ability of the 19 SNP GS compared to the CHD in T2D GSs (no difference in AUROC curve, all p>0.05). The impact of the adding the 19 SNP GS to QRISK2 was not assessed due to the poor calbration of QRISK2 in the UCLEB T2D participants.

Table 61: Mean updated 19 SNP GS in UCLEB participants who did and did not develop CHD during follow-up

Study		No CHD	CHD	p-value
BRHS	Weighted GS (sd)	2.02 (0.24)	2.11 (0.22)	4.0×10^{-3}
	Unweighted GS (sd)	15.76 (2.03)	16.39 (2.04)	0.03
BWHHS	Weighted GS (sd)	2.03 (0.23)	2.04 (0.15)	0.73
	Unweighted GS (sd)	15.78 (2.51)	16.46 (2.63)	0.40
CAPS	Weighted GS (sd)	2.05 (0.30)	2.12 (0.20)	0.41
	Unweighted GS (sd)	17.55 (2.28)	17.44 (1.97)	0.88
EAS	Weighted GS (sd)	2.13 (0.26)	2.07 (0.14)	0.31
	Unweighted GS (sd)	16.96 (2.49)	16.54 (1.81)	0.51
ELSA	Weighted GS (sd)	2.06 (0.24)	2.07 (0.24)	0.93
	Unweighted GS (sd)	15.91 (2.26)	16.29 (2.43)	0.70
ET2DS	Weighted GS (sd)	2.05 (0.25)	2.07 (0.21)	0.71
	Unweighted GS (sd)	16.41 (2.31)	16.97 (2.30)	0.19
MRC1946	Weighted GS (sd)	2.07 (0.23)	2.09 (0.14)	0.62
	Unweighted GS (sd)	16.44 (2.13)	17.40 (1.34)	0.21
WHII	Weighted GS (sd)	2.09 (0.26)	2.20 (0.15)	0.34
	Unweighted GS (sd)	17.07 (2.39)	16.33 (0.58)	0.21
Combined	Weighted GS (sd)			0.03
	Unweighted GS (sd)			0.03

Mean weighted and unweighted 6 SNP GS in UCLEB T2D participants who did and did not go on to develop CHD. The mean (standard deviation) for each study is shown individually. Mean gene score between the CHD and no CHD groups were compared using Welch's t-tests in individual studies and by ANOVA (with study as a factor) for the combined analysis. GS=gene score.

Figure 21: Association between updated 19 SNP GS and CHD in UCLEB participants with T2D



The unadjusted effect size (95% CI) determined in each study, per standard deviation of GS, is shown along with the combined effect size. The meta-analysis was performed using the R package "metafor" (Viechtbauer 2010). A FE model was used as there was no evidence of heterogeneity between the studies (I^2=0, p=0.78). GS=gene score. CI=confidence interval. FE=fixed effects.

4.2.4 Functional analysis of CHD in T2D risk variant rs10911021

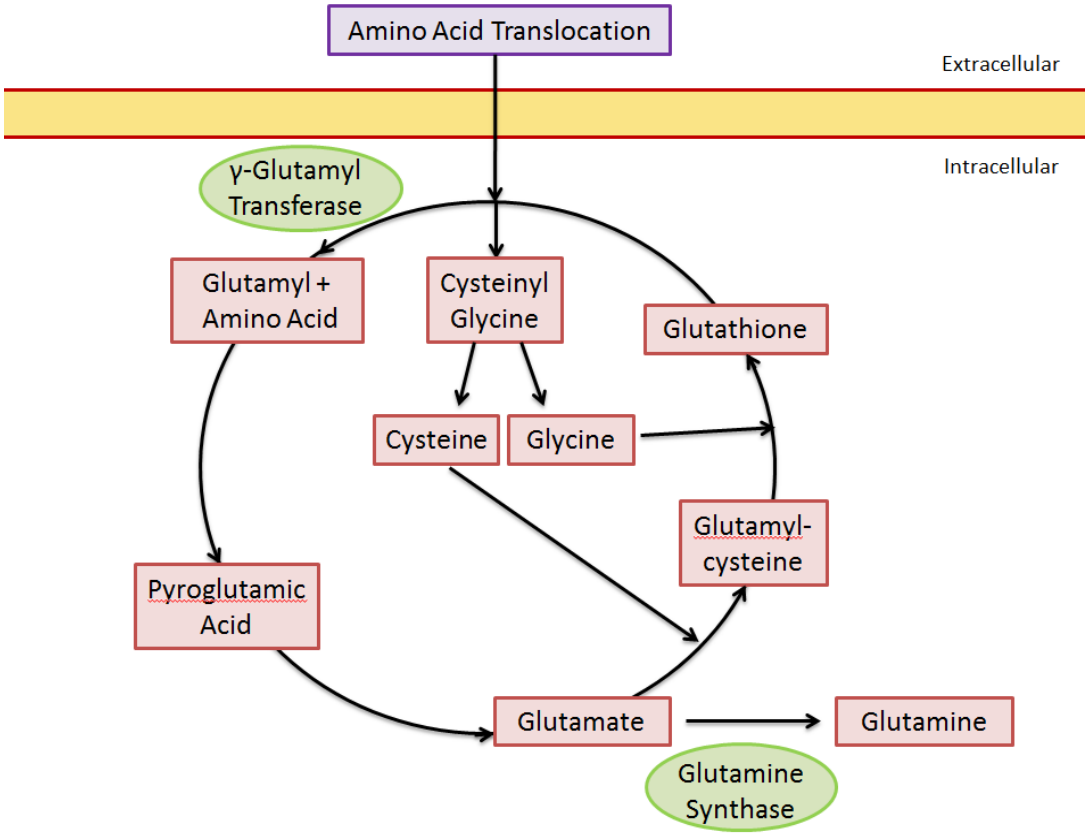
4.2.4.1 rs10911021 and CHD

Only one of the variants found to be associated with CHD in T2D, rs10911021 on chromosome 1, has not previously been identified as a CHD risk locus in the general population (Qi, Qi et al. 2013). The minor allele was found to be protective with an OR of 0.74 (95% CI 0.66-0.82, 1517 CHD cases, 2671 controls). Subsequently, the association was replicated in the Look AHEAD study (2016) and the SNP was found to be associated with all-cause mortality in the same T2D cohorts used for the original GWAS (Prudente, Shah et al. 2015).

The closest downstream gene to rs10911021 is *GLUL* which encodes glutamine synthase, an enzyme which catalyses the conversion of glutamate to glutamine. The authors of the original study also observed that endothelial cells homozygous for the risk allele had 32 % lower expression of *GLUL* compared to those homozygous for the protective allele. Moreover, while no association between rs10911021 and glutamate or glutamine was observed, individuals homozygous for the risk allele were found to have a lower pyroglutamic acid/glutamate ratio compared to those homozygous for the protective allele. Both metabolites are part of the ɣ-glutamyl cycle (Figure 22) which is involved in amino acid uptake and in the homeostasis of the anti-oxidant glutathione (Meister 1973). Thus, the authors hypothesised that the presence of the risk allele may impair the ɣ-glutamyl cycle resulting in a lesser availability of glutathione. Intracellular glutathione is known to be lower in those with T2D (Yoshida, Hirokawa et al. 1995).

The CHD in T2D risk locus falls close to a locus robustly associated with HDL-cholesterol levels (lead SNP rs1689800) (Teslovich, Musunuru et al. 2010). The close proximity between the SNPs raises the possibility that this risk locus may be involved in HDL metabolism. However, the degree of LD between the lead SNPs rs10911021 and rs1689800 is low (r^2=0.03 and D'=0.22, calculated from the CEU group of 1000 Genomes pilot). Moreover, rs10911021 was not found to associated with HDL-cholesterol levels in either the general population (Global Lipids Genetics Consortium data (p=0.05) (Willer, Schmidt et al. 2013)) or in overweight/obese individuals with T2D (Look AHEAD study (2016)).

Figure 22: Simplified diagram of the ɣ-glutamyl cycle



Other enzymes involved in the cycle and some metabolites have been omitted for clarity.

4.2.4.2 rs10911021 and CHD in UCLEB

The relationship between rs10911021 and CHD risk in those with T2D was investigated using data from the UCLEB consortium. The T2D group included only those with prevalent diabetes (either self-reported or clinically confirmed as described in (Shah, Engmann et al. 2013)). The basic characteristics and selected T2D-CHD risk factors of the UCLEB participants are presented in Table 62, separated by T2D status. As would be expected those with T2D had higher BMI, triglycerides, blood pressure, fasting glucose, glycated haemoglobin, insulin, whereas those without T2D had higher cholesterol (both LDL-cholesterol and HDL-cholesterol) and a greater proportion were male.

Table 62: Baseline characteristics for UCLEB participants separated by T2D status

Trait	Non-T2D participants		T2D-participants		p-value
	n	Mean (SD)	n	Mean (SD)	
Age (years)	13015	61.1 (6.0)	1803	61.3 (8.1)	0.32
Sex (percentage male)	8068	62.00 %	1053	58.4%	0.003
BMI (kg/m^2)	12803	26.7 (4.3)	1747	28.6 (5.80)	1.346×10^{-36}
Triglycerides* (mmol/l)	12022	0.43 (0.55)	1563	0.67 (0.75)	8.461×10^{-33}
TC (mmol/l)	12736	6.28 (1.24)	1784	6.04 (1.65)	4.484×10^{-8}
HDL-cholesterol (mmol/l)	12493	1.42 (0.38)	1757	1.25 (0.51)	2.114×10^{-34}
LDL-cholesterol (mmol/l)	12385	4.00 (1.07)	1607	3.62 (1.43)	1.573×10^{-21}
Systolic blood pressure (mmHg)	12739	139.90 (22.80)	1783	148.00 (30.60)	1.650×10^{-23}
Diastolic blood pressure (mmHg)	12722	81.70 (12.90)	1782	84.40 (17.30)	3.716×10^{-9}
Fasting glucose* (mmol/l)	12741	1.69 (0.15)	1670	1.98 (0.19)	2.54×10^{-303}
Insulin* (μIU/ml)	7732	1.89 (0.62)	456	2.50 (0.66)	1.686×10^{-80}
Glycated haemoglobin (%)	8711	5.37 (0.65)	1807	6.80 (0.98)	8.14×10^{-265}

Mean and standard deviation, where appropriate, are shown. P-values were determined using a chi-squared test for sex and with regression model (adjusted for age and sex) for the other variables. *Variables were log transformed.TC=total cholesterol.

Eight studies had data on diabetes status, rs10911021 imputation and CHD outcome. As shown in Table 63, no association was observed between those without T2D and rs10911021 OR=1.00 (95%CIs 0.92-1.10). Whereas in those with T2D, while the association between the SNP and CHD was not statistically significant it was directionally similar OR=0.80 (95%CIs 0.60-1.06, p=0.13) to the published data. The results from UCLEB were then meta-analysed with the published data using both FE and RE models. Both meta-analyses gave a strongly statistically significant p-value and a similar effect size, (FE: OR=0.74, 95% CIs 0.68-0.82 $p=8.22 \times 10^{-10}$ and RE: OR=0.75 95% CIs 0.67-0.84, $p=1.61 \times 10^{-6}$). Heterogeneity between the studies was low (I^2=18%) The sensitivity analysis is shown in Table 64 and Table 65.

Table 63: Relationship between the minor allele of rs10911021 and CHD for UCLEB participants with and without T2D

No T2D	BRHS	BWHHS	CAPS	EAS	ELSA	ET2DS	MRC1946	WHII	Combined
MAF no CHD	0.30 (1544)	0.32 (1528)	0.31 (1022)	0.31 (553)	0.30 (1426)	-	0.32 (2294)	0.31 (2851)	0.31 (8665)
MAF CHD	0.30 (378)	0.31 (285)	0.28 (239)	0.29 (132)	0.29 (114)	-	0.31 (65)	0.35 (161)	0.30 (1677)
OR (95% CI)	1.02 (0.85-1.22)	1.01 (0.79-1.28)	0.82 (0.65-1.04)	0.90 (0.67-1.23)	1.10 (0.81-1.49)	-	1.00 (0.68-1.47)	1.23 (0.97-1.56)	1.00 (0.92-1.10)
p value	0.81	0.94	0.10	0.64	0.54	-	0.43	0.09	0.93
T2D									
MAF no CHD	0.31 (190)	0.34 (94)	0.30 (20)	0.23 (46)	0.32 (160)	0.30 (793)	0.28 (45)	0.31 (29)	0.30 (1377)
MAF CHD	0.18 (72)	0.20 (13)	0.29 (16)	0.24 (13)	0.29 (7)	0.32 (31)	0.40 (5)	0.30 (3)	0.26 (160)
OR (95% CI)	0.44 (0.26-0.74)	0.48 (0.17-1.33)	1.43 (0.51-4.00)	1.05 (0.36-3.03)	0.85 (0.25-2.94)	1.35 (0.80-2.33)	1.69 (0.46-6.25)	1.01 (0.52-1.96)	0.80 (0.60-1.06)
p value	$2x10^{-3}$	0.16	0.49	0.95	0.80	0.26	0.43	0.87	0.13

Minor allele frequency (MAF) is shown separately for those who did and did not go on to develop CHD. Number of participants is shown in brackets. The odds ratio (OR) adjusted for sex for the association between rs10911021 and CHD in T2D is also shown with its 95% confidence intervals (95% CI).

Table 64: Sensitivity analysis for fixed-effects meta-analysis of the association between rs10911021 and CHD in T2D

Study Source	Study Left Out	OR	p-value
Qi, Qi et al (Qi, Qi et al. 2013)	NHS	0.75	$3.54x10^{-7}$
Qi, Qi et al.(Qi, Qi et al. 2013)	HPFS	0.76	$4.20x10^{-7}$
Qi, Qi et al.(Qi, Qi et al. 2013)	JHS	0.73	$5.33x10^{-9}$
Qi, Qi et al.(Qi, Qi et al. 2013)	GHS	0.75	$3.37x10^{-8}$
Qi, Qi et al.(Qi, Qi et al. 2013)	CS	0.74	$1.91x10^{-9}$
UCLEB	BRHS	0.76	$1.43x10^{-8}$
UCLEB	BWHHS	0.75	$1.58x10^{-9}$
UCLEB	CAPS	0.74	$4.69x10^{-10}$
UCLEB	EAS	0.74	$6.70x10^{-10}$
UCLEB	ELSA	0.74	$8.28x10^{-10}$
UCLEB	ET2DS	0.73	$1.22x10^{-10}$
UCLEB	MRC1946	0.74	$5.19x10^{-10}$
UCLEB	WHII	0.74	$7.84x10^{-10}$

OR for the effect relating to the minor allele is shown along with the p-value, when the meta-analysis was performed without individual studies. The meta-analyses were performed using the R "metafor" package (Viechtbauer 2010). NHS=Nurses' Health Study, HPFS=Health Professionals Follow-up Study, JHS=Joslin Heart Study, GHS=Gargano Heart Study, CS=Catanzaro Study. OR=odds ratio.

Table 65: Sensitivity analysis for random-effects meta-analysis of the association between rs10911021 and CHD in T2D

Study Source	Study Left Out	OR	p-value	Heterogeneity (I^2 (%))
Qi, Qi et al. (Qi, Qi et al. 2013)	NHS	0.76	$3.67x10^{-4}$	24.57
Qi, Qi et al.(Qi, Qi et al. 2013)	HPFS	0.77	$1.39x10^{-4}$	17.83
Qi, Qi et al.(Qi, Qi et al. 2013)	JHS	0.74	$3.48x10^{-5}$	21.88
Qi, Qi et al.(Qi, Qi et al. 2013)	GHS	0.76	$1.51x10^{-4}$	24.39
Qi, Qi et al.(Qi, Qi et al. 2013)	CS	0.75	$2.03x10^{-5}$	24.19
UCLEB	BRHS	0.76	$1.43x10^{-8}$	0
UCLEB	BWHHS	0.76	$5.08x10^{-6}$	20.79
UCLEB	CAPS	0.74	$3.42x10^{-7}$	15.73
UCLEB	EAS	0.75	$3.62x10^{-6}$	22.55
UCLEB	ELSA	0.75	$7.29x10^{-6}$	24.47
UCLEB	ET2DS	0.73	$1.22x10^{-10}$	0
UCLEB	MRC1946	0.75	$3.98x10^{-7}$	15.98
UCLEB	WHII	0.75	$6.05x10^{-6}$	24.10

OR for the effect relating to the minor allele is shown along with the p-value, when the meta-analysis was performed without individual studies. The meta-analyses were performed using the R "metafor" package (Viechtbauer 2010). NHS=Nurses' Health Study, HPFS=Health Professionals Follow-up Study, JHS=Joslin Heart Study, GHS=Gargano Heart Study, CS=Catanzaro Study. OR=odds ratio.

4.2.4.3 rs10911021 and the ɣ-glutamyl cycle in T2D

The relationship between rs10911021 and the ɣ-glutamyl cycle was investigated by assessing whether the SNP was associated with levels of amino acids which are taken up into cells by it. Data on the levels of nine amino acids, determined using an NMR-metabolomics platform, was available for four UCLEB studies: BWHHS, ET2DS, MRC1946 and WHII. These included glutathione constituent glycine and glutamine (the product of the reaction catalysed by glutamine synthase). there was no association between rs10911021 and any of the amino acids measured in either those with T2D ($p>0.05$) as shown in Table 66, nor in those without T2D ($p>0.05$, data not shown).

Table 66: Relationship between rs10911021 and NMR-determined amino acid levels in those with T2D

Trait	Beta-coefficient (se)	p-value
Alanine (mmol/l)	-0.007 (0.07)	0.94
Glutamine (mmol/l)	0.005 (0.08)	0.94
Glycine (mmol/l)	0.003 (0.07)	0.97
Histidine (mmol/l)	0.03 (0.07)	0.66
Isoleucine (mmol/l)	0.02 (0.07)	0.74
Leucine (mmol/l)	-0.005 (0.07)	0.94
Valine (mmol/l)	0.06 (0.07)	0.44
Phenylalanine (mmol/l)	0.04 (0.07)	0.58
Tyrosine (mmol/l)	-0.03 (0.07)	0.65

Beta-effects corresponding to the minor allele from the linear regression are shown – adjusted for lipid lowering medication, along with the standard errors (se). Prior to analysis the metabolomics measures were adjusted for age, age^2 and sex and inverse rank transformed.

4.2.4.4 rs10911021 and T2D-CHD risk factors

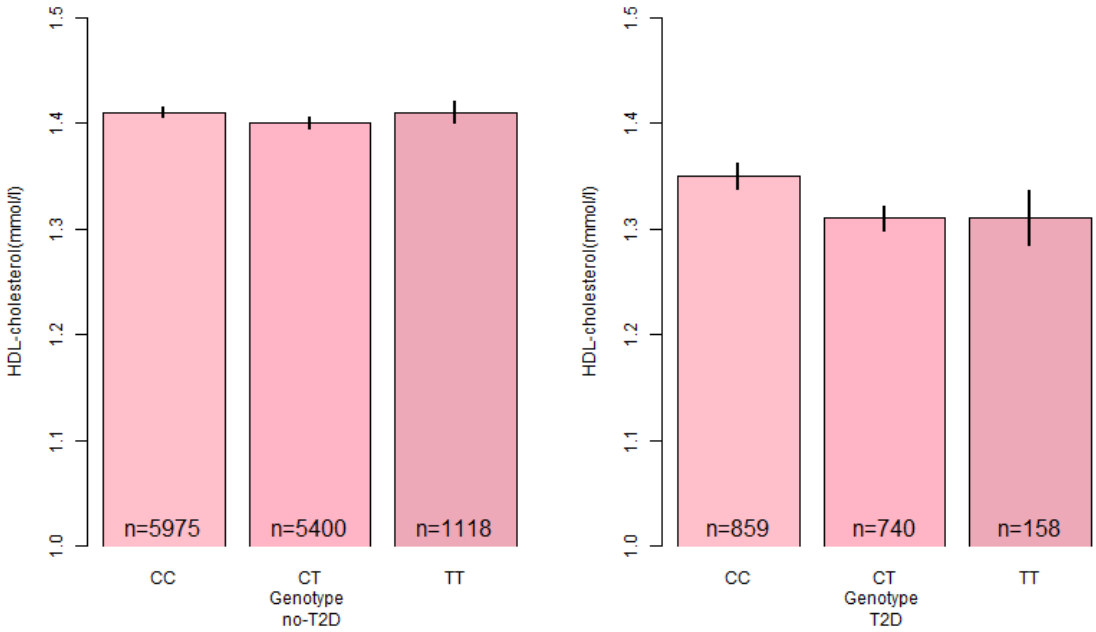
The relationship between rs10911021 and T2D-CHD CRFs was then investigated, again using the UCLEB data. The SNP was not associated with any T2D-CHD CRFs in the no-T2D group (all p>0.05; Table 67). In the T2D participants, rs10911021 was associated with HDL-cholesterol levels ($p=5x10^{-4}$, Table 67). Unexpectedly the minor allele, which was found to be "protective" for CHD, was associated with 0.034 mmol/l lower HDL-cholesterol levels. Mean HDL-cholesterol levels by genotype (adjusted for sex and study) are shown in Figure 23. The effect appears to be recessive.

Table 67: Relationship between rs10911021 and T2D-CHD risk factors in UCLEB in those with and without T2D

Trait	Non-T2D UCLEB participants			T2D UCLEB participants		
	n	Beta-coefficient (se)	p-value	n	Beta-coefficient (se)	p-value
BMI (kg/m^2)	12803	-0.032 (0.055)	0.56	1747	-0.055 (0.178)	0.76
Triglycerides* (mmol/l)	12022	0.007 (0.007)	0.34	1563	0.030 (0.020)	0.87
TC (mmol/l)	12736	-0.011 (0.016)	0.25	1784	0.026 (0.043)	0.54
HDL-cholesterol (mmol/l)	12493	-0.001 (0.005)	0.86	1757	-0.034 (0.012)	$5x10^{-4}$
LDL-cholesterol (mmol/l)	12385	-0.018 (0.014)	0.21	1607	0.070 (0.037)	0.06
Systolic blood pressure (mmHg)	12739	0.045 (0.298)	0.88	1783	0.056 (0.794)	0.94
Diastolic blood pressure (mmHg)	12722	0.052 (0.170)	0.76	1782	-0.510 (0.432)	0.24
Fasting glucose* (mmol/l)	12740	0.001 (0.002)	0.61	1670	-0.011 (0.009)	0.21
Insulin[a] (μIU/ml)	7732	-0.019 (0.011)	0.09	456	0.039 (0.063)	0.53
Glycated haemoglobin (%)	8711	-0.003 (0.008)	0.73	1317	0.032 (0.040)	0.42

Mean and standard error (se) for each trait in those with and without T2D is shown. The beta effect relating to the minor allele is shown. *Variables were log transformed. TC=total cholesterol.

Figure 23: Mean HDL-cholesterol by rs10911021 genotype in UCLEB participants with and without T2D

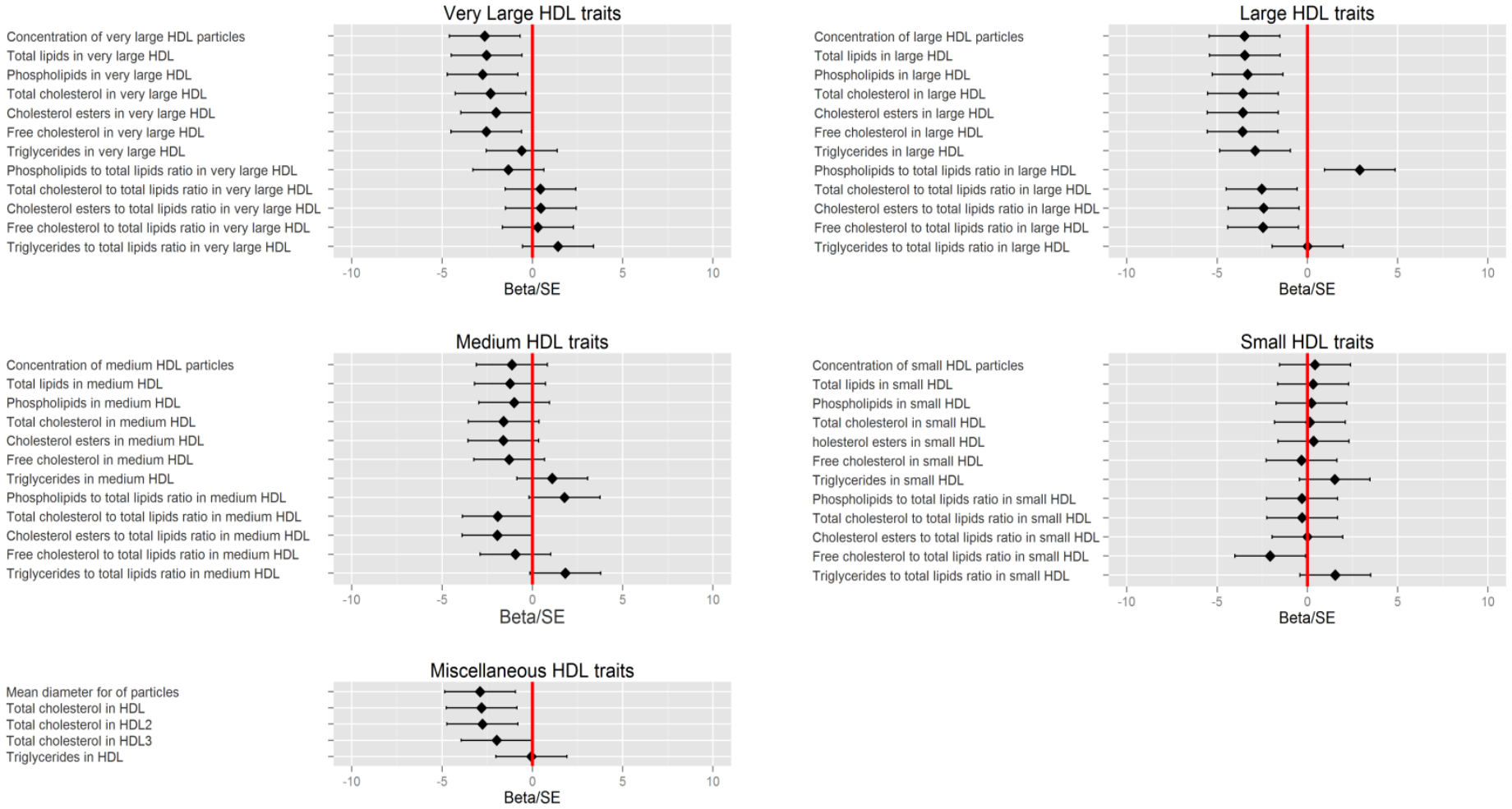


The means were adjusted for sex and study. The error bars represent standard error. C is the common, "CHD risk" allele.

Four UCLEB studies had HDL-cholesterol metabolomics data available. The NMR-based platform can separate HDL by size into four subclasses (very large, large, medium and small) with twelve separate traits pertaining to particle composition available for each one. Overall mean HDL particle diameter, concentrations of HDL-cholesterol and the sub-fractions HDL2 and HDL3 and the triglyceride content of HDL particles were also measured. A summary of the results in those with T2D is shown in Figure 24. As shown in Table 68, in those with T2D, six metabolic measures, all relating to large HDL particles showed an association with rs10911021 with an FDR adjusted p-value $p<0.05$. A further 16 HDL metabolic measures had an unadjusted p-value below $p=0.05$ (Table 69). By contrast, no association between rs10911021 and any of the HDL measurements in non-T2D participants was observed (unadjusted $p>0.05$). Figure 25 is a representative forest plot of large HDL particle concentration showing a consistent lower level associated with the minor allele of rs10911021 in those with T2D in the four studies.

Given the close proximity of a GWAS hit (lead SNP rs1689800) for HDL-cholesterol levels to the CHD in T2D locus, it was hypothesised that the association observed between rs10911021 and HDL traits could involve this locus. However, conditional analysis was performed and found similar results as in the unadjusted model (Table 68).

Figure 24: Relationship between HDL metabolomic traits and the minor allele of rs10911021 in those with T2D



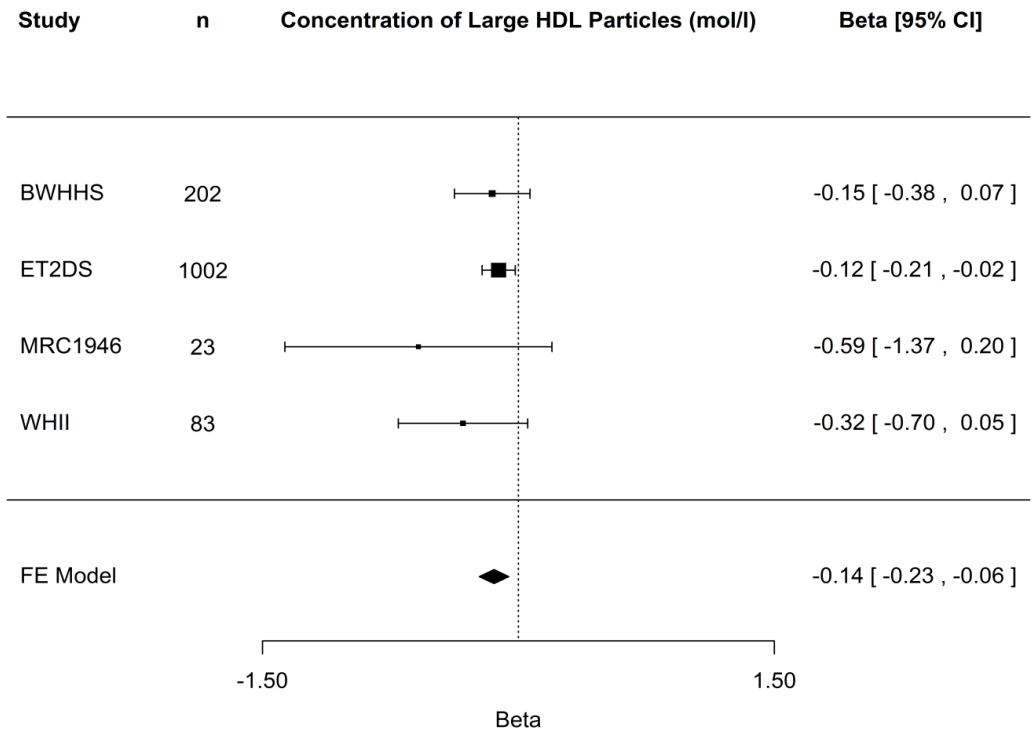
SE=standard error. Beta=beta-coefficient.

Table 68: Metabolomic HDL traits associated with rs10911021 in those with T2D

Trait	Non-T2D Participants					T2D participants						
	n	Beta-coefficient (se)	p-value	FDR adjusted p-value	Heterogeneity (I^2 (%))	n	Beta-coefficient (se)	p-value	FDR adjusted p-value	Heterogeneity (I^2 (%))	p-value for conditional analysis with rs1689800	FDR adjusted p-value for conditional analysis with rs1689800
Concentration of large HDL particles (mol/l)	5221	0.01 (0.02)	0.59	1	0	1310	-0.15 (0.04)	0.0005	0.03	0	0.001	0.07
Total lipids in large HDL (mmol/l)	5229	0.01 (0.02)	0.62	1	0	1310	-0.15 (0.04)	0.0005	0.03	0	0.001	0.07
Phospholipids in large HDL (mmol/l)	5223	0.01 (0.02)	0.59	1	0	1310	-0.14 (0.04)	0.0009	0.03	0	0.002	0.09
Total cholesterol in large HDL (mmol/l)	5223	0.008 (0.02)	0.71	1	0	1310	-0.15 (0.04)	0.0004	0.04	0	0.001	0.07
Cholesterol esters in large HDL (mmol/l)	5221	0.009 (0.02)	0.67	1	0	1310	-0.15 (0.04)	0.0004	0.03	0	0.001	0.07
Free cholesterol in large HDL (mmol/l)	5221	0.005 (0.02)	0.83	1	0	1310	-0.16 (0.04)	0.0003	0.03	0	0.0009	0.07

Beta-effects corresponding to the minor allele from the linear regression are shown – adjusted for lipid lowering medication, along with the standard errors (se). Prior to analysis the metabolomics measures were adjusted for age, age^2 and sex and inverse rank transformed. FDR analysis was performed using the Benjamini-Hochberg-Yekutieli method (Benjamini and Yekutieli 2001). FDR=false discovery rate

Figure 25: Forest plot for the meta-analysis of large HDL particle concentration and the minor allele rs10911021 in those with T2D



The meta-analysis was performed using the "metaphor" R package. Beta=beta-coefficient. CI=confidence interval. FE=fixed effects.

Table 69 :Metabolomic HDL traits which did not show an association with rs10911021 in those with and without T2D

Trait	Non-T2D participants					T2D participants						
	n	Beta-coefficient (se)	p-value	FDR adjusted p-value	Heterogeneity (I^2 (%))	n	Beta-coefficient (se)	p-value	FDR adjusted p-value	Heterogeneity (I^2 (%))	p-value for conditional analysis with rs1689800	FDR adjusted p-value for conditional analysis with rs1689800
Triglycerides in very large HDL (mmol/l)	5221	-0.006 (0.02)	0.77	1	0	1310	-0.03 (0.04)	0.56	1	0	0.69	1
Concentration of medium HDL particles (mol/l)	5229	0.03 (0.02)	0.17	1	0	1310	-0.05 (0.04)	0.26	1	0	0.42	1
Total lipids in medium HDL (mmol/l)	5224	0.03 (0.02)	0.16	1	0	1310	-0.05 (0.04)	0.21	1	0	0.36	1
Phospholipids in medium HDL (mmol/l)	5224	0.03 (0.02)	0.17	1	0	1310	-0.04 (0.04)	0.31	1	0	0.49	1
Total cholesterol in medium HDL (mmol/l)	5222	0.03 (0.02)	0.19	1	0	1310	-0.07 (0.04)	0.11	0.96	0.94	0.20	1
Cholesterol esters in medium HDL (mmol/l)	5224	0.03 (0.02)	0.21	1	0	1310	-0.07 (0.04)	0.11	0.96	0	0.19	1
Free	5224	0.03	0.17	1	0	1310	-0.06	0.20	1	0	0.31	1

cholesterol in medium HDL (mmol/l)		(0.02)					(0.04)					
Triglycerides in medium HDL (mmol/l)	5221	0.02 (0.02)	0.38	1	0	1310	0.05 (0.04)	0.27	1	0	0.21	1
Concentration of small HDL particles (mol/l)	5229	0.02 (0.02)	0.29	1	0	1310	0.02 (0.04)	0.67	1	0	0.62	1
Total lipids in small HDL (mmol/l)	5222	0.02 (0.02)	0.38	1	0	1310	0.01 (0.04)	0.75	1	0	0.71	1
Phospholipids in small HDL (mmol/l)	5222	0.03 (0.02)	0.11	1	0	1310	0.01 (0.04)	0.82	1	0	0.67	1
Total cholesterol in small HDL (mmol/l)	5221	-0.004 (0.02)	0.83	1	44.90	1310	0.01 (0.04)	0.89	1	0	0.98	1
Cholesterol esters in small HDL (mmol/l)	5221	-0.01 (0.02)	0.58	1	59.27	1310	0.01 (0.04)	0.73	1	0	0.91	1
Free cholesterol in small HDL (mmol/l)	5221	0.02 (0.02)	0.43	1	0	1310	-0.01 (0.04)	0.74	1	0	0.85	1
Triglycerides in very small HDL (mmol/l)	5221	-0.006 (0.02)	0.78	1	0	1310	0.07 (0.04)	0.13	1	0	0.17	1
Phospholipids to total lipids	5143	0.0004 (0.02)	0.99	1	0	958	-0.07 (0.05)	0.18	1	16.56	0.31	1

ratio in very large HDL (%)												
Total cholesterol to total lipids ratio in very large HDL (%)	5142	0.008 (0.02)	0.72	1	0	958	0.02 (0.05)	0.66	1	0	0.90	
Cholesterol esters to total lipids ration in very large HDL (%)	5142	0.0004 (0.02)	0.98	1	0	958	0.02 (0.05)	0.64	1	52.42	0.83	
Free cholesterol to total lipids ratio in very large HDL (%)	5142	0.02 (0.02)	0.41	1	0	958	0.02 (0.05)	0.76	1	66.02	0.84	1
Triglycerides to total lipids ratio in very large HDL (%)	5142	-0.01 (0.02)	0.51	1	15.40	958	0.02 (0.05)	0.16	1	0	0.13	1
Triglycerides to total lipids ratio in large HDL (%)	5140	0.003 (0.02)	0.88	1	0	1056	0.0005 (0.05)	0.99	1	8.14	0.89	1
Phospholipids to total lipids ratio in medium HDL (%)	5221	-0.01 (0.02)	0.50	1	53.87	1309	0.08 (0.04)	0.08	0.70	0	0.12	1
Total cholesterol to total lipids	5218	0.01 (0.02)	0.63	1	21.40	1309	-0.08 (0.04)	0.05	0.55	0	0.09	0.90

ratio in medium HDL (%)												
Cholesterol esters to total lipids ration in medium HDL (%)	5221	0.009 (0.02)	0.69	1	24.66	1309	-0.08 (0.04)	0.05	0.55	0	0.08	0.84
Free cholesterol to total lipids ratio in medium HDL (%)	5221	0.01 (0.02)	0.50	1	0	1309	-0.04 (0.04)	0.34	1	0	0.42	1
Triglycerides to total lipids ratio in medium HDL (%)	5218	-0.0009 (0.02)	0.97	1	0	1309	0.08 (0.04)	0.07	0.66	0	0.10	0.94
Phospholipids to total lipids ratio in small HDL (%)	5218	0.04 (0.02)	0.06	1	60.62	1310	-0.0 (0.04)	0.77	1	0	0.92	1
Total cholesterol to total lipids ratio in small HDL (%)	5217	-0.03 (0.02)	0.13	1	66.43	1310	-0.01 (0.04)	0.77	1	0	0.55	1
Cholesterol esters to total lipids ration in small HDL (%)	5217	-0.03 (0.02)	0.15	1	60.49	1310	-0.00005 (0.04)	0.9996	1	0	0.74	1
Triglycerides	5217	-0.01	0.56	1	6.76	1310	0.07	0.12	1	3.23	0.18	1

to total lipids ratio in small HDL (%)		(0.02)					(0.04)					
Triglycerides in HDL (mmol/l)	5219	0.01 (0.02)	0.60	1	0	1310	-0.002 (0.04)	0.96	1	0	0.94	1

Beta-effects corresponding to the minor allele from the linear regression are shown – adjusted for lipid lowering medication, along with the standard errors (se). Prior to analysis the metabolomics measures were adjusted for age, age^2 and sex and inverse rank transformed. FDR analysis was performed using the Benjamini-Hochberg-Yekutieli method (Benjamini and Yekutieli 2001). FDR=false discovery rate.

4.3 Discussion

The association between a GS comprised of five risk SNPs and CHD in T2D observed by Qi, Parast et al. was replicated in the UCLEB data set. There was also evidence for an association between a 6 SNP GS (the five original SNPs plus a subsequently identified risk SNP rs10911021 (Qi, Qi et al. 2013)) and CHD in T2D. Furthermore, it was found that both the 5 SNP and 6 SNP GSs were not associated with CHD in NPHSII participants free of T2D at baseline. The general CHD 19 SNP GS was also associated with CHD in those with T2D indicating that it would be suitable to use this tool in the diabetic population. The effect size for each GS is similar and there was no difference in the AUROC between the scores. This suggests that a specific CHD in T2D GS based on current knowledge would not improve CHD risk prediction over and above a general CHD GS. However, a major limitation of this work is that no meaningful assessment of whether addition of the GSs to a CRF risk score gives improved performance compared to the CRF score alone could be performed. QRISK2 data was available but this score was found to be very poorly calibrated in the UCLEB T2D participants. Unsurprisingly given their relatively modest effect sizes, addition of the CHD in T2D GSs did not improve calibration. It is unclear why QRISK2 greatly overestimated CHD risk in this group. The developers of the QRISK score published a document detailing the validation of QRISK2 (2014 version as was calculated in UCLEB) in those with T2D on the QRISK2 website (http://www.qrisk.org/). Data from the QRESEARCH group comprising almost 80,000 individuals with T2D was used. While statistics pertaining to the calibration of QRISK2 were not given, it is clear from the calibration plot shown, that the model is much better fitting than observed in the UCLEB data set. However, external validation is required. Only once the GSs have been assessed in combination with CRFs risk score can firm conclusions about their potential utility be drawn.

Five of the six SNPs used in the CHD in T2D GSs are GWAS hits for CHD and four were confirmed as CHD risk SNPs in the CARDIoGRAMplusC4D meta-analysis (Deloukas, Kanoni et al. 2013). Therefore, it would be expected that a GSs composed of these SNPs would be associated with CHD in the general population. That no such association was observed in NPHSII could simply be because the sample was underpowered to detect the effect. Thus it appears that the risk alleles have a much greater impact on CHD risk in those with T2D compared to the general population. It is not surprising therefore, that the effect size SNP is larger for four or the five SNPs (for the 9p21 SNP it is the same) in those with T2D, although these effect sizes may be inflated as the number of participants in the studies

they were derived from was relatively small compared to the tens of thousands included in the CARDIoGRAMplusC4D meta-analysis.

As the functional mechanisms of CHD risk variants are elucidated, it will become possible to determine if the variants associated with CHD in both the general population and in T2D affect risk through the same mechanism or if the diabetic state leads to as yet unknown consequences. Two of the CHD in T2D risk variants (rs646776 close to *SORT1* and rs11206510 close to *PCSK9*) are known to be associated with LDL-cholesterol levels (Teslovich, Musunuru et al. 2010). This would explain the suggestive association between the 5 SNP GSs and LDL-cholesterol. It seems reasonable to speculate therefore that each risk allele may be associated with a greater increase in LDL-cholesterol in the diabetic state compared to the non-diabetic state. Thus, presence of these risk alleles may contribute to diabetic dyslipidaemia which is characterised by high triglyceride levels, a high concentration of small dense LDL particles and a low HDL-cholesterol concentration (Wu and Parhofer 2014). This may also be the case for the risk locus in *HNF1A*, where the lead SNP rs2259816 is in moderate LD with the lead SNP (r^2=0.48, taken from the 1000 Genomes phase 3 EUR panel) at a confirmed LDL-cholesterol associated locus (rs1169288) (Teslovich, Musunuru et al. 2010). Moreover, mutations in *HNF1A* are also the most common cause of Mendelian diabetes (Gardner and Tai 2012) and a GWAS hit for T2D is located in this gene (Voight, Scott et al. 2010). *HNF1A* encodes hepatocyte nuclear factor 1-alpha, which is involved in regulating the expression of many genes, particularly in the pancreas and liver (Courtois, Morgan et al. 1987), suggesting this gene could influence risk of both CHD and T2D through a number of different metabolic pathways. For the two remaining loci originally identified by (Qi, Parast et al. 2011) (lead SNPs: rs4977574 at the 9p21 locus and rs12526453 in *PHACTR1*) there is no obvious mechanism of action. The relationship between the 9p21 locus and CHD was discussed in 1.4.3.1. Notably a GWAS identified risk locus for T2D is also present on chromosome 9p21. In addition to CHD, the SNP in *PHACTR1* is associated with coronary artery calcification (Pechlivanis, Muhleisen et al. 2013) which is also known to be greater in those with T2D (Erbel, Lehmann et al. 2014). *PHACTR1* encodes PHACTR-1 which thought to be play a key role in endothelial cell function (Jarray, Allain et al. 2011) and angiogenesis (Allain, Jarray et al. 2012) but the biology of this protein is not well understood.

The identification of a CHD risk variant in those with T2D only (Qi, Qi et al. 2013) further suggests that the diabetic state itself is pro-atherogenic. The UCLEB data lacked the power to replicate this association (237 CHD cases and 2038 CHD controls would be required for 80% power to detect the same effect). However, a consistent protective but statistically insignificant association between the minor allele of rs10911021 and CHD in T2D was observed. The more modest effect size found here is not unexpected as the initial report of an association is likely to be inflated due to the "winner's curse"(Ioannidis 2008). Indeed, the association between CVD and rs10911021 observed in the Look AHEAD study had a smaller effect than the original report (2016), although the outcome definition was broader which may also partly account for this. When the data presented here were meta-analysed with the data from Qi, Parast et al. using a FE meta-analysis, the p-value was lower compared to the original findings indicating that our data support the original observation. A meta-analysis using a RE model was also performed although the p-value was higher than in the original study. However, sensitivity analysis (Table 64 and Table 65) shows that this is being driven by one study and as heterogeneity is relatively low between the studies a FE model is satisfactory.

The authors of the original study implicated impairment of the ɣ-glutamyl cycle and thus glutathione availability as the mechanism through which rs10911021 affects CHD risk. They observed that subjects homozygous for the risk allele of rs10911021 had a lower pyroglutamic acid to glutamate (substrate of *GLUL*) ratio, and that endothelial cells with this genotype had lower expression of the enzyme glutamine synthase (encoded by *GLUL*). Variants in the T2D/CHD in T2D risk gene *HNF1A* have been found to be associated with levels of ɣ-glutamyl transferase, another enzyme involved in the ɣ-glutamyl cycle (Yuan, Waterworth et al. 2008) (Figure 22), further implicating this pathway in the development of CHD in T2D. Neither pyroglutamic acid nor glutamate were measured by the NMR-metabolomics platform used in this study nor were cysteine and glutamate which are crucial to glutathione levels (Liu, Hyde et al. 2014). However, nine amino acids, including glycine (another constituent of glutathione (Liu, Hyde et al. 2014)) and glutamine, the product of the reaction catalysed by glutamine synthase, were measured but no association was observed in those with T2D. Therefore, a relationship between the ɣ-glutamyl cycle and rs10911021 cannot be discounted but if so the results herein indicate that it is not through limiting the availability of glycine or by inhibiting general amino acid translocation into the cell.

There was no association between rs10911021 and any of the classical CHD risk factors in those without T2D, while in those with T2D only an association with HDL-cholesterol was observed. This is contrary to the findings of the Look AHEAD study which found no association between the SNP and HDL-cholesterol in their overweight/obese T2D cohort. In UCLEB, the association remained when the analysis was restricted to those with BMI equal to or greater than 25 kg/m^2. Curiously, the protective minor allele was associated with lower HDL-cholesterol, which has long been associated with an increased risk of CHD. More in-depth analysis with the metabolomics data found an association between rs10911021 and six large-HDL traits, again only in those with T2D. There were also suggestive associations between the SNP and a further 16 HDL traits, mostly relating to large and very large HDL particles. These associations were found to be independent of the nearby HDL GWAS hit marked by rs1689800.

The relationship between HDL-cholesterol and CHD remains to be clarified. Mendelian randomisation studies have failed to find a relationship between genetically low HDL-cholesterol and CHD (Assimes, Holm et al. 2010; Holmes, Asselbergs et al. 2015) and HDL-cholesterol raising therapies have failed to improve cardiovascular outcome (Keene, Price et al. 2014). The failure to confirm a causal association between lower levels of HDL-cholesterol and CHD has moved the focus from HDL-cholesterol concentration towards HDL particle subclasses. While increased levels of small HDL particles have been associated with increased risk of CHD the converse is true of large HDL particles (Rosenson, Otvos et al. 2002; Morgan, Carey et al. 2004; Musunuru, Orho-Melander et al. 2009). In this analysis an association between the minor (previously identified as CHD "protective") allele and lower levels of large HDL particle traits including concentration and cholesterol content were observed, which is the opposite of what would be expected for a protective gene variant. A variant with a similar phenotype (HDL-cholesterol raising but also associated with CHD) was recently identified in the *SCARB1* gene (Zanoni, Khetarpal et al. 2016) providing further evidence that high HDL-cholesterol is not necessarily protective and may in some circumstances promote CHD.

It is also unclear why rs10911021 should be associated with HDL traits in T2D but not in the general population. As previously mentioned one feature of diabetic dyslipidemia is low HDL-cholesterol mostly driven by a potentially pro-atherogenic reduction in the presence

of larger HDL particles (Krauss 2004). It may be that presence of the minor allele of rs10911021 leads to changes in the expression of protein(s) involved in HDL metabolism altering the composition of large HDL particles, creating slightly less pro-atherogenic particles compared to carriers of the risk allele. Of course this presumes that large HDL particles do play a protective role and are not simply a biomarker and/or confounded by another causal factor.

There are several limitations to this study. One study, ET2DS, contributed the majority of participants in our metabolomic analysis of those with T2D. All suggestive associations were lost when this study was left out of the meta-analysis as power was greatly reduced. While the results were adjusted for use of any lipid-lowering medications, data on the specific medication used was not available for analysis and this may have led to residual confounding. It has long been known that the relationship between a particular lipid-lowering medication and HDL-cholesterol varies greatly. For example, rosuvastatin and simvastatin have been found to have a much greater HDL-cholesterol raising ability compared to atorvastatin (Barter, Brandrup-Wognsen et al. 2010). It is unknown how lipid-lowering medications may affect the HDL sub-fractions measured here. A study investigating the impact of statin use on the HDL traits measured here found that the concentration of very large HDL particles increased and the concentration of small HDL particles decreased while the concentration of large and medium HDL particles was largely unaffected (Wurtz, Wang et al. 2016) but this study did not assess individual statins. Due to the very high proportion of the T2D group that were on lipid-lowering medication we were unable to perform any meaningful analysis after exclusion of those on lipid lowering medication.

4.4 Conclusion to the chapter

In summary, both a CHD in T2D specific GS and a general CHD GS were found to be associated with CHD in UCLEB participants with T2D. In addition, data from the UCLEB consortium supported an association between rs10911021 and CHD in T2D. However, our results indicate that rs10911021 does not impact upon CHD risk by limiting the availability of the glutathione constituent glycine or by inhibiting general amino acid translocation into the cells. Furthermore, rs10911021 was found to be associated with classically measured HDL-cholesterol levels and a number of large HDL particle traits in those with T2D only. Counterintuitively, the minor "protective" allele was associated with the atherogenic phenotype in both classically measured HDL-cholesterol and the metabolomics large HDL traits pointing to a potentially novel mechanism through which HDL particles promote CHD pathogenesis.

5 The CoRDia study

5.1 Introduction

While T2D confers an increased risk of CHD, good glycaemic control can minimise this risk (as discussed in Chapter 4.1). Thus reducing the CHD risk associated with the diabetic state can be achieved by taking regular exercise, having a balanced diet and taking medication as directed. Such behaviours may also reduce the likelihood or severity of other diabetic complications such as retinopathy (Zhang, Zhao et al. 2015) and renal disease (Holman, Paul et al. 2008). The importance of good diabetes management from both an individual and a public health perspective is underlined in the most recent NICE guidelines which recommend that patients attend structured education classes covering this at diagnosis, with annual follow-up thereafter (2015). Such self-management interventions (SMI) seek to provide participants with motivation to make behavioural changes, a forum to learn problem solving skills and to increase confidence, all of which are key to sustainable lifestyle change (Barlow, Wright et al. 2002). Systematic reviews of clinical trial data have found that SMI attendance improves diabetes management (Steinsbekk, Rygg et al. 2012; Chrvala, Sherr et al. 2015) although effectiveness in "real-world" clinical settings remains to be demonstrated. An examination of the Canadian registry data found attendance led to better quality of care but no improvement in the rate of diabetic complications or mortality after a median of 5.3 years follow-up (Shah, Hwee et al. 2015).

The CoRDia study seeks to investigate the impact of attendance at SMI sessions with and without provision of personalised CHD risk information compared to usual care, in patients with poorly managed T2D. This will be assessed using two primary outcomes: glycated haemoglobin levels and comparison of CHD risk (as determined using the UKPDS risk score, see Chapter 4.1) across the study groups. These parameters were determined at baseline and will be re-assessed at 6-month and 12-month post-recruitment. A number of behavioural and clinical secondary outcomes (such as smoking cessation and cholesterol levels) will also be assessed.

CoRDia recruitment commenced in December 2013 and follow-up is due to finish in June 2016. Therefore, the aim of the work presented here is to i) assess the baseline T2D-CRF characteristics of the recruits and ii) to determine the genetic characteristics of the participants in the SMI plus risk profile group. Furthermore, how CHD risk in the CoRDia recruits compares to a cross-sectional study of individuals with T2D was also assessed.

5.2 CoRDia trial protocol

The full protocol for the CoRDia study has been published elsewhere (Davies, McGale et al. 2015) and is depicted diagrammatically in Figure 29. Briefly, participants were recruited from primary care centres in the East of England. All were defined as having poorly controlled T2D based as on glycated haemoglobin levels (Hba1c >6.45 %). The age range of participants was 25-74 years and those of European, Afro-Caribbean, Asian Indian, mixed European/Afro-Caribbean or mixed European/Asian Indian ethnicity were included. Participants were randomised into one of three groups (i) usual care, (ii) SMI only or (iii) SMI-plus-personalised-CHD-risk. CHD risk was estimated using the UKPDS risk score (Stevens, Kothari et al. 2001). When calculating the UKPDS score, those of mixed ethnicity were treated as non-Afro-Caribbean as suggested by the developers of the UKPDS risk score. The UKPDS risk score was combined with the weighted 19 SNP GS (as described in Chapter 2.5.2, score calculated using the original weightings) to give a combined CHD risk for participants in the SMI plus risk profile group. All participants were free of CVD at baseline. Ethical approval for this study has been granted by the East of England Research Ethics Committee (ref 12/EE/0437) and the study complied with the ethical principles underlying the Declaration of Helsinki. This study has been registered at ClinicalTrials.gov; registration identifier NCT01891786.

Personalised CHD risk information in the form of a risk report was delivered to participants in the SMI plus risk profile group in one-to-one sessions with a researcher. In the report, genetic CHD risk (relative to population CHD risk) was displayed graphically along with average genetic risk (set at zero). A statement contextualising genetic risk was also included, and the wording depended on the individual's genetic CHD risk score. The statement read - "Your risk is (adjective used based on cut-offs shown in Table 70) was higher/lower than the average person". The UKPDS CHD risk - percentage risk of CHD in the next ten years - was referred to as "lifestyle risk" and was displayed graphically alongside the UKPDS score for an "average person" of the same age (and with the same duration of diabetes), sex, smoking status and ethnicity as the participant. The values used for the other variables are given in Table 71, the systolic blood pressure and lipid ratio were determined using data from UDACS (Stephens, Hurel et al. 2004) and the glycated haemoglobin value was set under guidance from a group clinician. Combined CHD risk was also displayed in this manner.

Table 70: Adjectives used in CoRDia study risk reports to describe genetic CHD risk relative to the average population risk

Displayed Genetic Risk	Value
>0.8	"Considerably"
0.3 - 0.7999	"Moderately"
0.051-0.299	"Slightly"
0-0.05	"Minimally"
0	"Same as average"

The same adjective was used whether displayed genetic risk was positive or negative.

Table 71: Values for variables in the UKPDS score used to calculate risk in an "average" person of the same age, sex, smoking status and ethnicity as the participant in the CoRDia study risk reports.

Trait	Value
Systolic Blood Pressure	136 mmHg
HbA1c	6.7 %
Lipid Ratio	3.71

Follow-up time-points were immediately after the cessation of the SMI (2-4 months post-recruitment, behavioural measures only), at 6 months and 12 months post recruitment with questionnaires and clinical measures taken at these time-points. Data collection from follow-up is not yet complete and data from intermediate time points has not yet been fully collated and analysed. Thus it is not available for presentation in this thesis.

Figure 26: Example of the depiction of genetic CHD risk in the CoRDia study risk reports

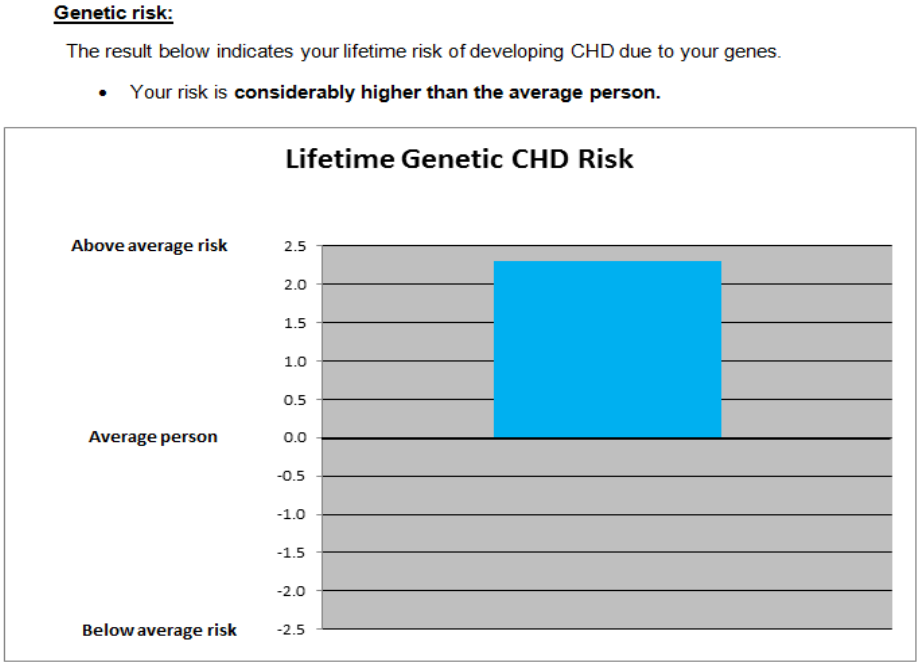


Figure 27: Example of the depiction of ten-year CHD risk, as determined by the UKPDS risk score and referred to as "lifestyle risk", in the CoRDia study risk reports

Lifestyle risk:

- **Your estimated lifestyle-related risk of developing CHD in the next 10 years is 22.9%.**
- The risk for an average person with diabetes, of the same gender, age and ethnicity as you is **18.7%, therefore your estimated risk is 122.4% higher than the average person.**
- **Your lifestyle-related risk of CHD can be reduced.**

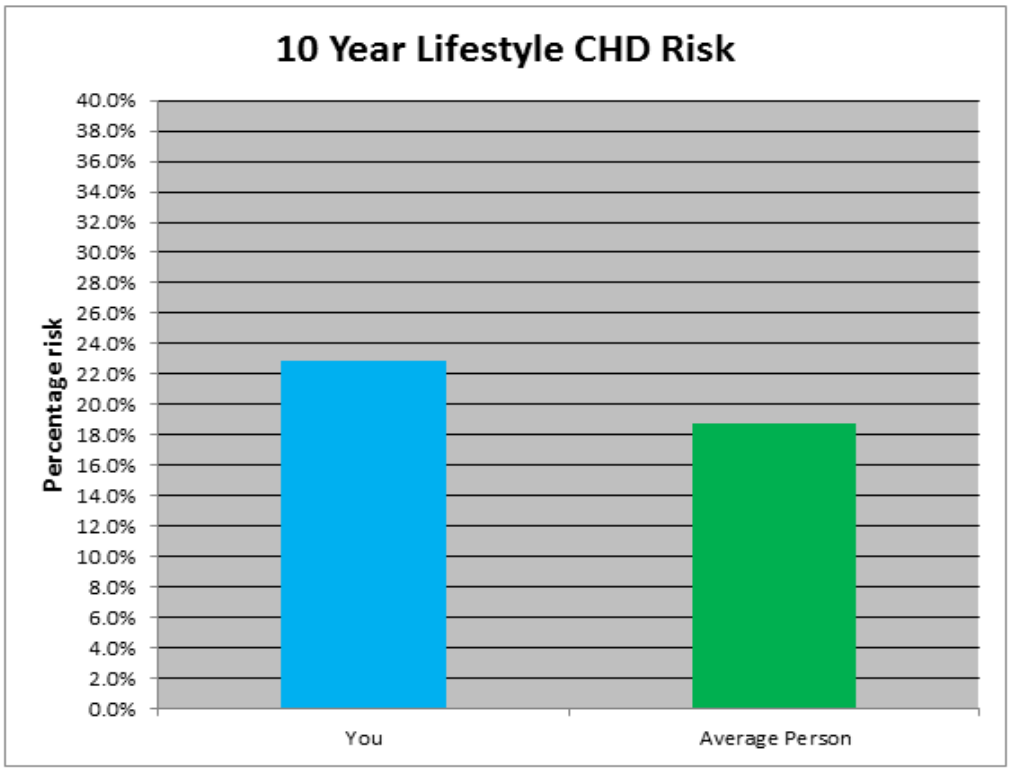


Figure 28: Example of the depiction of combined ten-year CHD risk, as determined by the UKPDS risk score plus the genetic risk, in the CoRDia study risk reports

Overall risk:

- **Your estimated overall risk of CHD in the next 10 years is 39.8%.**
- The risk for an average person with diabetes of the same gender, age and ethnicity as you is **18.7%, therefore your estimated risk of CHD in the next 10 years is 212.8% higher than the average person.**

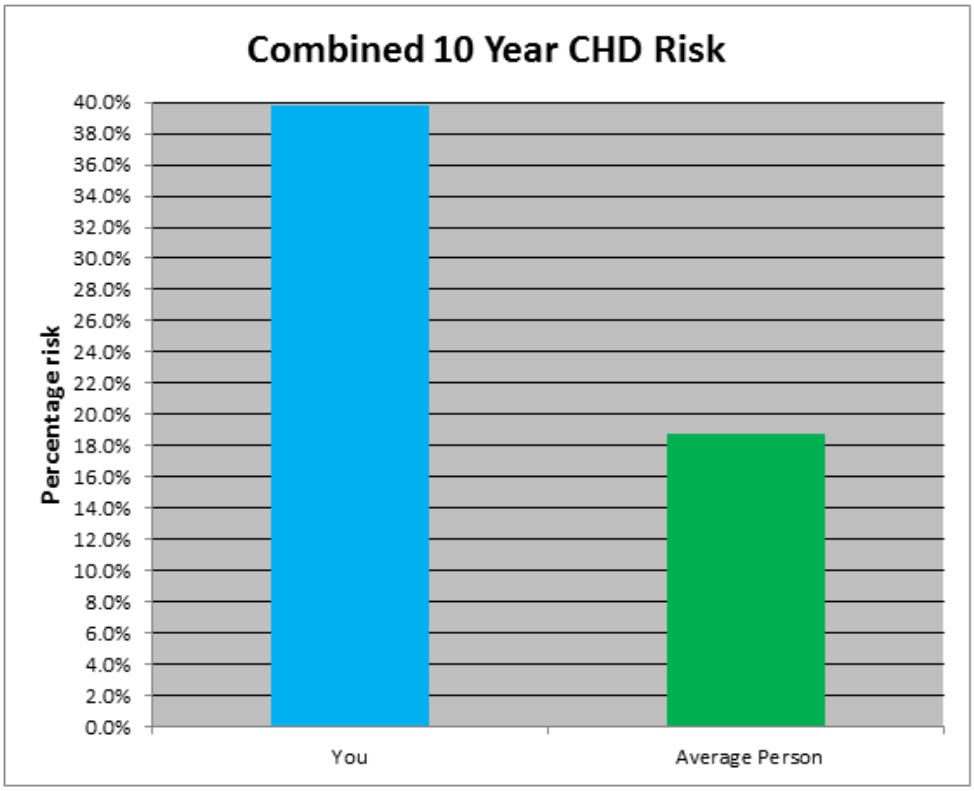


Figure 29: Flow-chart representing the CoRDia study protocol

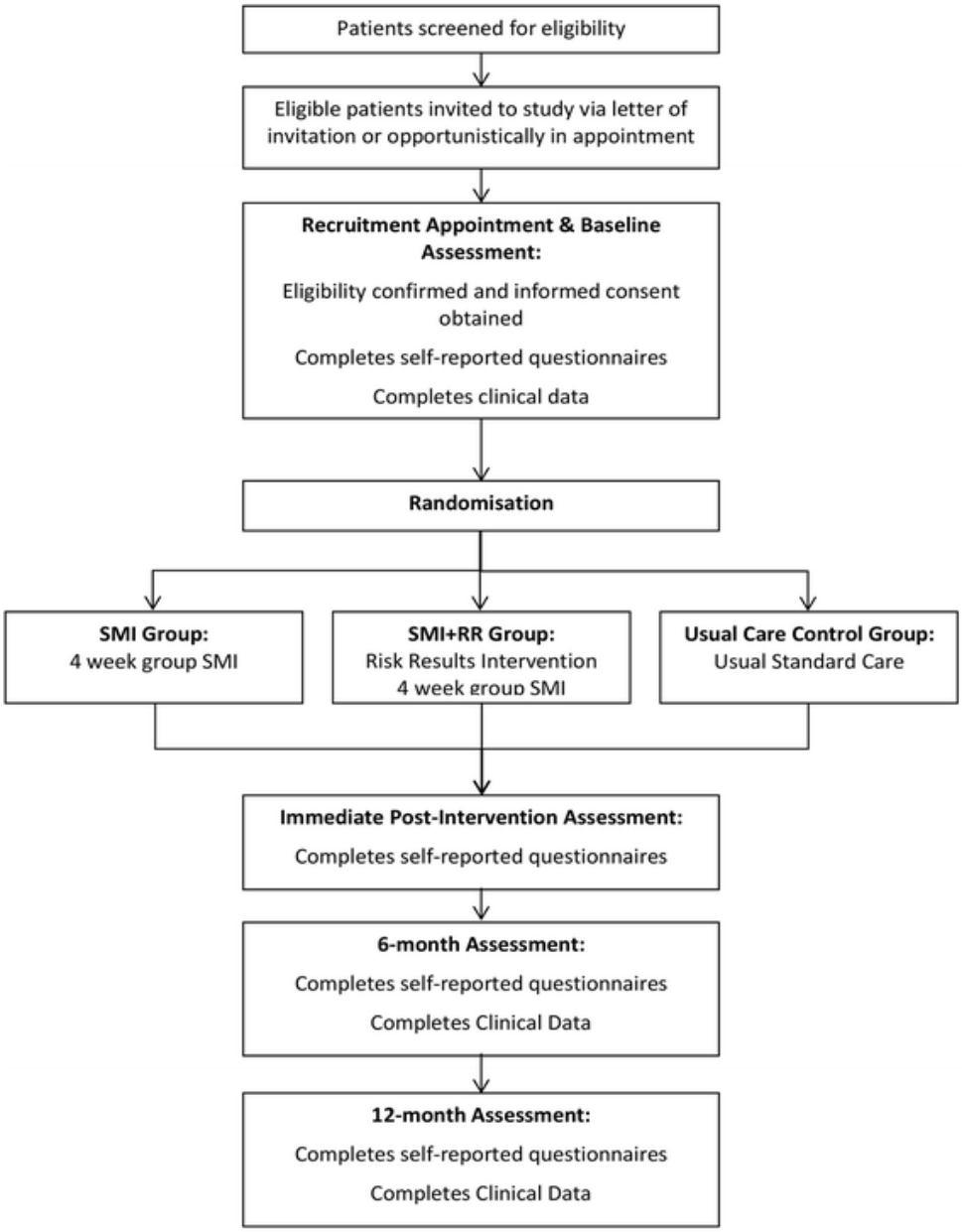


Figure originally published by BioMed Central in (Davies, McGale et al. 2015).

5.3 Results:

5.3.1 Baseline characteristics of the CoRDia participants

Participant recruitment was conducted at 14 GP surgeries and one community diabetes clinic in the East of England from December 2013 to June 2015 (Figure 30). The baseline T2D-CHD risk factor characteristics of the CoRDia recruits are shown by randomisation group in Table 72. None of the risk factors differed between the three groups nor did ten-year risk of CHD (a p>0.05).

Figure 30: Map of CoRDia recruitment sites

Recruitment sites are indicated by orange dots. Image created using data from "OpenStreetMap" available under Open Database License. ©OpenStreetMap contributors. (http://www.openstreetmap.org/copyright.org).

Table 72: Baseline characteristics of the CoRDia study participants by randomisation group

Trait	CoRDia Control Group (n=67)	CoRDia SMI only Group (n=74)	CoRDia SMI+risk profile Group (n=70)	p-value
Age (years)	61.40 (10.08)	61.28 (9.11)	62.36 (7.42)	0.53
Sex (% Female)	48 % (n=33)	46 % (n=40)	34 % (n=24)	0.05
Ethnicity (% European)	92 % (n=60)	92 % (n=68)	94 % (n=66)	0.65
TC (mmol/l)*	4.18 (0.97)	4.36 (0.89)	4.29 (1.03)	0.55
HDL-cholesterol (mmol/l)*	1.27 (0.29)	1.23 (0.37)	1.16 (0.29)	0.05
Systolic blood pressure (mmHg)	133.93 (12.56)	134.36 (14.41)	133.44 (12.97)	0.97
Duration of T2D (years)**	5.60 (0-23)	5.60 (0-39)	6.33 (0-23)	0.38
Age of T2D Onset (years)	55.00 (10.87)	54.99 (10.17)	54.90 (8.36)	0.95
Glycated haemoglobin (%)*	7.73 (1.29)	7.60 (1.03)	7.58 (0.99)	0.41
Smoking % (n)	7 % (n=5)	5 % (n=4)	16 % (n=11)	0.09
UKPDS risk score (%)*	12.70 % (9.56)	13.18 % (9.44)	16.06 (10.47)	0.13

The mean plus standard deviation is shown, where appropriate. Numerical variables were compared using ANOVA and categorical variables by Chi-squared tests. *Variable was log transformed. Geometric mean and approximate standard deviation are shown.**Variable was square-root transformed, means were transformed back and the range is shown. TC=total cholesterol.

To ascertain how ten-year CHD risk observed in the CoRDia recruits compared to other cohorts from the T2D population, the UKPDS score was compared to that observed in the UDACS study (in those with no history of CHD). Unlike CoRDia, where participants were recruited from primary care, the UDACS participants were recruited from a hospital diabetes clinic. CHD risk was found to be higher in the UDACS recruits compared to the CoRDia (median 14.89 % v 20.85 % $p=<2.20x10^{-16}$, Figure 31). Comparing the T2D-CHD CRFs between the two studies reveals that the groups have very different risk profiles (Table 73). The UDACS participants were older, had been diagnosed with T2D at an earlier age and the duration of T2D was longer in this group. In addition, total cholesterol, HDL-cholesterol and systolic blood pressure were higher in UDACS, as was the proportion of smokers but the proportion of those of European ethnicity was lower. The lower cholesterol levels in the CoRDia recruits can be accounted for by the much higher proportion of CoRDia participants on lipid lowering therapy compared to the UDACS participants (76% v 20% $p<2.20x10^{-16}$). It is unsurprising therefore that CHD risk was also higher. Glycated haemoglobin did not differ between the CoRDia participants and the UDACS participants. To qualify for inclusion in the CoRDia study recruits were required to have poorly controlled T2D as defined by their

glycated haemoglobin level. Therefore, this suggests that a majority of the UDACS participants may also have poorly controlled diabetes but this would assume that diabetes management was equally effective when the UDACS participants were recruited in 2001-2002, as now.

Figure 31: Ten-year CHD risk as determined by the UKPDS score in A) the CoRDia participants and B) UDACS participants

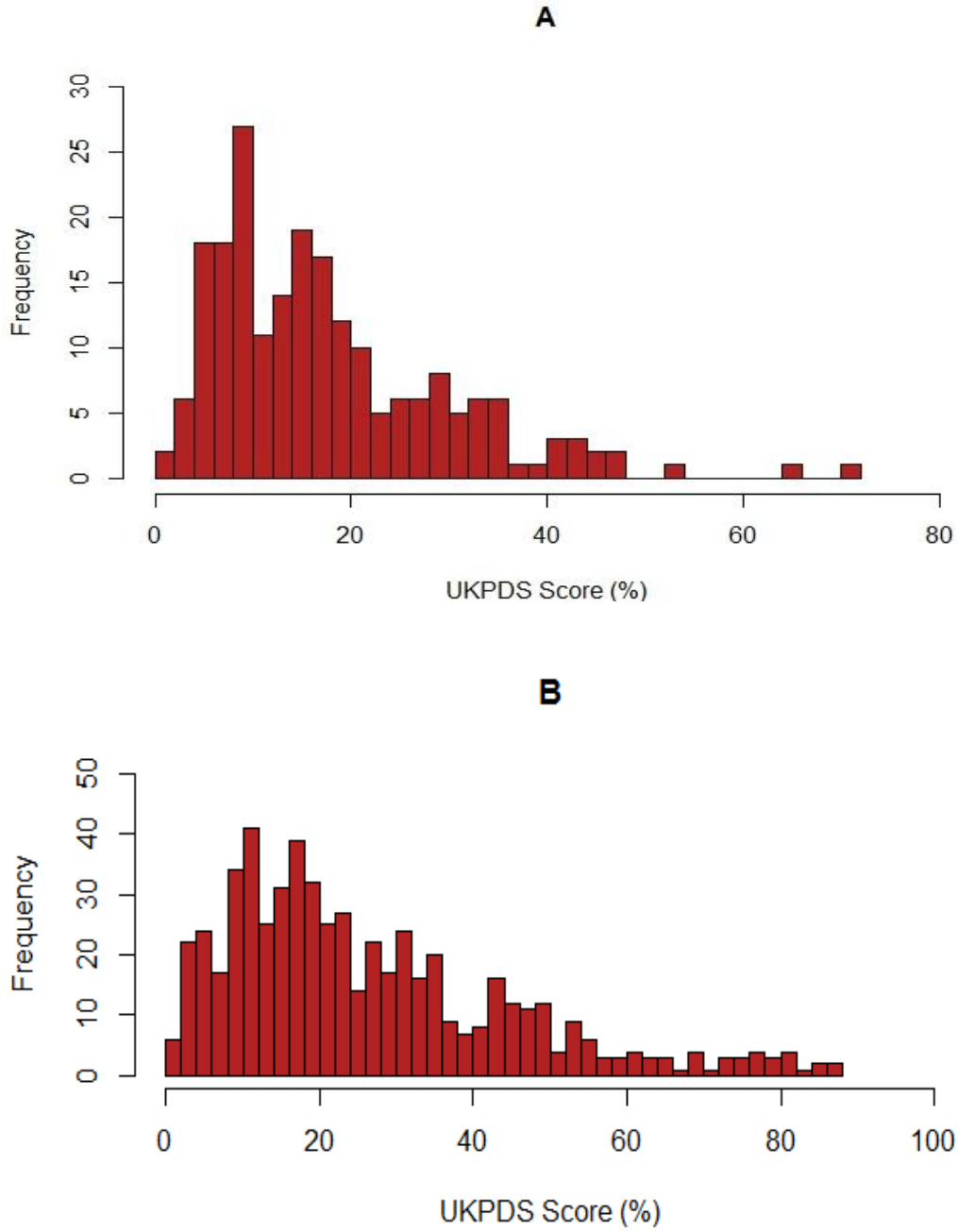


Table 73: Baseline characteristics of the CoRDia participants and UDACS participants without CHD at recruitment

Trait	CoRDia (n=211)	UDACS (n=597)	p-value
Age (years)	61.68 (8.90)	64.02 (11.55)	$2.55x10^{-3}$
Sex (% Female)	45 % (n=96)	40 % (n=240)	0.22
Ethnicity (% European^{+})	98 % (n=207)	74 % (n=442)	$9.61x10^{-8}$
TC (mmol/l)$^{\&}$	4.20 (3.73-4.89)	5.10 (4.40-5.80)	$<2.20x10^{-16}$
HDL-cholesterol (mmol/l)*	1.21 (0.32)	1.29 (0.38)	$9.55x10^{-3}$
Systolic blood pressure (mmHg)$^{\&}$	134.00 (125.0-140.0)	141.00 (130.10-151.50)	$9.09x10^{-11}$
Duration of T2D (years)**	5.80 (0-39)	9.16 (0-60)	$3.20x10^{-10}$
Age of T2D Onset (years)	54.96 (9.80)	52.92 (13.17)	0.01
Glycated haemoglobin (%)*	7.63 (1.10)	7.83 (1.71)	0.06
Smoking % (n)	10 % (20)	17% (97)	0.02
UKPDS risk score (%)$^{\&}$	14.89 (8.34-23.53)	20.85 (12.65-34.53)	$<2.20x10^{-16}$

The mean plus standard deviation is shown where appropriate. *Variable was log transformed. Geometric mean and approximate standard deviation are shown (except for UKPDS risk score where the interquartile range is shown). **Variable was square root transformed, mean was transformed back and range is also shown. $^{\&}$Medians and interquartile range shown as distribution of the variable appeared to differ between the studies. These variables were compared using a Mann-Whiney Wilcox test. The other numeric variables were compared using Welch's t-test and categorical variables were compared using chi-square tests. TC=total cholesterol. $^{+}$For the CoRDia group this refers to non-Afro Caribbean participants.

5.3.2 Genetic CHD risk in CoRDia

Genetic CHD risk was determined in the SMI plus risk profile group using the 19 SNP GS with the original weightings (Chapter 3.1). To assess how the genetic CHD risk of the recruits compared to that of the non-T2D population, the 19 SNP GS was compared between the CoRDia SMI plus risk profile group and NPHSII. Furthermore, the UCLEB participants with T2D were used to represent a the T2D population. The distribution of risk alleles in NPSHII (n=1360) and the participants of European ethnicity in the CoRDia SMI plus risk profile group (n=66 as there were four participants not of European ethnicity in this group) is shown in Figure 32. Furthermore, there was no difference in mean GS between the participants of European ethnicity of the CoRDia group and NPSHII (3.14 v 3.17 p=0.80, Figure 33) or between the participants of European ethnicity of the CoRDia group the UCLEB participants with T2D (3.14 v 3.27 p=0.14, Figure 33).

Figure 32: Histogram of unweighted GS in A) NPHSII B) UCLEB participants with T2D and C) CoRDia SMI plus risk profile group

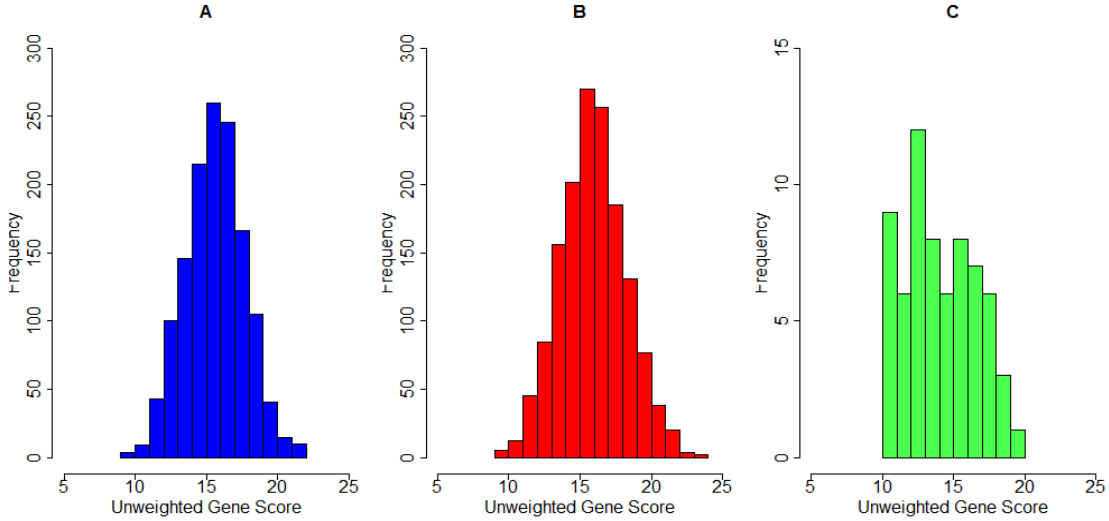
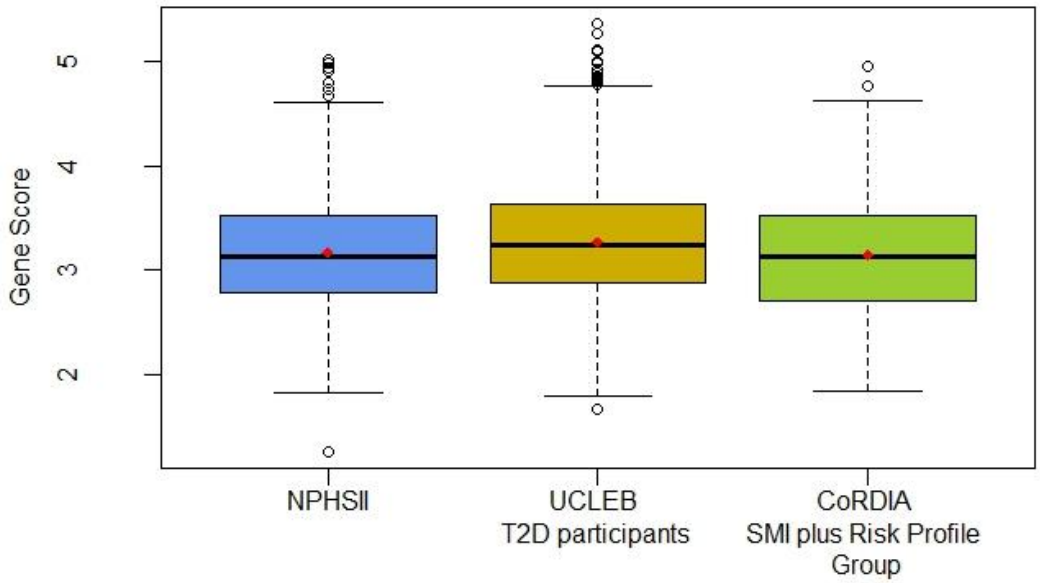


Figure 33: Boxplot of GS in NPSII, the UCLEB participants with T2D and the CoRDia SMI plus Risk Profile group



The mean GS (gene score) for each group is marked in red (horizontal line represents the median).

5.4 Discussion

The CoRDia study seeks to investigate how attendance at SMI sessions with and without knowledge of personal CHD risk affects clinical and behavioural outcomes compared to usual T2D care. CHD risk was determined using the T2D-specific UKPDS score (Stevens, Kothari et al. 2001) as well as a combined UKPDS plus genetic risk score. As has been discussed previously (Chapter 4.1), the UKPDS score has been found to overestimate CHD risk. However, it was recommended for use by the NICE guidelines (2008) when the study was commenced and therefore was used to estimate CHD risk in the participants. Furthermore, the primary purpose of the study is not to evaluate the risk tool per se but to investigate how knowledge of personal CHD risk impacts upon behavioural and clinical characteristics in the context of attendance at SMI sessions. Thus while the ideal would be to use a well calibrated risk score (and the use of a poorly calibrated one is a limitation of the study), using the UKPDS will still allow the research aims to be met. Another limitation in relation to using the UKPDS score is how to categorise those of mixed ethnicity. The UKPDS score assigns those of Afro-Caribbean ethnicity with a lower CHD risk compared to those of European or South Asian ethnicity and Afro-Caribbean ethnicity is coded as a binary variable. However, how to code those of mixed ethnicity is not clear. Under guidance from the developers of the UKPDS score, those of mixed race (mixed European and Afro-Caribbean) were assigned with non-Afro-Caribbean ethnicity with the caveat that this may underestimate risk. Only two participants were of mixed European and Afro-Caribbean ethnicity were recruited (both in the usual case group) so this is unlikely to affect the findings.

Subjects were randomised using a pre-specified procedure (Davies, McGale et al. 2015) and none of the T2D-CRFs nor ten-year CHD risk was found to differ between the three study groups. The genetic CHD risk in the SMI plus risk profile arm was found to be similar to that in a non-T2D cohort (NPHSII) and a T2D cohort (UCLEB T2D participants). This suggests that the genetic CHD risk of the recruits in the CoRDia SMI plus risk profile group is reflective of the general UK population.

In the CoRDia trial the personalised risk information given was a combined CRF and GS risk score. Therefore, the effect of including genetic information in addition to CRF risk cannot be assessed. The MI-GENES trial compared LDL-cholesterol levels between two groups of individuals with intermediate CHD risk (Kullo, Jouni et al. 2016). One group was given CHD risk as determined by a CRF risk score. The other group was given the CRF score and how it compared to a combined risk score which had genetic risk incorporated into it. Genetic risk was determined using a GS comprising 28 SNPs robustly associated with CHD, but not with CRFs – 4 of these SNPs are included in the 19 SNP GS. An individualised graphic displaying how many people in 100 with the same risk profile would have an MI in the next ten years was used to communicate CHD risk to participants. A similar graphic was used to visualise the impact of genetics upon CHD risk (e.g. if the person had low genetic risk it would show how many MIs would be prevented as a result) and the reduction in risk that would result from statin use (shown as number of events prevented by the medication). All participants then had a meeting with a clinician, primarily to discuss whether to initiate statin treatment. At 6-month follow-up those in the CRF plus GS group were found to have lower LDL-cholesterol and a greater proportion had initiated statin treatment (a likely cause of the former). Statin use was highest and LDL-cholesterol lowest in those with high genetic risk, indicating that knowledge of personalised genetic information increases motivation to start and adhere to risk-lowering medication. There was no difference in dietary fat intake or physical activity between the groups, suggesting that knowledge of personalised risk alone does not help sustain healthier lifestyle choices. Thus other strategies (such as attendance at SMI sessions) must be employed to help patients do this and the CoRDia study will assess if access to personalised CHD risk information will improve outcomes over-and–above SMI attendance alone.

5.5 Conclusion to chapter

Analysis of the baseline CoRDia data found that glycated haemoglobin level and ten-year CHD risk were similar between the three randomisation groups. The CHD risk observed in the CoRDia subjects recruited from primary care was lower than in the UDACS participants recruited from a hospital based clinic. The genetic CHD risk of the SMI plus risk profile group was found to be similar to other cohorts with and without T2D. Follow-up will be completed in June 2016.

6 Functional analysis of CHD risk locus 21q22

6.1 Introduction

The CHD risk locus on chromosome 21q22 was first identified in a GWAS of early-onset MI (Kathiresan, Altschuler et al. 2009). The association has since been confirmed in the CARDIoGRAMplusC4D meta-analysis (Deloukas, Kanoni et al. 2013) where each copy of the minor allele was associated with an increase in CHD risk (OR=1.13). However, the mechanism through which this locus affects CHD risk remains obscure. Like many of the GWAS identified CHD risk loci, 21q22 is not associated with any CRFs for CHD (Deloukas, Kanoni et al. 2013). Moreover, it lies in a "gene desert" (Figure 34). The closest upstream genes are *SLC5A3* and *MRSP6*. These genes share an exon which is in the open reading frame for *MRPS6* but not *SLC5A3* (Gardiner, Slavov et al. 2002). *SLC5A3* (solute carrier family 5-inositol transporter) encodes a sodium myo-inositol transporter which is involved in the response to hypertonic stress (Berry, Mallee et al. 1995). *MRPS6* (mitochondrial ribosomal protein 6) encodes a subunit of the mitochondrial ribosome (Suzuki, Terasaki et al. 2001). The locus containing these two genes was identified in a GWAS concerning red blood cell traits as being associated with packed cell volume (van der Harst, Zhang et al. 2012). The closest downstream gene is the potassium channel subunit encoding *KCNE2* and mutations in this protein are known to cause long-QT syndrome (Abbott, Sesti et al. 1999). Long-QT syndrome is associated with arrhythmias and sudden cardiac death. Furthermore, even within the normal range, longer QT-interval has been associated with increased risk of CHD mortality (Zhang, Post et al. 2011). However, it is unclear whether longer-QT interval has a causal relationship with CHD. Variants in or close to *KCNE2* have also been associated with lung function (Soler Artigas, Loth et al. 2011), height (Lango Allen, Estrada et al. 2010) and BMI in Hispanic post-menopausal women who smoke (Velez Edwards, Naj et al. 2013). However, none of these points to a plausible pathway to account for the association with CHD and thus genomic location does not suggest any obvious mechanism through which this CHD risk locus is acting.

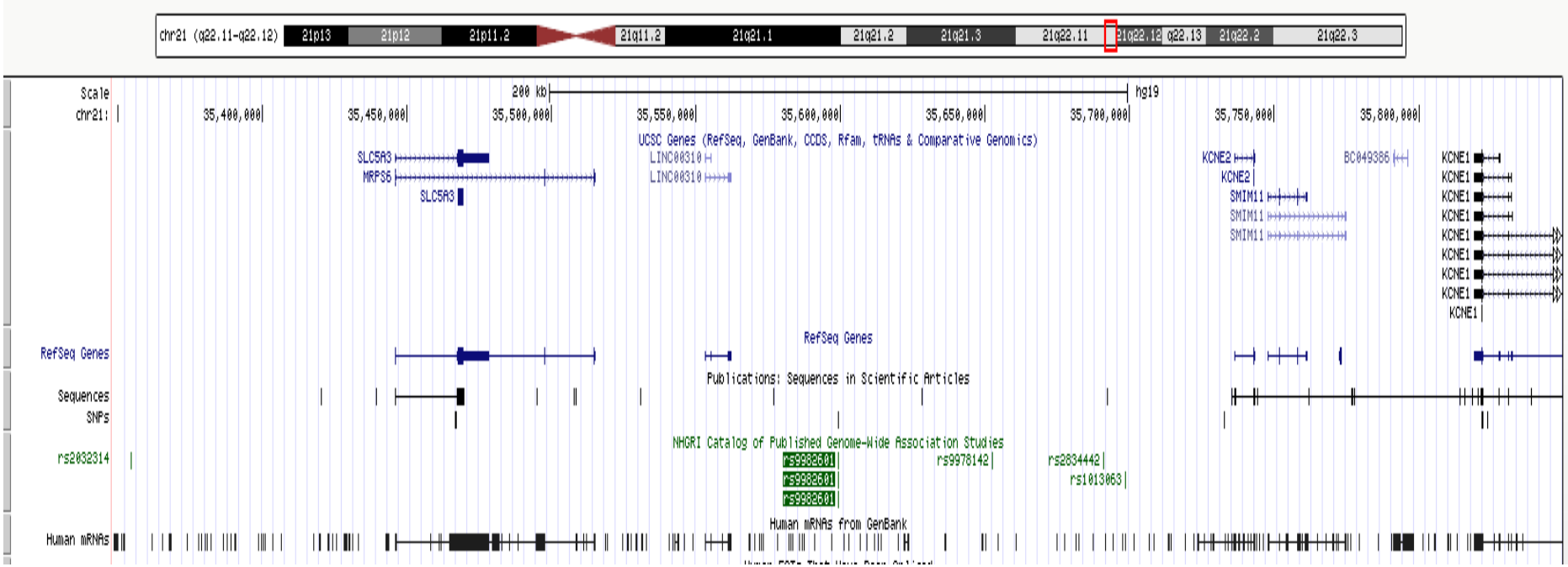
Being located within a gene desert, neither the lead SNP at the 21q22 risk locus nor any SNPs in strong LD result in changes to the protein coding sequence. Rather, it is likely that the locus impacts upon risk of CHD through involvement in the regulation of gene expression (Hindorff, Sethupathy et al. 2009). A SNP that is located within open, transcriptionally active chromatin is more likely to be functional than a SNP located within heterochromatin and thus investigating the genomic context of a locus can provide insight into its functionality. Publically available bioinformatics data, both experimentally

determined and predicted, can be used to study this. HaploReg (Ward and Kellis 2012) can be used to search all SNPs in LD with a query SNP (r^2>=0.2 as determined in 1000 Genomes phase 1) and its output combines information concerning these SNPs from a variety of sources. This includes data on transcription factor binding, presence of DNAse I hypersensitivity sites and chromatin marks from ENCODE (Bernstein, Birney et al. 2012), predicted regulatory chromatin state from the Roadmap Epigenomics Project (Kundaje, Meuleman et al. 2015) and eQTL data from GTEX (2013). More detailed information from the ENCODE data can be obtained using the UCSC genome browser (Kent, Sugnet et al. 2002) where results of interest (e.g. presence of enhancer chromatin marks in a relevant cell line) can be displayed alongside genomic location or the GTEX browser itself. Thus such tools can be used to identify candidate functional SNPs.

Candidate functional SNPs identified from bioinformatics analysis require to be investigated experimentally. A crucial aspect of DNA variation is how it affects transcription factor binding. If a SNP lies within the binding site for a particular transcription factor, presence of one allele could lessen or enhance the binding affinity of the protein for that sequence compared to the other allele. Ultimately this could result in differential gene expression. Therefore, allele-specific binding suggests functionality and this can be studied *in vitro* with EMSAs. EMSAs can be used to screen for functional variants as many variants can be studied in the same experiment. If a particular variant shows consistent allele-specific binding, then its effect on gene expression can be investigated using a dual-reporter luciferase assay. Here the impact of each allele on expression of the reporter luciferase gene is investigated. Should a SNP show allele-specific binding and affect gene expression, this is strong evidence of functionality. The focus can then shift to the identification of the specific molecular pathways (i.e. transcription factors and target genes) involved.

Identification of the functional SNP(s) at a particular risk locus and the mechanism through which is impacts upon CHD risk can provide an important insight into the pathogenesis of CHD. In this context, the aim of this study to was investigate the 21q22 CHD risk locus using bioinformatics analysis and *in vitro* functional assays to identify the candidate functional SNP(s).

Figure 34: Schematic image of the 21q22 CHD risk locus taken from the UCSC Genome Browser (http://genome.ucsc.edu).



The lead SNP at the CHD risk locus, rs9982601 is highlighted in green. Genes present at the locus are shown in blue with *SLC5A3* and *MRPS6* upstream of the risk locus and *KCNE2* downstream of the risk locus. GWAS hits from the NHGRI catalog are also shown in green. Displayed is rs9978142, a GWAS hit the ratio of forced expiratory volume in 1 second/forced vital capacity (Soler Artigas, Loth et al. 2011) and rs2834442, a GWAS hit for height (Lango Allen, Estrada et al. 2010). Also shown is rs1013063, a SNP found to be associated with BMI measures with smoking in Hispanic women in a study looking at gene environment interactions (Velez Edwards, Naj et al. 2013).

6.2 Results

6.2.1 Association of the 21q22 CHD risk locus with CHD risk factors

In order to assess whether the 21q22 risk locus was associated with any traits that might provide an insight into how it affects CHD, a phenome scan of the UCLEB data set was performed. This included many inflammatory markers, lipid traits and a number of physiological phenotypes. The cohort had been genotyped using the Metabochip platform (designed to cover regions associated with cardiometabolic disease (Voight, Kang et al. 2012)). Four SNPs at the 21q22 locus lwere analysed, the lead SNP rs9982601 and three SNPs in moderate LD, rs8131284 (r^2=0.78), rs7278204 (r^2=0.76) and rs973754 (r^2=0.75) - r^2 was determined using the 1000 Genomes phase 1 data (Abecasis, Auton et al. 2012). Over one hundred traits were tested but none met the Bonferroni-adjusted significance threshold ($p=4.72x10^{-4}$). Only one trait, QT interval, showed a suggestive association ($p<0.05$) for all four SNPs with the effect in the same direction (Table 74). The CHD risk allele was nominally associated with longer QT interval. This putative association is of interest due to the close proximity of the potassium ion channel gene *KCNE2* to the risk locus.

The p-value for height was also below 0.05 for all four SNPs but for 3 SNPs the rare allele was nominally associated with greater height while for rs9982601, the effect was in the opposite direction. It is noteworthy that rs2834442 (GWAS hit for height (Lango Allen, Estrada et al. 2010)) is also present at this locus (Table 75) and is in weak LD with the lead SNP (r^2=0.23, 1000 Genomes pilot data).

Table 74: The association between four SNPs at the CHD risk locus on chromosome 21q22 and mean QT interval in UCLEB.

SNP					Beta-coefficient	p-value
rs9982601	Genotype	CC	TC	TT	1.83 (0.90)	0.04
	n (frequency)	5329 (0.75)	1643 (0.23)	130 (0.02)		
	Mean QT interval (ms)	402.8 (37.35)	404.5 (39.28)	409.1 (41.08)		
rs8131284	Genotype	TT	CT	CC	2.14 (0.89)	0.02
	n (frequency)	5293 (0.75)	1670 (0.24)	142 (0.02)		
	Mean QT interval (ms)	402.7 (37.41)	404.9 (39.20)	408.1 (38.97)		
rs7278204	Genotype	AA	GA	GG	2.07 (0.89)	0.02
	n (frequency)	5290 (0.74)	1673 (0.24)	141 (0.02)		
	Mean QT interval (ms)	402.7 (37.73)	404.9 (38.19)	407.9 (39.26)		
rs973754	Genotype	AA	GA	GG	1.85 (0.89)	0.02
	n (frequency)	5300 (0.75)	1663 (0.23)	142 (0.02)		
	Mean QT interval (ms)	402.7 (37.75)	404.6 (38.18)	408.2 (38.84)		

Mean QT interval is shown by genotype as well as the beta coefficient for the minor allele (±standard error).

Table 75: The association between four SNPs at the CHD risk locus on chromosome 21q22 and mean height n UCLEB

SNP					Beta-coefficient	p-value
rs9982601	Genotype	**CC**	**TC**	**TT**	-0.002 ($9x10^{-4}$)	0.02
	n (frequency)	9481 (0.75)	2899 (0.23)	212 (0.02)		
	Mean height (m)	1.680 (0.10)	1.679 (0.10)	1.678 (0.10)		
rs8131284	Genotype	**TT**	**CT**	**CC**	0.003 ($9x10^{-4}$)	0.004
	n (frequency)	9424 (0.75)	2924 (0.23)	248 (0.02)		
	Mean height (m)	1.68 (0.10)	1.68 (0.10)	1.681 (0.10)		
rs7278204	Genotype	**AA**	**GA**	**GG**	0.003 ($9x10^{-4}$)	0.003
	n (frequency)	9417 (0.75)	2931 (0.23)	246 (0.02)		
	Mean height (m)	1.68 (0.10)	1.68 (0.10)	1.681 (0.10)		
rs973754	Genotype	**AA**	**GA**	**GG**	0.003 ($9x10^{-4}$)	0.002
	n (frequency)	9433 (0.75)	2917 (0.23)	247 (0.02)		
	Mean height (m)	1.68 (0.10)	1.68 (0.10)	1.68 (0.10)		

Mean height is shown by genotype as well as the beta coefficient for the minor allele (±standard error).

6.2.2 Identification of a putative functional SNP

6.2.2.1 Bioinformatics analysis of the CHD risk locus 21q22

HaploReg v2 (Ward and Kellis 2012) was used to identify SNPs that are in strong LD (r^2>0.8 as calculated in the 1000 Genomes phase 1 EUR data) with the lead SNP, rs9982601. The output is shown in Figure 35. Five such SNPs were identified. The genomic context of rs9982601 and plus these five SNPs was assessed using the UCSC Genome Brower (Figure 36- 40) (Kent, Sugnet et al. 2002). One SNP (rs28451064) showed strong evidence of residing within open chromatin in some cell types. Data from the ENCODE project showed that this SNP is located in DNase I hypersensitivity sites in HepG2, Huh-7 and human umbilical vein endothelial cell (HUVEC) cell lines. This SNP was also found to be positioned within a site bound by multiple transcription factors including specificity protein 1(SP1) and forkhead box A2 (FOXA2) (Figure 36).

6.2.2.2 Assessment of allele-specific binding

The lead SNP plus the five in strong LD were investigated for allele-specific binding of nuclear proteins using EMSAs. The assays were performed with nuclear extracts from two hepatocyte carcinoma cell lines, HepG2 and Huh-7, as the only enhancer chromatin marks found for any of the SNPs at this locus were in HepG2 cells. Five of SNPs showed allele specific binding in the initial experiments (Figure 41) but this was only consistent for rs28451064 when replicates were performed (Figure 42). Therefore, the bioinformatics analysis together with the EMSA results shows that rs28451064 is a strong candidate to be the functional SNP.

6.2.2.3 Predicted impact of rs28451064 on transcription factor binding

In order to assess whether how the presence of the two alleles of rs28451064 might affect transcription factor binding, the genomatix software suite (Genomatix Software GmbH, Munich, Germany) which details predicted binding sites, was used. Presence of the minor “A” allele rather than the major “G” allele was predicted to abolish a vitamin D receptor-retinoid X receptor heterodimer (VDR-RXR) binding site and a homeodomain protein H6 family member 3 (HMX3) binding site. The software also predicted the creation of a forkhead-related transcription factor 4 (FREAC4) binding site in the presence of the A allele.

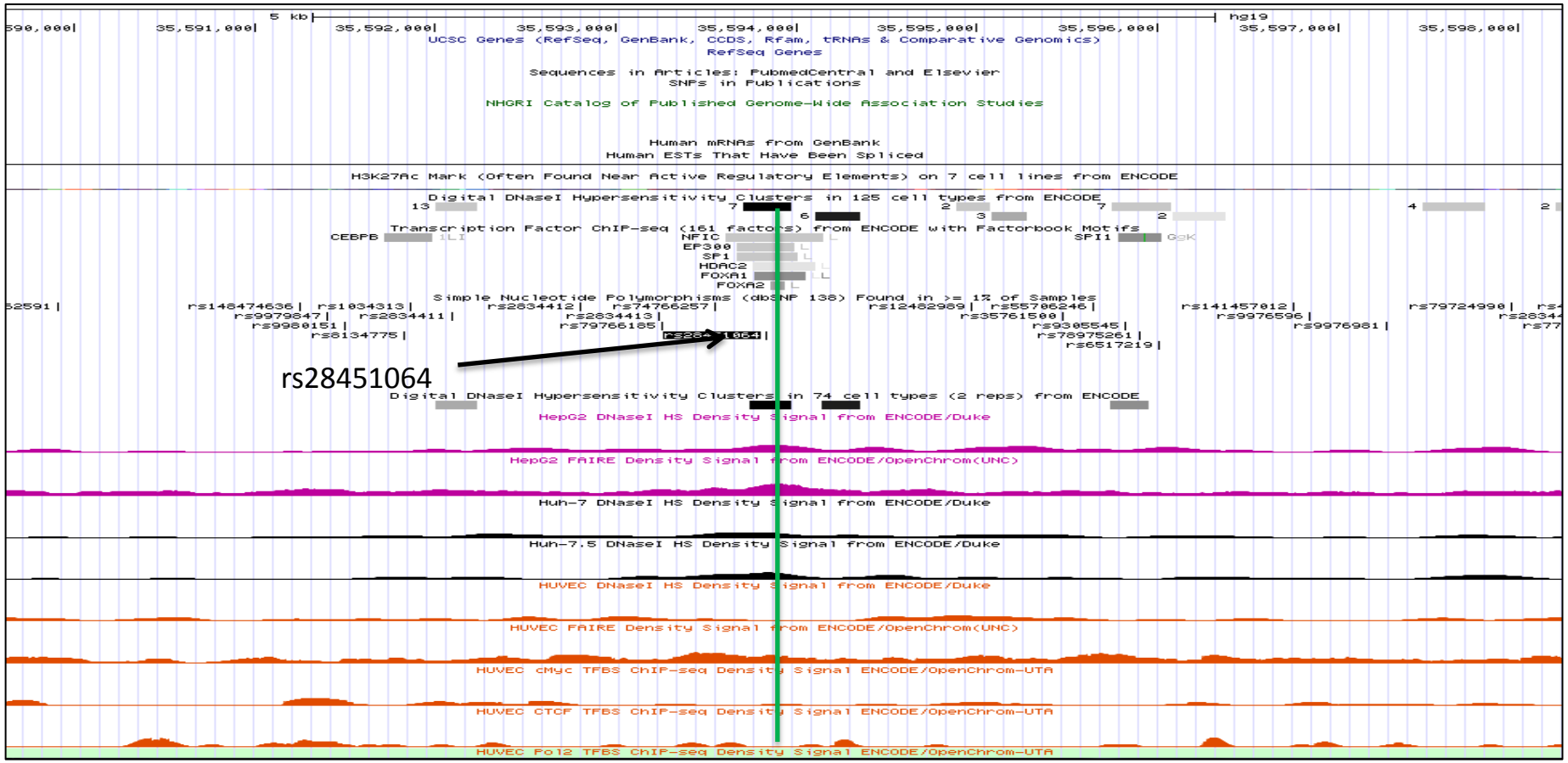
Figure 35: HaploReg v2 output for query SNP rs9982601 (http://www.broadinstitute.org/mammals/haploreg/haploreg_v2.php)

Query SNP: rs9982601 and variants with r^2 >= 0.8

chr	pos (hg19)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	GENCODE genes	dbSNP func annot
21	35593827	0.82	0.92	rs28451064	G	A	0.01	0.08	0.00	0.13			HepG2	7 cell types	FOXA2,SP1		Dobox4,PPAR,RXRA	AP000318.2	
21	35599128	1	1	rs9982601	C	T	0.21	0.10	0.00	0.13				Osteobl			4 altered motifs	AP000318.2	
21	35600505	1	1	rs9980618	C	T	0.20	0.10	0.00	0.13				Th1			Pax-5	AP000318.2	
21	35602268	0.88	1	rs60687229	T	C	0.35	0.12	0.00	0.14							CTCF	AP000318.2	
21	35606199	0.86	0.97	rs9977419	T	A	0.20	0.10	0.00	0.12				CMK	PU1			AP000318.2	
21	35606344	0.86	0.97	rs9977093	G	A	0.20	0.10	0.00	0.12					PU1		16 altered motifs	AP000318.2	

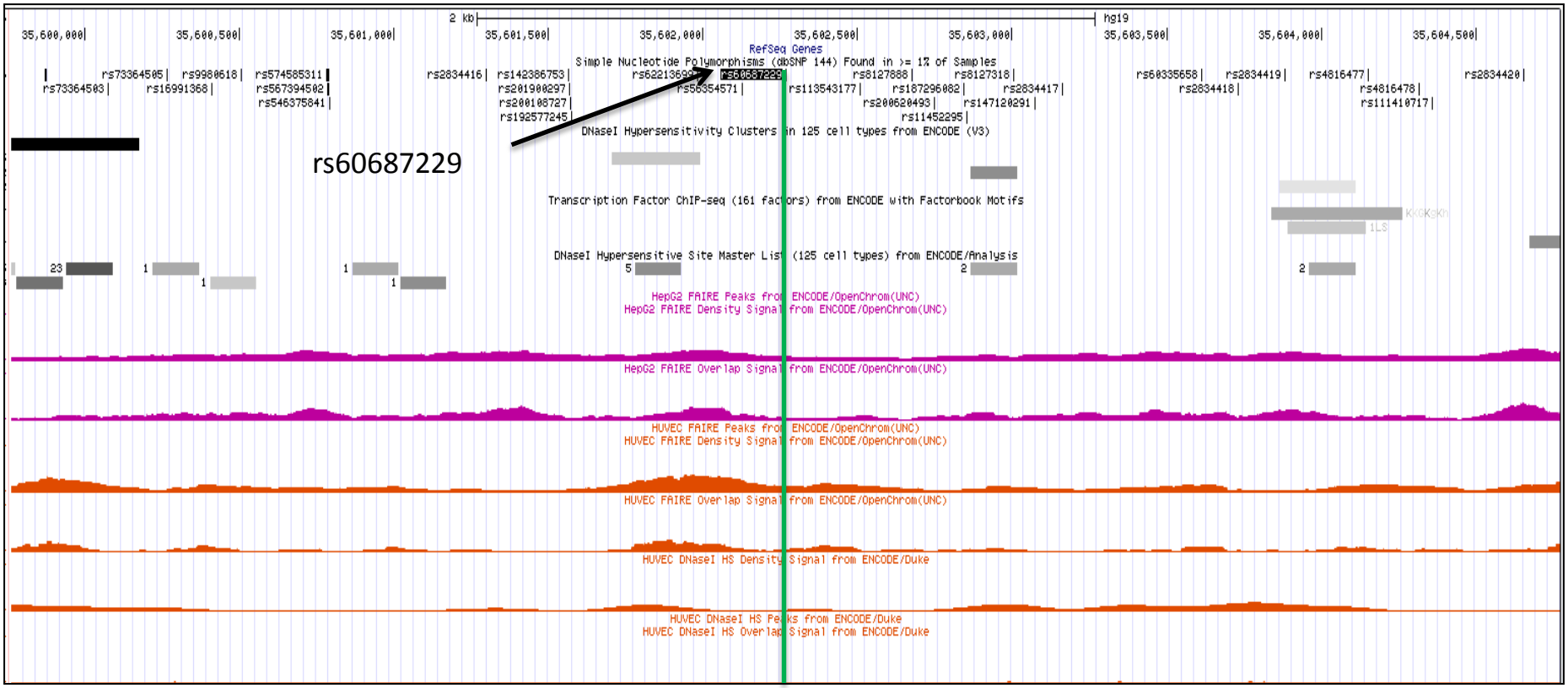
Screenshot taken from HaploReg version 2 output for the query SNP rs9982601. HaploReg (Ward and Kellis 2012) displays frequency information from four ethnic groups from the 1000 genomes project (this is also where the linkage disequilibrium (LD) information is sourced), conservation information is taken from a combination of SNPinfo and TRANSFAC. Promoter histone marks, enhancer histone marks, DNAse and protein binding information comes from the ENCODE project. eQTL data is from the GTEx project and motif alteration predictions are based on a library constructed from the literature. Gene information is taken from GENECODE. The lead SNP at the locus is shown in red. The list of SNPs was limited to SNPs in strong LD with the query SNP (r^2>=0.8 as calculated in the 1000 Genomes phase 1 EUR data).

Figure 36: The genomic environment of rs28451064, taken from the UCSC Genome Browser (http://genome.ucsc.edu;GR37/hg19)



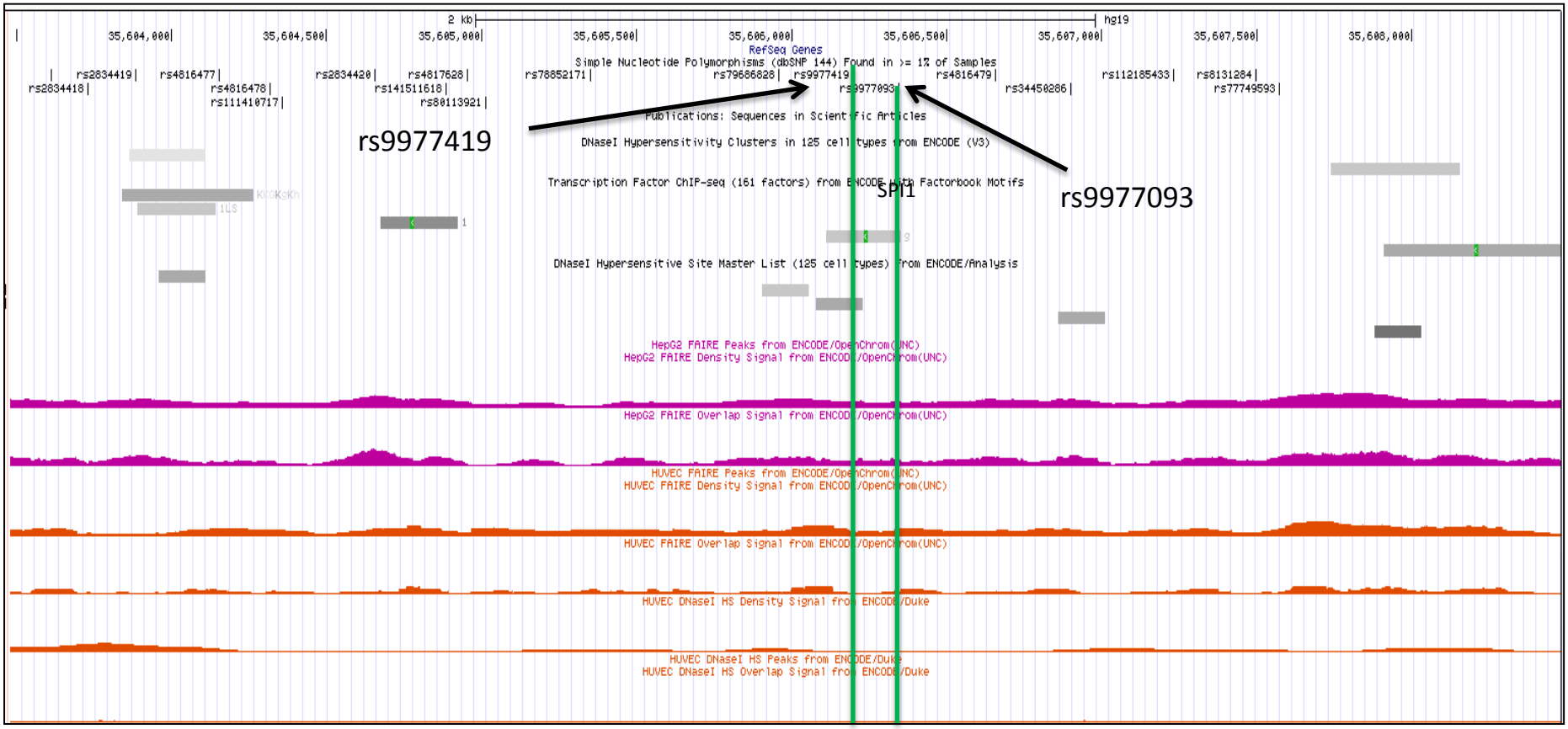
Presence of DNAse I hypersensitivity sites were determined using *in vitro* assays as part of the ENCODE project. The results are shown both as a composite of results in all cell lines tested and in HepG2, Huh-7 and HUVEC cell lines specifically (shown in pink, black and orange respectively). Transcription factor binding sites identified by chip-seq assays performed as part of the ENCODE are also shown (grey boxes). All of the transcription factor binding site were found in HepG2 cells. The position of rs28451064 is indicated by the green line. HUVEC= human umbilical vein endothelial cell.

Figure 37: The genomic environment of rs60687229, taken from the UCSC Genome Browser (http://genome.ucsc.edu; GR37/hg19)



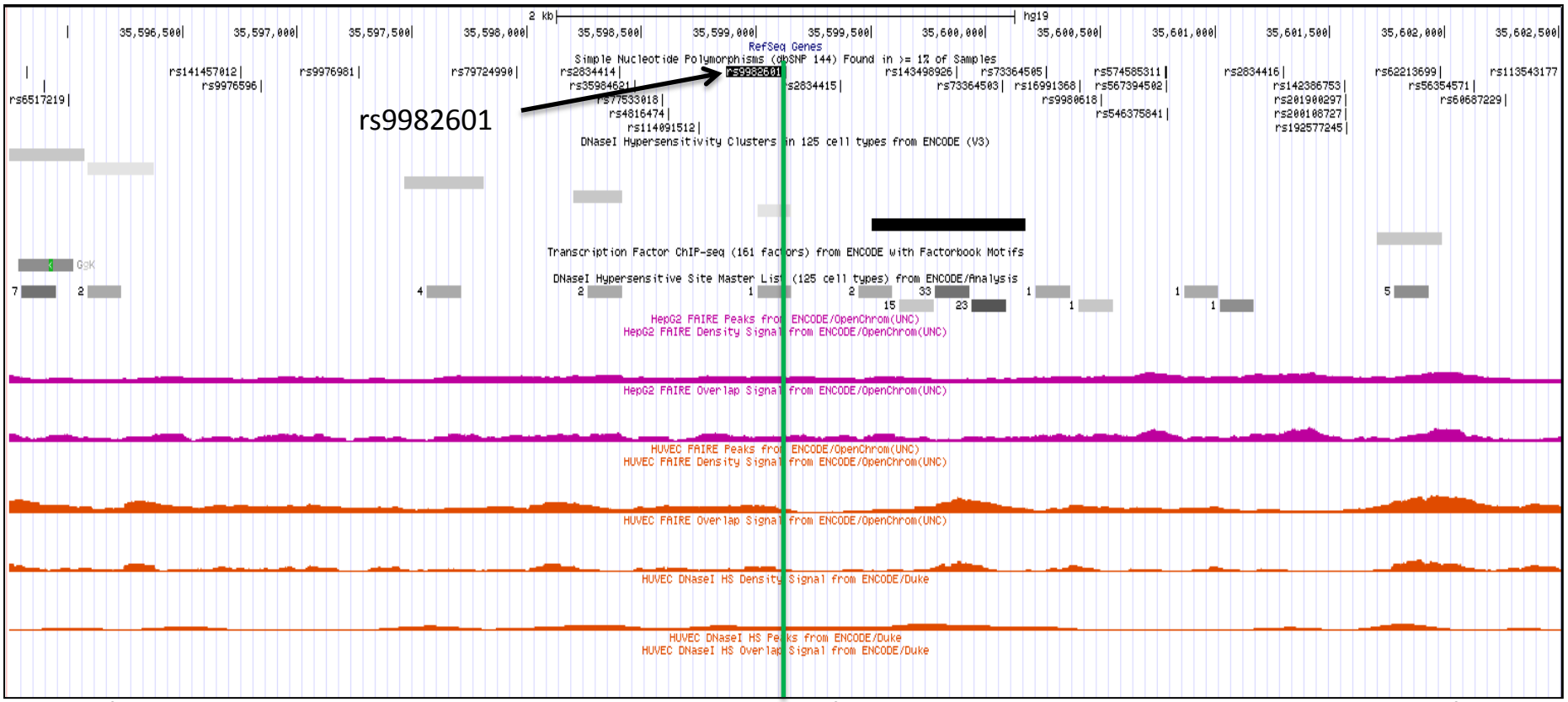
Presence of DNAse I hypersensitivity sites were determined using *in vitro* assays as part of the ENCODE project. The results are shown both as a composite of results in all cell lines tested and in HepG2 and HUVEC cell lines specifically (shown in pink and orange respectively). Transcription factor binding sites identified by chip-seq assays performed as part of the ENCODE are also shown (grey boxes). The position of rs60687229 is indicated by the green line. HUVEC=human umbilical vein endothelial cell.

Figure 38: The genomic environment of rs9977419 and rs9977093, taken from the UCSC Genome Browser (http://genome.ucsc.edu; GR37/hg19)



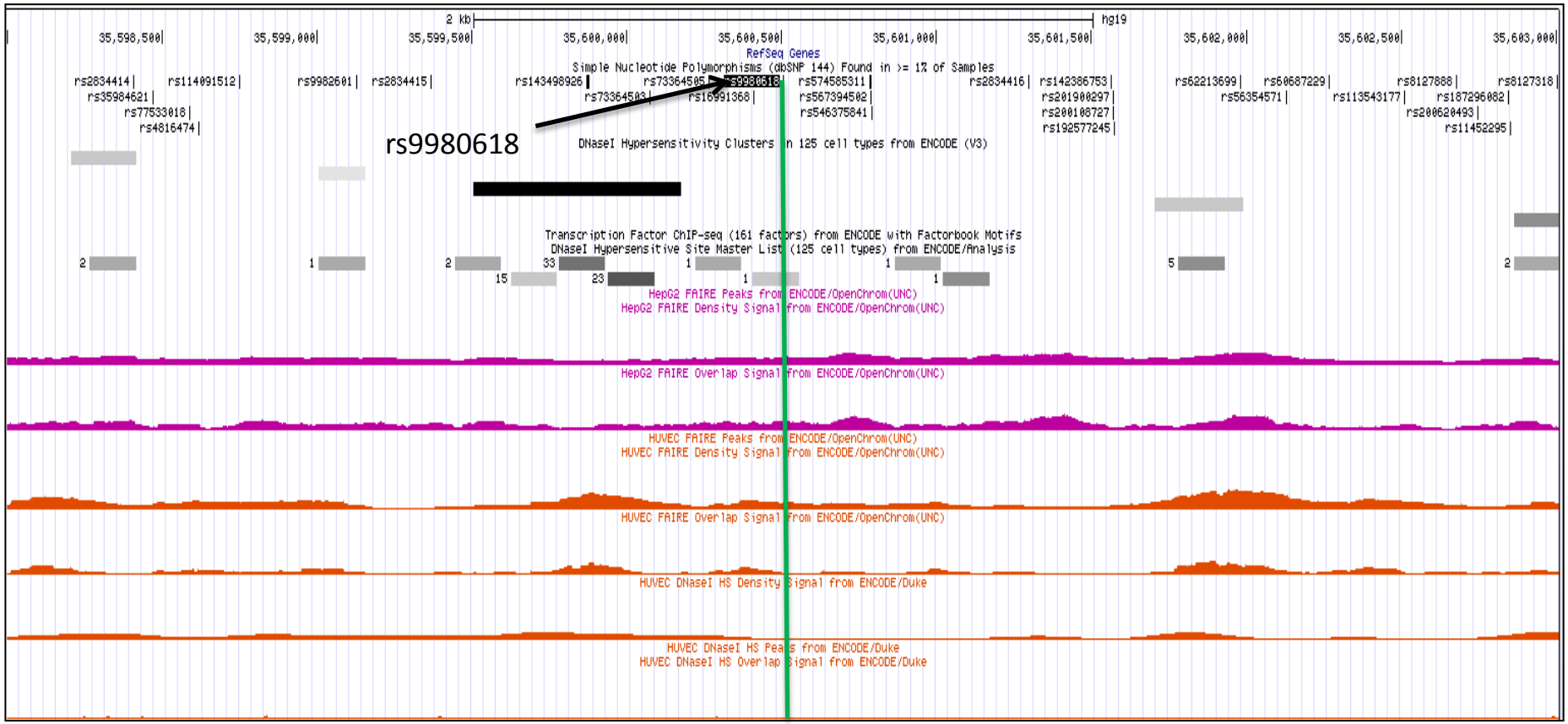
Presence of DNAse I hypersensitivity sites were determined using *in vitro* assays as part of the ENCODE project. The results are shown both as a composite of results in all cell lines tested and in HepG2 and HUVEC cell lines specifically (shown in pink and orange respectively). Transcription factor binding sites identified by chip-seq assays performed as part of the ENCODE are also shown (grey boxes). The positions of rs9977419 and rs9977093 are indicated by the green lines. HUVEC=human umbilical vein endothelial cell.

Figure 39: The genomic environment of rs9982601, taken from the UCSC Genome Browser (http://genome.ucsc.edu;GR37/hg19)



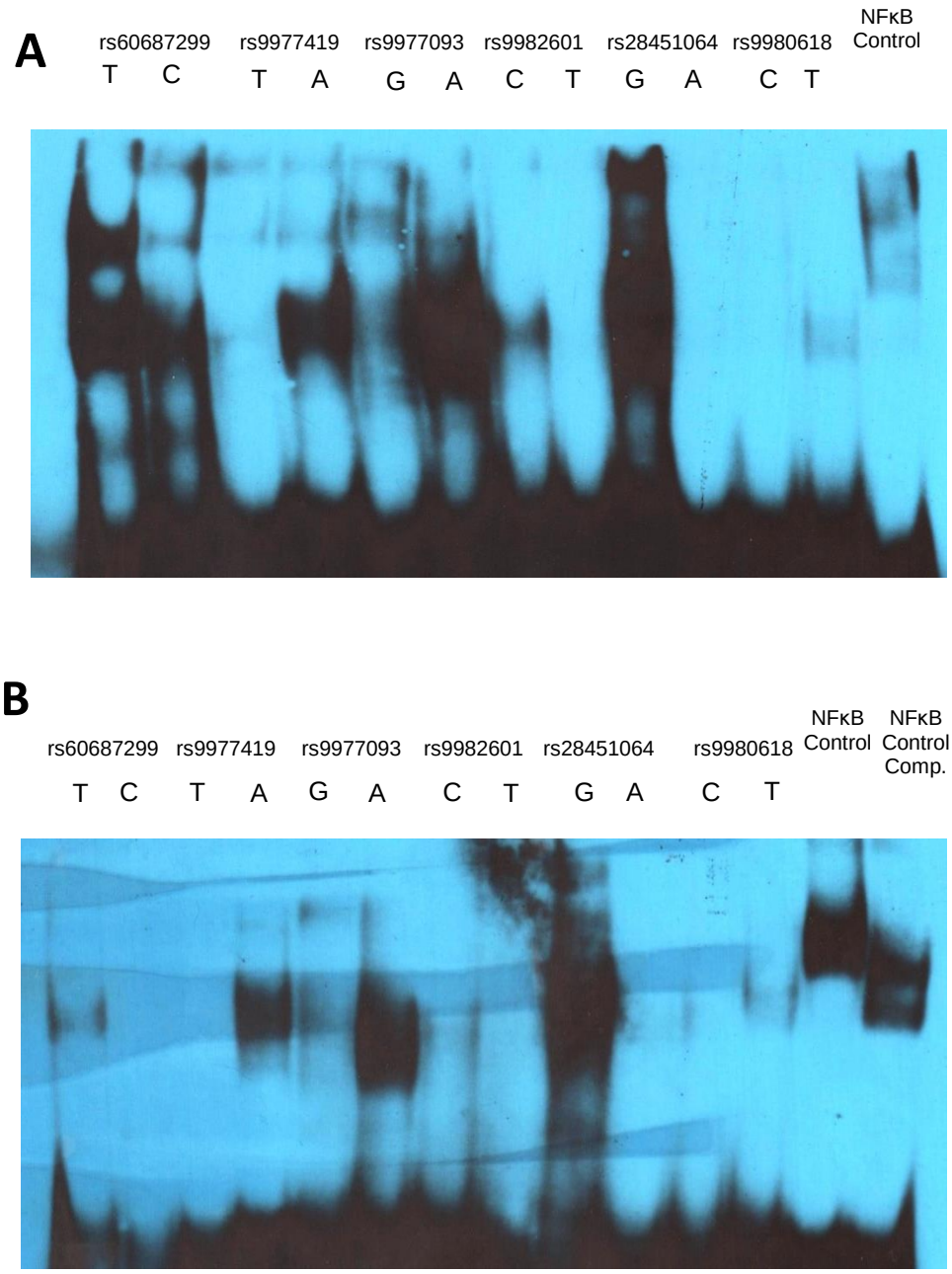
Presence of DNAse I hypersensitivity sites were determined using *in vitro* assays as part of the ENCODE project. The results are shown both as a composite of results in all cell lines tested and in HepG2 and HUVEC cell lines specifically (shown in pink and orange respectively). Transcription factor binding sites identified by chip-seq assays performed as part of the ENCODE are also shown (grey boxes). The position of rs9982601 is indicated by the green line. HUVEC=human umbilical vein endothelial cell.

Figure 40: The genomic environment of rs9980618, taken from the UCSC Genome Browser (http://genome.ucsc.edu; GR37/hg19)



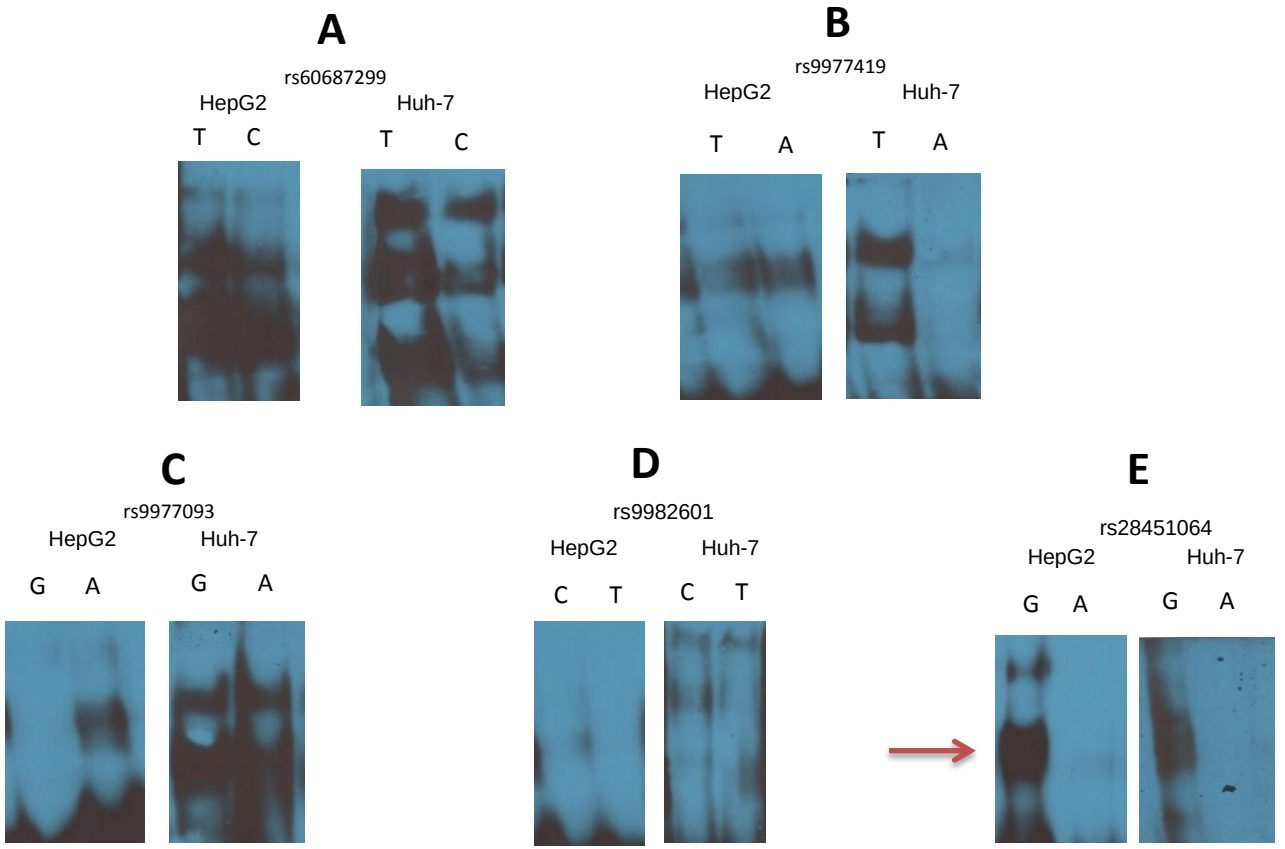
Presence of DNAse I hypersensitivity sites were determined using *in vitro* assays as part of the ENCODE project. The results are shown both as a composite of results in all cell lines tested and in HepG2 and HUVEC cell lines specifically (shown in pink and orange respectively). Transcription factor binding sites identified by chip-seq assays performed as part of the ENCODE are also shown (grey boxes). The position of rs9980618 is indicated by the green line. HUVEC=human umbilical vein endothelial cell.

Figure 41: EMSA results for assays performed with A) HepG2 nuclear extract and B) Huh-7 nuclear extract.



Allele-specific binding can be observed particularly with rs28451064. Both assays used NFkB as a positive control. In the right-most lane of (B) unlabelled NFkB oligo has been added and it can be seen that the original band has been competed out and the oligo has bound to proteins of a different size.

Figure 42: Replication of EMSAs with A) rs60687299 B) rs977419 C) rs977093 D) rs9982601 and E) rs28451064 probes using HepG2 and Huh-7 nuclear extract

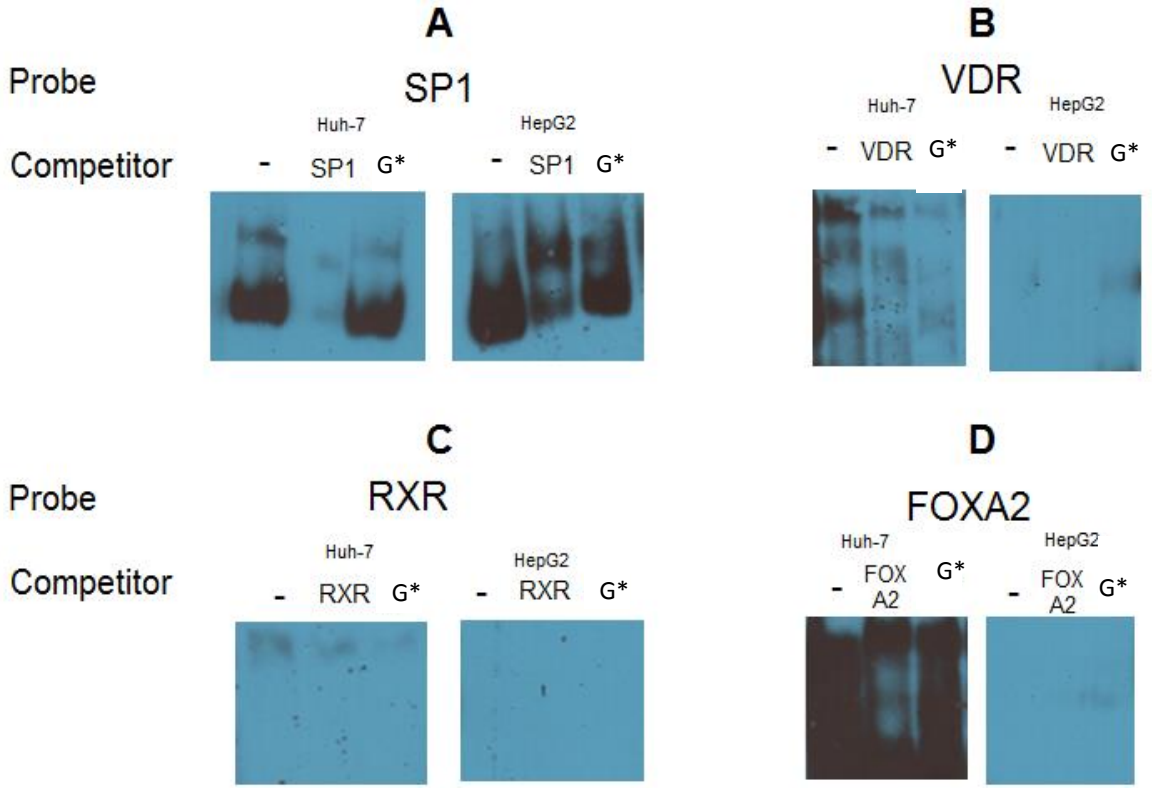


Binding by both alleles was compared for all five SNPs. Only in E (rs28451064) is there strong binding for one allele but complete absence of binding for the other in both cell lines.

6.2.3 Assessment of possible transcription factors

In order to investigate which transcription factors may be binding to rs28451064, competitor EMSAs were performed, again with Huh-7 and HepG2 nuclear extracts (for proteins where consensus sequence probes were available). Four transcription factors were assessed (Figure 43). SP1 and FOXA2 were investigated as these proteins had been found to bind at this locus in the HepG2 cell line in the ENCODE project. VDR and RXR were assessed as the A allele of rs28451064 was predicted to disrupt the binding site for the VDR:RXR heterodimer. The RXR consensus probe did not bind proteins in either extract, indicating the binding conditions were not optimal. However, its binding to the rs28451064-G probe could not be ruled out as the conditions required may differ between the rs28451064-G and the RXR consensus sequence probe. SP1 binding to its consensus sequence was not competed out by the addition of the rs28451064-G probe. This makes it unlikely that SP1 is involved in the allele-specific binding observed previously (Figure 41 and Figure 42). The results with FOXA2 and VDR were inconsistent, with binding observed on some occasions but not others and thus their involvement could neither be discounted nor confirmed.

Figure 43: Competitor EMSA results from assays performed with A) SP1 probes B) VDR probes C) RXR probes and D) FOXA2 probes and HepG2 and Huh-7 nuclear extracts



*G allele of rs28451064.

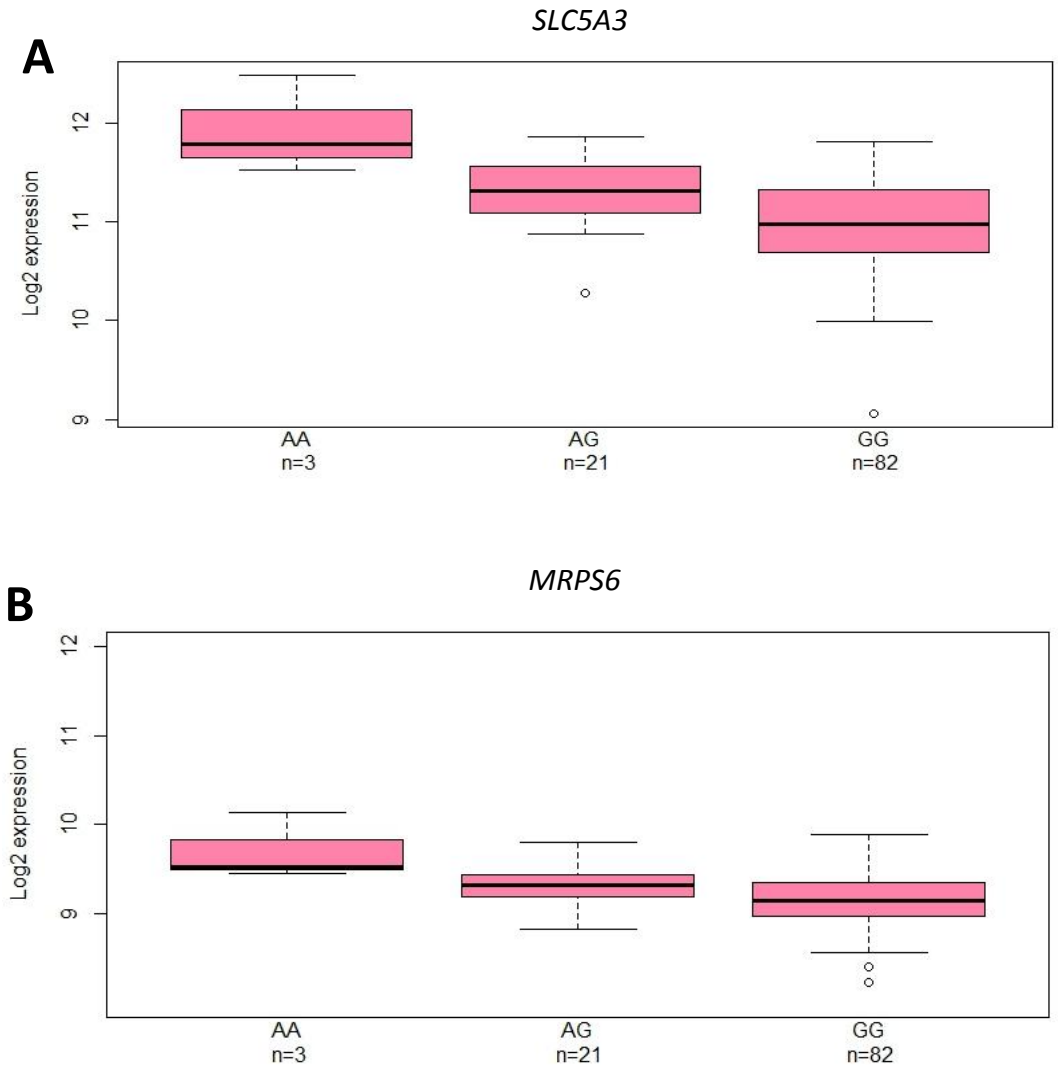
6.2.4 Impact of the 21q22 CHD risk locus on gene expression

The relationship between the risk locus and gene expression also required to be considered. One eQTL for this locus was identified from the literature. The risk allele of the lead SNP, rs9982601, was found to be associated with higher expression of *MRPS6* (closest upstream gene) in blood in the deCODE cohort (Schunkert, Konig et al. 2011). Other data sources were investigated to assess other tissues.

6.2.4.1 ASAP Study

The relationship between the lead SNP at the risk locus and expression of its three closest genes *MRPS6*, *SLC5A3* and *KCNE2* was examined using data from the ASAP study (Folkersen, van't Hooft et al. 2010). Expression data was available for five tissues (liver, mammary artery, aortic adventitia, aortic intima media and heart). Genotyping data for rs9982601 was available for 106 ASAP participants. For both *SLC5A3* and *MRPS6*, the minor CHD risk allele was associated with higher expression of the mRNA transcript in aortic intima media (*SLC5A3* 1.30 fold (95% CIs 1.16-1.47) per A allele $p=3.98x10^{-5}$; *MRPS6* 1.15 fold (95% CIs 1.06-1.25) per A allele $p=9.60x10^{-4}$, Figure 44). No association was observed for *KCNE2*.

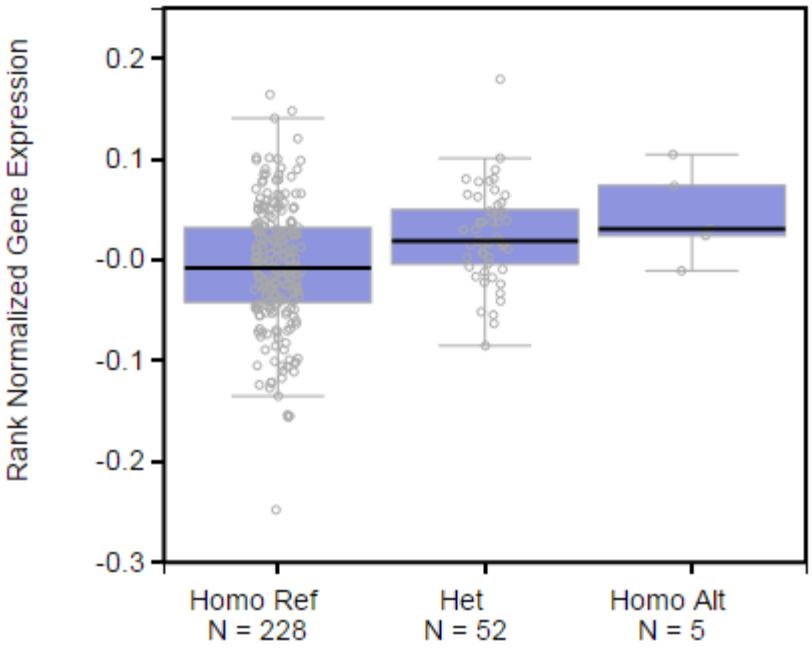
Figure 44: Expression of A) *SLC5A3* and B) *MRPS6* in aortic intima media presented by rs9982601 genotype in the ASAP study.



6.2.4.2 GTEX project

The relationship between the 21q22 CHD risk locus and gene expression was further studied using data from the GTEX project (http://www.gtexportal.org/)(2013). No genes met the significance threshold for single tissue eQTL with either the lead SNP rs9982601 or the putative functional SNP rs28451064. The search was then narrowed to consider the relationship between the risk locus and the three genes located most closely to it, *KCNE2*, *MRPS6* and *SLC5A3*, in seven tissues (Table 76). This gives a Bonferroni-adjusted p-value of $4x10^{-3}$. In agreement with the ASAP results, the minor allele of rs28451064 was found to be associated with higher expression of *MRPS6* in the aortic ($p=1.2x10^{-3}$) and tibial arteries ($p=1.1x10^{-4}$, Figure 45), although not in the coronary artery. There was a suggestive association between the minor allele and lower expression of *MRPS6* in whole blood (p=0.04). Similar results were obtained for rs9982601. There were suggestive associations between the minor allele and higher expression of *KCEN2* in aortic and tibial artery tissue (both p=0.02). Expression data for *SLC5A3* was not available.

Figure 45: Expression of *MRPS6* by rs28451064 genotype in the tibial artery



The graph was created using data from GTEX (2013) (http://www.gtexportal.org/home/).

Table 76: Relationship between rs28451064 and expression of selected genes in seven tissues from GTEX (http://www.gtexportal.org/home/)

Gene	Tissue	n	Effect Size	p-value
MRPS6	Aortic Artery	197	0.17	$1.2x10^{-3}$
MRPS6	Coronary Artery	118	0.10	0.42
MRPS6	Tibial Artery	285	0.21	$1.1x10^{-4}$
MRPS6	Atrial Appendage	159	-0.05	0.67
MRPS6	Left Ventricle (Heart)	190	-0.09	0.42
MRPS6	Liver	97	-0.12	0.48
MRPS6	Whole Blood	338	-0.05	0.04
KCNE2	Aortic Artery	197	0.17	0.02
KCNE2	Coronary Artery	118	0.22	0.06
KCNE2	Tibial Artery	285	0.19	0.02
KCNE2	Atrial Appendage	159	-0.06	0.29
KCNE2	Left Ventricle (Heart)	190	-0.09	0.72
KCNE2	Liver	97	-0.05	0.36
KCNE2	Whole Blood	338	0.16	0.54

Effect sizes refer to the change in expression per minor allele.

6.2.4.3 Impact of rs28451064 on reporter gene expression

To investigate how the rs28451064 affects gene expression, dual luciferase reporter assays were used. To do this the SNP was cloned into the enhancer site of the pGL3 promoter vector which contains the SV40 promoter sequence (Figure 46). The pGL3 plasmids were then transfected into Huh-7 cells. The results are shown in Figure 47 and are from four different experiments each with eleven or twelve replicates.

Figure 46: Schematic diagram of the pGL3 vector with the sequence surrounding rs28451064 inserted downstream of the luciferase gene

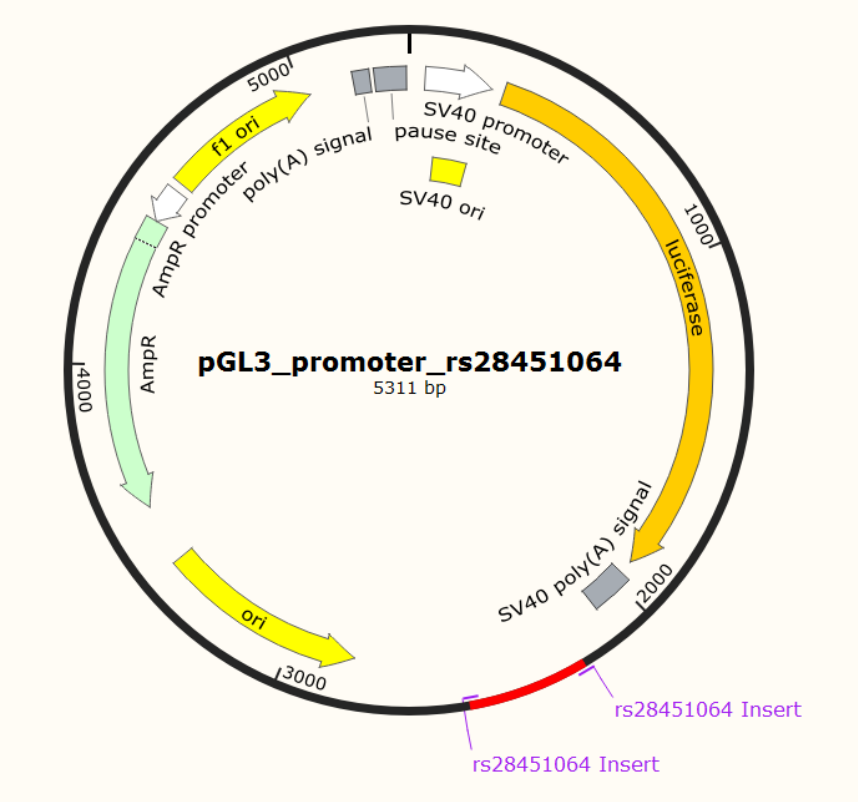
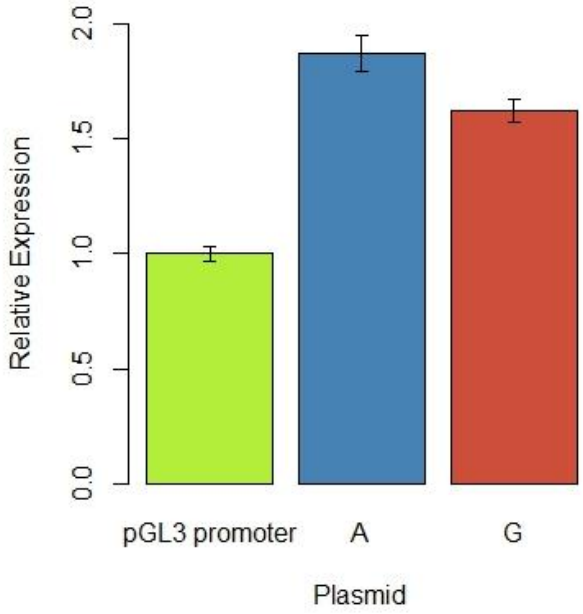


Image create using SnapGene software (from GSL biotech; available at snapgene.com).

Both plasmids containing the sequence surrounding rs28451064 showed higher expression (A allele 87 % higher $p=1.90x10^{-15}$, G allele 62 % higher $p=9.74x10^{-15}$, Figure 47) than the pGL3 promoter plasmid, indicating that this region acts as an enhancer. Furthermore, the minor A (risk) allele was found to have 12 % higher expression compared to the G allele ($p=4.82x10^{-3}$, Figure 47), which is in agreement with the eQTL data.

Figure 47: Relative expression of a vector containing the rs284510654 A allele and rs28451064 G allele normalised to the pGL3 promoter expression



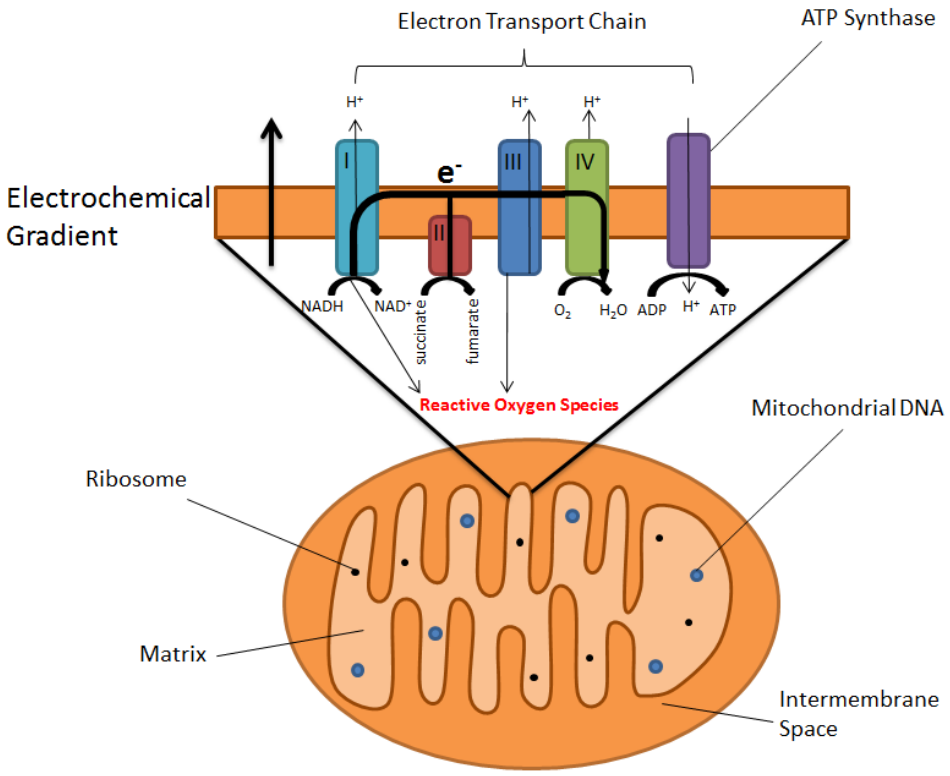
Relative expression was compared using paired t-tests. A=rs28451064 A allele, G=rs28451064 G allele. The errors bars represent 95% confidence intervals.

6.3 Discussion

A locus on chromosome 21q22 has been consistently associated with CHD. However, like the majority of the confirmed GWAS loci for CHD, how this locus affects CHD risk is not clear (Deloukas, Kanoni et al. 2013). In this study, rs28451064 was identified as a putative functional SNP at the locus. The minor CHD risk allele was found to show less protein binding and be associated with higher gene expression *in vitro*. This allele was also found to be associated with higher expression of the two closest upstream genes (*MRPS6* and *SLC5A3*) in a number of tissues. In agreement with previous studies no association between the lead SNP rs9982601 and CRFs for CHD was observed. A suggestive association between the risk locus and QT interval was observed, indicating that it may be impacting CHD risk through regulating the expression of the potassium channel subunit gene *KCNE2*, the closest downstream gene to the risk locus. However, while a suggestive association between rs9982601 and *KCNE2* expression in the aortic and tibial arteries was observed in the GTEX data set, the evidence for the risk locus being involved in the regulation of *MRPS6* and *SLC5A3* was more consistent.

How increased expression of any of the nearby genes might affect CHD risk is unclear. As a constituent part of the mitochondrial ribosome, the gene product of *MRPS6* plays a key role in the synthesis of the thirteen proteins encoded in mitochondrial DNA, all of which are involved in oxidative phosphorylation (Taanman 1999)(Figure 48). An important by-product of oxidative phosphorylation is the generation of reactive oxygen species (ROS) (Murphy 2009). Overproduction of ROS by dysfunctional mitochondria has been associated with multiple pro-atherogeneic consequences including the activation of inflammatory pathways and endothelial dysfunction, but whether this is a causal relationship remains to be determined (Wang and Tabas 2014). If so, it may be that increased expression of *MRPS6* caused by presence of the risk allele disrupts the translation of the genes encoded by the mitochondrial DNA, increasing the risk of mitochondrial dysfunction and ultimately oxidative stress.

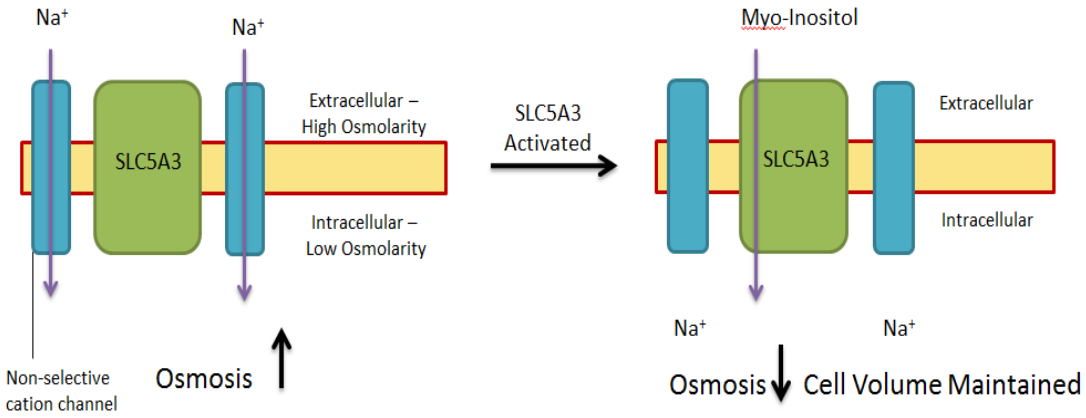
Figure 48: Simplified schematic diagram of the mitochondrion and key proteins involved in oxidative phosphorylation



The sequence of *SLC5A3* lies completely within that of *MRPS6*. The protein encoded by *SLC5A3* is a sodium-myoinositol co-transporter (SLC5A3). This transporter plays an important role in the maintenance of cell volume in response to hyperosmotic stress. As osmolarity of the extracellular fluid increases, non-selective cation channels are activated causing sodium ions to enter the cell, disrupting cellular ion homeostasis. In response solute carrier family proteins (including SLC5A3) are activated causing increased transport of small organic molecules such as myo-inositol, referred to as "compatible osmolytes" to replace inorganic ions (Brocker, Thompson et al. 2012) (Figure 49). Recent evidence has suggested that SLC5A3 is also involved in the response to hypotonic stress but this is much less well understood (Andronic, Shirakashi et al. 2015). Work in mouse models has found that SLC5A3 is involved in the development of the peripheral nervous system and respiratory gas exchange (Chau, Lee et al. 2005; Buccafusca, Venditti et al. 2008). From the current knowledge there is no obvious mechanism to link SLC5A3 with the pathogenesis of CHD. However, as it appears that there may be a number of pathways which contribute to atherosclerosis yet to be elucidated (indicated by the number of GWAS hits with unknown mechanisms), the involvement of SLC5A3 cannot be discounted. Alternatively, the

association between the risk locus and expression of this gene may simply be a consequence of its sharing an exon with *MRPS6* (Gardiner, Slavov et al. 2002).

Figure 49: Simplified schematic representation of the involvement of SLC5A3 in the cellular response to hyperosmotic stress



The relationship between *KCNE2* and CHD appears to be complex. The ion channel subunit encoded by *KCNE2* has a long established relationship with QT interval (and thus with the electrical activity of the ventricles). How this may relate to CHD risk is yet to be elucidated. Recently, deletion of the gene was found to promote spontaneous atherosclerotic lesions in mice (Lee, Nguyen et al. 2015). In addition, $Kcne2^{-/-}$ mice were also found to have raised LDL-cholesterol and impaired glucose tolerance, both pro-atherogenic characteristics (Hu, Kant et al. 2014). While results from mice are not directly translatable to humans, this does provide preliminary evidence of a causal relationship between *KCNE2* and CHD. However, only weak evidence for a relationship between the 21q22 risk locus and the gene was observed in this study. Of course, it may be that risk locus acts through multiple molecular pathways.

While a putative functional SNP (rs28451064) was identified in this study none of the transcription factors involved could be identified. Presence of rs28451064 minor (risk) allele was predicted to abolish a binding site for the VDR-RXR heterodimer transcription factor complex. A large scale analysis performed in lymphoblastoid cells found that expression of both *MRPS6* and *SLC5A3* increased in response to treatment with calcitriol (a bioactive form of vitamin D) (Ramagopalan, Heger et al. 2010). This indicates that the VDR pathway is involved in expression of these two genes. How expression of the two genes, rs28451064 and VDR might be related is unclear - higher expression of both genes is associated with the minor (risk) allele which is also predicted to abolish the VDR binding

site. Here, competitor EMSAs were performed in an attempt to investigate which transcription factors were binding the rs28451064-G probe in the EMSAs. However, while some binding of VDR was detected in the Huh-7 nuclear extract, no binding was observed in the HepG2 nuclear extract. It could be that the VDR protein is not expressed in this cell line or alternately that the protein is expressed but does not bind the consensus sequence probe under the conditions used in the experiment. Whether VDR is involved could be more directly studied by using a supershift EMSA. Here, a VDR-binding antibody is included in the reaction mix and binding is detected by the presence of a "supershift" band as this complex will move more slowly when run on a polyacrylamide gel. Mass spectroscopy could also be used to identify the proteins bound to the DNA probe (Stead, Keen et al. 2006). Once the DNA-protein complex has been run on the polyacrylamide gel, the relevant region can be excised and the proteins present identified. This is particularly useful when there is no candidate transcription factor as it essentially "hypothesis-free" in this regard.

This work has a number of limitations. The functional molecular assays were performed in hepatocyte carcinoma cell lines and these may not be the most appropriate cellular model. However, given that the mechanism through which this locus impacts upon risk remains obscure, it is not clear which cell type would serve as the most appropriate model. Moreover, the putative functional SNP, rs28451064, was found to lie in a DNAse I hypersensitivity site, transcription binding sites and have enhancer chromatin marks in HepG2 (hepatocyte) cells. This indicates that the SNP lies in open chromatin in this cell line and thus may be influencing gene expression. The luciferase assays were performed using the pGL3 promoter vector which contains a general SV40 bacterial promoter. It would have been preferable to use the promoter of either *MRPS6* or *SLC5A3*. However, the *MRPS6* promoter is not well characterised and attempts to clone the *SLC5A3* promoter sequence into the pGL3 basic vector were unsuccessful.

Both EMSA and the luciferase reporter assays are *in vitro* techniques which cannot account for chromatin state or long range interactions and thus only provide a guide as to the true situation occurring within the cell. In recent years a number of "chromatin capture" methods which enable DNA interactions in the native state to be studied have been developed. The first such was 3C (chromatin confirmation capture) which can be used to investigate whether two distant genomic sequences interact (Dekker, Rippe et al. 2002). DNA binding proteins are formaldehyde cross-linked to the DNA and then a restriction

enzyme digestion is performed. The cross-linked DNA is then ligated creating "ligation junctions". The cross-links are removed, the DNA purified and selected ligation junctions quantified using PCR with primers specific to the genomic loci being studied. Should a particular combination (e.g. sequences from the 21q22 risk locus and the *MRPS6* promoter) form more ligation junctions than proximal sequences, this demonstrates that a chromatin loop exists between these two sites (Simonis, Kooren et al. 2007). This technique is appropriate when investigating if two particular sequences interact. However, it may be desirable to investigate all the DNA sequences with which the 21q22 CHD risk locus is interacting. To do this 4C (chromosome confirmation capture (3C) on-chip) can be used (Simonis, Klous et al. 2006). This follows similar protocol to 3C, except that there is a second round of restriction enzyme digestion and ligation to get the DNA into a suitable form for PCR, as one of the DNA sequences in the ligation junction will be unknown.

In recent years, there has been the development of "genome editing" methodologies, most notably the CRISPR/Cas9 system (Ran, Hsu et al. 2013). This uses RNA-guided nucleases to introduce double-strand breaks in the DNA, activating the cell's non-homologous end joining pathway or the homology directed repair pathway (although this is only active in dividing cells) to repair this, allowing insertion or deletion of a small DNA fragment. This technology can be used to generate cell lines or model organisms with a particular genotype and thus the impact of a single variant (e.g. on gene expression) can be investigated.

6.4 Conclusion to chapter

Functional analysis of the 21q22 CHD risk locus was performed using both bioinformatics tools and *in vitro* functional assays. A putative functional SNP, rs28451064, was identified but the affected gene(s) and transcription factor(s) remain obscure. Future work should focus on identifying the pathway(s) through which this locus influences CHD risk, specifically the transcription factors and genomic loci involved.

7 General discussion

7.1 Overview

In addition to increasing the knowledge of the biology of common diseases such as CHD, investigating their genetics has two further aims, use in risk prediction and identification of therapeutic targets. For example, genetic studies of FH led to the identification of *PCSK9*, as a third gene (after *LDLR* and *APOB*) in which FH-causing mutations are found (Abifadel, Varret et al. 2003). Subsequent studies using animal models and in human subjects have led to the development of PCSK9 inhibitors. These have been found to reduce LDL-cholesterol by more than 50% and are currently undergoing phase 3 clinical trials (Marais, Kim et al. 2015). With the advent of the post-GWAS era, the use of genetics in risk prediction has focused on GSs, often using lead SNPs from GWAS-identified risk loci, with functional analysis to identify the molecular mechanisms involved also ongoing. Identifying the functional variant(s) can also be beneficial in risk prediction as this will refine the association, capturing all of the risk effect held by a particular locus. Furthermore, it should make the results more easily translatable between ethnic groups by removing the issue of differing levels of LD between the functional SNP and the proxy used in the GS.

7.2 Risk prediction

One of the major aims of this thesis was to investigate and optimise the use of a 19 SNP CHD risk GS in CHD risk prediction. This GS (and a smaller a 14 SNP GS derived from it) was found to have potential clinical in UK men (Chapter 3.2.3.2.3), in those with T2D (Chapter 4.2.3.3) and those of Afro-Caribbean origin (Chapter 3.2.3.2.4). Data from the South Asian cohorts was inconsistent as discussed in Chapter 3.3 and thus at present there is no evidence of clinical utility in this ethnic group. A kit to genotype these SNPs and ultimately provide an estimate of CHD risk incorporating the GS – the Cardiac Risk Prediction array (Randox Laboratories, Crumlin, Co Antrim, UK; Chapter 2.3.3) - is currently undergoing the "CE marking" procedure.

Overall, demonstrating the clinical utility of including genetic information in CHD risk prediction has proved challenging. The Joint British Societies' consensus recommendations for the prevention of CVD (JBS3) did not advocate the use of genetics risk in CVD prevention, as it was felt that the available evidence showed tools including genetic information performed more poorly than CRF based tools (2014). This had been underlined by the relatively disappointing performance of risk scores including GSs comprised of the variants identified in the CARDIoGRAMplusC4D meta-analysis (Deloukas,

Kanoni et al. 2013), which was identified in this thesis as the best source of risk variants for inclusion in CHD risk prediction. The results from the prospective Rotterdam study (de Vries, Kavousi et al. 2015) and the UCLEB consortium (Morris, Cooper et al. 2016) found very limited benefit in the population-wide inclusion of the GS in risk prediction, although improvements in both discrimination and reclassification were observed in a meta-analysis of six Swedish prospective cohorts (Ganna, Magnusson et al. 2013) and in the Malmo Diet and Cancer (MDC) study (Tada, Melander et al. 2016). Only the reclassification analysis in the MDC study was performed using the most recent guidelines however this was based on the US guidelines from the ACC/AHA (Goff, Lloyd-Jones et al. 2013) rather than the 10 % high risk cut-off recommended in the most recent NICE guidelines in the UK (2014). Therefore, direct comparison with the results observed here cannot be made. It has been suggested that due to the nature of case selection in GWASs, many of the variants identified in the CARDIoGRAMplusC4D meta-analysis are actually associated with CHD *survival* rather than an incident CHD event itself. This is supported by data from both the Rotterdam study and UCLEB consortium where the gene score was more strongly associated with prevalent rather than incident disease (de Vries, Kavousi et al. 2015),(Morris, Cooper et al. 2016). This indicates that the weightings used may not accurately reflect the impact of each variant on incident CHD risk and thus effect sizes obtained from a prospective cohort should be used. This strategy was used by Ganna, Magnusson et al. and a better performance was observed with the inclusion of the GS (Ganna, Magnusson et al. 2013). This issue is likely to be more pertinent for the CARDIoGRAMplusC4D SNPs whereas the majority of SNPs included in the 19 SNP GS (and indeed 14 SNP GS) have a clear mechanism of action to impact CHD and rather than purely CHD survival. This may partly explain the relatively strong performance of the updated 19 and 14 SNP GSs in NPHSII compared to the relatively poor performance of the CARDIoGRAMplusC4D GSs in much larger studies. Ultimately a large-scale well powered prospective study is required to alleviate the problem of survival bias in genetic association studies. If such data became available this could be used to provide the weights for the GS assessed in this thesis as it would be hypothesised that this would improve its performance.

It has also been suggested given the life-long nature of genetic risk, that it might be advantageous to identify those with high genetic CHD risk at a relatively young age (say early-middle age) but the data concerning this is inconsistent. Analysis performed in the

MDC found that risk estimates for the GSs assessed were higher in those below the median age (Tada, Melander et al. 2016) however there was no difference in reclassification or discrimination between those above and below the median age with the addition of the GS to QRISK2 in UCLEB (Morris, Cooper et al. 2016).

As with many risk factors and as observed herein, most GSs are normally distributed. Thus the majority of individuals have *intermediate* genetic risk, which will consequently have little impact on overall CHD risk. Therefore, the primary purpose of using GSs is to identify those individuals who have an *intermediate* risk according to the CRF score but who carry a *high* genetic CHD risk. Both (Ganna, Magnusson et al. 2013) and (Morris, Cooper et al. 2016) investigated the impact of restricting the inclusion genetic risk to those in the intermediate risk group (10-20 %, with prescription of statins at ≥20 % risk). They observed that this would postpone one event for 318 and 462 individuals screened in this manner respectively ((Ganna, Magnusson et al. 2013),(Morris, Cooper et al. 2016). When a similar strategy was assessed in the MI-GENES study (Kullo, Jouni et al. 2016), it was found that a greater proportion of those with a high genetic risk but intermediate CRF risk (i.e. those who would likely move into the high risk category using the combined CRF plus GS risk score) initiated statin treatment leading to lower LDL-cholesterol levels in this group compared to the controls. Larger trials with a long follow-up are required to determine if this finding can be replicated and whether it translates into a clinically relevant reduction in CHD risk. However, the results suggest that knowledge of high genetic risk may help individuals take appropriate steps to lower their overall CHD risk. Moreover, a meta-analysis of RCTs performed with statin therapy found that those in the top quintile of the GS (using 27 SNPs from the CARDIoGRAM GWAS (Schunkert, Konig et al. 2011)) had the greatest reduction in both relative and absolute in risk (Mega, Stitziel et al. 2015). This indicates that those with a high risk genetic risk may derive greater clinical benefit from statin use, increasing the potential benefit of identifying these individuals. The limited evidence available suggests that uptake of statins in those with >10% risk may be much lower than estimated in the NICE guidelines (Usher-Smith, Pritchard et al. 2015). Maximising statin uptake in those who are eligible is important to ensure the greatest benefit possible is derived from the guidelines and the results of the MI-GENES study show that provision of genetic risk may have a role to play in this. However, knowledge of genetic risk did not alter physical activity levels or dietary fat intake between the groups demonstrating that other strategies are required (such as the self-management

interventions used in the CoRDia trial) to help individuals adopt and sustain healthier lifestyles.

In this thesis the inclusion of genetic testing in CHD/CVD risk prediction has been within a clinical context only, most likely primary care such as general practice in the UK. Here the genetic risk information (as well as the CRF information) can be relayed to the individual by healthcare professionals to ensure correct interpretation of the results. The available evidence suggests that this model does not lead to unnecessary fatalism or false assurance (Collins, Wright et al. 2011). However, genetic testing for a variety of traits is now available to the general public in the UK most notably through "23andme"(Mullard 2015). This is despite the Food and Drug Administration ordering the cessation of health-related genetic testing services by 23andme in the USA (2014). It is unclear how, or indeed if, receiving genetic risk information outwith a clinical setting will impact on lifestyle choices and indeed how widespread uptake of the service will be.

The study of genetics of CHD in T2D presented here were mostly in agreement with previously published results (Qi, Parast et al. 2011; Qi, Qi et al. 2013). Efforts should now focus on investigating whether including the GSs in CHD in T2D risk prediction provides any additional benefit over using QRISK2 (as recommended in the guidelines), which could not be assessed herein. Additionally this will show if there is any benefit in using a specific CHD in T2D for those with T2D compared to a general CHD GS, as the results presented in this thesis also indicated that the 19 SNP GS was suitable for use in those with T2D.

7.3 Functional analysis

Functional analysis of two variants was performed in this thesis. One variant, rs10911021 associated with CHD in T2D, was found to be associated with HDL-cholesterol in T2D. Thus the analysis focussed on the relationship between the risk variant and this CRF. Counterintuitively, the CHD protective allele was associated with lower HDL-cholesterol and also large HDL particle traits, pointing to a potentially novel pathogenic mechanism, pending robust replication. The full implications of this association will only be understood when the relationship between HDL (particularly large HDL particles) and CHD is clarified. However, this does not preclude the functional analysis (possibly using a similar strategy as was applied to CHD risk locus 21q22) of the locus to establish the molecular pathway through which presence of the minor allele results in lower HDL-cholesterol in the diabetic state.

The other risk locus studied (chromosome 21q22) in this thesis was not associated with any CRFs for CHD. Therefore, a different strategy had to be employed. This focussed on identifying a candidate functional SNP and possible target genes. The phenomenon of LD which enables widespread coverage of the genome also creates the problem of having an entire locus associated with the trait, often with no obvious functional SNP. However, the availability of large datasets such as ENCODE and GTEX combined with *in vitro* functional assays was used to overcome this issue. This study identified rs28451064 as a candidate functional SNP and implicated the involvement of the genes *MRPS6* and *SLC5A3* (and possibly *KCNE2*). Developments such as chromatin capture techniques to analyse interactions between DNA elements and genome editing methods (to create model organisms/cell types with a specific genotype) will help to elucidate the molecular mechanisms involved.

The results from this thesis indicate that both variants investigated influence CHD risk through novel mechanisms. Such studies build on the available knowledge and can provide a fuller picture of the pathogenesis of CHD (Edwards, Beesley et al. 2013) and could ultimately identify new therapeutic targets.

8 References:

(2003). "The International HapMap Project." Nature **426**(6968): 789-796.

(2008). Type 2 Diabetes: National Clinical Guideline for Management in Primary and Secondary Care (Update). London.

(2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

(2013). "The Genotype-Tissue Expression (GTEx) project." Nature Genetics **45**(6): 580-585.

(2014). Lipid Modification: Cardiovascular Risk Assessment and the Modification of Blood Lipids for the Primary and Secondary Prevention of Cardiovascular Disease. London.

(2014). "Irresistible force meets immoveable object." Nature Biotechnology **32**(1): 1.

(2014). "Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3)." Heart **100 Suppl 2**: ii1-ii67.

(2014). "National Institute for Health and Care Excellence : Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease."

(2015). "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013." Lancet **385**(9963): 117-171.

(2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." Science **348**(6235): 648-660.

(2015). Type 2 diabetes in adults: management. London.

(2016). "Prospective Association of GLUL rs10911021 With Cardiovascular Morbidity and Mortality Among Individuals With Type 2 Diabetes: The Look AHEAD Study." Diabetes **65**(1): 297-302.

Abbott, G. W., F. Sesti, et al. (1999). "MiRP1 forms IKr potassium channels with HERG and is associated with cardiac arrhythmia." Cell **97**(2): 175-187.

Abecasis, G. R., D. Altshuler, et al. (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Abecasis, G. R., A. Auton, et al. (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

Abifadel, M., M. Varret, et al. (2003). "Mutations in PCSK9 cause autosomal dominant hypercholesterolemia." Nature Genetics **34**(2): 154-156.

Agyemang, C., J. Addo, et al. (2009). "Cardiovascular disease, diabetes and established risk factors among populations of sub-Saharan African descent in Europe: a literature review." Global Health **5**: 7.

Ahmad, N. and R. Bhopal (2005). "Is coronary heart disease rising in India? A systematic review based on ECG defined coronary heart disease." Heart **91**(6): 719-725.

Ahmad, O. S., J. A. Morris, et al. (2015). "A Mendelian randomization study of the effect of type-2 diabetes on coronary heart disease." Nat Commun **6**: 7060.

Alberti, K. G. and P. Z. Zimmet (1998). "New diagnostic criteria and classification of diabetes--again?" Diabetic Medicine **15**(7): 535-536.

Allain, B., R. Jarray, et al. (2012). "Neuropilin-1 regulates a new VEGF-induced gene, Phactr-1, which controls tubulogenesis and modulates lamellipodial dynamics in human endothelial cells." Cellular Signalling **24**(1): 214-223.

Almontashiri, N. A., M. Fan, et al. (2013). "Interferon-gamma activates expression of p15 and p16 regardless of 9p21.3 coronary artery disease risk genotype." Journal of the American College of Cardiology **61**(2): 143-147.

Andronic, J., R. Shirakashi, et al. (2015). "Hypotonic activation of the myo-inositol transporter SLC5A3 in HEK293 cells probed by cell volumetry, confocal and super-resolution microscopy." Plos One **10**(3): e0119990.

Angelakopoulou, A., T. Shah, 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.

Appelman, Y., B. B. van Rijn, et al. (2015). "Sex differences in cardiovascular risk factors and disease prevention." Atherosclerosis **241**(1): 211-218.

Assimes, T. L., H. Holm, et al. (2010). "Lack of Association Between the Trp719Arg Polymorphism in Kinesin-Like Protein-6 and Coronary Artery Disease in 19 Case-Control Studies." Journal of the American College of Cardiology **56**(19): 1552-1563.

Auton, A., L. D. Brooks, et al. (2015). "A global reference for human genetic variation." Nature **526**(7571): 68-74.

Baigent, C., L. Blackwell, 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.

Bannister, C. A., C. D. Poole, et al. (2014). "External validation of the UKPDS risk engine in incident type 2 diabetes: a need for new type 2 diabetes-specific risk equations." Diabetes Care **37**(2): 537-545.

Barlow, J., C. Wright, et al. (2002). "Self-management approaches for people with chronic conditions: a review." Patient Education and Counseling **48**(2): 177-187.

Barter, P. J., G. Brandrup-Wognsen, et al. (2010). "Effect of statins on HDL-C: a complex process unrelated to changes in LDL-C: analysis of the VOYAGER Database." Journal of Lipid Research **51**(6): 1546-1553.

Barzi, F., A. Patel, et al. (2007). "Cardiovascular risk prediction tools for populations in Asia." Journal of Epidemiology and Community Health **61**(2): 115-121.

Benjamini, Y. and D. Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency." Annals of Statistics **29**(4): 1165-1188.

Benn, M., B. G. Nordestgaard, et al. (2010). "PCSK9 R46L, low-density lipoprotein cholesterol levels, and risk of ischemic heart disease: 3 independent studies and meta-analyses." Journal of the American College of Cardiology **55**(25): 2833-2842.

Bennet, A. M., E. Di Angelantonio, et al. (2007). "Association of apolipoprotein E genotypes with lipid levels and coronary risk." JAMA **298**(11): 1300-1311.

Bernstein, B. E., E. Birney, et al. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Berry, G. T., J. J. Mallee, et al. (1995). "The human osmoregulatory Na+/myo-inositol cotransporter gene (SLC5A3): molecular cloning and localization to chromosome 21." Genomics **25**(2): 507-513.

Blekhman, R., O. Man, et al. (2008). "Natural selection on genes that underlie human disease susceptibility." Current Biology **18**(12): 883-889.

Blom, G. (1958). Statistical estimates and transformed beta-variables. New York, Wiley.

Boekholdt, S. M., N. R. Bijsterveld, et al. (2001). "Genetic variation in coagulation and fibrinolytic proteins and their relation with acute myocardial infarction: a systematic review." Circulation **104**(25): 3063-3068.

Boekholdt, S. M., R. J. Peters, et al. (2003). "Molecular variation at the apolipoprotein B gene locus in relation to lipids and cardiovascular disease: a systematic meta-analysis." Human Genetics **113**(5): 417-425.

Boekholdt, S. M., F. M. Sacks, et al. (2005). "Cholesteryl ester transfer protein TaqIB variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of

pravastatin treatment - Individual patient meta-analysis of 13,677 subjects." Circulation **111**(3): 278-287.

Boerwinkle, E. and L. Chan (1989). "A three codon insertion/deletion polymorphism in the signal peptide region of the human apolipoprotein B (APOB) gene directly typed by the polymerase chain reaction." Nucleic Acids Res **17**(10): 4003.

Bradshaw, D., P. Groenewald, et al. (2003). "Initial burden of disease estimates for South Africa, 2000." South African Medical Journal **93**(9): 682-688.

Braun-Dullaeus, R. C., M. J. Mann, et al. (1998). "Cell cycle progression: new therapeutic target for vascular proliferative disease." Circulation **98**(1): 82-89.

Brautbar, A., C. M. Ballantyne, et al. (2009). "Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the Atherosclerosis Risk in Communities study." Circ Cardiovasc Genet **2**(3): 279-285.

Brindle, P., J. Emberson, et al. (2003). "Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study." BMJ **327**(7426): 1267.

Brindle, P. M., A. McConnachie, et al. (2005). "The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study." British Journal of General Practice **55**(520): 838-845.

Brocker, C., D. C. Thompson, et al. (2012). "The role of hyperosmotic stress in inflammation and disease." Biomol Concepts **3**(4): 345-364.

Bruckert, E., G. Hayem, et al. (2005). "Mild to moderate muscular symptoms with high-dosage statin therapy in hyperlipidemic patients--the PRIMO study." Cardiovascular Drugs and Therapy **19**(6): 403-414.

Buccafusca, R., C. P. Venditti, et al. (2008). "Characterization of the null murine sodium/myo-inositol cotransporter 1 (Smit1 or Slc5a3) phenotype: myo-inositol rescue is independent of expression of its cognate mitochondrial ribosomal protein subunit 6 (Mrps6) gene and of phosphatidylinositol levels in neonatal brain." Molecular Genetics and Metabolism **95**(1-2): 81-95.

Burton, P. R., D. G. Clayton, et al. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

Businaro, R., A. Tagliani, et al. (2012). "Cellular and molecular players in the atherosclerotic plaque progression." Annals of the New York Academy of Sciences **1262**: 134-141.

Butterworth, A. S., P. S. Braund, et al. (2011). "Large-Scale Gene-Centric Analysis Identifies Novel Variants for Coronary Artery Disease." PLoS Genetics **7**(9).

Buysschaert, I., K. F. Carruthers, et al. (2010). "A variant at chromosome 9p21 is associated with recurrent myocardial infarction and cardiac death after acute coronary syndrome: The GRACE Genetics Study." European Heart Journal **31**(9): 1132-1141.

Capewell, S. and M. O'Flaherty (2011). "Rapid mortality falls after risk-factor changes in populations." Lancet **378**(9793): 752-753.

Casas, J. P., L. E. Bautista, et al. (2004). "Endothelial nitric oxide synthase genotype and ischemic heart disease: meta-analysis of 26 studies involving 23028 subjects." Circulation **109**(11): 1359-1365.

Casas, J. P., G. L. Cavalleri, et al. (2006). "Endothelial nitric oxide synthase gene polymorphisms and cardiovascular disease: a HuGE review." American Journal of Epidemiology **164**(10): 921-935.

Casas, J. P., J. Cooper, et al. (2006). "Investigating the genetic determinants of cardiovascular disease using candidate genes and meta-analysis of association studies." Annals of Human Genetics **70**(Pt 2): 145-169.

Chamnan, P., R. K. Simmons, et al. (2009). "Cardiovascular risk assessment scores for people with diabetes: a systematic review." Diab tologia **52**(10): 2001-2014.

Chau, J. F., M. K. Lee, et al. (2005). "Sodium/myo-inositol cotransporter-1 is essential for the development and function of the peripheral nerves." FASEB Journal **19**(13): 1887-1889.

Chen, X., L. Li, et al. (2014). "Prevalence of hypertension in rural areas of china: a meta-analysis of published studies." Plos One **9**(12): e115462.

Chiodini, B. D., S. Barlera, et al. (2003). "APO B gene polymorphisms and coronary artery disease: a meta-analysis." Atherosclerosis **167**(2): 355-366.

Chrvala, C. A., D. Sherr, et al. (2015). "Diabetes self-management education for adults with type 2 diabetes mellitus: A systematic review of the effect on glycemic control." Patient Education and Counseling.

Clarke, R., J. F. Peden, et al. (2009). "Genetic Variants Associated with Lp(a) Lipoprotein Level and Coronary Disease." New England Journal of Medicine **361**(26): 2518-2528.

Colhoun, H. M., P. M. McKeigue, et al. (2003). "Problems of reporting genetic associations with complex outcomes." Lancet **361**(9360): 865-872.

Collins, G. S. and D. G. Altman (2010). "An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study." BMJ **340**: c2442.

Collins, R. E., A. J. Wright, et al. (2011). "Impact of communicating personalized genetic risk information on perceived control over the risk: a systematic review." Genet Med **13**(4): 273-277.

Congrains, A., K. Kamide, et al. (2012). "Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B." Atherosclerosis **220**(2): 449-455.

Conroy, R. M., K. Pyorala, et al. (2003). "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project." European Heart Journal **24**(11): 987-1003.

Cooper, A. and N. O'Flynn (2008). "Risk assessment and lipid modification for primary and secondary prevention of cardiovascular disease: summary of NICE guidance." BMJ **336**(7655): 1246-1248.

Courtois, G., J. G. Morgan, et al. (1987). "Interaction of a liver-specific nuclear factor with the fibrinogen and alpha 1-antitrypsin promoters." Science **238**(4827): 688-692.

Crook, E. D., B. L. Clark, et al. (2003). "From 1960s Evans County Georgia to present-day Jackson, Mississippi: an exploration of the evolution of cardiovascular disease in African Americans." American Journal of the Medical Sciences **325**(6): 307-314.

Cunnington, M. S., M. Santibanez Koref, et al. (2010). "Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression." PLoS Genet **6**(4): e1000899.

Cuppen, E. (2007). "Genotyping by Allele-Specific Amplification (KASPar)." CSH Protoc **2007**: pdb prot4841.

D'Agostino, R. B., S. Grundy, et al. (2001). "Validation of the Framingham Coronary Heart Disease prediction scores - Results of a multiple ethnic groups investigation." Jama-Journal of the American Medical Association **286**(2): 180-187.

Dandona, S., A. F. Stewart, et al. (2010). "Genomics in coronary artery disease: past, present and future." Canadian Journal of Cardiology **26 Suppl A**: 56A-59A.

Davies, A. K., N. McGale, et al. (2015). "Effectiveness of a self-management intervention with personalised genetic and lifestyle-related risk information on coronary heart disease and diabetes-related risk in type 2 diabetes (CoRDia): study protocol for a randomised controlled trial." Trials **16**(1): 547.

Davies, R. W., G. A. Wells, 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.

Dawber, T. R., G. F. Meadors, et al. (1951). "Epidemiological approaches to heart disease: the Framingham Study." American Journal of Public Health and the Nations Health **41**(3): 279-281.

de Vries, P. S., M. Kavousi, et al. (2015). "Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: the Rotterdam Study." International Journal of Epidemiology **44**(2): 682-688.

Dekker, J., K. Rippe, et al. (2002). "Capturing chromosome conformation." Science **295**(5558): 1306-1311.

Deloukas, P., S. Kanoni, et al. (2013). "Large-scale association analysis identifies new risk loci for coronary artery disease." Nature Genetics **45**(1): 25-U52.

Di Angelantonio, E. and A. S. Butterworth (2012). "Clinical utility of genetic variants for cardiovascular risk prediction: a futile exercise or insufficient data?" Circ Cardiovasc Genet **5**(4): 387-390.

Di Angelantonio, E., N. Sarwar, et al. (2009). "Major lipids, apolipoproteins, and risk of vascular disease." JAMA **302**(18): 1993-2000.

Dickson, S. P., K. Wang, et al. (2010). "Rare variants create synthetic genome-wide associations." PLoS Biol **8**(1): e1000294.

Diehl, A. G. and A. P. Boyle (2016). "Deciphering ENCODE." Trends in Genetics.

Dudbridge, F. (2013). "Power and predictive accuracy of polygenic risk scores." PLoS Genet **9**(3): e1003348.

Edwards, S. L., J. Beesley, et al. (2013). "Beyond GWASs: illuminating the dark road from association to function." Am J Hum Genet **93**(5): 779-797.

Erbel, R., N. Lehmann, et al. (2014). "Progression of coronary artery calcification seems to be inevitable, but predictable - results of the Heinz Nixdorf Recall (HNR) study." European Heart Journal **35**(42): 2960-2971.

Erdmann, J., A. Grosshennig, et al. (2009). "New susceptibility locus for coronary artery disease on chromosome 3q22.3." Nature Genetics **41**(3): 280-282.

Erqou, S., C. C. Lee, et al. (2014). "Statins and glycaemic control in individuals with diabetes: a systematic review and meta-analysis." Diab tologia **57**(12): 2444-2452.

Erridge, C., J. Gracey, et al. (2013). "The 9p21 locus does not affect risk of coronary artery disease through induction of type 1 interferons." Journal of the American College of Cardiology **62**(15): 1376-1381.

Falk, E., M. Nakano, et al. (2013). "Update on acute coronary syndromes: the pathologists' view." European Heart Journal **34**(10): 719-728.

Fawcett, T. (2006). "An introduction to ROC analysis." Pattern Recognition Letters **27**(8): 861-874.

Finegold, J. A., C. H. Manisty, et al. (2014). "What proportion of symptomatic side effects in patients taking statins are genuinely caused by the drug? Systematic review of randomized placebo-controlled trials to aid individual patient choice." Eur J Prev Cardiol **21**(4): 464-474.

Folkersen, L., F. van't Hooft, et al. (2010). "Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease." Circ Cardiovasc Genet **3**(4): 365-373.

Ford, E. S., V. L. Roger, et al. (2014). "Challenges of ascertaining national trends in the incidence of coronary heart disease in the United States." J Am Heart Assoc **3**(6): e001097.

Fox, C. S., L. Sullivan, et al. (2004). "The significant effect of diabetes duration on coronary heart disease mortality: the Framingham Heart Study." Diabetes Care **27**(3): 704-708.

Franceschini, N., Y. Hu, et al. (2014). "Prospective associations of coronary heart disease loci in African Americans using the MetaboChip: the PAGE study." Plos One **9**(12): e113203.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.

Futema, M., R. A. Whittall, et al. (2013). "Analysis of the frequency and spectrum of mutations recognised to cause familial hypercholesterolaemia in routine clinical practice in a UK specialist hospital lipid clinic." Atherosclerosis **229**(1): 161-168.

Ganna, A., P. K. Magnusson, et al. (2013). "Multilocus genetic risk scores for coronary heart disease prediction." Arteriosclerosis, Thrombosis, and Vascular Biology **33**(9): 2267-2272.

Gardiner, K., D. Slavov, et al. (2002). "Annotation of human chromosome 21 for relevance to Down syndrome: gene structure and expression analysis." Genomics **79**(6): 833-843.

Gardner, D. S. and E. S. Tai (2012). "Clinical features and treatment of maturity onset diabetes of the young (MODY)." Diabetes Metab Syndr Obes **5**: 101-108.

Gauderman, W. and J. Morrison (2001). QUANTO documentation. (Technical report no. 157). Los Angeles, CA, Department of Preventive Medicine, University of Southern California.

Gerstein, M. B., A. Kundaje, et al. (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.

Glagov, S., E. Weisenberg, et al. (1987). "COMPENSATORY ENLARGEMENT OF HUMAN ATHEROSCLEROTIC CORONARY-ARTERIES." New England Journal of Medicine **316**(22): 1371-1375.

Goff, D. C., Jr., D. M. Lloyd-Jones, et al. (2013). "2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines." Circulation.

Goldacre, B. (2014). "Meta-analysis of side effects of statins shows need for trial transparency." BMJ **348**: g2940.

Goldacre, B. and L. Smeeth (2014). "Mass treatment with statins." BMJ **349**: g4745.

Grallert, H., J. Dupuis, et al. (2012). "Eight genetic loci associated with variation in lipoprotein-associated phospholipase A2 mass and activity and coronary heart disease: meta-analysis of genome-wide association studies from five community-based studies." European Heart Journal **33**(2): 238-251.

Guella, I., R. Asselta, et al. (2010). "Effects of PCSK9 genetic variants on plasma LDL cholesterol levels and risk of premature myocardial infarction in the Italian population." Journal of Lipid Research **51**(11): 3342-3349.

Hames, C. G. and K. J. Greenlund (1996). "Ethnicity and cardiovascular disease: The Evans County heart study." American Journal of the Medical Sciences **311**(3): 130-134.

Hannou, S. A., K. Wouters, et al. (2015). "Functional genomics of the CDKN2A/B locus in cardiovascular and metabolic disease: what have we learned from GWASs?" Trends Endocrinol Metab **26**(4): 176-184.

Harismendy, O., D. Notani, et al. (2011). "9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response." Nature **470**(7333): 264-268.

Harrison, S. C., J. A. Cooper, et al. (2012). "Association of a sequence variant in DAB2IP with coronary heart disease." European Heart Journal **33**(7): 881-888.

Hatano, S. (1989). "Changing CHD mortality and its causes in Japan during 1955-1985." International Journal of Epidemiology **18**(3 Suppl 1): S149-158.

Hellman, L. M. and M. G. Fried (2007). "Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions." Nat Protoc **2**(8): 1849-1861.

Hense, H. W., H. Schulte, et al. (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.

Hilden, J. and T. A. Gerds (2014). "A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index." Statistics in Medicine **33**(19): 3405-3414.

Hindorff, L. A., P. Sethupathy, et al. (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.

Hingorani, A. and S. Humphries (2005). "Nature's randomised trials." Lancet **366**(9501): 1906-1908.

Hippisley-Cox, J. and C. Coupland (2010). "Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database." BMJ **340**: c2197.

Hippisley-Cox, J., C. Coupland, et al. (2007). "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study." BMJ **335**(7611): 136.

Hippisley-Cox, J., C. Coupland, et al. (2008). "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2." BMJ **336**(7659): 1475-1482.

Holdt, L. M., F. Beutner, et al. (2010). "ANRIL expression is associated with atherosclerosis risk at chromosome 9p21." Arteriosclerosis, Thrombosis, and Vascular Biology **30**(3): 620-627.

Holman, R. R., S. K. Paul, et al. (2008). "10-year follow-up of intensive glucose control in type 2 diabetes." New England Journal of Medicine **359**(15): 1577-1589.

Holmes, M. V., F. W. Asselbergs, et al. (2015). "Mendelian randomization of blood lipids for coronary heart disease." European Heart Journal **36**(9): 539-550.

Hosmer, D. W. and S. Lemeshow (1980). "GOODNESS OF FIT TESTS FOR THE MULTIPLE LOGISTIC REGRESSION-MODEL." Communications in Statistics Part a-Theory and Methods **9**(10): 1043-1069.

Hu, Z., R. Kant, et al. (2014). "Kcne2 deletion creates a multisystem syndrome predisposing to sudden cardiac death." Circ Cardiovasc Genet **7**(1): 33-42.

Huang, Y., L. Gao, et al. (2014). "Epidemiology of dyslipidemia in Chinese adults: meta-analysis of prevalence, awareness, treatment, and control." Popul Health Metr **12**(1): 28.

Humphries, S. E., R. A. Whittall, et al. (2006). "Genetic causes of familial hypercholesterolaemia in patients in the UK: relation to plasma lipid levels and coronary heart disease risk." Journal of Medical Genetics **43**(12): 943-949.

Huxley, R., F. Barzi, et al. (2006). "Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies." BMJ **332**(7533): 73-78.

Imamura, M. and S. Maeda (2011). "Genetics of type 2 diabetes: the GWAS era and future perspectives [Review]." Endocrine Journal **58**(9): 723-739.

Ioannidis, J. P. (2008). "Why most discovered true associations are inflated." Epidemiology **19**(5): 640-648.

Ioannidis, J. P., E. E. Ntzani, et al. (2001). "Replication validity of genetic association studies." Nature Genetics **29**(3): 306-309.

Jansen, H., C. Loley, et al. (2015). "Genetic variants primarily associated with type 2 diabetes are related to coronary artery disease risk." Atherosclerosis **241**(2): 419-426.

Jarray, R., B. Allain, et al. (2011). "Depletion of the novel protein PHACTR-1 from human endothelial cells abolishes tube formation and induces cell death receptor apoptosis." Biochimie **93**(10): 1668-1675.

Joshi, P., S. Islam, et al. (2007). "Risk factors for early myocardial infarction in South Asians compared with individuals in other countries." JAMA **297**(3): 286-294.

Jousilahti, P., E. Vartiainen, et al. (1999). "Sex, age, cardiovascular risk factors, and coronary heart disease - A prospective follow-up study of 14 786 middle-aged men and women in Finland." Circulation **99**(9): 1165-1172.

Kamstrup, P. R., A. Tybjaerg-Hansen, et al. (2013). "Extreme lipoprotein(a) levels and improved cardiovascular risk prediction." Journal of the American College of Cardiology **61**(11): 1146-1156.

Kamstrup, P. R., A. Tybjaerg-Hansen, et al. (2009). "Genetically elevated lipoprotein(a) and increased risk of myocardial infarction." JAMA **301**(22): 2331-2339.

Kannel, W. B., A. Kagan, et al. (1961). "FACTORS OF RISK IN DEVELOPMENT OF CORONARY HEART DISEASE - 6-YEAR FOLLOW-UP EXPERIENCE." Annals of Internal Medicine **55**(1): 33-&.

Kathiresan, S., D. Altschuler, 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.

Kearney, P. M., L. Blackwell, et al. (2008). "Efficacy of cholesterol-lowering therapy in 18,686 people with diabetes in 14 randomised trials of statins: a meta-analysis." Lancet **371**(9607): 117-125.

Keavney, B., J. Danesh, et al. (2006). "Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'." International Journal of Epidemiology **35**(4): 935-943.

Keene, D., C. Price, et al. (2014). "Effect on cardiovascular risk of high density lipoprotein targeted drug treatments niacin, fibrates, and CETP inhibitors: meta-analysis of randomised controlled trials including 117,411 patients." BMJ **349**: g4379.

Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Research **12**(6): 996-1006.

Khamis, A., J. Palmen, et al. (2015). "Functional analysis of four LDLR 5'UTR and promoter variants in patients with familial hypercholesterolaemia." European Journal of Human Genetics **23**(6): 790-795.

Kitamura, A., S. Sato, et al. (2008). "Trends in the incidence of coronary heart disease and stroke and their risk factors in Japan, 1964 to 2003: the Akita-Osaka study." Journal of the American College of Cardiology **52**(1): 71-79.

Kraft, P. (2008). "Curses--winner's and otherwise--in genetic epidemiology." Epidemiology **19**(5): 649-651; discussion 657-648.

Krauss, R. M. (2004). "Lipids and lipoproteins in patients with type 2 diabetes." Diabetes Care **27**(6): 1496-1504.

Krishnan, M. N., G. Zachariah, et al. (2016). "Prevalence of coronary artery disease and its risk factors in Kerala, South India: a community-based cross-sectional study." BMC Cardiovasc Disord **16**(1): 12.

Kullo, I. J., H. Jouni, et al. (2016). "Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates: Effect on Low-Density Lipoprotein Cholesterol Levels (the MI-GENES Clinical Trial)." Circulation **133**(12): 1181-1188.

Kundaje, A., W. Meuleman, et al. (2015). "Integrative analysis of 111 reference human epigenomes." Nature **518**(7539): 317-330.

Kundu, R. K., Y. S. Aulchenko, et al. (2014). "PredictABEL: Assessment of Risk Prediction Models." R package version 1.2-2.

Kundu, S., Y. S. Aulchenko, et al. (2011). "PredictABEL: an R package for the assessment of risk prediction models." European Journal of Epidemiology **26**(4): 261-264.

Lango Allen, H., K. Estrada, et al. (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height." Nature **467**(7317): 832-838.

Larifla, L., K. E. Beaney, et al. (2016). "Influence of Genetic Risk Factors on Coronary Heart Disease Occurrence in Afro-Caribbeans." Canadian Journal of Cardiology.

Lawlor, D. A., R. M. Harbord, et al. (2008). "Mendelian randomization: using genes as instruments for making causal inferences in epidemiology." Statistics in Medicine **27**(8): 1133-1163.

Lawlor, D. A., R. M. Harbord, et al. (2008). "Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology." Statistics in Medicine **27**(8): 1133-1163.

Lee, S. M., D. Nguyen, et al. (2015). "Kcne2 deletion promotes atherosclerosis and diet-dependent sudden death." Journal of Molecular and Cellular Cardiology **87**: 148-151.

Leening, M. J., M. M. Vedder, et al. (2014). "Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide." Annals of Internal Medicine **160**(2): 122-131.

Lele, S. R., J. L. Keim, et al. (2014). "ResourceSelection: Resource Selection (Probability) Functions for Use-Availability Data." R package version 0.2-4.

Lettre, G., C. D. Palmer, et al. (2011). "Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project." PLoS Genet **7**(2): e1001300.

Li, N., H. Wang, et al. (2012). "Ethnic disparities in the clustering of risk factors for cardiovascular disease among the Kazakh, Uygur, Mongolian and Han populations of Xinjiang: a cross-sectional study." BMC Public Health **12**: 499.

Li, Y. H., M. M. Luke, et al. (2011). "Genetic Variants in the Apolipoprotein(a) Gene and Coronary Heart Disease." Circulation-Cardiovascular Genetics **4**(5): 565-573.

Libby, P., P. M. Ridker, et al. (2011). "Progress and challenges in translating the biology of atherosclerosis." Nature **473**(7347): 317-325.

Libby, P. and P. Theroux (2005). "Pathophysiology of coronary artery disease." Circulation **111**(25): 3481-3488.

Linsel-Nitschke, P., A. Gotz, 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).

Liu, Y., A. S. Hyde, et al. (2014). "Emerging regulatory paradigms in glutathione metabolism." Advances in Cancer Research **122**: 69-101.

Lloyd-Jones, D. M., J. C. Evans, et al. (2005). "Hypertension in adults across the age spectrum - Current outcomes and control in the community." Jama-Journal of the American Medical Association **294**(4): 466-472.

Lloyd-Jones, D. M., M. G. Larson, et al. (1999). "Lifetime risk of developing coronary heart disease." Lancet **353**(9147): 89-92.

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

Ma, Y. Q., W. H. Mei, et al. (2013). "Prevalence of hypertension in Chinese cities: a meta-analysis of published studies." Plos One **8**(3): e58302.

Madjid, M. and J. T. Willerson (2011). "Inflammatory markers in coronary heart disease." British Medical Bulletin **100**: 23-38.

Mannucci, E., M. Monami, et al. (2009). "Prevention of cardiovascular disease through glycemic control in type 2 diabetes: a meta-analysis of randomized clinical trials." Nutr Metab Cardiovasc Dis **19**(9): 604-612.

Marais, A. D., J. B. Kim, et al. (2015). "PCSK9 inhibition in LDL cholesterol reduction: genetics and therapeutic implications of very low plasma lipoprotein levels." Pharmacology and Therapeutics **145**: 58-66.

McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nat Rev Genet **9**(5): 356-369.

McPherson, R., A. Pertsemlidis, et al. (2007). "A common allele on chromosome 9 associated with coronary heart disease." Science **316**(5830): 1488-1491.

Mega, J. L., N. O. Stitziel, et al. (2015). "Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials." Lancet **385**(9984): 2264-2271.

Meister, A. (1973). "On the enzymology of amino acid transport." Science **180**(4081): 33-39.

Mihaylova, B., J. Emberson, et al. (2012). "The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials." Lancet **380**(9841): 581-590.

Miller, G. J., K. A. Bauer, et al. (1995). "The effects of quality and timing of venepuncture on markers of blood coagulation in healthy middle-aged men." Thrombosis and Haemostasis **73**(1): 82-86.

Mohan, V., R. Deepa, et al. (2001). "Prevalence of coronary artery disease and its relationship to lipids in a selected population in South India: The Chennai Urban Population Study (CUPS No. 5)." Journal of the American College of Cardiology **38**(3): 682-687.

Morgan, J., C. Carey, et al. (2004). "High-density lipoprotein subfractions and risk of coronary artery disease." Curr Atheroscler Rep **6**(5): 359-365.

Morgan, T. M., C. S. Coffey, et al. (2003). "Overestimation of genetic risks owing to small sample sizes in cardiovascular studies." Clinical Genetics **64**(1): 7-17.

Morris, D. R., J. V. Moxon, et al. (2012). "Meta-analysis of the association between transforming growth factor-beta polymorphisms and complications of coronary heart disease." Plos One **7**(5): e37878.

Morris, R. W., J. A. Cooper, et al. (2016). "Marginal role for 53 common genetic variants in cardiovascular disease prediction." Heart.

Mudau, M., A. Genis, et al. (2012). "Endothelial dysfunction: the early predictor of atherosclerosis." Cardiovascular Journal of Africa **23**(4).

Mullard, A. (2015). "23andMe sets sights on UK/Canada, signs up Genentech." Nature Biotechnology **33**(2): 119.

Murphy, M. P. (2009). "How mitochondria produce reactive oxygen species." Biochemical Journal **417**(1): 1-13.

Musunuru, K., M. Orho-Melander, et al. (2009). "Ion mobility analysis of lipoprotein subfractions identifies three independent axes of cardiovascular risk." Arteriosclerosis, Thrombosis, and Vascular Biology **29**(11): 1975-1980.

Nabel, E. G. and E. Braunwald (2012). "A tale of coronary artery disease and myocardial infarction." New England Journal of Medicine **366**(1): 54-63.

Nakashima, Y., H. Fujii, et al. (2007). "Early human atherosclerosis: accumulation of lipid and proteoglycans in intimal thickenings followed by macrophage infiltration." Arteriosclerosis, Thrombosis, and Vascular Biology **27**(5): 1159-1165.

Neph, S., J. Vierstra, et al. (2012). "An expansive human regulatory lexicon encoded in transcription factor footprints." Nature **489**(7414): 83-90.

Nichols, G. A. and C. E. Koro (2007). "Does statin therapy initiation increase the risk for myopathy? An observational study of 32,225 diabetic and nondiabetic patients." Clinical Therapeutics **29**(8): 1761-1770.

Nichols, M., N. Townsend, et al. (2014). "Cardiovascular disease in Europe 2014: epidemiological update." European Heart Journal **35**(42): 2950-2959.

Nikpay, M., A. Goel, et al. (2015). "A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease." Nature Genetics **47**(10): 1121-1130.

Niu, W., Y. Liu, et al. (2012). "Association of interleukin-6 circulating levels with coronary artery disease: a meta-analysis implementing mendelian randomization approach." International Journal of Cardiology **157**(2): 243-252.

Nyboe, J., G. Jensen, et al. (1991). "Smoking and the risk of first acute myocardial infarction." American Heart Journal **122**(2): 438-447.

Onen, C. L. (2013). "Epidemiology of ischaemic heart disease in sub-Saharan Africa." Cardiovasc J Afr **24**(2): 34-42.

Pasmant, E., I. Laurendeau, et al. (2007). "Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF." Cancer Research **67**(8): 3963-3969.

Paynter, N. P., D. I. Chasman, et al. (2009). "Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3." Annals of Internal Medicine **150**(2): 65-72.

Paynter, N. P., D. I. Chasman, et al. (2010). "Association between a literature-based genetic risk score and cardiovascular events in women." JAMA **303**(7): 631-637.

Pechlivanis, S., T. W. Muhleisen, et al. (2013). "Risk loci for coronary artery calcification replicated at 9p21 and 6q24 in the Heinz Nixdorf Recall Study." BMC Med Genet **14**: 23.

Pencina, M. J., R. B. D'Agostino, Sr., et al. (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; discussion 207-112.

Peng, P., J. Lian, et al. (2012). "Meta-analyses of KIF6 Trp719Arg in coronary heart disease and statin therapeutic effect." Plos One **7**(12): e50126.

Perk, J., G. De Backer, et al. (2012). "European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts)." European Heart Journal **33**(13): 1635-1701.

Pohjola-Sintonen, S., A. Rissanen, et al. (1998). "Family history as a risk factor of coronary heart disease in patients under 60 years of age." European Heart Journal **19**(2): 235-239.

Preiss, D., S. R. Seshasai, et al. (2011). "Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: a meta-analysis." JAMA **305**(24): 2556-2564.

Prudente, S., H. Shah, et al. (2015). "Genetic Variant at the GLUL Locus Predicts All-Cause Mortality in Patients With Type 2 Diabetes." Diabetes **64**(7): 2658-2663.

Qi, L., L. Parast, et al. (2011). "Genetic Susceptibility to Coronary Heart Disease in Type 2 Diabetes 3 Independent Studies." Journal of the American College of Cardiology **58**(25): 2675-2682.

Qi, L., Q. Qi, et al. (2013). "Association between a genetic variant related to glutamic acid metabolism and coronary heart disease in individuals with type 2 diabetes." JAMA **310**(8): 821-828.

R Core Team (2015). R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Raitoharju, E., I. Seppala, et al. (2011). "Common variation in the ADAM8 gene affects serum sADAM8 concentrations and the risk of myocardial infarction in two independent cohorts." Atherosclerosis **218**(1): 127-133.

Ramagopalan, S. V., A. Heger, et al. (2010). "A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution." Genome Research **20**(10): 1352-1360.

Ran, F. A., P. D. Hsu, et al. (2013). "Genome engineering using the CRISPR-Cas9 system." Nat Protoc **8**(11): 2281-2308.

Rana, A., R. J. de Souza, et al. (2014). "Cardiovascular risk among South Asians living in Canada: a systematic review and meta-analysis." CMAJ Open **2**(3): E183-191.

Ray, K. K., S. R. Seshasai, et al. (2009). "Effect of intensive control of glucose on cardiovascular outcomes and death in patients with diabetes mellitus: a meta-analysis of randomised controlled trials." Lancet **373**(9677): 1765-1772.

Ripatti, S., E. Tikkanen, et al. (2010). "A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses." Lancet **376**(9750): 1393-1400.

Robin, X., N. Turck, et al. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." BMC Bioinformatics **12**: 77.

Rose, G. (1981). "Strategy of prevention: lessons from cardiovascular disease." British Medical Journal (Clinical Research Ed.) **282**(6279): 1847-1851.

Rosenson, R. S., J. D. Otvos, et al. (2002). "Relations of lipoprotein subclass levels and low-density lipoprotein size to progression of coronary artery disease in the Pravastatin Limitation of Atherosclerosis in the Coronary Arteries (PLAC-I) trial." American Journal of Cardiology **90**(2): 89-94.

Ross, R. (1999). "Mechanisms of disease - Atherosclerosis - An inflammatory disease." New England Journal of Medicine **340**(2): 115-126.

Rumana, N., Y. Kita, et al. (2008). "Trend of increase in the incidence of acute myocardial infarction in a Japanese population: Takashima AMI Registry, 1990-2001." American Journal of Epidemiology **167**(11): 1358-1364.

Sagoo, G. S., I. Tatt, et al. (2008). "Seven lipoprotein lipase gene polymorphisms, lipid fractions, and coronary disease: a HuGE association review and meta-analysis." American Journal of Epidemiology **168**(11): 1233-1246.

Samani, N. J., J. Erdmann, et al. (2007). "Genomewide association analysis of coronary artery disease." New England Journal of Medicine **357**(5): 443-453.

Sarwar, N., J. Danesh, et al. (2007). "Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies." Circulation **115**(4): 450-458.

Sarwar, N., P. Gao, et al. (2010). "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies." Lancet **375**(9733): 2215-2222.

Sarwar, N., M. S. Sandhu, et al. (2010). "Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies." Lancet **375**(9726): 1634-1639.

Sattar, N., D. Preiss, et al. (2010). "Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials." Lancet **375**(9716): 735-742.

Scarborough, P., K. Wickramasinghe, et al. (2011). Trends in coronary heart disease, 1961-2011. London, British Heart Foundation.

Schildkraut, J. M., R. H. Myers, et al. (1989). "Coronary risk associated with age and sex of parental heart disease in the Framingham Study." American Journal of Cardiology **64**(10): 555-559.

Schunkert, H., J. Erdmann, et al. (2010). "Genetics of myocardial infarction: a progress report." European Heart Journal **31**(8): 918-925.

Schunkert, H., I. R. Konig, et al. (2011). "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease." Nature Genetics **43**(4): 333-U153.

Selvin, E., S. Marinopoulos, et al. (2004). "Meta-analysis: glycosylated hemoglobin and cardiovascular disease in diabetes mellitus." Annals of Internal Medicine **141**(6): 421-431.

Shah, B. R., J. Hwee, et al. (2015). "Diabetes self-management education is not associated with a reduction in long-term diabetes complications: an effectiveness study in an elderly population." Journal of Evaluation in Clinical Practice **21**(4): 656-661.

Shah, T., J. Engmann, et al. (2013). "Population genomics of cardiometabolic traits: design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium." Plos One **8**(8): e71345.

Shearman, A. M., J. A. Cooper, et al. (2006). "Estrogen receptor alpha gene variation is associated with risk of myocardial infarction in more than seven thousand men from five cohorts." Circulation Research **98**(5): 590-592.

Shen, G. Q., K. G. Abdullah, et al. (2009). "The TaqMan method for SNP genotyping." Methods in Molecular Biology **578**: 293-306.

Sherf, B. A., S. A. Navarro, et al. (1996). "Dual-luciferase reporter assay: an advanced co-reporter technology integrating firefly and Renilla luciferase assays." Promega Notes **57**: 2-8.

Simonis, M., P. Klous, et al. (2006). "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)." Nature Genetics **38**(11): 1348-1354.

Simonis, M., J. Kooren, et al. (2007). "An evaluation of 3C-based methods to capture DNA interactions." Nat Methods **4**(11): 895-901.

Simons, L. A., J. Simons, et al. (2003). "Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study." Medical Journal of Australia **178**(3): 113-116.

Smith, A. J., F. D'Aiuto, et al. (2008). "Association of serum interleukin-6 concentration with a functional IL6 -6331T>C polymorphism." Clinical Chemistry **54**(5): 841-850.

Smith, A. J., S. E. Humphries, et al. (2015). "Identifying functional noncoding variants from genome-wide association studies for cardiovascular disease and related traits." Current Opinion in Lipidology **26**(2): 120-126.

Smith, A. V. (2008). "Retrieving HapMap Data Using HapMart." CSH Protoc **2008**: pdb prot5026.

Smolina, K., F. L. Wright, et al. (2012). "Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study." BMJ **344**: d8059.

Snowden, C. B., P. M. McNamara, et al. (1982). "Predicting coronary heart disease in siblings--a multivariate assessment: the Framingham Heart Study." American Journal of Epidemiology **115**(2): 217-222.

Soler Artigas, M., D. W. Loth, et al. (2011). "Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function." Nature Genetics **43**(11): 1082-1090.

Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." Journal of Molecular Biology **98**(3): 503-517.

Speicher, M. R., J. B. Geigl, et al. (2010). "Effect of genome-wide association studies, direct-to-consumer genetic testing, and high-speed sequencing technologies on predictive genetic counselling for cancer risk." Lancet Oncol **11**(9): 890-898.

Stamler, J., A. R. Dyer, et al. (1993). "Relationship of baseline major risk factors to coronary and all-cause mortality, and to longevity: findings from long-term follow-up of Chicago cohorts." Cardiology **82**(2-3): 191-222.

Stampfer, M. J., F. B. Hu, et al. (2000). "Primary prevention of coronary heart disease in women through diet and lifestyle." New England Journal of Medicine **343**(1): 16-22.

StataCorp (2013). Stata Statistical Software: Release 13. College Station, TX, StataCorp LP.

Stead, J. A., J. N. Keen, et al. (2006). "The identification of nucleic acid-interacting proteins using a simple proteomics-based approach that directly incorporates the electrophoretic mobility shift assay." Mol Cell Proteomics **5**(9): 1697-1702.

Steinberg, D. (2002). "Atherogenesis in perspective: hypercholesterolemia and inflammation as partners in crime." Nature Medicine **8**(11): 1211-1217.

Steinsbekk, A., L. O. Rygg, et al. (2012). "Group based diabetes self-management education compared to routine treatment for people with type 2 diabetes mellitus. A systematic review with meta-analysis." Bmc Health Services Research **12**: 213.

Stephens, J. W., S. J. Hurel, et al. (2004). "An interaction between the interleukin-6 -174G>C gene variant and urinary protein excretion influences plasma oxidative stress in subjects with type 2 diabetes." Cardiovasc Diabetol **3**: 2.

Stettler, C., S. Allemann, et al. (2006). "Glycemic control and macrovascular disease in types 1 and 2 diabetes mellitus: Meta-analysis of randomized trials." American Heart Journal **152**(1): 27-38.

Stevens, R. J., V. Kothari, et al. (2001). "The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56)." Clin Sci (Lond) **101**(6): 671-679.

Strong, J. P., G. T. Malcom, et al. (1999). "Prevalence and extent of atherosclerosis in adolescents and young adults: implications for prevention from the Pathobiological Determinants of Atherosclerosis in Youth Study." JAMA **281**(8): 727-735.

Suzuki, T., M. Terasaki, et al. (2001). "Proteomic analysis of the mammalian mitochondrial ribosome. Identification of protein components in the 28 S small subunit." Journal of Biological Chemistry **276**(35): 33181-33195.

Swerdlow, D. I., D. Preiss, et al. (2015). "HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials." Lancet **385**(9965): 351-361.

Taanman, J. W. (1999). "The mitochondrial genome: structure, transcription, translation and replication." Biochimica et Biophysica Acta **1410**(2): 103-123.

Tabor, H. K., N. J. Risch, et al. (2002). "Candidate-gene approaches for studying complex genetic traits: practical considerations." Nat Rev Genet **3**(5): 391-397.

Tada, H., O. Melander, et al. (2016). "Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history." European Heart Journal **37**(6): 561-567.

Talmud, P. J., J. A. Cooper, et al. (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., M. Smart, et al. (2008). "ANGPTL4 E40K and T266M Effects on Plasma Triglyceride and HDL Levels, Postprandial Responses, and CHD Risk." Arteriosclerosis Thrombosis and Vascular Biology **28**(12): 2319-U2284.

Taylor, F., M. D. Huffman, et al. (2013). "Statins for the primary prevention of cardiovascular disease." Cochrane Database Syst Rev **1**: CD004816.

Teslovich, T. M., K. Musunuru, et al. (2010). "Biological, clinical and population relevance of 95 loci for blood lipids." Nature **466**(7307): 707-713.

Teupser, D., R. Baber, et al. (2010). "Genetic Regulation of Serum Phytosterol Levels and Risk of Coronary Artery Disease." Circulation-Cardiovascular Genetics **3**(4): 331-339.

Thompson, A., E. Di Angelantonio, et al. (2008). "Association of cholesteryl ester transfer protein genotypes with CETP mass and activity, lipid levels, and coronary risk." JAMA **299**(23): 2777-2788.

Thurman, R. E., E. Rynes, et al. (2012). "The accessible chromatin landscape of the human genome." Nature **489**(7414): 75-82.

Tillin, T., A. D. Hughes, et al. (2013). "The relationship between metabolic risk factors and incident cardiovascular disease in Europeans, South Asians, and African Caribbeans: SABRE (Southall and Brent Revisited) -- a prospective population-based study." Journal of the American College of Cardiology **61**(17): 1777-1786.

Tillin, T., A. D. Hughes, et al. (2014). "Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a U.K. tri-ethnic prospective cohort study (SABRE--Southall And Brent REvisited)." Heart **100**(1): 60-67.

Tunstall-Pedoe, H. and M. Woodward (2006). "By neglecting deprivation, cardiovascular risk scoring will exacerbate social gradients in disease." Heart **92**(3): 307-310.

Ueshima, H., K. Tatara, et al. (1987). "Declining mortality from ischemic heart disease and changes in coronary risk factors in Japan, 1956-1980." American Journal of Epidemiology **125**(1): 62-72.

Usher-Smith, J. A., J. Pritchard, et al. (2015). "Offering statins to a population attending health checks with a 10-year cardiovascular disease risk between 10% and 20." International Journal of Clinical Practice **69**(12): 1457-1464.

Vaarhorst, A. A., Y. Lu, et al. (2012). "Literature-based genetic risk scores for coronary heart disease: the Cardiovascular Registry Maastricht (CAREMA) prospective cohort study." Circ Cardiovasc Genet **5**(2): 202-209.

van der Harst, P., W. Zhang, et al. (2012). "Seventy-five genetic loci influencing the human red blood cell." Nature **492**(7429): 369-375.

VanderLaan, P. A., C. A. Reardon, et al. (2004). "Site specificity of atherosclerosis: site-selective responses to atherosclerotic modulators." Arteriosclerosis, Thrombosis, and Vascular Biology **24**(1): 12-22.

Velez Edwards, D. R., A. C. Naj, et al. (2013). "Gene-environment interactions and obesity traits among postmenopausal African-American and Hispanic women in the Women's Health Initiative SHARe Study." Human Genetics **132**(3): 323-336.

Viechtbauer, W. (2010). "Conducting Meta-Analyses in R with the metafor Package." Journal of Statistical Software **36**(3): 1-48.

Vogel, U., M. K. Jensen, et al. (2011). "The NFKB1 ATTG ins/del polymorphism and risk of coronary heart disease in three independent populations." Atherosclerosis **219**(1): 200-204.

Voight, B. F., H. M. Kang, et al. (2012). "The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits." PLoS Genet **8**(8): e1002793.

Voight, B. F., L. J. Scott, et al. (2010). "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis." Nature Genetics **42**(7): 579-589.

Voko, Z., Z. Bereczky, et al. (2007). "Factor XIIIVa134Leu variant protects against coronary artery disease - A meta-analysis." Thrombosis and Haemostasis **97**(3): 458-463.

Wang, Y. and I. Tabas (2014). "Emerging roles of mitochondria ROS in atherosclerotic lesions: causation or association?" Journal of Atherosclerosis and Thrombosis **21**(5): 381-390.

Wang, Y., J. Zheng, et al. (2012). "Association between the Interleukin 10-1082G > A polymorphism and coronary heart disease risk in a Caucasian population: a meta-analysis." International Journal of Immunogenetics **39**(2): 144-150.

Wang, Y. Y., W. L. Zhang, et al. (2011). "Genetic variants of the monocyte chemoattractant protein-1 gene and its receptor CCR2 and risk of coronary artery disease: A meta-analysis." Atherosclerosis **219**(1): 224-230.

Wannamethee, S. G., A. G. Shaper, et al. (2011). "Impact of diabetes on cardiovascular disease risk and all-cause mortality in older men: influence of age at onset, diabetes duration, and established and novel risk factors." Archives of Internal Medicine **171**(5): 404-410.

Ward, L. D. and M. Kellis (2012). "HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants." Nucleic Acids Res **40**(Database issue): D930-934.

Weber, C. and H. Noels (2011). "Atherosclerosis: current pathogenesis and therapeutic options." Nature Medicine **17**(11): 1410-1422.

Welter, D., J. MacArthur, et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." Nucleic Acids Res **42**(Database issue): D1001-1006.

Wensley, F., P. Gao, et al. (2011). "Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data." BMJ **342**: d548.

Wheeler, J. G., B. D. Keavney, et al. (2004). "Four paraoxonase gene polymorphisms in 11,212 cases of coronary heart disease and 12,786 controls: meta-analysis of 43 studies." Lancet **363**(9410): 689-695.

Wickham, H. (2009). ggplot2 : elegant graphics for data analysis. New York ; London, Springer Science + Business Media.

Wienke, A., A. M. Herskind, et al. (2005). "The heritability of CHD mortality in Danish twins after controlling for smoking and BMI." Twin Research and Human Genetics **8**(1): 53-59.

Wild, S. and P. McKeigue (1997). "Cross sectional analysis of mortality by country of birth in England and Wales, 1970-92." BMJ **314**(7082): 705-710.

Willer, C. J., E. M. Schmidt, et al. (2013). "Discovery and refinement of loci associated with lipid levels." Nature Genetics **45**(11): 1274-1283.

Wilson, P. W., R. B. D'Agostino, et al. (1998). "Prediction of coronary heart disease using risk factor categories." Circulation **97**(18): 1837-1847.

Wittrup, H. H., A. Tybjaerg-Hansen, et al. (1999). "Lipoprotein lipase mutations, plasma lipids and lipoproteins, and risk of ischemic heart disease - A meta-analysis." Circulation **99**(22): 2901-2907.

Woodward, M., P. Brindle, et al. (2007). "Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC)." Heart **93**(2): 172-176.

Woodward, M., X. Zhang, et al. (2003). "The effects of diabetes on the risks of major cardiovascular diseases and death in the Asia-Pacific region." Diabetes Care **26**(2): 360-366.

Wu, L. and K. G. Parhofer (2014). "Diabetic dyslipidemia." Metabolism **63**(12): 1469-1479.

Wu, O., N. Bayoumi, et al. (2008). "ABO(H) blood groups and vascular disease: a systematic review and meta-analysis." J Thromb Haemost **6**(1): 62-69.

Wu, Z. J., Y. Q. Lou, et al. (2012). "The Pro12Ala Polymorphism in the Peroxisome Proliferator-Activated Receptor Gamma-2 Gene (PPAR gamma 2) Is Associated with Increased Risk of Coronary Artery Disease: A Meta-Analysis." Plos One **7**(12).

Wurtz, P., Q. Wang, et al. (2016). "Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase." Journal of the American College of Cardiology **67**(10): 1200-1210.

Xu, M., P. Sham, et al. (2010). "A1166C genetic variation of the angiotensin II type I receptor gene and susceptibility to coronary heart disease: collaborative of 53 studies with 20,435 cases and 23,674 controls." Atherosclerosis **213**(1): 191-199.

Xuan, C., X. Y. Bai, et al. (2011). "Association Between Polymorphism of Methylenetetrahydrofolate Reductase (MTHFR) C677T and Risk of Myocardial Infarction: A Meta-analysis for 8,140 Cases and 10,522 Controls." Archives of Medical Research **42**(8): 677-685.

Yang, G., L. Kong, et al. (2008). "Emergence of chronic non-communicable diseases in China." Lancet **372**(9650): 1697-1705.

Yoshida, K., J. Hirokawa, et al. (1995). "Weakened cellular scavenging activity against oxidative stress in diabetes mellitus: regulation of glutathione synthesis and efflux." Diab tologia **38**(2): 201-210.

Yuan, X., D. Waterworth, et al. (2008). "Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes." Am J Hum Genet **83**(4): 520-528.

Zafarmand, M. H., Y. T. van der Schouw, et al. (2008). "The M235T Polymorphism in the AGT Gene and CHD Risk: Evidence of a Hardy-Weinberg Equilibrium Violation and Publication Bias in a Meta-Analysis." Plos One **3**(6).

Zaman, M. J., P. Philipson, et al. (2013). "South Asians and coronary disease: is there discordance between effects on incidence and prognosis?" Heart **99**(10): 729-736.

Zanoni, P., S. A. Khetarpal, et al. (2016). "Rare variant in scavenger receptor BI raises HDL cholesterol and increases risk of coronary heart disease." Science **351**(6278): 1166-1171.

Zdravkovic, S., A. Wienke, et al. (2002). "Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins." Journal of Internal Medicine **252**(3): 247-254.

Zeggini, E. and J. P. A. Ioannidis (2009). "Meta-analysis in genome-wide association studies." Pharmacogenomics **10**(2): 191-201.

Zeggini, E., L. J. Scott, et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nature Genetics **40**(5): 638-645.

Zeggini, E., M. N. Weedon, et al. (2007). "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes." Science **316**(5829): 1336-1341.

Zhang, H. F., S. L. Xie, et al. (2011). "Tumor necrosis factor-alpha G-308A gene polymorphism and coronary heart disease susceptibility: An updated meta-analysis." Thrombosis Research **127**(5): 400-405.

Zhang, X., J. Zhao, et al. (2015). "Effects of intensive glycemic control in ocular complications in patients with type 2 diabetes: a meta-analysis of randomized clinical trials." Endocrine **49**(1): 78-89.

Zhang, Y., W. S. Post, et al. (2011). "Electrocardiographic QT interval and mortality: a meta-analysis." Epidemiology **22**(5): 660-670.

Zhou, L., B. Xi, et al. (2012). "Association between adiponectin gene polymorphisms and coronary artery disease across different populations." Thrombosis Research **130**(1): 52-57.

Zintzaras, E., G. Raman, et al. (2008). "Angiotensin-converting enzyme insertion/deletion gene polymorphic variant as a marker of coronary artery disease: a meta-analysis." Archives of Internal Medicine **168**(10): 1077-1089.

Zuo, H., Z. Shi, et al. (2014). "Prevalence, trends and risk factors for the diabetes epidemic in China: a systematic review and meta-analysis." Diabetes Research and Clinical Practice **104**(1): 63-72.